Business Analytics & Data Mining Modeling Using R

Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

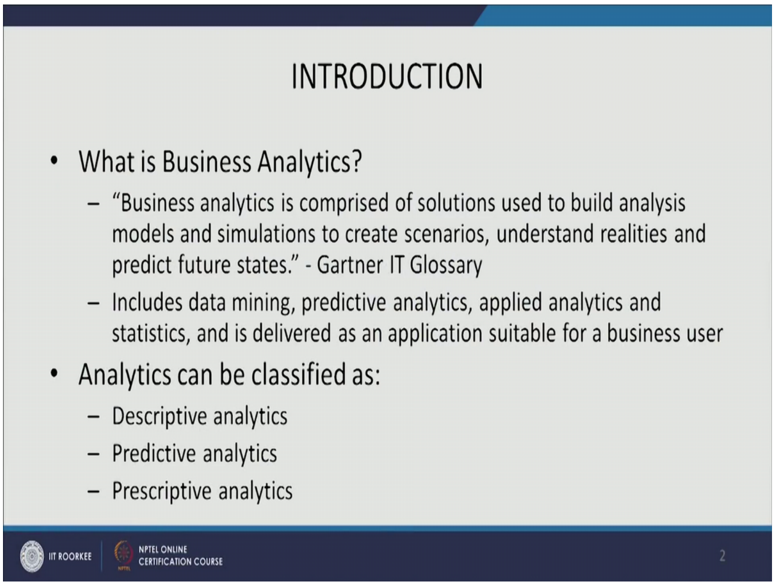
Lecture - 01

Introduction

Welcome to the course Business Analytics and Data Mining Modeling Using R. This is the very first lecture and we are going to cover the introduction part of it. First we need to understand why we need to study this course. If you follow the industry news and related reports you would see that there is his requirement of data scientist, data engineers, business analyst and other positions other relevant positions which require expertise in domains like business analytics, data mining R and other related areas.

So, we are going to cover some things in this course. So, let us start. So, what is business analytics?

(Refer Slide Time: 01:07)



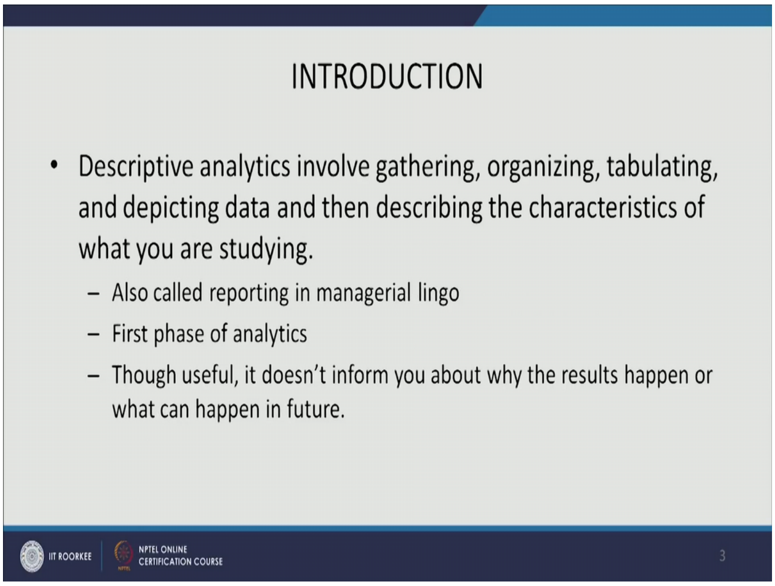
INTRODUCTION

- What is Business Analytics?
  - "Business analytics is comprised of solutions used to build analysis models and simulations to create scenarios, understand realities and predict future states." - Gartner IT Glossary
  - Includes data mining, predictive analytics, applied analytics and statistics, and is delivered as an application suitable for a business user
- Analytics can be classified as:
  - Descriptive analytics
  - Predictive analytics
  - Prescriptive analytics

The primary purpose of business analytics is to assist and aid in and drive in decision making activities of a business organization. Now, if you look at the Gartner's definition they have defined business analytics as comprised of solutions used to build analysis, models and simulations to create scenarios to understand realties and predict future sales.

The domains which are actually part of business analytics are data mining predictive analytics, applied analytics, test statistics and generally the solution is delivered in a application format which is suitable for business user. Now, if we look at analytics as such it can be classified as 3 categories descriptive analytics, predictive analytics and prescriptive analytics.

(Refer Slide Time: 02:01)



INTRODUCTION

- Descriptive analytics involve gathering, organizing, tabulating, and depicting data and then describing the characteristics of what you are studying.
  - Also called reporting in managerial lingo
  - First phase of analytics
  - Though useful, it doesn't inform you about why the results happen or what can happen in future.

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
3

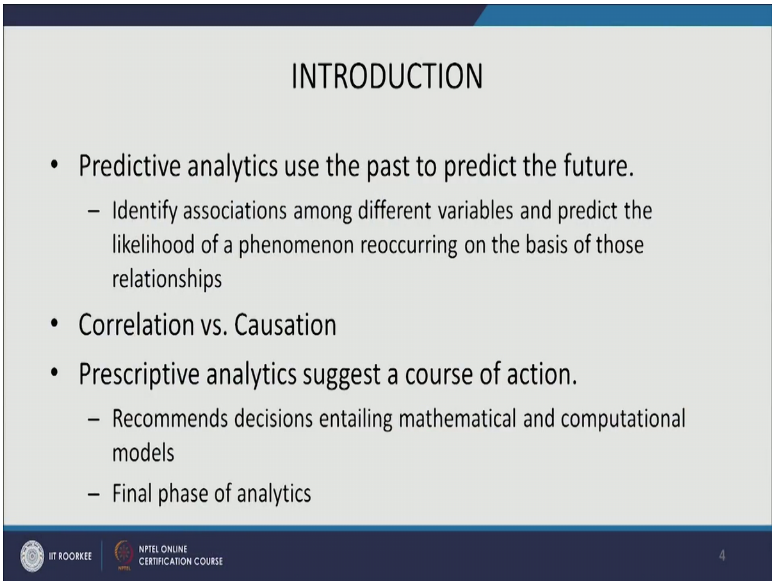
Now, first let us understand the descriptive analytics. Descriptive analytics mainly revolves around gathering, organizing, tabulating, presenting and depicting data and describing the characteristics of what you are studying. This is about a descriptive analytics is mainly about what is happening. So, we have to we try to answer the question what is happening in a particular context looking at the data.

Now, this is also called reporting and managerial lingo reason being because we generally look at the sales number in cost numbers, revenues number etcetera are different ratios. So, we try to understand how our business is performing, how the company organization is performing or how the industry or country economy is growing in a overall sense. So, therefore, what is happening is covered in descriptive analytics.

This is also first phase of analytics, this is where you actually start, you try to gather the sense of what is happening and then you look for other things other categories of analytics. Now, this descriptive analytics can be useful in the sense to inform why the results are happening, but you know they do not inform why the results are happening or what can happen in future. To answer these questions we need phase of analytics which is predictive.

So, predictive analysis can be defined as you know, predictive analytics can be used to predict the past two predict the future. Now, the main idea is to identify association among different variables and predict the likelihood of a phenomena occurring on the basis of those relationships.

(Refer Slide Time: 03:56)



INTRODUCTION

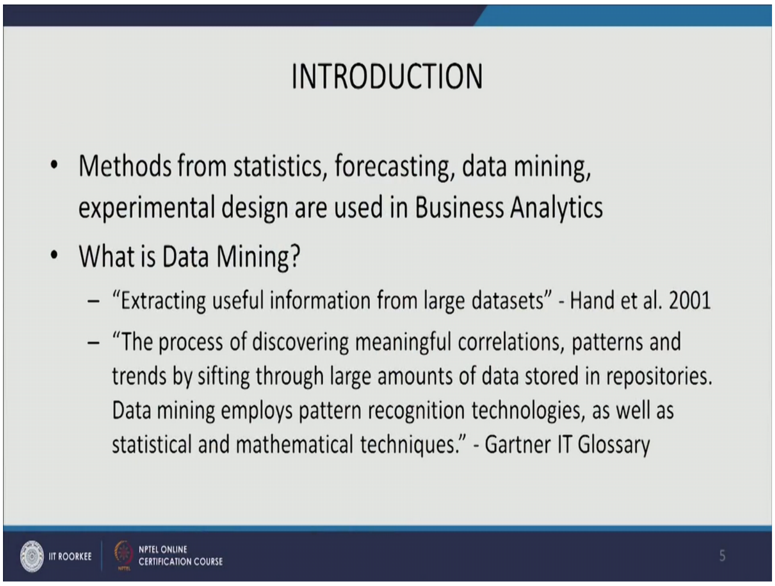
- Predictive analytics use the past to predict the future.
  - Identify associations among different variables and predict the likelihood of a phenomenon reoccurring on the basis of those relationships
- Correlation vs. Causation
- Prescriptive analytics suggest a course of action.
  - Recommends decisions entailing mathematical and computational models
  - Final phase of analytics

Now, at this point we need to understand the concept of correlation versus causation. Now, correlation is something you know if something is happening if x is happening, if x is correlated with y that is sufficient for us to you know predict something.

But if we are looking for what we should be doing about it if we know that what is going to happen in the future which is going to be you know part of predictive analysis then what we are going to do about it is becomes part of prescriptive analytics. Now, prescriptive analytics is where the cause and effect relationship that comes into the fact and the main idea about main idea of the prescriptive analytics is to suggest a course of action. So, generally you know prescriptive analytics is about recommending decisions and which entailes generally in mathematical and computational model you do lots of simulations optimizations to find out what can be done about this future scenario of the business or relevant topic.

So, prescriptive analytics is also defined as final phase of analytics. Now, methods from disciplines like statistics, forecasting, data mining, experimental design they are used in business analytics. Now, next part that brings us to the next part data mining the core of this particular course is data mining. So, let us understand what is data mining. Brief definition of data mining could be extracting useful information from large data sets.

(Refer Slide Time: 05:42)



INTRODUCTION

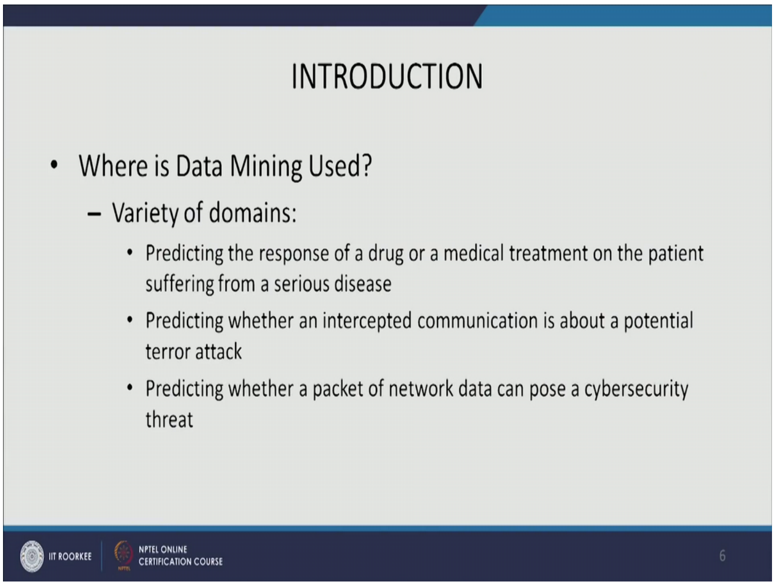
- Methods from statistics, forecasting, data mining, experimental design are used in Business Analytics
- What is Data Mining?
  - "Extracting useful information from large datasets" - Hand et al. 2001
  - "The process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies, as well as statistical and mathematical techniques." - Gartner IT Glossary

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 5

Now, if you look at the how Gartner is defining data mining. So, they define data mining as the process of discovering meaningful correlations, patterns and trends by shifting through large amount of data stored in repositories.

