Mini Project : 2

Members :
1. Puru Jaiswal (NetID : PXJ200018)
2. Sara Tabassi (NetID : SXT200083)

Both the team members worked together to finish the project. Collaborated to learn R and then write the code. Both of us worked on each of the problems independently and then reviewed each other's answers to determine which solution was optimal.
Both members worked efficiently to complete the project requirements.

Q1) 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 the read.csv function.

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 barplot function for this. What can we conclude from the plot? Back up your conclusions with relevant summary statistics.

```
# Read the data from roadrace.csv

data.1 <- read.csv("./Desktop/UTD\ Course\ Work/Statistical\ Methods\ For\
Data\ Science\ CS_6313/Mini\ Projects/Mini\ Project\ 2/roadrace.csv")

# Apply barplot function on column 11 and generate plots for variables Maine
and Away

barplot(c(sum(data.1$Maine == 'Maine'), sum(data.1$Maine == 'Away')), names.arg
= c('Maine', 'Away'), space = 0.2, ylab = 'Runners')
```

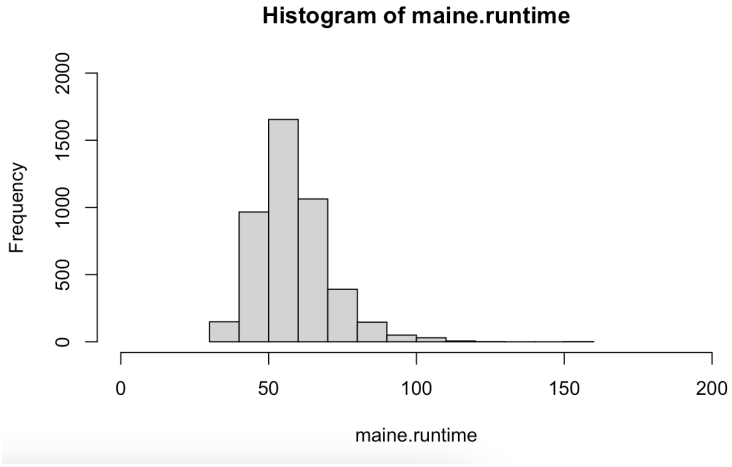
```
> maine.sum = sum(data.1$Maine=='Maine')
> away.sum = sum(data.1$Maine=='Away')
> print(maine.sum)
[1] 4458
> print(away.sum)
[1] 1417
> #Total Runners
> total <- sum(data.1$Maine=='Maine')+sum(data.1$Maine=='Away')
> print(total)
[1] 5875
> #Calculate percentage
> maine.percent <- (maine.sum/total)*100
> print(maine.percent)
[1] 75.88085
> away.percent <- (away.sum/total)*100
> print(away.percent)
[1] 24.11915
>
```

Conclusion : It is clear from above Bar Plot that the Maine group is greater than the total runners from the 'Away' group.

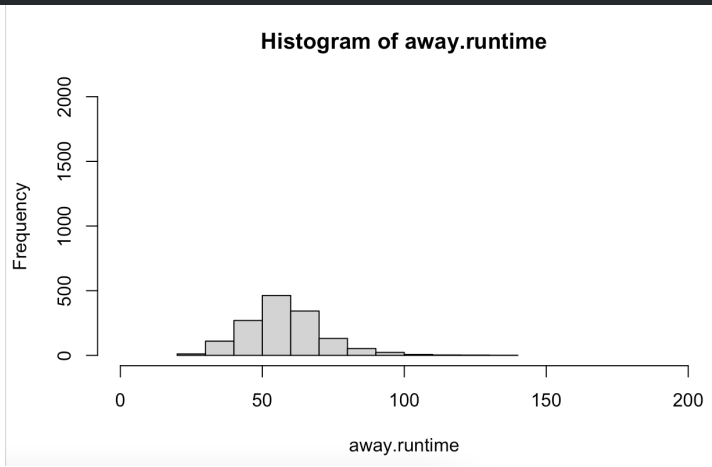
The Maine group holds for 75.8%, while the Away group holds for 24.2%.

