FML_Assignment.4

Jyothsna P - 811251679

2023-03-19

```
##install all required packages
##Load the required packages
#install.packages("factoextra")
#install.packages("flexclust")
#install.packages("cluster")
#install.packages("FactoMineR")
library(readr)
library(ISLR)
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(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2
## --

## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v stringr 1.5.0
## v tidyr   1.3.0     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tinytex)
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: lattice
## Loading required package: modeltools
## Loading required package: stats4
```

```
library(cluster)
library(FactoMineR)
library(ggcorrplot)
library(caret)
```

```
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
##Import the Pharmaceuticals data to R environment

Pharmaceuticals_Data<- read.csv("C:/Users/peddi/OneDrive/Desktop/Spring 2023/FML/Module 6/Pharmaceutica

head(Pharmaceuticals_Data)
```

```
##   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
```

```
#To display the summary of the Pharmaceuticals data.
summary(Pharmaceuticals_Data)
```

```
##     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  
##                    
##                    
##                    
```

```
#To find the type of data present in the Pharmaceuticals dataset
sapply(Pharmaceuticals_Data,class)
```

```
##                Symbol                  Name            Market_Cap 
##           "character"           "character"             "numeric" 
##                  Beta              PE_Ratio                   ROE 
##             "numeric"             "numeric"             "numeric" 
##                   ROA        Asset_Turnover              Leverage 
##             "numeric"             "numeric"             "numeric" 
##            Rev_Growth     Net_Profit_Margin Median_Recommendation 
##             "numeric"             "numeric"           "character" 
##              Location              Exchange 
##           "character"           "character"
```

```
#Finding out if there any missing or null values present in the dataset.

colMeans(is.na(Pharmaceuticals_Data))
```

```
##                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
```

There are no missing values in the dataset.

Question: 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.

```
#Using only the numerical variables (1 to 9) to cluster the 21 firms

Pharmaceuticals_1<-Pharmaceuticals_Data[c(1,3:11)]
row.names(Pharmaceuticals_1)<- Pharmaceuticals_1[,1]
Pharmaceuticals_1<- Pharmaceuticals_1[,-1]
head(Pharmaceuticals_1)
```

```
##     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
```

