Leadership

ANALYTICS PROGRAMMING - R

Naveen Kumar

Week 3

Agenda

- Week 2 Summary and Business Apps Presentation
- Getting Familiar with Datasets
- Review Assignment 1
- Quiz 1
- Project Q&A
- Types of Analytics
- Descriptive Analytics (with Business Apps)
- Cran R and Descriptive Analytics
- Summary and Conclusions

Getting Familiar With the Dataset

Use the dataset “BillingInfo” to perform data pre-processing in the upcoming slides.

#Reading the dataset

read.csv(): Reads the file in .csv format and creates a data frame with lines (rows) and variables (columns) same as the .csv file.

If the dataset has a header, enter header = TRUE in read.csv()

Since it is a .csv file, separator is “,”

```
med.data <- read.csv("BillingInfo.csv", header = TRUE, sep = ",")
```

Getting Familiar With the Dataset

dim(): Determines the number of dimensions (rows and columns) in the data.

summary(): Summary function returns the basic descriptive statistics of the vector, matrix or dataframe (like Minimum, 1st Quartile, Mean, Median, 3rd Quartile and Maximum)

dim(med.data)

class(med.data)

summary(med.data)

```
> dim(med.data)
[1] 20  9
> class(med.data)
[1] "data.frame"
> summary(med.data)
       ID          AgeInYears    Gender    Opinion    ChargesInDollars VisitTimeInMin
 Min.   : 1.00   Min.   :21.00   F:11   Min.   :1.0   Min.   : 24.00   Min.   : 31.00
 1st Qu.: 5.75   1st Qu.:32.75   M: 9   1st Qu.:1.0   1st Qu.: 45.75   1st Qu.: 55.50
 Median :10.50   Median :52.00          Median :2.0   Median : 58.50   Median : 66.50
 Mean   :10.50   Mean   :46.30          Mean   :2.6   Mean   : 65.90   Mean   : 72.95
 3rd Qu.:15.25   3rd Qu.:60.25          3rd Qu.:4.0   3rd Qu.: 92.75   3rd Qu.:102.50
 Max.   :20.00   Max.   :65.00          Max.   :5.0   Max.   :114.00   Max.   :120.00

   Insurance  PriorVisits         Date
 BCBS    :5   Min.   : 31.00   1/1/14 : 1
 Medicaid:4   1st Qu.: 55.50   1/15/14: 1
 Private :7   Median : 66.50   1/25/14: 1
 Self Pay:4   Mean   : 72.95   1/5/14 : 1
              3rd Qu.:102.50   2/13/14: 1
              Max.   :120.00   2/14/14: 1
                               (Other):14
```

