Mini project #2

Group_22 Member: Chaoran Li, Wenting Wang

Contribution of each member:

Firstly, we discussed the mathematical models. Then, we divided the project into two part and finished our respective work. Wenting Wang mainly worked on Q1 while Chaoran Li worked on Q2 and trial in to use ggplot2 to draw interleaved histograms in Q1-b. Then, we merged our code and worked on optimizing the diagrams. Finally, we gathered all we got into this report.

Each member makes contribution to each sub task of this project and combines all to finish this project, as the details shown in table 1.

	Question1-a	Question1-b	Question1-c	Question1-d	Question2
Chaoran li	30%	50%	30%	30%	70%
Wenting wang	70%	50%	70%	70%	30%

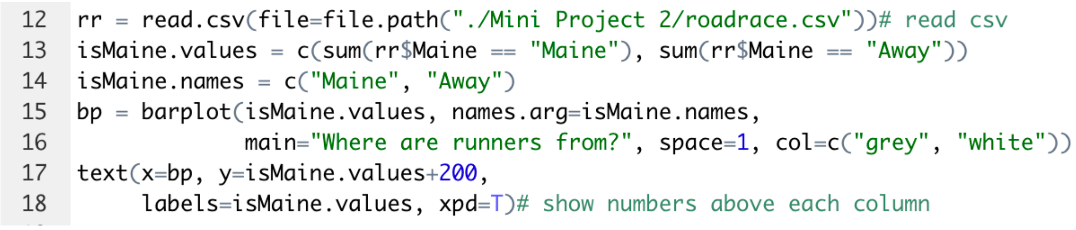
Table 1: Member contribution table

Question 1:

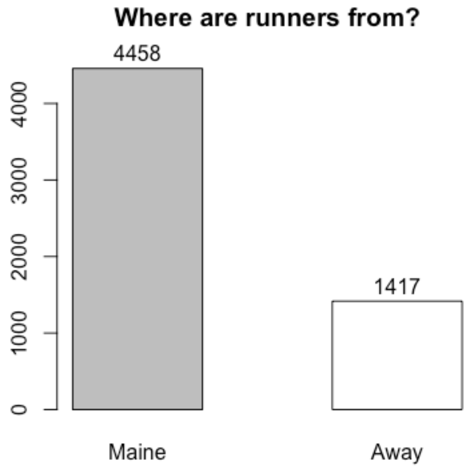
Consider the dataset roadrace.csv posted on eLearning. It contains observations on 5875 runners who finished the 2010 Beach to Beacon 10K Road Race in Cape Elizabeth, Maine. You can read the dataset in R using read.csv function.

(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 barplot function for this. What can we conclude from the plot? Back up your conclusions with relevant summary statistics.

Solution:



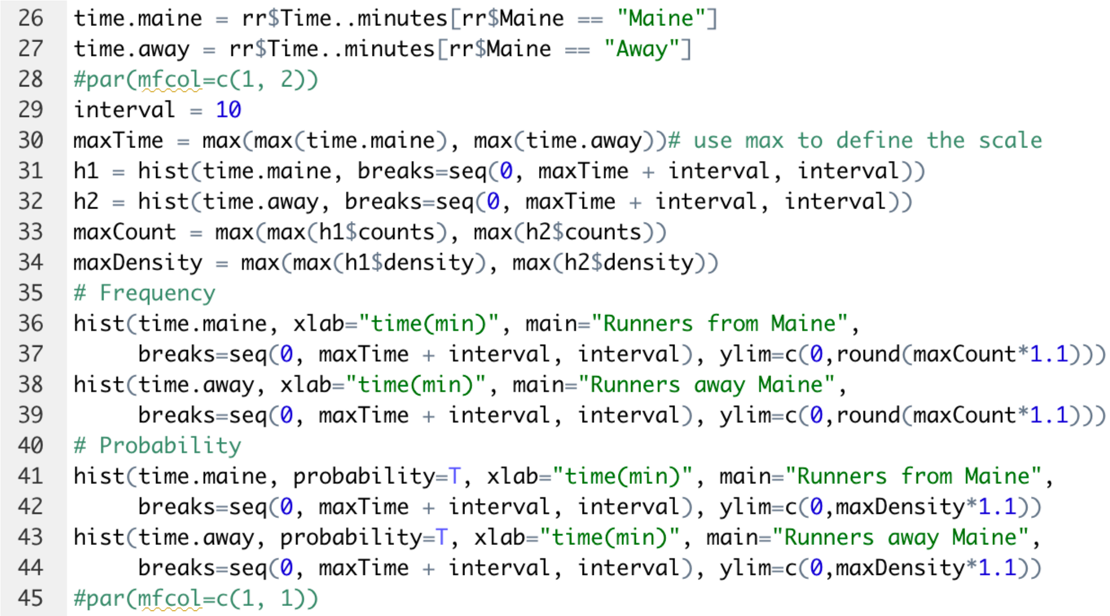
```r
rr = read.csv(file=file.path("./Mini Project 2/roadrace.csv"))# read csv
isMaine.values = c(sum(rr$Maine == "Maine"), sum(rr$Maine == "Away"))
isMaine.names = c("Maine", "Away")
bp = barplot(isMaine.values, names.arg=isMaine.names,
             main="Where are runners from?", space=1, col=c("grey", "white"))
text(x=bp, y=isMaine.values+200,
     labels=isMaine.values, xpd=T)# show numbers above each column
```



