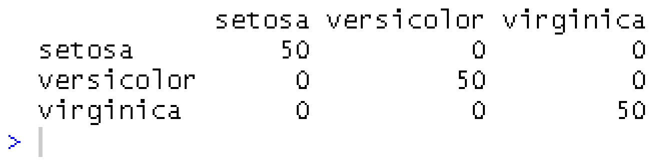
HW 2

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.

In marketing course, I remember learning about market segmentation. A clustering model could be used in identifying which market segmentation does a consumer product like a shampoo manufacturer belong too. Predictors consumer's age, income, ratings of the product, product ingredients, and how long has the product been in the market (which might be helpful in predicting consumer loyalty).

4.2 The iris data set iris.txt contains 150 data points, each with four predictor variables and one categorical response. The predictors are the width and length of the sepal and petal of flowers and the response is the type of flower. The data is available from the R library datasets and can be accessed with iris once the library is loaded. It is also available at the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Iris). The response values are only given to see how well a specific method performed and should not be used to build the model.
Use the R function kmeans to cluster the points as well as possible. Report the best combination of predictors, your suggested value of k, and how well your best clustering predicts flower type.

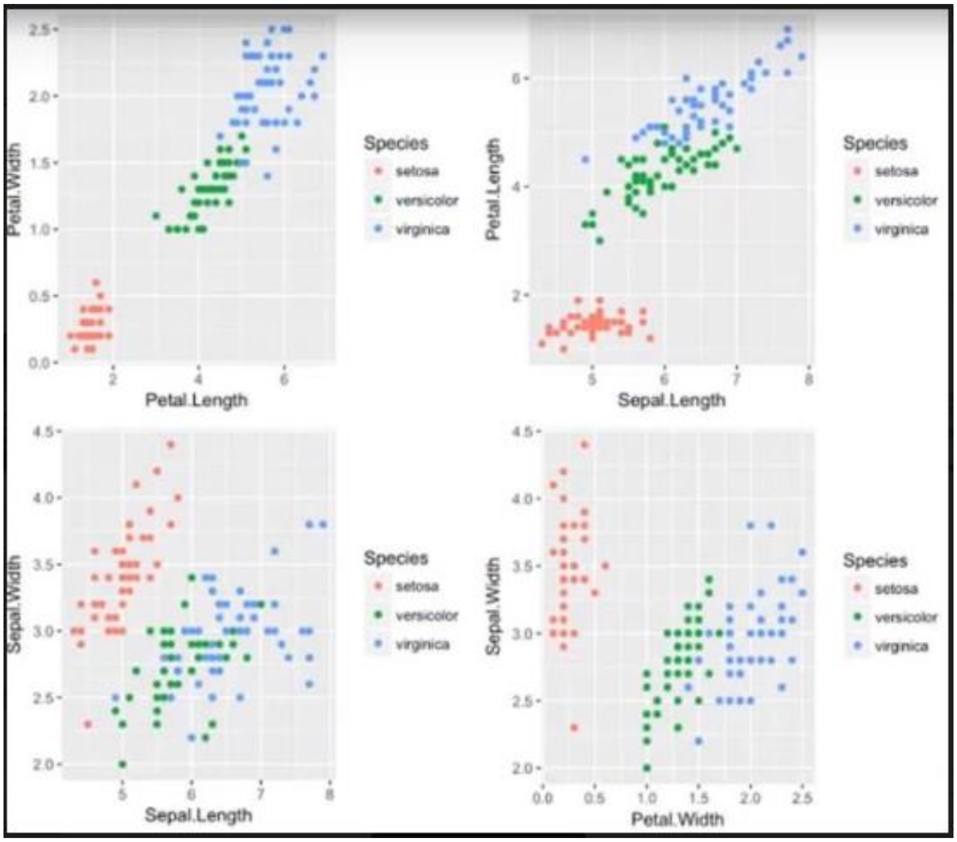
```
#Open and read iris data table.
data <- read.table("4.2irisSummer2018.txt", stringsAsFactors = FALSE, header = TRUE)
#Total Number of elements  each flower type
table(data[,5], data[,5])
```



