Homework 3

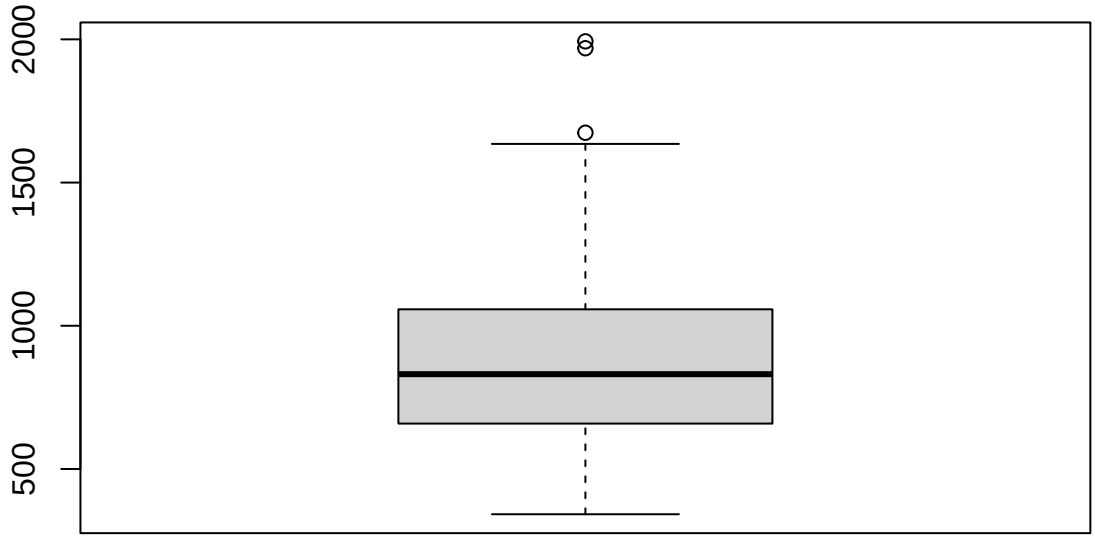
2024-01-29

Question 5.1

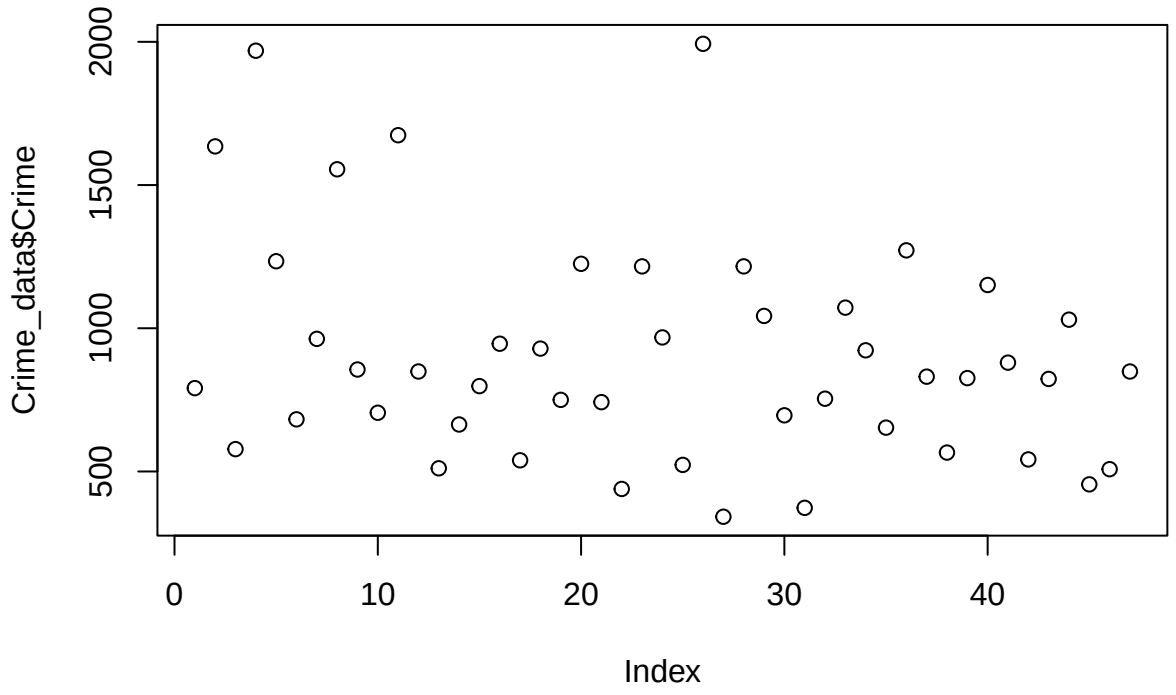
```
Crime_data = read.table(file= "C:\\Users\\sheya\\OneDrive\\Desktop\\uscrime.txt",
                        header = TRUE)
#Checking two rows of data
head(Crime_data, 2)
```

```
##      M So   Ed  Po1 Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq     Prob
## 1 15.1  1  9.1  5.8 5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1 0.084602
## 2 14.3  0 11.3 10.3 9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4 0.029599
##      Time Crime
## 1 26.2011   791
## 2 25.2999  1635
```

