ML_assignment_KMeans

Abhishek_Sau

19/03/2021

R Markdown

```
setwd("D:\\Study\\Assignments\\MachLearn\\MachLearnAssignment4_KmeansClustering")
Pdata <- read.csv("Pharmaceuticals.csv")
head(Pdata)
```

```
##   Symbol                Name Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover
## 1    ABT Abbott Laboratories      68.44 0.32     24.7 26.4 11.8            0.7
## 2    AGN      Allergan, Inc.       7.58 0.41     82.5 12.9  5.5            0.9
## 3    AHM        Amersham plc       6.30 0.46     20.7 14.9  7.8            0.9
## 4    AZN     AstraZeneca PLC      67.63 0.52     21.5 27.4 15.4            0.9
## 5    AVE             Aventis      47.16 0.32     20.1 21.8  7.5            0.6
## 6    BAY            Bayer AG      16.90 1.11     27.9  3.9  1.4            0.6
##   Leverage Rev_Growth Net_Profit_Margin Median_Recommendation Location Exchange
## 1     0.42       7.54              16.1          Moderate Buy       US     NYSE
## 2     0.60       9.16               5.5          Moderate Buy   CANADA     NYSE
## 3     0.27       7.05              11.2            Strong Buy       UK     NYSE
## 4     0.00      15.00              18.0         Moderate Sell       UK     NYSE
## 5     0.34      26.81              12.9          Moderate Buy   FRANCE     NYSE
## 6     0.00      -3.17               2.6                  Hold  GERMANY     NYSE
```

```
dim(Pdata) #Dataframe has 14 variables and 21 records
```

```
## [1] 21 14
```

```
str(Pdata) #Shows the No. of variable and their datatypes
```

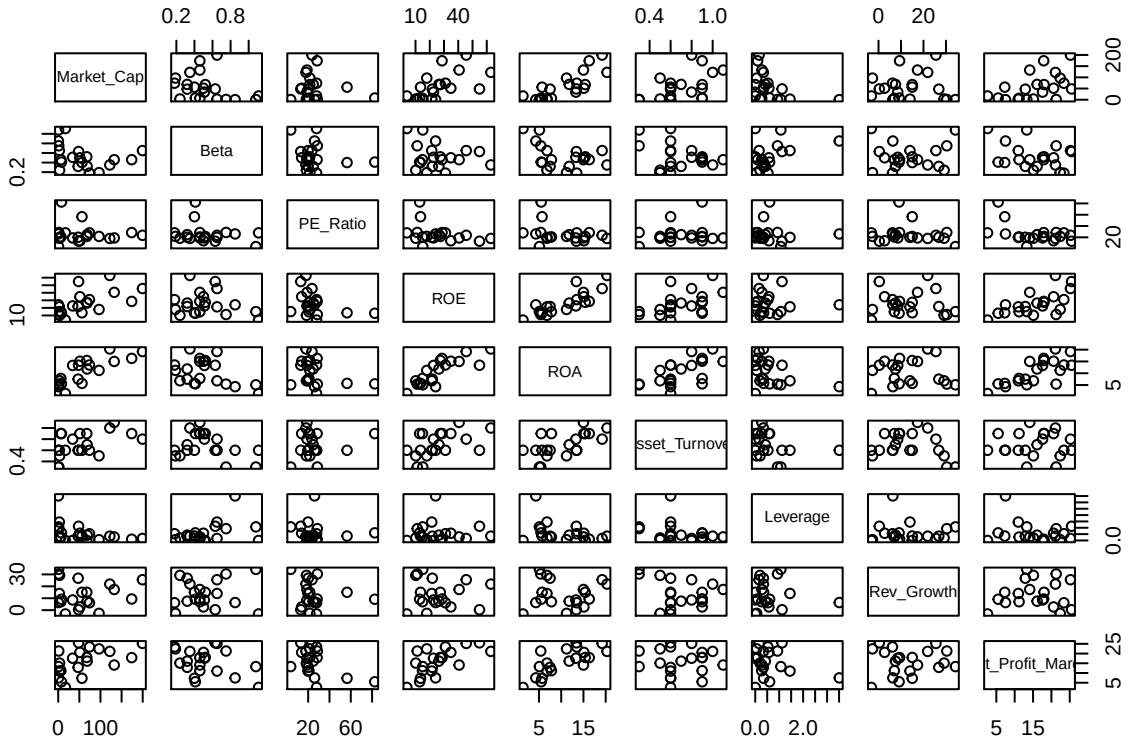
```
## 'data.frame':    21 obs. of  14 variables:
##  $ Symbol               : chr  "ABT" "AGN" "AHM" "AZN" ...
##  $ Name                 : chr  "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZeneca PL
##  $ Market_Cap           : num  68.44 7.58 6.3 67.63 47.16 ...
##  $ Beta                 : num  0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...
##  $ PE_Ratio             : num  24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...
##  $ ROE                  : num  26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...
##  $ ROA                  : num  11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...
##  $ Asset_Turnover       : num  0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
##  $ Leverage             : num  0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...
##  $ Rev_Growth           : num  7.54 9.16 7.05 15 26.81 ...
##  $ Net_Profit_Margin    : num  16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...
##  $ Median_Recommendation: chr  "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sell" ...
##  $ Location             : chr  "US" "CANADA" "UK" "UK" ...
##  $ Exchange             : chr  "NYSE" "NYSE" "NYSE" "NYSE" ...
```

```
U<-unique(Pdata) # This shows there are no duplicate rows
dim(U)
```

```
## [1] 21 14
```

```
pairs(Pdata[3:11]) #All possible combinations of Discrete variables plotted

library(DataExplorer)
```

```
## Warning: package 'DataExplorer' was built under R version 4.0.4
```



