ISYE-6501, Fall 2022, Homework 3

Sep 14, 2022

Question 5.1

Goal

Test crime data to see whether there are any outliers in the variable *number of crimes per 100,000 people* using the *grubbs.test* R function.

Methodology

The data is the Effect of Punishment Regimes on Crime Rates data set available from the Australasian Data and Story Library (OzDASL). The data set was provided as part of this homework assignment and not sourced directly from OzDASL.

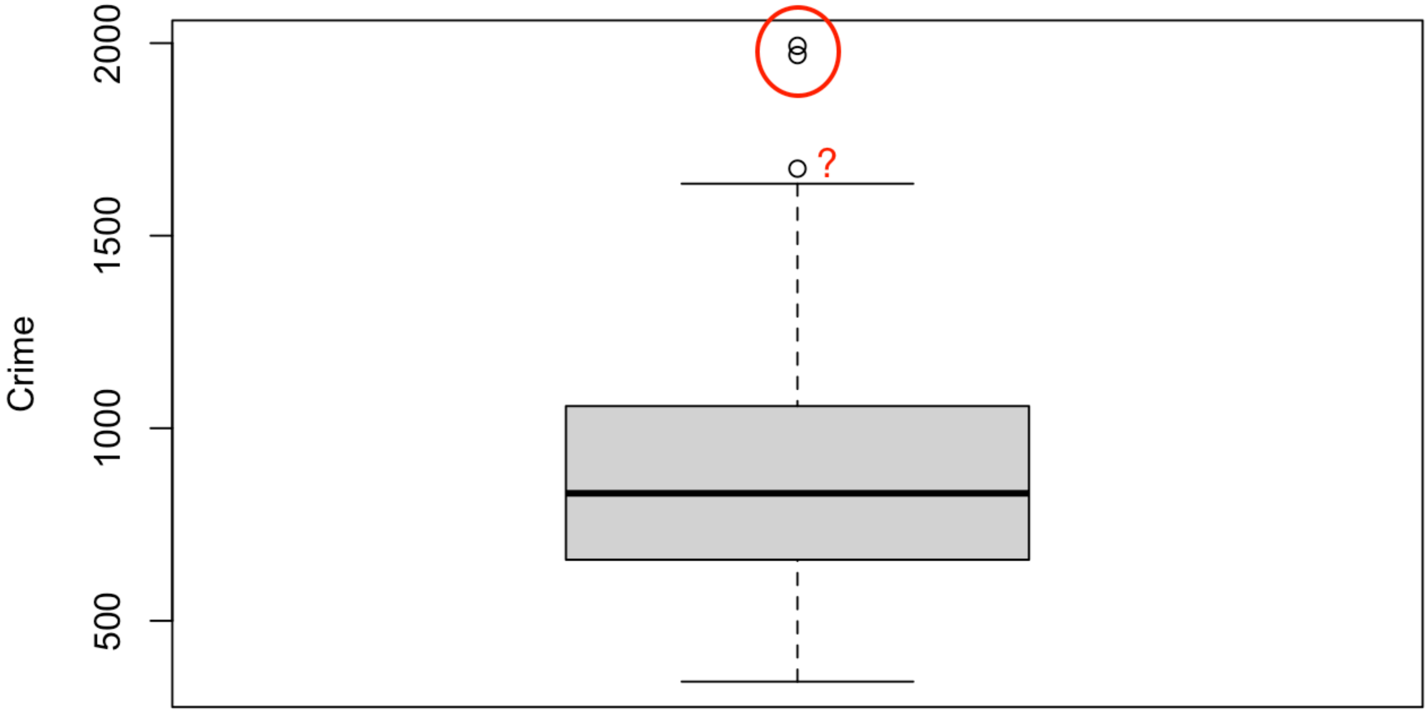
It contains aggregated crime data from 47 U.S. states for 1960 and consists of 47 data points with 16 variables. The target variable for this assignment is *Crime*, the last variable, which represents the crime rate as the number of offenses per 100,000 population.

In ascending order, the crime values are:

342, 373, 439, 455, 508, 511, 523, 539, 542, 566, 578, 653, 664, 682, 696, 705, 742, 750, 754, 791, 798, 823, 826, 831, 849, 849, 856, 880, 923, 929, 946, 963, 968, 1030, 1043, 1072, 1151, 1216, 1216, 1225, 1234, 1272, 1555, 1635, 1674, 1969, 1993.

There are two duplicate values, 849 and 1216.

I constructed a box plot to visually see if any outliers may be present.

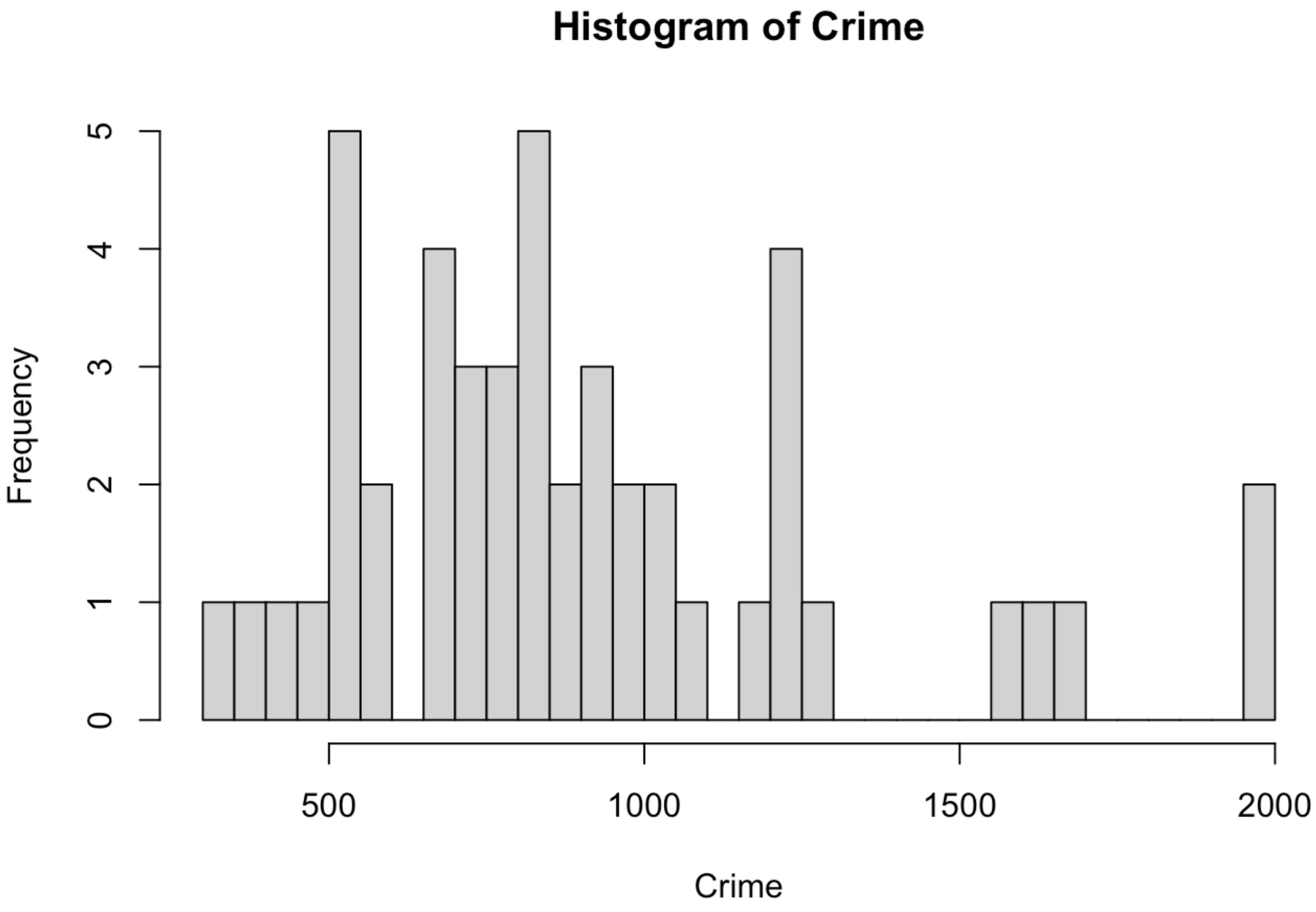


There does indeed appear to be two, possibly three, outliers on the large end of values. There do not appear to be any outliers on the small end of values.

