**Clustering Pharmaceutical Firms**

An equities analyst is studying the pharmaceutical industry and would like your help in exploring and understanding the financial data collected by her firm. Her main objective is to understand the structure of the pharmaceutical industry using some basic financial measures.

Financial data gathered on 21 firms in the pharmaceutical industry are available in the file Pharmaceuticals.xls.

For each firm, the following variables are recorded.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Model Role** | **Data Type** | **Description** |
| Symbol | Ignore | Categoric | Company stock symbol |
| Name | Ignore | Categoric | Company name |
| Market\_Cap | Input | Numeric | Market capitalization (in billions of dollars) |
| Beta | Input | Numeric | Beta |
| PE\_Ratio | Input | Numeric | Price to earnings ratio |
| ROE | Input | Numeric | Return on equity |
| ROA | Input | Numeric | Return on investment |
| Asset\_Turnover | Input | Numeric | Asset turnover |
| Leverage | Input | Numeric | Leverage |
| Rev\_Growth | Input | Numeric | Estimated revenue growth |
| Net\_Profit | Input | Numeric | Net profit margin |
| Median\_  Recommendation | Ignore | Categoric | Median recommendations (across major brokerages) |
| Location | Ignore | Categoric | Location of company headquarters |
| Exchange | Ignore | Categoric | Stock exchange on which the firm is listed |

> getwd()

[1] "C:/Users/aaroh/Documents/BA with R"

> setwd("C:/Users/aaroh/Documents/BA with R")

> pharm.df <- read.csv("Pharmaceuticals(1).csv",header = TRUE)

> row.names(pharm.df) <- pharm.df[,1]

> # removing non-numerical values

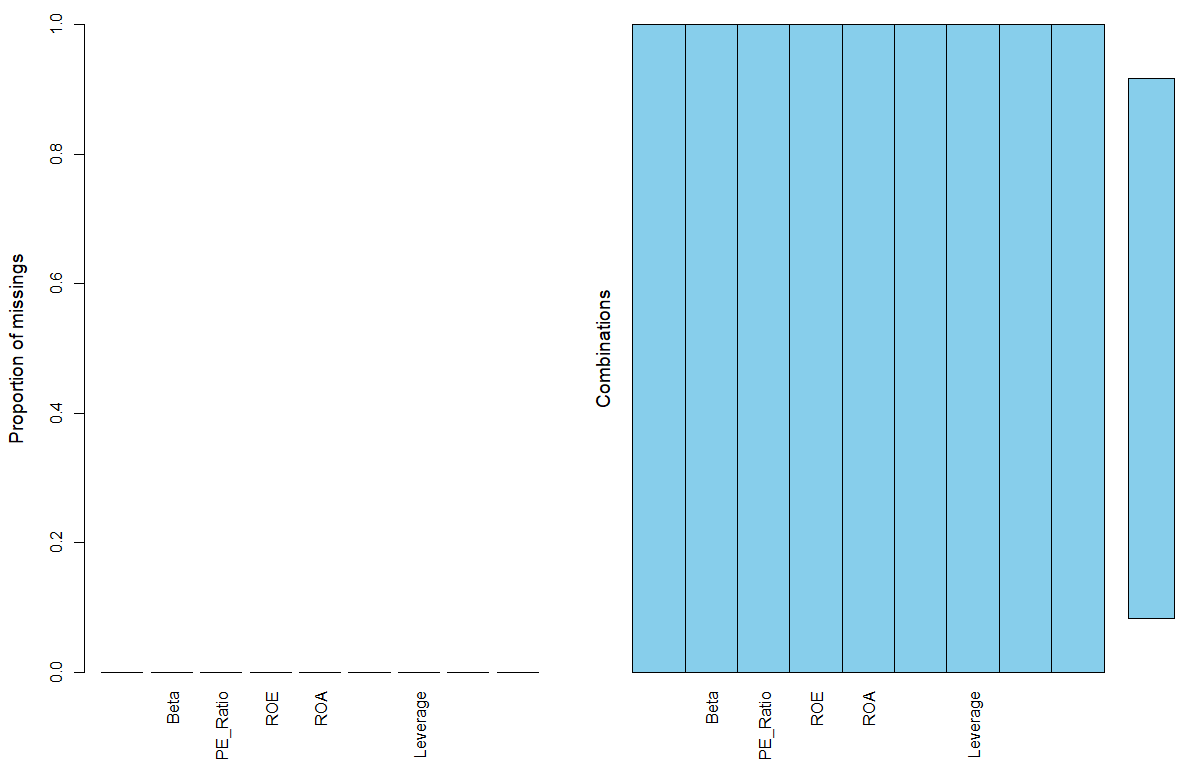
> pharm.df <- pharm.df[,-c(1,2,12,13,14)]

Now, we need to check for missing values using aggr() function from VIM package which tells us the number of missing values for each variable and for the combinations. Based on the plot below, we can say that there are no missing values.

