**DESCRIPTIVE STATISTICS  
DHRUVIN KOTHARI**

**ARVIND PAWAR**

**NORTHEASTERN UNIVERSITY**

**ALY6015 INTERMEDIATE ANALYTICS**

**INSTRUCTOR NAME- TENGLONG LI**

**DATE- 02/27/2018**



**PART 1:**

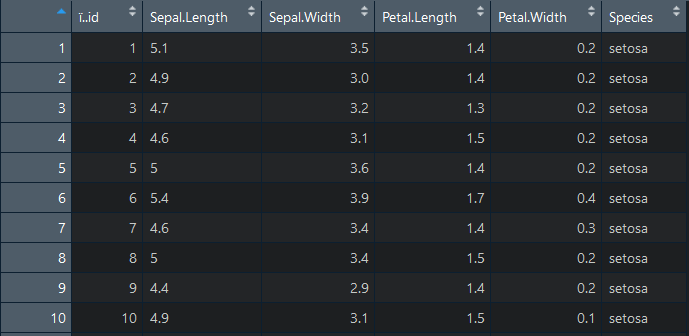
1. **Download the attached dataset “assignment1.csv, set your working directly to the location of the file, and import the dataset using R code.**

**CODE 1:**

> setwd("C:/Users/Arvind/Desktop/Intermediate Analytics/Week 1")

> df <- read.csv('assignment1.csv')

> View(df)

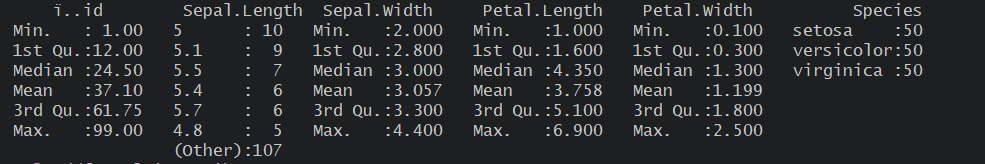
****

**Using descriptive statistics, visualize this data numerically and graphically.**

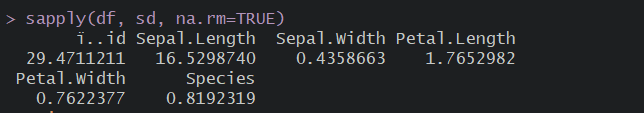
**Solution:**

**Descriptive Statistics numerically:**

> summary(df)

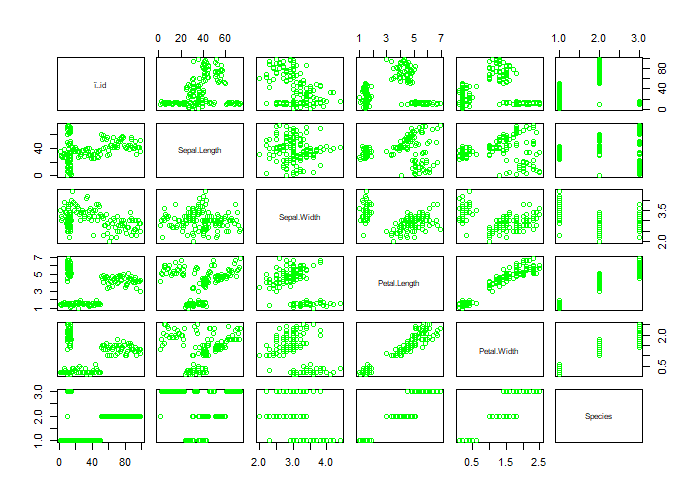
****

> sapply(df, sd, na.rm=TRUE)

****

**Descriptive Statistics Graphically:**

> plot(df, col='green')

**Figure:1 Graphical Representation of Descriptive Statistics**

**In particular, please provide summary statistics (mean, median, standard deviation, min, max, Q1, and Q3) for variable “Sepal.Width”.**

**Solution:**

>summary(df$Sepal.Width)

