HW3

5.1 - Outliers (Crime Rates)

First, to identify the outliers I created three different plots as shown below:

1) Histogram 2) Box plot and 3) Line graph. In looking at these graphs, I can see five data points are clearly above the others. Specifically, two data points are close to a crime rate of about 2,000 (1,969 and 1,993) that seem like outliers.
   Also, three other data points are over 1,500 (1,635, 1,555, 1,674) which could be considered as outliers as well, but most likely not.

So, I used the Grubbs test to further determine if any of the above-mentioned data points could be indeed outliers.

First, I performed the Grubbs test to see if the highest point is an outlier. As shown below, the P-value for that test is 0.07887 which is greater than 0.05, hence we can accept the null hypothesis that 1,993 is not an outlier.

The fact that 1,993 is not an outlier indicates that other 4 points mentioned above are not outliers either, so there is no need to test for them.

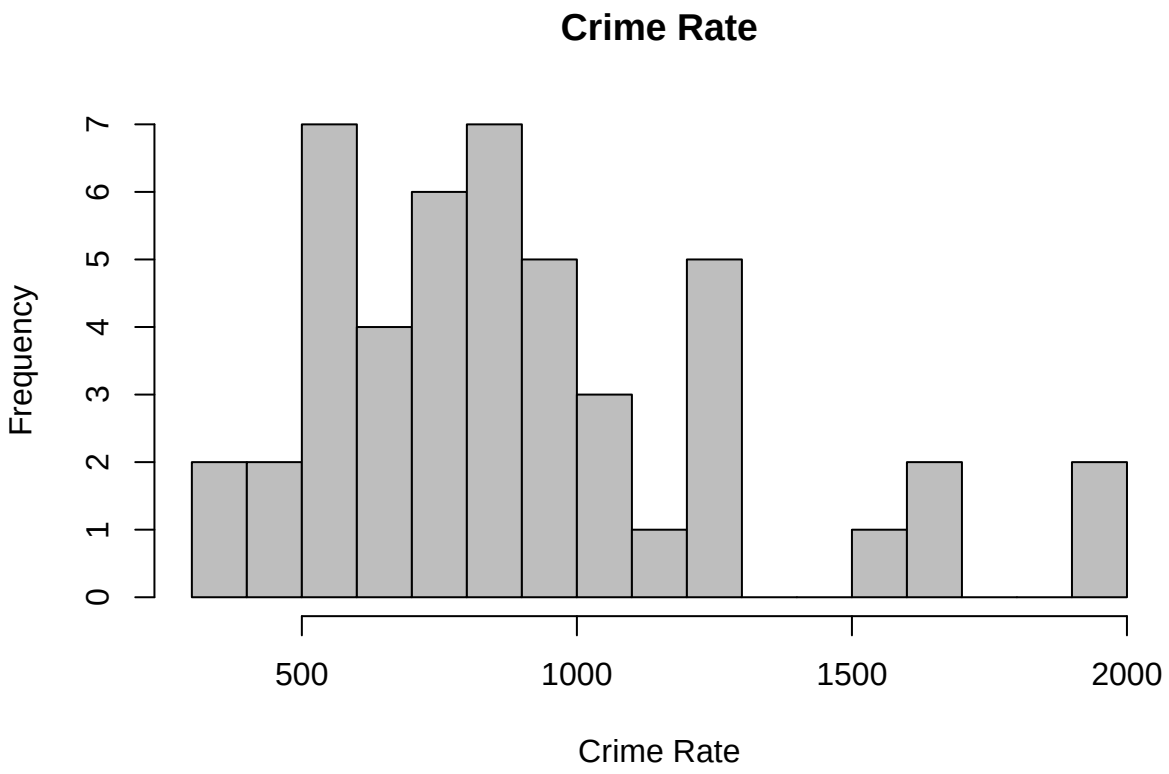
For the sake of completeness, I performed the Grubbs test to see if the extreme points on both sides of the tail are outliers. As shown below the P value obtained is 1, which clearly indicates that both those points are not outliers.

