Project 2

Khiem Pham

```
#Install devtools and cardiomoon/webr to plot p-value
install.packages('devtools', repos="http://cran.us.r-project.org", type="source")
```

```
## Installing package into 'C:/Users/Khiem/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## also installing the dependencies 'usethis', 'lifecycle', 'rlang'
```

```
## Warning in install.packages("devtools", repos = "http://cran.us.r-
## project.org", : installation of package 'rlang' had non-zero exit status
```

```
## Warning in install.packages("devtools", repos = "http://cran.us.r-
## project.org", : installation of package 'lifecycle' had non-zero exit status
```

```
## Warning in install.packages("devtools", repos = "http://cran.us.r-
## project.org", : installation of package 'usethis' had non-zero exit status
```

```
## Warning in install.packages("devtools", repos = "http://cran.us.r-
## project.org", : installation of package 'devtools' had non-zero exit status
```

```
library('devtools')
```

```
## Loading required package: usethis
```

```
## Warning: package 'usethis' was built under R version 4.0.3
```

```
devtools::install_github("cardiomoon/webr")
```

```
## Skipping install of 'webr' from a github remote, the SHA1 (6e411bdc) has not changed since la
st install.
##   Use `force = TRUE` to force installation
```

```
# Read data
data <- read.csv("Data Facebook Friends.csv", header = F)
unlistData <- unlist(data)

#Structure
str(unlistData)
```

```
##  Named int [1:294] 317 1225 1192 0 715 1066 485 609 658 1640 ...
##  - attr(*, "names")= chr [1:294] "V11" "V12" "V13" "V14" ...
```

```
#Define variables for reference

#Sample Median
sampleMedian <- median(unlistData)
sampleMedian
```

```
## [1] 747.5
```

```
#Sample Mean
sampleMean <- mean(unlistData)
sampleMean
```

```
## [1] 751.4864
```

```
#Standard Deviation
sd_data <- sd(unlistData)
sd_data
```

```
## [1] 357.1355
```

Sample Median: 747.50

Sample Mean: 751.49

Standard Deviation: 357.14

Number of Observations: 294

a) Write appropriate hypothesis for the text

Null hypothesis H0: The mean of facebook friends is the same as the student's high school, which is 649

H0: μ = 649

Alternative hypothesis HA: The mean of facebook friends is higher at the student's school.

HA: μ > 649

b) The test statistics is

$x^- - \mu/(\sigma/\sqrt{n})$

x is the sample mean
μ is the population mean
σ is the standard deviation of population
n is the number of observations

```
#Population Mean - from the question
populationMean <- 649

# Divide mean by x
TS <- (sampleMean-populationMean)/(sd_data/sqrt(294))
TS
```

```
## [1] 4.920473
```

The test statistic is 4.92

c) The p-value of the test statistics is

```
require(moonBook)
```

```
## Loading required package: moonBook
```

```
## Warning: package 'moonBook' was built under R version 4.0.3
```

```
require(webr)
```

```
## Loading required package: webr
```

```
# Manually solve to find p value
p <- pt(-abs(TS), df=293)
p
```

```
## [1] 7.212136e-07
```

