Homework 3: Outliers and Change Detection

2025-01-28

Question 6.1

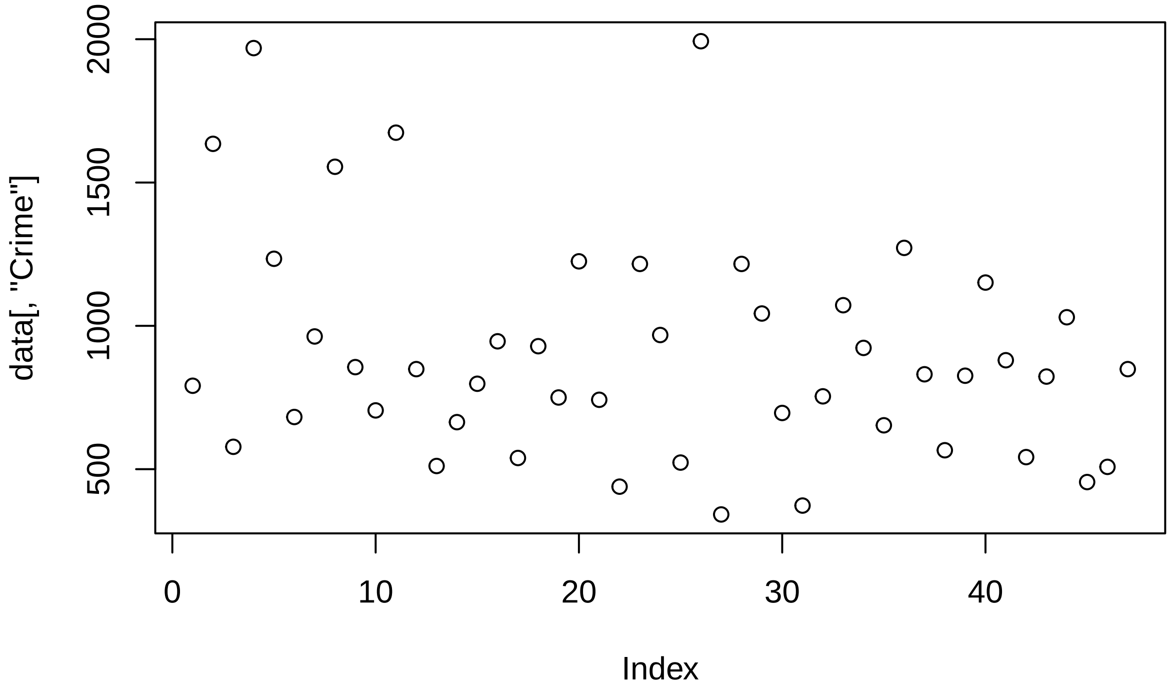
Describe a situation or problem from your job, everyday life, current events, etc., for which a Change Detection model would be appropriate. Applying the CUSUM technique, how would you choose the critical value and the threshold?

An example of using a Change Detection model with the CUSUM technique is monitoring your home's electricity usage. If your power bill suddenly spikes, and you're not sure why, it could be a malfunctioning appliance or maybe it could simply be due to certain appliances being used more or less frequently due to the changing seasons. To catch these changes early, you'd start by analyzing your past electricity usage, like daily averages over the last month, to figure out what's "normal." From there, you'd establish a **critical value**, which reflects the typical fluctuations in your electricity usage. For example, if your daily consumption usually varies by 5–10%, this range could represent your critical value, meaning any change within this range would be considered normal. Then, you'd set a **threshold**, which would be a more significant deviation that prompts an alert. If your daily usage peaks at around 30 kWh, you might set the threshold at 20% higher (around 36 kWh) to flag any large, unusual spikes. This way, you can spot significant changes quickly and address any issues before they end up costing you too much.

