ISYE6501 Homework week 2

Dylan Peters

Question 1

A situation from real life where a clustering model would be appropriate:

I recently worked on a project that sought to organize customer marketing data into clusters by the customers' interests and spending habits. Cluster parameters would be: How much did the customer spend in the Fast Food/Fast Casual/Fine Dining segments? Do they spend regularly for Travel? Do they have a mobile phone? Do they view pay TV e.g. HBO? Do they spend on sporting events, or rock concerts, or symphonies? These parameters would enable us to cluster the customers into similar groups and thus provide more targeted advertising, or perhaps find similar interests for recommendation.

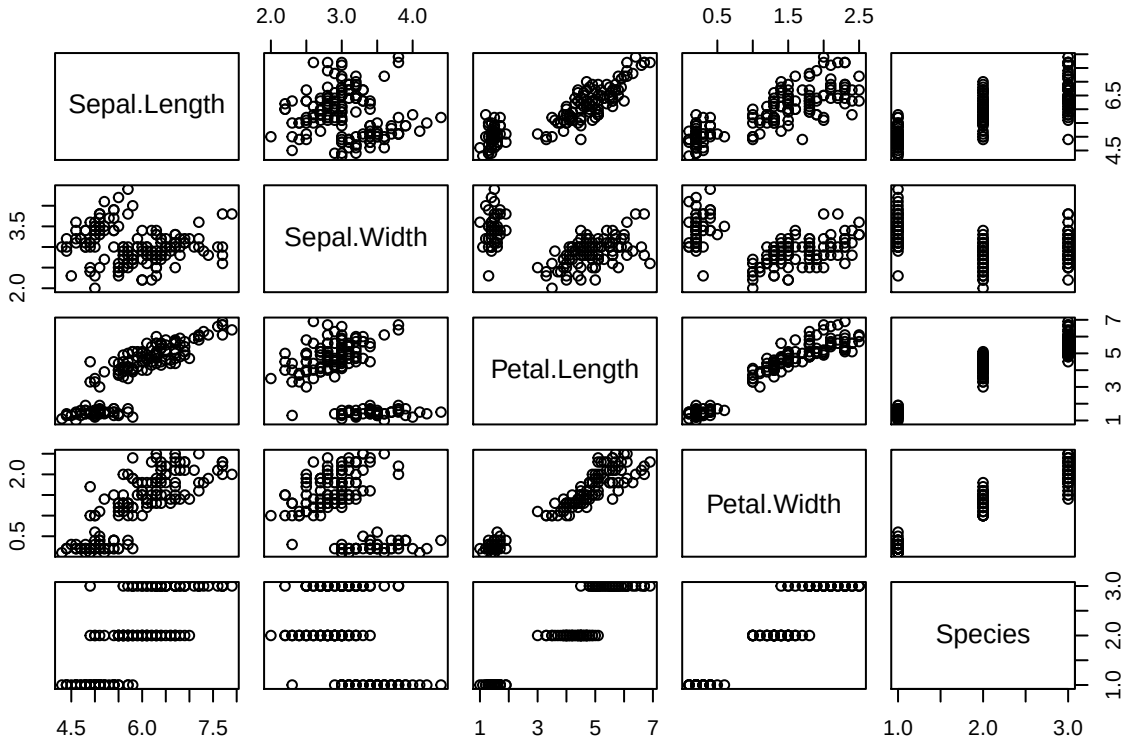
Question 2

Use kmeans to analyze the iris dataset:

First load the data and get a general idea of the features:

```
data(iris)

plot(iris)
```



