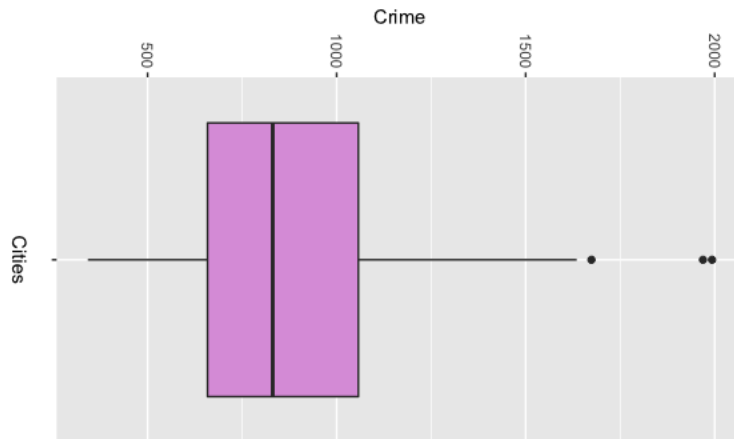


## 5.1

Please review the attached file 5.1.R along with this explanation.

First, I loaded the data and inspected the data, then created a box plot of crime rates. From the initial plot, it looks like there are 3 potential outliers on the high end of the data, with none on the low end.



I then ran a one-sided grubbs test on the crime data. The p-value of 0.07887 for the highest point (1993) indicates that the data point is not an outlier at .05. I decided to re-run the test at a p-value of .1. At this the significance level 1993 is of course an outlier. I removed the first outlier and re-ran the grubbs test. The next highest value, 1969, is also an outlier at this significance level so I removed it as well and re-ran the test. The third highest point, 1674, is not an outlier at this significance level. The grubbs test for the lowest crime cities did not indicate an outlier at any reasonable p-value.

In conclusion, neither of the two highest data points are outliers at a p-value of .05, but both are at .1. I think you could make a reasonable argument for either removing both data points or keeping both, depending on whether removing good data points or failing to remove a true outlier would be bigger risk to your business. Additionally, because the two potential outliers are spaced closely with each other they are likely influencing the test results, further arguing that it may be reasonable to remove them despite the relatively high p-value.

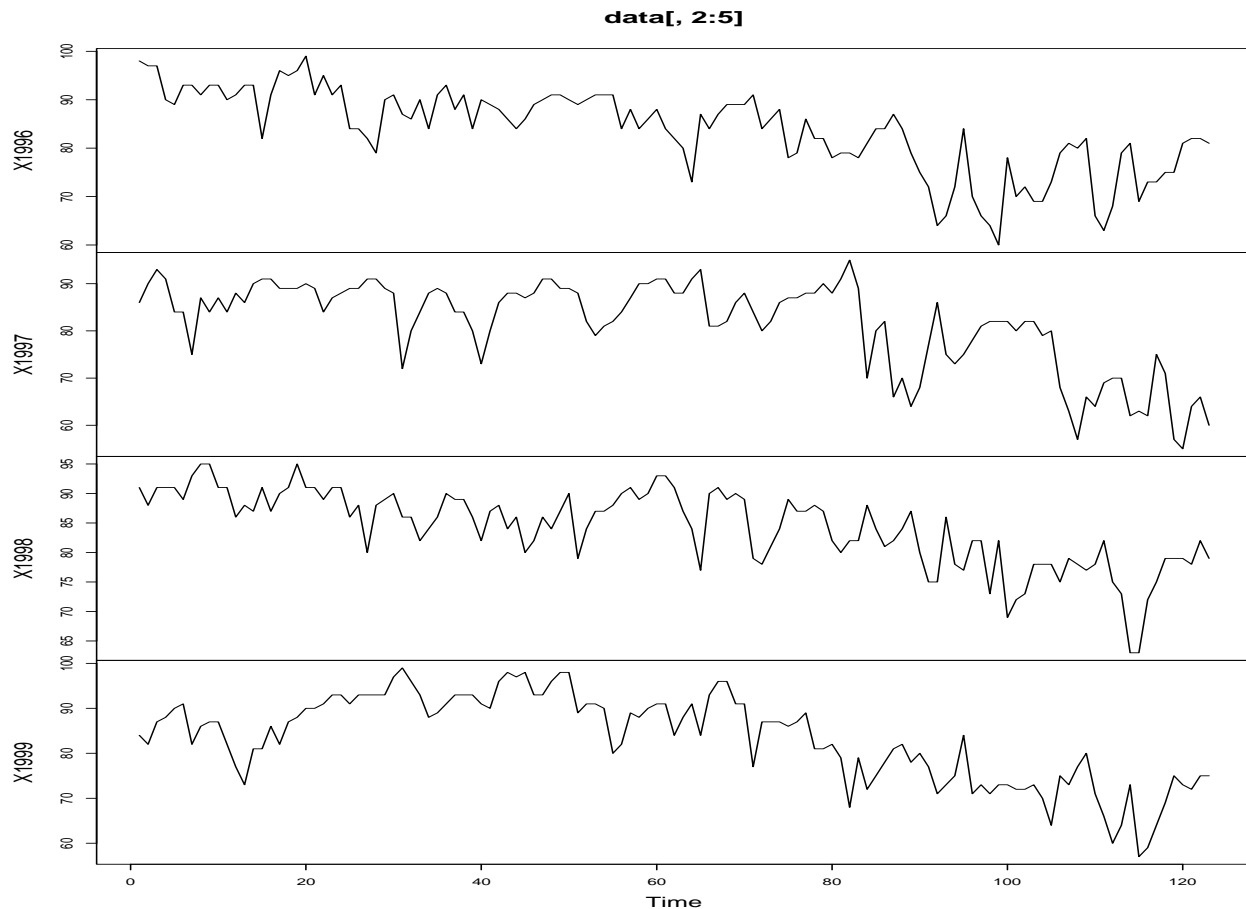
## 6.1

In my personal life I could use a change detection model in my cycling training to determine if I need to take time to recover. I track daily resting heart rate and heart rate and power output during training sessions, all of which can deviate from my known baselines due to fatigue or sickness. A change detection model could potentially detect these symptoms before they get too bad, allowing me to take a break and recover sooner.

