

DATA DRIVEN DECISION MAKING IN DIGITAL WORLD



JUMIN KAMKI



JUMIN KAMKI





Notion Press

Old No. 38, New No. 6 McNichols Road, Chetpet Chennai - 600 031

First Published by Notion Press 2016 Copyright © Jumin Kamki 2016 All Rights Reserved.

ISBN 978-1-946556-20-2

This book has been published with all efforts taken to make the material errorfree after the consent of the author. However, the author and the publisher do not assume and hereby disclaim any liability to any party for any loss, damage, or disruption caused by errors or omissions, whether such errors or omissions result from negligence, accident, or any other cause.

No part of this book may be used, reproduced in any manner whatsoever without written permission from the author, except in the case of brief quotations embodied in critical articles and reviews.

Dedicated

То

My Wife Kasturika Saikia

&

My Son Eevan aka Minyansh Kamki

ABOUT THE AUTHOR JUMIN KAMKI

PGDM from IIM Ahmedabad with more than 10 years of experience in Retail and Ecommerce industries across different verticals. Keen interest in Analytics and Data Driven decision making, he has been spearheading Analytics initiatives in multiple organizations focused on building digital capabilities to transform the organization and bringing strategic shift towards data driven decision making. His area of interest includes Game Theory, Machine Learning, Digital Marketing and Economics. He has worked with multiple organization – Reliance Industries Limited, Aditya Birla Retail Ltd, Utsavfashion.com, Askme Group and Tata Insights and Quants in his professional career. More about him on

Blog: www.juminkamki.com

LinkedIn: https://in.linkedin.com/in/juminkamki

CONTENTS

<u>Preface</u> <u>Acknowledgement</u>

Chapter - I : DATA ANALYTICS FOUNDATION

Section - I : MEASURING CENTRAL TENDENCY AND DISPERSION Section - II : PROBABILITY THEORY Section - III : SAMPLING AND HYPOTHESIS TESTING Section - IV : LINEAR PROGRAMMING

Chapter - II : ANALYTICS SYSTEM

Section - I : BUSINESS INTELLIGENCE SYSTEM Section - II : R BASICS

Chapter - III : WEB ANALYTICS

Section - I : GOOGLE ANALYTICS

Chapter - IV : CUSTOMER ANALYTICS

Section - I : CUSTOMER ANALYTICS

Chapter - V : DIGITAL MARKETING

Section - I : DIGITAL MARKETING BASIC Section- II : DIGITAL CHANNEL OPTIMIZATION

Chapter - VI : FORECASTING AND PREDICTION

Section - I : REGRESSION Section - II : TIME SERIES FORECASTING

Chapter - VII : INVENTORY MANAGEMENT

Section - I : INVENTORY MODEL Section -II : TRANSPORTATION PROBLEM **Chapter - VIII :** ADVANCED TOPICS

PREFACE

The genesis of this book is an idea that comes to me while interacting with many young aspirants who want to build career in analytics. Most of the candidates and colleagues I interacted have good knowledge on the certain areas of the analytics but there is gap in understanding the basic knowledge of the analytics from overall analytics domain perspective. This book is an attempt to provide a comprehensive guide to the readers who want to build a career in analytics. As such analytics is a vast domain with each industry having different practices due to demand of the system and processes of that industry but there is underlying common thread in analytics that cut across all industries that is the digital side of the analytics. The orientation of the book is more towards ecommerce industry and retail industry but it is equally useful for those in insurance and finance domain doing digital analytics.

The book is comprehensive in sense that all areas of analytics being covered to some extent. Intent is to provide as many aspect of digital analytics as possible in a single book; such that a reader should not need to consult any other book while going through this book. Respecting the freedom and style of each analytics person, I have not prescribe a single model to be followed; It has been left for reader to explore each topic and figure out their own model and style from knowledge gained from the book.

My intent was to create something that a person without prior knowledge of analytics can enter the book and come out as an expert in the digital analytics at the end of the book. This book is intended for people with no knowledge of digital analytics and people with some level of digital analytics and want to enhance it. Some people can use this book as a reference in their work as well. This book is not intended for someone looking for advance topic in analytics such as big data, machine learning, and internet of thing and so on.

Flow of the Book

The beauty of this book is that one can read a chapter as an independent unit because each chapter is self-contain and start with very basic understanding of the concept and then it has been taken to a higher level of understanding. However there are many interconnection at the overall scheme of the book. For example you would need analytics foundation chapter to understand concept of probability distribution and central limit theorem which has been used in say Inventory Management chapter. Similarly you would need knowledge of R in Analytics System Chapter to use R codes in Customer Analytics chapter. Therefore depending on the level of knowledge one has, she can pick any chapter and gain something out of the chapter.

My intent of covering vast topic is to make reader without knowledge of analytics or with little analytics knowledge, a rock star in Digital Analytics.

CHAPTER I: First chapter of the book provides one with basic foundation of the Analytics worlds that is the statistical knowledge. Section I of the chapter deals with the concept of central tendencies- mean, mode & median, variance & standard deviation, correlation and graphs. Section II of the chapter talks about probability concepts, probability distribution and central limit theorem. Section III deals with the hypothesis testing, chi-square and ANOVA. Section IV provides concept of optimization techniques called Linear Programming.

CHAPTER II: Second chapter of book is about how to build and use analytics system. As an example of the analytics system in section I, I have used Pentaho stack to provide overall understanding of concept of Extraction, Transformation & Loading (ETL), Online Analytics Processing (OLAP) cube and Dashboard. Section II provides understanding of Analytics System called R which is open source tool for statistical analysis.

CHAPTER III: Third chapter is extension of second chapter as it deals with another analytics system known as Google Analytics. Google Analytics is different from previous system because it deals with collection and reporting of website data.

CHAPTER IV: Fourth chapter deals talks about ways and means to analyze online and offline customer data. It not only provide theoretical concept but R codes with examples for self-practice.

CHAPTER V: Fifth chapter provides rigorous understanding of the digital marketing concepts and tools in the section I. Using concept from section I, the optimization of the digital channel and attribution modeling is discussed in section II.

CHAPTER VI: Sixth chapter is all about prediction and forecasting. Section I deals with predictive models such as regression and section II talks about

time series forecasting including smoothening and ARIMA model.

CHAPTER VII: Seventh chapter deals with the operation part of the retail and ecommerce companies. Section I is all about optimizing inventory and section II is about optimizing the transportation. Both section combine provides ammunition to optimize overall supply chain optimization.

CHAPTER VIII: Eight chapter is all about more advanced topic which are not covered in detail in this book but give a glimpse of the subject matter so that interested readers can pick up from here to specific book on the subject.

ACKNOWLEDGEMENT

I would like to thank my colleagues who have spent their valuable time in painstakingly going through the manuscript and providing me with their critical inputs. Without their support and feedback the book would not have been as interesting and useful as it is now. Thank to Mr. Soyinka Majumdar, Mr Deepak Verma, Mr Anuj Dhandkar, Ms Priyamvada Joshi and Ms Shruti Gupta for their inputs and feedbacks. I would also like to thank all those who have been working closely with me in last 10 years and those who have help me developed as a good people manager. Last but not least I would like thank my wife Kasturika Saika for constant source of encouragement and my son Minyansh for sparing his play time all these days.

Chapter - I

DATA ANALYTICS FOUNDATION

"Know thy self, know thy enemy. A thousand battles, a thousand victories."

-by Sun Tzu

During height of the World War II, Germany and its allies were winning most of the battles --entire Europe except British Isles & Russia was overrun by German Army, North Africa was overrun by German army under one of the greatest general of all time Erwin Rommel (also known as Desert Fox), Japan was winning China and South East Asian battles. Germany's U boat was sinking the ship carrying cargo and weapon supplies to Britain from North America especially US. To maintain steady supply of oil and food Germany was planning to take over the Suez Canal from allies. All the message from Hitler and German army was coded using machine called Enigma; and the messages can be decipher at the receiver end using specialized machine known to German only. British scientists working under Alan Turing in Bletchley Park could able to decipher the German coded message. The message sent to the U-Boats and Rommel was decoded, its help allies knowing the exact location and the plans from German army; finally U-boats were neutralized in Atlantic Ocean and Rommel's Panzer Army was defeated in El Alamein in North Africa. It is no exaggeration to say that the decoding of Enigma message help Allies turn the tide of the World War II.

There are numerous other instances in the history where the great victories were achieved by clever use of information even though victorious armies were smaller and had lesser resources. As old adage goes *Knowledge is power*. Since the dawn of the civilization the information has been used for the advancement of the society through application of the data for the decision making in every sphere of social life. The ancient civilization like Greek and Indian civilization developed and used number system to solve their daily problem related to the business, transportation, town planning, military and so on. Along with the advancement of the science and technology the rate of usage of the data increased every generation. The amount of the data being processed has been increasing at compounded rate since the advent of the computer

system. The current mobile phone processing power is equivalent to the mainframe machine of the say 30 years ago. The internet has revolutionized the way people communicate and interact with the business on day to day life. The ecommerce has revolutionized the way people buy goods and services from the internet. It not only offers convenience but also reduces the price of the goods and services. This is mainly because of the disintermediation of the certain layers in the supply chain. It also increases the competition among the seller to offer goods and services at minimum prices because the cost of the information search in the internet has just reduces to very low level. At any instant the buyer can compare the price, specification and services for the products from different seller in the internet on the click of mouse.

It would not be wrong to say that there is no company in the world which doesn't have website irrespective of whether it sells a goods or a services through internet. At present any new company would create a website before creating a physical infrastructure. The purpose of having a website may differ from the company to company. Some company would have website just for the information to the potential customer or for client to look at company's information; they would not even bother to capture the basic information like email or phone number. They are not actively using website for lead generation. The next set of the companies are those who list the services and products so that potential customer can browse through its listing to gather more information about the products or services; and later those customer may buy from the physical retail store or from other ecommerce website which list the company's products. These sites don't have facilities to do the transaction but objective is to increase the offline transaction indirectly. The third sets of companies are those selling products or services through the website. They have facilities to transact through website. This site will invariably have listing of the products, facilities to do online payment, space to enter personal information for the billing and shipping, tracking the delivery of the products etc.

Irrespective of the type of the sites there are certain goals for each types of the site. To understand whether site is performing as per the designed objectives of the company there is need to measure the user's behavior in the site. This is invariably done through web analytics tools which capture the page level information, sources of traffic, exits and entry pages, number of goals achieved, the time spend on site, bounces from site and so on. As the need for understanding the customer behavior better increases, the companies capture each and every events of the site to track the behavior of the customer. This lead to the explosion of the data that is being captured.

Along with the Web Analytics, the companies will have system such as Enterprises Resources Planning (ERP) system or Customer Relationship Management (CRM) system or any other system to capture the transaction information. These data are mostly captured in RDBMS database. Then there are systems to support the site with the data related to products catalogue, their inventory, products features etc. There is needed to keep these data from disparate system into one repository for better data analytics. The companies invest on data warehousing to have one repository of the data required for the reporting and analysis purpose. Along with the team doing the data analysis and the end users, the final decision makers need to understand the science behind the analysis to understand the fit fall of the decision using the data. The need to understand the theory and processes of data analysis is felt at all level for the better decision making of all companies.

The companies are increasingly turning to data for their decision making. The automation of the processes through intelligence system reduces the requirement of the manpower. The building intelligence from the information need better data input and data analytics. The companies are investing on the data analytics to understand their customer, predict the sales and inventory, understand competitors and so on. The investment in data analytics is no longer a luxury but a necessity for any companies who want to survive in the market.

Having good data platform is no longer a luxury but a necessity for all the companies big or small. The companies are deriving competitive advantages through data analytics. The judgment based on gut feel no longer work in data driven environment. The understanding of data and its output is no longer reserve for the people at the top and analytics department. Every single person in the company has to be data savvy. The technology and analytics department is enabler for the real decision makers to look for insight from the data.

The focus of this chapter is to provide readers the basic foundation of the data analytics. The chapter is divided into four sections covering all theoretical foundation for the sound data analytics.

Section I: This section is meant for the understanding of the central tendency of the data - How to measure mean, mode and median. It also covers dispersion from central tendency – variance and standard deviation, correlation and how to create graphs in excel.

Section II: This section is to make readers understand the basic of probability, probability theory, probability distribution and central limit theorem.

Section III: This section covers hypothesis testing under different condition, Chi-Square test and ANOVA

Section IV: This section is covers linear programming using graphical method and using MS Excel.

Section - I

MEASURING CENTRAL TENDENCY AND DISPERSION

In our day to day routine it is common tendency to calculate the averages of data such as salary of group of people, rate of sales of last seven days, average temperature of the city, and average mileage of car, average monthly spend on the grocery, time taken to reach office and so on. Common thread running through all these measurement is one number or a concept known as mean or average of numbers. The mean gives approximate central value of the parameter measured over the period or across other dimensions. These numbers are quoted and used for taking daily decision like before leaving office you look at your watch and think on an average it takes say 45 minutes to reach office so you leave office at 8 am to reach office to attend at 9 am meeting, keeping a buffer of 15 minutes. This buffer of 15 minutes has been kept because the average gives you a number which approximately says 50% of the days you take more than 45 minutes to reach office and 50% of the days you take less than 45 minutes to reach office. The 15 minutes buffer will take care of the bad days when you take more than 45 minutes to reach office. The buffer can be estimate more precisely if you know the dispersion of the time taken around the average of the time taken to reach office.

The average number is used so commonly that people hardly think about the implication of the decision being taken on the basis of the mean. As decision makers in the daily business we have to be careful in using the number on the face value, the average can hide many inconvenient truth. This section has been designed to takes readers through the process of the measurement of the central tendency for the sample set of data. The measurement of central tendency doesn't complete without understanding the dispersion of the numbers around its mean, hence the readers will learn how to calculate the variance and standard deviation of the data set and what is the meaning of the standard deviation. We will also learn how to measure the correlation of the two set of

data and how to make sense out of the number. Towards the end of the section we will learn basic graphs and how to draw graphs using excel or spreadsheet.

1.1.1 Measuring the Central Tendency

The measurement of the central tendency typically means finding the central value around which a group of data tends to cluster around. There are three main measure of the central tendency –mean, median and mode. The mean is commonly used in our analysis as it consider all the data point, however median and mode has their own properties which makes them attractive in certain cases where mean may not be good measure. We will explore the pros and cons of each measure of central tendency.

1.1.2 Mean

The simple mean is measures as the central point of all the points in the space by assigning all point equal weightage in the calculation of the central point.

The mean is calculates as $\mu = [x_1 + x_2 + ... + x_n]/N$. Mean is commonly denoted by μ and $x_1, x_2..x_n$ are set of N numbers.

Simple $\mu = \sum D_t / N$

Where $\sum D_t$ is the sum of all data point and N is the number of the data point.

For example, if there are four person with age 22, 30, 35 and 45 years in a team, the simple mean of the age of the team is just μ = [22+30+35+45]/4=132/4=33 years.

The weightage mean is use to give some kind of importance or priority to the few members of the data set. Weightage average is calculated as $\mu = [w_1^*x_1^+w_2^*x_2^+...+w_3^*x_3]/[w_1^+w_2^+...+w_n]$. The weightage mean are useful in the situation where some members are to be given more weightage based on certain criteria. For example if we have to calculate the weightage average of the members of a team giving more weightage to the lower age group as they have lower risk of say disease in calculating the discount given in the insurance premium.

For example in above group we can assign 35% to 22 years, 30% to 30 years, 20% to 35 years and 15% to 45%, then weightage average is

x'=[0.35*22+0.30*30+0.20*35+0.15*45]/[0.35+0.30+0.20+0.15]= 30.45 years. A you can see the more weightage given to younger members bring

down the overall mean of the group to 30.45 years from simple mean of 33 years.

In case of the grouped data that is frequency distribution the mean is calculated as

 $\mu = \sum f_x/n$

Where f is the frequency of each class, x is the mid-point of each class and n is the number of class.

Take example of the below frequency distribution table

Class	Frequency
0-49	3
50-99	5
100-149	10
150-199	20
200-249	6
250-299	3
300-349	1

Class	Frequency(f _n)	Mid Point(x)	$f_x = x * f_n$
1-50	3	25	75
51-100	5	75	375
101-150	10	125	1250
151-200	20	175	3500
201-250	6	225	1350
251-300	3	275	825
301-350	1	325	325

Mean $\mu = \sum f_x/n = 7700/48 = 160.42$

As you can see mean is very easy to understand and calculate. It is intuitive and useful in many situations. However under certain condition mean may give skewed result due to outliers. Mean doesn't tell me how many number in the set is above the mean and how many of them are below the mean.

1.1.3 Median

The median is the number that is at the halfway of the group of data points if arranged in ascending or descending order. Therefore the median tend to consider only one or two number point from the data unlike that of mean which consider all the numbers in the data. One of the main drawbacks of the mean is the inability to exclude the outliers. One or two outlier can skewed that mean and hence impact the decision making. From median you can always say that 50% of number are above median and 50% are below the median.

An example of the impact of the outlier data point is showing the picture below. Here outlier is number 100. Without outlier the mean would have been 32 as average salary of the group. However the average salary of the group is now 40.6 due to one person having very high salary of 100. Now let's say the benefit of the group is based on the average salary of the group. A group with higher than 40k average salary will get allowance of 5% and average below 40k will get 10%. In above case one person's salary skewed the data so much that all other members are deprived of the benefit which they are entitles as an individual members provided that outlier is removed from the group. Such cases are often encounter when few individual are super rich and rest are poor in the society; average income tend to show high because of the outliers whereas majority of people are poor.



The **median** of a set of data is the middlemost number in the set. The median is also the number that is halfway into the set. To find the median, the data should first be arranged in order from least to greatest.

For example 1: Median calculation with odd numbers of data point

23, 12, 45, 56, 40, 66, 32, 21, 70

Arrange the numbers is ascending order

12, 21, 23, 32, 40, 45, 56, 66, 70

Median is 40. However if number of the data point is even then calculation is done as mean of two middlemost data point.

Example 2: median calculation with even numbers of data point

23, 12, 45, 56, 40, 66, 32, 21, 70, 46

Order in ascending order

12, 21, 23, 32, 40, 45, 46, 56, 66, 70

Two middle data point is 40 and 45. Median is (40+45)/2=42.5

Median is not affected by the extreme values or outliers. It is also useful in cases where the number is ordinal like ranking. However median has certain cons like calculation can be tedious in case of large numbers, it is not representative, and we cannot give further algebraic treatment to the median.

1.1.4 Mode

The mode of the data is number that occur most number of times irrespective of the position of the data point and the number of other data point. It just counts the frequency of each data point and the number with the highest occurrence is the mode. The mode can be used for measuring the central tendency of nominal data.

For example:

101, 123, 45, 56, 250, 123, 200, 324, 250, 300, 123

Arrange the data in ascending order

45, 56, 101, 123, 123, 123, 200, 250, 250, 300, 324

Number 123 is the mode as occurs higher number of times in the data set.

The mode may not be unique number. If two numbers has same number of occurrence the it can be bimodal. More than two number with same number of occurrence and is highest then it's called as multimodal.

Mode is easy to understand and calculate. Like median mode is also not affected by extreme values or outliers. However mode may not be unique number and it is not representative of all data. It is not well defined and the mode is not capable of further algebraic treatment.

1.1.5 Measurement of Dispersion

The measure of the central tendency provides us the central value around which the data point tend to cluster. The dispersion provided how spread the data points are from the central point. Two set of the data can have same mean but different dispersion due to distribution of its members in the space.



In above example both have same mean but the skewedness and kurtosis of each of the distribution are different. It is clear from the distribution graphs that the graph B has larger dispersion than graph A.

The **skewedness** is the measure of the symmetry or rather lack of symmetry in the data. In second graph A is symmetric but B is not symmetric and skewed toward right. The **kurtosis** is the measure of the whether the data are long tails or short tail. In first graphs we can see A has short tail and B has long tail.

The dispersion of the data from its mean is measured by the variance of the data. The variance is calculated as

 $\begin{aligned} \mathbf{6}^2 &= \sum (\mu - x) / N \\ \text{Where } \mathbf{6}^2 &= \text{ is the variance of the population} \\ \mu \text{ is the mean of the population} \\ x \text{ is the data points} \\ \text{N is the number of member in the population} \\ \text{For example} \\ 5, 12, 15, 20, 3, 6, 8, 24, 34, 32, 25, 16 \\ \mu &= (5 + 12 + 15 + 20 + 3 + 6 + 8 + 24 + 34 + 32 + 25 + 16) / 12 = 14 \\ \mathbf{6}^2 &= [(14 - 5)^2 + (14 - 12)^2 + (14 - 15)^2 + (14 - 20)^2 + (14 - 3)^2 + (14 - 8)^2 + (14 - 8)^2 + (14 - 24)^2 + (14 - 34)^2 + (14 - 32)^2 + (14 - 25)^2 + (14 - 16)^2] / 12 \\ &= (81 + 4 + 1 + 36 + 121 + 64 + 36 + 100 \\ + 400 + 324 + 121 + 4) / 12 = 1292 / 12 \\ &= 107.66 \end{aligned}$

The variance provided the average of the sum of the square of the deviation from the mean which gives position number. Without squaring the deviation would be positives and negatives which cancel out on summation. The variance overcome that problem, however it gives more weightage to the higher difference due to the squatting of the deviation. The square root of variance is known as the standard deviation of the population.

For above example the standard deviation is $\mathbf{6}$ =10.37

The standard deviation will be used in most of our subsequent discussion in the book. In simple term standard deviation measure the variability of the data point around its mean. The standard deviation is measured at same unit as data which make it easier to understand than variance.

1.1.6 Correlation

In the previous section we have seen the properties of univariate numbers like mean, median, mode and variance; in this section I would like to introduce one very important concept of bi-variate data – correlation. The correlation

measures the extent of the linear relationship between two set of data. For instance the relationship between sales and discount, relationship between salary and number of year of experience, relationship between stock prices and profit of company and so on. In statistical sense the correlation is measured on the dependency of each set of the number from the two data, however in real world a statistical dependency may not always mean causal relationship. There the correlation data has to be interpreted with the real world's causal relationship in the mind. It may not really make sense to find the correlation between numbers of child birth with number of death in a year even though data may show some kind of positive correlation.

The correlation of two data X and Y is generally denoted by ρ and is calculated as

 $\rho(X,Y)$ =covariance(X,Y)/ $[\sigma_x \sigma_y]$ = E[(X- μ_x)(Y- μ_y)]/($\sigma_x \sigma_y$)

where μ_x =mean of the data set X

 μ_v =mean of the data set Y

 σ_x =standard deviation of data X

 σ_v =standard deviation of data Y

the correlation of two data set $X=x_i$ and $Y=y_i$ is calculated as

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

The correlation value ranges from -1 to 1. The correlation 1 mean perfect positive relationship between X and Y and correlation of -1 means perfectly negative relationship between two data X and Y. the correlation of 0 means there is no linear relationship between two data X and Y but it doesn't mean there is no relationship in the real world; it can have no linear relationship.

The correlation of more than 0.5 or less than -0.5 is generally considered to be strong relationship between X and Y for positive and negative respectively. The correlation between -0.5 to 0.5 is generally considered to be weak between two set of numbers.

Price	Sales (Y)	X-ux	Y-uy	(X-ux) ²	(Y-uy) ²	(X-ux)*(Y-uy)
14	215	-6.09	-228.09	37.10	52,025.46	1,389.28
16	325	-4.09	-118.09	16.74	13,945.46	483.10
12	185	-8.09	-258.09	65.46	66,610.92	2,088.19
15	332	-5.09	-111.09	25.92	12,341.19	565.55
18	400	-2.09	-43.09	4.37	1,856.83	90.10
22	522	1.91	78.91	3.64	6,226.64	150.64
20	415	-0.09	-28.09	0.01	789.10	2.55
25	610	4.91	166.91	24.10	27,858.64	819.37
23	540	2.91	96.91	8.46	9,391.37	281.92
26	650	5.91	206.91	34.92	42,811.37	1,222.64
30	680	9.91	236.91	98.19	56,125.92	2,347.55
			Total	318.91	289,982.91	9,440.91
Mean of X	20.09		-			
Mean of Y	443.09					

 $r_{xy} = 9440.91 / [\sqrt{(318.91)} \times \sqrt{(289982.91)}] = 0.98.$

The two data set X and Y has very strong positive linear relationship between each other.

From the scatter plot itself we can see the good linear relationship between two data set.



The linear trend linear for two data shows straight line touching almost all point. That is the reason why correlation was high.



1.1.7 Graph

The graph is a pictorial representation of the data in the two or three dimensional space. The graphical representation conveys interpretation of the data better than collection of the data point in many cases. The graph is easier to read and understand. There are many types of graphs; we will be doing graphs of the commonly used one which are available in Excel or Google Spreadsheet. The reader will also understand the usage of the graphs along with the context where a particular type of graphs should be used. The graphs are indispensable part of the data analysis and presentations in day to day business working.

Bar Graphs

The bar graph is used in displaying the visual representation of the categorical data. In below example the monthly sales of a company are displayed using bar graphs. The bar graphs can be horizontal or vertical. The vertical bar graphs are generally called column graphs. However the purpose of both is same.

Month	Sales (INR Crore)
Jan	350
Feb	230
Mar	400
Apr	500

May	670
Jun	380
Jul	450



We can add the data label in the bars to clearly indicate the number to the users. This provides additional information for interpretation of the data.



Pie Chart

The pie chart is a circle which divided into number of sectors based on the relative weight of each member of the graphs. The Pie chart is useful for showing the contribution of the members or category to the overall data. It can be represented in sector with their absolute data as shown in Picture – or in the percentage contribution of the category as shown in the picture



Line Graphs

The line graphs connect each point of the graphs through lines. The line graphs representation provides slopes of each data point which helps in the understanding the growth or decline of a particular parameter. The line graphs is generally suitable for time series data.

Date	Sales Qty
1-Jan-16	2020
2-Jan-16	2300
3-Jan-16	3400
4-Jan-16	5000
5-Jan-16	4500

6-Jan-16	2300
7-Jan-16	5600
8-Jan-16	6050
9-Jan-16	4700
10-Jan-16	3600
11-Jan-16	4500
12-Jan-16	2900
13-Jan-16	3540
14-Jan-16	4120
15-Jan-16	4300
16-Jan-16	3450
17-Jan-16	3400
18-Jan-16	4200
19-Jan-16	3400
20-Jan-16	4600
21-Jan-16	4200
22-Jan-16	3400
23-Jan-16	4400
24-Jan-16	4500
25-Jan-16	4500
26-Jan-16	2300
27-Jan-16	3700

28-Jan-16	3200
29-Jan-16	4320
30-Jan-16	6500
31-Jan-16	8000

In this example we can clearly makes out the growth and decline in the sales quantity each day.



Scatter Plot

The scatter plot is generally user to see the relationship between two data set. For example you may want to see the linear relationship between data. The scatter plot also help in visualizing the concentration of the data point in the two dimensional space.

Date	Sales Qty	Discount
1-Jan-16	2020	5%
2-Jan-16	2300	4%
3-Jan-16	3400	10%
4-Jan-16	5000	15%
5-Jan-16	4500	12%
6-Jan-16	2300	6%
7-Jan-16	5600	16%
8-Jan-16	6050	20%

9-Jan-16	4700	17%	
10-Jan-16	3600	14%	
11-Jan-16	4500	16%	
12-Jan-16	2900	10%	
13-Jan-16	3540	12%	
14-Jan-16	4120	14%	
15-Jan-16	4300	15%	
16-Jan-16	3450	10%	
17-Jan-16	3400	8%	
18-Jan-16	4200	11%	
19-Jan-16	3400	9%	
20-Jan-16	4600	16%	
21-Jan-16	4200	15%	
22-Jan-16	3400	15%	
23-Jan-16	4400	14%	
24-Jan-16	4500	16%	
25-Jan-16	4500	17%	
26-Jan-16	2300	4%	
27-Jan-16	3700	7%	
28-Jan-16	3200	6%	
29-Jan-16	4320	17%	
30-Jan-16	6500	20%	

31-Jan-16	8000	25%
-----------	------	-----

In this example we can clearly see some linear relationship between quantities sold and discount%. The same is not very apparent in the data.



In below graph I plotted linear trend line in the graphs.



Area Graphs

The area graph helps in the displaying the quantitative data in term of area occupied by each members starting with the base lines. The area graph showing the trends of each member over the time in one graphs based on their area for each period. In below example sales of five category of a company over months are shown using area graphs. This graph shows the absolute value of each category over the time.

Category	Jan	Feb	Mar	Apr	May
Mobiles	230	450	300	500	610
Footwear	100	120	134	150	165
Fashion	200	250	270	300	340
Furniture	130	145	165	140	160
Books	120	130	125	140	156



Another area graphs representation as shown in the next graphs is the relative contribution of each category for the period. It shown percentage contribution of each category to the sales. Both graphs have different utility based on requirement.



Stacked Graphs

In stack graph the number from the members are stacked over previous ember till all members are represented. Like area graph this can show the absolute number and the relative number. In first graph the absolute number is stacked over each other which represent salesfor the month for each category. We can add data into each stack to make it more clear. In the second graph the relative weight of each member is stackedover each other leading to 100% for each time period.



Bubble Chart

The bubble chart is used for representing three dimensional data with third dimension being the size of the bubble. In below example X axis is Sales, Y axis is margin and the bubble size is the Average selling price of the category

Category	Sales	Margin	Average Sales Price
Mobiles	230	5%	8000
Footwear	100	12%	1250
Fashion	200	15%	2000
Furniture	130	10%	5500
Books	120	7%	250



Learning from this chapter

- The concept of central tendency and how various central numbers are measured
- Meaning, calculation, shortcoming and utility of mean, mode and median.
- Importance of dispersion, calculation of variance and standard deviation
- Correlation, how is it measured and its meaning
- MS excel basic graphs in daily usage

Section - II

PROBABILITY THEORY

Once a doctor told a patient that his change of survival is 80%, then patient said that means Dr 80% of my body will function and 20% will not be functional. Right?

Probability is one of the misunderstood words in many contexts in daily life because the person is not been able to relate theory with the application. In the simple term a probability is defined as the chances of occurrence of the events by the universe of the events possible. The real events in the world have certain randomness in it. The future cannot be predicted with precision, hence while modeling real world's events it is necessary to factor in the possibility of the randomness in the model. For example the possibility of snowfall is predicted with much precision but there is always possibility that the prediction is wrong due to some other random events such as strong wind in the area. Therefore while predicting the snowfall the weather department will mention the probability of the snowfall with some precision such as strong, weak, mild etc. rather than with 100% surety. Similarly while forecasting the sale of the product; we can predict with certain probability rather that 100% accuracy because we cannot be 100% sure about the customer's behavior. The monsoon prediction in country has much effect on the government policy and the farmer's decisions. As we know most of the prediction done at the time beginning of the monsoon are revised as time progressed as the new input and much precise data is available as the time progresses. Still the prediction can never be accurate, it gives us some number which is a probability say 80% chances that there will be 120 mm rain in this region and so on.

While probability is a very big subject in itself we will for our purpose try to understand the basic probability and the probability distribution which will help us model our events in the subsequent chapters. The readers can read other books on probability to understand the advanced topics of the probability. I have covered those probability theory and distribution which are commonly used in our daily analysis. My intention is to guide you through probability theories which are frequently used in the real data analysis in the day to day work.

1.2.1 Probability Meaning

We will start with basic definition and the notation of the probability theory in this section. The set of all possible outcomes in an experiment is known as the sample space of the experiment. We will denote it by S. For example flipping of a coin will have

 $S{=}$ {H, T}; H for head and T for Tail landing on the floor on tossing of a coin.

Flipping of two coins together will have outcome equal to

 $S = \{HH, HT, TH, TT\}$

Throwing of dice will have sample space

 $S=\{1, 2, 3, 4, 5, 6\}$

Throwing of a coin and a dice will have sample space

S={1H,1T,2H,2T,3H,3T,4H,4T,5H,5T,6H,6T}

Any subset of the sample space is known as events and is denoted by E. For example, events that a head will appear in the flipping of a coin

 $E=\{H\}$

Even number will appear in the throwing of a dice

 $E=\{2, 4, 6\}$



Union and Intersection of Events

For any two events E and F of a sample space S, the union of the event is EUF consist of all elements which are either in E or F.

For example, event E is the elements consist of event number in a throw of a dice, Event F is the elements consist of number less than 3.

E= {2, 4, 6}
F= {1, 2, 3} EUF= {1, 2, 3, 4, 6}



Intersection of the events E and F consist of all elements which are in both E and F.

E intersection $F=\{2\}$

In toss of coin the events E is head appearing $\{H\}$ and events F is tail appearing $\{T\}$. Intersection of E and F events in this case is null set. E and E are mutually exclusive events; there are no common elements between two events.

1.2.2 Basic Probability Definition

For each event E of the sample space S the probability of the event E is P(E) which given by formula

Probability P (E)=Events / Space=E/S

For example in toss of coin the probability of head appearing is P (H)={H}/ {H, T} =1/2.

The probability of an events satisfy three conditions

- 1. Probability is always between 0 and 1 which is $0 \le P(E) \le 1$
- 2. Probability of sample space is 1 which is P(S)=1
- 3. For mutually exclusively E_1 , E_2 , $E_n P(UE_n)$ =summation $P(E_n)$

The probability is always between 0 and 1. The 0 probability means that the events has zero changes of occurrence whereas 1 probability means that the events will happen for all possible universal space.

The reverse of the probability is 1-P, which is also called complement of events and is denoted by Ec in this book. If the probability of an event being occurrence is 0.20 then probability of not happening is 0.8 (1-0.20).

P (Ec) =1-P (E)

```
Union of probability of events E and F is

P(E) + P(F)=P(EUF) + P(EF)

P(EUF)= P(E) + P(F) - P(EF)

For mutually exclusive events the P(EF)=null, so P(EUF)= P(E) + P(F)

For example, in toss of two unbiased coin

S=\{HH, HT, TH, TT\}

E= \{HH, HT\}

F=\{HH, TH\}

P(EUF)=P(E) + P(F) -P(EF)

= \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}
```

1.2.3 Conditional Probability

In the examples that we have worked in the earlier part of the chapter all probabilities were unconditional. There were no special conditions attached to it. In case of conditional probability we have additional knowledge about the event that can affect its outcome.

For example, in toss of a fair die the probability of events P(E) that even number comes up is $\frac{1}{2}$. However if we know than on a particular throw of a die the result was number less than and equal to 3. The additional knowledge reduces the sample space to 3 from 6. The number of even number between 1 to 3 is 1 hence probability of even number appearing is 1/3.

P(A|B)=P(AB)/P(B)=P(2) / [P(1)+P(2)+P(3)]=1/6/[3/6]=1/3.

Take another example. Given below probabilities of cancer

Simple events	Probabilities
AC	0.15
AC'	0.25
A'C	0.10
A'C'	0.50

A=events that individual smoke

C= individual develops cancer

AC=events that an individual is a smoker and develops cancer

AC'=events that an individual is a smoker and does not develops cancer

Probability that an individual develops cancer given that he smokes is P(C/A)=P(AC)/P(A)

P(A)=P(AC) + P(AC')=0.15 + 0.25=0.40

P(C/A)=0.15 / 0.4 =0.375

Hence the probability of an individual develops cancer for a smoker is 0.375.

1.2.4 Bayes Theorem

The Bayes theorem is used in the understanding how the probability that events is true is affected by the new pieces of evidence. The Bayes theorem is mathematically stated as

P(X|A)=P(A|X)P(X)/P(A)

Where P(A) and P(X) are probability of events A and X

P(A|X) is conditional probability that is probability of A given X occurs.

P(X|A) is probability of X given A is true

Suppose that we have a method of testing cancer in a person with below data

	Cancer (1% of population)	No Cancer (99% of population)
Test Positive	80%	5%
Test Negative	20%	95%

In the given population the 1% of people has cancer and 99% of people do not have cancer. If a person has cancer then test gives position result for 80% of time and negative result for 20% of time. If a person does not have cancer then the test throws positive result in 5% of times and negative results in 95% of times. What is the change of having cancer if the test shows positive result?

Let P(X) be the probability of having cancer

P(A) be the probability of test showing positive result

P(A/X) be the probability of showing positive result for a cancer person

P(X') be the probability of not having a cancer

P(A/X') be the probability of positive test given that person do not have cancer.

Now P(X|A) the probability of having cancer given that the test is positive is given by

$$\begin{split} P(X/A) = P(A/X)P(X)/P(A) = P(A/X)P(X)/[P(A/X)P(X) + P(A/X')P(X')] \\ = (0.8*0.01)/[0.8*0.01 + 0.05*0.99) = 13.91\% \end{split}$$

In this case the P(X) the probability of cancer is prior probability which is 1% of the population. With the new evidence of the positive test the P(X/A) is the posterior probability of the X. The new evidence increase the probability of the cancer to 13.91%. The idea is that P(X|A) represents the probability assigned to X after taking into account the new piece of evidence, A. To calculate this we need, in addition to the prior probability P(X), two further conditional probabilities indicating how probable our piece of evidence is depending on whether our theory is or is not true. We can represent these as P(A|X) and P(A|X'), where X' is the negation of X.

1.2.5 Probability Distribution

The basic probability that we have gone through in previous sections are basic foundation but they are not very useful in the real application. The data tend to be in a sample with certain properties. The probability distribution is way to understand the underlying properties of the population. The particular distribution exhibits certain characteristics which are very useful in the analysis of the data from the population or population itself. So in remaining sections we will learn some of the most useful probability distribution, their properties and application in using very simple examples.

A random variable is a rule that assign one numerical value to each simple event of experiments. There are two types of random variables discrete and continuous random variable. The probability distribution is the function that assigns value to each value of the random variable or cumulative value for the point of the random variable. A random variable that can assume a countable number of values are called discrete random variable. A random variable that can assume value corresponding to any of the points contain in one or more intervals are called continuous random variables.

Examples of discrete random variables are

- 1. Sales by a salesperson in a given week
- 2. No of customer waiting in the ticket lines
- 3. The number of hit to a website

Examples of continuous random variables

- 1. The length of time between arrivals of patient in the hospital
- 2. The depth at which a successful oil drilling venture first struck oil.

The probability distribution of a discrete random variable is a graph, table or formula that specifies the probability associated with each possible value the random variable can assume. It is denoted by p(x).

Simple rule of discrete random variable is

 $p(x) \ge 0$ for all value of x

 $\sum p(x)=1$

For example, if we toss two unbiased coin the probability associated with each combination of H and T is $\frac{1}{4}$. The random variable can assume value 0, 1, 2 for appearance of H.

 $P(x=0)=P(TT)=\frac{1}{4}$ $P(x=1)=p(HT) + p(TH)=\frac{1}{4} + \frac{1}{4}=\frac{1}{2}$ $P(x=2)=p(HH)=\frac{1}{4}$

X	p(x)
0	1/4
1	1/2
2	1/4



Mean or expected value of a discrete random variable x is

$$\mu = E(x) = \sum xp(x)$$

For example in above coin throwing experiment the expected value is

 $\mu = 0*1/4 + 1*1/2 + 2*1/4 = 1$

The variable σ^2 is defined as the average of the square distances of x from the population mean μ . Since x is a random variable the square distance $(x-\mu)^2$ is also a random variable.

E[$(x-\mu)^2$]= $\sum [(x-\mu)^{2*}p(x)$ Standard deviation is the square root of the variance. $\mathbf{6}=\sqrt{(\sigma^2)}$

1.2.5.1 Binomial Distribution

A binomial random variable is one with only two possible outcomes - yes or no, pass or fail, win or lose and so on. The binomial distribution is used to model those situations where we know the outcome is dichotomous. Basic characteristics of the Binomial Random variable are

- 1. The experiment consists of n identical trials.
- 2. There are only two possible outcomes of each trial. We denote one outcome by S for success and another by F for failure.
- 3. The probability of S remains the same from trial to trials. The probability of success is denoted by p and probability of failure is denoted by q where q=1-p.
- 4. The trials are independent
- 5. The binomial random variable x is the number of S in n trials.

For example in a market survey respondent either chooses new product or the existing products.

If there are n trials with each trials has probability p of success then probability distribution with x success is

 $P(x)=n!/[x!(n-x)!]*p^{x*q^{(n-x)}}$

The term n!/[x!(n-x)!] is nothing but the combination of x elements from the n elements. In short this id denoted by (n,x).

 $P(x) = (n,x) p^{x} q^{(n-x)}$

The expected value of discrete random variable us $\mu = \sum x^* p(x)$ for binomial random variable $\mu = pq$ and variance $\sigma^2 = npq$, standard deviation $\sigma = \sqrt{(npq)}$.

Example:

In a quality inspection 60% of the products pass the quality standard and 40% fail to pass the quality test.

What is the probability that 3 out of 5 sample will pass the test?

Here p=0.6, q=0.4, $P(3) = (5,3)p^3q^2$

 $=5!/3!2! * (0.6)^{3}(0.4)^{2} = 34.56\%$

What is probability that at least 3 sample will pass the test?

The probability of atleast three sample pass the test=probability of exactly 3 sample pass the test + probability of exactly 4 sample pass the test + probability of exactly 5 sample pass the test.

 $P(x \ge 3) = P(3) + P(4) + P(5) = 5!/3!2! *(0.6)^3(0.4)^2 + 5!/4!1! *(0.6)^4(0.4) + 5!/5!0! *(0.6)^5(0.4)^0$

=34.56% + 25.92% + 7.78% = 68.26%

1.2.5.2 Poisson distribution

A type of probability distribution that is often useful in describing the number of events that will occur in a specific period of time or in a specific area or volume in the Poisson distribution. Examples of Poisson random variables are

- 1. The number of traffic accident per month at an intersection.
- 2. The part per million of some toxicant found in the water or air emission from a manufacturing plants.

Probability distribution is represented by

P(x)= $\lambda^x e^{-\lambda} / x!$ Where x= 0,1,2,3.... For Poisson distribution $\mu = \lambda$, $\sigma^2 = \lambda$

Example:

The average number of home sold by a reality company is 2 homes per day. What is the probability that exactly 3 units will be sold tomorrow.

λ=2, x=3, e=2.71828 P(x=3)=[e^-2.71828 * 2^3] / 3!=0.180.

1.2.5.3 Continuous Distribution

We will explore some of the commonly used continuous random variable which will be used in some form in our analysis in subsequent chapters.

The probability distribution for a continuous random variable x is a smooth curve that appear like below



F(x) is the function of x is called probability density function, frequency function or a probability distribution.

The area under a probability distribution corresponds to probability for x. For example in above distribution function, area A is between point a and b. the probability x assume a value between a and b (a < x < b) ~ a <=x <=b. since point a or b in continuous variable is just a point and hence p(x=a)=p(x=b)=0.

```
\begin{split} P(A) = p(a < x < b) = integration \text{ from a to b } f(x) dx \\ \text{Where } f(x) >= 0. \\ \text{Integration from } -\infty \text{ to } + \infty \int f(x) dx = 1 \end{split}
```

1.2.5.4 Uniform Distribution

In uniform distribution all point appear to have equally likelihood outcomes over their range of possible values The possible value of x consist of all point in the interval between point c and point d. the height of f(x) is constant in that interval and equal to 1/(d-c).

Total area of rectangle is given by base x height

$$=(d-c)*1/(d-c)=1$$

Mean of uniform distribution is u=(c+d)/2

Standard deviation of uniform distribution is sigma=(d-c)/sqrt(12)



1.2.5.5 Exponential Distribution

The exponential distribution is used to describe events such as length of time between earthquakes, distance travel by a person between two events, the length of time between breakdowns of manufacturing equipment's.

Probability distribution of exponential distribution is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0, \\ 0 & x < 0. \end{cases}$$

where λ is called the rate parameter.

To calculate the probabilities for exponential random variable we need to be able to find areas under the exponential probability distribution



The cumulative distribution function of the exponential distribution is

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \ge 0, \\ 0 & x < 0. \end{cases}$$

The probability or the area A of the distribution is $p(x \ge a) = e^{-\lambda x}$

The mean of the distribution is $\mu = 1/\lambda$, variance= $1/\lambda^2$

Example: Suppose length of time (in hours) between emergency arrivals at a certain hospital is 2, what is the probability that more than 5 hours pass without an emergency arrivals?

Here $\lambda = 1/2$

Here $A=e^{-\lambda x}=e^{(-5/2)}=0.082085$

The probability that more than 5 hours pass without emergency arrivals is about 0.08 for the hospital.

The exponential distribution has memoryless properties which means that the probability of the events happening in the future doesn't depends on the history of the events in the time that has already elapsed.

The conditional probability distribution of Y=X-t given X > t is

$$G(Y)=P(Y\leq=y, X>t)$$

$$= P(X \le y + t/x > t)$$

 $=1-e^{(-1)}$

G(y) is independent and identical to the original exponential distribution of x. the distribution of the remaining life does not depend on how long the device has been operating.

1.2.5.6 Normal Distribution

One of the most commonly used continuous distribution is normal distribution in most of our analysis we uses normal distribution or assume normal distribution sometime without being aware of it. The normal distribution plays very important role in the science of statistical inferences. Many natural processes generate random variables with probability distribution that are very well approximated by a normal distribution. For example the error made in measuring a person's blood pressure may be a normal random variable. Another example is the yearly rainfall in a certain region might be approximated by a normal probability distribution. Being one of the most important distributions in this chapter we will spend good amount of time

understanding the basics of the normal distribution. Normal distribution is known by bell shaped curve in most of the analysis.

The probability distribution for a Normal Random Variable x $F(x)=1/\sigma^*\sqrt{(2\pi)^*e^{-1/2[(x-\mu)/\sigma]^2}}$

Where μ = mean of the normal random variable x

 σ =standard deviation of the population

The probability of the normal curve is the area between two points in the curve. Since it is complicate we will be using normal table to find the probability of the area in a normal distribution. In practice there can be infinity large number of normal curves, one for each pair of mean u and standard deviation sigma can form a single table that is applicable to any normal curves. This is done be constructing the table of area on a function of the z-scores.

Z-score is very important concept you have to understand. The population zscore for a measurement is defined as the distance between the measurement and the population mean divide by the population standard deviation.

 $Z=[x-\mu]/\sigma$



Suppose the length of time x between charges of a pocket calculator has a normal distribution with mean of 50 hours and a standard deviation of 15 hours. If we were to observe the length of time that elapses before the need for the next charges what is the probability that this measurement will assume a value between 50 and 70 hours.

The probability between 50 and 70 is the area A in the above picture. The z score of the normal distribution with mean 50 and standard deviation 15 is calculated as

 $Z=(x-\mu)/\sigma=(70-50)/50=1.33$

Refer Normal Table, to find the area corresponding to a z-score of 1.33 we locate 1.3 in the left hand side and 0.03 in the columns. Form the table area

A=0.4082

The use of z-score simplifies the calculation of normal probabilities because if x is normally distributed with any mean and standard deviation z is always a normal random variable with mean 0 and standard deviation of 1. For the reason z is often referred to as a standard normal random variable.

If we were to find the area B of the above picture than we have to find area from 1.33 standard deviation to the infinity. However we know that the normal curve is symmetric to the mean, so area to the left and right is 0.5. Hence Area of B is 0.5 - area of A=0.5 - 0.4082=0.0908

To find the area between 1.33 standard deviation to the -1.33 standard deviation, we have to just find the area from one side and add it up because area is same due to symmetric properties.

Hence area between -1.33 to 1.33 is 2x0.4082=0.8164Let us take another example find the probability that a standard normal random variable exceed 1.64 i.e p(z>=1.64).



Area to the right side of the mean is 0.5. To find area of A2, we have to find the area of A1 and calculate area of A2.

 $A2=0.5 - p(z \ge 1.64) = 0.5 - 0.4495$ (from table)=0.0505

For the standard normal curve some of the area makes special significance as we often use them in our statistical inference such as hypothesis testing, confidence interval, creating control band etc.



The area under -1 to 1 standard deviation in a normal curve is 68% of the total area.



The area under -2 to 2 standard deviation of normal curve is 95%.



The area under -3 and 3 standard deviation of normal curve is 99.7%.

1.2.6 Central Limit Theorem

The central limit theorem states that the distribution of the sum (or average) of a large number of independent and identically distributed samples will be approximately normal regardless of the underlying distribution. The second definition of the central limit theorem states that the as the sample size increases, the sampling distribution of the mean will approach normality regardless of the shape of the population distribution. As Central Limit Theorem is integral to most of the analysis we do I would like to spend good amount of space to explain and make all readers understand the concept and its applications. The importance of the central limit theorem is hard to overstate; indeed it is the reason that many statistical procedures work. Let $X_1, X_2, ..., X_n$ be n random variable that are independent and identically distributed with mean μ and standard deviation σ . It help us makes inference about the population without knowing anything about the distribution of the population.

X'= $(X_1 + X_2 + X_3 + +X_n) / n$ is the sample mean

 $[X' - \mu']/(\sigma/\sqrt{n}) \to N(0,1) \text{ as } n \to \infty$

Which implies that $x' \rightarrow N(\mu, \sigma^2/\sqrt{n})$

The sample mean can be approximated with a normal random variable with mean μ and standard deviation $\sigma/\sqrt{(n)}$.

The first step in understanding the theorem is to prove it practical significance. Let's us use excel to prove the above claims.

In sheet1, name cell A1 as Sl. No. and B1 as Population. From A2 to A201 get serial number 1 to 200. In B2 write formula =INT(RAND()*100) and drag it till B201. Copy entire range B2:B201, copy it and paste special -> value. This is to get rid of frequent change in the number due to RAND() function.



Plot the frequency table of the observation using pivot table. Make a bar graph of the frequency distribution. My distribution is like one below. Is it bell shape? No. Definitely this is not a normal distribution. Well this is a random numbers so we cannot expect any defined distribution. So our population is 200 numbers between 0 to 100 and with no defined distribution.



Population mean is 46.36 and standard deviation is 28.68 for above observations.

Now Open Sheet2. Name A1 as Sample No. and from B1 to AY1 name cell as Ob1, Ob2...Ob50. We are going to use sample size (n) =50. From A2 to A10001 name the sample as 1 to 10000. We are going to take 10,000 samples from the population.

In cell B2 write formula =VLOOKUP(INT(RAND ()*200),Sheet1! \$A:\$B,2,0). This will help us pick one number from 200 randomly. Copy and paste this formula in all cells in the range B2:AY10001. Now we have 10,000 samples with sample size 50 each. From AZ2 to AZ10001 take the average of each sample. Take the average and standard deviation of the sample averages that is range AZ2:AZ10001. For simplicity I have rounded off each sample average to nearest integer.

Average of sample mean=46 ~population mean

Standard deviation =4.06

If we use the formula $\sigma/\sqrt{(n)}$ then 28.68/ $\sqrt{(50)}$ =4.06 which is equal to standard deviation of the sample means.





Now Plot the frequency distribution of the sample means starting with 100 samples and then 500, 1000, 2000, 5000 and 10000 samples. As you can see from the plot of the sample means distribution tend to normal shapes as we increases the number of samples.

As we mentioned in our discussion of normal curve that

About 68% of the sample mean will be within the range of μ - σ to μ + σ

- About 95% of the sample mean will be within the range of μ -2 σ to μ +2 σ

- About 99% of the sample mean will be within the range of μ -3 σ to μ +3 σ Same being derived from the distribution of the means as below table. This is another proof that the central limit theorem is true

Range	Lower Limit	Upper Limit	No of Sample mean	Total Samples	% sample in this range	Reference %
μ -σ to μ +σ	42	50	7379	10000	73.8%	68%
μ-2σ to μ +2σ	38	54	9646	10000	96.5%	95%
μ -30 to μ +30	34	58	9985	10000	99.9%	99%

Another variant of this could be keeping number of sample same say 2000 and start with sample size and increase it till 180 or so and see the difference in the shape of the sample means distribution. This will prove the point that as sample size increases the sample mean distribution tends to be normal distribution.

CLT will be used in many of the ensuing chapters. I will mentioned the theorem in those places. In this section let's try to understand one application

of the CLT in the production process



You are manufacturer a wrench which is 2 cm in the space using a machine. A tolerance of 0.5 mm is allowed on either side that is 1.95 cm to 2.05 cm space width is acceptable. Since the machine produce thousands of them every day it is not possible for you to measure each one of them. As a process you decided to take sample of 30 wrench everyday randomly and take average of those sample and plot it see how process is behaving- whether it is within tolerance range or it is out of range or going out of range. Historically the width of the space has standard deviation of the 1 mm around the center that is 2 cm. If the sample mean is 99% within the allowed tolerance the process is in synch else you have to get machine recalibrated by a technician from other company. How would you go about it?

Here again you can take help of Central Limit Theorem (CLT). As we know as sample size increases the sample means standard deviation is equal to population standard deviation divided by sample size.

As n -> ∞ the distribution of X' is normal with mean μ and standard deviation σ/\sqrt{n}

So sample mean standard deviation is $1/\sqrt{30} = 0.183$ mm

The confidence interval of 99% around mean is -2.58*0.183 and +2.58*0.183 i.e. -0.47 mm and 0.47 mm. So for 99% confidence interval control the lower bound is 2 cm -0.47 mm and upper bound is 2 cm +0.47 mm. As $0.47 \sim 0.5$ mm our process has 99% confidence interval is within 0.5 mm tolerance in both sides, so we can go with it.



In above two patterns first one has one point outside the limit but pattern is more stable. Second one is within the limit but pattern shows that it is going outside with clear trends. Perhaps it time to recalibrate before it goes out of the control.

Learning from the Chapter

- Understanding the concept of probability events, universe, complementary events, union & intersection of events.
- Conditional probability and Bayes theorem.
- Formulation of probability distribution, difference between discrete and continuous probability distribution.
- Understanding of concept and usage of commonly used distribution Binomial distribution, Poisson distribution, uniform distribution, exponential distribution and normal distribution.
- Area under normal curve, Z Score, uniform normal distribution
- Central limit Theorem concept, application and formulation.

Table: Normal table

					North	4101		ş.		
			2 03		/					
			1		/					
			£ 01.	l	/					
			12.2		·		-			
					1 11		1 3			
						×				
			Are	a under th	he Normal	Curve fr	om 0 to X			
x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0,00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.03567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.33993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0,44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0,47558	0.47615	0.47670
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0,49813	0,49819	0.49825	0,49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0,49893	0,49896	0.49900
3.1	0.49903	0,49906	0.49910	0.49913	0.49916	0.49918	0.49921	0.49924	0,49926	0.49929
3.2	0,49931	0.49934	0,49936	0.49938	0.49940	0.49942	0.49944	0.49946	0.49948	0.49950
3.3	0.49952	0.49953	0.49955	0.49957	0.49958	0.49960	0.49961	0.49962	0.49964	0.49965
3.4	0.49966	0.49968	0.49969	0.49970	0.49971	0.49972	0.49973	0.49974	0.49975	0.49976
3.5	0.49977	0.49978	0.49978	0.49979	0.49980	0.49981	0.49981	0.49982	0.49983	0.49983
3.6	0.49984	0.49985	0.49985	0.49986	0.49986	0.49987	0.49987	0.49988	0.49988	0.49989
3.7	0.49989	0.49990	0.49990	0.49990	0.49991	0.49991	0.49992	0.49992	0.49992	0.49992
3.8	0.49993	0,49993	0.49993	0.49994	0.49994	0.49994	0.49994	0.49995	0.49995	0.49995
3.9	0.49995	0.49995	0.49996	0.49996	0.49996	0.49996	0.49996	0.49996	0.49997	0.49997
4.0	0.49997	0.49997	0.49997	0.49997	0.49997	0.49997	0.49998	0.49998	0.49998	0.49998

Section - III

SAMPLING AND HYPOTHESIS TESTING

1.3.1 Sampling

Sampling is an important part of the data analysis. Sampling is the method of collecting information from a sub-section of a large group or population. If we can study every unit of elements of the population then the survey is known as census survey whereas if we study only a part of the population then the survey is known as the sample survey.

The sample survey is used in almost all fields. It is not always possible to have read all the population's data point for making all decision. Instead we analyze the sample from the population and make decisions. The sampling has many advantages like

- 1. It reduces the cost as we have to study on smaller number of elements.
- 2. The collection and analysis of data takes lesser time for the sample.
- 3. In destructive test, sampling is the only way to test the performance of the products.
- 4. In many cases we do not know the true extent of the true population. The only inferences can be made about the population is through the sample of the population.

Even though sample study is the fact of the life, we need to be aware of its limitation.

- 1. The sample provides very limited information with respect to population.
- 2. There is likely to have error due to the projection of the sample characteristics to the population. This error is known as sampling error.

In real life the decisions are made using the sample data available at that point of time due to advantages listed above.

There are many kind of sampling. Broadly sampling method can be classified into two groups – Probabilistic Sampling methods and Non-Probability

Sampling methods. In probability sampling methods every element in the population has an equal chance of being included in the sample. The Non-probability sampling method includes variety of the techniques ranging from simple method of sampling as per convenience to complex method of allocating quota on the basis of characteristics of the population.

Some of the frequently used sampling method will be described here:-

- 1. **Simple random Sampling**: The simple random sampling is the most simple and frequently used sampling method. In this method we select n elements from population of N elements such that each and every possible sample of size has the same chance. Once an element has be drawn from the population all the remaining elements has the equal chances of being included in the sample in next draw.
- 2. **Cluster Sampling**: In this method the population is often divided into cluster by the properties of the population or time. For instance, if surveying households within a city, we might choose to select 100 city blocks and then interview every household within the selected blocks.
- 3. **Stratified Sampling**: This is used where population can be classified into different strata. Each stratum is then sampled as an independent subpopulation, out of which individual elements can be randomly selected. The difference between stratified sampling and cluster sampling is that in cluster each and every member of selected clusters are included in the test but in stratified sampling the random sampling is applied to each strata (cluster) of the population to select n number of members from each strata.
- 4. **Systematic Sampling**: Systematic sampling (also known as interval sampling) relies on arranging the study population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every kth element. In this case, k= (population size/sample size). It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the kth element in the list. A simple example would be to select every 10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 10')

5. Quota Sampling: In quota sampling, the population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified proportion. Each member of the sub groups has known probability of being selected that is the selection of the members are not based on the random sampling like that of stratified sampling.

Let us consider a hypothetic population to provide you an example of how different sampling method works. In a city there are 5 blocks which is composed of following 2 block belong to Lower Income Group (LIG) with 1000 families in each block

- 1. 2 block belong to Middle Income Group (MIG) with 500 families in each block
- 2. 1 block belong to High Income Group (HIG) with 500 families.

So total population of the city is 2x100 + 2x500 + 500=3500 families with distinct house number 1 to 3500. A survey is to be conducted to find the household goods of the city by an NGO to access the living standard of the people in the city. The representative of the NGO has limited time and manpower, so he considered the sample survey instead of census with sample size of 500. He considered following survey methods

Simple Random Sampling: In this method he will randomly choose 500 out of the 3500 families irrespective of the block in which a family lives. In the first draw each family has 500/3500 probability of being chosen; in the second draw each of remaining 3499 families has 499/3499 probability of being chosen and so on till on 500th draw each of the remaining 3001 families has 1/3001 probabilities of being chosen.

Cluster Sampling: In this method the cluster of 5 was created based on the block and then every household of chosen cluster will be surveyed. For example 1 block of MIG can be selected for the survey and each of 500 families are part of the survey.

Stratified Sampling: In this method city can be divided into three strata – LIG, MIG and HIG. Since we have 3500 families, then number of families from LIG should be 2000/3500*500= 286, number of families from MIG should be

1000/3500*500=143 and number of families from HIG should be 500/3500*500=71. Here we have given equal proportion to the each strata of the population.

Systematic Sampling: In this method we can use house number as the counter for sampling. Since we have to sample 500 from 3500, the counter can be 7. Say house number 1 was selected then other members are house No 1+7, 1+7+7 till 500 members are selected.

Quota Sampling: In this method we can divided the city into three cluster l - LIG, MIG and HIG like stratified sampling. However we have been mandated to include 300 LIG, 150 MIG and 50 HIG. This makes sampling non probabilistic as it forces inclusion of predetermined number of each cluster or sub group.

1.3.2 Hypothesis Testing

A hypothesis test is a method of testing a claim or a hypothesis about the parameter of the population using a sample data from the population. The hypothesis is an assumption about the population which needs to be tested. The assumption or hypothesis is tested using the sample data and determine the difference between the hypothesized value and the actual value of the sample mean. The difference between the hypothesized value and the mean of the sample provides the value that is used to accept the assumption or not. The smaller the difference the greater the likelihood that hypothesized value for the population mean is correct whereas larger the difference the smaller the likelihood. In real world we often come across many situation in which we have set of data from different population and have to determine query such as which sample is better, are both population same in statistical term, can we conclude sample 1 has higher content than sample 2 and so on. The hypothesis test can be used in almost every field.

In ecommerce industry we have to come across many A/B testing data which try to understand the difference between two set of pages or advertisement. In medical fields the fields testing of medicines, testing of medicines with different combination etc. can be very useful. In this chapter we will understand the method of defining the hypothesis and how to test them. The learning from this chapter will be useful in the subsequent chapter related to A/B testing, customer analytics and so on. The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. In previous section we study the reason for using sample. Using sample as representative of the population, we make the inference. This is possible due to the powerful theory we have learnt that is Central Limit Theorem (CLT).

Generally we deal with two samples in hypothesis. The question that arise are of nature

- 1. Is mean of population greater/less than value x?
- 2. Is population A and population B similar?
- 3. Is population A better/higher than population B?
- 4. Is population A lesser/lower than population B?

We will be generally dealing with two types of data in hypothesis testing numeric unit and proportion (p-value). In case of the numeric unit data like age of population deals with the numeric mean whereas in case of proportion data like percentage of people with income greater than 1 lakhs deals with the percentage data or proportion. The general principle of the testing will be same but way sample mean and standard deviation is calculated will be different.

Number Type	Number of Sample	Which Tail
Mean Value	One Sample Two Sample	One Tail Two Tail
P-Value	One Sample Two Sample	One Tail Two Tail

Broadly Hypothesis Testing can be classified into following types

As shown in the first figure the one tail test is either done on the left side or right side based on the hypothesis. Assuming we have to test whether x is significantly greater than μ with 95% confidence interval. We just need to calculate the Z-score of the x from the μ with σ_s standard deviation where σ_s is sample standard deviation.

 $Z\!\!=\!\![x\!\!-\!\!\mu]/\sigma_s$

If Z is greater than 1.96 σ (95% confidence interval) then we can reject the hypothesis that x is similar to mean else we reject the hypothesis. In case of Two-tail test we have to create two boundary $1.96\sigma_s$ and $-1.96\sigma_s$ and check if sample Z-score falls inside the boundary; if yes than we can conclude that null hypothesis is accepted else it is rejected. Here σ_s (standard error of the sample) is σ/\sqrt{n} where σ is population standard deviation and n is sample mean.



Two Tail Hypothesis Testing

The broad steps that are followed in hypothesis testing are

1. We identify the claim or hypothesis that is to be tested. We have to formulate null hypothesis H_0 and alternative hypothesis H_1 . In null hypothesis we presume that the population mean is true or mean of two populations is same is true. This is somewhat similar to court procedure where under trial are presumed to be innocent at the beginning. Note that only reason we are testing the null hypothesis is that we think it is wrong. Using same courtroom analogy the very reason that a court is having hearing is to prove that under trial is not innocent else court has already assumed it to be innocent from the beginning.

- 2. Set the criteria for testing. We state the level of significance of the test. Generally we use 95% level of significance however it is not fixed. For some test like in medicines it could be 99.99%. We know that sample mean is equal to population mean on average if null hypothesis is true.
- 3. Compute the test statistics. The test statistics is the z-score we have discussed earlier. We have the boundary condition based on the level of significance. For 95% significance boundary is set at 1.96SD for one tail test and -1.96SD to 1.96SD for two tail test.
- 4. Make the decision based on the test statistics. If the probability of sample mean falling within the boundary condition is less than 5% then null hypothesis is rejected.

The above logic of the hypothesis testing is based on two important assumptions

- 1. The sample mean is unbiased estimator of the population mean
- 2. Regardless of the population distribution the sample mean is normally distributed due to Central Limit Theorem. Hence the probability of all other sample means can be selected as normally distributed.
- 3. In Hypothesis we can make error as it can never achieve 100% result, hence we assume some level of confidence.

		Decision		
		Accept null Hypothesis	Reject Null Hypothesis	
Population Truth Position	True	CORRECT 1-a	TYPE I Error α	
	False	TYPE II Error β	CORRECT 1-β (POWER)	
	~			

The TYPE I error is the probability of rejecting the null hypothesis that is actually true. Making this type of error is analogous to pronouncing someone guilty instead of being actually innocent. α that is level of significance is the largest possible probability of making TYPE I error.

We control the probability of the TYPE I error through level of significance α . The correct decision is to reject the null hypothesis. There is always some probability that we decide that the null hypothesis is false when indeed it is false. This decision is called the power of the decision making process. It is called power denoted by 1- β because it is the decision we aimed for. Remember that we only tested the null hypothesis because we wanted to prove it is wrong.

There is always tradeoff between Type-I error and Type-II error. In some test like medical where the drugs are dangerous we have to make Type-II error close to nil where as if you are doing a market research for product promotion then Type-II error can be more.

1.3.2.1 One Sample Hypothesis Testing

In one sample hypothesis testing there are only one population and the sample belong to that population. The test involved claims such as the average is greater than x, average is less than x and average is equal to x.

Let's explain this with example. A nursery school claims that average students of the school studies 6 hours a day. We want to test the claim of the school using hypothesis testing using 95% confidence interval.

Null Hypothesis $H_0: \mu_0 = 6$

Alternate Hypothesis $H_1: \mu_1 <> 6$

Assume we take the sample of 45 students and found the average to be 5.6 and standard deviation of the sample is 2 hours.

Here we have to calculate the z-score using the mean and standard error of the mean

Standard Error of the mean $\sigma_x = \sigma/\sqrt{(n)} = 2/\sqrt{(45)} = 0.3$

$$Z=(\mu - x')/\sigma_x=(6 - 5.6)/0.3=1.33$$



Since Z score 1.33 lies within +-1.96SD of the distribution, the null hypothesis cannot be rejected. Hence we cannot conclude that average study hours of the students are not 6 hours.

Let us assume now that the mean of the study hours was 7 hours from sample of 45. In this case the z-score would be

Z=(7-6)/0.3=3.03

Since 3.03 lies outside the 95% confidence interval +-1.96SD (here SD is 1) we can reject the null hypothesis that the average study hours of the nursery classes are 6 hours.

Above example is a case of two-tail test. In case of the claim such that the average studies of nursery is more than 6 hours then Null Hypothesis would have been H₀: μ =6 and Alternate Hypothesis would have been H₁: μ < 6

In this case will be looking into the right hand side of the mean in the diagram only. In the previous example let us assume that the mean was 5.2.

Z score=(6-5.2)/0.3=2.33

We can see from the diagram the z score is above -1.96; hence we cannot reject the null hypothesis. Anything less than -1.96 would have led to rejection of the null hypothesis.



1.3.2.2 Two Sample Hypothesis Testing

Two-sample hypothesis testing is statistical analysis designed to test if there is a difference between two means from two different populations. For example, a two-sample hypothesis could be used to test if there is a difference in the mean salary between male and female doctors in the New York City area. A twosample hypothesis test could also be used to test if the mean number of defective parts produced using assembly line A is greater than the mean number of defective parts produced using assembly line B. Similar to onesample hypothesis tests, a one-tailed and two-tailed test of the null hypothesis can be performed on two-sample hypothesis testing as well. The two-sample hypothesis test of no difference between the mean salaries of male and female doctors in the New York City area is an example of a two-tailed test. The test of whether or not the mean number of defective parts produced on assembly line A is greater than the mean number of defective parts produced on assembly line B is an example of a one-tailed test.

Similar to one sample testing, the two sample testing also start with the stating the null hypothesis and alternate hypothesis. In two sample test hypothesis is generally state as

 $H_0: \mu_1 = \mu_2$

 $H_1: \mu_1 \diamondsuit \mu_2$

Where μ_1 is mean of sample 1 and μ_2 is mean of sample 2.

In two sample the variance for sample is calculated as

$$\sigma_{x}^{2} = \sigma_{1} / \sqrt{(n_{1})} + \sigma_{2} / \sqrt{(n_{2})}$$

If we know that the standard deviation of population 1 is σ_1 and standard deviation of the sample 2 is σ_2 .

In case population variance is not known then we calculate the standard error of the sample mean difference as

 $\mathbf{5}_{x1} = \sqrt{(sp^{2*}(1/n_1 + 1/n_2))}$

Where pooled variance sp id calculated as

 $Sp^{2} = [(n_{1}-1)s_{1}^{2} + (n_{2}-1)s_{2}^{2}]/(n_{1}+n_{2}-2)$

Where s_1 is sample1 standard deviation and s_2 is sample 2 standard deviation.

Z score is calculate as

 $Z\!\!=\!\!\left[\mu_1 \text{ - } \mu_2\right]/\!\sigma_{x1} \text{ - }_{x2}$

If Z fall within the ySD as per confidence then null hypothesis is accepted else null hypothesis is rejected.

For example in a class there is two sections A and B. The institute claimed that section A being taught by outside faculty score higher than section B at 95% confidence interval.

Number of student in Section A=50

Number of student in Section B=65 Mean score of Section A=93.5 Mean score of Section B=87 Standard Deviation of the section A=10 Standard deviation of Section B=15 We have calculate the pooled variance first Sp =[$(n_1-1)s_1^2 + (n_2-1)s_2^2$]/ $(n_1 + n_2 - 2)$ =[$(50-1)*10^2 + (65-1)*15^2$] /(50+65-2)=170.8 Standard Error of the sample mean difference $\sigma_{x1-x2} = \sqrt{(sp^{2*}(1/n_1 + 1/n_2))}$

 $=\sqrt{(170.8*(1/50 + 1/65))}$ =2.5 Z score=Z=[$\mu_1 - \mu_2$] / σ_{x1-x2} =(93.5-87)/2.5=2.59



Since the Z score lies outside the 1.9SD of the sample mean difference we can reject null hypothesis that there is no difference between students of Section A and Section B. Hence we conclude that the claim by the school is correct at 95% confidence interval.

1.3.2.3 Two Samples from same Population

Sometime we have to compare the two sample from same population to ascertain if the claims are true or false. Let's take an example of Gym weight loose programme.

One fine day you decided to shed some kilos you have accumulated in last few month. One gym instructor explain you the process of weight reduction technique in his Gym and promise to shed at least 5 kilos in 6 months' time. In the Gym wall is pasted with examples of people who have reduces weight in this gym like one below. You are sure about the authenticity of the photo but is after photo taken before or after.



Before committing your money to this gym you decided to find out the real fact for yourself. You know few friends who are members of this gym, so you decided to take their help. Using your friends help you gather data of 30 members who has completed 6 months of this programme their weight at the time of registration and their weight after the registration.

Reading during registration are { 62, 77, 68, 80, 57, 66, 56, 55, 70, 71, 85, 78, 60, 61, 53, 58, 63, 66, 74, 69, 72, 66, 64, 81, 69, 58, 63, 75, 80, 82 } and reading after size months are { 60, 72, 60, 76, 52, 58, 50, 53, 65, 67, 77, 77, 52, 55, 48, 56, 56, 60, 70, 63, 70, 62, 60, 75, 62, 57, 60, 66, 75, 69}.

From both number we get the reduction in the weight as $\{2, 5, 8, 4, 5, 8, 6, 2, 5, 4, 8, 1, 8, 6, 5, 2, 7, 6, 4, 6, 2, 4, 4, 6, 7, 1, 3, 9, 5, 13\}$. The average weight loss is 5.2 Kg and 18 people lost 5 or more kg of weight and 12 people lost less than 5 kg of weight.

Now you are confused what to do. But you are determined that you will join only if 99% sure that the weight reduction technique work. You decided to go for hypothesis testing using this data

H₀: µ=5

 $H_1: \mu > 5$

Since you want 99% confidence here α =0.01. Standard deviation of the weight reduction is 2.66 Kg. The standard error of mean reduction= σ /=2.66/=0.49

 $Z=(\mu - 5)/\sigma=(5.2-5)/0.49=0.41$

99% standard Z score is 2.33. As 0.41 < 2.33 we cannot reject the null hypothesis. Hence claim made gym cannot be accepted. This helps you take the decision. Isn't it very easy way of proving and disproving a claim made by someone with just handful of data points?



1.3.2.4 Hypothesis Testing of the Proportion

In proportional hypothesis testing the test is done to disprove the claim of the population proportion using sample proportion and standard deviation with certain confidence interval. For convenience we will continue using 95% confidence in this section as well.

For the one sample the testing are of the nature that the population proportion is x or less than or greater than x. The hypothesis is generally formulated as

 $H_0: p_0 = x$

 $H_1: p_0 <> x$

Let us take an example to show the steps involved. Proportion of the student absent at 4 pm class in the school is 0.35 as per school management. The sample of 30 students randomly checked at 4 pm found the proportion to be 0.42.

H₀: p₀=0.35

H₁:p₁ <> 0.35

Standard deviation is calculated as $\sigma_p = \sqrt{(pq/n)} = \sqrt{(0.35*0.65/30)} = 0.087$ Z score= $(p_0 - x) / \sigma_p = (0.35 - 0.42)/0.087 = -0.805$



The Z score is within the 1.96 SD of the distribution, hence we cannot reject the null hypothesis.

1.3.2.5 Two population proportion Hypothesis Testing

The two population testing is done to find if the two groups are same or are they different. Is one group better/lower/higher than second group?

In case the population proportion is known then the standard error of the sample proportion difference is calculated as

 $\mathbf{6}_{p_1-p_2} = \sqrt{(p_1q_1/n_1 + p_2q_2/n_2)}$

We are to find the different in the proportion between two populations then the sample standard error of the proportion is calculated as

$$P_{\text{pooled}} = [(n_1-1)p_1 + (n_2-1)p_2]/(n_1+n_2-2)$$

Standard error of the Proportion difference is calculated as

 $\mathbf{6}_{p1-p2} = \sqrt{(P_{pooled} * Q_{pooled} / n_1 + P_{pooled} * Q_{pooled} / n_2)}$

The Z score is calculate as $Z=(p_1-p_2)/\sigma_{p_1-p_2}$

If -1.96SD<=Z<=1.96SD the null hypothesis is not rejected else null hypothesis is rejected at 95% confidence interval.

I am not adding any example, readers can come up with their own hypothesis and test as a practice

1.3.3 Chi-square Test

The **chi-squared distribution** with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. It is a special case of the gamma distribution and is one of the most widely used probability distributions in inferential statistics that is in hypothesis testing or in construction of confidence intervals.

If Y_i are distributed with mean 0 and standard deviation 1 then

 $\chi^2 = \sum Y_i^2$ where i=1 to k

Then is χ^2 Chi-Square distribution with k degree of freedom.

The Chi-Square distribution is generally not used for modeling the natural phenomenon but generally used for hypothesis testing. The chi-square hypothesis is used in Test of Independence in Contingency tables and Goodness of fit of observed data to the hypothetical distribution. There are many other hypothesis testing using chi-square but in this book we still be learning Test of Independence and goodness of fit testing using chi-square.

3.3.1 Test of independence

The test of independence is done for two categorical variables. Like any other hypothesis testing the null hypothesis assume two variables are independent. The alternative hypothesis is to find that both variables are elated.

H₀: Two categorical variables are Independent

H₁: two categorical variables are related

The categorical data are tabulated in contingency table with r rows and c columns with the observed count. Using observed count we find the expected count assuming the variables are independent. The chi-square is calculated as

 $\chi_2 = \sum (Oi-Ei)_2/Ei$

where Oi is the observed count, Ei is expected count.

The expected count is calculated as E=(total row)*(total column) / sample size=r*c /n

	Favor	Indifferent	Opposed	Total
Student	135	80	65	280
Teacher	65	68	80	213
Total	200	148	145	493

In the above table the opinion of the Students and teachers in the university about the changes in the syllabus of the subject.

The expected count is calculated as

	Favor	Indifferent	Opposed	Total	
Student	280*200/493	280*148/493	280*145/493	280	
	= 113.59	=84	=82.35		
Teacher	213*200/493	213*148/493	213*145/493	213	
	=86.40	=63.94	=62.65		
Total	200	148	145	493	

The degree of freedom=(r-1)*(c-1)=(2-1)*(3-1)=2

 $\chi^{2=(135-113.59)^{2}/113.59}$ + (80-84) ²/84 + (65-82.35) ²/82.35 + (65-86.40) ²/86.40 +

 $(68-63.94)^{2}/63.94 + (80-62.65)^{2}/62.65$

=18.24



At 95% confidence interval (α =0.05) with 2 degree of freedom the value is 5.991. The calculated chi-square is 18.24 which is outside the acceptance region. Hence we can reject the hypothesis that the opinion of students and teachers are independent.

3.3.2 Goodness of fit Test

The chi-square test is used to test if a sample of data came from a population with a specific distribution.

For example we want to find if a diesis biased. We rolled die 60 times and noted the observed frequencies as below

Die Face	1	2	3	4	5	6
Frequency	8	10	14	7	13	8

H₀: the die is fair

H₁: Die is not fair

Significance level α =0.05

Degree of freedom=v-1=6-1=5

Die Face	1	2	3	4	5	6
Frequency	8	10	14	7	13	8
Expected Fre- quency	10	10	10	10	10	10
Oi-Ei	-2	0	4	-3	3	-2
(Oi-Ei) ²	4	0	16	9	9	4
(Oi-Ei) ² /Ei	0.4	0	1.6	0.9	0.9	0.4

$\chi 2 = 4.2$

the chi-square at 5 degree of freedom is 11.071 from the table



As 4.2 (calculated) < 11.07 (from table) we cannot reject the null hypothesis. There is no evidence to suggest at 95% confidence interval that die is not fair.

				TABLE IV							
Degrees of		Chl-Square (χ^2) Distribution Area to the Right of Critical Value									
Freedom	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01			
1 2 3 4 5	0.020 0.115 0.297 0.554	0.001 0.051 0.216 0.484 0.831	0.004 0.103 0.352 0.711 1.145	0.016 0.211 0.584 1.064 1.610	2.706 4.605 6.251 7.779 9.236	3.841 5.991 7.815 9.488 11.071	5.024 7.378 9.348 11.143 12.833	6.635 9.210 11.345 13.277 15.086			
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812			
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475			
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090			
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666			
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209			
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725			
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217			
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688			
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141			
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578			
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000			
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409			
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805			
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191			
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566			
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932			
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289			
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638			
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980			
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314			
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642			
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963			
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278			
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588			
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892			
3.4 Analysis of Variance (ANOVA)

In previous sections we have done the hypothesis test of one sample or two samples from the populations. When we have to test the parameter of more than two populations through the hypothesis testing then Analysis of Variance (ANOVA) is used for testing the significance of the difference among the sample means. For example the sample of people from different metro cities of India for testing the significance of difference among their weight to test whether weight of the population of cities have same mean. The hypothesis of ANOVA is written as

Null Hypothesis H₀: $\mu_1 = \mu_2 = \mu_3 = ..= \mu_n$

Alternative Hypothesis H_1 : μ_1 , μ_2 , μ_3 and μ_n are not equal

The hypothesis is tested by comparing the variance among the sample means and the variance within the samples.

Let x_T =mean of all the data point in the samples which is called grand mean

 x_{T_i} =mean of sample j

k=number of samples

 n_i =number of items in the sample j

 $x_{ii} = i^{th}$ member of j^{th} sample

 $x_{T_i} = \sum x_{ij}/n_i$ where I =1 to nj

 $x_T = \sum x_{ij} / \sum n_j$ where j =1 to k and i=1 to nj

Variance among the sample mean $\sigma_T^2 = \sum n_j (x_{Tj} - x_T)^2 / k-1$

Variance within the samples $\sigma_{w}^{2} = \sum (n_{j}-1)/(n_{T}-k)*s_{j}^{2}$

Total sample size $n_T = \sum n_j$

 s_{i}^{2} =sample variance of the jth sample= $\sum (x-x_{Ti})^{2}/(n_{i}-1)$

F statistics=variance among sample mean / variance within the samples

= between-column variance/within-column variance= σ_T^2 / σ_w^2

The hypothesis is likely to be true in case the variance among sample means and variance within the sample are equal or nearly equal which shows that the sample comes from populations with same mean and F tends to be 1.