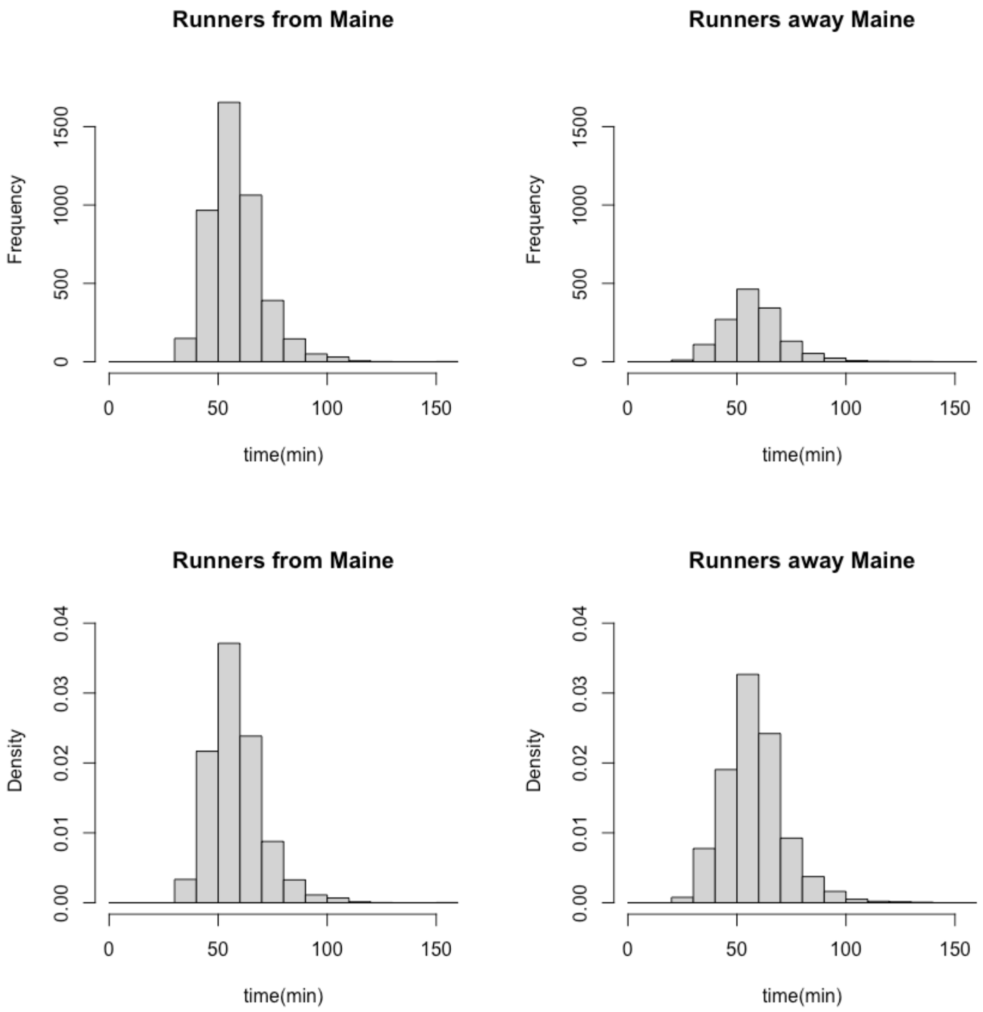
Conclusion: From the bar graph above, we can conclude that the runners from Maine is more than that from other places. The number of runners from Maine is 4458 about 75.8% of the total number of runners, but the number of from Away places is only 24.2% of the total number of runners.

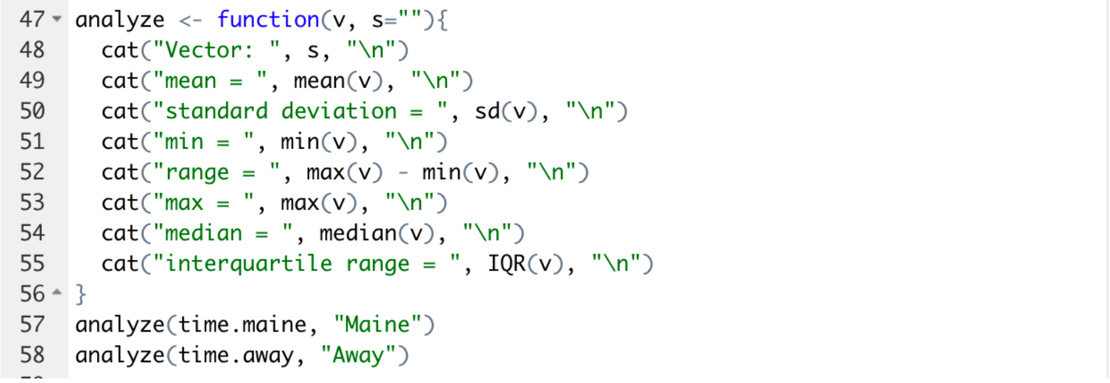
(b) **Create two histograms the runners' times (given in minutes) — one for the Maine group and the second for the Away group. Make sure that the histograms on the same scale. What can we conclude about the two distributions? Back up your conclusions with relevant summary statistics, including mean, standard deviation, range, median, and interquartile range.**

Solution:

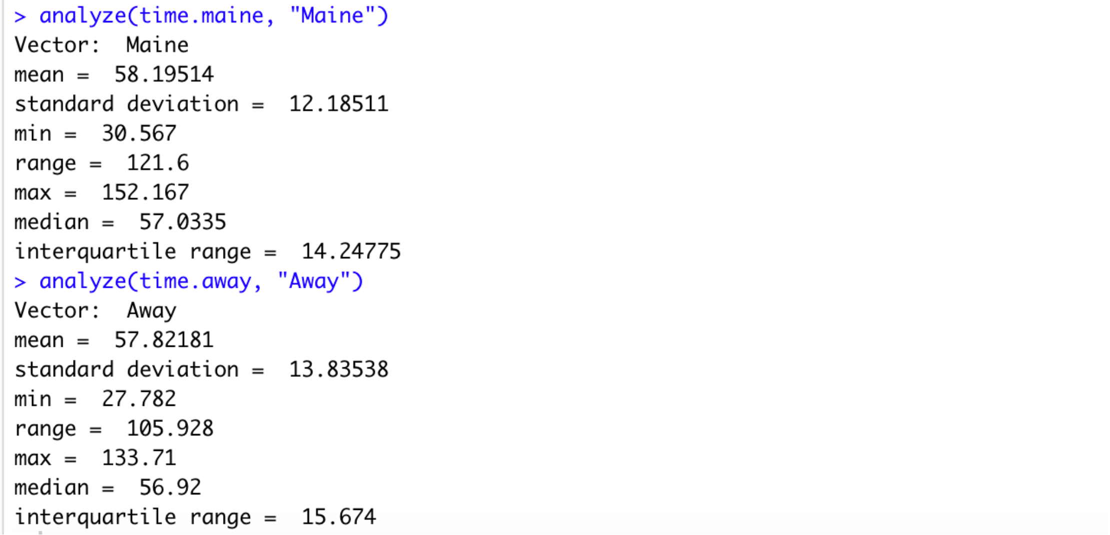


```
time.maine = rr$Time..minutes[rr$Maine == "Maine"]
time.away = rr$Time..minutes[rr$Maine == "Away"]
#par(mfcol=c(1, 2))
interval = 10
maxTime = max(max(time.maine), max(time.away))# use max to define the scale
h1 = hist(time.maine, breaks=seq(0, maxTime + interval, interval))
h2 = hist(time.away, breaks=seq(0, maxTime + interval, interval))
maxCount = max(max(h1$counts), max(h2$counts))
maxDensity = max(max(h1$density), max(h2$density))
# Frequency
hist(time.maine, xlab="time(min)", main="Runners from Maine",
     breaks=seq(0, maxTime + interval, interval), ylim=c(0,round(maxCount*1.1)))
hist(time.away, xlab="time(min)", main="Runners away Maine",
     breaks=seq(0, maxTime + interval, interval), ylim=c(0,round(maxCount*1.1)))
# Probability
hist(time.maine, probability=T, xlab="time(min)", main="Runners from Maine",
     breaks=seq(0, maxTime + interval, interval), ylim=c(0,maxDensity*1.1))
hist(time.away, probability=T, xlab="time(min)", main="Runners away Maine",
     breaks=seq(0, maxTime + interval, interval), ylim=c(0,maxDensity*1.1))
#par(mfcol=c(1, 1))
```





```r
analyze <- function(v, s=""){
  cat("Vector: ", s, "\n")
  cat("mean = ", mean(v), "\n")
  cat("standard deviation = ", sd(v), "\n")
  cat("min = ", min(v), "\n")
  cat("range = ", max(v) - min(v), "\n")
  cat("max = ", max(v), "\n")
  cat("median = ", median(v), "\n")
  cat("interquartile range = ", IQR(v), "\n")
}
analyze(time.maine, "Maine")
analyze(time.away, "Away")
```



```
> analyze(time.maine, "Maine")
Vector:  Maine
mean =  58.19514
standard deviation =  12.18511
min =  30.567
range =  121.6
max =  152.167
median =  57.0335
interquartile range =  14.24775
> analyze(time.away, "Away")
Vector:  Away
mean =  57.82181
standard deviation =  13.83538
min =  27.782
range =  105.928
max =  133.71
median =  56.92
interquartile range =  15.674
```

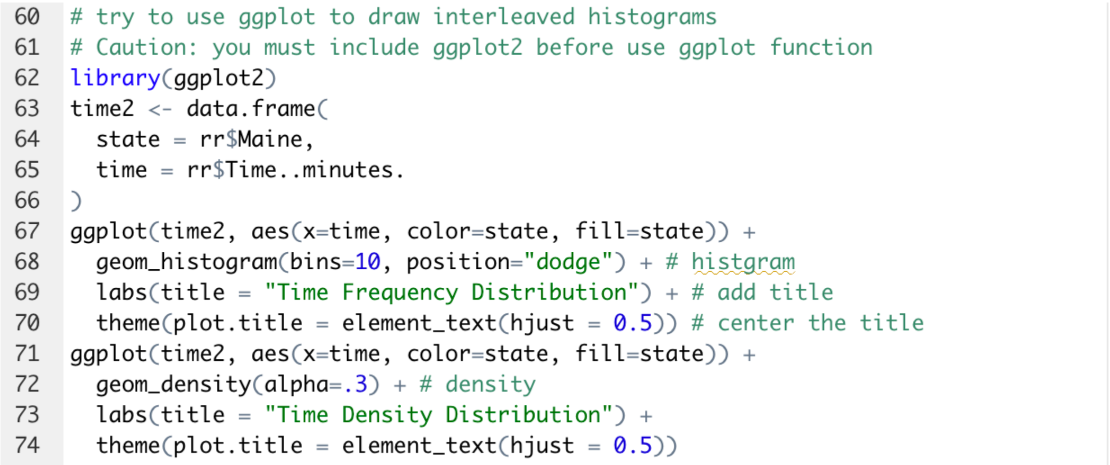
State	Mean	SD	Min	Range	Max	Median	IQR
Maine	58.19514	12.18511	30.657	121.6	152.167	57.0335	14.24775
Away	57.82181	13.83538	27.782	105.928	133.71	56.92	15.674