There is a decent amount of random variance in these values on a daily basis, so I would use a relatively high c-value, maybe 1 std. deviation for each data type. However, it would be much more valuable to detect a false positive early and inspect more closely rather than missing out on a potential change, so I would use a relatively small T value.

## 6.2.1

I first began loading the data in R and plotting a few years' worth of daily temp data to get a feel for the shape of the annual trends. As expected, the temperature trends downward from July to October, but with significant variability day to day.



After exploring R's CUSUM functionality I decided it would just be easier to explore everything in excel due to the small size of the data set and simplicity of the CUSUM formula. I used the formula described in the lectures to calculate the CUSUM separately for each year and used one C value and one T value across all years. I used a rolling mean from July 1 to the date being tested for the Mu value in each calculation. I had initially used all the data points to calculate the mean for Mu, but decided that a rolling mean made the most sense as it uses all available data up to the date being tested without looking at future data. This would be the case in a real world application where we would not know what future values would be. I also considered just using the mean temperatures for July, but decided that it made more sense to incorporate more data, as the July mean may not accurately represent summer temperatures, especially for particularly warm or cold Julys.

I tested a variety of C and T values, from 0 to 5 standard deviations for each. Ultimately, I arrived at a C value of about 6.7 (one standard deviation) and a T value of 21 (about 3 standard deviations). These values seem to correctly flag the beginnings of sustained temperature decline without finding early false positives other than one in 2013 that was unavoidable without a significant negative impact on finding other true changes. This was largely a qualitative decision based on balancing the two factors.

[illegible]

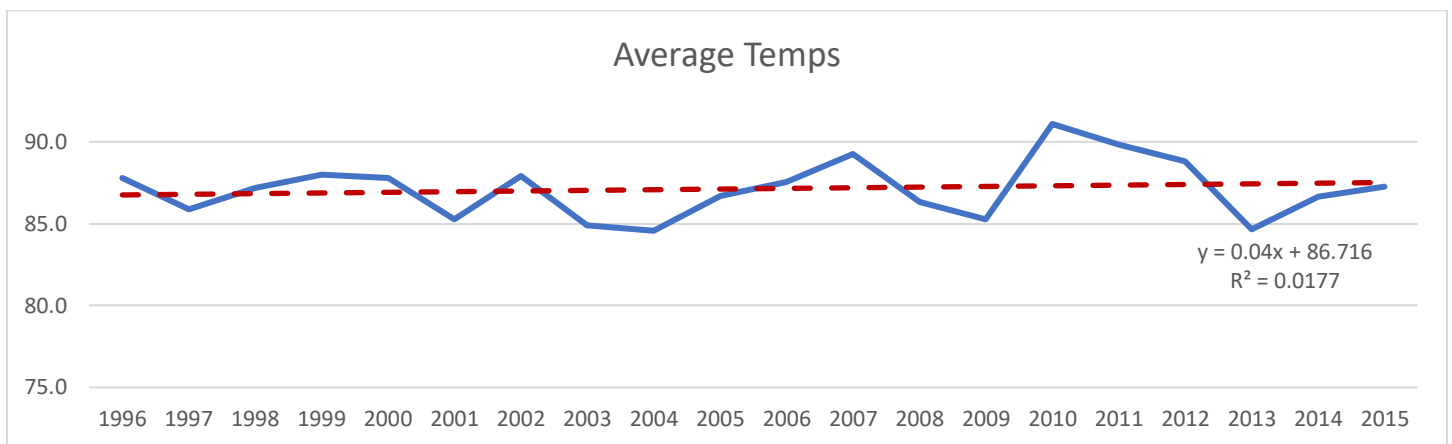
1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
9/30/18	9/26/18	10/9/18	9/20/18	9/6/18	9/25/18	9/25/18	9/30/18	10/12/18	10/8/18
2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
9/21/18	9/18/18	10/18/18	10/4/18	9/29/18	9/6/18	10/2/18	8/17/18	9/28/18	9/26/18

With the C and T values above, the average end of summer is 87 days from July 1, or September 25<sup>th</sup>. The latest end date is 109 days, or October 17<sup>th</sup> in 2008. The earliest is August 17<sup>th</sup> from the false positive in 2013, but with that false positive removed the earliest is September 6<sup>th</sup>.

## 6.2.2

I took two different approaches to see if the summer climate has warmed over the time period in the data set. First I applied CUSUM to the average annual summer length, then I applied it to average annual summer temperatures over the dataset. See tab 6.2 of the included spreadsheet for the calculations and results of those models

There is a very, very, slight (slope of .04) upward temperature trend over the time period:



However, I don't think there is enough data to conclusively say that the summer climate has warmed over such a short period. With low C and T values, both CUSUMs described above find changes after the long 2008 summer and the three warm years in 2010-2012 respectively, but at higher T values (about 2 and 3 standard deviations respectively) neither model detects a change. Given that average annual summer temperatures and lengths both declined following these potential change detections, and the changes were only detected at low thresholds, I think it is too soon to conclusively say that summers warmed over the time period.