Homework3

Question 5.1

Outlier (uscrime.txt):

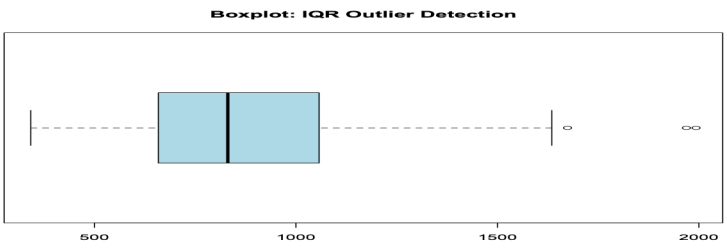
Methodology:

I took a structured approach by progressing from robust descriptive statistics to formal hypothesis testing:

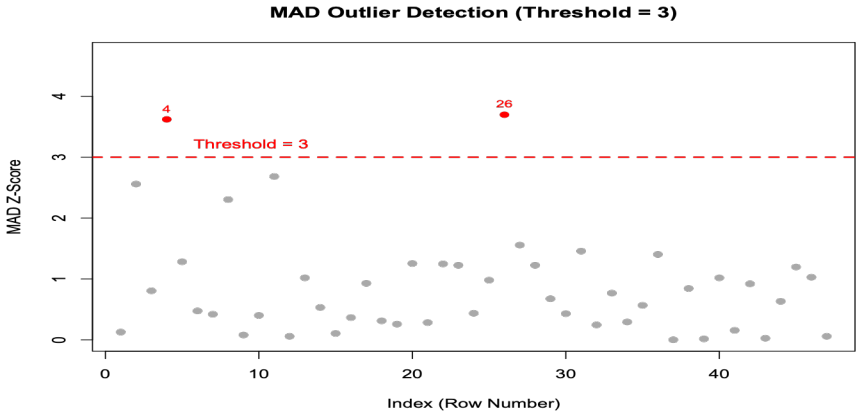
- **Exploratory Heuristics (IQR & MAD):** These were used to establish a baseline. The IQR(Interquartile range) method identifies the values falling outside the "whiskers" of a boxplot (Q1 - 1.5 * IQR) to (Q3 + 1.5 * IQR). The MAD (Median absolute deviation) provides a more robust center, using a threshold of 3 units from the median to flag potential anomalies without the skewing influence of outliers themselves.
- **Normality assessment:** Since many statistical tests (like Grubbs') assume a normal distribution, we conducted a Shapiro-Wilk test. We supplemented this with a Q-Q Plot and a 95% Confidence Envelope to visually identify which data points were responsible for pushing the data away from normality.
- **Formal Statistical Testing:** * Iterative Grubbs' Test**:** I used this to identify and "peel" away single outliers one at a time. This allowed me to observe how the p-value changed as extreme values were removed.
    - **Rosner's Test:** Employed as the "Gold Standard" for this dataset. It is more sophisticated because it tests for multiple outliers (up to k = 5) simultaneously, preventing "masking" where two extreme values might hide one another.

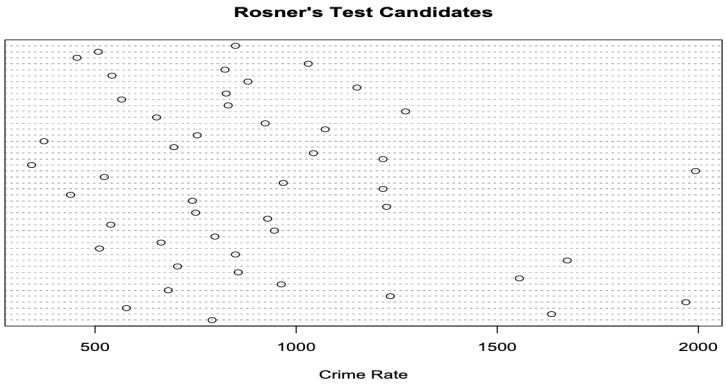
Results & Analysis:

- **Primary Outlier (Row 26):** Every test, from IQR to Rosner's, flagged **1993** as a significant outlier. In the Q-Q plot, it is the most extreme point, sitting well outside the confidence envelope.

