CS 6313 Statistical Methods in Data Science

Name: Gaurav Joshi (Net ID: gsj150030), Murakonda Srujana (Net ID: SXM200270)

Date: March 4th, 2022

Mini Project 2

Contributions: Gaurav and Srujana both had equal contributions in completing the assignment.

Question 1:

(a).  
The bar graph is plotted using the barplot function in R. For better visualization, the away runners are shaded as red, and the Maine runners are shaded as blue.

# Read the data from csv file

road\_race\_data **<-** read.csv**(**'roadrace.csv'**)**

# Store the data associated with Maine in a table

road\_race\_table **<-** table**(**road\_race\_data**$**Maine**)**

# Plot the data as bar graph

barplot**(**road\_race\_table, main **=** "Road Race Data for Maine", ylab **=** "Race Participants", col **=** c**(**'red', 'blue'**))**Chart

Description automatically generated

# More Information about the runners

A picture containing graphical user interface

Description automatically generated

From the table, we can see that there were 4458 runners from Maine and 1417 runners away from Maine. In terms of percentages, about 75.88% of runners were from Maine while only 24.12% of runners were from away. We can conclude that the runners from Maine are heavily represented in the data set.

(b).  
  
More information about the runners can be calculated based on the runners’ times in minutes.

We use the subset function to gather a subset of the road\_race\_data based on whether the runners were from Maine or away. We further constraint this subset by filtering only the runners’ times in minutes and storing it into separate variable.

# Store details of racers from Maine

road\_racers\_from\_Maine **<-** subset**(**road\_race\_data, road\_race\_data**$**Maine **==** 'Maine'**)**

maine\_racers\_time **<-** road\_racers\_from\_maine**$**Time..minutes.

# Store details of racers not from Maine

road\_racers\_from\_Away **<-** subset**(**road\_race\_data, road\_race\_data**$**Maine **==** 'Away'**)**

away\_racers\_time **<-** road\_racers\_from\_away**$**Time..minutes.  
  
We calculate required information for both the Maine runners and the Away runners.

# Relevant information about Maine Runners

Text

Description automatically generated

# Relevant Information about Away Runners

Text

Description automatically generated

A generalized summary can be calculated using the summary function.

# General Summary

Text

Description automatically generated

# Plot Histogram of the two groups based on Time in Minutes

hist**(**maine\_racers\_time**)**

hist**(**away\_racers\_time**)**

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

(c)  
  
We use the boxplot function in R and pass the two variables representing runners from Maine and away to draw a side-by-side boxplot.

# Plot BoxPlot of the two groups based on Time in Minutes

boxplot**(**maine\_racers\_time, away\_racers\_time, names = c(“Maine Runners”, “Away Runners”)**)**

Chart, diagram, box and whisker chart

Description automatically generated  
The box plot helps us identify outliers in each of the groups. For example, we can see that one runner from Maine took significantly longer to finish the race compared to the rest of the group.

(d) The road\_runners variable for male and female dataset stores a subset of data filtered on gender of the runner. The runners variable further filters male and female runners based on their age. The data is converted from character to numeric type as a requirement for drawing a boxplot.

# Create a subset of all the male runners from the data set

male\_road\_runners **<-** subset**(**road\_race\_data, road\_race\_data**$**Sex **==** 'M'**)**

#Filter the male runners based on their age

male\_runners **<-** as.numeric**(**as.character**(**male\_road\_runners**$**Age**))**

# Create a subset of all the female runners from the data set

female\_road\_runners **<-** subset**(**road\_race\_data, road\_race\_data**$**Sex **==** 'F'**)**

#Filter the female runners based on their age

female\_runners **<-** as.numeric**(**as.character**(**female\_road\_runners**$**Age**))**

# Draw a side-to-side box plot of male and female runners

boxplot**(**male\_runners, female\_runners, names **=** c**(**"Male Runners", "Female Runners"**)**, ylab **=** "Age in Years"**)**

Chart, box and whisker chart

Description automatically generated

To get a better understanding of this data set, we use the mean, median, IQR, range and standard deviation of both the male and female runners.