In case of the sample from populations with different means the variance among sample means tend to be larger than the variance within the samples and F value tend to be large.

F statistics is a family of distribution. It has two degree of freedom one each from numerator and denominator.

Numerator degree of freedom=number of samples -1=k-1

Denominator degree of freedom=number of elements in sample – number of samples



Where 1- alpha is the confidence interval.

ANOVA Table

Source of Vari- ation	Sums of Squares	Degree of Freedom	Mean of squares	F
Between Group	$SSB = \sum n_j (x_{Tj} - x_T)^2$	k-1	MSB= SSB/k-1	F=MSB/MSE
Within group	$SSE = \sum \sum (x - x_{Tj})^2$	n-k	MSE= SSE/N-k	
Total variation	$SST = \sum \sum (x - xT)^2$	N-1		

Where $N = \sum n_i$

Example: In below example the sample from four cities on are compare to find out if there is difference in the income of the family from four cities. With alpha value of 0.05 the F statistics from the F distribution table with degree of freedom Numerator =3 and Denominator=38 is 2.84. F-statistics from the calculation is 0.27 which is less than F value from table for the given degrees of freedom, therefore we do not reject the null hypothesis that the samples come from population with same mean.

LE HOME INSEE	FACE LANSA	IORAAA	0474	RVDE	VOV	ANDA IN					1 10 -
X Cat	+ 14	180		e. 5	Toplat	General	-	E		In Be The Easter - by	15
10 City - 0 /	**	3+A+		+C +C []	Marga & Cantor +	5 - % +	12.25	andrand Fa	mata Cal	Inat Dates Farmat Aller Set &	Feeth
Cation &				-				construction of the	cen your	Con Antone	Select-1
		- 4	- 0		4		1	1			N 1
		Same	ples				* 0	6-3	10.1		1
	Dethi	Mumbai	Overeal	Kolkata		Delhi	Mumbai	Otennai	Kolkata	6 ⁴ ~	372.83
	20	35	12	7		173.4	5.6	236.7	436.3	64,	101.21
	13	14	23	13		406.7	346.9	19.2	221.7	\$	0.27
	25	24	78	42		66.7	74.4	2561.9	199.1	Numerator degree of freedom	3
	17	53		35		261.4	415.1	818.8	50.6	Denomenator degree of Ereedom	38
	-45	65	43	41		140.0	1048.1	243.8	171.9		1.00
	12	52	15	18		448.0	375.4	153.4	97.8		
	78	50	24			2010.0	511.9	11.5	0.0		
	45	8	31	3.5		140.0	606.4	13.1	26.1		
	50		.11	34		283.4		268.5	37.3		
	42		8	1		78.0	-	375.8			
	40		16			46.7		129.6	-		
	11		24			491.4		11.5			
			15	-	1			153.4			
Mean (a-)	33.17	32.63	27.38	27.89	51.4.3	4545.7	3383.9	4997.1	1240.9		
sample size (n)	\$2		13	9	2x++)//(=1)	413.2	483.4	416.4	155.1		
Number of sample (k)	4			-		6		2			
Grand Mean (us)	30.14										
All sample elements (N	0 43										
											1
· · Sheeti Mer	st Sheets	00						547			
and the second se											

ANOVA can be calculated with pre-defined function in the Excel. Using same sample data we have calculated the ANOVA in the screenshot below

C H ST	NORT	NEWMONT	POMALA	0454	ADVEN.	VEN	ANDA In	al.							. * 1		-
Rate Parts	(aller Factor # 7	-]n ¥ • ≡ •	3. A - 3	- 1	₽- 5+ << 0+	nge Tani Inge & Conton in	(anes) 5 + % +	19.0	E		21	2× (1)	NA RA	Arlan -			
Option		Fait	×		argenere		8,454			Styles.		-046		1.000	×		- 2
-	Tes and	6	and the second s		-	÷									-	9	- 3
· DHPU	Mumbus	Cherneys	COLIFA	-													
×X	35	12		1													
1 11	5 14	23	1	-													
*25	5 24	78	- 4,														
1	r] 53	- 56															
e -45	5 65	43	- 41														
F 12	2 52	15	1	4													
el 79	4 10	24	21	4													
- 45	5 8	31	3														
s) 52	2	11	3	-													
42	2		-	1													
c 40	0	16		1													
1 11	1	24		1													
4	-	16		1													
13	-			-													
18																	
0																	
-4																	
2																	
13																	
29																	
24																	
10101	Sheet1 Sheet	sheet)	. 8							+							
RACK.																	-

Go to Data Analysis Tool option in Excel



Select ANOVA Single factor. Here we are working on single factor.

2 2 9 + 0 - 7 5 + 4	ANDVA - Microsoft Excel Annual Vera Exercise			
A DA			* [] * [] ill ill * [true too too too too	S Salar S Den Inagon Inajon
A B C D 1 Perfect Vendor Vendor Vendor 2 Perfect Vendor Vendor Vendor Vendor 2 Perfect Vendor Vendor	Anone Single Factor Anone Single Factor Part Part of the sector of th	A L I	и . о Р	a
8 8 8 8 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1		04		aa

Add Alpha(α), input range and output range

Output of the ANOVA is as below

Anova: Single

Factor

SUMMARY

Groups	Count	Sum	Average	Variance	
Delhi	12	398	33.17	413.25	
Mumbai	8	261	32.63	483.42	
Chennai	13	356	27.38	416.43	
Kolkata	9	251	27.89	155.11	

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	303.64	3	101.21	0.271	0.846	2.852
Within Groups	14167.51	38	372.83			
Total	14471.14	41				

Table F The F Distribu	tion
------------------------	------

									α =	.05
dí _N dí _D	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
00	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Learning from Chapter

- Important of the sampling- why sampling is necessary
- Different method of sampling Simple random Sampling, Cluster Sampling, Stratified Sampling, Systematic Sampling and Quota Sampling
- What is hypothesis Testing and its utility
- How hypothesis testing is formulated and the rejection/acceptance of hypothesis
- Different types of hypothesis testing one sample, two sample, hypothesis testing of proportion, two sample from same population
- Type-I and Type-II error- how it is measured and what is the significance of the errors in arriving at rejection or acceptance of hypothesis testing
- What is Chi-Square test and where it is used
- What is ANOVA and its utility

Section - IV

LINEAR PROGRAMMING

Linear programming is one of the leading traditional methods of the optimization where we have defined goals and set of the constraints on the variables. In every field we have goals like revenue, margin, cost, quantity sales, leads generated and so on. The linear programming is an excellent tool for planning and day to day optimization of the resources based on the set goals. One of the simplest examples is for a marketing team there is a revenue target with predefined budget. The team has different channel to generate orders with different characteristics such as average selling price of each channel and cost of generating orders, maximum and minimum orders expected from the order. With revenue as objective function and cost and channel order generation capacity as constraint the linear programming can provide you the cost allocation along with the maximum revenue that is possible from the given budget and the available channels. We can have many variations of these problems. Similarly category team can allocate the discount budget to achieved sales of the month and at the same time achieve target margin by realigning discount budget among the categories. In this section we will learn how to formulate a linear programming problem and solve those in graphical method and using MS excel. Learning from this chapter will be used in other chapters such as Transportation model and digital channel optimization.

1.4.1 Linear Programming Formulation

In order to understand the linear programming let's start with basic two variable problems with limited constraints. We will solve the problem in graphical method and find the sensitivity of the solution. Once we are familiar with the graphical solution then we will solve the same problem in the excel solver.

Here machine hours are the resources and objective is to maximize the profit from manufacturing of Paint A and Paint B using given machine hours available in Machine 1 and Machine 2.

Paint A (Hours required	Paint B(Hours required	Maximum
to finish a unit)	to finish a unit)	Hours per day

Machine1	2	1	8
Machine2	1	3	8
Profit per unit	30	20	

This LP can be formulated as

Objective function: Maximize $Z=30x_1 + 20x_2$, where x_1 =units of Paint A and x_2 is units of Paint B

Constraints: Subject to $2x_1 + x_2 \le 8$ (Machine1 constraint)

 $x_1 + 3x_2 \le 8$ (Machine2 constraint)

 $x_1, x_2 \ge 0$ (Non Negativity constraints)

In this problem we are trying to maximize the profit using given resources. The solution will provide the maximum amount of the profit for the given constraints that is machine hours.

A LP can be a minimization problem with set of constraints. For example, an individual need minimum quantity of nutrients for survival. The nutrients can be calories, proteins, calcium, iron, vitamin etc. These can be obtained from different food items. The problem is to choose least cost combination of food items that gives at least the minimum requirement of the body.

Daily Allowance of Nutrients for a person

Nutrients	Daily Allowance
Calories	3000 calories
Protein	70 gms
Calcium	0.8 gms
Vitamin	5000 units

Table: Nutritive Values of Foods per dollar expenditure

Food	Calories (1000)	Proteins (gm)	Calcium (gm)	Vitamin (1000)

Wheat	44.7	1411	2.0	0
Cheese	7.4	448	16.4	28.1
Liver	2.2	333	0.2	169.2

Let X₁=Dollar Spend on wheat flour

X₂=Dollar Spend on cheese

X₃=Dollar Spend on liver in a day

Objective function: minimize $Z=X_1 + X_2 + X_3$

Subject to

Calories: $44.7X_1 + 7.4X_2 + 22X_3 \ge 3$

Protein: $1411X_1 + 448X_2 + 333X_3 \ge 70$

Calcium: $2X_1 + 16.4X_2 + 0.2X_3 \ge 0.8$

Liver : $28.1X_2 + 169.2X_3 \ge 5$

Non negativity: $X_1, X_2, X_3 \ge 0$

We can use LP formulation for solving various cases such as budget allocation, transportation problems, investment decisions etc. In next section we will learn how to solve a linear programming problem in Graphical Method. The general structure of LP is

Objective function Min/Max Z=C₁X₁+C₂X₂+ ...+C_nX_n Subject to Constraint: $a_{11}X_1 + a_{12}X_2 + ... + a_{1n}X_n \le M_1$ (Constraint 1) $a_{21}X_1 + a_{22}X_2 + ... + a_{2n}X_n \le M_2$ (Constraint 2) $a_{m1}X_1 + a_{m2}X_2 + ... + a_{mn}X_n \le M_m$ (Constraint n) $a_{11}, a_{12} ... a_{m1}... a_{mn} \ge 0$ (Non negativity constraints)

1.4.2 Graphical Solutions

A Linear Programming with two variables can be solved graphically although in practice we hardly get LP with two variables. In this example we demonstrate the graphical solution so that readers are familiar with the process.

In graphical we find the corner of the all feasible solutions. In example in the graphs, A, B, C and D are corner solution of all feasible solutions. We find out the Z (objective function) for each corner solution and select one with the highest Z as the optimum solution.



Not all the problem definition and constraint will have feasible solution there are many cases where we do not have feasible solutions. You need to take care of such problems. The two such cased are unbounded solution and non-feasible solution. The topic is beyond the scope of this book. Interested reader can explore further about the dual-primal and solving problem using simplex method.

1.4.3 Sensitivity Analysis

In LP the parameters can be changed within certain limits without causing the optimum solution to change. The analysis is known as sensitivity analysis. To understand the sensitivity analysis we will use the paint and machine problem to get general idea of sensitivity.

- 1. Sensitivity of the optimum solution to the changes in the availability of the resources (Right hand side constraints).
- 2. Sensitivity of the optimum solution to the change in unit's profits or unit cost (coefficient of the objective function).



The capacity of the machine1 is increased from 8 hours to 9 hours the new optimum will occur at G. The rate of changes in the z resulting from changing machine 1 capacity from 8 hours to 9 hours can be computed as

Delta Z= [Zg-Zc]/capacity changes = (142-128)/(9-8) = 14 per hours.

This means that the unit increase in the machine1 capacity will increase the profit by \$14. In LP term this is called dual or shadow price. If you see the graph the G can be moved to F and to B.

At B the machine 1 capacity is B(0,2.67) = 2x0 + 2.67x1 = 2.67 hours

At F the machine 1 capacity is F(8,0)=8x2+0x1=16

Thus we can conclude that the shadow price of \$14 per hours for machine 1 will remain valid from range $2.67 \le$ machine1 capacity \le 16.the changes outside this range will produce different shadow price.

Similarly we can calculate the shadow price of machine 2. It is \$2 per hours and its range is

4 hours <= machine 2 capacity <= 24 hours.

The computed limits for machine 1 and machine 2 are referred to as the feasible ranges.

Now take the case of changing the coefficient of the objective functions. Our objective function is Maximize $Z=c_1x_1+c_2x_2$

Imagine that the line z is pivoted at C and it can rotate clockwise and counter-clockwise. The optimum solution will remain at C as long as $z=c_1x_1 + c_2x_1 + c_3x_1 + c_3x_2 + c_3x_3 + c_3$

 c_2x_2 lies between the two lines $x_1 + 3x_2=8$ and $2x_1 + x_2=8$. This means that the ratio c_1/c_2 can vary between 1/3 and 2/1 which yields

 $1/3 \le c_1/c_2 \le 2/1$ or $0.333 \le c_1/c_2 \le 2$.



1.4.4 Linear Programming in Excel

We will solve the same LP problem using Excel solvers. In this section we will go step by step on how to formulate and solve a LP problem in Excel. Once LP is executed we will interpret the various reports being provided by Excel Solver.

Objective function: Maximize $Z=30x_1 + 20x_2$, where x_1 =units of Paint A and x_2 is units of Paint B

Constraints: Subject to $2x_1 + x_2 \le 8$ (Machine1 constraint)

 $x_1 + 3x_2 \le 8$ (Machine2 constraint)

 $x_1, x_2 \ge 0$ (Non Negativity constraints)

Step 1: Set variable as 0 for initial state.

HONE	HORN PAGELAND	in interior	0414 4	VEN YEA	- 10	our out the									1
XCa RhCay + Flored Fa	(alla -	H - K Z - A-A-		Diffing fait		5 - 71 - 7	 1	P	Nai	Denne A	NIN NIN	Antim - Mar Case	27 A	***	
A	4	¢ .	.p	1	ì	6	1	1	۰.	1	м	N	0		
-	Variable	X1 (Paint A)	X2 (Paint 8)	Max Profit (2)											
	Value	0	0	0											
	Contribution (Rs)	30	20	05		805									
onstraiets M	Machine 1(Hours)	2	1	0	-	8									
proprieta	Machine2(Hours)	1	3	0	-	. 8									

Step 2: Calculate the Objective function by multiplying contribution with the variable. Use sumproduct formula as shown in screenshot

G: 1	- 21 C	la manan m			Boot Wark - Do	e .						+ 3	- #
	Active Court	Content R Source R So				107 m 117 m 100 m	ee Sole by Sole notemaa Soothey not Window Poster	10 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	12				Spra
1	A		-suwwobuch) C	D		,	6	SH.	1	1	ĸ		м
		Variable	X1 (Paint A)	X2 (Paint 8)	Max Profit (2)								
		Value	0	0	*SUMPRODUCT	10101	(CS-05)						
		Contribution (Rs)	30	20	LHS		RHS						
¢	onstraints	Machine1(Hours) Machine2(Hours)	2	3	0	a	8						
F													

Step 3: Similarly for LHS calculate the resources used by multiplying variable with the machine hours required for each machines

C H S C.	n National and	MALAS SATA	ADVENT VENT	bod Not- be	e .						1.1	с – 8 ж Браф
Anna fage base free	Content Reserve Reserve					ne Sala by Sala Networks Scott and Weston Part	1001					
-		-5,4890004	01010707									
		c	Ð		+	6		1	1	κ.	5	мЗ
2	Variable	X1 (Paint A)	X2 (Paint 8)	Max Profit (2)								
3	Value	0	0	0								
5	Contribution (Rs)	30	20	LHS		RHS						
7 Constraints	Machine1(Hours)	2	1	SUMPRODUCT	(0.0)	c7:07)						
-	Machine2(Hours)	1	3	0	<i i<="" td=""><td></td><td>8</td><td></td><td></td><td></td><td></td><td></td></i>		8					
10												
11												
13												
13												
17												
19												

Step 4: Go to Data, Click on Solver



Step 5: Put E3 in the objective function. Select Max as we are solving maximization function

111 ····	· S (=). mark free is	and formular				Coor in	Wprk -	Foresoft Excel			- 0
1	the franchise of the	1	gi Lamadana gi Angantan wa Katalana	11 (11) 51 (14)	1	M. Cont Generation Sciences		HI III II- III	N N M .	Prom Datat Mary Tanat	
	of Educat Data	0	reactions		Sof & P	ta:	-	(bela holi)	Outine	in Marbolt	
		A Constraints						50%	er Parameters		
	-	N (Palet A) (C	(Pale 1)	the traffic (2			Seg Openhoes (#12)			
	Contration (%)	P I	् स	UK5	8	Red.		N Kipe Ong O Briterpresente One	per ou [
Constraint	Machine3(Mours) Machine3(Mours)	1	- 1		1			Sylpect to the Carolinantia		N	
								0.0000		8m	
										Quega	
										Deve	
										gent 4	
										Lotters	
								2 Mgr Uncentrated Tanàlés Ton-Ingat	•		
								Spint a Saling Method CRC Ion	-	Ogtore	
								Solving Helfool Soluti The ORCHINETING angles for Solvin IV angles for Integr Solvin Hallance, and added representation.	ubless hal art possili seritoan Na Doktorary angre fa Salar	Select the UP Singles publicity that are	
***	0.0.0						-	99	gine .	Oger	

Step 6: Put variable cells in the By Changing Variables codes. These are variable whose value has to be calculated.

C 12 9 . C .	1 (min) mark - Reprint	nd formula				Excel 1	Work -	Microsoft Excel	- 0
11 11 11 11 11 11 11 11 11 11 11 11 11	to the second second		g Connections g Anna Anna ar Cast Looks machines	11 (14) 11 (14)	The set at	Harris States	inter and a second	Harris Contract States	
63	+ (*) +	6 Constraints		_			_	Solver Parameters	-
2 A	vielable Velue	c In presentațion R	0 2000.00	E Man Propfit (, 2.0			September (#32) Nr Kiger Ong Operat (*	
5 0 7 Construints	Controluction (Ita) Machines(Mours) Machine2(Mours)	м 2	1	945	100	8+5 8		b dwys meth dw scortod M	
9 10 11								- An Owen	
10								(para)	
18 17								- Juilles	
18								Regis frequencies and lands the registre Special a Saling Reduct Optime Optime	
22 22 28								Shing Heffel Solar for SEContrace anyon for Solar Autilians had an anosh naminer. Solar for it's logist anyon for Solar Automations, and solar the Solatorary anyon for Solar publicits that are non-anosh.	
And D	19.9							gela Gyar	
1 (1)	0 2	6 🗉		0	w,	0		0.004	-

Step 7: Add Constraints one by one.

0 12	S (re) w	und formula		New Ye		Cricel And And	Work -	Microsoft Excel	0 = 0
191	h Barris Cherry Barris		Anne for Presentes No fatt (onla freedom	11 III 11 III	The set at	N. Cont Se Annuals Se Advanced Se	1000 1000 1000 1000	Bit Bit <th></th>	
		& Constraints						Solver Parameters	_
A			D.			6			
2	Variable	N3 (Palet A) 10	(Paint B)	Max Profile	2			lej Opelier #13	
	pranye	9	9		9			The Right Crieg Cliphands	
-	Contribution (%)	N N	.8	UK5		845		E-Charges tradit Cels	
Constraints	Machine 2(Mours)	1		_	11			Marin for Contaction	
								- MI	
1								Queros	
								(pere	
								gant 4	
								Lutters	
								Z Hige Department Variables Non-Registrie	
-								Spintabling-method distantment v Optime	
								Shang Nethol Seine For (Mit Services angue for Soline Andreas Rat are award) nonlines. Seine For V Service angue for lawer Soline Andreas, and website For Solineary angue for Soline (indices, the are non-model).	
ANN Pasts	10.22							ne pie Que	101

Machine 1 constraint

12	1) (in the second secon	and formula		nan Van	Der	Encer W	iprik + M	resoft	Excel								- 0
11	A Date (Mar da		Stannahala Stanjatisi si (at jasi) reatists	11 III 11 III	T.	Hi, Cinar Go Arnanaly Go Astronomik M	India India India	-	Tota Tota	12	100	1	Caller Caller	4] 2000 	Contant Contant	Jahor .	
67		& Centrams														_	_
		¢	D		1	6			1		. 6	M		0		.0	
	Warlable Warlab	N3 (Paint A) 10	(Paint II) A	Han Propile (2)													
	Communer (N)	18	20	UK.		6+5											
Constraints	Machine2(Mouri) Machine3(Mouri)	1		4	EE												
											Ad	d Constrain	¢.				
									C(FALM MUT	-			-soet		N		
										×)		-		Great			
A R. Darf	10.0																

Machine2 constraints



Non Negativity constraint

1	To an and the second	nd famous	100			E-cel	Wprk - N	ficesoft	Excel								- 0
1211	h h	1 (a)	i Lamadan I Propisian na Lati (mis) median	11 11 11 10	- T	K. Cont Gertransty Sectorements	And the local division of the local division	H.	Euro Data	12	No.	10	and the second	1 1 1 1 1	Coluit 1	a Selver	
A7	*(*)	6 Constraints		_			_										
- A		0	0					1	A	ĸ		м	. N	0		0	
-	vuluble Volve	NT (Part A) AC	(Puint II)	Max Profit	92												
	Contribution (%)	M	20	96		845											
Combulata	Machine((Mount) Machine2(Mount)	1	-	_	-												
											Ad	d Constrain		3			
									GENA	renae.			age we				
									80.60		0	** *	۹		N.		
										×		-		Qent			
1																	
a a a Deal	1.10.12									141	_			1.00.0			

140.00	3.0-14				Excel Work	· Moresoft	Encel							- 0	
ALC: No. AL	mark Reprik	out formate 8	eta Review Vi	-	toper Add.mo									- 0 -	00
	A BA		anten 11 11	T.	A Cont Cont Cont Cont Cont Cont Cont Cont	C BI	100	13-	100	10 3		4] (100) 4] (100)	100 2 10	-	
47		6 Constraints													-
		< 1			6 1	0.1.0	1	. К.	. 6	M		0		Q 8	
	Value	13 (Paint A) 12 (Paint 0	el B) Max Propilit C	2											_
	Contraction (%)	8	20 145		845										
Constraints	Machine3(Mours) Machine3(Mours)		-	a:E	-										
									Ad	d Constrain	e				
							0/14/6 (010	-	-		ingen and		8		
							-	×)		-		Center			
*** Per	0.00.00						1						-	-	-

Step 8: Change the setting in Options to set as per your requirement

049-0	· To get a	and formula	-			Exten	Wprk -	Konsoft Ease	- 0
	the function of		an fan jost	51 (1) 51 (1)	7-1	A Cont Bolicana) De Alexande	100	Annual Statements State State	
	Let External Bata	() 6. 6. 6. 6. 6. 6. 6. 6. 6. 6. 6. 6. 6. 6	readiate		tod & P	figi	-	Gela Taolti Guillea S. Anajoti	() · · · · · ·
-	-	6	D					Sover Parameters	
1 2 3 4 5 6	Versite Versite Contraction (%)	01 (Part A) (O 0 0	(Pacer 11) (2) (2)	Was Profile (2)	2	845		Nel Olestre III Nel IIII Nel III NEI IIII NEI IIIIII NEI IIII NEI IIIIIII NEI IIIIIII NEI IIIIIII NEI IIIIIII NEI I	
7 Canadraliat	Machines(Mours)	1		2	2-1		i i	K080 N	
	Machine 3(Mours)		- 1		200	-	6	Agent to the Construction	
38								6.0 ta 0 60 ta 0	
11								Rill 1+ Kill Darge	
8								Deve	
5								grant of	
17								- Uniting	
18								Z high choreshared fandes for Argebre	
28								Split a Salvey Nethod OAL normean (*) Optime	
8 22 23 24 24								Solvay Shehol Salah de Salah Salah yang ke Salah Andrea And an prosti naringan. Salah Yenyi Salah yang ke Salah Andrea, yai salah te Salah yang ke Salah yang ke salah te sa salah prosti.	
	Q. 40.13							pés Que	2008

Go to options and make changes if required

0 M 4 - 6 - 8 20 4	Excel Work - Microsoft Excel -	0
Name Name Page Light Parame Parame<	Contract Laboration Contra	0-0
A7 K Executive A B C D Verside N3 (Prior A) (2) (Prior A) Name Nutriti (2) Verside N3 (Prior A) D Name Nutriti (2) Verside N3 (Prior A) Name Nutriti (2) Verside Construints Machime (Prior A) Name Verside Construints Machime (Prior A) Name Verside Construints Machime (Prior A) Name Name R Name Name Name Name Name Name Name Name Name Name Name Name Name Name Name Name Name Name Name Name Name Nam Nam <	Conservation (Service)	*
2 2 3 3 • • • • 9++0 • • • • 9+		

Select Simplex Linear model for linear problems

011-0-0	2.0-04					Code	Work -	Microsoft Eacel	- 0
	handber at		Constant Constant Set Satisate	11 (11) 11 (11)	Y	A Cont Sectors	in the second	Harris Contraction States Contraction Cont	
		& constraints			-			Solver Parameters	
A		¢	Ð			6			
	Variable Value	N3 (Parine A) (C)	(Build B)	Was Profit	0			Ne Rev Org Operation	
_	Contrabution (%)	N I	20	US.	-	845		britegeg testik bis Konstel	
Construints	Machine 2040-01		- 1		40			Sugard to the Constraints	
								8.07 (+-1) 200 (+-1)	
1								dist in Kill Owige	
								Deve	
								Brand Ad	
3								- justice	
								2 Mgr Uncentrated Tanàlés Non-Negative	
								Spin a Scheng Herholt OLI Linkings III - Opport Scheng Herholt Berlander - Opport Scheng Herholt Berlander - Opport Scheng Herholt Schen Herholt Hart Jack Herholt Her	
AXA Dest	10/10						-	940 (JAA 0ge	

Step 10: Click on Solve button

	Y 2= /*	od female	Des New		Excel W	ork - Micror	oft Excel					- 0
	trans (Mar Land		Cannediana 7 Augustus et Lational	Li San Pan	Garante a	The first of the local division of the local	Data			10 10 1	Inter Selat	
		6 Constraints	CAN LABOR DE						Solver Parame	ters .		
A		6	0	. E								
	variable value	NI (Parint A) NI D	(Paint B) Ma	Popla (2)			Opertor # Des	and the	Operation		N	
	Contribution (4)	N	10	945	845		Campo Tanalia	-				
Constraints	Machine (Johns)	-		-	-		0.00	1111			8	
						- 6	12 1+0				-	
							17 co 8647 8 co 868				Quege	
						- 11					Quera	
											grant 48	
										100	Lucifice	
							Hele Unionation	ed Variables Name	and a state			
							att a Salary Metho	• 10	100		Options	
							ning Nethol delt Ne (H), turk gine for trans (H) manach.	nas engre fo b	ing Problems Parl of adult Par (Indultan		Select Ren (* Singles publics Ret av	
AT AT DAME	14.0					-					1.1	-
100 100	and the second second						94			Bet	Olive	1-Querry

Select Keep solver Solution if value is to be populated in variable cells.

C C - C - Y Serve	and female (br) from	Exter Work - Microsoft Excel	- 0
The Part Part Part Part	Internet Secondary Grandson Constant	Set 1 for the constant of the	Colley (Marcola College)
A2 + A 8 1 A 2 Variable 3 Control variable 4 Control variable 5 Control variable 7 Control variable 8 Variable 9 Variable 10 Variable 11 Variable 12 Variable 13 Variable 14 Variable 15 Variable 16 Variable 17 Variable 18 Variable 19 Variable 10 Variable 11 Variable 12 Variable 13 Variable 14 Variable 15 Variable 16 Variable 17 Variable 18 Variable 19 Variable	St. Conventioning 0 D 01 Protect Al: NO (Partie H) 02 3.0 02 2 02 2 03 10	Solver Results & M Solver Results So	N 0 P 0 N
a		84	

Select reports you would like to see. I have selected all reports.

Q 10 - P-	Y 24 Y				factor in	lock - Microsoft	Eanel								- 1	,
	man Appropriat	Namutal Da	to Berley 1	tes Destage											- 0 -	
1 12 11	5 13 I B	(a) (a)	metters 11 1	TY		2 11	12	10	123	*	18		how Detail	S Salver		
Fram. Fra	a free Otter Letting	Antonia un Las	1 10	e new year	anost .	forthe Ramove Columni Dupmuter	Data Vandation *	Canandada	Anatypes	-	ingras be	and at 1				
64	Edward Data	Catrantia	-	Soli & Film	_		Data faun	_			0.4			Market .	_	-
*		C D		1. 4	0.1115		1	×.	-	м		0		. 0		1
	Contrato Data	and a line of	and here		-	Solver Results										
	Value	12	3.4	other found a solution	-	matrix and apt	nally									
	Contribution (%s)	NI.	28	Parana Parana			and and a second	1	- 11							
and whether	Mechanel(Mouri)	1		Oterration			100									
	Machine2(Mount)	-		0 Beer offer a	-											
			6	3 fignum to Solver P	*****	0.eing	Dev	ine Reports								
				¢K.	General		- 20	e Scanaria	-							
				olver found a solution	n. All Co	to show and succe	ulty undit	-								
				nished. Then the DAG engine	-	Somer has found a										
				plution, Bhan Singi plimal splution,	er# 3 +	set, this means to	ner has four	a s from								
																-
																- 1
* Deef	10.0							I • (_	_	_		-	an (7)-1	0	1
1 2. 6	and arrest					bot that - bod								*	× -	2.2
0.0	P P	Ph 22 Ge	callers at 77		2	A. 110	2.2		0	172	-5	10	100	11 10	25	4.01
LO LI	hun (the faire	Adult Des	nta la la	A Ada Sea		Tatto Fah	ETC.	Des (11	the t		- 5mg	CII I	10.10	2.	
the fail	Status Connection	a Alt Citor	1000 M	Set & Sale	Larval	Columns 74	Duplicates Ve	Editor." Editoria		Analyse *		1080	- Cultria			
	•] / [A	1.5														-
A.		c		с.			w. 1	1.1		κ.			N.			4
	1000															
	Variable	X1 (Peint A)	X2 (Paint B)	Max Profit (Z)	4											
	(Table				·	1.1										
	Contribution (Rs)	30	20	05	-	8945										
	Constant For															
nstraints	Machine1(Hours)	2	1	1	•											
nstrainte	Machine1(Hours) Machine2(Hours)	2	1		-											
nstrainta	Machine3(Hours) Machine2(Hours)	2	3		4											
protrainte	Machine1(Hours) Machine2(Hours)	2	1		4	8										
onstraints	Machine1(Hours) Machine2(Hours)	3	1		4											
onstraints	Machine2(Hours) Machine2(Hours)	2	1		4											
ionstraints	Machine3(Hours) Machine3(Hours)	3	3		0.0											
enstraints	Machine3(Hours) Machine3(Hours)	3	1		4											-
ionotrainte	Machine2(Hours)	3	1		6 4 6											
Constraints	Machine (Hours) Machine 2 Hours)	3	3		0.4											-

Final output is X1=3.2, X2=1.6 and Z=128. The same result was obtained in Graphical Method.

How to Interpret Reports

Answer Report

Objectives cells provide you the final result of the Z.

Variable Cells provides the final value of each of the variables.

Constraints section contains the binding status of the constraints. In the problem both Machine1 and Machine 2 is binding which means all available machine hours is used to obtain Z=128. If there is slack then there is some available capacity which is still not used.

H N	NE NORT FACELANDUS	PORMANS DA	the Advent Ve	a borth	v4 bod								1 10	- 0 5pr
X Cal	antinene (anten el)e antinene (el ¥ + 2 ¥ + ⊡ + antinene (el × 1 × 1)		0 - 5 mm	National Allantar - S + 1 N N	•• 9.8	(Landiana	- 144-	28.0	22 · 3		E Autolas E fait - R Chart	tana tana tana	A	
	* A V													
		. D			- 10	- 10		F. 1	2.6	1		N	0.1	
20jectiv	re Cell (Max)	October 1 Mahar	Read States											
5151	Volue Max Profit (7)	Congress value	O I	78										
3133	A 1976 Mark From (23		v 1											
(wishle	Cells													
Cell	Name	Original Value	Final Value	Integer										
\$C\$3	Value X1 (Paint A)		0 3	1.2 Contin										
5053	Value X2 (Paint 8)		0 1	1.6 Contin										
	1													
Cell	Name	Cell Value	formula	Status	Slack									
5157	Machine Eliterat LHS		# \$1574-5057	Rinding	0									
5158	Machine201oundUK5		8 51584-5658	Binding	0									
\$(\$3	Value X5 (Paint A)		3.2 5532-0	Not Binding	3.2									
\$053	Value X2 (Paint 8)		1.6 5053>-0	Not Binding	1.6									
	Annual Report 1 Denibul	ty Report 1 Limbs	Report 1 UP	8			*							
													_	-

Sensitivity Report

In section Adjustable cells columns Final Value give the optimum solution to the problem.

Allowable increase and allowable decrease of coefficient of each variable indicates range within which the current optimum solution will remain optimum.

For Example for X1, current optimum solution will remain optimum for coefficient range of 30 - 23.33 to 30 + 10.

In the constraints section final value indicates the utilization of the resources. Allowable increase and allowable decrease of RHS value of each constraint gives the range of the value within which the current optimum will remain optimum. For example, the current optimum solution will remain optimum for the machine1 capacity range of 8 - 5.3 to 8 + 8.

Shadow price is the rate of change of optimum value of the objective function with respect to the change in the values of RHS of the constraints. The shadow price is valid for binding constraints only. For example in the machine 1 constraints shadow price is 14, allowable increase is 8 and allowable decrease is 5.33. Hence within 8 - 5.33 to 8 + 8the unit change in the machine hours of machine1 will change the value of objective function by 14. Similarly for Machine2 the objective function will changes by 2 for each unit changes in machine 2 capacity within range of 8 - 4 to 8 + 16. This is the same result we obtained in Graphical method.

Conton - (14 + 7 4 - 2 	S x x = = S → A = = = A = = a sport ta7 PM	a de Serai E este Este Agenti	hap fast i fa hap & Conse - 5 4	4	The second secon		× 🗵	E Adoler E Cart	tring	N P
14.0 Sensitivity Re el Work.siss().P 12/23/2015 11:57	A port	Kiyaant		Auto (N .		(m) (talong M	N TR
14.0 Sensitivity Re el Worksits(LP 12/23/2015 11:57	A iport	t		4				- 1 SK		N [0
(14.0 Sensitivity Re el Work.siss)(P 12/23/2015 11:57	арант 1.37 РМ	,		4	H			k	M	N 0
14.0 Sensitivity Re el Work.xisx)JP 12/23/2015 11:57	nport 1:37 PM									
	Final	Reduced	Objective	Allowable	Allowable					
Name	Value	Cost	Coefficient	Increase	Decrease					
X3 (Paint A)	3.2	0	30	10	23.3					
X2 (Paint 8)	1.6	0	20	70	5					
100 M 100	Final	Shadow	Constraint	Allowable	Allowable					
Name	Value	Price	R.H. Side	Increase	Decrease					
ne1(Hours) LHS		14			5.3					
we2(mount) this		2		19	4					
	Name K1 (Paint A) K2 (Paint B) Name ne10hours) UrS ne20hours) UrS	Final Final Name Value K1 (Paint A) 3.2 K2 (Paint B) 1.6 Final Final Name Value net(Droard) U/S 8 nc2(Pount) U/S 8	Final Reduced Name Value Cost K1 (Paint A) 3.2 0 K2 (Paint B) 1.6 0 Final Shadow Price Name Value Price net(Priours) Ur5 8 2	Final Reduced Objective Name Value Cost Coefficient K1 (Paint A) 3.2 0 30 K2 (Paint B) 1.6 0 20 Final Shadow Constraint Name Value Price R.M. Side net/Prices 8 2 8	Final Reduced Cost Objective Coefficient Allowable Increase Name Value Coefficient Increase X2 (Paint B) 3.2 0 30 10 X2 (Paint B) 1.6 0 20 70 Final Shadow Constraint RM, Side Allowable Increase Name Value Price R.M, Side Increase ne(2)fourn) UPS 8 2 8 16	Final Reduced Objective Allowable Allowable Name Value Cost Coefficient Increase Decrease X1 (Paint A) 3.2 0 30 10 23.3 X2 (Paint B) 1.6 0 20 70 5 Final Shadow Constraint Allowable Allowable Decrease Name Value Price R.H.Side Increase Decrease Name Value Price R.H.Side Increase Decrease Name/Unloss 8 14 8 8 5-3 nc2(hours)Uris 8 2 8 16 4	Final Reduced Objective Allowable Allowable Name Value Cost Coefficient Increase Decrease X2 (Paint R) 1.6 0 20 20 23.3 X2 (Paint R) 1.6 0 20 20 5 Final Shadow Constraint Allowable Decrease Decrease Name Value Price R.K.Side Increase Decrease Name Value 916 2 8 16 4	Final Reduced Objective Allowable Allowable Name Value Cost Coefficient Increase Decrease X1 (raint A) 3.2 Cost Coefficient Increase Decrease X2 (Paint B) 1.6 0 20 20 5 Final Shadow Constraint Allowable Allowable Name Value Price R.N. Side Increase Name Value Price R.M. Side Secrease Name Value Price R.M. Side Secrease Name Value Side 8 5.3 No(Nourci LIYS) 8 2 8 16 4	Final Reduced Objective Allowable Allowable Name Value Cost Coefficient Increase Decrease X1 (raint A) X2 (Paint B) 1.6 0 20 20 5 Final Shadow Constraint Allowable Allowable Decrease Final Shadow Constraint Allowable Allowable Decrease Name Value Price RM.536 B 5.3 Name(Nours) Ur5 8 2 8 16 4	Final Reduced Cost Objective Coefficient Allowable Increase Allowable Decrease X1 (raint A) X2 (Paint B) 1.6 0 20 20 5 X2 (Paint B) 1.6 0 20 20 5 Final Shadow Constraint Allowable Allowable Name Value Price RUL Side Increase Name Value Price RUL Side Decrease Name Value Price RUL Side Allowable Name Value Price RUL Side Allowable Name Value Price RUL Side Allowable

Limit Report

This report simply provides the value of the variables and the objective function.

HOME	POINT FROELEHOUT	PORMAN	DATA ADV	W VEW											Sep
X Cat Bit Cater	Caller - [14 0.7 10 - 21 - 2	3x x = 1		Difficulture Difficulture	- 5 - % +	1 1 Core		2	1	Serie Farma	2 A.A.	fum +	ET I	A	
Datend	4 844		Ange	and its	1 1.00	- 1	ityles	. she.		(m.	6.00	tain	1441154	fact 1	
	* 1 A V	A 1													
					er ar	10			м	N			0		
	Objective														
	Name	Value													
Cell	Makes Have Beerford's	8.24													
5453	Value Max Profit (2)	178													
5(5)	Value Max Profit (2) Variable	128	Lower	Objective	Upper	Objective									
Cell Cell	Value Max Profit (2) Variable Name	128 Value	Lower	Objective Result	Upper Limit	Objective Result									
Cell (Cell (CS3	Value Max Profit (2) Variable Name Value X3 (Paint A)	128 Value 3.2	Lower Limit	Object/ve Result 0 32	Upper Limit 3.2	Objective Result 128									
Cell (C53 (053	Value Max Profit (2) Variable Name Value X1 (Paint A) Value X2 (Paint B)	128 Value 3.2 1.6	Limit Limit	Objective Result 0 32 0 96	Upper Limit 3.2 1.6	Objective Result 128 128									
Cell (CS3 (053	Value Max Profit (2) Variable Name Value X1 (Paint A) Value X2 (Paint B)	128 Value 3.2 1.6	Lower Limit	Objective Result 0 32 0 96	Upper Limit 3.2 1.6	Objective Result 128 128									
Cell (153 Cell 5(53 5053	Value Max Profit (2) Variable Name Value X3 (Paint A) Value X2 (Paint B)	128 Value 3.2 1.6	Lower Limin (Objective Result 0 32 0 96	Upper Limit 3.2 1.6	Objective Result 128 128									
Cell (CS3 (053	Value Max Profit (2) Variable Name Value X1 (Paint A) Value X2 (Paint B)	178 Velue 3.2 1.6	Lower Limit (Objective Result 0 32 0 96	Upper Limit 3.2 1.6	Objective Result 128 128									
Cell 5(53 5(53 5(53 5(53)	Value Max Profit (2) Variable Name Value X1 (Paint A) Value X2 (Paint B)	128 Value 3.2 1.6	Lower Limit	Objective Result 0 32 0 96	Upper Limit 3.2 1.6	Objective Result 128								_	_

Readers can solve the diet problem using MS excel as a practice.

Learning from the chapter

- How to formulate problem into linear programming problem
- Graphical representation of the Linear Programming and solving using Graphical solution
- How to do sensitivity Analysis in graphical solution
- Formulation and solving of LP using Excel
- How to interpret the answers from the Excel output
- Understanding of the shadow price, sensitivity, Limit in the reports

Chapter - II ANALYTICS SYSTEM

"Give me six hours to chop down a tree and I will spend the first four sharpening the axe." -By Abraham Lincoln



Understanding the analytics system, collecting the data from disparate sources and preparing the data input to the analytics system takes lot more time than actual time analyzing the data. Therefore it is really necessary for the analyst to know the data requirement to meet the business objectives, availability and format of the data from different sources, how to create data warehouse from different data sources, how to prepare data as input to the algorithm, how to diagnose the error in the system and how to interpret the output of the analysis. Understanding the algorithm and execution of algorithm is often very small part of the overall process. In this context this chapter is an attempt to make users familiarize with the some of the commonly used analytics system and applications.

Every company has a reporting system in one form or another to track the past performances and make future prediction for decision making. In this book we will call reporting system analytics system in line with the theme of the book even though there is thin line difference between both. The small organization can have very basic system like maintaining data in MS excel or other spreadsheet while bigger organization do often have elaborate system with data being maintained in database or data warehouse. Even company/organization/unit with no computer system like small kirana (shop) does keep track of the sales, credit and inventory through register in the notepad using pen and paper. The analytics system is generally derived system because it doesn't generate any data of its own apart from the derived numbers. The basic data invariable comes from fields, website, order system, ERP, CRM and so on. Therefore the objectives of analytics system are to keep data (data warehousing), generate report for the users as per their requirement, and perform analysis to get insight from the data.

Every organization has different way of looking at the data but there are certain common features that an analytics system must have to be called a BI/analytics system

- 1. Single Database containing data from different system to enable one sources of information
- 2. The system should have some mechanism to disseminate the information to the users. This can be in the form of spread sheet report, a monitor, a printout or a pdf form.
- 3. The system should be capable of extracting data from different sources and warehouse the data in one central database.
- 4. The system should have capability to generate reports and analysis as per user's requirements.

In this book we will learn three analytics system-

Pentaho BI for understanding the implementation and application of BI system using database, ETL, OLAP cube and Dashboarding.

- 1. Google Analytics for understanding web analytics
- 2. R for statistical analysis

All above application are open source and are among the best and commonly used business world. I have chosen open sources as readers can actually build their own system with minimum of the hardware and software requirement. Pentaho has a community edition as well a professional edition. We will be using community edition as it is free to build a simple system with all basic functionality like database, Extraction Transformation and Loading (ETL), Reporting Cube, Report designer, Automatic report scheduling and so on.

In second section of the chapter we will learn the analytics tools R and their basic functionality. The R is an open source tools extensively used for statistical analysis and machine learning. What we will be learning and using would not even amount to scratching the surface. Nevertheless as per our requirement in each of the chapter we will learn how to use R for the algorithm or model in the question. The readers are suggested to read more detail of the R from the R specific book to get better handle of the R. There are many online resources available in the internet; the readers can use those as well.

Section - I

BUSINESS INTELLIGENCE SYSTEM

In this section we will learn a Business Intelligence tool using Pentaho BI system. Pentaho is an Open Source Business Intelligence solution that has all the component of a business intelligence system. The major components of the Pentaho that will be discussed in this chapter are

- Pentaho Server- Community Edition
- Extraction, Transformation and Loading (ETL), also called Kettle
- OLAP Services Mondrian
- Reporting and Dashboarding Community Dashboard
- Web based front end called Saiku

Before the actual installation of the Pentaho Solution in the system, user has to prepare the operating system with necessary resources to install the entire component. The Pentaho manual provides extensive information on the system requirement. The amount of the resources in term of memory and CPU depends on expected umber of the users and amount of data processing that is expected to be done in the system. However for the user who want to set up reporting system for learning purpose, the system can be as small as a virtual machine with Linux Operating System inside a windows systems in your laptop or desktop. In this book we will be using Amazon EC2 server with Centos 6.5 Operating System. The user can install Centos Operating System in virtual machine in their Windows machine. The real BI system for users can have different configuration based on number of active users and other task to be handled by the server.

In this book we will not be showing the actual installation steps which are available in the Pentaho Support site, manual and bloggers sites. We will assume user has installed latest version of MySQL database in the same machine or different machine, java version required by Pentaho Version, Pentaho Server Community Edition, Pentaho Data Integration (PDI), Schema Workbench or Mondrian, Saiku Analytics, Reporting, Dishoarding and so on. We are not describing the detail of installation as most of the time installation requires technical resources; we will learn as a functional user of the system who is concern with report, OLAP and dashboard.

Pentaho Data Integration has to be installed separately. In fact PDI can be used as independent application without installing Pentaho BI server. PDI windows version can be downloaded in Windows machine and used as it is by adding MySQL driver in the lib file. For rest of the component the installation happen from the Market Place after the BI server is installed.

After installing Pentaho BI server, one has to open the 8080 port to access it from different machine.

The Linux command to start the Pentaho Server is *\$./start-pentaho.sh* and command to stop Pentaho is *\$./pentaho.sh*. After installing you have to restart the server.



Once the server is up, enter URL http://xxx.xxx.xxx.8080/pentaho in the browser to open the login windows. The Login window is as shown in picture below.



In the first instance you have to login using admin user id. Once you login into Pentaho, you will see below windows.

Browse Files	Pental	ho Busine	ss Analytic	S Try Enterprise Edition
Manage Data Sources	Learn .	Contribute	Meet the Family	Upgrade to Enterprise Edition
Documentation	The Pentaho Bur	iness Analytics Platform delivers an		
rts	Ether by provide	g ad-hoc tools to analyze and		1
test saku 🔅	visualize data or different content dashboards, data created plugns c	by serving pre-created content, types including analyse, reports, mining and even community an be seamlessly connected.		

As we have to installed different components, go to Browse **Files** -> **Marketplace**. The market place contains all components free and paid available with the version of Pentaho.

Browse Files				admin +
NOTE		11223	100000000	
Brokoe Files	0	roes	Folder Actors	
Opened Some		There are no fire in this labor.	New Folder_	
Marketplace			Move to Trach	
Schedules			Returne.	
Administration				
			upload	
			Download	
			Properties	

Install the required components from the list. If some components are not likely to be used you can leave it.

Nat - Junix Kanto - Dut : # 💍 anime procey - Doople 1: # . 🐨 Pendite - Wispedia, We 1: # - 🙃 Doople Acces	ns x/ @ Renalis User Console x	1000 · 0 1130
		▼☆ 単
ter ven toet nete serbeveren Marketplace ~		admin +
Available Installed		0
Type At V SuperAt V Such :	٩	O labor are staged
Percaho Marketplace Nentaho	() mader	s sa duar 🖌
Community Dashboards Ramework medical	Installed	p to Conte
Community Data Access methodala	() Least-Mintern	y to Date 🖌
Community Deshboard Editor	() testaled	to 16 Durine 🚽
Community Graphics Generator memorals	() testinger	y ta Dute 🛹
Taula Baseline Sectores Bullet	A Linear	

An example of installation of a component.

Do you want to Instal? Two are about to restal motel, Analysis, Do you want to		
proceet	red .	E.
	144	6
	1000	
	nest -	5
		C.
	Treed 1	

Once all components are installed, we are ready for using the Pentaho for real business cases.

2.1.1 Pentaho Data Integration (PDI)

Pentaho Data Integration can be downloaded from the same site as Pentaho BI server. The download file is in zip (.tar) format. Once downloaded the file has to be unzipped into the folder where you want to keep the PDI. It is advisable to keep it in the same folder as BI server. Once the unzip is completed, go to the data-integration. The command to start PDI is *\$./spoon.sh*



By default spoon comes with very minimal memory configuration. Hence it is advised to increase the memory as per available RAM in the server. Just open spoon.sh file in vi or vim editor and change the memory setting as shown in pictures below.



Change the MaxPermSize as shown in below picture. I have changed it from 256MB to 5000 MB.



Start the PDI by typing the command as shown in below picture.



The first screen of PDI as system opens the windows.



The transformations are the basic building block of PDI. Any activity is either a transformation or a collection of transformation. The collection of one or more elements which does a unit task can be called as transformation. The group of transformation in a sequence completing a particular process forms a job. The job can also contain a job. Before delving deep into the transformations and jobs, lets us create an environment or a business case (dummy data) for using PDI for solving a problem. This data will be used for other components like OLAP cube and Dashboarding.



We will be working on the start schema as shown above. The Sales data of XYZ Company is linked to product catalog by the productid, customer by customerid and calendar by date. For ease of implementation we will be using smaller data set but enough to make meaningful data. The sample of each data set is as below. The table will have data related to customer name, city and marital status.

CustomerId	Name	City	MaritalStatus
CC00001	Customer1	New Delhi	Married
CC00002	Customer2	Bangalore	Married
CC00003	Customer3	Hyderabad	Married
CC00004	Customer4	Chennai	Single
CC00005	Customer5	Mumbai	Married
CC00006	Customer6	New Delhi	Single

Calendar Table has information related to the year, month, financial year, quarter, weekday etc.

Date	Year	FY_Year	Month_ Name	Year- Month	Quar- ter	Day	Weekday
4/1/2015	2015	2015_16	04_Apr	2015_04	Q1	1	Wednes- day
4/2/2015	2015	2015_16	04_Apr	2015_04	Q1	2	Thursday
4/3/2015	2015	2015_16	04_Apr	2015_04	Q1	3	Friday

4/4/2015	2015	2015_16	04_Apr	2015_04	Q1	4	Saturday
4/5/2015	2015	2015_16	04_Apr	2015_04	Q1	5	Sunday
4/6/2015	2015	2015_16	04_Apr	2015_04	Q1	6	Monday
4/7/2015	2015	2015_16	04_Apr	2015_04	Q1	7	Tuesday
4/8/2015	2015	2015_16	04_Apr	2015_04	Q1	8	Wednes- day
4/9/2015	2015	2015_16	04_Apr	2015_04	Q1	9	Thursday
4/10/2015	2015	2015_16	04_Apr	2015_04	Q1	10	Friday
4/11/2015	2015	2015_16	04_Apr	2015_04	Q1	11	Saturday
4/12/2015	2015	2015_16	04_Apr	2015_04	Q1	12	Sunday
4/13/2015	2015	2015_16	04_Apr	2015_04	Q1	13	Monday

Product table has product name, category of product, subcategory and the listed price.

Productid	ProductName	Category	SubCategory	Price
100000	Product name 1	Mobiles	Smartphones	20000
100001	Product name 2	Mobiles	Featured Phones	2400
100002	Product name 3	Mobiles	Smartphones	25000
100003	Product name 4	Mobiles	Smartphones	15000
100004	Product name 5	Mobiles	Featured Phones	1800
100005	Product name 6	Mobiles	Featured Phones	3400
100006	Product name 7	Mobiles	Featured Phones	3000
100007	Product name 8	Laptop	Windows	23000
100008	Product name 9	Laptop	Windows	35000
100009	Product name 10	Laptop	Windows	43000
100010	Product name 11	Laptop	Mac	56000
100011	Product name 12	Laptop	Mac	50000
100012	Product name 13	Laptop	Mac	46000
100013	Product name 14	Apparels	Sarees	1800
100014	Product name 15	Apparels	Jeans	2200
100015	Product name 16	Apparels	Sarees	900
100016	Product name 17	Apparels	Jeans	1200
100017	Product name 18	Apparels	Sarees	1400

Productid	ProductName	Category	SubCategory	Price
100000	Product name 1	Mobiles	Smartphones	20000
100001	Product name 2	Mobiles	Featured Phones	2400
100002	Product name 3	Mobiles	Smartphones	25000
100003	Product name 4	Mobiles	Smartphones	15000
100004	Product name 5	Mobiles	Featured Phones	1800
100005	Product name 6	Mobiles	Featured Phones	3400
100006	Product name 7	Mobiles	Featured Phones	3000
100007	Product name 8	Laptop	Windows	23000
100008	Product name 9	Laptop	Windows	35000
100009	Product name 10	Laptop	Windows	43000
100010	Product name 11	Laptop	Mac	56000
100011	Product name 12	Laptop	Mac	50000
100012	Product name 13	Laptop	Mac	46000
100013	Product name 14	Apparels	Sarees	1800
100014	Product name 15	Apparels	Jeans	2200
100015	Product name 16	Apparels	Sarees	900
100016	Product name 17	Apparels	Jeans	1200
100017	Product name 18	Apparels	Sarees	1400

The Sales table will have the quantity sold, customerid, productid, date of sales, amount, discount and the status of order.

Orderid	Customer Id	Order Date	Prod- uct id	Quan- tity	Amount	Dis- count	Status
OD000001	CC00096	1-Apr- 15	100015	2	1800	0	Returned
OD000002	CC00043	2-Apr- 15	100022	2	1040	125	Completed
OD000003	CC00098	3-Apr- 15	100008	2	70000	3500	Cancelled
OD000004	CC00068	4-Apr- 15	100013	3	3600	180	Completed
OD000005	CC00086	5- Apr-15	100013	3	3600	0	Completed
OD000006	CC00006	6-Apr- 15	100011	1	100000	12000	Completed
OD000007	CC00061	7-Apr- 15	100002	1	50000	0	Completed

		8-Apr-	1000000	7728	-01210101		
OD000008	CC00087	15	100029	1	3000	360	Completed
		9-Apr-					
OD000009	CC00067	15	100003	3	30000	3600	Returned
		10-					
OD000010	CC00094	Apr-15	100002	3	50000	0	Completed
Contraction of the second	the same for	11-				1.01	
OD000011	CC00073	Apr-15	100018	1	4600	552	Completed
		12-					
OD000012	CC00014	Apr-15	100000	1	40000	0	Completed
		13-					
OD000013	CC00069	Apr-15	100007	2	46000	0	Completed
		14-					
OD000014	CC00034	Apr-15	100025	3	4600	0	Completed
		15-					1
OD000015	CC00094	Apr-15	100023	2	1000	0	Completed
		16-					
OD000016	CC00026	Apr-15	100007	2	46000	0	Completed
		17-					
OD000017	CC00059	Apr-15	100016	1	2400	120	Completed

Use below MySQL query to create table in the **Pentaho** database. I am assuming you have installed MySQL and have sufficient access right to do read, write, delete, and modify operations. Use below query to create respective tables

Create Table Customer

```
(CustomerId varchar(50) primary key,
```

```
Customername varchar(100),
```

```
City varchar(30),
```

```
maritalStatus varchar(20)
```

)

```
Create Table Calender
(
Date date Primary Key,
Year INT,
```

```
FY_Year varchar(12),
Month_Name varchar(20),
YearMonth varchar(20),
Quarter varchar(10),
Day INT,
Weekday varchar(20)
)
```

```
Create Table Product
(
Productid INT primary key,
ProductName varchar(50),
Category varchar(50),
SubCategory varchar(50),
Price INT
)
```

```
Create Table Sales
(
Orderid varchar(20) primary key,
CustomerId varchar(20),
OrderDate date,
Productid INT,
Quantity INT,
Amount Numeric,
Discount Numeric,
Status varchar(30)
)
```

Once the tables are created the next logical step is to upload the data files to the respective tables. We will use PDI to upload the data into the tables. The basic steps is to connect the MS excel files located at predefined location, read the data from excel file and write the data into the table. For this operation we will require two elements – an excel input elements and table output elements. Both elements will be hold together in a transformation.

Go to File -> New -> Transformation



In left hand side you will see all the available elements classified as per the functions it performs. The transformation window is blank as of now.

V2	ip-10-0-0-87.ap-southeast-1.compute.internal:1 (ec2-user) - VNC Vie	wer = 0
×	Spoon - Transformation 1	_ 0 X
File Edit View Action Tool	is Help	
		Perspective: 👯 Data Integration
View / Design	Welcomet III Transformation 1 11	
Steps 🔞 🖌 🖬	▷ II □ ● Ø ▷ IX, © Ø Q, III 100% ~	
> 🗀 input	3	6
Output		
Transform		
Utility		
> 🛅 Flow		
Scripting		
BA Server		
Lookup		
Joins		
Data Warehouse		
Validation	Drag & drop from the	
Statistics	design panel to begin.	1
👂 🛅 Big Data		
> 🗅 Agile		
Cryptography		
P 🗀 Palo		
Open ERP		
o 🗖 job		
Mapping		
Bulk loading		Authenti
a Per talipa	B 2	

On expansion of input we can see all possible input file system.



Look for Microsoft Excel Input from the input list, select it, drag and drop it in the transformation windows.

0	Spoon - Transformation 1 (changed)	- 0 -
Fin fat time Action facts Help		Perspective 🔠 Data Integration
B Ver / Design	O Intranel (III Sandamation 1)	
Sept. La El	DIDOODXCBC II WA +	
Gorge Anaples Gorge Anaple Gorge	the work for that	

Click on Microsoft Excel Input the pop up will open. In the files tab select the file from the location and add it.

file Edit New Actual Tools Mary	Spoon - Transformation II (changed)	- 0 1
		getting () Outs hespation
B Ver Deign Ottetand 12	landomation 1 II	
141: DIO.000	KOBO II III I	
15 Country In Pr.	Microsoft Excel input - 0	×.
D, 42# CN	Top Farls Mounth Eval Typet And Farlie	
W Ann Ann File Sheets Corner Brow Handing Ha	ah Additional autput falah	
O station Serve	Erheit von Empirel (bour Eh-2000 K.S. (PL)	*
Q. LDR Here	file or directory (C. Sweder Selector	
P. Landflier	Reple Lynnier	writty to the list of thes download
P. Monatt Belo	in Tegele Egension	
P. Monatt	Selected files a " File/Constany Wildcard (RegDig) Exclude existent Required Include subhidaes	
C? Mondran	1	and a second sec
C? OLATINA	Acad General fam previou data	
C. Property 1	Accept filmames from previous step []	100 C
(ST_ RIS legue)	Step to mod Versenan how	(c)
P, to christ	Rails of the papel to see to filmane [9
(A), Salasharas	Res Barredt	
P. SHP Head	Provide and a second seco	
E Sid hand	OK Province must Cannot	
C, Table Juge (C) Pergr		
Et fait the reput		
R Alless input		
ER, MA, Hyur Shawn (SUII)		
Tani Input		

Go to Sheets tab, click on Get Sheet name.

			500	ion - Transform	nation 1 ich	anged)	- 0
	tion may						Pergettine (1) Data Integration
Ver Deige		O minuner H Stanformation 1 11	_				
ince i	2. 21	DID.00 X0801		· ·			
The surgery of				Mote	of Dicel ing	M	- 0
E. 42# Ch		Day same stat fully	Mon	A Los Topal			
Contraction of	Nim Stants Co	mane Ever Handling Tates Additional subject for	-				
C (247 sp) C (247 sp) C (247 sp) C (247 sp) C (248 sp) Monath C (248 sp) C (248 sp) Monath C (248 sp) C (248 sp) Monath C (248 sp) C (248		List of shares to need		Past same	Bart nux	Bat solution	
C Sali Input			Sen	ox n	-	Canol	
P. Marine P. Marine R. MA, head the P. Tearlingsof	().00) 						

All the available sheet will show up. Select the correct sheet and transfer it into selection pane.

0		Spoon - Transformation 1 (changed)	
Bar data Yana Antona Yanta Data Di Cal Cal Cal	nay		Perspective 👯 Data Integration
B Ver & Design	O Treasure II Baselernation 1	N	
Steps	LI: DIDOODXDB	1 10 Wh v	
(F. Gauge & R.		Microsoft Excel input	- 0
D sover	See .	Monath load input	
W has been film	Sheets Contest Encotherating States Additional out	urfulde	
O, sold inge	List of sheets to	and a H Enter Int - O	
O, tild input		1 Australia Visur selection	
C, Lauffler		Tales	
P. Monart		Preduct Caurey	
P. Mennet		Dec.	
C. Mandrian		121	
CT CLAPTOP			
C. Papety (
D Champed		Propa angenta	
C. Scarse		Cent Cent	
P titland			
C Stiller		Get shadrama()	
EP, Table Iner			
D. Tentier Or	wite	OK Permanent Canod	
D. Maning	and the second se		
R M& hput loware	9449		
Tantingat	2000 C		
1	10		19.7

Go to Fields tab, click on get fields from headers rows

				hanged)	nation 1 (d	Transform	Spoon -							•
Outs Integration	Pergestive (22.0x											1404		3 8
								nation 1 22	(I hander	O Helane			Ø Design	Bver
						*	1076	080	. 00 ×	DII C	5.11			Steps
	- 0			put.	oft Excel in	Microse						P,	D. Gauge h	
						al input	Moreart (or	Bep rame Addition					D. 62# CN	
7	3						eta)	ional autput fo	ng Falls Add	t Erer Hands	eets Cantan	Aim St	Se anno anno	
		Grouping	Decimal	Curancy	Format	Asper	Trimition	Pacion	Length	Type	Name	17	O, LOAP app	
												1	O, she have	
													P, Lost fier	
													P. Mount	
													C Manham	
													C. OLATIN	
													P. Popatyl	
													(F), Rillingue	
													6 DOM	
													S. Selectores	
				Qu.,	ten Cyster	General							E Saland	
-				elab from head	Que fa			-			C117	-	D. Lawing	
				Cane	LOBOR HERE	1.54	0X				6	Om	D. factor	
													La upor all	

All the columns will show up with data type. Here you can modify the data type if required.
>						5	ipoon - Tran	Normation	1 (change	ed)				- 0
14 EF	I Vine Arten	. Tech	****										Provide Manual	interests
3 63	10 80 99		-										Andress of cars	and were
	S. Davida	_	-	C II C A C	O Y D C		100							
(inde		-	74 E	Paces	P AGE								- 0.000	
	F. Geogle In	2						COSOT EX	OR INDUX					
	E covor				20	grane Mo	work Destings	4						
	(), HLTHAN	Ann D	anto Canton	a Arrest Manadatan Mada	A Antional o	day facto					_			
	P, has her				1							An other		
	Q, LOW Her	1.	Cumone	ne line	reader	Pacesa		N	Page 1	Caranda	(acres	author a		
	C, COP Have	2	Name	Shing			none							
	C. Loog the c	3	Chy	Saing			-							
	D Mound			and here				-						
	C Manhier													
	CT OLAF INF													
	[] Papatyl													
	R. REPART													
	P. sicting													
	St. Selectors													
	C Shirted						100							
	P SAS Input	L					Ca	Carlas From F	ander on					
	Table ingr	Training of	10			100	64.	FIRING 1	-	Lance				
	C, texting r	(C) He	*I								-			
	C neeres													
	(P) Name in state													
-	and the state			S										
		-						_	_		_			_
	11 A	Z	16	W- 🖬		0	21 3	8					A 6 6 6 6	

You can click on preview row to view sample data to ascertain the data accuracy and the format accuracy.

0	R Ves Adapt	Testa	Page 1			5	poon - Tran	domatio	1 (chang	ed)				
0 15	988												Perquities (1) Data	'integration
8 Yes	@ Design	_]	O Welsond 1	Sundarantes	1.11								
Selat		-	41:	PILOOO	DURDE	101 1	100% 41							
	F. Googe to	2						CONNE	oe input					
	C, SEP CSV				54	grame Mc	work box typ							
	(); HL7Hput	files 2	hads Contant	Error Handling Hald	Additional of	dist fails								
	C data p	New Dama Content for Interfered Interfere Additional databal facts. # * Name Frage Length Paction Stimstype Report Famult Converty Builtoni Grouping 1 Converting Stimstype Anne No Stimstype Stimstype No 2 Stimstype Stimstype Anne No Stimstype Stimstype												
	P. SICIVIA													
	Statement													
	C Galland						Get	feld from 1	eader ton	3				
	D. febrier	-				-								
	P, fetfier	0	4				OK.	were a	-	Canod	1			
	P. Mara ing.													
	(C) 104 legel	lease (***	0										

Output of previews of 4 rows

0					Examine preview data	- 0 🔤
-	mp Moreart &	of legal (it must	e			
	Customeriel CC00001 CC00002 CC00008 CC00004	Name Customet Customet Customet	City New Defix Bangarlow Hydenabad Chevrasi	Martaflatus Mariad Mariad Mariad Single		
					Clean Showlang	Þ

Go to Output, expand it. Select table output, drag and drop it into the transformation windows. Connect Microsoft Exvel Input file to Table Output by keeping Shift Key Pressed and drag from Excel Inout to Table Output; the arrow will join shhowing the direction of the data flow.



Click on Table Output, the pop-up will open. Since this is the first time we are connecting to the database, we have to create a JDBC connection. For that click on new in connection option; another pop-up will open. Select MySQL as database, add appropriate connection name. Here we called it Pentaho. For same machine host name can be localhost if database is in same server else enter the server IP. Add database name, Port Number which is 3306 for MySQL, username and password of database.

0		Spoon - Transformation 1	(changed)		- * 80
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	14	Table output		Part	perfect () Data Integration
ter of Reine ter ter of Reine ter of Reine ter of Reine ter of Rei	Sign same in a second s	An August Constant fame Page Connection fame Page Connection fame Page Connection fame Page Connection fame Page Connection fame Page Connection fame Page Connection fame Page Connection Page Conn	Alex. Name Winds.		

Click on test to check the connection.

0		Spoon - Transformation 1 (changed)		
B tel 1 fan Anten berb indy B tel 10 G ID	0	Dutations Connection	121	Respective 💥 Data Integration
Constant of the second se	In the function of the functio	Conception Tage Penning Conception Tage Michanas Mic		
Statutette	Omp	Test Testure Laboration		
CT THE MAN	×	04	Cancel	

Once the connection is established, we have to select the table. For that click on Target Table browse button. Select the table **Customer**



Tick the Truncate table option for the first time. However for subsequent upload Tick it only if fresh data is required. Tick the **Specify database Fields** to map the fields from excel to Table

5		Spoon - Transfo	mation 1 (changed)		- 0
the fair New Action Tools Help					
9 2 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9	14	Table output		0	Pergedire () Data Integration
B Ver / Deign				Charles and Charle	
Sept 1	Suprama Page and	Table output			
· D Invit		record	* ddt. New.	Ward.	
· CI Origina	Terpi shena			Broot,	
III Adonatic Documentaria	Target table	Cutomer		Broom.	
D, Deles	Commit size	1000		•	
30 Invest/Update	Trynoste table	8			
Section and	Specify debuters fields	9 D			
V LOW Durput	Main systems Detabase Setab				
Microsoft Access Output					
Mount Loc Date	Patition data aver table	0			
Mount Los line	Pattoning fait			- 0	
Percente Reporting Cuty	Parking data per manth				
Properties Output	Lander bester of	5 G			
E REDup	Use batch update for insets	8			
SI Ne Dugut	to the name of the table defined in a field	0			
F. Selectorce Delate	Full that contains starts of fables				
Selectores inset	Store the fairlenance field	M.			
(A), Saleshoca Update	feters and appropriate in	0			
(a) Salashara Sprad	faces of acts generated ing fails			•	
13 SQL File Output	() may	OK Cancel 923.			
C. Louissia da mara					
20 fails extent					
No. a. a.					

Click on Get fields to get the Fields from Incoming Input. Click on Enter field mapping to map fields one by one. If the fields name is same the system usually maps it automatically. For different field names you have to select the fields and map it.

0			5	poon - Transformation 1 (ch	anged)		
Fit fait You Active Task Weig			10.00				Property of Data Managing
Di tita (d' fanine)	58		Table	NUM .			and a second sec
Varue (2) Design Terr Terr Terr Conput Co	Main aptions (Deta Ratio to invest • Table 1 1 Colorer 2 Name	Step name Connaction Target scheme Target scheme Connecte sche Connecte sche Specify detaesen finds Base fields eff Desem for werdt Desem for Some	hans ouput Fundea Conturner 1000 12 12 12 12 14 14 14 14 14 14 14 14 14 14 14 14 14	*)	let. Non. Hard. • Brant. • Brant. • Statut See Sector		
	C Huge	Balan Maritalo	en 0x Ca	10 Seven fields Auto target solution? (2) Miles support solution? (2)	End Target Fable. Anto servera adhetican ² () Hida seripmet target Fable? ()	er Mapping Aug Boose C	- 0 C Maging Copy-Co Copy-Co Mandollario manifoldura Manet - Culture D D Daniel

Not our transformation is ready. Save the transformation in the folder and name it appropriately to distinguish it later.



Click on Forward Button to run the transformation. The Execution result windows will show the rows being read and rows written, time taken etc. In case of any error you can look into logging tab to find the cause of the error in the execution.



The Customer table is populated with data from the excel file. You can check it by writing query in the data base Table Customer.

ect * from Co	stones hout 10		and the second se	
an Ireal end		AN DEPT RATE	the executed a life F	verse (Press to 1
1000) (L.				and from a l
Contoner16	Custementane	CSTY	mailfalStatus	
00000000	Contonesi	New Delibit	Rectifi	
0000000	Customer 2	Bangaloce	Mactied	
5,000003	Customer 3	Hyderabad.	Married	
4 0000004	Customer4	Chesnes	Tungue	
0000005	Customers	Mushel	Married	
0000006	Custinetal	New Deling.	Single .	
5000007	Customer?	Bangaloge	Maculed	
0000008	Customers	Ryder 40-46	Single	
Eccoece#	Customers?	Chettata	Bingle	
00000000	Custement10	Madel	Single	
Concernance of the				N
				ন্দ

Similarly follow the above steps to upload data into Calendar table.

are - mon	Careboard stream to							
23 233	2/14/16 10:48:36	AR 207) Deuty	s executed -	(P) Exces	(7144	(292ex)		
Bata	Teat FT Teat	Roath Name	Teachanth	Quarter	Der	Weekday [10.00	
314/1/2015	2015 2015 14	04.404	2015 04	44		1 Wednesday	Б	
24/2/2015	2015 2015_14	04,401	2015_04	01		2 Thursday		
8 4/3/2015	2013 2015 14	04, 401	2013.04	01		1 Palifier		
4 4/4/2015	2015 2015 14	04 Apr	2015.04	01		4 Saturday		
\$14/5/2015	2015 2015 14	04,401	2015_04	25		5 Dunday		
414/6/2015	2015 2015_14	04,401	2015_04	00		(Number		
1 4/1/2015	2013 2015 14	04,404	2015,04	45		77488642		
# 4/9/2015	2015 2015 14	04,301	1015_04	01		t Vednesday		
8 4/9/2015	2015 2015_14	04,405	2015_04	44		5 Thursday		
and the second second	2015 2015_14	04,401	2015_04	- 25	3	5 Felder		
# 4/9/2015 # 4/9/2015	2015 2015 14 2015 2015 14 2015 2015 14	04,401 04,401 04,401	2015_04 2015_04 2011_04	01 01 01	,	t Veduceday 5 Thursday 5 Filday		

Also upload customer data and sales data into the respective tables using same process. Note that name the transformation in such a way that it is easily recognizable.

Sample customer data

et.	* from Product 15441 51							
_					100 C 10			
1	[289] [2/14/14 10:50:29.3	A 197) Seco	ipt executed - (8)	frees [Time	27(ar.)			
5	reductid Frederillane	Category	2.0 Category	Pater	Non-Personal Procession			
r	100000 Furdant mast 1	Solutes.	Date tphotes	20000				
	100001 Frederit name 1	Mobiles.	Testured Bones	2400				
	100002 Product name 3	Rebubes	Taxa tobaces	28500				
	100003 Fundart heat 4	Mabubes	Tax toboase	15000				
	100004 Jundant near 1	Mabiles	Featured Shones	1800				
1	100005 Furdact near d	Robiles	Peabured Phones	3400				
-	10000d Forduct near 7	Nabiles	Featured Phines	2000				
	100007 Foodurt near 8	Laptop	Windows	23000				
	100008 Fundant meat 5	Laption	Vindows	35000				
1	100000 Freduct near 10	Laptop	Windows	43000				
1								
						- 21		
						L2		

Sample Sales data

9 [193]	(1/14/16 10:5	5:05 AM 1977)	Sccipt execut	ed - (0) E	(Tis	ei 840as 3		
Oudeu14	Customente	Orderbate	Fundareld	Quantity	Annual	Discount	Deature	
100000041	0000064	4/1/2018	100013	and the second se	1800	4	Januaned	
10000002	CC00043	4/1/2015	000022	· · · · ·	1 1040	\$25	Completed	
10000003	0000098	4/3/2018	100008		2 70000	3500	Cancelled	
(\$0000004)	0000068	4/4/3015	100013		2 2400	100	Completed.	
10000005	0000064	4/1/2015	100013		3400	4	Completed.	
10000004	0000004	4/6/2015	000013		1 100000	\$2000	Completed	
10000007	CC000641	4/7/2018	100000	1	\$0000		Completed.	
10000008	-0000087	4/1/2015	100025	8	5000	340	Completed	
00000009	0000047	4/8/2015	200003		30000	3400	Patiened	
000000330	0000014	4/10/2015	100000		1 10000		Completed.	

Now that we have uploaded the required data into the table for the excel file let us explore other elements of the PDI. Let's us practice one Table to Table data transfer. The database to database is actually a Table to table transfer only difference is there will be two database connections. For this example create a Table name OLAP with additional columns such as City and Margin. Create the table using below query

Create Table OLAP

(
Orderid y	varchar(20) primary key,
CustomerId	varchar(20),
OrderDate d	ate,
Productid	INT,
Quantity	INT,
Amount	Numeric,
Discount	Numeric,
Status	varchar(30),
City var	char(30),
Margin	Numeric

)

Select a table Input and Table output and connect it in the transformation windows.



Connect the database using Pentaho connection we created in previous exercise. Add the Query into the SQL box. Here you can see I have added city from customer Table and margin is calculated as 10% of Sales Amount.

			Table input				
	Step name Table ing	d.					
	Connection Puncha					\$81. Nov.	Ward.
10.						Get SQL when	network.
elleri S. Order el, E. S. Order el, E.	over13						
Cover 10 Column 30	auth bar converting D						
Taylo	ce variables in script						
	Inset data from dap						*
	Internet of the sector of the						
	Sandrabar (P						•
(C) Help		OK .	Penies Ca	ercel.			

Data previews from the query

						0.0	nine previ	en data			. 0
	tep Table inpu	did-mark									
	C-devid 00000001 00000000 00000000 00000004 000000054	Culture CCIIIIII CCIIIIII CCIIIIIII CCIIIIIII CCIIIIII	OnderCate 2015/04/01 00:00:00.000 2015/04/02 00:00:00.000 2015/04:02 00:00:00 2015/04:05 00:00:00 2015/04:05 00:00:00 2015/04:05 00:00:00	Polucial 108015 10802 10808 10803 10803	Quartity 2 2 2 3 3 3	And the second s	0 125 3900 180 0	Status Returned Completed Cancelled Completed Completed	City Nano Delhi Hyderaitaid Hyderaitaid Myderaitaid Nano Delhi	Megin 100 103 103 300 300	
_	-				_				-		

In table output select Pentaho connection, select OLAP table and map the fields.

0			Spoon - Tran	noformation 1 (changed)				- 0 - 0
die Litt Van Adam Tauls Help	(
9 5 11 2 2 2	14		Table output			0		Perspective 32 Data integration
Brain P Design 1 Brain 1 1 Distribution 1 1	for Their sphere (beater	Ship have Connection Target solars Target solar Connection Target solar Connection	Technologia Technologia Ultit SE	*)[60.] (• •	Nard. Base.		
Sanatura Dates Selectura Inaet Selectura Inaet Sanatura Updas Sanatura Updas Sanatura Ina Sanatura Ina Sanatura Ina Sanatura Ina Sanatura Ina	Table for care	Bream Stree		Outsbase Explorer	1		Of. inter-	
C Systemize the range Take subpri S Take subpri S Take Subpri S State Subpri S States S St	(Ome)		Custome Burg Produkt Sales Sales Spheryme					

Save the transformation and execute it. The result from query windows.

28 [285]	12/14/16 11/1	7142 AR 1973	Script execut	44 - (D) Ex	THE TIME	ei 207an]				
Dedeci4	Customerla	OcdesDate.	Productid	Quantity	James	Discount	Deature	City	Rergin	
1.05000001	0000046	4/1/2015	100015	1	1800	01	Petuched.	New Belits	100	
2 00000002	CC00043	4/2/2015	100022		1045	125	Completed	Sydes about	204	
5,0000005	0000088	4/3/2015	100008		70000	3500	Cabdelled	Syderabad	1000	
4-10-000004	0000048	4/4/2015	100013	1	3400	180	Completed	Ryderabad	369	
5-coopers	0000086	4/5/2015	100013		5400		Completed	New Dollar.	360	
4-100000004	CC00004	4/8/2015	100011	1	100000	12000	Completed	New Dellas.	10000	
5-120000007	0000043	4/7/2013	100002	-	10000		Completed	New Deliks	\$000	
8-10000004	0000087	4/8/2015	100025	1	3000	340-	Completed	Sancal-ce	300	
9-00000009	0000047	4/9/2015	100005	3	30000	54003	faturned.	Zangaloge	3000	
10-10000010	0000094	4/18/2018	100002		58000		Consisted	Ontonal	\$000	
8j100000005	CC00087 CC00084	4/10/2015	100003	-	30000	9400	Completed	Sangaloge Chemital	3000	

Let's practice one more elements which will be useful in every situation. The database often need update, we will learn how to update the tables. In the entire process of ETL the transformation included database updation. In this case let us update City Bangalore to Bengaluru in OLAP table.

Open a new transformation, drag and drop Execute SQL script from Scripting Tab. Select the connection, enter the Update query and execute it

0			Spoon - City_Update			- 0
And that they define their re-					Pergettie	() Outs hetergradiant
B Ver & Beign		HOLF HOULENN				
Stepi	1.11	DICEOP KOSS S	905 V			
r C3 input	100		Execute SQL statements		0	1
ED Over		Ship na	ene Geoute VOL script			
> C Instein		Connect	for Persite	V 581. Nov.	Ward.	
· Co shiey	34	X script to execute, (statements separated by () Quest	ten maks will be replaced by arguments.	and a second discount		
CD Fee	1	place OLAP			- *	
 C) Sugging 		tere City+ Biopriors'				
The best we so so and						
It famile					and of	
C Madded Jan Scipilia C Mage Subatton C Rule Scientific C Rule Scientific		El Column 19 Escuto for coch cont () Escuto ex a single determine () Veriale substantion () End presenter ()				
E User Defined In-a Class	2 h	Dave Stop/12				CX.
1 Une Defined two Depart	1	* Feld name to be used as argument	-			
- Clister			Field to contain inset state			
C ann			Katol to existain Update stato			ingut/indput
1 🗂 Onto Warehouse			Field to contain Oxfere stats		- ×	3
F C3 Valence	10	and end	religita contain faced data			
E Distation	12	D Hele	Of Canad Cat Sets			
i 🗋 Sigilita	11		OK I			
1. 1. 440						

Execute the update transformation

0	Spoon - City_Update - 0
the Balt View Advan Tools Help B CS (1) (2) (2) (2)	Pargentine 🔡 Data Integration
B Ver / Dely	HOW HOUSEN
Compart C	Encort23 torp
E Varbefreit in a Spreasan	Execution Results (C x) Company (C Stay Marin) (C Antoniana Graph (C Marin) (B Frainc day)
C Loning C Loning C Loning C Lonin D Data Illustrians D Data Illustrians	Copyrer Read Written Input Output Spatial Report Error Active Time Speed()() input/vdput Demotr3Qtaript 0 0 1 0 0 0 0 0 facthed Site Qt +

As shown in the result the Bangalore has been changed to Bengaluru.

