Change Detection Models & Outlier Detection

Outlier Detection

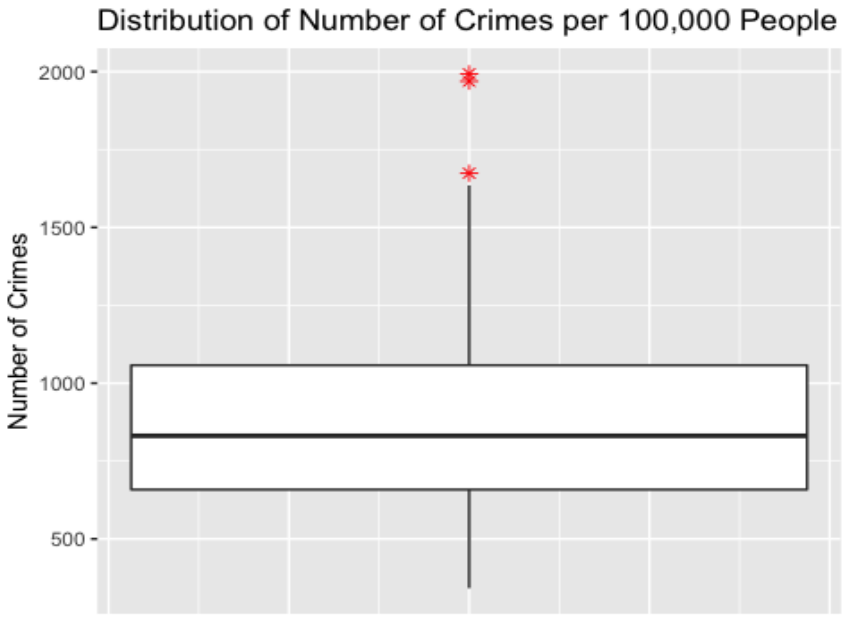
The goal of this question was to determine if there were any outliers present in the *uscrimes* data set, specifically, in the last column containing the number of crimes per 100,000 people. To do so, three types of outliers must be investigated to determine their presence.

First, it was ruled out that any contextual outliers were present, since the data in question is not a time series. Any data point selected could not be far from a point nearby in time, since the data was not recorded with time intervals.

Next, it was determined that no collective outliers were present. As it can be seen below there are no missing points in the data set. Additionally, since the data is not a time series no pattern could be analyzed to determine if there was missing or incorrect data at any time interval.

```
sum(is.na(uscrime))

## [1] 0
```

Lastly, it was determined that there were three possible point outliers with values of 1993, 1969, and 1674. Possible point outliers were determined using the Tukey method of calculating whiskers (75th percentile + 1.5*IQR, 25th percentile - 1.5*IQR). Any points outside of these whiskers were determined to be possible point outliers. This can be seen in the box plot below:



Next, it needed to be determined whether these three possible point outliers were actually outliers. To do so, the Grubbs test was used to find the probability of a point being an outlier. To begin, the test was done with the highest point, 1993. If this point was concluded to be an outlier, then the test would take place for the smaller possible point outiers as well to determine if they were also outliers. If 1993 was not determined to be an outlier, however, then it could also be determined that the smaller possible point outliers were also not outliers. H_0 for this test was that 1993 was not an outlier, and H_A was that 1993 was an outlier. A significance level of 0.05 was used for this test, and any p-values smaller than this would reject the null hypothesis. The results for the Grubbs test for the point 1993 can be seen below:

```
grubbs.test(x = uscrime$Crime,
            type = 10,
            opposite = F,
            two.sided = F)

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

It can be seen that for the Grubbs test that 1993 was an outlier, there was a p-value of 0.07887. From this, we fail to reject the null hypothesis and cannot conclude that 1993 is a point outlier. Because 1993 is not a point outlier, we can also assume that the two smaller possible point outliers, 1969 and 1674, are also not outliers. With this understanding, we can conclude that there are no outliers present in the data for the number of crimes per 100,000 people in the *uscrimes* data set.

The limitations of the approach to identifying outliers in this situation should be noted. First, using a different method for calculating whiskers in the boxplot could have revealed no possible point outliers, compared to the three identified using the Tukey method. If higher percentiles were using for the whiskers, then the use of the Grubbs test could have been ancillary in this situation. Additionally, the conclusion that 1993 was not an outlier from the Grubbs test came with a p-value that was still relatively small, though not below the significance level set. A different test could have revealed this point and possibly more as outliers, however, and therefore the influence of these points should still be examined when fitting any models.

CUSUM Approach to Change Detection

Number 1

The goal of this question was to determine when the weather starts cooling off (the unofficial end of summer) using July through October daily-high-temperature data for

Atlanta for 1996 through 2015. The task was to do so using the CUSUM approach for Change Detection, and specifically, detecting a decrease for this situation.

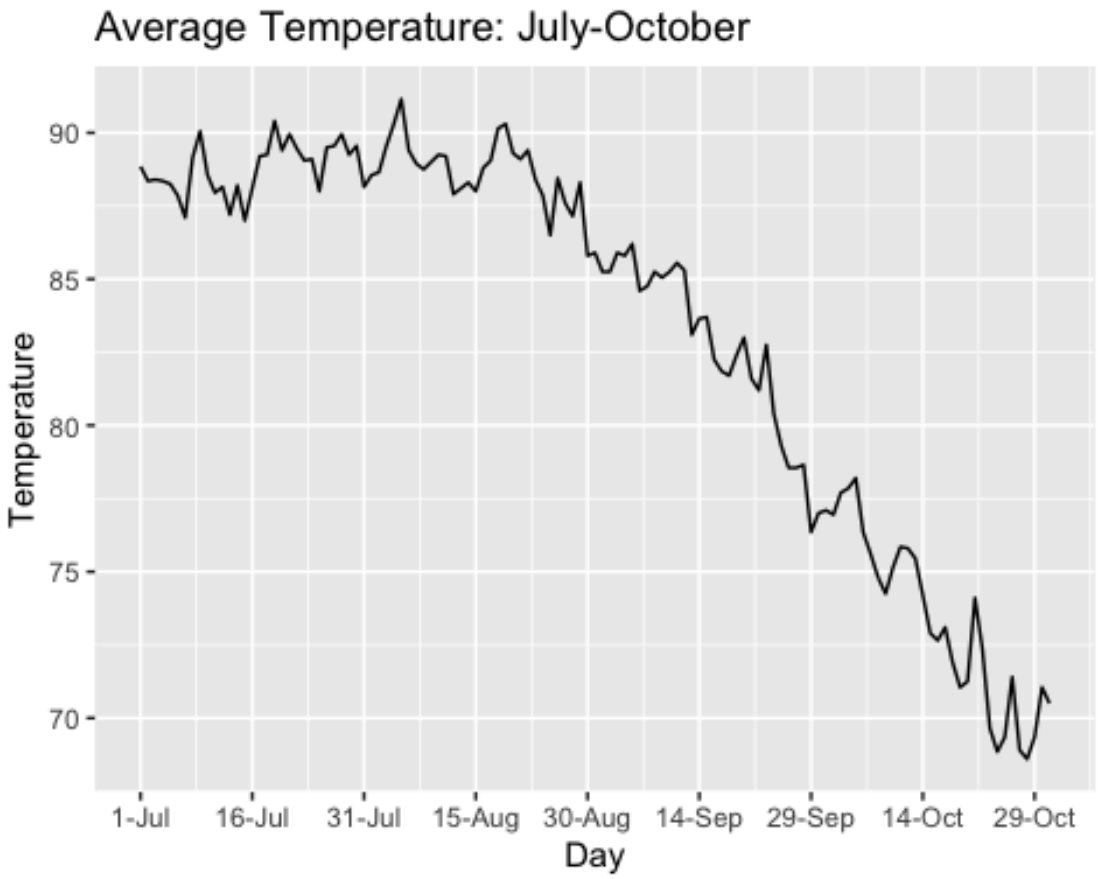
To answer this question, an approach was chosen that used the average temperature from every day between July and October to find the average unofficial end of summer for this time period. This method was chosen to hopefully provide better insight about when temperatures start to cool off by summarizing this for the entire time period, rather than reporting each year separately which could vary and provide less useful insight.