Getting Familiar With the Dataset

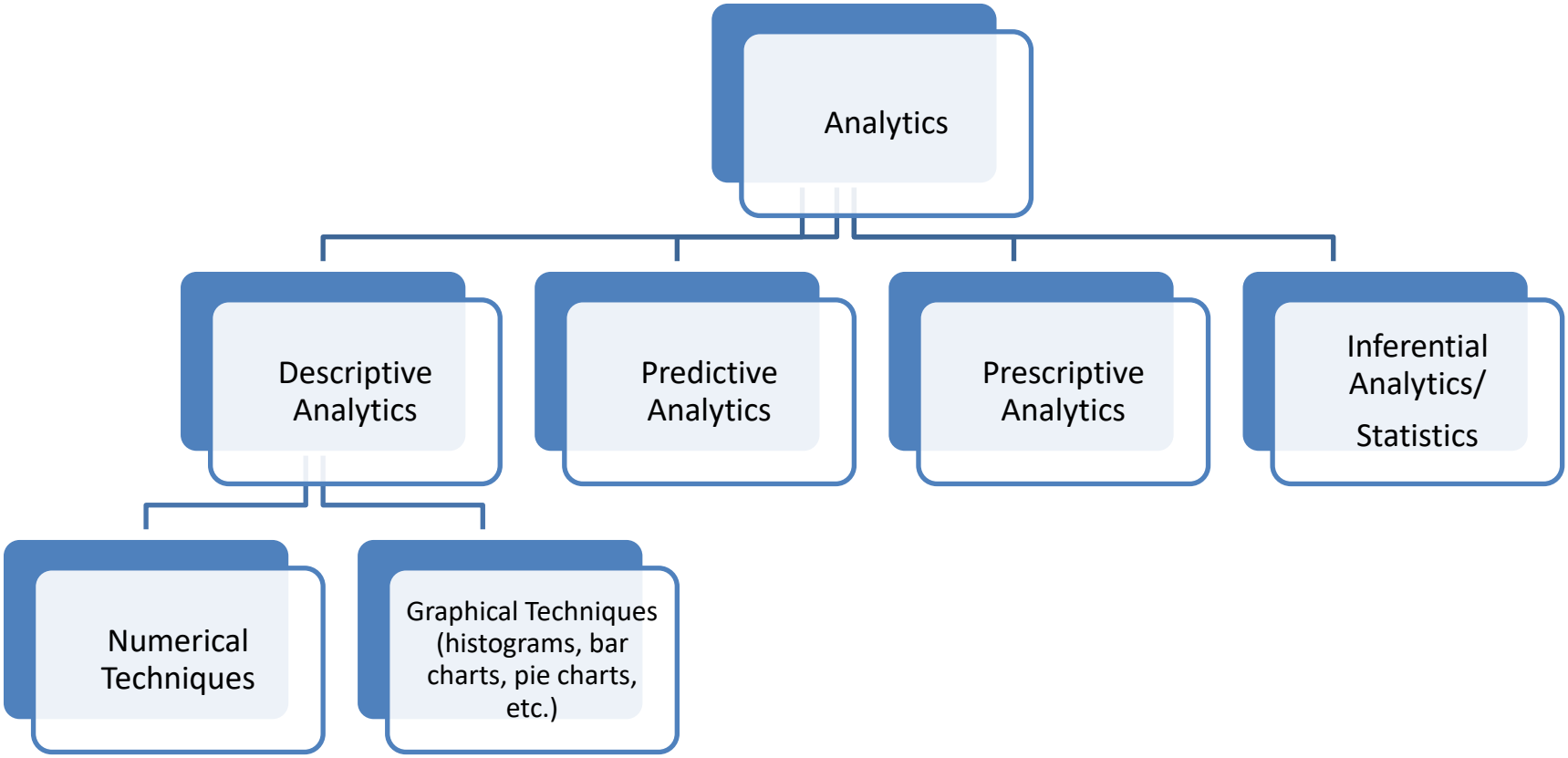
#Unique Insurance Type

unique() – Returns the unique values in a vector

```
> employee.gender <- c("male", "female", "female", "male", "female")
> unique(employee.gender)
[1] "male" "female“
```

To select specific column of a dataset, notation will be “**dataset$columnname”**

Types of Analytics



Descriptive Analytics

- Arranging, summarizing, and presenting data in a convenient and informative way

- Simple, yet powerful techniques for uncovering patterns that offer insights

- Examples:
  - Determining the sales performance in different regions of a company
  - Categorizing customers by their likely product preferences
  - Identifying the performance of the call center teams in different locations etc.
  - Monitoring Covid-19 trends

Descriptive Analytics

- **Numerical Techniques**
  - To summarize data
- **Graphical Techniques**
  - Presents data in ways that make it easy to extract useful information

Predictive Analytics

- Predict future event(s) of interest based on current or historical dataset.

- Examples:
  - Weather prediction
  - Sales prediction
  - COVID-19 Fatalities prediction

- Key Techniques:
  - Supervised Machine Learning such as Logistic Regression

Prescriptive Analytics

- Maximize/minimize specific characteristics of interest under some constraints
- Example:
  - Maximize revenue by determining right product mix
- Really valuable, but largely underused
- According to Gartner, 13 percent of organizations are using predictive but only 3 percent are using prescriptive analytics
- Techniques:
  - Linear Programing, Integer Programing

Inferential Analytics

Making an estimate, prediction, or decision about a population based on a sample

Population

- The group of all items of interest
- Frequently very large; sometime infinite
- Example: all NY voters

Sample

- Subset of data taken from the population
- Potentially very large, but less than the population
- Example: A subset of NY voters polled on election day

