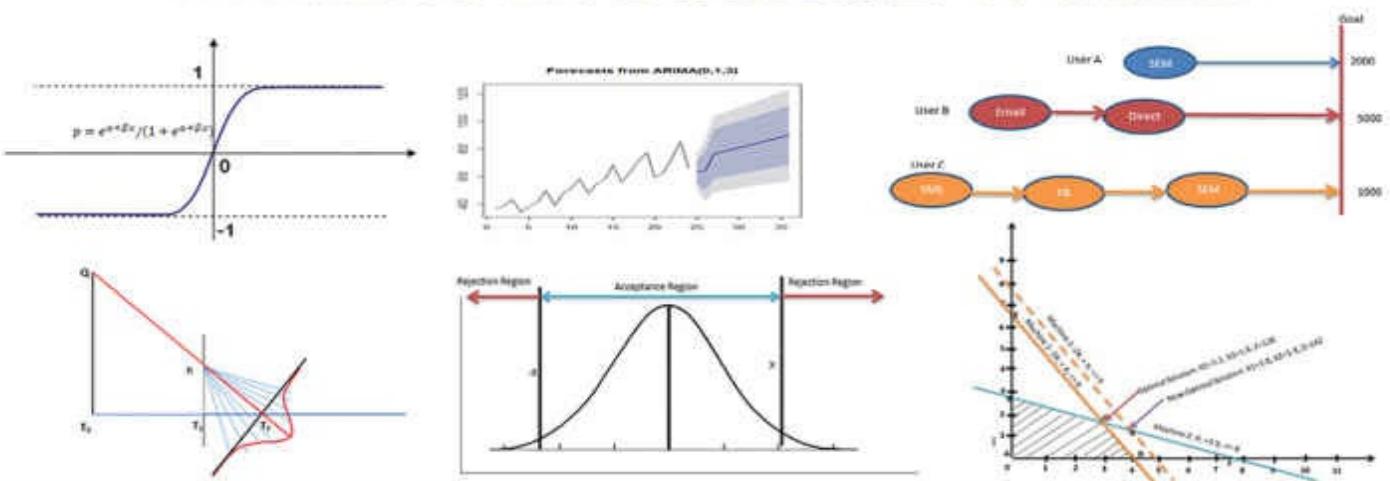


## DATA DRIVEN DECISION MAKING IN DIGITAL WORLD



JUMIN KAMKI

# DIGITAL ANALYTICS

DATA DRIVEN DECISION  
MAKING IN DIGITAL WORLD

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 error-free 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**

To

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](http://www.juminkamki.com)

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

# **CONTENTS**

Preface

Acknowledgement

## **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  $\mu = [x_1 + x_2 + \dots + x_n]/N$ . Mean is commonly denoted by  $\mu$  and  $x_1, x_2, \dots, x_n$  are set of  $N$  numbers.

$$\text{Simple } \mu = \sum D_t / N$$

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  $\mu = [22+30+35+45]/4=132/4=33$  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  $\mu = [w_1 * x_1 + w_2 * x_2 + \dots + w_n * x_n] / [w_1 + w_2 + \dots + w_n]$ . 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

$$x' = [0.35 * 22 + 0.30 * 30 + 0.20 * 35 + 0.15 * 45] / [0.35 + 0.30 + 0.20 + 0.15] = 30.45 \text{ years.}$$

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

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

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 in 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

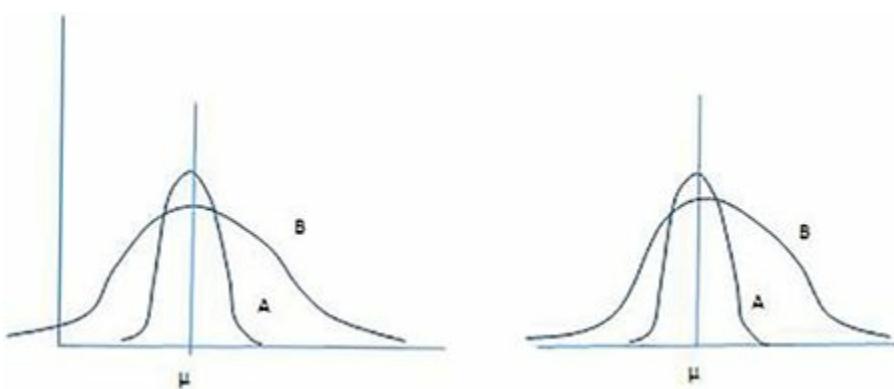
Number 123 is the mode as occurs higher number of times in the data set.

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

$$\sigma^2 = \sum(\mu-x)/N$$

Where  $\sigma^2$  = is the variance of the population

$\mu$  is the mean of the population

x is the data points

N is the number of member in the population

For example

5, 12, 15, 20, 3, 6, 8, 24, 34, 32, 25, 16

$$\mu = (5+12+15+20+3+6+8+24+34+32+25+16)/12 = 14$$

$$\sigma^2 = [(14-5)^2 + (14-12)^2 + (14-15)^2 + (14-20)^2 + (14-3)^2 + (14-6)^2 + (14-8)^2 + (14-24)^2 + (14-34)^2 + (14-32)^2 + (14-25)^2 + (14-16)^2]/12$$

$$= (81+4+1+36+121+64+36+100$$

$$+400+324+121+4)/12 = 1292/12$$

$$= 107.66$$

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.

$$\sigma = \sqrt{(\sum(\mu-x)/N)}$$

For above example the standard deviation is  $\sigma = 10.37$

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  $\rho$  and is calculated as

$$\rho(X,Y) = \text{covariance}(X,Y)/[\sigma_x\sigma_y] = E[(X-\mu_x)(Y-\mu_y)]/(\sigma_x\sigma_y)$$

where  $\mu_x$ =mean of the data set X

$\mu_y$  =mean of the data set Y

$\sigma_x$ =standard deviation of data X

$\sigma_y$  =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.

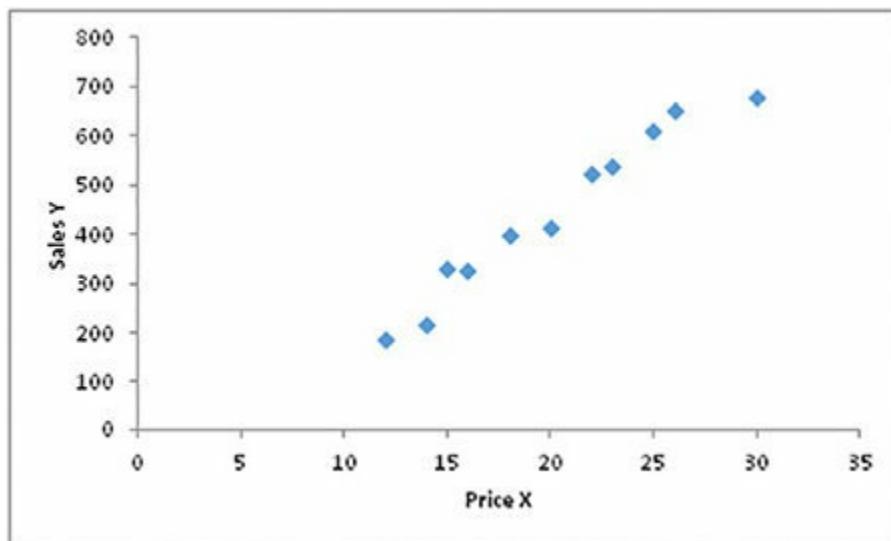
The correlation of more than 0.5 or less than -0.5 is generally considered to be strong relationship between X and Y for positive and negative respectively. The correlation between -0.5 to 0.5 is generally considered to be weak between two set of numbers.

Price (X)	Sales (Y)	X-ux	Y-uy	(X-ux) <sup>2</sup>	(Y-uy) <sup>2</sup>	(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					

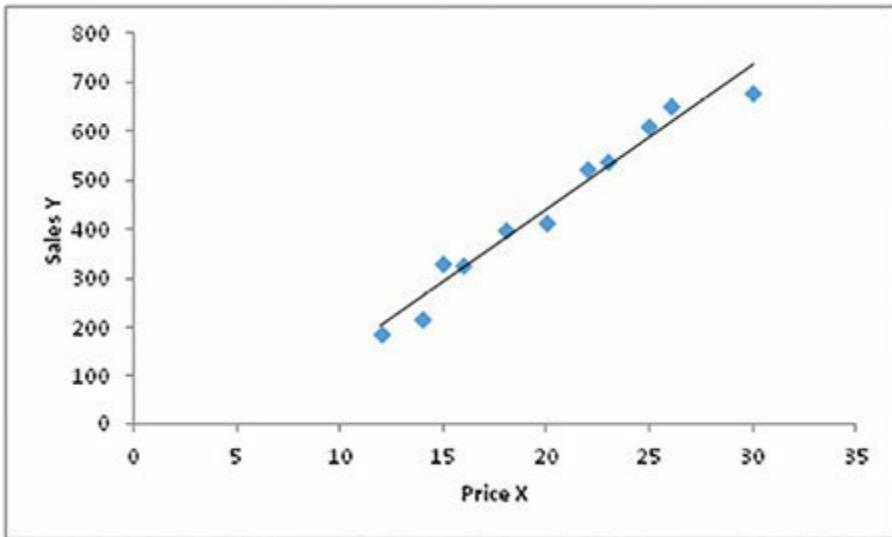
$$r_{xy} = 9440.91 / [\sqrt{318.91} * \sqrt{289982.91}] = 0.98.$$

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 line 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

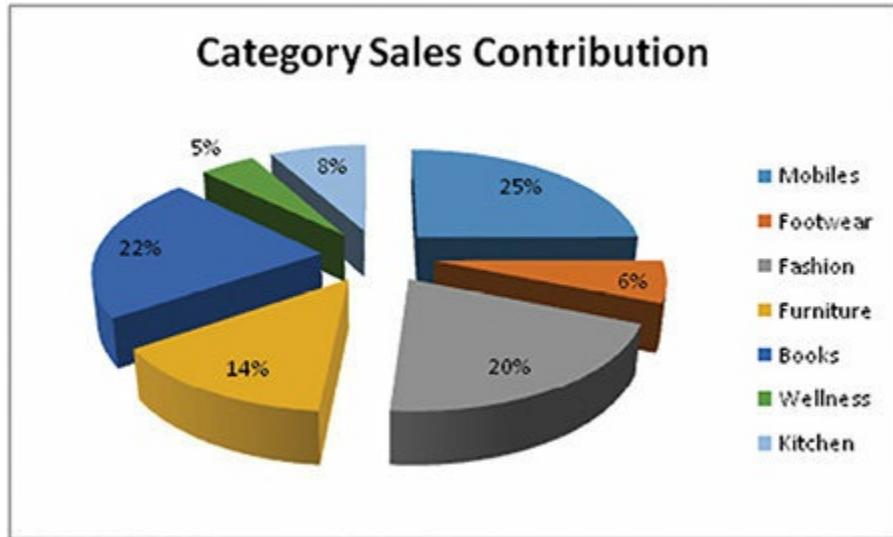
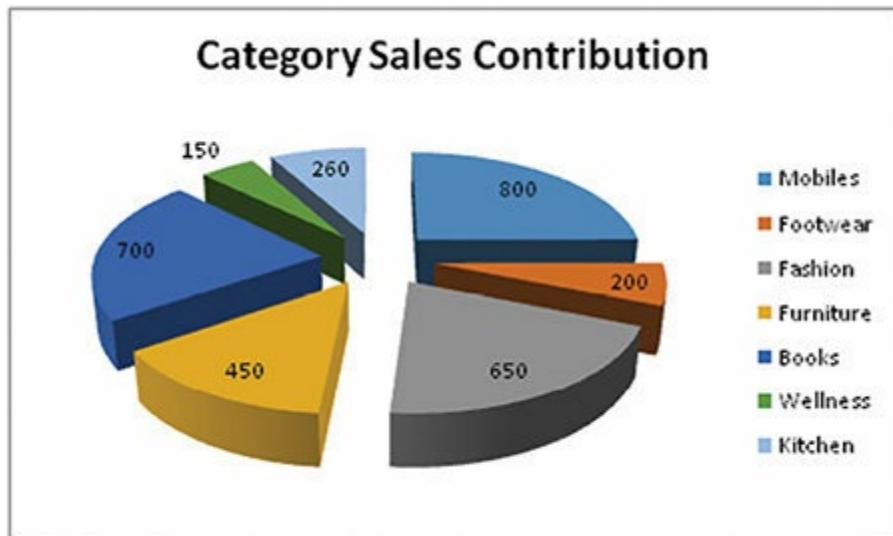


We can add the data label in the bars to clearly indicate the number to the users. This provides additional information for interpretation of the data.



## 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 make out the growth and decline in the sales quantity each day.



## Scatter Plot

The scatter plot is generally used to see the relationship between two data sets. For example you may want to see the linear relationship between data. The scatter plot also helps 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

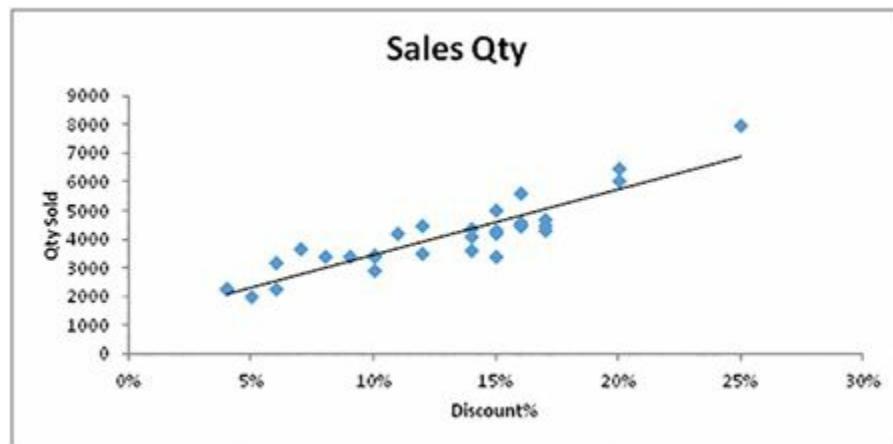
8000

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.



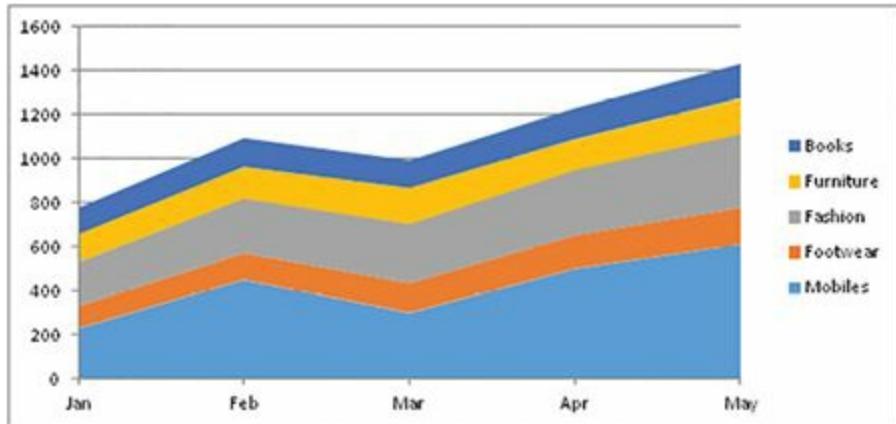
In below graph I plotted linear trend line in the graphs.



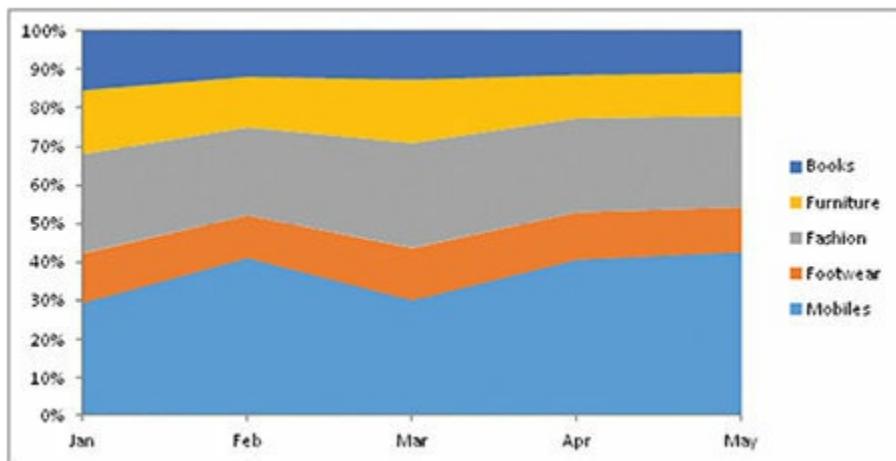
## Area Graphs

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

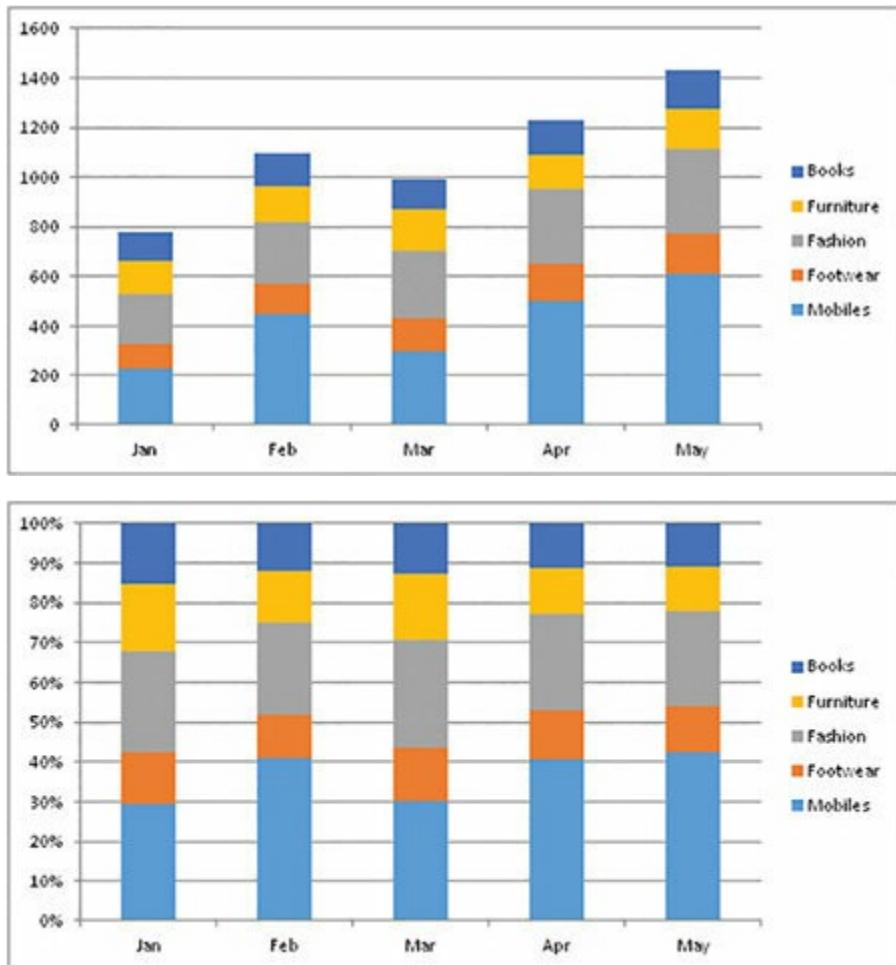


Another area graphs representation as shown in the next graphs is the relative contribution of each category for the period. It shown percentage contribution of each category to the sales. Both graphs have different utility based on requirement.



## Stacked Graphs

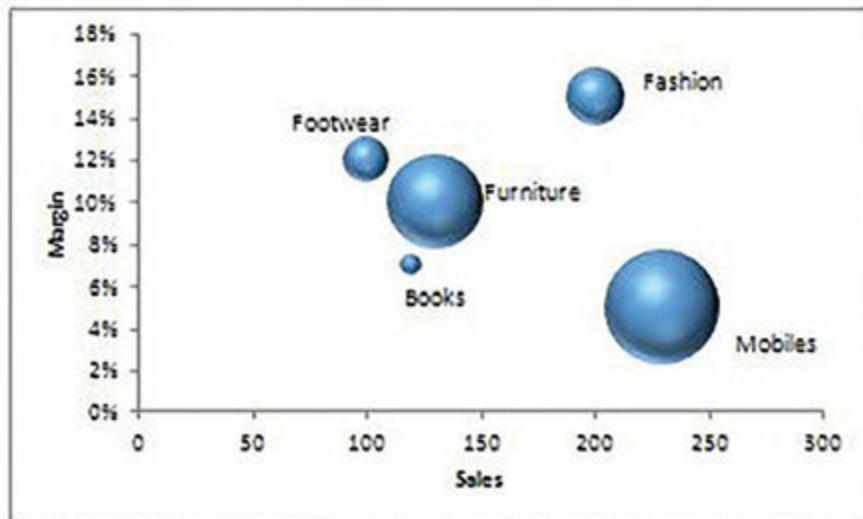
In stack graph the number from the members are stacked over previous member till all members are represented. Like area graph this can show the absolute number and the relative number. In first graph the absolute number is stacked over each other which represent salesfor the month for each category. 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 chance 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 than 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

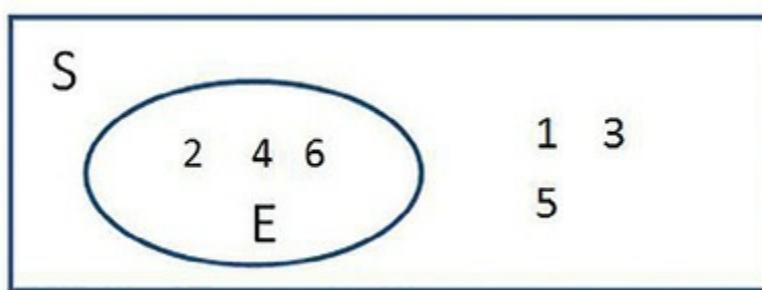
$S = \{1H, 1T, 2H, 2T, 3H, 3T, 4H, 4T, 5H, 5T, 6H, 6T\}$

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

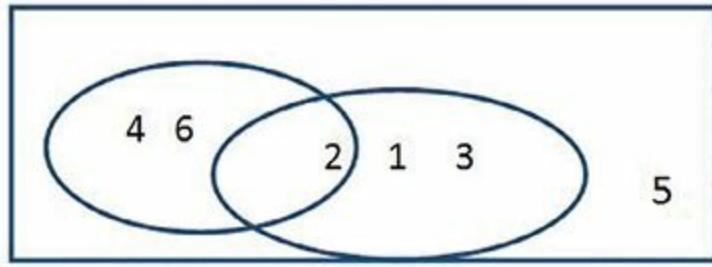
For any two events E and F of a sample space S, the union of the event is EUF consist of all elements which are either in E or F.

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.

$$E \text{ intersection } F = \{2\}$$

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 F 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

$$\text{Probability } P(E) = \text{Events / Space} = E/S$$

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  $0 \leq P(E) \leq 1$
2. Probability of sample space is 1 which is  $P(S) = 1$
3. For mutually exclusively  $E_1, E_2, \dots, E_n$  –  $P(UE_n) = \text{summation } P(E_n)$

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  $E_c$  in this book. If the probability of an event being occurrence is 0.20 then probability of not happening is 0.8 ( $1-0.20$ ).

$$P(E_c) = 1 - P(E)$$

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  $P(EF)=\text{null}$ , 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  $\frac{1}{3}$ .

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

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 develop 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 developing 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 positive result for 80% of time and negative result for 20% of time. If a person does not have cancer then the test gives positive result in 5% of times and negative results in 95% of times. What is the chance 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

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

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

$$p(x) \geq 0 \text{ for all value of } x$$

$$\sum p(x) = 1$$

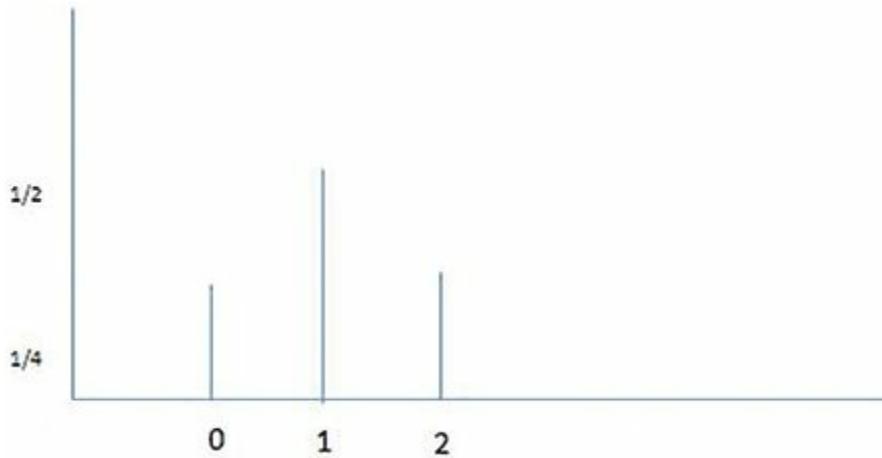
For example, if we toss two unbiased coin the probability associated with each combination of H and T is  $\frac{1}{4}$ . The random variable can assume value 0, 1, 2 for appearance of H.

$$P(x=0)=P(TT)=\frac{1}{4}$$

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

$$P(x=2)=p(HH)=\frac{1}{4}$$

x	p(x)
0	$\frac{1}{4}$
1	$\frac{1}{2}$
2	$\frac{1}{4}$



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  $\sigma^2$  is defined as the average of the square distances of  $x$  from the population mean  $\mu$ . 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.

$$\sigma = \sqrt{(\sigma^2)}$$

### 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  $q = 1 - p$ .
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 is denoted by  $(n,x)$ .

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

The expected value of discrete random variable is  $\mu = \sum x * p(x)$  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?

$$\text{Here } p=0.6, q=0.4, P(3) = (5,3)p^3q^2$$

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

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.

$$\begin{aligned} P(x \geq 3) &= P(3) + P(4) + P(5) = 5!/3!2! * (0.6)^3(0.4)^2 + 5!/4!1! * (0.6)^4(0.4) + \\ &5!/5!0! * (0.6)^5(0.4)^0 \\ &= 34.56\% + 25.92\% + 7.78\% = 68.26\% \end{aligned}$$

### 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  $x = 0, 1, 2, 3, \dots$

For Poisson distribution  $\mu = \lambda$ ,  $\sigma^2 = \lambda$

Example:

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

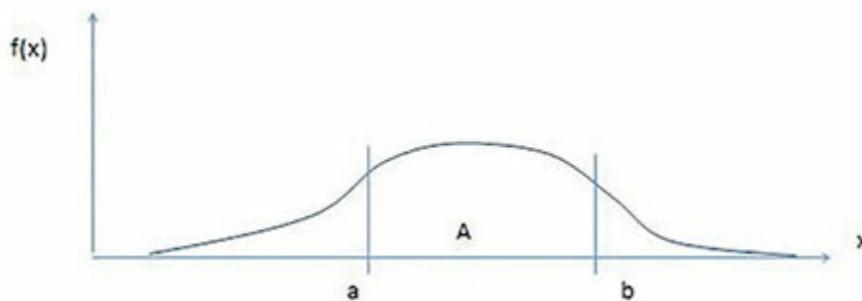
$$\lambda = 2, x = 3, e = 2.71828$$

$$P(x=3) = [e^{-2.71828} * 2^3] / 3! = 0.180.$$

### 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 appears 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  $b$  ( $a < x < b$ )  $\sim a \leq x \leq b$ . since point  $a$  or  $b$  in continuous variable is just a point and hence  $p(x=a)=p(x=b)=0$ .

$$P(A) = p(a < x < b) = \text{integration from } a \text{ to } b f(x)dx$$

Where  $f(x) \geq 0$ .

$$\text{Integration from } -\infty \text{ to } +\infty \int f(x)dx = 1$$

### 1.2.5.4 Uniform Distribution

In uniform distribution all points appear to have equally likely outcomes over their range of possible values

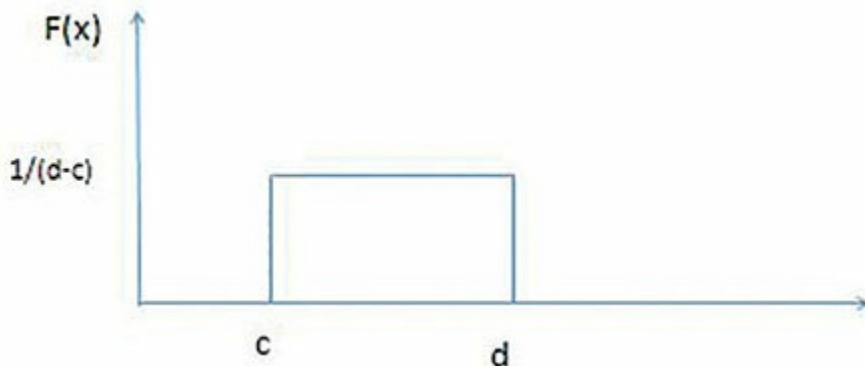
The possible value of  $x$  consist of all point in the interval between point  $c$  and point  $d$ . the height of  $f(x)$  is constant in that interval and equal to  $1/(d-c)$ .

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

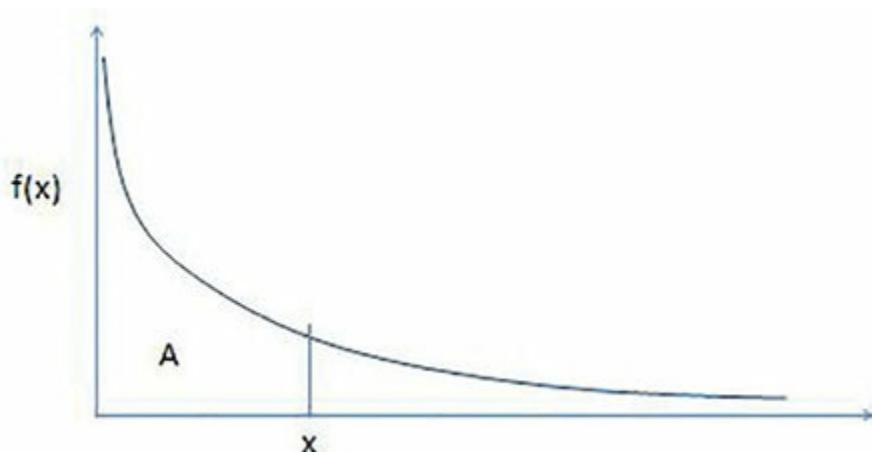
The exponential distribution is used to describe events such as length of time between earthquakes, distance travel by a person between two events, the length of time between breakdowns of manufacturing equipment's.

Probability distribution of exponential distribution is

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

where  $\lambda$  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

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

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

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 depend 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

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

$$= P(X \leq y+t | X > t)$$

$$= 1 - e^{-(\lambda(y+t))}$$

$G(y)$  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 use 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

$$F(x) = \frac{1}{\sigma} \sqrt{2\pi} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

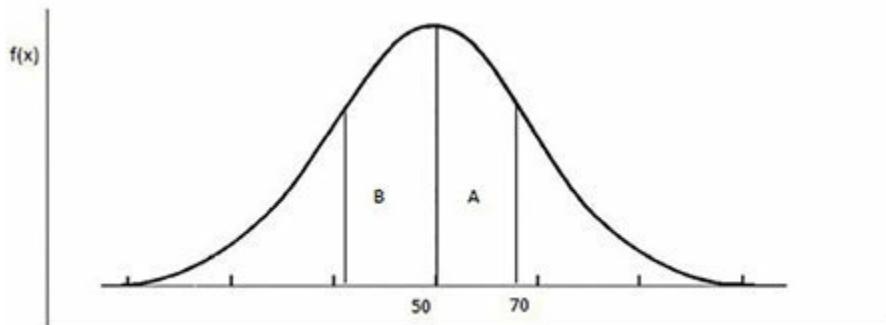
Where  $\mu$ = mean of the normal random variable x

$\sigma$ =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 z-score for a measurement is defined as the distance between the measurement and the population mean divide by the population standard deviation.

$$Z = \frac{x - \mu}{\sigma}$$



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

$$Z = \frac{x - \mu}{\sigma} = \frac{70 - 50}{15} = 1.33$$

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

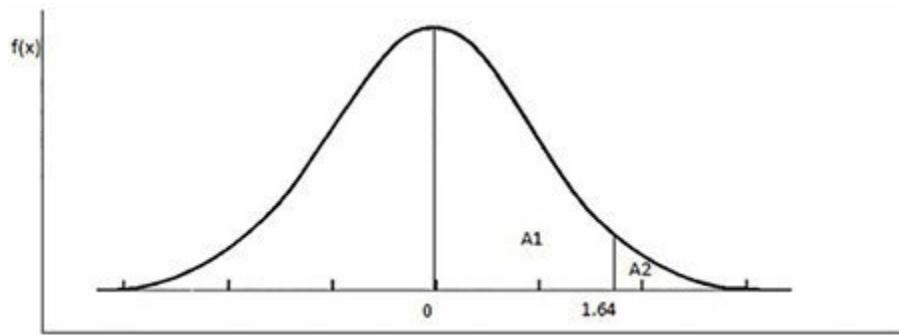
$$A=0.4082$$

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  $0.5 - \text{area of } A=0.5 - 0.4082=0.0908$

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.

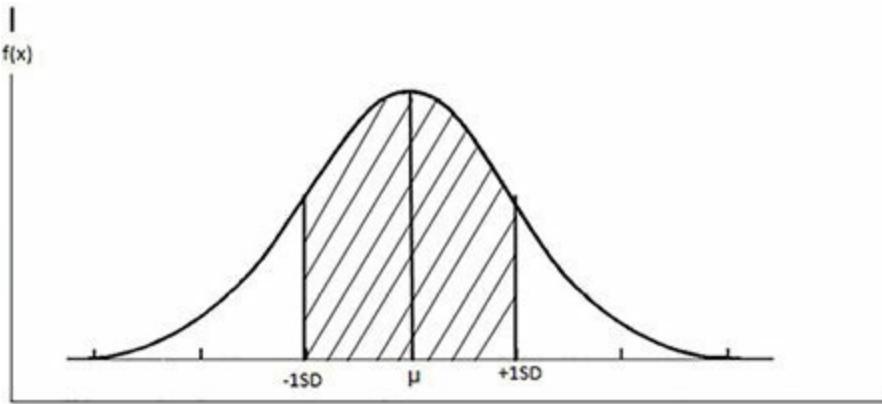
Hence area between -1.33 to 1.33 is  $2 \times 0.4082=0.8164$  Let us take another example find the probability that a standard normal random variable exceed 1.64 i.e  $p(z \geq 1.64)$ .



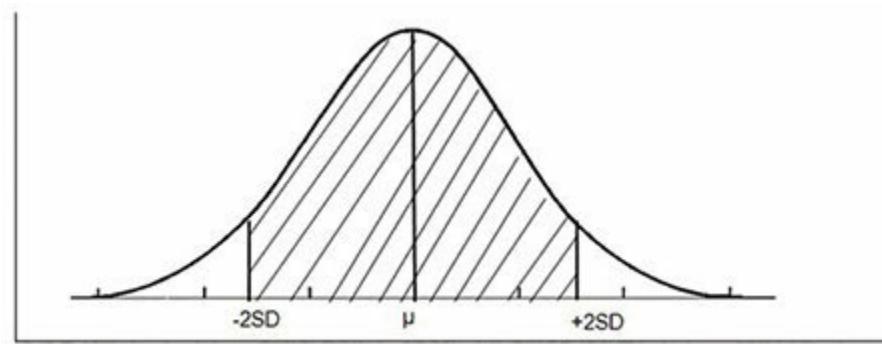
Area to the right side of the mean is 0.5. To find area of  $A_2$ , we have to find the area of  $A_1$  and calculate area of  $A_2$ .

$$A_2=0.5 - p(z \geq 1.64) = 0.5 - 0.4495 \text{ (from table)} = 0.0505$$

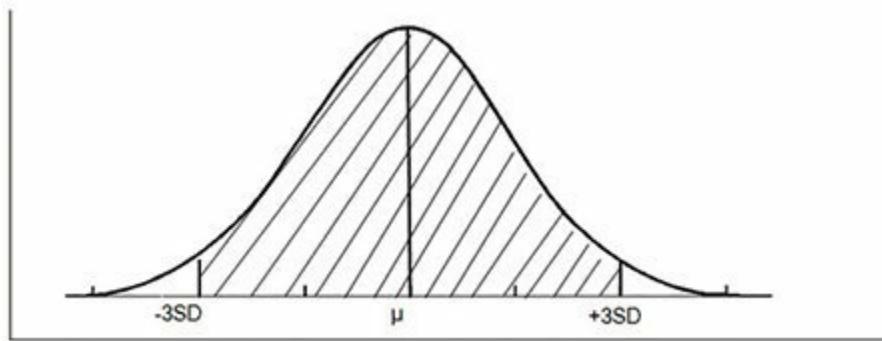
For the standard normal curve some of the area makes special significance as we often use them in our statistical inference such as hypothesis testing, confidence interval, creating control band etc.



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 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  $X_1, X_2, \dots, X_n$  be  $n$  random variable that are independent and identically distributed with mean  $\mu$  and standard deviation  $\sigma$ . It help us makes inference about the population without knowing anything about the distribution of the population.

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

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

Which implies that  $X' \rightarrow N(\mu, \sigma^2 / n)$

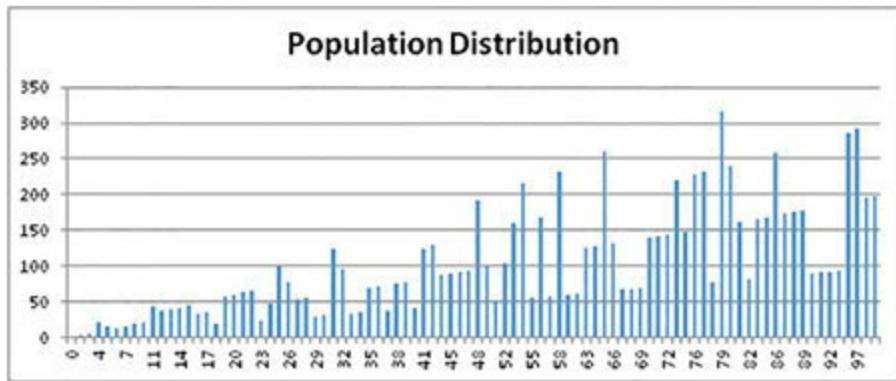
The sample mean can be approximated with a normal random variable with mean  $\mu$  and standard deviation  $\sigma / \sqrt{n}$ .

The first step in understanding the theorem is to prove its practical significance. Let's 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.

Sl. No.	Population
1	0
2	1
3	1
4	2
5	2
6	4
7	4
8	4
9	6
10	6
11	8
12	9
13	9
14	9
15	9
16	9
17	9
18	9
19	9
20	9
21	9
22	11
23	11
24	11
25	11
26	11
27	11
28	11
29	11
30	11
31	11
32	11
33	11
34	11
35	11
36	11
37	11
38	11
39	11
40	11
41	11
42	11
43	11
44	11
45	11
46	11
47	11
48	11
49	11
50	11
51	11
52	11
53	11
54	11
55	11
56	11
57	11
58	11
59	11
60	11
61	11
62	11
63	11
64	11
65	11
66	11
67	11
68	11
69	11
70	11
71	11
72	11
73	11
74	11
75	11
76	11
77	11
78	11
79	11
80	11
81	11
82	11
83	11
84	11
85	11
86	11
87	11
88	11
89	11
90	11
91	11
92	11
93	11
94	11
95	11
96	11
97	11
98	11
99	11
100	11

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.

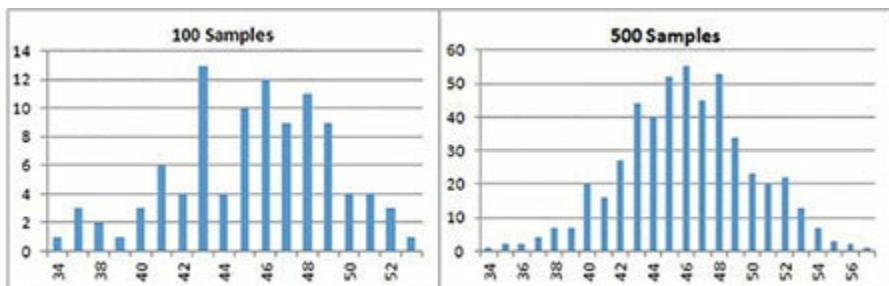
Now Open Sheet2. Name A1 as Sample No. and from B1 to AY1 name cell as Ob1, Ob2...Ob50. We are going to use sample size ( $n$ ) =50. From A2 to A10001 name the sample as 1 to 10000. We are going to take 10,000 samples from the population.

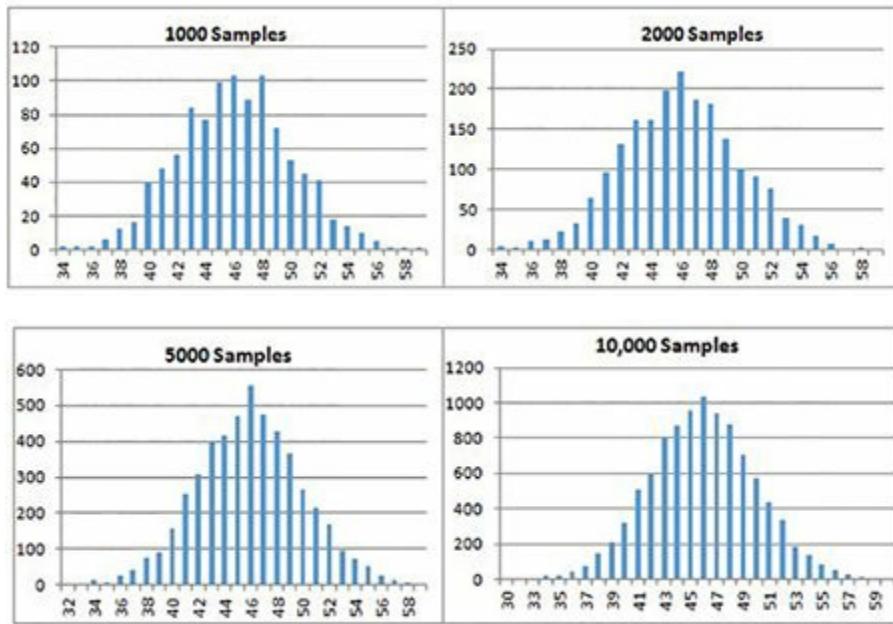
In cell B2 write formula =VLOOKUP(INT(RAND() \* 200), Sheet1!\$A:\$B, 2, 0). This will help us pick one number from 200 randomly. Copy and paste this formula in all cells in the range B2:AY10001. Now we have 10,000 samples with sample size 50 each. From AZ2 to AZ10001 take the average of each sample. 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  $\mu - \sigma$  to  $\mu + \sigma$

- About 95% of the sample mean will be within the range of  $\mu - 2\sigma$  to  $\mu + 2\sigma$
- About 99% of the sample mean will be within the range of  $\mu - 3\sigma$  to  $\mu + 3\sigma$

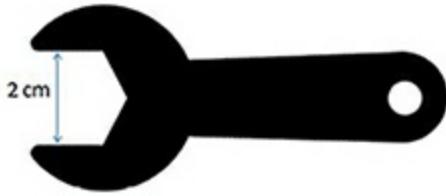
Same being derived from the distribution of the means as below table. This is another proof that the central limit theorem is true

Range	Lower Limit	Upper Limit	No of Sample mean	Total Samples	% sample in this range	Reference %
$\mu - \sigma$ to $\mu + \sigma$	42	50	7379	10000	73.8%	68%
$\mu - 2\sigma$ to $\mu + 2\sigma$	38	54	9646	10000	96.5%	95%
$\mu - 3\sigma$ to $\mu + 3\sigma$	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



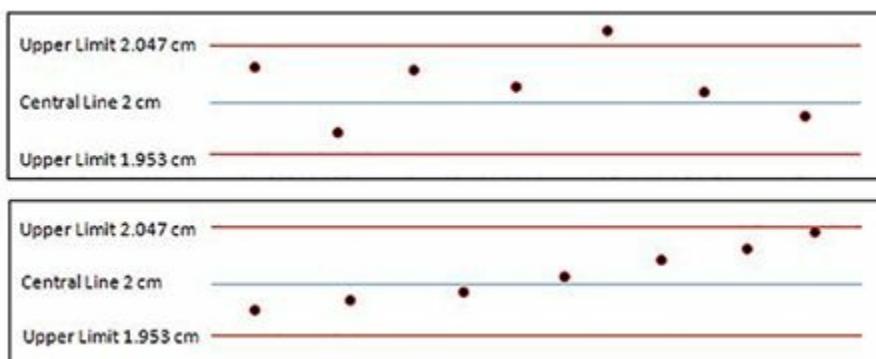
You are manufacturer a wrench which is 2 cm in the space using a machine. A tolerance of 0.5 mm is allowed on either side that is 1.95 cm to 2.05 cm space width is acceptable. Since the machine produce thousands of them every day it is not possible for you to measure each one of them. As a process you 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  $n \rightarrow \infty$  the distribution of  $X'$  is normal with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$

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 2cm + 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

Area under the Normal Curve from 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	0.00359	0.00798	0.01197	0.01595	0.01964	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	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	0.21224	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24531	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	0.31856	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	0.35993	0.36214
1.1	0.36433	0.36656	0.36864	0.37076	0.37286	0.37495	0.37694	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39434	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41303	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	0.45332	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	0.46483	0.46562	0.46638	0.46712	0.46784	0.46854	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	0.49303	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	0.49991	0.49991	0.49992	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 been drawn from the population all the remaining elements have the equal chances of being included in the sample in next draw.
2. **Cluster Sampling:** In this method the population is often divided into clusters 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 sub-population, 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  $k^{\text{th}}$  element. In this case,  $k = (\text{population size}/\text{sample size})$ . 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  $k^{\text{th}}$  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  $2 \times 100 + 2 \times 500 + 500 = 3500$  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 500<sup>th</sup> 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 1 – 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.

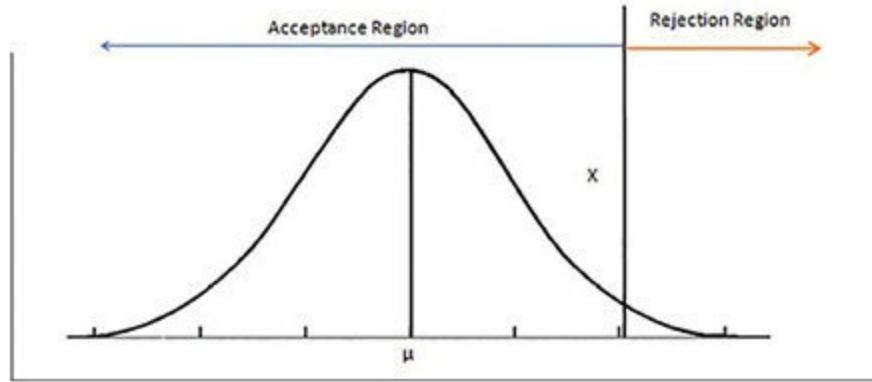
Broadly Hypothesis Testing can be classified into following types

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

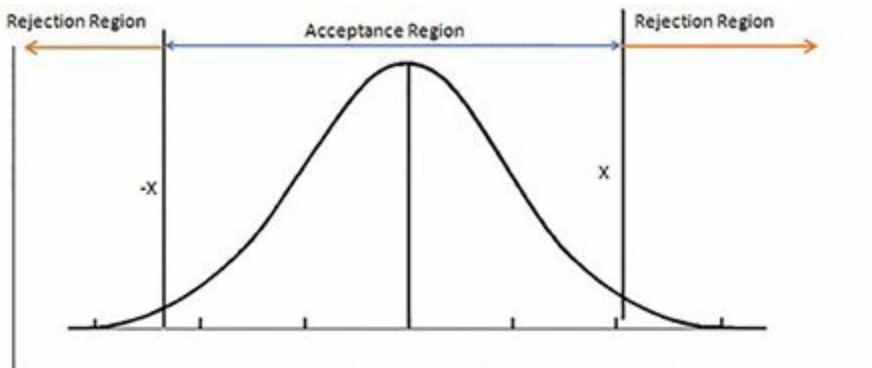
As shown in the first figure the one tail test is either done on the left side or right side based on the hypothesis. Assuming we have to test whether x is significantly greater than  $\mu$  with 95% confidence interval. We just need to calculate the Z-score of the x from the  $\mu$  with  $\sigma_s$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

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

If  $Z$  is greater than  $1.96\sigma$  (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 then we can conclude that null hypothesis is accepted else it is rejected. Here  $\sigma_s$  (standard error of the sample) is  $\sigma/\sqrt{n}$  where  $\sigma$  is population standard deviation and  $n$  is sample size.



### One Tail Hypothesis Testing.



### 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- $\alpha$	TYPE I Error $\alpha$	
	False	TYPE II Error $\beta$	CORRECT 1- $\beta$ (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.  $\alpha$  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  $\alpha$ . 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  $1-\beta$  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  $H_0: \mu_0 = 6$

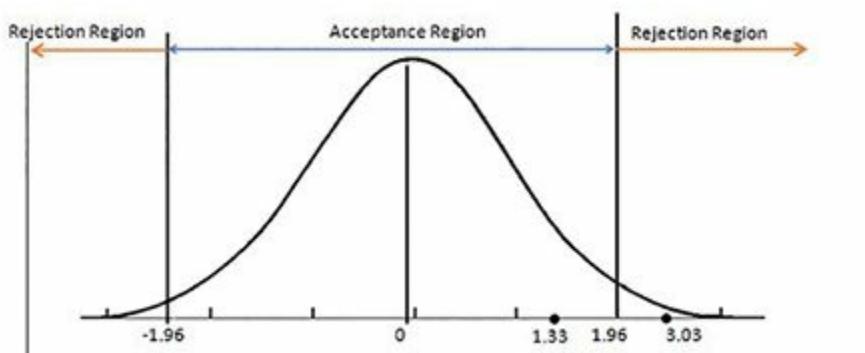
Alternate Hypothesis  $H_1: \mu_1 < 6$

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

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

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



Since Z score 1.33 lies within  $\pm 1.96$ SD 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

$$Z = (7 - 6) / 0.3 = 3.03$$

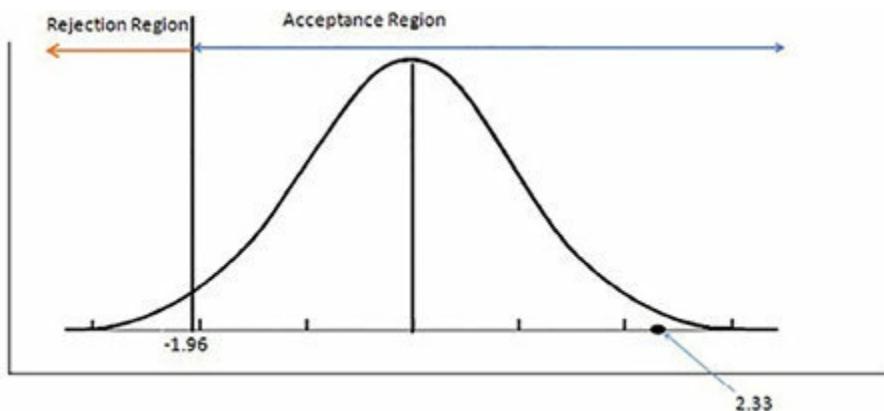
Since 3.03 lies outside the 95% confidence interval  $\pm 1.96$ SD (here SD is 1) we can reject the null hypothesis that the average study hours of the nursery classes are 6 hours.

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  $H_0: \mu = 6$  and Alternate Hypothesis would have been  $H_1: \mu < 6$

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.

$$Z \text{ score} = (6 - 5.2) / 0.3 = 2.33$$

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 two-sample 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 one-

sample 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

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Where  $\mu_1$  is mean of sample 1 and  $\mu_2$  is mean of sample 2.

In two sample the variance for sample is calculated as

$$\sigma_x^2 = \sigma_1^2/n_1 + \sigma_2^2/n_2$$

If we know that the standard deviation of population 1 is  $\sigma_1$  and standard deviation of the sample 2 is  $\sigma_2$ .

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

$$\sigma_{x_1 - x_2} = \sqrt{(s_p^2 * (1/n_1 + 1/n_2))}$$

Where pooled variance  $s_p$  is calculated as

$$s_p^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2] / (n_1 + n_2 - 2)$$

Where  $s_1$  is sample 1 standard deviation and  $s_2$  is sample 2 standard deviation.

Z score is calculate as

$$Z = [\mu_1 - \mu_2] / \sigma_{x_1 - x_2}$$

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 are 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

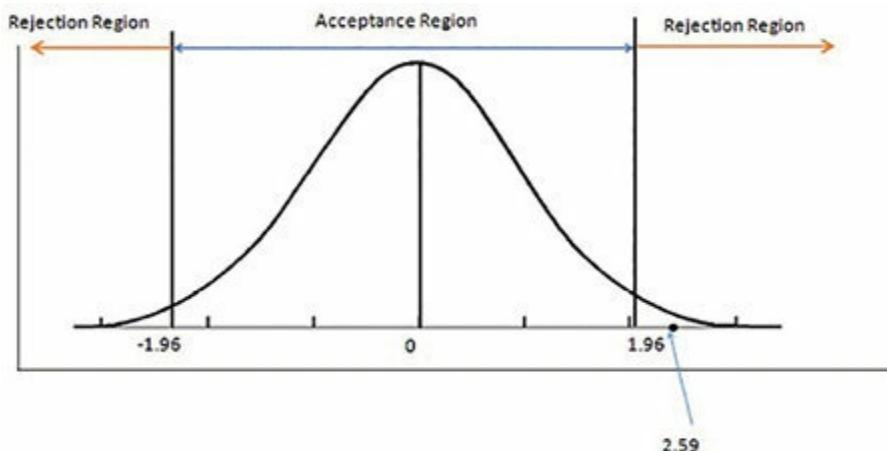
Standard deviation of Section B=15

We have calculate the pooled variance first

$$\begin{aligned} Sp &= [(n_1-1)s_1^2 + (n_2-1)s_2^2] / (n_1 + n_2 - 2) \\ &= [(50-1)*10^2 + (65-1)*15^2] / (50+65-2) \\ &= 170.8 \end{aligned}$$

$$\begin{aligned} \text{Standard Error of the sample mean difference } \sigma_{x1-x2} &= \sqrt{(sp^2 * (1/n_1 + 1/n_2))} \\ &= \sqrt{(170.8 * (1/50 + 1/65))} \\ &= 2.5 \end{aligned}$$

$$\begin{aligned} Z \text{ score} &= Z = [\mu_1 - \mu_2] / \sigma_{x1-x2} \\ &= (93.5 - 87) / 2.5 = 2.59 \end{aligned}$$



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.

### Lose Weight



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

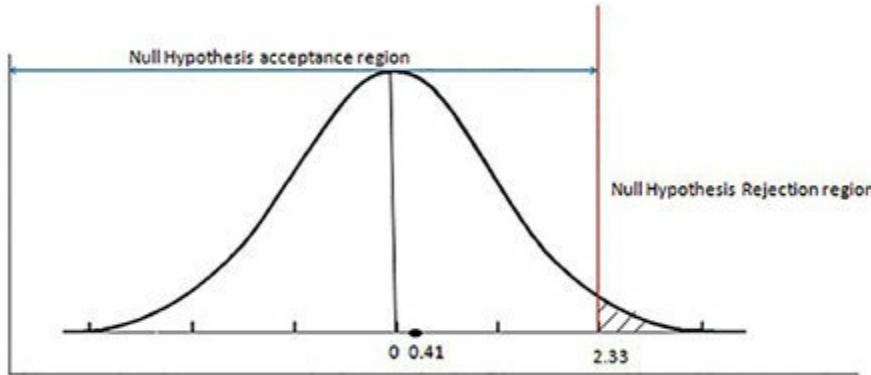
$$H_0: \mu=5$$

$$H_1: \mu > 5$$

Since you want 99% confidence here  $\alpha=0.01$ . Standard deviation of the weight reduction is 2.66 Kg. The standard error of mean reduction= $\sigma/\sqrt{2.66}=0.49$

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

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

$$H_0: p_0 = x$$

$$H_1: p_0 \neq x$$

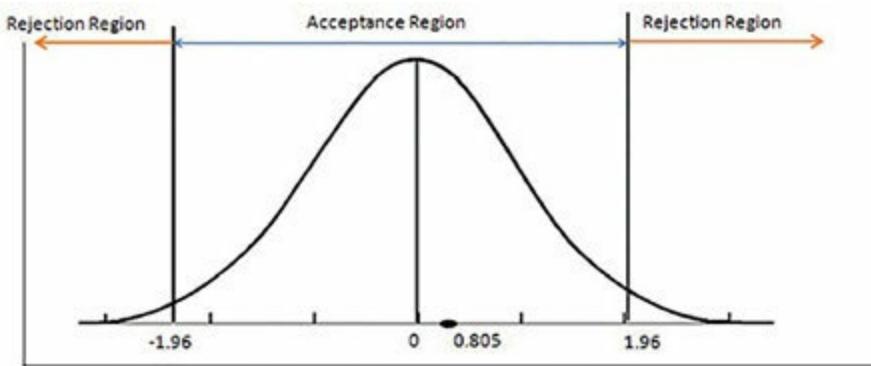
Let us take an example to show the steps involved. Proportion of the student absent at 4 pm class in the school is 0.35 as per school management. The sample of 30 students randomly checked at 4 pm found the proportion to be 0.42.

$$H_0: p_0 = 0.35$$

$$H_1: p_1 \neq 0.35$$

Standard deviation is calculated as  $\sigma_p = \sqrt{(pq/n)} = \sqrt{(0.35 * 0.65 / 30)} = 0.087$

$$Z \text{ score} = (p_0 - x) / \sigma_p = (0.35 - 0.42) / 0.087 = -0.805$$



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

$$\sigma_{p_1-p_2} = \sqrt{(p_1 q_1 / n_1 + p_2 q_2 / n_2)}$$

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

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

$$Q_{\text{pooled}} = 1 - P_{\text{pooled}}$$

Standard error of the Proportion difference is calculated as

$$\sigma_{p_1-p_2} = \sqrt{(P_{\text{pooled}} * Q_{\text{pooled}} / n_1 + P_{\text{pooled}} * Q_{\text{pooled}} / n_2)}$$

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

If  $-1.96 \leq Z \leq 1.96$  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

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

Then is  $\chi^2$  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.

$H_0$ : Two categorical variables are Independent

$H_1$ : 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

$$\chi^2 = \sum (O_i - E_i)^2 / E_i$$

where  $O_i$  is the observed count,  $E_i$  is expected count.

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

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

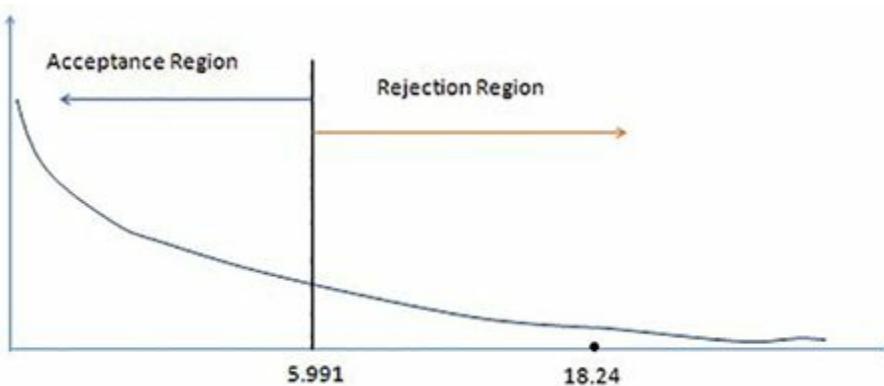
In the above table the opinion of the Students and teachers in the university about the changes in the syllabus of the subject.

The expected count is calculated as

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

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

$$\chi^2 = \frac{(135-113.59)^2}{113.59} + \frac{(80-84)^2}{84} + \frac{(65-82.35)^2}{82.35} + \frac{(65-86.40)^2}{86.40} + \frac{(68-63.94)^2}{63.94} + \frac{(80-62.65)^2}{62.65} \\ = 18.24$$



At 95% confidence interval ( $\alpha=0.05$ ) 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 die is 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

$H_0$ : the die is fair

$H_1$ : Die is not fair

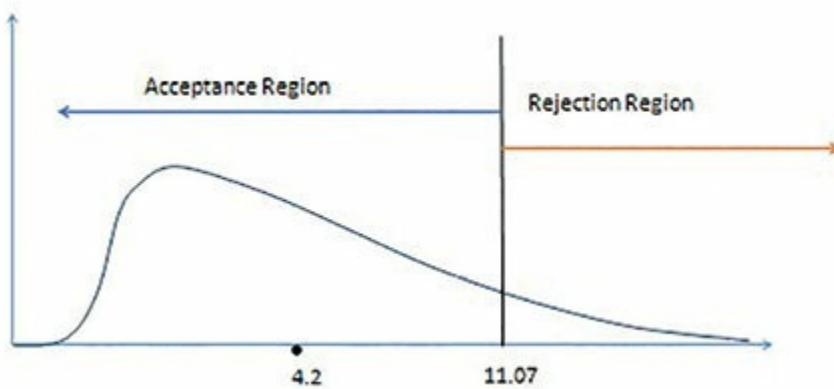
Significance level  $\alpha=0.05$

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

Die Face	1	2	3	4	5	6
Frequency	8	10	14	7	13	8
Expected Frequency	10	10	10	10	10	10
Oi-Ei	-2	0	4	-3	3	-2
(Oi-Ei) <sup>2</sup>	4	0	16	9	9	4
(Oi-Ei) <sup>2</sup> /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.

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	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  $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_n$

Alternative Hypothesis  $H_1: \mu_1, \mu_2, \mu_3 \text{ and } \mu_n \text{ are not equal}$

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

$x_{Tj}$ =mean of sample j

k=number of samples

$n_j$ =number of items in the sample j

$x_{ij}$ =i<sup>th</sup> member of j<sup>th</sup> sample

$x_{Tj} = \sum x_{ij} / n_j$  where I = 1 to  $n_j$

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

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

Variance within the samples  $\sigma^2_w = \sum (n_j - 1) / (n_T - k) * s_j^2$

Total sample size  $n_T = \sum n_j$

$s_j^2$ =sample variance of the j<sup>th</sup> sample= $\sum (x - x_{Tj})^2 / (n_j - 1)$

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

= between-column variance/within-column variance= $\sigma^2_T / \sigma^2_w$

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

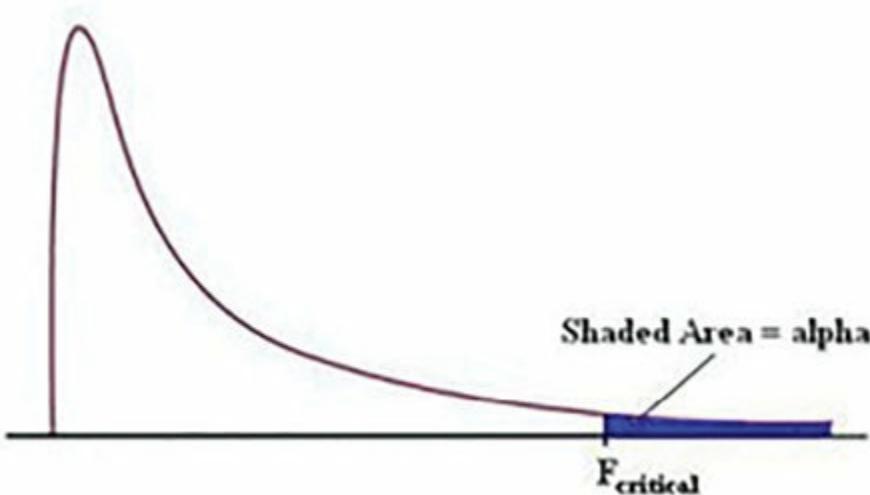
F value tend to be large.

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

$$= \sum(n_j - 1) - k$$



Where  $1 - \alpha$  is the confidence interval.

### ANOVA Table

Source of Variation	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	$SST = \sum \sum (x - x_T)^2$	N-1		

Where  $N = \sum n_j$

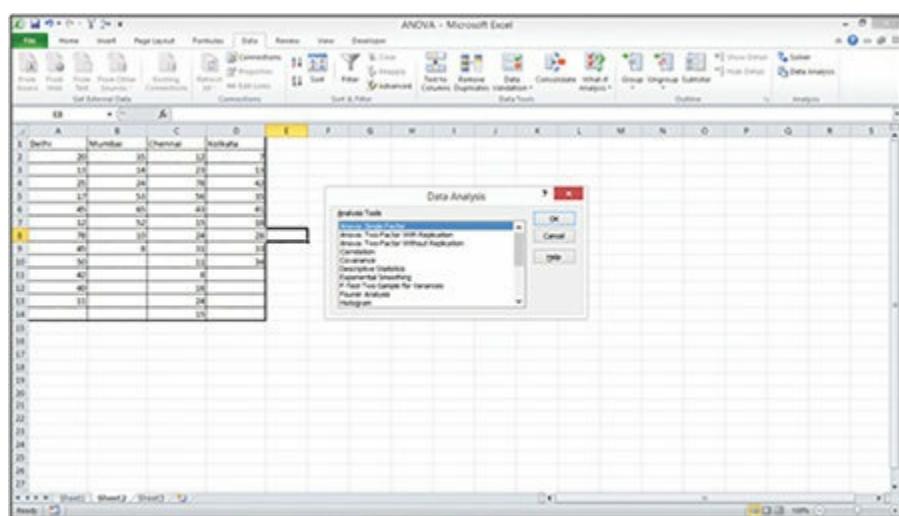
Example: In below example the sample from four cities are compare to find out if there is difference in the income of the family from four cities. With alpha value of 0.05 the F statistics from the F distribution table with degree of freedom Numerator =3 and Denominator=38 is 2.84. F-statistics from 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.

Samples				$S^2_{\text{between}}$			
Delhi	Mumbai	Chennai	Kolkata	Delhi	Mumbai	Chennai	Kolkata
20	35	22	7	173.4	5.6	236.7	436.3
13	14	23	33	406.7	346.9	19.2	221.7
25	24	28	42	66.7	24.4	2561.9	199.1
17	53	56	35	261.4	455.1	818.8	50.6
45	65	43	41	140.0	1048.1	243.8	171.9
12	52	15	18	448.0	375.4	153.6	97.8
28	10	24	28	2010.0	511.9	11.5	0.0
45	8	31	33	140.0	606.4	13.1	26.1
50		11	34	283.4		268.5	37.3
42		8		28.0		375.8	
40		16		46.7		329.6	
11		24		491.4		11.5	
		15					153.4
Mean ( $\bar{x}_{ij}$ )				$S^2_{\text{within}}$	4545.8	3383.9	4997.1
sample size ( $n_j$ )				$S^2_{\text{within}}/(n-1)$	413.2	483.4	416.4
Number of sample ( $k$ )					155.1		
Grand Mean ( $\bar{x}_{..}$ )							
All sample elements ( $N$ )							

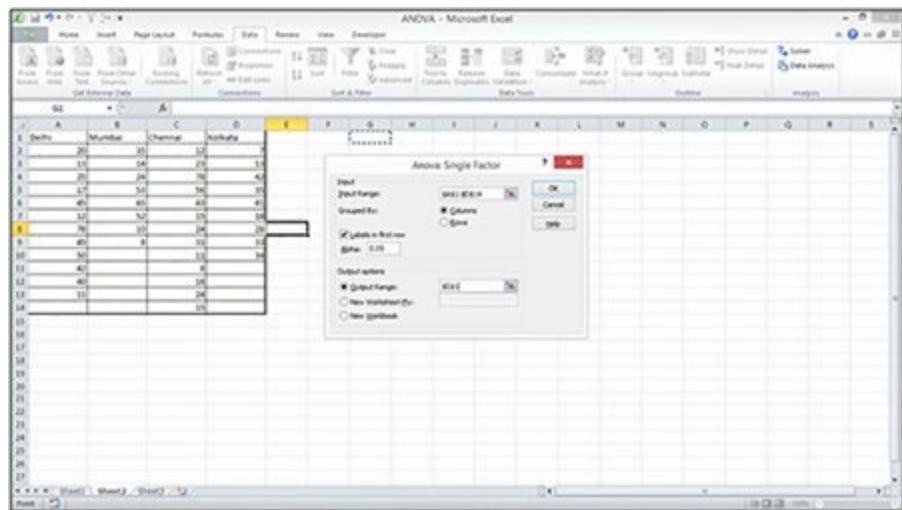
ANOVA can be calculated with pre-defined function in the Excel. Using same sample data we have calculated the ANOVA in the screenshot below

Delhi	Mumbai	Chennai	Kolkata
20	35	22	7
13	14	23	13
25	24	28	42
17	53	56	35
45	65	43	41
12	52	15	18
28	10	24	28
45	8	31	33
50		11	34
42		8	
40		16	
11		24	
		15	

Go to **Data Analysis Tool** option in Excel



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



Add Alpha( $\alpha$ ), 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* Distribution

		$\alpha = .05$									
$df_N$	$df_D$	1	2	3	4	5	6	7	8	9	10
1	1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
$\infty$	$\infty$	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 to finish a unit)	Paint B(Hours required to finish a unit)	Maximum 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  $2x_1 + x_2 \leq 8$  (Machine1 constraint)

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

$x_1, x_2 \geq 0$  (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  $X_1$ =Dollar Spend on wheat flour

$X_2$ =Dollar Spend on cheese

$X_3$ =Dollar Spend on liver in a day

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

Subject to

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

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

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

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

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

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  $Z=C_1X_1+C_2X_2+ \dots +C_nX_n$

Subject to Constraint:

$a_{11}X_1 + a_{12}X_2+ \dots +a_{1n}X_n \leq/ \geq M_1$  (Constraint 1)

$a_{21}X_1 + a_{22}X_2+ \dots +a_{2n}X_n \leq/ \geq M_2$  (Constraint 2)

....

