Assignment 04 -FML

Chathurani Ekanayake

2023-03-19

install.packages("factoextra") install.packages("flexclust")

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   1.0.1
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.3.0      ✓ stringr 1.5.0
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(factoextra)
```

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

```
library(ISLR)
library(flexclust)
```

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

```
library(tinytex)
library(cluster)
library(FactoMineR)
library(ggcorrplot)
library(NbClust)
```

```
library(readxl)
Pharmaceuticals <- read_excel("F:/1st sem/ML/Assignment     04/Pharmaceuticals.xlsx")
View(Pharmaceuticals)
```

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.

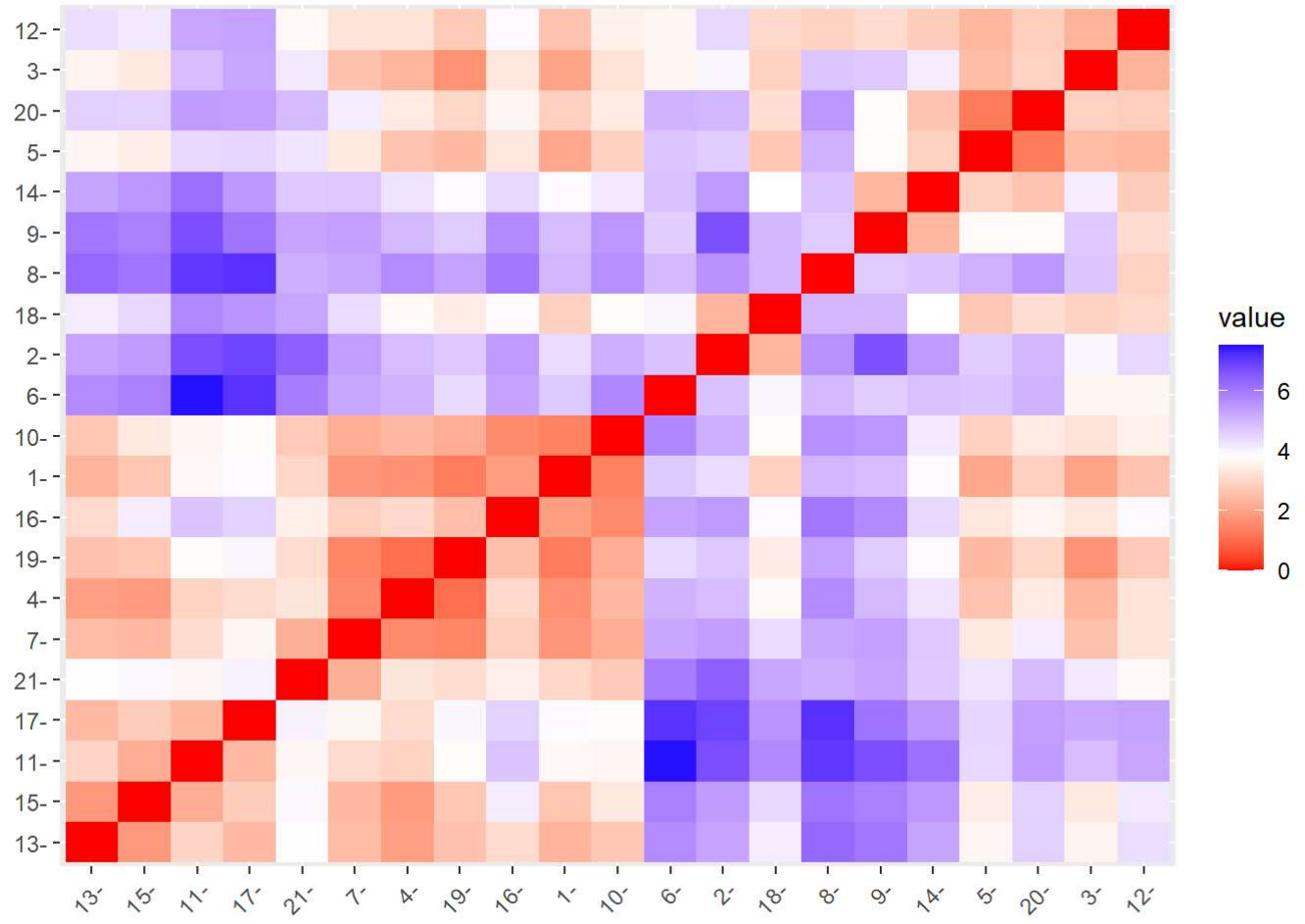
```
t(t(names(Pharmaceuticals)))
```

