ML Assignment 4

Abhinay Chiranjeeth Marneni

2022-11-01

loading library functions packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(flexclust)
```

```
## Warning: package 'flexclust' was built under R version 4.2.2
```

```
## Loading required package: grid
## Loading required package: lattice
## Loading required package: modeltools
## Loading required package: stats4
```

```
library(cluster)
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.2.2
```

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

```
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.2.2
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.2.2
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.2.2
```

```
library(dplyr)
```

Importing the Pharmaceuticals.csv file

```
data<- read.csv("C:/Users/abhin/OneDrive/Documents/Assigments Buss 1sem/ML/Pharmaceuticals.csv") # impo
view(data) # using view function to display the whole table
head(data) # Using head function to view the 6rows of dataset
```

```
##   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
```

A.Use only the numerical variables (1 to 9) to cluster the 21 firms. Justify the various choices made in

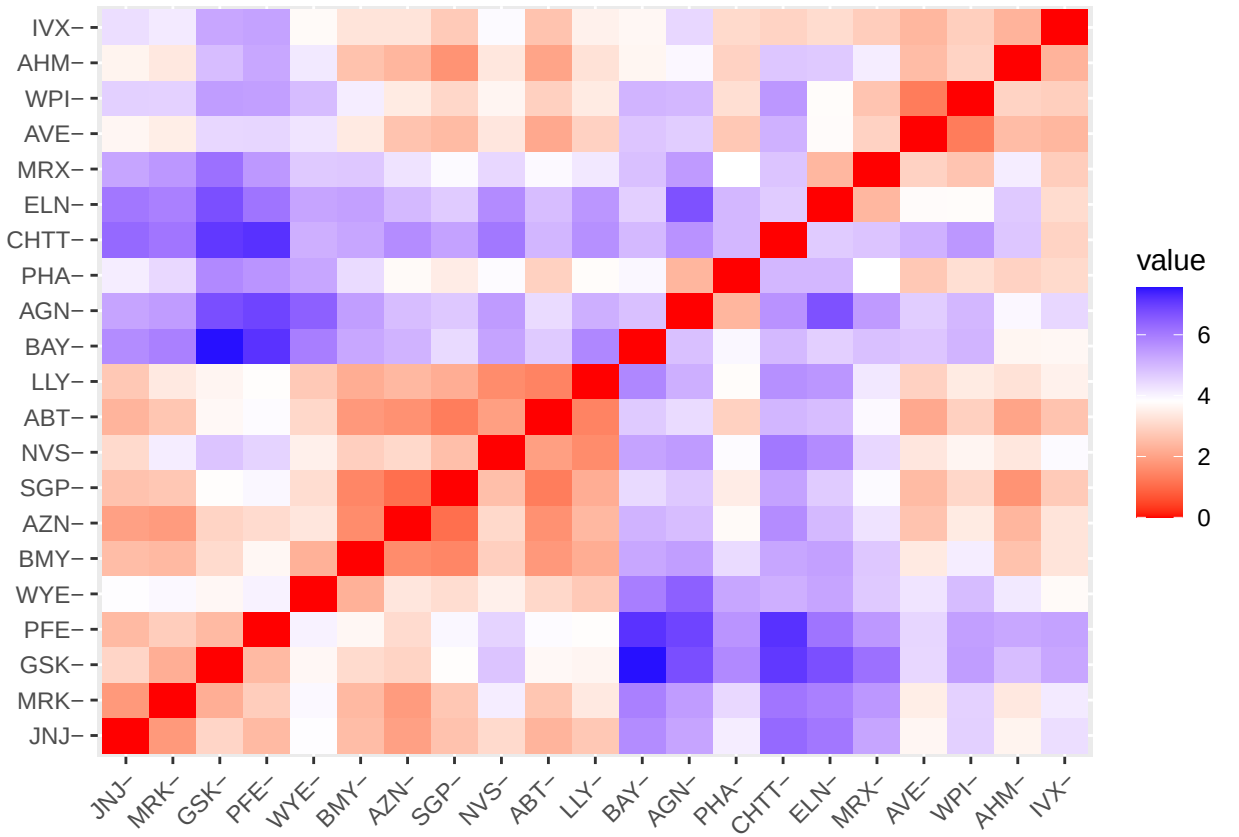
conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on.

```
set.seed(450)
da <- data [,3:11]
summary(da)
```

```
##    Market_Cap          Beta           PE_Ratio          ROE
##  Min.   :  0.41   Min.   :0.1800   Min.   : 3.60   Min.   : 3.9
##  1st Qu.:  6.30   1st Qu.:0.3500   1st Qu.:18.90   1st Qu.:14.9
```

```
##  Median : 48.19   Median :0.4600   Median :21.50   Median :22.6
##  Mean   : 57.65   Mean   :0.5257   Mean   :25.46   Mean   :25.8
##  3rd Qu.: 73.84   3rd Qu.:0.6500   3rd Qu.:27.90   3rd Qu.:31.0
##  Max.   :199.47   Max.   :1.1100   Max.   :82.50   Max.   :62.9
##       ROA         Asset_Turnover    Leverage        Rev_Growth
##  Min.   : 1.40   Min.   :0.3    Min.   :0.0000   Min.   :-3.17
##  1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600   1st Qu.: 6.38
##  Median :11.20   Median :0.6    Median :0.3400   Median : 9.37
##  Mean   :10.51   Mean   :0.7    Mean   :0.5857   Mean   :13.37
##  3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000   3rd Qu.:21.87
##  Max.   :20.30   Max.   :1.1    Max.   :3.5100   Max.   :34.21
##  Net_Profit_Margin
##  Min.   : 2.6
##  1st Qu.:11.2
##  Median :16.1
##  Mean   :15.7
##  3rd Qu.:21.1
##  Max.   :25.5
```

