Georgia Tech - SU18: Introduction to Analytics Modeling

Homework Assignment: Week 2

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.

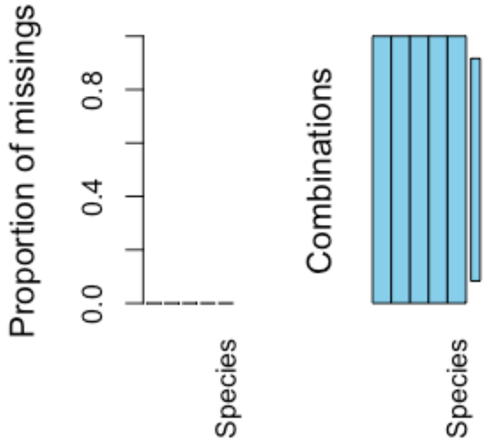
I work as a portfolio manager investing mostly in funds. We invest in a range of strategies and asset classes. As we build portfolios for our clients, we would look to have a diversified portfolio across asset classes and strategies, but we have also clients who want strategy specific strategies. Too often, it is difficult to label a fund’s strategy because everyone has their own tweak or approach, also because of the use of different financial instruments (within the same strategy – derivatives vs cash instruments) or the use of leverage or not, the concentration of the portfolio, and hence as a result risk / return profiles of these products are not necessarily the same. In a very simple way, we attempt to cluster these funds to better understand if we have more of the same in the portfolio (just with different labels) or we are truly achieving diversification.

Predictors: volatility, returns, leverage.

Question 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.

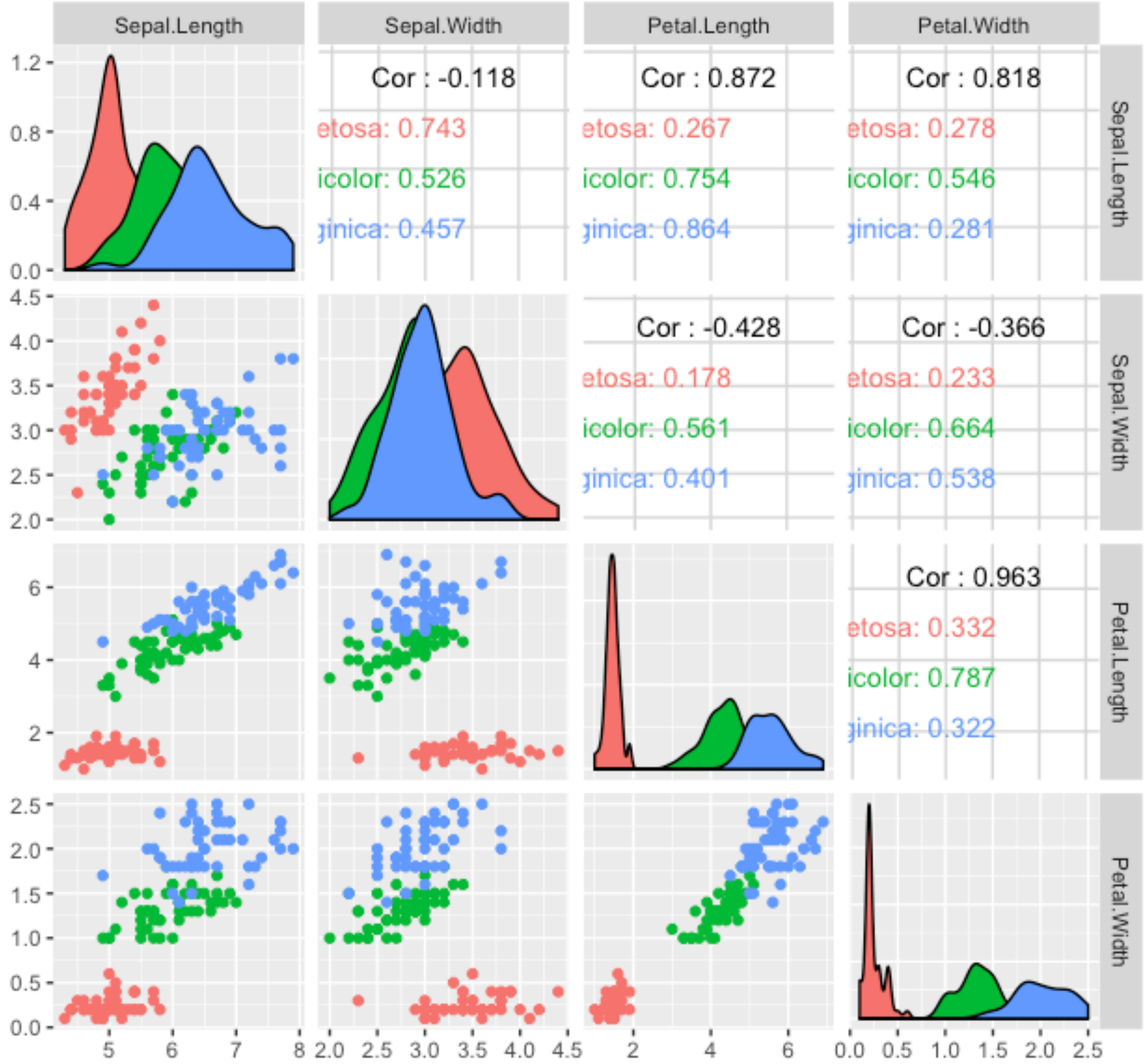
Report the best combination of predictors, your suggested value of k, and how well your best clustering predicts flower type.