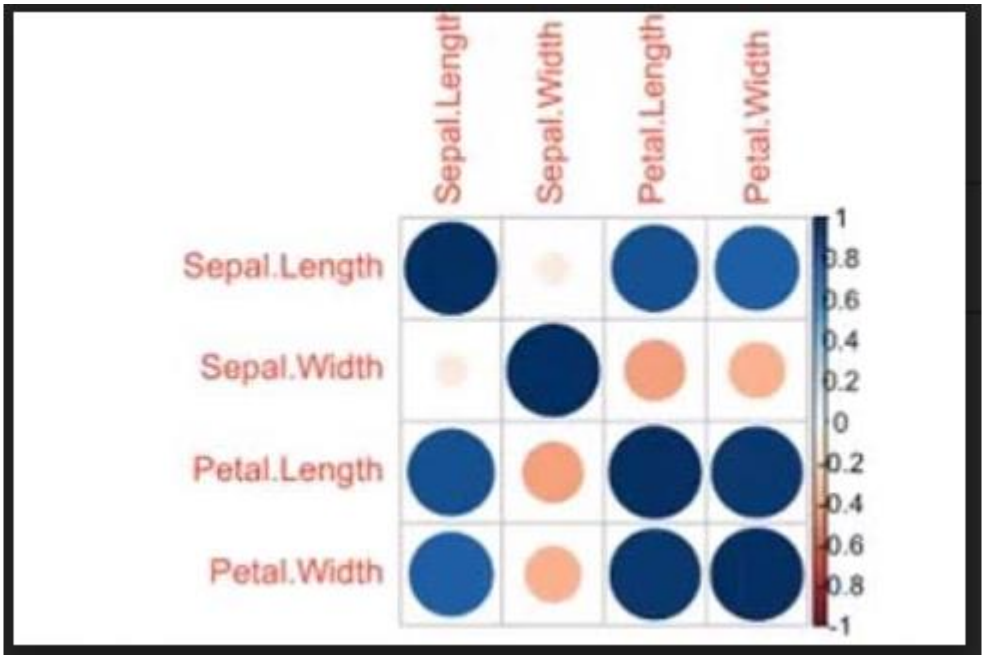
```
             setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         50         0
  virginica       0          0        50
> 
```

#Before doing the Kmeans model, we can look at the variable through ggplot and corrplot.
#Plotting different predictor variables in a graph gives the below results. From the below graph, I can estimate that there is a correlation between, petal length and width, sepal length and petal length, and petal width and sepal width. There is a week correlation between sepal length and sepal width since the colored points are mixed up and not segregated very clearly.

```
p1 <- ggplot(data,aes(Petal.Length, Petal.Width, color = Species)) + geom_point()
p2 <- ggplot(data,aes(Sepal.Length, Setal.Width, color = Species)) + geom_point()
p3 <- ggplot(data,aes(Sepal.Length, Petal.Length, color = Species)) + geom_point()
p4 <- ggplot(data,aes(Petal.Width, Sepal.Width, color = Species)) + geom_point()
multiplot(p1, cols=2)
```



# From the corrplot, you can see a strong correlation between variables with darker blue circles.
corrplot(cor(data[,2:5]))



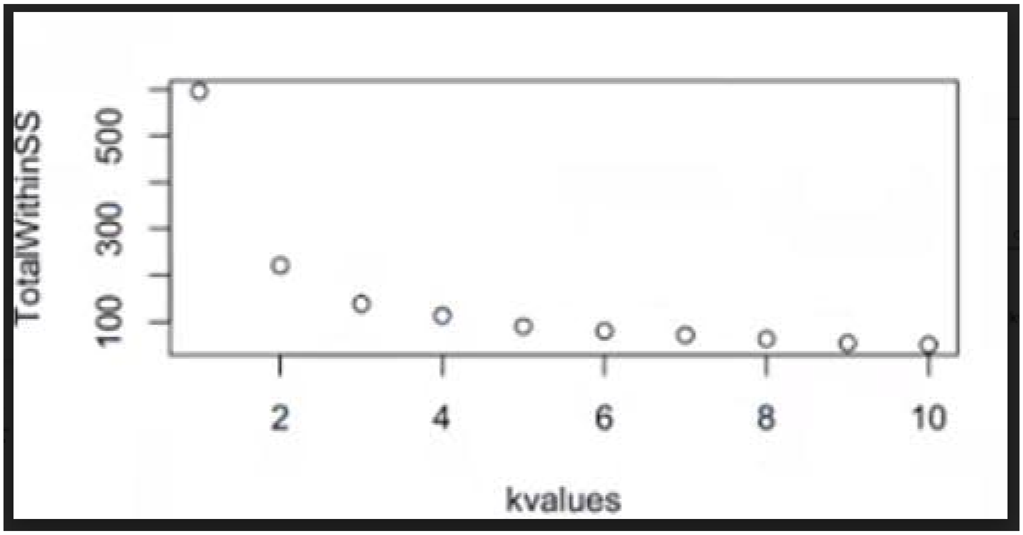
# Using K means cluster model, we can create a model like this below:
scaled <- scale(data[,2:5])
#The diffrence between cluster center and the point in a cluster
irisCluster <- vector(mode = "list", length = 10)
TotalSS <- rep(0, length = 10)

```
TotalWithinSS <- rep(0, length = 10)

for (k in 1:10) {
  irisCluster[[k]] <- kmeans(scaled, k, nstart = 20)
  TotalSS[k] <- irisCluster[[k]]$totss
  TotalWithinSS[k] <- irisCluster[[k]]$tot.withinss
  kvalues[k] <- k
}
```

