

Mini Project # 2

Names of group members: (Group-18)

Deepika Mamidipelly

Preethi Pasunuri

Contribution of each group member

Both group members contributed equally to the inputs for both questions and best is chosen among them to solve these problems. Collaboratively learned R, ran the scripts, and assessed the results. Some of the scripts written by Deepika were analysed and finalized by Preethi and similarly scripts written by Preethi were assessed and finalized by Deepika. Both group members distribute equal amounts of report documentation, which is then integrated into a single final document. Members of the group worked diligently to meet all the project criteria.

Question 1:

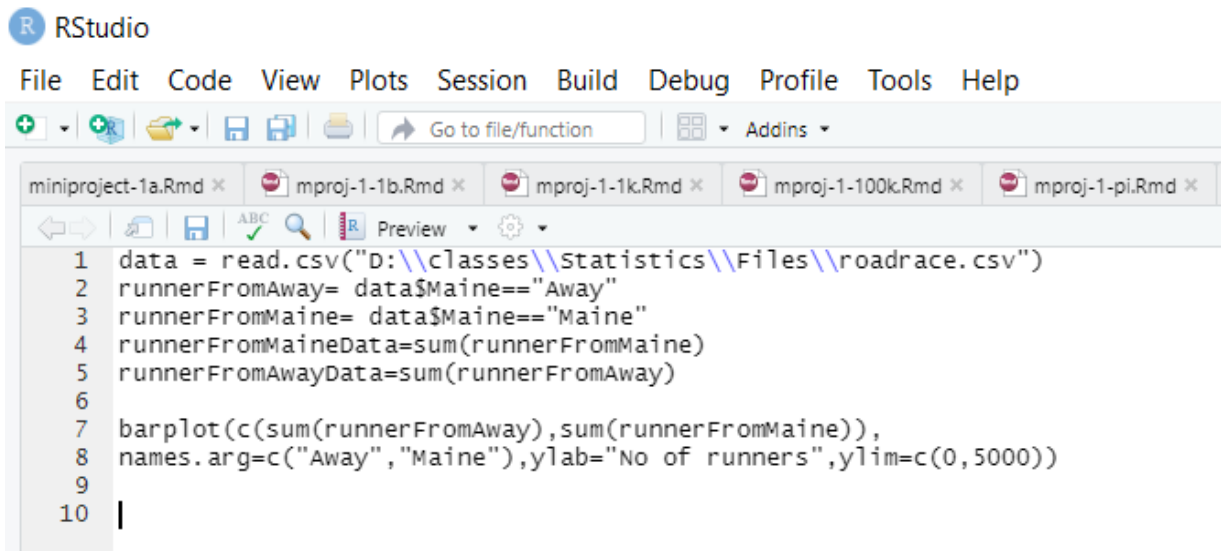
1) Solution:

Firstly, we downloaded the **roadrace.csv** file from eLearning and saved it in the “D:\\classes\\statistics\\Files\\roadrace.csv” path. Then, read the file in R code using the file location by command **read.csv**.

- a) We extracted the data of the runners whether they are from “Maine” or “Away (Other place)” from the Maine column in the roadrace.csv file into the respective variables.

Then we computed the number of runners from the Maine and Away separately.
We Used the bar plot function to plot the bar graph of the data obtained.

Rcode Input:

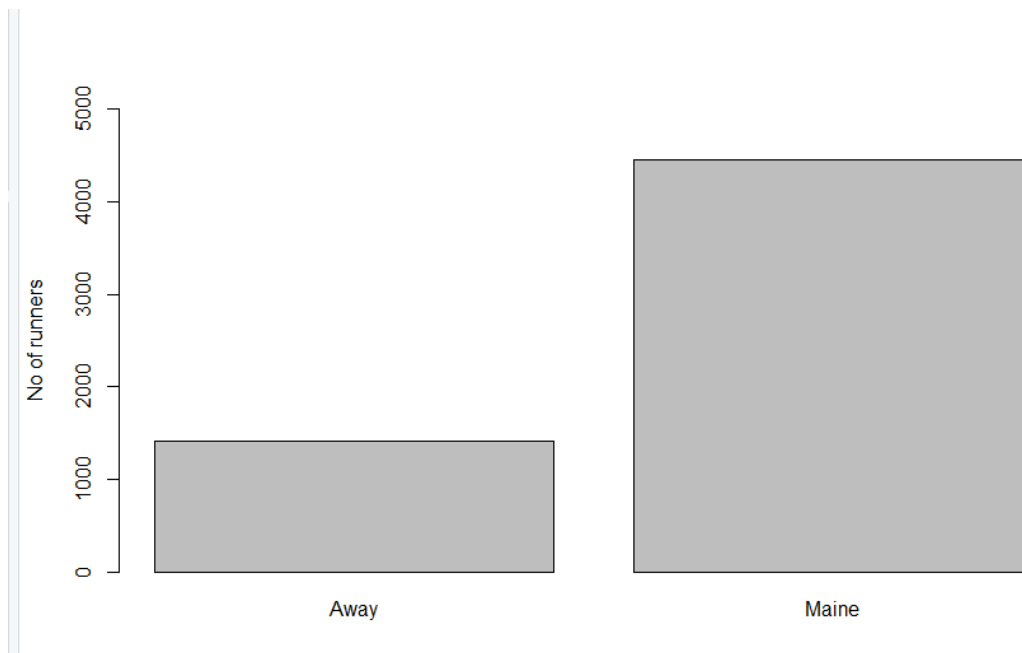


```
1 data = read.csv("D:\\classes\\statistics\\Files\\roadrace.csv")
2 runnerFromAway= data$Maine=="Away"
3 runnerFromMaine= data$Maine=="Maine"
4 runnerFromMaineData=sum(runnerFromMaine)
5 runnerFromAwayData=sum(runnerFromAway)
6
7 barplot(c(sum(runnerFromAway),sum(runnerFromMaine)),
8 names.arg=c("Away","Maine"),ylab="No of runners",ylim=c(0,5000))
9
10 |
```

Output:

Environment	History	Connections	Tutorial
Import Dataset 74 MiB			
R Global Environment			
Data			
data	5875 obs. of 12 variables		
Values			
runnerFromAway	logi [1:5875] TRUE TRUE TRUE TRUE TRUE TRUE ...		
runnerFromAwayData	1417L		
runnerFromMaine	logi [1:5875] FALSE FALSE FALSE FALSE FALSE FALSE ...		
runnerFromMaineData	4458L		

BAR GRAPH



RCode:

#read the file using the file location

```
data = read.csv("D:\\classes\\Statistics\\Files\\roadrace.csv")
```

#extracting the data of Maine column from read.csv file into variables

```
runnerFromAway= data$Maine=="Away"
```

```
runnerFromMaine= data$Maine=="Maine"
```

#computing the values of Away and Maine from Maine column using Rcode

```
runnerFromMaineData=sum(runnerFromMaine)
```

```
runnerFromAwayData=sum(runnerFromAway)
```

#plotting the computed values using barplot.

```
barplot(c(sum(runnerFromAway),sum(runnerFromMaine)),  
names.arg=c("Away","Maine"),ylab="No of runners",ylim=c(0,5000))
```

Observations:

We observed that the majority of runners are from Maine. 1417 runners are from Away which constitutes for 24.11 % of the total runners and Maine group constitutes 75.8% which is 4458 runners of total 5875 runners.

b) Rcode Input:

```
1 data = read.csv("D:\\classes\\Statistics\\Files\\roadrace.csv")  
2  
3 AwayGroupRunnerTime= data[which(data$Maine=="Away"),]$Time.minutes  
4 MaineGroupRunnerTime= data[which(data$Maine=="Maine"),]$Time.minutes  
5  
6 hist(AwayGroupRunnerTime,xlab="Away",ylab="No of  
runners",xlim=c(0,200),ylim=c(0,2500))  
7 hist(MainGroupRunnerTime,xlab="Maine",ylab="No of  
runners",xlim=c(0,200),ylim=c(0,2500))  
8  
9 summary(AwayGroupRunnerTime)  
10 summary(MainGroupRunnerTime)  
11  
12 sd(MainGroupRunnerTime)  
13 sd(AwayGroupRunnerTime)  
14  
15 range(MainGroupRunnerTime)  
16 range(AwayGroupRunnerTime)  
17  
18 IQR(MainGroupRunnerTime)  
19 IQR(AwayGroupRunnerTime)
```

