STATISTICAL METHODS FOR DATA SCIENCE   Mini-Project 2

Duo Group #23

Members: Hima Sri Tipirineni

Nithin Pingili

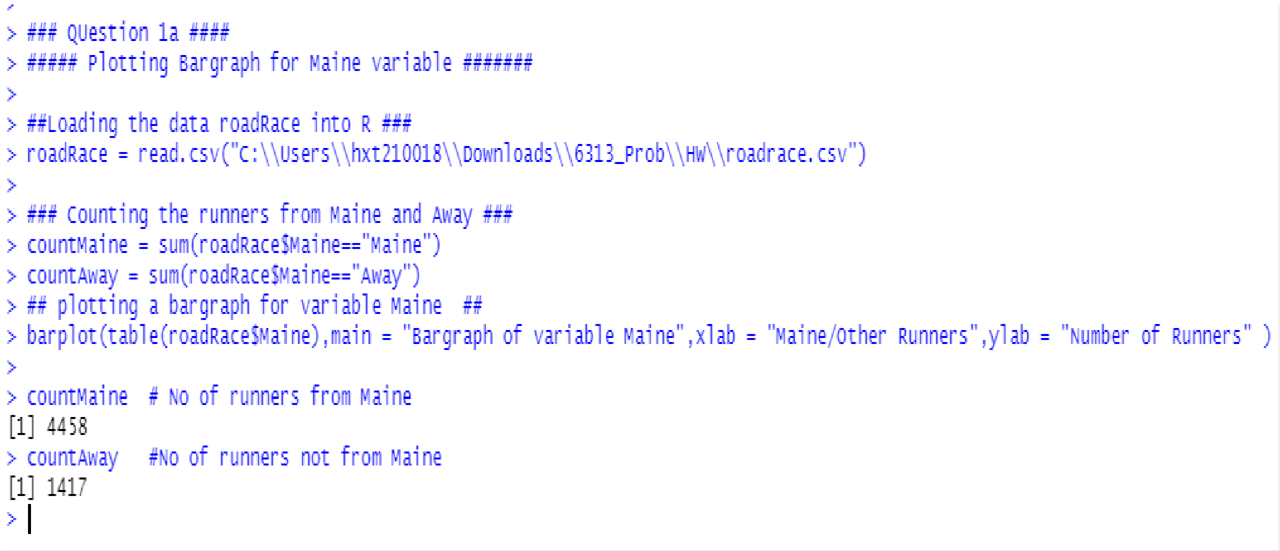
Contribution of each team member:

Hima Sri and Nithin worked together to complete both the questions. Collaborated to learn R and then worked on plotting the bar graphs , boxplots and histograms for the given data distributions. Both worked together to answer the questions and report all the findings. Hima Sri wrote R code and annotated the code and Nithin worked to check the accuracy of the R code and added the observations . Both worked efficiently to complete all sections of the project.

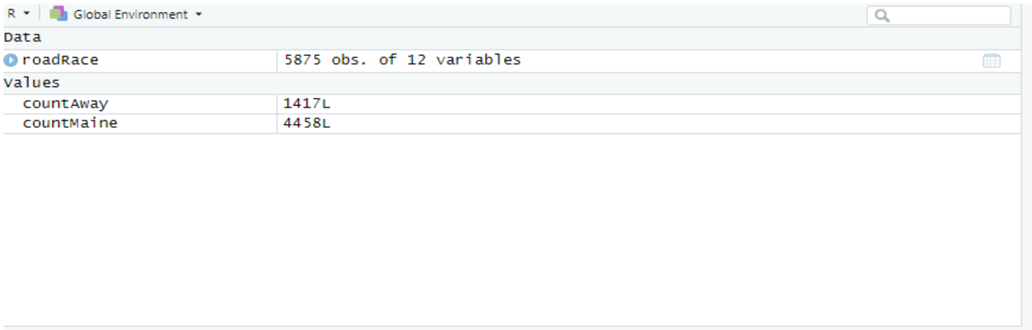
Question 1:

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.

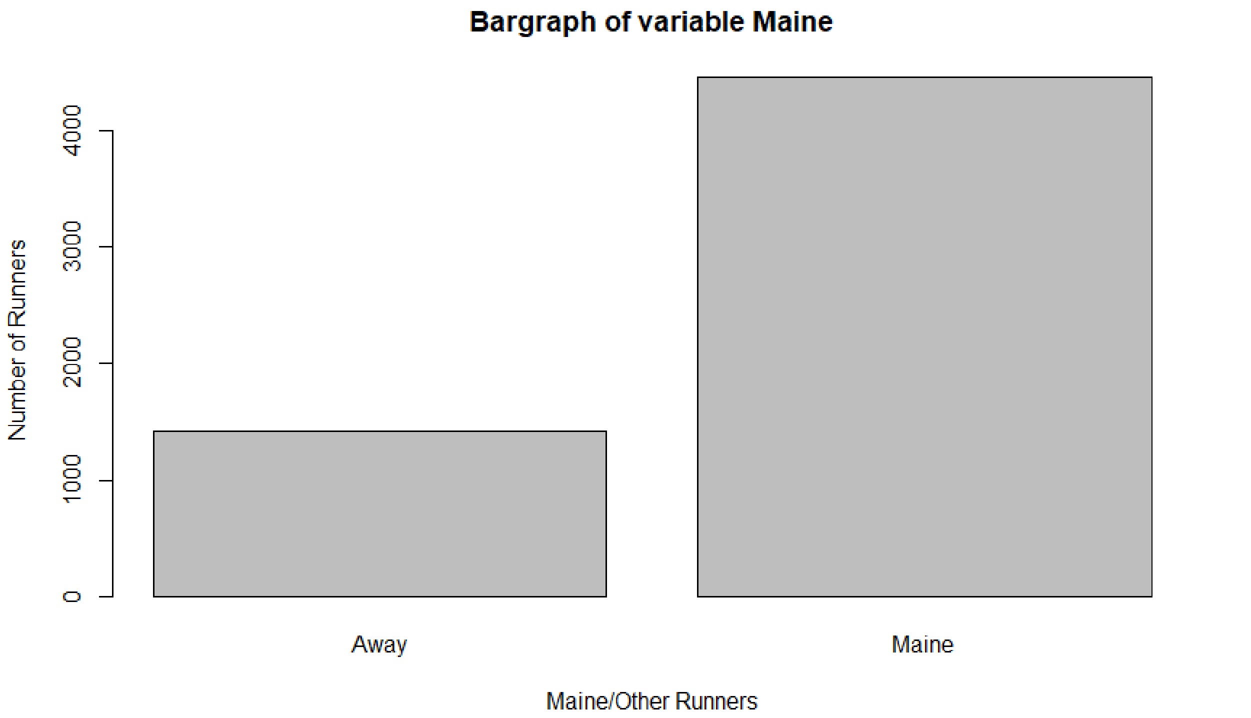
Solution :



```
> ### QUestion 1a ####
> ##### Plotting Bargraph for Maine variable #######
>
> ##Loading the data roadRace into R ###
> roadRace = read.csv("C:\\Users\\hxt210018\\Downloads\\6313_Prob\\HW\\roadrace.csv")
>
> ### Counting the runners from Maine and Away ###
> countMaine = sum(roadRace$Maine=="Maine")
> countAway = sum(roadRace$Maine=="Away")
> ## plotting a bargraph for variable Maine  ##
> barplot(table(roadRace$Maine),main = "Bargraph of variable Maine",xlab = "Maine/Other Runners",ylab = "Number of Runners" )
>
> countMaine  # No of runners from Maine
[1] 4458
> countAway   #No of runners not from Maine
[1] 1417
> 
```



