Project -2

Statistical Methods for Data Science (2019)

Group Members -

Lakshmi Sindhuri Pallapothu (LXP170004)

Gnaneswar Gandu (GXG170000)

Contributions -

- Solved 1 and 2 together.
- R code for Question 1 was done by Gnaneswar.
- R code for Question 2 and report were done by Sindhuri

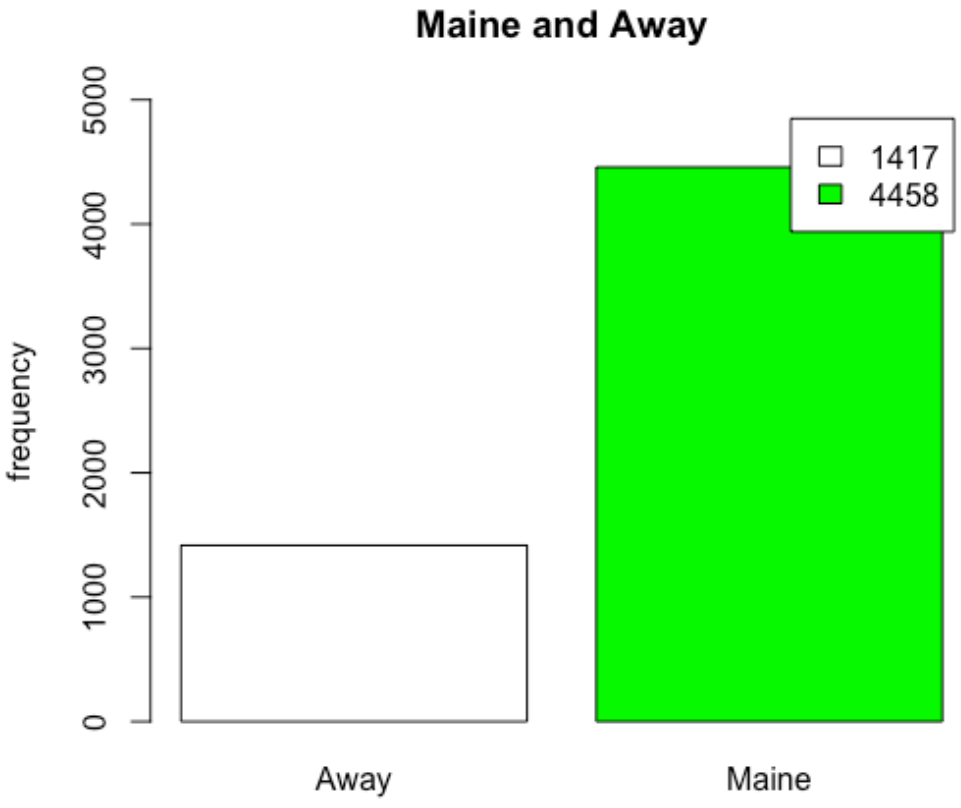
Question 1

Consider the dataset roadrace.csv posted on eLearning. It contains observations on 5875 runners who finished the 2010 Beach to Beacon 10K Road Race in Cape Elizabeth, Maine. You can read the dataset in R using read.csv function.

Steps -

- Place the csv file in a folder and make that folder the working directory in R.
- After that read the csv file and store it in a variable so we can access it many times.

(a) Create a bar graph of the variable Maine, which identifies whether a runner is from Maine or from somewhere else (stated using Maine and Away). You can use bar plot function for this. What can we conclude from the plot? Back up your conclusions with relevant summary statistics.

