CS 6313.001

Statistical Methods for Data Science

By

Prof. Min Chen

MINI PROJECT-2

DUO GROUP-41

1.Amit Kumar	AXK210047
2.Vedant Paresh Shah	VXS200021

Contribution of each group member:

Both worked together and finished the questions as instructed. First went through all the details required, followed the class Note and textbook, learned R then wrote down the scripts. Both of us worked efficiently to complete the required project and finished it on time.

Question #1:

(12 points) 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.

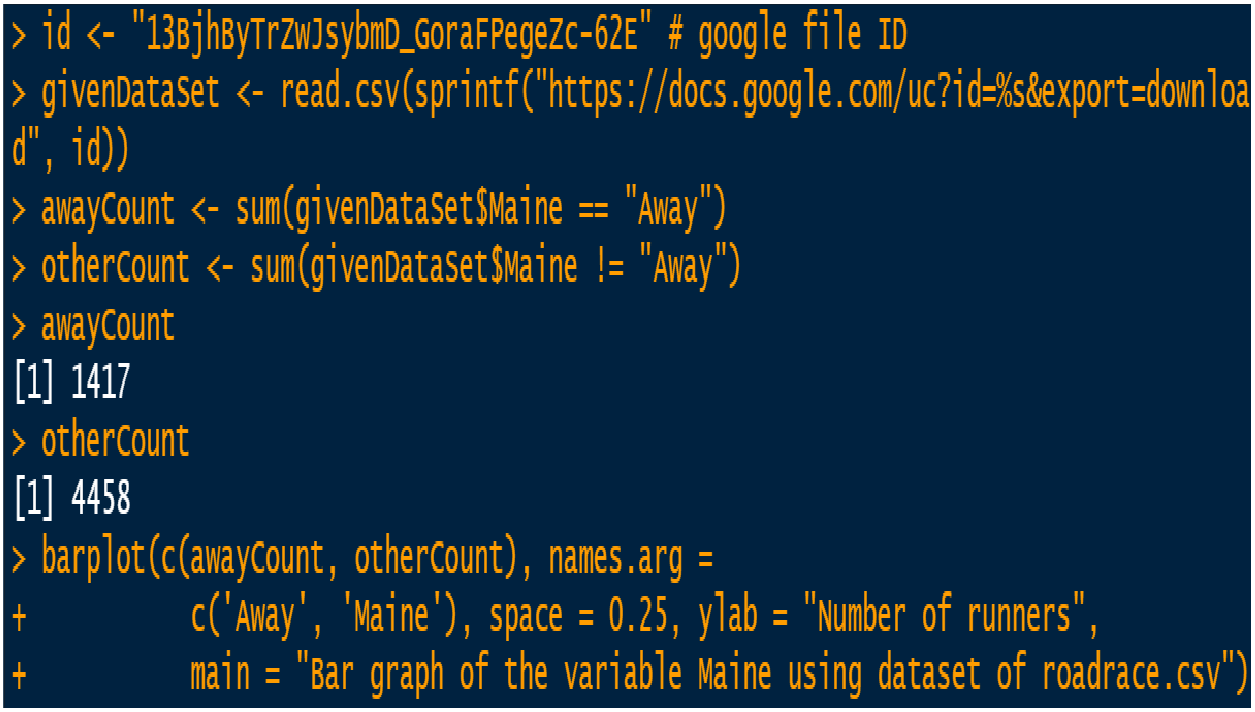
Answer:

To create a bar plot of Maine and Away we need to extract the data from the given dataset roadrace.csv file. Then we need to find which runner are from Maine and Away from Maine. After achieving this data we can create a bar plot.

Code Snippet:

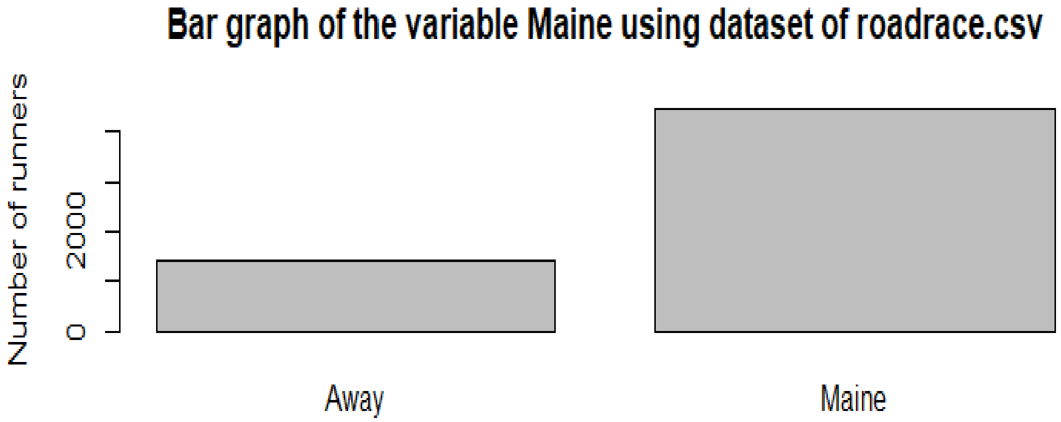
```
> id <- "13BjhByTrZwJsybmD_GoraFPegeZc-62E" # Google Drive file ID

>givenDataSet<-
read.csv(sprintf("https://docs.google.com/ucid=%s&export=download",id))

> awayCount <- sum(givenDataSet$Maine == "Away")

> otherCount <- sum(givenDataSet$Maine != "Away")

> awayCount

> otherCount

>barplot(c(awayCount,otherCount), names.arg=c('Away','Maine'),space= 0.25,

  ylab = "Number of runners",main = "Bar graph of the variable Maine using dataset
of roadrace.csv")
```

