CS-6313 – Statistical Methods for Data Science

Mini Project #2
Group No - 5
Manan Dalal (MUD200000)
Lipi Patel (LDP210000)

Contribution of Team Members:

Worked together with each other to solve both the problems. Manan worked on the script and made them efficient and readable. Lipi worked on documenting the work and both worked together on checking and giving the final touch to 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 bar plot function for this. What can we conclude from the plot? Back up your conclusions with relevant summary statistics.**

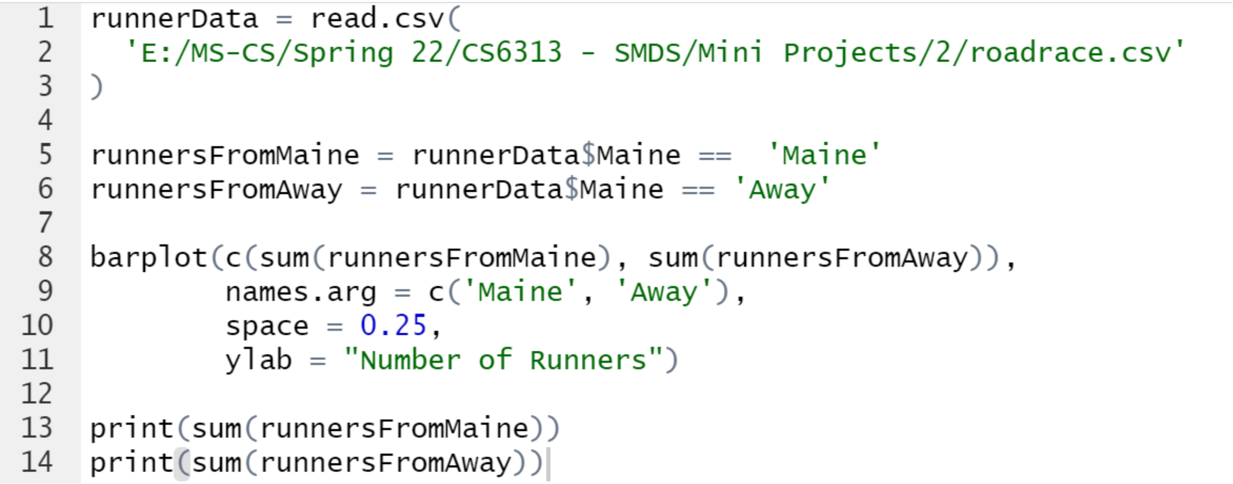
Solution

⇨ Firstly, we need to read the CSV file using read.csv function.

⇨ Then by using the which operator on the column 'Maine', we filtered out the runners that were from Maine and those who were not.

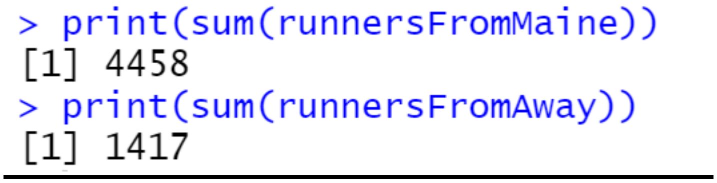
⇨ Then, we feed the number of runners from and not from main to the bar plot function to create a bar graph.