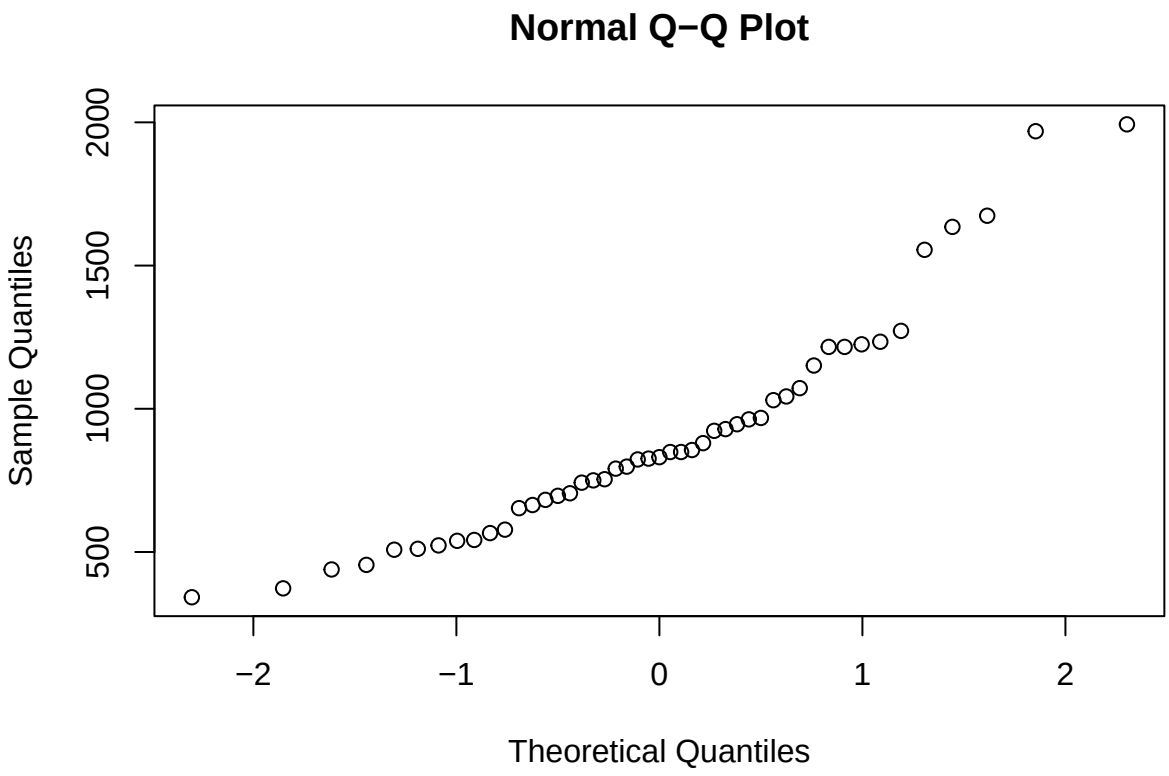
```
#Boxplot of the original data in column 'Crime' that visualizes three possible outliers
boxplot(Crime_data$Crime)
```



```r
#Plot graph of the original data with column 'Crime' that visualizes outliers exists
plot(Crime_data$Crime)
```



