# **Statistical Methods for Data Science**

Project 2

By: Chirag Shahi cxs180005

Krishnan Vijayaraghavan kxv190006

Contribution:

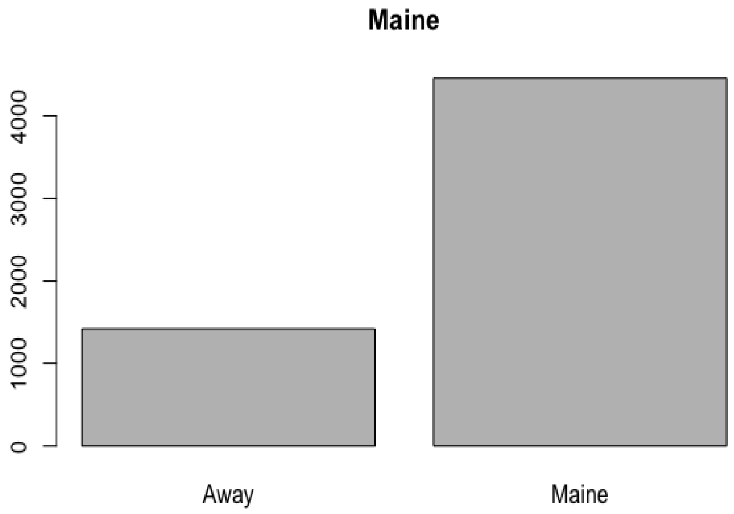
Chirag: Question 1

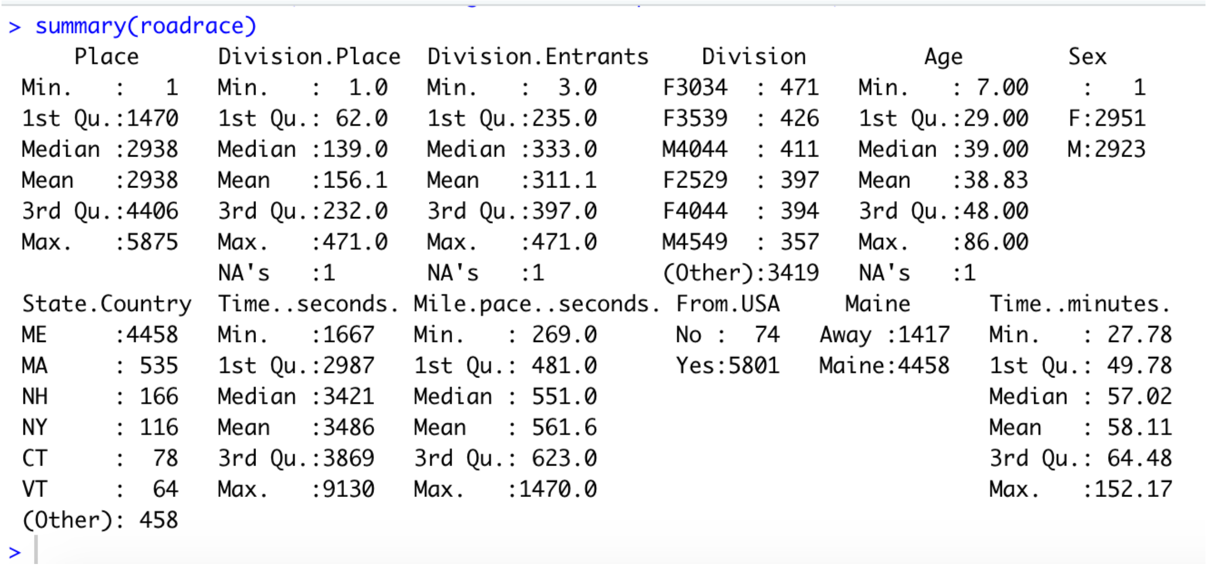
Krishnan: Question 2

# SECTION 1

Answers:

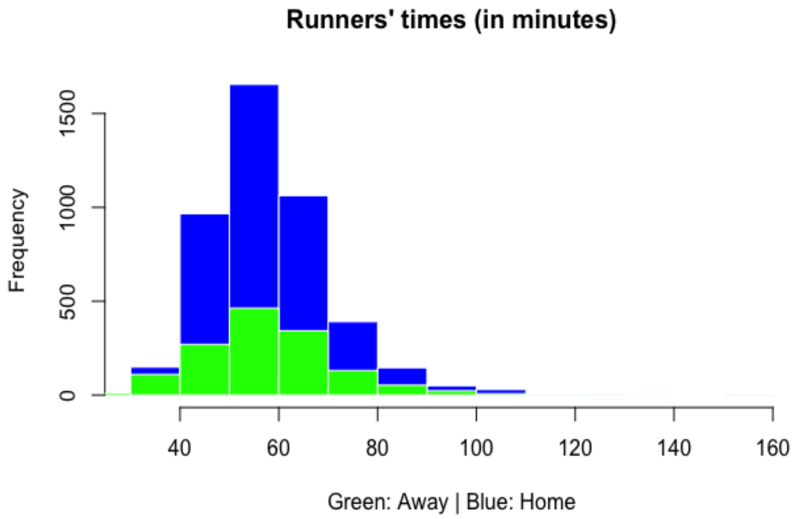
1. (a)