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.

```r
#Histogram for Maine group run time

maine.runtime <- data.1$Time..minutes.[which(data.1$Maine=='Maine')]

hist(maine.runtime, xlim = range(0,200), ylim = range(0,2000))
```



```r
#Histogram for Away group run time

away.runtime <- data.1$Time..minutes.[which(data.1$Maine=='Away')]

hist(away.runtime, xlim = range(0,200), ylim = range(0,2000))
```



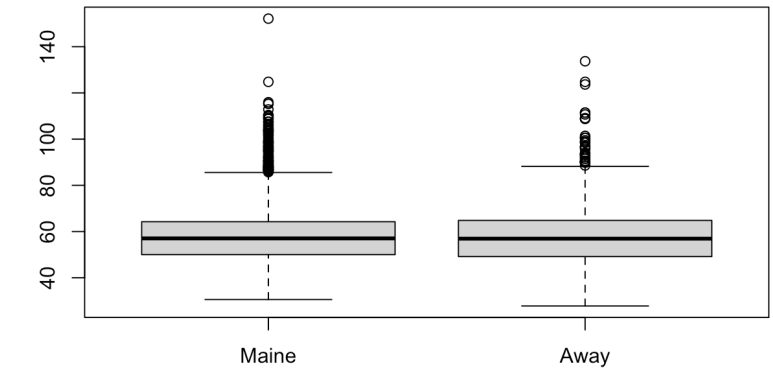
Statistics Summary :

```
> summary(maine.runtime)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.57   50.00   57.03   58.20   64.24  152.17
> summary(away.runtime)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.78   49.15   56.92   57.82   64.83  133.71
> mean(maine.runtime)
[1] 58.19514
> mean(away.runtime)
[1] 57.82181
> range(maine.runtime)
[1]  30.567 152.167
> range(away.runtime)
[1]  27.782 133.710
> sd(maine.runtime)
[1] 12.18511
> sd(away.runtime)
[1] 13.83538
> IQR(maine.runtime)
[1] 14.24775
> IQR(away.runtime)
[1] 15.674
```

“From the above summary we can conclude that both the distributions are skewed towards right as the mean is greater than median in both cases Maine and Away.”

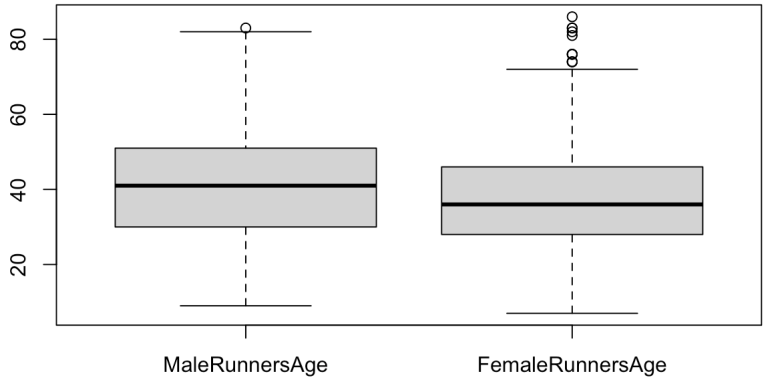
c. Repeat (b) but with side-by-side boxplots.

```
#boxplot

boxplot(maine.runtime, away.runtime, names=c('Maine', 'Away'))
```



d. Create side-by-side boxplots 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.

```
> male.runner.age = strtoi(data.1$Age[which(data.1$Sex=='M')])
> female.runner.age = strtoi(data.1$Age[which(data.1$Sex=='F')])
> boxplot(male.runner.age, female.runner.age, names=c('MaleRunnersAge', 'FemaleRunnersAge'))
```



Statistics Summary

```
> summary(male.runner.age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   30.00   41.00   40.45   51.00   83.00
> mean(male.runner.age)
[1] 40.4468
> range(male.runner.age)
[1]  9 83
> sd(male.runner.age)
[1] 13.99289
> IQR(male.runner.age)
[1] 21
> summary(female.runner.age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   28.00   36.00   37.24   46.00   86.00
> mean(female.runner.age)
[1] 37.23653
> range(female.runner.age)
[1]  7 86
> sd(female.runner.age)
[1] 12.26925
> IQR(female.runner.age)
[1] 18
```

Based on the above stats, males seem to be more actively participating in races, however it can also be seen that the max age of male participants(age = 83) is lower than female max age which is 86.