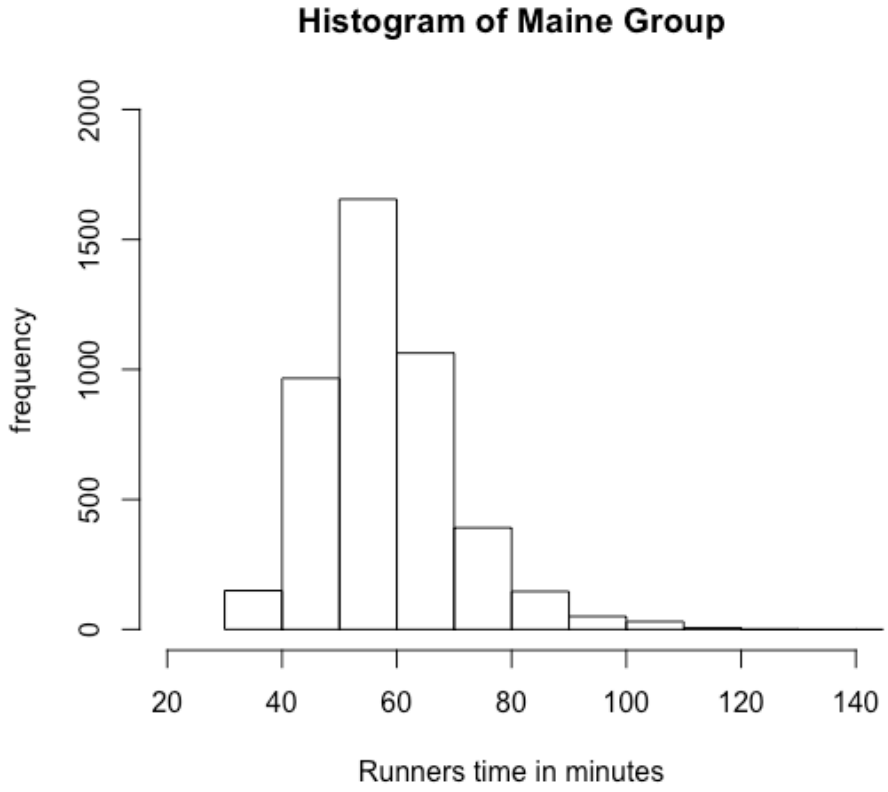


BAR GRAPH OF VARIABLE MAINE

Conclusion - It can be concluded from the graph that number of runners from **Marine** group are **4458** and number of runners from **Away** group are **1417.** The Proportion of Maine group is approximately 50% more than that of of Away group.

R code for the above is available in R Code section.

(b) Create two histograms the runners' times (given in minutes) — one for the Maine group and the second for the Away group. Make sure that the histograms on the same scale. What can we conclude about the two distributions? Back up your conclusions with relevant summary statistics, including mean, standard deviation, range, median, and interquartile range.

