

FML-Assignment-4

shiva gadila

2023-03-18

#Importing the Dataset

```
Pharmaceuticals <- read.csv("~/Downloads/Pharmaceuticals.csv")
summary(Pharmaceuticals)

##      Symbol      Name      Market_Cap      Beta
## Length:21      Length:21      Min.   : 0.41      Min.   :0.1800
## Class :character Class :character 1st Qu.: 6.30      1st Qu.:0.3500
## Mode  :character Mode  :character Median : 48.19      Median :0.4600
##                                     Mean  : 57.65      Mean   :0.5257
##                                     3rd Qu.: 73.84      3rd Qu.:0.6500
##                                     Max.   :199.47      Max.   :1.1100
##      PE_Ratio      ROE      ROA      Asset_Turnover      Leverage
## Min.   : 3.60      Min.   : 3.9      Min.   : 1.40      Min.   :0.3      Min.
## :0.0000
## 1st Qu.:18.90      1st Qu.:14.9      1st Qu.: 5.70      1st Qu.:0.6      1st
## Qu.:0.1600
## Median :21.50      Median :22.6      Median :11.20      Median :0.6      Median
## :0.3400
## Mean   :25.46      Mean   :25.8      Mean   :10.51      Mean   :0.7      Mean
## :0.5857
## 3rd Qu.:27.90      3rd Qu.:31.0      3rd Qu.:15.00      3rd Qu.:0.9      3rd
## Qu.:0.6000
## Max.   :82.50      Max.   :62.9      Max.   :20.30      Max.   :1.1      Max.
## :3.5100
##      Rev_Growth      Net_Profit_Margin      Median_Recommendation      Location
## Min.   : -3.17      Min.   : 2.6      Length:21      Length:21
## 1st Qu.: 6.38      1st Qu.:11.2      Class :character      Class :character
## Median : 9.37      Median :16.1      Mode  :character      Mode  :character
## Mean   :13.37      Mean   :15.7
## 3rd Qu.:21.87      3rd Qu.:21.1
## Max.   :34.21      Max.   :25.5
##      Exchange
## Length:21
## Class :character
## Mode  :character
##
##
##
str(Pharmaceuticals)
```

```
## 'data.frame':    21 obs. of  14 variables:
## $ Symbol          : chr  "ABT" "AGN" "AHM" "AZN" ...
## $ Name            : chr  "Abbott Laboratories" "Allergan, Inc."
"Amersham plc" "AstraZeneca PLC" ...
## $ 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" ...
```

#Loading the Packages

```
library(readr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(tidyverse)

## — Attaching core tidyverse packages ————— tidyverse
2.0.0 —
## ✓ forcats    1.0.0      ✓ stringr    1.5.0
```

```
## ✓ lubridate 1.9.2      ✓ tibble      3.1.8
## ✓ purrr      1.0.1      ✓ tidyr      1.3.0

## — Conflicts —————
tidyverse_conflicts() —
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()      masks stats::lag()
## ✗ purrr::lift()     masks caret::lift()
## ⓘ Use the [8;;http://conflicted.r-lib.org/conflicted-package]8;; to force
all conflicts to become errors

library(cluster)
library(gridExtra)

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

library(ggrepel)
library(factoextra)

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

library(flexclust)

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

library(ggcorrplot)
library(FactoMineR)
```