```
##       [,1]
##  [1,] "Symbol"
##  [2,] "Name"
##  [3,] "Market_Cap"
##  [4,] "Beta"
##  [5,] "PE_Ratio"
##  [6,] "ROE"
##  [7,] "ROA"
##  [8,] "Asset_Turnover"
##  [9,] "Leverage"
## [10,] "Rev_Growth"
## [11,] "Net_Profit_Margin"
## [12,] "Median_Recommendation"
## [13,] "Location"
## [14,] "Exchange"
```

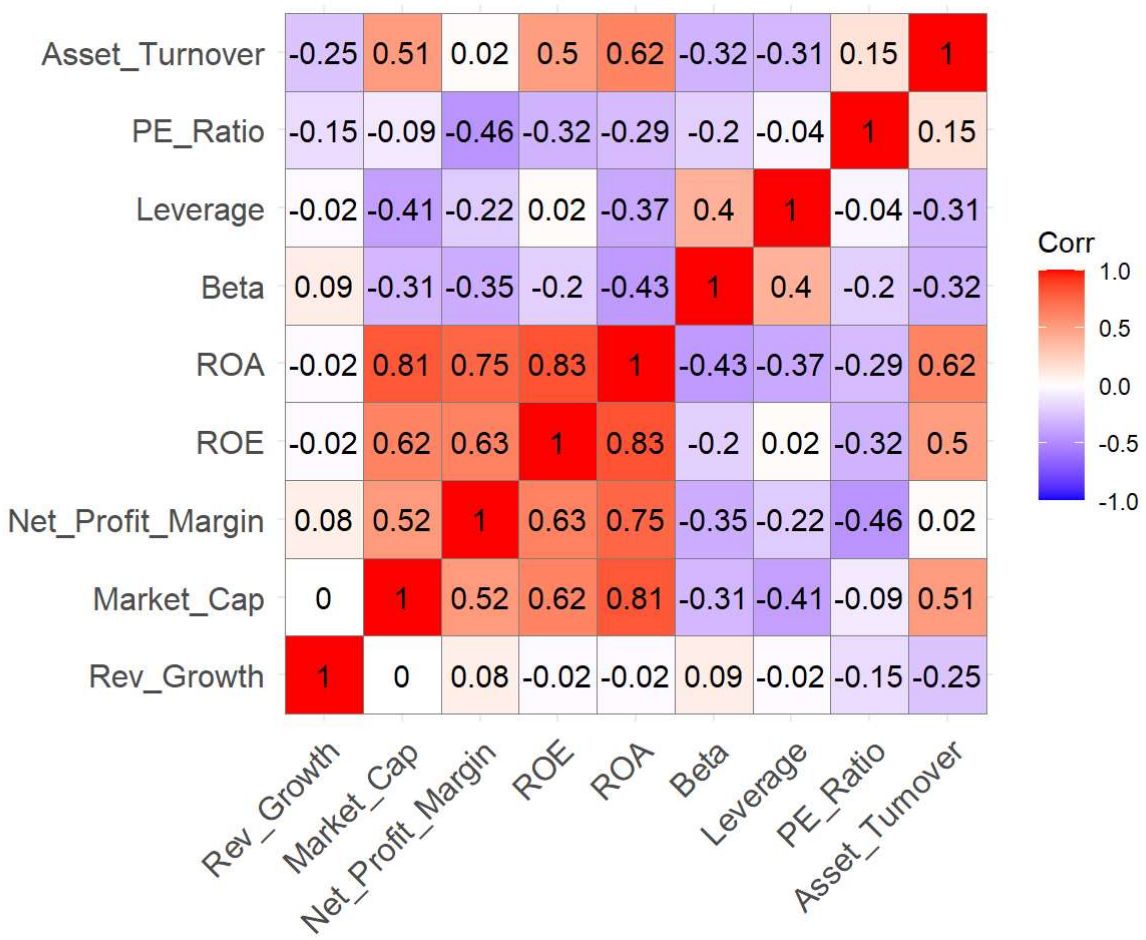
```
# Selecting numerical variables
df <- Pharmaceuticals[,c(3:11)]
t(t(names(df)))
```