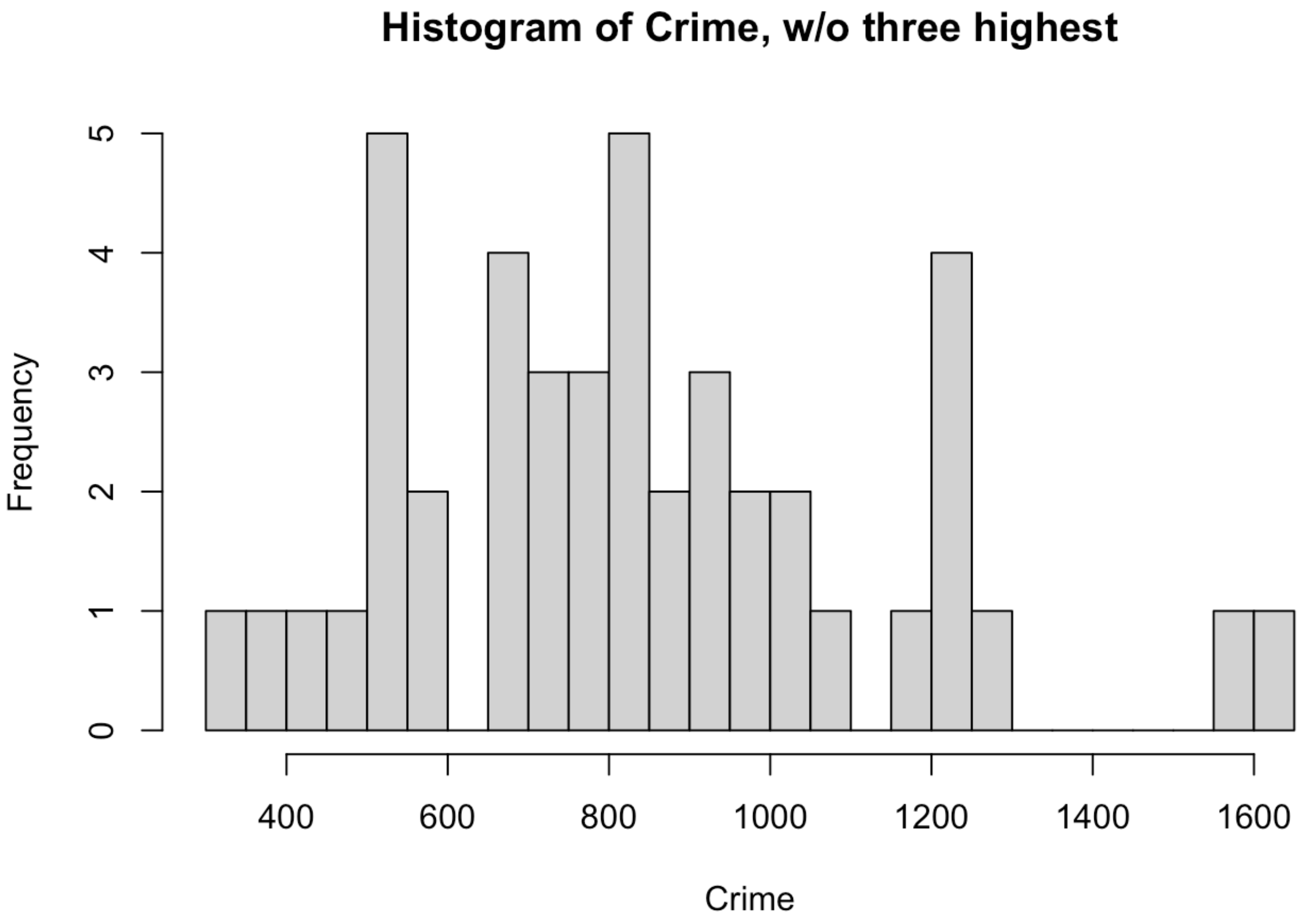
The box plot is a visual indication to the presence of outliers. The Grubbs test can provide quantitative support for the presence of outliers.

According to Grubbs (1969) one of the assumptions for this test is that the data should be normally distributed.

A histogram of this data shows several peaks. This crime data does not appear normally distributed to me.



Even if I remove the three highest values 1674, 1969, 1993 (the values that stood out on the box plot as possible outliers), the distribution still does not look any better to me. There are several peaks.



Given the distribution of this data, the Grubbs test may not be the best method to use.

Results

I ran several Grubbs one-sided tests. The results are as follows.

Test #1

```
#One-tail test on largest value
mod <- grubbs.test(crime, two.sided=FALSE)
print(mod)
```

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

In this hypothesis test:

- Null hypothesis: Highest value 1993 is not an outlier
- Alternative hypothesis: Highest value 1993 is an outlier

The p-value is **0.0788749**. If we use a significance level of 5%, a commonly used level of significance for hypothesis testing, the p-value is greater than 0.05, so we should fail to reject the null hypothesis and conclude that 1993 is not an outlier.

If however, we use a significance level of 10%, then we should reject the null hypothesis and conclude that 1993 is an outlier. It would be up to the situation as to whether this level of significance is acceptable or not.

A more holistic approach may be used to determine if this value should be treated as an outlier. By examining some of the the other data in this set, something interesting stood out.

The state with the crime rate of 1993 has a population of 300k (the variable Pop). That is the smallest population value in the data set. There is only one other state in this data set that has a population of 300k. This other state has a crime rate of 849.

They are both non-southern states, so we could conclude that weather may not be a factor in crime rate, e.g., less crime in states with harsher weather because people stay indoors more vs. more crime in states with temperate climates.

If we examine the variable Po1, the per capita expenditure for police protection in 1960, for these two states, we can see that the state with the lower crime rate has spent a little over half (9) what the state with the highest crime rate spent (16) in 1960.

Row Number	Population	Police Funding	Crime
26	3	16	1993
46	3	9	849

In other words, the state with the highest crime rate out of all 47 states, spent almost 2 times as much in police protection as another state with similar population, but it had almost twice the crime rate. That seems odd given that a reasonable person would expect a negative correlation between police funding and crime rate. This does not mean that data is incorrect; it may be real data. But as a whole, something is not adding up and therefore this data point should be investigated further.

Test #2

To test the second highest value, 1969, we can remove 1993 from the data and run the one-tail test again.

```
#One-tail test on second largest value
mod <- grubbs.test(sort(crime)[-length(crime)], two.sided=FALSE)
print(mod)
```

```
##
##  Grubbs test for one outlier
##
## data:  sort(crime)[-length(crime)]
## G = 3.06343, U = 0.78682, p-value = 0.02848
## alternative hypothesis: highest value 1969 is an outlier
```

In this hypothesis test:

- Null hypothesis: Highest value 1969 is not an outlier
- Alternative hypothesis: Highest value 1969 is an outlier

The p-value is **0.0284782**. The p-value is less than 0.05 so we should reject the null hypothesis and conclude that 1969 is an outlier.

