Jordan Mayer
January 25, 2018
STAT 350
Lab 02

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) +
```

Jordan Mayer
January 25, 2018
STAT 350
Lab 02

```
  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

```
> # print five-number summary
> FNS <- fivenum(TestScore)
> FNS
[1] 1377.151 1563.421 1604.037 1643.348 1754.276
```

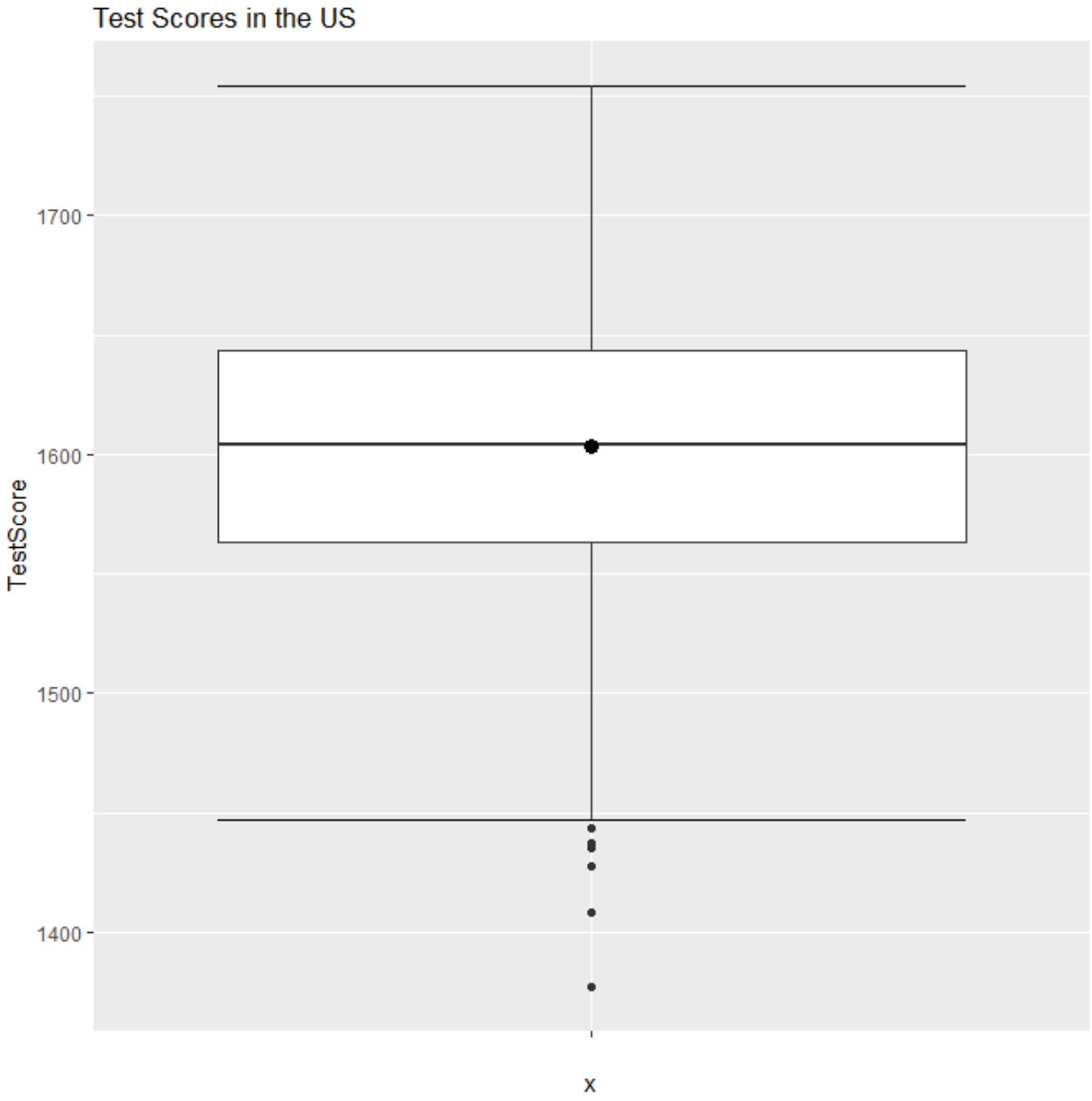
3. Outliers

```
> IF_U
[1] 1763.239
> IF_L
[1] 1443.53
> Outlier_Index <- which(TestScore < IF_L | TestScore > IF_U)
> Outliers <- TestScore[Outlier_Index]
> Outliers
[1] 1436.830 1435.146 1437.173 1443.278 1377.151 1443.493 1427.778 1408.523
```

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.

