Business Data Mining (IDS 572)

Homework 1-Solution

Question 1

(a) There are 6 attributes and 37 instances in this data set. These information can be obtained in R using dim() function.

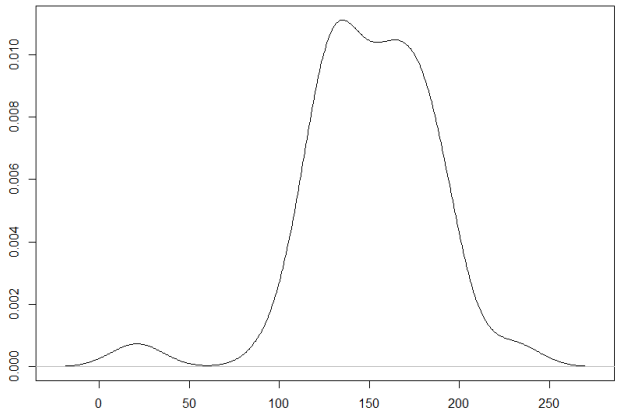
(b) Weight: Continuous (Integer), Height: Continuous (Integer), Gender: Binary (Factor with 2 levels: Female and Male), Exercise Per Week: Continuous (Integer), Regular Exercise: Binary (Factor with 2 levels: Yes and No), Ate Fried Food: Nominal (Factor with 3 levels: At least 3 times, Less than 3 times, and None). These information can be obtained in R using str() function.

(c) Mean = 150.5, Standard Deviation = 35.45, Variance = 1256.977 and Median = 154.

These information can be obtained in R using mean(), sd(), var(), median(), or summary() functions.

(d) No. Below is the density of the variable Weight. As you can see the distribution is not symmetrical bell-shaped.

plot(density(Data$Weight))



(e) Using the function table(Data$AteFriedFood)/nrow(Data) we get

Ate Fried Food	Relative Frequency
None	40.54%
Less than 3 times	51.35%
At least 3 times	8.11%

(f) Using table(Data$ExercisePerWeek)/nrow(Data) we have

Exercise Per Week	Relative Frequency
0 Day	0%
1 Day	10.81%
2 Days	21.62%
3 Days	18.92%

4 Days	21.62%
5 Days	24.32%
6 Days	0%
7 Days	2.70%

(g) Using the function shapiro.test(Data$Weight) we get

```
 Shapiro-Wilk normality test

data:  Data$Weight
W = 0.9151, p-value = 0.007959
```

The p-value of Shapiro test is 0.0079, which is more than 0.05. We can therefore conclude that the data is unlikely from a normal distribution.

To answer to this question, you could also use Q-Q Plot.

(h) Mean = 147.29, Standard Deviation = 25.08, Median = 140

We can find these information in R using the following code:

```
aggregate(Data$Weight, by = list(Data$Gender), FUN = mean)
aggregate(Data$Weight, by = list(Data$Gender), FUN = sd)
aggregate(Data$Weight, by = list(Data$Gender), FUN = median)
```

Question 2

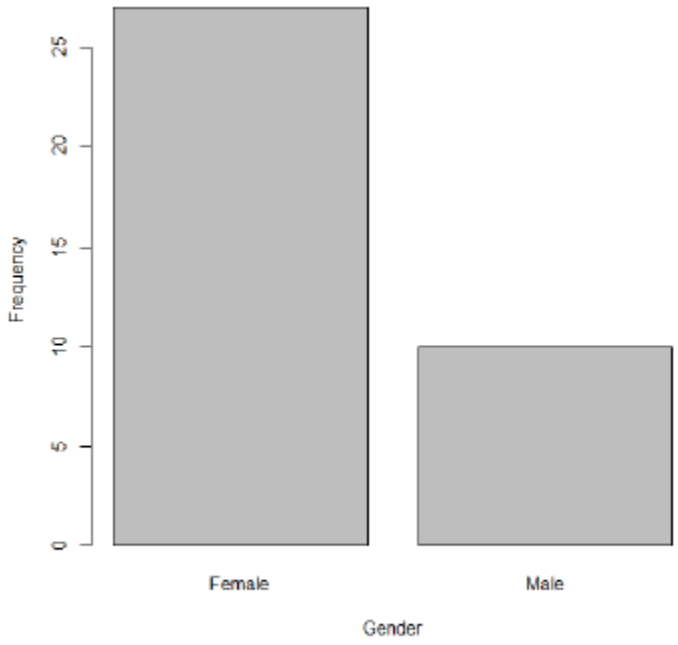
(a) To get the frequency of a categorical variable we can use the table() function in R.