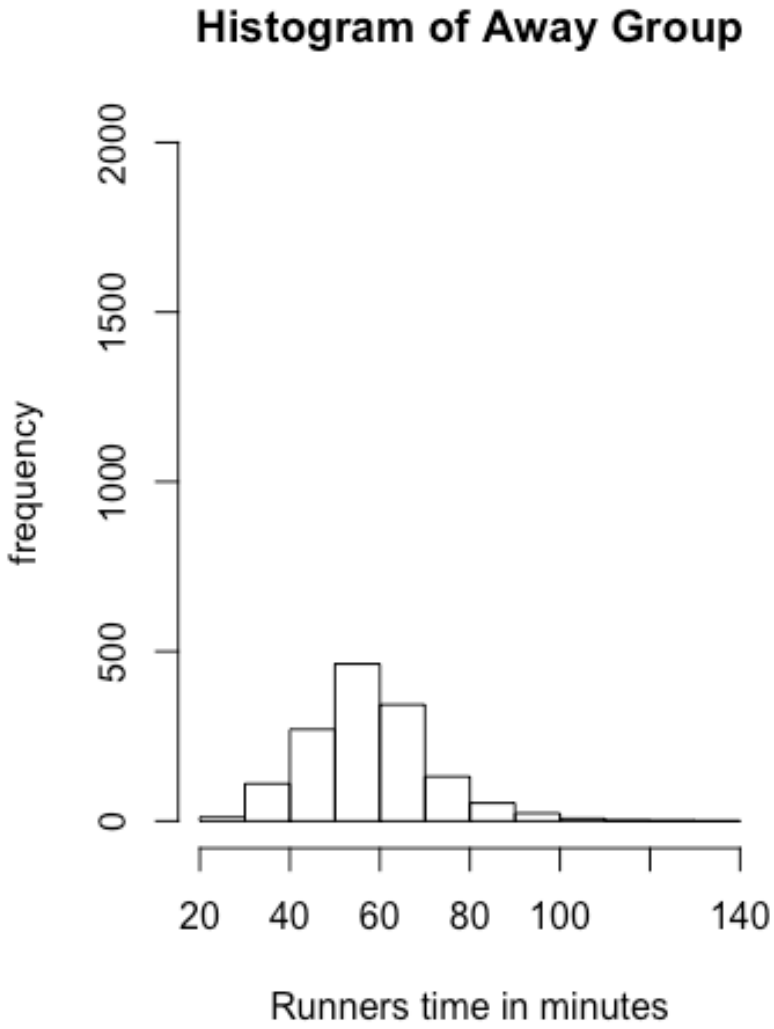


Statistics for Maine Group -

Summary -

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
30.57	50.00	57.03	58.20	64.24	152.17

Mean - 58.20
Standard Deviation - 12.18511
Range - 30.567 - 152.167
Median - 57.03
Inter Quartile Range - 14.24775



Statistics for Away Group -

Summary -

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
27.78	49.15	56.92	57.82	64.83	133.71

Mean - 57.82
Standard Deviation - 12.18511
Range - 27.782 - 133.710
Median - 56.92
Inter Quartile Range - 15.674

Conclusions -

From the Histograms and statistical data, the below information can be concluded

Histograms – Both are Right sewed

Mean – Mean for both the groups is almost similar (Maine – 58.20, Away – 57.03)

Standard Deviation – It is observed that deviation of Away group is slightly more than that of Maine group (Maine – 12.18, Away – 13.83)

Range - Range which is equal to the difference of max and min values is also greater for Maine group (121.6) when compared to Away group (105.93)

Median – Median for both the groups is almost similar (Maine – 57.03, Away – 56.92)

Inter Quartile Range – It is observed that deviation of Away group is slightly more than that of Maine group (Maine – 14.24, Away – 15.67)

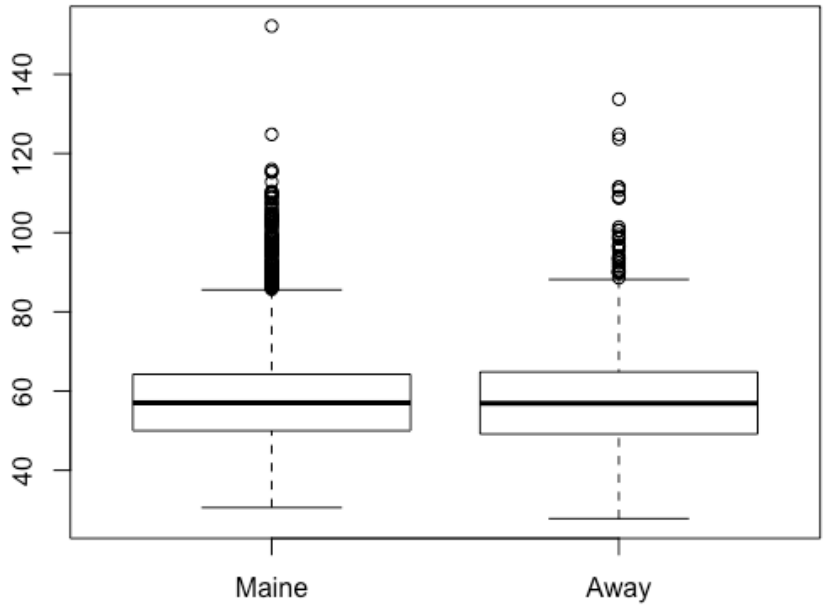
Maximum – Maximum value from both the groups was observed in Maine Group (152.167)