Output:



```
> id <- "13BjhByTrZwJsybmD_GoraFPegeZc-62E" # google file ID
> givenDataSet <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=downloa
d", id))
> awayCount <- sum(givenDataSet$Maine == "Away")
> otherCount <- sum(givenDataSet$Maine != "Away")
> awayCount
[1] 1417
> otherCount
[1] 4458
> barplot(c(awayCount, otherCount), names.arg =
+           c('Away', 'Maine'), space = 0.25, ylab = "Number of runners",
+           main = "Bar graph of the variable Maine using dataset of roadrace.csv")
```

Bar plot:



Maine Runners = 4458

Away Runners = 1417

Based on the graph, it can be concluded that Maine group runner data is greater than Away group runner data. Maine group count is 4458 and Away group runner data is 1417, if we find the probability of Away group runner data is 24.11%, it is also clearly conferred from the bar graph.

(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.

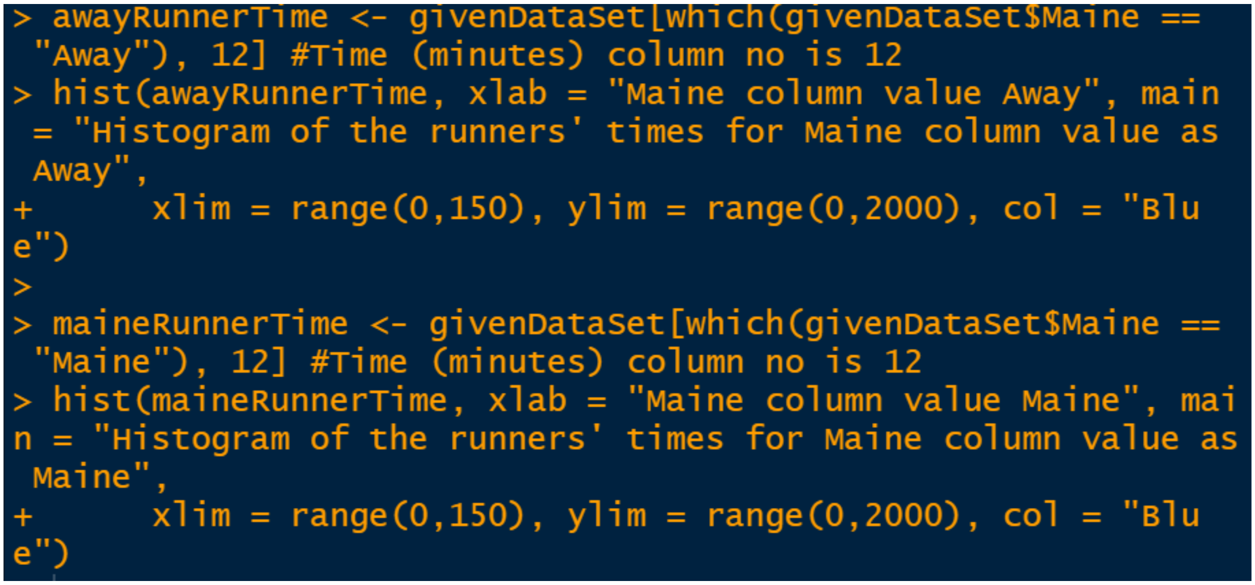
Answer:

To create histogram of runner time from Maine and Away we use data from Maine column and Time column from dataset roadrace.csv

Code Snippet:

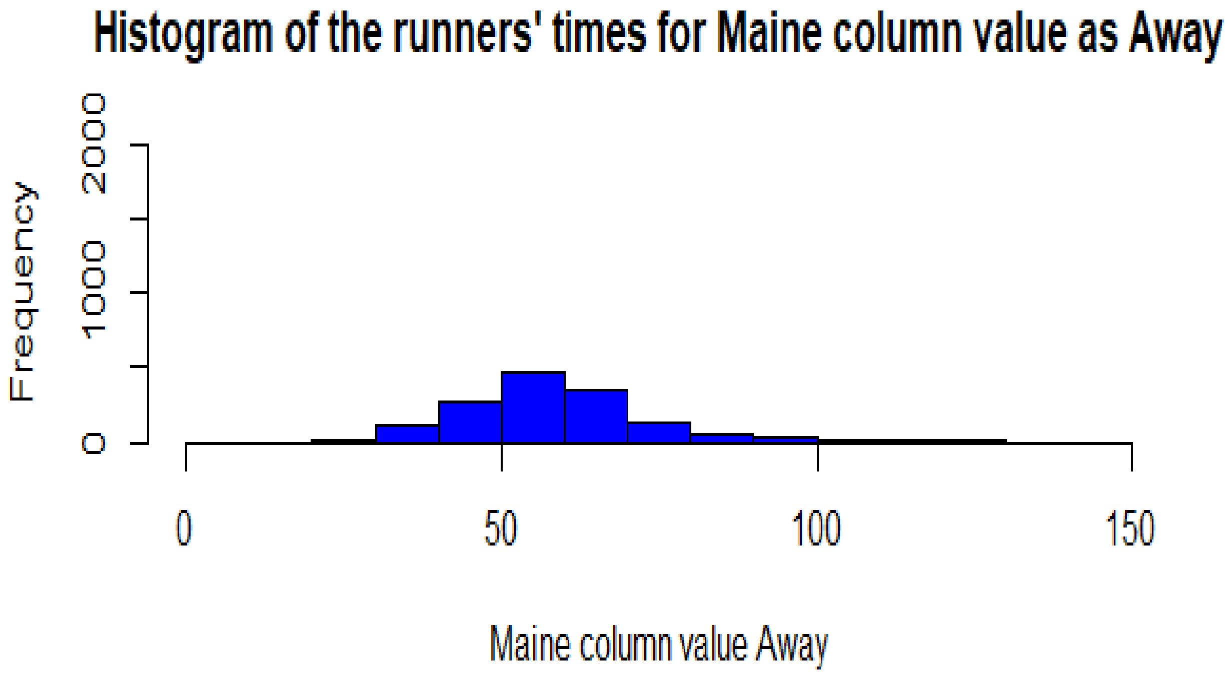
```
#Time (minutes) column no is 12

> awayRunnerTime <- givenDataSet[which(givenDataSet$Maine == "Away"), 12]

> hist(awayRunnerTime, xlab = "Maine column value Away", main = "Histogram of the runners' times for Maine column value as Away", xlim = range(0,150), ylim=range(0,2000),col="Blue")

> maineRunnerTime <- givenDataSet[which(givenDataSet$Maine == "Maine"), 12]

> hist(maineRunnerTime, xlab = "Maine column value Maine", main = "Histogram of the runners' times for Maine column value as Maine", xlim = range(0,150), ylim = range(0,2000), col = "Blue")
```