Code Snippet



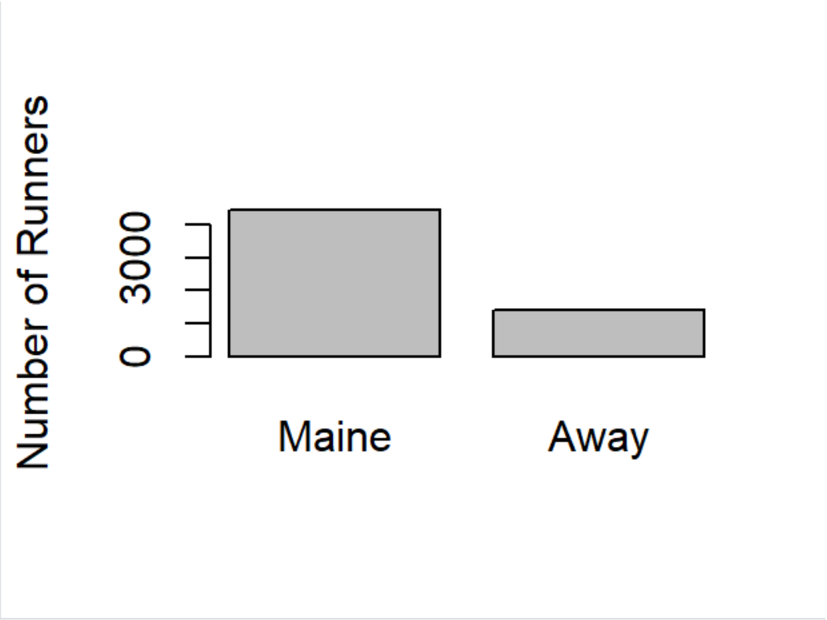
```r
runnerData = read.csv(
  'E:/MS-CS/Spring 22/CS6313 - SMDS/Mini Projects/2/roadrace.csv'
)

runnersFromMaine = runnerData$Maine ==  'Maine'
runnersFromAway = runnerData$Maine == 'Away'

barplot(c(sum(runnersFromMaine), sum(runnersFromAway)),
        names.arg = c('Maine', 'Away'),
        space = 0.25,
        ylab = "Number of Runners")

print(sum(runnersFromMaine))
print(sum(runnersFromAway))
```

Output



```
> print(sum(runnersFromMaine))
[1] 4458
> print(sum(runnersFromAway))
[1] 1417
```

Bar Graph:



Observations

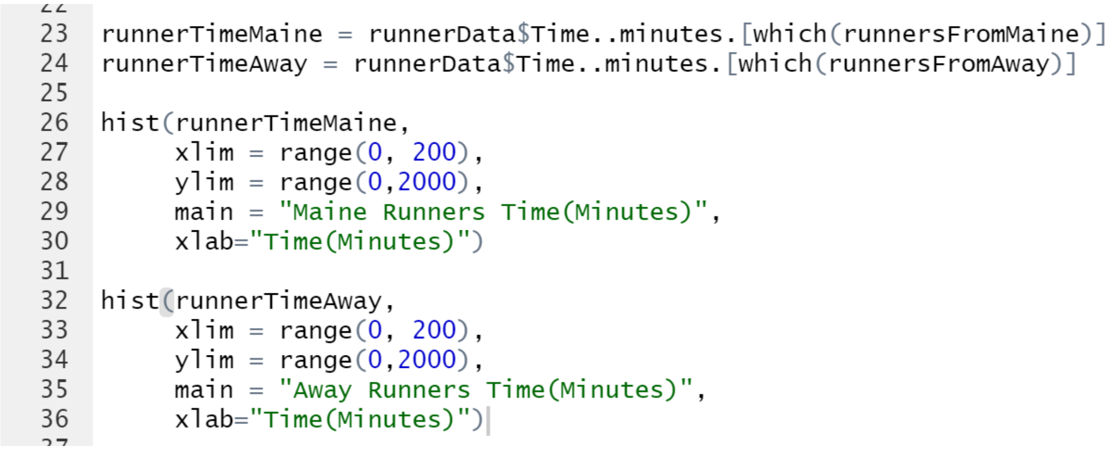
- From the bar graphs it can be concluded that the Maine group is greater than the total number of runners from the away group.
- It can be concluded that the Maine group account for 75.8% of the portion while the away group account for 24.2% of the portion out of a total 5875 runners.

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.