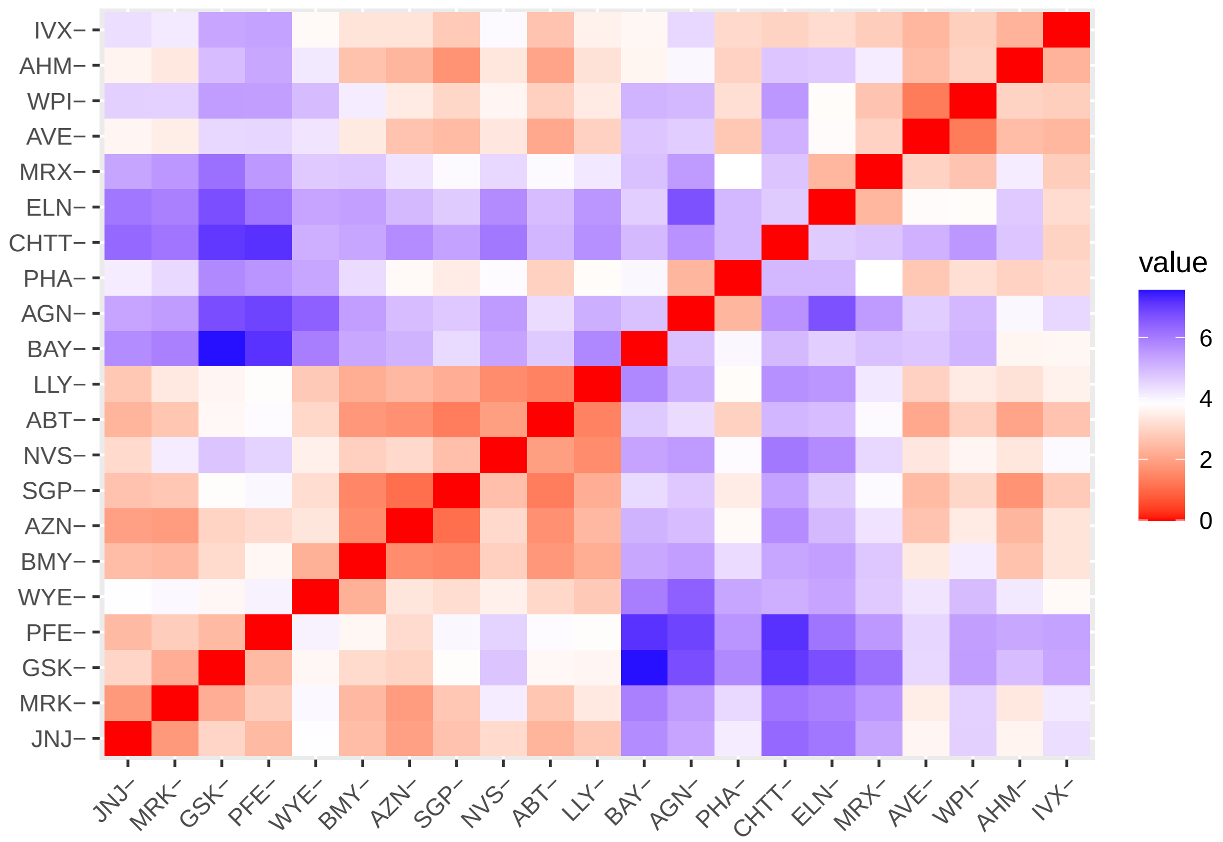
```
# Excluding the Charecter variables Name, Median_Recommendation, Exchange, Loacation.
colnames(Pharmaceuticals_1)
```

```
## [1] "Market_Cap"        "Beta"              "PE_Ratio"
## [4] "ROE"               "ROA"               "Asset_Turnover"
## [7] "Leverage"          "Rev_Growth"        "Net_Profit_Margin"
```

```
sapply(Pharmaceuticals_1,class)
```