**Min. 1st Qu. Median Mean 3rd Qu. Max.**

**2.000 2.800 3.000 3.057 3.300 4.400**

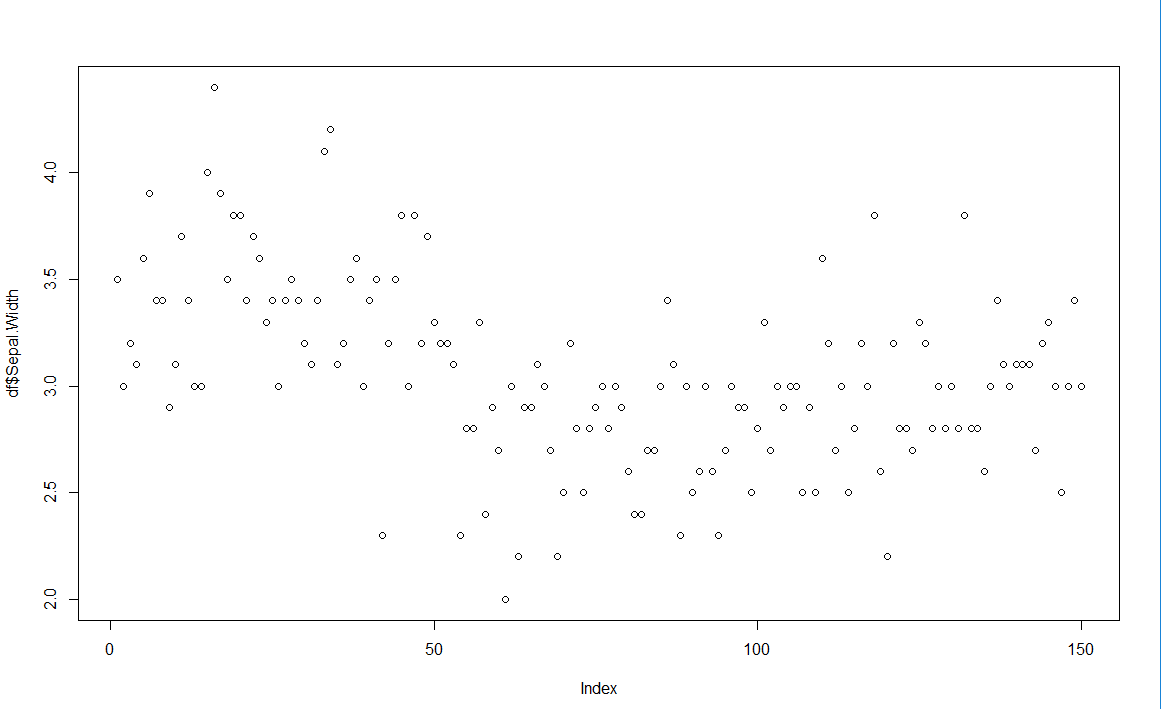
> #standard Deviation

> std <- sd(df$Sepal.Width)

**> std**

[1] 0.4358663

> plot(df$Sepal.Width)

Figure:2 Graphical Representation of Summary statistics for Sepal.Width

**Also, create a histogram for “Sepal.Width”. You can use the following resource as a guide: section 2.5 (p. 15) and 3.5 (p.25) from “Using R for Data Analysis and Graphics”.**

**Solution:**

> hist(df$Sepal.Width, main = "Histogram for Sepal.Width column", xlab = "Sepal.Width")

> hist(df$Sepal.Width, main = "Histogram for Sepal.Width", xlab = "Sepal.Width")

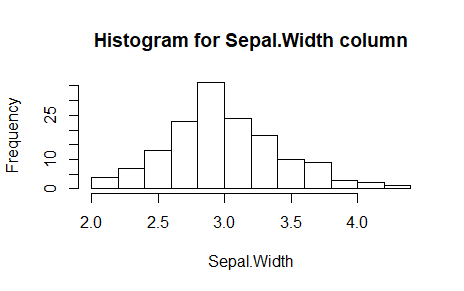


Figure:3 Histogram of Sepal.Width.