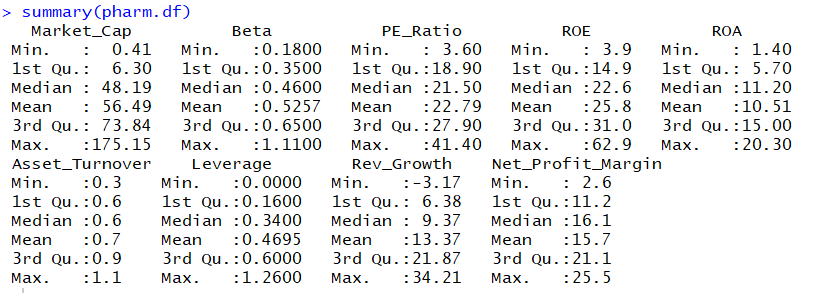
> library("VIM")

> aggr(pharm.df)

Hit <Return> to see next plot:



**OUTLIERS:**

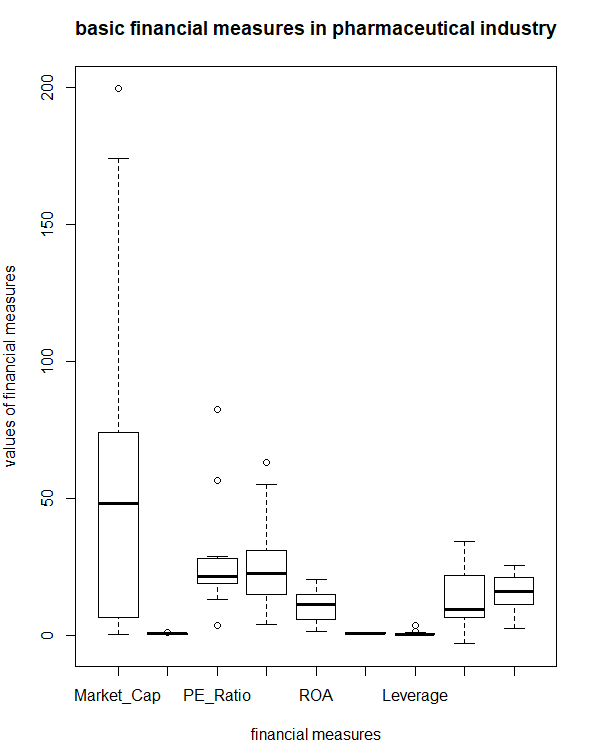


Based on the box plot, we can say that there are very littile outliers which we need to determine and impute. These outliers cannot be removed since the data collected is very less which will lead to poor model. Thus, we impute the values of outliers. We can see that outliers are present in 3 variables:

• Market\_Cap

• PE\_Ration

• Leverage



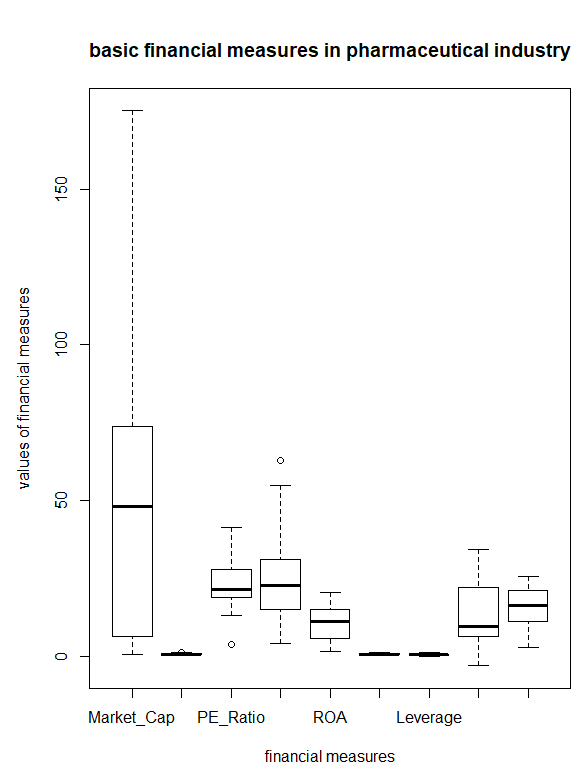
To remove outliers we have used the winsorizing techniques where we eliminate the upper and lower 1% of the data checking if it is greater than third quartile + 1.5\* Interquartile range or first quartile – 1.5\*Interquartile range.

> pharm.df$Market\_Cap[pharm.df$Market\_Cap > 73.84 + 1.5\*IQR(pharm.df$Market\_Cap)] <- 73.84 + 1.5\*IQR(pharm.df$Market\_Cap)

> pharm.df$Leverage[pharm.df$Leverage > 0.6+1.5\*IQR(pharm.df$Leverage)] <- 0.6+1.5\*IQR(pharm.df$Leverage)

> pharm.df$PE\_Ratio[pharm.df$PE\_Ratio > 27.90+ 1.5\*IQR(pharm.df$PE\_Ratio)] <- 27.90+ 1.5\*IQR(pharm.df$PE\_Ratio)

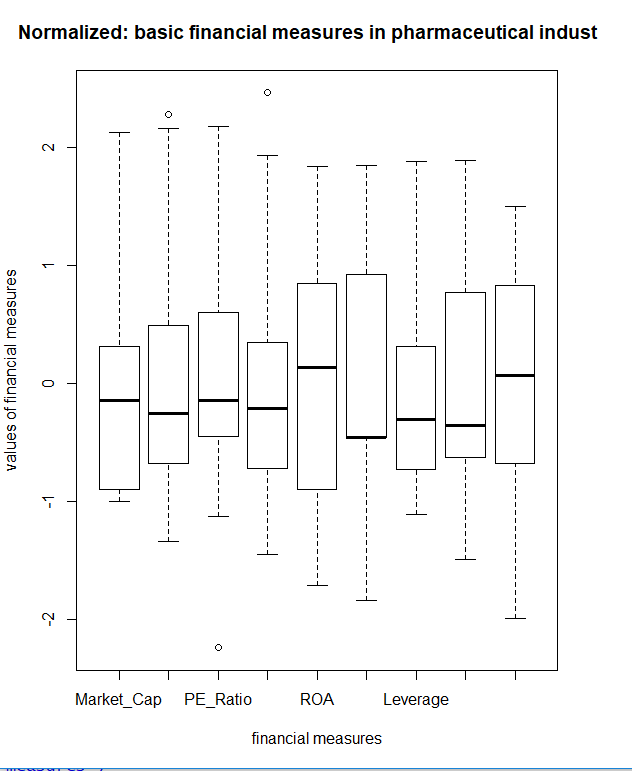
> boxplot(pharm.df, main="basic financial measures in pharmaceutical industry", xlab="financial measures", ylab="values of financial measures")



Now, we perform standardization of variables using scaling and create a boxplot of scaled imputed values.