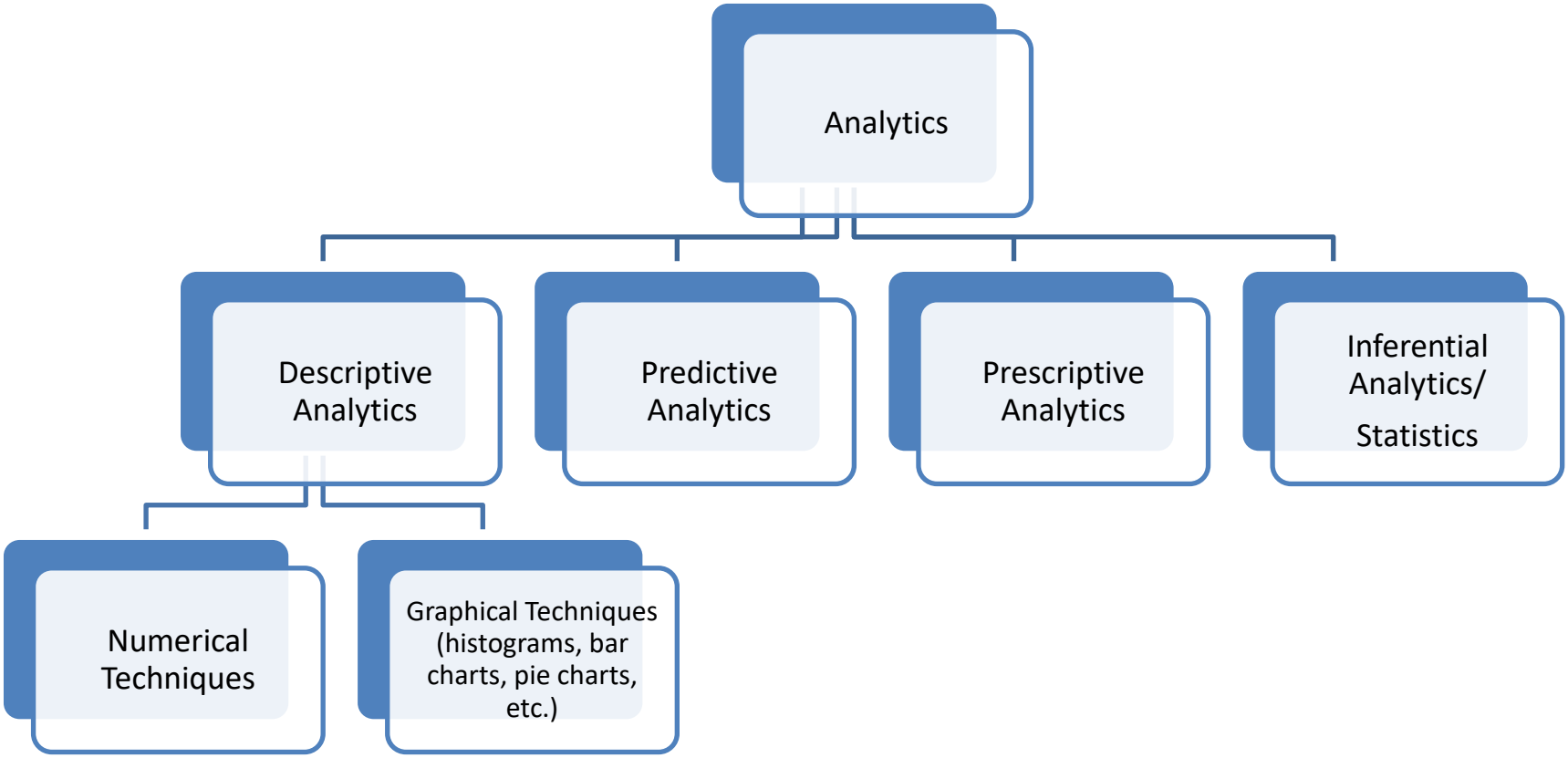
Techniques

- Hypothesis Testing

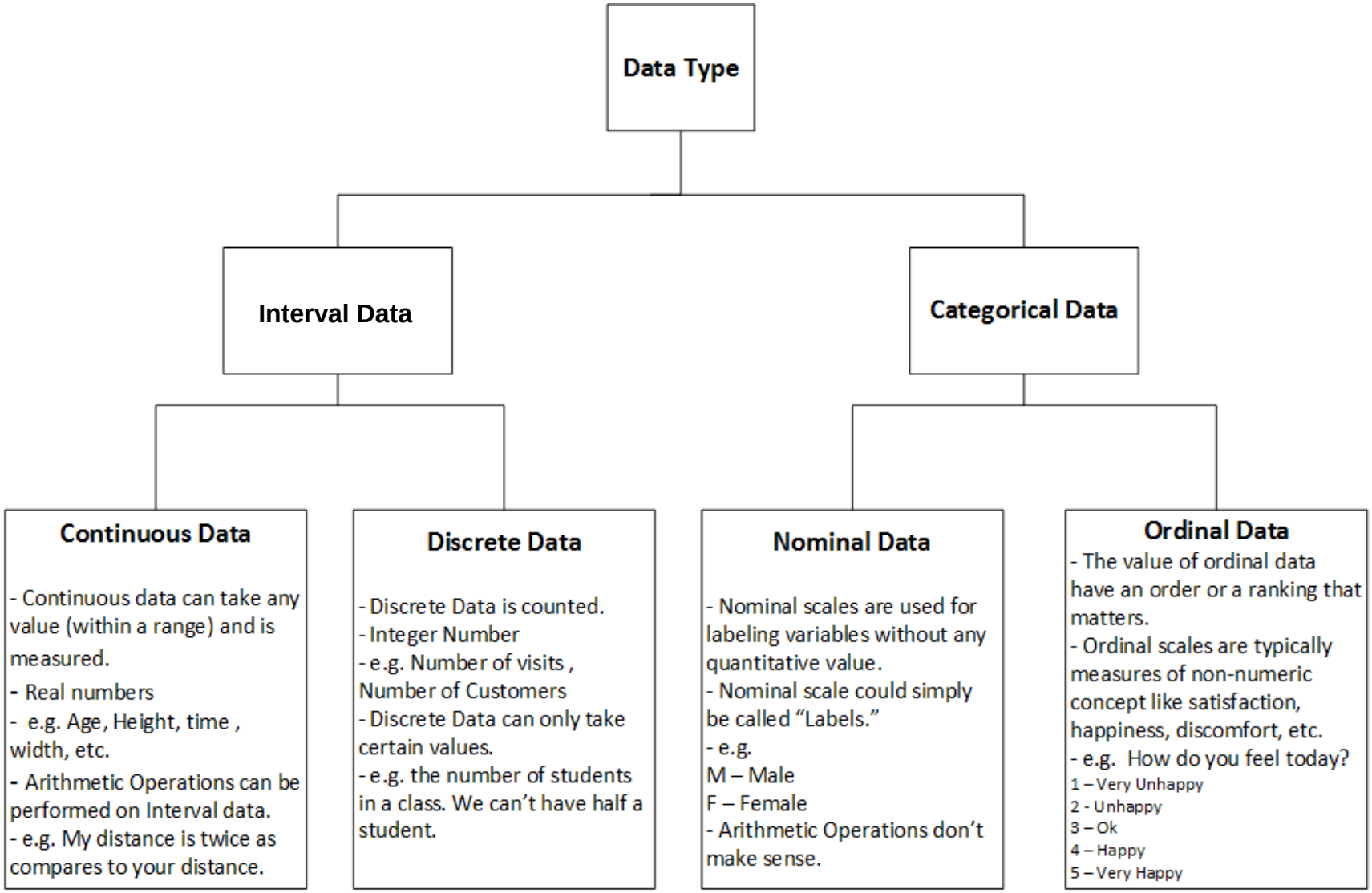
Cran R

- Programing language
- Produces high quality analysis and graphics
- Open-source
- Free
- Runs on any operating system
- User types commands directly into user “console”

Analytics



Data Type



Interval Data

Continuous Data

- Real numbers:
  - Examples: age, height, width, average weight time
- Arithmetic can be performed on continuous data

Discrete Data

- Integer numbers:
  - Examples: number of visits, number of customers
- Arithmetic can be performed on discrete data

Categorical Data

Nominal Data

- Values are not quantitative and do not have any rank order
  - Example: Marital status coded as Single = 1, Married = 2, Divorced = 3
- Cannot be used for arithmetic
  - (e.g. does Single * 2 = Married?!)