Now, data mining generally it employs a pattern recognition technology as well as a statistical and mathematical technique. Now, where is data mining used? What are the domains where data mining can actually be used?

(Refer Slide Time: 06:11)



INTRODUCTION

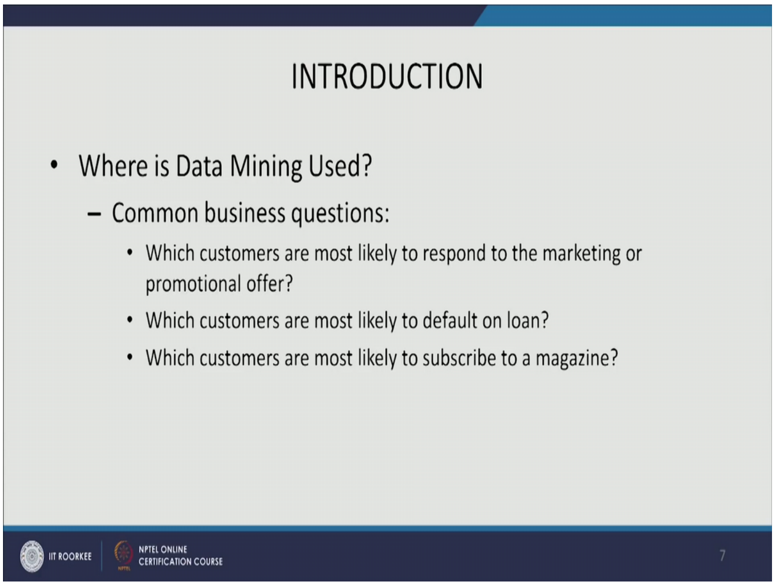
- Where is Data Mining Used?
  - Variety of domains:
    - Predicting the response of a drug or a medical treatment on the patient suffering from a serious disease
    - Predicting whether an intercepted communication is about a potential terror attack
    - Predicting whether a packet of network data can pose a cybersecurity threat

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
6

If you look at some of the examples given here first one is related to the medical field. So, data mining can actually help us in predicting the response of a drug or a medical treatment on the patient suffering from a serious disease or illness. Another example could be in the security domain where data mining can help us predicting whether an intercepted communication is about a potential terror attack. Now, another application of data mining could be in computer or network security field where it can help us predicting whether a packet of network data can pose a cyber security threat.

Now, our main interest is in the business domain. So, let us look at some of the examples where data mining can help in business context. So, common business questions where data mining can help could be like this which customers are most likely to respond to the marketing or promotional offer.

(Refer Slide Time: 07:12)



INTRODUCTION

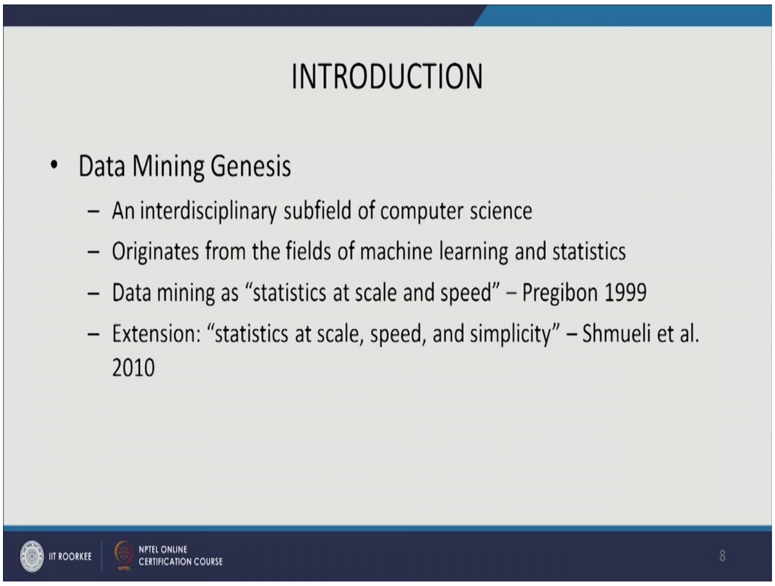
- Where is Data Mining Used?
  - Common business questions:
    - Which customers are most likely to respond to the marketing or promotional offer?
    - Which customers are most likely to default on loan?
    - Which customers are most likely to subscribe to a magazine?

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
7

Another example could be which customers are most likely to default on loan. So, banking and financial institutions they might be worried about some of their customers defaulting on loan. So, they would like to identify such customers and then take appropriate actions.

Now, another question could be which customers are most likely to likely to subscribe to a magazine. So, for a magazine if you know if they are running advertising and you know marketing promotional offer it would be important for them to understand the customer which are likely to subscribe to the kind of content that they are publishing or selling through their magazine. So, all these are some of the, some of the flavour of you know kind of questions where data mining can actually help us.

(Refer Slide Time: 08:05)

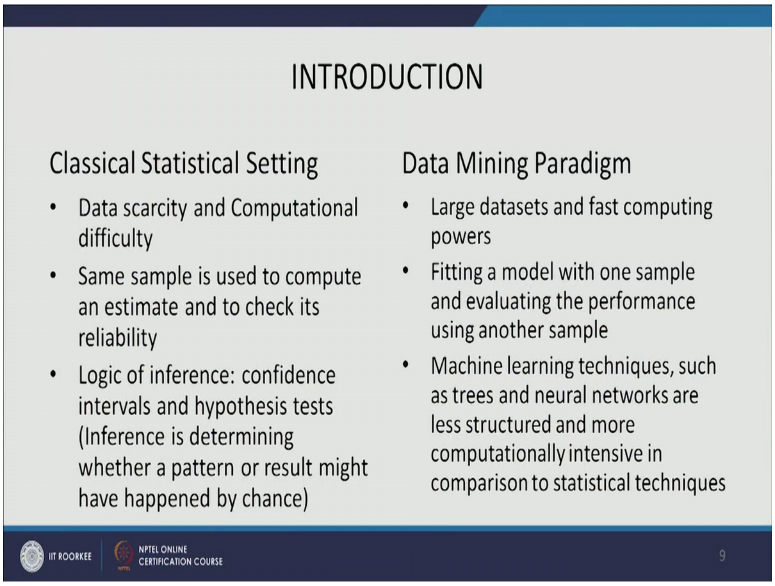


INTRODUCTION

- Data Mining Genesis
  - An interdisciplinary subfield of computer science
  - Originates from the fields of machine learning and statistics
  - Data mining as "statistics at scale and speed" – Pregibon 1999
  - Extension: "statistics at scale, speed, and simplicity" – Shmueli et al. 2010

Now, let us look at the origins of data mining. So, data mining genesis it is mainly an interdisciplinary field of computer science and it originates from the fields of machine learning and statistics. Now, some researchers have also define data mining as a statistics at scale and speed. Some people also define it has a statistics at scale speed and simplicity main reason being that in data mining we generally do not use the concepts of confidence and logic of entrance rather than rather we rely on partitioning and using different samples to test about models. So, that makes the whole process simpler. So, that is why this simplicity comes from.

(Refer Slide Time: 08:59)



INTRODUCTION

Classical Statistical Setting

- Data scarcity and Computational difficulty
- Same sample is used to compute an estimate and to check its reliability
- Logic of inference: confidence intervals and hypothesis tests (Inference is determining whether a pattern or result might have happened by chance)

Data Mining Paradigm

- Large datasets and fast computing powers
- Fitting a model with one sample and evaluating the performance using another sample
- Machine learning techniques, such as trees and neural networks are less structured and more computationally intensive in comparison to statistical techniques

Now, let us compare the classical statistical setting and data mining paradigm. Now, classical statistical setting is mainly about you know data scarcity and computational difficulty.

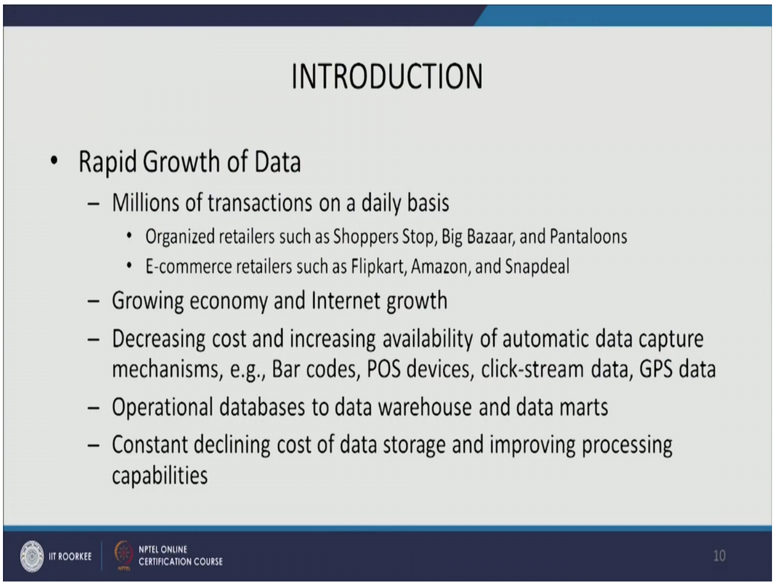
So, generally in a statistical setting you are dealing with a statistical question where you are looking for primary data and that is of course, costly to collect. So, there is always you are going to face this situation where data is not enough or it is very difficult to get the data. And if we look at the times manage statistical it is an old discipline old discipline, so the time and statistic statistical studies you know there used a face there was a time and they used to face lot of computational difficulty where most of the mathematical computation they have to perform manually. So, that was the time and the statistical setting originated. So, they generally face is classical statistical setting is generally they face the data problem, the data scarcity problem and the computational problem. When will look at the data mining paradigm it mainly its relatively a newer area and this as mainly you know developed or evolved because of the availability of large data set and ever improving computing powers. So, therefore, in data mining paradigm we are not faced with the problem of data sets or the computational problems.

Now, let us look at the another points you know in same sample in classical statistical setting same sample you used to compute an estimate and to check its reliability, while if you look at the data mining paradigm because we do not have any problem in terms of data points or data sets therefore, we can fit a fit a model with one sample and then evaluate the performance of the model with another sample. So, that is one another different. Now, third one is the logic of influence. Now, we need confidence intervals and hypothesis test to actually because we are using the same sample to compute the estimate and then check its reliability therefore, we need to eliminate this option where a pattern or a relationship can developed can we see in because of a chance. So, chance variable has to be eliminated. So, therefore, logic of inference it is important to the statistical setting and much stricter conditions are actually placed in a statistical modeling.