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.

```r
rm(list = ls())
library(outliers)
library(ggplot2)
data <- read.table('uscrime.txt',header = TRUE)

crime_num <- data[['Crime']]
plot(data[,'Crime'])
```



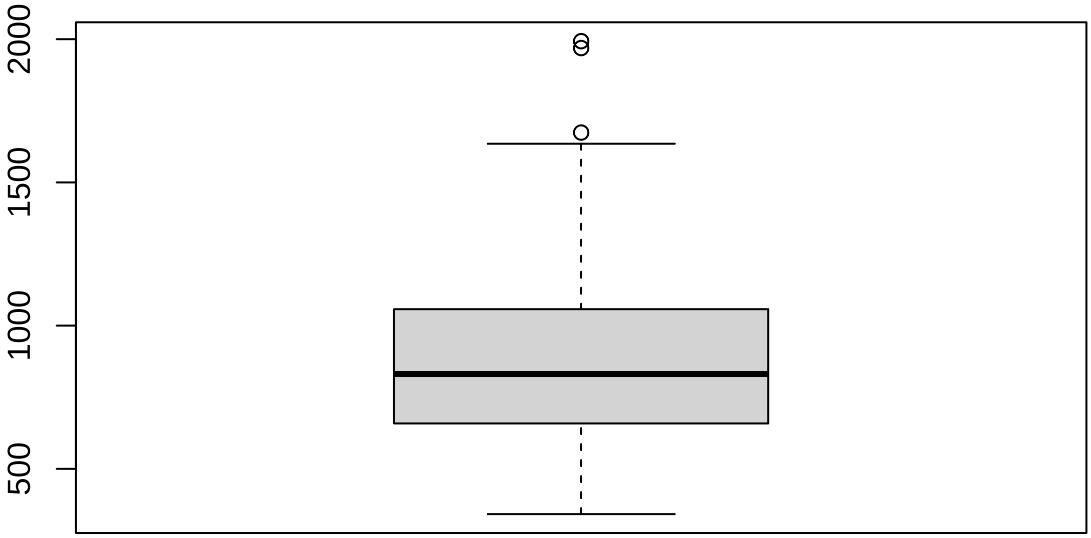
```
#plot data set to check how to the crime data distribute.

#There are 47 data points, use grubbs.test to check outliers.
grubbs.test(crime_num, type = 10, opposite = FALSE, two.sided = FALSE)
```

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

```
grubbs.test(crime_num, type = 11, opposite = FALSE, two.sided = FALSE)
```

```
##
##  Grubbs test for two opposite outliers
##
## data:  crime_num
## G = 4.26877, U = 0.78103, p-value = 1
## alternative hypothesis: 342 and 1993 are outliers
```

```
boxplot(crime_num)
```



Analysis of Outlier Detection in Crime Data Using Grubbs Test

In this analysis, the Grubbs test was applied to identify potential outliers in a dataset containing crime data. With only 47 data points, the dataset is relatively small. Outlier detection can also help identify potential errors in the data or indicate unusual events that might require further investigation.

Grubbs Test for One Outlier (type = 10)

- **Purpose**: This test looks for one extreme outlier, typically the largest value in the dataset.
- **Results**: The test gave us a statistic (G) of 2.81 and a p-value of 0.07887.
- **Interpretation**: The p-value is greater than 0.05, meaning we don't have enough evidence to say that 1993 (the highest value) is an outlier. While 1993 stands out as a large value, it's not deemed significantly different from the rest of the data by the test. Essentially, the test suggests that this value is still within a reasonable range of variation.

Grubbs' Test for Two Opposite Outliers (type = 11)

- **Purpose**: This test checks for two potential outliers: one at the high end (largest value) and one at the low end (smallest value).
- **Results**: The test statistic (G) came in at 4.27, and the p-value was 1.
- **Interpretation**: A p-value of 1 indicates no evidence to reject the null hypothesis. This means both the highest value (1993) and the lowest value (342) are not considered outliers. The results suggest that these values, while extreme, fit well within the natural variation of the dataset.