```
##       [,1]
##  [1,] "Market_Cap"
##  [2,] "Beta"
##  [3,] "PE_Ratio"
##  [4,] "ROE"
##  [5,] "ROA"
##  [6,] "Asset_Turnover"
##  [7,] "Leverage"
##  [8,] "Rev_Growth"
##  [9,] "Net_Profit_Margin"
```

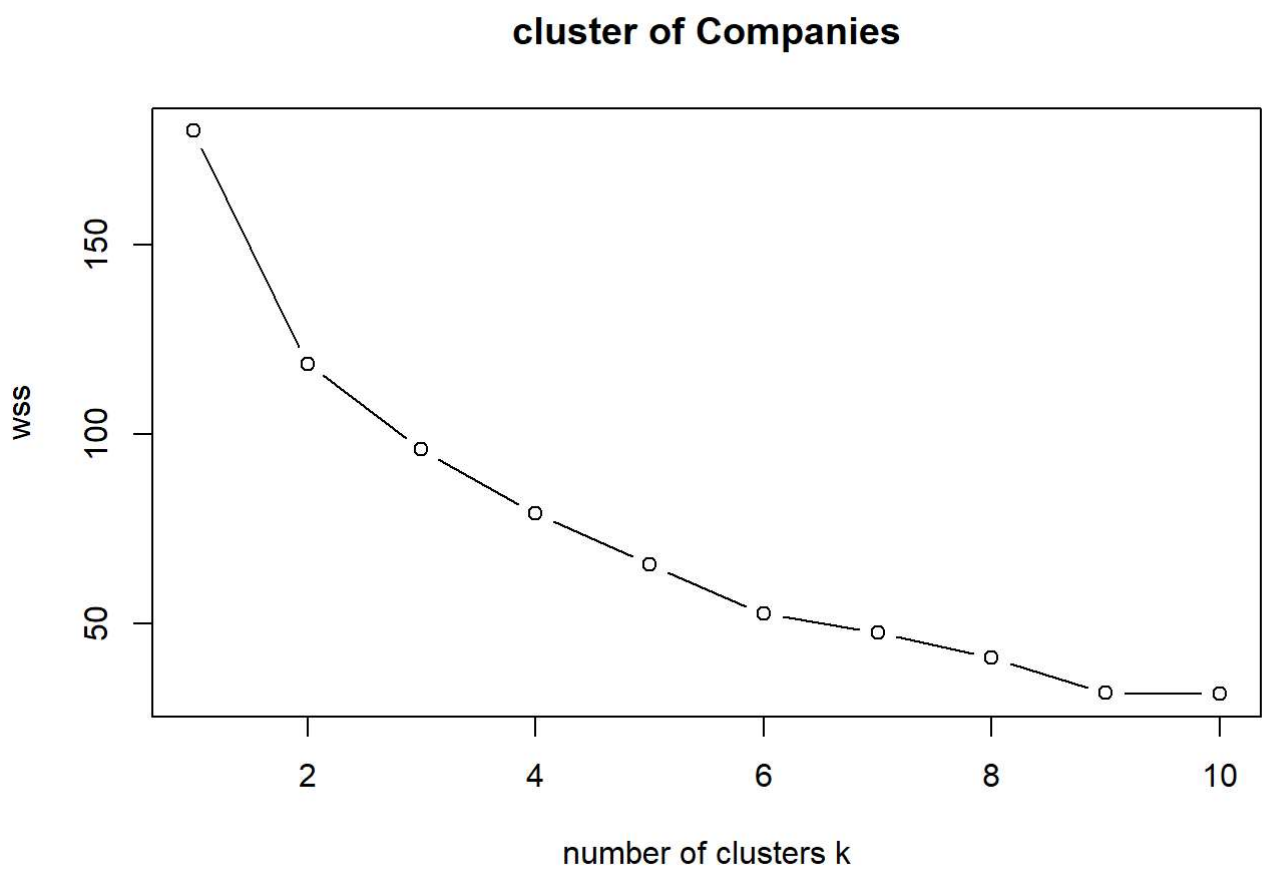
```
# Normalizing the data
df1 <- scale(df)
distance <- get_dist(df1)
fviz_dist(distance,)
```



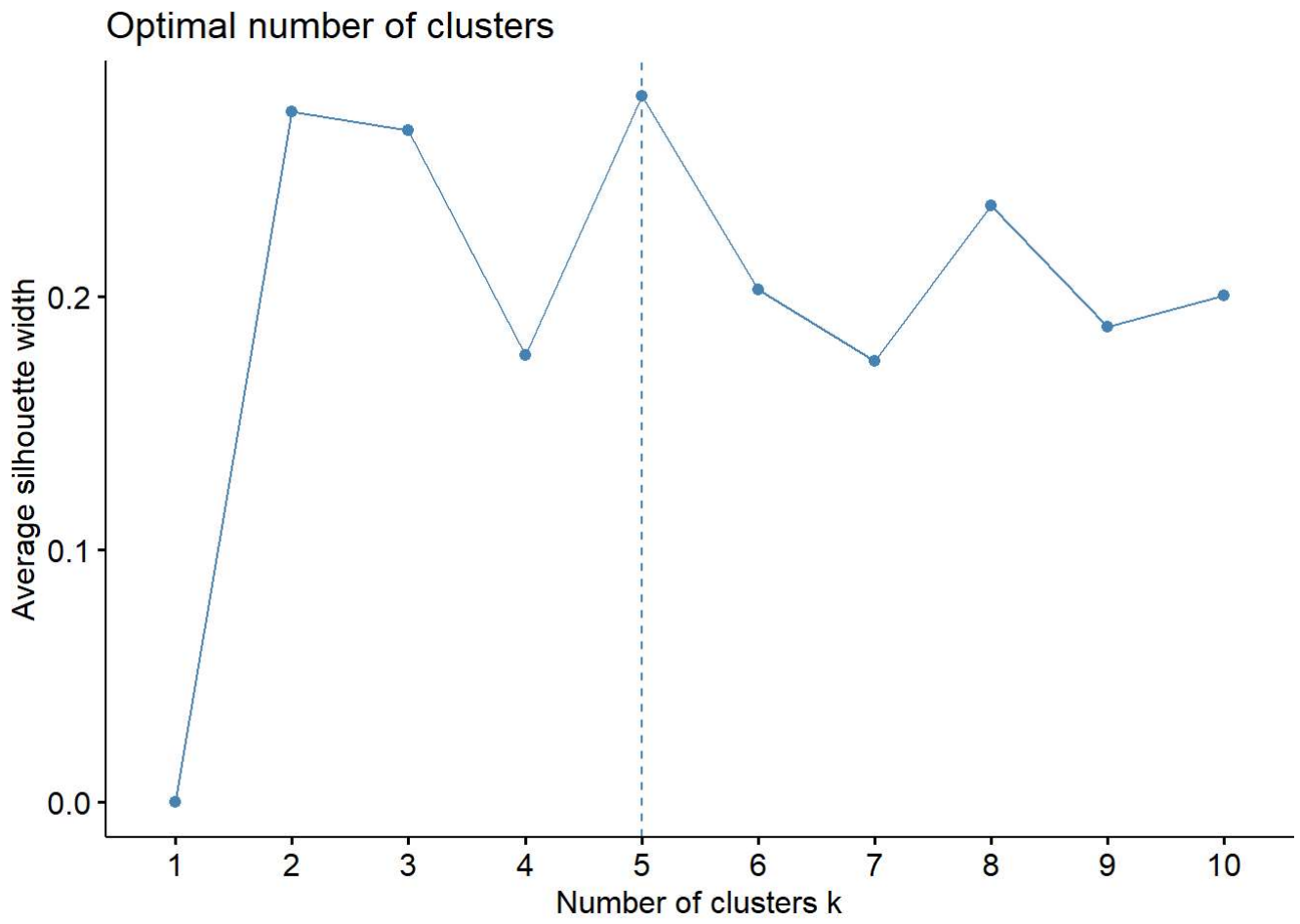

