Week 3 Assignment - Outliers and CUSUM Change Detection

Omer Farooq (EDx ID: mfarooq4)

1/25/2020

Table of Contents

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.

Loaded needed libraries.

```
library(outliers)
library(dplyr)
library(ggpubr)
library(grDevices)
library(climtrends)
```

Next, loaded the US Crime data set and separated out the last column i.e. Crime per 100,000 people.

```
set.seed(101)

#loading data
my_data <- read.delim("data_5.1/uscrime.txt")
head(my_data)

##      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
## 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
## 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
## 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
## 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
##      Time Crime
## 1 26.2011   791
## 2 25.2999  1635
## 3 24.3006   578
## 4 29.9012  1969
## 5 21.2998  1234
## 6 20.9995   682
```

```
#basic stats of the crime per 100k data
crime_data <- my_data$Crime
summary(crime_data)
```

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

```
sd(crime_data)
```

```
## [1] 386.7627
```

The basic statistics of the Crime data showed that data has a long tail on the maximum value side given the max value is pretty distant from the 3rd quartile compared to how distant the min value is from the 1st quartile.

Next, I applied different Grubbs tests to the Crime data to check outliers in the data. It is important to keep in mind that Grubbs test tests one outlier (or two in some cases) on either extremes of the data. I applied the following test options:

- Grubbs Test type 10 - tests one outlier. Opposite value TRUE or FALSE allows to check maximum and minimum value as outlier.
- Grubbs Test type 11 - tests two outliers on both extremes. This test will only identify outliers if both data points are outliers. If one of the two is not, it will fail to reject the NULL hypothesis which is “Both Values are not outliers”.
- Grubbs Test type 20 - tests the most extreme and the next extreme value as outliers. Opposite parameter allows to test both ends.

Results of all these 6 tests were logged in a matrix and printed below.

```
#empty matrix to log results of grubbs test
test_results <- matrix(NA, nrow=6, ncol=4)
colnames(test_results) <- c("Test Method","Alternative Hypothesis","P-
Value","Verdict")



#different grubbs tests to check outliers on both extremes
test1 <- grubbs.test(crime_data, type=10, opposite = FALSE, two.sided =
FALSE)
test_results[1,] <- c(test1$method, test1$alternative,round(test1$p.value,
digit=3),"Potentially Outlier")

test2 <- grubbs.test(crime_data, type=10, opposite = TRUE, two.sided = FALSE)
test_results[2,] <- c(test2$method,
test2$alternative,round(test2$p.value,digit=3),"Not Outlier")

test3 <- grubbs.test(crime_data, type=11, opposite = FALSE, two.sided =
FALSE)
```

```r
test_results[3,] <- c(test3$method,
test3$alternative,round(test3$p.value,digit=3),"Not Outlier")

test4 <- grubbs.test(crime_data, type=11, opposite = TRUE, two.sided = FALSE)
test_results[4,] <- c(test4$method,
test4$alternative,round(test4$p.value,digit=3), "Not Outlier")

#type 20 of grubbs test works only on 3-30 sample size. Selecting top 30 and
then bottom 30 data points to use in this test.
sorted_data <- my_data[order(my_data$Crime),]
topn30_data<-top_n(sorted_data,30)
topn30_data<-topn30_data$Crime

test5 <- grubbs.test(topn30_data, type=20, opposite = FALSE, two.sided =
FALSE)
test_results[5,] <- c(test5$method,
test5$alternative,round(test5$p.value,digit=3), "Outlier")

bottomn30_data<-top_n(sorted_data,-30)
bottomn30_data<-bottomn30_data$Crime

test6 <- grubbs.test(bottomn30_data, type=20, opposite = FALSE, two.sided =
FALSE)
test_results[6,] <- c(test6$method,
test6$alternative,round(test6$p.value,digit=3), "Not Outlier")

test_results

##      Test Method
## [1,] "Grubbs test for one outlier"
## [2,] "Grubbs test for one outlier"
## [3,] "Grubbs test for two opposite outliers"
## [4,] "Grubbs test for two opposite outliers"
## [5,] "Grubbs test for two outliers"
## [6,] "Grubbs test for two outliers"
##      Alternative Hypothesis                  P-Value Verdict
## [1,] "highest value 1993 is an outlier"        "0.079" "Potentially Outlier"
## [2,] "lowest value 342 is an outlier"          "1"     "Not Outlier"
## [3,] "342 and 1993 are outliers"               "1"     "Not Outlier"
## [4,] "342 and 1993 are outliers"               "1"     "Not Outlier"
## [5,] "highest values 1969 , 1993 are outliers" "0.008" "Outlier"
## [6,] "lowest values 342 , 373 are outliers"    "0.402" "Not Outlier"
```

