Statistical Methods for Data Science
Mini-Project #2

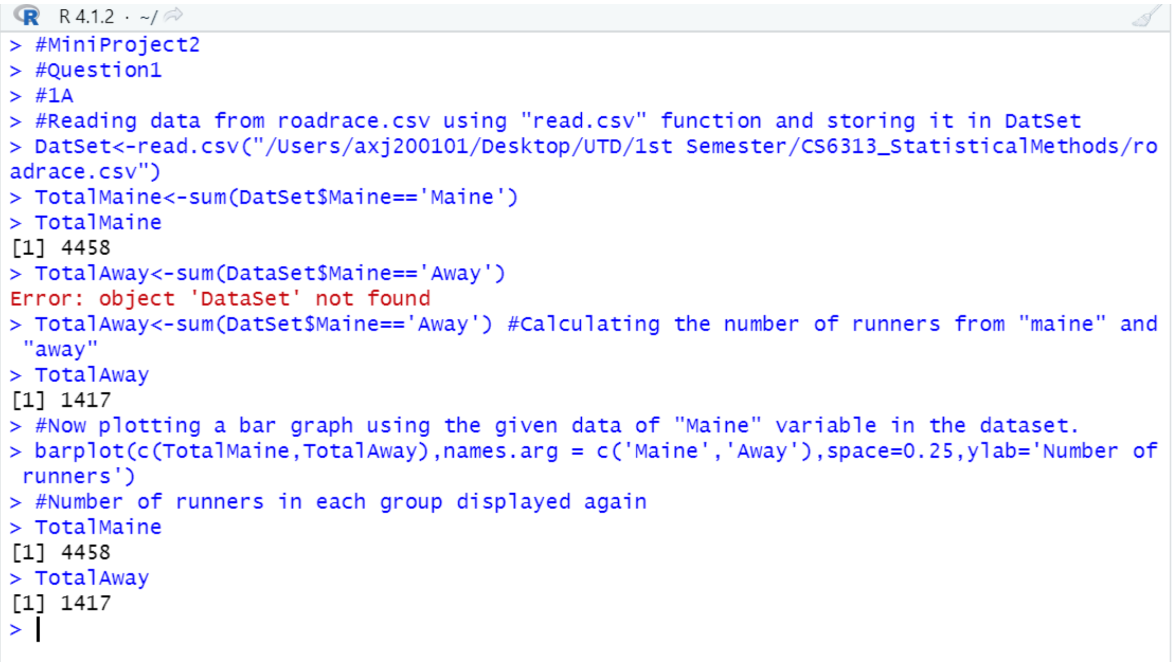
Group-46
Anish Joshi(UTD ID-2021591978), Aneena Manoj(UTD ID-2021623362)

Contributions – The mini-project was discussed in detail initially by both of us. Further, Anish and Aneena studied, discussed, and wrote the R code/script for both the questions and came up with the output, and wrote the report together. Both the partners worked efficiently to complete the mini-project.

Question 1

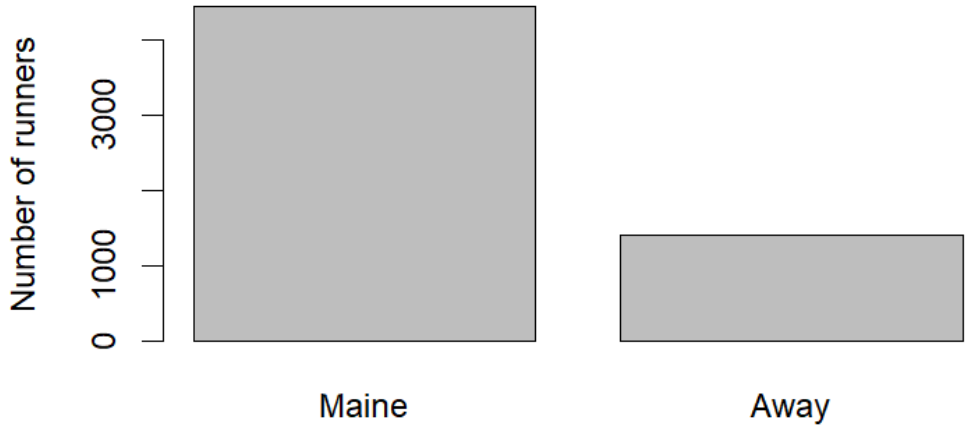
Starting the project by downloading the "roadrace.csv" file and reading through the data, and making note of the column headings in the dataset. The next step is to read the file into R using the function "read.csv" as mentioned in the question.