Table 2: statistics for runners from Maine and Away

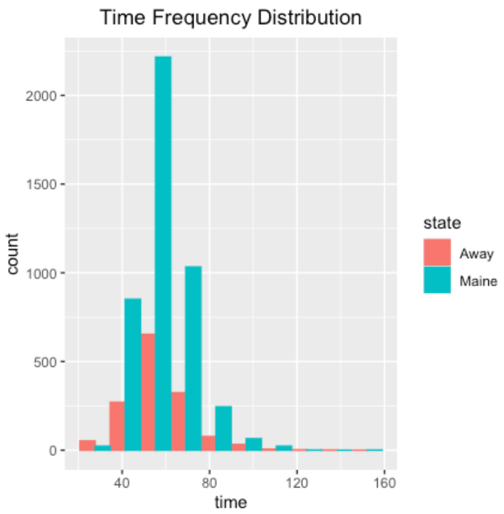
Conclusion: From the graph and table 2 we can conclude that the average time for Maine runners is more than Away runners. Moreover, the winner should be in the Away Group, because the minimum time for Away Group is smaller than Group Maine. The slowest runner is in the Maine Group. But the time for Maine Group is a little steady than Away Group, as the standard deviation and interquartile range of Maine Group are smaller.

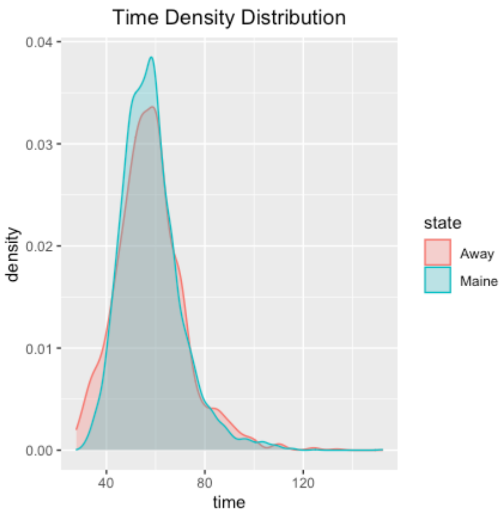
Let’s combinate the data with reality. Usually, runners from other state will have higher competitive level. Because a beginner would seldom drive far to join this competition. Therefore, the mean, median and minimum of the Away Group are relatively small. But the local runner would suffer less from accidents. Then, the Maine Group's results are more stable, which reflects in the smaller standard deviation and interquartile range.

We also use package ggplot to draw two more intuitive diagrams.



```r
# try to use ggplot to draw interleaved histograms
# Caution: you must include ggplot2 before use ggplot function
library(ggplot2)
time2 <- data.frame(
  state = rr$Maine,
  time = rr$Time..minutes.
)
ggplot(time2, aes(x=time, color=state, fill=state)) +
  geom_histogram(bins=10, position="dodge") + # histgram
  labs(title = "Time Frequency Distribution") + # add title
  theme(plot.title = element_text(hjust = 0.5)) # center the title
ggplot(time2, aes(x=time, color=state, fill=state)) +
  geom_density(alpha=.3) + # density
  labs(title = "Time Density Distribution") +
  theme(plot.title = element_text(hjust = 0.5))
```

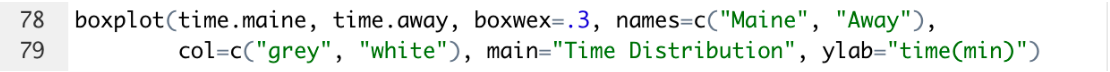




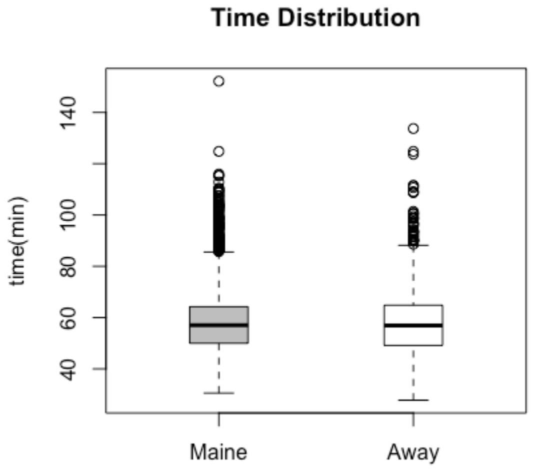
From these two diagrams, we can clearly see that the distributions of Maine Group and Away Group have similar shape. Because the participants are normal runners. But Maine Group has more data points.