Minimum – Minimum value from both the groups was observed in Away Group (27.782)

R code is Available in Code Section

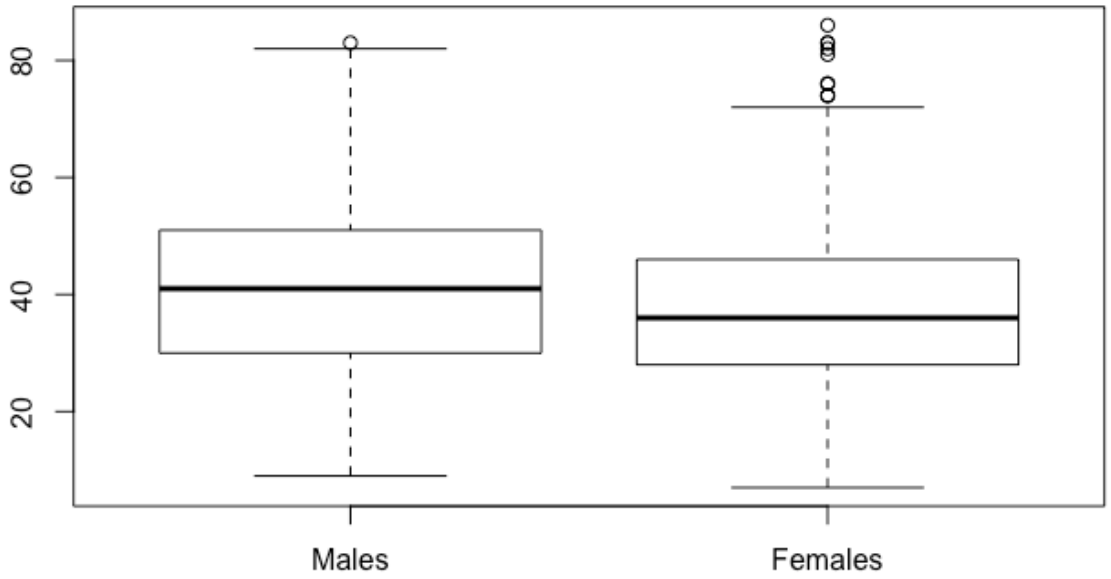
(c) Repeat (b) with side-by-side box-plots

Side-By-Side Plot - Maine VS Away



(d) Create side-by-side box-plots for the runners' ages (given in years) for male and female runners. What can we conclude about the two distributions? Back up your conclusions with relevant summary statistics, including mean, standard deviation, range, median, and interquartile range.

Side-By-Side Plot -



Statistics for Male Group -

Summary –

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
9.00	30.00	41.00	40.45	51.00	83.00

Mean - 40.45

Standard Deviation - 13.99

Range - 9 - 83

Median - 41.00

Inter Quartile Range - 21

Statistics for Female Group -

Summary –

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
7.00	28.00	36.00	37.24	46.00	86.00

Mean - 37.24

Standard Deviation - 12.26

Range - 7 - 86

Median - 36.00

Inter Quartile Range - 18

Conclusions -

From the side by side plot and statistical data, the below information can be concluded

Mean - Mean for male group is higher than female group (Male – 40.45, Female – 37.24)

Standard Deviation – It is observed that deviation of Male group is slightly more than that of Female group (Male – 13.99, Female – 12.26)

Range - Range which is equal to the difference of max and min values is lesser for Male group (74) when compared to Female group (79)

Median – Median is also observed to be greater for male group (Male – 41, Female – 36)

Inter Quartile Range – It is observed that deviation of Male group is slightly more than that of Female group (Male – 21, Female – 18)

