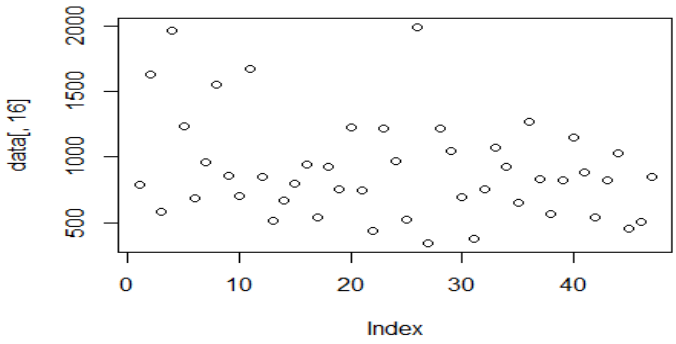
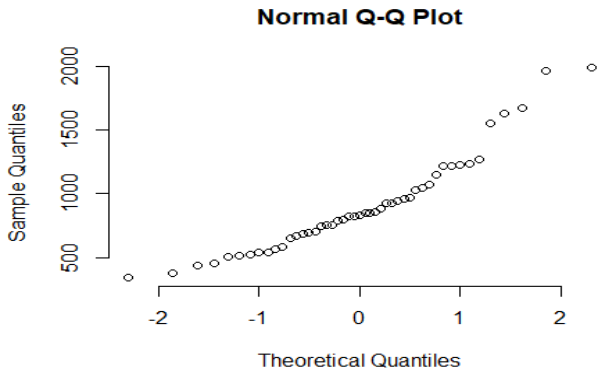
Question 5.1

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.

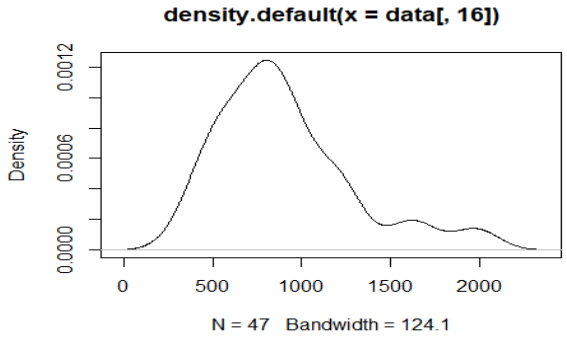
The code for everything in this solution can be found in appendix 5.1. In this question, I wanted to test a few things. So, to begin, I first plotted the data in the crime column to see if visually I could identify any points that may be outliers. The plotted data is as follows:



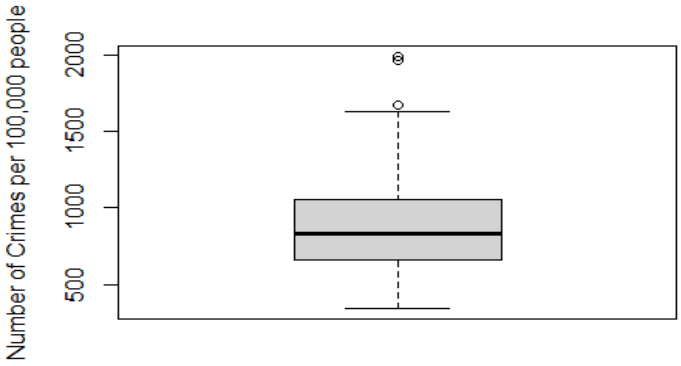
As a first impression from this plot, there seem to be a few outliers near the top. Now, as the Grubbs' test can only be performed if the data is of a normal distribution, I also ran a Shapiro-Wilk test and obtained a p-value of 0.001882, which indicates that the data is not normally distributed. However, to see why this may be, I looked at a Q-Q plot:



The vast majority of the points here in the middle seem to follow a normal distribution and the ends seem to be outliers. These outliers may be the reason the Shapiro-Wilk test said the data was not normal and as such, I decided to continue with the Grubb's test as planned (since the middle looks normally distributed). However, I also wanted to visualize these outliers some more and get a few more metrics for the data. As a result, I created a density plot:



From this plot, it is apparent that there is a right skew in the data. Because of this, I measured the skewness of the data and obtained the value 1.08848, which fits what we see visually. I also looked at the kurtosis of the data and obtained a value of 3.943658, which indicates that it is leptokurtic. Leptokurtic means that the distribution has more data in the tails, thus indicating the possibility of more outliers. Continuing, I decided to perform a Grubbs' test with type=11 to identify opposite outliers. The identified outliers were 342 and 1993 (here, 1993 matches what you get if you use the outlier function on the crime column). However, the p-value was 1, which in this situation indicates that we cannot accept the alternate hypothesis that both are outliers. Based on the density plot, skewness, and kurtosis calculated earlier, I thought it would be more likely that there is an outlier in the right tail. For this reason, I ran the Grubbs' test again with type=10, which is a test for one outlier. The obtained p-value was 0.07887 and the alternative hypothesis was that the highest value 1993 is an outlier. However, as the p-value is still above 0.05, we fail to reject the null hypothesis that the highest value is not an outlier. Still, as the p-value is close to significance, I wanted to explore this situation a bit more. For that reason, I created a box-and-whisker plot of the data:



As can be seen, the plot shows that there are likely two outliers near the top. Extracting the values of the points, one can see they are 1969, 1674, and 1993. In this situation, whether or not one wants to remove the outliers from the data depends on their risk tolerance. If they want to go strictly off of the p-value in the Grubb's test, they might feel compelled to not exclude it. However, if they look at the box plot, they may feel compelled to exclude the 1993 and 1969 values. Personally, as there is some confusion regarding whether or not we should exclude the

point, I decided to experiment by removing the 1993 point and running the Grubb's test again. Here, I got a p-value of 0.02848 with the alternative hypothesis that 1969 is an outlier. As the p-value is below 0.05, we can indeed reject the null hypothesis and accept the alternative hypothesis. As such, I might feel more confident in removing that data point.

Finally, as a fun little experiment, because there might have been more than one outlier, I also performed the Generalized Extreme Studentized Deviate Many Outlier test. By doing so, I found that when two outliers were tested (the outliers being 1993 and 1969), a p-value of 0.05695641 was obtained, which could (due to how close it is to 0.05) indicate that there may be more than one outlier in the data. This would match what we have noticed in our analysis.

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?

This is an interesting question. As I did my undergrad in the medical sciences, I think that a pandemic would be a great example of where a Change Detection model would be helpful. If we take the COVID-19 pandemic, we can begin by monitoring the number of cases. When the cases reach a certain threshold in an area, it could indicate that more actions need to be taken due to an outbreak, whether that be a lockdown or something else.

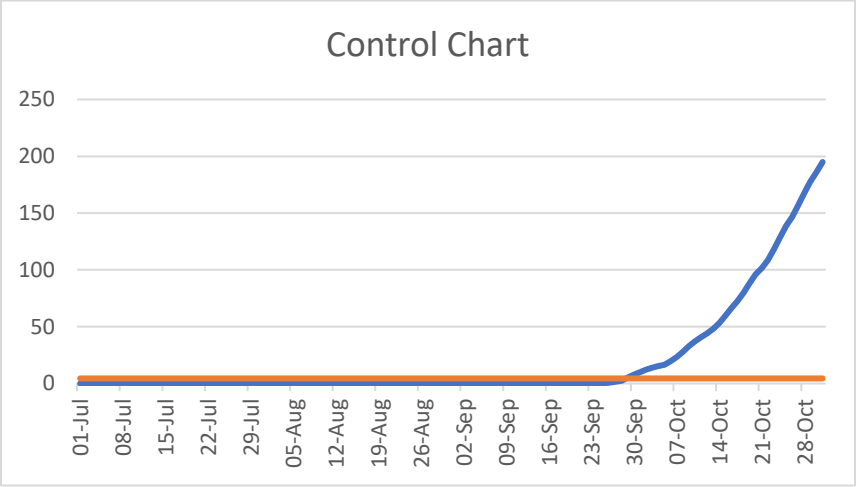
To choose a critical value and a threshold value, I would look at previous research done in this field. It might indicate values to start off with. However, if I notice that the virus is particularly contagious, I might be willing to take even less risk. For this reason, I would decrease the C value and threshold even more. It would be better to have false positives (e.g., detect there is an outbreak when there isn't and try to stop it) than false negatives (e.g., detect no outbreak when there really is one and you don't stop it in time).

