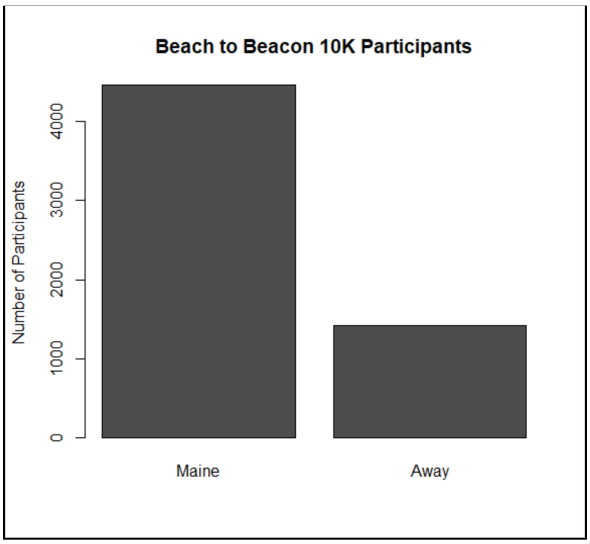
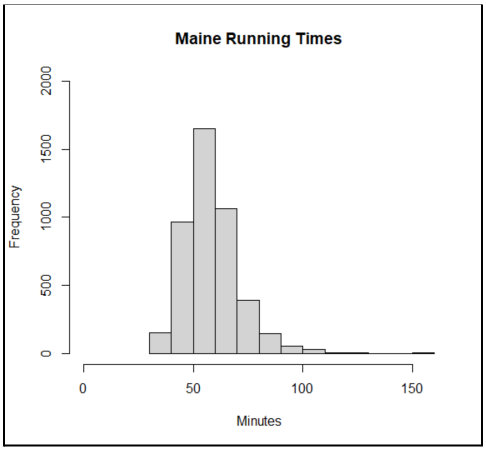
Mini Project #2
David McCormick (DTM190000)
I am in a solo group.

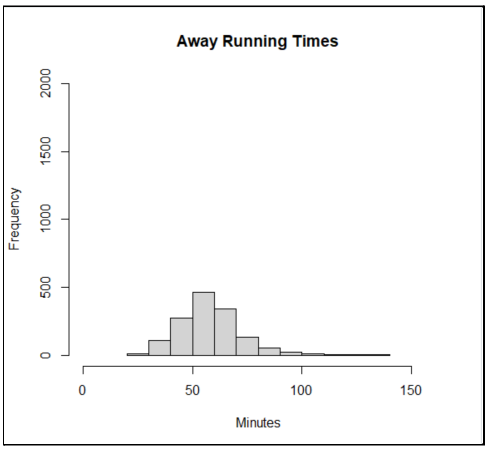
Section 1

1. Part 1
    a. The bar graph separates runners who finished the 2010 Beach to Beacon 10K on whether or not they are from Maine. The bar graph indicates that more people participated from Maine than did from outside of the state. There were, in fact, 3041 more people who participated from Maine than from other places as indicated by the range of the data (4458 - 1417 = 3041).

