ISYE 6051 : Homework 3
2/10/2021

Table of Contents

5.1 Outliers using the grubbs.test function

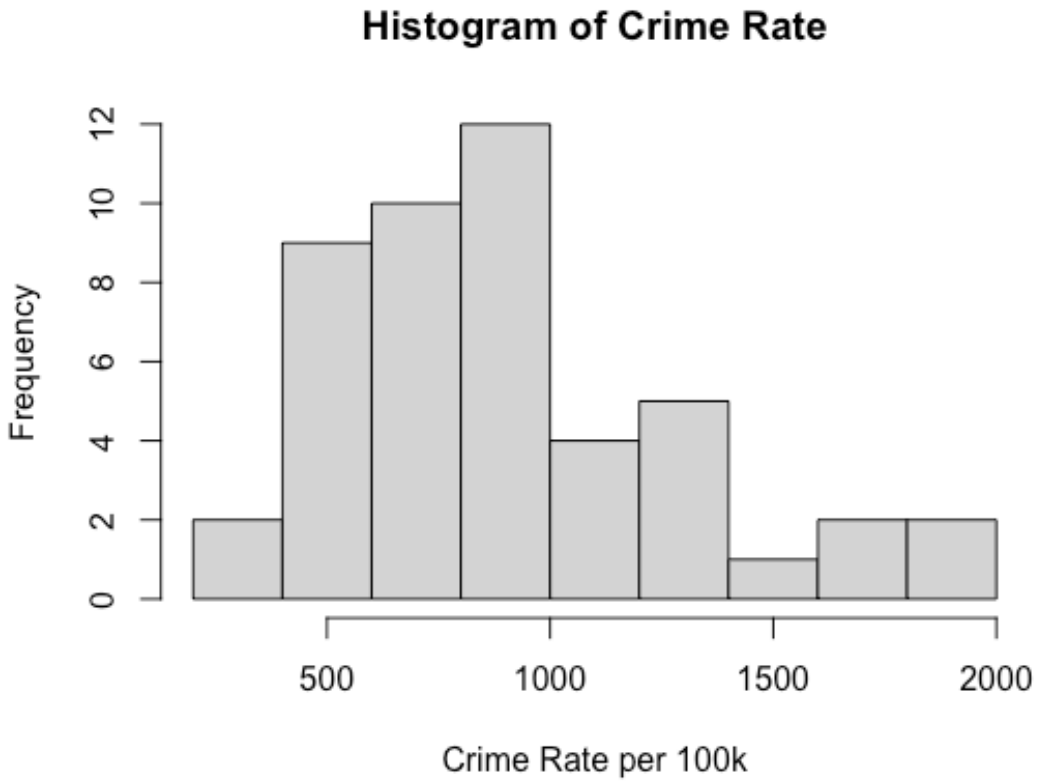
Dataset uscrimes contains 16 variables with 47 observations. The first assignment is to determine any outliers utilizing the grubbs.test function. Data must be first normalized in order to accurately utilize the function and per additional information utilizing histograms to visualize the data against a normal curve.

```
set.seed(17)

datapoints <- select(CrimeData,
c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16))

print(datapoints)
summary(datapoints)

hist(datapoints$Crime,
    xlab = 'Crime Rate per 100k',
    main = "Histogram of Crime Rate",
    breaks = sqrt(nrow(datapoints)))
```



