**BUSINESS DATA MINING**

**(IDS 572)**

**Solutions to Homework 1**

**Group Members**

* Amey Pophali (apopha2@uic.edu)
* Karthik Varanasi (vvaran3@uic.edu)
* Mrinal Dhawan (mdhawa3@uic.edu)

**Problem 1 -**

* **How many attributes and instances are in this data set? How do you get this information in RStudio?**

**Ans.** We can get the values by **dim(hw1)** command.

*Attributes = 6,*

*Instances = 37*

Note: hw1 is the name of the dataset.

* **Write down the measurements of all the variables.**

**Ans.** The following output can be viewed using **str(hw1)** command.

*Weight: INT*

*Height: INT*

*Gender: Factor*

*Average Exercise: INT*

*Exercise Yes or NO: Factor*

*Fried Items Frequency: Factor*

We can get the measurements of the dataset using the command **attributes(hw1)** in R Studio.

The unit of measurements of the variables are as follows –

*Weight: Pounds*

*Height: Inches*

*Gender: Male or Female*

*Exercise Average: Days*

*Exercise Regularly: Yes or No*

*Fried Items Frequency: None, Less than 3 times or More than 3 times*

* **Find the overall mean, median, variance and sample standard deviation of weight variable in**

**this data set.**

**Ans.** We can get the values using the following commands –

weights=hw1$Weight

mean(weights)

median(weights)

var(weights)

sd(weights)

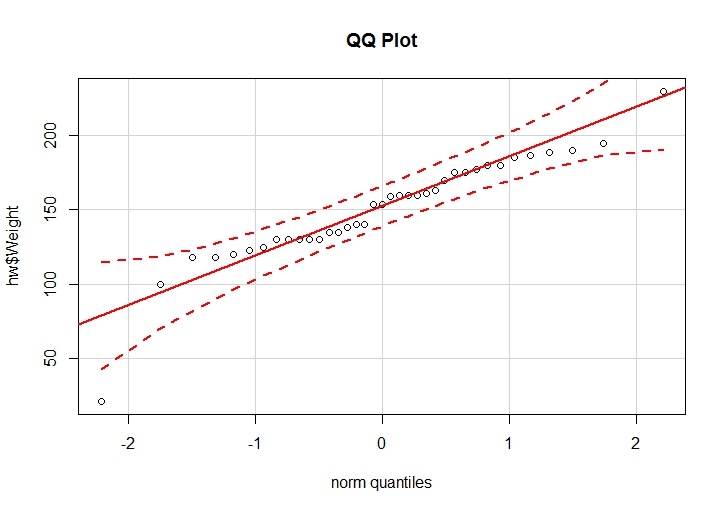
*Mean = 150.4594, Median = 154, SD = 35.45388, Variance = 1256.978*

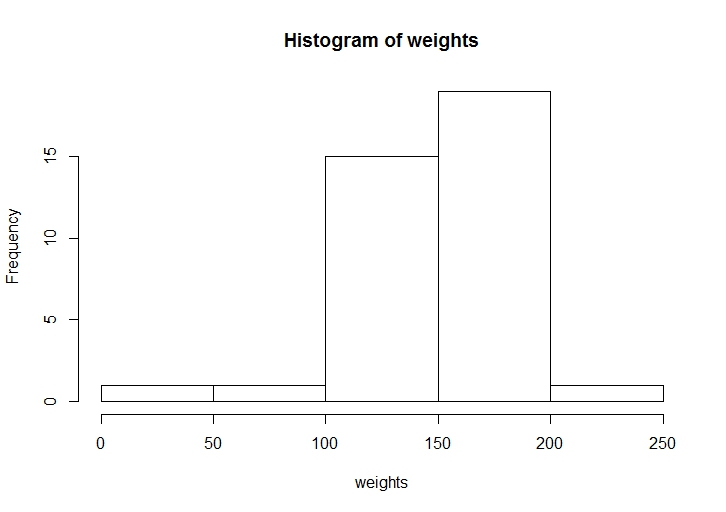
* **Does the distribution of the weight data have a symmetrical belled-shape? Justify your answer.**

**Ans.**

We can draw a QQ plot using the command **qqnorm(hw1$weights)** and **qqline(hw1$weights)**. The weight value QQ plot is shown. As can be seen from the QQ plot, the distribution of the weight variable is normal i.e. it is Bell shaped. But it does have outliers. Hence it is difficult to say if the data is normal or not.