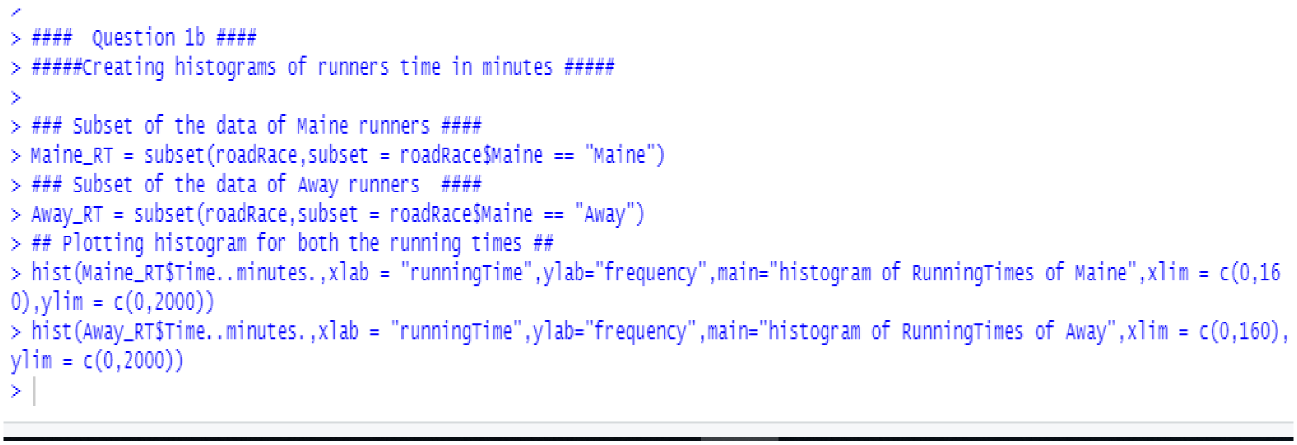
```
Data
roadRace      5875 obs. of 12 variables
Values
  countAway   1417L
  countMaine  4458L
```



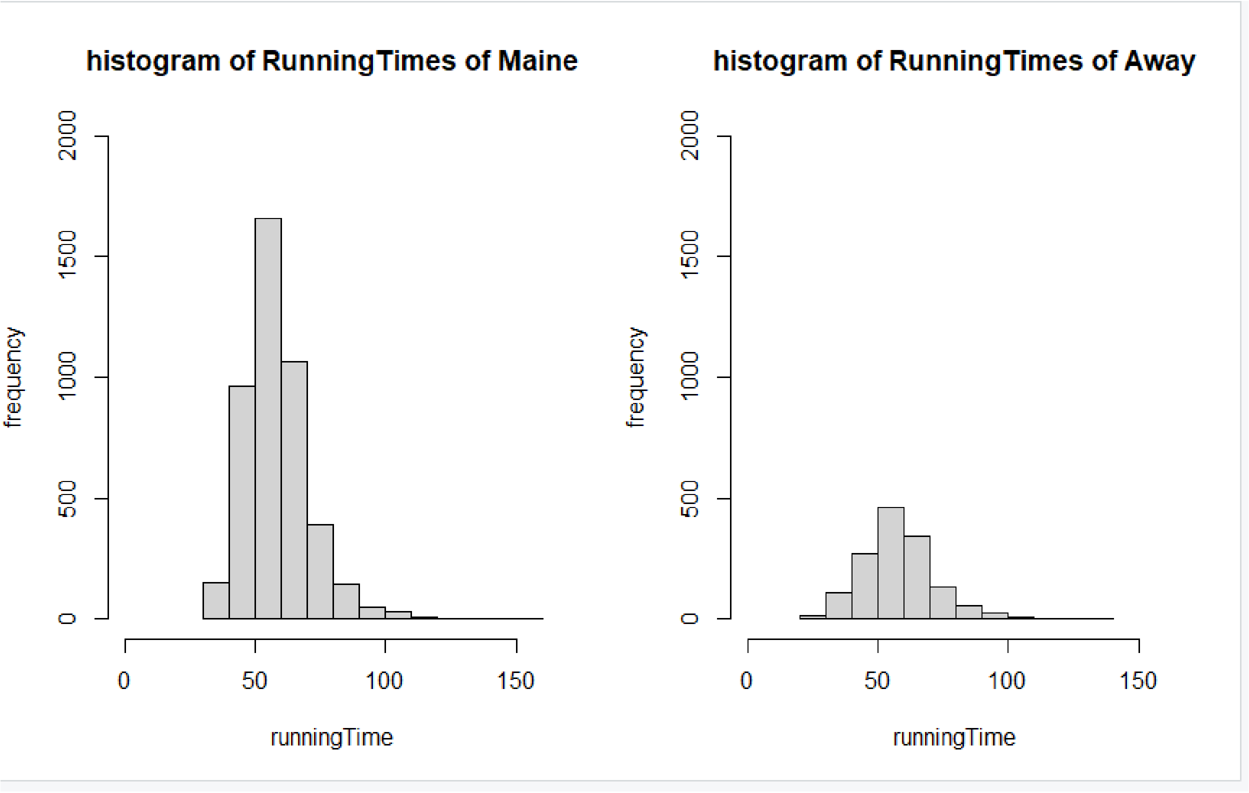
From the plot we can conclude that runners from Maine outnumber runners from somewhere else. Out of the 5875 runners, 75.88% of the runners are from Maine and 24.12% of the runners are from somewhere else.