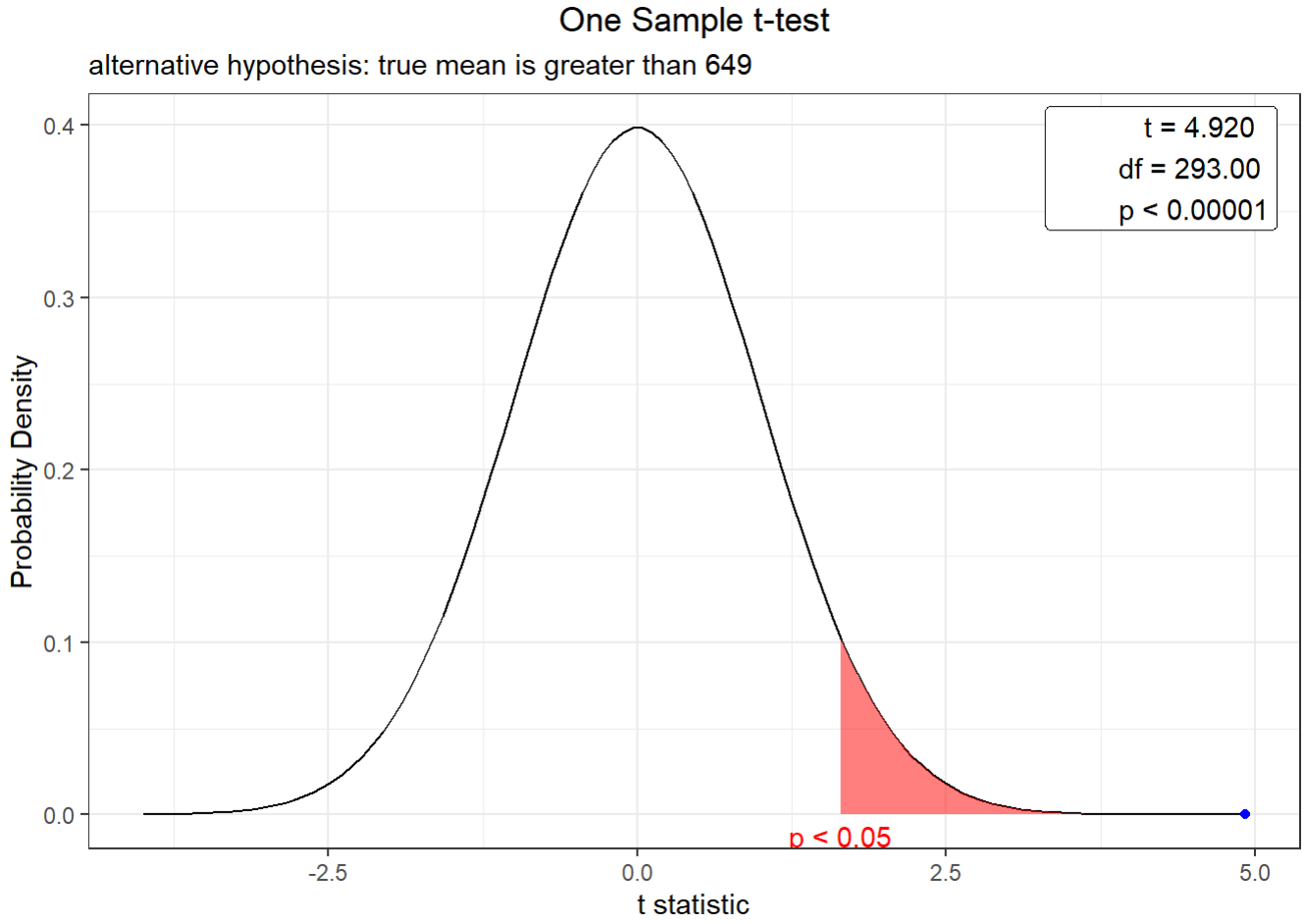
```
# Using t-test to double-check p value
test <- t.test(unlistData, mu=649, alternative="greater")
test
```

```
##
##  One Sample t-test
##
## data:  unlistData
## t = 4.9205, df = 293, p-value = 7.212e-07
## alternative hypothesis: true mean is greater than 649
## 95 percent confidence interval:
##  717.1178      Inf
## sample estimates:
## mean of x
##  751.4864
```

p-value = 7.21e-07

p-value < .01, so the test is highly significant.

```
# Plot p value
plot(test)
```



d) State your decision using both rejection region approach and p-value approach: (Use alpha = .05)

Rejection Region approach

Right-Tailed Test: Reject H0 if the test statistic is greater than or equal to the critical value

P-Value Approach

If p value is less than or equal to alpha, enough evidence to reject null hypothesis

If p value greater than alpha, not enough evidence to reject null hypothesis

```
# Degrees of freedom: Number of items - 1
df <-293

# Calculate Critical Value of Right Tail
CVR <- qt(.05,df, lower.tail = FALSE)
CVR
```

```
## [1] 1.650071
```

Critical Value of Right Tail: 1.65
Test statistic: 4.92

Rejection Region Approach

Right-Tailed Test: Since test statistics is greater than cvr, we reject the null hypothesis.

p-value: 7.21e-07
alpha: .05

P-Value Approach

Since p value is less than alpha, we reject the null hypothesis.

e) Interpret your decision made in part d) There –?– sufficient evidence to conclude that the mean number of Facebook friends for students at the school is –?– 649

Based on the results from d, we have enough evidence to reject the null hypothesis.