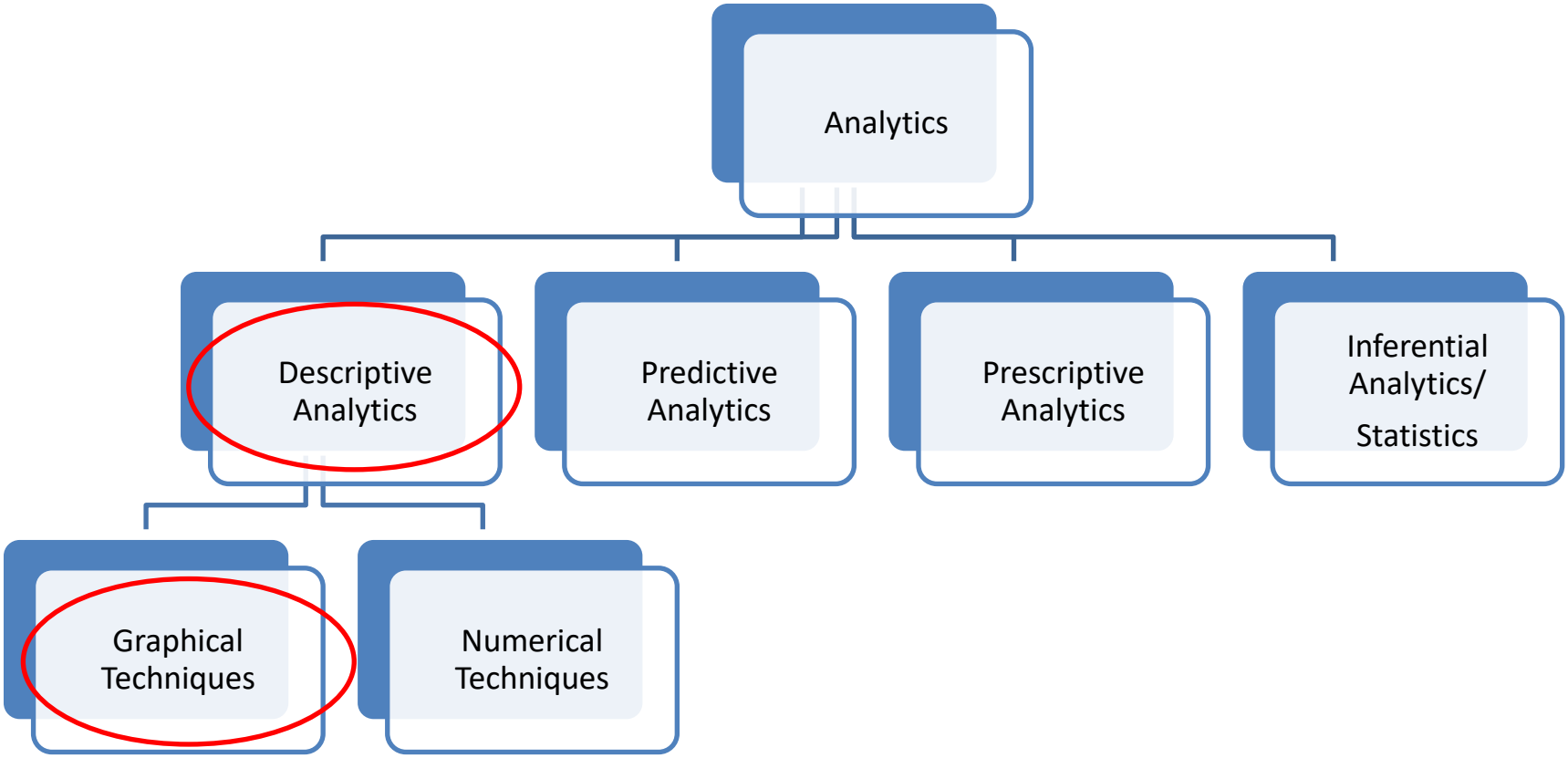
Ordinal Data

- Values have a rank order
  - Example: College course rating system, poor = 1, fair = 2, good = 3, very good = 4, excellent = 5
- While it's still not meaningful to do arithmetic on this data (e.g. does 2*fair = very good?), we can say: excellent > poor or fair < very good.
- Order is maintained no matter what numeric values are assigned to each category

Key Points

- All calculations are permitted on interval data
- Only a ranking process is allowed for ordinal data
- Typically, no calculations are performed on nominal data, except for counting the number of observations in each category

Analytics: Graphical Techniques

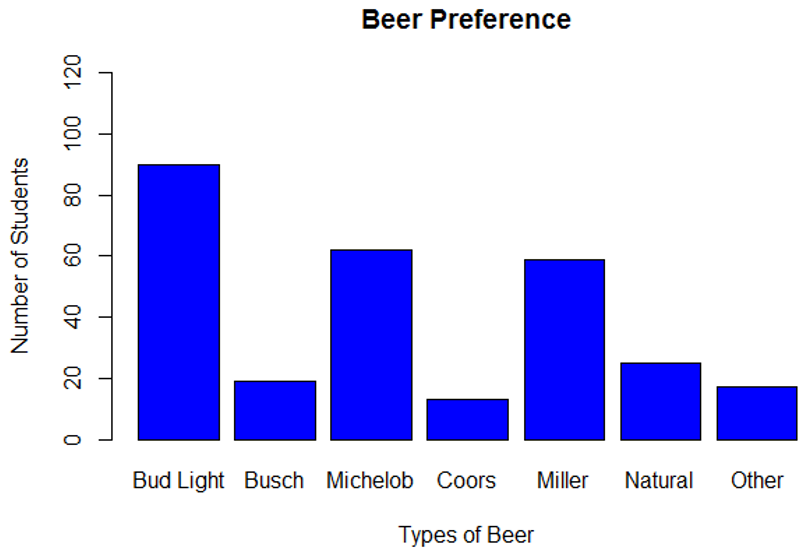


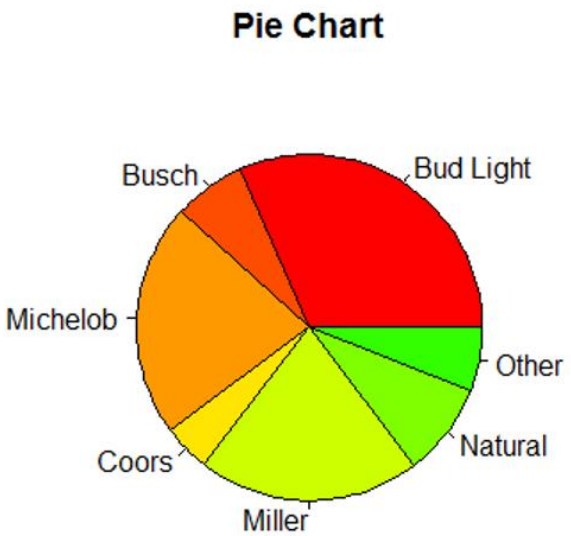
Graphical Techniques

- Presents data in ways that make it easy to extract useful information
- Objectives
  - Understand and use appropriate graphical methods suitable for a given set of data
  - Transform raw data into information through graphical display using prominent graphical methods
  - Describe the relationship between two variables

