Homework 3

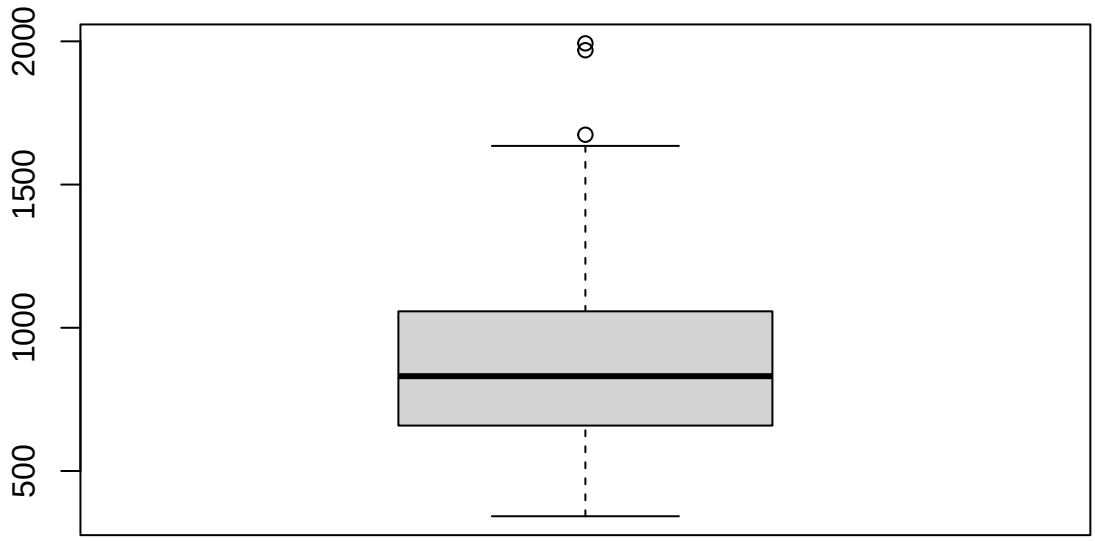
Question 5.1

"Using crime data from the file uscrime.txt (http://www.statsci.org/data/general/uscrime.txt,description at http://www.statsci.org/data/general/uscrime.html), test to see whether there are any outliers in the last column (number of crimes per 100,000 people). Use the grubbs.test function in the outliers package in R."

```
rm(list=ls())
library(outliers)
filepath<-"C:/Users/raque/OneDrive/Desktop/OMSA/ISYE 6501/HW/Homework 3/uscrime.txt"
crime_data<-read.table(filepath, header=TRUE)
summary(crime_data)
```