Output:

```
> hist(AwayGroupRunnerTime,xlab= Away ,ylab= No of runners ,xlim=c(0,200),ylim= (0,2500))  
> summary(AwayGroupRunnerTime)  
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
 27.78  49.15   56.92   57.82  64.83  133.71   
> summary(MainGroupRunnerTime)  
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
 30.57  50.00   57.03   58.20  64.24  152.17   
>  
> sd(MainGroupRunnerTime)  
[1] 12.18511  
> sd(AwayGroupRunnerTime)  
[1] 13.83538  
>  
> range(MainGroupRunnerTime)  
[1] 30.567 152.167  
> range(AwayGroupRunnerTime)  
[1] 27.782 133.710  
>  
> IQR(MainGroupRunnerTime)  
[1] 14.24775  
> IQR(AwayGroupRunnerTime)  
[1] 15.674  
>
```

Project: (None)

Environment

History

Connections

Tutorial

Import Dataset

117 MiB

List

R

Global Environment

Data

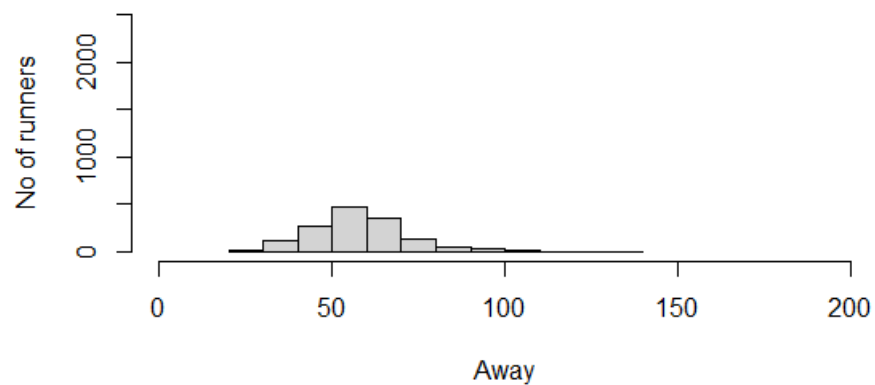
data5875 obs. of 12 variables

Values

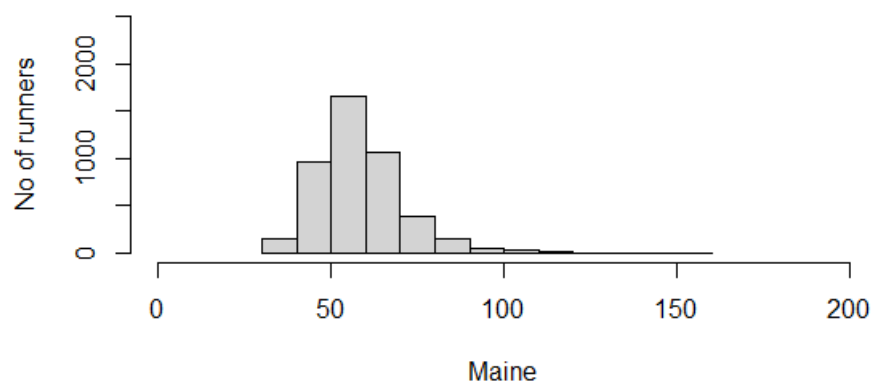
AwayGroupRunnerTi...	num [1:1417]	27.8 27.9 28 28 28.2 ...
MaineGroupRunnerT...	num [1:4458]	30.6 30.7 31.5 31.9 32.1 ...

Histograms:

Histogram of AwayGroupRunnerTime



Histogram of MaineGroupRunnerTime



Rcode:**#read the file using file location**

```
data = read.csv("D:\\classes\\Statistics\\Files\\roadrace.csv")
```

#Extracting runners time of the runners who are from Maine and from other places using which() into variables

```
AwayGroupRunnerTime= data[which(data$Maine=="Away"),]$Time.minutes
```

```
MaineGroupRunnerTime= data[which(data$Maine=="Maine"),]$Time.minutes
```

#using hist() to plot histograms of the data obtained

```
hist(AwayGroupRunnerTime,xlab="Away",ylab="Noof  
runners",xlim=c(0,200),ylim=c(0,2500))
```

```
hist(MaineGroupRunnerTime,xlab="Maine",ylab="Noof  
runners",xlim=c(0,200),ylim=c(0,2500))
```

#using summary() to get summary statistics of the data

```
summary(AwayGroupRunnerTime)
```

```
summary(MaineGroupRunnerTime)
```

```
sd(MaineGroupRunnerTime)
```

```
sd(AwayGroupRunnerTime)
```

```
range(MaineGroupRunnerTime)
```

```
range(AwayGroupRunnerTime)
```

```
IQR(MaineGroupRunnerTime)
```

```
IQR(AwayGroupRunnerTime)
```