Categorical Data: Nominal

- Typically, the permissible calculation on nominal data is to count the frequency of each value or variable
- Can be visualized with bar graph or pie chart



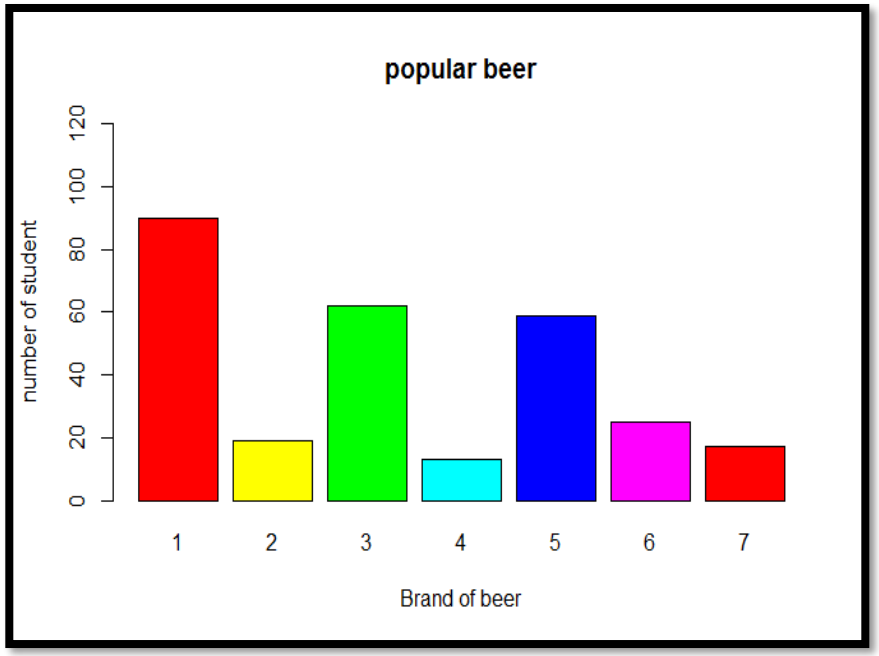


Bar Graph

- **Syntax:**

barplot(table(beer.data$Brand), ylim = c(0,120), xlab = "Brand of beer", ylab = "number of student", main = "popular beer", col = rainbow(6))

- xlab , ylab : used to put labels on X and Y axis respectively
- ylim : define the values on Y-axis
- main : is used to put the main heading (label)
- col : used to define the color of the bar



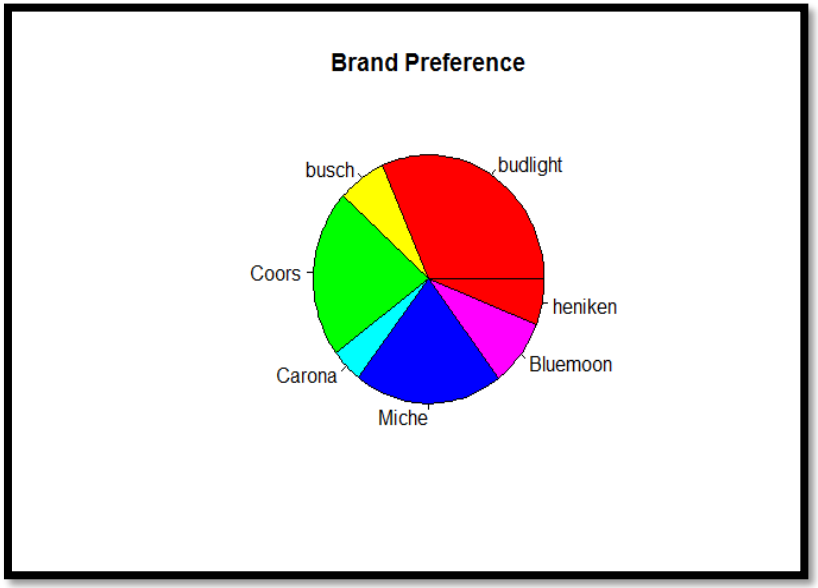
Pie Chart

- **Syntax:**

piechart.labels <-
c("budlight","busch","Coors","Carona","Miche","Bluemoon","heniken")

