Given below are pre-requisites for participants attending R workshop:-

Here is the detail (including system requirement)

1. Participants should bring their laptop (preferably Windows 7 or higher/ Mac OS installed).
2. Participants should have latest version of R and Rstudio installed on their system.
3. First Install R and then R Studio. Latest version of both softwares can be found at:

   - R Can be downloaded from: https://cran.r-project.org/
   - RStudio can be downloaded from: https://www.rstudio.com/products/rstudio/download

Moreover:

- Participants should have basic programming skills and should be in able to understand the scripting language.
- High speed internet connection will be provided to participants during the training hours at IIMB.

System Requirements:

- OS: Mac OS X (any machine built since 2008 but not before that) or Linux or Windows (Version XP or later) is required.
- The Installation will approximately consume 150 MB of disk space.
- Minimum 1 GB RAM on the system is preferable.

== End of instructions ==

Venue, Date & Workshop timings:-

C-11	**On Right side after IIMB entrance and behind Auditorium**
Dates	**15th & 16th December 2015**
8:30 to 09:00	Fulfill software upload and other requirements
9:00 to 10:15	Session 1
10:30 to 11:45	Session 2
12:00 to 1:15	Session 3
1:15 to 2:15	Lunch @ MDC
2:15 to 3:30	Session 4
3:45 to 5:00	Session 5

For any more queries and clarifications, send an email to rahul235@gmail.com

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Introduction to Business Analytics

Overview

Objective

After completing this lesson you will be able to:

- Describe business analytics
- Explain the components of business analytics
- Explain the usage of business analytics in various domains

In God we trust, all other must bring data

- W Edward Deming

Highest Paid Person's Opinion

Business Analytics–Definition

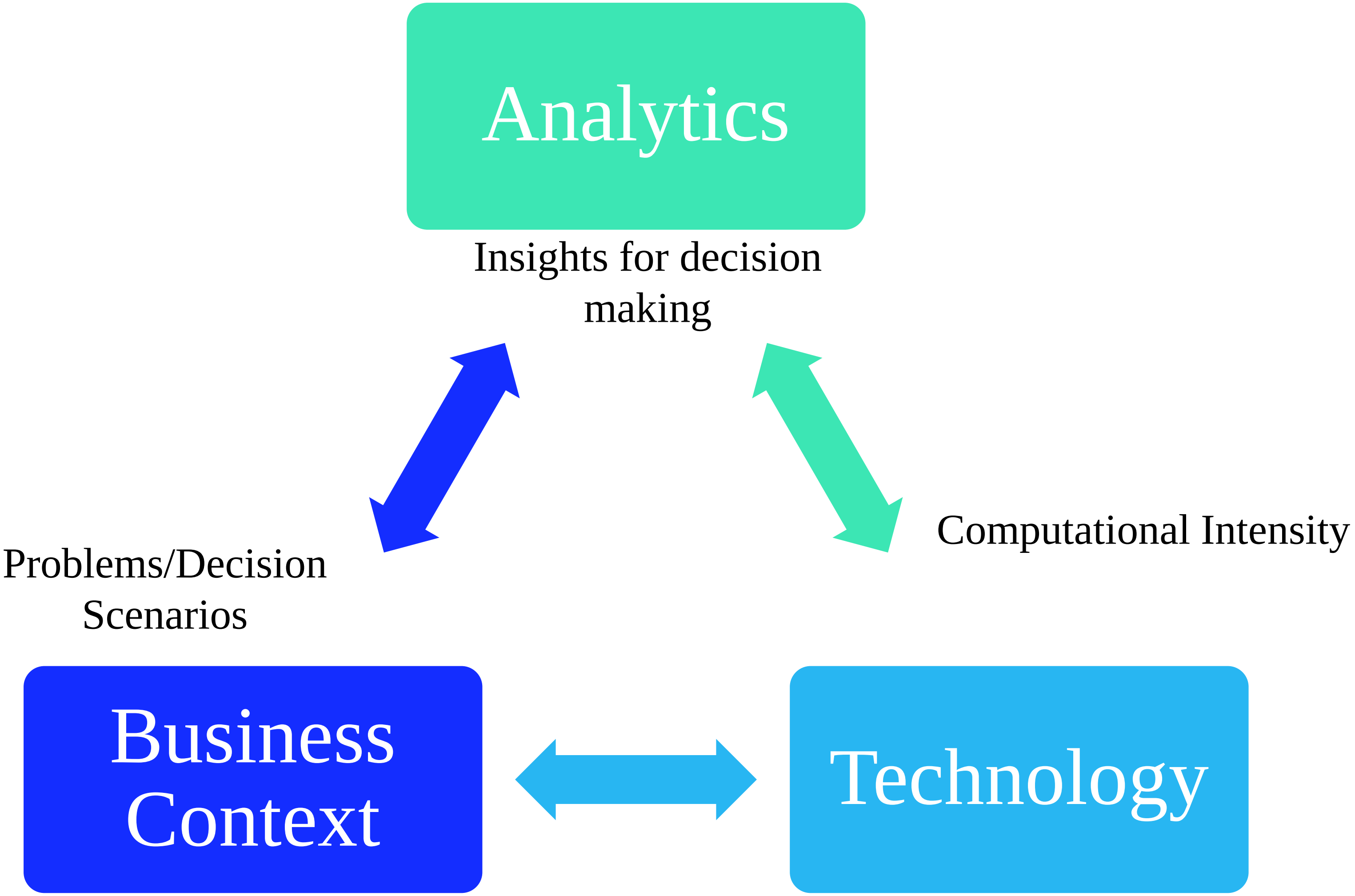
- Business analytics (BA) refers to the tools, techniques and processes for continuous exploration and investigation of past data to gain insights and help in decision making.

- Business Analytics is an integration between science, technology and business context that assist data driven decision making.

Data Explosion

- About seven billion shares change hand in US equity markets everyday.
- About 10 billion photos are uploaded every hour in the facebook.
- Amount of credit card debt in US: $793.1 billion.
- Total amount of credit card fraud worldwide: $5.5 billion.
- Percentage of US credit card holders who have been victims of credit card fraud: 10%
- Every week, about 100 million customers visit Walmart stores.

Analytics Trilogy



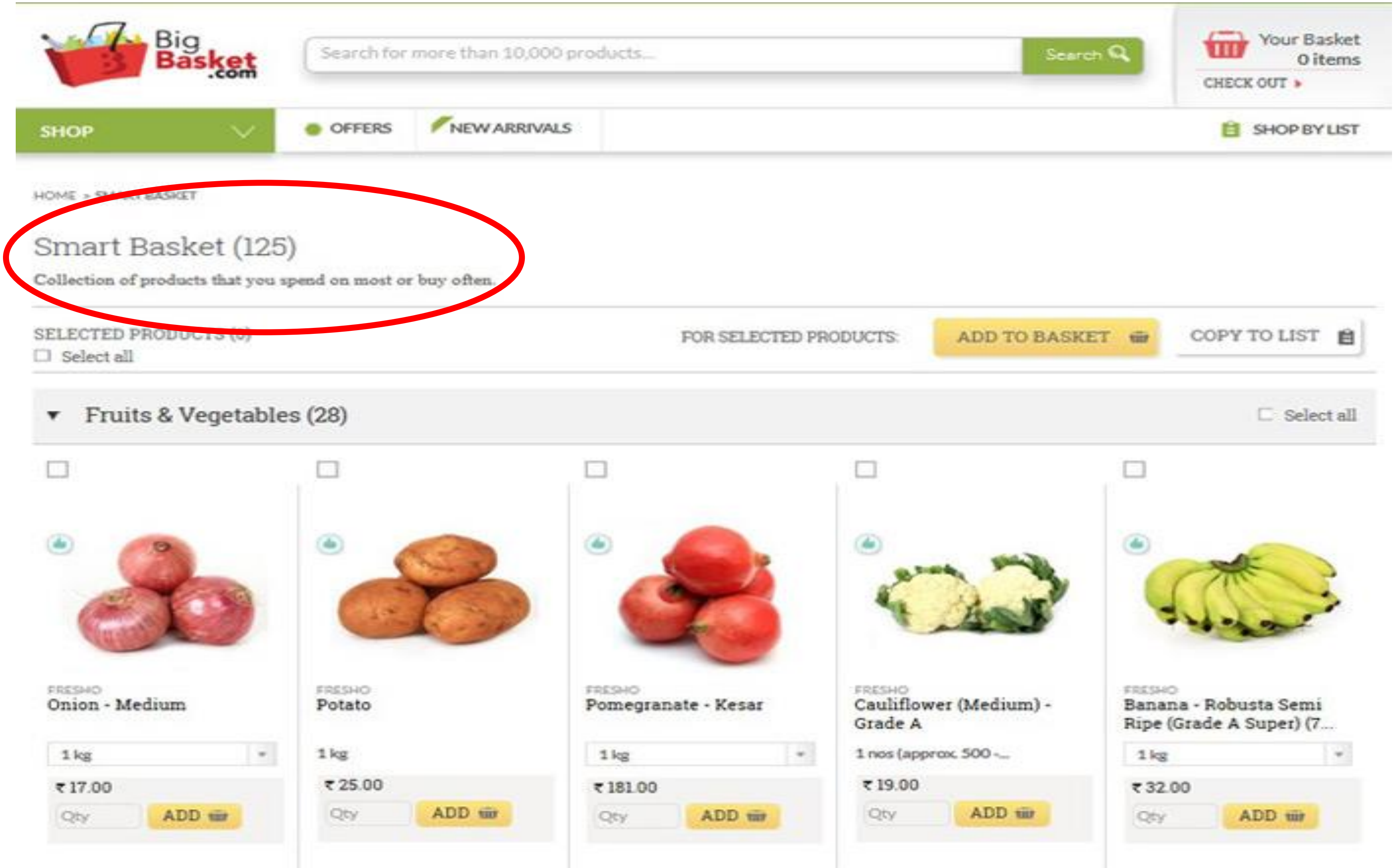
Why Analytics

- Analytics provides competitive advantage.
- Analytics removes inefficiency in the system/organization.
- Provides ability to make better decisions.

Analytics in Use–Flipkart

- Forecast demand for each SKU.
- Predict customer cancellations and returns.
- Predict customer contacts at the customer service.
- Predict what a customer is likely to purchase in the future?
- How to optimize the delivery system?

Analytics in Use–Big Basket



How would you solve this?

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Trak.in
Business of Tech, Mobile & Startups in India

HOME | BUSINESS | TECHNOLOGY | INTERNET | TELECOM | MOBILE | STARTUP | OTHERS | ABOUT

Home / Internet / Ecommerce / Flipkart delivers 2 stones instead of iPod to a user! What if…

Flipkart delivers 2 stones instead of iPod to a user! What if…

Posted by: Arun Prabhudesai | In Ecommerce, India | March 11, 2013 | 46 Comments

Now, this is outrageous – A twitter user tweeted today that his sister had ordered an iPod, but to her surprise she was delivered 2 stones inside the box.

Twitter user with handle @nikhilsekhar tweeted the following accompanying picture of the said delivery.

> @flipkart My sister got two stones instead of the INR 20K iPod that she ordered from Flipkart. What is wrong with you twitter.com/nikhilsekhar/s…
>
> — Nikhil (@nikhilsekhar) March 10, 2013

Here is the photo accompanying that tweet:

The good part is Flipkart promptly replied the user with following tweets:

> We are very sorry about the incident with our customer's iPod purchase, we're taking up this issue extremely seriously. 1/3
>
> — Flipkart (@Flipkart) March 10, 2013

> We have spoken to our customer and will make sure a replacement is sent over right away. 3/3
>
> — Flipkart (@Flipkart) March 10, 2013

But what if…

Now, this kind of a thing happening is outrageous – and it is good to see that Flipkart has owned the responsibility and is going to offer an immediate replacement of the product.

However, I have a question in mind – What if the customer is lying?

Please bear in mind, I am not talking about this incidence, but putting across a theoretical situation.

What if a buyer actually replaces the delivery (after he has received the correct product) and then alleges that Flipkart has delivered him with stones (or whatever) inside it? What happens in that case. If you think that such kind of thing will not happen….you are wrong. There are many out there waiting to take advantage of the system..

Flipkart is in no position to actually contend that they have delivered it correctly, neither can they challenge the customer (and if they do, it will be one big social media mess).

OH…IS IT?

FOLLOW US

45.6k Follows

RSS Feed	23.8k Followers	Facebook	16.4k Followers
Twitter	3.4k Followers	LinkedIn	0 Followers
Pinterest	775 Followers	Google+	1k Followers

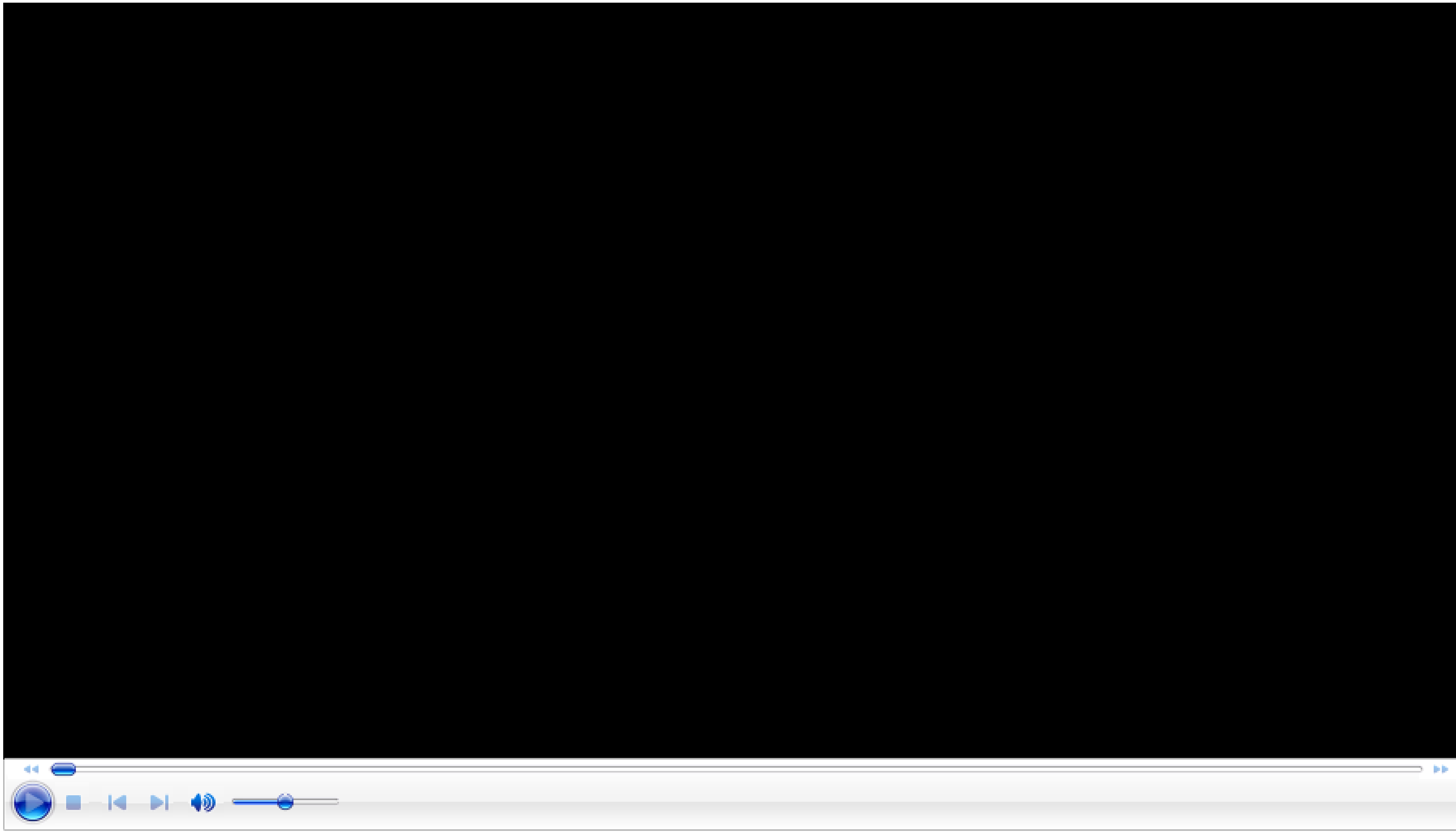
FIND US ON FACEBOOK

http://trak.in/tags/business/2013/03/11/flipkart-delivers-stones-instead-ipod/

Decision Making–The Monty Hall Problem

After having seen “What lies besides door 1”, Would you like to switch?

Decision Making–Beer Launcher



If you are a venture capitalist, will you fund this?

https://www.youtube.com/watch?v=bKBGbCVBv8M

The Game Changers

- Google
  - Used Markov chains to rank pages.
- Proctor and Gamble
  - Analytics as competitive strategy.
- Target
  - Predicts customer pregnancy.
- Capital One
  - Identifies the most profitable customer.
- Hewlett Packard
  - Developed “flight risk score” for 3,30,000 employees.
- Obama’s 2012 presidential campaign.
  - Persuasion Modelling.

The Innovators

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

- OKCupid: Predicts which online dating messages is most likely to get a response!
- Polyphonic HMI: Uses “hit song science” to predict commercial success of a song.
- Netflix: Predicts movie ratings by customers (RMSE is 1%).
- Amazon.com: 35% of sales come from product recommendations.
- Citizens Bank: Predicted fraudulent cheques.
- Divorce360.com: Predicting success of a marriage!

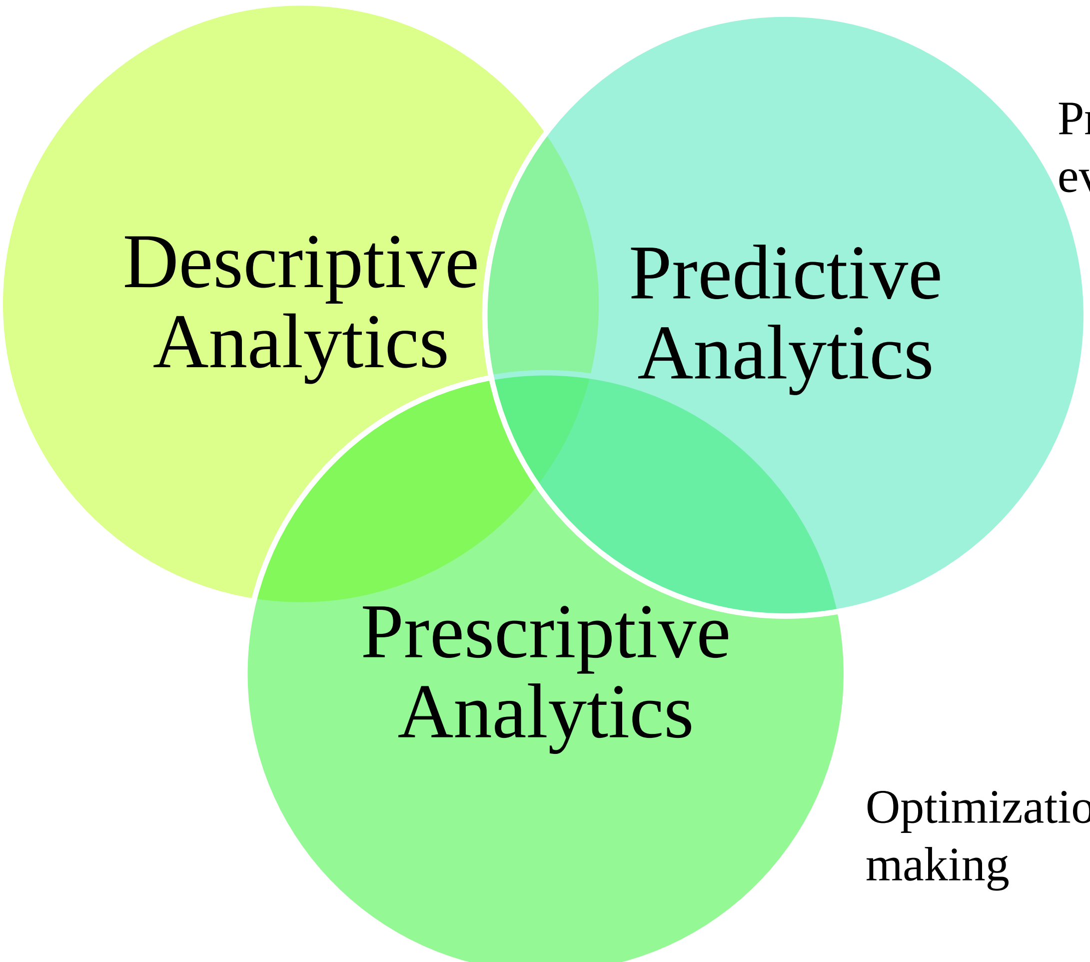
Data Scientists will be the sexiest job of 21st century

Harvard Business Review 2012

Components of Business Analytics

Data synthesis and Visualization

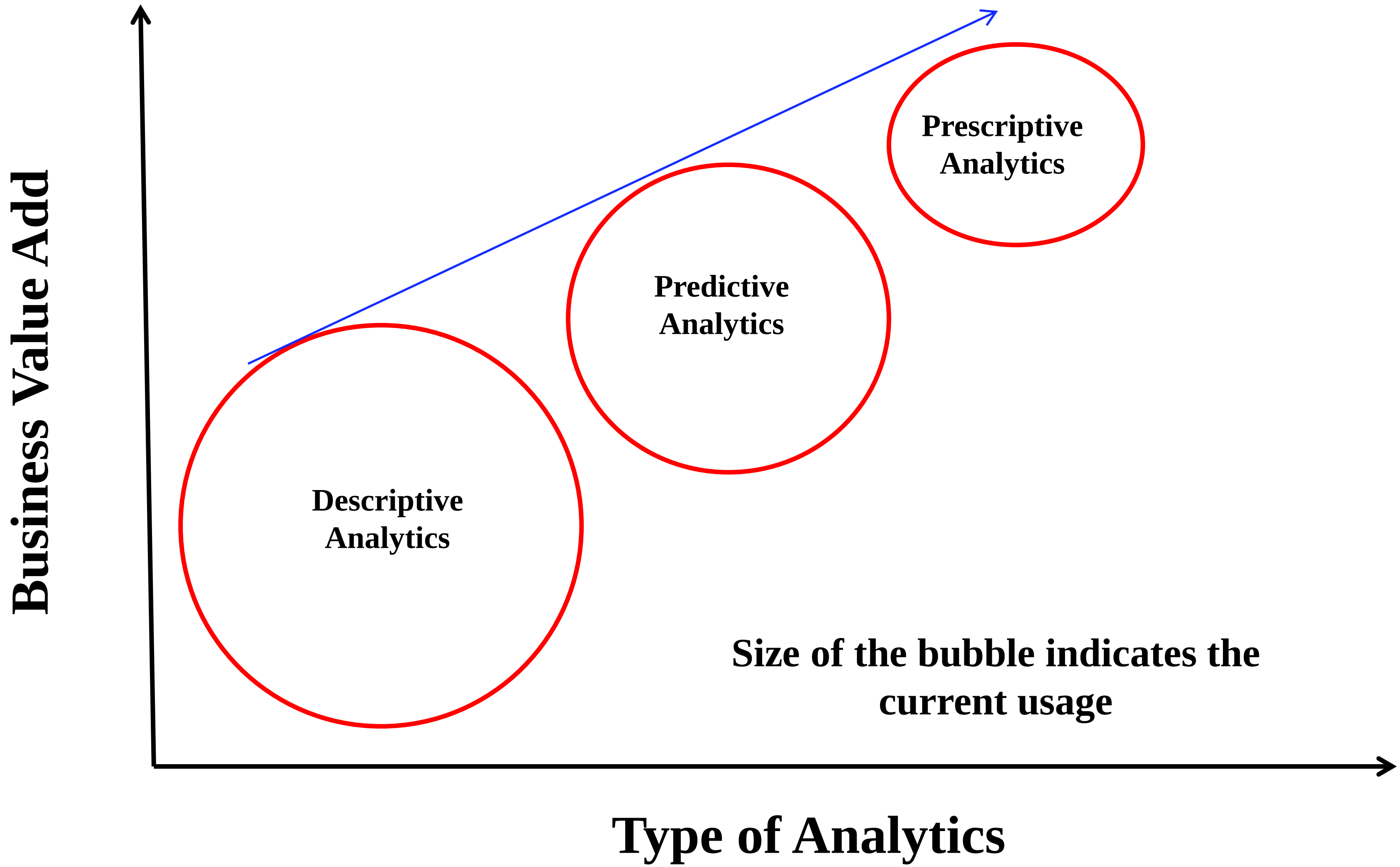
Predicting future events



Optimization and decision making

Components of Business Analytics

Descriptive analytics	• Communicates the hidden facts and trends in the data • Simple analysis of data can lead to business practices that result in financial rewards • Helps organizations uncover inefficiencies and eliminate them
Predictive analytics	• Predicts the probability of occurrence of a future event • Helps organizations to plan their future course of action • Most frequently used type of analytics across several industries
Prescriptive analytics	• Assists users in finding the optimal solution to a problem • In most cases, provides an optimal solution/decision to the problem • Inventory management is one of the problems that are most frequently addressed


Business Value Add
Prescriptive Analytics
Predictive Analytics
Descriptive Analytics
Size of the bubble indicates the current usage
Type of Analytics

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Power of Descriptive Analytics

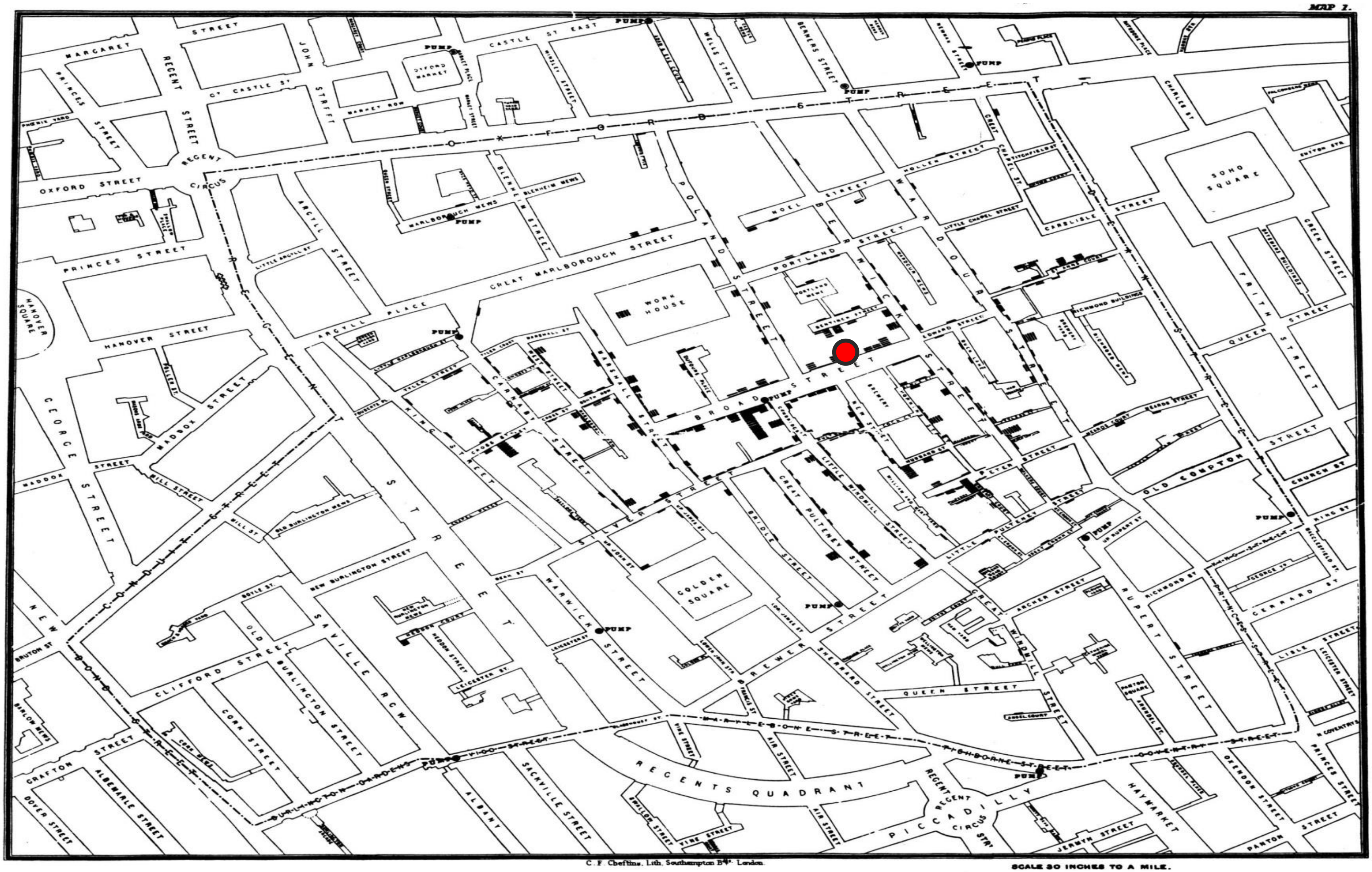
London Cholera Outbreak–1854

- Severe outbreak of cholera that occurred near Broad Street (now Broadwick street) in Soho district of London in 1854.

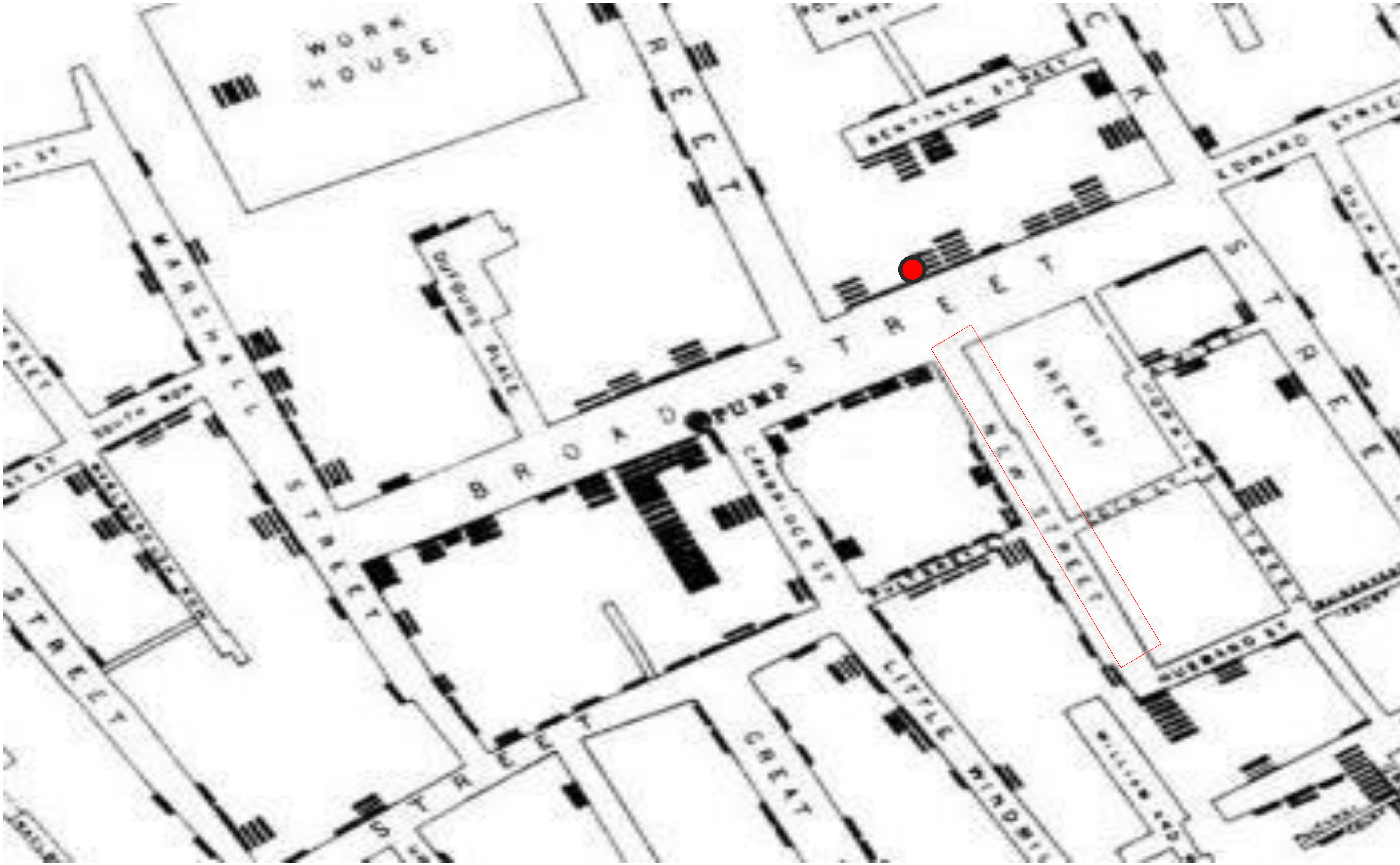
- More than 500 people died within 10 days of the outbreak, the mortality rate in some parts of the city was as high as 12.8%.

https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

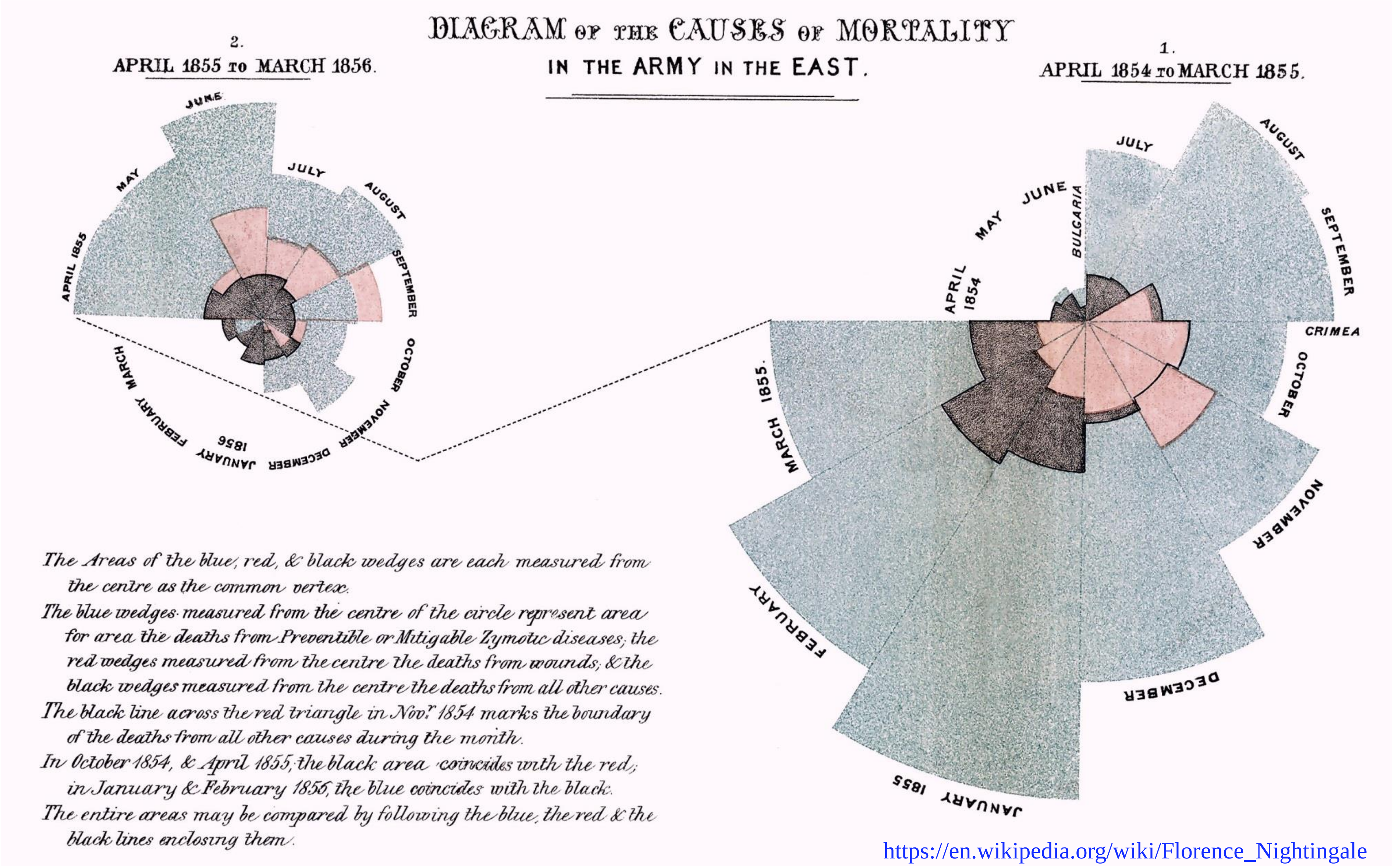
Visual Map of the Area



Visual Map of the Street

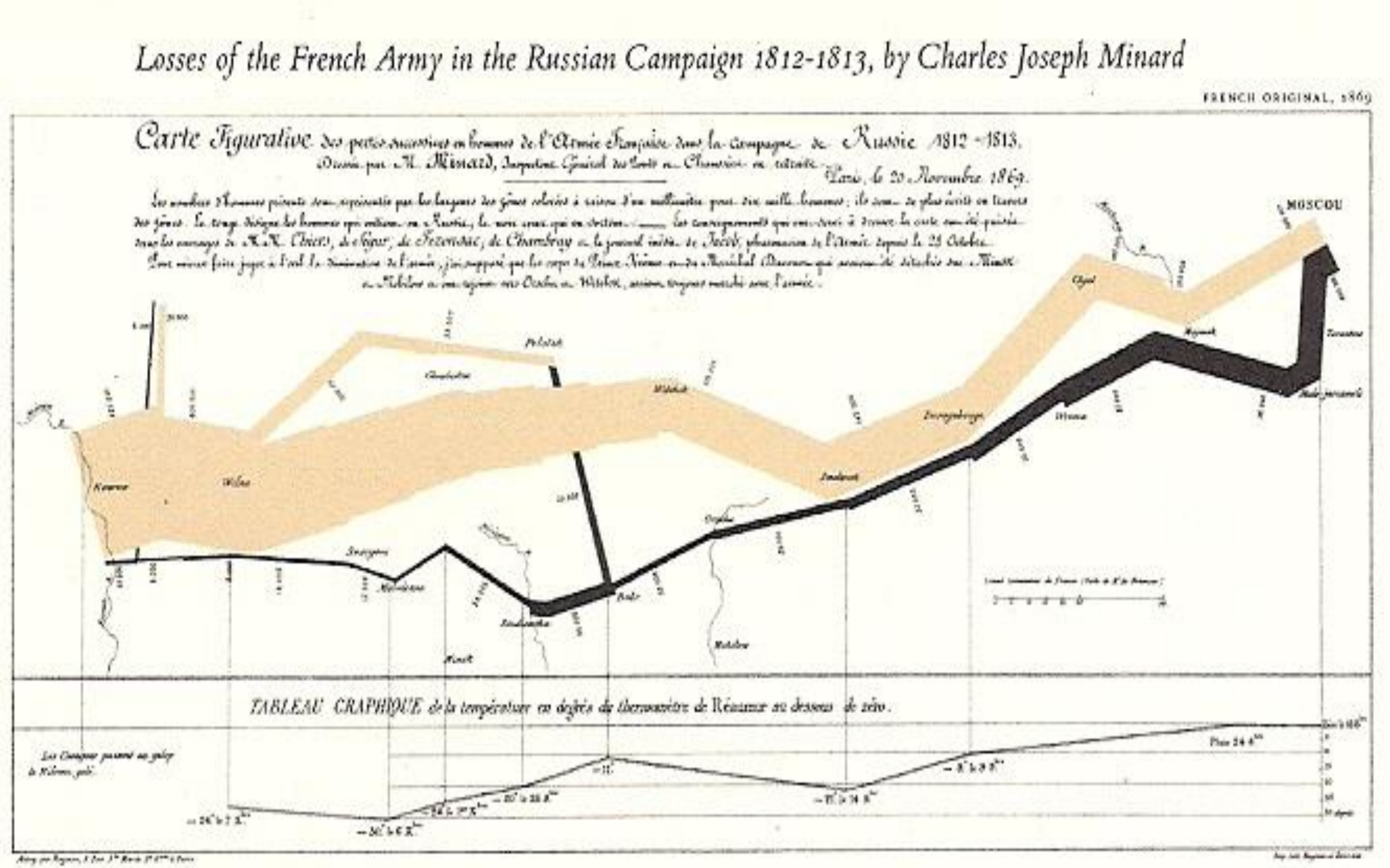


Florence Nightingale's Pie Chart– Crimean War Data (1853-1856)