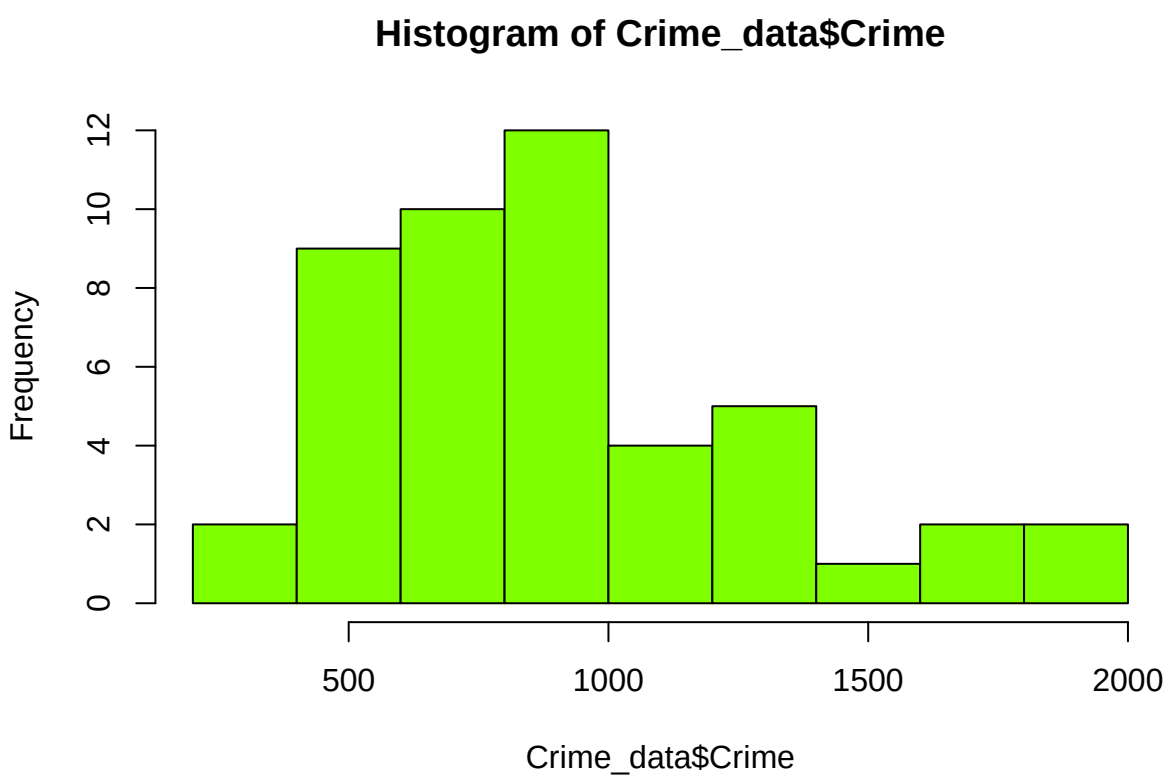

```r
#Checking if the Crimes column in the data is normal with a Q-Q plot graph
qqnorm(Crime_data$Crime)
```




```
#According to the Q-Q plot graph, it is not straight
#Which would indicate non-normal data

#In order to use Grubbs' Test, a dataset should be approximately normally distributed
#Here I use the Shapiro-Wilk Test to check normality

#Checking normality on 100 random values on a 'normal' distribution
set.seed(0)
Crime_data_ss <- rnorm(100)

shapiro.test(Crime_data_ss)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Crime_data_ss
## W = 0.98957, p-value = 0.6303
```

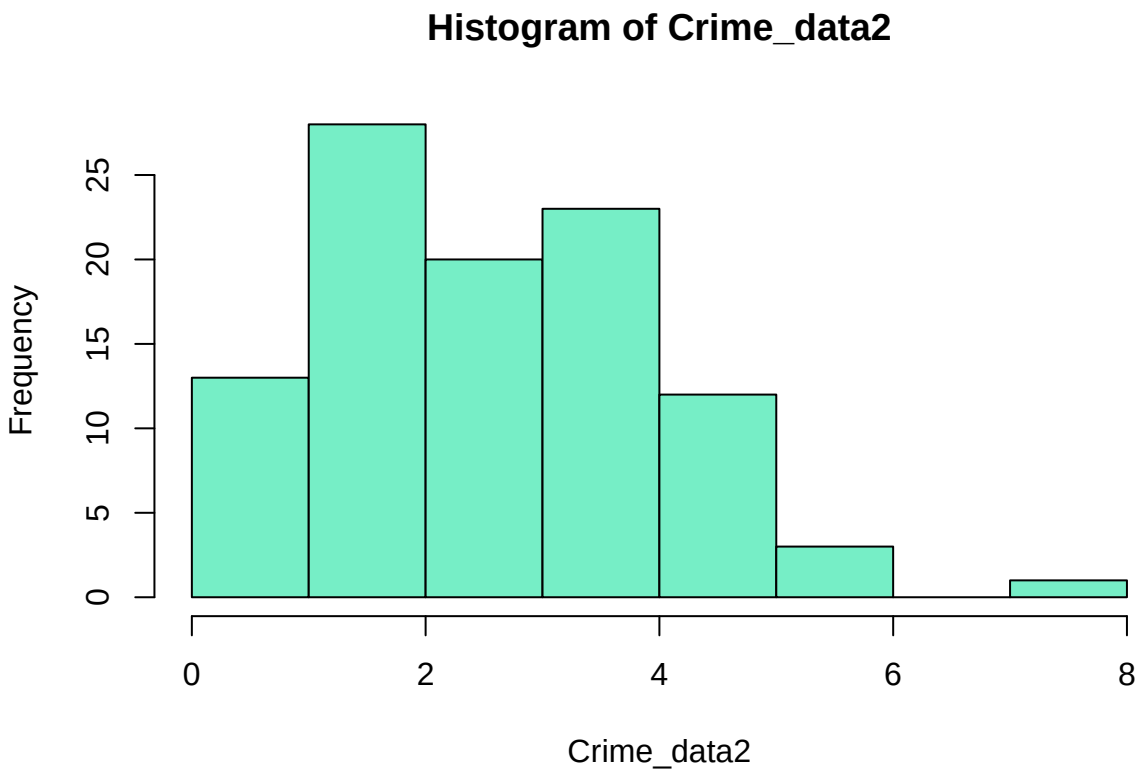
```
hist(Crime_data$Crime, col = 'chartreuse')
```




```
#Based on the shape of the histogram, it is not bell-shaped
#this indicates a non-normal data set
##########

#Checking sample non-normality
set.seed(1)

Crime_data2 <- rpois(n = 100, lambda = 3)

shapiro.test(Crime_data2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Crime_data2
## W = 0.94406, p-value = 0.0003439
```