Commentary on the Grubbs Test Results

- Row #1 indicated that highest value 1993 is possibly an outlier. P-Value needed to be less than 0.05 for NULL Hypothesis to be rejected and it's pretty close to it.
- Row #2 indicated that lowest value 342 is not an outlier given P-Value is high.
- Rows #3 and 4 were not relevant because test type 11 checked for both values to be outliers at the same time and as rows 1 and 2 showed, only one of the two extreme

values is an outlier, these tests were bound to fail and they did (as shownby high P-Values).

- Row 5 and 6 used test type 20 with Top 30 and Bottom 30 rows of data to test 1st and 2nd extreme data points. Using just 30 out of total rows affected the distribution of data so results of the test could be skewed but directionally it provided good insights. It showed that highest values 1969 and 1993 were outliers (very low P-Values). On the other hand, lowest values 342 and 373 ereot outliers (high P-Values).

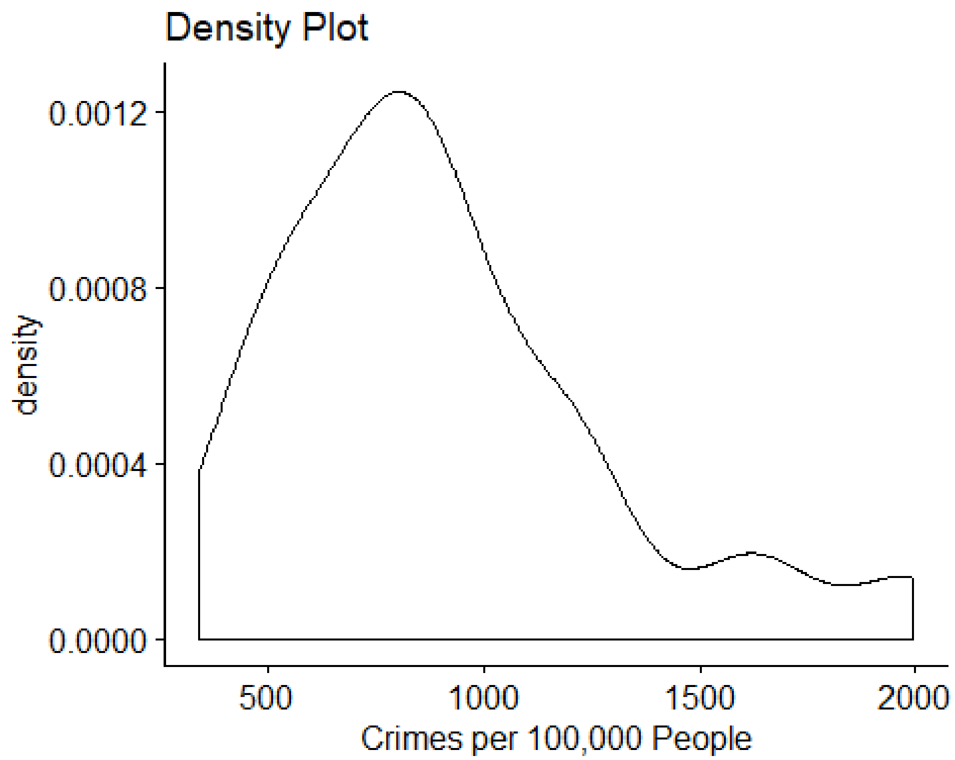
Grubbs test indicated at least 2 outliers (highest and 2nd highest value). But are there are two pending questions:

- Are these results reliable? And,
- Are there other outliers as well?