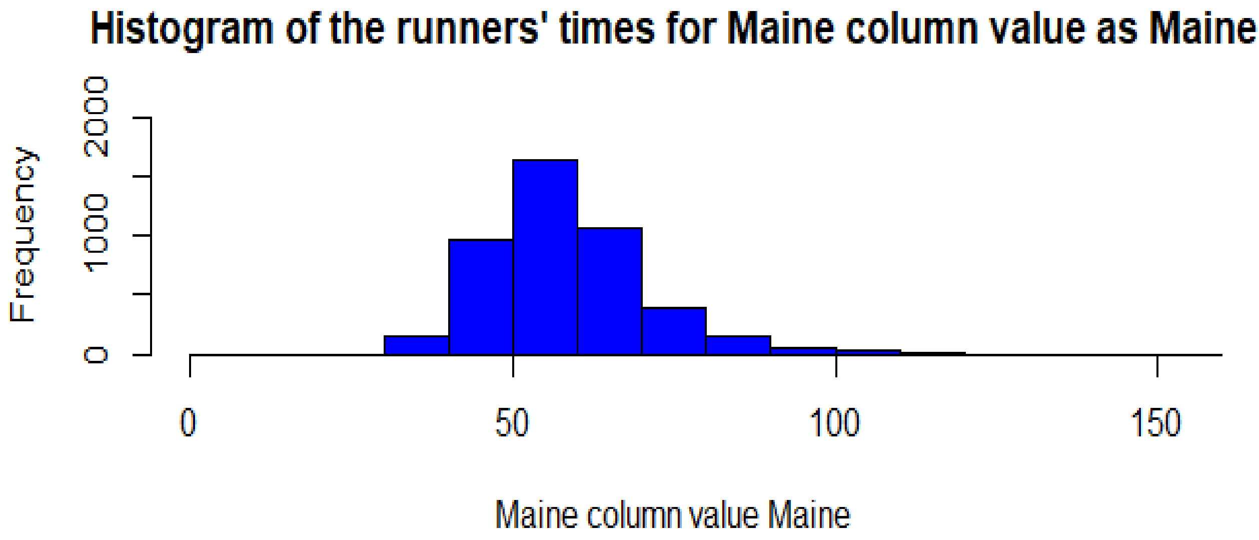
Output:



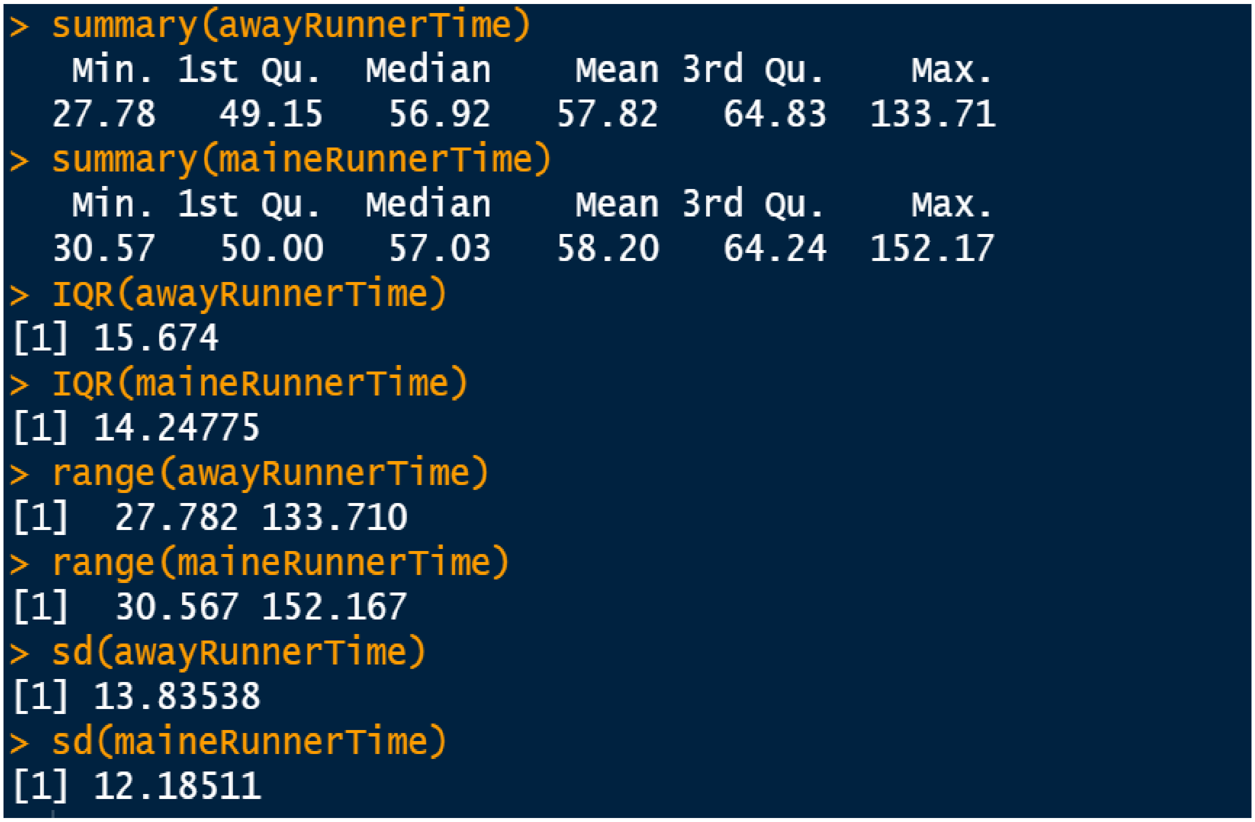
```
> awayRunnerTime <- givenDataSet[which(givenDataSet$Maine ==
 "Away"), 12] #Time (minutes) column no is 12
> hist(awayRunnerTime, xlab = "Maine column value Away", main
 = "Histogram of the runners' times for Maine column value as
 Away",
+       xlim = range(0,150), ylim = range(0,2000), col = "Blu
e")
>
> maineRunnerTime <- givenDataSet[which(givenDataSet$Maine ==
 "Maine"), 12] #Time (minutes) column no is 12
> hist(maineRunnerTime, xlab = "Maine column value Maine", mai
n = "Histogram of the runners' times for Maine column value as
 Maine",
+       xlim = range(0,150), ylim = range(0,2000), col = "Blu
e")
```

Histogram:





Summary:



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

Observation:

Going through the graph, it can be concluded that both distributions are right-skewed. We used the summary function of R to find out the Min, Max, 1st Q, Median, Mean, 3rd Q.

Summary Stats:

Maine	Min.	1st Q	Med	Mean	3rd Q	Max
Away	27.78	49.15	56.92	57.82	64.83	133.71
Maine	30.57	50.00	57.03	58.20	64.24	152.17

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

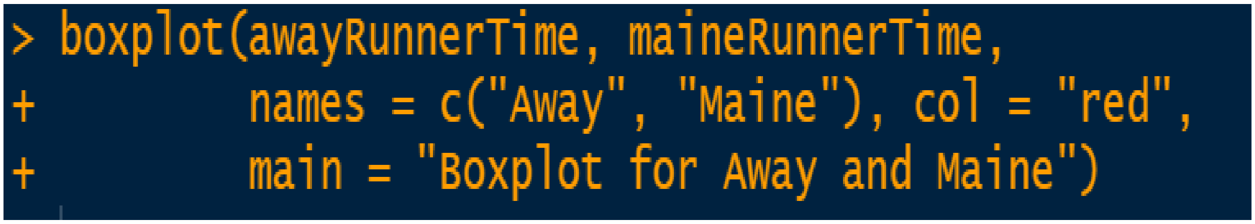
Answer:

To create side-by-side boxplots of runner time from Maine and Away we use data from Maine column and Time column from dataset roadrace.csv

Code Snippet:

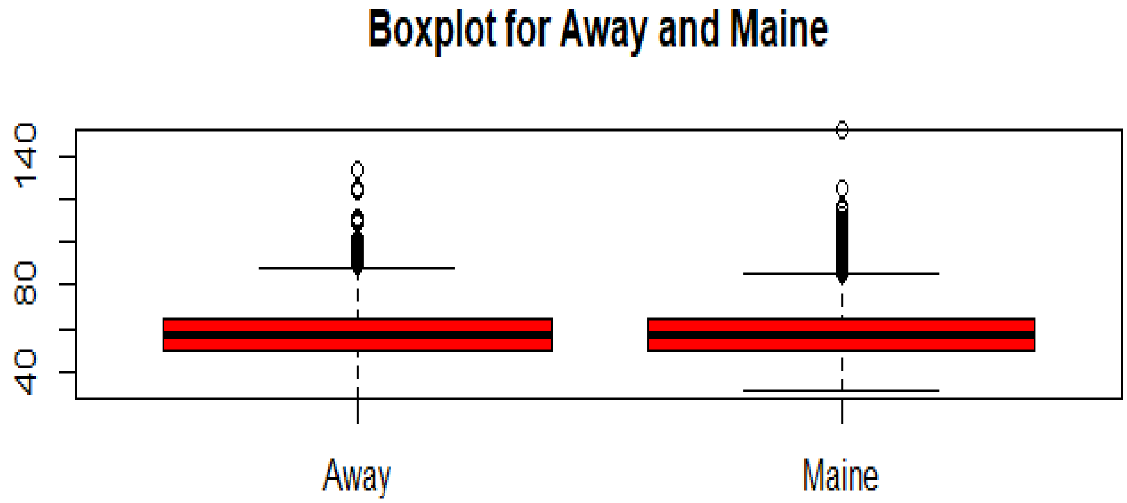
boxplot(awayRunnerTime,maineRunnerTime,names= c("Away", "Maine"), col = "red",main = "Boxplot for Away and Maine")

Output:



```
> boxplot(awayRunnerTime, maineRunnerTime,
+          names = c("Away", "Maine"), col = "red",
+          main = "Boxplot for Away and Maine")
```

Box Plot:



(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.