> pharm.df.norm <- sapply(pharm.df.new,scale)

> boxplot(pharm.df.norm,main="Normalized: basic financial measures in pharmaceutical industry", xlab="financial measures", ylab="values of financial measures")



Use hierarchical cluster analysis to explore and analyze the given dataset (only the numeric variables as shown in the table above) and answer the following questions:

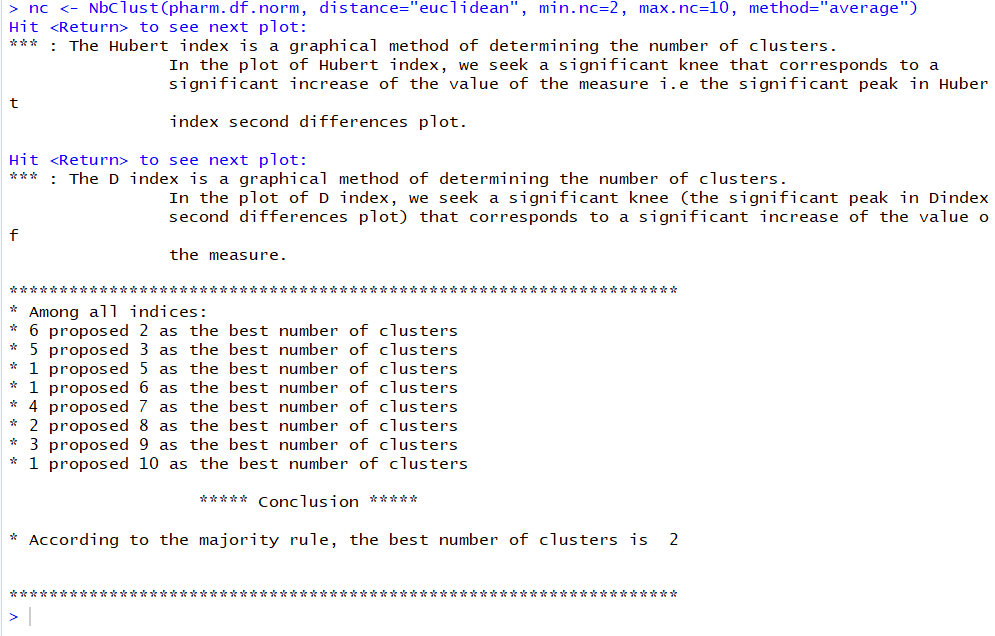
1. Based on the dendogram, what is the optimal number of clusters? How did you come to that determination?

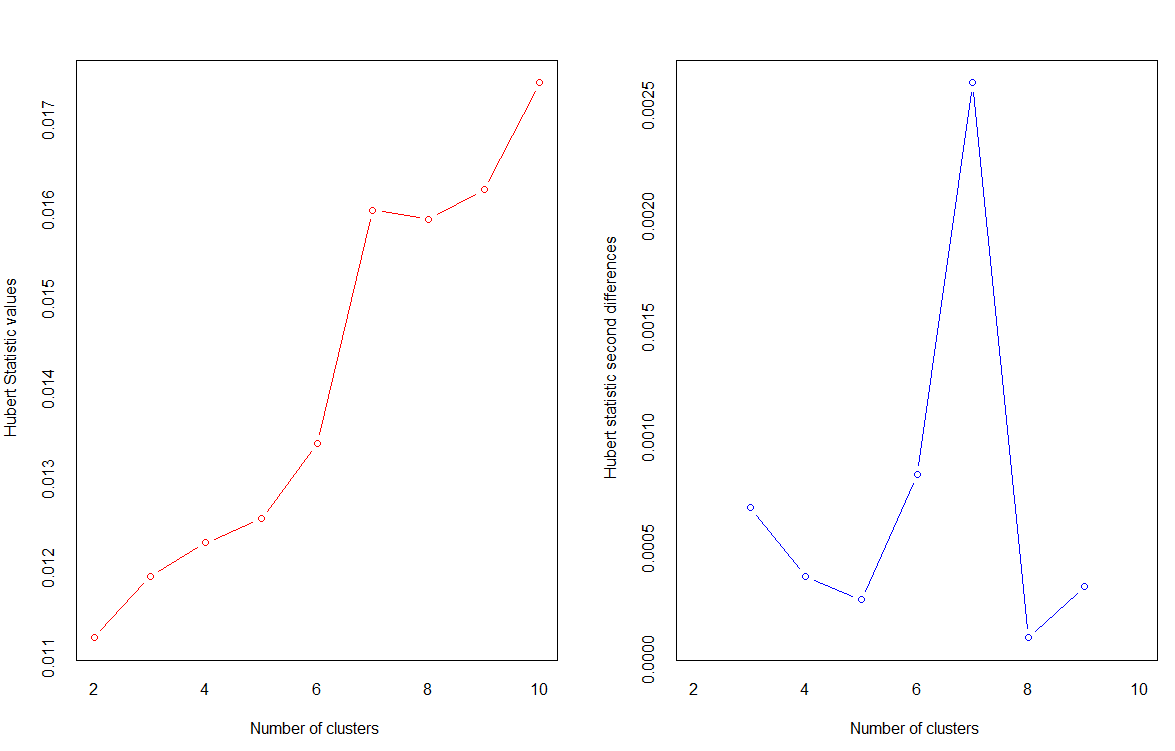
The optimal number of clusters is 2.