#(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

#Selecting the numerical variables and removing the dataset's null values.
colSums(is.na(Pharmaceuticals))

```
##           Symbol           Name           Market_Cap
##           0             0             0
##           Beta           PE_Ratio          ROE
##           0             0             0
##           ROA           Asset_Turnover      Leverage
##           0             0             0
##           Rev_Growth     Net_Profit_Margin  Median_Recommendation
##           0             0             0
```

```

##           Location           Exchange
##           0             0

row.names(Pharmaceuticals)<- Pharmaceuticals[,1]
Pharmaceuticals_data_num<- Pharmaceuticals[, 3:11]
head(Pharmaceuticals_data_num)

##      Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## ABT      68.44 0.32   24.7 26.4 11.8           0.7      0.42      7.54
## AGN       7.58 0.41   82.5 12.9  5.5           0.9      0.60      9.16
## AHM       6.30 0.46   20.7 14.9  7.8           0.9      0.27      7.05
## AZN      67.63 0.52   21.5 27.4 15.4           0.9      0.00     15.00
## AVE      47.16 0.32   20.1 21.8  7.5           0.6      0.34     26.81
## BAY      16.90 1.11   27.9  3.9  1.4           0.6      0.00     -3.17
##      Net_Profit_Margin
## ABT              16.1
## AGN              5.5
## AHM             11.2
## AZN             18.0
## AVE             12.9
## BAY              2.6

# Normalizing and scaling the dataset.
Pharmaceuticals_scale <- scale(Pharmaceuticals_data_num)
head(Pharmaceuticals_scale)

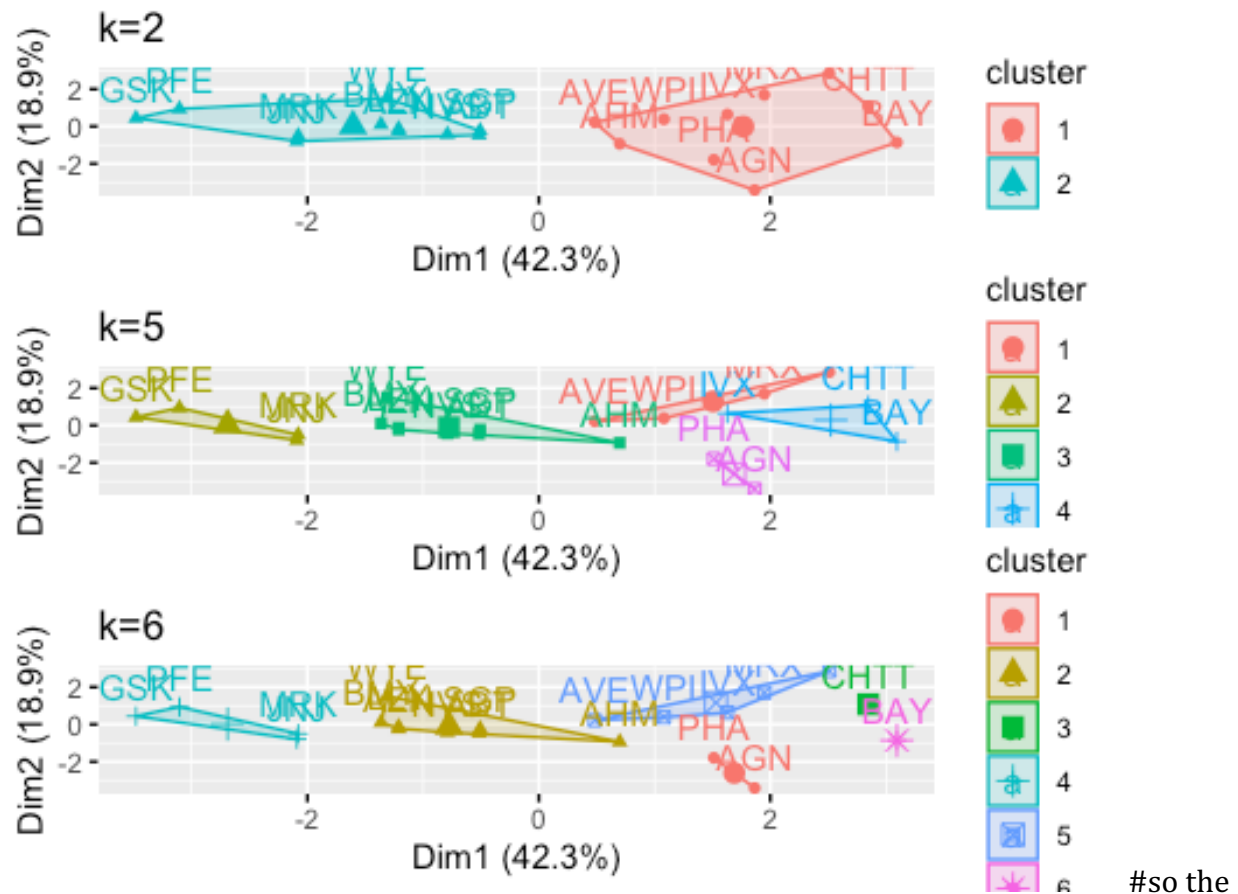
##      Market_Cap      Beta      PE_Ratio      ROE      ROA
Asset_Turnover
## ABT  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121 -5.121077e-
16
## AGN -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871  9.225312e-
01
## AHM -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700  9.225312e-
01
## AZN  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259  9.225312e-
01
## AVE -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461 -4.612656e-
01
## BAY -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612 -4.612656e-
01
##      Leverage Rev_Growth Net_Profit_Margin
## ABT -0.2120979 -0.5277675      0.06168225
## AGN  0.0182843 -0.3811391     -1.55366706
## AHM -0.4040831 -0.5721181     -0.68503583
## AZN -0.7496565  0.1474473      0.35122600
## AVE -0.3144900  1.2163867     -0.42597037
## BAY -0.7496565 -1.4971443     -1.99560225

normalization_data <- as.data.frame(scale(Pharmaceuticals_data_num))