Test #3

According to the initial box plot, there may be a third outlier, although that one was close call. I tested it. I ran the test again, this time removing the two largest values 1993 and 1969.

```
#One-tail test on third largest value
mod <- grubbs.test(sort(crime)[-c(length(crime), length(crime)-1)], two.sided=FALSE)
print(mod)
```

```
##
##  Grubbs test for one outlier
##
## data:  sort(crime)[-c(length(crime), length(crime) - 1)]
## G = 2.56457, U = 0.84712, p-value = 0.1781
## alternative hypothesis: highest value 1674 is an outlier
```

In this hypothesis test:

- Null hypothesis: Highest value 1674 is not an outlier
- Alternative hypothesis: Highest value 1674 is an outlier

The p-value is **0.1780797**. The p-value is not less than 0.05 so we should fail to reject the null hypothesis and conclude that 1674 is not an outlier.

Test #4

So far, all of the tests have focused on the largest values because that is what the initial box plot indicated.

I forced the grubbs.test function to test whether the minimum value, 342, was an outlier by using the *opposite* parameter.

```
#One-tail test on smallest value
mod <- grubbs.test(crime, two.sided=FALSE, opposite=TRUE)
print(mod)
```

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

In this hypothesis test:

- Null hypothesis: Lowest value 342 is not an outlier
- Alternative hypothesis: Lowest value 342 is an outlier

The p-value is **1**. The p-value is not less than 0.05 so we should fail to reject the null hypothesis and conclude that 342 is not an outlier.

Conclusion

A box plot of the crime data indicates that outliers may be present.

The crime data does not appear to be a normal distribution and so the Grubbs test may not be the best method to use.

A holistic approach is beneficial when trying to make sense of unexpected (or expected) test results.

Grubbs test suggests that 1993 and 1969 are outliers, while 1969 and 342 are not outliers.

R Code

Base R version 4.2.1 (2022-06-23)

```
#Clear environment
rm(list=ls())

#Load the data
uscrime <- read.table("~/RWork/Data/uscrime.txt", header=TRUE)

#Inspect data
class(uscrime)
dim(uscrime)
str(uscrime)
summary(uscrime)
head(uscrime)
tail(uscrime)

#Isolate Crime vector
crime <- uscrime$Crime

#Inspect crime values
summary(crime)
sort(crime)
sort(unique(crime))

#Create box plot to see if any outliers may be present
boxplot(crime, ylab="Crime")

#Check rows with largest three values
uscrime[which(uscrime$Crime %in% c(1674, 1969, 1993)),]

#Create histogram to check distribution of data
```

```
hist(crime, breaks=40, main="Histogram of Crime", xlab="Crime")
hist(sort(crime)[-c(45,46,47)], breaks=40, main='Histogram of Crime, w/o three highest', xlab="Crime")

#Load "outliers" package
library(outliers)

#One-tail test on largest value
grubbs.test(crime, two.sided=FALSE)

#Check data, order by population; is crime low or high for lowest populated states?
head(uscrime[order(uscrime$Pop),], n=10)

#Inspect rows with population less than or equal to 3
uscrime[which(uscrime$Pop<=3),]

#One-tail test on second largest value
grubbs.test(sort(crime)[-length(crime)], two.sided=FALSE)

#One-tail test on third largest value
grubbs.test(sort(crime)[-c(length(crime), length(crime)-1)], two.sided=FALSE)

#One-tail test on smallest value
grubbs.test(crime, two.sided=FALSE, opposite=TRUE)

#Detach outliers library
detach("package:outliers", unload = TRUE)
```

References

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. Technometrics, 11(1), 1–21. https://doi.org/10.2307/1266761 (https://doi.org/10.2307/1266761)

Smyth, GK (2011). Effect of Punishment Regimes on Crime Rates. Australasian Data and Story Library (OzDASL). http://www.statsci.org/data/general/uscrime.html (http://www.statsci.org/data/general/uscrime.html).

Question 6.1

My department conducts mailing campaigns to help educate customers on aberrant billing patterns with the hopes that they adjust their processes and fall into compliance with accepted standards.

CUSUM techniques may be used to detect changes in billing patterns over time to determine whether or not the mailing campaigns have a positive, negative, or no effect at all. This information could help decision makers determine if the campaigns should be continued, modified, or stopped all together.

Suppose we start recording billing patterns on January 1 and mail letters on April 1 of that same year. Suppose also that from the time the letters are mailed, you expect it to take two weeks for transit and receipt of the letters, and an additional six weeks to allow time for behaviors to change, i.e., you must allow some time for people to evaluate, decide, and then act upon the request. So let’s say 8 weeks (2 months total), putting us at June 1. This is when we could expect to start to see a change. Thus, the period from Jan 1 until Jun 1 would be the baseline, non-changing period.

Both the threshold and critical value would need to be fine tuned according to the nuances in the data and business needs through trial-and-error.

In this example, one method to get a starting value for threshold could be by using three standard deviations of the observed billing metrics from Jan 1 to Jun 1 as a threshold to detect significant changes.

One method to find a starting value for the critical value may be to just set it at 0 initially and see what changes are detected. If the model is too sensitive, and there are too many false-positives, we could begin to increase the value of this parameter to try and reduce them in a trial-and-error approach.

For example, in this scenario, if the model detected changes prior to April 1 when the letter were mailed, we may want to try increasing the critical value to tune these out. At the same time, you would have to be mindful that increasing this parameter by too much could potentially make the model so insensitive to changes, that you could actually miss detecting a change that would have been meaningful.

Question 6.2

Goal

Part 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.

Part 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).

Methodology

I conducted this analysis in Excel. I started by importing into a blank Excel workbook the raw data provided in the "temps.txt" file. I used the built-in Excel data import functionality to do this.

I examined the data and concluded it appeared clean to me. It consisted of the daily high temperature in Atlanta, Georgia, USA from July 1 to October 31 (123 days) for each year from 1996 to 2015 (20 years), for a grand total of 2460 temperature observations.

There are two questions at hand:

1. When is the unofficial end of summer based on temperature?
2. Is Atlanta's climate getting warmer?

To begin, I wanted to get an of expectation of when summer ends based on other definitions.

I used both Labor Day, a U.S. holiday and the traditional (cultural) end of summer, as well as the end of the Summer Solstice, the astronomical end of summer, to give me rough idea of what "end of summer" means.

I used an online resource (timeanddate, n.d.) to look up the dates for Labor Day and the end of the Summer Solstice for each of the 20 years.

Year	Labor Day	End of Summer Solstice
1996	2-Sep	21-Sep
1997	1-Sep	21-Sep
1998	7-Sep	22-Sep
1999	6-Sep	22-Sep
2000	4-Sep	21-Sep
2001	3-Sep	21-Sep
2002	2-Sep	22-Sep
2003	1-Sep	22-Sep
2004	6-Sep	21-Sep
2005	5-Sep	21-Sep
2006	4-Sep	22-Sep
2007	3-Sep	22-Sep
2008	1-Sep	21-Sep

