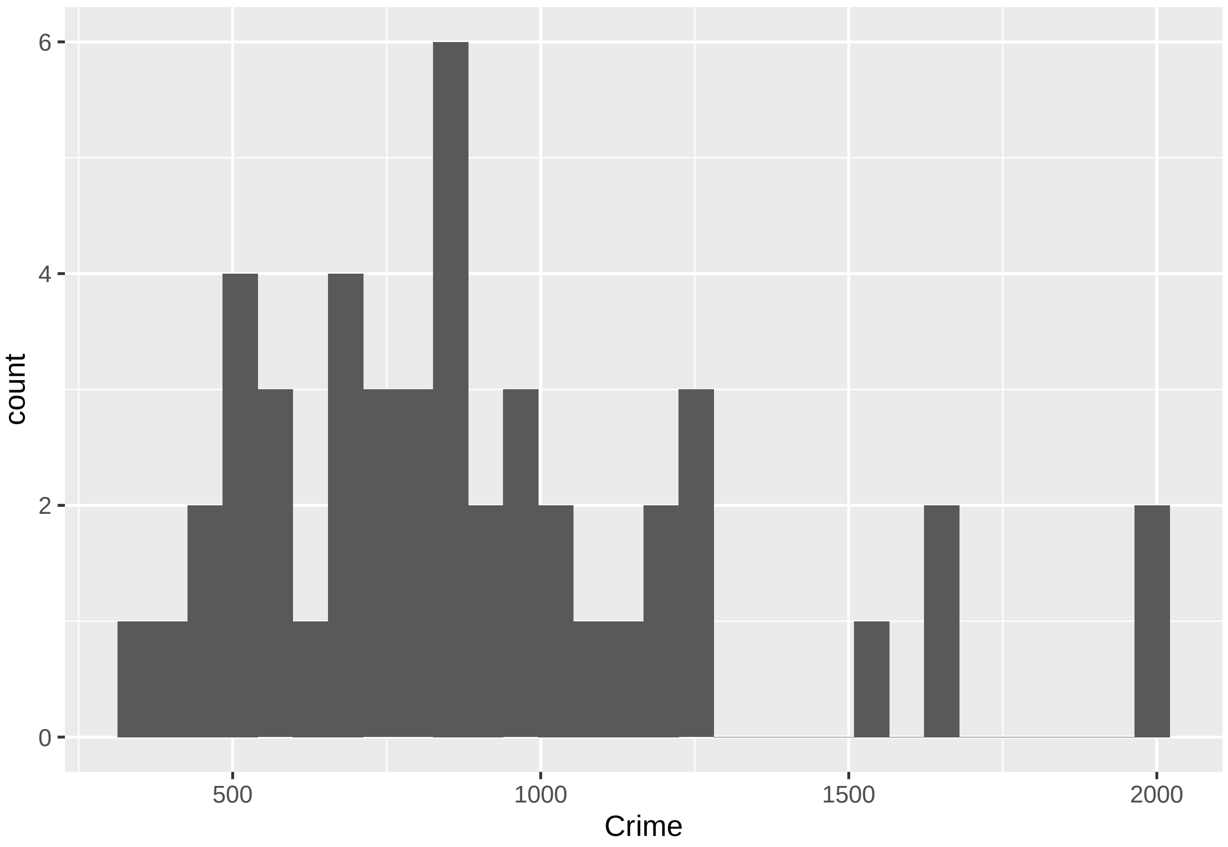
Assignment 3

Question 5.1

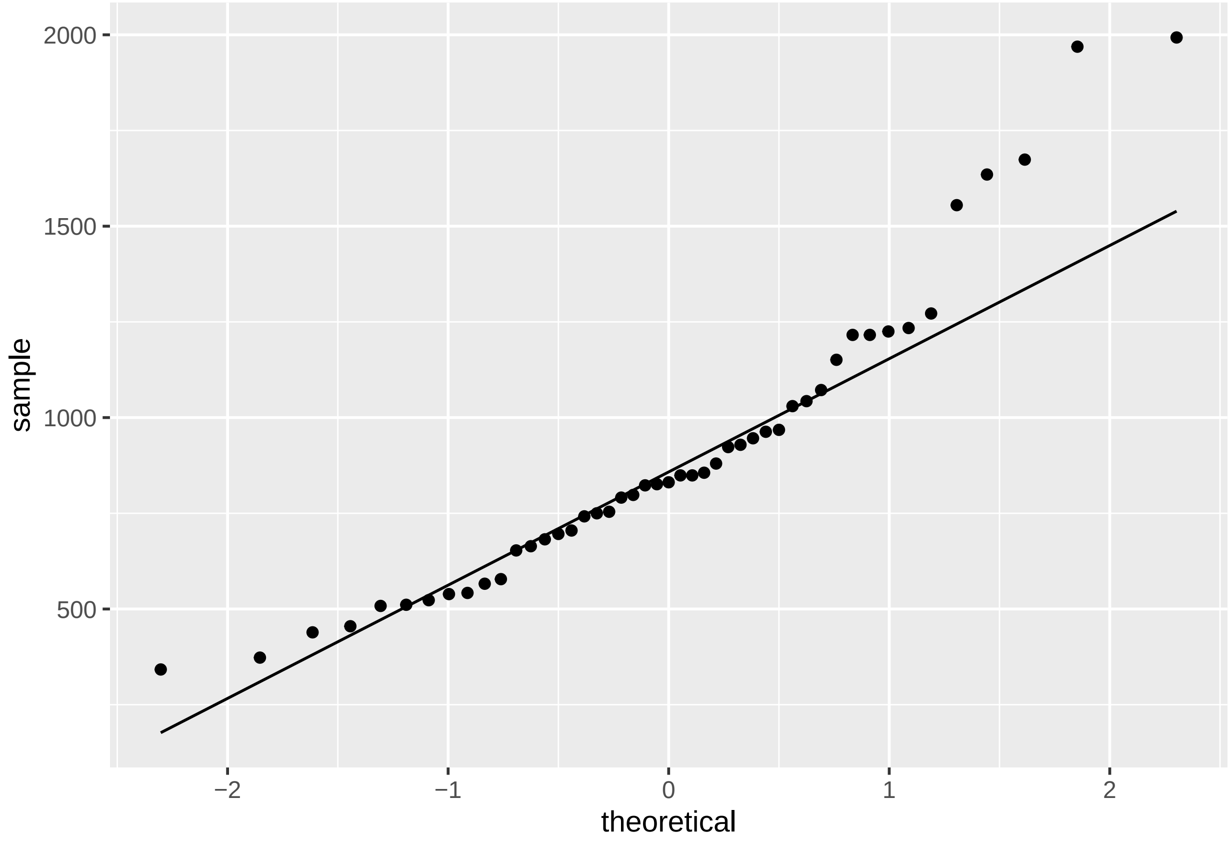
Using crime data to see whether there are any outliers in the last column (number of crimes per 100,000 people). Use the grubbs.test function in the outliers package in R.

Answer 5.1

```r
library(outliers)
library(ggplot2)

#Getting the data
data<-read.table('Crime Data.txt', sep = "", header = TRUE)

#Testing if the dataset (Crime rate) in question is normally distributed before
#applying grubbs test. Normality is tested using Histogram, Q-Q Plot,
#Shapiro-WilK test

#Histogram
ggplot(data=data, aes(Crime)) + geom_histogram()
```



```
#Q_Q Plot
ggplot(data, aes(sample=Crime))+ stat_qq() + stat_qq_line()
```