So, the conclusion is that any analysis we perform on the dataset should include all points.

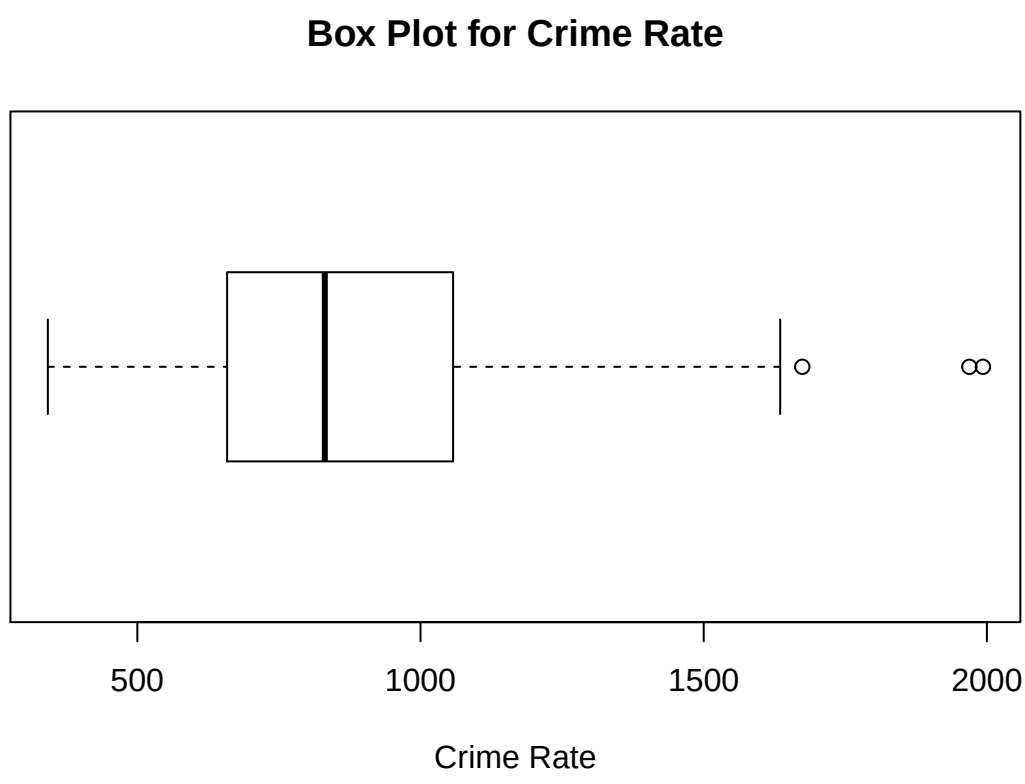
```
crime_data = read.table("uscrime.txt", header=TRUE)
crime = crime_data[,16]

print("\n")
```

```
## [1] "\n"
```