Question 6.2

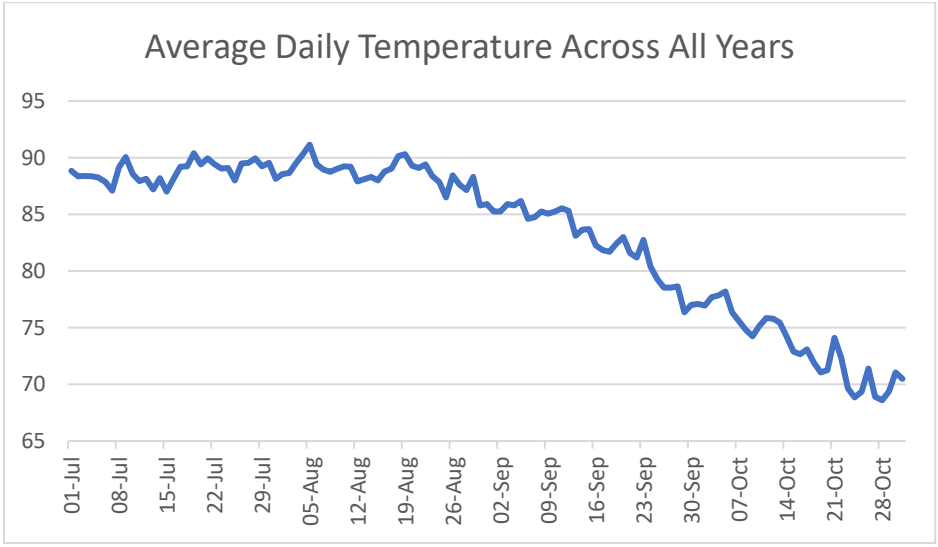
1. Using July through October daily-high-temperature data for Atlanta for 1996 through 2015, use a CUSUM approach to identify when unofficial summer ends (i.e., when the weather starts cooling off) each year. You can get the data that you need from the file temps.txt or online, for example at http://www.iweathernet.com/atlanta-weather-records 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.

   The data and approach for this problem can be found in the attached Excel document in sheet 6.2.1. In order to approach this problem, I found out the mean temperature for each day across all the years. Then, as I wanted to detect a decrease in temperatures, I set up the equation for S_t such that it would detect a decrease. In particular, this was where $S_t = \max\{0, S_{t-1} + (\mu - x_t - C)\}$. I played around quite a bit with what values I should choose for C and T. Ultimately, I decided that I would find the standard deviation of the mean temperatures for each day across all the years already obtained until the July 31st date.

My reasoning for this was because I wanted a measure that would capture the "official summer" months. With this measure, then I could see when the temperature starts to change from it using the CUSUM approach. Correspondingly, I set C to 0.5 times the value for standard deviation obtained. Threshold was set to 5 times the standard deviation value. Based on these parameters, I was able to obtain the following CUSUM control chart:

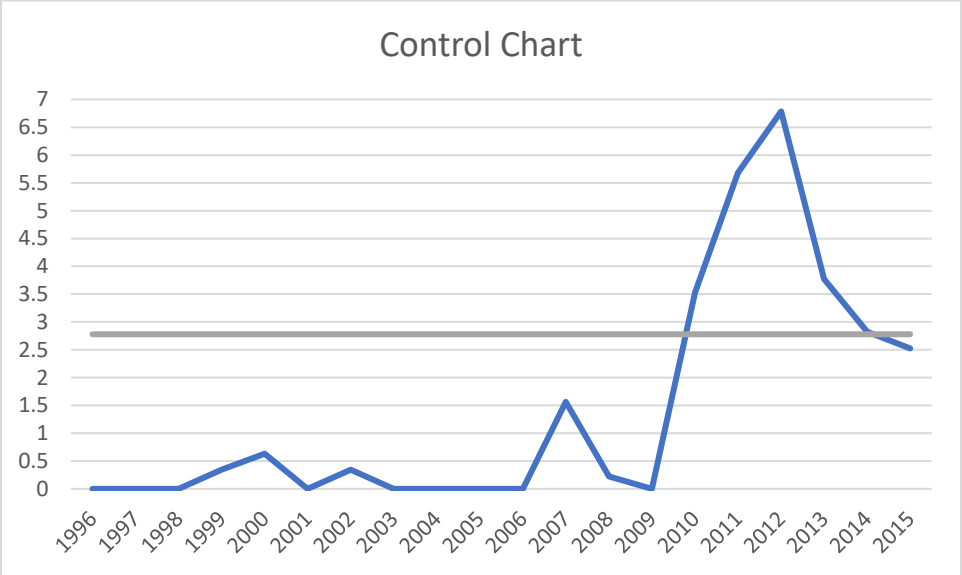


As per this chart, a change was detected around September 25th, which is what I would give as my answer for when "unofficial summer" ends and the weather starts cooling off. As we can see, the dates after that still remain above threshold, which indicates that cooling likely continues. In connection, the following plot of the average daily temperature across all years offers a nice visual companion to the control chart above:

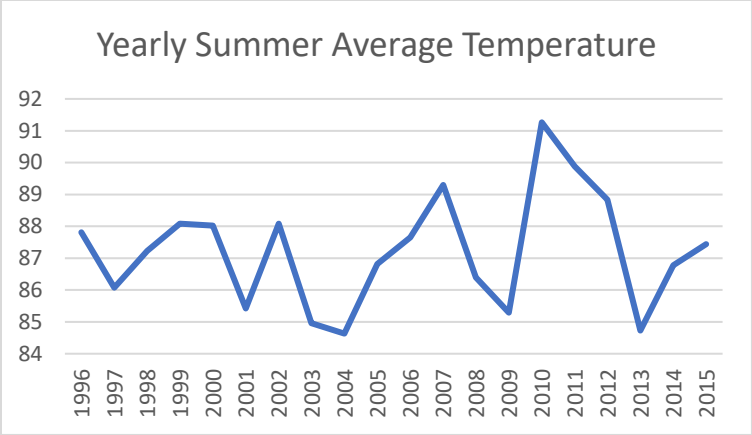


2. Use a CUSUM approach to make a judgment of whether Atlanta's summer climate has gotten warmer in that time (and if so, when).

   The data and approach for this problem can be found in the attached Excel document in sheet 6.2.2. For this question, I began by getting the mean of the summer months (so from July 1st to when I calculated my end of unofficial summer in the previous question). From there, I also calculated the standard deviation of that time period. Initially, I intended to use a C value of 0.5 times the standard deviation; however, that value was still too large and so I opted to use 0.5 instead. I also decided to set my threshold to be 0.5 times the standard deviation obtained. The next step I needed to do was set up an S_t equation to detect an increase as I wanted to see if the summer climate has gotten warmer over time. The formula used was $S_t = \max\{0, S_{t-1} + (x_t - \mu - C)\}$. I also calculated the average temperature across the summer months that I defined per year. This way I would have the average summer temperatures from 1996 to 2015. Ultimately, once I put everything into the S_t formula and plotted it, I was able to obtain this CUSUM control chart:



   Based on this chart, a change is detected around 2010. For comparison purposes, I also created a plot of yearly summer average temperature:



By comparing the two, it is possible to see that around 2010 there is a general spike in summer temperatures. However, I would not be so confident to say that Atlanta's summer temperature has generally gotten warmer in the time frame described. This is because in the control chart, we see that the S_t values seem to be trending down to below the threshold after 2012. In the yearly summer average temperature chart, we can also see that after 2010, the temperature seems to drop pretty significantly before rebounding to a moderate temperature in 2015. Because of these observations, I would be inclined to say the changes on the control chart were more so due to a temporary temperature increase than the summer climate truly getting warmer.

Appendix

Question 5.1

```
#Load library
library(outliers); library(moments); library(PMCMRplus)

##
## Attaching package: 'PMCMRplus'

## The following objects are masked from 'package:outliers':
##
##     pcochran, pgrubbs, qcochran, qgrubbs

#Load data
data <- read.table("C:\\Users\\User\\OneDrive\\Desktop\\Data 5.1\\uscrime.txt
", stringsAsFactors = F, header = T)

#Test for normality
shapiro.test(data[,16])

##
##  Shapiro-Wilk normality test
##
## data:  data[, 16]
## W = 0.91273, p-value = 0.001882

#Plot the data and its Q-Q plot
plot(data[,16])

qqnorm(data[,16], pch=1, frame=FALSE)

#Create a density plot for the data to visualize it
#Also calculate outliers, skew, and kurtosis of the data
plot(density(data[,16]))

skewness(data[,16])

## [1] 1.08848

kurtosis(data[,16]) #Leptokurtic

## [1] 3.943658

outlier(data[,16])

## [1] 1993

#Perform the Grubbs' Test
outliers <- grubbs.test(data[,16], type = 11)
outliers

##
##  Grubbs test for two opposite outliers
##
```

```
## data:  data[, 16]
## G = 4.26877, U = 0.78103, p-value = 1
## alternative hypothesis: 342 and 1993 are outliers

#Perform with type=10 since in the density plot, we do notice a right skew
outliers1 <- grubbs.test(data[,16], type = 10)
outliers1

##
##  Grubbs test for one outlier
##
## data:  data[, 16]
## G = 2.81287, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier

#Remove the largest data point and try again
outliers2 <- grubbs.test(data[-26,16], type = 10)
outliers2

##
##  Grubbs test for one outlier
##
## data:  data[-26, 16]
## G = 3.06343, U = 0.78682, p-value = 0.02848
## alternative hypothesis: highest value 1969 is an outlier

#Create a box-and-whisker plot for the data
boxplot(data[,16], ylab="Number of Crimes per 100,000 people" )

#Check what are the outliers from the boxplot
boxplot(data[,16], plot=FALSE)$out

## [1] 1969 1674 1993

#Because based on the plot, this dataset had more than one outlier
#I decided to use the GESD to investigate
a <- gesdTest(data[,16], 7)
a

##
##   GESD multiple outlier test
##
## Nr. of outliers tested:
##      p-value
## 1 0.15774973
## 2 0.05695641
## 3 0.35615936
## 4 0.22779846
## 5 0.21631147
## 6 1.00000000
## 7 1.00000000
```

```
##
## alternative hypothesis:  two.sided

summary(a)

##
##   GESD multiple outlier test
##
## Outliers tested:
##                    i        R Pr(>|R|)
## 1                 26 2.812874 0.157750
## 2               26 4 3.063425 0.056956 .
## 3             26 4 11 2.564571 0.356159
## 4           26 4 11 2 2.685609 0.227798
## 5         26 4 11 2 8 2.691071 0.216311
## 6     26 4 11 2 8 36 1.871331 1.000000
## 7  26 4 11 2 8 36 27 1.852226 1.000000
##
## alternative hypothesis:  two.sided
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```