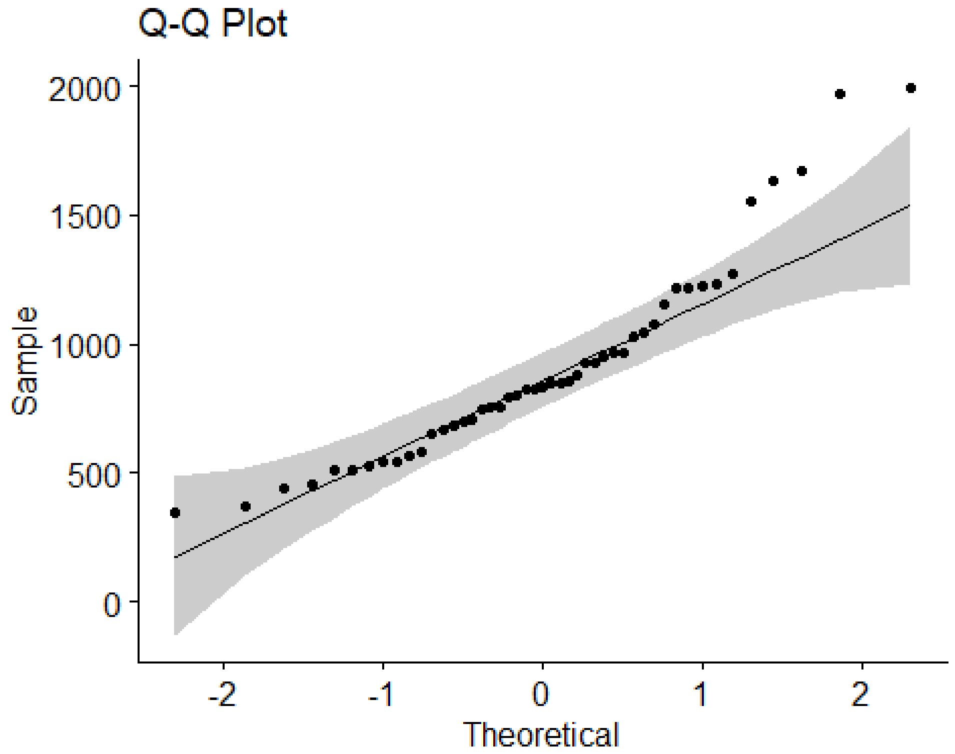
First on reliability, Grubbs test assumes that data has normal distribution (ref: https://en.wikipedia.org/wiki/Grubbs%27s_test_for_outliers). We did not check this condition up front. I checked normality of the data to ensure that our results from the Grubbs test are reliable.