Jordan Mayer
January 25, 2018
STAT 350
Lab 02

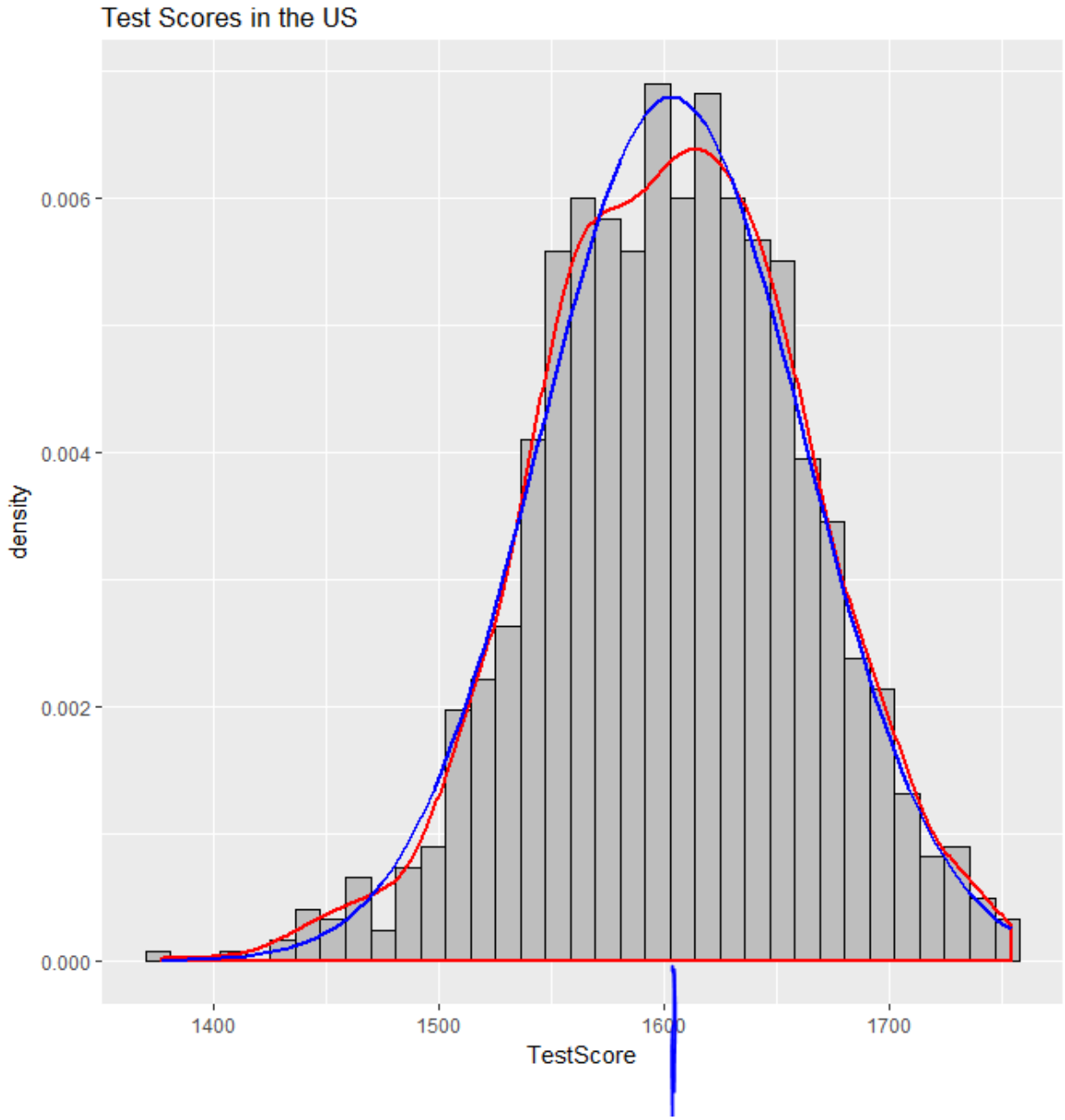
4. Modified Boxplot



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.