**PART 2:**

Using  [height\_weight\_byGender.csv](https://northeastern.blackboard.com/bbcswebdav/pid-18179710-dt-content-rid-41422766_1/xid-41422766_1) data file, plot a histogram of men’s and women’s height:

**CODE 2:**

> #read CSV file

> Height\_Weight\_by\_Gender <- read.csv('height\_weight\_by\_gender.csv')

> #view data from CSV file

> View(Height\_Weight\_by\_Gender)

> #view the column names

> names(Height\_Weight\_by\_Gender)

[1] "ï..Gender"       "Height..inches." "Weight..lbs."

**Column names were not proper, so renamed the columns as follows.**

> colnames(Height\_Weight\_by\_Gender)[colnames(Height\_Weight\_by\_Gender)=="ï..Gender"] <- "Gender"

> colnames(Height\_Weight\_by\_Gender)[colnames(Height\_Weight\_by\_Gender)=="Height..inches."] <- "Height(Inches)"

> colnames(Height\_Weight\_by\_Gender)[colnames(Height\_Weight\_by\_Gender)=="Weight..lbs."] <- "Weight(lbs)"

**> #Check the renamed columns**

> names(Height\_Weight\_by\_Gender)

[1] "Gender"         "Height(Inches)" "Weight(lbs)"

**We observed that in Weight column some values are -1, -2 as well as values less than 40, which is impossible and may mislead the graphical representation of the data. So, we decided to clean the data. We replaced those values with median values as follows.**

>Height\_Weight\_by\_Gender$`Weight(lbs)`[Height\_Weight\_by\_Gender$`Weight(lbs)`<0]= median(Height\_Weight\_by\_Gender$`Weight(lbs)`)

>Height\_Weight\_by\_Gender$`Weight(lbs)`[Height\_Weight\_by\_Gender$`Weight(lbs)`<50]=median(Height\_Weight\_by\_Gender$`Weight(lbs)`)

**Then, we Installed and loaded the ggplot2 and colorspace libraries that are useful in visualization of data.**

> #load the packages

> library(ggplot2)

> library(colorspace)

> #**Histogram for men's and women's height**

> ggplot(Height\_Weight\_by\_Gender, aes(x= Height\_Weight\_by\_Gender$`Height(Inches)`))+

+   geom\_histogram(aes(shape=factor(Gender), fill=Gender))+

+   ggtitle("Histogram of men's and women's height")+

+   xlab("Height in Inches")+

+   ylab("Count")

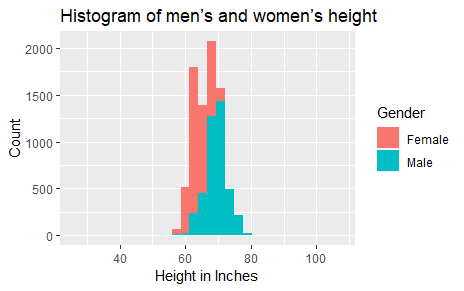


Figure:4 Histogram of Height for Men and Women.

**> #Histogram for men's and women's weight**

> ggplot(Height\_Weight\_by\_Gender, aes(x= Height\_Weight\_by\_Gender$`Weight(lbs)`))+

+   geom\_histogram(aes(shape=factor(Gender), fill=Gender))+

+   ggtitle("Histogram of men's and women's weight")+

+   xlab("Weight in lbs")+ +   ylab("Count")

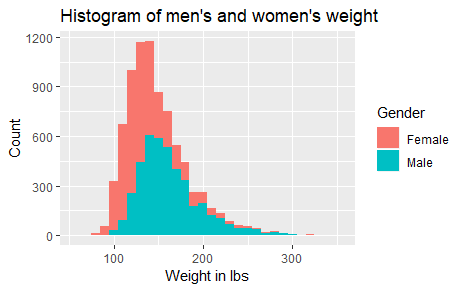


Figure:5 Histogram of Weight for Men and Women.