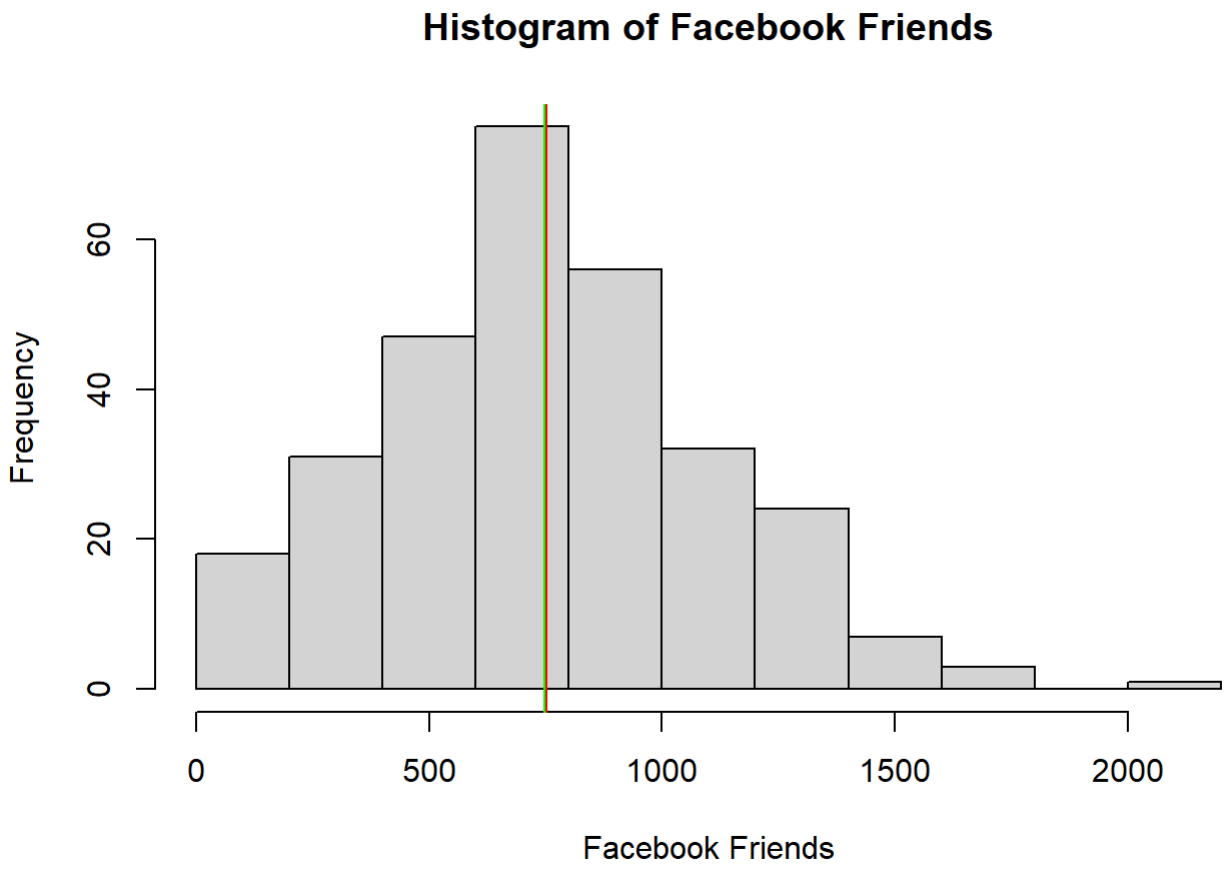
There is sufficient evidence to conclude that the mean number of Facebook friends for students at the school is greater than 649.

f) Provide a histogram of the sample

```
hist(unlistData, xlab = "Facebook Friends", main="Histogram of Facebook Friends")

# Plot Median
abline(v=sampleMedian, col = "Green")

# Plot Mean
abline(v=sampleMean, col = "Red")
```