```
# To scale the data from the variables are measured using various weights throughout the rows
set.seed(450)
data1 <- scale(da)
row.names(data1) <- data[,1]
distance <- get_dist(data1)
fviz_dist(distance)
```

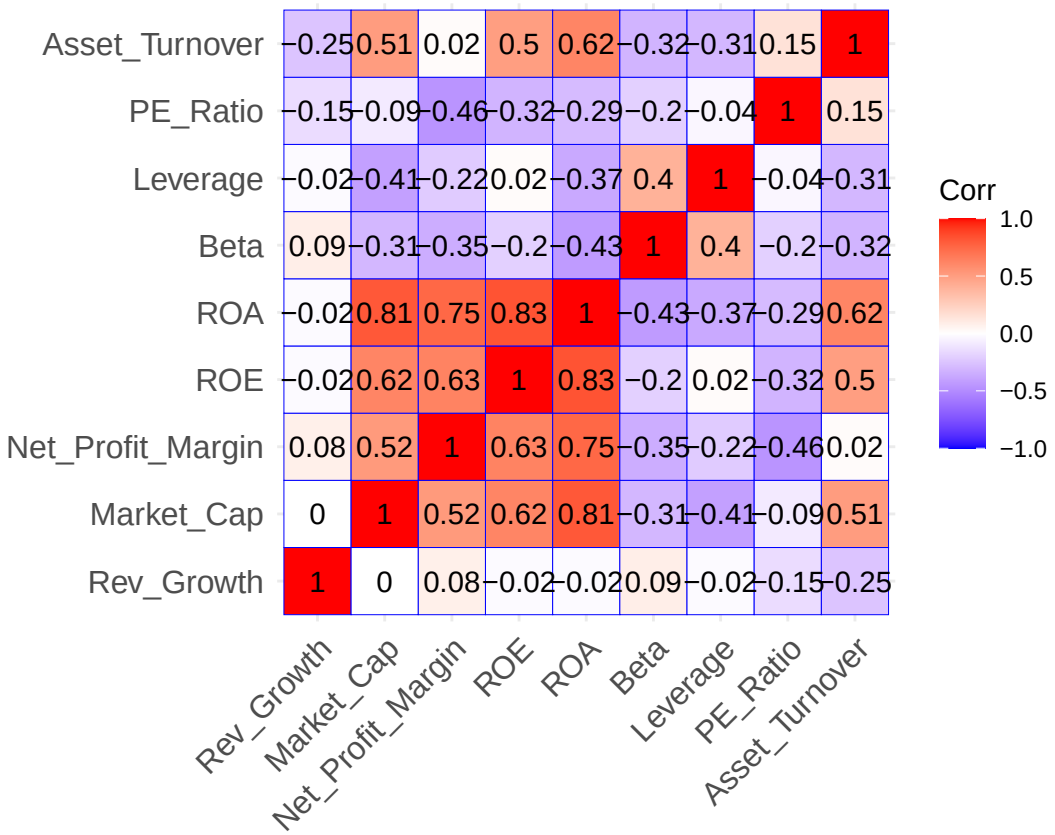


```
head(round(as.matrix(distance), 2),4)# To calculating the distance between the rows of data and viewing
```

```
##      ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK
## ABT 0.00 4.42 2.02 1.67 2.11 4.69 1.81 5.02 4.90 1.42 3.69 2.62 2.33 3.92 2.68
## AGN 4.42 0.00 3.95 4.91 4.64 4.85 5.42 5.61 6.70 5.14 6.75 4.47 5.32 5.48 5.44
## AHM 2.02 3.95 0.00 2.36 2.49 3.64 2.60 4.76 4.70 3.24 4.90 2.32 3.59 4.12 3.36
## AZN 1.67 4.91 2.36 0.00 2.63 5.07 1.57 5.72 4.97 2.41 2.96 3.28 1.96 4.27 1.86
##      NVS  PFE  PHA  SGP  WPI  WYE
## ABT 1.92 3.89 2.91 1.31 2.88 3.04
## AGN 5.47 6.91 2.37 4.73 5.01 6.45
## AHM 3.33 5.27 2.93 1.70 2.94 4.19
## AZN 3.06 3.11 3.72 1.08 3.41 3.32
```

To check the major variables, I using the correlation matrix method

```
set.seed(450)
corr <- cor(data1)
ggcorrplot(corr, outline.color = "blue", lab = TRUE, hc.order = TRUE, type = "full")
```



The k-means algorithm can be used to manually adjust K numbers to observe how the dataset clusters. I've chosen k values at random from 2,3,4 and 5, and a restart value of 25.