Q.2. Consider the dataset motorcycle.csv posted on eLearning. It contains the number of fatal motorcycle accidents that occurred in each county of South Car- olina during 2009. Create a boxplot of data and provide relevant summary statistics. Discuss the features of the data distribution. Identify which counties may be con- sidered outliers. Why might these counties have the highest numbers of motorcycle fatalities in South Carolina?

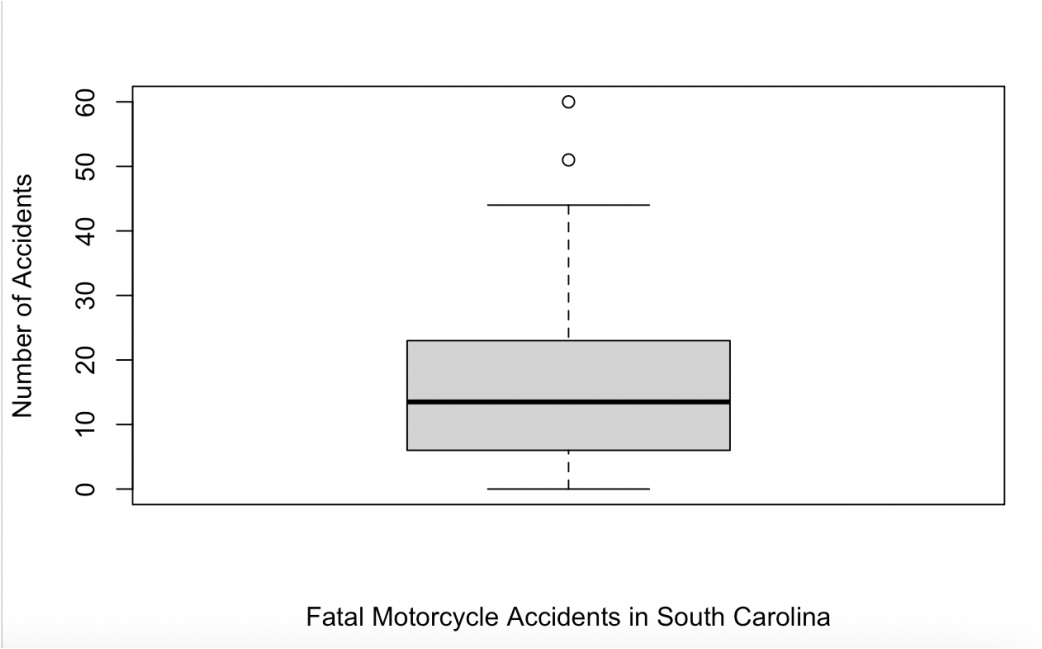
A. Since the mean (17.02) is greater than the median number of accidents (13.5), the distribution is **right skewed**. The standard deviation is 13.5, with an IQR of 17 and range of 60. There are two values which fall above 48.5 1.5IQR+Q3), which makes these two values outliers. The two counties are Horry (60 accidents) and Greenville (51 accidents). One possible reason why these counties are outliers could be poor road

conditions. There could also be lack of patrol on the roads, leading people to drive more recklessly. It could also be that these counties are more densely populated than other counties, so therefore the number of accidents will naturally be more.

summary(motorcyle1)

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

0.00 6.00 13.50 17.02 23.00 60.00



Code Snippet:

```
> motorcycle1 = read.csv("/Users/saratabassi/Desktop/STATS/MiniProj-2/motorcycle.csv")
> FatalAccidents = motorcycle1$Fatal.Motorcycle.Accidents
> boxplot(FatalAccidents, xlab = 'Fatal Motorcycle Accidents in South Carolina', ylab = 'Number of Accidents')
> summary(motorcycle1)
    County          Fatal.Motorcycle.Accidents
 Length:48          Min.   : 0.00
 Class :character   1st Qu.: 6.00
 Mode  :character   Median :13.50
                    Mean   :17.02
                    3rd Qu.:23.00
                    Max.   :60.00
> 
```