Mini Project 2

Name: Adarsh Hegde -AXH190002

Contribution: Question 1-a,b,c

Name: Keerti - KXK190012

Contribution: Question 2 , Question 1-c

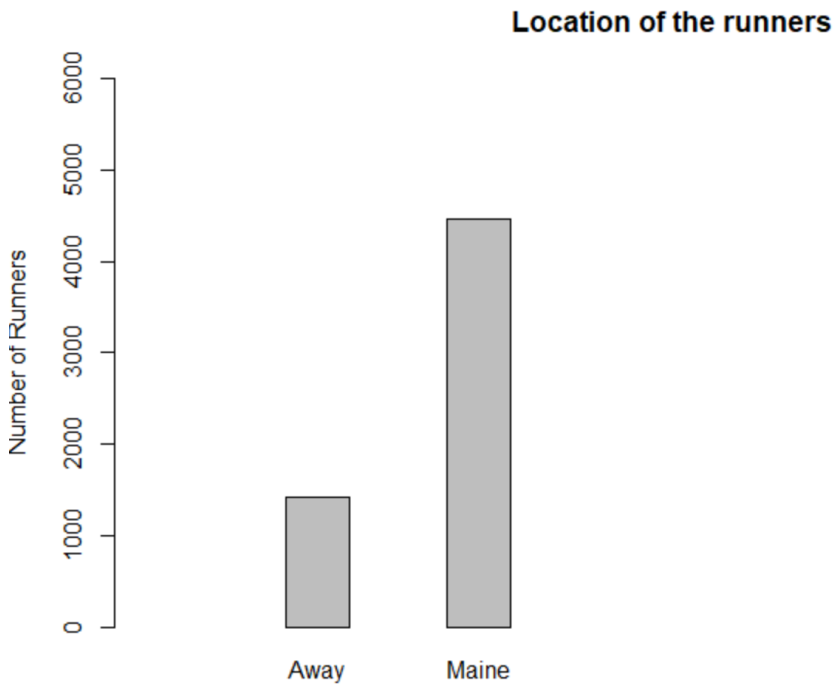
Question 1] Consider the dataset roadrace.csv posted on eLearning. It contains observations on 5875 runners who finished the 2010 Beach to Beacon 10K Road Race in Cape Elizabeth, Maine. You can read the dataset in R using read.csv function.

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 roadrace.csv file
race <- read.csv("roadrace.csv")

## Fetch the column Maine from the csv file
maine_frequency <- table(race$Maine)

## Plot the bar graph
barplot(maine_frequency,main="Where a Runner is from",ylim =
c(0,6000),xlim=c(0,16),ylab = "Number of Runners",space=c(0,0.25))
```



The graph shows that runners from the graph ~ 4800 & Away ~ 1500. From the R calculations, we can see that the value populated at Maine = 4458 and Away = 1417. Thus, we can conclude that the graph estimations are right.

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.

##Categorizing the runners based on Maine and Away in Maine Attribute

```
maine_runners <- race[which(race$Maine=="Maine"), names(race) %in% c("Time..minutes.")]

away_runners <- race[which(race$Maine=="Away"), names(race) %in% c("Time..minutes.")]
```

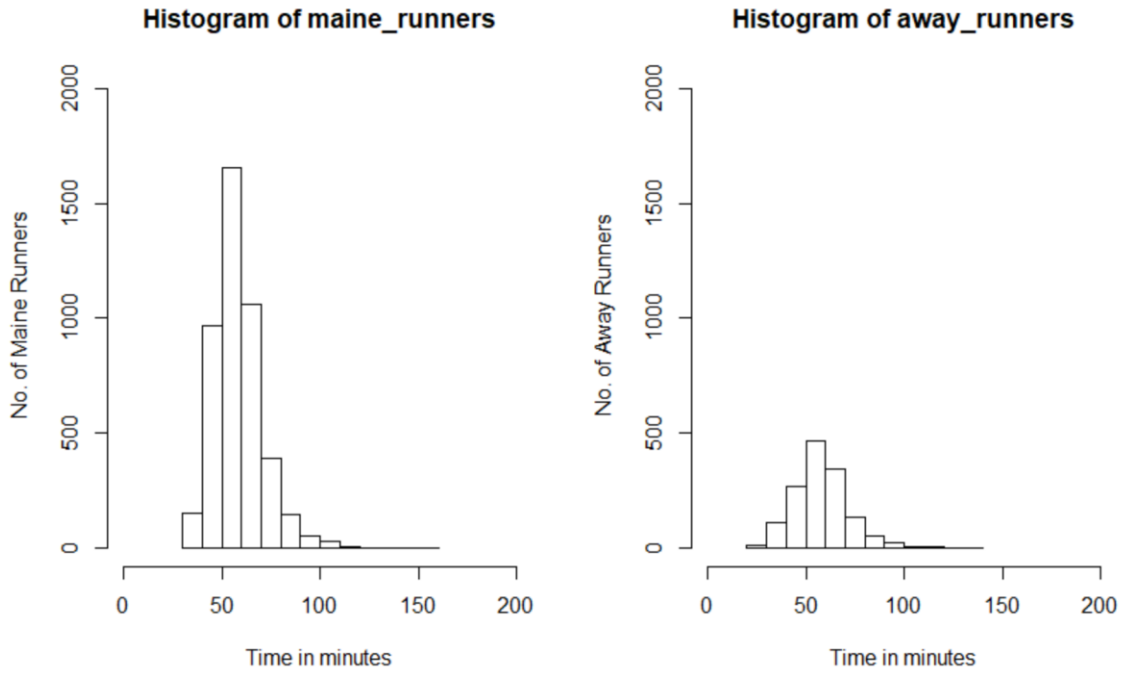
Plot 2 graphs in 1 row

