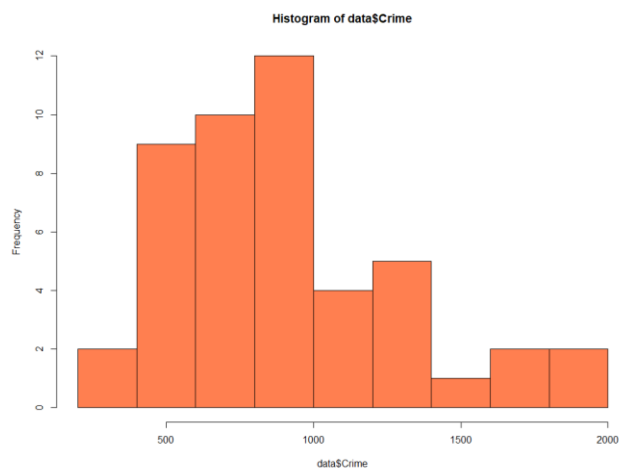
ISYE6501 Homework 3

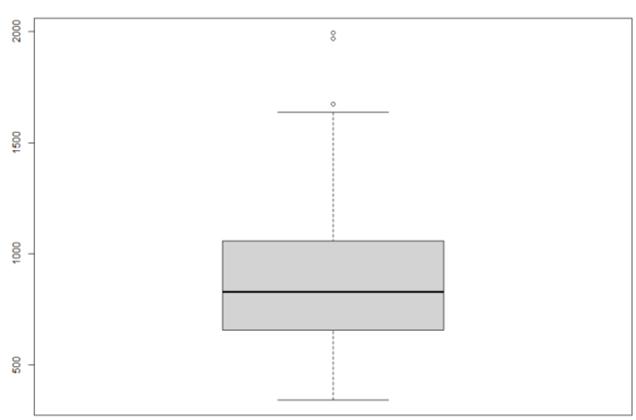
5.1)

The table studies effect of punishment regimes on crime rates for 47states based on aggregate data of 15 predictors. To analyze if we have an outlier in our crime rate, the following approach was used:

a) The crime rate datapoints were plotted in an histogram to check if the data had a normal distribution. The histogram shows that our dataset is left skewed with a long tail in the right. This visually indicates that our data probably has a higher end or an upper limit that could create outliers.



b) A box plot was then plotted to identify potential outliers visually. Based on the plot, we seem to have 3 datapoints that are outside our upper limits and therefore potential outliers. However, since the data represents crime data of a 47 different states. It is possible that a few states have much higher crime rate than the other states and therefore a direct comparison with other states in terms of distribution may not be accurate. So I continued to analyze the data one more time with a different method



c) Using the *grubbs.test* for one outlier check on the upper limit, we noticed a p-value of 0.07887. Since the p-value is not <0.05, we fail to reject the null hypothesis and do not have sufficient evidence to say that the maximum value of 1993 is an outlier. A similar check using *grubbs.test* was performed to check with the lowest value is an outlier. With a p-value of 1 we again did not have sufficient evidence to sat that the minimum value of 342 is an outlier.

   Based on the *grubbs.test* I concluded that the minimum datapoint and the 3 potential outliers from boxplot are within the expected standard deviation and will not be accounted as outliers.

d) Based on the available data information, there should not be any accounted outliers in this dataset. The crime rate documented in each state is not a variable based on other states crime rate. To exclude one point as an outlier could potentially be harmful and lead us to neglect on states with higher or slower crime rates. So logically, even though the grubbs test is close to 0.05, I lean towards not an outlier

```
~/
>
> rm(list=ls())
> set.seed(100)
>
> #Call Required Library
> library(outliers)
>
> #Sort last column into required data
> data=data=read.table("M:/OMSA/ISYE6501/HW3/uscrime.txt",stringsAsFactor=FALSE,header=TRUE)
> head(data)
     M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq     Prob    Time Crime
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 26.2011   791
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 25.2999  1635
3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0 0.083401 24.3006   578
4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7 0.015801 29.9012  1969
5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4 0.041399 21.2998  1234
6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6 0.034201 20.9995   682
> crime_data=data[,16]
>
> #Check if data is normally distributed using histogram
> hist(data$Crime,col="coral")
>
> #Check for potential outliers using a boxplot
> boxplot(crime_data)
>
> #use grubbs test for minimum and mixximum values in the dataset
> grubbs.test(crime_data)

        Grubbs test for one outlier

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

> grubbs.test(crime_data,opposite=TRUE)

        Grubbs test for one outlier

data:  crime_data
G = 1.45589, U = 0.95292, p-value = 1
alternative hypothesis: lowest value 342 is an outlier
```

6.1]

In semiconductor manufacturing, the etch time on a wafer is a critical process step that can either cause electrical shorts or holes on the wafer causing multiple chips on a wafer to either have low yield or lead to complete scrap (Waste). In order to reduce this costly error, a change detection model can be used to monitor process health. The following parameters will be critical:

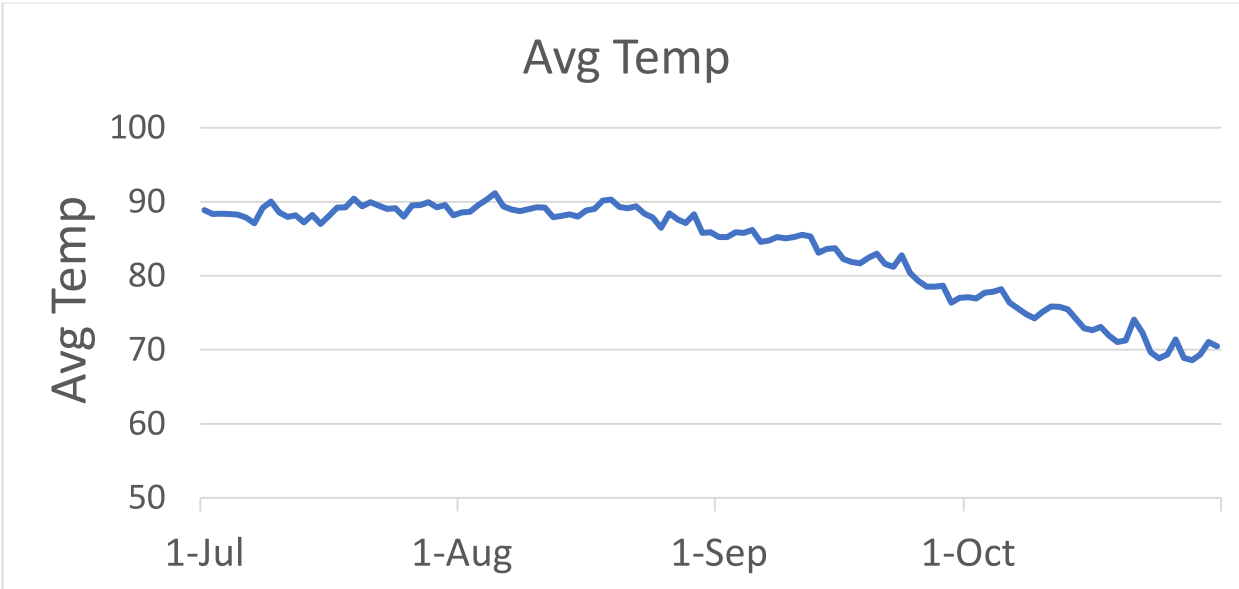
1) Monitor tool etch rate/etch time based on end point detection or time based etching methods
2) Setup product specific monitoring charts instead of tool specific since every product will have a different etch time
3) Using the CUSUM technique, I would choose a critical value based on standard deviation from historically valid dataset
4) The threshold for time based and end point detection based etching would be dependent on layer thickness at a particular processing step, for example, a layer of 100 angstroms thickness might only need 40-60angstroms to be etched, so the threshold could be set at 65 or 70 angstroms depending on process criticality

6.2]

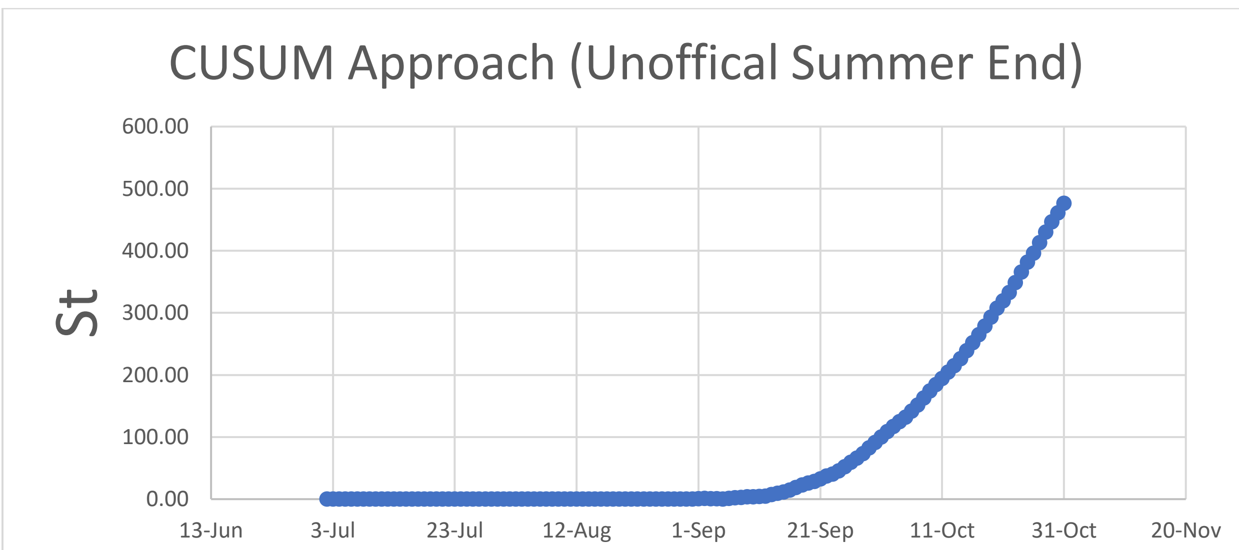
1)

In order to find the average temperature during our control period, the average temp on each day every year was calculated. The graph (1) below shows that the temperature seems to be fairly linear and consistent across July, 1st to August, 23rd. I took the mean during this period as a Mu value of 87.39.