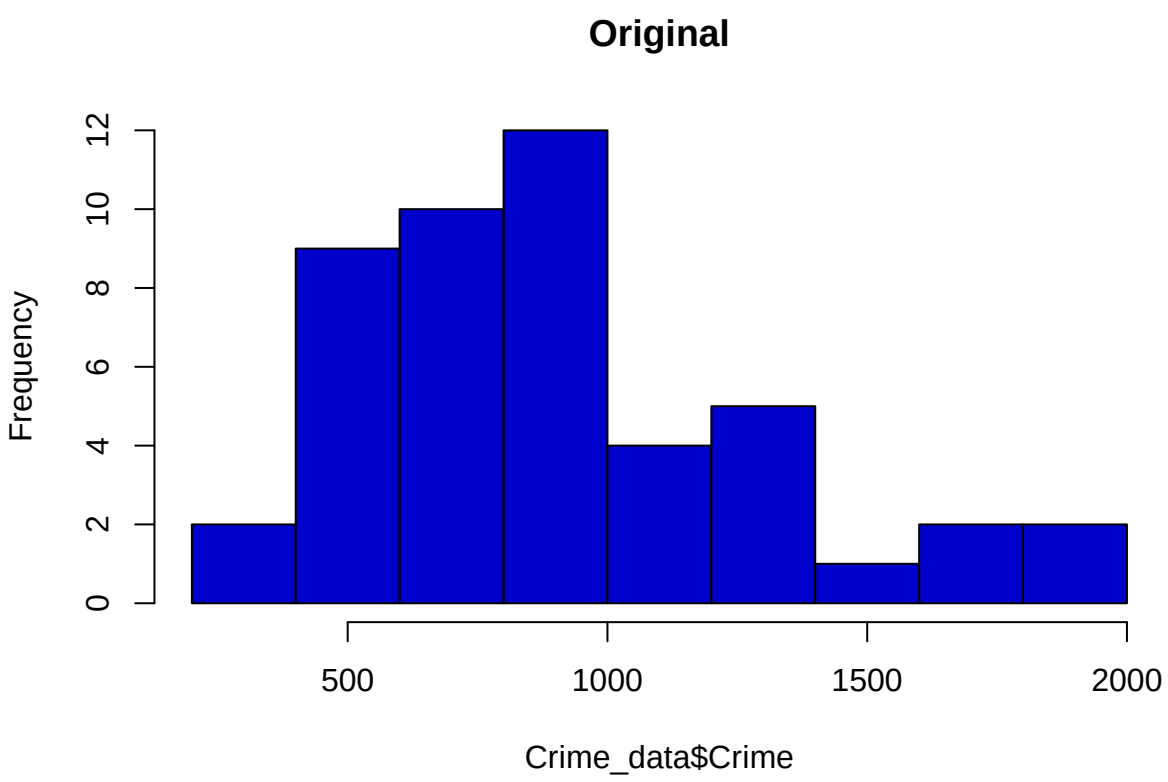
```
#The p-value of the test is 0.0003439. Since this value is less than 0.05,
#there is sufficient evidence to say that the sample data does not come from
#a population that is normally distributed

hist(Crime_data2, col = 'aquamarine2')
```




```
#Based on the histogram, it is not bell shaped.




#Perform Shapiro-Wilk Test on original data
shapiro.test(Crime_data$Crime)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Crime_data$Crime
## W = 0.91273, p-value = 0.001882
```