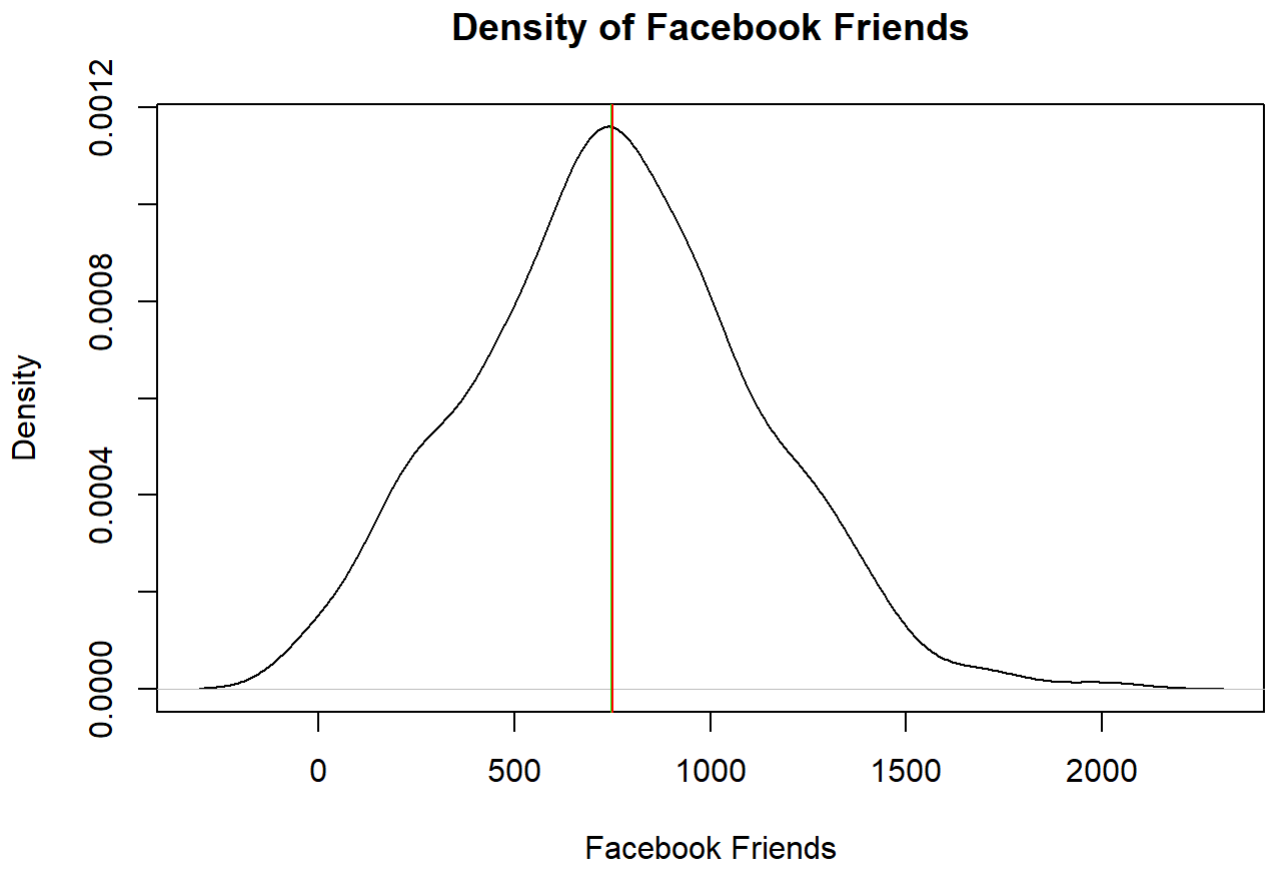
g) Comment on distribution shape of the sample

```
densityData <- density(unlistData)

plot(densityData, xlab = "Facebook Friends", main = "Density of Facebook Friends")

# Plot Median
abline(v=sampleMedian, col = "Green")

# Plot Mean
abline(v=sampleMean, col = "Red")
```



The distribution looks like a standard normal distribution with a bell-shaped curve. It seems to be right-skewed.

h) Apply an appropriate test to statistically confirm if data is normally distributed. State name of the test and the null and alternative hypothesis for this test

Null hypothesis H0: The data are normally distributed

Alternative hypothesis HA: The data are not normally distributed

If p value is less than alpha, H0 is rejected and data is not normally distributed.