Observation:

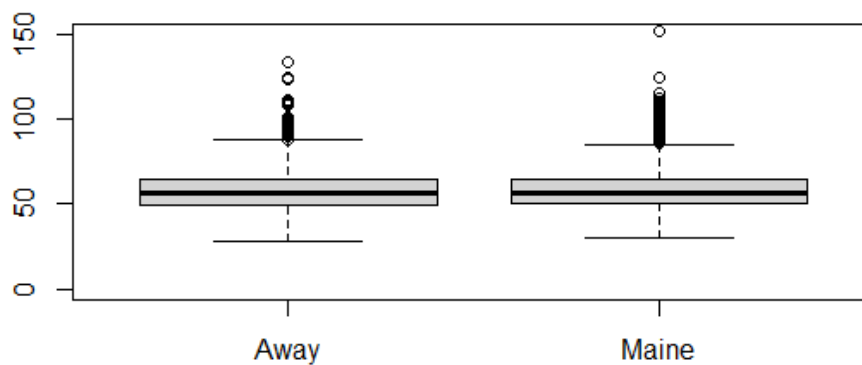
When we plot median on the histograms, we observed that data is little bit right skewed. We observed from the summary statistics of both groups that the minimum, Q1, median, mean, max are higher for the Maine group and Q3, Standard Deviation, Inter Quartile Range(IQR) are higher for away group.

c)Created box plot for the runners time using boxplot() in r

Rcode Input:

```
1 data = read.csv("D:\\classes\\Statistics\\Files\\roadrace.csv")
2
3 AwayGroupRunnerTime= data[which(data$Maine=="Away"),]$Time.minutes
4 MaineGroupRunnerTime= data[which(data$Maine=="Maine"),]$Time.minutes
5
6 boxplot(AwayGroupRunnerTime,MaineGroupRunnerTime,names=c("Away","Maine"),
7 ylim=c(0,150))
```

Boxplot:



RCode:

#read the file using file location

```
data = read.csv("D:\\classes\\Statistics\\Files\\roadrace.csv")
```

#Extracting runners time of the runners who are from Maine and from other places using which() into variables

```
AwayGroupRunnerTime= data[which(data$Maine=="Away"),]$Time.minutes
```

```
MaineGroupRunnerTime= data[which(data$Maine=="Maine"),]$Time.minutes
```

#using boxplot() to plot box plot

```
boxplot(AwayGroupRunnerTime,MaineGroupRunnerTime,names=c("Away","Maine"),
ylim=c(0,150))
```

d)Now, we need to create box plot for runners age based on their gender

Rcode Input:

```
1 data = read.csv("D:\\classes\\Statistics\\Files\\roadrace.csv")
2
3 maleRunnersAge= data$Age[which(data$Sex=="M")]
4 maleRunnersAgeNumeric=as.numeric(unlist(maleRunnersAge))
5 femaleRunnersAge= data$Age[which(data$Sex=="F")]
6 femaleRunnersAgeNumeric=as.numeric(unlist(femaleRunnersAge))
7
8 boxplot(maleRunnersAgeNumeric,femaleRunnersAgeNumeric,names=c("MaleRunners",
9 "FemaleRunners"),ylab="No. of runners")
10
11 summary(femaleRunnersAgeNumeric)
12 summary(maleRunnersAgeNumeric)
13
14 sd(femaleRunnersAge)
15 sd(maleRunnersAge)
16
17 range(femaleRunnersAge)
18 range(maleRunnersAge)
19
20 IQR(femaleRunnersAge)
21 IQR(maleRunnersAge)
```

Output:

```
> boxplot(maleRunnersAgeNumeric,femaleRunnersAgeNumeric,names=c("MaleRunners",
+ "FemaleRunners"),ylab="No. of runners")
>
> summary(femaleRunnersAgeNumeric)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.00  28.00   36.00  37.24  46.00   86.00
> summary(maleRunnersAgeNumeric)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  9.00  30.00   41.00  40.45  51.00   83.00
>
> sd(femaleRunnersAge)
[1] 12.26925
> sd(maleRunnersAge)
[1] 13.99289
>
> range(femaleRunnersAge)
[1] "10" "86"
> range(maleRunnersAge)
[1] "10" "9"
>
> IQR(femaleRunnersAge)
[1] 18
> IQR(maleRunnersAge)
[1] 21
>
```