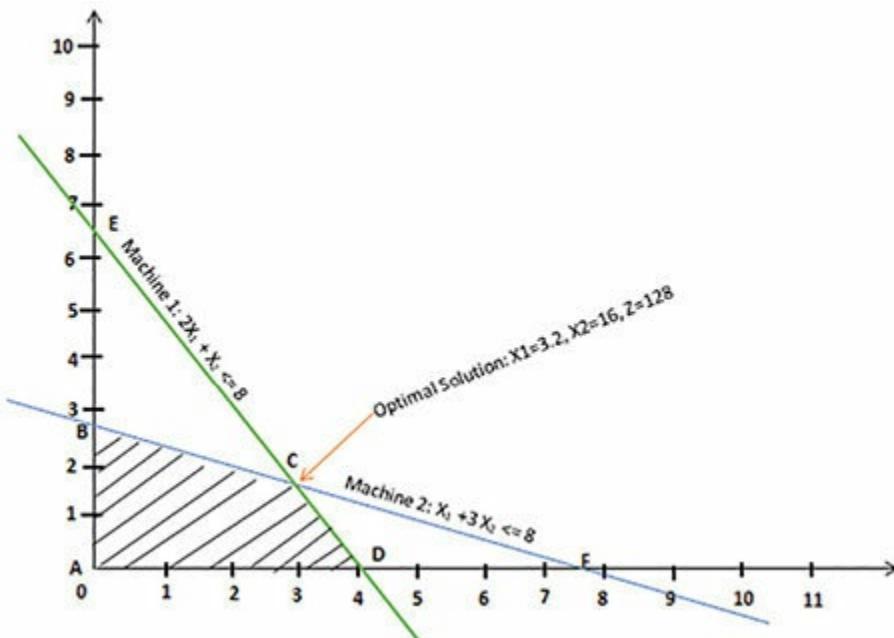
$a_{m1}X_1 + a_{m2}X_2+ \dots +a_{mn}X_n \leq/ \geq M_m$  (Constraint n)

$a_{11}, a_{12} \dots a_{m1} \dots a_{mn} \geq 0$  (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.

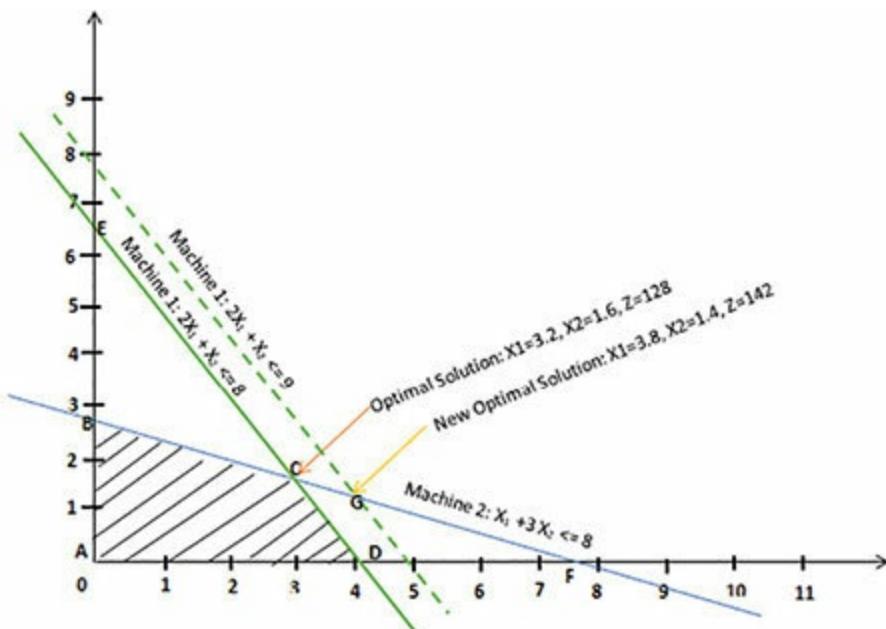


Not all the problem definition and constraint will have feasible solution there are many cases where we do not have feasible solutions. You need to take care of such problems. The two such cased are unbounded solution and non-feasible solution. The topic is beyond the scope of this book. Interested reader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

$$\Delta Z = [Z_g - Z_c] / \text{capacity changes} = (142 - 128) / (9 - 8) = 14 \text{ per hours.}$$

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  $B(0,2.67) = 2x_0 + 2.67x_1 = 2.67$  hours

At F the machine 1 capacity is  $F(8,0) = 8x_2 + 0x_1 = 16$

Thus we can conclude that the shadow price of \$14 per hours for machine 1 will remain valid from range  $2.67 \leq \text{machine1 capacity} \leq 16$ .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

$4 \leq \text{machine 2 capacity} \leq 24$  hours.

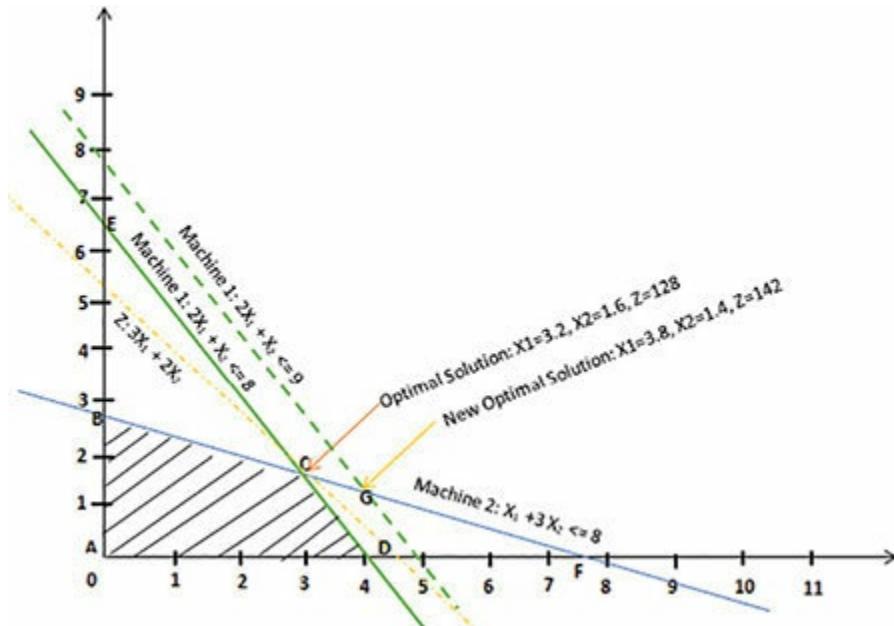
The computed limits for machine 1 and machine 2 are referred to as the feasible ranges.

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  $z = c_1x_1 +$

$c_2x_2$  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

$$1/3 \leq c_1/c_2 \leq 2/1 \text{ or } 0.333 \leq c_1/c_2 \leq 2.$$



#### 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  $2x_1 + x_2 \leq 8$  (Machine1 constraint)

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

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

**Step 1:** Set variable as 0 for initial state.

Excel Work - Excel					
	Variable	X1 (Paint A)	X2 (Paint B)	Max Profit (Z)	
	Value	0	0	0	
	Contribution (Rs.)	30	20	LHS	RHS
	Constraints	Machine 1(Hours)	2	1	0
				<=	8
		Machine 2(Hours)	1	3	0
				<=	8

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

	A	B	C	D	E	F	G	H	I	J	K	L	M
SUM					=SUMPRODUCT(C3:D3,C5:D5)								
1													
2	Variable	X1 (Paint A)	X2 (Paint B)	Max Profit (Z)									
3	Value	0	0	=SUMPRODUCT(C3:D3,C5:D5)									
4	Contribution (Rs)	30	20		LHS								
5	Constraints	Machine1(Hours)	2	1	0	<=	8						
6		Machine2(Hours)	1	3	0	<=	8						
7													
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													
18													
19													

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

	A	B	C	D	E	F	G	H	I	J	K	L	M
SUM					=SUMPRODUCT(C3:D3,C5:D5)								
1													
2	Variable	X1 (Paint A)	X2 (Paint B)	Max Profit (Z)									
3	Value	0	0	0									
4	Contribution (Rs)	30	20		LHS								
5	Constraints	Machine1(Hours)	2	1	=SUMPRODUCT(C3:D3,C7:D7)								
6		Machine2(Hours)	1	3	0	<=	8						
7													
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													
18													
19													

**Step 4:** Go to Data, Click on Solver

	A	B	C	D	E	F	G	H	I	J	K	L	M
SUM													
1													
2	Variable	X1 (Paint A)	X2 (Paint B)	Max Profit (Z)									
3	Value	0	0	0									
4	Contribution (Rs)	30	20		LHS								
5	Constraints	Machine1(Hours)	2	1	=SUMPRODUCT(C3:D3,C7:D7)								
6		Machine2(Hours)	1	3	0	<=	8						
7													
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													
18													
19													

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

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

**Step 7:** Add Constraints one by one.

Machine 1 constraint

The screenshot shows the Microsoft Excel Solver interface. The worksheet is titled "Constraints". It includes columns for Variable (X1, X2), Contribution (\$), and Max Profit (\$). There are two rows for constraints: one for Machine 2 (2X1 + 2X2 ≤ 8) and one for Machine 4 (2X1 + 3X2 ≤ 6). An "Add Constraint" dialog box is displayed, showing the formula  $\leq$  \$B\$8 for the constraint. The Solver ribbon tab is selected.

## Machine2 constraints

The screenshot shows the Microsoft Excel Solver interface. The worksheet is titled "Constraints". It includes columns for Variable (X1, X2), Contribution (\$), and Max Profit (\$). There are two rows for constraints: one for Machine 2 (2X1 + 2X2 ≤ 8) and one for Machine 4 (2X1 + 3X2 ≤ 6). An "Add Constraint" dialog box is displayed, showing the formula  $\leq$  \$B\$8 for the constraint. The Solver ribbon tab is selected.

## Non Negativity constraint

The screenshot shows the Microsoft Excel Solver interface. The worksheet is titled "Constraints". It includes columns for Variable (X1, X2), Contribution (\$), and Max Profit (\$). There are two rows for constraints: one for Machine 2 (2X1 + 2X2 ≤ 8) and one for Machine 4 (2X1 + 3X2 ≤ 6). An "Add Constraint" dialog box is displayed, showing the formula  $\geq$  \$C\$7 for the constraint. The Solver ribbon tab is selected.

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

Go to options and make changes if required

Select Simplex Linear model for linear problems

## Step 10: Click on Solve button

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

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

The screenshot shows two overlapping dialog boxes in Microsoft Excel:

- Solver Results**: Shows the message "Solver found a solution. All constraints and optimality conditions are satisfied." with options to "Keep Solver Solution" (selected), "Generate Original Values", "Return to Solver Parameters Dialog", and "Online Reports".
- LP Solver Parameters**: Shows the setup for the linear programming problem. The objective function is set to "Max Profit (Z)" with coefficients [X1 (Point A)] = 3.2 and [X2 (Point B)] = 1.6. The constraints are defined as follows:
 

Constraints	Machine1(Hours)	Machine2(Hours)	RHS
Machine1(Hours)	2	1	8
Machine2(Hours)	1	3	8

Final output is  $X_1=3.2$ ,  $X_2=1.6$  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.

The screenshot shows the 'Solver Results' dialog box in Microsoft Excel. The 'Report' tab is selected, displaying the 'Sensitivity Report'. The report includes three tables: 'Objective Cell (Max)', 'Variable Cells', and 'Constraints'.

Objective Cell (Max)					
Cell	Name	Original Value	Final Value		
\$E\$3	Value Max Profit (Z)	0	128		

Variable Cells					
Cell	Name	Original Value	Final Value	Integer	
\$C\$3	Value X1 (Paint A)	0	3.2 Contin		
\$D\$3	Value X2 (Paint B)	0	1.6 Contin		

Constraints					
Cell	Name	Cell Value	Formula	Status	Slack
\$E\$7	Machine 1(Hours) LHS	8	\$A\$7<=\$G\$7	Binding	0
\$E\$8	Machine 2(Hours) LHS	8	\$A\$8<=\$G\$8	Binding	0
\$C\$3	Value X1 (Paint A)	3.2	\$C\$3>=0	Not Binding	3.2
\$D\$3	Value X2 (Paint B)	1.6	\$D\$3>=0	Not Binding	1.6

## 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 X<sub>1</sub>, 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 + 8 the 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.

**Variable Cells**

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
\$C\$3	Value X1 (Paint A)	3.2	0	30	10	23.3
\$D\$3	Value X2 (Paint B)	1.6	0	20	10	5

**Constraints**

Cell	Name	Final Value	Shadow Price	Constraint R.H.S.	Allowable Increase	Allowable Decrease
\$E\$7	Machine1(hours) LHS	8	14	8	8	5.3
\$E\$8	Machine2(hours) LHS	8	2	8	16	4

## Limit Report

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

**Objective**

Cell	Name	Value
\$E\$3	Value Max Profit (Z)	128

**Variable**

Cell	Variable Name	Value	Lower Limit	Objective Result	Upper Limit	Objective Result
\$C\$3	Value X1 (Paint A)	3.2	0	32	3.2	128
\$D\$3	Value X2 (Paint B)	1.6	0	96	1.6	128

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

## Learning from the chapter

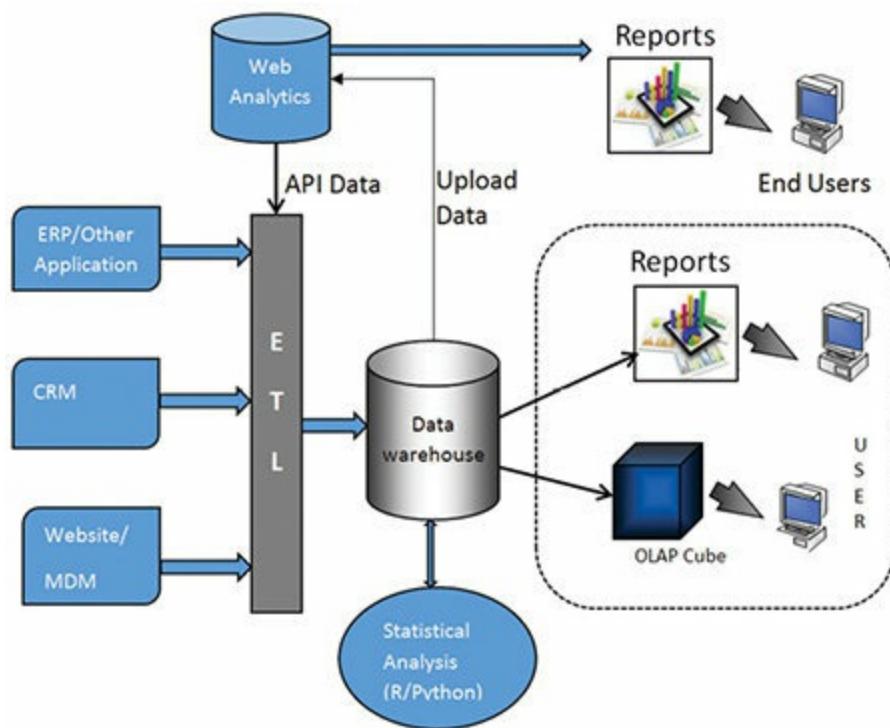
- How to formulate problem into linear programming problem
- Graphical representation of the Linear Programming and solving using Graphical solution
- How to do sensitivity Analysis in graphical solution
- Formulation and solving of LP using Excel
- How to interpret the answers from the Excel output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

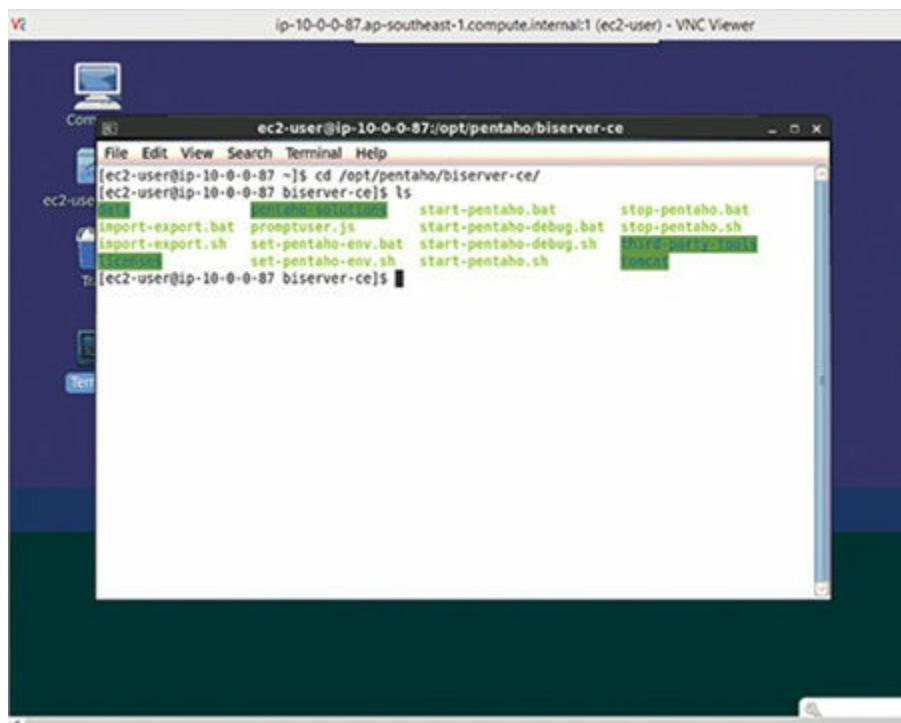
In this book we will not be showing the actual installation steps which are available in the Pentaho Support site, manual and bloggers sites. We will assume user has installed latest version of MySQL database in the same machine or different machine, java version required by Pentaho Version, Pentaho 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.



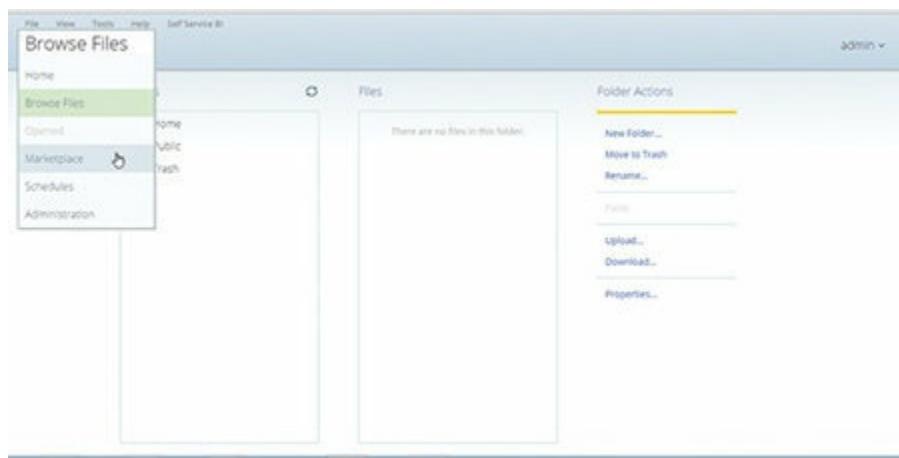
Once the server is up, enter URL <http://xxx.xxx.xxx.xxx:8080/pentaho> in the browser to open the login windows. The Login window is as shown in picture below.



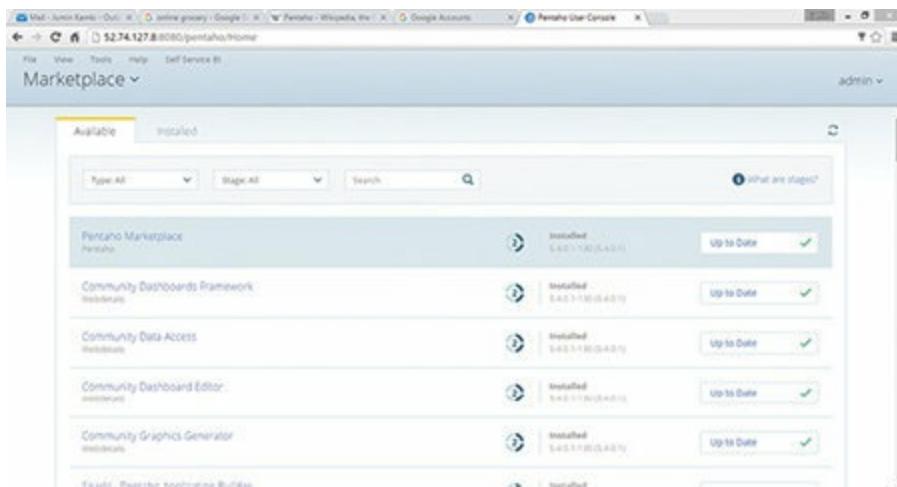
In the first instance you have to login using admin user id. Once you login into Pentaho, you will see below windows.



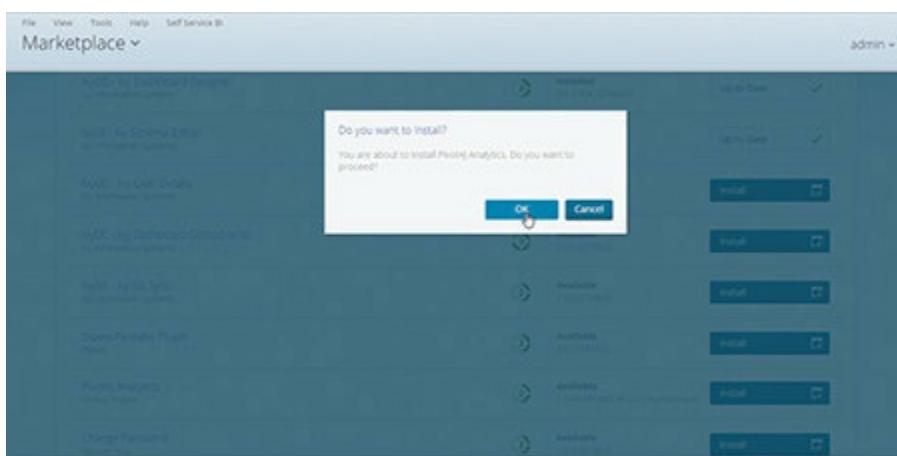
As we have to installed different components, go to **Browse Files** -> **Marketplace**. The market place contains all components free and paid available with the version of Pentaho.



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



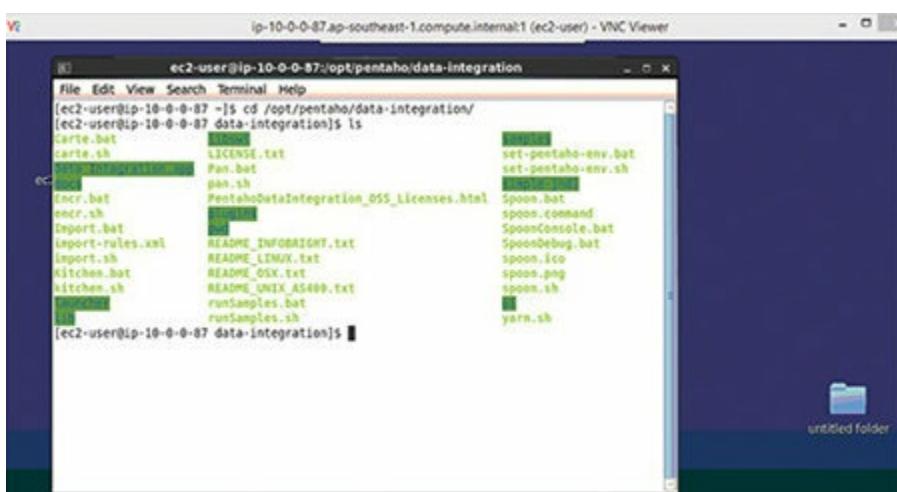
An example of installation of a component.



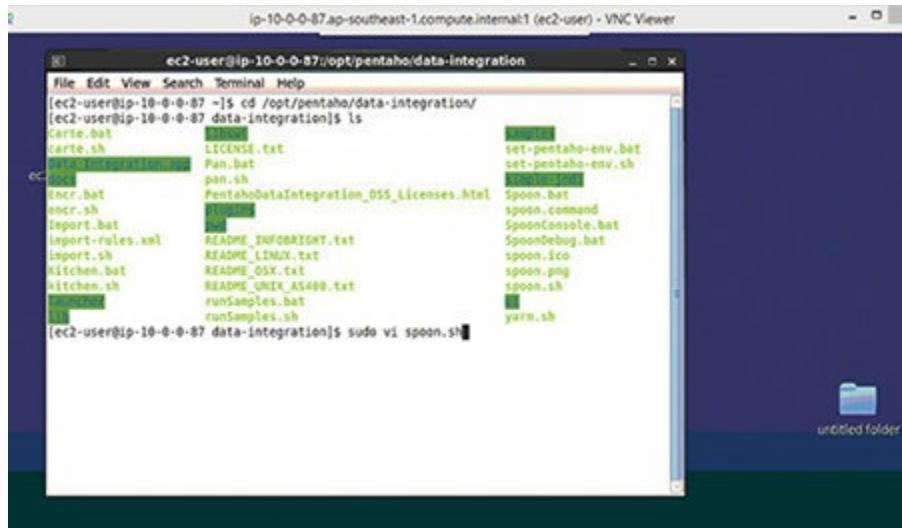
Once all components are installed, we are ready for using the Pentaho for real business cases.

### 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`



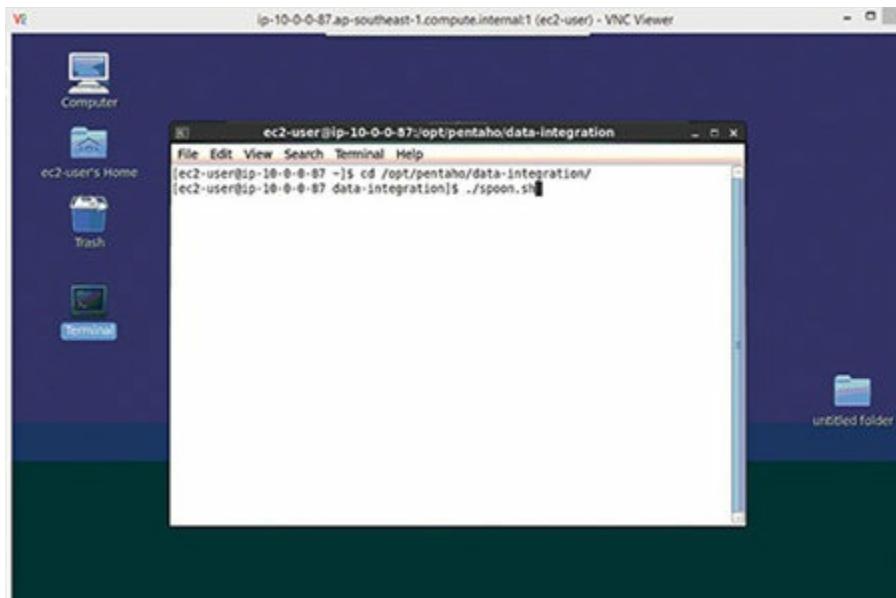
By default spoon comes with very minimal memory configuration. Hence it is advised to increase the memory as per available RAM in the server. Just open spoon.sh file in vi or vim editor and change the memory setting as shown in pictures below.



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

A screenshot of a terminal window titled "ec2-user@ip-10-0-0-87:/opt/pentaho/data-integration". The window displays the contents of the spoon.sh file. A specific line of code, "# PENTaho DI JAVA OPTIONS=-Xmx3g -XX:MaxPermSize=5000m", is highlighted with a green selection bar.

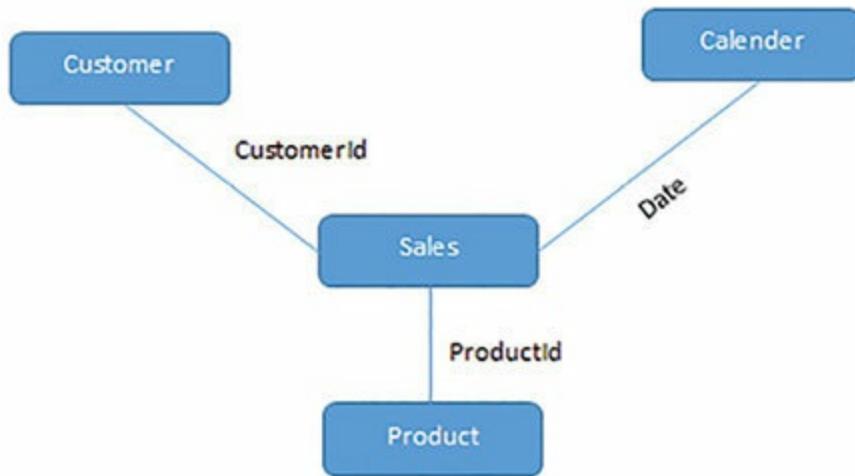
Start the PDI by typing the command as shown in below picture.



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	Quarter	Day	Weekday
4/1/2015	2015	2015_16	04_Apr	2015_04	Q1	1	Wednesday
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	Wednesday
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

OD000008	CC00087	8-Apr-15	100029	1	3000	360	Completed
OD000009	CC00067	9-Apr-15	100003	3	30000	3600	Returned
OD000010	CC00094	10-Apr-15	100002	3	50000	0	Completed
OD000011	CC00073	11-Apr-15	100018	1	4600	552	Completed
OD000012	CC00014	12-Apr-15	100000	1	40000	0	Completed
OD000013	CC00069	13-Apr-15	100007	2	46000	0	Completed
OD000014	CC00034	14-Apr-15	100025	3	4600	0	Completed
OD000015	CC00094	15-Apr-15	100023	2	1000	0	Completed
OD000016	CC00026	16-Apr-15	100007	2	46000	0	Completed
OD000017	CC00059	17-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  
)
```

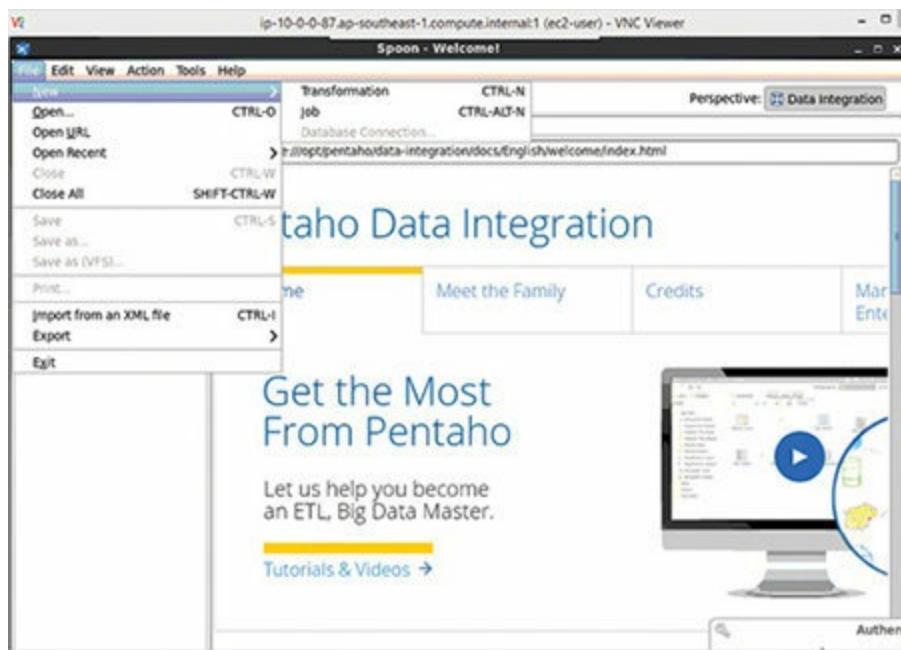
### Create Table **Sales**

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

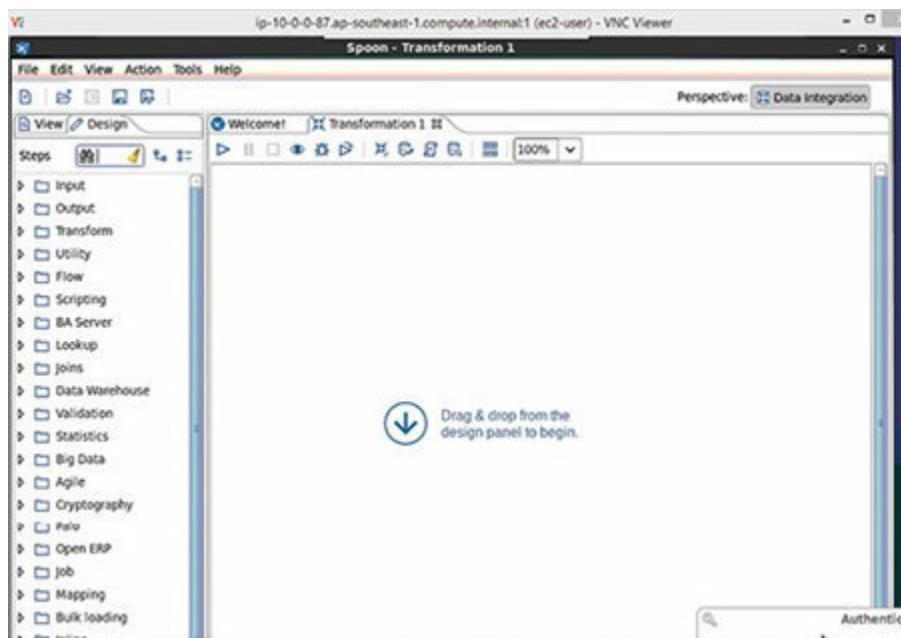
Once the tables are created the next logical step is to upload the data files to the respective tables. We will use PDI to upload the data into the tables. The basic steps is to connect the MS excel files located at predefined location, read the data from excel file and write the data into the table. 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.

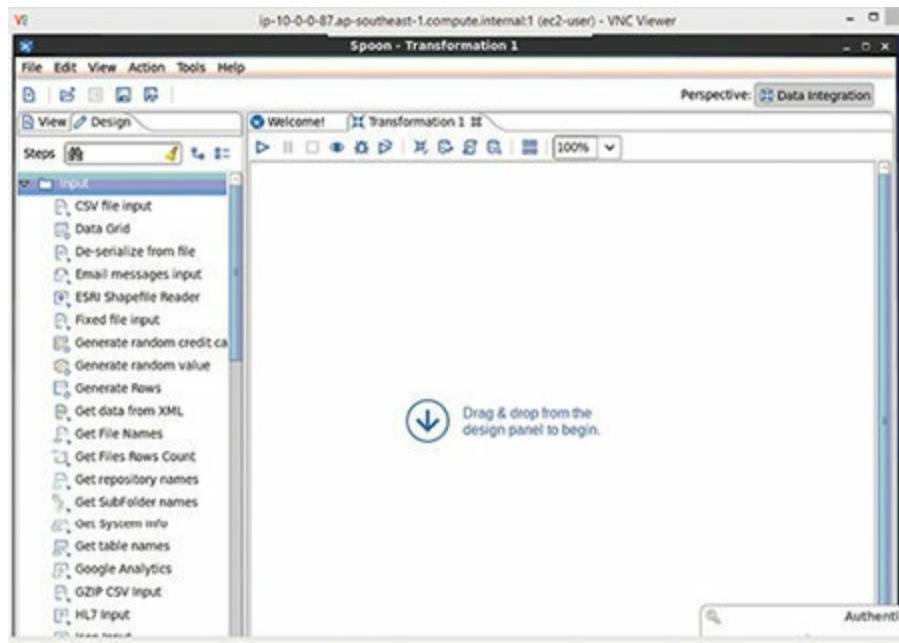
Go to File -> New -> Transformation



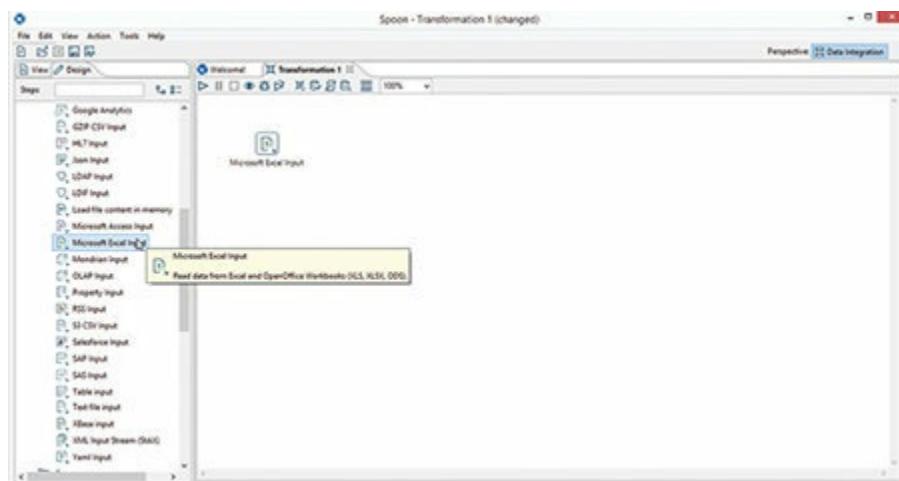
In left hand side you will see all the available elements classified as per the functions it performs. The transformation window is blank as of now.



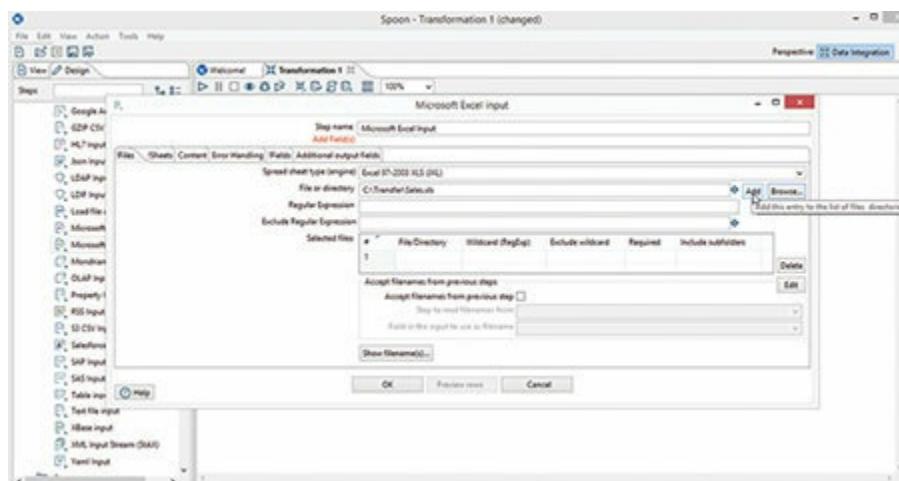
On expansion of input we can see all possible input file system.



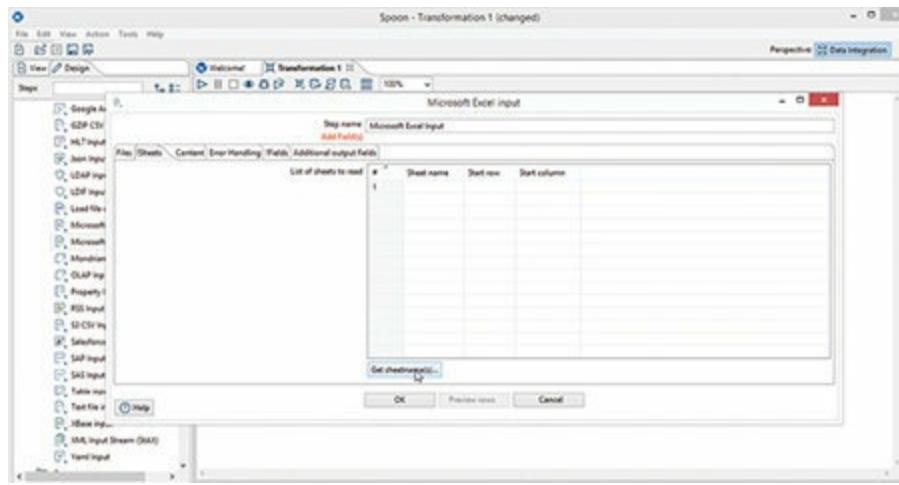
Look for Microsoft Excel Input from the input list, select it, drag and drop it in the transformation windows.



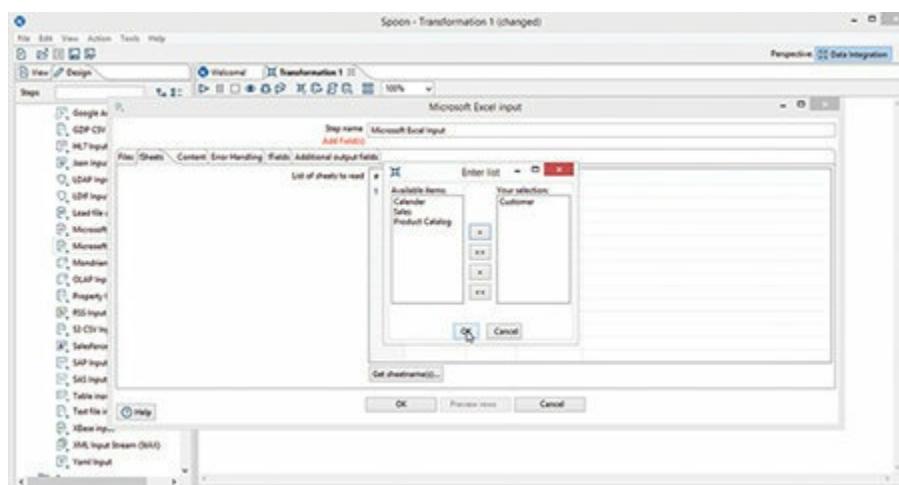
Click on Microsoft Excel Input the pop up will open. In the files tab select the file from the location and add it.



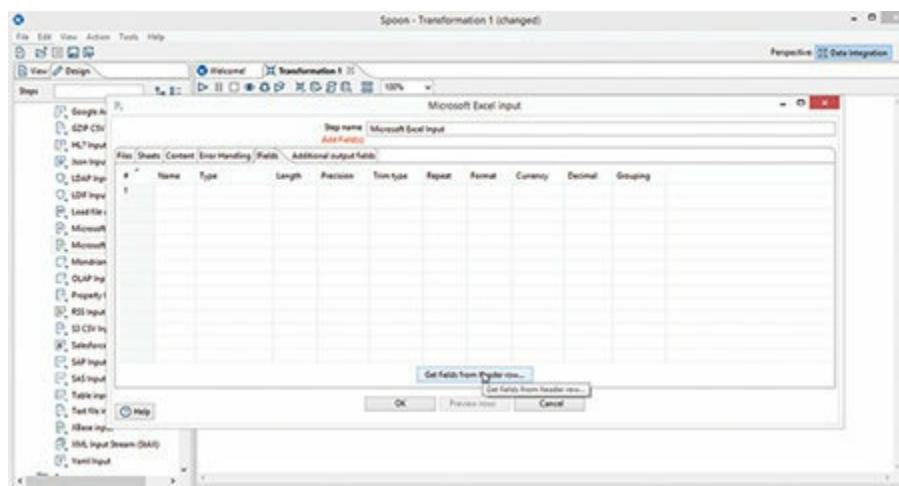
Go to Sheets tab, click on Get Sheet name.



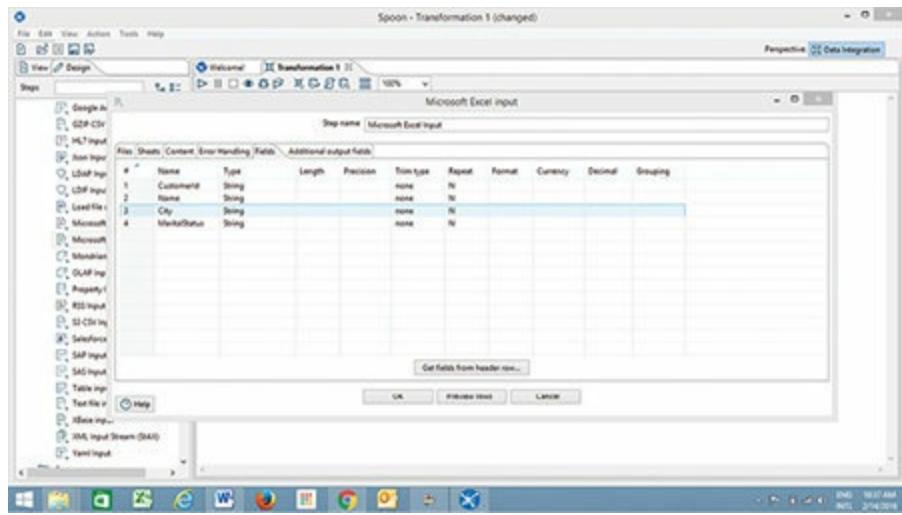
All the available sheet will show up. Select the correct sheet and transfer it into selection pane.



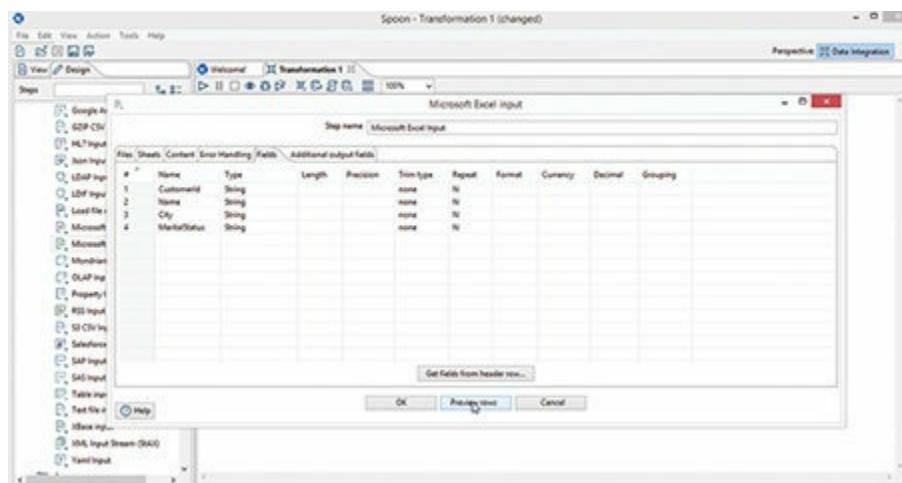
Go to Fields tab, click on get fields from headers rows



All the columns will show up with data type. Here you can modify the data type if required.



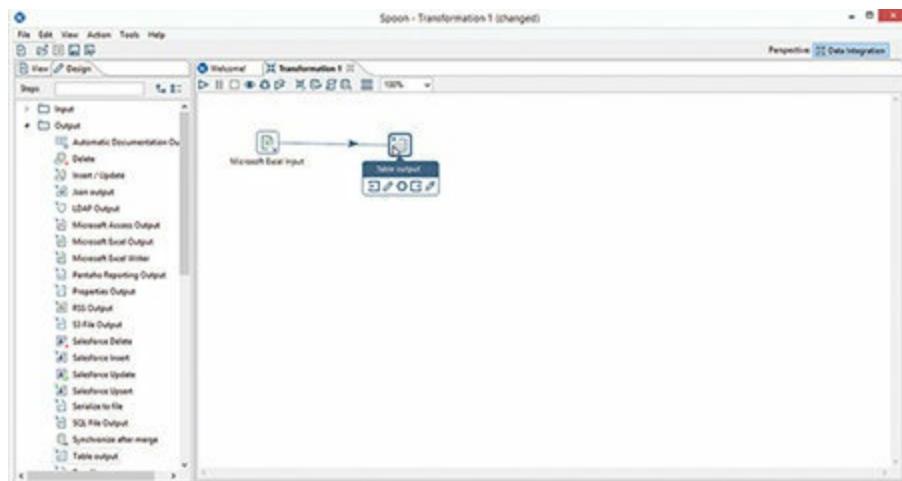
You can click on preview row to view sample data to ascertain the data accuracy and the format accuracy.



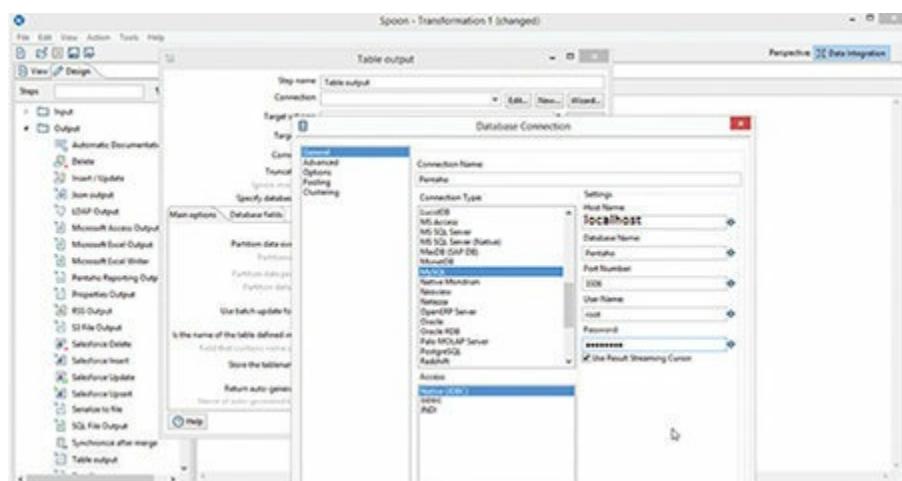
Output of previews of 4 rows

Rows of step Microsoft Excel Input (4 rows)				
#	CustomerId	Name	City	MaritalStatus
1	C000001	Customer1	New Delhi	Married
2	C000002	Customer2	Bangalore	Married
3	C000003	Customer3	Mumbai	Married
4	C000004	Customer4	Chennai	Single

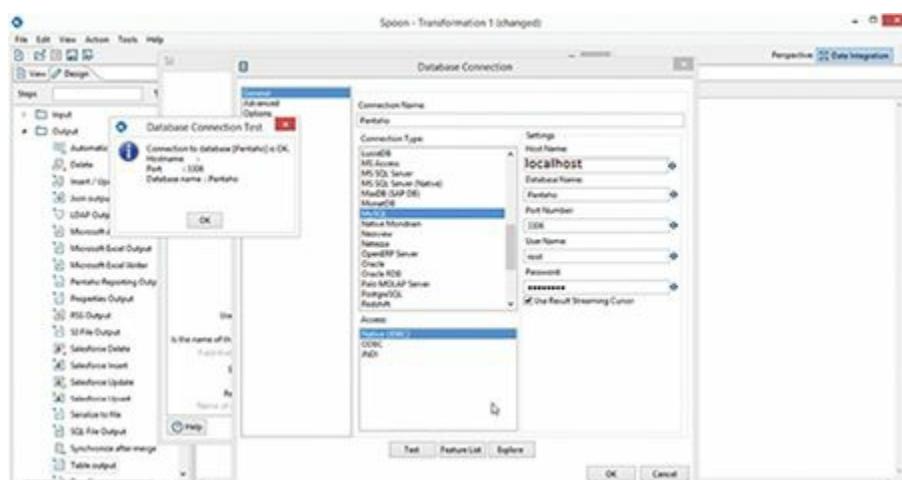
Go to Output, expand it. Select table output, drag and drop it into the transformation windows. Connect Microsoft Excel Input file to Table Output by keeping Shift Key Pressed and drag from Excel Inout to Table Output; the arrow will join showing the direction of the data flow.



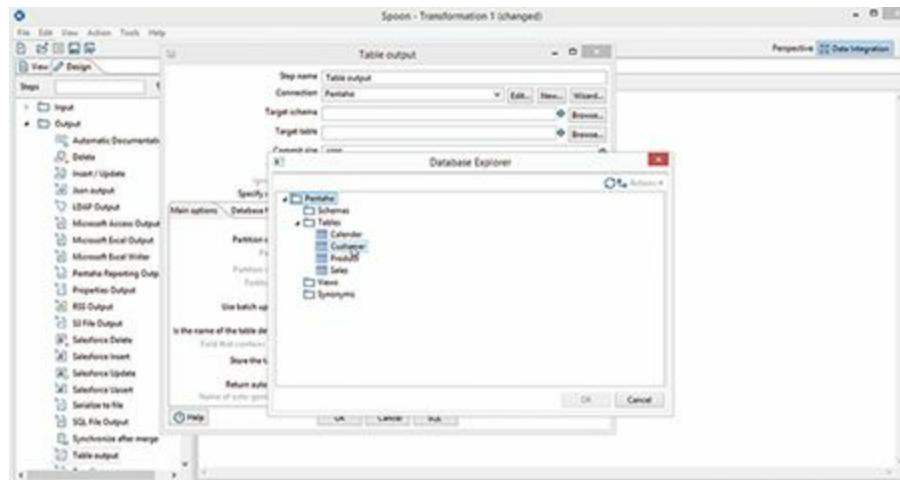
Click on Table Output, the pop-up will open. Since this is the first time we are connecting to the database, we have to create a JDBC connection. For that click on new in connection option; another pop-up will open. Select MySQL as database, add appropriate connection name. Here we called it Pentaho. 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.



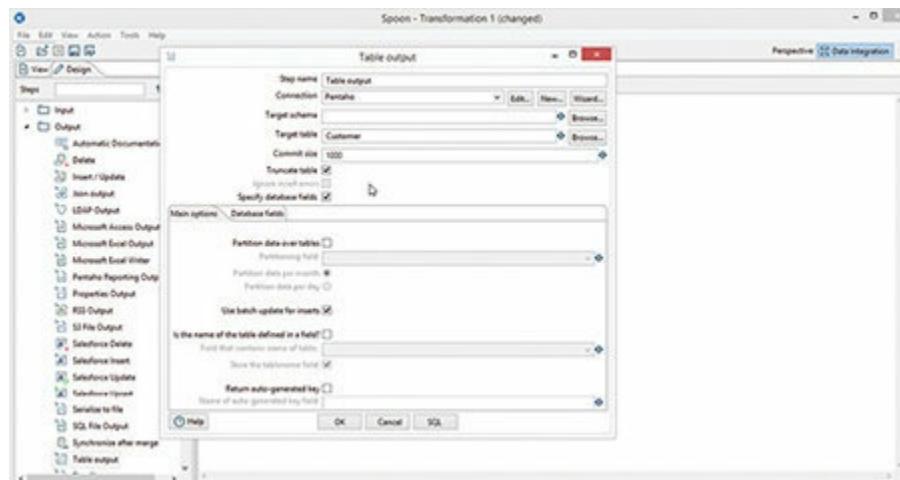
Click on test to check the connection.



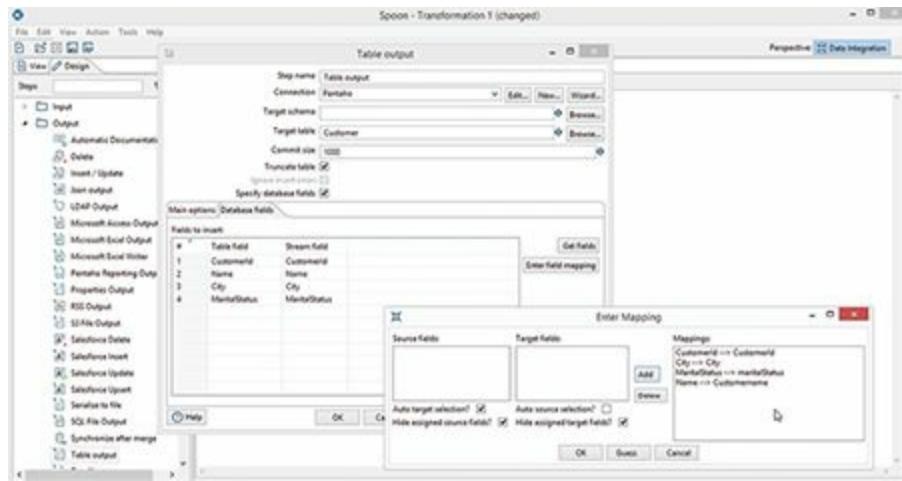
Once the connection is established, we have to select the table. For that click on Target Table browse button. Select the table **Customer**



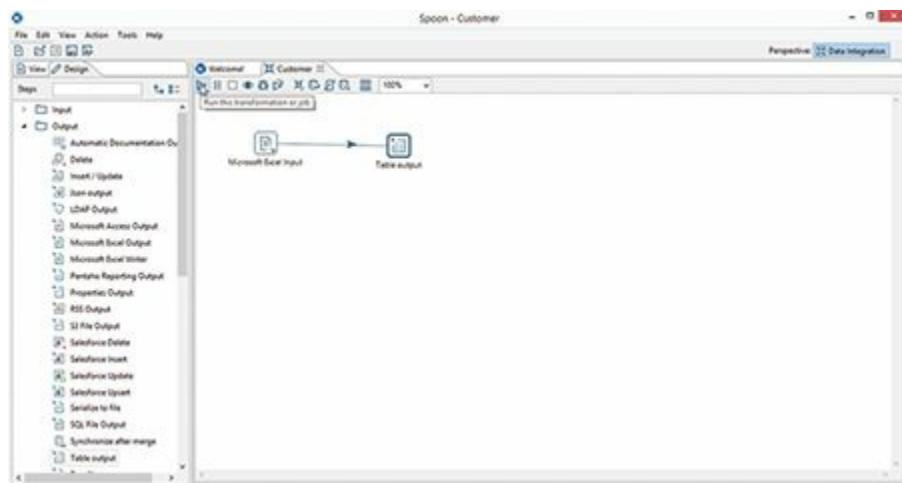
Tick the Truncate table option for the first time. However for subsequent upload Tick it only if fresh data is required. Tick the **Specify database Fields** to map the fields from excel to Table



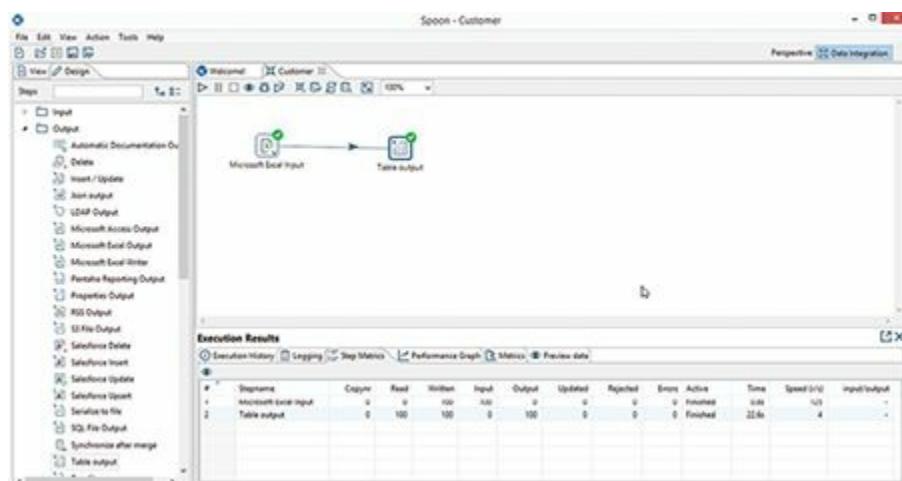
Click on Get fields to get the Fields from Incoming Input. Click on Enter field mapping to map fields one by one. If the fields name is same the system usually maps it automatically. For different field names you have to select the fields and map it.



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.

```

Database: Pentaho
Schema: root Username: root Connection ID: 527
select * from Customer limit 10

```

CustomerID	CustomerName	City	MaritalStatus
1 C0000001	Customer1	New Delhi	Married
2 C0000002	Customer2	Bangalore	Married
3 C0000003	Customer3	Mumbai	Married
4 C0000004	Customer4	Chennai	Single
5 C0000005	Customer5	Kolkata	Married
6 C0000006	Customer6	New Delhi	Single
7 C0000007	Customer7	Bangalore	Married
8 C0000008	Customer8	Mumbai	Single
9 C0000009	Customer9	Chennai	Single
10 C0000010	Customer10	Kolkata	Single

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

```

Database: Pentaho
Schema: root Username: root Connection ID: 527
select * from Calendar limit 10

```

Date	Year	FY_Year	Month_Name	YearMonth	Quarter	Day	Weekday
3/4/2015	2015	2015	Mar	2015_03	1	1	Wednesday
3/5/2015	2015	2015	Mar	2015_03	2	2	Thursday
3/6/2015	2015	2015	Mar	2015_03	3	3	Friday
3/7/2015	2015	2015	Mar	2015_03	4	4	Saturday
3/8/2015	2015	2015	Mar	2015_03	5	5	Sunday
3/9/2015	2015	2015	Mar	2015_03	6	6	Monday
3/10/2015	2015	2015	Mar	2015_03	7	7	Tuesday
3/11/2015	2015	2015	Mar	2015_03	8	8	Wednesday
3/12/2015	2015	2015	Mar	2015_03	9	9	Thursday
3/13/2015	2015	2015	Mar	2015_03	10	10	Friday
3/14/2015	2015	2015	Mar	2015_03	11		

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

```

Database: Pentaho
Schema: root Username: root Connection ID: 527
select * from Product limit 10

```

ProductID	ProductName	Category	SubCategory	Price
1 100000	Product name 1	Mobiles	Smartphones	20000
2 100001	Product name 2	Mobiles	Featured Phones	2400
3 100002	Product name 3	Mobiles	Smartphones	28000
4 100003	Product name 4	Mobiles	Smartphones	15000
5 100004	Product name 5	Mobiles	Featured Phones	18000
6 100005	Product name 6	Mobiles	Featured Phones	3400
7 100006	Product name 7	Mobiles	Featured Phones	3000
8 100007	Product name 8	Laptops	Windows	23000
9 100008	Product name 9	Laptops	Windows	35000
10 100009	Product name 10	Laptops	Windows	65000

### Sample Sales data

The screenshot shows a MySQL command-line interface window. At the top, it displays the connection details: Database: Pentaho, Schema: root, Username: root, Connection ID: 56. Below this, a SQL command is entered: `select * from Sales limit 10`. The results are displayed in a table with the following columns: OrderId, CustomerId, OrderDate, ProductId, Quantity, Amount, Discount, and Status. The data consists of 10 rows of sales information.

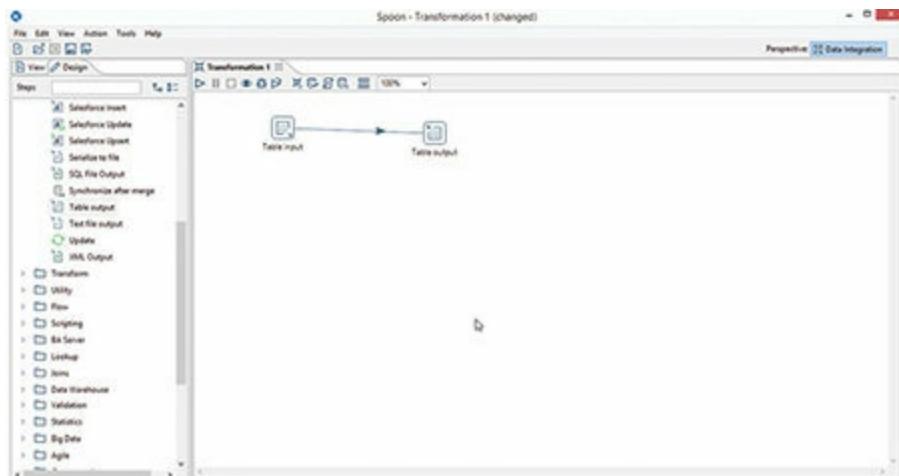
OrderId	CustomerId	OrderDate	ProductId	Quantity	Amount	Discount	Status
1000005	CC00045	4/1/2015	100015	2	1800	0	Returned
1000006	CC00043	4/2/2015	100012	2	1040	133	Completed
1000007	CC00048	4/3/2015	100008	2	70000	3500	Canceled
1000008	CC00044	4/4/2015	100017	3	2000	100	Completed
1000009	CC00049	4/5/2015	100013	2	3400	0	Completed
1000010	CC00050	4/6/2015	100011	1	100000	12000	Completed
1000011	CC00051	4/7/2015	100002	1	50000	0	Completed
1000012	CC00052	4/8/2015	100029	1	3000	300	Completed
1000013	CC00053	4/9/2015	100003	2	30000	3600	Returned
1000014	CC00054	4/10/2015	100002	3	50000	0	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      varchar(20) primary key,  
CustomerId    varchar(20),  
OrderDate    date,  
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.



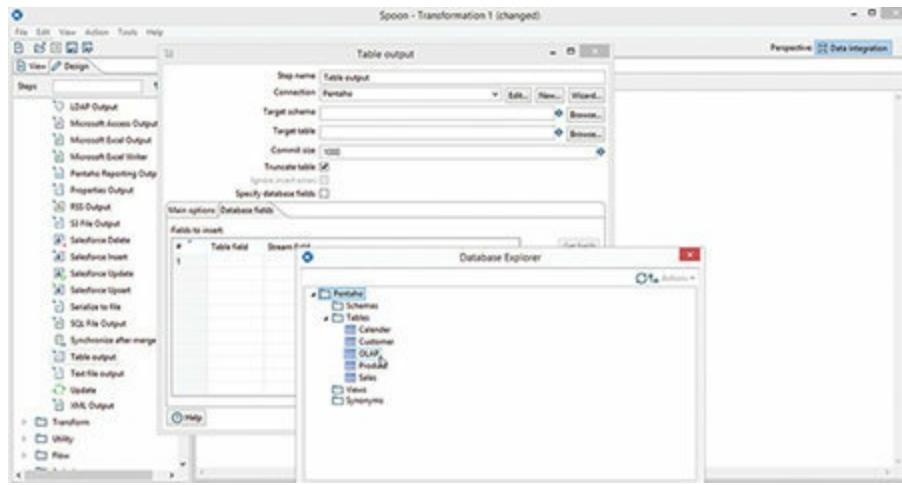
Connect the database using Pentaho connection we created in previous exercise. Add the Query into the SQL box. Here you can see I have added city from customer Table and margin is calculated as 10% of Sales Amount.



Data previews from the query

Root of step Table input (5 rows)								
#	OrderId	CustomerId	OrderDate	ProductId	Quantity	Amount	Discount	Status
1	O0000001	C0000001	2015-04-01 00:00:00.000	100013	2	1000	0	Returned
2	O0000002	C0000002	2015-04-01 00:00:00.000	100022	2	1040	125	Completed
3	O0000003	C0000003	2015-04-01 00:00:00.000	100013	2	1000	100	Hydrated
4	O0000004	C0000004	2015-04-01 00:00:00.000	100013	3	1500	180	Completed
5	O0000005	C0000005	2015-04-01 00:00:00.000	100013	3	1500	0	Completed

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

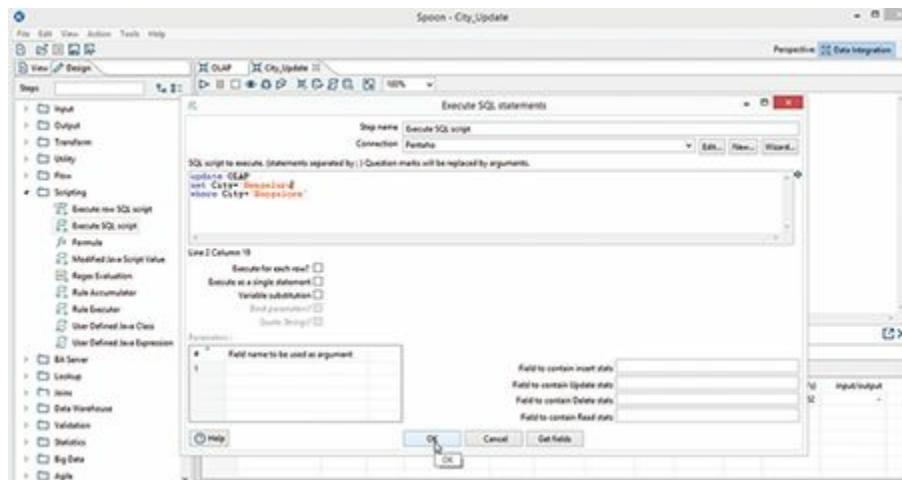


Save the transformation and execute it. The result from query windows.

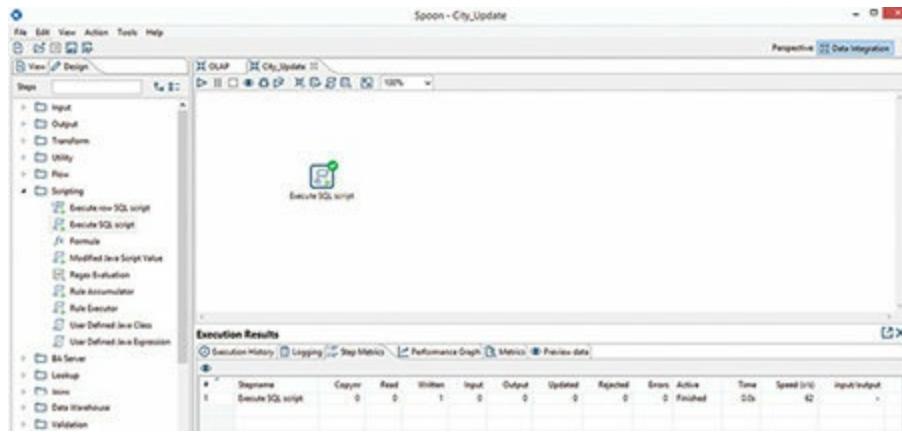
1/28 [DB: [2/14/16 11:17:42 AM EST]] Script executed - (0) Errors (Time: 29ms )									
OrderID	CustomerID	OrderDate	ProductID	Quantity	Amount	Discount	Status	City	Region
1/00000001	C000047	4/1/2015	100015	2	1800	0	Shipped	New Delhi	100
2/00000002	C000043	4/2/2015	100022	2	1040	115	Completed	Mumbai	204
3/00000003	C000098	4/3/2015	100008	2	7000	300	Completed	Mumbai	7000
4/00000004	C000048	4/4/2015	100018	3	8400	180	Completed	Mumbai	360
5/00000005	C000049	4/5/2015	100019	2	3200	0	Shipped	Kolkata	160
6/00000006	C000046	4/6/2015	100011	1	12000	12000	Completed	New Delhi	10000
7/00000007	C000043	4/7/2015	100002	1	3000	0	Completed	New Delhi	3000
8/00000008	C000087	4/8/2015	100028	1	3005	365	Completed	Bangalore	300
9/00000009	C000047	4/9/2015	100003	3	30000	3600	Shipped	Bangalore	3000
10/00000010	C000094	4/10/2015	100002	3	58000	0	Completed	Chennai	5000

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



Execute the update transformation



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

Database: Pentaho Schema root: Username: host Connection ID: 591											
select * from GAP_Sales											
1:28 [DB] [2/14/16 11:21:00 AM IST] Script executed = (0) Errors [Time: 23sec ]											
OrderDate	CustomerID	OrderDate	ProductID	Quantity	Amount	Discount	Status	City	Margin		
2/10/2005	CUSTOMER4	4/1/2015	100015	2	1000	0	Shipped	New Delhi	100		
2/10/2005	CUSTOMER3	4/2/2015	100002	2	1040	125	Completed	Hyderabad	104		
2/10/2005	CUSTOMER4	4/3/2015	100006	2	70000	3500	Cancelled	Hyderabad	7000		
2/10/2005	CUSTOMER4	4/4/2015	100015	3	3400	180	Completed	Hyderabad	340		
2/10/2005	CUSTOMER4	4/5/2015	100013	3	3400	0	Completed	New Delhi	340		
2/10/2005	CUSTOMER6	4/6/2015	100011	1	200000	12000	Completed	New Delhi	180000		
2/10/2005	CUSTOMER7	4/7/2015	100002	1	50000	0	Completed	New Delhi	50000		
2/10/2005	CUSTOMER8	4/8/2015	100029	1	20000	300	Completed	Bangalore	19700		
2/10/2005	CUSTOMER9	4/9/2015	100002	2	30000	3000	Returned	Bangalore	29700		
2/10/2005	CUSTOMER4	4/10/2015	100002	2	50000	0	Completed	Chennai	50000		

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

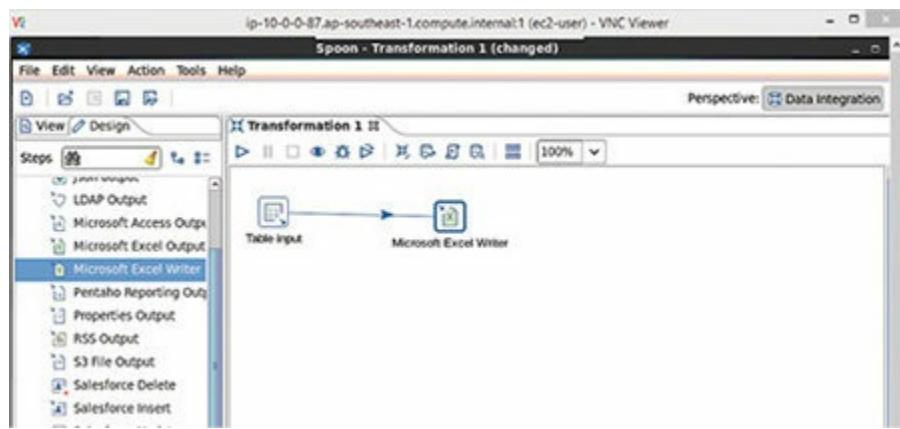
```

---

```

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

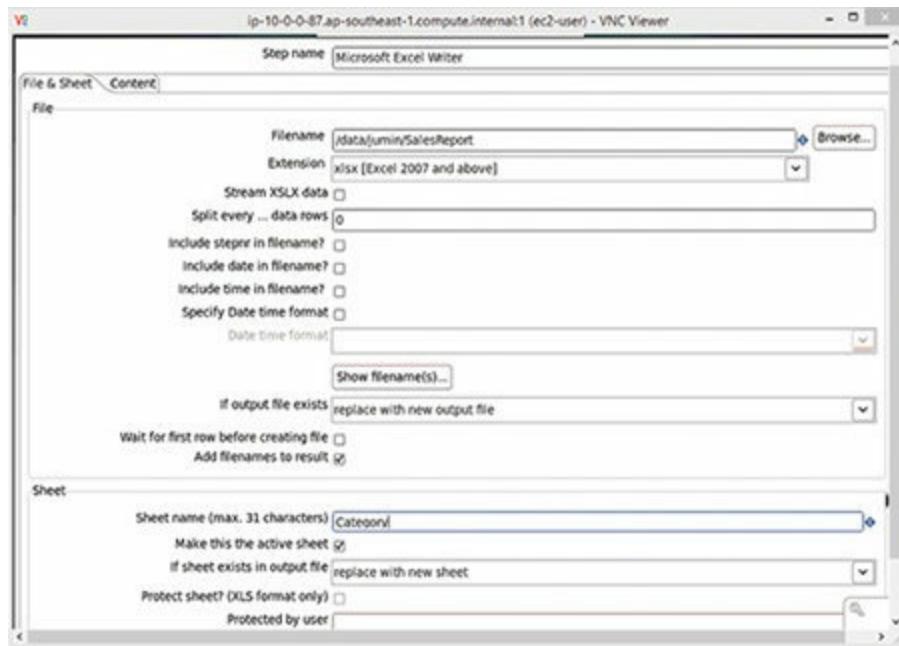
```



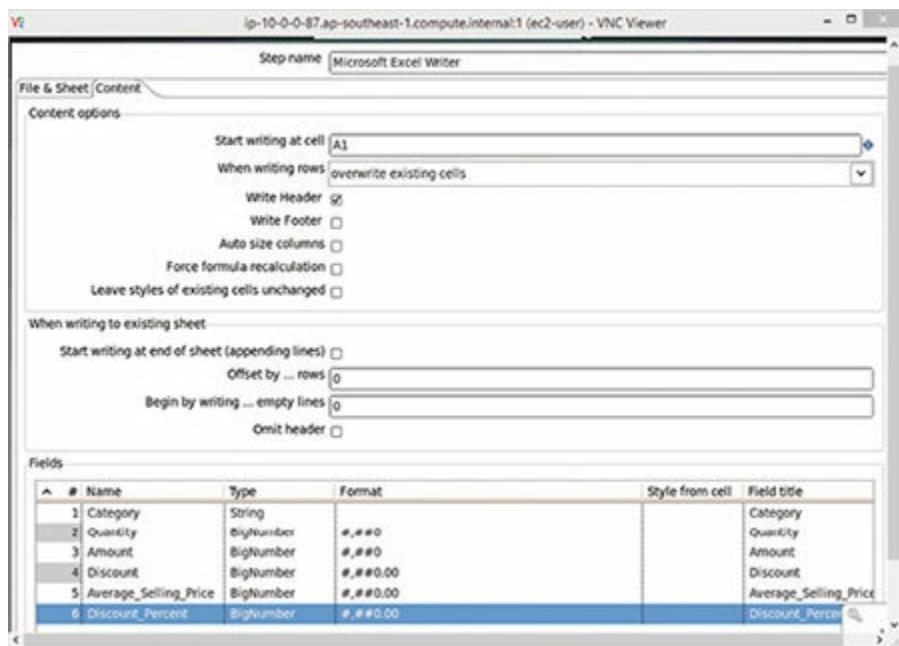
Select a table Input and Microsoft Excel Output. Add Query into the Table Input and check the output using preview rows option.