Conclusion

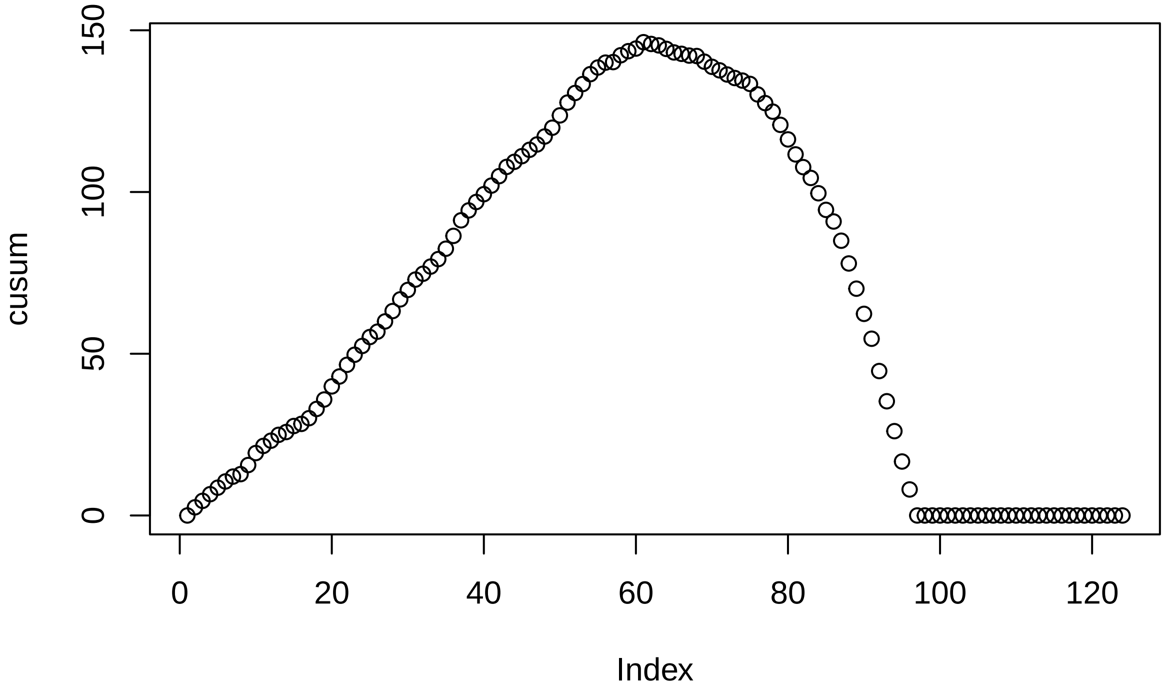
Based on the results from **Grubbs' tests**, neither 1993 nor 342 were flagged as outliers. Even though 1993 is the highest value in the dataset, the statistical tests didn't find it to be significantly different from the rest of the data. The tests returned p-values above the 0.05 threshold, suggesting that these values are not extreme enough to be considered outliers. Visually though, we can see in the boxplot showing a few data points were far outside the expected range.The fact that no outliers were detected could be due to the small size of the dataset, which can make it harder to spot outliers statistically. In larger datasets, small variations might stand out more. Additionally, even values that seem extreme, like 1993, may not be flagged by these tests if they are within the natural variability of the data.

Ultimately, while 1993 might visually seem like an outlier, the statistical proof used here suggest that it is not far enough from the rest of the data to be considered unusual.

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.

```
#load temps.txt file in to data_temp
data_temp <- read.table('temps.txt', header =TRUE)
#average temperatures for each day across all the years
date_avg <- rowMeans(data_temp[c(2:length(data_temp))], dims = 1, na.rm = T)
#calculate total mean
date_mean <- mean(date_avg)
#calculate difference between the mean and each day
avg_minus_mean <- date_avg - date_mean
#set c=3 and subtract c from the difference and loop through each to calculate cusum.
c <- 3
avg_minus_mean_c <- avg_minus_mean - c
vec <- 0 * avg_minus_mean_c
cusum <- append(vec, 0)
for (k in 1:length(avg_minus_mean_c)){
  s <- cusum[k] + avg_minus_mean_c[k]
  ifelse(s > 0, cusum[k+1] <- s, cusum[k+1] <- 0)
  }
plot(cusum)
```