	Characterized in the particular in				-			
1003	[1/14/14 11/1	1-00 AM 197)	Scalpt execut	el - ili Er	THE [TIM	*: 235e#]		
Ouder14	Customer14	Cudexbate.	Freduct14	Quantity	Anisat	Discount Status City	Macgin	
12000004	0000044	4/1/2018	100015		1400	dirturned New Delhi	185	
20000000	0000043	4/2/2014	100022	3	1040	125 Completed Mydeumbed	304	
10000003	0000048	4/3/2018	100004	1	70000	1500 Cancelled Spirashed	7000	
00000004	0000068	4/4/2015	100013	3	5400	180 Completed Mydezabas	240	
10000005	0000044	4/5/2015	100013	3	3/00	d Completed New Delbi	240	
10000004	0000004	4/4/2014	100011		100000	12000 Completed New Belhi	10000	
10406687	0000041	4/1/2015	100002	1	50000	d Completed New Belhi	5000	
10000008	0000087	4/8/2015	100009	. 3	3000	300 Completed Mongalumu	200	
00000000	CC00047	4/9/2015	100003	3	30000	bitt Returned Bengaluru	3000	
COLORADO B.	00000000	A-110-12018	1.55500		STARS.	Street start Channel		

In the real business situation there will be multiple data sources. The ETL tool PDI can handle all business requirements. It has almost all the connection possible and data processing that one can think of. You just need to have right credential, rest is drag and drop and sequencing it right.

We have seen some example of data input using PDI, lets us explore an output from the data base and its distribution. The PDI can well handle reporting requirement at very complex level. In this example I will explain how to create transformation and job to accomplish this.

Assuming we have to generate two report

- 1. Categorywise Sales quantity, Amount and Discount of last 7 days on daily basis
- 2. Categorywise City level Sales, Amount, Discount, ASP and Discount percentage for last 7 days on daily basis.

Both report are to be mailed to users daily 9 am in a single excel file. The query for both report are as below.

```
select Category, sum(Quantity) Quantity, sum(Amount) Amount,
sum(Discount) Discount,
sum(Amount) / sum(Quantity) Average_Selling_Price, (sum(Discount) /
sum(Amount))*100 Discount_Percent
from Sales S
inner join Product P on S.Productid=P.Productid
where S.OrderDate >= date(date_sub(curdate(), interval 7 day))
group by Category
```

select Category, City, sum(Quantity) Quantity, sum(Amount) Amount, sum(Discount) Discount,

```
sum(Amount) / sum(Quantity) Average_Selling_Price, (sum(Discount) /
sum(Amount))*100 Discount_Percent
```

from Sales S

```
inner join Product P on S.Productid=P.Productid
```

```
inner join Customer C on C.CustomerId=S.CustomerId
```

```
where S.OrderDate >= date(date_sub(curdate(), interval 7 day))
```

group by Category, City



Select a table Input and Microsoft Excel Output. Add Query into the Table Input and check the output using preview rows option.

Click on Microsoft Excel Output, in the popup Enter Filename, File Extension and add Sheet name. In case if you want to generate a new output file every time then select "*replace with new output file*" option in the If

output File exists options. Similarly for sheet select "*replace with new sheet*" if you want to generate new sheet every time.

8	ip-10-0-87.a	ap-southeast-1.compute.internal:1 (ec2-user) - VNC Viewer		- 0
	Step name	Microsoft Excel Writer		
file & Sheet Conte	nk)			
File				
	Filename	/data/jumin/SalesReport	•	Browse
	Extension	xisx [Excel 2007 and above]	~	
	Stream XSLX data	0	- Contract	
	Split every data rows	0		
	Include stepnr in filename?	0		
	Include date in filename?	0		
	Include time in filename?	0		
	Specify Date time format	0		
	Date time format			
		Show filename(s)		
	If output file exists	replace with new output file		~
VUsi	t for first row before creating file Add filenames to result	8		
sheet				
5	heet name (max. 31 characters)	Category		-
	Make this the active sheet	8		
	If sheet exists in output file	replace with new sheet		~
1 1	Protect sheet? (XLS format only)	0		(6)
	Protected by user	F		

In Content tab select the starting Cell where Report will start. Tick Header option if there are header rows in the report. Scroll down to the end, click on add columns. The entire column will be shown with their data type. You can remove come of the column if not required. Change the data format as per your reports requirement.

			Step name	Microsoft Excel Writer		
e & 1	hee	t Content	5 A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.			
Conte	nt e	ptions				
		5	tart writing at cel	[A]		10
			When writing rows	everwrite existing cells		
			White Header	0		12
			Write Footer			
			Auto size column			
		Force for	nula recalculation			
		Leave styles of existin	g cells unchanged	10		
	int	ting to existing sheet				
		and the shared of chard	(second on Fear			
	XA	t writing at eno of sheet	(appending lines			
			Crister by rows	*[0		
		Begin by writ	ing empty lines	\$[0		
			Omit heade	0		
ield						
*		Name	Type	Format	Style from cell	Field title
	2	Category	String			Category
	2	Quantity	Bighumber	*,**0		Quartity
	3	Amount	BigNumber	0,000		Amount
	4	Discount	BigNumber	*,**0.00		Discount
	5	Average_Selling_Price	BigNumber	#,##0.00		Average_Selling_Pric

Similarly for category city report create a transformation, add the second query in the Table Input. In the excel output file tab enter the sale file name, select "*Use existing file for writing*". This is because we create a new file in

first transformation with sheet name category, for the Catcity report we just have to add a new sheet CatCity.

8	ip-10-0-0-87.ap- step name-p	-southeast-1.compute.internal:1 (ec2-user) - VNC Viewer ficrosoft Excer Writer		- 0
File & Sheet Content				
File				
	Filename R	Sata/umin/SalesReport	0	Browse
	Extension x	isx (Excel 2007 and above)	~	
	Stream XSLX data)		
	Split every data rows			
	Include stepre in filename?)		_
	Include date in filename?			
	Include time in filename?	1		
	Specify Date time format			
	Date time format			2
	(5	how filename(s)		
	if output file exists U	se existing file for writing		~
Wait fo	r first row before creating file Add filenames to result @			
Sheet				
She	et name (max. 31 characters)	atCity		0
	Make this the active sheet g)		
	If sheet exists in output file	eplace with new sheet		~
Pro	tect sheet? (XLS format only)			
	Protected by user			

Add your fields and do formatting.

-			ip-10-0-0-81	ap-southeast-1.compute.internal1 (ec2-	user) - VNC Viewer	- 08
1				Excel Writer Step v1.2 by type-ex	it.org	
			Step name	Microsoft Excel Writer		
le 6 1	Sheel	Content				
Contr	ent o	ptions				
		5	tart writing at ce	a Al		
		,	When writing row	s overwrite existing cells		
			Write Heade	10		
			Write Foote			
		4	Auto size column	6 D		
		Force for	nula recalculatio	0		
		Leave styles of existin	g cells unchange	40		
when	writ	ting to existing sheet				
	Start	t writing at end of sheet	(appending line	00		
			Offset by row	\$ 0		
		Begin by writ	ing empty line	50		
			Omithead	40 [v		
				.0		
Field	5					
^		Name	Type	Format	Style from cell	Field title
	3	Category	String			Category
	2	City	String	a second as		City
	3	Quantity	BigNumber	0.000		Quantity
	-4	Amount	BigNumber	8,880		Amount
	5	Discount	BigNumber	#,##0.00		Discount
100	ĸ	Buarana Kallina Brira	Rinklumber	1 # ##0.00	1	Buarana Katlina III

Now that we have created two transformations, we have to join both transformations together. This is done through job. Go to File -> New -> job

In the job window add elements. All job starts with job start button, so add it as first elements in the sequence of transformation. Add two transformations in sequence.



Click on Transformation, the popup opens. Enter a friendly name in job entry, browse and select the transformation from the location in the system file. Provide a descriptive name to each of the transformation.

	ip-10-0-	0-87.ap-southeast-1.compute.internal:1 (ec2-user) - VNC Viewer	- 0
		Spoon - Job 1 (changed)	- 0
le Edit View Action To	ols Help		(management)
0 1 25 13 14 14 1	Warne a s	Pe	spective: 🐯 Data Integration
View O Design	[] Job 1 I		
	Name of job	Job entry details for this transformation:	
		Capabor	
ransformation specification	Advanced Logg	ng settings Argument Parameters	
 Transformation filename 	5	/data/jumin/SalesCategory.ktr	• [3
O Specify by name and do	vectory	[•
		[
C. Coarily by relations			36
C. short of reserve		l.	14
Энер		OK Cancel	
	ip-10-0-	0-87.ap-southeast-1.compute.internal:1 (ec2-user) - VNC Viewer	- 0
		Spoon - job 1 (changed)	
le Edit View Action To	ols Help		- 0
e Edit View Action To	ols Help	Pe	spective: III Data Integration
le Edit View Action To	ols Help	Pe	spective: 😂 Data Integration
le Edit View Action To	ols Help	Per Job entry details for this transformation:	spective: 🕃 Data Integration
le Edit View Action To	ols Help	Per Job entry details for this transformation: entry: [catCity	spective: (2) Data Integration
ile Edit View Action To B B B B B View Obsign tansformation specification	ols Help	Per Job entry details for this transformation: entry: [CanCity ing settings] Argument] Parameters]	spective: 🕃 Data Integration
ile Edit View Action To Edit View Design View Design transformation specification * Transformation filename	Name of job	Per Job entry details for this transformation: entry: [CatCity ing settings] Argument] Parameters] [/dstafjumin/SolesCatCity.ktr	spective: E Data Integration
Ile Edit View Action To Edit View Design View Design transformation specification * Transformation filename	ols Help 22 Job 1 1 Name of job Advanced Logg	Per Job entry details for this transformation: entry: [catCity ing settings] Argument) Parameters] [/data/jumin/SalesCatCity.ktr	spective: (2) Data integration
Edit View Action To Edit View Action To View / Design Transformation specification Transformation filename O specify by name and de	Name of job Advanced Logg tr weboy	Per Job entry details for this transformation: entry: [cacOty ing settings] Argument] Parameters] [rdstafjumin/SolesCatOtyXtr	spective: (E Data Integration) (2) (2) (2)
ile Edit View Action To Edit View Control To View Control To View Control To Transformation specification Transformation filename Control Specify by name and do	Advanced Logg	Per Job entry details for this transformation: entry: [catCity ing settings Argument Parameters [/dstafumin/SolesCatCity.ktr	spective: (2 Data integration

The job is ready not but one element is missing – a system to mail the report to the users. For the mailing system create a transformation, add a Table input

and Mail. In the Table Input you connect to Pentaho or any other live connection and add a query as shown. In the query you can add To Emails, From emails, Subject of the mail, mail body, SMTP, Port, mail userid and password.

V? ip	-10-0-0-87.4	p-southeast-1.compute.internal:1 (ec	2-user) - VNC Viewer		- 0	
		Table input				^
	Step name	Table input				
	Connection	Prod		Edt	New	Wizar
sou				Get SQL	select sta	temer
<pre>'Tomprofiles Tender' Tenden' 'Tomprofiles Tender' 'Tomprofiles Tender' 'Tomprofiles Tender' 'Sales Report' Tender' 'Sales Report' Tender' 'Tamper Dorr Mor Tend 'Domprof Dorr Mor Tend 'Domprofiles Tender' 'Domprofiles Tender' 'Domprofiles Tender' 'Sales Tenders' 'Demoken Dorr Mor Tender' 'Demoken Dorr Mor Tender' 'Sales' Francesco</pre>	يەمل.					

Map table input to the mail elements. Add from the dropdown the correct fields defined in the query.

0	Spoon - MaliSale	sReport	- 0
File Edit View Action Tools	Help		
		Perspect	ve: 🐹 Data Integration
View / Design	8 -	Send mail	×
Steps 🗿 🔥 🖬	E Step name	201	
input	Addresses Server EMail Message	Attached Files Embedded images	
C Output	Destination		
Transform	Destination address	The Ferral	[w]
D Utility		w_cmax	
E Flow		CC_Email	×
Scripting	BCC		¥
BA Server	E-mail center		
C Lookup	Contraction of the second		1000
D Joins	Sender name	from_Email	×
🗀 Data Warehouse	Sender address	From_Email	¥
Validation	8 · · · · · · · · · · · · · · · · · · ·		
Statistics	Reply to	Reply_Email	~
🗀 Big Data	Contact		~
Agile	Contact phone	1	
Cryptography			
C] Palo			
Open ERP			
lop			
mapping			
Brik loading	HULL I.		, (f)

Add SMTP, Port, userid and password. Select secure authentication as your mailing server configuration.

C Elle Edit View Artiss Tasle	Spoon - MailSalesReport _ 0
	Perspective: 🕃 Data integration
View / Design	🗲 🔹 Send mail 🛛 🗙
Steps 🗿 👍 🖬	D Step name 201
input	Addresses Server EMail Message Attached Files Embedded images
Output	SMTP Server
Transform	SMTP Server Gurra
Utility	And Star
Flow	Port Port
Scripting	Authentication
BA Server	Use authentication co
Lookup	Authentication user and an and and
b 🗀 joins	ose_io V
Data Warehouse	Authentication password Password
Validation	Use secure authentication go
Statistics	Secure connection type TLS
Dig Data	
Aglie	
Cryptography	
Palo	
Open ERP	
b Ca Job	
Mapping	
Bulk loading	

Add subject and mail body if any in comment.

n	ip-10-0-0-87.ap-southeast-1.compute	e.internal:1 (ec2-user) - VNC Viewer	- 0
File Edit View Action Tools H	Spoon - MailSale	esReport	- 0
0 8 8 8 8 9		Perspective	e: 🗮 Data Integratio
View @ Design	e -	Send mail	
Steps 🙀 👍 🚛	D Step name	Mail	
> 🗀 input	Addresses Server EMail Message	Attached Files Embedded images	
> Contract	Message Settings		
Transform	Include date in message	20	
Usity	Only send comment in mail body	20	
Flow	Use HTML format in mail body	2.00	
Scripting	Encoding	UTF-8	
> 🗀 8A Server	Manage of oth		
E Lookup	Planage priority		
> 🗀 joins		Normal	
Data Warehouse	Importance	e Normal	v .
Validation	Sensitivity	Vivornal	(v)
Statistics		Sector Contraction of	
🕨 🛅 Big Data	Message		
> 🗅 Agile	Subject	t Subject	~
Cryptography	Commen	e	
P C1 P60		1	
D Open ERP			
dot 🗂 4			
Mapping			
Bulk loading			(m)

Add attachment of the mail from the output file from previous two transformations.

-	Sneen - MallSalesBanert
File Edit View Action Tools H	spoon - mansaresneport
	Perspective: 00 Data Integration
View @ Design	🖸 🔹 Send mail 🛛 🗙
Steps 🚯 👍 🚛 💷	Step name Mail
> 🗀 input	Addresses Server EMail Message Attached Files Embedded images
Output	Attached content file
Transform	Attach content file co
Usility	Content fieldsame
C Flow	
C Scripting	Fienanie fieldnanie
BA Server	Attached Resource
Lookup	Acadomy intrames
D Joins	Dynamic filenames?
Data Warehouse	Filename held Attachment
Validation	Wildcard field Cur. Date
Statistics	Filename foldername (users 6 wird folge forend wird) (6 (forend fore)
🗀 Big Data	Addaption participation and a state of the second state of the second state of the state of the second sta
Aglie	include subraiders
Cryptography	Wildcard
C_ Palo	Zip files
D Open ERP	Zofies n
dol 🖸	is zie Nename dynamic? ro
Mapping	Ziphiehame field
Bulk loading	

Save the Job and execute it.

	4	the second se		1974 - C	
View Ø Design	SalesReport #				
Entrie 🌆 🔥 🚛 🚛	00886	100%			
Coneral START OUMMY Dummy Example plugin O) Job Set variables v Success	START C	allepon CarCey	•••	E Carl Report	
II Tansformation	a				
Mail	Execution Results				6
File management	History 🚺 Logging 🚑	Job metrics R Metrics			_
Conditions	job / job Entry	Comment	Result	Reason	Filena
Scripting		000002010-0000000			
Bulk loading	Job: SalesReport	Start of job execution		start	
Big Data	START	Start of job execution		start	
> Ch XML	START	job execution finished	Success	followed up could be add link	adates!
D Utility	Catheront	ish execution finished	Success	Politikeo arconoromanink	idatal
Becository	CALCIN	Start of job execution		Followed link after success	Matal
b C File transfer	CatCity	Job execution finished	Success		/data/
h Co file execution	MailReport	Start of job execution		Followed link after success	fle III
h Co Palo	MailReport	Job execution finished	Success		fie.m
Deprecated	job: SalesReport	Job execution finished	Success	finished	
				Current workspace: "We	orkspace

The mail received in the mailbox

•		0	ii.	101	· 4	More *			
Saless F	Report .	Index x	5						5 19
± 10 m							@ 12.05 PM (3 minutes age) 😳	*	
- Roll-	18887								
×	SalesReport	Loha 🖉							
							N		

The file output is as shown below

state manual inter	a hor	and 1	and the	teres 1	ine bee	500	paport anis	- Miciqu	OT, DADRI								- 0	
A CA	0.000 E		- A-	-		e face pe la Carritar +	Genera 5			Na la	10 10 10 10 10 10 10 10 10 10 10 10 10 1	shi.	Deve fure	E auto 2 Com	- 2	7 A		
	0	<i>f</i> .	-				-			-	-	-			-			
A 8	¢	0	6 - C			6		1	1	- K.		м		0		g		1
Category Quantity a	30.30E	1 Mail OF	enerate Second	E MOR DISEBUI	6.64													
Books 20	4.100	112.00	-	346.00	1.42													
LADTED 28	1,628,000	53,245.00		4,495.45	5.58													
Mobiles 36	151,200	4,572.00		8,456.00	3.42													
Tey1 25	81,200	354.00		1.348.00	1.78													
(1990) (1990)								_										
								_										
				0														
* * Category / C	Selley 1 P.	1								84								
1 (r - Y)e	in National	uning 1	umuta (14)	feres 1	144 (ber	Sales nam há	Report	- Micros	nt Excel							1.00% (-		
A Cat	r Papel	uquint - 1 -) 10	ender bei	feire 1	Den Car	Sales name tra name	Report e-m General	- Microw	nt Decer	N I		3-	ə (8)	E sone	··· 2	r a	.0	
A Cat Cat Case - Y 2m A Cat Cat Case -	t hat	and 1 +)s (+) =+	ана (ал. 25 б. с. 2 - <u>д</u> - 1		144 (144 127974) 18 121944	Sales reper to rbue a & Center *	Report sou General S - 1	- Mees	n bor	Na Internet Pro-	and Cat	-1 -k	in an	E Auto	· · · · ·	T A	•0	
A CA Catology - Catology - Catology - Catology - Catology -	n nye Galan 1 Z B	ania + 11 + 11 + ⊒ + Anit		-	na (na ∑-ma ≇ ∰-ma	Sales naper na nac Center -	Report sou Source Source	- Micros	n bor	Maria Internet	an a	ş.	Degra Farma	E Auto Ta ta Q Dee	- 12 tetter	7 (A)	• 0	
NLL · C· Y Se	t Papel Californi B 2 12 (*	n ben n + + □ + + □ + het A	· · · · · ·		ina (na Sirina F Mang	Sales nam na Nat a & Center - V	Report time Servera 1 - 1	- Micros 	n bor	Na la	an Spe-	1-1-	in the second	E Auto 2 Dee	ten - 2	T A	• 0	
A B	n Papel Galleri B Z B (*	inter + 11 + 21 +				Salar ngan ha Nat a & Cantar - 2 6	Report one Served S - 1 N	- Micros	n Doel	Na la Starpe Jo	and Cat Min - Rate-	-1 -1	Dente France	E Auto 2 feat 2 feat 2 feat 6	ten - 2 . 2 ten-y	A forst	- 0	
Mile Control C	a Rojet Galast B Z S Guantity	Amburt	- A A	farme i Friday	Para Dana Distance Para Distance Para Distance	Saler ngar Ad Nat a & Contar - Changant	Report Served Served	- Meres	A bor		Carl Spectrum	-1- ×	Dente Francis On	E suto 2 Corr 0	ten - A	7 (A)	- 0	
A Ge Crasting Control	r Popel Gallest B Z U C Guarding	14000 1 + 101 + 101 - 101	1 1 1 1 1 1 1 1 1 1 1 1 1 1	Terrer 1	ina (no Dimo Milliona	Salen najer ka Not e & Center 1	Report son source source	- Meese	A bor	National Providence of the State	Con Space		Company Property Costs	E suto 2 Our 0	ton - 2	T A	- 0	
A De Cr. Y Su Toma Nut A De Group - Group - Mil - A B diagory Cy pparts Bargiore pparts Connor	r Papel Galles B Z U Counting	1 - 11 - 11	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Factors 1	The December The December 147 149	Sales reper to fore e & Center 1 4 Chemant & S2 E2300	Report sou s y	- Meter	n bor I S St	Na I Marana Par Marana Na Par	and Col		Decis Face on	E sono 2 One 0	ton - N	C A		
A de A	a Roat B Z U Colori C Coorting	14/00 1 - 111 - 1111 - 1111 - 111 - 111 - 111 - 111 - 111 - 111 - 11	1 1 1 1 1 1 1 1 1 1 1 1 1 1	Factors 1	ina (na Prina	Salver Tool 1 tool 4 A Center 1 4 G G G G G G G G G G G G G G G G G G	Report s-ni Served	- Microso 	n bor	Na I Marine - Ar Are	and Cat	N N	in the second se	E sono 2 One 0	tota - 2	C A	.0	
A Cot A	a Roja Galari B Z U C Guantity	Ampund F - 10 Ampund Ampund Ampund 2,000 4,200 4,200 4,4200 4,4200	1 1000-001 1000 1 1000-001 1 1000-000 1 1000-000	Factors 1 Factors (F) F F F F F F F F F F F F F	ina (na Tran Hurry Hurry Karal K	Salver Inger Ad Tool & A.Conter - Generati & S2 S2.10 & 6.39 & 7.30 & 7.300 & 7.30 & 7.30 & 7.300 & 7.3000 & 7.30000 & 7.30000 & 7.30000 & 7.30000 & 7.30000 & 7.30000 & 7.30000 & 7.30000 & 7.300000 & 7.30000000 & 7.3000000000000000000000000000000000000	Report s-ni Servera #	- Microso 5 + - 54 	rt bor	Ra I Store - Av	met Cer Mite - Date-	N N	Decision for the second	E son 2 Our 0	teriory	7 A	- 0	
A B A B A B A B A B A B A B A B	a Papel Gallen B Z U Guarding	Annual An	****** (*** * A' A' * Decourt Aver \$28.00 106.00 473.00 288.00 106.00 403.00 288.00 106.00	Factors 1 	144 (144 27 1744) 28 1744 29 1744 20 1744 2	Sales nue 10 104 6 Center 1 1230 539 639 639 639 639 639	Report som s	- Microso 6. + - 54 	A bor	Ra literation for the second	total Cat	1. ×	in a second seco	E sone 2 Over 0	terting	4 (MA)	. 0	
O = Y 2x Tona Num G Con G Con G Con G Con G Con G Con	e Ruje Gales B Z U Guersin	Amount A - 111 - 111		Former 1	1440 (1444 1447 1447 1447 1447 1447 1447 1447 1447 1440 1400	Sales nue a Aus a & Center - 4 Chement 8.52 82.59 8.20 8.20 8.20 9.20 9.20 9.20 9.20 9.20 9.20 9.20 9	Report som s	• Meren	A bor	National Providence of the second sec	Ca San - Spec-	- 1 - N	Decent Factor	E sum 2 fra - 2 for 0	terres	C (A)	. 0	
A B Barrier Control of	t Rup I Gales B Z U Guercity	Implied Implied * 100 * 100 * 100 * 100 * 100 * 100 * 1000 * 1000 * 1000	1 0 · A · A · 1 0 · A · A · 1 1 1 1 1 1 1 1 1 1 1 1 1 1	7 7 7 7 7 7 7 7 7 7 7 7 7 7	1440 (044	Sales nue 14 nue 14 nue 14 center 1 c. Parsant 12.00 4.04 4.04 1.74 0.00 7.00 0.00	Report e-mi General 9 - 1 9	- Microso bi Sa biotector	oft Decel	Ra I other Providence - Pro- Northern - Pro- N	see Cri	a in the	Sector Astron	E subre 2 Cour 0	· 2	C (1)		
A B A Can A Can A Can A Can A Can A Can A B A B A B A B A B A B A B A B	Calles Calles	August 7 - 10 - 10	ници (на - А́ А́ А́ В́ - А́ А́ А́ В́ - А́ А́ В́ - А́ А́ В́ - А́ А́ - В́ - А́ А́ В́ - А́ - В́ - В - В - В - В - В - В - В - В	Factors 1 F C C C C C C C C C C C C C C C C C C C	100 (000 F 100 F 100 F 100 F 100 100 100 100 100 100 100 100	Sales New 34 Autor 1 Autor 1 A	Report sens sens s	- Microso h. + : 53 unerer	rt Doel	Ra atom- Av Are	Cateron Spectra	10 P	200 N	E sole 2 Corr 0	inter a state	r A		
Marken Service	Caller Caller Caller Country	140000 F - 141 - 141	Constant (1999)	Former 1 Former 2000 Former 2000 200 2000 2	100 (000 F 100) F 100 F 100 F 100 100 100 100 100 100 100 100	54/m Frager 24 Frage 24 Frage 6 C Ferment 8.52 12.00 6.00 1.05 6.00 7.00 0.00 6.000 6.0000 6.000 6.0000 6.0000 6.0000 6.0000 6.0000 6.0000 6.0000 6.0000 6.0000 6.0000 6.00000 6.0000 6.00000 6.00000 6.000000 6.0000000000	Asport sou Govern sou s	- Microso 	n bor	Kallander Franklinger Ber	Contraction of the second seco	10 I. ×	2 (1) (1) (2) (2) (2) (2) (2) (2) (2) (2) (2) (2	E sono 2 Car 0	tering	a Part		
A cet A	Rupe Galden B Z B C C C C C C	Copyling F - -<	1 100 100 1	Farmer V Frage Setting Pro Property Setting Pro Property Setting Pro Property Setting Pro Property Setting Pro Property Setting Pro- Property Setting Pro- Property Setting Pro- Property Setting Pro- Property Setting Pro- Pr	Terrar Terrar	Sales regim 3 ab 76ml & A.Conter = - - - - - - - - - - - - - - - - - - -	Report ens ferrer f	- Microso 6 52 unesee	et Door	Ra la tester Ar	See See	3-1- 2	2 2 2 2 2	E sono 2 Corr	ten - 2 . See fatting	a har sout		
A - Cr - Y - Se Inter - Transmission A - Cr	Content Content Content Contenting	April 4 April 4 Apr	Example	r F F F F F F F F F F F F F	(ma Country) (ma Country) (m	Salver regen Ad Total & Contex C. Parenett & S2 \$22.00 \$22.00 \$22.00 \$22.00 \$22.00 \$22.00 \$22.00 \$22.00 \$22.00 \$22.00 \$20.000 \$20.000 \$20.000 \$20.000 \$20.000 \$20.000 \$20.000 \$20.000 \$20.000 \$20.000 \$20.000000 \$20.0000000 \$20.0000 \$20.000000 \$20.000000	Report one forest f	• Micros	at Door	Na literative de la constante	s	5-1- ×		E subre 2 Over	ten - 2 . Jan fatting	C (A)	* 0	
A Control Cont	Calles B / B	April 1 April 1 Apr		Foreign Control of Con	Terrar (Income Terrar (Income	Sales num 10 /bet 6 (/beneet 8.32 12.80 6.32 12.80 6.32 12.80 6.32 12.80 6.32 12.80 6.32 6.32 6.33 6.30 6.00 6.00 6.00 6.00 6.00 6.00	Report ensi Servi X	- Meres 5 5	A bor	Real Francesco	near Cet Mitter - Bases -	July N	20 S	E sone 2 Corr	ten * 2 · Participante P	a (11)	- 6	
A Car X See X	Callen	Anno Anno Anno Anno Anno Anno Anno Anno	International (1999) International (Foreign (1997) Foreign (1997)	1440 (1440) F 1000 F 1000 F 1000 100 100 100 100 100 100 100	Sales reger to the s & Center 1 (Permett 8.52 12.30 6.00 7.00 6.00 6.00 6.00 6.00 6.00 6.0	Report son Sones N	- Microso	at bor	Ra Harrison Part	See See	N II		E sono 2 One 0	ten - 2 . Bar fattag	9 (A)		
	Content	Anno an		Freese 7	1000 (mm 2000) 2000) 2000 200 2000 2	Salver reger to Autor Au	Report ensi general g	- Microso	ant loce	Real Providence -	S S Con	An I-	2 2 2	I som	ton - 2 . States	a .		
O • • • • • • • • • • • • • • • • •	Y Rent Gales K Z U Counting Co	Amburit F Amburit F	International (1999) International (Farmer 1 Farmer 2 F F F F F F F F F F F F F	Tess (Inco	Sales nue in Autority (Center i Sales (Center i Sales	Report ensi General F	- Morea	ant Door	Na II States Are	see Span-	Martin Carlos		I Addi	ten - 2 . Ber	Q.		
	V Rept	1,000 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		Forest C Forest C	1000 (mar 2000) 2000 200 2000 2	5 Allen nager A A Neet 4 A Center 1 5	Report ensi server	- Moreau 5 Sa tarate	ant boer	Re la	S S	×	2 	E Adda	ten - 2 	4		
Orient Y 20 Orient Y	V Rept	Initial I I I I I I I I I I I I I I I I I I I		Farmer 1 Farmer 2 F F F F F F F F F F F F F	Terrer Terre	54/80 144 5 7 7 7 6 7 7 6 7 7 6 7 7 8 12 10 10 10 10 10 10 10 10 10 10	Report oni Source N	- Moreau 5 Sa turner	in Door	Ra II anteres Ex	N Sge	N N		B hata	ten - 2 . Sec. fam. P	C C C C C C C C C C C C C C C C C C C	. 6	
Construction C	Y Rept	10,443 3 10,443 3 10,443 4 10,444		Factors 1 Factors 1 F F F F F F F F F F F F F	(not (not	Saless Input in a 1 had a & Center i generative 1 have 1 h	Report son Soned S	- More	A Dor	Ra Fa	5	No.		8 Anto 2 Char 0	ten - 2 . Batting fatting	4 4 4	. 6	
A Constant of the constan	* Rani Gales B 2 S Guarding S	initial i iii i iii iii f iii iii f iiii iiii f iiiii iiiiii f iiiiiii iiiiiiii f iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii		Factor 1 Factor	(nor (nor) (nor)	Saless Autor Saless	Report ens Server S	- Moos	at bot	Real Former	i i i i i i i i i i i i i i i i i i i	5. I.		E hotel 2 Der	ten - 2 . Jan Integ	Q.	. 6	
Construction C	e Rantin Galeri E d'united Galering E	Initial I * [11] * * [12] * * [12] * * [12] * * [12] * * [12] * * [12] * * [12] * * [12] * * [14] * * [14] * * [14] * * [14] * * [14] * * [14] * * [14] * * [14] * * [14] * * [14] * * [14] * * [14] * * [14] * * [14] * * * * * * * * * * * * * * * * * * * * *<	Control Contro Control Control Control Control Control C	Factors 1 Factors 1 F F F F F F F F F F F F F	Control ((Contro)))))))))))	Salense August A	Asport Server Server S	- Moose	i a ga	Ra Francisco Alexandro Ale	S S	July N	3	E setti 2 (tar	ter - 2 terry	2 (A)	. 6	
A construction of the second sec	e Rent Gales B 2 S Gales Controls Contr	+ 11 + 11 + 12 -		Farmer 1 Farmer 1 Farmer 1 Farmer 1 Farmer 2 1 1 1 1 1 1 1 1 1 1 1 1 1	test	Salense ad Trans A d 5 a d	Report ensi s	- Moos	t bor	Re Former and State	and the second s	Jali ×		E Andre Tra 1 2 Char 0	ter - 2 terry	9 04 0	. 6	
Construction C	e Auri	Initial I • 11 • 11 • 12 • 13 • 14 • 12 • 12 • 12 • 12 • 12 1	1	Factor 1 Factor 1 F 2000 2	Control	540mm 144 540mm 540m	Report seni seni seni seni seni	• Morea	A Bor	Re la constante de la constant	Sector Space	5-1 · ×	Construction	E note 2 Out	ton - 2 - Second former	Q.		

The job can be scheduled using cron job in the linux



In the crobtab –e space add the minutes, hour, week, month and year. The job is executed using **kitchen.sh** in the PDI, so enter the kitchen.sh pathname and the job file. For scheduling a transformation you have to use **pan.sh** instead of kitchen.sh. The readers can learn more about the cron timing options from the internet, plenty of material is available.



In case if you have predefined format with color coding and formula in the excel sheet then you can dump the data into the sheet and link it to the main dashboard sheet. You can create as complex report as possible using these tools. The advantage of the reporting is that the regular users are more comfortable working in the spreadsheet. The users can quickly do their own additional calculation.

2.1.2 Mondrian OLAP Cube

In the previous section we have learn about PDI ETL tools for ETL function and as a reporting agents. In this section we will learn how to design an OLAP cube using Mondrian (Workbench Schema) and how to render it to the Saiku Analytics Front end tools for the end users. On a very simplistic term an OLAP cube is similar to Excel Pivot table with data residing in the database. The user access to that data through MDX query by dragging and dropping the dimensions and metrics in the browser workspace. The main component of the OLAP cube is Dimension and Metrics. The Dimensions are the categorical data, the structure of the report that forms the rows, columns and filters. The Metrics are the facts that fill up the report body. You can form hierarchy of the dimensions in the design like Category -> Subcategory, Country ->State -> City and so on.

We will use the OLAP table as the main cube table and the Product and Calender Table as the additional dimension outside the cube. The dimensions from Product and calendar will be linked with Sales Cube. The external dimension can be shared by multiple cubes. For example a Product Dimension like Category can be shared by Sales Cube, Purchase Cube, and Inventory Cube and so on.

To start the Mondrian, go to the folder where it is being installed. In this instance we have installed it in the same location as BI server which is advisable.

File Edit View Search Terminal Help ec2-user@ip-10-0-0.87 ~]\$ cd /opt/pentaho/sc cchema-workbench/ script/ ec2-user@ip-10-0-0.87 ~]\$ cd /opt/pentaho/schema-workbench/ ec2-user@ip-10-0-0.87 ~]\$ cd /opt/pentaho/schema-workbench/ ec2-user@ip-10-0-0.87 ~]\$ cd /opt/pentaho/schema-workbench/ ec2-user@ip-10-0-0.87 chema-workbench]\$ ls pappend.bat LICENSE.html tice set-pentaho-env.bat od mondrian.properties rivers PentahoSchemaWorkbench_OSS_Licenses.html SAPI.properties plugins vy.xml readme.txt workbench.sh schema-workbench]\$ ec2-user@ip-10-0-0.87 schema-workbench]\$	ec2-user@ip-10-0-0-87:/opt/pentaho/schema	vorkbench _
ec2-user@ip-10-0-0.87 ~]\$ cd /opt/pentaho/sc chema-workbench/ script/ ec2-user@ip-10-0-0.87 ~]\$ cd /opt/pentaho/schema-workbench/ ec2-user@ip-10-0-0.87 schema-workbench]\$ ls pappend.bat LICENSE.html set-pentaho-env.bat crisce log4j.xml set-pentaho-env.sh od mondrian.properties workbench.bat rivers PentahoSchemaWorkbench_OSS_Licenses.html workbench.command workbench.sh Vy.xml readme.txt bchema Workbench.app ec2-user@ip-10-0-0.87 schema-workbench]\$	File Edit View Search Terminal Help	
	ec2-user@ip-10-0-0-87 ~]\$ cd /opt/pentaho/sc chema-workbench/ script/ ec2-user@ip-10-0-0-87 ~]\$ cd /opt/pentaho/schema-work ec2-user@ip-10-0-0-87 schema-workbench]\$ ls pappend.bat LICENSE.html log4j.xml d mondrian.properties IVERS PentahoSchemaWorkbench_OSS_Licenses. SAPI.properties Elugins vy.xml readme.txt E Schema Workbench_app ec2-user@ip-10-0-0-87 schema-workbench]\$	ench/ set-pentaho-env.bat set-pentaho-env.sh workbench.bat tml workbench.command workbench.sh

Start the Mondrian with **\$./workbench.sh** command.



The Mondrian windows look like below.



First you have to connect the database. Here we connect to Pentaho database using same credential as we used in PDI



Dile Edit Yiew Option	s Windows Help	Schema Workbench		- 0 ×	
0.000	1	Database Connection			
Schema - New Sch 公人人内 Schema	beneral Udvanced Dyblions Paoling Clustering	Connection Name Pentaho Connection Type: MaxDB (SAP DB) MenetOB MySOL Native Mondrian Neoview Netezza Oracle RDB Pentaho Data Services Pentaho Data S	Settings Host Name Jocahost Database Name Pertabo Pertabo Port Name 3306 User Name Peot Password Was Result Streaming		Cicied fai

Once connection is tested we are ready for designing a cube. Click on Schema and name it by company name. Here I named it XYZ. Right click on Schema you will see lots of add items list. To start with Click on add cube. Name the cube as Sales.

Sche 9	人大	C (Schema	t uor	CH.	0	÷ģ	0.	Q¥_	9	因	5	2	X	0	0	1	07	e" []		
Scha	Add cub		1.				-			_	4	Schen	1.8	Value					1.	
	Add Dim	ension			trapu		212	6						VAID	-					
	Add Use	r Defined	Funct	ion	Capto	09						_								
1	Add Virt Add Role	ual Cube			-		_													
	Add Par	ameter																		
			T		1) -															
																			title	id fold
																				otle

As I mentioned Cube is made of two components the Dimensions and the Metrics. Right click on the Sale cube add Table. In between you will see many red colored warning at the bottom, ignore it at this stage. The final error will be known at the time of publishing the cube.

₿ · 6	0 6 5									
Schema R L	XYZ (Schemal	uor cu <	. [66]	a (9)		• 1		111	e' C' 13	
AF 74	24 mg		Ψ[:+]	9 9	NR 100 123	Cube	• • u	l'a l'a	*	
1 aSale	1	Attr	ibute				Value		1	
*	Add Dimension			Sales						
	Add Dimension	Usage								
	Add Calculated	Member		× ×						
	Add Named Se			2						
	Add View									
	Add Inline Tabl	le								
	Add Annotatio	05								
5										-
										titled fol

In the Table space, add database name Pentaho in the schema and Table name OLAP in name. If you want to keep an alias of table you can add alias name.

1.xmD*		o" tf 🔯
not cat 👌 🚓	4 4 🖸 💹 🔝 🔔 👗 🖏 🖄 🎁 🕼	
-	Table for 'Sales' Cube	-
Attribute	Value	
name	OU#	
akas		
		titled
	ug cg viii Attribute schema name allas	veg eq 🔖 iiù o 🕰 🔯 🐹 🔊 🙇 💥 🕲 🖄 🗊

Right click on Sales cube you will have option to add Dimensions and Metrics (Measures). You can start with any one of them. Here I will start with the measures.

The first measures I add is Sales Amount. Add the name as Amount, in aggregator add sum from dropdown, add column name from table which is Amount in OLAP table and add data type.

e Edit Viel	Options 6 5:	Windows Help		
Schema - X	YZ (Schema	1.xm0*		o' 17 🔯
9 1 7	4 KD M3	vor or 0 the		
Schema			Measure for 'Sales' Cube	-
@ sales		Attribute	Value	
- Table	CLAP.	description	Anova	
Amou	1	adduedwor	Sym.	
v 1000		Kolumn	Amount	
		datatype	Numeric	
		formatter		
		caption .	8	
				titled (
	Bantaha /	1011		

Similarly add Sales Quantity as shown below.

V?	ip-10-0-0 Windows Help	-87.ap-southeast-1.compute.internal:1 (ec2-user) - VNC Viewer Schema Workbench	- 0 ×
G - G B F: Schema - XYZ (Schem 安 保 人 大 心 *	al.xm0+ 5 ucy cy 0 +++	< < < < < < < < < < < < < < < < < < <	e" 17 (13)
Solvena Sol	Attribute hame description isopresider column formatisming detatype formatism caption visible	Measure for 'Sales' Cube Value Quantity Sum Quantity Refere 2	asted folder
t 'et gatabase - Pentaho	(MySQL)		

Now that we have added two measures lets learn how to add dimensions. Right click on Sales cube, select Add Dimensions.

Right click on Dimensions and select Add Hierarchy.

	1 60 M	uoy (0 10	4Q -	. 9	8			x	0	TT.	EY.	e' tř . 🛛
schema	+ -+	10	+ +		• •	000	Dimensio	n for 1	Sales	Cube	10.00		1.
D xsules		name	Attribs	te	New Dim	ension (0			alue			
Anter	Nemension 0 Add Hier ar	desci foreig chy	iption nKey Vefa		Standard	Dimensi	ion						
O ours	Add Annot	stions	-		9								

You will see two Hierarchies in the dimensions. You can delete one of them if not required. Right Click on Hierarchy and select add Table

Q 人 大 内 xSchema Stable: OLAP 大 XNew Dimension 合力 XNew Heers 合力 XNew Heers 人 XNew Heers の A mount く Quantity	ng on 0 echudi Add L Add Ju Add se Add A Delet	uge cu	the form	Q Q Hiera Neen Hierart Z	Image: Second se	ew Dimens	K D Dim Value	ension	•	citled fo

Again right click on Hierarchy and select add Level. The table name is not actually required as we are creating dimension out of same Table as Sales cube. The table has been added so that we can add level. Once level is added delete the Table immediately to avoid conflict with cube Table.

-			5	of cat	0 "	\$ 3	<i>A</i>	52	1	2	. X	0	0	a.	U.			
				-	Itchute	-	Hie	rarch	y for '	New Di	mensio	Value	Dimen	sion	_	-		-
res				name	tripute	N	eur Hier	archy d					_					
Mare C	Newaria	0.40		description and an and an	Dra .	9	8	1000										-
17. 184	w Hier		L	alMenbe	vitame													
	Table :	Add L	evel	_	ne					_			_					
th and	w Hier	Addjo	sin	×	aber-			_									_	
Amoun	e	Add V	iew	332	Table													
Quantit	W.	A661	sline anal	Table	r			_			_	_	_	_	_	_		
		Delet			-		6											
																		titled

In the level add the Level Name, Table as OLAP, column as city and in Type add as string.

Schema - XYZ (Schem		* d' 🛛
A the formation of the second of the se	Image: Construction Image: Construction Image: Construction Construction Image: Construction	In a circle of

Select Level Type as Regular. Rest of the fields can be blanks as we can do without those values. You may fill values if required.

Schema - XYZ (Schema)	Lxml)*		o* 17 🖾
日人大的時	not of 0 43	<pre>< < I I I I I I I I I I I I I I I I I I</pre>	
hema	•	Level for 'New Hierarchy O' Hierarchy	•
sales	Attribute	Value	
Table: OLAP	description	Oryfame	
A star Dimetrica A	table	CUP	
V MARKA CARRENDON O	Kolumn	ON .	
ATA New Herarchy O	nameColumn		
- **** ObyName	nulParentValue		
Table CLAP	ordinalColumn		
And allow Merseller 1	1004	String	
DO MARK MARKON I	internalType		
O Amount	unqueMembers	Manadar	
O Quantity	hideMembert	Bendar	
	approxRowCount	TimeYears	
	c aption	TimeHalfYears	
	k ap6onColumn	TeseHalfrear	
	Formacter	TimeQuarters	
	a state	Tanemontos Ministradas	
		Tradius	
		construction of the second sec	

Now add values in the Hierarchy. Here again add name, table name and the table primary key as shown below.

Scheme - XYZ (Schema)	Lxmi)• uợ cự ⊘ ôĝ		o' t' 🖾
xSchema	Attribute Attribute Attribute descrption AnsAl alMemberCapton allereberCapton allereberCapton allereberCapton primarXer/Capton primarXer/Capton visible	Hierarchy for 'New Dimension 0' Dimension Value Cry OLAP Ordend	tries fo

Got to Dimension and add name and type. For the internal dimension these two values are required. For external dimension we will learn in later part.

G. G G G S: Schema - XYZ (Schemal xm Q人大心、吗。	t c# Ø +↓ 10•	4 % <u>8 8 8 8 8</u>	6000	e" cf _ 10
 County Instein CLAP Instein CLAP Init Coplance Init Coplance<th>Attribute anne becropson oreignKey rspe sagetrefus ageton sisble</th><th>City StandardDimension Ki</th><th>Value</th><th>the state</th>	Attribute anne becropson oreignKey rspe sagetrefus ageton sisble	City StandardDimension Ki	Value	the state

Finally delete the Table from dimensions.

8	ip-10-0-0-87.ap-southeast-1.compute.internal:1 (ec2-user) - VNC Viewer	- 0
le Edit View Options G - cs D B . f: Schema - XYZ (Schema)	Schema Workbench Windows Help	- 0 X
会人大的 Mi Schema		
B states Table OUP A sCey + dta sCey -	Sobera Pestaho name OLAP	steet fai
Database - Pentaho (H	Schema Pentaha doos not exist ySQL)	

Now we have basic component call Sales cube as OLAP cube and publish it. It has two basic components – Dimensions and Measures.

Save the Cube in the folder of your choice.

3.600.0			
Schema - XYZ (Schem 日本の×	al xeels g uug uug 🖉 mig og 💖 😽		e* 6* 😒
xSchema ② xSdes □ Table OLV T 人 Cty ↓ Cby	Seve ja: Cjunin		1.
O Anoue O Anoue O Oversty	el		DEbed folder
		Save	

Click on File -> Publish

V1	ip-10-0-0-	87.ap-southeast-1.compute.internal.1 (ec2-user) - VNC Viewer	- 0
A	Windows Hele	Schema Workbench	o x
New Open	,		
Save Save As	Ce 5 100	• 9 8 8 8 • V 5 • 9 7	d. 13
Pentaho.ml		Hierarchy for "City" Dimension	1.
	Attribute tre scription	Value City	
+ CA Day 	all-famber Same all Aranbar Capton all area Name default-famber member Marker primary Key Capton Visible	cue Crient K	
			stied falder"

Add Server URL, User and password.

4.90.6			
Schema - XYZ (Pentaho.x)	ml)	*'d' 🛛	
日本的吗。	1 CH 5		
Schema States States A Cay + Ch Coy + Ch Coy	Atte Rama description hasAll alMenbach alLevelham defactben primium/Kep primium/Kep primium/Kep primium/Kep	nsion • Nue e: e: sh Cancel	ed fold

The cube is Published Successfully.

R	ip-10-0-0-87.ap-southeast-1.compute.internal:1 (ec2-user) - VNC Viewer	- 0
ile Edit View Options)	Schema Workbench _ O X Windows Help	
Schema - XYZ (Pentaho		×
Schema Schema Table OLAP Arcay Amount Country	Attribute Schema Value Schema	urettes folds
Database - Production	(MySQL)	
	¢	

Now go to browser login. Open Saiku Analytics.



Select Create a New Query.

Welcome Fea	tures Get Help En	Terprise Explore Data. Visualise. Act.
Saiku and an and a state of the	e allowal prior Subartness and matter descenses. Easter proviets no argumentation. Charling and easter priorities and does not been any later	Thank you Our statements to fails Reporting too save fully factors. This is presentence to the fails proton and with the an is fault and a protone. Save fails Reporting protons with the fee these are consisted. (Into base for more update) <u>Consistent</u> <u>Consistent</u>
Curck Links Credit a time party Unit the cateline	un a dachtaand Prijstikking	

In the workspace click on select a cube dropdown. We can see our cube with Schema XYZ and Cubename Sales. Select sales cube.



You can expand Measures and see two measures we created – Amount and Quantity. Similarly open dimensions you will see city. Now just you have to drag and drop to create a report.

Saliu Analytics -						
Cubes 5	s 8 0 📝	• • • •	2 4 4 4 × 34	A. & W. A. J. A.	0 0	101.4
ten •	You are using Solic Coron	nty Editor, places consider upper	ing to Lake Colorantes or antering a p Indexe affe	antantia anatori ali a 4 mant a Kanz	started Scientific is annual controls of	-
Manaret Add	Basers	•				1
Conversions	Column	-				
* On	Rent	·				
	Pau	•			b	

Add Amount and Quantity in the measures and Cityname in the Rows. The report is displayed as shown below

Salky Analytics			
den S	980 P + F	CORCERSENCE	100 4
(m •)	The are using fails. Conversity Ballion pile	na surante appinte la <u>labalistazion</u> e atterio a <u>suranzi in Australia di un sepet de superet (a surabin la santa au suranda ani</u> Internativa	1
headins Add		Bellevil Seally	-
Arrant Burnty	Annual Guerrity	Bengden J*9.70 10 Demos 1403 90 100	
alignations .		Hydenatual 1.000.403 (45) Harebal 1.760.403 (55)	
Cay (M) Ch/hane	Caluma	New Balls Cont, 700 Up	
	Auro *		
	Oyfure Oy		
	-	D ₂	

Sine we have only one dimensions and two measures, there is hardly anything we can do on the cube. Let's add more Dimensions and Measures. Right Click on Sales cube and select Add Calculated Member.

ile Edit	yiew g	Dptions B 8:	Window	Scher s <u>H</u> elp	na Workb	ench		_ 0 X	- 0
Schen	na - XYZ 人 大 ma	(Pentaho	uer c	ų S Aû	o _↓ 9¥	Schem	Value	ं त छ 0 1 07	
2 2 0 0 0 × 0	Add Di Add Me Add Ca Add Na Add Ta Add Ta Add Ta Add In Add In	mension basure iculated med Set ble two ine Table notation	Usage Member o	on sCaption Je	XYZ				
	Delete			_					

The calculated members are created out of existing measures. Here we are creating average selling price (ASP) which is calculated as Amount by Quantity. In the name section add ASP, Select Measures in dimension; add formula in the Formula space as [Measures].[Amount]/[Measures].[Quantity].

Vi Die Ldit View Options Die Col Die To	ip-10-0-0- Scham Windows Help	17 ap-southeast-1 compute.internal:1 (ec a Workbench	2-user) - VIVC Viewer _ O X	- • ×	- •
Schema - XYZ (Pentaho Q A A A A A A A XSchema + Q xSales - Table: OLAP - A xCBy - A mount - O Countby - O Countby - O Discount - O Discount	xm0* Uge Cg Q ing ing Attribute name description caption demession Merecky parent visible formula formulation formula formulation	Image: Second State Image: Second St	er of B B B B B D obe		unklöged folder

Similarly add Discount percent by adding Discount as measures and calculate discount percent as 100*[Measures].[Discount]/[Measures]. [Amount].

R.	ip-10-0-0-8	7.ap-southeast-1.compute.internal:1 (ec2-user) - VNC Viewer	- 0
C Ele Edit View Options) C	ip-10-0-0-8 Scheim Alteribute Parme Gescription Captio	IT ap-southeast-1.compute.intenait1 (ec2-user) - VNC Viewer Workbeach Catculated Member for "Sales" Cube Osceant Measures Not-Theasures][Cosceant](Measures][Amount]	- C X
Amount Quantity Discount CM ASP CM Discount	Caption demension herarchy parent visible formula form	Measures 20 100=[Measures][Qiscount][Measures][Amount]	untitied fair

Let us add global dimensions. Right click on schema, select add Dimension.

File Edit Yiew Options Y 日一日日日 E Schema - XYZ (Pentaho 日本山 和	(indows (ml)* uor cw	Rep.	4 Q	9,	8		2	X	0 0	'd 8		110	
Standard Cube Add Cube Add Dimension Add User Defined F Add Virtual Cube Add Virtual Cube Add Faremeter Add Faremeter Add Annotations CM Discount	unction	tribute 0 Caption 0	*	72		Schem	Veh	20				0	unitized fo
-	-				Cub	e Sales is	invalid .			 -			

Select table as Product, add category dimension. The dimension adding is same as describe earlier, hence I am not showing each steps henceforth.

Edit View Options Vindows Help Edit View Options Vindows Help Schema V72 (Pentabo xml)* Anount Table: OLAP Anount Anount Column		
Q cal (C) (G) (Z) Schema - XYZ (Pentabo.xml)* Q) (A, A, (A) (A) (A) (Q) (A) (A) Y states Table: COP A mount A mount Discourt (A) A (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) (A) <th>ema Workbench _ O X</th> <th>- 0 ×</th>	ema Workbench _ O X	- 0 ×
Schema - XYZ (Pentaho.xml)* G & to		10
Schema - XYZ (Pentabo xm0)* Q A TA A M M US A TA Schema Sch		
Image: Solution of the soluti	e ^x 0 ^r 🖾	
Schema Attribute Sales Attribute Table: OLAF Schema Column Annount Discount Discount Obscount Discount CM Discount Discount Chin Mawu Hierarchy 0 Capbon Cable Nonuer Hierarchy 1 Column Capbon Cable Column Database - Production (MySQL) Column Column Column Column Cable Column <td< th=""><th>1 4 4 🖸 🚺 🔝 🔍 🖉 💥 🕲 🖞 🖓</th><th></th></td<>	1 4 4 🖸 🚺 🔝 🔍 🖉 💥 🕲 🖞 🖓	
Image: States Attribute Table: CLAP Amount Amount Discount Outatly Column Discount Discount Manout D	Level for 'New Hierarchy O' Hierarchy -	
Table: CLAF CGY Anount Counting	Value	
Anouet Albee Anouet Anouet Anouet Anouet Discourt Anouet Discourt BareeColumn CM ASP BareeColumn Column BareeColumn Schema Yesters Schema Schema Schema Anoust Schema Anoust Schema Anoust Schema Anoust Schema Anoust Schema Calone Schema Calone Schema Schema	Category	
Namouet Namouet Ouantity Discoutt Discoutt ParentColumn CM ASP ParentColumn CM ASP ParentColumn CM Discoutt	Product Category	
> Outstray > Discourt > Discourt > Office Court > Office Court > Schema > Table: Production (MySQL)		
Control Control Cont App Control Cont App Control Control Control <td></td> <td></td>		
CM Discourt Were Demension 0 An Alker Herarchy 0 An Alker Herarchy 0 An Alker Herarchy 1 Communication An Alker Herarchy 1 Communication Cathology Table: Production (MySQL) Communication Cathology	String	
Afters Demension 0 Image: State Sta	0	
An when Hierarchy 0 approximation Court cachine Cachin Cachine Cachin Cachine Cachine Cachine Cachine Cachin Ca	Regise	-
Action of the second of t		
Table: Product Seminator ATA information Seminator ATA information Seminator Schema Schema Atable: CLAP Attribut Schema Attribut Schema Attribut Answet Attribut Attribut Attribut Attribut Attribut Attribut Attribut Attribut Attribut Attribut Attribut <td></td> <td>untitled for</td>		untitled for
A.A. INSERV Heavy (by 1 Image: Solution (MySQL) Database - Production (MySQL) Database - Production (MySQL) Column Schema Sch	2	1
Database - Production (MySQL) Ig-10- Schema - XY2 (Pentaho.xml)* @ /, d // di % @ C () Schema - XY2 (Pentaho.xml)* @ /, d // di % @ C () Schema - XY2 (Pentaho.xml)* @ /, d // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () Schema - XY2 (Pentaho.xml)* @ // di // di % @ C () // di // di // di % @ C () // di // di	. ex.	
B etal B. E. Schema - XY2 (Pentaho.xml)* Q J. J. J. M. M. Weg. Cell Y stales Table: OLAP Attribuet. Table: OLAP Anount Discourt CM Discourt CM Discourt CM Discourt Attribuet. Attribuet. CM Discourt Attribuet. CM Discourt Attribuet. Attribuet. </th <th>ema Workbench _ 🗆 X</th> <th>- 0 ×</th>	ema Workbench _ 🗆 X	- 0 ×
Schema - XY2 (Pentaho.xml)*		1
Qi Å ★ Ait (Ai) Mit usige (Cit) Ait (Ait) States Table: OLAP Ait (Ait) Ait (Ait) Ait (Ait) Qi Discourt Ait (Ait) Ait (Ait) Qi Discourt Cit (Ait) Cit (Ait) Ait (Ait) Cit (Ait) Cit (Ait)	*'d' 🖂	
Schema S	÷ < < 2 2 X 0 0 1 0 0	
Attribut	Hierarchy for 'New Dimension & Dimension	
A cost	Category	
Anoune Anoune Quanty Discount CM ASP CM Discount ANNew Dimension 0 STA Actegory Table: Product ATD affeet Micrarchy 1	8	
Quantity Discount CM ASP CM Discount A SP CM Discount C A SP CM Discount A SP CM Discount C A SP CM Discount C A SP CM Discount C A SP C A Discount C A Discount	(
Discount Getaddeebee Getaddeebeebee Getaddeebee Getaddeebee Getaddeebee Getaddeeb		
CM Discourt CM Discourt A xNew Dimension 0 A xNew Dimension 0 A xNew Dimension 0 A xNew Dimension 0 Caption A xNew Dimension 0 Caption Cap	55	
Capton Capton A xNew Dimension 0 visible A xNew Dimension 0 A xCategory Table: Product ATA stress Hierarchy 1	Product Productd	
Ath Acceleratory	2	
Table: Product	10	
Table: Product		untitled for
- Eh aften Herardhy 1		
1 11 Colora		
Column 1		
Catabase - Pantaba (Mr.501)	roductid defined in field primaryKey does not exist in table Product	
Cartering - Faultano (197941)		

I have named all level, hierarch and dimension as category. Here we will not delete the table because the Dimension will pick up data from this table not from the Sales cube.

Vi A File Edit View Options V	ip-10-0-0- Schem	Tap-southeast-1.compute.internal:1 (ec2-user) - VNC Viewer Workbench	
B Cl D B. 5: Schema - XYZ (Pentaho. Q A. A. AQ M. Schema Schem	cm0+ vog ce se initial Attribute Rame description foreigney tote capton visible	e' o'	unotical folder
Database - Pentaho (My	รจม	Herarchy Category is invalid	

Similarly add Year from the Calender Table.

0.000			
Schema - XYZ (Pentaho.	xm0+		
04 14 74 149 14	och ch 1 1	4 4 M M M M M M M	
*Schema	-	Level for New Hierarchy & Hierarchy	
💓 xSales	Attribute	Value	
Table: OLAP	description		
- 1 cer	table	Calender	
A second	column name/alumn	Year	
O whome	parentColumn		
Ocanity .	nultrarentivalue		
- O Discount	ordinalColumn		
ON ASP	1/24	String	
CM Discourt	microaldambars	0	
1 sections	eveltype	Hair	
V screedory	hideHenbert		
* ATA Category	approxilowCount		
++++ Category	Cap6o4		untitled to
Table: Product	formatter		
A New Cimension 1	visible	×	
Chatter Harriston			
Taxa anew retractly 0			
····· Year			
Table: Calender			

Add primary key table and primary key of the table

e Edit Yiew Options	<u>Windows H</u> elp		
t . a 0 8 1:			
Schema - XYZ (Pentaho 会人人人 你 你 XSchema States 一面 Table: OLAP	xmD* Vor Cu o iii - Hit Attribute name description	* * <th></th>	
Arrount Arrount Orushity Discount CM ASP CM Discount	hasAl alMemberKame alMemberCaption alLevelName defaultMember memberReaderClass primaryKeyTable primaryKey	Ki Calender Date	
人 xCategory ATh xCategory 前前 Category 同 Table: Product 人 xNew Dimension 1	visible	8	untitled fo
• ATA stream ····· Year Table: Calender			

Similarly we have added month as dimensions.

Ca Edit View Options Window	s Help		
Schema - XYZ (Pentaho,xm)* Q /s /s /s /s /s /s /s /s /s eschema Add Dimension Add Dimension Usage	N Attribute	of of S Cube Value Sales	
Add Measure Add Measure Add Named Set Add Table Add Table Add Table Add Inline Table Add Inline Table Add Annetations Delta	goon che ubled sble	8	
+ & Kategory + Gh sCategory - III Category - III Table: Product + & where + Constant			
Table: Calender	Fact tabl	e OLAP does not exist in dutabase schema Pentaha	

The global dimensions created can be used by any cube under Schema XYZ. Here we have only one cube Sales. To use the global dimensions in the cube, so to cube, right click and select add Dimension Usage.

In the Dimension Usage windows we have to add name of dimension for the cube, foreign key and Source of the dimension. The Foreign key is the primary key of the dimension table in the cubes table. Here the productid is primary key of the Product table that link with OLAP table.

t Cing & Q		
name foreidaKey souit@ visapitefix capton visible	Value Category Producted Category Tesi Rooth Roo	unitided fo
	sagehrefix Capton visible	Septerix rear Caption team vicible to Source must be set

Similarly add Year and Month using similar process. Here the foreign key is Orderdate.

Q L to M	ung cu 🖉 ng	< < < < < < < < < < < < < < < < < < <	D X						
3 xSchema	*	Dimension Usage for 'Sales' Cube +							
Table: OLAP	name foreignKey	Year OrderDate							
- 人 xCey	source level	Vear							
大 Year	Capbon Vicible	9	e:/						
大 Month	1000								
O Quantity			ora						
S Discount			tes						
CM Discount Percent			100 C						
人 *Category			pra Cristi						
- A stream									
V									
Schema - XYZ (Pentah	uer ce 🛇 ñó	<u>.</u>			X IB	0	ೆ ರ ಟ ಗೇ [)/	5 X	
----------------------	--	-------------------------------------	-----------	-------	------	---	-----------------	--	-----
	Attribute name foreignKey foor(e feed icagehefix capton visible	North Code/Date Month Roth	usage for	Value	cube			e=/ e:/ ora_ ora_ ora_ ora_	Let

Publish the cube again and open the browser after successfully publishing. Now you can see other measures and dimensions we have just added.

Salio Analytics -														
laten S	- 8 0 Z + F	1		H 9	1+	н		4 #		\mathcal{R}	v.	0	0	1014
ten •	The are using Sales Constanty Billion pla			Laborations			-	Allerant	10.10	-	-	mart ()	added and a second and	
Restartes Add	-	Calegory	-										Junio 1752 2 3 1 4 7 6 88	~
Arrest Courts	Arrest ASP	Apparents Desites	101.000 201.02	100.0										
Aller Aller Channel	Columna ·	Market Market	100.00	1.001										
Nenamalaria														
Campiny (48)	nus •													
Campuny Day VAR Couhanne	Cangary Cangary													
10	****													

You can drag and drop and try out different combination. In fact there is charting option in the right side. Explore the options.

aku Analytica								
6es 5		1	9	e 9 4		A & #	x & e 0 0	101.4
an •)	You are using Talka Caronauthy Ballian, pr	ana consiliar	uppedag to	Into Education and	-	standa arestato	the design of the second of th	-
MA REPORT				12.27	-		and an U.S. A had a differ	-
	Basaria *			2816		2814		~~
Access The Party of Street	Arrest	Category	Annut	Shinet, Petert	Annut	Doctor, Parcent		
Quantity	Decost, Pecart	Apparets	180,000	180	30.800	8791		1.00
Decent.		Botta	40.000	4.92	10.840	8.494		
ALP .	6.00mm	CANE	+ 302 800	041	100.000	4.08		
And and Andrea		Int	100.405	1.01	80.400	1.00		
analota	100	10000						
Category	1.00							
Calegory	Acres .							
Chy	Compro							
140	Calegory							
Storth								
Tast	Contract in the second second							
100	104						1.20	

Salku Analytics -																
uten	5		2	+ (#	14	9.6	9	1 16		8	8	8.0	0	0		100.4
-		The set using Sale	Corne	in Editor, pa	-	approving to La	RA.SHOULD		nunita a	meterlaß.			grant ()	contribute in an internal of	in doors	
	4.04							banades	anti					Artist 17.50 1 1		-
	1000	Monores						205		2010					000.032	~
Arrange		Annual			Calegory	Ol/tare	interest.	Decout, Percet	Anapet.	Owner, Pe	-					
County		Declark Pecar	-		Apres	Bergatura	21,400	10%	1,210		104					x
ADP						Pythesisted	38.800	1.001	16,209		1.040					
Dataset, Percent		Columns		•		three	31.200	2.4%	*.400		5-894					
amazona .		100 C		764	Sec.	Res-Debi	28.800	1470	2.810		+10					
Catalogue and Catalogue an		198				Owna	10.100	4.04	1.810							
140						Hyderabat	8.400	2 898	4.529							P.
Category		Area		•		New Cast	8,800	110	1.040		-					
LAR				Category	Lane	Bergature	NUM-000	1.07	182.000							
Cliniane		Category				Owne	86×.000	140	175.00		4318					
140		Children		CN/		Wyderstad.	\$04,000	174	885.000		1.000					
						Sec. Lake	1.074.000	* 1962	80.008		-					
Sec.																
Tear (48)					-	Bergature	10.000	4947	16,399		1279					
in the second		Filmer			Victor	Bergature Charrow Reconstruct	10.00	4.94 ⁴ 2.868	15,200		1279				- 1	
Dpened ~	10	Press and Service III			Valle	Bergatera Charma Rydenation Klarma Nas-Satti	212,800 103,800 104,800 104,800	4.94 ⁴ 2.849 6.849 6.759 6.759	19,209 101,800 86,800 86,800 101,800		121					admir
the vec hash Opened ~	Sally Ara	Anne and Service III (year (2)			Value	Bergalon Channa Rydenstaf Ukertas Nau-Sati	212 800 125 800 127 800 106 800 106 800	4 84 ⁴ 2 488 4 49 4 49 4 49 4 49 4 49 4 49 4 49	10,000 (01,000 86,000 (01,000		120					admir
the vec test Opened ~	Sala Ara	Anne And Sarveys II Anne (2)	2	•	(k) (Bengatun Channa Rydonatad Hanna Nav-Calit	11 H 11 H 12 H 1	1	4.00 44.00 14.00 14.00 14.00		× * * * *	5 0	0	σ		admir 100 (
the test of the test of the test of the test of test o	Salu Ara	Firm Inf Servers II (Con [2] (Con [2] (Con [2])	(at) Come	•		Bengatun Oseran Ryskestant Ben Dath	40 400 40 400 40 400 40 400 40 400		11.00 11.00		100 100 100 100 100 100 100 100 100 100	5 Q	0	Q.	annah.as	admin III 4
and the second s	Sala Ara Sala Ara Ata	Vera (2)	(a) Correct	•		Bengative Oversor Rystendard United New Date:	40.40 47.40 10.40 10.40 10.40 10.40	- 19 - 19 - 19 - 19 - 19	4.30 4.40 1.40 1.40 1.40 1.40 1.40 1.40 1.4		1275 1276 1277 1277 1277 1277 1277	× 0	0	U 101.00.00	antraste and 2+4 / 4 m	admir III (
alle Andre Sale	Soliv Ara Soliv Ara Att		(J) Correct	• • •	Compose	Bengative Overage Rystembert Warriss New Delti	40 400 40 400 40 400 50 400 50 400	- 687 - 199 - 720 - 720	4.30 4.40 4.40 4.40 4.40 4.40 4.40 4.40		100 100 100 100	5 0	0	C cardinal is seen as	1000k.at	admir M
and the second s	Salia Ang Salia Ang Salia Ang Ang		(a) Correct	• • •		Benjative Okarana Nyakanale Nava-Sakk Nava-Sak	42 400 97 400 10 400 10 400	- 48" - 188 - 479 - 179 - 179	4.20 4.20 10.20 10.20 10.20		1279 1279 1270 1270	× 0	0	U 2010 - 100 - 100	1000k.ad (11/100	admir Million No.
and the second s	Sala Ara Sala Ara Sala Ara Sala Ara	Free art Servers II (CG. (2) (CG. (CG. (CG. (CG. (CG. (CG. (CG. (C	2	•	Competence Competence	Bergature Cherana Ry-Constant Res-Datits Res	40 400 97 400 10 400 10 400	- 18 - 19 - 19 - 19 - 19 - 19 - 19 - 19 - 19	4,00 4,00 4,00 4,00 4,00		100 100 100 100 100 100	× 0	0	U antici anti-	1999-1994 - 1994	admir Million T
and and a second	Sala Ana Sala Ana Ana		(a) a Carma	•		Bergature Cherana Ry-Constant Res-Datits Res	102.000 (0.000 (0.000 (0.000	4 44 1 44 1 47 1 47 1 47 1 47 1 47 1 47	4.00 4.00 4.00 4.00 4.00		100 100 100 100 100		0	C and a star of	111/1 11	admir III (L
tern the tern	Salar Ana Salar Ana Ana	Free Internet internet interne	8	• • •		Bengatun Cherana Mycheatada Haritas Haritas Haritas Haritas Haritas Haritas Haritas Haritas Haritas Haritas Haritas Haritas Haritas Haritas	на на 10.400 47.400 10.400 (0.400	4 44 4 44 4 44 4 44 4 44 4 44 4 44 4 4	4.00 40.00 40.00 40.00 40.00		129 140 140 1 1 1	R D	0	U antibit discussion and U I	antractic and 2 + 4 + 4 the	admir IIII IIII IIIII IIIIIII IIIIIIIIIIII
	Sofie Ara Att	Free and Service II (Colored Colored C	Concession of the second	• • •	Katala Ka	Bengatury Cherman Warthan Haro Dath Haro Dath	на на на на ф* на 10 400 10 400	1 00 1 00 1 10 1 10 1 10 1 10 1 10 1 10	4.40 44.40 44.40 40.40 40.40		100 100 100 100 100 100	R P	0	D anti- and a	10000 and 211/100	admir III (III) II
and the second s	Sofie Ara	Free Inf Service II (Content Service II Service III Service II Service II Service II Service I	Conservation of the second sec	• • •	Malan E Galger Aganth Gan Saya Taya	Bengative (Decem Physical States Harden Harden Harden Harden Harden Harden Harden Harden Harden Harden Harden Harden Harden Harden	на на на на ф* на 10 400 10 400	1 00 1 00 1 00 1 00 1 00 1 00 1 00 1 00	4.40 44.40 44.40 40.40 40.40		129 130 130 130 130 130 130 130 130 130 130	R 0	0	C and all i	2000 A 40	adme III (III) II
territoria territ	Salu Ara Salu Ara An	Free Marca B Marca B M	Conservation of the second sec	e e e e e e e e e e e e e e e e e e e	Maller E Colgan Galanti Galant	Bergatory (Decroit Hydrosofte Hanclaski Hancla	10 40 10 40 10 40 10 40 10 40 10 40		4.10 41.40 10.40 10.40	a a a a a a a a a a a a a a a a a a a	100 100 100 100 100 100 100 100 100 100	X 0	0	O antida a series as anti- terr	eresta et	adme III (
and the second s	Salie Ara Salie Ara Att	Fire Fire (Colored Colored Colored Colored Colored Colored Colored Colored	Conservation of the second sec	e e e e e e e e e e e e e e e e e e e	Maller E Colgon Galanti Galant	Bergatory Overson Physical and Wantane New Deski New Deski Physical Physica	00 400 10 400 10 400 10 400 10 400	1 19 1 19 1 19 1 19 1 19 1 19 1 19 1 19	4.10 41.40 10.40 10.40	H H	1275 1275 1275 1275 1275 1275 1275 1275	X 0	0	Q set un	100000.001 (111/1.00	astrik Militari T
nem har in the second s	Salie Ana Salie Ana Ana	Free Free VICE (2) - Free	2 Comme	t (in the second	Maller E Colper Aparta Daris Taya	Breaker Orania Normali Normal Normal Normal Sciences Sciences Sciences Sciences Sciences	402 400 103 400 104 400 104 400 104 400		4.00 4.00 4.00 4.00 4.00 4.00		127 140 140 140 140 140 140	× 0	0	C and and a	anticula, and 2+4 + 8 the	admir IIII IIII IIIII IIIII IIIIII IIIIIII
anan taka taka taka taka taka Depended ~ anan anan taka	E Shine And Shine And And	Fine Fine	(e) Conner	e in the case	Malan	Bengdon Charata Wantas NavaCati Wantas NavaCati Walking Walkin	102 400 103 400 104 400 104 400		4.00 4.00 10.00 10.00 10.00		1271 1410 1410 1410 1410	5 P	0	C and a difference	1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.	adme Militari E

You can save the format in your folder. Next time you want to see same report just browse and open the file.

Folders
D (% D)am D)am D 10/y D 10/y D 10/y D 10/y D 10/y D 10/y

The result of the report will be displayed. You can edit it if required and save again.



This pretty much explain how to create a simple cube and render them in the browser. For the real time user the data will be much more complex and size will be very big. You have to plan all dimensions and metrics by speaking to the users before starting the cube designing.

Another important topic I will just mention here with regard to cube is that the company will have data policy where certain users are to given access to certain data. For that segregation you can create roles in the user management windows. There is option to provide different roles. Before starting plan out the Roles and create it with different right. When users are created just add those roles.

Users & Roles	Users / Roles		
Mail Server	Manage Users Manage Roles Sys	com Roles	
Setting	 Rotes + X Controls Administrative Report Aufford Report Aufford Report Aufford Report Aufford Controls Contro	or Professions robustication order Statuty of Conserver Are good Data Sources Are Conserver to C	eod b

The users roles defined in the Administration windows can be used in the cube to restrict the cube or Dimensions.



I will not explain the entire process but users can try out different combination and experiment with cube. It's not difficult.

2.1.3 Dashboards

The dashboard are also a report but what makes it different from usual report is that is more summarized form with more pictorial representation meant for the top management. Select CDE Dashboard from the Create New



For designing a good dashboard you will need to have good knowledge of html and css. In community dashboard we have three main components – layout structure which provides the structure of the dashboard, component panel which is the elements of the dashboard lie chart, tables and the datasources which provides which data sources to be used for a component panel.



We will design very basic dashboards. Select a layout. I have selected 2x2 template.

E Duthbard -		
CDE tee tee tee and family New Dashboard*	Q	5
Apple Structure C C C C C C C C Trapelles		
< >		
These Educates Translates From Educates Translates		

You can change the name of dashboard by changing the html header as shown

Editing Sales Dashboard					
CDE see tere to	everal. Reload Settings Sales Dashboard		E 2	0 2	6
Layout Structure	0.0010878733	R Properties			
Term .		Property.	1944		
+ Now	Reader	Name	Tak		
+ Column		and,	 Ind syle="calor #10.0.1 		0
- Head		Color	0		
t Ase	Budy	Cos Ciesa			
+ how	Factor				
ta Van Tools Help	Juli jeven li				
opened - 🗗	Jefferson II D-				admin +
ne vee teen ree Opened ~ 🗗 Storg Sales Dechoord	bet jarvar it. D			4	admin +
te Ven Tool Hele Opened ~ 🗗 StogSales Dathboard - CDE ten Stort Se	Selfserver B D- Seles Dashboard Seles Dashboard		a 2	, D	admin +
n ver too ner Opened - E Forg Sales Denboard CDE ner Sale Sa	Sales Deshboard	Q Properties	a 2	, , ,	admin +
n vive tools mee Opened ~ 📑 Song Sales Danboard CDE too Sale to Arood Structure	Sales Deshboard	C Angeles	E 2 ×	, no	admin •
In the Toph Indo Depended V III Story Sales Dashboard CDE Into Sale Ta Aprod Diructure III Tata	Settlerver II C state finited Setting Sales Dashboard C C C C C C C C C C C C C C C C C 664 1 av espensar mesor non-11 fertered sour	D Provins	2 4	, 10	admin + Ge
to Vew Tools relation Opened ~ E Story Sales Deshboard - CDE too Sale Sa Layod Structure C Ass - Staam	Sefferme B Constant Servery Sales Dashboard COD COD COD COD COD COD Edit Color Cod code constant, reacht festererijkawy	C Aspertis	E 4 *	, n	admin +
In You Tools Inte Oppendid V Poi Storg Sales Dashboard - CDE Inte Store for Layout Structure Inte Asso - Classes - Stare	Inflorma II D a so. Annuel Servey. Sales Desiboard C D D D D D D D D D D D D D Edit I av agle-base states, -a.Mt tensord, r.v.y	C Paperini de	E 2	, G	admin +
In two Tools Inte Oppened ~ P Story Sales Deshboard CDE two South Sa Layout Shuthure - South - South - South - South - South - South - South	Set Server II Contract Setting Contract Setting Contract Setting Contract Setting Sales Deshboard Setting Sales Deshboard Setting Sales Deshboard Setting Sales Deshboard Setting Sales Deshboard Setting Sales Deshboard Setting Seting Setting Setting Setting Setting Seti	C Projectinis	E 2	. 0	admin +
In You Tools Indo Opened ~ If RingSales Dashboard - CDE Into South 10 Jood Douture - Reason - Reasonn -	Sefferme II Control feeling Sales Dashboard COD COD COD COD COD COD 664 Control code researcy - sulfit feetime-duraux	C Properties	2 ×	, G	edmin +
In Yes Tools Help Oppened ~ P Strig Sales Deshboard - CDE	Sefferme II Constant Serrep Sales Dashboard COD COD COD COD COD COD Edit	C Augestes	E 4	, G	admin +



Go to component panel and select Wizards -> Saiku OLAP wizard

Opened - 🖻 🖻		admin ~
CDE see tes		
SUIND OLDER WILLING SUIND OLDER WILLING	Dama sound labo Davy	
KONTTAL Querra KITCL/Instrum Status Status Status Data Contaction		(mar) (0. (0)

Create chart using OLAP cube data we just created. We have four panels add name of panel, title, chart type and Panel. Here we have selected first panel panel_1



ong Sales Dashboard					
DE tree Lave Save at, Autour	Sale Deterources	s Destooerd* ふておし	e manne		Ģ
OLAP Selector wound OLAP Chart wound Selec OLAP Wound	+ Crise nds wer mendrierij	WOI Queres et Cengory/stimQuery	Arran Lood Jol Martine Lood Jol Mardine Liberte	CanggrySdracQuary Public Pariatio NG	
Community Data Access			Query	when NOV DAPTH(D ()	
Lagery Defensions			Randod Mode	D Comann	
NYSE Endpoints.			Colouiseast Columns	8	
MOR Queries			Columna		
DUARE Queries			Output Maile	n	
Compound Queries			Cortra Kaya	0	
SCRIPTING Queres			Carlie Duration Carlie	2400 True	
NETTLE Queries					
100, 0x810					
SQL Queries					

On preview we can see the chart in the first panel. The chart is distorted. User has to do dimension setting to get it to the right size.

gSales Dashboard		
E has been total array Some Derrichert"		C A D
XYZ Dashbo	ard	in the second
CaregorySeen	1	
· · ·		
Area 1		
	-	
L.	L D	

Create another chart – pie chart for panel_2.



The preview will show both chart. You can do adjustment in the panel size, chart size to get it to aesthetic shape.

ing saves cash occura	
DE une tare large terrer taren Destinant	
Mark Research	i]
XYZ Dashboard	
Caregory Sales	
Lange Contraction of	
here a	
a called patient called called Called Called	
	-
City	
1982 - State	
	6
- 100.00	
- 00.00 	

I will not be adding more components. You can save it and share it to the users. There are many options to make dashboard a truly great dashboard. One has to spend time to design it.