The standard deviation for the dataset was calculated to be 6.7 which was used as the baseline for setting the values for C and T. Since T is the threshold after which the model would identify a change of pattern, I used a value of 2 times the standard deviation since 14 degrees temperature change is a big difference. The value of C which helps add an error margin to avoid false identification of pattern change can make the model too sensitive. Since temperature variance is normal due to temporary climatic changes , however not long lasting, I choose a value of C to be 2.345 (25% of the Std Deviation). The St values were plotted against time as per in graph (2). This indicated that the unofficial summer ended during mid September



Graph: (1)



Graph: (2)

6.2]

2)

I did a yearly average of atlanta's summer days between 1996 to 2015. The Mu was calculated as the average of all the yearly averages and set to 83.33. The standard deviation calculated for the dataset was 1.58. Based on similar assumptions and reasoning, the C was set at 25% of the Std Dev(0.39) and T was 2 times the std dev (3.17).

Based on graph (3) Below, Atlanta's summer days have gotten considerably warmer since the year 2010.

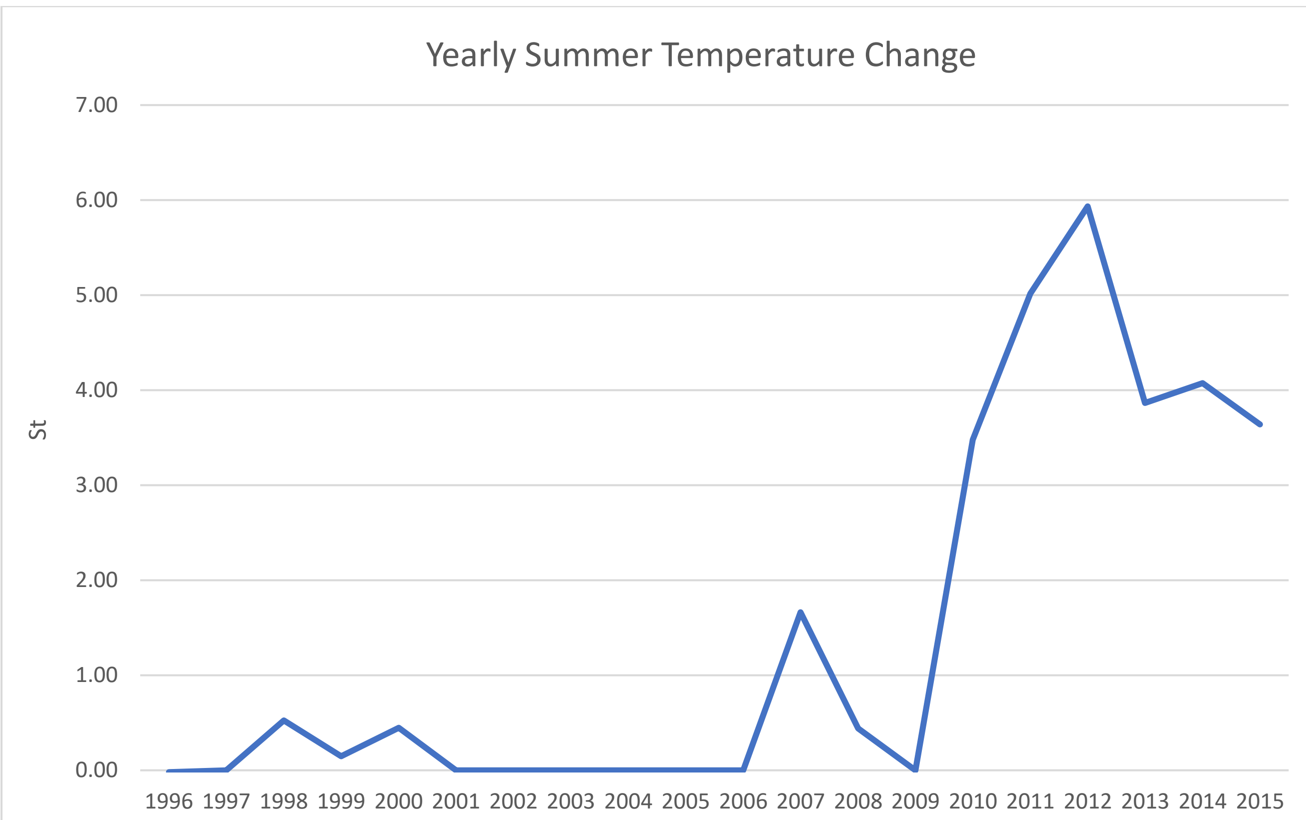


Figure: (3)