Outliers

ISYE6501 – Week 3 Assignment

Purpose

The purpose of this assignment is to analyze outliers in the given data sets. An outlier is a value that's very different than the other observations. This can be due to extreme variability (possible but unlikely response) or simply measurement or other errors. So, dealing with outliers has both statistical and philosophical context.

Q 3.1 – Crime Data

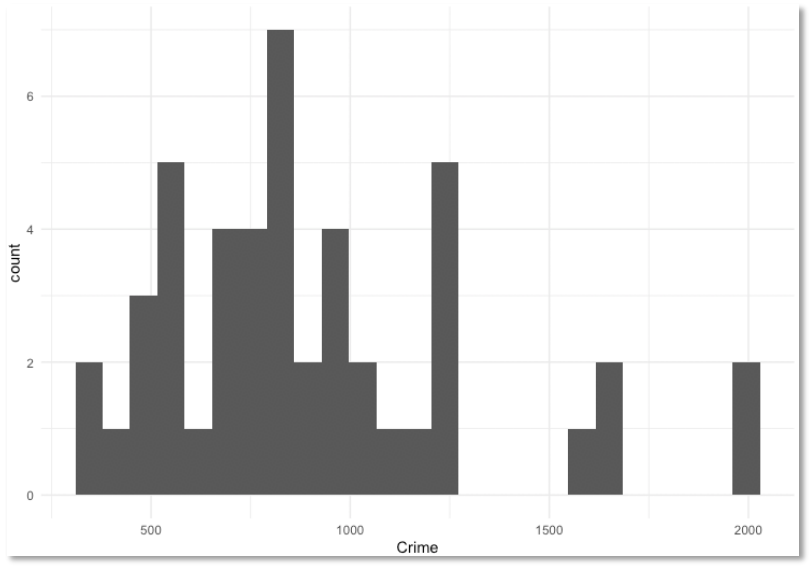
Using crime data from the file `uscrime.txt`, test to see whether there are any outliers in the last column (number of crimes per 100,000 people). Use the `grubbs.test` function in the `outliers` package in R.

The first step after reading the data in would be to do some descriptive statistics by using basic R functions such as summary(), str(). This will give us the range of predictors as well general idea of the spread of the data. One useful option is to plot the histogram & boxplot of values as visual inspections often catch our attention. Below, Iam examining one attribute (Crime), but this can be done for any / all attributes.

```
> summary(data)
       M               So              Ed              Po1
Po2               LF              M.F
 Min.   :11.90   Min.   :0.0000   Min.   : 8.70   Min.   : 4.50
Min.   : 4.100   Min.   :0.4800   Min.   : 93.40
 1st Qu.:13.00   1st Qu.:0.0000   1st Qu.: 9.75   1st Qu.: 6.25
1st Qu.: 5.850   1st Qu.:0.5305   1st Qu.: 96.45
 Median :13.60   Median :0.0000   Median :10.80   Median : 7.80
Median : 7.300   Median :0.5600   Median : 97.70
 Mean   :13.86   Mean   :0.3404   Mean   :10.56   Mean   : 8.50
Mean   : 8.023   Mean   :0.5612   Mean   : 98.30
 3rd Qu.:14.60   3rd Qu.:1.0000   3rd Qu.:11.45   3rd Qu.:10.45
3rd Qu.: 9.700   3rd Qu.:0.5930   3rd Qu.: 99.20
 Max.   :17.70   Max.   :1.0000   Max.   :12.20   Max.   :16.60
Max.   :15.700   Max.   :0.6410   Max.   :107.10
      Pop              NW              U1               U2
Wealth          Ineq            Prob
 Min.   :  3.00   Min.   : 0.20   Min.   :0.07000   Min.   :2.000
Min.   :2880   Min.   :12.60   Min.   :0.00690
 1st Qu.: 10.00   1st Qu.: 2.40   1st Qu.:0.08050   1st Qu.:2.750
1st Qu.:4595   1st Qu.:16.55   1st Qu.:0.03270
 Median : 25.00   Median : 7.60   Median :0.09200   Median :3.400
Median :5370   Median :17.60   Median :0.04210
 Mean   : 36.62   Mean   :10.11   Mean   :0.09547   Mean   :3.398
Mean   :5254   Mean   :19.40   Mean   :0.04709
 3rd Qu.: 41.50   3rd Qu.:13.25   3rd Qu.:0.10400   3rd Qu.:3.850
3rd Qu.:5915   3rd Qu.:22.75   3rd Qu.:0.05445
```