Frequency distribution table for Gender variable

Gender	Frequency
Male	27
Female	10

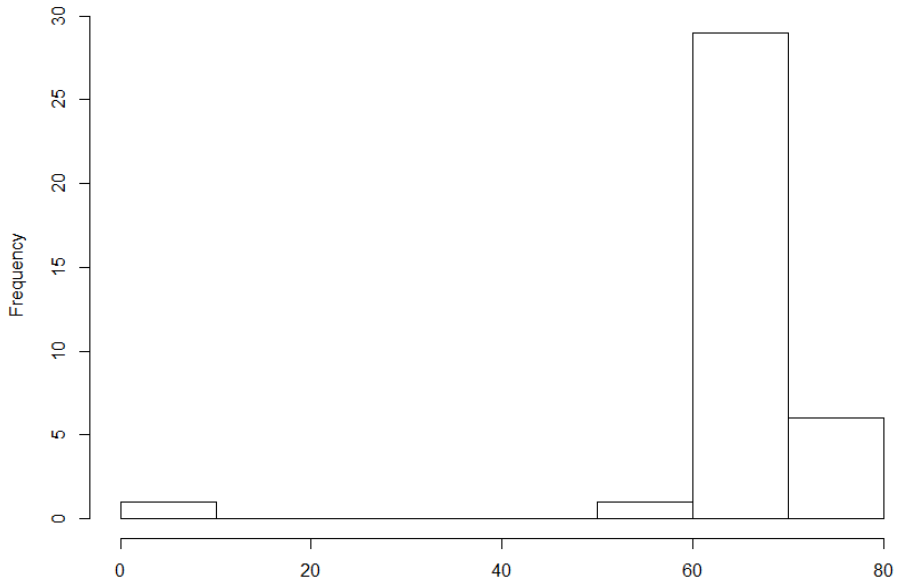
The result shows that the majority of instances are Male.

(b) Using the function barplot(table(Data$Gender)) we have

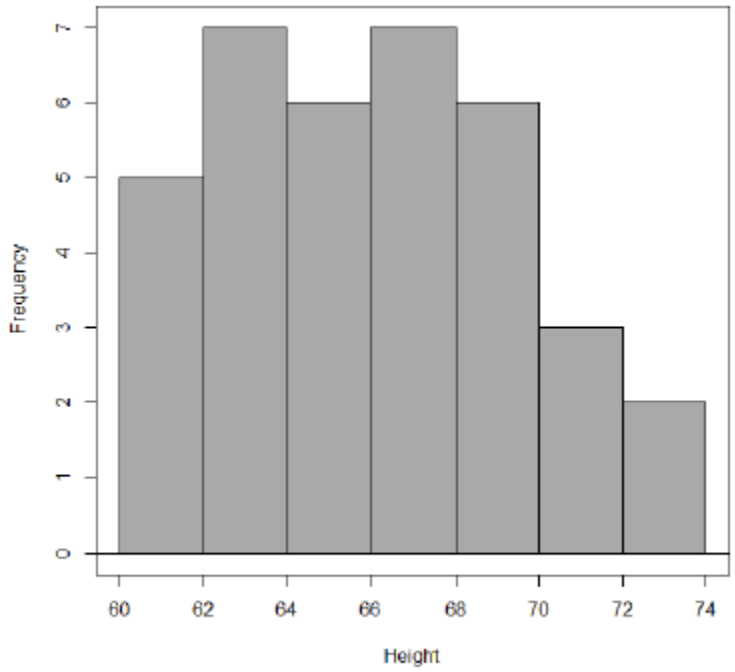


Frequency distribution of males and females.