**Scatterplot comparing the weight and height of men and women.**

**Solution:**

> ggplot(Height\_Weight\_by\_Gender, aes(x=`Weight(lbs)`, y=`Height(Inches)`))+

+ geom\_point(aes(shape=factor(Gender), color=factor(Gender)))+

+ labs(title="Scatterplot for comparing the weight and height of men and women",

+ x="Weight", y="Height")

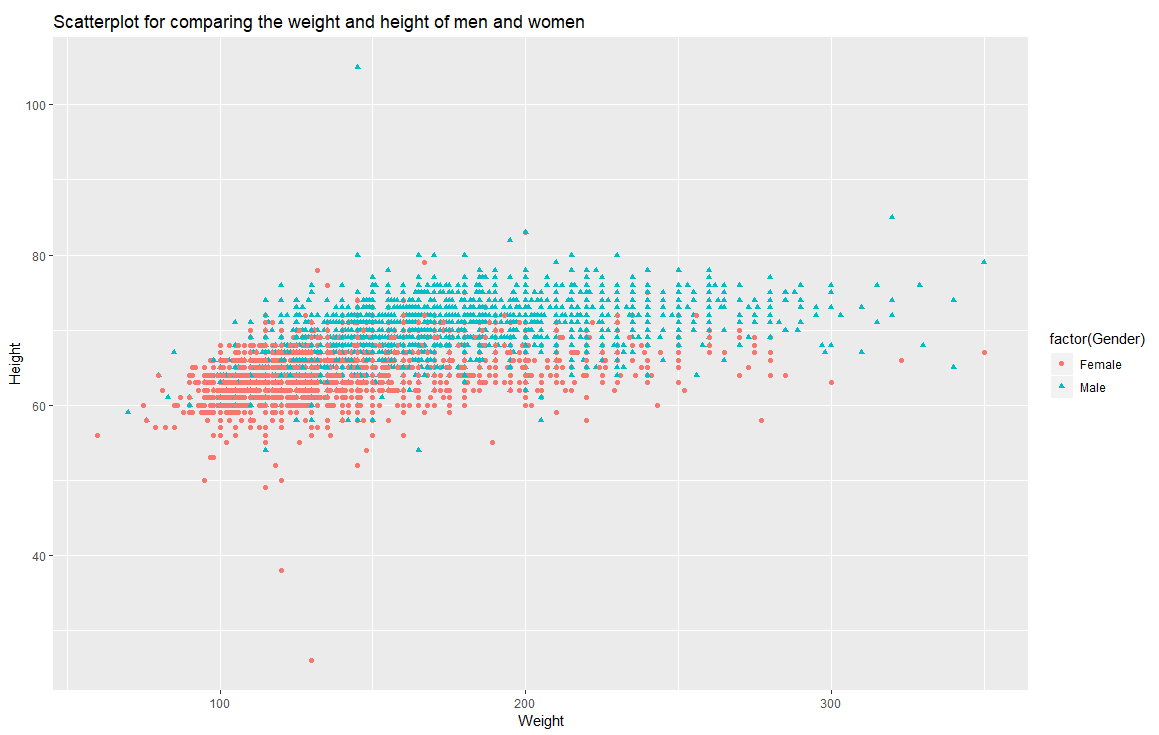


Figure:6 Scatterplot for comparing Weight and Height for Men and Women.

**PART 3:**

**Then, using the height\_weight\_byGender.csv file, we plotted a scatterplot of height by weight for either men or women using the plot() R function.**

**CODE:**

**> #Scatterplot**

>plot(Height\_Weight\_by\_Gender$`Height(Inches)`[Height\_Weight\_by\_Gender$Gender=="Male"],

+      Height\_Weight\_by\_Gender$`Weight(lbs)`[Height\_Weight\_by\_Gender$Gender=="Male"],

+      xlab="Height(inches)",ylab="Weight(lbs)",main="Scatterplot of Height and Weight for Men",col=c("red", "blue"))

> legend("topright", cex=0.8, pch=1, pt.cex = 0.5, legend=c("Height","Weight"), lwd=c(2,2), col=c("blue","red"))