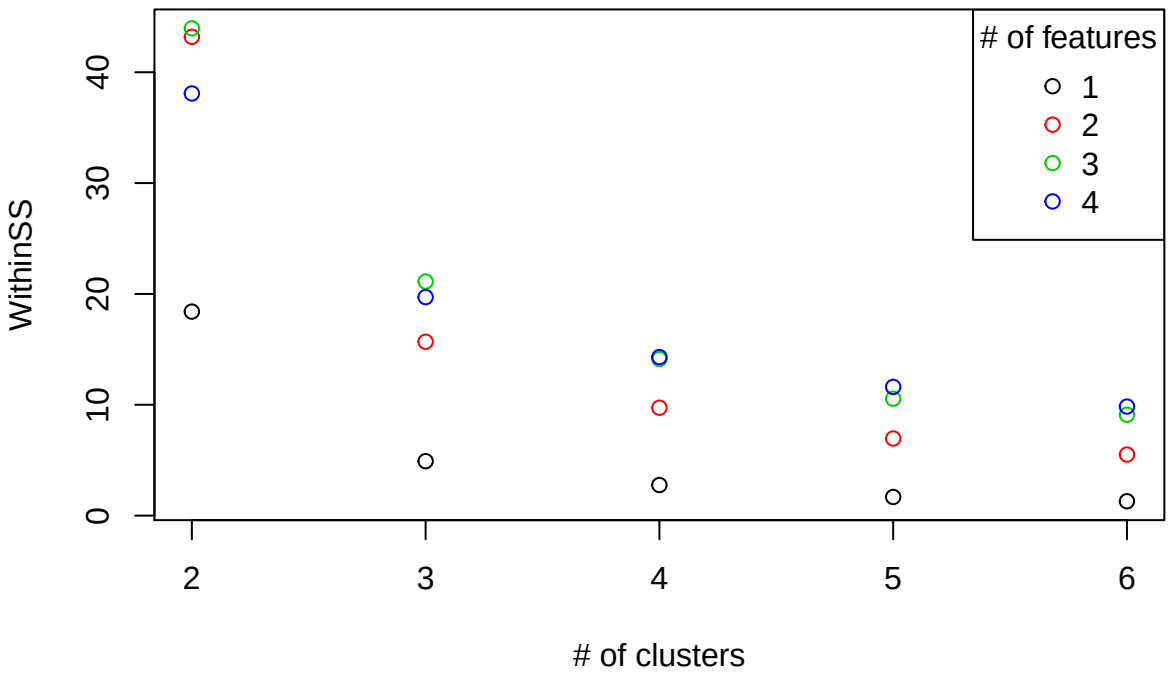
It looks like Petal.Length and Petal.Width are good clustering candidates. Of the other two, perhaps Sepal.Length could be useful. I run it here for 1-4 features and 2-6 clusters. To measure I use the total within SS divided by the number of features, in order to control for the higher error with more dimensions.

```
klist <- c(2,3,4,5,6)
featurelist <- list(c("Petal.Width"),
                    c("Petal.Length","Petal.Width"),
                    c("Sepal.Length","Petal.Length","Petal.Width"),
                    c("Sepal.Length","Sepal.Width","Petal.Length","Petal.Width"))

results <- data.frame(FeatureCount=integer(),
                      KValue=integer(),
                      WithinSS=double(),
                      BetweenSS=double())

for (features in featurelist)
{
  for (k in klist)
  {
      kmeans.model <- kmeans(iris[,features], k, nstart=20)
      results <- rbind(results,c(length(features), k, kmeans.model$tot.withinss / length(features), kmea
  }
}
colnames(results) <- c("FeatureCount","KValue","WithinSS","BetweenSS")
```

This plot shows the decrease in total Within-SS for each cluster model:



The best number of clusters is 3 or possibly 4. The single feature (Petal.Width) has the best within cluster sum-of-squares, even when controling for number of features.

This shows how well the 1 feature, 3 cluster model clusters the species correctly:

```
kmeans.model <- kmeans(iris[,"Petal.Width"], 3, nstart=20)
table(iris$Species, kmeans.model$cluster)
```

```
##
##                1  2  3
##   setosa      50  0  0
##   versicolor   0 48  2
##   virginica    0  4 46
```

Setosa is properly classified, and versicolor and virginica are mostly classified into single clusters. With only 6 items misclassified, the model has an accuracy of 96%.