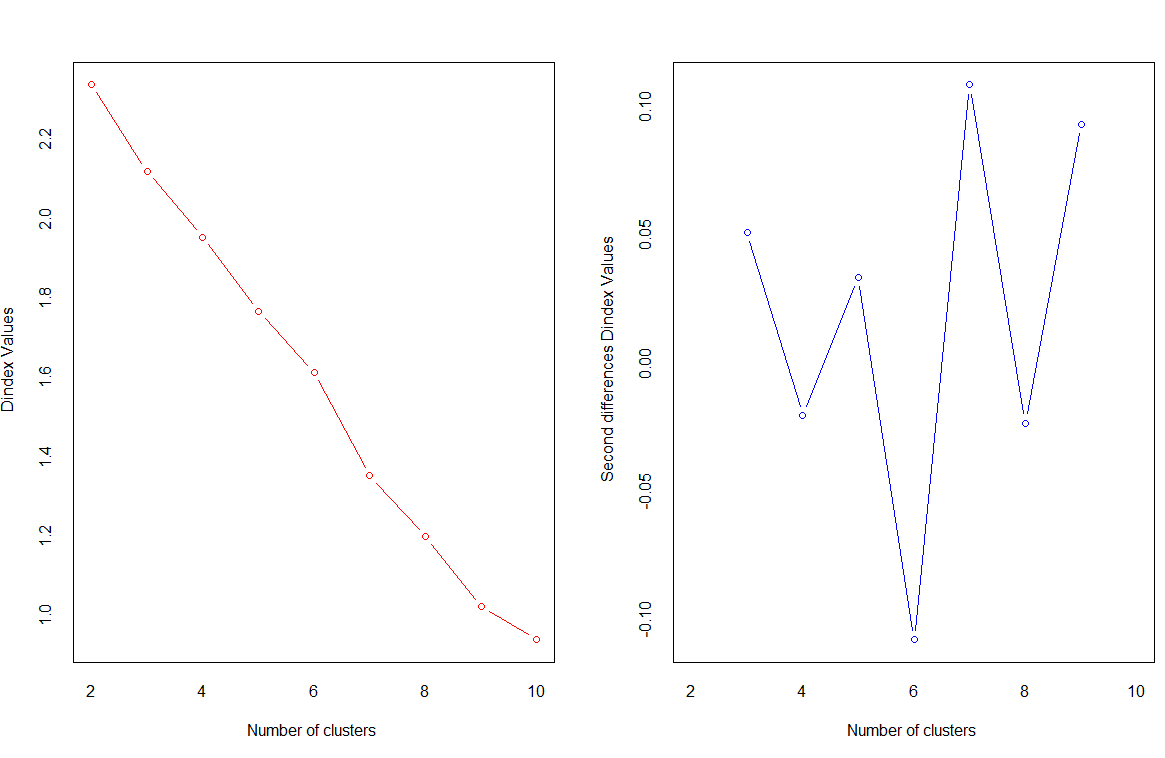
Hierarchical clustering has been performed in the following LOC using average linkage clustering method. This has been performed in following steps:

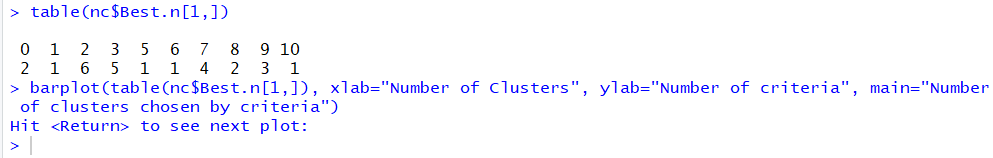
* Data has been standardized with mean 0 and variance as 1.
* Euclidean distance is calculated for each element which represents a cluster.
* Two clusters are combined having the smallest distance.
* B and C are repeated until all clusters merge into a single cluster.

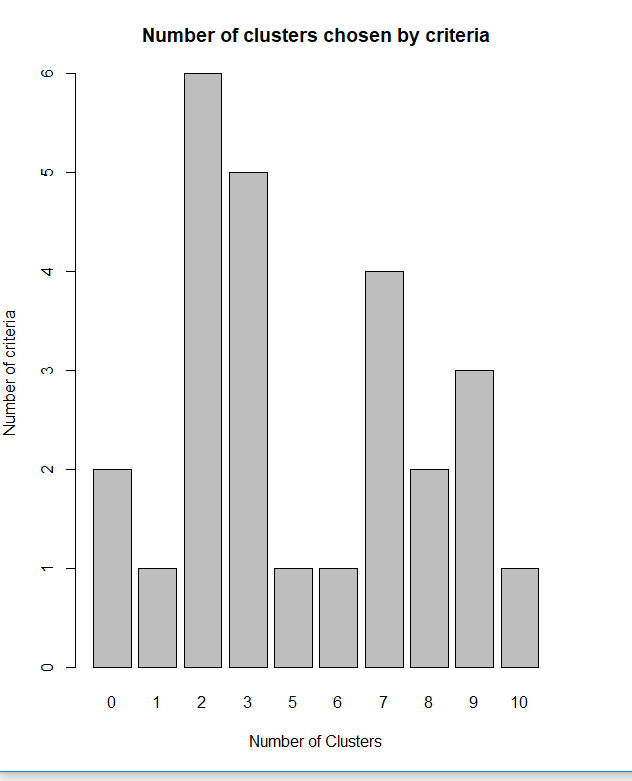
We determine the number of clusters using NbClust package.