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.

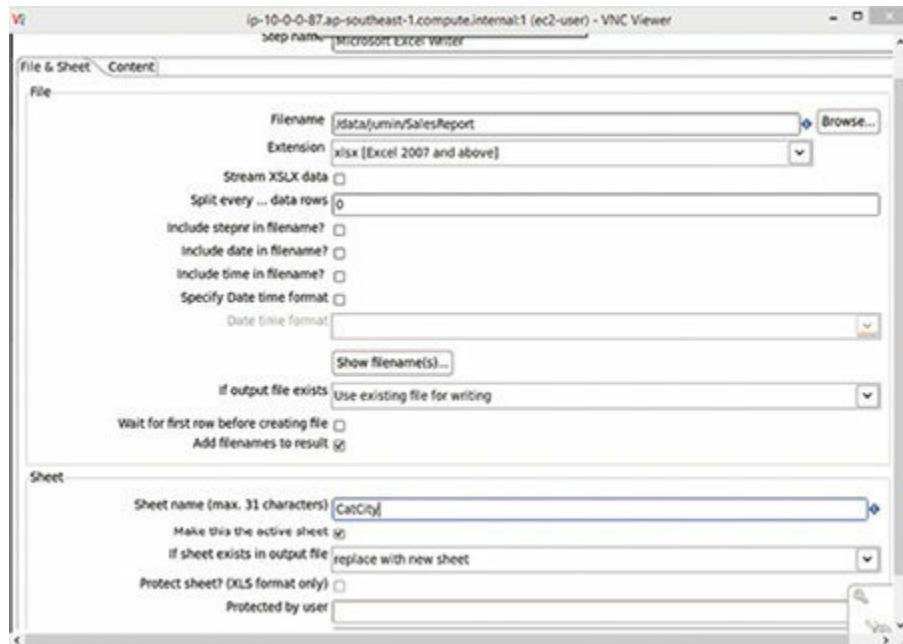


In Content tab select the starting Cell where Report will start. Tick Header option if there are header rows in the report. Scroll down to the end, click on add columns. The entire column will be shown with their data type. You can remove some of the column if not required. Change the data format as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

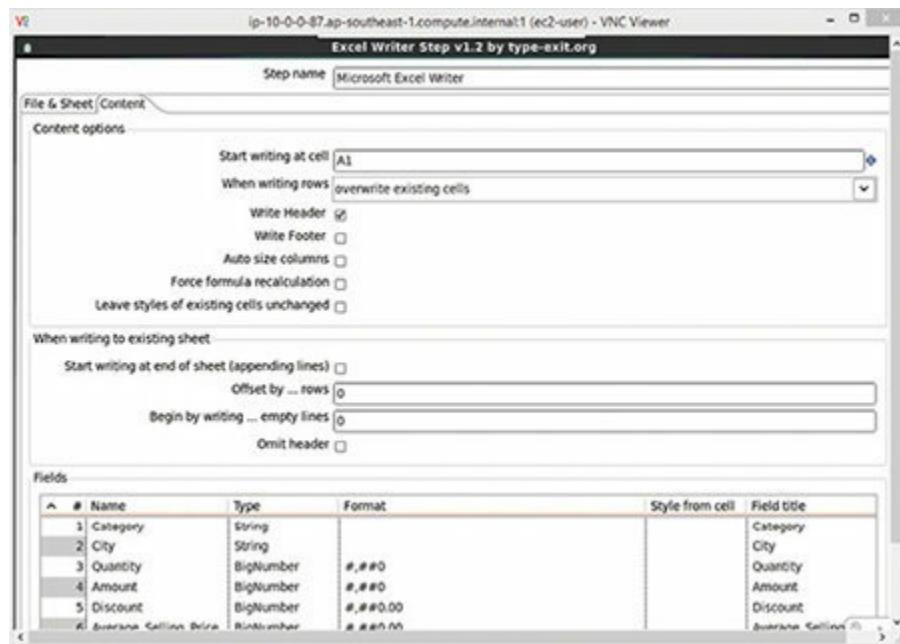


Similarly for category city report create a transformation, add the second query in the Table Input. In the excel output file tab enter the sale file name, select “*Use existing file for writing*”. This is because we create a new file in

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

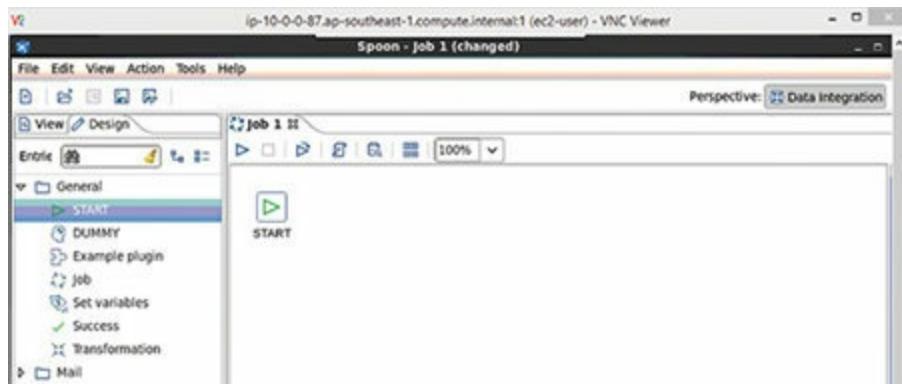


Add your fields and do formatting.

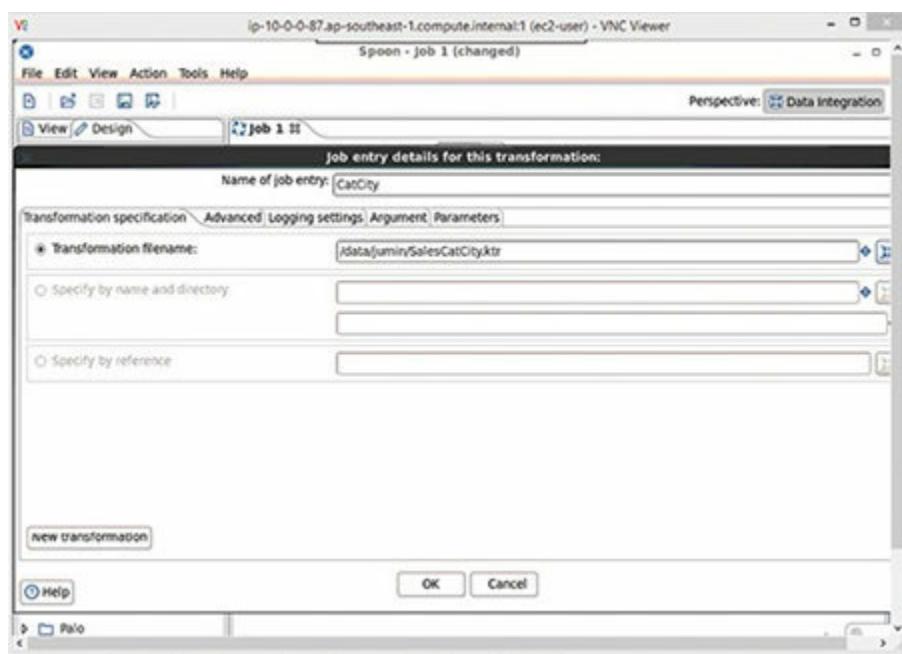
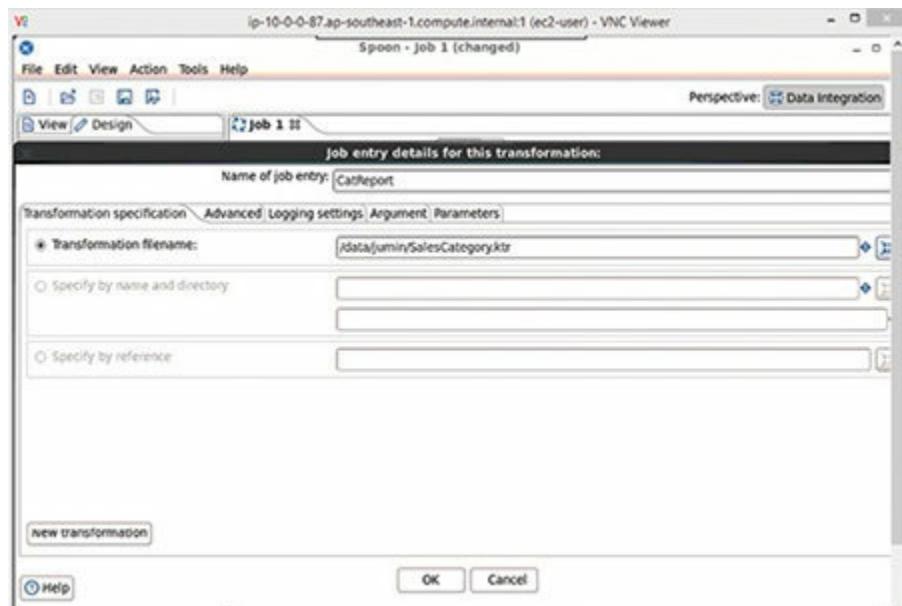


Now that we have created two transformations, we have to join both transformations together. This is done through job. Go to File -> New -> job

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.



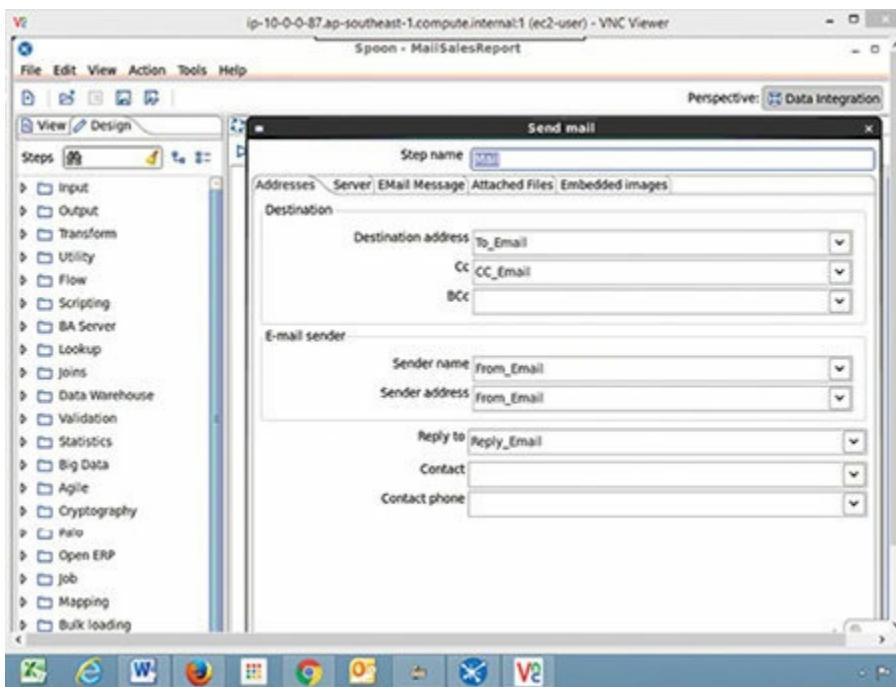
The job is ready not but one element is missing – a system to mail the report to the users. For the mailing system create a transformation, add a Table input

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.

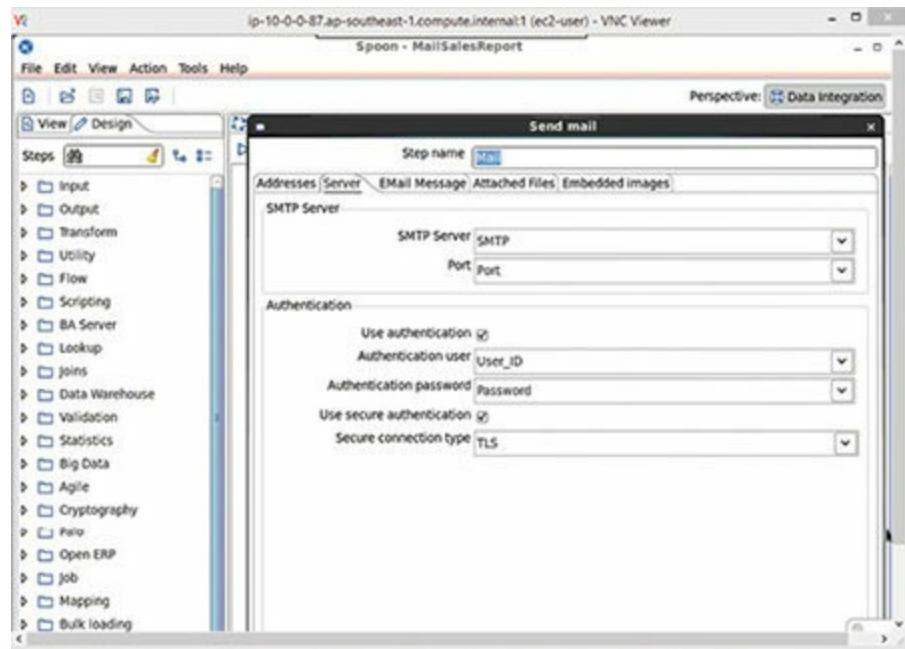


```
Step name: Table input
Connection: Prod
SQL:
SELECT * FROM "companyname.com" To_Email
, * FROM "companyname.com" From_Email
, * FROM "companyname.com" Sender
, * FROM "companyname.com" Reply_Email
, * FROM "companyname.com" Subject
, * FROM "companyname.com" Body
, * FROM "companyname.com" Attachment
, * FROM "companyname.com" SMTP
, * FROM "companyname.com" Port
, * FROM "companyname.com" User_ID
, * FROM "companyname.com" Password
```

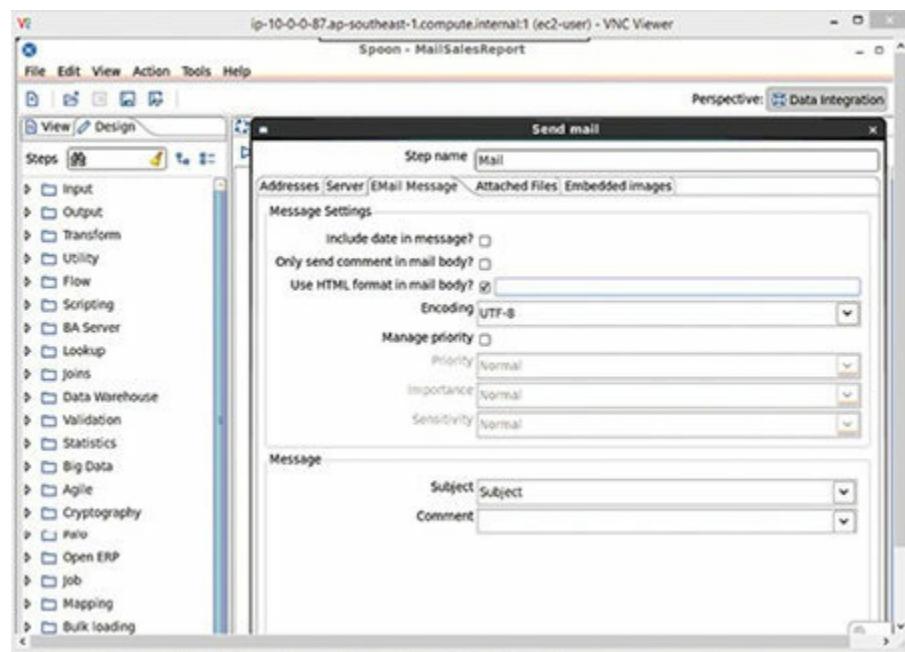
Map table input to the mail elements. Add from the dropdown the correct fields defined in the query.



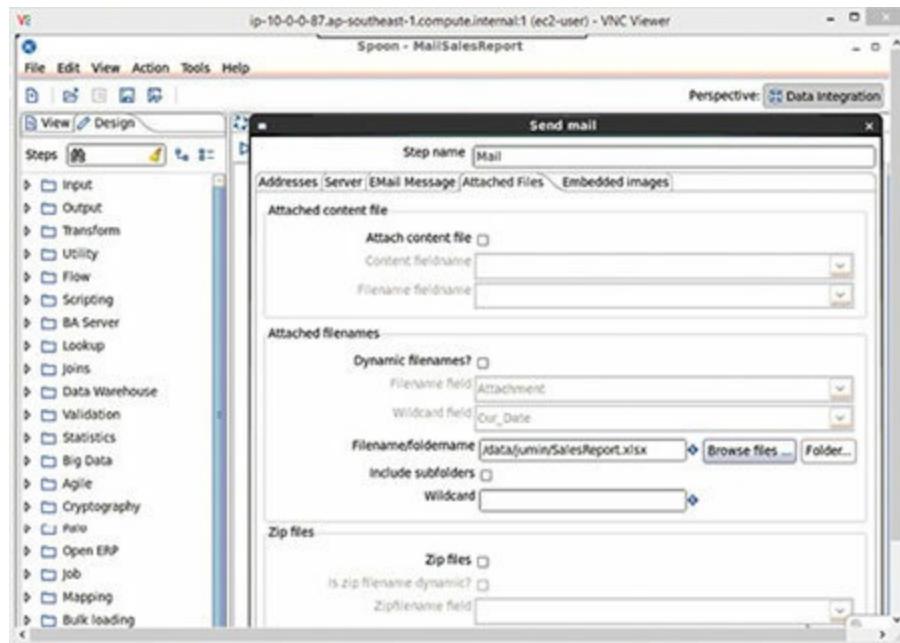
Add SMTP, Port, userid and password. Select secure authentication as your mailing server configuration.



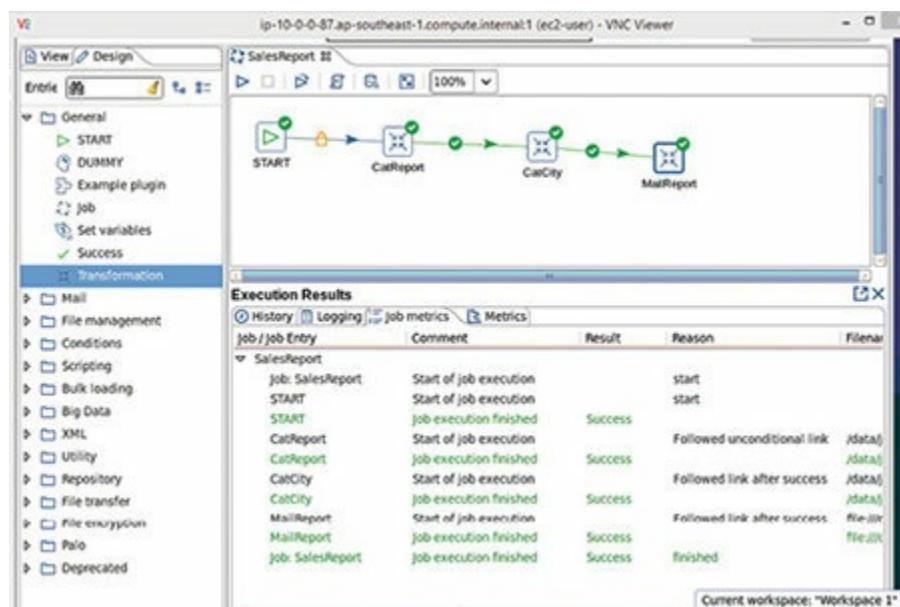
Add subject and mail body if any in comment.



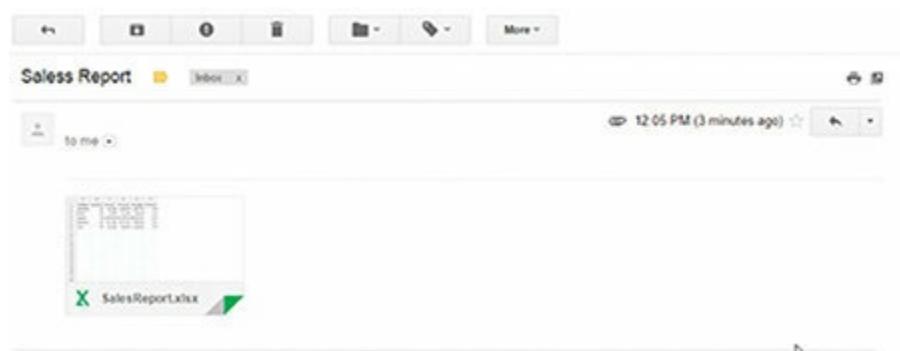
Add attachment of the mail from the output file from previous two transformations.



Save the Job and execute it.



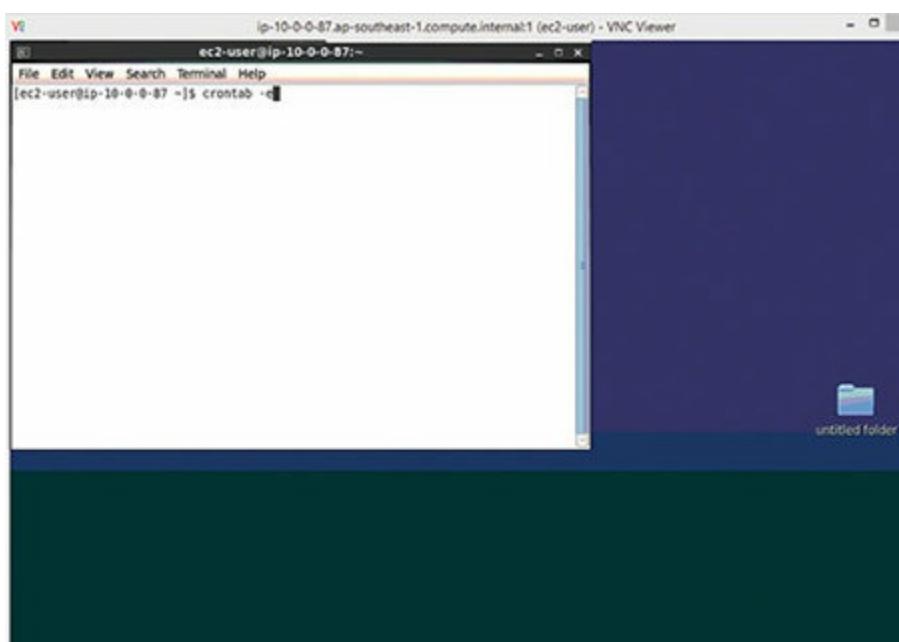
The mail received in the mailbox



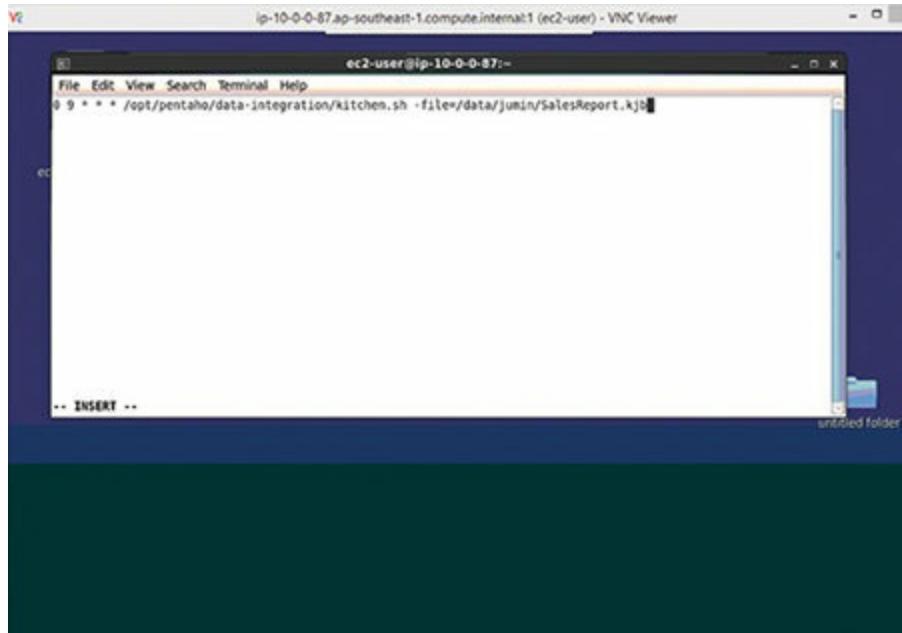
The file output is as shown below

The image contains two side-by-side screenshots of Microsoft Excel spreadsheets. Both are titled 'SalesReport - Microsoft Excel'.  
The top spreadsheet has columns A through K. It includes a header row and data rows 2 through 6. The data shows sales for categories like Apparel, Books, Laptop, Mobiles, and Toys across different quantities and amounts.  
The bottom spreadsheet also has columns A through K. It includes a header row and data rows 2 through 25. This version of the report includes an additional column 'City' and shows sales data grouped by both Category and City.

The job can be scheduled using cron job in the linux



In the crontab –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.

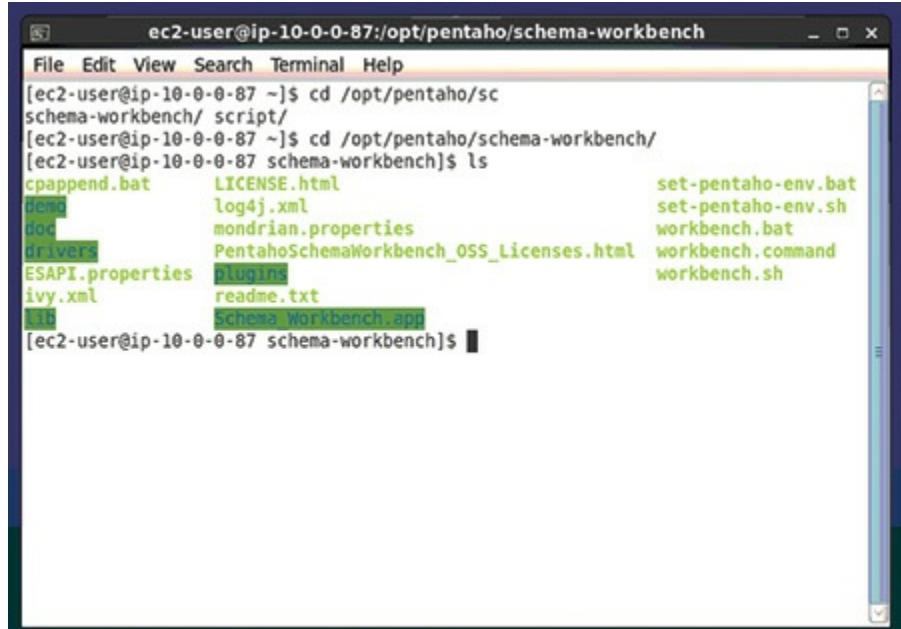
### 2.1.2 Mondrian OLAP Cube

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

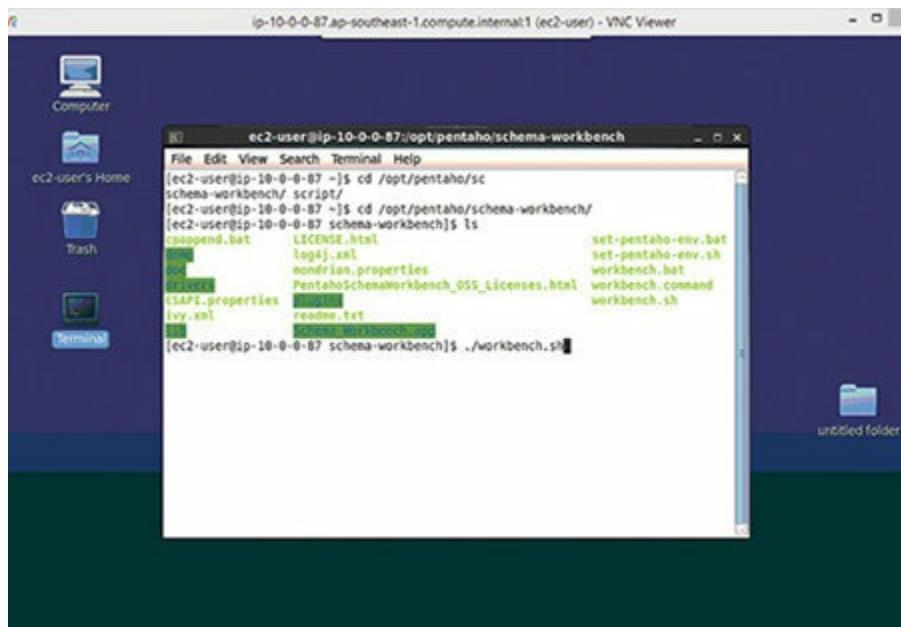
like Category can be shared by Sales Cube, Purchase Cube, and Inventory Cube and so on.

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.



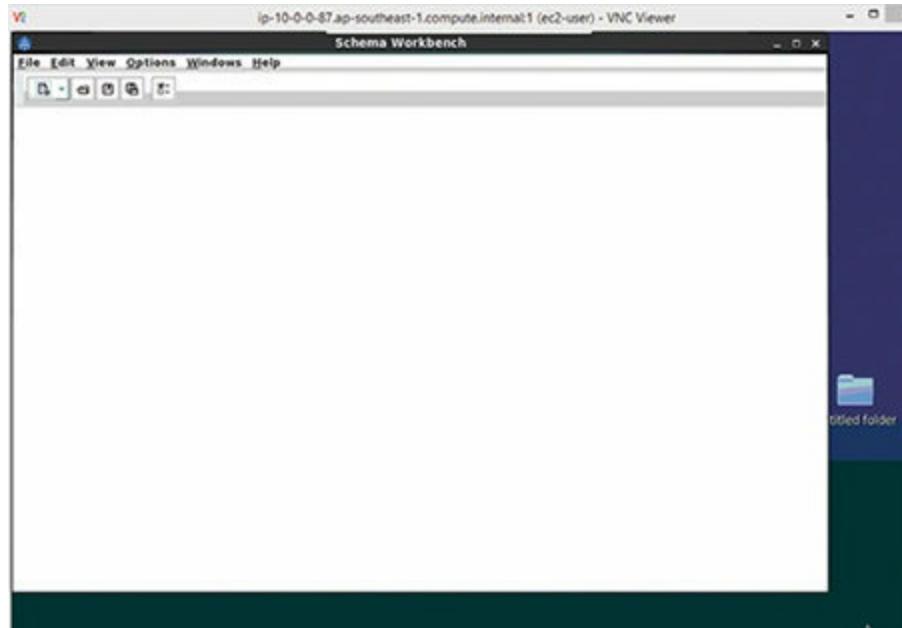
```
ec2-user@ip-10-0-0-87:/opt/pentaho/schema-workbench
File Edit View Search Terminal Help
[ec2-user@ip-10-0-0-87 ~]$ cd /opt/pentaho/sc
schema-workbench/ script/
[ec2-user@ip-10-0-0-87 ~]$ cd /opt/pentaho/schema-workbench/
[ec2-user@ip-10-0-0-87 schema-workbench]$ ls
cpappend.bat      LICENSE.html
demo             log4j.xml
doc              mondrian.properties
drivers          PentahoSchemaWorkbench_OSS_Licenses.html
ESAPI.properties  plugins
ivy.xml           readme.txt
lib              schema Workbench.app
set-pentaho-env.bat
set-pentaho-env.sh
workbench.bat
workbench.command
workbench.sh
[ec2-user@ip-10-0-0-87 schema-workbench]$
```

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

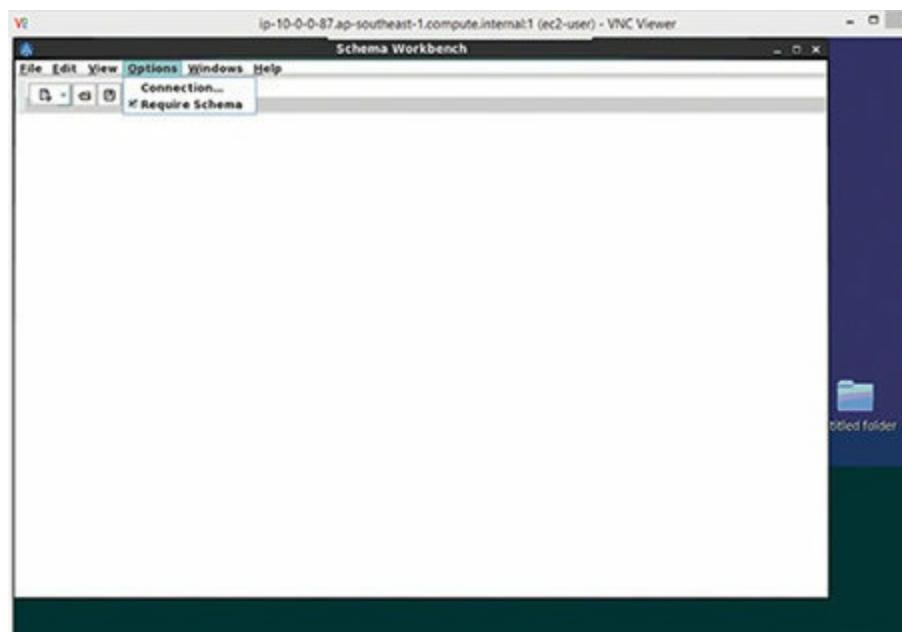


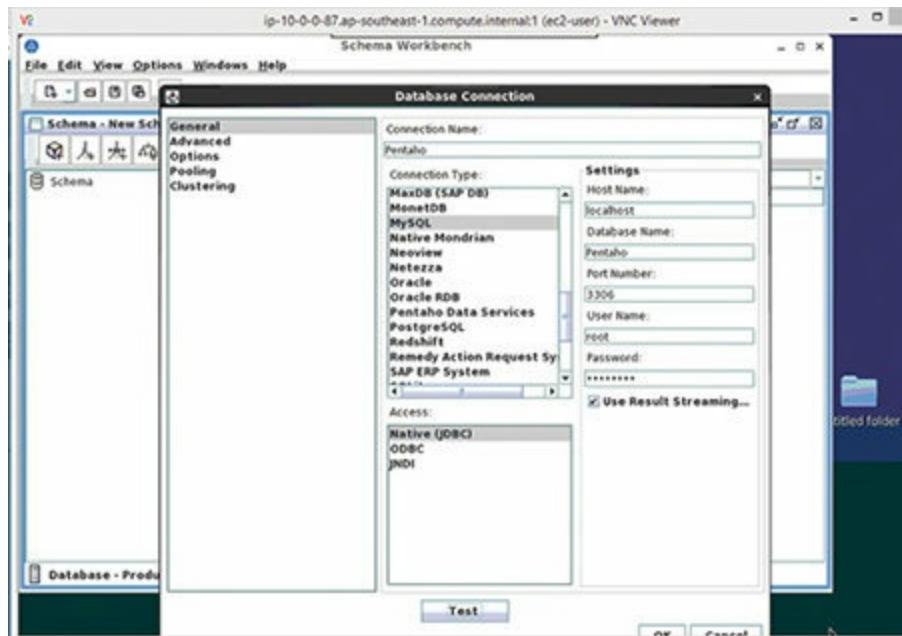
```
ip-10-0-0-87.ap-southeast-1.compute.internal (ec2-user) - VNC Viewer
File Edit View Search Terminal Help
[ec2-user@ip-10-0-0-87 ~]$ cd /opt/pentaho/sc
schema-workbench/ script/
[ec2-user@ip-10-0-0-87 ~]$ cd /opt/pentaho/schema-workbench/
[ec2-user@ip-10-0-0-87 schema-workbench]$ ls
cpappend.bat      LICENSE.html
demo             log4j.xml
doc              mondrian.properties
drivers          PentahoSchemaWorkbench_OSS_Licenses.html
ESAPI.properties  plugins
ivy.xml           readme.txt
lib              schema Workbench.app
set-pentaho-env.bat
set-pentaho-env.sh
workbench.bat
workbench.command
workbench.sh
[ec2-user@ip-10-0-0-87 schema-workbench]$ ./workbench.sh
```

The Mondrian windows look like below.

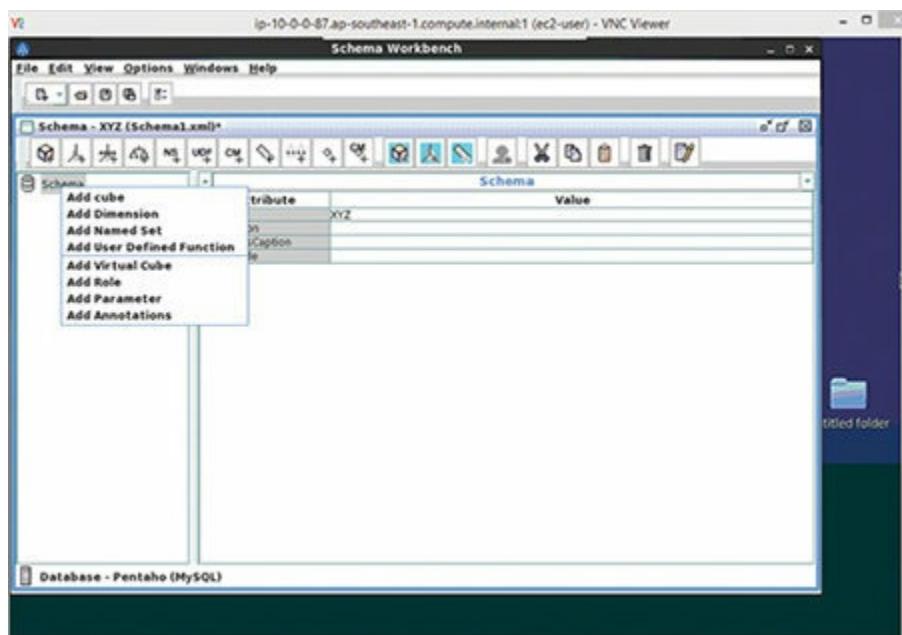


First you have to connect the database. Here we connect to Pentaho database using same credential as we used in PDI

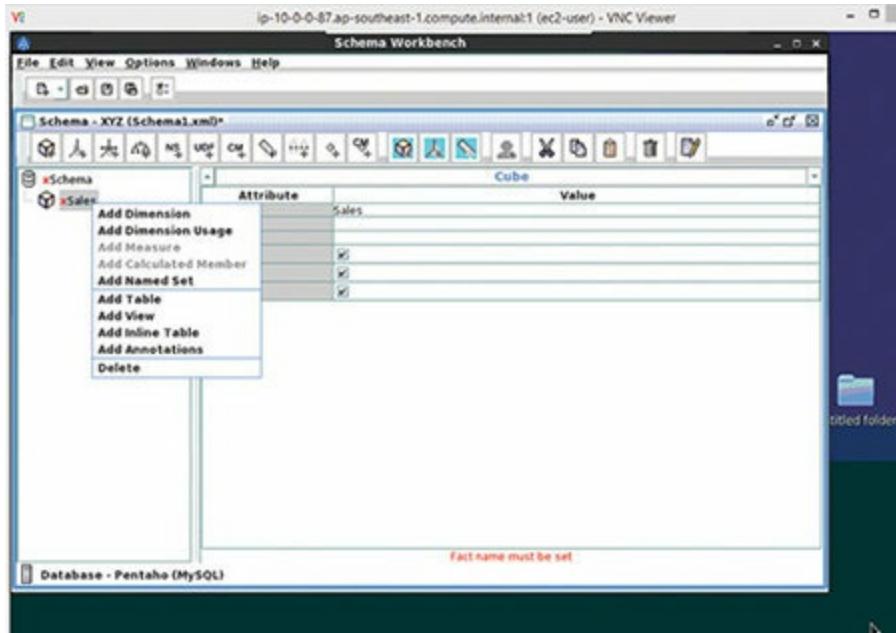




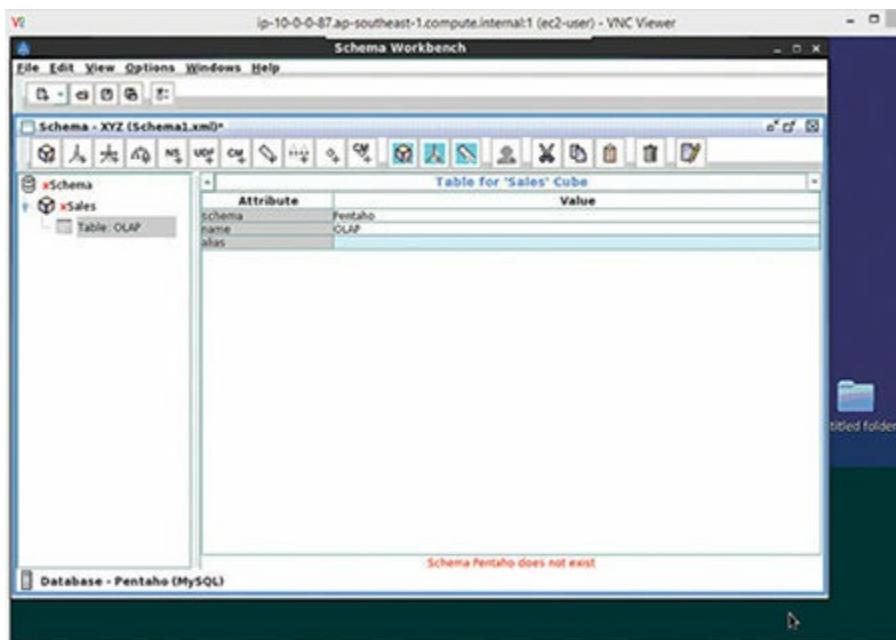
Once connection is tested we are ready for designing a cube. Click on Schema and name it by company name. Here I named it XYZ. Right click on Schema you will see lots of add items list. To start with Click on add cube. Name the cube as Sales.



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.

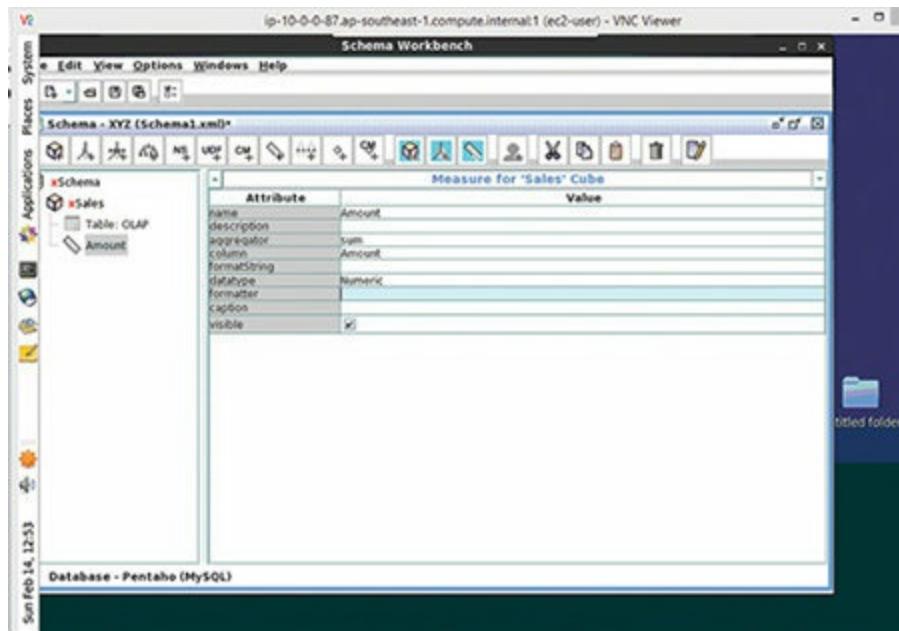


In the Table space, add database name Pentaho in the schema and Table name OLAP in name. If you want to keep an alias of table you can add alias name.

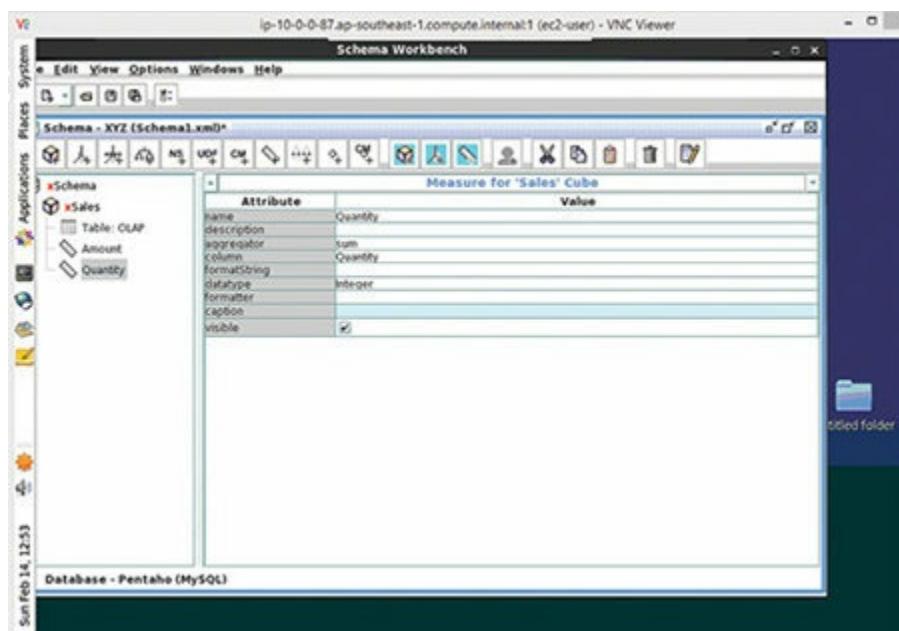


Right click on Sales cube you will have option to add Dimensions and Metrics (Measures). You can start with any one of them. Here I will start with the measures.

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.

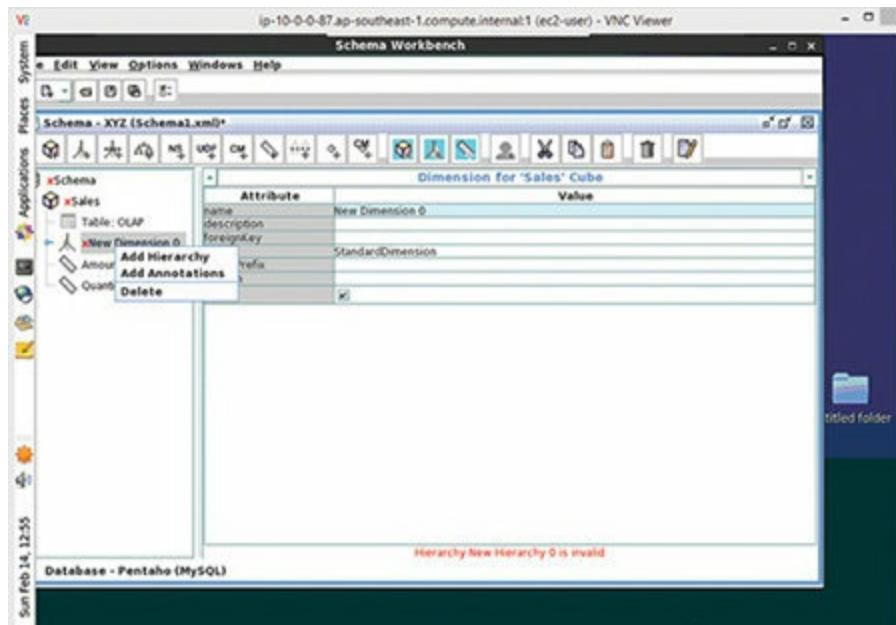


Similarly add Sales Quantity as shown below.

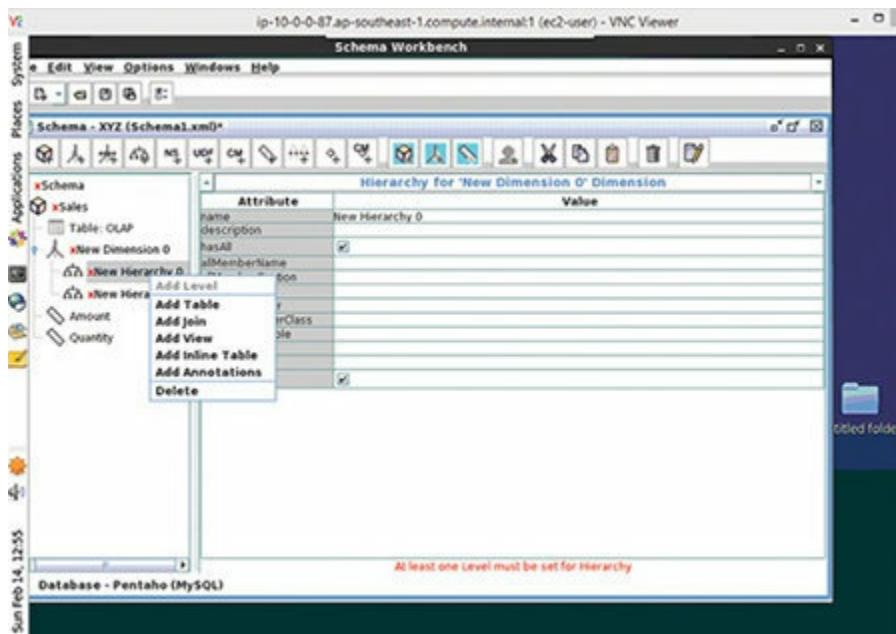


Now that we have added two measures lets learn how to add dimensions.  
Right click on Sales cube, select Add Dimensions.

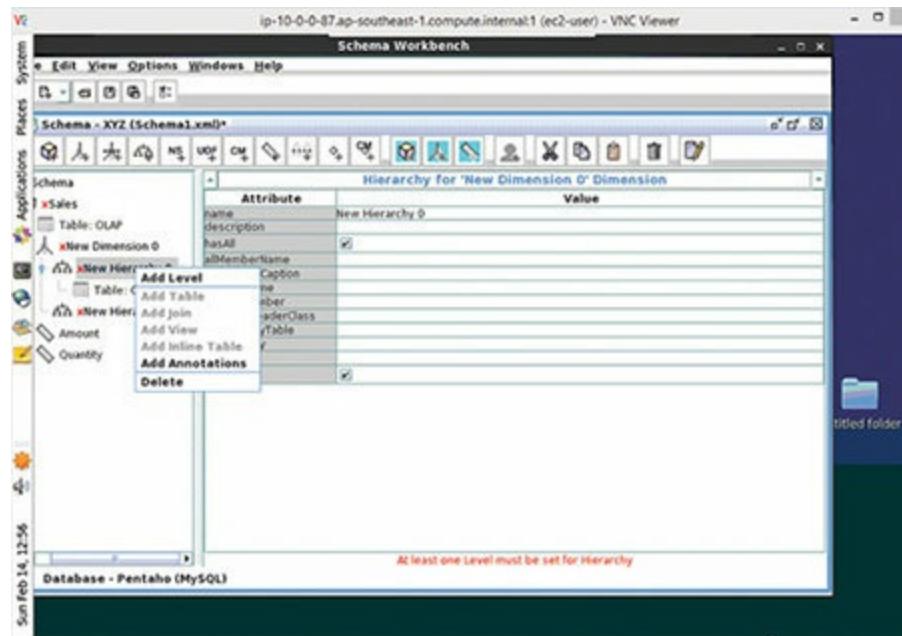
Right click on Dimensions and select Add Hierarchy.



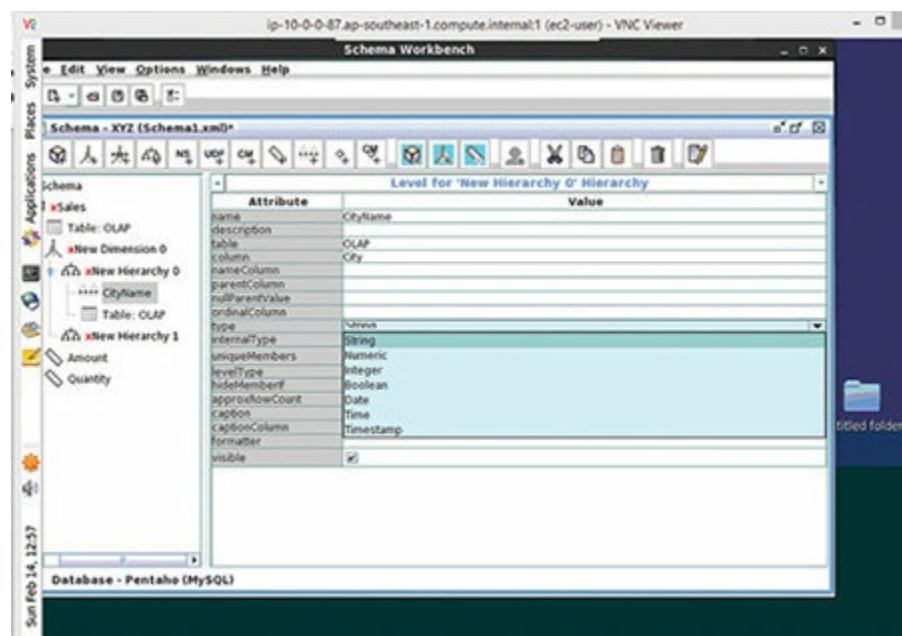
You will see two Hierarchies in the dimensions. You can delete one of them if not required. Right Click on Hierarchy and select add Table



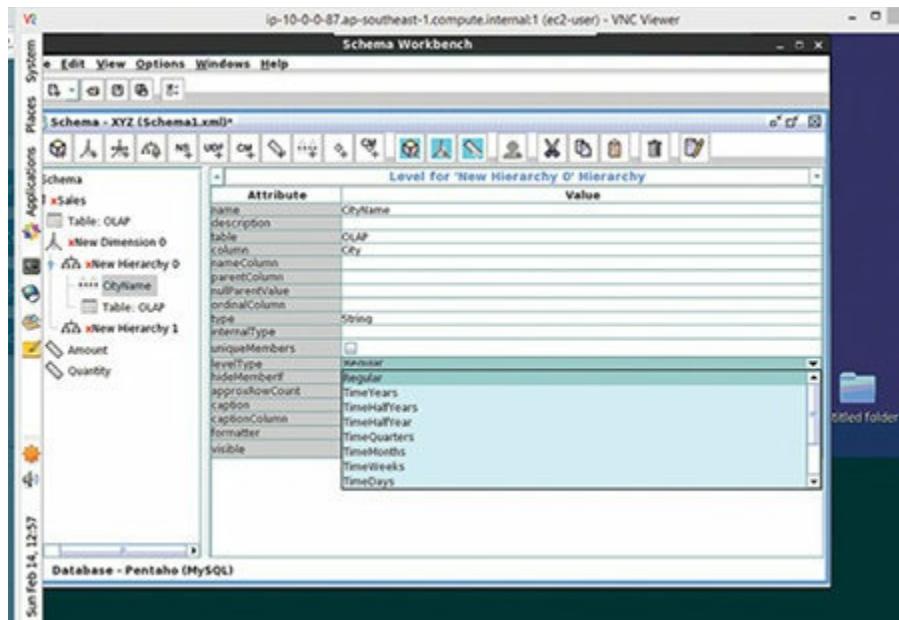
Again right click on Hierarchy and select add Level. The table name is not actually required as we are creating dimension out of same Table as Sales cube. The table has been added so that we can add level. Once level is added delete the Table immediately to avoid conflict with cube Table.



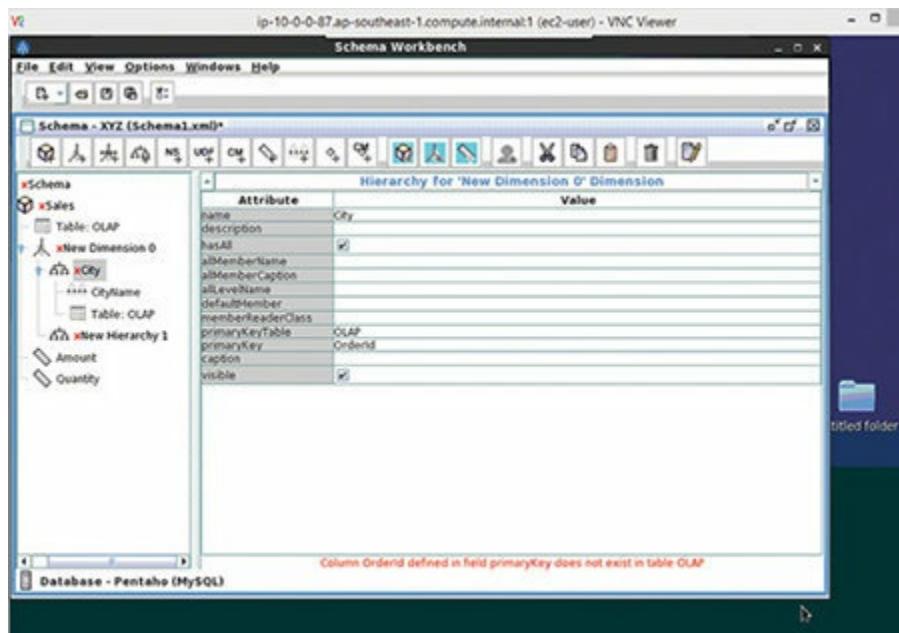
In the level add the Level Name, Table as OLAP, column as city and in Type add as string.



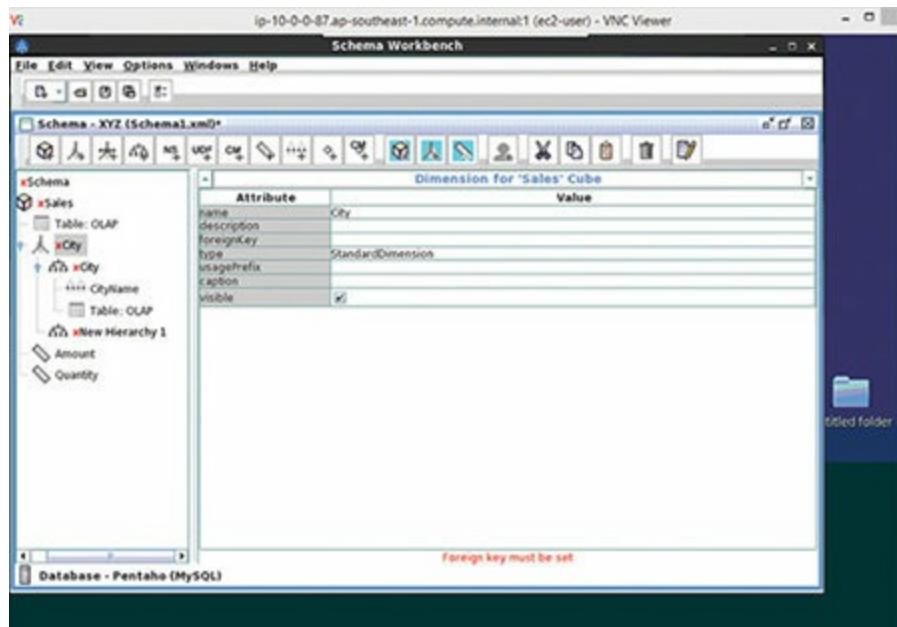
Select Level Type as Regular. Rest of the fields can be blanks as we can do without those values. You may fill values if required.



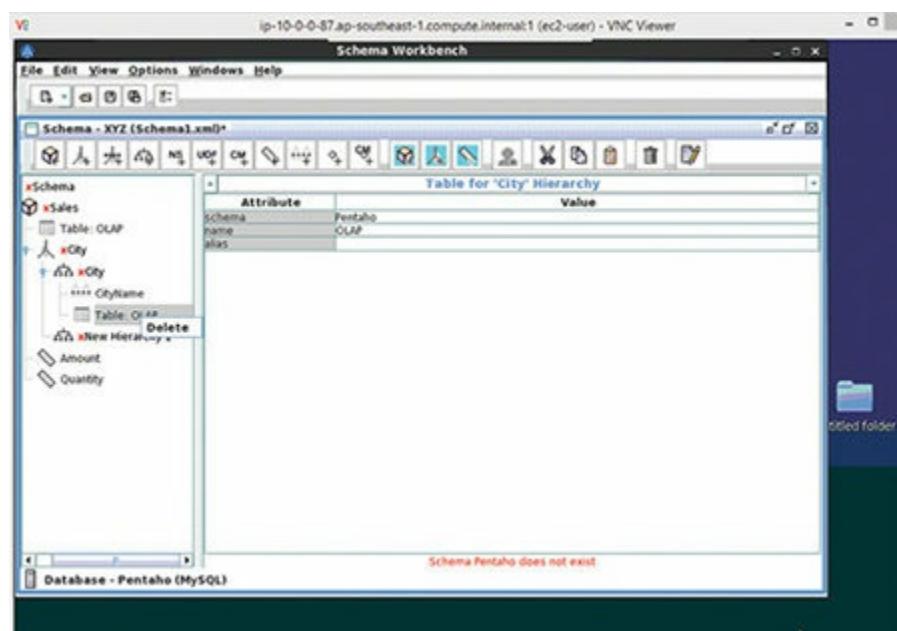
Now add values in the Hierarchy. Here again add name, table name and the table primary key as shown below.



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.

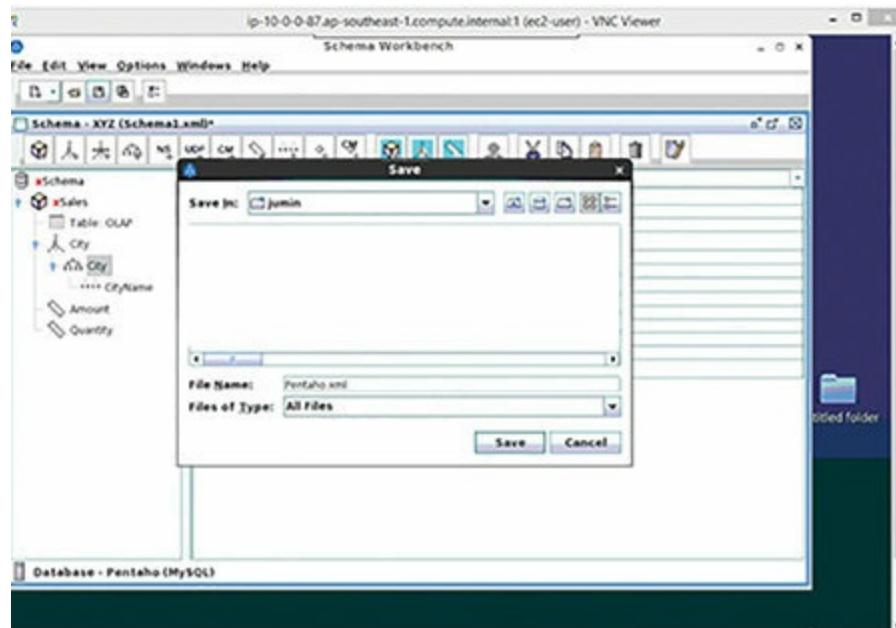


Finally delete the Table from dimensions.

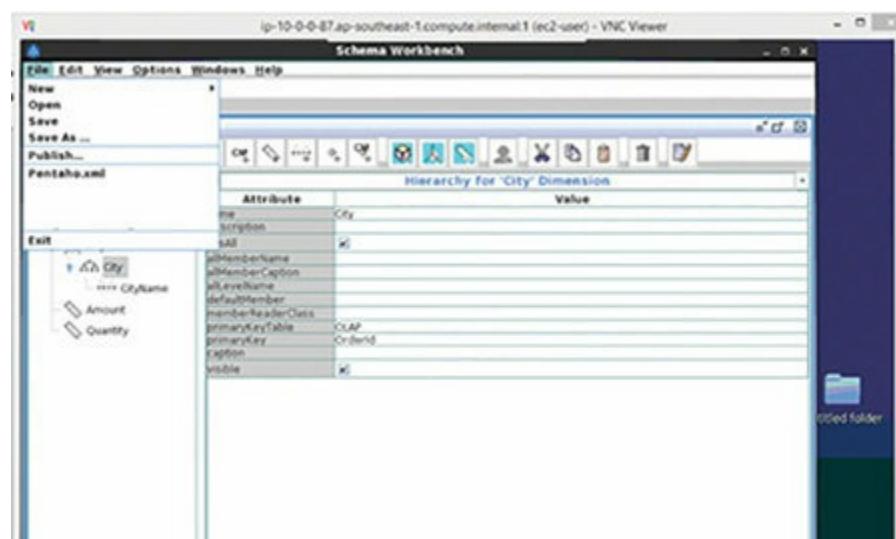


Now we have basic component call Sales cube as OLAP cube and publish it.  
It has two basic components – Dimensions and Measures.

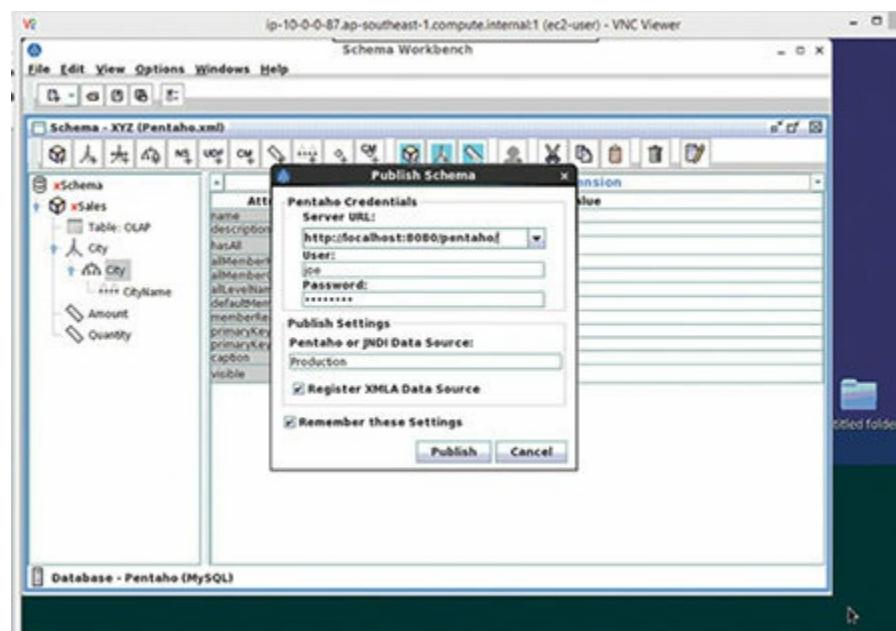
Save the Cube in the folder of your choice.