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

Solution:



```r
boxplot(time.maine, time.away, boxwex=.3, names=c("Maine", "Away"),
        col=c("grey", "white"), main="Time Distribution", ylab="time(min)")
```

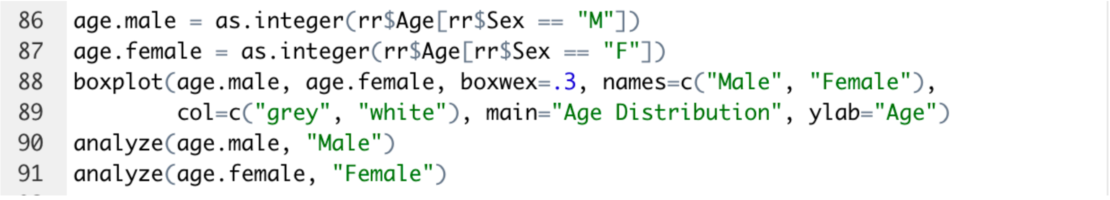


Conclusion: Beside the conclusions we get from (b), we can get that there are some outliers in both Maine and Away. The reason behind that deserve us to discover. Even though this might not be appropriate to use the results in part (d), but the data shows that many old runners participate in this activity and might earned the maximum score.

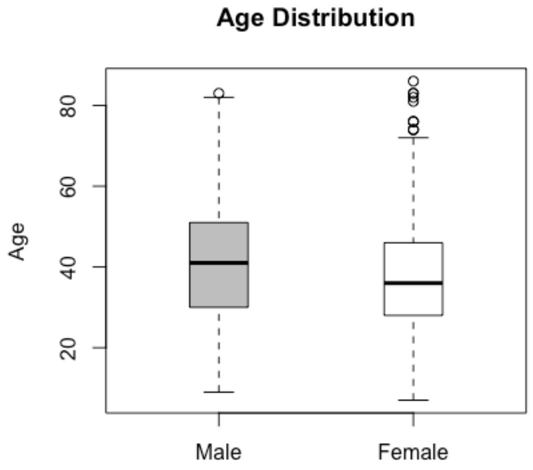
Histograms show more information about the general shape of distribution while the boxplots are more distinct and intuitive for comparation.

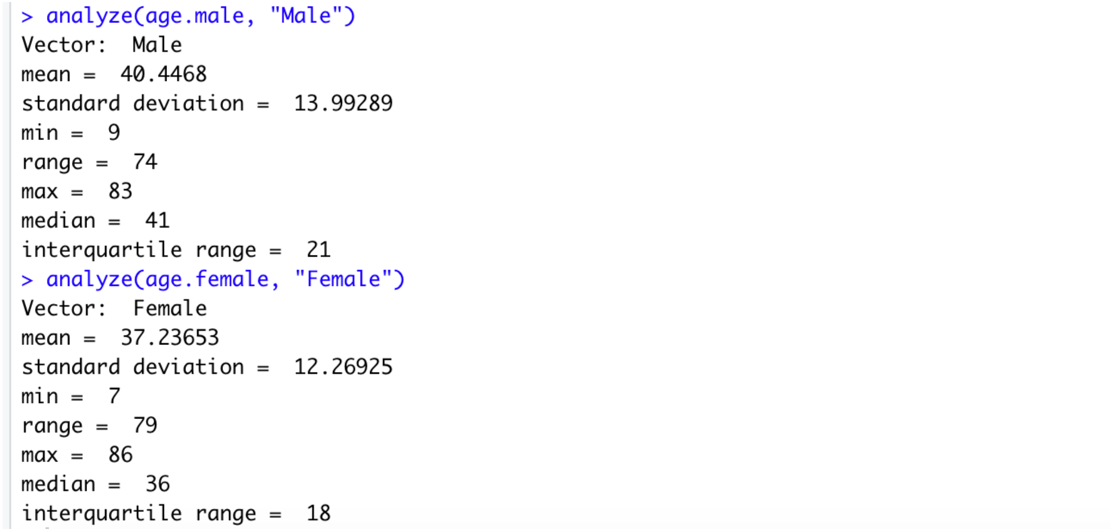
(d) Create side-by-side boxplots for the runners' ages (given in years) for male and female runners. What can we conclude about the two distributions? Back up your conclusions with relevant summary statistics, including mean, standard deviation, range, median, and interquartile range.

Solution:



```r
age.male = as.integer(rr$Age[rr$Sex == "M"])
age.female = as.integer(rr$Age[rr$Sex == "F"])
boxplot(age.male, age.female, boxwex=.3, names=c("Male", "Female"),
        col=c("grey", "white"), main="Age Distribution", ylab="Age")
analyze(age.male, "Male")
analyze(age.female, "Female")
```





```
> analyze(age.male, "Male")
Vector:  Male
mean =  40.4468
standard deviation =  13.99289
min =  9
range =  74
max =  83
median =  41
interquartile range =  21
> analyze(age.female, "Female")
Vector:  Female
mean =  37.23653
standard deviation =  12.26925
min =  7
range =  79
max =  86
median =  36
interquartile range =  18
```

Gender	Mean	SD	Min	Range	Max	Median	IQR
Male	40.45	13.99	9	74	83	41.00	21
Female	37.24	12.26	7	79	86	36.00	18