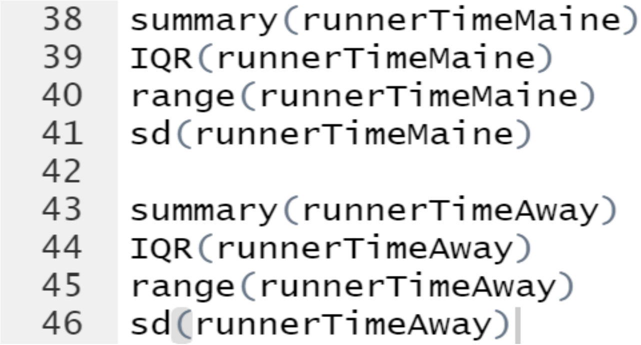
Solution

- ⇨ Firstly, we extract the Time. Minutes column and filter it on the basis of the Maine column.
- ⇨ Then, we create 2 histograms, one for Maine and one for Away. We set the range of minutes to [0, 200] and the frequency to [0, 2000] for both the graphs.

Code Snippet



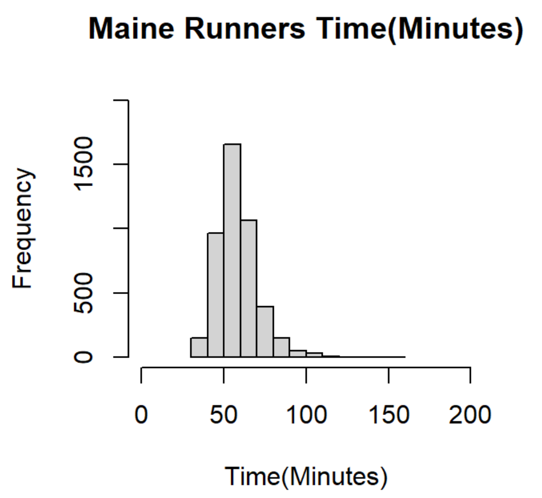
```
runnerTimeMaine = runnerData$Time..minutes.[which(runnersFromMaine)]
runnerTimeAway = runnerData$Time..minutes.[which(runnersFromAway)]

hist(runnerTimeMaine,
     xlim = range(0, 200),
     ylim = range(0,2000),
     main = "Maine Runners Time(Minutes)",
     xlab="Time(Minutes)")

hist(runnerTimeAway,
     xlim = range(0, 200),
     ylim = range(0,2000),
     main = "Away Runners Time(Minutes)",
     xlab="Time(Minutes)")
```

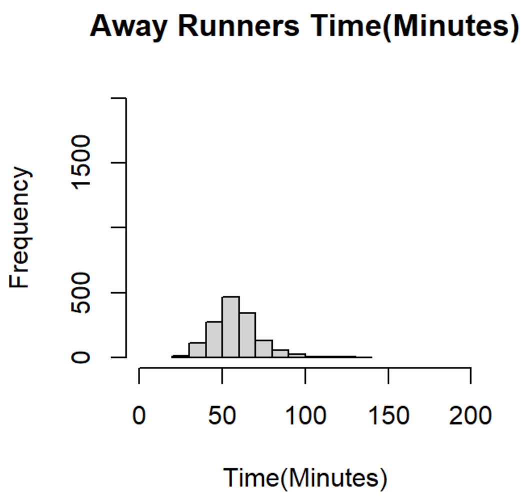


```
summary(runnerTimeMaine)
IQR(runnerTimeMaine)
range(runnerTimeMaine)
sd(runnerTimeMaine)

summary(runnerTimeAway)
IQR(runnerTimeAway)
range(runnerTimeAway)
sd(runnerTimeAway)
```