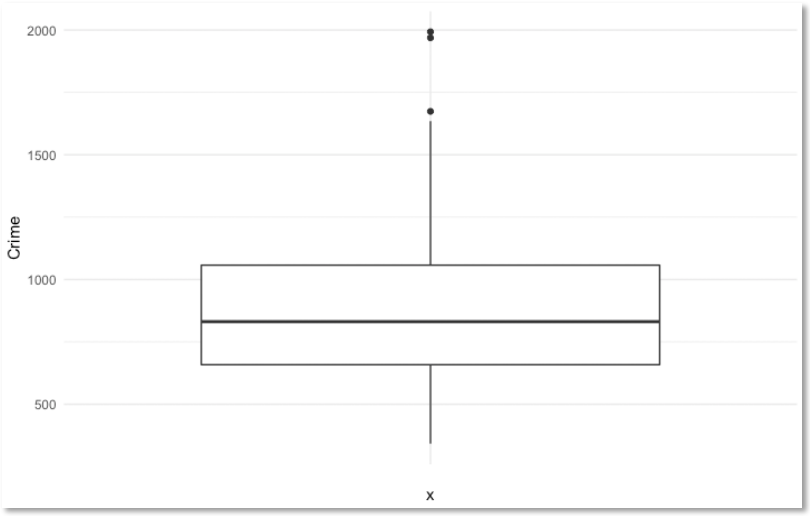
```
 Max.   :168.00   Max.   :42.30   Max.   :0.14200   Max.   :5.800
Max.   :6890   Max.   :27.60   Max.   :0.11980
      Time           Crime
 Min.   :12.20   Min.   : 342.0
 1st Qu.:21.60   1st Qu.: 658.5
 Median :25.80   Median : 831.0
 Mean   :26.60   Mean   : 905.1
 3rd Qu.:30.45   3rd Qu.:1057.5
 Max.   :44.00   Max.   :1993.0
```

```
ggplot(data) +
  aes(x = Crime) +
  geom_histogram(bins = 25L) +
  theme_minimal()
```



The boxplot also confirms this.

```
ggplot(data) +
  aes(x = "", y = Crime) +
  geom_boxplot() +
  theme_minimal()
```



From the above plot, we know there are couple of outliers in far-right part of the plot.

Boxplot Stats

Using boxplot.stats() lets examine if there are outliers. This provides a convenient summary since Grubbs test detects on a single value.

```
out <- boxplot.stats(data$Crime)$out
> out
[1] 1969 1674 1993
```

Grubbs test

Grubs test enables us to detect if outlier is found in either side of the range – minimum or maximum. This is a good measure for point outliers (but not necessarily contextual or collective outliers)

Commented [RN1]: Read about p-value and t=20?
Remove outlier and rerun
When do you stop ?

```
> test <- grubbs.test(data$Crime)
> test

      Grubbs test for one outlier

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

```
> grubbs.test(data$Crime, opposite = T)

        Grubbs test for one outlier

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

> grubbs.test(data$Crime, type=11)

        Grubbs test for two opposite outliers

data:  data$Crime
G = 4.26877, U = 0.78103, p-value = 1
alternative hypothesis: 342 and 1993 are outliers
```

As you can see below, in addition to the possible outliers, Grubb's test also prints the p-value, which is the probability of rejecting the null hypothesis when it's true. Smaller the p-value is, stronger the evidence is against the null hypothesis, i.e there is more evidence to go with alternate hypothesis.

Null Hypothesis (in one-sided) test: smallest (one-tailed lower) or largest (one-tailed upper) value in the sample is an NOT outlier. Conversely, the alternative hypothesis is that this value is an outlier.

Lets see if there are other outliers by successively removing outlier