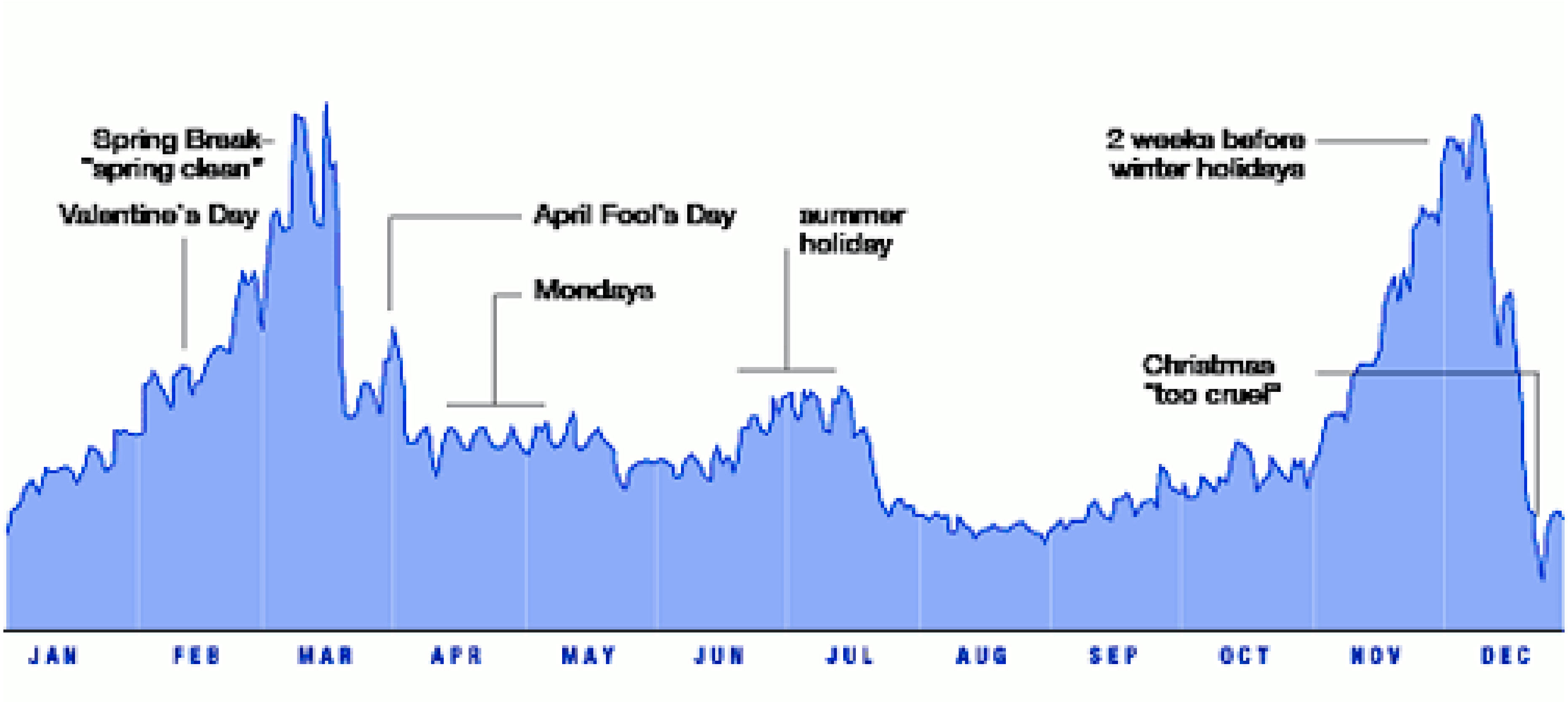
https://en.wikipedia.org/wiki/Florence_Nightingale

Napoleon's Invasion of Russia



https://en.wikipedia.org/wiki/French_invasion_of_Russia

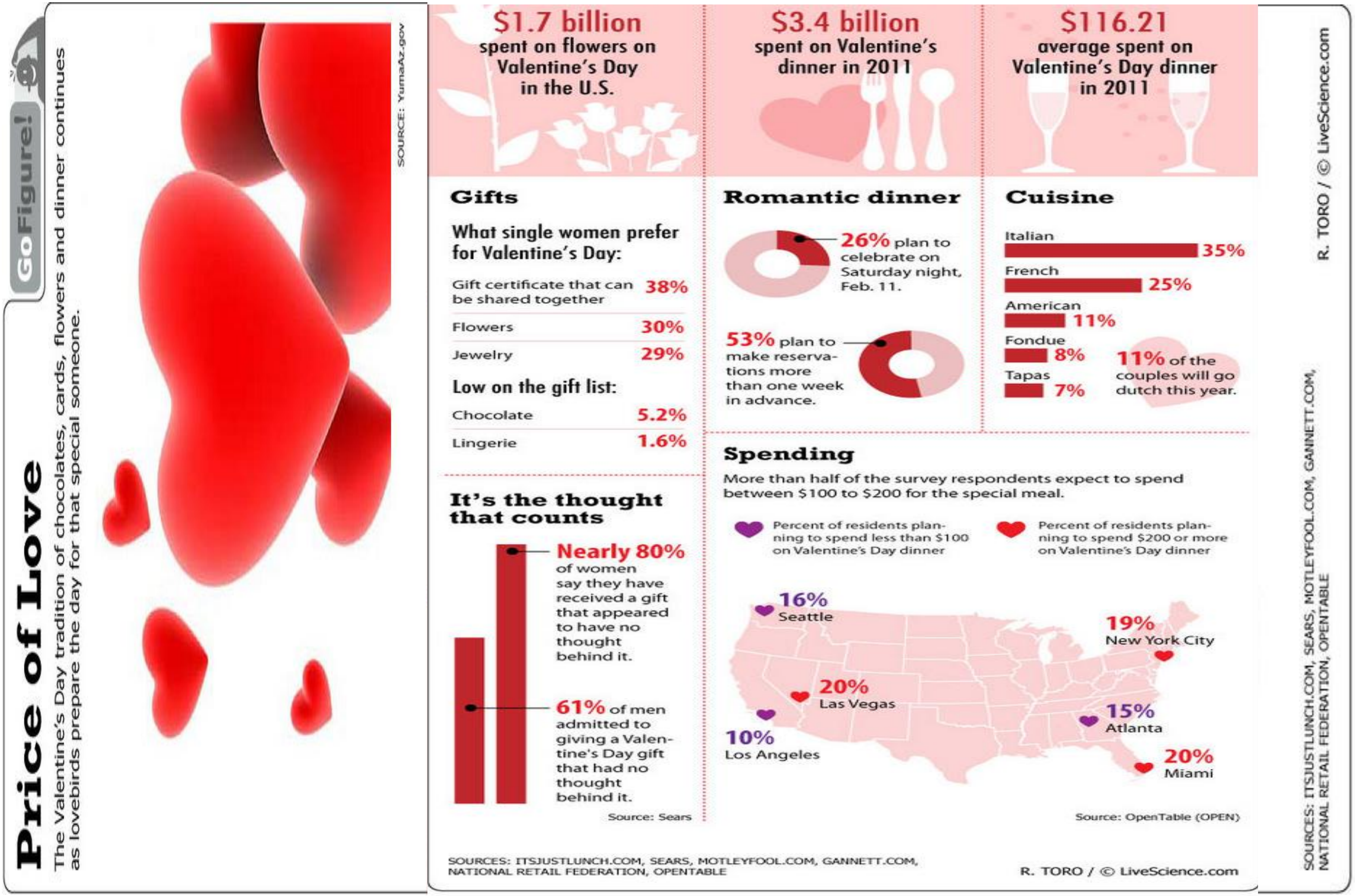
Relationship Breakups Status Update on Facebook



Facebook Relationship Breakups

Infographics–Price of Love

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE



Descriptive Analytics Applications

- Most shoppers turn towards right when they enter the a retail store.
- Conversion rate of women shoppers is higher than male shoppers among electronic gadgets purchasers (Radio Shack).
- Strawberry pop-tarts sell 7 times more during hurricane compared to regular period (Wal Mart).
- Women car buyers prefer women sales person.

Predictive Analytics Problems

- Which product the customer is likely to buy in his next purchase (recommender system).
- Which customer is likely to default in his/her loan payment.
- Who is likely to cancel the product that was ordered through e-commerce portal.

Prescriptive Analytics Problem

- What is the optimal route for a delivery truck.
- Whether a company should introduce a new product?
- What is the optimal product mix?
- How to manage the fleet of vehicles owned by a company for employee drop and pick up?

Framework For Decision Making

Opportunity Identification

- Domain knowledge is very important at this stage of the analytics project. This will be a major challenge for many companies who do not know the capabilities of analytics.

Collection of relevant data

- Once the problem is defined clearly, the project team should identify and collect the relevant data.
- This may be an interactive process since "relevant data" may not be known in advance in many analytics projects. The existence of ERP systems will be very useful at this stage.

Data Pre-processing

- Data preparation and data processing forms a significant proportion of any analytics project.
- This would include data imputation and the creation of additional variables such as interaction variables and dummy variables in the case of predictive analytics projects.

Model Building

- Analytics model building is an iterative process that aims to find the best model. Several analytical tools and solution procedures will be used to find the best analytical model in this stage.

Communication of the data analysis

- The communication of the analytics output to the top management and clients plays a crucial role.
- Innovative data visualization techniques may be used in this stage.

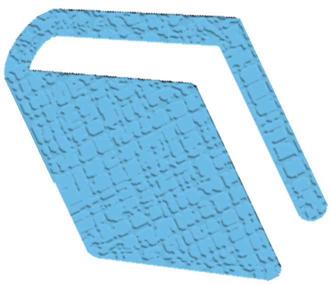
Industry Wide Application of Analytics

Manufacturing	Retail	Healthcare	Service	Banking and Finance	IT and ITES (IT enabled services)
Supply chain analytics	*Assortment Planning*	*Clinical Care*	*Demand Forecasting*	*Service Demand Analysis*	*Demand for Analytics Services*
Quality and Process improvement	*Promotion Planning*	*Hospitality related data*	*Service Quality Analysis*	*Customer Transaction Analysis*	*Software Development Cycle Time*
Revenue and Cost Management	*Demand Forecasting*		*Customer Segmentation*	*Credit Scoring*	
	Market Basket Analysis		*Promotion*		
	Customer Segmentation				

**Primary sources of data and secondary sources to be used in solving these analytical problems

Summary

Summary of the topics covered in this lesson:



- With the data explosion across industry, the usage of analytics in decision making will become the most critical factor for being competitive in business.
- Descriptive analytics becomes the stepping stone to all the complex problems which can be solved using analytics.

End of Lesson–Introduction to Business Analytics

भारतीय प्रबंध संस्थान बेंगलूर

INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Data Science Using R

Lesson01–Overview of R

Objective

After completing this lesson you will be able to:

- Describe the genesis of R
- Locate and install R in the system
- Explain R Studio interface
- Install packages from the repositories

Overview and History of R

R is a dialect of the S language. S is a language that was developed by John Chambers and others at Bell Labs.

Features:

- Runs on almost any standard computing platform/OS
- Frequent releases (annual + bug fix releases); active development
- Quite lean, as far as software goes; functionality is divided into modular packages
- Very sophisticated graphics capabilities; better than most stat packages
- Useful for interactive work, but contains a powerful programming language for developing new tools
- Very active and vibrant user community; contains R-help and R-developer mailing lists and Stack Overflow

R is an official part of the Free Software Foundation's GNU project.
RStudio and Shiny are affiliated projects of the Foundation for Open Access Statistics.

Design of the R System

- The R system is divided into two conceptual parts:

Base R	R Extension
• Base R can be downloaded from CRAN (http://cran.r-project.org). • The base R system contains, among other things, the base package, which is required to run R and contains the most fundamental functions.	• There are about 4000 packages on CRAN that have been developed by users and programmers around the world. • There are also many packages associated with the Bioconductor project (http://bioconductor.org).

R Resources

- Number of references and reading material can be found on R.

R resources available from CRAN

- An Introduction to R
- Writing R Extensions
- R Data Import/Export
- R Installation and Administration (mostly for building R from sources)
- R Internals (not for the faint of heart)

Standard books

- Chambers (2008). Software for Data Analysis, Springer (your textbook)
- Chambers (1998). Programming with Data, Springer
- Venables & Ripley (2002). Modern Applied Statistics with S, Springer
- Venables & Ripley (2000). S Programming, Springer
- Pinheiro & Bates (2000). Mixed-Effects Models in S and S-PLUS, Springer
- Murrell (2005). R Graphics, Chapman & Hall/CRC Press

R Dataset and Tools

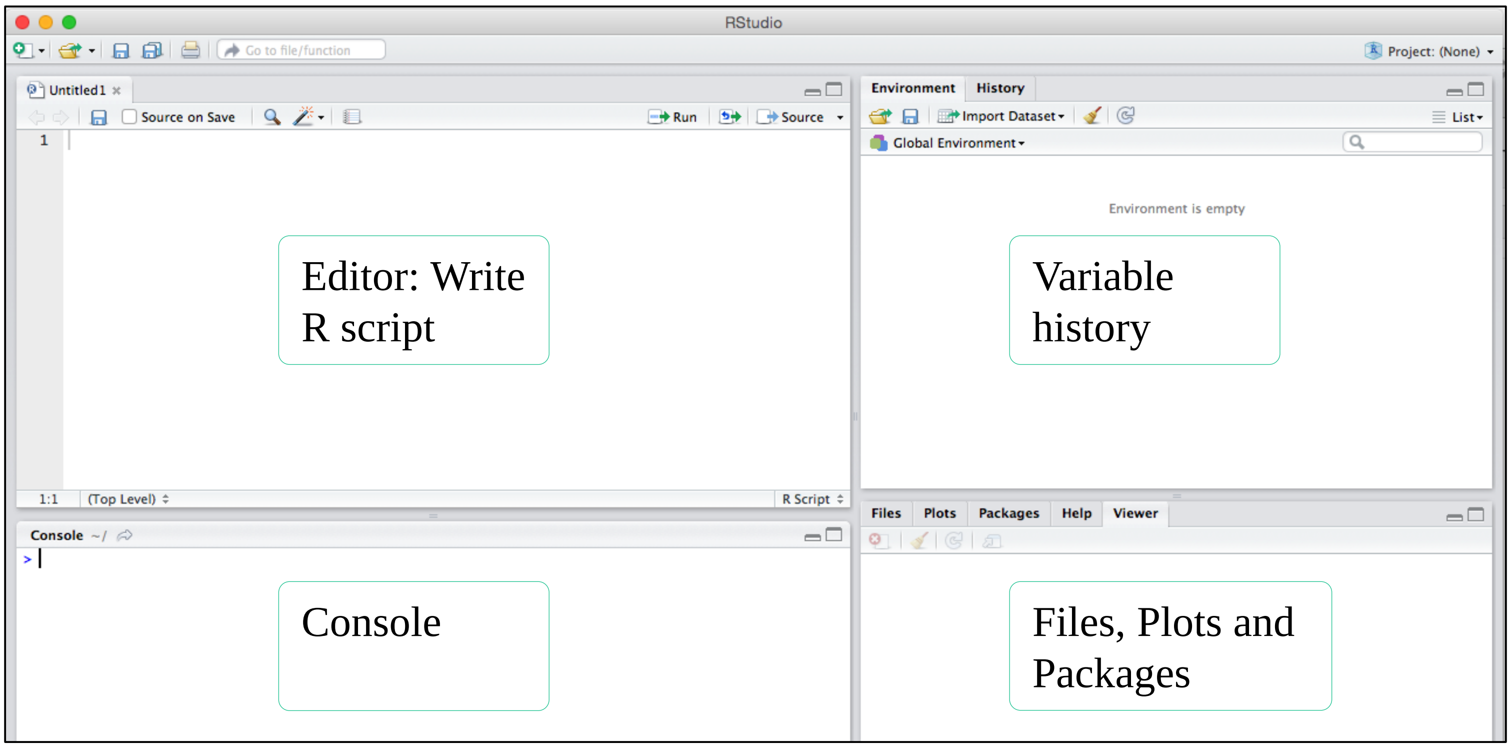
- R comes with a number of sample datasets that you can experiment with.

On R Studio Console:

```
data( ) #to see the available datasets in the installed packages
help(datasetname)  #for details on a sample dataset
```

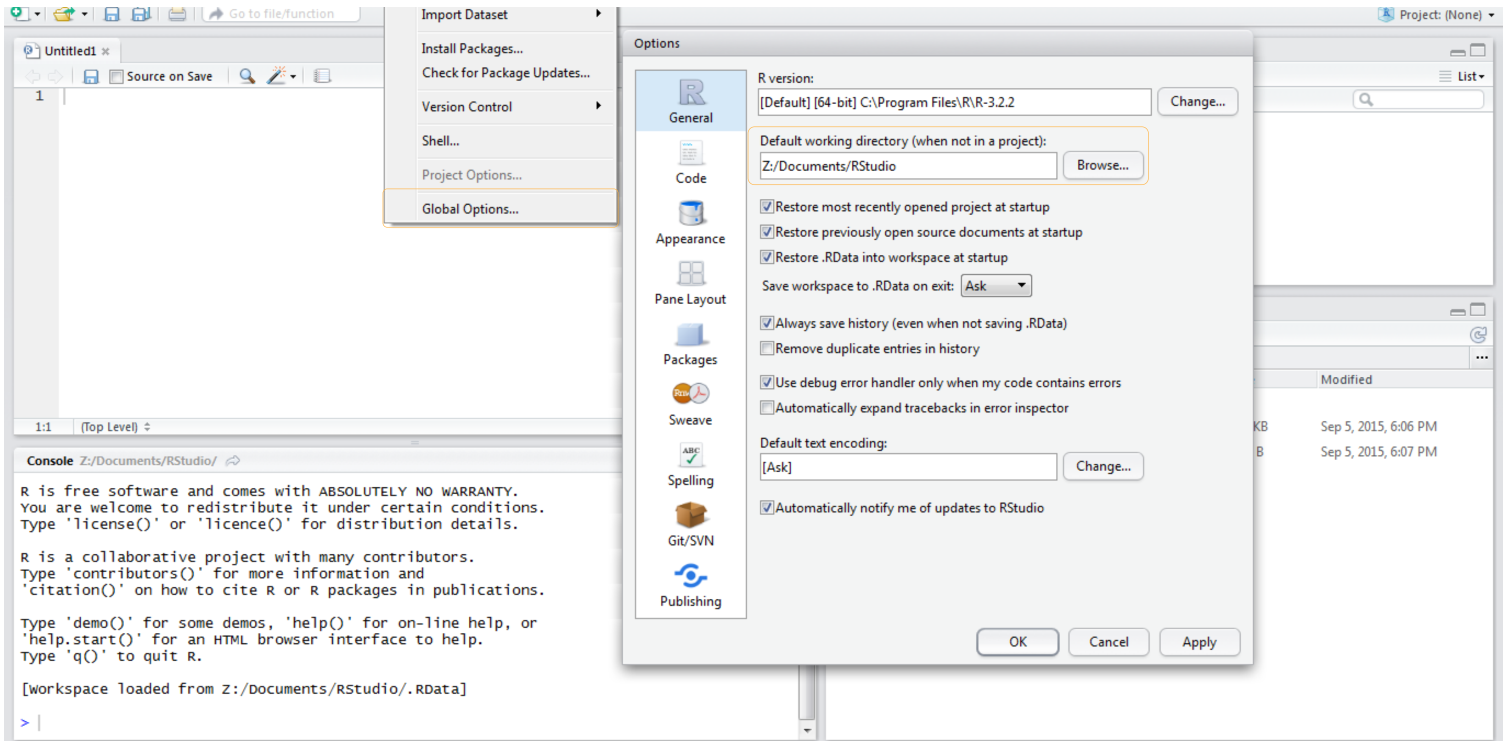
Introducing R Studio

- R Studio (version 0.99.442) console has four primary blocks:



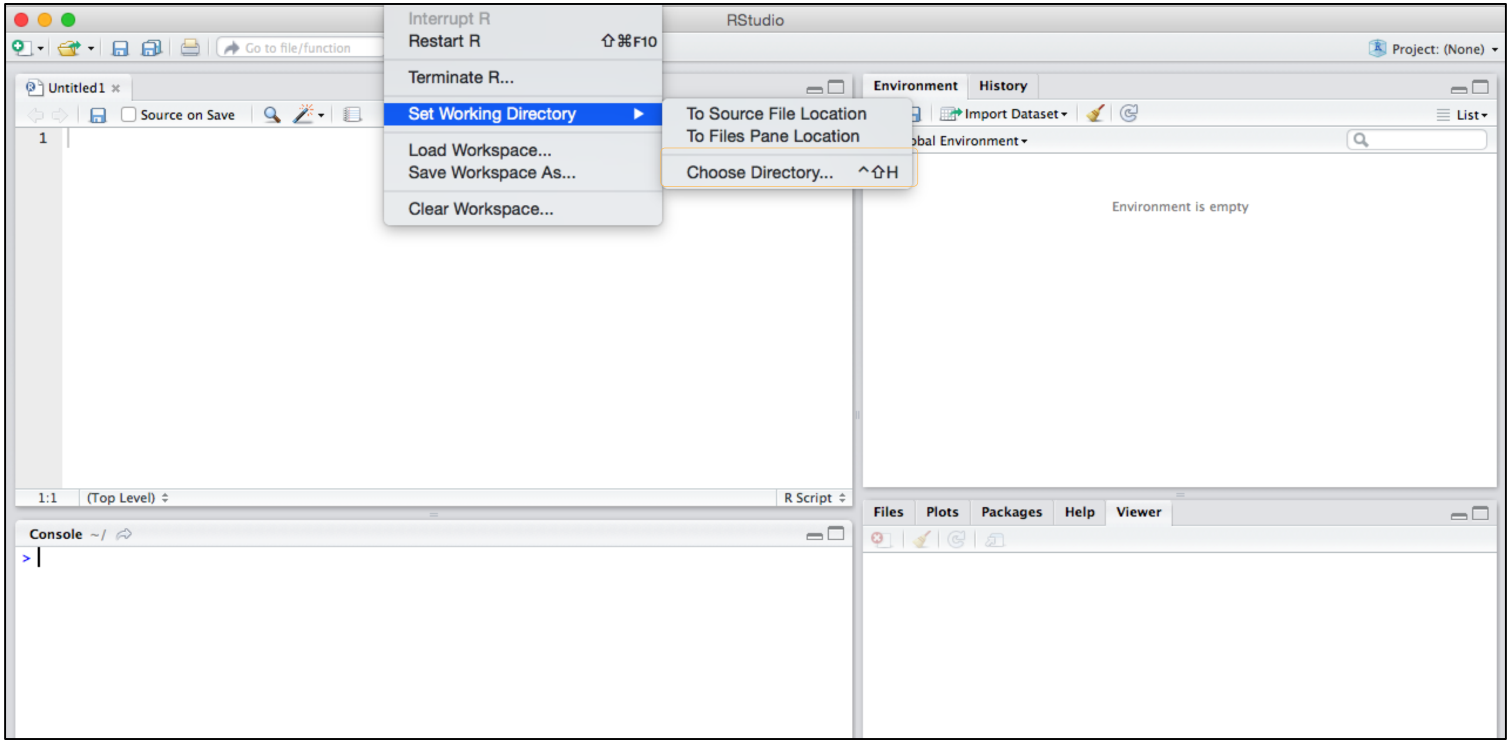
R Studio—Setting Global Working Directory

- Set the global working directory through the option below:



R Studio—Setting Local Working Directory

- Local working directory can be set as shown below. Two useful commands: getwd() and setwd()



Choosing a CRAN Mirror and Setting Repositories

- CRAN mirrors contain R packages that can extend the functionality of R.
- Choose a mirror located close to you as that will most likely give you the fastest downloads
- Repositories host the packages. Some of the examples of repositories are:

CRAN, CRANextra, BioCsoft, BioCann, BioCexp, BioCext, Omegahat, R-Forge and rforge.net

```
Console Z:/Documents/RStudio/
help.start() for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from Z:/Documents/RStudio/.RData]

> setRepositories()
--- Please select repositories for use in this session ---


 1: + CRAN                 2:   BioC software        3:   BioC annotation
 4:   BioC experiment      5:   BioC extra           6: + CRAN (extras)
 7:   Omegahat             8:   R-Forge              9:   rforge.net
10:   CRAN (extras, https) 11:   R-Forge [https]     12:   rforge.net [https]

Enter one or more numbers separated by spaces, or an empty line to cancel
1:
```

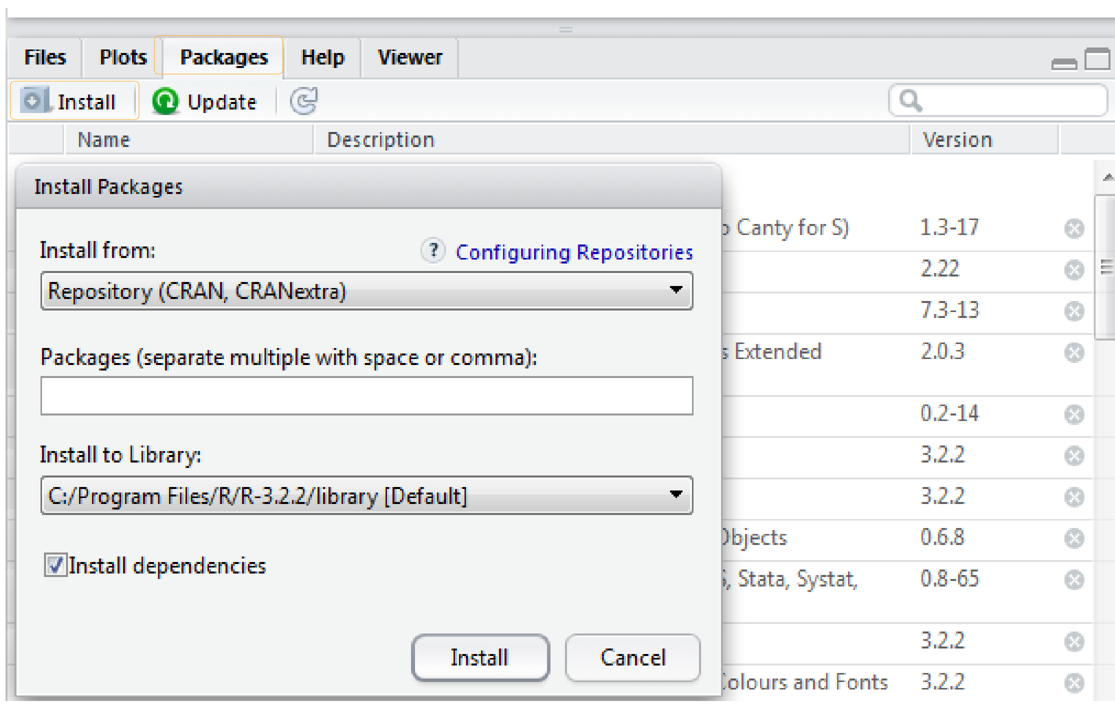
- Use this code to set your repositories: `setRepositories()`

Package TAB

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Packages can be installed as and when required.

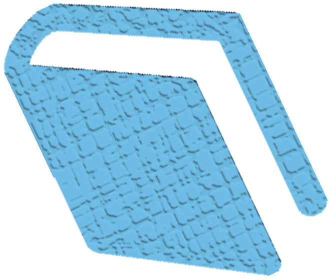
- The Packages tab lets you see what packages are installed
- A package must be loaded before you can use it
  - In Rstudio, this is accomplished by clicking the checkbox next to the package name in the package tab



This demo will show the R Studio features.

Summary

Summary of the topics covered in this lesson:



- R is a dialect of the S language. S is a language that was developed by John Chambers and others at Bell Labs.
- R comes with many built in dataset which can be used to practice the analytical concepts.

QUIZ TIME

Quiz Question 1

Quiz 1

Which of the following is true about R?

Select all that apply.

a. *Runs on almost any standard computing platform/OS*

b. *R has a very active and vibrant user community*

c. *R has very sophisticated graphics capabilities; better than most stat packages*

d. *User community does not provide help on R*

Quiz Question 1

Quiz 1 Which of the following is true about R?

Select all that apply.

a. *Runs on almost any standard computing platform/OS*

b. *R has a very active and vibrant user community*

c. *R has very sophisticated graphics capabilities; better than most stat packages*

d. *User community does not provide help on R*

Correct answer is:

a , b & c

All the options are correct except d. User community is helpful.

Quiz Question 2

Quiz 2	Which of the following is a name of the repository from where packages can be downloaded for use in R? *Select all that apply.*

a. *CRAN*

b. *Cranberry*

c. *R-Forge*

d. *Cran soft*

Quiz Question 2

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Quiz 2	Which of the following is a name of the repository from where packages can be downloaded for use in R? *Select all that apply.*

a. *CRAN*

b. *Cranberry*

c. *R-Forge*

d. *Cran soft*

Correct answer is:

a & c

b and d are not the name of repositories in R.

End of Lesson01–Overview of R

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Data Science Using R

Lesson02–Fundamentals of R

Objective

After completing this lesson you will be able to:

- Import data files into an R system
- Perform basic data manipulation

Reading Data into R

Data can be imported to R in multiple ways. Two commonly used ways are:

From a csv file

```
myData <-
read.table(file.choose(),hea
der=TRUE, sep=",")
# first row of the csv being
#read, contains variable
#names, comma is separator.
```

From an excel file

```
library(xlsx)
myData <-
read.xlsx(file.choose(),head
er = TRUE, sheetIndex = 1)
# first row of the excel
#being read contains
#variable names.
```

The best way to read an Excel file is to export it to a comma delimited file and import it using read.table() or read.csv()

Creating and Renaming Variables

INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Creating and renaming variables are two important aspects in R.

Creating New Variables

- New variables are created by using the assignment operator '<-'.

```
mydata <- list(x1= 2, x2=5)
mydata$sum <- mydata$x1 +
mydata$x2
mydata$mean <- (mydata$x1 +
mydata$x2)/2

attach(mydata)
mydata$sum <- x1 + x2
mydata$mean <- (x1 + x2)/2
detach(mydata)
```

Renaming variables

- Variables can be renamed programmatically or interactively

```
# rename interactively
t <- list(name= “roy",
age=30, gender="M")
fix(t) # results are saved on
close

# rename programmatically
library(reshape)
mydata <- rename(mydata,
c(oldname="newname"))
```

Understanding Data Types—Vectors and Matrix

There is a fine distinction between vectors and matrix.

- Vectors: All elements must be of the same type. **Example:**

```
name <- c("Mike", "Lucy", "John") #vector of string
or character
age <- c(20, 25, 30) #vector of integers
```

- Matrix: A special kind of vector with two additional attributes i.e. the number of rows and the number of columns. **Example**:

```
x <- matrix(c(1,2,3,4), nrow=2, ncol=2)
```

Understanding Data Types—List

List is an ordered collection of objects which allows to gather a variety of (possibly unrelated) objects under one name.

- **Example** of a list with 4 components: A string, a numeric vector, a matrix and a scaler.

```
w <- list(name="Fred", mynumbers=a, mymatrix=y,
age=5.3)
```

- **Example** of a list containing four vectors:

```
n = c(2, 3, 5)
s = c("aa", "bb", "cc", "dd", "ee")
b = c(TRUE, FALSE, TRUE, FALSE, FALSE)
x = list(n, s, b, 3)   # x is a list which contains
copies of n, s, b
```

Understanding Data Types—Factors

A factor stores the nominal values as a vector of integers in the range [1... k] (where k is the number of unique values in the nominal variable) and an internal vector of character strings (the original values) mapped to these integers.

- **Example**: Variable gender with 20 "male" entries and 30 "female" entries

```
gender <- c(rep("male",20), rep("female", 30))
gender <- factor(gender) # stores gender as 20 1s and 30 2s and
associates. 1=female, 2=male.

#R now treats gender as a nominal variable
summary(gender)
```

- The order of the levels can be set using the levels argument to factor(). This can be important in linear modelling.

Understanding Data Types—DataFrames

Data frames are used to store tabular information.

- They are represented as a special type of list, where every element of the list has to have the same length
- Each element of the list can be thought of as a column and the length of each element of the list is the number of rows
- Unlike matrices, data frames can store different classes of objects in each column (just like lists); matrices must have every element of the same class
- Data frames also have a special attribute called row.names
- Data frames are usually created by calling read.table() or read.csv()
- Can be converted to a matrix by calling data.matrix()

Data Sub setting in R

R has robust subsetting feature which can be used for selecting or excluding variables from a dataset.

- Below are the examples for selecting or excluding variables

Examples for selection:

```
# select variables mpg, cyl,
disp from mtcars dataset
myvars <- c("mpg", "cyl",
"disp")
newdata <- mtcars[myvars]


# select 1st and 7th through
11th variables
newdata <- mtcars[c(1,7:11)]
```

Examples for exclusion:

```
# exclude variables mpg, cyl,
disp from mtcars dataset
myvars <- names(mtcars) %in%
c("mpg", "cyl", "disp")
newdata <- mtcars[!myvars]

# exclude 1st and 7th
variables
newdata <- mtcars[c(-1,-7)]
```

Data Sub setting in R

Subsetting feature in R can be used for selecting or excluding observations as well.

- Below are the examples for selecting observations from a dataset:

Examples for selection:

```
# first 5 observations of
#mtcars dataset across all
#variables
newdata <- mtcars[1:5,]

# select observations of
#mtcars dataset based on
#condition
newdata <-
mtcars[which(mtcars$hp >100 &
mtcars$cyl > 4),]
```

Examples for selection using subset:

```
#select hp and cyl from
#mtcars
newdata <- subset(mtcars, hp
> 100 | cyl < 10,
select=c(hp, cyl))]

# select 1st through 6th
#variable from mtcars dataset
newdata <- subset(mtcars, hp
> 100 | cyl < 10,
select=c(1:6))
```

Data Manipulation—Sort

To sort a dataframe in R, use the order( ) function. By default, sorting is ASCENDING. Prepend the sorting variable by a minus sign to indicate the DESCENDING order.

Example: *Sorting examples using the mtcars dataset*

```
data(mtcars)
# sort by mpg
newdata = mtcars[order(mtcars$mpg),]
#sort by mpg and cyl
newdata <- mtcars[order(mtcars$mpg, mtcars$cyl),]
#sort by mpg (ascending) and cyl (descending)
newdata <- mtcars[order(mtcars$mpg, -mtcars$cyl),]
```

Data Manipulation—Merge

To merge two DataFrames horizontally, use the merge function. Typically, DataFrames are joined by one or more common key variables. Merge two data frames by ID

```
total <- merge(dataframeA,dataframeB,by="ID")
```

Delete the extra variables in dataframeA or

```
total <- merge(dataframeA,dataframeB,by=c("ID","Country"))
```

Create the additional variables in dataframeB and set them to NA (missing) before joining

```
total <- rbind(dataframeA, dataframeB)
```

For merging vertically, two dataframes must have same variables. If not then:

- Merge two data frames by ID and Country
- To join two dataframes (datasets) vertically, use the rbind function.