1A]
In this question, a bar graph is to be plotted using the given conditions. We will start by initializing variables and storing the required data into them. In the "read.csv" function, we need to specify the original pathway of where the file is present in the system. The function to plot a bar graph is barplot().
Using R,
The total number of runners belonging to the 'maine' group=4458,
The total number of runners belonging to the 'away' group=1417.
Code –



```
> #MiniProject2
> #Question1
> #1A
> #Reading data from roadrace.csv using "read.csv" function and storing it in DatSet
> DatSet<-read.csv("/Users/axj200101/Desktop/UTD/1st Semester/CS6313_StatisticalMethods/ro
adrace.csv")
> TotalMaine<-sum(DatSet$Maine=='Maine')
> TotalMaine
[1] 4458
> TotalAway<-sum(DataSet$Maine=='Away')
Error: object 'DataSet' not found
> TotalAway<-sum(DatSet$Maine=='Away') #Calculating the number of runners from "maine" and
 "away"
> TotalAway
[1] 1417
> #Now plotting a bar graph using the given data of "Maine" variable in the dataset.
> barplot(c(TotalMaine,TotalAway),names.arg = c('Maine','Away'),space=0.25,ylab='Number of
 runners')
> #Number of runners in each group displayed again
> TotalMaine
[1] 4458
> TotalAway
[1] 1417
> |
```

Plotted Bar Graph –



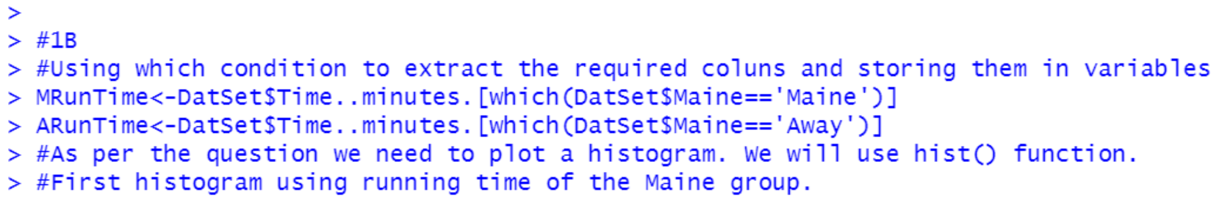
There are two conclusions that can be drawn from the plotted bar graph –

1. The number of runners in the Maine group(4458) is greater than the number of runners in the Away group(1417).
2. Out of 5875 runners, 75.88% of the total runners belong to the Maine group and the remaining 24.12% belong to the Away group.

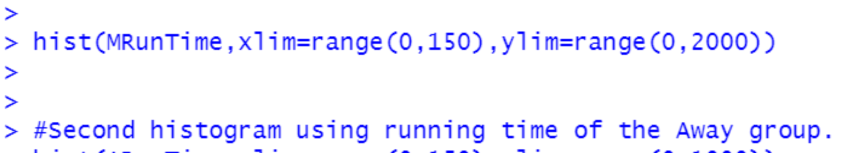
1B]

To solve this particular problem, first we have to extract the required columns into variables. To do the following, we need to use the “which” condition to extract the running time of both the groups into the variables. The variable name to denote the running time of the Maine group and Away group is MRunTime and ARunTime respectively.

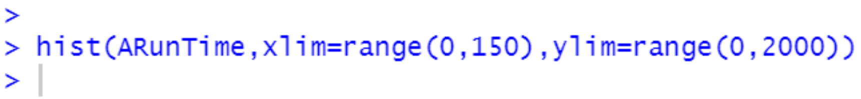
Code –



```
>
> #1B
> #Using which condition to extract the required coluns and storing them in variables
> MRunTime<-DatSet$Time..minutes.[which(DatSet$Maine=='Maine')]
> ARunTime<-DatSet$Time..minutes.[which(DatSet$Maine=='Away')]
> #As per the question we need to plot a histogram. We will use hist() function.
> #First histogram using running time of the Maine group.
```

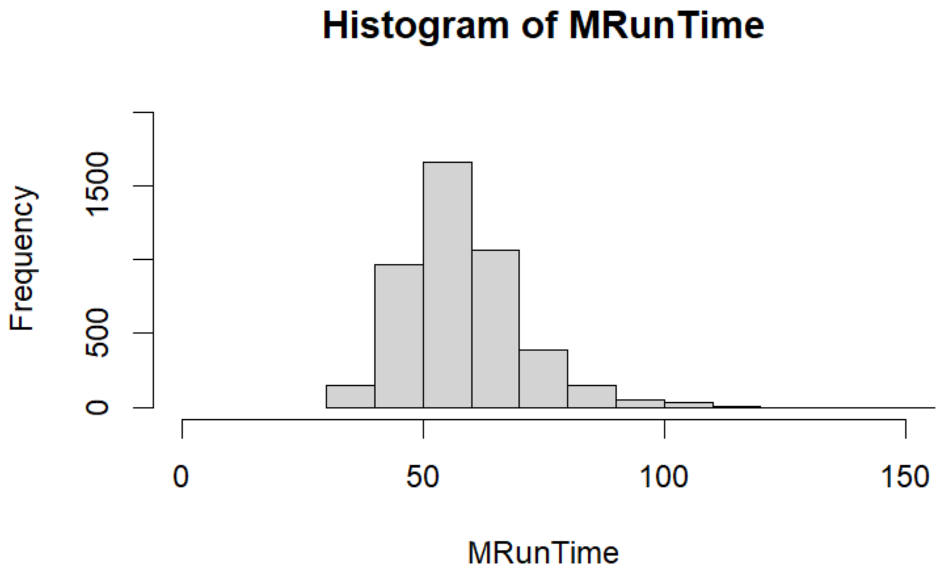


```
>
> hist(MRunTime,xlim=range(0,150),ylim=range(0,2000))
>
>
> #Second histogram using running time of the Away group.
```