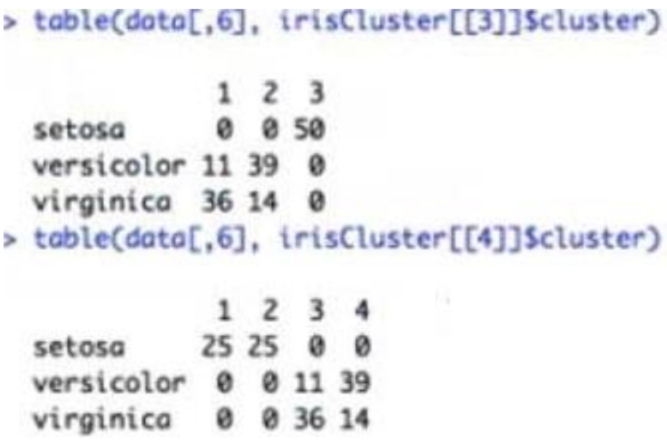
#From the code below, I can plot an elbow diagaram. From this graph, we can see that k values should be 3 or 4, since the change is minimum

```
plot(kvalues, TotalWithinSS)
```



# In order to determine the flower type, I run the following code for k values 3 and 4.

```
table(data[,5], irisCluster[[3]]$cluster)
table(data[,5], irisCluster[[4]]$cluster)
```



```
> table(data[,6], irisCluster[[3]]$cluster)

              1  2  3
  setosa      0  0 50
  versicolor 11 39  0
  virginica  36 14  0
> table(data[,6], irisCluster[[4]]$cluster)

              1  2  3  4
  setosa     25 25  0  0
  versicolor  0  0 11 39
  virginica   0  0 36 14
```

#I ran the below code to determine accuracy to get a result around 3.

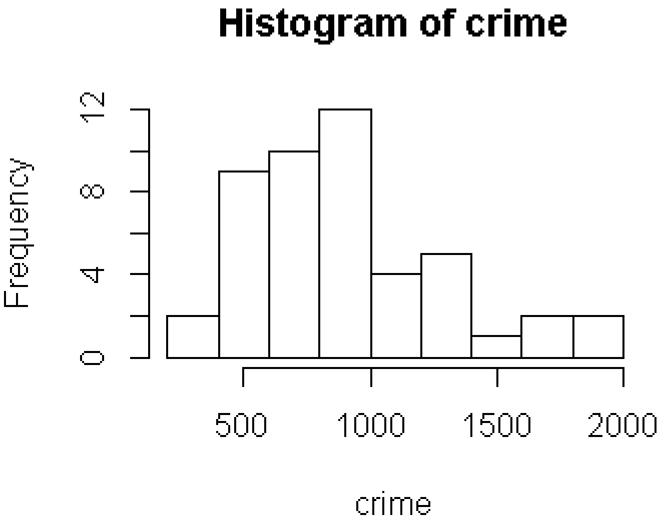
```
(max(confusion_matrix[,1]+max(confusion_matrix[,2]+max(confusion_matrix[,3]))/nrow(data)
```