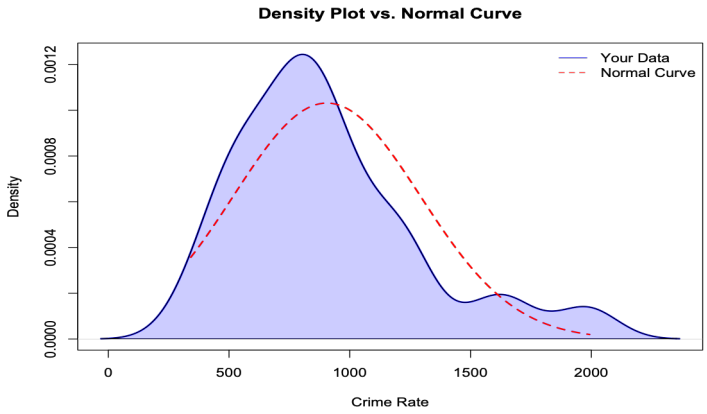


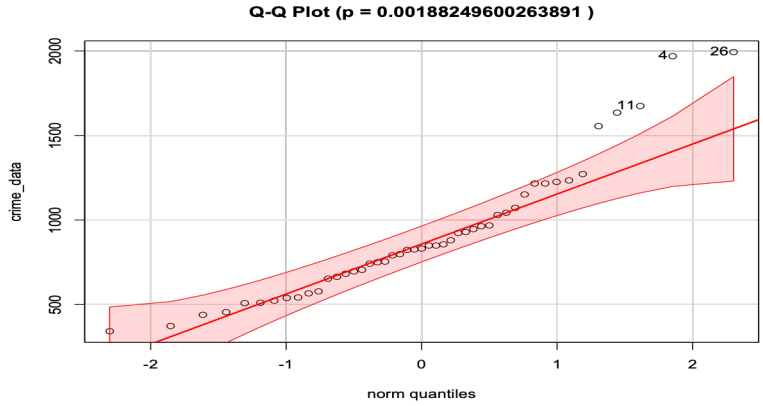
- **Secondary Candidates (Row 25 & 11):** Values **1969** and **1555** were frequently flagged by the more sensitive tests (MAD and Rosner's). These represent a "cluster" of high-crime observations that shift the mean significantly.





- **Normality Verdict:** The Shapiro-Wilk test p-value = 0.0018 ($p < 0.05$) indicates that the Crime column is **not** strictly normal. However, the Q-Q plot shows the "core" of the data follows the line well, suggesting the non-normality is caused by a few specific "contaminated" outliers rather than a different underlying distribution (like log-normal).





For the statistical significance test for normality in my code, I used 0.01 while typically p = 0.05 is a standard choice. Using 0.01 means that you are demanding 99% confidence before you declare the normality test as statistically insignificant. So, with 99% confidence we can say that the normality hypothesis can be rejected, and this means that we can reject the normality hypothesis at a lower confidence of 95% which corresponds to p = 0.05. In our Grubbs' test, the p-value for 1993 is approximately **0.078**. This means at a 95% confidence level, it *barely* misses the cut, but at a 90% level, it is a clear outlier.

Comparative Summary of Outlier Detection Results

Method	Flagged Values	Rationale / Result
IQR Method	1993, 1969, 1635	Points outside 1.5 × IQR upper fence.
MAD Method	1993, 1969, 1635	Values with a Z-score > 3 using robust median.
Shapiro-Wilk	Global Result	p = 0.0018 (Reject Normality at α=0.05).
Q-Q Plot (car)	Row 26, Row 25	Values significantly beyond 95% confidence envelope
Grubbs' (Iterative)	1993	Identified as the single most extreme point.
Rosner's Test	1993, 1969, 1635	Flagged up to k=3 as a cluster of high-end outliers.