Answer:

Now we will create gender specific box plot of runners using the given dataset of roadrace.csv

Code Snippet:

> maleAgeGroup <- givenDataSet[which(givenDataSet$Sex == "M"), 5] #Age column no is 5

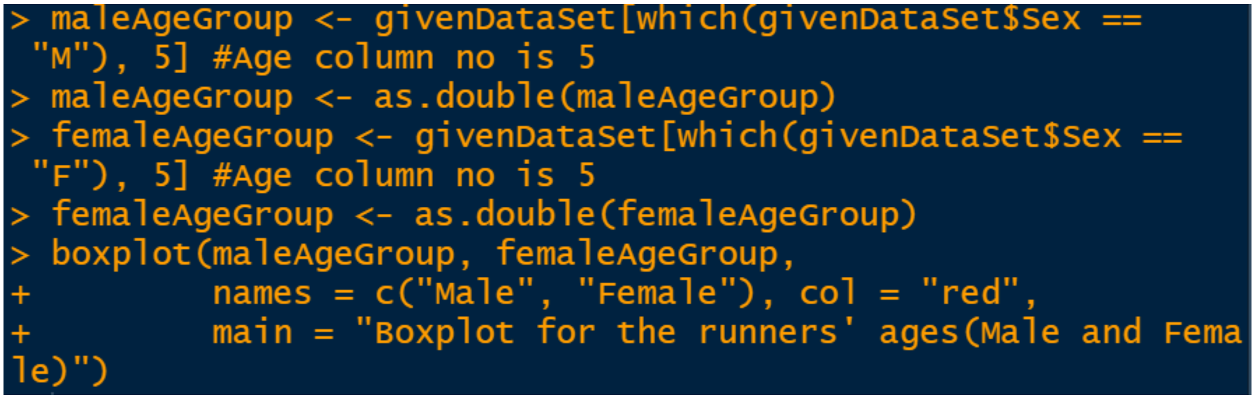
> maleAgeGroup <- as.double(maleAgeGroup)

> femaleAgeGroup <- givenDataSet[which(givenDataSet$Sex == "F"), 5] #Age column no is 5

> femaleAgeGroup <- as.double(femaleAgeGroup)

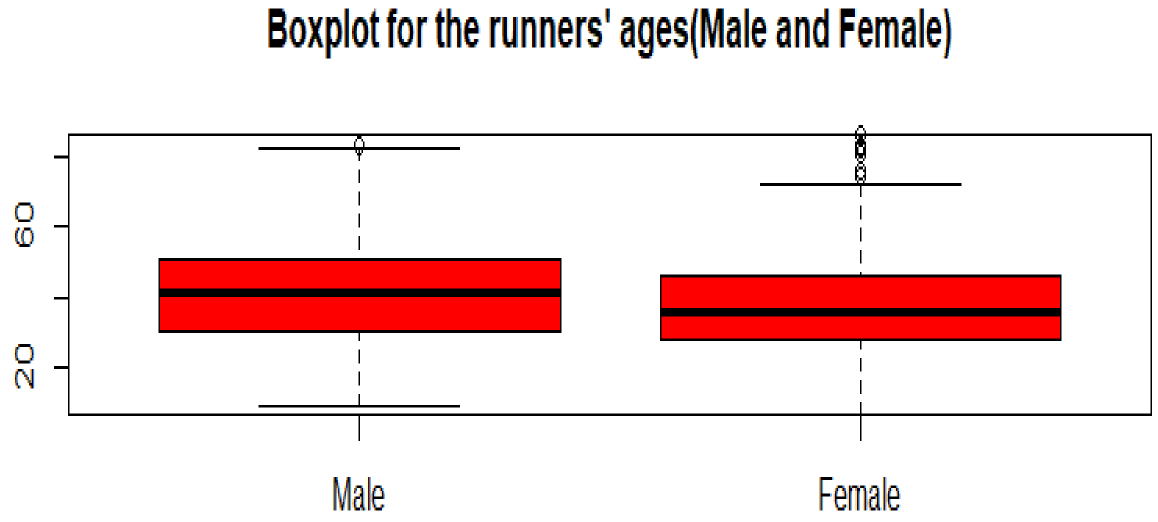
>boxplot(maleAgeGroup, femaleAgeGroup, names = c("Male", "Female"), col = "red",main = "Boxplot for the runners' ages(Male and Female)")

Output:



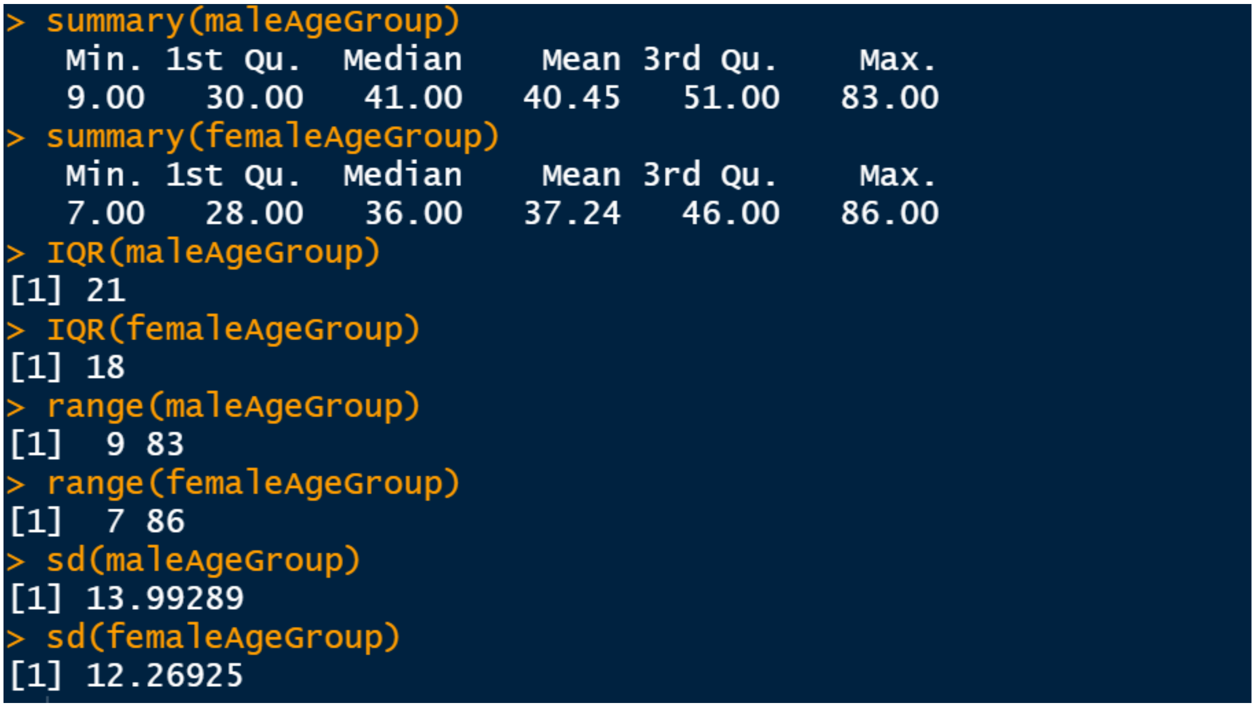
```
> maleAgeGroup <- givenDataSet[which(givenDataSet$Sex ==
 "M"), 5] #Age column no is 5
> maleAgeGroup <- as.double(maleAgeGroup)
> femaleAgeGroup <- givenDataSet[which(givenDataSet$Sex ==
 "F"), 5] #Age column no is 5
> femaleAgeGroup <- as.double(femaleAgeGroup)
> boxplot(maleAgeGroup, femaleAgeGroup,
+         names = c("Male", "Female"), col = "red",
+         main = "Boxplot for the runners' ages(Male and Fema
le)")
```

Box Plot:



Summary Details:

```
> summary(maleAgeGroup)

> summary(femaleAgeGroup)

> IQR(maleAgeGroup)

> IQR(femaleAgeGroup)

> range(maleAgeGroup)

> range(femaleAgeGroup)

> sd(maleAgeGroup)

> sd(femaleAgeGroup)
```



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

Summary Stats:

Sex	**Min.**	**1st Q**	**Med**	**Mean**	**3rd Q**	**Max**
Male	9.00	30.00	41.00	40.45	51.00	83.00
Female	7.00	28.00	36.00	37.24	46.00	86.00

Observation:

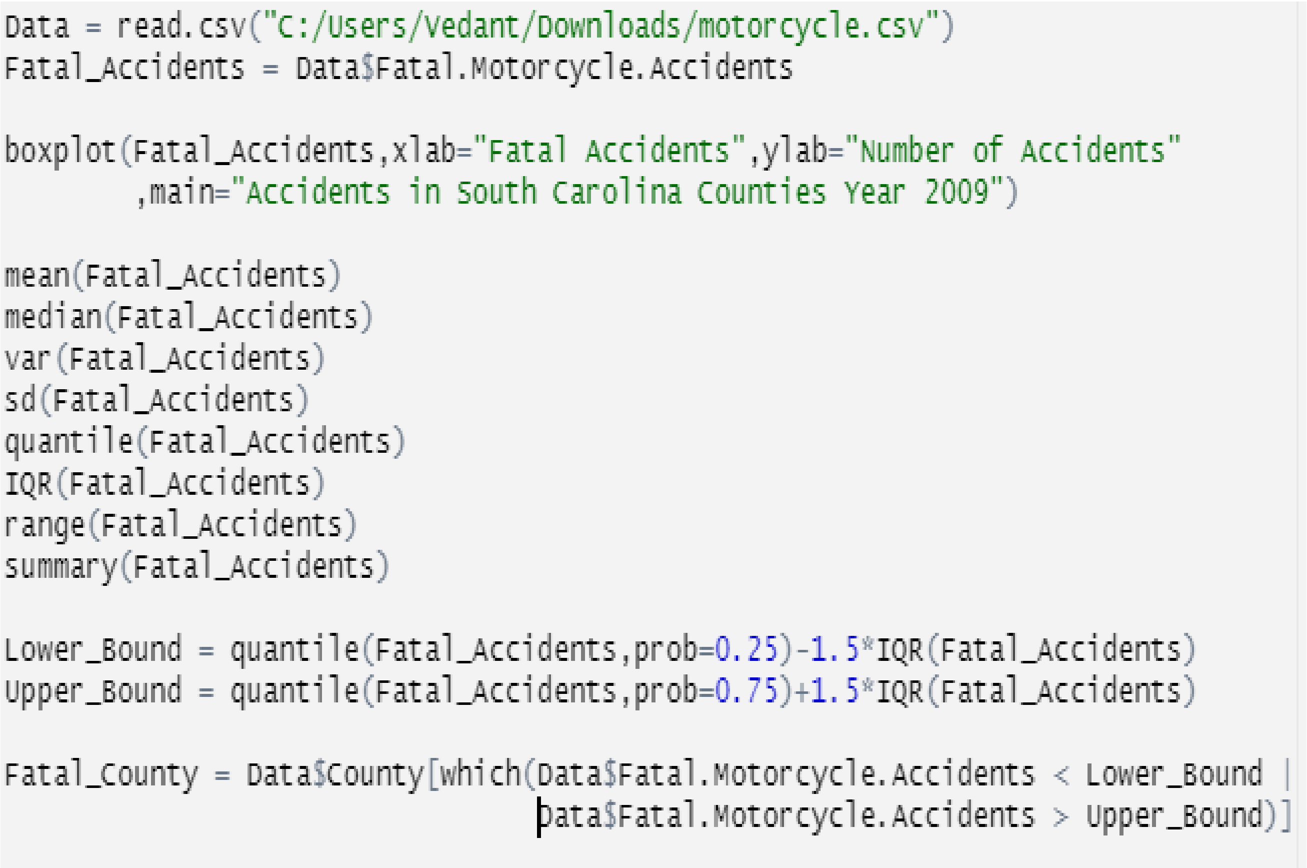
Going through the graph as well as stats, Female of age group 86 and more are more actively participating in the race.