There are multiple ways of checking normality of data (ref: http://www.sthda.com/english/wiki/normality-test-in-r). SOme are visual and some based on tests. Let’s look at several options below.

```
#density plot
ggdensity(crime_data,
          main = "Density Plot",
          xlab= "Crimes per 100,000 People")
```



```
#Q-Q plot
ggqqplot(crime_data,
         main = "Q-Q Plot")
```



```
#shapiro-wilk test
shapiro.test(crime_data)

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

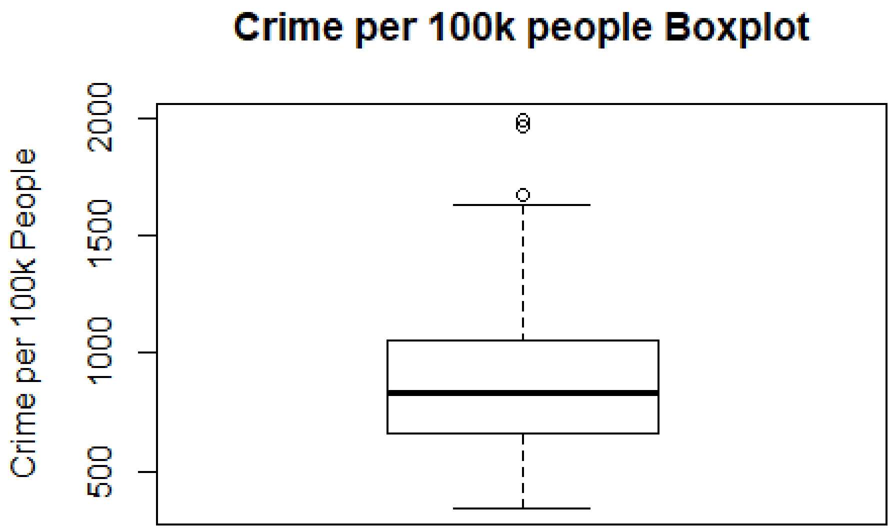
Density Plot showed similar trend to what we found from the basis statistics of the crime data i.e. data was normal for the most part except one side of the distribution.

Q-Q plot (quantile-quantile plot) draws the correlation between a given sample and the normal distribution. A 45-degree reference line is also plotted. It showed most points were within the normal distribution range except a few high value points (this was similar information to what Density plot showed).

Lastly, I checked the **Shaprio-Wilk Normality Test**. The P-Value was 0.0018 which is very low compared to 0.05, thus NULL hypothesis that data is normal is rejected.

I plotted the **boxplot** of the data as well to validate the information above and it presented similar insights. Data showed outliers on one extreme whcih aligned with the outcome of Grubbs tests.

```
#box plot to visually check outliers
boxplot(crime_data,
        main = "Crime per 100k people Boxplot",
        ylab = "Crime per 100k People"
        )
```



Given we saw portion of normal behavior with some outliers on one edge, there is some validty in the Grubbs test, though not 100% reliable but the outcome of the tests is mostly supported by the visual analysis above.

Secondly, Grubbs test only checks one or two points at a time, what about other possible outliers? I used a few more broader tests to check possible other outliers.

```
#values outside the boxplot
boxplot.stats(crime_data)$out

## [1] 1969 1674 1993

#outliers based on Z-Score test
#ref:
https://www.rdocumentation.org/packages/climtrends/versions/1.0.6/topics/Find
OutliersZscore

crime_data[FindOutliersZscore(crime_data, coef=2.5)]

## [1] 1969 1993

#Outliers based on the absolute deviation around the median
#ref:
https://www.rdocumentation.org/packages/climtrends/versions/1.0.6/topics/Find
OutliersMAD

crime_data[FindOutliersMAD(crime_data,coef = 3)] #coef=3, very conservative
```

```
## [1] 1969 1993

crime_data[FindOutliersMAD(crime_data,coef = 2.5)] #coef=2.5, moderately
conservative