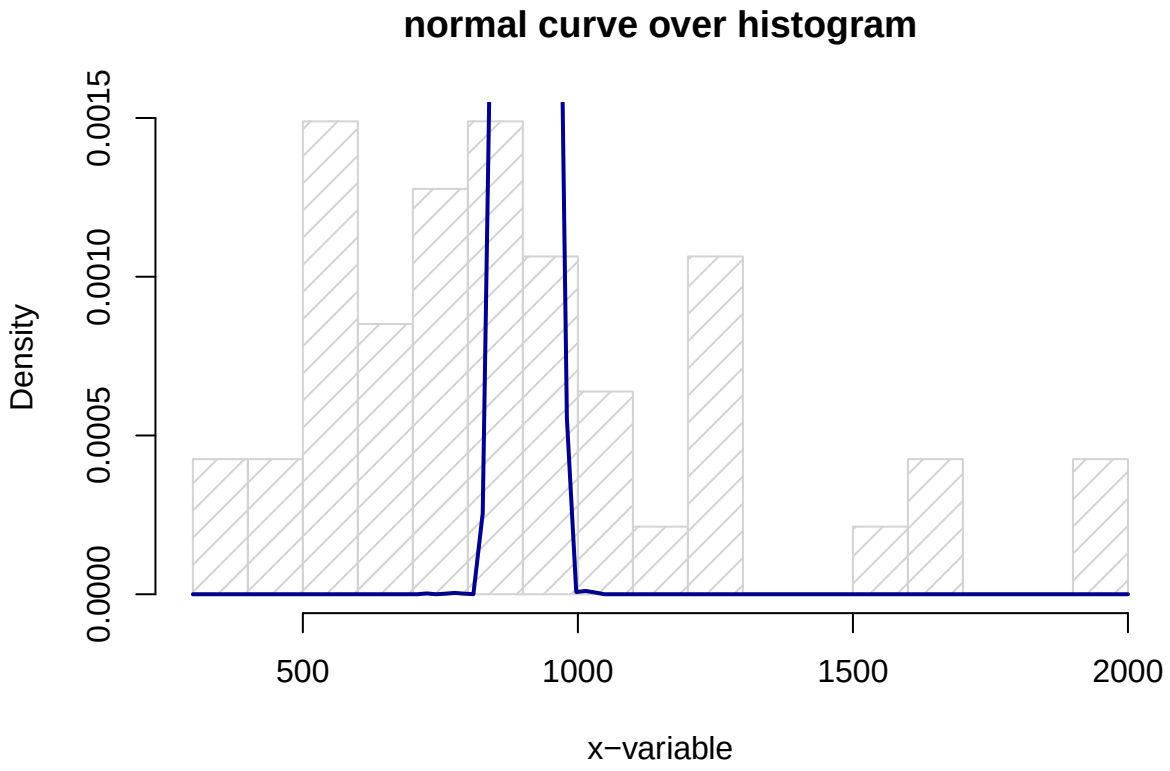
```
##        M               So              Ed             Po1
##  Min.   :11.90   Min.   :0.0000   Min.   : 8.70   Min.   : 4.50
##  1st Qu.:13.00   1st Qu.:0.0000   1st Qu.: 9.75   1st Qu.: 6.25
##  Median :13.60   Median :0.0000   Median :10.80   Median : 7.80
##  Mean   :13.86   Mean   :0.3404   Mean   :10.56   Mean   : 8.50
##  3rd Qu.:14.60   3rd Qu.:1.0000   3rd Qu.:11.45   3rd Qu.:10.45
##  Max.   :17.70   Max.   :1.0000   Max.   :12.20   Max.   :16.60
##       Po2               LF              M.F              Pop
##  Min.   : 4.100   Min.   :0.4800   Min.   : 93.40   Min.   :  3.00
##  1st Qu.: 5.850   1st Qu.:0.5305   1st Qu.: 96.45   1st Qu.: 10.00
##  Median : 7.300   Median :0.5600   Median : 97.70   Median : 25.00
##  Mean   : 8.023   Mean   :0.5612   Mean   : 98.30   Mean   : 36.62
##  3rd Qu.: 9.700   3rd Qu.:0.5930   3rd Qu.: 99.20   3rd Qu.: 41.50
##  Max.   :15.700   Max.   :0.6410   Max.   :107.10   Max.   :168.00
##        NW               U1               U2            Wealth
##  Min.   : 0.20   Min.   :0.07000   Min.   :2.000   Min.   :2880
##  1st Qu.: 2.40   1st Qu.:0.08050   1st Qu.:2.750   1st Qu.:4595
##  Median : 7.60   Median :0.09200   Median :3.400   Median :5370
##  Mean   :10.11   Mean   :0.09547   Mean   :3.398   Mean   :5254
##  3rd Qu.:13.25   3rd Qu.:0.10400   3rd Qu.:3.850   3rd Qu.:5915
##  Max.   :42.30   Max.   :0.14200   Max.   :5.800   Max.   :6890
##       Ineq             Prob             Time           Crime
##  Min.   :12.60   Min.   :0.00690   Min.   :12.20   Min.   : 342.0
##  1st Qu.:16.55   1st Qu.:0.03270   1st Qu.:21.60   1st Qu.: 658.5
##  Median :17.60   Median :0.04210   Median :25.80   Median : 831.0
##  Mean   :19.40   Mean   :0.04709   Mean   :26.60   Mean   : 905.1
##  3rd Qu.:22.75   3rd Qu.:0.05445   3rd Qu.:30.45   3rd Qu.:1057.5
##  Max.   :27.60   Max.   :0.11980   Max.   :44.00   Max.   :1993.0
```

```
set.seed(123)
boxplot(crime_data[,16])
```