(c) We can get the distribution of Height variable using hist() function



As can be seen, there are outliers in data. After removing the outliers we get

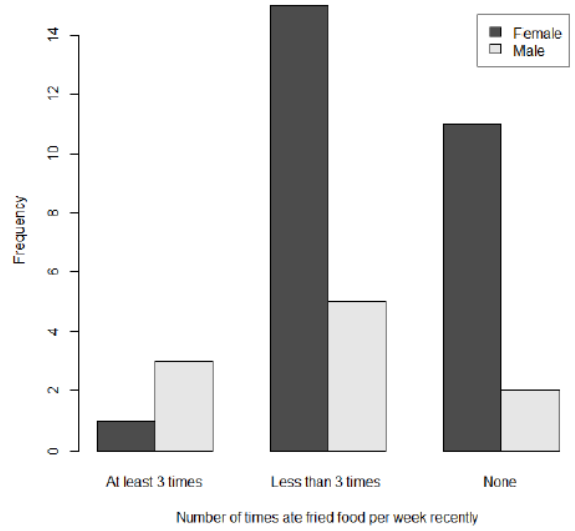


Frequency Distribution of Height.

This distribution is skewed to the right. Majority of instances have height between 62 and 68 inches.

(d) To get cluster barchart we can use the following code

Barplot(table(Data$Gender, Data$AteFriedFood), xlab = “Number of items ate fried food per week recently”, ylab = “Frequency”, legend = rownames(counts), beside = TRUE)

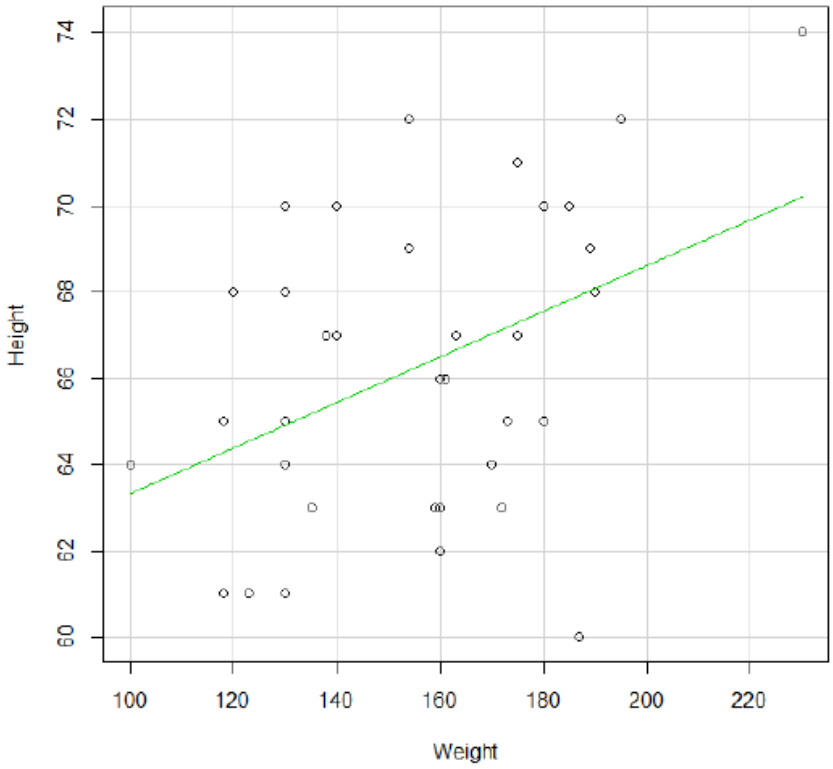


Cluster Bar Chart for Number of Fried Foods Eaten by Gender.

This graph indicates that there are few females who eat fried foods at least three times; majority of them eat fried food less than 3 times or none. But the distribution of Ate Fried Food is more uniform for men.

