STATISTICAL METHODS FOR DATA SCIENCE - MINI PROJECT 2

Names of group members: 1. Manneyaa Jayasanker Net ID: MXJ180040
2. Venkatesh Sankar Net ID: VXS200014

Contribution of each group member:

Manneyaa Jayasanker:
- Worked on R code for 1(a) ,1(b) and 1(c).
- Worked on conclusion for 1(a) and 1(b).
- Wrote documentation for 1(a) ,1(b) and 1(c).

Venkatesh Sankar:
- Worked on R code for 1(d) and 2.
- Worked on conclusion for 1(d) and 2.
- Wrote documentation for 1(d) question 2.

==========

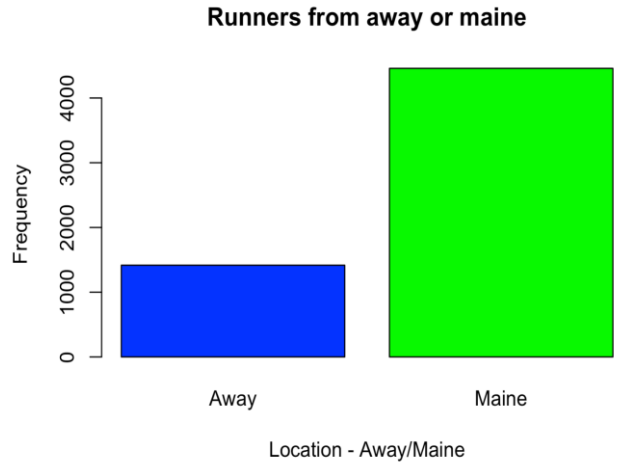
1a)

##Reading data from roadrace.csv

```
> dataset<- read.csv(file="/Users/manneyaajayasanker/Downloads/Stats 3/Projects/MiniProject2/roadrace.csv")
```

##Plotting Barplot for Maine and Away

```
> barplot(table(dataset$Maine),main = "Runners from away or maine",xlab = "Location
 - Away/Maine",ylab = "Frequency",col=c("blue","green"));
```



##Summary statistics for Maine and Away

```
> summary <- table(dataset$Maine)
> summary

 Away Maine
 1417  4458
```

Observations:

We can observe from the bar plot that the number of participants from Maine is at least three times greater when compared to the number of away participants. This conclusion is backed up when we look at the summary statistics of the data. We can also conclude that 75.8% of the total runners are maine runners while only 24.2% of runners are away runners.

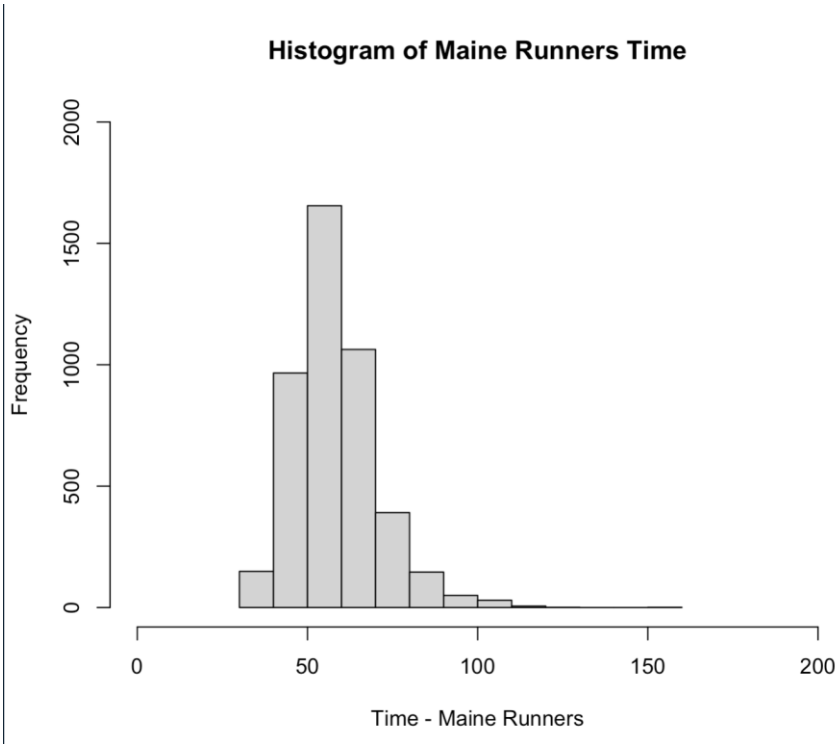
1b)