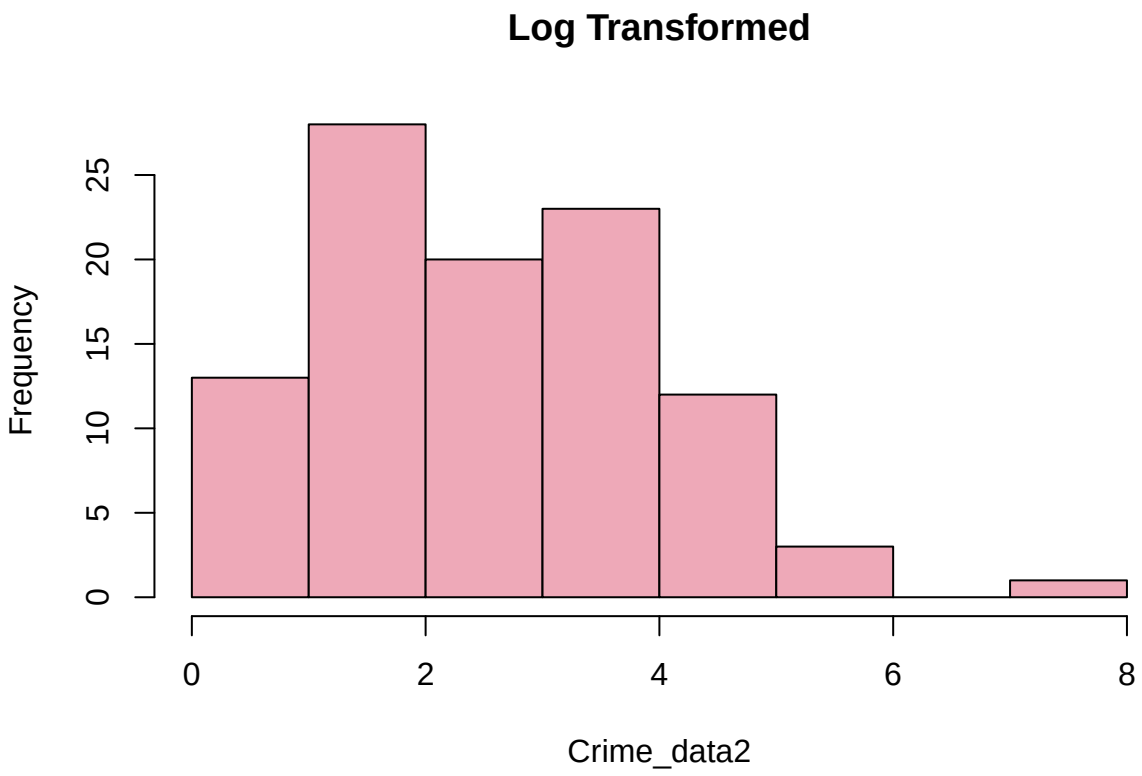
```
hist(Crime_data$Crime, col = 'blue3', main = 'Original')
```




```
#The distribution does not show a bell shape
#Therefore, using other transformations for the non-normal dataset


#Now performing a Shapiro-Wilk test on each distribution


#Perform Shapiro-Wilk Test on log-transformed data
logged_crimes <- log(Crime_data$Crime)

shapiro.test(logged_crimes)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  logged_crimes
## W = 0.98709, p-value = 0.8778
```

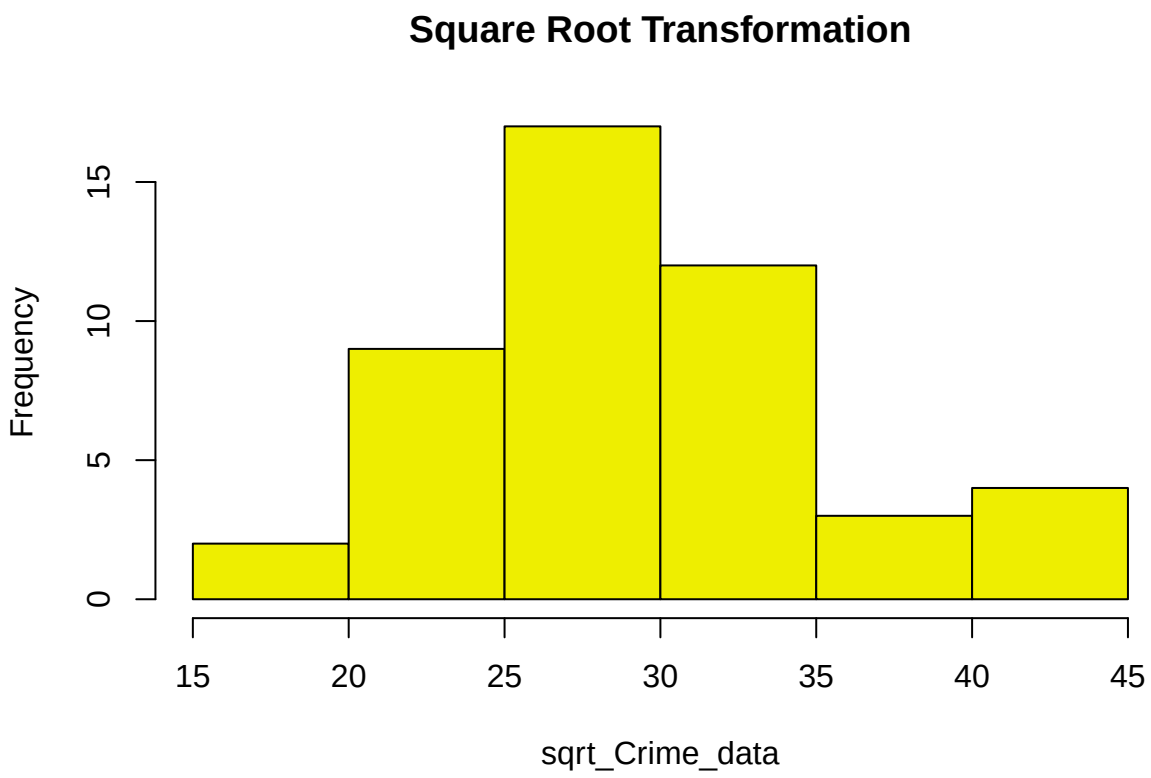
```
hist(Crime_data2, col = 'pink2', main= 'Log Transformed')
```




```
#The Log Transformation looks better but is not fully bell shaped



#Perform The Square Root Transformation
sqrt_Crime_data <- sqrt(Crime_data$Crime)

shapiro.test(sqrt_Crime_data)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sqrt_Crime_data
## W = 0.96448, p-value = 0.1619
```