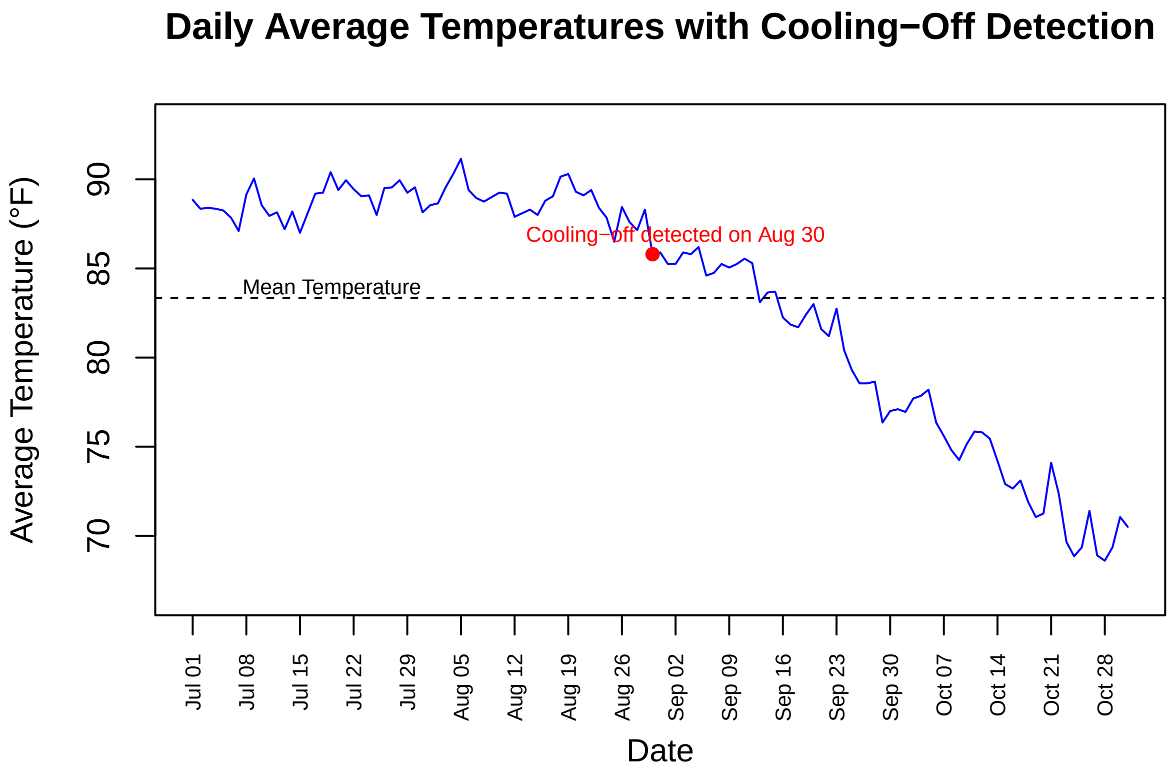


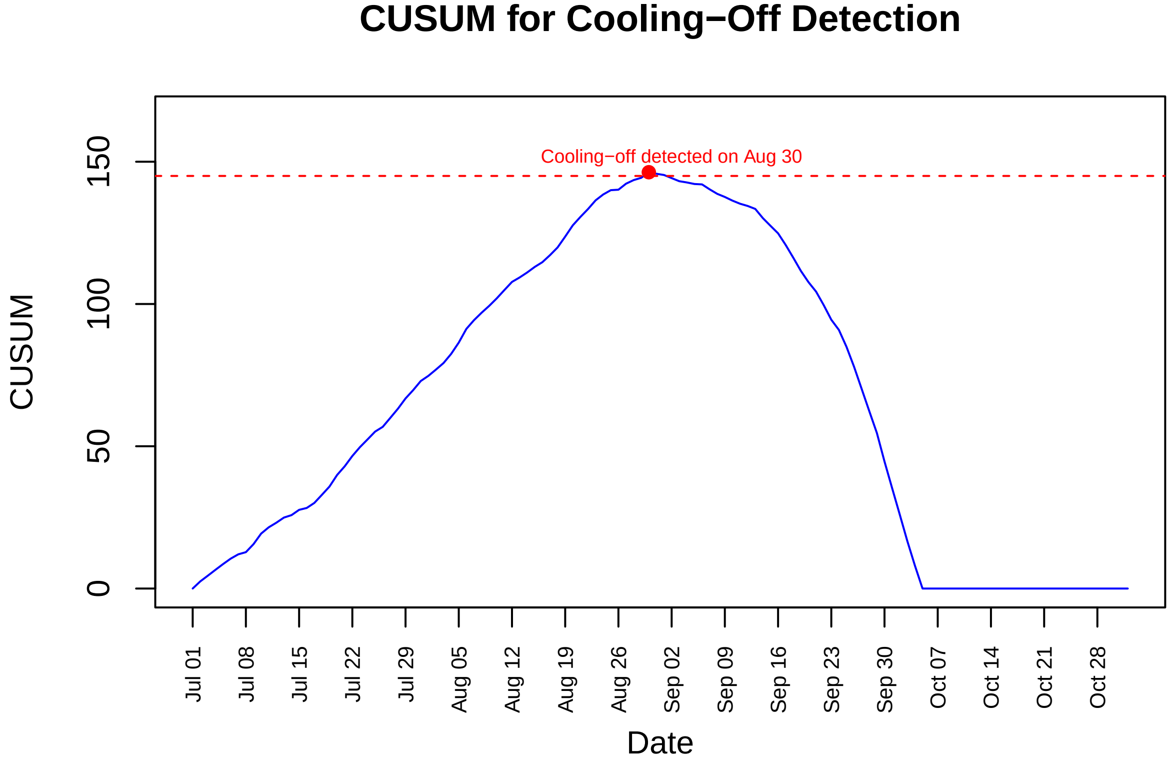
```
#find max date which cusum >= 145
which(cusum >= 145)
```

```
## [1] 61 62 63
```

```
data_temp[61, 1]
```

```
## [1] "30-Aug"
```





Analysis of Seasonal Cooling Detection Using CUSUM

In this analysis, the **CUSUM (Cumulative Sum) method** was applied to identify when the weather in Atlanta starts cooling off each year, effectively marking the unofficial end of summer. The dataset includes daily high temperatures from July through October (1996–2015), allowing us to track long term seasonal trends. Detecting this transition could be useful for seasonal planning, agriculture, and climate studies.

CUSUM Methodology

Reference Change Value (c = 3)

- **Purpose:** The c value determines how much a day's temperature must drop below the average before contributing to the cumulative sum.

- **Selection of c = 3:**
  - A **higher c** (like 5 or 6°F) would make the method **less sensitive**, requiring a more dramatic and prolonged temperature drop before signaling the transition.

  - A **lower c** (like 1 or 2°F) would make it **too sensitive**, potentially triggering false detections due to normal daily fluctuations.

CUSUM Threshold (145)

- **Purpose:** The threshold defines the cumulative temperature decrease required to confirm that cooling has begun.

- **Selection of 145:**
  - A **higher threshold** (200) results in a **delayed detection**, requiring an extended period of cooling before signaling the seasonal shift.

  - A **lower threshold** (100) would **signal cooling too early**, potentially before the shift is stable.