```
temps$Average = rowMeans(temps[,-1])
```

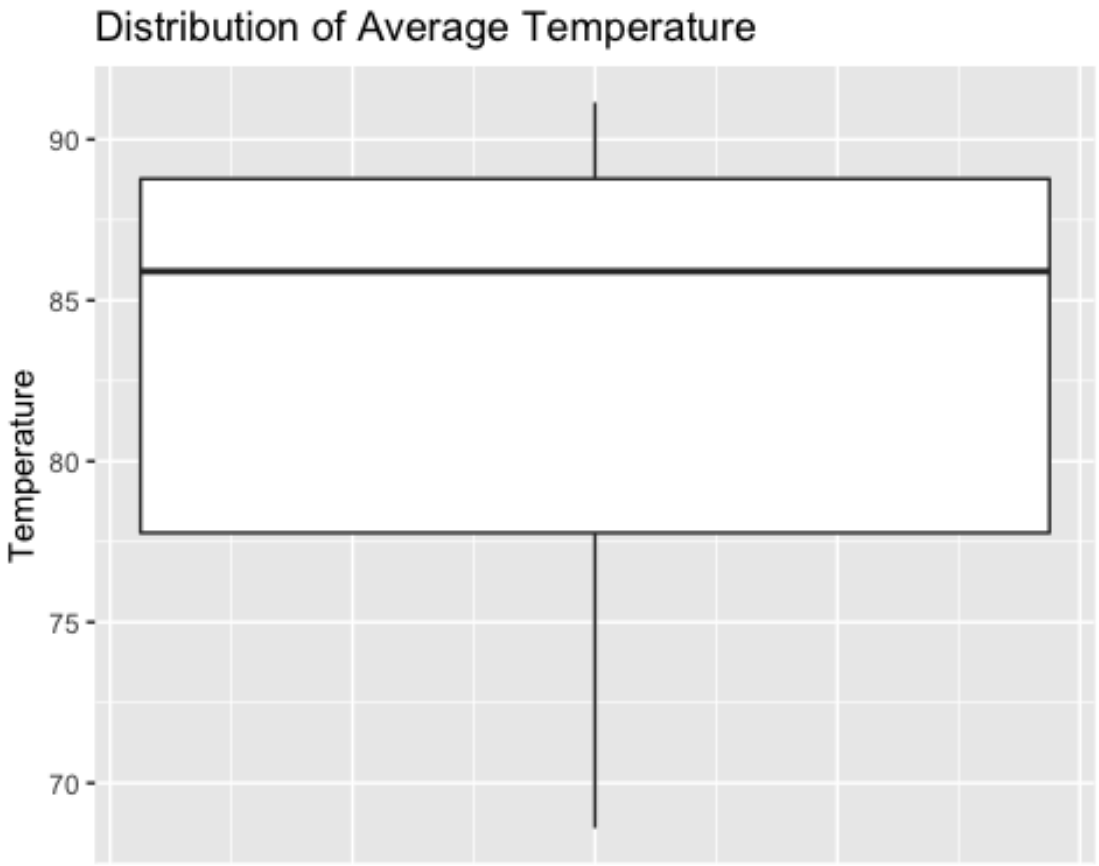
Before fitting a change detection model, the presence of outliers needed to be examined. To begin, it was determined that no collective outliers were present in the data as no points were missing. This also confirms that no outliers were present when calculating average temperatures above.

```
sum(is.na(temps))