The final recommendation is that given the consistent flagging of Row 26 (1993) across all three statistical frameworks, it is recommended to treat this point as an outlier. Further

analysis should determine if this represents a data entry error or a unique high-crime jurisdiction that requires its own specialized model.

Question 6.1

A real-life example in my field of chip design is **chip manufacturing** use case for CUSUM. In semiconductor fabrication, "yield" refers to the percentage of functional chips on a silicon wafer. Because a single wafer can contain thousands of chips and the manufacturing process involves hundreds of chemical and physical steps, even a tiny "drift" in a machine's calibration can cost millions of dollars in lost product. Here is how we could apply a Change Detection model to this specific problem.

The Problem: Yield Excursion in Photolithography

Imagine your "baseline" yield is **92%**. You are monitoring the daily output. You don't care about random daily bounces (e.g., 91.8% one day, 92.2% the next), but you are terrified of a **process shift** where a lens becomes slightly dirty, causing the yield to drop to **88%** and stay there.

Applying the CUSUM Technique: The CUSUM formula is:

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

*(Note: Since we are tracking a **drop** in yield, we look at how much the actual value falls below the target.)*

1. **Choosing the Critical Value (*C*):** The critical value acts as a "filter" for noise.

- **The Logic:** You want to detect a shift from 92% to 88%. The total shift size is 4%.
- **The Choice:** A standard rule of thumb is setting to half the shift you want to detect.
- **Application:** You would set C = 2% .
- This means that if a day's yield is 91% , the model sees (92-91-2) = -1. Since it's negative, the S_t stays at 0. The model ignores it as noise. However, if the yield drops to 89%, the model sees (92-89-2) = +1. The "cumulative sum" starts to build.

2. Choosing the Threshold (*T*): The threshold is your "Line in the Sand" for calling a technician to stop the machines.

- **The Logic:** You choose based on your tolerance for False Alarms vs. Late Detection.
- **Selection via Standard Deviation:** In chip labs, we calculate the standard deviation (σ) of the yield. T is typically set to 4σ or 5σ.
- **Business Context:** * If stopping the line is incredibly expensive (takes 24 hours to restart), you choose a **High** (e.g., 7σ). You want to be *sure* there is a problem.

- If a yield drop is catastrophic to your quarterly profits, you choose a **Low** (e.g., 3σ) to catch the drift as fast as possible, accepting that you might occasionally stop the line for no reason.

Why CUSUM is better than a Simple Limit

In chip manufacturing, a "Simple Limit" (like "Stop if yield < 87%") is dangerous. If the machine drifts and the yields stay at 87.5% for a month, a simple limit never triggers. CUSUM will catch this**.** Because 87.5% is consistently below the 92% target, the sum will keep growing day after day until it eventually crashes through the threshold (*T*), sounding the alarm.

Parameter	Value in Example	Goal
Target (μ0)	92%	The "Perfect" state.
Critical Value (C)	2%	Ignore small, random fluctuations.
Threshold (H)	5σ (e.g., 10)	Signal an "Excursion" (stop the line).

Question 6.2

The following analysis details the methodology, results, and conclusions derived from applying the CUSUM (Cumulative Sum) control chart technique to Atlanta temperature data from 1996 to 2015. The study is divided into two parts: detecting the seasonal transition from summer to fall (cooling trend) and evaluating long-term summer temperature increases (warming trend).

