Homework 3

2024-09-10

R Markdown File

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.

First thing is to load necessary libraries. I use rio to import datasets easier. The others are for cleaner formatting for the plots and dates.

```
#housekeeping
library(pacman)
pacman::p_load(rio,outliers, ggplot2, tidyverse, tidyr, lubridate, repr,
reshape)
```

Importing the dataset into a vector called crimeData. Since we are only looking for outliers in the last column, I thought it would be helpful to quickly look at a summary for that column. This just gives us a few numbers to look at ie. min, max, median, mean.

```
#Question 5.1
#load data
crimeData <- import("D:/…/uscrime.txt")
head(crimeData) #preview of data
```

```
##      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
```

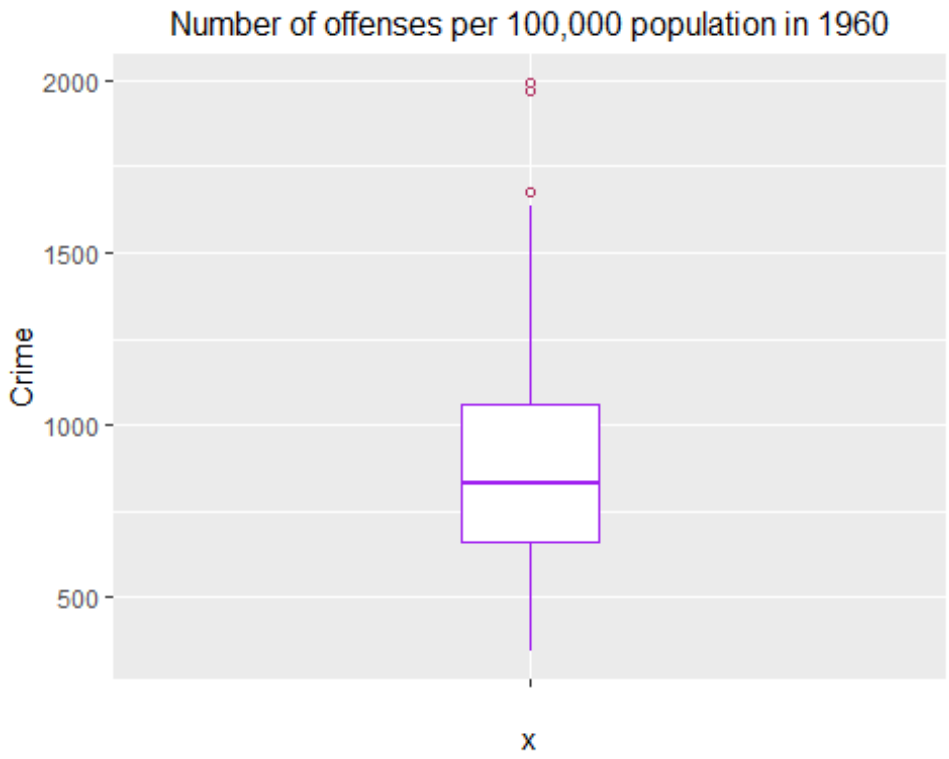
```
## 4 29.9012  1969
## 5 21.2998  1234
## 6 20.9995   682
```

```
summary(crimeData$Crime) #summary of the last column (crime rate: number of offenses per 100,000 population in 1960)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   342.0   658.5   831.0   905.1  1057.5  1993.0
```

Before using the grubbs.test function, I can also create a box-and-whisker plot to check for outliers via visualization. The code below makes the plot and I've added some extra color and formatting just to make it prettier. As you can see, the plot indicates that there are three data points which are far out of the expected range. At first glance, one might consider to write these points off as outliers, but I believe these are valid observations coming from real data. We can use the grubbs.test function to verify.

```
#visualization (box and whisker)
options(repr.plot.width=2, repr.plot.height=5) #height and width of plot
ggplot(crimeData, aes(x="",y=Crime))+ #use crime dataset for plot, y-axis is the crime, x-axis empty
  geom_boxplot(width=0.2,outlier.color="maroon",outlier.shape = 1, color = "purple")+ #initialize boxplot + color and shape
  labs(title="Number of offenses per 100,000 population in 1960")+ #title
  theme(plot.title = element_text(hjust=0.5,size=12)) #center title, change font
```