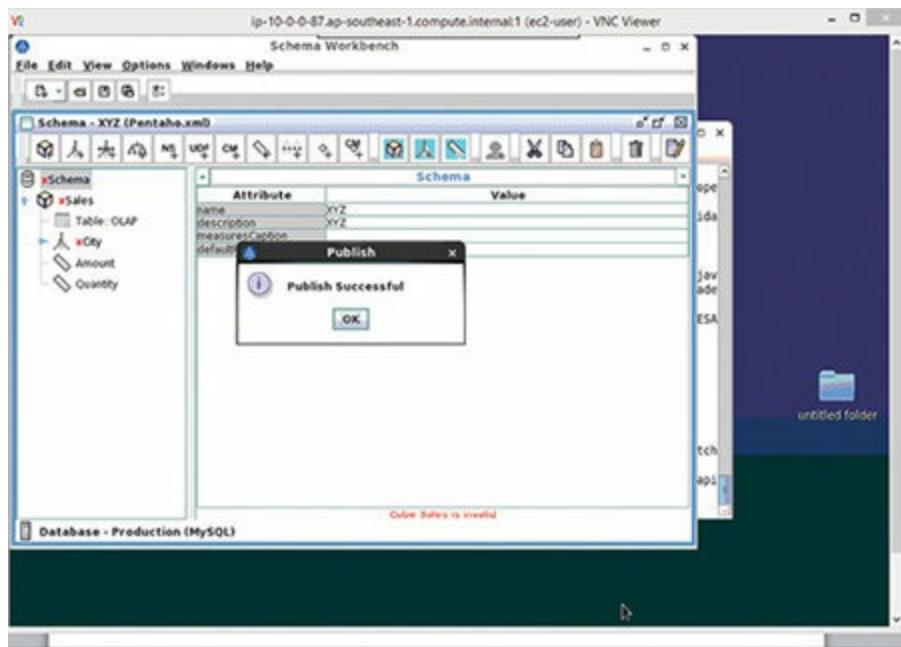
Click on File -> Publish



Add Server URL, User and password.



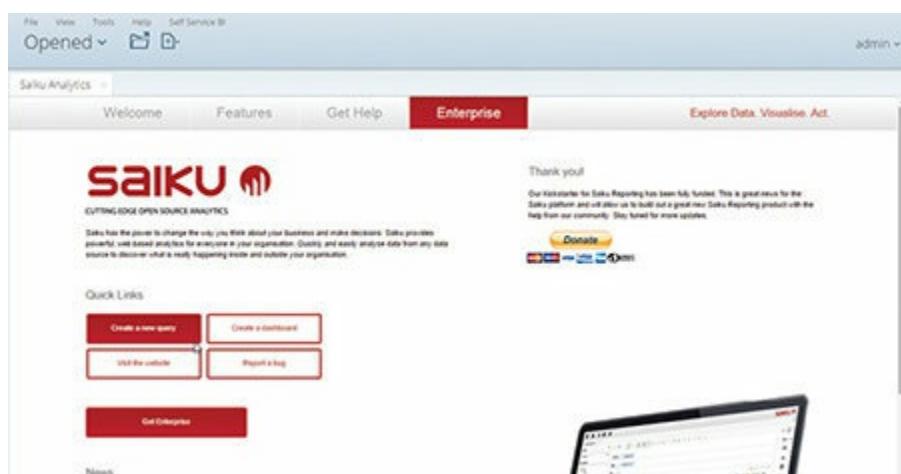
The cube is Published Successfully.



Now go to browser login. Open Saiku Analytics.



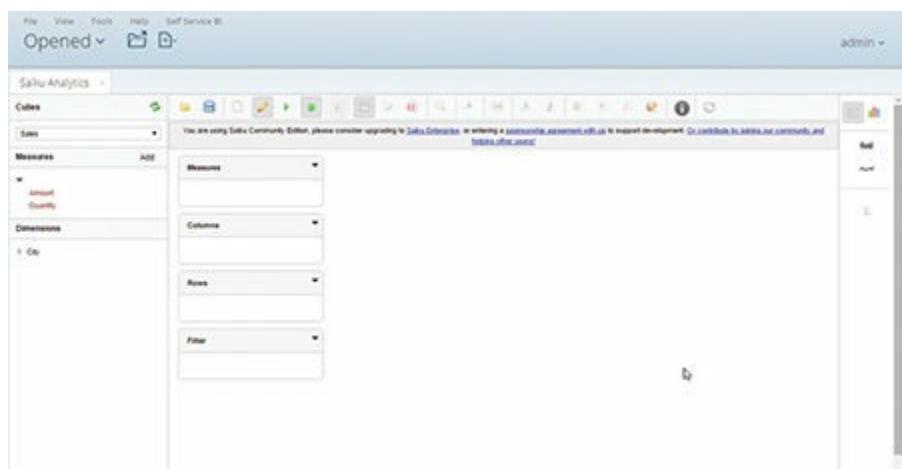
Select Create a New Query.



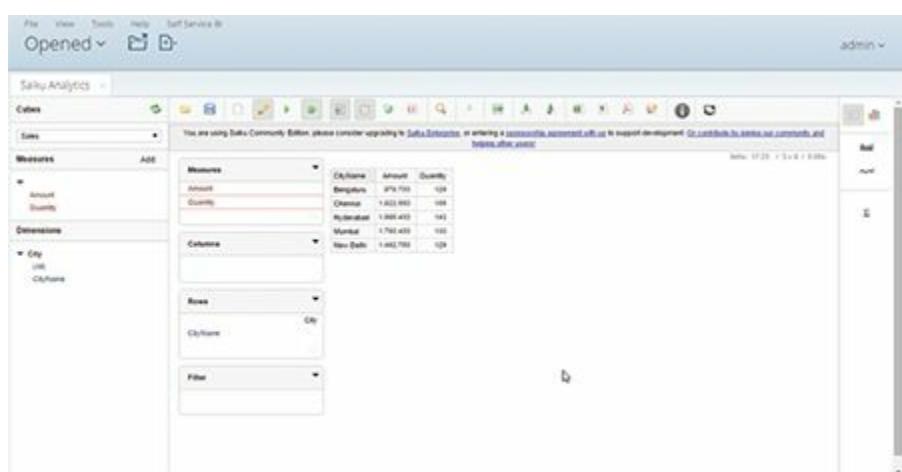
In the workspace click on select a cube dropdown. We can see our cube with Schema XYZ and Cubename Sales. Select sales cube.



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.

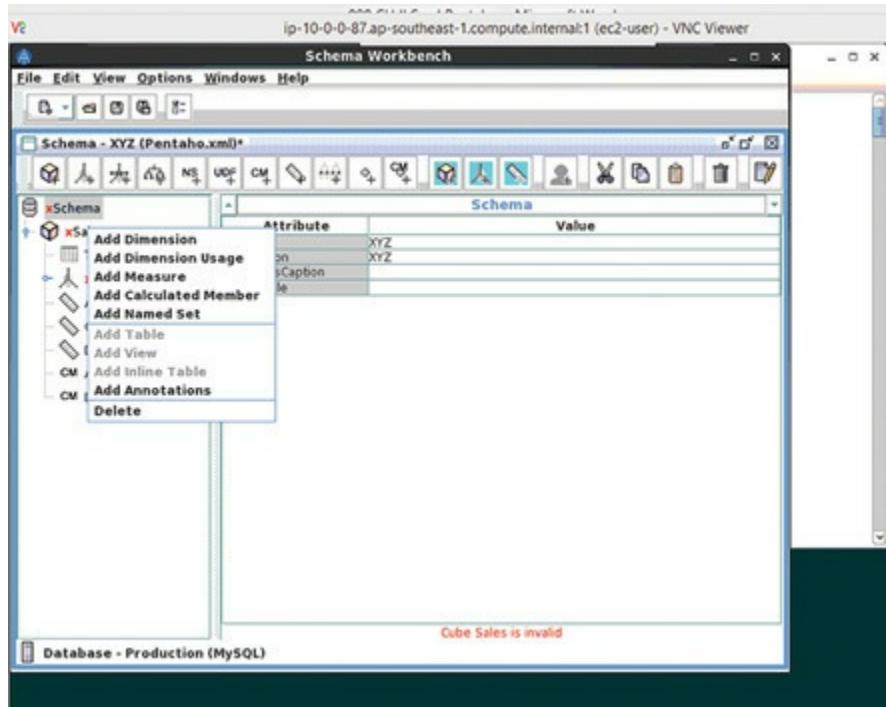


Add Amount and Quantity in the measures and Cityname in the Rows. The report is displayed as shown below

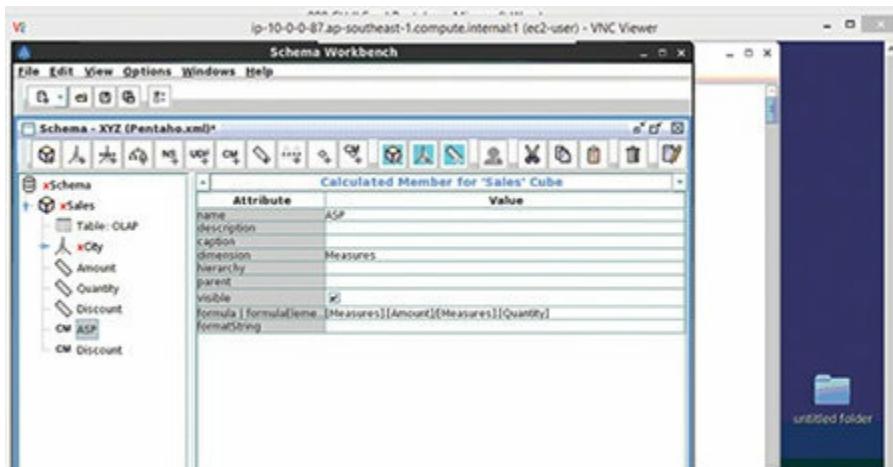


CityName	Amount	Quantity
Bangalore	1074700	125
Chennai	1422400	145
Hyderabad	13605400	143
Mumbai	1792400	133
New Delhi	1462700	129

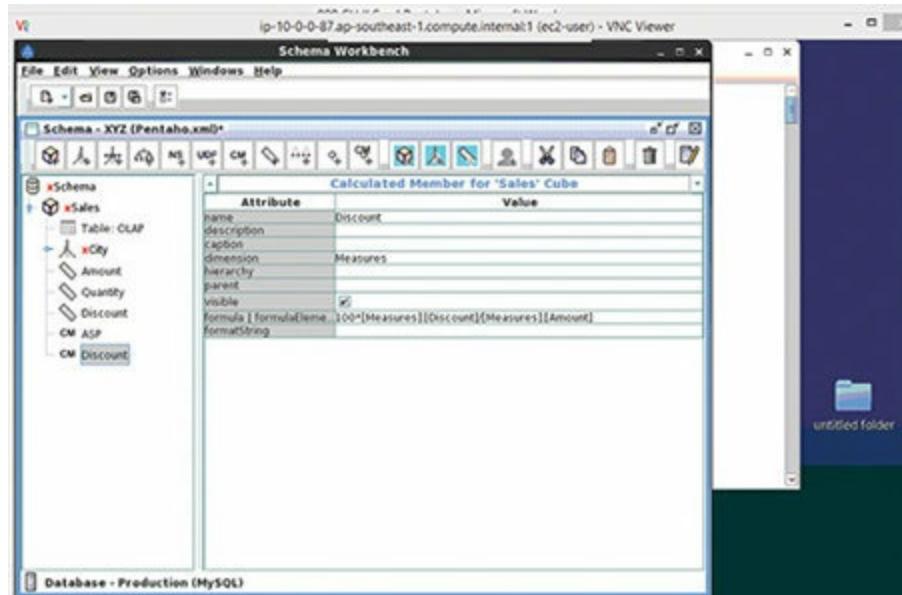
Sine we have only one dimensions and two measures, there is hardly anything we can do on the cube. Let's add more Dimensions and Measures. Right Click on Sales cube and select Add Calculated Member.



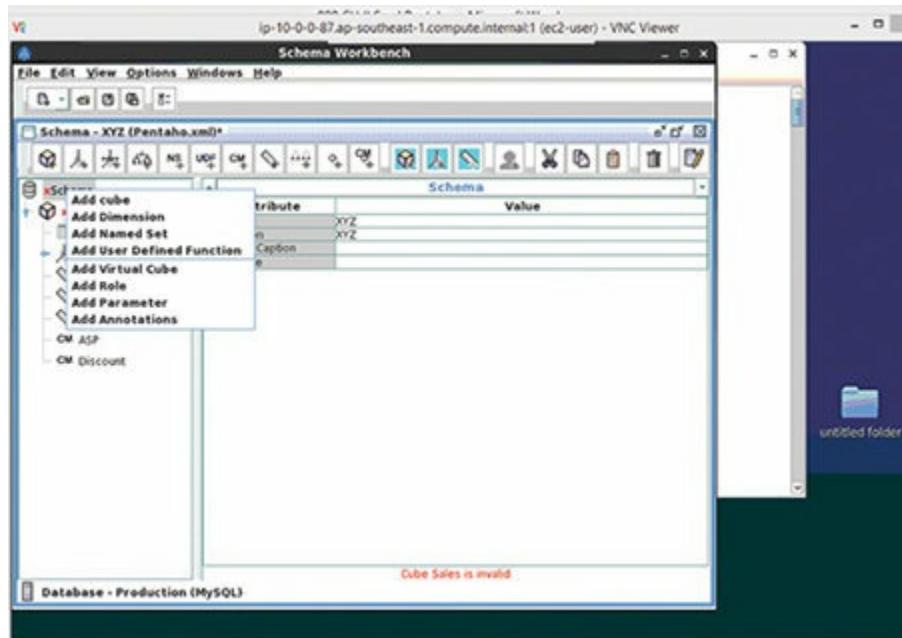
The calculated members are created out of existing measures. Here we are creating average selling price (ASP) which is calculated as Amount by Quantity. In the name section add ASP, Select Measures in dimension; add formula in the Formula space as [Measures].[Amount]/[Measures].[Quantity].



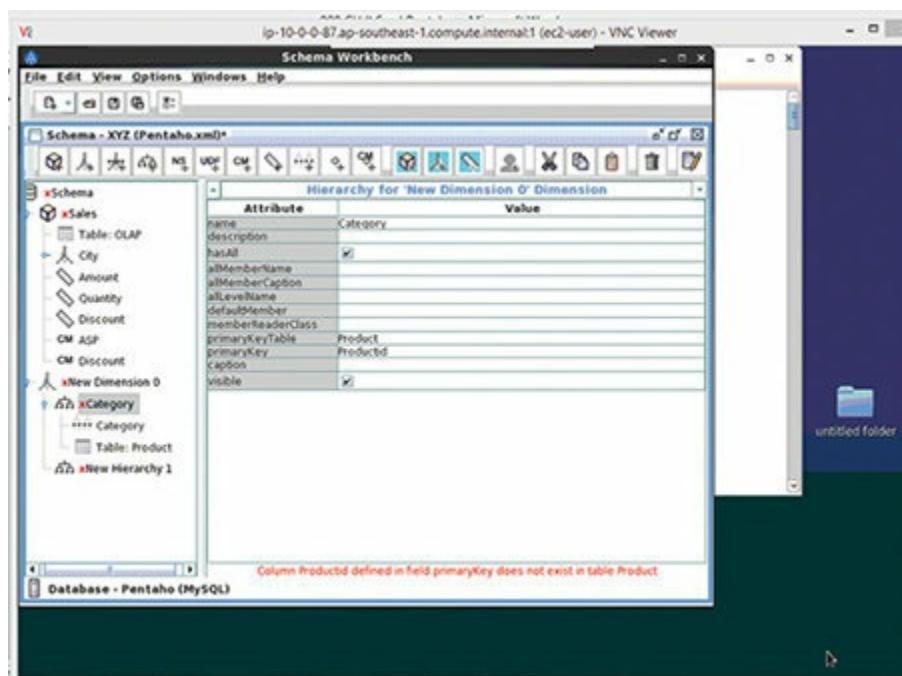
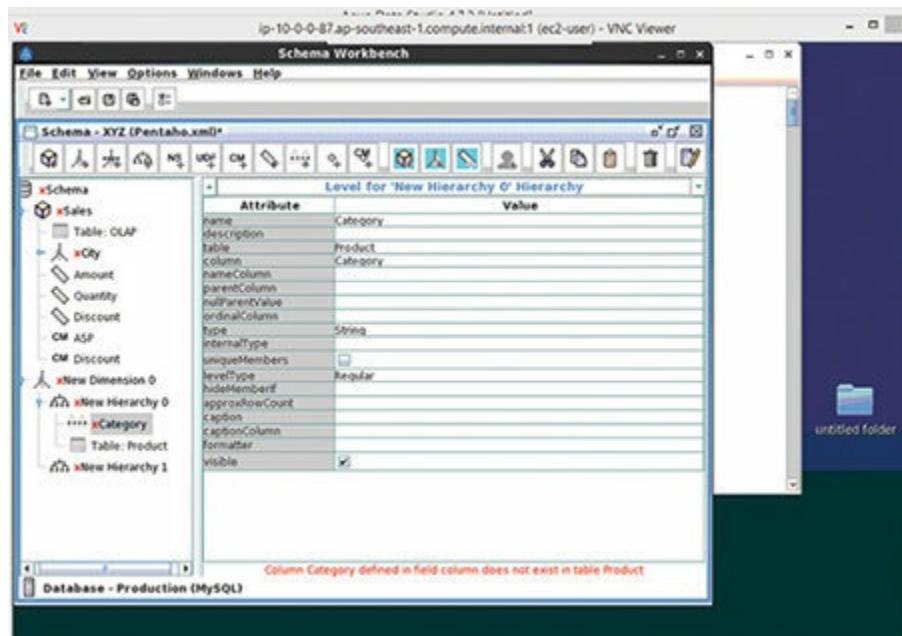
Similarly add Discount percent by adding Discount as measures and calculate discount percent as  $100 * [\text{Measures}].[\text{Discount}] / [\text{Measures}].[\text{Amount}]$ .



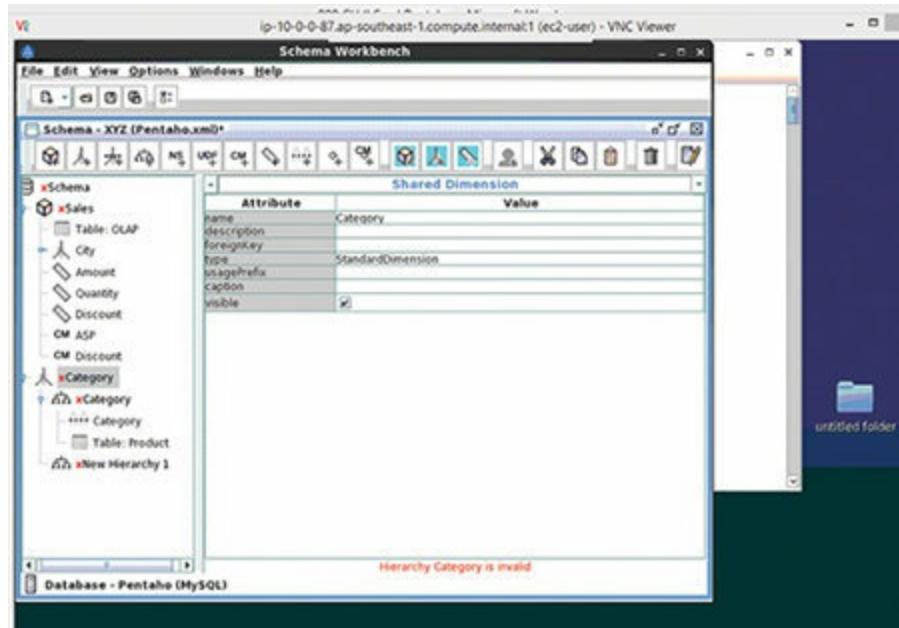
Let us add global dimensions. Right click on schema, select add Dimension.



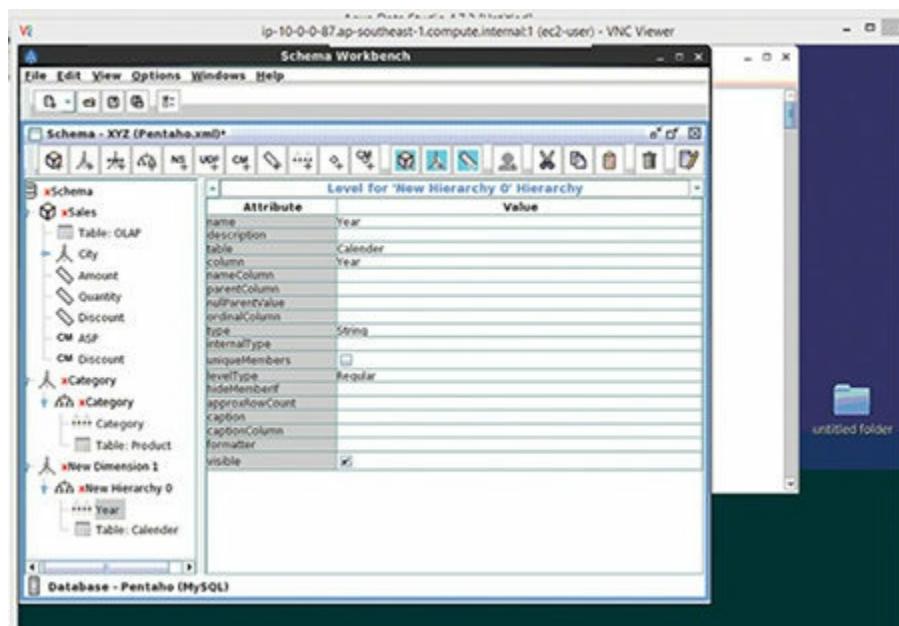
Select table as Product, add category dimension. The dimension adding is same as describe earlier, hence I am not showing each steps henceforth.



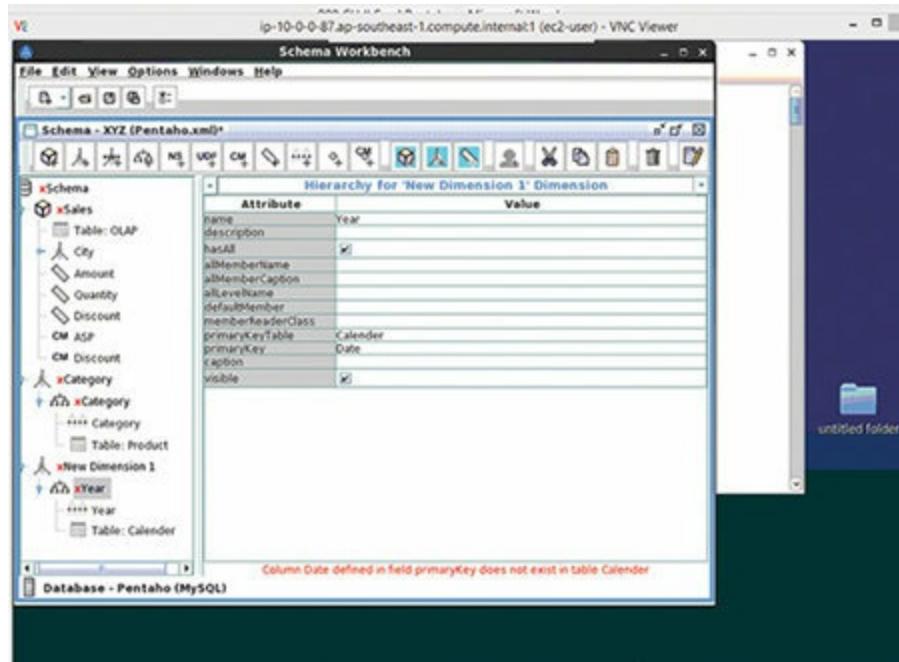
I have named all level, hierarch and dimension as category. Here we will not delete the table because the Dimension will pick up data from this table not from the Sales cube.



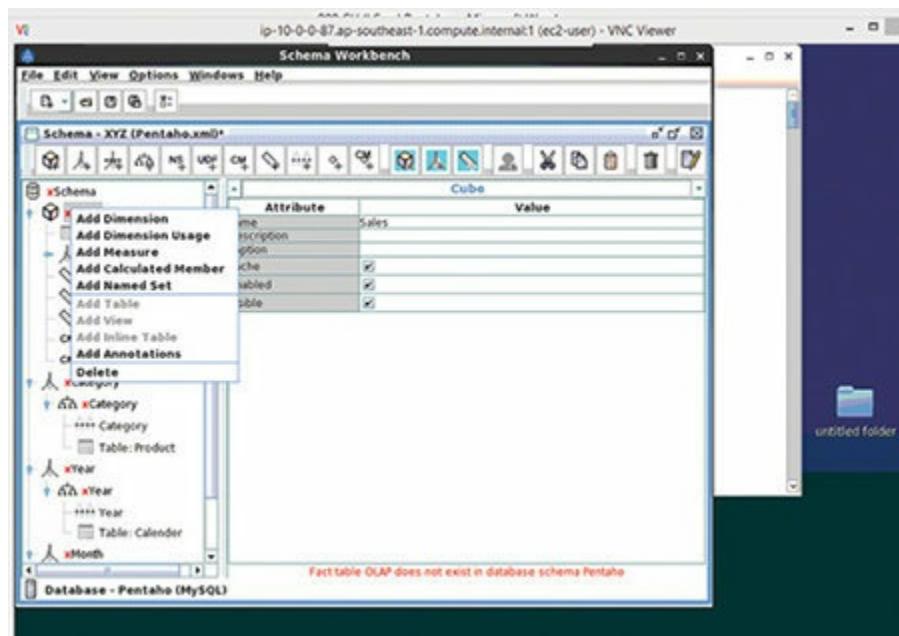
Similarly add Year from the Calender Table.



Add primary key table and primary key of the table

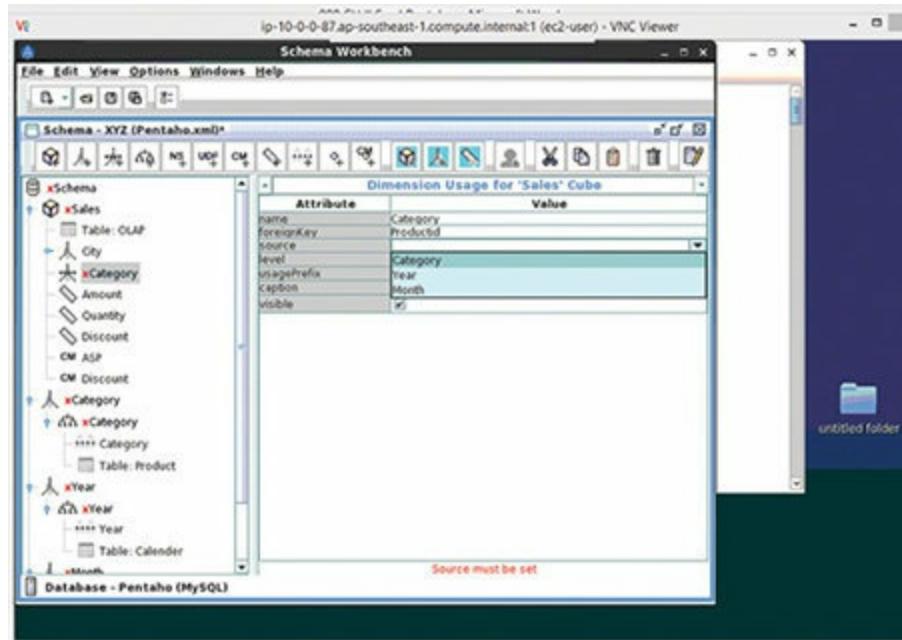


Similarly we have added month as dimensions.

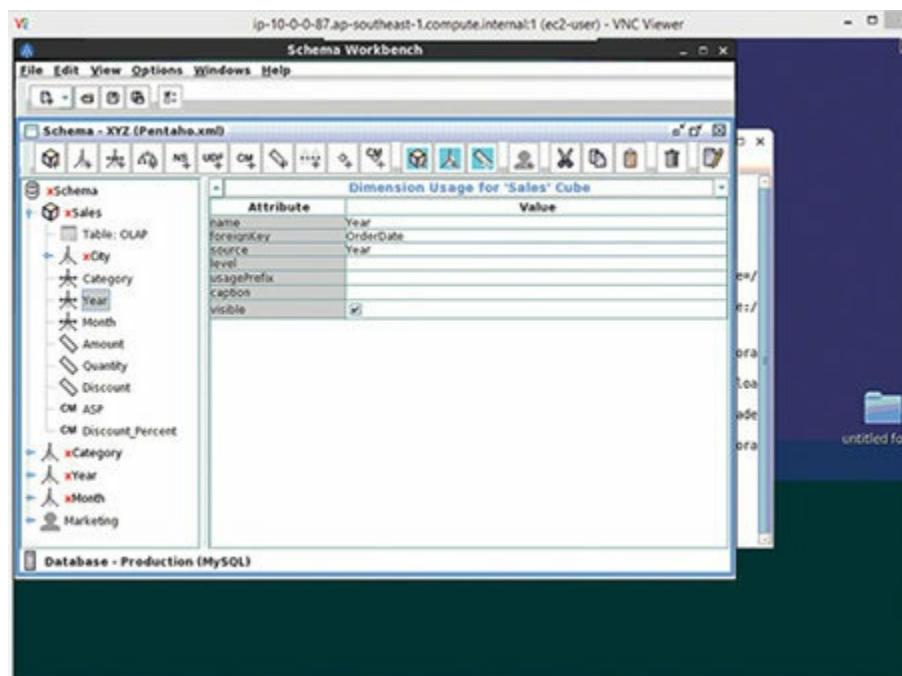


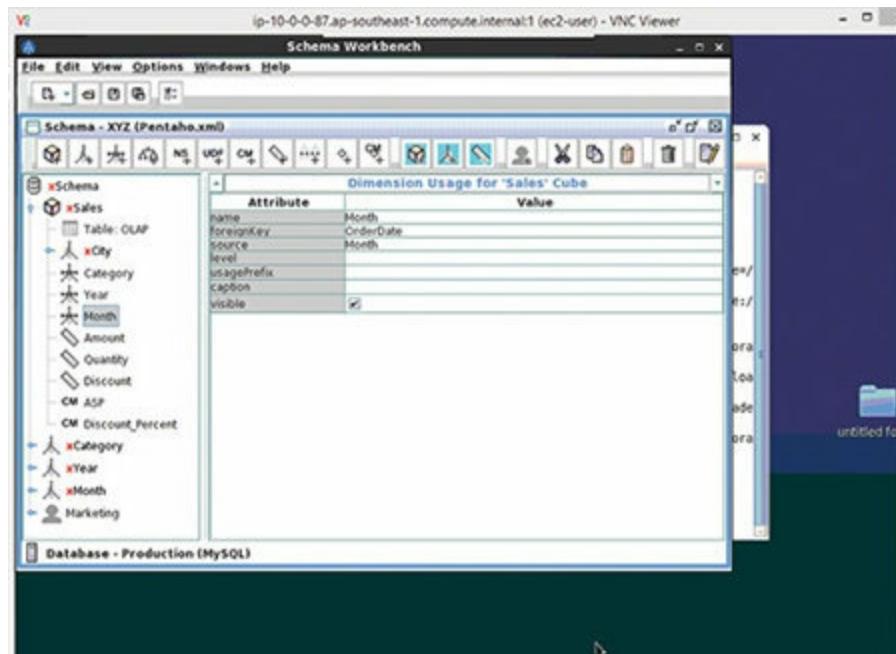
The global dimensions created can be used by any cube under Schema XYZ. Here we have only one cube Sales. To use the global dimensions in the cube, so to cube, right click and select add Dimension Usage.

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.



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





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

The screenshot shows the Salsu Analytics interface. On the left, there are sections for 'Cubes' (with 'Sales' selected), 'Measures' (listing 'Amount', 'Quantity', 'Discount', 'ASP', and 'Discount\_Percent'), and 'Dimensions' (listing 'Category', 'City', 'Mark', 'Year', 'Month', and 'Year'). The main area displays a data grid with columns 'Category', 'Amount', and 'ASP'. The data shows categories like Apparel, Books, Laptop, and Mobiles with their respective amounts and ASP values. A message at the top of the grid area says, 'You are using Salsu Community Edition, please consider upgrading to Salsu Enterprise, or entering a password@salsoft.net:8080 to support development. Go contribute to https://www.salsu.com and https://github.com/salsu'.

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

This screenshot shows the Salsu Analytics interface with a similar layout to the previous one. The left sidebar shows 'Cubes' (Sales), 'Measures' (Amount, Quantity, Discount, ASP, Discount\_Percent), and 'Dimensions' (Category, City, Mark, Year, Month, Year). The main area displays a data grid with columns 'Category', 'Amount', 'Discount\_Percent', '2015', and '2016'. The data shows categories like Apparel, Books, Laptop, and Mobiles with their respective amounts and discount percentages for both years. A message at the top of the grid area says, 'You are using Salsu Community Edition, please consider upgrading to Salsu Enterprise, or entering a password@salsoft.net:8080 to support development. Go contribute to https://www.salsu.com and https://github.com/salsu'.

The screenshot shows two separate Power BI reports. Both reports have a similar layout with a left sidebar for 'Codes' (Measures and Dimensions) and a right main area for data visualization.

**Report 1 (Top):**

- Measures:** Amount, Discount, Discount\_Percent.
- Dimensions:** Category (All, Company), City (All, CityName), Year (All, Year).
- Data:** A table showing sales data for 2015 and 2016 across various cities. The table includes columns for Category, CityName, Amount, Discount\_Percent, and a calculated column for 2016.

Category	CityName	2015		2016	
		Amount	Discount_Percent	Amount	Discount_Percent
Apparels	Bengaluru	31,400	9.37%	8,200	8.81%
	Chennai	42,800	4.21%	2,800	-1%
	Hyderabad	28,800	8.88%	10,200	5.00%
	Mumbai	31,200	9.47%	7,400	5.88%
	New Delhi	28,800	9.45%	8,800	1.76%
Books	Bengaluru	8,800	9.87%	2,800	4.32%
	Chennai	10,100	4.10%	1,800	0
	Hyderabad	8,800	2.88%	4,800	0
	Mumbai	8,800	2.82%	1,240	0
	New Delhi	8,800	9.12%	4,800	0
Laptops	Bengaluru	928,100	1.12%	182,000	0
	Chennai	884,200	2.48%	170,000	4.00%
	Hyderabad	924,200	2.71%	180,000	4.00%
	Mumbai	924,200	1.76%	208,000	2.04%
	New Delhi	924,200	8.86%	17,000	10
Mobiles	Bengaluru	618,800	4.86%	16,000	5.07%
	Chennai	618,800	2.88%	181,000	4.00%
	Hyderabad	617,800	3.88%	82,000	10.80%
	Mumbai	618,800	9.74%	34,000	0
	New Delhi	614,400	1.72%	151,000	4.00%

**Report 2 (Bottom):**

- Measures:** Amount.
- Dimensions:** Category (All, Company), City (All, CityName), Year (All, Year).
- Data:** A table showing sales data for 2016 only across various cities. The table includes columns for Category, CityName, and Amount.

Category	Amount
Apparels	181,800
Books	10,800
Laptops	9,781,000
Mobiles	1,831,200
Tots	210,000

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

The screenshot shows the 'Browse Files' interface in Power BI. It displays a file tree on the left and a list of files on the right. A context menu is open over the 'Test.xlsx' file in the list.

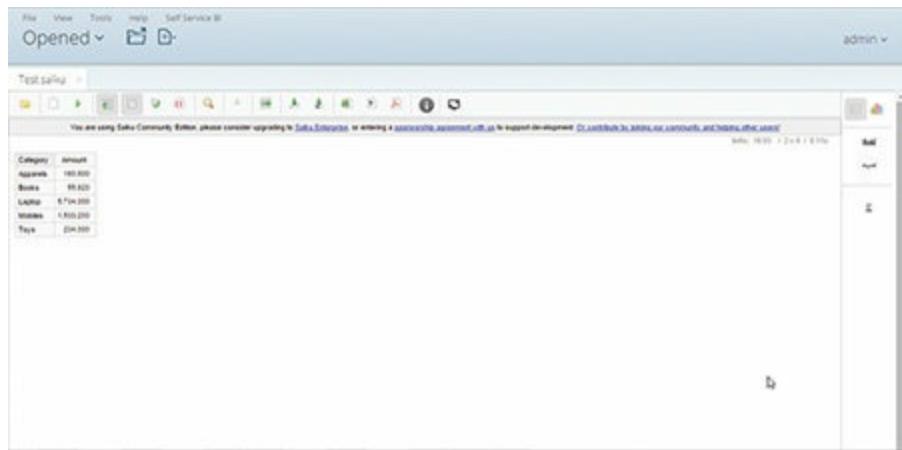
**File Tree:**

- Folders:** joe, Public.
- Files:** Disc\_Percent.xlsx, MTD\_Order.xlsx, Test.xlsx.

**Context Menu (Open over Test.xlsx):**

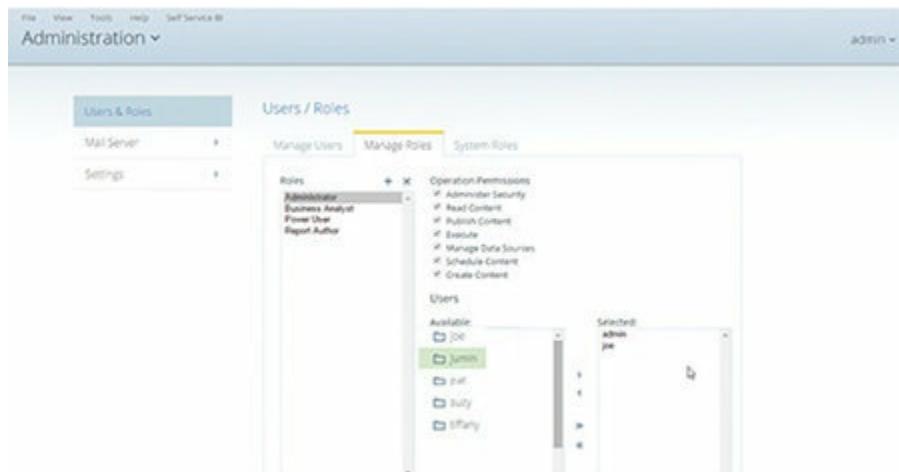
- New Folder...
- Move to Trash
- Rename...
- Upload...
- Download...
- Properties...

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

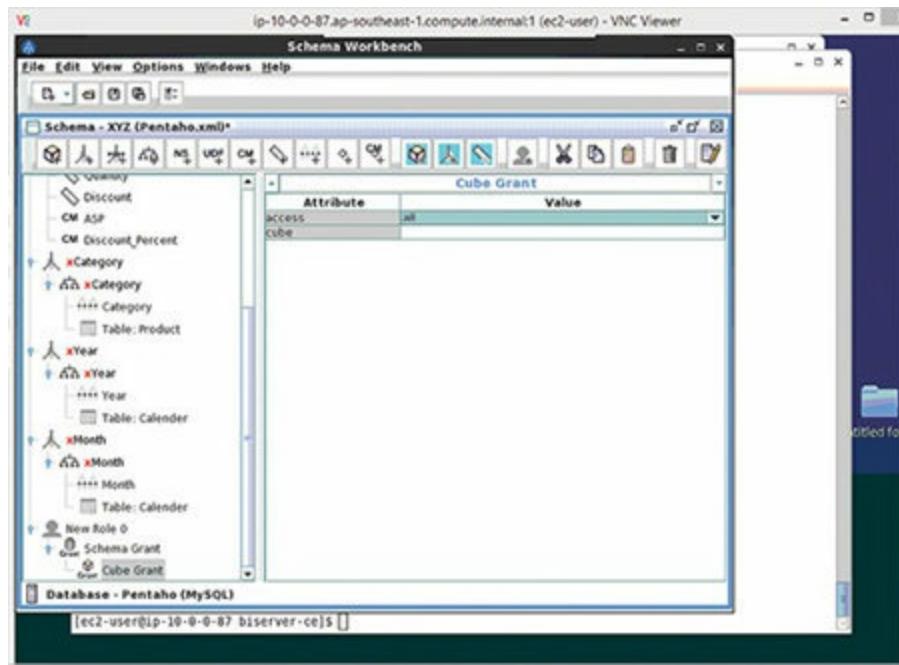


This pretty much explain how to create a simple cube and render them in the browser. For the real time user the data will be much more complex and size will be very big. You have to plan all dimensions and metrics by speaking to the users before starting the cube designing.

Another important topic I will just mention here with regard to cube is that the company will have data policy where certain users are 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.



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



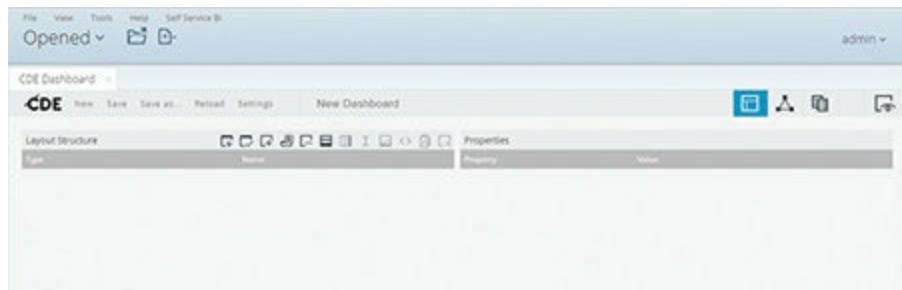
I will not explain the entire process but users can try out different combination and experiment with cube. It's not difficult.

### 2.1.3 Dashboards

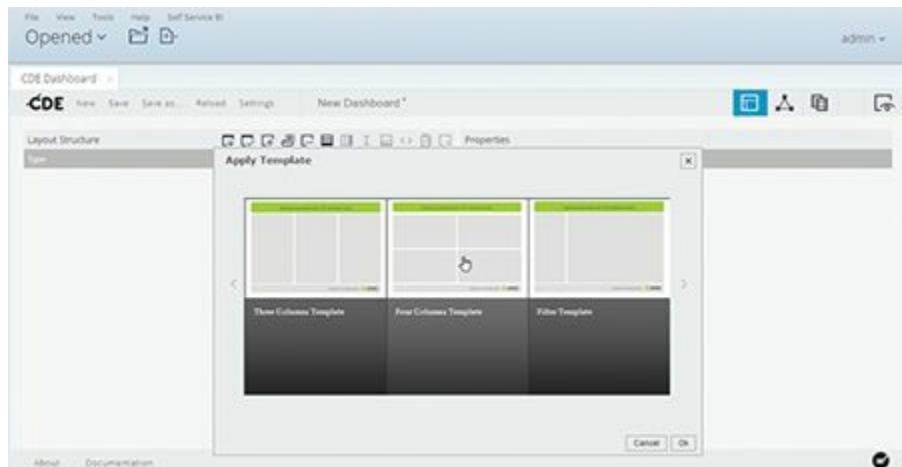
The dashboard are also a report but what makes it different from usual report is that is more summarized form with more pictorial representation meant for the top management. Select CDE Dashboard from the Create New



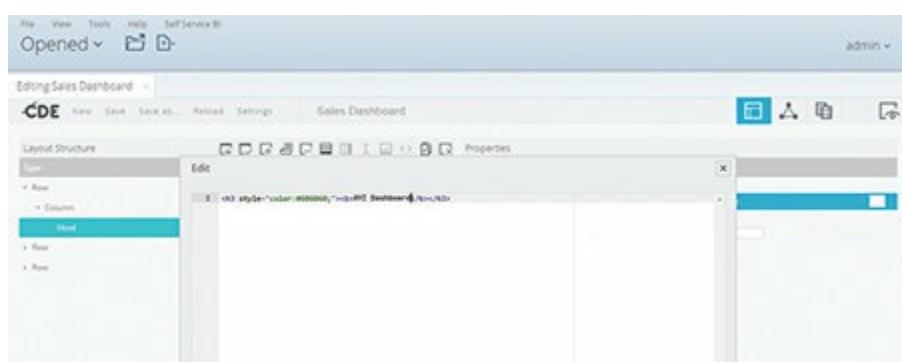
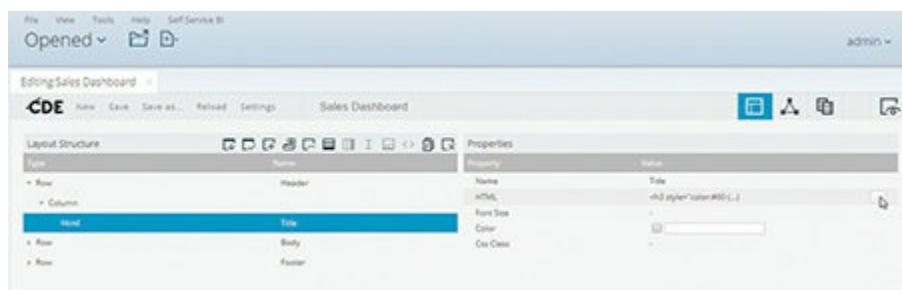
For designing a good dashboard you will need to have good knowledge of html and css. In community dashboard we have three main components – layout structure which provides the structure of the dashboard, component panel which is the elements of the dashboard lie chart, tables and the datasources which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

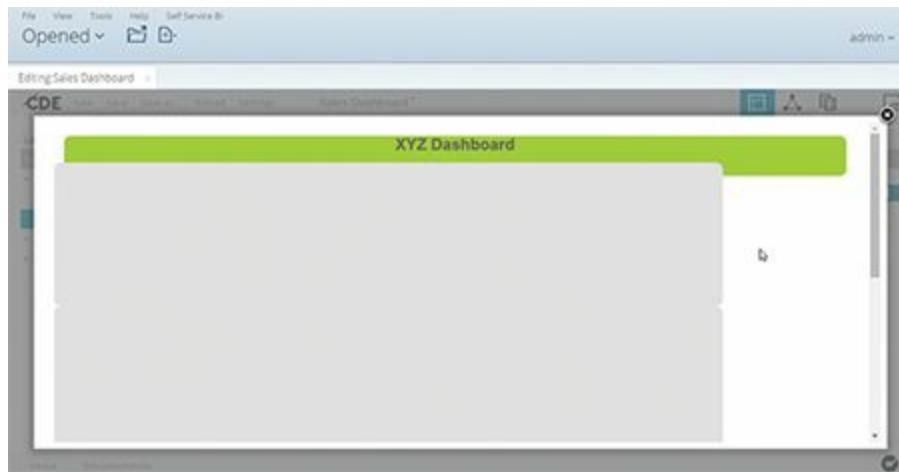


We will design very basic dashboards. Select a layout. I have selected 2x2 template.



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

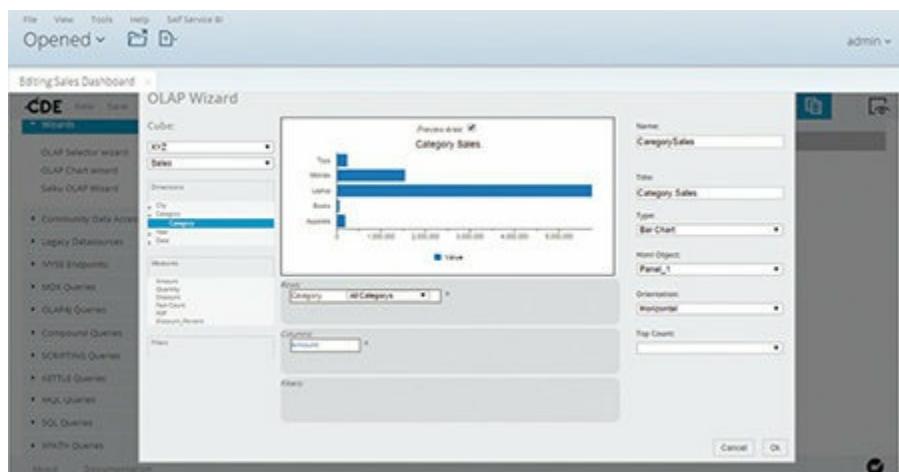


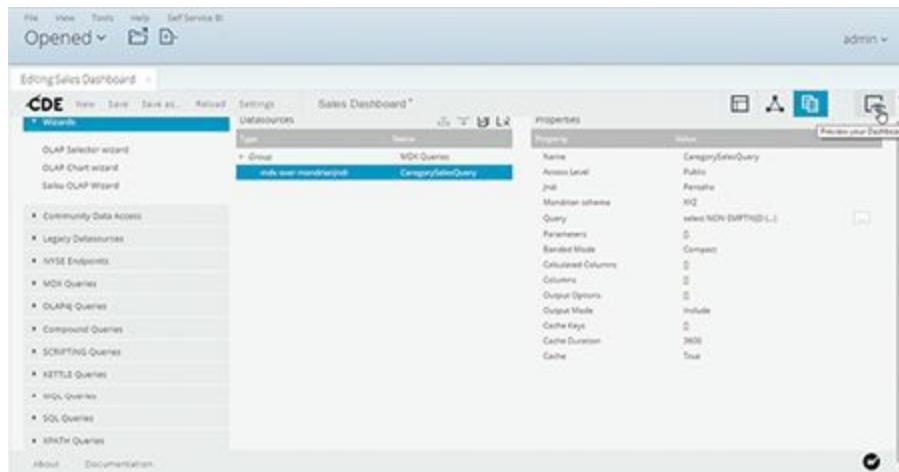


Go to component panel and select Wizards -> Saiku OLAP wizard

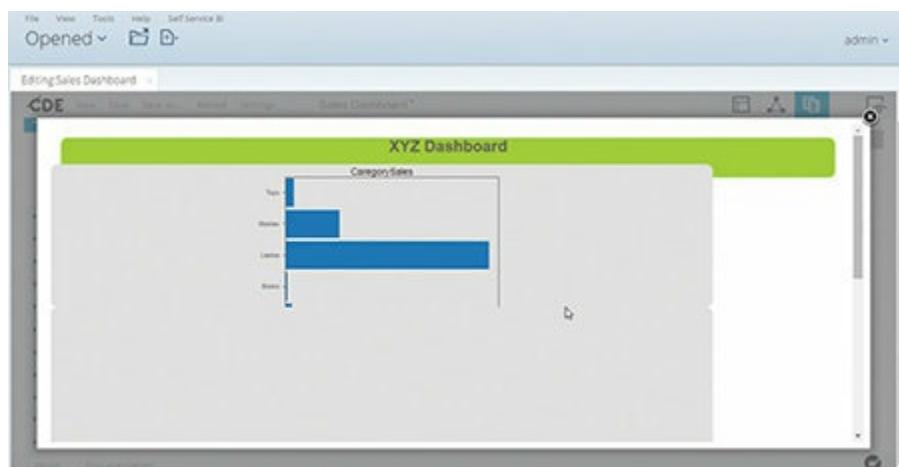


Create chart using OLAP cube data we just created. We have four panels add name of panel, title, chart type and Panel. Here we have selected first panel panel\_1

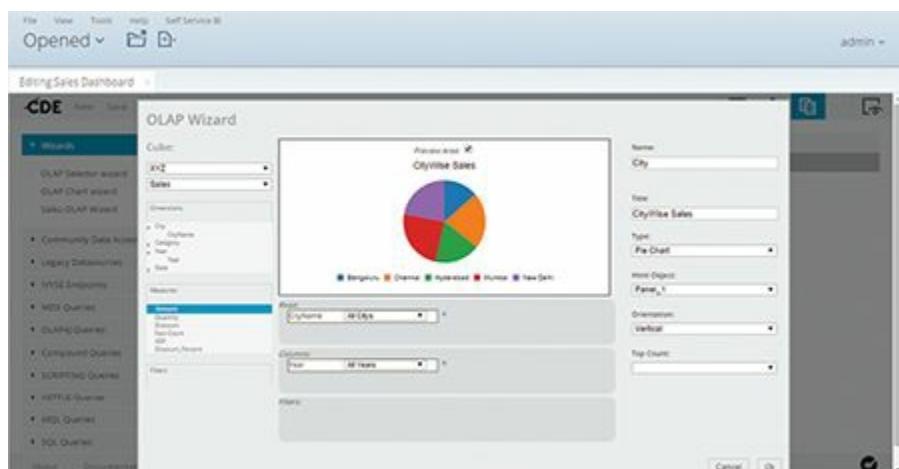




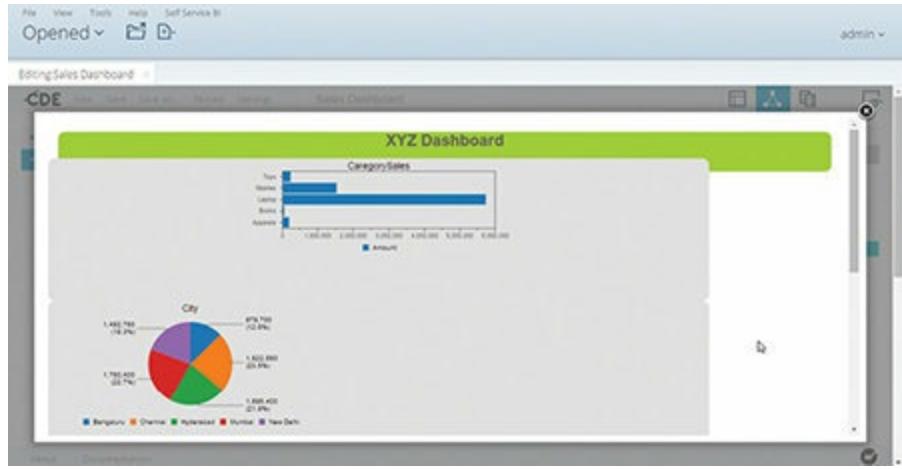
On preview we can see the chart in the first panel. The chart is distorted. User has to do dimension setting to get it to the right size.



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.



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.

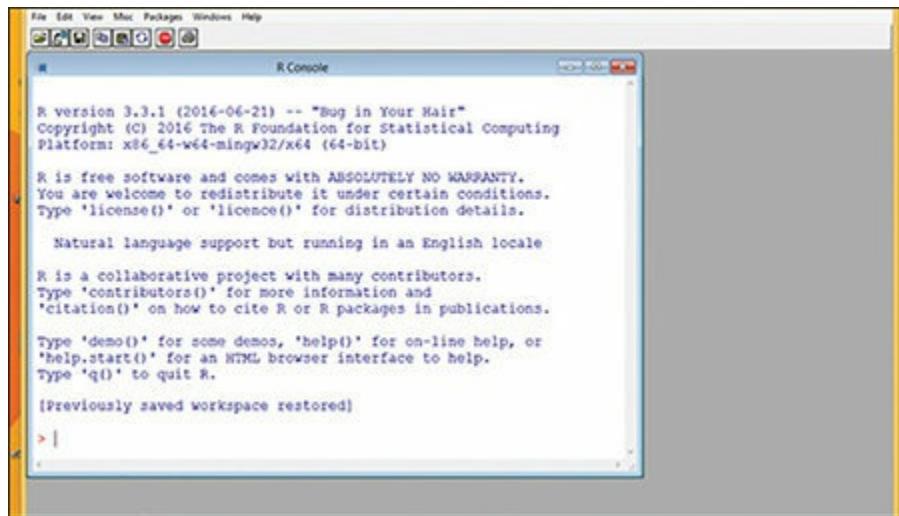
## Learning from Chapter

- 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 as 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 <- 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

command 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
1 1 2
2 2 3
3 3 4
4 4 5
5 5 6
> 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
```

a	b
---	---

Min.	:1	Min.	:2
------	----	------	----

1st Qu.	:2	1st Qu.	:3
---------	----	---------	----

Median	:3	Median	:4
--------	----	--------	----

Mean	:3	Mean	:4
------	----	------	----

3rd Qu.	:4	3rd Qu.	:5
---------	----	---------	----

Max.	:5	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
```

```
> A
```

	X	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.

	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

```
> 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

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

```
> F
```

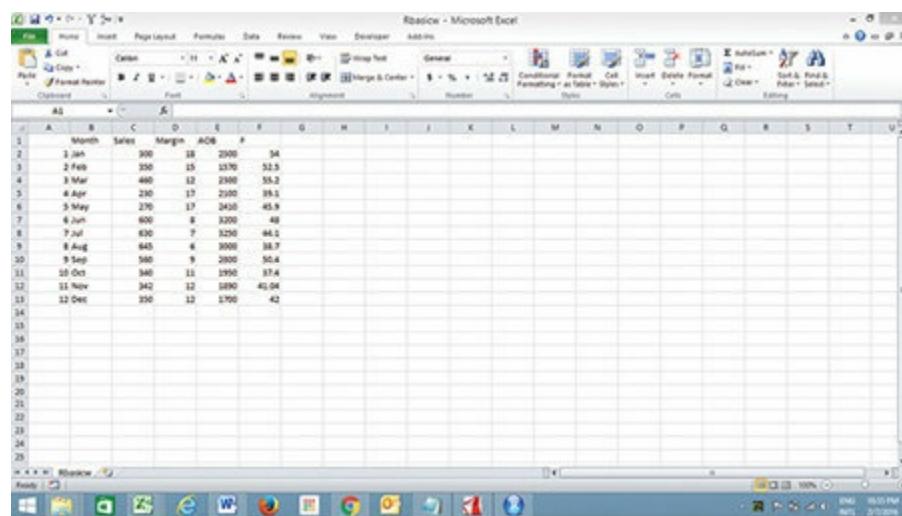
```
[1] 54.00 52.50 55.20 39.10 45.90 48.00 44.10 38.70 50.40 37.40 41.04  
[12] 42.00
```

```
> FL <- cbind(SLS, F)
```

```
> FL
```

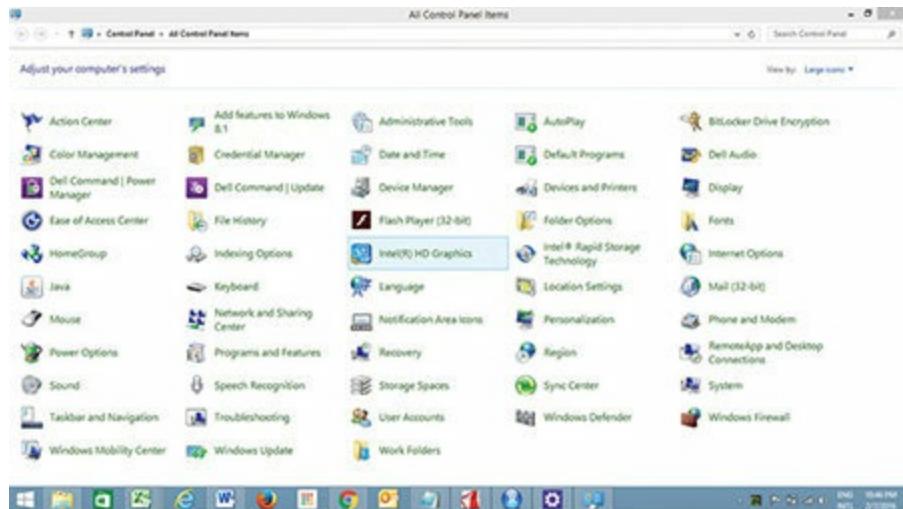
	Month	Sales	Margin	AOB	F
1	Jan	300	18	2500	54.00
2	Feb	350	15	1570	52.50
3	Mar	460	12	2300	55.20
4	Apr	230	17	2100	39.10
5	May	270	17	2410	45.90
6	Jun	600	8	3200	48.00
7	Jul	630	7	3250	44.10
8	Aug	645	6	3000	38.70
9	Sep	560	9	2800	50.40
10	Oct	340	11	1950	37.40
11	Nov	342	12	1890	41.04
12	Dec	350	12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

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

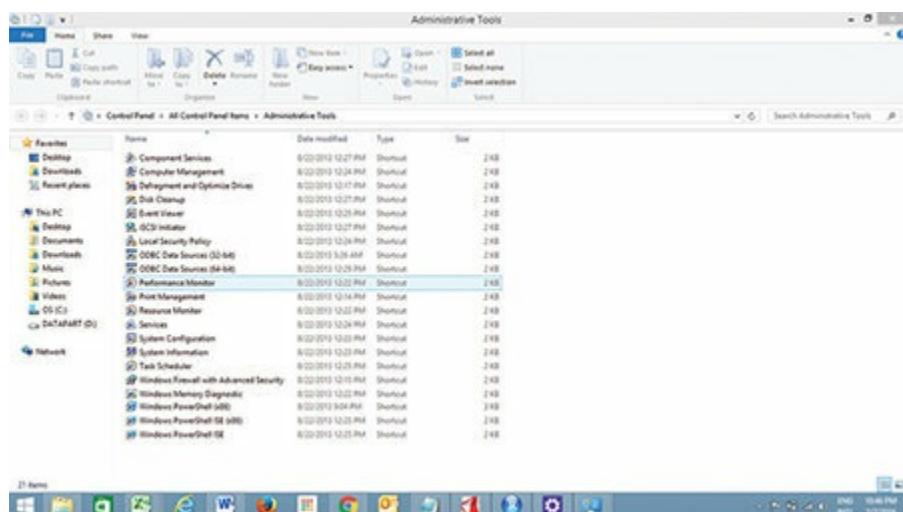


## 2.2.4 R Database Connection

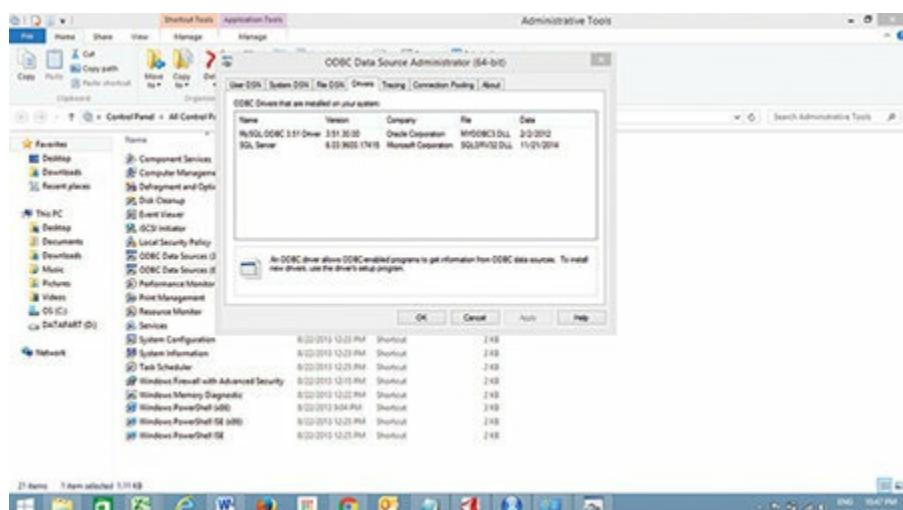
R connects to database through ODBC interface in windows. Here I am going to show how to connect MySQL database. Go to **control panel**.



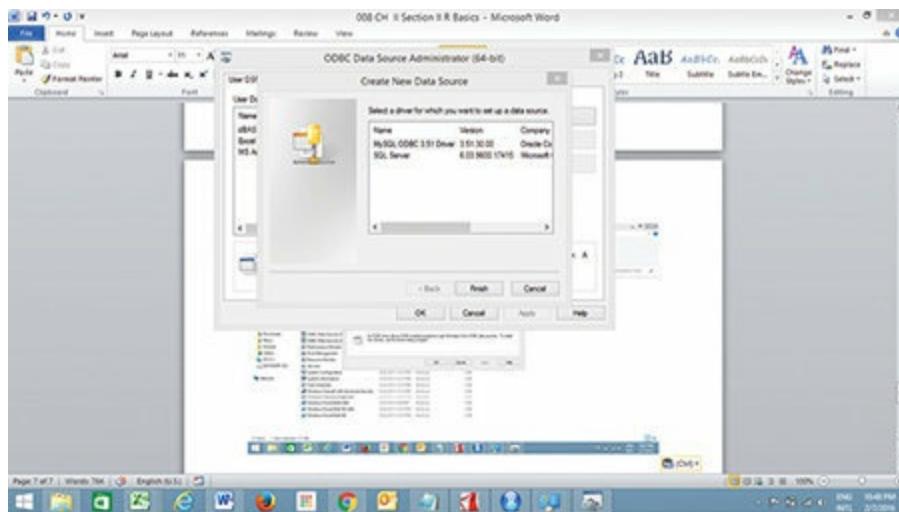
## Select ODBC data source



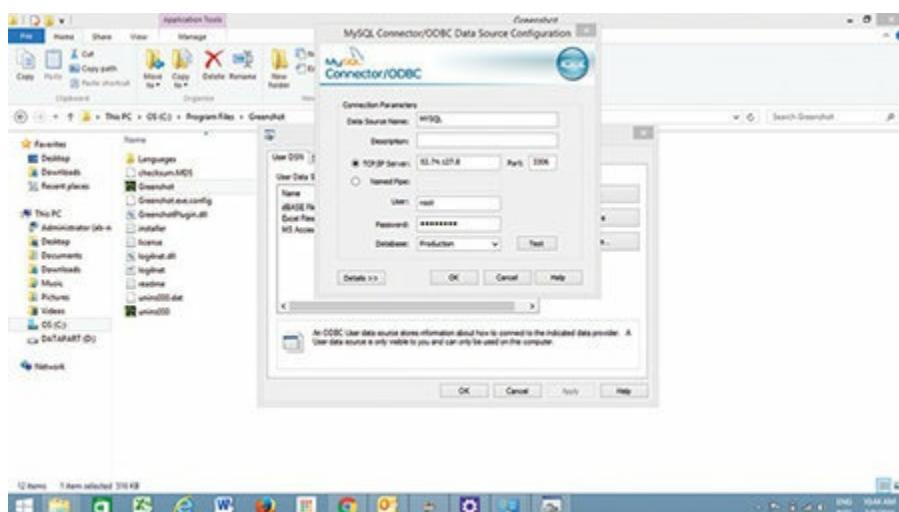
## Click on Driver->Add new Data Source



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



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



## 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;")
> 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
44	1	Saturday					2014_15_11
2	2014-11-02	2014	2014_15	11	11_Nov	08_Nov	2014_11
45	2	Sunday					2014_15_11

```
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")



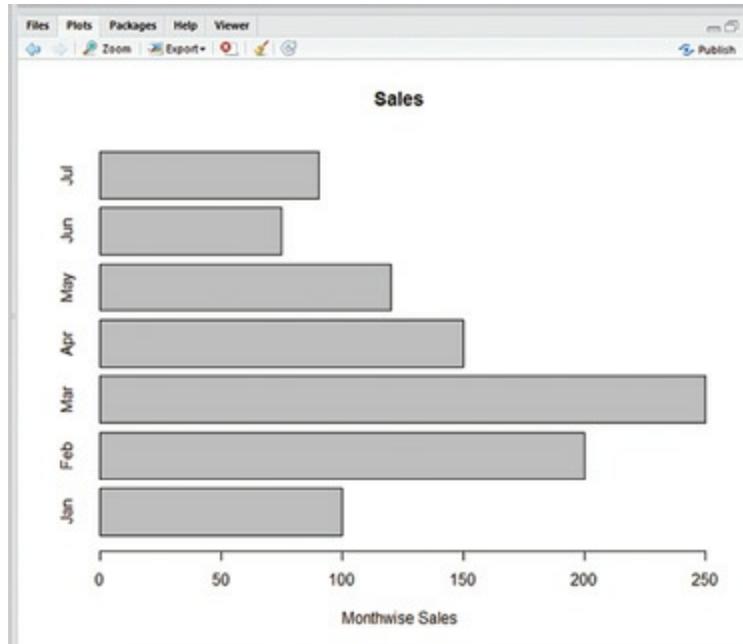
Add month in the bars

```
> 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"))
```



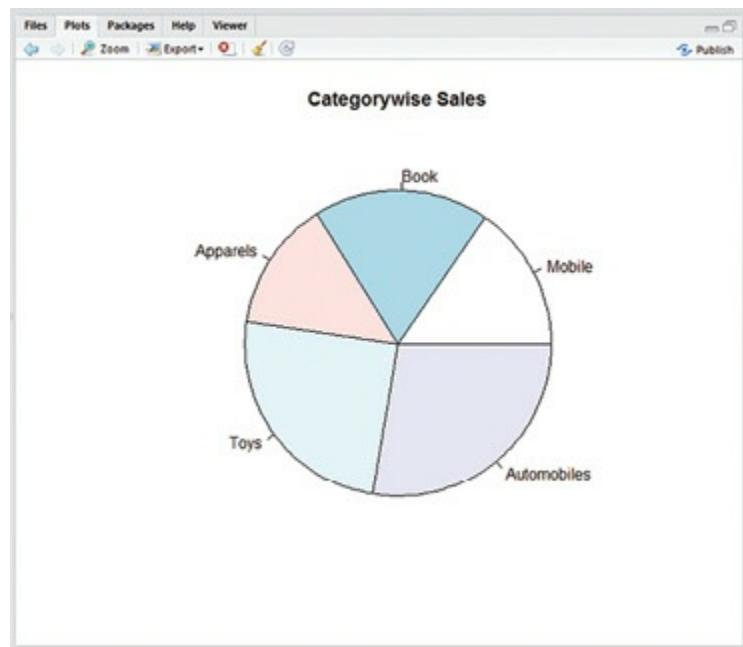
## Pie Graph

Pie chart of the Categorywise sales

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

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

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

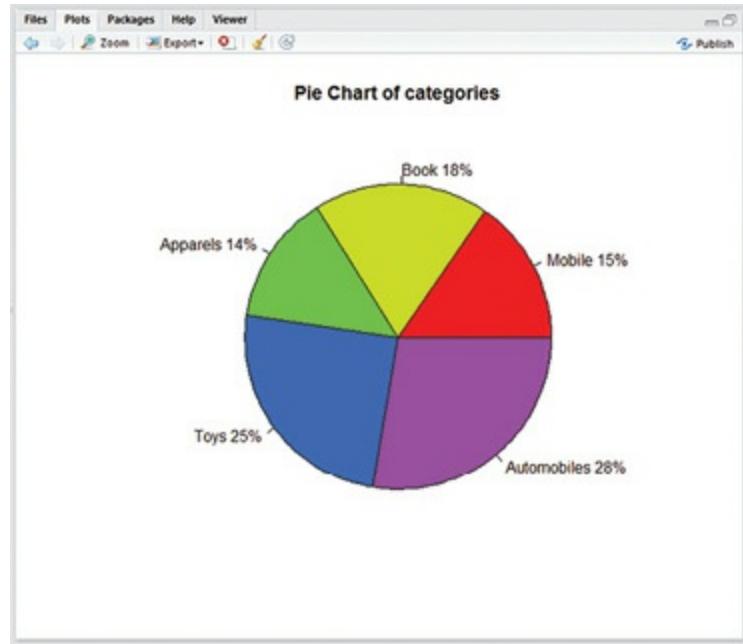


```
> Percent <- round(Pie/sum(Pie)*100)
```

```
> Cat <- paste(Cat, Percent)
```

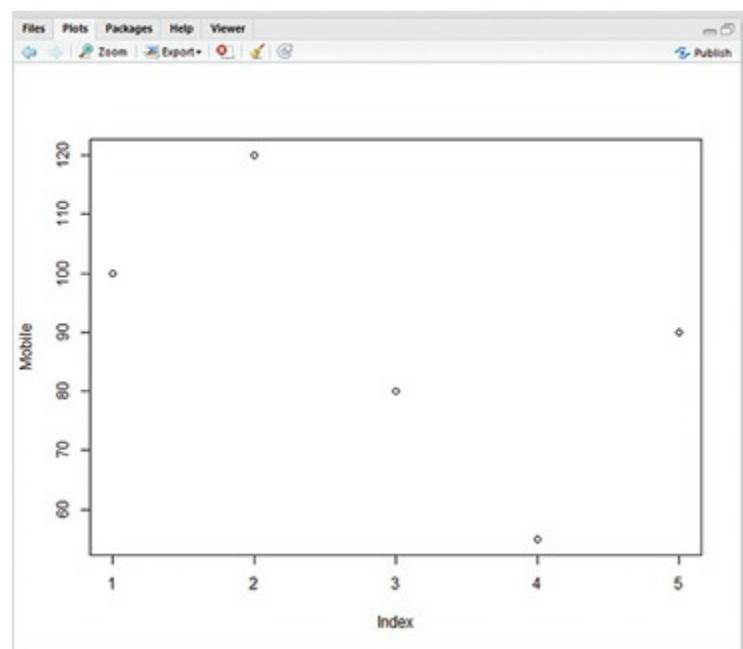
```
> Cat <- paste(Cat,"%",sep="")
```

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

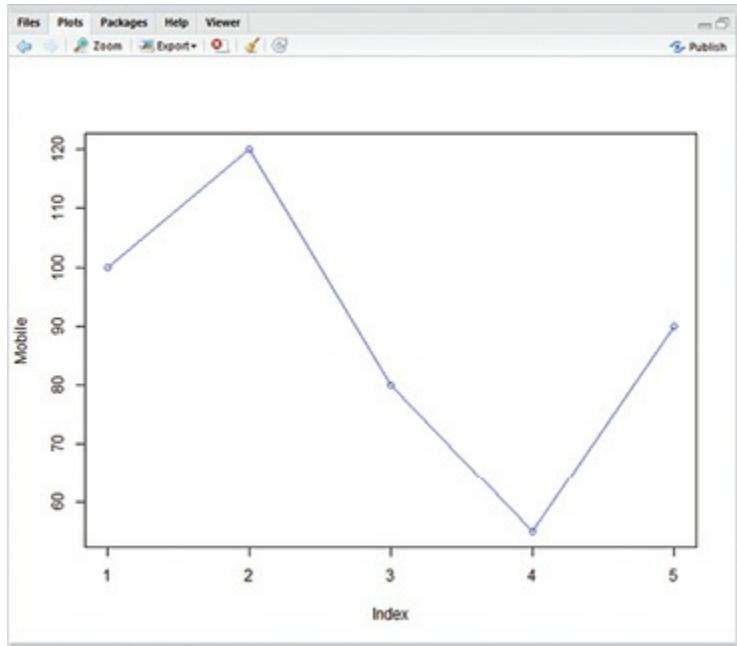


## Line chart

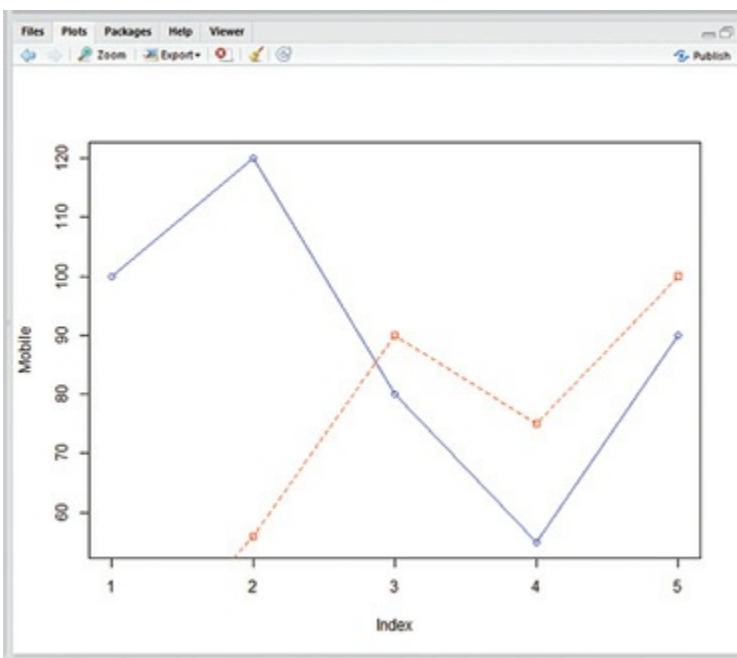
```
> Mobile <- c(100, 120, 80, 55, 90)  
> plot(Mobile)
```



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



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

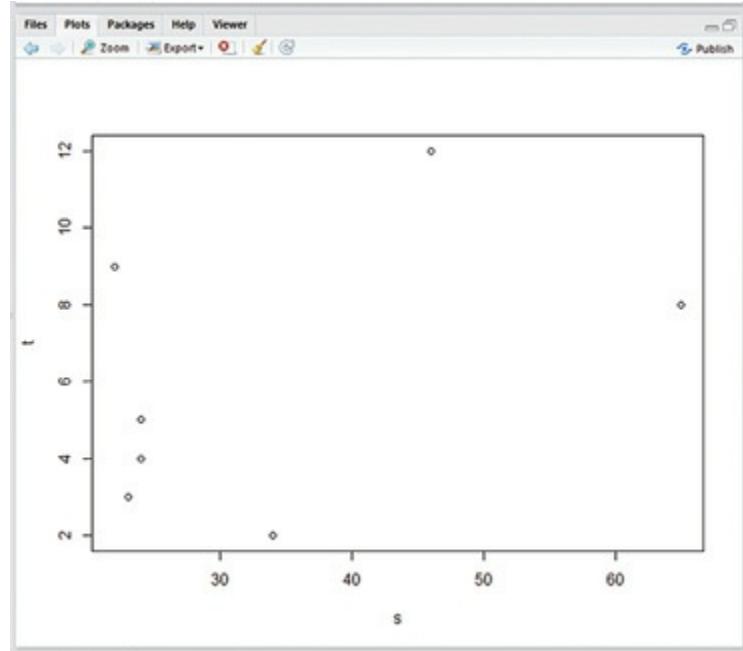


## Scatter Plot

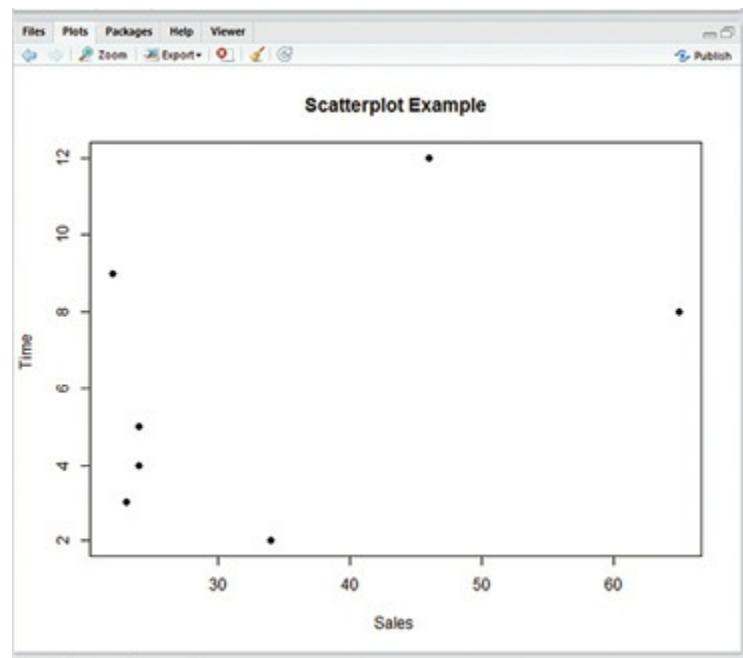
```
> SC <- data.frame(s=c(24, 23, 34, 22, 24, 65, 46), t=c(4, 3, 2, 9, 5, 8, 12))
> SC
```

	s	t
1	24	4
2	23	3
3	34	2
4	22	9

```
5 24 5  
6 65 8  
7 46 12  
> plot(SC)
```



```
> 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)
```

+ {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.