```
set.seed(450)
k2 <- kmeans(data1, centers = 2, nstart = 25)
k3 <- kmeans(data1, centers = 3, nstart = 25)
k4 <- kmeans(data1, centers = 4, nstart = 25)
k5 <- kmeans(data1, centers = 5, nstart = 25)
k2$size
```

```
## [1] 11 10
```

```
k3$size
```

```
## [1]  4  6 11
```

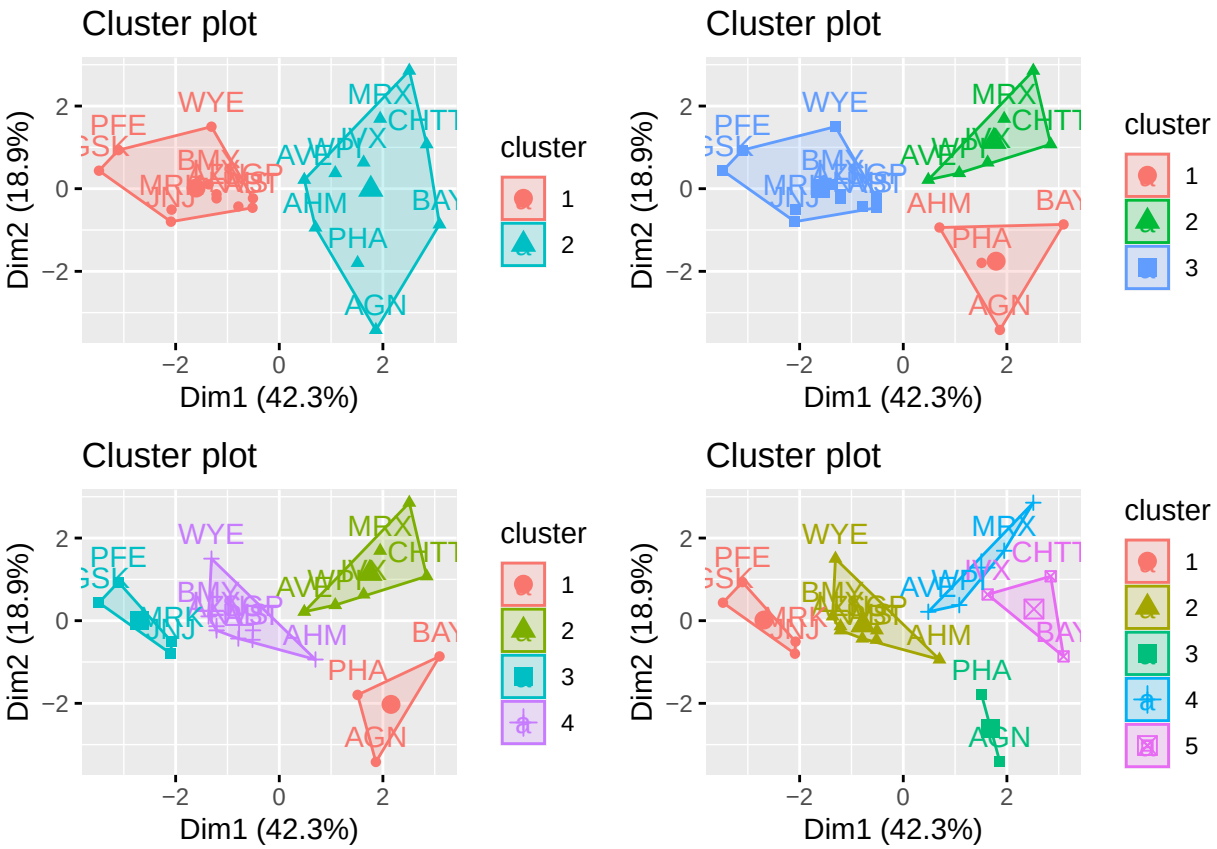
```
k4$size
```

```
## [1] 3 6 4 8
```

```
k5$size
```

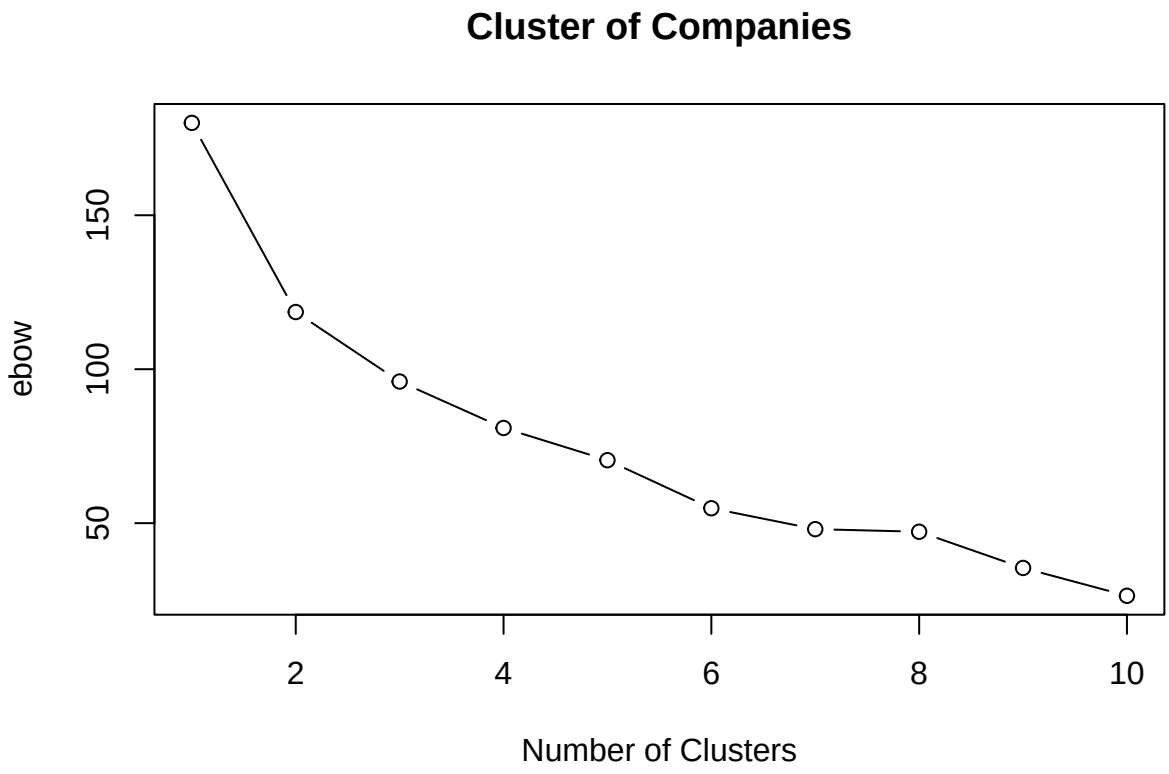
```
## [1] 4 8 2 4 3
```