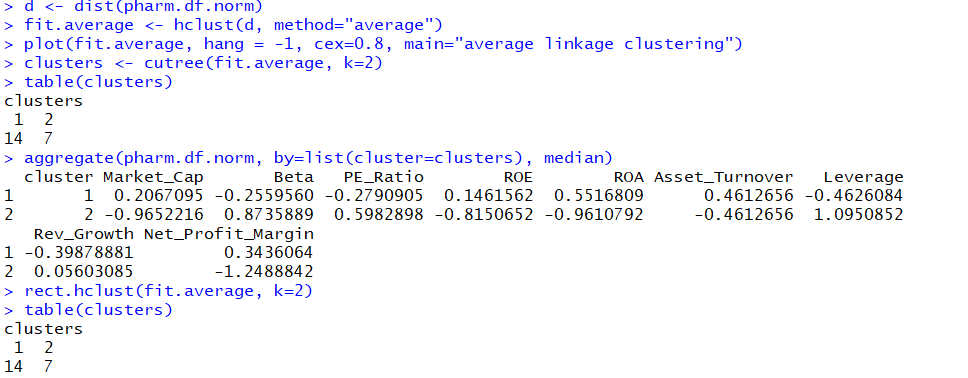


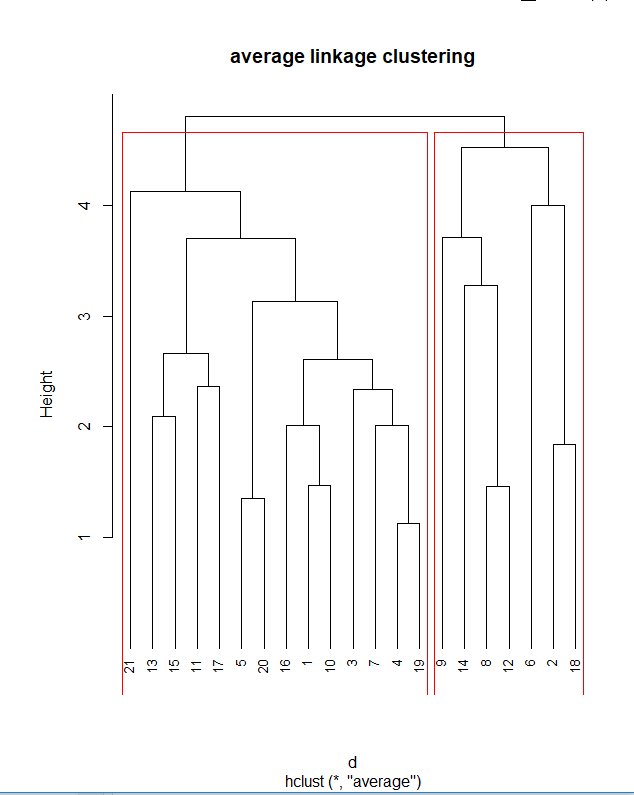










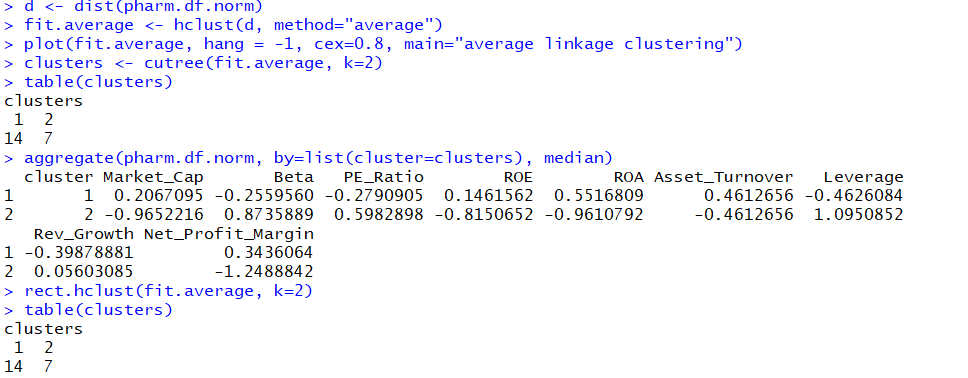


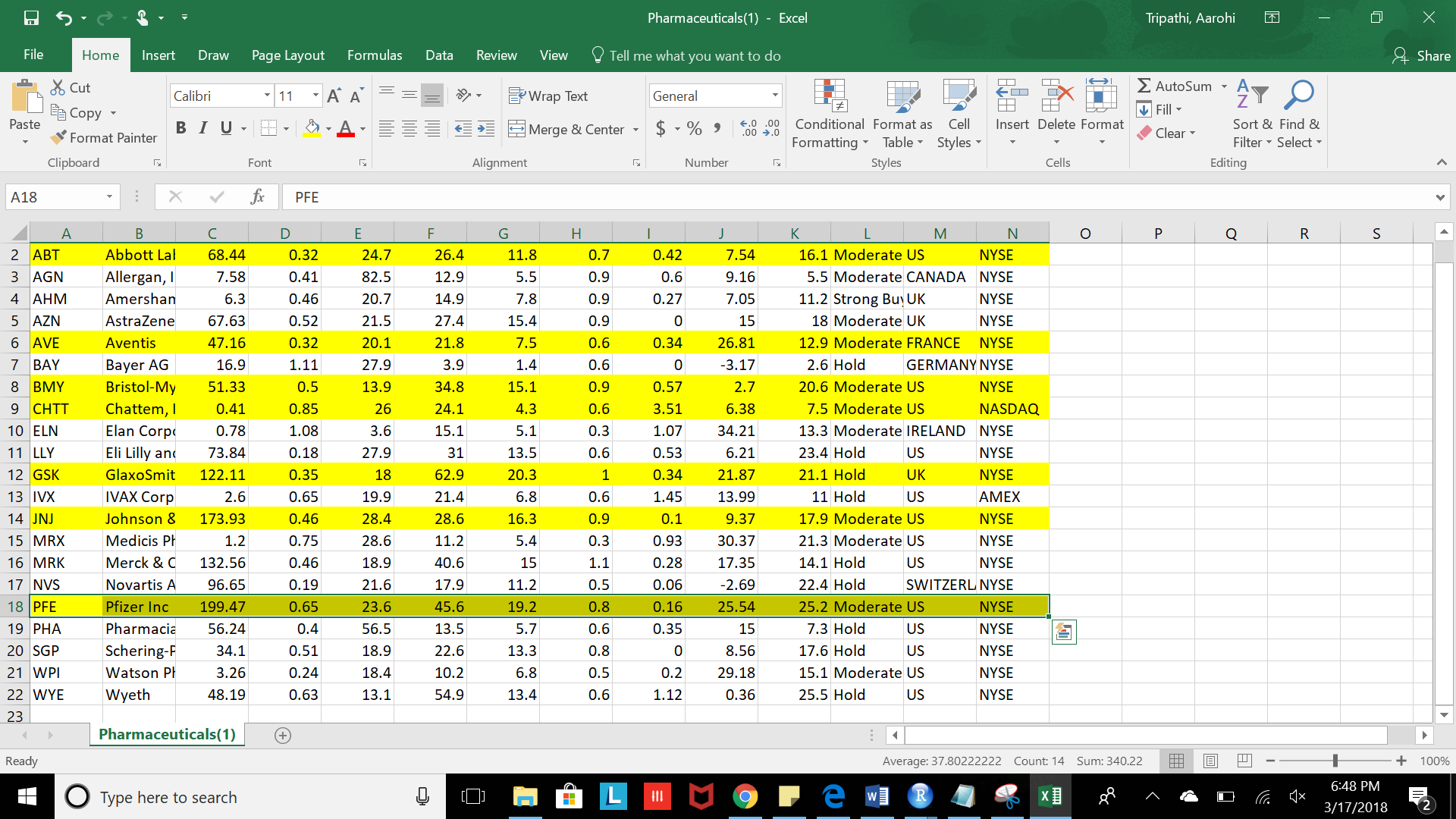
now, the dendogram created above is cut to create 2 clusters depicted by red region above. This creates a hierarchical view of the data. The Cutree() function cuts the tree into three clusters and assigns cases.

We observe that first cluster has 14 observations, second cluster has 7 observations. The aggregate() function provides the median profile for each cases.

rect.hclust() function is used to superimpose the two cluster solution.

1. Interpret the clusters with respect to the variables that were used in their formation. Provide an appropriate name for each cluster using any or all of the variables in the dataset.

