Mini-Project 2: CS6313.001

Rutvij Shah, rds1900000.

Question 1

(a) Maine vs Away

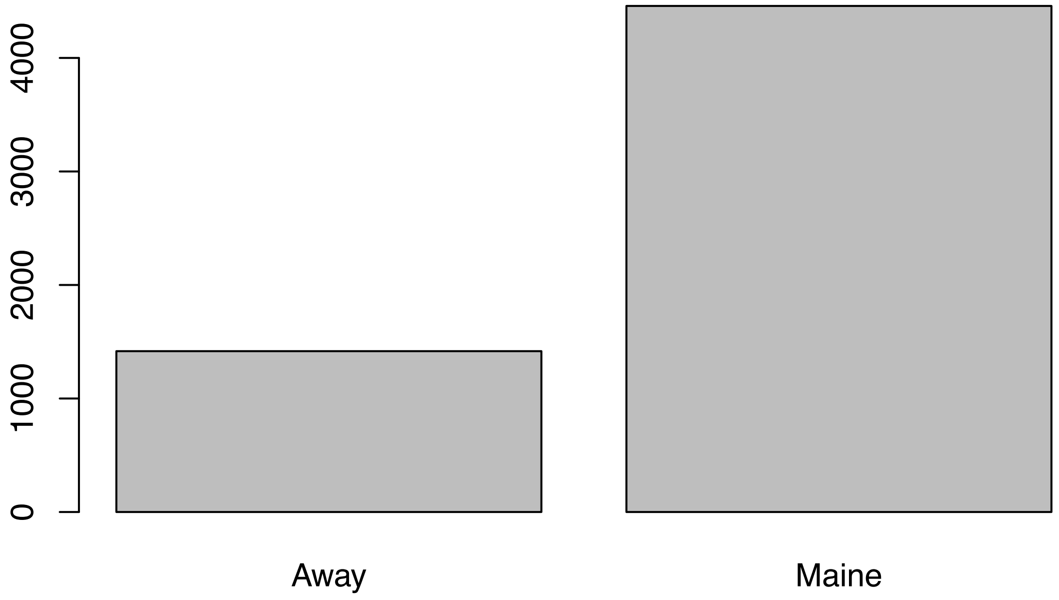
Loading data & and a split of runners from Maine and those not from Maine (away)

```
# A function to calculate mode
mode <- function(codes){
  which.max(tabulate(codes))
}
roadrace = read.csv("/Users/rutvijshah/Downloads/roadrace.csv")
# A library to make table of summary stats
suppressMessages(library(vtable))
# Extract the Maine vs Away column from the dataset
runners_split = table(roadrace["Maine"])
runners_split
```

```
##
##  Away Maine
##  1417  4458
```

Plot of the number of runners from Maine and Away.

