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.

Using Shiny, Outliers, GGplot2, and some assistance from GPT O1 and the outliers documentation, I wrote a small webapp that displays both a single outlier and two outlier Grubbs test. Based on the output of my program, I can not say with certainty that any values were outliers beyond a reasonable doubt.

The Grubbs test supposes different alternative hypotheses that are accepted in my program if their associated p-values are below 0.05 (or are at 95% degree of confidence). While the single test was close, holding a p-value of 0.078875 when suggesting that 1993 was an outlier, it did not quite cross the threshold.

The two-tailed Grubbs test possessed a p-value of 1 because the alternative hypothesis '342 and 1993 are outliers' is blatantly false. Value 342 is not an outlier.

Directly below this section are attached images of my webpage. You may notice that the graphs have different presentations on the X axis. This is arbitrary and solely done such that all datapoints are viewable.

Code is present in R file: "5dot1Crime.R". Make sure the data is present in the proper subdirectory folder labeled 'data' before attempting to run this program on a local machine.

Crime Data Outlier Detection with Grubbs' Test

Select Grubbs' Test Type:

Single Outlier

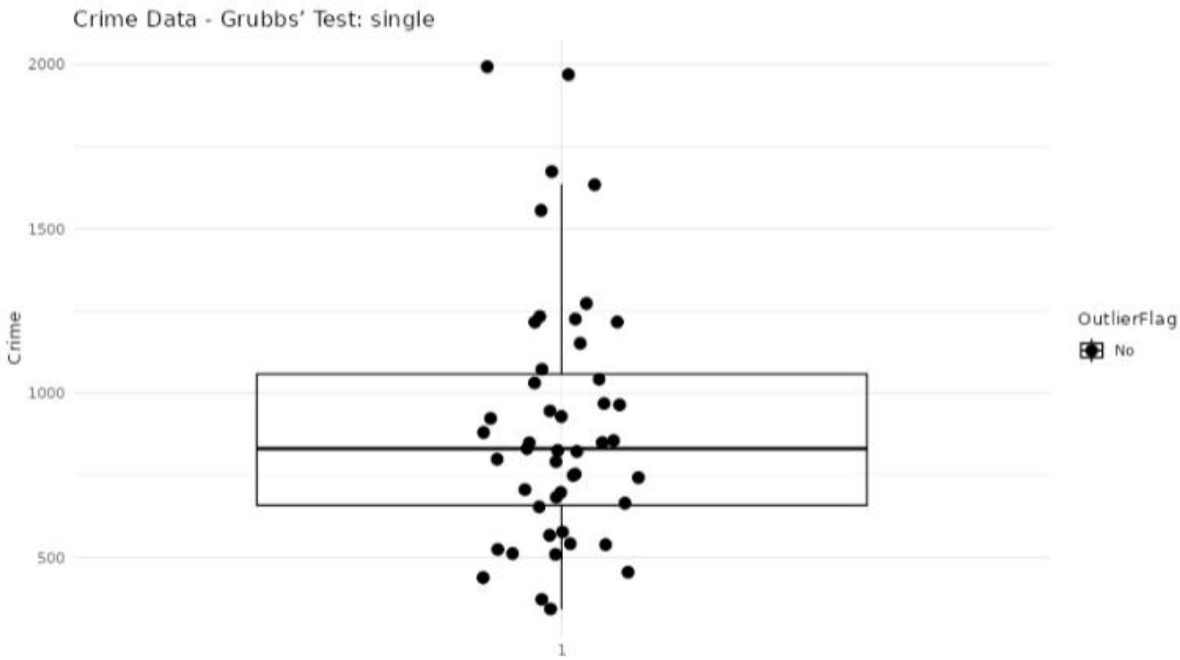


Figure Caption:

Points are flagged as 'Yes' (red) if $p < 0.05$ from Grubbs' test. Here, G = 2.8129, p-value = 0.078875, alt. hypothesis: 'highest value 1993 is an outlier'. All other points are black.