```
corr<-cor(df1)
ggcorrplot(corr,outline.color = "grey50",lab = TRUE,hc.order = TRUE,type = "full")
```




```
# determine the optimal value of clustering using elbow method & silhouette index.
set.seed(10)
wss <- vector()
for(i in 1:10) wss[i]<- sum(kmeans(df1,i)$withinss)
plot(1:10, wss, type="b", main=paste("cluster of Companies"), xlab = "number of clusters
k", ylab="wss")
```




```
# According to the above graph it shows the optimal number of clusters is 6

fviz_nbclust(df1, stats::kmeans, method = "silhouette")
```

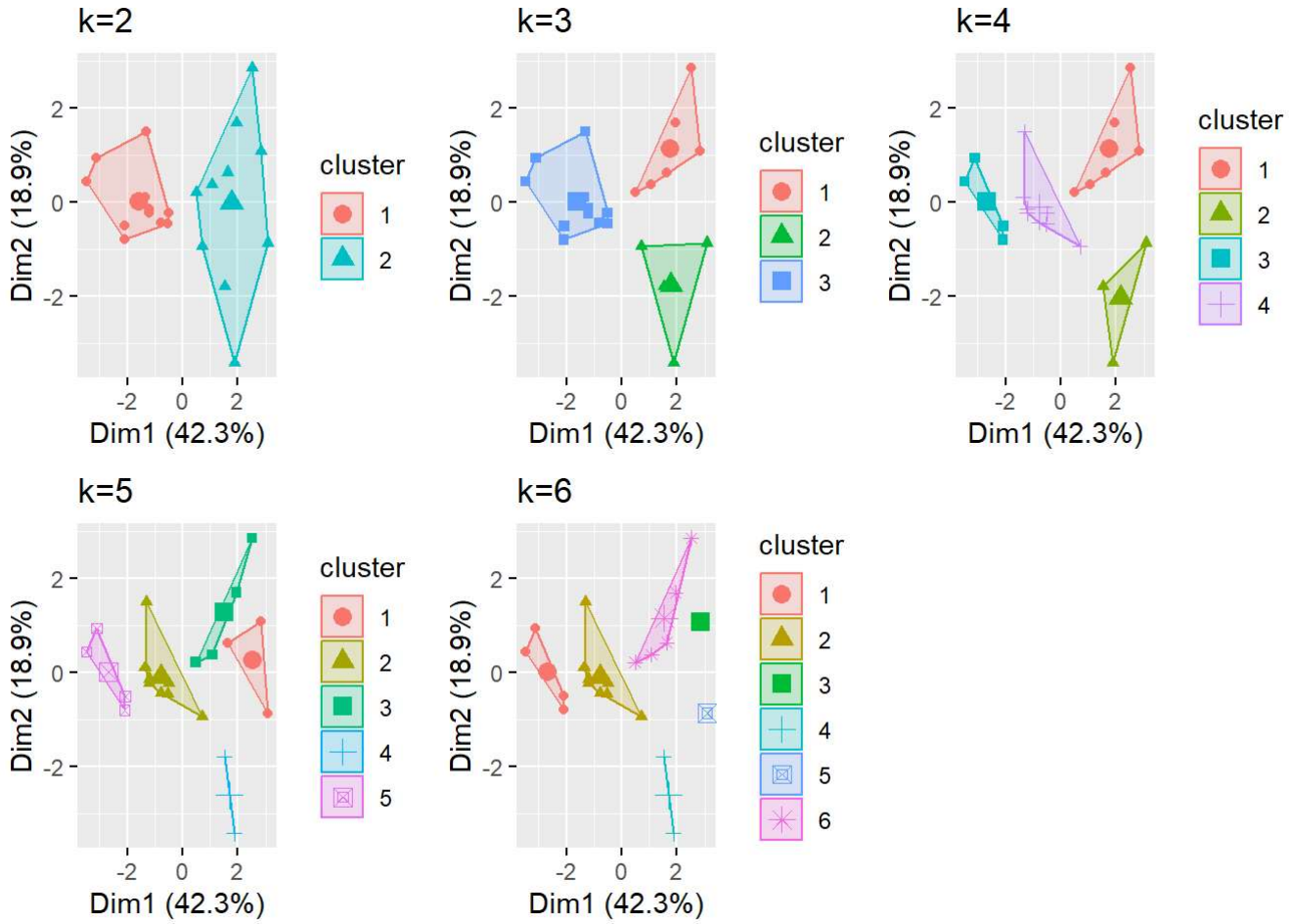



```
#According to the above plot the optimal number of cluster is 5.

# Perform K-means clustering with k= 4 to 7
k2<-kmeans(df1, centers=2, nstart=25)
k3<-kmeans(df1, centers=3, nstart=25)
k4<-kmeans(df1, centers=4, nstart=25)
k5<-kmeans(df1, centers=5, nstart=25)
k6<-kmeans(df1, centers=6, nstart=25)

p1<-fviz_cluster(k2,geom="point", data=df1)+ggtitle("k=2")
p2<-fviz_cluster(k3,geom="point", data=df1)+ggtitle("k=3")
p3<-fviz_cluster(k4,geom="point", data=df1)+ggtitle("k=4")
p4<-fviz_cluster(k5,geom="point", data=df1)+ggtitle("k=5")
p5<-fviz_cluster(k6,geom="point", data=df1)+ggtitle("k=6")

library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

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

```
grid.arrange(p1,p2,p3,p4,p5,nrow=2)
```




```
# Based on the above calculations Select K=5 as the optimal number of clusters based on t
he silhouette method.
```