```
k21 <- fviz_cluster(k2, data = data1)
k31 <- fviz_cluster(k3, data = data1)
k41 <- fviz_cluster(k4, data = data1)
k51 <- fviz_cluster(k5, data = data1)
grid.arrange(k21,k31,k41,k51)
```



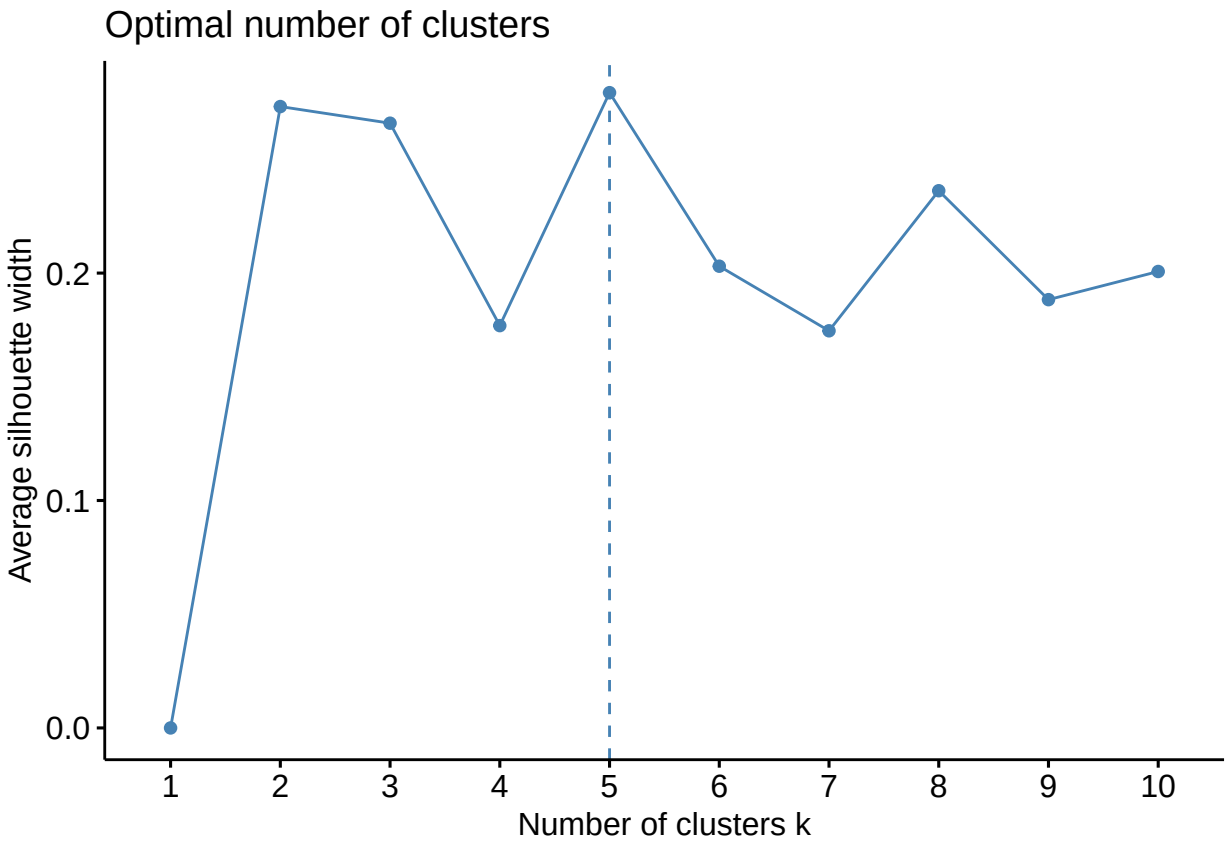
Elbow Method