## [1] 0
```

This was supported by the findings from the plot below, which did not show any real pattern in the time series data were points may be missing. Additionally, from this plot it could be determined that there were no significant contextual outliers. While there certainly were spikes and dips in average temperature, this variance is present throughout the entire data and no one point sticks out as a significant contextual outlier. This can be seen in the plot below:



Finally, it was determined that there were no point outliers present in the data. As it can be seen from the box plot below, no data points lie outside the whiskers of the box plot (calculated with the Tukey method). With the conclusion that there were no outliers present in the data, a Change Detection model could be fit.



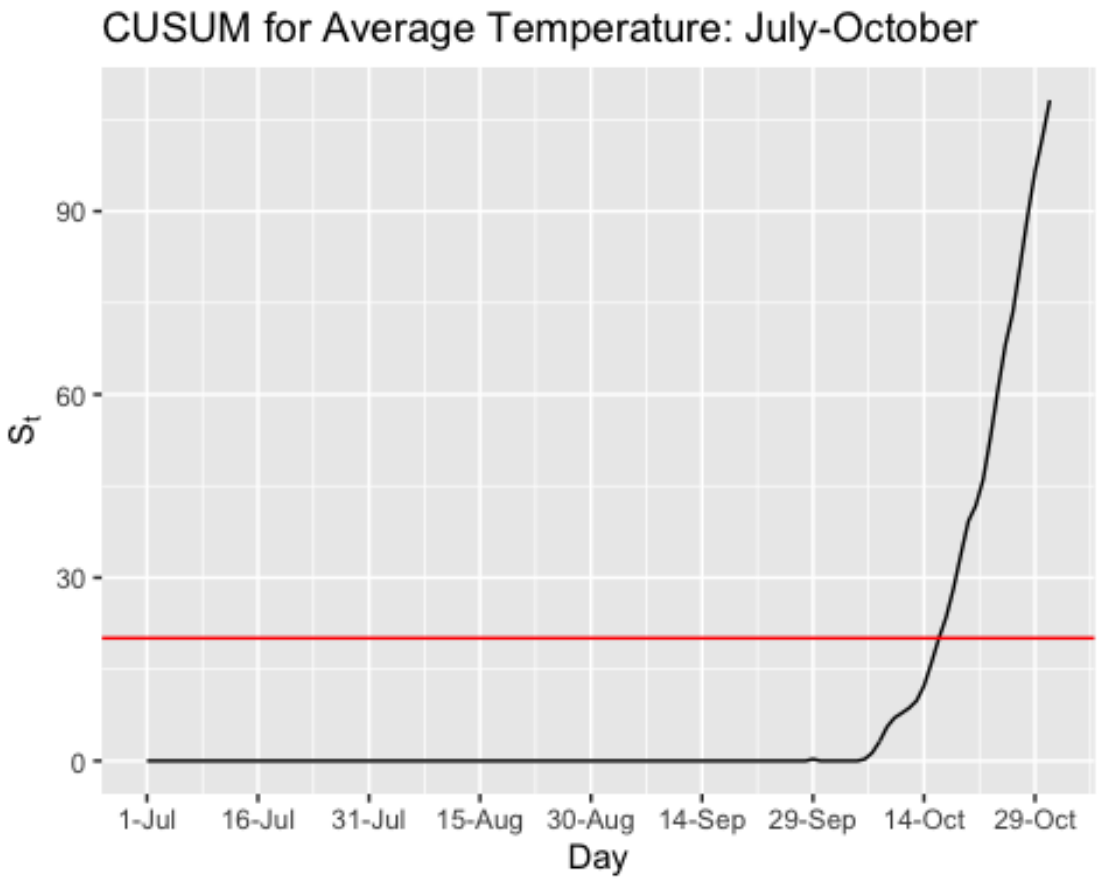
To set the parameters for the CUSUM approach, the mean, critical level, and threshold needed to be calculated. The threshold was set at three standard deviations which is a common choice since assuming a normal distribution, three standard deviations from the mean in each direction covers 99.7% of the observed data. As it can be seen below, this choice covers all of the observed data for the average temperature. To accompany the threshold, a critical level of one standard deviation was set as it is also a common choice and will keep the model from being too optimistic. Given that the data was sufficiently covered using these values, it was not needed to raise the values even higher given that the cost for false postives using the CUSUM approach was not high.

```
average_mean = mean(temps$Average)
average_c = sd(temps$Average)
T.Value = 3*average_c
nrow(temps[temps$Average <= (mean(temps$Average)+3*average_c) &
             temps$Average >= (mean(temps$Average)-3*average_c),])/nrow(temps
)