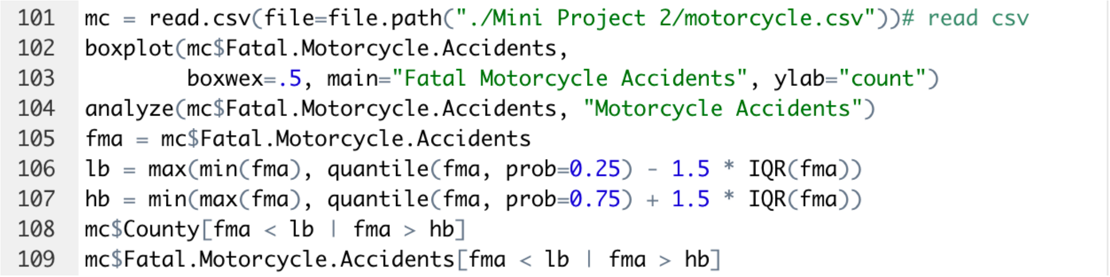
Table 3: statistics for runners of Male and Female

Conclusion: From the graph and table 3 we can conclude that the age for male runners who are interested in joining road race is older than Female runners. But there are still some older women over 70 are interested in joining road race. This might show that women have some advantages on bodies than men when they grow old.

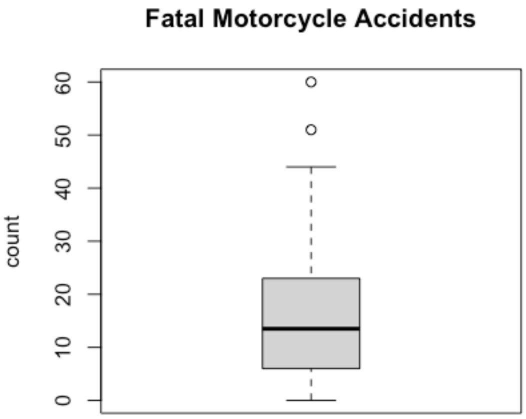
Question　2：

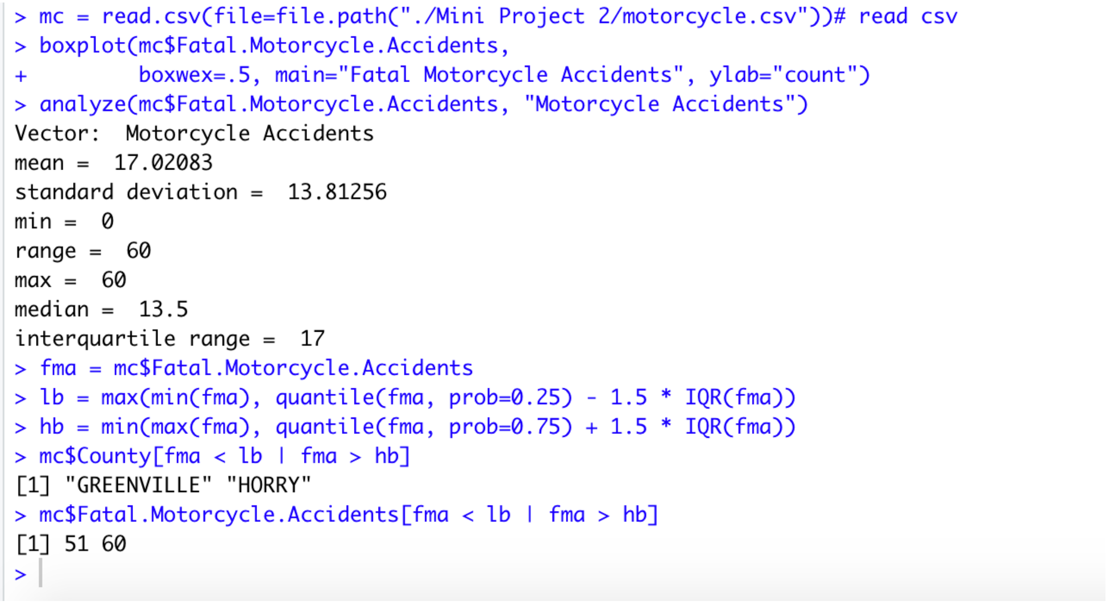
Consider the dataset motorcycle.csv posted on eLearning. It contains the number of fatal motorcycle accidents that occurred in each county of South Carolina during 2009. Create a boxplot of data and provide relevant summary statistics. Discuss the features of the data distribution. Identify which counties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

Solution:



```
mc = read.csv(file=file.path("./Mini Project 2/motorcycle.csv"))# read csv
boxplot(mc$Fatal.Motorcycle.Accidents,
        boxwex=.5, main="Fatal Motorcycle Accidents", ylab="count")
analyze(mc$Fatal.Motorcycle.Accidents, "Motorcycle Accidents")
fma = mc$Fatal.Motorcycle.Accidents
lb = max(min(fma), quantile(fma, prob=0.25) - 1.5 * IQR(fma))
hb = min(max(fma), quantile(fma, prob=0.75) + 1.5 * IQR(fma))
mc$County[fma < lb | fma > hb]
mc$Fatal.Motorcycle.Accidents[fma < lb | fma > hb]
```