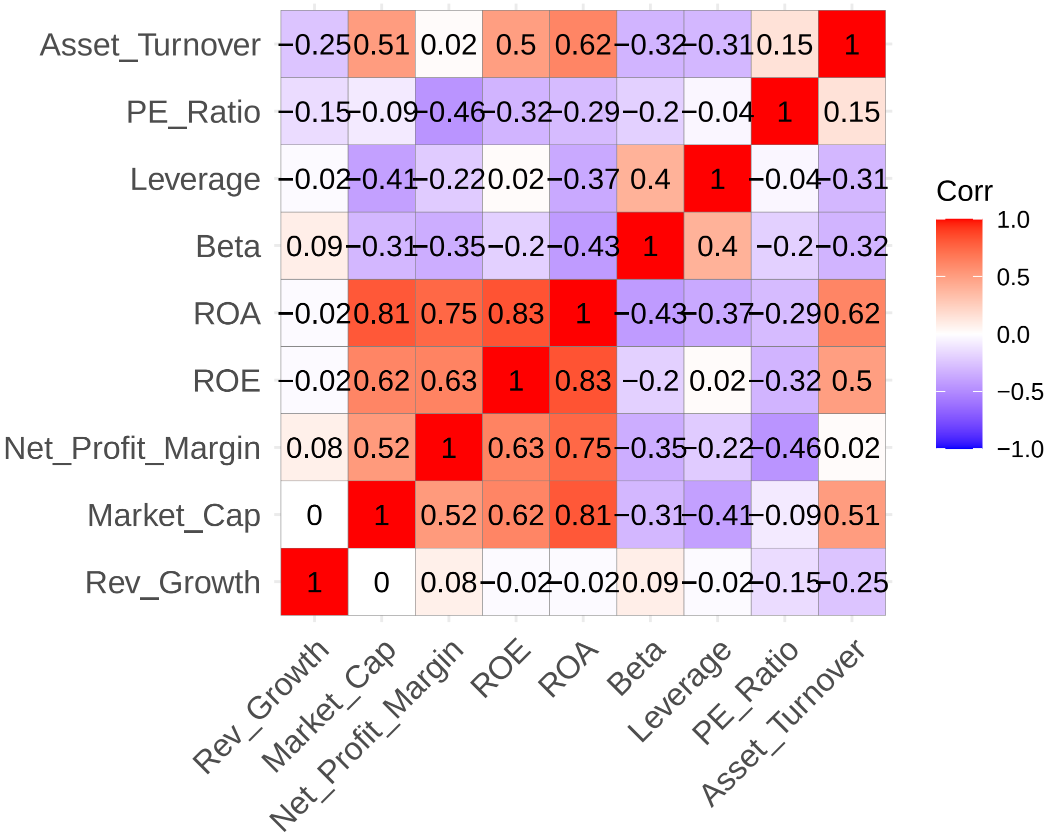
```
##        Market_Cap              Beta          PE_Ratio               ROE
##         "numeric"         "numeric"         "numeric"         "numeric"
##               ROA    Asset_Turnover          Leverage        Rev_Growth
##         "numeric"         "numeric"         "numeric"         "numeric"
## Net_Profit_Margin
##         "numeric"
```

```
#clustering the data by using the Euclidean distance formula and plotting the graph

##Euclidean distance = sqrt[(x2-x1)^2+(y2-y1)^2]

set.seed(110)

Pharamaceuticals_Norm<- scale(Pharmaceuticals_1)#Normalizing the numerical variables from the dataset
Pharmaceutical_Distance<- get_dist(Pharamaceuticals_Norm)#Uses Euclidean distance formula by default.
fviz_dist(Pharmaceutical_Distance, order = TRUE, show_labels = TRUE)
```



The heatmap color intensity shows the increase and decrease of distance between the observations in the dataset.

