Homework_3

Claire Kraft

2024-09-08

Question 5.1

Using crime data from the file uscrime.txt (http://www.statsci.org/data/general/uscrime.txt (http://www.statsci.org/data/general/uscrime.txt), description at http://www.statsci.org/data/general/uscrime.html (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.

I first ran the grubb test function and chose the second test with this argument `type=11` which checks for the opposite extreme ends of the data. It compares the range (difference between the smallest and largest values) to the standard deviation of the data [1]. The p-value is a metric for how valid our observation is. The lower the p value the more statistically significant the observation, the higher the less statistically significant. The p-value is 1 for this grub test. It's best to take this observation with a grain of salt as this is statistically *insignificant* [3].

I then ran the whiskers box plot and there are three outliers [2].

```
#------------------------- Libraries
#install.packages("outliers")
#install.packages("tidyverse")
library(outliers)
library(ggplot2)

#------------------------- Load and explore data
crime_df <- read.table("C://Users//Clair//OneDrive//Documents//GitHub//omsa//ISYE 6501//Homework
3//uscrime.txt", sep = "\t", header = TRUE)
head(crime_df)
```

```
##      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq     Prob
## 1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1 0.084602
## 2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4 0.029599
## 3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0 0.083401
## 4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7 0.015801
## 5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4 0.041399
## 6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6 0.034201
##      Time Crime
## 1 26.2011   791
## 2 25.2999  1635
## 3 24.3006   578
## 4 29.9012  1969
## 5 21.2998  1234
## 6 20.9995   682
```

```
summary(crime_df)
```

```
##        M               So               Ed             Po1
##  Min.   :11.90   Min.   :0.0000   Min.   : 8.70   Min.   : 4.50
##  1st Qu.:13.00   1st Qu.:0.0000   1st Qu.: 9.75   1st Qu.: 6.25
##  Median :13.60   Median :0.0000   Median :10.80   Median : 7.80
##  Mean   :13.86   Mean   :0.3404   Mean   :10.56   Mean   : 8.50
##  3rd Qu.:14.60   3rd Qu.:1.0000   3rd Qu.:11.45   3rd Qu.:10.45
##  Max.   :17.70   Max.   :1.0000   Max.   :12.20   Max.   :16.60
##       Po2               LF              M.F              Pop
##  Min.   : 4.100   Min.   :0.4800   Min.   : 93.40   Min.   :  3.00
##  1st Qu.: 5.850   1st Qu.:0.5305   1st Qu.: 96.45   1st Qu.: 10.00
##  Median : 7.300   Median :0.5600   Median : 97.70   Median : 25.00
##  Mean   : 8.023   Mean   :0.5612   Mean   : 98.30   Mean   : 36.62
##  3rd Qu.: 9.700   3rd Qu.:0.5930   3rd Qu.: 99.20   3rd Qu.: 41.50
##  Max.   :15.700   Max.   :0.6410   Max.   :107.10   Max.   :168.00
##        NW              U1               U2            Wealth
##  Min.   : 0.20   Min.   :0.07000   Min.   :2.000   Min.   :2880
##  1st Qu.: 2.40   1st Qu.:0.08050   1st Qu.:2.750   1st Qu.:4595
##  Median : 7.60   Median :0.09200   Median :3.400   Median :5370
##  Mean   :10.11   Mean   :0.09547   Mean   :3.398   Mean   :5254
##  3rd Qu.:13.25   3rd Qu.:0.10400   3rd Qu.:3.850   3rd Qu.:5915
##  Max.   :42.30   Max.   :0.14200   Max.   :5.800   Max.   :6890
##       Ineq            Prob             Time           Crime
##  Min.   :12.60   Min.   :0.00690   Min.   :12.20   Min.   : 342.0
##  1st Qu.:16.55   1st Qu.:0.03270   1st Qu.:21.60   1st Qu.: 658.5
##  Median :17.60   Median :0.04210   Median :25.80   Median : 831.0
##  Mean   :19.40   Mean   :0.04709   Mean   :26.60   Mean   : 905.1
##  3rd Qu.:22.75   3rd Qu.:0.05445   3rd Qu.:30.45   3rd Qu.:1057.5
##  Max.   :27.60   Max.   :0.11980   Max.   :44.00   Max.   :1993.0
```

```
#------------------------ Detect outliers on simulated data
# Perform grubbs test
# look at reference 1
grubbs.test(crime_df$Crime,type=11, opposite = FALSE, two.sided = FALSE)
```