## [1] 1635 1969 1674 1993

crime_data[FindOutliersMAD(crime_data,coef = 2)] #coef=2, poorly conservative

## [1] 1635 1969 1555 1674 1993
```

The **Boxplot.stats** function returns the data points that were outside the extremes of the whiskers. It showed clearly that top 3 high value data points were returned to be outside the whisker extreme as outliers.

I then tried the **FindOutliersZscore** function from the climtrends package. Using the default coef of 2.5, the two highest points were returned as outliers (same as Grubbs tests).

Lastly, I tried the **FindOutliersMAD** function from the climTrends package. This function identified outliers based on MAD and coef values determined how convservative or open we wanted to be in outliers identification. All coef options identified outliers on the maximum end of the distribution with conversation coef identifying only two highest points and the other coef including a few more points as outliers.

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?

At my job at the T-Mobile HQ in the Seattle area, my team helps get analytics products built for our procurement and supply chain teams. One of the key insights we provide to our organization is tracking of Key Performance Indictors (KPIs) and operational metrics. These measures keep the organizational leaders informed regarding the performance of the teams and whether performance towards key business goals is on track. Examples of these KPIs and metrics are **deal cycle time** (time between kick off of a deal negotiation and final contract signatures), **Key Supplier Performance Scores**, **Purchase Requisiton Approval Time**, **Material Fulfillment Cycle Time** (time between a project is allocated material to and material is picked in the warehouse and staged for shipping) etc. There are targets for each of these KPIs which teams track religiously. Teams would want to know quickly if a certain SLA or metric is slipping. CUSUM could be used to monitor these KPIs and detect when a change above or below the threshold.