Jordan Mayer
January 25, 2018
STAT 350
Lab 02

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.

Jordan Mayer
January 25, 2018
STAT 350
Lab 02

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

```
> xbar
[1] 1603.538
> xmed
[1] 1604.037
> s
[1] 58.74789
```

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.

```
> range(TestScore)
[1] 1377.151 1754.276
```

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.

Jordan Mayer
January 25, 2018
STAT 350
Lab 02

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)
```

Jordan Mayer
January 25, 2018
STAT 350
Lab 02

2. Five-number summary

```
> FNS
[1]  170.16 2028.42 3071.26 4370.17 12274.59
```

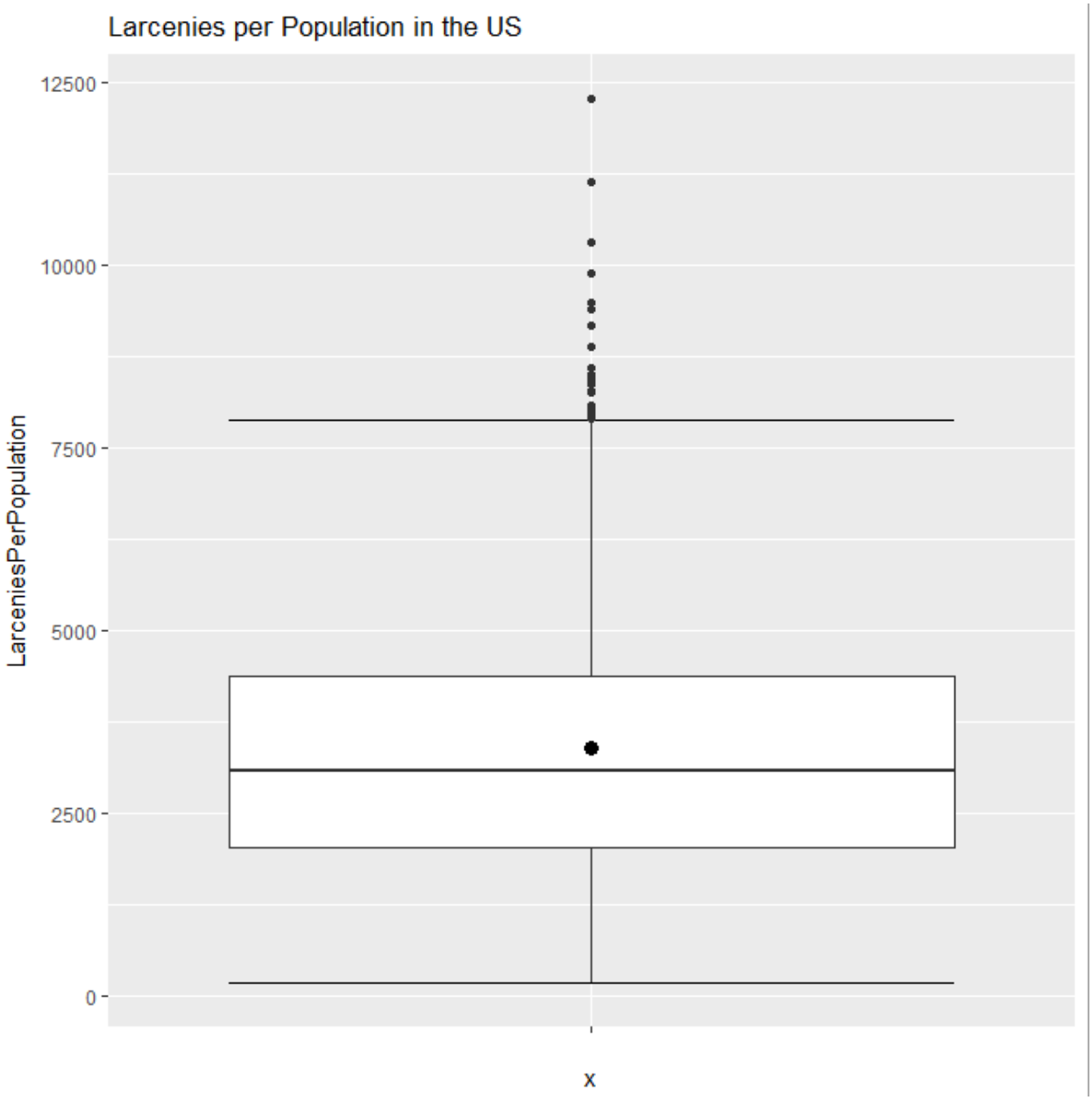
3. Outliers