The Pentaho has many other components. There are specialized book on the Pentaho BI tools. The intent of this book is to make readers aware of how a BI system work and give readers a first-hand experience of creating their own small BI system using pentaho. The learning from this chapter can be user to understand more detail of the Pentaho. You can consider this chapter as scratching the surface of Pentaho's capability. Typically any BI system will contain above components. The working process is more or less same but the way system is built is different. The understandings of Pentaho provide you a structure in your mind about the overall architecture of the BI system.

Learning from Chapter

- Overall Architecture of Business Intelligence
- Pentaho and Its components
- Pentaho Data Integration and Reporting using it
- Concept of OLAP cube and how to create one using Mondrian
- Dashboarding using Pentaho dashboard

Section - II

R BASICS

R is an open source statistical computing language with contribution from large number of the developer across the globe. R can be downloaded from https://cran.r-project.org. You have to install the R after downloading it and must have right to install component in your computer. In this book we will be using windows version of the R. Once R is installed open the R, you will see R window with command prompt > in the screen. The screen will typically look like below picture.



Unlike other statistical computing software R is command based rather than GUI based. In this book we will be using R studio which has more advanced GUI than basic R screen. You can download and install R studio as well form internet free of cost.

2.2.1 R Data Type

Before doing any data analysis using R it is important to know the data type, its notation and structure. Like any other programming language R has its own way of storing the data in a structure that helps in easy of input and analysis. The basic data type such an integer, numeric, string, logical, complex and character remains same as any other programming language. The most of the R data analysis will be collection of those basic data type in a set called R-objects. Some of the R-Objects are

Vector: vector is a collection of one or more elements from same data type. For example

A \leq c(1,4,6,7,8,8,9) is example of numeric vector

B <- c("Delhi", "Mumbai", "Kolkata", "Chennai") is example of character vector

C <- c(TRUE, FALSE, TRUE, TRUE) is example of logical vector

Note that a vector wills elements from same data type only.

Factor: A factor is a special case of vector that is solely used for representing nominal variables

For example: Gender <- factor(c("MALE", "FEMALE", "MALE"))

List: A list, is used for storing an ordered set of values. However, unlike a vector that requires all elements to be the same type, a list allows different types of values to be collected

Example: weather <- list("Shimla", 23, "Rainfall", TRUE)

Matrices: A matrix is a two-dimensional rectangular data set. It can be created using a vector input to the matrix function

Example: A=matrix(c(10,20,30,40,50,60), nrow=2, ncol=3, byrow=TRUE) print(M)

[,1] [,2] [,3]

- [1,] 10 20 30
- [2,] 40 50 60

Arrays: A matrices is two dimensional data set whereas array can take n numbers of dimensions.

Example: A <- array(c('India', 'Delhi'), Pincode= c(110011,110012, 110013))

Data Frames: Data frames are tabular data objects. Unlike a matrix in data frame each column can contain different types of data. The first row can be character, second rows can be logical and third row can be numeric.

Example: Candidate <-data.frame(Name = c("Tom","Evan","Roy"), Pass=c(TRUE, TRUE, FALSE), Marks=c(90, 78, 34))

It is highly recommended that readers practice different data type or data objects in R command mode to familiarize with the R system.

2.2.2 R Basic commands

Basic R command can be executed in command mode. Also series of R

commend can be added in file to execute them in sequential style called R script.

```
> X<-5 //assign number 5 to variable X
> Y<-6 //Assign number 6 to variable Y
> X+Y // sum of X and Y
[1] 11
> X*Y // multiplication of X and Y
[1] 30
> X/Y
[1] 0.8333333
> X*5 + Y*3
[1] 43
> Z<-X+Y //Assign sum of X and Y to Z
> Z
[1] 11
> a <- c(1:5) //Create a vector a with elements 1,2,3,4,5
> a
[1] 1 2 3 4 5
> b < -c(2,3,4,5,6) //create a vector b with elements 2, 3, 4, 5, 6
>b
[1] 2 3 4 5 6
> c<-data.frame(a,b) //create a data frame from vector a and b
> c
a b
112
223
334
445
556
> d < -a * b
> d
[1] 2 6 12 20 30
```

> c\$a //print a column from dataframe c
[1] 1 2 3 4 5

> summary(c) //summary statistics of dataframe C b a Min. :1 Min. :2 1st Qu. :2 1st Qu. :3 Median :3 Median :4 Mean :3 Mean :4 3rd Qu. :4 3rd Qu. :5 :5 Max. Max. :6 Creating a matrix > A = matrix(c(2, 3, 5, 10, 7, 9), nrow=2, ncol=3, byrow=TRUE) > A [,1] [,2] [,3] [1,] 2 3 5 [2,] 10 7 9 > A[1,3] //print row=1, col=3 elements from matrix A [1] 5 > A[2,] //print entire row=2[1] 10 7 9 > dimnames(A)=list(c("Car", "Bike"), c("X","Y","Z")) //assigning names to rows and columns >AX Y Z

Car 2 3 5 Bike 10 7 9

2.2.3 Reading and Writing data in the file

R can read and write file from the location of the computer system. When data input is large it is not possible to assign data from the R command mode, hence we have to upload data from the file or connect to the database directly. Though R can work on many file types like tables, CSV and excel; here we will discuss reading and writing file from and to csv file. This is because we will be primarily using csv file in subsequent chapters.



> SLS <- read.csv(file="C:/RData/Rbasic.csv") # read data from csv file > SLS

	Month	Sales	Margin	AOB
1	Jan	300	18	2500
2	Feb	350	15	1570
3	Mar	460	12	2300
4	Apr	230	17	2100
5	May	270	17	2410
6	Jun	600	8	3200
7	Jul	630	7	3250
8	Aug	645	6	3000
9	Sep	560	9	2800
10	Oct	340	11	1950
11	Nov	342	12	1890
12	Dec	350	12	1700
			11.0.0	

> F <- SLS\$Sales*SLS\$Margin/100

> F

[1] 54.00 52.50 55.20 39.10 45.90 48.00 44.10 38.70 50.40 37.40 41.04 [12] 42.00

> FL <-cbind(SLS, F)

> FL

	Month	Sales	Margin	AOB	F
1	Jan	300	18	2500	54.00
2	Feb	350	15	1570	52.50
3	Mar	460	12	2300	55.20
4	Apr	230	17	2100	39.10
5	May	270	17	2410	45.90
6	Jun	600	8	3200	48.00
7	Jul	630	7	3250	44.10
8	Aug	645	6	3000	38.70
9	Sep	560	9	2800	50.40
10	Oct	340	11	1950	37.40
11	Nov	342	12	1890	41.04
12	Dec	350	12	1700	42.00

> write.csv(FL, file="C:/RData/Rbasicw.csv") #write data to csv file

1	Runs In & Cat Cathon - Cathone Runter	Catton B Z 1	riayed +(* E * / E * Fait	- K K - A -				nay hat nga bat wga B. Certar	General Server	- - -	4.5	No.	10 10 10 10 10 10 10 10 10 10 10 10 10 1	3. 1.	Beete Free		Anton - 2 - 2 Later	r A	• • •	
	A5		5			_				-				-						-
27		C.	0			0								9		0			τ.	
5	1.04	100	and a																	
2	2.640	11		11300	42.5															
21	3 Mar	-	b 13	2350																
21	8.644	216	5 17	2100	28.2															
21	5 May	29	5 17	3458	45.9															
7	6.045	60		1200	48															
1	7.24	630	0 7	3,250	46.2															
	8 Aug	64	5 6	1000	34.7															
10	9 Sep	540	5 9	2800	50.4															
11	10 Oct	346	0 11	1998	82.4															
12	11 NOV	343	2 12	1890	45.04															
15	12 046	33	5 13	1700	42															
14																				
15																				
8																				
17																				
18																				
0																				
20																				
1																				
2																				
28																				
٥.												(Dec.)				_		_		-
	and the second s											1.e.		_	_	-	100	1.100 (2)	-	-
	and the second second	-		and in case of	-	_	-	-	-	-	1000	-	_	_	_	_				-

2.2.4 R Database Connection

R connects to database through ODBC interface in windows. Here I am going to show how to connect MySQL database. Go to **control panel.**

		All Control Panel Ite	716	- 0
- T 😝 - Control Fand + A	E Control Panel Namo			w 🚯 Search Correct Panel
Bust your computer's settings				New By: Lege turn: *
Action Center	Add features to Windows	Administrative Teols	AutoPlay	ংশ্ব্ Bitlacker Drive Encryption
Color Management	Credential Manager	Dute and Time	Default Programs	Del Audio
Del Command Power Manager	Dell Command Update	J Device Manager	Big Devices and Printers	Tispiay Display
Fase of Access Center	Re History	Flash Player (32-64)	E felder Options	A Fores
3 HomeGroup	De Indexing Options	invel(R) HD Graphics	Intel # Rapid Storage Technology	C Internet Options
S. Java	Kryboard	🛒 Language	🔯 Location Settings	(3) Mail (12-64)
7 Mouse	Metwork and Sharing Center	Notification Area loons	S Penonalization	A Phone and Modern
Power Options	R Programs and Features	Recovery	🔗 Region	Remonskop and Desinop Connections
9 Sound	👌 Speech Recognition	Storage Spaces	Sync Center	1 System
Taskbar and Navigation	Troubleshooting	Stor Accounts	Mindows Defender	Windows Firewall
and the second of the second of the	17% Windows Undate	Work Folders		

Select ODBC data source

n nes Se	n Var		Admin	strative Tools		- 0
Casy Parts & Cat	an Andrew Carlos Andrew	Chan ben -	Den Cart	Statut at Statut name Statut satutation Tatut		
0 0 - 1 0-	Control Panel + All Control Panel Name + Admin	stative Task				 Sanch Admonstrative Type
tereiter	Parts .	Data modified	hor .	Ser		
Desiting Description	Component Services Secondar Management No. Defense and Contents Drives	6-02-0210 12-01-044 8-02-0210 12-04-044 8-02-0210 12-04-044	Dontext Shortcut	2×8 2×8 2×8		
M The PC	2 Dia Dang 2 Dia Teang	800000 001.04 800000 001.04	Shonouk Shonouk	248 248		
2 Decuments 2 Decuments	A Good Security Pality South Control Security Pality Structure (32-54)	8/20/2012 COT MAR 8/20/2012 SOLEMAR	Shortual Shortual	148 148 148		
D Mase E Februar	 CORE Data Sources (M-MR) References Menter References Menter 	800-003-0-03944 800-003-0-02944 800-003-0-04444	Shortcut Shortcut Shortcut	248 248 248		
Ca Datafait (D)	S Asserve Manter & Serves	6/20/2013 (0/22 PM 8/20/2013 12/24 PM	Shonout Shonout	248 248		
S Never	System Congration System Information Of Task Scheduler	\$10,000 0,00 M	Donest Donest	248		
	If Hindows Reseal with Advanced Desurby In Hindows Memory Daparetic If Hindows PercerStat 1000	8/00/00/0 (0/0 MM) 8/00/00/0 (0/0 MM) 8/00/00/0 (0/0 MM)	Shancut Shancut Dhancut	348 548 548		
	(e) Mindows Provided 51 (48) (e) Mindows Provided 51	#10-01-025.9M	Shortcut Shortcut	248		
2 ³ during						(E) 4
a 🎮 🖸) 🖄 🍐 🖄 😺		0 1	1 0	9	- P G d C Mi Maile

Click on Driver->Add new Data Source

00 00 5 m	Manage	-	0.00			-		
New Care Care		ODEC Data	Source Admini	trator (64-0-0)				
and the life of	iter 0.05 June	DON No DON Drives	Twing Covedo	Pairs And				
Departure of the local data	COSC Drives the	an installed or your syste						
orbei Pand + All Control Pr	Same	Version	Censory	Te .	0ee			✓ Ø Stan(h Administrative Typ)
Parts	WysioLooec 1	110mm 111.000 4.01 MIL 111	Oracle Corporation 1. Horsey Corporation	MINDOWCH DU	1. 11/01/001			
Parate Norre E belfstep (b) Compared Services, (b) Dontifieds (b) Compared Managerer (b) Dontified (b) Dontified (b) Dontified (c) Dontified (b) Dontified (b) Dontified (c) Dontified (b)						~		
Gold Constantial Parloy Collect Constantial Parloy Collect Constantial Collect Constantiation Software Structures (E) Performance Monitor Sign Priorit Management	a) 23	WC driver places 100 Krans Frank, une fine driver's antig	dini yayana is pita yayan	fornation Trans (108	C data anyone	Tread		
S Resource Maniter			04	Grod	Auto	140		
Suten Cardipastien		8-20-2010 (3-20-PM	Inorinal	248				
55 System Information		NO 00 000 M	Dotted	248				
D Tall Scheduler		\$20,0015 \$2,00 PM	Dotod	-248				
If Hindows Frewall with Ad	anced Security	8-20-2013 (0.15 MM	Shorted	240				
St Kindeus Menery Dagne	-	a colorista de Par	Detail	148				
18 Hindows Promittell 54 1	-	8/02/2010 12/25 PM	Profest	218				
and the second sec		ACCOUNT OUT IN	Shuthing	248				
	Annual Construction of the second sec	All Control Particle All Control Part	COPIC Data Copy Data	Complex Service A demission Complex Service Complex Service A demission Complex Service Complex S	Composition Status Composition Co	Complex Strategy Comp	CODEC Data Source Administrator (Sd-Data) Compare Name Compar	Comport Name C

Select MySQL driver. If you do not have driver please download it from internet.

A tor and the second	* 5	ODBC Data Source Administrator (64-bit)		AaB sation sation . A	A Mine-
Captored to Fait	- tarb	Create New Data Source III.III Select a draw for which you work to set up a data source.			nir g bear s teng
		Name XMMA Company MyXXX,004C131 Dear 151:3031 Onter D 305 Server 613 MOS 11/16 Microsoft	-	-	
		c ,	-	-	
	-	- Seck Read			
		Of Great Auto	140		
	ini)				
				B.041	
A					10 CO 11 CO

Enter database IP, database name, userid and password. Also name the connection.

The Part Day	Var Hange		Averative MySQL Connector/ODIIC Data Source Configuration	- 0
Case Party (Case and	Hand Capy Dede Reserve	Neder City	Connector/ODEC	
(Sphered	Departor	-	Consultar Parameters	
(e) = + + + + + + +	<pre>xPC + OS(C) + Population + G</pre>	territer	Data Source Name: MSQ.	v 6 tech0endut
in favorites	Ante		Bergten	
E Delitap	👗 Languages	One 504 [1	8 1010 Series 10.71 (27.8 Avs. 2014	
A Dearlines	Control of	Gerbes 6	O tanet/type	
	Gentatesanta	Nate -	Not not	
IN THE PC	St Grenchethuge.ett	Don fire	Proved Manager	
P Administrator (sto-a	E setate	9/5 Accies		
2 Decements	N lopher.dl		Designed (and the second secon	
A Descripto	in legiture		Anna an	
Maria Maria	andre .		Canadra Constant Constant	
3 Videos	animal00	4	1.2	
Colica Coloradation		0	CDC claim data survey divers information about two to conversit to the indicated data provide. A claim so only reduce to go, and can only be used on the computer	
			OK Groat Anny Mag	
				12 m
Contract of the second	m			
			S S S S S S	A REAL PROPERTY AND ADDRESS

To connect to MySQL Database

> con <- dbConnect(MySQL(), host="XX.XX.XX.XX", port= 3306, user="username", password = "password", dbname="DB Name")

To get data from database to R through MySQL Query

> MCal <- dbGetQuery(con, "select * from Master_Calender;")</pre>

> head(MCal, nrow=10) //Print first 10 rows od data

Date Year FY_Year Month Month_Name FYMonth YearMonth FY_Month Quarter Week Day Weekday

1 2014-11-01 2014 2014_15 11 11_Nov 08_Nov 2014_11 2014_15_11 Q3 44 1 Saturday

2 2014-11-02 2014 2014_15 11 11_Nov 08_Nov 2014_11 2014_15_11 Q3 45 2 Sunday

3 2014-11-03 2014 2014_15 11 11_Nov 08_Nov 2014_11 2014_15_11 Q3 45 3 Monday

4 2014-11-04 2014 2014_15 11 11_Nov 08_Nov 2014_11 2014_15_11 Q3 45 4 Tuesday

5 2014-11-05 2014 2014_15 11 11_Nov 08_Nov 2014_11 2014_15_11 Q3 45 5 Wednesday

6 2014-11-06 2014 2014_15 11 11_Nov 08_Nov 2014_11 2014_15_11 Q3 45 6 Thursday

- MTD YTD Day_Passed
- 1 NA YTD 19107
- 2 NA YTD 19106
- 3 NA YTD 19105
- 4 NA YTD 19104
- 5 NA YTD 19103
- 6 NA YTD 19102

Further Queries that can be used To list the tables of database >dbListTables(con)

To write data into MySQL Table from data frame >dbWriteTable(con, name='table_name', value=data.frame.name)

2.2.5 Graph Plot in R

> bar<-c(100,200, 250, 150, 120, 75,90)

> barplot(bar, main="Sales", xlab="Monthwise Sales")



Add month in the bars

> barplot(bar, main="Sales",

+xlab="Monthwise Sales", names.arg=c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul"))



Horizontal Bar Graph for the same data

> barplot(bar, main="Sales",

+xlab="Monthwise Sales", horiz = TRUE, names.arg=c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul"))



Pie Graph

Pie chart of the Categorywise sales

Pie <- c(50, 60, 45, 80, 90)

```
> Cat <- c("Mobile","Book","Apparels","Toys","Automobiles")</pre>
```

> pie(Pie, labels = Cat, main="Categorywise Sales")



- > Percent <- round(Pie/sum(Pie)*100)</pre>
- > Cat <- paste(Cat, Percent)</pre>
- > Cat <- paste(Cat,"%",sep="")</pre>

> pie(Pie,labels = Cat, col=rainbow(length(Cat)),main="Pie Chart of categories")



Line chart

> Mobile <- c(100, 120, 80, 55, 90)

> plot(Mobile)



> plot(Mobile, type="o", col="blue")



> Fashion <- c(30, 56, 90, 75, 100)
> lines(Fashion, type="o", pch=22, lty=2, col="red")



Scatter Plot

> SC <- data.frame(s=c(24, 23, 34,22, 24, 65,46), t=c(4, 3, 2, 9, 5, 8, 12)) > SC

S

t

- 1 24 4
- 2 23 3
- 3 34 2
- 4 22 9



> plot(SC, main="Scatterplot Example", + xlab="Sales", ylab="Time", pch=19)



Function in R

User defined function helps in creating commands that are not readily available. Below is an example of how to create a function

> SQ <- function(x, y)</pre>

+ {z <- x + y + return(z)} > SQ(4,5) [1] 9

Chapter - III WEB ANALYTICS

"Wisdom begins in wonder"

By Socrates

In previous chapter we learnt R and Pentaho BI tool. In this chapter we will learn Web Analytics tool called Google Analytics. Web analytics is a general term used for collection, processing and reporting of the website activities. The desire to know the website behavior led to the development of the application such as Google Analytics and Adobe Omniture. In the early phase the website data collection and processing the companies used to retrieve website data from the webserver logfiles. The logfile analysis has been increasingly being replaced with webpage tagging due to various advantages tagging has over the logfile. The tagging technology such as Google Tag Manager (GTM) and Dynamic Tag Manager (DTM) are extensively used. The tagging makes the implementation of the web analytics lighter, faster and easier.

There are many web analytics tools in the market with different process of capturing and reporting the data. The capabilities of the tools also differ. For this book we will be using Google Analytics as the primary web analytics tools. The Google Analytics is free Web Analytics tools provided by Google. It has premium version as well but we will be working on free Google Analytics. In the chapter we will understand the basic features of Google Analytics. We will assume that the developer has added the tags in the required pages of locations the site. The actual tagging of the pages is out of scope of this book but users interested can develop their site and learn how to tag the site. There are books and online resources to help you out with your experiments on tagging.

One important point to be noted by the reader is that Web Analytics tools will not capture 100% of the data because of the various reasons. The accuracy will depend on the website design and the implementation. The good implementation would typically capture more than 90% of the traffic and conversion data. There might be apprehension among the user how to make decision without capturing all the visitors and transactions. However keep it in

mind the analytics is all about the sampling and trends. You may consider web analytics data as random large sample. The trends of the traffic, conversion, goals, and events are to be used for decision making. You may spend innumerable amount of time and effort to capture 100% data but the cost of the achieving 100% would likely to be higher than gain from addition 5% to 10% of data. The lost opportunity from the decision not taken from the sample data would also been higher than gain from more accurate data.

The primary motive of learning Google analytics is to provide the readers with web analytics tools for analyzing the customer behavior data in the website and help users track the marketing channel performance which we will learn in Digital Marketing chapter. We analytics system is not beginning to be central system for digital marketing for better segmentation and targeting. For instance Adobe Media Optimizer use Adobe Omniture data for segmentation and targeting; similarly Adwords uses Google Analytics data for creating audience for targeting.

Section - I

GOOGLE ANALYTICS

3.1.1 Metrics and Dimension

In the section we will assume that the Google Analytics has been implemented in the site. The focus will be more on the reporting side and the way to interpret the reports in the Google Analytics. Most of the data in the Google Analytics are self-explanatory but to get the meaning and how the data is defined readers can go to Google Analytics Support site to get detail of any dimensions and Metrics. The dimensions are the categorical data and the metrics are the facts being captured (https://developers.google.com/analytics/devguides/reporting/core/dimsmets). In fact this page will again be useful for the API integration with Google Analytics data which will be discussed later.

3 Googe Developers	Q Anaptica Cara Reporting		
Analytics Core F	Reporting API		
Janualisrigit Quickshart	Dimensions & Me	etrics Explorer	
Fundamentals Rendermentals	The dimensions and metrics explore Core Reporting API. Use this referen	er lists and describes all the dimensions and metrics available through the Ket to:	
Sevenanes & Metters (Splane) Septembe	Explore all the dimensions and met quickly find the name you're looking	ries - Click the plus box to see dimensions and metrics by feature. Search to for.	
Segments Overview Using Segments	Valid Combinations – Not all doman metrics can be used together to one other values that can be combined in	stons and metrics can be queried together. Only certain dimensions and are valid combinations. Select a dimension or metric checkbox to see all the in the same query.	
Authorization Baltifung Ferformance Ton	Modes - Switch modes to organize names.	and view dimensions and metrics in different ways based on AP1 and UI	
Common Queries	Muse API Names	Taget	
) Google Developers	Q Analytica Care Naparting . 🖗) Bank	BON R
Google Developers	Q Anaples Care Aspertage Q) Seeh	BOA N
Google Developers	Q Anapos Care Aspertag. Q	() Search	BOR R
Coogle Developers	Q Anapos Care Asympton Q	Branch Branch Drothow allowed in segmentes Directure allowed for segmentes Directure depresented	BON R
Coogle Developers COS APPENDENCE ON Annatorial Californian Annatorial Californian Annatorial Californian Annatorial Californian Colored California	Q (Anaposa Cara Auguring () arrost	Branch Branch Drifty show allowed in segmentes Drifty show allowed in segmentes	BOA A
Cocogle Developers COC APPENDENC ON Annabergin Chaldwater main Appe Annabergin Chaldwater Annabergin Chaldwat	C Anapora Cara Auguring . C second Main Names C Spand al C Spand al C Spand al Disconduras	Brench Brench Drifty show allowed in segmentes Drifty show allowed in segmentes Drifty show allowed in Methods	BGA B
Cocogle Devetopers CC REPERDENT R Annaboling Eduktori mit Ages Anderentials Restang Edit Die APA Bernantik Segmentik Segmentik Segmentik Segmentik	C Anapose Care Awarding . C	Search	. Bion a
Cocogle Devetopers COC MOTORNO M Anatoria Calebratan min Agos Anatoria Calebratan min Agos Semananos L. Anatoria Segmenta Segmenta Calebraice Segmenta Segme	C Analytics Care Howevilley . C	Search	SGA P
Coccele Developers April 2010 April 20	Andered Care Anywring (Their Annes Capend all Capend all Deser Deser parameters parameters parameters parameters parameters	Branch Describ Describ Descr	500 m
Coccele Deverlopers Koccele Deverlopers Application Application Application Application Separatio Sep	Anapora Care Auguring .	Brench Dresch Dresch	500 m
Society Development Montaneous Annonesis Annon	C Anapose Care Awarding . C	Search Drath Only show allowed in segments Include depresated Metrics Include depresated Metrics Include depresated Include depresated Include depresated Include depresated Include depresated Include depresated	500+ H
Coccele Developers Acceleration A	Audote Care Auswing (Second all Order Deser Second all Deser Second all Second al	Branch Describ Describ Describ Describ Describ Describ Describ Describ Describ Describ Describ Describ Describ Describ Describ Describ Describ Describ Describ Des	. 500 m
Society Developers ACC REPORT ACC REPORT	Q (Added Card Aquering . ()		500 m
Societ Developeration April 2000 (April 200	August Constanting Market Market Market Market Caped al Caped al		500 m
Coccele Developers Annovember Annovember Annovember Coccele Cocce	Andread Care Housening (Second all Other Deserved all Deserved all Deserved all Deserved all or operational careformation or operational careformation proceeding careformation proceding careformation proceeding careformation proceding c		. 500 m
Coccele Developers Control Coccele Coccele Control Coccele Cocc	August Cara Augusting . Marries Al Names O Visames Coperd al O User Deservicita prover Type or Support Caral Streams	Branch Drawch	500 m
Coccele Development Montania Montania	Contract Construction Construct Cons	Search Drawn Orly show allowed in segments Product regressed Product regressed Op a series Op a serie	500 m

Google Developer				BOX IN
CUDES REFERENCE	SUPPORT			
Januforigi Gulokatari Keli Appi Pundamenlah Kening with the API Dimensioni & Motors Explore	garner Ul Nam The nu Anvibu	wUsers le: New Users mber of users whose sessi tes:	on was marked as a first time session.	
Segments Segments Overview	Outs T	Type	Allowed in Segments	
Using Segments	INTEG	CR	Tes	
Resources Authorization Batching Performance Tigs Common Queries Client Libraries Limits and Queries	gache Status Uas ga	wVicilio Deprecated		
Damos & Taala Guary Robinar (2) GA Spreadsheet Addion (2)	ga:1d	ayUsers we 1 Day Active User		

The Google Analytics has certain inbuilt reports but users can create its own dashboards using dashboard builders. We will go through main tabs of Google Analytics to understand the data available for analysis. Most of the dimensions will be explained through the tab. Before going into the Google Analytics screens let us understand the key metrics that will be used across the web analytics domain.

Metrics	Basic Definition
Sessions	Sessions are defined as a period of consecutive activity by the same user. By default, in Google Analytics, a session persists until a user stops interacting with the site for 30 minutes. We call this the session timeout length
Users	The metric called "visitors" or "users" measures the number of unique users that visit your site during a certain time period. This metric is most commonly used to understand the overall size of your audience.
New Users	The users which come to the site for the first time in the period are the new user for the period.
Bounces	The bounces are the session with only one page. The user closed the browsers in the first pages itself.
Bounce Rate	Bounces rate is the number of sessions with bounces divided by number of sessions.
	The average duration is the time user spends on the site.

Avg. Session Duration	
Pageviews	The page views are the number of pages visited by the users in the session.
Pages / Session	The pages viewed divided by the number of session in the period is the Pages / Session.
Unique Pageviews	The number of unique page views. Page views within different sessions count as separate unique page views. This takes into account both the pagePath and pageTitle to determine uniqueness
Exits	The number of exits from those pages.
% Exit	The percentage of exits from your property that occurred out of the total page views.
Transactions	The number of transaction in the period
Ecommerce Conversion Rate	The conversion rate is calculated as number of transactions divided by the number of sessions.
Revenue	The total sale revenue provided in the transaction excluding shipping and tax.
Quantity	The total number of items purchased. For example, if users purchase 2 Frisbees and 5 tennis balls, 7 items have been purchased.

The login screen of the Google Analytics will look like below picture

of heards where a cost	Audience Overvi	ew			11 Jan 2016 - 10 Feb 2016
Daltarth	Inst Equal + Addis	Instituted Named			•
a Shittab	Al Services		O - 40	t Segment	
and the second	Overview				
Intelligence Counts	Second + 18 (second	1 Partie			North Bar Mark Mark
Red Tree	· factors				
Autoria	100				
Overslew					++++++
Overslew Active Users			~~~~		
Oversleer Active Union Cohort Analysis ^{wite}		126	224	7.00	174
Overslew Active Users Calcul Analysis ¹⁰⁷ • Demographics • Interests		ine .	2 international and the second	810	154 New Yorks # Reservey Yorks
Overslee Active Users Colout Analysis ¹⁰¹¹ + Demographics + Heavity + Geo		um 93,631	Frage Vision 1,193,705	Bite	174 New Yorks # Reserve Yorks
Overslee Active Users Coloris Analysis ¹⁰¹⁰ • Decographics • Decographics • Decographics • Decographics • Decographics • Decographics • Decographics • Decographics	5mmm 116,466	um 93,631	27 Am Fage Views 1,193,705	Bite	170 • Tere Yilder • Financing Yular

The date range selection can be done at the top right. You have inbuilt customized ranges like last month, last 7 days, yesterday and so on or there is option to selected data from the calendar.

S. Seath news \$140	Audience Overs	iew.			11 Jan 2016 - 10 Feb 2016 -
	Enal Equat + Adda	Partners Partner	*17111		Title Datase
• Shrinin	O Al Bessiers		******		
• Intelligence Events	Courses				
Ø Red Time	· Sestina	\$11£28			Contraction of the Contraction
🔹 Autorea					
Overview Active Users				*****	
Cohort Analysis With		rt.sei	10 Jan	21.00	1fe
+ interests	Services	Users	Face Views		Raw Victor Bataning Victor
+ Geo	116,466	93,631	1,193,705		
+ Technology					D. ST Record (D. Pr)

You can compare with two time range by selecting compare and select the comparisons range. By default the compare range is same number of previous period from selected period but you can override by selecting you own range.

C. Santon march & help	Audience Overv	iew		1	11 Jan 2016 - 10 Feb 2016 -
E Centeren	Beal Equal + All's	Darlinet Postal	*******		000 Tanja (2017/001 - 1) (4 740 2018 - 10 740 2018
era Dation	O Al Sessions		*******		(In section () + () the street ()
	Overview			**********	Anny Secon
• Intelligence Events	Second + 15 Sales				Norty Day must Munit
Ø Real Time	• Sessions				6
Aufonce Overvlaw Active Users			~		
Cohot Analysis ***		1.1at	12 im	28 i m	17m
+ Denspaphics + Internis + Oee + Bahavioar	116,466	Uses 93,631	Page Views 1,193,705	• • • • • •	ter # Recentling Valuer
+ Tacheology + Multila	Paper Section	Arg. Session Duration	Bource Rate		

In real time Tab you can current traffic, location, sources of traffic, which pages the traffic is and so on. The real time traffic is very useful for monitoring traffic after doing changes in marketing campaign, activation of ATL/BTL campaign, for monitoring site issues and estimating the sales etc.

C, bench reports 6 hole	Right r	now	Page Vews Per minute	Per second
Costboarth	15	5		
Intelligence Events	45% (199	or site	mar it.	1
Ø Red Time	SAUGRADANA UNITE		10	
Overview			Vering May Cath	
Locations	lines Tatar 55	e.	and the second s	
Traffic Sources	Country	Active Union &		
Contract	1. Indu	10 05.57%		
Events	2. Wyannar (burna)	1 647%		
Conversions	3. Palazba	1 6.67%		
	4 Bault Astern	1 6476		
Audence Audence	5. United Arab Erstated	1 647%	X Y Y Y	
Overview	& Lines Dates	1 647%		
Active Users				
Cahort Analysis 1078				

The Audience Section provides general information about traffic. The overview tab provided consolidated sessions, users, page views, bounces rate and new or old users' information. You can view trend report at hour, day, and month or year level for some of the metrics.

Autorea					Kary
Overview Active Users	2.584				
Colloct Analysis Wite	-	the 4	tua tutua	the .	5740 VE74
+ See	Session 24,284	20,899	Page Verst 220,701	•	es Victor · Returning Victor
+ Tachnology + Mubile + Castian	Pages/Dessian 9.09	Arg Sessier Duration 00:02:50	28.66%		
+ Declorating	% New Dessions				

The Active Users Tab provides number of active users in 1 day, 7 days, 14 days and 30 days.

a an in the state	· · · · · · · · · · · · · · · · · · ·						
C, Deput space & log	Active Users						
Autorea	2 10ay Active Users	T Day Active Users	16 Day Active Users	30-Exy Active Uses			
Overview	1.00						
Active Users							
Cabot Analysis 818							8
+ Demographics							/
+ Indexedta	100					/	
+ Oes		2211					
+ Exhaviour						and the second	
* Technology							
+ Mutola	-	144	114	14-2	16-2	164	
	1000000000	The ba	a Then	Man Artistica	area.	Miles Artice Chart	
+ Custum		- A Martin Party Party	Carl Contract of Carl C	TRACE AND ADDRESS COMPANY		- Structure - County	
+ Custom + Decuberanting	3 505	24 10	2	40 270		92 261	
+ Contain + Decolemation Usan Floor	3,505	24,10	2 Harry Carrier	40,279		92,261	
+ Castan + Deschmaking Uses Flow	3,505 Surface the area (2,00)	24,10. Surface	2 Mark (AND)	40,279 % of hear following (st.)**	74.4	92,261 5 of the 101,001,00,200	Ratherst Regard

3.1.2 Cohort

The cohort analysis shows the repeat visitor to the site. The cohort tells that in a day if 100% users were there for the metrics selected, then in subsequent days how many of them come back or completed those events. We can create cohort for different time size – day, week or month, select metrics and select data ranges.

G, Skett name & hep	All Sessors + And Segment		
	Colort Analysis		
Anderes Overview	Colloca Type Colloca Size Manay Dana Range Anganatan Saw - Ne der - Saw Transition - Saw Transition -		
Active Users Cohort Analysis ¹⁰¹⁶	Acquisition Date onlocks by User Retention 1 setund -		
+ Demographics	e at because		
+ Geo			
+ Behaviour + Tacheslage			
+ Matala	B HS		
+ Custom			
Users Filme			
Acquisition	6mit 6mit 6mit 6mit	Day N	4417

The cohort can be viewed in triangular table as below. This view provides period-wise trends and is good for comparisons.

R, basel agent to key											
Active Users											
Cohort Analysis ***	a bart	fag 1	842	In I	_	Bay 4	da.	\$	Day 6	- tay	
+ Demographics + Interests		0 m 1	Gar 3	Sec 2	941	Day 4	Der 6	Der 6	1 art	Do 4	ş
+ Gent	All Sessions	100.00%	2.81%	0.99%	0.55%	0.74%	0,45%	0.00%			
+ Behaviour	4 Feb 2016	10.075	110	4795	10%	1.0%	145	115			
+ Mulda	5 Feb 2016	10.075	2476	0.07%	110	110	1.01%				
+ Gustam	6 Fut 2016	10.075	165	1.855	195	1.075					
+ panonianing	7 Fab 2014	10.05	188	1.015	110						
Users Fine	8 Feb 2016	10.075	2200	1.055							
🔹 Acashitan	5 Feb 2016	10.075	1.105								
-	10 Feb 2015	110									

3.1.3 Major Tabs

The demographics tab provides information about age and gender distribution of the visitors.



The interest Tab provides information about the visitor's life style, affinity to particular thinks etc.

	* Key Metric Dessare *			
C. Depty reputs & terp	Affinity Category (reach)	91 S2% of Total receivers	In Market Segment	57.87% of Total seconds
Autors Overview Active Over Other Analysis ¹⁰⁴⁴ - Strengthis - Interests Overview Attrity Orlapties builded fearmer	2 195 5 795 4 095 4 095 4 095 3 095 3 095 3 095 3 095 3 095 3 095	Morie Lowes South Aster Film Fans. Nano Juvisos & Antif Banders Extendement & Calabitry Neno Juvisos Struteyos Struteyos Struteyos Sangers (Sophaltics Taront Buffs Cashing Billwaters Aspeng Cheft Social Morie Enthesiants Fachagitas Bandy Marves	4 20% 3 42% 3 42% 2 42% 2 42% 2 42% 2 42% 2 21% 2 21% 2 21% 2 21%	Englisymeit Agrand & Accessions Terrethritekti & Accessions Benediter/Networks Agrand & Accessions/Terret's Agrand Consume Exclusive Molder Promes Translikh Transi Translikh Transi Beardy Proberts & Senten Education/Terr Properties & Nationg Oth & Consultany/Wedding Planning
Other Categories - Geo - Barbaniser - Tacheology - Mobile - Coaltern - Sanctimating	Other Category 742% 442% 445% 299% 247%	E1.59% of Third Honore Sectors 2 Ann. & Extendioment Global South Allan Pan. Ann. & Extendioment Guideltine & Extendioment Rous Ann. & Extendioment V & 5 Volao Colline Villes Fund & Dosc Costing & Recipion Collines South Aslan Constantindo Coloris		

The Geography tab provides information about the visitor's location – country, city and continent.

11, Search reports & harp		Acquisition			Belantest			Constance Br	ineranta en
🛓 Audorca	CNy	Senters 4	5. Rea Joseffern	Res (See	Boarce Rate	Pages Sension	Arg. Second Decetion	Terration .	Research
Overview Active Users		24,284 5,4764 (0.0%)	73.81%	17,923	28.66% http://www. 24.40% 0.40%	9,09 http://www.ibidi //tightig	00.02.50	15 15	Rs.226,451.00
Called Analysis with	1. New Delhi	5,040 (01474)	35.35%	2240 (21474)	36.71%	6.12	80 12 26	205396	Rs 2701.00 ct enco
Demographics	2 Manhai	GABE INTER-	11.05	tesi ama	24.87%	16 33	80 83 20	40000	Ro. 57, 678, 20 years and
- Geo	3 Sergelars	1.40 (1.1%)	72.74%	1854 (1996)	25.25%	8.16	01245	# 100m	Pa.8.00 (1995)
Language	4 (M140)	1274 minu	62.17%	1.051 (1993)	42.87%	5.17	10 11 47	4	Pa.8.00 (0.011)
Location	S. Notura	85 0.0%	0.6%	\$6 anni	24545	8.0	80 03 35	4.00%	R. 8.35 (1975)
+ Deheslow	6 Para	NZ (SAM)	73.60%	543 c.mm	27.30%	147	00.02.45	\$ 1100s	Rs 6.00 (2.17%)
+ Technology	7 Hybridian	567 (2.17%)	72 66%	412 (2.076)	26.67%	10.36	0112.54	210.000	Ro. 81.525.00 ct. error
* Materie	1 Putney	546 (2,20%)	84.25%	40 0100	44.69%	4.27	86.01.25	8 (1994)	Padda units
+ Custom	9 Chernel	508 (2494)	76.97%	391 dama	25.79%	8.61	10.11.12	8,000	R. 5.18 (1975)
+ Banchmarking	ND Dubai	474 (1.004)	75.32%	30 000	29.75%	1.0	00.12.25	8.00m	Pa.2.12 (1111)

The technology tab provides information about the Operating System & Browsers from which users are visiting the site and the network also.

C, Sarris sports & help			1745	£7.6		7740		174		1746	12748
	. Pos	nary Denersian	Browned Operating System	n : Sonn familian S	unan Coloren	Fast viewer, CO	No				
Audience			onting dimension + . Sort	tor brief +					٩	atomat [8]	0 2 2 1
Overview Active Users				Acquisition			Behaviour			-	* 10000
Cohot Analysis **** + Demographics		Browner		Section 4	N. New Internet	Res (term	Brunce Rate	Pages Season	Aug. Session Densities	Summerious	Annese -
+ bitareste + Cese				24,284 5 of Total 100 Min. (14,04)	73.81%	17,923	28.66% Ang to Very It are parts	9.09 Ang ter Harry 1.00 (0.00%)	00:02:50 2010 000	15 Norther Markets	Rs 226,451.0
• Deharitor	0	1. Onne	i.	12,240 (0.471)	75.8%	9,200 (m. 20%)	24.80%	9.25	00 03 00	6(0.01)	Pa 73.328-04 (11.1
Browner & 05	0	2 Salat		4340 (0.010)	6166	3.119 (0.49%)	21.52%	12.00	90 53 37	8101254	Rp. 101.774-00 (11.1
Network	0	3. UC 844		t.Mt (Long	83.52%	156 ans	46.01%	3.60	60 01 21	8 (1974)	Ra 8-00 -0-0
+ Multile	0	4 Opera k	Rei	4.505 (1.500)	95 KPN	1.115	14 12%	+ 79		8 (com)	B. 8.00
+ Custom	0	S. Android	Denter	1,019 (1,211)	75.27%	32.000	62.10%	4.50	80 01 31	B. colored	Pa 8 10 - 11
+ Banchmarking	0	& Finites		993 (A.M.)	75 36%	794 (1475)	17 22%	10.82	68 63 32	8 0.000	Pa 8 48 (11
Quen Fine	0	7. Salari i	in-aqui	915 (1.7%)	46.52%	605 ci.smi	23.01%	7.65	00 12 15	# 10.00%	Pa \$ 10 - 11
	0	8. inuted	Englane	SNE salama	77.47%	\$21 anno	17.21%	12.67	80 53 37	1 04250	Ro 21 248 00 11-1

The Mobile tab provides information about the device category from which users has visited the site – computer, mobile or a tablet. The particular devices model can be identified for the mobile users.

		Sessions									
C, Dearch squark & Nala		6.000									
			_	-					_	-	
Audience		1.54									
Overview											
Active Users		5746	-	1.0	viii i	7.64		8748		3746	164
Calut Analysis											
* Demographics	1	ray Several Dence Category	21277	Carlos and				-		Territoria de la constitución de la	
+ Interests		Secondry developed *	Sart 1, pa	Darrent .a.					đ	staring [B]	0.5.5.11
+ Geo				Acquisition			Behaviour			Conversions E-	ormen *
+ Behaviour		Benice Category		tenne 4	1. Sea	Real Dans	Bearing Sala	Pages Section	Aug. Senator	forestore .	Annual .
+ Making											
Overview				24,284	73.81%	17,923	28.66%	9.09	00:02:50	15	Rs.226,451.
Decices				100.01% (24.294)	75.87% (9.19%)	100.00% (17,004)	20.00% 10.00%	(1995)	88-42 M	102.00% (15)	(81.25.5)
+ Custom	0	1. mobile		54,369 cm/ma	73.80%	10.504 (m. mm)	34.27%	7.04	95 42 48	# (10.0%)	Rs.142.972.02 (c)
+ Sanchmanking	0	2 decidage		6.625 paseta	78.21%	6,116 (01/076)	28.29%	11.53	00.03.48	2 (11.30%)	\$121.365.00 m
Uses Flow	0	3 table		1.890 (0.75%)	63.65%	1,200 (0.11%)	2196%	54.30	40.04.05	\$-03.0%	Ph. 62 108 49 cm
Sound much a tem		an an and the second seco	See Free	Defeat -				-	9.	shared (II)	0 5 3 5
, South reports & tong		an an anna Mara anna an a'	Burt Type	Eviluat =					٩	ational 📳	0 E 3 m
, Suarch reports & tang		na politik Blan post na -	Surf Type	(ulul +			Belaviour		٩	atomoti 🗐 Conveniens Rom	0 E % F
Search reports & traig	•	National Second Se	But Fype	(stat +		New York	Balances Refe	Pages Section	Q.	atorial II	0 2 2 F
, Scarth republic time Audience Orandiar		an an <mark>1900 Ma</mark> na aman an ' Ta t Security (travelar 1 Malah Jawa Ma	Set Fype	Default = Acquarters	<u></u>		Salanina Basesa Rata	Pagas Sensition	A Ang banta Sector	attanced III Convenient Rev Parameters	0 C 2 m
Audence Crundiar Active Users		na ya afala Barana ana ana ana ana ana ana ana ana an	Set Type	(vint + Augustins Instant + 16.237	<u></u> 72.66%	Mare theme 11,798	References Bosonies Radia 322, 78%		9. Ang lemite Denter 00.02.21	atorial II Convenies For Networks	0 2 2 m
Charth reports & help Audience Cranklee Active Users Colort Analysis ¹⁰		na ya afala ka ana ana ana ana ana ana ana ana ana	Set Type	(read + Arquation (readown + 16,237 (readown +	12.00%	ter then 11,798	Seturious Bosines Rates 32,78%	Pagas Inness 7,88 Ang In 100	9. 49. 10000 10002-21	attance () Convenient for Neuralism School	0 E 2 F
Audence Audence Coantier Active Users Colont Analysis *Demographics *Demographics		National States and any of the second s	See Type	(print + konnention 16.237 16.237 1.275 2.558 120 marrier	2.66% 172.66%	New York 11,798 11,798 11,798 11,798 11,798 11,718	Between the Baserse Turks 32,78% 32,78% 32,0000 32,0000 32,0000 32,0000 32,0000 32,0000 32,0000 32,0000 32,0000 32,0000 32,000000000000000000000000000000000000	Pagas Sasana 7.858 Ang No (10.10) (10.10) (10.10)	Ang. Senter Dester 00.02-21 Ang Senter Sente	abaroot ()) Commission for Networks 13 Notices National National	C C C C C C C C C C C C C
Summer regioners & transport Constrainers Active Usana Colourt Analysis ¹⁹⁷⁰ • Demographics • Stemans		Mathe Server of	See Type (B)	(Mark +) Augustus 16,237 16,237 16,237 3,279 3,	S Bar Minister 72,665% South Yes States Cases Cases	New Own 11,798 Nother 117200 2114 or Are 2114 or Are	Belance Benco Res 32,78% (0.3%) 25.3% 55.2%	7,88 7,88 8494 (1011) 1011 1011 1011 1011	44 batter 144 batter 100.02211 141 b the 111 bit 111	atorical III Consentian For Nametrica 13 13 13 13 13 13 13 13 13 13 13 13 13	C E E F
Audence Audence Creatier Autor Uans Calos Andysis ¹⁰ - Demographics - Interest - One - Detection		Rade Densel V	3 m Type (6	Delast + Arquestion Section + 16.237 ¹⁰ strate Colline 3.209 (colored 1.359 (colored	3. Mare Interiors 72.005% degree trans 11.20% 64.05% 64.05%	New Own 11,798 ¹ of large 2114 of large 1414 of large	Reference Reserve Auto 32,778% Martin 25,37% S5,27% S5,27%	7,88 77,88 2424 (1011) 1011 1011 1011 1011	9, Ang Institu Doctor 00.02221 Ang Trans 00.0221 Ang Trans 00.0221 Ang Trans 00.0221 Ang Trans 00.0221 Ang Trans 00.0221 Ang Trans 0.029 Ang 0	atorical III Commissi Err Namation 13 13 13 13 13 13 13 13 13 13 13 13 13	C C C C C C C C C C
Audence Overliee Active Usen Orbert Andyste ¹⁰⁰ • Description • Description • Description • Description • Description • Description • Description		Rado Security American and American Security American * Rado Security American * 1 Apple Phone 2 por set 1 Apple Phone 2 por set	Beet Tope	bank + kopenites 16.237 States 3.200 party USB party USB party USB party	3 Mars 3 Mars 72,0075 1019	Non-Ober 11,798 11,798 2114 (* 1476) 2114 (*	Reference Reserve Auto 322,785% Notific Auto 25,39% 45,25% 45,25% 45,25%	7.88 https://www. 7.88 http://www.star. 1934 334 334 344 344 344	4, Ang beats Destine 00,02,21 Market 00,02,20 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,02,00 Market 00,000 Market	abaroot ()) Community () Community () Commun	C C C C C C C C C C C C C C C C C
Sourch reports & heap Audience Oranise Audience Oranise Autorose Gene Software Analysis with + Oranogaphics + Insteach * Gal * Balaniser * Tablaise		Scholar Belevander and	5 m Tupe (6)	Read + Augustine Institut 4 16,237 Scriet Sc	** Nov Manifest 72,00% 1,100 64,0% 64,0% 64,0% 51,1%	New Own 11,798 5,000 0,000 0,000 0,000 1400 0,000 14	Between Benner Hate 32,78% 19,27% 33,37% 51,27% 19,27% 19,27% 19,27%	Paper Sensors 7.88 ang the Sensors 3.91 3.91 3.91 3.91 3.91 3.91 3.91 3.91	4, Ang Institut Destitut 00,0224 Martine Marti	ational III Connections For Parameters 13 Notices Bactures 4 (2029) 2 (2029) 2 (2029)	© E E T Researce Res. 2005, 102 C Value All Park All All All Park All All Park All All All All All Park All All All All All All All All Park All All All All All All All All All Al
Auderse Oversite Auderse Oversite Oversite Oversite Oversite Oversite Oversite Oversite		Apple Phone 4	bet Type (h) (h) (h) (h) (h) (h) (h) (h) (h) (h)	Event + Augustion 10,237 5,538 5,239	3. Nov. Mainten 72.665% 1.125% 6.42% 9.15% 9.15% 72.90%	New Own 11,798 11,798 2114 or Ayro 1488 or Arrow 1488 or Arrow 148 or Arrow 148 or Arrow 148 or Arrow 148 or Arrow	Reference Researce Anter 32,78% 19,27% 19,37% 19,27% 19,27% 19,27% 19,27% 19,27% 19,27%	Pages Seattine 7.88 Anglet Seattine 2012 2020 554 3.077 8.67 2.52 2.52 2.52 2.52 2.52 2.52 2.52 2.5	9. Ang Senter Doctor 00.02221 https://www. 10.1298 00.129 00.0218 00.0218 00.0218 00.0218 00.0218 00.0218 00.0218 00.0218	ationed ()) Connection for Network Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences	© [2] [2; [2]] menter = Res. 205, 102, 0 % 1 (20, 0) % 2 (20, 0)
Anders reports & troup Consolair Active Usans Caluer Analysis ¹⁹⁷⁰ • Demographics • Demographics • Demographics • Definition • Caluer • Caluer Analysis • Demographics • Definition • Caluer • Caluer • Demographics • Definition • Caluer • Caluer • Demographics • Definition • Caluer • Caluer		Apple Phone Secure Q 500/7112 Galaxy Secure Q 500/7112 Galaxy	Set Type	brink + brink + brink + 16.237 ¹ 5.239 (2.29) 1.299 (1.24%) 1.299 (1.24%) 129 (1.24%) 214 (1.2%) 214 (1.2%)	3. Nov Mathies 72.665% 10.00% 64.6% 63.5% 63.5% 72.9% 83.5% 72.9% 83.4%	New York 11,798 11,798 2114 (* 2,794) 2114 (* 2,794) 11,614 (* 2,794) 11,614 (* 2,794) 11,614 (* 2,794) 11,614 (* 2,794)	Relation Research Relation 32,78% 15,39% 1	Project Sensitive 7.555 Angi Weight Sensitive 2.512 3.517 8.627 3.52 9.544 3.517 9.545 3.547 9.547 9.557 9.545 9.547 9.5	9. Ang Senter Dooten 00.02211 ^{Ang Senter 10.1298 00.129 00.129 00.129 00.129 00.129 00.129}	ationed III Connection for Numerical Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences	C 2 2 11 Remon Rs 205 102 S 30
 Inserte reporte à trege Andence Oranitair Andencie Oranitair Andencie Cahon Analysis ¹⁶⁷⁶ Demographica Intensita Oan Statulaire Technology Balaite Overviere Beisen Cabann 		Apple Phone 5 Samuerg Stat/Titly Galaxy	Set Type (6) (7) (7) (7) (7) (7) (7) (7) (7) (7) (7	рилл - Анриника Ф 16,237 5,239 (слок) 2,299 (слок) 2,299 (слок) 2,297 (слок) 2,397	5.800 5.800 72.6655 10.975 10.975 64.675 50.975 50.975 10.475 10.475	New York 11,798 301200 2014 (01200) 2014 (01200) 2014 (01200) 301 (01200) 301 (01200) 301 (01200) 301 (01200) 301 (01200) 301 (01200)	Relation Research Relation 32,78% Note Control 23,39% 32,59% 32,59% 32,59% 32,39% 34,39% 34,39% 34,39% 34,39% 34,39% 34,39% 34,39% 34,39% 34,39% 34,39% 34,	Paper boson 7,88 Seq to 2010	9, Ang. Seator Boation 00,022 51 00,02 58 00,02 58 00,02 00,02 58 00,02 00,00000000	adurocat (1)) Consensions for Neuroscience 113 133 145 (1400) 14 (6 C 1 (17)
 South reports it trap Audience Oranitar Audien Usans Oranographica Interestin Ore Balanitoria Tachtening Balanitoria Banditariading 		Apple Phone 5 Semang SMA/1016 Calary Apple Phone 5 Semang SMA/1016 Calary Apple Phone 5 Semang SMA/1016 Calary Apple Phone 5	3et 1ge 66 68 7 (8) 7 (8) 69 7 (8) 60 9 45 (8)	енин - лороните 16.217 % стар 2016 239 (нато) 239 (нато) 241 (нато) 194 (нато)	5 Mars 72,005% 11,200 64,85% 89,55% 89,55% 12,95% 10,45	Кон Оник 11,7508 Зайтан 2114 от Арто 1414 от Арто	Reference Ante Researce Ante 32,78% 19,278% 19,278	Paper Sources 7.88 Ang Ur (2011) 101 3.51 3.52 3.52 4.65 4.55 4.52 10.24 10.25	4, 449 beens beets 111 201 111 201	atoroa III Connector Fr Internet Intere	6 (() () () () () () () () ()
C. Source reports & trap Autorice Autor Uses Containe Autor Uses Containe Autor Uses Containe Stemation Containe Containe Containe Devices Contain Success Contain Con		Apple Phone Semang SSA2005 Color	341 Type (6) (7) (6) (7) (6) (7) (6) (7) (6) (7) (6) (7) (6) (7)	енин - лороните Марините 16,237 3,247 3,247 3,247 3,247 2,247 2,247 2,247 2,247 2,247 2,247 2,247 2,247 2,247 2,247 2,247 2,247 2,247 2,247 2,247 2,247 2,247 2,247 2,3	5.50 72.60% 72.60% 7.50% 6.65% 6.65% 6.65% 6.65% 6.52%	Man Shan 11,7558 Surface 2114 or 2010 1414 or 2010	Between Ante Beence Ante 32,78% 15,27% 15,	Paper bosons 7,888 Any to 2000 100 554 3.077 8.07 2.02 6.65 6.55 6.20 10.36	440 heatin bandim 000.02.21 00.02.21 00.02.28 00.02.28 00.02.28 00.02.28 00.02.20 00.00.20 00.02.20 00.00 00.02.20 00.02.20 00.02.20 00.02.20 00.02.20 00.02.20 00.02.20 00.02.20 00.02.20 00.02.20 00.00 00.02.20 00.00 000000	aturosi III Connectore Ter Parametere 113 114 114 114 114 114 114 114 114 114	6 () () () () () () () () () (

The Acquisition section shows the sources of the traffic to the site. It has many tabs showing different way of classification of the traffic sources. As shown below the classification is done based on the Google classification of the traffic.

C, Secch April & Ma	Acquisition Overview		4 Feb 2016 - 10 Feb 2016 -
• Instance Long	Anal Aport - Sharbad		*
O Real Time	O Al fersions	Add Segnent	
🚊 Autoro	Primary Distances in Constraints		
Acquisition Oversiteer • Ad Turfle • AdMoth • Search Engine Optimisation • Social • Companyin	Top Channes Copert Search Come Come Come	Sessons • Sessons Lite 2.00	Conversions • Econverse Conversion Rate Sinth Display of the Conversion Rate
E belaviour			

We can further drill down to have more detailed number of each channel. The bar graphs with actual numbers help comprehend the information faster.

C, Seere Apple & Long		Acquisition Seesee	1. Non-	Rea (1997)	Behaviour Bosna fate	Aperton.	Arg. benter Dentities	Conversions E-conversions	Sumation	Arrente -
Ø Real Time	1 B Cronic Search	24,294	73.69%	17,854	21.0%	10	00 02 50	105		Ro 226,453.00
🛓 Andrew	2 # Own	4.10	-		10.32%			6 10%		
Acquisition Overview	3 = (50w)	2,495			200	-		1345		
+ All Turfis + Additurds	a = Digity	1.665			14.955			185		
Optimization + Social	E at Erest	158	i -		20.36%	-		0.00%		
E Behaviour	7 in Faland	\$17	1		NDS	-		0.00%		
E Conversions	e e Sala	445	1		32.96%			9.22%		

The Source/medium tab is one of the widely used tabs for the traffic classification. Here the traffic is classified based on the sources tag provided by the users. We called it UTM in web analytics parlance. This code contains sources, medium and campaign. Whenever a user's comes to the site with the UTM codes, the Google Analytics automatically assign source, medium and campaign values in the UTM to the Sources/Medium tab.

		1748	174		750		8746		3740		10.7
Intelligence Events	***	and anterferrers second party									
Bad Time		Secondary devenues + Sec	Type Debut +					Q, and	mont [III]	e Σ ≤	m
			Acquisition			Behaviour			Commission	Lones +	
Auderce		louis Median		1. Sea	Are then	Brence Rate	Paperlansian	Arg Senter	- L	Transactions	
Acquisition								Duration	Rete		
Overview + All Traffic			24,284	73.81%	17.923	28.66%	9.09 Ang tar Osta 1.15 (3.07%)	00.02.50	0.06%	15 1. at Table	R1.2
Channels	0	1. google / wyards	12,996 (0.194)	17.14%	\$364 (15 19%)	21.31%		10.02.05	4175	\$ 10 mm	Ph. 111
Treamages Security Medium	0	2 (direct) / drome)	4,169 (10,0%)	75 25%	2,254 (18-0%)	33.32%	8.29	00.02.34	6.155	40000	Ph 103
Referats	0	3. geogle/cpc	2,858 (11,774)	64.45%	1.842 (19.27%)	\$4.0%	4.58	00 01 28	0.00%	1 in erec	
+ Additionals	0	A. Bestageting ((nit set)	1321 (1411)	75.38%	1,032 (1.71%)	9.62%	8.79	98 92 45	0.00%	£	
+ Search Engine	0	4. Apter/set set)	987 (1.774)	77.84%	706 (1.54%)	05.56%	3.94	00.011.15	0.00%	4	1
- Seciel	0	4. EMale/anal	\$56 (a.a.m.)	32.00%	125 (1994)	29.36%	16.13	00 04 55	0.00%	4 stars	
+ Campaigns	0	7. yahos / organis	224 (0.074)	6473%	545 (0.01%)	18.30%	13.48	00.04.41	8.00%	4.00%	

As Adwords and Google Analytics are from same stable there is auto linkage from Adwords to Analytics. Initial linking has to be done; afterwards the campaign information starts flowing in from adwords to Google Analytics.

Disease and	5.00	ardination Canadan inthing									
of preside strate & pub-		Sectory Brance + Set Typ	(brod +					Q stand	11 0 E	2 6 17	
• Inducta Conta			Arquisture				Between		Constants	1	
O fine Time		Campaign	caa - 4	Cost -	(H) -	hanne -	Bounce Rate	Pages 30000	1	-	
Andreas			4,685	Rs.35,867.62	Rs.7.66	2,858	54.41%	4.58 hepter/dest 0.00 cm 0.00	0.00%	0 5.41465303	
Ownlaw	.0	t beauting	0.000 (0.001)	Rs.7.680.54 (2117)	Ph 8.22	1.642 (27.10%)	75.9%	213	105	1.000	1
+ All Twife	0	2 Take 60A RBCT	5.89 (marks	Ba 7.575.34 (21.576)	Re 5.18	8.000	10%	0.00	0.0%	\$ (1995)	ł
+ Adliante	0	3 Seath Indu 8863	646 (m.m.)	Rs 15,206.67 (4) (m)	84.17.57	713 10 10 10	20.00%	4.75	0.02%	4.000	ł
Campatigns	0	4 Suits and Kurtle India.	379 0.000	Re.4.353.81 (comu	86.1373	336 (10.17%)	22.99%	5.74	440%	\$ anni	ł
Treatings	0	5 Real - Lakel MDA	B card	Rs. 808.74 crosse	81.853	8 (1994)	105	1.0	10%	1.00m	ł
Eid Adjustments	0	6 Brind Suits Online Nill	4 0.000	Re 633-69 (1776)	85 13 78	54 .0.000	8.05	5.85	1075	8 (1996)	ł
Rayworth Sealar	0	7 Seath	Harris	Pa 28.40 (comu	19-120	. 17 . 2.100.0	17.67%	12.76	0.07%	4.000	þ
Plane of Easy	0	A put set)	8 c.m.	Aug 10	Ro. 1.00	2 1000	34.00%	2.00	1475	1	1
Destination URLs	0	· Distry 2055 Remarkating	+	A.110	Ph 0.00	1.000	105	3.54	105	0 a 100	k
Display Targeting		-					-			1	

The **Behavior tab** provides information about the customer flow, pages viewed, the landing pages, exit pages and so on.

To Search reach & help	1	-	9	LP40	6746	-	1746	6746		1749	10.0
v		Penayl	warmen Page Page 1	te Otie -							
		-	- Decertary deverse	n • Settion	Debut +				9	abanced III 6	1 5 T 0
Autorea			nan -		Page 2044	Unique Page Vision	Ang. Tana on Page	Entering 1	Bearing Tate	5.64	Page Value
Acposton					220,701	58,100	00.00.21	24,284	28.66%	11.00%	Rs.44.0
D Defaultur			Anna		MANAMANA	4.358 (1.30%)	00.00.18	166.0.00	20.74%	12.37%	Pa.2.22
Overview	1	1	Auto		25.7% crame	3.82 0.00	00-00 17	2120 (10 101)	28.0%	145	Pa 19 32 10 10
Behaviour Filter		1.	in suits		26,669 (******	3.00 0.04	00.00 17	152 (111)	22.57%	7305	Po 24 57 (sear
+ Site Context			fuels.	0	21.050 (5.5%)	25% 3.0%	00 00 17	256 4.04	29.27%	1475	Fail 10
Contant Delidoum	1	0 5	Aure .		15,561 (1.01%)	9,354 (16.054)	90 99 32	5475 (2 10c)	\$ 90%	17.52%	Ro 36.63 (01.07)
Landing Pages		1 4	Alamak salts		15,100 (1.51)	25/8 (4104)	00.00 18	1.759 (1.0%)	22.40%	10.34%	Ry 28 44 (sector
Ext Pages	1	1 1	idesigner-Califies		6.00 cars	50 (10)	00.00 14	20	19.21%	DIATA	Ro 26-07 (mint
+ Site Speed	1	1.	Adeepatrical		1,540 camp	87.000	00.00 14	127 (1.5%)	34.09%	4775	Pa 2 50
+ Site Search	1		- Conditionantia		2,49 0.004	1.125 (1.714)	00.00 14	30	21.0%	6.32%	Re 41.55 (HUT)
+ Events			Advergendentilt		2414	379	00.00.14	9 4100	11.115	1475	Pa 2 10

The landing page is the page where customer lands to the site. The landing page is important for analyzing the traffic for the putting up the banners/ads.

To best much king	1.00	taria D	manual Landing-Page (Man 1									
v		~	Secondary deserves at the Type	Drive +					9,	abatet 🗐	0 5 3 11	
				Argentine	Angelation Ba					Constraint Economic +		
Analasian		1	anding Page	-	1. New	Now York	Brance Sales	Pages Season	Aug. Senator Desetors	Tanantan	Annual I	
E fataciae				24,284	73.81%	17,923	28.66%	9.09 1010-101 1010-101	00.02 50	15 5.4754 10.075/70	Rs 226,451.0	
Dominur Day	0	٩.	Note	5.475 cm erro	0.05	3,279 (1.974)	3.975	15.25	00.04.34	10.00 m	Ra 158 105 01 (m.m.	
+ Sile Carbot	0	2	Marga 0	3,646 (10.00)	87 19%	3,179 (11) (10)	20.74%	107	00.02.10	1.000	B101 111	
Al Pages	0		hat	2.528 (10.001	63%	2.114 (1.7%)	28.52%	7.14	00.62.64	1.00%	Re 16,858-30 -1-10	
Contard Deltabore	0	4	100 0	206 0.00	0.05	1039	21276	8.88	0.02.36	8.000	ALT:	
Landreg Pages	0	5	-	1.50 000	75.674	1.456 -0.176	2217%	8.45	10 12 33	9	Pa 1 11	
Ext Pape	0	4	nimmada di	179 2200	81305	1471 1 1910	22 87%	8.19	00 02 24	1.00%	Ro.11396.00 15/0	
+ Site Speed	0	1	Indentive contracturities a	#2 .cm	7475	373 0 000	45.90%	3.46	00 89 23	1	A+11	
+ Site Search + Events	0		this is a set in closed out in high sing and the	4 9 (1.11)	61.10%	286 (1474)	MIPS	135	66.66.39	8.000	Re. 8.10	
+ Publisher		*	Induktion and a	30.000	81.54%	275 11 11 11	21574	5.94	10.11.42	8.000	Pa \$ 28	

The page timing provides the average time for the page to load. The higher the page load time more likely the users will bounces hence the web designed tries to reduce the size of pages to increase the page load time.

Care and a start start		1746	6746	 1740	8716	3/4	10.7
				10. March			
	-	uni Omericani Page Page 186 (1858-1					
a reserve		Incodes meson + bollyse datat +				6 , m	11 2 11 ton
Acquister		Page		Auge class	• •	(ing Reprint Tracial) #	(compared to all average)
1 Selector				Notice 1	220,701		13.41
Orabilas		Autoriga			8.05)	D7.20%
Bahaviour Flow	2	auria			25,716		205.81%
+ Site Context	3	reals			24,665		23.295
+ Site Speed		Auto			21,010	-65365	
Overview	. 5	nome			19,561)	12.71%
Fage Tenings	- 6	induition of suits			15,500	00 10%	
Speed Suggestions	. 7	Mexigner clutters			8,030	-48.55%	
User Timings		Advergational			7,50	40.545.3	
+ Site Search		numerican series			7,815	.47.89%	
+ Exarts	-	Anhanger (Investig			2,614	41.7254	
+ Publisher						Barries in	And Inc.

The users searches at the site is captured in this section. The site search is important because it provides infromation about what kind of product users looks for in the site. You can find the information such as number of session with search, the time spend after the searches, the average dept of the search etc. In the Search Terms tab you can find the actual keywords being searched in the site.

 A standard space to tag A standard space to tag 		-		-	
🚔 Autorea	_	174	the The	8746	3.fat. 15.f
* Acquistion	0.56% of your visit	ts used site search			
Balanteer Overview Betaniour Plans + Site Content + Site Speet - Site Speet	135 5 Search Edits 22,68%	Total Unique Searches 194 % Search References 25.00%	finado Page Vinos Search 1.32 Tree After Search 00:04:32	Visis Wear 5	to Sect. B Vick We Ste Sect.
Overview Unape Search Terris Pages	Average Search Depth 10.11	-			an D
+ Events + Publisher	She Canteria		Search Term		Total Unique St Total Unique

The conversion section is about the bottom of the funnel. This section tracks the visitors action such as completion of a goals, completion of the purchase for ecommerce site, the products being purchase etc. the important metrics available are transaction, conversion rates, revenue, average revenue per order and so on. This section is very important because the intent of the website is being captured here.

C. Sand much king	* Quantine				
Anderen Argehöten Beheriner Conversions	Round - Vi Commercianes • Revenue - Economica Comm Roccionia ************************************	e Kon + jo nius Ras tifus	17m 87m	7009 (by) Red 174	8.0% 6.0%
+ Commerce • Economerce Overview • Shapping Analysis Product Parloma. Salas Parlomance	Revenue & Conversion Rate Revenue R5.226,451.00 Marketing	nemete Convesion Res 10 06%	Arrange Onter Value 5 RS.15,096.7	3	
Product List Parls . • Matheting • Multi-Channel Funnels • Attribution	Campaigne 1 Transactions R5.2,850,00 Transact R5.2,850,00 Average Celler Value Tran Sellers	Internal Properties 0 Impretations	Order Coopen Code © Transactions RS.0.00 Revenue RS.0.00 Average Criter Value	Attilution 15 Transactions R5.226,451,00 Revenue R5.15,096,73 Average Onter Value Product	b

3.1.4 Goals & Funnels

The goals are the goalpost in the site to measure the completion of the specific action by the customers. The website can have multiple goals based on the objectives set by the company. The goals helps in the measuring the effectiveness of the marketing activities and communication. The typical goals are events like Email subscription Completion, the purchase events completion, providing mobile number as a lead etc. In the section we will learn how to set up the goals and measure it in Google Analytics. After the goals are created we will learn how to create a Funnel. Funnel is visualization of traffic flows like the typical sales funnel. The simple funnel could be HomePage - > Product Page->Checkout ->Purchase.

For creating a goals go to Admin, click on the Goals.

🚰 Google Analytics Norte Ray	nting Cummication			· o 40
and and a second se		man	ine .	
Testarte		Test Site	 lest Site	
The Management		C Property sample	The Wangerset	
Y Althen		(A. Sourcepiele	M Grain	
O Owgenitation		modul(1 cmm)	A Const Despity	
1 Test Car		INC Advisors Lanung	T films	
		Atlana Labora	[5] Daniel Setimp.	
		At Octory Letting	H Common Settings	
		Al Probatis	Def. Calculated Warnes and to	
		D futers		
Click on +New Goals or you can edit existing goals by setting editing it.



The goal template will provide several options- the predefined templates like place an order, create an account etc. You can create custom ones by using custom radio button and continue with next steps. Here we will create checkout completion goals.

Template		
Select a template to start with a pre-filled configuration		
REVENUE		
Place an order : Completed purchase or pre-order request.		
Checkout complete Constants a purpose	₽.	
Acquisition		
Create an account Succeedul oprival account, or view present		
mquitty		
O Meet more - Venesi product or service details		
Contact us . Viewed phone number, directions, shet in small		
See available Channel inventory or schedule		
See deals		
C Read reviews . Viewel within and strips		
ENGAGEMENT		
Compare Information Compared Natures, products, or systems		
Share / social context. Shared to a social research or ameliad.		
 Newsletter sign up - signal up for revolution or mailings 		
Add to Call. Saved product to shopping set.		
) Smat Goal		
Maasura the most angaged visits to your velocitie and automatically turn these vo death to improve your AdMards bidding. Learn more	is inst blaik. Than use these	
Custom		

Check the checkout complete option

 Template 	
Select a template to start with a pre-filled configuration	
REVENUE	
Place an order Completed purchase or pre-order request	
Checkout complete Completed a purchase	
ACQUISITION	
 Create an account Successful sign up, account, or view greated 	
INGUIRY	
View more Viewed product or service details	
 Contact us Viewed phone number, directions, chat or email 	
See available Checked inventory or schedule	
See deals Viewed current sales or promotions	
Read reviews Viewed reviews and ratings	B
ENGAGEMENT	
Compare information Compared teatures, products, or options	
Share / social connect Shared to a social network or emailed	
 Newsletter sign up Signed up for newsletter or maillings 	
Add to cart Served product to shopping cent	
Smart Goal	
Measure the most engaged visits to your website and automatically turn those visits into Goals. Then use those	

In the goal detail you have you provide the url of the goals. Assuming out checkout page is <u>www.test.com/checkout</u>, we add /checkout in the url.

Goal desc	iption tai		
Goaltype De	strator		
Goal detail Destination	5		
Equals to +	ideotat 🖸 Ca	se sensitive	
For example, a	as by Soreen for an app and charayou how instead of even available concharayou how fo	r a ven sege	
Value antors			ь
OFF	Aprign a monatery value to the conversion		
Funnel surv	wi		
Fennel serve	1		
Fennel syny Orr Standy a jach ynur Osal	ul you waped halfs to take lowerin the Sectoration. Use it is analyse the extension and exit p	tent fra intent	
Fennel synty Orr Standy & John your Over Verify this Co	uel pris sequed traffic to task lowers the Ballmannin. Use it to enables the entering and exit p M - Bae fine other this Baal initial takes servaried based or your bala from the part ? or	uning that interest	

Click on save. The goal is created.

Goal	+ W	Past 7 day conversions	Recording
Checkout complete	Goal ID 2 / Goal Set 1	۰.	COK
Place an order	Goal ID 1/ Goal Set 1	2140	01

Now you can view your goals in all the tabs and select the goals in case you want to see the goals completion.

				Angelater			Belancer .			Carvaniana Board Deated annual -				
Austine				beause - 4	honors & Silver Australian		Barra 648	Recent for Pages' Any Recent		Ouchout complete dis 1 Conversor Rates	eCommente All'Goals Goal 1, Place an or	3		
Overview Active Usars	L			74,849 6.47 Teser 700.000 (70.000)	68.74% hig to take 0.17%	51,448 5.4 Tax	33.92% Aug for Year 0.02% 0.12%	3.43 Aq37 Tar 140 B B W	00:01:53 40 57 000 0 07 00 0 07 00	0.00	✓ Gran 2 Chemodra	argana .	10.00	
Concernances	0		Ovone	40.676	67.32%	32,772,00 1010	37.39%	3.32	000153	6.00	5 8.00m	R100		
+ Marcala	0	1	Proto	6746	64.79%	4,375 (4.11%)	45.40%	2 29	000153	0.00	6 8.000	R4.00		
+ Geo	0	3	UC Browley	\$2727 at 1000	71.52%	4.992	7.62%	5.42	00.02.43	8.00	• • • • • • •	A100	10	
- Behavior			Opera Mm	4,638 (0.20%)	18.22%	3,472 (1111)	4.00%	4.45	0001.40	6.00	 8.530% 	Rs.0.0	10 in 11%	
+ Technology	0	1	Android Browner	3,590 (4.17%)	73.58%	2340	1.47%	3.13	00.01.20	8.00	6 0.00m	Ra00		
Browner & OS	0		Suferi	2,378 (0.10%)	77.12%	1834	7.44%	3.85	00.01.10	8.80	6 . 0 minu	AL. 0.0	10	
Network			Internet Explorer	UN dim	72.52%	1,648 (2110)	10.00%	234	000128	0.00	5 5.00m	R100		
+ 1010318	0		Opera	1,000 (1.30%)	1230%	623 miles	43.20%	2.84	00.01.58	8.00	6 8 mm	Ru00		
+ Custom	0		Safet (mage)	295 0.075	75.01%	200	2.36%	4.80	00.01.15	6.00	6 0 miles	PL.00		
- Benchmarking Users Frie		15	Edge	100 0.070	10.07%	101.0124	54.17%	5.05	40.05.08	6.00	6 8.00m	Rs.0.0	10 ann	

For funnel we have to Put Funnel Option ON for the goals. Let's add funnel to the placed an Order Goals. Assuming home page is where the maximum traffic visits, the first steps in the funnel are home page and second step in the funnel is checkout. Save the funnel once steps are defined.

0	Goal d	fetails flos			
	Repla	expression - checkout/succes	5	Case sensitive	
	For search	iple, use My Screen for an app and ithan	type the indext of several angle.	con-Manayou Amil for a vest page.	
	Value -	optional Assign a monetary value to the	Burketin		
	Funnel	optional po parean name plong or a reab page DM and instance of arrow aparticle communities	, for each step. For exempte, use to you html for a web secs.	ly Screen for an equilant	
	Step	Name	ScreenPage	Required?	
	0	1	Home Page	NO	
	Θ	Index.php?route+checkoutiche	Checkout	0	
	• Ad	ld another Deep			
	Verify Dr	is Goal See have aftern this Quart woul	d have sofverted based on your de	da Torn Pre padi 7 daya.	
Ca	nost .				

Þ

From Funnel Visualization you can view the funnel created. The funnel just created will take some time to populate.

A free sum time	1	Place an order				
		This Gearwas (ampleted in 17,454 sessions) 100 00% fund	nel conversion i	ute .		
Contant Drittown Landing Pages	1			6		
ExtPages			0 >		> 0	
+ Site Speed				Y		
- Site Search				0-(0.00%)		
1040				prosedul la trata pre factor d'actual d'actual		
+ Publiches				1		
Equitients						
In Page Analytics				index.php?route=checkout checkout		
Conversions			• >		Þ 0	
+ Geam						
Overview				e (6.00%) prosected to Sector or other		
Oner URL e						
Reverse Geal Path				+		
Funnet Visualiza				Place an order		
Goal From				17,454		
Breat Goals Har			17,454 >>	185.00% Screet commencies rate		W.
· Exemute		Notes and Provident Particular Interioral	61.142			
Hill/B-Channel Furnets		tedas pig ¹ tiude-accountiectium				

The data populated for a day in checkout.



From goals flow tab you can view the goals completion for any dimensions. In example below browser is chosen. You can choose any dimension.



For the ecommerce site Google analytics has enhanced ecommerce which helps in better tracking of the customer funnel, product analysis, category analysis etc. The interested reader can explore enhance ecommerce.

3.1.5 Integrating Google Analytics Data with BI System

Quite often we need to pull data out of the Google Analytics in the Excel to do more complex data analytics. For a limited data set excel export works but when data set is large and the dimension is more than two it is really difficult situation. The Pentaho data integration with Google Analytics is quite handy tools to extract data from Google Analytics and put it into the tables in the BI system. One can do incremental data transfer and schedule it to get incremental data from Google Analytics to the Tables. The cases where the data from Google Analytics are to be used in the data analysis of the order engine database like finding the sources of traffic for the orders having Google Analytics data in the BI system really helps. In this section we will learn how to make the connection and show some sample data transfer. The users can create more complex system as per their requirement. Again the Google Analytics Dimension and Metrics explorer will be useful in finding the API field names and data types. Select API Names mode and find the required metrics and dimension for the API names and their data types.



The stepwise process to create Google API is given in the http://wiki.pentaho.com/display/EAI/Google+Analytics

Go to <u>http://console.developers.google.com</u> and login using the Google Analytics Credentials from which you want to create connection to Pentaho. In the login there is option to use Google API. First create a project and name it as per your convenience.

the Demonstrate Rest Rest Rest Rest Rest Rest Rest Re	nu x/		0 0 0 A 1
Type on the action that and you'll get \$200 in a well and \$2 days to explore the	iga Coul Patty in Laws may	DEMES	NON OF FOR FREE TRAC.
Google Developers Console		Belef sympet + 0 Ø	0 8 1 🎯
Getting started			
10 martine territori		Documentation	
Line upoge APre	enveronment)	E Steeps David Mattern decommittee U	
and stage	Create and exploy a metallitisme appli- schede scenying about the underlying	E Good Partient and anno 12	
RPE Ender and manage NPa	superconduction on the finger research	Contraction of the	
Distance of the second	II And Experimentation (2)		
Try Compute Engine	a fine the large test and through	Start with a coalcult solution	
Sperup ortical machines using Banger	down that the law in April or good of contraction	Shareh year fanante sattware jeal ager	
Computer Engine Route you introduction the complete a guardinate again the guided evaluation	🐡 Stat the Polt general satisfyingh	an Sought Could Pathons with pat a few wishes	
(Bet stated	🤣 Start for Fyrlun " golael valid ooph	A knowperspec	
• State or imply point	Try App Engine (managed VM environment)		
	Nong plur favorite language, daphis a sample approachise using thereaper tilds		

Create a new project



Select Google API section for enabling the API.