```
set.seed(450)
ebow <- vector()
for(i in 1:10) ebow[i] <- sum(kmeans(data1,i)$withinss)
plot(1:10, ebow , type = "b" , main = paste('Cluster of Companies') , xlab = "Number of Clusters", ylab=
```



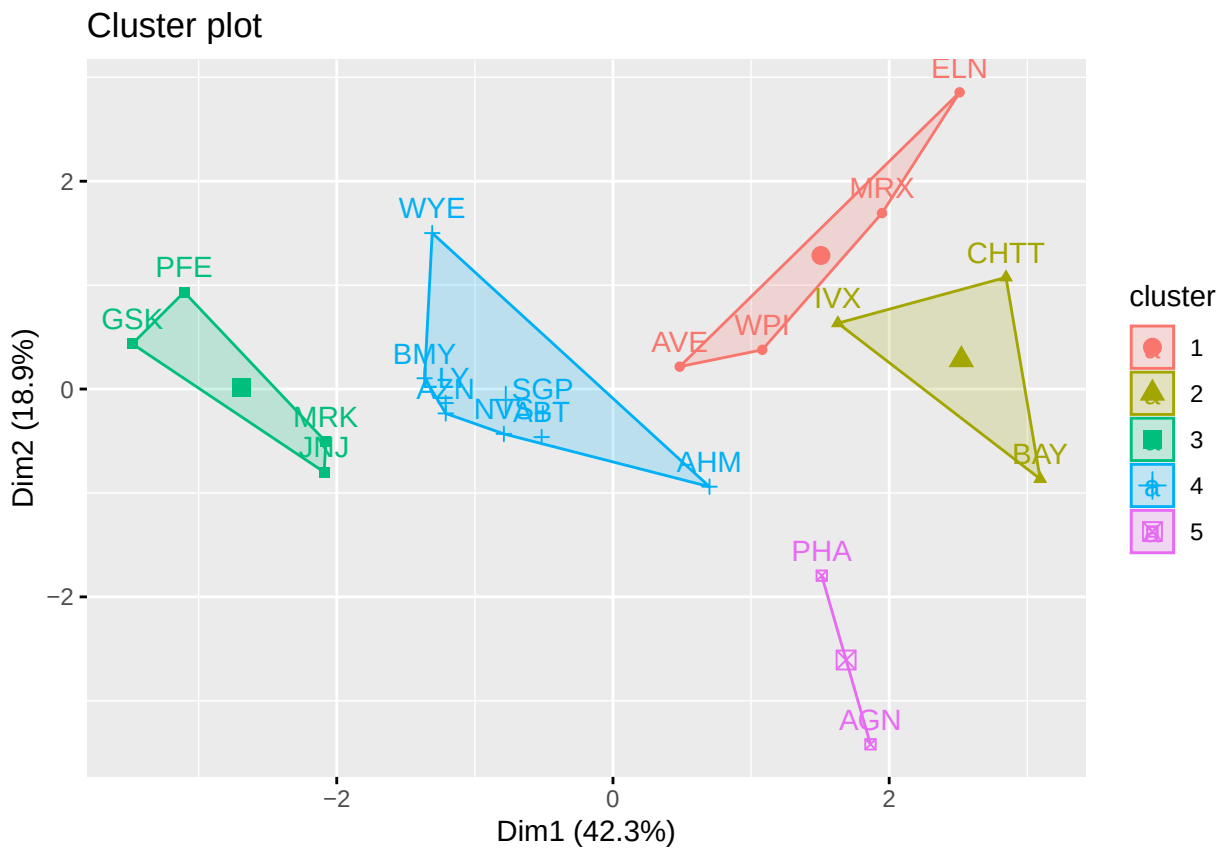
Silhouette Method

```
set.seed(450)
fviz_nbclust(data1, kmeans, method = "silhouette")
```



From above graph we find K clusters value "5"

```
set.seed(450)
k5 <- kmeans(data1, centers = 5, nstart = 25)
fviz_cluster(k5, data = data1)
```



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

```
print(k5)
```

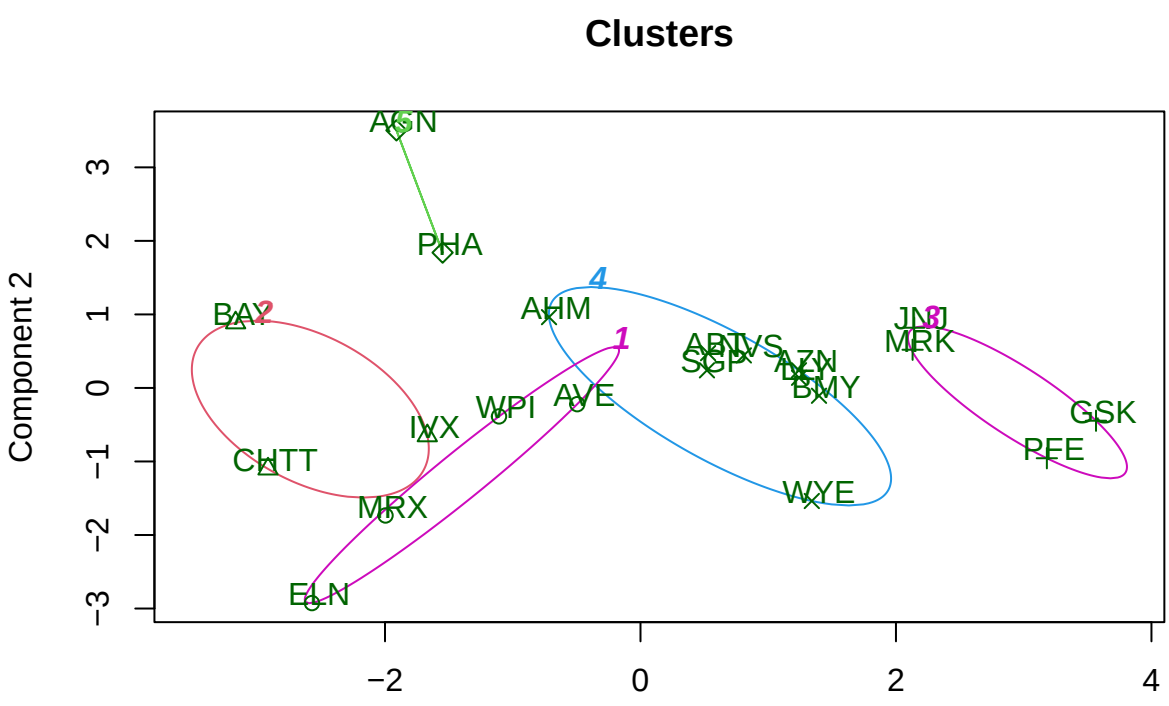
```
## K-means clustering with 5 clusters of sizes 4, 3, 4, 8, 2
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 3  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 4 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 5 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
##      Leverage Rev_Growth Net_Profit_Margin
## 1  0.06308085  1.5180158      -0.006893899
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.46807818  0.4671788       0.591242521
## 4 -0.27449312 -0.7041516       0.556954446
## 5 -0.14170336 -0.1168459      -1.416514761
##
## Clustering vector:
```