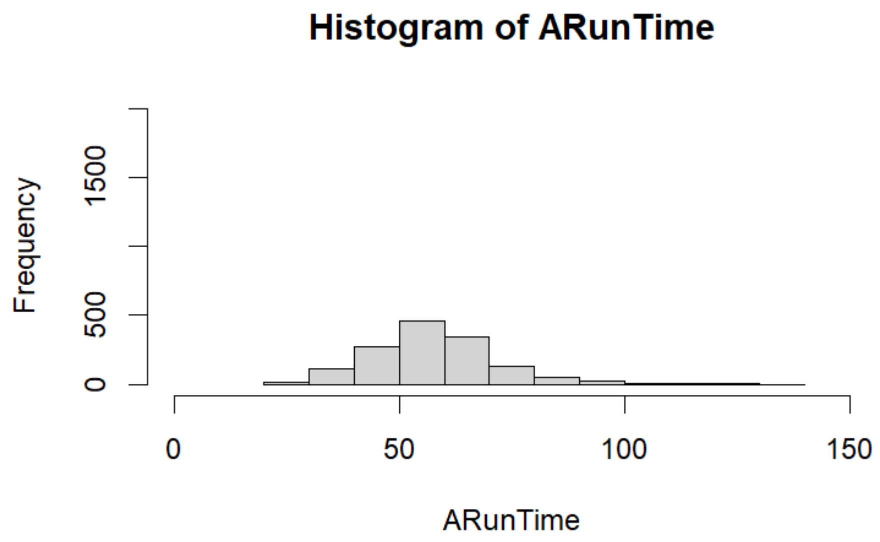


```
>
> hist(ARunTime,xlim=range(0,150),ylim=range(0,2000))
> 
```

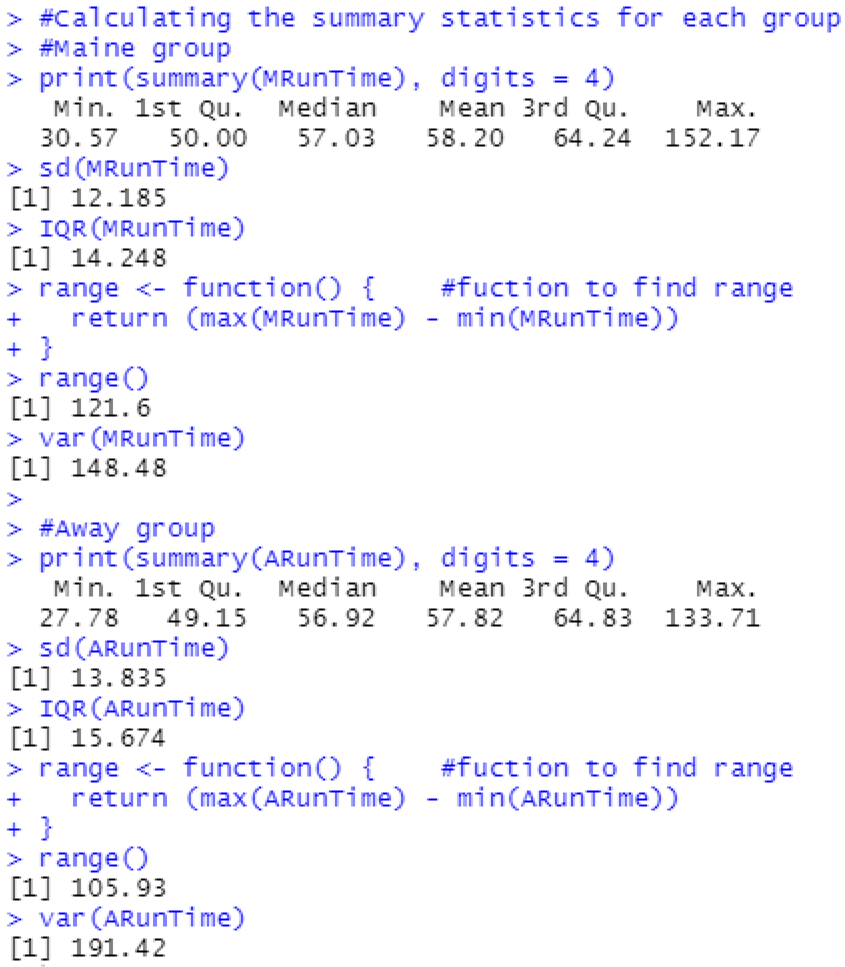
Histogram for Running time of Maine group –