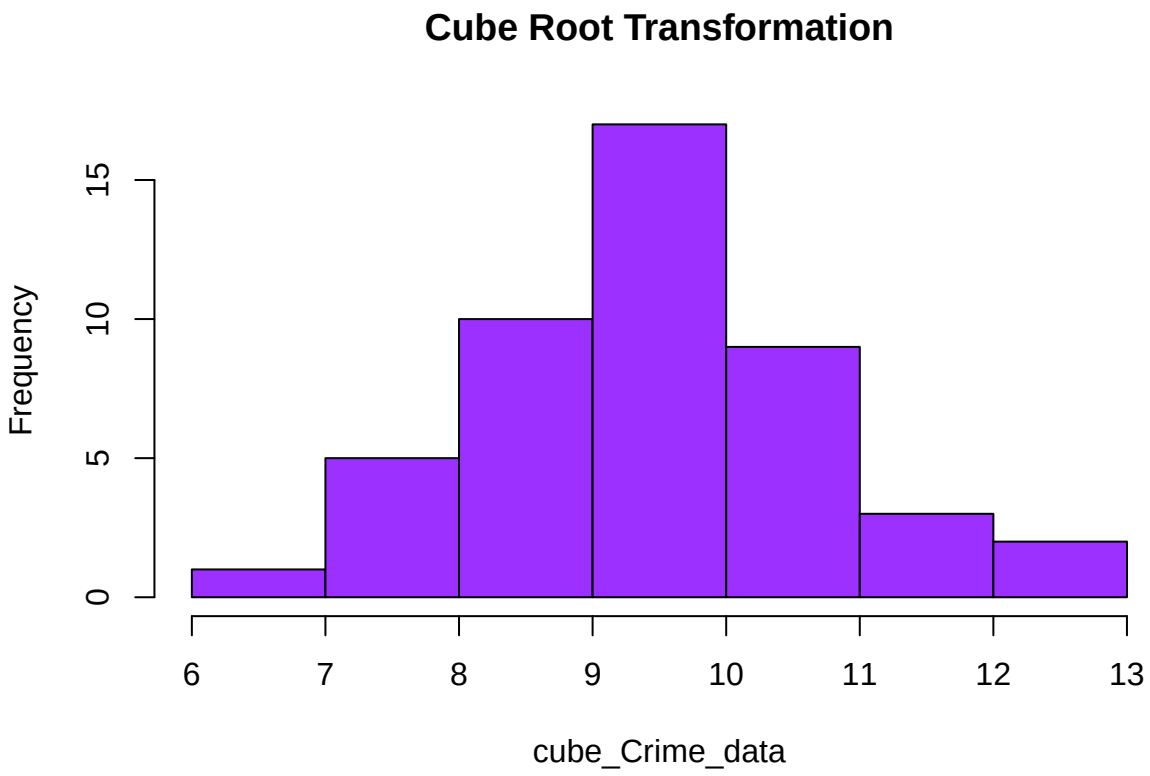
```
hist(sqrt_Crime_data, col='yellow2', main='Square Root Transformation')
```




```
#The Square Root Transformation is much more normally distributed with a bell shape



#Perform The Cube Root Transformation
cube_Crime_data <- Crime_data$Crime^(1/3)

shapiro.test(cube_Crime_data)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cube_Crime_data
## W = 0.97564, p-value = 0.4265
```

```
hist(cube_Crime_data, col= 'purple1', main = 'Cube Root Transformation')
```




```
#The Cube Root Transformation is the best normally distributed and is perfectly bell shaped
#By performing these transformations, the response variable typically becomes closer to
#normally distributed.

#Based on the transformations and histograms to view the distribution
#the transformation reveals that the dataset is now normally distributed
#The dataset has the best transforamtion into a more normally distribution using
#The Cube Root Transformation.


#Calculating the stats of the UScrime dataset with the header Crime
summary(Crime_data$Crime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   342.0   658.5   831.0   905.1  1057.5  1993.0
```

```
#Outliers test
library(outliers)

#Now that the Shapiro Test has been executed, We are able to continue with the Grubbs Test

#Grubbs test to determine whether the max value is an outlier
grubbs.test(Crime_data$Crime, type = 10, opposite = FALSE, two.sided = FALSE)
```

```
##
```

```
##  Grubbs test for one outlier
##
## data:  Crime_data$Crime
## G = 2.81287, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
```

```
#The Grubbs test had a p-value of 0.07887 which is not less than 0.05. Therefore will reject
#the null hypothesis and conclude that the max value of 1993 is not an outlier

#Grubbs test whether the lowest value is an outlier
grubbs.test(Crime_data$Crime, type = 10, opposite = TRUE, two.sided = FALSE)
```

```
##
##  Grubbs test for one outlier
##
## data:  Crime_data$Crime
## G = 1.45589, U = 0.95292, p-value = 1
## alternative hypothesis: lowest value 342 is an outlier
```

```
#The Grubbs test has a p-value of 1 which is not less than 0.05
#Therefore, we will reject the null hypothesis.
#There is not enough evidence to say the minimum value of 342 is an outlier.


#Grubbs Test within a closed set to check for an outlier
Crime_data3 <- Crime_data[-26,16]
grubbs.test(Crime_data3, type = 10, opposite = FALSE, two.sided = FALSE)
```

```
##
##  Grubbs test for one outlier
##
## data:  Crime_data3
## G = 3.06343, U = 0.78682, p-value = 0.02848
## alternative hypothesis: highest value 1969 is an outlier
```

```
#The Grubbs test has a p-value of 0.02848 which is less than 0.05
#Therefore, there is enough evidence to say the value 1969 is an outlier
#From the range in Crime_data3
```

