Week 2 HW - ISYE 6501

May 30, 2018

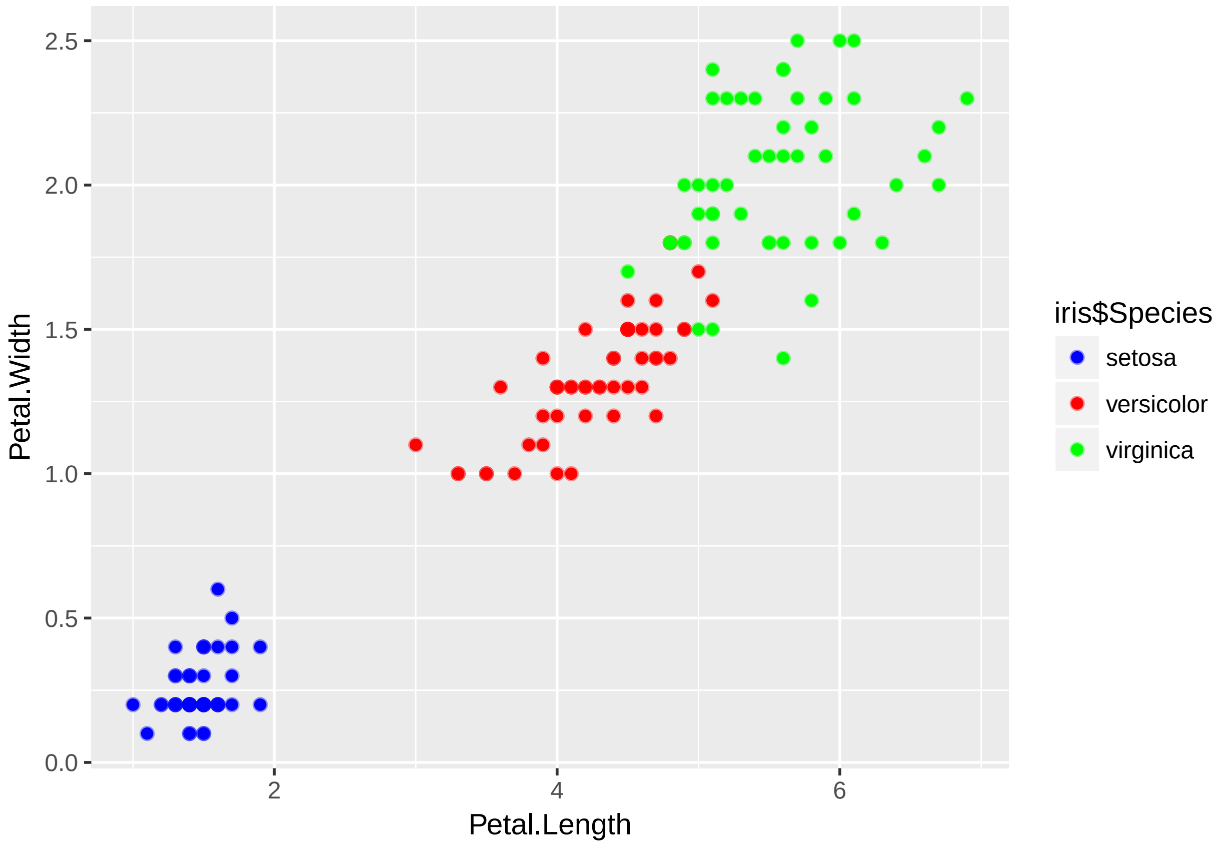
Question 4.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a clustering model would be appropriate. List some (up to 5) predictors that you might use.

Answer: We would like to cluster classroom students together based on academic performance. Then we could assign students to advanced, regular and remedial classes that are more fitting to their ability level. Predictors we would use include standardized test scores, attendance data, and classroom grades from the previous year.