2. (8 points) 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?

Answer:

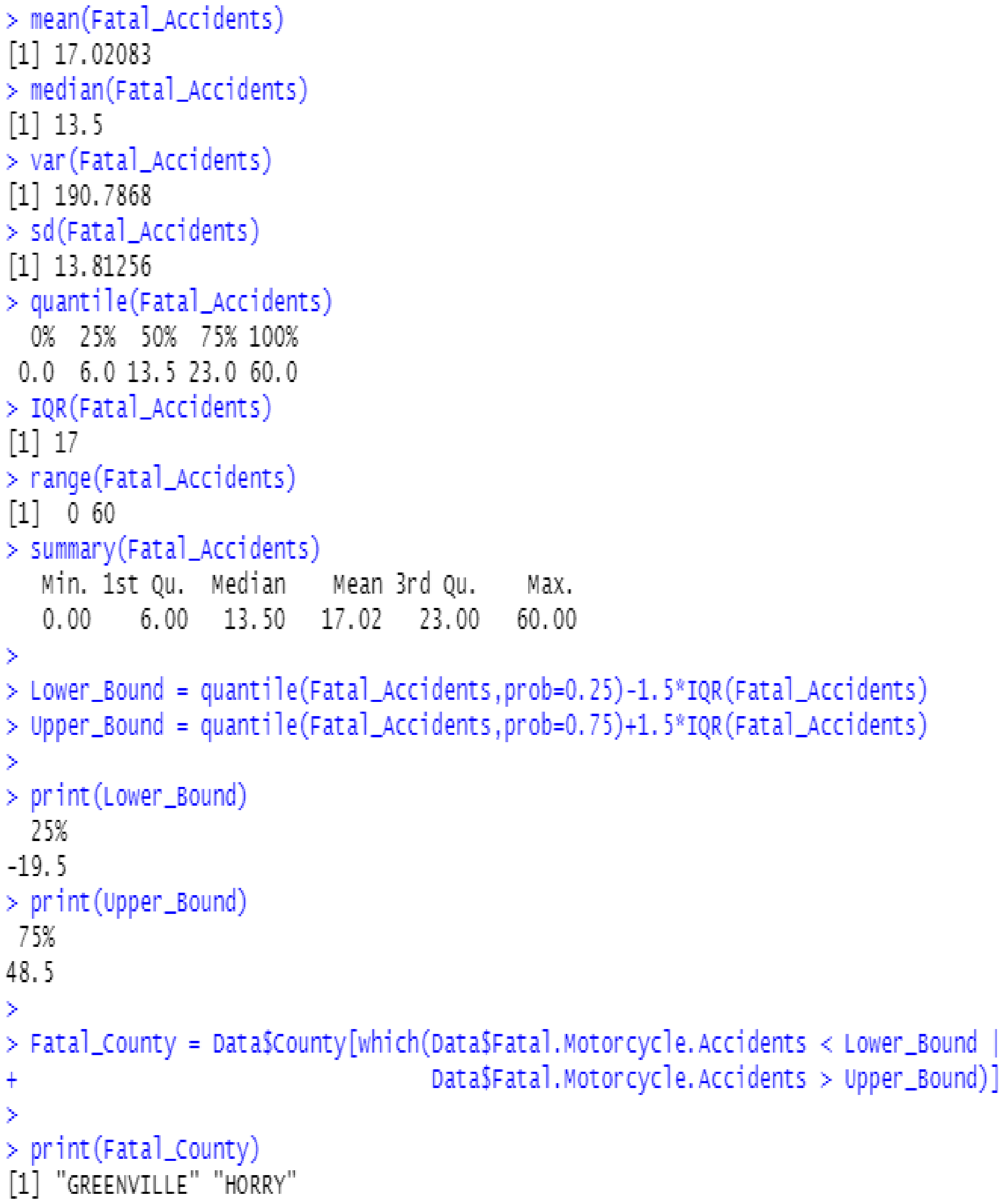
First we extract Fatal Motorcycle Accidents column from given dataset of motorcycle.csv and then create bar plot and also perform summary statistics.

