Project 3

Marcus McKenzie

3.1

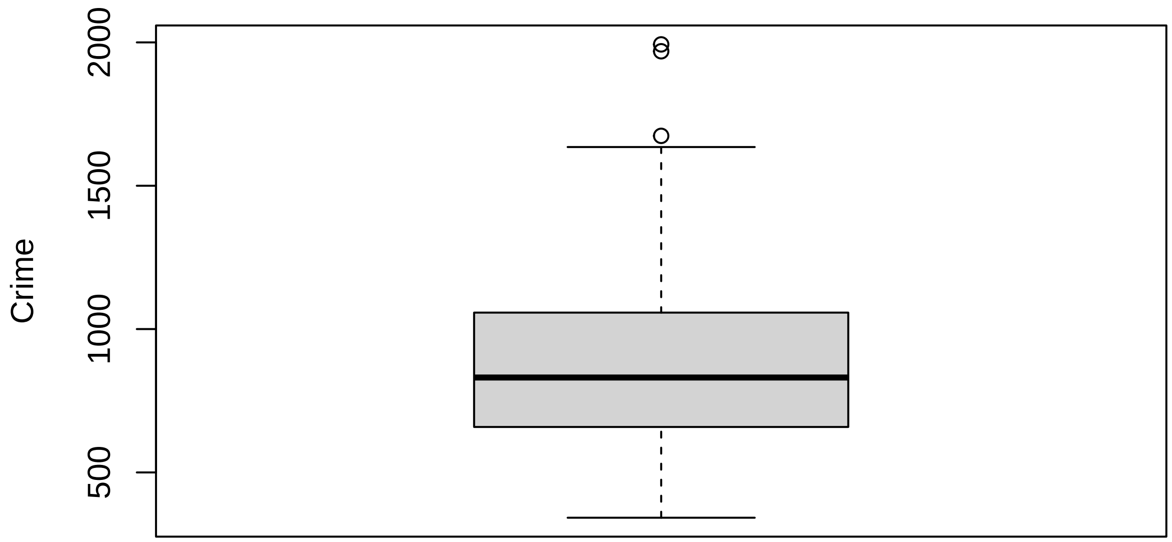
Using crime data test to see whether there are any outliers in the last column (number of crimes per 100,000 people). Use the grubbs.test function in the outliers package in R

Load data:

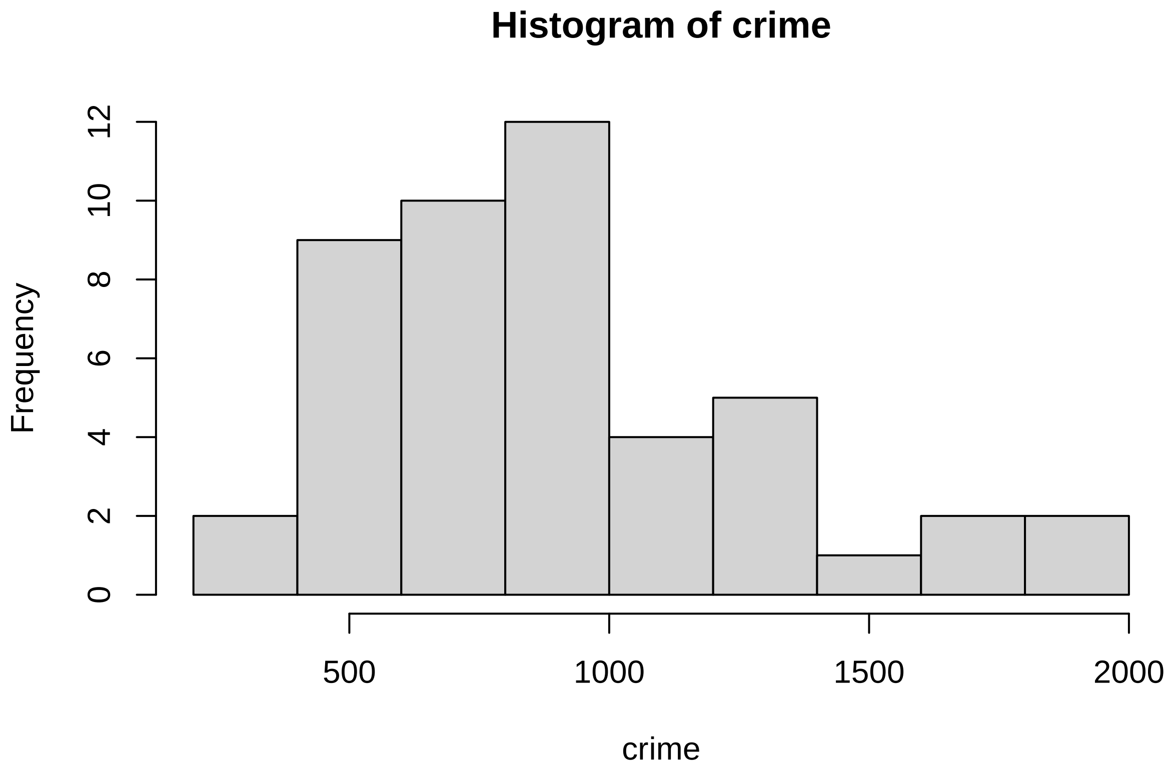
```
library(outliers)

rm(list = ls())

#data <- read.table("Documents/OMSCS/Analytics Modeling/Assignments/Assignment3/uscrime.txt", header= T

data <- read.table("uscrime.txt", header=TRUE, stringsAsFactors = FALSE)


crime <- data$Crime
```