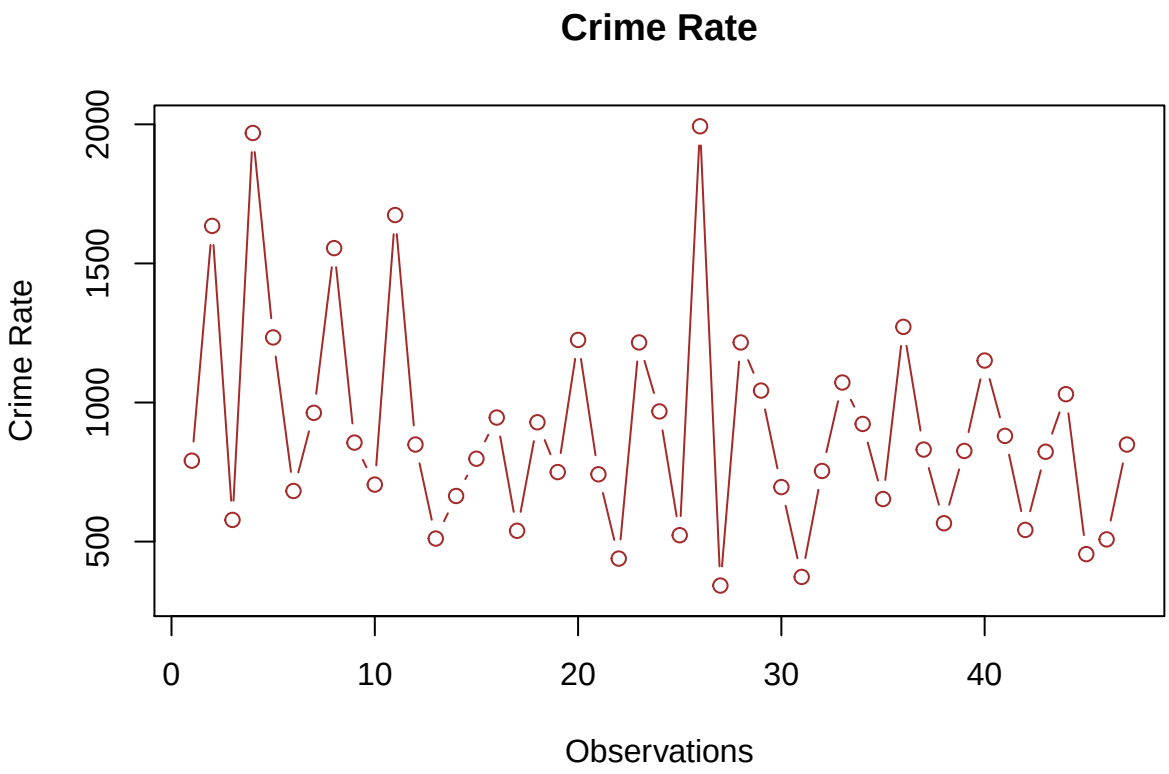
```
h = hist(as.numeric(crime), col="Grey", breaks=20, main="Crime Rate", xlab="Crime Rate")
```




```
boxplot(as.numeric(crime_data[,16]), horizontal=TRUE, main="Box Plot for Crime Rate", xlab="Crime Rate"
```




```
plot(as.numeric(crime_data[,16]), main="Crime Rate", xlab="Observations", ylab="Crime Rate", type='b',
```




```
library(outliers)
grubbs.test(as.numeric(crime_data[,16]), type=10)
```

```
##
##  Grubbs test for one outlier
##
## data:  as.numeric(crime_data[, 16])
## G = 2.81290, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
```

```
grubbs.test(as.numeric(crime_data[,16]), type=11)
```

```
##
##  Grubbs test for two opposite outliers
##
## data:  as.numeric(crime_data[, 16])
## G = 4.26880, U = 0.78103, p-value = 1
## alternative hypothesis: 342 and 1993 are outliers
```

6.1 - Change detection

A1C is a parameter that is used to detect if someone has turned diabetic or not. A1C test measures the blood sugar level over a period - usually 3 months. This is different than a random blood sugar test which measures the blood sugar level at that instant, for example in the morning before eating anything. If doctors

suspect that someone has turned diabetic, then they will prescribe an A1C test to see the blood sugar level over a period.

If I have the data of blood sugar levels for someone over a period, I would indeed use CUSUM to see when the A1C level exceeds the threshold to diagnose that person as diabetic. This would really be helpful because if someone has turned diabetic and if we deduct the condition after many months, a good amount of treatment time can be missed. On the other hand, if we deduct the condition/change in the A1C levels right after it happens, we can use that information to come up with a better treatment/cure for the disease.

In terms of choosing the appropriate C and threshold values, I would just iterate over a range of values and select the value that is prevalent in most of those runs. This is the approach I used for the later part of this assignment.

6.2 - Part 1

The idea here is to use CUSUM approach to determine when unofficial summer ends in each year based on the drop in temperature.