```
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    4    5    4    4    1    2    4    2    1    4    3    2    3    1    3    4
##  PFE  PHA  SGP  WPI  WYE
##    3    5    4    1    4
##
## Within cluster sum of squares by cluster:
## [1] 12.791257 15.595925  9.284424 21.879320  2.803505
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
da %>% mutate(Cluster = k5$cluster) %>% group_by(Cluster) %>% summarise_all("mean")
```

```
## # A tibble: 5 x 10
##   Cluster Market_Cap  Beta PE_Ratio   ROE   ROA Asset_~1 Lever~2 Rev_G~3 Net_P~4
##     <int>      <dbl> <dbl>    <dbl> <dbl> <dbl>    <dbl>   <dbl>   <dbl>   <dbl>
## 1       1      13.1  0.598     17.7  14.6  6.2     0.425   0.635   30.1    15.6
## 2       2       6.64 0.87      24.6  16.5  4.17    0.6     1.65     5.73    7.03
## 3       3     157.   0.48      22.2  44.4 17.7     0.95    0.22    18.5    19.6
## 4       4      55.8  0.414     20.3  28.7 12.7     0.738   0.371    5.59   19.4
## 5       5      31.9  0.405     69.5  13.2  5.6     0.75    0.475   12.1     6.4
## # ... with abbreviated variable names 1: Asset_Turnover, 2: Leverage,
## #   3: Rev_Growth, 4: Net_Profit_Margin
```

```
clusplot(data1,k5$cluster, main="Clusters",color = TRUE, labels = 2,lines = 0)
```



These two components explain 61.23 % of the point variability.

From above plot and data the companies are group into a varous clusters

##Cluster 1: MRK,PFE,GSK and JNJ ##Cluster 2: ABT,AHM,AZN,BMY,LLY,NVS,SGP and WYE ##Cluster 3: AGN and PHA ##Cluster 4: AVE,ELN,MRX and WYE ##Cluster 5: BAY,CHTT and IVX

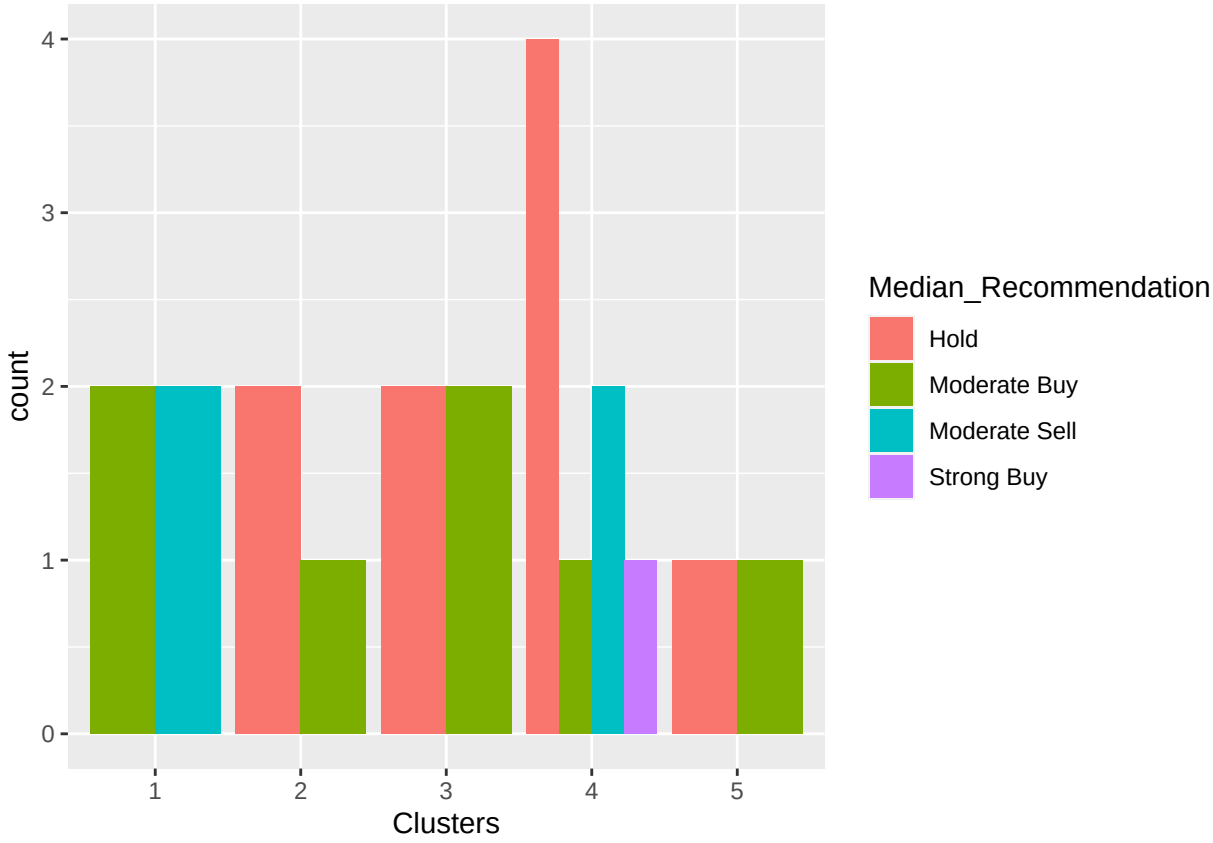
C. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)?