The screenshot shows the 'Dimensions & Metrics Explorer' page of the Google Developers Analytics Core Reporting API documentation. The left sidebar includes links for JavaScript Quickstart, Web App, Fundamentals, Working with the API, Dimensions & Metrics Explorer (which is selected), Segments, Segments Overview, Using Segments, Resources, Authorization, Batching, Performance Tips, Common Queries, Client Libraries, Limits and Quotas, Demos & Tools, Query Explorer, and GA Spreadsheet Addon. The main content area has tabs for MODE (set to API Names) and SEARCH. It displays a table with two columns: Dimensions and Metrics. Under Dimensions, there are checkboxes for ga:userId, ga:sessionCount, ga:dateSinceLastSession, and ga:customDefinedValue. Under Metrics, there are checkboxes for ga:users, ga:newUsers, ga:percentNewSessions, ga:1dayUsers, ga:7dayUsers, ga:30dayUsers, ga:whiteList, and ga:sessionsPerUser. Below the table, there are sections for Session and Traffic Sources.

The screenshot shows the Google Developers API Reference page for the Analytics Core Reporting API. The left sidebar lists various metrics and dimensions. The main content area is focused on the 'ga:newUsers' metric. It defines 'ga:newUsers' as 'The number of users whose session was marked as a first-time session.' Below this, there's a table for 'Attributes' with 'Data Type' as 'INTEGER' and 'Allowed in Segments' as 'Yes'. Further down, it shows 'ga:newVisits' with a status of 'Deprecated' and a note to use 'ga:newUsers' instead. At the bottom, 'ga:1dayUsers' is listed with a UI name of '1 Day Active Users'.

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



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.



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.



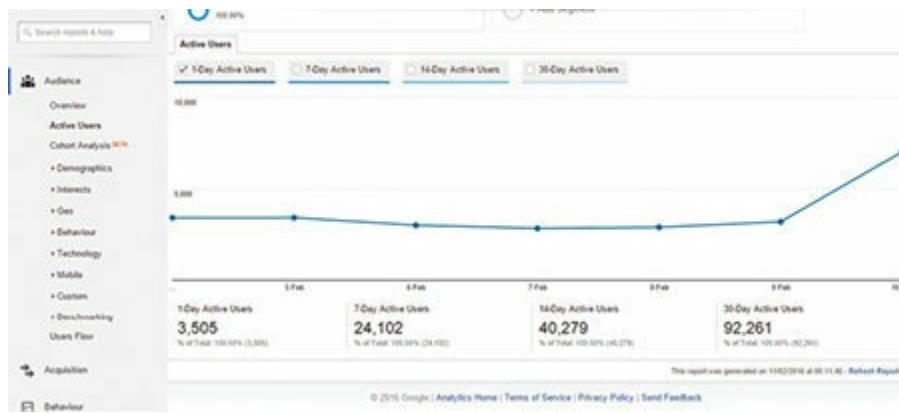
In real time Tab you can current traffic, location, sources of traffic, which pages the traffic is and so on. The real time traffic is very useful for monitoring traffic after doing changes in marketing campaign, activation of ATL/BTL campaign, for monitoring site issues and estimating the sales etc.



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.

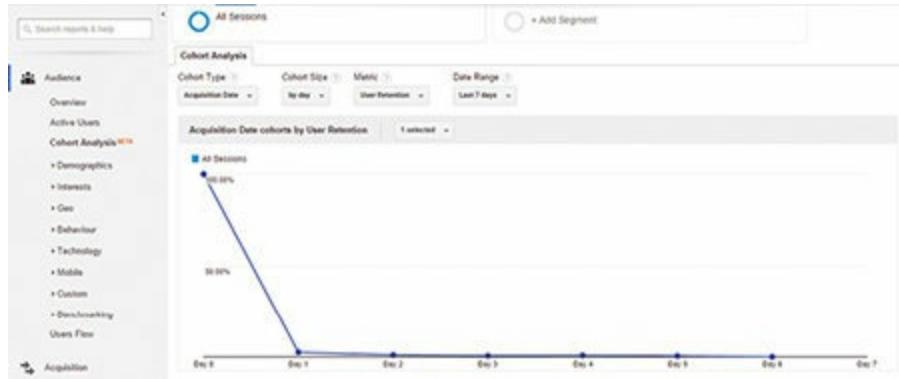


The Active Users Tab provides number of active users in 1 day, 7 days, 14 days and 30 days.



### 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.



The cohort can be viewed in triangular table as below. This view provides period-wise trends and is good for comparisons.



### 3.1.3 Major Tabs

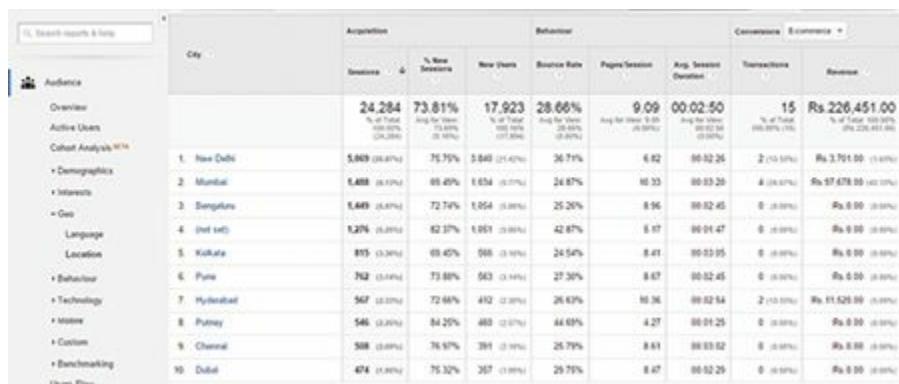
The demographics tab provides information about age and gender distribution of the visitors.



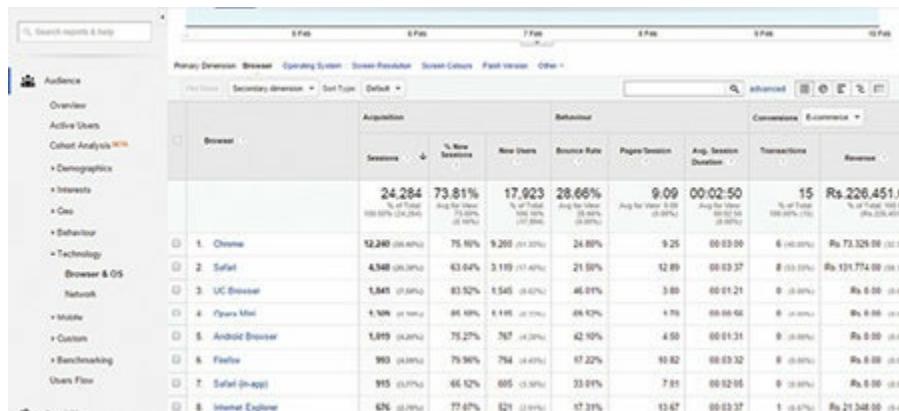
The interest Tab provides information about the visitor's life style, affinity to particular thinks etc.



The Geography tab provides information about the visitor's location – country, city and continent.



The technology tab provides information about the Operating System & Browsers from which users are visiting the site and the network also.



The Mobile tab provides information about the device category from which users have visited the site – computer, mobile or a tablet. The particular devices model can be identified for the mobile users.

The first screenshot shows the 'Device Category' report with the following data:

Device Category	Sessions	% Total Sessions	New Users	Bounce Rate	Pages/Session	Avg. Session Duration	Transactions	Revenue
1. mobile	14,369 (38.47%)	73.81%	10,604 (34.98%)	34.21%	7.04	00:02:50	8 (0.03%)	Rs. 142,972.00 (0.03%)
2. desktop	8,625 (31.09%)	46.21%	6,116 (20.12%)	26.29%	11.53	00:03:48	2 (0.03%)	Rs. 21,345.00 (0.03%)
3. tablet	1,890 (0.79%)	63.65%	1,203 (0.77%)	21.96%	54.30	00:04:05	3 (0.03%)	Rs. 42,130.00 (0.03%)

The second screenshot shows the 'Mobile Device Info' report with the following data:

Mobile Device Info	Sessions	% Total Sessions	New Users	Bounce Rate	Pages/Session	Avg. Session Duration	Transactions	Revenue
1. Apple iPhone	3,259 (10.00%)	64.87%	2,114 (7.01%)	25.36%	9.94	00:02:58	5 (0.01%)	Rs. 84,484.00 (0.01%)
2. (not set)	1,356 (3.92%)	26.57%	1,618 (5.01%)	55.23%	3.77	00:01:20	0 (0.00%)	Rs. 0.00 (0.00%)
3. Apple iPad	1,297 (3.69%)	60.60%	706 (2.22%)	19.20%	16.67	00:04:41	4 (0.01%)	Rs. 50,140.00 (0.01%)
4. Apple iPhone 5	252 (0.76%)	50.92%	151 (0.39%)	32.54%	7.12	00:02:16	0 (0.00%)	Rs. 0.00 (0.00%)
5. Samsung SM-J7102 Galaxy Grand 2	214 (0.62%)	72.95%	154 (0.39%)	28.50%	6.65	00:02:08	0 (0.00%)	Rs. 0.00 (0.00%)
6. Samsung SM-J700F Galaxy J7	181 (0.51%)	83.43%	155 (0.39%)	30.36%	6.56	00:02:21	0 (0.00%)	Rs. 0.00 (0.00%)
7. Apple iPhone 6s	145 (0.40%)	43.45%	92 (0.26%)	28.28%	8.23	00:02:39	0 (0.00%)	Rs. 0.00 (0.00%)
8. Samsung SMA300G Galaxy A3 (2015)	134 (0.38%)	70.90%	86 (0.24%)	27.61%	10.36	00:03:00	0 (0.00%)	Rs. 0.00 (0.00%)
9. Samsung GT-A7105 Galaxy	133 (0.37%)	69.92%	83 (0.24%)	33.98%	7.65	00:02:43	0 (0.00%)	Rs. 0.00 (0.00%)
10. Motorola MotoG3	129 (0.37%)	82.72%	108 (0.32%)	27.12%	5.39	00:01:34	0 (0.00%)	Rs. 0.00 (0.00%)

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.



We can further drill down to have more detailed number of each channel. The bar graphs with actual numbers help comprehend the information faster.

The Acquisition report provides a detailed breakdown of traffic sources:

Acquisition	Sessions	% New Sessions	New Users	Bounce Rate	Pages/Sess...	Avg. Session Duration	Conversions	E-commerce Conversion Rate	Transactions	Revenue
1. Organic Search	24,254	73.69%	17,894	28.56%	9.09	00:02:50	8.96%	15	Rs. 226,451.00	
2. Direct	13,268			21.19%	1.00		0.07%	1		
3. (Other)	4,169			33.32%	1.00		0.10%	1		
4. Display	2,466			35.84%	1.00		0.04%	0		
5. Paid Search	1,203			26.43%	1.00		0.00%	0		
6. Email	598			20.36%	1.00		0.00%	0		
7. Referral	517			14.51%	1.00		0.00%	0		
8. Social	446			32.96%	1.00		0.22%	1		

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.

Source/Medium	Acquisition			Behavior			Conversions			E-commerce	
	Sessions	% New Sessions	New Users	Bounce Rate	Pages/Sessions	Avg. Session Duration	E-commerce Conversion Rate	Transactions	Revenue	Visits	
1. google / organic	24,284 (0.00%)	73.81%	17,923 (0.00%)	28.66%	9.09	00:02:50	0.06%	15	Rs.2.00	100.00% (15)	
2. (direct) / none	4,169 (0.00%)	79.29%	3,264 (0.00%)	33.32%	8.39	00:02:34	0.00%	4	Rs.103.00	100.00% (4)	
3. google / cpc	2,258 (0.00%)	64.45%	1,842 (0.00%)	54.41%	4.58	00:01:28	0.00%	0	Rs.0.00	0.00% (0)	
4. Retargeting / (not set)	1,371 (0.00%)	75.36%	1,033 (0.00%)	18.82%	8.79	00:02:45	0.00%	0	Rs.0.00	0.00% (0)	
5. Referral / (not set)	167 (0.00%)	77.84%	106 (0.00%)	59.58%	3.56	00:01:15	0.00%	0	Rs.0.00	0.00% (0)	
6. Email / email	596 (0.00%)	32.09%	476 (0.00%)	29.36%	16.13	00:04:55	0.00%	0	Rs.0.00	0.00% (0)	
7. yahoo / organic	224 (0.00%)	84.73%	145 (0.00%)	16.36%	13.48	00:04:49	0.00%	0	Rs.0.00	0.00% (0)	

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.

Campaign	Acquisition			Behavior			Conversions			E-commerce	
	Code	Cost	CPC	Sessions	Bounce Rate	Pages/Sessions	Avg. Session Duration	E-commerce Conversion Rate	Transactions	Revenue	
1. Marketing	1,029 (0.00%)	Rs.7,682.54 (0.00%)	Rs.7.66	2,858 (0.00%)	54.41%	4.58	00:00:00	0.00%	0	0.00%	0.00% (0)
2. Latent INDIA RISK	1,052 (0.00%)	Rs.7,676.34 (0.00%)	Rs.7.66	8.00%	0.00%	0.00	0.00%	0.00%	0	0.00%	0.00% (0)
3. Search - India RISK	846 (0.00%)	Rs.15,204.67 (0.00%)	Rs.17.97	777 (0.00%)	24.56%	0.78	0.00%	0.00%	0	0.00%	0.00% (0)
4. Suits & Kurtis India	579 (0.00%)	Rs.4,523.81 (0.00%)	Rs.7.71	335 (0.00%)	22.99%	5.74	0.00%	0.00%	0	0.00%	0.00% (0)
5. Retail - Latent INDIA	89 (0.00%)	Rs.408.74 (0.00%)	Rs.4.50	8.00%	0.00%	0.00	0.00%	0.00%	0	0.00%	0.00% (0)
6. Retail Suits Online MR	46 (0.00%)	Rs.533.69 (0.00%)	Rs.11.78	52 (0.00%)	58.00%	0.86	0.00%	0.00%	0	0.00%	0.00% (0)
7. Search	22 (0.00%)	Rs.20.43 (0.00%)	Rs.1.29	0.00%	67.05%	12.76	0.00%	0.00%	0	0.00%	0.00% (0)
8. (ind out)	8 (0.00%)	Rs.0.00 (0.00%)	Rs.0.00	2 (0.00%)	50.00%	2.00	0.00%	0.00%	0	0.00%	0.00% (0)
9. Display/EDGS Remarketing	0 (0.00%)	Rs.0.00 (0.00%)	Rs.0.00	0.00%	0.00%	0.00	0.00%	0.00%	0	0.00%	0.00% (0)

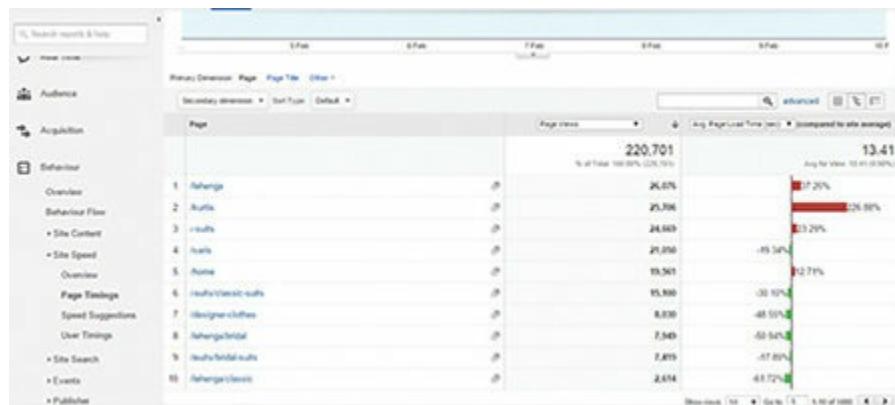
The **Behavior** tab provides information about the customer flow, pages viewed, the landing pages, exit pages and so on.

Page	Page Views	Unique Page Views	Avg. Time on Page	Entrances	Bounce Rate	% Exit	Page Value	
							% of Total Page Views	% of Total Unique Page Views
1. About	220,701 (0.00%)	58,100 (0.00%)	00:00:21	24,284 (0.00%)	28.66%	11.00%	Rs.44.07	100.00% (0.00%)
2. Products	26,679 (0.00%)	4,358 (0.00%)	00:00:18	3,645 (0.00%)	29.78%	12.30%	Rs.0.00	100.00% (0.00%)
3. Home	25,706 (0.00%)	3,862 (0.00%)	00:00:18	2,539 (0.00%)	28.12%	9.68%	Rs.19.32	100.00% (0.00%)
4. Services	24,659 (0.00%)	3,111 (0.00%)	00:00:17	1,920 (0.00%)	22.83%	7.56%	Rs.24.67	100.00% (0.00%)
5. Home	19,561 (0.00%)	9,314 (0.00%)	00:00:32	5,475 (0.00%)	5.50%	17.52%	Rs.36.03	100.00% (0.00%)
6. About-us	15,190 (0.00%)	2,518 (0.00%)	00:00:18	1,795 (0.00%)	22.40%	10.34%	Rs.28.44	100.00% (0.00%)
7. Designer-clothes	8,850 (0.00%)	1,545 (0.00%)	00:00:14	263 (0.00%)	19.01%	15.43%	Rs.26.07	100.00% (0.00%)
8. Designer-brands	7,340 (0.00%)	1,057 (0.00%)	00:00:14	137 (0.00%)	24.09%	4.70%	Rs.0.00	100.00% (0.00%)
9. Apparel-store	7,019 (0.00%)	1,026 (0.00%)	00:00:14	343 (0.00%)	21.57%	6.32%	Rs.41.55	100.00% (0.00%)
10. Authentic-classic	2,818 (0.00%)	379 (0.00%)	00:00:14	9 (0.00%)	11.11%	3.43%	Rs.0.00	100.00% (0.00%)

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.

Primary Dimension	Landing Page	Acquisition			Behavior			Conversions		
		Sessions	% New Sessions	New Users	bounce Rate	Page Session	Avg. Session Duration	Transactions	Revenue	
	Landing Page	24,284	73.81%	17,923	28.66%	9.09	00:02:50	15	Rs.226,451.00	
	1 - Home	5,475 (19%)	69.52%	3,279 (13%)	9.95%	10.28	00:04:34	00:00:00	Rs.150,315.00 (~70%)	
	2 - Change	3,446 (13%)	87.99%	2,114 (9%)	20.74%	7.17	00:02:10	0	Rs.0.00 (~0%)	
	3 - Ruffle	2,329 (8%)	83.26%	2,114 (9%)	28.52%	7.14	00:02:04	1	Rs.1,858.00 (~1%)	
	4 - Ruffles	2,045 (7%)	80.95%	1,629 (7%)	29.27%	8.88	00:02:36	0	Rs.0.00 (~0%)	
	5 - ruffles	1,500 (5%)	75.83%	1,456 (6%)	22.53%	8.48	00:02:33	0	Rs.0.00 (~0%)	
	6 - ruffles ruffles	1,759 (6%)	81.30%	1,421 (6%)	22.40%	8.19	00:02:24	1	Rs.1,950.00 (~1%)	
	7 - Indian blue cotton cult online	467 (2%)	79.87%	373 (2%)	45.95%	3.46	00:01:23	0	Rs.0.00 (~0%)	
	8 - India's embroidered cult indigo clothing	499 (2%)	68.80%	286 (1%)	34.01%	8.88	00:00:39	0	Rs.0.00 (~0%)	
	9 - Youth fashion outfit	343 (1%)	81.34%	279 (1%)	24.57%	5.54	00:01:42	0	Rs.0.00 (~0%)	

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.



The users searches at the site is captured in this section. The site search is important because it provides information 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.



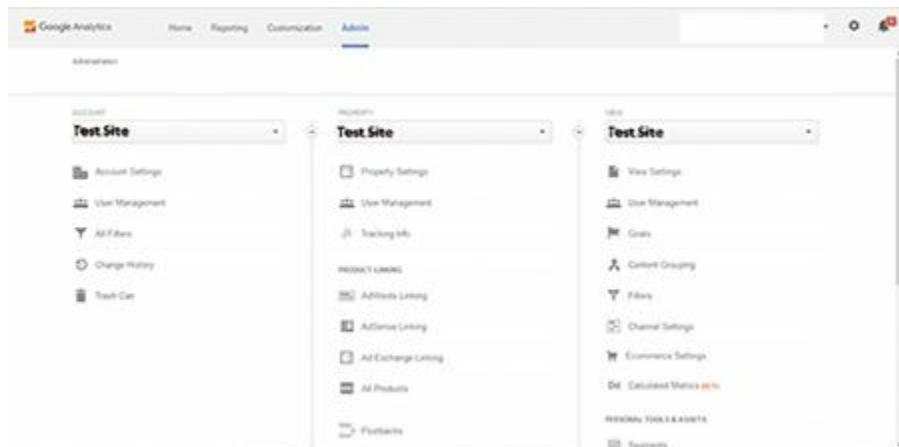
The conversion section is about the bottom of the funnel. This section tracks the visitors action such as completion of a goals, completion of the purchase for ecommerce site, the products being purchase etc. the important metrics available are transaction, conversion rates, revenue, average revenue per order and so on. This section is very important because the intent of the website is being captured here.



### 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.



Click on +New Goals or you can edit existing goals by setting editing it.

The screenshot shows the AdWords interface with the 'Test Site' selected. On the left, there's a sidebar with options like 'View Settings', 'User Management', 'Goals' (which is highlighted in red), 'Content Grouping', 'Filters', 'Channel Settings', 'Ecommerce Settings', 'Calculated Metrics', 'Personal Tools & Assets', and 'Help'. The main area has a header 'Import from Analytics' with a 'New Goal' button. Below it is a table with one row:

Goal	ID	Period	Action
Place an order	Goal ID 1 / Goal Set 1	Past 7-day convergence	Recording

At the bottom left, it says '19 goals left'.

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.

The screenshot shows the 'Goal setup' step. At the top, there's a title 'Goal setup' with a back arrow. Below it is a section titled 'Template' with the sub-section 'Select a template to start with a pre-filled configuration'. There are four main categories with sub-options:

- 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 creation.
- INQUIRY**:
  - 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.
- ENGAGEMENT**:
  - Compare information: Compared features, products, or options.
  - Share / social connect: Shared to a social network or emailed.
  - Newsletter sign up: Signed up for newsletter or mailings.
  - Add to cart: Saved product to shopping cart.

Below these categories, there's a section for 'Smart Goal' with a link to learn more about measuring engaged visitors. At the bottom, there's a 'Custom' option.

Check the checkout complete option

**Goal setup**

**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 creation

**INQUIRY**

- 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

**ENGAGEMENT**

- Compare information Compared features, products, or options
- Share / social connect Shared to a social network or emailed
- Newsletter sign up Signed up for newsletter or mailings
- Add to cart Saved product to shopping cart

**Smart Goal**

Measure the most engaged visits to your website and automatically turn those visits into Goals. Then use those Goals to improve your AdWords bidding. [Learn more](#)

In the goal detail you have you provide the url of the goals. Assuming our checkout page is [www.test.com/checkout](http://www.test.com/checkout), we add /checkout in the url.

**Goal setup** Edit  
Template: Checkout complete

**Goal description** Edit  
Name: Checkout complete  
Goal type: Destination

**Goal details**  
**Destination**  
Equals to   Case sensitive  
For example, use My Screen for an app and thank-you.html instead of www.example.com/thank-you.html for a web page.

**Value** optional  
 OFF Assign a monetary value to the conversion.

**Path** optional  
 OFF Specify a path you expect traffic to take towards the destination. Use it to analyze the entrance and exit points that impact your Goal.

**Verify this Goal** See how often this Goal would have converted based on your data from the past 7 days.

**Save** **Cancel**

Click on save. The goal is created.

<b>NEW GOAL</b>		Import from gallery	<input type="button" value="Search"/>	
<input type="checkbox"/>	Goal	ID	past 7 day conversions	Recording
<input type="checkbox"/>	Checkout complete	Goal ID 2 / Goal Set 1	0	<input type="checkbox"/> ON
<input type="checkbox"/>	Place an order	Goal ID 1 / Goal Set 1	2140	<input type="checkbox"/> ON

18 goals left

Now you can view your goals in all the tabs and select the goals in case you want to see the goals completion.

Real Time		Acquisition						Behavior						Conversions					
Audience		Browser						Page / Session						Conversion					
Overview		Sessions			% New Sessions			New Users			Source Rate			Avg. Session Duration			Checkout complete (2 Conversions)		
Demographics		Sessions			% New Sessions			New Users			Source Rate			Avg. Session Duration			Checkout complete (2 Conversions)		
1. Chrome	48,876 (33%)	67.3%	32,772 (31%)	37.39%	3.32	00:01:53	0.30%	0	0	0	0.00	0	0	0.00	0	0	0	0	0
2. Firefox	4,376 (3%)	64.7%	4,376 (4%)	65.40%	2.29	00:01:53	0.30%	0	0	0	0.00	0	0	0.00	0	0	0	0	0
3. UC Browser	3,737 (3%)	71.52%	4,102 (4%)	7.62%	5.42	00:02:43	0.30%	0	0	0	0.00	0	0	0.00	0	0	0	0	0
4. Opera Mini	4,426 (3%)	78.2%	3,472 (3%)	4.30%	4.48	00:01:40	0.30%	0	0	0	0.00	0	0	0.00	0	0	0	0	0
5. Android Browser	3,588 (3%)	73.58%	2,549 (2%)	6.45%	3.13	00:01:20	0.30%	0	0	0	0.00	0	0	0.00	0	0	0	0	0
6. Safari	2,278 (2%)	77.12%	1,834 (2%)	7.44%	3.81	00:01:18	0.30%	0	0	0	0.00	0	0	0.00	0	0	0	0	0
7. Internet Explorer	1,894 (2%)	72.2%	1,448 (2%)	59.08%	2.34	00:01:28	0.30%	0	0	0	0.00	0	0	0.00	0	0	0	0	0
8. Opera	1,066 (1%)	62.30%	825 (1%)	43.20%	2.84	00:01:58	0.30%	0	0	0	0.00	0	0	0.00	0	0	0	0	0
9. Safari (in-APP)	296 (0.2%)	70.1%	209 (0.2%)	2.36%	4.89	00:01:15	0.30%	0	0	0	0.00	0	0	0.00	0	0	0	0	0
10. Edge	168 (0.2%)	66.07%	115 (0.1%)	54.17%	5.66	00:00:59	0.30%	0	0	0	0.00	0	0	0.00	0	0	0	0	0

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.

Goal details

Destination: `REGULAR EXPRESSION` `/CHECKOUT/SUCCESS` Case sensitive

Value: `OFF` Assign a monetary value to the conversion.

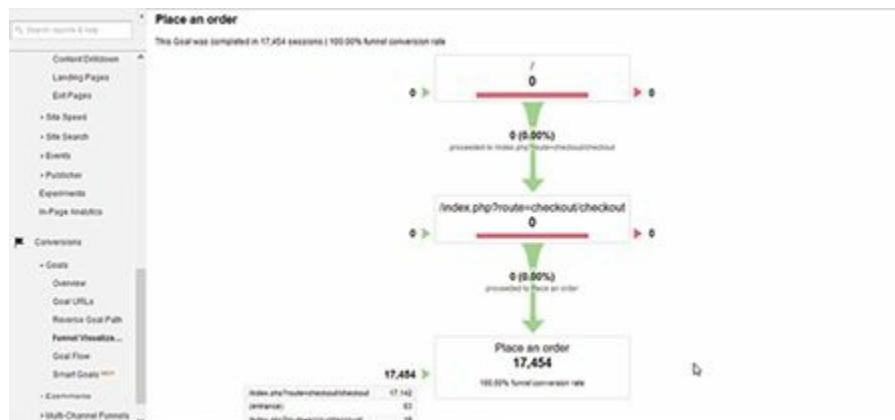
Funnel: `ON` Use an app screen name string or a web page URL for each step. For example, use `My Screen` for an app and `thankyou.html` instead of `www.example.com/thankyou.html` for a web page.

Step Name	Screen/Page	Required?
1	Home Page	NO
2	Index.php?route=checkout/checkout	Checkout

Verify this Goal: See how often this Goal would have converted based on your data from the past 7 days.

**Save** **Cancel**

From Funnel Visualization you can view the funnel created. The funnel just created will take some time to populate.



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

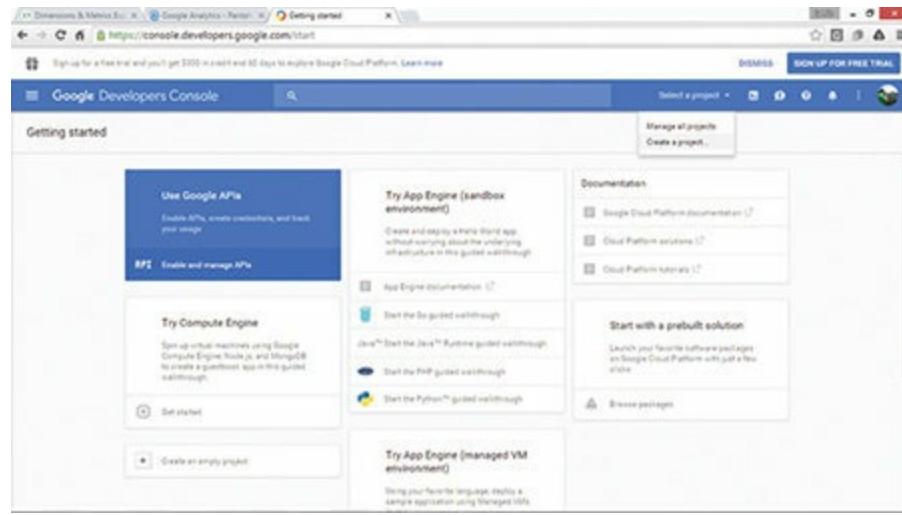
Quite often we need to pull data out of the Google Analytics in the Excel to do more complex data analytics. For a limited data set excel export works but when data set is large and the dimension is more than two it is really difficult situation. The Pentaho data integration with Google Analytics is 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.

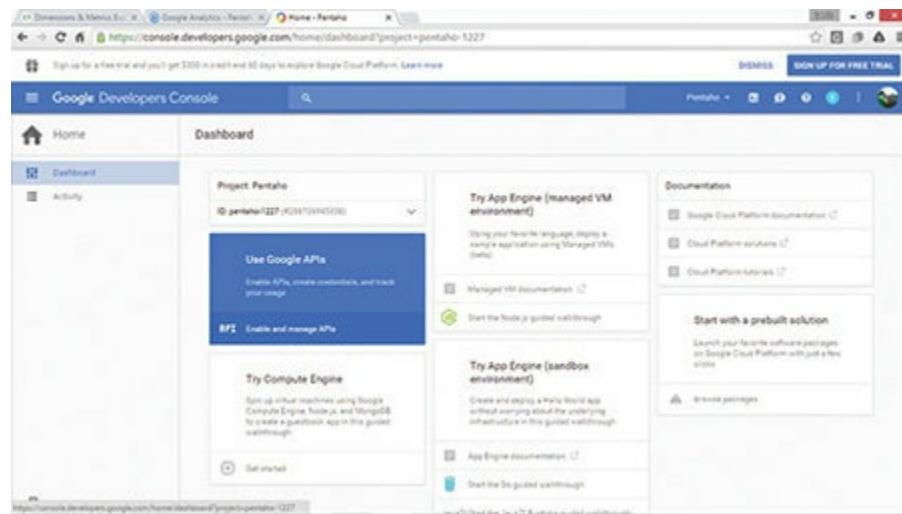
The stepwise process to create Google API is given in the <http://wiki.pentaho.com/display/EAI/Google+Analytics>

Go to <http://console.developers.google.com> 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.

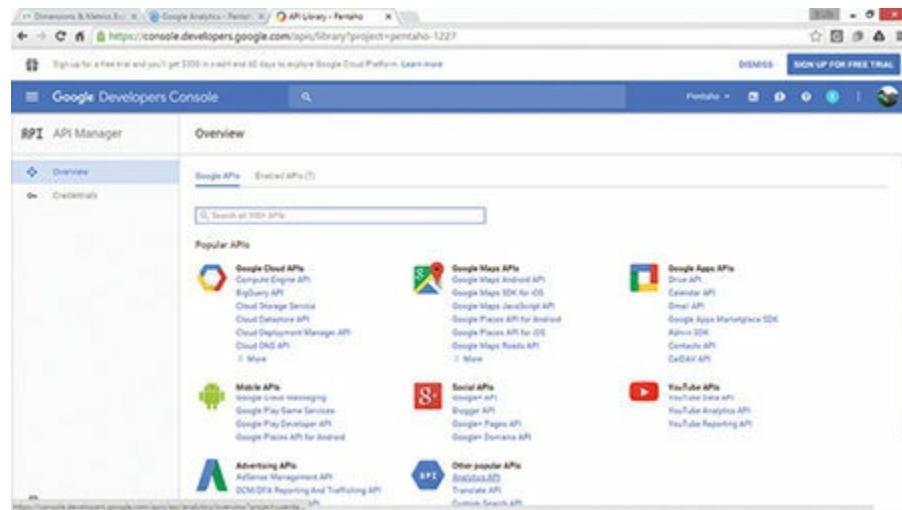
Create a new project



Select Google API section for enabling the API.



Select Analytics API



Enable the API by clicking on the Enable button

The screenshot shows the Google Developers Console API Manager Overview page for the Analytics API. It includes sections for enabling the API, using credentials, and server-to-server interactions. A diagram illustrates the OAuth 2.0 flow: Your App → User consent → User data, and Your service → Authorization → Google service.

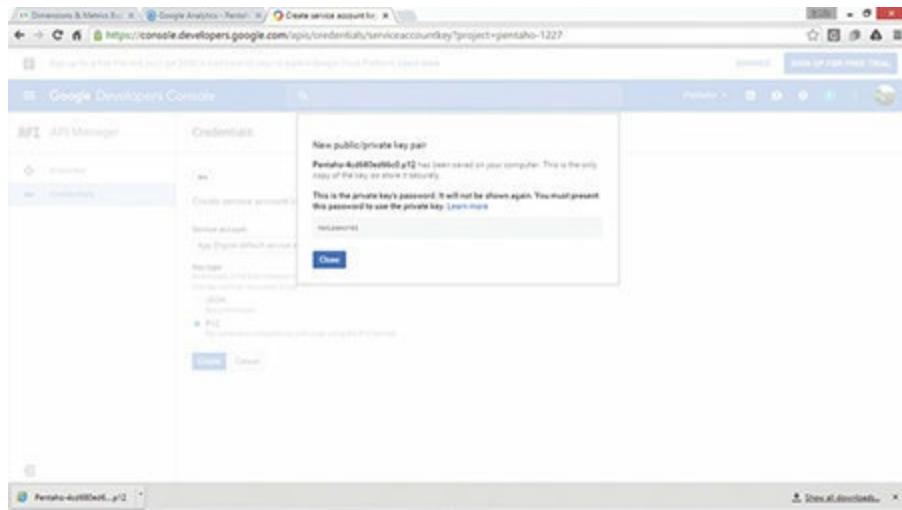
**Once the API is enabled, you have to create credential for using it in the API call. Select Service Account Key**

The screenshot shows the Google Developers Console Credentials page. Under the 'Create credentials' dropdown, 'Service account key' is selected. A client ID is shown as 28873943238-uhfhe7ka7m0l3ec0gjhd4entc.apps.googleusercontent.com.

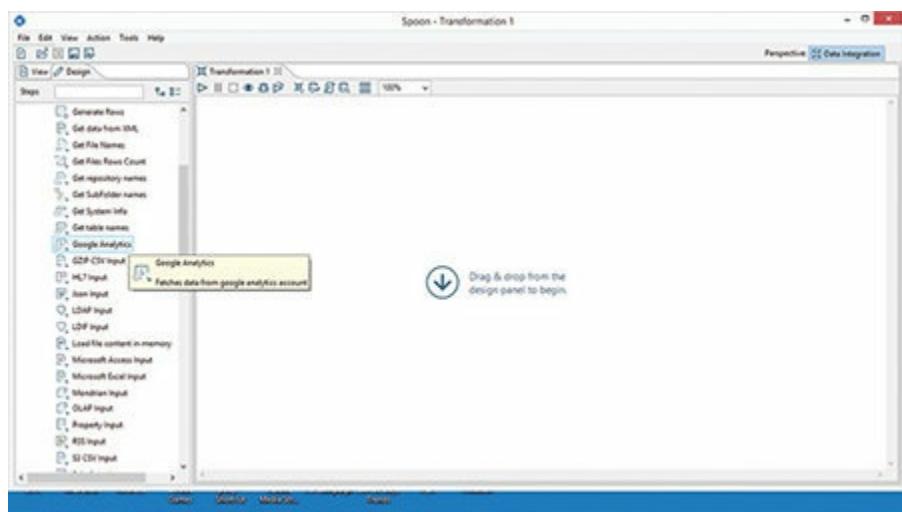
**Select service account and Key type as P12.**

The screenshot shows the Google Developers Console Create service account key page. It displays a form for creating a service account key. The 'Service account' dropdown is set to 'App Engine default service account'. The 'Key type' dropdown has two options: 'JSON' (selected) and 'P12' (radio button). Below the 'P12' option is the note: 'For backwards compatibility with tools using the PKCS12 format.'

P12 file will be downloaded into your system. Keep it safe in a location which is relatively permanent.



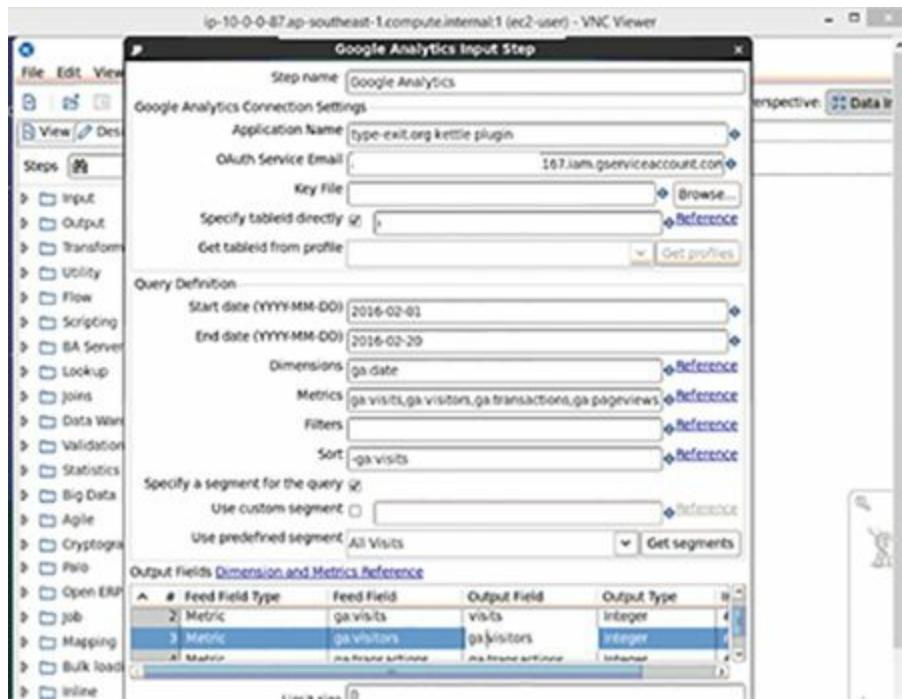
Now our API related steps are completed open Pentaho Data Integration. Create a new transformation, select Google Analytic from the Input.



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.



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.

#	ga-date	ga-visits	ga:visitors	ga:transactions	ga:pageviews
1	2016/02/20 00:00:00.000	913669	759299	26850	1882
2	2016/02/19 00:00:00.000	859427	700673	28204	1865
3	2016/02/18 00:00:00.000	852740	679188	49526	2052
4	2016/02/10 00:00:00.000	775776	641899	20030	1566

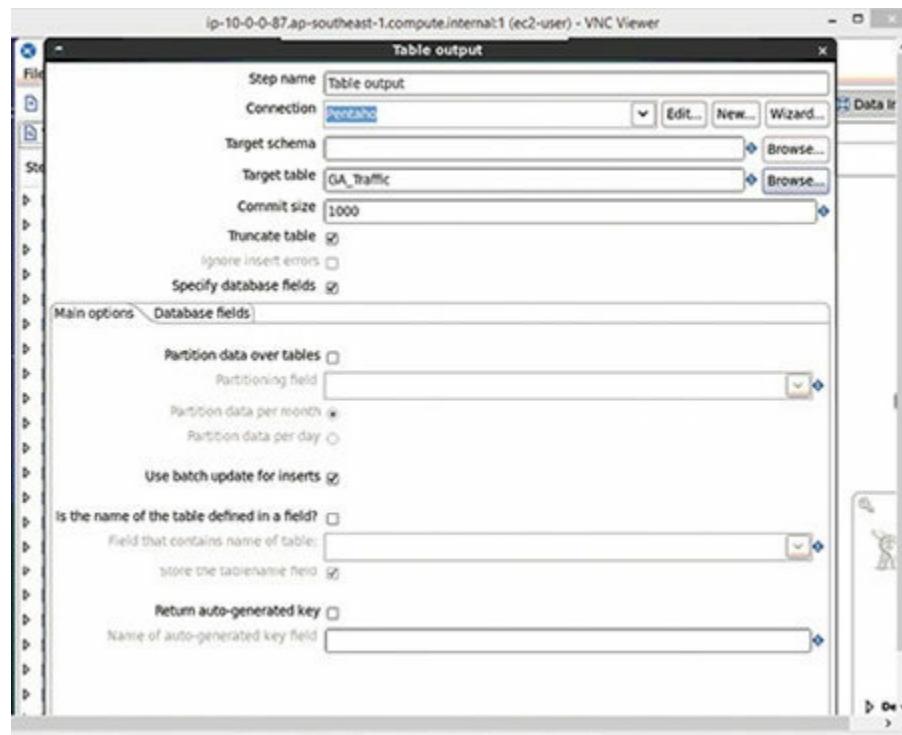
Now our Google Analytics input is ready, add a table output in the transformation. Make sure that a table with same columns is created in the database. In this example we have created a table GA\_Traffic with date, visits, visitors, transactions and pageviews in Pentaho database. Connect the database and select the table GA\_Traffic.

```
V2 ip-10-0-0-87.ap-southeast-1.compute.internal:1 (ec2-user) - VNC Viewer
ec2-user@ip-10-0-0-87:~ 
File Edit View Search Terminal Help
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

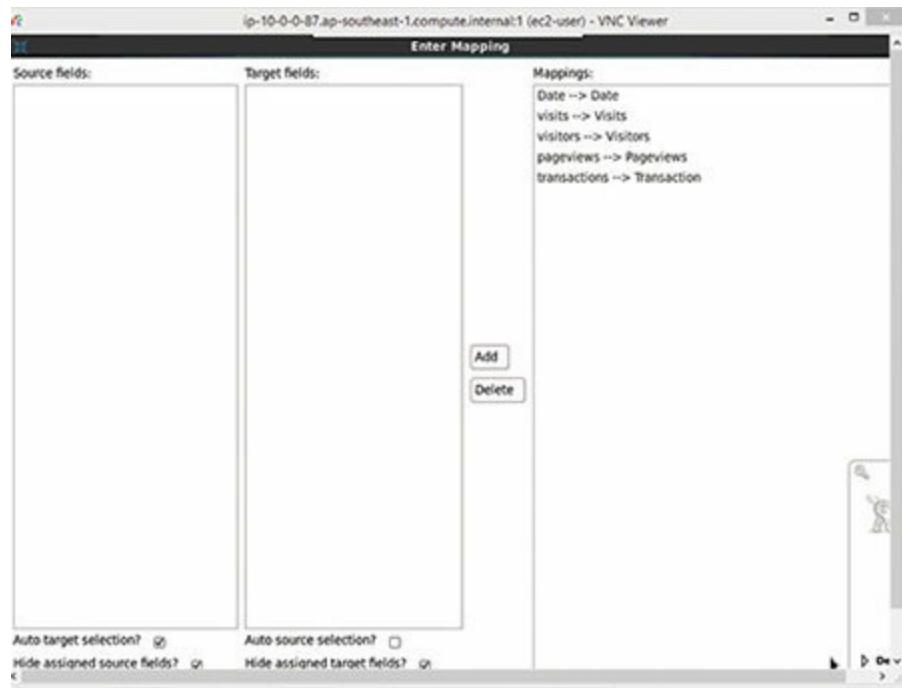
mysql> use Pentaho;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+-----+
| Tables_in_Pentaho |
+-----+
| Calender          |
| Customer          |
| OLAP              |
| Product           |
| Sales              |
+-----+
5 rows in set (0.00 sec)

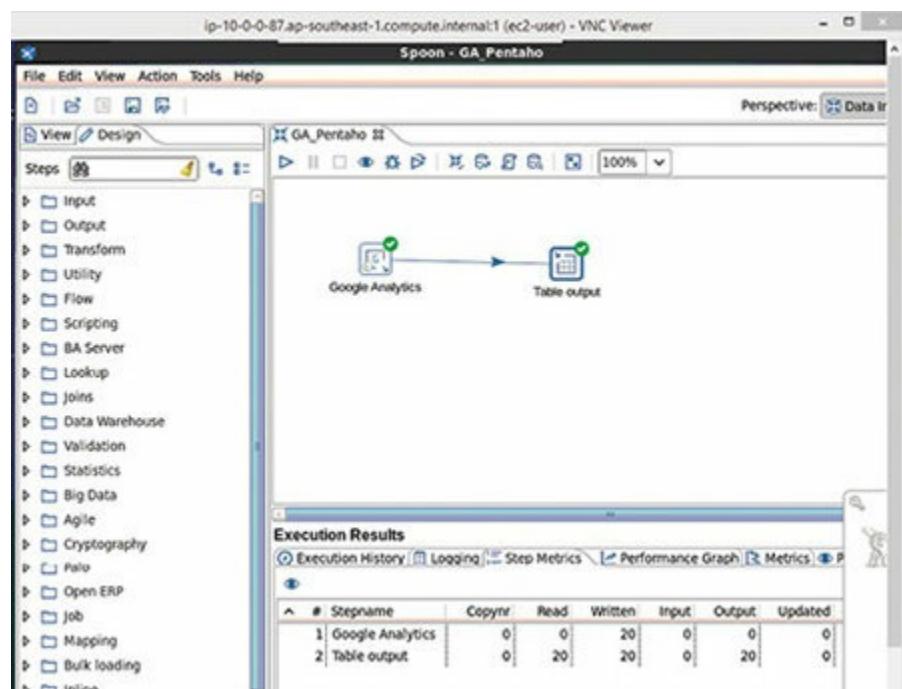
mysql> create table GA_Traffic ( Date date,
->   Visits Integer,
->   Visitors Integer,
->   Transaction Integer,
->   Pageviews Integer);
```



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.



Verify this by querying on table GA\_Traffic, we can see 20 days day wise visits, visitor, transaction and pageviews.

MySQL Workbench screenshot showing the results of a query against the GA\_Traffic table. The table has columns: Date, Visits, Visitors, Transaction, and Pageviews. The data spans from February 10, 2016, to February 20, 2016.

Date	Visits	Visitors	Transaction	Pageviews
2016-02-10	913669	759299	26850	1882859
2016-02-19	859427	700673	28204	1865327
2016-02-18	852740	679188	49526	2052848
2016-02-17	775776	641899	20030	1566930
2016-02-11	742065	618829	19471	1552858
2016-02-16	712183	575563	19179	1554952
2016-02-15	644087	519729	17664	1397920
2016-02-12	627280	498936	18346	1366582
2016-02-13	546825	421086	16379	1231180
2016-02-02	533430	397649	17071	1231563
2016-02-08	532786	417059	18328	1262262
2016-02-05	528441	407172	19959	1258673
2016-02-04	517442	385717	18674	1203873
2016-02-03	515171	384178	16831	1194755
2016-02-01	508299	392233	18679	1243282
2016-02-09	507589	410410	16429	1148346
2016-02-06	502168	392861	17030	1196537
2016-02-07	485968	381425	16532	1147681
2016-02-14	464078	367287	15488	1049551

20 rows in set (0.00 sec)

Now let's add one more dimensions ga:devicecategory.

MySQL Workbench screenshot showing the Google Analytics Connection Settings dialog. The connection is named 'type-exit.org\_kettle plugin' and is configured to use OAuth Service Account authentication with email 'askmebaazar@askmebaazar-1167.iam.gserviceaccount.com'. The table ID is specified as 'ga:85464358'. The Query Definition section shows the start date as 2016-02-01, end date as 2016-02-20, dimensions as 'ga:date, ga:devicecategory', metrics as 'ga:visits, ga:visitors, ga:transactions, ga:pageviews', and sort by 'ga:visits'. The Output Fields section shows the feed field type as 'Dimension' and the output field as 'Device'.

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.

The screenshot shows the PDI Data Integration interface. At the top, it displays 'Google Analytics Connection Settings' with an application name 'type-exit.ecn.kettle.plugin'. Below this is a preview window titled 'Rows of step: Google Analytics (4 rows)'. The data is presented in a table with columns: #, Date, Device, visits, visitors, transactions, and pageviews. The data rows are:

#	Date	Device	visits	visitors	transactions	pageviews
1	2016/02/20 00:00:00.000	desktop	519747	437738	15740	9421
2	2016/02/19 00:00:00.000	desktop	447638	367484	13585	8816
3	2016/02/18 00:00:00.000	desktop	419868	336654	26267	9471
4	2016/02/10 00:00:00.000	desktop	411622	339664	8572	7831

At the bottom of the interface, there are tabs for 'Bulk load', 'Inline', and 'Experiments', along with buttons for 'OK', 'Get Fields', 'Preview', and 'Cancel'.

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

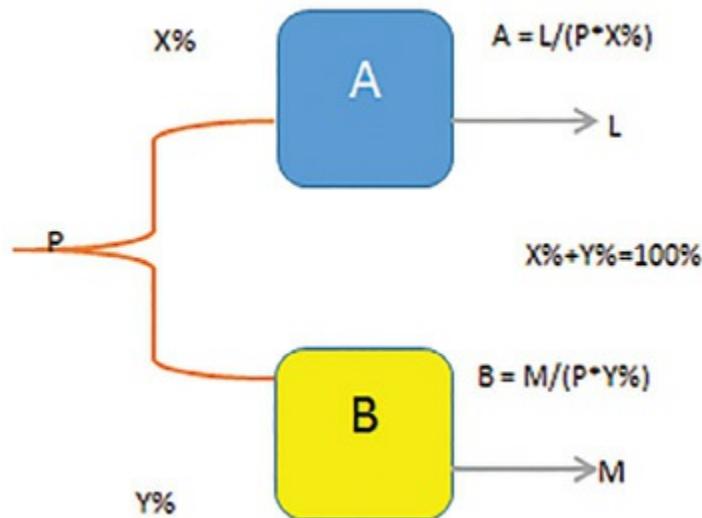
When you are not sure about which pair of shoe to wear in an occasion or when you are not sure about how to tell a sad news to your relatives or when you are not sure how your parent will react to your grades or when you are not sure about how to say four letter word to your dearest one and if you have 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 kind 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  $H_0: p_a = p_b$

Alternative Hypothesis  $H_1: p_a \neq p_b$

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  $p_t = [n_1 p_a + n_2 p_b] / (n_1 + n_2)$

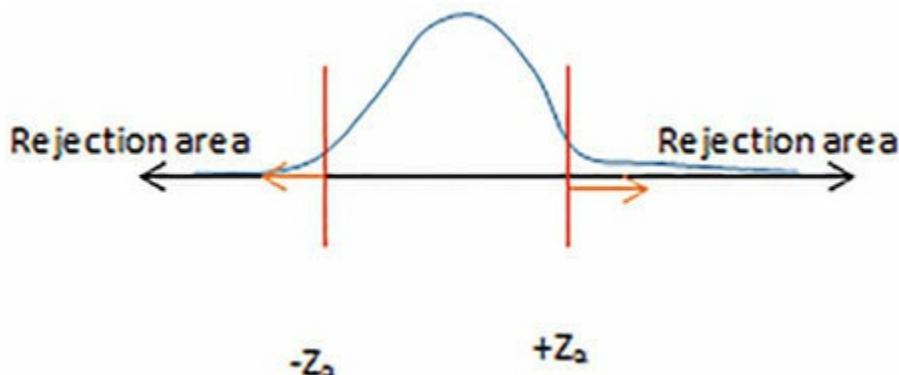
Standard error of the Proportion difference is calculated as

$$\sigma_{p_a - p_b} = \sqrt{(p_t * (1-p_t) / n_1) + (p_t * (1-p_t) / n_2)}$$

The Z score is calculate as  $Z_1 = (p_a - p_b) / \sigma_{p_a - p_b}$

Based on the confidence interval we can get alpha  $\alpha = 100\% - \text{confidence interval}$  and the critical value  $Z_\alpha$  using Standardized Normal distribution table.

Reject null hypothesis if  $Z_1 > Z_\alpha$  or  $Z_1 < -Z_\alpha$



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 $1-\alpha$	TYPE I Error $\alpha$
	False	TYPE II Error $\beta$	CORRECT $1-\beta$ (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.  $\alpha$  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  $\alpha$  ( $1 - \alpha$  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.

Example: A/B testing of a Page is done with below data

	Sample 1 (n1)	Sample 2(n2)
Traffic	100	120
Click	5	7
Click through Rate (p)	$5/100 = 5\%$	$9/120 = 7.5\%$

As  $p_1 < p_2$  we can conclude that experiment 2 is better than experiment 1 but we have to prove it statistically. Right?

$$pt = [(100*0.05) + (120*0.075)] / (120+100) = 6.3\%$$

$$\text{SE}_{p_a - p_b} = \sqrt{[6.3\% * (1 - 6.3\%) / 100] + [6.3\% * (1 - 6.3\%) / 120]} = 3.2\%$$

$$Z_1 = (7.5\% - 5\%) / 3.2\% = 0.78$$

Assume 95% confidence interval,  $\alpha = 5\%$  (level of significance).

$$Z_{\alpha} = +1.96 \text{ or } -1.96.$$

As 0.78 is within  $\pm 1.96$ , 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.

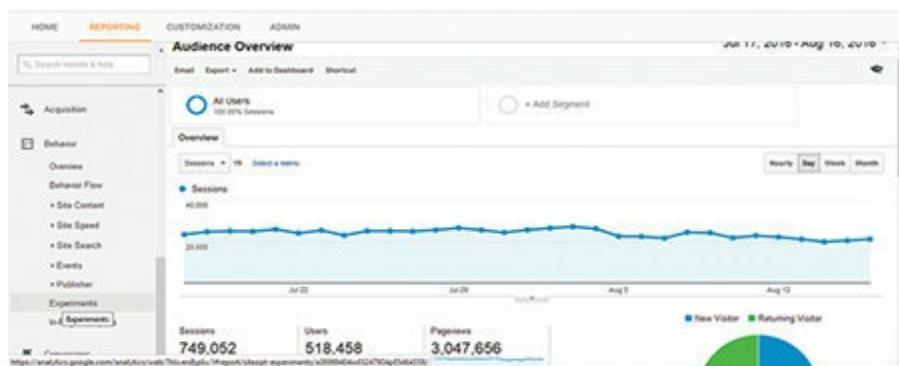
Hope this example is useful. Here I have taken CTR as the result; you could have taken conversion or any other success events as the goal. The calculation will remain same but one just needs to change the numbers in the formula.

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**



Click on create experiment button to start a new experiment



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.

In Step-2, add the original pages and its variants.

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.

```

<!-- Google Analytics Content Experiment code -->
<script>function utm(e,a,b,c,d,f){var g="";g+=e?e:"";g+=a?"/"+a:"";g+=b?"/"+b:"";g+=c?"/"+c:"";g+=d?"/"+d:"";g+=f?"/"+f:"";document.cookie="utm="+g+";path=/";};function utma(e,a,b,c,d,f){var g="";g+=e?e:"";g+=a?"/"+a:"";g+=b?"/"+b:"";g+=c?"/"+c:"";g+=d?"/"+d:"";g+=f?"/"+f:"";document.cookie="utma="+g+";path=/";};function utmc(e,a,b,c,d,f){var g="";g+=e?e:"";g+=a?"/"+a:"";g+=b?"/"+b:"";g+=c?"/"+c:"";g+=d?"/"+d:"";g+=f?"/"+f:"";document.cookie="utmc="+g+";path=/";};function utmd(e,a,b,c,d,f){var g="";g+=e?e:"";g+=a?"/"+a:"";g+=b?"/"+b:"";g+=c?"/"+c:"";g+=d?"/"+d:"";g+=f?"/"+f:"";document.cookie="utmd="+g+";path=/";};function utme(e,a,b,c,d,f){var g="";g+=e?e:"";g+=a?"/"+a:"";g+=b?"/"+b:"";g+=c?"/"+c:"";g+=d?"/"+d:"";g+=f?"/"+f:"";document.cookie="utme="+g+";path=/";};function utmt(e,a,b,c,d,f){var g="";g+=e?e:"";g+=a?"/"+a:"";g+=b?"/"+b:"";g+=c?"/"+c:"";g+=d?"/"+d:"";g+=f?"/"+f:"";document.cookie="utmt="+g+";path=/";};function utmz(e,a,b,c,d,f){var g="";g+=e?e:"";g+=a?"/"+a:"";g+=b?"/"+b:"";g+=c?"/"+c:"";g+=d?"/"+d:"";g+=f?"/"+f:"";document.cookie="utmz="+g+";path=/";};function utmz(e,a,b,c,d,f){var g="";g+=e?e:"";g+=a?"/"+a:"";g+=b?"/"+b:"";g+=c?"/"+c:"";g+=d?"/"+d:"";g+=f?"/"+f:"";document.cookie="utmz="+g+";path=/";};</script>
<!-- End of Google Analytics Content Experiment code -->

```

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 “I'm 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 sold together as combo, 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.

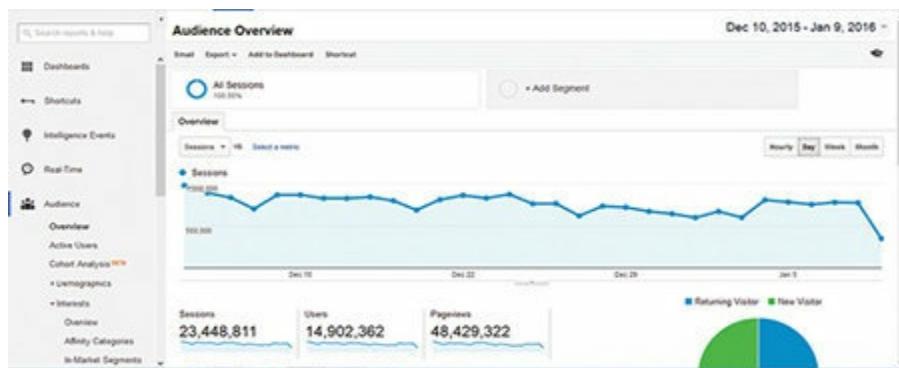
# Section - I

## 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**.



In the Segment Panel you will see all the current available segments for the user. There are many inbuilt segments in Google Analytics.

The screenshot shows the 'Audience Overview' interface in Google Analytics. On the left, there's a sidebar with navigation links for Dashboards, Shortcuts, Intelligence Events, Real-Time, Audience (Overview, Active Users, Cohort Analysis), Demographics, Interests, Geo, Behavior, Technology, Mobile, and Custom. The main area is titled 'Audience Overview' with a date range of 'Feb 22, 2016 - Feb 22, 2016'. A modal window titled '+ New Segment' is open, showing a list of segments under 'View dimensions'. The 'All' dimension is selected, and the 'All Users' segment is highlighted. Other segments listed include Bounced Sessions, Converters, Direct Traffic, Emailer, Google CPC, Made a Purchase, Mobile and Tablet Traffic, Mobile Traffic, Multi-session Users, New Users, Non-bounce Sessions, Non-Converters, Organic Traffic, Paid Traffic, Performed Site Search, Referral Traffic, and Returning Users. At the bottom of the modal are 'Apply' and 'Cancel' buttons.

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**

This screenshot is identical to the one above, showing the '+ New Segment' modal in the Google Analytics Audience Overview interface. It displays the same list of segments under 'View dimensions' with 'All' selected and 'All Users' highlighted. The modal includes 'Apply' and 'Cancel' buttons at the bottom.

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.

This screenshot shows the '+ Add Segment' modal in the Google Analytics Audience Overview interface. The 'Demographics' tab is selected. On the left, there are tabs for Demographics, Technology, Behavior, Date of First Session, Traffic Sources, Ecommerce, Advanced, Conditions, and Sequences. On the right, there's a 'Summary' section with a large circle showing '100.00% of users'. Below it, it says '20.14%' and '100.00% of sessions'. At the bottom right of the summary section is a link 'No Filters'. The modal has 'Next' and 'Cancel' buttons at the bottom.

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.

The screenshot shows the Google Analytics Segment builder. On the left, there's a sidebar with navigation links like Dashboards, Shortcuts, Intelligence Events, Real-Time, Audience, Overview, Active Users, Cohort Analysis, Demographics, Interests, Geo, Behavior, Technology, Mobile, Custom, and Benchmarking. The main area is titled 'All Users' and shows a 'Demographics' segment. Under 'Demographics', the 'Age' filter is set to '25-34' and '35-44'. The summary on the right indicates 20.14% of users.

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.

This screenshot shows the same Segment builder interface as the previous one, but the 'Age' filter has been changed to '25-34' and '35-44'. The summary now shows 4.26% of users.

Let's us create another segment with medium as Organic and name it Organic Traffic.

This screenshot shows the Segment builder with a 'Traffic Sources' segment selected. The 'Medium' filter is set to 'organic'. The summary indicates 30.61% of users.

Now we have used Organic Traffic segment, all the information in any tab will show only organic traffic.



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.



In any tab you will see the comparison of the segments selected.