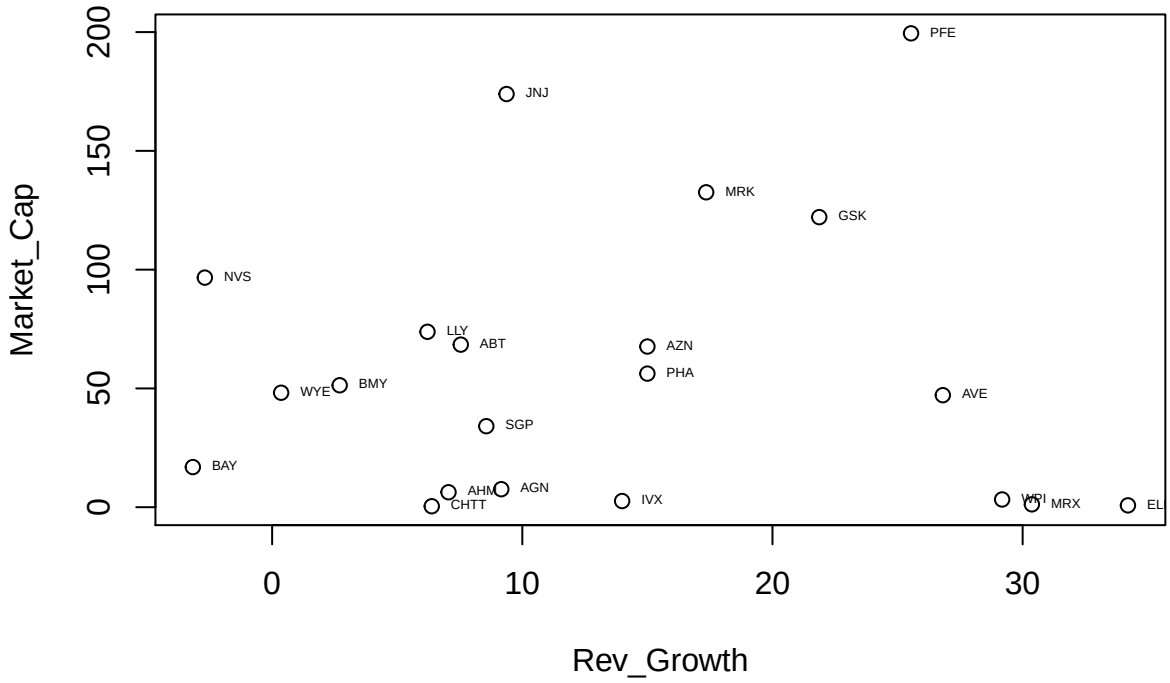

```
introduce(Pdata) #Shows the number of Discrete/continuous variables and missing value, if any.
```

```
##   rows columns discrete_columns continuous_columns all_missing_columns
## 1   21      14                5                  9                   0
##   total_missing_values complete_rows total_observations memory_usage
## 1                    0            21                294         9680
```