Project: (None)

Environment History Connections Tutorial

Import Dataset 134 MiB

R Global Environment

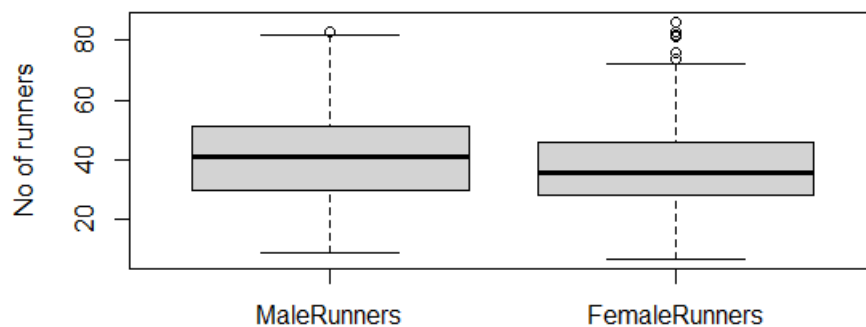
Data

data 5875 obs. of 12 variables

Values

femaleRunnersAge	chr [1:2951]	"25" "25" "23" "26" "32" "29" "31" ...
femaleRunnersAgeN...	num [1:2951]	25 25 23 26 32 29 31 30 29 40 ...
maleRunnersAge	chr [1:2923]	"25" "21" "28" "25" "21" "22" "29" ...
maleRunnersAgeNum...	num [1:2923]	25 21 28 25 21 22 29 28 25 33 ...

BoxPlot:



RCode:

#read the file using file location

```
data = read.csv("D:\\classes\\Statistics\\Files\\roadrace.csv")
```

#Extracting runners age based on their gender using which() into variables

```
maleRunnersAge= data$Age[which(data$Sex=="M")]
```

```
maleRunnersAgeNumeric=as.numeric(unlist(maleRunnersAge))
```

```
femaleRunnersAge= data$Age[which(data$Sex=="F")]
```

```
femaleRunnersAgeNumeric=as.numeric(unlist(femaleRunnersAge))
```

#box plotting using boxplot()

```
boxplot(maleRunnersAgeNumeric,femaleRunnersAgeNumeric,names=c("MaleRunners","FemaleRunners"),ylab="No of runners")
```


#calculating summary statistics using summary()

```
summary(femaleRunnersAgeNumeric)
```

```
summary(maleRunnersAgeNumeric)
```

```
sd(femaleRunnersAge)
```

```
sd(maleRunnersAge)
```

```
range(femaleRunnersAge)
```

```
range(maleRunnersAge)
```

```
IQR(femaleRunnersAge)
```

```
IQR(maleRunnersAge)
```

Observation:

Statistical values of the male runners are higher than the female runners.

Question 2:

Initially, started off by reading the motorcycle.csv file into the R code. Storing the Fatal Motorcycle Accidents column values into a variable to access further in the code.

As asked in the question, plotting the boxplot of data using boxplot()

We need to find out the outliers:

A value in the data can be considered as an outlier if it is more than $1.5 \times \text{IQR}$ away from Q1 and Q3 quartiles.

Hence to find minBound, we have

minBound=Q1-1.5*IQR

maxBound=Q3+1.5*IQR

we calculate Q1, Q3 using quantile().

Therefore, any value that lies below the minBound and above the maxBound are considered as outliers.

Rcode Input:

```
1 data= read.csv("D:\\classes\\Statistics\\Files\\motorcycle.csv")
2 noOfFatalAccidents= data$Fatal.Motorcycle.Accidents
3
4 boxplot(noOfFatalAccidents,xlab="Fatal Motorcycle Accidents",ylab="No of
  Motorcycle Accidents")
5
6 minBound=max(min(noOfFatalAccidents),quantile(noOfFatalAccidents,prob=0.25)
7             -(1.5*IQR(noOfFatalAccidents)))
8 maxBound=min(max(noOfFatalAccidents),
9             (quantile(noOfFatalAccidents,prob=0.75)+(1.5*IQR(noOfFatalAccidents)))
10
11 countyWithMoreFatalAccidents=data[which(data$Fatal.Motorcycle.Accidents
12     <minBound | data$Fatal.Motorcycle.Accidents>maxBound),]$County
13
14 countyWithMoreFatalAccidents
15 summary(noOfFatalAccidents)
16 sd(noOfFatalAccidents)
17 range(noOfFatalAccidents)
18 IQR(noOfFatalAccidents)
```