Device Category	Acquisition			Behavior			Conversions			eCommerce	
	Sessions	% New Sessions	New Users	Bounce Rate	Pages / Session	Avg. Session Duration	Transactions	Revenue	E-commerce Revenue	E-commerce Conversion Rate	
Organic Traffic	375	45.87%	172	10.40%	21.63	00:10:15	1	Rs.307.00	Rs.307.00	0.27%	
All Users	1,073	61.88%	664	19.94%	17.09	00:07:46	2	Rs.1,306.00	Rs.1,306.00	0.19%	
1. desktop											
Organic traffic	264	48.52%	99	10.16%	18.14%	00:06:21	1	Rs.307.00	Rs.307.00	0.49%	
All Users	867	64.89%	369	10.16%	24.00%	00:09:43	2	Rs.1,306.00	Rs.1,306.00	0.34%	
2. mobile											
Organic traffic	167	41.82%	79	10.16%	9.80%	00:18:14	0	Rs.0.00	Rs.0.00	0.00%	
All Users	458	57.58%	219	10.16%	2.22%	00:16:52	0	Rs.0.00	Rs.0.00	0.00%	
3. tablet											
Organic traffic	4	75.00%	3	10.00%	25.00%	00:00:56	0	Rs.0.00	Rs.0.00	0.00%	
All Users	16	66.67%	13	10.00%	3.88%	00:00:49	0	Rs.0.00	Rs.0.00	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 0	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
All Users 13,631 users	100.00%	3.89%	1.7%	1.11%	0.51%	0.38%	0.35%	0.18%	0.19%	0.20%	0.11%
Feb 9, 2016 1,344 users	100.00%	2.53%	1.71%	1.05%	0.97%	0.85%	0.87%	0.22%	0.10%	0.10%	0.22%
Feb 10, 2016 1,235 users	100.00%	2.66%	3.41%	1.14%	0.49%	0.49%	0.49%	0.33%	0.24%	0.09%	0.09%
Feb 11, 2016 1,333 users	100.00%	3.01%	1.30%	0.79%	0.29%	0.36%	0.29%	0.22%	0.59%	0.36%	0.14%
Feb 12, 2016 1,294 users	100.00%	2.62%	1.15%	0.82%	0.22%	0.13%	0.13%	0.13%	0.29%	0.02%	0.04%
Feb 13, 2016 1,244 users	100.00%	2.66%	0.75%	0.27%	0.36%	0.27%	0.33%	0.11%	0.11%	0.02%	0.31%
Feb 14, 2016 1,214 users	100.00%	1.45%	0.62%	0.25%	0.16%	0.41%	0.28%	0.16%	0.08%	0.05%	
Feb 15, 2016 542 users	100.00%	2.86%	0.37%	0.56%	0.19%	0.35%	0.19%	0.19%	0.02%		
Feb 16, 2016 623 users	100.00%	3.77%	0.94%	0.54%	0.71%	0.71%	0.34%	0.03%			
Feb 17, 2016 572 users	100.00%	8.17%	2.47%	3.09%	1.25%	1.85%	0.59%				
Feb 18, 2016 547 users	100.00%	15.48%	3.77%	3.99%	2.33%	0.56%					
Feb 19, 2016 610 users	100.00%	9.59%	4.85%	3.63%	0.00%						
Feb 20, 2016 642 users	100.00%	8.75%	2.49%	0.00%							
Feb 21, 2016 715 users	100.00%	8.54%	0.00%								
Feb 22, 2016 633 users	100.00%	0.00%									

In this table for 9<sup>th</sup> Feb 1344 users are in the selected cohort. For this row 9<sup>th</sup> Feb is day 0. On day 1 that is 10<sup>th</sup> Feb 2.53% of users from 1344 that is 34 users again completed the goals or events selected. On day2 that is 11<sup>th</sup> 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 10<sup>th</sup> Feb. The all users is the weightage average of the cohort 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	1,688
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 9<sup>th</sup> 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							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

Number 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	46,762,489	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						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					2,432	4,484	3,327
May-16						2,160	2,578
Jun-16							2,047

As we can see from 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 learned 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.

	TransactionID	CustomerID	Date of Trans	Quantity	Amount
1	TRX00001	CT10003	31-Apr-15	3	18000
2	TRX00007	CT10003	23-May-15	6	42000
3	TRX00016	CT10003	22-Jul-15	1	4500
4	TRX00029	CT10001	8-Sep-15	2	8000
5	TRX00030	CT10002	17-Apr-15	1	1000
6	TRX00022	CT10002	15-Aug-15	6	20000
7	TRX00003	CT10003	20-Apr-15	1	2000
8	TRX00006	CT10003	4-May-15	5	30000
9	TRX00009	CT10003	30-May-15	2	10000
10	TRX00004	CT10004	25-Apr-15	6	12000
11	TRX00005	CT10005	4-May-15	2	24000
12	TRX00012	CT10005	16-Jun-15	4	24000
13	TRX00019	CT10005	31-Jul-15	3	45000
14	TRX00008	CT10006	23-May-15	1	5000
15	TRX00010	CT10006	10-Jun-15	7	49000
16	TRX00023	CT10006	19-Aug-15	1	450
17	TRX00011	CT10007	15-Jun-15	8	21000
18	TRX00015	CT10007	6-Jul-15	2	8000
19	TRX00013	CT10008	20-Jun-15	20	20000

	TransactionID	CustomerID	Date of Transaction	Quantity	Amount
1	TRX00025	CT10008	31-Aug-15	1	38000
2	TRX0018	CT10009	30-Jun-15	3	3500
3	TRX0027	CT10009	5-Sep-15	6	36000
4	TRX0017	CT10010	25-Jul-15	3	15000
5	TRX0021	CT10010	15-Aug-15	1	6400
6	TRX0018	CT10011	28-Jul-15	12	240000
7	TRX0009	CT10012	1-Aug-15	1	250
8	TRX0024	CT10013	28-Aug-15	5	100000
9	TRX0028	CT10013	6-Sep-15	4	16000
10	TRX0026	CT10014	3-Sep-15	1	40000
11	TRX0030	CT10015	12-Sep-15	1	24000

Sort on CustomerID and then by date of Transaction as shown below.

TransactionID	CustomerID	Date of Transaction	Quantity	Amount
TRX0001	CT10001	15-Apr-15	3	8000
TRX0002	CT10002	15-Apr-15	3	8000
TRX0003	CT10003	21-May-15	6	42000
TRX0004	CT10004	22-Jul-15	3	4500
TRX0005	CT10005	8-Sep-15	2	8000
TRX0006	CT10006	17-Aug-15	3	10000
TRX0007	CT10007	15-Aug-15	4	20000
TRX0008	CT10008	29-Aug-15	12	24000
TRX0009	CT10009	31-Aug-15	3	15000
TRX0010	CT10010	15-Aug-15	1	4000
TRX0011	CT10011	20-Jun-15	4	24000
TRX0012	CT10012	20-Jun-15	20	200000
TRX0013	CT10013	21-May-15	3	12000
TRX0014	CT10014	21-May-15	3	12000
TRX0015	CT10015	22-Jun-15	3	12000
TRX0016	CT10016	22-Jun-15	3	12000
TRX0017	CT10017	22-Jun-15	3	12000
TRX0018	CT10018	22-Jun-15	3	12000
TRX0019	CT10019	22-Jun-15	3	12000
TRX0020	CT10020	22-Jun-15	3	12000
TRX0021	CT10021	22-Jun-15	3	12000
TRX0022	CT10022	22-Jun-15	3	12000
TRX0023	CT10023	22-Jun-15	3	12000
TRX0024	CT10024	22-Jun-15	3	12000
TRX0025	CT10025	22-Jun-15	3	12000
TRX0026	CT10026	22-Jun-15	3	12000
TRX0027	CT10027	22-Jun-15	3	12000
TRX0028	CT10028	22-Jun-15	3	12000
TRX0029	CT10029	22-Jun-15	3	12000
TRX0030	CT10030	22-Jun-15	3	12000

Let's create number of purchase of customer. First purchase as 1, second purchase as 2 and so on.

TransactionID	CustomerID	Date of Transaction	Quantity	Amount	No to Purch.
TRX0001	CT10001	15-Apr-15	3	8000	1
TRX0007	CT10007	21-May-15	6	42000	=IF(E2>D2,F2+1,1)
TRX0016	CT10001	22-Jun-15	3	12000	2
TRX0029	CT10005	8-Sep-15	2	8000	3
TRX0002	CT10002	17-Aug-15	3	10000	1
TRX0022	CT10002	15-Aug-15	4	20000	2
TRX0003	CT10003	20-Aug-15	3	20000	1
TRX0006	CT10003	4-May-15	5	30000	2
TRX0009	CT10003	30-May-15	2	10000	3
TRX0004	CT10004	25-Aug-15	6	12000	1
TRX0005	CT10005	8-May-15	2	2400	1
TRX0012	CT10005	16-Jun-15	8	24000	2
TRX0019	CT10009	31-Jul-15	3	4500	3
TRX0008	CT10006	23-May-15	3	5000	1
TRX0010	CT10006	10-Jun-15	7	49000	2
TRX0023	CT10006	19-Aug-15	3	450	3
TRX0011	CT10007	15-Jun-15	3	2100	1
TRX0015	CT10007	6-Jul-15	2	800	2
TRX0018	CT10008	20-Jun-15	20	200000	1

All the order has been assign N<sup>th</sup> purchase of the customer.

TransactionID	CustomerID	Date of Transaction	Quantity	Amount	No to Purch.
TRX0001	CT10001	15-Apr-15	3	8000	1
TRX0007	CT10007	21-May-15	6	42000	2
TRX0016	CT10001	22-Jun-15	3	12000	3
TRX0029	CT10005	8-Sep-15	2	8000	4
TRX0002	CT10002	17-Aug-15	3	10000	1
TRX0022	CT10002	15-Aug-15	4	20000	2
TRX0003	CT10003	20-Aug-15	3	20000	1
TRX0006	CT10003	4-May-15	5	30000	2
TRX0009	CT10003	30-May-15	2	10000	3
TRX0004	CT10004	25-Aug-15	6	12000	1
TRX0005	CT10005	8-May-15	2	2400	1
TRX0012	CT10005	16-Jun-15	8	24000	2
TRX0019	CT10009	31-Jul-15	3	4500	3
TRX0008	CT10006	23-May-15	3	5000	1
TRX0010	CT10006	10-Jun-15	7	49000	2
TRX0023	CT10006	19-Aug-15	3	450	3
TRX0011	CT10007	15-Jun-15	3	2100	1
TRX0015	CT10007	6-Jul-15	2	800	2
TRX0018	CT10008	20-Jun-15	20	200000	1

Let us take first purchase as the date of Joining for the customer to the system.

	TransactionID	CustomerID	Date of Transaction	Quantity	Amount	No to Purch	Joining Date	Month of Purchase
1	TR00001	CT10001	15-Apr-15	3	1800	3	15-Apr-15	
2	TR00007	CT10001	21-May-15	6	4200	2	15-Apr-15	
3	TR00016	CT10001	22-Jul-15	3	4500	3	15-Apr-15	
4	TR00029	CT10001	8-Sep-15	2	8000	4	15-Apr-15	
5	TR00002	CT10002	17-Apr-15	3	1000	3	17-Apr-15	
6	TR00022	CT10002	15-Aug-15	4	2000	2	17-Apr-15	
7	TR00003	CT10003	20-Aug-15	3	2000	3	20-Aug-15	
8	TR00006	CT10003	4-May-15	5	3000	2	20-Aug-15	
9	TR00009	CT10003	30-May-15	2	8000	3	20-Aug-15	
10	TR00004	CT10004	25-Aug-15	4	1200	3	25-Aug-15	
11	TR00005	CT10005	4-May-15	2	2400	3	25-Aug-15	
12	TR00012	CT10005	16-Jun-15	4	2400	2	25-Aug-15	
13	TR00019	CT10005	31-Jul-15	3	4500	3	25-Aug-15	
14	TR00008	CT10006	23-May-15	3	5000	3	23-May-15	
15	TR00010	CT10006	10-Jun-15	7	49000	2	23-May-15	
16	TR00023	CT10006	19-Aug-15	3	450	3	23-May-15	
17	TR00011	CT10007	15-Jun-15	3	2100	3	15-Jun-15	
18	TR00015	CT10007	6-Jul-15	2	800	2	15-Jun-15	
19	TR00018	CT10008	20-Jun-15	20	20000	3	20-Jun-15	

Basically for one customer all subsequent purchase will share same data of Joining as the first purchase.

	TransactionID	CustomerID	Date of Transaction	Quantity	Amount	No to Purch	Joining Date	Month of Purchase
1	TR00001	CT10001	15-Apr-15	3	1800	3	15-Apr-15	Apr
2	TR00007	CT10001	21-May-15	6	4200	2	15-Apr-15	May
3	TR00016	CT10001	22-Jul-15	3	4500	3	15-Apr-15	Jul
4	TR00029	CT10001	8-Sep-15	2	8000	4	15-Apr-15	Sep
5	TR00002	CT10002	17-Apr-15	3	1000	3	17-Apr-15	
6	TR00022	CT10002	15-Aug-15	4	2000	2	17-Apr-15	
7	TR00003	CT10003	20-Aug-15	3	2000	3	20-Aug-15	
8	TR00006	CT10003	4-May-15	5	3000	2	20-Aug-15	
9	TR00009	CT10003	30-May-15	2	8000	3	20-Aug-15	
10	TR00004	CT10004	25-Aug-15	4	1200	3	25-Aug-15	
11	TR00005	CT10005	4-May-15	2	2400	3	25-Aug-15	
12	TR00012	CT10005	16-Jun-15	4	2400	2	25-Aug-15	
13	TR00019	CT10005	31-Jul-15	3	4500	3	25-Aug-15	
14	TR00008	CT10006	23-May-15	3	5000	3	23-May-15	
15	TR00010	CT10006	10-Jun-15	7	49000	2	23-May-15	
16	TR00023	CT10006	19-Aug-15	3	450	3	23-May-15	
17	TR00011	CT10007	15-Jun-15	3	2100	3	15-Jun-15	
18	TR00015	CT10007	6-Jul-15	2	800	2	15-Jun-15	
19	TR00018	CT10008	20-Jun-15	20	20000	3	20-Jun-15	

Extract month name from the date of joining to get the month of joining.

	TransactionID	CustomerID	Date of Transaction	Quantity	Amount	No to Purch	Joining Date	Month of Purchase
1	TR00001	CT10001	15-Apr-15	3	1800	3	15-Apr-15	Apr
2	TR00007	CT10001	21-May-15	6	4200	2	15-Apr-15	=TEXT(J2,"mmm")
3	TR00016	CT10001	22-Jul-15	3	4500	3	15-Apr-15	
4	TR00029	CT10001	8-Sep-15	2	8000	4	15-Apr-15	
5	TR00002	CT10002	17-Apr-15	3	1000	3	17-Apr-15	
6	TR00022	CT10002	15-Aug-15	4	2000	2	17-Apr-15	
7	TR00003	CT10003	20-Aug-15	3	2000	3	20-Aug-15	
8	TR00006	CT10003	4-May-15	5	3000	2	20-Aug-15	
9	TR00009	CT10003	30-May-15	2	8000	3	20-Aug-15	
10	TR00004	CT10004	25-Aug-15	4	1200	3	25-Aug-15	
11	TR00005	CT10005	4-May-15	2	2400	3	25-Aug-15	
12	TR00012	CT10005	16-Jun-15	4	2400	2	25-Aug-15	
13	TR00019	CT10005	31-Jul-15	3	4500	3	25-Aug-15	
14	TR00008	CT10006	23-May-15	3	5000	3	23-May-15	
15	TR00010	CT10006	10-Jun-15	7	49000	2	23-May-15	
16	TR00023	CT10006	19-Aug-15	3	450	3	23-May-15	
17	TR00011	CT10007	15-Jun-15	3	2100	3	15-Jun-15	
18	TR00015	CT10007	6-Jul-15	2	800	2	15-Jun-15	
19	TR00018	CT10008	20-Jun-15	20	20000	3	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.

	TransactionID	CustomerID	Date of Transact.	Quantity	Amount	Is to Purchase	Joining Date	JoinMonth	PurchaseMonth
1	TRX00001	CT10001	15-Apr-15	3	1800	3	15-Apr-15	Apr	Apr
2	TRX00007	CT10001	25-May-15	6	4200	2	15-Apr-15	Apr	May
3	TRX00016	CT10001	22-July-15	3	4500	3	15-Apr-15	Apr	Jul
4	TRX00029	CT10001	8-Sep-15	2	8000	4	15-Apr-15	Apr	Sep
5	TRX00002	CT10002	17-Apr-15	3	1000	3	17-Apr-15	Apr	Apr
6	TRX00022	CT10002	15-Aug-15	4	2000	2	17-Apr-15	Apr	Aug
7	TRX00003	CT10003	20-Aug-15	3	2000	3	20-Aug-15	Aug	Aug
8	TRX00006	CT10003	4-May-15	5	3000	2	20-Aug-15	Aug	May
9	TRX00009	CT10003	30-May-15	2	10000	3	20-Aug-15	Aug	May
10	TRX00004	CT10004	25-Aug-15	4	1200	3	25-Aug-15	Aug	Aug
11	TRX00005	CT10005	8-May-15	2	2400	3	4-May-15	May	May
12	TRX00012	CT10005	16-Jun-15	8	2400	2	4-May-15	May	Jun
13	TRX00019	CT10005	31-Jul-15	3	4500	3	4-May-15	May	Jul
14	TRX00008	CT10006	23-May-15	3	5000	3	23-May-15	May	May
15	TRX00023	CT10006	10-Jun-15	7	49000	2	23-May-15	May	Jun
16	TRX00023	CT10006	19-Aug-15	3	450	3	23-May-15	May	Aug
17	TRX00011	CT10007	15-Jun-15	3	2100	3	15-Jun-15	Jun	Jun
18	TRX00015	CT10007	6-Jul-15	2	800	2	15-Jun-15	Jun	Jul
19	TRX00018	CT10008	20-Jun-15	20	20000	3	20-Jun-15	Jun	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.

	Count of CustomerID	Column Labels	Apr	May	Jun	Jul	Aug	Sep
	Row Labels							
5	Apr		4	3		1	1	1
6	May			2	2	1	1	
7	Jun				3	1	1	1
8	Jul					2	1	
9	Aug						2	1
10	Sep							2
11	Grand Total		4	5	5	5	6	5

Month	Apr	May	Jun	Jul	Aug	Sep	
15	Apr	100%	75%	0%	25%	25%	25%
16	May		100%	100%	50%	50%	0%
17	Jun			100%	33%	33%	33%
18	Jul				100%	50%	0%
19	Aug					100%	50%
20	Sep						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 month-wise 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 quite useful for understanding the customer loyalty and effectiveness of marketing campaign.

	Count of CustomerID	Column Labels	Apr	May	Jun	Jul	Aug	Sep
	Row Labels							
4	Apr		4	3	5	1	1	1
5	May			2	2	1	1	
6	Jun			3	1	1		
7	Jul				2	1		
8	Aug					2	1	
9	Sep						2	
10	Grand Total		4	5	5	5	6	5

Month	Apr	May	Jun	Jul	Aug	Sep
11	100%	75%	0%	25%	25%	25%
12	May	100%	100%	50%	50%	0%
13	Jun		100%	33%	33%	33%
14	Jul			100%	50%	0%
15	Aug				100%	50%
16	Sep					100%

	Sum of Amount	Column Labels	Apr	May	Jun	Jul	Aug	Sep
	Row Labels							
17	Apr		6000	17200	4500	2000	8000	
18	May			2400	51400	4500	450	
19	Jun				22450	800	3800	3600
20	Jul					25500	680	
21	Aug						30250	1600
22	Sep							6400
23	Grand Total		6000	24600	73850	35500	17380	19400

New Customer	6000	7400	22450	25500	30250	6400
Old Customer	0	17200	51400	9800	6930	13200
New Customer AOV	1,600	3,700	7,483	12,750	15,125	3,200
Old Customer AOV	-	5,713	25,700	1,267	1,713	4,400

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.

The screenshot shows a Microsoft Excel spreadsheet titled "Cohort Work - Sheet1". The PivotTable is structured as follows:

Row Labels	Column Labels				
	Apr	May	Jun	Jul	Aug
Apr	4	3	1	1	1
May		2	2	1	1
Jun			3	1	1
Jul				2	1
Aug					2
Sep					2
<b>Grand Total</b>	<b>4</b>	<b>5</b>	<b>5</b>	<b>5</b>	<b>5</b>

Below the PivotTable, there is a summary section with the following data:

	Sum of Amount	Column Labels			
Row Labels		Apr May Jun Jul Aug Sep			
Apr	6000	17200 4500 2000 8000			
May	7400	51400 4500 450 0			
Jun		22450 800 3800 3600			
Jul			25500 680 0		
Aug				10250 1600 0	
Sep					6400 0
<b>Grand Total</b>	<b>6000</b>	<b>24600</b> <b>73850</b> <b>35300</b> <b>17180</b> <b>19600</b> <b>0</b>			

Summary statistics:

- New Customer: 6000, 7400, 22450, 25500, 10250, 6400
- OldCustomer: 0, 17200, 51400, 4500, 450, 0
- New Customer AOV: \$1,500, 3,200, 7,483, 17,750, 5,125, 3,200
- Old Customer AOV: 0, 5,733, 25,700, 3,267, 1,733, 4,400

Additional metrics:

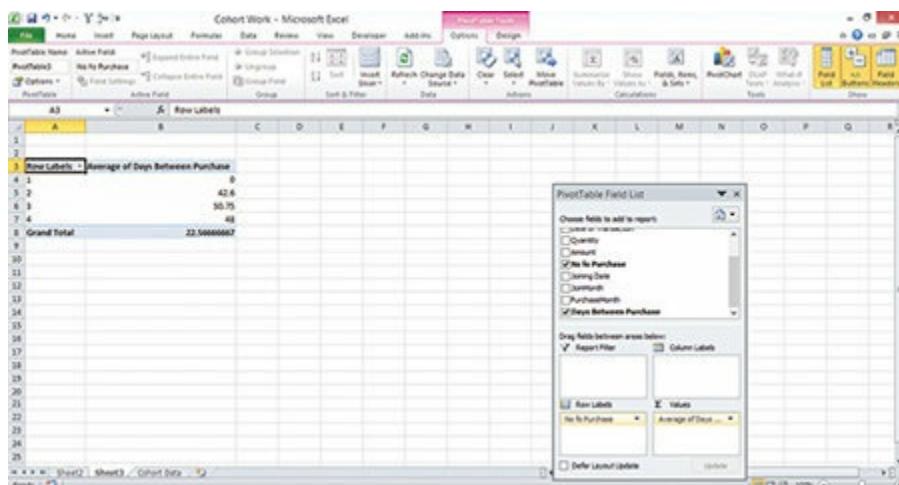
- No of Customer six month old: 4
- No of transaction done: 10
- Amount of Transaction: 37700
- Average Transaction per customer: 2.5
- Average Transaction Amount per customer: 9437.5

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.

The screenshot shows a Microsoft Excel spreadsheet titled "Cohort Work - Sheet1". The data table has an additional column "Days Between Purchase" calculated using the formula  $=IF(I2>1,0,INT(E2-C2))$ .

TransactionID	CustomerID	Date of Transaction	Quantity	Amount	No of Purchases	Joining Date	JoinMonth	PurchaseMonth	Days Between Purchase
TRX0001	CT10001	15-Apr-15	3	1800	3	15-Apr-15	Apr	Apr	0
TRX0002	CT10002	21-May-15	6	4200	2	15-Apr-15	Apr	May	=IF(D2>1,0,INT(E2-C2))
TRX0003	CT10003	22-Jul-15	5	4500	3	15-Apr-15	Apr	Jul	
TRX0004	CT10004	8-Sep-15	2	8000	4	15-Apr-15	Apr	Sep	
TRX0005	CT10002	17-Aug-15	9	1000	3	17-Aug-15	Aug	Aug	
TRX0006	CT10002	15-Aug-15	4	2000	2	17-Aug-15	Aug	Aug	
TRX0007	CT10003	20-Aug-15	3	2000	3	20-Aug-15	Aug	Aug	
TRX0008	CT10004	4-May-15	5	3000	2	20-Apr-15	Apr	May	
TRX0009	CT10003	30-May-15	2	10000	3	20-Aug-15	Aug	Aug	
TRX0010	CT10004	25-Apr-15	8	1200	3	25-Apr-15	Apr	Apr	
TRX0011	CT10005	4-May-15	2	2400	3	4-May-15	May	May	
TRX0012	CT10005	16-Jun-15	4	2400	2	4-May-15	May	Jun	
TRX0013	CT10005	31-Jul-15	3	4500	3	4-May-15	May	Jul	
TRX0014	CT10008	23-May-15	5	5000	5	23-May-15	May	May	
TRX0015	CT10006	10-Jun-15	7	49000	2	23-May-15	May	Jun	
TRX0016	CT10006	19-Aug-15	5	450	3	13-May-15	May	Aug	
TRX0017	CT10007	15-Jun-15	3	2100	3	15-Jun-15	Jun	Jun	
TRX0018	CT10007	6-Jul-15	2	800	2	15-Jun-15	Jun	Jul	
TRX0019	CT10008	20-Jun-15	20	20000	3	20-Jun-15	Jun	Jun	
TRX0020	CT10008	31-Aug-15	3	3800	2	20-Jun-15	Jun	Aug	
TRX0021	CT10009	30-Jun-15	3	350	3	30-Jun-15	Jun	Jun	

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.



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.

A screenshot of Microsoft Excel titled "Cohort Work - Microsoft Excel". The PivotTable has "Row Labels" set to "Average of Days Between Purchase" and "Values" set to "Sum of Days Between Purchase". The data shows three rows of values: 42.6, 30.75, and 48, with a Grand Total of 22.56666667 and a value of 30.7 in the bottom right corner.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
3	Row Labels	- Average of Days Between Purchase	Sum of Days Between Purchase												
4	1		0												
5	2		42.6												
6	3		30.75												
7	4		48												
8	Grand Total		22.56666667												
9			30.7												
10															
11															
12															
13															
14															
15															
16															
17															
18															
19															
20															
21															
22															
23															
24															
25															
26															
27															

If you add amount in the pivot it shows the average purchase value of those who are purchasing first time, second time and so on.

A screenshot of Microsoft Excel titled "Cohort Work - Microsoft Excel". The PivotTable has "Row Labels" set to "Average of Days Between Purchase" and "Values" set to "Average of Amount". The data shows three rows of values: 5200, 7959, and 4882.5, with a Grand Total of 22.56666667 and a value of \$884,333333 in the bottom right corner. A PivotTable Field List window is open, showing fields like "Amount", "No Purchase", "Joining Date", "Purchase Month", and "Days Between Purchase".

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
3	Row Labels	- Average of Days Between Purchase	Average of Amount												
4	1		0	5200											
5	2		42.6	7959											
6	3		30.75	4882.5											
7	4		48	8000											
8	Grand Total		22.56666667	\$884,333333											
9			30.7												
10															
11															
12															
13															
14															
15															
16															
17															
18															
19															
20															
21															
22															
23															
24															
25															
26															
27															

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 life 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  $R_1, R_2, \dots, R_n$  revenue and cost of acquisition and retention is  $Q_1, Q_2, \dots, Q_n$ . Assuming the company has  $x\%$  margin and weighted average cost of capital is  $r$ , we can calculate customer life time value as

$$CLV = (R_1 * x\% - Q_1) / (1+r) + (R_2 * x\% - Q_2) / (1+r)^2 + \dots + (R_n * x\% - Q_n) / (1+r)^n$$

$$\text{Or } CLV = P_1 / (1+r) + P_2 / (1+r)^2 + \dots + P_n / (1+r)^n$$

$$CLV = \sum_{i=1}^N P_i / (1+r)^i$$

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 Cohort			

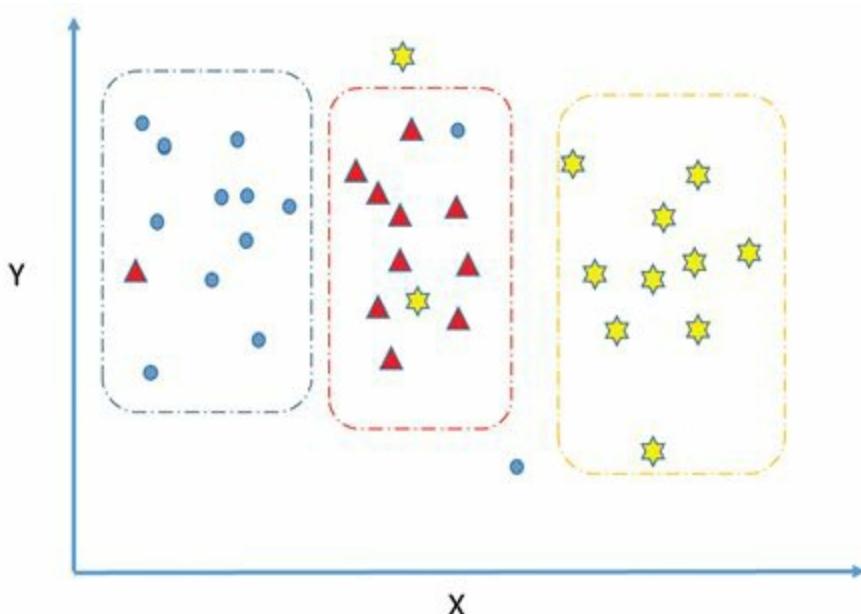
Subsequent Month	Average margin per customer	Average Repeat rate	Final Value
Month01	2,300	100.0%	2,300.00
Month02	4,000	15.0%	600.00
Month03	3,500	10.0%	350.00
Month04	3,400	8.0%	272.00
Month05	2,400	6.0%	144.00
Month06	2,000	6.0%	120.00
Month07	1,900	5.0%	95.00
Month08	2,100	5.0%	105.00
Month09	1,850	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
		<b>Total</b>	<b>4,182.00</b>

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.



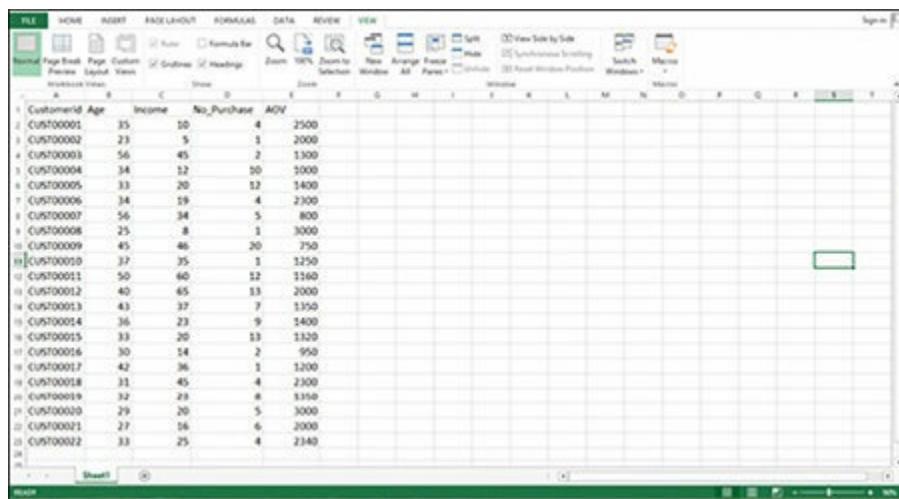
However in most of the situation we will have multiple dimension on which a population has to be segmented. In many of the situation with good knowledge of business requirement if-else type of condition can be used to filter out the segment using different dimensions. The complex segmentation need more 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.



	CustomerID	Age	Income	No_Purchase	AGV
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	1360
12	CUST00012	40	45	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	1150
20	CUST00020	29	20	5	3000
21	CUST00021	27	16	6	2000
22	CUST00022	33	25	4	2340

Upload the data into R

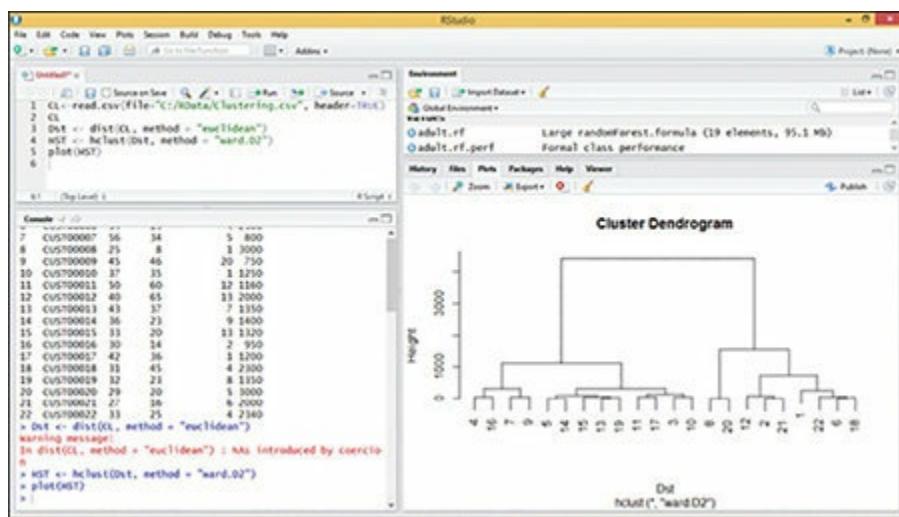
```

File Edit Code View Plots Session Build Debug Tools Help
Go to file/function

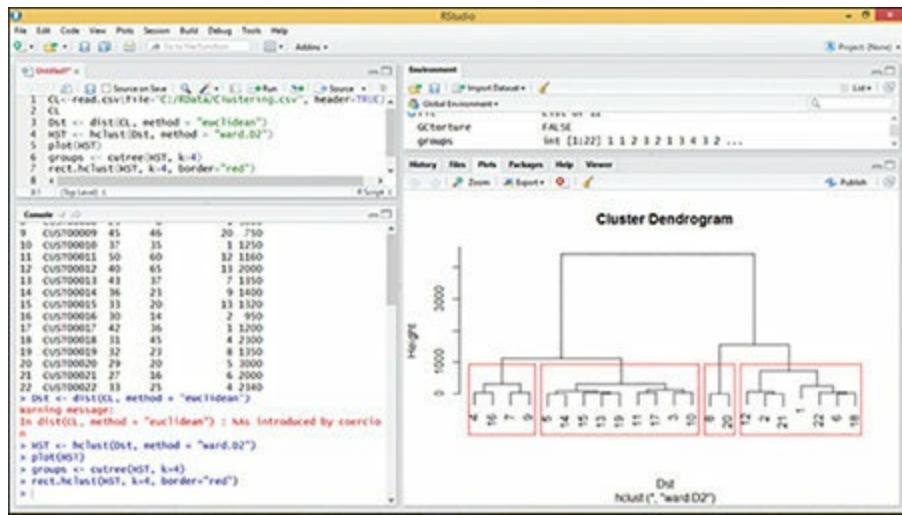
Console - / 
> CL <- read.csv(file = "C:/RData/Clustering.csv", header = TRUE)
> CL
   CustomerId Age Income No_Purchase AOV
1  CUST00001  35    10        4 2500
2  CUST00002  23     5        1 2000
3  CUST00003  56    45        2 1300
4  CUST00004  34    12        10 1000
5  CUST00005  33    20        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.



## 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 =  $x_1, x_2, x_3 \dots x_n$

Cluster sets =  $C_1, C_2, \dots, C_k$  and Mean of each cluster as  $\mu_1, \mu_2, \dots, \mu_k$

$$\text{minimize} \sum_{j=1}^k \sum_{i=1}^n (x_{ij} - \mu_j)^2$$

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.

```

File Edit Code View Plots Session Build Debug Tools Help
Go to File Function Addins

Untitled1.R

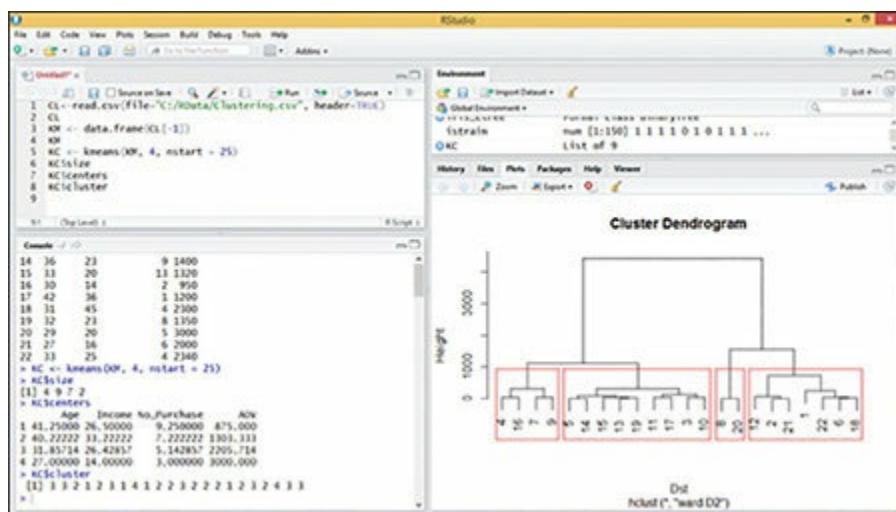
1 CL<-read.csv(file="C:/RData/Clustering.csv", header=TRUE)
2 CL
3 KM <- data.frame(CL[-1])
4 KM
5 KC <- kmeans(KM, 4, nstart = 25)
6 KC$size
7 KC$centers
8

R Script

Console -/ />
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      5 3000
21 27 16      6 2000
22 33 25      4 2340
> KC <- kmeans(KM, 4, nstart = 25)
> KC$size
[1] 4 9 7 2
> KC$centers
   Age Income No._Purchase      ADW
1 41.25000 26.50000 9.250000 875.000
2 40.22222 33.22222 7.222222 1303.333
3 31.85714 26.42857 5.142857 2205.714
4 27.00000 14.00000 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 Cluster4 = {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  $I = \{i_1, i_2, \dots, i_n\}$

Transactions are instances of groups of items bought together. For each transaction, then, we have an item set  $T_n = \{i_1, i_2, \dots, i_k\}$

Rules are statements of the form  $\{i_1, i_2, \dots\} \Rightarrow \{i_k\}$

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.

The image contains two screenshots of Microsoft Excel. Both screenshots show a table with three columns: OrderID, Product, and Quantity. The data in both tables is identical, representing 15 different orders for products A through F. The top table has rows numbered 1 through 15. The bottom table has rows numbered 1 through 15. The data is as follows:

OrderID	Product	Quantity
00001	A	1
00001	B	2
00001	C	3
00001	D	2
00002	A	1
00002	B	4
00002	D	5
00003	D	2
00004	B	1
00004	C	6
00004	D	2
00005	A	5
00005	B	6
00005	C	9
00005	D	2
00005	E	2
00005	F	4
00006	A	5
00006	B	3
00006	D	7
00006	F	3

Upload the data in R

The screenshot shows the RStudio interface. The top panel displays the workspace with the variable 'MB' defined as a data frame. The bottom panel shows the console output of the command 'head(MB)' which prints the first 17 rows of the data frame.

```
1 MB <- read.csv(file = "C:/RData/MBA.csv", header=TRUE)
2 MB
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
```

	OrderID	Product	Quantity
1	00001	A	1
2	00001	B	2
3	00001	C	3
4	00001	D	2
5	00002	A	1
6	00002	B	4
7	00002	D	5
8	00003	D	2
9	00004	B	1
10	00004	C	6
11	00004	D	2
12	00005	A	5
13	00005	B	6
14	00005	C	9
15	00005	D	2
16	00005	E	2
17	00005	C	8

The input data is in the linear format. This has to be converted to comma separated format for each item in the orders.

The screenshot shows the RStudio interface. The top panel displays the workspace with variables 'MB', 'MB1', and 'MB2'. The bottom panel shows the console output of the commands used to convert the data from a single row per order to multiple rows per order.

```
1 MB <- read.csv(file = "C:/RData/MBA.csv", header=TRUE)
2 MB
3 write.csv(MB, file="C:/RData/Temp.csv", row.names = FALSE)
4 MB1 <- read.transactions(file ="C:/RData/Temp.csv", format="single", sep=",", cols=c("OrderID","Product"))
5
```

```
30 00009    D   5
31 00009    E   4
32 00009    F   6
33 00010    A   3
34 00010    E   2
35 00010    F   3
36 00011    B   1
37 00011    C   3
38 00011    D   5
39 00011    E   4
40 00011    F   4
41 00012    C   5
42 00012    D   3
43 00012    F   1
44 00013    E   2
45 00013    F   4
```

```
> write.csv(MB1, file="C:/RData/Temp.csv", row.names = FALSE)
> MB1 <- read.transactions(file ="C:/RData/Temp.csv", format="single", sep=",", cols=c("OrderID","Product"))
>
```

You can see the output for each order. For example Order OD001 has 4 items A, B, C, D.

The screenshot shows the RStudio interface with two panes. The top pane is a code editor with the following R script:

```
1 MB <- read.csv(file = "C:/RData/MBA.csv", header=TRUE)
2 MB
3 write.csv(MB, file="C:/RData/Temp.csv", row.names = FALSE)
4 MB1 <- read.transactions(file ="C:/RData/Temp.csv", format="single", sep=",", cols=c("OrderID","Product"))
5 inspect(MB1)
6
```

The bottom pane is the R Console, showing the output of the inspect command:

```
6.1 (Top Level) >
Console <- >
45 00013   F    4
> write.csv(MB, file="C:/RData/Temp.csv", row.names = FALSE)
> MB1 <- read.transactions(file ="C:/RData/Temp.csv", format="single", sep=",", cols=c("OrderID","Product"))
> inspect(MB1)
  items      transactionID
[1] {A,B,C,D} 00001
[2] {A,B,D}   00002
[3] {D}        00003
[4] {B,C,D}   00004
[5] {A,B,C,D,E,F} 00005
[6] {A,B,D,F} 00006
[7] {D,E,F}   00007
[8] {A,D,E,F} 00008
[9] {B,D,E,F} 00009
[10] {A,E,F}  00010
[11] {B,C,D,E,F} 00011
[12] {C,D,F}  00012
[13] {E,F}    00013
>
```

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.

The screenshot shows the RStudio interface with two panes. The top pane is a code editor with the following R script:

```
1 MB <- read.csv(file = "C:/RData/MBA.csv", header=TRUE)
2 MB
3 write.csv(MB, file="C:/RData/Temp.csv", row.names = FALSE)
4 MB1 <- read.transactions(file ="C:/RData/Temp.csv", format="single", sep=",", cols=c("OrderID","Product"))
5 inspect(MB1)
6 MOut <- apriori(MB1, parameter=list(support=0.001, confidence=0.25, minlen=1))
7
```

The bottom pane is the R Console, showing the output of the apriori command and its parameters:

```
7.1 (Top Level) >
Console <- >
> MOut <- apriori(MB1, parameter=list(support=0.001, confidence=0.25, minlen=1))
Apriori
Parameter specification:
confidence minval maxval arm originalSupport maxtime support minlen maxlen target ext
0.25     0.1    1 none FALSE           TRUE     5  0.001     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
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[6 item(s), 13 transaction(s)] done [0.00s].
sorting and recoding items ... [6 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.00s].
writing ... [192 rule(s)] done [0.00s].
creating 54 object ... done [0.00s].
>
```

System has generated 192 rules.

The screenshot shows the RStudio interface with the code editor containing the following R script:

```
1 MB <- read.csv(file = "C:/RData/MBA.csv", header=TRUE)
2 MB
3 write.csv(MB, file="C:/RData/Tmp.csv", row.names = FALSE)
4 MBI <- read.transactions(file = "C:/RData/Tmp.csv", format="single", sep=",", cols=c("OrderID","Product"))
5 inspect(MBI)
6 MOut <- apriori(MBI, parameter=list(support=0.001, confidence=0.25, minlen=1))
7 MOut
8 |
```

The console output shows the parameter specification and algorithmic control:

```
Parameter specification:
confidence minval maxval arm eval originalSupport maxtime support minlen maxlen target ext
      0.25     0.1    1 none FALSE           TRUE      5   0.001     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
```

Execution details:

```
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[6 item(s), 13 transaction(s)] done [0.00s].
sorting and recoding items ... [6 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.00s].
writing ... [192 rule(s)] done [0.00s].
creating 54 object ... done [0.00s].
> MOut
set of 192 rules
> |
```

You can view the rules in the R screen or export it to Excel.

The screenshot shows the RStudio interface with the code editor containing the same R script as above, followed by the command:

```
> MOut
set of 192 rules
> inspect(MOut)
```

The console output displays the rules in a tabular format:

lhs	rhs	support	confidence	lift
[1]	{}	=> (C)	0.38461538	0.3846154
[2]	{}	=> (A)	0.46153846	0.4615385
[3]	{}	=> (E)	0.53846154	0.5384615
[4]	{}	=> (B)	0.53846154	0.5384615
[5]	{}	=> (F)	0.69230769	0.6923077
[6]	{}	=> (D)	0.84615385	0.8461538
[7]	(C)	=> (A)	0.15184615	0.4000000
[8]	(A)	=> (C)	0.15184615	0.1111111
[9]	(C)	=> (E)	0.15184615	0.4000000
[10]	(E)	=> (C)	0.15184615	0.2857143
[11]	(C)	=> (B)	0.30769231	0.8000000
[12]	(B)	=> (C)	0.30769231	0.5714286
[13]	(C)	=> (F)	0.23076923	0.6000000
[14]	(F)	=> (C)	0.23076923	0.3333333
[15]	(C)	=> (D)	0.38461538	1.0000000
[16]	(D)	=> (C)	0.38461538	0.4545455
[17]	(A)	=> (E)	0.23076923	0.5000000
[18]	(E)	=> (A)	0.33076923	0.3333333
[19]	(A)	=> (F)	0.33076923	0.6785714

```

File Edit Code View Plots Session Build Debug Tools Help
File Source Save Run Source
Untitled.R
1 MB <- read.csv(file = "C:/RData/MBA.csv", header=TRUE)
2 MB1 <- read.transactions(file = "C:/RData/Temp.csv", format="single", sep=",", cols=c("OrderID","Product"))
3 inspect(MB1)
4 MOut <- apriori(MB1, parameter=list(support=0.001, confidence=0.25, minlen=1))
5 inspect(MOut)
6 opt = inspect(MOut)
7 write.csv(opt, file="C:/RData/MBAout1.csv")
8
9 (Top Level) R Script
Console -> 
[177] "[A,C,E,F] => [D] 0.15384615 1.0000000 1.1818182
[178] "[B,C,D,F] => [F] 0.15384615 1.0000000 1.4444444
[179] "[C,D,E,F] => [B] 0.15384615 1.0000000 1.8571429
[180] "[B,C,D,F] => [E] 0.15384615 1.0000000 1.8571429
[181] "[B,D,E,F] => [C] 0.15384615 0.6666667 1.7333333
[182] "[A,B,E,F] => [D] 0.07692308 1.0000000 1.1818182
[183] "[A,B,D,F] => [F] 0.07692308 1.0000000 1.4444444
[184] "[A,D,E,F] => [B] 0.07692308 0.5000000 0.9285714
[185] "[A,B,D,F] => [E] 0.07692308 0.5000000 0.9285714
[186] "[B,D,E,F] => [A] 0.07692308 0.1111111 0.7222222
[187] "[A,B,C,E,F] => [D] 0.07692308 1.0000000 1.1818182
[188] "[A,B,C,D,F] => [F] 0.07692308 1.0000000 1.4444444
[189] "[A,C,D,E,F] => [B] 0.07692308 1.0000000 1.8571429
[190] "[A,B,C,D,F] => [E] 0.07692308 1.0000000 1.8571429
[191] "[C,D,E,F] => [A] 0.07692308 0.5000000 1.0833333
[192] "[A,B,D,E,F] => [C] 0.07692308 1.0000000 2.6000000
> write.csv(opt, file="C:/RData/MBAout1.csv")
Error in is.data.frame(x) : object 'opt' not found
> write.csv(opt, file="C:/RData/MBAout1.csv")
> opt = inspect(MOut)

```

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.

	Bns	rhs	support	confidence	lift
1	[1]	[C]	0.38	0.38	1.00
2	[2]	[A]	0.46	0.46	1.00
3	[3]	[E]	0.54	0.54	1.00
4	[4]	[B]	0.54	0.54	1.00
5	[5]	[F]	0.69	0.69	1.00
6	[6]	[D]	0.85	0.85	1.00
7	[7]	[A]	0.15	0.40	0.87
8	[8]	[C]	0.15	0.33	0.87
9	[9]	[F]	0.15	0.40	0.74
10	[10]	[C]	0.15	0.29	0.74
11	[11]	[B]	0.31	0.80	1.49
12	[12]	[B]	0.31	0.57	1.49
13	[13]	[F]	0.23	0.60	0.87
14	[14]	[F]	0.23	0.33	0.87
15	[15]	[D]	0.38	1.00	1.18
16	[16]	[C]	0.38	0.45	1.18
17	[17]	[A]	0.23	0.50	0.93
18	[18]	[A]	0.23	0.43	0.93
19	[19]	[B]	0.31	0.67	1.24
20	[20]	[A]	0.31	0.57	1.24
21	[21]	[F]	0.31	0.67	0.96

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

$$\text{Log odd} = \text{log}(p/(1-p))$$

Assuming that log odd can be represented in linear format,  
 $\text{Log}(p/(1-p)) = \sum b_j x_j$  where  $j=0$  to  $k$ ,  $b$  is coefficient and  $x$  is independent variables

Taking exponential in both sides we get

$$p/(1-p) = \pi \exp(b_j x_j)$$

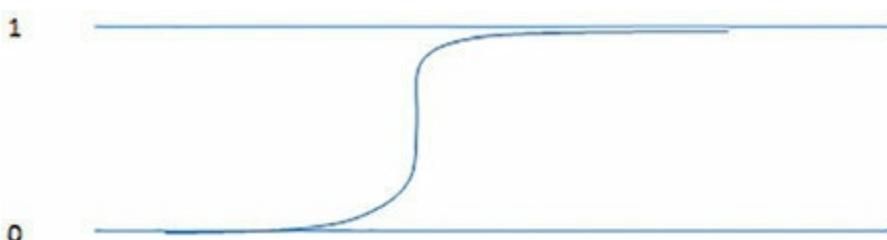
The expression  $\exp(b_j)$  tells us how the odd of the response being true increase (or decrease) as  $x_j$  increase by one unit, all other thing being equal.

Assuming  $j=1$

$$p/(1-p) = \exp(b_0 + b_1 x)$$

$$p = \exp(b_0 + b_1 x) / (1 + \exp(b_0 + b_1 x))$$

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.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	CustomerID	Income(\$k)	Purchase																
1	Cust000001	10	1																
2	Cust000002	5	0																
3	Cust000003	4	0																
4	Cust000004	12	0																
5	Cust000005	20	1																
6	Cust000006	19	1																
7	Cust000007	12	0																
8	Cust000008	8	0																
9	Cust000009	46	1																
10	Cust000010	35	1																
11	Cust000011	40	1																
12	Cust000012	65	1																
13	Cust000013	37	0																
14	Cust000014	23	0																
15	Cust000015	20	1																
16	Cust000016	30	1																
17	Cust000017	30	1																
18	Cust000018	36	1																
19	Cust000019	45	1																
20	Cust000020	23	1																
21	Cust000021	29	1																
22	Cust000022	16	0																
23	Cust000023	25	1																

Add data into the R. Since data set is small I have just input the data using data frame.

The screenshot shows the RStudio interface. In the top-left pane, there is an 'Untitled.R' script with the following code:

```

1 lg<-data.frame(Income=c(10, 5, 4, 12, 20, 19, 12, 8, 46, 35, 60, 65, 37, 23, 20, 70, 36, 45, 23, 20, 16, 25),
2 Purchase=c(1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1))
3 lg
4

```

In the bottom-left pane, the 'Console' output shows the printed data frame:

	Income	Purchase
1	10	1
2	5	0
3	4	0
4	12	0
5	20	1
6	19	
7	12	0
8	8	0
9	46	1
10	35	1
11	60	1
12	65	1
13	37	1
14	23	0
15	20	1
16	70	1
17	36	1

The right-hand sidebar shows the 'Environment' tab with objects like 'lg', 'a111.mdf', 'logit', 'cars', and 'CRK' listed.

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.

The screenshot shows the RStudio interface. In the top-left pane, there is an 'Untitled.R' script with the following code:

```

1 lg<-data.frame(Income=c(10, 5, 4, 12, 20, 19, 12, 8, 46, 35, 60, 65, 37, 23, 20, 70, 36, 45, 23, 20, 16, 25),
2 Purchase=c(1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1))
3 lg
4 lt<-glm(Purchase ~ Income, data=lg, family = "binomial")
5

```

In the bottom-left pane, the 'Console' output shows the results of the logistic regression:

```

5
6 20
7 19
8 12
9 8
10 46
11 35
12 60
13 65
14 37
15 23
16 20
17 70
18 36
19 45
20 23
21 20
22 16
23 25
> lt<-glm(Purchase ~ Income, data=lg, family = "binomial")
> opt = inspect(lt$out)

```

The intercept is -3.9414 and coefficient is 0.2491.

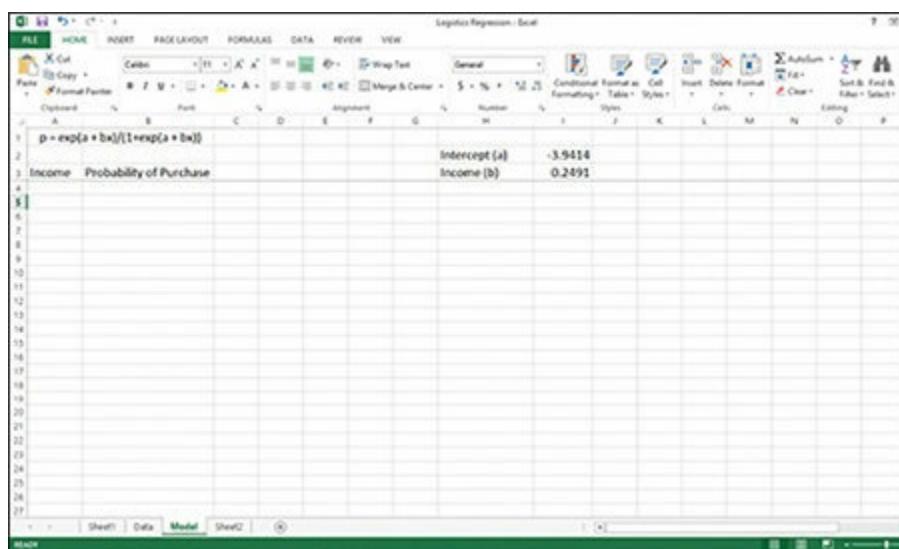
The screenshot shows the RStudio interface. In the top-left pane, there is an 'Untitled.R' script with the following R code:

```
1 Tg<-data.frame(Income=c(10, 5, 4, 12, 20, 19, 12, 8, 46, 35, 60, 65, 37, 23, 20, 70, 36, 45, 23, 20, 16, 25),
2 Purchase=c(1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1))
3 lg
4 lt<-glm(Purchase ~ Income, data=lg, family = "binomial")
5 lt
```

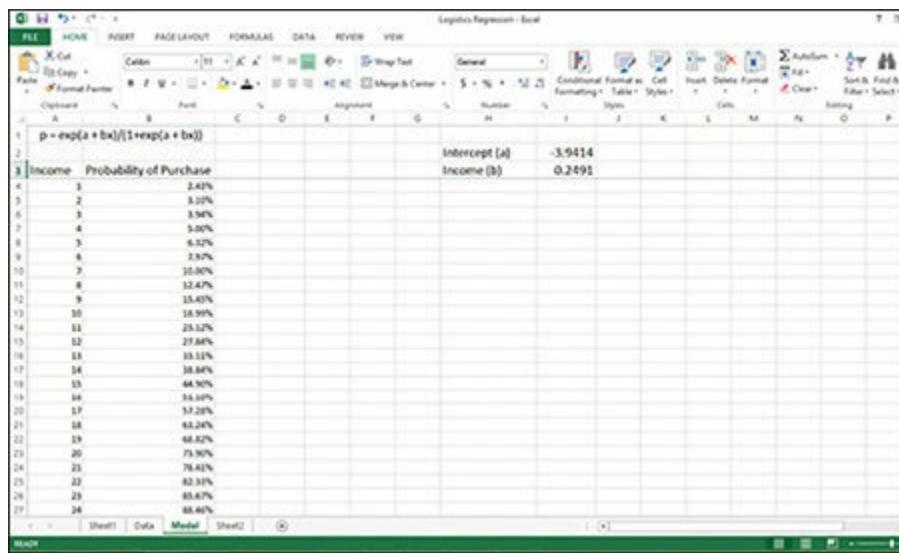
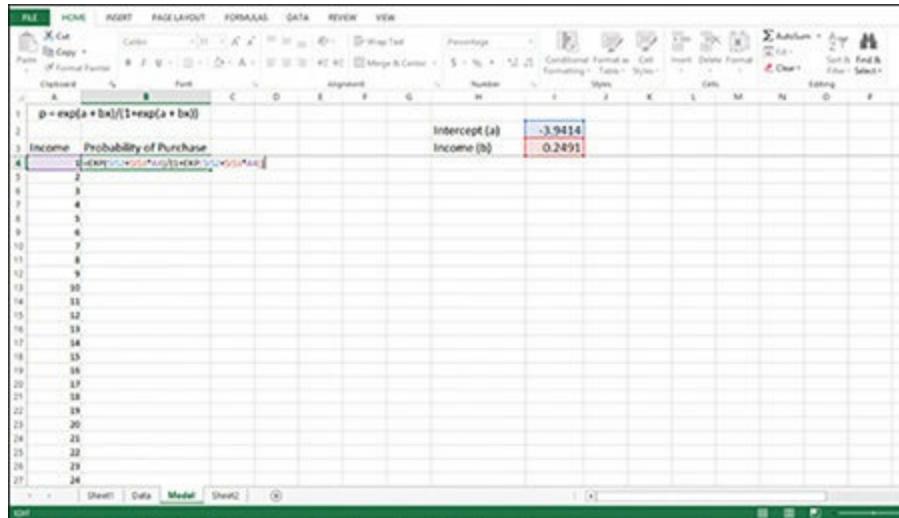
In the bottom-right pane, the 'Console' tab displays the output of the R code:

```
16 70 1
17 36 1
18 45 1
19 23 1
20 20 1
21 16 0
22 25 1
> lt<-glm(Purchase ~ Income, data=lg, family = "binomial")
> lt
Call: glm(formula = Purchase ~ Income, family = "binomial", data = lg)
Coefficients:
(Intercept) Income
-3.9414      0.2491
Degrees of Freedom: 21 Total (i.e. Null); 20 Residual
Null Deviance: 27.52
Residual Deviance: 13.55    AIC: 17.55
>
```

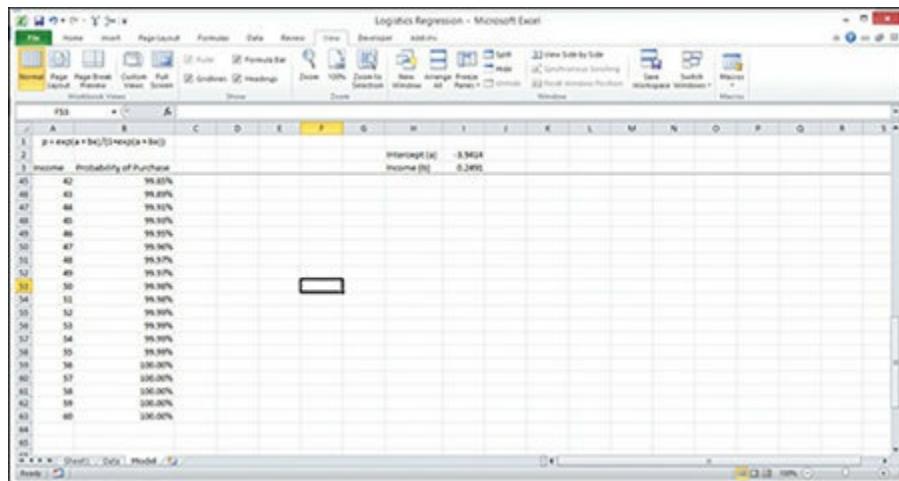
Using the intercept and the coefficient let us calculate the final probability for different income level.



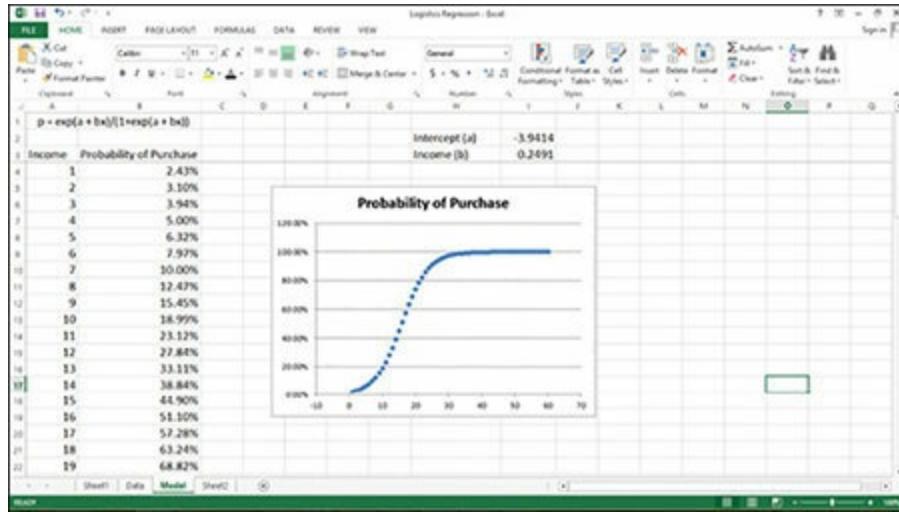
Use  $p = \exp(b_0 + b_1 x) / (1 + \exp(b_0 + b_1 x))$  to calculate the probabilities.



The probability till 60 lakhs income has been calculated using the output from the input data.



On plotting the probability on the income we get an s-shape curve which is nothing but a sigmoid function.



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.

```

File Edit Code View Plots Session Build Debug Tools Help
Untitled.Rnw
File Edit Code View Plots Session Build Debug Tools Help
Untitled.Rnw
1 Tg<-data.frame(Income=c(10, 5, 4, 12, 20, 19, 12, 8, 46, 35, 60, 65, 37, 23, 20, 70, 36, 45, 23, 20, 16, 25),
2 Purchase<-c(1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1)
3 lg
4 lt<-glm(Purchase ~ Income, data=Tg, family = "binomial")
5 lt
6 nd<-data.frame(Income=c(1,4,7,8,10,17,18,23,27,30,33,45,56,60))
7 Prdt <- predict(lt, newdata=nd, type="response")
8

B1 (log level) 1
Console <-->
18   45    1
19   23    1
20   20    1
21   16    0
22   25    1
> lt<-glm(Purchase ~ Income, data=lg, family = "binomial")
> lt
Call: glm(formula = Purchase ~ Income, family = "binomial", data = lg)

Coefficients:
(Intercept)      Income
-3.9414        0.2493

Degrees of Freedom: 21 Total (i.e. Null): 20 Residual
Null Deviance: 27.52
Residual Deviance: 13.55          AIC: 17.55
> nd<-data.frame(Income=c(1,4,7,8,10,17,18,23,27,30,33,45,56,60))
> Prdt <- predict(lt, newdata=nd, type="response")
>

```

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.

```

File Edit Code View Plots Session Build Debug Tools Help
Untitled.Rnw Addins
Untitled.Rnw
1 Tg<-data.frame(Income=c(10, 5, 4, 12, 20, 19, 12, 8, 46, 35, 60, 65, 37, 23, 20, 70, 36, 45, 21, 20, 16, 25),
2 Purchase=c(1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1))
3 lg
4 lt<-glm(Purchase ~ Income, data=lg, family = "binomial")
5 lt
6 nd<-data.frame(Income=c(1,4,7,8,10,17,18,23,27,30,33,45,56,60))
7 Prdt <- predict(lt, newdata=nd, type="response")
8 Prdt
9 

D:\R Scripts\Untitled.Rnw:1:1-1:1: warning: unclosed brace or bracket

```

Call: glm(formula = Purchase ~ Income, data = lg, family = "binomial")

Coefficients:

(Intercept)	Income
-3.9424	0.2491

Degrees of Freedom: 21 Total (i.e. Null); 20 Residual

Null Deviance: 27.52

Residual Deviance: 13.55 AIC: 17.55

> nd<-data.frame(Income=c(1,4,7,8,10,17,18,23,27,30,33,45,56,60))

> Prdt <- predict(lt, newdata=nd, type="response")

> Prdt

1	2	3	4	5	6	7	8	9	10
0.02430764	0.04996843	0.09994363	0.12468789	0.18990951	0.57272836	0.63229517	0.85662014	0.94179597	0.97155946
11	12	13	14						
0.98632396	0.99930253	0.99995491	0.99998336						

```

File Edit Code View Plots Session Build Debug Tools Help
Untitled.Rnw Addins
Untitled.Rnw
2 Purchase=c(1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)
3 lg
4 lt<-glm(Purchase ~ Income, data=lg, family = "binomial")
5 lt
6 nd<-data.frame(Income=c(1,4,7,8,10,17,18,23,27,30,33,45,56,60))
7 Prdt <- predict(lt, newdata=nd, type="response")
8 Prdt
9 Opt<-data.frame(nd$Income, Prdt)
10 Opt
11 

D:\R Scripts\Untitled.Rnw:1:1-1:1: warning: unclosed brace or bracket

```

Call: glm(formula = Purchase ~ Income, data = lg, family = "binomial")

Coefficients:

(Intercept)	Income
11.1	0.02430764

Degrees of Freedom: 21 Total (i.e. Null); 20 Residual

Null Deviance: 27.52

Residual Deviance: 13.55 AIC: 17.55

> Opt<-data.frame(nd\$Income, Prdt)

> Opt

nd\$Income	Prdt
1	0.02430764
2	0.04996843
3	0.09994363
4	0.12468789
5	0.18990951
6	0.57272836
7	0.63229517
8	0.85662014
9	0.94179597
10	0.97155946
11	0.98632396
12	0.99930253
13	0.99995493
14	0.99998336

Logistics Regression - Microsoft Excel

File Home Insert Page Layout Formulas Data Review View Developer Add-ins

Formulas

Clipboard

Sum

A1 B1 C1 D1 E1 F1 G1 H1 I1 J1 K1 L1 M1 N1 O1 P1 Q1 R1 S1

1 p = exp(x + b) / (1 + exp(x + b))

2 Intercept [A] -0.5624

3 Income Probability of Purchase

4 Income [B]

5 10 99.8%

6 11 99.8%

7 12 99.9%

8 13 99.9%

9 14 99.9%

10 15 99.9%

11 16 99.9%

12 17 99.9%

13 18 99.9%

14 19 99.9%

15 20 99.9%

16 21 99.9%

17 22 99.9%

18 23 99.9%

19 24 99.9%

20 25 99.9%

21 26 99.9%

22 27 99.9%

23 28 99.9%

24 29 99.9%

25 30 99.9%

26 31 99.9%

27 32 99.9%

28 33 99.9%

29 34 99.9%

30 35 99.9%

31 36 99.9%

32 37 99.9%

33 38 99.9%

34 39 99.9%

35 40 99.9%

36 41 99.9%

37 42 99.9%

38 43 99.9%

39 44 99.9%

40 45 99.9%

41 46 99.9%

42 47 99.9%

43 48 99.9%

44 49 99.9%

45 50 99.9%

46 51 99.9%

47 52 99.9%

48 53 99.9%

49 54 99.9%

50 55 99.9%

51 56 99.9%

52 57 99.9%

53 58 99.9%

54 59 99.9%

55 60 99.9%

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.

	Ind.Income Prdt	Manually Calculated
1	3	2.4%
2	4	5.0%
3	7	10.0%
4	8	12.5%
5	10	19.0%
6	17	57.3%
7	18	63.2%
8	23	85.7%
9	27	94.2%
10	30	97.2%
11	33	98.6%
12	45	99.9%
13	56	100.0%
14	60	100.0%
15		
16		
17		
18		
19		
20		
21		
22		
23		
24		
25		
26		
27		
28		
29		
30		
31		
32		
33		
34		
35		
36		
37		
38		
39		
40		
41		
42		
43		
44		
45		
46		
47		
48		
49		
50		
51		
52		
53		
54		
55		
56		
57		
58		
59		
60		

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.

```

RStudio
File Edit Code View Plots Session Build Debug Tools Help
Console ~/j0
> lt <- glm(Purchase ~ Income, data = lg, family = "binomial")
> summary(lt)

Call:
glm(formula = Purchase ~ Income, family = "binomial", data = lg)

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|)    
(Intercept) -3.9414    2.0782 -1.897   0.0579 .  
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
> |

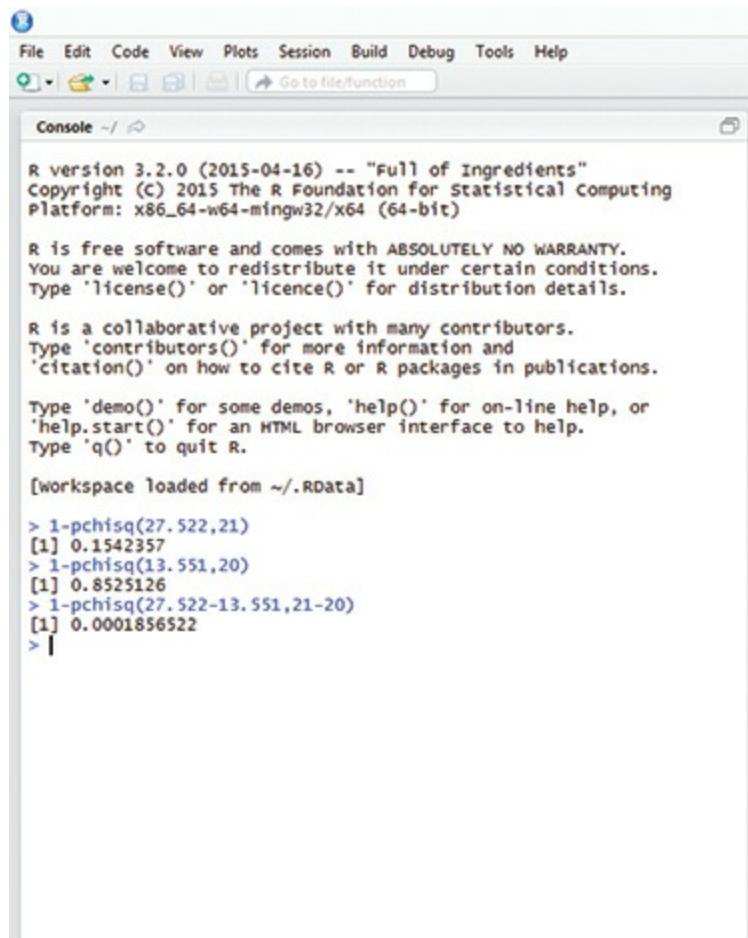
```

The first data point of interest is  $\text{Pr}(>|z|)$  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.