```
> which(crime_data==1993)
[1] 26
> crime_data <- crime_data[-26]
> grubbs.test(crime_data, opposite = F)

        Grubbs test for one outlier

data:  crime_data
G = 3.06343, U = 0.78682, p-value = 0.02848
alternative hypothesis: highest value 1969 is an outlier

> which(crime_data==1969)
[1] 4
> crime_data <- crime_data[-4]
> grubbs.test(crime_data, opposite = F)

        Grubbs test for one outlier

data:  crime_data
G = 2.56457, U = 0.84712, p-value = 0.1781
alternative hypothesis: highest value 1674 is an outlier

> which(crime_data==1674)
[1] 10
```

```
> crime_data <- crime_data[-10]
> grubbs.test(crime_data, opposite = F)

      Grubbs test for one outlier

data:  crime_data
G = 2.68561, U = 0.82837, p-value = 0.1139
alternative hypothesis: highest value 1635 is an outlier
```

For one-sided test for largest value, though p-value (0.07) is more than the threshold of 0.05, its still to closer than rejecting the null hypothesis complete. In subsequent tests, we find 1969 with low enough p-value of 0.02. **I conclude that 1993 and 1969 are outliers in the upper tail**.

On the lower end, since p-value is 1, we fail to reject the null hypothesis. Hence, I summarize that that **there is no sufficient evidence to conclude that an outlier exists in lower tail**.

Outliers and other attributes

Commented [RN2]: List +ve and -ve corrrlations seperately

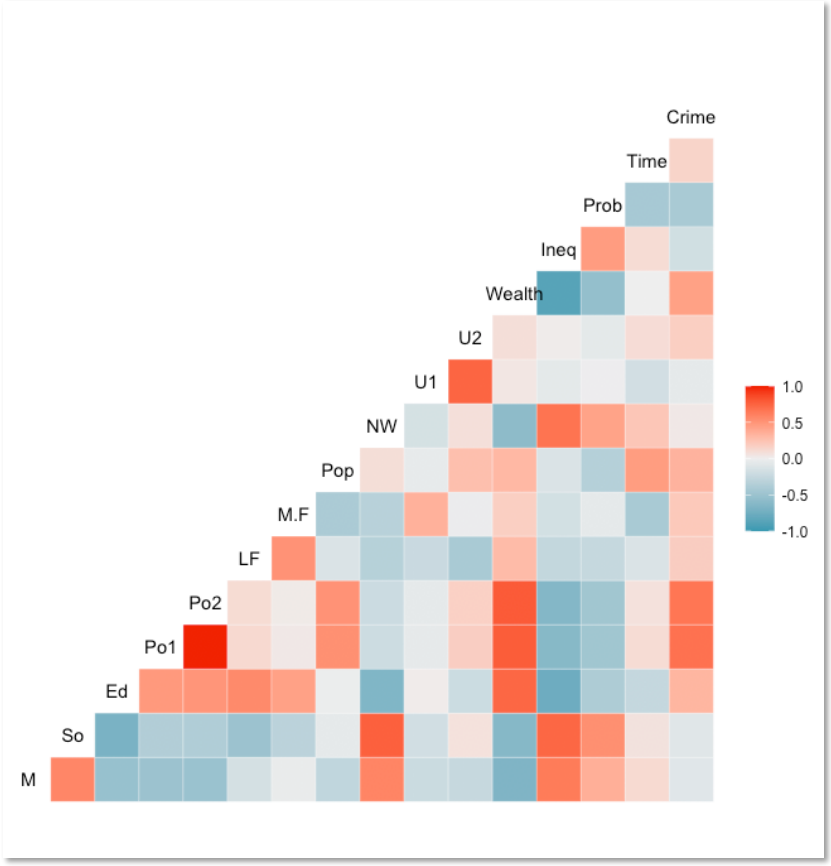
Looking at other attributes for entries with Crime outlier measurement

```
out_ind <- which(data$Crime %in% c(out))