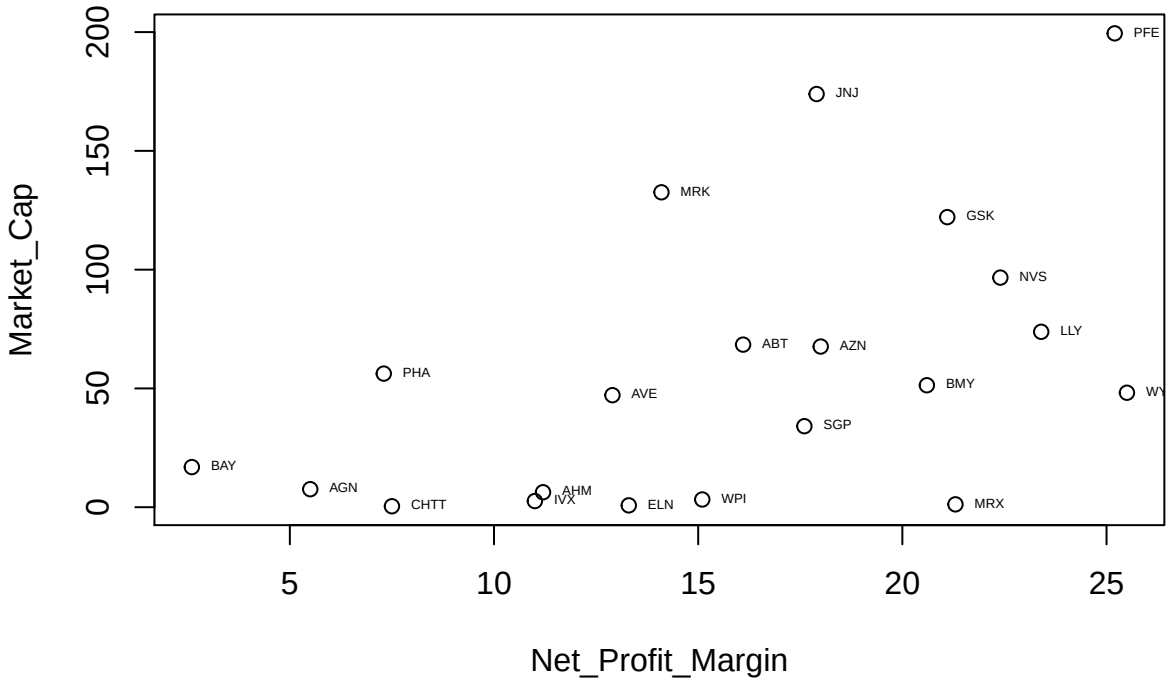
```
#There are 5 discrete Variables, 9 Continuous variables and no missing variables.
```

#Data exploration

```
set.seed(123)
plot(Market_Cap~Rev_Growth, Pdata)
with(Pdata, text(Market_Cap~Rev_Growth, labels=Symbol, pos=4, cex=0.4)) #A rough estimate of how the di
```




```
plot(Market_Cap~Net_Profit_Margin, Pdata)
with(Pdata, text(Market_Cap~Net_Profit_Margin, labels=Symbol, pos=4, cex=0.4))
```



K-means clustering model

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.0.4
```

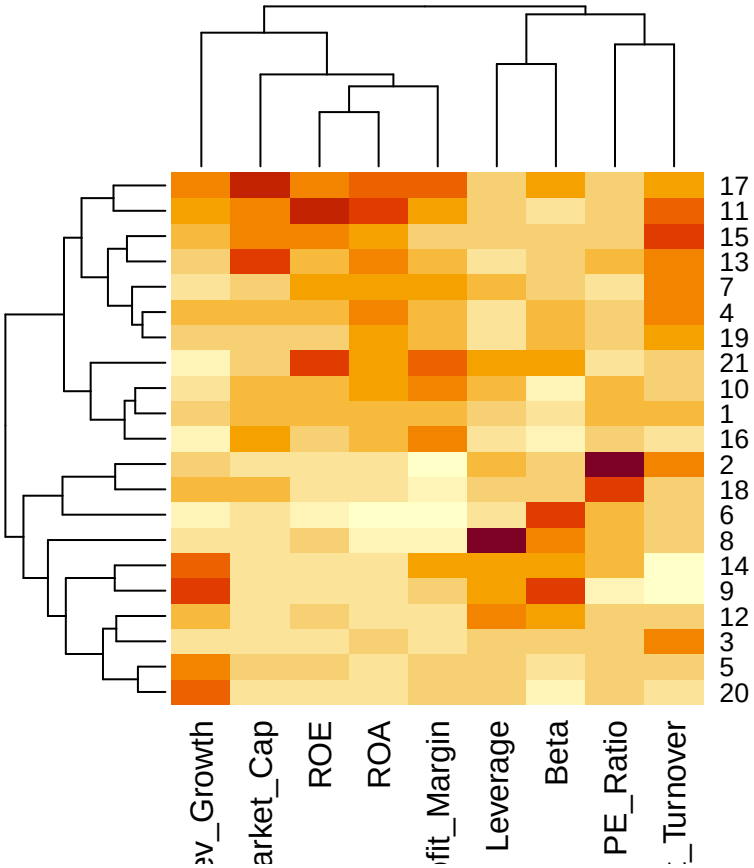
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(ISLR)

Pdata1 <- scale(Pdata[, c(3:11)], center = TRUE, scale = TRUE)

dist_pdata <- get_dist(Pdata1)
print(dist_pdata, digits = 3)
```

```
##      1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 2 4.42
```

```
## 3  2.02 3.95
## 4  1.67 4.91 2.36
## 5  2.11 4.64 2.49 2.63
## 6  4.69 4.85 3.64 5.07 4.76
## 7  1.81 5.42 2.60 1.57 3.40 5.27
## 8  5.02 5.61 4.76 5.72 5.10 4.97 5.29
## 9  4.90 6.70 4.70 4.97 3.75 4.61 5.38 4.68
## 10 1.42 5.14 3.24 2.41 2.91 5.80 2.19 5.66 5.55
## 11 3.69 6.75 4.90 2.96 4.48 7.55 3.10 7.08 6.73 3.63
## 12 2.62 4.47 2.32 3.28 2.39 3.66 3.28 2.95 3.12 3.54 5.28
## 13 2.33 5.32 3.59 1.96 3.64 5.72 2.51 6.31 6.07 2.72 2.99 4.35
## 14 3.92 5.48 4.12 4.27 2.93 4.85 4.73 4.79 2.39 4.19 6.19 2.83 5.31
## 15 2.68 5.44 3.36 1.86 3.47 5.92 2.43 6.10 5.92 3.38 2.22 4.16 1.81 5.53
## 16 1.92 5.47 3.33 3.06 3.33 5.33 2.87 6.06 5.73 1.58 4.78 3.90 3.08 4.48 4.11
## 17 3.89 6.91 5.27 3.11 4.50 7.16 3.67 7.18 6.12 3.78 2.45 5.36 2.45 5.52 2.83
## 18 2.91 2.37 2.93 3.72 2.72 3.96 4.41 5.00 5.01 3.75 5.77 3.07 4.11 3.83 4.45
## 19 1.31 4.73 1.70 1.08 2.46 4.43 1.48 5.35 4.67 2.21 3.78 2.76 2.60 3.91 2.71
## 20 2.88 5.01 2.94 3.41 1.30 5.06 4.12 5.54 3.76 3.41 5.44 2.86 4.59 2.65 4.57
## 21 3.04 6.45 4.19 3.32 4.25 5.95 2.27 5.13 5.31 2.75 3.67 3.72 3.86 4.71 3.94
##      16   17   18   19   20
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17 4.54
## 18 3.88 5.59
## 19 2.54 3.96 3.45
## 20 3.63 5.40 3.17 3.03
## 21 3.53 4.03 5.29 3.15 4.92
```