Output

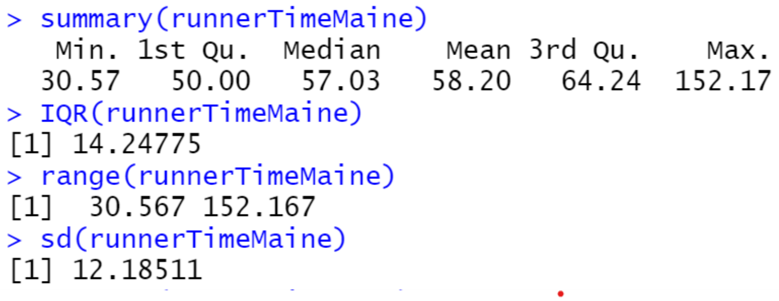
Histograms





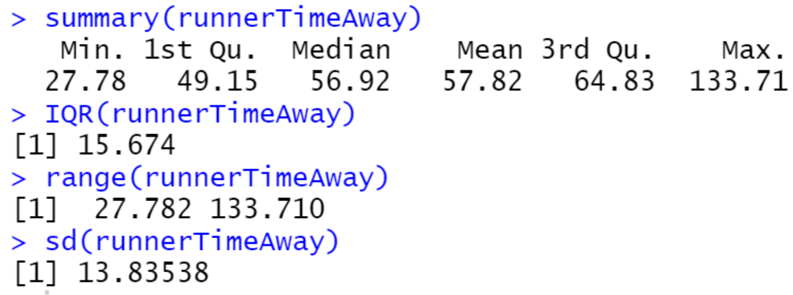
Statistics

⇨ For Maine Runners



```
> summary(runnerTimeMaine)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  30.57   50.00   57.03   58.20   64.24  152.17 
> IQR(runnerTimeMaine)
[1] 14.24775
> range(runnerTimeMaine)
[1]  30.567 152.167
> sd(runnerTimeMaine)
[1] 12.18511
```

⇨ For Away Runners



```
> summary(runnerTimeAway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  27.78   49.15   56.92   57.82   64.83  133.71 
> IQR(runnerTimeAway)
[1] 15.674
> range(runnerTimeAway)
[1]  27.782 133.710
> sd(runnerTimeAway)
[1] 13.83538
```

Observations

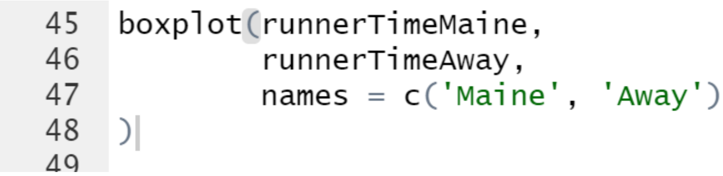
- ⇨ From the graphs it is evident that both distributions and skewed to the right.
- ⇨ By looking at the statistics, we can observe that the values of min, 1st Q, median, mean and max are higher for the Maine group but 3rd Q, IQR and Standard deviation are higher for the Away group.

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

Solution

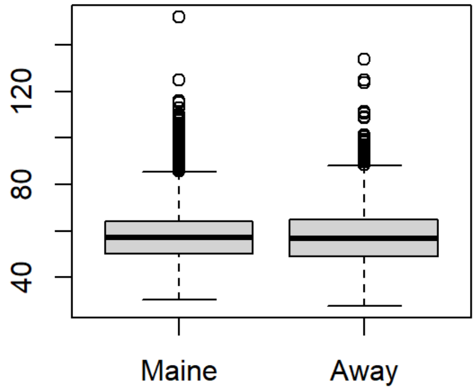
- ⇨ We create a single graph for both box plots using the boxplot function.

Code Snippet



```r
boxplot(runnerTimeMaine,
        runnerTimeAway,
        names = c('Maine', 'Away')
)
```