## [1] 1
```

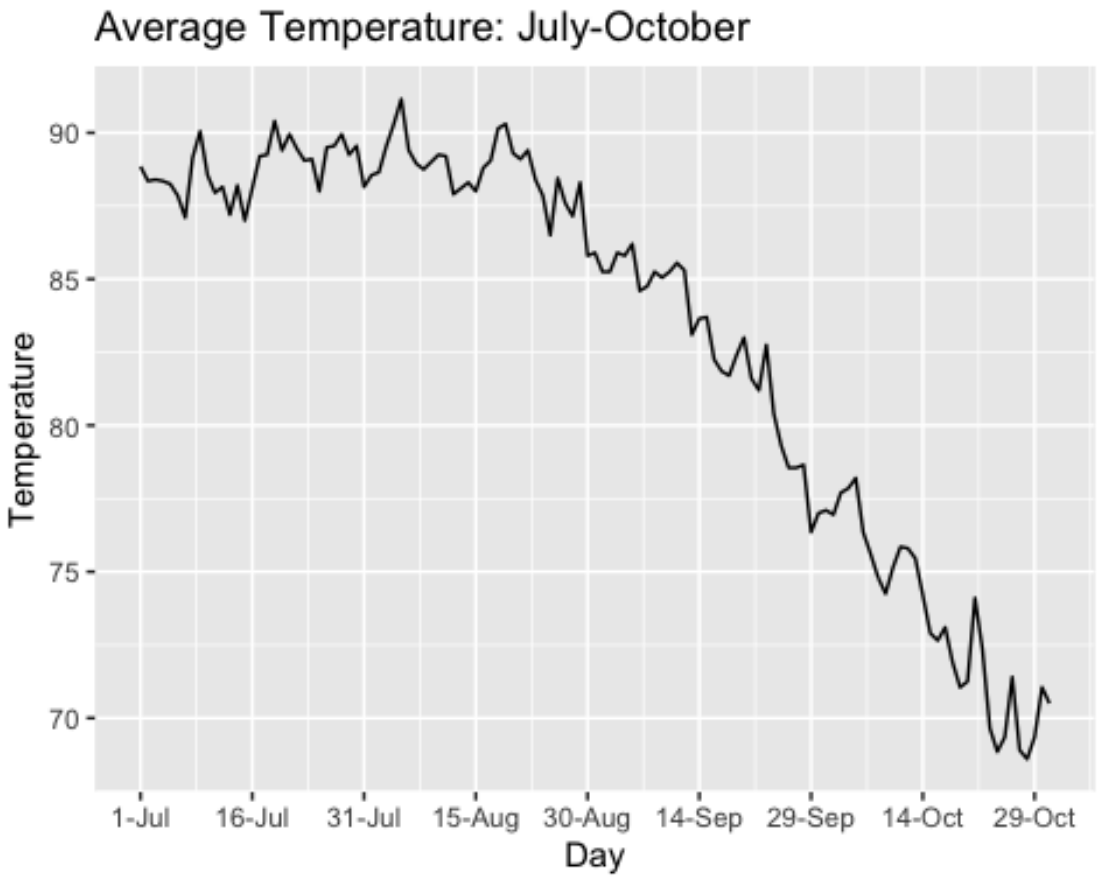
Next, the CUSUM model was fit. The S_t value was calculated for each day between July and October using the average temperatures and parameters specified above. Next, the first S_t value to be greater than or equal to the threshold was reported, as it was the first detected change by the model.

```
# Generate St
temps[1,"S.Value"] = 0
for(i in 2:nrow(temps)){
  temps[i,"S.Value"] = max(0,temps[i-1,"S.Value"]+(average_mean-temps[i,"Aver
age"]-average_c))
}
temps[temps$S.Value >= T.Value,]$Day[1]

## [1] "17-Oct"
```

It was determined that October 17th was the first detected change by the CUSUM model. From this, it was concluded that October 17th was the average unofficial end of summer for Atlanta between 1996 and 2015. From the plot below, the S_t value for each day between July and October can be seen, and when these values crossed the threshold.



This can be compared with the plot of Average Temperatures by Day to see when the CUSUM model detected the unofficial end of summer (October 17th).



Number 2

The goal of this final question was to determine whether or not Atlanta's summer climate has gotten hotter during the time period of 1996 to 2005. To do so, the average summer temperature would need to be calculated for each year, then a CUSUM approach would need to be used to determerine if an increase in average summer temperature was detected.

To begin answering this question, first the summer season needed to be determined for each year. This was done by first calculating the unofficial end of summer for each year using the CUSUM approach.