Plot data:

```
boxplot(crime, ylab= "Crime")
```

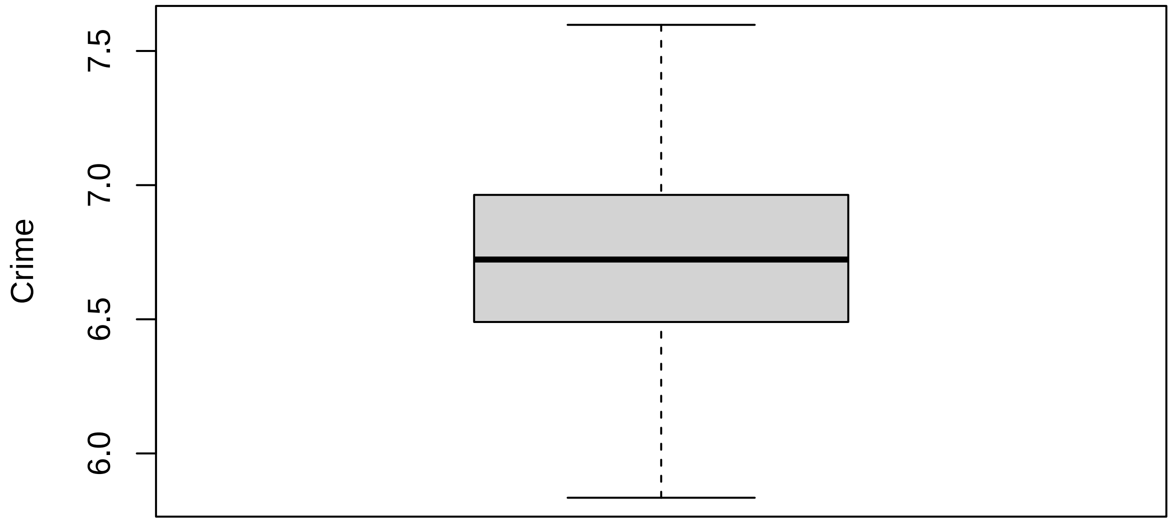


```r
hist(crime, breaks= 10)
```

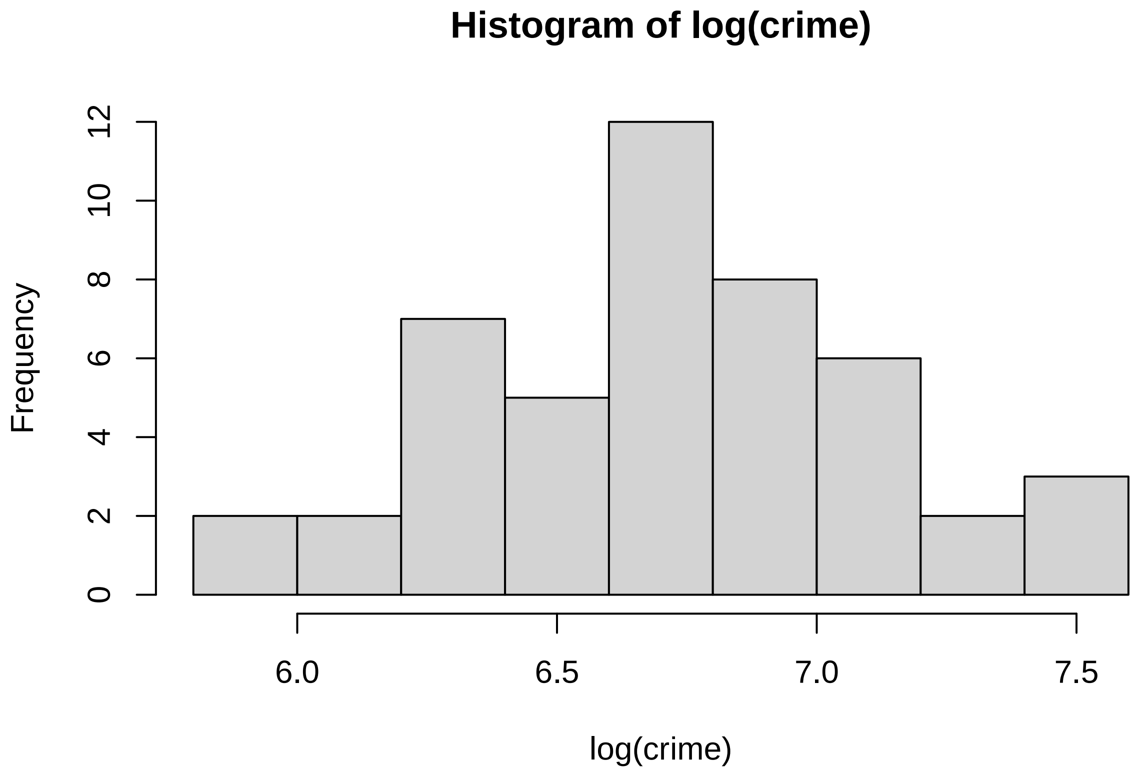


Plot log of data:

```
boxplot(log(crime), ylab= "Crime")
```



```r
hist(log(crime), breaks= 10)
```



Test for normal distribution:

```
shapiro.test(crime)
```

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

```
shapiro.test(log(crime))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(crime)
## W = 0.98709, p-value = 0.8778
```

Implement Grubbs test:

```
grubbs.test(log(crime), type = 10, opposite = FALSE, two.sided = FALSE)
```

```
##
##  Grubbs test for one outlier
```

```
##
## data:  log(crime)
## G = 2.16544, U = 0.89585, p-value = 0.6329
## alternative hypothesis: lowest value 5.8348107370626 is an outlier
```

```r
grubbs.test(log(crime), type = 10, opposite = TRUE, two.sided = FALSE)
```

```
##
##  Grubbs test for one outlier
##
## data:  log(crime)
## G = 2.12247, U = 0.89994, p-value = 0.712
## alternative hypothesis: highest value 7.59739632021279 is an outlier
```

Conclusion

From the crime data we were able to determine that there existed some outliers in the provided crime data. We were also able to demonstrate that the boxplot was fairly accurate but not as precise as the grubbs test. In this case the log of the data was more useful than the original provided data.

3.2

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?

Air pollution could be a great use for a change detection model. It would be useful to analyze how air pollution changes, depending on the day of the week or the month of the year. It would especially be interesting to see how air pollution has changed since the coronavirus, as energy consumption is down. The critical value, C, for this model would have to depend on what constitutes a dangerous amount of pollution to a populaiton. Since false positives are not costly, a lower value may be more ideal. The threshold value in this case may simlply be the average of the more air polluted areas throughout the world.