piechart.labels : used to assign labels

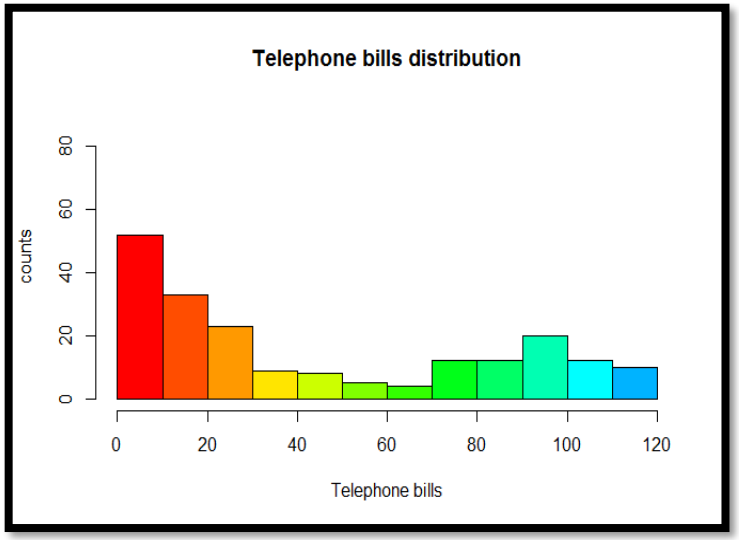
pie(table(beer.data$Brand),main = "Brand Preference" , col = rainbow(6), labels = piechart.labels)



Graphical Measures: Interval Data

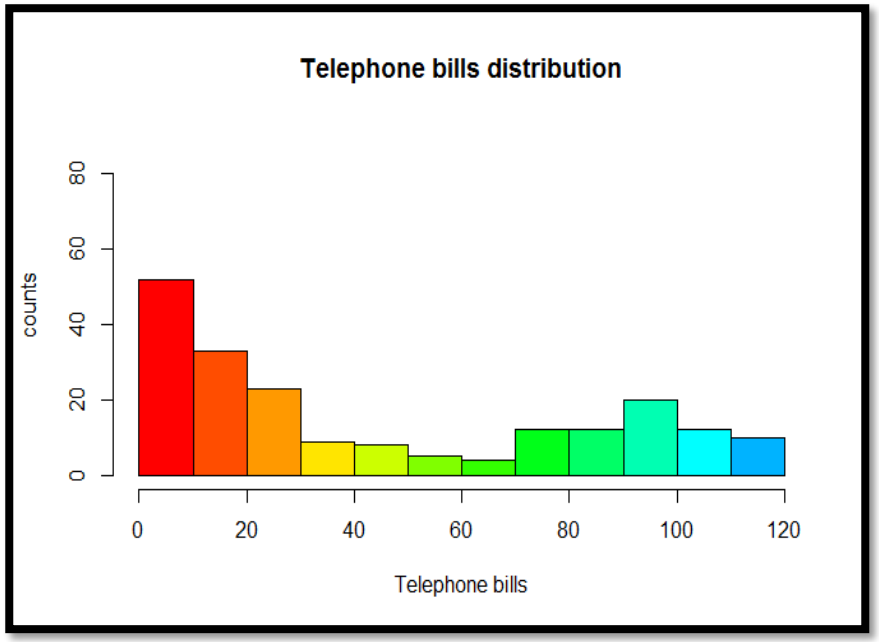
Histogram

- To graphically describe interval data
- Construct a frequency distribution from which a histogram can be drawn
- Create sets of intervals (called bins) that cover the complete range of observations and count the number of observations that fall into each set



Histogram

- Syntax:
  - **hist**(bills.data$Bills, ylim = c(0,90), xlab = "Telephone bills", ylab = "counts", main = "Telephone bills distribution", col = rainbow(20))
    - xlab : used to put labels on X respectively
    - ylim : define the values on Y-axis
    - main : is used to put the main heading (label)
    - col : use to define the color of the bar



Histogram Shapes

- ***Symmetric***: draw a vertical line down the center of the histogram and the sides should be identical in shape and size

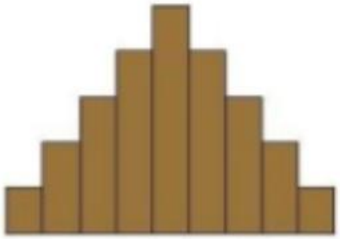
- ***Skewed***: a long tail extending either to the right or left

- ***Modality***: A unimodal histogram has a *single peak*, while a bimodal histogram has *two peaks*

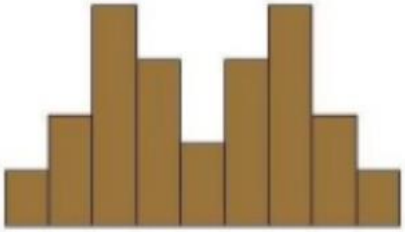
- ***Bell shaped***: a special type of symmetric unimodal histogram is one that is bell shaped
  - Many analytical techniques require that the population be bell shaped
  - Drawing the histogram helps to verify the shape of the distribution of a variable in a population

Histogram Shapes

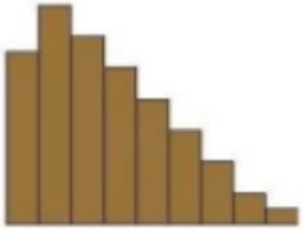
Bell-shaped: A bell-shaped usually presents a normal distribution (symmetric unimodal)



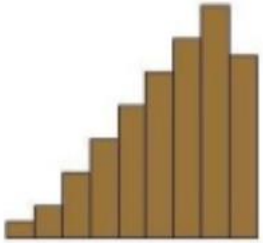
Bimodal: A bimodal shape has two peaks. This shape may show that the data has come from two different systems. If this shape occurs, the two sources should be separated and analyzed separately.