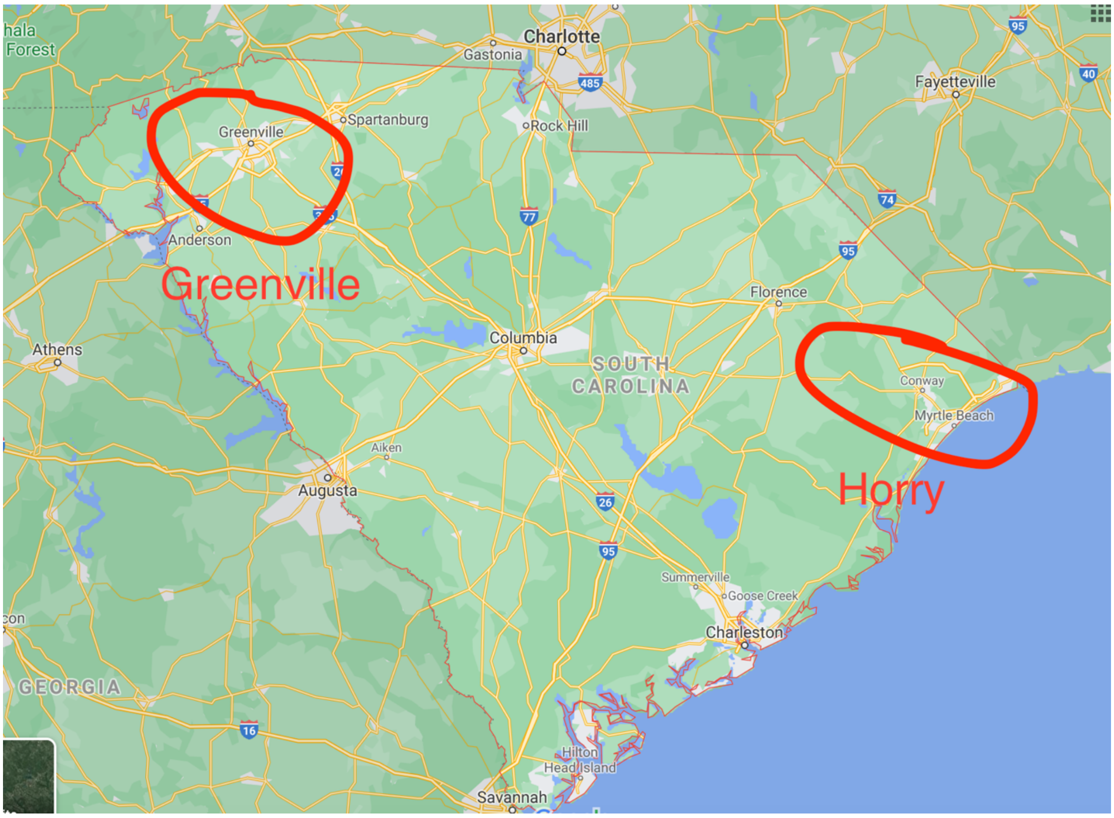
```
> mc = read.csv(file=file.path("./Mini Project 2/motorcycle.csv"))# read csv
> boxplot(mc$Fatal.Motorcycle.Accidents,
+         boxwex=.5, main="Fatal Motorcycle Accidents", ylab="count")
> analyze(mc$Fatal.Motorcycle.Accidents, "Motorcycle Accidents")
Vector:  Motorcycle Accidents
mean =  17.02083
standard deviation =  13.81256
min =  0
range =  60
max =  60
median =  13.5
interquartile range =  17
> fma = mc$Fatal.Motorcycle.Accidents
> lb = max(min(fma), quantile(fma, prob=0.25) - 1.5 * IQR(fma))
> hb = min(max(fma), quantile(fma, prob=0.75) + 1.5 * IQR(fma))
> mc$County[fma < lb | fma > hb]
[1] "GREENVILLE" "HORRY"
> mc$Fatal.Motorcycle.Accidents[fma < lb | fma > hb]
[1] 51 60
> 
```

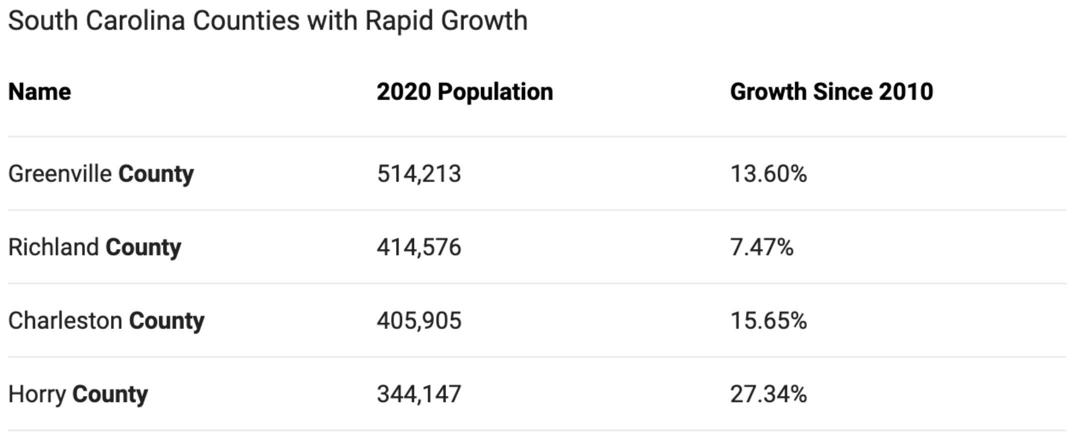
	Mean	SD	Min	Range	Max	Median	IQR
Accidents	17.02	13.81	0	60	60	13.50	17

Table 4: statistics for accidents number of South Carolina in 2009

Conclusion: From the graph and table 4 we can conclude that the average number of accidents is 17.02 and the median is 13.50. Compared with the maximum, which is 60, there will exist some extremely large number which correspond with the outliers in boxplot.