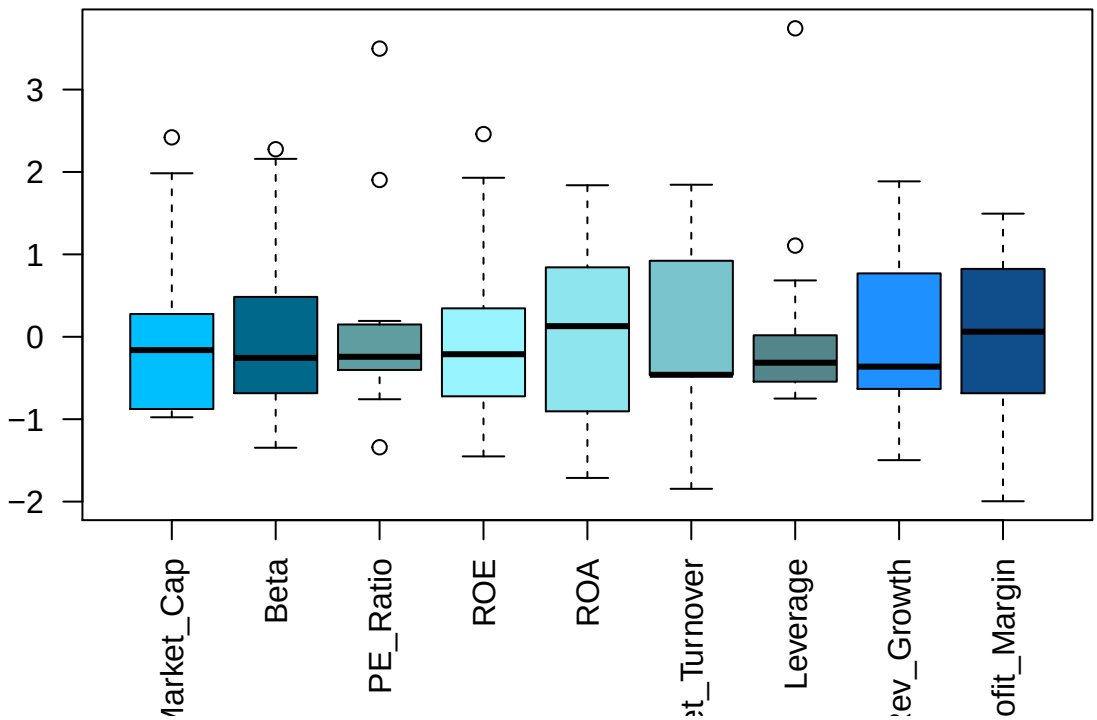
```
heatmap(Pdata1, scale = "column")
```




```
#Company 8 and 17 are the furthest from each other as the distance between them is the largest i.e 7.18
Pdata[c(8,17),] #From this we can see that there is a huge difference in terms of Market capital, Rev_G
```