Histogram for Running time of Away group –



Now the next part of the question tells us to find the summary statistics including mean, standard deviation, range, median and interquartile range.
Code –



```
> #Calculating the summary statistics for each group
> #Maine group
> print(summary(MRunTime), digits = 4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.57   50.00   57.03   58.20   64.24  152.17
> sd(MRunTime)
[1] 12.185
> IQR(MRunTime)
[1] 14.248
> range <- function() {    #fuction to find range
+   return (max(MRunTime) - min(MRunTime))
+ }
> range()
[1] 121.6
> var(MRunTime)
[1] 148.48
>
> #Away group
> print(summary(ARunTime), digits = 4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.78   49.15   56.92   57.82   64.83  133.71
> sd(ARunTime)
[1] 13.835
> IQR(ARunTime)
[1] 15.674
> range <- function() {    #fuction to find range
+   return (max(ARunTime) - min(ARunTime))
+ }
> range()
[1] 105.93
> var(ARunTime)
[1] 191.42
```

Table depicting all the relevant summary statistics –

Groups	Min	1st Quar.	Median	Mean	3rd Quar.	Max	SD	range	IQR	Var
Maine	30.57	50.00	57.03	58.20	64.24	152.17	12.185	121.6	14.248	148.48
Away	27.78	49.15	56.92	57.82	64.83	133.71	13.835	105.93	15.674	191.42

From the above table, it can be observed that the values of **Min, 1st Quar., Median, Mean and Max** is higher for the "Maine" group whereas the values of **3rd Quar., SD, IQR, and Variance** is higher for the "Away" group.

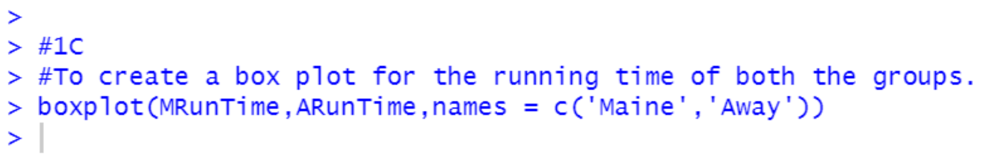
We can see from the Maine Group histogram plot that the distribution is **Right-Skewed** distribution as its peak is slightly off-center and the tail stretches away to the right. The shape of the histogram is asymmetrical and hence skewed. Also, since the mean is greater than the median, the histograms are right-skewed.

The time taken for Away runners has a better symmetry than the time taken for Maine runners but this also is not perfectly symmetrical and the tail stretches a bit to the right along with the peak being slightly off-centered. The mean is slightly greater than the median. Hence, the distribution is slightly **right-skewed** distribution.

1C]

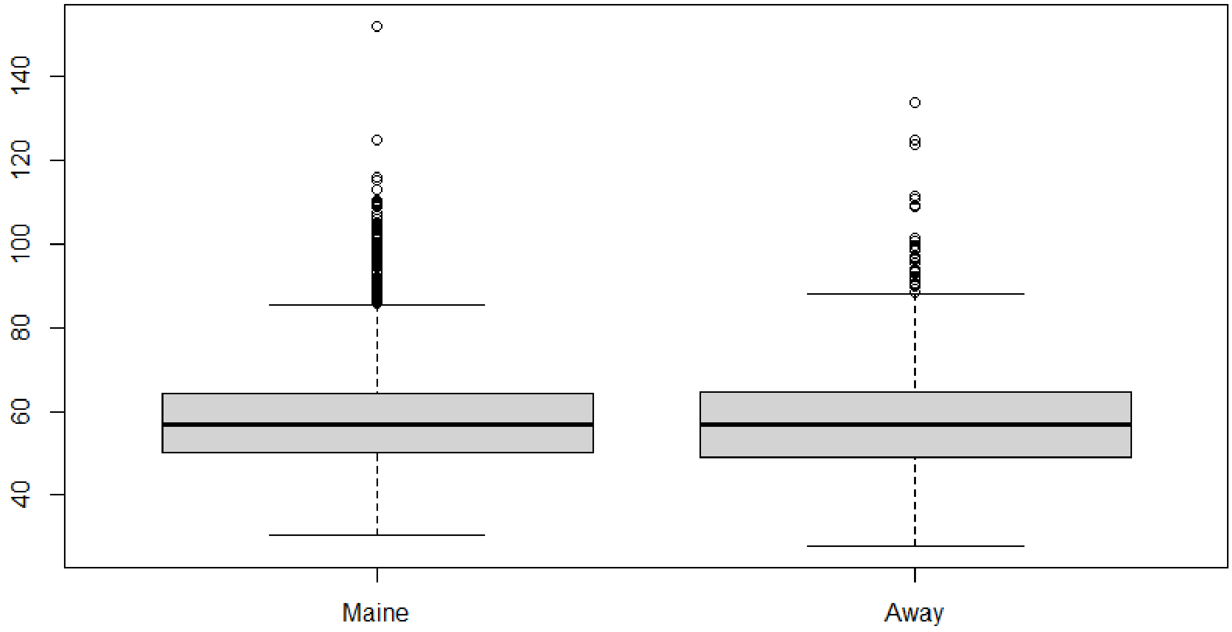
In this question, a box plot is to be created for the running times of both groups. boxplot() is the function in R to create boxplots.