There are two important parameters that could affect the results in the CUSUM approach: 1) C 2) Threshold.

As shown below, I analyzed the temperature data for each of different C and threshold values. In each case, I found the exact date the temperature dropped below the threshold. Finally, I selected the date that appeared the greatest number of times from all the runs. For example, for a C value of 1 and threshold of 10, 2-Oct could be the date the temperature dropped, however for a different C and threshold value the result obtained could be different. The idea is to choose the date that appeared the most in all combination of C and threshold values.

```
library("plyr")
getwd()
```

```
## [1] "/Users/Mani/Desktop/Analytics/HW3"
```

```
temp = read.table("temps.txt",header=TRUE, check.names=FALSE)

c = 1
t = 30
ends = c()
s = rep(0, nrow(temp))
for (j in 2:ncol(temp))
  {
    l = c()
    for (c in 1:5)
      {
        for (t in 5:50)
          {
            d =  mean(as.numeric(temp[,j])) - as.numeric(temp[,j]) - c
            for (i in 1:nrow(temp))
              {
                if (i == 1)
                  s[i] = max(0, d[i])
                else
                  s[i] = max(0, d[i]+s[i-1])
              }
            f = temp[which(s>t),1]
```

```
            l = c(l, as.character(f[1]))
        }
    }


  g = count(l)
  print(g)

  cat("The temperature shifts downward in", colnames(temp[j]),"is: ", as.character(g[which.max(g[,2]),1
  cat("\n")
  ends = c(ends, match(as.character(g[which.max(g[,2]),1]), temp[,1]))
}
```

```
##         x freq
## 1   1-Oct   52
## 2  19-Sep    2
## 3   2-Oct   18
## 4   2-Sep   22
## 5  20-Sep    3
## 6  21-Sep    9
## 7  22-Sep    3
## 8  28-Sep    3
## 9  29-Sep   38
## 10 30-Sep   74
## 11  4-Oct    4
## 12  5-Oct    2
## The temperature shifts downward in 1996 is:  30-Sep
##
##         x freq
## 1   1-Aug    1
## 2  16-Oct    8
## 3  17-Oct    8
## 4  22-Sep    9
## 5  23-Sep    1
## 6  25-Sep   41
## 7  26-Sep   45
## 8  27-Sep   73
## 9  28-Sep   31
## 10 29-Sep    1
## 11  3-Oct    2
## 12 31-Jul    9
## 13  7-Jul    1
## The temperature shifts downward in 1997 is:  27-Sep
##
##         x freq
## 1  10-Oct   19
## 2  10-Sep    6
## 3  11-Oct    3
## 4  11-Sep    3
## 5  12-Oct    3
## 6  13-Oct    5
## 7  14-Oct    8
## 8  16-Oct    1
```

```
## 9  17-Oct    3
## 10 18-Oct    1
## 11  2-Oct    3
## 12 20-Oct    1
## 13 21-Oct    7
## 14 22-Oct    2
## 15 29-Sep    5
## 16  3-Oct   15
## 17  3-Sep    3
## 18 30-Sep   30
## 19  4-Oct    1
## 20  5-Oct    2
## 21  6-Oct   30
## 22  7-Oct    2
## 23  8-Oct   45
## 24  9-Oct   32
## The temperature shifts downward in 1998 is:  8-Oct
##
##         x freq
## 1   1-Oct   18
## 2  12-Jul    1
## 3  13-Jul   30
## 4  14-Jul    2
## 5  15-Jul    1
## 6   2-Oct    8
## 7  20-Sep   15
## 8  21-Sep   10
## 9  22-Sep   40
## 10 23-Sep   25
## 11 24-Sep   14
## 12 25-Sep    1
## 13 27-Sep    6
## 14 28-Sep    1
## 15 29-Sep    8
## 16 30-Sep   22
## 17  4-Oct    5
## 18  5-Oct    9
## 19  6-Oct    8
## 20  7-Oct    5
## 21  8-Oct    1
## The temperature shifts downward in 1999 is:  22-Sep
##
##         x freq
## 1   1-Oct    2
## 2  17-Sep    3
## 3  18-Sep    8
## 4  22-Sep    3
## 5  25-Jul   11
## 6  26-Jul    1
## 7  26-Sep    7
## 8  27-Sep    7
## 9  28-Sep    5
## 10 29-Sep    5
## 11 30-Sep    4
```