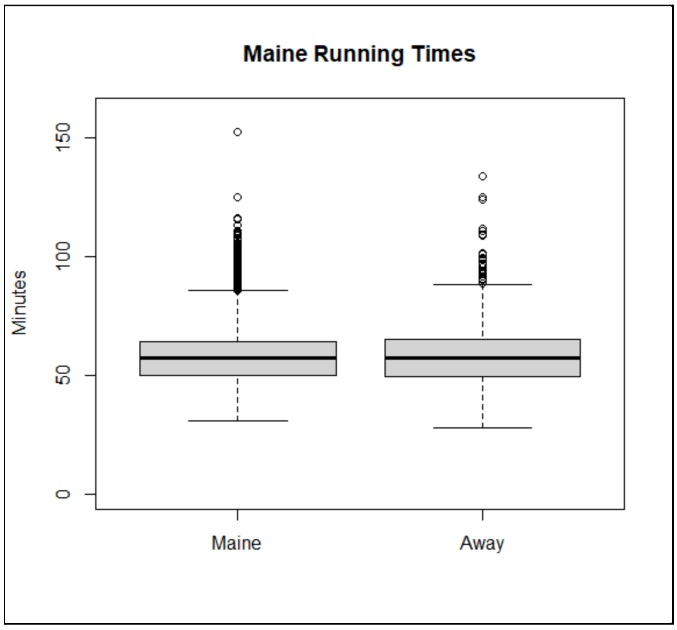


    b. The two distributions are both slightly skewed right as their means are both greater than their medians. The spread for both groups of runners appear to be similar, with close means of 58.20 minutes and 57.82 minutes and standard deviations of 12.19 minutes and 13.84 minutes for Maine and Away groups respectively. The Maine runner category has a greater range, as the maximum value is 152.167 minutes whereas the Away runner category's maximum value is 133.71 minutes.

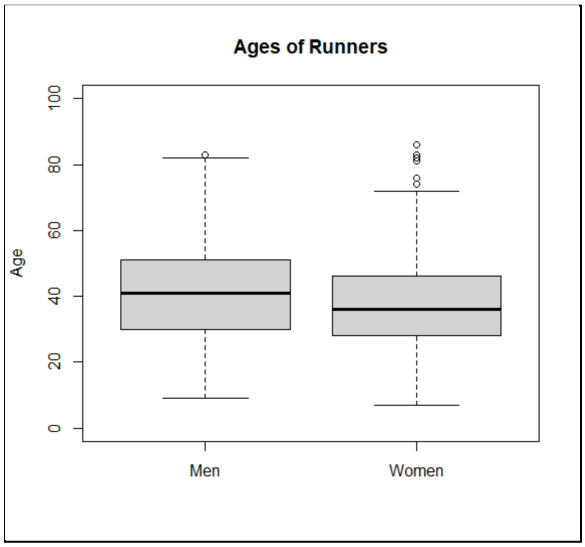




c. The side-by-side boxplots support the previous conclusions. It is easier to note that the distributions with boxplots as they have very similar spreads: similar medians and interquartile ranges. It can be said that the Maine distribution has more outliers than the Away group, and the maximum value for the Maine distribution, 152.167 minutes, is much greater than the maximum value for the Away distribution of 133.71 minutes.

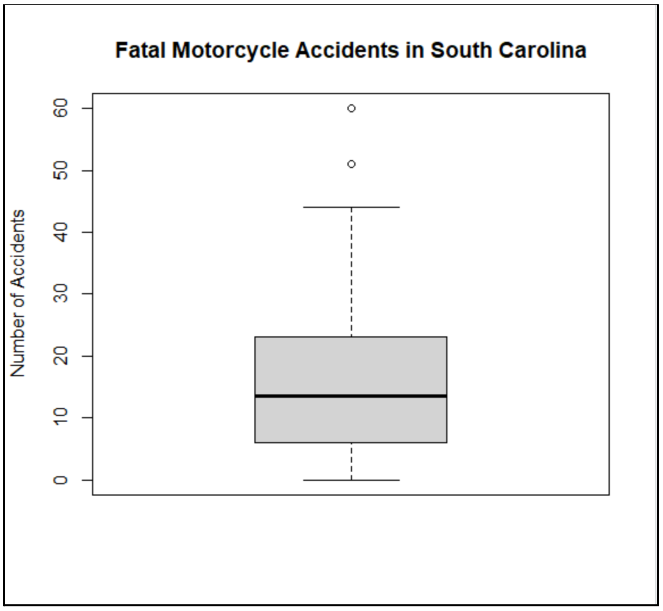


d. According to the boxplot of men's ages versus women's ages, more younger women participated than younger men. The median age for female runners is 36 years whereas the mean age for male runners lies at 41 years. The interquartile range for male runners is also greater than the interquartile range for female runners at 21 years and 18 years respectively, representing a larger spread.



2.

a. The data for motorcycle accidents in South Carolina appears to be skewed right with a mean of 17.02 accidents and a median of 13.5 accidents. One of the counties experienced 0 accidents whereas another experienced 60 accidents, which represent the range of motorcycle accidents.