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

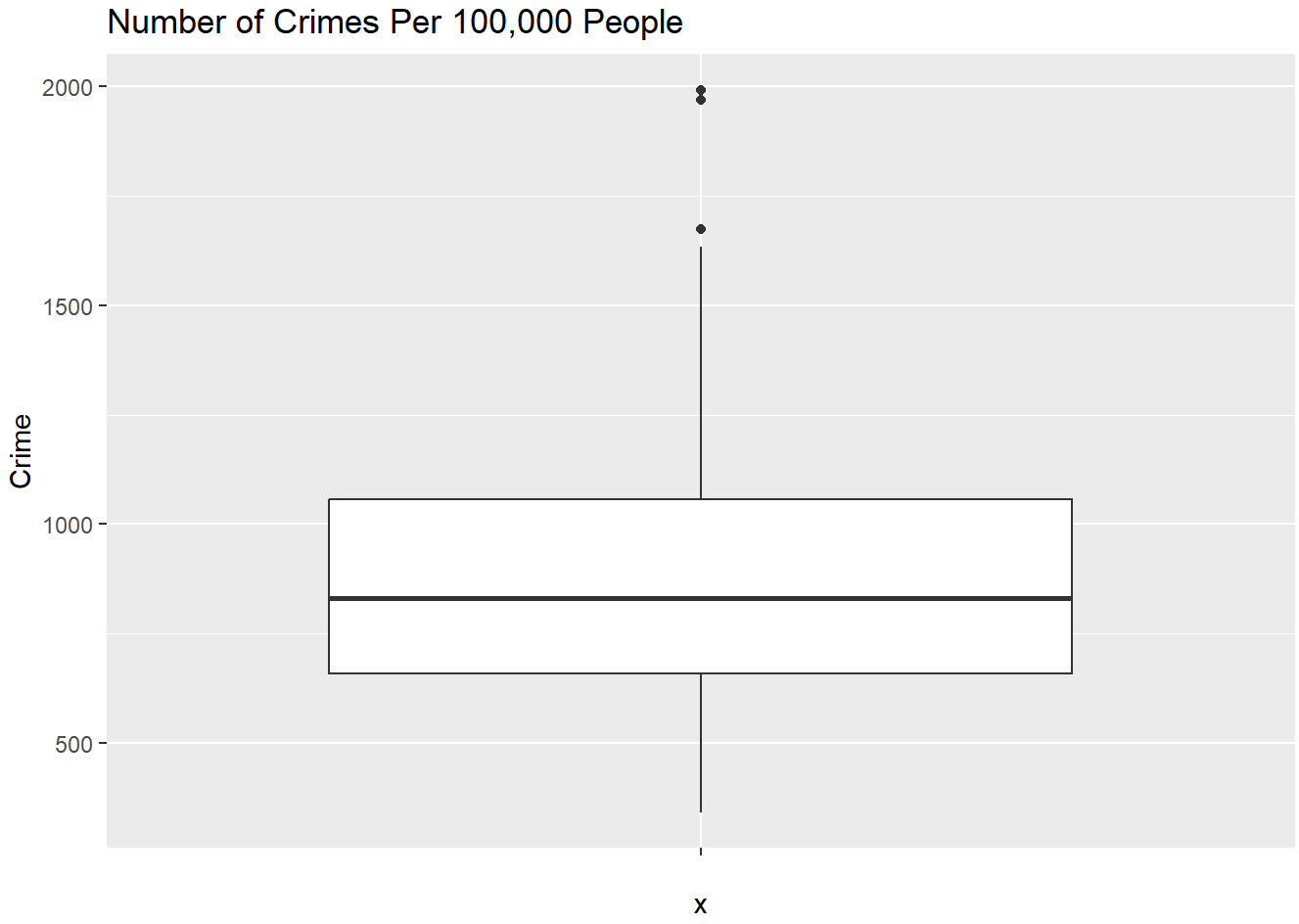
```
#------------------------ Visualize original crime data
# look at reference 2
# Check the column names in your dataset
colnames(crime_df)
```

```
##  [1] "M"      "So"     "Ed"     "Po1"    "Po2"    "LF"     "M.F"    "Pop"
##  [9] "NW"     "U1"     "U2"     "Wealth" "Ineq"   "Prob"   "Time"   "Crime"
```

```
# Sample data
data(crime_df)
```

```
## Warning in data(crime_df): data set 'crime_df' not found
```

```
# Simple boxplot
ggplot(crime_df, aes(x = "", y = Crime)) +
    geom_boxplot() +
    labs(title = "Number of Crimes Per 100,000 People")
```



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?