```
> IF_U
[1] 7882.795
> IF_L
[1] -1484.205
> Outlier_Index <- which(LarceniesPerPopulation < IF_L | LarceniesPerPopulation >
IF_U)
> Outliers <- LarceniesPerPopulation[Outlier_Index]
> Outliers
 [1]  7982.35  8273.66 11122.09  9398.81  8070.60  8445.63  7955.76  8242.13
 [9]  8415.45  7898.68  9167.91  8415.84  7927.45  8490.87 10298.30  8593.62
[17]  9888.57  8356.42 12274.59  8060.57  8880.63  7936.72  8261.41  9486.86
[25]  8016.48
```

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.

Jordan Mayer
January 25, 2018
STAT 350
Lab 02

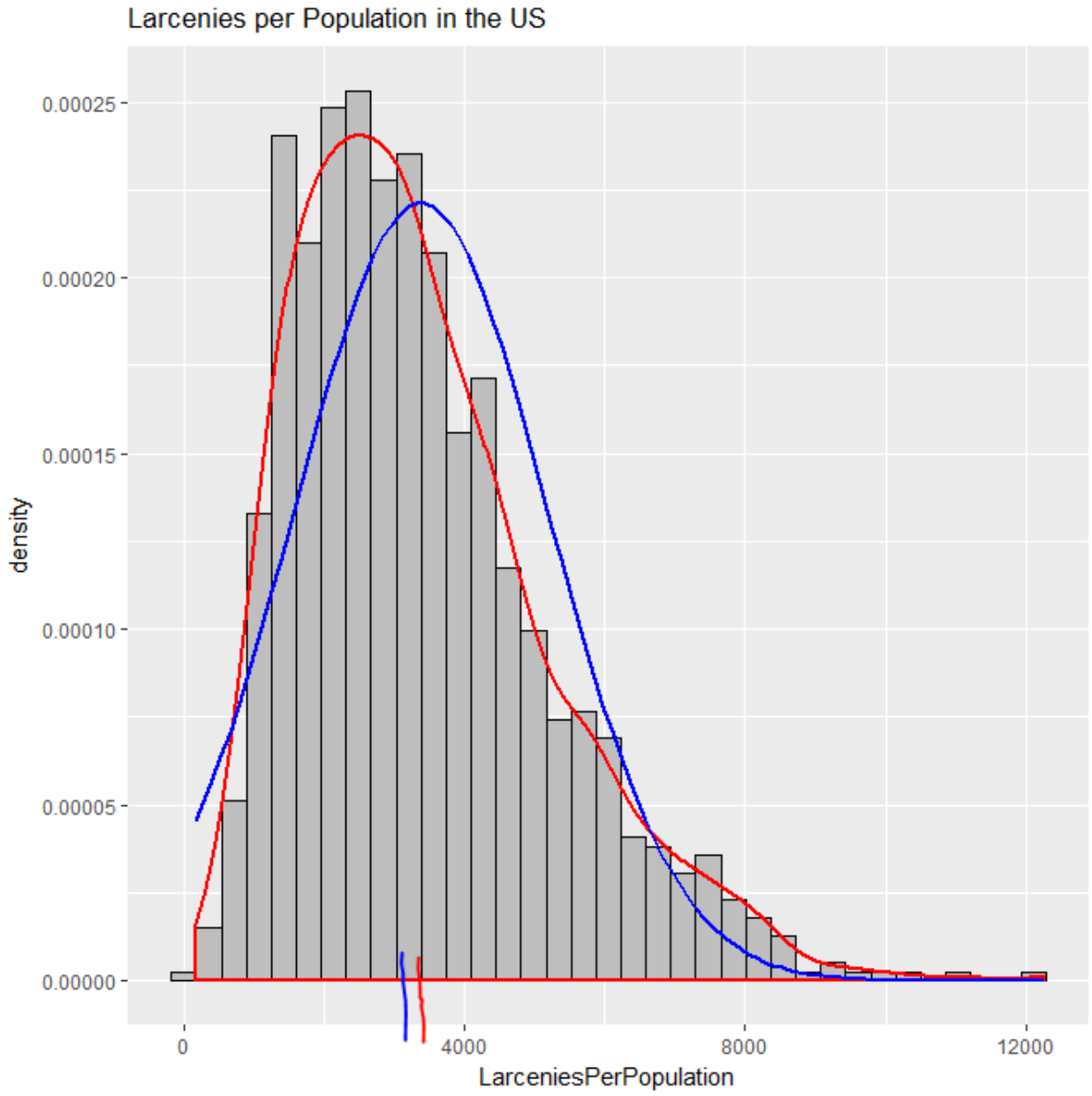
4. Modified Boxplot



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.

Jordan Mayer
January 25, 2018
STAT 350
Lab 02

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.

Jordan Mayer
January 25, 2018
STAT 350
Lab 02

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