← → C 6 & https:	 B Coope Avenue - Terrari, N/ Q Paree - Fertana x //console developers google com, "home/dashboard"pro 	yett-pentaba-1227	1888 - 0 1 1 A 1	
1 Teristerteres	and you'll get \$200 in a solid and 60 days to explore Boogle Court Pather	- Lean mar	DISMISS BOW OF FOR FREE TRAL	
Google Devel	opers Console 9.		resse - 🗷 🖉 🖉 🖉 l 🎯	
A Home	Dashboard			
12 Conterest	(beneficial		1000000	
E Alley	Propert Party of Control Control of Control	Try App Engine (managed VM	-	
	the period of the contraction	The second secon	Bengie Deut Particie Sourcestation C	
	The County Ally	sample application pring Managed (Ma- (being)	Deal Parties existence C	
	Logis Of a cost statistics, and had	Under Google APIs		
	and and a second se	El Weiger Withenstein C		
	BF2 Endle and manage MPa	Start the Scale jo guided wait through	Start with a prebuilt solution	
	Try Compute Engine	Try App Engine (sandbox environment)	Laurch pair face face to collecte participation for Songle Could Pathon with path 4 few of the	
	fam up inter machines using though Compute English, toologis, and Minipublish for small is a particular, angli intera partici anticitanage.	Over and exproj a walk fitted app estimat everyong allow fire orderlying orderstructure in time guided waldbrough	V needender	
	0.1000	🔯 Aze Engrie Assumementen, U		
	(c) the result	👹 - Start Har Do posted and Honough		
an	un hore instanced property periate (227			

Select Analytics API

19 Disasten & Metric Ec. W.	Coople Analytics - Terrary . IN / Q. Al	filinay-heala x 📖			BBB - 0 - 1				
← → C 6 8 https://c	onsole developers google com/i	spis,58nary1project+pentaho	-1227		0004				
1 Spintersteers and	profil get \$200 manufil and \$2 days to an	Jos Stop Coul Pattern Games		DISMISS INCOLUMN TO FIRE T					
Google Develop	ers Console	٩		reste - 🖬 🕖	o o i 🍲				
RPI APi Manager	Overview								
- Conver	Bergle Mile - Distant all	9 (T)							
 Determine 	Q. toward of 1000 MVL Popular APIs Void March Control Status Complex Control Status Control Control Control Control Control Control Control Control Control Control Control Control Control Control Control Control Control Control Control Contrel Contre	Art Design After	Provide Marce APIA Secure Marce And AN Secure Marce And AN Secure Marce And And Secure Marce And And Secure Marce And And Secure Process AN for Andread Secure Process AN for Andread	Security Access MPIN Strom ART Science (MP) Brend ART Brend ART Brend ART Articles Strot Constants ART Constants ART					
	Australia Anto Marcina Anto Complete Transition Complete Tran	enageng te Services degree API Priter Andread Annuel API annuel API	Social AFIs Social article Social article Social art Social	The Allow Mark Allow For Section 2014 Allow Allow Allow Allow Allow Allow Allow Allow Allow For					

Enable the API by clicking on the Enable button

0	Equiption to the trail and pro-	to get \$200 manufit and \$6	lays to explore Boogle Court Party in Grant Hare				DISMISS		BONG	-		
	Google Developer	s Console	٩			Periodic -		0	۰	۰	1	*
RPI	API Manager	Overview										
4	Dramar	- 1999										
8	Overantian	Analytics API The Analytics API Land Heat Anne Thy the API is AP Accessing uses A Viscous access diam 2 proved formation are the Land Heat	provides access to instant to construct and equat para. 4 Cystem : C ² As with Year 24 on the construct and on the construct and a cons	2	- V	-1	Line: data					
		Tenuer fan servier fins Yna can ster fins Instanten e wielden an witch enstellen an witch enstellen an	devided IP genter and source interaction for extension photomore a strong to anyone their reset a strong resource constraints and a strong to anyone the strong resource anyone applications. Next they because a strong resource anyone applications for anyone (parts more anyone) anyone anyone (parts more anyone) anyone (parts more anyone) (parts more anyone) (parts more anyone) (parts more an	a and a second	Atherselen	- 0						

Once the API is enabled, you have to create credential for using it in the API call. Select Service Account Key

0	The interaction of and you	Type 2001 manufit and 60	days to explore Boogle Cloud Pathors, Gene mare		DEMO	TRAL				
	Google Developer	s Console	٩	Person -		0	۰	۰	1	۲
RPI	API Manager	Credentials								
۰	0.01100	Cubritan C	Autoprise Detail and other							
8		40 kay Namita San Kabu Kalan Garan Kaba Mana Kaba Mana Kaba Mana Kaba Mana Kaba Mana Kaba Mana Kaba	ngan Laing a anggi A ³ ing ta diau ta para ang ang ang Inggi Tawaran anggi Tawaran anggi Tawaran anggi Laingi Ang Ang Canada Ma	er sinnin Cere C 2017/04/203-uglen/fra/fradfijlens/flightfarserd.ppr.prod		te di ce	•			

Select service account and Key type as P12.

 All and the second se Second second se			growth open states
The state state of the	gh up fur in from trial and poull get \$200 minord find bill days to exploye Boogle Cloud Photform, Soarn more		BON OF FOR FREE TRA
Google Develop	rs Console 9.	Particle - 🖬 🛛	0 0 0 1 🚳
I APi Manager	Credentials		
Contentials	Oreste service account key		
	And a second		
	App Drgine default service account		
	Kon sear To solve the first sectors the public protecting pair. How the first party is source that the public protecting pair is the first sector that the public protecting pair is the public protecting of		
	Gran Garan		

P12 file will be downloaded into your system. Keep it safe in a location which is relatively permanent.

- Director & Merici Sc. # C 6 @ https://d	😸 čespis Analytica - Termini, iki / 🥥 Česek service a ornačie developers, googie com i spis, ico do	assurble: *\ mlab_hanvior.accountiley?project+pentabo-1227	1000 - 0 - 0
[]			
AT ATMongs	Contentials (a) Control encourse account of A on Proceedings and account of A on A on Proceedings and account of A on Proceedings and account of A on A on Proceedings	Alle (private larg pair * AddBhritish (st) on the state state of pair simples. This is the units for the private state states; the private states is a state of the state of the state states and states are states and the private states. We state the private states are states and the states are states and the state states are states and the states are states are states and the states are states are states are states are	
Persta-Autoballp12			± iterationistic ×

Now our API related steps are completed open Pentaho Data Integration. Create a new transformation, select Google Analytic from the Input.

	Spoon - Transformation 1	- 0
ed the Ader Test May		Perspective 👯 Cata Integration
Ver & Design	T fandemater 1 II	
tere to 12	DIDEEP XOSC II WALLY	
C decrete from B de de las tes 154, de las tes 1	nagha beat han gangka analytics assured	

Add service email account available in the credential in the OAuth Service Email. Browse and add P12 file in the Key file. Click on get profiles, you will see all Google Analytics Profiles associated with this service account in the list. Select the profile for which data is required.

In the query definition area add start date and end data in YYYY-MM-DD format. In Dimension add dimension you want to extract, in this example we have taken Date; it has to be done in format ga:date (ga:<dimensions>). For multiple dimensions just put comma between dimensions. Add metrics in the similar fashion in the Metrics area (ga:<metrics>). Each metrics separated by a comma.

In case if you want data for specific segments from Google Analytics you can specify that in the segments area.

K	*	s Input Step	oogle Analytic	• •	•
1		1	Google Analytic	Step name	file Edit View
erspective It Data			ngs	Google Analytics Connection Setti	8 8 6
		ttie plugin	type-exit.org ke	Application Name	View @ Des
	m.gserviceaccourt.con	167.44	6	OAuth Service Email	Steps (A)
1	Browse			Key File	Printer P
	oBeference		9.6	Specify tableid directly	D Ottot
			~ e	Get tableid from profile	m Banaform
8	· OR provies				Co Unity
				Query Definition	Film
	\$		2016-02-01	Start date (YYYY-MM-DO)	
	0		2016-02-29	End date (MMM-DD)	BA Server
6	oBeference		ga date	Dimensions	D Lockup
	a coopying a Belerence	tors oa transactions o	Da visits oa visi	Metrics	aniot C 4
	ABelerence			Filters	Data Wan
	Balanacca			6.00	Validation
			-987/5/5	fourth a comment for the most	Statistics
(0,			×	specify a segment for the query	Big Data
	◆ DELESSE		° [Use custom segment	> C Agle
	✓ Get segments		All Visits	Use predefined segment	Crystogra
			cs. Reference	Output Fields Dimension and Metri	Paro
3	Output Type a	Output Field	ed field	A # Feed Field Type Fe	Copen ERP
	integer 4	vata	a visita	2 Metric gi	dot 🖸 🕯
	rteger r	os prostors	A VISITORS	3 Metric gi	Mapping

Once dimension and Metrics are added, click on get fields, all fields will be populated in the Output field's area. You can delete a row, change name of output fields or change the data type. Once the fields are ready you can preview the rows.

4	-	ip-10-0	-u-ar.ap-sout	seast-1.compute.memai:1 (ec2-user) - v	N. Viewer	
	City Edit Mary	85.	10213-000	boogle Analytics Input Step	×	
	File East view		Step name	Google Analytics		
	0 6 0	Google Analytics C	onnection Set	tings	ers	pective: 🔀 Data I
	View / Desi	Ap	plication Name	type-exit ora kettle plugin	0	
¥				Examine preview data		
lows o	f step: Google Analy	tics (4 rows)				
	ga:date	gavisits	ga:visitors	ga transactions		ga pagevia
1	2016/02/20 00:00	00.000 913669	759299	26850		1882
-	2 2016/02/19 00:00	00.000 859427	700673	28204		1865
1	3 2016/02/18 00:00	00.000 852740	679188	49526		2052
- 24	4 2016/02/10 00:00	00.000 775776	641899	20030		1566
						a y
				Cine Show Los		B

Now our Google Analytics input is ready, add a table output in the transformation. Make sure that a table with same columns is created in the database. In this example we have created a table GA_Traffic with date, visits, visitors, transactions and pageviews in Pentaho database. Connect the database and select the table GA Traffic.



Connects the fields from Google Analytics Input to Table Output



Execute the transformation we can see 20 rows been transferred from Google Analytics to Table GA_Traffic.

ip	10-0-0-87.ap-southeast-1.compute.internal:1 (ec2-user) - VNC Viewer - 🔍
*	Spoon - GA_Pentaho
File Edit View Action Tool	Help
0 8 8 8 8 8	Perspective: 🔯 Data I
View / Design	IX GA_Pentaho 21
Steps 🙀 👍 t	
Þ 🛅 Input	
Dutput	
Transform	
Utility	
Þ 🛅 Flow	Google Analytics Table output
Scripting	
D BA Server	
Lookup	
p ioins	
Data Warehouse	
Validation	
Statistics	
🕨 🛅 Big Data	(a)
Agile	
Cryptography	Execution Results
P 🗀 Palo	(2) Execution History [1] Logging Step Metrics \ [2] Performance Graph [2] Metrics @ P
Dig Open ERP	
dol 🖂 🕯	▲ # Stepname Copynr Read Written Input Output Updated
Mapping	1 Google Analytics 0 0 20 0 0 0
Bulk loading	Z Table output 0 20 20 0 20 0
b C3 Inline	

Verify this by querying on table GA_Traffic, we can see 20 days day wise visits, visitor, transaction and pageviews.

Ele felt View Search Terr	inal Help	
Duery OK, 0 rows affected i	0.03 sec)	Perspective: 00 Data Ir
ysql> select * from GA_Tra	ffic	
<pre>>> : (mpty set (0.00 sec)</pre>		
ysql> select * from GA_Tra	ffic;	
Date Visits Vis	itors Transaction Pageviews	
2016-02-20 913009 7	00673 1 28304 1 1865327 1	
2016-02-18 852740 6	79188 49526 2052848	
2016-02-10 775776 6	41099 20030 1555930 1	
2016-02-17 749066 6	08429 22512 1617410	
2016-02-11 742005 6	10829 19471 1552858	
2016-02-16 712103 5	75563 19179 1554952	
2016-02-15 644087 5	19729 17664 1397920	-
2016-02-12 627280 4	98936 18346 1366582	
2016-02-13 540825 4	21806 16379 1231100	
2010-02-02 533430 1	97649 17071 1231563	
2010-02-08 532780 4	17929 18328 1202202	
2010-02-03 5220441 4	0/1/2 19939 12300/3	
2010-02-04 317442 3	84178 1 16831 1 1164365 1	0
2016-02-03 500200 3	041/0 1 10630 1 1343363 1	
2016-02-01 503589 1	16430 1 16430 1 1348346 1	No.
2016-02-06 502105 1	92661 1 17030 1 1196537 1	hand the statistics of the
2016-02-07 485968 3	81425 1 14532 1 1147681 1	praph Le Metrics (d) P
2016-02-14 464078 3	67287 1 15488 1 1049551 1	
		Output Updated
to rear in set (oros set)		0 0
vsal>		
P 1 80K 1030/00		20 0
		10 - Wites Costs - 50
P 🗖 baline		b Ge
Experimental		
h an Annual		
Contraction of the local division of the loc	and present the second s	1
X- (2) W (2)	📰 🎧 📭 🛸 🔀 V2	- D

Now let's add one more dimensions ga:devicecategory.

- 01	viewer	act (ers-aper) - the	contest occupationer	- 10 - 0 - 10 - 10		
erspective: 🕃 Data			Settings	Google Analytics Connection S		rite to
	•	olugin	ame type-exit.org kettle	Application Nat	View Ø Des	BV
	wiceaccoure.com	ebazaar-1167.jam.gse	mail askmebazaar@askm	OAuth Service Em	teps M	·> Step
	Browse_	16667c96/1.p12	File Idata/Askmebazaar	Key f	C Input	pty si
	· Seference		ctly @ [ga:85464358	Specify tableid direc	D Output	rsql> P C
	v Get pridies		ofile	Get tableid from pro	Transform	Date P D
	lent and a second			August Datables	U Utility	2016-1 P C
			00) [2016 02 02	Start date (VYY-MM-C	E Flow	2016-1 0 0
			2010-02-01	fed data (CONVINIA	Scripting	2016-1 P C
	•		2016-02-20	End date (TTTTPINI-C	BA Server	2016-1 0 0
	♦Beference	tegory	ions gardate, gardeviceca	Dimensio	Lookup	2016-1 0 0
	eviews ofference	ga transactions, ga pag	trics gavisits, gavisitors,	Metr	D Joins	2016-1 0 0
	o Beference		iters	Filb	Data Ware	2016-1 0 0
	oReference		Sort -gavisits	5	Validation	2016-1 0 0
_			very g	Specify a segment for the que	Statistics	2016-1 0
A	oference		nent O	Use custom segme	Big Data	2016-1
	u Catumente		nent all scalar	Use predefined seam	Agie Agie	2016-1 P C
Metrics @ P	• Get segments		AI VISIS		Cryptogra	2016-
	10		tetrics Reference	Output Fields Dimension and Mi	Contra ERR	2016-1
t Updated	Output Type	Output Field	Feed Field	 # Feed Field Type Dimension 	C ish	· · · · · ·
0 0	30000	Device	gaidevicecategory	2 Dimension	Macoing	
0 0	Subana C	l scielles	- Andrew	stance .	Bulk load	/sque P C
			0	Linuit et	C Inline	P 0
P			stat (-	Dunic	D Experime	P D
10000114	ecz-usenthip-1	tview Cancel	Get Fields Pri	CK CK	Pa Presserves	

We can see device category column in the preview. You can create a new table or add device category column in the existing table to transfer this data into the table.

file	E	9 B B	6000	Anabelies	Connection	Lettinost			erspective:	C Data Ir
hier	y 0	D Man (& Dec)		A State Stat	tolication Na	me land	and any fermine of ania	10		
×.		THE DESIT		-		Exar	mine preview data	0		
Rear I	. 11	stery Google Analy	tire IA e	(Land		-				
~		Date		Device	visits	visitors	transactions		_	pagevie
	1	2016/02/20 00:00 0	000.000	desktop	519747	437738	15740			942
	2	2016/02/19 00:00 0	000.000	desktop	447638	367484	13585			881
	3	2016/02/18 00:00:0	000.000	desktop	419868	336054	26267			947
	4	2016/02/10 00:00:0	000.000	desktop	411622	339664	8572			783
										a good
							Close Show Log			and p
		b 🛅 Bulk load	6				Close Show Log	•		
		5 ⊡ Bulk load 6 ⊡ Inine	(8)		Umit	uze 0	Close Show Log)		2 (Jac) (Jac
		Bulk load brine brine brine	6	5	Umit:	uize 0	Close Show Log	D		a poor

PDI Google Analytics integration is very convenient tool for extracting data from Google Analytics and directly add data to database. Quite often we need data from Google Analytics in our data warehouse so that it can be combined with sales data and do more meaningful analysis.

3.1.6 A/B Testing using Google Analytics

When you are not sure about which pair of shoe to wear in an occasion or when you are not sure about how to tell a sad news to your relatives or when you are not sure how your parent will react to your grades or when you are not sure about how to say four letter word to your dearest one and if you have more than one option, didn't you ever wish you can try all option one by one on someone else and finally use the best word in the best possible ways for the final ones. Alas! That is not possible in real world but definitely possible in virtual world. Yes! We are talking about the A/B Testing and multi-variate testing.

Most of the A/B testing is done through some king of testing tools. For instance A/B testing of ads using Google Adwords testing section, A/B testing of pages using Google Analytics, A/B testing using Adobe Test & Target and so on. There are many other tools out there in the market for A/B Testing. The good things about these tools are that they help in the creation of test with ease and they provide result in very readable format which any one can understand.

When you create ads in Adwords, sometime you wait for few days and makes changes to see if there is improvement in performances. If there is improvement in performances then keep the changes else we make more changes or revert back to old ones. Here unknowingly we are doing A/B testing of ads in sequential fashion. Similarly when we makes changes in website pages and compare the result from old page to new page we are doing A/B testing. We do not call it A/B testing probably because we are not using a testing tool and we are not taking users through both experiment simultaneously as all testing tools will divide traffic into predetermined percentage and send it to experiment A and B simultaneously. Nevertheless they are also A/B testing.

One of the most important assumptions of the A/B testing is that the traffic is divided into two samples using random sampling method. Does random sampling necessarily need to be done simultaneously? No, not necessarily, we can call people coming to our site/ads on day 1 as sample 1 and the people coming to our site/ads on day 2 as sample 2. This is sequential way of sampling but is as good as simultaneously random sampling for testing your ads and pages; only issues being your inability to pre-determine the % of traffic for both experiments in advance. Hence we can really do A/B testing without any tools and in fact we unknowingly do lots of A/B testing.



How do you determine the result of A/B testing when no formal tools are being used? If A > B then can we say experiment A is better than experiment B? Not exactly! Remember here we are dealing with the samples from the population. We can never be 100% sure about the result from the sample will match the population characteristics. However we can say with 50% confidence or 80% confidence or 99% confidence about the population characteristics using sample data. The confidence interval you want to use is simply based on your requirement from the test; how much type-II error you can live with. This is again another use of Central Limit Theorem we discussed in chapter II.

Formally we can define A/B testing as

Null Hypothesis H_o: p_a=p_b

Alternative Hypothesis $H_1: p_a \Leftrightarrow p_b$

Assuming sample 1 as $n_1 = P^*X\%$ and sample 2 as $n_2 = P^*Y\%$

 $p_a = L/(P^*X\%)$, $p_b = L/(P^*Y\%)$

Pooled proportion $p_t = [n_1p_a + n_2p_b]/(n_1+n_2)$

Standard error of the Proportion difference is calculated as

 $\mathbf{6}\mathbf{p}_{a}-\mathbf{p}_{b} = \sqrt{(\mathbf{p}_{t}^{*}(1-\mathbf{p}_{t})/\mathbf{n}_{1} + \mathbf{p}_{t}^{*}(1-\mathbf{p}_{t})/\mathbf{n}_{2})}$

The Z score is calculate as $Z1 = (p_a-p_b)/\sigma p_a-p_b$

Based on the confidence interval we can get alpha α =100%-confidence interval and the critical value Z α using Standardized Normal distribution table.

Reject null hypothesis if $Z1 > Z\alpha$ or $Z1 < Z\alpha$



In Hypothesis we can make error as it can never achieve 100% result, hence we assume some level of confidence

		De	cision
		Accept null Hypothesis	Reject Null Hypothesis
Population Truth	True	CORRECT 1-α	TYPE I Error α
Position	False	TYPE II Error β	CORRECT 1-β (POWER)

The TYPE I error is the probability of rejecting the null hypothesis that is actually true. Making this type of error is analogous to pronouncing someone guilty instead of being actually innocent. α that is level of significance is the largest possible probability of making TYPE I error.

We control the probability of the TYPE I error through level of significance α (1- α is confidence interval). There is always some probability that we decide that the null hypothesis is false when indeed it is false. This decision is called the power of the decision making process. It is called power because it is the decision we aimed for. Remember that we only tested the null hypothesis because we wanted to prove it is wrong.

	Sample 1 (n1)	Sample 2(n2)
Traffic	100	120
Click	5	7
Click through Rate (p)	5/100 = 5%	9/120 = 7.5%

Example: A/B testing of a Page is done with below data

As p1<p2 we can conclude that experiment 2 is better than experiment 1 but we have to prove it statistically. Right?

pt=[(100*0.05)+(120*0.075)]/(120+100)=6.3%

 $\sigma_{p_a-p_b} = \sqrt{[6.3\%*(1-6.3\%)/100+6.3\%*(1-6.3\%)/120]} = 3.2\%$

 $Z_1 = (7.5\% - 5\%)/3.2\% = 0.78$

Assume 95% confidence interval, α =5% (level of significance).

 Z_{α} =+1.96 or -1.96.

As 0.78 is within +-1.96, we cannot reject the null hypothesis that $p_a=p_b$. Visually experiment 2 was looking better but the statistically proven result contradicts the visual result.

Hope this example is useful. Here I have taken CTR as the result; you could have taken conversion or any other success events as the goal. The calculation will remain same but one just needs to change the numbers in the formula.

There are many criteria to be considered for the testing, I am leaving behind two important points:

- 1. When you are doing A/B testing of existing page then it might sometime make sense for the sampling of the returning traffic only and result to be considered based on returning traffic only. This is because both pages will be anyway new for the new traffic and may not make sense from their reaction to the page.
- 2. When you are doing A/B testing of an important page it may be risky proposition to send 50% of traffic through new pages. Instead you can start with 10% of traffic through new experiment and 90% of traffic through controlled environment. If result is encouraging the rise it further by 10% and continue it till traffic through new pages is more than 50%. With more than 50% traffic through new pages and result is positive then you can switch to new pages.

A/B Testing in Google Analytics

The Google Analytics has the mechanism to carry out A/B testing of pages and its variants. Go to Acquisition->Experiments

	, Audience Overvi	ew			20111, 2010 - Aug 10, 2010
Servit month & first	Bruel Deport + Addition	Sealthcard Shortcat			•
Acquister	Al Users		() + AM	Segment	
Details	Overview				
Overview	Season + vil. Same	-			Bourly Bay Stack Hards
Dehavior Flow	· Sessions				
+ Site Content	10.000				
+ Site Speed					-
+ Site Search	24.000			11.1.1.1	
+ Events					
		342	2425	241	AqR
+ Publisher					
• Publisher Experiments					
+ Pullater Experiments 11 (Approxim)	families	Det	Parries		liee Vistor 🔳 Returning Vistor

Click on create experiment button to start a new experiment

	Create aspectment					
Acquistion	Esperiment Name	Startus	Ovtails	Sealora	Start Date	End Date
and a second	Home Page	Sela	Step 3	-		-
Baharar				Des	1000 32 W 000 1	L DOM K.
Oramiere						
Dehalor Flow						
+ Sile Content						
+ Site Speed						
+ Site Search						
+ Eretta						
+ Publisher						
Experiments						

In first Step, Name the A/B Testing, Select a metrics to be measured and mentioned the 100% of traffic to be allowed to pass through it.

Ch, Balant short & hell	Content Experiments - Create a new experiment Choose an experiment objective	Í
😘 Acquister	A Rene for this experiment US68ed experiment	
Detranse Overview Detranse Pere + Site Content + Site Speed + Site Speed	Objective for this experiment is beind a new is in the objective Proceedage of buffic to experiment is interime is consistent ordination for important changes or or or or	
+ Eurota + Publisher Experiments In-Page Analytics	Advanced Options	

In Step-2, add the original pages and its variants.

C. September & See	gure your experiment		
Acquisition Control of	Igned Page of the with drawn provinces are assumed. (1) in page to experiment (1) more for the page regard Consolidate experiment for other parties (1)	Pipe Preview	
Site Speed Site Speed Site Speed Dents Philaber Caperheem InPage Autors W	nan t 0 ht pape to experiment :: nor in the top pape mont 1	Page Preview	

In Step-3, add experiment codes in the original mentioned in the Step-2.



You can copy and paste the codes in the Original Page and send code to web master as instructed.

Q. September 5 Mill	Setting up your experiment code	î
1 Acquister	These do you want his and up your experiences calls? Instant, sever the case, a locate use is constructed.	
B Behavior Overview	Adding uning code to provide and or the page 10 1 More any new expose and varied pages have Code Annual Annual Code Annual Code Annual Code 2. They parts the component page immediately also the page of your anyonic page.	
Dataser Peer + Dis Conset + Dis Speel + Dis Seach + Dires + Publisher Foundation	1.4. Grappia readphilds (Control Dispational roles	
to Page Analytics	Additional information for your experiment (ode: Experiment Gr. 1990;62:770:0100; gu62;0 Experiment Kay: 1996;63:03	

In Step-4 once the codes are added you just need to review and start the experiment



Let the experiment run for time period and check the result for the hypothesis testing

Learning from the Chapter

- Understanding Google Analytics dimensions and matrices
- How to look at the report and explanation of all report tabs
- Cohort analysis in Google Analytics
- How to create goals and funnel and how to interpret the reports
- How to use Google Analytics API to extract data using Pentaho Data Integration
- Concept of A/B testing and how to carry out A/B testing using Google Analytics

Chapter - IV CUSTOMER ANALYTICS

"Being on par in terms of price and quality only gets you into the game. Service wins the game." by Tony Allesandra

A young man was thrown out of a company selling handsets because he couldn't make single sale in the booming market. When he was asked to leave he drove to an industrial estate, sat in his car and cried. He called his mother to say "Tm useless; I need to jack this in." Within two years he was able to buy a Ferrari for himself doing the same job of selling handsets. What exactly he did differently from his earlier job. When he initially tried selling handset the customers would refuse because their contract for the renewal was not immediately due. What he did was list down all the companies he had contacted and note their contract renewal dates. Knowing the exact date of renewal help him contact the right person at right time with better products than the existing products.

"The customer is the king" for any business. The business which doesn't respect their customer and create a customer loyalty will not be sustainable in the long run. The studies has found that it cost way higher to acquire a new customer than to win over existing customer for repeat purchase. Therefore it is imperative that the companies maintain the customer information for understanding their behavior, design marketing offering to suit their needs. For this there is requirement to collect customer information for segmentation and targeting. In the online environment the buyer has to login with certain information like address, email and phone numbers. In case of offline also the stores collect customer's information through different means though data may not be as rich as online data.

All the money spends on the marketing programs are directed either to win new customer or to induce customer for repeat purchase. The effectiveness of the marketing programs is measured either in term of awareness creation that is brand visibility or customer acquisition or increase in sales. Therefore understanding customer and creating the marketing campaign according to customer behavior is more likely to be successful. In ecommerce there is enough data from the web analytics, CRM system, Order Engines and Master Data to create very effective segmentation for any digital marketing campaign.

Understanding of the customer behavior is important for the improving the site design and flow. What products to be displayed, what banner to be shown and where, what products to be recommends and so on are very important decision which are purely done using the customer web analytics information. The companies resort to A/B testing and multi-variate testing to understand the customer preference for color, text, image, products and other contents.

The customer analytics helps the sourcing managers and category managers to prioritize their product offering. What products sells at what price point, what product has better preference in which location, what time of the day a particular category of products sells, out of many products in the catalogue which products are likely to sell more for making promotion decision, what products to be displayed in category page and home pages and so on. All these decision are linked back to the customer behavior data.

The importance of the customer information can be overemphasized. For any company customer information is a gold mine. How company uses would obvious vary. The customer information comes from different applications of a company. It cannot be after thought. The initial design of any systems should be done with customer data capturing in the mind else some of the important data point can be missed. If customer is the most important person and first preference of the company why not customer data collection is first thought in the system design.

In this chapter we will learn how to create customer cohort and analyze it. The cohort analysis will be extended to included calculate the customer lifetime value for the cohort. Cohort analysis is important for understanding the behavior of different types of the customer segments. As mentioned earlier customer segmentation and effective targeting is important for any marketing program we will learn how to segment customer using clustering in R. The prediction of customer behavior is another important analysis for situation like whether these customers are likely to buy, whether these customers are churning out and so on. We will learn how to use logistics regression to predict the customer behavior. Another important data point is the cart data – what kind of products are bought together. The question like what products should be recommended when

customer added product X into cart, which product should be promoted together and so on. The market basket analysis help us understand the customer affinity to the products. The detail of each analysis will be explained with examples in the respective section. One thing the reader need to remember is that each company has different objectives these analysis will give you broad idea of how to approach the customer analysis but there are many other algorithm. In fact there is no end to the customer analysis, one just need to decide what is important and is in line with company's objectives.

Section - I

CUSTOMER ANALYTICS

4.1.1 Segmentation in Google Analytics

In web Analytics Chapter we have learnt about Web analytics especially Google Analytics – tabs and how to interpret the data. In this section we will learn how to create segments in Google analytics. The purpose of the understanding segment in Google Analytics is that it helps in creating a subset of the users in the website with certain criteria for studying behaviors of those user set only. Such subsets or segments can be used for effective remarketing, A/B testing and retargeting. Any web analytics system has means to create segments because the customer behavior analysis needs triangulation of data at various points. For instance if we want to compare the customer flow in desktop and mobile then we have to create desktop segment and mobile segment for comparison. The segmentation creation in Google Analytics is very easy and self-explanatory; however I will take you through few segment creations to familiarize you with the process. The purpose of the section is to make readers aware of the process and how we can use segments for different objectives.

By default Google Analytics show all session segments which includes all sessions in the period. To add any segment click on +Add Segment.

E. Dealth myselfs & here	Audience Overvier	*			Dec 10, 2015 - Jan 9, 2016
Daubtearts	a brust Expert + Addition	alboard Sharlout			•
· Shotcuts	Al Sessions		- AN In	gront	
Intelligence Direnta	Destries				Nurth Bay Starts Starts
Realized	• Sessers				The construction of the second second
Autore	~			~~~~	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Ovendew	100.000				
Active Users					
Active Users Cahort Analysis ^(K1)		Dec 10	Dei 32	Gei 28	3e1
Active Users Cohort Analysis ^{MTM} + Unemographics + Interests		2mi 18	5H 22	04129 • Re	241 furning Visitor 🕷 New Visitor
Active Users Colourt Analysis ^{MEN} + Usersgraphics + Itilizents Overview Affety Categories	5ecces 23,448,811	Uners 14,902,362	Pagesteen 48,429,322	54(2) # 54	Jan 1 Kuning Visiter III New Visiter

In the Segment Panel you will see all the current available segments for the user. There are many inbuilt segments in Google Analytics.

G, Harris morta & hear	Audience Overview			Fe	5 22, 2016 - Feb 22	2016
Durboards	a Beat Equit. Antichan	tere Barbat				11 +
- Stotuts	All Users		Chasse segment have been			
Malipera Denta		matery thereases		~ = 11	1, back spects	a
D Realine	ver selver/s	Al Ders	 Bounced becomes 	Converters -		Î
Audence Overview	NF Ejsten	· Ored Traffic	Enaler	 Geoge OPC 		
Adve Users Oxfort Analysis IPT	Custom Shared	· Son afwitten	None and Table Traffic	··· Hosis Tatla		
- Demographics	Barel	· Multi-second Users	Nextons	10 Not-bource Beasines		
+ Geo		E Non-Convertiers	Organic Traffic	e Pastato -		- 1
Technology Update		 Performed Site Search 	 Releval Traffic 	 Refurring Users 		
+ Oustern	And Emol					

You can create a new segment or add segment from the gallery by importing it. For the custom segments you have to create your own segments as per your requirement. To add any new segment just click on +New Segment

The first of reserving it have	Audience Overview				Feb	22, 2016 - Feb 22	2916
Centosaria	Bast Square Astronom	Read Robal					10 •
Sharkula	Artises		Contra televisi presi del presi				
• Integration Events		angeles Banagaan		Var	= [1]	1. Seech aspects	
D Repfore		Arbun	Bounced Second	Conettes			i
Autoria Overview	Al Dynkers	· Devt Trafic	Enalw	Google CPC			
Adva Users Cohot Analysis ***	Culture	Rade a PortPase	Brow and Talout Traffic	more Total			
- Demographica	Bankt	 Multi-session Users 	heriters	1 Non-Service Second			
+ Geo	1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1	I Non-Converters	Organic Traffic	Pactole			- 1
+ Technology		Performant She Search	Autoral Traffic	Returning Using			
- Shote - Custom							

The segment creation windows contain all the dimensions we discussed in the Google Analytics Chapter – Demographics, Technology, Traffic Sources, Ecommerce and Behavior. If you see the summary in the right hand side of below screen it shows 100% which means there is no filtration.

To Stand Starts & State	Audience Overview			Feb 22, 2016 - Feb 22, 2016 -
an function	A Bear Brant - Adda Barr	and Burbat		21 •
+- thomas	O Al Users		C + Att Beginn	
🕈 Intelligence Events	September 4		ten land Parker	Reports only if any the Darge
Ø Real-Time	Denographics	Demographics		Summery
	Technology	September your scores by a	terregische offertation.	\cap
· Audience	Balante	Apr 1	0 1942 0 25:34 0 25:44 0 45:44 0 19544 0 195-	100.00%
Overview .	Date of First Secolory	danter 1	D'fenge Ditate D'Owneen	d and
Cohot Augusta min	Traffic Bources	Language (1)	annes a	\sim
- Demographica - Internets	Escenaria	alfority Calegory (wash)	(mm 1)	Decement a
- Gee	Conditions	In-Manat Segment 1	(annum (+)	100 DPh of seconds
- Behaver - Technology - Motes	Importan	Other Callegory -	anda a	
- Custom				Stafillaria

We add location segment city = New Delhi in the segment, the summary now shows 20.14% which means the segment with city as New Delhi will contain 20.14% of the total traffic of the period.

D, National Street	* Erest Expert - Antic Deather	a Dated			22 •
E Cartonem	· O AD Users		All Seguet		
Storbulk	- Departure		ten Canad Maxima		begreen to come or any Veron Diargo
• Interpret Conte	Conspience (C)	Demographics			Summary
D Real-Time	Technology	Exprise your uters by	Care Care Care Care Care Care		
a notaria	Cale of First Second	Gender 1	Direnas Dilate Direnten		20.14%
Overvee	Tuffclouros	Larguege ()	and a		
Cahert Anatons ***	Econeca	Afterly Category (Health)	(mmm +)		175
- Demographica	AbsActed	initiated Segment	and a		202 11.175 of second
> Geo	Separate	Other Calegory (and a		Demographics 7
- behaver - Sectoratege		Location ::	Ny - main - NavQells]+	
- Matolike				Þ	
+ Custom	Overview				

Now we select 25-34 years and 35-44 years in the age dimension (25-44 years) the summary shows 4.26% of the traffic.

To Security Advances	" Bear Barry Adda Barting	ant Martial			2 e
E Custown	Al Users		O + Add Segment		
++ Dotob	Inprectane		teres Parter		Report a collect any feer through
mangana Carata Asar Tana Asar Tana Asar Sana Asar Sana Asar Sana	Demopration () Technolog Behave Date of Fast Tecation Table Sources	Conceptantics Depret your store by Apr 1 Gender 1 Language 1	benografie eternation □ 1926 18 2926 18 2944 □ 4946 □ 594 □ Renge □ 1936 □ (1996aus 1999au - 1	4 C 4+	4.265 d'aires
Cutot Anapos ***	Examiners Adversed Conditions Betweenee	Mints Calegory (mail) In Hannat Tegment (Other Calegory ()			37 Excelore 43 45 A December Demographics Age 2016 (Merc City services Van Defer
+ Technology + Wolfe + Coatom		Location ::	(ty -) main -) tax Calls		Þ

Let's us create another segment with medium as Organic and name it Organic Traffic.

States and Advances	* Enal Equit- Articlation	and Married				2 •
E Deshtoorts	Al Users			- Add begree	4	
++ Statute	Seprentiane			Canal Param		Segments with it any their (harps
• Independents	Denoptation	Tatle lour				Summary
O Basilina	Technology	Express yes	users by how t	they finand your		
	Belavior	Filter Incoming	file lines			30.61%
🛓 Audence	Cale of Fest Section		in the second second			d men
Overview	Tufic Devices 💿	Chutheda /	compte -			
Adhe Users Cahot Analysis Mile	Economica	Statum 1	-	esenc	0	266
- Demographics	Adapted	Serie 1	-			Second and
- Manufa	Conditiona	Report 1				34.30% of severates
- Gere	Deguences					Traffic Sources
- Behavior						
+ Technology						
- Mode						
+ Custom						
+ Benchmaning	Overview					

Now we have used Organic Traffic segment, all the information in any tab will show only organic traffic.

A. Second married to be	Audience Overview					Feb 22, 2016 - Feb 22, 2016
. Custown	A Shart Ergent - Anti-Da	attand Barbad				88 •
- Potos	O Coganic Traffic		· O ·	Add Segment		
antigence beens	Destries	terns .				Territy Day 1000 1000
Austine	· Sessions					Rosenid
Andence			1000		~	
Overview Active Upera	<				-	~
Overview Adhe Users Cahot Analysis ***	1	10.40		411.00		10.70
Overview Active Users Cotront Analysis ^{with} > Demographics > Interests	-	t al sur Vers Descen Tarle	Times	Agent Page	610 Po	10 Mc Returning Under
Oversiew Adlus Usaris Calculat Analysis *** > Demographics > Demographics > Demographics > Demographics > Demographics > Demographics	Seasons Organic balls 375	tal.op User Dipart Safe 266	nn-er Poperane Dipare Tuño 8,110	Visit Page Pages (Second Organic Traffic 21.63	CE Pr	ED N/ Existing Value - Exercicate
Dearteau Actua Usors Control Analysis Illino - Demographics - Demographics - Demographics - Demographics - Behavior - Technology	testine Operations 375	tarian Opera Sufis 266	rome Processes Oppose Trade 8,110	rate for Paper Floration Organization 21.63		Ein re Biskeng teler Biskertuker
Countralieur Action Losans Contol Analysis ⁽¹⁰⁰⁾ - Consignations - Sens - Bahavann - Bahavann - Bahavann - Bahavann - Bahavann - Bahavann - Bahavann	testime Dysec ball: 375	naf no User Organis Suth 266 Brance Rate Organis Suth Organis Suth Organis Suth	EIIIII Papenese Organis Tanti 8,110 Nature Despans Organis Tanti Organis Tanti	Vite Provide State	- 11 PU	Ein re Bistong tabe Bistoriuster

To compare two segments add segment in segment bar. Here I have added All traffic and Organic traffic segment. The data comparison of both segments is shown across. This is very useful features for the comparison across the segments. We can compare more than two segments.

0, 10411 10111 1 101	Audience Overvie	*			Feb.	22, 2016 - Feb 22, 2016 -
III Continents	A Beat Equit- Adda.D	etters Barbar				8 *
++ thotats	Organic Taffe		0	All Charts Till Bith Search	- (x)	Þ
• instigator Early	Overview					
Q Realities	Descens (Digent fra	no · Sealors (HV)+	ni -			CONTRACTOR OF CONTRACTOR
Autoros Orantes Atter Chert Catert Anglans III - Cerrographics	5	10.40		UNI-PE	*() Fu	
+ tribera pla	Seasons Organic Traffic	Uses Organic Traffic	Pageneses Grane Traffic	Pages / Session Organic Traffic	Organic Traffic	AR Users

In any tab you will see the comparison of the segments selected.

N, Section county is here.	112	Baarias diversor +	tet fam. Setup + 1					C		atariat [II] @	1.6.3
			Angeleter			Balantar				erren i	
O feature		Sector Campoy	Beneral de	t.ter because	Tee Itees	810000 Fam	Pages / Berniet	Any beater Denter	****	Received -	Conserved and Server a
Autoria Oranian Attac Usars Const Anauss - Conseption - Standts	Ē	Organic Traffic	375 9.0 mm	45.87%	172	10.40%	21.63	00:10:15	1 1.112 1.112	Rs.307.00	0.27%
	L	All Deers	1,073	61,88%	664 5 x 7 10 x 10 10 10 10 10	19.94%	17.09	00.07.46		Rs.1.306.00	0.19%
+ 040		1. evenue									
- Behavior		Organic Traffic	204 (10.07%)	48.12%	10.0100	10.14%	7.30	00.08.21	1,001000	Ra 307.00	0.475
+ Technology		AllUsers	847 co uno	44.00%	-	34.00%	7.81	00.05.43	2.000	Ra 1,306-00	6.34%
+ Webble	0	1 mobile									
Oversieve		Organic Traffic	167 (10.07%)	41.52%	79.000	110%	38.53	62 15 14	81.000	Re201-114	1075
Destas		All Using	458 (47.5 (10))	\$7.56%	208 (10.014)	222%	38.16	40.16.52	811000	#a.8.00 in res.	6.02%
Custom Sendimating Does Fire	0	3. Solut Organic Traffic	4,0000	75.00%	3.114	25.00%	5.00	00.00.54	\$	Ra. 2.30 - 5 (mm)	0.00%
-		40110am	Materia	15 MM	48.0000	1476	4.47	65.64.45	A	#x.8.00	0.00%

Google Analytics user can create as many segment in the profile for analysis and reporting.

4.1.2 COHORT ANALYSIS

At some point time in life, many of us would have spent some time comparing the achievement of the people from the same batch in the college - how many of them are working & in which organization they are working, how many of them are entrepreneurs and how many of them are married, how many of them are still bachelor and so on. Right? What we are doing here? Basically we are doing a comparative analysis of group of people having similar characteristics; the characteristics is people pass out from college Z in year XXXX. Knowingly or unknowingly you are doing a sort of cohort analysis in the process.

In a simple term, Cohort is defined as group of subjects sharing similar characteristics; characteristics can be time based like people sharing similar events in a time period or other features of the subjects like gender male or people from city A or people born in a particular year. In statistics the time based characterization is most popularly used. In this book we will be heavily using time based characteristics like customer purchase in particular month, customer purchase on particular offer period and using the offer code and so on. There are many ways of defining a cohort but make sure that the cohort is meaningful and significant. In this section we will revisit Google Analytics Cohort and then move on to customers' cohort from the purchase behavior. We have already gone through the different options available in the Google Analytics for cohort analysis; we are just revisiting to look at the format and how to interpret the data from the cohort table.

	Day 9	Dep 1	Oxy 2	Dep 2	Dec.+	Day 5	Day 9	Day 7	Day 8	Day 9	Day 10
All Users	100.00%	3.89%	1.76%	1.11%	0.51%	0.38%	0.35%	0.10%	0.59%	0.20%	0.11%
Feb 9, 2016	100.00%	2.03%	1.71%	1.00%	6.97%	1.0%	1.07%	0.22%	0.12%	0.17%	0.22%
Feb 10, 2016	100.00%	2.02%	2.47%	1.14%	0.49%	0.49%	0.49%	0.32%	0.24%	0.00%	-
Feb 11, 2015	100.00%	1.0**	1.57%	1794	0.29%	5.0%	0.29%	9.22%	1.12%	6.00%	0.14%
Feb 12, 2016	100.00%	1.03%	1.13%	1.05	0.22%	0.12%	0.13%	4.0%	0.09%	1.07%	0.04%
feb 13, 2016	100.00%	2.00%	8.75%	0.27%	4.38%	827%	0.32%	.0.17%	0.17%	1.075	1.07%
Feb 14, 2016	100.00%	1.40%	0.02%	1.25%	0.10%	0.47%	0.28%	0.10%	-	0.00%	
feo 15, 2016	100.00%	1.00%	1.57%	1.0%	4.19%	1.01%	0.12%	4.19%			
Feb 16, 2016	105.00%	1778	0.5-76	0.54%	67%	8.71%	634%	630%			
Feb 17, 2016	100.00%	8.17%	105	3.09%	1.20%	-	0.00%				
Feb 18, 2016	100.00%	13.40%	3.77%	3.859	2.30%	105					Q.
feb 19, 2016	100.00%	1.05	4.80%	1.03	0.00%						
F#0 29, 2919	100.00%	1758	2.49%	0.00%							
Feb 21, 2016	100.00%	8.3em	0.00%								
feb 22, 2016	100.00%	0.00%									

In this table for 9th Feb 1344 users are in the selected cohort. For this row 9th Feb is day 0. On day 1 that is 10th Feb 2.53% of users from 1344 that is 34 users again completed the goals or events selected. On day2 that is 11th Feb 1.71% of the 1344 users that is 23 users again completed the action but remember that 23 is from 1344 not necessarily from 34 users of 10th Feb. The all users is the weightage average of the cohost for 14 days. This cohort clearly shows how a particular set of users repeat the same action in different point of

time. Here the characteristics of population for cohort is the people completing a particular event on day 0.

Let us look at a cohort from real environment. Below is a cohort from company XYZ. Assuming that company start working from Oct 2015, the cohort has been prepared till June 2016. This is month wise cohort with first purchase as the date of joining the site. Some of the basic interpretation from table

- In October 2015, 17899 people purchase for the first time in the site. Out of those 17899 customers, 2633 came back in November 2015, 2294 came back in December 2015 and so on. Total customer base were 17,899 and new customer are 17,899 as there is not repeat customer in the first month.
- In Nov 2015, 34,908 new customer purchase from site and 2633 existing customer purchase from site making total unique customer as 37,541. So 7% of customers were existing customer and 93% customers were new customers. Total customer base became 52,807.
- In June 2016, 16,165 new customer purchase and 17418 existing customer purchased. Total customer base for 9 month of operation became 329,723.

Customers Count	Dec-15	Jan-16	Feb-16	Mar-16	Apr-16	May-16	Jun-16
Dec-15	67,727	10,809	8,561	6.580	5,765	5,254	2,474
Jan-16		36,465	8,198	5,807	4,508	4,106	1,984
Feb-16			31,838	6,739	4,896	4,331	1,888
Mar-16				31,445	6,297	4,853	2,132
Apr-16					29.353	6,056	2,495
May-16						63,923	4,309
Jun-16							16,165
Repeat Customer	0	10.809	16.759	19,126	21,466	24,600	15,282
New Customer	67,727	36,465	31,838	31,445	29,353	63,923	16,165
Repeat Customer %	0.0%	22.9%	34.5%	37.8%	42.2%	27.8%	48.6%
New Customer %	100.0%	77.1%	65.5%	62.2%	57.8%	72.2%	51.4%
Total Unique Customer for month	67,727	47,274	48,597	50,571	50,819	88,523	31,447
Cumulative Total customer Base	67,727	104,192	136.030	167,475	196,828	260,751	276,916

Now let's look at the Retention rate, some of the observations are

- 14.7% of the new customer in October came back to site in November. This is calculated as 2633 / 17899.
- In December 12.8 customers from October New customer came back, 14% of November new customer came back.
- The rate of retention from a Cohort decreases as time passes. In 9th month the customer who join site in Oct 15, 3.4% came back.

Retention rate	Dec-15	Jan-16	Feb-16	Mar-16	Apr-16	May-16	Jun-16
Dec-15	100.0%	16.0%	12.6%	9.7%	8.5%	7.8%	3.7%
Jan-16		100.0%	22.5%	15.9%	12.4%	11.3%	5.4%
Feb-16			100.0%	21.2%	15.4%	13.6%	5.9%
Mar-16				100.0%	20.0%	15.4%	6.8%
Apr-16					100.0%	20.6%	8.5%
May-16						100.0%	6.7%
Jun-16	1.1.1						100.0%

We can further extend the cohort to include number of orders, revenue and average order per customer, and average revenue per order and so on. Sample is given below

Numbe of Order	Dec-15	Jan-16	Feb-16	Mar-16	Apr-16	May-16	Jun-16
Dec-15	75,858	26,656	35,350	31,019	22,959	20,874	8,900
Jan-16		55,399	35,806	28,115	18,392	16,280	6,812
Feb-16			59,949	27,682	18,399	15,651	5,851
Mar-16				54,420	22,754	16,689	6,522
Apr-16					49,832	21,207	6,960
May-16						87,561	9,451
Jun-16							20,285

Revenue cohort

Revenue	Dec-15	Jan-16	Feb-16	Mar-16	Apr-16	May-16	Jun-16
Dec-15	219,916,230	67,776,205	73,642,202	51,193,022	37,065,164	28,950,751	9,709,302
Jan-16		119,053,267	70,010,507	47,656,284	29,965,346	22,846,204	10,142,270
Feb-16			103,401,588	48,762,489	29,948,528	21,650,989	6,872,288
Mar-16				90,208,819	35,894,551	21,347,991	7,101,481
Apr-16					71,372,420	27,152,165	8,301,803
May-16	S	S				138,044,683	11,106,608
Jun-16							33,089,025

Average Bill Value

Avg Bill Value	Dec-15	Jan-16	Feb-16	Mar-16	Apr-16	May-16	Jun-16
Dec-15	2.899	2,543	2.083	1,650	1.614	1.387	1,091
Jan-16		2,149	1,955	1,695	1,629	1,403	1,489
Feb-16			1,725	1,762	1,628	1,383	1,175
Mar-16				1,658	1,578	1,279	1,089
Apr-16					1,432	1,280	1,193
May-16						1,577	1,175
Jun-16							1,631

Average bill per customer

Avg Bill Per Customer	Dec-15	Jan-16	Feb-16	Mar-16	Apr-16	May-16	Jun-16
Dec-15	3,247	6,270	8,602	7,780	6,429	5,510	3,925
Jan-16		3,265	8,540	8,207	6,647	5,564	5,112
Feb-16			3.248	7.236	6.117	4,999	3.640
Mar-16				2,869	5,700	4,399	3,331
Apr-16				1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1	2.432	4,484	3.327
May-16						2.160	2,578
Jun-16							2.047

As we can see form the above examples there are many way of developing cohort. We can clearly see trends for each cohort and then cohort wise comparison of the key parameters. I would urge readers to make at least 5 inferences out of above cohort tables. Now that we have learn what cohort is, let us do a calculation on small data set to understand the underlying steps in the cohort table preparation. Here again I have used month-wise cohort. As an analyst you can have different time window based on your requirement.

4.1.3 Cohort working with Excel

The cohort analysis from the Google Analytics just seen about has one major drawback that is we do not have complete information of each of the customer visiting the site. The site usually captures the customer detail from the login or at the checkout where customer fill-up the form containing the details. Any ecommerce site typically have conversion rate of as low as 0.1% to as high as 5-6%, which means that 95% of the customer in the Google Analytics cohort data are not identifiable as they have not placed the order or logged-in. those customers for which we do not have identity cannot be really used for many of the marketing program such as sending Email, Sending Text SMS, sending newsletter etc.

However the data available from the order contains information about the real customer who has purchase hence doing analysis on these customer is worth it. We have to live with the facts that all website visitors will not provide their identity for sure, so we can exclude them.

Let us start with simple example of a sales data. The real life data will be very large; our motive here is to understand the process of making a cohort and how to interpret it. For the large data set it is advisable to make cohort in the database itself rather than in excel. In data below we have transaction data from the POS/checkout with information about Transactionid, customerid, date of transaction, quantity and Amount. Here customer ID is unique identified to link to the larger customer information.

RL KM	noitti fac	ELANDUT FORM	4.14) Ok14	NYCH VOR								Sign in
A Ker	Caller # / W	- [4 - 2 - 2 - 4		er. Dimart	ati h Carta	- 5 - % + 3		22	2	X Adolan R Sat- Com-	27 M	
ilghood	6	Fard .	4	Algebore		is matter	4	Uphs .	(alta	- 64	nng	
W18		2. 0. 5										
- A		¢	. e.				- MC-	1. COM		- e		M
Transactio	ni Customenio	Date of Trans	Quantity	Amount		TransactionID	Customeril	Date of Transaction	Quantity	Amount		
10000001	CT10001	15-Apr-15	,	1800		TRX0025	CT10008	31-Aug-15	1	5800		
11000007	CT100001	21-May-15	- 6	4200		19040014	CT10009	30-Jun-15	- 3	350		
TR000016	CT10001	22-34-15	1	4500		TRX0027	CT10009	5-5ep-15	- 6	3600		
18060029	CT10001	8-5ep-15	2	8000		TRX0017	CT10010	25-34-15	3	1500		
TR000002	C110002	17-Apr-15	1	1000		TR000021	CT10010	15 Aug-15	1	680		
TR060022	CT10002	15-Aug-15	4	2000		TRXD018	CT10011	28-34-15	12	24000		
11000003	CT10003	20 Apr-15	1	2000		TRX0020	CT10012	1-Aug-15	1	250		
17060006	CT10003	4-May-15	5	3000		19060024	CT10013	23-Aug-15	5	10000		
TR060009	CT10003	30 May-15	2	10000		TRXID028	CT10013	6-Sep-15	4	1600		
1000004	CT100004	25-Apr-15	4	1200		19040026	CT10014	3-5ep-15	1	4000		
11010005	CT100005	4-May-15	2	2400		18060030	CT10015	17-5ep-15	1	2400		
15060012	CT10005	16.365-15	. 4	2400						-		
18060019	CT10005	31-54-15)	4500								
11060008	CT10006	23-May-15	1	5000								
18060000	CT10006	10-Jun-15	7	49000								
TR080023	CT10006	19-Aug-15	1	450								
TR060001	CT100007	15-Jun-15)	2100								
18060015	C1100007	6-348-15	2	800								
TR000013	CT10008	20-Jun-15	20	20000								

Sort on CustomerID and then by date of Transaction as shown below.

	A Line Office	a Batting a	Convert	1 (and)	Art Statement Construction Cons	Conter Lander	
43	• (*)	,6 TRX00	65				
anadatt.	Common D	tions of frameworks	lowering.	Anne			x
4+0004	1010083	15-64-X	N	No. Later			
100012	3750062	L ^b Apr 1	A	a labor		1 1000	
10000	Country of the local division of the local d	20-April	1	Ann	Sort		
5x0004	271080	25-Met X		1.00	The second	(paper and construction)	
1 miles	3730HB	4-14er2		200	"Titution Advances addresses v . Open"	K. Hy dete he geoders	
140005	SCHOOL	4-5/M/ X		1000	Antar Antar Antar	and a second s	
Ex0007	1710988.	().49er/	10000	4,000	and our		
1-0004	2730888	19-11-12		North Street	perter CurbrelD (r) Telus (r) A to 2	(M)	
	S. North	Notine 1	1 million (about 1	Tentile Date of Tanas Tentes () Orbot to fee	and with the	
LADESE	22,000	38-549-2	1 ······	4800			
440011	2730987	iban (S	2089			
140014	27,0008	16.003	-	240			
1000	St. Name	20.00 3		Anne			
1000 A	57.0000	N-brt I	-	100			
10003	ACCREASE OF	+ NC 1	h	800			
	proves	22.MCX		4000		Construction of Construction o	
1000	20,000	(B-N-3		194	0	Cartost	
-	PC PT P	anex.	6 0	1000			
-	1. A	- BACI	-	-			
-	C.C.C.	244		- 21			
Contract of Contra	Change	1040	-	-			
-	Crusse			-			
-	Contraction of the local division of the loc	Children of Childr	-	-			
August 1	Column	Di Ang I		-			
100.00	a Partie	A Sec. 1	-	-			
August.	AT LODGE	b-Sep-1	-	a annu			
840628	STAMA.	6140-1		· Long			
100.00	Values.	A lan i	-	and a state of			
Indexe	ATUMEA	17.049-2	-	2400			
				the second se			

Let's create number of purchase of customer. First purchase as 1, second purchase as 2 and so on.

CH DIO	NORT R	AGELENDAT PORMAN	Ce VEW	NATION IN	of .								4 30	- 6	ň		
R Kor		$- u \rightarrow \chi^{\prime} \chi^{\prime}$ $\epsilon + z \rightarrow A + c$ $\epsilon + c$		- 3- 	nge Tani Inge Se Cantar - 1	- 15 + Autor	ч.а.	1 <u>F</u>]	100 100 100 100 100 100 100 100 100 100	211	in The state	No.	Charles	- 41 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	A		
3UM		3 4 5 4	181-82/2-63	0													
A		٤				4	110	н		12.5					- 14	- N	2
1 Transactio	- Customei	Oute of Transact	Qyantity -	Amoun -	No fo Purch -		6	- (a)				(*)					
2 TR0600001	CT10001	15-Apr-15	5 3	1800	1					1							
3 TRX00007	CT10001	21-May-15	5 6	4200	+18(0.3+82,57	+1,1)											
+ 19X0056	CT10001	22-348-15	1	4500	, ,												
3 TRX0029	CT10001	8-Sep-15	5 2	8000	4	1											
 TR060002 	CT10002	17-Apr-15	1	1000	1												
> TRX0022	CT10002	15-Aug-15	4	2000	2												
a TRX0003	CT10003	20 Apr-15	1	2000	3												
TR00006	CT10003	4-May-15	5 5	3000	2	1											
 TRX0009 	CT10003	30 May 15	5 2	10000	1												
H 1880004	CT10004	25-Apr-11	4	1200	3												
1 TRX0005	CT10005	4-May-15	2	2400	1					1							
TR000012	CT10005	16-Jun-11	S 34	2400	2	0				1							
14 TRX0019	CT10005	31-04-15	5 3	4500	1												
IS TRACOODE	CT10006	23-May-15	5 1	5000	1		_										
TR00000	CT10006	10-Jun-11	1 7	49000	2	1											
TRX0023	CT10006	19-Aug-15	1	450	1												
I TROKOOD1	CT10007	15-Jun-15	5)	2100	3			_									
·· TRX0015	CT10007	6-348-11	2	800	2												
70 TR060013	CT10008	20 Jun 11	20	20000	3	1											
100 1 1 1 1 1 M	1					-				-							- 1

All the order has been assign Nth purchase of the customer.

CH St C	 	and the second	1000	-	C.e	of Not-Lot						1 10	1.2.2
Rate Stand		$\frac{1}{2} \frac{(\mathbf{u}_{1} + \mathbf{x}_{2})}{\mathbf{x}_{1}} \frac{\mathbf{x}_{2}}{\mathbf{x}_{2}} \frac{\mathbf{x}_{2}}{\mathbf{x}_{2}} \mathbf{x}_{1} \mathbf{x}_{2}}{\mathbf{x}_{2}} \mathbf{x}_{2} $		- 5- - 5-	Ingelast (in Angelastense + 3	- % + %		Parte Sta	that been	E KA		A	
N00	•	$ A_{-} = A_{-} $											
A				101.8		- 16 - 11	H				C 1 4	- 14	1.0
Transactio	Customer -	Date of Transact -	Quantity -	Amoun	- No fo Purch -	(4)	(a)	- 54		(4)			
3 140600001	CT10001	15-Apr-15		180	0 1								
1 TRX0007	CT10001	21-May-15	6	430	0 2								
a 18060056	CT10001	22-346-15	1	450	0 3								
TRX0029	CT10001	8.5rp-15	1	800	0 4								
 TR000002 	CT10002	17-Apr-15	1	100	0 1								
1 14060022	CT10002	15-Aug-15	4	200	0 2								
TR000003	CT10003	20 Apr-15	1	200	0 1								
 TR000006 	CT10003	4-May-15	5	300	0 2								1
10 11060009	CT10003	30-May-15	1 2	\$000	0 3								
1000004	CT10004	25-Apr-15	4	120	0 1								_
11000005	CT10005	4-May-15	1 2	240	0 1								
1 TR060012	CT10005	16-Jun-15	1. 4	240	0 2								
+ 15000019	CT10005	31-04-15		450	0 3								
1 TRADOOD	CT10006	23-May-15	1	500	0 3								
TRXX0050	CT10006	10-Jun-15	1	4900	0 2								
TROUDOUR	CT10006	19 Aug-15	1	45	0 3								
· THOROOD1	CT10007	15-Aut-15)	210	0 1								
- TROKOODS	CT10007	6-348-15	2	80	0 2								
IN TROBUCCE	CT10008	20 Jun-15	20	2000	0 1								

Let us take first purchase as the date of Joining for the customer to the system.

OH 5 C	** *)	e server a server		New York	4	parties bee								1 10	- 6
RE HOME	NORT F	ACELANDUT PORMULAS	24/16	REVER VEN											Synm
A Car	Caller	-[n-r,d,d]	* = _ (4	- Drawla	÷ 1	Carlon	- B	1.12	2	2-2-	1	E Adata Tana	• · §7	*	
Farm Strengt		FILE DIAL	0.000	at Sheep A	Carton 1	\$1.56 + 151	A Sente	and Turnet	- 54	Intern Series 6		Con.	in the second	Feeth	
Options	4	Aut 1		Angenet St.		marine .		Uples.		Com.			times.		
	101														
5.4			(PECON)												
A		6					H.						6		N.
Transactio	- Customer-	Date of Transact -	Qyansity -	Amoun - No	fa Purch	- Joining Du -	- 14	1	- 14		- (*)				
1 14060001	CT10001	15-Apr-15	3	1800		1 15-Apr-15			_						
1 1600007	CT10001	21-May-15	6	4300		2 -04(13+1,0)	(07)		_						
a 18060056	CT10001	22-34-15	1	4500	_	3									
 TR000029 	CT10001	8-5ep-15	2	8000		4									
 TR060002 	CT10002	17-Apr-15	5	1000		1									
F 14060022	CT10002	15-Aug-15	4	2000		2			-						
TR00003	CT10003	20 Apr-15	1 1	2000		1			-						
 TRX0006 	CT10003	4-May-15	5	3000		2		_	_						
1000009	CT10003	30-May-15	2	10000		3			_						
100004	CT10004	25-Apr-15	4	1200		1			_						
1 1680005	CT10005	4 May 15	2	2400	_	1			-						
1 18060012	CT10005	16 ho 15	4	2400		2	_		-						
10000029	CT10005	31.5.6.15	- 1	4500		1			-						
TEXACOR	CT10006	21.May 15	1	5000		1		-	-						
TEXODOD	CT10006	10 80 15	7	49000		2			-						
18060023	CT10006	19 Aut 15	1	450		3		-	-						
· TIOROOT1	CT10002	15-8-0-15		2100		1		-	_						
TENDORS	CT10002	6.56.15		800		2			-						
TEXADOR	CT10008	10 he 15	10	20000				-	-						
	1.110008	20 805 15				3			-						

Basically for one customer all subsequent purchase will share same data of Joining as the first purchase.

CI H St I	and a	AND AND A DOMAGN		and a		Not Bus - Deal				4.36	- * *
Para Cupy -	5000 1 1 1	-(<u>1</u> -1) <i>X</i> X + = = <u>1</u> + <u>1</u> + <u>1</u> + Net 5		- 3-00 - 200	erter (6 gescene - 1		Parata Col Salar Sylar Syla	Σ. 	ation - An - Lin terry	A forth	
9423	· •	2. 1. 6									
A		£		. I		4		 	4	м	N .7
1 Transactio	- Customei -	Date of Transact -	Qyantity -	Amoun -	No fo Purch -	Joining Dr -	 14	(1)			
2 18060001	CT10001	15-Apr-15	3	1800	1	15-Apr-15	 				
1 18000007	CT10001	21-May-15	6	4200	2	15-Apr-15					
« TRX0016	CT10001	22-348-15	1	4500		15-Apr-15					
5 TRX0029	CT10001	8-Sep-15	2	8000	4	15-Apr-15					
 18060002 	CT10002	17-Apr-15	1	\$000	1	17-Apr-15					
7 TR060022	CT10002	15-Aug-15	4	2000	2	17-Apr-15					
a 18060003	CT10003	20-Apr-15	1	2000	1	20-Apr-15					
+ TR000006	CT10003	4-May-15	. 5	3000	2	20-Apr-15					
·· TR000009	CT10003	30-May-15	. 2	\$0000	3	20-Apr-15					
11 TR060004	CT10004	25-Apr-15	. 4	1200	1	25-Apr-15					
0 18060005	CT10005	4-May-15	2	2400	1	4-May-15					
18060012	CT10005	16-hrs-15	- 4	2400	2	4-May-15					
14 TR000019	CT10005	31-348-15	3	4500	3	4-May-15					_
1 TRADOOR	CT10006	23-May-15	1	5000	1	23-May-15					
IN TRACOUSD	CT10006	10 Aut 15	7	49000	2	23-May 15					
17 TR060023	CT10006	19-Aug-15	1	450	3	23-May-15					
10 TR060033	CT10007	15-Jun-15	3	2100	1	15-Jun 15					
·· 18060015	CT10007	6-34-15	2	800	2	15-Jun 15					
IN TRADOLS	CT10008	20-hm-15	20	20000	1	20-Jun-15					

Extract month name from the date of joining to get the month of joining.

0	H 5 C	NORT 11	ISLANDA PORMUM	0474	NUMBER OF	си Ки	Autility for	3						1 10	- 8	ň
110	X Car RE Casy - F Lunch Calcul		- (N - K - K - (D - A - Aut -)		5	urbai geti Cartar I	S - No + No Notes	- IE	1	28.6	8	2 4.44 2 14 1 2 Cap	- 22 22	A		
344	6		X 4 5 4	0x7(123,"mmm")	1											-
1		· ·	4				0.10	COLUMN ST				ж.			11 N	
+ 3	inansactio	- Customer -	Date of Transact -	Quantity -	Amoun -	No fo Purch	- Joining Dr -	JoinMa -			(4)					
20	1000001	CT10001	15-Apr-15	3	1800		1 15-Apr-15	Apr								
1 1	800007	CT10001	21-May-15	6	4200		2 15-Apr-15	+TIDCH(S.)	"mmm"							
4.1	800056	CT10001	22-34-15	1	4500		3 15-Apr-15									
3.3	8060029	CT1000t	8-5ep-15	2	8000		4 15-Apr-15									
+ 3	1000002	CT10002	17-Apr-15	1	1000		1 17-Apr-15									
2.3	1000022	CT10002	15-Aug-15	4	2000		2 17-Apr-15									
10	100003	CT10003	20 Apr-15	1	2000		1 20-Apr-15									
+1	800006	CT10003	4 May 15	5	3000		2 20-Apr-15									
-1	1000009	CT10003	30 May 15	2	10000		3 20-Apr-15									
110	1000004	CT10004	25-Apr-15		1200		1 25-Apr-15									
123	1000005	CT10005	4 May 15		2400		1 4-May-15									
10	1060012	CT10005	16-Jun-15	- 4	2400		2 4-May 15									
143	8000019	CT10005	31-54-15	3	4500	1	3 4-May-15									
-10	8000308	CT10006	23-May-15	1	\$000		1 23-May-15									
143	1000030	CT10006	10 Jun 15	. 7	49000		2 23 May 15									
103	1000023	CT10006	19 Aug-15	1	450		3 23-May-15									
43	100001	CT10007	15-hitt-15		2100	-	1 15-Nn-15									
93	16060015	CT10007	6-34-15	2	800		2 15-Jun-15									
10	1000053	CT10008	20 Jun 15	20	20000		1 20 Jun 15									

Similarly find the month of purchase from the date of transaction. The first transaction will have same date of joining and purchase date.

CH ST		and and and a		-		ut nut -bot	1					1 10		ň
Fam Kont	• • • •	- (n -),< x + = - <u>0</u> + <u>4</u> +		- 5-	utur ja utaliene + 3	- 1. 1. 12	-		in the face	X Anton	- 27	A		
Options	6	Aut is		Algente		n,eise	4	99,94	045		imng .			
43		2 0 5												-
2. 4						0.00			S				- N	
Transactio	Customer -	Date of Transact -	Quantity -	Amoun -	No fo Purch -	Joining Dr -	JoinMa	- PurchaseN -						
14060001	CT10001	15-Apr-15	3	1800	1	15-Apr-15	Apr	Apr						
14060007	CT10001	21-May-15	6	4200	2	15-Apr-15	Apr	May						
18060056	CT10001	22-34-15	1	4500	3	15-Apr-15	Apr	Pul .						
TR060029	CT10001	8-5ep-15	2	8000	4	15-Apr-15	Apr	Sep						
TR000002	CT10002	17-Apr-15	1	1000	1	17-Apr-15	Apr	Apr						
18080022	CT10002	15-Aug-15	- 4	2000	2	17-Apr-15	Apr	Aug						
TR00003	CT10003	20 Apr 15	- 1	2000	1	20-Apr-15	Apr	Apr.						
1KX00006	CT10003	4.May 15	5	3000	2	20-Apr-15	Apr	May						
TR000009	CT10003	30-May-15	2	10000	3	20-Apr-15	Apr	May						
11000004	CT10004	25-Apr-15		1200	1	25-Apr-15	Apr	Apr						
11000005	CT10005	4 May 15	2	2400	1	4-May-15	May	May .						
1 TR060052	CT10005	16-Jun-15	4	2400	2	4-May 15	May	2un						
a TAXX0019	CT10005	31-34-15	3	4500	3	4-May-15	May	Pul I						
TROCOCOB	CT10006	23-May-15	1	\$000	1	23-May-15	May	May						
TROKOODD	CT10006	10 Jun 15	. 7	49000	2	23 May 15	May	Jun						
TR06002.8	CT10006	19 Aug-15	1	450)	23-May-15	May	Aug						
TR060011	CT10007	15-No-15	- B	2100	1	15-Nn-15	Run	2 Min						
B TR060015	CT10007	6-348-15	2	800	2	15-Jun-15	But	b _i d						
TROUDOS 3	CT10008	20 Jun 15	20	20000	1	20 Jun 15	high	Jun						

Run a pivot.



Add Join Month on the row and purchase month on the column and the any other column on the value and make it count. Now you can see triangular shape data appearance. This shape can be compare with the Google Analytics cohort table.



Calculate the percentage of the customer count from first month to the all subsequent month. For example in April we have 4 customers this is base for the April cohort. Now all purchases done by these four customers in subsequent month are divided by base month to get the cohort in percentage term as shown below.

onnal	Page Break Pa Preview Lay Workbook View	ge Custom out Views	S Ruler	□ Form	nula Bar Sings	Q Zoom	100% Zoom	orn to ection V	New
-	A		8	c	0	E	F	G	-
R	unt of Custome	Apr	in Labels 💌	May	Jun	Jul	Aug	Sep	
Ar	M		4	3		1	1	1	
M	34			2	2	1	1		
Ju	n				3	1	1	1	
Ju	1					2	1		
AL	le la						2	1	
Se	P							2	
Gr	and Total		4	5	5	5	6	5	i.
M	onth	Apr		May	Jun	lut	Aug	Sep	1
A	M		100%	75%	0%	25%	25%	25%	
M	зү			100%	100%	50%	50%	0%	
Ju	n				100%	33%	33%	33%	
Ju	1					100%	50%	0%	
A	8						100%	50%	
Se	P							100%	

Similarly we can create cohort for the Amount by adding the amount in the value fields as shown below. From these cohort table you can derive monthwise number of new customer are old customers. All numbers on the diagonal are new and all data above diagonal are for repeat purchase. In below sheet we have calculated new customer and old customer's % month wise and AOV month wise. This information is quote useful for understanding the customer loyalty and effectiveness of marketing campaign.

HL HOME IN	s VSRT FACE	union.	ion	u.e	0474	ADV	in i	GANTINA	-Base						4.16	- 0) 3400]
Racial Carl	• / *	+]0 - ⊞ + 7+0	-X 2+ A	1.0	- 1	8-1 42.42 249	-	ng-Taul (Seried) nga-3-Center + S + No. 1 No. No.41	W S Contractor	Panet a light	233	at bons I	10 Mar	Katshan Far- Cher-	Sera Freda Taur Saut	
kU?		2.1	.6													
A	· · ·			1		6		and the second	F	×.	4	188			1 4	-
Court of Contorner®	Esturne Labels	•		1000				Sum of Amount	Column Labels	• 124 A						
Row Labels	Apr	May	Jun	M	Are.	Sep		Row Labels	- Apr	May	Jun	24	Are	Sep		
Apr		•				100		Apr	600	0 17200		4500	2000	8000		
North State			1					North State		1900	31410	400	430	1400		
								h.			114.30	25500	440	2000		
Aut.								Aut					10250	1600		
544								540						6400		
Grand Tatal		4 1	5 5	5			£)	Grand Total	600	0 24609	73850	35300	17180	19600		
Month	Apr	May	Jun -	M	Ave	Sep	1	Sew Cultomer	400	7400	22450	25500	10750	6400		
Apr	200	75.0	0%	25%	25%	25%	1	OleCustomer		0 17200	51400	9800	6930	13200		
May		1009	100%	50%	50%	05	<u>.</u>	New Customer AOV	1,500	1,700	7,483	12,750	5,125	3,200		
Aut .			100%	11%	33%	3.3%		Old Customer AOV	1 T	5.733	25,700	3,267	1,733	4,400		
м				100%	50%	.0%	1									
Ast.		-			100%	100										
542		_	_			100%										
		-	-	-			÷ .									
New Customer		4	-	- 1		- 1	1									
OldCustamer		9	1		4	-	£									
N Nerw Cushomer	100	404	60%	404	114	1.00	£									
a concretional		741 604	6.04	604	0.7%	1.60%	1									

The cohort also shows the quality of the customer that is being acquired. The

month wise comparison of the repeat purchase and the AOV is one indicator. You can do many more additional calculation out of this excel sheet.

第二日 ゆう (で、)									Celuit But - Lost							1 10 -	
FLE HOME INSE	HT 1	FREELANDS	A (10	454,3,45	045	4 MOV	64 1	rew .									See a F
	13	St. Autor	1) for	-	9	1 5	a.	-8	8 10 24	Oview Sole by Sole	B	7 0	1				
Anna Pageltout, Page C Pantes Sacod	News	Course	214	denge .	200m	Wh D	on toi	New	Anange Frank	Anat Western Party	Sect.	A Mari					
Wolfsoyk Year.			(mare			Inte				-		(there					
A			£.		. 8		- 6			1				M.	14		
Court of CostomentD	Colum	ntabels +							Sum of Amount	Column Lab	bels 🕒						
Row Labels -	Apr		May	Jun	м	Ave	Sep		Row Labels	- Apr		May	Jun	м	Aug	Sep	
Apr		- 4	3			1	1	1	Apr		6000	17200		4500	2000	8000	
May			2	2	1	1 1			May			7400	51400	4500	450		
Jun				3	1	1 1	1	t -	Jun				22450	800	3800	3600	
24					3	1			put.					25500	680		
Ang						2	1	1	And						10250	1600	
Sep							2	1	Sep							6400	
Grand Total		4	5	5		6		£	Grand Total		6000	24600	73850	35300	17180	19600	
Month	Apr		May	Jun	M	Aug	Sep	1	New Customer		6000	7400	22450	25500	10250	6400	
Apr		300%	75%	0%	255	25%	255	1	OldCashomer		0	17200	51400	9800	6930	13200	
May			\$00%	100%	50%	50%	0%	1	New Customer AOV	17 I I I I I I I I I I I I I I I I I I I	1,500	3,700	7,483	12,750	5,125	3,200	
r hin				100%	331	33%	33%	1	Old Customer AOV		14	5,733	25,700	3,267	1,733	4,400	
M					100%	50%	0%	1	-								
Aug						100%	50%	1	No of Outomer six m	worth old		- 4					
Sep							100%	1	No of Transaction do	one .		10					
1 (C)									Amount of Ghransac	tion .		37700					
New Customer		- 4	. 2		1	2	1	7	Average Transaction	per customer		2.5	2				
OldCustomer		0	3	2		1	1	1	Average Transaction	Amount Jer cur	tomer	9425	1				
N New Customer	1	100%	40%	647%	40%	33%	40%					-					
% Old Customer		0%	60%	40%	60%	63%	60%										
· · Sheet2	See	a con	104a	Sheet	1.0.4	1 6	0			1.(4)							
tain .															D •		

Let us add one more column in the excel sheet – Days between purchase. How long customer takes to repeat the purchase. The repeat purchase days will vary across the category.

H HOME	** 1 NOR 140	LANOUT PORMANE	OA1a	even ven	Cellur	Clink-Dod	1					1 10 -	La P
Part Access Part Rep. Part	De la	Rates Distant	14 <u>33</u> 14 <u>55</u>	$\frac{\overline{T}}{T} \stackrel{\overline{V}_{1}}{\underset{\overline{V}_{1}}{\underset{\overline{V}_{2}}{\underset{2}}{\underset{\overline{V}_{2}}{\underset{\overline{V}_{2}}$	1	E B	1.2	Consider a	The Part Streement on Streem	*(i)	1	Balan Arnadiya Balanan	
Grite	ternar bete	Contectuals		Lucanes.		14		da Tenti		0.04		Anaripe's	
Transartia	Customert C	ate of Transaction	anning [Amount Dia fe	Durchard	alaine Det	Ininte	of Burchasel	to Days Between Pur	chase		-	
140600001	CT10001	15-Apr-15	3	1800	1	15-Apr-15	Acr	Ace		0			
TR000007	CT10001	21-May-15	6	4200	2	15-Apr-15	Apr	May	+IF0 3+1.0.INT/CB	(2)			
TR000056	CT10001	22-348-35	1	4500	3	15-Apr-15	Apr	but		-			
TRX0029	CT10001	8-Sep-15	2	8000	4	15-Apr-15	Apr	Sep					
TR000002	CT10002	17-Apr-15	1	1000	1	17-Apr-15	Apr	Apr					
18060022	CT10002	15-Aug-15	4	2000	2	17-Apr-15	Apr	Aug					
TR000003	CT10003	20 Apr-15	- 5	2000	3	20-Apr-15	Apr	Apr					
11060006	CT10003	4-May-15	- 5	3000	2	20-Apr-15	Apr	May					
TRX0009	CT10003	30-May-15	2	\$0000	3	20-Apr-15	Apr	May					
1000004	CT10004	25-Apr-15	- 4	1200	1	25-Apr-15	Apr	Apr					
140400005	CT10005	4-May-25	2	2400	3	4-May-15	May	May					
TR000012	CT10005	16-Nrs-15	4	2400	2	4 May-15	May	Buth					
TAX0019	CT10005	31-54-15	3	4500	3	4 May 15	May	Ref					
TRADOOR	CT10006	23-May-15	1	5000	1 2	23 May 15	May	May					
TRX00050	CT10006	10-Jun-13	7	49000	27	23-May-15	May	245					
TR060023	CT10006	19 Aug-15	1	450	3 7	23-May-15	May	Aug					
TRX00011	CT10007	15 Jun 15	3	2100	1	15-Jun-15	hut	Durk .					
18060015	CT10007	6-348-33	2	800	3	15-Jun 15	but	but					
TR000013	CT10008	20 Auto-15	20	20000	1	20-Jun 15	Purh	2 yrs					
TRX0025	CT10008	31-Aug-15	1	3800	2	20 Jun 15	Run	Aug					
18060014	CT10009	30 Jun 15	1	350	3	30-Jun 15	Buth	bun	1				

The average time taken for user to purchase second time is 42.6 days; third time is 50.75 days after second purchase and so on.

	inst Population Formula	rt Work	- Microsofterme	rt Excel Ven 1	-		Cutton I	durings.									
erfalse Name Adi orfheinig Na Carlans - S Austrass	And Facility of Compared System Faces	a Coup S a Coup to B Coup to Other			inat A	an Chang		2		E Constantion	North Constantion	Anna Anna		1000	22		102
43	+ - A Revubels																
A		c	0					- U.	1		- 6 V	M	N	0	. #	0	1
Revisite's -	Average of Days Between Purchase																
1									Te.	and table for	and the second	-		1			
1	50.75								-18	NOT ROOK TO	eesse:						
4	41								9	Parate Relits to a	all in caper		(D) *				
Grand Total	22.50000007								1	Contraction of the local division of the loc	A.40			-			
									18	Anave			- 13	-			
									-18	(In to Parch			- 8				
									19	Jarring Darm			- 11				
									- lî	Peterter			- 10				
									- 12	Citeys Betwe	es Parcha	-	*				
										og feld-beline	er in site bei	-					
										/ Report File		III GAMES	-	-			
														-			
									-13	A Rev Liberts	-	E 18645					
										a crystee		Array of	and an other	-			
0																	
										The loss	-		-				
A B B Port2	sheet3 / Ghert bits . 5)								12.	, pee blac	and a		-	_			

Now add standard deviation of the days between the purchase, there is huge variability in the first to second purchase, second to third purchase seems to be more stable.