Output

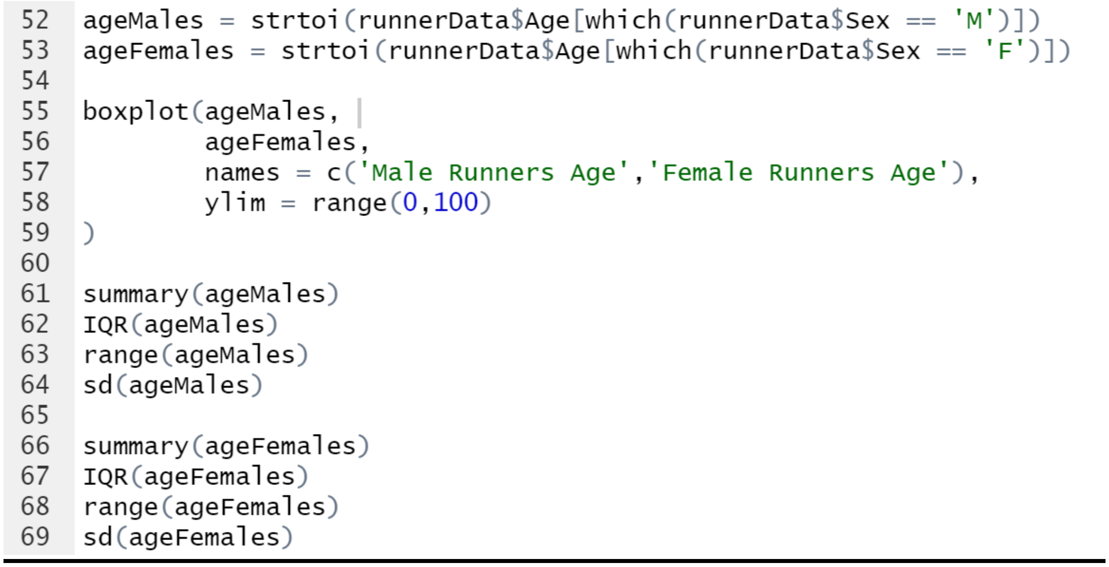


d) Create side-by-side boxplots for the runners' ages (given in years) for male and female runners. What can we conclude about the two distributions? Back up your conclusions with relevant summary statistics, including mean, standard deviation, range, median, and interquartile range.

Solution

- For generating the desired box plot, we first extract the Age column and filter it through the Sex column into two separate variables.
- Then, we create a box plot for both these variables.

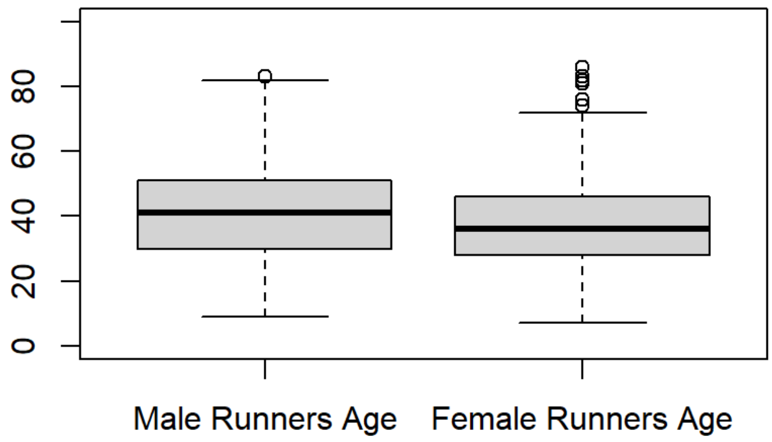
Code Snippet



```
ageMales = strtoi(runnerData$Age[which(runnerData$Sex == 'M')])
ageFemales = strtoi(runnerData$Age[which(runnerData$Sex == 'F')])

boxplot(ageMales,
        ageFemales,
        names = c('Male Runners Age','Female Runners Age'),
        ylim = range(0,100)
)

summary(ageMales)
IQR(ageMales)
range(ageMales)
sd(ageMales)

summary(ageFemales)
IQR(ageFemales)
range(ageFemales)
sd(ageFemales)
```