If p value is greater than alpha, we do not reject H0 and data is normally distributed.

Shapiro-Wilk's method

```
shapiro.test(unlistData)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  unlistData
## W = 0.99231, p-value = 0.1324
```

p value: 0.13
alpha: .05

Since p value is greater than alpha, we don't reject the null hypothesis and can conclude that the data is normally distributed.

i) Calculate 90% and 99% Confidence Intervals for the population mean. Interpret each of the Confidence Intervals.

```
n <- 294

# Manually Calculate 90% Confidence Intervals
critical90 <- qt(p=.05, df = 293, lower.tail=FALSE)

confidence90high <- sampleMean - (critical90 * (sd_data/sqrt(n)))
confidence90high
```

```
## [1] 717.1178
```

```
confidence90low <- sampleMean + (critical90 * (sd_data/sqrt(n)))
confidence90low
```

```
## [1] 785.855
```

```
# Manually Calculate 99% Confidence Intervals
critical99 <- qt(p=.01/2, df=293, lower.tail=FALSE)

confidence99low <- sampleMean - (critical99 * (sd_data/sqrt(n)))
confidence99low
```

```
## [1] 697.4839
```

```
confidence99high <- sampleMean + (critical99 * (sd_data/sqrt(n)))
confidence99high
```

```
## [1] 805.4889
```

```
# Double-check if 90% confidence intervals are correct
intNinety <- t.test(unlistData, mu=sampleMean, conf.level = 0.90)
intNinety
```

```
##
##  One Sample t-test
##
## data:  unlistData
## t = 0, df = 293, p-value = 1
## alternative hypothesis: true mean is not equal to 751.4864
## 90 percent confidence interval:
##  717.1178 785.8550
## sample estimates:
## mean of x
##  751.4864
```

```
# Double-check if 99% confidence intervals are correct
intNinetyNine <- t.test(unlistData, mu=sampleMean, conf.level = 0.99)
intNinetyNine
```

```
##
##  One Sample t-test
##
## data:  unlistData
## t = 0, df = 293, p-value = 1
## alternative hypothesis: true mean is not equal to 751.4864
## 99 percent confidence interval:
##  697.4839 805.4889
## sample estimates:
## mean of x
##  751.4864
```

For 90% Confidence Interval, the average mean of facebook friends at his school is between 717.12 and 785.86. This means that we are 90% confident that this interval contains the mean number of friends at the school. If the student took many random samples, 90% of the intervals will include the true population mean of facebook friends at this school.