Code Snippet:



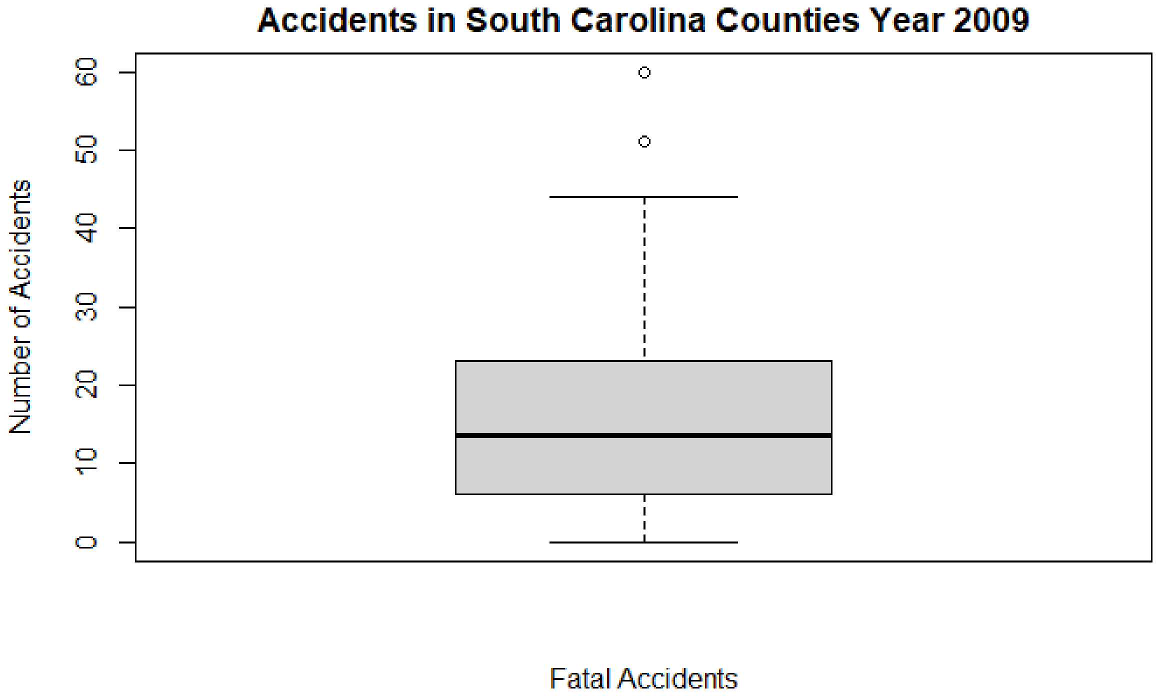
```
Data = read.csv("C:/Users/Vedant/Downloads/motorcycle.csv")
Fatal_Accidents = Data$Fatal.Motorcycle.Accidents

boxplot(Fatal_Accidents,xlab="Fatal Accidents",ylab="Number of Accidents"
        ,main="Accidents in South Carolina Counties Year 2009")

mean(Fatal_Accidents)
median(Fatal_Accidents)
var(Fatal_Accidents)
sd(Fatal_Accidents)
quantile(Fatal_Accidents)
IQR(Fatal_Accidents)
range(Fatal_Accidents)
summary(Fatal_Accidents)

Lower_Bound = quantile(Fatal_Accidents,prob=0.25)-1.5*IQR(Fatal_Accidents)
Upper_Bound = quantile(Fatal_Accidents,prob=0.75)+1.5*IQR(Fatal_Accidents)

Fatal_County = Data$County[which(Data$Fatal.Motorcycle.Accidents < Lower_Bound |
                                 Data$Fatal.Motorcycle.Accidents > Upper_Bound)]
```

Output:



```
> mean(Fatal_Accidents)
[1] 17.02083
> median(Fatal_Accidents)
[1] 13.5
> var(Fatal_Accidents)
[1] 190.7868
> sd(Fatal_Accidents)
[1] 13.81256
> quantile(Fatal_Accidents)
  0%  25%  50%  75% 100% 
 0.0  6.0 13.5 23.0 60.0 
> IQR(Fatal_Accidents)
[1] 17
> range(Fatal_Accidents)
[1]  0 60
> summary(Fatal_Accidents)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    6.00   13.50   17.02   23.00   60.00 
>
> Lower_Bound = quantile(Fatal_Accidents,prob=0.25)-1.5*IQR(Fatal_Accidents)
> Upper_Bound = quantile(Fatal_Accidents,prob=0.75)+1.5*IQR(Fatal_Accidents)
>
> print(Lower_Bound)
  25% 
-19.5 
> print(Upper_Bound)
 75% 
48.5 
>
> Fatal_County = Data$County[which(Data$Fatal.Motorcycle.Accidents < Lower_Bound |
+                                  Data$Fatal.Motorcycle.Accidents > Upper_Bound)]
>
> print(Fatal_County)
[1] "GREENVILLE" "HORRY"
```

Box plot:



Summary:

	Min.	1st Q	Med	Mean	3rd Q	Max	IQR	Range	SD
Accidents	0.00	6.0	13.5	17.02	23.00	60.00	17	0-60	13.81

Observation:

As seen in the box plot we can see that the distribution is right-skewed since more values are at the right end. Also, the mean is higher than the median implying right skewness of distribution. The quartiles also show right skewness as the gap between data of Q1 and Q3 is close and Q3 and Q4 is very large. The range of data is 0-60 and range with IQR*1.5 is 68.

We can find the Outliers Counties by finding counties that are not in range of upper bound and lower bound of data.

Outlier Counties = Greenville and Horry

Greenville has 51 Accidents and Horry has 60 Accidents.

The Reason for the Fatal Accidents on motorcycles in South Carolina during the year9 could be due to the poor road and highway construction and maintenance. Also, the reason could be negligence from the driver's side.