**B. Average Test Scores**

**1. Code**

# prepare ggplot library

library(ggplot2)

# set working directory and import US Data

setwd("C:/Users/jordan/Google Drive/Courses Spring 2018/STAT 350/STAT 350 Labs/Lab 02")

USData <- read.table("USData\_Spring.txt", header=TRUE, sep="\t")

# clean US Data

USData\_clean <- USData[complete.cases(USData),]

# use clean dataset

attach(USData\_clean)

### PART B: TestScore ###

# print five-number summary

FNS <- fivenum(TestScore)

FNS

# print 1.5 IQR limits and find outliers

IQR <- FNS[4] - FNS[2]

IF\_U <- FNS[4] + 1.5\*IQR # upper limit

IF\_L <- FNS[2] - 1.5\*IQR # lower limit

IF\_U

IF\_L

Outlier\_Index <- which(TestScore < IF\_L | TestScore > IF\_U)

Outliers <- TestScore[Outlier\_Index]

Outliers

# make modified boxplot

windows()

ggplot(USData\_clean, aes(x = "", y = TestScore)) +

stat\_boxplot(geom = "errorbar") +

geom\_boxplot() +

ggtitle("Test Scores in the US") +

stat\_summary(fun.y = mean, col = "black", geom = "point", size = 3)

# make histogram

windows()

xbar <- mean(TestScore)

xmed <- median(TestScore)

s <- sd(TestScore)

ggplot(USData\_clean, aes(TestScore)) +

geom\_histogram(aes(y = ..density..),

bins = sqrt(nrow(USData\_clean))+2,

fill = "grey", col = "black") +

geom\_density(col = "red", lwd = 1) +

stat\_function(fun = dnorm, args = list(mean = xbar,

sd = s),

col = "blue", lwd = 1) +

ggtitle("Test Scores in the US")

xbar

xmed

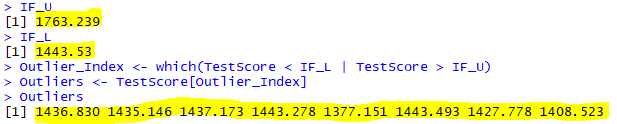
s

range(TestScore)

**2. Five-number summary**

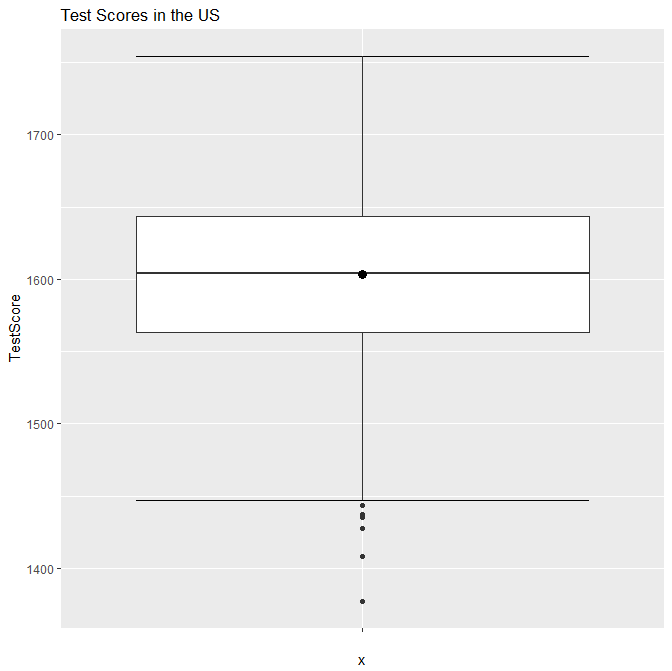


**3. Outliers**



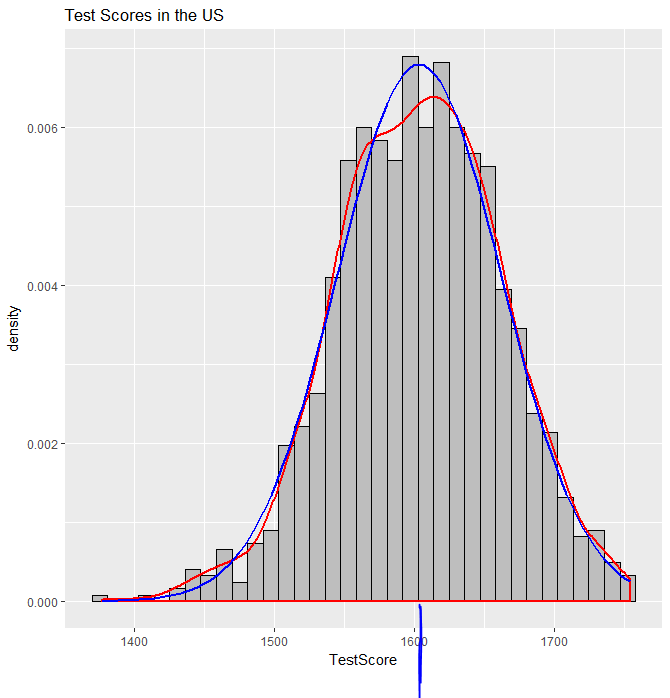
**Yes**, there are outliers. I know this because my R code shows that values exist outside of the 1.5 IQR upper and lower limits.

