Question 5.1

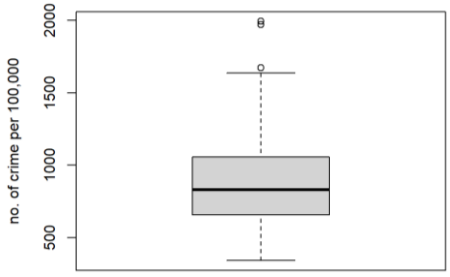
Using crime data from the file `uscrime.txt` (http://www.statsci.org/data/general/uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), test to see whether there are any outliers in the last column (number of crimes per 100,000 people). Use the `grubbs.test` function in the `outliers` package in R.

Answer:

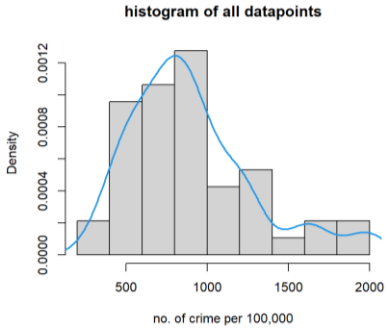
In this exercise, I plotted the data to see if there are any visible outliers, tested the data if they are normally distributed, and conducted Grubbs test by eliminating the candidate data points for the outliers.

First, I checked the quantiles, max and min value of the crime column and plotted the data in the box plot. In the boxplot, we could observe two data points with one of the highest values that seem like outliers.

```
> summary(uscrime$Crime)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  342.0   658.5   831.0   905.1  1057.5  1993.0
```



Second, as Grubbs' test assumes approximately a normal distribution and aims to find a single outlier, we need to check if the dataset is normally distributed in advance. I tried plotting histogram and ran Shapiro-test to test normal distribution. The histogram shows that the dataset is skewed to the right, and as the p-value of the Shapiro test is less than 0.05 (alpha), we reject the H0 and take H1, where we assume that the dataset is not normally distributed.




```
> shapiro.test(uscrime$Crime) ## rejects H0 that it is normally distributed

        Shapiro-Wilk normality test

data:  uscrime$Crime
W = 0.91273, p-value = 0.001882
```

As the dataset is skewed to the right, I assumed that the highest value is the outlier. So, I deleted the highest value and ran Shapiro test again, and the results suggests that the dataset is not normally distributed.

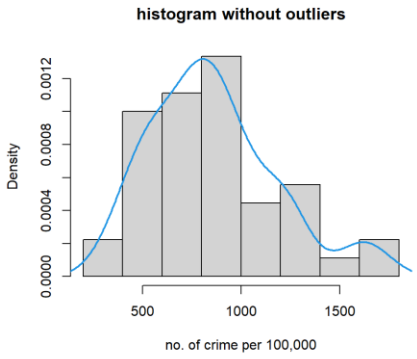
```
> ### get rid of the higest value
> uscrime_2 <- uscrime[-which.max(uscrime$Crime),]
> shapiro.test(uscrime$Crime)

        Shapiro-Wilk normality test

data:  uscrime$Crime
W = 0.91273, p-value = 0.001882
```

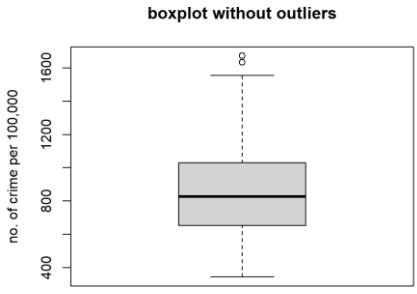
I deleted the highest value again, and this time got the p-value of 0.05634 from the Shapiro test. Depending on the alpha we choose, for example if we choose 0.1, we have the possibility to accept H0 that the dataset is normally distributed. Therefore, I ran Grubb's test, where we fail to reject H0 that there are no outliers in the data. Therefore, the outliers were the two highest data, 1993 and 1969.