```

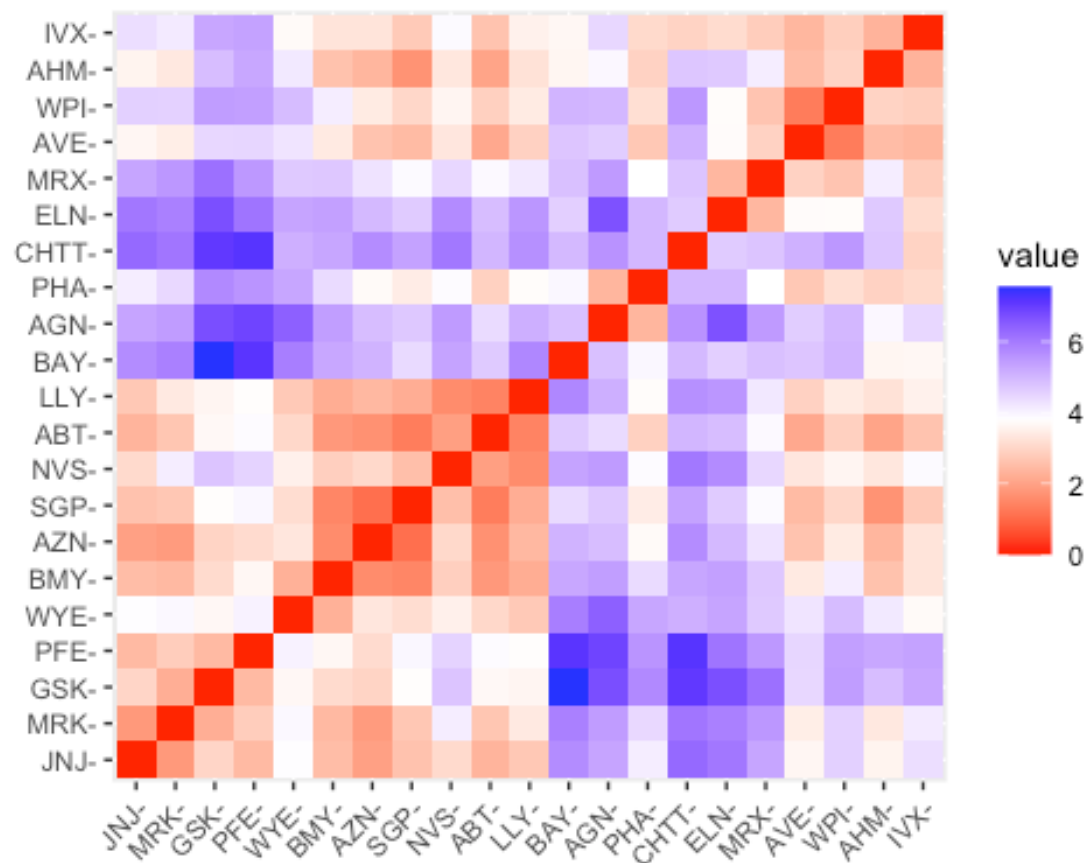
Using multiple K values, compute K-means clustering for various centers, and compare the results.

```
kmeans_1 <- kmeans(Pharmaceuticals_scale, centers = 2, nstart = 30)
kmeans_2<- kmeans(Pharmaceuticals_scale, centers = 5, nstart = 30)
kmeans_3<- kmeans(Pharmaceuticals_scale, centers = 6, nstart = 30)
Plot_1<-fviz_cluster(kmeans_1, data = Pharmaceuticals_scale)+ggtitle("k=2")
plot_2<-fviz_cluster(kmeans_2, data = Pharmaceuticals_scale)+ggtitle("k=5")
plot_3<-fviz_cluster(kmeans_3, data = Pharmaceuticals_scale)+ggtitle("k=6")
grid.arrange(Plot_1,plot_2,plot_3, nrow = 3)
```