(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? Backup your conclusions with relevant summary statistics, including mean, standard deviation, range, median, and interquartile range.

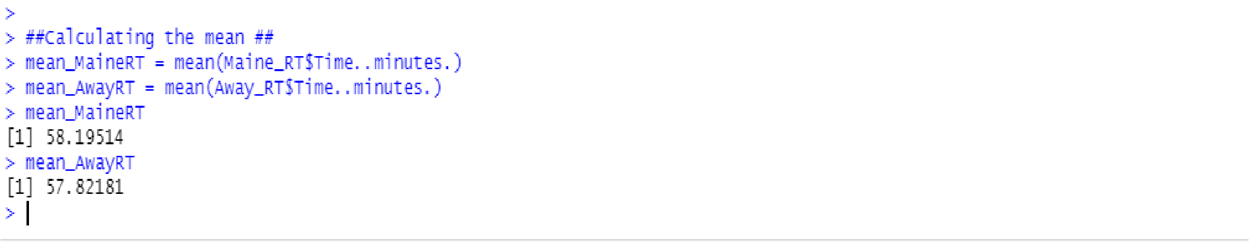
Solution :



```
> ####  Question 1b ####
> #####Creating histograms of runners time in minutes #####
>
> ### Subset of the data of Maine runners ####
> Maine_RT = subset(roadRace,subset = roadRace$Maine == "Maine")
> ### Subset of the data of Away runners  ####
> Away_RT = subset(roadRace,subset = roadRace$Maine == "Away")
> ## Plotting histogram for both the running times ##
> hist(Maine_RT$Time..minutes.,xlab = "runningTime",ylab="frequency",main="histogram of RunningTimes of Maine",xlim = c(0,16
0),ylim = c(0,2000))
> hist(Away_RT$Time..minutes.,xlab = "runningTime",ylab="frequency",main="histogram of RunningTimes of Away",xlim = c(0,160),
ylim = c(0,2000))
> 
```

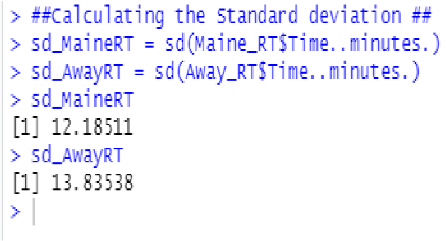


Mean :



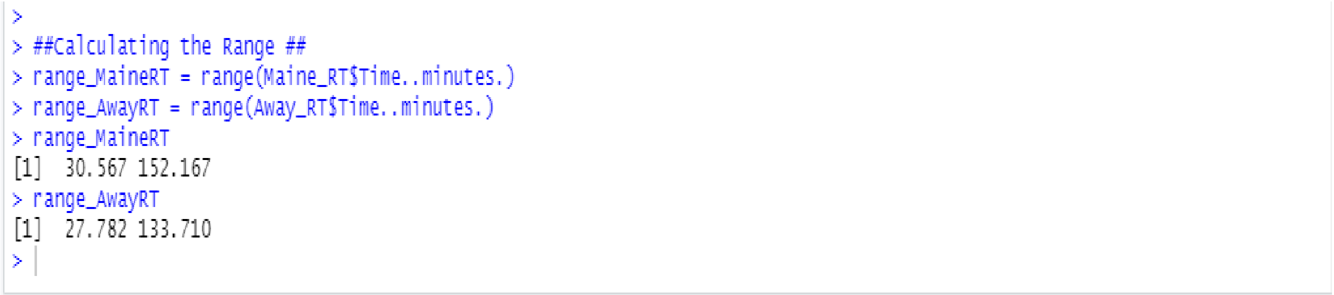
```
>
> ##Calculating the mean ##
> mean_MaineRT = mean(Maine_RT$Time..minutes.)
> mean_AwayRT = mean(Away_RT$Time..minutes.)
> mean_MaineRT
[1] 58.19514
> mean_AwayRT
[1] 57.82181
> 
```

Standard Deviation :



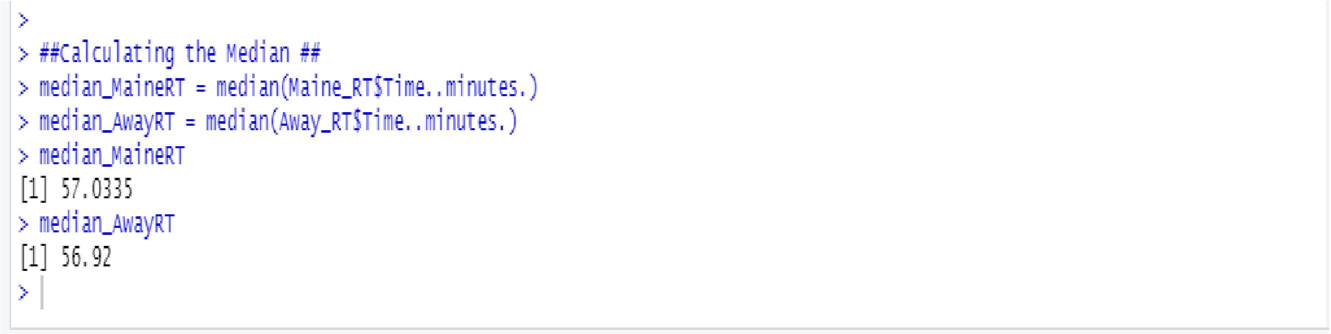
```
> ##Calculating the Standard deviation ##
> sd_MaineRT = sd(Maine_RT$Time..minutes.)
> sd_AwayRT = sd(Away_RT$Time..minutes.)
> sd_MaineRT
[1] 12.18511
> sd_AwayRT
[1] 13.83538
> 
```

Range :