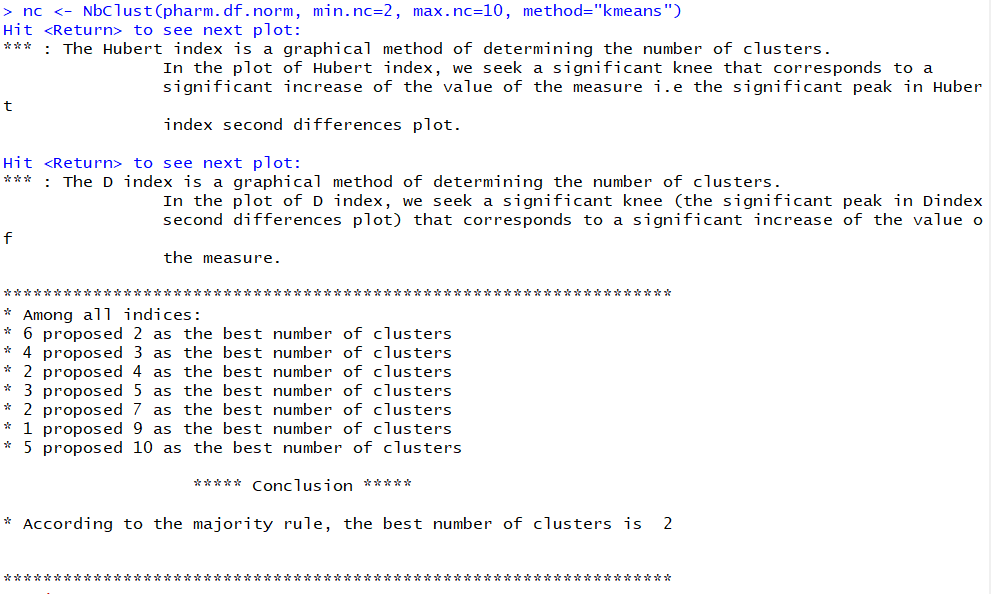


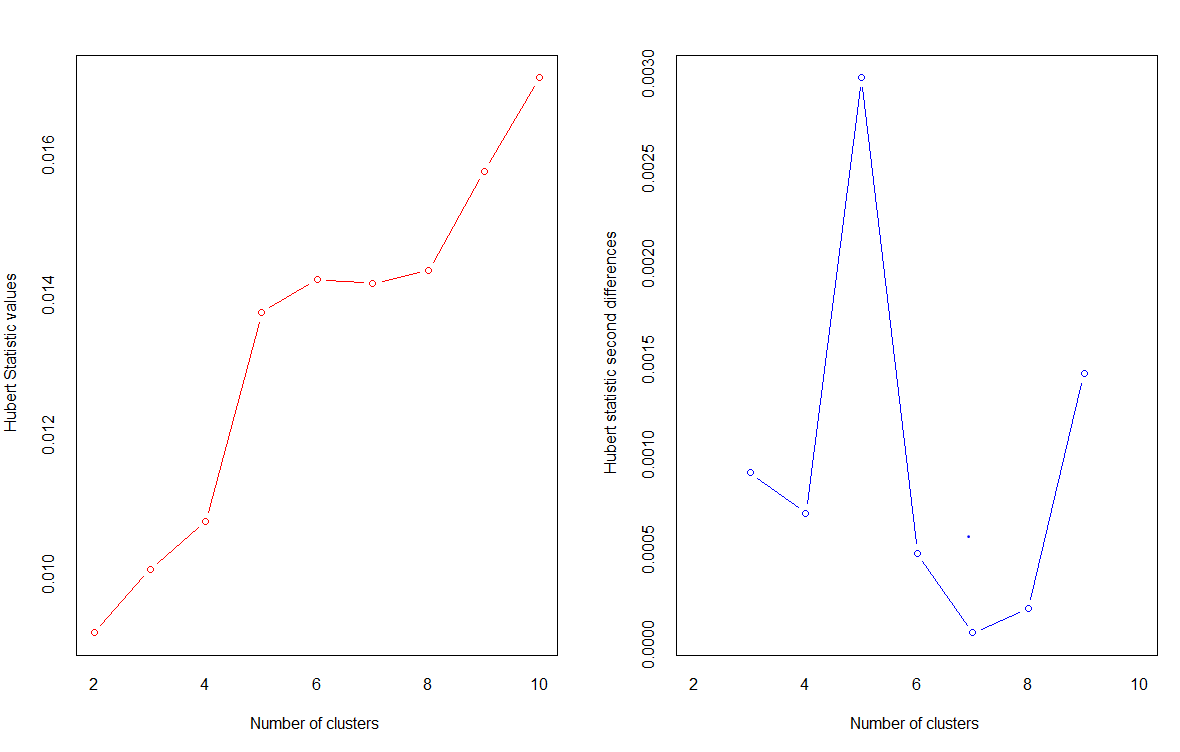


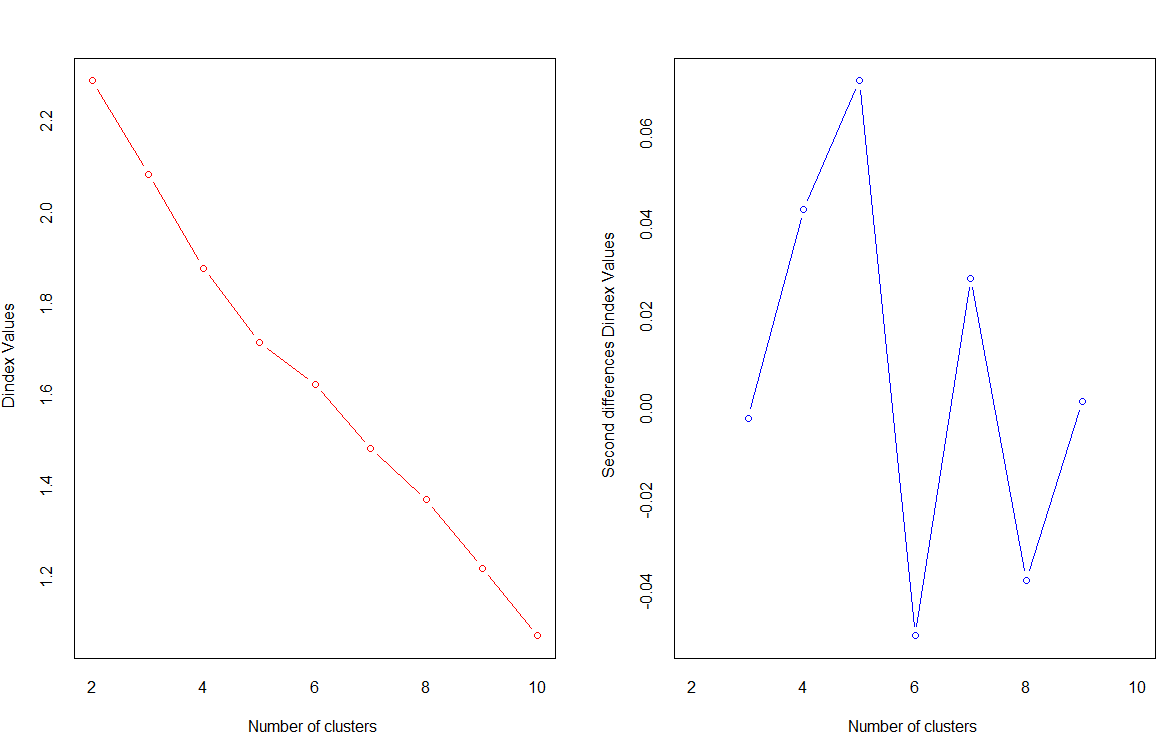
The highlighted are part of second cluster and rest are part of first. We see that cluster 1 centroid is closer to Market\_Cap, ROE, ROA, Asset Turnover and Net Profit Margin on the basis of average Euclidian distance.