The screenshot shows the RStudio interface with the R console tab active. The console window displays the R startup message and a series of R commands and their outputs. The commands involve the use of the `pchisq` function to calculate p-values for chi-square tests. The first command is `> 1-pchisq(27.522,21)`, followed by its output [1] 0.1542357. Subsequent commands show the calculation of p-values for different degrees of freedom (13.551, 20) and the subtraction of two chi-square values (27.522 - 13.551) with 21 degrees of freedom, resulting in a p-value of 0.0001856522.

```
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

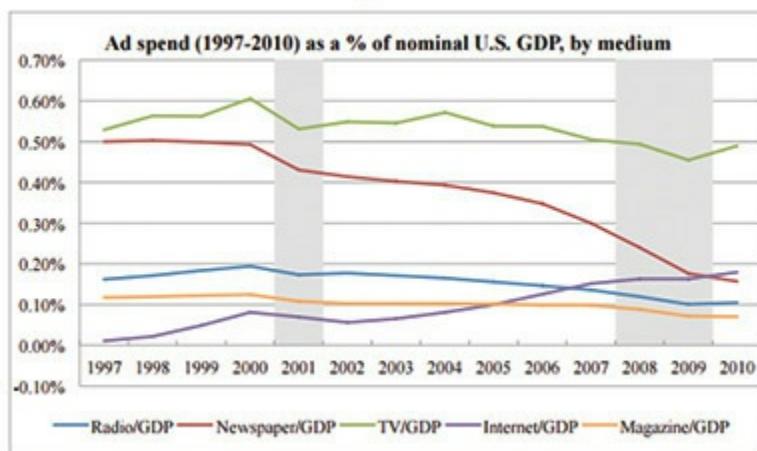
## DIGITAL MARKETING

“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? Invariably you will get answer as 4Ps and STP. Let me start with some basic understanding of 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 answered 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 optimize 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 non-paid 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.

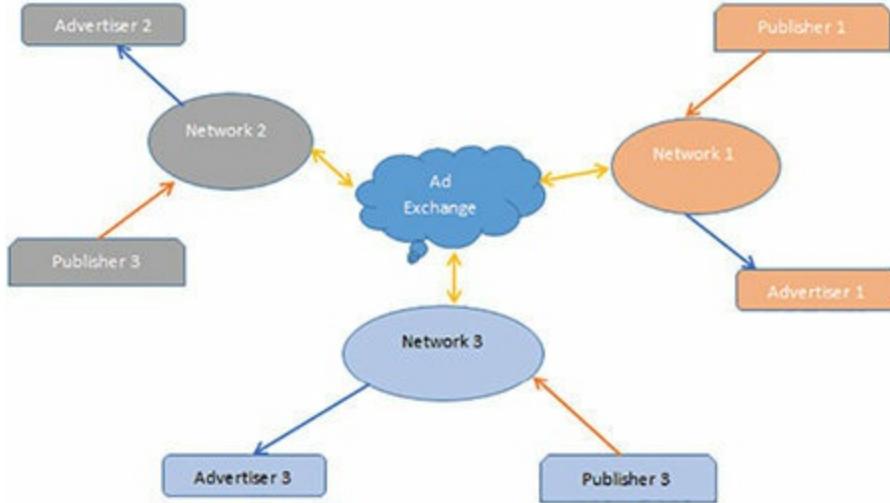
The focus of this chapter will be on the performance marketing side. Before we start analyzing the data for the performance of each channel we have to understand the mechanics of each channel and what are the key performance levers of the channels and how do we optimize for the best performance.

### **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.

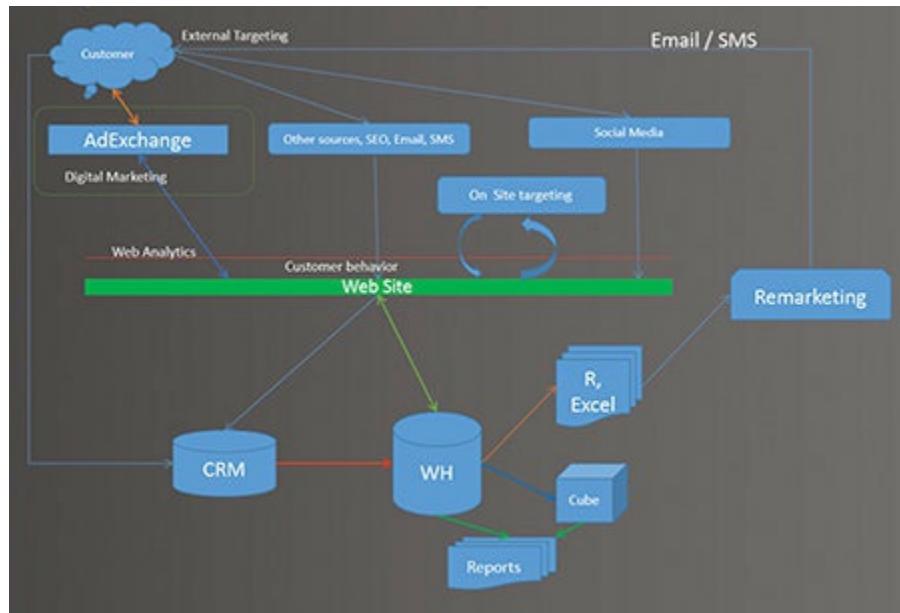
**Advertisers:** The individual or company who want to advertise their product or services or brand etc. Components of the ad ecosystem are

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).

$$\text{CPO} = \text{cost} / \text{order} = \text{avg. CPC} * \text{clicks} / \text{order} = \text{Avg CPC} * \text{clicks} / \text{clicks} * \text{CR} = \text{Avg. CPC} / \text{CR}$$

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

Below is the sample of a campaign run by [www.myweb.com](http://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	Budget	Campaign type	Impressions	Clicks	CTR	Avg. CPC	Cost	Converted	Cost/converted click	Click conversion rate	Conversions	Conversion Rate	View-through conv.
Display_Myweb_Jeans	2,000	Display Only	14898	84	0.56%	3.39	452.78	0	0	0.00%	0	0.00%	0
Display_Myweb_Tshirt	2,000	Display Only	88511	32	0.36%	3.73	196.98	0	0	0.00%	0	0.00%	0
Display_Myweb_Shorts	8,000	Display Only	51650	264	0.51%	4.76	1257.94	2	418.97	0.16%	2	1.89%	2
Display_Myweb_Men_Shoes	2,000	Display Only	20126	71	0.35%	5.49	389.96	0	0	0.00%	0	0.00%	0
Display_Myweb_Women_Shoes	80,000	Display Only	187508	418	0.35%	10.87	4,512.34	4	1,081.37	0.88%	4	1.27%	4
Display_Myweb_Bags	2,000	Display Only	91668	20	0.08%	4.25	88.09	0	0	0.00%	0	0.00%	0
Display_Myweb_Purses	2,000	Display Only	70121	42	0.56%	7.33	300.36	0	0	0.00%	0	0.00%	0
Display_Myweb_Formal_Shirts	80,000	Display Only	558794	857	0.26%	8.44	7,048.58	11	442.14	1.31%	11	2.27%	11
Display_Myweb_Formal_Pants	20,000	Display Only	134307	750	0.54%	8.53	6,096.37	1	8,095.37	0.14%	1	0.27%	1
Display_Myweb_Belts	5,000	Display Only	129096	487	0.36%	4.93	2,396.90	1	2,396.90	0.21%	1	0.45%	1
Display_Myweb_Plus_Sizes	3,000	Display Only	67882	239	0.37%	6.64	2,333.93	0	0	0.00%	0	0.00%	0
Display_Myweb_Albis	25,000	Display Only	190071	2892	0.91%	8.23	20,413.01	22	925.81	0.37%	22	0.30%	22
Total			1247463	5995	0.46%	6.79	17,747.51	32	1,479.61	0.57%	32	0.88%	32

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 for 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.

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 determined by quality score and the bidding. The Google maintains quality score of each keyword based on the CTR, landing pages and ads relevance. In below picture I have added quality score columns.

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%. 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.

Campaign	Keyword	Impressions	Clicks	Avg. CPC	Cost	Converted clicks	Cost/converted click	Click conversion rate	Quality score	Avg. position	Search share	
Search_Mixed_Laptop	Laptops shopping online	2,000	68	3.40%	4.56	310.1	1	310.1	1.47%	10	2.5	50.00%
Search_Mixed_Laptop	New laptop buying	4,587	105	2.30%	10.45	1,097.7	1	1,097.7	0.93%	6	4.5	100.00%
Search_Mixed_Laptop	New laptop	2,345	70	3.00%	29.59	515.9	2	515.9	4.29%	8	4.1	45.65%
Search_Mixed_Laptop	Best laptop to buy	28,007	975	2.90%	18.67	9,012.0	5	1,802.6	0.87%	6	2.7	42.37%
Search_Mixed_Laptop	laptop offer	2,456	195	3.90%	25.69	4,789.3	8	592.6	4.23%	8	3.7	88.92%
Search_Mixed_Laptop	New new laptop	89,787	4,040	4.90%	9.87	39,878.9	84	474.7	2.09%	7	4.3	42.86%
Search_Mixed_Laptop	Best best buy	3,476	75	2.17%	38.15	2,877.6	2	1,438.8	2.65%	7	3.1	49.88%
Search_Mixed_Laptop	Laptops for sale	3,310	74	3.20%	45.23	3,992.4	3	997.5	4.03%	7	4.6	85.00%
Search_Mixed_Laptop	Best buy laptop	78,800	3,916	2.67%	12.45	26,388.7	20	819.5	1.52%	8	4	32.50%
Total		209,748	6,509	3.10%	12.13	78,951	127	621.7	1.95%	4.0	34.0%	

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	Avg. CPC	Cost	Converted clicks	Cost / converted click	Click conversion 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%
PLA_Mobiles	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	2.49%	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 lead 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 collects 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.

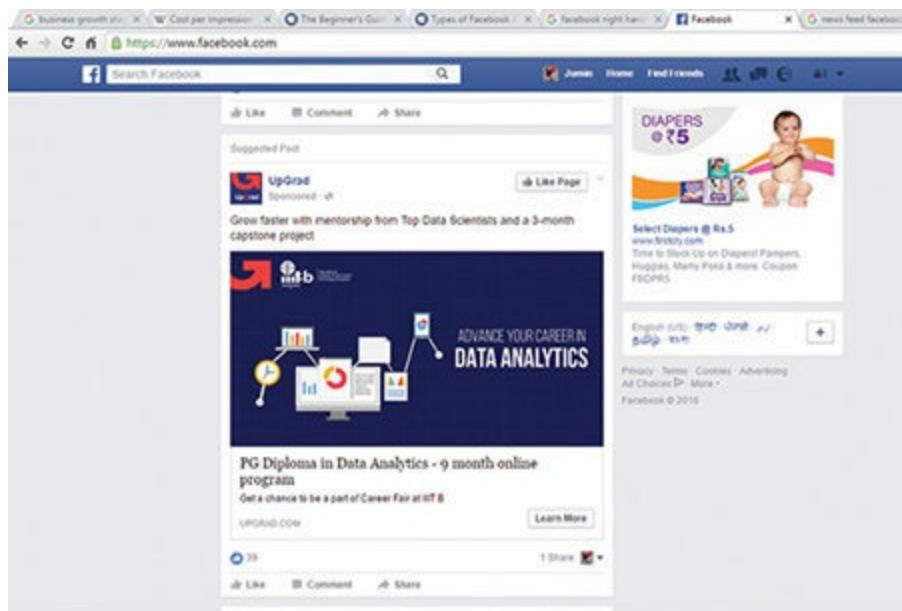
#### 5.1.4.1 Domain Ads

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.



### 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.

Jasper's Market  
Sponsored

Jasper's is more than just groceries! You can now find all of your cooking tools and utensils on our website.

**Set of 5 White Ceramic White Bowls**  
jaspers-market.com

**White Ceramic Mortar and Pestle**  
jaspers-market.com

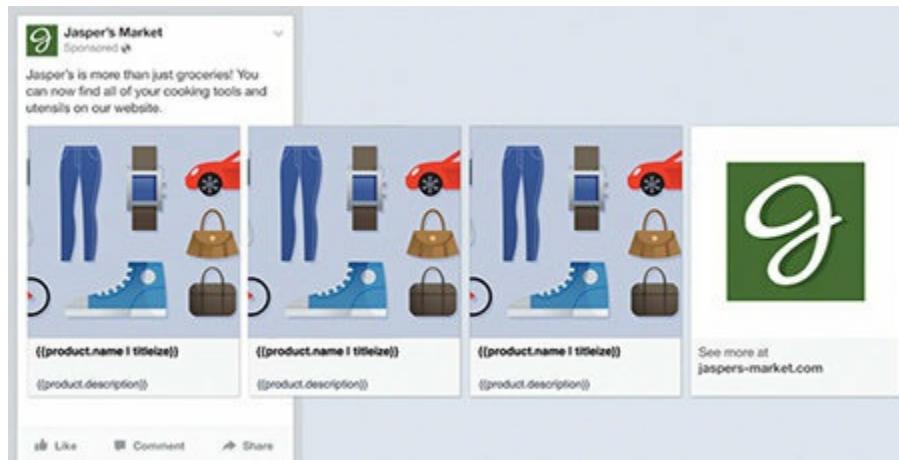
**Set of 2 Wooden Serving Utensils**  
jaspers-market.com

See more at [Jaspers-market.com](http://Jaspers-market.com)

Like · Comment · Share

#### 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 (All)	Amount spent (INR)	CPC (All) (INR)	Conversion (1 Day)	Frequency	Reach
Facebook Apparel	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	24090
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 Apparels Category meant for Diwali Festival, I can tag my url [www.myweb.com/apparels/diwali](http://www.myweb.com/apparels/diwali) with utm\_source = Facebook&utm\_medium = Paid & utm\_campaign = Diwali is added to url ([www.myweb.com/apparels/diwali&utm\\_source=Facebook&utm\\_medium=Paid&utm\\_campaign=Diwali](http://www.myweb.com/apparels/diwali&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

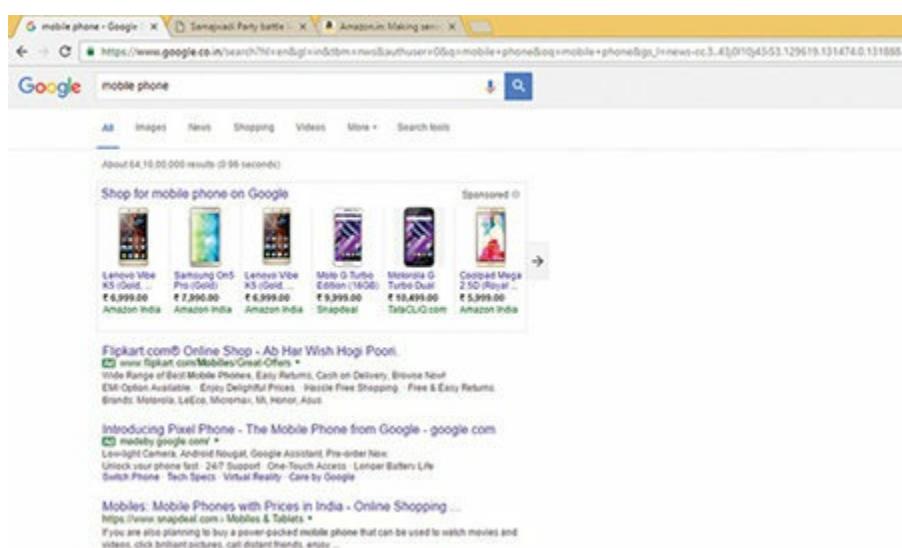
tab of the Google Analytics. The tab will provide the number of traffic, transaction and revenue from the sources in the period. Same data can be extracted to the database directly using the Pentaho Data Integration API interface. The quality of the classification of the channels will entirely depends 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)

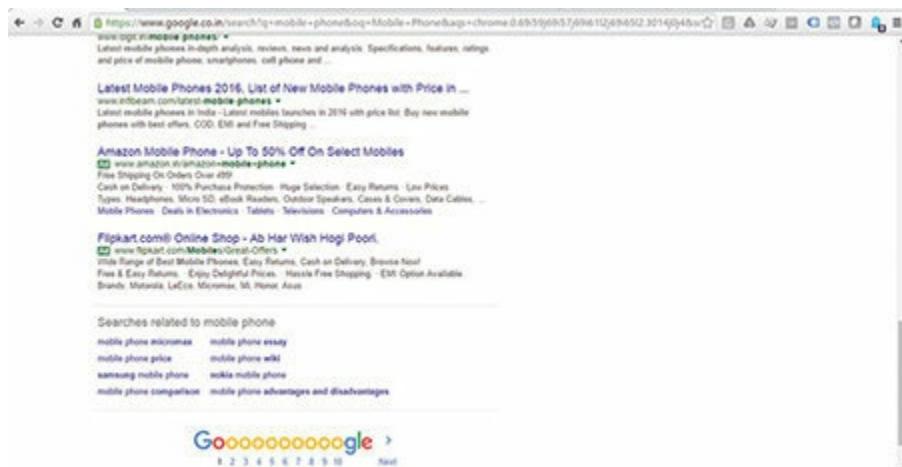
According to Webopedia, Search engine optimization is a methodology of strategies, techniques and tactics used to increase the amount of visitors to a website by obtaining a high-ranking placement in the search results page of a search engine (SERP) -- including Google, Bing, Yahoo and other search engines.

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.

$$\text{Objectives Function Max } Z = 2000X_1 + 1400X_2 + 1500X_3 + 1000X_4$$

Subject to

$$650X_1 + 600X_2 + 450X_3 + 500X_4 \leq 100000$$

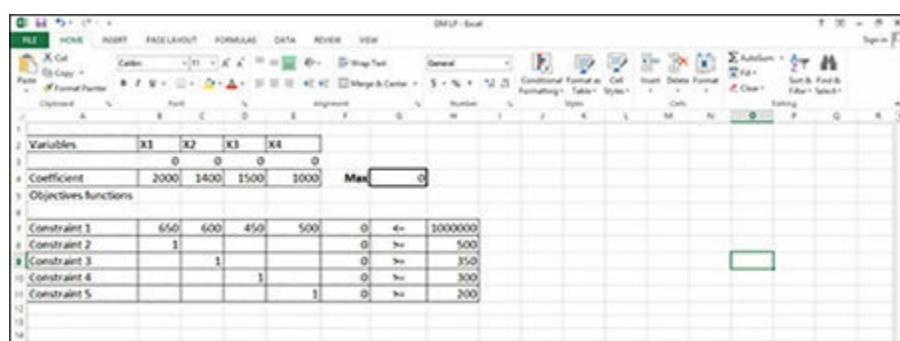
$$X_1 \geq 500$$

$$X_2 \geq 350$$

$$X_3 \geq 300$$

$$X_4 \geq 200$$

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.



In solver we added G4 as the max objectives and added all constraint.

The screenshot shows the Microsoft Excel interface with the Solver Parameters dialog box open. The objective cell is \$G4, and the maximum value is specified. The changing variables are \$B1:\$B4. The constraints are defined by the range \$F1:\$F5, which must be greater than or equal to the values in \$B1:\$B4. The engine selected is Simplex LP.

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.

The screenshot shows the Microsoft Excel interface with the Solver Results dialog box open. It displays the message "Solver found a solution. All constraints and optimality conditions are satisfied." The global optimum value is 2900000. The dialog box also includes options to keep Solver Solution, restore original values, and run Solver Parameters again.

Let us change the objectives function a bit. We have now objective of achieving a revenue target with minimum cost.

$$\text{Objectives function } \min Z = 650X_1 + 600X_2 + 450X_3 + 500X_4$$

Subject to

$$2000X_1 + 1400X_2 + 1500X_3 + 1000X_4 \geq 5000000$$

$$X_1 \geq 500$$

$$X_2 \geq 350$$

$$X_3 \geq 300$$

$$X_4 \geq 200$$

The formulation is shown above

The constraint has been changed

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.

	Channel wise Average Bill Value				Channel 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 bring 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 underperforms. Channel 1 has no 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

Now let us formulate the above problem in the Linear Programming Problem. The objective function is assumed to be maximizing the revenue.

$$\begin{aligned}
 \text{Objective Function Max } Z = & 500X_{11} + 350X_{12} + 370X_{13} + 400X_{14} \\
 & + 12000X_{21} + 8000X_{22} + 750X_{23} + 6000X_{24} \\
 & + 1200X_{31} + 950X_{32} + 1000X_{33} + 800X_{34} \\
 & + 6000X_{41} + 7500X_{42} + 6500X_{43} + 9000X_{44} \\
 & + 560X_{51} + 600X_{52} + 750X_{53} + 800X_{54}
 \end{aligned}$$

$$\text{Subject to } X_{11} + X_{12} + X_{13} + X_{14} \geq 5000 \text{ (Book constraints)}$$

$$X_{21} + X_{22} + X_{23} + X_{24} \geq 8000 \text{ (Mobile constraints)}$$

$$X_{31} + X_{32} + X_{33} + X_{34} \geq 10000 \text{ (Fashion constraints)}$$

$$X_{41} + X_{42} + X_{43} + X_{44} \geq 6000 \text{ (Furniture constraints)}$$

$$X_{51} + X_{52} + X_{53} + X_{54} \geq 2500 \text{ (Kitchen constraints)}$$

$$X_{11} + X_{21} + X_{31} + X_{41} + X_{51} \geq 1000 \text{ (Minimum channel 1 constraint)}$$

$$X_{12} + X_{22} + X_{32} + X_{42} + X_{52} \geq 500 \text{ (Minimum channel 2 constraint)}$$

$$X_{13} + X_{23} + X_{33} + X_{43} + X_{53} \geq 250 \text{ (Minimum channel 3 constraint)}$$

$$X_{14} + X_{24} + X_{34} + X_{44} + X_{54} \geq 250 \text{ (Minimum channel 4 constraint)}$$

$$X_{12} + X_{22} + X_{32} + X_{42} + X_{52} \leq 15000 \text{ (Maximum channel 2 constraint)}$$

$$X_{13} + X_{23} + X_{33} + X_{43} + X_{53} \leq 12000 \text{ (Maximum channel 3 constraint)}$$

$$X_{14} + X_{24} + X_{34} + X_{44} + X_{54} \leq 10000 \text{ (Maximum channel 4 constraint)}$$

$$200X_{11} + 120X_{12} + 150X_{13} + 230X_{14}$$

$$+ 650X_{21} + 450X_{22} + 700X_{23} + 500X_{24}$$

$$+ 240X_{31} + 250X_{32} + 300X_{33} + 260X_{34}$$

$$+ 850X_{41} + 700X_{42} + 600X_{43} + 1000X_{44}$$

$$+ 140X_{51} + 200X_{52} + 150X_{53} + 250X_{54} \leq 50000000$$

$$X_{ij} \geq 0 \text{ (Greater than 0 constraints)}$$

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.

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.

Category	Margin%	Channel wise Average Bill Value				Channel wise cost per Order			
		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  $Z = 20\%*500X_{11} + 20\%*350X_{12} + 20\%*370X_{13} + 20\%*400X_{14}$

$$\begin{aligned}
 & + 5\%*12000X_{21} + 5\%*8000X_{22} + 5\%*750X_{23} + 5\%*6000X_{24} \\
 & + 35\%*1200X_{31} + 35\%*950X_{32} + 35\%*1000X_{33} + 35\%*800X_{34} \\
 & + 25\%*6000X_{41} + 25\%*7500X_{42} + 25\%*6500X_{43} + 25\%*9000X_{44} \\
 & + 15\%*560X_{51} + 15\%*600X_{52} + 15\%*750X_{53} + 15\%*800X_{54}
 \end{aligned}$$

Subject to

$$X_{11} + X_{12} + X_{13} + X_{14} \geq 5000 \text{ (Book constraints)}$$

$$X_{21} + X_{22} + X_{23} + X_{24} \geq 8000 \text{ (Mobile constraints)}$$

$$X_{31} + X_{32} + X_{33} + X_{34} \geq 10000 \text{ (Fashion constraints)}$$

$$X_{41} + X_{42} + X_{43} + X_{44} \geq 6000 \text{ (Furniture constraints)}$$

$$X_{51} + X_{52} + X_{53} + X_{54} \geq 2500 \text{ (Kitchen constraints)}$$

$$X_{11} + X_{21} + X_{31} + X_{41} + X_{51} \geq 1000 \text{ (Minimum channel1 constraint)}$$

$$X_{12} + X_{22} + X_{32} + X_{42} + X_{52} \geq 500 \text{ (Minimum channel2 constraint)}$$

$$X_{13} + X_{23} + X_{33} + X_{43} + X_{53} \geq 250 \text{ (Minimum channel3 constraint)}$$

$$X_{14} + X_{24} + X_{34} + X_{44} + X_{54} \geq 250 \text{ (Minimum channel4 constraint)}$$

$$X_{12} + X_{22} + X_{32} + X_{42} + X_{52} \leq 15000 \text{ (Maximum channel2 constraint)}$$

$$X_{13} + X_{23} + X_{33} + X_{43} + X_{53} \leq 12000 \text{ (Maximum channel3 constraint)}$$

$$X_{14} + X_{24} + X_{34} + X_{44} + X_{54} \leq 10000 \text{ (Maximum channel 4 constraint)}$$

$$\begin{aligned} & 200X_{11} + 120X_{12} + 150X_{13} + 230X_{14} \\ & + 650X_{21} + 450X_{22} + 700X_{23} + 500X_{24} \\ & + 240X_{31} + 250X_{32} + 300X_{33} + 260X_{34} \\ & + 850X_{41} + 700X_{42} + 600X_{43} + 1000X_{44} \\ & + 140X_{51} + 200X_{52} + 150X_{53} + 250X_{54} \leq 50000000 \end{aligned}$$

$$X_{ij} \geq 0 \text{ (Greater than 0 constraints)}$$

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.

**Variables:**

	Average order per value	Margin% Order	CPO
X11	500	10%	50
X12	350	10%	35
X13	370	10%	37
X14	400	10%	40
X21	12000	5%	600
X22	8000	5%	400
X23	7500	5%	375
X24	6000	5%	300
X31	1200	35%	420
X32	950	35%	332.5
X33	1000	35%	350
X34	800	35%	280
X41	6000	25%	1500
X42	7500	25%	1875
X43	6500	25%	1625
X44	9000	25%	2250
X51	560	15%	84
X52	600	15%	90
X53	750	15%	112.5
X54	800	15%	120

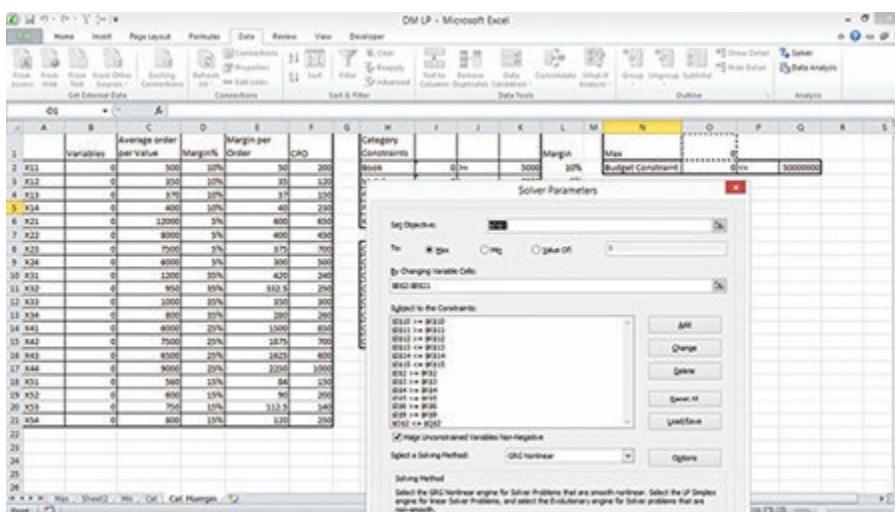
**Category Constraints:**

	Category	Constraints	Margin
1	Book	0 >= 5000	10%
2	Mobiles	0 >= 8000	5%
3	Fashion	0 >= 10000	35%
4	Furniture	0 >= 6000	25%
5	Kitchens	0 >= 2500	15%

**Channel Constraints:**

	Channel	Constraints	Max
1	C1	0 >= 3000	3000
2	C2	0 >= 500	500
3	C3	0 >= 250	250
4	C4	0 <= 15000	15000
5	C5	0 <= 12000	12000
6	C6	0 <= 10000	10000

The constraint will remain same across the category



	Average order	Margin	Margin per Order	CPO	Category	Constraints	Margin	Max	103099705.9
X11	5000	500	10%	50	200	Book	5000 >= 5000	10%	
X12	0	350	10%	35	120	Mobiles	8000 >= 8000	5%	
X13	0	370	10%	37	150	Fashion	10000 >= 10000	35%	
X14	0	400	10%	40	230	Furniture	52675 >= 6000	25%	
X21	8000	12000	5%	600	650	Kitchen	2500 >= 2500	15%	
X22	0	8000	5%	400	450				
X23	0	2500	5%	375	700				
X24	0	6000	5%	300	500				
X31	10000	1200	35%	420	240	C1	41176 >= 10000		
X32	0	950	35%	332.5	250	C2	15000 >= 950		
X33	0	1000	35%	350	300	C3	32000 >= 1000		
X34	0	800	35%	280	260	C4	10000 >= 800		
X41	15676	6000	25%	1500	850	C2	15000 <= 15676		
X42	25000	2500	25%	1875	700	C3	12000 <= 25000		
X43	32000	8500	25%	1625	600	C4	10000 <= 32000		
X44	10000	9000	25%	2250	1000				
X51	2500	560	15%	88	350				
X52	0	600	15%	90	200				
X53	0	250	15%	112.5	340				
X54	0	800	15%	120	250				

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

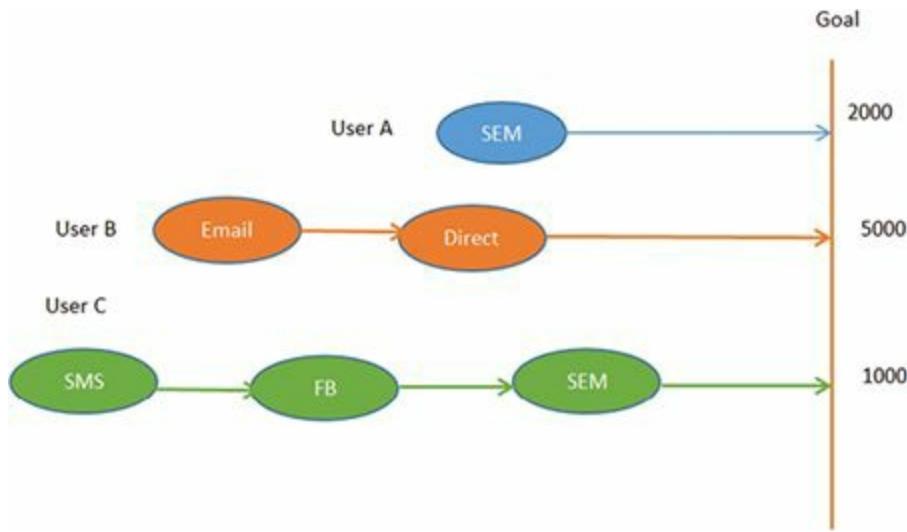
X44	0	10000
X51	0	2500
X52	0	0
X53	2500	0
X54	0	0

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.

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.

The screenshot shows the Google Analytics Reporting interface with the 'Model Comparison Tool' selected. The left sidebar includes sections for Home, Reporting, Customization, Admin, Acquisition, Behavior, and Conversions. Under Conversions, 'Model Comparison...' is highlighted. The main area displays a comparison between 'Last Interaction' and 'Linear' models. The primary dimension is set to 'MCF Channel Grouping'. The table lists three MCF Channel Groupings: Direct, Organic Search, and (Other). The 'Last Interaction' model shows values of 9,158.00 (0.47%), 5,348.00 (0.47%), and 2,189.00 (0.47%) respectively. The 'Linear' model shows values of €4,338,792.00 (0.00%), €1,709,537.00 (0.00%), and €708,232.00 (0.00%).

You can select two models and compare using the comparison tool.

This screenshot is identical to the one above, showing the Google Analytics Model Comparison Tool interface. It compares 'Last Interaction' and 'Linear' models across MCF Channel Groupings: Direct, Organic Search, and (Other). The 'Last Interaction' model values are 9,158.00, 5,348.00, and 2,189.00 respectively. The 'Linear' model values are €4,338,792.00, €1,709,537.00, and €708,232.00 respectively.

The setting of the attribution model is done in Admin -> Attribution Models

The screenshot shows the Google Analytics Admin interface with the 'Attribution Models' section selected. On the left, there are navigation links for Home, Reporting, Customization, and Admin. The right side shows a list of attribution models: 'All Products', 'Pixelbacks', 'Audience Definitions', 'Custom Definitions', 'Data Import', and 'Social Settings'. Below these are 'PERSONAL TOOLS & ASSETS' sections for 'Segments', 'Annotations', 'Attribution Models' (which is highlighted in red), 'Custom Channel Groupings', 'Metrics', 'Custom Alerts', 'Scheduled Emails', 'Shortcuts', and 'Share Assets'.

You can build your custom attribution models by changing the look back windows, % attribution in position based and half-life time.

The screenshot shows the Google Analytics Admin interface with the 'Attribution Models' section selected. On the left, there are navigation links for Home, Reporting, Customization, and Admin. The right side shows a form for creating a new model. The 'Model Name' field is empty. The 'Baseline Model' dropdown is set to 'Linear'. There are several options listed under 'Linear': 'Linear', 'First Interaction', 'Last Interaction', 'Time Decay', and 'Position Based'. To the right of these options are checkboxes for 'Send metrics from the resulting visit' and 'My model is defined'.

In time decays model you can change the time period of half-life of the decay

The screenshot shows the 'Baseline Model' section under the 'ADMIN' tab. It is set to 'Time Decay'. A note says: 'An assist click occurring this many days prior to conversion will receive 1/2 the credit of a click on the day of conversion.' A 'Set half-life of decay' input field shows '7 days'. There are three toggle switches: 'Lockback Window' (On), 'Adjust credit based on user engagement' (Off), and 'Apply custom credit rules' (Off). A 'Save' button is at the bottom.

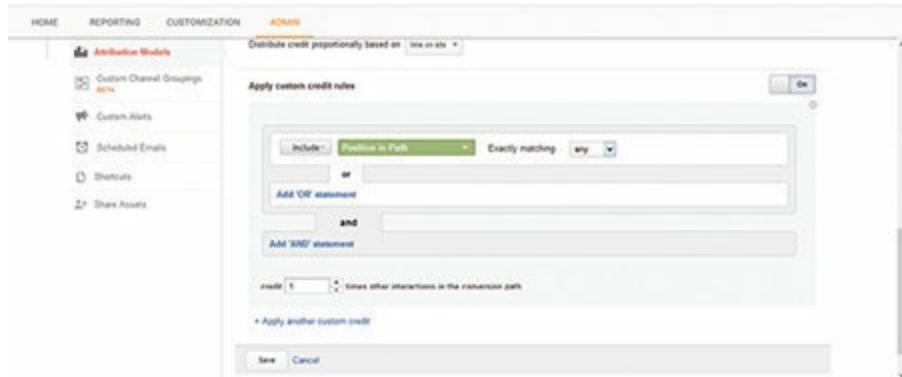
In position based model you can change the % attribution of the first, last and other channels. The first and last is assign and then remaining channels are linearly distributed. In below model 40% is assign to first and 40% to last and remaining 20% to the remaining channels.

The screenshot shows the 'Baseline Model' section under the 'ADMIN' tab. It is set to 'Position Based'. A note says: 'Specify the amount of conversion credit based on the position.' It shows: 'First interaction: 40%', 'Middle interactions: 20%' (with a note 'This will be distributed evenly to all middle interactions'), and 'Last interaction: 40%'. A 'Total: 100%' note says 'Value for 1000'. There are three toggle switches: 'Lockback Window' (On), 'Adjust credit based on user engagement' (Off), and 'Apply custom credit rules' (Off). A 'Save' button is at the bottom.

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.

The screenshot shows the 'Attribution Models' section under the 'ADMIN' tab. It includes: 'Lockback Window' (Set 90 days prior to conversion, On), 'Adjust credit based on user engagement' (Distribute credit proportionally based on 'time-on-site', On), and 'Apply custom credit rules' (Off). A 'Save' button is at the bottom.

You can also define custom credit rules.



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 even 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  $Y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n$

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).

$$Y = \exp(a + b_1x_1 + b_2x_2 + \dots + b_nx_n) / (1 + \exp(a + b_1x_1 + b_2x_2 + \dots + b_nx_n))$$

**Polynomial Regression:** If power of any of the dependent variable is more than 1 then regression becomes polynomial regression

$$\text{E.g } Y = 1 + b_1x_1 + cx^2$$

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  $\sigma_e^2$ .

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

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

where  $\bar{x}_m$  is mean of X and  $\bar{y}_m$  is mean of Y

$$a = y_m - bx_m$$

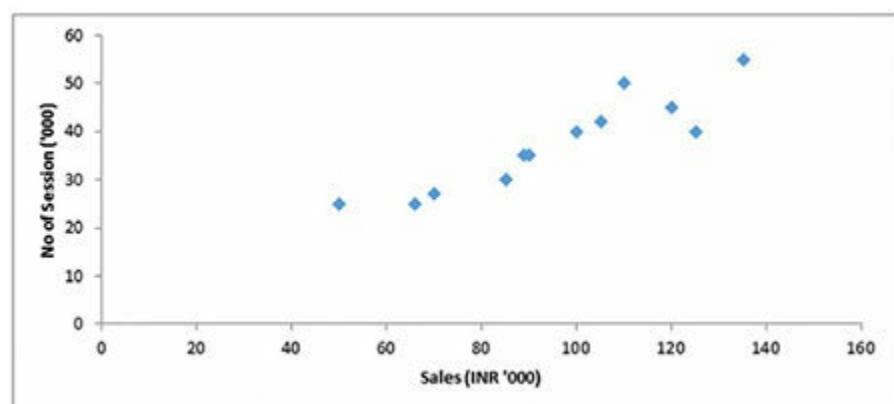
Let us start with a simple linear regression example. Use below data point where sales of a website is dependent on the number of session in a day.

	Sales (INR '000)	No. of Session (000)
2	120	45
3	50	25
4	70	27
5	90	35
6	125	40
7	85	30
8	66	25
9	100	40
10	105	42
11	110	50
12	135	55
13	89	35
14		
15		
16		
17		
18		
19		

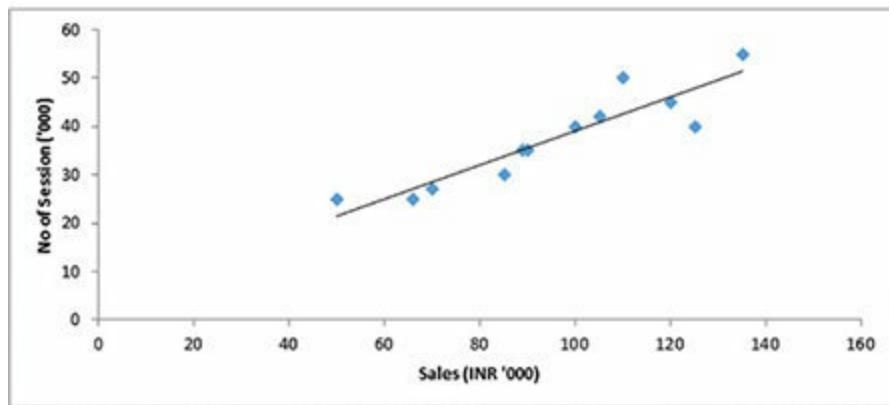
Using least square calculation we have calculated a and b as per formula discussed above.

	Sales (INR '000)	No. of Session (000)	$x_i - \bar{x}_m$	$y_i - \bar{y}_m$	$(x_i - \bar{x}_m)(y_i - \bar{y}_m)$	$(x_i - \bar{x}_m)^2$
2	120	45	2.58	24.58	58.42	57.53
3	50	25	-12.42	-45.42	565.92	154.17
4	70	27	-10.42	-25.42	264.76	108.53
5	90	35	-2.42	-5.42	13.09	5.84
6	125	40	2.58	29.58	76.42	6.67
7	85	30	-2.42	-10.42	77.26	5.93
8	66	25	-12.42	-29.42	365.26	154.17
9	100	40	2.58	4.58	11.84	6.67
10	105	42	4.58	9.58	43.92	21.01
11	110	50	12.58	14.58	183.51	158.34
12	135	55	27.58	39.58	696.01	750.56
13	89	35	-2.42	-6.42	15.53	5.84
14	Total	1145	449		2497.92	1042.92
15	Mean	95.42	37.42			
16	b	2.40				
17	a	5.80				

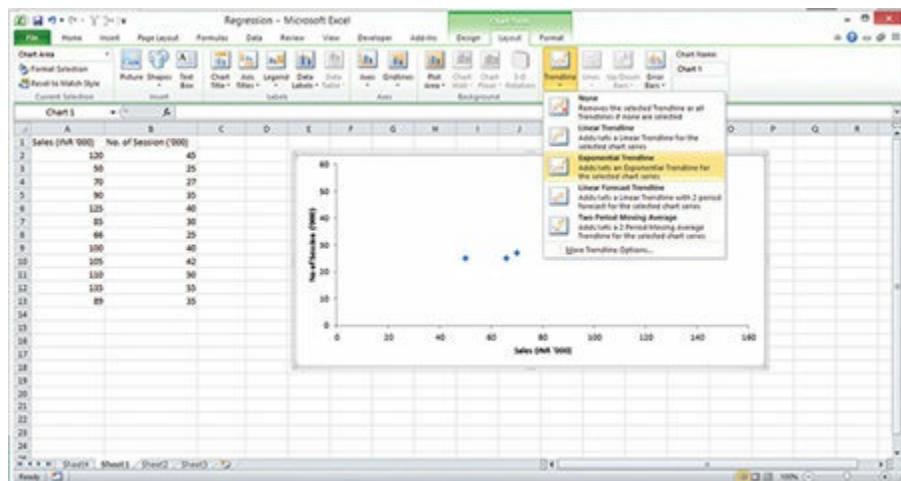
The first thing one has to do for a regression data is to plot the scatter plot in Excel or R. We will learn R process later, here focus is on excel



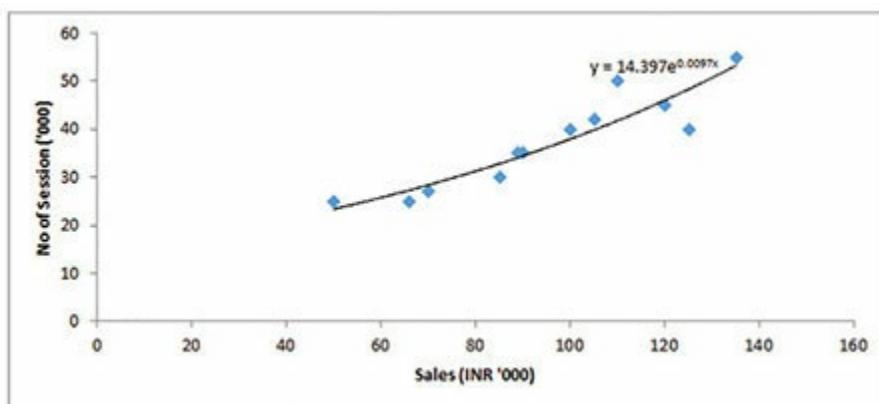
One can add trend line to the scatter plot to check if there is actually linear relationship existed between independent and dependent variables.



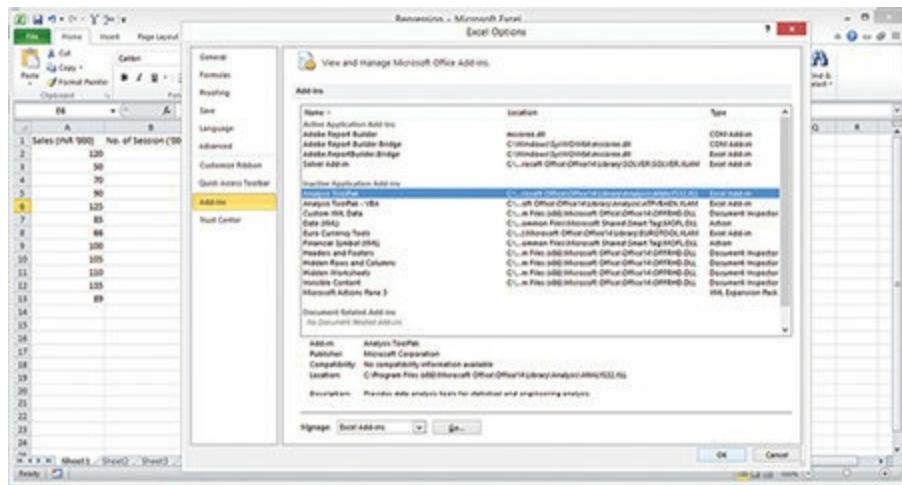
You can add other trend line as well. We added exponential trendline in the scatter plot



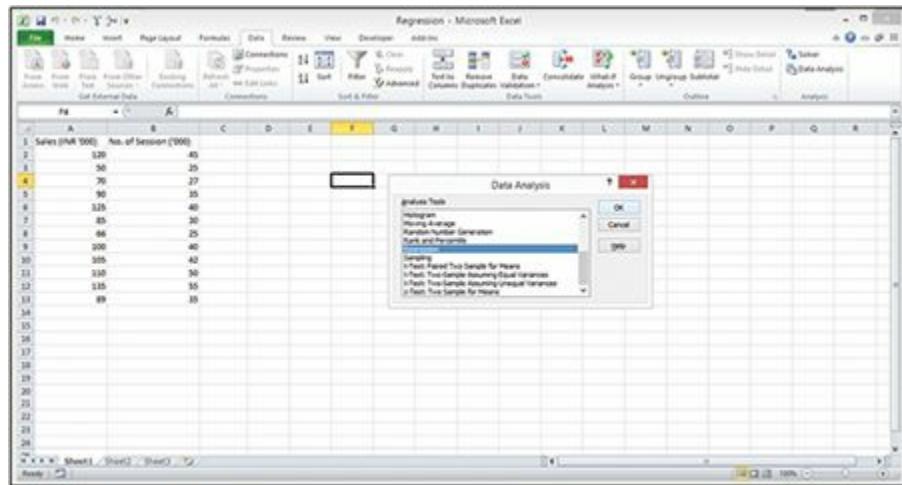
The plot shows the trend line and the equation.



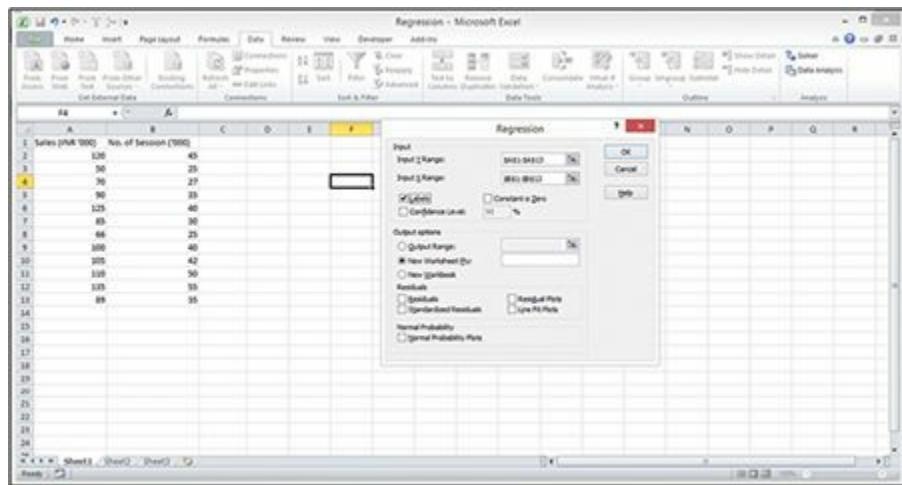
From the linear trend line we can clearly see a good linear relationship between two variables. Let us run actual regression analysis to find the extent of the relationship.



Select regression from the list



Select label as we have headers in the data selection. Confidence interval has been kept at default 95%.



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  $\sum(y_i - \hat{y}_i)^2$ , where  $y_i$  is the value of Y for  $i^{\text{th}}$  observation and  $\hat{y}_i$  is the calculated y from the regression  $\hat{y}_i = a + bx_i$ . This is also called Error Sum of Squares (ESS).

Total Sum of Square (TSS) is calculated as  $\sum(y_i - \bar{y})^2$ . The Regression Sum of Square (RSS) is calculated as the difference between Total Sum of Square minus Residual Sum of Square.

$R^2$  is calculated as  $RSS / TSS = 5982.82 / 7124.91 = 0.839 \sim 83.9\%$ .

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 explain 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.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	5.7990411151	12.76053945	0.454451097	0.659209196	-22.63321257	34.23129487	-22.63321257	34.23129487
No. of Session ('000)	2.395125849	0.33092206	7.237734001	2.79866E-05	1.657785549	3.132466149	1.657785549	3.132466149

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

The screenshot shows the RStudio interface with the following code and output:

```
File Edit Code View Plots Session Build Debug Tools Help
Untitled1.R
1 Reg<-read.csv(file="C:/RData/SimpleReg.csv", header=TRUE)
2 Reg
3

3.1 (Top Level) 1
Console <-->
    abbreviate, write
> Reg<-read.csv(file="C:/RData/SimpleReg.csv", header=TRUE)
> Reg
   Sales Session
1    120      45
2     50      25
3     70      27
4     90      35
5    125      40
6     85      30
7     66      25
8    100      40
9    105      42
10    110      50
11    135      55
12     89      35
>
```

Install packages required for the regression

The screenshot shows the RStudio interface with the following code:

```
File Edit Code View Plots Session Build Debug Tools Help
Untitled1.R
1 Reg<-read.csv(file="C:/RData/SimpleReg.csv", header=TRUE)
2 Reg
3 install.packages("car")
4 library("car")
5
```

```

file<-read.csv(file="C:/RData/SimpleReg.csv", header=TRUE)
library("car")
OP<-lm(Sales~Session, data=Reg)
summary(OP)

Call:
lm(formula = Sales ~ Session, data = Reg)

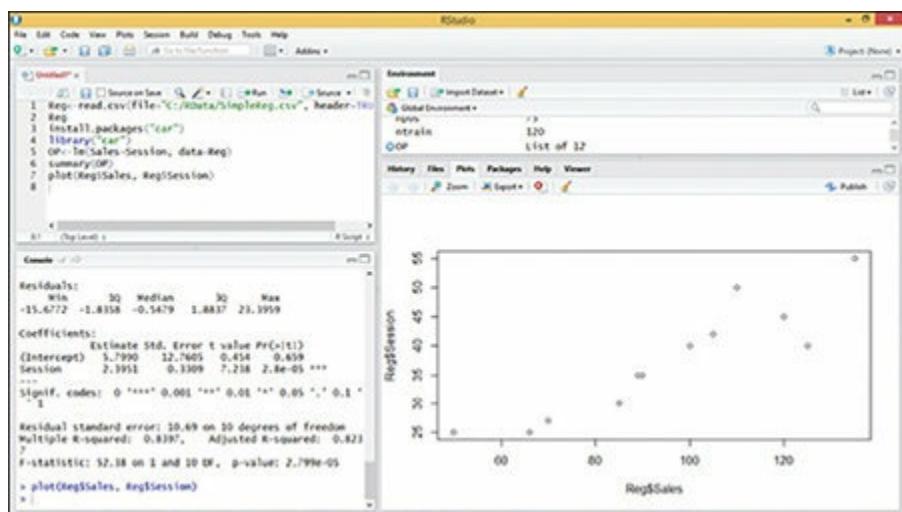
Residuals:
    Min      1Q  Median      3Q     Max 
-15.6772 -1.8358 -0.5479  1.8837 23.3959 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.7990    12.7605   0.454   0.659    
Session     2.3951    0.3309   7.238  2.8e-05 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.69 on 10 degrees of freedom
Multiple R-squared:  0.8397, Adjusted R-squared:  0.8237 
F-statistic: 52.38 on 1 and 10 DF,  p-value: 2.799e-05

```

The regression output is same as we get in excel. For the large data set excel may not be 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

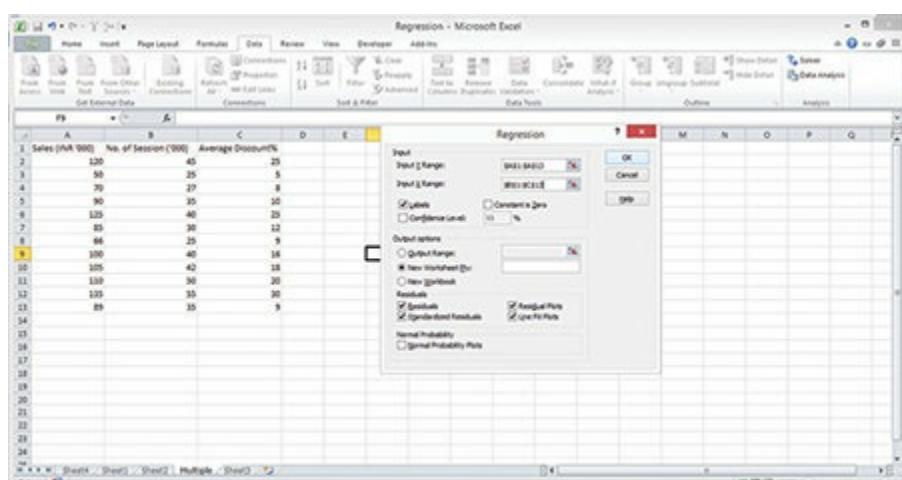
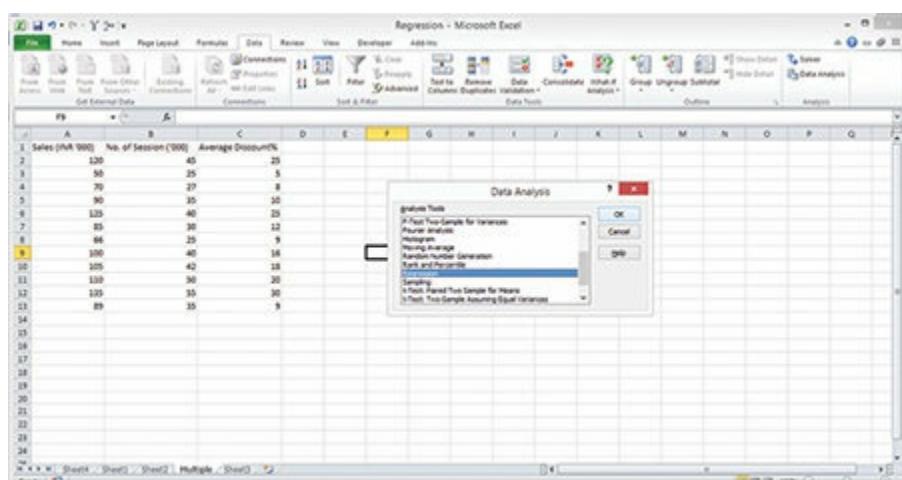
$$Y = a + b_1 X_1 + b_2 X_2 + \dots + b_n X_n + e$$

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.

Regression - Excel

	A	B	C
1	Sales (INR '000)	No. of Session ('000)	Average Discount%
2	120	45	25
3	50	25	5
4	70	27	8
5	90	35	10
6	125	40	25
7	85	30	12
8	66	25	9
9	100	40	16
10	105	42	18
11	110	50	20
12	135	55	30
13	89	35	9
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			

## Select Regression



**SUMMARY OUTPUT**

Regression 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 Stat	PValue	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 Discounts	2.10	0.60	3.52	0.01	0.75	3.45	0.75	3.45

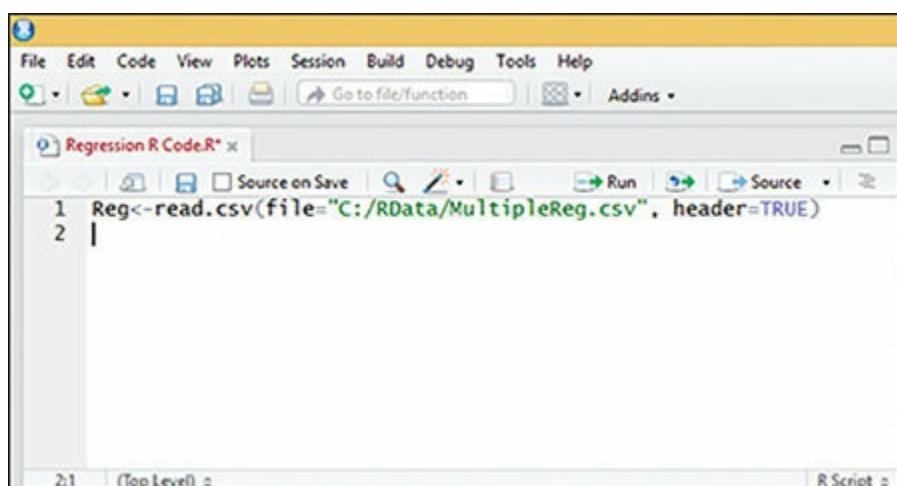
Equation is  $Y = 30.74 + 0.85 X_1 + 2.10 X_2$ .

Assuming number of session is 50K and discount is 19% then Sales

$$Y = 30.74 + 0.85 * 50 + 2.10 * 19$$

$$= 113.4 (\text{'000 INR})$$

## Multiple Regressions in R



```
Regression R Code.R*
File Edit Code View Plots Session Build Debug Tools Help
Source on Save Run Source
1 Reg<-read.csv(file="C:/RData/MultipleReg.csv", header=TRUE)
2 |
```

Using same data set as excel example.

The screenshot shows the RStudio interface. The top panel is titled "Regression R Code.R" and contains the following R code:1 Reg<-read.csv(file="C:/RData/MultipleReg.csv", header=TRUE)
2 RegThe bottom panel is titled "Console" and displays the output of the code:> Reg<-read.csv(file="C:/RData/MultipleReg.csv", header=TRUE)
> Reg
 Sales Session Discount
1 120 45 25
2 50 25 5
3 70 27 8
4 90 35 10
5 125 40 25
6 85 30 12
7 66 25 9
8 100 40 16
9 105 42 18
10 110 50 20
11 135 55 30
12 89 35 9

The screenshot shows the RStudio interface. The top panel is titled "Regression R Code.R" and contains the following R code:1 Reg<-read.csv(file="C:/RData/MultipleReg.csv", header=TRUE)
2 Reg
3 Mout<-lm(Sales ~ Session + Discount, data = Reg)
4The bottom panel is titled "Console" and displays the output of the code:> Reg
 Sales Session Discount
1 120 45 25
2 50 25 5
3 70 27 8
4 90 35 10
5 125 40 25
6 85 30 12
7 66 25 9
8 100 40 16
9 105 42 18
10 110 50 20
11 135 55 30
12 89 35 9
> Mout<-lm(Sales ~ Session + Discount, data = Reg)
>

The output is same as excel output.

```

File Edit Code View Plots Session Build Debug Tools Help
Go to file/function Addins

Regression.R* x
Source on Save Run Source
1 Reg<-read.csv(file="C:/RData/MultipleReg.csv", header=TRUE)
2 Reg
3 Mout<-lm(Sales ~ Session + Discount, data = Reg)
4 summary(Mout)
5

S1 (Top Level) R Script

Console > 
> Mout<-lm(Sales ~ Session + Discount, data = Reg)
> summary(Mout)

Call:
lm(formula = Sales ~ Session + Discount, data = Reg)

Residuals:
    Min      1Q  Median      3Q     Max
-12.5774 -5.0984  0.0018  4.4774  9.4846

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.7439   11.2425   2.735  0.02305 *
Session      0.8529    0.4932   1.729  0.11784
Discount     2.1023    0.5974   3.519  0.00653 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.308 on 9 degrees of freedom
Multiple R-squared:  0.9325, Adjusted R-squared:  0.9175
F-statistic: 62.2 on 2 and 9 DF, p-value: 5.381e-06
>

```

### 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

$$Y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n + e$$

Here expected value of error term e is assumed to be 0

In GLM the relationship between Y and X is assumed to be

$$Y = g(a + b_1x_1 + b_2x_2 + \dots + b_nx_n) + e$$

Where e is error and g is a function  $g(x)$ . The inverse of g is called link function denoted by  $f(x)$

$$f(\mu_y) = a + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

$\mu_y$  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 based on historical relationship between dependent and 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.

$$F_t = (1/N) (D_{t-1} + D_{t-2} + \dots + D_{t-n})$$

In below example we calculated 5 days SMA and 10 days SMA.

Date	Price	5-Day SMA	10-day SMA
24-Mar-10	22.27		
25-Mar-10	22.19		
26-Mar-10	22.06		
29-Mar-10	22.17		
30-Mar-10	22.18	22.18	
31-Mar-10	22.13	22.15	
1-Apr-10	22.23	22.16	
5-Apr-10	22.43	22.23	
6-Apr-10	22.24	22.25	
7-Apr-10	22.29	22.27	22.22
8-Apr-10	22.15	22.27	22.21
9-Apr-10	22.39	22.30	22.23
12-Apr-10	22.38	22.29	22.26
13-Apr-10	22.61	22.37	22.31
14-Apr-10	23.36	22.58	22.42
15-Apr-10	24.05	22.96	22.61
16-Apr-10	23.75	23.23	22.77
19-Apr-10	23.83	23.52	22.91
20-Apr-10	23.95	23.79	23.08
21-Apr-10	23.63	23.84	23.21
22-Apr-10	23.82	23.80	23.38
23-Apr-10	23.87	23.82	23.53
24-Apr-10	23.65	23.79	23.65
25-Apr-10	23.19	23.63	23.71
26-Apr-10	23.10	23.53	23.69
27-Apr-10	23.33	23.43	23.61
30-Apr-10	22.68	23.19	23.51
1-May-10	23.10	23.08	23.43
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.



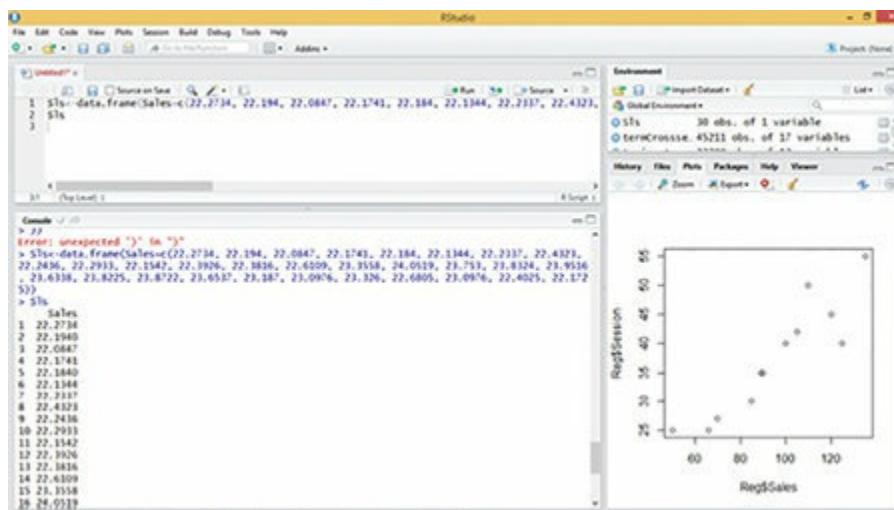
Plot 5 days SMA and 10 days SMA along with the price in the same graph

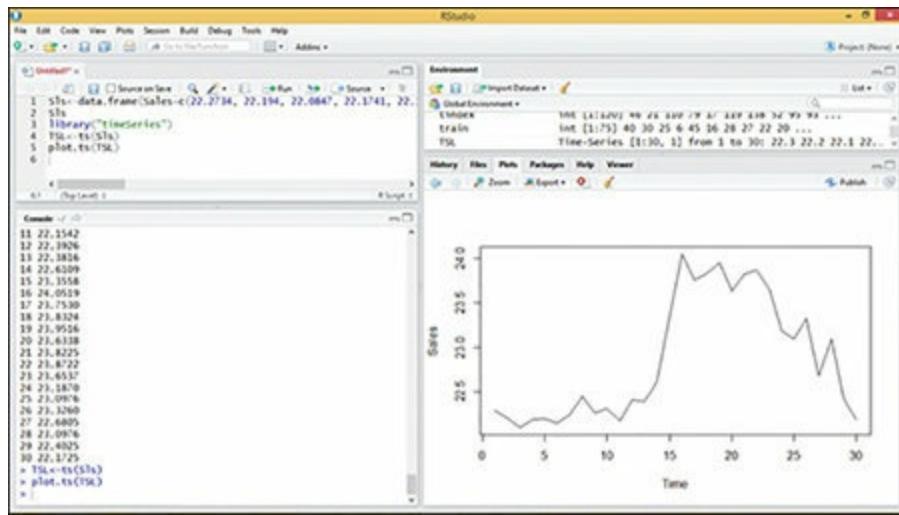


The longer the period or size of the window more the lag in the data. 10 days SMA has lag over 5 days SMA and is smoother also. A moving average is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles. The threshold between short-term and long-term depends on the application, and the parameters of the moving average will be set accordingly.

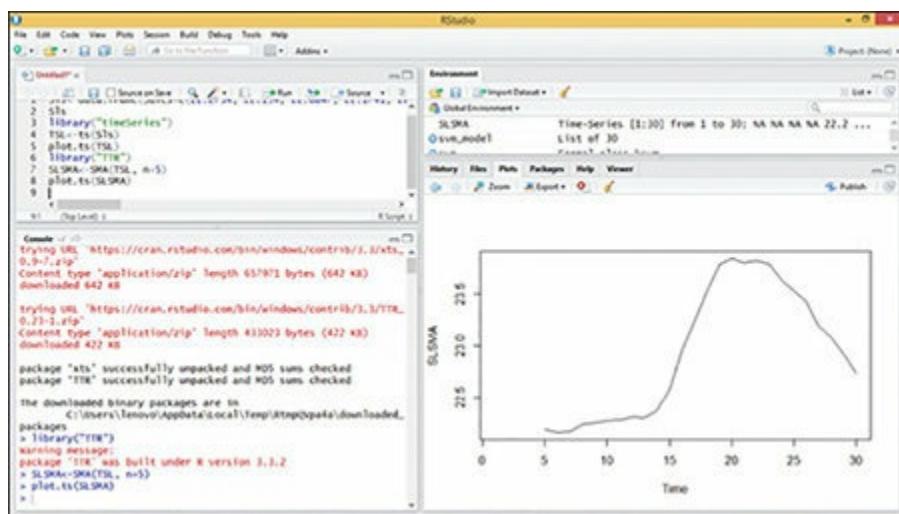
## SMA in R

We will use same data set to show how to calculate and plot SMA in R.

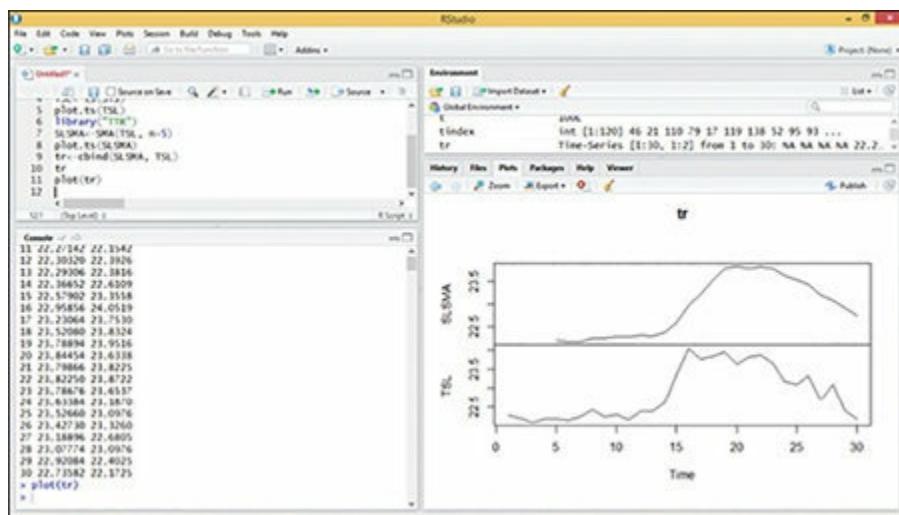




Plot of the actual data.



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  $\alpha$ . 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

$$F_t = \alpha \cdot X_t + (1-\alpha) F_{t-1}$$

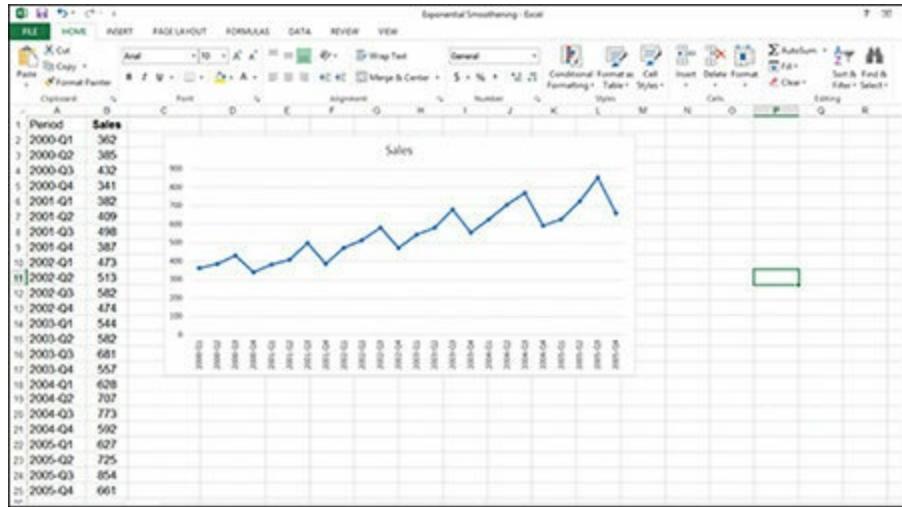
Where  $F_t$  is the weighted average of the recent observation  $X_t$  and the previous smoothed data  $F_{t-1}$ ,  $\alpha$  is the smoothening factor which ranges from  $0 < \alpha < 1$ .

The higher value of  $\alpha$  gives more weightages to the recent observation and the lower value of  $\alpha$  gives more weightages to the older observation in the time series data. There is no ideal value of  $\alpha$  but one can do trial and error to get the required smoothening effects. One statistical method to arrive at  $\alpha$  is to find the value of  $\alpha$  that minimizes the  $(F_{t-1} - X_{t-2})^2$

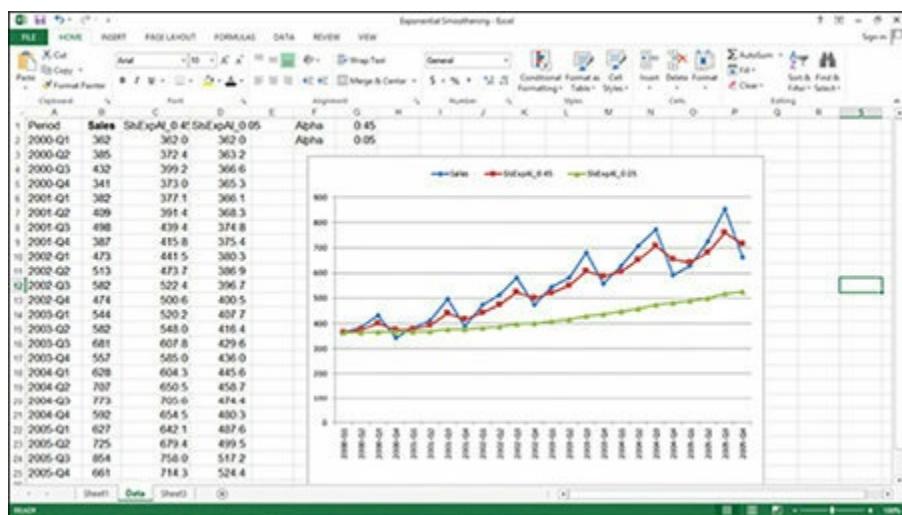
Small values of  $\alpha$  means that the forecasted value will be stable (show low variability) and increases the lag of the forecast to the actual data if a trend is present, whereas large values of  $\alpha$  mean that the forecast will more closely track the actual time series data.

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  $\alpha$  to see how it impact the smoothed 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

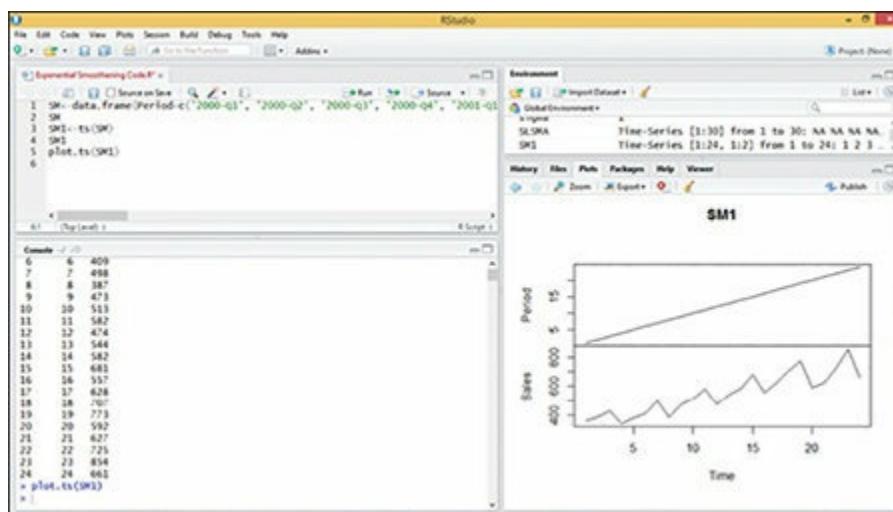
Using same data series to show EWA work in R.

RStudio

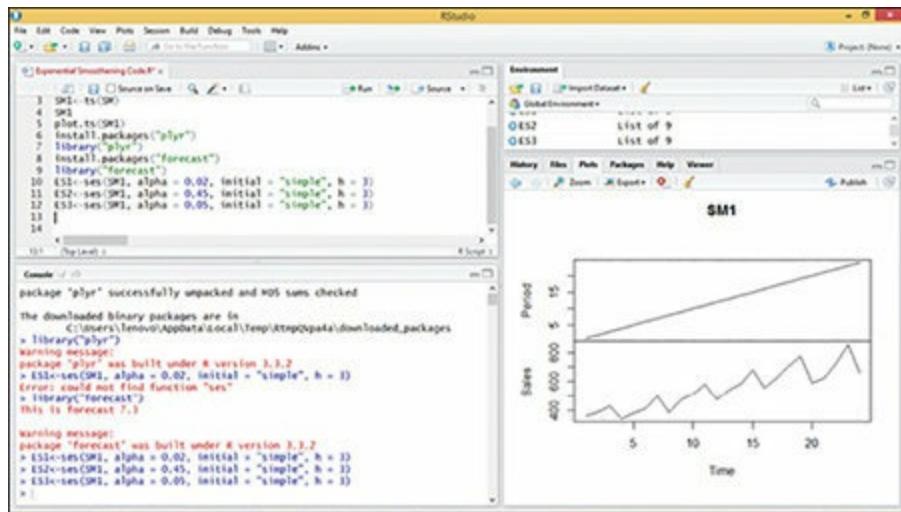
```
File Edit Code View Plots Session Build Debug Tools Help
Source on Save Go to File Functions Addins
Exponential Smoothening Code.R*
1 SM<-data.frame(Period=c('2000-Q1', '2000-Q2', '2000-Q3', '2000-Q4', '2001-Q1',
2 SM
3