When we look at the data mining modeling we do not face with the problems of influence and related problems reason being that different samples are being used. So, therefore, reliability or robustness of the model is automatically taken care of because the model is built on one sample and it is evaluated on the different sample different partition. We look at the machine learning techniques such as trees neural networks they are also less structured, but more computationally intensive in compare comparison to statistical techniques. So, therefore, they might require more running time more compute more computation time and they are less structured while we look at the statistical technique regression and logistic regression discriminant analysis they are highly structured techniques.

Now, another important aspect of emergence of this field business analytics data mining big data and related fields is rapid growth of data.

(Refer Slide Time: 12:38)



INTRODUCTION

- Rapid Growth of Data
  - Millions of transactions on a daily basis
    - Organized retailers such as Shoppers Stop, Big Bazaar, and Pantaloons
    - E-commerce retailers such as Flipkart, Amazon, and Snapdeal
  - Growing economy and Internet growth
  - Decreasing cost and increasing availability of automatic data capture mechanisms, e.g., Bar codes, POS devices, click-stream data, GPS data
  - Operational databases to data warehouse and data marts
  - Constant declining cost of data storage and improving processing capabilities

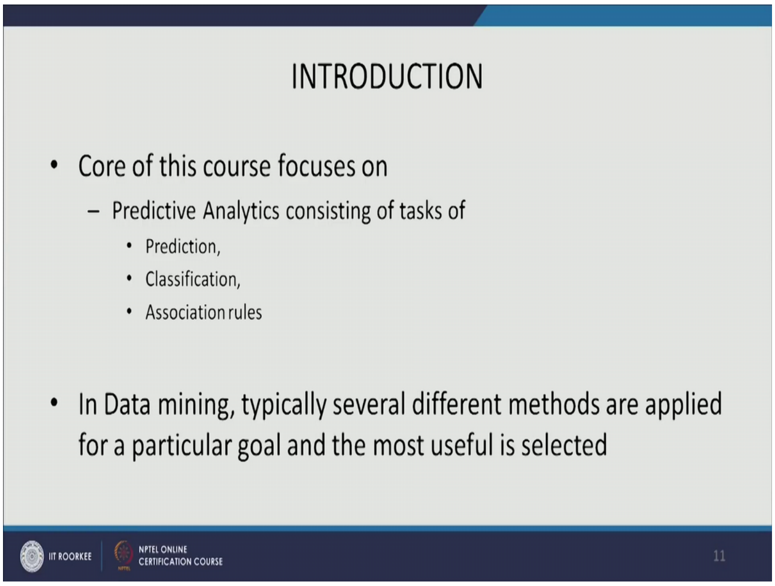
IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 10

Nowadays millions of transactions are recorded on a daily basis we have organized retailers such as shoppers stop big bazaar pantaloons where on a daily basis lots of the transactions are being recorded. We also have E-commerce retailers like flipkart, amazon, snapdeal. So, all these organizations have huge amount of data because of the transactions which are being recorded.

Our economy is has also been growing there has been growth in the internet infrastructure that has also let too many people using internet and digital technologies. So, that has also lead to growth of data. Automatic data capture mechanisms for example, bar code, POS devices, click-stream data, GPS data, that is also has led to rapid growth of data. Now, operational databases the ones we talked about whenever you visit a retail stores or whatever you items that you want to buy all those datas they are can be considered you know as transactions between the business and the customers all those days you know transactions or recorded in the operational database. Now, for the data analytics purposes or business analytics purposes these operational data base they have to now, brought into a data warehouse because you cannot actually do any kind of meaning full analysis on operational data bases. So, therefore, the data has to be brought into data warehouse and data marts, from there the data can then be sampled out for the analysis.

Now, another reason for rapid growth of data is constant declining cost of data storage and improving processing capabilities so that has also you know even smaller organization can nowadays invest in related IT infrastructure and have the analytical capabilities and developed analytical capabilities to improve their business. Now, core of this force core of this cores focuses on predictive analysis mainly we focus on three task prediction classification and association rules.

(Refer Slide Time: 14:58)



INTRODUCTION

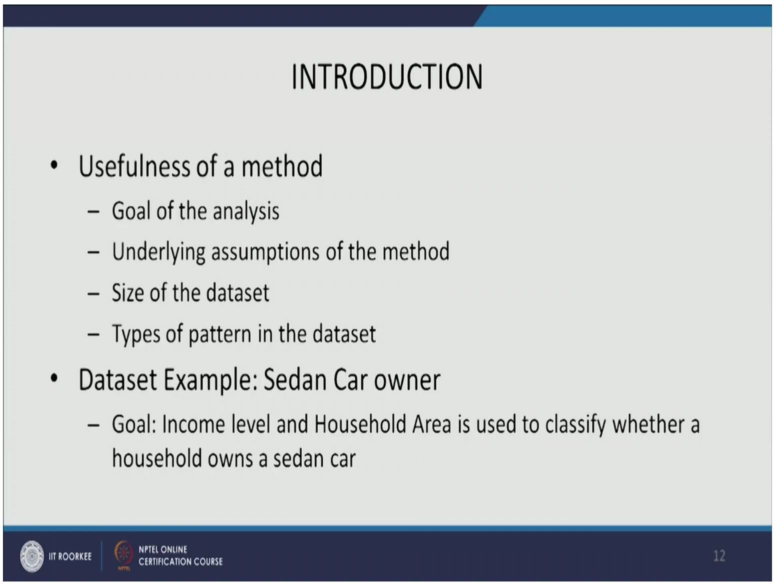
- Core of this course focuses on
  - Predictive Analytics consisting of tasks of
    - Prediction,
    - Classification,
    - Association rules
- In Data mining, typically several different methods are applied for a particular goal and the most useful is selected

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
11

Prediction is mainly when we are trying to predict value of a variable will discuss in detail different terminology and concepts later in this lecture. Classification is a task where in you are we are trying to predict we are trying to predict type of a particular variable. Association and rules is where we are trying to find out the association between different items in transactions.

Now, another important thing related to data mining is data mining and data mining process as such is we generally try different several methods for a particular goal and then a useful method is finally, selected and then used in the production systems.

(Refer Slide Time: 15:46)



INTRODUCTION

- Usefulness of a method
  - Goal of the analysis
  - Underlying assumptions of the method
  - Size of the dataset
  - Types of pattern in the dataset
- Dataset Example: Sedan Car owner
  - Goal: Income level and Household Area is used to classify whether a household owns a sedan car

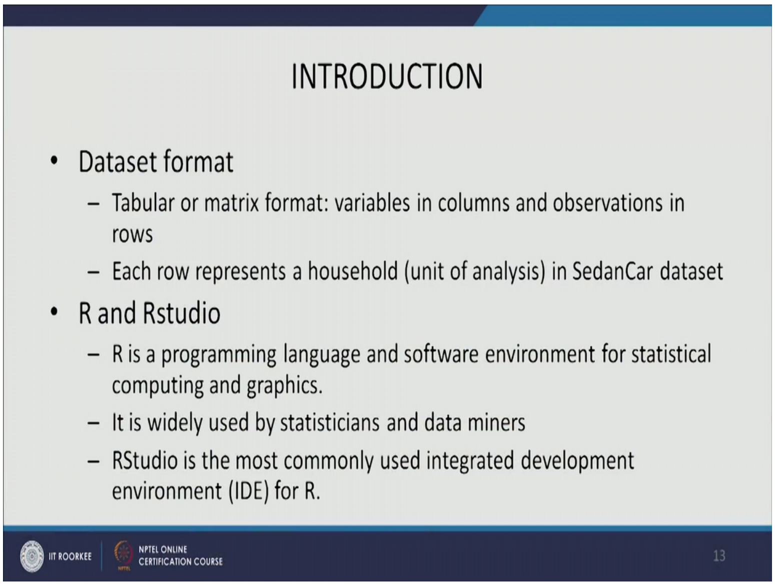
IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
12

Now, how do we define usefulness of a method? Now, the methods that we select to perform a particular class of course, they have to be relevant with respect to goal of the analysis. underlying assumptions the matter method they also have to be there, there should also meet the requirement of the goal and the problem. Size of the data set different methods and algorithms they are going to impose their own restrictions on number of variables and the number of records that are going to be used for analysis. So, size of the data set is also going to determine.

Types of pattern in the data set. So, different methods or algorithm they are suitable for suitable for finding out or understanding different types pattern. So, therefore, type of pattern in a data set is also going to determine the usefulness of a method with respect to a particular goal.

Let us look at an example to develop a better understanding. Now, we have this hypothetical example sedan car owner. So, where in main goal is income level and household area is used to classify whether a household owns a sedan car. So, we have these two variables income level and household area and we are trying to classify. So, this is essentially a classification problem and we want to classify whether a particular household owns a sedan car or not.

(Refer Slide Time: 17:24)



INTRODUCTION

- Dataset format
  - Tabular or matrix format: variables in columns and observations in rows
  - Each row represents a household (unit of analysis) in SedanCar dataset
- R and Rstudio
  - R is a programming language and software environment for statistical computing and graphics.
  - It is widely used by statisticians and data miners
  - RStudio is the most commonly used integrated development environment (IDE) for R.

Now, let us understand the data set format that is typically used in data mining and business analytics. Now, generally the data that we use is in tabular or matrix format variables are generally or in columns and observations are in rows.

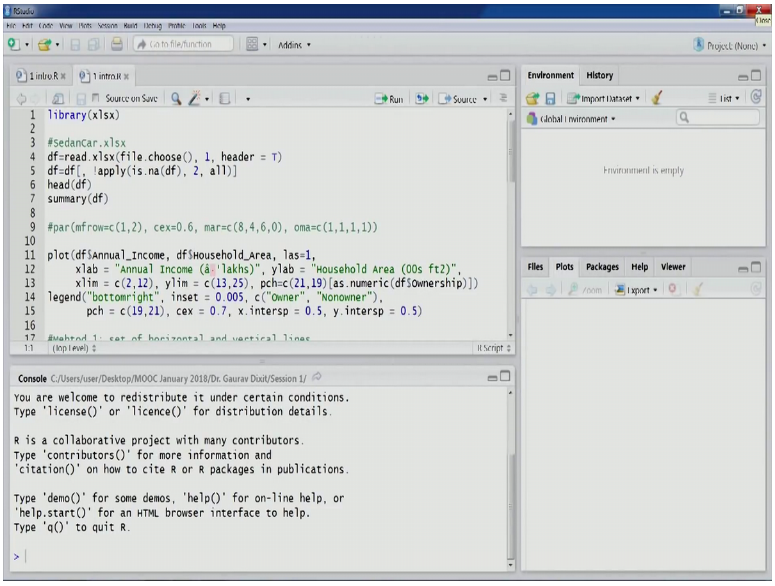
Now, another important thing is each row represents a household this is unit of analysis for example, in this case the sedan car data set case unit of analysis house hold. So, all the data, all the variables that are there in the data set there about this particular household, so household being the unit of analysis.

Now, the statistical software data mining software that we are going to use in this courses R and R studio. So, R what is R? R is a programming language and software environment for a statistical computing and graphics, widely used by statistician across the world and also by data miners. So, there are many packages available for different kinds of functionalities, related to statistical techniques and also related to data mining techniques.

R studio is another you know this is again most commonly used integrated development environment for R. So, it might be difficult for some users to directly start using with R because of the interface they might not be very much comfortable with the interface of

R, R studio bridges this gap and provides a much better interface to perform your data mining, modeling or statistical modeling using R.