1. Cooling Detection: The "End of Unofficial Summer"

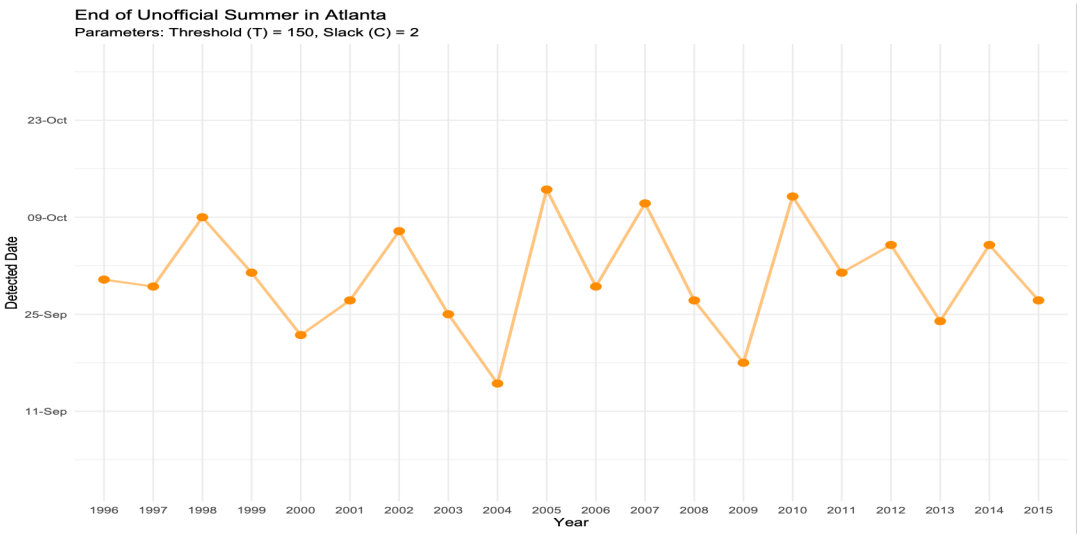
Methodology:

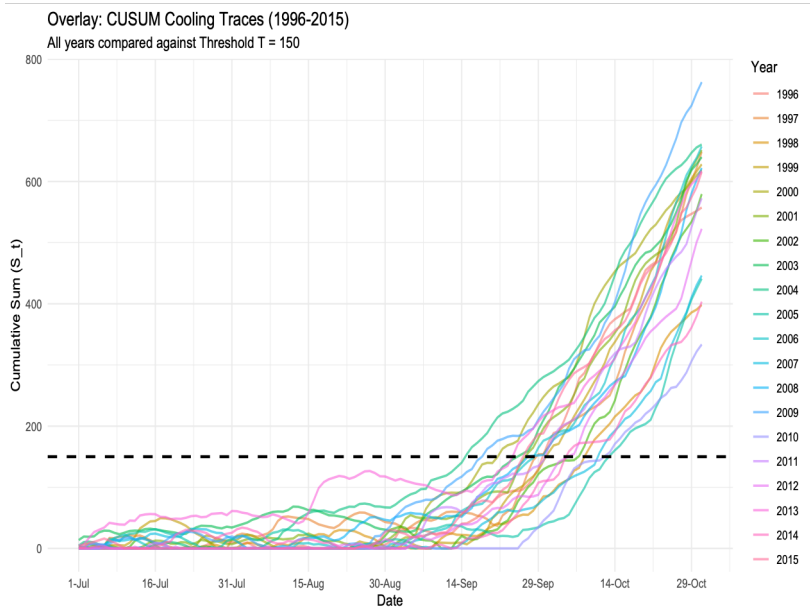
The goal was to identify the specific date each year when summer temperatures definitively transitioned to a cooler autumn pattern.

- **Baseline Establishment:** A "Global Summer Baseline" was calculated as the mean temperature of all days in July and August across the entire 20-year dataset.
- **CUSUM Calculation:** For each year, a cumulative sum (S_t) was tracked starting from July 1st. The formula used was: $S_t = \max(0,\ S_{t-1} + (\mu_{summer} - Actual\ Temp_t - C_1))$
- **Parameters:** * **Slack (C_1) = 2:** This ensures the algorithm ignores minor daily fluctuations and only accumulates substantial temperature drops (more than 2 degrees below the summer mean).
    - **Threshold (T_1) = 150:** This acts as the "tripwire." Once the cumulative drop exceeds 150, the transition to fall is officially detected.

Results and Analysis

The detection dates across the two decades varied significantly, generally falling between mid-September and mid-October:





- **Earliest End of Summer:** 2004 (September 15th).
- **Latest End of Summer:** 2005 (October 13th).
- **Trace Visualization:** By overlaying the CUSUM traces for all years, we observe that some years exhibit a "sharp cliff" where temperatures drop rapidly, while others show a gradual, staggered decline into autumn.