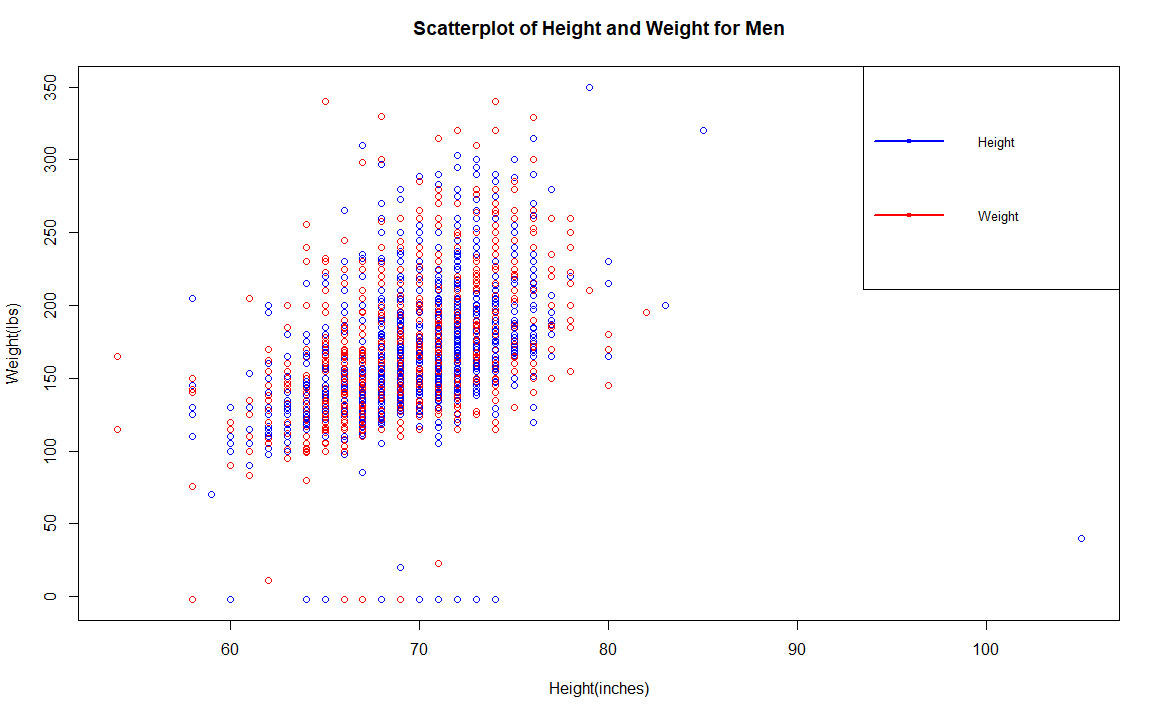


Figure:7 Scatterplot comparing the weight and height of men.

**PART 4:**

**Descriptive statistics is all about describing a sample of data (or a population). Above, you utilized descriptive statistics to describe the height and weight of individuals within the “height\_weight” dataset. What is an example of a statistical inference question we may ask next?**

**Solution:**

1. What is the range of our true value?
2. On basis of the given hypothesis our hypothesis is accepted or not?

References:

* R Histograms. (2019). www.tutorialspoint.com. Retrieved 27 February 2019, from <https://www.tutorialspoint.com/r/r_histograms.htm>
* Programming with R: Reading and Writing CSV Files. (2019). Swcarpentry.github.io. Retrieved 27 February 2019, from <https://swcarpentry.github.io/r-novice-inflammation/11-supp-read-write-csv/>
* contributors, D. (2019). Data visualization with ggplot2. Datacarpentry.org. Retrieved 27 February 2019, from <https://datacarpentry.org/R-ecology-lesson/04-visualization-ggplot2.html>
* R Scatterplots. (2019). www.tutorialspoint.com. Retrieved 27 February 2019, from <https://www.tutorialspoint.com/r/r_scatterplots.htm>