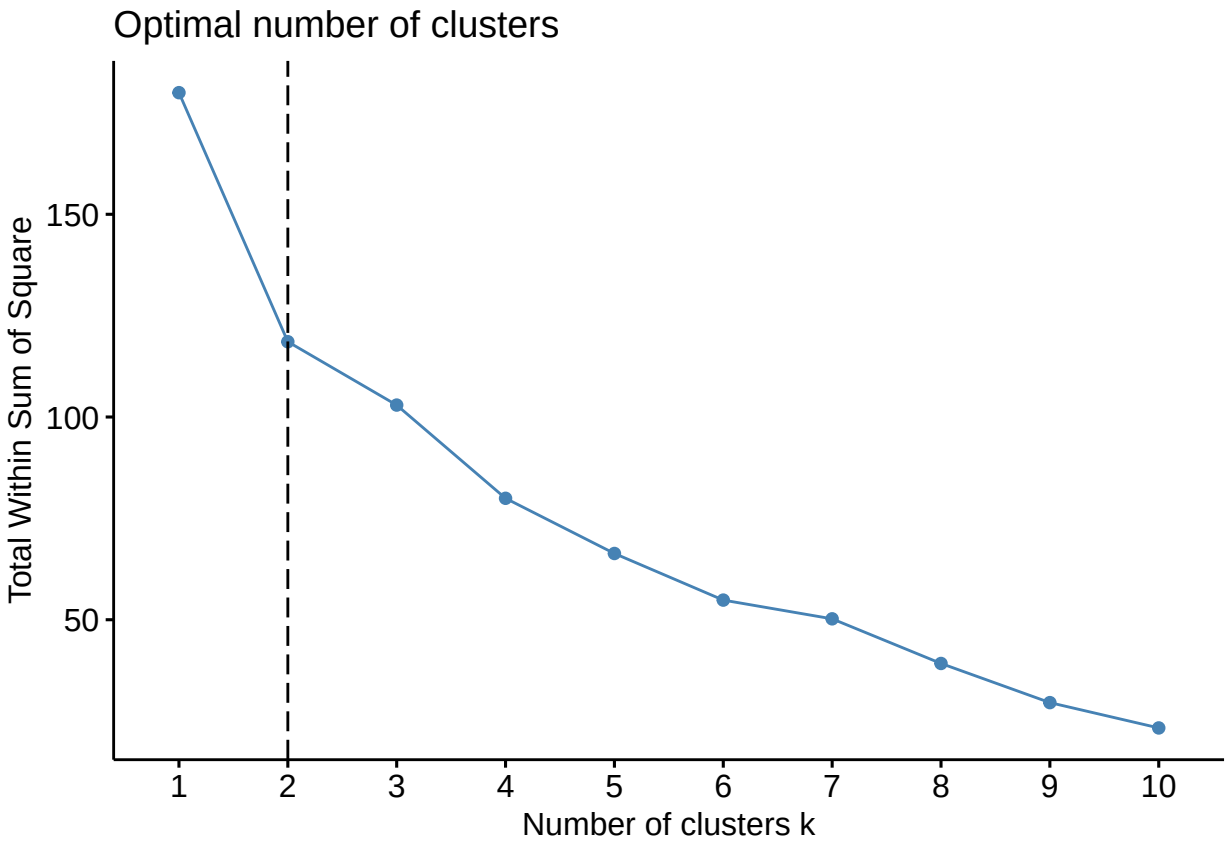
```
##    Symbol         Name Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover
## 8    CHTT Chattem, Inc       0.41 0.85     26.0 24.1  4.3            0.6
## 17    PFE   Pfizer Inc     199.47 0.65     23.6 45.6 19.2            0.8
##    Leverage Rev_Growth Net_Profit_Margin Median_Recommendation Location
## 8      3.51       6.38                7.5          Moderate Buy       US
## 17     0.16      25.54               25.2          Moderate Buy       US
##    Exchange
## 8    NASDAQ
## 17     NYSE
```

```
boxplot(Pdata1, col=c("deepskyblue","deepskyblue4","cadetblue","cadetblue1","cadetblue2","cadetblue3","
```