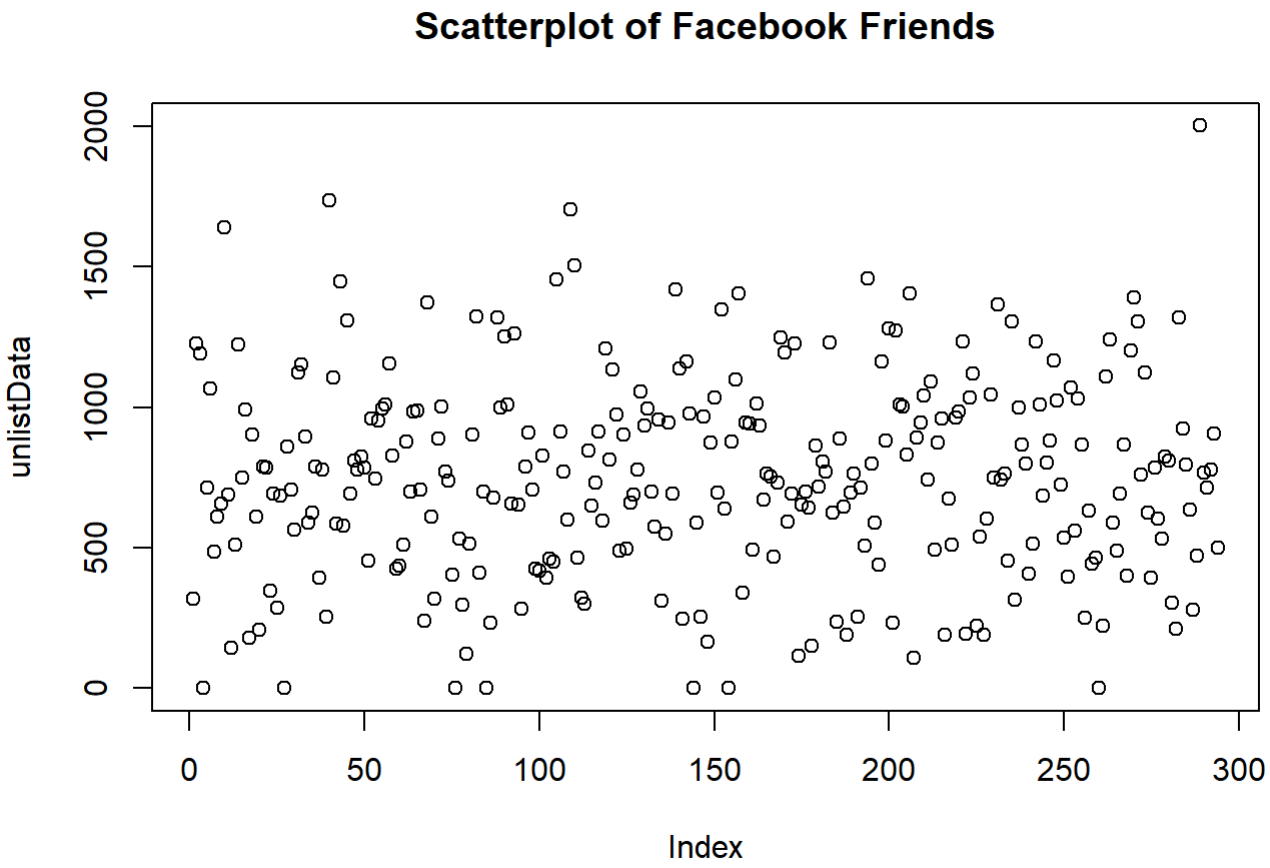
For 99% Confidence Interval, the average mean of facebook friends at his school is between 697.48 and 805.49. This means that we are 99% confident that this interval contains the mean number of friends at the school. If the student took many random samples, 99% of the intervals will include the true population mean of facebook friends at his school.

j) Compare the 90% and 99% Confidence Intervals for the population mean. Which one gives a better idea about what the population mean is? Explain

90% confidence has lower difference between the intervals. 99% confidence has higher difference between the intervals. 90% confidence intervals probably gives a better, more accurate idea about what the population mean is since it has a narrower range.

k) Provide scatter plot of the sample observations. Identify the outliers in this sample. (Hint: Computer z-scores of your observations and consider any z-scores bigger than 3 or less than -3 as outliers)

```
plot(unlistData, main="Scatterplot of Facebook Friends")
```




```
# Calculate z-scores
zs <- (unlistData - sampleMean)/sd_data

# Find rows that have z scores greater than 3
which(zs > 3)
```

```
## V1289
##   289
```

```
# Find rows that have z scores less than -3
which(zs < -3)
```

```
## named integer(0)
```

```
# Find z score at row 289
zs[289]
```