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

```
# Plot K=5 cluster indicating with the index number
k5 <-kmeans(df1,centers = 5, nstart = 25)

# Add cluster labels to the original data set to make it clear to understand assignment o
f each company for the clusters.
Pharmaceuticals$cluster <- k5$cluster
print(Pharmaceuticals)
```

```
## # A tibble: 21 × 15
##    Symbol Name         Marke…¹  Beta PE_Ra…²   ROE   ROA Asset…³ Lever…⁴ Rev_G…⁵
##    <chr>  <chr>          <dbl> <dbl>   <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl>
##  1 ABT    Abbott Labo…   68.4   0.32    24.7  26.4  11.8     0.7    0.42    7.54
##  2 AGN    Allergan, I…    7.58  0.41    82.5  12.9   5.5     0.9    0.6     9.16
##  3 AHM    Amersham plc    6.3   0.46    20.7  14.9   7.8     0.9    0.27    7.05
##  4 AZN    AstraZeneca…   67.6   0.52    21.5  27.4  15.4     0.9    0      15
##  5 AVE    Aventis        47.2   0.32    20.1  21.8   7.5     0.6    0.34   26.8
##  6 BAY    Bayer AG       16.9   1.11    27.9   3.9   1.4     0.6    0      -3.17
##  7 BMY    Bristol-Mye…   51.3   0.5     13.9  34.8  15.1     0.9    0.57    2.7
##  8 CHTT   Chattem, Inc    0.41  0.85    26    24.1   4.3     0.6    3.51    6.38
##  9 ELN    Elan Corpor…    0.78  1.08     3.6  15.1   5.1     0.3    1.07   34.2
## 10 LLY    Eli Lilly a…   73.8   0.18    27.9  31    13.5     0.6    0.53    6.21
## # … with 11 more rows, 5 more variables: Net_Profit_Margin <dbl>,
## #   Median_Recommendation <chr>, Location <chr>, Exchange <chr>, cluster <int>,
## #   and abbreviated variable names ¹Market_Cap, ²PE_Ratio, ³Asset_Turnover,
## #   ⁴Leverage, ⁵Rev_Growth
```

```
# Calculate the mean values,centers and size of each numerical variable for each cluster
aggregate(Pharmaceuticals[,3:11], by=list(Pharmaceuticals$cluster),mean)
```

```
##   Group.1 Market_Cap    Beta PE_Ratio      ROE       ROA Asset_Turnover
## 1       1  55.810000 0.41375  20.2875 28.73750 12.687500         0.7375
## 2       2  31.910000 0.40500  69.5000 13.20000  5.600000         0.7500
## 3       3  13.100000 0.59750  17.6750 14.57500  6.200000         0.4250
## 4       4   6.636667 0.87000  24.6000 16.46667  4.166667         0.6000
## 5       5 157.017500 0.48000  22.2250 44.42500 17.700000         0.9500
##   Leverage Rev_Growth Net_Profit_Margin
## 1 0.371250   5.591250         19.350000
## 2 0.475000  12.080000          6.400000
## 3 0.635000  30.142500         15.650000
## 4 1.653333   5.733333          7.033333
## 5 0.220000  18.532500         19.575000
```

```
k5$centers
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 2 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 3 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 4 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 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 -0.14170336 -0.1168459      -1.416514761
## 3  0.06308085  1.5180158      -0.006893899
## 4  1.36644699 -0.6912914      -1.320000179
## 5 -0.46807818  0.4671788       0.591242521
```

```
k5$size
```