Year	EndDay	Year	EndDay
1996	30-Sep	2006	29-Sep
1997	29-Sep	2007	11-Oct
1998	9-Oct	2008	27-Sep
1999	1-Oct	2009	18-Sep
2000	22-Sep	2010	12-Oct
2001	27-Sep	2011	1-Oct
2002	7-Oct	2012	5-Oct
2003	25-Sep	2013	24-Sep
2004	15-Sep	2014	5-Oct
2005	13-Oct	2015	27-Sep

2. Summer Warming Trend Analysis

Methodology:

While the first part looked at daily changes within a year, this part evaluates whether the overall summer heat (July and August) in Atlanta showed a sustained increase from 1996 to 2015.

- **Data Aggregation:** The mean temperature for July and August was calculated for each individual year.
- **Warming CUSUM:** A new CUSUM was applied to these annual means to detect a sustained shift above the global baseline.

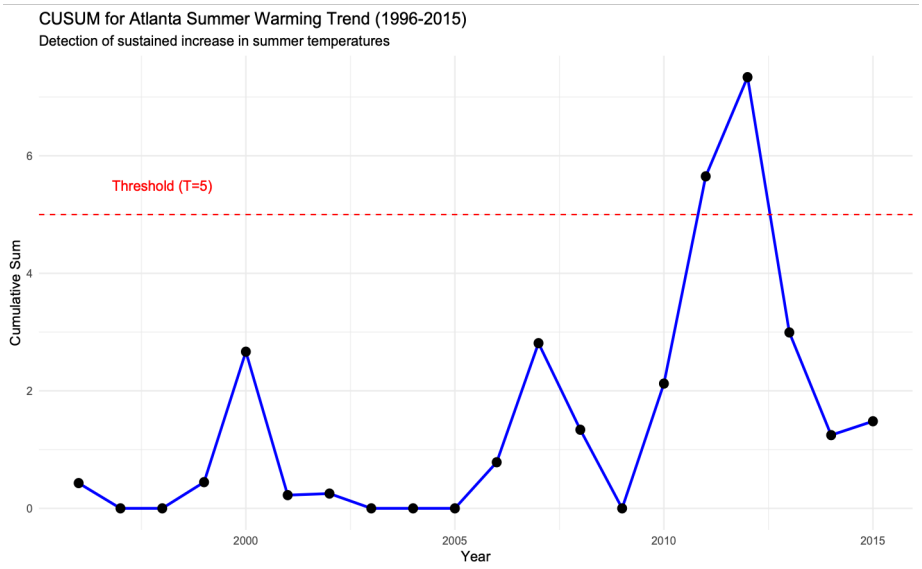
$$S_{warming} = \max(0, S_{i-1} + (Yearly\ Mean_i - \mu_{summer} - C_2))$$

- **Parameters:**
  - **Slack (C_2) = 0.5:** This helps filter out years that are only slightly warmer than average.
  - **Threshold (T_2) = 5:** A cumulative shift of 5 units across the years signals a significant climatic warming trend.

Results and Analysis:

The warming trend CUSUM provided the following insights:

- **Trend Accumulation:** The analysis tracked the CUSUM column in the warming_df to see if the values continued to climb over time.
- **Threshold Crossing:** By plotting these values against the threshold, the study identifies precisely which year Atlanta's summer heat reached a "statistically significant" higher state compared to the historical baseline.



Conclusions:

1. **Seasonal Consistency:** Atlanta's summer typically ends between late September and early October. The high threshold (T = 150) ensures that a single "cold snap" does not trigger a false detection of fall; rather, it requires a sustained shift in the weather pattern.
2. **Climate Shift Detection:** The warming trend analysis demonstrates that CUSUM is an effective tool for climate change monitoring, as it can detect subtle, incremental changes in annual means that might be missed by simply looking at a standard line graph of temperatures.