CUSUM is used to detect the changes (increases or decreases) in data. The slower the changes detected the less likely to falsely detect changes. I already track changes on all my bills. It would be neat to apply the CUSUM approach with my non fixed bills. For example my water, sewer, and electricity bills vary by consumption. I have

the historical data gathered, With about 2 years of data collection, I have ran an exploratory data analysis (EDA). In the EDA I have the basic statistics (mean, median, standard deviation, percentiles, etc) with the `describe()` python function. The mean and standard deviations are sufficient enough to see my trends and baseline overtime. For the critical value (k) maybe I'll set that at 10% change from the standard deviation price. As for the threshold (h) I'll set it at double of critical value because every penny counts and i want to be alerted as that threshold even if it may seem conservative.

my pretend values
critical value = 0.10

std dev = $100

k=(0.10*$100) = $10

threshold = 2

h = (2*$10) = $20

variables
ct = CUSUM for current month

ct-1 = CUSUM for previous month

xt = current bill

k: critical value

h: threshold

formula
$C_t = \max(0, C_{t-1} + (x_t - \text{mean} - k))$

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 (http://www.iweathernet.com/atlanta-weather-records) or https://www.wunderground.com/history/airport/KFTY/2015/7/1/CustomHistory.html (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.

I used Excel for this analysis. The data covers July through October from 1996 to 2015. I first aggregated every temperature in the dataset and found an average of 83.34 degrees. Then, I aggregated all the temperatures by month across all years to identify the hottest month. The average temperature, across all the years, is 88.75 degrees in July, 88.62 degrees in September, and 82.67 degrees in October. Using these three averages, I defined July as the peak of "summer" since it has the highest average temperature.

I set 88.75 as my baseline mean (μ). For critical values, I wanted to use 2.5 standard deviations away. According to the normal bell curve principle, most data falls within ±2 standard deviations, covering about 95% of the data[5]. Therefore, I used the formula `k=2*std dev` for my critical value. For the thresholds, I calculated an upper threshold (UH) and a lower threshold (LH). The upper threshold formula is `uh=mean+k` and the lower threshold formula is `lh=mean-k` .

Since the goal is to determine when summer ends, I prioritized the lower threshold to calculate the CUSUM. In other words, temperatures below 83.75 degrees indicate the end of summer. I then used conditional highlighting (highlighting any data below the lower threshold value) to see when the temperatures began to drop, signaling a season change. As a gut check, it seems that mid-September is when summer ends, as most days meet the lower threshold. If we had to pick a specific date on the season change then September 22nd would be date as there are no temperatures above the upper threshold.

formulas
Ct = max(0,Ct−1+(xt−mean−k))