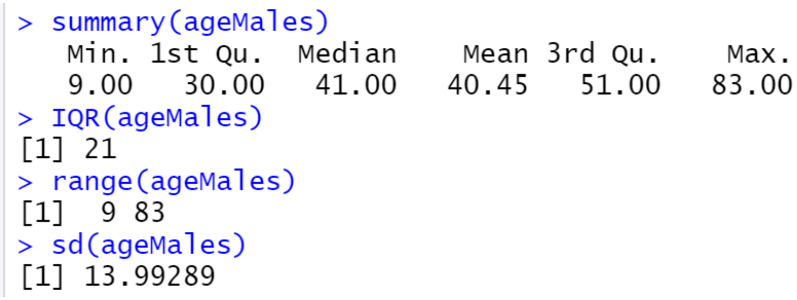
Output

Boxplot



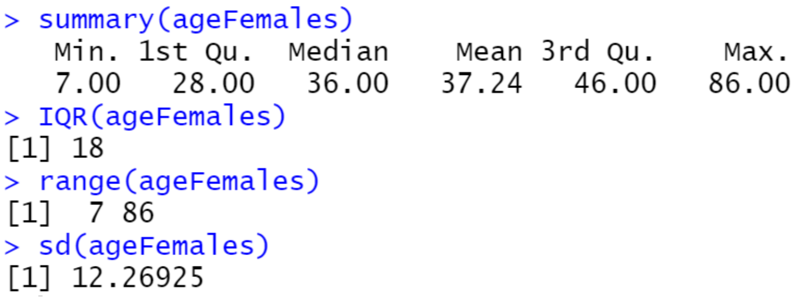
Statistics

⇨ For Male Runners



```
> summary(ageMales)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   30.00   41.00   40.45   51.00   83.00
> IQR(ageMales)
[1] 21
> range(ageMales)
[1]  9 83
> sd(ageMales)
[1] 13.99289
```

⇨ For Female Runners



```
> summary(ageFemales)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   28.00   36.00   37.24   46.00   86.00
> IQR(ageFemales)
[1] 18
> range(ageFemales)
[1]  7 86
> sd(ageFemales)
[1] 12.26925
```

Observations

- ⇨ The statistical values of males are higher than those of females.
- ⇨ Females who are of age 80 and up seem to be active participants in races.

Question-2

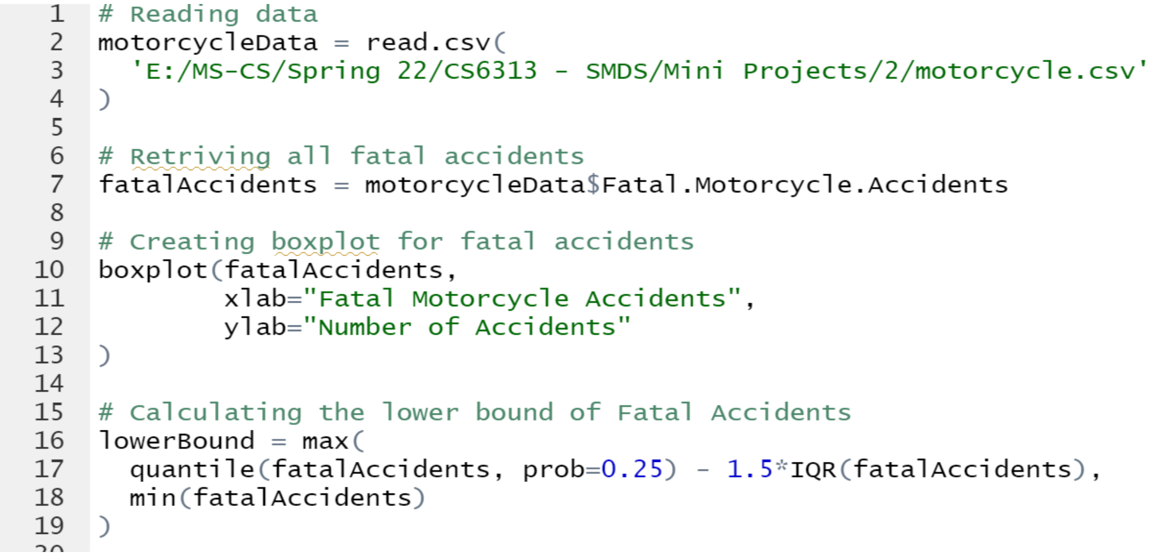
Consider the dataset motorcycle.csv posted on eLearning. It contains the number of fatal motorcycle accidents that occurred in each county of South Carolina during 2009. Create a boxplot of data and provide relevant summary statistics. Discuss the features of the data distribution. Identify which counties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