First, I loaded the data(iris) and did a health check of the data (checking for missing data) and check what kind of data are available.

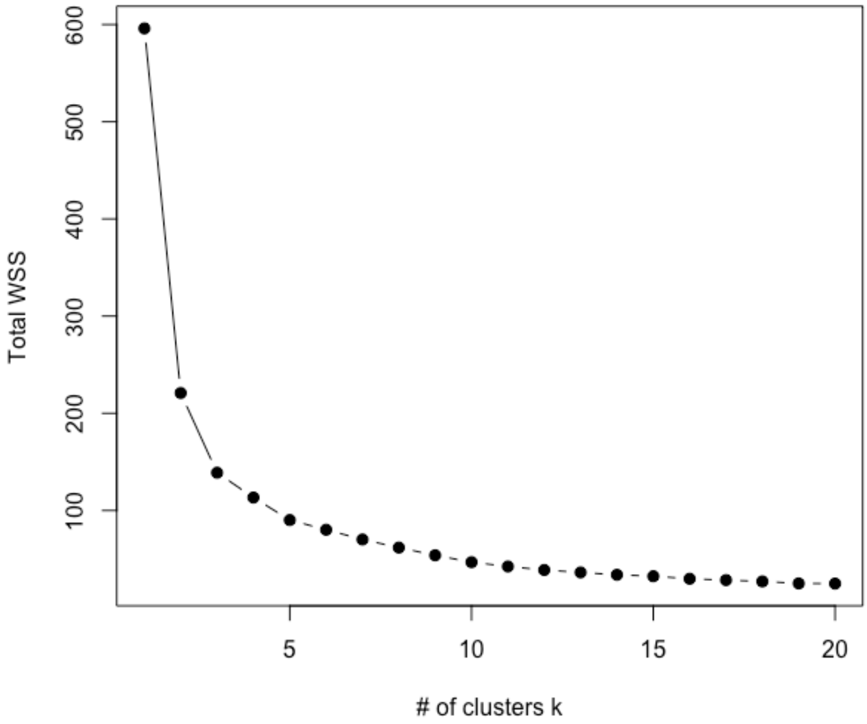


```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

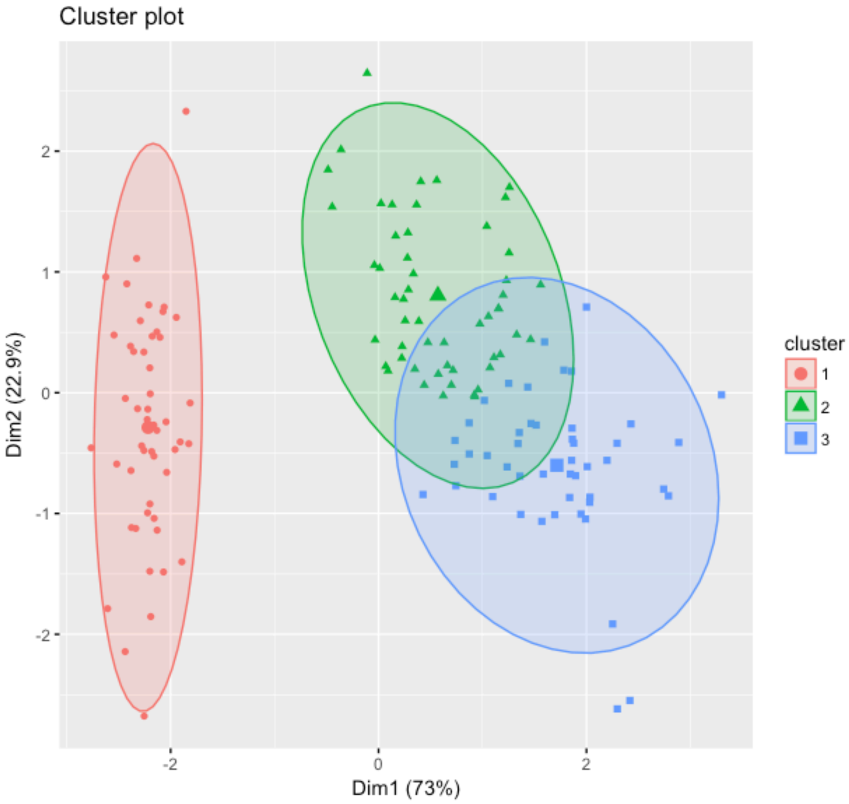
I then visualize all the data showing all combination to have a first impression of the data.



Looking at the graph, we can clearly see three defined clusters, but with 2 somewhat overlapping each other. I then finally made the “elbow method” assessment to assess how many clusters there should be.



As we can see in the graph, the point k = 3 is the appropriate number. Finally, I ran the code for kmeans clustering and plot it in a graph.

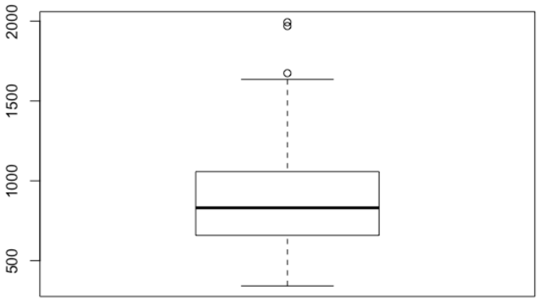


Code:

```
#import data
data("iris")

#check data
library(VIM)
str(iris)
summary(iris)
head(iris)
aggr(iris) #check even number of data
table(iris$Species) #other way to check number of data is even
#Plot data to visualize
library(ggplot2)
set.seed(42)
ggpairs(iris, columns = 1:4, mapping=aes(colour=Species))

#kmeans - elbow
iris.scaled <- scale(iris[, -5]) #scale data
k.max <- 20 #max k-clusters
wss <-sapply(1:k.max, function(k){kmeans(iris.scaled, k,nstart=50)$tot.withinss})
wss

plot(1:k.max,wss, type="b", pch = 19, xlab="# of clusters k", ylab="Total WSS")

#kmeans clustering
kmiris <- kmeans(iris.scaled, 3, nstart=50)
kmiris

#visualize the data
library(factoextra)
fviz_cluster(kmiris, data = iris.scaled, geom = "point", ellipse.type="norm")
```

Question 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.

I first check the data using similar procedure to the previous question. I then extracted the crime data and created a box plot (see below) to see the outliers. Seeing at the data, it's clear you don't need to look for outliers on both extremes.



I then ran the grubbs.test and got the following results:

```
Grubbs test for one outlier

data:  crime
G = 2.81290, U = 0.82426, p-value = 0.07887
alternative hypothesis: highest value 1993 is an outlier
```

The p-value is small and validates the hypothesis. This also visually confirmed through the graph.

Code:

```
#import Data
uscrimedata <- read.table("http://www.statsci.org/data/general/uscrime.txt", header=TRUE,
sep = "")

#check data
library(VIM)
str(uscrimedata)
head(uscrimedata)
summary(uscrimedata)
aggr(uscrimedata)

#Plot data to visualize
set.seed(42)
uscrimedata.scaled <- scale(uscrimedata)
head(uscrimedata.scaled)
boxplot(uscrimedata.scaled)


#assess crime data only
crime<-uscrimedata$Crime
summary(crime)
boxplot(crime)

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

Question 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?