The Grubbs' test is based on the difference between the most extreme value (either min or max) and the mean of the dataset, normalized by the standard deviation. So basically, find the extreme value, calculate how far it is from the average of the dataset, then divide by the standard deviation. It returns a test statistic, critical value, and p-value. G is the test statistic and it represents how far the most extreme point is from the mean, relative to the standard deviation. Larger g = more likely a point is an outlier. U is the ciritical value/threshold, (which depends on the size of dataset). It is used to evaluate the test statistic. Basically, its whats considered an extreme enough deivation.

p-value is used to decide if something is an outlier. A p-value < 0.05 indicates that the observed extreme value is significantly different from the rest of the data, implying an outlier. (So one can conclude either the min, max, or both are outliers) A p-value >= 0.05 means there is no evidence to suggest that the extreme values are outliers.(No evidence the min nor max are outliers.)

The code below runs the Grubbs' test on the Crime column in the crimeData dataset. I did not specify the type since the box and whisker plot indicated potential outliers are all on one side. The p-value returned is 0.07887, which means there is no evidence that the max (highest value 1993) is an outlier. (I had also tried type 11, which checks two points, min and max. The p-value was 1, which would also mean there is no evidence that the points are outliers.)

```
#grubbs
#?grubbs.test
grubbs.test(crimeData$Crime) #default(type 10) checks single outlier either
min or max)

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

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 can apply a change detection model/CUSUM approach to monitor my internet usage, since my usage depends on my activities. If I'm doing schoolwork/studying I use less internet. On the other hand, if I am relaxing, my usage spikes as I am doing way more things/have more things running at once. a CUSUM model can detect when a shift occurs, signaling different activities. I would set a baseline for average internet usage and track deviations. CUSUM will accumulate deviations from the baseline. C value would represent expected variation ie. usage will probably fluctuate naturally by 1GB when doing school work so a lower C value (like 1) would be used. Threshold will determine when CUSUM detects significant

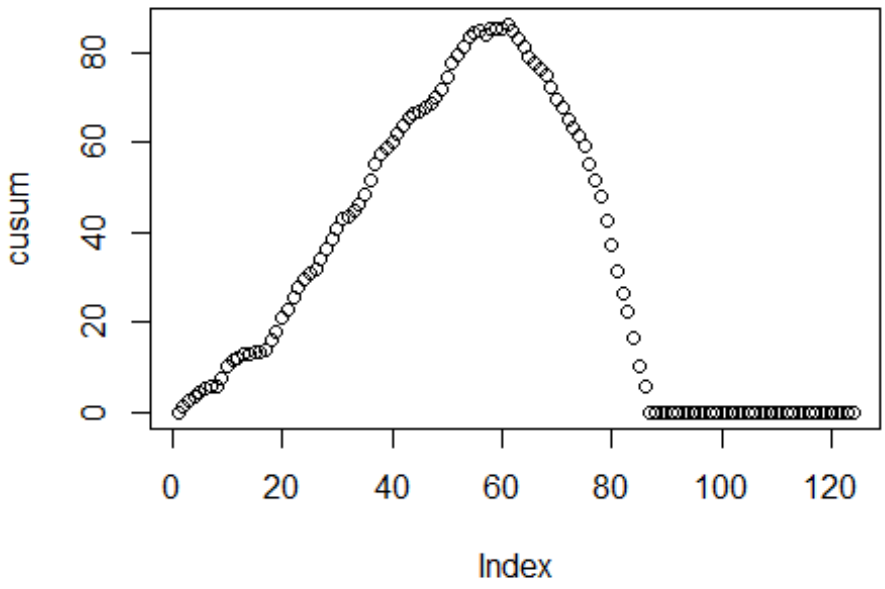
shif. In my case I typically do school work in the morning - afternoon, then relax in the evening - night. I would set a lower threshold to detect more frequent/daily shifts.

Question 6.2.1

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.