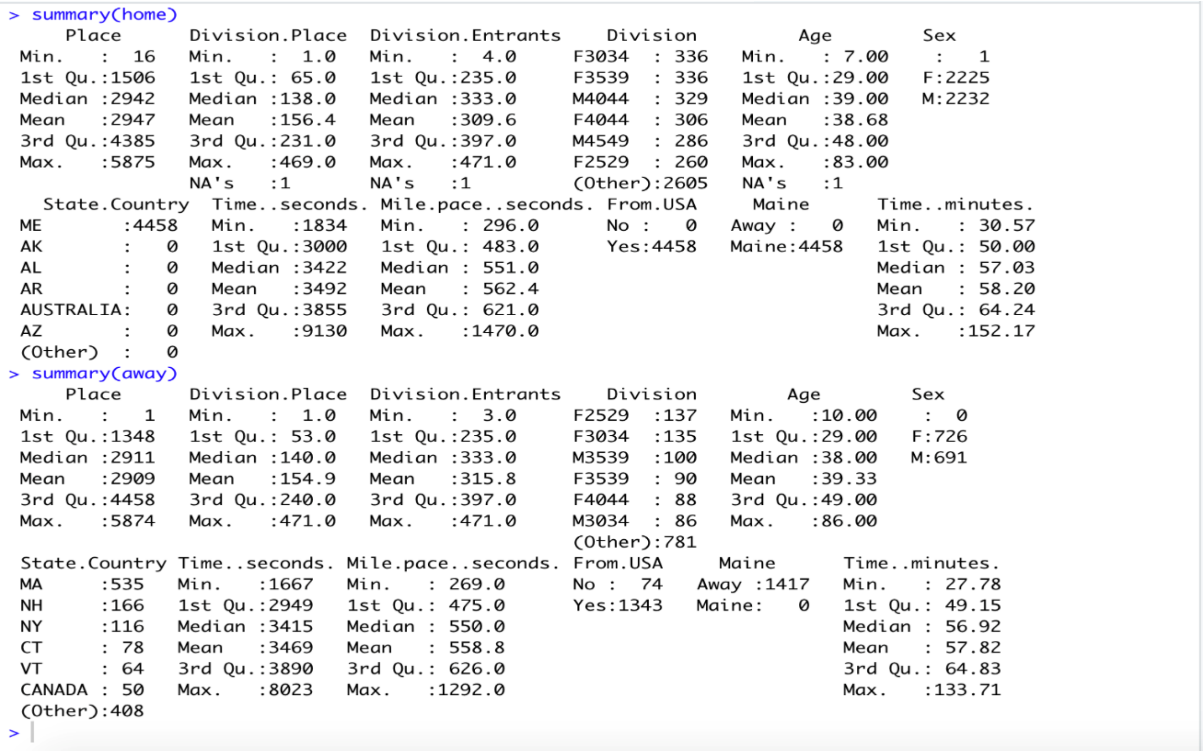




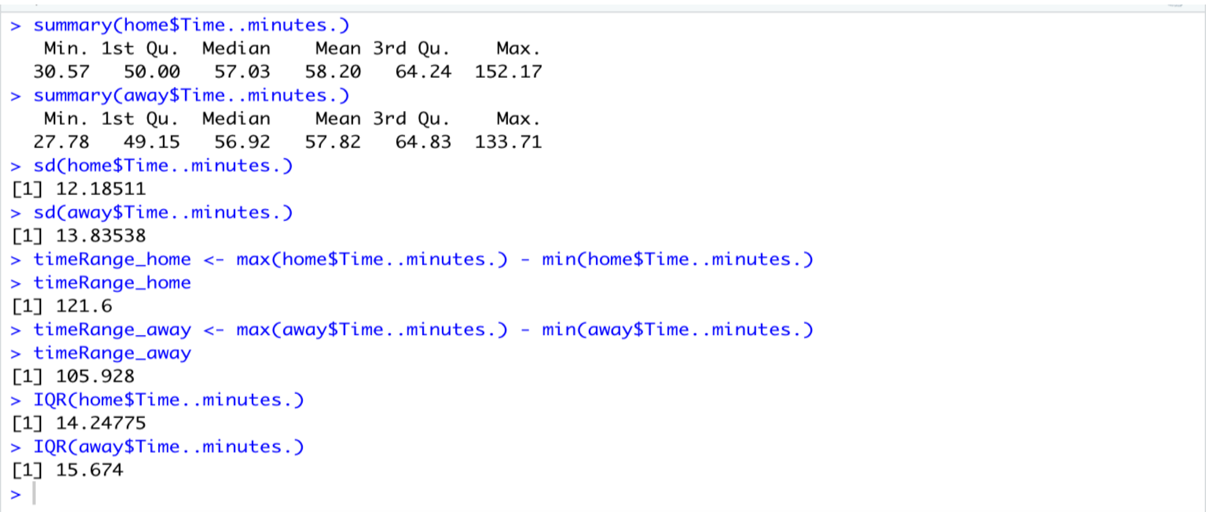
```
> summary(roadrace)
     Place        Division.Place  Division.Entrants    Division          Age          Sex
 Min.   :   1    Min.   :  1.0    Min.   :  3.0     F3034   : 471   Min.   : 7.00    :   1
 1st Qu.:1470    1st Qu.: 62.0    1st Qu.:235.0     F3539   : 426   1st Qu.:29.00   F:2951
 Median :2938    Median :139.0    Median :333.0     M4044   : 411   Median :39.00   M:2923
 Mean   :2938    Mean   :156.1    Mean   :311.1     F2529   : 397   Mean   :38.83
 3rd Qu.:4406    3rd Qu.:232.0    3rd Qu.:397.0     F4044   : 394   3rd Qu.:48.00
 Max.   :5875    Max.   :471.0    Max.   :471.0     M4549   : 357   Max.   :86.00
                 NA's   :1        NA's   :1         (Other):3419    NA's   :1
 State.Country   Time..seconds.  Mile.pace..seconds.  From.USA     Maine        Time..minutes.
 ME      :4458   Min.   :1667   Min.   : 269.0     No :  74   Away :1417    Min.   : 27.78
 MA      : 535   1st Qu.:2987   1st Qu.: 481.0     Yes:5801   Maine:4458    1st Qu.: 49.78
 NH      : 166   Median :3421   Median : 551.0                              Median : 57.02
 NY      : 116   Mean   :3486   Mean   : 561.6                              Mean   : 58.11
 CT      :  78   3rd Qu.:3869   3rd Qu.: 623.0                              3rd Qu.: 64.48
 VT      :  64   Max.   :9130   Max.   :1470.0                              Max.   :152.17
 (Other): 458
>
```

1. b)