Data Manipulation—Aggregate

Aggregate function is used to summarize the data by a variable

```
# aggregate dataframe iris by species and return
means for numeric variables
attach(iris)
aggdata <-aggregate(iris, by=list(iris$Species),
FUN=mean, na.rm=TRUE)
print(aggdata)
```

When using the aggregate() function, the **'by'** variables must be in a list (even if there is only one). The function can be built-in or user provided.

Data Manipulation—Aggregate

Other ways to aggregate data is by using summarize() function available in the Hmisc package.

- Example 1:

```
summarize(iris$Sepal.Length,iris$Species,FUN=mean)
```

- Example 2:

```
set.seed(1)
temperature <- rnorm(300, 70, 10)
month <- sample(1:12, 300, TRUE)
year  <- sample(2000:2001, 300, TRUE)
g <-
function(x)c(Mean=mean(x,na.rm=TRUE),Median=median(x,na.rm=
TRUE))
summarize(temperature, month, g)
```

Data Manipulation—Data Conversion

- Data conversion in R can be done using the pre-defined functions.
- For example, adding a character string to a numeric vector converts all the elements in the vector to character.

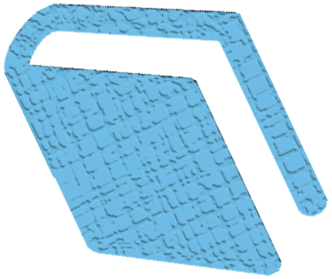
Some useful functions for type conversion:

is.numeric(), is.character(), is.vector(), is.matrix(), is.data.frame(), as.numeric(), as.character(), as.vector(), as.matrix(), as.data.frame()

This demo will show the use of data operations in R using Rstudio.

Summary

Summary of the topics covered in this lesson:



- Vectors, Matrix, List, Factors and Dataframes are different data types which can be used to store datasets.
- Read, Sort, Merge, Aggregate are some of the basic data manipulation techniques which can be helpful in data analysis.
- R has robust subsetting feature which can be used for selecting or excluding variables or observations from a dataset.

QUIZ TIME

Quiz Question 1

Quiz 1

What will be the data type of age in the following code:
age <- c(20, 25, 30). *Select all that apply.*

a. *Vector of integers*

b. *List*

c. *DataFrame*

d. *Matrix*

Quiz Question 1

Quiz 1

What will be the data type of age in the following code:
age <- c(20, 25, 30). *Select all that apply.*

a. *Vector of integers*

b. *List*

c. *DataFrame*

d. *Matrix*

Correct answer is: *a*

age is a vector of integers.

Quiz Question 2

Quiz 2	Which command when typed in console will show the inbuilt dataset in R?

a. *dataset()*

b. *dataframe()*

c. *data()*

d. *showData()*

Quiz Question 2

Quiz 2 Which command when typed in console will show the inbuilt dataset in R?

a. *dataset()*

b. *dataframe()*

c. *data()*

d. *showData()*

Correct answer is: ***a***

dataset() shows the inbuilt dataset in R when typed in console; others are not commands to show built in dataset in R.

Quiz Question 3

Quiz 3

What will the following code give as output: *newdata <- iris[c(-1,-2)]*

a. *Will create a variable newdata with the first two columns included.*

b. *Will create a variable newdata with the first two columns excluded.*

c. *Will create a variable iris with the first two columns excluded.*

d. *Will create a variable iris with the first two columns included.*

Quiz Question 3

Quiz 3 What will the following code give as output: *newdata <- iris[c(-1,-2)]*

a. *Will create a variable newdata with the first two columns included.*

b. *Will create a variable newdata with the first two columns excluded.*

c. *Will create a variable iris with the first two columns excluded.*

d. *Will create a variable iris with the first two columns included.*

Correct answer is: ***b***

iris dataset is inbuilt in R and the above command will exclude the first two columns.

End of Lesson02–Fundamentals of R

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Data Science Using R

Lesson03–Using Functions and Loops in R

Objective

After completing this lesson you will be able to:

- Use the control structures in R
- Create a user defined function in R
- Identify the built in functions in R

Control Structures

Control structures in R allow you to control the flow of execution of the program, depending on runtime conditions. Common structures are:

- *if else — testing a condition*
- *for — execute a loop a fixed number of times*
- *while — execute a loop while a condition is true · repeat: execute an infinite loop*
- *break — break the execution of a loop*
- *next — skip an interaction of a loop*
- *return — exit a function*

Most control structures are not used in interactive sessions but when writing functions or longer expressions.

Control Structures—If Condition

General construct of “if” control structure is given below.

Construct 1:

```
if(<condition>) {
## do something
}
else{
## do something else
}

else clause is not necessary.
if(<condition1>) { }
if(<condition2>) { }
```

Construct 2:

```
if(<condition1>) {
## do something
}
else if(<condition2>) {
## do something different
}
else{
## do something different
}
```

Control Structures—For Loop

Example of “for” control structure is given below.

Example:

```
for(i in 1:10) { print(i)
}
```

The above loop takes the i variable and in each iteration of the loop gives it values 1, 2, 3, ..., 10, and then exits.

Example *with break statement*:

```
for(i in 1:10) {
  print(i)
  if (i==2){
    break
  }}
```

Print i and break as soon as i equals 2

For loops are most commonly used for iterating over the elements of an object (list, vector, etc.)

Control Structures—Nested For Loop

Example of “for” control structure is given below.

Example with nested for loops*:*

```
x <- matrix(1:6, 2, 3)
for(i in seq_len(nrow(x))) {
  for(j in seq_len(ncol(x))) {
    print(x[i, j])
    }
  }
```

*X is a 2*3 matrix. The for loop is being used to print the values of the matrix row wise.*

Be careful with nesting though. Nesting beyond 2–3 levels is often very difficult to read or understand the nested loops.

Control Structures—While Loops

Example of “while” control structure is given below.

ε

Example *with one condition*:

```
count <- 0
while(count < 10) {
    print(count)
    count <- count + 1
    }
```

While loops begin by testing a condition. If it is true, then execute the loop body. Once the loop body is executed, the condition is tested again, and so forth.

ε

Example with multiple condition:

```
z<-5
while(z>=3&&z<=10){
print(z)
coin <- rbinom(1, 1, 0.5)
#generate 0 or 1
if(coin == 1) {
#do not change z }
else{ z<-z-1  }}
```

Loop, print z till the condition is satisfied. decrease the value of z when the coin toss results in 0.

While loops can potentially result in infinite loops if not written properly. Use with care.
Conditions are always evaluated from left to right.

Loop functions in R

R has inbuilt functions to implement loops. These functions takes away the complexity of writing “for” or “while” loops.

- **Apply function**: Function over the margins of an array

```
y <- matrix (rnorm (100), 10, 5)
apply (y, 2, mean) # 2 means column
wise
```

- **Lapply function**: Loop over a list and evaluate a function on each element

```
y <- list(i = 1:5, n = rnorm(10))
lapply (y, mean)
```

- **Sapply function**: Same as lapply but tries to simplify the result

```
y <- list(i = 1:5, n =
rnorm(10))
sapply (y, mean)
```

- **Tapply function**: Apply a function over subsets of a vector

```
x <- rnorm(30)
f <- gl(3, 10)
df <-data.frame(x,f)
tapply(df$x, df$f, mean)
```

- **Mapply function**: Multivariate version of lapply

User Defined Functions

R gives the flexibility of writing custom function.

- Below is a structure of a function followed with example:

Structure of a user defined function:

```
newfunction <- function(arg1,
arg2, ... ){
statements
return(object)
}
```

Example of a user defined function:

```
 summarize <- function(x) {
    center <- mean(x); spread
<- sd(x)
    cat("Mean is", center,
"\n", "Std dev is", spread,
"\n")
    result <-
list(center=center,
spread=spread)
    return(result)
  }
  set.seed(12345)
  x <- rnorm(500)
  y <- summarize(x)
```

R Built-in Functions

- Numeric functions:

Function	Description
abs(x)	absolute value
sqrt(x)	square root
ceiling(x)	ceiling(3.475) is 4
floor(x)	floor(3.475) is 3
trunc(x)	trunc(5.99) is 5
round(x, digits=n)	round(3.475, digits=2) is 3.48
signif(x, digits=n)	signif(3.475, digits=2) is 3.5
cos(x), sin(x), tan(x)	also acos(x), cosh(x), acosh(x), etc.
log(x)	natural logarithm
log10(x)	common logarithm
exp(x)	e^x

R Built-in Functions

- Character functions:

Function	Description
substr(x, start=n1, stop=n2)	Extract or replace substrings in a character vector. x <- "abcdef" substr(x, 2, 4) is "bcd" substr(x, 2, 4) <- "22222" is "a222ef"
grep(pattern, x , ignore.case=FALSE, fixed=FALSE)	Search for pattern in x. If fixed =FALSE then pattern is a regular expression. If fixed=TRUE then pattern is a text string. Returns matching indices. grep("A", c("b","A","c"), fixed=TRUE) returns 2
sub(pattern, replacement, x, ignore.case =FALSE, fixed=FALSE)	Find pattern in x and replace with replacement text. If fixed=FALSE then pattern is a regular expression. If fixed = T then pattern is a text string. sub("\\s",".","Hello There") returns "Hello.There"

- Character functions:

Function	Description
strsplit(x, split)	Split the elements of character vector x at split. strsplit("abc", "") returns 3 element vector "a","b","c"
paste(..., sep="")	Concatenate strings after using sep string to seperate them. paste("x",1:3,sep="") returns c("x1","x2" "x3") paste("x",1:3,sep="M") returns c("xM1","xM2" "xM3") paste("Today is", date())
toupper(x)	Uppercase
tolower(x)	Lowercase

- Stat functions:

Function	Description

dnorm(x)	normal density function (by default m=0 sd=1)
	# plot standard normal curve
	x <- pretty(c(-3,3), 30)
	y <- dnorm(x)
	plot(x, y, type='l', xlab="Normal Deviate", ylab="Density", yaxs="i")

pnorm(q)	cumulative normal probability for q (area under the normal curve to the right of q) pnorm(1.96) is 0.975
qnorm(p)	normal quantile. value at the p percentile of normal distribution qnorm(.9) is 1.28 # 90th percentile
rnorm(n, m=0,sd=1)	n random normal deviates with mean m and standard deviation sd. #50 random normal variates with mean=50, sd=10 x <- rnorm(50, m=50, sd=10)

R Built-in Functions

- Stat functions:

Function	Description

dbinom(x, size, prob)
pbinom(q, size, prob)
qbinom(p, size, prob)
rbinom(n, size, prob)
	binomial distribution where size is the sample size and prob is the probability of a heads (pi)
	# prob of 0 to 5 heads of fair coin out of 10 flips
	dbinom(0:5, 10, .5)
	# prob of 5 or less heads of fair coin out of 10 flips
	pbinom(5, 10, .5)

dpois(x, lamda)
ppois(q, lamda)
qpois(p, lamda)
rpois(n, lamda)
	poisson distribution with m=std=lamda
	#probability of 0,1, or 2 events with lamda=4
	dpois(0:2, 4)
	# probability of at least 3 events with lamda=4
	1- ppois(2,4)

dunif(x, min=0, max=1) punif(q, min=0, max=1) qunif(p, min=0, max=1) runif(n, min=0, max=1)	uniform distribution, follows the same pattern as the normal distribution above. #10 uniform random variates x <- runif(10)

R Built-in Functions

- Stat functions:

Function	Description
mean(x, trim=0, na.rm=FALSE)	mean of object x # trimmed mean, removing any missing values and # 5 percent of highest and lowest scores mx <- mean(x,trim=.05,na.rm=TRUE)
sd(x)	standard deviation of object(x). also look at var(x) for variance and mad(x) for median absolute deviation.
median(x)	median
mean(x, trim=0, na.rm=FALSE)	mean of object x # trimmed mean, removing any missing values and # 5 percent of highest and lowest scores mx <- mean(x,trim=.05,na.rm=TRUE)

- Stat functions:

Function	Description
quantile(x, probs)	quantiles where x is the numeric vector whose quantiles are desired and probs is a numeric vector with probabilities in [0,1]. # 30th and 84th percentiles of x y <- quantile(x, c(.3,.84))
range(x)	range
sum(x)	sum
diff(x, lag=1)	lagged differences, with lag indicating which lag to use
min(x)	minimum
max(x)	maximum

R Built-in Functions

- Stat functions:

Function	Description
scale(x, center=TRUE, scale=TRUE)	column center or standardize a matrix.
seq(from , to, by)	generate a sequence indices <- seq(1,10,2) #indices is c(1, 3, 5, 7, 9)
rep(x, ntimes)	repeat x n times y <- rep(1:3, 2) # y is c(1, 2, 3, 1, 2, 3)
cut(x, n)	divide continuous variable in factor with n levels y <- cut(x, 5)

This demo will show the use of concepts covered in this lesson using Rstudio.

Summary

Summary of the topics covered in this lesson:

- If/else, for, while are typical control structures available in R. R also has specific loop functions like apply, tapply, mapply etc. which works exactly like the control structures.
- User defined functions give the flexibility of writing generic functions which can be used to structure a complex code.
- The in-built functions in R gives flexibility to perform complex data manipulations while analyzing datasets.

QUIZ TIME

Quiz Question 1

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Quiz 1

Which of the following is a loop function in R?

Select all that apply.

a. *apply*

b. *gapply*

c. *tapply*

d. *mapply*

Quiz Question 1

Quiz 1 Which of the following is a loop function in R?
Select all that apply.

a. *apply*

b. *gapply*

c. *tapply*

d. *mapply*

Correct answer is: **a, c & d**

All the options are correct except b. gapply is not a function in R.

Quiz Question 2

Quiz 2	What is the output of following code? count <- 0 while(count < 10) { count <- count + 1 } print(count)

a. *10*

b. *11*

c. *9*

d. *8*

Quiz Question 2

Quiz 2	What is the output of following code? count <- 0 while(count < 10) { count <- count + 1 } print(count)

a. *10*

b. *11*

c. *9*

d. *8*

Correct answer is:

a

The code when run in R will give 10 as an output.

Quiz Question 3

Quiz 3	What will the following code give as output: rnorm(40, 50, 10)

a. *Generate 40 random numbers with a mean of 50 and std. dev of 10*

b. *Generate 50 random numbers with a mean of 40 and std. dev of 10*

c. *Generate 10 random numbers with a mean of 40 and std. dev of 50*

d. *Generate 40 random numbers with a mean of 10 and std. dev of 50*

Quiz Question 3

INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Quiz 3	What will the following code give as output: rnorm(40, 50, 10)

a. *Generate 40 random numbers with a mean of 50 and std. dev of 10*

b. *Generate 50 random numbers with a mean of 40 and std. dev of 10*

c. *Generate 10 random numbers with a mean of 40 and std. dev of 50*

d. *Generate 40 random numbers with a mean of 10 and std. dev of 50*

Correct answer is: ***a***

First variable is the number of observation, second is the mean and third is std dev.

Quiz Question 4

Quiz 4 Which of the following is a function in R?

a. *abs(), sqrt(), sub(), dnorm()*

b. *abs(), sqrt(), subtract(), dnorm()*

c. *abs(), sqrted(), sub(), dnorm()*

d. *abs(), sqrt(), sub(), wnorm()*

Quiz Question 4

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Quiz 4	What will the following code give as output: rnorm(40, 50, 10)

a. *abs(), sqrt(), sub(), dnorm()*

b. *abs(), sqrt(), subtract(), dnorm()*

c. *abs(), sqrted(), sub(), dnorm()*

d. *abs(), sqrt(), sub(), wnorm()*

Correct answer is:

a

Other options have one or more options which are not functions in R

End of Lesson03–Using Loop and Functions in R

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Data Science Using R

Lesson04–Data Visualization using R

Objective

After completing this lesson you will be able to:

- Explain the importance of Data Visualization
- Create bar chart, pie chart, mosaic plot using R
- Create scatter plot, histogram and correlation plot in R
- Create box plot and other advanced plotting using R

Exercise 1

How many cells with revenue greater than 15 lacs?

10	23	12	14	10	13	123	12	8	2
20	2	14	19	13	43	12	56	5	4
4.5	12	16	20	31	56	3	7	2	2
10	13	12	12	42	7	5	6	134	7
11	7	13	6	5	8	12	4	150	5
12	14	15	7	7	3	4	18	7	2
3	18	15	8	12	12	87	2	12	12
8	12	14	4	13	1	3	5	12	5
13	3	17	12	12	4	15	5	3	23
17	5	12	10	11	8	8	12	5	45
1	9	3	12	10	12	2	13	7	12
2	12	10	14	2	9	13	6	6	6

Consider the above table to be revenue in lacs from various technologies(rows) in different domains (columns) for an IT firm.

Exercise 1…

How many cells with revenue greater than 15 lacs?									
10	23	12	14	10	13	123	12	8	2
20	2	14	19	13	43	12	56	5	4
4.5	12	16	20	31	56	3	7	2	2
10	13	12	12	42	7	5	6	134	7
11	7	13	6	5	8	12	4	150	5
12	14	15	7	7	3	4	18	7	2
3	18	15	8	12	12	87	2	12	12
8	12	14	4	13	1	3	5	12	5
13	3	17	12	12	4	15	5	3	23
17	5	12	10	11	8	8	12	5	45
1	9	3	12	10	12	2	13	7	12
2	12	10	14	2	9	13	6	6	6

Color the numbers greater than 15

Exercise 2

Which BUs generate comparable revenue from every technology?									
10	23	12	14	10	13	123	12	8	2
20	2	14	19	13	43	12	56	5	4
4.5	12	16	20	31	56	3	7	2	2
10	13	12	12	42	7	5	6	134	7
11	7	13	6	5	8	12	4	150	5
12	14	15	7	7	3	4	18	7	2
3	18	15	8	12	12	87	2	12	12
8	12	14	4	13	1	3	5	12	5
13	3	17	12	12	4	15	5	3	23
17	5	12	10	11	8	8	12	5	45
1	9	3	12	10	12	2	13	7	12
2	12	10	14	2	9	13	6	6	6

Consider the above table to be revenue in lacs from various technologies(rows) in different domains (columns) for an IT firm.

Exercise 2…

INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Which BUs generate comparable revenue from every technology?									
10	23	12	14	10	13	123	12	8	2
20	2	14	19	13	43	12	56	5	4
4.5	12	16	20	31	56	3	7	2	2
10	13	12	12	42	7	5	6	134	7
11	7	13	6	5	8	12	4	150	5
12	14	15	7	7	3	4	18	7	2
3	18	15	8	12	12	87	2	12	12
8	12	14	4	13	1	3	5	12	5
13	3	17	12	12	4	15	5	3	23
17	5	12	10	11	8	8	12	5	45
1	9	3	12	10	12	2	13	7	12
2	12	10	14	2	9	13	6	6	6

Data bars for each of the columns separately.

Which BUs generate comparable revenue from every technology?

10	23	12	14	10	13	123	12	8	2
20	2	14	19	13	43	12	56	5	4
4.5	12	16	20	31	56	3	7	2	2
10	13	12	12	42	7	5	6	134	7
11	7	13	6	5	8	12	4	150	5
12	14	15	7	7	3	4	18	7	2
3	18	15	8	12	12	87	2	12	12
8	12	14	4	13	1	3	5	12	5
13	3	17	12	12	4	15	5	3	23
17	5	12	10	11	8	8	12	5	45
1	9	3	12	10	12	2	13	7	12
2	12	10	14	2	9	13	6	6	6

Gradient fill of green (min to max) for each of the columns separately.

Exercise 3

How is the firm performing from overall revenue perspective?									
10	23	12	14	10	13	123	12	8	2
20	2	14	19	13	43	12	56	5	4
4.5	12	16	20	31	56	3	7	2	2
10	13	12	12	42	7	5	6	134	7
11	7	13	6	5	8	12	4	150	5
12	14	15	7	7	3	4	18	7	2
3	18	15	8	12	12	87	2	12	12
8	12	14	4	13	1	3	5	12	5
13	3	17	12	12	4	15	5	3	23
17	5	12	10	11	8	8	12	5	45
1	9	3	12	10	12	2	13	7	12
2	12	10	14	2	9	13	6	6	6

Consider the above table to be revenue in lacs from various technologies(rows) in different domains (columns) for an IT firm.

Exercise 3…

INDIAN INSTITUTE OF MANAGEMENT BANGALORE

How is the firm performing from overall revenue perspective?									
10	23	12	14	10	13	123	12	8	2
20	2	14	19	13	43	12	56	5	4
4.5	12	16	20	31	56	3	7	2	2
10	13	12	12	42	7	5	6	134	7
11	7	13	6	5	8	12	4	150	5
12	14	15	7	7	3	4	18	7	2
3	18	15	8	12	12	87	2	12	12
8	12	14	4	13	1	3	5	12	5
13	3	17	12	12	4	15	5	3	23
17	5	12	10	11	8	8	12	5	45
1	9	3	12	10	12	2	13	7	12
2	12	10	14	2	9	13	6	6	6

Gradient fill from red (min of values) to green (max of the values)

Video: Best stats you have ever seen

Why Visualization

- Data visualization shifts the balance between seeing (perception) and thinking (cognition) to take maximum advantage of how brain functions.

- Studies in attention and memory have revealed that humans have limited ability to hold multiple items simultaneously in awareness.
  - Encoding information visually, allows more information to be chunked together into the limited slots available in working memory.
  - Several views of information in front of eyes at one time, extends ability to explore data from multiple dimension and from multiple perspectives.

More notes at: Data Visualization for human perception

Basic Points for Effective Visualization

- Human eye can read linear distances more effectively than circular distances.
- Human eyes are tuned to pick up red, green and blue colors instantly than any other color.
  - Coloring based on the gradient shades of green, blue or red brings more meaning to the data being represented.
- We live in a 3 dimensional space and thus are tuned to recognize 2 dimensional charts easily. But what after that?

> - First two dimensions can be visualized through co-ordinates
> - Color intensity may form the third dimension
> - Size or length may form the fourth dimension
> - Shape may form the fifth dimension
> - Texture, angle…

- Numbers after decimals may not be needed when analyzing large data set.
- 3D rendering of charts often complicates comparison as perspective skews relative shape and size.
- Legends in graphs with many options/colors to select becomes non-intuitive.

Descriptive Statistics and Data Visualization

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

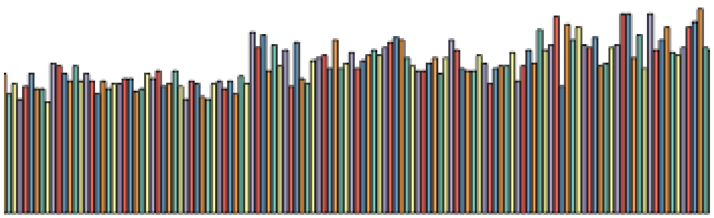
- Descriptive statistics is a field in analytics which caters to summarizing data and extracting information from the data.
- Data Visualization may form the building block for descriptive statistics.
- R provides the flexibility and robustness in data visualization. Some notable features of R which aids in data visualization are:

- Powerful environment for visualizing data
- Integrated graphics and statistics infrastructure
- Fully programmable and highly reproducible
- Vast number of R packages with graphics utilities

Data visualization is only successful to the degree to which it encodes information in a manner that our eyes can discern and our brains can understand.

Bar Charts

- Used to show comparison of quantities over different categorical variables in the dataset. Examples to generate bar chart from Iris dataset.

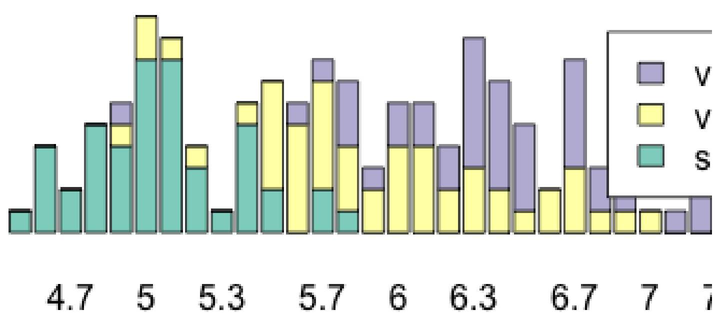


 # Simple bar charts . Uses graphics() library.

```
library(RColorBrewer)
barplot(iris$Sepal.Length,col =
brewer.pal(3,"Set1"))
```

#stacked bar charts

```
library(RColorBrewer)
barplot(table(iris$Species,iris$Sepal.Le
ngth),col = brewer.pal(4,"Set3"),
legend.text = TRUE)
```



Pie Charts

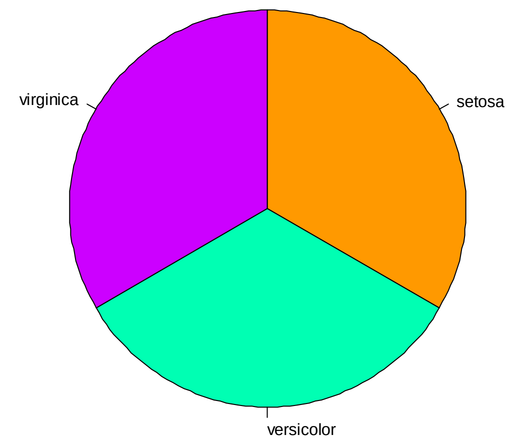
- Used to show share of categorical variables in the overall dataset. Examples to generate pie chart from Iris dataset.

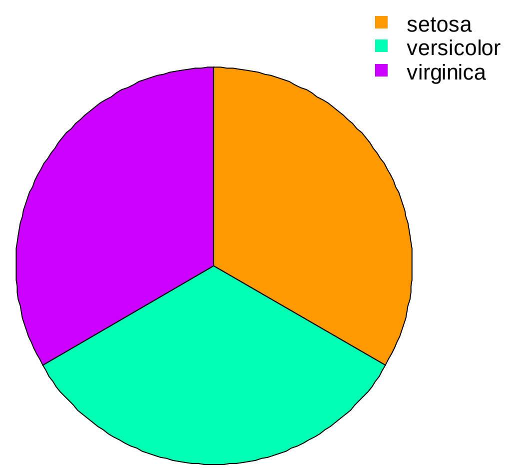
 # Plots a simple pie chart. Uses graphics() library.

```
y <- table(iris$Species)
pie(y, col=rainbow(length(y), start=0.1,
end=0.8), main="Pie Chart", clockwise=T)
```

#plot a pie chart with legends

```
pie(y, col=rainbow(length(y), start=0.1,
end=0.8), labels=NA, main="Pie Chart",
clockwise=T); legend("topright",
legend=row.names(y), cex=1.3, bty="n",
pch=15, pt.cex=1.8,
col=rainbow(length(y), start=0.1,
end=0.8), ncol=1)
```





- Pie chart may not be a useful way to represent any data.

Mosaic Plot

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT
BANGALORE

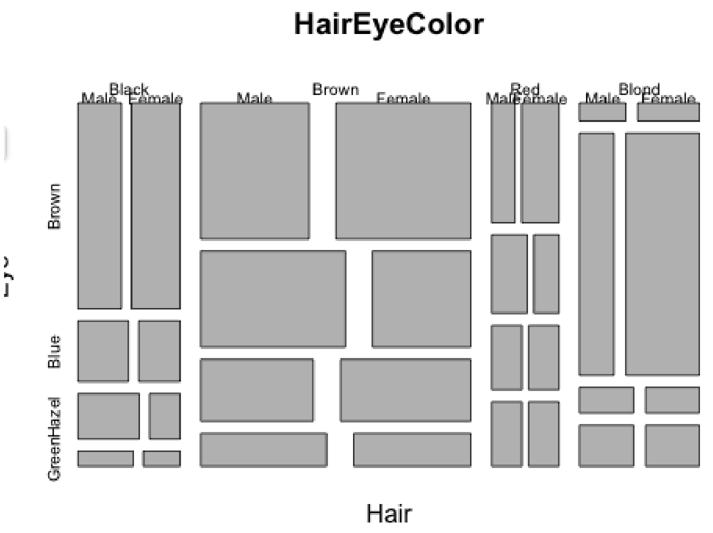
- Used for plotting large set of categorical data where area of the tile shows relative proportion. Examples to generate mosaic plot from Iris dataset.

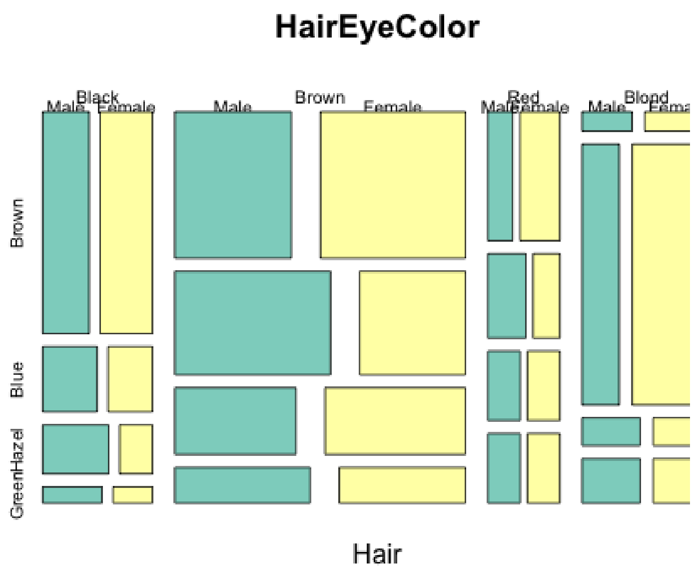
 # Mosaic plot without color. Uses graphics() library.

```
mosaicplot(HairEyeColor)
```

 # Mosaic plot with color.

```
library(RColorBrewer)
mosaicplot(HairEyeColor,col  =
brewer.pal(6,"Set3"))
```

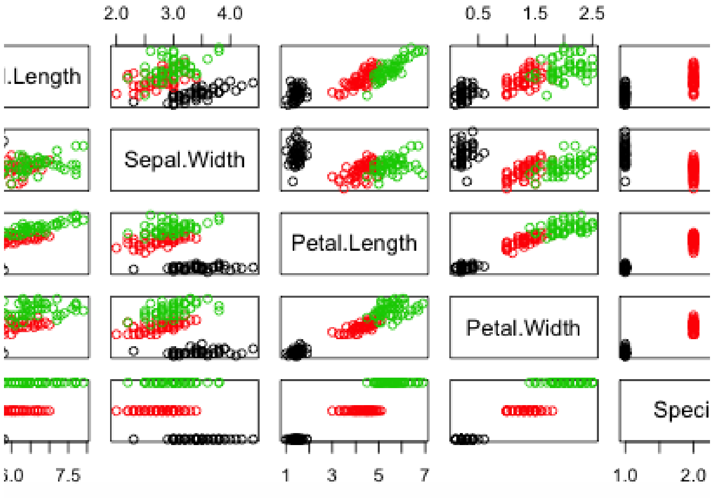


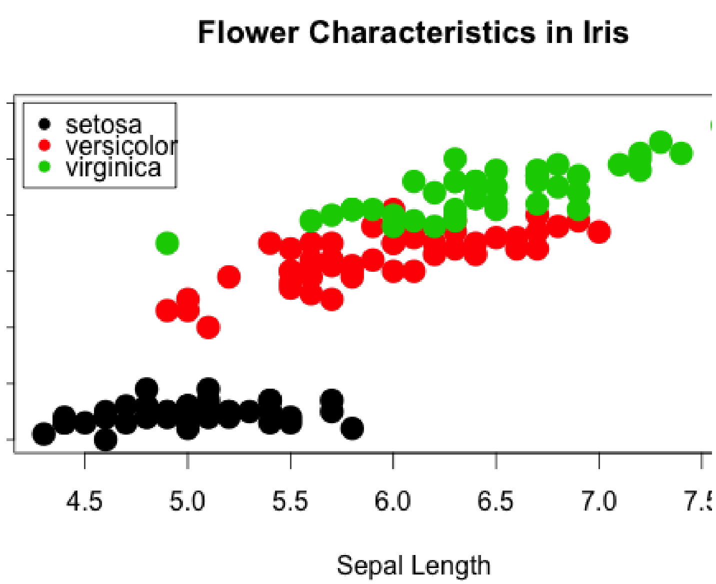


Pair plot or Scatter Plot

- Used to show joint variation of numeric data which can be segregated by categorical variable. Examples to generate pair plot from Iris dataset.

```
# scatter plot matrix with iris dataset. Uses graphics()
#library.
data(iris)
pairs(iris, col = iris$Species) #pair
plot with color
#plot of all variables with color
plot(iris$Sepal.Length, iris$Petal.Length,  # x & y variable
col = iris$Species,                         # color by species
pch = 16,                                   # type of point to use
cex = 2,                                    # size of point to use
xlab = "Sepal Length",                      # x axis label
ylab = "Petal Length",                      # y axis label
main = "Flower Characteristics in Iris")    # plot title
legend (x = 4.2, y = 7, legend = levels(iris$Species), col =
c(1:3), pch = 16)
```



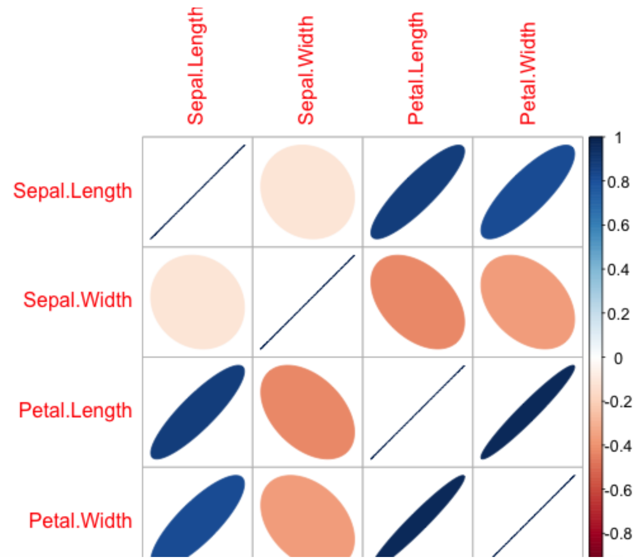


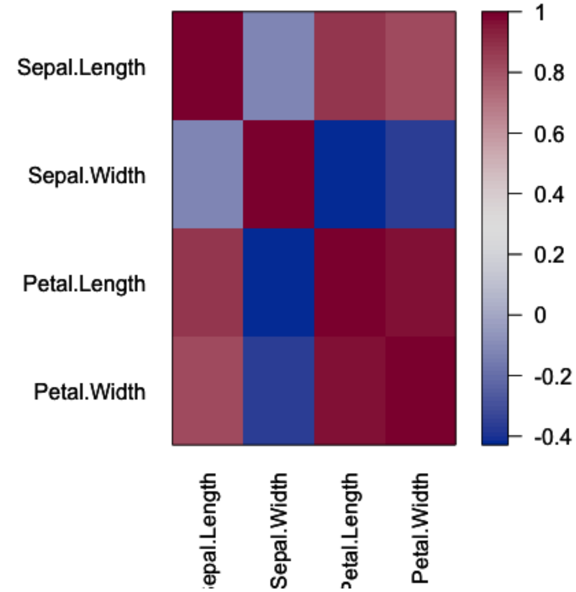
Correlation Plot

- Correlation plot shows the degree of variation between two numeric variables. Examples to generate correlation plot.

```
#correlation plot with iris dataset
library(corrplot)
iris_matrix <- as.matrix(iris[,1:4])
corrplot(cor(iris_matrix),
method="ellipse")

#correlation plot with a different library
library(seriation)
iris_matrix <- as.matrix(iris[,1:4])
pimage(cor(iris_matrix), colorkey=TRUE,
range=c(-1,1), col=diverge_hcl(100))
```





Histogram and Box Plot

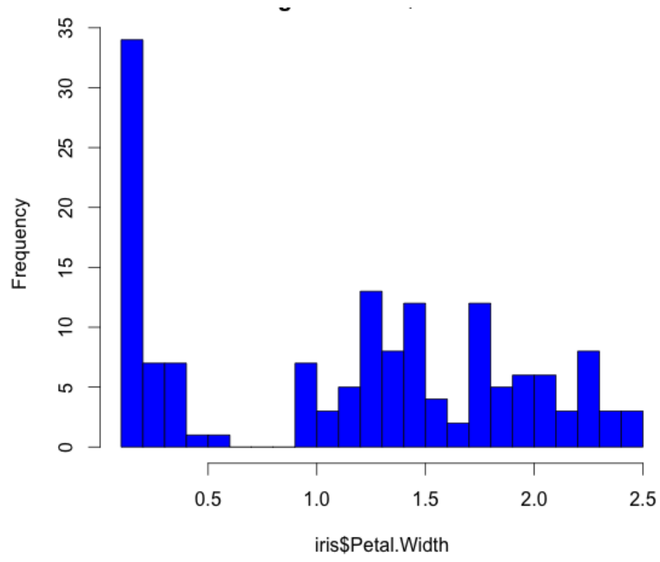
- Both used to summarize numeric data.
  - Histogram is used to bin the data and understand the underlying pattern in data.
  - Boxplot can be used to identify outliers in the dataset. Examples to generate histogram and box plot.