If you add amount in the pivot it shoes the average purchase value of those who are purchasing first time, second time and so on.



The above information about gap between purchases can be used for sending communication to customer for the repeat purchase. If average time taken for repeat purchase for a category is 30 days then probably the communication can be sent at 28/29 days after the previous purchase. There are many other useful information can derived out of the above cohort information.

In the example I have used month as the time for the cohort, we can have different time period and time windows. For example we can create a cohort of customer who purchase before a festival offers and cohort of customer who purchase on the festive offer. We can have cohort of customer who comes through a traffic sources such as Facebook. We can have cohort of customer who buy a particular category in a month. We can have cohort of customer who
participate in a survey. Basically we can create as many cohorts as possible; it really depend on what kind of analysis is to be done and what is expected out of it.

If the volume of data is huge then you might find it difficult to work with excel then you can create same structure using database.

4.1.4 Customer Lifetime Value

The customer lite time value is total value a customer provides to the company over a period of time. There are several methods of calculating the customer life time value (we call it CLV in short). In very simple term is CLV is present value of the profit that customer have given to the company is the time period. Now you can assume appropriate time period or assume customer to be of going concern. Let say customer stay for N months, can is expected to give R₁, R₂,...R_n revenue and cost of acquisition and retention is Q₁, Q₂,...Q_n. Assuming the company has x% margin and weighted average cost of capital is r, we can calculate customer life time value as

$$\begin{aligned} & \text{CLV}=&(R_1 \ast x \% - Q_1)/(1 + r) + (R_2 \ast x \% - Q_2)/(1 + r)^2 + ... + (R_n \ast x \% - Q_n)/(1 + r)^n \\ & \text{Or CLV}=& P_1/(1 + r) + P_2/(1 + r)^2 + + P_n/(1 + r)^n \\ & \text{CLV}=&_{i=1}\sum^N P_i/(1 + r)^i \end{aligned}$$

Where P_i are the profits from the customer in period i.

Here I want to use cohort analysis to calculate the customer lifetime value and how to use cohort to predict the CLV of a customer. From the cohort analysis we can find the customer rate of return month after month and the average purchase amount per customer month after the month. Assuming the time period for calculating the CLV of the customer 1 years and if cohort data provides following information, then we can calculate CLV as average margin per customer for the month and the average repeat rate. CLV is coming as 4182 by margin without using discounting factor. Here we can do month wise discounting or more aggregated yearly discount using the company's cost of capital. Assuming company has 10% as annual cost of capital, the CLV can be calculated as 4182/(1+0.1) = 3801 at yearly basis.

Base month Coho	ort	

Subsequent	Average margin per	Average Repeat	Final
Month	customer	rate	Value
Month01	2,300	100.0%	2,300.00
Month02	4,000	15.0%	600.00
Month03	3,500	10.0%	350.00
Month04	3,400	8.0%	272.00
Month05	2,400	6.0%	144.00
Month06	2,000	6.0%	120.00
Month07	1,900	5.0%	95.00
Month08	2,100	5.0%	105.00
Month09	1,850	4.0%	74.00
Month10	1,500	3.0%	45.00
Month11	1,700	3.0%	51.00
Month12	1,300	2.0%	26.00
		Total	4,182.00

This is a generic CLV calculation for all the customers. However there will be many different types of the customer. For prediction of the CLV of the new customer it is better to classify the customer into the one the segment that company has already created and apply cohort data of that segment to provide better prediction. Also the cohort based CLV will help understand the increasing or decreasing loyalty of the customer over the period.

4.1.5 Customer Segmentation by Clustering

Segmentation is perhaps one of the most important term in the digital analytics because any marketing activities involved targeting and targeting cannot be precise without good segmentation. The segmentation is intuitive and very easy to understand. For example in below picture, we have three distinct type of shapes in the population; the job of segmentation is to find the way in which the population can be divided into three segment using its shape and color.



However in most of the situation we will have multiple dimension on which a population has to be segmented. In many of the situation with good knowledge of business requirement if-else type of condition can be used to filter out the segment using different dimensions. The complex segmentation need more sophisticated algorithm.

In the Google Analytics segmentation section we learned how to create segment using different values of the dimensions available in the Google Analytics. The segment thus created can be viewed as a separate entity and further analysis can be done on it. The segment is based on the known dimensions and data collection being done as per the possible segments. In case of the order data the segment cannot be created out of the box, the segment has to be created out of the data from different systems. Another problem with segment by filtering the dimension with known value is that there could be hidden segment which we do not visualize. In this section the readers will be taken through two of the popular clustering technique that is **hierarchical clustering** and **k-mean clustering**. The clustering is method of creating clusters from the data in such a way that the data points with similar characteristics are in one group or cluster. In hierarchical clustering the number of cluster is not fixed, instead the result is dendrogram. By looking at the dendrogram the clusters can be created. In **k-means** we fixed the number of the

clusters at the beginning and the system will create that many clusters using distance measures. The readers must be aware of the facts that the clustering is a trial and error method in some way; one can hardly get a great segmentation in the first attempt. Also one has to be aware of the level at which segment has to be created. The segment can be start from an individual data point to the entire population. From utility point of view in marketing the segment should be distinctive and significant enough to be target precisely with require communication.

Hierarchical Clustering

In this section we will work on R to generate the clustering output and try to derive clusters out of that output. Assume that we have below sample of customer purchase history. We have to create clusters out of the given data. Before starting the clustering process you can create scatter diagram for two dimensional data point to see if some data are creating a distinctive cloud in the space. However for multi-dimensional visual effect cannot be used directly, so we rely on the clustering algorithm.

THE HOME O	NGRE 1	HADE LIKE OVT	NUMMER	DATA REVEN	VEW	L.		-								Sepa F
	121	Rea De	innate (2 12 18		E	10 2	Lett.	Divisor Se	kty5dk	85	2				
Institute Projectionals, Proje	· Cuton	Victoria Via	Andrea A	are 1975 June	to New	here	funit	-	22 2000	rana branter	Sub h	Marrie				
Patrice Liph	of Xeek			Select	en Windo	* 44	Parist 1	-	10.444	region Paulos	* Wednes	*				
		1.000		1 1		-		- 7			-	N			1.1	
I Customerid Apr	e ind	one No I	Purchase A	ov											-	
2 CUST00001	35	10	4	2500												
CU5700002	23	5	1	2000												
+ CU5700003	56	45	2	1300												
1 CUST00004	34	12	50	2000												
a CUST00005	33	20	12	3400												
7 CUST00006	34	29	4	2300												
CUS700007	56	34	5	800												
+ CUST00008	25		1	3000												
 CUST00009 	45	45	20	250												
+ CUST00010	32	35	1	1250												
u cu5700011	50	60	12	1160											-	
US700012	40	65	13	2000												
w CUS200013	43	37	7	1350												
5 CU5700014	36	23	9	1400												
# CUST00015	33	20	13	1320												
r CUST00016	30	14	2	950												
= CUST00017	42	36	1	1200												
0 CU5700018	31	45	4	2300												
- CUST00019	32	28		1350												
CUST00020	29	20	5	3000												
22 CU5700021	27	26	6	2008												
s Cu5700022	33	25	4	2340												
· · · Dead	4 6									1.047						2114
and the second	and a	_	_		_			-	-					-	_	

Upload the data into R

3									RStudio
File	Edit Code 1	liew P	lots Session	n Build D	ebug To	ols Help			
0	A.00		de Go ta f			1			
-			1 Automatical Contraction		-				
Co	nsole -/ 🕫								6
> (L <- read. c	sv(fi	le = "c:/	RData/clu	stering	.csv", hea	ader = TRU	(B)	
> 0	L		1.000						
	CustomerId	Age In	ncome No_	Purchase	AOV				
1	CUST00001	35	10	4	2500				
2	CUST00002	23	5	1	2000				
3	CUST00003	56	45	2	1300				
4	CUST00004	34	12	10	1000				
5	CUST00005	33	20	12	1400				
6	CUST00006	34	19	4	2300				
7	CUST00007	56	34	5	800				
8	CUST00008	25	8	1	3000				
9	CUST00009	45	46	20	750				
10	CUST00010	37	35	1	1250				
11	CUST00011	50	60	12	1160				
12	CUST00012	40	65	13	2000				
13	CUST00013	43	37	7	1350				
14	CUST00014	36	23	9	1400				
15	CUST00015	33	20	13	1320				
16	CUST00016	30	14	2	950				
17	CUST00017	42	36	1	1200				
18	CUST00018	31	45	4	2300				
19	CUST00019	32	23	8	1350				
20	CUST00020	29	20	5	3000				
21	CUST00021	27	16	6	2000				
22	CUST00022	33	25	4	2340				
>									

Using Euclidian distance method of calculating the cluster member's nearness we created hierarchical cluster using wards method. In the right windows we can see the clustering output in Dendrogram. The number of levels of hierarchy is different for different members.



At the lowest each members are a segment or a cluster. However any segment which are not significant has no use in the marketing activities, we must segment them at some level. In this example I created four clusters out of the 22 members and bordered them with red color. Now {4, 16, 7, 9} is cluster 1, {5, 14, 15, 13, 19, 11, 17, 3, 10} is cluster 2, {8, 20} is cluster 3 and {12, 2, 21, 1, 22, 6, 18} is cluster 4. Similarly you can create a 5 cluster or 3 cluster out of this Dendrogram.

0	RStudio	
His Edit Code View Parts Session Buff Debug Tank Help		
💽 👷 🕫 🕼 🖓 🖓 (Alterheterie) 👘 🖓 Adder		S Popul Novel +
*) (maint's	- Indexed	-0
2 B Cleventer & Z + E 34a 39 3tors + 3	at 12 Streptions 4	11 Se+ (🚱
1 CL fead.csv(file-'CL/RDatk/Clustering.csv', header-TRUC) ,	Code Ducestart -	4
J Dst dist(CL, method = "euclidean")	Gitariure 18.5	
4 HST hclust(Dst, eethod = "ward.02")	groups int (1:22) 1 1 2 3 2 1 3 4 3 2	
6 groups - cutree(HST, k=4)		
7 rect.hclust(HGT, k=4, border="red")	Manay tals Parts Parkages may Verson	mO
81 (Related 6 Prove 1	and a series of the series of	S Pass 107
1 m h i A		
T AND AND AN A A AND A A	Cluster Dendrogram	
9 CUS100009 45 45 20 750 10 CUS100016 17 15 1 1250		
11 CUS700011 50 60 12 1160		
12 CUS700612 40 65 13 2000	1	
14 CUCTODD14 36 21 9 1400	8	
15 CUST00015 13 20 13 1320	81	
16 CUST00006 30 14 2 950		
18 CUUTODELS 11 45 4 2100	5 1	
19 CV5700629 32 23 8 1350	1 2	
20 CV5700829 29 20 5 3000	§ 1 1 1 1 1 1 1	
22 CV5700022 11 23 4 2140		
> DSt <- dist(CL, method + "ewclidean")	• ^ 6 6 6 6 6 6 6 6	
sarving methods, "methods,") this introduced by contribu-	NORDANNARPHODON	A
a critical second a contracte y therein		64 54 a.
> MST <- hclust(0st, method = "ward.02")		
+ plot(951)		
+ rect.hclust0457, k-4, border="red")	Det	
*	houst (*, "ward 02")	

K-Mean Clustering

K mean clustering is to create a k cluster from the n data point. Assuming there k dimension in the data set the k mean algorithm objective is to minimize the distance from the data point to the mean of the cluster.

Data Set = $x_1, x_2, x_3 .. x_n$

Cluster sets = $C_1, C_2, ..., C_k$ and Mean of each cluster as $\mu_1, \mu_2 ..., \mu_k$

minmize
$$\sum_{j=1}^{k} \sum_{i=1}^{n} (xij - \mu j)$$

It is an iterative process cluster mean are recalculate and data point are reallocated till minimum distance is achieved.

In the example we will use k-mean clustering algorithm to cluster the same set of the data point. In this example the number of cluster is chosen as 4 by default. There is various method of deciding the number of clusters which are more of heuristics or thumb of rule.

2	Uncicled1*	×		-0
-	5	G Sere	enfere Q Z . D	La 50 = 50.000 + ≥
1	CL c	read.csv(fil	e="C:/RData/Clustering.cs	w", header-TRUE)
2	CL			,
3	KM <	- data.frame	(CL[-1])	
- 4	826			
5	KC <	- kmeans(KM.	4. nstart = 25)	
6	KCSs	ize		
7	KCSC	enters		
8	1			
.8	1 (%	plevel) s		R Script a
Con	sole -/	0		
12	40	65	13 2000	
13	43	37	7 1350	
14	36	23	9 1400	
15	33	20	13 1320	
16	30	14	2 950	
17	42	36	1 1200	
18	31	45	4 2300	
19	32	23	8 1350	
20	29	20	\$ 3000	
21	27	16	6 2000	
22	33	25	4 2340	
> K	C <- X	means (KR. 4.	nstart = 25)	
	4 9 7	2		
	(Scent	ars		
1	An	e Income N	o Purchase AOV	
×K		0 26,50000	9,250000 875,000	
× K	1.2500	2 33, 22222	7.222222 1303.333	
1 4	1.2500			
14233	1.2500 0.2222 1.8571	4 26,42857	5,142857 2205,714	
1 4 2 4 3 3 4 2	1.2500 0.2222 1.8571 7.0000	4 26.42857	5.142857 2205.714 3.000000 3000.000	

The clusters has been 4 cluster created cluster $1 = \{3, 5, 10, 11, 13, 14, 15, 17, 19\}$, cluster $2 = \{8, 20\}$, cluster $3 = \{4, 7, 9, 16\}$ and Cluster $4 = \{1, 2, 6, 12, 18, 21, 22\}$. The clusters created are exactly similar to the one we achieved in the hierarchical method. This is because both method uses Euclidean distance measures, the same dataset was used and number of clusters are same.



The real life data will have millions of the rows of data; hence it requires iterative process to arrive at the final clusters. The objective of this section was to show how to run the algorithm and interpret it using very small data set. More than any other algorithm, the clustering requires good understanding of the data input.

4.1.6 Market Basket Analysis

The market basket is the basket of the customer in their purchase. The cart of the customer order contains all the information on the product purchase by each customer. Combining catalog information like brand and category in the cart provide data which can be used for segmentation of the customers.

Items are the objects that we are identifying associations between. For an online retailer, each item is a product in the site. For a publisher, each item might be an article, a blog post, a video etc. A group of items is an item set I= $\{i_1, i_2, ..., i_n\}$

Transactions are instances of groups of items bought together. For each transaction, then, we have an item set $T_n = \{i_i, i_j, \dots, i_k\}$

Rules are statements of the form $\{i_1, i_2, ...\} \Rightarrow \{i_k\}$

That is if you have the items in item set I_1 and I_2 on the left hand side of the rule, then it is likely that a visitor will be interested in the item on the right hand side (RHS i.e. $\{i_k\}$. The example could be $\{\text{flour, sugar}\} \Rightarrow \{\text{eggs}\}$. This means that people who bought flour and sugar is likely to buy eggs also. The output of a market basket analysis is generally a set of rules that we can then exploit to make business decisions. Before delving deep into the actual algorithm we have to understand the few terms commonly associated with Market Basket Analysis.

Support: The percentage of transactions that contain all of the items in an itemsets (e.g., pencil, paper and rubber). The higher the support the more frequently the item set occurs. Rules with a high support are preferred since they are likely to be applicable to a large number of future transactions.

Confidence: The probability that a transaction that contains the items on the left hand side of the rule (in our example, pencil and paper) also contains the item on the right hand side (a rubber). The higher the confidence, the greater the likelihood that the item on the right hand side will be purchased or, in other words, the greater the return rates you can expect for a given rule.

Lift: The probability of all of the items in a rule occurring together (otherwise known as the support) divided by the product of the probabilities of the items on the left and right hand side occurring as if there was no association between them. For example, if pencil, paper and rubber occurred together in 2.5% of all transactions, pencil and paper in 10% of transactions and rubber in 8% of transactions, then the lift would be: 0.025/(0.1*0.08) = 3.125. A lift of more than 1 suggests that the presence of pencil and paper increases the probability that a rubber will also occur in the transaction. Overall, lift summarizes the strength of association between the products on the left and right hand side of the rule; the larger the lift the greater the link between the two products.

Now that terms are clear to us let us work on one example in R to understand the process of generating the data and how to interpret the data. We will use below simple data set – ordered, products associated with order and Quantity Sold.



Upload the data in R

0			and a second	RStudio
ie E	de Code Ver	Plots 1	lession Build Debug Tools Help	
	· 88		#8 Gatellia function () Addea +	
0)0	atter a			
-	0.0	-	A 4 0	Case of the United States of the
123	MS <- rea MS	L.csv(f)	le = "C:/RData/MBA.csv", header-TRUE)	34 9 344 0 4
31	(tig Level))			tion of
	c. read.c	wffile	= "C:/RData/HEA.cov", header=TRUE)	
> 31				
(rderID Pro	duct Qui	intity	
1	00001	A	1	
2	00001	8	2	
3	00001	C.	1	
4	00001	0	2	
3	00002		1	
<u>e</u>	00002		1	
<u> </u>	00002	0	2	
8	00003	0	4	
2	00004	0	1	
-	C C C C C C C C C C C C C C C C C C C	e.		
10				
11	00004	D	4	
11 12	00004	A		
11 12 13	00004 00005 00005	DAS	5	
11 12 13 14	00004 00005 00005 00005	DASC	5	
11 12 13 14 15	00004 00005 00005 00005 00005	04800	692	

The input data is in the linear format. This has to be converted to comma separated format for each item in the orders.

0				Kšudio
Fig 8	ldit Code Vie	ew Phyla S	ession Build Debug Tools Help	
• •	· 8	8 81	# Genetic function	
5	Dettern a			-0
	40 8	Severe	nine Q.Z. E	Setter Se Defense + 2
1	NS <- rea	ad.csv(fi	<pre>le = "C:/RData/WEA.csv", header=TRUE)</pre>	
2	MB			
45	981 c r	ead.trans	actions(file ="C:/RData/Temp.csv", format="single	", sep-",", cols-c("OrderIO","Product"))
51	(by Leve)	1		A Scopt 1
10	00000	0		
11	00009	6	4	
12	00009	ě.	6	
5	00010		i i	
34	00010	6	2	
35	00010	E.	1	
36	00011		1	
17	00011	c	ì	
38	00011	D	5	
39	00011	6	4	
40	00011	6	4	
41	00012	C	5	
42	00012	0	1	
43	00012		1	
44	00013	0	2	
45	00013			
2.00	rite.csv(M	I. file-	C:/KData/Temp.csv", row.mames = FALSE)	
SN	11 c- read	transact	tions(file ="C:/RData/Temp.csv", format="simple",	sep=",", cols=c("Order10","Product"))
1				
1		_		×

You can the output for each order. For example Order OD001 has 4 items A, B, C, D.

	and the second	RStudio				
lie Edit Code View P	tura Session Build Debug Tools Help					
	Addre • Addre •					
() United at a		-0				
0 0 0	Severen See 9 Z + D	The best states and the second at the				
1 MB <- read.c	sv(file = "C:/RData/MEA.csv", header=TRUE)					
4 MB1 <- read. 5 inspect(MB1) 6	transactions(file -"C:/#Data/Temp.csv", format-"single"	", sep-",", cols-c("OrderIO","Product"))				
61 (liplest) s		- Contra				
45 00013 > write.csv(MB, f > MB1 <- read.tra > inspect(MB1) items [1] (A,B,C,D) [2] (A,B,D)	<pre>Fle="C:/RData/Temp.csv", row.names = FALSE) nsactions(file ="C:/RData/Temp.csv", format="single", s transactionID 00001 00002 00002</pre>	sep=",", cols=c("OrderID","Product"))				

Run the Apriori algorithm in R. Add minimum support and confidence level. I have added very low number because the data set is very small but for real data set the support and confidence has to be higher for any good decision.

0	RStudio
For Edit Code View Plots Session Build Debug Tools Help	
🐑 + 🔐 + 🔒 🕼 (Alteriation technologies)) 🔛 + Addee +	
P Desteril x	-0
2 B Disence and a Z . E	Gefan 39 (Pfever + ≥
1 NB <- read.csv(file = "C:/RData/MBA.csv", header=TRUE)	
<pre>write.csv0%6, file="C:/RData/Temp.csv", row.names = FALS() % M81 <- read.transactions(file ="C:/RData/Temp.csv", format="sing) inspect(M81) % MOut <- apriori(M81, parameter=list(support=0.001, confidence=0.2 7 </pre>	le", sep=",", cols=c("Order10","Product")) [5, minlen=1))
21 (bplant) s	t Soyt (
Consule -/ +D	-0
<pre>> MOut <- aprioriOMB1, parameter=list(support=0.001, confidence=0.25, Apriori</pre>	minlen-1))
Parameter specification: confidence minval smax arem aval originalSupport maxtime support min 0.25 0.1 1 none FALSE TRUE 5 0.001	nlen maxlen target ext 1 10 rules FALSE
Algorithmic control: filter tree heap memopt load sort verbose 0.1 TRUE TRUE FALSE TRUE 2 TRUE	
Absolute minimum support count: 0	
<pre>set item appearances[0 item(s)] done [0.005]. set transactions[6 item(s), 11 transaction(s)] done [0.005]. sorting and recoding items [6 item(s)] done [0.005]. creating transaction tree done [0.005]. checking subsets of size 1 2 3 4 5 6 done [0.005]. writing [192 rule(s)] done [0.005]. creating 54 object done [0.005].</pre>	

System has generated 192 rules.



You can view the rules in the R screen or export it to Excel.

Ū	R	Shudio
Fie Edit Code V	ex Pots Session Build Debug Tools Help	
Q · 🛫 · 🔒	🚱 🔮 🖓 Gete Networken 👘 🔛 • 🛛 Addes •	
O Destative .		- 7
E		
40 6	Severensee 9 Z . L	29 Fun / 39 (19 Source + 1 2)
1 MB <- re	ad.csv(file = 'C:/RData/MGA.csv', header=IRUE)	
1 write.cs	with file-"C:/RData/Teno.csv", fow manes - FALSE)	
4 M81 <- r	ead.transactions(file ="C:/RData/Temp.csv", format="single",	sep=",", cols=c("Order10","Product"))
5 inspect(M81)	
6 MOUT <-	<pre>apriori(M01, parameter=list(support=0.001, confidence=0.25,</pre>	minlen=1))
/ sout	Mine)	
e inspecti	AGR()	
91 (lipland	P.8.	KSoyt :
Annah 1.10		
Console -/		
> NOUE	1	
set of 192 ru		
lhs	the support confidence lift	
[1] ()	=> (C) 0, 18461518 0, 1846154 1,0000000	
(2) ()	«> (A) 0.46153846 0.4615385 1.0000000	
[3] ()	*> (£) 0.53846154 0.5384615 1.0000000	
[4] ()	(8) 0.53846154 0.5384615 1.0000000	
61 0	*> (F) 0.69230769 0.6923077 1.0000000	
161 ()	(D) 0.84613385 0.8461338 1.0000000 (A) 0.35284435 0.4000000 0.8444442	
	(A) 0.15384615 0.4000000 0.80000007	
191 101	(c) 0.15184615 0.4000000 0.7428571	
(10) (6)	=> (C) 0.15184615 0.2857143 0.7428571	
[11] (c)	=> (6) 0.30769231 0.8000000 1.4857143	
(12) (6)	=> (c) 0.30769231 0.5714286 1.4857143	
(1)) (c)	-> (F) 0.23076923 0.6000000 0.8666667	
(14) (6)	«> (C) 0.23076923 0.3333333 0.86666667	
[15] (C)	«> (0) 0.38461538 1.0000000 1.1818182	
[16] (0)	=> (C) 0.38461538 0.4545455 1.1818182	
(1/) (A)	(6) 0.23076923 0.5000000 0.9285714	
100 /0		

0 85	do
lie Edit Code View Pluts Session Build Debug Tools Help	
- C + B B A CA Contratuctor B+ Addes +	
6) DetHell, x	-0
2 B Dievernier Q Z . E	- fun Se Disease + 2
1 MB <- read.csv(file = "C:/RData/MBA.csv", header=TRAC)	
2 MB	
J write.csv(MB, file="C1/RData/Temp.csv", row.names = FALSE)	and T. P. and a Conductor Streed at The
<pre>* Mo1 < read.transactions(tile = C:/Nuita/Temp.csv , Format= single ,</pre>	sep coisie(prograp . Product))
6 MOUT <- apriori(M81, parameter=list(support=0.001, confidence=0.25,	minlen:1))
7 Mout	
<pre>8 inspect(Mout)</pre>	
9 opt = inspect(Mout)	
10 write.csv(opt, file) c:/#Data/#DAOULLCSV)	12
N1 (helest) (Elevel 1
Conside -(/)	-0
[177] (B,C,E,F) => (D) 0.15384615 1.0000000 1.1818182	
[178] (B,C,D,E) ⇒ (F) 0.15384615 1.0000000 1.4444444	
[179] (C.O.E.F) => (6) 0.15384615 1.0000000 1.8571429	
[180] (B,C,D,F) ⇒ (E) 0.15384615 1.0000000 1.8571429	
[182] (A, B, C, C) => (0) 0.07692108 1.0000000 1.1818182	
[183] (A.B.D.E) => (E) 0.07692108 1.0000000 1.4444444	
[184] (A,D,E,F) => (8) 0.07692308 0.5000000 0.9285714	
[185] (A,B,D,F) => (E) 0.07692308 0.5000000 0.9285714	
[186] (8,0,E,F) -> (A) 0.07692308 0.3333333 0.7222222	
[187] [A.S.C.E.F] => [D] 0.07692308 1.0000000 1.1818182	
[155] (A,6,C,0,E) => (F) 0.0/692305 1.0000000 1.4444444	
[190] [A B C D E] -> [E] 0.07692108 1.0000000 1.8571429	
[191] (B.C.D.E.F) => (A) 0.07692308 0.5000000 1.0831333	
[192] (A,B,D,E,F) => (C) 0.07692308 1.0000000 2.6000000	
> write.csv(0pt, file="C:/RData/MEAOut1.csv")	
Error in is.data.frame(x) : object 'Opt' not found	
> write.csv(opt, file="C:/RData/WEAOutl.csv")	
s opt = inspect/pour/	

In excel output we can filter out on any value of support and confidence for arriving at the decision. The system generated rules for single items also but we might be interested in only combination of 3 to 4 items.

C H	HOME	NORT	FROM LANK	NT POMAA	s DATA ADV	ter vew	MER	Datt - Beat									1 10	- o x
D Bo	a ay i	-		1 X X - H		Dimentaria	-			E	2.5	2	Sector For	2 2	lukylum M+	-	A	
	and fare						123		10	mattings 1	Sear Style	6× ×	1.4		Chur -	1.841 - 5	aller B. I	
-		2.1	1.0		-						-		- 14		- 1	-		
200	Pri.		ebs. e	transfer in	confidence -	501				-			-			-		- 1
	0	- 1	073	0.18	0.18	1.00												
1 (21	0	1.0	(4)	0.44	0.46	1.00												
1 198	0	-	111	054	0.54	100												
141	0	-	185	0.54	0.54	1.00												
4 193	0	-	143	0.69	0.69	1.00												
+ 163	0	-	100	0.85	0.85	1.00												
. (2)	603	-	(4)	0.15	0.40	0.82												
141	143		10	0.15	0.11	0.87												
10.040	10	1.0	013	015	0.40	0.74												
IN CAME	(11)		3/3	015	0.29	0.74												
11111	163	-	195	0.11	0.80	1.49												
10.0230	(83		473	0.11	0.57	1.49												
14 12 21	1/2	-	111	0.23	0.60	0.82												
- (54)	043	-	403	0.21	0.11	0.87												
11111	103	-	101	0.38	1.00	1.18												
10 1567	000	-	001	0.18	0.45	1.18												
10 (17)	GAJ	-	013	0.23	0.50	0.93										1		
10 (28)	013	-	(A)	0.23	0.43	0.93										-	-	
10 1251	(A)	-0	080	0.31	0.67	1.24												
11 1200	082	-0	GAD	0.31	0.57	1.24												
22 (221)	GAD	-0	043	0.31	0.67	0.96												
	-	10.00	-							1.65								-
-	-	and the second	0		_		-	-	_	1.01					-	-	-	-

The market basket analysis output can be used in various business decisions such as what are the items that customer frequently bought together. This can be used for making combo. The website recommendation can be based on the product being added in the cart. For example $\{A, B\} \rightarrow H$ then if a customer add A and B then user can be shown H.

4.1.7 Logistics Regression

In many situations the decision are made based on binary data TRUE or FALSE, yes or no. A bank will decide whether to extend load to a customer in Yes/No form. There is nothing in between or number involved in decision. In such situation logistics regression is ideal for making prediction and help

decision making. For instance bank decision to extend loan or not is based on customer's likelihood of default or not. Using the customers information about income, location, education, age, property etc. the bank can predict whether a person is likely to default or not and then extend loan to those only who are not likely to default.

The logistics can take more than binary outcome as dependent variable but in this book we will consider cases which are binary in nature. The output of logistics regression is a probability of an event which is either success or failure, Yes/No. The logistics can take non numeric independent variables like male/female, national/foreigners and so on. This makes it more generic and help predicting the probability using both numeric and non-numeric data. These characteristics make logistics regression ideal solution for predicting the outcome of the customer behavior. If we know the behavior of customer leading the come events in the fast, we can use it to predict the outcome using logistics regression. One of the simplest examples is the probability of purchase based on the customer's information and site behavior.

Before started using the logistics regression lets us understand the basic mathematical background of the logistics regression. We will focus more on the application side of the logistics regression rather than the mathematics side. However we need to understand the process of the arriving at the logistics regression result and be able to interpret the result from the R.

Assuming you are betting on the outcome of a football match between two clubs, one of the clubs is your favorite and you mostly bet for that club only. You will bet on the basis of the odd or the likelihood of winning team. Let's assume in previous 10 matches against same club your favorite club won 7 times and lost 3 times.

So odd of the winning is 7/3 which is 2.667.

In statistics the probability of winning is 7/(3+10)=0.7.

Denote probability of p then odd of the winning can be represented as p/(1-p).

Hence odd of an event= p/(1-p). (Probability of Success / Probability of Failure)

If we take the natural log on both side then $L_{a,a} = L_{a,a} \left(\frac{1}{2} \right)$

Log odd = loge(p/(1-p))

Assuming that log odd can be represented in linear format,

 $Log(p/(1-p)) = \sum b_j x_j$ where j=0 to k, b is coefficient and x is independent variables

Taking exponential in both sides we get

 $p/(1-p)=\pi exp(b_j x_j)$

The expression exp b_j tells us how the odd of the response being true increase (or decrease) as x_j increase by one unit, all other thing being equal.

```
Assuming j=1
```

```
p/(1-p)=exp(b_0+b_1x)
```

 $p = exp(b_0 + b_1 x)/(1 + exp(b_0 + b_1 x))$

The output of the logistics regression is in s-shape graphs called sigmoid function.



In this section we will show how to run logistics regression in R as an example. Below data point will be used in the example. The purchase column is 0/1, 0 is for not purchase and 1 is for purchase. Another column Income is the independent variable in the logistics regression. For the sake of simplicity we will keep one independent variable only.



Add data into the R. Since data set is small I have just input the data using data frame.



Run the logistics regression with income as independent variable and purchase as dependent variable. Use binomial family as we want to have two types of output.

Ū		RStudio
Fie Ede	Code View Plots Session Build Debug Tools Help	
2.00	• 🔒 🚱 🚔 (a) Carlo The Turchan	
6) freese	67 x	
1 10 2 1 10	2 Sourcester Q. Z. L. p-data.frame(Income-c(10, 5, 4, 12, 20, 19, 12, 8, 46, Purchase-c(1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, Purchase-c(1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,	→ fan 30 → forst · 2 35, 60, 65, 17, 2), 20, 70, 16, 45, 2), 20, 16, 25), 1, 1, 0, 1, 1, 1, 1, 1, 0, 1))
4 10	<pre>c-glm(Purchase ~ Income, data-lg, family = "binomial")</pre>	
1.1		
	Autor A.	16-11
	(4) (44) (
Console -	1.0	-0
s	20 1	
6	19 1	
7	12 0	
8	8 0	
9.	46 1	
10	35 1	
11	60 1	
12	65 1	
13	17 1	
14	23 0	
15	20 1	
16	70 1	
17	36 1	
18	45 1	
19	23 1	
20	20 1	
21	16 0	
22	25 1	
> 11<-0	plm(Purchase ~ Income, data=lg, family = "binomial")	
> opt =	(inspect(Mout)	
		•

The intercept is -3.9414 and coefficient is 0.2491.



Using the intercept and the coefficient let us calculate the final probability for different income level.

085.4.1	-			-	2 12	- 13	Lagistics Regression - Be	*							7.1
A Col Carro	H +[0 1 y + ⊡ + Put	· A .		8-1 12-12 10-12	E Nag	n Tant A Cantar	(ana) - 5 - % + %	a Constant	Tata -	28 S		10	∑ Adula R fd+ € Cha+	Sata Sata	A
. A		¢	D ·	1		4				. 6	4	8.0	N	0	
p = exp(a + bx)/(1+	exp(a + bx3)														
2							Intercent (a)	-3.9414							
Income Deckshills	of Descharts						Income (b)	0.3404							
a lincome Processing	of Perchase			-			surficiente (10)	0.2491							
1															
1															
0															
1															
0															
13															
4															
5															
6															
7															
8															
9															
ф															
19															
12															
(3															
24															
13															
24															
11															
· · Sheett 0	dia Model	Sheet2							*						

Use $p=\exp(b_0+b_1x)/(1+\exp(b_0+b_1x))$ to calculate the probabilities.

	4 NT	61 -	1 m - 1	0- D-	and the	Presentation	- IK	0.02	5013	5+ 3×	701	Saddon	1.64	A
- SIL CARY -							A Conditions	a farmed as	21	an prov	101	Sec.	Sec. 1	End B
(* Found Farmer					organic carrie		** famating	- Tapis - 1	9,00			2 Our1	Ete .	Select
A		e 1	D	i i	- 4			- A	*	1		N	0	
p - expla + bxl/(1+	exp(a + bx))													
						Intercept (a)	-3.9414							
Income Probability	of Purchase					Income (b)	0.2491							
1-(10) +/	SPACE REPORT	opined.												
-														
3														
90														
11														
13														
14														
13														
12														
14														
13														
21														
22														
29														
24	and the second		_											
Deep 1.0	an mean p	ewc 1 - 6	9								_	_	1.1	
H St Ct - 1 HOME MORT	PAGELAHDUT	FORMUS F.C. C.	0414 	10100 0- 5	VEW Road Test	Legito Represent				Q= 3	× N	244	dun + g	-
H HONE MORT	PAGELAHOUT Bas - [11 7 9 + = +	- 10640.45 ⊡ X X = 1 9 - 1 - 1	0414 *	=====================================	vew Rog Test Weye & Cer	General General	teat				× ()	Σ.v.s	dum • g	T
HONE NORT	PAGELAHOUT BAI - [11 F W - II - 1 Anni	- 1000000 - 2 2 3 - 2 - 2	0414 	Rves ♦- 5 •: •: 5: •: •: 5:	vew Roge Test Merge & Cer	General General Services	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	and family	n Cel Syles	-1 and -	× ()	E AAA	dum • 4 • 5	T ANTS
A Part Cart	PAGE LAHOUT des +[11 2' W + = + Partit 8	товили Э.К.К. 1 Э <u>А</u> 1 С	0414 * = 11 # 12 12	Rves €- 3- +: -: 3: separat	vew Roge Test Marga & Car G	Expetics Represent General Ner + \$ + % + % Bustler H		ing take	- Ca - Syles - 4	-1 B	×	X AM	dum - 40 - 50 - 50 - 50 - 50 - 50 - 50 - 50 - 5	
A Col A Col B Coly Col Col Col Col Col Col Col Col	PAGE LAHOUT des +[11 2 ⊈ + ⊟ + 1 Parti 8 exp(a + bx())	тониция - Д - Д - 1 	0414 * = =	सरस्थ 6- क्र बद्द ब्द क्ष ब्राह्मवास्य 1	VEN Roge Test Merge & Cer G	Separtics Represent The Second Secon	teri N 2 Conto Israel	ng tang Ngan Jan Jan	- Cel - Syles - 4	- 2 0 C		E AM		
A star and a star a sta	*************************************	тонацая - А. А 1 <u>А</u> 1 	0414 ** ==]]] 0	Roca ⊕- ⋽ •< < ⊡ • •	ven Ruge Taal Marge & Car	Separation Regression ner - S + Si + S	teat 14 25 Consta 14 25 Consta 1 -3.941	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	- Sylar - Sylar - K	1 1 ·	× ()	E AAA E Court	dum + 4 - 5 tating 0	T
Par C* - a A	PAGE LANDUT IN [11 2 ¥ + = 2 Host 2 Host 2 Host 2 Host 3 Host 2 Host 3 Host 2 Host 3 Host 2 Host 3 Host 3 Host 3 Host 4 Hos	ποθαλλά Ξ.Κ.Κ. Π Ξ.Υ.Δ. Π Κ.	0414 	Roca ⊕- ⋽ •< < ⊡ • •	VEN Roge Test Merge & Cer	Experies Represent	1 2 Condition	nna fanna ng- Take Nym I 4	2 Cd 5 Spin		× ()	Σ Auto Transformed and the second s	fam + de station 0	T A F
A Con Control of Contr	PAGELAHOUT des - [11 2 ¥ - □ -] 2 ves 4 ves	тонала - Д. Д 1 	0 0	ночен Ф- Б- 42 42 (2) алунын 1 1	VEH Roge Test Marga & Car G	(aptic) feynmen (denuel ne \$ 5 % * 5 % masses ni Intercept (a) Income (b)	5 cm	nga fanad nga fada Ngan J 4	2 Cel . 575. 4	-1 10		Σ.A.A Tata 2 Con	fam + 4 5 tating 0	7
XCu none XCu none Tomar hume Cathere p = exp(a + ba)/(1+ tecome Probabilit 3	PAGE LANDUT der • [11 7 9 • = • 1 Freis 8 exp(a + bx()) y of Purchase 2.41% 1.51% 1.51%	тонца - (, к , к) - (, к	0 0 0	Rice ⊕- 5 << 2 square (1	VEW Roge Test Margar & Car G	(applic) Represent General Intercept (a) Income (b)	5.04 1.1 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2	nga fanad nga falis Ngan J 4	- Se	-1 15 a		Σ AAA Tata- e Cou	fam + 4 s tan 0	T
ki \$2+ C++ = sOM - sOM - SCM	PAGELANDUT des - [11 2 ¥ +]] + see exp(a + bx)) y of Purchase 2,24% 3,5% 3,5% 3,5%	10940.45 ⊡ K K 1 2- ▲ - 1 	2414 * = = = 0	Rota ⊕- ⋽ +< < ⊡ agenes 1	VEW Roge Test G	(aquito fagement (deneal 5 % % *) 5 % monte 8 Intercept (a) Income (b)	100 13 Contro Summer -3.941 0.249	ing faile Nym 1	Carlos v	- 1 10 1 ·		ZAA ZAA ZOU	ture + 4 5 tama 0	T Mar S Nort S
Kon noor Kon noor Kon noor Kon noor Kon noor Kon noor Prove(a+ba)(11 noor Probablis 2 3 4 5	PAGE LANOUT der • [1] 2 y · [] ·] 3 fund 1 fund 1 fund 2 Although 2 Al	10940.45 ⊡A:A: 01-▲- 0	2414 * = = = 0	Rees ⊕- 5 +c ∈ 2 square 1	VEW Prop Test Margar & Car G	Impires Represent General Intercept (a) Income (b)	-3.941 0.249	ing faile Nym 1	2 Ca 22	491	× (8)	ZAA ZAA ZOU	ture • • •	T Internet
Scat 2007	PAGE LANOUT dos - [1] f V - [] + [form form	тонала -] К К - 	0	River P · S ec ec S separate ()	VEW Prop Test Margar & Car G	Impiles Represent	Seat Seat Seat -3.941 0.249	nut fand nya fand Nya j	- 34- ×	49 1 ·		∑ Auto Trans d Con N	fam + 4 3 tamp	ter b l
K	PAGE LANDUT ass + [1] 2 ¥ 4 [] +	тоницая - (<i>К</i> - К - 1) - (-) -	0 0	River P - D +C +C D Appende (VEW Roge Test Mege & Ce G	(aquito Represent (dense) (a S + N) + 1 N Boston (a) Intercept (a) Income (b)	500 1 3 6000 1 3,941 0,249	and fand right fand light 1 4	- 53 · ·	100 °.		Σ Auto Taria Con N	fam · da saturno o	TT I
Kon noor Kon noor Kon noor Kon noor Kon noor Kon noor Provide ten S	PAGE LANOUT das - [11 f V - [11 for V - [11] f V - [11] for V - [11] for V - [11] for V - [11] j V - [11]	10000.00 	0 0	River e- S e< C square t	VEW Prog. Text Unrope & Car G	Inputs Toymout General Toy 5 + 56 + 1 S Summer N Intercept (a) Income (b)	5 (k) 1	ng fand ng faile Nan J	1 2 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4	100 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	× (8)	E AM	dan · · · · · · · · · · · · · · · · · · ·	TT III
ki 9- (* -) scott scott SCar car Stand Partie Stand Partie Proposition Probabilit 2 3 4 5 5 5 5 5 5 5 5 5 5 5 5 5	MAR LANOUT as '[1] 1 y + [] + 1 extp(++bx)) y of Purchase 2.41% 3.50% 5.50%	10000.00 	0414 1	Ref# ⊕- 5 << 2 square 1	VEW Prog Test Margar & Car G	Impiles Represent	500 91 2 500 1 -3.941 0.249	ng Table Ngan J	10 C S	10 2 °	× () 	E AM		TT III
ki *)- (* sour sour Kon sour Ricey - Providence A p-exp(a+ba)(1- ncome Probabilit 1 2 3 4 5 5 5 5 5 5 6 1 1 1 1 1 1 1 1 1 1 1 1 1	PAGL LANOUT das (11) f y + (-1) f y + (-1) f exep(a + bx)) y of Purchase 5,200	Конца - Д. К. С. Г. 5 - Д 1 - С. Г.	0 0	R×68 ♦- 3 << 2 4 1	VEW Broge Tast Ukegar Si Car G	(apito figment (apito figment apito figment apit	500 3 2 5 500 5 500 5 500 500	A 2	100 N		No a fare fare fare fare fare fare fare fa	E ALA Tata Con	fam • 4 10 10 0	
Kon none Propasse S	PAGE LAPOUT das • [11 1 y + [1] + 1 y + 1 y + [1] + 1 y +	- 0000000 ⊡ K X 1 00-▲- 1 0	0 0	River P · S +c +c (2) toposet t	VEW Bray Tad Ukrye & Cer	Inputs Represent	500 31 23 famili -3.941 0.249	A 1	× 60	State of the second sec	No a face i	E Con	fan - A	
K ← 1 K ← 1 K	MAR LANOUT as ' 1 1 ¥ + ::	- 1094444 ⊡ & X = 1 0 - ▲ - 1 2	0444 H = H U U U 0	€005 Ф: 5 42 42 (2) барная 1	VDH Roge Sch G	Impriss Represent	teat 3 8 3 6000 1 3 6000 1 4 -3.941 0.249		* Cd • Syler	1997 - "		E Ann	fan - 4 5 5 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
ki *)- C+ score score R(core - core * floored score * floored scored sc	PAGLUNOUT das • [1] f ¥ + [-1] fuel texp(a+ba)) y of Purchase 1 2,405 2	1094444 → A A 1 → A + 1 <	2474 9 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	River D- S- ec e S square 1	VEW Prog. Text Unegot & Cor G	Inputs Represent General Inter-S-N-N- S- Intercept (a) Income (b)		A A	. Ca . 93%	10 2 °	× (E	Zan Cou	dan + dan + dan + dan	
Kon Probabilit Provence Probabilit S	PAGE LAPOUT data [1] f y + [1]	товица -] К К - 	9414 9 9 9 9 0	Rvt# 0- 3- << 2 1 1 1 1 1 1 1 1 1 1 1 1 1	VEW Hop Tet G	Inputs Represent	1001 12 2 Godde 1 -3.941 0.249	and frame name y take y 4	« Cd • 3940.*	da 3. "	and from the second sec	2 AM	dan + dan + dan series se	1 1 1 1 1 1
K ← 1 K ← 1	MARE LANGUT 46 • [1] 1 1 <	10840.45	9444 9 10 10 10 9 10 10	River 0- 5 < < 2 1	VDW Bray Tad Harpe & Co	Inputs Represent	tead 3 8 3 3 6 1 -3.941 0.249	and found yes yes 4 2	- Ca - Syler - K	Ellin G Hust Or 1		∑an gas d d Cu	fan - 4 - 5 - 5 - 5 - 5 - 5 - 5 - 5 - 5 - 5 - 5	1 1 1 1 1 1
kil 9-0-1 scoat scoat Ricery 6 Without how 6 Optimum 6 S 6 S 6 S 6 S 6 S 6 S 6 S 6 S 6 S 7 S 6 S 7 S 6 S 7 S 7 S 7 S 7 S 7 S 7 S 7 S 7 S 7 S 7 S 7 S 7 S 7 S 7 S 7 S 7 S 7 IS 8 S <td>PAGE LANGUT das • [11] f y + [1] + [11] f y + [1] + [11] f y + [11] + [11] f y + [11] f y + [11] + [11</td> <td>10880.85 ⊇ ∧ ∧ 1 > ▲ 1 <</td> <td>0</td> <td>Rives 0 - 5 - 5 - 5 - 5 - 5 - 5 - 5 - 5</td> <td>VEW Bray Sul G</td> <td>Inputs Topmont General Top 5 + 55 + 1 5 - Sector Altercept (a) Income (b)</td> <td>3.941 0.249</td> <td>4 2</td> <td>* Sylar</td> <td>Can the Cr</td> <td>An an an</td> <td>E Entropy State</td> <td>dan - Article -</td> <td>tr i</td>	PAGE LANGUT das • [11] f y + [1] + [11] f y + [1] + [11] f y + [11] + [11] f y + [11] f y + [11] + [11	10880.85 ⊇ ∧ ∧ 1 > ▲ 1 <	0	Rives 0 - 5 - 5 - 5 - 5 - 5 - 5 - 5 - 5	VEW Bray Sul G	Inputs Topmont General Top 5 + 55 + 1 5 - Sector Altercept (a) Income (b)	3.941 0.249	4 2	* Sylar	Can the Cr	An	E Entropy State	dan - Article -	tr i
K Ga → C4 → A0007 Force A none K Ga → C4 → A0007 K Ga →	Add Labout An An An An	1000.04 3.4 ≤ 1 2.4 ≤ 1 4.5 4.5	0.114	Rever ⊕: 5 sec = 2 sec = 1 s	VOX Bray Ted G	Inputs Toymon General Toy - 5 + 5 + 1 - Second S	1000 11 23 Conten 1 -3.941 0.249	nge hale Store 2	- Ce - Syler	Sa tak	×	E San Contraction	fan - 4 - 5 - 5 - 5 - 5 - 5 - 5 - 5 - 5 - 5 - 5	1
Content and a content and	Indel.Lanout as • [1] I I inext Inext i • [1] y of Parchase 2.41% 2.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.52% 3.54% 3.54% 3.54% 3.54% 3.54% 3.54% 3.54% 3.54% 3.54% 3.54% 3.54% 3.54%	1000004 	0414 H = 10 H =	Recti 한 : 월 : 로 42 월 : 1	V(sr Brup Tad Harpe & Cr G	Inputs Represent	teat 3 8 4 3 6 6 1 -3.941 0.249	ng tang ng tang ng tak	- Span- - Span- - C	da 1.	in and a second	E San Tes Cost	ten - erste stanyo	T
kil *C+ *C *CM PODE *C *SCM *S *C *SCM *S *C *SCM *S *S	PAGE LANOUT das • [11] 1 Y + [14] Foot 1 Percel 1 P	1000004 10 / X / 1 2 - A - 1 2 - A - 1	0414	Rece 0- B coport	VEW Bray Ted G	Inputs Toymout General Toy 5 + 55 + 1 5 - Sector Hotercept (a) Income (b)	3.941 0.249	A 2	• 994	da J.) Σια « ετα ν	fan - 4 - 5 - 5 - 5 - 5 - 5 - 5 - 5 - 5 - 5 - 5	T

The probability till 60 lakhs income has been calculated using the output from the input data.

01	9.0.8					-	1	ogities Region	sion - Mic	nosoft D	iori.						- 0	
			Si nun Si tunin	2 47 2 47 2 47	nate stre	9 G 2000 50 20		3	100	-	All free 3 all free 3 All free 4 Minutes	en to Sele arteur Serek		1 8				
-	193		1				-								0			-
80	p + extia + bel	College and the local division of the			-	-		-	-	-		-				 	-	-
								Intercept (a)	-3.9624									
1	come Probe	hity of hardware						income (h)	6.34%									
5	42	95.85%																
6	- 60	99,89%																
6	44	99,91%																
	-6	95.50%																
	46	95,95%																
ē.,	67	99.96%																
ι.	48	95.57%																
1	40	95,37%				_												
	50	95.95																
٢.	. 10	95.56%																
	10	95,90%																
ι.	58	79.70%																
	54	99,99%																
	50	35,99%																
		100.00%																
÷	10	100.00%																
5.	10	100.00%																
	-	10.00																
1.	-	16.6.4																
											-							
2	Peril.	DOL Hold / D									-11 C.		_	_		 -	-	

On plotting the probability on the income we get an s-shape curve which is nothing but a sigmoid function.

10 H	HOME NERT	FIGURED FORM	U.46 0	the ADA		tw.	Lagisha	Asymmet (and the							1 10	- di
01.	Cat Cat Cater - Format Factor -	en ila ila Z ¥ - ⊡ - <u>A</u> - <u>A</u> nei	2 = = - = =		5 m	yr fant yr A Card	- 5	- 5 - 5			· · · · ·	the last	* 10	E Autolar Tata- R Chart		A	
1	4	8 C		1		4				1			84	N	0		-0
· p-	exp(a + bx)/[1+	exp(a + bx3)															
t							Inder	cept (a)	-3.941	4							
1400	me Probabilit	y of Purchase					Inco	me (b)	0.249	1							
C	1	2,43%															
	2	3.10%	1								1						
	3	3.94%			. 1	robal	bility (of Purcha	se								
	4	5.00%	1.1	-													
	5	6.32%															
	6	7.97%	1	00 M %				-		-							
1	7	10.00%					/				_						
	8	12.47%		80.0%			1			_	_						
1	9	15.45%									_						
	10	18.99%															
	11	23.12%															
	12	27.84%		112							_						
	13	33.11%		111N		1					-						
1	14	38.84%				/					_				2		
	15	44.90%		1000 L-1		-		24 44									
	26	51.10%	- L								-						
	17	57.28%															
	18	63.24%															
2	19	68.82%															
	theatt f	here Martine Should	-							1.1.5							1.1.1
-	- 2000 T 1	wantered proved	-0/	_	_	_	_		_		_	_	_	-	-	-	-

This is one way of calculating the probabilities of the new members using the output from the past data. However excel work is not most convenient way of calculating the probabilities when the number of the input is large. Since we have created logistics model from a past data, the past data output became training set for the new input. The probabilities of the new inputs can be calculated by training the new input using past model. Let us add few data point in a data frame and in R with some data point.

Ũ		Kitudio
File Edit Code Ver	· Plots Session Build Debug Tools Help	
Q · @ · B (👪 🖓 🖓 Geta Bachana 🔛 🗍 🔛 🖬 🗛	Mes •
() (heatest? +		-0
2 9 1 lg-data. 2 3 lg 4 lt-glm(P 5 lt	□ SourcenSee 9, Z. • □ frame(Income-c(10, 5, 4, 12, 20, 1 Purchase-c(1, 0, 0, 0, 1, 1, 0 urchase - Income, data-lg, family	
6 nd<-data. 7 Prdt <- p 8	<pre>frame(Income-c(1,4,7,8,10,17,18,2) redict(lt, newdata-nd, type="respo "response in the test of the test of the test of te</pre>	,27,10,11,45,56,40)) ese")
81 (log (mm)	1.	KSigt 1
Console -/ 10		-0
18 45 19 23 20 20 21 16 22 25 > ltc-glm(Purc > lt	1 1 0 1 hase ~ Income, data-10, family = "	binomial")
Call: glm(for	mula + Purchase ~ Income, family +	"binomial", data = lg)
Coefficients: (Intercept) -3.9414	Income 0.2491	
Degrees of Fre Null Deviance: Residual Devia > ndc-data.fra > Prdt <- pred	edon: 21 Total (i.e. Null); 20 Re 27.52 ALC: 17.55 mcG: 13.55 ALC: 17.55 mcGincone-c(1.4.7,8.10,17.18,23,27 lict(lt, newdata-nd, type="response	sidual .10,13,45,56,60)) ''

Let us predict the probabilities of the new inputs in nd data frame using the old data using predict function.

The probabilities of each new input are generated. Let's us compare this data with the data we have calculated in the excel sheet.





The predicted probabilities from the Excel and the R are same as shown below. This is expected as the model used to calculate the new probabilities is same data input.

RE I	HOME	NERT FROELEND	UT PORMALAS D	414	NOVEH .	vew				1951	-	100			Sector			Sprin
0.	1 mm	Caller +	B + X X ==		1 21	tap fact	54	red		186	12	2.1	1 - SA			2 <u>7</u>	•	
-	Carlo Carlo		· · · · · · · ·	10.40	*: EX	lege & Carl	w - 5	1. 16 1	2.5	Contrana	Format as	Cel 0	nat been	Frond		Sec.8.	Teth	
1			_				1	-		-	index.	silver.	·		eres.	144-	Select -	
-			- C	0			- 6	141	1			1.1	- 14	74			- 0	
nd.	ncome Pro	t Man	ually Calculated															
	1	2.4%	2.4%															
	4	5.0%	5.0%															
	7	10.0%	10.0%															
	8	12.5%	12.5N															
	10	19.0%	19.0%									7						
	17	57.3%	\$7.3%								_							
	18	63.2%	63.2%															
	23	85.7N	85.7%															
	27	94.2%	94.2%															
	30	97.2%	97.2%															
	33	98.6N	98.6N															
	45	99.9%	99.9%															
	56	100.0%	100.0%															
	60	100.0%	100.0%															
	1.00-0	A L Rep J Married	Terret Lines															

In the above example we could have added many more variable like male/female status of the customer, customers average monthly spend on house hold goods, customers household size etc. The readers can try out different combination for better understanding of the algorithm.

Now that you have run the logistics regression in R, how do you interpret the result and how to you test whether the model is good, bad or ugly.

0	RStudio
File Edit Code View Plots Session Build Debug Tools Help	
💽 • I 🔐 🕼 🕼 🕼 Getetlik fundien	
Console -/ @	0
<pre>> lt <- glm(Purchase ~ Income, data = lg, family = "binomial") > summary(lt)</pre>	
<pre>call: glm(formula = Purchase ~ Income, family = "binomial", data = lg)</pre>	
Deviance Residuals:	
Min 1Q Median 3Q Max -1.97092 -0.35108 0.03517 0.52729 1.82275	
coefficients:	
Estimate Std. Error z value Pr(> z)	
Income 0.2491 0.1185 2.102 0.0356 *	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	
(Dispersion parameter for binomial family taken to be 1)	
Null deviance: 27.522 on 21 degrees of freedom Residual deviance: 13.551 on 20 degrees of freedom AIC: 17.551	
Number of Fisher Scoring iterations: 7	
>1	

The first data point of interest is Pr(>|z|) in the out coefficient section. Since the value of Income variable is less than 0.05 we can say this factor is significant at 95% confidence interval.

Likelihood ratio test

Null deviance number is the chi-square for the model with only constant term and residual deviance is one with all coefficients in the model. For the null deviance the p value is 0.1542 which is greater than 0.05 which means the model has not comes from the logistics regression with only constant term.

```
0
File Edit Code View Plots Session Build Debug Tools Help
🔍 • 🔐 • 🔒 🔒 🖂 🗖 🐼 Go to file. To
                                                                                                            5
   Console ~/ 🖒
 R version 3.2.0 (2015-04-16) -- "Full of Ingredients"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
  R is free software and comes with ABSOLUTELY NO WARRANTY.
  You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
  R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
  Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
  [workspace loaded from ~/.RData]
     1-pchisq(27.522,21)
  [1] 0.1542357
     1-pchisq(13.551,20)
  [1] 0.8525126
    1-pchisq(27.522-13.551,21-20)
  [1] 0.0001856522
```

Similarly for residual deviance p value is 0.852 which is again greater than 0.05; which means that the model comes from logistics regression with constant term and other independent variables.

The difference between the chi-square of model with constant term and with independent variables p value is 0.00018, which is less than 0.05 which means we can reject the hypothesis that the deviance of the model with constant term and deviance of model with one variable added to it is same.

Hosmer Lameshow goodness of fit Test

In R you have to install Resource Selection package using below command

```
>install.packages("ResourceSelection")
>library(ResourceSelection)
> h1<-hoslem.test(lg$Purchase, fitted(lt), g=8)
> h1
Hosmer and Lemeshow goodness of fit (GOF) test
```

data: lg\$Purchase, fitted(lt)

X-squared = 2.7009, df = 6, p-value = 0.8453

The P value is >0.05 the model pass the test.

The logistics regression is very useful algorithm for the prediction of the binary output like what is probability of customer defaulting, what is the probability of customer churning, what is the probability customer will repeat the purchase and so on. It is also useful in the other areas such as prediction of the product sales probabilities selection of catalogue, the prediction of the probabilities of the product display and so on.

Learning for the Chapter

- How to segment using Google Analytics for behavioral studies onsite
- Cohort analysis- what is cohort and how to do simple cohort in excel
- Understand the meaning of Customer Lifetime Value (CLV) and how it is calculated
- Segmentation by clustering- algorithms and how to do clustering using R
- Market Basket Analysis to understand the association of the product in the baskets
- Logistics regression concept and utility in customer analytics

Chapter - V

DIGITAL MARKETING

"The Internet is becoming the town square for the global village of tomorrow"

By Bill Gates

Encyclopedia Britannica halts print publication after 244 years and go online only in 2012. By 2010 the print version of Encyclopedia Britannica was contribution to only 1% of sales. Britannica president Jorge Cauz says "The print set is an icon. But it's an icon that doesn't do justice to how much we've changed over the years," on being question on the how nostalgic he is about the print version.

Encyclopedia Britannica is not only iconic brand moving from print to digital version. You can look at many other brands in the printing business that are going to digital alone or majority of the readership is coming from the digital space. This shift in the trend has been happening from last 3 decades. The advertisement spend follows where readers goes, so digital marketing spend has been growing every years with respect to the traditional advertisement platform. As you can see from below graph the internet ads spend in 1997 was close to zero, and is only internet is a medium which is growing every year with respect to GDP in US. All other countries have similar trends.



The digital marketing channels are being increasing used by all types of the companies for their marketing activities due to the reach the digital channel provided and the ease of the measurement of the performance of the digital channels. The traditional marketing channels are still very much in demand but

the ratio of spend has clearly shifted to the digital side the actual ratio depends on the industry to industry. The digital channels are not just being used for the online transaction but for branding also, so companies without a proper website are also spending money on online branding.

The digital marketing has many advantages which a traditional channel does not have – in digital channels the company can instantly measure the performance of the campaign at great accuracy and hence there is flexibility to change the campaign or stop the campaign if it doesn't work. In traditional channels the performance from the campaign are mostly guesswork and the numbers are collected after the campaign are over; there is no time to react. In the ecommerce sites the sales from any campaign can be measures at any instant, the cost involved the number of users acquired, the number of users coming to the site. One can compare the performance of different digital channels at any moment and take decision to shift the budget. This flexibility makes optimization of the digital marketing prime objectives of the companies as the cost of acquisition has gone up and the customers are saturated with digital marketing ads.

The marketing department will always grapple with the different objectives to be achieved with the budget at disposal. Some of the digital channels can consume the budget very fast without providing significant return on investment. There will be mix of digital marketing channels and the category where the target has to be achieved.

In this chapter we will learn traditional optimization technique Linear Programming (LP) and use it for optimizing the budget allocation decision. LP can be used at the planning stages and later used to optimize the channels and category mixes based on the results from the digital campaign performances. Each digital channels has different ways of optimization at the advertisement level, we will not be talking about that optimization. In fact we will assume that each channel are sufficiently optimized using the digital marketing tools available in the market. The LP will be used for the allocation of the resources to achieve certain revenue or profit or cost objectives.

The basic of digital marketing will be explained for the readers not familiar with the digital marketing in section I and section II is about digital budget optimization.

Section - I

DIGITAL MARKETING BASIC

Ask any marketing students what are the basic of marketing? Invariable you will get answer as 4Ps and STP. Let me start with some basic understanding marketing and later relates these to the understanding of the digital marketing. The 4Ps in marketing stands for Products, Place, Price and Promotion and STP stands for Segmentation, Targeting and Positioning. Basically like traditional marketing the digital marketing also starts with products - what is product's functionalities, features and after sales services; Prices - how competitive is the prices in the internet space, is there any discount or offers; Places – where do customers look for the products Google Search or Display or Site or Facebook, how do I make products available to customer; and Promotion -How do I promote the products, which channel is more effective. Once above question has been answer the marketer has to start segmenting their customers based on different criteria - demographics, geographic and psychographic. The segments which are significant and have potential are targeted with the advertisement for that segment. The digital marketing is no different from traditional marketing in the above context. However the way it is carried out is totally different. For example for running a TV ads on women wear the target channel will be entertainment channel, TV serials on family Drama, 7 to 11 pm. The same audience will be targeted in digital channel say Facebook by Gender - Female, Age -18 to 45 years, time - 9 am to 10 pm, location - cities.

There are hundreds of the digital marketing channels existed in the market. Instead of going through each of the channel separately, we will classify the channels into the different types based on the how the cost of the advertisement is accounted for. For each type of channels we will learn how the key metrics of the performances are defined. This is important for the harmonization of the different marketing channels and their objectives into one single number that can show the relative performances of each channels and how we can optimized the channels based on the harmonized metrics.

The company has certain plan for the marketing activities in advance and according budget is allocated for the branding and sales. The branding spends are not expected to provide instant result to the sales of the company but it certainly has long term impact on the sales. On the traditional marketing side we can broadly classify marketing activities as Above The Line (ATL) and Below The Line (BTL). The ATL are those which target the mass audience like TV ads, radio ads and Newspaper ads. The BTL are more of one to one direct marketing like distribution of coupon, sample testing, pamphlet, kiosk etc. The BTL are expected to provide more instant sales boost and ATL are more towards branding side and has long term impact on the sales.

In digital marketing also there are certain channels which are more branding oriented and there are channels which are performance oriented. The company decides the digital spends in each channel according to the objectives of the branding and sales push. Along with the paid channels company spends money on the Organic Search and Social Media channels. The Organic search is nonpaid customer but there is range of activities to be carried out to increase the keywords ranking in the organic searches. Similarly the social activities traffic is non-paid but the activities involve cost to the company.

The focus of this chapter will be on the performance marketing side. Before we start analyzing the data for the performance of each channel we have to understand the mechanics of each channel and what are the key performance levers of the channels and how do we optimize for the best performance.

5.1.1 Digital Marketing Ecosystem

Before going into the specifics of the digital marketing it would be good idea to understand what actually is digital marketing all about? What are the component of the digital marketing and why it is different from traditional marketing?

There is no single term that defines the digital marketing. Broadly digital marketing comprises of activities such as search engine optimization (SEO), search engine marketing (SEM), content marketing, influencer marketing, content automation, campaign marketing, and e-commerce marketing, social media marketing, social media optimization, e-mail direct marketing, display advertising, e-books, optical disks and games, and any other form of digital media. It also extends to non-Internet channels that provide digital media, such as mobile phones (SMS and MMS), callback and on-hold mobile ring tones.

That is too broad to understand. Let's focus on the paid marketing.



Main players in the digital advertisement ecosystems are

Ad Exchange: An ad exchange is a technology platform that facilitates the buying and selling of media advertising inventory from multiple ad networks. Prices for the inventory are determined through bidding. It's basically a competitive bidding platform for the media inventory. Examples of Ad Exchanges are DoubleClick from Google, Microsoft Ad Exchange, Rubicon etc.

Network: An online advertising network or ad network is a company that connects advertisers to web sites that want to host advertisements. The key function of an ad network is aggregation of ad space supply from publishers and matching it with advertiser demand

Publisher: Publishers are one who has the space for ads to sell it to advertisers.

Advertisers: The individual or company who want to advertise their product or services or brand etc. Components of the ad ecosystem are

Basically there are series of publishers with ad spaces to sell and there are series of advertisers who want to use that space. The ad exchange help buying and selling on real time through competitive bidding process.

An example is that you are a regular blogger on latest mobile phone. Since you have good traffic, you want to rent available space to ecommerce companies selling mobile phone. Since you cannot contact all advertisers on your own, you would take help of network to rent your space through ad exchanges. Here you are publisher.

How do you integrate data from end to end that is from acquisition of the customer through digital marketing channel, website behavior of those

customers, their interaction with CRM and then recycle them through remarketing effort such as personalized mailers by understanding their profile is the routine work of any digital marketing manager. Below picture is one of the typical flow of information or customer in the digital world. Note that digital marketing is not confine to external customer only but it also should include on-site targeting through such tools available like doing A/B testing, Personalization of communication, cart abandonment follow-up and product recommendation. All such activities can be called in-site targeting.



Now we have good idea of the overall digital marketing world, let's look at each of the digital channel

5.1.2 Costs per Click (CPC) Advertising

The cost per click advertisement is among the oldest and the most widely used advertisement methods in the digital space. In CPC the cost is incurred by the advertiser once a viewer click on the advertisement. The cost of each click is function of the bidding at the ad exchanges or any other real time bidding systems. The cost for the period is the sum of all the cost of the clicks for that period. The cost per click is likely to vary for each click depending on the dynamics of the bids. However the average cost per click is the metrics one has to use for measuring the cost of the advertisement. In CPC system each ads click will cost the advertiser but the objectives of the advertisement may not be achieved in each clicks. The clicks on the ads lead to the advertiser's website where the users are expected to perform certain objectives like purchase or fill a form or spend time on the pages and so on. There is always a real possibility that the users clicks on the ads closes the browsers before landing on the advertisers website or the landing pages not responding when customer clicked on the ads. For the advertisers each click should translate into a visit in the website but due to various technical reasons the visits numbers will always be lesser than the clicks numbers. In CPC the advertisers has to pay for the clicks irrespective of the clicks translating into a sessions and fulfilling the objectives of the advertising.

Now let's look at the other metrics in the CPC form of the advertisement. The number of the eyeball that advertisements garner is called **Impression** in the common terminology. The number of **clicks** per impressions are called **click through rate** (CTR) which is a percentages. The clicks that reached website and fulfill the objectives or goals of the campaign are called conversions. Say the campaign was to increase the orders from the site; the number of the orders from the campaign is called conversions. The conversion per clicks are called **conversion rate (CR)**. The cost per conversion is important number you have to remember. The cost is the sum of all the CPCs and the conversion is the goal completion for that period, hence cost per conversion gives information about how much to money has to be spend to get as many orders. For easy of notion cost per conversion will be demoted as CPO (cost per order).

CPO = cost / order = avg. CPC * clicks / order = Avg CPC*clicks / clicks*CR = Avg. CPC / CR

From the above formula the CPO is inversely proportional to CR. Higher the CR lower the CPO. Hence sometime the higher bid ads may have lower CPO if CR is higher.

To summarize the key matrices and formula involved in CPC advertising:-

Impression: No of times ads was displayed to the users

Clicks: The count of the ads being clicked by the users

Cost per Clicks (CPC): The cost of the clicks as per bidding system of the channel

Conversion: The number of clicks that translate into the fulfillment of the goals set in the site.

Click through Rate (CTR) = clicks / Impression **Conversion Rate** (CR) = Conversion / Clicks Average cost per clicks = cost / clicks

Cost per conversion = cost / conversion

Revenue Per conversion: revenue from the channel / conversion from the channel

Cost per Acquisition: Cost / New Customer Acquired

Return of Investment = Revenue per Conversion * gross margin% / cost per conversion

Each channel has much variation of the above matrices. For example Google Adwords have three conversion numbers – converted clicks, conversion and view through conversion. Similarly Facebook has conversion in day and within 28 days. These lead to many ways of interpreting the conversion number. Hence companies tend to use attribution modeling to come up with the actual and assisted conversion and what weightage to be given to each channel from the conversion. Similarly for other metrics also there are many ways of interpreting and each company follow certain ways of using it. The cost per clicks ads are the oldest and most widely used in digital marketing world, hence we will look at the major channels with their mechanics and how to optimize the campaign.

5.1.3 Google Adwords

5.1.3.1 Display networks

The display is one of the common forms of advertisement. As shown in picture below the display usually have an image with the pictures of the products and services with the message of the advertisement. The communication is very important for the display ads because the display ads are push form of the advertisement where users are not actively seeking products unlike that of search ads. Display is push form of advertising but it is not mass targeting, the advertisers has the option to choose the target group of the customers based on the users segment created by ad exchanges. For instance if you want to sell a sport goods then you can target those users who follows sports, love outdoor or members of sports group. The display is usually used for branding and spreading message very fast. The display can acquire impression pretty fast as the user base is large and the ads can be pushed to any users in the networks irrespective of whether users are interested in the products and services. When display is solely used for the branding purpose the another metrics that can be used is the CPM(cost per mille) or Cost per impression where advertiser has to pay per impression irrespective of the clicks and sessions generated.

Below is an example of display ads.



You can add or delete certain metrics for the display

lodity columns			10000000000000		
Select multica			Drag and drop to reorder		
Performence =	(min	Add all columns	Caritada		
Conversions +	inge.		Entry .		
Tool Metros -	CTR .		famalititus		
Call Infolds .	Als OPC	-			
Federitarios photos Viento -	Arg CPM		Caulanda Proston X		
Congle Analytics +	Aug Pes.		Cion X		
Change history +	figure cent		and X		
Genal Mathica	Invalid clicks		(TR X		
Custom columns =	Engagements		: Ag 0FC 34		
			Cast X		
			Ang Pag		
			(mand data		
			×		
			Concernance N		
			Cick conversion rate X		
			Conversions X		
All enabled campaigns +	Segment v	Filter - C	olumna • 🛛 🛓 Find car	npaigns Q Wew Chang	pe History
All enabled campaigns + Modify columns Select metrics Attributes	Segment v	Filter - C	ohanna + E A Find car	Drag and drop to reorder Campaign	pe History
All enabled comparises + Modify columns Select metrics Attributes Performance	Segment +	Filter + C	Add all columns	Drag and drop to reorder Campaign Budget	pe History
All enabled campaigns + Modify columns Select metrics Atributes Performance Conversions	Segment v	Filter + C	Add all columns	Drag and drop to reorder Campaign Budget Status	pe History
All enabled campaigns + Modify columns Select metrics Attributes Performance Conversions Attribution	Segment -	Filter - C	Add all columns	Drag and drop to reorder Campaign Budget Status	pe History
All enabled campaigns + Modify columns Select metrics Attributes Performance Conversions Attribution Social Metrics	Segneri -	Filter - C Converted clicks Cost / converted Click converted Cost / converted	Add all columns	Drag and drop to reorder Campaign Budget Status Campaign type	pe History X
All enabled campaigns + Modify columns Select metrics Attributes Performance Conversions Attribution Social Metrics Call details Correction profet	Segment -	Filter - C Converted clicks Cost/ converted Click conversion Conversions	Add all columns Add all columns Click rate	Drag and drop to reorder Campaign Budget Status Campaign type Campaign type	pe History X
All enabled campaigns + Modify columns Select metrics Atributes Performance Conversions Atribution Social details Compative metric Performance (Activ	Segment -	Filter - C Converted clicks Cost/ converted Click converted Click conversions All conv.	Add all columns Add all columns Click	Drag and drop to reorder Campaign Budget Status Campaign type Campaign subtype	pe History X X
All enabled campaigns + Modify columns Select motrics Attributes Performance Conversions Attribution Social Metrics Call dealts Competitive metric Performance (Acth Decelar dealts)	Segment v in in in in in in in in in in in in in	Filter - C Converted clicka Cost / converted Click conversion Conversion Conversion All conv. View-through co	Add all columns click • rate • tr. •	Drag and drop to reorder Campaign Budget Status Campaign type Campaign subtype Caicks	pe History X X
All enabled campaigns + Modify columns Select metrics Attributes Performance Conversions Attribution Social Metrics Call details Competitive metric Performance (Acth Google Analytics Basch Metrics	Segment +	Filter - C Converted clicks Cost / converted Click converted Conversions All corv. View-through co Cost / all corv.	Add all columns Add all columns Click Clic	Drag and drop to reorder Campaign Budget Status Campaign type Campaign subtype Clicka	pe History X X X
All enabled campaigns + Modify columns Select metrics Attributes Performance Conversions Attribution Social Metrics Call details Competitive metric Performance (Addi Google Analytics Reach Metrics Chance Ristory	Segment +	Falter - C Converted clicks Cost / converted Click conversion Conversions All conv. View-brough co Cost / all conv. All conv. rate	Add all columns Add all columns Click C C C C C C C C C C C C C C C C C C C	Drag and drop to reorder Campaign Budget Status Campaign type Campaign subtype Clicks Impr.	pe History × × × ×
All enabled campaigns + Modify columns Select metrics Attributes Performance Conversions Attribution Social Metrics Call details Competitive metric Performance (Adb Google Analytics Reach Metrics Change Initiony Graul Metrics	Segment v Segment v	Fatter + C Converted clicks Cost / converted Click converted Click conversion Conversions All conv. Vies=brough co Cost / all conv. All conv. value	Add all columns Add all columns Click rate rate rate	Drag and drop to reorder Campaign Budget Status Campaign type Campaign subtype Clicks Inpt CTE	pe History X X X X
All enabled campaigns + Modify columns Select metrics Attributes Performance Conversions Attributen Social Metrics Call details Competitive metric Performance (Attributes) Reach Metrics Change history Gmail Metrics	Segment +	Filter - C Converted clicks Cost / converted Click conversion Conversions All conv. View-through co Cost / all conv. All conv. rate All conv. vate	Add all columns Add all columns Click Charter Content	Drag and drop to reorder Campaign Budget Status Campaign type Campaign subtype Campaign subtype Campaign subtype Campaign subtype Campaign subtype Campaign subtype	pe History × × × ×
All enabled campaigns + Modify columns Select metrics Attributes Performance Conversions Attribution Social Metrics Campetive metric Performance (Acth Coogle Analytics Reach Metrics Change history Graal Metrics YouTube Earned a Curster columns	Segment +	Filter - C Converted clicks Cost / converted Click conversion Conversions All conv. View-through co Cost / all conv. All conv. value All conv. value All conv. value	Add all columns Add all columns Click Charter Ch	Drag and drop to reorder Campaign Budget Status Campaign type Campaign subtype Cicks Impc CTR Avg. CPC	pe History × × × × ×
All enabled campaigns + Modify columns Select metrics Athoues Performance Conversions Athousion Social Metrics Call details Competitive metric Performance (Ach Coogle Analytics Reach Metrics Change History Gmail Metrics YouTube Earned a Coustorn columns	Segment +	Filter + C Converted clicks Cost/converted Click conversion Conversions All conv. View-brough co Cost / all conv. All conv. rate All conv. value Value / all conv.	Add all columns Add all columns Colick Colic	Drag and drop to reorder Campaign Budget Status Campaign hype Campaign subtype Clicks Impr. CTR Avg. CPC	pe History × × × × ×
All enabled campaigns + Modify columns Select metrics Attributes Performance Conversions Attribution Social Metrics Call dealls Competitive metric Performance (Adb Google Analytics Reach Metrics Change history Gmail Metrics YouTube Earned a Custors columns	Segment +	Fater + C Converted clicks Cost/ converted Click converted Click conversion Conversions All conv. View-brough co Cost / all conv. All conv. value All conv. value All conv. value Value / all conv.	Add all columns Add all columns Add all columns Click	Drag and drop to reorder Campaign Budget Status Campaign type Campaign subtype Campaign Sub	pe History × × × × × × ×
All enabled campaigns + Modify columns Select metrics Attributes Performance Conversions Attributen Social Metrics Call dealls Competitive metric Performance (Acth Google Analytics Reach Metrics Change history Gmail Metrics YouTube Earned a Custorn columns	Segment +	Filter + C Converted clicks Cost / converted Click converted Click converted Conversions All conv. All conv. All conv. value All conv. value Value / all conv.	Add all columns Add all columns Click Characteristic Click Characteristi	Drag and drop to reorder Campaign Budget Status Campaign type Campaign type Campaign subtype Campaign Subtype	pe History × × × × × × × × × ×
All enabled campaigns + Modify columns Select metrics Attributes Performance Conversions Attribution Bocial Metrics Call dealls Competitive metric Performance (Acth Google Analysics Reach Metrics Change history Gmail Metrics YouTube Earned a Custors columns	Segment +	Falter + C Converted clicks Cost / converted Click converted Click converted Conversions All conv. Vies=brough co Cost / all conv. All conv. value All conv. value All conv. value All conv. value	Add all columns Add all columns Click Add all columns Click Add all columns Click Add all columns Add all colu	Drag and drop to reorder Campaign Budget Status Campaign type Campaign type Campaign subtype Campaign subtyp	pe History × × × × × × × × ×
All enabled campaigns + Modify columns Select motrics Attributes Performance Conversions Attribution Social Metrics Call dealls Competitive metric Performance (Adb Google Analytics Reach Metrics Change history Graal Metrics YouTube Earned a Custors columns	Segment +	Fatur + C Converted clicks Cost/ converted Click converted Click conversion Conversions All conv. View-Enrough co Cost / all conv. All conv. value All conv. value All conv. value All conv. value	Add all columns Add all columns Add all columns Click	Drag and drop to reorder Campaign Budget Status Campaign type Campaign type Campaign subtype Campaign subtyp	e History × × × × × × × × ×
All enabled campaigns + Modify columns Select metrics Attributes Performance Conversions Attribution Social Metrics Call dealls Competitive metric Performance (Adb Google Analytics Reach Metrics Change history Graal Metrics YouTube Earned a Custors columns	Segment +	Falter + C Converted Clicks Cost/ converted Click converted Click conversion Conversions All conv. View-through co Cost / all conv. View-through co Cost / all conv. All conv. value All conv. value Value / all conv.	Add all columns Add all column	Drag and drop to reorder Campaign Budget Status Campaign type Campaign subtype Clicks Impc CTR Avg. CPC Cost Avg. Pos. Converted clicks Cost/converted click	pe History × × × × × × × × × × ×

Below is the sample of a campaign run by www.myweb.com for its apparels category in Google Adwords display ads. The portfolio of all apparels display campaign have

CTR of 0.46% which is calculated from clicks/impressions (5995 / 1247463)

- Converted clicks is 32
- Click conversion is 0.57% which is calculated as unique conversion / clicks (32 / 5994).
- Averages CPC is calculated as total cost by clicks (37747.51 / 5994) = 6.75 and
- Cost per clicks conversion is 1179.61 which is calculated as cost / click conversion.