Storing all the values where Runners are from Maine

```
> maine = subset(dataset, Maine == "Maine")
```

Plotting a histogram

```
> hist(maine[,12],xlab = "Time - Maine Runners", xlim = c(0,200), ylim = c(0,2000), main="Histogram of Maine Runners Time")
```

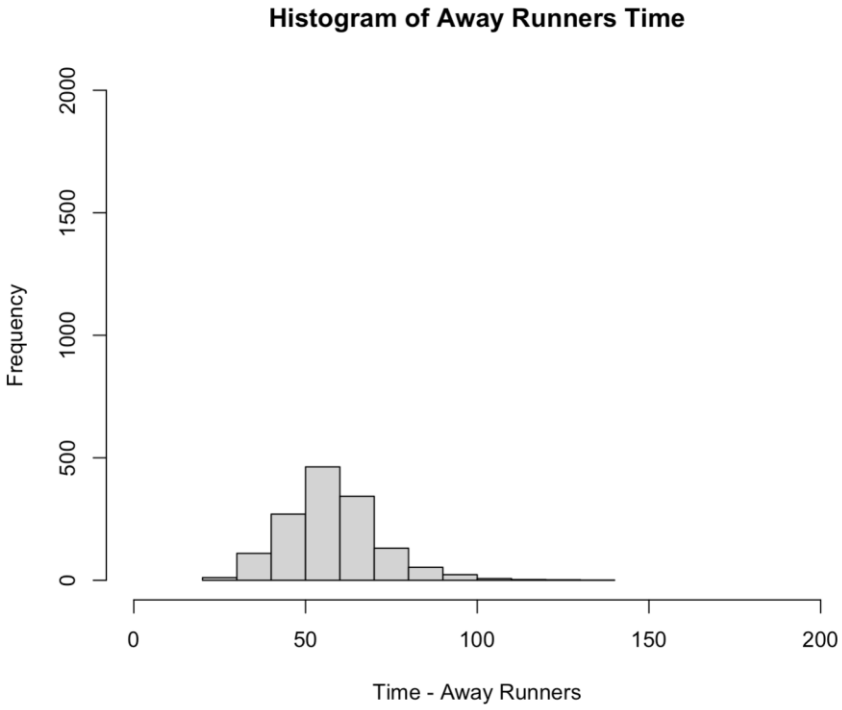


Storing all the values where Runners are from Away

```
> away = subset(dataset, Maine == "Away")
```

Plotting a histogram

```
> hist(away[,12],xlab = "Time - Away Runners", xlim = c(0,200), ylim = c(0,2000), main="Histogram of Away Runners Time")
```



Summary statistics of times of Runners from Maine

```
> summary(maine[,12])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.57   50.00   57.03   58.20   64.24  152.17
```

Range of the runtimes of runners from Maine

```
> range = max(maine[,12])-min(maine[,12])
> range
[1] 121.6
```

Interquartile range of runtime of runners from Maine

```
>  IQR(maine[,12])
[1] 14.24775
```

Standard Deviation of runtime of runners from Maine

```
> sd(maine[,12])
[1] 12.18511
```

Summary statistics of times of Runners from Away

```
> summary(away[,12])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  27.78   49.15   56.92   57.82   64.83  133.71 
```

Range of the runtimes of runners from Away

```
> range = max(away[,12])-min(away[,12])
> range
[1] 105.928
```

Interquartile range of runtime of runners from Away

```
>  IQR(away[,12])
[1] 15.674
```

Standard Deviation of runtime of runners from Away

```
>  sd(away[,12])
[1] 13.83538
```