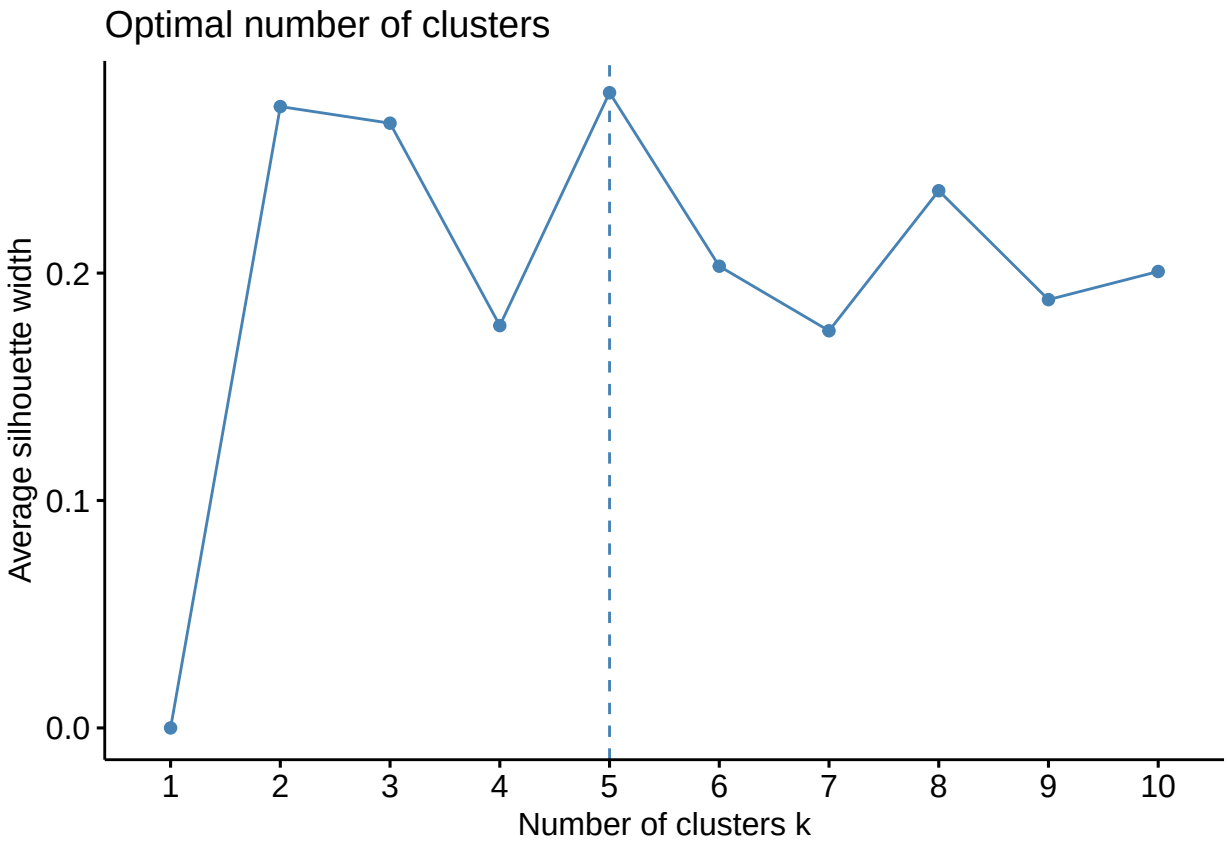


Finding optimal number for K as per Elbow method

```
fviz_nbclust(Pdata1, kmeans, method = "wss") +
 geom_vline(xintercept = 2, linetype = 5)       #Elbow method
```




```
fviz_nbclust(Pdata1, kmeans, method = "silhouette") #Silhouette Method
```

