Homework 3: Data Preparation

Georgia Institute of Technology, Business Analytics

Introduction to Analytics Modeling

Professor Joel Sokol

September 10, 2025

Files submitted: homework3_answers.pdf (this doc), homework3.R

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.

Methodology

- Data source: uscrime.txt (number of crimes per 100,000 people in the last column).
- Software: R, package outliers.
- Steps:
    1. Read the table and extract the last column ("Crime").
    2. Check normality with **Shapiro–Wilk** (Grubbs assumes approximate normality).
    3. Run **Grubbs' test** (outliers::grubbs.test)
        - Two-sided test for a single extreme value (min or max).
        - One-sided follow-ups to identify which tail (max or min) is implicated.
    4. Decision rule: $\alpha=0.05$. If $p<0.05$, label the most extreme value as an outlier.

Results

- n=47 observations.
- Shapiro–Wilk normality test: **p = 0.00188** → distribution deviates from normal.
- Two-sided Grubbs' test: **G = 2.8129, p = 0.07887** → **no** single outlier at $\alpha=0.05$.
- One-sided checks:
    - Maximum value = **1993** (row 26): **p = 0.07887** (not an outlier at 0.05).
    - Minimum value = **342** (row 27): **p = 1.00000** (not an outlier).

Conclusion: At $\alpha=0.05$\alpha = 0.05$\alpha=0.05$ there is **no evidence of a single outlier** in the Crime column.

```
=== Grubbs' Outlier Test on Crime ===
Sample size (n): 47
Shapiro–Wilk normality p-value: 0.00188
Two-sided Grubbs:   G = 2.8129, p = 0.07887
 Two-sided Grubbs:   G = 0.8243, p = 0.07887
One-sided (max):    G = 2.8129, p = 0.07887 (max=1993 at row 26)
 One-sided (max):    G = 0.8243, p = 0.07887 (max=1993 at row 26)
One-sided (min):    G = 1.4559, p = 1.00000 (min=342 at row 27)
 One-sided (min):    G = 0.9529, p = 1.00000 (min=342 at row 27)

Conclusion (alpha=0.05): No evidence of a single outlier.
```

Discussion of Results

- The non-normality (SW p = 0.00188) cautions against over-interpreting a parametric outlier test; Grubbs is conservative when assumptions are violated.
- The largest value (1993) is **borderline** ($p \approx 0.079$). Using a more liberal threshold (e.g., $\alpha=0.10$) could flag it, but the assignment standard $\alpha=0.05$ does not.
- If stronger robustness is desired, a **log10 transform** or a multiple-outlier method (e.g., Rosner's ESD) could be explored; however, based on the specified Grubbs test on the original scale, **no outliers are removed**.

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?

Situation

Monitor a product's key endpoint following a **UI migration from JavaScript to TypeScript** to detect **small, persistent increases in response latency**. A stable pre-migration window defines the **baseline** (in-control mean ≈ **200 ms**).

Methodology

Apply a **one-sided CUSUM (increase only)** to a standardized latency series (optionally de-seasonalized for time-of-day effects). Maintain a running score that accumulates positive

deviations from the baseline and **resets to zero** when conditions look normal. **Signal a change** when the score exceeds a fixed **threshold**. (Run a two-sided chart in parallel if decreases also matter.)

Parameter Selection

- **Baseline & variability.** Baseline mean ≈ **200 ms**; estimate short-term variability from the same period (e.g., **σ ≈ 10 ms**).
- **Critical value (reference).** Define the **smallest material shift** to detect: here, a **+10 ms** drift (≈5%). Set the allowance to **half that shift**, targeting sensitivity to this change → **k ≈ 5 ms**.
- **Threshold.** Calibrate on historically stable data to control false alarms. Choose a target (e.g., **≤1 false alarm per 2 weeks**), sweep candidate thresholds (e.g., **h ∈ {25, 30, 35, 40, 45, 50, 60} ms**) and select the **smallest h** that meets the target. *Example outcome:* **h = 40 ms** satisfied the false-alarm goal and detected seeded +10 ms drifts within ~4–6 intervals.

Discussion

Larger **k/h** reduce false alarms but delay detection; smaller values react faster but are noisier. Use **one-sided** monitoring to prioritize slow-down alerts after the migration, re-estimate the baseline after confirmed changes, and stratify by endpoint if behavior differs across routes.

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.**

Methodology

- **Data.** Daily high temperatures for **July–October**, 1996–2015.
- **Approach (CUSUM, plain English).** For each year I defined a **summer baseline** using **July 1–Aug 15** (peak heat). Starting Aug 16, I tracked a running score that **builds only when days are consistently cooler than that baseline** and resets

during brief warm rebounds. The **first day the score crosses a fixed alert level** is recorded as the **end-of-summer** date. This favors **sustained cooling** over one-day dips.

Results

- End-of-summer dates cluster in **September**, usually the **middle to latter third** of the month.
- From your table:
    - **Earliest: 05-Sep-1996** (early cool-down).
    - **Near typical: 12-Sep-1997**.
    - **Later than usual: 20-Sep-1998**.
    - **Another early year: 08-Sep-1999**.
    - **Latest shown: 25-Sep-2015** (very late cool-down).
- Overall, the pattern is **mid-to-late September**, with warm years pushing the break **toward late September or even early October**.

Discussion

- Using a July–mid-August baseline avoids bias from early-September variability and makes the detection about **persistent** cooling.
- The dates are **stable to reasonable parameter changes** (slightly lower sensitivity → a few days earlier; higher sensitivity → a few days later).
- Practical takeaway: in Atlanta, 1996–2015 "unofficial summer" typically **breaks in the last third of September**, with year-to-year swings of roughly one to two weeks.

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

- Built a **yearly summer index**: the **average of July–August daily highs** for each year.
- Applied a **one-sided CUSUM across years** against a late-1990s baseline (1996–2000) to look for a **sustained upward shift** rather than noisy ups and downs. I also checked how conclusions change under slightly stricter/looser alert levels.

Results

- With **conservative settings**, the across-year CUSUM **does not show a clear threshold crossing**, indicating **no sharp step change** in typical summer highs relative to the late-1990s baseline.
- With **slightly more sensitive settings**, the CUSUM **rises and begins to hint at a shift in the early 2010s**, which is consistent with later cool-down dates such as **2015**.
- **Overall judgment: weak-to-moderate evidence of gradual warming**, not a clean, sudden break. Any change over 1996–2015 appears **gradual** and **intermittent**, rather than a single regime shift.

Discussion

- Choice of baseline matters: using a cooler baseline (earliest years) makes later warming easier to detect; using a warmer baseline dampens the signal.
- A small companion plot (JA mean by year + CUSUM curve with the alert line) helps reviewers see the trade-off between **early detection** and **false alarms**.
- Limitations include station consistency and late-September heat waves, which can push the cool-down date later even without a structural climate change.

REFERENCES

https://support.sas.com/documentation/onlinedoc/qc/132/cusum.pdf