Code –



```
>
> #1C
> #To create a box plot for the running time of both the groups.
> boxplot(MRunTime,ARunTime,names = c('Maine','Away'))
>
```

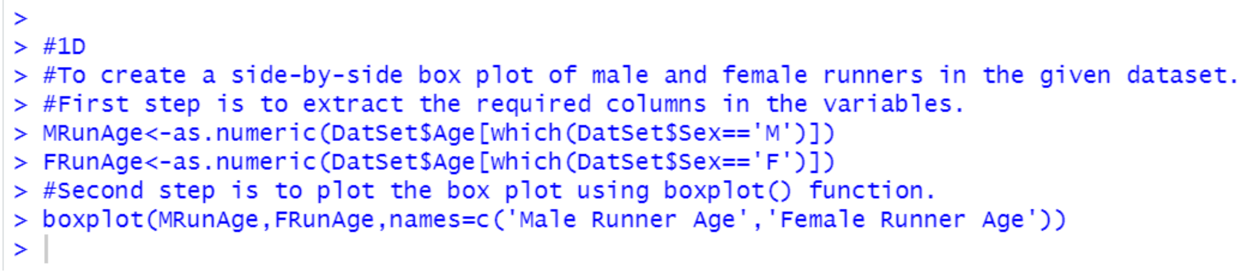
Boxplot –



1D]

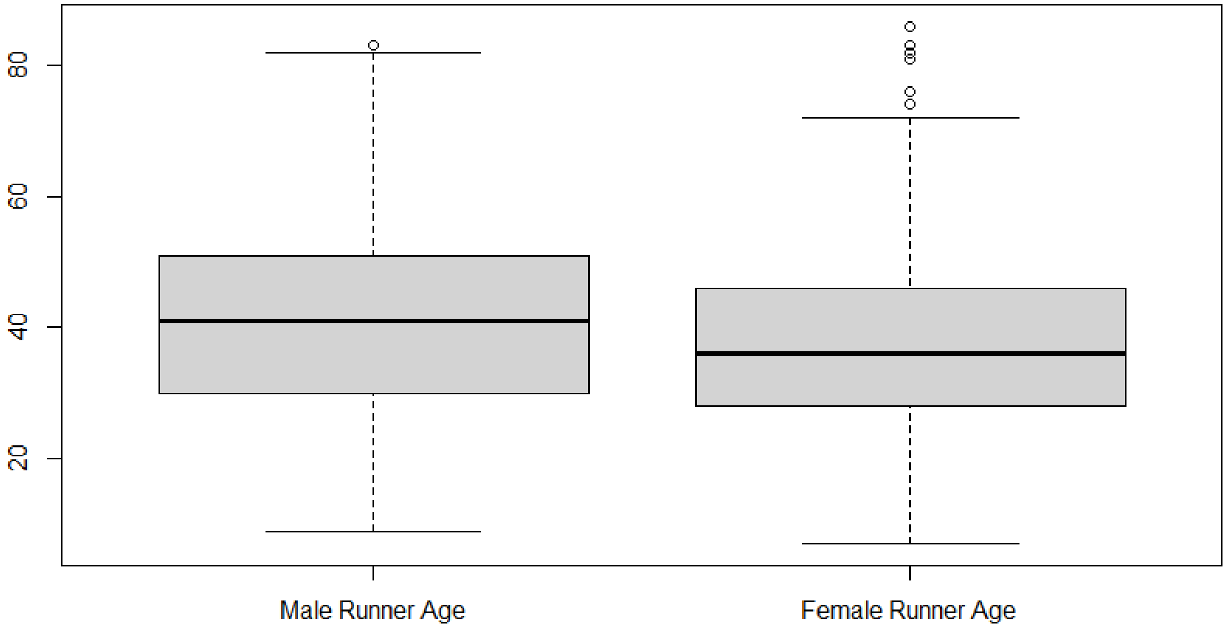
To solve this question, first we need to extract the required columns in variables. The side-by-side boxplot to be created is going to be based on the gender of the runners in the dataset. The variable name is taken as MRunAge for the male runners and FRunAge for the female runners.