Question 4.2

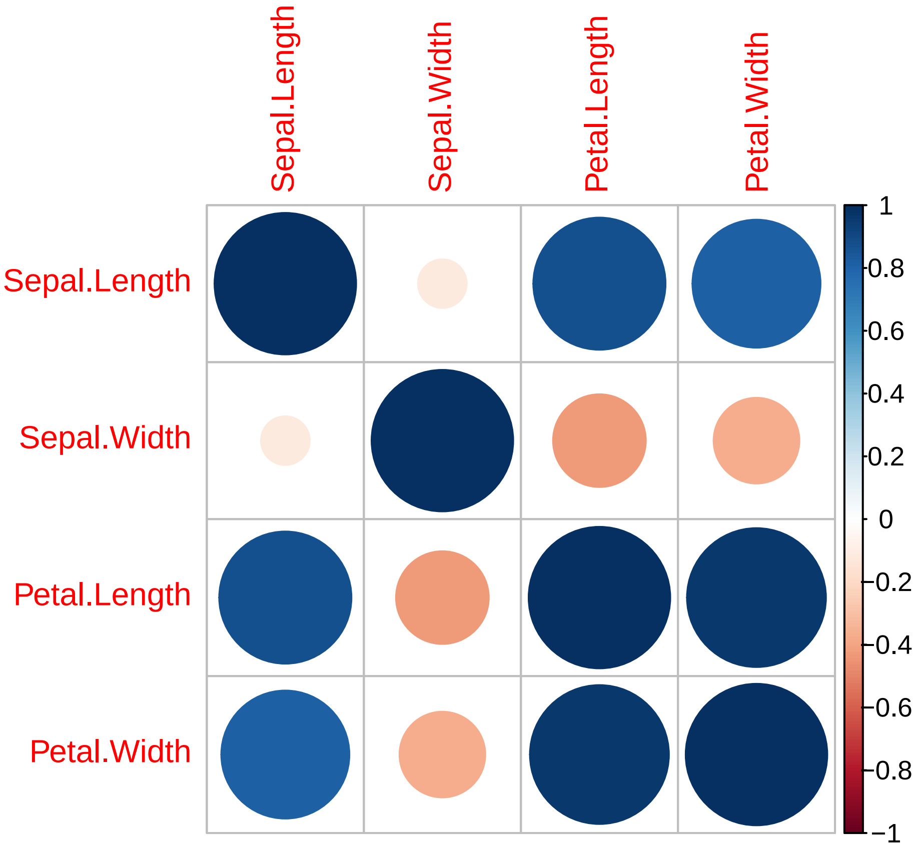
Solution: Possible predictors are sepal.width, sepal.length, petal.width, and petal.length. The response is a categorical variable indicating the type of flower for each data point. We plot Petal.Length against Petal.Width colored by species to illustrate:



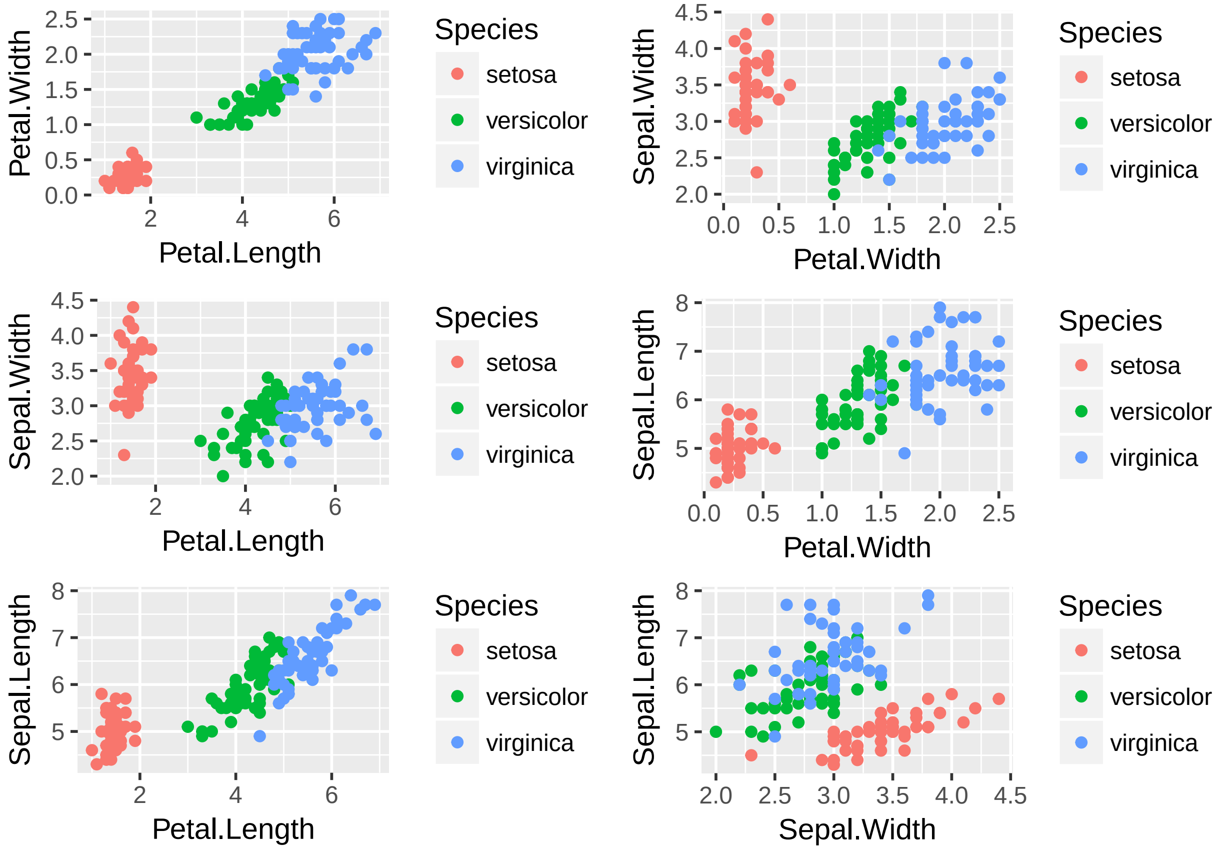
We are not going to use the Species information to help us cluster the data, but it helps us illustrate what the problems will be for the unsupervised methods. While the Setosa species is located a large distance away from either the Versicolor or Virginica data, the latter two are much closer together so there is going to be some amount of error in the predictions.

Data Prep: Scale the data:

Correlation plot of the data:

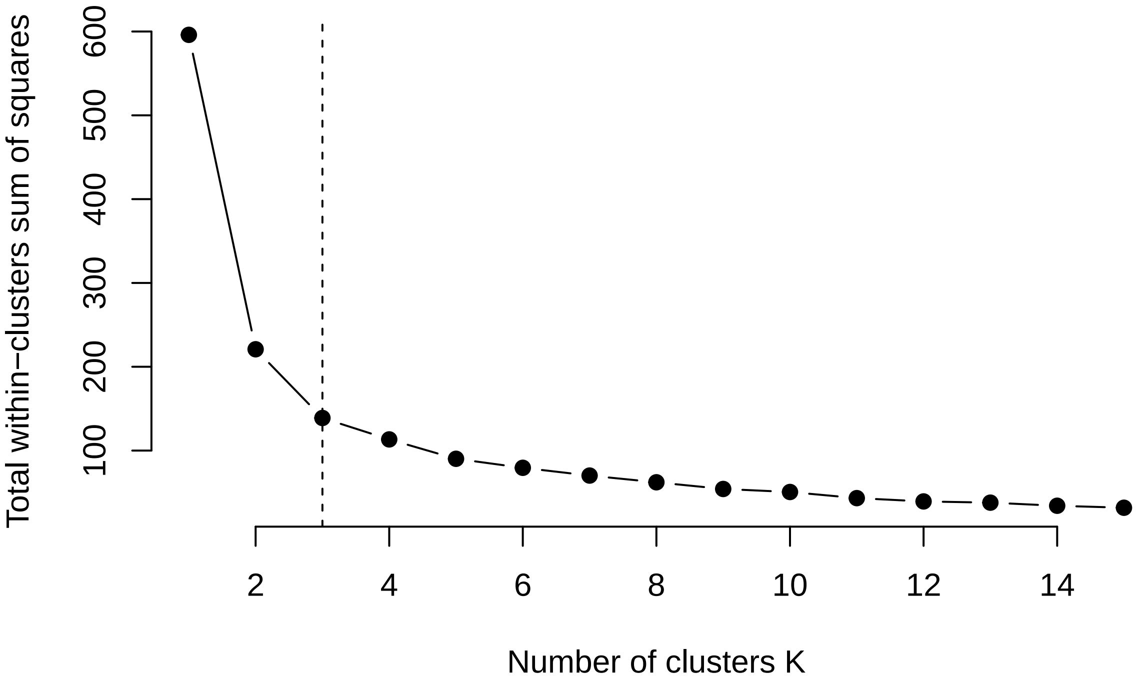


Graphs of the potential pairs of predictors (unscaled):



The two predictors with the highest correlation are Sepal.Length and Petal.Length. We will investigate later the different combinations of predictors that yield the greatest model accuracy.

Now we select the number of clusters based on the Elbow method (using scaled data):



This means that we should choose $k = 3$. Now we try the base model with 4 predictors on the data:

```
## [1] 0.8933333
```

The question now is whether we can achieve a higher accuracy from a smaller set of predictors. We first look at each predictor individually:

```
## [1] 4
```

```
## [1] 0.96
```

Our results show that by using Petal.Width as the lone predictor, we achieve an accuracy of 0.96. Now we try combinations of 2 predictors:

```
## [1] 1
```

```
## [1] 0.9466667
```

The model that yielded the highest accuracy was Model 1 (Petal.Length, Petal.Width) with an accuracy of 0.9466667.

Now we try 3 predictors (thankfully only 4 combinations!) similarly:

```
## [1] 4
```

```
## [1] 0.94
```

The 3-predictor model with highest accuracy was Model 4 (Petal.Width, Petal.Length,Sepal.Width) with an accuracy of 0.94.

Summary: our best overall model was with Petal.Width as the lone predictor. This yielded an accuracy of 0.96.

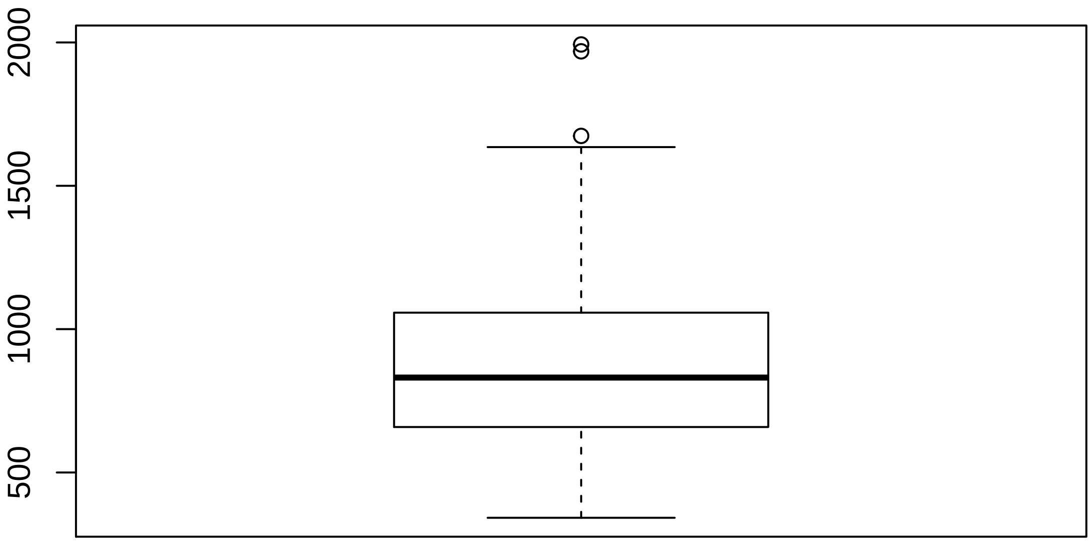
Further work ideas: I would have liked to create an elbow graph for each of the 15 subsets of predictors and then plotted them all on either the same graph or a small number of graphs, but this seems like a lot of work and I don't know how informative it would be.