The threshold for each KPI and metric would vary depending on the cost of not detecting early and false detection. Statistically speaking, typical critical value is half of standard deviation and threshold is set at 4 times the standard deviation (*source*: https://www.spcforexcel.com/knowledge/variable-control-charts/keeping-processtarget-cusum-charts) . But as noted above, depending on the trade off between early detection and false alarm, the critical value and threshold could be adjusted. For example, material fulfillment cycle is critical (if material doesn't ship on-time, construction of the cellular network gets delayed) and C & T values for this KPI would be lower. Whereas deal cycle time has more wiggle room and C & T values for this measure would be relatively higher.

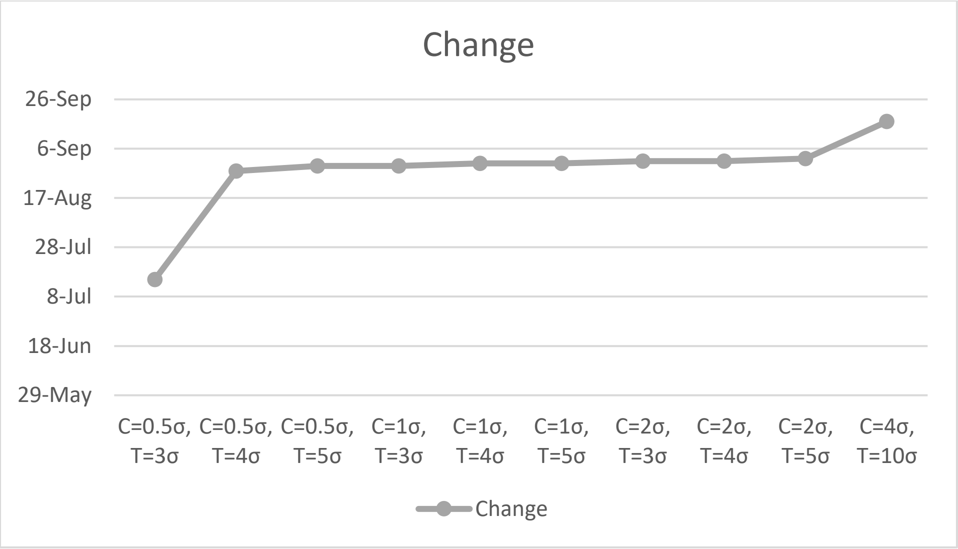
QUESTION 6.2

- **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.**

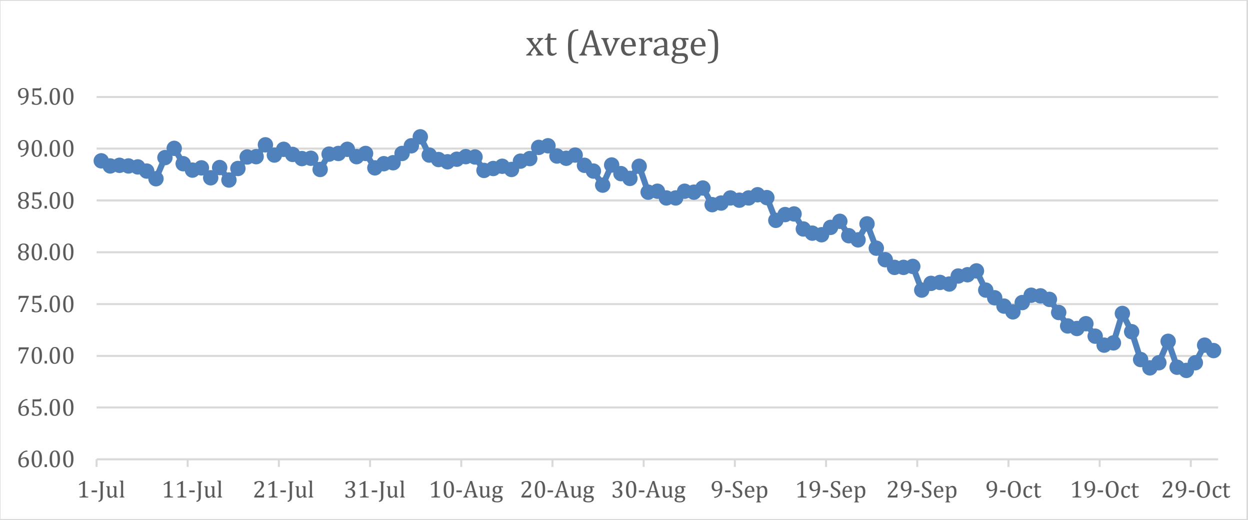
The approach I took for this question is as following:

- For each day in July 1st to Oct 31st, I used the average temperature across years.
- I calculated the average (mu) for July averaged temperatures only. This makes sense because CUSUM requires a mean for no change data. July is the least change summer month.
- I calculated the standard deviation for the same July average temperatures.
- Statistically, C should be 0.5 times standard deviation and T should be 4 times standard deviation. But if the data is too sensitives or less sensitive, higher or lower values might be needed. To check for this, I calculated S_t for a range of C and T values to see where change is indicated.
- Since we are detecting a decrease, the S_t formula for decrease is used.

$$S_t = \max\{0, S_{t-1} + (\mu - x_t - C)\}$$



The line chart above is the plot of dates when summer end was indicated (Y-Axis) for a given combination of C and T Values. It shows that end of August to early September is the commonly detected timeframe when temperatures started to decrease indicating end of summer. For a significantly large C and T (last data point of the plot above), the change is detected for near September end. This means that even for a very high C and T, summer end is not later than Sep end.