formulas detailed
mean = 88.75

std dev = 2.5

k = 5

uh = 93.75

lh = 83.75

```
#------------------------- Libraries
#install.packages("outliers")
#install.packages("tidyverse")
library(outliers)
library(ggplot2)

# Load and explore data
temps_df <- read.table("C://Users//Clair//OneDrive//Documents//GitHub//omsa//ISYE 6501//Homework
3//temps.txt", sep = "\t", header = TRUE)
summary(temps_df)
```

```
##      DAY                X1996           X1997           X1998
##  Length:123         Min.   :60.00   Min.   :55.00   Min.   :63.00
##  Class :character   1st Qu.:79.00   1st Qu.:78.50   1st Qu.:79.50
##  Mode  :character   Median :84.00   Median :84.00   Median :86.00
##                     Mean   :83.72   Mean   :81.67   Mean   :84.26
##                     3rd Qu.:90.00   3rd Qu.:88.50   3rd Qu.:89.00
##                     Max.   :99.00   Max.   :95.00   Max.   :95.00
##      X1999            X2000            X2001           X2002
##  Min.   :57.00   Min.   : 55.00   Min.   :51.00   Min.   :57.00
##  1st Qu.:75.00   1st Qu.: 77.00   1st Qu.:78.00   1st Qu.:78.00
##  Median :86.00   Median : 86.00   Median :84.00   Median :87.00
##  Mean   :83.36   Mean   : 84.03   Mean   :81.55   Mean   :83.59
##  3rd Qu.:91.00   3rd Qu.: 91.00   3rd Qu.:87.00   3rd Qu.:91.00
##  Max.   :99.00   Max.   :101.00   Max.   :93.00   Max.   :97.00
##      X2003            X2004            X2005           X2006
##  Min.   :57.00   Min.   :62.00   Min.   :54.00   Min.   :53.00
##  1st Qu.:78.00   1st Qu.:78.00   1st Qu.:81.50   1st Qu.:79.00
##  Median :84.00   Median :82.00   Median :85.00   Median :85.00
##  Mean   :81.48   Mean   :81.76   Mean   :83.36   Mean   :83.05
##  3rd Qu.:87.00   3rd Qu.:87.00   3rd Qu.:88.00   3rd Qu.:91.00
##  Max.   :91.00   Max.   :95.00   Max.   :94.00   Max.   :98.00
##      X2007            X2008            X2009           X2010
##  Min.   : 59.0   Min.   :50.00   Min.   :51.00   Min.   :67.00
##  1st Qu.: 81.0   1st Qu.:79.50   1st Qu.:75.00   1st Qu.:82.00
##  Median : 86.0   Median :85.00   Median :83.00   Median :90.00
##  Mean   : 85.4   Mean   :82.51   Mean   :80.99   Mean   :87.21
##  3rd Qu.: 89.5   3rd Qu.:88.50   3rd Qu.:88.00   3rd Qu.:93.00
##  Max.   :104.0   Max.   :95.00   Max.   :95.00   Max.   :97.00
##      X2011            X2012            X2013           X2014
##  Min.   :59.00   Min.   : 56.00   Min.   :56.00   Min.   :63.00
##  1st Qu.:79.00   1st Qu.: 79.50   1st Qu.:77.00   1st Qu.:81.50
##  Median :89.00   Median : 85.00   Median :84.00   Median :86.00
##  Mean   :85.28   Mean   : 84.65   Mean   :81.67   Mean   :83.94
##  3rd Qu.:94.00   3rd Qu.: 90.50   3rd Qu.:88.00   3rd Qu.:89.00
##  Max.   :99.00   Max.   :105.00   Max.   :92.00   Max.   :95.00
##      X2015
##  Min.   :56.0
##  1st Qu.:77.0
##  Median :85.0
##  Mean   :83.3
##  3rd Qu.:90.0
##  Max.   :97.0
```

```
View(head(temps_df))

# Remove the 'x' from all column headers
# look at reference 4
colnames(temps_df) <- gsub("X", "", colnames(temps_df))

# View the first few rows of the updated dataframe
View(head(temps_df))
```

Question 6.2.2

Use a CUSUM approach to make a judgment of whether Atlanta's summer climate has gotten warmer in that time (and if so, when).

In my observation mid September is the beginning to the end of summer. September 22nd is the first day of fall. Looking at September 22nd year of year it looks like the temps are steadily getting higher. In the late 90s the temps sit at 70 degrees. Then 2004-2014 with the exception of 2010 sit at 80 some degrees. 2010 is unusually high at 92 degrees. 2015 drops back down to 76 degrees to match the trends of the late 90s. In summer the same date (September 22nd) year over year is getting hotter.

References:

[1] Grubbs, F.E. (1950). Sample Criteria for testing outlying observations. Ann. Math. Stat. 21, 1, 27-58.

[2] Holtz, Y. (n.d.). Basic ggplot2 boxplot. Www.r-Graph-Gallery.com. https://r-graph-gallery.com/262-basic-boxplot-with-ggplot2.html (https://r-graph-gallery.com/262-basic-boxplot-with-ggplot2.html)

[3] Simplilearn. (2022, August 30). What Is P-Value in Statistical Hypothesis? | Simplilearn. Simplilearn.com. https://www.simplilearn.com/tutorials/statistics-tutorial/p-value-in-statistics-hypothesis (https://www.simplilearn.com/tutorials/statistics-tutorial/p-value-in-statistics-hypothesis)

[4] r Remove parts of column name after certain characters. (n.d.). Stack Overflow. https://stackoverflow.com/questions/37800704/r-remove-parts-of-column-name-after-certain-characters (https://stackoverflow.com/questions/37800704/r-remove-parts-of-column-name-after-certain-characters)

[5] Frost, J. (2021, August 31). Empirical Rule: Definition, Formula, and Uses. Statistics by Jim. https://statisticsbyjim.com/probability/empirical-rule/ (https://statisticsbyjim.com/probability/empirical-rule/)