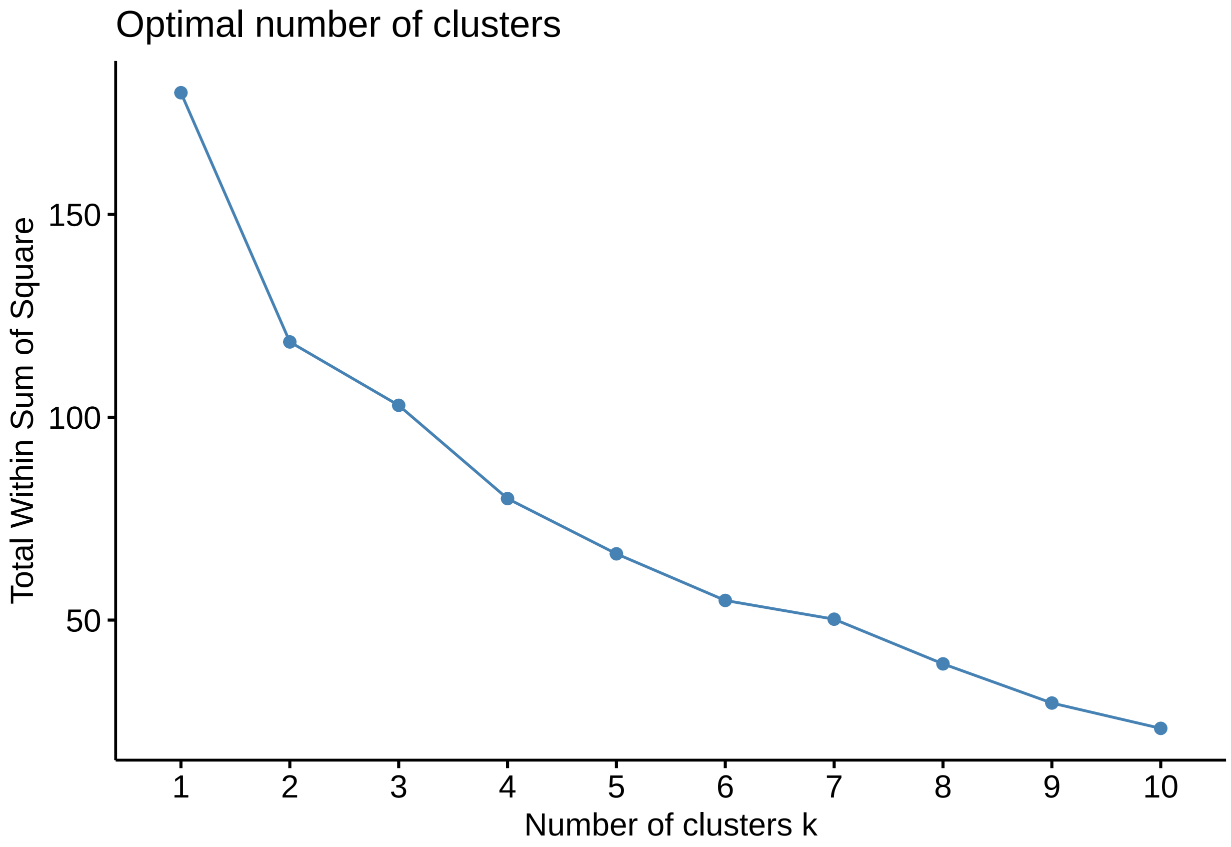
```
#plotting to find if there is any correlation between the variables.

correlation<- cor(Pharamaceuticals_Norm)
ggcorrplot(correlation,outline.color = "gray50",lab = TRUE,hc.order = TRUE, type = "full")
```



There is correlation between the variables,like ROA has the high positive correalation with Market__cap,Net__Profit__Margin,ROE,Asset__Turnover,and whereas ROE has positive correaltion with Market cap Net profit margin and ROA which means if the increase or decrease in one effects the other variables that are correlated.

