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:
Given below are the steps I performed to calculate outliers in the USCRIME dataset.

Step 1: Load all the Libraries required

CODE:
```
library(data.table)
install.packages("outliers")
library(outliers)
install.packages("histogram")
library(histogram)
install.packages("shapiro")
library(shapiro)
set.seed(42)
```

Step 2: Load the uscrime.txt dataset

CODE:
```
uscrime = read.table("D://MS Georgia Tech/Introduction to Analytics/HW3/uscrime.txt", header = TRUE, sep = '\t')
```

Step 2: Running shapiro normality test on the crime column of the uscrime dataframe

CODE:
```
shapiro.test(uscrime$Crime)
```

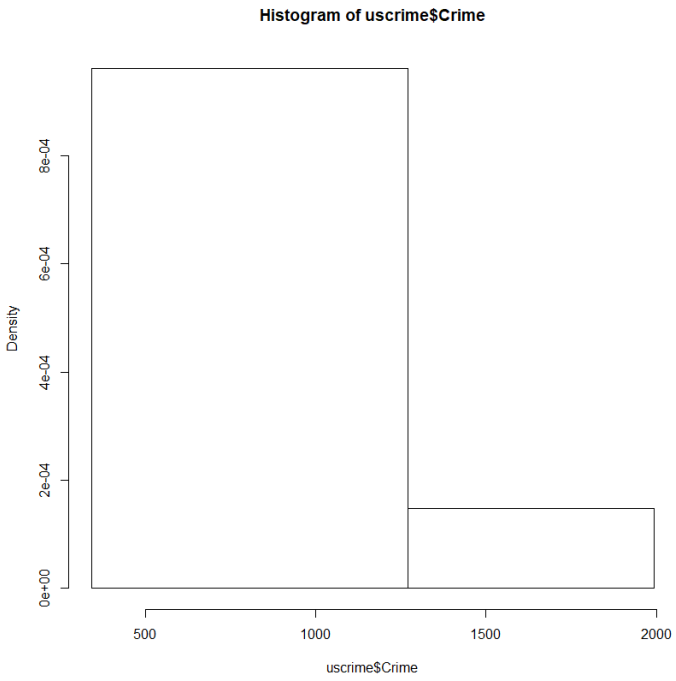
OUTPUT:
```
        Shapiro-Wilk normality test

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