In my demonstrations, I used a plot graph and boxplot to visualize any outliers and in those visuals, possible outliers were discovered. Next, I used a Q-Q plot graph to check normality of the Crimes column. The Q-Q plot graph showed a non-linear dataset which indicated the dataset was non-normal.

Moreover, I needed to test the dataset using the Grubbs' Test although, in order to use the Grubbs' Test, the dataset should be approximately normally distributed. I demonstrated this using Shapiro-Wilk test, transformations, histograms, as well as a smaller range in the dataset. I was able to transform the dataset using log, square root, and cube root transformations to create normality in the data. I also used a smaller dataset (Crimes_data3) to create normality in the data to calculate a more accurate outlier.

Overall, there was enough evidence in my findings to sufficiently reveal an outlier of 1969 with a p-value of 0.02848 in a smaller dataset.

Question 6.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a Change Detection model would be appropriate. Applying the CUSUM technique, how would you choose the critical value and the threshold?

A situation where a Change Detection model would be efficient is outdoor cannabis growers deciding which locations are best to cultivate based on temperature changes. Filtering out extreme temperatures (outliers), cannabis growers can plant in either natural soils or pots that are pre-made. To generate optimum quantities of THC-containing resin, the plant needs a fertile soil and long hours of daylight.

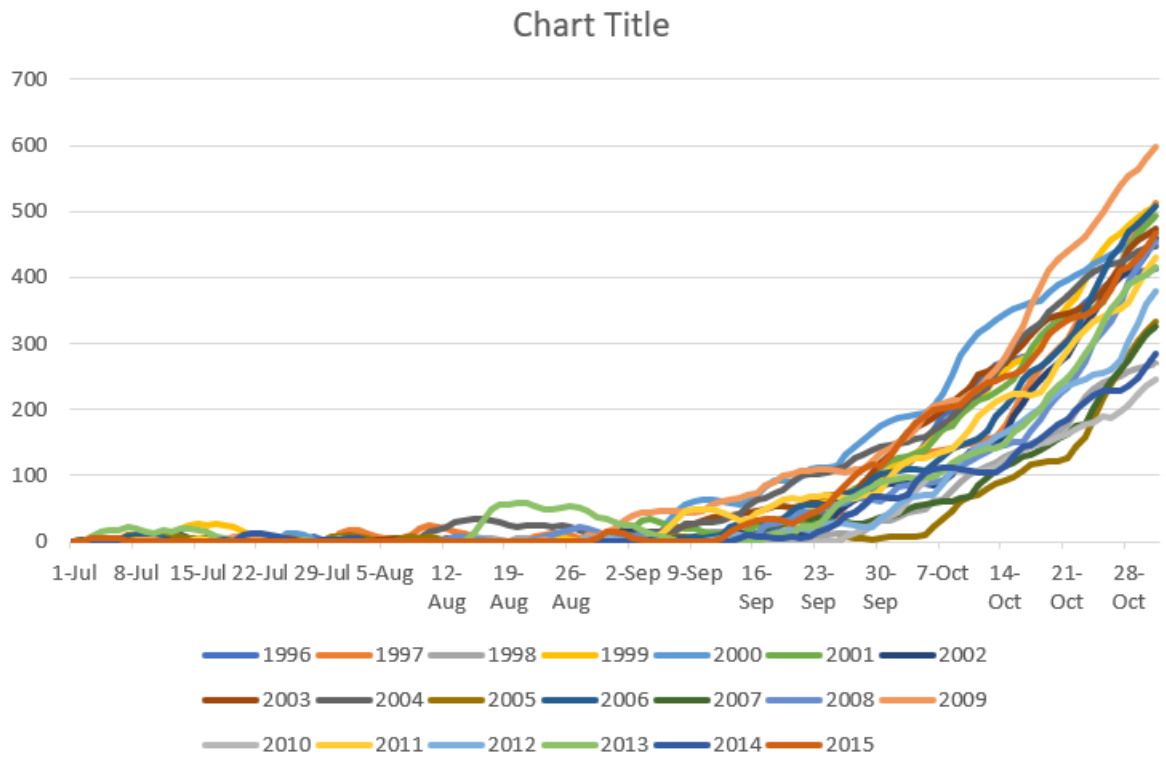
- Threshold- Average temperatures from different regions from prior years that recorded the best THC-containing resin levels at a given time of the month and year.
- Critical value- Setting a large value can bypass a wide range of various temperature readings that can mislead the best areas to cultivate.

