Homework-3

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.

Import Packages

```
library(outliers)
```

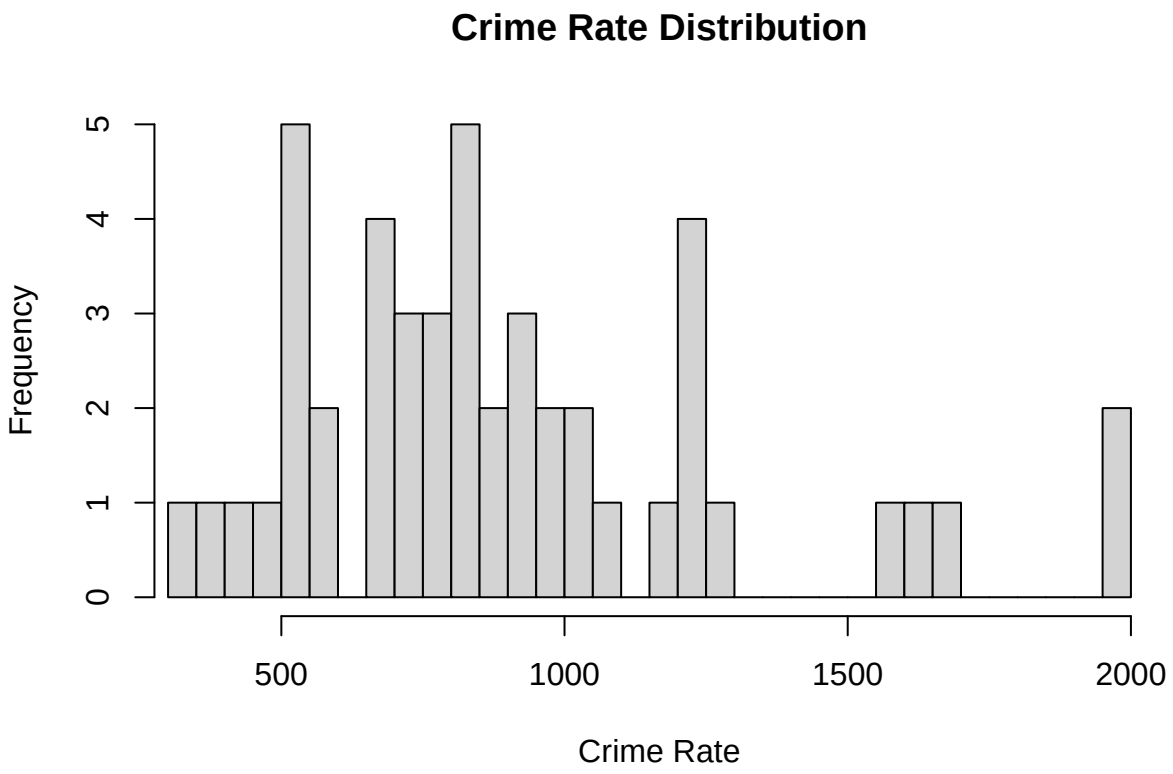
Load Data

```
data = read.table("C:/Users/Admin/Desktop/MM/Homework 3/uscrime.txt",
                  stringsAsFactors = FALSE,
                  header = TRUE)
head(data)
```