```
## 12  6-Sep   49
## 13  7-Oct   11
## 14  7-Sep   75
## 15  8-Oct    3
## 16  8-Sep   30
## 17  9-Sep    6
## The temperature shifts downward in 2000 is:  7-Sep
##
##         x freq
## 1   1-Oct    8
## 2  10-Oct    4
## 3  11-Oct    1
## 4  15-Oct    1
## 5   2-Sep    2
## 6  24-Sep    3
## 7  25-Sep   62
## 8  26-Sep   30
## 9  27-Sep   20
## 10 28-Sep    3
## 11 29-Sep   35
## 12  3-Sep   23
## 13 30-Sep   22
## 14  7-Oct    8
## 15  9-Oct    8
## The temperature shifts downward in 2001 is:  25-Sep
##
##         x freq
## 1  10-Oct    9
## 2  11-Oct    2
## 3  12-Jul    1
## 4  13-Jul    1
## 5  14-Oct   13
## 6  14-Sep    7
## 7  15-Oct   19
## 8  24-Sep    7
## 9  25-Sep   60
## 10 26-Sep   25
## 11 27-Sep   30
## 12 28-Sep   10
## 13 29-Sep   28
## 14 30-Sep    4
## 15 31-Aug   11
## 16  9-Oct    3
## The temperature shifts downward in 2002 is:  25-Sep
##
##         x freq
## 1   1-Jul    6
## 2   1-Oct   30
## 3   2-Oct   52
## 4  29-Sep   46
## 5   3-Oct   32
## 6  30-Sep   40
## 7   4-Oct    1
## 8   6-Oct    1
```

```
## 9   7-Oct    3
## 10  7-Sep   14
## 11  8-Oct    4
## 12  9-Oct    1
## The temperature shifts downward in 2003 is:  2-Oct
##
##         x freq
## 1  10-Aug    1
## 2  10-Oct   19
## 3  11-Oct    3
## 4  12-Oct   15
## 5  13-Aug    5
## 6  13-Oct   39
## 7  14-Oct   20
## 8  15-Oct    2
## 9  15-Sep    4
## 10 16-Sep   19
## 11 18-Sep    2
## 12 19-Sep    5
## 13 20-Sep   21
## 14 21-Sep   16
## 15 22-Sep    1
## 16 27-Sep    7
## 17 28-Sep    5
## 18 29-Sep    3
## 19 30-Sep    1
## 20  6-Oct    2
## 21  7-Oct    4
## 22  8-Oct    8
## 23  8-Sep    2
## 24  9-Oct   26
## The temperature shifts downward in 2004 is:  13-Oct
##
##         x freq
## 1  10-Oct   17
## 2  12-Oct   16
## 3  13-Oct    7
## 4  14-Oct    6
## 5  15-Oct    1
## 6  16-Oct    5
## 7  17-Oct    9
## 8  22-Oct    3
## 9  23-Oct    6
## 10 24-Oct    9
## 11  6-Oct   18
## 12  7-Jul    3
## 13  7-Oct   42
## 14  8-Oct   38
## 15  9-Oct   50
## The temperature shifts downward in 2005 is:  9-Oct
##
##         x freq
## 1  12-Oct   17
## 2  13-Oct   45
```

```
## 3  13-Sep   30
## 4  14-Oct   19
## 5  14-Sep    3
## 6  15-Oct    8
## 7  16-Oct    1
## 8  20-Sep    5
## 9  21-Sep   16
## 10 25-Sep    1
## 11 26-Sep    4
## 12 27-Sep    5
## 13 28-Sep    1
## 14 29-Sep   30
## 15 30-Sep   21
## 16  7-Oct    4
## 17  8-Oct   17
## 18  9-Oct    3
## The temperature shifts downward in 2006 is:  13-Oct
##
##         x freq
## 1   1-Oct    5
## 2  10-Oct    2
## 3  11-Oct   29
## 4  12-Oct   27
## 5  13-Oct   20
## 6  14-Oct    6
## 7  15-Oct    5
## 8  16-Oct   11
## 9  16-Sep   18
## 10 17-Sep   20
## 11 18-Oct    2
## 12 18-Sep   17
## 13 19-Oct    4
## 14 19-Sep    6
## 15  2-Oct    2
## 16 20-Oct    5
## 17 20-Sep    6
## 18 21-Oct    2
## 19 21-Sep   21
## 20 22-Jul    1
## 21 23-Jul    3
## 22  3-Oct    7
## 23  4-Oct    4
## 24  5-Oct    2
## 25  6-Oct    2
## 26  9-Oct    3
## The temperature shifts downward in 2007 is:  11-Oct
##
##         x freq
## 1  10-Oct   11
## 2  11-Oct    5
## 3  12-Oct    7
## 4  13-Oct    9
## 5  14-Oct    1
## 6  17-Oct   14
```