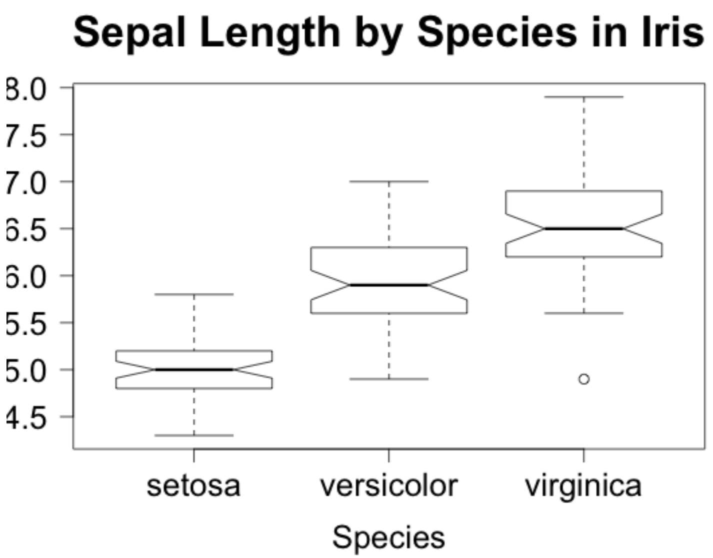
#histogram plot with iris dataset. Uses graphics() library.

```
hist(iris$Petal.Width, breaks=20,
col="blue")
```

#box plot of all variables

```
boxplot(iris$Sepal.Length ~ iris$Species,   #
x &y variable,
notch = T,    # Draw notch
las = 1,      # Orientate the axis tick labels
xlab = "Species",        # X-axis label
ylab = "Sepal Length",   # Y-axis label
main = "Sepal Length by Species in Iris",
cex.lab = 1.5,  # Size of axis labels
cex.axis = 1.5, # Size of the tick mark labels
cex.main = 2)   #Size of the plot title
```





Data Matrix Visualization

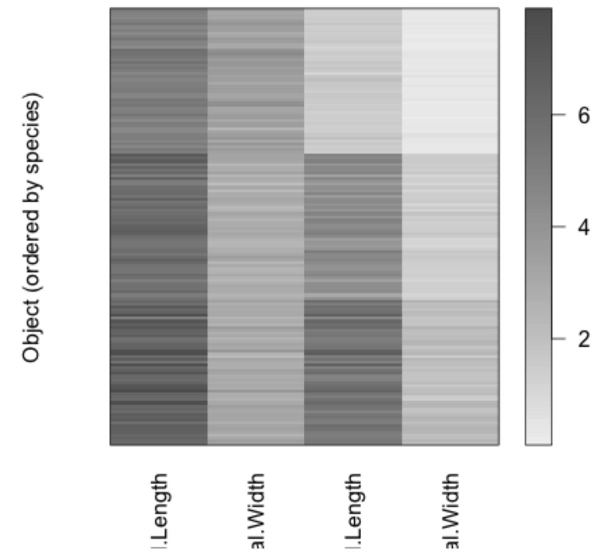
- Gradation of the color signifies the varied levels in the dataset. Examples to plot individual values of a dataset

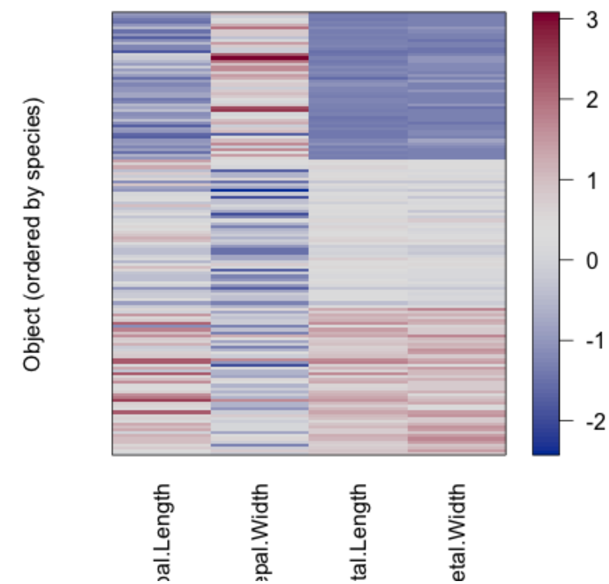
#plotting individual values of the iris dataset

```
library(seriation) #for pimage
iris_matrix <- as.matrix(iris[,1:4])
pimage(iris_matrix, ylab="Object (ordered by species)", main="Original values",
colorkey=TRUE)
```

#values smaller than the average are blue and larger ones are red

```
library("colorspace") ### for diverge_hcl
library(seriation) #for pimage
iris_matrix <- as.matrix(iris[,1:4])
pimage(scale(iris_matrix), ylab="Object (ordered by species)",
main="Standard deviations from the feature mean",
range=c(-3.5,3.5), col=diverge_hcl(100),
colorkey=TRUE)
```





Saving Graphs to Files

- Saving graphs to a file follows a specific sequence of commands. Below are some of the examples:

ε

 # Saving a jpeg file in the working directory. The actual image data are not written to the file #until the 'dev.off()' command is executed!

```
jpeg("test.jpeg"); plot(1:10, 1:10); dev.off()
```

 # Same as above, but for pdf format. The pdf format provides often the best image quality, #since it scales to any size.

```
pdf("test.pdf"); plot(1:10, 1:10); dev.off()
```

 # Same as above, but for png format.

```
png("test.png"); plot(1:10, 1:10); dev.off()
```

 # Same as above, but for PostScript format.

```
postscript("test.ps"); plot(1:10, 1:10); dev.off()
```

- The following options can be used inside the graph function to control text and symbol size in graphs.

option	description
cex	number indicating the amount by which plotting text and symbols should be scaled relative to the default. 1=default, 1.5 is 50% larger, 0.5 is 50% smaller, etc.
cex.axis	magnification of axis annotation relative to cex
cex.lab	magnification of x and y labels relative to cex
cex.main	magnification of titles relative to cex
cex.sub	magnification of subtitles relative to cex

Advanced Visualization Using R

- Many libraries in R which provides the capability of advanced data visualization.

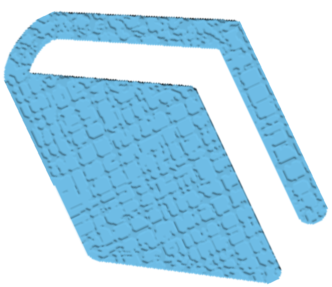
- **`tabplotd3()`** – *visualization for large dataset with both categorical and numeric variables*
- **`metricsgraphics()`** – *for advanced scatterplot*
- **`dygraphs()`** – *Time series plot with basic forecasting using holts winter technique*
- **`d3heatmap()`** – *heat map with clustering of similar groups*
- **`treemap()`** – *visualization of large dataset*
- **`networkd3()`** – *network graphs. Earlier it was d3network()*

More about networkd3() at: https://christophergandrud.github.io/networkD3/

Demo of Sales Dashboard

Summary

Summary of the topics covered in this lesson:



- Data visualization and Descriptive statistics goes hand in hand to summarize and extract useful information from data.
- R provides umpteen number of libraries which can be used to visualize any dataset.
- Scatter plot, box plot, histogram, correlation plot are some of the statistical plots useful in summarizing data.
- The graphs generated using the graph functions can be saved in different file formats using R commands.

QUIZ TIME

Quiz Question 1

Quiz 1

What will be plotted on x-axis and y-axis with the following command?
boxplot(iris$Sepal.Length ~ iris$Species)

a. Sepal Length on x-axis and Species on y-axis.

b. Sepal Length on y-axis and Species on x-axis.

c. Syntax incomplete. Graph will not be plotted.

d. Syntax complete but x and y axis plot not defined.

Quiz Question 1

Quiz 1

What will be plotted on x-axis and y-axis with the following command?
boxplot(iris$Sepal.Length ~ iris$Species)

a. Sepal Length on x-axis and Species on y-axis.

b. Sepal Length on y-axis and Species on x-axis.

c. Syntax incomplete. Graph will not be plotted.

d. Syntax complete but x and y axis plot not defined.

Correct answer is: ***b***

The first parameter in the boxplot represents y-axis variable and second parameter represents x-axis variable.

End of Lesson04–Data Visualization using R

© Copyright 2015 ARIMA RESEARCH LLP. All rights reserved.

भारतीय प्रबंध संस्थान बेंगलूर

INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Data Science Using R

Lesson05–Understanding Data Attributes

Objective

After completing this lesson you will be able to:

- Understand the building blocks of statistics
- Describe the location, dispersion and shape attributes of a data through the use of sample cases

Case: Types of Data variables

Romanov, an Analytics consultant works with Credit One bank. His manager gave him a list having the name of bank's customers. Further he has been asked to pull the information from bank's database pertaining to the customer list. The information will be around the credit cards issued by the bank. He needs to define the variable types and the type of value each one of them will contain. Romanov, who has just started his professional career, doesn't has a good idea about different variable types.

Now, suppose after extracting data he approached you and asked your help in categorizing the different variables. Help Romanov in variable categorization.

Case: Types of Data variables (Data snapshot)

Sl No	Name of Customer	Customer ID	Number of Credit Cards	Age of Customer (Last Birthday)	Gender of the Customer	Marital Status of the Customer	Annual Salary (in USD)	Monthly Credit Card Usage
1	Josh	111669	5	42	F	Never Married	88,001	Low
2	Janice	146861	6	25	F	Married	592,489	Low
3	Dandre	171690	3	50	M	Divorced	272,304	Low
4	Aiden	161721	6	37	M	Married	726,593	Low
5	Celine	170359	7	50	F	Never Married	612,075	Low
6	Emilio	175646	5	41	M	Never Married	490,356	Low
7	Joaquin	180732	2	62	F	Divorced	164,732	Low
8	Justus	113136	7	26	F	Never Married	510,321	Low
9	Chaya	169254	4	24	M	Never Married	358,534	Low
10	Justyn	149771	4	35	M	Married	140,400	Low
11	Jadon	166226	7	36	M	Never Married	105,259	Low

Case: Types of Data variables

Information to be extracted by Romanov

Variable Name	Name of Customer	Customer ID	Number of Credit Cards	Age of Customer Last Birthday	Gender of Customer	Marital Status of Customer	Annual Salary	Monthly Credit Card Usage
Value Stored	?	?	?	?	?	?	?	?
Variable Type	?	?	?	?	?	?	?	?
Remarks								

Case: Types of Data variables

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Information to be extracted by Romanov

Variable Name	**Name of Customer**	**Customer ID**	**Number of Credit Cards**	**Age of Customer Last Birthday**	**Gender of Customer**	**Marital Status of Customer**	**Annual Salary**	**Monthly Credit Card Usage**
Value Stored	Name of the individual customer	Unique identifier	1, 2, 3…	18, 19, 20…	Male / Female	Married / Divorced / Never Married	Amount	Low(<25%) / Medium(<50%) / High(<75%) / Very High(>75%)
Variable Type	?	?	?	?	?	?	?	?
Remarks								

Types of Data Variables

Data consists of a combination of "variables" which actually contain the values. Variables at a high level are of two types depending on the kind of values they store:

Numerical variables

Discrete

- Arises from counting. Can take only a set of particular values including negative and fractional values
- Examples: Credit score, number of credit cards owned by a person, number of states in a country, charge on electron etc.

Continuous

- Arises from measuring. Can take any value with in a specified range
- Examples: Height, Amount of money, Age etc.

Categorical variables

Binary (or Dichotomous)

- Has only two categories
- Examples: yes/no, male/female, pass/fail etc.

Nominal

- Has several unordered category
- Examples: Type of bank account, type of insurance policy etc.

Ordinal

- Has several ordered category
- Examples: questionnaire responses such as "strongly in favour / … / strongly against".

Case: Types of Data variables (Revisited)

INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Information to be extracted by Romanov

Variable Name	**Name of Customer**	**Customer ID**	**Number of Credit Cards**	**Age of Customer Last Birthday**	**Gender of Customer**	**Marital Status of Customer**	**Annual Salary**	**Monthly Credit Card Usage**
Value Stored	Name of the individual customer	Unique identifier	1, 2, 3…	18, 19, 20…	Male / Female	Married / Divorced / Never Married	Amount	Low(<25%) / Medium(<50%) / High(<75%) / Very High(>75%)
Variable Type	--	--	Numerical (Discrete)	Numerical (Discrete)	Categorical (Binary)	Categorical (Nominal)	Numerical (Continuous)	Categorical (Ordinal)
Remarks	Identifier	Identifier	Arises from counting. Takes certain discrete values in a given range	Arises from counting. Takes certain discrete values in a given range	Only two categories	Several ordered category	Takes many values in a given range	Several ordered category

Case: Summarizing Data

Romanov, an Analytics consultant works with Credit One bank. His manager gave him some data around credit cards relating to number of credit cards issued to a set of customers and the credit limit of the cards. Further he has been tasked to summarize the data in a presentable form and prepare the report. Romanov, who has just started his professional career, has never played around with such kind of data, so he is clueless about the different summarizing techniques.

Now, suppose he approached you and asked your help in preparing the report. Help Romanov in summarizing the data and preparing the report.

Comments: Summarizing Data

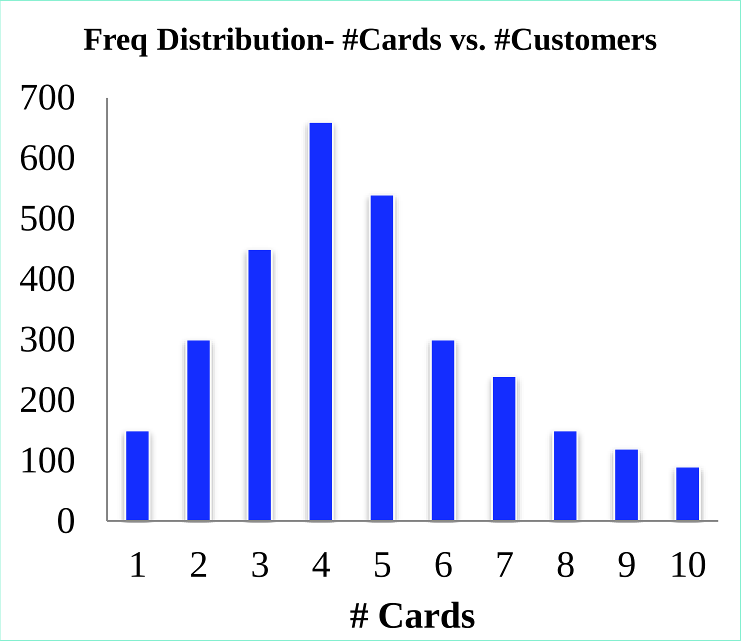
There are various ways to summarize data. Some of them are

- Frequency distribution
- Grouped frequency distribution
- Cumulative frequency distribution
- Stem leaf diagram
- Line plots

Summarizing Data–Frequency distribution

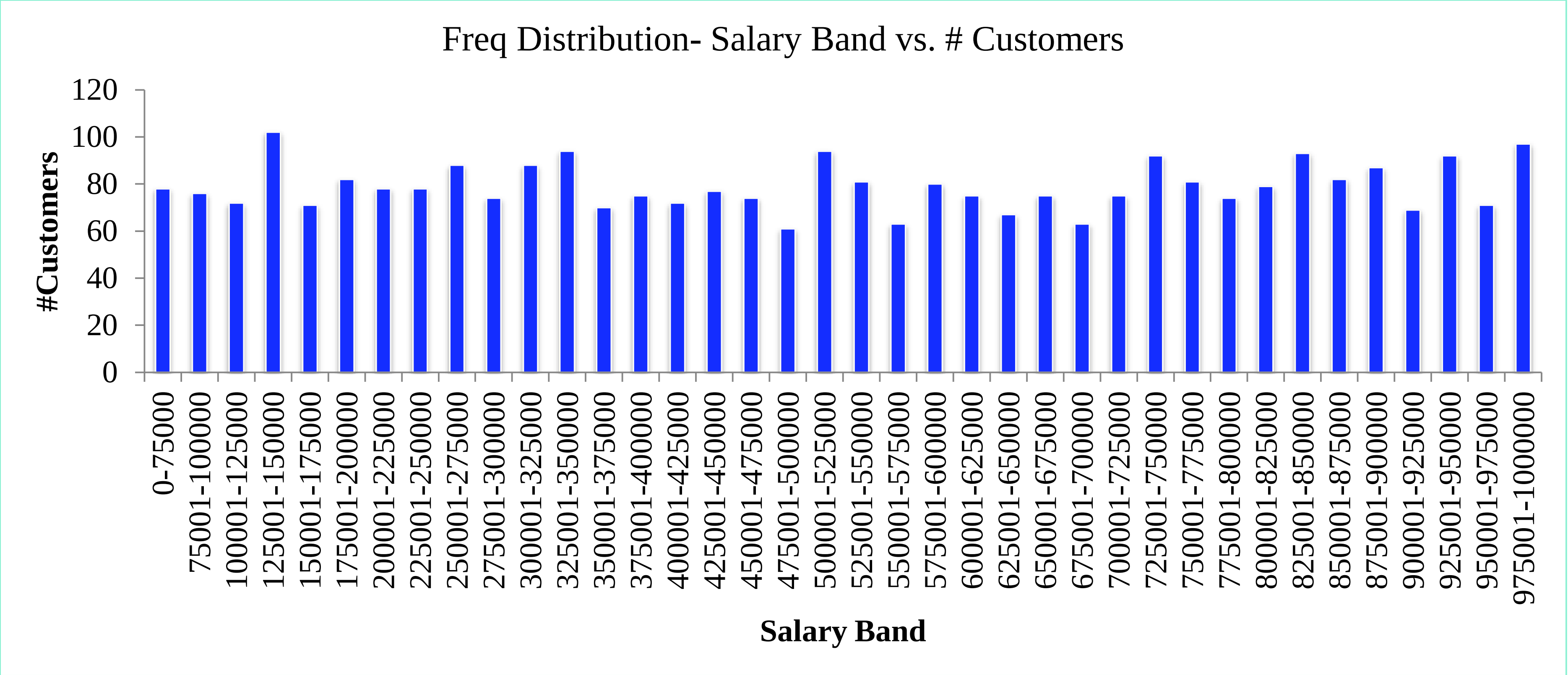
- A technique to summarize discrete data
- A simple process which involves counting of distinct discrete values
- The representation can be either tabular or graphical
- Example: Number of credit cards owned in a sample of 3000 individuals

Number of Credit Cards	# Customers
1	150
2	300
3	450
4	660
5	540
6	300
7	240
8	150
9	120
10	90



Summarizing Data–Grouped Frequency Distribution

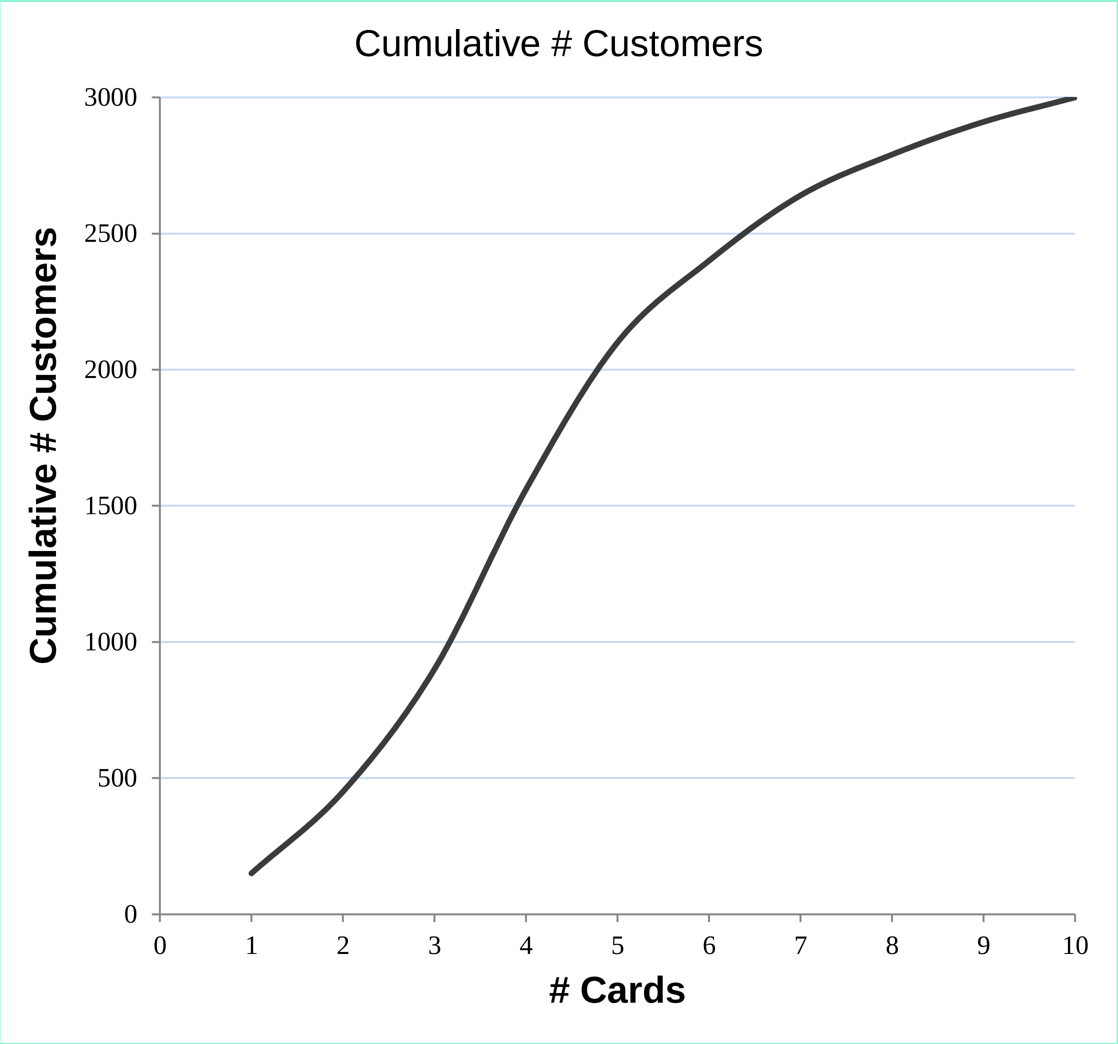
- A technique to summarize continuous data or discrete data having large number of observations and an extended range
- A simple process which involves counting of values falling under the different intervals (grouped)
- Example: Number of customers falling under different Salary groups



Summarizing Data–Cumulative Frequency Distribution

- Cumulative frequencies are obtained by accumulating the frequencies to give the total number of observations up to and including the value or group in question.
- Example: Cumulative number of cards in the sample of 3000 individuals

Number of Credit Cards Up to	Cumulative # Customers
1	150
2	450
3	900
4	1560
5	2100
6	2400
7	2640
8	2790
9	2910
10	3000



Case: Measure of Central Tendency/Location

After Romanov presented the summarized data to his manager at Credit One, he was asked to produce the various measures of Central Tendency of the Credit Card data.

Now, Romanov being unaware of the term "central tendency" again approached you and asked your help in calculating the central tendency of the data in question. Help Romanov in carrying out his task.

Measure of Central Tendency/Location

- There are a number of different quantities, which can be used to estimate the central point of a sample.
- These are called measures of central tendency or measures of location.
- Just different ways of calculating the "average" value of dataset.

Three ways to summarize the central tendency
- Mean
- Median
- Mode

Measure of Central Tendency/Location

- Mean or average of a list of values is given by:

Mean = Sum of values/Count of values

- Median is that value which splits list of numbers into two equal halves. Median of a list of numbers is calculated after sorting the numbers in increasing/ascending order:

Count of numbers is odd: Median is the middle value
Count of numbers is even: Median is the sum of two middle values divided by 2

- Mode is the value in the list of numbers which occurs most frequently. For ease, sort the value in increasing/ascending order:

Count the value which occurs most number of times.

Measure of Central Tendency/Location

The measure of central tendency for the customer's age:

Sl No	Name of Customer	Age of Customer (Last Birthday)	Mean	Median	Mode
1	Josh	42	39	37	50
2	Janice	25			
3	Dandre	50			
4	Aiden	37			
5	Celine	50			
6	Emilio	41			
7	Joaquin	62			
8	Justus	26			
9	Chaya	24			
10	Justyn	35			
11	Jadon	36			

Case: Measure of Spread

After Romanov presented the summarized data along with "measures of Central tendency" to his manager at Credit One, he was further asked to add the various measures of spread to the report.

Now, Romanov being unaware of the term "measures of spread" again approached you and asked for your help. Help Romanov in carrying out his task.

Measure of Spread

- The central tendency of a data set is usually the main feature of interest. But another feature of interest is the spread (or variability or dispersion or scatter).
- Spread determines how widely scattered the data is about the mean (or other measure of location).

Three ways to summarize the spread are:

- Variance and Standard Deviation
- The Range
- The Inter quartile range

Measure of Spread

- Standard deviation is a measure to show how far on average the observations are from the mean.
- The range is a measure to show the spread as a difference between the largest and smallest observations in the data set.

Range of a dataset = (Max value in dataset –Min value in dataset)

- Interquartile range is a measure of spread and is calculated based on the quartiles of a data set.
  - Quartile divides the data set into 4 quarters and is denoted by Q1, Q2 and Q3.
  - Interquartile range is given by Q3-Q1.
  - Use Quartile function in excel to compute the quartile values.

- Standard deviation is the most commonly used metric for measure of spread.
- Range is a poor measure of spread as it relies on the extreme values.
- Interquartile range is similar to range but is not affected by extreme values.

Measure of Spread

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

The measure of dispersion for the customer's annual salary (Mean salary is USD 369, 188):

Sl No (N)	Name of Customer	A= Annual Salary (in USD)	Deviation D=(A-Mean)	E=Square(D)	Variance V= Sum(E)/N	Standard Deviation= SQRT(V)	Range (Min-Max)	IQR = Q3-Q1
1	Josh	88,001	-281,187	79,065,924,469				
2	Janice	592,489	223,301	49,863,499,002				
3	Dandre	272,304	-96,884	9,386,438,995				
4	Aiden	726,593	357,405	127,738,593,956				
5	Celine	612,075	242,887	58,994,271,414				
6	Emilio	490,356	121,168	14,681,772,346	47,596,985,579	218,167	638,592	398,839
7	Joaquin	164,732	-204,456	41,802,107,241				
8	Justus	510,321	141,133	19,918,626,331				
9	Chaya	358,534	-10,654	113,499,968				
10	Justyn	140,400	-228,788	52,343,782,553				
11	Jadon	105,259	-263,929	69,658,325,093				

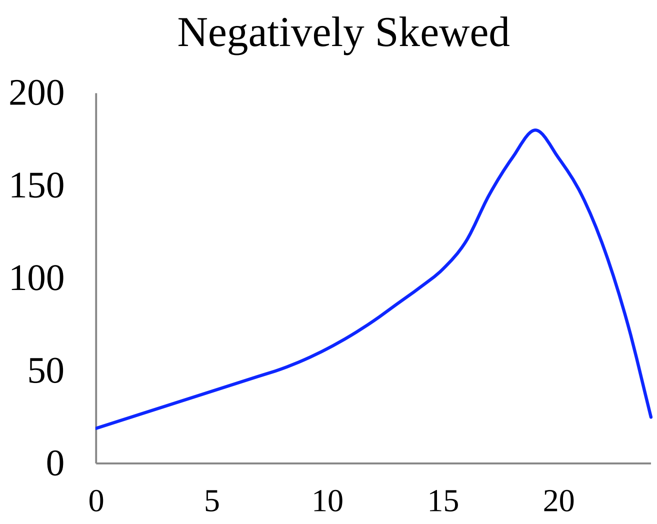
Case: Symmetry and skewness of data

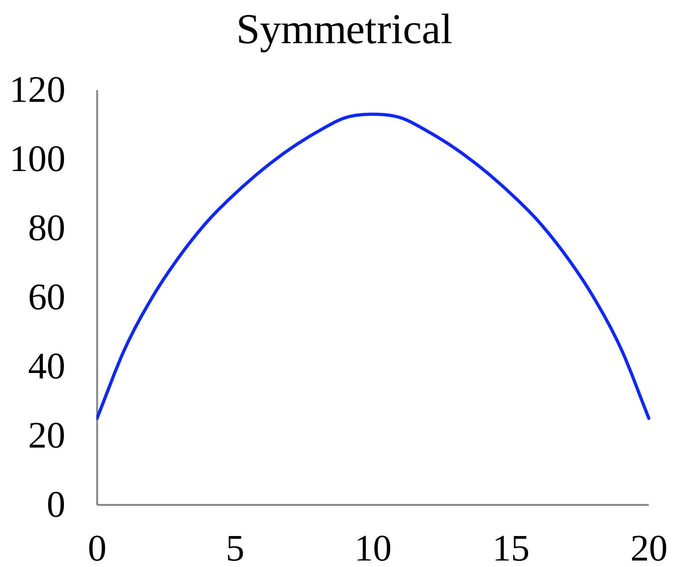
Romanov got appreciations after he presented the summarized data along with "measures of Central tendency" and "measure of spread" to his manager at Credit One. But, he was further asked to create an illustration around symmetry and skewness of data. Following that carry out the analysis of credit card data

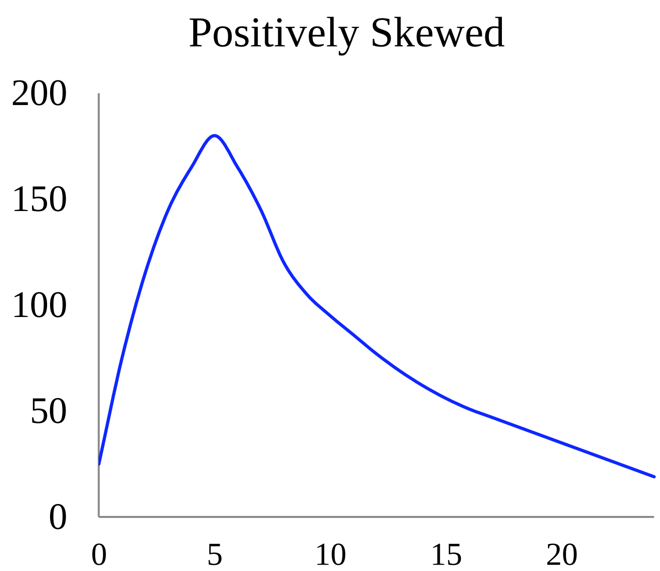
Now, Romanov being unaware of the term "symmetry and skewness" again approached you and asked for your help. In return he promised to gift you a bottle of Champagne. Help Romanov in carrying out his task.

Symmetry and skewness

- Symmetry and skewness deals with the shape of the distribution of a data set.
- The approximate shape of a distribution can be determined by looking at a histogram.
- Density plot is a better representation to analyze the shape of the distribution







Symmetry and Skewness

The measure of shape for the customer's age:

<table>
<tr><th>Sl No</th><th>Name of Customer</th><th>Age of Customer (Last Birthday)</th><th>Mean</th><th>Median</th><th>Mode</th></tr>
<tr><td>1</td><td>Josh</td><td>42</td><td rowspan="11">39</td><td rowspan="11">37</td><td rowspan="11">50</td></tr>
<tr><td>2</td><td>Janice</td><td>25</td></tr>
<tr><td>3</td><td>Dandre</td><td>50</td></tr>
<tr><td>4</td><td>Aiden</td><td>37</td></tr>
<tr><td>5</td><td>Celine</td><td>50</td></tr>
<tr><td>6</td><td>Emilio</td><td>41</td></tr>
<tr><td>7</td><td>Joaquin</td><td>62</td></tr>
<tr><td>8</td><td>Justus</td><td>26</td></tr>
<tr><td>9</td><td>Chaya</td><td>24</td></tr>
<tr><td>10</td><td>Justyn</td><td>35</td></tr>
<tr><td>11</td><td>Jadon</td><td>36</td></tr>
</table>

Symmetrical: Mean = Median = Mode
Positively Skewed: Mean > Median > Mode
Negatively Skewed: Mean < Median < Mode

Summary

Summary of the topics covered in this lesson:

- Data behavior is explained through location, spread and shape or distribution of the data.
- Mean, Median and Mode are the three attributes which explains the location or central tendency of the data.
- Standard deviation is a measure to understand the spread of the data. This is the most commonly used attribute apart from range.
- Shape of the data is explained by histogram plot. However, histogram may be misleading in understanding the shape/distribution of data. Density plot is a better representation to understand data distribution.

QUIZ TIME

Quiz Question 1

Quiz 1	What are the attributes to understand the central tendency of data? *Select all that apply.*

a. Mean

b. Variance

c. Median

d. Standard deviation

Quiz Question 1

Quiz 1	What are the attributes to understand the central tendency of data? *Select all that apply.*

a. Mean

b. Variance

c. Median

d. Standard deviation

Correct answer is:

a & c

Mean and Median are the two attributes to understand central tendency. The other attribute is Mode.

End of Lesson05–Understanding Data Attributes

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Data Science Using R

Lesson06–Data Pre processing

Objective

After completing this lesson you will be able to:

- Describe the importance of data pre-processing and its impact on the analysis
- Understand the various techniques of data pre-processing

Why Data Preprocessing?

Data in the real world is dirty

- incomplete: missing attribute values, lack of certain attributes of interest, or containing only aggregate data
  - e.g., occupation="“”"
- noisy: containing errors or outliers
  - e.g., Salary="-10"
- inconsistent: containing discrepancies in codes or names
  - e.g., Age="42" Birthday="03/07/1997"
  - e.g., Was rating "1,2,3", now rating "A, B, C"
  - e.g., discrepancy between duplicate records

Why Is Data Preprocessing Important?

No quality data, no quality results!

- Quality decisions must be based on quality data
- e.g., duplicate or missing data may cause incorrect or even misleading statistics.

Data preparation, cleaning, and transformation comprises the majority of the work in a data analytics project (~60%).

Major Tasks in Data Preprocessing

- Data integration
  - Integration of multiple databases, or files

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies

- Data transformation
  - Normalization and aggregation

Data Cleaning

Data cleaning tasks

- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data

Data Cleaning–Missing Data

Data is not always available

- E.g., many tuples have no recorded values for several attributes, such as customer income in sales data

Missing data may be due to

- equipment malfunction
- inconsistent with other recorded data and thus deleted
- data not entered due to misunderstanding
- certain data may not be considered important at the time of entry
- not registered history or changes of the data

Missing data may need to be inferred.

Data Cleaning–Handling Missing Data

- Ignore the tuple: usually done when class label is missing
  - assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., “unknown”, a new class?
- Use the attribute mean to fill in the missing value
- Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree

Data Cleaning–Missing Value Imputation

There are a variety of techniques for missing value imputation; but these should be considered more as scenario-specific than just being a set of pure alternative choices.

There are several missing value imputation techniques:

- Impute Missing Values with ZERO
- Impute Missing Values with MEDIAN
- Impute Missing Values with MEAN
- Impute Missing Values with MODE
- Information based Segmentation
- Impute using Regression on other Non-Missing Predictors
- Logical imputation

Data Cleaning–Noisy Data

Noise: random error or variance in a measured variable

Incorrect attribute values may due to

- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation
- inconsistency in naming convention

Other data problems which requires data cleaning

- duplicate records
- incomplete data
- inconsistent data

Data Cleaning–Handling Noisy Data

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

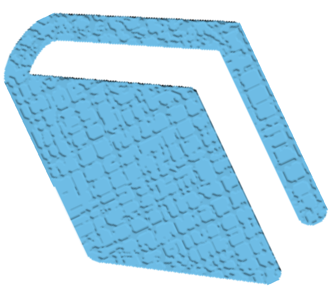
- Clustering
  - detect and remove outliers

- Combined computer and human inspection
  - detect suspicious values and check by human

- Regression
  - smooth by fitting the data into regression functions

Summary

Summary of the topics covered in this lesson:



- Data preparation is a time taking activity and majority of the time in an analytics projects is typically spent in this phase.
- There are several techniques available to improve the quality of the data i.e. data completeness and data consistency.

QUIZ TIME

Quiz Question 1

Quiz 1	What are the typical reasons for missing data?

a. Data not entered due to misunderstanding.

b. Certain data may not be considered important at the time of entry.

c. Inconsistent with other recorded data and thus deleted.

d. All the above.

Quiz Question 1

Quiz 1	What are the typical reasons for missing data?

a. Data not entered due to misunderstanding.

b. Certain data may not be considered important at the time of entry.

c. Inconsistent with other recorded data and thus deleted.

d. All the above.

Correct answer is:

d

There can be many more reasons for missing data but all the above factors into those reasons as well.

End of Lesson06–Data Pre processing

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Data Science Using R

Lesson07–Basic of Statistics

Objective

After completing this lesson you will be able to:

- Explain the basic concepts of statistics
- Understand the application of these concepts in statistical modelling

Population and Sample

The objective of sampling and further analysis on the sampled data is to understand the population parameter.

Population

- Includes all elements from a set of data
- Measurable characteristic is called parameter (μ, σ)
- Costly to collect e.g. customer satisfaction

Sample

- One or more observations from the dataset
- Measurable characteristic is called statistics $(\bar{x}, s)$
- Effective and cheaper way to understand the population parameter

The objective of sampling and further analysis on the sampled data is to understand the population parameter. The sample must be random in order to use statistics to learn things about the population

Study of data

Three basic components of business analytics

Descriptive analytics	Predictive analytics	Prescriptive analytics
• Condenses data into smaller, useful nuggets of useful information. • Looks at past performance and finds reasons for success or failures.	• Uses variety of statistical, modeling, data mining, and machine learning techniques to study recent and historical data	• Use optimization and simulation algorithms to advice on possible outcomes. • Suggests decision options and continuously takes new data to re-predict and re-prescribe. • Transition is fuzzy.
• What happened and why did it happen?	• What might happen?	• What should we do?
• Reports that provide insights into finance, operations, sales etc.	• Sentiment analysis, credit scoring, predicting what items customer will buy together etc.	• Optimize production, recommendation engines etc.

Inferential statistics: enables to make an educated guess about a population parameter based on a statistic computed from a sample randomly drawn from that population.

Central Limit Theorem

- Generally, the sample mean ($\bar{X}$) derived in repeated sampling from a normally distributed population with mean μ and standard deviation σ will follow a normal distribution with mean $\bar{X} = \mu$ and standard deviation $SD(\bar{X}) = \frac{\sigma}{\sum n}$ for any sample size n.

- Central limit theorem: The sample mean ($\bar{X}$) derived in repeated sampling from a population with mean μ and standard deviation σ will follow a normal distribution with mean $\bar{X} = \mu$ and standard deviation $SD(\bar{X}) = \frac{\sigma}{\sum n}$ for large sample size n > 30.