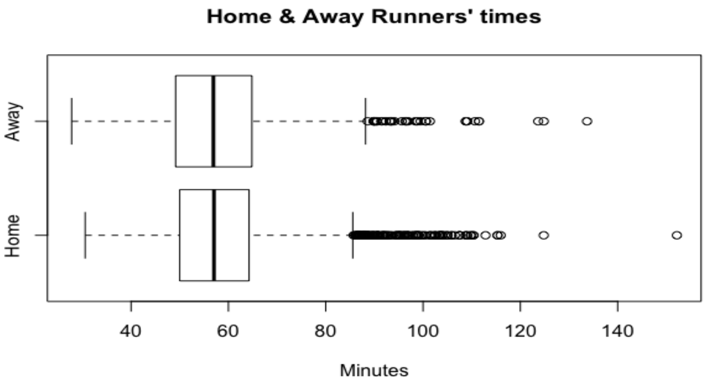


```
> summary(home)
     Place        Division.Place   Division.Entrants     Division          Age           Sex    
 Min.   :  16   Min.   :  1.0   Min.   :  4.0     F3034   : 336   Min.   : 7.00    :   1  
 1st Qu.:1506   1st Qu.: 65.0   1st Qu.:235.0     F3539   : 336   1st Qu.:29.00   F:2225  
 Median :2942   Median :138.0   Median :333.0     M4044   : 329   Median :39.00   M:2232  
 Mean   :2947   Mean   :156.4   Mean   :309.6     F4044   : 306   Mean   :38.68           
 3rd Qu.:4385   3rd Qu.:231.0   3rd Qu.:397.0     M4549   : 286   3rd Qu.:48.00           
 Max.   :5875   Max.   :469.0   Max.   :471.0     F2529   : 260   Max.   :83.00           
                NA's   :1       NA's   :1         (Other):2605   NA's   :1               
   State.Country  Time..seconds.  Mile.pace..seconds.  From.USA      Maine       Time..minutes.  
 ME       :4458   Min.   :1834   Min.   : 296.0       No :   0   Away :   0   Min.   : 30.57  
 AK       :   0   1st Qu.:3000   1st Qu.: 483.0       Yes:4458   Maine:4458   1st Qu.: 50.00  
 AL       :   0   Median :3422   Median : 551.0                               Median : 57.03  
 AR       :   0   Mean   :3492   Mean   : 562.4                               Mean   : 58.20  
 AUSTRALIA:   0   3rd Qu.:3855   3rd Qu.: 621.0                               3rd Qu.: 64.24  
 AZ       :   0   Max.   :9130   Max.   :1470.0                               Max.   :152.17  
 (Other)  :   0                                                                              
> summary(away)
     Place        Division.Place   Division.Entrants     Division         Age           Sex   
 Min.   :   1   Min.   :  1.0   Min.   :  3.0     F2529   :137   Min.   :10.00    :  0  
 1st Qu.:1348   1st Qu.: 53.0   1st Qu.:235.0     F3034   :135   1st Qu.:29.00   F:726  
 Median :2911   Median :140.0   Median :333.0     M3539   :100   Median :38.00   M:691  
 Mean   :2909   Mean   :154.9   Mean   :315.8     F3539   : 90   Mean   :39.33          
 3rd Qu.:4458   3rd Qu.:240.0   3rd Qu.:397.0     F4044   : 88   3rd Qu.:49.00          
 Max.   :5874   Max.   :471.0   Max.   :471.0     M3034   : 86   Max.   :86.00          
                                                  (Other):781                           
 State.Country Time..seconds.  Mile.pace..seconds. From.USA      Maine       Time..minutes.  
 MA     :535   Min.   :1667   Min.   : 269.0      No :  74   Away :1417   Min.   : 27.78  
 NH     :166   1st Qu.:2949   1st Qu.: 475.0      Yes:1343   Maine:   0   1st Qu.: 49.15  
 NY     :116   Median :3415   Median : 550.0                              Median : 56.92  
 CT     : 78   Mean   :3469   Mean   : 558.8                              Mean   : 57.82  
 VT     : 64   3rd Qu.:3890   3rd Qu.: 626.0                              3rd Qu.: 64.83  
 CANADA : 50   Max.   :8023   Max.   :1292.0                              Max.   :133.71  
 (Other):408                                                                             
> 
```