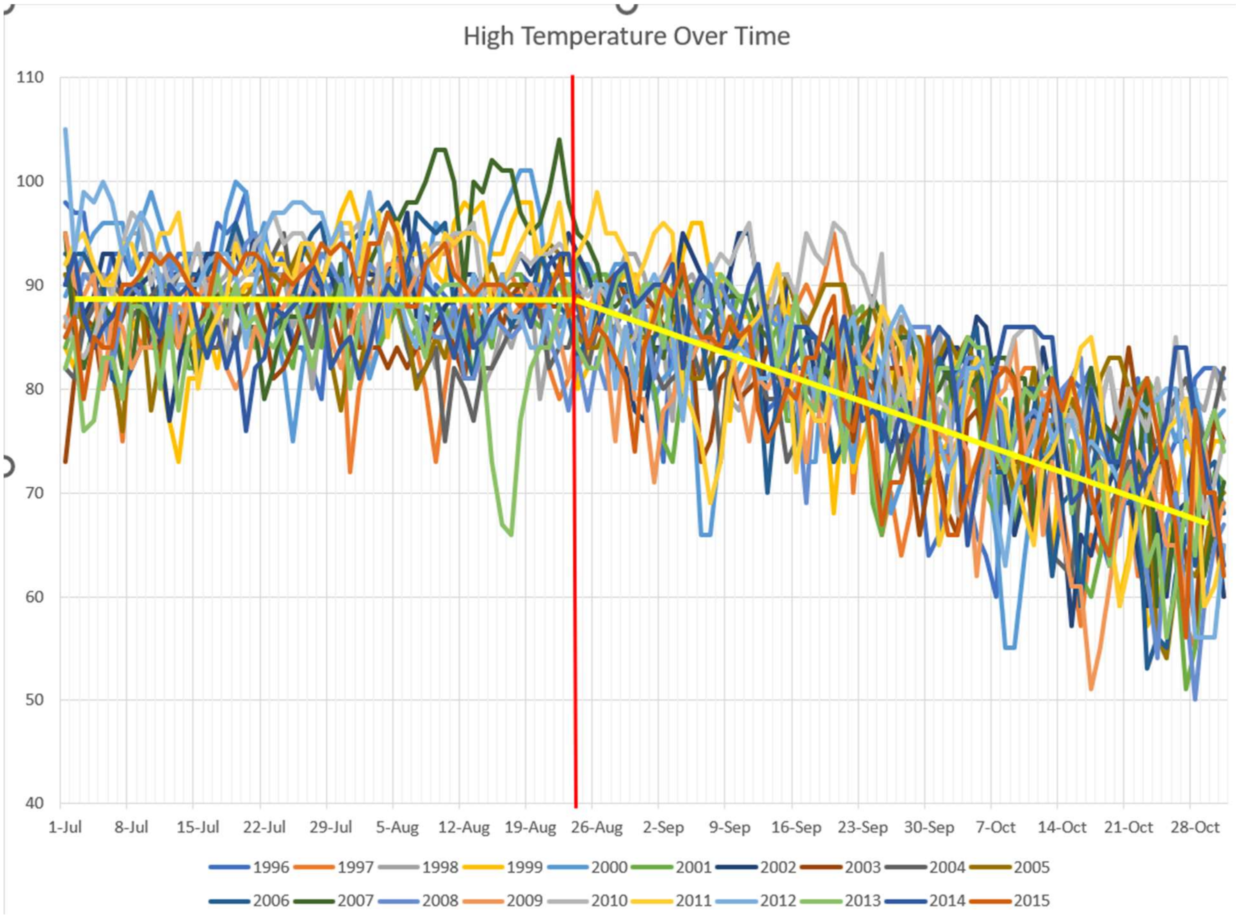
2009	7-Sep	21-Sep
2010	6-Sep	21-Sep
2011	5-Sep	22-Sep
2012	3-Sep	21-Sep
2013	2-Sep	21-Sep
2014	1-Sep	21-Sep
2015	7-Sep	22-Sep

I then took the mode of both and arrived at Sep 1 for Labor Day and Sep 21 for Summer Solstice. This gave me a rough idea of what I should be looking for.

To apply the cumulative sum method, I needed to find three things:

1. μ: The average of the unchanged temperature
2. C: A buffer value to account for random variation
3. T: The threshold that triggers that a change has occurred

To find μ, I used the suggested method that was presented in the ISYE 6501 office hours on Monday Sept 12, 2022 (Hsu, 41:53). I first plotted all 2046 temperature values over time.



From this plot I could see a relatively flat area up until around Aug 24 where all of the yearly line plots start exhibiting a downward trend.

To get μ for each year, I calculated the average temperature for each year from Jul 1 to Aug 24. There are 55 such values for each year in the data set. For easier explanation (I hope) I will call these the *baseline non-changing temperatures*.

For the threshold, T, I had to decide on a value that would indicate a large enough deviation from the baseline non-changing temperatures to indicate a significant enough change in the temperature so that the "end of summer" could be found.

For this value of T, I decided to take sample standard deviation of the 55 baseline non-changing temperatures, and multiply it by three. Although there doesn't seem to be a standard cutoff for how many standard deviations from the mean is considered an outlier, three seems to be a commonly agreed upon value, so I went with it for this assignment.

The buffer, C, was determined by trail-and-error for each year. By this I mean, I had to adjust that number for each year based on the specific randomness that occurred for that year. For example, some years were warmer for longer periods of time at the beginning. Some years started off cooler than others. Some years exhibited a gradual decrease in temperatures. And in some years, the decrease in temperatures was sharp and easily detectible.

The basic approach I used was to start with a C of zero and see how many false-positive hits I got. Given the initial intuition from examining the Labor Day and Summer Solstice dates, if I got an indicator of the end of summer in July, I would consider that a false-positive and adjust the C value upward by a standard increment of 0.5. I kept doing this until the false-positive either disappeared entirely, or were reduced to just one.

Although my initial strategy was to reduce the false-positive completely by adjusting C, I found that this was not practical because in some years the temperatures would spike upwards or downwards very early in the year. I would get a false-positive very early on, sometimes even on Jul 1. Eliminating these types of false-positives was almost impossible without adjusting the value of C so high that it made the rest of the model very insensitive, and I risked false-negatives instead, i.e., not identifying the end of summer early enough.

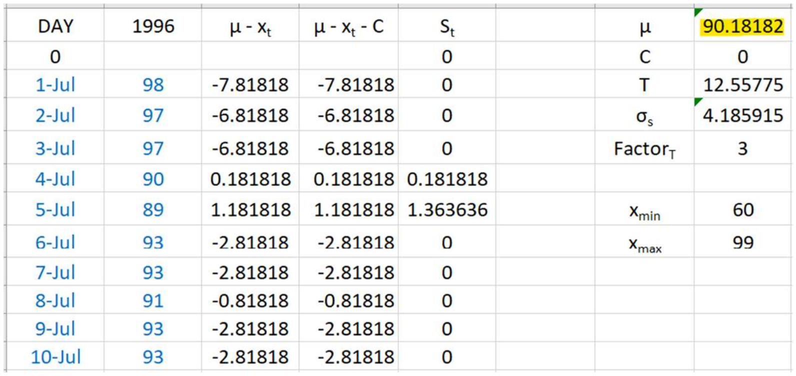
Since I am trying to test for decreasing values, in order to calculate S_t, I had to subtract x_t from μ.

The equation used to find S_t was:

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

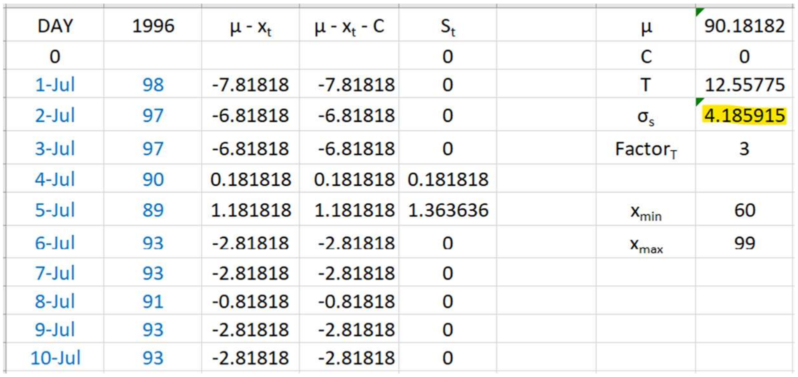
I used this strategy over and over for each year. Using 1996 as an example, these were the steps I used:

1. Calculate the mean for the 55 values from Jul 1 to Aug 24. The mean is 90.2.



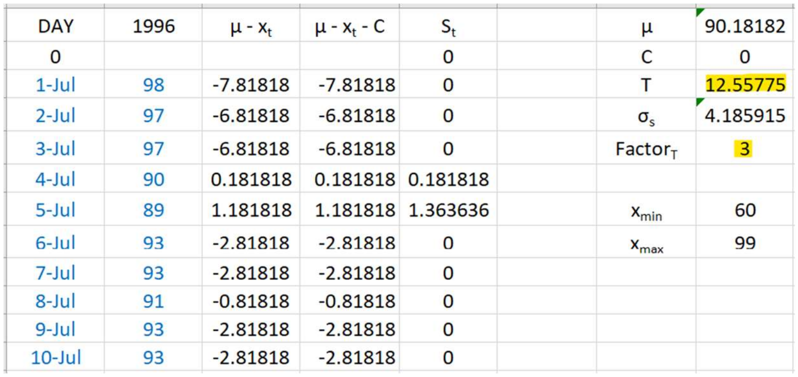
DAY	1996	$\mu - x_t$	$\mu - x_t - C$	S_t		μ	90.18182
0				0		C	0
1-Jul	98	-7.81818	-7.81818	0		T	12.55775
2-Jul	97	-6.81818	-6.81818	0		σ_s	4.185915
3-Jul	97	-6.81818	-6.81818	0		Factor_T	3
4-Jul	90	0.181818	0.181818	0.181818			
5-Jul	89	1.181818	1.181818	1.363636		x_{min}	60
6-Jul	93	-2.81818	-2.81818	0		x_{max}	99
7-Jul	93	-2.81818	-2.81818	0			
8-Jul	91	-0.81818	-0.81818	0			
9-Jul	93	-2.81818	-2.81818	0			
10-Jul	93	-2.81818	-2.81818	0			

2. Calculate the sample standard deviation for all 55 values from Jul 1 to Aug 24. The standard deviation is 4.2.



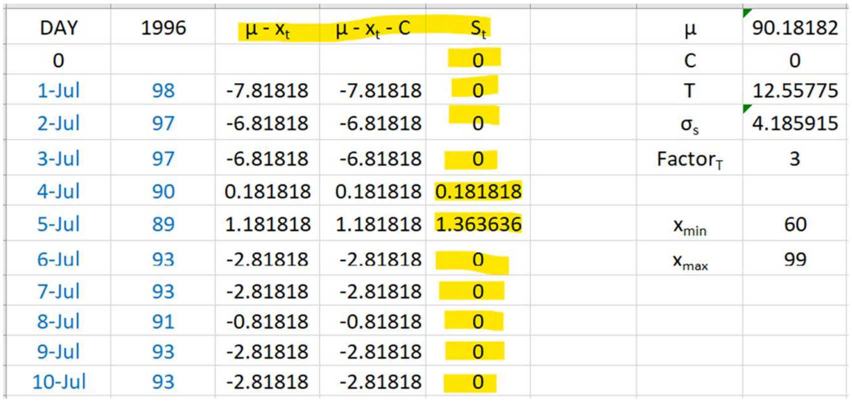
DAY	1996	$\mu - x_t$	$\mu - x_t - C$	S_t		μ	90.18182
0				0		C	0
1-Jul	98	-7.81818	-7.81818	0		T	12.55775
2-Jul	97	-6.81818	-6.81818	0		σ_s	4.185915
3-Jul	97	-6.81818	-6.81818	0		Factor_T	3
4-Jul	90	0.181818	0.181818	0.181818			
5-Jul	89	1.181818	1.181818	1.363636		x_{min}	60
6-Jul	93	-2.81818	-2.81818	0		x_{max}	99
7-Jul	93	-2.81818	-2.81818	0			
8-Jul	91	-0.81818	-0.81818	0			
9-Jul	93	-2.81818	-2.81818	0			
10-Jul	93	-2.81818	-2.81818	0			

3. Multiply the sample standard deviation by 3. This is the threshold value T. Here it is 12.6.



DAY	1996	$\mu - x_t$	$\mu - x_t - C$	S_t		μ	90.18182
0				0		C	0
1-Jul	98	-7.81818	-7.81818	0		T	12.55775
2-Jul	97	-6.81818	-6.81818	0		σ_s	4.185915
3-Jul	97	-6.81818	-6.81818	0		Factor_T	3
4-Jul	90	0.181818	0.181818	0.181818			
5-Jul	89	1.181818	1.181818	1.363636		x_{min}	60
6-Jul	93	-2.81818	-2.81818	0		x_{max}	99
7-Jul	93	-2.81818	-2.81818	0			
8-Jul	91	-0.81818	-0.81818	0			
9-Jul	93	-2.81818	-2.81818	0			
10-Jul	93	-2.81818	-2.81818	0			

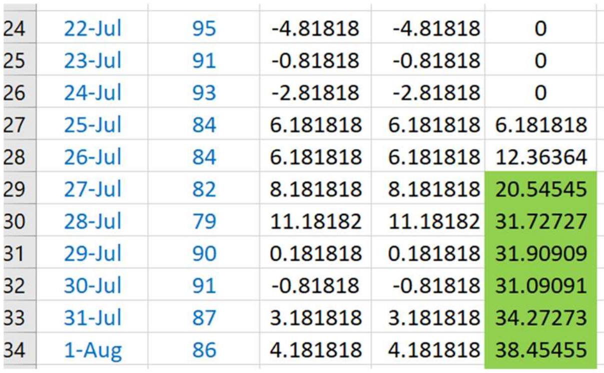
4. For each value x at time t, calculate S_t and take the maximum value of either S_t or 0. Do this for all 123 temperature values in the year.



DAY	1996	$\mu - x_t$	$\mu - x_t - C$	S_t			
0				0		C	0
1-Jul	98	-7.81818	-7.81818	0		T	12.55775
2-Jul	97	-6.81818	-6.81818	0		σ_s	4.185915
3-Jul	97	-6.81818	-6.81818	0		Factor_T	3
4-Jul	90	0.181818	0.181818	0.181818			
5-Jul	89	1.181818	1.181818	1.363636		x_{min}	60
6-Jul	93	-2.81818	-2.81818	0		x_{max}	99
7-Jul	93	-2.81818	-2.81818	0			
8-Jul	91	-0.81818	-0.81818	0			
9-Jul	93	-2.81818	-2.81818	0			
10-Jul	93	-2.81818	-2.81818	0			

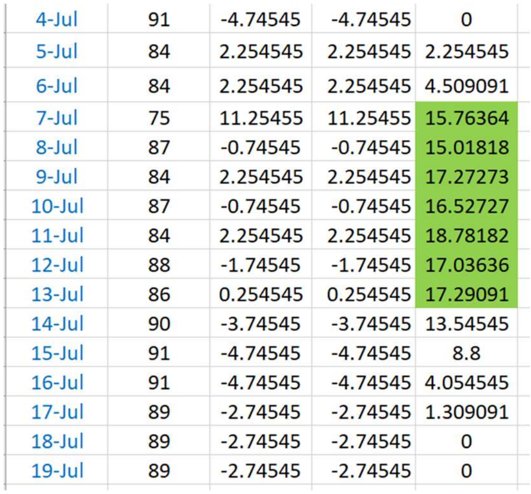
The first row of the right-hand block reads: μ | 90.18182

5. Compare S_t with the value of T and if S_t is greater than or equal to T, flag that temperature x at that t as a change. I accomplished this with conditional formatting in Excel. For example, using a C value of 0 for 1996, the first change that met or exceeded the threshold of 12.6 was registered on Jul 27.