```
par(mfrow=c(1,2))
```

Plotting histograms for both the variables using same limit for axes and apt labels

hist(maine_runners, xlim = c(0,200), ylim = c(0,2000), ylab = " No. of Maine Runners", xlab = "Time in minutes", breaks = 10)

hist(away_runners, xlim = c(0,200), ylim = c(0,2000), ylab = " No. of Away Runners", xlab = "Time in minutes", breaks = 10)



The above graphs suggest that the both the graphs follow same distribution, almost normal but slightly right skewed.

Summary:

Maine :

> mean(maine_runners)

[1] 58.19514

> median(maine_runners)

[1] 57.0335

> sd(maine_runners)

[1] 12.18511

> range(maine_runners)

[1] 30.567 152.167

> IQR(maine_runners)

[1] 14.24775

Away:

> mean(away_runners)

[1] 57.82181

> median(away_runners)

[1] 56.92

> sd(away_runners)

[1] 13.83538

> range(away_runners)

[1] 27.782 133.710

> IQR(away_runners)

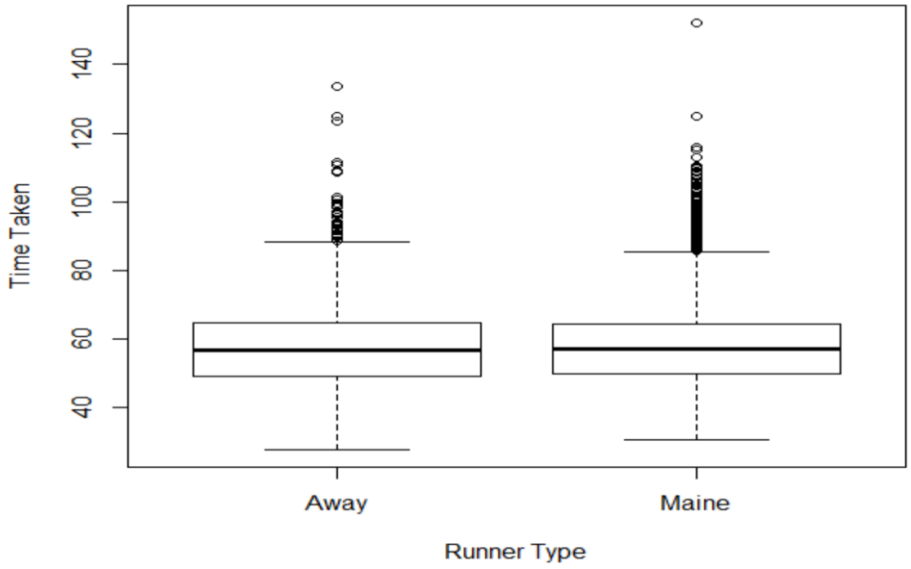
[1] 15.674

From the above calculations, we can conclude that the initial graph estimations were right about being almost normal yet right skewed.

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

Plotting both Maine and Away side by side.

boxplot(race$Time..minutes. ~ race$Maine,xlab="Runner Type", ylab="Time Taken")



Estimated summary values:

Maine: The estimated set of 5 values for Maine is (30,50,60, 65,150). Away: The estimated set of 5 values for Away is (25,50,60,65,130).

Calculated summary values:

Maine:

```
> summary(maine_runners)
   Min.  1st Qu.  Median    Mean   3rd Qu.  Max.
  30.57   50.00   57.03   58.20   64.24  152.17
```

Away:

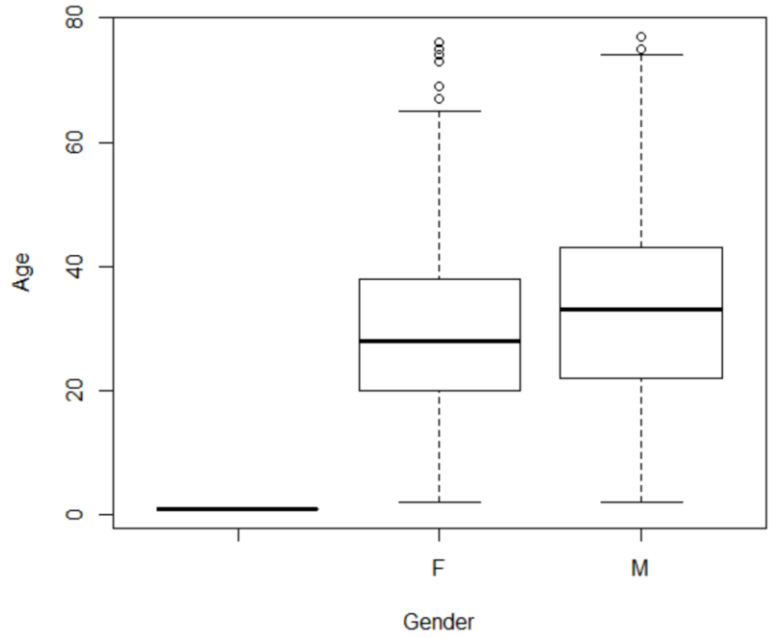
```
> summary(away_runners)
   Min.  1st Qu.  Median    Mean   3rd Qu.  Max.
  27.78   49.15   56.92   57.82   64.83  133.71
```

As the estimated values are very close to the calculated values, the graph is right.

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.

##Creating boxplot of using the column Sex which categorizes based on Age.

boxplot(as.numeric(race$Age) ~ race$Sex, xlab = "Age",ylab = "Sex")



Since we have a random '*' in column Sex, we have an additional plot for that.

Plotting the graph side by side

par(mfcol=c(1,2))

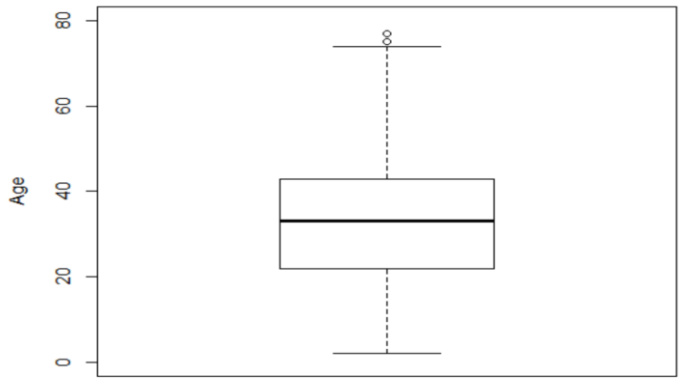
Creating M and F for male and female runners:

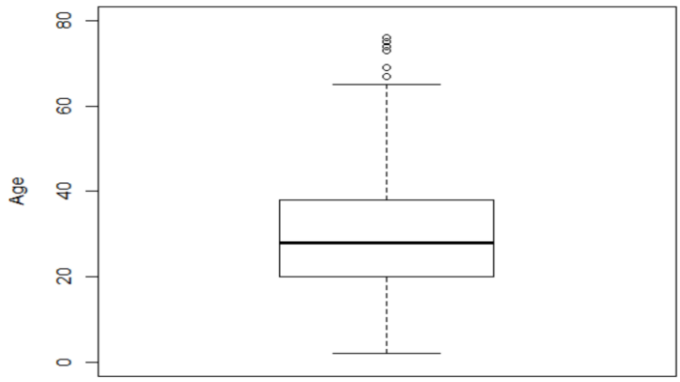
>Male <- race[which(race$Sex=="M"), names(race) %in% c("Age")];

>FeMale <- race[which(race$Sex=="F"), names(race) %in% c("Age")];

> boxplot(as.numeric(Male), xlab = 'Male', ylab = 'Age', ylim = c (0,80));

> boxplot(as.numeric(FeMale), xlab = 'Female', ylab = 'Age', ylim = c(0,80));





For Male:

Estimated values: The set of 5 values estimated from the graph for Males are (3,21,32,42,76). IQR = 42 – 21 = 21

For Female:

Estimated values: The set of 5 values estimated from the graph for Males are (3,20,30,39,75). IQR = 39 – 20 = 19

Calculated Values:

Male:

> mean(as.numeric(Male))

[1] 32.56312

> median(as.numeric(Male))

[1] 33

> sd(as.numeric(Male))

[1] 14.07031

> IQR(as.numeric(Male))

[1] 21

> range(as.numeric(Male))

[1] 2 77

Female:

> mean(as.numeric(FeMale))