3.3

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.

Load temperature data:

```r
rm(list = ls())

#data <- read.table("Documents/OMSCS/Analytics Modeling/Assignments/Assignment3/temps.txt", header= TRU

data <- read.table("temps.txt", header=TRUE, stringsAsFactors = FALSE)
```

```r
cusum <- function(d, c=.5*sd(d)){

  u <- mean(d)
  s <- c(0)

  for (i in 2:length(d)){

    s[i] <- min(0, s[i-1] + (d[i] -u -c))

  }

  return(s)

}
```

Create function to select dates:

```
dates <- function(dt, c=.5, t=-5){

  col <- colnames(dt)
  date <- c(0)

  for (i in 2:length(col)){

    d <- data[,i]

    C <- c*sd(d)
    S <- cusum(d,C)

    threshold <- t*sd(d)

    b <- S < threshold
    m <- min(which(b == TRUE))

    day <- data$DAY[m]
    date[i] <- day

  }

  return(date)

}
```

Implement functions:

```
index <- dates(data, c=.1, t=-5)

for (j in 2:length(index)){

  #dt <- droplevels(data$DAY[index[j]])
  #print(dt)

}
```

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).

Initialize data:

```r
cols = colnames(data)
mean_tmp <- c(0)
```

Determine average temps:

```r
for (i in 2:length(cols)){

  u <- mean(data[,i])
  mean_tmp[i-1] <- u

}
```

Create cusum function for climate:

```r
cc <- function(dt){

  u <- mean(dt)
  s <- c(0)
  c <- .5*sd(dt)

  for (j in 2:length(dt)){

    s[j] <- max(0, s[j-1] + dt[j] - u -c)

  }

  return(s)
}
```

Determine whether climate change is significant:

```r
t <- 5*sd(mean_tmp)
cc(mean_tmp) > t
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

Conclusion

The temperature values provided in the data and the results above shows that the temperature has not significantly increased since the time the data was recorded. However, the same data could show different results if different values for the functions were implemented.