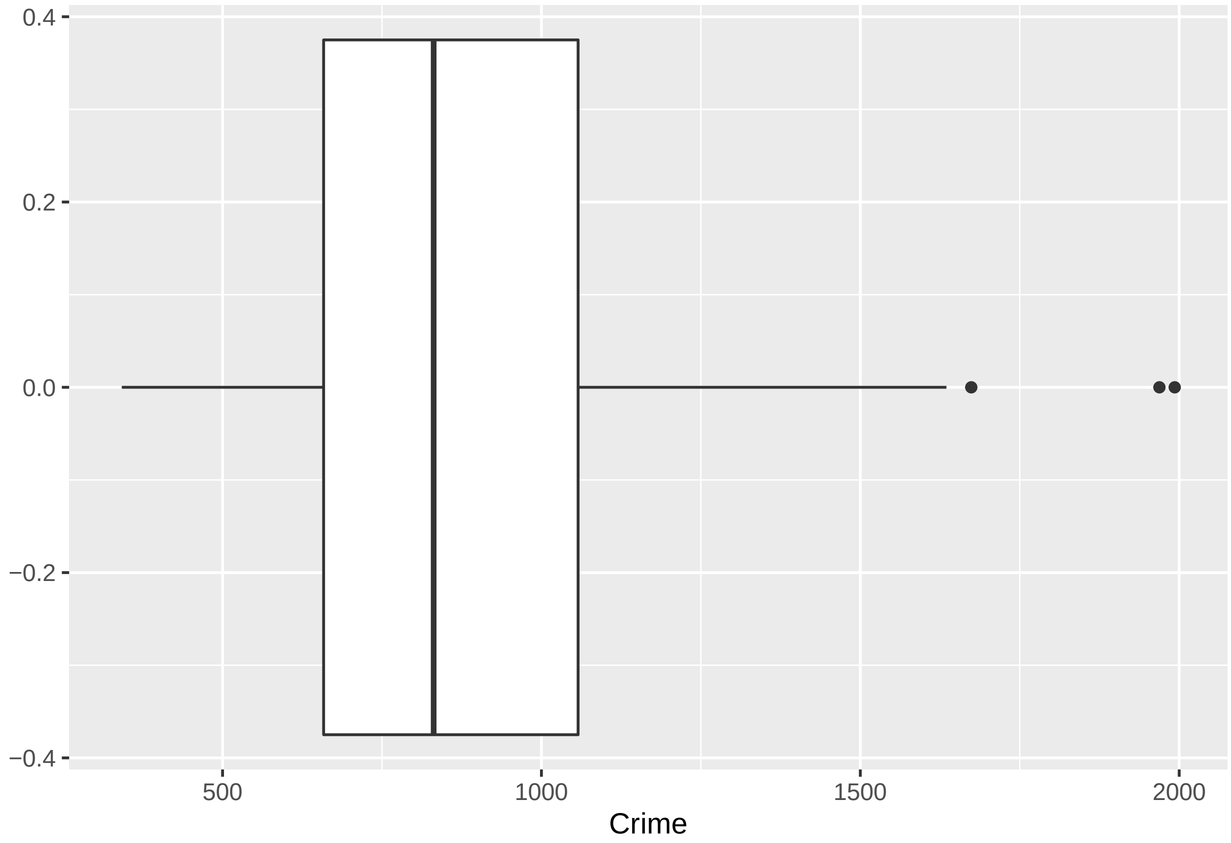


```
#Shapiro-WilK test

#Stating the hypothesis, threshold alpha=0.05 (significance level)
#H0: The data is normally distributed
#H1: The data is not normally distributed.

#Test 1
norm_test<-shapiro.test(data$Crime)
norm_test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Crime
## W = 0.91273, p-value = 0.001882
```

```
#Since p value for the test is less than than the chosen threshold the null
#hypotheses is rejected and the data used in this test is not normally
#distributed. A log normal transformation can be achieve normality or by
#removing the outliers as Shapiro test is sensitive to sample size, meaning if
# sample size is sufficiently large this test may detect even trivial departures
```

```
#from the null hypothesis. So we will assume the data to to normally distributed
#as also evident from the Q-Q plot where most of the data lies in a st. line.

#Outlier detection using Box plot

ggplot(data,aes(Crime))+geom_boxplot()
```



```
#The box plot shows that there might be few points as outliers. We can now
#perform Grubbs Test

#Using Grubbs test (alpha(sigificance level)=0.05)
test1<-grubbs.test(data$Crime, type = 10, opposite = FALSE,
                   two.sided = FALSE)
test1
```