When I ignored the normal distribution assumption and ran the dataset with Grubb's test twice (full dataset and without 1993), it showed the same result where I could reject H0 and accept the alternative hypothesis that the highest value (1993 and 1969 each) is an outlier.

```
> ### again get read of the highest value and try the Grubbs' test
> uscrime_3 <- uscrime_2[-which.max(uscrime_2$Crime),]
> shapiro.test(uscrime_3$Crime)

        Shapiro-Wilk normality test

data:  uscrime_3$Crime
W = 0.95119, p-value = 0.05634

> grubbs.test(uscrime_3$Crime, type = 10) # test if max value is outlier

        Grubbs test for one outlier

data:  uscrime_3$Crime
G = 2.56457, U = 0.84712, p-value = 0.1781
alternative hypothesis: highest value 1674 is an outlier
```

Here is the final histogram and box plot for the new dataset without outliers. We can see the difference from the first box plot, that now the highest data points are much closer to the other data points.





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?

Answer:

CUSUM can be very useful for environmental monitoring, because it detects change very fast. One of the examples of application area for CUSUM can be detecting increase in the air pollution level, such as carbon monoxide or nitrogen dioxide concentrations. The baseline standards can be decided based on a long-term trend, in order to be independent from the annual fluctuations from the effect of meteorology. Critical value should consider the harmful level as well as the unexpected fluctuation, and we could consider seasonal max. and min. value for the unexpected fluctuation. For deciding the threshold, standard deviation can be considered, and the level of concentration that scientifically proven or controlled by regulations, which mark the level that can be a threat to human health or environment.

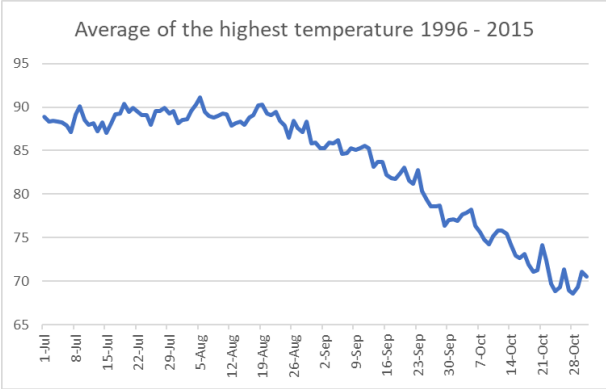
Question 6.2

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.

Answer:

For this exercise, I had to find a base average mu, and a good value for C and T. Also, I tried to find the unofficial summer ends for each year. For the analysis tool, I used Excel spreadsheet.

First, in order to find a good mu, C and T value, I found the average of the highest temperature of each day from year 1996 to 2015. For mu, I thought July can be a good baseline where we can assume as there will be no change, as it is in the mid of summer. July highest temperature has an average 89°F, with standard deviation 0.87, min 87 and max 90.



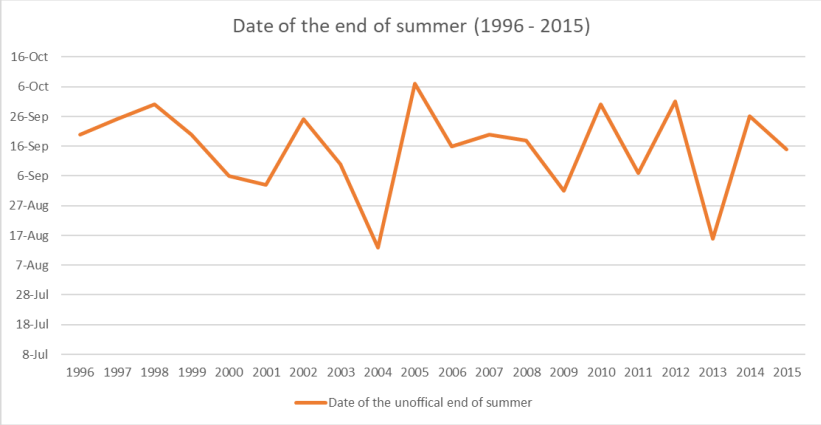
In order to find a good C value, we have to consider that we can have anomalies while we are still in the same season. I started C value from 2.6, which is 3 times the standard deviation in July.