Grubbs' Test Output:

Crime Data Outlier Detection with Grubbs' Test

Select Grubbs' Test Type:

Two Outliers

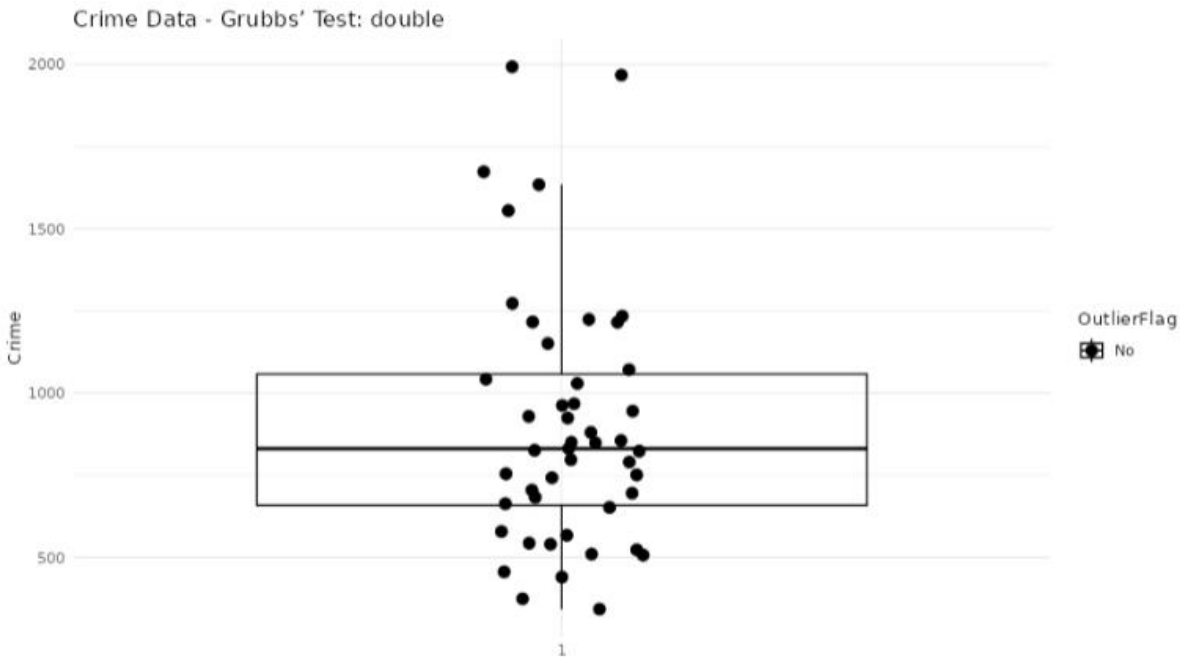


Figure Caption:

Points are flagged as 'Yes' (red) if p < 0.05 from Grubbs' test. Here, G = 4.2688, p-value = 1, alt. hypothesis: '342 and 1993 are outliers'. All other points are black.

Grubbs' Test Output:

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?

I'm interested in eventually becoming involved with quantitative trading which absolutely utilizes change detection models to determine whether a stock is volatile or not. Due to the variety of trading options available, traders can potentially generate a profit if they are better able to predict different features of various investment opportunities. Scanning through a stock's current price, comparing those values to past records, and identifying outliers may prompt further investigation which can be exploited for profit.

If I were to apply the CUSUM technique, the threshold would likely be set at a range slightly beyond any given stock's usual fluctuation, while the C value would be set relatively high. It would be important to minimize false positives, especially if I were observing a volatile stock.

Using these techniques may help me to identify unique investment opportunities. Combining these tools with semantic analysis of newly published legal documents, news articles, and other predictive modeling techniques may reveal key information that helps me to get ahead of the market.

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.
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 solution to question 6.2 is a little ghoulish overkill.

Using the same tech stack, I constructed a web app that displays answers to the questions asked depending on a selected year with the assistance of three graphs and the CUSUM method. In short, the end of summer can be detected by finding the maximum CUSUM value. This is done iteratively every year selected.