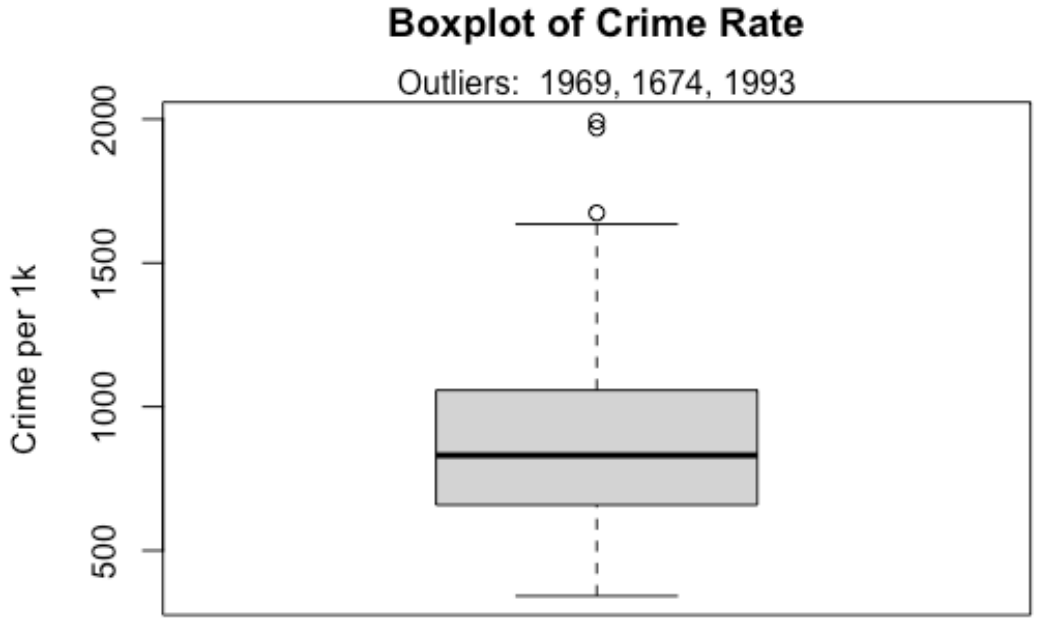
The median value of the variable – crime is equal to 831.0 as reflected in the summary of the dataset datapoints and as the highest point in the bell curved historigram. Data is therefore normalized.

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

To start by visualization techniques, a boxplot can be used to evaluate if there are outliers using an IQR of 25th – 95th percentile.

```
boxplot(datapoints$Crime,
        ylab='Crime per 1k',
        main = "Boxplot of Crime Rate")
outlierdata <- boxplot.stats(datapoints$Crime)$out
outlierrows <- which(datapoints$Crime %in% c(outlierdata))

mtext(paste("Outliers: ", paste(outlierdata, collapse = ", ")))
```



The boxplot shows outliers = 1969,1674 and 1993.
The grubbs.test function provides information on the min or max possible outlier that is furthest away from the mean. The hypothesis states that the max point is not an outlier with the alternative hypothesis being that the max point is an outlier.

```
grubbs.test(datapoints$Crime, two.sided=TRUE)
```

```
        Grubbs test for one outlier

data:  datapoints$Crime
G = 2.81287, U = 0.82426, p-value = 0.1577
alternative hypothesis: highest value 1993 is an outlier
```

p-value is a value between 0 and 1 that measures the evidence that a datapoint is in fact an outlier. The closer to 0 is stronger evidence that it is an outlier. The closer to 1 is weaker evidence that the point is an outlier. The grubbs test indicates that the alternative hypothesis was chosen (1993 is an outlier) which aligns with the boxplot analysis showing 1993 as an outlier.

6.1 Detecting comorbidities using the CUSUM technique

The risk of having one or multiple comorbidities (high blood pressure, high cholesterol, type-2 diabetes) have long been assumed to be linked to an increase in weight across ethnic groups. Public health officials and individuals have looked for that magic number to determine optimal weight by factors such as ethnicity, height, bone structure, etc.