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

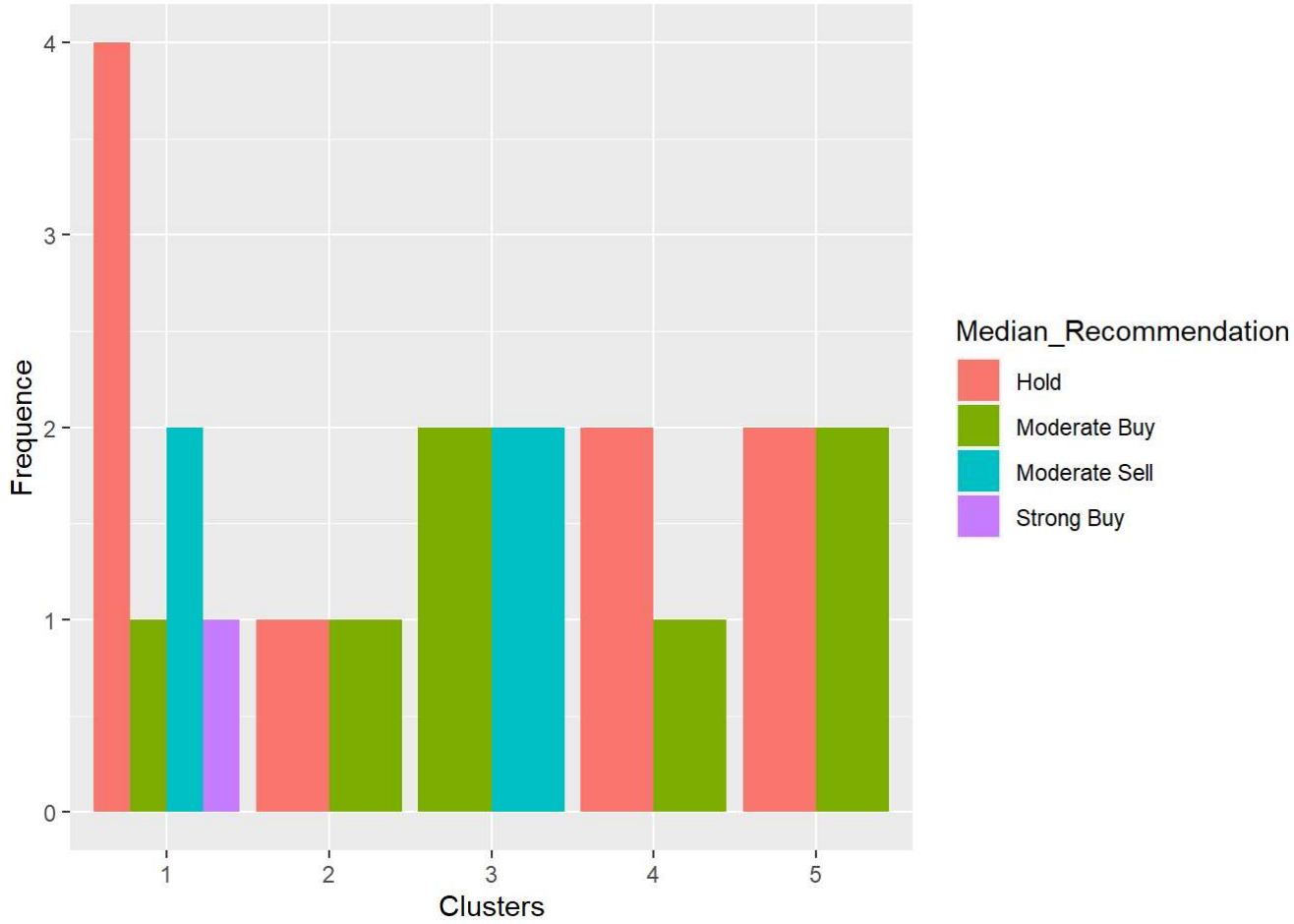
```
# Cluster 1 - This cluster contains four companies. 11. GlaxoSmithKline plc, 13. Johnson & Johnson, 15. Merck & Co., 17. Inc., Pfizer Inc
# According to the above calculations, this cluster has the highest market capital, highest Return On Equity (ROE), highest Return On Asset (ROA), highest value of Asset Turnover and highest Net Profit Margin.Therefore this cluster represents large and stable pharmaceutical companies with high profitability and low financial risk.

# Cluster 2 - This cluster contains three companies. 6.Bayer AG, 8.Chattem, Inc & 12.IVAX Corporation.
# According to the above calculations, this cluster has the lowest market capital,highest beta,lowest Return on Asset (ROA),but it has highest leverage. Comparatively this cluster has a lower net profit margin. According to that, this cluster contains small risky pharmaceutical companies with low profitability but high growth potential.

# Cluster 3 - This cluster contains eight companies.1.Abbott Laboratories, 3.Amersham plc, 4.AstraZeneca PLC, 7.Bristol-Myers Squibb Company, 10.Eli Lilly and Company, 16.Novartis AG, 19.Schering-Plough Corporation, 21.Wyeth
# According to the above calculations, this cluster has moderate values for all numerical variables and it has the lowest rev growth rate. Therefore this cluster can be identifies as group of pharmaceutical companies with moderate profitability and financial risk.

# Cluster 4 - This cluster contains four companies. 5.Aventis, 9.Elan Corporation, plc, 14.Medicis Pharmaceutical Corporation, 20.Medicis Pharmaceutical Corporation
# According to the above calculations, this cluster has the lowest market capital, lowest Profit Earning (PE) ratio, lowest asset turnover with highest revenue growth.Therefore this cluster can be identifies as high-growth, high efficiency and profitable market with low  financial stability.

# Cluster 5 - This cluster contains two companies. 2.Allergan, Inc., 18.Pharmacia Corporation
# According to the above calculations, this cluster has the lowest beta value, Highest Profit Earning (PE) ratio, lowest Return On  Investment Ratio (ROE) and lowest net profit margin. Therefore this cluster contains pharmaceutical companies with high growth potential but low profitability.
```

C) Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

```
# Median_Recommenation
pattern <- Pharmaceuticals %>% select(c(12,13,14)) %>% mutate(cluster = k5$cluster)
print(pattern)
```

```
## # A tibble: 21 × 4
##    Median_Recommendation Location Exchange cluster
##    <chr>                 <chr>    <chr>      <int>
##  1 Moderate Buy          US       NYSE           1
##  2 Moderate Buy          CANADA   NYSE           2
##  3 Strong Buy            UK       NYSE           1
##  4 Moderate Sell         UK       NYSE           1
##  5 Moderate Buy          FRANCE   NYSE           3
##  6 Hold                  GERMANY  NYSE           4
##  7 Moderate Sell         US       NYSE           1
##  8 Moderate Buy          US       NASDAQ         4
##  9 Moderate Sell         IRELAND  NYSE           3
## 10 Hold                  US       NYSE           1
## # … with 11 more rows
```