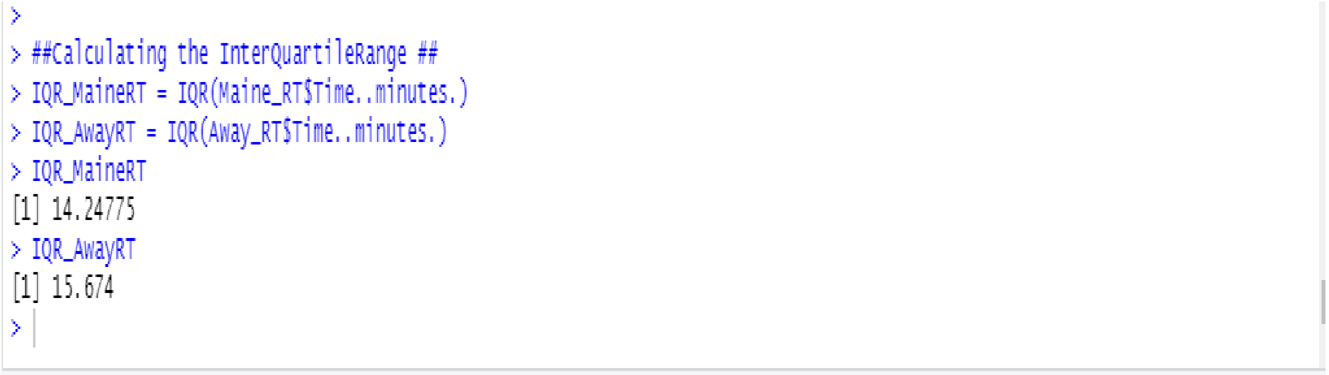
```
> 
> ##Calculating the Range ##
> range_MaineRT = range(Maine_RT$Time..minutes.)
> range_AwayRT = range(Away_RT$Time..minutes.)
> range_MaineRT
[1]  30.567 152.167
> range_AwayRT
[1]  27.782 133.710
> 
```

Median :



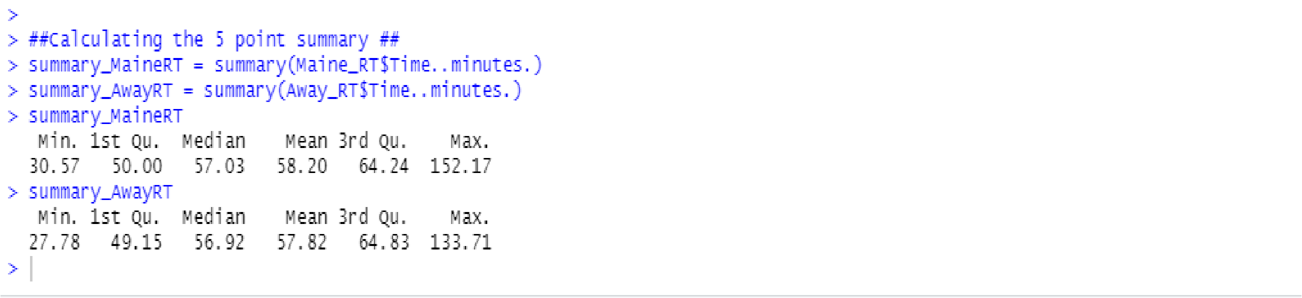
```
> 
> ##Calculating the Median ##
> median_MaineRT = median(Maine_RT$Time..minutes.)
> median_AwayRT = median(Away_RT$Time..minutes.)
> median_MaineRT
[1] 57.0335
> median_AwayRT
[1] 56.92
> 
```

InterQuartileRange :



```
> 
> ##Calculating the InterQuartileRange ##
> IQR_MaineRT = IQR(Maine_RT$Time..minutes.)
> IQR_AwayRT = IQR(Away_RT$Time..minutes.)
> IQR_MaineRT
[1] 14.24775
> IQR_AwayRT
[1] 15.674
> 
```

Summary :

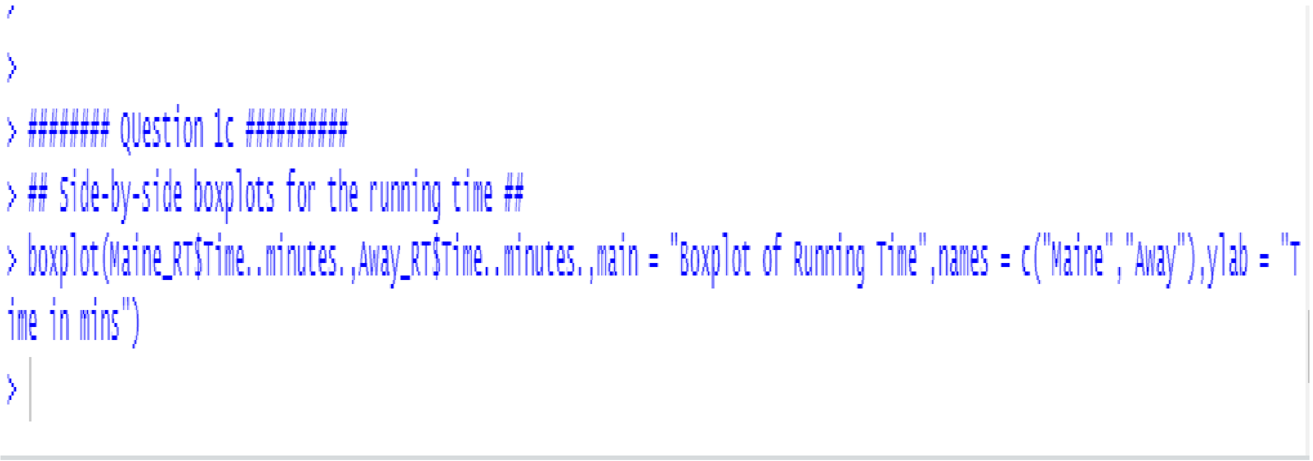


```
> 
> ##Calculating the 5 point summary ##
> summary_MaineRT = summary(Maine_RT$Time..minutes.)
> summary_AwayRT = summary(Away_RT$Time..minutes.)
> summary_MaineRT
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  30.57   50.00   57.03   58.20   64.24  152.17 
> summary_AwayRT
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  27.78   49.15   56.92   57.82   64.83  133.71 
> 
```