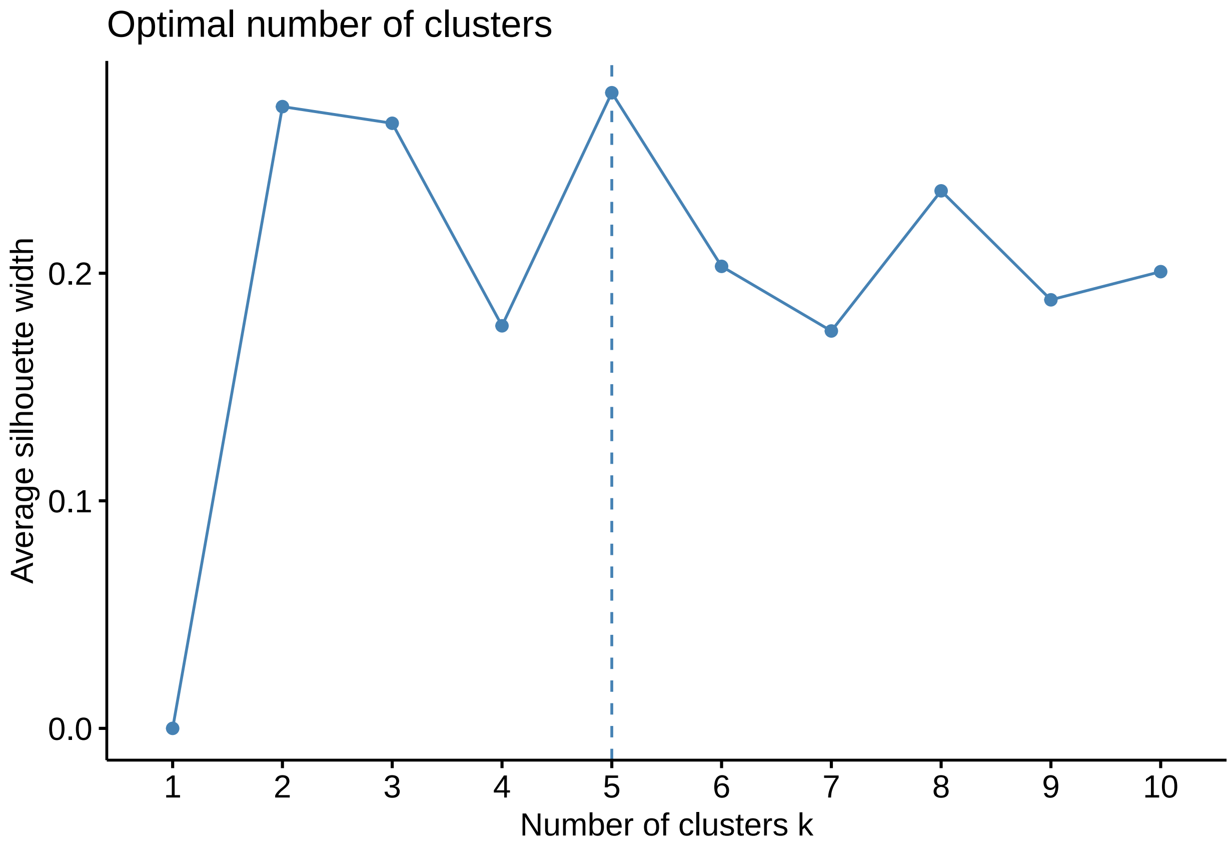
```
# finding the k value using Elbow Method
Elbow_Method<-fviz_nbclust(Pharamaceuticals_Norm, kmeans, method = "wss")
plot(Elbow_Method)
```

For finding the number of clusters,which means for finding of K value there are many clustering methods but,Elbow Method and Shilhouette Method are two main and widely used methods.



#The elbow method is showing the optimal value of K=2 or 6.

```
# Finding the k value using the Silhouette Method
Silhouette_Method<-fviz_nbclust(Pharamaceuticals_Norm, kmeans, method = "silhouette")
plot(Silhouette_Method)
```