**4. Modified Boxplot**



This distribution is **left skewed**. We know this because **there are outliers**, and they all lie on the “left” of the plot (toward the lower TestScore values). The outliers are shown as small closed circles.

**5. Histogram**



This distribution is **left skewed**. We know this from the general shape as shown by the blue line, and because **there are outliers** on the left side of the histogram. The outliers are clear from the frequencies to the left of the bulk of the data.

**6. Histogram vs Boxplot**

The outliers shown in these figures appear to be the **same** – both the histogram and the boxplot show outliers in the lower TestScore range.

**7. Mean, Median, Standard Deviation**



The locations of the mean and median are roughly indicated in the histogram figure above. Note that these values are extremely close together, so much so that the lines in the figure are essentially indistinguishable.

**8. Median and Mean**

As noted above, the mean and median are **very close** on the scale of the histogram. However, we do not need a figure to know this – the Test Scores range from approximately 1,377 to approximately 1,754, as shown in the output below. Mathematically, the median and mean differ by only approximately 0.5 – this is a very small amount compared to the total range of data.



**9. Median or Mean**

These values are very similar, but if I had to choose I would use the **median** to characterize the data, because this is more resistant to outliers. If a few data points for test score were extremely low or extremely high, it could have a significant effect on the mean, but the median would be relatively unaffected.

**B. Average Test Scores**

**1. Code**

### PART C: Larcenies ###

# print five-number summary

FNS <- fivenum(LarceniesPerPopulation)

FNS

# print 1.5 IQR limits and find outliers

IQR <- FNS[4] - FNS[2]

IF\_U <- FNS[4] + 1.5\*IQR # upper limit

IF\_L <- FNS[2] - 1.5\*IQR # lower limit

IF\_U

IF\_L

Outlier\_Index <- which(LarceniesPerPopulation < IF\_L | LarceniesPerPopulation > IF\_U)

Outliers <- LarceniesPerPopulation[Outlier\_Index]

Outliers

# make modified boxplot

windows()

ggplot(USData\_clean, aes(x = "", y = LarceniesPerPopulation)) +

stat\_boxplot(geom = "errorbar") +

geom\_boxplot() +

ggtitle("Larcenies per Population in the US") +

stat\_summary(fun.y = mean, col = "black", geom = "point", size = 3)

# make histogram

windows()

xbar <- mean(LarceniesPerPopulation)

xmed <- median(LarceniesPerPopulation)

s <- sd(LarceniesPerPopulation)

ggplot(USData\_clean, aes(LarceniesPerPopulation)) +

geom\_histogram(aes(y = ..density..),

bins = sqrt(nrow(USData\_clean))+2,

fill = "grey", col = "black") +

geom\_density(col = "red", lwd = 1) +

stat\_function(fun = dnorm, args = list(mean = xbar,

sd = s),

col = "blue", lwd = 1) +

ggtitle("Larcenies per Population in the US")

xbar

xmed

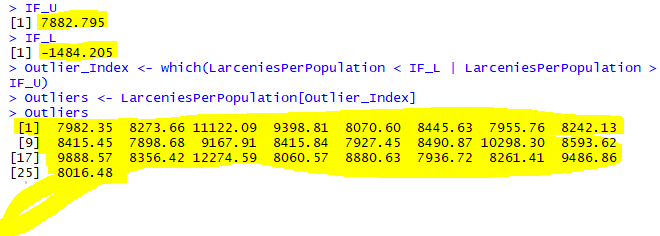
s

range(LarceniesPerPopulation)