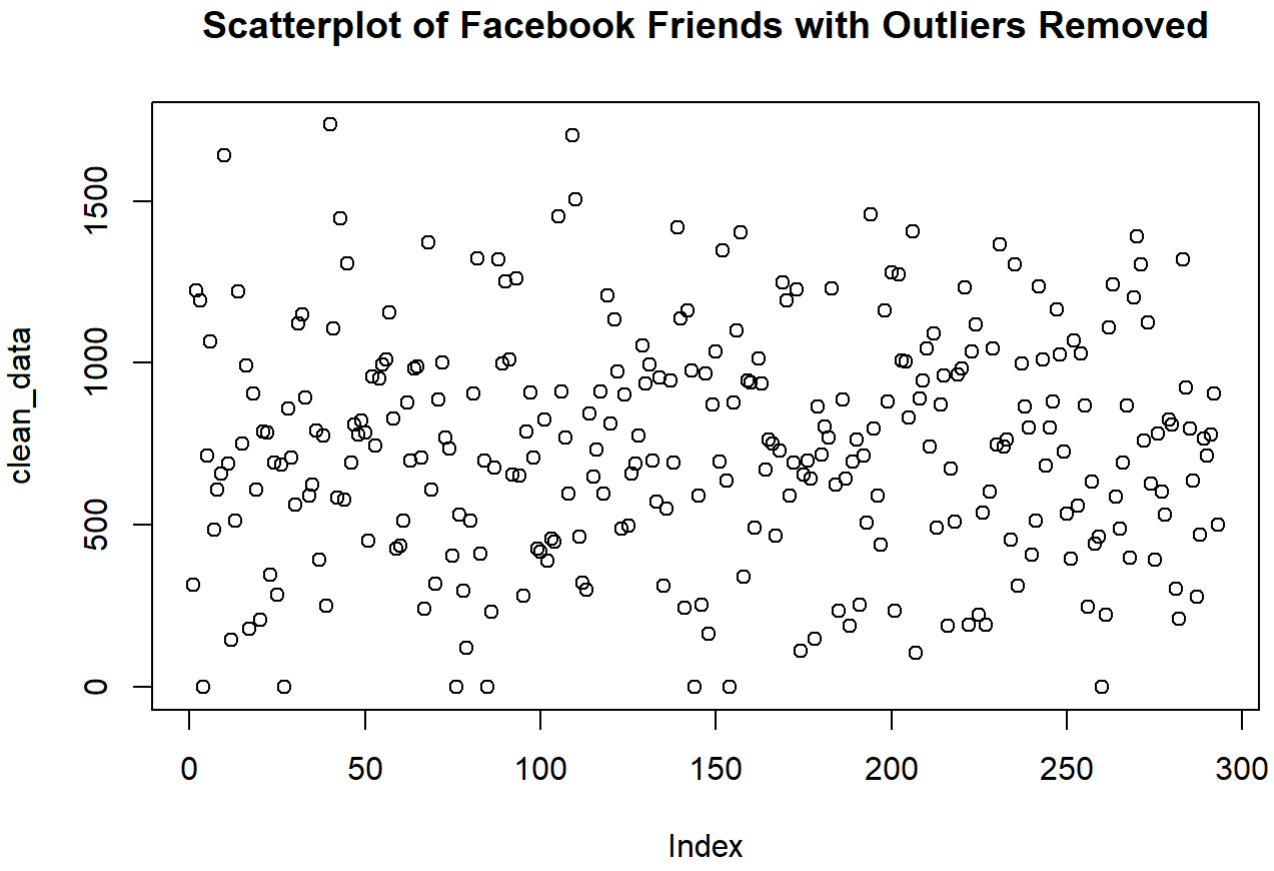
```
##   V1289
## 3.50151
```

There seems to be an outlier at row 289 with a z-score of 3.50.

l) Drop the identified outliers in part k) and provide scatter plot of the observations again.

```
# Data cleaned of outliers
clean_data <- unlistData[zs>-3 & zs<3]

plot(clean_data, main="Scatterplot of Facebook Friends with Outliers Removed")
```



m) After dropping the outliers, answer parts h), i) and j) again. Explain if removing the outliers affected your findings in part h), i), j)? Explain all the statistical

differences that you might have observed, if any

m-h) Apply an appropriate test to statistically confirm if data is normally distributed. State name of the test and the null and alternative hypothesis for this test

Null Hypothesis H0: The data is normally distributed

Alternative hypothesis: The data are not normally distributed

If p value is less than alpha, H0 is rejected and data is not normally distributed.

If p value is greater than alpha, H0 isn't rejected and the data is normally distributed.

Shapiro-Wilk's Method

```
shapiro.test(clean_data)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  clean_data
## W = 0.99322, p-value = 0.208
```

p value: .21
alpha: .05

Since p value is greater than alpha, we don't reject the null hypothesis and the data is normally distributed.