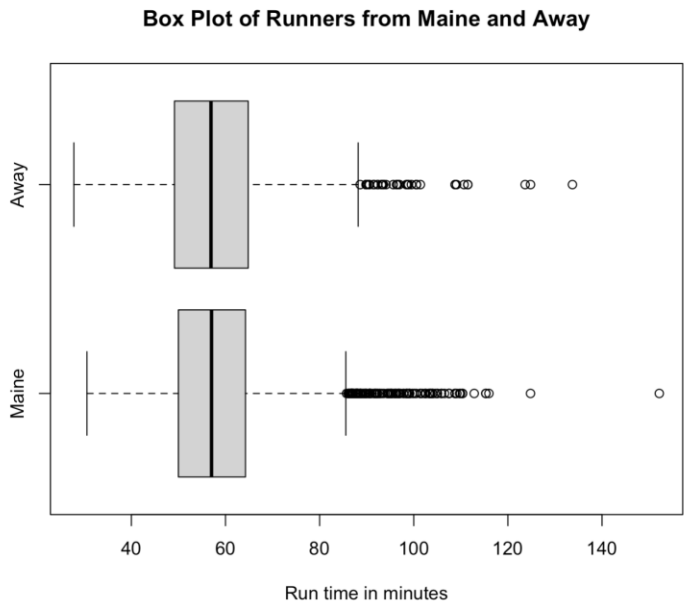
Observations:

Both the histograms are approximately normal distributions. It can also be seen that the range of runner time would be around 130 for the away runners. In the case of Maine runners time, the average is most likely to be around 50-60 with the median in the 50-60 rangeIt can be observed that the mean of away runners time would be around 50-60 with the median also in the same range.

1c)

Plotting side by side box plot for Maine and Away

```
> boxplot(maine[,12],away[,12],names=c("Maine","Away"),horizontal=TRUE, main="Box Plot of Runners from Maine and Away",xlab="Run
 time in minutes")
```

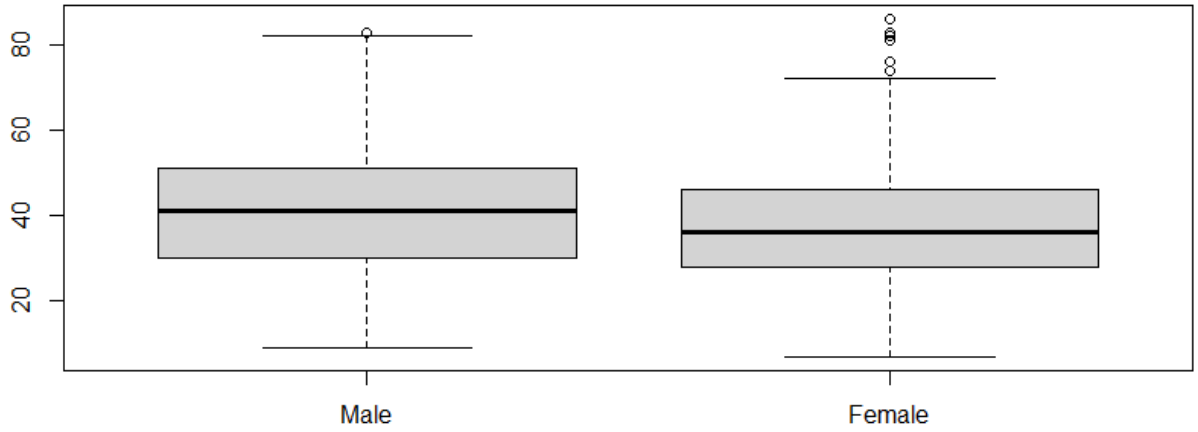


1d) In order to find the ages of male and female runners, we first filter out male and female runners based on 'Sex' column and then boxplot the ages of them based on 'Age' column.