# Relevant Information about Male Runners

Text

Description automatically generated

# Relevant Information about Female Runners

Text

Description automatically generated

A generalized summary can be calculated using the summary function.

# General Summary

A picture containing text

Description automatically generated

From the box plot, we can see that predominantly, most male runners are between the ages of 30 and 51. So most of the data is densely populated between second and third quartile for male runners, more so than the female runners.  
   
Additionally, female runners clearly have more outliers. Female runners have lower median, lower standard deviation, and lower inter-quartile range. However, the outliers (shown by dotted points) are clearly more in female runners.

Q 2.

Before drawing the boxplot, the data is extracted from csv file. This data is based to the boxplot filtered on the column representing the number of accidents.

# Read the data from csv file

motorcycle\_fatalities **<-** read.csv**(**'motorcycle.csv'**)**

number\_of\_accidents **<-** motorcycle\_fatalities**$**Fatal.Motorcycle.Accidents

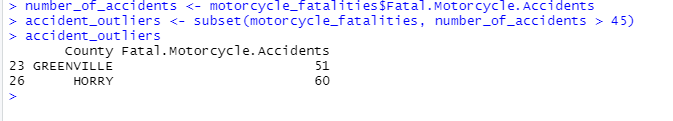
# Draw a boxplot of number of accidents in South Carolina Counties

boxplot**(**number\_of\_accidents, ylab = "Number of Motorcycle Accidents"**)**

Chart, box and whisker chart

Description automatically generated

From the boxplot, we can clearly see **two outliers** where the number of accidents is at least more than 45. Informally, we can compute these two outliers by using the subset function.



Formally, an outlier can be calculated by calculating all the values that lie at least 1.5 times lower than the inter-quartile range of first quartile and 1.5 times greater than the interquartile range of the third quartile.   
  
We build a relevant statistics table like the question one before formally computing the outliers.

# Relevant Information About Number of Accidents

Text

Description automatically generated

To get further information, we can use the summary function.

# Relevant Information About Number of Accidents

Text

Description automatically generated

Compute the outlier by first computing the lower bound and upper bound based on the inter-quartile range and building a subset of all the values that match this condition.

# Get the first and third quartiles of number of accidents dataset

first\_quartile **<-** summary**(**number\_of\_accidents**)[**2**]**

third\_quartile **<-** summary**(**number\_of\_accidents**)[**5**]**

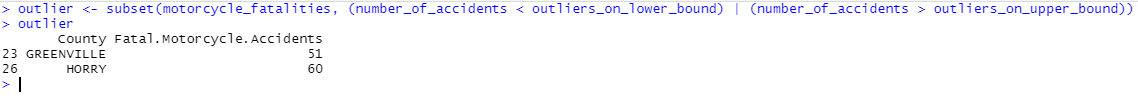
# Calculate the lower and upper bound within which a value is acceptable

outliers\_on\_lower\_bound **<-** first\_quartile **-** IQR**(**number\_of\_accidents**)** **\*** 1.5

outliers\_on\_upper\_bound **<-** third\_quartile **+** IQR**(**number\_of\_accidents**)** **\*** 1.5

# Filter out acceptable values and store the outliers

outlier\_data **<-** subset**(**motorcycle\_fatalities, number\_of\_accidents **<** outliers\_on\_lower\_bound **|** number\_of\_accidents **>** outliers\_on\_upper\_bound**)**

  
  
Thus, Greenville and Horry counties in South Carolina had exceeding number of motorcycle accidents compared to the rest of the counties in 2009. Upon further research, we discovered, that Horry County is a bit of a tourist spot, so it makes sense for a tourist spot to have more accidents. Additional reasons could include negligence, poor infrastructure, maintenance issues or having accident prone zones.