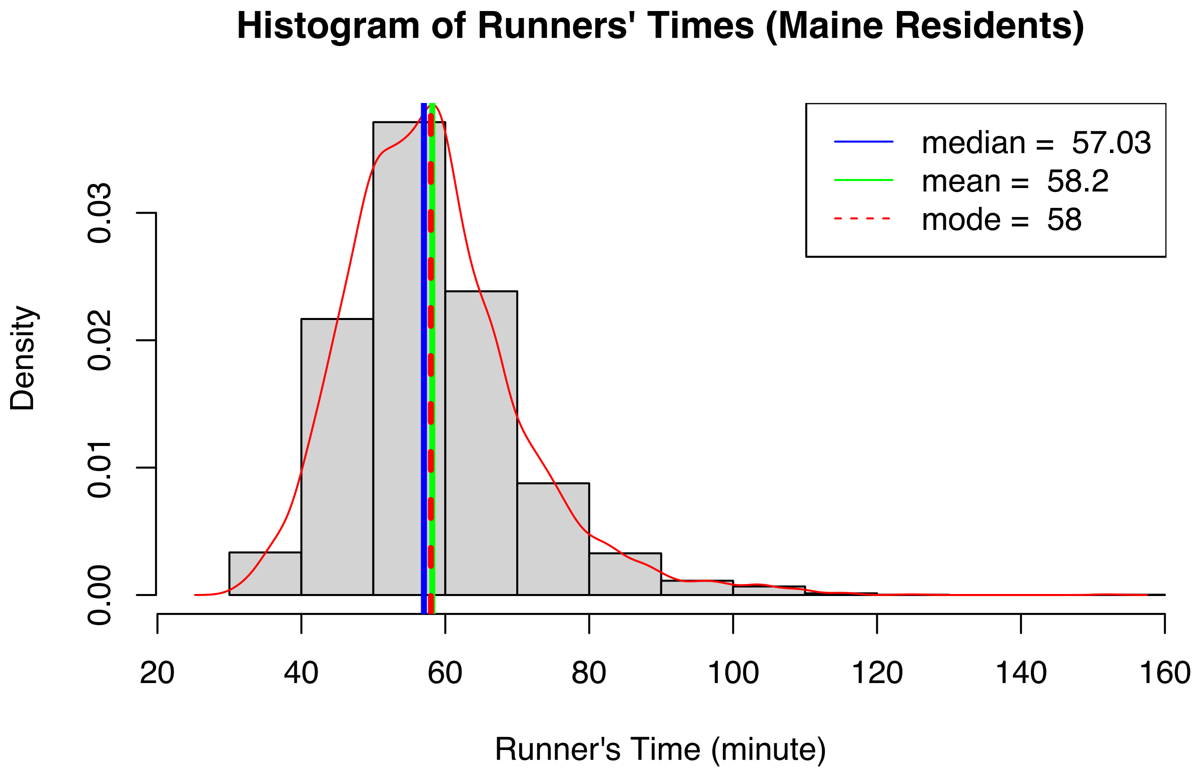
```
barplot(runners_split)
```



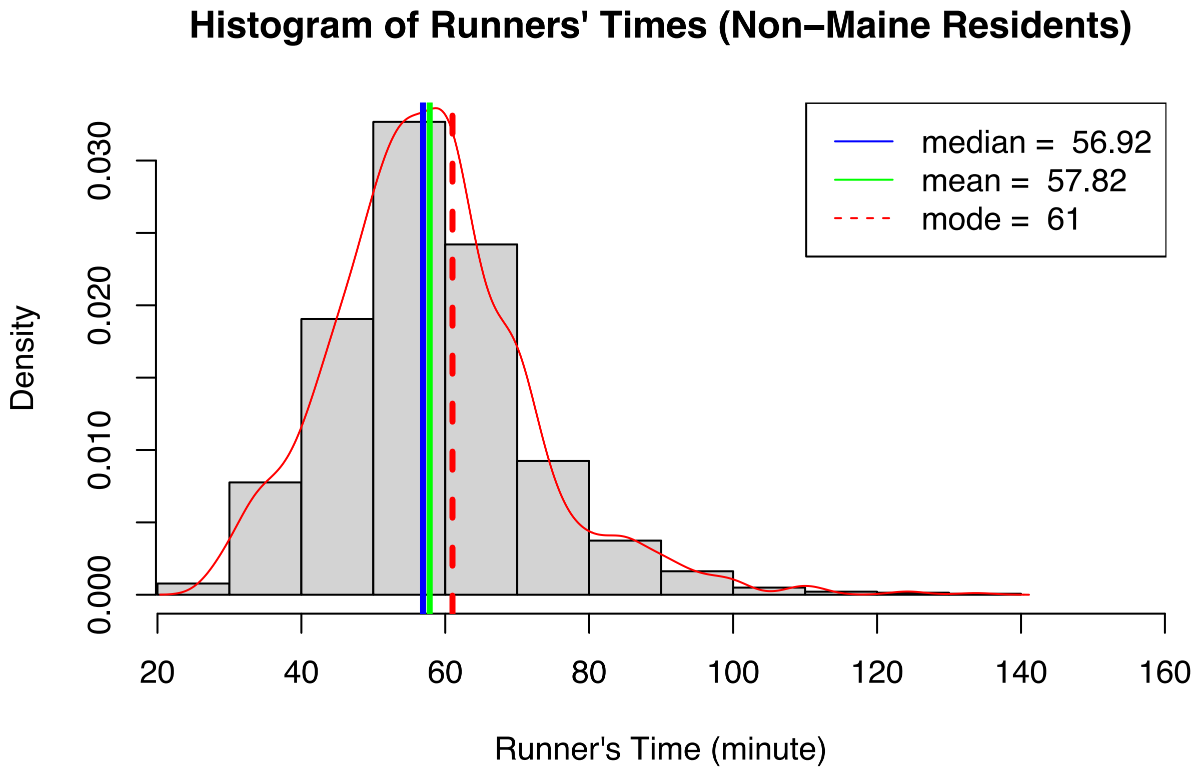
Conclusions This plot highlights the fact that the majority of the runners in this road race are residents of Maine.