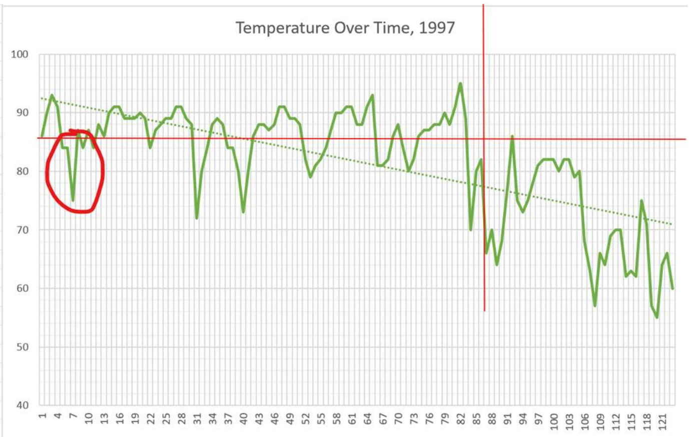


24	22-Jul	95	-4.81818	-4.81818	0
25	23-Jul	91	-0.81818	-0.81818	0
26	24-Jul	93	-2.81818	-2.81818	0
27	25-Jul	84	6.181818	6.181818	6.181818
28	26-Jul	84	6.181818	6.181818	12.36364
29	27-Jul	82	8.181818	8.181818	20.54545
30	28-Jul	79	11.18182	11.18182	31.72727
31	29-Jul	90	0.181818	0.181818	31.90909
32	30-Jul	91	-0.81818	-0.81818	31.09091
33	31-Jul	87	3.181818	3.181818	34.27273
34	1-Aug	86	4.181818	4.181818	38.45455

6. Adjust C upward in increments of 0.5 so as to eliminate any obvious false-positives. For example, in this data from the year 1997, C = 0 led to some very early false-positives to be registered on Jul 7 due to a drop in temperatures in July of that year.

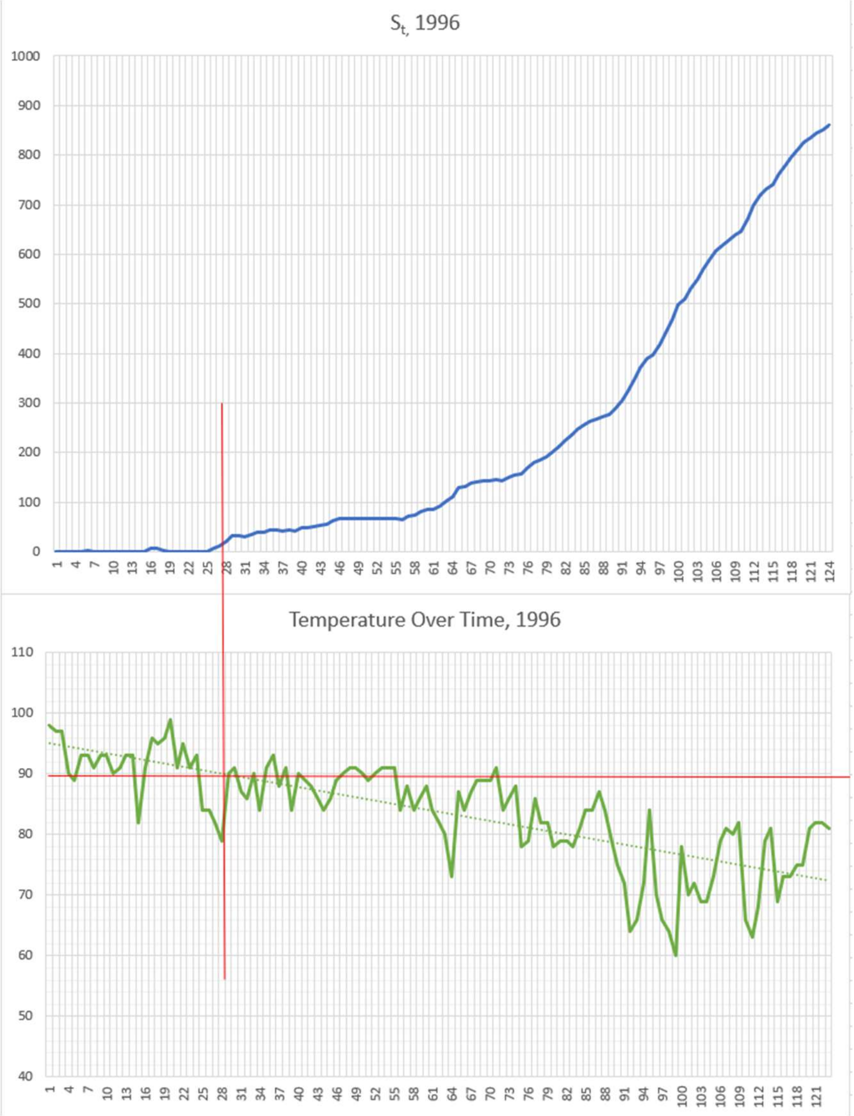


4-Jul	91	-4.74545	-4.74545	0
5-Jul	84	2.254545	2.254545	2.254545
6-Jul	84	2.254545	2.254545	4.509091
7-Jul	75	11.25455	11.25455	15.76364
8-Jul	87	-0.74545	-0.74545	15.01818
9-Jul	84	2.254545	2.254545	17.27273
10-Jul	87	-0.74545	-0.74545	16.52727
11-Jul	84	2.254545	2.254545	18.78182
12-Jul	88	-1.74545	-1.74545	17.03636
13-Jul	86	0.254545	0.254545	17.29091
14-Jul	90	-3.74545	-3.74545	13.54545
15-Jul	91	-4.74545	-4.74545	8.8
16-Jul	91	-4.74545	-4.74545	4.054545
17-Jul	89	-2.74545	-2.74545	1.309091
18-Jul	89	-2.74545	-2.74545	0
19-Jul	89	-2.74545	-2.74545	0



7. Once the optimal value of C was established, and again this was mainly done by intuition, I plotted both S_t and temperature just to get a visual idea of what was happening. I would plot µ on the temperature chart (the red horizontal line on the bottom chart) and extend a line down from the overhead plot of S_t where the change was detected to see if it makes sense.

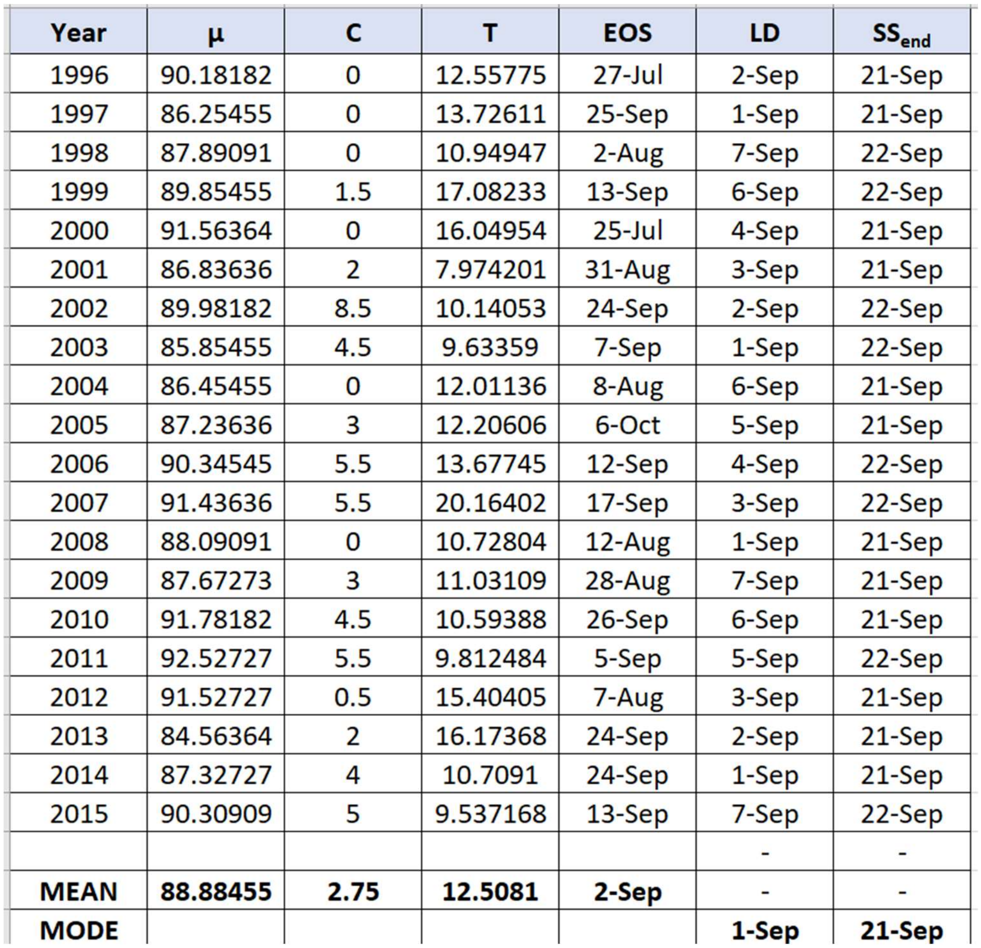
   For example, in the plots for 1996, the change was detected on Jul 27, that is the bump upward in the St chart (the top chart). Extending the line down, I can see that this corresponds to around the time when the temperatures began to trend (the green dotted diagonal line) below the non-changing mean (the red horizontal line). *Note: This was not always the case however.*