For question 2, I concatenated all the daily data from all years into a single time series and utilized the CUSUM method to search for significant changes in slope that would indicate a climate shift. This indicated that the maximal shift occurred between 2009-10-31 and 2012-09-16. After observing the chart, there certainly seems to be an upwards trend of CUSUM values throughout this period, indicating that warming is occurring.

To get larger scope on how climate was trending, I implemented a linear regression model that suggests that temperature is increasing at a rate of ~0.053 degrees/year. The P value of this line is 0.40027, which is significantly above the acceptable threshold of under 0.05.

Therefore even with the provided analysis, it's difficult to say whether or not Atlanta's summer climate has gotten warmer throughout these years. More definitively than anything, I think I can suggest that 2010 was freakishly hot.

Attached below this writeup section is a screenshot of my webapp in 1998, 2010, and 2014. I'm hoping this will prove beyond a reasonable doubt that my code accomplishes the provided task without requiring you to run it on your local machine. If you plan to, make sure that the data is in the same subdirectory labeled 'data' mentioned earlier.

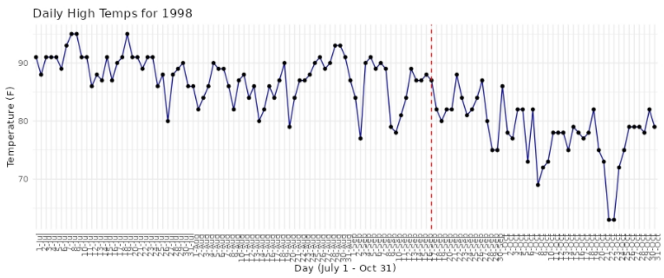
Code is present in R file '6dot2Temps.R'.

Atlanta Temperature CUSUM Analysis (1996-2015)

Select a Year:

1998

Reads daily high temperatures from data/temps.txt for July 1 - Oct 31, 1996-2015. A per-year CUSUM identifies the day each year cools off. A global CUSUM checks long-term warming.



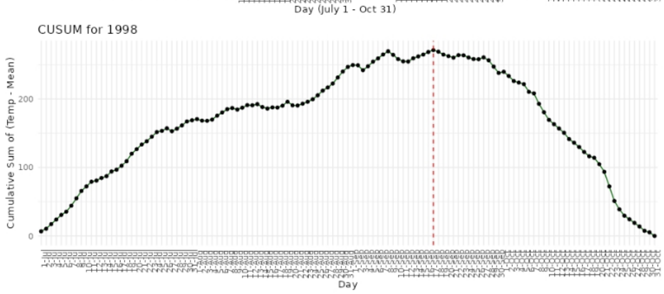
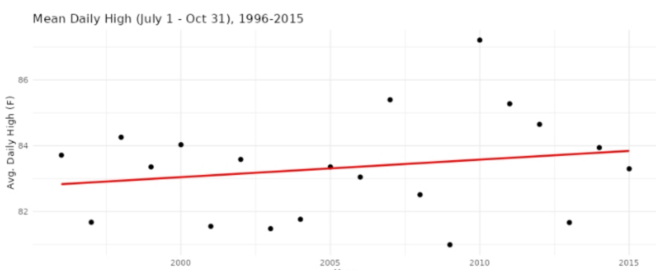


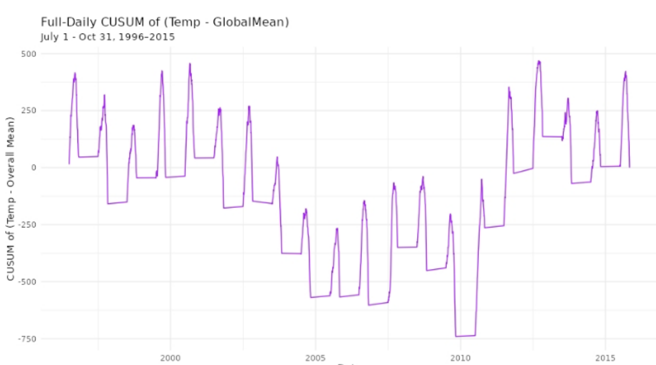
Figure Caption (Selected Year):

For 1998, the average July-Oct daily high was ~84.3 F. A per-year CUSUM was cumsum(Temp - 84.3). The maximum CUSUM (red dashed line) indicates 'unofficial summer ends' after that day. Detected end-of-summer: **17-Sep**.