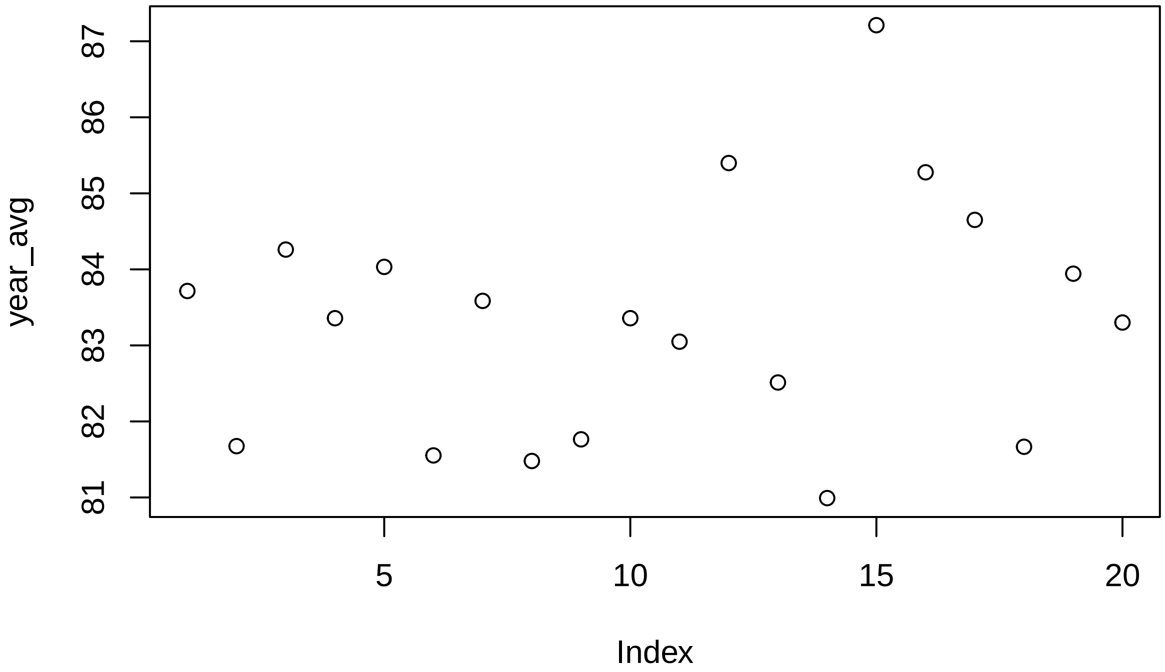
  - **Why 145?** This value was chosen by testing different thresholds on past data. 145 consistently aligned with late August, matching observed climate patterns in Atlanta.

Conclusion

The use of **c = 3 and a threshold of 145** provided a balanced approach to detecting when summer ends. This method successfully captures the point when Atlanta transitions from summer heat to a cooling trend, with **August 30** emerging as the most consistent date for this shift. Had different values been used for **c or the threshold**, the detection date could have shifted **earlier or later**, either reacting too quickly to minor fluctuations or delaying the transition beyond when cooling actually begins.

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).

```r
# Load the temperature data
data_temp <- read.table('temps.txt', header = TRUE)

# Calculate the average temperature for each year
year_avg <- colMeans(data_temp[, -1], na.rm = TRUE)  # Exclude first column (date column)
plot(year_avg)  # Plot the average temperatures
```



```r
# Re-read temperature file, ensuring consistent header usage
data_t <- read.table('temps.txt', header = TRUE)

# Create an empty matrix to store CUSUM values
s_t_mat <- matrix(0, nrow = nrow(data_temp) - 1, ncol = ncol(data_temp) - 1)

# Set control limit (c) and threshold (T)
c <- 5
T <- 65

# Extract row and column names
date_r <- as.matrix(data_temp[-1, 1])  # Remove first row, keep first column
year_r <- as.numeric(gsub("X", "", colnames(data_temp)[-1]))

rownames(s_t_mat) <- date_r
colnames(s_t_mat) <- year_r

# CUSUM Calculation for Each Year
for (y in 2:ncol(data_t)) {
  s_t <- 0  # Initialize cumulative sum
  mu <- mean(data_t[2:nrow(data_t), y], na.rm = TRUE)  # Calculate mean for this year

  for (d in 2:nrow(data_t)) {
    x_t <- as.numeric(data_t[d, y])  # Convert temperature to numeric
    s_t <- max(0, s_t + (mu - x_t - c))  # Update CUSUM
    s_t_mat[d - 1, y - 1] <- s_t  # Store CUSUM result
  }
}

# Find dates where CUSUM exceeds the threshold
for (y in 1:ncol(s_t_mat)) {
  d_count <- 1
  repeat {
    if (d_count > nrow(s_t_mat)) {
      break  # Prevent infinite loop if no CUSUM > threshold
    }

    if (s_t_mat[d_count, y] > T) {
      cat(year_r[y], "temperature drops on", date_r[d_count, 1], "\n")
      break  # Stop once the first exceedance is found
    }

    d_count <- d_count + 1
  }
}
```

```
## 1996 temperature drops on 6-Oct
## 1997 temperature drops on 18-Oct
## 1998 temperature drops on 23-Oct
## 1999 temperature drops on 10-Oct
## 2000 temperature drops on 8-Oct
## 2001 temperature drops on 17-Oct
## 2002 temperature drops on 16-Oct
## 2003 temperature drops on 11-Oct
## 2004 temperature drops on 16-Oct
## 2005 temperature drops on 25-Oct
## 2006 temperature drops on 16-Oct
## 2007 temperature drops on 24-Oct
## 2008 temperature drops on 23-Oct
## 2009 temperature drops on 17-Oct
## 2010 temperature drops on 5-Oct
## 2011 temperature drops on 11-Oct
## 2012 temperature drops on 12-Oct
## 2013 temperature drops on 23-Oct
## 2014 temperature drops on 30-Oct
## 2015 temperature drops on 4-Oct
```