> out_ind
[1]  4 11 26

> data[out_ind,]
      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq     Prob    Time Crime
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
11 12.4  0 10.5 12.1 11.6 0.580  96.6 101 10.6 0.077 3.5   6570 17.0 0.016201 41.6000  1674
26 13.1  0 12.1 16.0 14.3 0.631 107.1   3  7.7 0.102 4.1   6740 15.2 0.041698 22.1005  1993
```

- None of these are southern states
- Education level (ED) is relatively high, above the median (10.0)
- Per capita police expenditure is quite high. This is an useful metrics to further look into since this can change policy and budget decisions.
- Except for one outlier, the population density (Pop) is quite high in these states. Another useful metrics look into as this can lead to city planning and resource allocation.
- Percentage of non whites (NW) is low in these states, but not too low. This observation could be further analysed for changes in policies on rational profiling, etc.
- Unemployment rate (U1 & U2) is quite high. This could be useful to further analyze for resource allocation for youth training, job assistance, etc.
- Surprisingly, these states are not poor states. They are above median wealth states.
- Income inequality is also not too high – all at or below the mean.
- Avg. prison time is quite relatively high on these states. We can infer that these states spend more on policing and prison collectively. As mentioned before, the data can be further analyzed to see to aid policy decisions.

General correlation