From the calculation, we found two outliers: Horry with number 60 and Greenville with number 51. I did some primary research about the possible reasons why these two cities have so many fatal motorcycle accidents.



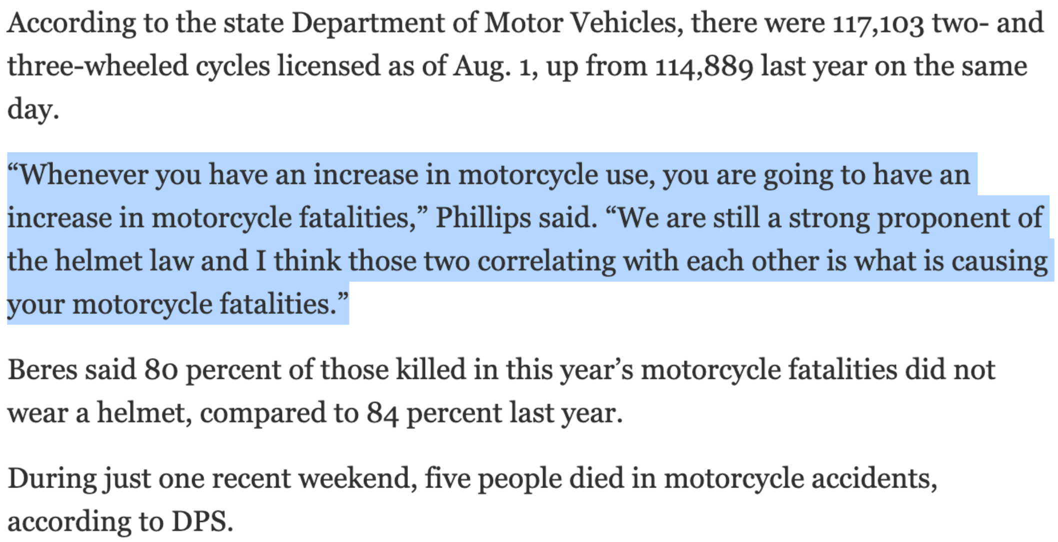


South Carolina Counties with Rapid Growth

Name	2020 Population	Growth Since 2010
Greenville **County**	514,213	13.60%
Richland **County**	414,576	7.47%
Charleston **County**	405,905	15.65%
Horry **County**	344,147	27.34%

The census shows that the Greenville County and Horry County have a large number of populations which might cause the large number of accidents.

Then I found a local news which also talked about the topic of fatal motorcycle accidents. And I made a screenshot and listed it below.



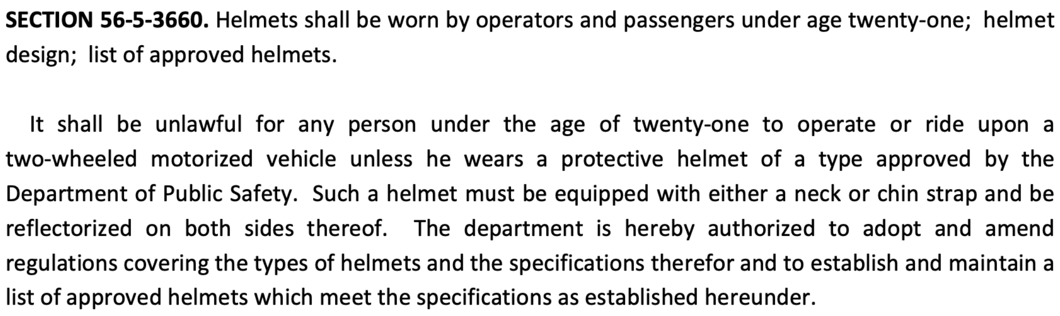
According to the state Department of Motor Vehicles, there were 117,103 two- and three-wheeled cycles licensed as of Aug. 1, up from 114,889 last year on the same day.

"Whenever you have an increase in motorcycle use, you are going to have an increase in motorcycle fatalities," Phillips said. "We are still a strong proponent of the helmet law and I think those two correlating with each other is what is causing your motorcycle fatalities."

Beres said 80 percent of those killed in this year's motorcycle fatalities did not wear a helmet, compared to 84 percent last year.

During just one recent weekend, five people died in motorcycle accidents, according to DPS.

I also googled the South Carolina Motorcycle Laws and found that only people under age twenty-one are forced to wear a helmet.



SECTION 56-5-3660. Helmets shall be worn by operators and passengers under age twenty-one; helmet design; list of approved helmets.

It shall be unlawful for any person under the age of twenty-one to operate or ride upon a two-wheeled motorized vehicle unless he wears a protective helmet of a type approved by the Department of Public Safety. Such a helmet must be equipped with either a neck or chin strap and be reflectorized on both sides thereof. The department is hereby authorized to adopt and amend regulations covering the types of helmets and the specifications therefor and to establish and maintain a list of approved helmets which meet the specifications as established hereunder.

To sum up, I found about two possible reasons for why Greenville and Horry County had so many fatal motorcycle accidents in 2009:

1). Large number of populations may cause large number of accidents if we assume that the probability of happening an accident is equal.

2) Many victims lost their lives because they did not wear helmets. Helmets are not forced to wear in South Carolina for grown driver and warm weather would contribute to it.

Further research is still required if this situation still occur now in South Carolina.