```
# Displaying the matrix and combining the clusters with the non-selected features
da1 <- data[,c(1,12,13,14)]
da2 <- as.data.frame(list(k5$cluster))
da1$cluster <- da2[,c(1)]
da1 %>% arrange(desc(da1$cluster))
```

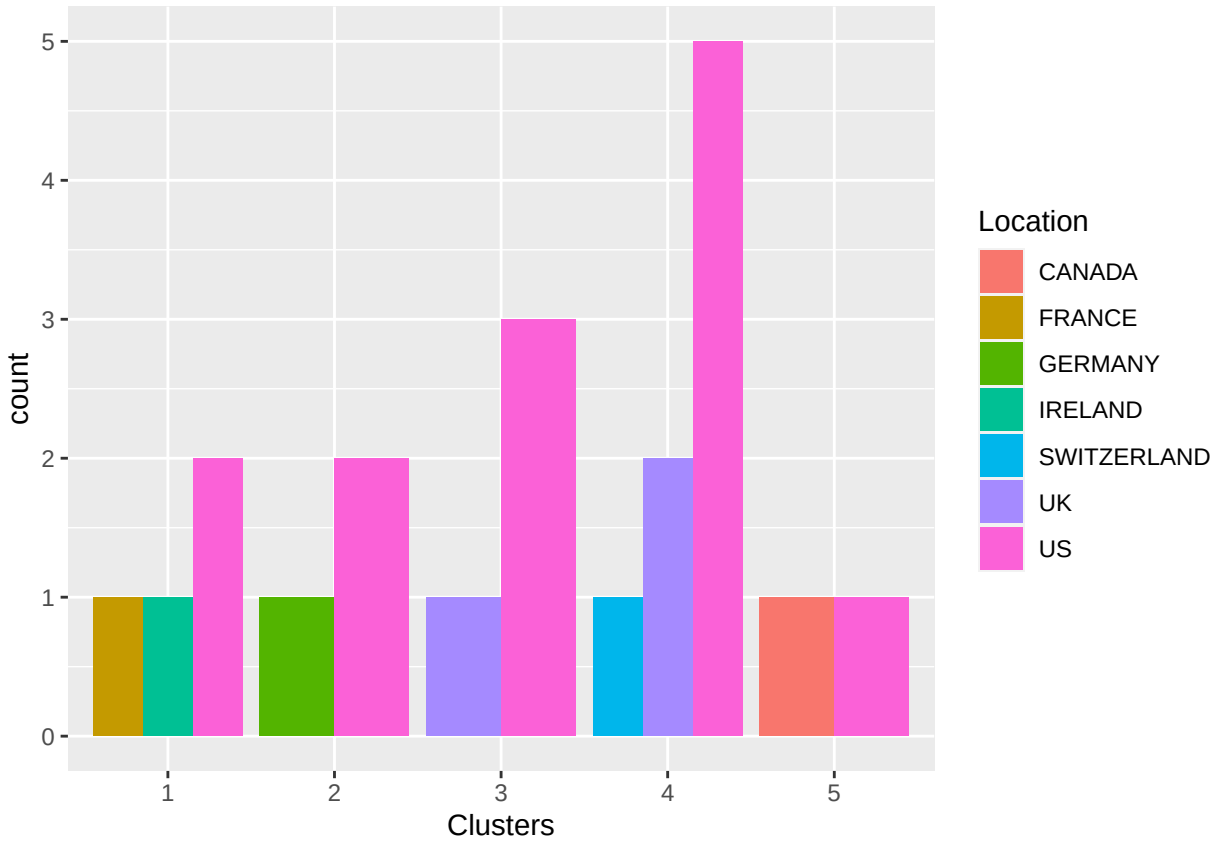
```
##    Symbol Median_Recommendation    Location Exchange cluster
## 1     AGN          Moderate Buy      CANADA     NYSE       5
## 2     PHA                  Hold          US     NYSE       5
## 3     ABT          Moderate Buy          US     NYSE       4
## 4     AHM            Strong Buy          UK     NYSE       4
## 5     AZN         Moderate Sell          UK     NYSE       4
## 6     BMY         Moderate Sell          US     NYSE       4
## 7     LLY                  Hold          US     NYSE       4
## 8     NVS                  Hold SWITZERLAND     NYSE       4
```