Question 3

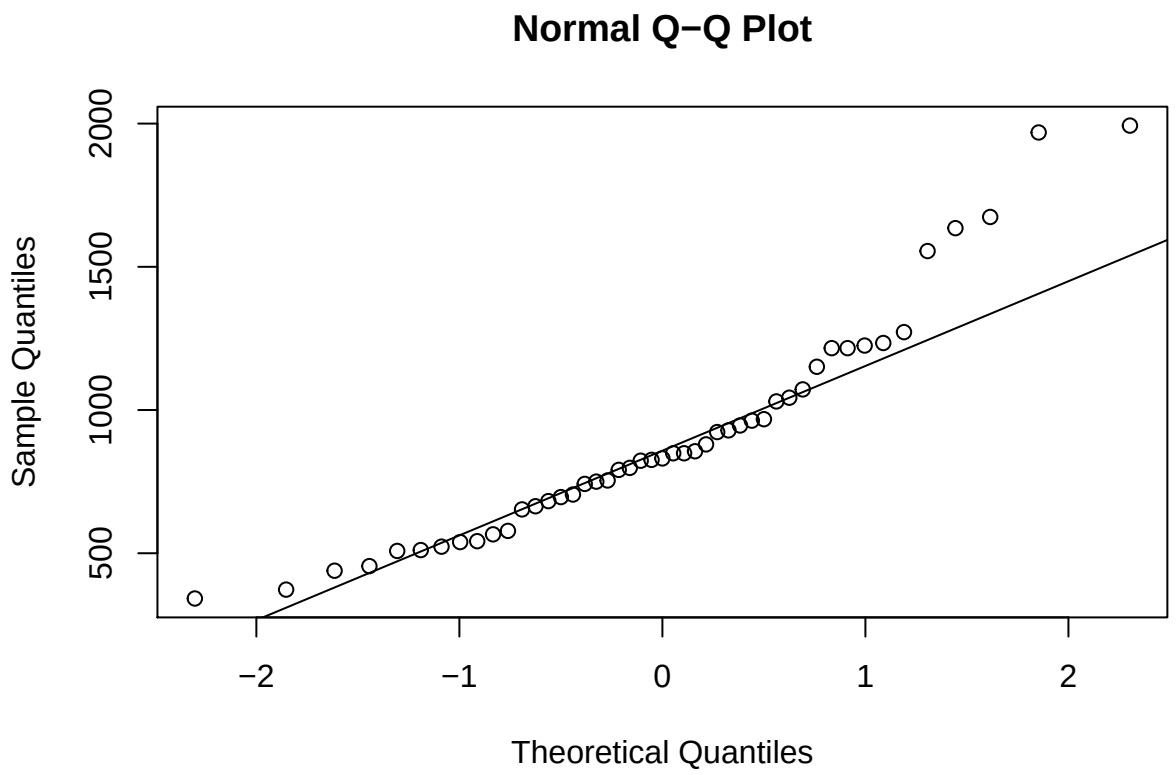
Crime Data and outliers:

Load the data:

```
crime <- read.delim("http://www.statsci.org/data/general/uscrime.txt")
```

Run a QQ plot to check for normality:

```
qqnorm(crime$Crime)
qqline(crime$Crime)
```



There is definitely some non-normality at the high end of the plot.

```
library(outliers)
grubbs.test(crime$Crime, type=10)
```

```
##
##  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
```

The highest value 1993 is not technically an outlier, but it is very close. Since $p > 0.5$, we fail to reject the null hypothesis that there are no outliers.

Question 4

Describe a problem where a Change Detection model would be appropriate:

At a previous employer, we provided an Internet interface to clients to connect to our system. We had a general expectation that response times would be less than 2 seconds. Also, we might have a specific SLA that said 90% of requests would be less than 1 second and 99% less than 2 seconds. We could use a CUSUM model to send alerts when the average time (per minute) became over a certain threshold, but we wouldn't want to send an alert just because one or two requests took longer than 2 seconds. If we set C to 2 and T to

4, then we could ignore any averages less than 2 seconds, but alert when the cumulative sum reaches 4, which would indicate that the system is slowing down and needs attention.

Question 5

1. Apply CUSUM to Atlanta temperature data to find when the summer temperature starts to cool down:

I used a Google Sheet to do the analysis. It can be found here: [https://docs.google.com/spreadsheets/d/15Pi_iS9m1wUfoNLx_y-tkaRkBLEMT7V-RWNJsZ8rMu4/edit?usp=sharing]

Steps: 1. Formatted the data into columns 2. Added extra columns with the CUSUM formula: =MAX(0, E2-(D3-$B4−B$3)) [The calculation is subtracted rather than added since we are looking for a decrease from the mean.] a. E2 is the previous calculation (S_t-1) b. D3 is current observation (x_t) c. B4 is the mean. I used 91 which was the average July temperature for 1996. d. B3 is the C value. After adjusting the C value, I landed on a value of -8. This eliminated the noise in the early observations and made a good curve that did not suddenly rise up. 3. Added a chart of the CUSUM calculations 4. Added a column for the average CUSUM for each day (over the years) and added a chart for that. 5. Determined threshold. I chose a threshold of 40 based on where the curve started to bend faster from 0.

For a threshold of 40, the cutoff date is September 26, which is just after the official equinox. An argument could be made for an earlier date, such as September 8 or 9, based on where the CUSUM average starts to move steadily away from 0.