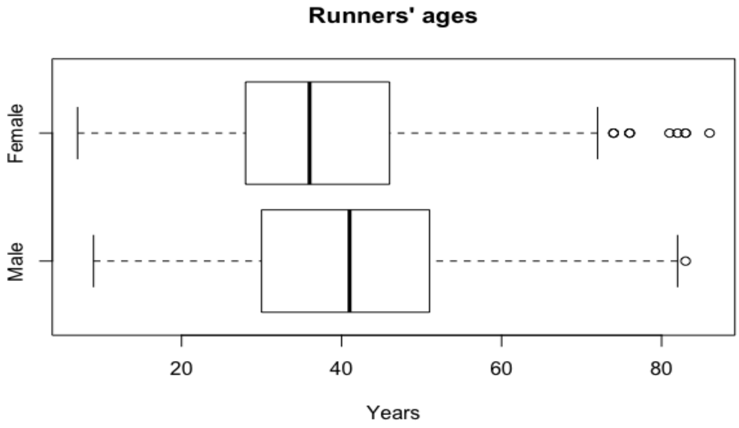


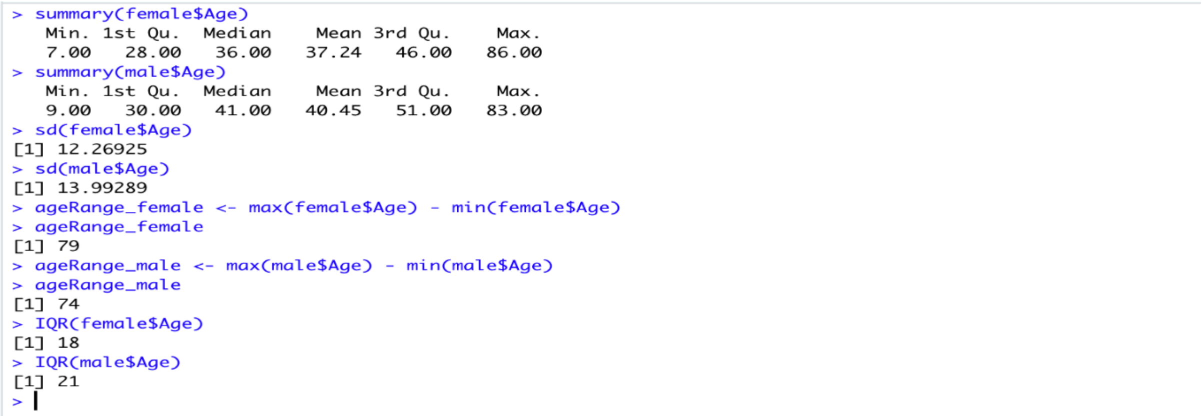
```
> summary(home$Time..minutes.)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  30.57   50.00   57.03   58.20   64.24  152.17 
> summary(away$Time..minutes.)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  27.78   49.15   56.92   57.82   64.83  133.71 
> sd(home$Time..minutes.)
[1] 12.18511
> sd(away$Time..minutes.)
[1] 13.83538
> timeRange_home <- max(home$Time..minutes.) - min(home$Time..minutes.)
> timeRange_home
[1] 121.6
> timeRange_away <- max(away$Time..minutes.) - min(away$Time..minutes.)
> timeRange_away
[1] 105.928
> IQR(home$Time..minutes.)
[1] 14.24775
> IQR(away$Time..minutes.)
[1] 15.674
> 
```

1. (c)



1. (d)

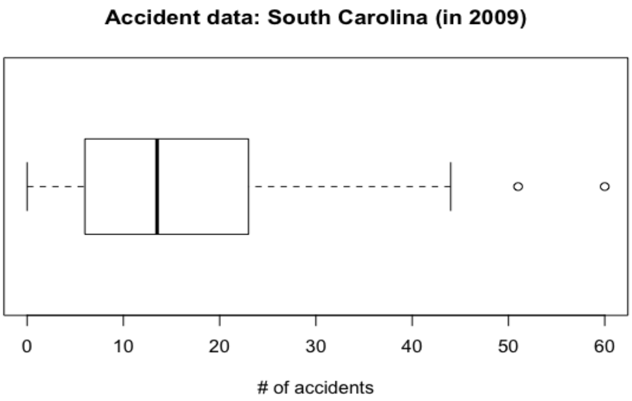


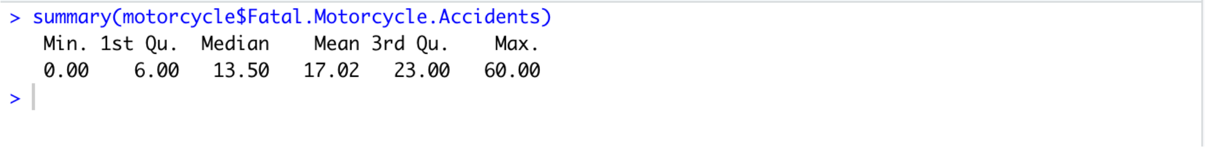


```
> summary(female$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   7.00   28.00   36.00   37.24   46.00   86.00 
> summary(male$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   9.00   30.00   41.00   40.45   51.00   83.00 
> sd(female$Age)
[1] 12.26925
> sd(male$Age)
[1] 13.99289
> ageRange_female <- max(female$Age) - min(female$Age)
> ageRange_female
[1] 79
> ageRange_male <- max(male$Age) - min(male$Age)
> ageRange_male
[1] 74
> IQR(female$Age)
[1] 18
> IQR(male$Age)
[1] 21
> 
```