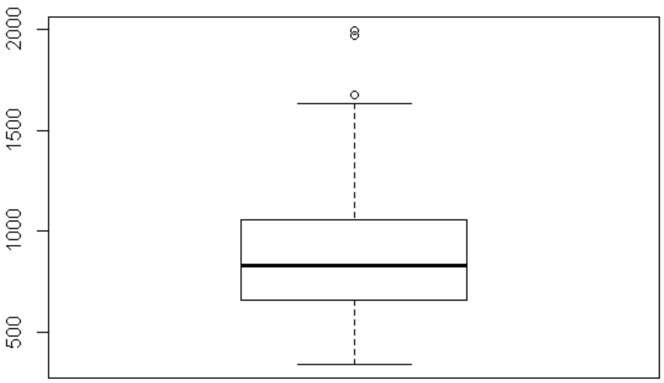
5.1 Using crime data from the file `uscrime.txt` (http://www.statsci.org/data/general/uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), test to see whether there are any outliers

in the last column (number of crimes per 100,000 people). Use the `grubbs.test` function in the `outliers` package in R.

```
# open US Crime data for summer 2018.
data2 <- read.table("5.1uscrimeSummer2018.txt", stringsAsFactors = FALSE, header = TRUE)
# Show values under Crime column only
crime <- data2[,"Crime"]

#Seeing outliers in Crime data via histograms and box plots gives us a rough idea on how many outliers there might be. It looks like between 1500 and 2000 there are roughly around 3 outliers.
hist(crime)
boxplot(crime)
```





```
#Run the Grubbs test for 2 outliers on opposite tail
test <- grubbs.test(crime, type = 11, opposite = TRUE)
test
#G = 4.26880, U = 0.78103, p-value = 1. This means the low and the high point are not outliers. If there are outliers, the p values will be very low like .5%.
#alternative hypothesis: 342 and 1993 are outliers

#Run the Grubbs test for 2 outliers
test <- grubbs.test(crime, type = 11, opposite = FALSE)
test
#G = 4.26880, U = 0.78103, p-value = 1. No outliers.
#alternative hypothesis: 342 and 1993 are outliers

#Run the Grubbs test for 1 outliers on one tail
test <- grubbs.test(crime, type = 10, opposite = TRUE)
test
#G = 1.45590, U = 0.95292, p-value = 1. No outliers.
#alternative hypothesis: lowest value 342 is an outlier

#Run the Grubbs test for 1 outliers on one tail
test <- grubbs.test(crime, type = 10, opposite = FALSE)
test
#G = 2.81290, U = 0.82426, p-value = 0.07887. Since p value is much lower, we can accept that there are outliers. As per this model, there should be roughly around three outliers which is very similar to what is
```

displayed in the graph. The highest value of the outlier is also close to 2000 which is displayed in the box plot graph.
#alternative hypothesis: highest value 1993 is an outlier

6.1 Describe a situation or problem from your job, everyday life, current events, etc., for which a Change Detection model would be appropriate. Applying the CUSUM technique, how would you choose the critical value and the threshold?