Reading data from roadrace.csv

```
> roadrace <- read.csv(file="/Users/manneyaajayasanker/Downloads/Stats 3/Projects/MiniProject2/roadrace.csv")
```

```
> male = subset(roadrace, Sex == 'M') #filtering out male based on Sex column
> female = subset(roadrace, Sex == 'F') #filtering out female based on Sex column
> boxplot(as.numeric(male$Age), as.numeric(female$Age), names = c("Male", "Female")) #boxplot of male and female Age coln
```



Summary Statistics:

```
> summary(as.numeric(male$Age)) #summary of male Age coln
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   30.00   41.00   40.45   51.00   83.00
> summary(as.numeric(female$Age)) #summary of female Age coln
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   28.00   36.00   37.24   46.00   86.00
```

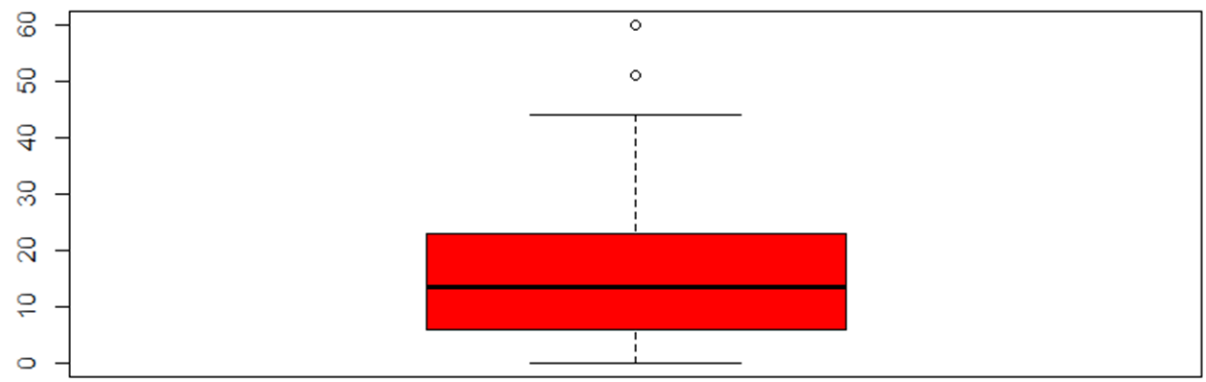
```
> IQR(as.numeric(male$Age)) #InterQuartile range of ages of male runners
[1] 21
> IQR(as.numeric(female$Age)) #InterQuartile range of ages of male runners
[1] 18
```

Observations:

Based on the boxplot and the summary statistics like mean, median, IQR etc, we can find that age of male and female have different distribution of statistics. There are more number of outliers in age of female runners. Age of male runners distribution seem to be left skewed whereas the age of female runners seem to be right skewed.

2. In order to provide observations and analysis of the motorcycle accident dataset, we first read the dataset using read.csv() function and then use the boxplot() to draw boxplot of the Fatal.motorcycle.Accidents column.

```
> setwd("D:/Academics/Spring 21/Statistical Methods Data Science/MP2") #sets working directory
> motorcycle = read.csv("motorcycle.csv") #reads dataset
> boxplot(motorcycle$Fatal.Motorcycle.Accidents, col = 'red')
```



Summary statistics:

```
> summary(motorcycle$Fatal.Motorcycle.Accidents)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    6.00   13.50   17.02   23.00   60.00
```

Observations:

Based on the summary and boxplot, we can observe that motorcycle accident distribution seem to be right skewed. Two counties have the highest number of accidents(greater than 45) whereas the accidents in all other counties fall below 45. There are some counties with zero accidents based on the min value and the maximum is 60.

Outliers:

In order to identify the counties which may be considered outliers, we use $out from boxplot data which gives the values of data points which lie beyond the extremes of the whiskers.

```
> boxplot = boxplot(motorcycle$Fatal.Motorcycle.Accidents)
> boxplot$out #prints outliers values from boxplot
[1] 51 60
```

The values of 51 and 60 correspond to Greenville and Horry county respectively. So these two counties can be considered outliers based on the increased number of motorcycle accidents.

The reason for increased number of motorcycle accidents in these counties can be the negligence and recklessness of the people while driving, road and climatic conditions etc.