```
ggcorr(data)
```



Looking at the correlation of all the attributes, in general, the following attributes have more influence on the Crime:

- Po1, Po2: Police expenditure.
- Population
- Adult unemployment (U2)
- Ironically, Wealth and Education level as well. But, this may be a colinear with Population attribute.

Q 6.1 – Outlier's example

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?

I work in the eventing / messaging space where applications use a central messaging system for mediating and exchanging events. Consumers of these events use "queues" (a persistent storage abstract) to consume these events on their pace and let the messages pile up on the queues when processing is slow or when consuming applications are down for maintenance or other reasons.
More info on this technology here: https://en.wikipedia.org/wiki/Message_broker

One critical measure of impending issue is when messages pile up on the queue for too long. This is an indication that the consuming application has an issue or offline. Not all such queue build ups are however bad or need to be alerted on – since its possible that the publishers of these events had an unexpected load and filled up the Queue quicky. As long as we have enough storage to hold these messages until they are processed, its ok. On the other hand, there could be some time sensitive event (eg: alert an airplane of sudden tarmac closure) and queuing these messages beyond acceptable limit can be costly, risky, and subject to legal consequences. Also depending on the event exchange patterns that change over time (eg: summer season sales vs winter sales), the thresholds keep changing over the time.

This problem lends itself to not using a "one threshold fit all" setting or even one threshold per application / queue setting. Since the thresholds are based on the current consumption capabilities of the consumers, a CUSUM algorithm, which can detect shifts in the mean measure, can be used effectively to tune threshold for each queue as well as adopt to changing event traffic scenarios.

Q 6.2 Atlanta Temperature data

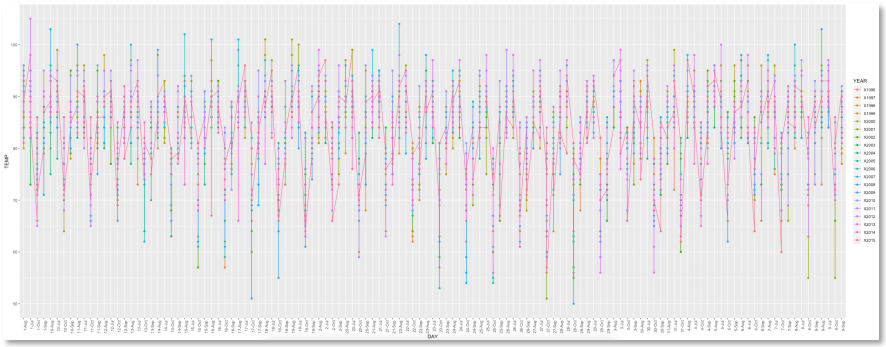
Q 6.2.1 Unofficial summer end

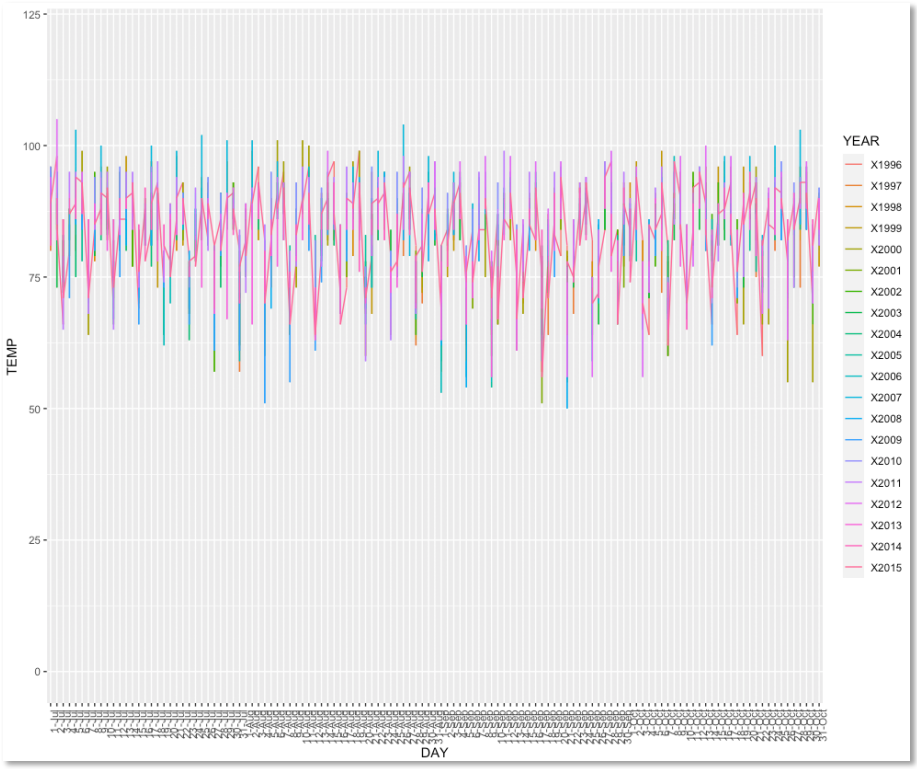
1. Using July through October daily-high-temperature data for Atlanta for 1996 through 2015, use a CUSUM approach to identify when unofficial summer ends (i.e., when the weather starts cooling off) each year.

First, lets see how the temperatures are distributed along the year across all years. Its useful to melt the data to narrow form and use ggplot grouping to plot across years.

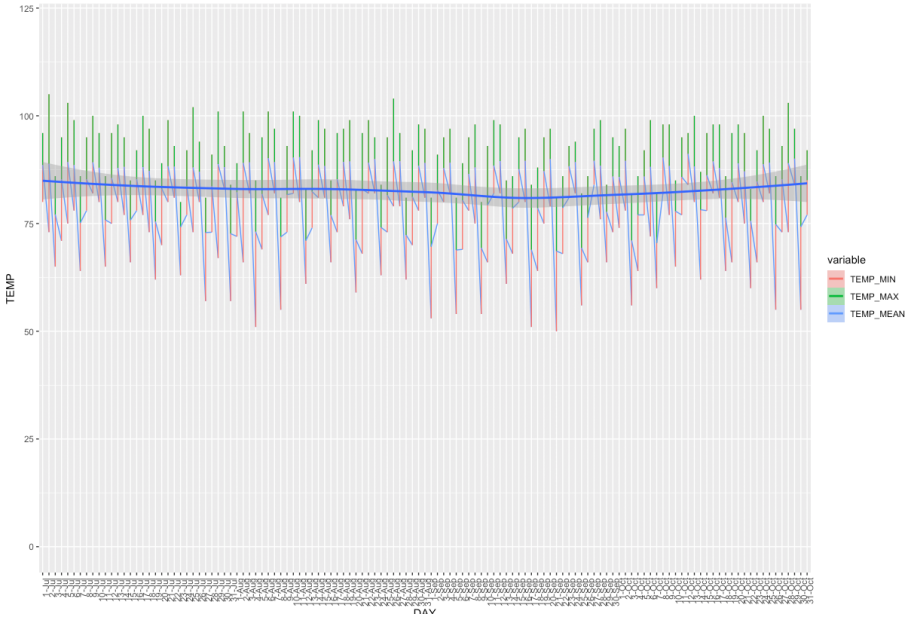
```
temps <-   read.table('temps.txt',
                          stringsAsFactors = FALSE,
                          header = TRUE)
```