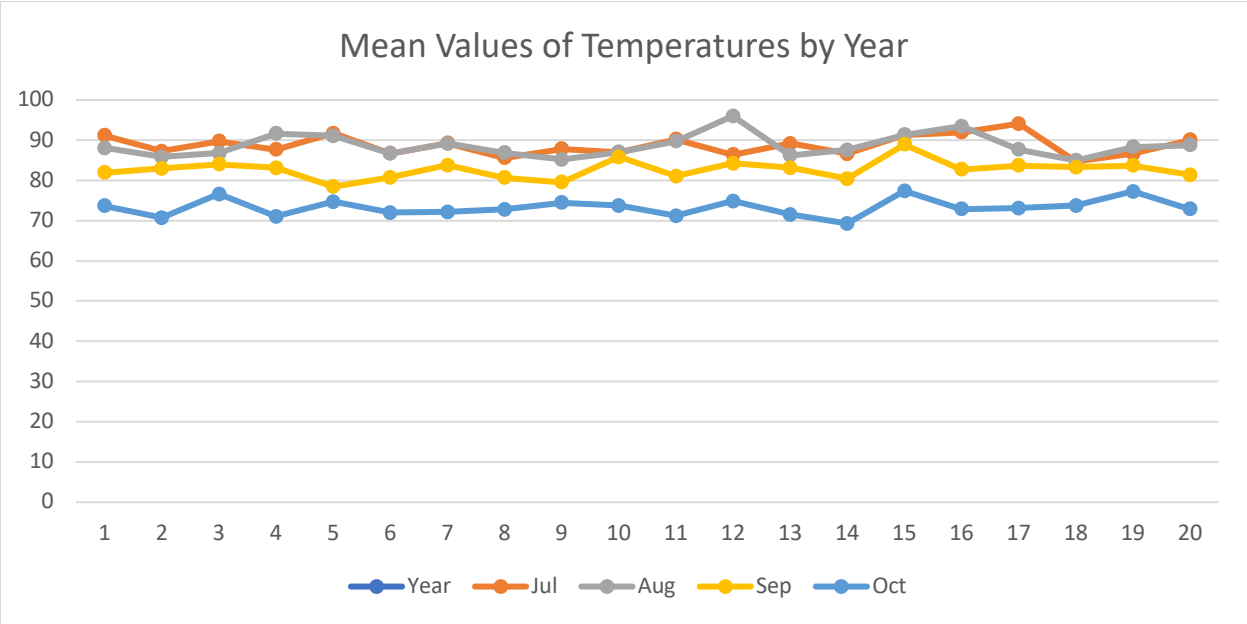
CUSUM can be used to monitor increases in comoborbidities as weight increases. Since even small weight changes have been shown to impact an individual's risk to introduce comorbidities, the C and T values should be conservative. Instead of C being ½ of the standard deviation and T being 5 times the standard deviation, I would look to start with significantly smaller values in order to catch slight weight changes that can be reduced much easier. False positive are better than leeting weight increase so quickly that reduction becomes prohibitively challenging.

6.2.1 Evaluating temperature fluctuations in Atlanta, GA

CUSUM modeling provides a mechanism to determine if the mean of an observation has gone beyond a critical level. In this example, I used data from temps.txt and by observation noted that the largest temperature increases occur during the July – September timeframe. Therefore, raw data through October 31st is used since this will provide the average of values before an observed change in temperature to determine if there is evidence of hotter summers.

To visualize the possibilities with data I created a tab on the excel spreadsheet (view both raw data and CUSUM tabs at Atlantatemperatures.pdf) to plot the mean of temperature by year for a 20-year timeframe of 1996-2015.

Year	Jul	Aug	Sep	Oct
1996	91.1935484	88.0322581	81.9333333	73.6451613
1997	87.2580645	85.8064516	82.9333333	70.7419355
1998	89.7096774	86.7741935	83.9666667	76.5806452
1998	87.6451613	91.6129032	83.1333333	71.0322581
2000	91.7419355	91.0645161	78.4333333	74.7096774
2001	86.7419355	86.7419355	80.7	72
2002	89.2580645	89.1612903	83.7666667	72.1612903
2003	85.5806452	86.8709677	80.6333333	72.8064516
2004	87.8387097	85.1612903	79.5666667	74.4193548
2005	86.9354839	87.0322581	85.8333333	73.7096774
2006	90.1935484	89.7419355	81.0333333	71.1612903
2007	86.4193548	96	84.2666667	74.8709677
2008	89.1612903	86.2580645	83.1	71.5483871
2009	86.6451613	87.5806452	80.4333333	69.2903226
2010	91.2580645	91.3548387	88.9	77.3870968
2011	91.9354839	93.483871	82.7666667	72.8387097
2012	94.0967742	87.6451613	83.7	73.1290323
2013	84.7096774	84.9677419	83.2666667	73.7741935
2014	86.6129032	88.2580645	83.6666667	77.2258065
2015	90.0645161	88.7741935	81.4	72.9032258



With the exception of a few spikes in data for years 2007 and 2011, visually there does not appear to be a significant increase in temperature during the summer months over this timeframe. Temperatures are generally within a +/- range of 10 degrees with no indication that temperatures are rising significantly over time.