[1] 29.26296

> median(as.numeric(FeMale))

[1] 28

> sd(as.numeric(FeMale))

[1] 12.28545

> IQR(as.numeric(FeMale))

[1] 18

> range(as.numeric(FeMale))

[1] 2 76

Thus, we can conclude that the estimation is right based on the exact calculations.

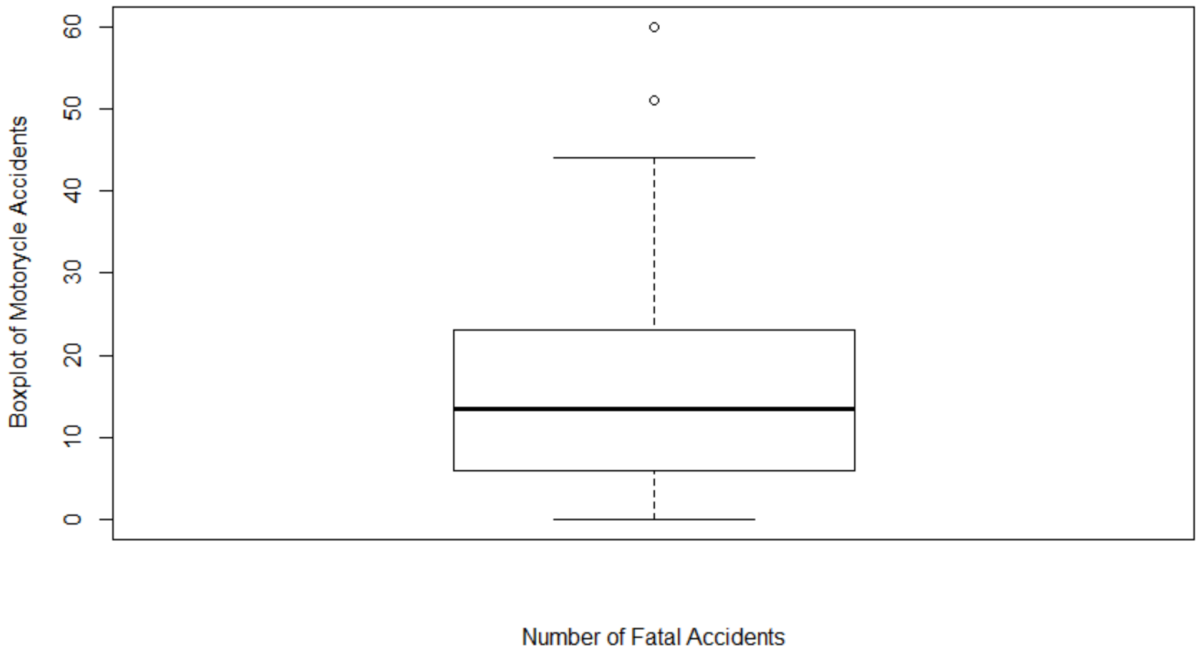
Question 2] Consider the dataset motorcycle.csv posted on eLearning. It contains the number of fatal motorcycle accidents that occurred in each county of South Carolina during 2009. Create a boxplot of data and provide relevant summary statistics. Discuss the features of the data distribution. Identify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

Read the csv file into acc

>acc <- read.csv("motorcycle.csv")

Plot the boxplot of the fatal accidents

> boxplot(acc$Fatal.Motorcycle.Accidents,xlab = "Number of Fatal Accidents",ylab = "Boxplot of Motorycle Accidents")



Estimated Values: The 5-number summary of the data: (0,5,12,23,60)

IQR = 23-5 = 18

Calculated Values: summary(acc$Fatal.Motorcycle.Accidents)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	6.00	13.50	17.02	23.00	60.00

By comparing the actual and estimated values, we can conclude that the values are accurate.

From the above boxplot, we can estimate that the values above the whisker 45 are considered as outliers.

Calculating the Outliers

```
outlier_counties <- acc[which(acc$Fatal.Motorcycle.Accidents  > 45),names(acc)%in%c("County")]
```

outlier_counties

```
[1] GREENVILLE HORRY
```

This concludes that Greenville and Horry counties are outliers in the data set and on verifying, we see that they indeed have the accidents of 51 and 60 which in fact are outliers.