```
# df to store dates where threshold crossed for each year
yearly_dates = c()
# CUSUM Yearly
for(y in 2:21){
  # subset df
  yearly_temps = data.frame(Day = temps$Day,
                            Temp = temps[,y])
  # yearly mean
  yearly_mean = mean(yearly_temps$Temp)
  # yearly C
  yearly_c = sd(yearly_temps$Temp)
```

```
  # yearly T
  yearly_T = yearly_c*3
  # Generate St
  yearly_temps[1,"S.Value"] = 0
  for(i in 2:nrow(yearly_temps)){
    yearly_temps[i,"S.Value"] = max(0,yearly_temps[i-1,"S.Value"]+(yearly_mea
n-yearly_temps[i,"Temp"]-yearly_c))
  }
  yearly_dates[y-1] = yearly_temps[yearly_temps$S.Value > yearly_T,]$Day[1]
}
```

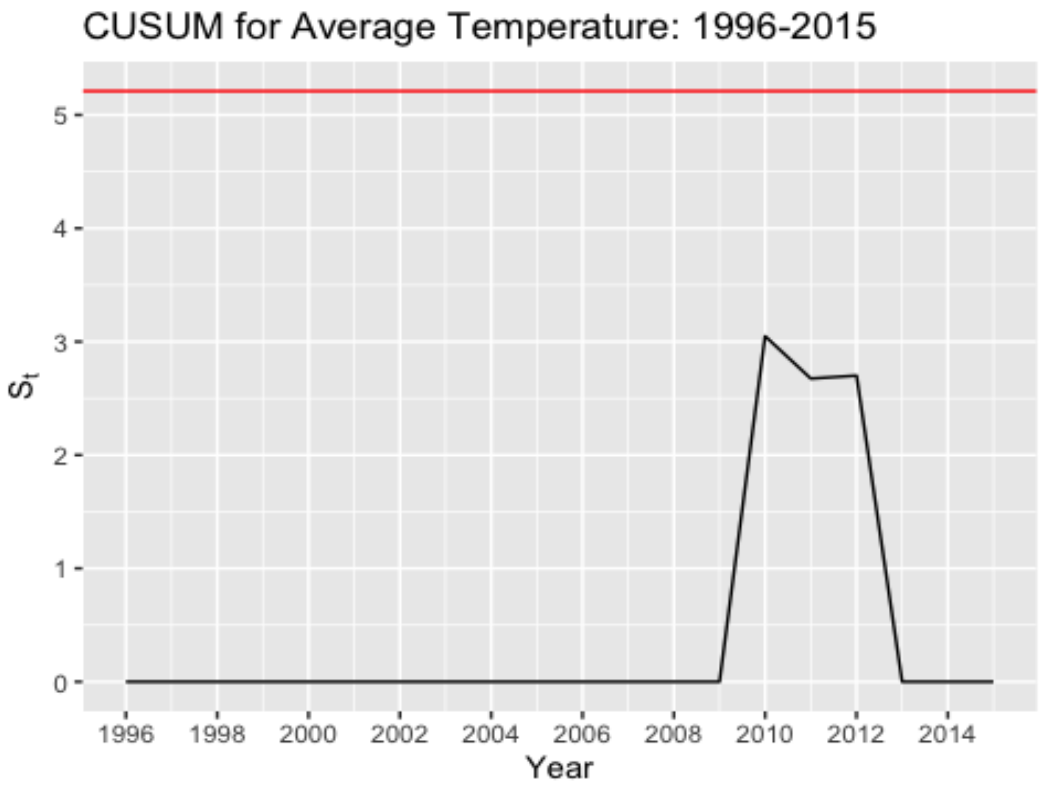
Assuming the summer season had already started by July, the summer period for each year could then be determined by taking the first day of July up until the unofficial end of summer for each year. After this, the average temperature for each summer between 1996 and 2015 could be calculated.