ANALYSIS: With a very low p-value of 0.001882 (lower than 0.05 threshold) suggests a very strong evidence of non-normality in the data i.e. data appears to be skewed as shown in the density histogram below

CODE:
```
histogram(uscrime$Crime)
```

OUTPUT:



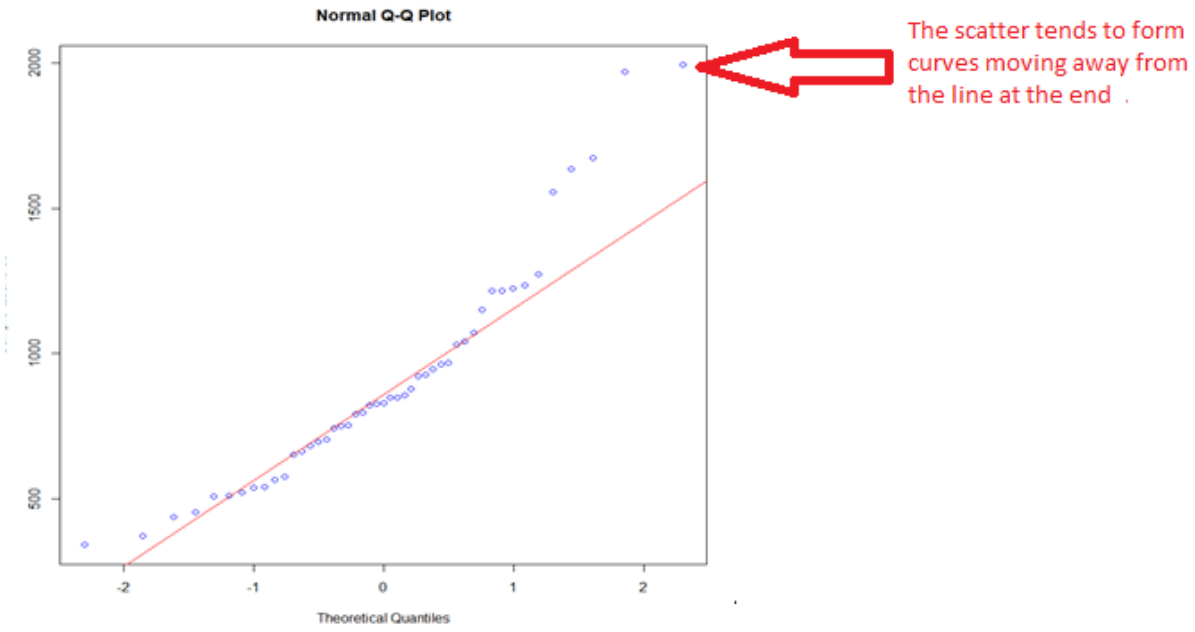
The density vs US$Crime Histogram suggests that the data is skewed towards right. SO the outliers appear more on the right tail.

Lets have a look at the normal QQ – PLOT and assess the normality

CODE:

qqnorm(uscrime$Crime, col = "blue")
qqline(uscrime$Crime, col = "red")

OUTPUT:



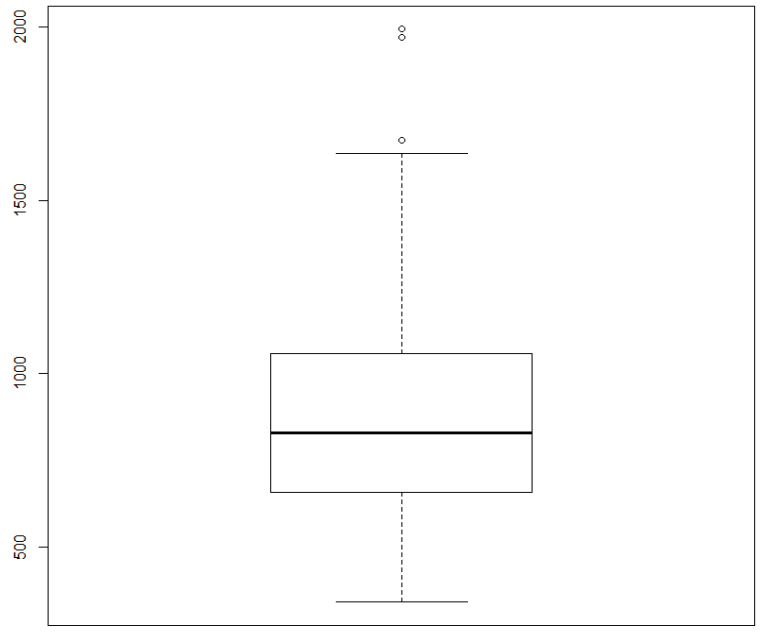
The scatter points in the right end moves away from the line which makes the data fail normal distribution. So those two points at extreme end are outliers. The left tail does not seem to have outliers.

Lets confirm the above theory using Box and Whisker plot.

CODE:

```
boxplot(uscrime$Crime)
```

OUTPUT:



The Box and Whisker plot above shows that there are higher values of crime which act as outlier and there are no outliers for lower values. There is a total of 3 outliers.

Step 4: Outlier Detection using Grubbs Test

Based on the analysis in test Step 3, we need to perform Grubbs test on the data which has outlier on one tail.

Type = 10 in Grubbs.Test formula is a test for one outlier (side is detected automatically and can be reversed by opposite parameter). 20 is test for two outliers in one tail. 20 can only be used when we have a data set less that 31 values.but we have a data set of 47 records.

So we use Type = 10

CODE:

grubbs.test(uscrime$Crime, type = 10)

OUTPUT:

```
        Grubbs test for one outlier

data:  uscrime$Crime
G = 2.81290, U = 0.82426, p-value = 0.07887
alternative hypothesis: highest value 1993 is an outlier
```

ANALYSIS: Looking at the results of Grubbs test 1993 appears to be an Outlier on the higher end (right tail) and a p-value of 0.07887 is greater than the threshold 0.05 which makes the null hypotheses fail. Since the Grubbs test is pulling one outlier at a time where Type = 10, we can run the Grubbs test 3 times to detect all the three outliers using the loop.

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:

I would like to give an example of Dam in my village for which change detection model is appropriate. The village along with several other villages is situated on the banks of a river. The government there has built a Dam which controls the flow of the water in the river. It is because of this dam, the waters do not flow fast and the villagers can benefit for the agriculture process. The level of water there is measured between 0 -1000 units where if water level reaches 900 units the dam has to release water on the other side or else it can impact the ability of the dam to hold water. Level 800 is still not good for the villagers as well as it may end up flooding the village even though the dam can handle that level. So the government will still release the waters at 800 units. 700 units is the best scenario with no damage on either side with surplus water for the villagers. So releasing the DAM at 700 is advisable as no one is affected at 700. The place has 4 months of monsoon from June to September where change detection is critical. This model is not supposed to be very sensitive as the probability of very heavy rains were considered while building the Dam. thOn a normal day when there is no heavy rain, water level stays between 400-500 units.

A change detection method is critical here when there is heavy rain detected. My data for the change detection model will be water levels in monsoon for last 15 years. When there is a heavy rain it reaches up to 800. So in my estimation my MU will be somewhere around 600 Units per day during monsoon. My SD would be around 100 Units. This model is not supposed to be very sensitive so I would use a high value of C and set T = 700 and make it difficult for the St to touch 700.

6.1.1 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.

Temps.xlsx

Answer. CUSUM method was used on the temp.txt dataset on spreadsheet embedded along with this file.

CUSUM formula to detect change in decreasing temperature. The formula used was
St = MAX(0,St-1(mu-xt-C) where

MU = Mean taken for the summer period which is estimated up to 31st August.

Sd = Standard Deviation for a sample of temperatures up to 31st August.

T = 5 * Average(sd)

C = 0.5 * Average(sd)

The part highlighted in blue represents data where St >= T

Given below is the chart the output table that provides cooling of period each year from 1996 to 2015.

Year	Cooling Off Period
1996	28-Jul
1997	23-Sep
1998	9-Aug
1999	19-Sep
2000	2-Sep
2001	2-Sep
2002	30-Aug
2003	8-Sep
2004	10-Aug
2005	6-Oct
2006	31-Aug
2007	17-Sep
2008	14-Aug
2009	31-Aug
2010	27-Sep
2011	5-Sep
2012	7-Aug
2013	24-Sep
2014	24-Sep
2015	29-Aug

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: When you look at the embedded excel, there are two sheets for this question. They are named as

1. Q6.2.2 Increase – Tells you if there is an increase in temperature each year.
2. Q6.2.2 Decrease – Tells you if there is a decrease in temperature year.

Based on the change in temperature observed in question 6.1.1, it safe to assume that on 5th September the winter has started for most of the years (14 out of the 20 years). So time before 5th September can have a period of summer. So our Dataset 6.2.2 has data up to 4th September.

When you look at both Increase and Decrease sheet, you will see thy complement each other. i.e. when one sheet says there is an increase the other says there is a decrease for a particular year.

After comparing two sheets, the table below gives a high level analysis for each year.

Year	Analysis
1996	There is no increase in weather but a clear decrease starting 27th July till 4th of September
1997	Shows a pattern of decrease starting 1st Aug with small changes. But no pattern of increase
1998	Clear pattern of decrease starting 2nd August.No Pattern of increase.
1999	Weather gets warmer this year starting 23 July onwards.
2000	Clear pattern of decrease starting 10th August. No Pattern of increase.
2001	The weather starts to decrease from 2nd September with no pattern of increase.
2002	Clear pattern of decrease starting 30th August. No Pattern of increase.
2003	Weather gets warmer this year starting 29 Aug onwards.
2004	Clear pattern of decrease starting 10th August. No Pattern of increase.
2005	Warmer period between 26th Jul - 29 July and 21-Aug -24th Aug.
2006	Mixed pattern of increase and Decrease. Warmer period between 3-Aug to 14th Aug.
2007	Weather gets warmer this year starting 3rd Aug onwards.
2008	Clear pattern of decrease starting 14th August. No Pattern of increase.
2009	Mixed pattern of increase and Decrease. Warmer period between 9-Aug to 26th Aug.
2010	Weather gets warmer between 26th July and 30th Aug.
2011	Weather gets warmer on 3rd and 4th Sep.
2012	Clear pattern of decrease starting 9th August. No Pattern of increase.
2013	Mixed pattern of increase and Decrease. Warmer period between 11-Aug to 14th Aug.
2014	Weather gets warmer on 4th September.
2015	Mixed pattern of increase and Decrease. Warmer period between 4-Aug to 16th Aug.

Summary

1. Atlanta's summer has gotten warmer in the years 1999, 2003, 2005, 2007 and 2010.
2. Year 2006, 2009, 2013 and 2015 have got mixed patterns of increasing weather and decreasing weather.
3. Rest of the years not mentioned in above two points has shown decreasing patterns which make sense as they are heading towards winter.