```
## 9     SGP                 Hold          US     NYSE       4
## 10    WYE                 Hold          US     NYSE       4
## 11    GSK                 Hold          UK     NYSE       3
## 12    JNJ          Moderate Buy          US     NYSE       3
## 13    MRK                 Hold          US     NYSE       3
## 14    PFE          Moderate Buy          US     NYSE       3
## 15    BAY                 Hold     GERMANY     NYSE       2
## 16   CHTT          Moderate Buy          US   NASDAQ       2
## 17    IVX                 Hold          US     AMEX       2
## 18    AVE          Moderate Buy      FRANCE     NYSE       1
## 19    ELN         Moderate Sell     IRELAND     NYSE       1
## 20    MRX          Moderate Buy          US     NYSE       1
## 21    WPI         Moderate Sell          US     NYSE       1
```

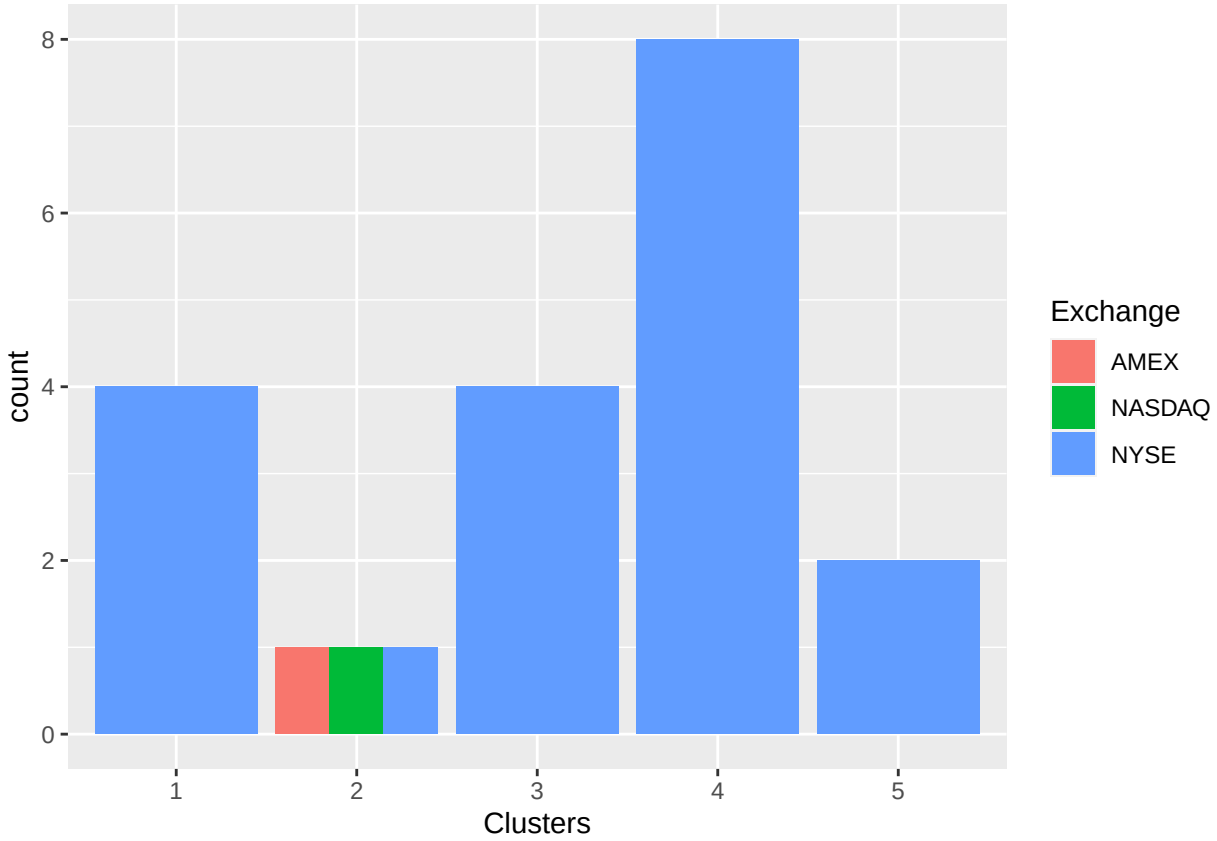
```
data2 <- data[12:14] %>% mutate(Clusters=k5$cluster)
ggplot(data2, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position='dodge')+
```




```
ggplot(data2, mapping = aes(factor(Clusters),fill = Location))+geom_bar(position = 'dodge')+labs(x ='Cl
```




```
ggplot(data2, mapping = aes(factor(Clusters),fill = Exchange))+geom_bar(position = 'dodge')+labs(x ='Cl
```



the Median Recommendation and Cluster

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

Cluster 1: Best Market

Cluster 2: Uncontrolled

Cluster 3: excellent

Cluster 4: take a risk

Cluster 5: Workable