Solution

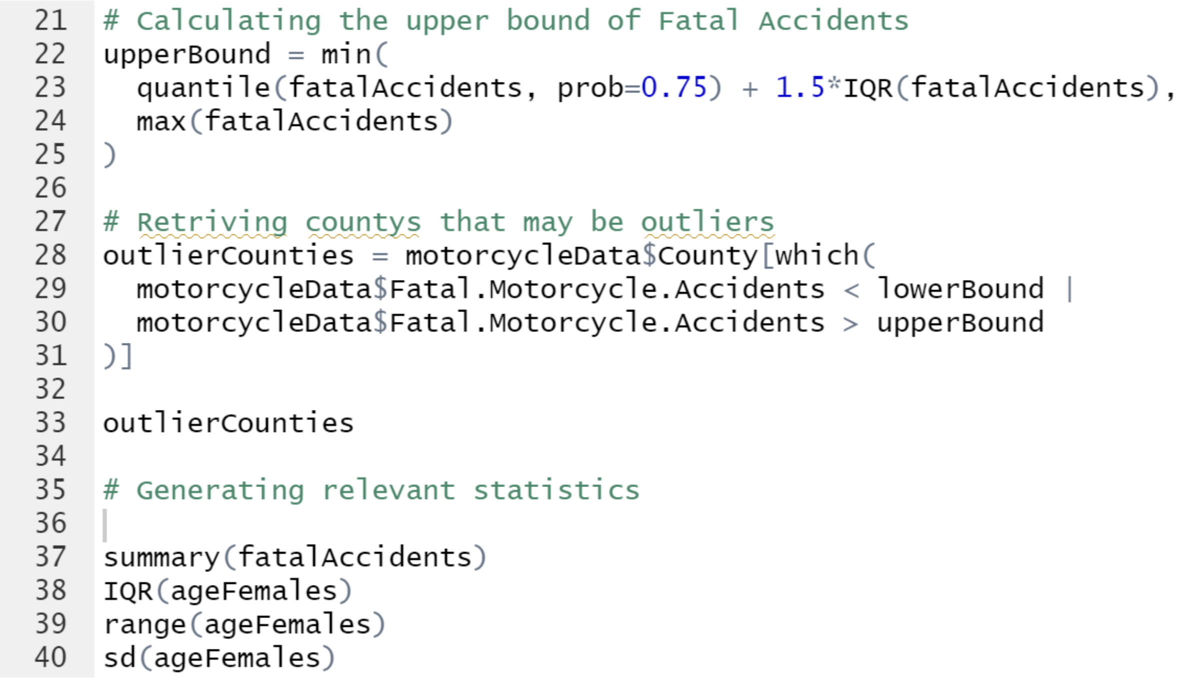
- ⇨ Firstly, we would read the data from the file and retrieve all the fatal accidents.
- ⇨ They, we create a boxplot to visualize the distribution of data.
- ⇨ In, order to find outliers, we must first find the lower and upper bounds of the distribution.
- ⇨ A value is an outlier if it’s more than 1.5*IQR away from the 25th and 75th quantiles.
- ⇨ Hence lower bound: 25th percentile - 1.5IQR and upper bound: 75th percentile + 1.5IQR.

- ⇨ Now, any value of accidents lower than the lower bound or higher than the upper bound are outliers.