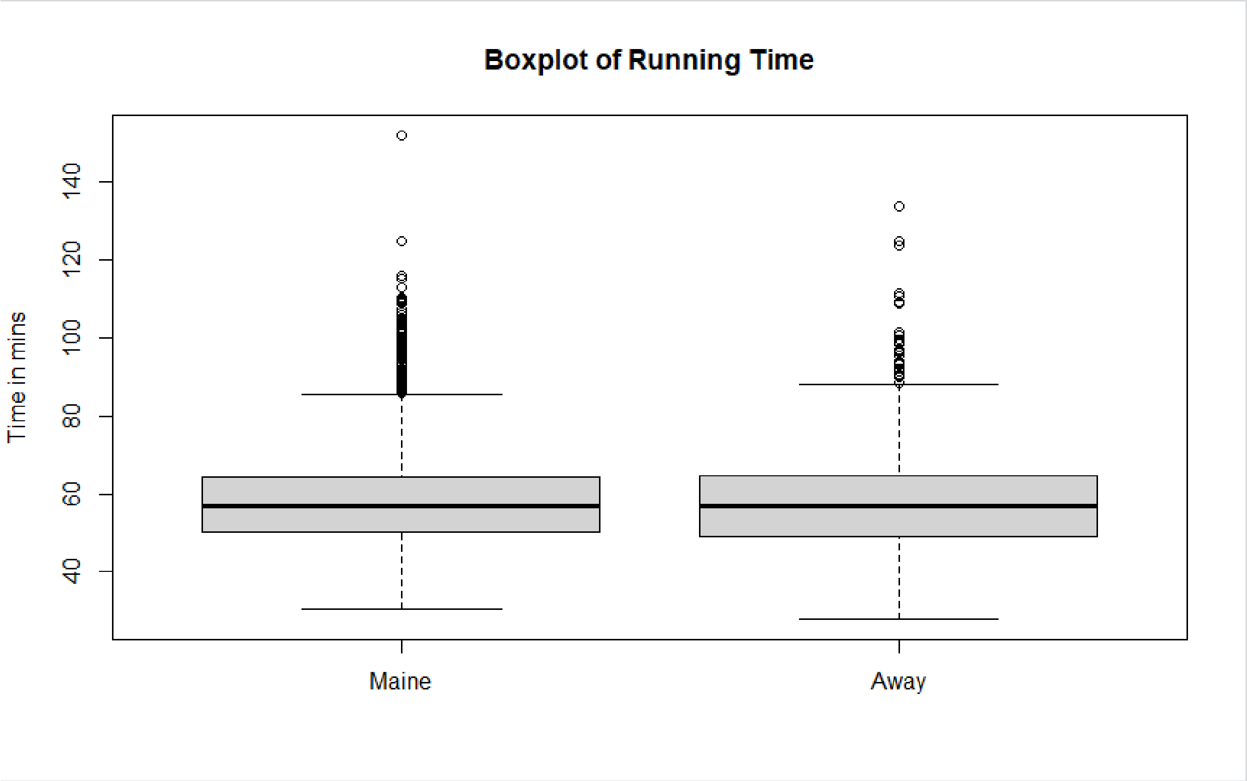
	Mean	**SD**	**Range**	**IQR**	**Min**	**1st Qu**	**Median**	**3rd Qu**	**Max**
MaineRT	58.20	12.18511	[30.57,152.17]	14.25	30.57	50.00	57.03	64.24	152.17
AwayRT	57.82	13.83538	[27.78,133.71]	15.67	27.78	49.15	56.92	64.83	133.71

Both distributions are skewed right. It is because the median is less than mean value. For both distributions, the median is closer to the min than the max. Also, the difference between the 1st Quartile and min is smaller than the difference between 3rd quartile and max for both distributions. These observations show that both distributions are skewed right.

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

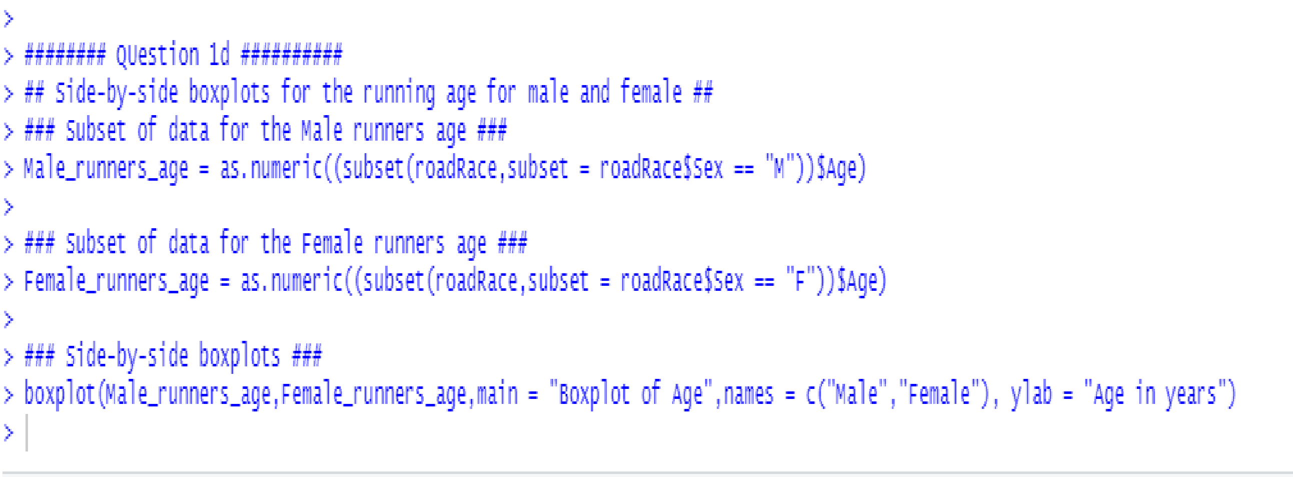


```
>
> ####### Question 1c #########
> ## Side-by-side boxplots for the running time ##
> boxplot(Maine_RT$Time..minutes.,Away_RT$Time..minutes.,main = "Boxplot of Running Time",names = c("Maine","Away"),ylab = "T
ime in mins")
> 
```