To determine when summer unofficially ends, I chose to use a sample of the data based on visualization to calculate the mean and stdev for each year. Mid-September appeared to be when the first sustainable drop in temperature took place each year. Therefore, the sample data was 80 out of 183 observations.
C = B2
T = B3
Row 5 = mean of sample data
Rows 6 = stdev of sample data
Row 7 = date of change
W7 = Average date of change over the 20-year period

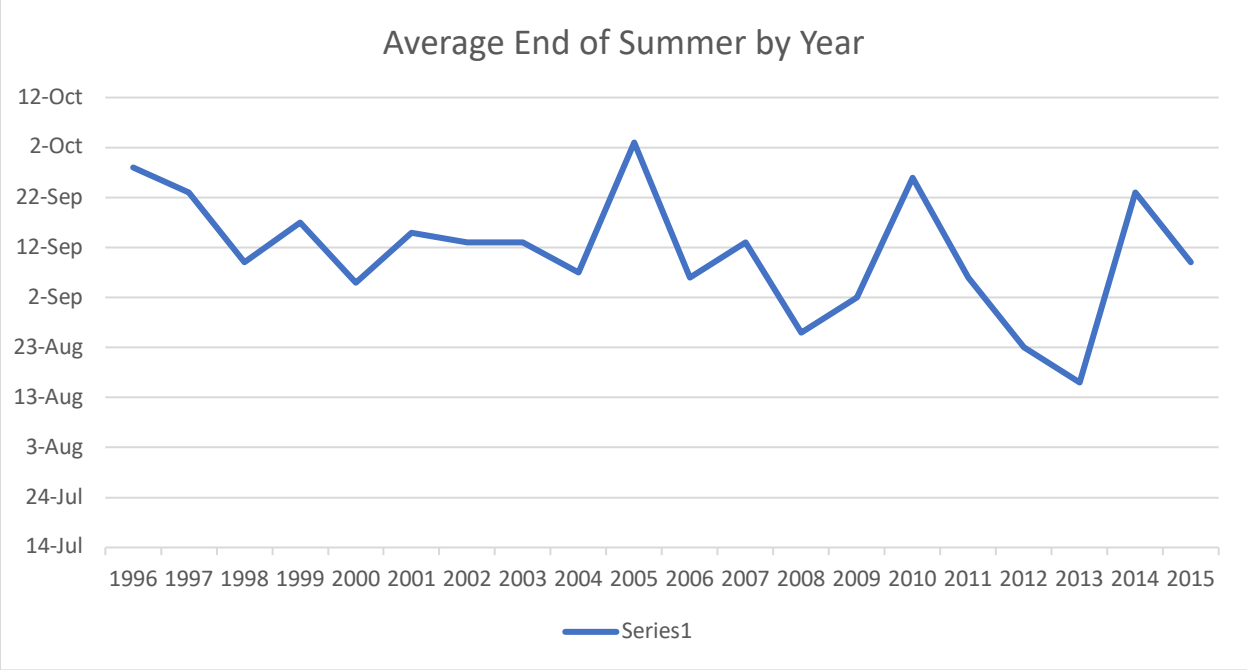
Using multiple values of C (~ ½ average stdev for all years, 1 stdev for all years rounded up) and increments of 5, 15, 25, 50 for T as the threshold.