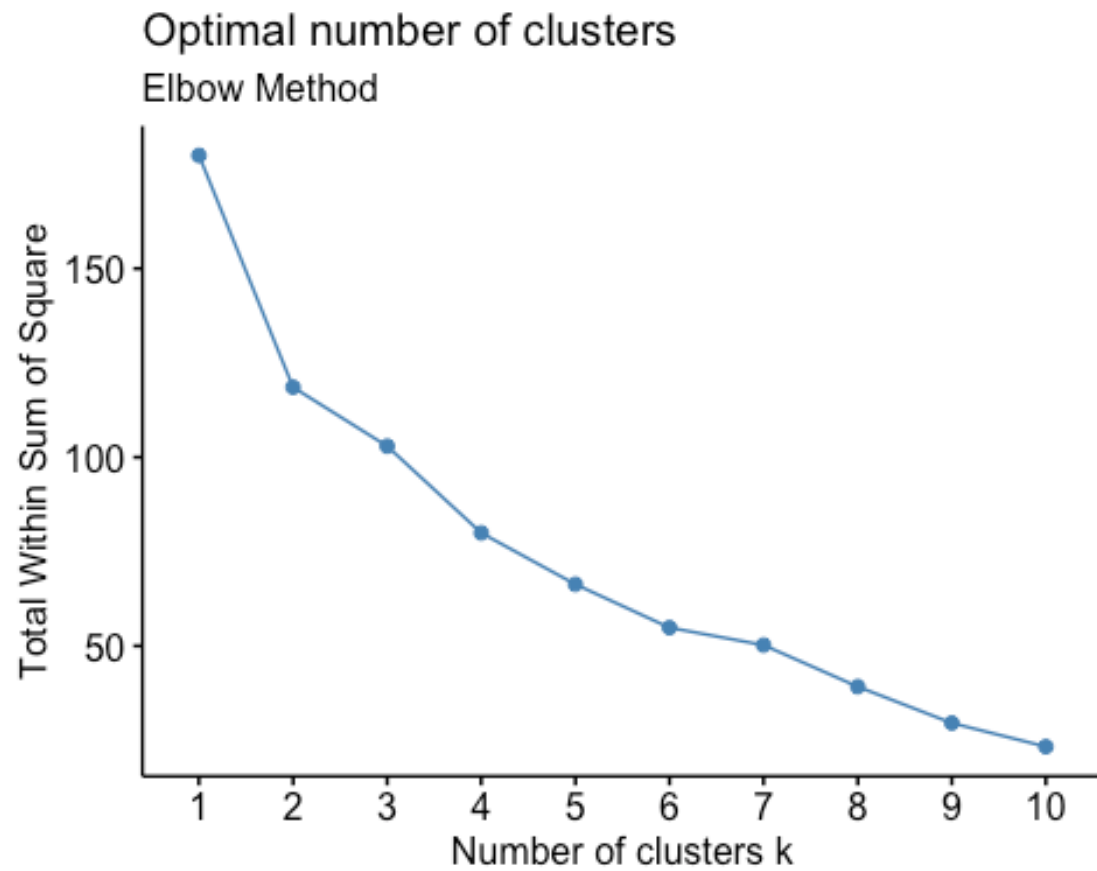


recommended number of clusters is k=2 i.e plot2

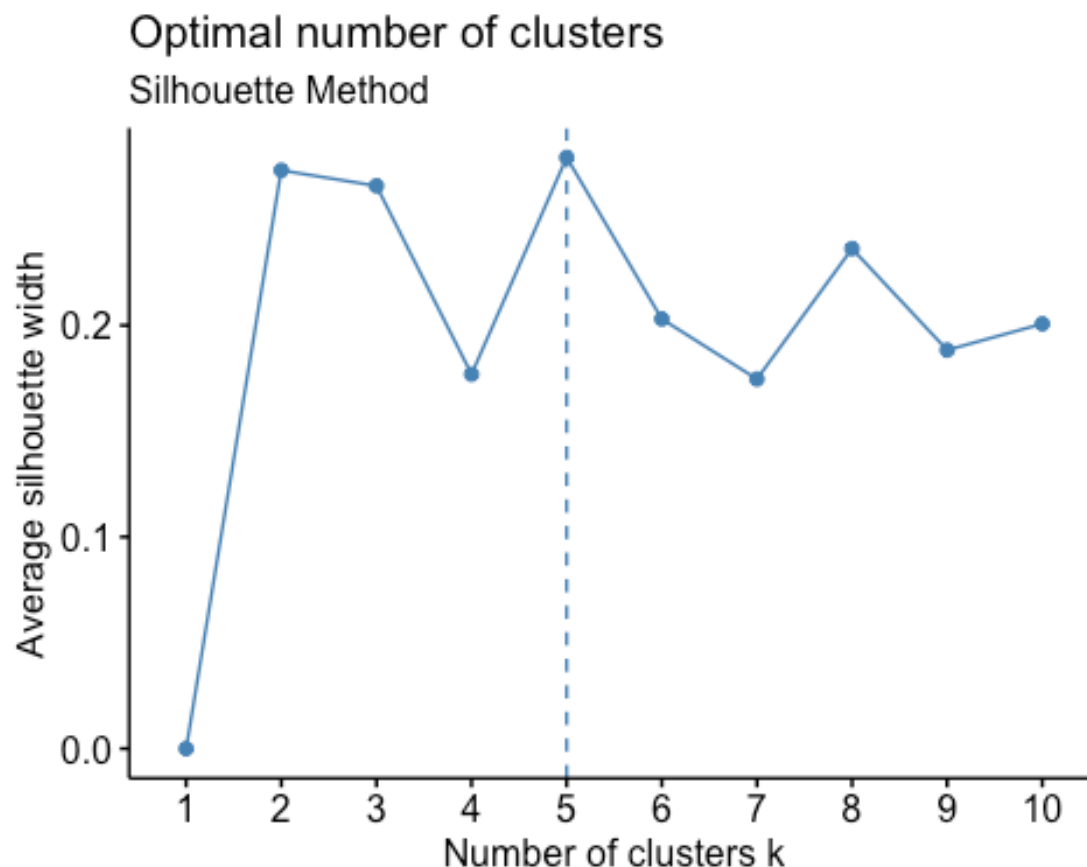
```
distance<- dist(Pharmaceuticals_scale, method = "euclidean")
fviz_dist(distance)
```



```
# Estimating the number of clusters
# Scaling the data using the Elbow Method to determine k's value
fviz_nbclust(normalization_data, FUNcluster = kmeans, method = "wss") +
labs(subtitle = "Elbow Method")
```



```
# The number of clusters is calculated by scaling the data using the  
Silhouette Method.  
fviz_nbclust(normalization_data,FUNcluster = kmeans,method =  
"silhouette")+labs(subtitle="Silhouette Method")
```