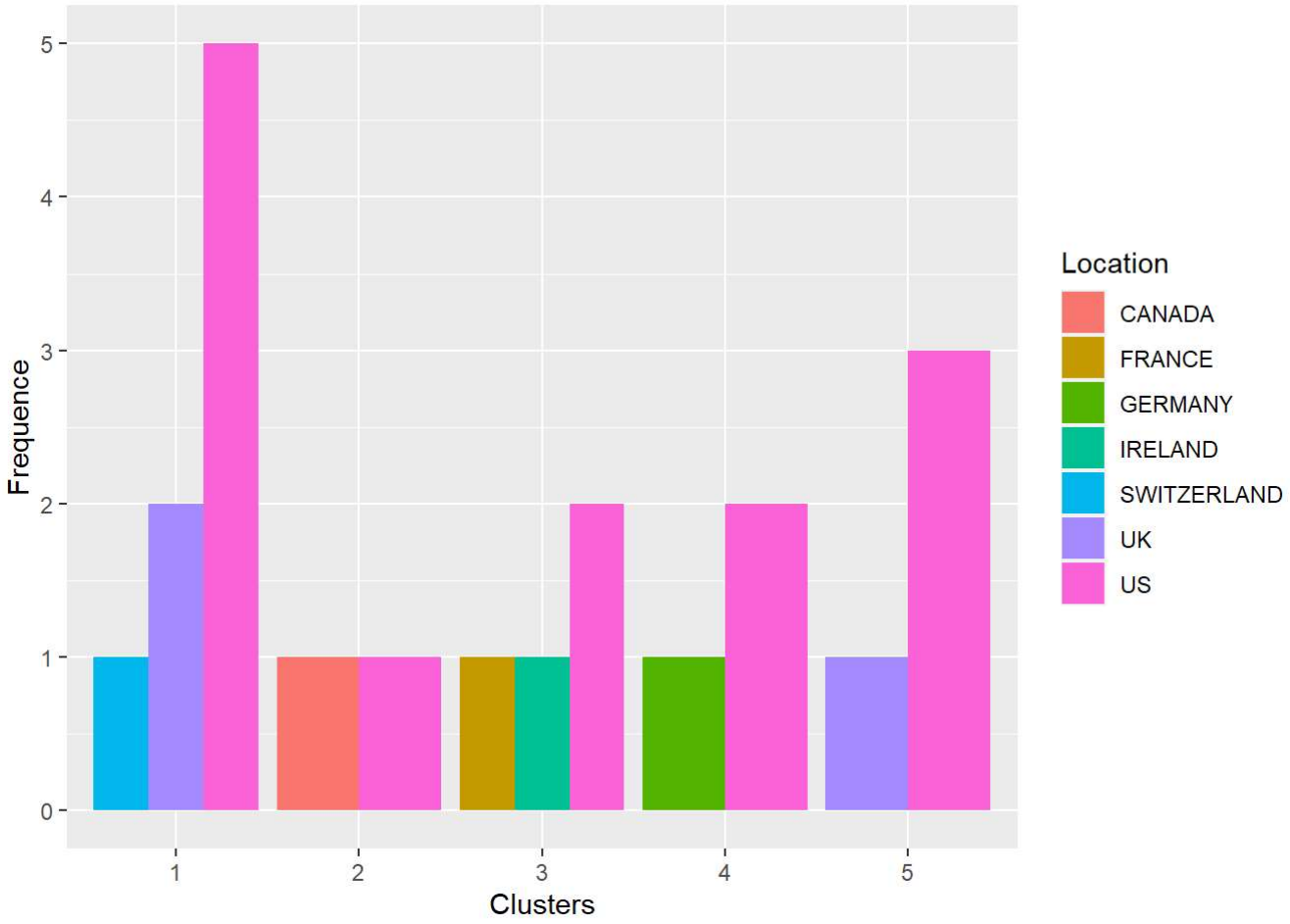
```
Median_Recommenation <- ggplot(pattern, mapping = aes(factor(cluster), fill=Median_Recommendation)) + geom_bar(position = 'dodge') + labs(x='Clusters', y='Frequence')
Median_Recommenation
```



```
#Answer: According to the above bar graph, for "Hold" is so high in the second cluster. M
oderate buy is equal for third an fifth clusters, as well as it is equal for all the othe
r clusters. Moderate buy can be seen in all clusters.The strong buy can be seen only in t
he second cluster.

# Location
pattern <- Pharmaceuticals %>% select(c(12,13,14)) %>% mutate(cluster = k5$cluster)
print(pattern)
```

```
## # A tibble: 21 × 4
##    Median_Recommendation Location Exchange cluster
##    <chr>                 <chr>    <chr>      <int>
##  1 Moderate Buy          US       NYSE           1
##  2 Moderate Buy          CANADA   NYSE           2
##  3 Strong Buy            UK       NYSE           1
##  4 Moderate Sell         UK       NYSE           1
##  5 Moderate Buy          FRANCE   NYSE           3
##  6 Hold                  GERMANY  NYSE           4
##  7 Moderate Sell         US       NYSE           1
##  8 Moderate Buy          US       NASDAQ         4
##  9 Moderate Sell         IRELAND  NYSE           3
## 10 Hold                  US       NYSE           1
## # … with 11 more rows
```