```
temps_m <- melt(temps, id.vars = "DAY", variable.name = 'YEAR',
value.name = 'TEMP')
temps_m$YEAR <- as.character(as.character(temps_m$YEAR))

temps_m %>% ggplot(aes(x=DAY, y=TEMP, color=YEAR, group = 1)) +
  geom_point() +
  geom_line() +
  theme(axis.text.x=element_text(angle=90, hjust=1))
```





Plotting Min, Max and Mean across month



These plots do show a slight downward trend in temperature distribution over the months, which is expected as we move from summer to the winter.

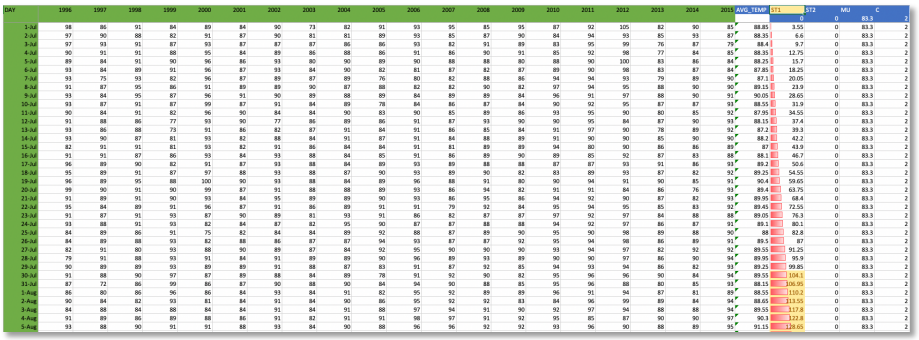
CUSUM with EXCEL

Lets examine the change and if there exists a point where it turns colder using cusum. For this, I created an average of all years per data point for analysis. I generated two additional columns, one for detecting increase & another for detecting decrease.

Macros

Macro for Increase	W3=MAX(0, W2+(V3-Y3-Z3))	W2: St-1 V3: Xt Y3: Mu Z3:C
Macro for decrease	X3=MAX(0, X2+(Y3-V3-Z3))	X2: St-1 V3: Xt Y3: Mu Z3:C

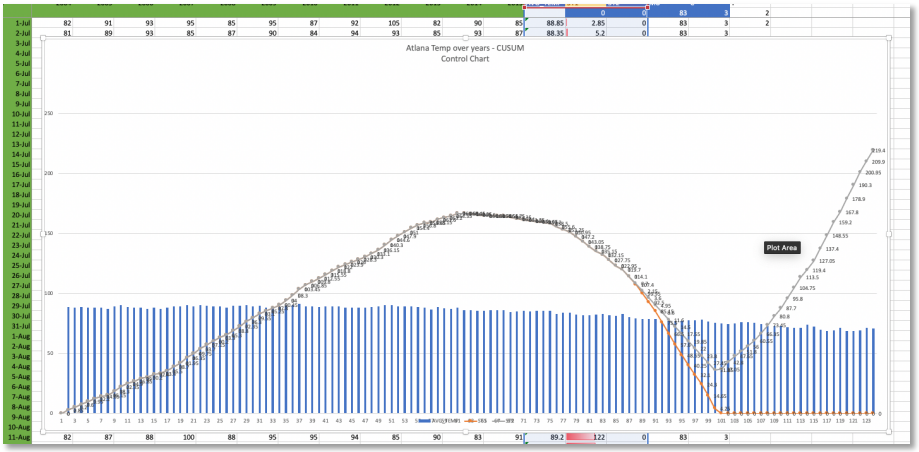
With Static conditional formatting, I can view rows where the threshold is hit.



Here are observations for various Mu, C and T values.

Mu	C	T	Threshold Reach
80	1	100	Jul-14
83	1	100	Jul-22
83	2	100	Jul-28
83	4	100	Aug-22
83	3	150	Aug-21

I also created a clustered column series chart with corresponding Xi, S1 (increase) and S2 (decrease) change detection to see the effects visually.

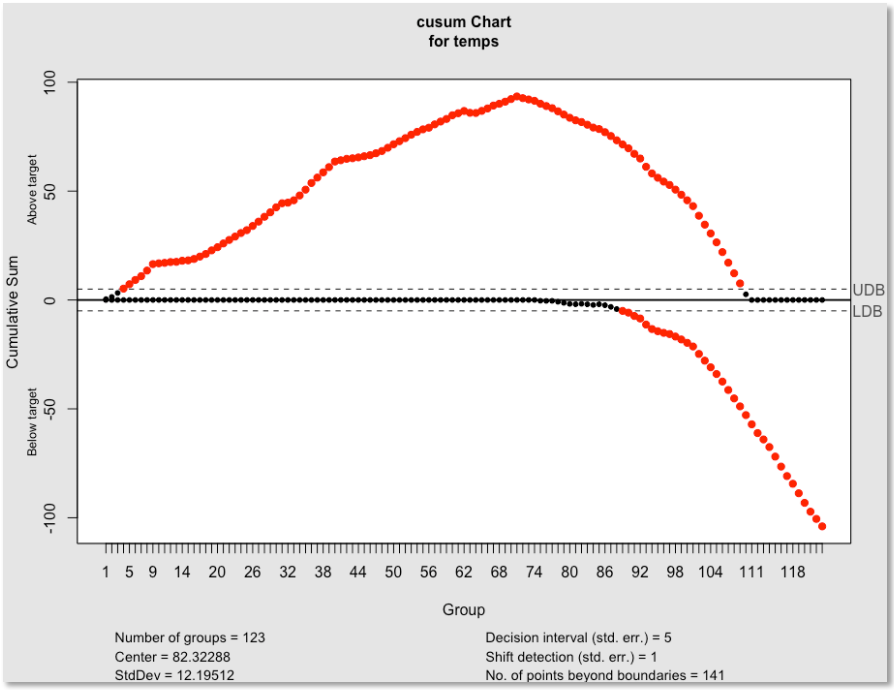


Based on these, we can conclude that:

- Increasing C makes it less sensitive to the changes and takes longer to detect.
- Larger values of T also make it less sensitive to detecting the changes.

CUSUM in R