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

```
#test1 show point 1993 is an outlier is an outlier as is close to alpha.Next we will remove
#this point and continue to investigate if we have more outliers.
```

```
#Removing 1993 data
data2<-data[-which.max(data$Crime),]

test2<-grubbs.test(data2$Crime, type = 10, opposite = FALSE,
                   two.sided = FALSE)
test2
```

```
##
##  Grubbs test for one outlier
##
## data:  data2$Crime
## G = 3.06343, U = 0.78682, p-value = 0.02848
## alternative hypothesis: highest value 1969 is an outlier
```

```
#test1 show point 1969 is an outlier as is close to alpha.Next we will remove
#this point and continue to investigate if we have more outliers.

#Removing 1969 data
data3<-data2[-which.max(data2$Crime),]

test3<-grubbs.test(data3$Crime, type = 10, opposite = FALSE,
                   two.sided = FALSE)
test3
```

```
##
##  Grubbs test for one outlier
##
## data:  data3$Crime
## G = 2.56457, U = 0.84712, p-value = 0.1781
## alternative hypothesis: highest value 1674 is an outlier
```

```
#test1 show point 1664 is not an outlier as p is much greater than alpha.
#We have now removed all the outliers
```

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 5.1

We use sensors to detect pump failures using vibration data. Thus, it is very essential for us to know if the pump need any service before they breakdown and cause huge losses. The CUSUM method can be used here to detect early if the pump need to undergo some servicing based on when the vibration reaches a critical value. The critical value can be determined by the threshold of failure given by the manufacturer. A reasonable cut of value can be decided to detect a change and proper action can be taken.

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 youneed from the file temps.txt. You can use R if you'd like, but it's straightforward enough that an Excel spreadsheet can easily do the job too.

Answer 6.2.1: Please refer to the excel file for the graph and calculations. Temperature data tab for understanding the data and building the initial intuition and the 6.2.1-CUSUM tab for calculation

For the first problem I started with the raw data to build the intuition around when the temperature starts to drop and get an idea about the summer end time. To do this I used plotted the temp data against time with the temp data on log scale (Fig 1). From the figure it was deduced that the temperature starts to fall somewhere around the first to second week of September. We will now use this information to build our CUSUM model.

The first step is to calculate mu and SD: It was calculated using all the temperature data until 9th of Sep. The initial value of C and T is calculated by the using the rule of thumb with C half of SD and threshold is 5 times the SD. Beyond this different C and T values are used to establish the intuition that summer temperatures starts decreasing in September with the summers official ending in September. The data is analyzed using the conditionally formatting tool in excel.

The results are summarized in the table in the excel file with the value of C=4 and T=45 for this analysis. The analysis itself is really open ended and can have many solutions. So with this it can be concluded that the temperature starts dropping for most of the years from the first and second week of September and with summers officially ending in September except for one the year 2005 which does not comply with the model results. Further looking into the data it shows that this year (2005) had the longest spell of hot weather.

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).

Answer 6.2.2

- Approach 1: Basic data analysis

This problem was started by analyzing the average of the temperatures for the summer months (July, Aug and Sep) across all the years to find answers if the climate has gotten warmer by analyzing the average temperature trends.

Refer to Fig 1 in the 6.2.2 Average Approach tab in the excel file for this problem. No solid conclusions can be drawn from this chart as the average temperatures have been fluctuating and its difficult to say if the climate has gotten warmer.

- Approach 2: CUSUM (Positive)

To goal is to find if the climate is getting warmer. To do this analysis µ was calculated using the first five years for the summer months (Jul, Aug and Sep). CUSUM was applied to see if there is an increasing trend if the climate is getting hotter. Again conditional formatting was used to determine this trend. Ideal situation would have been where we could see a consistent trend of increase or neutral trend in the temperature trend to come up with a conclusion. Using different combinations of C and T no consistent trend could be established and thus the analysis could not reach to a confident conclusion whether the climate has gotten warmer in Atlanta.

Table results are summerized in the excel.