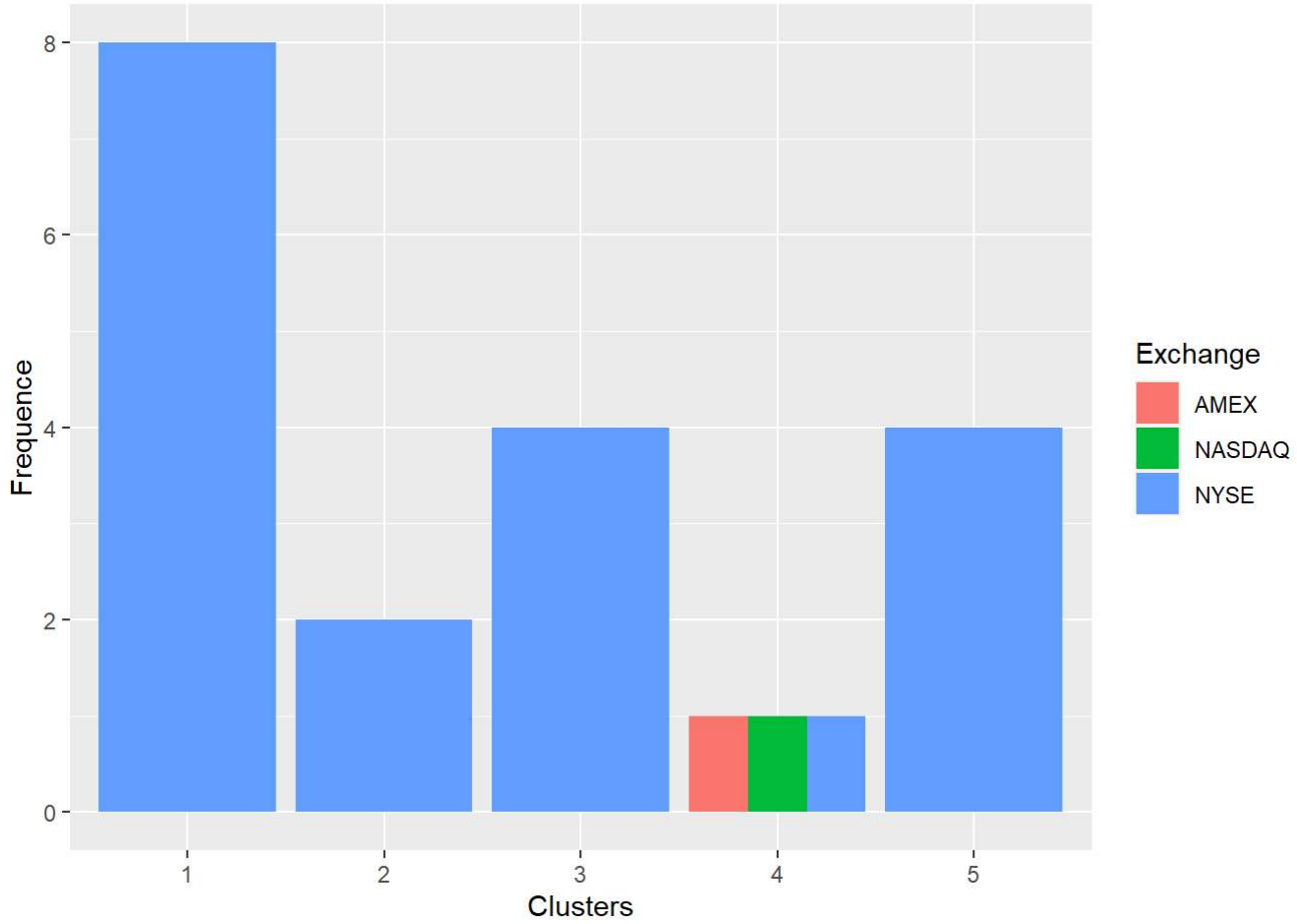
```
Location <- ggplot(pattern, mapping = aes(factor(cluster), fill=Location)) + geom_bar(pos
ition = 'dodge') + labs(x='Clusters', y='Frequence')
Location
```




```
# Answer: According to the above bar graph, The majority of second cluster comes from US, and Us represents all the clusters. Germany comes only in the first cluster. Switzerland comes only in the second cluster. France and Ireland can be seen only in the third cluster. Canada comes only in the fourth cluster. UK comes in the second and fifth cluster only.

# Exchange
pattern <- Pharmaceuticals %>% select(c(12,13,14)) %>% mutate(cluster = k5$cluster)
print(pattern)
```

```
## # A tibble: 21 × 4
##    Median_Recommendation Location Exchange cluster
##    <chr>                 <chr>    <chr>      <int>
##  1 Moderate Buy          US       NYSE           1
##  2 Moderate Buy          CANADA   NYSE           2
##  3 Strong Buy            UK       NYSE           1
##  4 Moderate Sell         UK       NYSE           1
##  5 Moderate Buy          FRANCE   NYSE           3
##  6 Hold                  GERMANY  NYSE           4
##  7 Moderate Sell         US       NYSE           1
##  8 Moderate Buy          US       NASDAQ         4
##  9 Moderate Sell         IRELAND  NYSE           3
## 10 Hold                  US       NYSE           1
## # … with 11 more rows
```

```
Exchange <- ggplot(pattern, mapping = aes(factor(cluster), fill=Exchange)) + geom_bar(pos
ition = 'dodge') + labs(x='Clusters', y='Frequence')
Exchange
```



```
# Answer:According to the above bar graph, All the exchanges are used in the first cluste
r. In the second cluster NYSE is so high.

# Summary:
# According to the above graphs for the second cluster, all the three variables (Median_R
ecommendation, Location & Exchange) have done a big influence. the variable "Location" ha
s done a big influence for all the clusters and some countries comes only in specific clu
ster.When consider about the "Exchange",except for the cluster one, only NYSE is influenc
ing for the others. Therefore the influence of "Exchange" for the clusters is minimal.
```

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

```
# Cluster 1: "Stable Profit Leaders" - This cluster represents large and stable pharmaceu
tical companies ,high market capital with high profitability and low financial risk.

# Cluster 2: "High Growth Risk Takers" - This cluster represents small and risky pharmace
utical companies with low profitability but high growth potential.

# Cluster 3: "Moderate Performers" - This cluster represents pharmaceutical companies wit
h moderate profitability and financial risk.

# Cluster 4: "Efficient but Risky" - This cluster represents pharmaceutical companies wit
h high efficiency and profitability, but low financial stability.

# Cluster 5: "Growth-Driven Under performers" - This cluster represents pharmaceutical co
mpanies with high growth potential but low profitability.
```