There are two other conversion metrics – conversion which is based number of total conversions within time period of 30 days and view-through conversion which shows people converted after viewing the ads (impression) without click on that impression; basically it attribute according to last impression. In the web analytics system the conversion from the Adwords will typically be tracked at last touch interaction for cross channel attribution, hence conversion shows in the Adwords are likely to have slight difference from Web analytics system. However there are other methods such as first touch attribution, linear attribution etc.

Campaign	Suipe	Campaign type	-pressions	0 ets	CTR	Arg. 090	Com	Convert edicions	Cost / converte diclick	Olek conversio nrate	00-14-50 75	conversio n Rate	Verior shroug h conv.
Daplay Myweb Jeans	2,000	Display Only	\$4895	84	0.56%	5.39	452.78	0	0	0.00%	0	0.00%	
Dapisy Myweb Takint	2,000	Diapisy Only	5851	52	0.89%	3.71	192.62	0	¢	0.00%	0	0.00%	1
Daplay_Myweb_Shore	8,000	Display Only	\$1650	264	0.51%	4.76	1,257.94	2	628.97	0.76%		1.89%	
Daplay_Myweb_Men Shoes	2,000	Display Only	20109	71	0.55%	5.49	3 89.66	0	6	0.00%	0	0.00%	
Display, Myweb, WomenShoes	50,000	Display Only	187708	628	0.55%	10.37	6.912.24	1	1.045.17	0.96%		1.27%	13
Daplay_Myweb_Bags	2.000	Diaplay Only	3166	20	0.63%	4.25	45.C2	0	4	0.00%	0	0.00%	
Daplay_Myseb_Purse	2,000	Diapity Only	7014	41	0.58%	7.33	300.36	0	6	0.00%	- 0	0.00%	
Daplay, Mysee, Formal Shire	50,000	Display Only	\$35794	837	0.25%	8.44	7,048.58	11	64214	151%	- 19	2.27%	12
Daplay_Myset_Formal Parts	10,000	Display Only	154307	730	0.54%	8.35	6.096.37	1	6.095.37	0.14%	1	0.27%	- 14
Daplay_Myweb_Sets	5.000	Display Only	129056	487	0.36%	4.95	2,398.90	1	2.395.90	0.21%	1	0.45%	1
Daplay, Myweb, Plus Spes	3,000	Diaplay Only	67342	309	0.57%	6.64	2.543.95	0	0	0.00%	0	0.00%	1
Daplay_Myweb_Ads	15.000	Diability City	220071	1992	0.59%	5.23	10,415.05		P45.32	0.33%	13	0.30%	1
Total			1247463	5595	0.45%	6.75	37,747.51	32	1,179.61	0.57%	52	0.93%	81

Based on the objectives of the display ads we can calculate the performance of the ads. In case of the CPM we are more concern with the number of impressions and hence conversion data are not much of importance whereas in case of campaign with goals of transaction, the conversion data are important. For customer acquisition campaigns we can find the new customers acquired from the other system and calculate the cost / new customer acquired. Typically CPM will be KPI of cost / thousand Impressions, conversion campaign will have Margin or Revenue / cost and customer acquisition campaign will have customer life time value / cost per customer acquired. There can be multiple objectives for different portfolio. The optimization and targeting foe each campaign will be based on any of the above objectives.

5.1.3.2 Search Engine Marketing (SEM)

The search advertisements are pull based kind of the advertisement where users search for particular keywords in the Search engine like Google and Bing. The advertisement is displayed as per the users search keywords. The advertisers have to bid on the keywords and the search engine display ads based on the algorithm in the search engine. On searching user cars I can see at four ads in the sequence. The ads at the top are the most coveted spot. The position of the ads in the search engine marketing ads display is based on many factors decided by search engine. For example the Google consider the quality score, bidding and expected impact of extensions and other ad formats. The quality score is further calculated on the click through rate, landing pages and the ad relevance. The high bidding itself does not guarantee higher ad ranks. If I click on the ads I will be taken to the landing page which is the page in the advertisers' website where the users are expected to land on clicking of the ads from the particular keywords. The keywords and the landing pages products and services relevance is very important criteria for increasing the quality score. The keywords research is one of the important components of the search ads which can be done using keyword planner in adwords. The click through rate and conversion rate are expected to be better than display form of advertising because the users are actively seeking the products or services by searching the product. However the scalability of the search ads solely depends on the search volume of the keywords in the search engine. Below picture are search ads from key used cars. There are four ads in the result.



In search ads we have additional metrics added in the output for analysis purpose. In below picture I have added Search Impression share, search exact match Impression Share and Search Lost impression Share which is used for measuring the competitive bidding and market share of the search in the period and location.

Campadora /	of groups	Seconda	-	Asymorth	Authences	A2 415		Denenariona			
Repeats Neg	ative heywords	Dearch	terms								
All but removed i	eywords +	Segm	ent =	Filter +	Columns +	l:	A			٩	View Change History
Modify col	umns										
Select me	trics							0	rag and	drop to r	eorder
Attributes					Re	move all	columns		Keynord		
Performan	CØ.			Search Impr. sh	are			i i	Campalo	n .	
Conversio	NS			Search Exact m	atch 15				Ad entrue		
Eld simula	la l			Search Lost IS	rank)				Crack of	-	
Competitiv	re metrics	-									
1 ALARA STATE									Max CPI	ę.	
									Campa	ign type	×
									Clicks		×
									Impe		×
									CTR		×
									Avg. CF	PC	×
									Cost		×
									Ave. Po	1d.	×

In search ads position of the ads are very important. The click through rate for the first position can be significantly higher hence every player will aim to achieve first ads position in each keyword. The position of the keywords is determine be quality score and the bidding. The Google maintain quality score of each keyword based on the CTR, landing pages and ads relevance. In below picture I have added quality score columns.

Caywords Negative keywords	Dearth term	•							
All but removed keywords +	Segnent -	Filler +	Columns +	Ŀ	A			٩	View Change History
Modify columns									
Select metrics							Drag and dro	10 TO 11	eorder
Attributes	Laute					Keyword			
Performance	Eld strategy					Campaign			
Conversions	Bid strategy type				1 3	Ad oroup			
Rid creately a		Final URL					Status .		
Competitive metrics		Tracking template			1		11. 000		
		Custom parameter				8	Max.OPC		
		Qual score			-		Campaign	504	×
		Est. first page t	Ded .		1	1 2	Circles		×
		Est top page b	NG						~
		Est, first positio	on bid				ange.		×
		moton type			10		CTR		×
		Nayword ID			1.4	L			
							AND OPC		×

Below table is an example of performance of a Search campaign of laptop with different combination of keywords. Additional metrics Quality score, average position of ads and Impression share of the keywords. The keywords ads are mostly performance oriented. The first keyword has quality score of 10, Average CPC of 4.56 and impression share of 50%, cost per clicks of Rs. 310 and CTR of 3.4%. There is scope of increasing the bidding to capture the rest of the impression share without increasing the avg cpc much. Assuming average price of laptop of Rs. 30000 the Revenue by cost is 97 which is good performance. Similarly second keywords has already achieved 100% of impression share hence no scope of increasing the clicks by increasing the bids, moreover it has revenue by cost 27. Hence instead of second keywords we can plough more money in first keywords. There are many ways of bidding and in market there are specialized tools to do this allocation automatically based on different performance metrics.



Keywords ads are most scientifically done ads with whole lots of metrics that can improve the performance. There are not set method each campaign managers follow different strategy based on the objectives of the company. Above example is just to provide you an understanding of the key metrics to be tracked and used. The analytics have lots to play in performance of the
search ads. There are many tools which does the automatic bidding based on the objectives of the company.

5.1.3.3 Product Listing Advertising (PLA)

The product listing ads are similar to SEM in the way it works. The difference is that instead of the bidding on the keywords the bidding is done automatically based on the products in the list. The products title and description contains many keywords the search engine matches the keywords searched with the products title and description keywords. As shown in the screenshot below the products images and prices are displayed on the search engine pages itself. When user clicked on the ads it takes user to the product pages instead of any dedicated landing pages. This is another point of difference between PLA and SEM.



In below table the PLA campaign for different category is shown. For the performance of PLA campaign the feed optimization through product title, description, Google Product category and offers are important parameter which is critical for better CTR. The product prices and images are shown for different competitor in a row. There is not bidding for the position of the ads but the impression share is important parameters. The bidding can be done at the product levels which make it really important for the feed optimization. The objective of the PLA campaign is to land customer to the product page and complete the transaction. Therefore the most important metrics for the PLA campaign would be product level Selling Price to cost per orders or margin at product level to cost per order.

Campaign	Campaign type	Impressions	Clicks	CTR	AVE CPC	Cost	Converted	Cost / converte d click	Click conversi on rate	Search Impr. share
PLA_books	Shopping	606,535	10,420	1.72%	6.73	70,175	137	512.22	1.31%	89.47%
PLA_Bike Accessories	Shopping	224,752	3,336	1.48%	6.99	23.315	14	1,665.34	0.42%	79.14%
PLA_TOY	Shopping	63,638	1,567	2.46%	13.37	20,954	11	1,904.88	0.70%	93.61%
PLA_PET	Shopping	229,119	5,387	2.35%	10.1	54,412	43	1,265.39	0.80%	86.72%
PLA_Apparels	Shopping	100,020	1,962	1.96%	9.72	19,068	10	1,906.82	0.51%	43.21%
PLAMobiles	Shopping	147,576	5,558	3.77%	4.2	23,340	98	238.16	1.76%	24,40%
PLA_Watches	Shopping	155,726	2,974	1.91%	8.12	24,137	27	893.98	0.91%	56.26%
PLA_Footwears	Shopping	53,746	1,331	2.48%	18.44	24,548	14	1,753.46	1.05%	80.29%

Since competitors' product with their prices are also displayed higher CTR is possible if your product is competitive. It is always good to create separate list of product like top seller, lowest price, long tail items etc. and bid accordingly. These list can be refresh every day.

5.1.3.4 Remarketing

Remarketing is a very broad term which can include many things; here I will use remarketing in the display advertising form. The remarketing is basically targeted to the users who have visited the site before. The advertiser can set the rules under which a particular product is to be shown to the users. For example if user viewed products and did not buy then remarketing can be configured to show same products or similar products. Or if person buy a mobile then remarketing can be configured to show accessories related to that mobile phones. Each companies running remarketing has different algorithms but basic constructs is that it read the users behavior and target them with products that is most likely to be bought. The look and feel of remarketing is same as display ads so I am not attaching any screenshot here. The remarketing campaign can also be used to upsell and cross sell the products based on the browsing and purchase behavior of the customer.

Remarketing list for search ads is another remarketing feature which allows Google to target and adjust the bids based on the behavior of the users who have previously visited the site. As remarketing metrics are similar to the display ads I will not add example in the section. The optimization of the remarketing lies in the audience creation through better segmentation and targeting. The frequency of ads display and time period of the display would be important parameter for the remarketing. For example a buying an IPhone can be remarketed with IPhone accessories immediately with frequency of 2/3 views per day for next 30 days. A person buying Grocery items can be shown same stuff after say 7 days with 2/3 frequency per day. It is important to understand the customer behavior related to the category of product to be remarketed. Random targeting with unlimited frequency will led to costly click but no conversion and customer might also get irritated. The crux of remarketing is precise segmentation and retargeting.

5.1.4 Facebook Advertising

The Facebook has different form of advertising but I will discuss them one by one in detail because the mechanics of the Facebook advertisement are all similar. The Facebook as a platform collect lots of user's data, the segmentation on those data and targeting is the mainstay of the advertisement. Unlike Google Facebook doesn't have pull based ads, it's all push based because people do actively not search for any products or services in the Facebook. The searches happen in Facebook are mostly related to the profile pages or a particular topics/group and so on. Below is one form of Facebook advertisement. The click on the ads will take user to the landing pages.

5.1.4.1 Domain Ads

The Facebook ads shown at the right column of the Facebook page. This is simple ads with a title, a short description, and the URL to be displayed. Nowadays, it usually underperforms in terms of its click-through rate (CTR), but its cost can be very cheap in comparison to other ad types in this list.



5.1.4.2 Page Post Link (Newsfeed)

This is the most common of all the Facebook Ads types. It's ideal to promote your external website. Page Post Link ads feature a big image that's great to catch a user's attention. These ads perform really well and have the side benefit of generating Likes for your page

f Search Facebook	Q	🖁 Janua Hanne Handlinnish: 🏦 🖓 🕞 👘
	de Like 🔠 Commant 🍂 Share	DIAPERS
	Suggested Park	075
	Updrad to the total	n Page *
	Grow faster with mentorship from Top Data Scientists and a 3-r capstone project	Select Depens @ Rs.5 www.fistury.com
	🍠 🎥 🕬	Huppes, Marry Pera & nore. Coupon FEOPIS
		REEN Share and the dest of
		Privacy Terms Costess Advertising Ad Chalces P Mare - Viscotes 0 2016
	PG Diploma in Data Analytics - 9 month online program	
	Vector charace to be a part of career hair at let to	ars More
	0.12	Day H-

5.1.4.3 Multi Products Ads

The Multi Product Ads are carousel with five products with their landing page links, images a, description etc. in the ads. Extremely useful for advertisers looking to promote multiple products from their store or for marketers looking to promote different posts and offers to see what works and attracts leads with their audience. The ad type, available only in the Newsfeed, desktop and mobile, is very similar to the normal page post link ad, but instead of featuring only one link, it supports up to five products with links showcased. With a single ad we can now promote five products, each one with its own picture, link and title making clicks on the ad much more likely.



5.1.4.4 Dynamic Product Ads (DPA)

Facebook's dynamic product ads are like re-marketing display ads on steroids. They target users based on past actions (or inactions) on your website or application with a perfectly timed ad. All you have to do is upload your product catalogue to Facebook and double-check that your Facebook Pixel is installed correctly on your site's pages. Facebook handles the automation and re-targeting. The product feed from Google PLA can be directly used in the Facebook DPA feed.



Like Google Adwords campaign, Facebook campaign performance can be measured using usual CTR, CPC, CR, Reach etc. As such Facebook is an engagement tools along with the performance hence the objectives of the ads need to be defined for the correct measurement. Below is an example of Facebook campaign with performance parameter.

Campaign name	Impressions	Clicks (All)	CTR (AII)	Amount spent (INR)	CPC (All) (INR)	Conversion (1 Day)	Frequency	Reach
Facebook Apparels	13503	62	0.46%	1.989	32.08	6	1.00	13503
Facebook books	24331	370	1.52%	4.000	10.81	111	1.01	24093
Facebook Mobile	123782	1108	0.90%	20,000	18.05	417	1.00	123782
Facebook Furniture	45446	209	0.46%	14.346	68.64	12	1.04	43895
Facebook Footwear	3214	536	16.68%	440	0.82	1	1.03	3127
Facebook Kids	10066	23	0.23%	2,475	107.60	2	1.12	8972

5.1.5 Pay per Order Channels

The pay per orders channels is form of advertisement in which the payment is made based on the transaction happening in the site. The term like impression, a click has no meaning. The number of sessions that partner brings has no meaning because the payment will be based on the transaction only. These forms of the advertising are mainly affiliates. The ecosystem of affiliates is complex but for us we just need to concern with the orders it brings to the site. There could be different rate for the orders generated such as rate based on selling price of product or margin of the products. There can be rate based on the lead and final delivery of the products.

5.1.6 Bulk Form of Advertising

I would classify Email and SMS as the bulk form the advertisement because both systems involves huge customer base and payment is usually done of the basis of the SMS sent or email sent irrespective of the receivers opening the mail/SMS or not. There may be some form of the segmentation but the emails and SMS are sending to many users at the time. The responses are usually low for the number of SMS or emails sent. Like any other channels the performance is based on the message and the offering. The key metrics for emails are open rate, click rate and conversions rate. Since the payment is made in bulk we will be more concern with the cost / order from the sources to calculate the CPO. The increase in open rate and click rate is likely to increase the conversion rate but it is function of many other variables. Similarly for the SMS the CPO is the bulk cost / orders from the SMS in the period.

I hope readers are now more familiar with the digital marketing worlds. I have briefly described it to make you familiar with the basic terms. The detail optimization and working of each form of the advertising can fill a book. As mentioned earlier the objective of this section is to get all channels to level playing fields and use same metrics for the measurement of the performances. The same metrics will be used for planning and optimizing in the next section.

5.1.7 Channel-wise Google Analytics Data

The web analytics tools including Google Analytics captures the sources of the traffic based on the tagging of the url of the landing pages. The tagging is done on each sources using UTM parameter where source, medium and campaign name is defined. For example if I am running a campaign in Facebook for meant Category for Diwali Festival, I can Apparels tag my url www.myweb.com/apparels/diwali with utm source = Facebook&utm medium utm campaign Paid Diwali added to & is url = (<u>www.myweb.com/apparels/diwali</u>&utm_source = Facebook&utm_medium = Paid & utm_campaign = Diwali). This provided unique way of tracking the sources of traffic even if multiple sources bringing the traffic to the same landing pages. In fact UTM parameter is not related to the landing pages. The sources of the traffic can be extracted from the Acquisition -> Source/Medium tab of the Google Analytics. The tab will provide the number of traffic, transaction and revenue from the sources in the period. Same data can be extracted to the database directly using the Pentaho Data Integration API interface. The quality of the classification of the channels will entirely depends on the way UTM are added. The cost data is not available in the Google Analytics except for the Adwords (Google) campaigns. At the sources level the cost data are available on the real time. For the pay per order sources the commission is fixed, hence the cost can be calculated instantly.

The return of the investment of the marketing for each channel can now be calculated using the revenue and the cost data available. The comparative ROI for each channel will provide us with information about which channel is giving me better return and so on. The ROI data can be used to shift the budget from the lower ROI to higher ROI channels.

5.1.8 Search Engine Optimization (SEO)

According to Webopedia, Search engine optimization is a methodology of strategies, techniques and tactics used to increase the amount of visitors to a website by obtaining a high-ranking placement in the search results page of a search engine (SERP) -- including Google, Bing, Yahoo and other search engines.

When someone searches on search engine such as Google.com the websites are displayed as the result. The search engine index pages of websites and threw result according to the algorithms considering multiple factors. The search engine tries to have most relevant pages on the top so that users find the relevant information it is looking for.



In the above screenshot I searched for Mobile Phone in Google search engine. The Google display advertisement at the top – PLA ads and Search Engine Marketing ads. Once the ads section is completed the organic results are displayed. At the top is snapdeal site, followed by ndtv gadget, Wikipedia and so on. The click on the organic result is free for the sites whereas the clicks on the ads cost companies. Hence SEO being free traffic every companies want to have top listing in the result.

At the bottom you can see pages 1,2,3.... The number of results can runs into thousands and millions. But who cares after few pages. After 2-3 pages the pages displayed tends to be less relevant. People mostly tend to look at first few pages only. So it is important to have the SEO listing in the first pages at least if not at the top.



What all is required to achieve the top ranking. There are many tools and techniques to improve the ranking. I will not discuss the detail of the SEO techniques. For this book remember that a company tends to spend money on the manpower and tools required for SEO so it's not entirely free traffic. A company has to spend good deal of technology and manpower on the SEO optimization. Therefore SEO is one of the least expensive traffic and has long term impact. It has nothing to do with your SEM spending. Hence from long term company's channel strategy SEO is definitely one of the most important marketing channels.

5.1.9 Social Media Optimization

Social media is another important digital channel for the organic or non-paid traffic. The company's social strategy is long term investment. The social media may not be great channel for the transaction performance but definitely great tool for the branding. The traffic on the social pages may not translate

into the traffic in the site but the awareness created in the social pages can increase the direct traffic. The most commonly used social media are Facebook, Twitter and Instagram.

From the marketing analytics view point the social can be analyzed in to places

The traffic coming from the social pages and social campaign to site. The analysis of the performance and behavior from this channel s done at website web analytics tools such as Google Analytics. These analyses are similar to all other analysis related to marketing channel.

- 1. The analysis of the users' behaviors in the social pages itself. Those are nothing to do with your website but definitely linked to the company as a whole. The analysis such as sentiment analysis, engagement, amplification and time on the pages.
- 2. From the long term strategy of the marketing channel social is an important channel. It does involved cost in term of manpower to manage the pages, post social campaign and moderation of comments. Hence social traffic is free in term of the variable cost component but the fixed cost remains.

Learning from the Chapter

- Concept of digital marketing, its ecosystem and advantages over traditional marketing
- Definition of the major metrics in digital marketing and their formula
- Google Display, Google Search Engine Marketing, Google Product Listing ads and remarketing
- Facebook ads Multi Product Ads, new feeds, Dynamic product Ads
- Bulk form of advertising email and SMS
- Pay per order types of advertisement affiliate channels
- How to identify the channel and campaign using UTM parameter in the campaign
- What is SEO and why it is important
- What is Social Media and how to measure the social media performances?

Section- II

DIGITAL CHANNEL OPTIMIZATION

5.2.1. Channel Optimization using LP

In the digital world the optimization is the catchword for anyone executing the digital campaign. There are different ways of optimizing the digital campaign like optimizing on the cost per lead, optimizing on revenue per lead, cost per acquisition, cost per new user acquired, optimizing on the CTR, optimizing on CR and so on. The lever for the optimization depends on the objective that the company has from the digital campaign. Each channel may have different optimization objectives. For example affiliate channels can be used for getting new users as much as possible and then use Adwords remarketing for retargeting those users; then objectives of affiliate channels are optimization on the new users and optimization on Adwords remarketing will be revenue per lead or reducing cost per lead. Another optimization is the cross channel optimization for the overall company's digital objectives. Assuming each channel is optimized using all available lever for that channel, then digital strategy is to optimize at the portfolio or the company level how to achieved maximum throughput from the available budget. You can think of channel specific optimization as local minima and overall marketing channel mix optimization as global minima. In this section we will be using Linear Programming to optimize the multiple channels objectives. The objective of company could be maximize the gross merchandize value (revenue) or maximize the profit or minimize the cost; all falls in the category of problem which can be optimized using the Linear Programming.

In last section we have learnt the basics of the digital marketing, the different channels and how to extract the revenue and cost data for each channels. In this section we will be using the same knowledge for the planning and optimization of the digital. Like any other Linear Programming problems the marketing channel mix decision has certain objectives such as revenue or profit using the given resources that is mainly budgeting (cost). There are many constraints in the decisions making because each category has to be given certain fixed budget allocation. The channels have constraints like maximum impression possible or order possible in the given period and so on. Let us start with very basic formulation of linear programming. The objective is to maximize the revenue using marketing four available channels and the budget constraints. The coefficient of the objectives function is the average order value of each channel. The coefficient of first constraint is the cost per order for each channel. Rests of the constraint are the minimum order that is to be generated from each channel. Consider below channel mix

Channels Orders	Cost per Acquisition	GMV per Order	Minimum Order
X1	650	2000	500
X2	600	1400	350
X3	450	1500	300
X4	500	1000	200

The objective is to maximize the GMV of the site using different channel mix.

Objectives Function Max $Z = 2000X_1 + 1400X_2 + 1500X_3 + 1000X_4$

Subject to

 $650X_1 + 600X_2 + 450X_3 + 500X_4 \le 100000$

 $X_1 >= 500$

X₂ >=350

```
X_3 >= 300
```

$$X_4 >= 200$$

The LP is plotted in the Excel with below. Max is the sumproduct of the B3:E3 and B4:E4. Similarly each constraint is sumproduct of the variables are the coefficient.

0 H 5. C							DHU-box	6							1	100	- 0	ж
FLE HOME INSP	FASELA	rout Po	454,1,45	OATA AD	VER VER	é											Spin	Р
Farme Frankfranker		- (n -)	() = ▲• ≡		Direct Direct	all A.Centar =	Ceneral 5 + % +	* ** 8	R	2	No.	1 200 C	10	EAddam Rfd- Chart	ST .	A		
Optional 4	54	4		40	-		Burline .			Trypes		-049			- 10			+
4 4		6							1	. 4		м		0		. 9		
Variables	81	12	10	X4														
1	0	0	0	0														
< Coefficient	2000	1400	1500	1000	Max	-												
Objectives function	95																	
r Constraint 1	650	600	450	500	0	£+.	1000000											
a Constraint 2	1				0	34	500											
Constraint 3		1	5 - C		0	34	350											
io Constraint 4			- 3		0	34	300											11
II Constraint S				1	0	ha	200											11
12		-			-		-											11
10																		11
																		1.1

In solver we added G4 as the max objectives and added all constraint.

a stop high have	Other Bush	14 A.A.		unter SI	1 m 2 m 3 m	harved of	dans fan		Solver Parameters				
64 .	- 5	-	Contras		101.1.100	_		Sej Objectives	100			8	-
A		¢	0		1 6	- 11	1		No Option	5			3
	La La			-				B. Channel Sameth Cale					
a-apres	PA P							-				100	
selficient	2000	34000	1500	1000	Man C							-	
Buschwes Functions								Syland is the Constraints:		-			
	1000		10.00	10.00		1.000		Paul in Paul		- C	84		
protraint 1	450	000	400	500	4	1000000		PET CA PET			Owner		
professional 2	1	-	-	_	60-4	500		P31 1= 210					
protrained 8			-			230		CONCERN.			Dev		
propriet a	+	-+											
		-	-	-	-					1.15	Erest all		
											Louifere		
								R his contained to a	dies fair ringelt a		-		
								Sport a Solvey Wether	(h) wher		Options		
								Sanag Reflect			in the second	2.1	
								ergine for locar Julyer And	mmi, and select the E-sk-lierary er	gra far Jalver	problems that are		
								re-mark.					
								pe		See .	Oge		

On solving we can see the maximum revenue of the 29 lakhs. The budget is fully consumed and channel 3 being most efficient has been allocated maximum orders whereas other channels are just left with their minimum values.

A fund fund for the fund for th	a Other Bad aron - Gales a Data	12 C	Canada	ppillus) i (podes ang	1 54 1 54 10	New Syn		and the Asso	Parties 10	Data Car Nation - Data Torn	antere a	19 1917 - 6 1911 -	the particular	Bill Subtrace Culture	-1 man tra	5.00	*****	
		c	0			. 6		1	1	×		w	N	0		9		- 3
(acables	ha h	a 1	a 1															
	500	350	823.3157	300								4	A and Raw in					
outfloant .	2005	54005	13000	2000	Mar	2904444.7							over nesus			_		
Reactives function										Salver few	nd a service	en All Cent		utrains in				
entranet 1	400	800	400	300	3000000		10000000			1	10.5			10	krainer .			
seutraint 2					500	-	500			0.041	14-# 56.M				and to be			
anatranit 3		1			350	-	100			Ogen	e Drighel he	6.00			100			
onstraint 4			. 2		403.033	-	100											
andrand 3					.2015	ha	200			Deput	to Selver Pr	anameters 2	and a	0	Option Inc	***		
										-	-					-		
											_	(pencel			gave Scen	· .		
										page and	10.7 million		and a second	2000		-		
										Solar fou	nd a solution	n, All Cand	valets and ap	timelity 10	ruffilans are			
										What the	and ergine	in some to		dist land				
										services, t	then bings	er Wild with	4.9-3	Server he	found + gro	ter.		
										aprile a	iveau.							

Let us change the objectives function a bit. We have now objective of achieving a revenue target with minimum cost.

Objectives function min $Z = 650X_1 + 600X_2 + 450X_3 + 500X_4$

Subject to

 $2000X_1 + 1400X_2 + 1500X_3 + 1000X_4 >= 5000000$ $X_1 >= 500$ $X_2 >= 350$ $X_3 >= 300$ $X_4 >= 200$

HOME NORT	FACELANDS	. PDM	1.1.45 0	ATA ADV	IN VE												Sep
X Ce Bh Cey - Ce Frank Farmer	n () 7 ¥ - ⊡ -	0 - A		0 +C +C	Dawy Dawy	a Center -	(men) 5 - % +	N.A	E	-	283	The State	E.	X Autolaum K Autolaum Come -	- AT Sat	A	
A	1	. 6		1		4		1				44	N	0	7	0	
Variables	xt	x2	x3	x4													
	0	0	0	0													
oefficient	650	600	450	500	Min	0											_
objectives function	705	1400	1500	1000	0		5000000										
Constraint 2	1	1400	1.000	1000	0	31	500										
constraint 3	- 1	1			0	31	350										
constraint 4			1		0	31	300										
Constraint 5				1	0	>-	200										
	14	-	1		-	-	-										

The formulation is shown above

	r)# et Poprios	nd Par	uni (1	to term		lawinger -	DM UP + M	ficrosof	t Decei						0-01
	a Other an	1		anation 3 gentus tonto	100 Z	The Const The Property Sectores		E.	18 12	59	Solver Paramet	()) *] :::=: (es	total & betw		
Cel Deuro	- 1		Carlotan		541.0	Adul		-	lag Objection					N	-
4 A		c	0	1		a .					Others				1 5
1	1000										C party				1
2 Variables	141	k2 .	0	6.6					B-Owgrep terrate C	ala:				-	
1									80.80					100	
4 Coefficient	410	900	400	500	Mat				Advent in the Constan						
5 Objectives function									P10 1+ P10			-	4.44		
									Para an Para				-		
Contract 1	200	3460	1900	2000			1000		W100 1+ 2100				Que		
Conversion 2		-			-				1 (1) = 1 (1)						
Constraint &	-	- 1							0.000000000				Dena		
L1 Canadraped 5		-	- 1	1.1	0.00	-	200								
1	-	_			-	_						1.1	Growt all		
3													Louiser		
H.									Projet Suppliers	(Tarialitas Inc.	(market				
8															
6									spect story were		12.00.048		Optore		
0									Salarg/Rethod						
28 29 20									Select the GHS have anyone for these Salar represents	an argue for 1	lation Phablems Rail a Cathol: The Column	r, anges for John	Select the UP Sing problems that are	-	
22									Contraction of the				1.0		
28									99		-	- pre	olie		
16															
8															
15															

The constraint has been changed

Anna Anna Anna Anna Anna Anna Anna Anna	ill Bart an Carro bra		Si gran	roman 1	14 (2)(2) (1 (5+4) (5-4)	100 30 100	manine d	ted to 1	anates y	Data Car andatus - Data Team	unitaria in An	65 1417 -	EU C	Cutton	and a second	5.	la kraljski urtyra	
		¢	0			. 6		1	1.1	- 15	4	M		¢		0		5
anatres	10 b	9	2206.447	300								5	o'ver Res	-				
bellicent Rjøctives functions	450	900	400	500	Ma	2429000				Salver faur eand tions	dia selutio are set of a	n All Car		d antimat	Nagara Angara			
propriet 1	2000	3400	1300	3000	3000000		Scotting			@ peo t	ine Salar	1			Selectivity			
and part 1		-			110	-	210			Ogene	- Departed	-						
onstraint 4			3		2256.667		300			Denne	a later Pa		Dalas		The starting law			
E Insettant				28	300	ha l	200								C. open and	-		
										94		çanori			jave form	#'8.		
										Salver faur until field Unter the I services To services To	od a solution CAS angine Chan Simple Nation.	is used, 5 is UP is up	invalence and investment for ed. this mass		en a local agen tas local agen	14		

The revenue target of 50 lakhs can be achieved using Rs. 16,28,000 as the digital marketing budget.

In the above two example we assume that each channel work at the same revenue per order and the same cost per order for all the products in the site. We can now expand the complexity of the problem by including the product category in the problem. The performances of the category channel-wise would be different based on the customers profile in the channels. For the sake of simplicity let's assume that the site has five categories and four marketing channels. We can prepare the chart as shown below with each category average bill value and the cost per orders for the channels.

	Chann Value	el wise .	Bill	Chan	nel wise	cost per	Order	
Category	C1	C2	C3	C4	C1	C2	C3	C4
Books	500	350	370	400	200	120	150	230
Mobiles	12000	8000	7500	6000	650	450	700	500
Fashion	1200	950	1000	800	240	250	300	260
Furniture	6000	7500	6500	9000	850	700	600	1000
Kitchen	560	600	750	800	150	200	140	250

As a company we have to grow each of the categories hence efficiency alone would not guide the allocation of budget. The category will have some minimum budget to achieve the presence it required.

Category	Minimum Order
Books	5000
Mobiles	8000
Fashion	10000
Furniture	6000
Kitchen	2500

As each channel has constraints on the number of orders it can brings we can put constraint on minimum and maximum based on the historical data. The minimum is assumed because the company spends time and money on the channel association hence would like to maintain the presence. The channel can be discontinuing if it underperform. Channel 1 has not maximum constraint as it can expand as the budget is increased.

Channel	Minimum Order	Maximum Order
C1	1000	No Constraint
C2	500	15000
C3	250	12000

C4	250	10000

Now let us formulate the above problem in the Linear Programming Problem. The objective function is assumed to be maximizing the revenue.

Objective Function Max $Z = 500X_{11} + 350X_{12} + 370X_{13} + 400X_{14}$ $+ 12000X_{21} + 8000X_{22} + 750X_{23} + 6000X_{24}$ $+ 1200X_{31} + 950X_{32} + 1000X_{33} + 800X_{34}$ $+ 6000X_{41} + 7500X_{42} + 6500X_{43} + 9000X_{44}$ $+ 560X_{51} + 600X_{52} + 750X_{53} + 800X_{54}$ Subject to $X_{11} + X_{12} + X_{13} + X_{14} >= 5000$ (Book constraints) $X_{21} + X_{22} + X_{23} + X_{24} >= 8000$ (Mobile constraints) $X_{31} + X_{32} + X_{33} + X_{34} >= 10000$ (Fashion constraints) $X_{41} + X_{42} + X_{43} + X_{44} >= 6000$ (Furniture constraints) $X_{51} + X_{52} + X_{53} + X_{54} >= 2500$ (Kitchen constraints) $X_{11}+X_{21}+X_{31}+X_{41}+X_{51} \ge 1000$ (Minimum channel1 constraint) $X_{12}+X_{22}+X_{32}+X_{42}+X_{52} \ge 500$ (Minimum channel2 constraint) $X_{13}+X_{23}+X_{33}+X_{43}+X_{53} \ge 250$ (Minimum channel3 constraint) $X_{14}+X_{24}+X_{34}+X_{44}+X_{54} >= 250$ (Minimum channel4 constraint) $X_{12}+X_{22}+X_{32}+X_{42}+X_{52} \le 15000$ (Maximum channel2 constraint) $X_{13}+X_{23}+X_{33}+X_{43}+X_{53} \le 12000$ (Maximum channel3 constraint) $X_{14}+X_{24}+X_{34}+X_{44}+X_{54} \le 10000$ (Maximum channel4 constraint) $200X_{11} + 120X_{12} + 150X_{13} + 230X_{14}$ $+650X_{21} + 450X_{22} + 700X_{23} + 500X_{24}$ $+240X_{31} + 250X_{32} + 300X_{33} + 260X_{34}$ $+850X_{41} + 700X_{42} + 600X_{43} + 1000X_{44}$ $+ 140X_{51} + 200X_{52} + 150X_{53} + 250X_{54} \le 50000000$ $X_{ii} \ge 0$ (Greater than 0 constraints)

In excel you can formulate same LP as shown below. You can have different format but keep all the constraints.



Add all constraints and the objectives functions as shown above.

	April 1948	Peprint	Aurosta (Data Review	-	terrige W. N. C.	446.04			+00	+00	en els	au faiter -	2. total	• 0	- 4
in in	Anna Anna	B UB	(G) a	Tangantan SI	348 N		many fail	Annos Data Ca	tentere atura	1. desa 0	All 1		in Servi	1300 m	60	
	Get Devenant	infra	Carte	and and	Sec.	Aller		Data faish	- and the		0.0	Direl .		Analysis		
M		- A	-			-		and the second second	111							
4 A		c	0	1 1			1.4.1.1.1	J	1 L- 1		N	0		Q.		
		Objective Coefficient		Category												
1	Variables	(AOV)	CHO	Canabrainte	-			Max	4492013223.3	_						
2 101		500	200	Beek	5000	14	5990	Budget Constrain	2 30000000	ia I	5000000					
1.112	200	100	122	Number and	44225.04	24	. #200									
4 313		1.75	224	Tashiph	20000	H	20000									
5 924		40	230	10000	0000	-	4000		50	iver Result	10		100			
6 325	44223.4768	12000	600	CONTRACT.	2000	H	2500									
7 322		4000	400	-	-			Selver found an	Internation all Cons	two and a	uting ity					
1 120	-	798	700	Channel Cor	Traces of		1	conditions are	setsfel.			age to				
9 934		4000	500	0	79473-04	14	1000	Renter	to an			Artiste'				
10 A34	175	120	- 243	-	5000							100				
1 111	+	1 100		<u>-</u>	1900	-	- 25	Ogenerativ	profit factors							
1 114	+			1	100		11000									
LA MAIL		400		1	8500	-	10000	C Apura ta Se	Let Parameters 2	aing	5	Outine Rep				
13 842	-	700	700	Ci I	210		10000		-							
14 842	4000	4300	400	-	-	-			generi			jest Sort	a/a.			
17 144		\$000	1000						and the second second							
18 354		540	150					Salvary Inund a 1	ub, tiph, All Canal	raids and ap	planality o	and/iana and				
18 x32		400	200					When the life	status is used for	and an income		-				
10 X54	250	79	140					selution When	Singles IP Is use	A this means	a Balver ta	a found a giv	iter .			
12 854		800	290					April March Service	A.							
12	-															
34																
	No. Sheet2	No. Shorth	0					0	×							

The output of the solver is shown above. The allocation of the order each channel for each category meeting all criteria is the number in the variables. We can multiple the costs per order and variable to get the budget per category per channels. The allocation of budget for the category and the channel can be calculated by corresponding order numbers and cost per order. Similarly we can calculate the revenue for each channel and each category.

In above example we optimized on the revenue from the given cost (budget). If the objective of the company is to maximize profitability rather than revenue then we have to change the objective function with margin maximization. Assuming same categories with revenue per order and cost per order; we know the margin % of each categories and each category has same margin structure across the SKUs.

		Chann Value	el wise	Average	Char Orde	Channel wise cost per Order			
Category	Mar- gin%	C1	C2	C3	C4	C1	C2	C3	C4
Books	20%	500	350	370	400	200	120	150	230
Mobiles	5%	12000	8000	7500	6000	650	450	700	500
Fashion	35%	1200	950	1000	800	240	250	300	260
Furniture	25%	6000	7500	6500	9000	850	700	600	1000
Kitchen	15%	560	600	750	800	150	200	140	250

Objective Function Max Z = $20\%*500X_{11} + 20\%*350X_{12} + 20\%*370X_{13} + 20\%*400X_{14}$

 $+ 5\%*12000X_{21} + 5\%*8000X_{22} + 5\%*750X_{23} + 5\%*6000X_{24}$

 $+ 35\%*1200X_{31} + 35\%*950X_{32} + 35\%*1000X_{33} + 35\%*800X_{34} \\$

 $+\ 25\%*6000X_{41}+25\%*7500X_{42}+25\%*6500X_{43}+25\%*9000X_{44}$

 $+ 15\%*560X_{51} + 15\%*600X_{52} + 15\%*750X_{53} + 15\%*800X_{54}$

Subject to

$$\begin{split} &X_{11} + X_{12} + X_{13} + X_{14} \ge 5000 (\text{Book constraints}) \\ &X_{21} + X_{22} + X_{23} + X_{24} \ge 8000 (\text{Mobile constraints}) \\ &X_{31} + X_{32} + X_{33} + X_{34} \ge 10000 (\text{Fashion constraints}) \\ &X_{41} + X_{42} + X_{43} + X_{44} \ge 6000 (\text{Furniture constraints}) \\ &X_{51} + X_{52} + X_{53} + X_{54} \ge 2500 (\text{Kitchen constraints}) \\ &X_{11} + X_{21} + X_{31} + X_{41} + X_{51} \ge 1000 (\text{Minimum channel1 constraint}) \\ &X_{12} + X_{22} + X_{32} + X_{42} + X_{52} \ge 500 (\text{Minimum channel2 constraint}) \\ &X_{13} + X_{23} + X_{33} + X_{43} + X_{53} \ge 250 (\text{Minimum channel3 constraint}) \\ &X_{14} + X_{24} + X_{34} + X_{44} + X_{54} \ge 250 (\text{Minimum channel4 constraint}) \\ &X_{12} + X_{22} + X_{32} + X_{42} + X_{53} \le 250 (\text{Minimum channel4 constraint}) \\ &X_{12} + X_{22} + X_{32} + X_{42} + X_{53} \le 15000 (\text{Maximum channel2 constraint}) \\ &X_{13} + X_{23} + X_{33} + X_{43} + X_{53} \le 12000 (\text{Maximum channel3 constraint}) \end{split}$$

X₁₄+X₂₄+X₃₄+X₄₄+X₅₄ <=10000(Maximum channel4 constraint)

$$200X_{11} + 120X_{12} + 150X_{13} + 230X_{14} + 650X_{21} + 450X_{22} + 700X_{23} + 500X_{24} + 240X_{31} + 250X_{32} + 300X_{33} + 260X_{34} + 850X_{41} + 700X_{42} + 600X_{43} + 1000X_{44} + 140X_{51} + 200X_{52} + 150X_{53} + 250X_{54} <= 50000000$$

 $X_{ij} \ge 0$ (Greater than 0 constraints)

The objective function has changes from sumproduct of the order and AOV to sumproduct of the order and margin as shown in Excel sheet below.

1.1	HOME INS	RT FACELANDUS	PDRMUA	6 0414 A	-	DATA.	but					,	8 - 1 be
Xaxa	Cat Caty - Furnat Parties	Caller -]N ● 7 ¥ - □ - Full	() ≤ K K () ≤ K K			rhCenter + S + No N Center + S + No	• 12	102	E	2.2		Ander - CT M	
	Variables	Average order per Value	MarginN	Margin per Order	00	Category Constraints	, I	1		Margin	Max	0	9
411	0	500	30%	10	200	Book	0	34	5000	10%	Budget Constraint	0 4- 50	000000
X12	0	350	30%	35	120	Mobiles	0	34	8000	5N			
X13	0	370	30%	37	150	Fashion	0	34	10000	35N			
x14	0	400	30%	40	230	Furniture	0	34	6000	25%			
X21	0	12000	55	600	650	Kitchen	0	34	2500	15N			
322	0	8000	5%	400	450	- C				(m. 19)			
323	0	2500	5%	375	200	Channel Constr	raints						
324	0	6000	5%	300	500	C1	0	30	1000				
X31	0	1200	35%	420	240	Q	0	34	500				
X32	0	950	35%	332.5	250	C3	0	34	250				
X33	0	\$000	35%	350	300	C4	0	. 34	250				
×34	0	800	35%	280	260	02	0	٠.	15000			_	
x41	0	6000	25%	1500	850	0	0	4.	12000				
3.42	0	7500	25%	1875	200	C4	0	. e+	10000				
X43	0	6500	25%	1625	600								
3.44	0	9000	25%	2250	1000								
254	0	560	15%		150								
225	0	600	25%	90	200								
723	0	750	15%	112.5	140								

The constraint will remain same across the category



3 H S	HOME IN	SHT PAGELANDA	. PORM	14.45 0474	ADVER	VEW	040-1	ind.					7 10 - 8 bein
Xaxa		Caller	4 - X 4 - A		6- 5 4 4 5	They fait Norge & Conter -	5 - %					E States - States	A
1	Variables	Average order per Value	Margin	Margin per Order	00	Category Constraints				Margin	Max	103099705.9	
311	5000	500	10%	50	200	Book	5000	34	5000	30%	Budget Constraint	50000000 <-	50000000
X12	0	850	10%	35	120	Mobiles	8000	34	8000	5%			
X13	0	370	10%	37	150	Fashion	\$0000	34	10000	35%			
334	0	400	10%	-40	230	Furniture	52676	34	6000	25%			
X21	8000	12000	5%	600	650	Glichen	2500	34	2500	35%			
322	0	8000	5%	400	450					1			
323	0	7500	5%	375	200	Channel Con	atraints						
324	0	6000	5%	300	-500	<1	41176	34	1000				
X31	10000	1200	35%	420	240	02	15000	.94	500				
X32	0	\$50	35%	332.5	250	C)	12000	.94	250				
X33	0	1000	35%	350	300	<4	\$0000	74	250				
×34	.0	800	35N	280	260	0	15000	61	15000				
×41	15676	6000	25%	1500	850	0	12000	61	12000				
842	15000	7500	25%	1875	700	64	10000	61	10000	1			
X43	1,2000	6500	25%	16,75	600								
x44	10000	9000	25%	2250	1000								
X54	2500	540	15%	84	150								
X52	0	600	15%	. 90	200								_
123	0	750	15%	112.5	140								- C
354	0	800	15%	120	250								1

From both output we can see the different number of orders coming from different objective function. This is because some of the category-channel mixes were good for generating revenue but poor in generating margin. The number of order from GMV optimization is 89723 and margin maximization is 78176 which show how the variable changes with change in objective function.

	Using AOV	Using Margin
X11	0	5000
X12	5000	0
X13	0	0
X14	0	0
X21	66223	8000
X22	0	0
X23	0	0
X24	0	0
X31	9750	10000
X32	0	0
X33	0	0
X34	250	0
X41	0	15676
X42	0	15000
X43	6000	12000

X44	0	10000
X51	0	2500
X52	0	0
X53	2500	0
X54	0	0

By now the readers must have understand the mechanics of the linear programming and how it can be used to solve an optimization problem. LP is powerful tool for planning allocation of resources when we have multiple objectives and the resources constraints. In this chapter my attempt is to use LP for planning marketing budget allocation. Since situation is dynamic in the real environment the coefficient and the constraint can be changed at any moment and rerun the LP to get new allocation. The LP can be run every day to decide the budget allocation for the next day and also finding the efficiency of the allocation of the past data.

We have used linear programming assuming only digital channel but for long run planning we can include SEO, social, ATL and BTL channel for the optimization. The measurement of ATL and BTL channels might not be very easy but assumption can be made for each activities to arrive at reasonable return on the spend.

5.2.2 Attribution Modeling

Attribution modeling is the method of assigning the value to the marketing channel based on the participation of the channels to the journey of the users which lead to finishing of a goal(s). For the simplicity we will assume that the site's goal is the transaction amount that a customer done through the site.



To understand the attribution modeling let us consider the journey of above three users A, B and C which completed the goals of the site. The customer A purchase 2000 worth of goods from the site, customer B purchase 5000 worth of good and customer c purchase 1000 worth of goods from the site.

- Customer A came through Search Engine Marketing and purchased on the same session.
- Customer B came through email in the first instance, purchase nothing on that session. In the second session customer B came through direct channel and purchased.
- Customer C came to site through SMS on the first instance and purchases nothing. On the second instance C came through FB but again purchase nothing. On the third instance C came through SEM and purchased.

The purchase done by A can be attributed to SEM directly but what about B and C which had multiple channels as touch point. How much value should be assigned to the different channels and why? The attribution modeling is all about answering these types of questions.

There are different methods of attribution available in the text. Each web analytics tools have some models or the other. In this book we will focus on Google Analytics only. In Google Analytics there are 7 attribution modeling.

- 1. Last Interaction: Assign all value to the last channel leading to the goal completion
- 2. **First Interaction**: Assign all value to the first channel of interaction of the customer which led to completion the transaction.

- 3. Linear: In linear all channels being interacted on the way to conversion are given equal weightage.
- 4. **Time Decay**: The most recent channels are given more weightage and second most recent channel is given lower than recent channel and so on. User can define decay function
- 5. **Position Based**: in this method you can assign certain percentage to first and last interaction and rest is assign linearly by the system to all channels between first and last interaction.
- 6. Last non-direct clicks: In this method the last channels which is not direct is given 100% credit.
- 7. Last AdWords Click: the last AdWords click—in this case, the first and only click to the Paid Search channel —would receive 100% of the credit for the sale.

Using above example of customer A, B & C lets us try to attribute the sales in three cases for different models

Customer A: for customer A the only channel was SEM so all attribution gives SEM as the only channel with 100% credit for sales

Models	SEM	FB	EMAIL	SMS	DIRECT
Last Interaction	2000				
First Interaction	2000				
Linear	2000				
Time Decay	2000				
Position Based	2000				
Last non-direct clicks	2000				
Last Ad words Clicks	2000				

Customer B: For the position based the credit is 40%, 20% and 40% for first, middle and last interaction respectively. The time decay period is say 7 days half-life of decay.

Г

Models	SEM	FB	EMAIL	SMS	DIRECT
Last Interaction					5000
First Interaction			5000		
Linear			2500		2500
Time Decay			2500		2500
Position Based			2500		2500
Last non-direct clicks			5000		
Last Ad words Clicks	NA				

Customer C: For the position based the credit is 40%, 20% and 40% for first, middle and last interaction respectively. The time decay period is say 7 days half-life of decay.

Models	SEM	FB	EMAIL	SMS	DIRECT
Last Interaction	1000				
First Interaction				1000	
Linear	333	333		333	
Time Decay	500	250		250	
Position Based	400	200		400	
Last non-direct clicks	1000				
Last Ad words Clicks	1000				

I am sure the above attribution makes sense to you. If not practice it again and again till you gets it.

In Google Analytics the attribution comparisons are done in Conversion -> Attribution -> Model comparisons Tool.

	weights X O server branches 27 X O track where the	N	posite subur shown	X	out - X O week the	eroes. X. e
• 0 a teperiodyte gauges	www.codytechede.700.ardipte/14-spect/64-ser-scienterers/200545646207506/2546207					e 🖉 1
askmebazaar.com	•				6*	II I 🕲
HOME REPORTING	CUSTOMIZATION ADMIN					
S. Depth-sports & hep	* Model Comparison Tool				Jul 17, 2016 - Au	g 16, 2016 -
10000	Conversion Septemb Export - Shortcat					•
Succession Uses File Acquister Behave	At All Matteria Bel (22) days prior to cohereses In of cohereses Vid cohereses Vid cohereses Interface	a Sues 10		er brugege -		
Comercions	Security disease +				C	A abanced
+ Grada	MCF Channel Graging 3	Apand anisted to a mouth	Lastaneuchan Parameters () 4	Gast Interaction CPA	Last Mariathian Concerning Value 12	Last Interaction Mint ()
+ Economice				-	54,338,792,00 an array	-
+ Economics + Multi-Channel Funnalis	1. Deut	-	Without a line of			
+ Economica + Multi-Channel Pannels + Attribution	1. Dent. 2. Organi franth	-	5,548.00 (0.474	-	\$1,709,537.00 (px.) mp	-

You can select two models and compare using the comparison tool.

CL Search shall	Last interaction - ve	Serverman	1			
+ Cestan	A Annaly Dramater. MCF Channel Grouping . Debut Dramat Brough	1	9	al Breamps -		
+ Derchmarking	Secondary dimension +	- Default Models				Q, abanced
Uses Floe	NCF Channel Grouping	Last Interaction	0	Last teneraction CAs	Cast Interaction Conversion Value	Cast Interaction BOAS
destantes :	A Deet	Last Advised City	- 1	-	\$4,338,792.00 micros	-
E Belaisr	2 Organic Search	XIII -		-	\$1,709,537.00 (school)	-
Conversions	3 (094)	L. Pre margina	× .	-	1798.232.08 1794	
+ Onath	4. Rebral	Linear	0		E05.177.00 (2.47%)	
+ Econometer	5 Social Televolt	Time Decay	0	-	11.018.00 million	
+ Multi-Channel Funnels	6 Other Advecturing	Foston Based	0	-	\$2.0% DI	
Abribution	7. Part Seath	- Custon Models		-	\$3,081.00 (CHA)	
Wodel Comparia	8 Enal	Charle free during model		-	1215.00 c.m.	-

The setting of the attribution model is done in Admin -> Attribution Models

 Sec. Sector Statements		
 Postaria Postaria Postaria Autoria Debataria 	Did Calculated Damos and HEDDINAL TOOL & A STOTTA IID Segments CEI Annotatione	
Die Des Ingen	Addition Makes Contain Charant Groupings Sets	
	Scheduled Emails Scheduled Emails	
	Z* Dian Assets	

You can build your custom attribution models by changing the look back windows, % attribution in position based and half-life time.

NEW .			
admebacaan.com •	Hodel Name		
1 Yes Setters	Baseline Wodel		
the Use Manapamant	Linear		
Mr. Gun	Q		
A Content Drouging	i and i a		
₹ fdan	I for spectra		
2 Dave Server	L	left matrics from the working esit.	
H Economic Integs	Time Decay	iay your model is defined:	
Did. Calculated Metrics agrin.	Poston Based		
Manfalmi Toriy S.S. accords	had states		

In time decays model you can change the time period of half-life of the decay

State Union Management Time Decay • Mic Grass Bat habitish of decays (?) days • A canadard Grouping An assist click socienting this many days prior to conversion will receive 1/2 the credit of a click on the day of conversion. • Thins: Content Grouping An assist click socienting this many days prior to conversion will receive 1/2 the credit of a click on the day of conversion. Thins: Content Statings Content Statings Conversion Statings Adjust credit based and uses for attribution. Conversion Statings Adjust credit based on time supportment memory ham the cut ham the click societ on the stating and, click	
Mar Bet haldship of decays () deps K Centers Grouping Mar An asset click second pills many days prior to conversion will receive 1/2 the credit of a click on the day of conversion. T Frient Construct Entropie Specify the transit back and used for attribution. Construct Entropie Adaptat Credit back days of one used for attribution. T Construct Entropie Adaptat Credit back days of one used for attribution. Adaptat Credit back days of one used for attribution. T Construct Units and your click back and used for attribution. Adaptat Credit back and used for a tribution. Del Construct Units angles and use the analysis transit one time angles ment memory. How the resulting any. Del Construct Units angles Adaptat credit back	
Consert Grouping An assist click sectoring this many days prior to convenient will receive 1/2 the credit of a click on the day of convenient. Ty Fritary Consert Grouping Or Control Eletropic Specify the tree back for other and by articlation. Ty Consert Eletropic Adjust Credit hand on your exappament. Adjust Credit hand on your exappament. Adjust Credit hand on your exappament. Adjust Credit hand on your exappament. Adjust credit need on your exappament. Od Construct Writers 1974 Adjust credit need on your exappament.	64
T / Fing Leadback Window Channel Settings Specify the text back panel used for attribution. The Construct Settings Adjust credit based on your engagement Apply multiple weighting wise based on site ungrapment metrics from the resulting opt. Del Catulated Winnes 1974 Adjust credit have Adjust credit normality Adjust credit have of an user engagement	04
Channel Settings Specify the fee back period used for attribution. Adjust credit based on year angagement Accounters Settings Adjust credit based on year angagement Accounters Settings Adjust credit based on sets engagement memory has been generated with the resulting equ. Adjust credit based Adjust credit based on sets engagement memory has been generated with the resulting equ. Adjust credit based Adjust credit based	
Adjust Credit based on user angagement Apply multiple weighting takes based on ante angagement metrics have the resulting and. Def. Calculated Unices as tw.	
Dd. Cambrid Metrica agrae Apply customs could rules	04
	07
PENDANI, TOKAR & ASSISTS	C.C. Mail
R) teprets tee Carol	

In position based model you can change the % attribution of the first, last and other channels. The first and last is assign and then remaining channels are linearly distributed. In below model 40% is assign to first and 40% to last and remaining 20% to the remaining channels.

E Van Detrop	Exceline Model	
un Unr Management	Lus Postion Based -	
M dam	Specify the amount of conversion credit based on the position.	
X Contact Grouping	Prot interaction: 40 % Models interaction: 20 % The only or anticipation of the second	
T flas	Last elevation 40 %	
[5] Oserel Setings	Tenal 100 % mercernen	
H Economics Suttings	The second se	122103
D.E. Calculated Metrics and the	Looklash Wedge Specify the Yook back period used for attribution	
PERSONAL TIMES & ADDRESS	Adjust (meet based as user angegement	· · · · · · · · · · · · · · · · · · ·
30 Septem	which we have a set of the	
(E) American	Apply custom could rules	0.0

You can change the look back windows. The system default is 90 days. You can also adjust based on user engagement – time on the site and page depth while crediting the sales to the channels.

 Bell Calculated Metrics lettings 	Lookback Window Set 10 Rays play between	-
PERSONAL TOOLS & ASSETS	a	
50 Septents	40 35 46 46 38 15 6.6vy(s)	
(B) Annextons	Adjust credit based on user angagement	Cu
da Antibulus Madeh	Datribute credit proportionally based on are an e-	
Custors Diamed Groupings	Apply certains conditioned index	(ee
🕫 Custon Alets		
Scheduled Erreits	tee Carol	
D Ehercuts		
2.4 Day Aug		

You can also define custom credit rules.

🛃 Anthony Balata	Debibule credit poportionally based an investment *	
Custon Diareel Groupege	Apply cestan credit rules	
🕫 Curran Riana		
S Scheduled Emails	Include: Produce in Park - Exactly matching any 😺	
D Illustouts	· · · · · ·	
L* Share Access	Add With element	
	and	
	Add SND statement	
	sum () C met met met met met met met met	
	+ Apply another system smill	

Why attribution model is important and deserve such a long and detailed conversation.

- 1. When you have to pay vendor based on the sales done through their channel the look back windows and attribution model is important. This is especially import for non-click based cost.
- 2. Each interaction is important in the customer's buying decision making point of view. For high value items customer tend to make several visits before final purchase is done. In such scenario it is really important that other channels are also given credit for the sales.
- 3. Every interaction with customer adds to the customers brand recall of the site. Hence each channel deserves some credit for the final purchase.
- 4. The cost of customer acquisition is not just the final channels spend but the spend made in each interaction in the journey to the purchase.
- 5. The channel decision may changes with different attribution. Hence it is important to use different model so that important channels which contributes to the purchase are not left out event if the last click attribution does not assign any value to that channel.

There are several other reasons why attribution is important. Irrespective of the attribution model being used, the channel optimization using linear programming will remain same but will provide different channel mix.

Learning from the Chapter

- How to optimize the Digital channel mix using linear programming
- LP formulation under different objectives and constraints
- What is attribution modeling
- Why attribution is important

• Different attribution modeling in Google Analytics and its example

Chapter - VI

FORECASTING AND PREDICTION

"It is far better to foresee even without certainty than not to foresee at all."

-By Henri Poincare

In 1999 a strong cyclone classified as Class V type cycle hit Orissa coastline at the highest wind speed of around 260 km per hours. Around 10,000 people were dead from the direct storm hit. In 2013 a cyclone name Phailin with similar intensity hit same Orissa coast, but this time the number of fatalities were just 45. So what went right in 2013 preparation with respect to 1999, which could help reduce fatalities from 10,000 to 45? In the decade time India government invested heavily on the radar and other technologies for forecasting the weather with greater accuracy. The government officials were able to forecast the exact time and location of landfall along with the intensity of the likely wind speed with high precision. Apart from better forecasting, the disaster management teams were trained to handle situation better after their experience in 1999 and years after. Looking at this example we can guess how important it is to forecast weather phenomenon especially for country like India where millions of people are dependent on monsoon. The business of prediction and forecasting is equally important in the corporate world where fortune is made or lost from numbers being forecasted.



In simple term the forecasting is a process of predicting the future value of the variable based on the historical data and its trends. The forecasting is very common activity in our daily work such as what would be sale of any item tomorrow if were to discount the item by 5%, when a Product will go out of stock using the current rate of sales, what would be sales of a category next quarter and so on.

Most of the forecasting work done on the daily basis would be simple and more of directional rather than the value have very significant meaning; the calculations are done on the fly in spreadsheet using mean and trends lines. However if the future value of the variable is of great importance the higher accuracy of the prediction is required with the confidence interval so that the risk from the wrong prediction can be minimized. The casual way of forecasting can lead to many wrong decisions; hence it is important for the forecaster to know how to use a particular forecasting method in a situation and how to interpret the numbers.

In cases where the data is not available the forecasting is done on the basis of the qualitative understanding of the experts. Such method is called qualitative method. Delphi method is one of the prominent qualitative methods. In this book we will not discuss qualitative method of forecasting. We will focus on the data driven forecasting only.

The prediction of any events or likelihood of events is important decision making process of the companies. In forecasting we generally take time into account whereas there are many areas time is not important but the probability of events or the outcome of the events is important. The focus of this book will be to provide users a handle on the popular forecasting and prediction techniques that are being used in the business environment. The first section of the chapter deals with the prediction using variables called independent variables like regression, logistics regression and generalized linear model. The second section of the chapter deals with pure forecasting using time series methods such as moving average, exponential smoothening and ARIMA. Nowadays machine learning is increasingly used for the prediction of the variables value.

Section - I

REGRESSION

In many area of statistical analysis there are requirement to establish relationship between variables in such a way that knowing the value of one variable help establish the second variable. One common example would be establishing relationship between rainfall and moisture content of the atmosphere. Establishing relationship between price and the sales quantity is another example. In simple term Regression is method of establishing relationship between the independent variable and the dependent variable. In above example the independent variable will be price and the dependent variable will be sales quantity. The regression enables us to predict the average value of the dependent variable given a set of independent variables. Regression can be broadly divided into following broad type:-

Linear Regression: The linear regression is the simplest of the regressions. IN linear regression the dependent variable Y is predicted using independent variables X(s) using linear predictor function. Linear regression is mostly fitted with least square method. In simple term linear regression fitted line is straight in the form $Y = a + b_1x_1 + b_2x_2 + ... + b_nx_n$

There are two types of linear Regression:

- 1. **Simple Regression**: In simple regression the number of independent variable is only one. The one independent variable is used to predict the average value of the dependent variables.
- 2. **Multiple-Regression**: In multiple regressions the prediction of the average value of dependent variable is done using more than one independent variable.

Logistics Regression: We have learnt about logistics regression in chapter IV. In logistics regression the dependent variable are binary (yes/no, success/failure).

 $Y = \exp(a + b_1 x_1 + b_2 x_2 + . + b_n x_n) / (1 + \exp(a + b_1 x_1 + b_2 x_2 + . + b_n x_n))$

Polynomial Regression: If power of any of the dependent variable is more than 1 then regression becomes polynomial regression

E.g Y= $1 + b_1 x_1 + c x^2$

There are infact many other types of regression but in our daily usage we will be most likely using Linear and Logistics regression; so will learn about linear regression in this section.

6.1.1 Simple Linear Regression

Suppose a company is interested in predicting the sale of the new product. If the companies feel that sales would depend only on price then a simple regression model can be used to predict the expected sales. However if company feels that the sales would also dependent on manpower apart from prices then multiple-regression can be used with both price and manpower as independent variables and sales as dependent variable.

We will begin with simple regression to understand the mechanics of the regression techniques and then move on to multiple regressions. In theoretical basic there is no difference but we need to keep certain checks in multiple regression.

The simple regression is denoted as

Y=a+bX+e

Where Y is dependent variable

X is independent variable

e is the error involved in using the linear model to predict the value of Y.

a is the intercept

b is the coefficient of X.

Generally we assume error term e to be random variable with mean 0 and variable σ_e^2 .

One important point to be remembered is that any two variables can show some result in regression analysis but it may not necessarily have any causal relationship between two variables. For any regression it is necessary to have actual relationship between dependent and independent variables in real life.

The value of a and b can be calculated by Least Square Estimation as

$$b = \sum (x_i - x_m)(y_i - y_m) / \sum (x_i - x_m)^2$$

where \boldsymbol{x}_m is mean of X and \boldsymbol{y}_m is mean of Y

 $a = y_m - bx_m$

Let us start with a simple linear regression example. Use below data point where sales of a website is dependent on the number of session in a day.

۵:	8 2. 4. 1	- ACANAN - ANNO					Aspenio	A - Excel						
. 10	E HOME POER	T PAGELANDUT FORMU	LAS DATA	REVE	W VEW									
Ê	Xca 8	(albei - 11 - A' A		ę.	Semap 1	at .	General		• R	3	1	11-	*	E Autolum
Pate	Frond Parter	I X × □ · △·▲·	553	42.42	E Marger	A Certiar +	5 - %	• 12	3 Conditi	onal Format	a Call	hat i	Delate Format	2 Capt
	Options 5	Faret	6	Ango	hand	4		-	16	39/45			Celli.	
10	A.	1	¢	0		,	6	н		1		4	M	N
* 5	sales (INR '000)	No. of Session ('000)												
2	120	45												
3	50	25												
4	70	27												
5	90	35												
4	125	40	().											
2	85	30												
	66	25												
	100	40												
10	105	42												
11	110	50	1											
12	135	55												
12	89	35												
14														
2														
17														
18														
39				_							_	_		

Using least square calculation we have calculated a and b as per formula discussed above.

H S	DME INSERT FIL	OF LANSING PORMANE	04% AD	ter ven	Represent to								1 18	- d Spin
Xc.		- (a +)x x = + (a + <u>a</u> + <u>a</u> + <u>a</u> +		Dimptor Ethiopitic		1 5 Carton	- 144	2.2	the last	* () 	E Co		T A	
		100 C		1			-	1	1			M		0
	Sales (INR 10000	No. of Session ('000)	X-X-	X-X-	(X-X_X)(Y-Y_0)	(x,x_)2								-
	120	45	7.58	24.58	186.42	57.51								
	50	25	-12.42	-45.42	\$63.92	154.17								
	70	27	-10.42	-25.42	264.76	108.51								
	90	35	-2.42	-5.42	11.09	5.84								
	125	40	2.58	29.58	76.42	6.67								
	#5	30	-7.42	-10.42	77.26	55.05								
	66	25	-12.42	-29.42	365.26	154.17								
	100	40	2.58	4.58	11.84	6.67								
	205	42	4.58	9.58	43.92	21.01								
	110	50	12.58	14.58	183.51	158.34								
	135	55	17.58	39.58	696.01	309.17								
	87	35	-2.42	-6.42	15.54	5.84								
fotal	1145	449		1.00	2497.92	1042.92								1 2
Mean	95.42	37.42		-										
	6	2.40	1											
		5.80	1											
	Sheetii Sheett	Sheet2 Sheet6 Ma	Agin Sheet	Sheet 1			(4)							
		and the second second second			Carrier Carrier		-							

The first thing one has to do for a regression data is to plot the scatter plot in Excel or R. We will learn R process later, here focus is on excel



One can add trend line to the scatter plot to check if there is actually linear relationship existed between independent and dependent variables.



You can add other trend line as well. We added exponential trendline in the scatter plot



The plot shows the trend line and the equation.



From the linear trend line we can clearly see a good linear relationship between two variables. Let us run actual regression analysis to find the extent of the relationship.

C G + C - Y je +		Benerica - Microsof Dice	h Parai Optione	,		
Park Park <t< th=""><th>General Promotes Reporting Date Manager Cathores Redation Genet Access Participation Next Center</th><th>Excel Were and manage Microsoft Office Add on Add to The Micro Agent South of the Add to Micro Agent South of the Micro Agent Add to Micr</th><th>Depicional Institute Mentina Million Common Million Common</th><th>Ber A Core Alaria Core Alaria Ber Alaria Ber Alaria Doc Alaria Doc</th><th>a g to And Send Q R</th><th>* ************************************</th></t<>	General Promotes Reporting Date Manager Cathores Redation Genet Access Participation Next Center	Excel Were and manage Microsoft Office Add on Add to The Micro Agent South of the Add to Micro Agent South of the Micro Agent Add to Micr	Depicional Institute Mentina Million Common	Ber A Core Alaria Core Alaria Ber Alaria Ber Alaria Doc	a g to And Send Q R	* ************************************
17 18 23 28 28 28 28 28 28 28 28 28 28		Nonsul Graynblan Gargolithi Landon Dronatse Dronatse Nanton del union ten to del Nanton del union ten to del Signapo Borl Adérez V (p	offenering services and services	CK Canad		

Select regression from the list

20 H	47 - (K.)	¥ 20 1	Partend		044		nu put	feg	ession + h	Acresoft	Excel								
1	11 and	terre la	to the second	1198	Contaillan Pagartas 1981 (onka)	11 m	Ser a rea	E. Carp G. Rendorts Ø Advanced	Hil	B-T	14	Constitutes	ES .		1 II 		a land	Sana Analy Analysis	•
_	.74	+ (<i>A</i>				_												
1.00	A.			¢	0	1	1			1		κ.	- 6C		N	0		Q	
1.50	et jave 500	1. 10.1	1 Section (1998)	-															
Q1-		30		-															
		76	20								ata Anata			*					
3		90	35					1.000	-		en reary			- North					
×		125	4					1.53					01						
51-		85						108	ng dianaga	Concerne of			Calv						
<u>- 1</u>								1 80	and for our		_	_	1000						
10		205						1200				_	100						
11		130	N					120	et: Parel Tel et: Tel: Sand	o Sengle for Ne housing	Party laws								
12		135	30	(12	et Toplany at Toplany	de Antonio y	(regarine)								
58													-						
10																			
2																			
17																			
18																			
29																			
20																			
22																			
28																			
24																			
a e e Reaty	e' post	1.2996	. Pett. 9									84C	_			-	Q.0.10	N (C) -	8 2

Select label as we have headers in the data selection. Confidence interval has been kept at default 95%.

20 H 9 . P . T ?	at barrent frame	The Average State	Repression - Microsoft Excel		
		Townshine II III III III III	Corrections and the second sec	Real of the second seco	yentere 19 Data mettere 19 parts
1 A	8 C	0 1 /	Repression	• • • • • •	0
	20 20 20 20 20 20 20 20 20 20 20 20 20 2		Ind. Ind.25mp: Ind.2410 IN Ind.25mp: Ind.25mp:	and	
22 23 24 24 24 24 24 24 24 24 24 24 24 24 24	el, their, the		D•1		

The output of the regression analysis has three main component – summary, ANOVA and coefficient output.

Regression Statistics				
Multiple R	0.916353931			
R Square	0.839704527			
Adjusted R Square	0.823674979			
Standard Error	10.68686993			
Observations	12			

R square of 83.9% is a good fit. This means that 83.9% of the variation of sales is explained by the number of sessions.

ANOVA						
	df		SS	MS	F	Significance F
Regression		1	5982.824777	5982.824777	52.38479347	2.79866E-05
Residual		10	1142.09189	114.209189		
Total		11	7124.916667			

The significance of less than 0.05 is good because we have kept confident interval of 95%. In our case the significance is 0.000028

The residual Error sum of squares (ESS) is calculated as $\sum (y_i \cdot y_j)^2$, where y_i is the value of Y for ith observation and y_j is the calculated y from the regression $y_j = a + bx_j$. This is also called Error Sum of Squares (ESS).

Total Sum of Square (TSS) is calculated as $\sum (y_i - y_m)^2$. The Regression Sum of Square (RSS) is calculates as the difference between Total Sum of Square minus Residual Sum of Square.

 R^2 is calculated as RSS / TSS = 5982.82/7124.91 = 0.839 ~ 83.9%.

A value of $R^2 = 1$ implies that the total sum of square and the regression sum of the square is same (TSS=RSS) and hence regression model explains all the variances of Y.

A value of $R^2 = 0$ implies that TSS=ESS and RSS=0 the regression model fails to explains any variance in present in the data.

For high value of R^2 say greater than 0.75 the positive value for R indicates that the slope is positive and negative value indicates that the slope is negative.



The intercept a is 5.8 and coefficient b is 2.39, hence equation is Y = 5.8 + 2.39 * X.

The intercept and coefficient value are same as one we got in Least Square calculation.

For any value of X we can predict the average value of Y using this equation.

Lets say Number of session (X) = 60

The Sales Y = 5.8 + 2.39 * 60 = 149.2 Thousand INR

Regression in R



Install packages required for the regression




The regression output is same as we get in excel. For the large data set excel may not the best tools hence R will come to be handy in many situation.



6.1.2 Multiple Regressions

The actual problem in the real situation would not be restricted to the single independent variables determining the value of dependent variables. There would be multiple independent variables, we call such regression as multiple regression. The formula for the multiple regressions can be expressed as

 $Y = a + b_1 X_1 + b_2 X_2 + .. + b_n X_n + e$

Where a is intercept of the slope and the b_i is the coefficient of the independent variable X_i . The process of executing the regression analysis is same as simple regression. In below example we added one more variable average discount% as second independent variables.

G	H St C	FACELONDUST FORMUL	o onte novem i	VEW	Asymptotic &	and .							5	7 30	- 0 X
1	Kot (al	e sja skri r x - ⊒- <u>∆</u> - <u>∆</u> -		nge fansk organisk Carrier i i	Canad 5 + % +	9.0 1	(E)	Tabi S	233	- Deles	0	E Adeban Rea-		A	
	Optood %	Aut is	Anground	-	Real of			Styles.		04		44	-		
1	A short Dark Street Bar	a of Casalon PODM	Concerns Discovered					-			-	U		0	-
2f	uses pres boot is	to, or session (cool)	merage concouncy												
1	120	40	0												
	30	10													
	10	27	10												
0ł	135	33	20												
	125		0												
6	85	30	12												
	100	10													
	100	40	10												
3	105	42	18												
	110	50	20												
3	135	33	30												
3	87	35	,												
18]															
2															
14															
2															
10															
72															1 1

Select Regression

				Repression - M	crosoft Excel							- 0	
Parts Incol	A Peprispent A	formular Data Au	nian Van Berel	lager Add-Int								-0-	. 0
BBI	a là	Cartana Cartana	HE Y	A.Com	Internet Cate Contra	- 12	1	1	1	ne Detait	To take		
to the first him	ros - Constant	Ap. and all press	Sort & Fills	Granance Column Column	Eutofesi Vatilation * Euto funti	Analysis *		0.00		-	Anatori		
	4								-				_
A	- 1	¢	0 6	1 6	H 1 2	× 1	1.	м		0		Q	
Sales (INR 1990) No	s, of Session (1998)	Average Discount%											
130	45												
70	22	1			Data Bashala								
90	35	30		maha	Care weather		-						
125	40	25		Print Tractoryle	for its lenses								
86	15	1		Phone and one		Carl	e						
100	40	16	C	Arriter for the	Contraction of the local data								
205	42	18		April and Percently	CASA A	- 1973							
139	30	20		Sampling In Test, Farmed Ture 1	Canada for Persons								
10	35	,		In fact, Two Garage	Assertig Spart Granus								
* Derti De	et Sheet? Hull	ple Sheet) . 2			Del								1.1
121 1.0.1 ()	• Period	familia (Eris) A	eiter Ven Ber	Repression - M	crosoft Decel					aas	en Oa		
	+ + Narleest + - Dar - Daris	Terminia Terla di Consentanti Consentanti Statuna antati sana		Repression - M new Address & Cross (- Cross (- Cross) (-)= 122 men 122 major	¥.	1	1.1.	en fotat in fotat	to tone Batter in	- 0 + 0 - 0	
	Page Laword	Ternator Data A		Repression - Million Balance Address Balance Statistics Statistics Statistics	couch Deal)- Rig Hanne Hanne	1	1	1	a a s	teresteres Antonio Antonio		
	Provisional Provision	Ternator Teris A Constant Sector Sector Sector Sector Constant	11 III Providence Prov	Repression - M law Addim Schweith Station Schweith Construction Construction	cosof Deal	in Right Martin Martin		1		en fotal ar fotal	Ban () Banar Banar Anator	- 0 + 0 0	
	Provincent	Termine Data 1	na Ve Ber 11 III (1990) 13 Set 100 Ber 1 Aug	Repression - M new Address & One Control of Pression Control of New York	crosoft Deel)- 12 	12.	1	- - -	en falat in falat in falat	the Color	- 0 + 0 0	
0 - 0 - 7 - 7	Preprintend	Arraya Sala Arraya Arraya Sala	nes Ves Ber II II II II II II II II II II II II II	Regression - M new Adding & Con Sciences (Consent) Consent Seed Teact Tange	Crosoft Deal Tel:			श्च । 	41 - 	en lidat ar bitat ar bitat	in () Se tana By tana Anatomi	- 8 + 9 10 	
Port Port	Poper Learned Poper L	tenute Dis It () () () () () () () () () ()	Dia Van Der 11 Jan Van Der Jan Anter Dia	Regression - M team Address Science Sciences Sci	count besi and the set and the set Asymptotic Marketo Marketo	· · · · · · · · · · · · · · · · · · ·			- 1- 	en linte ar binat o	the C	- 0 + 0 = 4	
9 + 0 - 7 34 Prote Bart 6 + 0 - 7 34 Prote Bart 6 + 0 - 7 34 10 + 0 - 7 - 7 34 10 + 0 - 7 - 7 - 7 - 7 - 7 - 7 - 7 - 7 - 7 -	Propriogram Propriogr	Ternate Data II Constanting Solution S	Den Ven Ber 14 Junior Ven 15 Junior Ven 16 Anton D E	Regression - M mare Adom 6 form 6 form 7 Adom 7 Ado	coset Lesi and Ed Provider (or Server) Regression Regression Provider Converta Jor	· · · · · · · · · · · · · · · · · · ·				en Delat in Senat	the C	- 0 + 0 = 	
9 + 0 - 7 24 Point Boot 6 - 0 - 7 24 Point Boot 6 - 0 - 0 10 10 10 10 10 10 10 10 10 1	Propi locad Propi loc	Annual Dis A Control of Control on Control of Control on Control of Control on Control of Control on Control of Control of Control Annual Control of Control Annual Control of Control Control of Control of Control of Control Control of Control of Control of Control Control of Control of Co	Den Van Der H ja set ja set bet AMA D E	Regression - M mer Anton Com Com Com Com Com Com Com Com	count Levi Anne Con Con Anne Con Con Con Van Repression Part Hol Part	in the second se		श्च । •		en Delat an Delat an Delat	in () i inne i j inne nega	- 8 + 9 - 0 8	
0 • 0 • 1 1 0 Note: Note: Noo	Pope Losend Pope Lose	Termine Brief Termine Termine Termine Termine Termine Termine Termine Termine Termine Termine Termine	No. Von Der 11 III III Provi III III III Provi III IIII IIII III IIII IIII III IIII III IIIII III IIII III IIIII III IIIII III IIIII III IIIII III IIIII III IIIII III IIIII III IIII III IIII III IIIII III IIIII III IIII III IIII III IIII III IIII III IIII III IIII III IIII III IIIII III IIII III IIII III IIIII III IIIII III IIIII III IIIII III IIIII III IIIII III IIII III IIIII III IIIII IIIIIIII	Repression - M men Advin Com Prevent	Coston Excer Transmission Construction Transmission Construction Transmission Construction Transmission Trans	9 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		या । •	. 1:	en Telat ar Detat	in () i inne i j inne nega	- 8 + 9 - 0 8	
	* Proprised 6 * Proprised 6 * * * * * * * * * * * * * * * * * * *	tempte Deta A Concentration American Deta A American Deta American Concentration Concentrat	Non You Pero II II II II II II II II II II II II II	Repression - M term Address Drawn Drawn Drawn Drawn D Drawn Drawn D Drawn Drawn D Drawn Drawn D Drawn Drawn D Drawn D D	count Less and a second secon			1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		en Telat ar Setat	tone Constants tone	- 8 + 9 0 non	
9 • 0 • 1 • 0 • 1 • 0 • 1 • 0 • 1 • 0 • 0	Proprised	Termine Determine Termine Deter	New Yes Person 11 New Yes 12 New Yes 14	Regression - M terr Adver 6.000 Press Pr	Count Deal	- 22 		2) i 2		en Total In Selan	in Contraction	- 8 + 9 0	
5 + (2 + 3) 2 = 2 These transformed and the second	* Page Looked * Page Looked * * * * * * * * * * * * * * * * * * *	Apenador Boto A Concentration Conc	No No No II II II II II II II II II II II II II	Regression - M term Asten - Annual - Annual	Concert Deal	in River and a second s		원 (·		o Triat	in C	- 0	
9 + 0 + 1 / 2 New 2 + 1 / 2 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 +	* Proprietant / * Proprietant / * * * * * * * * * * * * * * * * * * *	Average Docurts	New Yes Deve H II II II II II II II II II II II II Det A Mar D	Repression - M terr Addres 6 Com 6 Com 6 Com 6 Com 7 Addres 1 Addres	count Less Terrer Con Ver Regression Particular Regression Particular Converts Jon Converts J			일 (·		o line in the second se	Rink () The Server By Grant And Andrews P	- 8 + 0 - 1 	
9 + 0 + 1 / 2	Propriored Propriored Propriored Propriore Proprior	Apenden Des A		Repression - M terr Adver 6.000 Press Pr	Concert Even	in the second se				en Total In Long	Bin ()	- 6 + 0 - 1 - 6 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1	
0 + 0 + 1 / 2 Note: 4 - 1 / 2 Note: 4	Proprieted P	Annual Description	New Yes Per II II	Repression - M	creat Lesi Terren Scher, Gen Berren Scher, Gen B			21 I 1 1 1		en Detet in State 0	ing () Se tana Bi tana Angin P	- 8 4 0 10 4 0 10	
9 • 0 • 1 • 0 • 1 • 0 • 1 • 0 • 1 • 0 • 1 • 0 • 0	* Page Legend of Page Legend of Descention (************************************	Average DocurtS	P E	Regression - M ter Addres 6. Con (- Addres 6. Con (- Addres 1. C	count Leur Terrer Con Long Con Lon	- 22 		i i tra		en folge n folge	in contraction of the second s	- 8 * 9 to 49to	
9 • 0 • 1 • 1 • 1 • 1 • 1 • 1 • 1 • 1 • 1	Proprieted P	Apenden Beis R. Constantion C		Regression - M mer Astron Control Control Co	Concert Eases The second seco			2] [] 1944 M		en folde e Balan	inter constructions of the construction of the	- 0 + 0 10 nijes -	
	* Proprieted 1 * Proprieted 1	Temple Bris R Constant C	New Yes Percent II II	Repression - M max Astron Construction Provide Construction Provide Response Provide Response Pr	count Lesi There is a second	in the second se				ee babe na babe	inter construction of the second s	• • • • • • • • • • • • • • • • • • •	
P • 0 • 0 • 0 • 0 • 0 • 0 • 0 • 0 • 0 •	 Page Loosed Page Loo	Apended Description	De C	Regression - M are Address & Cons (count Lees Terrer Con Units Con Segretion Particular Particula	and an and a second sec		in the second se		ee Datat ie Souri	inter constructions of the second sec	• • • • • • • • • • • • • • • • • • •	
9 • 0 • 1 • 1 • 0 • 0 • 1 • 0 • 0 • 0 • 0	Proprieted P	Aprenative Bots Research Concentration Concentrat	Nor You Per H II II II II II II II II M II II II M II II N II II N II II N III N II N III N IIIII N III N IIII N III N III N III N III N III N I	Regression - M	An State Sta			New York		es bint es bint e Sinu	interest of the second se	• 8 • 0 0 • 10	
9 • 0 • 1 • 0 • 1 • 0 • 1 • 0 • 1 • 0 • 1 • 0 • 1 • 0 • 0	 Page Legend of Comments Description Description Description def Sension (1990) d	tempte Bris E Concentration Concen	Den Ven Der H II II II II III II III III III Det Alter Det E	Repression - M the Adding Composition - Marcine Composition - Marcine Composition - Marcine Composition - Marcine Part Starge Part Starge Part Starge Composition - Marcine Composition - Marcine C	count Leus Terren Con Con Con Ven Regression Paratel Con Ven Paratel Con Ven	in the second se		191 E		en Delet in Bout	nte C	• 0 • 0 • 0 • 0 • 0 • 0 • 0 • 0 • 0 • 0	
1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 10 0 0 0 10 0 0 0 10 0 0 0 10 0 0 0 10 0 0 0 10 0 0 0 10 0 0 0 10 0 0 0 10 0 0 0 10 0 0 0 10 0 0 0 10 0 0 0 10 0 0 0 10 0 0 0	Propriored P	Apenden Bes A		Regression - M mer Astron Control Control Co	Concert Deal					en Inde an Inde O	ne contra a series a	• 8 • 0 0 • 0	

SUMMARY OUTPUT					
Regressio	n Statistics				
Multiple R	0.965678475				
R Square	0.932534918				
Adjusted R Square	0.917542678				
Standard Error	7.308162008				
Observations	12				
ANOVA					
	df	SS	MS	F	Significance F
Regression	2	6644.233579	3322.11679	62.20117139	5.38091E-06
Residual	9	480.6830874	53.40923194		
Total	11	7124.916667			

The R Square has increased after adding one more independent variable. That means 93.25% of the variation of the sales can be explained by both independent variables.