```
c <- cusum(temps)
plot(c)
```



We observe a change in direction at about row 90, marking temperatures cooling which corresponds to Sep 28th. As you can see this is slightly behind the actual winter start since this model is not tuned with right T and C values and not very sensitive to changes immediately.

Q 6.2.2 Effect of global warming

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

To examine any change in temperature distribution over the year, lets first look at min/max/mean temperate over the years.

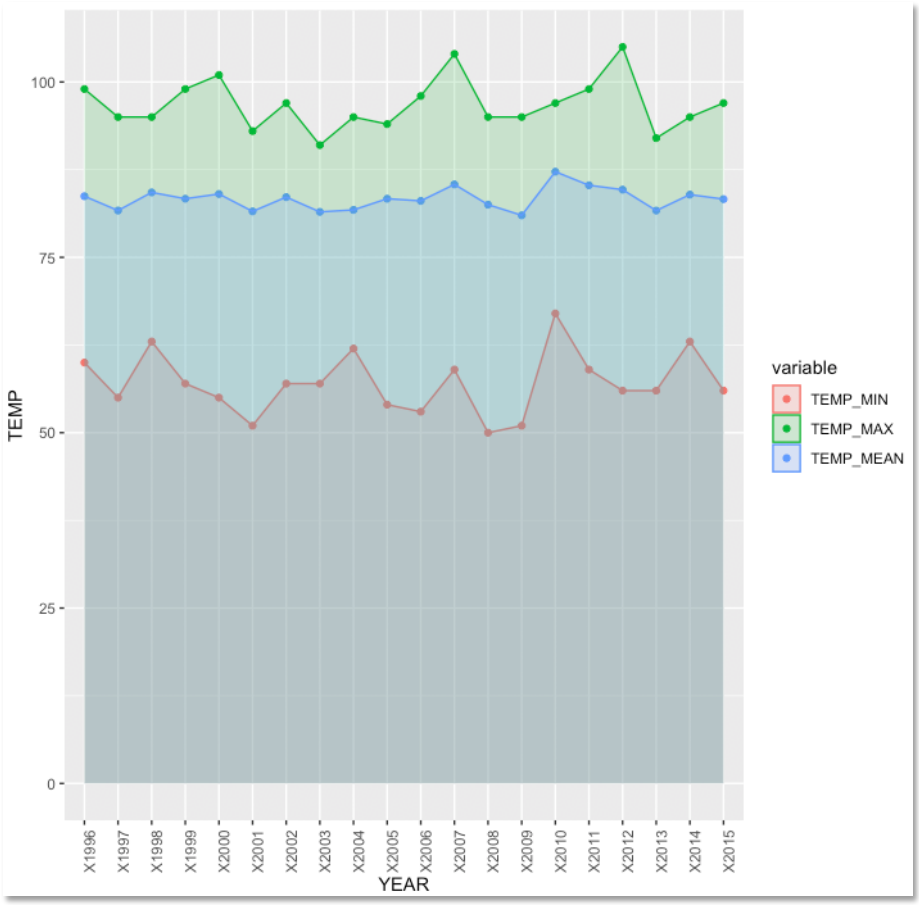
```
# avg temp over the  years
temp_mean <- aggregate(TEMP~YEAR, data=temps_m, mean);
colnames(temp_mean)[2] <- 'TEMP_MEAN'
temp_max <- aggregate(TEMP~YEAR, data=temps_m, max) ;
colnames(temp_max)[2] <- 'TEMP_MAX'
temp_min <- aggregate(TEMP~YEAR, data=temps_m, min) ;
colnames(temp_min)[2] <- 'TEMP_MIN'