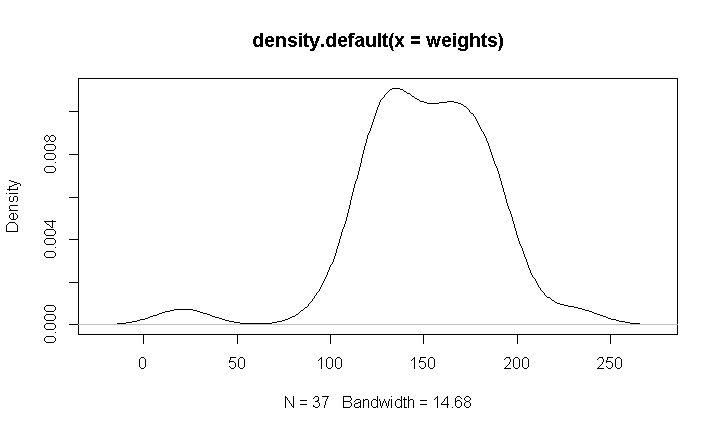
Going forward, we can draw a histogram of weight using the command **hist(weights)**. The histogram is shown below**.** By looking at the histogram we can declare that the data is left skewed and is not normal.





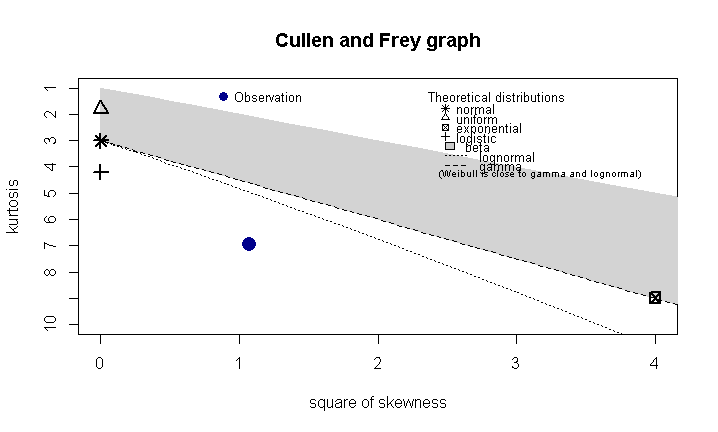
The density distribution of the weights is shown can be plotted as shown below.

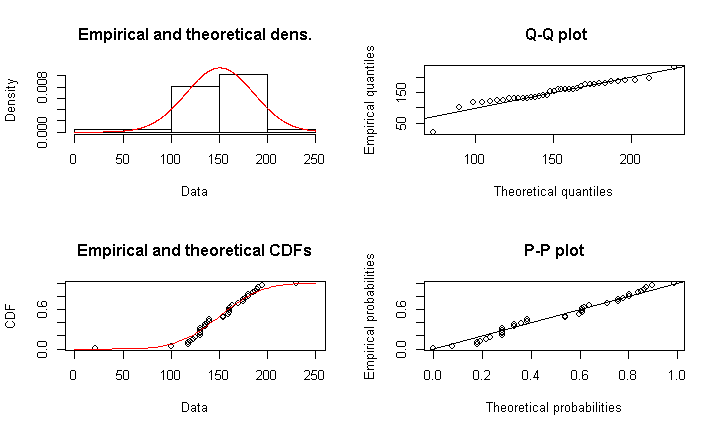
**plot(density(weights))**



Method 2:

By Using the package “**fitdistrplus**” 🡪 fits the distribution to the respective data.