m-i) Calculate 90% and 99% Confidence Intervals for the population mean. Interpret each of the Confidence Intervals.

```
n <- 293

sd_clean <- sd(clean_data)
sampleMean_clean <- mean(clean_data)

# Manually Calculate 90% Confidence Intervals
critical90 <- qt(p=.05, df = 292, lower.tail=FALSE)

confidence90high_clean <- sampleMean_clean - (critical90 * (sd_clean/sqrt(n)))
confidence90high_clean
```

```
## [1] 713.4637
```

```
confidence90low_clean <- sampleMean_clean + (critical90 * (sd_clean/sqrt(n)))
confidence90low_clean
```

```
## [1] 780.9731
```

```
# Manually Calculate 99% Confidence Intervals
critical99 <- qt(p=.01/2, df=292, lower.tail=FALSE)

confidence99low_clean <- sampleMean_clean - (critical99 * (sd_clean/sqrt(n)))
confidence99low_clean
```

```
## [1] 694.1799
```

```
confidence99high_clean <- sampleMean_clean + (critical99 * (sd_clean/sqrt(n)))
confidence99high_clean
```

```
## [1] 800.2569
```

```
# Double-check if 90% confidence intervals are correct
intNinety_clean <- t.test(clean_data, mu=mean(clean_data), conf.level = 0.90)
intNinety_clean
```

```
##
##  One Sample t-test
##
## data:  clean_data
## t = 0, df = 292, p-value = 1
## alternative hypothesis: true mean is not equal to 747.2184
## 90 percent confidence interval:
##  713.4637 780.9731
## sample estimates:
## mean of x
##  747.2184
```

```
# Double-check if 99% confidence intervals are correct
intNinetyNine_clean <- t.test(clean_data, mu=mean(clean_data), conf.level = 0.99)
intNinetyNine_clean
```

```
##
##  One Sample t-test
##
## data:  clean_data
## t = 0, df = 292, p-value = 1
## alternative hypothesis: true mean is not equal to 747.2184
## 99 percent confidence interval:
##  694.1799 800.2569
## sample estimates:
## mean of x
##  747.2184
```

For 90% Confidence Interval, the average mean of facebook friends at his school is between 713.46 and 780.97. This means that we are 90% confident that this interval contains the mean number of friends at the school. If the student took many random samples, 90% of the intervals will include the true population mean of facebook friends

at his school

For 99% Confidence Interval, the average mean of facebook friends at his school is between 694.18 and 800.26. This means that we are 99% confident that this interval contains the mean number of friends at the school. If the student took many random samples, 99% of the intervals will include the true population mean of facebook friends at his school.

m-j) Compare the 90% and 99% Confidence Intervals for the population mean. Which one gives a better idea about what the population mean is? Explain

90% confidence has lower difference between the intervals. 99% confidence has higher difference between the intervals. 90% confidence intervals probably gives a better, more accurate idea about what the population mean is since it has a narrower range.

m) - Conclusion

Part h - No outlier removed

p value: 0.13 alpha: .05

Since p value is greater than alpha, we don't reject the null hypothesis and can conclude that the data is normally distributed.

Part m-h - Outlier removed

p value: .21 alpha: .05

Since p value is greater than alpha, we don't reject the null hypothesis and can conclude that the data is normally distributed.

Part h Analysis

The p value increased slightly when we removed the outlier, but they both still had the same results, in which we don't reject the null hypothesis and conclude the data is normally distributed.

Part i - No outlier removed

For 90% Confidence Interval, the average mean of facebook friends at his school is between 717.12 and 785.86.

For 99% Confidence Interval, the average mean of facebook friends at his school is between 697.48 and 805.49.

Part m-i - Outlier removed

For 90% Confidence Interval, the average mean of facebook friends at his school is between 713.46 and 780.97

For 99% Confidence Interval, the average mean of facebook friends at his school is between 694.18 and 800.26.

Part i Analysis

Removing the outlier made the 90% and 99% intervals narrower and shifted them down a bit.

Part j vs m-j Analysis

This part remained the same. The 90% confidence still had narrower intervals than the 99% confidence so 90% confidence still gives us a better, more accurate idea about what the population mean is.

```
# Original Data
summary(unlistData)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   506.8   747.5   751.5   983.8  2002.0
```

```
sd(unlistData)
```

```
## [1] 357.1355
```

```
# Data with outlier removed
summary(clean_data)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   506.0   746.0   747.2   983.0  1735.0
```

```
sd(clean_data)
```

```
## [1] 350.1555
```

Looking at the above data, it shows that removing the outlier slightly decreased the mean, median, and standard deviation of the data.