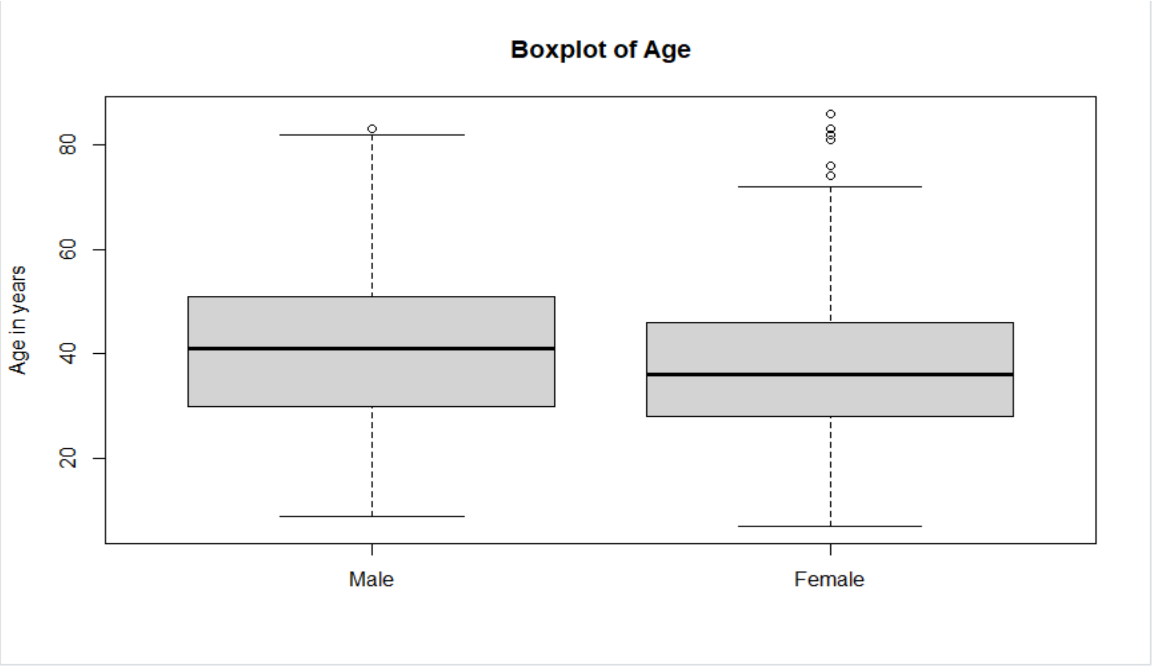


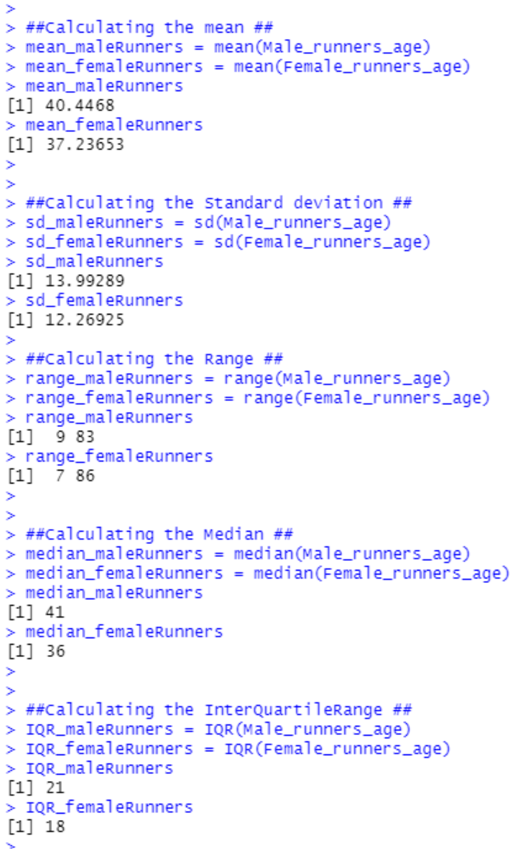
Both distributions are skewed right. It is because the median is less than mean value .For both distributions, the median is closer to the min than the max. Also, the difference between the 1st Quartile and min is smaller than the difference between 3rd quartile and max for both distributions. These observations show that both distributions are skewed 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.

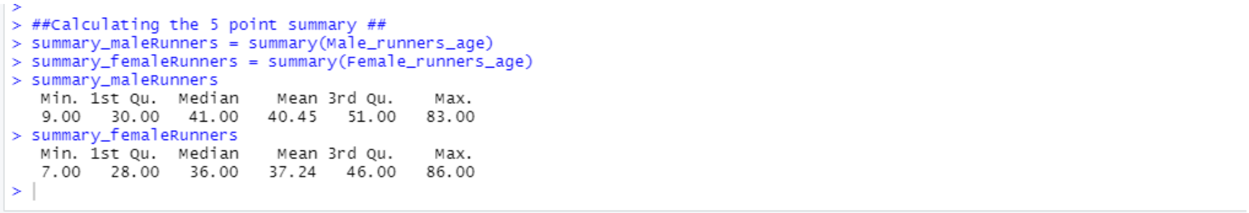


```
>
> ######## Question 1d ##########
> ## Side-by-side boxplots for the running age for male and female ##
> ### Subset of data for the Male runners age ###
> Male_runners_age = as.numeric((subset(roadRace,subset = roadRace$Sex == "M"))$Age)
>
> ### Subset of data for the Female runners age ###
> Female_runners_age = as.numeric((subset(roadRace,subset = roadRace$Sex == "F"))$Age)
>
> ### Side-by-side boxplots ###
> boxplot(Male_runners_age,Female_runners_age,main = "Boxplot of Age",names = c("Male","Female"), ylab = "Age in years")
> 
```





```
>
> ##Calculating the mean ##
> mean_maleRunners = mean(Male_runners_age)
> mean_femaleRunners = mean(Female_runners_age)
> mean_maleRunners
[1] 40.4468
> mean_femaleRunners
[1] 37.23653
>
>
> ##Calculating the Standard deviation ##
> sd_maleRunners = sd(Male_runners_age)
> sd_femaleRunners = sd(Female_runners_age)
> sd_maleRunners
[1] 13.99289
> sd_femaleRunners
[1] 12.26925
>
> ##Calculating the Range ##
> range_maleRunners = range(Male_runners_age)
> range_femaleRunners = range(Female_runners_age)
> range_maleRunners
[1]  9 83
> range_femaleRunners
[1]  7 86
>
>
> ##Calculating the Median ##
> median_maleRunners = median(Male_runners_age)
> median_femaleRunners = median(Female_runners_age)
> median_maleRunners
[1] 41
> median_femaleRunners
[1] 36
>
>
> ##Calculating the InterQuartileRange ##
> IQR_maleRunners = IQR(Male_runners_age)
> IQR_femaleRunners = IQR(Female_runners_age)
> IQR_maleRunners
[1] 21
> IQR_femaleRunners
[1] 18
>
```