(Refer Slide Time: 19:04)



```
library(xlsx)

#SedanCar.xlsx
df=read.xlsx(file.choose(), 1, header = T)
df=df[, !apply(is.na(df), 2, all)]
head(df)
summary(df)

#par(mfrow=c(1,2), cex=0.6, mar=c(8,4,6,0), oma=c(1,1,1,1))

plot(df$Annual_Income, df$Household_Area, las=1,
     xlab = "Annual Income (â·'lakhs)", ylab = "Household Area (00s ft2)",
     xlim = c(2,12), ylim = c(13,25), pch=c(21,19)[as.numeric(df$Ownership)])
legend("bottomright", inset = 0.005, c("Owner", "Nonowner"),
       pch = c(19,21), cex = 0.7, x.intersp = 0.5, y.intersp = 0.5)

#Mehtod 1: set of horizontal and vertical lines
```

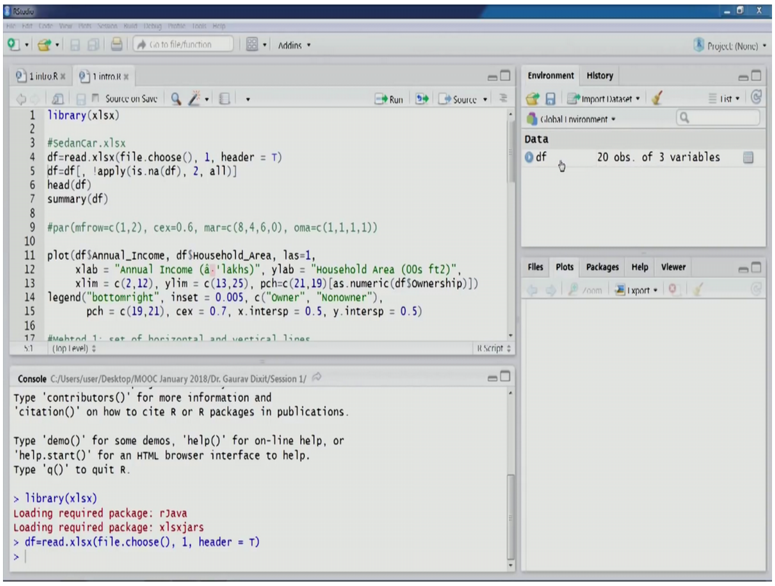
Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 1/

```
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

So, let us look at this example a sedan car. So, this is R studio environment. So, here you have these four parts here first part this part is about the R script, the R script the code is actually written here and this part is actually to run this commands that are given in this script. Then in this part you have in the environment where all your you know data and variables they would be loaded up if that if a particular data set is loaded into R studio it would be shown here, if a particular variable is loaded into R studio it would actually be shown here. In this part you have plots help page for a detailed before.

So, I recommend you to go through the supplementary lecture of introduction to R and basic statistics if you are not able to understand what is going on in this particular case. So, first we need to load this library (Refer Time: 20:16) because we are using the data set which is available in the axial format, you would see in this environment window that this data set has been loaded we have 20 observation of 3 variables. So, here you can look at the tabular data.

(Refer Slide Time: 20:36)



```
library(xlsx)

#SedanCar.xlsx
df=read.xlsx(file.choose(), 1, header = T)
df=df[, !apply(is.na(df), 2, all)]
head(df)
summary(df)

#par(mfrow=c(1,2), cex=0.6, mar=c(8,4,6,0), oma=c(1,1,1,1))

plot(df$Annual_Income, df$Household_Area, las=1,
     xlab = "Annual Income (â'lakhs)", ylab = "Household Area (00s ft2)",
     xlim = c(2,12), ylim = c(13,25), pch=c(21,19)[as.numeric(df$Ownership)])
legend("bottomright", inset = 0.005, c("Owner", "Nonowner"),
       pch = c(19,21), cex = 0.7, x.intersp = 0.5, y.intersp = 0.5)

#Mehtod 1: set of horizontal and vertical lines
```

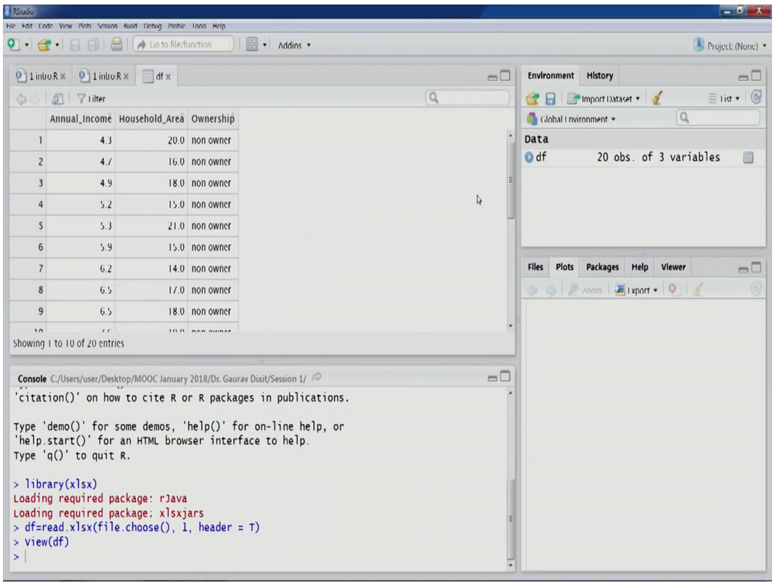
Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 1/

```
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(xlsx)
Loading required package: rJava
Loading required package: xlsxjars
> df=read.xlsx(file.choose(), 1, header = T)
> 
```

Data
df 20 obs. of 3 variables

So, this is the data that we were talking about. You can see that this is in the tabular format or matrix format.

(Refer Slide Time: 20:44)



	Annual_Income	Household_Area	Ownership
1	4.3	20.0	non owner
2	4.7	16.0	non owner
3	4.9	18.0	non owner
4	5.2	15.0	non owner
5	5.3	21.0	non owner
6	5.9	15.0	non owner
7	6.2	14.0	non owner
8	6.5	17.0	non owner
9	6.5	18.0	non owner

Showing 1 to 10 of 20 entries

Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 1/

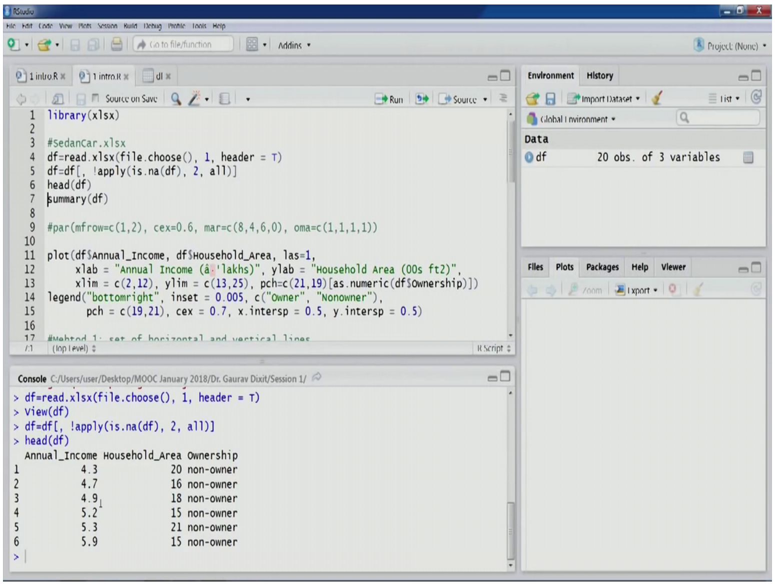
```
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(xlsx)
Loading required package: rJava
Loading required package: xlsxjars
> df=read.xlsx(file.choose(), 1, header = T)
> View(df)
> 
```

Data
df 20 obs. of 3 variables

So, you have these three variables. So, annual income and household area these are your predictor variables. Will discuss the termilogy later in the lecture, and this is your outcome variable ownership. So, we want to predict based on these two variables we want to predict the class of a sedan car you know means class of a you know household whether they own a sedan car or not. If we looking at the whole data set we have 20 observations and each observation is about a household, it depicts their annual income in rupees lakh and the household area in 100 a square feet.

Now, let us go back. Now, if we run this command head df we are going to get first 6 observation from the data set. So, if we you do not want to go and actually have a look at the full data set which might be a large file.

(Refer Slide Time: 22:05)



```
library(xlsx)

#SedanCar.xlsx
df=read.xlsx(file.choose(), 1, header = T)
df=df[, !apply(is.na(df), 2, all)]
head(df)
summary(df)

#par(mfrow=c(1,2), cex=0.6, mar=c(8,4,6,0), oma=c(1,1,1,1))

plot(df$Annual_Income, df$Household_Area, las=1,
     xlab = "Annual Income (â'lakhs)", ylab = "Household Area (00s ft2)",
     xlim = c(2,12), ylim = c(13,25), pch=c(21,19)[as.numeric(df$Ownership)])
legend("bottomright", inset = 0.005, c("Owner", "Nonowner"),
       pch = c(19,21), cex = 0.7, x.intersp = 0.5, y.intersp = 0.5)

#Mehtod 1: set of horizontal and vertical lines
```

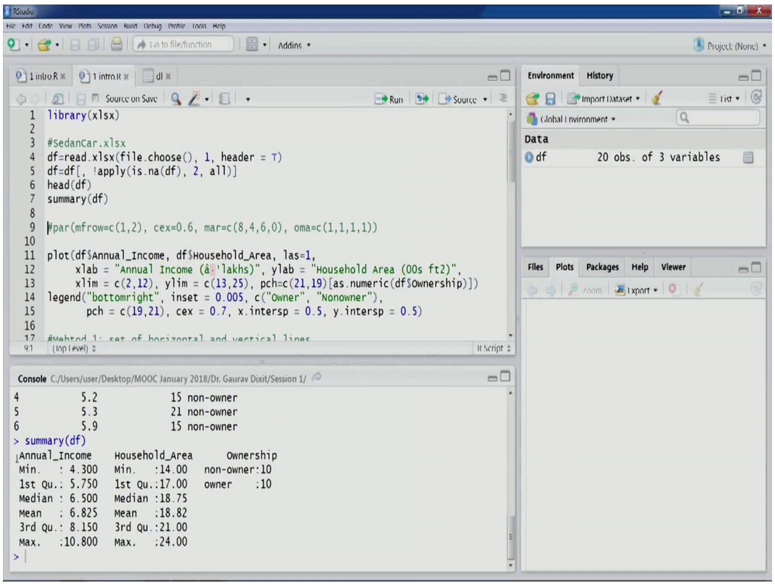
Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 1/

```
> df=read.xlsx(file.choose(), 1, header = T)
> View(df)
> df=df[, !apply(is.na(df), 2, all)]
> head(df)
  Annual_Income Household_Area Ownership
1           4.3             20 non-owner
2           4.7             16 non-owner
3           4.9             18 non-owner
4           5.2             15 non-owner
5           5.3             21 non-owner
6           5.9             15 non-owner
> 
```

Data
df 20 obs. of 3 variables

So, you might look at you might run this command head, the head df and then you can look at the first 6 observation of the data set.

(Refer Slide Time: 22:16)



```
library(xlsx)