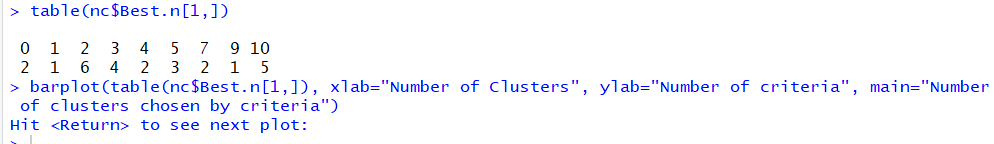
We see that the second cluster has observations with high Market Cap and ROA so we can name the cluster to be High Rated and the first cluster can be named as Low rated since it has low Market Cap and ROA.

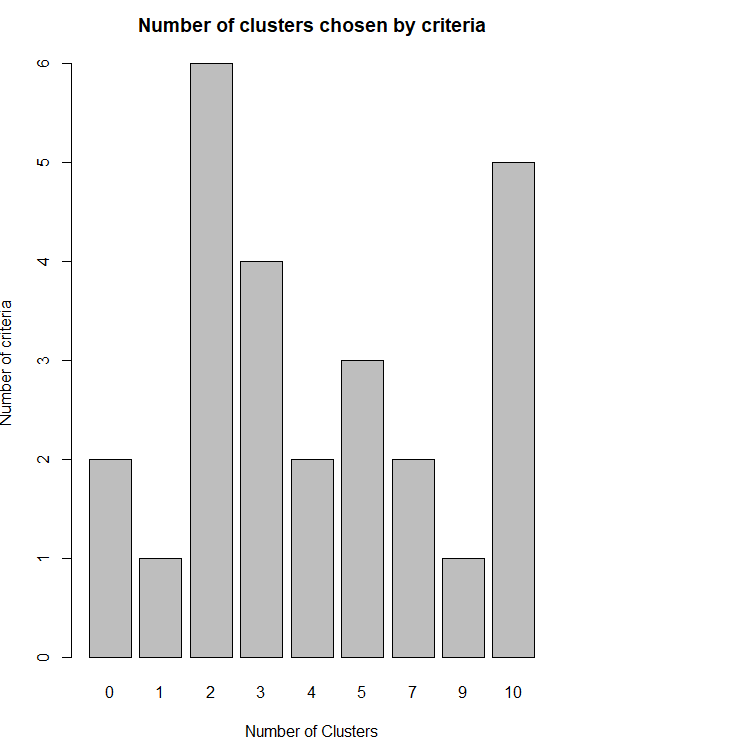
There are total 2 clusters formed which seem to be reasonable considering the number of observations we have is 21 and the observations when clustered show characteristics of high Market Cap and ROA in one cluster and low Market Cap and ROA in one cluster.

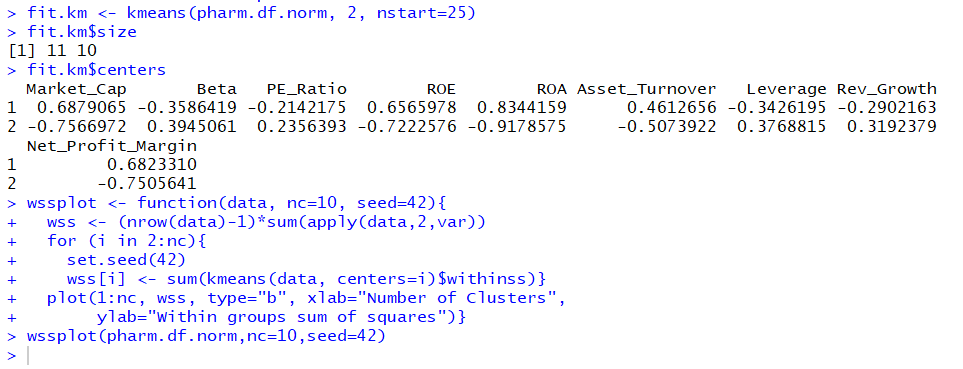


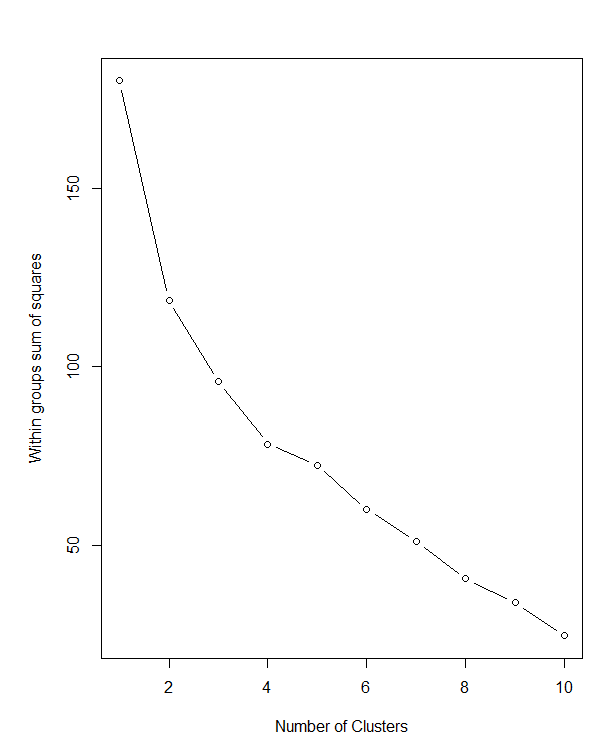












The above method proposed 2 as the best number of clusters when seed was set to 42. We see that the clusters have 11 and 10 observations respectively as compared to hierarchical clustering which shows 14 and 7 observations in the 2 clusters. This clustering looks good since we see the observations are equally divided.