Maximum – Maximum value from both the groups was observed in Female Group (86)

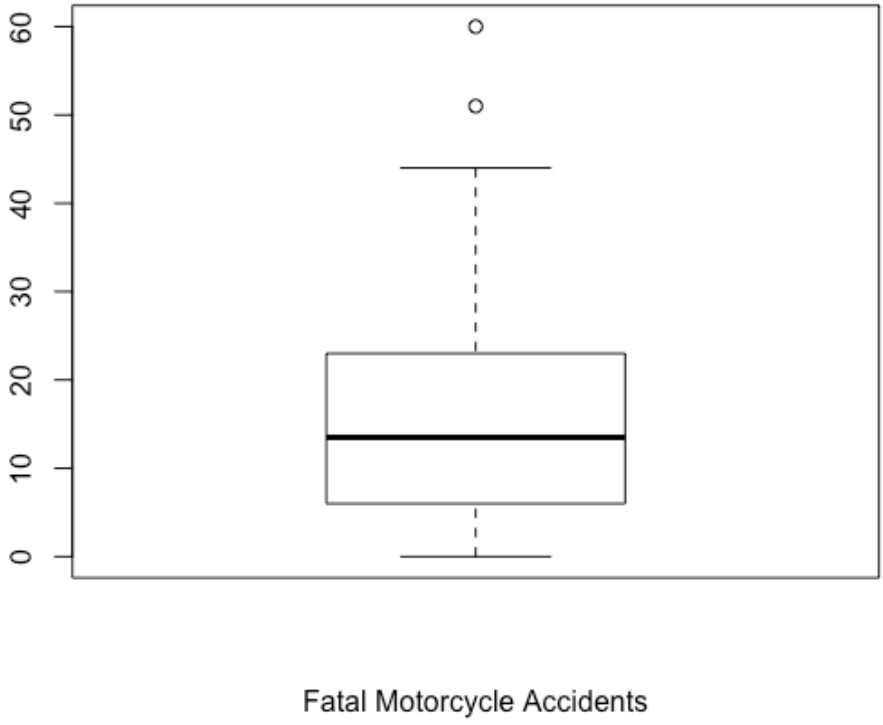
Minimum – Minimum value was also observed in Female Groups (86)

R code is available in code section.

Question 2

Consider the dataset motorcycle.csv posted on eLearning. It contains the number of fatal motorcycle accidents that occurred in each county of South Carolina during 2009. Create a box-plot of data and provide relevant summary statistics. Discuss the features of the data distribution. Identify which counties may be considered outliers. Why might these counties have the highest numbers of motorcycle fatalities in South Carolina?

A particular value is said to be outlier when it is not within the range of + 1.5*IQR() from 75th quantile and - 1.5*IQR() 25th quantile



Statistics -

Summary -

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

0.00 6.00 13.50 17.02 23.00 60.00

Mean -17.02

Standard Deviation - 13.81256

Range - 0 - 60

Median - 13.50

Inter Quartile Range - 17

Lower Boundary can be calculated using **max((quantile(f, prob=0.25) - 1.5*IQR(f)), min(f))** and

Upper Boundary can be calculated using **min((quantile(f, prob=0.75) + 1.5*IQR(f)),max(f))**.

Now any value which is lesser than lower boundary **or** greater than upper boundary will be considered as an outlier.

We used **which()** condition with OR condition to get final outliers.

Based on R command executions, The county's with highest number of motorcycle fatalities in South Carolina are **GREENVILLE** and **HORRY.**

Probable Reasons for GREENVILLE and HORRY having highest number of fatalities are –

1. Unsafe Line Changes
2. Poor Road Conditions
3. Speeding
4. Driver under influence
5. Alcohol and Drugs can be reason for losing control

R code is available in code section.