The output of the R is as below:

**descdist(weights,discrete=FALSE)**

*summary statistics*

*------*

*min: 21 max: 230*

*median: 154*

*mean: 150.4595*

*estimated sd: 35.45388*

*estimated skewness: -1.035904*

*estimated kurtosis: 6.946055*

**fit.norm <- fitdist(weights, "norm")**

**plot(fit.norm)**

# Fit of the normal distribution to the given data at QQ plot and find the normality of the function

**fit.norm$aic**

*[1] 372.0369*

* **Report the percentage distribution of the Ate Fried Food variable?**

**Ans.** We can use the following commands to get the output

1. **rel.vals = 100\*table(hw1$FriedItemsFreq)/nrow(hw1)**
2. **rel.vals**

Output -

*At least 3 times Less than 3 times None*

*8.108108 51.351351 40.540541*

|  |  |
| --- | --- |
| **Ate Fried Food** | **Relative Frequency** |
| None | 51.351351 |
| Less than 3 times | 40.540541 |
| At least 3 times | 8.108108 |

* **Report the percentage distribution of Exercise Per Week variable. Write your answer into a**

**table similar to the table in part (e).**

Commands

**rel.vals = 100\*table(hw1$ExerciseAv)/nrow(hw1)**

**rel.vals**

|  |  |
| --- | --- |
| **Exercise Average** | **Relative Frequency** |
| 1 | 10.810811 |
| 2 | 21.621622 |
| 3 | 18.918919 |
| 4 | 21.621622 |
| 5 | 24.324324 |
| 7 | 2.702703 |

* **Does the weight data suggest that it was from a normally distributed population? Perform a normality test and report the p-value of the test using 5% as the auto for decision making of the normality test.**

**Ans.** We can check the normality by Shapiro Wilk’s Normality Test

**shapiro.test(hw1$Weight)**

*Shapiro-Wilk normality test*

*data: hw1$Weight*

*W = 0.9151, p-value = 0.007959*

As the p-value is significantly less than 0.05 (the standard alpha value of 5%), we can reject the null hypothesis. Hence the data is not normal.

* **Report the mean, median and sample standard deviation of weight variable for female subjects in this data set.**

**Ans.** We can get the output using the commands -

**aggregate(hw1$Weight, by=list(hw1$Gender),FUN=mean)**

**aggregate(hw1$Weight, by=list(hw1$Gender),FUN=median)**

**aggregate(hw1$Weight, by=list(hw1$Gender),FUN=sd)**

Outputs –

*Mean = 147.29*

*Median = 140*

*SD = 25.08*

**Problem 2 –**

* **Make a frequency distribution table for the gender variable to see the frequency distribution.**

**Ans.** Commands -

**gen = hw1$gender**

**gen.freq = table(hw1$gender)**

**gen.freq**

Output -

*Female Male*

*27 10*

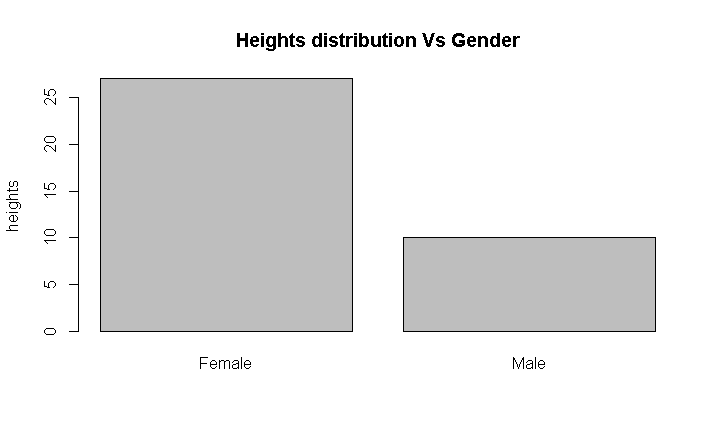
* **Make a bar chart for gender variable.**

**Ans.** We can get the barplot of the data using the command -

**barplot(gen.freq,ylab='heights',main ='Heights distribution Vs Gender')**

Note: ‘gen.freq’ is the variable from previous question.

The barplot looks as follows -

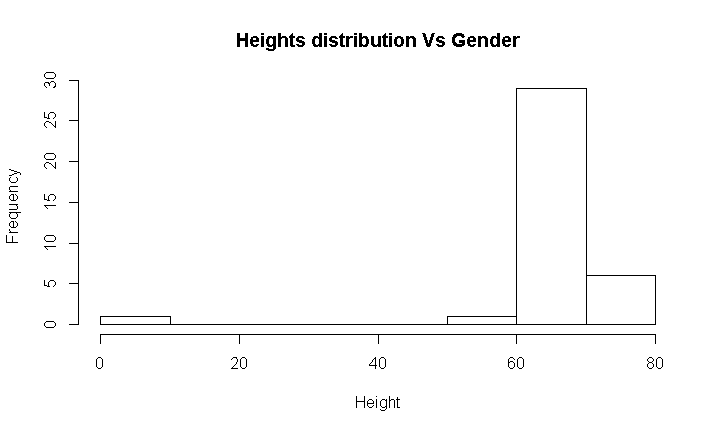


* **Make a histogram to display the distribution of the Height variable.**

**Ans.** Commands:

**ht = hw1$Height**

**hist(ht)**



* **Make a cluster bar chart (side-by-side bar chart) to examine the correlation between gender and Ate Fried Food variables.**