#SedanCar.xlsx
df=read.xlsx(file.choose(), 1, header = T)
df=df[, !apply(is.na(df), 2, all)]
head(df)
summary(df)

#par(mfrow=c(1,2), cex=0.6, mar=c(8,4,6,0), oma=c(1,1,1,1))

plot(df$Annual_Income, df$Household_Area, las=1,
     xlab = "Annual Income (â'lakhs)", ylab = "Household Area (00s ft2)",
     xlim = c(2,12), ylim = c(13,25), pch=c(21,19)[as.numeric(df$Ownership)])
legend("bottomright", inset = 0.005, c("Owner", "Nonowner"),
       pch = c(19,21), cex = 0.7, x.intersp = 0.5, y.intersp = 0.5)

#Mehtod 1: set of horizontal and vertical lines
```

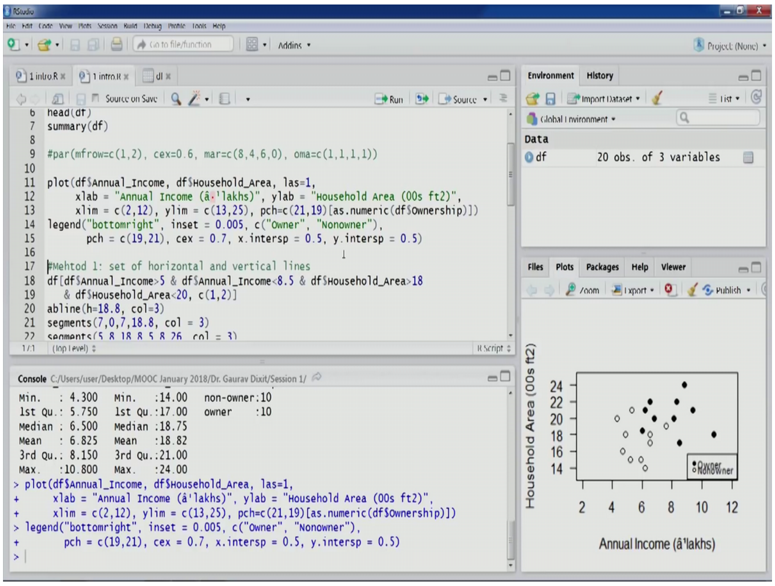
Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 1/

```
4           5.2             15 non-owner
5           5.3             21 non-owner
6           5.9             15 non-owner
> summary(df)
 Annual_Income    Household_Area      Ownership
 Min.   : 4.300   Min.   :14.00   non-owner:10
 1st Qu.: 5.750   1st Qu.:17.00   owner    :10
 Median : 6.500   Median :18.75
 Mean   : 6.825   Mean   :18.82
 3rd Qu.: 8.150   3rd Qu.:21.00
 Max.   :10.800   Max.   :24.00
> 
```

Data
df 20 obs. of 3 variables

Summary, this command is going to give you this these basic statistics about the data set variable for example, annual income we can see min max mean and other statistics similarly for household area and ownerships. So, you can see from the ownership data this is a categorical data we will discuss later in this lecture what it is a categorical variable. We can see 10 observation belong to the non-own category and intern observation belong to the owner class. We can also look at the other things.

(Refer Slide Time: 22:57)



```
head(df)
summary(df)

#par(mfrow=c(1,2), cex=0.6, mar=c(8,4,6,0), oma=c(1,1,1,1))

plot(df$Annual_Income, df$Household_Area, las=1,
     xlab = "Annual Income (â·'lakhs)", ylab = "Household Area (00s ft2)",
     xlim = c(2,12), ylim = c(13,25), pch=c(21,19)[as.numeric(df$Ownership)])
legend("bottomright", inset = 0.005, c("Owner", "Nonowner"),
       pch = c(19,21), cex = 0.7, x.intersp = 0.5, y.intersp = 0.5)

#Mehtod 1: set of horizontal and vertical lines
df[df$Annual_Income>5 & df$Annual_Income<8.5 & df$Household_Area>18
   & df$Household_Area<20, c(1,2)]
abline(h=18.8, col=3)
segments(7,0,7,18.8, col = 3)
segments(5.8,18.8,5.8,26, col = 3)
```

Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 1/

```
Min.   : 4.300   Min.   :14.00   non-owner:10
1st Qu.: 5.750   1st Qu.:17.00   owner    :10
Median : 6.500   Median :18.75
Mean   : 6.825   Mean   :18.82
3rd Qu.: 8.150   3rd Qu.:21.00
Max.   :10.800   Max.   :24.00
> plot(df$Annual_Income, df$Household_Area, las=1,
+      xlab = "Annual Income (â'lakhs)", ylab = "Household Area (00s ft2)",
+      xlim = c(2,12), ylim = c(13,25), pch=c(21,19)[as.numeric(df$Ownership)])
> legend("bottomright", inset = 0.005, c("Owner", "Nonowner"),
+        pch = c(19,21), cex = 0.7, x.intersp = 0.5, y.intersp = 0.5)
> |
```

Data
df 20 obs. of 3 variables

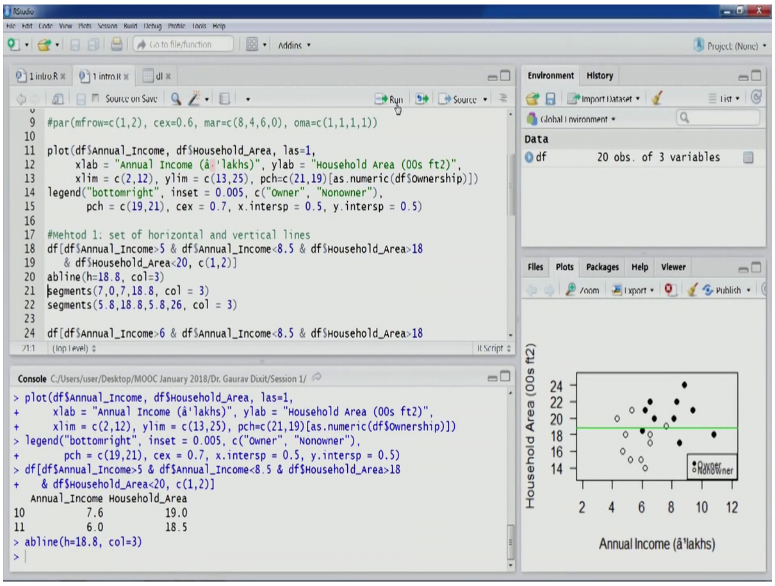
Now, let us plot a graph between you know household area and annual income. Annual income being on the x axis and household being in the on the y axis, so this is the plot, you can look at if you look at this plot some of the observation which are belonging to the a non-owner category non owner class they are mainly on this part and the observation belonging to owner they are mainly on this half. So, our goal is to classify the ownerships of a sedan car.

Now, as we talked about that different methods could be tried out in a data mining modeling and then the best one is generally selected. Now, one method in this case could be a set of horizontal and vertical lines. So, we can look at this data and we can create you know set of horizontal and vertical lines this could be a hypothetical method one which could, then be used to classify which than could be used to classify this observations. Another method could be a single diagonal line. So, we could draw a single diagonal line somewhere here and then it could also be used to classify these observations.

Now, if you look at this data if you are able to draw a line somewhere here most of the observations belong to owner call owner you know owner class they will be on the upper rectangular region and the most of the observation belonging to the non owner class they would be on the lower rectangular region. So, let us do that.

So, if we look at these two points. So, these two points are actually 7.6 and 6. So, these two points have been look at if a line can be a horizontal line can be drawn here then two partitions two rectangular regions can actually be created.

(Refer Slide Time: 25:20)

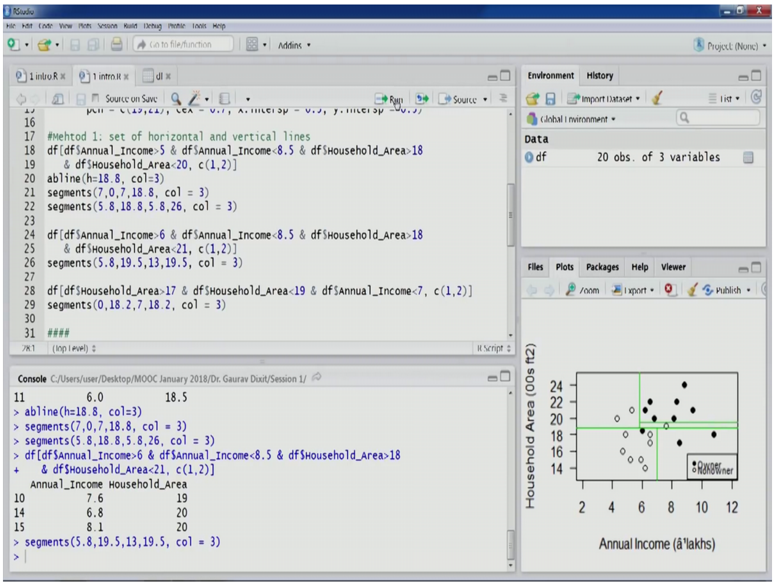


```
#par(mfrow=c(1,2), cex=0.6, mar=c(8,4,6,0), oma=c(1,1,1,1))

plot(df$Annual_Income, df$Household_Area, las=1,
     xlab = "Annual Income (â'lakhs)", ylab = "Household Area (00s ft2)",
     xlim = c(2,12), ylim = c(13,25), pch=c(21,19)[as.numeric(df$Ownership)])
legend("bottomright", inset = 0.005, c("Owner", "Nonowner"),
       pch = c(19,21), cex = 0.7, x.intersp = 0.5, y.intersp = 0.5)

#Mehtod 1: set of horizontal and vertical lines
df[df$Annual_Income>5 & df$Annual_Income<8.5 & df$Household_Area>18
   & df$Household_Area<20, c(1,2)]
abline(h=18.8, col=3)
segments(7,0,7,18.8, col = 3)
segments(5.8,18.8,5.8,26, col = 3)

df[df$Annual_Income>6 & df$Annual_Income<8.5 & df$Household_Area>18
```

```
> plot(df$Annual_Income, df$Household_Area, las=1,
+      xlab = "Annual Income (â'lakhs)", ylab = "Household Area (00s ft2)",
+      xlim = c(2,12), ylim = c(13,25), pch=c(21,19)[as.numeric(df$Ownership)])
> legend("bottomright", inset = 0.005, c("Owner", "Nonowner"),
+        pch = c(19,21), cex = 0.7, x.intersp = 0.5, y.intersp = 0.5)
> df[df$Annual_Income>5 & df$Annual_Income<8.5 & df$Household_Area>18
+    & df$Household_Area<20, c(1,2)]
   Annual_Income Household_Area
10           7.6           19.0
11           6.0           18.5
> abline(h=18.8, col=3)
> 
```

If you similarly keep on drawing the horizontal and vertical lines to keep on separating these observations you can see here this particular you know only 3 observation belonging to normal or there others are owner observations in this lower rectangular region most of the observation are belonging to the non owner and only 3 belongs to the category.

(Refer Slide Time: 25:48)



```
       pch = c(19,21), cex = 0.7, x.intersp = 0.5, y.intersp = 0.5)