Final analysis, extraction of data from five clusters, and presentation of the data

```
set.seed(300)
```

```
final_Cluster<- kmeans(Pharmaceuticals_scale, 5, nstart = 25)
```

```
print(final_Cluster)
```

```
## K-means clustering with 5 clusters of sizes 8, 3, 4, 2, 4
```

```
##
```

```
## Cluster means:
```

```
##   Market_Cap      Beta  PE_Ratio      ROE      ROA Asset_Turnover
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915    0.1729746
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478   -0.4612656
## 3 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428   -1.2684804
## 4 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951    0.2306328
## 5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431    1.1531640
```

```
##   Leverage Rev_Growth Net_Profit_Margin
```

```
## 1 -0.27449312 -0.7041516    0.556954446
## 2  1.36644699 -0.6912914   -1.320000179
## 3  0.06308085  1.5180158   -0.006893899
## 4 -0.14170336 -0.1168459   -1.416514761
## 5 -0.46807818  0.4671788    0.591242521
```

```
##
```

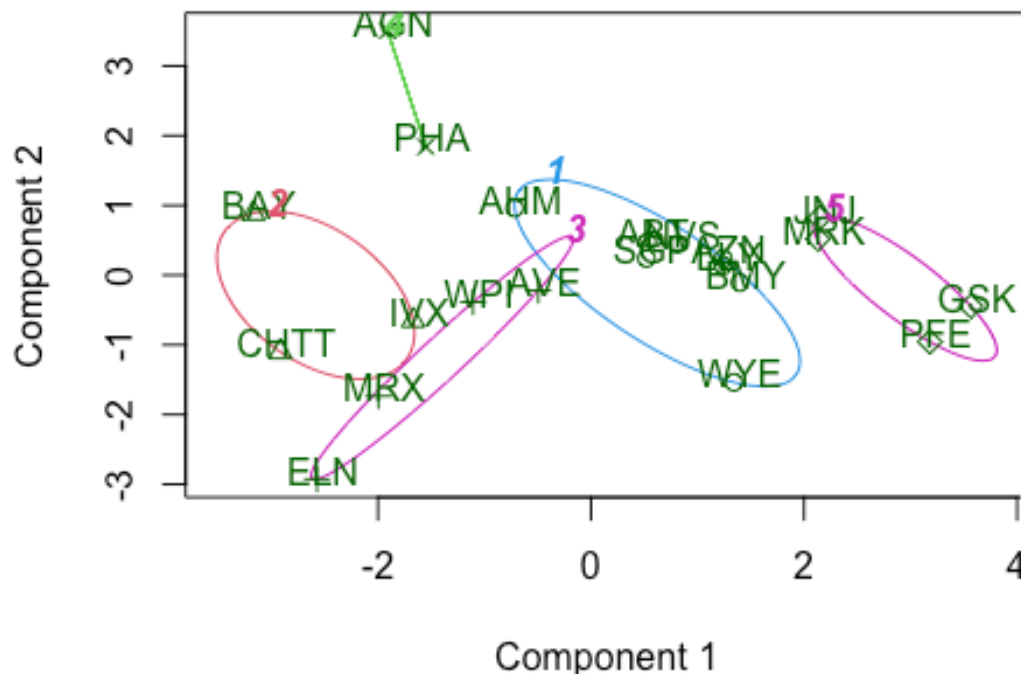
```
## Clustering vector:
```



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

clusplot(Pharmaceuticals_scale,final_Cluster$cluster, color = TRUE, labels =
2,lines = 0)
```

CLUSPLOT(Pharmaceuticals_scale)



These two components explain 61.23 % of the point variance #b)

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

#Cluster 1 - AHM,SGP,WYE,BMY,AZN, ABT, NVS, LLY (Lowest Market_Cap,Lowest Beta,Lowest PE_Ratio,highest Leverage,highest Rev_Growth.)

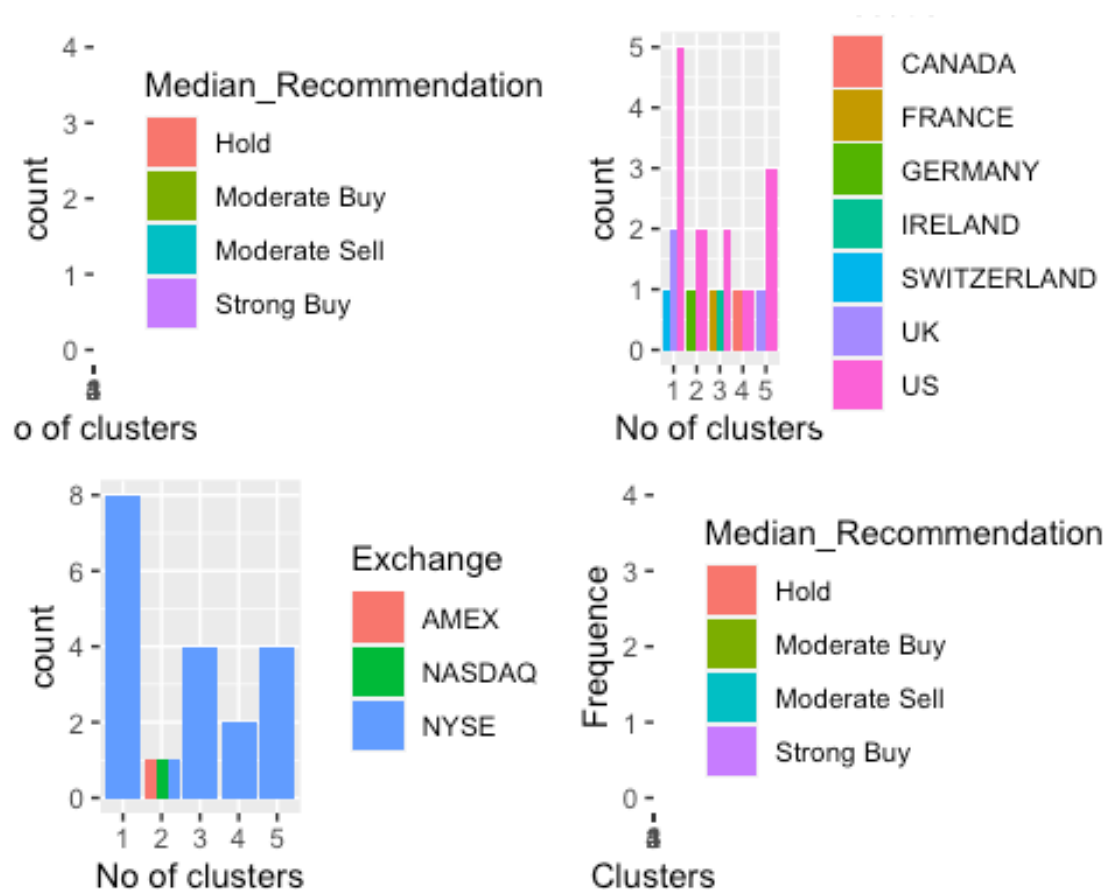
#Cluster 2 - BAY, CHTT, IVX (Lowest Rev_Growth, highest Beta and Leverage, lowest Net_Profit_Margin)
#Cluster 3 - WPI, MRX, ELN, AVE (Lowest PE_Ratio, highest ROE, lowest ROA, lowest Net_Profit_Margin, highest Rev_Growth)
#Cluster 4 - AGN, PHA (lowest Beta, lowest Asset_Turnover, highest PE Ratio)
#Cluster 5 - JNJ, MRK, PFE, GSK (Highest Market_Cap, ROE, ROA, Asset_Turnover Ratio and lowest Beta/PE Ratio)