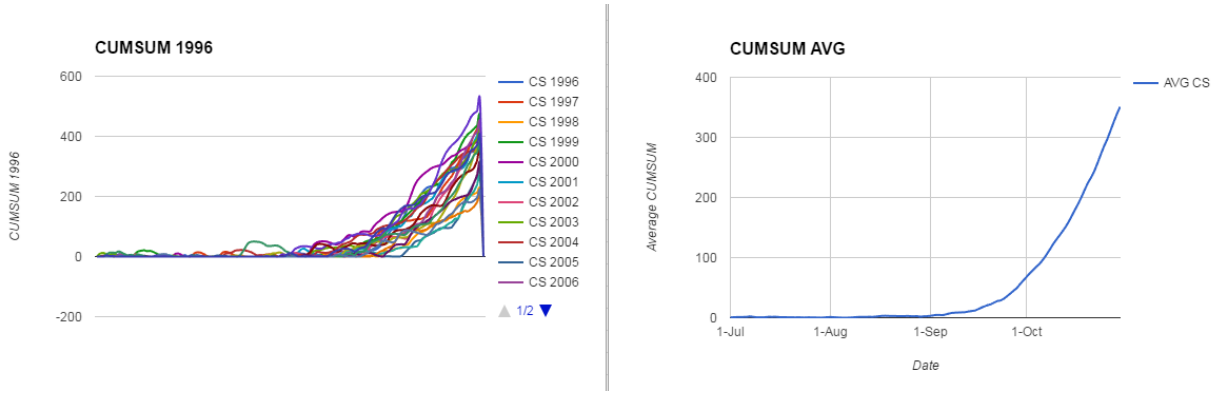


Figure 1: alt text

2. Apply CUSUM across years to determine whether summers have gotten hotter in Atlanta

Click on the second sheet in the workbook and scroll to the right ->

Steps: 1. Add data to a new sheet and add data for average temperature by month/year. 2. Add CUSUM formula: =MAX(0, Y7+(Z3-$Y12−Y$16)) a. Y7 is the previous calculation (S_t-1) b. Z3 is current observation (x_t) c. Y12 is the mean. I used the overall average for each month across years. d. Y16 is the C value. After playing with the C value, I decided on value of 0. Other values either gave an always increasing result or a result that always returned to 0. 3. Added a row for the average CUSUM for each year (over the 4 months) and added a chart 4. Determined threshold

With the months on a separate row and compared to their own average, it is easy to see how each month's CUSUM result moves–and they are often in tandem. This is consistent with the idea that it is the whole year

that is higher or lower than average and not just one month. I chose a threshold value of 6 since that's when the average CUSUM over the months moved consistently away from 0. This happened in 2011.

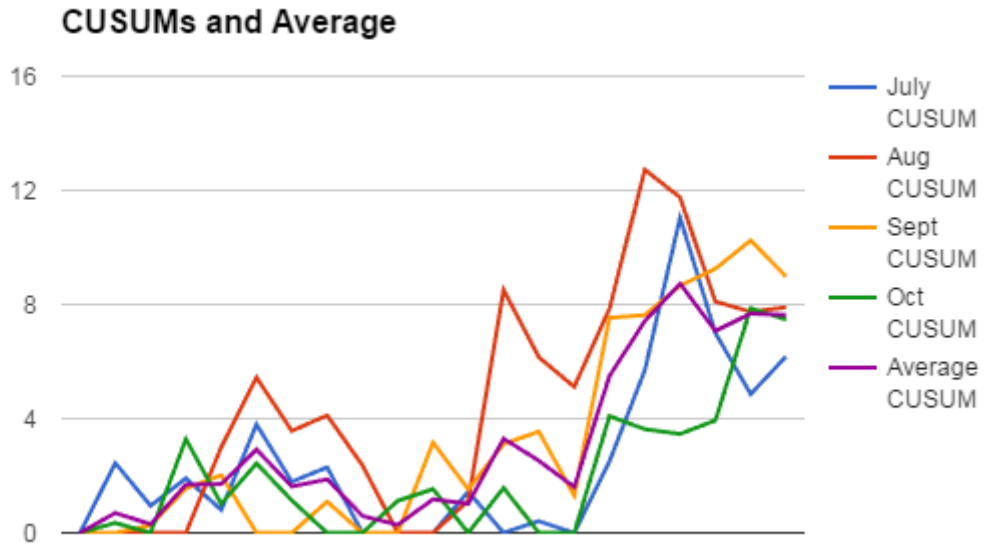


Figure 2: alt text

It is possible that the higher temperatures since 2010/2011 are just a bit of random noise, in which case a higher value of C would keep the result from reaching the threshold.