```
## 7  17-Sep   35
## 8  18-Oct   40
## 9  19-Oct   38
## 10  2-Oct    2
## 11 20-Oct   21
## 12 20-Sep    3
## 13 21-Oct    5
## 14 21-Sep   12
## 15 22-Oct    2
## 16 24-Sep    1
## 17 25-Aug    1
## 18 26-Sep    1
## 19 27-Sep    4
## 20  3-Oct    1
## 21  8-Oct    6
## 22  9-Oct   11
## The temperature shifts downward in 2008 is:  18-Oct
##
##         x freq
## 1   1-Oct   10
## 2   1-Sep   12
## 3  12-Oct    8
## 4  13-Oct    4
## 5  14-Oct   18
## 6  15-Oct   15
## 7  19-Sep    1
## 8   2-Oct   15
## 9   2-Sep    3
## 10  3-Oct   10
## 11  3-Sep    1
## 12 30-Sep    8
## 13  4-Oct   20
## 14  5-Oct   80
## 15  6-Oct   25
## The temperature shifts downward in 2009 is:  5-Oct
##
##         x freq
## 1   1-Oct   25
## 2   2-Oct   21
## 3  26-Sep   18
## 4  27-Sep   15
## 5  28-Sep   40
## 6  29-Sep   25
## 7   3-Jul    1
## 8   3-Oct   33
## 9  30-Sep   45
## 10  4-Jul    1
## 11  4-Oct    6
## The temperature shifts downward in 2010 is:  30-Sep
##
##         x freq
## 1   1-Oct    4
## 2  10-Oct    7
## 3  18-Sep    1
```

```
## 4  19-Sep    4
## 5   2-Oct   13
## 6  20-Sep    1
## 7  22-Sep    1
## 8  24-Sep    1
## 9   3-Oct   13
## 10  4-Oct    1
## 11  5-Sep   12
## 12  6-Sep   33
## 13  7-Oct    1
## 14  7-Sep   70
## 15  8-Oct    3
## 16  8-Sep   45
## 17  9-Oct   14
## 18  9-Sep    6
## The temperature shifts downward in 2011 is:  7-Sep
##
##         x freq
## 1   1-Oct   34
## 2  19-Sep    3
## 3   2-Oct   46
## 4  20-Sep    8
## 5  24-Sep    4
## 6  25-Sep    3
## 7   3-Oct   39
## 8  30-Sep   14
## 9   4-Oct    1
## 10  4-Sep    3
## 11  7-Oct   36
## 12  8-Oct   34
## 13  9-Oct    5
## The temperature shifts downward in 2012 is:  2-Oct
##
##         x freq
## 1  10-Oct    1
## 2  15-Aug    1
## 3  15-Oct    6
## 4  16-Aug   58
## 5  16-Oct    6
## 6  17-Aug   65
## 7  17-Oct    3
## 8  18-Aug    6
## 9  18-Oct    5
## 10 19-Oct   35
## 11 20-Oct   14
## 12 21-Oct    8
## 13 22-Oct    6
## 14 29-Sep    2
## 15 30-Sep    3
## 16  4-Jul    6
## 17  8-Oct    1
## 18  9-Oct    4
## The temperature shifts downward in 2013 is:  17-Aug
##
```

```
##         x freq
## 1  19-Oct    1
## 2  20-Jul    3
## 3  20-Oct   13
## 4  21-Jul    1
## 5  21-Oct    2
## 6  22-Oct    9
## 7  23-Oct    3
## 8  24-Sep    9
## 9  25-Sep   18
## 10 26-Sep   35
## 11 27-Sep   15
## 12 28-Sep   35
## 13 29-Sep   48
## 14  4-Oct   15
## 15  5-Oct   20
## 16  6-Oct    3
## The temperature shifts downward in 2014 is:  29-Sep
##
##         x freq
## 1   1-Oct    4
## 2  13-Sep    5
## 3  14-Sep   14
## 4  15-Sep    6
## 5  16-Sep    6
## 6   2-Oct   15
## 7  24-Sep   12
## 8  25-Sep   67
## 9  26-Sep   38
## 10 27-Sep   35
## 11 28-Sep   12
## 12 29-Sep    3
## 13  3-Oct    2
## 14 30-Aug   11
## The temperature shifts downward in 2015 is:  25-Sep
```