Demonstrate CLT through an example

Confidence Interval concept

- If sample is drawn from a normal population or a large sample is used, then by the rules of normal distribution, before the sampling, there is .95 probability that sample mean $\bar{X}$ will fall within the interval

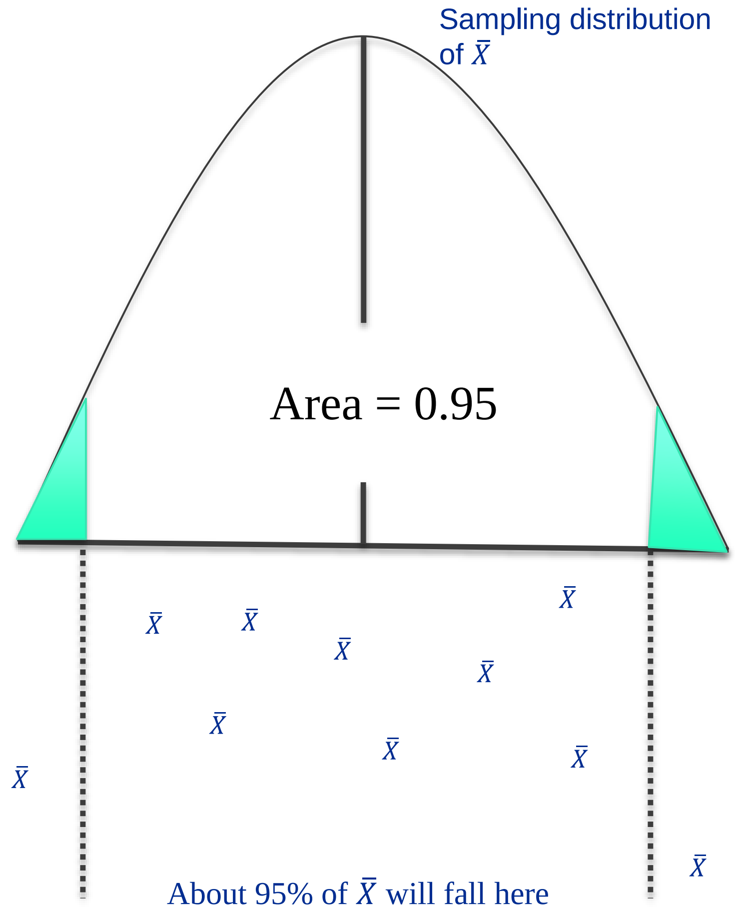
$$\mu \pm 1.96 * \frac{\sigma}{\sum n}$$

- After sampling, about 95% of the values of $\bar{X}$ obtained in large number of repeated sampling will fall in the interval defined by equation above.

- $\bar{X}$ falls within the interval defined above if and only if μ happens to be within

$$\bar{X} \pm 1.96 * \frac{\sigma}{\sum n}$$

- 95% confident that population mean lies within the above range. CLT in action.



Degree of Freedom

Objective is to estimate the population parameters (mean, std deviation etc.) from the sample. The degrees of freedom (df) of an estimate is the number of independent pieces of information on which the estimate is based.

Population mean (height in ft)= 6

- Sampling of one person: height is 8 ft
  - Variance = (8-6)^2 = 4

This estimate is based on one piece of information, so df = 1

- Sampling of one person: height is 5 ft
  - Variance = (5-6)^2 = 1

This estimate is based on one piece of information, so df = 1

- Population variance = 2.5 with $df = 2$

Population mean not known

- Sampling lead to 8ft and 5ft as two data points
  - Mean = 6.5
  - Variance estimate 1 = (8-6.5)^2 = 2.25
  - Variance estimate 2 = (5-6.5)^2 = 2.25
- The two estimates are not independent so $df \neq 2$ but $df = 1$

The degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated enroute to the estimate in question.

Hypothesis Testing

- The average price of a stock for the last 1 year for company XYZ is Rs. 1000. Randomly the stock price for 30 days are picked up and the average comes to Rs. 900. What can be concluded about the experiment?
  - These 30 stock prices are different from the stock prices of XYZ and thus there average performance is poor. May be it is from a different population.
  - There is no difference and it is due to random chance.

We can take any one of the following action:

- *Increase sample size and test again*
- *Test for another samples*
- *Calculate random chance probability*

- What is the probability that the sample would have an average stock price of Rs 900?

  - If the population distribution is normal, the characteristics of normal distribution can be directly applied to calculate the Z score and the probability value.
  - If the population is not normal, Central limit theorem can be applied to calculate the Z score and the probability value.

- What is the random chance probability comes to 40%? What if it is 2%?

  - Significance level (denoted by α) is a threshold value which helps to decide whether the random chance probability is due to pure chance or not.
  - If random chance probability is less than 5% (α being 5%), it can be concluded that stock price average of 1000 is from a different population than the sample of 30 stock prices whose average is Rs.900

Hypothesis Testing –Steps

1. **Set up the Hypothesis**
   - Null Hypothesis (H_0) –There is no difference between the sample and population behavior
   - Alternate Hypothesis (H_a) – There is a significant difference between sample and population behavior
2. **Set the Criteria for decision**
   - Define the level of significance at which decision would be made. Generally it is set at 5% but may be changed based on the business context.
3. **Compute the random chance probability**
   - Computed based on the formula. All software packages will report this probability. Higher probability has higher likelihood and enough evidence to accept the Null hypothesis.
4. **Make decision**
   - Compare p value with predefined significance level and if it is less than significance level, we reject Null hypothesis.

$$Rule: When\ p\ is\ less\ than\ \alpha, reject\ H_0$$

Covariance and Correlation Coefficient

- **Covariance** is a statistical measure of the degree to which the two variables move together. The sample covariance is calculated as :

$$\text{cov}_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

- **Correlation** coefficient is a measure of the strength of the linear relationship between two variables. The correlation coefficient is given by:

$$r_{xy} = \frac{\text{cov}_{xy}}{\sigma_x \sigma_y}$$

- Population correlation is denoted by ρ (rho). Sample correlation is denoted by r. Features of ρ and r
  - Unit free and ranges between -1 and 1
  - The closer to -1, the stronger the negative linear relationship
  - The closer to 1, the stronger the positive linear relationship
  - The closer to 0, the weaker the linear relationship

Y (Exp)	X (Inc)	Y' = Y-Y(Avg)	X' =X-X(Avg)	X'*Y"
700	800	-410	-900	369000
650	1000	-460	-700	322000
900	1200	-210	-500	105000
950	1400	-160	-300	48000
1100	1600	-10	-100	1000
1150	1800	40	100	4000
1200	2000	90	300	27000
1400	2200	290	500	145000
1550	2400	440	700	308000
1500	2600	390	900	351000
1110	**1700**			2E+06

Covaraince	186666.6667
Correlation	0.980847369

Partial and Semi-Partial Correlation

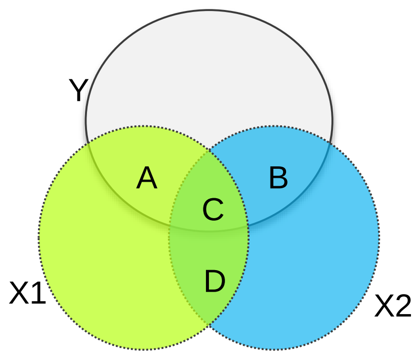
भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

- **Partial correlation** coefficient measures the relationship between two variables (say Y and X1) when the influence of all other variables (say X2, X3, …, Xn) connected with these two variables (Y and X1) are removed.

$$r_{12,3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Correlation between y1 and x2, when the influence of x3 is removed from both y1 and x2.

- **Part correlation** (or semi partial) coefficient measures the relationship between two variables (say Y and X1) when the influence of all other variables (say X2, X3, …, Xn) connected with these two variables (Y and X1) are removed from one of the variables (X1).



$$sr_{12,3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{23}^2}}$$

R^2			$1 - R^2$
Variance specific to X1 (part A)	Variance common to X1 & X2 (part C)	Variance specific to X2 (part B)	Left over

R is an official part of the Free Software Foundation's GNU project.
RStudio and Shiny are affiliated projects of the Foundation for Open Access Statistics.

Summary

Summary of the topics covered in this lesson:

- We always deal with sample dataset and the objective is to give a meaningful estimate of the population parameters through sample statistics.
- Central limit theorem is the building block for interpreting the outcome of many advanced statistical techniques.

QUIZ TIME

Quiz Question 1

Quiz 1 The population mean is known and 11 people from the population are selected at random to estimate the standard deviation. DF of standard deviation will be:

a. 11

b. 9

c. 10

d. None of the above

Quiz Question 1

Quiz 1	The population mean is known and 11 people from the population are selected at random to estimate the standard deviation. DF of standard deviation will be:

a. 11

b. 9

c. 10

d. None of the above

Correct answer is: *a*

Since population mean is known, there is no intermediate estimate to arrive at the estimate of standard deviation. Hence degree of freedom will be 11.

End of Lesson07–Basic of Statistics

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Data Science Using R

Lesson08–Regression Concepts

Objective

After completing this lesson you will be able to:

- Explain Regression analysis
- Describe the assumptions of linear regression
- Explain the need of transformations of data
- Understand the representation of qualitative variables in linear regression

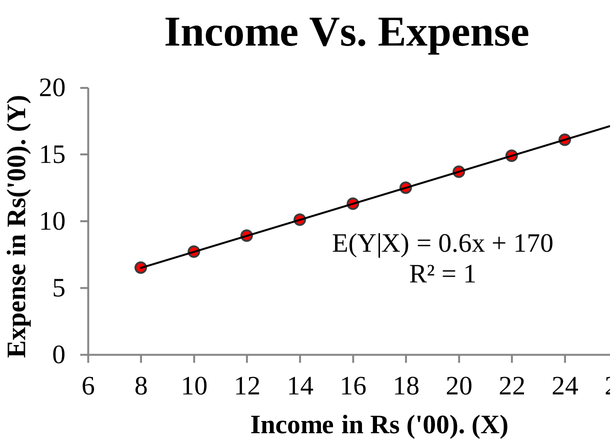
Regression–Building the Concept

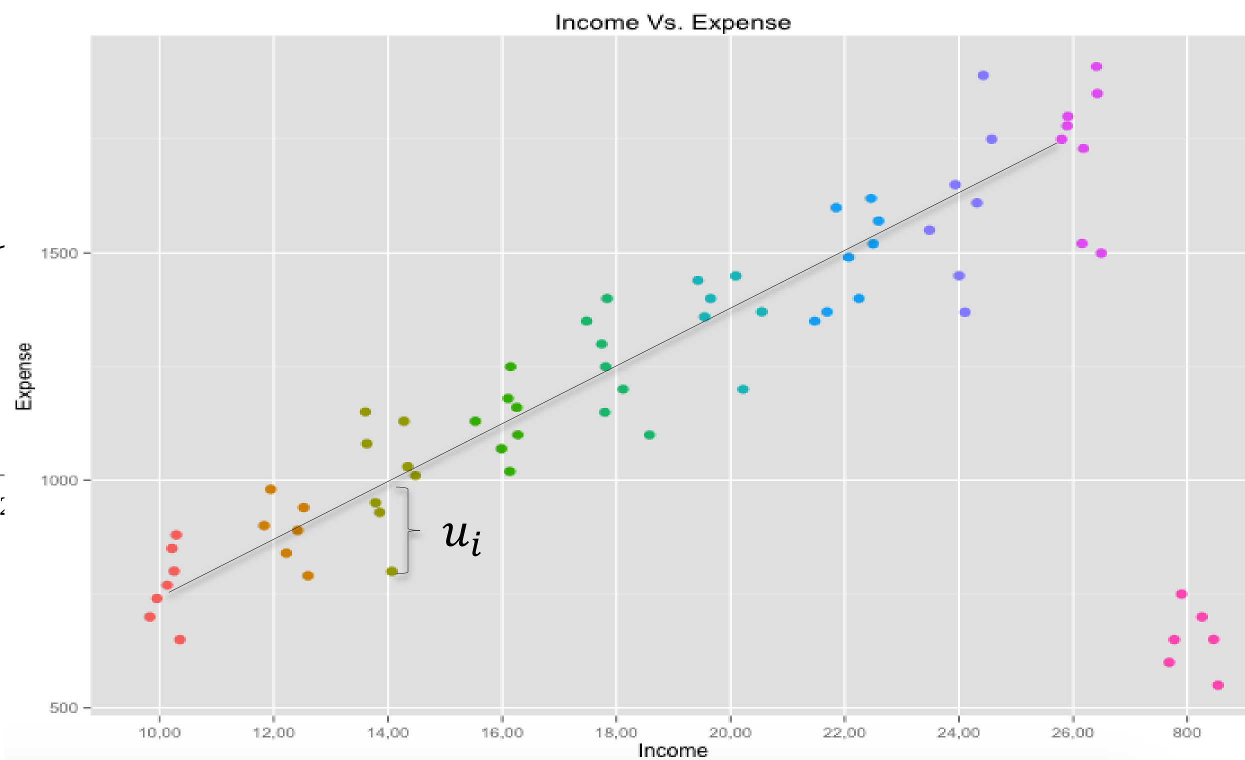
	Weekly family income X (Rs.)									
X	**800**	**1000**	**1200**	**1400**	**1600**	**1800**	**2000**	**2200**	**2400**	**2600**
Weekly expenditure (Rs.) Y	550	650	790	800	1020	1100	1200	1350	1370	1500
	600	700	840	930	1070	1150	1360	1370	1450	1520
	650	740	900	950	1100	1200	1400	1400	1550	1750
	700	800	940	1030	1160	1300	1450	1520	1650	1780
	750	850	980	1080	1180	1350	-	1570	1750	1800
	-	880	-	1130	1250	1400	-	1600	1890	1850
	-		-	1150	-	-	-	1620	-	1910
Total	3250	4620	4450	7070	6780	7500	6850	10430	9660	12110
E(Y|X)	650	770	890	1010	1130	1250	1370	1490	1610	1730

- The unconditional mean i.e. E(Y) = 72720/60 = 1212.
- The essence of regression analysis is to be use the knowledge of income level to better predict the weekly expenditure.

Regression–Population Regression Function

As family income increases, the average family consumption expenditure increases too. But does the individual family consumption expenditure increases too?





The deviation of individual family expenditure from the average family expenditure for a given X is denoted by

$$u_i = Y_i - E(Y|X_i)$$

E(Y|X) is called the population regression function and tells how the mean response of Y varies with X.

The first assumption of PRF is a linear function of X:

$$E(Y|X_i) = \beta_1 + \beta_2 * X_i$$

- β_1 is the estimated average value of Y when the value of X is zero. More often than not it does not have a physical interpretation
- β_2 is the estimated change in the average value of Y as a result of a one-unit change in X.

Linearity for regression assumes linearity in beta values and not in X variables.

Regression–Sample Regression Function

Generally the information available will be a randomly selected sample of Y values for fixed X values.

Y (Exp)	X (Inc)
700	800
650	1000
900	1200
950	1400
1100	1600
1150	1800
1200	2000
1400	2200
1550	2400
1500	2600

Y (Exp)	X (Inc)
550	800
880	1000
900	1200
800	1400
1180	1600
1200	1800
1450	2000
1350	2200
1450	2400
1750	2600

Sample regression function (SRF) takes the form:

$$\widehat{Y_i} = \widehat{\beta_1} + \widehat{\beta_2} * \widehat{X_i}$$

where

- $\widehat{Y_i}$ = estimator of E(Y|X_i)
- $\widehat{\beta_1}$ = estimator of β_1
- $\widehat{\beta_2}$ = estimator of β_2

Regression–Sample Regression Function

- Method of ordinary least squared (OLS) is used to choose SRF in such a way that

$$\sum \widehat{u_i}^{\,2} = \sum (Y_i - \widehat{Y_i})^2 \text{ is minimized.}$$

The equation obtained

$$\widehat{Y_i} = \widehat{\beta_1} + \widehat{\beta_2} * \widehat{X_i}$$

will have following properties:

- The sum of the squared residuals is a minimum.
- The sum of the residuals from the least squares regression line is 0.
- The simple regression line always passes through the sample mean of the Y and X variable.

Objective is to not only estimate $\widehat{\beta_1}$ and $\widehat{\beta_2}$ but also ensure it is close as possible to the true β_1 and β_2?

Regression Model–Validity and Usefulness

Will the model work on the population data? Is the model generalizable and useful?

Is the model valid?

- Use of co-efficient of determination to check the goodness of fit of regression.
- Precision of OLS estimates and t-tests to validate the beta coefficients are significant
- Analysis of Variance (ANOVA) and F test to check the overall fitness of the regression model.
- Residual analysis to check the model adequacies and Multicollinearity

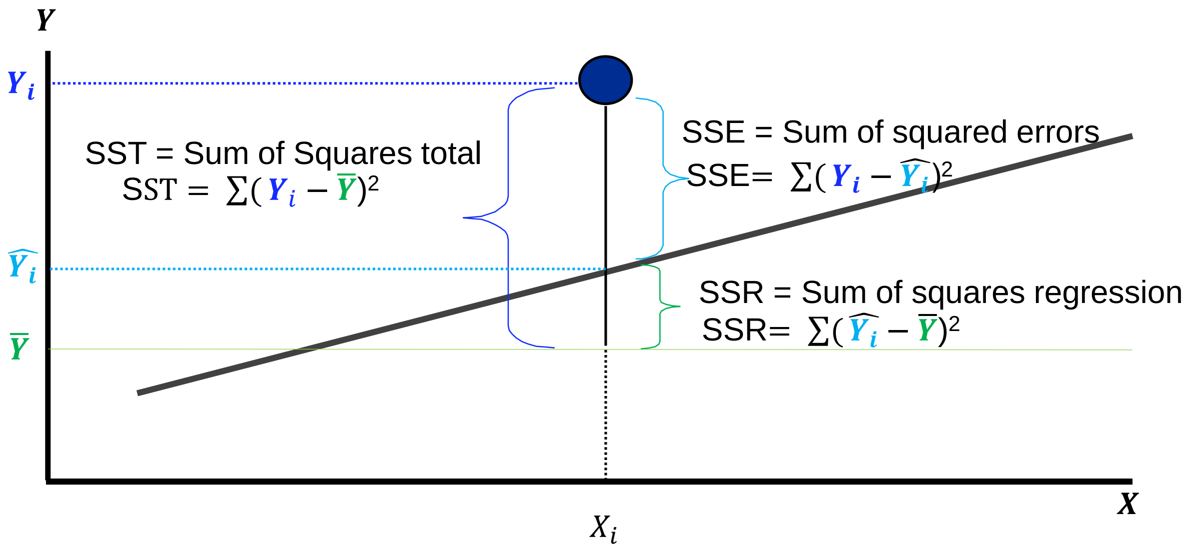
Is the model useful?

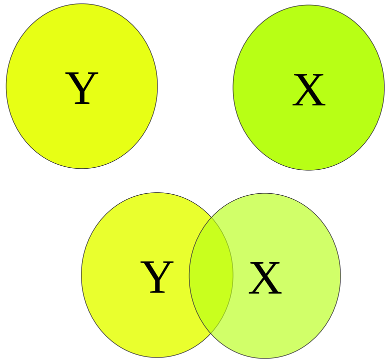
- Is the confidence interval estimating the average value of Y for a given value of X?
- Is the prediction interval estimating the individual Y for a given value of X?
- Is the prediction inline with the natural belief?

This is the gist of the assumptions of Classical Linear Regression Model (CLRM). More on the assumptions at: http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm

Is the model valid–Goodness of fit test

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE





Venn diagram representation

- Coefficient of determination is a measure of the extent to which the variation in Y is explained by X

- The r - squared (coefficient of determination) tells how well the sample regression line fits the data.

$$R^2 = \frac{SSR}{SST} \; where \; 0 \leq R^2 \leq 1$$

Closer the value of R^2 towards 1 more is the variation in Y explained by X.

Is the model valid–Goodness of fit test

Regression model for the Expense (Y) and Income (X):

ε

Y (Exp)	X (Inc)
700	800
650	1000
900	1200
950	1400
1100	1600
1150	1800
1200	2000
1400	2200
1550	2400
1500	2600

Y (Weekly Expense) = 244.5 + 0.509* X(Weekly Income)

Regression Statistics	
Multiple R	0.980847369
R Square	0.96206156
Adjusted R Square	0.957319256
Standard Error	64.93003227
Observations	10

Is the Model Valid–Precision of OLS Estimates

OLS estimates $(\widehat{\beta_1}, \widehat{\beta_2})$ are a function of sample data. If sample changes estimates will change. How to get the reliability of the estimate then?

- Precision or reliability of an estimate i.e. $\widehat{\beta_1}$ *and* $\widehat{\beta_2}$, is measured by the standard error of the $\widehat{\beta_1}$ *and* $\widehat{\beta_2}$

$$\mathrm{se}(\widehat{\beta_2}) = \widehat{\sigma} / \sqrt{\sum {x_i}^2}$$

where $\hat{\sigma} = \sqrt{\frac{\sum {u_i}^2}{n-2}}$; $\hat{\sigma}$ is the measure of standard deviation of y values about the estimated regression line. Also called standard error of the regression.

Standard error of beta coefficients goes on to decide the range in which the population beta coefficients may fall (in repeated sampling). Smaller the SE better is the range.

Is the Model Valid–Precision of OLS Estimates

How to know that the sample beta coefficient ($\widehat{\beta_2}$) is a true estimate of the population beta coefficient (β_2)?

- t-Test on the beta co-efficient with the following hypothesis:

$$\boldsymbol{H_0}\text{: } \beta_2 = 0; \boldsymbol{H_a}\text{: } \beta_2 \# 0$$

Decision rule: Reject H_0 if |t| >$t_{\alpha/2,df}$ where $|t| = (\widehat{\beta_2} - 0)/\text{s.e}(\widehat{\beta_2})$

P-value corresponding to the t-stats helps make decision. P < .05, reject the null hypothesis.

Is the Model Valid–Goodness of fit test

Regression model for the Expense (Y) and Income (X).

ε

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	244.5454545	64.13817299	3.812791091	0.005142172	96.64256241	392.4483467
Weekly_income (X)	0.509090909	0.035742806	14.24317115	5.75275E-07	0.42666785	0.591513968

Y (Exp)	X (Inc)
700	800
650	1000
900	1200
950	1400
1100	1600
1150	1800
1200	2000
1400	2200
1550	2400
1500	2600

- The confidence interval range (.4266, .5915) suggests that if we do repeated sampling, then in 95 out of 100 cases the above interval will contain the true beta.
- The larger the standard error, greater is the uncertainty of estimating true beta.

Calculating the confidence interval (in Excel):

Lower 95% = 0.50909 - T.INV.2T(0.05,8)*0.03574; upper 95% = 0.50909 - T.INV.2T(0.05,8)*0.03574

Is the Model Valid–ANNOVA and F Statistics

How to know that the sample coefficients which is an estimate of population coefficients are not simultaneously equal to zero?

Source of Variability	DoF	Sum of Squares	Mean Sum of Squares
Regression(Explained)	k	RSS	MSR=RSS/k
Error(Unexplained)	n-k-1	SSE	MSE=SSE/n-k-1
Total	n-1	SST=RSS+SSE	

- F-test is always a single tailed test while testing the hypothesis that the coefficients are simultaneously equal to zero. F statistics is given by:

$$F = \frac{MSR}{MSE} = \frac{RSS/k}{SSE/n-k-1}$$

Computed F value is compared with the critical F value from the F table. Or obtain the p-value. P < .05, reject the null hypothesis that all the beta values are simultaneously equal to zero. Valid for regression models with more than one X variables.

Is the Model Valid–ANNOVA and F Statistics

Regression model for the Expense (Y) and Income (X).

ε

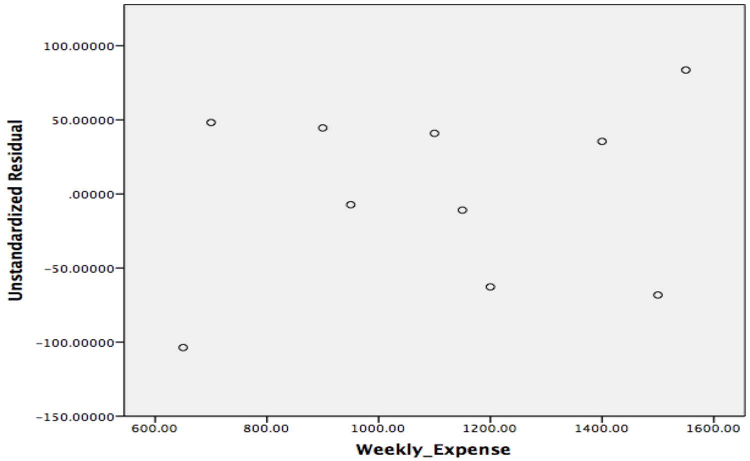
	df	SS	MS	F	Significance F
Regression	1	855272.7273	855272.7273	202.8679245	5.75275E-07
Residual	8	33727.27273	4215.909091		
Total	9	889000			

Y (Exp)	X (Inc)
700	800
650	1000
900	1200
950	1400
1100	1600
1150	1800
1200	2000
1400	2200
1550	2400
1500	2600

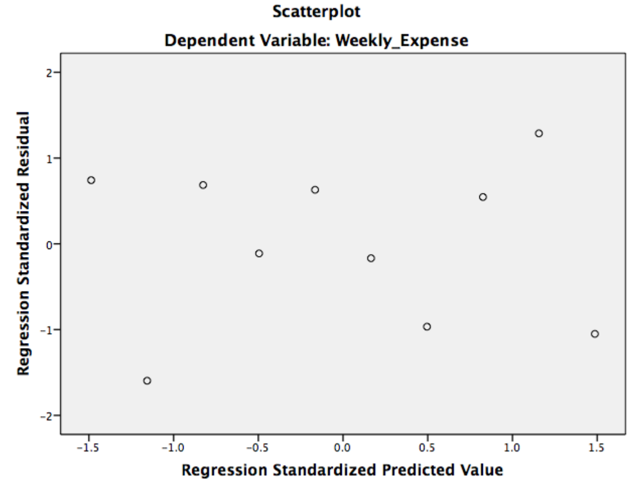
- Significance F (p-value) is less that 0.05 which signifies that the beta coefficients are not simultaneously equal to zero.

Is the Model Valid: Residual Analysis

No correlation between error term and X variable

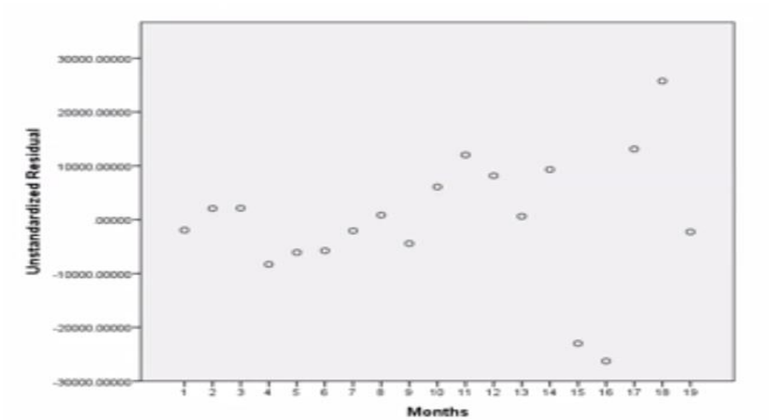


Heteroscedasticity problem (non constant variance for the error term)

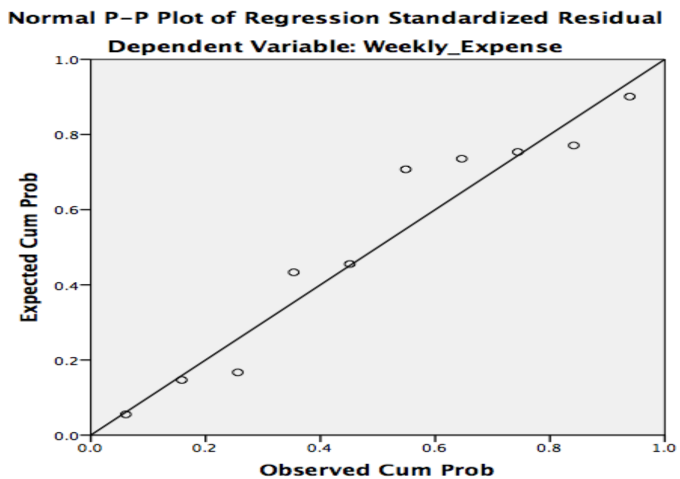


- White test
- Parks test

Autocorrelation in case of time series data (Plot of residual vs time)



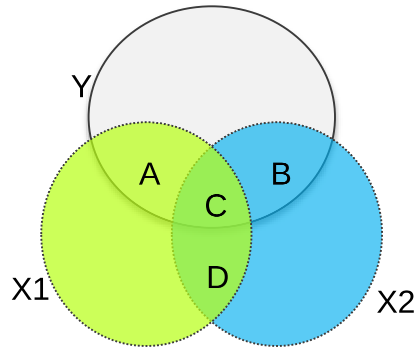
Normality of error term



- AD test

Is the Model Valid–Multicollinearity

The independent variables when correlated with each other leads to multicollinearity issue.



R^2			$1 - R^2$
Variance specific to X1 (part A)	Variance common to X1 & X2 (part C)	Variance specific to X2 (part B)	Left over

- Consequence of multicollinearity:
  - This correlation leads to larger variances and covariance's in the OLS estimators.
  - This can lead to wider CI of beta estimates and nullify the t-statistic for statistical significance.

Multicollinearity: If t-test concludes that the coefficients are not statistically different from zero but the F-test is significant and the coefficient of determination (R^2) is high. VIF (variance inflating factor) is a measure of MC. VIF>10 implies high degree of multicollinearity.

Is the model useful–Confidence and prediction interval

- Is the model depicting the general belief?
- Is the CI and PI encompassing the real world scenario?

Confidence Interval

- Confidence interval provides the interval estimate of the expected value of Y given X
- Used for interpolation of data within the range. $(1-\alpha) * 100\%$ CI for $E(Y|X)$ is:

$$\hat{Y}_i \pm t_{\frac{\alpha}{2},n-2} * S_e * \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SS_X}}$$

Where

$$S_e = standard\ error\ of\ regression$$

$$SS_X = \sum_{1}^{n} (X_i - \bar{X})^2$$

$$\hat{Y}_i = E(Y|X) = \widehat{\beta_1} + \widehat{\beta_2} * \hat{X}_i$$

Prediction Interval

- Prediction interval provides the interval estimate for Y given X
- Used for interpolation of data within the range. $(1-\alpha) * 100\%$ PI for Y is:

$$\hat{Y}_i \pm t_{\frac{\alpha}{2},n-2} * S_e * \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SS_X}}$$

Where

$$S_e = standard\ error\ of\ regression$$

$$SS_X = \sum_{1}^{n} (X_i - \bar{X})^2$$

$$\hat{Y}_i = E(Y|X) = \widehat{\beta_1} + \widehat{\beta_2} * \hat{X}_i$$

Transformations on Data

Transformations on the dataset may be needed to use linear regression techniques more effectively and achieve:

- Normality in error term.
- Homoscedasticity of variance.
- Normality of regression equation.
- Better strength of relationship between response and explanatory variables.

Relationship between σ^2 *and* $E(Y)$	**Transformation (Y')**
$\boldsymbol{\sigma}^{\mathbf{2}}$ is constant	$Y' = Y\ (no\ transformation)$
$\boldsymbol{\sigma}^{\mathbf{2}} \alpha\ \mathrm{E(Y)}$	$Y' = \sqrt{Y}$
$\boldsymbol{\sigma}^{\mathbf{2}} \alpha\ E(Y)^2$	$Y' = \ln(Y)$
$\boldsymbol{\sigma}^{\mathbf{2}} \alpha\ E(Y)^3$	$Y' = 1/\sqrt{Y}$
$\boldsymbol{\sigma}^{\mathbf{2}} \alpha\ E(Y)^4$	$Y' = 1/Y$

Representing Qualitative Factors

Representing Qualitative factors in a regression equation:

- By using 'dummy variables'. variables that take values of either 1 or 0, depending whether it is true or false.

Martial Status (MS)	MS_Married	MS_Single	MS_Divorced
Married	1	0	0
Single	0	1	0
Divorced	0	0	0

- If there are **n** factors, they can be represented by **n-1** dummy coded variables. This is derived from the concept of degrees of freedom.

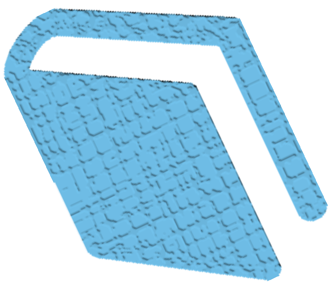
More on: http://www.ats.ucla.edu/stat/mult_pkg/faq/general/dummy.htm. Package 'dummy' can be used in R for Dummy Variable Coding.

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Regression in R Using an Example

Summary

Summary of the topics covered in this lesson:



- The intent of performing regression analysis is to predict the outcome of an event outside the sample dataset.
- Assumptions of linear regression needs to be satisfied to bring in better generalizability of the model.
- Usefulness of the model is understood by interpreting the signs of the coefficients and whether it is inline with the natural belief.

QUIZ TIME

Quiz Question 1

Quiz 1

Which library in R can be used for dummy variable coding?

Select all that apply.

a. dummy

b. *dummies*

c. *dummys*

d. *dumb*

Quiz Question 1

Quiz 1	Which library in R can be used for dummy variable coding? *Select all that apply.*

a. dummy

b. *dummies*

c. *dummys*

d. *dumb*

Correct answer is:

a & b

dummys and dumb are not defined packages in R.

End of Lesson08–Regression Concepts

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Data Science Using R

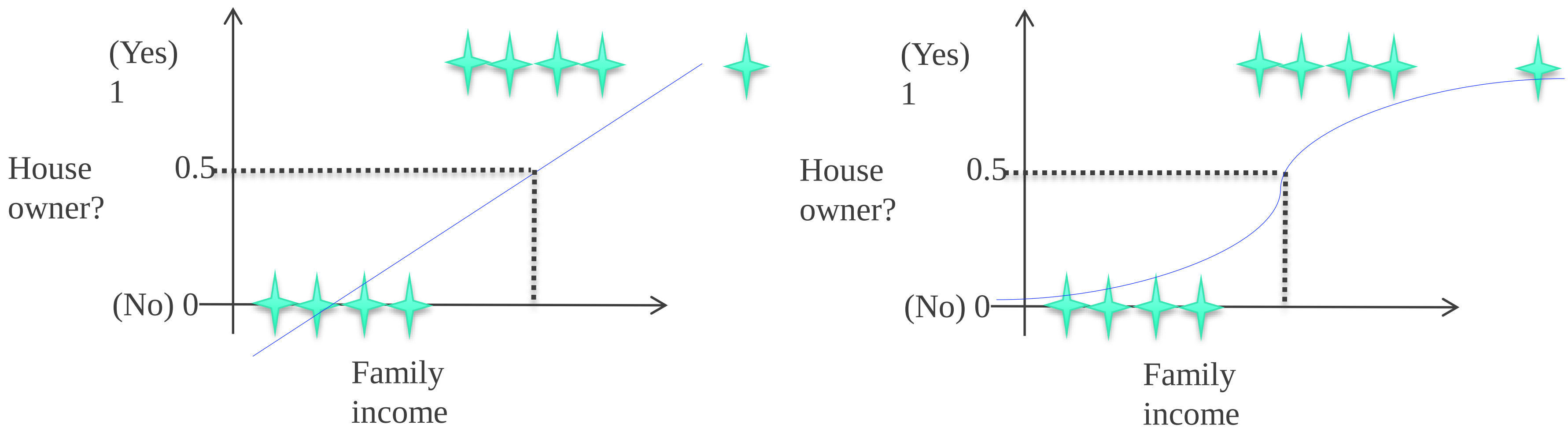
Lesson09–Logistic Regression Concepts

Objective

After completing this lesson you will be able to:

- Explain logistic regression analysis
- Describe the application areas of logistic regression
- Explain the various parameters derived to understand the validity of the model

Why Not Regression?



- Issues with regression to model qualitative response variable (owning a house at certain income level example):
  - Non-normality of error term
  - Heteroscedasticity in error term
  - Dependent variable beyond 0 and 1 values
  - Not logically attractive model.

Regression Vs. Logistic Regression

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Regression models	Logistic regression models
Objective is to estimate the expected or mean value given the independent variables.	Objective is to find the probability of an event given the independent variables.