Skewed Right: A distribution skewed to the right is said to be positively skewed.



Skewed Left: A distribution skewed to the left is said to be negatively skewed.



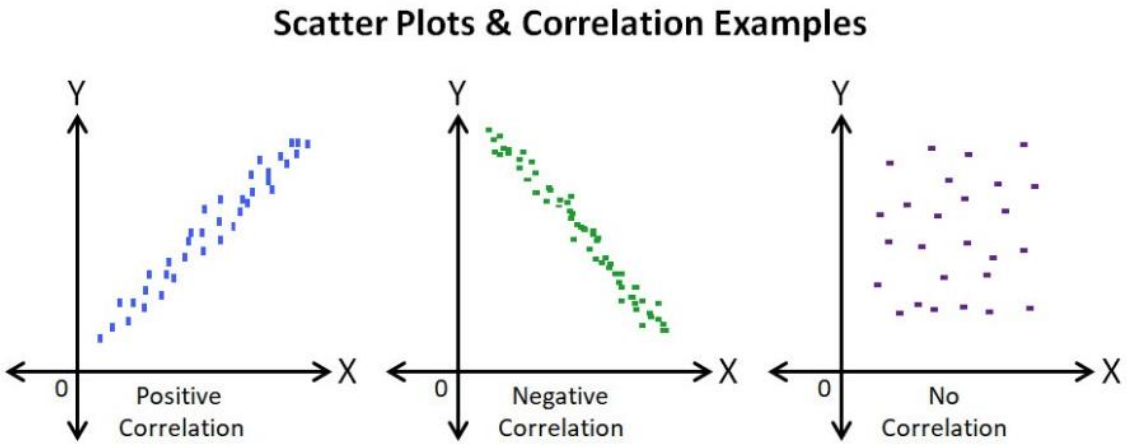
Bar Chart vs Histogram

- Difference between a Bar Chart and Histogram?
  - Histograms do not have gaps between adjacent columns because columns represent continuous, quantitative data

Relationship Between Two Variables

Scatter Diagram

- Plot two continuous variables against one another
  - Independent variable is labeled X: plotted on the horizontal axis
  - Dependent variable is labeled Y: plotted on the vertical axis
- We are interested in the linearity and direction of the scatter



Perfect positive linear relationship

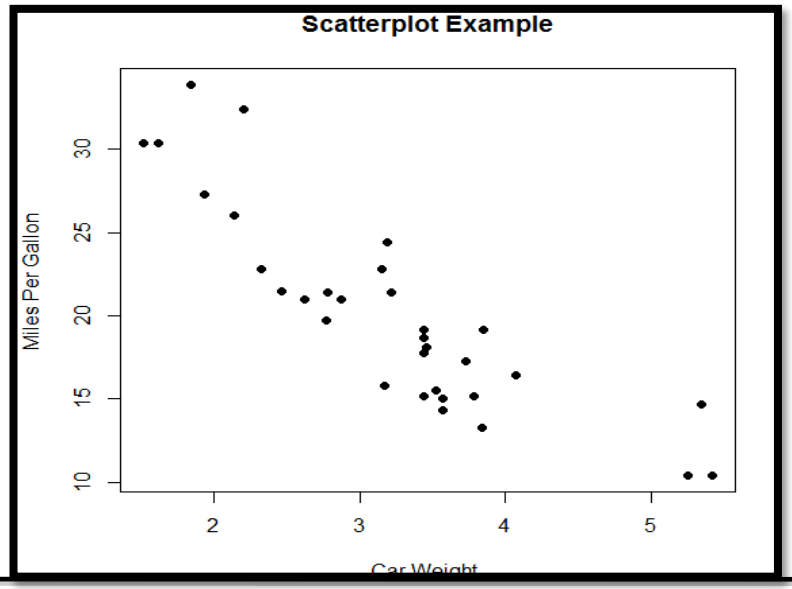
Perfect negative linear relationship

No linear relationship

Scatter plot

- **Syntax:**
  - data(“mtcars”)
  - attach(mtcars)
    plot(mtcars$wt, mtcars$mpg, main="Scatterplot Example",
    xlab="Car Weight ", ylab="Miles Per Gallon")

  - main : is used to put the main heading (label)
  - xlab ,Y lab : used to put labels on X and Y axis respectively
  - pch : used to define the type of points on scatter plot



Summary and Questions