(b) Runners' Time Distributions by State of Domicile I

```r
# min time overall
min_time = min(roadrace[,"Time..minutes."])
# max time overall
max_time = max(roadrace[,"Time..minutes."])
# find the closest multiple of 5 less than min
time_range_min = min_time-min_time%%5
# find the closest multiple of 5 greater than max
time_range_max = max_time+(5-max_time%%5)
# the scaled for the histograms
scale_times = c(time_range_min, time_range_max)

# filter times of runners' from Maine
times.runners_from_maine = roadrace[roadrace$"Maine" == "Maine", ][,"Time..minutes."]

hist(times.runners_from_maine,
     main="Histogram of Runners' Times (Maine Residents)",
     xlab = "Runner's Time (minute)",
     xlim = scale_times,
     freq = FALSE
     )

# overlay density plot to better visualized dist.
lines(density(times.runners_from_maine), col="red")
# find the median
maine_median = round(median(times.runners_from_maine), digits=2)
# plot the median
abline(v = maine_median, col="blue", lwd=3)
# find the mean
maine_mean = round(mean(times.runners_from_maine), digits=2)
# plot the mean
abline(v = maine_mean, col="green", lwd=3)
# find the mode
maine_mode = mode(times.runners_from_maine)
# plot the mode
abline(v = maine_mode, col="red", lwd=3, lty=2)
# plot legend
legend(x='topright',
       legend = c(paste("median = ", maine_median),
                  paste("mean = ", maine_mean),
                  paste("mode = ", maine_mode)
                  ),
       lty=c(1,1,2),
       col=c("blue","green","red"))
```



```
# filter times of runners' NOT from Maine
times.runners_from_away = roadrace[roadrace$"Maine" == "Away", ][,"Time..minutes."]
hist(times.runners_from_away,
     main="Histogram of Runners' Times (Non-Maine Residents)",
     xlab = "Runner's Time (minute)",
     xlim = scale_times,
     freq = FALSE)

# overlay density plot to better visualized dist.
lines(density(times.runners_from_away), col="red")
# find the median
away_median = round(median(times.runners_from_away), digits=2)
# plot the median
abline(v = away_median, col="blue", lwd=3)
# find the mean
away_mean = round(mean(times.runners_from_away), digits=2)
# plot the mean
abline(v = away_mean, col="green", lwd=3)
# find the mode
away_mode = mode(times.runners_from_away)
# plot the mode
abline(v = away_mode, col="red", lwd=3, lty=2)
# plot the legend
legend(x='topright',
       legend = c(paste("median = ", away_median),
                  paste("mean = ", away_mean),
                  paste("mode = ", away_mode)),
       lty=c(1,1,2),
       col=c("blue","green", "red"))
```



Comparative Summary Statistics

```
# plot summary table for runners; times, excluding any NAs, group on Maine
st(na.exclude(roadrace[, c("Maine", "Time..minutes.")]), group = "Maine", out = "return", digits=2, grou
```

```
##          Variable    N  Mean Std. Dev.   Min Pctl. 25 Pctl. 75    Max
## 1    Maine: Away
## 2 Time..minutes. 1417 57.82     13.84 27.78    49.15    64.83 133.71
## 3
## 4   Maine: Maine
## 5 Time..minutes. 4458  58.2     12.19 30.57       50    64.24 152.17
```

```
# display the medians for both the groups
paste("Maine's Runners' Median Time", median(times.runners_from_maine))
```

```
## [1] "Maine's Runners' Median Time 57.0335"
```

```
paste("Away Runners' Median Time", median(times.runners_from_away))
```

```
## [1] "Away Runners' Median Time 56.92"
```

Conclusions Both distributions are close to being a normal distribution with marginal variation between their mean, mode and median. But both also have their median to be slightly lower than their mean thus showing a slight positive skew.