Question 5.2

Using the crime dataset, 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.

This dataset gives state values of crime occurences for 47 states in the year 1960. We start our by plotting a boxplot of the Crime variable:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   342.0   658.5   831.0   905.1  1057.5  1993.0 
```



```
## [1] 26
```

```
##       M So   Ed Po1  Po2    LF   M.F Pop  NW    U1  U2 Wealth Ineq
## 26 13.1  0 12.1  16 14.3 0.631 107.1   3 7.7 0.102 4.1   6740 15.2
##        Prob    Time Crime
## 26 0.041698 22.1005  1993
```

We see that there are two potential outliers that have crime values of 1969 and 1993. Investigating these points further shows that the state with Crime equal to 1993 is a state with a population of roughly 300,000 and it's not a southern state. Consulting state population data from 1960, we see that the state is likely Wyoming. The state with Crime equal to 1969 however has a 1960 population of 15.7 million people - it surely has to be California. Now we test the 1993 outlier with the Grubbs test:

```
##
```

```
##  Grubbs test for one outlier
##
## data:  crime$Crime
## G = 2.81290, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
```

This yields a p-value of 0.07887 - while small, it is not below the standard threshold of significance (0.05) and thus we do not reject the null hypothesis which is that there are no outliers. Despite this statistical test, we would need further information about the desired application