The name, logistic regression, is derived from logistic function. Logistic regression or logit model is such that:

$$0 \leq f(x) \leq 1$$

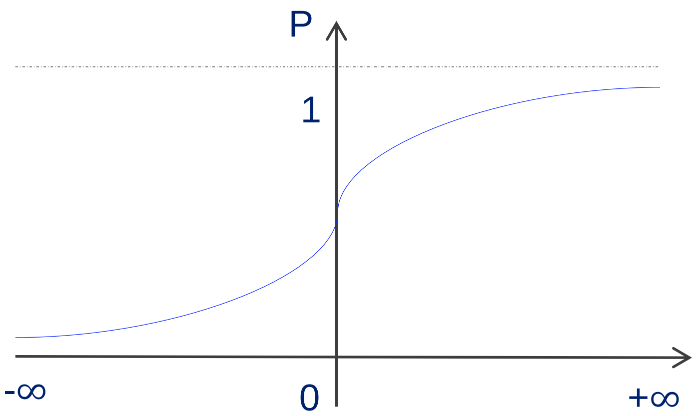
$$Y \in \{0,1\}$$

0: "Negative Class"
1: "Positive Class"

Applications of Logistic Regression

- Logistic regression is one of the most powerful technique to solve classification problem.
  - Email: Spam/Not Spam
  - Online Transaction: Fraudulent/Not Fraudulent (Yes/No)
  - HR Status: Joining/Not Joining
  - Credit Scoring: Defaulter/Non-defaulter

Deriving Logit Model



The logistic distribution function is given by:

$$P_i = P(Y=1|X_i) = \frac{1}{1+e^{-(\beta_1+\beta_2 * X_i)}}$$

Or

$$P_i = P(Y=1|X_i) = \frac{1}{1+e^{-Z_i}}$$

where $Z_i = \beta_1 + \beta_2 * X_i$

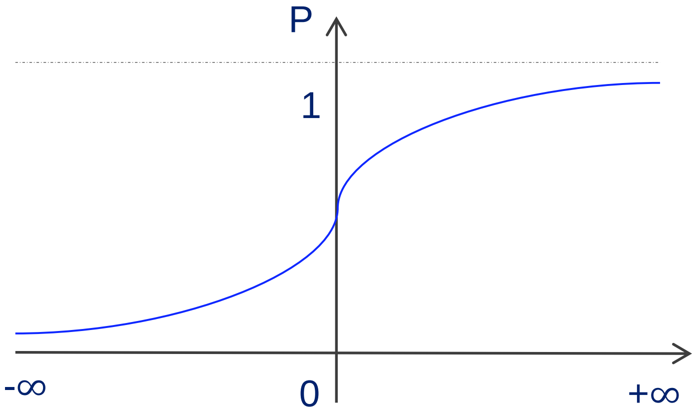
P_i is non-linear in βs. Linear transformation is required:

$$1 - P_i = P(Y=0|X_i) = \frac{1}{1+e^{Z_i}}$$

$$P_i/(1-P_i) = e^{Z_i}$$

$P_i/(1-P_i)$ is the odds ratio in favor of owning a house. The ratio of the probability that a family will own a house to the probability that it will not own a house.

Deriving Logit Model



The logistic distribution function is given by:

$$L_i = \ln(P_i/(1-P_i)) = Z_i$$

$$L_i = \ln(P_i/(1-P_i)) = \beta_1 + \beta_2 * X_i$$

L, the log of odds ratio is both linear in X and in parameters. This equation is called the logit model.

β_2, the slope, tells how the log odds in favor of say, owning a house change as income changes by a unit. If coefficient sign is positive, probability of owing a house increases. . If coefficient sign is negative, probability of owing a house decreases.

Odds Ratio Interpretation

- An example to illustrate the interpretation of coefficients.

ε

The logistic regression equation is

$$\ln(P_i/(1-P_i)) = -1.59474 + 0.07862 * X_i$$

$$P_i/(1-P_i) = e^{-1.59474} * e^{0.07862 * X_i}$$

where P is the probability of owning a house P(Y=1|X)

- $e^{0.07862} = 1.0817$ which means that every unit change in the income, the odds in favor of owning a house increases by 1.0817 or 8.17 %.

Estimation of the Logit Model

Family	Y	X ('000 in $)
1	0	8
2	1	16
3	1	18
4	0	11
5	0	12
6	1	19
7	1	20
8	0	13
9	0	9
10	0	10

Estimation will need value of X and Y.

$$\ln(P_i/(1-P_i)) = \beta_1 + \beta_2 * X_i + u_i$$

OLS will not work and maximum likelihood (ML) technique will be needed for estimating Logit. ML estimate holds good for large sample. Thumb rule >30 data points.

Logistic Regression–Model Validity

- Omnibus test of model coefficient:

Value less than 0.05 helps to reject the null hypothesis that that there is no difference between the model with only a constant and the model with independent variables

- Wald statistics: Equivalent of t – statistics in regression. Used to check the significance of individual explanatory variable.

If the P value corresponding to Wald statistics is < 0.05, the coefficient of the explanatory variable is not zero.

- Hosmer Lemeshow test: Test for overall fitness for binary logistic regression.

P value < 0.05 signifies bad fit for the model. P value > 0.05 the model is accepted

Logistic Regression–Model Validity

- Likelihood ratio: Equivalent of F statistics in regression. Used to test the null hypothesis that all the slope coefficients are simultaneously equal to zero.

  Deviance D = - 2 * (LL). LL implies log likelihood.

Measures the deviance from the perfect model. The larger the value of D, the worse the fit.
If the P value corresponding to D is < 0.05, the overall model is accepted.

- Conventional measure of R^2 is not meaningful. Different R^2 statistics prevalent:
  - McFadden R^2 value of .20 and above is considered good.
  - Cox and Snell R^2
  - Nagelkerke R^2: Modified Cox and Snell R^2 to maximum value of 1.
  - Count R^2 which is $\frac{no\ of\ correct\ predictions}{total\ number\ of\ observations}$ (If predicted probability is > 0.5 it is classified as 1 else as 0.)

Logistic Regression–Model Validity

$$Sensitivity = \left(\frac{TP}{TP + FN}\right) = \frac{4}{7} = 57.1\%$$

$$Specificity = \left(\frac{TN}{TN + FP}\right) = \frac{17}{17} = 100\%$$

Classification matrix

	Predicted	
Observed	**Class=1 (Positive)**	**Class=0 (Negative)**
Class =1 (Positive)	f_{11}= 4 [TP]	f_{10}= 3 [FN]
Class =0 (Negative)	f_{01}= 0 [FP]	f_{00}= 17 [TN]

$$Model\ accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) = \frac{21}{24} = 87.5\%$$

Sensitivity is the probability that predicted class is 1 when observed class is 1.
Specificity is the probability that the predicted class is 0 when the observed class is 0.

Logistic Regression–Influential Cases and Outliers

- An **outlier** is an observation whose dependent variable value is unusual given its values on the predictor variables.
  - Residual is the difference between the actual probability and predicted probability.

If a case has a standardized residual larger than 3.0 or smaller than -3.0, it is considered an outlier and a candidate for exclusion from the analysis

- An observation is said to be **influential** if removing the observation substantially changes the estimate of coefficients.
  - Cook's distance: is a measure of the influence which a case has on the solution

A case is identified as influential if its Cook's distance is greater than 1.0

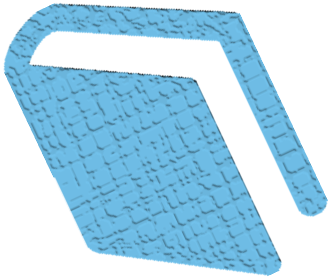
- An observation with an extreme value on a predictor variable is called a point with high **leverage**. Leverage is a measure of how far an observation deviates from the mean of that variable

If after removing the outliers and influential cases the model accuracy does not change by more than 2%, then retain the cases. More at: http://www.ats.ucla.edu/stat/r/dae/rreg.htm

Logistic Regression in R Using an Example

Summary

Summary of the topics covered in this lesson:



- The intent of performing logistic regression analysis is to predict the probability of an event outside the sample dataset.
- Validity of logistics regression is understood through various statistics.
- Validity is important to bring in better generalizability of the model.
- Usefulness of the model is understood by interpreting the signs of the coefficients and whether it is inline with the natural belief.

QUIZ TIME

Quiz Question 1

Quiz 1 What is the significance of Wald Statistics in logistic regression?

a. Used to check influential cases in the dataset.

b. Used to check the overall fit of the model.

c. Used to check the significance of individual explanatory variable.

d. None of the above.

Quiz Question 1

Quiz 1 What is the significance of Wald Statistics in logistic regression?

a. Used to check influential cases in the dataset.

b. Used to check the overall fit of the model.

c. Used to check the significance of individual explanatory variable.

d. None of the above.

Correct answer is: *c*

Equivalent of t – statistics in regression. Used to check the significance of individual explanatory variable.

End of Lesson09–Logistic Regression Concepts

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Data Science Using R

Lesson10–Decision Tree Concepts

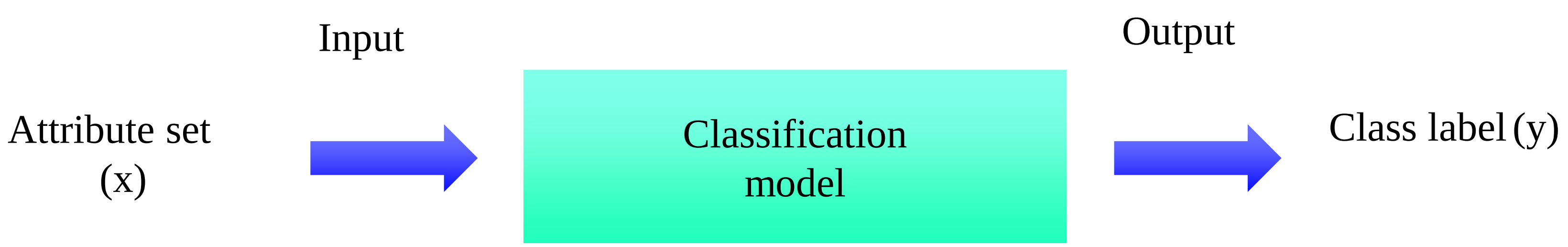
Objective

After completing this lesson you will be able to:

- Explain Decision Trees and its applications
- Explain the various parameters which are used to evaluate the outcome of the decision trees.

Decision Trees

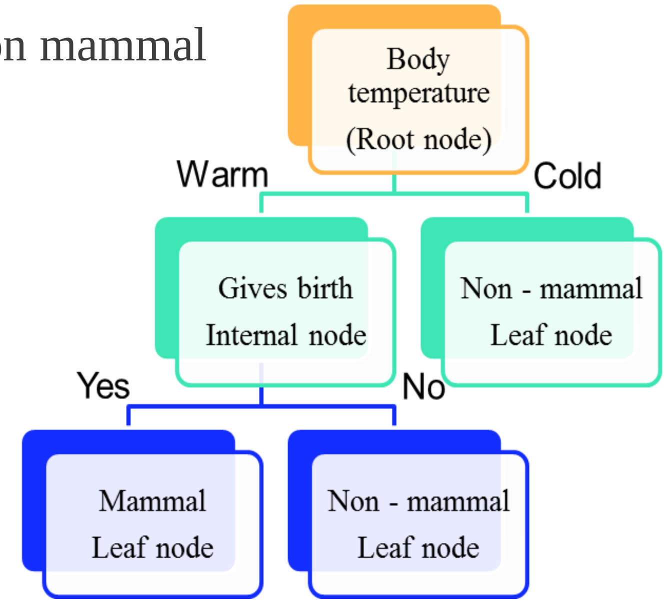
- Classification is a task of assigning objects to one of the several pre-defined categories.
  - Descriptive modelling: Can be used as an explanatory tool to distinguish between objects of different classes.
  - Predictive modelling: Can be used to predict the class label of unknown records.



- Objective is to build a learning algorithm with good generalization capability.

Decision Tree–Concept Development

Classifying species as mammal or non mammal



CART	C5.0	CHAID
Hunt’s algorithm	Hunt’s algorithm	CHAID algorithm
Split: Gini Index	Split: Entropy	Split: x^2 test

- Criteria for comparing different methods: Predictive accuracy, speed, robustness, scalability, Interpretability

Decision Tree–Classification Matrix

$$Sensitivity = \left(\frac{TP}{TP + FN}\right) = \frac{4}{7} = 57.1\%$$

$$Specificity = \left(\frac{TN}{TN + FP}\right) = \frac{17}{17} = 100\%$$

Classification matrix

	Predicted	
Observed	**Class=1 (Positive)**	**Class=0 (Negative)**
Class =1 (Positive)	f_{11}= 4 [TP]	f_{10}= 3 [FN]
Class =0 (Negative)	f_{01}= 0 [FP]	f_{00}= 17 [TN]

$$Model\ accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) = \frac{21}{24} = 87.5\%$$

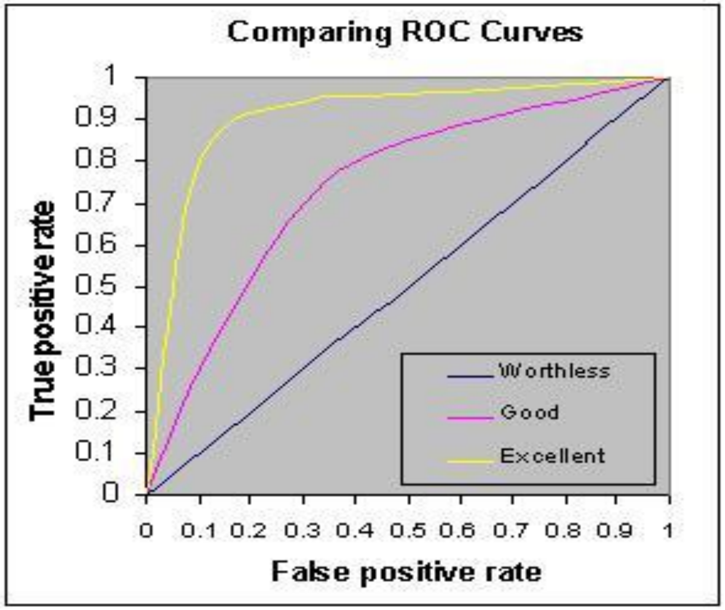
Sensitivity is the probability that predicted class is 1 when observed class is 1.
Specificity is the probability that the predicted class is 0 when the observed class is 0.

Decision Tree–ROC Curve

- Receiver operating characteristics (ROC) Curve is a useful way to cut-off point which maximizes sensitivity and specificity.
- Sensitivity and specificity measures are computed based on a sequence of cut-off points to be applied to the model for predicting observations into Positive or Negative.

An overall indication of the diagnostic accuracy of a ROC curve is the area under the curve (AUC). AUC values between:

- *0.9-1 indicate perfect sensitivity and specificity,*
- *0.8-0.9 indicate good sensitivity and specificity,*
- *0.7-0.8 indicate fair sensitivity and specificity,*
- *0.6-0.7 is poor*
- *0.6 and below indicate by chance outcome*



Decision Tree–Gain Chart and Lift Chart

- Lift and Gain chart measure how much better one can expect to do with the model comparing without a model.
- In contrast to the confusion/classification matrix that evaluates models on the whole population, gain or lift chart evaluates model performance in a portion of the population.

Steps to build Gain / Lift:

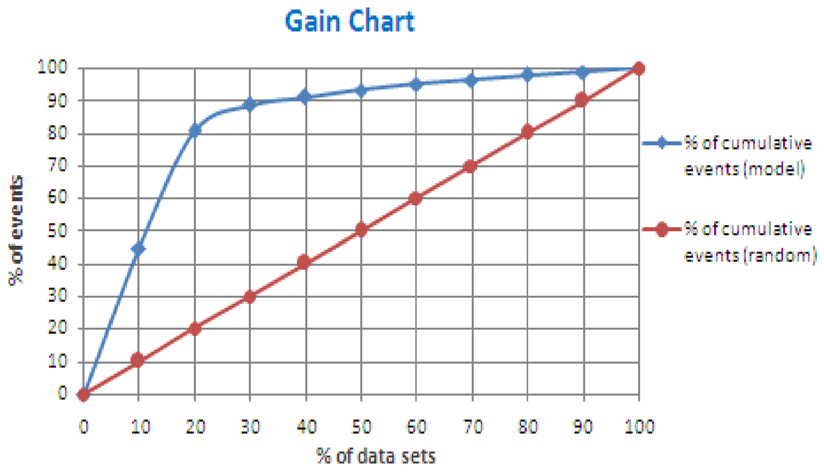
1. *Randomly split data into two samples (say): 80% = training sample, 20% = validation sample.*
2. *Score (predicted probability) the validation sample using the response model (training sample).*
3. *Rank the scored file, in descending order by probability.*
4. *Split the ranked file into 10 sections (deciles). Count the number of events in each section.*

Cumulative gains and lift charts are a graphical representation to depict the advantage of using a predictive model to choose which customers to contact.

Decision Tree–Gain Chart

	Input Values					
Decile	Number of Cases	Number of Responses	Cumulative Responses	% of events	Gain	Cumulative Lift
1	2500	2179	2179	44.71	44.71	4.47
2	2500	1753	3932	35.97	80.67	4.03
3	2500	396	4328	8.12	88.80	2.96
4	2500	111	4439	2.28	91.08	2.28
5	2500	110	4549	2.26	93.33	1.87
6	2500	85	4634	1.74	95.08	1.58
7	2500	67	4701	1.37	96.45	1.38
8	2500	69	4770	1.42	97.87	1.22
9	2500	49	4819	1.01	98.87	1.10
10	2500	55	4874	1.13	100.00	1.00
	25000	**4874**				

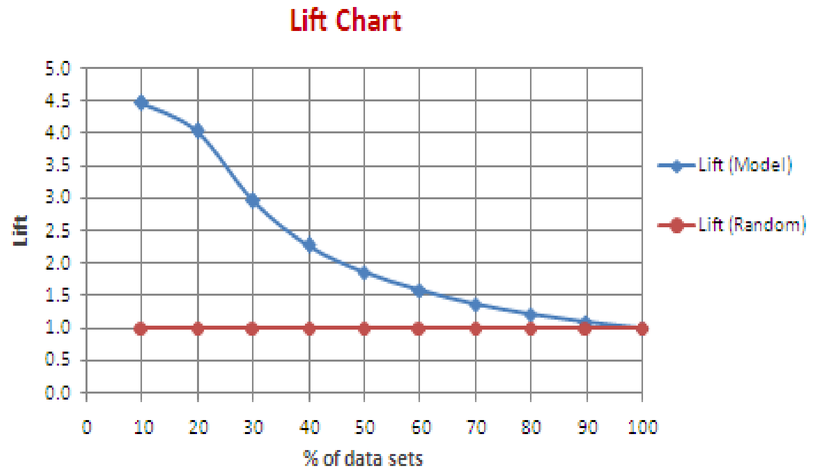
- Gain at a given decile level is the ratio of cumulative number of targets (events) up to that decile to the total number of targets (events) in the entire data set.



Source: http://www.listendata.com/2014/08/excel-template-gain-and-lift-charts.html

Decision Tree–Lift Chart

- Lift measures how much better one can expect to do with the model comparing without a model.
- It is the ratio of gain % to the random expectation at a given decile level. The random expectation at the xth decile is x%.



Interpretation:

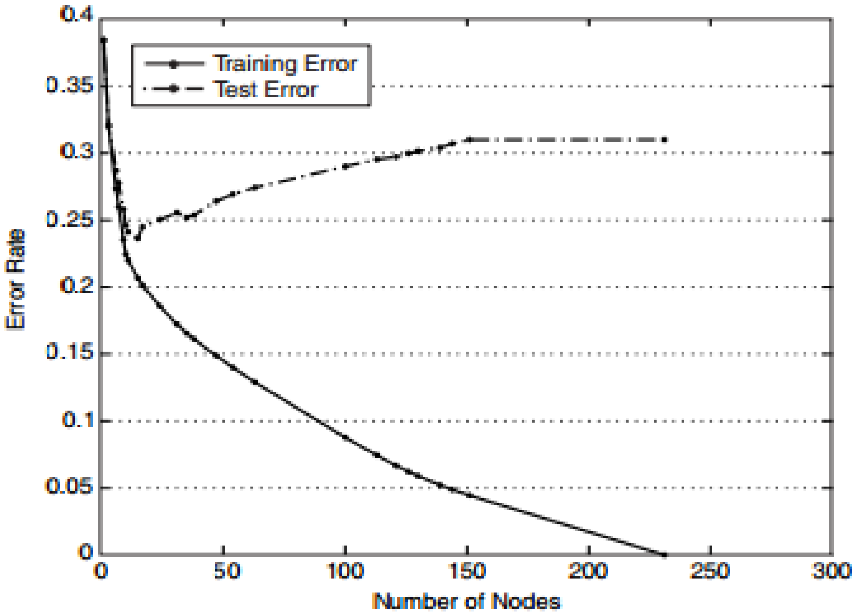
By contacting only 10% of customers, 4.5 times customers may respond.

To build Lift and Gain Chart in R. Refer to https://heuristically.wordpress.com/2009/12/18/plot-roc-curve-lift-chart-random-forest/

Decision Tree–Under Fitting and Over Fitting

- Model under fitting:
    - Model did not learn from the training set due to less data
    - Training and test error rate are large when the tree size is small

- Model overfitting:
    - Model has learned too much from the data and cannot be generalized.
    - As the number of nodes increases, the training error decreases but test error may increase
    - More complex trees than needed.



Model under fitting or over fitting leads to lack of generalizability and thus such decision tree models may not be useful in correct classification on unknown cases.

Decision Tree–Pruning

Pruning is applied to overcome the under fitting or over fitting issues in the decision tree model

Pre-pruning

Stop the algorithm before it becomes a fully grown tree:

- Stop if number of instances is less than some user specified threshold.
- Stop if expanding the current node does not improve impurity measures (e.g., Gini or information gain) by at least some threshold

This is more efficient but less accurate.

Post Pruning

Grow decision tree to its entirety. Trim the nodes of the decision tree in a bottom-up fashion

- If generalization error improves after trimming, replace sub-tree by a leaf node.
- Class label of leaf node is determined from majority class of instances in the sub-tree

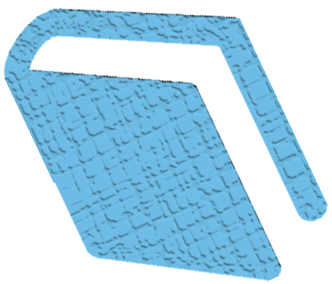
This is more accurate but less efficient.

Misclassification error pruning: Decision tree pruning stops when number of cases in a terminal node becomes less than a threshold

Decision Tree in R Using an Example

Summary

Summary of the topics covered in this lesson:



- Decision Tree is one of the most widely used data mining technique.
- The outcome of decision tree can be used for exploration of data as well as to build in predictive model.
- Unlike regression and logistic regression model, there are no statistical attributes which can suggest that the decision tree model is good and generalizable.

QUIZ TIME

Quiz Question 1

Quiz 1 Which of the below is a correct statement?

Select all that apply?

a. Sensitivity is the probability that predicted class is 1 when observed class is 1.

b. Specificity is the probability that the predicted class is 1 when the observed class is 0.

c. Specificity is the probability that the predicted class is 0 when the observed class is 0.

d. Sensitivity is the probability that predicted class is 0 when observed class is 1.

Quiz Question 1

Quiz 1

Which of the below is a correct statement?

Select all that apply?

a. Sensitivity is the probability that predicted class is 1 when observed class is 1.

b. Specificity is the probability that the predicted class is 1 when the observed class is 0.

c. Specificity is the probability that the predicted class is 0 when the observed class is 0.

d. Sensitivity is the probability that predicted class is 0 when observed class is 1.

Correct answer is:

a & c

b & d are incorrect statements.

End of Lesson10–Decision Tree Concepts

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Data Science Using R

Lesson11–Clustering and Segmentation

Objective

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

After completing this lesson you will be able to:

- Explain Clustering and its applications
- Describe hierarchical clustering and K means clustering.

Cluster Analysis

- Used in marketing for creating product segmentation and customer segmentation.
  - Is helpful to understand the product spread and understand which products are cannibalizing. Either internal or of the competitors.
  - Helps in creating customer profiles for targeted marketing.
  - The marketing expense can be optimized and utilized effectively.

- Clustering:
  - Putting similar things into one single group.
  - Clustering is performed by looking into different characteristics which may be helpful in bringing out a pattern.

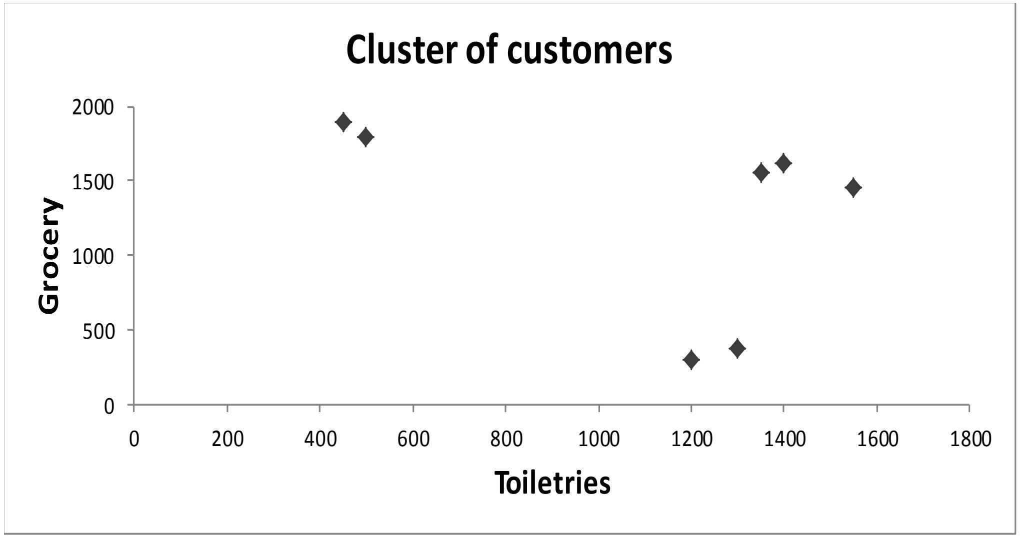
Types of clustering being discussed:
- *Hierarchical clustering*
- *K means clustering*

Hierarchical Clustering–Concept Development

Customer	Groceries	Toiletries
1	1200	300
2	1300	380
3	500	1800
4	450	1900
5	1350	1560
6	1400	1620
7	1550	1450

Three distinctively different categories:

- Low on toiletries and high on grocery
- High on toiletries and low on grocery
- High on toiletries and high on grocery



How do you do this for 10000 customers and 20 products?

Hierarchical Clustering–Concept Development

- Closer the points were, more similarity within the customers. Father the points, more dissimilarity within the customers.
  - Distance between the points is a measure of similarity or dissimilarity

- Calculate the linear distance between all the customers:
  - Distance between Customer 1 to Customer 2, 3, … , 7.
  - Distance between Customer 2 to Customer 1, 3, … , 7.
  - Distance between Customer 3 to Customer 1, 2, … , 7.
  - .
  - .
  - Distance between Customer 7 to Customer 1, 2, … , 6.

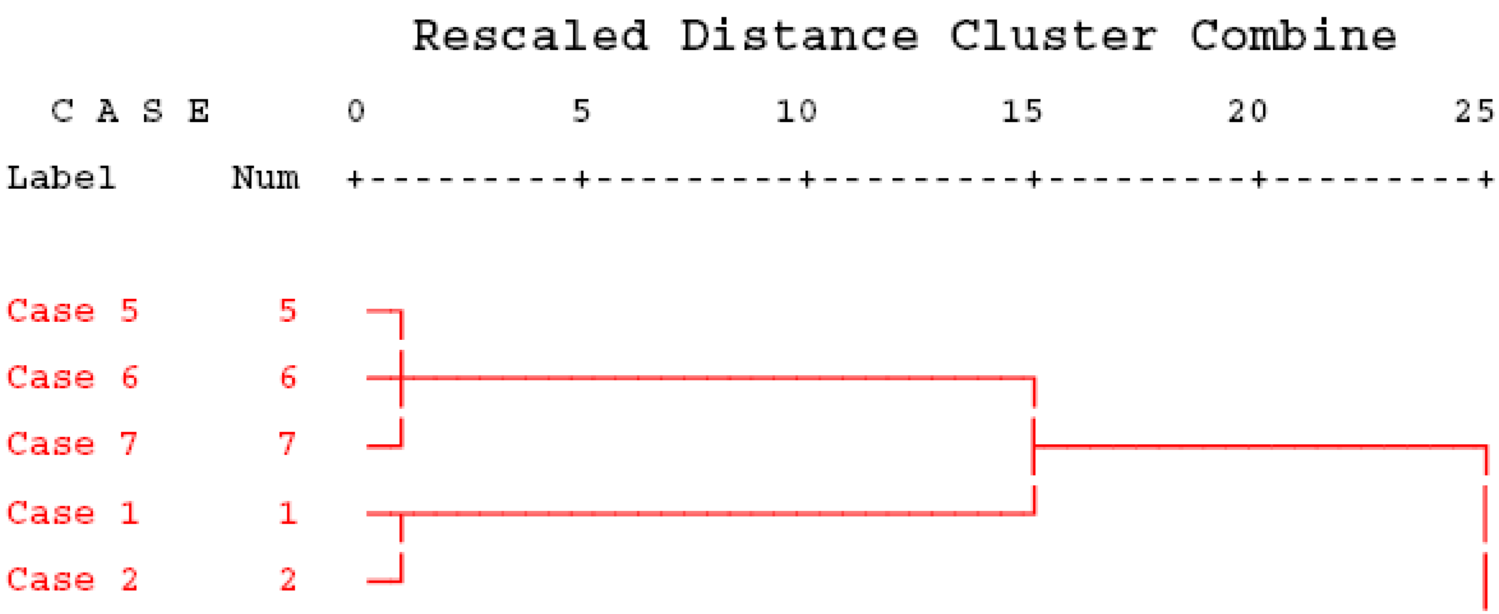
- 7*7 matrix is formed. Pick the smallest number. This forms the first cluster.

- Now one cluster and 5 customers. Total six entities for which above steps are repeated. Cluster formation happens in hierarchy and thus the name

Hierarchical Clustering–Concept Development

- When to stop the clustering:
  - The variation within the cluster is low and variation across cluster is very high.
  - Dendrogram gives this output in graphical form.
  - Farther distance travelled on dendrogram, more dissimilar entities are being clustered.

Dendrogram using Average Linkage (Between Groups)



Hierarchical Clustering–Beer Data Case

- 20 brands of beer with Calorie content, sodium content, Alcohol content and Cost.

Perform Hierarchical Clustering

ID	BEER	CAL	SOD	ALC	COST
1	Budweiser	144	15	4.7	0.43
2	Schlitz	151	19	4.9	0.43
3	Lowenbrau	157	15	4.9	0.48
4	Kronenbourg	170	7	5.2	0.73
5	Heineken	152	11	5	0.77
6	Old Mil	145	23	4.6	0.28
7	Augsburger	175	24	5.5	0.4
8	Strohs	149	27	4.7	0.42
9	Miller lite	99	10	4.3	0.43
10	Bud light	113	8	3.7	0.44
11	Coors	140	18	4.6	0.44
12	Coors lite	102	15	4.1	0.45
13	Michelob light	135	11	4.2	0.5
14	Becks	150	19	4.7	0.76
15	Kirin	149	6	5	0.79
16	Pabst	68	15	2.3	0.38
17	Hamms	136	19	4.4	0.43
18	Heilemans	144	24	4.9	0.43
19	Olympia	72	6	2.9	0.46
20	Schilitz lite	97	7	4.2	0.47

Hierarchical Clustering–Beer Data Case

- Linear distance from Budweiser
- These distances are to be calculated for each beer brand

ID	BEER	CAL	SOD	ALC	COST	Total
1	Budweiser	0	0	0	0	0
2	Schlitz	49	16	0.04	0	65.04
3	Lowenbrau	169	0	0.04	0.0025	169.0425
4	Kronenbourg	676	64	0.25	0.09	740.34
5	Heineken	64	16	0.09	0.1156	80.2056
6	Old Mil	1	64	0.01	0.0225	65.0325
7	Augsburger	961	81	0.64	0.0009	1042.641
8	Strohs	25	144	0	0.0001	169.0001
9	Miller lite	2025	25	0.16	0	2050.16
10	Bud light	961	49	1	0.0001	1011
11	Coors	16	9	0.01	0.0001	25.0101
12	Coors lite	1764	0	0.36	0.0004	1764.36
13	Michelob light	81	16	0.25	0.0049	97.2549
14	Becks	36	16	0	0.1089	52.1089
15	Kirin	25	81	0.09	0.1296	106.2196
16	Pabst	5776	0	5.76	0.0025	5781.763
17	Hamms	64	16	0.09	0	80.09
18	Heilemans	0	81	0.04	0	81.04
19	Olympia	5184	81	3.24	0.0009	5268.241
20	Schilitz lite	2209	64	0.25	0.0016	2273.252

Hierarchical Clustering–Beer Data Case

- Square of distance from bud wiser.
- The total column is called the Euclidian distance

ID	BEER	CAL	SOD	ALC	COST	Total
1	Budweiser	0	0	0	0	0
2	Schlitz	49	16	0.04	0	65.04
3	Lowenbrau	169	0	0.04	0.0025	169.0425
4	Kronenbourg	676	64	0.25	0.09	740.34
5	Heineken	64	16	0.09	0.1156	80.2056
6	Old Mil	1	64	0.01	0.0225	65.0325
7	Augsburger	961	81	0.64	0.0009	1042.641
8	Strohs	25	144	0	0.0001	169.0001
9	Miller lite	2025	25	0.16	0	2050.16
10	Bud light	961	49	1	0.0001	1011
11	Coors	16	9	0.01	0.0001	25.0101
12	Coors lite	1764	0	0.36	0.0004	1764.36
13	Michelob light	81	16	0.25	0.0049	97.2549
14	Becks	36	16	0	0.1089	52.1089
15	Kirin	25	81	0.09	0.1296	106.2196
16	Pabst	5776	0	5.76	0.0025	5781.763
17	Hamms	64	16	0.09	0	80.09
18	Heilemans	0	81	0.04	0	81.04
19	Olympia	5184	81	3.24	0.0009	5268.241
20	Schilitz lite	2209	64	0.25	0.0016	2273.252

Hierarchical Clustering–Beer Data Case

Euclidian distance matrix for all the brands

BEER	Budweise	Schlitz	Lowenbra	Kronenbo	Heineken	Old Mil	Augsburge	Strohs	Miller lite	Bud light	Coors
Budweise	0	65.04	169.04	740.34	80.21	65.03	1042.64	169	2050.16	1011	25.01
Schlitz	65.04	0	52	505.18	65.13	52.11	601.36	68.04	2785.36	1566.44	122.09
Lowenbra	169.04	52	0	233.15	41.09	208.13	405.37	208.04	3389.36	1986.44	298.09
Kronenbo	740.34	505.18	233.15	0	340.04	881.56	314.2	841.35	5050.9	3252.33	1021.44
Heineken	80.21	65.13	41.09	340.04	0	193.4	698.39	265.21	2810.61	1531.8	193.27
Old Mil	65.03	52.11	208.13	881.56	193.4	0	901.82	32.03	2285.11	1249.84	50.03
Augsburge	1042.64	601.36	405.37	314.2	698.39	901.82	0	685.64	5973.44	4103.24	1261.81
Strohs	169	68.04	208.04	841.35	265.21	32.03	685.64	0	2789.16	1658	162.01
Miller lite	2050.16	2785.36	3389.36	5050.9	2810.61	2285.11	5973.44	2789.16	0	200.36	1745.09
Bud light	1011	1566.44	1986.44	3252.33	1531.8	1249.84	4103.24	1658	200.36	0	829.81
Coors	25.01	122.09	298.09	1021.44	193.27	50.03	1261.81	162.01	1745.09	829.81	0
Coors lite	1764.36	2417.64	3025.64	4689.29	2516.91	1913.28	5411.96	2353.36	34.04	170.16	1453.25
Michelob	97.25	320.49	500.49	1242.05	289.71	244.21	1770.7	452.26	1297.01	493.25	74.16
Becks	52.11	1.15	65.12	544.25	68.09	41.24	650.77	65.12	2682.27	1491.1	101.11
Kirin	106.22	173.14	145.11	442.04	34	305.42	1000.4	441.23	2516.62	1301.81	225.28
Pabst	5781.76	6911.76	7927.77	10476.53	7079.44	5998.3	11540.24	6710.76	990	2075.96	5198.29
Hamms	80.09	225.25	457.25	1300.73	320.48	97.06	1547.21	233.09	1450.01	650.49	17.04
Heileman	81.04	74	250	965.18	233.13	2.11	961.36	34.04	2221.36	1218.44	52.09
Olympia	5268.24	6414	7310	9610.36	6429.51	5620.92	10939.76	6373.24	746.96	1685.64	4770.89
Schilitz lit	2273.25	64	0.25	0	2337.5	2560.2	6374.69	3104.25	13.01	257.25	1970.16

But there may be a problem if clustering is done without standardizing the data. Why?

Hierarchical Clustering–Beer Data Case

Amalgamation or Linkage Rules: Once several objects have been linked together, how do we determine the distances between those new clusters?