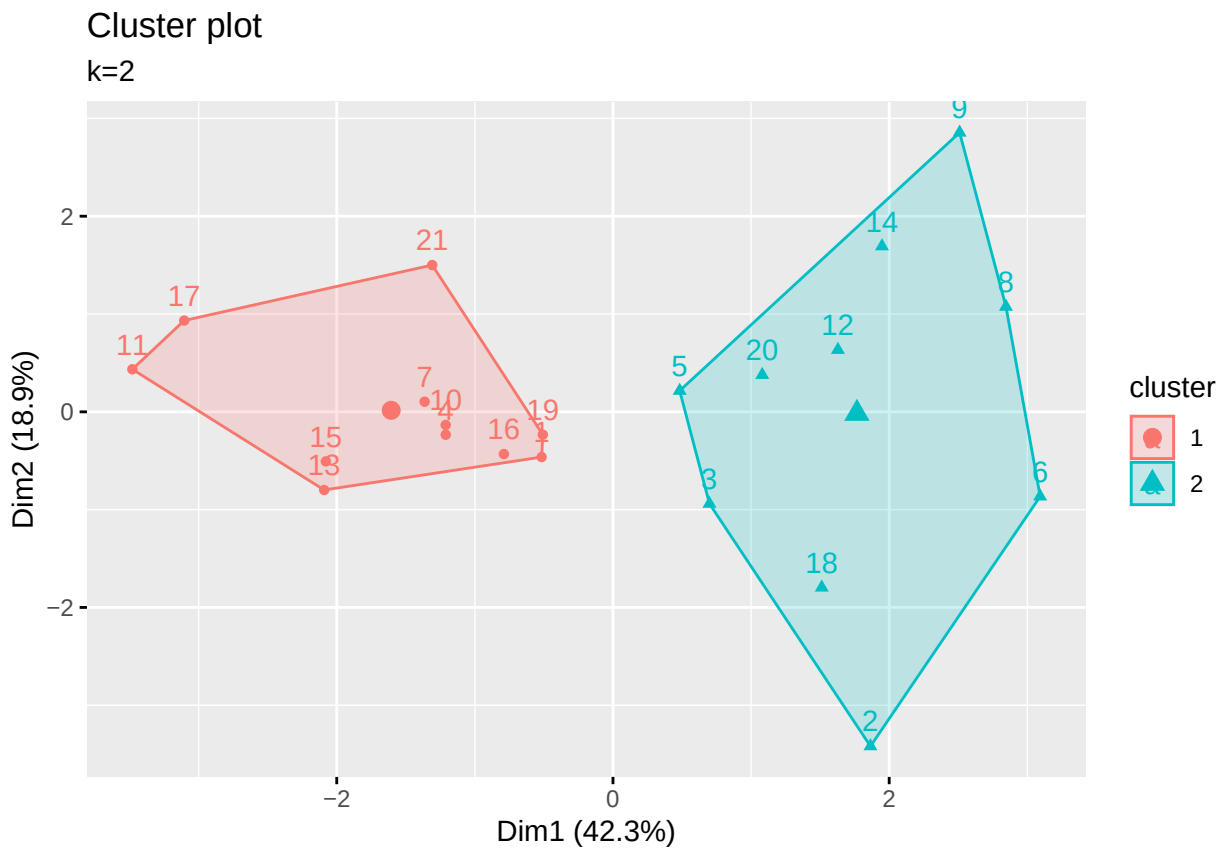


Silhouette method of finding optimal number of K is highly influenced by outliers. Since our data has quite a few number of outliers.

On removing the outliers, the optimal value of K using the silhouette method also came out to be 2.

Hence, We can consider the optimal number of K using the Elbow method.

```
#k=2
k_mean_model2 <- kmeans(Pdata1, centers = 2, nstart = 25)
fviz_cluster(k_mean_model2, data = Pdata1)+ labs(subtitle = "k=2")
```




```
k_mean_model2
```

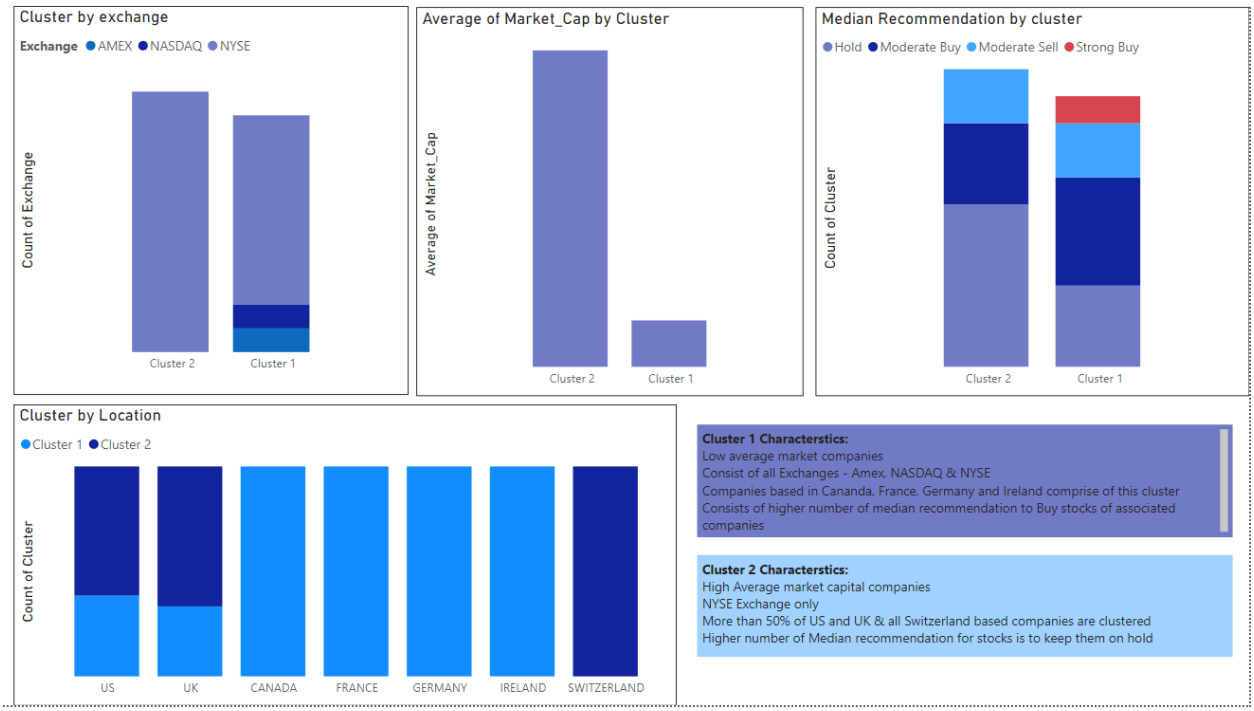
```
## K-means clustering with 2 clusters of sizes 11, 10
##
## Cluster means:
##   Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656
## 2 -0.7407208  0.3945061  0.3039863 -0.7222576 -0.9178575     -0.5073922
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.3331068 -0.2902163         0.6823310
## 2  0.3664175  0.3192379        -0.7505641
##
## Clustering vector:
##  [1] 1 2 2 1 2 2 1 2 2 1 1 2 1 2 1 1 1 2 1 2 1
##
## Within cluster sum of squares by cluster:
## [1] 43.30886 75.26049
##  (between_SS / total_SS =  34.1 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
Pdata2 <- Pdata[-c(3:11)] %>% mutate(data.frame(k_mean_model2$cluster))

library("writexl")
```