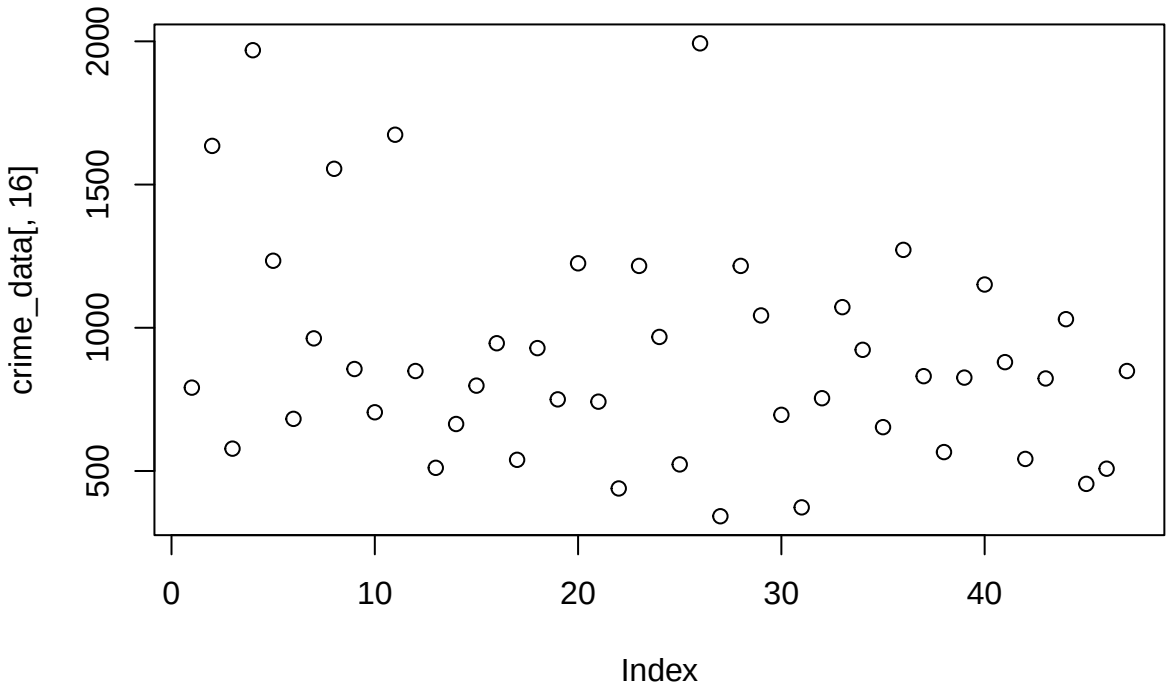
From my initial visual analysis, of column 16 (crime) from "crime_data", I noticed that the box plot only showed three possible outliers all above "1500". The maximum outlier being around the value "2000". I then assessed the grubbs test model based on the information I learned from the box plot.

```
g=(crime_data[,16])
m<-mean(g)
std<-sqrt(g)
hist(crime_data[,16],density=10, breaks=12,prob=TRUE,
      xlab="x-variable",
      main="normal curve over histogram")
curve(dnorm(x,mean=m, sd=std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```



I also noticed from the histogram, that there was a slight right skew, where the tail of the distribution curve was longer on the right side. This meant that the outliers of the distribution curve were further out towards the right and closer to the mean on the left.