6.2 - Part 2

The task here is to analyze how the summer temperature has been over the years - specifically, to judge whether Atlanta's summer climate has gotten warmer over the years since 1996.

First, I obtained the date when the weather cooled off in each year - this is basically the answer to the previous question.
Then for each year, I took the average of the temperatures until that date (not including the date when the change happened). The idea is that the date obtained in the previous question is when summer unofficially ended, so to understand if the summer has gotten worse over the years the temperatures until that date would be a good indicator. I took the simple average of the temperatures in each year until the date when the weather cooled off.

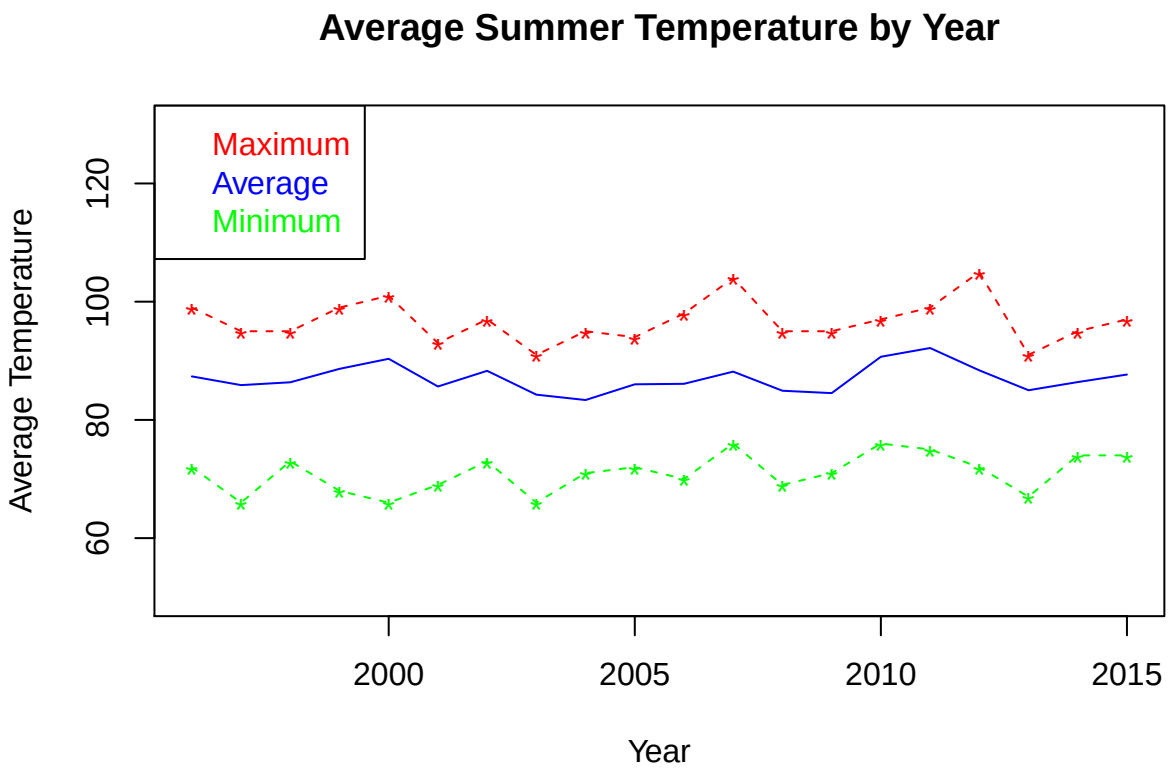
As shown below in the tables and graph, it would be hard to conclude that the summer temperature has gotten progressively warmer in Atlanta. As shown in the graph below the average temperature in each year is not a flat line nor does it go up and down significantly for long periods to conclude for sure. In other words, the average summer temperature goes slightly up and down which makes me conclude that the summer temperatures over these years have not gotten progressively worse. To substantiate my conclusion,