#Mehtod 1: set of horizontal and vertical lines
df[df$Annual_Income>5 & df$Annual_Income<8.5 & df$Household_Area>18
   & df$Household_Area<20, c(1,2)]
abline(h=18.8, col=3)
segments(7,0,7,18.8, col = 3)
segments(5.8,18.8,5.8,26, col = 3)

df[df$Annual_Income>6 & df$Annual_Income<8.5 & df$Household_Area>18
   & df$Household_Area<21, c(1,2)]
segments(5.8,19.5,13,19.5, col = 3)

df[df$Household_Area>17 & df$Household_Area<19 & df$Annual_Income<7, c(1,2)]
segments(0,18.2,7,18.2, col = 3)

####
```

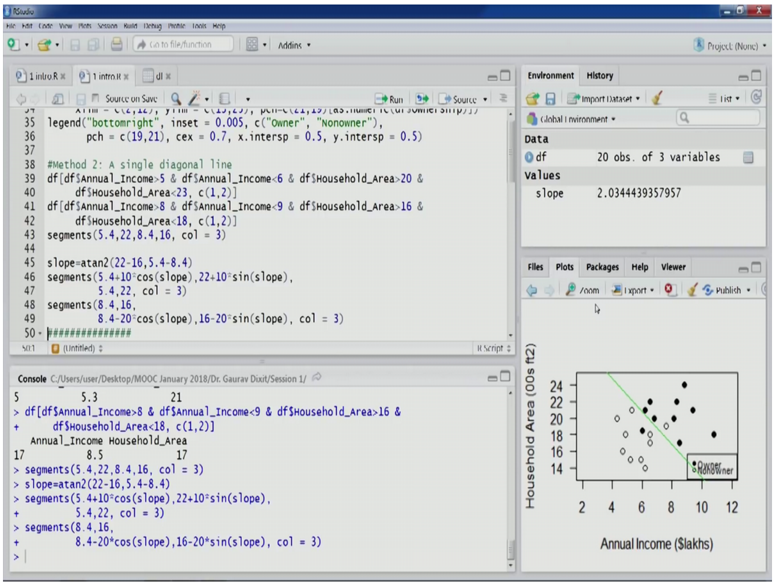
Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 1/

```
11          6.0          18.5
> abline(h=18.8, col=3)
> segments(7,0,7,18.8, col = 3)
> segments(5.8,18.8,5.8,26, col = 3)
> df[df$Annual_Income>6 & df$Annual_Income<8.5 & df$Household_Area>18
+    & df$Household_Area<21, c(1,2)]
   Annual_Income Household_Area
10          7.6            19
14          6.8            20
15          8.1            20
> segments(5.8,19.5,13,19.5, col = 3)
> 
```

Generally we can keep on creating similar lines to further classify. So, now finally, will end up with a particular graphics where each rectangular region is homogeneous; that means, it contains the observation belonging to only one class either owner or non owner. So, this could be these set up of horizontal and vertical lines could be a method, could be a model to classify these observations.

Similarly as we talked about the method two could be about finding a single diagonal line to separate these observations. Now, if we look at the again if we look the plot again, the line a diagonal line could be somewhere here it can go from somewhere here and then we will have homogenous partitions homogeneous rectangle regions. So, again a similar process can be adopted to find out that particular line you would see a line being drawn there we can extend this line and create partitions. This could be a single, this could be another method to actually classify the observation.

(Refer Slide Time: 27:09)



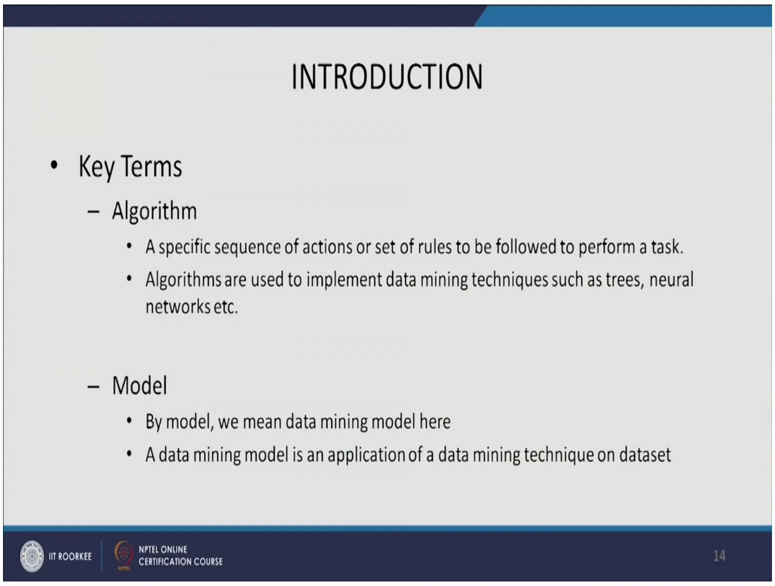
```
     xlim = c(2,12), ylim = c(13,25), pch=c(21,19)[as.numeric(df$Ownership)])
legend("bottomright", inset = 0.005, c("Owner", "Nonowner"),
       pch = c(19,21), cex = 0.7, x.intersp = 0.5, y.intersp = 0.5)

#Method 2: A single diagonal line
df[df$Annual_Income>5 & df$Annual_Income<6 & df$Household_Area>20 &
     df$Household_Area<23, c(1,2)]
df[df$Annual_Income>8 & df$Annual_Income<9 & df$Household_Area>16 &
     df$Household_Area<18, c(1,2)]
segments(5.4,22,8.4,16, col = 3)

slope=atan2(22-16,5.4-8.4)
segments(5.4+10*cos(slope),22+10*sin(slope),
         5.4,22, col = 3)
segments(8.4,16,
         8.4-20*cos(slope),16-20*sin(slope), col = 3)
###############
```