Code Snippet



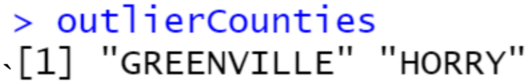
```r
# Reading data
motorcycleData = read.csv(
  'E:/MS-CS/Spring 22/CS6313 - SMDS/Mini Projects/2/motorcycle.csv'
)

# Retriving all fatal accidents
fatalAccidents = motorcycleData$Fatal.Motorcycle.Accidents

# Creating boxplot for fatal accidents
boxplot(fatalAccidents,
        xlab="Fatal Motorcycle Accidents",
        ylab="Number of Accidents"
)

# Calculating the lower bound of Fatal Accidents
lowerBound = max(
  quantile(fatalAccidents, prob=0.25) - 1.5*IQR(fatalAccidents),
  min(fatalAccidents)
)
```



```r
# Calculating the upper bound of Fatal Accidents
upperBound = min(
  quantile(fatalAccidents, prob=0.75) + 1.5*IQR(fatalAccidents),
  max(fatalAccidents)
)

# Retriving countys that may be outliers
outlierCounties = motorcycleData$County[which(
  motorcycleData$Fatal.Motorcycle.Accidents < lowerBound |
  motorcycleData$Fatal.Motorcycle.Accidents > upperBound
)]

outlierCounties

# Generating relevant statistics

summary(fatalAccidents)
IQR(ageFemales)
range(ageFemales)
sd(ageFemales)
```

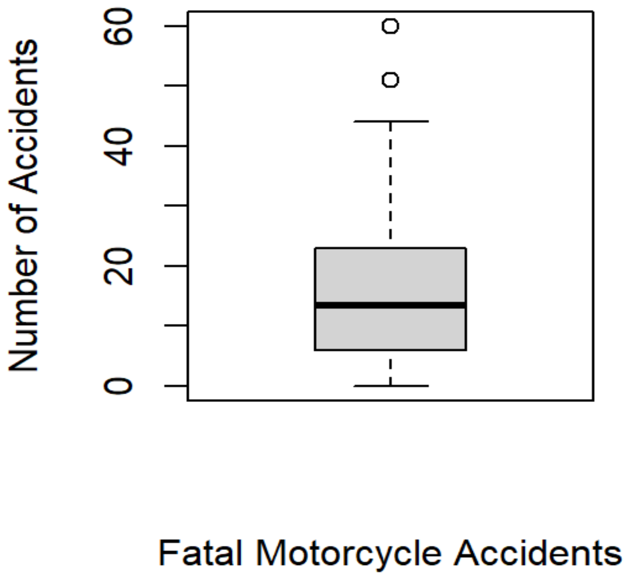
Output

Outlier Counties

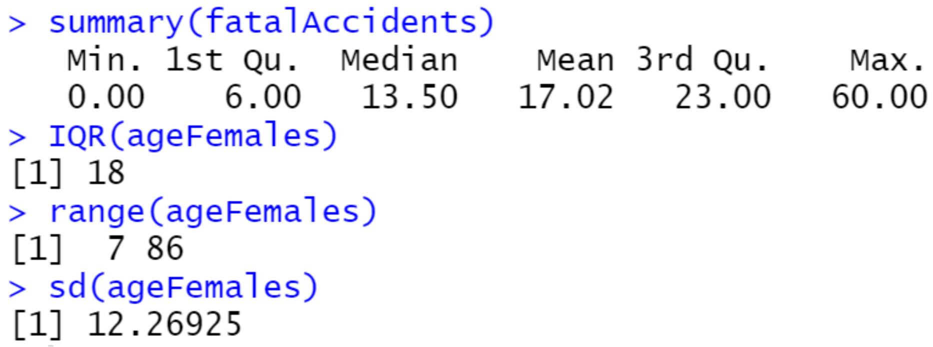


```
> outlierCounties
[1] "GREENVILLE" "HORRY"
```

Boxplot



Statistics



```
> summary(fatalAccidents)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    6.00   13.50   17.02   23.00   60.00 
> IQR(ageFemales)
[1] 18
> range(ageFemales)
[1]  7 86
> sd(ageFemales)
[1] 12.26925
```

Observations

- The counties highest number of motorcycle fatalities are in South Carolina in Greenville Horry.
- The reason for this could be poor road and highway maintenance as well as negligent/reckless drivers.