Section 2

```
stats <- function(column) {
  cat("Mean:\t", mean(column), "\n")
  cat("SD:\t", sd(column), "\n")
  cat("Range:\t", range(column), "\n")
  cat("Median:\t", median(column), "\n")
  cat("IQR:\t", IQR(column), "\n")
} # Displays relevant descriptive statistics

csv <- read.csv("C:/Users/David/Desktop/roadrace.csv") # Reads the roadrace CSV
fromMaine <- subset(csv, csv[["Maine"]] == "Maine") # Gets the Mainers from the Maine column
fromAway <- subset(csv, csv[["Maine"]] == "Away") # Gets the runners from Away from the Maine column
tab <- matrix(c(nrow(fromMaine), nrow(fromAway)), ncol=2) # Turns the data into a matrix
colnames(tab) <- c('Maine', 'Away') # Adds column names into the matrix
barplot(tab, main="Beach to Beacon 10K Participants", ylab="Number of Participants") # Graphs as a bar plot

stats(tab) # Displays the statistics for Maine/Away participants (not very useful except for range)

## Mean:    2937.5
## SD:   2150.312
## Range:    1417 4458
## Median:   2937.5
## IQR:  1520.5

hist(fromMaine$Time..minutes., xlim=(c(0,160)), ylim=(c(0,2000)), main="Maine Running Times", xlab = "Minutes", ylab = "Frequency") # Histogram of time in minutes for Maine runners

hist(fromAway$Time..minutes., xlim=(c(0,160)), ylim=(c(0,2000)), main="Away Running Times", xlab = "Minutes", ylab = "Frequency") # Histogram of time in minutes for Away runners

stats(fromMaine$Time..minutes.) # Displays the statistics for the time in minutes of Maine runners

## Mean:    58.19514
## SD:   12.18511
## Range:    30.567 152.167
## Median:   57.0335
## IQR:  14.24775

stats(fromAway$Time..minutes.) # Displays the statistics for the time in minutes of Away runners

## Mean:    57.82181
## SD:   13.83538
## Range:    27.782 133.71
## Median:   56.92
## IQR:  15.674
```

```
boxplot(fromMaine$Time..minutes., fromAway$Time..minutes., ylim=(c(0,160)), main="Maine Running Times", names = (c("Maine", "Away")), ylab = "Minutes") #Displays the data as a side-by-side boxplot

men <- subset(csv, csv[["Sex"]] == "M") # Separates men in Sex column
women <- subset(csv, csv[["Sex"]] == "F") # Separates women in Sex column
boxplot(as.integer(men$Age), as.integer(women$Age), ylim=(c(0,100)), main="Ages of Runners", names = (c("Men", "Women")), ylab="Age") # Side-by-side boxplot for ages of women and men participants

stats(as.integer(men$Age)) # Displays the statistics for age of men

## Mean:     40.4468
## SD:   13.99289
## Range:    9 83
## Median:   41
## IQR:  21

stats(as.integer(women$Age)) # Displays the statistics for age of women

## Mean:     37.23653
## SD:   12.26925
## Range:    7 86
## Median:   36
## IQR:  18

csv1 <- read.csv("C:/Users/David/Desktop/motorcycle.csv") # Reads the motorcycle csv
boxplot(csv1$Fatal.Motorcycle.Accidents, main = "Fatal Motorcycle Accidents in South Carolina", ylab = "Number of Accidents") # Boxplot for the number of accident in each county

outliers = which(csv1$Fatal.Motorcycle.Accidents > quantile(csv$Fatal.Motorcycle.Accidents, .75) + 1.5 * (IQR(csv$Fatal.Motorcycle.Accidents))) # Finds outliers (1.5IQR + Q3)
counties = csv1$County[outliers] # Gets the county names
stats(csv1$Fatal.Motorcycle.Accidents) # Displays the statistics for motorcycle accidents in SC

## Mean:     17.02083
## SD:   13.81256
## Range:    0 60
## Median:   13.5
## IQR:  17
```