t <- merge(temp_min, temp_max, by='YEAR')
temp_stats <- merge(t, temp_mean, by='YEAR')
```

The following lists the values sorted by mean temperature.

```
> temp_stats[order(temp_stats$TEMP_MEAN),]
    YEAR TEMP_MIN TEMP_MAX TEMP_MEAN
14 X2009       51       95  80.99187
8  X2003       57       91  81.47967
6  X2001       51       93  81.55285
18 X2013       56       92  81.66667
2  X1997       55       95  81.67480
9  X2004       62       95  81.76423
13 X2008       50       95  82.51220
11 X2006       53       98  83.04878
20 X2015       56       97  83.30081
4  X1999       57       99  83.35772
10 X2005       54       94  83.35772
7  X2002       57       97  83.58537
1  X1996       60       99  83.71545
19 X2014       63       95  83.94309
5  X2000       55      101  84.03252
3  X1998       63       95  84.26016
17 X2012       56      105  84.65041
16 X2011       59       99  85.27642
12 X2007       59      104  85.39837
15 X2010       67       97  87.21138
```

We can also plot these to get a visual summary.

```
temp_stats_m <- melt(temp_stats, id.vars = "YEAR", value.name =
'TEMP')

temp_means <- aggregate(TEMP~YEAR, data=temps_m, mean)
temp_means %>%
  ggplot(aes(x=YEAR, y=TEMP, group = 1)) +
  geom_point() +
  geom_area(aes(alpha=0.5 , size=.5,color='red')) +
  theme(axis.text.x=element_text(angle=90, hjust=1))
```



This data shows the overall temperature shift over the year:

- Max temperature had been on the rise peaking at 2012
- Average temp had been on the rise as well peaking at 2009
- 2009 – 2012 is hotter overall – with low, mean, and max temps above the prev. years.

Reference

- https://www.statlearning.com
- https://statsandr.com/blog/outliers-detection-in-r/
- https://rpubs.com/ssufian/658130
- https://www.extendoffice.com/documents/excel/1534-excel-cumulative-chart.html
- https://support.minitab.com/en-us/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/outlier-test/interpret-the-results/all-statistics-and-graphs/