```
plot(crime_data[,16])
```




```
grubbs.test(crime_data[,16],type=11,opposite=FALSE, two.sided=FALSE)
```

```
##
##  Grubbs test for two opposite outliers
##
## data:  crime_data[, 16]
## G = 4.26877, U = 0.78103, p-value = 1
## alternative hypothesis: 342 and 1993 are outliers
```

The grubbs.test function in R showed me the outliers in the crime data file. I set the x value to only show me the last column "number of crimes per 100,000 people", "type=11" and to check if both the minimum and maximum values are outliers. Then I set "opposite=FALSE" since I was not checking the opposite value, and also set"two.sided=FALSE" since I was not treating this as two sided. The results showed that the possible outliers were 342 and 1993. The test statistic was "G=4.26877" and the corresponding p-value was p=1. Since the value of p was greater than 0.05, we failed to reject the null hypothesis that there were no outliers in the dataset. Therefore, I still did not have sufficient evidence to say that the minimum value of "342" and that the maximum value of "1993" are true outliers.

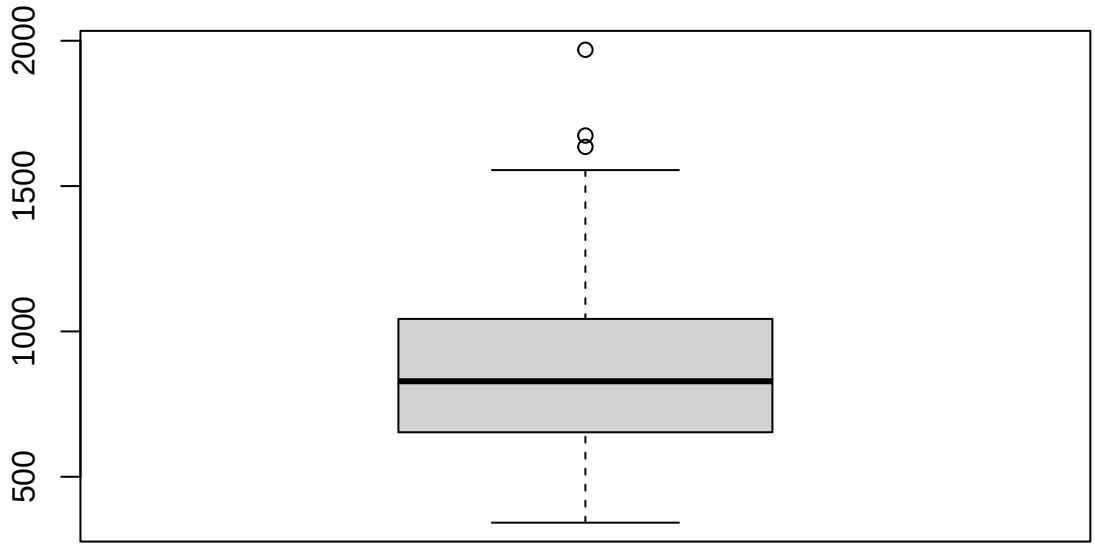
```
grubbs.test(crime_data[,16], type=10, opposite=FALSE, two.sided=FALSE)
```

```
##
##  Grubbs test for one outlier
##
## data:  crime_data[, 16]
```

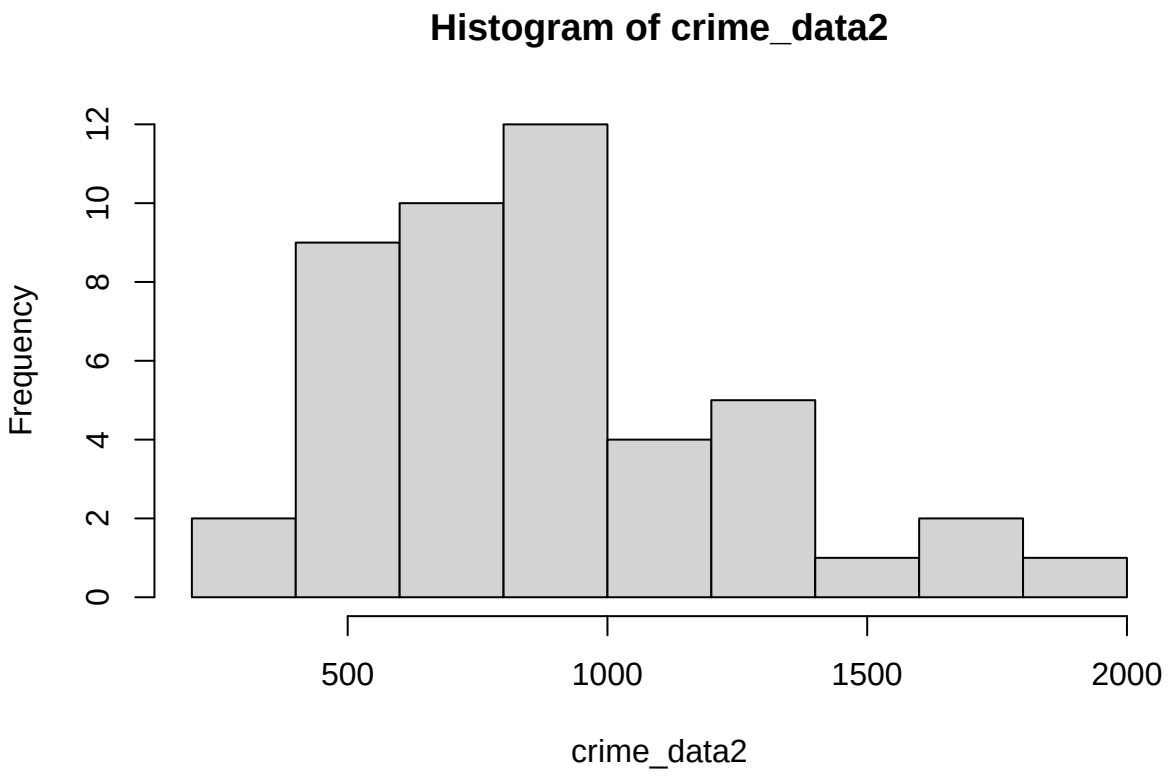
```
## G = 2.81287, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
```

From this grubbs test I noticed that the p-value was still above 0.05 so we still failed to reject the null hypotheis that there was no outlier in the dataset. Therefore, I still did not have sufficient evidence to say that the maximum value of "1993" was an outlier.Next, I decided to remove the "1993" point from the crime dataset and then tested the new dataset ("crime_data2) below using the grubbs.test function to see if I get an outlier with a p-value of 0.05 or below.