For T value, I tried to find when the highest temperature begins to drop from the average of the highest temperature 1996 – 2015 line graph. The temperature started to decrease more rapidly from September in the graph, which starts with the value about 5 degree less from our mu (89). So, for the initial C and T value, I set c = 2.6, T = 5 and tried CUSUM on each day of the average of the highest temperature from year 1996 to 2015, using $S(t) = \max \{ 0, S(t-1) + (mu - x(t) + C \}$.

The result is as in the table below: I encountered false alarm where the temperature went up to 89°F for several days after the summer ends, as well as belated alarm where I had already several days below 89°F before the alarm. The best suggestion after some trial was C=5, T = 25.

C	T	average date of end of summer	temp. of the day end of summer	Comments
2.6	5	28-Jul	90	false alarm
2.6	10	29-Jul	89	false alarm
3	5	28-Jul	90	false alarm
3	10	29-Jul	89	false alarm
3	15	3-Sep	86	false alarm
5	10	3-Sep	86	false alarm
5	15	3-Sep	86	false alarm
5	20	20-Sep	83	false alarm
5	25	21-Sep	82	no but attention on 2013
7	10	3-Sep	86	false alarm
7	15	23-Sep	83	not sensitive enough

As a result, here is a line graph for the dates of the unofficial summer ends for 1996 and 2015. From the data, we found out that the unofficial summer ends in 2004 and 2013 were exceptionally early.



1996	20-Sep
1997	25-Sep
1998	30-Sep
1999	20-Sep
2000	6-Sep
2001	3-Sep
2002	25-Sep
2003	10-Sep
2004	13-Aug
2005	7-Oct
2006	16-Sep
2007	20-Sep
2008	18-Sep
2009	1-Sep
2010	30-Sep
2011	7-Sep
2012	1-Oct
2013	16-Aug
2014	26-Sep
2015	15-Sep

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).

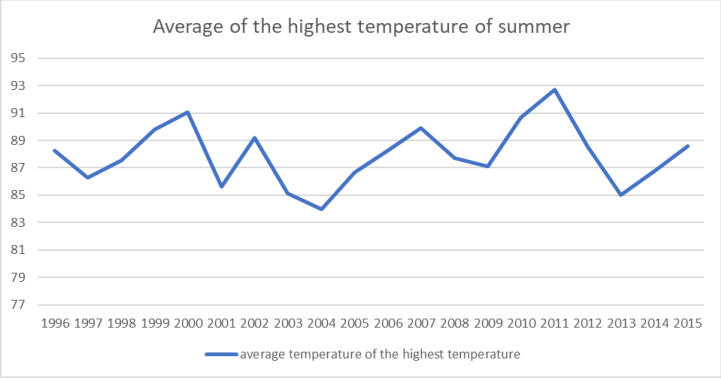
Answer:

For this exercise, I calculated average highest summer temperature for the "unofficial summer period" from July that we found from 6.2.1 for each year. The average of the summer highest temperature was 88, with max 93, min 84 and std 2.16.

I made some candidates starting with C= 0 and T = 5 (about 2 times the standard deviation). Then, in order to detect increase using S(t) = max {0, S(t-1) +(x(t) - mu + C}, I conducted CUSUM on the average for year 1996 to 2015.

The result does not clearly show a pattern of increase in the summer temperature. As in the table, I could observe an alarm for increase in 2011, but as you can see in the graph, the average temperature plummeted in the year 2013. Also, for other C and T values, I could not detect increase at all.

C	T	year with change observed	temp. value
0	5	2011	93
0	7	2011	93
0	10	-	-
3	5	-	-



Reference

https://www.statisticshowto.com/grubbs-test/#:~:text=Grubbs'%20test%20is%20used%20to,(excluding%20the%20potential%20outlier).

Lang, P. (2020). New approaches to the statistical analysis of air quality network data: insights from application to national and regional UK networks (Doctoral dissertation, University of York).