The summary stats for distribution of times for runners from Maine are:

Statistic	Dist. Runners not from Maine	Normal Dist. mean = 57.8, sd = 13.84
25th percentile	49.15	48.5
median	56.92	57.8
75th percentile	64.83	67.5

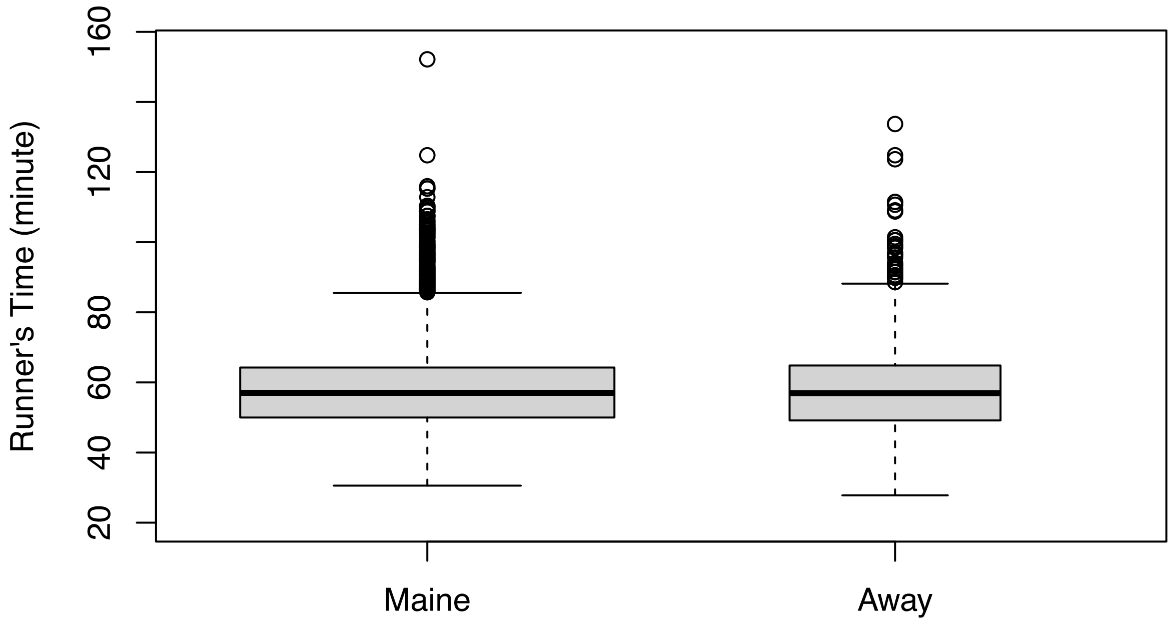
The table above suggests that the away runners' time distribution does almost resemble a normal dist.

Statistic	Dist. Runners from Maine	Normal Dist. mean = 58.2, sd = 12.19
25th percentile	50	50
median	57.03	58.2
75th percentile	64.24	66.5

The table above suggests that the Maine runners' time distribution does almost resemble a normal dist, with a slight right skew.

(c) Runners' Time Distributions by State of Domicile II

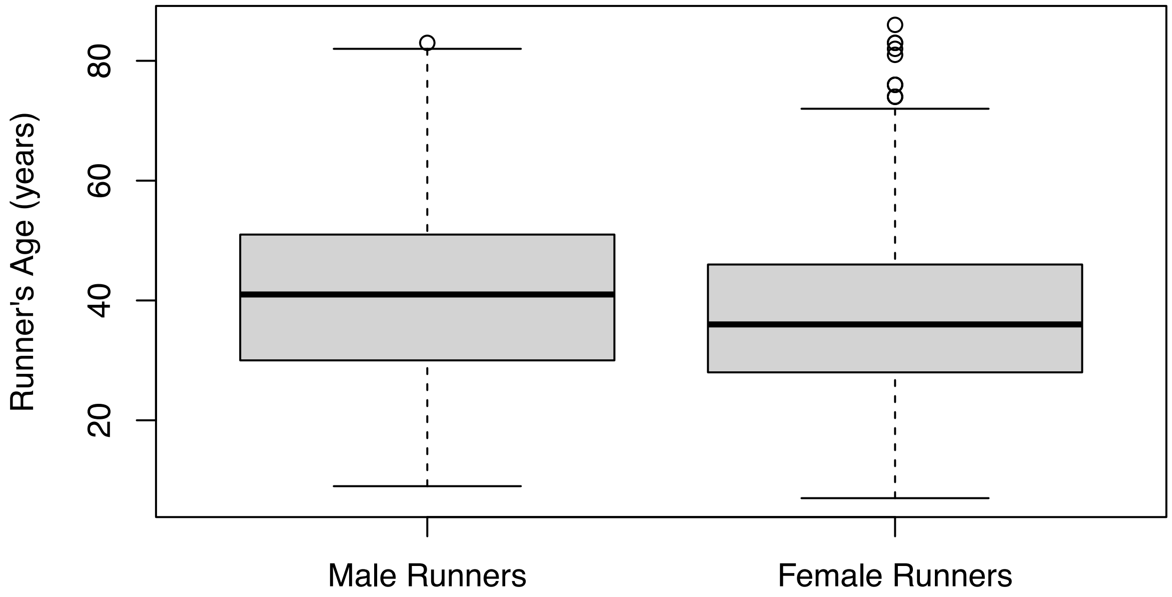
```r
boxplot(times.runners_from_maine,
        times.runners_from_away,
        ylim=c(20, 155),
        varwidth = TRUE,
        names = c("Maine", "Away"),
        ylab = "Runner's Time (minute)")
```