```
# yearly summer averages
summer.averages = data.frame(Year=1996:2015)
for(y in 2:21){
  summer = temps[1:which(temps$Day %in% yearly_dates[y-1]),y]
  summer.averages[y-1,'Average'] = mean(summer)
}
```

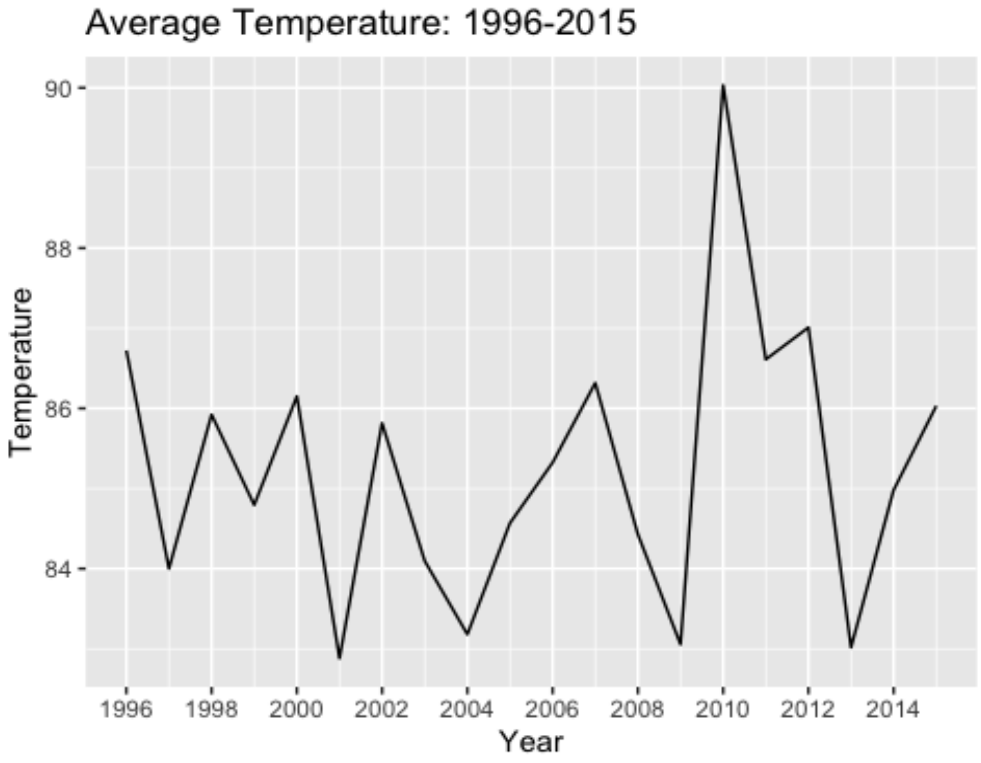
After this, a CUSUM approach could be used to determine if an increase in average summer temperature was detected. As used previously, the critical value was set at one standard deviation and the threshold was set at three standard deviations.

```
# Create Mu, C, T
average_mean = mean(summer.averages$Average)
average_c = sd(summer.averages$Average)
T.Value = 3*average_c
# Generate St
summer.averages[1,"S.Value"] = 0
for(i in 2:nrow(summer.averages)){
  summer.averages[i,"S.Value"] = max(0,summer.averages[i-1,"S.Value"]+(summer
.averages[i,"Average"]-average_mean-average_c))
}
summer.averages[summer.averages$S.Value >= T.Value,]$Year[1]

## [1] NA
```

It can be seen that the CUSUM model is not returning any S_t values which are above the threshold. When looking at the plot of S_t values by year below, it can be seen that indeed no S_t values passed the threshold.



Comparing this to the yearly average summer temperatures below, it can be seen that indeed while temperatures fluctuate they do no consistently increase.



Given this understanding, it can be concluded that Atlanta's summer climate did not gotten hotter during the time period of 1996 to 2015.

The limitations of the CUSUM approach used in this two questions must be noted along with the finding presented. Most notably, it is possible that the models were too optimistic in detecting an increase or decrease and that some false positives could have been detected. This could be remedied by increasing the critical value and threshold used throughout, though it should be noted that the justification used for choosing these values when fitting models was sound. On the contrary, models could have been too conservative as well and resulted in false negatives as a result of these values being too high. This is the give-and-take implied with using the CUSUM approach for a Change Detection model, and the consequences of specifying its parameters should be noted even if they were chosen with proper reasoning.