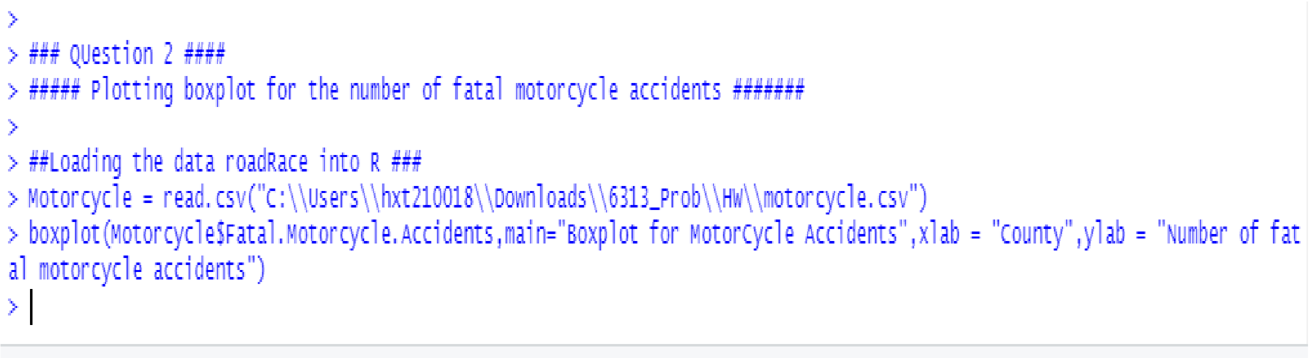
```
>
> ##Calculating the 5 point summary ##
> summary_maleRunners = summary(Male_runners_age)
> summary_femaleRunners = summary(Female_runners_age)
> summary_maleRunners
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   9.00   30.00   41.00   40.45   51.00   83.00 
> summary_femaleRunners
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   7.00   28.00   36.00   37.24   46.00   86.00 
> 
```

	Mean	**SD**	**Range**	**IQR**	**Min**	**1st Qu**	**Median**	**3rd Qu**	**Max**
Male Runners	40.45	13.99289	[9,83]	21	9.00	30.00	41.00	51.00	83.00
Female Runners	37.24	12.26925	[7,86]	18	7.00	28.00	36.00	46.00	86.00

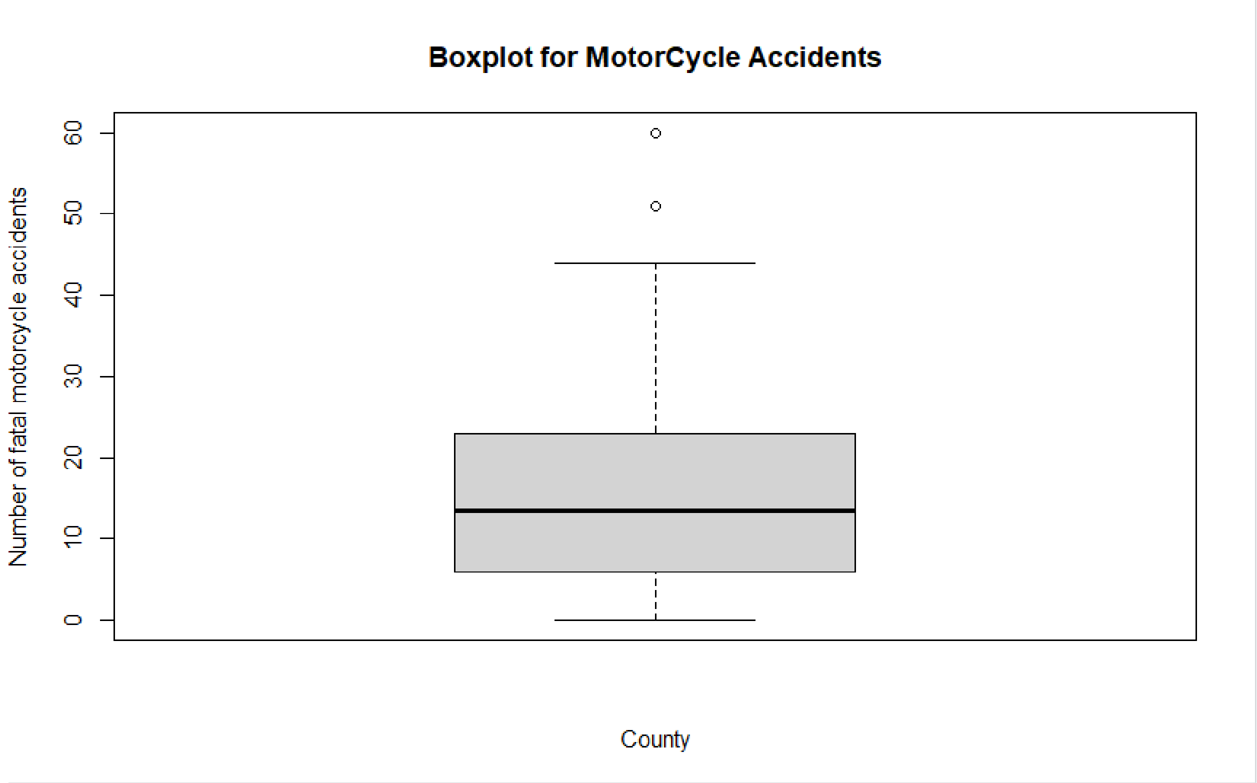
The two boxplots show that male runners tend to be older than female runners. This is because the 1st quartile, 3rd quartile and IQR are greater for male runners than female runners. Though, the oldest female runner is older than the oldest male runner because the max age of female runners is 86 and the max age of male runners is 83.

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?