I also plotted the maximum and minimum values over the years and they also pretty much follow the same pattern as the average values.

```
#ends
avg = c()
year = c()
maxi=c()
mini = c()
for (i in 1:length(ends))
{
  temp_var = mean(temp[1:ends[i]-1,i+1])
  maxi = c(maxi, max((temp[1:ends[i]-1,i+1])))
  mini = c(mini, min((temp[1:ends[i]-1,i+1])))

  avg = c(avg,temp_var)
  cat("The average summer temperature in", colnames(temp[i+1]), "is", avg[i], "\n")
  year = c(year,colnames(temp[i+1]))


}
```

```
## The average summer temperature in 1996 is 87.36264
## The average summer temperature in 1997 is 85.88636
## The average summer temperature in 1998 is 86.36364
## The average summer temperature in 1999 is 88.63855
## The average summer temperature in 2000 is 90.33824
## The average summer temperature in 2001 is 85.65116
## The average summer temperature in 2002 is 88.30233
## The average summer temperature in 2003 is 84.26882
## The average summer temperature in 2004 is 83.36538
## The average summer temperature in 2005 is 86.02
## The average summer temperature in 2006 is 86.10577
## The average summer temperature in 2007 is 88.16667
## The average summer temperature in 2008 is 84.93578
## The average summer temperature in 2009 is 84.53125
## The average summer temperature in 2010 is 90.68132
## The average summer temperature in 2011 is 92.16176
## The average summer temperature in 2012 is 88.3871
## The average summer temperature in 2013 is 85.02128
## The average summer temperature in 2014 is 86.4
## The average summer temperature in 2015 is 87.67442
```

```
plot(year, avg, xlab='Year', ylab='Average Temperature', main='Average Summer Temperature by Year', typ
#par(new=TRUE)
#plot(maxi, type='l', col='red')
#par(new=TRUE)
#plot(mini, type='l', col='green')

points(year, maxi, col="red", pch="*")
lines(year, maxi, col="red",lty=2)
points(year, mini, col="green", pch="*")
lines(year, mini, col="green",lty=2)
legend( "topleft", c("Maximum", "Average", "Minimum"),
text.col=c("red", "blue", "green") )
```


Average Summer Temperature by Year
Maximum
Average
Minimum
Average Temperature
60
80
100
120
2000
2005
2010
2015
Year