```
> xbar
[1] 3378.412
> xmed
[1] 3071.26
> s
[1] 1804.218
```

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.

```
> range(LarceniesPerPopulation)
[1]   170.16 12274.59
```

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.

Jordan Mayer
January 25, 2018
STAT 350
Lab 02

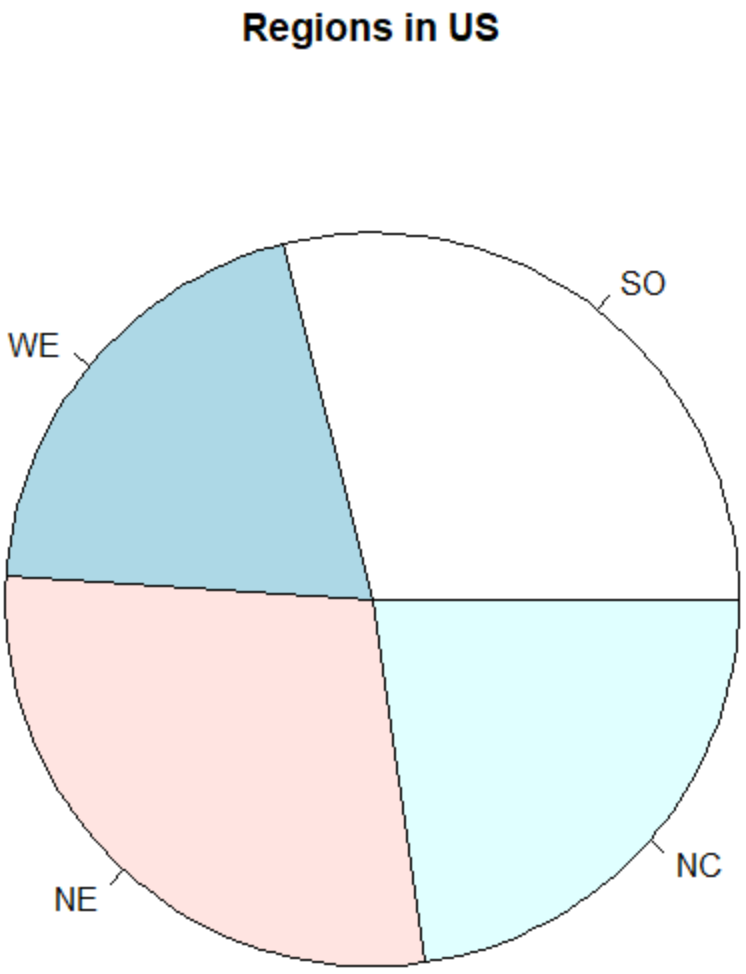
D. Bonus

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")
```

Jordan Mayer
January 25, 2018
STAT 350
Lab 02

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.

Jordan Mayer
January 25, 2018
STAT 350
Lab 02