< (Top Level) R Script

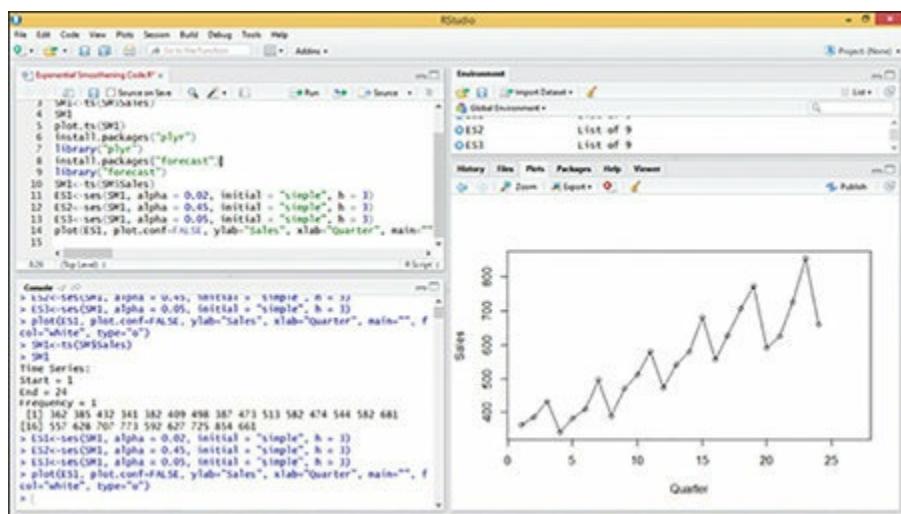
Console ~ / ~
> plot(tr)
> SM<-data.frame(Period=c('2000-Q1', '2000-Q2', '2000-Q3', '2000-Q4', '2001-Q1',
  '2001-Q2', '2001-Q3', '2001-Q4', '2002-Q1', '2002-Q2', '2002-Q3', '2002-Q4',
  '2003-Q1', '2003-Q2', '2003-Q3', '2003-Q4', '2004-Q1', '2004-Q2', '2004-Q3',
  '2004-Q4', '2005-Q1', '2005-Q2', '2005-Q3', '2005-Q4'), Sales<-c(362, 385, 432,
  341, 382, 409, 498, 387, 473, 513, 582, 474, 544, 582, 681, 557, 628, 707, 7
73, 592, 627, 725, 854, 661))
> SM
  Period Sales
1 2000-Q1 362
2 2000-Q2 385
3 2000-Q3 432
4 2000-Q4 341
5 2001-Q1 382
6 2001-Q2 409
7 2001-Q3 498
8 2001-Q4 387
9 2002-Q1 473
10 2002-Q2 513
11 2002-Q3 582
12 2002-Q4 474
*3...3...3...3...3...3...
```



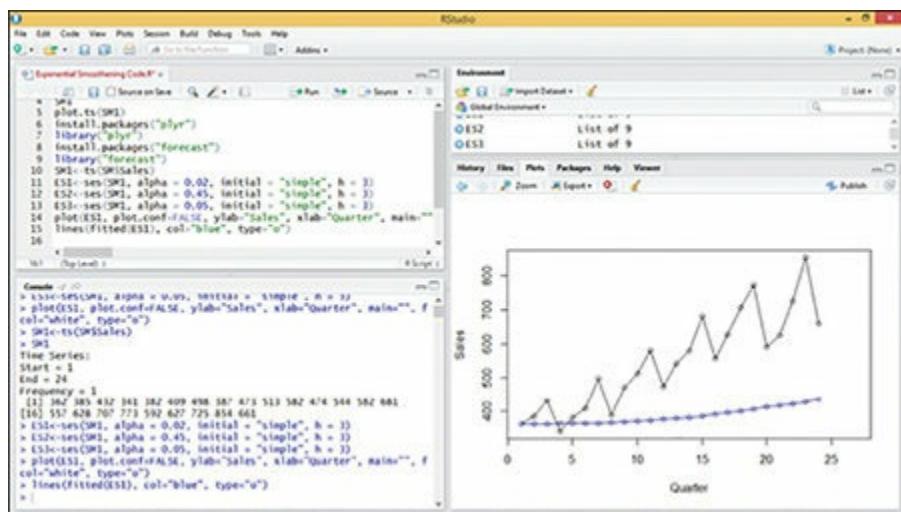
We plot Exponential smoothening using  $\alpha=0.02$  (ES1) ,  $\alpha=0.45$  (ES2) and  $\alpha=0.05$  (ES3)



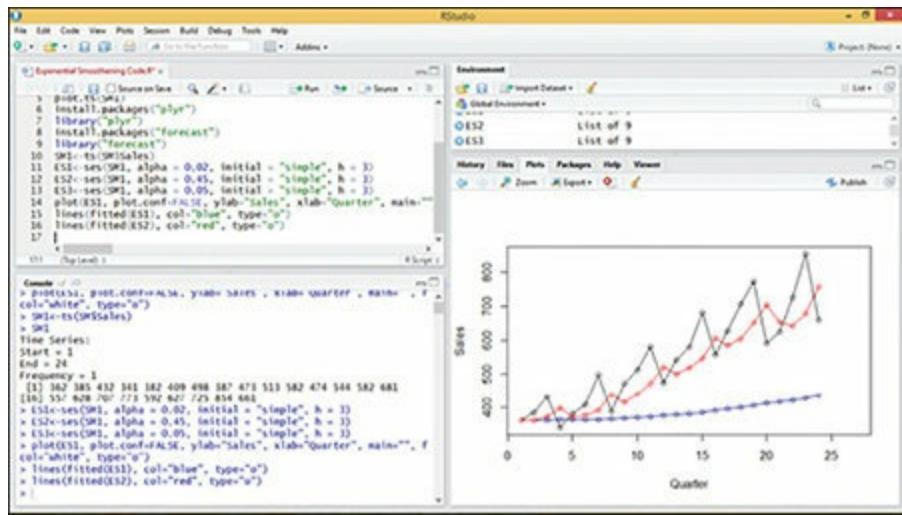
## Adding data point in the graph



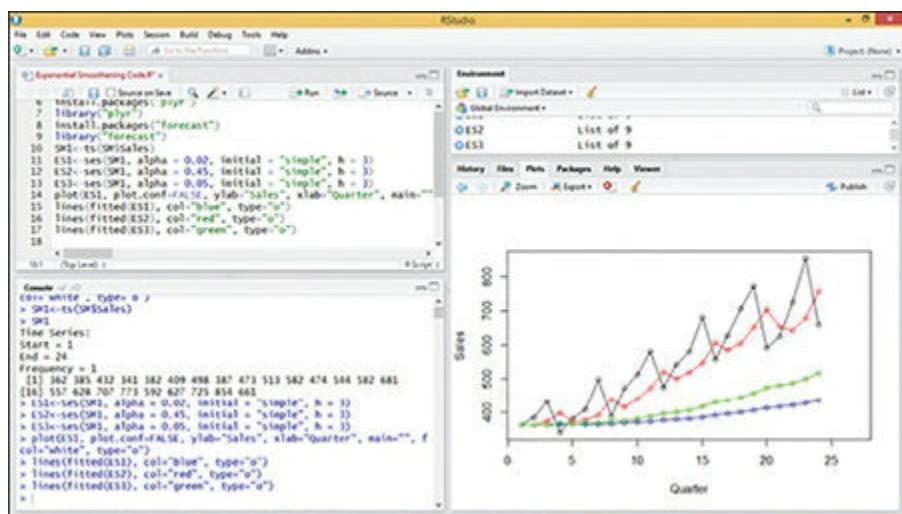
## Add ES1 line with blue color



## Adding ES2 line with red color

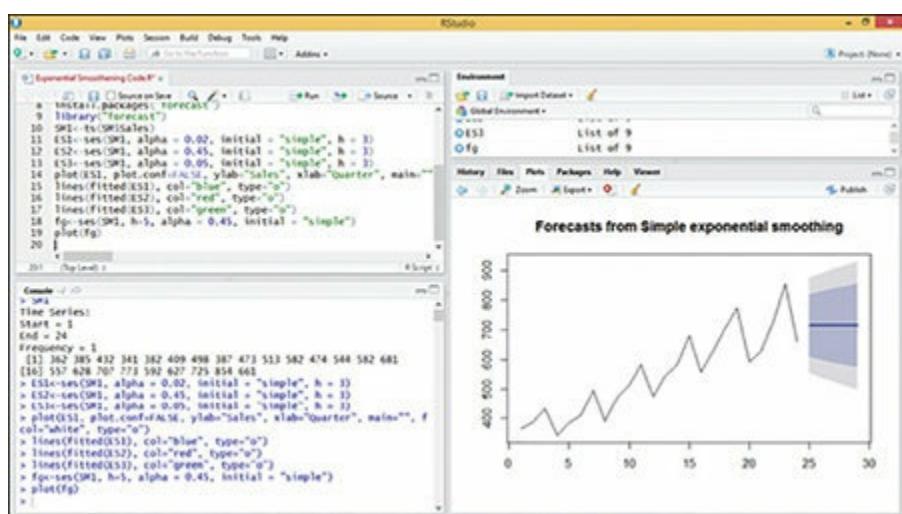


## Adding ES3 line with green color



The plot of  $\alpha=0.45$  and  $\alpha=0.05$  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.



### 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

$$Y_t = c + b_1 Y_{t-1} + b_2 Y_{t-2} + \dots + b_p Y_{t-p} + e_t$$

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.

$$Y_t = c + e_t + b_1 e_{t-1} + b_2 e_{t-2} + \dots + b_q e_{t-q}$$

$Y_t$  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

$y'_{t-m} = y_t - y_{t-m}$  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

$$Y'_t = c + b_1 Y'_{t-1} + b_2 Y'_{t-2} + \dots + b_p Y'_{t-p} + b_1 e_{t-1} + b_2 e_{t-2} + \dots + b_q e_{t-q} + e_t$$

where  $Y'_t$  is the differenced series.

Model with different value of p,q and d gives following types

White noise	ARIMA(0,0,0)
Random walk	ARIMA(0,1,0) with no constant
Random walk with drift	ARIMA(0,1,0) with a constant
Autoregression	ARIMA(p,0,0)
Moving average	ARIMA(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))
```

```

File Edit Code View Plots Session Build Debug Tools Help
Untitled1.R
Source on Save Run Source
1 SM<-data.frame(Period=c('2000-Q1', '2000-Q2', '2000-Q3', '2000-Q4',
2 SM
3

Console -> 
> fgees(SM1, h=5, alpha = 0.45, initial = "simple")
> plot(fg)
> 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', '2
  002-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, 5
13, 582, 474, 544, 582, 681, 557, 628, 707, 773, 592, 627, 725, 854,
661))
> SM
   Period Sales
1 2000-Q1 362
2 2000-Q2 385
3 2000-Q3 432
4 2000-Q4 341
5 2001-Q1 382
6 2001-Q2 409
7 2001-Q3 498

```

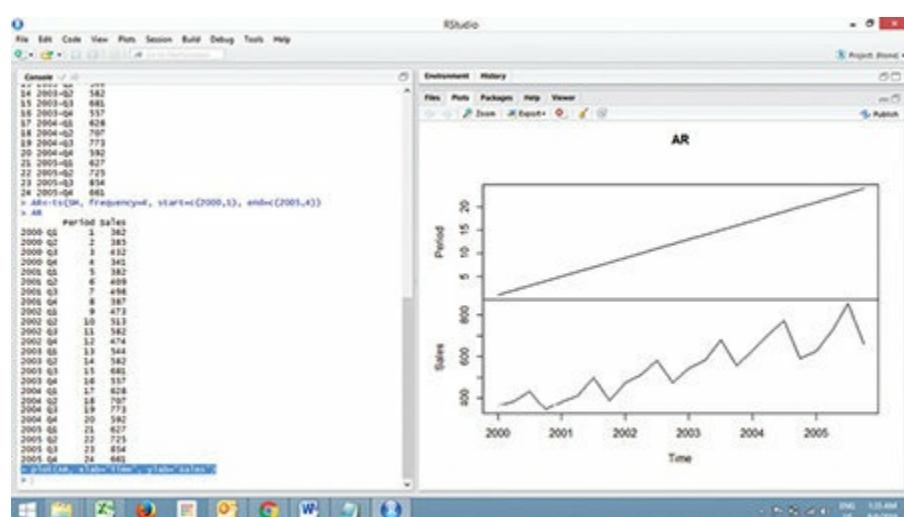
Convert data to time series data

> SM1<-ts(SM\$Sales)

> AR<-ts(SM1, frequency=4, start=c(2000,1), end=c(2005,4))

Plot the time series data

> plot(AR, xlab="Time", ylab="Sales")

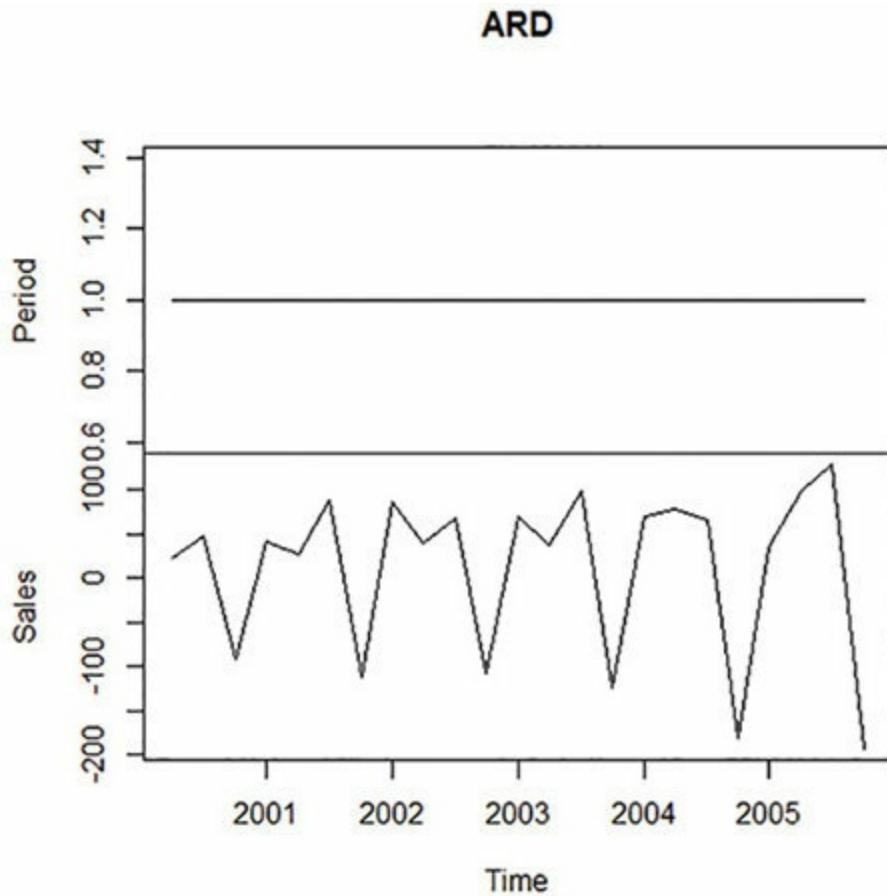


Differencing data to make it stationary as we can see there is clearly a seasonal factor in the data. Plot of differencing d=1

> ARD<-diff(AR)

> plot(ARD, xlab="Time", ylab="Sales")

The mean seems to be stationary but variance is not stationary

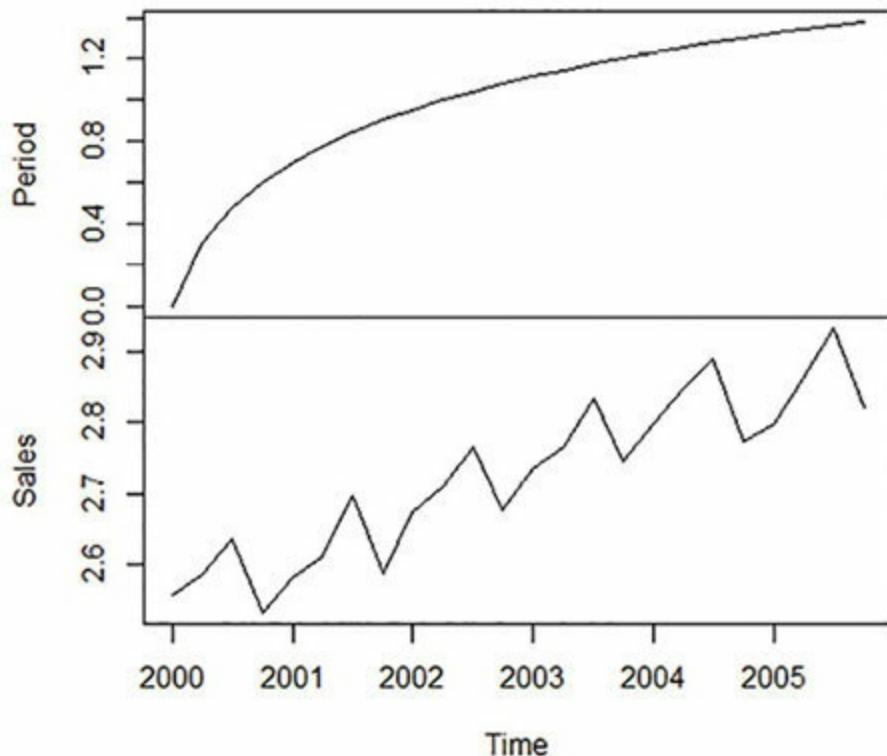


Tryout with logarithm of the data

```
> AR10 <-log10(AR)  
> plot(AR10, xlab="Time", ylab="Sales")
```

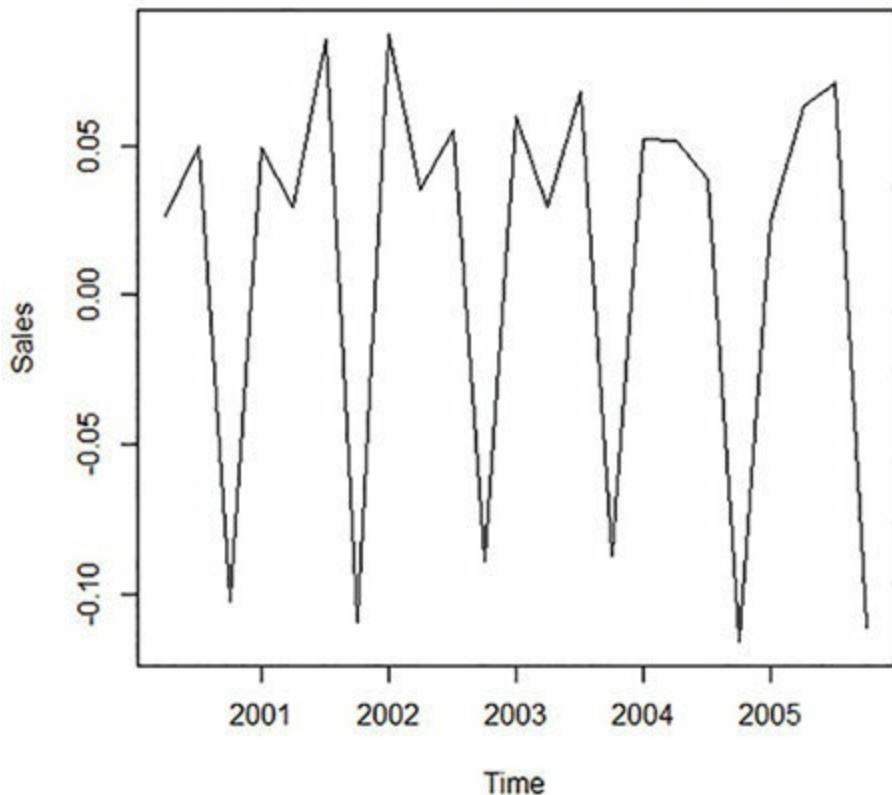
From graph we can see that the variance seems to be stationary but mean is not.

## AR10



The  $\log_{10}$  of the difference of data should make mean and variance stationary

```
> AR10 <-log10(AR)  
> ARDIF10 <-diff(ARDIF10)  
> plot(ARDIF10, xlab="Time", ylab="Sales")
```



Now let's try to find the best fit ARIMA model

```
>require(forecast)
> model<-auto.arima(AR)
> model
```

```
> model<-auto.arima(AR)
> model
Series: AR
ARIMA(0,0,1)(0,1,0)[4] with drift

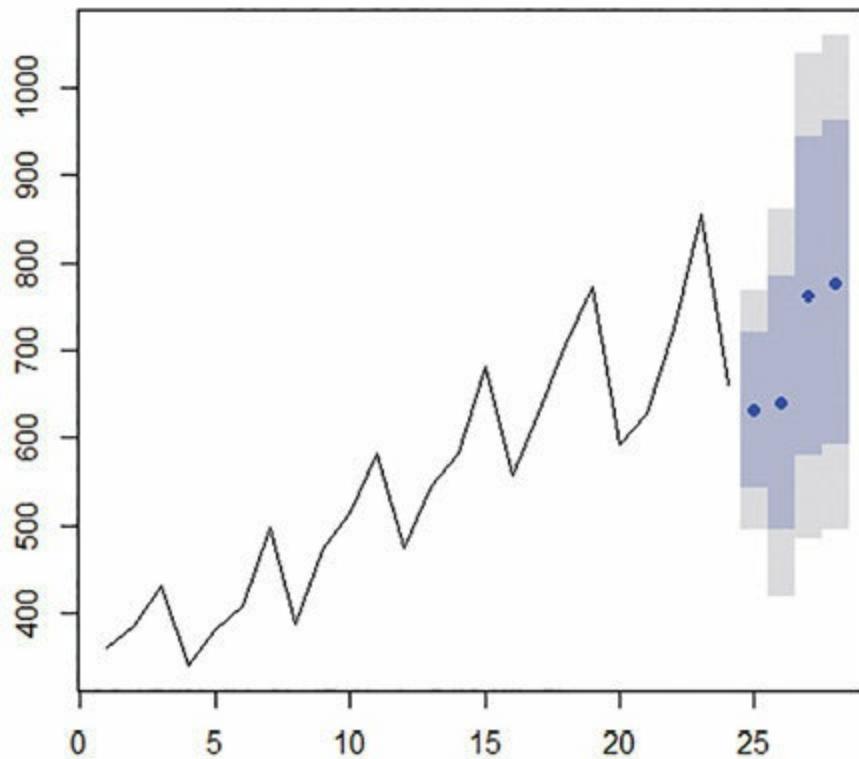
Coefficients:
      m1      drift
      0.9136  16.6876
  s.e.  0.2334   2.4785

sigma^2 estimated as 624.1:  log likelihood=-92.58
AIC=191.15  AICc=192.65  BIC=194.14
> |
```

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))
```

### Forecasts from ARIMA(0,1,3)



```
> forecast(fit, h=4)
```

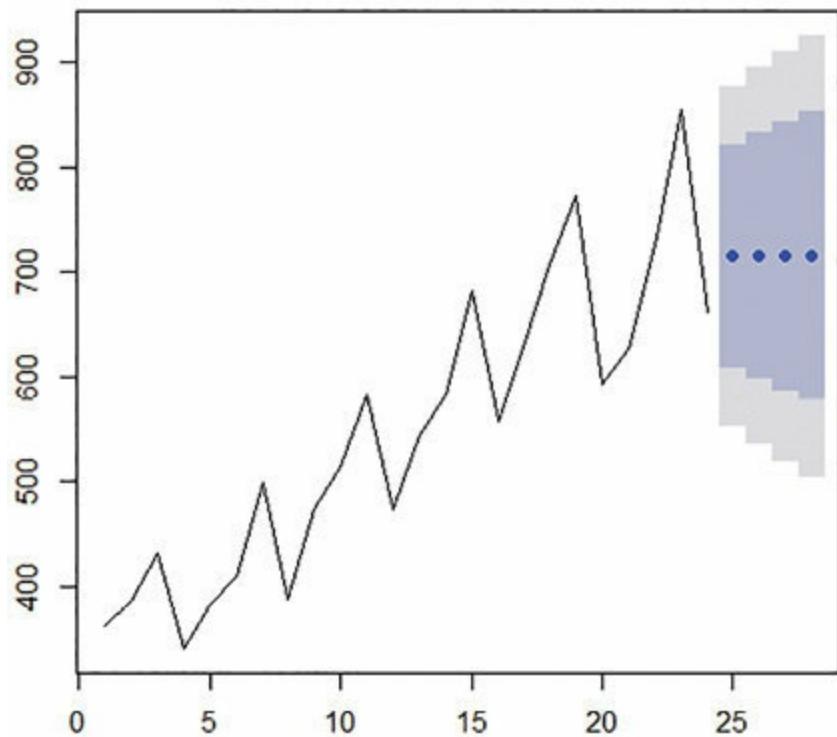
Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
25	633.3992	519.6330	747.1654	459.4088	807.3896
26	612.1237	468.4816	755.7658	392.4421	831.8053
27	595.7239	436.9711	754.4767	352.9325	838.5154
28	583.0825	415.9972	750.1678	327.5476	838.6174

With best fit suggested by system

```
> fit <- arima(SM1,order=c(0,0,1), seasonal =c(0,1,0))
```

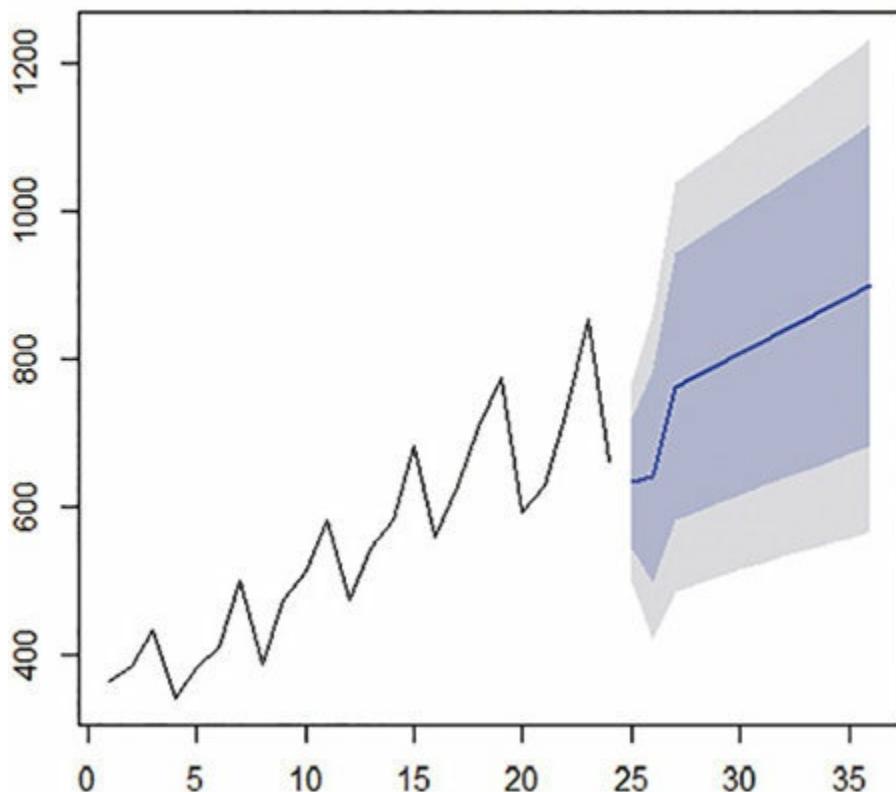
```
> plot(forecast(fit,4))
```

### Forecasts from ARIMA(0,0,1)



```
> plot(forecast(fit, h=12)) # for 12 quarter ahead forecast
```

### Forecasts from ARIMA(0,1,3)



```
> forecast(fit, h=12) # data point for 12 quarter forecast
```

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95

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.

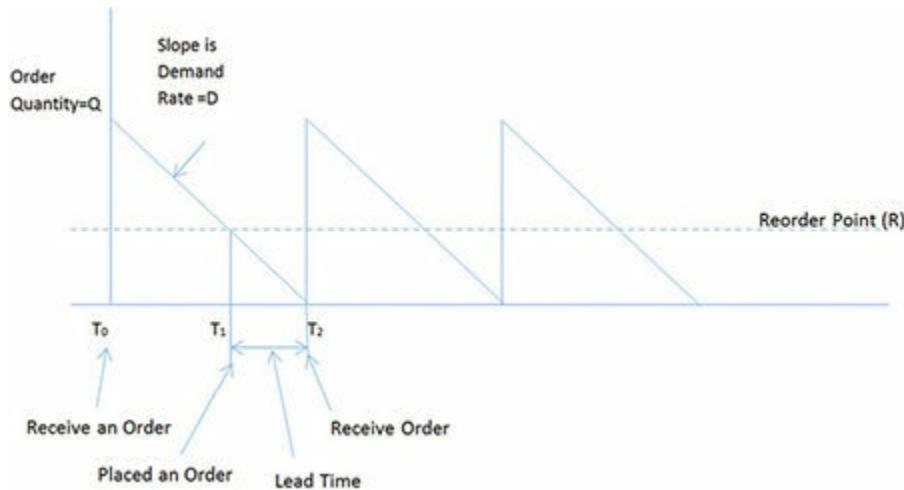
The variables that impact the inventory model are

**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 -

$$\text{Inventory Carrying cost} = Q/2 * H$$

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 -

$$\text{Cost of Ordering} = D/Q * S$$

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

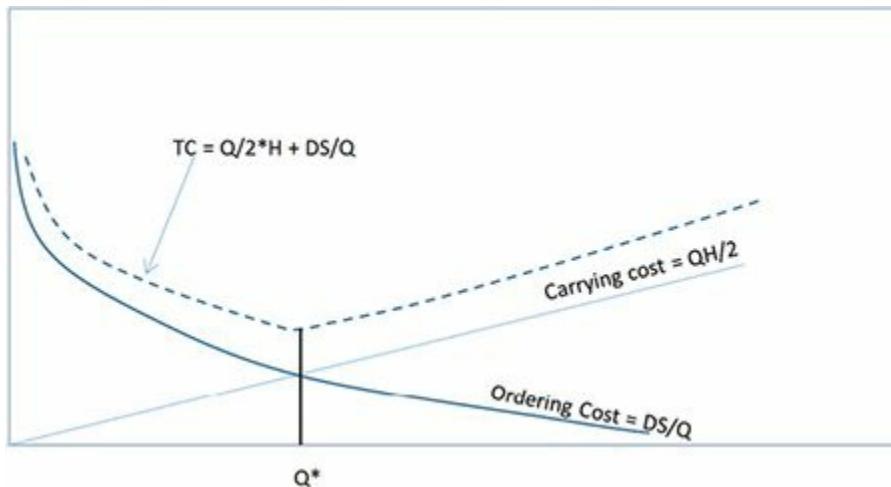
$$dTC/dQ = dQ/2*H + d(D/Q)S = H/2 - DS/Q^2$$

$$\text{at minima } dTC/dQ = 0$$

$$0 = H/2 - DS/Q^2$$

$$Q^2 = 2DS/H$$

$$Q = \sqrt{2DS/H}$$



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.

$$Q^* = \sqrt{2DS/H} = \sqrt{2*100*100/0.02} = 1000 \text{ bulbs.}$$

$$\text{Reorder Quantity} = 5*100=500.$$

The inventory is ordered when it falls just below 500.

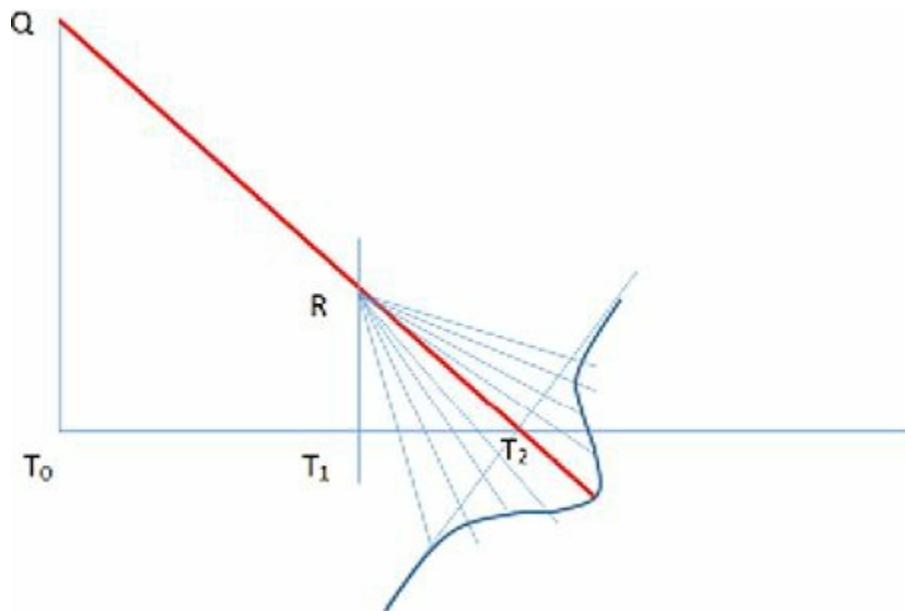
Total cost of the inventory on daily basis=  $[1000/2 * 0.02] + [100/(1000/100)] = \text{Rs. 20 per day}$

Frequency of order = years Demand / EOQ =  $100*365/1000 = 36.5 \sim 37$  order in a year

$$\text{Cost of Ordering} = \text{Ordering Cost} * \text{Frequency of Order} = 37*100 = 3700$$

## 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  $T_1$  and the order will be received in lead time ( $L$ ) that is  $T_2$ . 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.

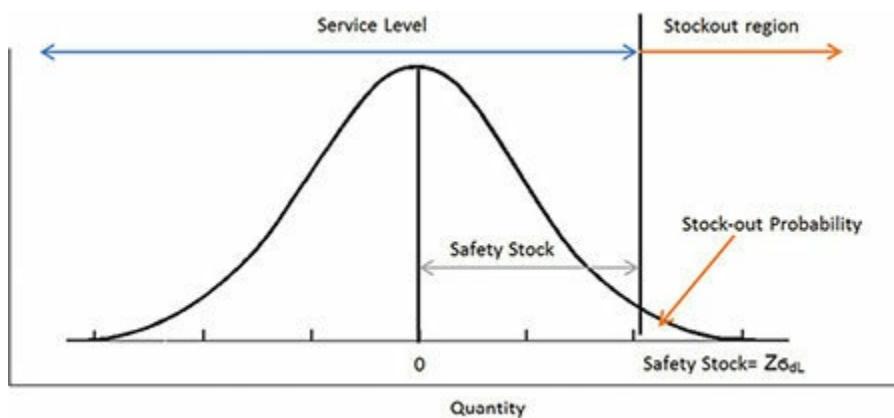
Since there is variability in the demand or lead time we have to add certain inventory to meet the unexpected demand surge. This is called safety stock. The amount of the safety stock is directly proportional to the service level of the inventory.

Now  $ROP = \text{Expected demand during lead time} + \text{Safety Stock} = DL + S$

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 or 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.

$$ROP = DL + Z\sigma_{dL}$$

$\sigma_{dL}$  = 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

$$ROP = D \cdot L + z \cdot d \cdot \sigma_L$$

$\sigma_L$  = standard deviation of the lead time in days

If both demand and lead time is variable then

$$ROP = D \cdot L + z^* \sqrt{(L^* \sigma_d^2 + d^2 \sigma_L^2)}$$

In this the demand and lead time are assumed 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

$$ROP = 100 \cdot 5 + 2.33 \cdot \sqrt{(5 \cdot 10^2 + 100^2 \cdot 2^2)} = 500 + 201$$

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 subsides 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 the case for many items.

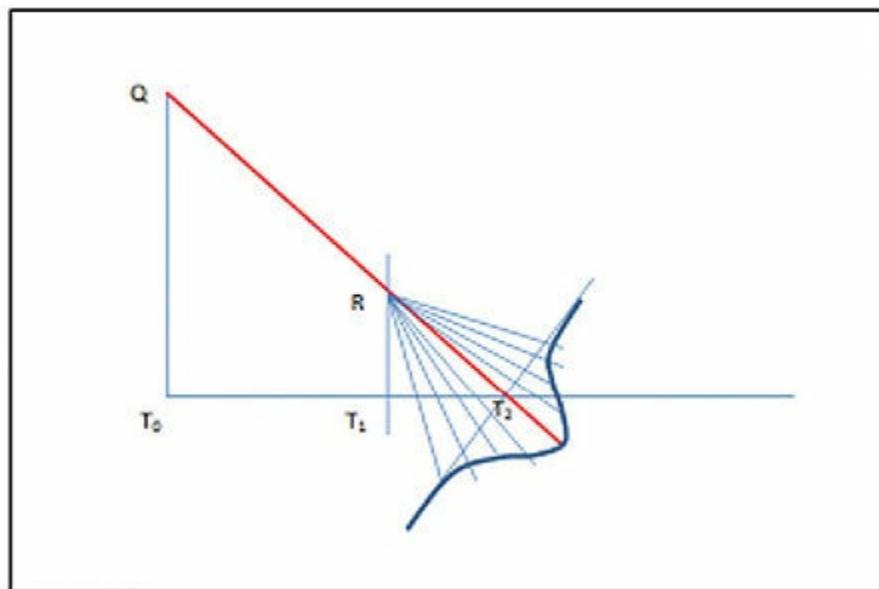
The idea of this chapter was to provide basic understanding of the inventory and the service level. The inventory management is a vast subject in itself. For more depth knowledge readers are encouraged to read books related to operation and inventory.

### **Example of Customer Service Level**

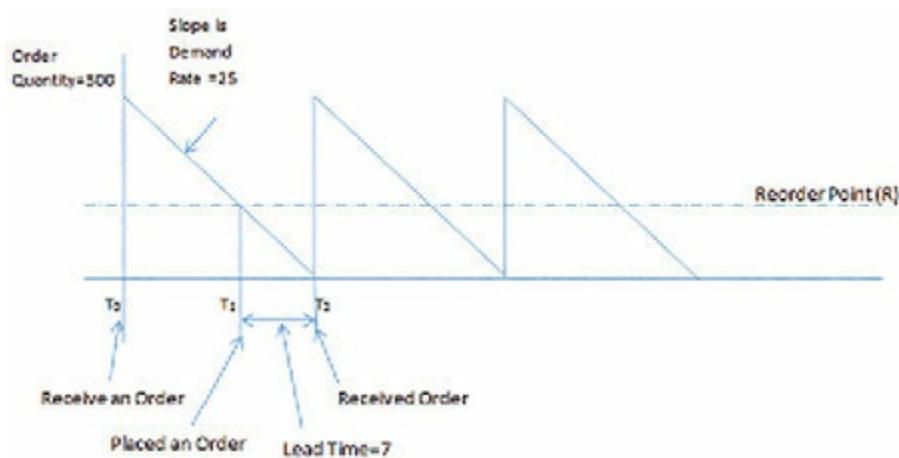
You are running a boutique in one small town. Most of the vendors are from bigger cities. One of the products which are the best-selling for your boutique has high demand variability. Sometimes your stock goes out of stock before the order from vendor reaches your location. As the product is high selling 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 points 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.



$$ROP = DL + Z\sigma_{dL}$$

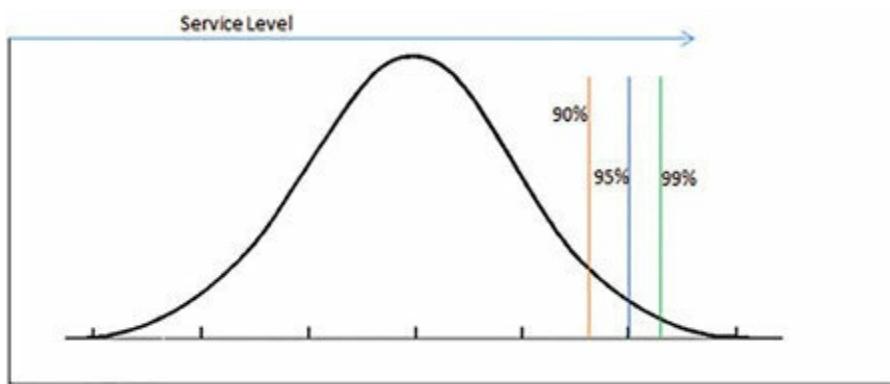
As you know Z for 99% is 2.58, you have to find the standard deviation of the mean. Here again you use Central Limit Theorem to find the standard deviation of the mean demand.

$$\text{Standard deviation} = \text{Sample Standard Deviation} / \sqrt{\text{Sample Size}} = 30 / \sqrt{40} = 4.74$$

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**

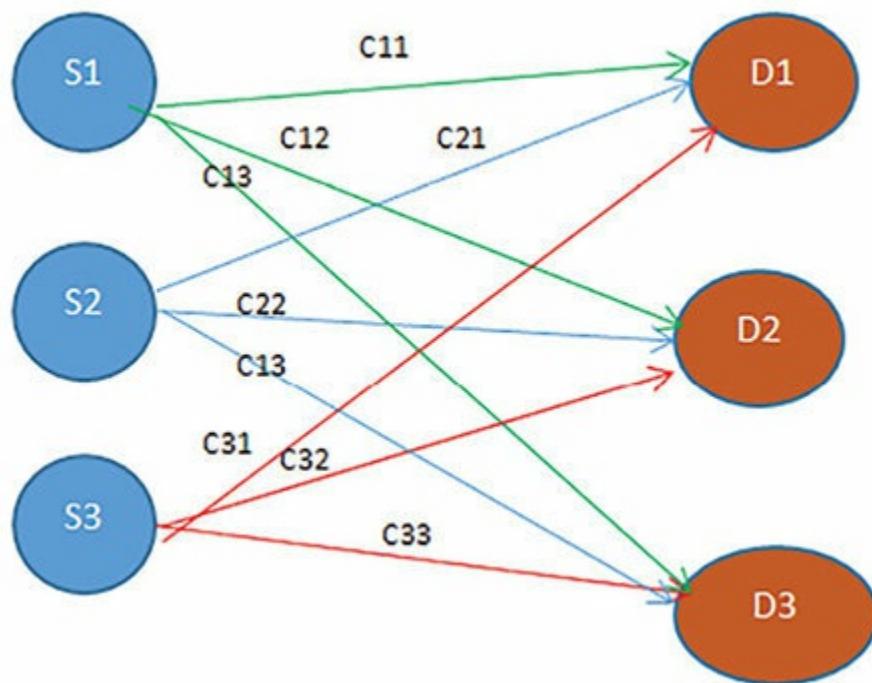
- Understanding of basic inventory replenishment model
- Various cost involved in the inventory decision point
- Economic Order Quantity (EOQ) model explanation and formulation
- Inventory replenishment under variable demand
- Concept and service level, safety Stock and Reorder point
- How to achieve 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  $T_{\text{cost}} = \sum C_{ij}X_{ij}$  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
	$X_{11}$	$X_{12}$	$X_{13}$	$X_{14}$	
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
	$X_{31}$	$X_{32}$	$X_{33}$	$X_{34}$	
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  $WH_1$  is exhausted, now move to next supply  $WH_2$  that is row 2. As demand for  $ST_1$  is exhausted, we start with  $ST_2$ . In  $ST_2$  we have unmet demand of 15 units, so we allocate  $X_{22}=15$  units. Demand for  $ST_2$  is exhausted; we move to demand of  $ST_3$ .  $WH_2$  still has 25 unallocated units, hence we allocate  $X_{33}=20$  units. The demand for  $ST_3$  is also exhausted; we move to demand for  $ST_4$ .  $WH_2$  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.

Let's assign cost to each pair of the Warehouse and Stores.

Warehouse/ Stores	ST1	ST2	ST3	ST4	Total Ware- house
WH1	$C_{11}=10$	$C_{12}=5$	$C_{13}=9$	$C_{14}=13$	40
	$X_{11}=30$	$X_{12}=10$	$X_{13}=0$	$X_{14}=0$	
WH2	$C_{21}=4$	$C_{22}=12$	$C_{23}=6$	$C_{24}=10$	45
	$X_{21}=0$	$X_{22}=15$	$X_{23}=20$	$X_{24}=10$	
WH3	$C_{31}=8$	$C_{32}=3$	$C_{33}=7$	$C_{34}=11$	25
	$X_{31}=0$	$X_{32}=0$	$X_{33}=0$	$X_{34}=25$	
Total Store	30	25	20	35	110

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	$C_{11}=10$ $X_{11}=30$ $D_{11}=$	$C_{12}=5$ $X_{12}=10$ $D_{12}=$	$C_{13}=9$ $X_{13}=0$ $D_{13}=$	$C_{14}=13$ $X_{14}=0$ $D_{14}=$	40	$u_1=0$
WH2	$C_{21}=4$ $X_{21}=0$ $D_{21}=$	$C_{22}=12$ $X_{22}=15$ $D_{22}=$	$C_{23}=6$ $X_{23}=20$ $D_{23}=$	$C_{24}=10$ $X_{24}=10$ $D_{24}=$	45	$U_2=7$
WH3	$C_{31}=8$ $X_{31}=0$ $D_{31}=$	$C_{32}=3$ $X_{32}=0$ $D_{32}=$	$C_{33}=7$ $X_{33}=0$ $D_{33}=$	$C_{34}=11$ $X_{34}=25$ $D_{34}=$	25	$U_3=8$
Total Store	30	25	20	35	110	
v	$v_1=10$	$v_2=5$	$v_3=-1$	$v_4=3$		

In the table above, set  $u_1=0$  as the initial value. For all the value where  $X_{1j}>0$  set  $v_j = C_{1j}-u_1$ .

$$v_1 = C_{11} - u_1 = 10 - 0 = 10$$

$$v_2 = C_{21} - u_1 = 5 - 0 = 5$$

as there are not more  $X_{1j}>0$ , we move to second row where  $X_{2j}>0$  set  $u_2=C_{2j}-v_{2j}$

$$u_2 = C_{22} - v_2 = 12 - 5 = 7$$

Similarly we calculate other value of v and u.

$$v_3 = C_{23} - u_2 = 6 - 7 = -1$$

$$v_4 = C_{24} - u_2 = 10 - 7 = 3$$

$$u_3 = C_{34} - v_4 = 11 - 3 = 8$$

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 <sub>11</sub> =10 X <sub>11</sub> =30 S <sub>11</sub> =10	C <sub>12</sub> =5 X <sub>12</sub> =10 S <sub>12</sub> =5	C <sub>13</sub> =9 X <sub>13</sub> =0 S <sub>13</sub> =-1	C <sub>14</sub> =13 X <sub>14</sub> =0 S <sub>14</sub> =3	40	u <sub>1</sub> =0
WH2	C <sub>21</sub> =4 X <sub>21</sub> =0 S <sub>21</sub> =17	C <sub>22</sub> =12 X <sub>22</sub> =15 S <sub>22</sub> =12	C <sub>23</sub> =6 X <sub>23</sub> =20 S <sub>23</sub> =6	C <sub>24</sub> =10 X <sub>24</sub> =10 S <sub>24</sub> =10	45	U <sub>2</sub> =7
WH3	C <sub>31</sub> =8 X <sub>31</sub> =0 S <sub>31</sub> =18	C <sub>32</sub> =3 X <sub>32</sub> =0 S <sub>32</sub> =13	C <sub>33</sub> =7 X <sub>33</sub> =0 S <sub>33</sub> =7	C <sub>34</sub> =11 X <sub>34</sub> =25 S <sub>34</sub> =11	25	U <sub>3</sub> =8
Total Store	30	25	20	35	110	
v	v <sub>1</sub> =10	v <sub>2</sub> =5	v <sub>3</sub> =-1	v <sub>4</sub> =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/ Stores	ST1	ST2	ST3	ST4	Total Ware-house	u
WH1	$C_{11}=10$ $X_{11}=0 (-30)$ $S_{11}=10$	$C_{12}=5$ $X_{12}=25 (+15)$ $S_{12}=5$	$C_{13}=9$ $X_{13}=15(+15)$ $S_{13}=-1$	$C_{14}=13$ $X_{14}=0$ $S_{14}=3$	40	$u_1=0$
WH2	$C_{21}=4$ $X_{21}=30(+30)$ $S_{21}=17$	$C_{22}=12$ $X_{22}=0(-15)$ $S_{22}=12$	$C_{23}=6$ $X_{23}=5(-15)$ $S_{23}=6$	$C_{24}=10$ $X_{24}=10$ $S_{24}=10$	45	$U_2=7$
WH3	$C_{31}=8$ $X_{31}=0$ $S_{31}=18$	$C_{32}=3$ $X_{32}=0$ $S_{32}=13$	$C_{33}=7$ $X_{33}=0$ $S_{33}=7$	$C_{34}=11$ $X_{34}=25$ $S_{34}=11$	25	$U_3=8$
Total Store	30	25	20	35	110	
v	$v_1=10$	$v_2=5$	$v_3=-1$	$v_4=3$		

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	$C_{21}=4$ $X_{21}=30$ $S_{21}=4$	$C_{22}=12$ $X_{22}=0$ $S_{22}=2$	$C_{23}=6$ $X_{23}=5$ $S_{23}=6$	$C_{24}=10$ $X_{24}=10$ $S_{24}=10$	45	$U_2=-3$
WH3	$C_{31}=8$ $X_{31}=0$ $S_{31}=5$	$C_{32}=3$ $X_{32}=0$ $S_{32}=3$	$C_{33}=7$ $X_{33}=0$ $S_{33}=7$	$C_{34}=11$ $X_{34}=25$ $S_{34}=11$	25	$U_3=-2$
Total Store	30	25	20	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

$$\text{Min } Z = \sum X_{ij} C_{ij} \text{ Total Transportation cost}$$

Subject to

$$X_{1j} \leq WH_1 \quad \text{warehouse1 capacity constraint}$$

$$X_{2j} \leq WH_2 \quad \text{warehouse2 capacity constraint}$$

$$X_{3j} \leq WH_3 \quad \text{warehouse3 capacity constraint}$$

$$X_{il} \leq ST_1 \quad \text{Store1 demand constraint}$$

$$X_{i2} \leq ST_2 \quad \text{Store2 demand constraint}$$

$$X_{i3} \leq ST_3 \quad \text{Store3 demand constraint}$$

$$X_{ij} \geq 0 \text{ and is integer}$$

Adding value in equations

$$\text{Min } Z = X_{11} * 10 + X_{12} * 5 + X_{13} * 9 + X_{14} * 13 + X_{21} * 4 + X_{22} * 12 + X_{23} * 6 + X_{24} * 10 + X_{31} * 8 + X_{32} * 3 + X_{33} * 7 + X_{34} * 11$$

$$\text{Subject to } X_{11} + X_{12} + X_{13} + X_{14} \leq 40$$

$$X_{21} + X_{22} + X_{23} + X_{24} \leq 45$$

$$X_{31} + X_{32} + X_{33} + X_{34} \leq 25$$

$$X_{11} + X_{21} + X_{31} \leq 30$$

$$X_{21} + X_{22} + X_{23} \leq 25$$

$$X_{31} + X_{32} + X_{33} \leq 20$$

$$X_{41} + X_{42} + X_{43} \leq 35$$

$$X_{ij} \geq 0 \text{ and is integer}$$

The problem can be constructed in excel as shown below. Total cost is nothing but sumproduct of variables X and cost.

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.

In Solver select **Min** and select **D1** cell as objective function.

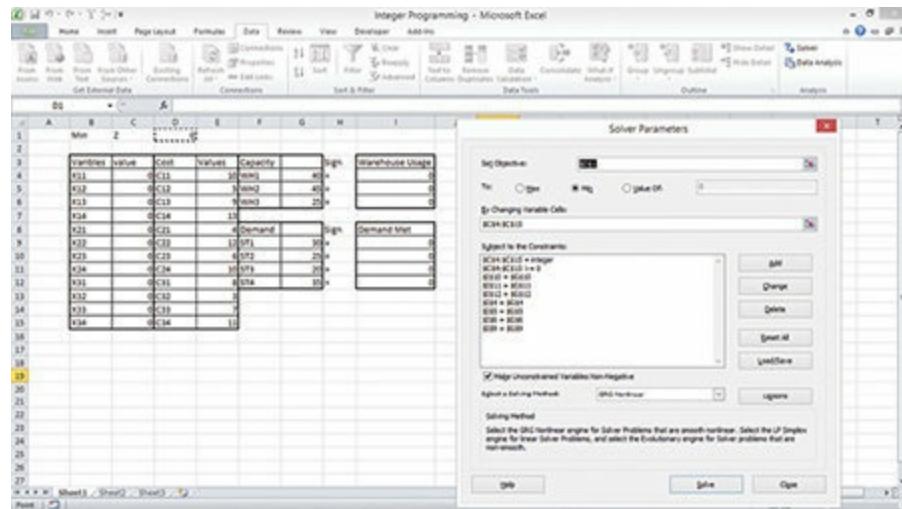
The screenshot shows the Microsoft Excel ribbon at the top with tabs like Home, Insert, Page Layout, Formulas, Data, Review, View, Developer, and Add-ins. The Data tab is selected. On the far right of the ribbon, there are icons for Data Tools (Advanced, What-If Analysis, Data Validation, Group, Subtotal, Outline, and Analysis) and Data Analysis (Data Analysis add-in).

The main area of the screen displays a spreadsheet titled "Integer Programming - Microsoft Excel". The spreadsheet contains data for variables, costs, values, capacity, and warehouse usage across multiple rows and columns.

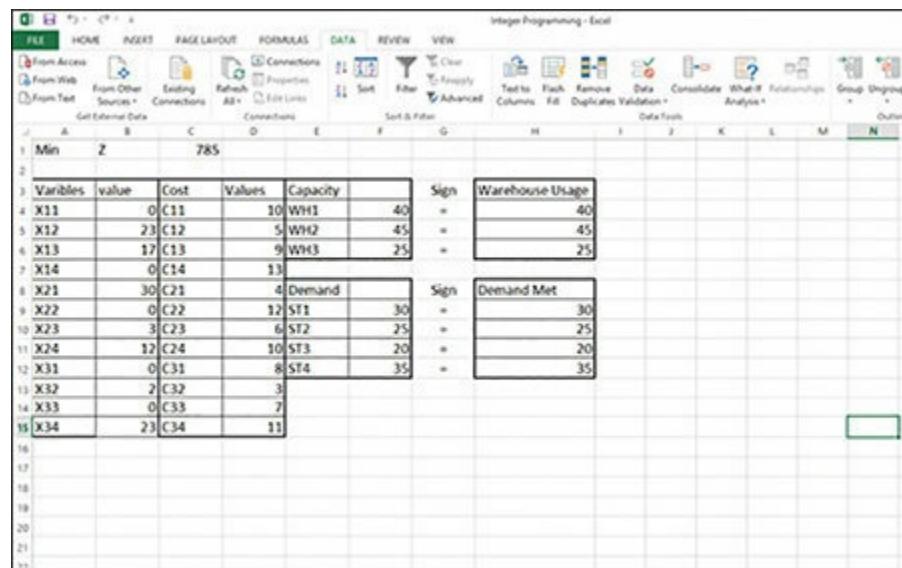
On the right side of the screen, the "Solver Parameters" dialog box is open. It includes fields for "Set Objective" (B10), "To" (Max), "By Changing Variable Cells" (B10:B14), and "Subject to the Constraints". There are also buttons for "Add", "Change", "Delete", "Save", and "Load/Save".

At the bottom of the dialog box, there is a section for "Solving Method" with the note: "Select the GRG Nonlinear engine for LP Problems that are smooth nonlinear. Select the LP Simplex engine for linear Problems, and select the Evolutionary engine for Solver problems that are non-smooth."

Add constraints related to Warehouse and Stores. Additionally add integer constraint and non-negative constraints.



On solving the D1 value is 785. Hence optimum transportation cost is 785. This number is same as we got in the previous section.



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
Prisoner A	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.

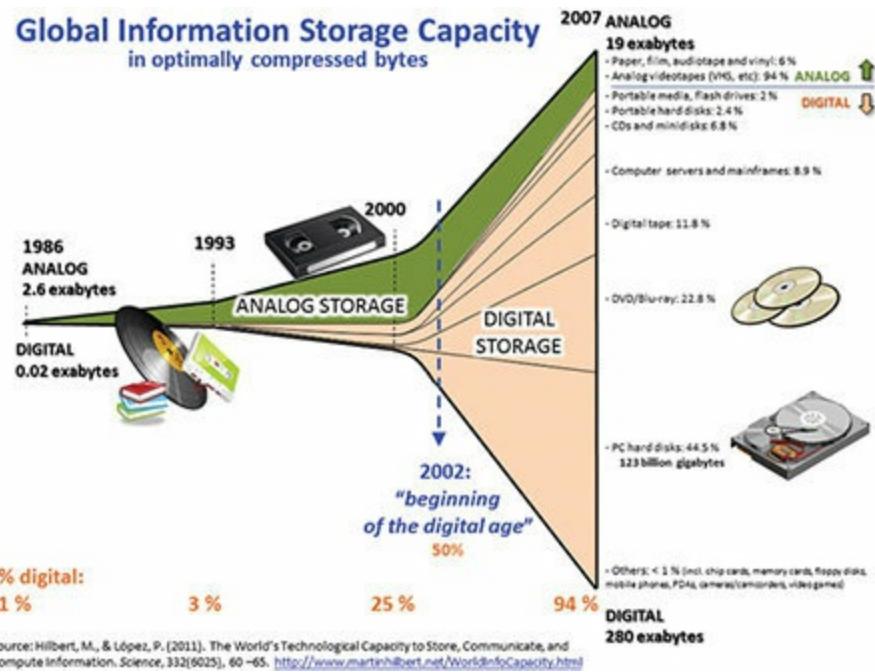
**Variety:** The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

**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.

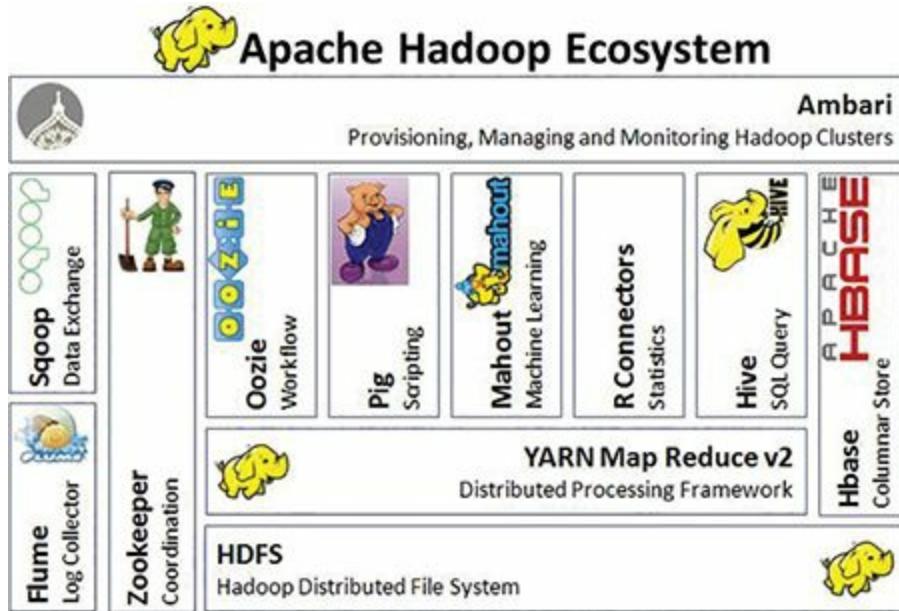
The growth of Digital data in the last two 3 decades as seen from the picture below show that the explosion of the data started somewhere in 2002 and its expanding at exponential phase.



Every organization collects digital data at each touch 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 discuss 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



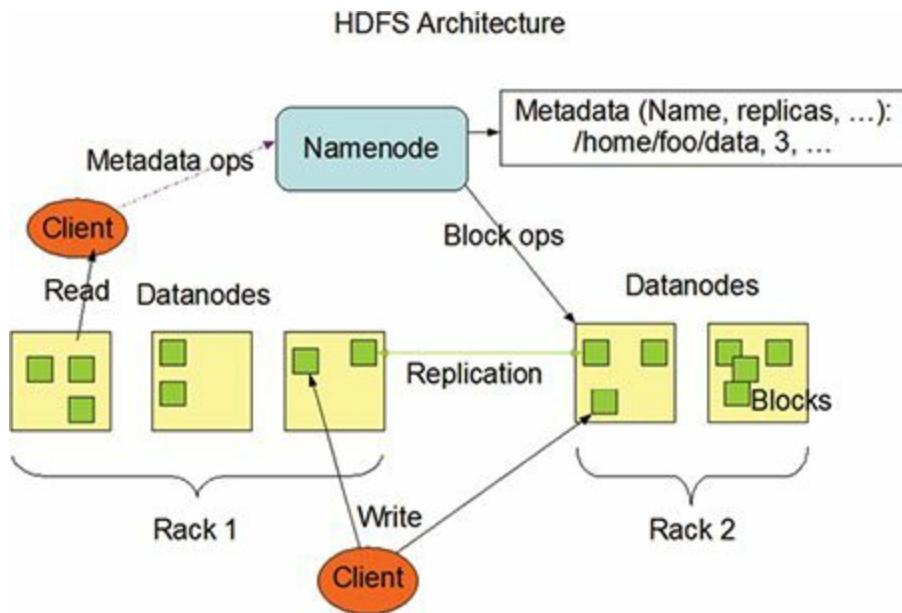
Important components of Hadoop Ecosystems are:

## 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](mailto: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.

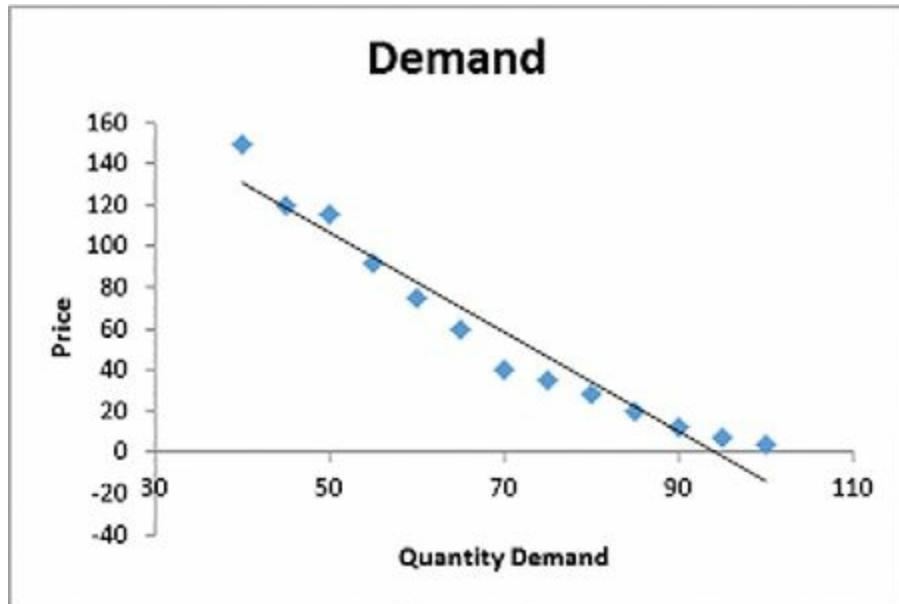
## 8.4 Pricing

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

55	18
50	13
45	7
40	3

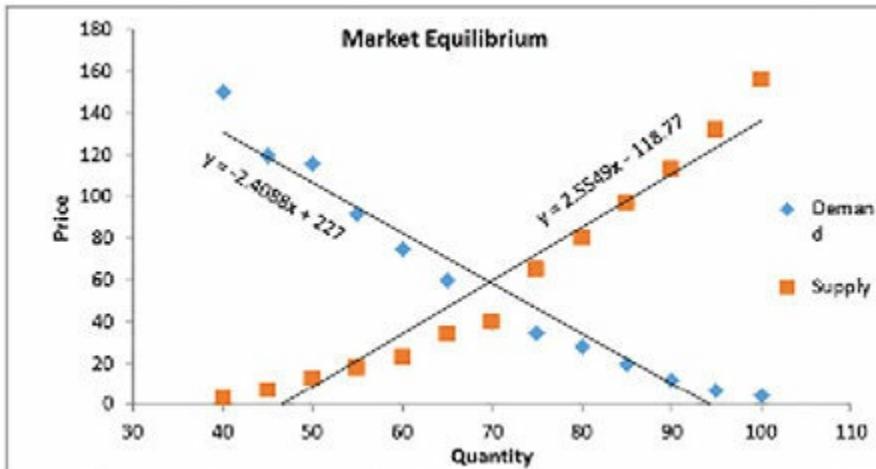


As you can see from above scatter plot the seller is willing to sell more and more quantity as the auction price is increases. The line is not straight but if we plot average of all the seller in the market then this line will be straight.

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 =  $227 - 2.4x$

Supply Equation =  $-118.77 + 2.55x$

Both equation will intersect at Demand = Supply

$$227 - 2.4x = -118.77 + 2.55x$$

$$2.55x + 2.4x = 227 + 118.77$$

$$5x = 346$$

$$X = 346/5 = 69$$

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 13<sup>th</sup> 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 13<sup>th</sup> 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. In 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.
2. **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
<b>Chapter - I : DATA ANALYTICS FOUNDATION</b>	<b>12</b>
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
<b>Chapter - II : ANALYTICS SYSTEM</b>	<b>92</b>
Section - I : BUSINESS INTELLIGENCE SYSTEM	95
Section - II : R BASICS	153
<b>Chapter - III : WEB ANALYTICS</b>	<b>168</b>
Section - I : GOOGLE ANALYTICS	170
<b>Chapter - IV : CUSTOMER ANALYTICS</b>	<b>200</b>
Section - I : CUSTOMER ANALYTICS	203
<b>Chapter - V : DIGITAL MARKETING</b>	<b>240</b>
Section - I : DIGITAL MARKETING BASIC	242
Section- II : DIGITAL CHANNEL OPTIMIZATION	262
<b>Chapter - VI : FORECASTING AND PREDICTION</b>	<b>280</b>
Section - I : REGRESSION	282
Section - II : TIME SERIES FORECASTING	295
<b>Chapter - VII : INVENTORY MANAGEMENT</b>	<b>313</b>
Section - I : INVENTORY MODEL	315
Section -II : TRANSPORTATION PROBLEM	323
<b>Chapter - VIII : ADVANCED TOPICS</b>	<b>333</b>