C	T	Avg Date of Change	Avg Temp
2	**5**	**Aug 1**	**88.28**
2	**15**	**Aug 19**	**86.76**
2	**25**	**Aug 28**	**86.88**
2	**50**	**Sep 10**	**79.78**
5	**5**	**Aug 3**	**83.93**
5	**15**	**Aug 20**	**87.01**
5	**25**	**Aug 29**	**86.61**
5	**50**	**Sep 11**	**83.08**

Based on the variables of C and T, C at either 2 (1/2 stdev) or 5 and T at 50 appear to give the most accurate date of change. This matches up with the visualization shown earlier.

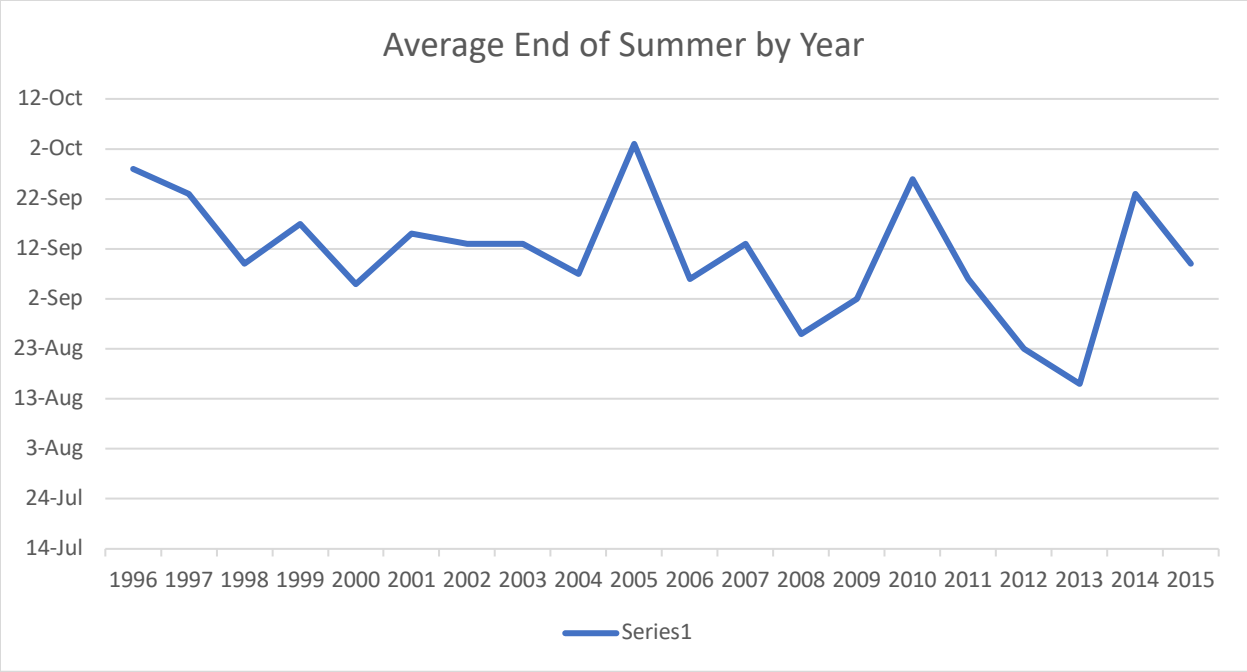
6.2.2 Determining Climate Changes

Using the variables of C at 2 and T at 50, the following plot shows the variation in temperature. September 10th was the average date of change.



Viewing the datapoints there are higher peaks during the years 2005, 2010 and 2014. However, there were also dips in average dates of change at 2008 and 2013.

Using the variables of C at 5 and T at 50, the following plot shows the variation in temperature. September 11th was the average date of change.



Once again viewing the datapoints there are higher peaks during the years 2005, 2010 and 2014. However, there were also dips in average dates of change at 2008 and 2013. These similarities are most likely due to the fact that C at 2, T at 50 and C at 5, T at 50 registered a date of change within 1 day.

Based on the CUSUM data, there is no obvious indication that the temperature is warming year over year. What is apparent though is that the variation in average temperature fluctuates at a much higher rate (highs and lows) as time increases.