Single linkage (nearest neighbor):

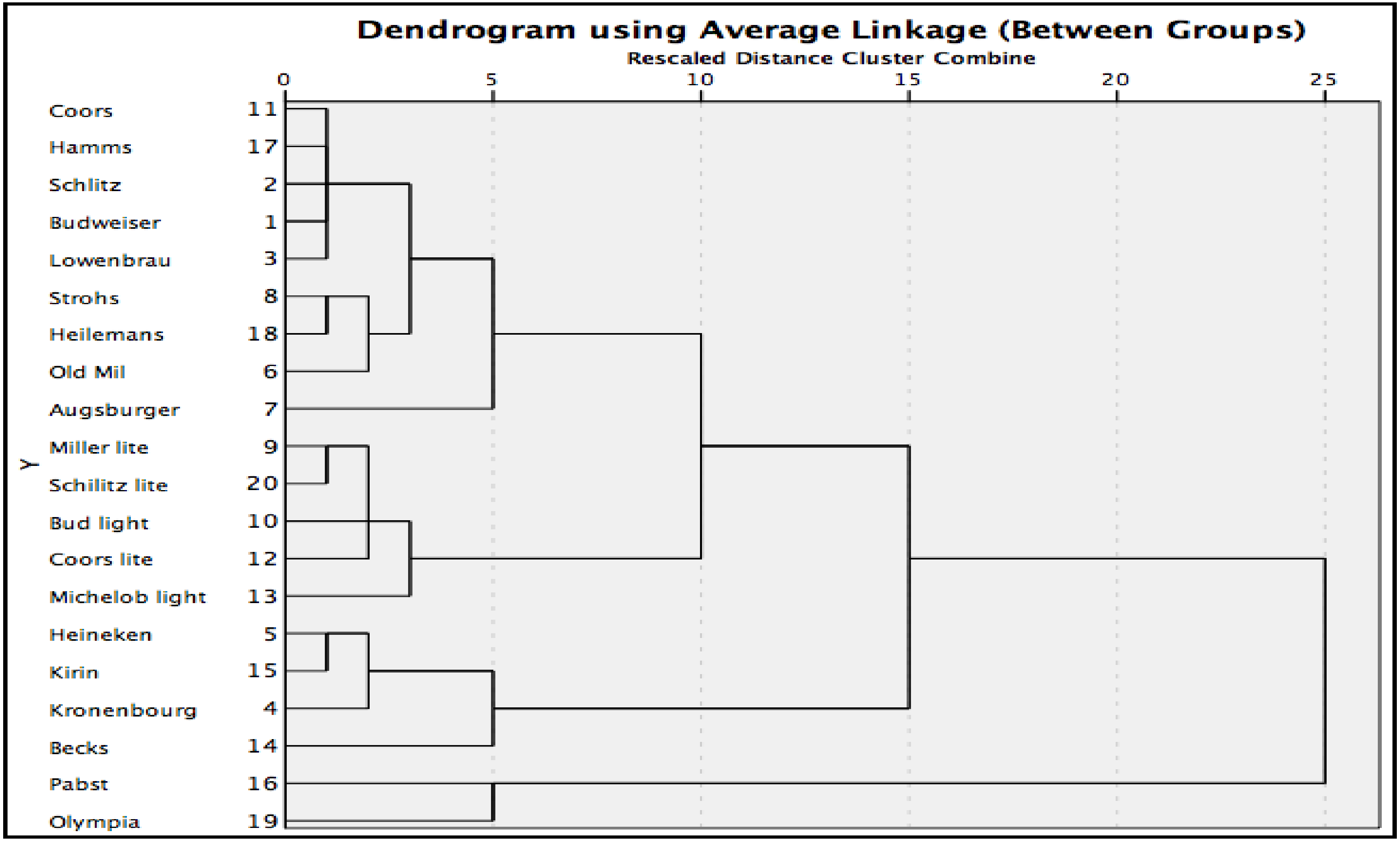
- The distance between two clusters is determined by the distance of the two closest objects (nearest neighbors) in the different clusters.
- This rule will, in a sense, string objects together to form clusters, and the resulting clusters tend to represent long "chains.“

Complete linkage (furthest neighbor):

- The distances between clusters are determined by the greatest distance between any two objects in the different clusters (i.e., by the "furthest neighbors").

Hierarchical Clustering Using Beer Data

Hierarchical Clustering–Beer Data Case

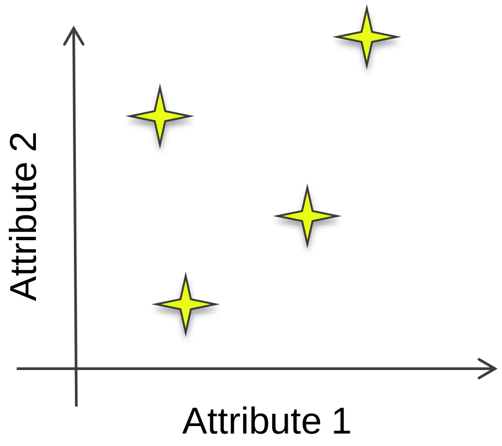


K Means Clustering–Concept Development

- Assume 70, 000 customer data having two attributes which needs to be segmented in 4 clusters.
- Here 4 depicts the K value for the cluster to be formed. Steps in K – means clustering
  1. Map 70K data points to 4 random numbers.
  2. Every mapping will keep shifting the centroids.
  3. Once 70K numbers are mapped, there will be four new centroids.
  4. Remove the data points and keep the new centroids.

- Repeat step 2 to 4 till the centroid movement stops.

Output is 4 clusters mapping 70, 000 customer data.



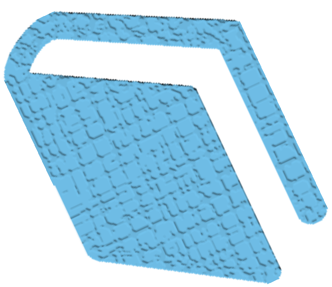
K Means Clustering Using Car Data

Clustering with Categorical Variables

- Hierarchical and K means clustering cannot handle categorical variables. Why?
  - Partitioning around Mediods (PAM) using 'gowers' as the distance measure rather than 'euclidian' as the distance measure.
  - Two step clustering technique (SPSS) can be applied to handle data with a mix of continuous and categorical variable.

Summary

Summary of the topics covered in this lesson:



- Clustering is one of the most used unsupervised learning algorithm.
- Hierarchical clustering is useful when comparing various brands, products on certain parameters.
- K means clustering is useful when the number of observations runs in thousands say customer footfall into supermarket, bank etc.
- Both Hierarchical and K means clustering cannot be used for grouping data with categorical variable.

QUIZ TIME

Quiz Question 1

Quiz 1 What is the distance measure for measuring dissimilarity between categorical variable?

a. Gower distance.

b. Euclidian distance.

c. Both Gower and Euclidian distance can be used.

d. Distance between categorical variable cannot be measured.

Quiz Question 1

Quiz 1	What is the distance measure for measuring dissimilarity between categorical variable?

a. Gower distance.

b. Euclidian distance.

c. Both Gower and Euclidian distance can be used.

d. Distance between categorical variable cannot be measured.

Correct answer is: *a*

Gower distance can be used as a distance measure in such cases.

End of Lesson11–Clustering and Segmentation

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Data Science Using R

Lesson12–Introduction to R Markdown and Rattle

Objective

After completing this lesson you will be able to:

- Describe R Markdown and Rattle
- Build a basic R Markdown document
- Explain the various features of Rattle
- Run a dataset in Rattle through a set of commonly used techniques of data analysis.

R Markdown–An Introduction

- R Markdown is an authoring format that enables easy creation of dynamic documents, presentations, and reports from R.
- R markdown can be used to create reports in the following format:

Report Format	Output Format
Document	HTML, PDF, WORD
Presentation	HTML(ioslides), HTML(Slidy), PDF(Beamer)
Interactive Shiny Report	Shiny Document, Shiny Presentation

- R Markdown documents can be automatically regenerated whenever underlying R code or data changes.

R Markdown–Install Package

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

- The first step to use R markdown is to install the package.

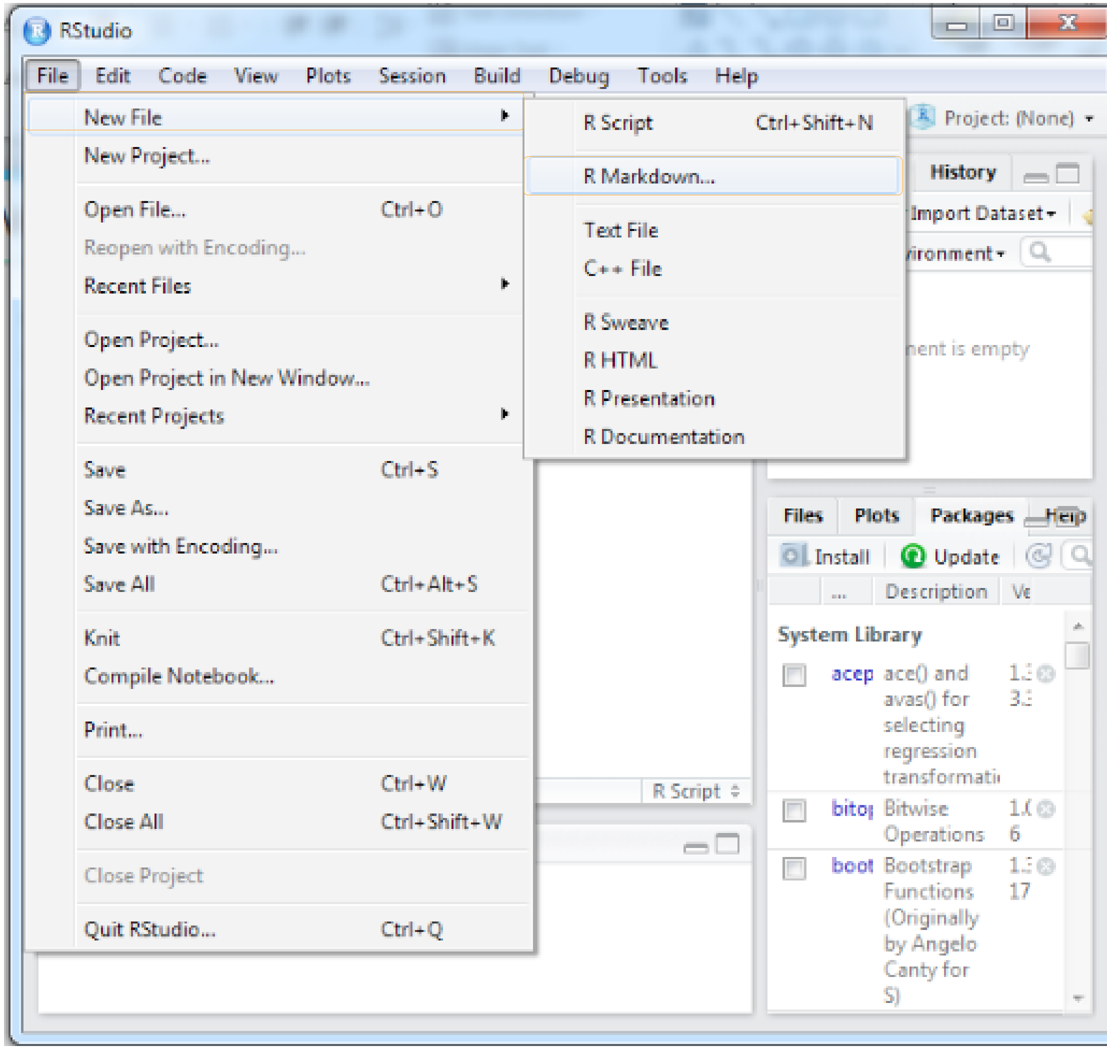
On R Studio Console:

```
>Install.packages(“rmarkdown”)
```

Or install using the Rstudio Install packages options

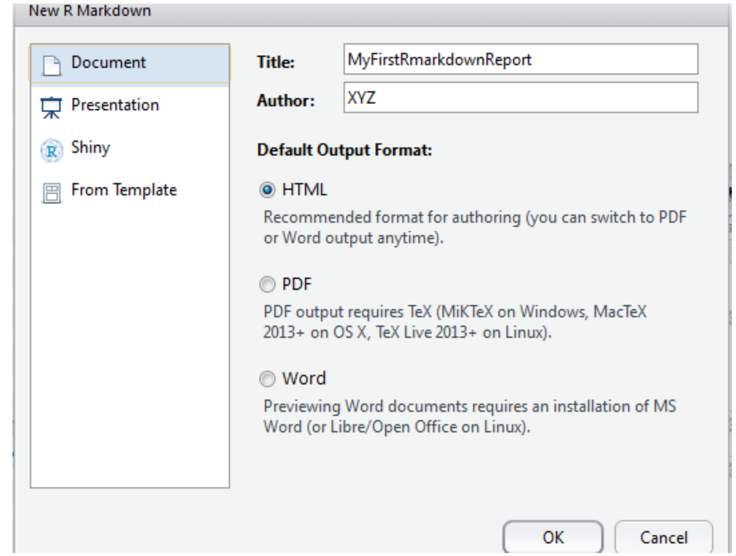
Working with R Markdown

- Open a new R markdown file from the R Studio file option.

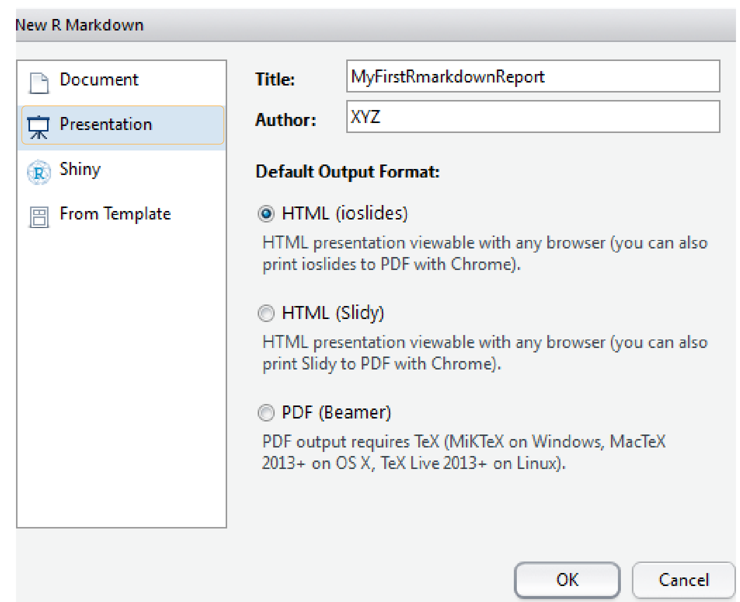


Working with R Markdown

- Select the type of report from the window that follows.



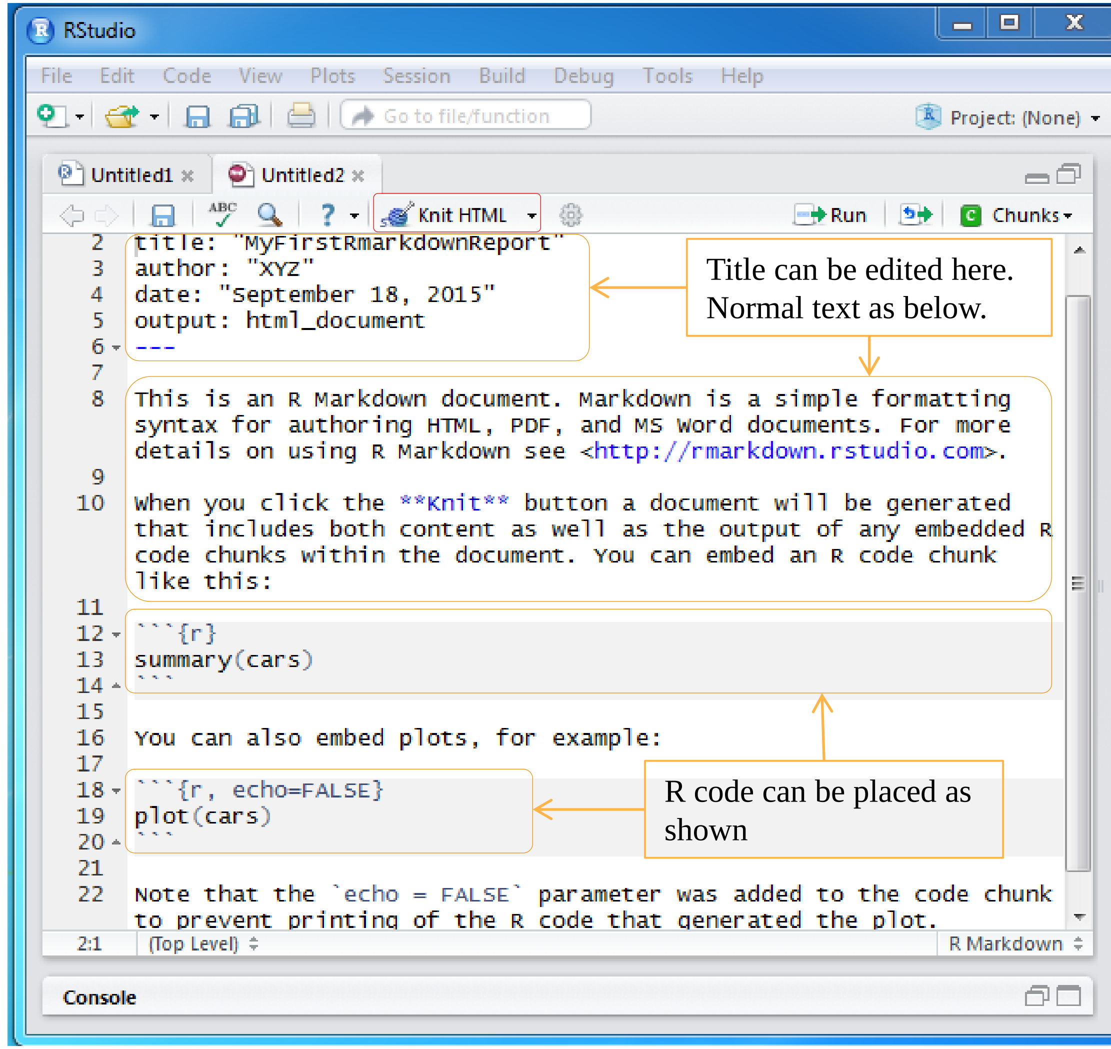
- *Select 'Document' as the report type if creating an HTML, PDF or Word document.*
- *Select 'Presentation' as the report type if creating HTML or PDF presentation.*
- *Select 'Shiny' as the report type if creating an interactive shiny report.*
- *There are specific templates which can be picked up to create report.*



My First R Markdown Code

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

- The R markdown code structure is simple to follow.
- Click on the Knit HTML icon to save the file.

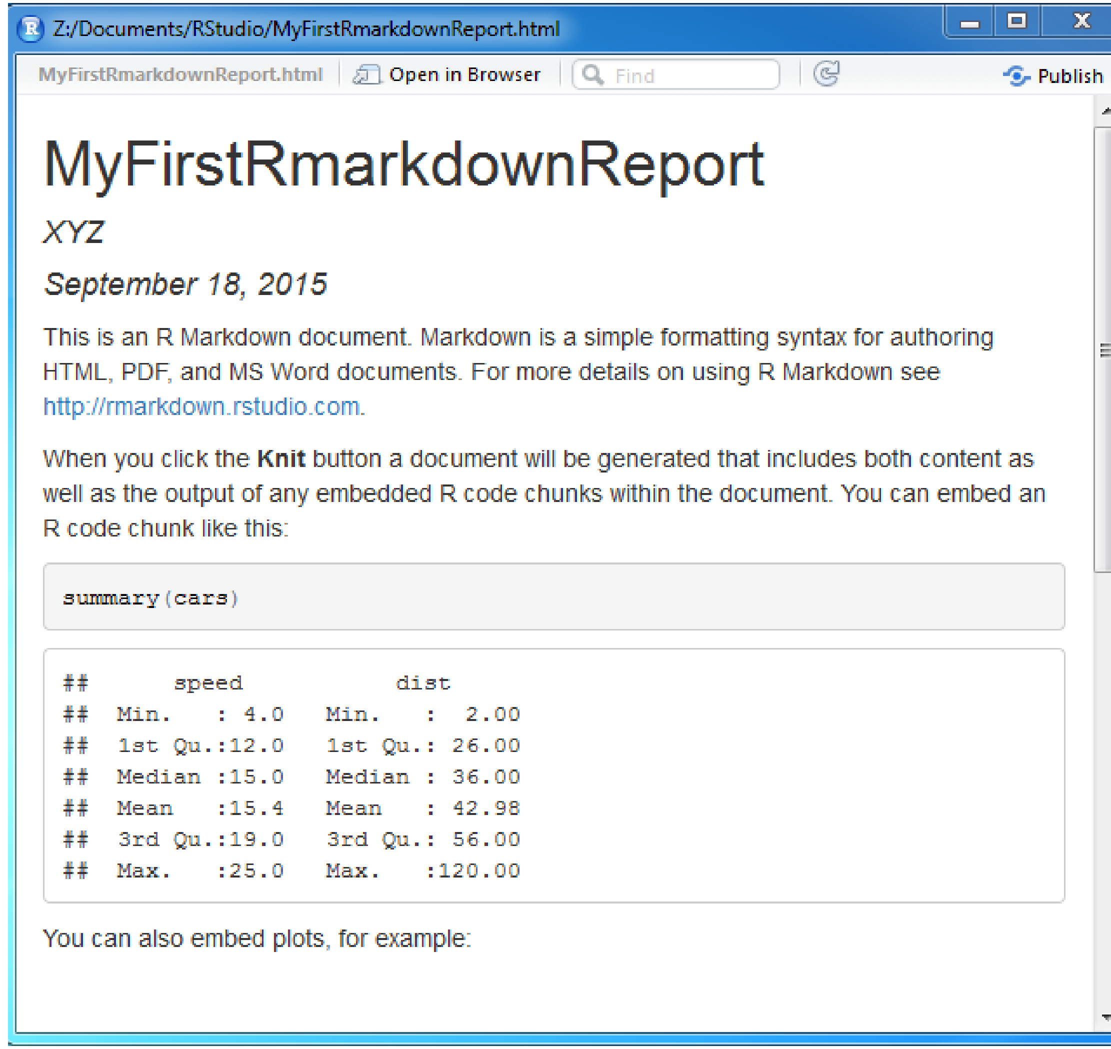


- File gets saved with '.Rmd' extension in the current working directory.
- Report can be opened up in a separate window or inside the R Studio viewer.

My First R Markdown Report

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT
BANGALORE

- The report will look like a formatted report.
- Very sophisticated formatting can be applied on the text including writing equations, hyperlinks, appending images etc.



Z:/Documents/RStudio/MyFirstRmarkdownReport.html

MyFirstRmarkdownReport.html | Open in Browser | Find | Publish

MyFirstRmarkdownReport

XYZ

September 18, 2015

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

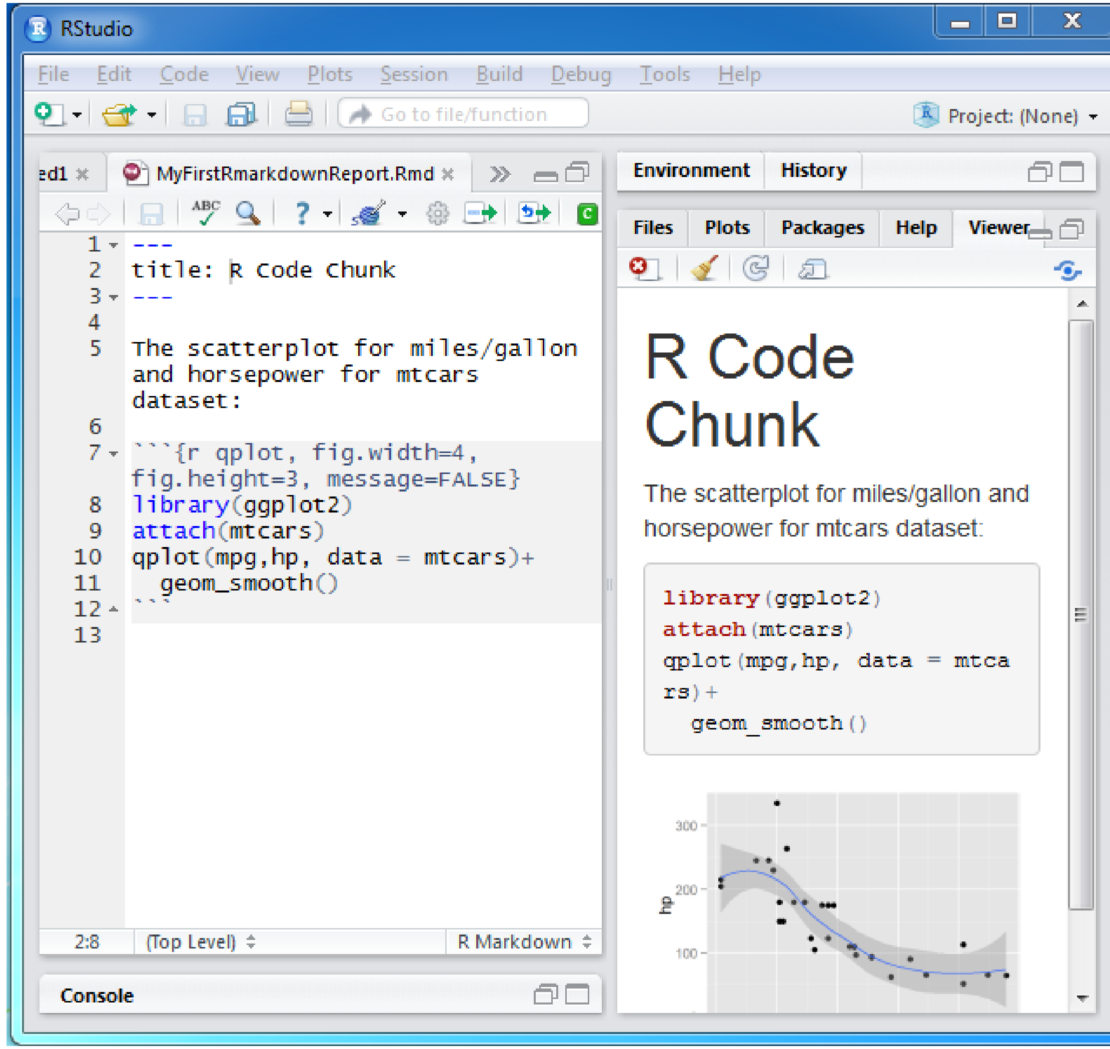
```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

You can also embed plots, for example:

R Markdown Code and Viewer

- The R code and viewer can be used side by side as a regular R scripting tool.
- The code for scatter plot and resulting output in the viewer is depicted here.



More on Rmarkdown at:
http://rmarkdown.rstudio.com/

Demo of the RMarkdown using an example dataset.

Rattle–An Introduction

- R Analytical Tool to Learn Easily (Rattle) is a user interface based data mining tool built on top of R.

***On R Studio Console*:**

```
>Install.packages(“rattle”)
```

To force the installation of all dependency:

```
>install.packages("rattle", dep=c("Suggests"))
```

Or install using the Rstudio Install packages options

- Rattle relies on extensive collection of R packages which powers the Rattle UI.

Dependent packages for Rattle are RGtk2, cairoDevice and XML. Troubleshooting at http://rattle.togaware.com/rattle-install-troubleshooting.html

Rattle User Interface

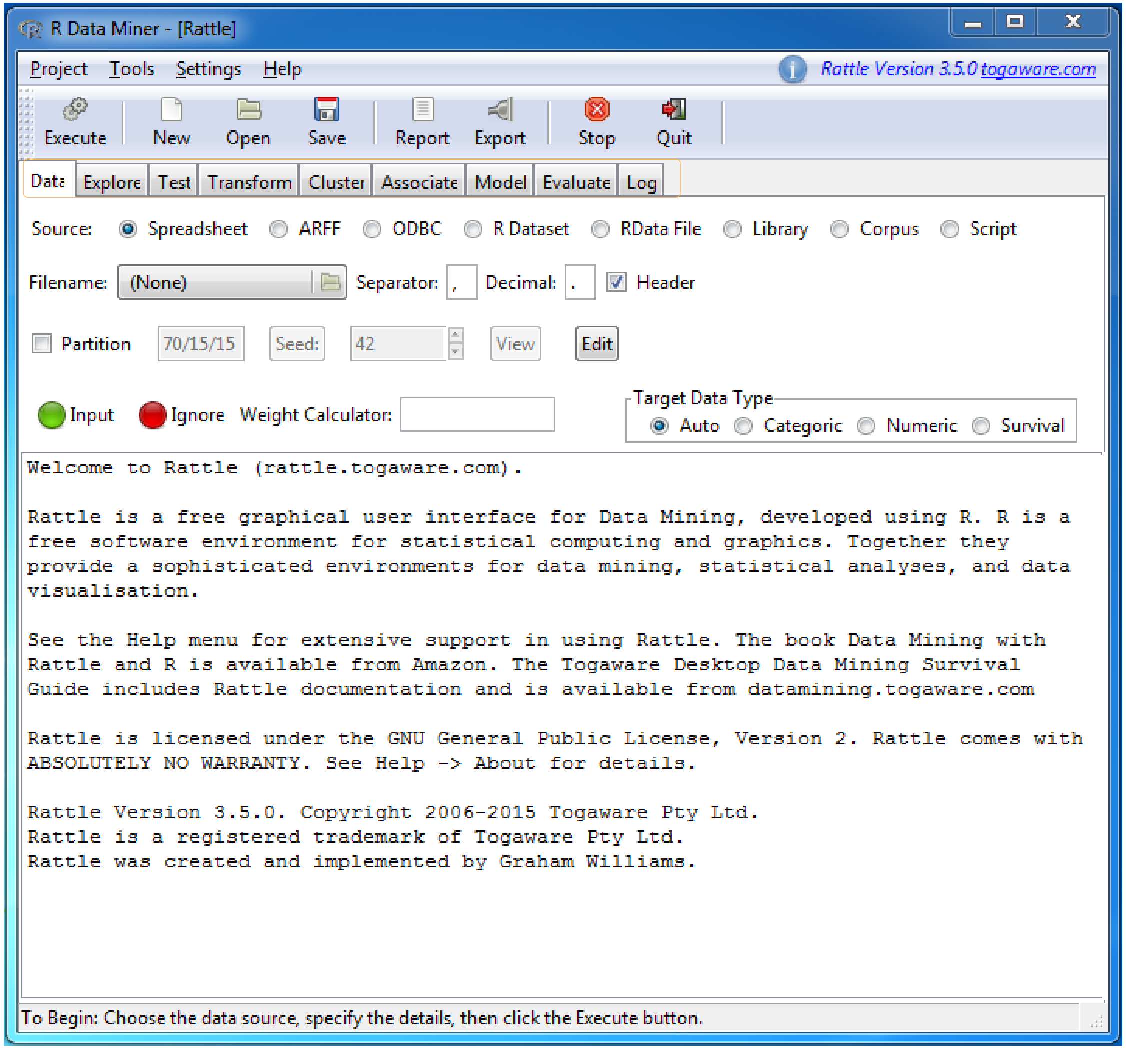
भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

- The user interface can be invoked as follows:

On R Studio Console:

```
>library(rattle)
>rattle()
```

- Tab based view with options to:
  - Load dataset
  - Explore dataset
  - Test distributions
  - Transform data
  - Clustering and association
  - Build models
  - Evaluate models
  - Code log

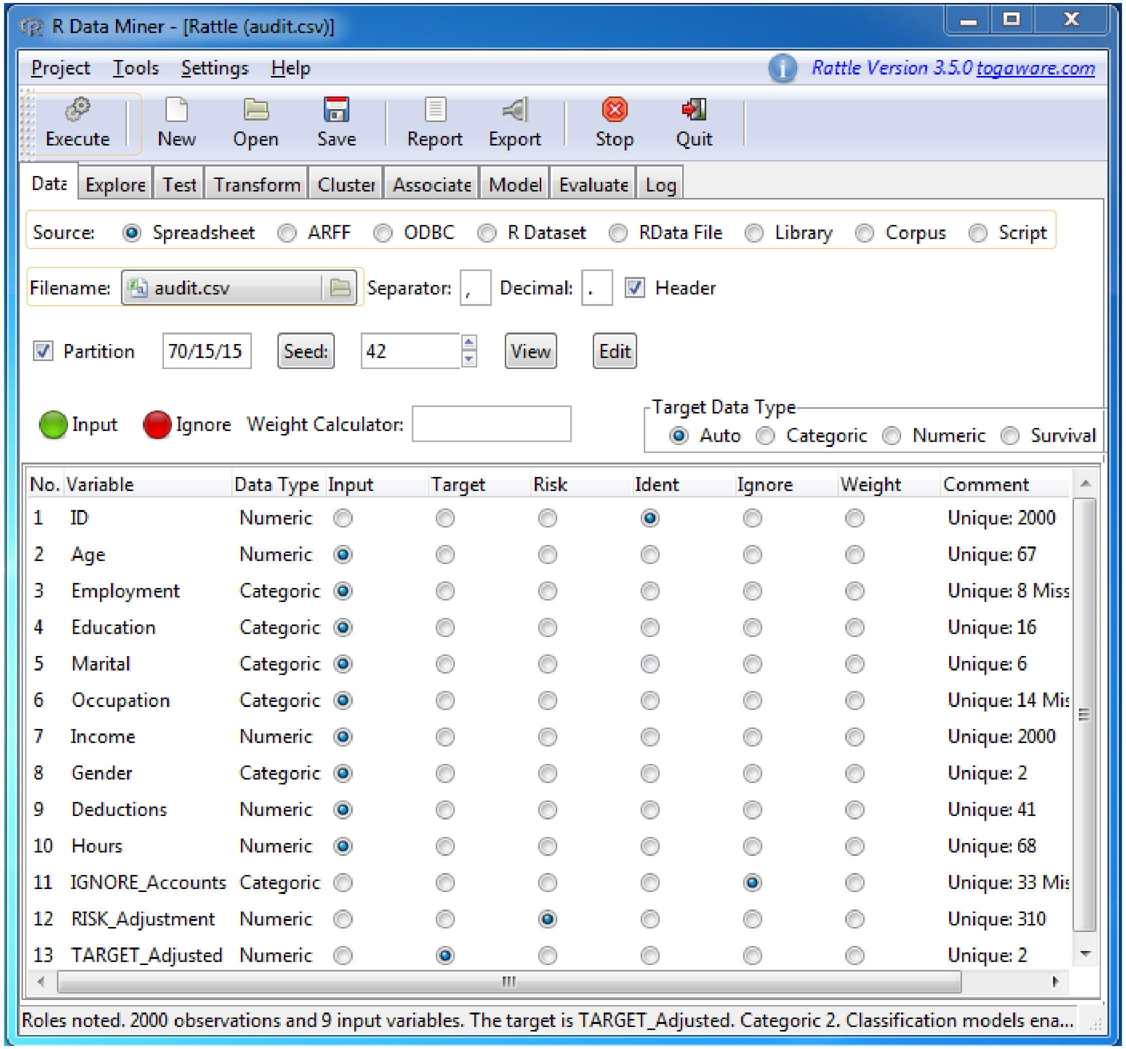


Rattle–Load Dataset

- A dataset is executed by the execute command.

- *If execute is clicked without any dataset, Rattle gives an option to load example dataset.*

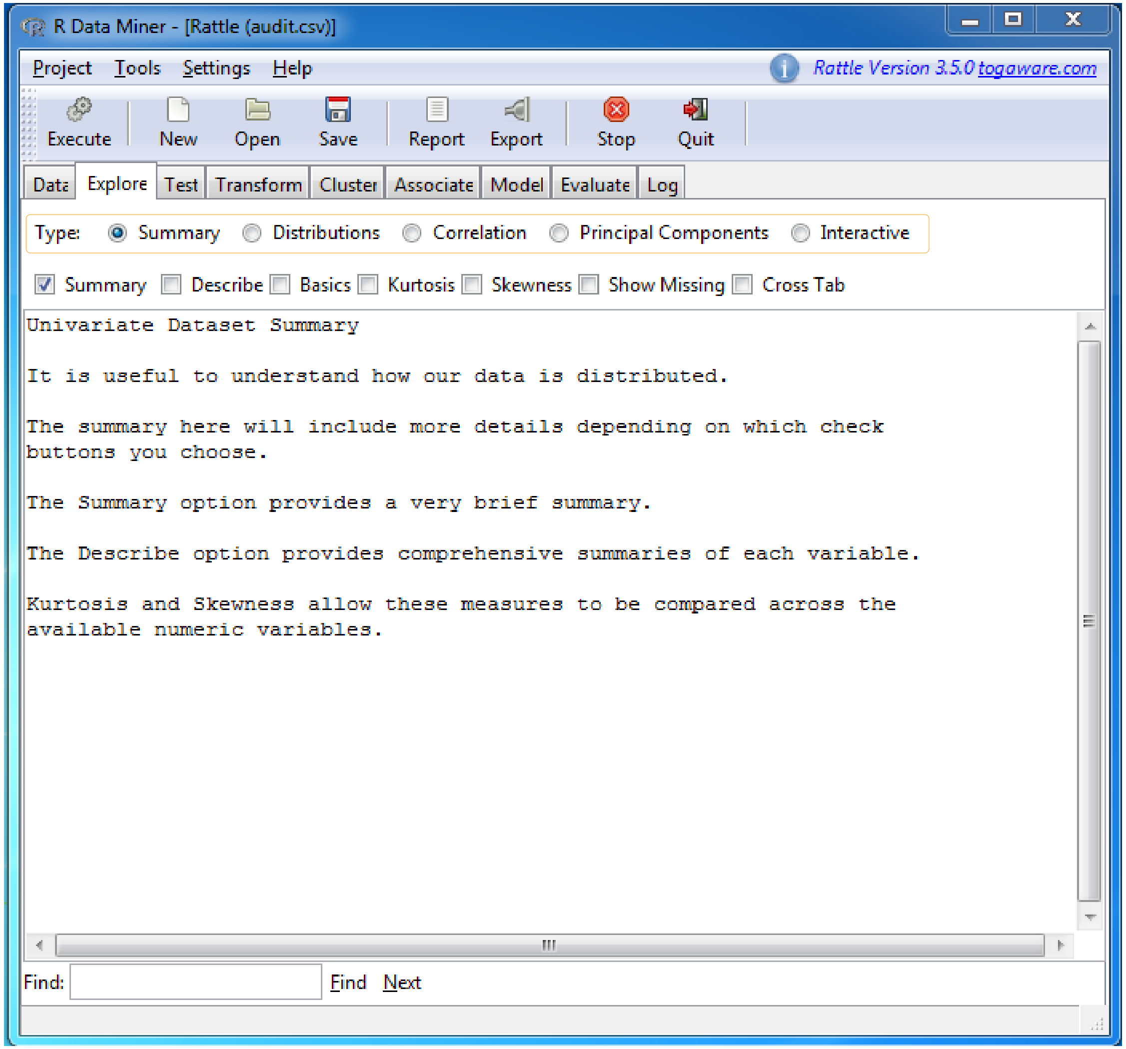
- *Rattle recognizes special pre-fixes for default variable role*
  - *'ID_'*
  - *'IGNORE_'*
  - *'RISK_' (measure of size of the target)*
  - *'IMP_'*
  - *'TARGET_'*



Rattle–Explore Dataset

- Explore tab provides various options for exploratory data analysis

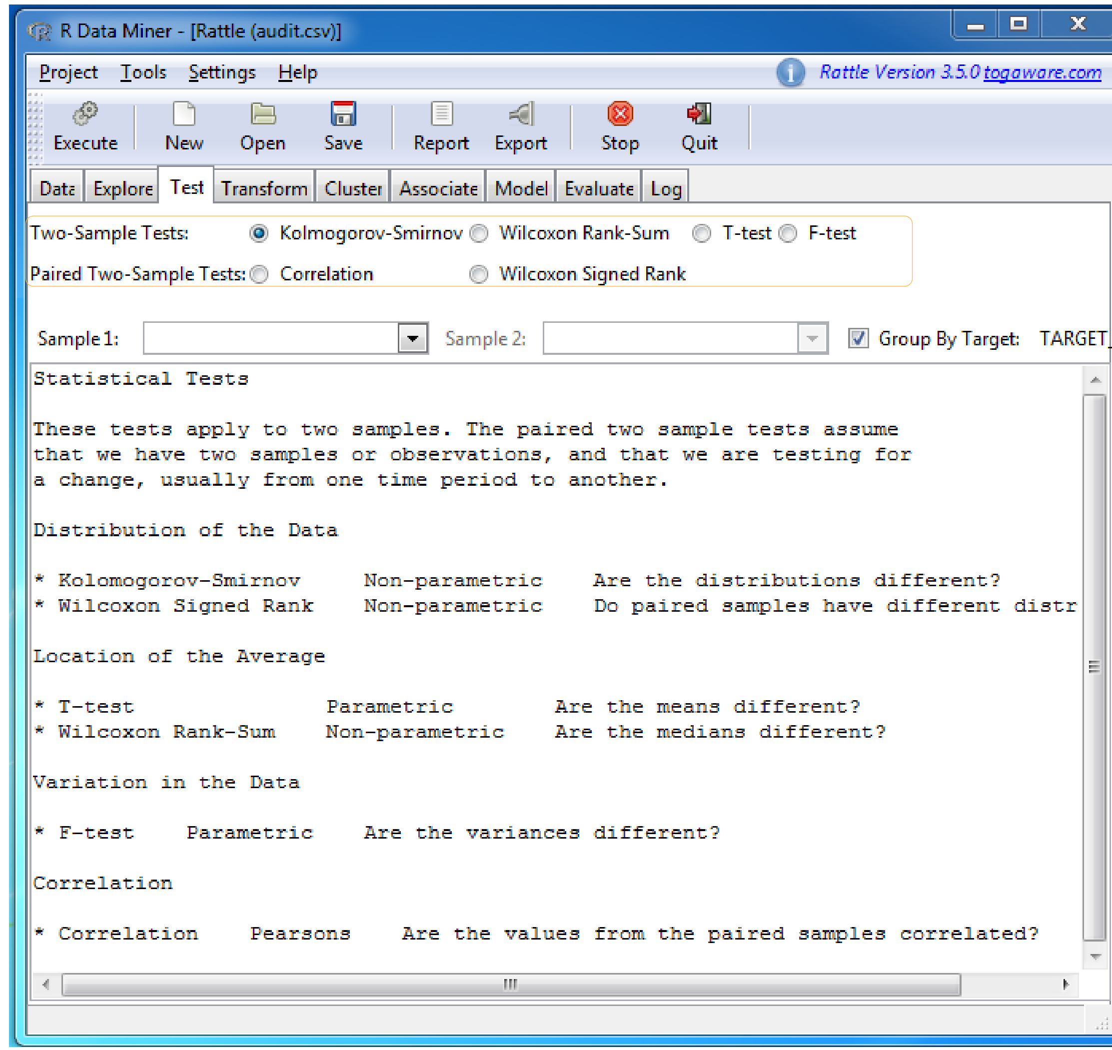
- ***Summary***: *Provides basic univariate summary and extended summary.*
- ***Distributions***: *Provide various plots for numeric as well as categorical data*
- ***Correlation***: *provides insights into the independence of the numeric input variables.*
- ***Principal component***: *Provides insight into the importance of variables in explaining the variation.*
- ***Interactive***: *Provides option for Interactive data exploration.*



Rattle–Test Dataset

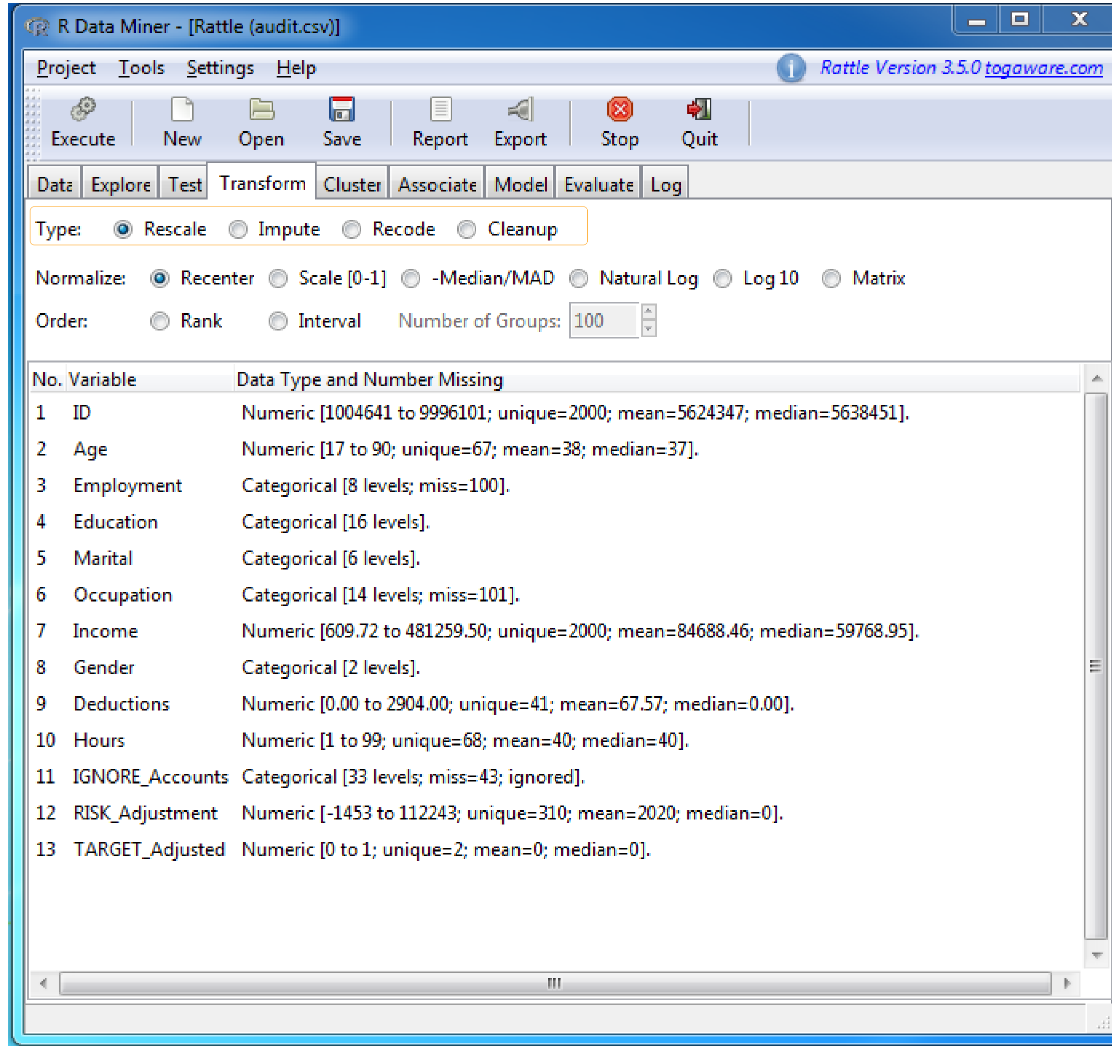
भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT
BANGALORE

- Provides access to number of statistical tests of distributions.



Rattle–Transform Dataset

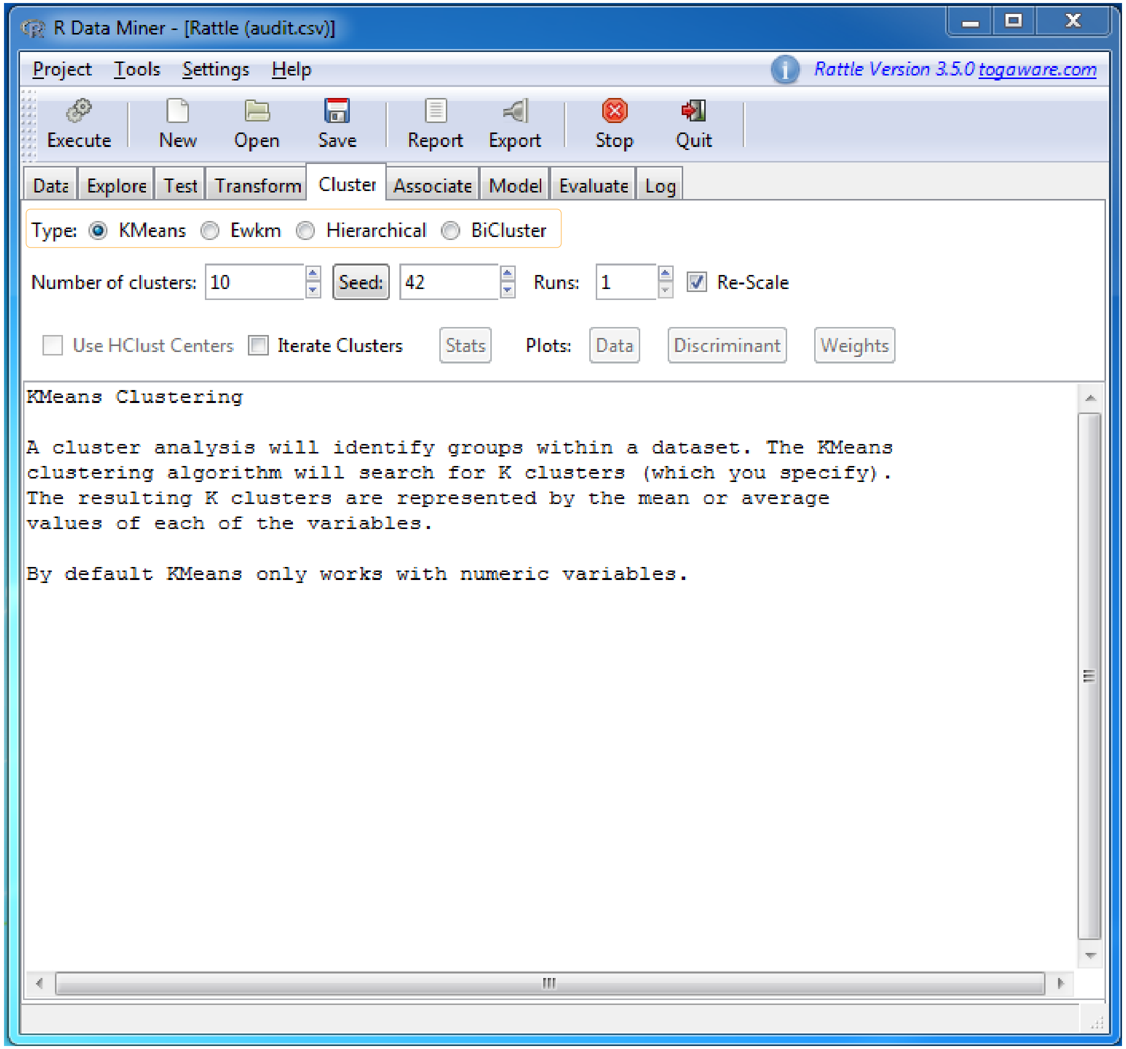
- Cleaning data and creating new features (derived variables) takes significant time in data analysis.



- ***Rescale***: *Provides options for re-centering and scaling around zero.*
- ***Impute***: *Provides basic imputation of missing values using mean, median and mode.*
- ***Recode***: *Provides options for recoding/binning the variables with a default of 4 bins.*
- ***Cleanup***: *Provides option to treat the missing values after having tried imputation etc.*

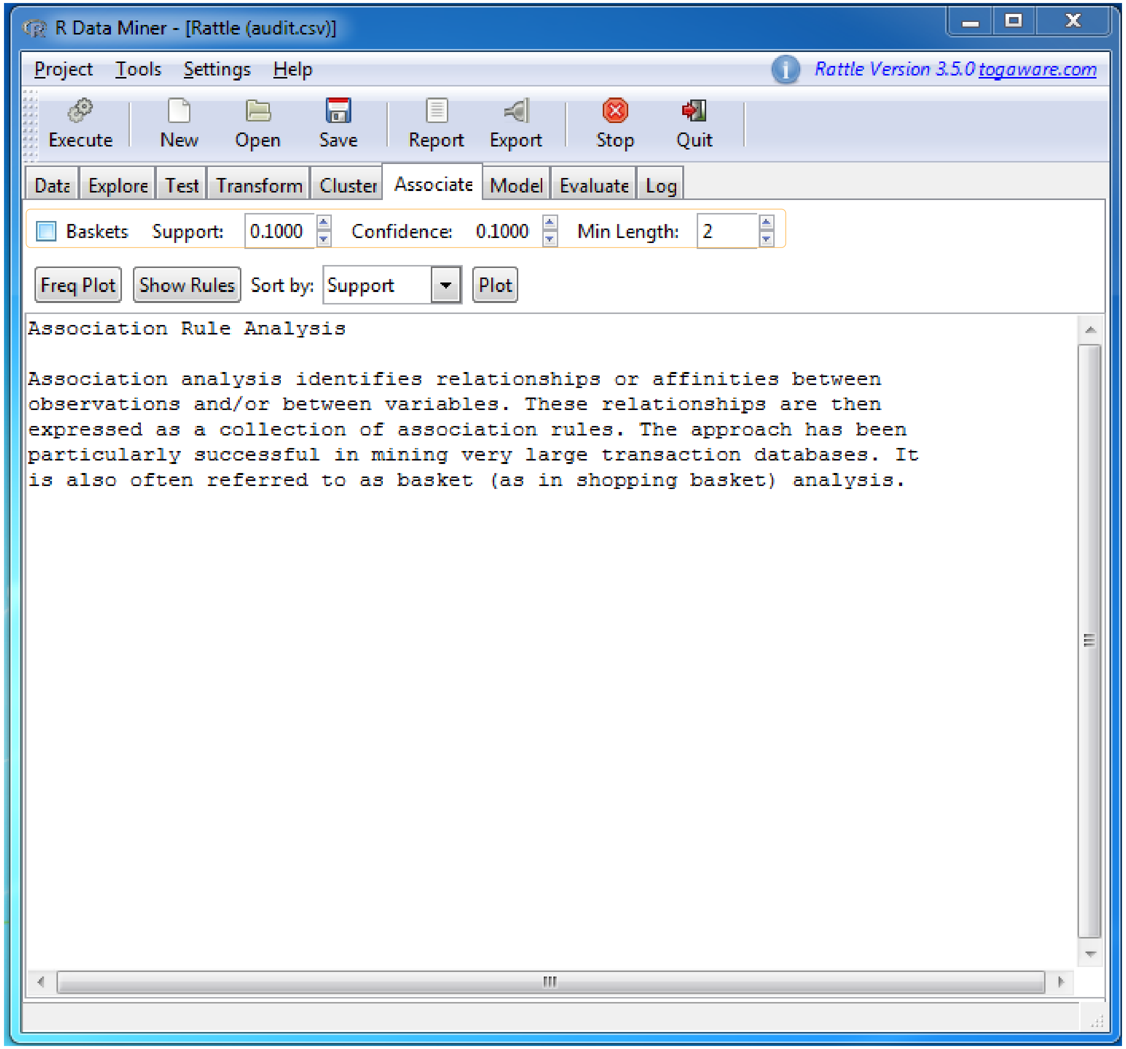
Rattle–Cluster Analysis

- Cluster tab provides option to build descriptive or unsupervised model.
- Several clustering algorithm available as options to identify groups within the dataset.



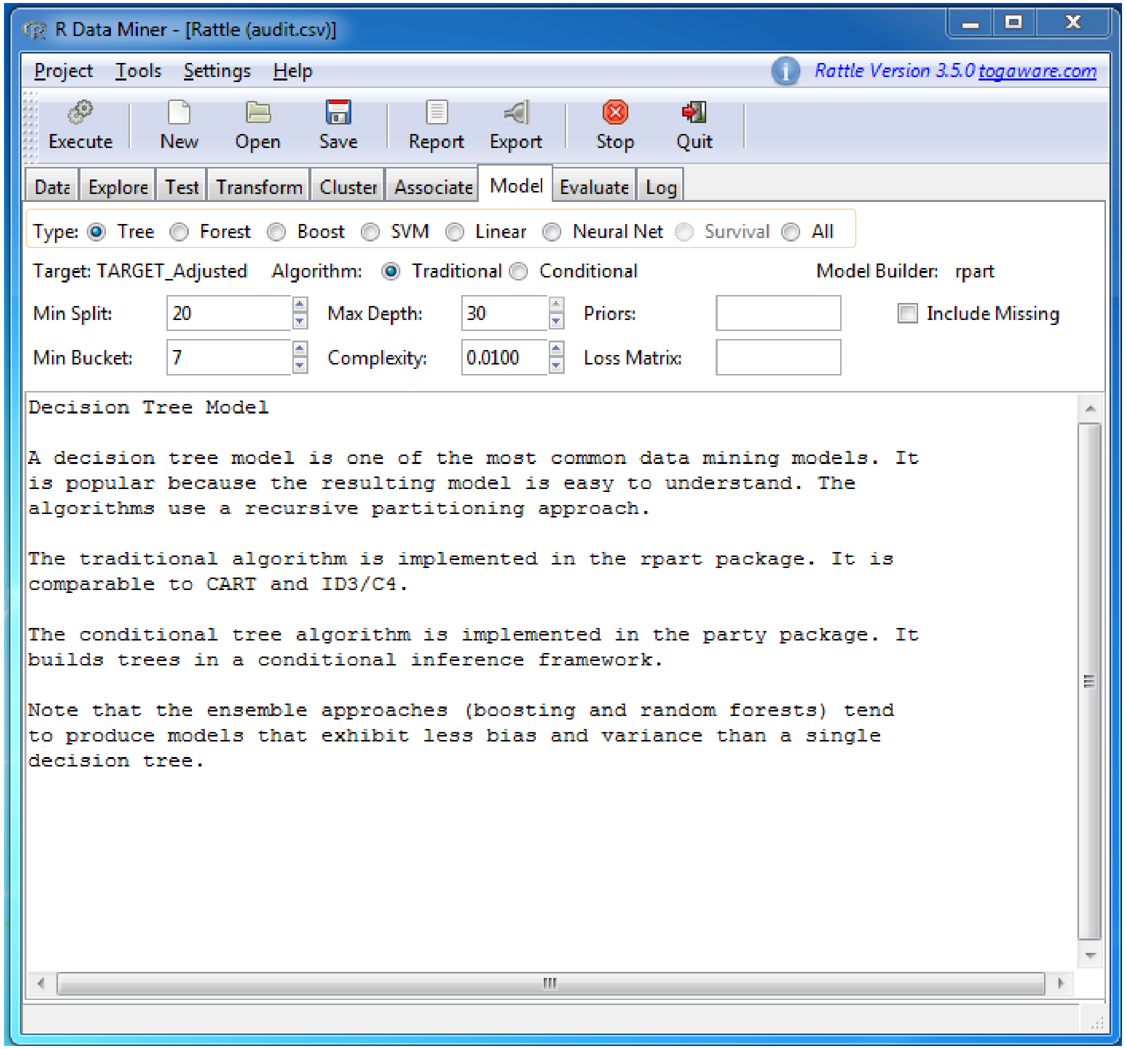
Rattle–Basket Analysis

- Associate tab gives another option to build descriptive or unsupervised model.
- Option available for market basket analysis to identify affinities between observations and/or between variables.



Rattle–Model Dataset

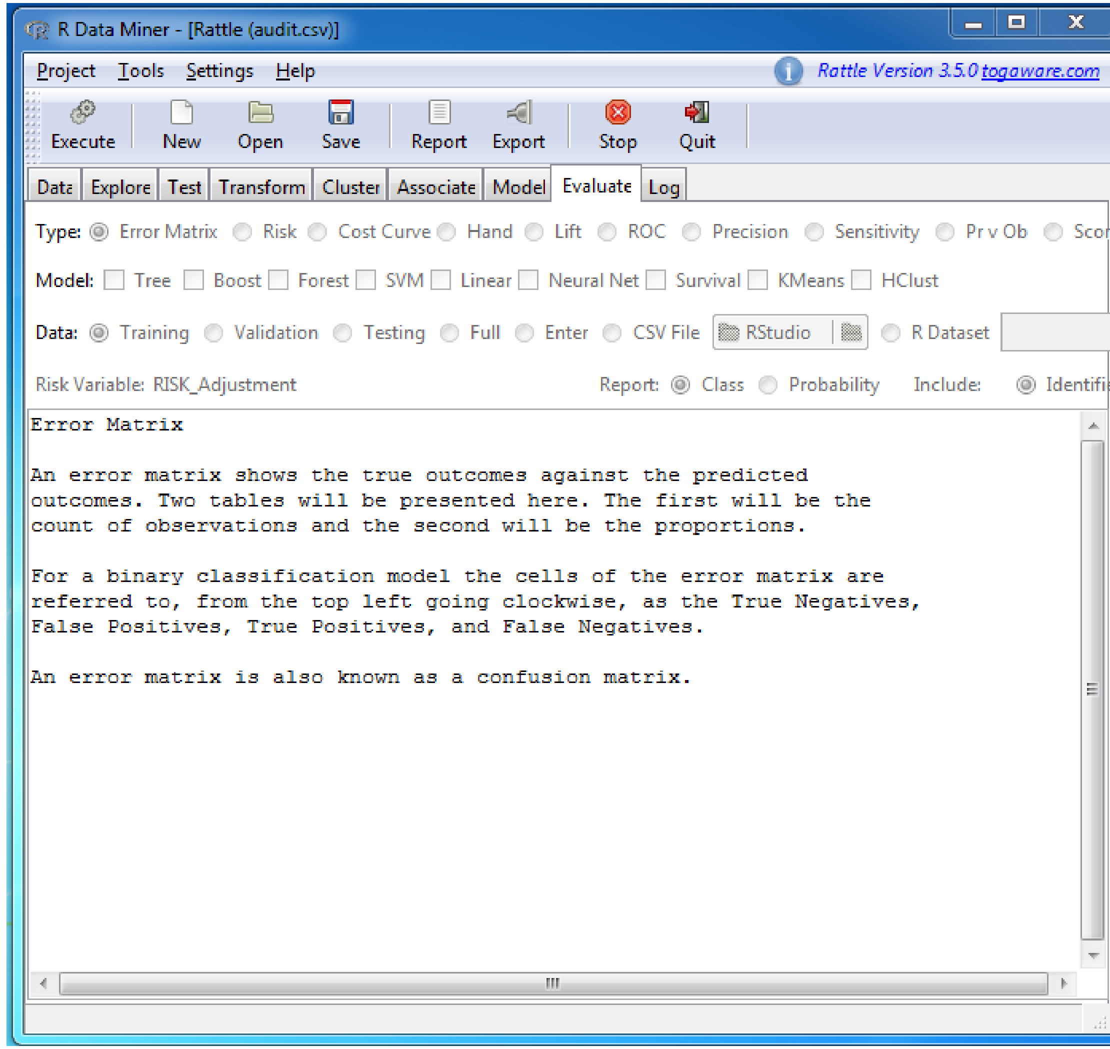
- Model tab provides a comprehensive list of techniques to build predictive models.



- *Provides an option to use all the model building techniques over the same dataset.*
- *The models can be evaluated for performance and the best model can be selected.*

Rattle–Evaluate Model

- Evaluate tab provides a collection of techniques for evaluating the performance of models



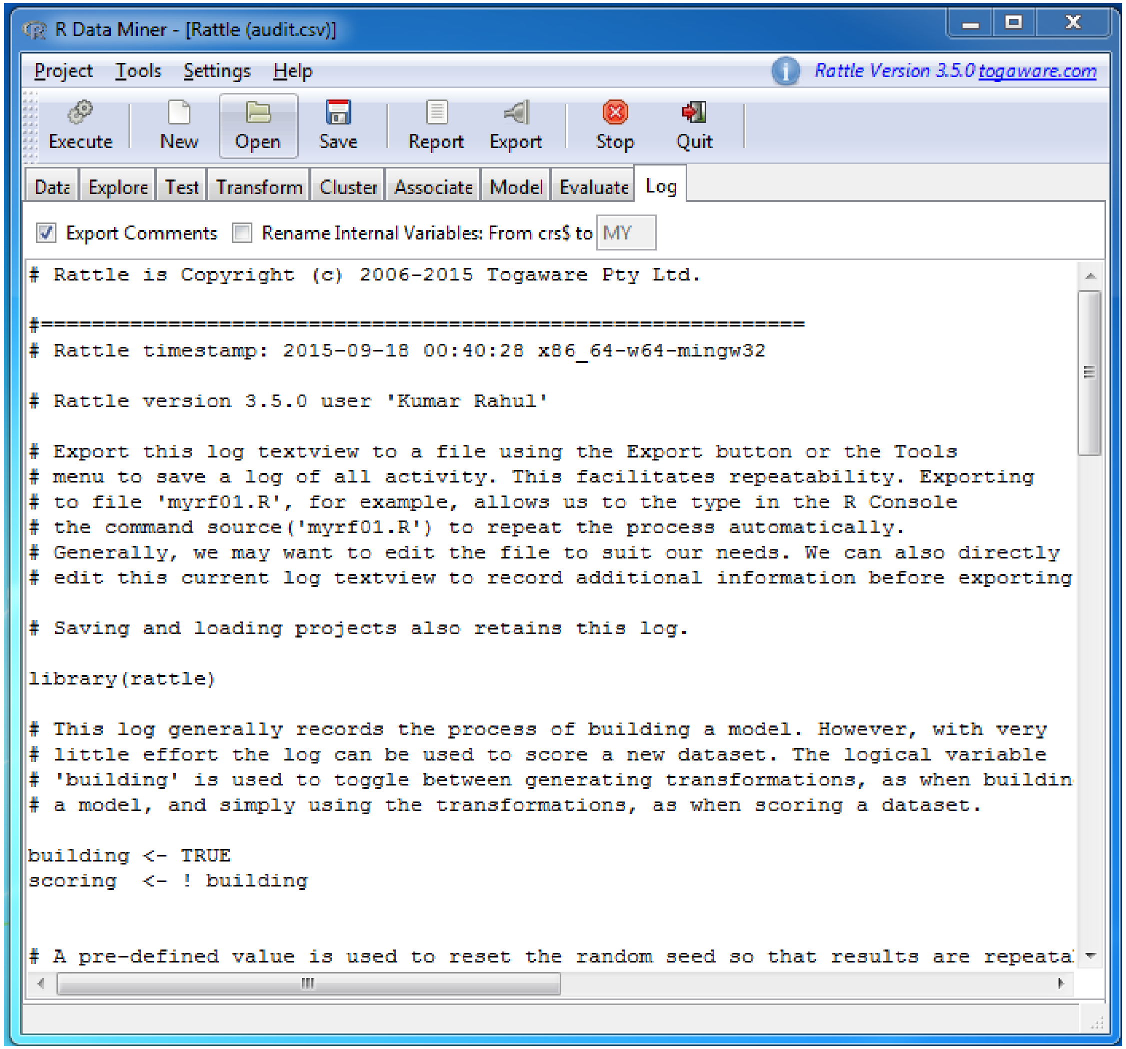
- *Some of the commonly used techniques for model comparison can be seen as options:*
  - *Error matrix*
  - *ROC curve*
  - *Lift Chart*

- *Rattle supports deployment of the model through the 'Score' option.*
  - *The complete model can be saved as a Rattle project and can later be used on the new dataset to score the*

Rattle–Log Generation

- Log tab records the process of building the model.
- The recorded script gives the flexibility to fine tune the analysis using R directly.

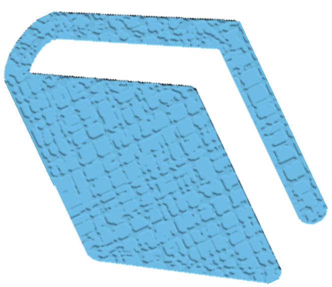
- *The log can be used for deployment to score a new dataset.*



Demo of the Rattle tool using an example dataset.

Summary

Summary of the topics covered in this lesson:



- R Analytical Tool to Learn Easily (Rattle) is a user interface based data mining tool built on top of R.
- Rattle provides a tab based options to load, explore, test, transform a dataset; followed by building and evaluating models.

QUIZ TIME

Quiz Question 1

Quiz 1

What is the command line syntax to install rattle?

Select all that apply.

a. *install.packages("rattle", dep=c("Suggests"))*

b. *install.packages("rattle“)*

c. *install.package(“rattle”)*

d. *install.package("rattle", dep=c("Suggests"))*

Quiz Question 1

Quiz 1	What is the command line syntax to install rattle? *Select all that apply.*

a. *install.packages("rattle", dep=c("Suggests"))*

b. *install.packages("rattle“)*

c. *install.package(“rattle”)*

d. *install.package("rattle", dep=c("Suggests"))*

Correct answer is:

a & b

Both a and b has the correct syntax. Option a has an optional argument of forcing the dependent packages to be installed.

End of Lesson12–Introduction to R Markdown and Rattle

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Data Science Using R

Lesson13–Introduction to Shiny R

Objective

After completing this lesson you will be able to:

- Explain the importance of Shiny R
- Describe the structure of Shiny application development using R
- Run the Shiny app from R Studio
- Deploy Shiny app on the web

Shiny R–Web Development Interface

- Shiny package provides a web development interface to R which can help build and run interactive web applications.
- The first step to build applications is by installing Shiny

On R Studio Console:

```
>Install.packages("Shiny")
```

Or install using the Rstudio Install packages options

Running the Shiny Examples

- There are eleven inbuilt examples in Shiny package each of which is a shiny app. An example can be accessed by the following command

Example 1: On R Studio Console:

```
>library("Shiny")
>runExample("01_Hello") #opens an interactive histogram
```

Example 2: On R Studio Console:

```
>library("Shiny")
>runExample("02_text")    #shows descriptive stats for
the selected datasets
```

- All shiny examples can be accessed by navigating to the location where R is installed.
- If R version 3.2.2 is installed in C drive then the path will be 'C:\Program Files\R\R-3.2.2\library\shiny\examples'.
- In Mac: 'Macintosh HD/library/Frameworks/Versions/Current/Resources/library/shiny/examples'

Shiny App Structure

- Shiny app has two basic components.

User Interface Script

- Controls the layout and appearance of the shiny app.
- The source script is named ‘ui.R’.

Server Script

- Responsible for the calculations which is to be performed to show the result on the UI.
- The source script is named as ‘server.R’.

The default working directory for shiny is the place where ui.R and server.R is saved for a specific shiny app.

Demo of the 01_Hello example

Making a Shiny App

- Every shiny app has the same structure of ui.R and server.R saved in one directory. To build your first shiny app, follow the below steps

1. *Copy '01_Hello' example and paste in your working directory.*
2. *Rename the 01_Helllo to any other folder name say, myapp.*
3. *Open the ui.R and server.R scripts in R Studio. Edit the server.R to change the color of the histogram from 'darkgrey' to 'blue'. Save the script.*
4. *On the console type:*
   >runApp("myapp")

Your first shiny app will be launched.

While giving name to the folder/directory ensure that the name is more than 3 characters long. If it is less than 3 character long, the ShinyApps throws error in deployment. More on deployment later.

Shiny App–UI Layout

- The ui.R has a basic structure as explained below:

```
# ui.R
shinyUI(fluidPage(            #creates a page which adjusts to browser
dimensions
  titlePanel("title panel"),          #title panel at the top

  sidebarLayout(                      #has two components as below
    sidebarPanel( "sidebar panel"),   #used to place the controls to
give interactions
    mainPanel("main panel")          #used to display the results
  )
))
```

More on how to write formatted paragraphs inside the sidebarPanel and mainPanel: http://shiny.rstudio.com/tutorial/lesson2/

Shiny App–UI Control Widgets

- Control widgets are used to send inputs by the user to shiny apps. The standard shiny widgets are:

Widgets	Functions
Action button	actionButton, submitButton
Checkboxes	checkboxInput, checkboxGroupInput
Date input	dateInput, dateRangeInput
File upload	fileInput
Field to enter input	numericInput, textInput
Radio buttons	radioButton
Slider bar	sliderInput
Box with choices to select from	selectInput

More on how to place control widgets inside the sidebarPanel and mainPanel: http://shiny.rstudio.com/tutorial/lesson3/

Shiny App–UI Reactive Output

- First step to make a reactive app is to add R objects using the output functions and control widgets through ui.R. Some of the output functions

Output Functions	Creates
htmlOutput	Raw HTML
imageOutput	Image
plotOutput	Plot
tableOutput	Table
textOutput	Text
uiOutput	Raw HTML
verbatimTextOutput	Text

In 01_Hello example, plotOutput("text1") was the output function used inside the main panel. More on this at : http://shiny.rstudio.com/tutorial/lesson4/

Shiny App–UI Reactive Output

- Second step is to build the object using the render function and pass the widget value to the code. This step is carried out in server.R. Some of the render function are:

Render Functions	Creates
renderImage	Images
renderPlot	Plot
renderPrint	Any printed output
renderTable	Data frame, Matrix, Other table like structure
renderText	Character strings
renderUI	Shiny tag object or HTML

Render function should correspond to the type of reactive object being made. In 01_Hello example, renderPlot was the inside the server.R to build the histogram. More on this at : http://shiny.rstudio.com/tutorial/lesson4/

Shiny App–Execution Flow

- The placement of code inside the server.R determines the efficiency at which the app will execute.

```
#server.R

#place to put a code

shinyServer(
  function(input,output) {

  #another place to put a code

   output$text1 <-renderText({
   #third place to put the code
   })
  }
)
```

Block 3 which is a render function runs on every change of widget value

Block 2 which is an unnamed function runs each time a user visits the app.

Block 1 which is server.R script runs once when the app is launched.

Load libraries, read datasets outside the shinyServer function, put user specific objects inside the unnamed function and control widget specific code in render function. More on this at : http://shiny.rstudio.com/tutorial/lesson5/

Shiny App–UI Reactive Expressions

- Reactive expressions are used to improve efficiency of code in server.R in cases where user controls the data upload through certain widgets.

```
#server.R
#place to put a code
shinyServer(
  function(input,output) {
  #another place to put a code
    reactiveInput <-reactive({
    #another place to put the code
    })

    output$text1 <-renderText({
    #reactiveInput used within
    })
  }
)
```

Block 4 which is a reactive function runs on certain widget value.

Block 3 which is a render function runs on every change of widget value.

Block 2 which is an unnamed function runs each time a user visits the app.

Block 1 which is server.R script runs once when the app is launched.

Shiny App–UI Reactive Expressions

- Code snippet to highlight the specific case of reading data from yahoo finance.

#without reactive function

```
output$plot <- renderPlot({
data <- getSymbols(input$symb,
src = "yahoo",
from = input$dates[1],
to = input$dates[2],
auto.assign = FALSE)

chartSeries(data, theme =
chartTheme("white"),
    type = "line", log.scale =
input$log, TA = NULL)
})
```

#with reactive function

```
dataInput <- reactive({
  getSymbols(input$symb, src =
"yahoo",
    from = input$dates[1],
    to = input$dates[2],
    auto.assign = FALSE)
})
output$plot <- renderPlot({
  chartSeries(dataInput(),
theme = chartTheme("white"),
    type = "line", log.scale =
input$log, TA = NULL)
})
```

More on reactive expressions at:
http://shiny.rstudio.com/tutorial/lesson6/

Shiny App Deployment

- The easiest way to make shiny app online is through shinyapps.io. Follow the below steps to create account at shinyapps and configure your system to host apps at shinyserver.

1. *Install devtools version 1.4 or later by running the command at the R studio console : install.packages('devtools')*
2. *Restart the R studio session and then install rsconnect through R studio console: devtools::install_github('rstudio/rsconnect')*
3. *Load rsconnect into the session through R studio console: library(rsconnect)*
4. *Create an account at http://www.shinyapps.io/*
5. *Configure rsconnect following the dashboard which appears after creating the account at step 4*
6. *Open a shiny app in your machine and from the console run the command to deploy the app*
   library(rsconnect)
   deployApp(<your app name>)

Or

shinyapps::deployApp(<your app name>)

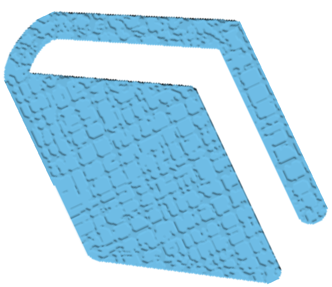
More on the deployment at: http://shiny.rstudio.com/articles/shinyapps.html

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Demo of the deployment on a shiny app

Summary

Summary of the topics covered in this lesson:



- Shiny is a web development interface to R which can be used to build and host applications online.
- All the applications developed using Shiny R will have two basic script: ui.R and server.R.
- ui.R helps build interactive UI using widgets whereas server.R helps in calculations whose results are shown on UI.
- The code within the render function is used to give interactivity to the applications.
- Reactive expressions are used to save on computing power of applications and make the app more efficient.
- Shiny apps can be deployed on the web by creating an account at shinyapps.io and configuring your system to deploy the apps.

QUIZ TIME

Quiz Question 1

Quiz 1 What is the command line syntax to install shiny?

Select all that apply.

a. install.packages(“Shiny”)

b. *install.package(“shiny”)*

c. *install.packages(‘Shiny’)*

d. *install.packages(“shiny”)*

Quiz Question 1

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Quiz 1	What is the command line syntax to install shiny? *Select all that apply.*

a. install.packages(“Shiny”)

b. *install.package(“shiny”)*

c. *install.packages(‘Shiny’)*

d. *install.packages(“shiny”)*

Correct answer is:

a & c

Both a and c has the correct syntax.

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

End of Lesson13–Introduction to Shiny R