Command:

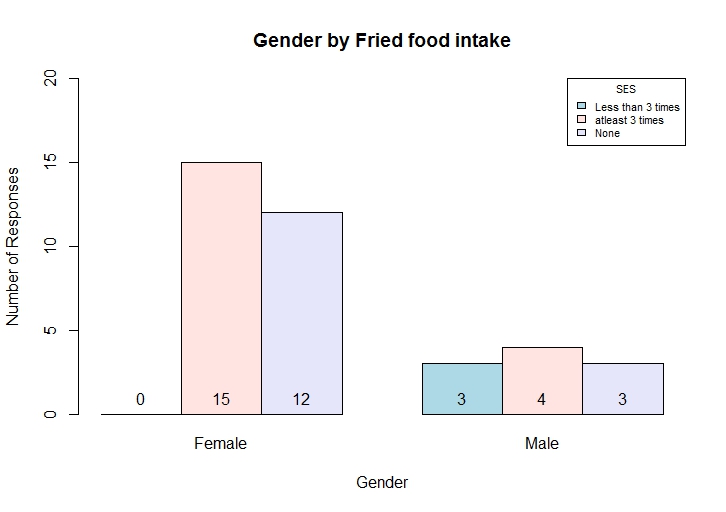
**bp<-barplot(table(hw1$FriedItemsFreq,hw1$Gender),beside=T, main = "Gender by Fried food intake", col = c("lightblue", "mistyrose", "lavender"),xlab = "Gender", names = c("Female", "Male"),**

**ylab = "Number of Responses", legend = c("Less than 3 times", "atleast 3 times", "None"),**

**args.legend = list(title = "SES", x = "topright", cex = .7), ylim = c(0, 20))**

**text(bp, 0, round(hw1$FriedItemsFreq,hw1$Gender), 1),cex=1,pos=3)**

Output:



* **Make a scatter plot to examine the correlation between Weight and Height variables, and write**

**a sentence to describe the trend you observed from the scatter plot.**

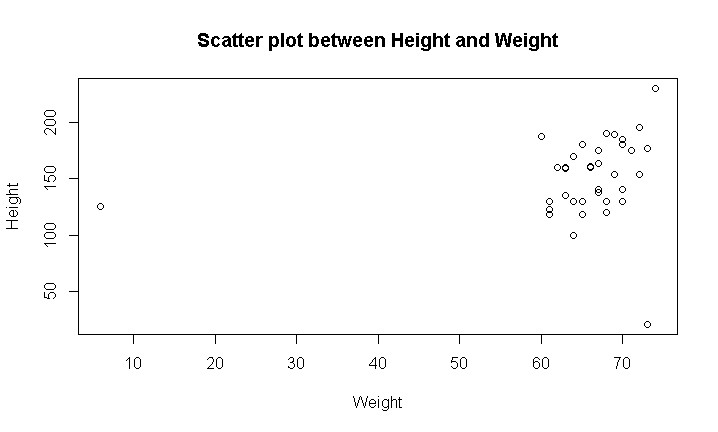
**Ans.**

Method 1

Command:

**plot(hw1$Height,hw1$Weight,xlab='Weight',ylab='Height',main="Scatter plot between Heig ht and Weight")**

Output -



Method 2

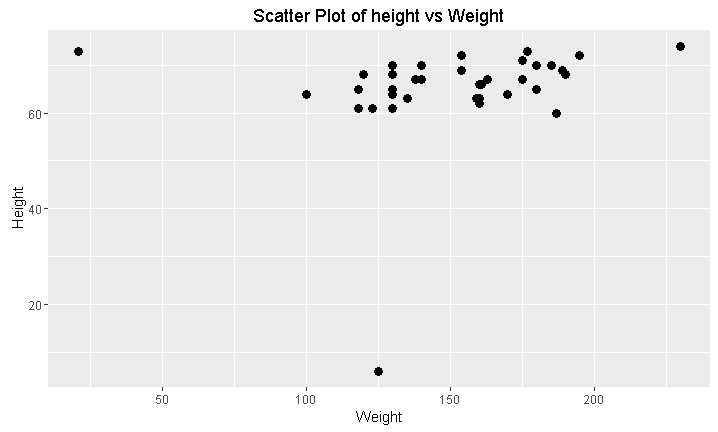
Command:

**library(ggplot2)**

**sc\_plot<-ggplot(hw1,aes(hw1$Weight,hw1$Height))**

**sc\_plot+geom\_point(size = 3)+xlab("Weight")+ylab("Height")+ggtitle("Scatter Plot of height vs Weight")**

Output -



* **Find the 5-number summary for the Height data and make a boxplot for the Height data with**

**mild and extreme outliers identified using inner and outer fences. Draw the boxplot.**

Code:

**summary(hw1$Height)**

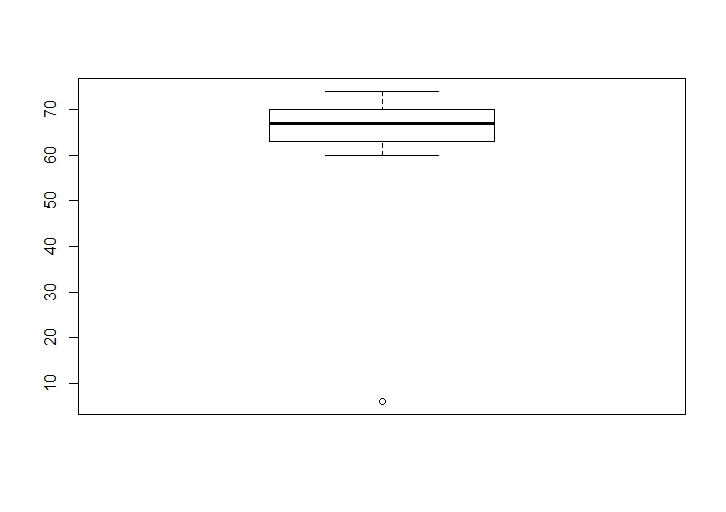
Output:

*summary(hw1$Height)*

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

*6.00 63.00 67.00 65.05 70.00 74.00*

Boxplot –

****

Extreme outliers are the points that are outside the limit Q1-3(IQR) or Q3+3(IQR).

Mild outliers are the points that are outside the limit Q1-1.5(IQR) or Q3+1.5(IQR).

**Problem 3 -**

* **Find the support and confidence for the rule**

**(Give Birth = no) ^ (Can Fly = yes) ! Birds**

**Ans.**

* Support = = proportion of records predicted correctly.
* Confidence = = proportion of correct predictions among records where it applies.

Here,

A = (Give Birth = no) ^ (Can Fly = yes)

B = (Can Fly = yes) -> Birds

Target = Birds

From the table, we see that they are independent events. Therefore,

Support = 3/20

Confidence = 1

* **Using the 1-rule method discussed in class, and the relevant sets of classification rules for the**

**target “Class" by testing each of the input attributes Blood Type, Give Birth, Can Fly, and**

**Live in Water. Which of these three sets of rules has the lowest misclassification rate?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Evaluating Attributes in the Animals Data | | | | |
|  | **Attribute** | **Rules** | **Errors** | **Total Errors** |
|  |  |  |  |  |
| 1. | Blood Type | Warm-> Mammals | 4/11 | 9/20 |
|  |  | Cold -> Reptiles | 5/9 |  |
| 2. | Gives Birth | Yes-> Mammals | 1/7 | 1/2 |
|  |  | No-> Reptiles | 9/13 |  |
|  |  |  |  |  |
| 3. | Can Fly |  |  |  |
|  |  | Yes -> Birds | 1/4 | 11/20 |
|  |  | No -> Mammals | 10/16 |  |
|  |  |  |  |  |
| 4. | Live in Water | Yes -> Fishes | 2/5 | 1/2 |
|  |  | No -> Mammals | 6/11 |  |
|  |  | Sometimes -> Amphibians | 2/4 |  |