To confirm this further, average temperature across years for each date (Y-Axis) is plotted against the dates to see when temperature change occurs. It can be seen that August end is when temperature really starts to drop, indicating an end of summer.

Thus, it can be concluded that end of August is the most probably end of summer with the range being end of August to end of September. (Detailed calculations are in the attached spreadsheet tab: Question 6.2.1).

- **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).**

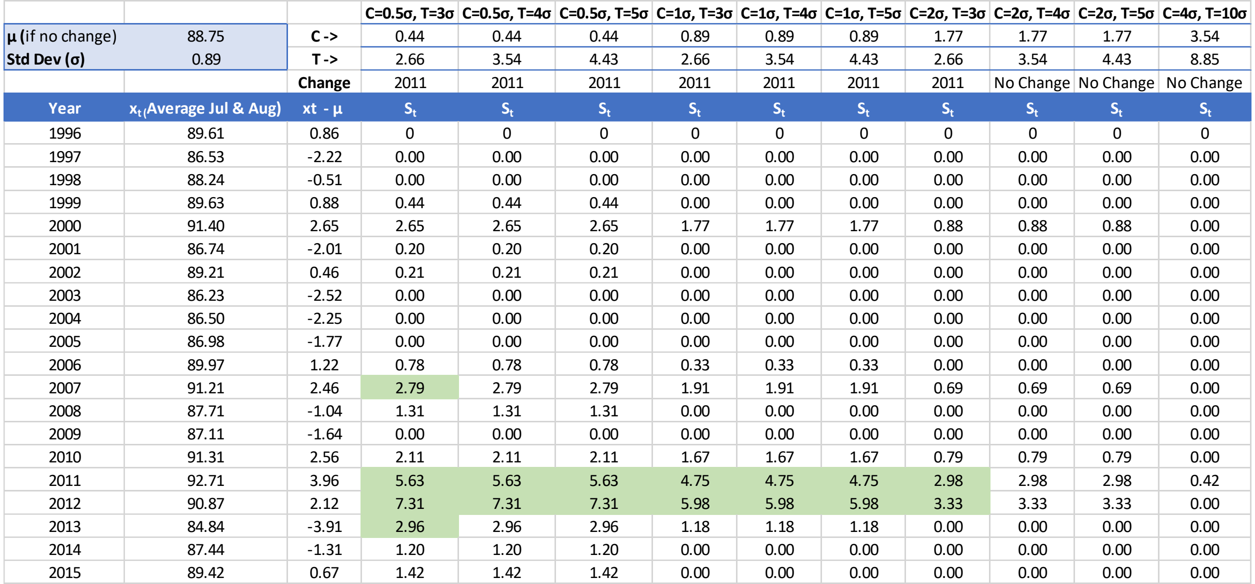
My understanding of this questions was that we have to determine whether summers in Atlanta over the years have gotten warmer. This means that we need to check whether for the same summer months, have temperatures started to increase over the years and if yes, since when.

I used the following approach for this question:

- Given the outcome of 6.2.1, I used July 1st to August 31st as the summer timeframe for each year.
- I averaged the temperature for each year from July 1st to August 31st.
- I used the same mean (mu) and standard deviation as 6.2.1 i.e. July average for each day in order to satisfy the no change requirement.
- I calculated S_t for the same set of C and T values as 6.2.1 question.
- I used the CUSUM formula to detect increase.

$$S_t = \max\{0, S_{t-1} + (x_t - \mu - C)\}$$

The table below is a screenshot of the calculation. Actual spreadsheet (Tab: Questions 6.2.2) is also attached with the submission.

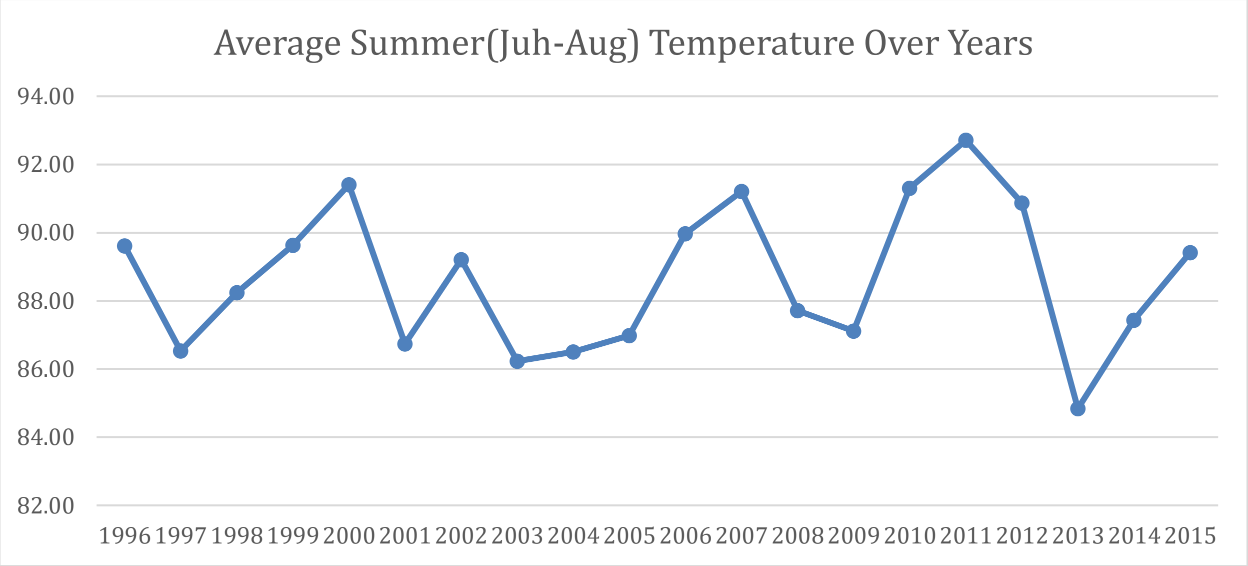


			C=0.5σ, T=3σ	C=0.5σ, T=4σ	C=0.5σ, T=5σ	C=1σ, T=3σ	C=1σ, T=4σ	C=1σ, T=5σ	C=2σ, T=3σ	C=2σ, T=4σ	C=2σ, T=5σ	C=4σ, T=10σ
μ (if no change)	88.75	**C ->**	0.44	0.44	0.44	0.89	0.89	0.89	1.77	1.77	1.77	3.54
Std Dev (σ)	0.89	**T ->**	2.66	3.54	4.43	2.66	3.54	4.43	2.66	3.54	4.43	8.85
		Change	2011	2011	2011	2011	2011	2011	2011	No Change	No Change	No Change
Year	**x_t(Average Jul & Aug)**	**xt - μ**	**S_t**	**S_t**	**S_t**	**S_t**	**S_t**	**S_t**	**S_t**	**S_t**	**S_t**	**S_t**
1996	89.61	0.86	0	0	0	0	0	0	0	0	0	0