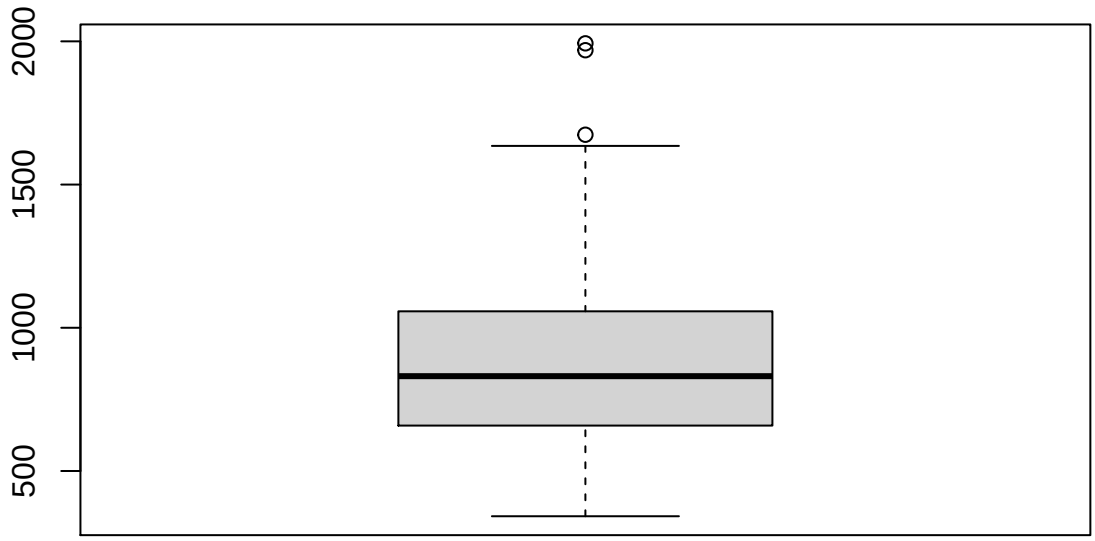
```
##      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq     Prob
## 1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1 0.084602
## 2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4 0.029599
## 3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0 0.083401
## 4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7 0.015801
## 5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4 0.041399
## 6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6 0.034201
##      Time Crime
## 1 26.2011   791
## 2 25.2999  1635
## 3 24.3006   578
## 4 29.9012  1969
## 5 21.2998  1234
## 6 20.9995   682
```

Visualize Data Distribution

```
# Histogram
hist(data['Crime'],
     main = 'Crime Rate Distribution',
     xlab = 'Crime Rate',
     breaks = nrow(data),
     xlim = c(min(data['Crime']),
              max(data['Crime'])))
```




```
# Boxplot
boxplot(data[,'Crime'])
```



2-tailed test

```
# Test for outliers on opposite tails of data distribution
grubbs.test(data[,'Crime'],type=11)
```

```
##
##  Grubbs test for two opposite outliers
##
## data:  data[, "Crime"]
## G = 4.26877, U = 0.78103, p-value = 1
## alternative hypothesis: 342 and 1993 are outliers
```

2-tailed test returns with p-value = 1, meaning null hypothesis of both 342 and 1993 are not outliers is not rejected.

1-tailed test

```
# Testing negative tail end outlier
grubbs.test(data[,'Crime'],type=10, opposite = TRUE)
```

```
##
```

```
##  Grubbs test for one outlier
##
## data:  data[, "Crime"]
## G = 1.45589, U = 0.95292, p-value = 1
## alternative hypothesis: lowest value 342 is an outlier
```

With p-value = 1, point 342 is not considered as an outlier.

```
# Testing positive tail end outlier
grubbs.test(data[,'Crime'],type=10, opposite = FALSE)
```

```
##
##  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
```

Depending on p-value threshold selected (0.1 or 0.05) we could choose to flag data point 1993 as an outlier or not.

```
# If we were to exclude data point 1993 and run the test for outlier again
data.new = data[-which.max(data[,'Crime']),]
grubbs.test(data.new[,'Crime'],type=10, opposite = FALSE)
```

```
##
##  Grubbs test for one outlier
##
## data:  data.new[, "Crime"]
## G = 3.06343, U = 0.78682, p-value = 0.02848
## alternative hypothesis: highest value 1969 is an outlier
```

Data point 1969 with p-value < 0.05 could also be considered as an outlier.

```
# Removing data point 1969
data.new = data.new[-which.max(data.new[,'Crime']),]
grubbs.test(data.new[,'Crime'],type=10, opposite = FALSE)
```

```
##
##  Grubbs test for one outlier
##
## data:  data.new[, "Crime"]
## G = 2.56457, U = 0.84712, p-value = 0.1781
## alternative hypothesis: highest value 1674 is an outlier
```

Data point 1674 with p-value > 0.05 and 0.1, not an outlier according to Grubbs test.