The boxplots draw greater emphasis towards the similarity of the distributions, and that even though Maine has more observations, both have a similar central tendency and 1.5*SD ranges.

(d) Runners' Age Distributions by Gender

```
# age data of males
ages.male = as.integer(roadrace[roadrace$"Sex" == "M", ][,"Age"])
# age data of females
ages.female = as.integer(roadrace[roadrace$"Sex" == "F", ][,"Age"])

boxplot(ages.male,
        ages.female,
        names = c("Male Runners", "Female Runners"),
        ylab = "Runner's Age (years)")
```



```
# suppress any warnings of NA when converting string age to int age
roadrace$Age = suppressWarnings(as.integer(roadrace$Age))
# plot summary table for runners' ages, excluding any NAs, group on Sex
st(na.exclude(roadrace[, c("Age", "Sex")]), group = "Sex", out = "return", digits=2, fixed.digits = TRU
```

```
##   Variable    N  Mean Std. Dev. Min Pctl. 25 Pctl. 75 Max
## 1   Sex: F
## 2      Age 2951 37.24     12.27   7       28       46  86
## 3
## 4   Sex: M
## 5      Age 2923 40.45     13.99   9       30       51  83
```

```
# display median info
paste("Female Runners' Median Age", median(ages.female))
```

```
## [1] "Female Runners' Median Age 36"
```

```
paste("Male Runners' Median Age", median(ages.male))
```

```
## [1] "Male Runners' Median Age 41"
```

Conclusions On average, Female runners were younger than Male runners and there were slightly more Female runners than Male.

The standard deviation of the ages of Female runners is also lower than that for Male runners while both the youngest and oldest runners overall were also Female.

The spread of Male runners was more uniform as evidenced by the single outlier aged 83. While there were greater number of outliers in the distribution of female runners.

Statistic	Age Female Runners	Normal Dist. mean = 37.24, sd = 12.27
25th percentile	28	29
median	36	37.24
75th percentile	46	45.5

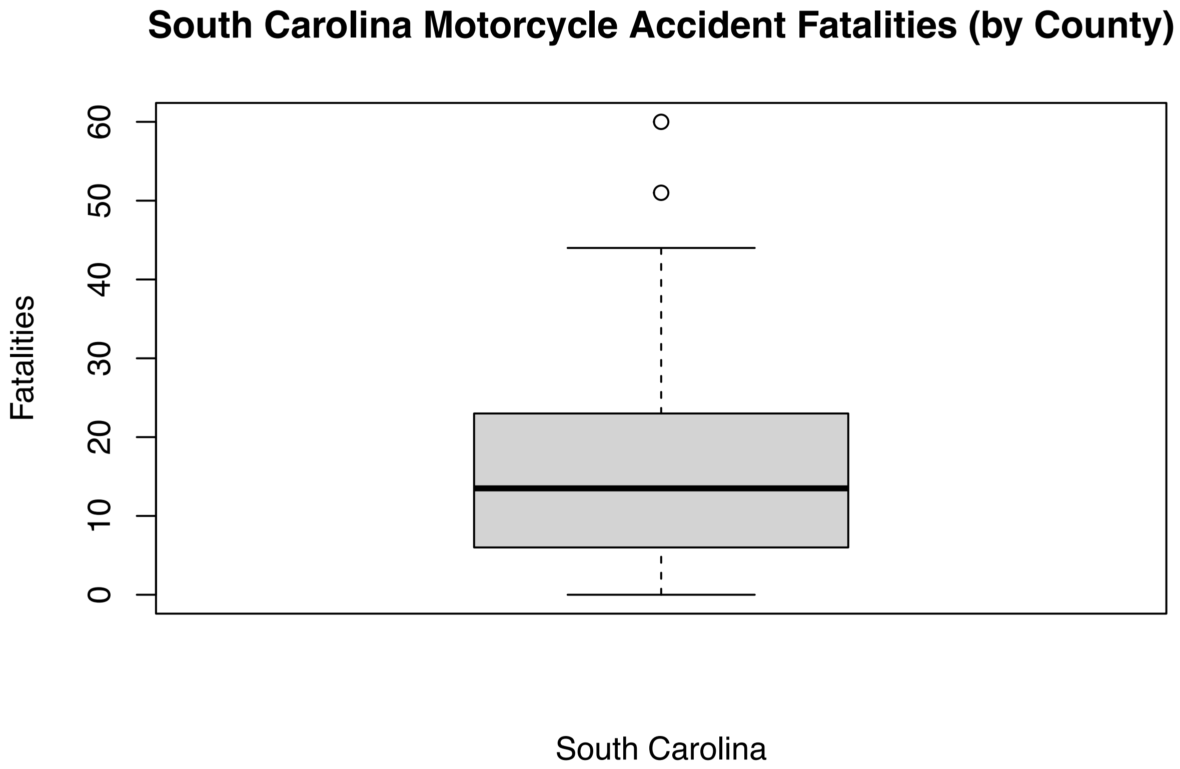
As evident from the table, the age dist. of female runners closely approximates a normal distribution.