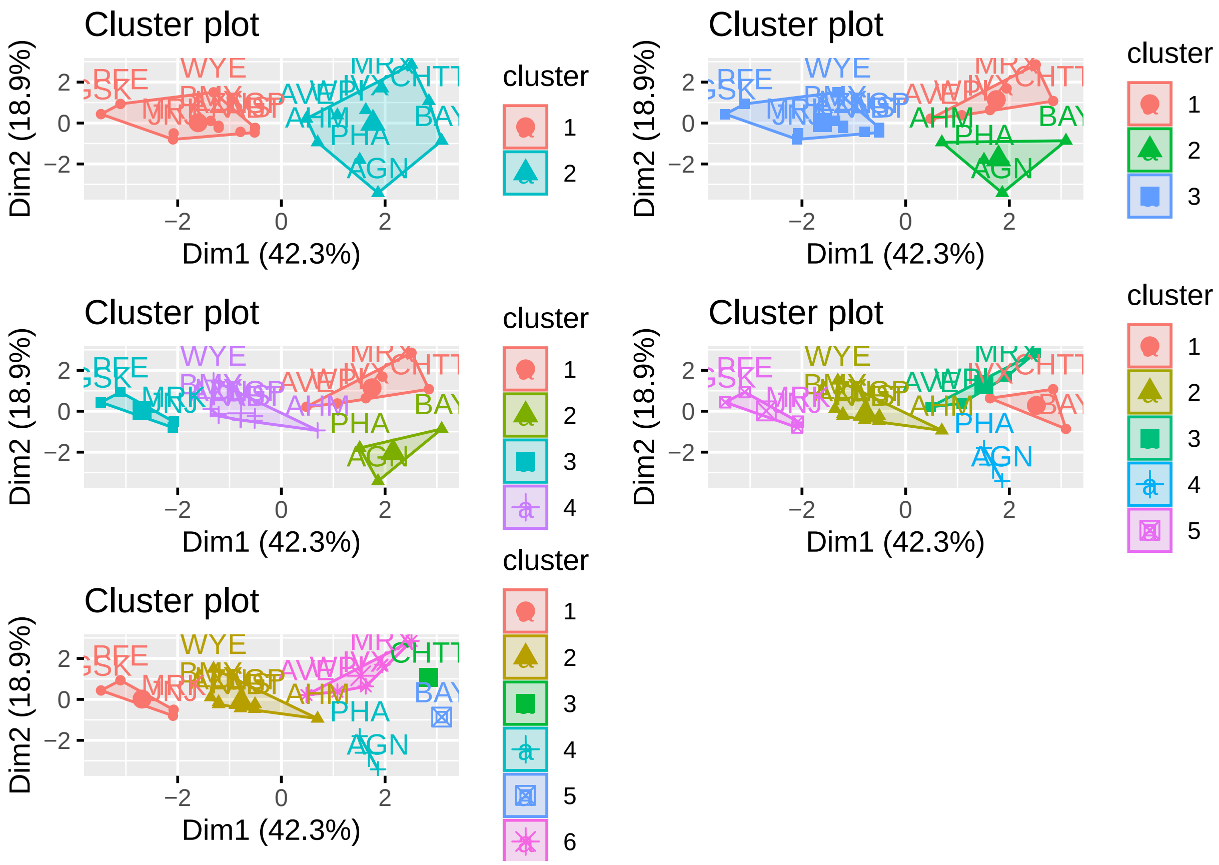


```
# Finding the values for all the K values from 2 to 6
#install.packages("gridExtra")
library(gridExtra)
```

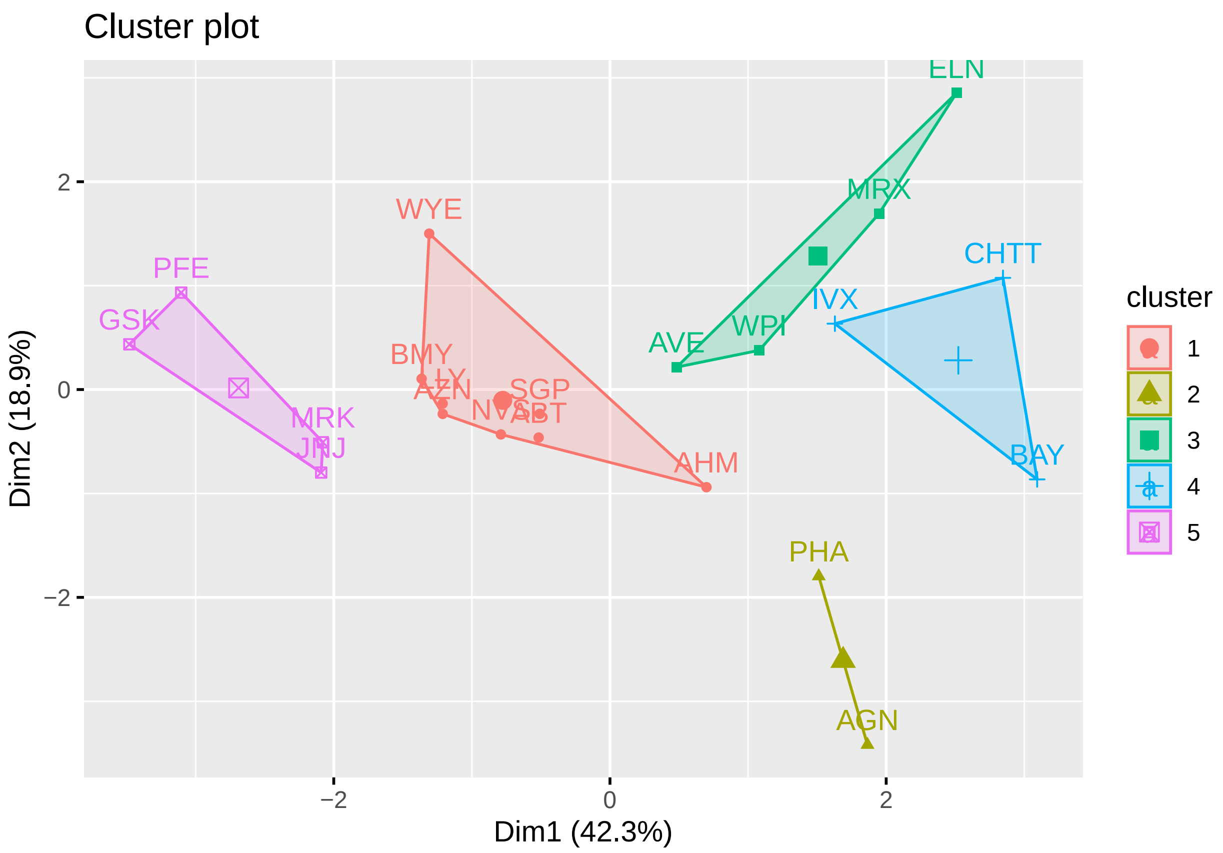
The Silhouette method is showing the optimal value of K=5