Code Section

R Code for Question 1

```
roadrace <- read.csv("roadrace.csv")

## 1st Part

creating a table with Maine

plotColumn <- table(roadrace$Maine)

barplot(plotColumn, ylab = "frequency", ylim = c(0,5000), col = c('white', 'green'), main = "Maine and Away", legend=plotColumn)

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

## 2nd Part

maineGroup = subset(roadrace$Time..minutes., roadrace$Maine=="Maine")

hist(maineGroup, xlab = "Runners time in minutes", ylab="frequency",ylim = c(0,2000), xlim = c(20, 140), main="Histogram of Maine Group")


summary(maineGroup)

IQR(maineGroup)

range(maineGroup)

sd(maineGroup)


> summary(maineGroup)

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

  30.57   50.00   57.03   58.20   64.24  152.17

> IQR(maineGroup)

[1] 14.24775

> range(maineGroup)

[1]  30.567 152.167

> sd(maineGroup)

[1] 12.18511
```

```
awayGroup = subset(roadrace$Time..minutes., roadrace$Maine=="Away")

hist(awayGroup, xlab = "Runners time in minutes", ylab="frequency",ylim = c(0,2000), xlim = c(20, 140), main="Histogram of Away Group")


summary(awayGroup)

IQR(awayGroup)

range(awayGroup)

sd(awayGroup)


> summary(awayGroup)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.78   49.15   56.92   57.82   64.83  133.71
> IQR(awayGroup)
[1] 15.674
> range(awayGroup)
[1]  27.782 133.710
> sd(awayGroup)
[1] 13.83538
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -


## 3rd Part

### With Outliers

boxplot(maineGroup, awayGroup, names = c("Maine", "Away"), outline = TRUE)

### Without Outliers

boxplot(maineGroup, awayGroup, names = c("Maine", "Away"), outline = FALSE)

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

## 4th Part

maleGroup = subset(roadrace$Age, roadrace$Sex=="M")
```

```
femaleGroup = subset(roadrace$Age, roadrace$Sex=="F")

##with Outliers
boxplot(maleGroup, femaleGroup, names = c("Males", "Females"), outline = TRUE)
## without Outliers
boxplot(maleGroup, femaleGroup, names = c("Males", "Females"), outline = FALSE)

summary(maleGroup)
IQR(maleGroup)
range(maleGroup)
sd(maleGroup)

> summary(maleGroup)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   30.00   41.00   40.45   51.00   83.00
> IQR(maleGroup)
[1] 21
> range(maleGroup)
[1]  9 83
> sd(maleGroup)
[1] 13.99289

summary(femaleGroup)
IQR(femaleGroup)
range(femaleGroup)
sd(femaleGroup)
```

```
> summary(femaleGroup)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   28.00   36.00   37.24   46.00   86.00
> IQR(femaleGroup)
[1] 18
> range(femaleGroup)
[1]  7 86
> sd(femaleGroup)
[1] 12.26925
```

- -

```
### Question 2

motorcycles = read.csv("motorcycle.csv")

boxplot(motorcycles)

fatal <- motorcycles$Fatal.Motorcycle.Accidents

boxplot(fatal, xlab="Fatal Motorcycle Accidents")

summary(fatal)
IQR(fatal)
range(fatal)
sd(fatal)
```

```
> summary(fatal)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    6.00   13.50   17.02   23.00   60.00
> IQR(fatal)
[1] 17
> range(fatal)
[1]  0 60
> sd(fatal)
[1] 13.81256

Lowerboundary = max((quantile(fatal, prob=0.25) - 1.5*IQR(fatal)), min(fatal))

Upperboundary = min((quantile(fatal, prob=0.75) + 1.5*IQR(fatal)), max(fatal))

result = fatal[which(fatal < Lowerboundary | fatal > Upperboundary)]

subset(motorcycles$County,motorcycles$Fatal.Motorcycle.Accidents == result)
> subset(motorcycles$County,motorcycles$Fatal.Motorcycle.Accidents == result)
[1] GREENVILLE HORRY
```