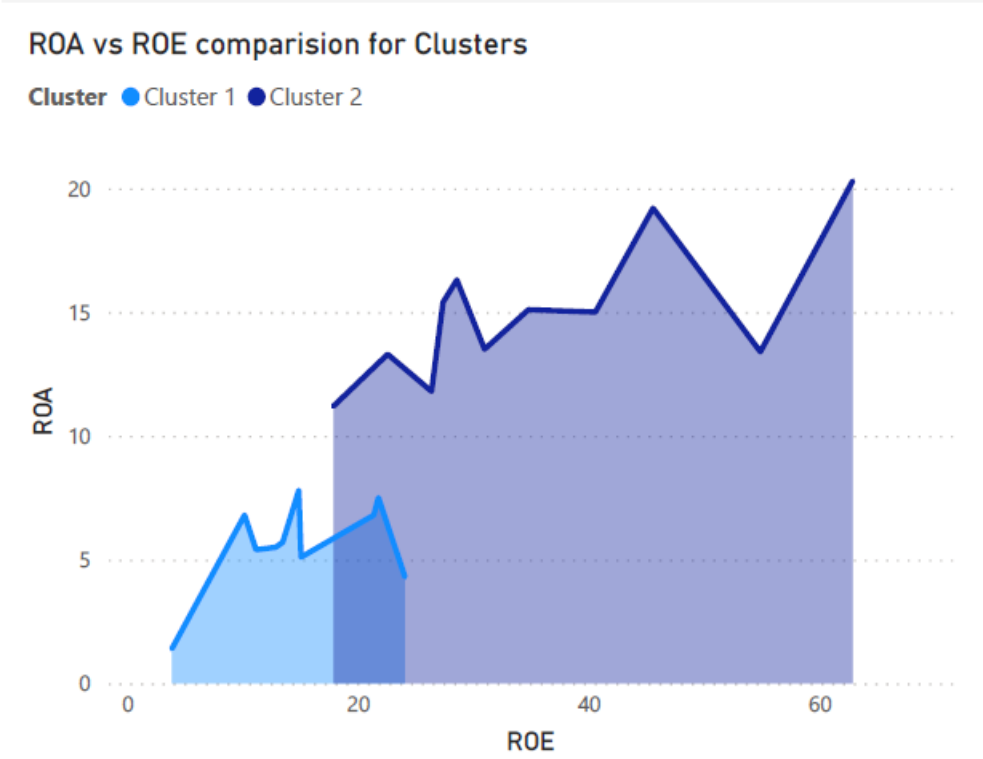
```
## Warning: package 'writexl' was built under R version 4.0.4
```

```
write_xlsx(Pdata2, "D:\\Study\\Assignments\\MachLearn\\MachLearnAssignment4_KmeansClustering\\Pdata2.xls
```

```
#Include Report image
knitr::include_graphics("Report.png")
```




```
knitr::include_graphics("ROEvsROA.png")
```



Q. Provide an appropriate name for each cluster using any or all of the variables in the dataset.

A- Low Market Cap companies are in cluster 1 and High Market cap companies are in Cluster 2 (As per the report above)

Q. Interpret the clusters with respect to the numerical variables used in forming the clusters.

A. For the same values of Return on equity, companies in cluster 2 have a higer Return on Equity than companies in cluster 1.