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

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

```
k2<- kmeans(Pharamaceuticals_Norm, centers = 2, nstart = 25)
k3<- kmeans(Pharamaceuticals_Norm, centers = 3, nstart = 25)
k4<- kmeans(Pharamaceuticals_Norm, centers = 4, nstart = 25)
k5<- kmeans(Pharamaceuticals_Norm, centers = 5, nstart = 25)
k6<- kmeans(Pharamaceuticals_Norm, centers = 6, nstart = 25)

plot.1=fviz_cluster(k2,data = Pharamaceuticals_Norm)
plot.2=fviz_cluster(k3,data = Pharamaceuticals_Norm)
plot.3=fviz_cluster(k4,data = Pharamaceuticals_Norm)
plot.4=fviz_cluster(k5,data = Pharamaceuticals_Norm)
plot.5=fviz_cluster(k6,data = Pharamaceuticals_Norm)

grid.arrange(plot.1,plot.2,plot.3,plot.4,plot.5)
```



```
#Plotting the clusters k=5 obtained from Silhouette
Silhouette_k5<- kmeans(Pharamaceuticals_Norm,centers = 5,nstart = 25)
Silhouette_plot<-fviz_cluster(Silhouette_k5,data=Pharamaceuticals_Norm)
plot(Silhouette_plot)
```



The total number of clusters formed are k=5.From Silhouette method and Elbow method approach, it is clear that k=5 has better silhouette width and the low withinness.

```
# finding the size of the cluster
Silhouette_k5$size
```

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

```
# Finding the withiness of cluster
Silhouette_k5$withinss
```

```
## [1] 21.879320  2.803505 12.791257 15.595925  9.284424
```

```
# Finding the cenetrs of the cluster
Silhouette_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
```

```
Silhouette_k5$cluster
```

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

```


# Finding the total withinss of the cluster
Silhouette_k5$tot.withinss
```

Above displayed is each observation is counties in the dataset belonging to the which Clusters.

```
## [1] 62.35443
```

```
# Finding the size,withinss and total withinss of the K=2 cluster from Elbow method
k2$size
```

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

```
k2$withinss
```

```
## [1] 43.30886 75.26049
```

```
k2$tot.withinss
```

```
## [1] 118.5693
```