before ultimately deciding if these points were outliers or not. For example, we could look at the distribution of crime over the years previous and following 1960 for this state, and that would give us a better idea if this value of 1993 was anomalous or not.

Question 6.1

We can use the CUSUM chart to detect increases of ocean temperatures over, say, a 20-year period. There is a lot of data about this; we could take daily data regarding ocean temperatures, collect the average over a period of time (a month, 6 months or a year) and finally we would then compute S_t looking for temperature increases. We would choose low values for both C and T because the observed effect would probably be quite small. As we see in the following problem, both C and T are functions of the variance. So we would choose $C = C_{mult} \times sd_{time}$ and similarly $T = T_{mult} \times sd_{time}$. We would guess that the multiplying constants will need to be very small (e.g. 0.01) to find an effect.

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 use R if you'd like, but it's straightforward enough that an Excel spreadsheet can easily do the job too.

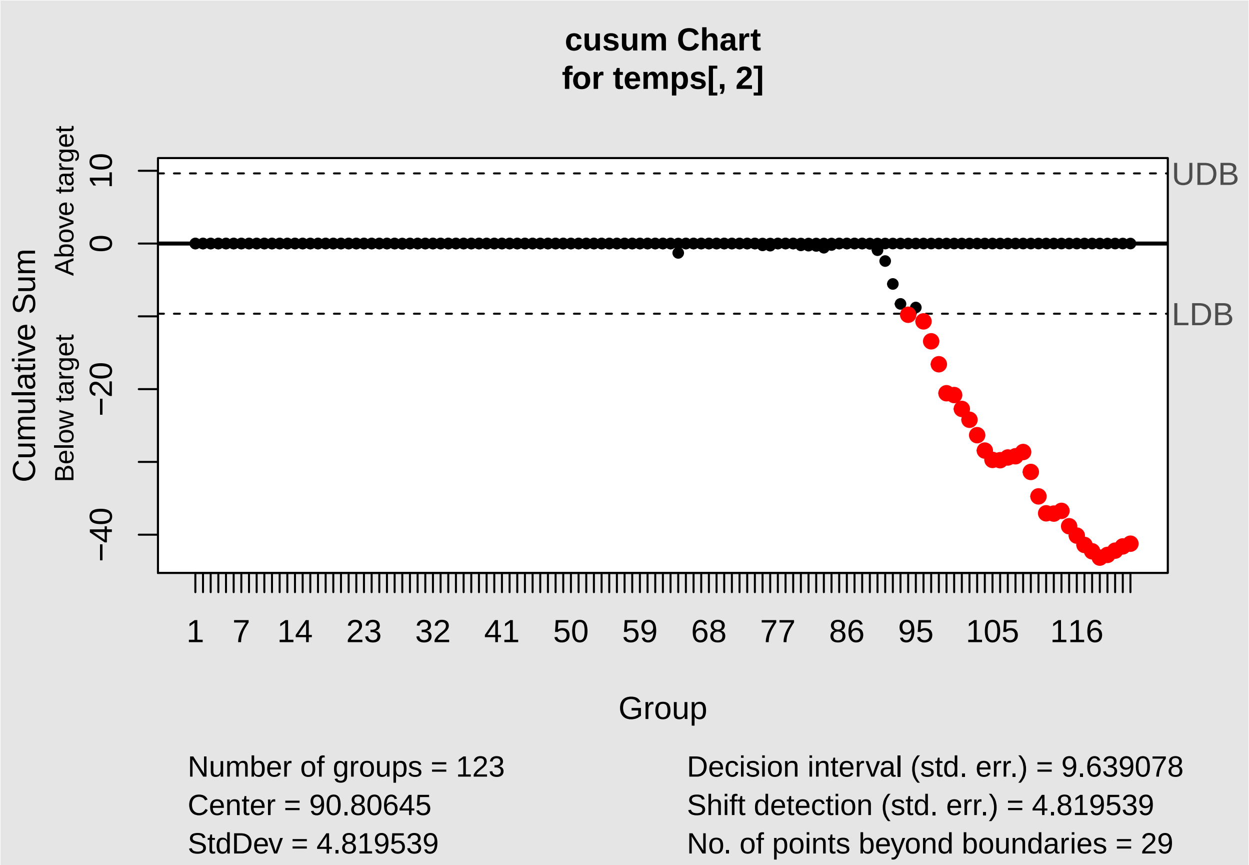
We find that the hottest month of the year for Atlanta is July, so we use the average of all July temperatures to get $\mu = 88.75$ with $\sigma = 4.6207109091$. The general CUSUM formula for decrease detection is as follows:

$$S_t = max\{0, S_{t-1} + (\mu - x_t - C)\}$$

Here is the CUSUM plot for the year 1996:

```
## Package 'qcc' version 2.7
```

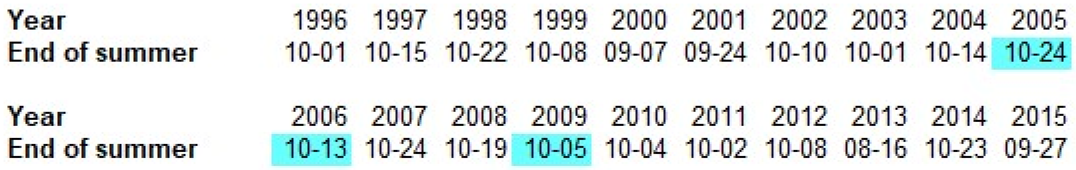
```
## Type 'citation("qcc")' for citing this R package in publications.
```



```
## List of 14
##  $ call              : language cusum(data = temps[, 2], center = july_mean_1996, std.dev = july_sd_19
##  $ type              : chr "cusum"
##  $ data.name         : chr "temps[, 2]"
##  $ data              : int [1:123, 1] 98 97 97 90 89 93 93 91 93 93 ...
##   ..- attr(*, "dimnames")=List of 2
##  $ statistics        : Named int [1:123] 98 97 97 90 89 93 93 91 93 93 ...
##   ..- attr(*, "names")= chr [1:123] "1" "2" "3" "4" ...
##  $ sizes             : int [1:123] 1 1 1 1 1 1 1 1 1 1 ...
##  $ center            : num 90.8
##  $ std.dev           : num 4.82
##  $ pos               : num [1:123] 0 0 0 0 0 0 0 0 0 0 ...
##  $ neg               : num [1:123] 0 0 0 0 0 0 0 0 0 0 ...
##  $ head.start        : num 0
##  $ decision.interval: num 9.64
##  $ se.shift          : num 4.82
##  $ violations        :List of 2
##  - attr(*, "class")= chr "cusum.qcc"
```

While this graph is nice, it would be a chore to produce 21 graphs and then analyze them for the relevant information. We turn to Open Office so we can do this more compactly; please reference the attached ods document for an implementation of the CUSUM method.

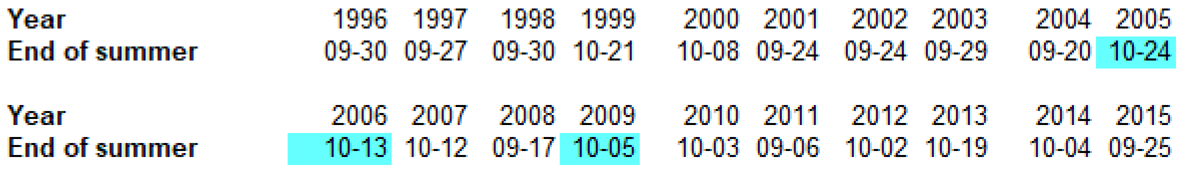
The first tab of this spreadsheet represents the situation where we have taken the cumulative mean and standard deviation for July in all of the different years. There are several problems with this approach however: choosing a large C value of $C = T = 3 \times \sigma$ produced several false positives in the detection process. We then tried another more careful approach.



Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
End of summer	10-01	10-15	10-22	10-08	09-07	09-24	10-10	10-01	10-14	10-24

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
End of summer	10-13	10-24	10-19	10-05	10-04	10-02	10-08	08-16	10-23	09-27

Figure 1: mu =88.75 for all years



Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
End of summer	09-30	09-27	09-30	10-21	10-08	09-24	09-24	09-29	09-20	10-24

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
End of summer	10-13	10-12	09-17	10-05	10-03	09-06	10-02	10-19	10-04	09-25

Figure 2: mu is calculated yearly

In this more careful approach, we calculate the mean and standard deviation for each year. Then we set up C and T to be some constant times the standard deviation for each year. We set C and T both to be 3 times the standard deviation, as before- we want to be sure that summer has ended and this is not a false positive. Our results for the two methods are as follows:

Method 1: Mean for entire 20-year span of July temps

Method 2: Mean calculated for each year

Interestingly enough, we found that 3 of the days predicted were exactly the same: in 2005, 2006, and 2009. However, most of the results for Method 2 seemed more reasonable. I am not familiar with the Atlanta region though, so I could not say for sure based on experience.

(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).

In the OpenOffice document tab 3, we take μ and σ to be the average temperature and standard deviation of July 1996, respectively. Then we see if we can detect an increase in the CUSUM in the values for the subsequent years. The model parameters T and C can both be adjusted by a factor times the standard deviation, as before; we need to use a smaller C value since the change in the process will be more negligible. I highlighted the C_{mult} to indicate that the values of S_t will all update when this parameter is changed.

We first choose $C = T = 1 \times sd_{1996}$. With these values, we observe that $S_t \geq T$ for much of 2012 - this was an exceptionally hot month, so this makes intuitive sense. Now we see if we can detect a smaller shift: we choose $C = 0.9 \times sd_{1996}$ while leaving $T = \times sd_{1996}$, and this results in small changes being detected from 19-22 of July in the year 2000, as well as the abnormally hot month of July 2012.

References:

Multiplot code: http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/

Question 4:

(1) Heirarchical Clustering in R https://www.r-bloggers.com/hierarchical-clustering-in-r-2/

(2) Kmeans Clustering in R https://www.r-bloggers.com/k-means-clustering-in-r/

(3) Determining the optimal number of clusters http://www.sthda.com/english/wiki/print.php?id=239

Question 5:

https://en.wikipedia.org/wiki/1960_United_States_Census