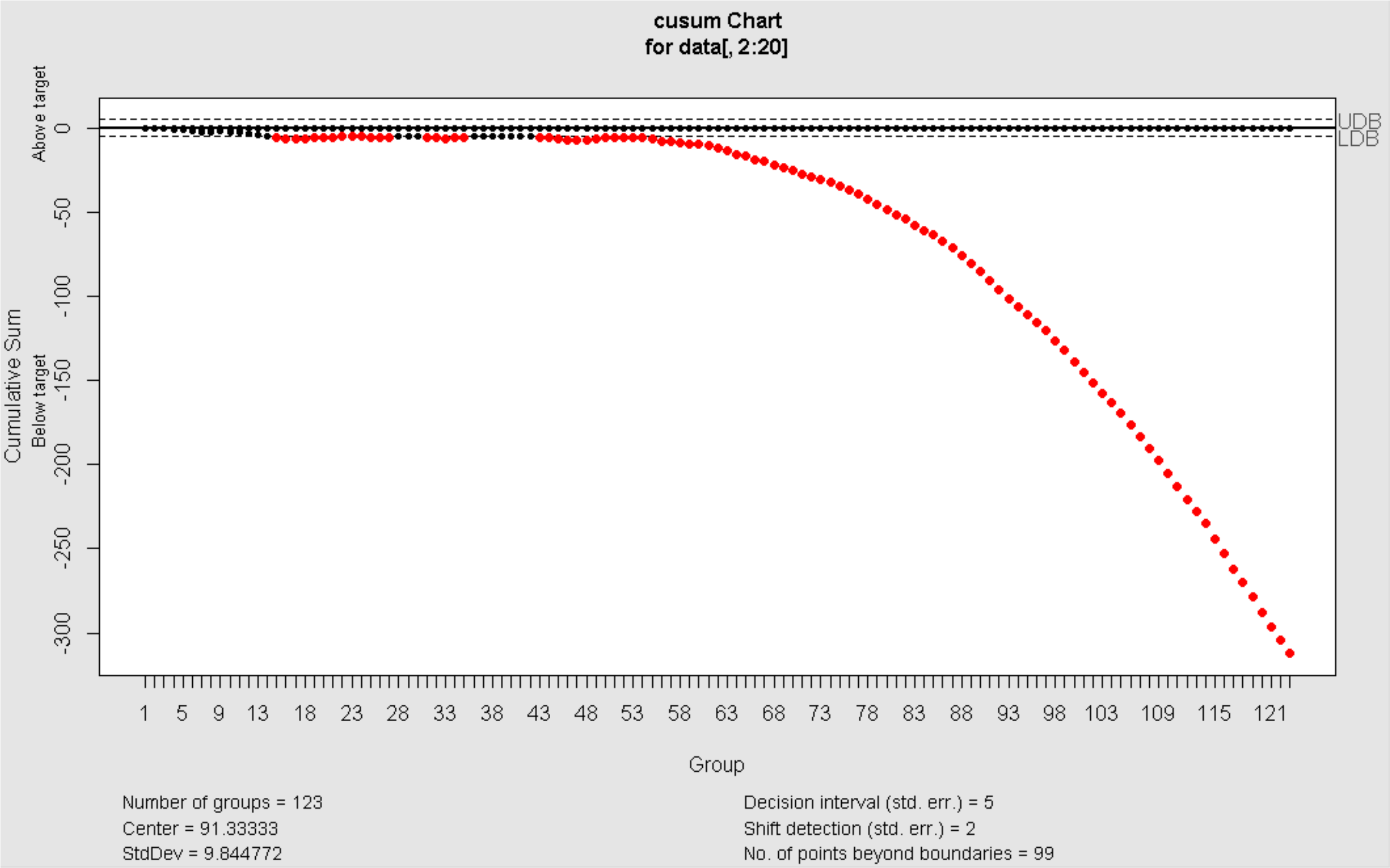
A change detection model could be used in detecting change of pH level when you add acidic solution to a base. The pH scale measures how acidic or basic a substance is. The pH scale ranges from 0 to 14. A pH of 7 is neutral. A pH less than 7 is acidic. I would choose my critical value to be 7 since it's the midpoint for a substance to either acidic or basic and the threshold will be 14 since it's the highest pH value a substance can have.

6.2.1Using 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 get the data that you need from the file temps.txt or online, for example at http://www.iweathernet.com/atlanta-weather-records or https://www.wunderground.com/history/airport/KFTY/2015/7/1/CustomHistory.html . You can use R if you'd like, but it's straightforward enough that an Excel spreadsheet can easily do the job too.

6.2.2Use a CUSUM approach to make a judgment of whether Atlanta's summer climate has gotten warmer in that time (and if so, when).

```
#I have combined 6.2.1 and 6.2.2 together
library(qcc)
# open 6.2tempsSummer2018.txt
data <- read.table("6.2tempsSummer2018.txt", stringsAsFactors = FALSE, header = TRUE)

#Use CUSUM function. center value is derived from the first month July.
centervalue <- mean(data[1:30,2])
stdev <- sd(data[1:30, 2])
# I am taking the cusum value of all the data points from July to oct from 1996 to 2015.
cusum(data[,2:20], center = centervalue, std.dev = 2*stdev, se.shift=2, plot = TRUE)
# From the graph below, it looks like there is a significant change starting from 73. Using excel, I took an average of all the temperatures per day from 1996 to 2015 and filtered from lowest to highest. The month of October 17th (onwards) would be a good estimate of when unofficial summer ends and winter starts. Before October 17th, the temperatures starts increasing.
```



Number of groups = 123
Center = 91.33333
StdDev = 9.844772

Decision interval (std. err.) = 5
Shift detection (std. err.) = 2
No. of points beyond boundaries = 99