Overall Trend Commentary (Linear Regression):

Linear regression slope: **0.053 deg/year**, p-value = 0.40027. If positive and significant (p < 0.05), suggests warming over 1996–2015.



Full-Daily CUSUM Commentary:

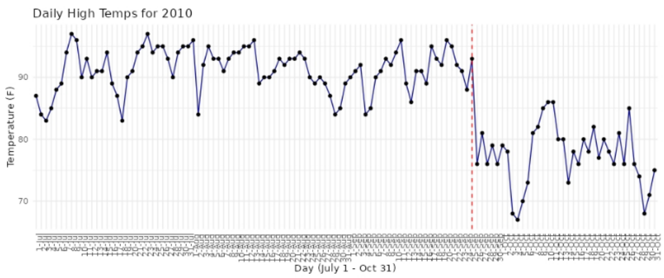
This chart shows cumsum(Temp - 83.34) across all days (1996–2015). Max CUSUM near 2012-09-16, min near 2009-10-31. Significant upward drift indicates warming.

Atlanta Temperature CUSUM Analysis (1996-2015)

Select a Year:

2010

Reads daily high temperatures from data/temps.txt for July 1 - Oct 31, 1996-2015. A per-year CUSUM identifies the day each year cools off. A global CUSUM checks long-term warming.



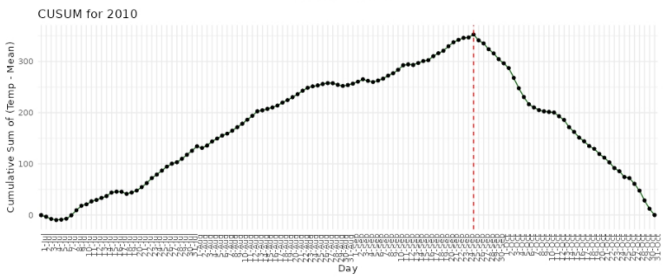
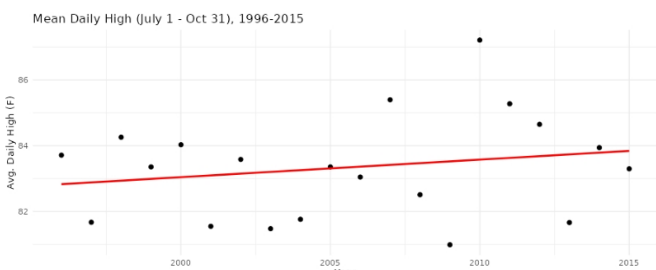


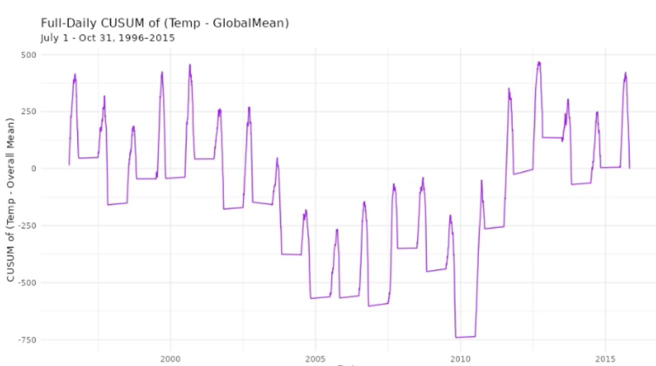
Figure Caption (Selected Year):

For 2010, the average July-Oct daily high was ~87.2 F. A per-year CUSUM was cumsum(Temp - 87.2). The maximum CUSUM (red dashed line) indicates 'unofficial summer ends' after that day. Detected end-of-summer: **25-Sep**.



Overall Trend Commentary (Linear Regression):

Linear regression slope: **0.053 deg/year**, p-value = 0.40027. If positive and significant ($p < 0.05$), suggests warming over 1996–2015.