Statistic	Age Female Runners	Normal Dist. mean = 40.45, sd = 13.99
25th percentile	30	31
median	41	40.45
75th percentile	51	50

As evident from the table, the age dist. of male runners closely approximates a normal distribution.

Question 2

```r
# load the motorcycle dataset
motorcycle = read.csv("/Users/rutvijshah/Downloads/motorcycle.csv")
boxplot(motorcycle$Fatal.Motorcycle.Accidents,
        xlab = "South Carolina",
        ylab = "Fatalities",
        main = "South Carolina Motorcycle Accident Fatalities (by County)")
```



```
# create summary table for motorcycle fatalities in South Carolina
st(motorcycle, out = "return")
```

```
##                     Variable  N   Mean Std. Dev. Min Pctl. 25 Pctl. 75 Max
## 1 Fatal.Motorcycle.Accidents 48 17.021    13.813   0        6       23  60
```

```
# find the upper quartile of the dist
upper_quartile = quantile(motorcycle$Fatal.Motorcycle.Accidents)[4]
# inter quartile range of the dist
iqr = IQR(motorcycle$Fatal.Motorcycle.Accidents)
# find the outliers
motorcycle[motorcycle$Fatal.Motorcycle.Accidents > upper_quartile+(1.5*iqr), ]
```