Notice that the above plot is drawn after removing the outliers.

(e) After removing the outliers from data set, the scatter plot is as follows:



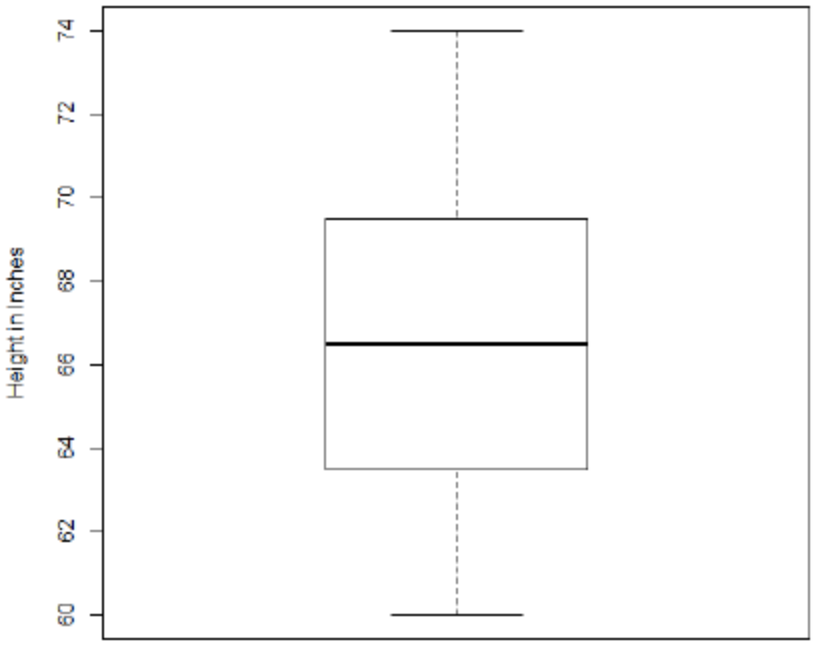
Scatter Plot for Weight versus Height Variables.

This graph indicates that there is a positive relationship between height and weight variables. Notice that if you do not remove the outliers from the data set, you may observe a weak correlation between these two variables.

To obtain the scatterplot the function plot(Data$Weight, Data$Height), abline(lsfit(Data$Weight, Data$Height), col = “green”) is used.

(f) Five number summary is

Min = 60, Q1 = 63.75, median = 66.5, Q3 = 69.25, max = 74



Question 3

(a) Let A = No giving birth and B = Can fly and C = Birds. Then

Support = P(A & B & C) = 3/20 = 0.15%

Confidence = P(C | A & B) = 3/3 = 100%

(b) The rules obtained by 1-rule method are of the following form:

Blood Type

- If Blood Type = warm then Mammals

(misclassification rate: 4/11)

- If Blood Type = cold then Reptiles

(misclassification rate: 5/9)

Therefore, the Blood Type attribute's error rate is: 11/20 × (4/11) + 9/20 × (5/9) = 9/20

Give Birth

- If Give Birth = yes then Mammals

(misclassification rate: 1/7)

- If Give Birth = no then Reptile

(misclassification rate: 9/13)

Therefore, the Give Birth attribute's error rate is: 7/20 × (1/7) + 13/20 × (9/13) = 10/20

Can Fly

- If Can Fly = yes then Birds

(misclassification rate: 1/4)

- If Can Fly = no then Mammals

(misclassification rate: 10/16)

Therefore, the Can Fly attribute's error rate is: 4/20 × (1/4) + 16/20 × (10/16) = 11/20

Live in Water

- If Live in Water = sometimes then Amphibians

(misclassification rate: 2/4)

- If Live in Water = yes then Fish

(misclassification rate: 2/5)

- If Live in Water = no then Mammals

(misclassification rate: 6/11)

Therefore, the Live in Water attribute's error rate is: $4/20 \times (2/4) + 5/20 \times (2/5) + 11/20 \times (6/11) = 10/20$

The set of rules for "Blood Type" has the lowest total misclassification rate.