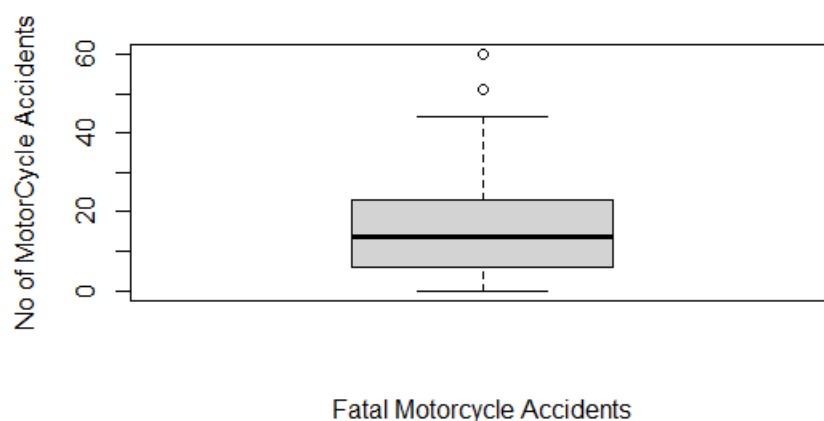
Output:

```
+ <minBound | data$Fatal.Motorcycle.Accidents>maxBound),]$County
>
> countyWithMoreFatalAccidents
[1] "GREENVILLE" "HORRY"
> summary(noOfFatalAccidents)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   6.00   13.50   17.02   23.00   60.00
> sd(noOfFatalAccidents)
[1] 13.81256
> range(noOfFatalAccidents)
[1] 0 60
> IQR(noOfFatalAccidents)
[1] 17
>
```

Project: (None)

Environment	History	Connections	Tutorial
<div> <div>Import Dataset</div> <div>135 MiB</div> <div>List</div> </div>			
<div> <div>R</div> <div>Global Environment</div> <div></div> </div>			
Data			
data	48 obs. of 2 variables		
Values			
countyWithMoreFa...	chr [1:2] "GREENVILLE" "HORRY"		
maxBound	48.5		
minBound	0		
noOfFatalAcciden...	int [1:48] 3 28 3 35 3 7 13 38 6 44 ...		

Boxplot:



Rcode:

#read the file using file location

```
data= read.csv("D:\\classes\\Statistics\\Files\\motorcycle.csv")
```

#extracting values of fatal accidents column for each county into a variable

```
noOfFatalAccidents= data$Fatal.Motorcycle.Accidents
```

#Plotting using boxplot()

```
boxplot(noOfFatalAccidents,xlab="Fatal Motorcycle Accidents",ylab="No of MotorCycle Accidents")
```

#calculating the lower and upper bounds of the data

```
minBound=max(min(noOfFatalAccidents),quantile(noOfFatalAccidents,prob=0.25)
(1.5*IQR(noOfFatalAccidents)))
```

```
maxBound=min(max(noOfFatalAccidents),
```

```
(quantile(noOfFatalAccidents,prob=0.75))+(1.5*IQR(noOfFatalAccidents)))
```

#finding the county which are outliers.

```
countyWithMoreFatalAccidents=data[which(data$Fatal.Motorcycle.Accidents  
    <minBound | data$Fatal.Motorcycle.Accidents>maxBound),]$County  
countyWithMoreFatalAccidents
```

#Calculating summary statistics of the data

```
summary(noOfFatalAccidents)  
sd(noOfFatalAccidents)  
range(noOfFatalAccidents)  
IQR(noOfFatalAccidents)
```

Observations:

The average number of fatal motorcycle accidents that occurred in south Carolina in 2009 is approximately 17.

The highest no of accidents occurred are 60 and there are counties with 0 fatal motorcycle accidents as well in the data.

As the minimum no of accidents occurred is 0, If there are any outliers below the lower bound, then the no of accidents occurred in that county will be negative value which is not possible. So the outliers we have, are above the upper bound and also are the counties with the highest number of fatal motorcycle accidents.

So the Counties with highest number of fatal motorcycle accidents are Greenville and Horry. The reasons for highest number of fatal moto cycle accidents in these counties are because of poor maintenance of roads and people not following traffic rules.