Code –



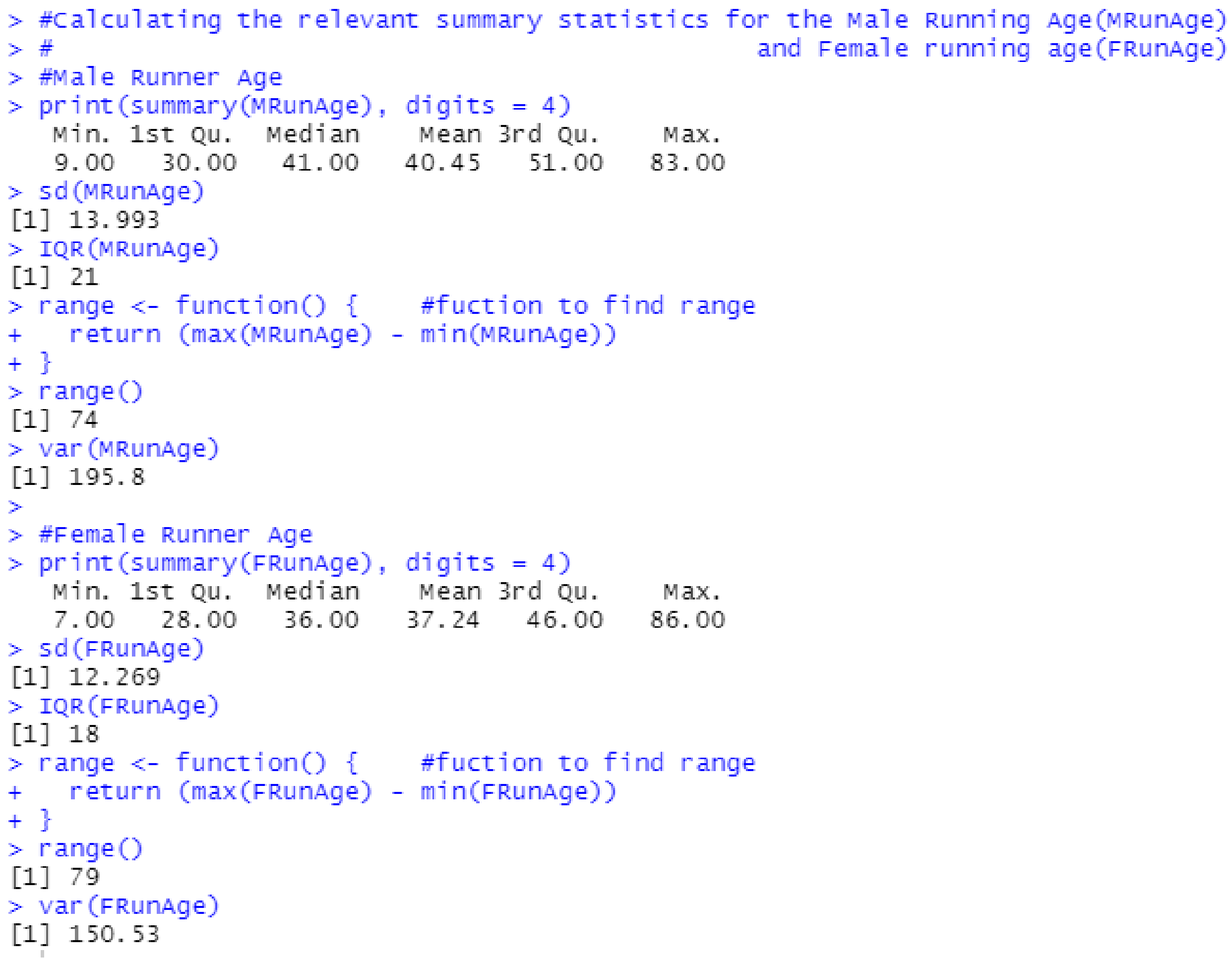
```
>
> #1D
> #To create a side-by-side box plot of male and female runners in the given dataset.
> #First step is to extract the required columns in the variables.
> MRunAge<-as.numeric(DatSet$Age[which(DatSet$Sex=='M')])
> FRunAge<-as.numeric(DatSet$Age[which(DatSet$Sex=='F')])
> #Second step is to plot the box plot using boxplot() function.
> boxplot(MRunAge,FRunAge,names=c('Male Runner Age','Female Runner Age'))
> 
```

Boxplot –



In the second part of the question, we need to find the summary statistics of the male running ages and female running ages.

Code –



```
> #Calculating the relevant summary statistics for the Male Running Age(MRunAge)
> #                                              and Female running age(FRunAge)
> #Male Runner Age
> print(summary(MRunAge), digits = 4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   9.00   30.00   41.00   40.45   51.00   83.00 
> sd(MRunAge)
[1] 13.993
> IQR(MRunAge)
[1] 21
> range <- function() {    #fuction to find range
+   return (max(MRunAge) - min(MRunAge))
+ }
> range()
[1] 74
> var(MRunAge)
[1] 195.8
> 
> #Female Runner Age
> print(summary(FRunAge), digits = 4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   7.00   28.00   36.00   37.24   46.00   86.00 
> sd(FRunAge)
[1] 12.269
> IQR(FRunAge)
[1] 18
> range <- function() {    #fuction to find range
+   return (max(FRunAge) - min(FRunAge))
+ }
> range()
[1] 79
> var(FRunAge)
[1] 150.53
```