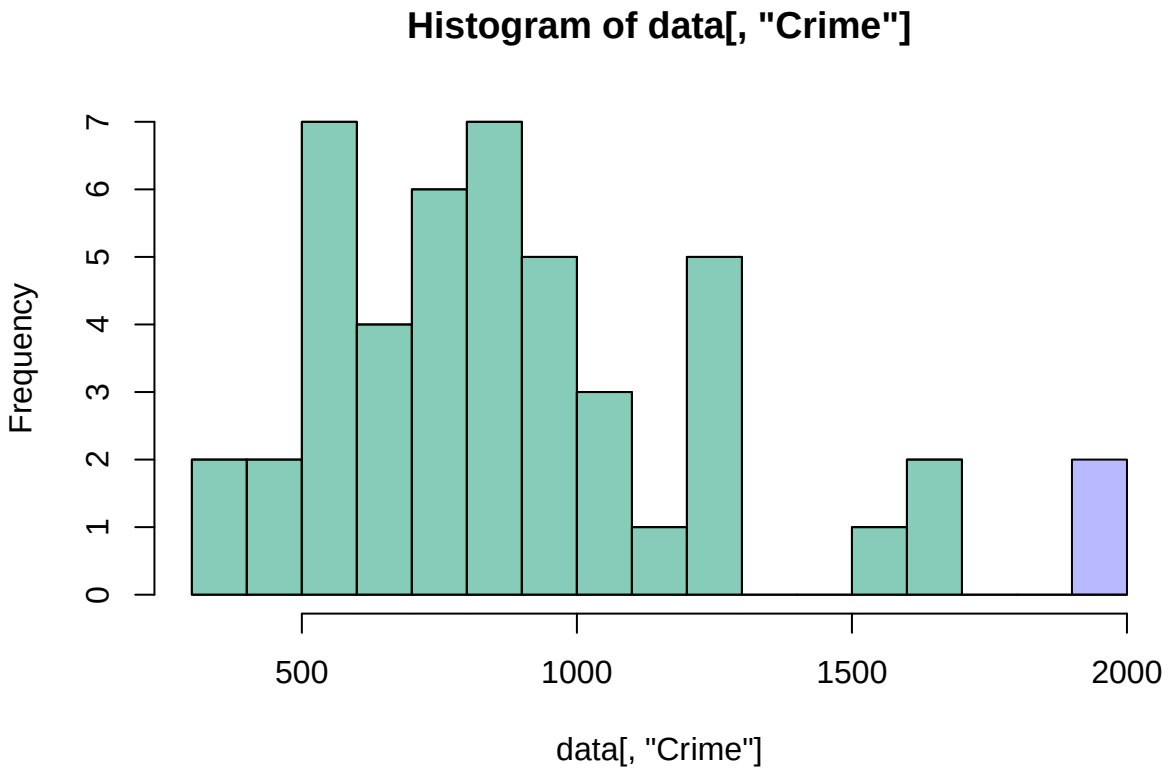
Visualize Data Distribution after removing outlier

```
b = min(c(data[,'Crime'], data.new[,'Crime']))
e = max(c(data[,'Crime'], data.new[,'Crime']))
ax = pretty(b:e, n=12)

hist1 = hist(data[,'Crime'], breaks = ax, plot = FALSE)
hist2 = hist(data.new[,'Crime'], breaks = ax, plot = FALSE)

c1 = rgb(0, 0, 255, max = 255, alpha = 70)
c2 = rgb(0, 255, 0, max = 255, alpha = 70)

plot(hist1, col = c1)
plot(hist2, col = c2, add = TRUE)
```



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?

In banks, cash movement and liquidity in all aspects of businesses are highly monitored to prevent potential bribery, corruption, fraud, terrorism financing, or even rogue trading. Retail banking in particular, it is important to track changes in transfer of funds across accounts, especially overseas. A change detection model in this circumstance would be insightful to help highlight these behaviors. Since there is little to no cost in a false positive detection, and on the other hand huge implications for not detecting such cases, it would make sense for the model to be fairly sensitive to large changes with low threshold level and a lower C value.

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.**
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).**

Please refer to excel sheet.