-

2.





```
> summary(motorcycle$Fatal.Motorcycle.Accidents)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    6.00   13.50   17.02   23.00   60.00 
> 
```

# SECTION 2

1. (a)

roadrace = read.csv("/Users/chiragshahi/Desktop/roadrace.csv")

#reads the data in csv file

summary(roadrace)

#summary function that gives the min, 1st quantile, median, mean, 3rd quantile and max of various attributes in roadrace dataset

maine <- table(roadrace$Maine)

#Frequency table with types of Maine variable

barplot(maine, main="Maine")

```
#Barplot function used to create a bar graph
```

The number of runners participating are more from Maine than anywhere else.

1. (b)

```
home <- roadrace[roadrace$Maine == "Maine",]
# Storing values Maine of Maine variable in home
away <- roadrace[roadrace$Maine == "Away",]
# Storing values Away of Maine variable in away
hist_home <- hist(home$Time..minutes., col='blue', border=F, xlab =
"Green: Away | Blue: Home", main = "Runners' times (in minutes)")
hist_away <- hist(away$Time..minutes., col='green', border=F, add=T)
# Creating two histograms of the runners' time for two different groups
using hist function
summary(home)
summary(away)
#summary function that gives the min, 1st quantile, median, mean, 3rd
quantile and max of two types in Maine attribute
summary(home$Time..minutes.)
summary(away$Time..minutes.)
# summary function that gives the min, 1st quantile, median, mean, 3rd
quantile and max of running time in minutes of the two groups.
sd(home$Time..minutes.)
sd(away$Time..minutes.)
# standard deviation of running time in minutes of the two
groups.Range_home <- max(home$Time..minutes.)-
min(home$Time..minutes.)
Range_home
Range_away <- max(away$Time..minutes.) - min(away$Time..minutes.)
Range_away
# Calculating range of running time in minutes that is the difference of
maximum and minimum values
IQR(home$Time..minutes.)
IQR(away$Time..minutes.)
```

```
# using IQR function to find inter-quantile range of running time in minutes of the two groups
```

The away runners have less running time than home runners. This means away runners are faster than home runners.

1. (c)

```
boxplot(home$Time..minutes., away$Time..minutes., names=c('Home', 'Away'), horizontal = TRUE, xlab='Minutes', main="Home & Away Runners' times")
# creating box plot using boxplot() function
```

1. (d)

```
male <- roadrace[roadrace$Sex == 'M',]
female <- roadrace[roadrace$Sex == 'F',]
boxplot(male$Age, female$Age, names = c('Male', 'Female'), horizontal = TRUE, xlab = 'Years', main = "Runners' ages")
# creating box plot for runners’ age for male and female runners
summary(female$Age)
summary(male$Age)
# summary function that gives the min, 1st quantile, median, mean, 3rd quantile and max of runners’ ages of the two groups


sd(female$Age)
sd(male$Age)
# standard deviation of runners’ ages of the two groups.
ageRange_female <- max(female$Age) - min(female$Age)
ageRange_female
ageRange_male <- max(male$Age) - min(male$Age)
ageRange_male
# Calculating range of runners’ ages that is the difference of maximum and minimum values.
IQR(female$Age)
IQR(male$Age)
# using IQR function to find inter-quantile range of runners’ ages of the two groups
```

We can conclude that the participating female runners are younger than the male runners on an average though the distribution of ages of female runners has outliers.

2.

```
motorcycle <- read.csv("/Users/chiragshahi/Desktop/motorcycle.csv")
# reading csv file
boxplot(motorcycle$Fatal.Motorcycle.Accidents, horizontal = TRUE, main = 'Accident data: South Carolina (in 2009)', xlab = '# of accidents')
#creating box plot
summary(motorcycle$Fatal.Motorcycle.Accidents)
# providing the summary
```

The distribution is left skewed and has two outliers. The median is between 10 and 15 and data is heavily spread between 7 and 23. The counties Greenville and Horry are outliers in the distribution.

There may be many factors contributing to the high motorcycle fatalities in the above mentioned counties such as higher population, road rules and so on.