Table depicting relevant summary statistics –

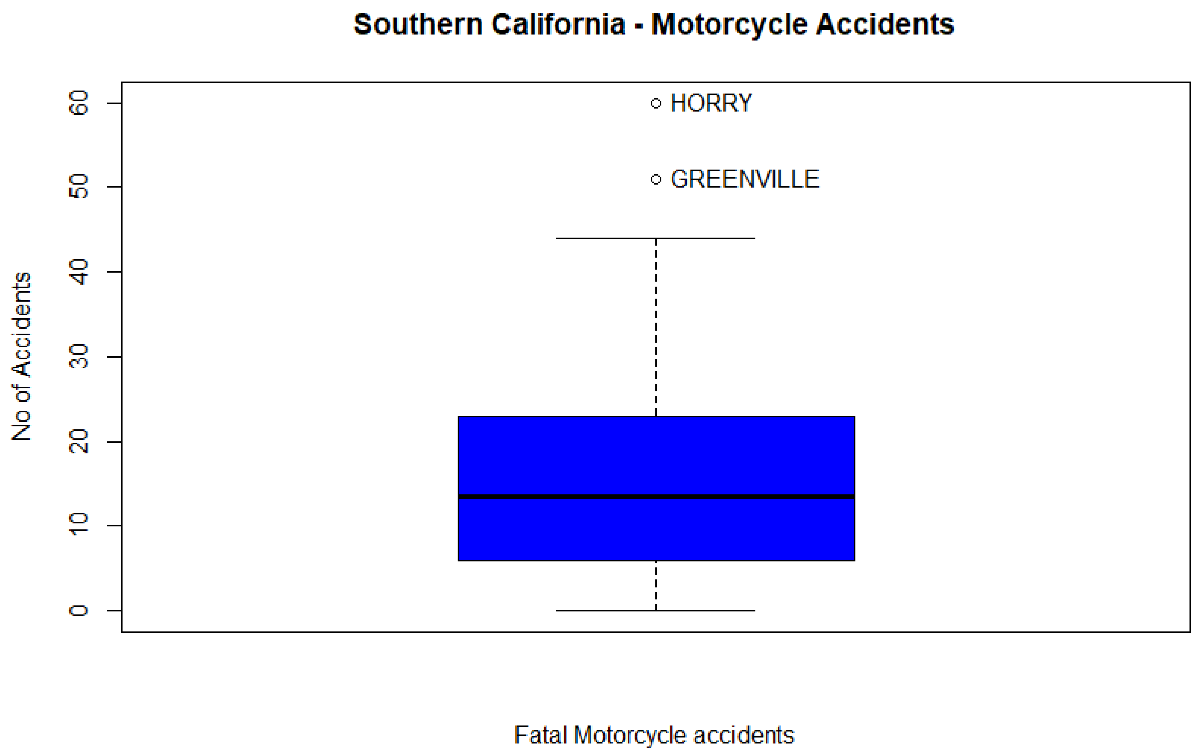
Gender	Min	1st Quar.	Median	Mean	3rd Quar.	Max	SD	range	IQR	Var
Male	9.00	30.00	41.00	40.45	51.00	83.00	13.993	74	21	195.80
Female	7.00	28.00	36.00	37.24	46.00	86.00	12.269	79	18	150.53

From the summary, we can conclude that Male age groups have a higher value for mean, median, and quartiles. The range of the female group is more than the male group because of the outlier. We can see multiple outliers in Female age graphs which I believe is impacting the value of the mean. From the two distributions, we can conclude that the male participants are on average older than the female participants with a few outliers on the female side. The youngest and the oldest participant of the race are both females.

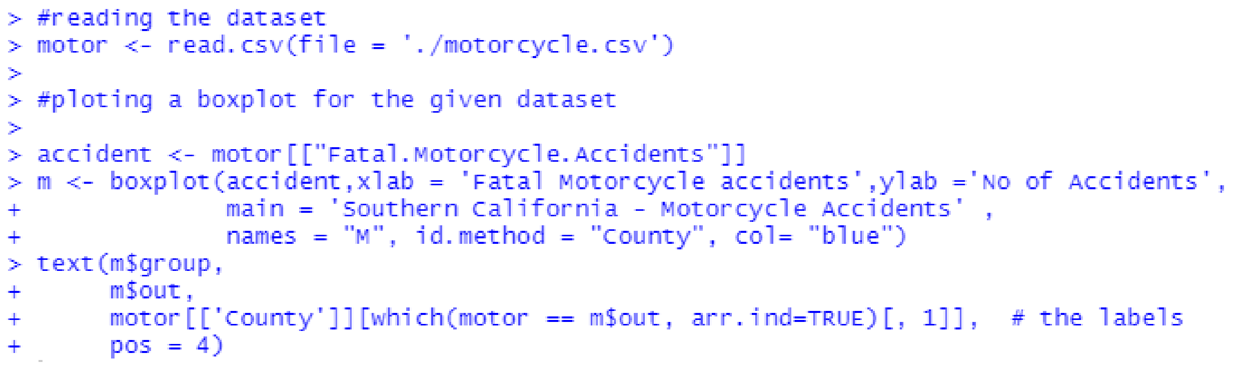
Question 2:

The dataset motorcycle.csv contains the number of fatal motorcycle accidents that occurred in each county of South Carolina during 2009. The plots for the accident by each county are as follows:

Creating Boxplot for Fatal Motorcycle Accidents

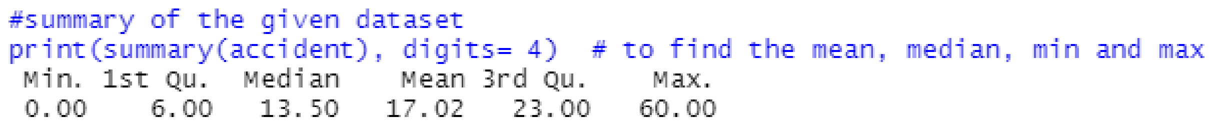


Code Snippet



```
> #reading the dataset
> motor <- read.csv(file = './motorcycle.csv')
>
> #ploting a boxplot for the given dataset
>
> accident <- motor[["Fatal.Motorcycle.Accidents"]]
> m <- boxplot(accident,xlab = 'Fatal Motorcycle accidents',ylab ='No of Accidents',
+              main = 'Southern California - Motorcycle Accidents' ,
+              names = "M", id.method = "County", col= "blue")
> text(m$group,
+      m$out,
+      motor[['County']][which(motor == m$out, arr.ind=TRUE)[, 1]],  # the labels
+      pos = 4)
```

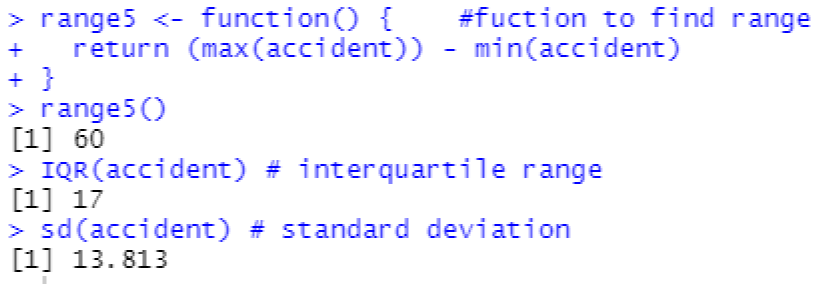
Tabular Representation of Summary Statistics:



```
#summary of the given dataset
print(summary(accident), digits= 4)  # to find the mean, median, min and max
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00    6.00   13.50   17.02   23.00   60.00
```

The distribution of fatal motorcycle accidents in each country of South Carolina during 2009 has a mean of 17.02 with a standard deviation of 13.813. The mean is higher than the median, as implied by the right-skewness of the distribution.
It has a median of 13.50. The first quartile is 6.00, and the 3rd quartile is 23.00. The lowest number of fatal motorcycle accidents in a county is 0, and the highest number of fatal motorcycle accidents in a county is 60.

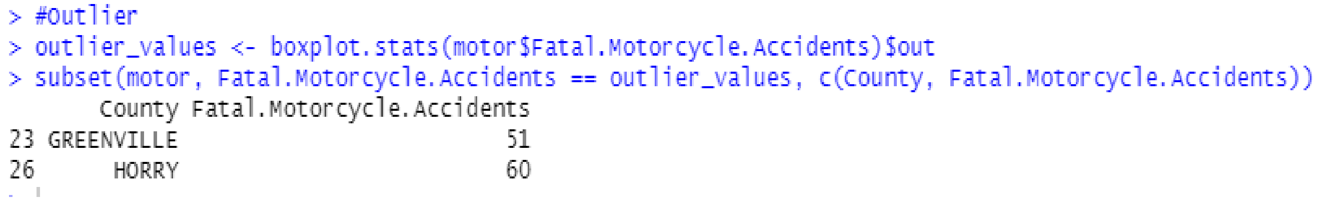
The data clearly has a **right-skewed** distribution since the mean (17.02) is higher than the median(13.5).



```
> range5 <- function() {    #fuction to find range
+   return (max(accident)) - min(accident)
+ }
> range5()
[1] 60
> IQR(accident) # interquartile range
[1] 17
> sd(accident) # standard deviation
[1] 13.813
```

	Min	1st Quar.	Median	Mean	3rd Quar.	Max	SD	range	IQR
Accidents	0.00	6.00	13.50	17.02	23.00	60.00	13.813	60	17

Outliers:



```
> #Outlier
> outlier_values <- boxplot.stats(motor$Fatal.Motorcycle.Accidents)$out
> subset(motor, Fatal.Motorcycle.Accidents == outlier_values, c(County, Fatal.Motorcycle.Accidents))
       County Fatal.Motorcycle.Accidents
23 GREENVILLE                         51
26      HORRY                         60
```

From the dataset and boxplot, it is clear that the two counties that are considered outliers are Greenville with 51 accidents and Horry with 60 accidents.
Since the data doesn’t provide the factors for the high fatality rate, we can’t assume for sure what was the reason behind these increased numbers of accidents.

These countries might have the highest no of fatalities rate due to:

1. Population
2. Type of terrain
3. Accuracy of reports
4. Rash driving by negligent drivers
5. Improper construction of roads or poor maintenance of the roads by the authorities.