1997	86.53	-2.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1998	88.24	-0.51	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1999	89.63	0.88	0.44	0.44	0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2000	91.40	2.65	2.65	2.65	2.65	1.77	1.77	1.77	0.88	0.88	0.88	0.00
2001	86.74	-2.01	0.20	0.20	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2002	89.21	0.46	0.21	0.21	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2003	86.23	-2.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2004	86.50	-2.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2005	86.98	-1.77	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2006	89.97	1.22	0.78	0.78	0.78	0.33	0.33	0.33	0.00	0.00	0.00	0.00
2007	91.21	2.46	2.79	2.79	2.79	1.91	1.91	1.91	0.69	0.69	0.69	0.00
2008	87.71	-1.04	1.31	1.31	1.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2009	87.11	-1.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2010	91.31	2.56	2.11	2.11	2.11	1.67	1.67	1.67	0.79	0.79	0.79	0.00
2011	92.71	3.96	5.63	5.63	5.63	4.75	4.75	4.75	2.98	2.98	2.98	0.42
2012	90.87	2.12	7.31	7.31	7.31	5.98	5.98	5.98	3.33	3.33	3.33	0.00
2013	84.84	-3.91	2.96	2.96	2.96	1.18	1.18	1.18	0.00	0.00	0.00	0.00
2014	87.44	-1.31	1.20	1.20	1.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2015	89.42	0.67	1.42	1.42	1.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00

The outcome is summarized in the table below. It can be seen that for most C and T values, change is detected in 2011.

	C	T	Change
C=0.5σ, T=3σ	0.44	2.66	2011
C=0.5σ, T=4σ	0.44	3.54	2011
C=0.5σ, T=5σ	0.44	4.43	2011
C=1σ, T=3σ	0.89	2.66	2011
C=1σ, T=4σ	0.89	3.54	2011
C=1σ, T=5σ	0.89	4.43	2011
C=2σ, T=3σ	1.77	2.66	2011
C=2σ, T=4σ	1.77	3.54	No Change
C=2σ, T=5σ	1.77	4.43	No Change
C=4σ, T=10σ	3.54	8.85	No Change

The confirm this, I plotted the July to Aug average temperature (Y-Axis) for each year to get an idea of the change. It can be seen that 2011 indeed indicated a change but it was followed by couple of years of average temperature drop before it picking back up.



Thus, 2011 does seem like the year when summer got warmer, but it dropped for couple of years after that as well which is why the higher C and T values do not show any change at all.