	Coefficients	Standard Error	t Stot	P-volve	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	30.74	11.24	2.73	0.02	5.31	56.18	5.31	56.18
No. of Session ('000)	0.85	0.49	1.73	0.12	-0.26	1.97	-0.26	1.97
Average Discounds	2.10	0.60	3.52	0.01	0.75	3.45	0.75	3.45

Equation is $Y=30.74 + 0.85 X_1 + 2.10 X_2$.

Assuming number of session is 50K and discount is 19% then Sales

```
Y = 30.74 + 0.85*50 + 2.10 * 19
```

= 113.4 ('000 INR)

Multiple Regressions in R



Using same data set as excel example.



The output is same as excel output.



6.1.3 Generalized Linear Model

There are two basic assumption of linear regression model Y=a+bX that is

- Y is continuous and has normal distribution.
- The effect of change in the independent variable on dependent variable is linear in nature. If X change by p then Y change by q; if X change by 2p then Y change by 2q.

In many situations there are cases where above two assumption does not hold

- The dependent variable Y is nominal or categorical e.g Success/failure
- The effect of change in independent variable X to dependent variable Y may not be linear. E.g. consumption of staple food will increase with increase in income. When income is at lower level then it may increase at higher rate but once income cross certain level the increase in consumption may be nil or very less.

Generalized linear model has been created to tackle above two problems in linear regression model. The values of the parameters (b_0 through b_n and the scale parameter) in the generalized linear model are obtained by maximum

likelihood (ML) estimation, which requires iterative computational procedures whereas linear regression uses least square method.

In linear regression model the dependent variable Y is associated with independent variable X variables by

 $Y = a + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n + e$

Here expected value of error term e is assumed to be 0

In GLM the relationship between Y and X is assumed to be

 $Y = g(a + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n) + e$

Where e is error and g is a function g(x). The inverse of g is called link function denoted by f(x)

 $f(\mu_y) = a + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$

 μ_v is expected value of Y

The link function can be any of the below

Family	Default Link Function in R
binomial	(link = "logit")
gaussian	(link = "identity")
Gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

Logistics regression is done using binomial family. Readers can check out other family and try out different combination of numbers to practice.

Learning from the Chapter

- What is regression -simple and multiple regression
- How to run regression in Excel and R
- How to interpret the regression program output from Excel and R
- Generalized Linear Model (GLM)

Section - II

TIME SERIES FORECASTING

A time series is a continuous timed interval data point with each time period having one data and the time intervals are uniform across the successive data points. An example of a time series data is stock price of a company for the day taken at day closing for each day over the period of years. The basic objective of the time series analysis is to find the autocorrelation between successive data point and the trends of the data points over the time. In regression analysis we learnt in previous chapter the dependent variable is predicted using independent variables. Here in time series the concept of dependent and independent variables are missing rather the prediction for the future data point are based on the historical data of the same data series. In this section we will learn some of the commonly used time series forecasting techniques.

6.2.1 Simple Moving Average (SMA)

One of the simplest methods of the time series forecasting is moving average. In this method the average of the certain set n is calculate. The window size of n is moved one number or point forward and next moving average is calculated. This process continues till all moving average is calculated. The moving average provides very good understanding of the long term trends and so on. It is also used for smoothening of the data series to smooth out the short term fluctuation and highlight the long term trends.

 $F_t = (1/N) (D_{t-1} + D_{t-2} + ... + D_{t-n})$

In below example we calculated 5 days SMA and 10 days SMA.

E) FR	E HOME	INSERT PA	IGE LAYOUT F	ORMULAS DATA	REVI	EW 1	new .		Forecastic	ng + Exc	el	
1	Xcu	Arial	- 10 -	x x ==	87.	-w	ap Tet		General			1
Pait	* Sormat Paint		ı • 🗇 • 🛕 •	▲· ■ ■ ■	42 42	E Me	rge & Ce	nter •	5 - %	•	52 22	Con
	Clipboard	6	Fort	6	Align	ment		· ~	N	mber		
1	A	0	c	0		1	G	н	1	J	×	
	Date	Price	5-Day SMA	10-day SMA	_							
	24-Mar-10	22.27	10000000000									
	25-Mar-10	22.19										
ŧΕ	26-Mar-10	22.08										
88	29-Mar-10	22.17										
ŧΕ	30-Mar-10	22.18	22.18									
ŧΕ	31-Mar-10	22.13	22.15									
11	1-Apr-10	22 23	22.16									
10	5-Apr-10	22.43	22.23									
11.	6-Apr-10	22.24	22.25	the second								
1	7-Apr-10	22 20	22 27	22.22								
10	8-Apr-10	22.15	22 27	22.21								
10	9-Apr-10	22 39	22 30	22 23								
	12-Apr-10	22 38	22 29	22.26								
1	13-Apr-10	22.61	22.37	22 31								
1	14-Apr-10	23.36	22.58	22.42	1							
10	15-Apr-10	24.05	22.96	22.61								
1	16-Apr-10	23.75	23 23	22 77								
1	19-Apr-10	23 83	23.52	22.91								
1	20-Apr-10	23.95	23 79	23.08	1							
1	21-Apr-10	23 63	23.84	23.21	1							
1	22-Apr-10	23 82	23.80	23.38	1							
5	23-Apr-10	23 87	23.82	23.53								
1	26-Apr-10	23 65	23 79	23.65				-				
1	27-Apr-10	23.19	23.63	23.71								
	28-Apr-10	23.10	23 53	23.69					-			
-	29-Apr-10	23 33	23.43	23.61								
1	30-Apr-10	22.68	23 19	23.51								
1	3-May-10	23 10	23.05	23.43	1							
-	4-May-10	22.40	22.92	23.28								
	5-May-10	22 17	22.74	23 13								

The 5 Days SMA is smoother than the actual data and there is lag in the SMA trends.



Plot 5 days SMA and 10 days SMA along with the price in the same graph



The longer the period or size of the window more the lag in the data. 10 days SMA has lag over 5 days SMA and is smoother also. A moving average is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles. The threshold between short-term and long-term depends on the application, and the parameters of the moving average will be set accordingly.

SMA in R

We will use same data set to show how to calculate and plot SMA in R.

0	e	Studie		- 5 -
Fin Lat Cole Yaw Puts Sesson B	Lafe Debug Tank Help Texturnet () () + Addep =			S Papet (fand
Conterre : Conterre : 1 Sis- data, franc Sales-c 2 Sis- 3	9. Z + 10 (n.v)4, n.166, n.660, n.v41, n.164, 1	*** *** ** ***** * * 2.114, 2.107, 2.407,	Entrement C II (Pingert Déset +) Dittel Inconnect+ O STs 30 obs. 4	f 1 variable
31 Depinet 1	-	they ?	Hatery files Parks Parkages	Net Years
Loss (-) D 271 seconcerted (-) (a. 2) 5.55x-dets, frame(sales, (-), 2) (a. 2) (b. 2) 27.2466, 27, 2713, 27, 25, 21, 47, 22 (b. 2) (b. 2) 5.16 (b. 2) (b. 2) (b. 2) 5.18 (b. 2) (b. 2) (b. 2) (b. 2) 5.18 (b. 2) (b. 2) (b. 2) (b. 2) (b. 2) 5.18 (b. 2) (b. 2)	.2734, 22.354, 32.0847, 22.1741, 22.384, 22. .2754, 22.354, 22.0547, 23.1554, 24.029, 23. .28.6547, 23.387, 23.0976, 23.324, 22.6805, 23.0976, 23.324, 22.6805, 23.0976, 23.324, 23.6805, 23.0976, 23.0	-C 1144, 22,2107, 32,409, 27,711, 27,8404, 23,9516 23,6956, 22,4025, 22,172		* * * * * * * * * * * * * * * * * * *



Plot of the actual data.

0	RStudio	
His Edit Code View Plats Sessen Build Delug Tools Help		
🔍 + 🎯 + 🔛 🕼 🖓 (Alexandronia) 🔯 + Addma +		B Popul (New) +
•	Entraneat	-0
E B Clearadar Q / - C ate 30 aleas - 2	🞯 🔛 (Pinget Delest + 🖌	31 bet + (🐨
2 sh	OuterEnvironment +	9
1 Horary, 'Heseries') 4 TBa-ts(35) 5 plot.ts(78)	SLIMA Time-series [1:30] from 1 to 30; MA NA N Oxim,model List of 30	u na 22.2
6 110rary("118") 7 SL988SHA(TSL, m-5)	Habery Has Path Packages Hole Venant	-0
8 plot. ts (SLSRA)	a a Plan Alaste Q d	S hand S
A Street of Contract of Contra	CREATE STREET MARKEN PARK	
NI (tiglan) I Klopt I		
0.3-7.249 0.3-7.249 Content type "application/aig" length 65/8/1 bytes (642 x8) demloaded 642 x8 trying URL "https://iran.rstudie.com/bin/window/contrib/3.3/TRL 0.27-3.25 Content type "application/aig" length 433022 bytes (422 x8) demloaded 22 x8 package "ats" successfully unpacked and 905 sums checked package "TR" successfully unpacked and 905 sums checked clumers/lenoso/appbsfaisocal/temp/dtspb0pa4a/downloaded. package "TR" ass built under % version 3.3.2 > \$2504-200(TR, n-5) = plat.	WNS75	w t

Plot of the 5 days SMA shows smooth trended graph.



Combining both price data and 5 Days SMA and plotting in the same space. The 5 days SMA is smoother and has some lag. Similarly we can plot 20 days SMA. SMA is useful in smoothening the trends and provides low pass filter.

6.2.2 Exponential Smoothening Methods

In Single Moving Averages the past observations are weighted equally, Exponential Smoothing assigns exponentially decreasing weights as the observation get older. In other words, recent observations are given relatively more weight in forecasting than the older observations. The relative weightage given to the recent observations are determined by the smoothening parameters α . In the case of moving averages, the weights assigned to the observations are the same and are equal to 1/N.

The simple form of Exponential smoothening is given by formula

 $F_t = \alpha$. $X_t + (1-\alpha) F_{t-1}$

Where F_t is the weighted average of the recent observation X_t and the previous smoothened data F_{t-1} , α is the smoothening factor which ranges from $0 < \alpha < 1$.

The higher value of α gives more weightages to the recent observation and the lower value of α gives more weightages to the older observation in the time series data. There is no ideal value of α but one can do trial and error to get the required smoothening effects. One statistical method to arrive at α is to find the value of α that minimizes the $(F_{t-1} - X_{t-2})^2$

Small values of α means that the forecasted value will be stable (show low variability) and increases the lag of the forecast to the actual data if a trend is present, whereas large values of α mean that the forecast will more closely track the actual time series data.

Exponential smoothing is commonly applied to smooth data, as many window functions are in signal processing, acting as low-pass filters to remove high frequency noise.

In below time series data we will apply different α to see how it impact the smoothened data.



Using $\alpha = 0.45$ and $\alpha = 0.05$ we created two exponential smoothening series. As we can see $\alpha = 0.45$ is very close to the actual data and $\alpha = 0.05$ has long lag and too smooth.



Exponential Smoothening in R

Using same data series to show EWA work in R.



We plot Exponential smoothening using $\alpha{=}0.02$ (ES1) , $\alpha{=}0.45$ (ES2) and $\alpha{=}0.05$ (ES3)



Adding data point in the graph



Add ES1 line with blue color



Adding ES2 line with red color



Adding ES3 line with green color

0 1	Studie	
Na Edit Cole Ver Park Sesen Bull Delug Tork Mig Q.+ 🐲 + 🔛 🔛 🔄 di la tradición 🔅 🛄 + Addes +		8 Popul David
C Special Insulant (abl	Indexement Indexement Interference - Interference	1 Las - 0
13		A

The plot of α =0.45 and α =0.05 shows same trends as we saw in excel output.

Now to forecast the next 5 value using the current time series data we can plot the value as shown below. You can change the h value to see different time period of the forecast.

0 0	Studie	
His Life Code View Parts Session Build Delvag Tools Help		
🧶 🐲 🗧 🕼 🕼 🕼 A ten televise 🛛 🔯 * Addes *		B Popul (New) +
Characteria (solution and solution and solut	Industant	-0
D Owners for Q. Q. Dense 1 Dense 1 Dense 1 Dense 2 Dense 2 Dense 3 Den	dt Bil (Propublicat- de Baar Incoment- 0153 List of 9 0fg List of 9	11 Ser - (@
11 (2)-tex(SP), alpha = 0.05, initial - "timple", h = 1) alpha(13), plot.com(all, ylab-"lab", xlab-"subarter", main." 11 Theory (fitted 32), col."year, "type"") 12 Theory (fitted 33), col."year, "type"") 13 Theory (fitted 33), col."year, "type"") 14 For-tex(SP1, h-5, alpha = 0.65, initial = "simple") 29 plot(fs)	Newy He Rei Arlant Ne Your 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	s tase 👘
31 Orginal 1 (* burg) 10 Orginal 1 (* burg) 11 String String String (* burg) 11 String String String (* burg) 11 String String (* burg) 12 String String (* burg) 13 String (* burg) 14 String (* bu	000 000 001 001 000 000 5 50 55 20	25 20

6.2.3 ARIMA Model

Exponential smoothing and ARIMA models are the two most widely-used approaches to time series forecasting, and provide complementary approaches to the problem. While exponential smoothing models were based on a description of trend and seasonality in the data, ARIMA models aim to describe the autocorrelations in the data

The ARMA has three components – Autoregressive (AR), Moving Average (MA) and Integrated (I). Let us understand each component before going into the full model.

Auto Regressive (AR)

Autoregressive (AR) models are models in which the value of a variable in one period is related to its value in previous periods. AR (p) is an autoregressive model with p lags can be denoted by

 $Y_t = c + b_1 Y_{t-1} + b_1 Y_{t-2} + \ldots + b_p Y_{t-p} + e_t$

Where c is constant and e is error term (white noise).

Moving Average (MA)

Moving average (MA) model account for the possibility of a relationship between a variable and the residuals from previous periods. Autoregressive moving average (ARMA) model combine both p autoregressive terms and q moving average terms.

 $Y_t = c + e_t + b_1 e_{t-1} + b_2 e_{t-2} + \dots + b_q e_{t-q}$

 $_{Yt}$ can be thought of as a weighted moving average of the past few forecast errors. This is not same moving average as moving average smoothening we have learnt earlier.

Integrated (I)

The time series forecasting with seasonality or trends are not stationary. The seasonality and trends will affect the value of time series at different time. The white noise e_t and cyclicity has been considered to be stationery as it has no effect on the data based on time at which item is observed.

Differencing is a method of making time series stationary. Differencing can help stabilize the mean of a time series by removing changes in the level of a time series, and so eliminating trend and seasonality.

The differenced series is the change between consecutive observations in the original series, and can be written as

 $y'_t = y_t - y_{t-1}$ Seasonal differencing can be done as

 $y'_{t=yt-yt-m}$ where m= number of seasons.

Non Seasonal ARIMA Model

ARIMA models are generally denoted ARIMA(p, d, q) where parameters p, d, and q are non-negative integers, p is the order of the Autoregressive model, d is the degree of differencing, and q is the order of the Moving-average model. If we combine differencing with autoregression and a moving average model, we obtain a non-seasonal ARIMA model

 $Y'_{t} = c + b_{1}Y'_{t-1} + b_{1}Y'_{t-2} + \ldots + b_{p}Y'_{t-p} + b_{1}e_{t-1} + b_{2}e_{t-2} + \ldots + b_{q}e_{t-q} + e_{t}$

where Y'_t is the differenced series.

Model with different value of p,q and d gives following types

RIMA(0,0,0)
RIMA(0,1,0) with no constant
RIMA(0,1,0) with a constant
RIMA(p,0,0)
RIMA(0,0,q)

ARIMA with R

Having understood the theoretical part of ARIMA let us try to work out an example of a forecasting. We will use same data series as Exponential Smoothening example.

> SM <-data.frame(Period=c('2000-Q1', '2000-Q2', '2000-Q3', '2000-Q4', '2001-Q1', '2001-Q2', '2001-Q3', '2001-Q4', '2002-Q1', '2002-Q2', '2002-Q3', '2002-Q4', '2003-Q1', '2003-Q2', '2003-Q3', '2003-Q4', '2004-Q1', '2004-Q2', '2004-Q3', '2004-Q4', '2005-Q1', '2005-Q2', '2005-Q3', '2005-Q4'), Sales = c(362, 385, 432, 341, 382, 409, 498, 387, 473, 513, 582, 474, 544, 582, 681, 557, 628, 707, 773, 592, 627, 725, 854, 661))</p>



Convert data to time series data

- > SM1<-ts(SM\$Sales)</pre>
- > AR<-ts(SM1, frequency=4, start=c(2000,1), end=c(2005,4))

Plot the time series data

> plot(AR, xlab="Time", ylab="Sales")



Differencing data to make it stationery as we can see there is clearly a seasonal factor in the data. Pot od differencing d=1

```
> ARD<-diff(AR)
```

```
> plot(ARD, xlab="Time", ylab="Sales")
```

The mean seems to be stationary but variance is not stationary



Tryout with logarithm of the data

> AR10 <-log10(AR)

> plot(AR10, xlab="Time", ylab="Sales")

From graph we can see that the variance seems to be stationary but mean is not.



The log10 of the difference of data should make mean and variance stationary

- > AR10 <-log10(AR)
- > ARDIF10 <-diff(AR10)

> plot(ARDIF10, xlab="Time", ylab="Sales")



Now let's try to find the best fit ARIMA model

>require(forecast)

```
> model<-auto.arima(AR)</pre>
```

> model



So best fit model is ARIMA (0,0,1)(0,1,0) with 4 additional differencing for seasonal time lag.

> fit <-arima(SM1,order=c(0,1,3), seasonal =c(0,1,1))

> plot(forecast(fit, h=4))



> fit <-arima(SM1,order=c(0,0,1), seasonal =c(0,1,0))

> plot(forecast(fit,4))



> plot(forecast(fit, h=12)) # for 12 quarter ahead forecast







25	633.3992	519.6330	747.1654	459.4088	807.3896
26	612.1237	468.4816	755.7658	392.4421	831.8053
27	595.7239	436.9711	754.4767	352.9325	838.5154
28	583.0825	415.9972	750.1678	327.5476	838.6174
29	573.3381	401.4931	745.1831	310.5239	836.1524
30	565.8269	391.2152	740.4386	298.7813	832.8724
31	560.0370	383.8019	736.2720	290.5088	829.5652
32	555.5739	378.3814	732.7665	284.5814	826.5665
33	552.1337	374.3747	729.8927	280.2748	823.9927
34	549.4819	371.3871	727.5766	277.1095	821.8543
35	547.4378	369.1438	725.7317	274.7607	820.1148
36	545.8621	367.4499	724.2743	273.0042	818.7200

You can try different combination of (p,d,q) and forecast for different time period to gain more familiarity.

Learning from Chapter

- Understanding of time series data
- Simple Moving average calculation in Excel and R and application
- Exponential Smoothening calculation in Excel and R and its application
- ARIMA Model understanding, formulation and calculation using R

Chapter - VII INVENTORY MANAGEMENT

"For want of a nail the shoe was lost, For want of a shoe the horse was lost, For want of a horse the knight was lost, For want of a knight the battle was lost, For want of a battle the kingdom was lost. So a kingdom was lost—all for want of a nail."

During second Punic war between Rome and Carthage, Hannibal Barca the Carthaginian general invade Roman mainland from Spain winning string of war in Italy. Roman capital Rome was within the grasp of Hannibal's army. He was running short of army, ammunition and siege weapon to conquer heavily fortified city Rome. He sends out request for more army and weapons to Carthage senate. After getting refusal from Carthage Hannibal tried using existing army to win as many allies as possible but Roman army knowing that they cannot defeat Hannibal in open war turned to hit and run tactics. Slowly Hannibal Army was reduced due to defection, Roman harassment and disease. Hannibal was finally called back and his army was defeated in battle of Zama by Scipio, the Roman General. Carthage lost the Second Punic War and was indebted to Rome for the next fifty years. Carthage was finally destroyed after third Punic war. Had Carthage sent require supplies to Hannibal on time, second Punic war could have been won by Hannibal and history could have been written by Carthage instead of Romans.

Likewise there are numerous instances in the history the army lost war due to shortage of supplies. In today's cut throat competition in the marketplace, the lack of stock in the store or site not only leads to loss of sales but also loss of customer's loyalty. This chapter is design to help readers understand some of the inventory management areas useful to us as an analyst.

An inventory is a stock or a store of goods. Inventory management is integral to the many of the business and customer satisfaction is highly impacted by how well inventory is managed. Not only in manufacturing organization but also in services organization inventory decision can be critical. For example a hospital carry inventory of drugs and blood supplies that might be needed in short notice. Being out of stock in on some of these could imperil the well-being of a patient.

For some readers the inventory management may seems out of the context in the digital analytics books. I have included the basic understanding of the inventory models and service level to develop understanding of how the inventory impacts the overall business level. Be it ecommerce or offline store the final transaction happens on the delivery of the products and services to the customer. The customer satisfaction comes from the products available on right time in right quality at right price. Along with the customer satisfaction there is huge business loss due to out of stock in every business. Reducing the out of stock by managing proper inventory through number driven analytics is the motto of this chapter. The correct and real-time inventory tracking is important for any digital marketing and site conversion.

Inventory management is very complex situation due to complexity of the demand and supply. Different companies employ different model to deal with the situation. In this chapter we start with the assumption that we know the demand and supply with definite probability to build the simple model. Once we understand the basic model we extend it to more complex situation. The key metrics of the inventory discussion in this chapter will be service level which is nothing but the probability that the customer will be provided with required product and quantity. The service level directly impacts the customer satisfaction and reduces the business loss due to out of stock situation.

The Section I deal with the classical Economic Order Quantity (EOQ) model to generate the optimum order quantity and the replenishment model under variable demand.

The Section II deals with the transportation model using Integer programming. The model is important for the retailers having multiple stores and multiple warehouses or marketplace ecommerce model with multiple sources of products and multiple customer demand.

Section - I

INVENTORY MODEL

7.1.1 Basic Inventory Model

Let's us assume a very simple business model where a store keep product A for sale. Store purchase product A from a single supplier in the vicinity. The customers come to store for purchasing product A. For store the customer demand are known in advance and supplier's lead time to deliver products to the store is constant. The first priority of the store owner is to fulfill all customers' order without any delay. At the same time store has to purchase only required quantity from the supplier to keep working capital minimum.



The tradeoff for the store is to satisfy the customer demand and keep inventory holding and ordering cost as low as possible. For every purchase order raised store has to incur cost in term of the manpower, printing cost, courier cost etc. The excess inventory will inventory holding cost. More the frequency of order the cost of ordering goes up whereas more that quantity orders the cost of holding inventory goes up. The objective is to find the optimal solution what keep the overall cost minimum. This is what Economic Order Quantity (EOQ) model does.

Total Inventory cost = [Purchasing cost] + [Ordering cost] + [Holding cost] + [Shortage Cost]

The objective of the store is to keep total cost minimized. The purchasing cost and shortage cost will be ignored for our basic model analysis because the purchase cost is constant irrespective of the number of orders and shortage cost is subjective.

The variables that impact the inventory model are

Lead Time (L): The time interval between ordering and receiving the inventory.

Holding Cost: the physical inventory carrying cost includes costs like interest cost, insurance cost, taxes, depreciation, obsolescence, pilferage etc.

Ordering Cost: The cost of ordering and receiving inventory. They are the cost that varies with the actual placement of an order. Probably in some cases transportation cost also need to be included.

Shortage Cost: This happen when demand exceed the supply of inventory on hand. These cost can includes the opportunity cost of not making a sale, loss of customer goodwill, late charges and so on.



Consider the above figure; Q is the quantity order in say time T_0 . With the fixed demand rate of D per period the inventory Q is consumed in T_2 - T_0 period. The supplier has lead time (L) of T_2 - T_1 that is order placed in T_1 will be delivered in T_2 . As we know the demand with definite certainty every time when quantity is reaches to level reordering level R the order is placed so that it is delivered when Q is fully consumed. Basically R is the inventory that is consumed in time T_2 - T_1 which is the lead time of the order; that means Q-R quantity are consumed in time period T_1 - T_0 . This is repeated for the second order and so on.

Assuming the demand is constant throughout the year the average inventory of the system would be Q/2. Let H be yearly interest rate the cost of carrying inventory is -

Inventory Carrying cost = Q/2 * H

As we can see from the formula lower the Q lesser the carrying cost but then frequency of the ordering would just go up. Assuming the ordering cost as S per order then ordering cost is -

Cost of Ordering = D/Q * S

Therefore total inventory cost = Carrying cost + Ordering Cost = (Q/2)H + (D/Q)S.

To find the minimum point of the curve we have to differentiate the Total cost with respect to the quantity

 $dTC/dQ = dQ/2*H + d(D/Q)S = H/2 - DS/Q^{2}$ at minima dTC/dQ = 0 $0 = H/2 - DS/Q^{2}$ $Q^{2} = 2DS/H$ $Q = \sqrt{(2DS/H)}$



The optimal quantity Q* is known as the Economic Order Quantity (EOQ) and this model is called as EOQ model.

Let us solve a simple example to show how the calculation is done

A demand for a Bulb is 100 units per day, it cost company 100 to initiate a purchase order, the cost of storage is Rs 0.02 per day and the lead time for delivery of bulb is 5 days.

 $Q^* = \sqrt{(2DS/H)} = \sqrt{(2^*100^*100/0.02)} = 1000$ bulbs.

Reorder Quantity = 5*100=500.

The inventory is ordered when it falls just below 500.

Total cost of the inventory on daily basis= [1000/2 * 0.02] + [100/(1000/100] = Rs. 20 per day

Frequency of order = years Demand / EOQ = $100*365/1000 = 36.5 \sim 37$ order in a year

Cost of Ordering = Ordering Cost * Frequency of Order = 37*100 = 3700

7.1.2 Probabilistic Model

The basic EOQ model we have just learnt is based on the assumption of the constant demand throughout the year and constant lead time. However in real life the demand is variable and the lead time is also variable. The variability in the demand and lead time introduce uncertainty in the ability to meet the demands. Given the variability in the demand the probability of meeting the demand of the customer is known as the service level.



Consider the above figure, in the deterministic demand condition the order will be placed on T1 and the order will be received in lead time (L) that is T2. However when the demand is variable as shown in the figure there will be shortages and overstocking. The amount of the shortage and the overstocking will be determined by how much is the variability of the demand that is standard deviation of the demand. More that variability higher the cases of stock out or overstocking. As a company we want to reduce the overstocking to reduce inventory carrying cost and obsolescence the stock out is directly impacting the customer satisfaction. *Order service level can be defined as the probability that demand will not exceed supply during the lead time*. If we can meet the 90% of customer demand on an average then service level is 90%. How do we increase the service level of the company?

The service level can be improved by keeping safety stock to meet the unexpected surge in the demand. The safety stock will definitely increase the inventory carrying cost but to meet the customer demand is more important for the company. Let us model the probabilistic demand and lead time to achieve certain service level. In our deterministic demand model the reorder point was calculated as ROP = DL which is demand during lead time.

Since there is variability in the demand or lead time we have to add certain inventory to meet the unexpected demand surge. This is called safety stock. The amount of the safety stock is directly proportional to the service level of the inventory.

Now ROP = Expected demand during lead time + Safety Stock = DL + S

The service level of 95% implies a probability in 95% of times the customer demand will not exceed supply during the lead time. For a given service level the greater the variability in either demand rate of lead time the greater the amount of the safety stock that will be needed to achieve that service level. Similarly for a given variability in demand and lead time the increasing the service level will increase the amount of the safety stock needed.

 $ROP = DL + Z\sigma_{dL}$

 σ_{dL} = The standard deviation of lead time demand.



For example to achieve 99% service level for the standard deviation of 10 quantities where demand is 100 per day and lead time is 5 days

ROP = 100*5 + 2.33*10 = 500 + 23

The company has to carry additional 23 quantity of inventory to achieve 99% service level. If the variability was more say standard deviation of demand during lead time was 40 then

ROP = 100*5 + 2.33*40 = 500 + 93

The company has to carry additional 93 quantity of inventory to achieve 99% service level.

In above example we assume that lead time is constant and demand is variable. In case when demand is constant and lead time is variable then reordering point is calculated as

 $ROP = D*L + zd\sigma_L$

 σ_L = standard deviation of the lead time in days

If both demand and lead time is variable then

 $ROP = D^*L + z^* \sqrt{(L^* \sigma_d^{\wedge 2} + d^{\wedge 2*} \sigma_L^{\wedge 2})}$

In this the demand and lead time are assume to be independent.

For example in the above example assume that standard deviation of lead time is 2 days and standard deviation of the demand is 10 quantities then at 99% service level

 $ROP = 100*5 + 2.33*\sqrt{(5*10^{2} + 100^{2}*2^{2})} = 500 + 201$

The company has to carry 201 quantities as safety stock to achieve 99% service level.

There may be many situations where the surge in the demand can be predicted beforehand – for seasonal items the demand will peak up in the season and then it subsidies when season ends. Similarly for the festivals there can be specific items which are to be promoted to drive the volume. Such demand should be accounted while actually forecasting the demand and inventory should be stocked as per the demand period. In the models we have studied we assume constant demand and standard deviation throughout the day which may not be case for many items.

The idea of this chapter was to provide basic understanding of the inventory and the service level. The inventory management is a vast subject in itself. For more depth knowledge readers are encouraged to read books related to operation and inventory.

Example of Customer Service Level

You are running a boutique in one small town. Most of the vendors are from bigger cities. One of the products which are the best-selling for your boutique has high demand variability. Sometime your stock goes out of stock before the order from vendor reach your location. As product is high selling price, you decided to do some scientific study on it. You decided to hire an intern from nearby MBA institute to work for you for a month to reduce the stock out.

He started working with you and gets some data point from your purchase order, customer order from records you maintained in your store book. As vendor is big, usually your demands quantity is always fulfilled and product is delivered to your doorstep in 7 days. The intern uses your cost structure to find the Economic Order Quantity for those products using some formula. The EOQ was coming to be 300 quantities. He looked at your last 40 days sales and found that the average order per day is 25 and standard deviation is 30. He told you that if you order 300 quantities then in 12 days your stock will be consumed but demand is highly variable so you have to keep some safety stock for the variability in the demand. So you decided to keep safety stock to cover at least 99% of the demand.



He introduce you the concept of service level. 99% service level is one in which 99% of the customer demands are fulfilled. Your regular order quantity is 300. As your lead time is 7 days the 175 quantity is the point at which you should reorder but demand as variable so you decided to keep reordering level at point where 99% of the demand is met.



 $ROP = DL + Z\sigma_{dL}$

As you know Z for 99% is 2.58, you have to find the standard deviation of the mean. Here again you use Central Limit Theorem to find the standard deviation of the mean demand.

Standard deviation=Sample Standard Deviation/ $\sqrt{(Sample Size)} = 30/\sqrt{(40)}=4.74$

The Reorder point is=demand during lead time+Safety Stock for 99% service level

=7*25+2.58*2.37=175+12=187

So as soon as your remaining stocks reach to 187 quantity you should order 300 quantity. This process has to be repeated to get 99% customer service level.



If you are ready to live with lower service level say 90% service level then you have to Reorder at Point = 175+1.64*4.74 = 175+8 = 183 quantity.

Learning from the Chapter

- Understanding of basic inventory replenishment model
- Various cost involved in the inventory decision point
- Economic Order Quantity (EOQ) model explanation and formulation
- Inventory replenishment under variable demand
- Concept and service level, safety Stock and Reorder point
- How to achieve certain service level

Section -II

TRANSPORTATION PROBLEM

7.2.1 Problem Formulation

Transportation problem is another form of resources allocation optimization in which the objective is to minimize the cost of transportation from multiple sources to multiple destinations. The class of transportation problem need not necessarily be actual transportation of a material from sources to destinations but it defines a general model with multiple sources and multiple sinks with each path having certain cost or opportunity and the optimization problem is to minimized the cost of fulfilling the demand and supply. There are definite sources and the definite destination with cost attached to the each pair of sources and destination. As shown in picture below, there are three sources S1, S2 & S3 and three Destination D1, D2 & D3. The sources can be warehouse or factory and destination can be store. Each pair of source and destination has cost associated with the transportation C_{ij} .



The amount of material transported from a source to a destination is denoted as X_{ij} . The total cost of transportation is $T_{cost} = \sum C_{ij}X_{ij}$ where i = 1 to N sources and j=1 to M destination.

In above case the total cost of transportation is given

 $T = X_{11}C_{11} + X_{12}C_{12} + X_{13}C_{13} + X_{21}C_{21} + X_{22}C_{22} + X_{23}C_{23} + X_{31}C_{31} + X_{32}C_{32} + X_{33}C_{33}$

The objective of the transportation problem is to minimize total cost T. The X_{ij} will have integer value; hence we cannot use linear programming directly. Instead we use integer programming for solving the transportation problem. The Integer programming is a special case of linear programming in which one or more variables are integer. Integer programming cannot be solved using simple graphical plot as we done in chapter I. we have to devise the method of solving the problem using simple pen and paper. Later in the chapter we will solve the problem using integer programming in excel.

Warehouse/Stores	ST1	ST2	ST3	ST4	Total Warehouse
WH1	C ₁₁	C ₁₂	C ₁₃	C ₁₄	40
	X	X ₁₂	X ₁₃	X.14	
WH2	C ₂₁	C ₂₂	C ₂₃	C_24	45
	X ₂₁	X ₂₂	X ₂₃	X.24	
WH3	C ₃₁	C ₃₂	C ₃₃	C ₃₄	25
	X.31	X.32	X.33	X	
Total Store	30	25	20	35	110

In above table there are three warehouses supplying to the four stores in the area. The capacity of warehouses and demand at the stores are matching with 110 quantities. In case where supply capacity and demand is not matching then dummy demand can be created with zero cost to balance the demand and supply for solving the problem.

7.2.2 North Western Corner Rule

Warehouse/ Stores	ST1	ST2	ST3	ST4	Total Ware- house
WH1	C ₁₁	C ₁₂	C ₁₃	C ₁₄	40
	$X_{11} = 30$	$X_{12} = 10$	$X_{13} = 0$	X ₁₄ =0	
WH2	C ₂₁	C ₂₂	C ₂₃	C ₂₄	45
	$X_{21} = 0$	X ₂₂ =15	X ₂₃ =20	X ₂₄ =10	
WH3	C ₃₁	C ₃₂	C ₃₃	C ₃₄	25
	$X_{31} = 0$	$X_{32} = 0$	X ₃₃ =0	X ₃₄ =25	
Total Store	30	25	20	35	110

There is various method of starting the allocation. The initial allocation helps in reducing the number of steps required for the solution. North Western Corner
rule is one such method of initial allocation. The rule is very simple. Start from the North and Western of the table. In this case it is X_{11} . Allocate as much as supply is available and demand is there. X_{11} is allocated 30 as supply is 40 and demand is 30, so we can allocate at max 30 in the cell. Once this cell is allocated move to right cell that is X_{12} , here supply remaining is 10 and demand is 25; so we allocate 10 to X_{12} . The supply WH₁ is exhausted, now move to next supply WH₂ that is row 2. As demand for ST₁ is exhausted, we start with ST₂. In ST₂ we have unmet demand of 15 units, so we allocate X_{22} =15 units. Demand for ST₂ is exhausted; we move to demand of ST₃. WH₂ still has 25 unallocated units, hence we allocate X_{33} =20 units. The demand for ST₃ is also exhausted; we move to demand for ST₄. WH₂ still has 5 units left, so we allocate X_{42} =5 units and move to X_{34} , allocate remaining units which is 25. The table above demand is equal to supply.

Warehouse/ Stores	ST1	ST2	ST3	ST4	Total Ware- house
WH1	$C_{11} = 10$ $X_{11} = 30$	C ₁₂ =5 X ₁₂ =10	$C_{13} = 9$ $X_{13} = 0$	$C_{14} = 13$ $X_{14} = 0$	40
WH2	$C_{21} = 4$ $X_{21} = 0$	$C_{22} = 12$ $X_{22} = 15$	$C_{23} = 6$ $X_{23} = 20$	$C_{24} = 10$ $X_{24} = 10$	45
WH3	$C_{31} = 8$ $X_{31} = 0$	$C_{32} = 3$ $X_{32} = 0$	C ₃₃ =7 X ₃₃ =0	C ₃₄ =11 X ₃₄ =25	25
Total Store	30	25	20	35	110

Let's assign cost to each pair of the Warehouse and Stores.

Total cost of the above table is T=10*30 + 5*10 + 12*15 + 6*20 + 10*5 + 11*25 = 975

The north western rule does not take cost in to account while allocating the units to each pair of supply and demand. Therefore we can improve the total cost by reallocation of the units among the pair. There are chances of getting lower cost of transportation by reallocating the units among the pair of WH and Stores. One such possibility is reallocating one unit to X_{21} from X_{11} and simultaneously allocating one unit from X_{23} to X_{13} . Decrease in cost in moving one unit from X_{11} to X_{21} is 10-4=6 and increase in the cost from moving unit from X_{23} to X_{13} is 9-6=3. Overall the cost will decrease by 6-3=3 from above unit movement. There may be many such possibilities. We cannot identify this

one by one; instead we develop method to optimize the solution. The solution is optimum when any unit movement from one pair to another increases the overall cost.

In order to find the feasible optimal solution we have to go through iteration of below steps till it is not possible to find lower overall transportation cost through reallocation.

Warehouse/ Stores	ST1	ST2	ST3	ST4	Total Warehouse	u
WH1	C ₁₁ =10	C ₁₂ =5	C ₁₃ =9	C ₁₄ =13	40	u ₁ =0
	X ₁₁ =30	X ₁₂ =10	X ₁₃ =0	X ₁₄ =0		
	D ₁₁ =	D ₁₁ =	D ₁₁ =	D ₁₁ =		
WH2	C ₂₁ =4	C ₂₂ =12	C ₂₃ =6	C ₂₄ =10	45	U ₂ =7
	X ₂₁ =0	X ₂₂ =15	X ₂₃ =20	X ₂₄ =10		
3	D ₁₁ =	D ₁₁ =	D ₁₁ =	D ₁₁ =		
WH3	C ₃₁ =8	C ₃₂ =3	C ₃₃ =7	C ₃₄ =11	25	U3=8
	X ₃₁ =0	X ₃₂ =0	X ₃₃ =0	X ₃₄ =25		
	D,,=	D ₁₁ =	D ₁₁ =	D ₁₁ =		
Total Store	30	25	20	35	110	
v	v,=10	v_=5	v,=-1	v_=3		

In the table above, set u1=0 as the initial value. For all the value where $X_{1j} > 0$ set $v_j = C_{1j} \cdot u_1$.

$$v_1 = C_{11} = u_1 = 10-0=10$$

 $v_2 = C_{21} = C_{12} - u_1 = 5-0=5$

as there are not more $X_{1j} \ge 0$, we move to second row where $X_{2j} \ge 0$ set $u_2 = C_{2j}$

$$\mathbf{v}_{2j}$$

$$u_2 = C_{22} - v_2 = 12 - 5 = 7$$

Similarly we calculate other value of v and u.

$$v3=C_{23}-u_2=6-7=-1$$

 $v4=C_{24}-u_2=10-7=3$
 $u3=C_{34}-v_4=11-3=8$

Now calculate the value S_{ij} for all cells with variable $X_{ij}=0$ in the initial allocation as

$$S_{13} = u_1 + v_3 = 0 + (-1) = -1$$

$$S_{14} = u_1 + v_4 = 0 + 3 = 3$$

$$S_{21} = u_2 + v_1 = 7 + 10 = 17$$

$$S_{31} = u_3 + v_1 = 8 + 10 = 18$$

$$S_{32} = u_3 + v_2 = 8 + 5 = 13$$

$$S_{33} = u_3 + v_3 = 8 + (-1) = 7$$

Warehouse/ Stores	ST1	ST2	ST3	ST4	Total Warehouse	u
WH1	$C_{11} = 10$ $X_{11} = 30$	C ₁₂ =5 X ₁₂ =10	C ₁₃ =9 X ₁₃ =0	C ₁₄ =13 X ₁₄ =0	40	u ₁ =0
WH2	$S_{11} = 10$ $C_{21} = 4$	$S_{12} = 5$ $C_{22} = 12$	$S_{13} = -1$ $C_{22} = 6$	$S_{14} = 3$ $C_{24} = 10$	45	U_=7
	X ₂₁ =0	X ₂₂ =15	X ₂₃ =20	$X_{24} = 10$		2
	$S_{21} = 17$	$S_{22} = 12$	S ₂₃ =6	$S_{24} = 10$		
WH3	$C_{31} = 8$	$C_{32} = 3$	C ₃₃ =7	$C_{34} = 11$	25	U ₃ =8
	X ₃₁ =0	X ₃₂ =0	X ₃₃ =0	X ₃₄ =25		
	$S_{31} = 18$	$S_{32} = 13$	S ₃₃ =7	$S_{34} = 11$		
Total Store	30	25	20	35	110	
v	v ₁ =10	$v_2 = 5$	v ₃ =-1	v ₄ =3		

Next step is to find all value where $S_{ij} > C_{ij}$, we can see following values

 $S_{21}>C_{21}$, $S_{31}>C_{31}$, $S_{32}>C_{32}$ which means we can improve the solution by increasing the value of X_{21} , X_{31} and X_{32} . However out of all such values we increase the value of X_{ij} which has maximum $S_{ij}-C_{ij}$.

$$S_{21}-C_{21}=17-4=13$$

 $S_{31}-C_{31}=18-8=10$

$$S_{32}-C_{32}=13-3=10$$

Since S_{21} - C_{21} has highest value we increase X_{21} . We increase the value of X_{21} by 30, decrease the value of X_{11} by 30

Ware- house/	ST1	ST2	ST3	ST4	Total Ware-	u
Stores	C 10			G 13	nouse	
WH1	$C_{11} = 10$	$C_{12} = 5$	C ₁₃ =9	$C_{14} = 13$	40	u ₁ =0
	X ₁₁ =0 (-30)	X ₁₂ =25 (+15)	$X_{13} = 15(+15)$	X ₁₄ =0		
	S ₁₁ =10	S ₁₂ =5	S ₁₃ =-1	S14=3		
WH2	C ₂₁ =4	C ₂₂ =12	C ₂₃ =6	$C_{24} = 10$	45	U ₂ =7
	$X_{21} = 30(+30)$	X ₂₂ =0(-15)	X ₂₃ =5(-15)	X ₂₄ =10		
	S ₂₁ =17	S ₂₂ =12	S ₂₃ =6	S ₂₄ =10		
WH3	C ₃₁ =8	$C_{32} = 3$	C ₃₃ =7	C ₃₄ =11	25	U ₃ =8
	X ₃₁ =0	X ₃₂ =0	X ₃₃ =0	X ₃₄ =25		
	S ₃₁ =18	S ₃₂ =13	S ₃₃ =7	S ₃₄ =11		
Total	30	25	20	35	110	
Store						
v	$v_1 = 10$	v ₂ =5	v ₃ =-1	v_=3		

With the new allocation the total cost is T=5*25+9*15+4*30+6*5+10*10+11*25 = 785. He we manage to reduce total cost by 975-785=190 after reallocation.

Now repeat the process till all $S_{ij} \leq C_{ij}$.

Warehouse/ Stores	ST1	ST2	ST3	ST4	Total Warehouse	u
WH1	$C_{11} = 10$ $X_{11} = 0$ $S_{11} = 7$	$C_{12} = 5$ $X_{12} = 25$ $S_{12} = 5$	$C_{13} = 9$ $X_{13} = 15$ $S_{13} = 9$	$C_{14} = 13$ $X_{14} = 0$ $S_{14} = 13$	40	u ₁ =0
WH2	$C_{21} = 4$ $X_{21} = 30$	$C_{22} = 12$ $X_{22} = 0$	$C_{23} = 6$ $X_{23} = 5$	$C_{24} = 10$ $X_{24} = 10$	45	U ₂ =-3
WH3	$S_{21} = 4$ $C_{31} = 8$ X = 0	$S_{22}=2$ $C_{32}=3$ $X_{2}=0$	$S_{23} = 6$ $C_{33} = 7$ X_{-0}	$S_{24} = 10$ $C_{34} = 11$ X_{-25}	25	U ₃ =-2
Total Store	$S_{31} = 5$ 30	$S_{32}=0$ $S_{32}=3$ 25	$A_{33} = 0$ $S_{33} = 7$ 20	$S_{34} = 11$ 35	110	
v	v ₁ =7	v ₂ =5	v ₃ =9	v ₄ =13		

From the above table we cannot find any more $S_{ij}>C_{ij}$, hence solution is optimal. Optimal cost of the transportation is 785.

7.2.3 Solving Transportation Problem using Integer Programming

The transportation problem is another form of linear programming with variables in the integer values which makes it special case of linear

programming. The basic construct and assumption of integer programming is same as general linear programming. Only difference stem from the assumption of the integer for the variables. The transportation programming we solved using North Western rule and Tables can be solved using Excel. The problem can be defined as below

Min $Z = \sum X_{ij}C_{ij}$ Total Transportation cost

Subject to

X _{1j} <=WH1	warehouse1	capacity	constraint

- $X_{2j} \le WH2$ warehouse2 capacity constraint
- $X_{3i} \le WH3$ warehouse3 capacity constraint
- X_{il}<=ST1 Store1 demand constraint
- X_{i2}<=ST2 Store2 demand constraint
- X_{i3}<=ST3 Store3 demand constraint

 $X_{ii} \ge 0$ and is integer

Adding value in equations

$$\begin{array}{ll} \text{Min} & Z=X_{11}*10+X_{12}*5+X_{13}*9+X_{14}*13+X_{21}*4+X_{22}*12+X_{23}*6+X_{24}*10+\\ X_{31}*8+X_{32}*3+X_{33}*7+_{X34}*11\\ \text{Subject to } X_{11}+X_{12}+X_{13}+X_{14}<=40\\ X_{21}+X_{22}+X_{23}+X_{24}<=45\\ X_{31}+X_{32}+X_{33}+X_{34}<=25\\ X_{11}+X_{21}+X_{31}<=30\\ X_{21}+X_{22}+X_{23}<=25\\ X_{31}+X_{32}+X_{33}<=20\\ X_{41}+X_{42}+X_{43}<=35\\ X_{ij}>=0 \text{ and is integer} \end{array}$$

The problem can be constructed in excel as shown below. Total cost is nothing but sumproduct of variables X and cost.

FUE HOA	AL MOUR	T PAGE	LAYOUT FORM	NRAS DATA	REVIEW	VOW						
Capy Stame	e di la constante di la consta	1 9 -	u + s $\equiv \cdot \Delta + A $	4 H H H H H H H H H H H	8. 5 10 10 10	Wap Test	General Sector + S + % + %	-	(F) conditional (annualizing -	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		Sales -
A		¢	0		1	6	н	1	1	ĸ	4	M
Min	z	+SUM	PRODUCT(84:	815,04:015)							
Varibles	value	Cost	Values	Capacity	_	Sien	Warehouse Usage					
X11		0 (11	10	WH1	40		0					
X12		0 C12	5	WH2	45		0					
X13		0 C13	9	WH3	25		0					
X14		0 C14	13	1								
X21	1	0 C21	4	Demand		Sign	Demand Met	2				
X22		0 C22	12	ST1	30		0					
X23		0 (23	6	ST2	25		0					
X24		0 (24	10	ST3	20	-	0					
X31		0 (31	8	ST4	35	-	0					
X32		0 C32	3									
X33		0 C33	7									
X34		0 C34	11									

Warehouse capacity constraint and the demand from the stores are added as shown below. WH1 should be equal to $X_{11}+X_{12}+X_{13}+X_{14}$. Similarly add demand constraints.

q	18.2.	6.1	in the data	and the	e nel torretto	-		Seager Programming - E	lacal .						
	LE HO	ME INSOR	T PAGELA	YOUT YOR	AULAS DAT	A REVEN	VEW								
10.0	Contraction	eturia I	albe A A A -	v = X v + X = A	2 ¹⁰ = 1 1 10 1	4-1 2 42 40 0	i Wag Ted Warge de	Consul Content - S - No. + 12		El International		dal.	Pro la	E Anti E Cos	2
3			6	0		1	6	н	1.1	1	×	4	M	N	0
1	Min	z	1	0											
2			_												
2	Varibles	value	Cost	Values	Capacity		Sign	Warehouse Usage							
4	X11		0 C11	10	WH1	40		+84+85+86+87							
5	X12		0 C12	5	WH2	45		0							
6	X13	1	0 C13	9	WH3	25		0							
7	X14	1	0 C14	13		1		King there is	1						
4	X21		0 (21	4	Demand	-	Sign	Demand Met							
9	X22		0 C22	12	ST1	30	-	0							
10	X23		0 (23	6	512	25	-	0							
11	X24	1	0 (24	10	\$13	20	-	0							
12	X31	1	0 (31	8	ST4	35		0							
13	X32		0 C32	3		01.0		1°							
14	X33		0 (33	7											
15	X34	. N	0 C34	11											
16	1	11-12		1											
17															
18															
19															
20															
21															
22															

In Solver select Min and select D1 cell as objective function.

80 H	4.6.1.2	1.14					Integer Pro	gramming - Microsoft Excel					- 0
1000	Note in	out hope	Layout	Parinulai	244 1	teres (as	Developer Add	-9%					+ Q = #
181	Co Law	an Other	-	Co fin	Contraction Theorem in Film (1990) Marchine	14 III 14 III	Y Kon tar Schentl Abbe	The second secon	De litte	No N N	Sam para	Sata analysis Analysis	
-	_	• • •	A		_							_	_
1	Mat	2	0			0 .			54	Aver Parameters			1
2	variation	Value	CHI	Natves	Capacity	-	Warehouse Usa	2 Sections	1011			8	
- B	611	-	CU .		Den 1			1		Contact II			
2	101		C13	1	NAND	20.4	-			Charles Com			
7	834	1	C14	1	13	_		Er Danging Variable	Onter			-	
	821	1.1	CR		e Demand	541	Demand Met	KIN KUI				96	
	122	1 24	532		2971	10-+		Sport to the Constr	-				
30	123		100	-	4972	27 **		9					
	104	-	C24	-	10575	200	-						
-	102	+	10 m	+	1		<u> </u>	3			Seb.		
34	100		C10	-	1						(prime)		
13	134	1 1	C14	1	18								
38	- 11	10.00	1	(T)	S						great all		
3.7													
<u>0</u> -								Winterman	Country No. 1		- Periore		
-								a set out the					
21								Egent a Salary Terra		whee (T)	ofers		
22								Stinghted					
33 24 25 3								Salact the GHC hard engine for lease fait non-second,	ner regns fo late a halana, avtai	n haldens fluit are presidentials at the Dolutionary angles for Sa	en John for V Inc	~	
D	Sheet3 / S	head 2 - Sha	10.00							. pte			
Bolat 1	2												

Add constraints related to Warehouse and Stores. Additionally add integer constraint and non-negative constraints.

юн «	9 - (* -) Mana	Delle Teat - A	or invest	funda	Des 1		-	integer Pro Developer Aut	ogrammin e m	g - Microsoft Escel						
181		Anna Ottor	tantra .	Co fi	Contractions Pringentines • East (onlice restlance	11 11		W Kons T Schoold Schoold	Sala Sala Sala	HI III Contra	100 100 100 100 100 100 100 100 100 100	No No	<u>레</u> :	tin brist	S Server S Serve Analysis Analysis	
-	00	• (*	<i>s</i> .	1.0									000			
1 4	Man	2		3		0		1	1		5	olver Parameters	37			1
3 4 5 6 7	1941 102 103 104	nes varue	0 C11 0 C12 0 C13 0 C13 0 C14	Norves 2 1	Caju(ity Iniki) ViskQ ViskQ	40 40 20	÷	Warehouse Us		Sig Dyantan Na Ogan Bi Danyog tanata Sala		Optech	2			
5	121		ecn ecl2		Demand (M)	10	Ser.	Demand Met		Agent is the Constraints						
11	134	+	ecu ecu	-	373	25				C14 (C13 + mage C14 (C13 + e) C14 (C13 + e) C14 + C13 + e) C14 + C14 + e) C14 + e) C				garge Darge		
54 35 36	233) 134		0 C33	1						500 - 550 500 - 550 500 - 550 500 - 550				Quinta Grant A		
17										(Call and the Call	-		-	Lotte	•	
26 21										Agent a Saling Period		land and a second		Lightere		
22 23 24 25 25										Salang Nethod Salad Ne (AC Northwar angle for Innor Salar In Northward).	22,125	e hallens hal en er kit fie boldener e	pt e tec	Select the UP 1 publicity that is	-	
27	Sheet).	See 2.2	Net3 / 5	2					11	99			jate .		-	31

On solving the D1 value is 785. Hence optimum transportation cost is 785. This number is same as we got in the previous section.

085	Ca. 1	5.221	22 22		201	22.5	Integer Programming - Ex	wi i				
Access From Access From Viels From Test	From Other Sources -	Easting Connections	Constant Canada	vectors [] peters [] rLong [] ors		The Court	Tatto Flack Anno Columns Fill Depks	e Deta	Consolider Consolider	Rhat # Road	8	inte Ungoing
A		c	0	t	1	- 0 -	н	1	¥. K	. L	м	N
1 Min	2	785										
Varibles	value	Cost	Values	Capacity		Sien	Warehouse Usage					
4 X11	0	C11	10	WHI	40		40					
5 X12	23	C12	5	WH2	45		45					
6 X13	17	(13	9	WH3	25		25					
7 X14	0	C14	13									
1 X21	30	(21	4	Demand		Sign	Demand Met					
9 X22	0	(22	12	STL	30		30					
10 X23	3	C23	6	512	25		25					
11 X24	12	C24	10	ST3	20		20					
12 X31	0	C31	8	ST4	35		35					
13 832	2	C32	3									
14 X33	0	C33	7									
15 X34	23	C34	11									
16		-	1	1								
17												
18												
19												
20												
21												
22												

This is a simple example of the transportation. The real life problems can have many more sources and destination. Also it can have many more constraints. However the basic premises remain same and same process can be used to solve complex problems.

Learning from the Chapter

- Understanding of Transportation problem
- Learn initial allocation method known as North Western Corner Rule
- How to solve transportation problem using pen and paper
- How to formulate transportation problem as a variant of Linear Programming called Integer Programming

• How to solve transportation problem using Integer Programming in Excel

Chapter - VIII ADVANCED TOPICS

"By 2025, we can expect the world to be completely digital. Paper books will be a thing of the past. Education will be delivered through analytics-based assessment tools and adaptive learning platforms." -By Osman Rashid

Analytics is a vast subject; hence no one book can cover all aspect of the analytics; this book is no exception. I have tried adding as many topics as possible in the single book but pages were running out. As my focus was mostly on the retail and ecommerce side I have left out many aspect of Analytics. Wish I could add some of the topics in these chapters but by themselves these topics are very big which will not be one or two chapter but a book in itself. The idea of adding them in to this book is to provide a glimpse of emerging fields in the analytics. The readers can explore these topics through net or other authoritative book in the respective fields. I have given the general idea and direction to be followed.

8.1 Game Theory

Game theory gets much needed attention among the general population after the Hollywood movie the beautiful mind based on life of John Nash was released and it gets Oscar awards for the best movie category. The game theory as a subject has been very much there in economics and management circle before that. The game theory as a subject start with the Jon Von Neumann book Theory of game and Economic behavior and later developed by John Nash the theory called Nash Equilibrium.

A game theory is concern with the interaction of players in a situation where competitive activities are carried out to maximize their returns. The game theory has wide application in the fields of economics, management, biology, political science and poker. The application of the game theory in management is emerging as the important fields due to high applicability and the tangible results from the activities. The companies use game theory for its pricing decision, negotiation with vendors, negotiation with its rivals, potential bidders in auction etc. In a very simple term, a game or a strategic game can be described as a competitive activity in which there are set of players; each player has set of actions and each action lead to some kind of payoff to the players. An assumption is that the players makes rational choices like if action A provides a profit of 10 and action B provides profit of 5 then we expect rational being to prefer action A over action B. Basically rational being is one who makes decision based on optimal solution; not from ego or emotion. If she prefers A over B and B over C then she is expected to prefer A over C. Now payoff can be a salary hike, a profit, a satisfaction or utility in economics term or winning auction bidding etc. etc.

		Prisoner B				
		Non Confess	Confess			
Prisoner A	Not Confess	1,1	5,0			
	Confess	0,5	3,3			

One of the most widely used examples of game theory is Prisoner's dilemma. In these game two prisoners A and B have two choices confess and not confess. The payoff for both is shown in table. If both A & B confess then 3 years jail term to each of them, if both do not confess then 1 year jail term to each and if one of them confess and another do not confess then 0 years term for one who confess and 5 years term for one who doesn't confess. If you are in same situation with one of your best friends; what would you do? Confess or do not confess!!!

Obviously being inspired by Bollywood movies like Sholay you would sacrifice and do not confess even if your friend crack under pressure. But there is no room for emotion in the rational choices. In game theory there is no room for emotion; it's all about rational choices. You have two conditions here –

1. Both A&B are allowed to consult; such games are called cooperative game and

2. There is no room for consultation (non-cooperative game).

In the first case obviously both of you would decide not to confess. Right? But can you trust your friend completely? What if he takes advantage of you? You are in dilemma now. Right? In second case both of you are not allowed to consult then assuming he doesn't confess you tend to confess because you will be free man. In case he confess, you will also confess because that way you will get lesser term (3 years) than that of being not confess (5 years). Similarly he will come to same conclusion and act based on your likely action. In non-cooperative games both would confess because whatever decisions other person makes I am better off by confessing.

Let us dig bit deeper. Put yourself in prisoner A's shoe and decide whether to confess or not. What if B confesses to get lighter punishment then A has to confess to reduce the potential punishment. In case B does not confess then A would be better off by confessing and get away without any punishment. Here A will confess irrespective of what B does. That means confess is the **dominant strategy** for A.

		Prisoner B	
Prisoner A		Non Confess	Confess
	Not Confess	1,1	5,0
	Confess	0,5	3,3

What about B's thinking process? B would also do the same. B also has confessed as dominant strategy that is whatever A does B is better off by confessing. Hence both prisoner would confess and get 3 years imprisonment each.

What if A and B doesn't have dominant strategy. Assuming that the police has no way of knowing A or B was involved in crime unless one of them confess then there is no punishment if both do not confess during interrogation and hence punishment is 0 years of imprisonment. If one of them confess and another do not confess then 1 year imprisonment for whoever confess and 5 years imprisonment for one who do not confess. In case both confesses then 3 year's imprisonment for both prisoners.

	Prisoner B	
Prisoner A	Non Confess	Confess

Not Confess	0,0	5,1
Confess	1,5	3,3

What would A and B do now? If B confess then A should confess and if B do not confess then A is better off not to confess. Similarly for B confess if A confesses and do not confess if A does not confess. Now this game has two strategy (confess, confess) and (non confess, non confess) which are stable. A will do his best by confess/non confess given what B is doing and B will do his best by confess or non confess given what A is doing.

What is difference between both games? In the first game A and B is doing its best irrespective of what another is doing whereas in second game A and B doing it's best given what another is doing. In the first game there is **dominant strategy** whereas in second case there is no **dominant strategy**.

The stable strategies like (confess, confess) and (non confess, non confess) in second game is known as **Nash Equilibrium**; named after John Nash. In Nash equilibrium each player is doing the best it can give the action of its competitor. There is no incentive for either to deviate from the chosen strategy. For example in second game if A confess then B would confess and if B confess then A would confess; both of better off by confessing if either of them confess. Similarly for non-confess both are better off if either of them do not confess.

Is there Nash Equilibrium in first game? Yes (confess, confess) is Nash Equilibrium. In face all dominant strategy is Nash Equilibrium but reverse is not true for all cases.

In below payoff matrix, company X and company Y producing certain commodity in a same market is feeling the pressure of raising input cost. Price increase is must but if one of them increase and another do not then one who increases the prices loses market and one who hold the price gain more. If both increase the price then both are better off. If both competitors are allow to collude then increase price the best option but such behavior is against the law of perfect competition in the market. Therefore (hold, hold) is dominant strategy for both company and is the Nash Equilibrium.

Payoff	Company Y	

Company X		Hold Price	Increase Price
	Hold Price	10, 8	25, 4
	Increase Price	4, 20	15, 12

What if the game is repeated many times? Both companies are going concern and have many products in the market on which they compete, therefore we can assume game will be repeated infinite times. If X increases the price and Y hold then X is worse off. Next time X would not increase the price, in the process both will be worse off. In the long run gain both can makes from increase the price will out weight the potential going from the undercutting the prices, so both companies will increase the prices.

What if game is not infinite but only N times? For Nth game X will think that he can undercut Y because there is no game left after N. Same strategy will followed by Y. That left both with (hold, hold) strategy. On N-1th game X will think that anyway Y will undercut me in N game I will undercut him in N-1th game. Same strategy will be followed by Y and hence (hold, hold) in N-1th game. For each preceding games same will hold and ultimately (hold, hold) is the strategy followed by both companies for N number of games.

Above examples are illustrations of the few popular game theory examples. Here you have learnt about the game theory, Nash Equilibrium and concept of Dominant Strategy. This section is to provide you introduction to game theory. I would suggest reader interested in the topic to read more through specific books in the field.

Well why you should know game theory and where you are likely to use it? Game theory is one of the emerging fields in the economics and management. It is moving from the library to the corporate board room and is driven by numbers. One of the simplest uses of game theory is that I help you structure the problem in familiar terrain and then you can look for solution. Just think of the two roadside vendor selling fruits, how they position themselves in term of location and compete on price; try structure the problem in term of 2x2 matrix like that of prisoner's dilemma. When you are negotiating with potential employer, negotiation with vendor, pricing of products with respect to competitor, decision to introduce products in the market, decision to add manufacturing capacity, decision to start a new business and so on has some

element of game in it. How well one use it depend on how well problems are structure and how much data is available.

8.2 Big Data

Big data is one of the most talked about term in the data analytics fields. The explosion of data in the last few years has spurned the need for the big data platform for managing the volume which cannot be handled by the traditional database and software technologies.

According to Webopedia, Big Data is a phrase used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity.

Big data can be described by the following characteristics

Volume: The quantity of generated and stored data. The size of the data determines the value and potential insight and whether it can actually be considered big data or not.

Variety: The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

Velocity: In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

Variability: Inconsistency of the data set can hamper processes to handle and manage it.

Veracity: The quality of captured data can vary greatly, affecting accurate analysis.

The growth of Digital data in the last two 3 decades as seen from the picture below show that the explosion of the data started somewhere in 2002 and its expanding at exponential phase.



Every organization collect digital data at each tough point for better decision making. The data sources of competitive advantages but managing the volume of data is challenges for the companies. The traditional databases and software are not equipped to handle the volume and speed of the data generation and analysis required. Hence big data is the answer for these companies. There is no clearly defined boundary for a data set to qualify as the big data. Big data is just not about the volume but the technology.

There are many tools and techniques related to the big data. For this book we will discussed the most popular and open source big technology called **hadoop**.

Apache Hadoop is an open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common and should be automatically handled by the framework. Hadoop has many components



Important components of Hadoop Ecosystems are:

HDFS (Hadoop Distributed File System)

HDFS, the Hadoop Distributed File System, is a distributed file system designed to hold very large amounts of data (terabytes or even petabytes), and provide high-throughput access to this information. Files are stored in a redundant fashion across multiple machines to ensure their durability to failure and high availability to very parallel applications. This module introduces the design of this distributed file system and instructions on how to operate it.

The primary objective of HDFS is to store data reliably even in the presence of failures including NameNode failures, DataNode failures and network partitions. The NameNode is a single point of failure for the HDFS cluster and a DataNode stores data in the Hadoop file management system.

HDFS uses a master/slave architecture in which one device (the master) controls one or more other devices (the slaves). The HDFS cluster consists of a single NameNode and a master server manages the file system namespace and regulates access to files.



MapReduce Framework (YARN)

Part of the core Hadoop project, YARN is the architectural center of Hadoop that allows multiple data processing engines such as interactive SQL, real-time streaming, data science and batch processing to handle data stored in a single platform, unlocking an entirely new approach to analytics.

HBASE

Apache HBase is an open source NoSQL database that provides real-time read/write access to those large datasets.

HBase scales linearly to handle huge data sets with billions of rows and millions of columns, and it easily combines data sources that use a wide variety of different structures and schemas. HBase is natively integrated with Hadoop and works seamlessly alongside other data access engines through YARN.

Hive

Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis

Pig

Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called Pig Latin. Pig can execute its Hadoop jobs in MapReduce, Apache Tez, or Apache Spark.

Mahout

Mahout is a library of scalable machine-learning algorithms, implemented on top of Apache Hadoop and using the MapReduce paradigm. Machine learning is a discipline of artificial intelligence focused on enabling machines to learn without being explicitly programmed, and it is commonly used to improve future performance based on previous outcomes.

Once big data is stored on the Hadoop Distributed File System (HDFS), Mahout provides the data science tools to automatically find meaningful patterns in those big data sets. The Apache Mahout project aims to make it faster and easier to turn big data into big information.

Oozie

Apache Oozie is a server-based workflow scheduling system to manage Hadoop jobs

Zookeeper

ZooKeeper is an open source Apache project that provides a centralized infrastructure and services that enable synchronization across a cluster. ZooKeeper maintains common objects needed in large cluster environments. ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services . In case of any partial failure clients can connect to any node and be assured that they will receive the correct, up-to-date information.

Sqoop

Apache Sqoop is a tool designed for bulk data transfers between relational databases and Hadoop. It is use for import and export to and from HDFS to relational database. From and to Hive/HBase to relational database.

Flume

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.

Big data is real and in application in almost all fields. The data volume has moved beyond conventional database technology, there is no looking back. Remember that big data doesn't provide you new insight or algorithm, it is just another way of manage unmanageable data. So whatever you have learned in the book will stand for big data as well. For example you have learnt Market Basket Analysis (MBA). The algorithm and insight from analysis will remain same whether you use hadoop or MySQL as backend; there difference will be the scale and speed at which big data can produce the result for your data. This is true for any analysis. In term of analytics flexibility big data has to catch up the conventional database as it doesn't provide much flexibility; however it is matter of time before big data technologies is simplified to the extent that users will not really care what technology is behind the data.

8.3 Machine Learning

Another hot topic in the analytics is the Machine Learning. There is not much of difference between Machine Learning, Statistical Modeling and Data Mining. The machine learning is concern with algorithm that transforms information into actionable insight. The machine learning usually is concerned with the known algorithm whereas task like data mining deals with finding unknown pattern in the data.

Some of the common usage of machine learning algorithm is

- 1. Predict the outcomes of elections
- 2. Identify and filter spam messages from e-mail for criminal activity
- 3. Automate traffic signals according to road conditions
- 4. Produce financial estimates of storms and natural disasters
- 5. Examine customer churn
- 6. Create auto-piloting planes and auto-driving cars
- 7. Identify individuals with the capacity to donate Target advertising to specific types of consumers

The machine learning typically involves training of the data set to provide a framework for generalization. The training of the data is updated with the additional data points. For example in spam filter the algorithm will use existing known pattern of spam to create a filter. The training of the existing data will provide a generalization such as word contains Viagra is spam, email from xyg@mmm.com is a spam and so on. Whenever new email comes the criteria (generalized framework) is applied to the email and classifies them as spam or non-spam. As spammer as new technique of overcoming those filters so algorithm has to relearn from the additional data points to update the generalized framework for filtering.

Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available to a learning system. These are [11]

- **Supervised learning**: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
- Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).
- **Reinforcement learning**: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal. Another example is learning to play a game by playing against an opponent.

Some of the commonly used machines Learning Algorithms are

- Nearest neighbor classification
- Bayes classification
- Neural networks
- Vector Machine
- Association Rules
- Logistics regression

We will not go into the specific of the algorithm. Some of the algorithm has been discussed in the previous chapters. Of the above algorithms Neural networks is very different from others and you might require much more understanding of the system before actually using it in real world.

8.4 Pricing

I would have loved to add chapter on pricing but the scope and space required for pricing theory is too big to be accommodated. In this section I would like to give brief understanding of the theory behind the pricing and then provide you the direction that one can follow to gain more insight. Some of you would have already done microeconomics in your academic curriculum. To have better understanding of the fundamental of pricing I will start with very basic experiment

Assume that you are selling apple at a busy junction of a city in a movable mini truck. There are enough people moving around in the junction from morning to evening. You open shop from 8 am to 9 pm every day. One fine day you decided to experiment the buying behavior of the customers. Here we assume all buyers are rational human being as per economic. All 13 hour see almost same type of customer and same number of potential customer. You started with high price and keep on dropping prices Rs. 5 per Kg each hour. You note down the sales per hours in Kg. You plot the price vs demand in an sheet of paper.

Hour	Price per Kg	Demand in Kg
8-9	100	4
9-10	95	7
10-11	90	12
11-12	85	20
12-13	80	28
13-14	75	35
14-15	70	40
15-16	65	60
16-17	60	75
17-18	55	92
18-19	50	116
19-20	45	120
20-21	40	150



As we can see from the above graph the demand increases with the decrease in the price of the apple. It is not a straight line. But if we plot all the demand and price for the entire market then line will be straight.

Nearby same market there is another experiment being carried out by your friends who has called few seller in a single location and run an auction. The auction starts with Rs. 40 per Kg of apple. At each incremental Rs. 5 the seller sells their apple to the willing buyers. The schedule of price and supply is as shown below

Price	Supply
100	156
95	132
90	113
85	97
80	80
75	65
70	40
65	34
60	23





As you can see from above scatter plot the seller is willing to sell more and more quantity as the auction price is increases. The line is not straight but if we plot average of all the seller in the market then this line will be straight.

Now if we merge all seller and buyers in the market together then there will be one single pot as shown below

Price	Demand	Supply
100	4	156
95	7	132
90	12	113
85	20	97
80	28	80
75	35	65
70	40	40

65	60	34
60	75	23
55	92	18
50	116	13
45	120	7
40	150	3



Demand equation is quantity = 227 - 2.4xSupply Equation = -118.77 + 2.55xBoth equation will intersect at Demand = Supply 227-2.4x = -118.77 + 2.55x2.55x+2.4x = 227 + 118.775x=346X = 346/5 = 69

The quantity at which demand and supply curve intersect is known as equilibrium price. This is the prevailing price in the perfectly competitive market. We will not go to nitty-gritty of perfectly competitive market, rather we will look into the application of the demand and supply curve.

Elasticity of Demand and Supply

The understanding of the demand schedule like above will help us understand the elasticity of the demand. For any pricing decision the price elasticity is an important consideration because there is tradeoff between margins from the unit sales versus overall margin from the volume of sales. The price elasticity also help in estimating discount or promotional offers required to attain certain volume of sales.

In simple term elasticity of demand is that change in the quantity demand with change in the price.

Demand Elasticity = % change in Demand Quantity / % Change in Price.

In above example if we want to find the elasticity of demand we can look at two price range and change in demand. When we changed the price from 75 to 80 then there was decrease in 7 unit sales.

So Price Elasticity of Demand = (-7/28) / (5/80) = -4.

If absolute value of elasticity is >1 then demand is elastic, <1 then demand is price inelastic.

Similarly price elasticity of supply is calculated as change in supply quantity with change in the price.

Price Discrimination

Another interesting derivative from the demand and supply schedule is the price discrimination. To understand the concept let us assume that there are one big room and 13 successive doors. The first door has Rs 100 per kg of apple; second door has Rs. 95 per Kg and so on till 13th door has Rs 40 per Kg of apple. In the room you have good number of people. None of the customer knows there are 13 doors. You instruct them to go through the door and let them buy apple. Those who bought in a particular door he will exit before next door and never interact with other potential customers.

In the first door you sell apple who think Rs 100 is work per Kg, in second door you sell apple to those who think Rs 90 per Kg is worth and so on till 13th door. So what you have just done is that you have sold apple at maximum possible price to each customer. This price is called reservation price. Not if you look at the market equilibrium price is Rs. 69, so you have sold at Rs 31 higher than market price to the customers who bought at Rs 100. This difference between customer reservation price and market price is called consumer surplus because when that customer goes to market even if he is willing to pay Rs 100 he just need to pay Rs. 69 due to prevailing market price so he earn surplus of Rs 31.

The price discrimination is way of extracting maximum customer surplus. The real world it is really difficult to get perfect price discrimination due to information being available to all consumers at same time. However it is always possible to attain imperfect price discrimination and extract some surplus from the customer.

There are three kind of price discrimination

- 1. First Degree Price discrimination: The pricing policy in which each customer are charged their reservation price thereby extracting maximum consumer surplus. In reality we do not know the reservation price of each customer so perfect price discrimination is out of question. We can charge different price to different customer based on their paying capacity like doctor charging different price to affluent and poor customers.
- Second Degree Price Discrimination: The pricing policy in which customer are charged different price per unit for same goods or services. Example is selling one unit of shirt at Rs 2000, bundle of two units at Rs. 1800, three units at 2400 etc.
- 3. Third Degree Price Discrimination: The process of creating different segment of customer and selling same product or services at different price. For example same vodka is sold at different price by applying different label. The quality is same but you can sell one label at higher price by making it exclusive and by branding. This is very common. For example same vendor manufacture same shirt for different brand, the label attached to the shirt from two different brands creates the price different rather than quality of the product.

There are many other pricing strategy which are commonly employed

1. **Inter-temporal Price Discrimination**: This is process of dividing consumers into high demand and low demand. Sell product at high price to the high demand customers and later sell same product at lower price to the low demand consumers. For example the mobile phone are sold at high price during introduction, later after months prices are reduced in stepwise fashion to capture all segments.

- 2. **Peak Load Pricing**: This is used for those products and services which have demand at peak time and lower demand later. For example Cinema hall will charge higher price per ticket during the weekend and that too in the evening as it is the peak hours for Cine-goers.
- 3. **Two-Part Tariff**: This is way of charging minimum entry fees and then extract from customer as per their ability to spend. For example some amusement parks have basic entry fees which are common across. Once the customer enters then they charge additional fees for each ride. This practice is followed in many bars as well.

The intention of this section was to provide you a basic understanding of pricing decision and to create an interest in you to explore this subject further. The pricing being one of the most important decision in the corporate world it is imperative for all of us to understand the theoretical basis of the pricing. I have just scratched a very thin layer on pricing in this section; the readers are encouraged to read further.

Table of Contents

Title	2
Copyright	3
Dedication	4
About the Author	5
Preface	8
Acknowledgement	11
Chapter - I : DATA ANALYTICS FOUNDATION	12
Section - I : MEASURING CENTRAL TENDENCY AND DISPERSION	16
Section - II : PROBABILITY THEORY	35
Section - III : SAMPLING AND HYPOTHESIS TESTING	56
Section - IV : LINEAR PROGRAMMING	78
Chapter - II : ANALYTICS SYSTEM	92
Section - I : BUSINESS INTELLIGENCE SYSTEM	95
Section - II : R BASICS	153
Chapter - III : WEB ANALYTICS	168
Section - I : GOOGLE ANALYTICS	170
Chapter - IV : CUSTOMER ANALYTICS	200
Section - I : CUSTOMER ANALYTICS	203
Chapter - V : DIGITAL MARKETING	240
Section - I : DIGITAL MARKETING BASIC	242
Section- II : DIGITAL CHANNEL OPTIMIZATION	262
Chapter - VI : FORECASTING AND PREDICTION	280
Section - I : REGRESSION	282
Section - II : TIME SERIES FORECASTING	295
Chapter - VII : INVENTORY MANAGEMENT	313
Section - I : INVENTORY MODEL	315
Section -II : TRANSPORTATION PROBLEM	323
Chapter - VIII : ADVANCED TOPICS	333