```
#Question 6.2a
temps <- import("D:/…/temps.txt")

avgTempDate <- rowMeans(temps[c(2:length(temps))], dims=1, na.rm=T) #average
temp for each day across the years in the txt
avgMean <- mean(avgTempDate) #mean temp of the avg time series
tempDev <- avgTempDate - avgMean #deviations for each day
C <- 4 #slack/allowance value
adjDev <- tempDev - C #adjusting deviation = tweaking sensitivity to detect
significant change
cusum <- numeric(length(adjDev) + 1) #empty vector, extra 0 for loop (i+1)
for (i in 1:length(adjDev)) #iterate through each day, check cusum, update
index with appropriate value
{
  checker <- cusum[i] + adjDev[i] #if check is positive, it is added to
cusum, otherwise reset to 0
  ifelse(checker > 0, cusum[i+1] <- checker, cusum[i+1] <- 0)
}
plot(cusum)
```



The above code implents the CUSUM approach to identify the unofficial end of summer. First, I load the data into a vector called temps. Then, I calculate the average daily high temperatures across multiple years (stored in avgTempDate). Using that vector (filled with the avg temp of the specified dates over the years) I get an overall average temperature (avgMean). This provides a reference for detecting deviations. The avgMean is 83.33. To calculate deviations (tempDev), I did avgTempDate - avgMean. This calculates how each day's average temperature deviates from the overall mean. Positive values indicate temperatures above the mean, and negative values indicate temperatures below the mean.

C is our slack/allowance value aka our expected deviation. After looking at the data, I went with a C value of 4 since temperature are usually within 4 degrees of each day on a given day. It will help me determine how much deviation from the mean is considered significant. I adjusted the deviation (adjDev) by subtracting C from tempDev. This adjustment makes the CUSUM calculation more sensitive to significant deviations. Deviations that are less than C will be reduced further, and those greater will be adjusted accordingly

I created the vector to store the CUSUM. The loop iterates through each day and updates the CUSUM. If checker is positive, it updates the next CUSUM value cusum[i+1] with checker. Otherwise, it sets cusum[i+1] to zero. This logic helps in detecting significant shifts by resetting the cumulative sum when it falls below zero.

Finally, I plotted the graph. We can see the graph follows a normal standard deviation pattern. This means the peak of the plot indicates significant changes in temperatures as well as a turning point (when the temperature starts dropping). The reason the peak indicates temperatures starting to fall is because when CUSUM starts to fall it means the

daily temperature is starting to fall closer to/under the overall mean temperature. The deviations will turn negative or become smaller and CUSUM stops increasing. The CUSUM peak represents the end of the phase where temperatures were accumulating above the mean temperature.

```
maxIndex <- which.max(cusum) #peak of cusum
temps[maxIndex, 1]  #the corresponding day

## [1] "30-Aug"

#which cusum is higher than 85 (threshold set based on graph)
which(cusum >= 85)

## [1] 56 58 59 60 61

temps[56, 1]

## [1] "25-Aug"

temps[58, 1]

## [1] "27-Aug"

temps[59, 1]

## [1] "28-Aug"

temps[60, 1]

## [1] "29-Aug"

temps[61, 1]

## [1] "30-Aug"

#peak + downward trend
```

This section simply (1) returns the peak of CUSUM and (2) returns index where CUSUM is above a threshold I set it to 85 by eyeballing the plot. The peak CUSUM is recorded on Aug 30. If you look at the graph, that point is a bit higher than the surrounding points, even though they are all above the threshold. The days that are at or above the threshold are Aug 25 and Aug 27-30. So basically, these few days (the end of August) signal the official end of summer (since temperatures start dropping). Obviously the weather is still hot, and it'll take some time before fall weather comes, but this is the period when the temperature starts cooling off.

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).

Looking at the above data (average daily temp over the years, deviations, etc), there is no evidence that summer climate has gotten warmer nor has summer gotten longer. For the years provided, it seems the temperature is always starting to cool off at the end of August.