```
5         5.3           21
> df[df$Annual_Income>8 & df$Annual_Income<9 & df$Household_Area>16 &
+      df$Household_Area<18, c(1,2)]
   Annual_Income Household_Area
17           8.5             17
> segments(5.4,22,8.4,16, col = 3)
> slope=atan2(22-16,5.4-8.4)
> segments(5.4+10*cos(slope),22+10*sin(slope),
+          5.4,22, col = 3)
> segments(8.4,16,
+          8.4-20*cos(slope),16-20*sin(slope), col = 3)
> 
```

Data
df 20 obs. of 3 variables
Values
slope 2.03444393579757

Now, as we you know see as we discussed these two methods, method one set of horizontal and vertical lines, and method two a single diagonal line. So, now we need to find out which is the most useful method which is the best method that can be done using different of assessment matrix that we are going to cover in later lectures.

(Refer Slide Time: 27:53)



INTRODUCTION

- Key Terms
  - Algorithm
    - A specific sequence of actions or set of rules to be followed to perform a task.
    - Algorithms are used to implement data mining techniques such as trees, neural networks etc.
  - Model
    - By model, we mean data mining model here
    - A data mining model is an application of a data mining technique on dataset

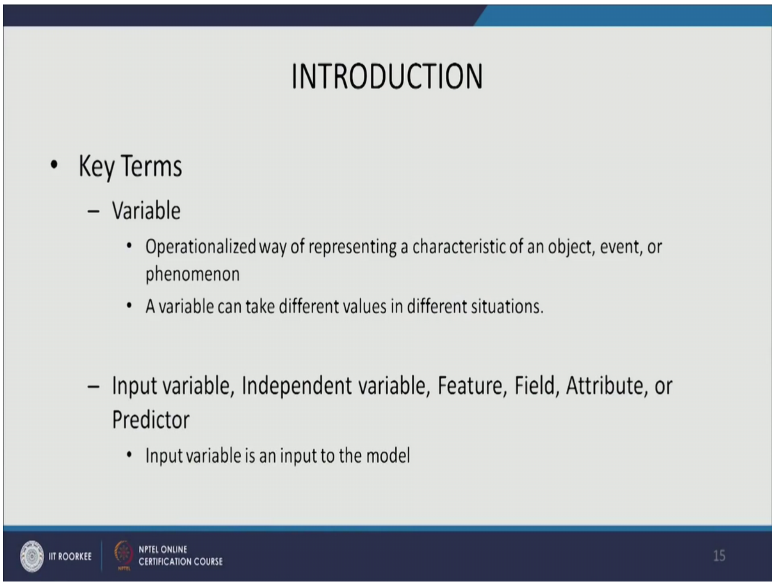
IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
14

Now, let us look at the key terms related to this course. As we go through other lectures we will come across many more terms and then we will discuss them where we need to first we need to discuss the key terms at this point.

So, first one is algorithm. So, we have been using this, this particular term quite often. So, algorithm can be defined as a specific sequence of actions or set of rules that has to be followed to perform task. Algorithms they are gently used to implement data mining techniques like trees, neural networks, for example, neural network we use back propagation algorithm that is part of neural network. So, there are a number of algorithm that are actually required to implement this techniques.

Next term is model. So, what we mean by model? So, here be by modeling we mean data mining model. Now, how we can define a model in a data mining contacts? A data mining model is an application of data mining technique on data set. So, when we apply some of these techniques like trees and neural networks which we are going to cover and later lectures on a data set then we get a model.

(Refer Slide Time: 29:06)



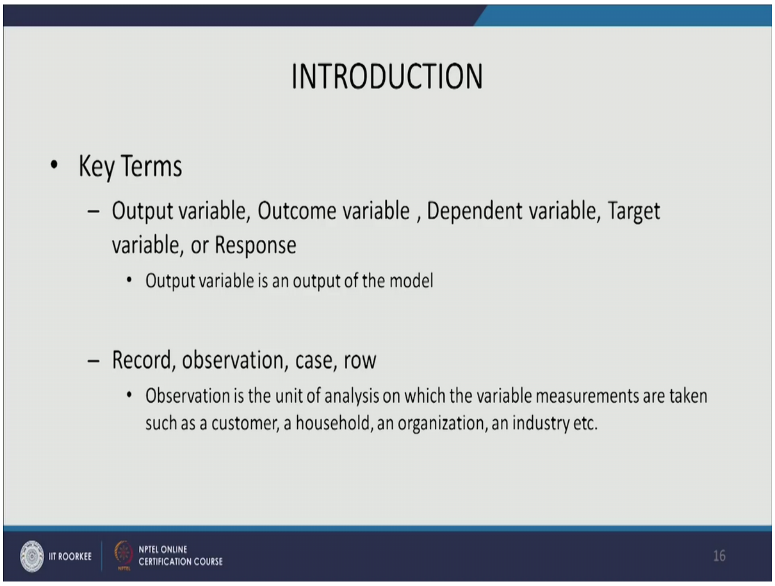
INTRODUCTION

- Key Terms
  - Variable
    - Operationalized way of representing a characteristic of an object, event, or phenomenon
    - A variable can take different values in different situations.
  - Input variable, Independent variable, Feature, Field, Attribute, or Predictor
    - Input variable is an input to the model

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
15

Our next term is, our next term is variable. A variable can be defined as the operationalize way of representing a characteristic of an object event or phenomenon. Now a variable can take different values in different situation. Now, there are generally two types of you know a variable that we going to deal with here is one is input variable it is also sometimes called as independent variable, feature, field, attribute or predictor. Essentially input variable is an input to the model.

(Refer Slide Time: 29:37)



IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
16

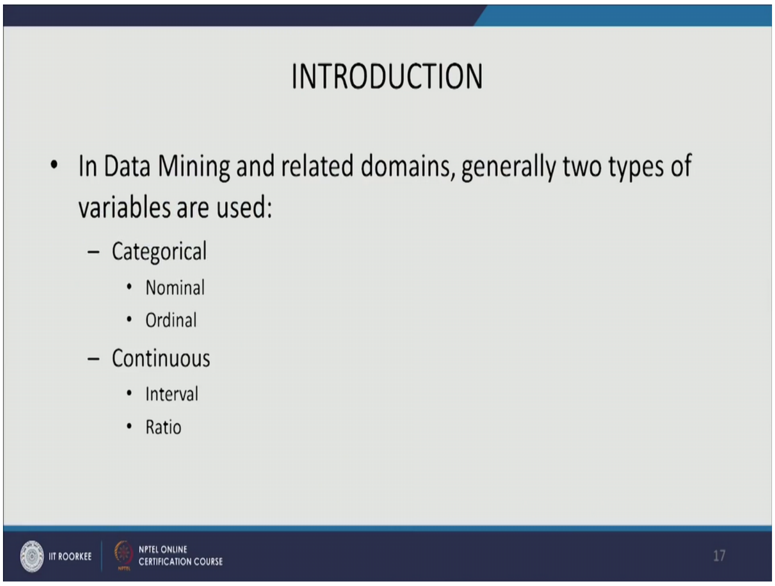
INTRODUCTION

- Key Terms
  - Output variable, Outcome variable , Dependent variable, Target variable, or Response
    - Output variable is an output of the model
  - Record, observation, case, row
    - Observation is the unit of analysis on which the variable measurements are taken such as a customer, a household, an organization, an industry etc.

The other type of variable is output variable which is generally, other names for this variable are outcome variable, dependent variable, target variable or response. So, output variable is an output to the, output of the model.

Another term that we come is across record observation, case or row. So, as we talked about the tabular data set of metric data set each row represented a record each row represented and observation or case. Now, how do we define it? So, observation is the unit of analysis on which the variable measurements that are in the column take can such as a customer, a household, an organization or an industry. For example, in our sedan car example case the unit of analysis was household and we had the variables like annual income and household area which actually major something related to household. Similarly customer organizational or industry you could also have these as a unit of analysis and the variables that are measured on these.

(Refer Slide Time: 30:53)



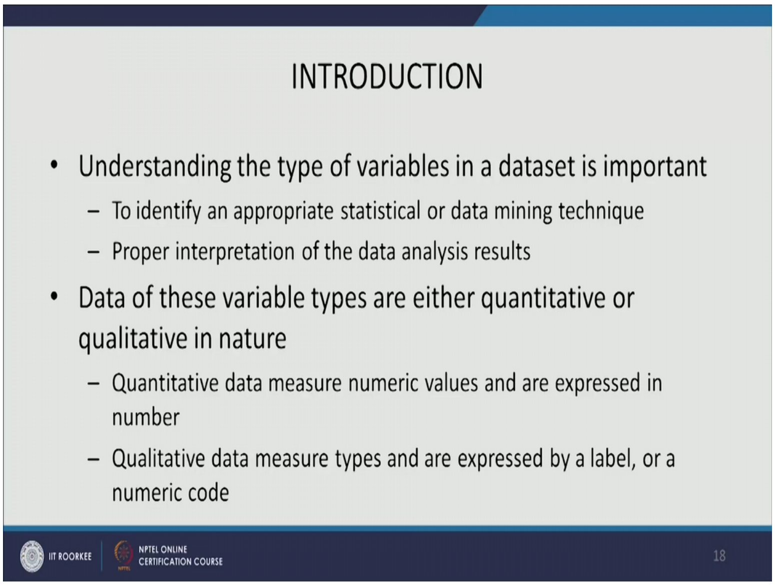
INTRODUCTION

- In Data Mining and related domains, generally two types of variables are used:
  - Categorical
    - Nominal
    - Ordinal
  - Continuous
    - Interval
    - Ratio

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 17

Now, let us the talk about the variables in detail. Not only two types of variables are used. So, other types that we talked about input variable and the output variable or outcome variable they are in the modeling sense. But here we are talking about more in the data sense. So, two types of variables are used categorical and continuous. Now, categorical variables can we further classified as nominal and ordinal, and continuous can further we classified into two categories interval and ratio variables. So, let us understand these variables.

(Refer Slide Time: 31:24)



INTRODUCTION

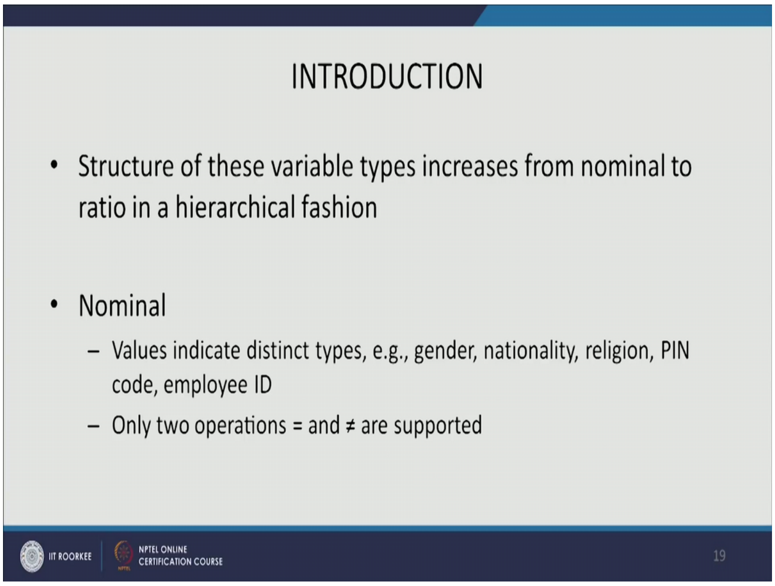
- Understanding the type of variables in a dataset is important
  - To identify an appropriate statistical or data mining technique
  - Proper interpretation of the data analysis results
- Data of these variable types are either quantitative or qualitative in nature
  - Quantitative data measure numeric values and are expressed in number
  - Qualitative data measure types and are expressed by a label, or a numeric code

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 18

Now, before we go into the details of what these 4 types of variable mean why we need to understand the type of variables in a data set. As we discussed before that we are go, in a data mining process in a data mining modeling we generally use a many methods and then the select the most useful one or the best one. So, therefore, it is important for us to identify an appropriate statistical or data mining technique and the understanding of variable type is part of this process.

Now, proper interpretation of the data analysis results, that also depends on the kind of technique that you are using and the kind of data that was actually analyzed. Now, another important thing is that data of these variable types or either quantitative or qualitative nature. For example, quantitative measure quantitative data numeric values and our expressed in a number, qualitative data they measure types and our expressed as label or a numeric code.

(Refer Slide Time: 32:31)



INTRODUCTION

- Structure of these variable types increases from nominal to ratio in a hierarchical fashion
- Nominal
  - Values indicate distinct types, e.g., gender, nationality, religion, PIN code, employee ID
  - Only two operations = and ≠ are supported

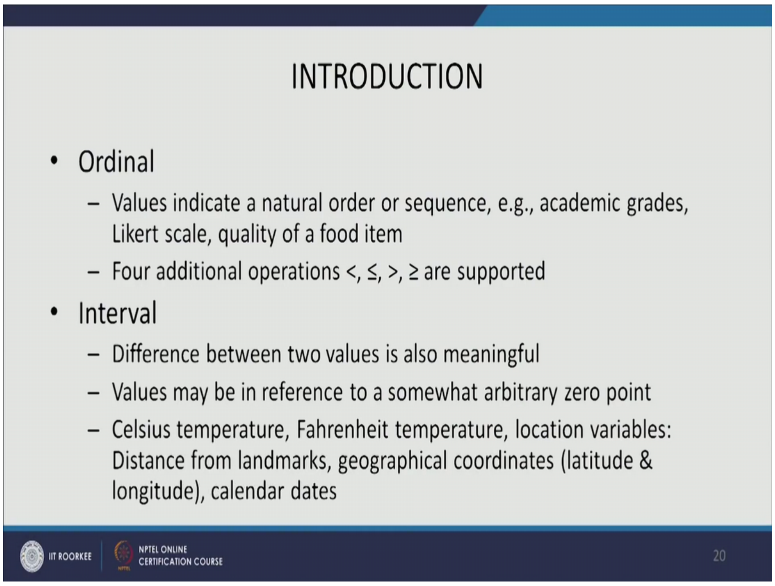
IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
19

Now, if we look at these 4 types nominal ordinal interval and ratio a structure of these variable types in gears from nominal to ratio in a hierarchical fashion. So, nominal is a least structured available type followed by ordinal followed by interval and then ratio is the most a structured variable type. Now, let us understand nominal variables.

Now, nominal values indicate distinct types for example, gender. So, gender values could be male or female. So, they indicate distinct types. Similarly nationality again get the values could be Indian, Pakistani, Bangladeshi, Sri Lankan and so all these values indicate they are distinct type. Similarly the region I could be another nominal variable, so the values will again indicate distinct types for example, Hindu or Muslim or Christian. Similarly, pin code could be another example for a nominal variable where each pin code actually indicate a distance distinct location. Employee ID could be an nominal variable because each employee ID would actually indicate a different person.