Moreover, we are experiencing hotter and longer summers than in recent years which can influence the extreme temperature readings. Although, if they do not, a lower c value can be used.

Question 6.2.1

Using July through October daily-high-temperature data for Atlanta for 1996 through 2015, use a CUSUM approach to identify when unofficial summer ends (i.e., when the weather starts cooling off) each year.

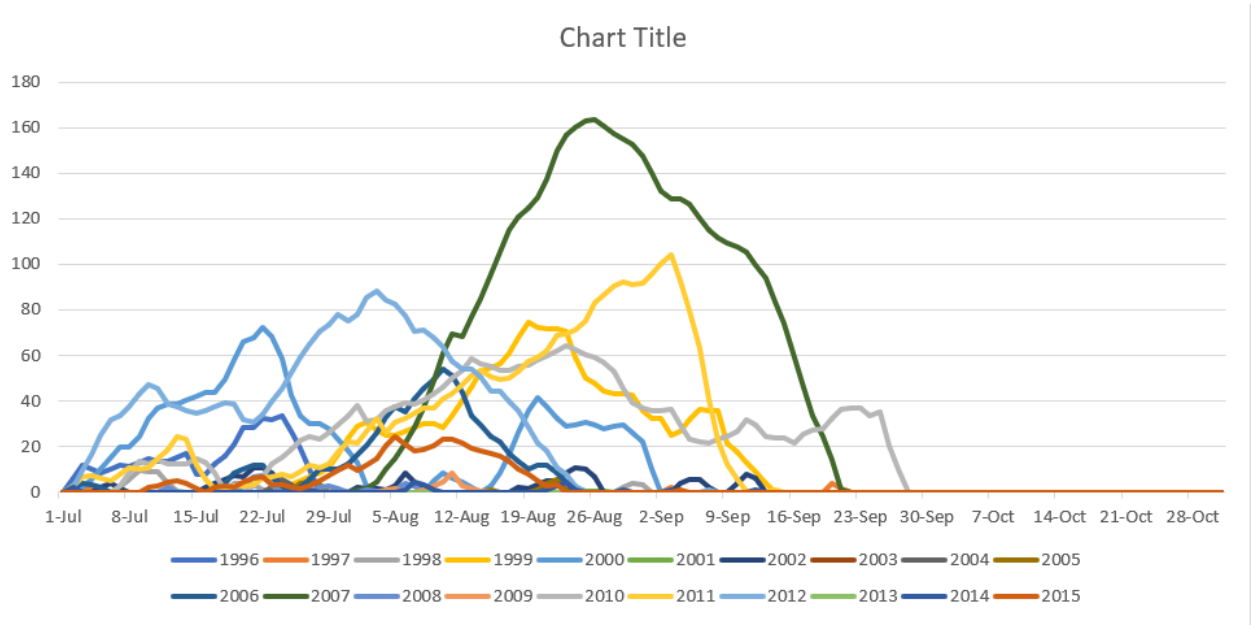
In my findings from the Temps data and using the decreasing formula, S_t = max{0,S_t-1 + (x_t - u - C)}, I was able to create the chart below, (I have attached my Excel worksheet for further analysis). As I set C = 3.5 and T = 35, it appears that there was a change in temperature starting the end of September and beginning of October for every year. It also appears that in 2015, Atlanta has been maintaining high temperatures longer. There has been discovery of potential global warming effecting the core of the earth which causes higher temperatures for a longer period of time. Only time will tell if the cause of global warming continues to increase the temperatures or if there are other factors for these measures.



Question 6.2.2

Use a CUSUM approach to make a judgment of whether Atlanta's summer climate has gotten warmer in that time (and if so, when).

As continued from 6.2.1, I used the increasing formula S_t = max{0,S_t-1 + (u - x_t - C)} to create the chat below that visualizes summers in Atlanta have continued to have high weather temperatures during the months July through August. Although, I did notice there has been a fluxuation in temperatures in 2014 from high temperatures that decreased in the middle of September but then rose back to high temperatures which then led to a significant decrease in the middle of October. This could be due to the climate changes. Ultimately, this effortlessly proves my findings in the previous question of when temperatures were starting to cool down as well as when Atlanta had gotten warmer.



Works Cited

How to Use summary() Function in R

https://www.statology.org/summary-function-in-r/

How to Perform Grubbs' Test in R

https://www.statology.org/grubbs-test-r/

How to Perform a Shapiro-Wilk Test in R

https://www.statology.org/shapiro-wilk-test-r/

How to Transform Data in R (Log, Square Root, Cube Root)

https://www.statology.org/transform-data-in-r/