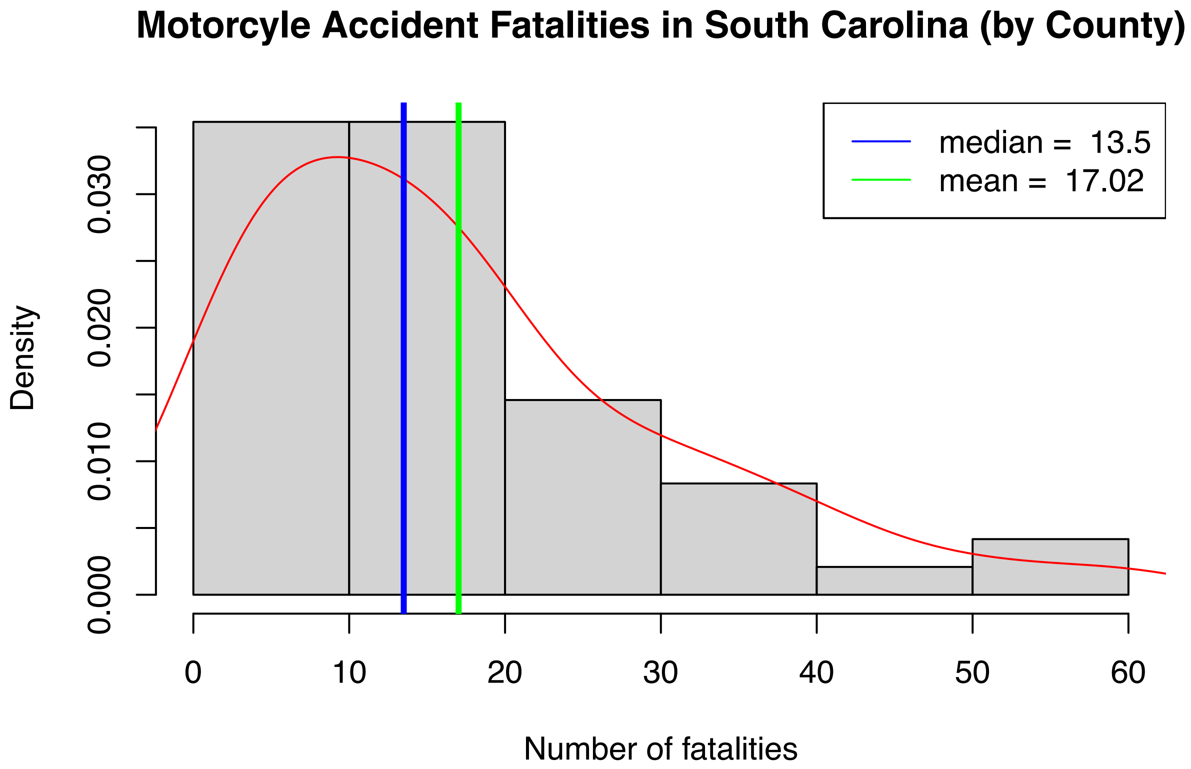
```
##        County Fatal.Motorcycle.Accidents
## 23 GREENVILLE                         51
## 26      HORRY                         60
```

```
# load population data for South Carolina's counties
csvData = read.csv("/Users/rutvijshah/Downloads/csvData.csv")
# string manipulation to remove trailing "County" from the county name string
counties = strsplit(csvData$CTYNAME, split = " ")
# unlist the split county strings and make a matrix
counties = matrix(unlist(counties), ncol=2, byrow = TRUE)
# extract the column of county names and make them uppercase
csvData$CTYNAME = toupper(counties[,1])
# merge the population and accidents data frames on county names
merged = merge(motorcycle, csvData, by.x="County", by.y="CTYNAME")
# find the correlation between the number of accidents and population of a county
```

```
paste("Correlation of Motorcycle Accidents to Population Density of a County:",
      round(cor(merged$Fatal.Motorcycle.Accidents, merged$popDensity, method="spearman"),
            4))
```

```
## [1] "Correlation of Motorcycle Accidents to Population Density of a County: 0.8903"
```

```
# plot a histogram of fatality counts
hist(motorcycle$Fatal.Motorcycle.Accidents,
     freq = FALSE,
     main = "Motorcyle Accident Fatalities in South Carolina (by County)",
     xlab = "Number of fatalities")

# plot a density overlay to better visualize the dist
lines(density(motorcycle$Fatal.Motorcycle.Accidents), col = "red")

# find median
median_acc = round(median(motorcycle$Fatal.Motorcycle.Accidents), digits=2)
# plot median
abline(v = median_acc, col="blue", lwd=3)
# find mean
mean_acc = round(mean(motorcycle$Fatal.Motorcycle.Accidents), digits=2)
# plot mean
abline(v = mean_acc, col="green", lwd=3)
# plot legend
legend(x='topright',
       legend = c(paste("median = ", median_acc), paste("mean = ", mean_acc)),
       lty=c(1,1),
       col=c("blue","green"))
```



Conclusions The distribution has a significant right skew, as evidenced by the mean being greater than the median.

The counties which can be considered outliers based on the dataset are Greenville and Horry counties.

My initial hypothesis for a significantly higher number of fatalities in these counties was that their populations might be higher.

Thus, I downloaded a dataset of populations of all South Carolina counties and though the population is from 2021, all counties within are region of interest (the top 5 most populous) so similar level of growths.

Based on the population size, Greenville and Horry are among the top 4 most populous counties and also in terms of population density, turns out all counties have been assumed as having similar land areas.

Due to the visual correlation of the counties' populations and fatalities, I ran a spearman correlation test on the population densities and number of fatalities, the degree of correlation was 0.89 showing a high degree of positive correlation and thus supporting the hypothesis that the greater number of accidents were as a result of a larger population size and density. Coupled with the fact that Greenville is at the heart of the largest metropolitan center in South Carolina while Horry seems to be more of the outlier despite its population.