Now, the two operations equal to and is not equal to or supported because these are distinct types you cannot say male is greater than female or so, greater multiplication greater than less than multiplication division all these observations are not supported only equal to R is not equal to is supported.

(Refer Slide Time: 34:06)



INTRODUCTION

- Ordinal
  - Values indicate a natural order or sequence, e.g., academic grades, Likert scale, quality of a food item
  - Four additional operations <, ≤, >, ≥ are supported
- Interval
  - Difference between two values is also meaningful
  - Values may be in reference to a somewhat arbitrary zero point
  - Celsius temperature, Fahrenheit temperature, location variables: Distance from landmarks, geographical coordinates (latitude & longitude), calendar dates

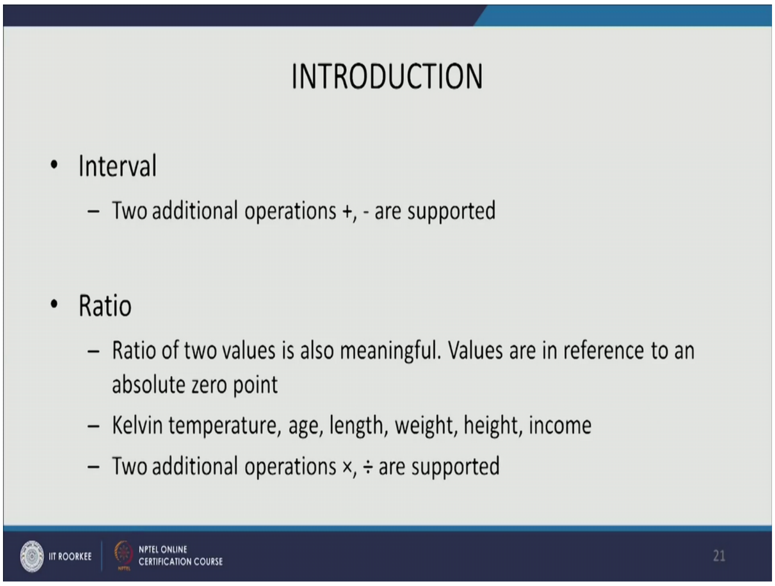
Let us look at the ordnance. So, values indicate a natural order of sequence. Natural order or sequence for example, academic grades. So, you have A grades like A B C D E F. So, they all these grades label they indicate you know A order, but essentially they are distinct from each other.

Similarly, likert scales, likert scale quality of a food item they are also examples of ordinal variables. In likert scale you come across whenever you are trying to fill a survey questionnaire, trying to reach to the respondents and trying to get the responses you always get the responses from a strongly disagree to a strongly agree or strongly agree to strongly disagree. So, all those points are actually part of likert scale and can be are actually can be defined as ordinal variable.

Similarly quality of a food item it could be better, it could be average, it could be you know poor so that can also be an ordinary variable. Now, 4 additional operations are supported because the values indicate a natural order. So, therefore, these less than or less than or equal to greater than or greater than or equal to operations are also supported.

If we look at the continuous variable types first one interval variable, now here apart from the apart from what we discussed a nominal ordinal difference between two values is also meaningful. Now, another important thing about interval variable is values may be in reference to a somewhat arbitrary zero point. So, for example, Celsius temperature, Fahrenheit temperature. So, Celsius temperature is generally used in India where we you know when we talk about the temperature is 35 degree Celsius 40 degree Celsius and those numbers. Now, when we talk about 0 degree Celsius it is not actually the absolute 0 point it is rather a can we called an arbitrary zero point. Similarly location specific wherever for example, distance from landmarks, a geographical coordinates they all are examples of interval variables.

(Refer Slide Time: 36:24)



INTRODUCTION

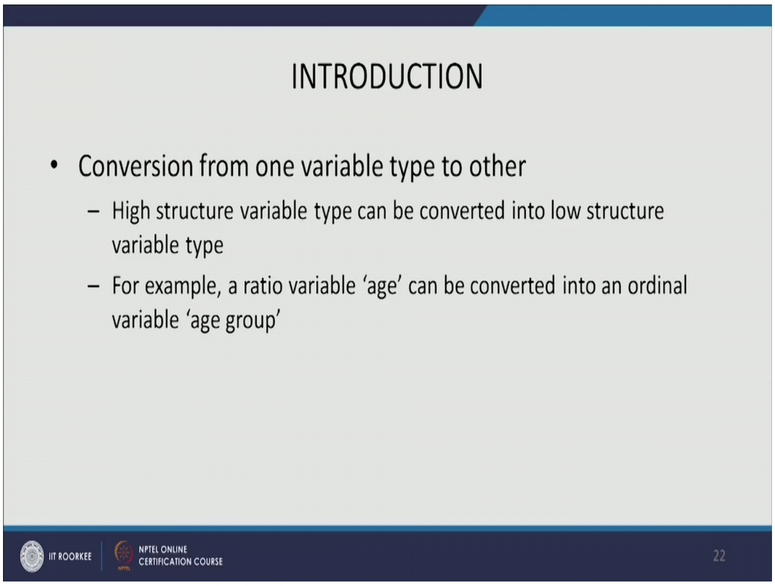
- Interval
  - Two additional operations +, - are supported
- Ratio
  - Ratio of two values is also meaningful. Values are in reference to an absolute zero point
  - Kelvin temperature, age, length, weight, height, income
  - Two additional operations ×, ÷ are supported

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
21

Now, next, the operations that are supported for interval variables are two additional operations apart from what we discussed for nominal and ordinal to additional operations are supported plus and minus addition and substraction. Now, the last variable type is ratio variable. Now, in ratio variables ration of two values is also meaning full. So, values and also values are in reference to an absolute 0 point example Kelvin temperature. So, in Kelvin temperature you have 0 degree Kelvin. So, when we say 0 degree Kelvin we are actually means is actually is actually mean 0 degree, 0 temperature, that is 0 in absolute in real sense. When we talked about the Celsius temperature or Fahrenheit temperature 0 degree centigrade or 0 degree Fahrenheit they do not actually mean a absolute 0 all right, real 0 in the absolute sense.

A is length of weight, height, income and we can also the examples of (Refer Time: 37:26) variables; apart from the operation that we discussed for nominal ordinal interval two additional operations division and multiplication or supported for these two variables.

(Refer Slide Time: 37:37)



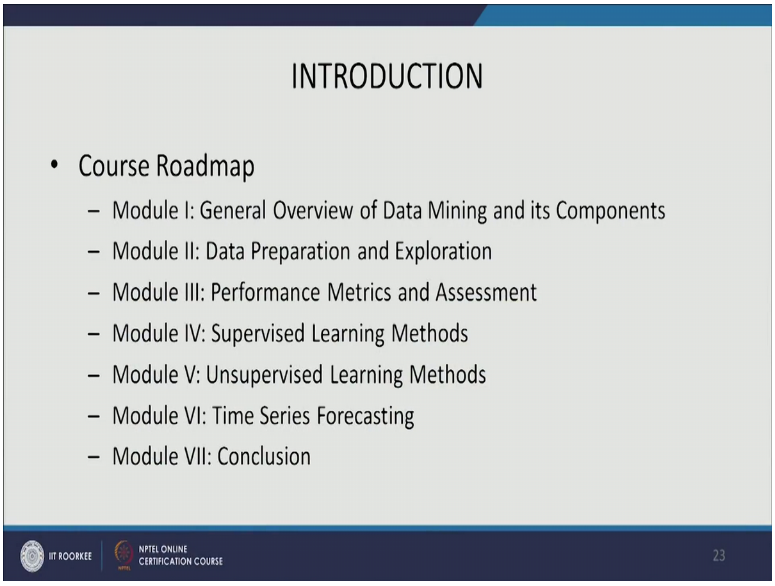
INTRODUCTION

- Conversion from one variable type to other
  - High structure variable type can be converted into low structure variable type
  - For example, a ratio variable 'age' can be converted into an ordinal variable 'age group'

Now, another important thing we lead to variable type is convergent from one variable type to other. Now, high structure variable type can always be converted because we discussed that these variable types they have you know hierarchy of a structure. So, therefore, you know high structure variable can always be converted into a low structure variable type, but we cannot convert a lower structure variable into a high structure variable. For example, a ratio variable is can we converted into an ordinal variable age group.

So, age group could be like you know you know adult, young, old, middle age, but age could also be you know a specific age could be 20, 21, 25, 40 and 45. So, this is actually being a ratio variable. Now, based on the these actual numbers you can actually convert this variable into a ordinal variable where you say that less than 20 is young then 20 to 40 its adult then later on middle age and then old age. So, that kind of grouping can actually be done. So, therefore, a high structure variable cannot always be converted into a lowest structure variable type.

(Refer Slide Time: 38:52)



INTRODUCTION

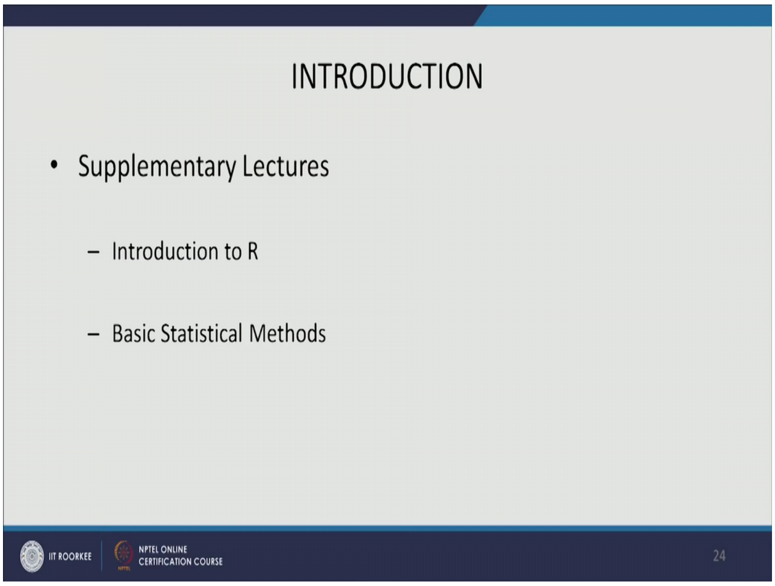
- Course Roadmap
  - Module I: General Overview of Data Mining and its Components
  - Module II: Data Preparation and Exploration
  - Module III: Performance Metrics and Assessment
  - Module IV: Supervised Learning Methods
  - Module V: Unsupervised Learning Methods
  - Module VI: Time Series Forecasting
  - Module VII: Conclusion

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
23

Now, let us discuss the road map of this particular course. So, module first which we started in this lecture is a general overview of data mining and its components which covers the introductory part and the data mining process. Then the module second it is about data preparation and exploration here we talk about the different you know a steps that are required to prepare the data, to explore the data, which lies the data and other techniques like dimension reduction etcetera.

Now, third module is about performance matrix and assessment where will try to understand the matrix that are actually used for task like do assess the performance of different models for task like classification and prediction. Now, the next module, fourth module is about supervised learning methods there will we are going to cover data mining and statistical techniques like regression and logistic regression, neural networks, trees. So, those techniques and many others are going to be covered in this particular module. Module fifth, is about unsupervised learning methods they are we are going to mainly cover clustering and association rules mining. Then next module is a time series forecasting there we are going to cover the time series handling regression based forecasting and a smoothing methods. Finally, in the last module we will have some final discussion and concluding remarks.

(Refer Slide Time: 40:29)



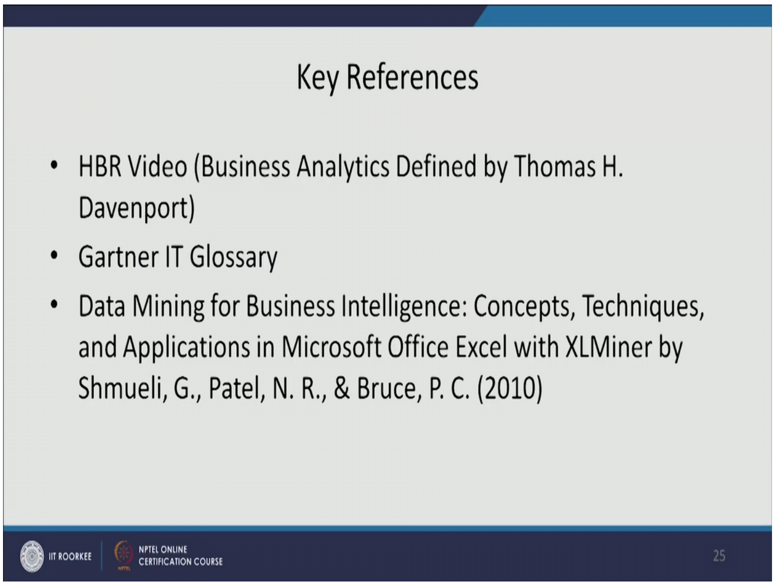
INTRODUCTION

- Supplementary Lectures
  - Introduction to R
  - Basic Statistical Methods

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 24

Now, apart from this road map will also have two supplementary lectures, one is on introduction to R, then other one on basic statistical methods. So, it is highly recommended that you go through these two lectures before proceeding further for next lecture.

(Refer Slide Time: 40:47)



Key References

- HBR Video (Business Analytics Defined by Thomas H. Davenport)
- Gartner IT Glossary
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 25

These are some of the key references.

Thank you.