**2. Five-number summary**

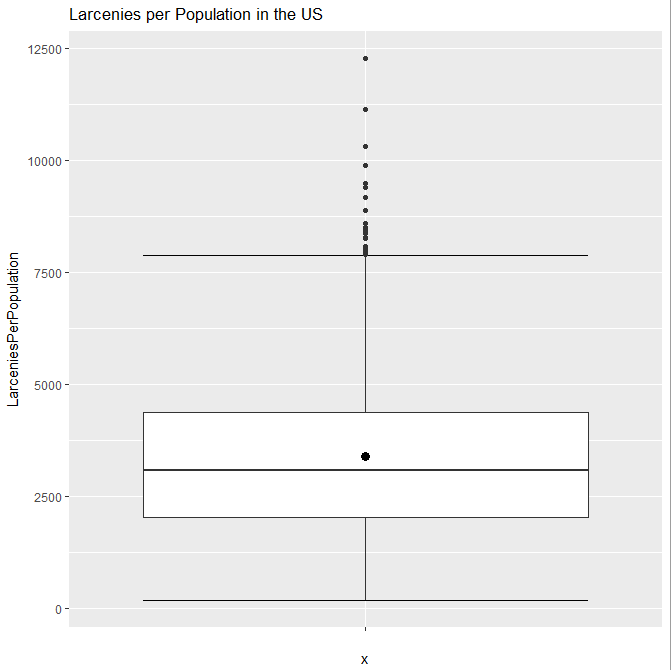


**3. Outliers**



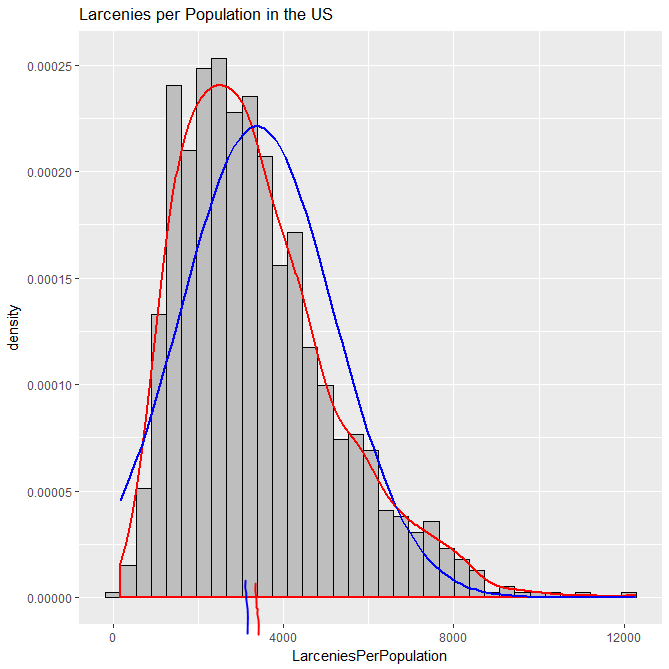
**Yes, there are outliers**. I know this because my R code shows that values exist outside of the 1.5 IQR upper and lower limits.

**4. Modified Boxplot**



This distribution is **right skewed**. This is clear because **there are outliers** in the “right” region of the plot – that is, for very high LarceniesPerPopulation values.

**5. Histogram**

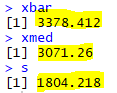


This histogram is **right skewed**. This is clear from the general shape of the blue and red lines, as well as the fact that **outliers are present** in the right of the graph, as is clear from the many frequencies that lie beyond the bulk of the data.

**6. Histogram vs Boxplot**

The data points that we see as outliers in the boxplot appear to show up as outliers again in the histogram – that is, the two figures show the **same** outliers.

**7. Mean, Median, Standard Deviation**



The mean and median are marked approximately on the histogram above – the median by the blue line and the mean by the red line.

**8. Median and Mean**

The median is **rather close** to the mean, but not extremely close. They are far enough apart to be distinguishable in the histogram above. However, the command below shows the data ranges from 170.16 to 12274.59, in light of which the difference between the mean and median of roughly 300 is rather small.



**9. Median or Mean**

If I only had one measure to describe the central location of this data, I would use the **median**. That is because, while the mean is significantly influence by outliers (as noted by the higher value of the mean in the presence of many upper outliers), the median is relatively resistant, providing a better sense of the data’s center.

**D. Bonus**

**1. Code**

### Part D: Bonus ###

so <- length(which(Region == "SO"))

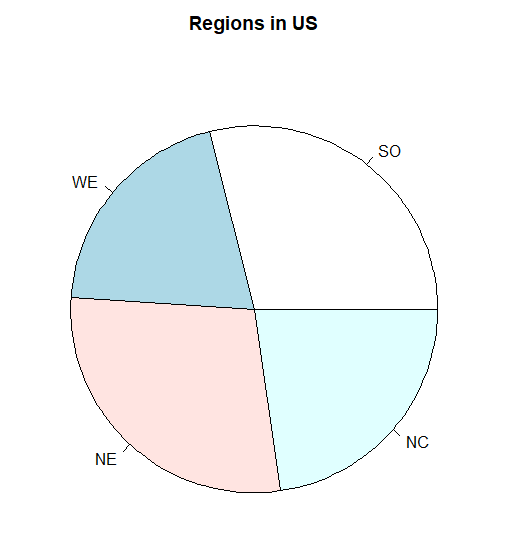
we <- length(which(Region == "WE"))

ne <- length(which(Region == "NE"))

nc <- length(which(Region == "NC"))

pie(c(so, we, ne, nc), c("SO", "WE", "NE", "NC"), main="Regions in US")

**2. Pie Chart**



From this chart, we can see that the US has a fairly even number of regions in each category, though there are somewhat more regions in the South (SO) and Northeast (NE) categories, with somewhat fewer in the West (WE) and North Central (NC) categories.