```
Pharmaceuticals_Cluster <- Pharmaceuticals[,c(12,13,14)]%>% mutate(clusters =
final_Cluster$cluster)%>% arrange(clusters, ascending = TRUE)
Pharmaceuticals_Cluster
```

##	Median_Recommendation	Location	Exchange	clusters
## ABT	Moderate Buy	US	NYSE	1
## AHM	Strong Buy	UK	NYSE	1
## AZN	Moderate Sell	UK	NYSE	1
## BMY	Moderate Sell	US	NYSE	1
## LLY	Hold	US	NYSE	1
## NVS	Hold	SWITZERLAND	NYSE	1
## SGP	Hold	US	NYSE	1
## WYE	Hold	US	NYSE	1
## BAY	Hold	GERMANY	NYSE	2
## CHTT	Moderate Buy	US	NASDAQ	2
## IVX	Hold	US	AMEX	2
## AVE	Moderate Buy	FRANCE	NYSE	3
## ELN	Moderate Sell	IRELAND	NYSE	3
## MRX	Moderate Buy	US	NYSE	3
## WPI	Moderate Sell	US	NYSE	3
## AGN	Moderate Buy	CANADA	NYSE	4
## PHA	Hold	US	NYSE	4
## GSK	Hold	UK	NYSE	5
## JNJ	Moderate Buy	US	NYSE	5
## MRK	Hold	US	NYSE	5
## PFE	Moderate Buy	US	NYSE	5

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

```
plot1<-ggplot(Pharmaceuticals_Cluster, mapping = aes(factor(clusters),
fill=Median_Recommendation))+geom_bar(position = 'dodge')+labs(x='No of
clusters')
plot2<- ggplot(Pharmaceuticals_Cluster, mapping = aes(factor(clusters), fill =
Location))+geom_bar(position = 'dodge')+labs(x='No of clusters')
plot3<- ggplot(Pharmaceuticals_Cluster, mapping = aes(factor(clusters), fill =
Exchange))+geom_bar(position = 'dodge')+labs(x='No of clusters')
plot4 <- ggplot(Pharmaceuticals_Cluster, mapping = aes(factor(clusters),
fill=Median_Recommendation)) + geom_bar(position = 'dodge') +
labs(x='Clusters', y='Frequency')
grid.arrange(plot1, plot2, plot3, plot4)
```



#1 Cluster: In this cluster, which also has medians for Hold, Moderate Buy, Moderate Sell, and Strong Buy, the Hold median is the highest. They hail from Switzerland, the United States, and are Listed on the NYSE.

#2 Cluster: Despite the fact that the companies are evenly distributed across the AMEX, NASDAQ, and NYSE, there is a distinct Hold and Moderate Buy median and a distinct count between the United States and Germany.

#3 Cluster: Listed on the NYSE, with separate counts for the United States, Ireland, and France, and moderate buy and sell medians that are equal.

#4, Cluster: distributed throughout the United States and the United Kingdom and listed in, shares the same hold and moderate buy medians

#Cluster 5: # only on the NYSE, equally distributed in the US and Canada, with medians of Hold and Moderate Buy

#The clusters follow a particular pattern in relation to the media recommendation variable:

#Hold Recommendation applies to Clusters 1 and 2.

#The buy recommendation for Clusters 3, 4, and 5 is moderate.

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

*#Cluster 1 :-Buy Cluster
#Cluster 2 :- Sceptical Cluster
#Cluster 3 :- Moderate Buy Cluster
#Cluster 4 :- Hold Cluster
#Cluster 5 :- High Hold Cluster*