Full-Daily CUSUM Commentary:

This chart shows cumsum(Temp - 83.34) across all days (1996–2015). Max CUSUM near 2012-09-16, min near 2009-10-31. Significant upward drift indicates warming.

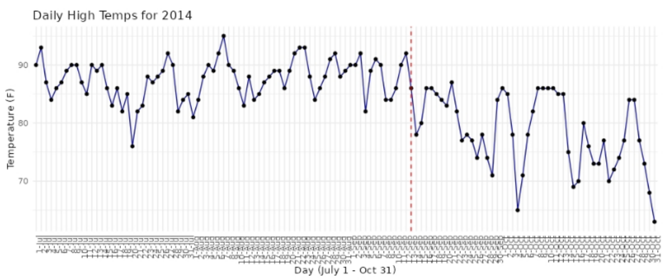
[redacted]

Atlanta Temperature CUSUM Analysis (1996-2015)

Select a Year:

2014

Reads daily high temperatures from data/temps.txt for July 1 - Oct 31, 1996-2015. A per-year CUSUM identifies the day each year cools off. A global CUSUM checks long-term warming.



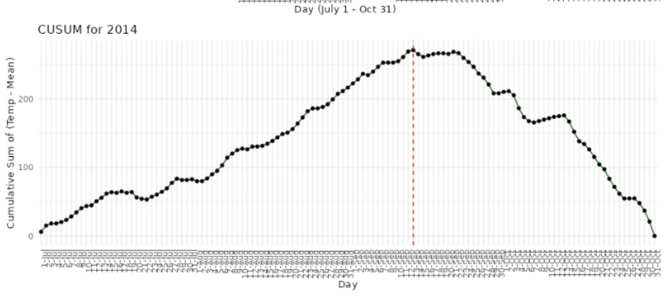
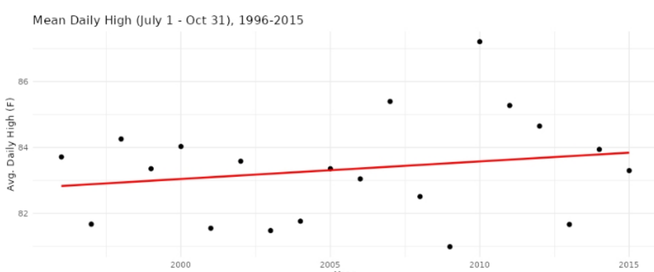


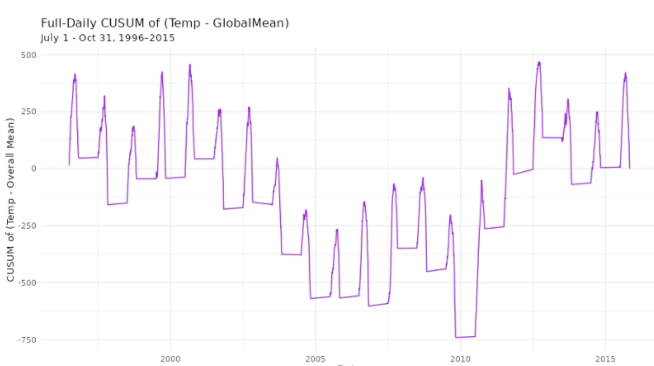
Figure Caption (Selected Year):

For 2014, the average July-Oct daily high was ~83.9 F. A per-year CUSUM was cumsum(Temp - 83.9). The maximum CUSUM (red dashed line) indicates 'unofficial summer ends' after that day. Detected end-of-summer: **13-Sep**.



Overall Trend Commentary (Linear Regression):

Linear regression slope: **0.053 deg/year**, p-value = 0.40027. If positive and significant ($p < 0.05$), suggests warming over 1996–2015.



Full-Daily CUSUM Commentary:

This chart shows cumsum(Temp - 83.34) across all days (1996–2015). Max CUSUM near 2012-09-16, min near 2009-10-31. Significant upward drift indicates warming.

Sources:

Outliers Documentation

GPT O1 Model for Code Refinement