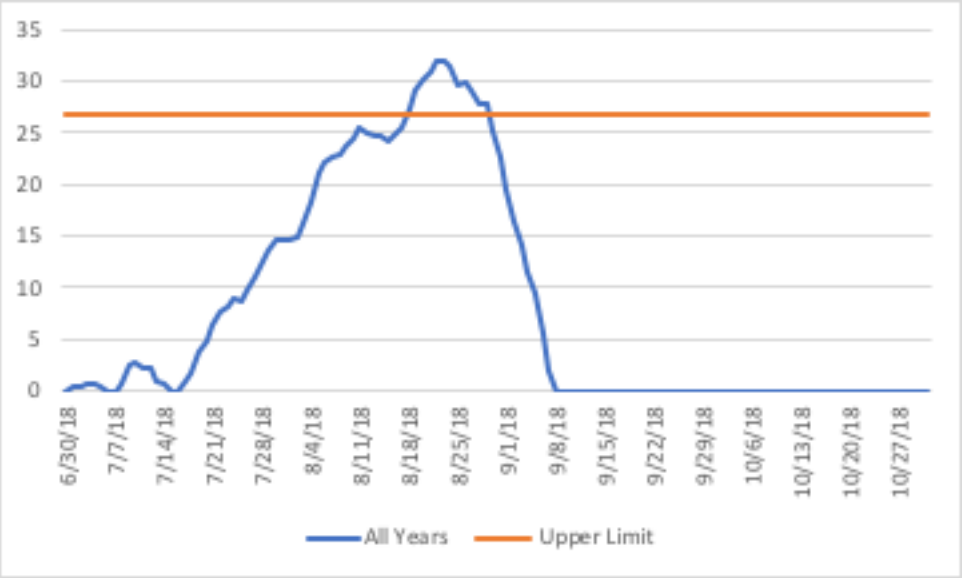
As part of my work, I monitor and use economic data to validate some of our market assumptions for our asset allocation models as well as making assessment if a change in trend is due to seasonality, earning cycle or truly structural / break in the trend.

If looking at seasonality, the threshold could be an average performance of historical calendar data (month to month for instance). Compared to the same past period, we could assess for changes in seasonality. This would potentially also allow us to identify if recent performance has been an outlier or not. Based on the exercises we are doing for this class, choosing the critical value as well as the threshold is a difficult one. For the critical value, I have been using 4 x the standard deviation, but as here we are looking at financial markets, a 4 standard deviation event is quite significant, hence may bringing down the value to a 3 standard deivation event might be more rational. As for the T-value, I have been testing using the median or mean of the data.

Question 6.2

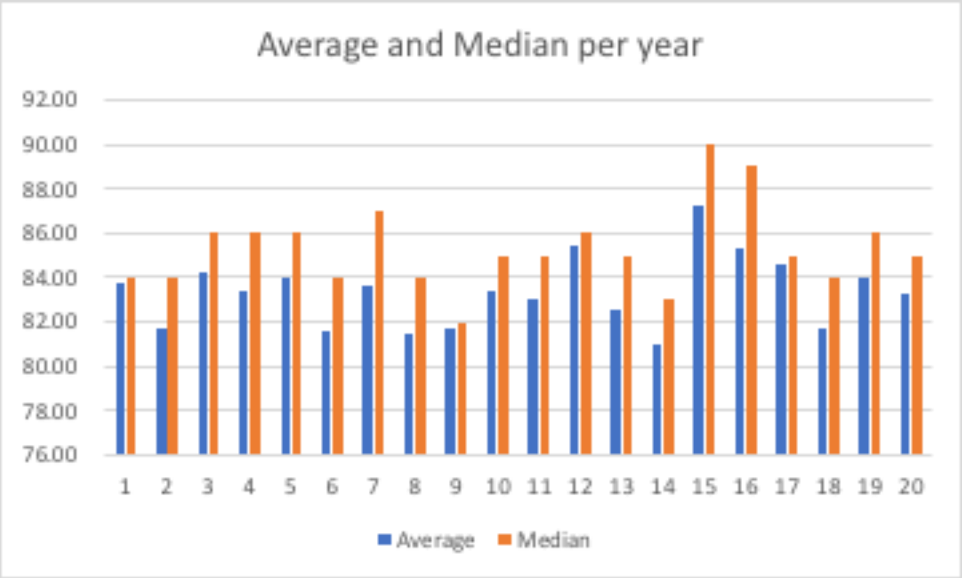
1. ***Using July through October daily-high-temperature data for Atlanta for 1996 through 2015, use a CUSUM approach to identify when unofficial summer ends (i.e., when the weather starts cooling off) each year. You can get the data that you need from the file temps.txt or online, for example at http://www.iweathernet.com/atlanta-weather-records 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.***

   I first aggregated the data by creating an average data stream for all periods. I then calculated the average, the standard deviation as well as the max and median of the data. As result, using a UCL of 26.81 (calculated as 4 sigma x standard deviation) and T of 85 (I first tried to use the median and mean but it seemed not to show anything so I brought down the number slightly. As a result, it seems on average summer peaks around the 24th of August and ends on August 30th (see graph below)



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).***

   It seems temperatures are going through a cycle of hot summers and moderate ones (as seen by the average and medians of each year), but it is also trending up with 2015 being the break up point (see graph below).


Average and Median per year
92.00
90.00
88.00
86.00
84.00
82.00
80.00
78.00
76.00
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Average
Median