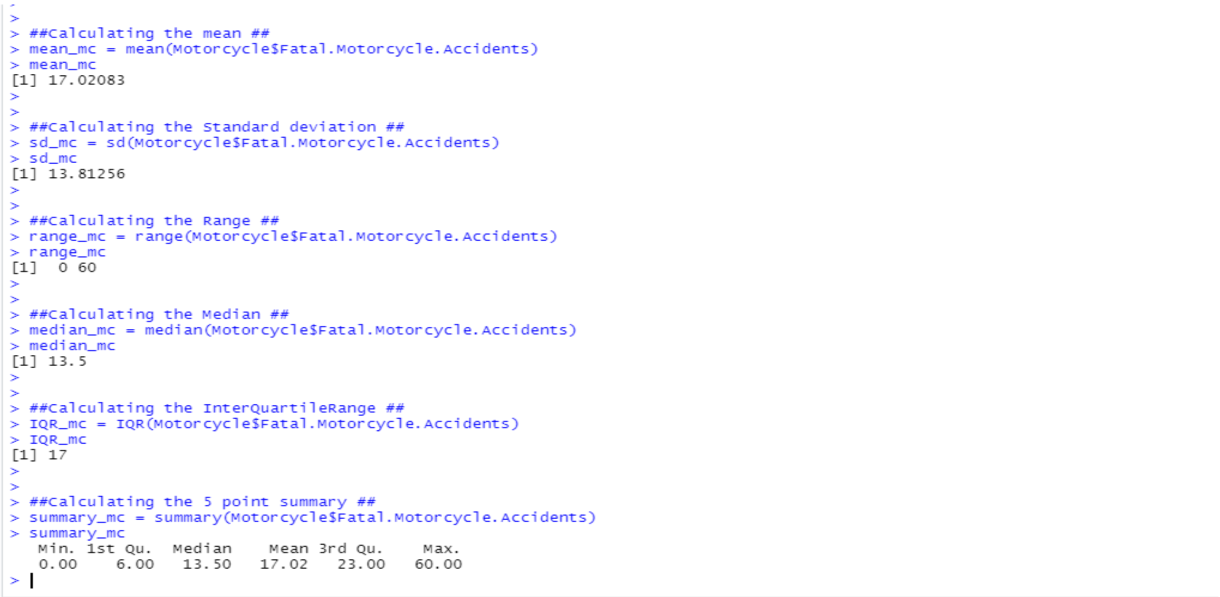
Solution:



```
>
> ### QUestion 2 ####
> ##### Plotting boxplot for the number of fatal motorcycle accidents #######
>
> ##Loading the data roadRace into R ###
> Motorcycle = read.csv("C:\\Users\\hxt210018\\Downloads\\6313_Prob\\HW\\motorcycle.csv")
> boxplot(Motorcycle$Fatal.Motorcycle.Accidents,main="Boxplot for MotorCycle Accidents",xlab = "County",ylab = "Number of fat
al motorcycle accidents")
> 
```



Summary Statistics :

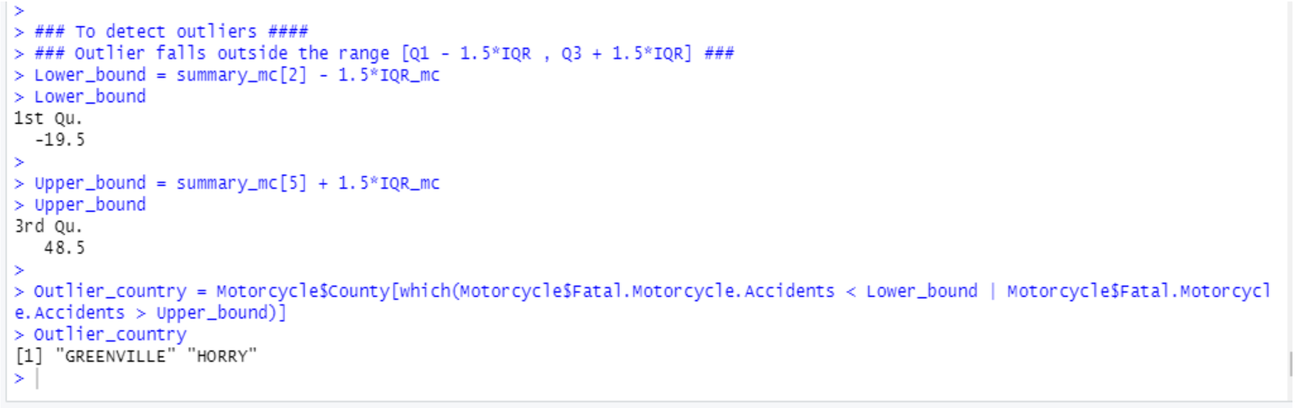


```
>
> ##Calculating the mean ##
> mean_mc = mean(Motorcycle$Fatal.Motorcycle.Accidents)
> mean_mc
[1] 17.02083
>
>
> ##Calculating the Standard deviation ##
> sd_mc = sd(Motorcycle$Fatal.Motorcycle.Accidents)
> sd_mc
[1] 13.81256
>
>
> ##Calculating the Range ##
> range_mc = range(Motorcycle$Fatal.Motorcycle.Accidents)
> range_mc
[1]  0 60
>
>
> ##Calculating the Median ##
> median_mc = median(Motorcycle$Fatal.Motorcycle.Accidents)
> median_mc
[1] 13.5
>
>
> ##Calculating the InterQuartileRange ##
> IQR_mc = IQR(Motorcycle$Fatal.Motorcycle.Accidents)
> IQR_mc
[1] 17
>
>
> ##Calculating the 5 point summary ##
> summary_mc = summary(Motorcycle$Fatal.Motorcycle.Accidents)
> summary_mc
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    6.00   13.50   17.02   23.00   60.00 
> |
```

	Mean	SD	Range	IQR	Min	1st Qu	Median	3rd Qu	Max
MotorCycleAccidents	17.02	13.81256	[0,60]	17	0.00	6.00	13.50	23.00	60.00

The boxplot shows that the distribution is skewed right. This is because the median is closer to the 1st quartile because the difference between median and 1st quartile is less than the difference between median and 3rd quartile. Also, the whisker is shorter on the lower end of box because the difference between min and 1st quartile is less than the difference between max and 3rd quartile. These observations show that distribution is skewed right.

Outlier detection :



```
>
> ### To detect outliers ####
> ### Outlier falls outside the range [Q1 - 1.5*IQR , Q3 + 1.5*IQR] ###
> Lower_bound = summary_mc[2] - 1.5*IQR_mc
> Lower_bound
1st Qu.
  -19.5
>
> Upper_bound = summary_mc[5] + 1.5*IQR_mc
> Upper_bound
3rd Qu.
   48.5
>
> Outlier_country = Motorcycle$County[which(Motorcycle$Fatal.Motorcycle.Accidents < Lower_bound | Motorcycle$Fatal.Motorcycl
e.Accidents > Upper_bound)]
> Outlier_country
[1] "GREENVILLE" "HORRY"
>
```

Greenville and Horry counties are considered outliers because the number of fatal motorcycle accidents in these counties, 51 and 60, falls outside of the range [-19.5, 48.5]

It's possible that Greenville and Horry have the highest numbers of motorcycle fatalities in South Carolina because motorcyclists in those counties do not wear helmets. In South Carolina, only riders under age 21 are required to wear a helmet. Also, it's possible the roads in these counties were in bad condition causing fatal accidents.