8. Record the date the changes were first detected by CUSUM as the unofficial end of summer.

9. Repeat these steps for the remaining years from 1997 to 2015.

I ended up with 20 values for µ, T, C, and the unofficial end of summer. These are summarized here:



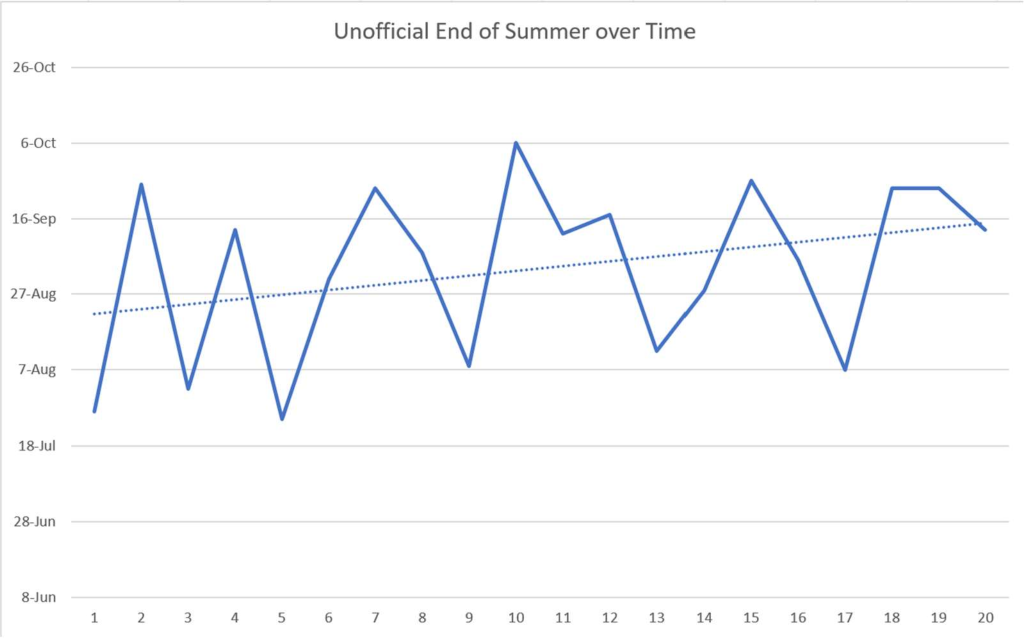
Year	µ	C	T	EOS	LD	SS_{end}
1996	90.18182	0	12.55775	27-Jul	2-Sep	21-Sep
1997	86.25455	0	13.72611	25-Sep	1-Sep	21-Sep
1998	87.89091	0	10.94947	2-Aug	7-Sep	22-Sep
1999	89.85455	1.5	17.08233	13-Sep	6-Sep	22-Sep
2000	91.56364	0	16.04954	25-Jul	4-Sep	21-Sep
2001	86.83636	2	7.974201	31-Aug	3-Sep	21-Sep
2002	89.98182	8.5	10.14053	24-Sep	2-Sep	22-Sep
2003	85.85455	4.5	9.63359	7-Sep	1-Sep	22-Sep
2004	86.45455	0	12.01136	8-Aug	6-Sep	21-Sep
2005	87.23636	3	12.20606	6-Oct	5-Sep	21-Sep
2006	90.34545	5.5	13.67745	12-Sep	4-Sep	22-Sep
2007	91.43636	5.5	20.16402	17-Sep	3-Sep	22-Sep
2008	88.09091	0	10.72804	12-Aug	1-Sep	21-Sep
2009	87.67273	3	11.03109	28-Aug	7-Sep	21-Sep
2010	91.78182	4.5	10.59388	26-Sep	6-Sep	21-Sep
2011	92.52727	5.5	9.812484	5-Sep	5-Sep	22-Sep
2012	91.52727	0.5	15.40405	7-Aug	3-Sep	21-Sep
2013	84.56364	2	16.17368	24-Sep	2-Sep	21-Sep
2014	87.32727	4	10.7091	24-Sep	1-Sep	21-Sep
2015	90.30909	5	9.537168	13-Sep	7-Sep	22-Sep
					-	-
MEAN	**88.88455**	**2.75**	**12.5081**	**2-Sep**	-	-
MODE					**1-Sep**	**21-Sep**

EOS stands for "End of Summer".

If you're wondering what LD and SS_{end} are, these are the Labor Day and Summer Solstice end dates respectively.

The average End of Summer for all twenty CUSUM's ended up being Sep 2. This is very close to the mode for Labor Day which ended up being Sep 1. I was surprised to see this result.

To answer the next question about whether Atlanta's summers are getting warmer, I tried plotting all twenty End of Summer values I found over time. I ended up with this:

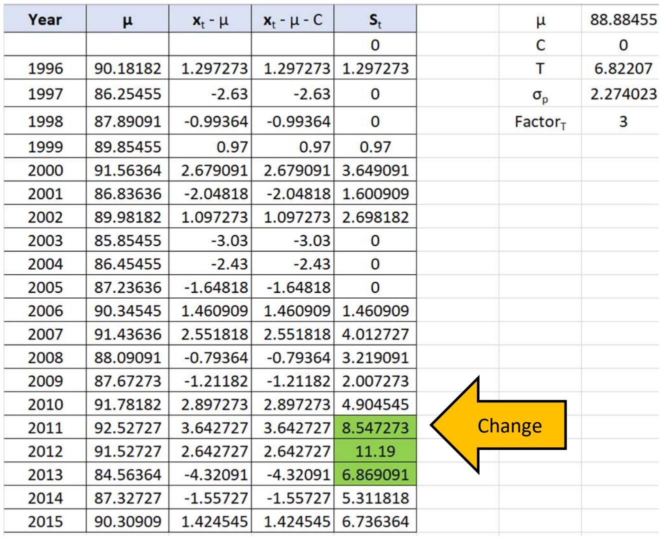


It is not very convincing. It does indicate an upward trend, the blue dotted line, but I can't tell based on this when that occurs definitively.

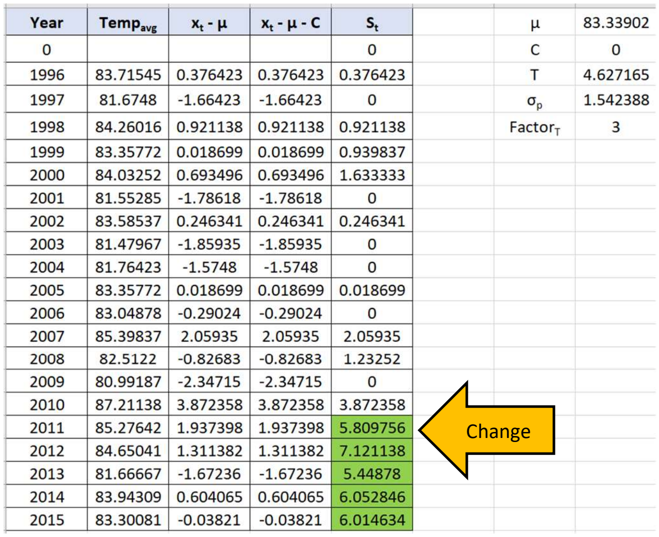
This may be due to the very random and unpredictable nature of the weather.