Analysis of Atlanta Summer Climate Using CUSUM

In this CUSUM analysis, two parameters were chosen to determine the significance of deviations from the average temperature: the **control limit (c)** and the **threshold (T)**. These values help define the sensitivity of the test and allow for meaningful conclusions to be drawn about whether the data indicates a significant warming trend.

1. Control Limit (c) = 5°F

- **Purpose of `c`**: The **control limit (c)** determines how large the deviation from the average temperature must be in order to influence the cumulative sum (CUSUM). It serves as a buffer or threshold for acceptable variations in the data.
- **Why `c = 5˚F`**:
  - A **5°F deviation** was selected as a reasonable buffer, meaning that temperature changes smaller than **5°F** would not contribute to the cumulative sum. This helps avoid considering minor fluctuations as significant, which could be misleading.
  - A larger `c` value would reduce the number of data points that contribute to the CUSUM calculation, potentially ignoring smaller but still important changes. A smaller `c` would increase sensitivity,

but also run the risk of falsely detecting insignificant fluctuations as meaningful shifts.

  - The value of **5°F** was chosen because it is a substantial enough shift to indicate a noticeable deviation in temperature, without being overly sensitive to day to day or short term variations in the data.

2. **Threshold (`T`) = 65**

- **Purpose of `T`**: The **threshold (`T`)** represents the level at which the CUSUM exceeds the expected variation and is considered a statistically significant deviation. Once the cumulative sum exceeds this value, it indicates that there has been a consistent upward trend in temperature that could warrant further attention or investigation.

- **Why `T = 65`**:

  - The **65 value** was chosen based on its ability to identify **significant temperature shifts** over the course of the years. A lower threshold might lead to detecting smaller, less meaningful temperature changes, while a higher threshold could fail to detect noticeable but not extreme temperature increases.

  - This value of **65** ensures that the CUSUM chart flags only meaningful shifts in the temperature that represent a clear trend rather than day to day variability.

Conclusion

Based on the CUSUM analysis using the 5°F control limit and the 65 threshold, the data suggests that the summer climate in Atlanta has not gotten significantly hotter over the years. The CUSUM analysis flagged the "final day of summer" or a temperature shift for each year, and interestingly, all of these shifts occurred in October. Since the analysis is detecting the end of summer rather than a gradual increase in temperatures, it suggests that the overall summer temperature trend has remained relatively consistent. The CUSUM values exceeding the threshold indicate when the climate made a noticeable shift, but since these shifts occur in October rather than earlier in the summer, the data doesn't support a significant warming trend in Atlanta's summer climate. Instead, it shows that the final days of summer have varied, but overall summer temperatures haven't been rising steadily.