```
crime_data2<-crime_data[-26,16]
boxplot(crime_data2)
```




```
hist(crime_data2)
```




```
grubbs.test(crime_data2, type=10)
```

```
##
##  Grubbs test for one outlier
##
## data:  crime_data2
## G = 3.06343, U = 0.78682, p-value = 0.02848
## alternative hypothesis: highest value 1969 is an outlier
```

Final Result:

From my initial visual analysis of the boxplot on the "crime_data2" dataset I observed that I still had 3 possible outliers, with the maximum outlier being close to the y-value of "2000". I also noticed that the p-value was now below 0.05. So, since we now rejected the null hypothesis, this showed me that the outlier was "1969" and all other prior evidence pointed to that as well.

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?"

One of the major current issues that our world faces is global warming. Therefore, I believe a change detection model is appropriate in detecting changes in the climate that are atypical. An example of this, is in the Arctic, where changes in temperatures have risen three times the global annual average. Since the Arctic is already very sensitive to any increase in temperatures, we can use the Change Detection model in tracking any increases in change over time. Since the cost of the temperature increasing in the Arctic is extremely costly, we can set the values of C and T to detect the changes faster, even if it is closer to falsely detecting the change earlier in the Change Detection model. This Change Detection model also involves trading off the costs of early false detection for reducing the risk of greater damage to the Arctic.

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. You can get the data that you need from the file temps.txt or online, for example at http://www.iweathernet.com/atlanta-weather-records or https://www.wunderground.com/history/airport/KFTY/2015/7/1/CustomHistory.html. You can use R if you'd like, but it's straightforward enough that an Excel spreadsheet can easily do the job too."

First, I began by doing an initial visual analysis of the Atlanta summer data using a line graph. I predicted that the period of no change was around the first twenty x values so I decided to use that when calculating my standard deviation in order to ensure my CUSUM model did not not detect changes late. From there I used Excel to create a CUSUM decrease detection formula using the average mean value (μ) of the period of no change, then subtracted the observed values from the certain time and the C buffer, 3. From there, I chose a threshold of 120, in order to to detect when the change of my results crossed the threshold of 120. I chose this threshold since the dataset had great variety throughout the years. Also, the changes were visible in the line graph around that month every summer. Overall, I identified using the CUSUM Decrease Detection Model that the unofficial summer in Atlanta begins to end around September 30th, with some years varying greatly (summer ending much sooner or later).

"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)."

Here I used a similar approach to the previous question, but with the CUSUM increase detection formula. From there I used Excel to create a formula from the observed values from a certain time, then subtracted the average mean value (μ) of the period of no change as well as the C buffer, 0. From there, I chose a threshold of 12, which indicated when the results were above the threshold. I had chosen this threshold value since the dataset includes summer temperatures and it did not take very long for the changes to show in the line graph and the CUSUM decrease dataset. Therefore, by using the CUSUM Increase Detection Model, I identified that the unofficial summer in Atlanta begins around July 15th, with some years beginning much sooner than others.

References

All analyses were performed using R Statistical Software (R version 4.3.2 (2023-10-31 ucrt)).

Ehrlich, I. (1973) Participation in illegitimate activities: a theoretical and empirical investigation. Journal of Political Economy 81, 521–565.

Vandaele, W. (1978) Participation in illegitimate activities: Ehrlich revisited. In Deterrence and Incapacitation, eds A. Blumstein, J. Cohen and D. Nagin, National Academy of Sciences, Washington DC, pp. 270–335.

Venables, W., and Ripley, B. (1998). Modern Applied Statistics with S-Plus, Second Edition. Springer-Verlag.