I then performed CUSUM analysis using the same basic steps I outlined above for all twenty years on **both** each years' average temp from Jul 1 to Aug 24 (µ) <u>and</u> the average temperature of all days from Jul 1 to Oct 31.

The results are below, the table on the left is the CUSUM for µ, and the table on the right is the CUSUM for average temperature:



Year	µ	x_t - µ	x_t - µ - C	S_t		µ	88.88455
				0		C	0
1996	90.18182	1.297273	1.297273	1.297273		T	6.82207
1997	86.25455	-2.63	-2.63	0		σ_p	2.274023
1998	87.89091	-0.99364	-0.99364	0		Factor_T	3
1999	89.85455	0.97	0.97	0.97			
2000	91.56364	2.679091	2.679091	3.649091			
2001	86.83636	-2.04818	-2.04818	1.600909			
2002	89.98182	1.097273	1.097273	2.698182			
2003	85.85455	-3.03	-3.03	0			
2004	86.45455	-2.43	-2.43	0			
2005	87.23636	-1.64818	-1.64818	0			
2006	90.34545	1.460909	1.460909	1.460909			
2007	91.43636	2.551818	2.551818	4.012727			
2008	88.09091	-0.79364	-0.79364	3.219091			
2009	87.67273	-1.21182	-1.21182	2.007273			
2010	91.78182	2.897273	2.897273	4.904545			
2011	92.52727	3.642727	3.642727	8.547273			
2012	91.52727	2.642727	2.642727	11.19			
2013	84.56364	-4.32091	-4.32091	6.869091			
2014	87.32727	-1.55727	-1.55727	5.311818			
2015	90.30909	1.424545	1.424545	6.736364			



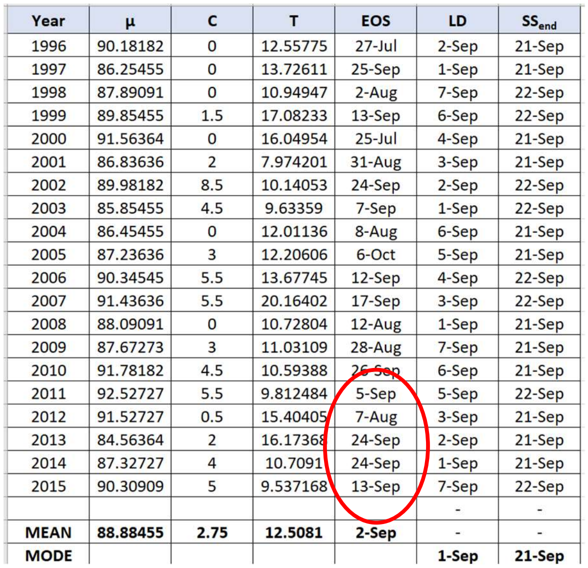
Year	Temp_{avg}	x_t - µ	x_t - µ - C	S_t		µ	83.33902
0				0		C	0
1996	83.71545	0.376423	0.376423	0.376423		T	4.627165
1997	81.6748	-1.66423	-1.66423	0		σ_p	1.542388
1998	84.26016	0.921138	0.921138	0.921138		Factor_T	3
1999	83.35772	0.018699	0.018699	0.939837			
2000	84.03252	0.693496	0.693496	1.633333			
2001	81.55285	-1.78618	-1.78618	0			
2002	83.58537	0.246341	0.246341	0.246341			
2003	81.47967	-1.85935	-1.85935	0			
2004	81.76423	-1.5748	-1.5748	0			
2005	83.35772	0.018699	0.018699	0.018699			
2006	83.04878	-0.29024	-0.29024	0			
2007	85.39837	2.05935	2.05935	2.05935			
2008	82.5122	-0.82683	-0.82683	1.23252			
2009	80.99187	-2.34715	-2.34715	0			
2010	87.21138	3.872358	3.872358	3.872358			
2011	85.27642	1.937398	1.937398	5.809756			
2012	84.65041	1.311382	1.311382	7.121138			
2013	81.66667	-1.67236	-1.67236	5.44878			
2014	83.94309	0.604065	0.604065	6.052846			
2015	83.30081	-0.03821	-0.03821	6.014634			

Note, that in these particular CUSUM analyses, I made one small change. Since I was trying to detect an increasing deviation from normal, the S_t calculation now subtract the mean from x_t. Also, I used the entire population to calculate the mean, μ, and the standard deviation calculation.

I tried plotting both the sets of data to see if I could establish a period of non-changing values, but I could not. The variation in values was to great and looked like the End of Summer plot above.

These plot seem to suggest that 2011 was possibly a year where the climate started to become warmer than in previous years.

Examining the unofficial end of summer values I previously calculated, the last five year from 2011-2015 all have unofficial end of summers in Sep for 4 out of the 5 years, which seems to suggest that summer lasted longer in those years as compared to the first five years for example.



Year	μ	C	T	EOS	LD	SS_{end}
1996	90.18182	0	12.55775	27-Jul	2-Sep	21-Sep
1997	86.25455	0	13.72611	25-Sep	1-Sep	21-Sep
1998	87.89091	0	10.94947	2-Aug	7-Sep	22-Sep
1999	89.85455	1.5	17.08233	13-Sep	6-Sep	22-Sep
2000	91.56364	0	16.04954	25-Jul	4-Sep	21-Sep
2001	86.83636	2	7.974201	31-Aug	3-Sep	21-Sep
2002	89.98182	8.5	10.14053	24-Sep	2-Sep	22-Sep
2003	85.85455	4.5	9.63359	7-Sep	1-Sep	22-Sep
2004	86.45455	0	12.01136	8-Aug	6-Sep	21-Sep
2005	87.23636	3	12.20606	6-Oct	5-Sep	21-Sep
2006	90.34545	5.5	13.67745	12-Sep	4-Sep	22-Sep
2007	91.43636	5.5	20.16402	17-Sep	3-Sep	22-Sep
2008	88.09091	0	10.72804	12-Aug	1-Sep	21-Sep
2009	87.67273	3	11.03109	28-Aug	7-Sep	21-Sep
2010	91.78182	4.5	10.59388	26-Sep	6-Sep	21-Sep
2011	92.52727	5.5	9.812484	5-Sep	5-Sep	22-Sep
2012	91.52727	0.5	15.40405	7-Aug	3-Sep	21-Sep
2013	84.56364	2	16.17368	24-Sep	2-Sep	21-Sep
2014	87.32727	4	10.7091	24-Sep	1-Sep	21-Sep
2015	90.30909	5	9.537168	13-Sep	7-Sep	22-Sep
					-	-
MEAN	**88.88455**	**2.75**	**12.5081**	**2-Sep**	-	-
MODE					**1-Sep**	**21-Sep**

Conclusion

The unofficial end of summer based on daily high temperature is **Sep 2**.

There appeared to be an upward trend in length of summer in Atlanta based on the computed unofficial end of summer.

Beginning in **2011**, the average temperature deviated by at least 3 deviations from the mean for the population, indicating a possible warming trend beginning around this time.

References

timeanddate. (n.d.). Labor Day Observances. https://www.timeanddate.com/holidays/us/labor-day timeanddate. (n.d.). Solstices & Equinoxes for Atlanta. https://www.timeanddate.com/calendar/seasons.html?year=2000&n=25

Hsu, P. (2022, September 12 Day). ISYE6501 Office Hours - September 12th [Recorded office hours]. Media Space. https://mediaspace.gatech.edu/media/ISYE6501+Office+Hours+-+September+12th+%28OMS+Su22%29/1_k3x5yn3z