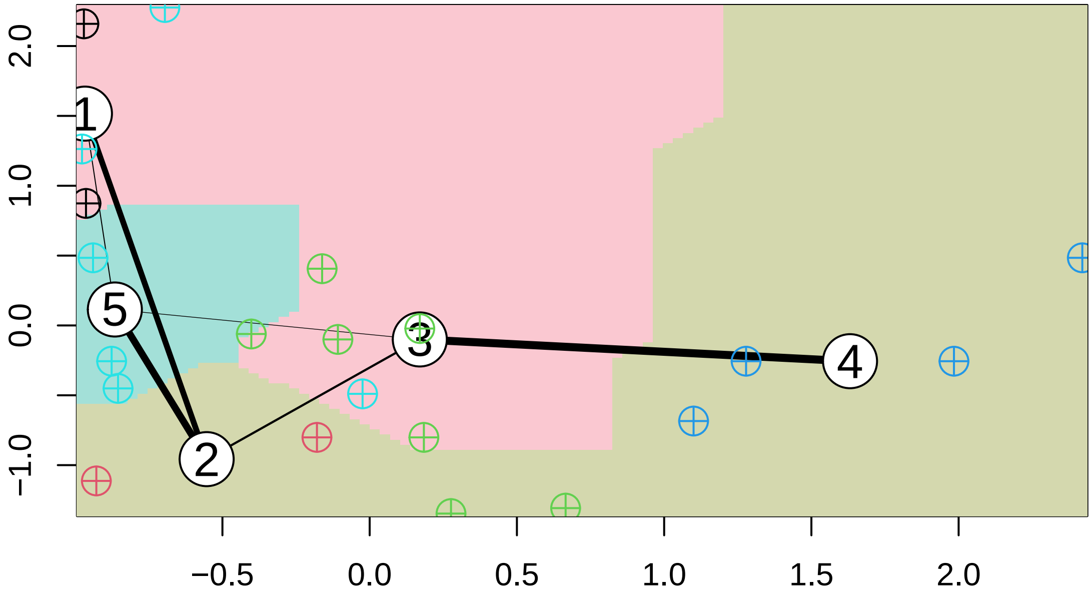
The total sum of squares within the Silhouette methos is 62.35 which is less than that of the value of total sum of squares within the Elbow method which is 118.56.Homogenous clusters is obtained when the sumof squares within the cluster is less. so I am choosing the silhouette method.where the optimal k value is K=5.

Here performing KCCA method of clustering using the Kmeans cluters k=5

```
set.seed(111)

pharmaceuticals_KCCA <- kcca(Pharamaceuticals_Norm, k = 5, kccaFamily("kmedians"))
pharmaceuticals_KCCA
```

```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = Pharamaceuticals_Norm, k = 5, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
## 1 2 3 4 5
## 2 2 7 4 6
```

```
clusters_index <- predict(pharmaceuticals_KCCA)
image(pharmaceuticals_KCCA)
points(Pharamaceuticals_Norm, col = clusters_index, pch = 10, cex = 2)
```



Both k-means and KCCA are clustering algorithms. KCCA is computationaly complex and requires more parameter tuning.where as k-means algorithm is a simple and widely used clustering algorithm that aims to partition a given set of observations into k clusters, where each observation belongs to the cluster with the nearest mean. so we will continue our analysis based on the K-means clustering algorithm.

Question: B

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

```
Silhouette_Group<- Silhouette_k5$cluster
Silhouette_Group<- as.data.frame(Silhouette_Group)
Silhouette_Pharmaceuticals= cbind(Pharmaceuticals_1,Silhouette_Group)
#Finding the mean of variables by clusters to understand the features of clusters.
Mean_of_Cluster= Silhouette_Pharmaceuticals %>% group_by(Silhouette_Group) %>% summarise_all("mean")
Mean_of_Cluster
```

```
## # A tibble: 5 x 10
##   Silhouette~1 Marke~2  Beta PE_Ra~3   ROE   ROA Asset~4 Lever~5 Rev_G~6 Net_P~7
##          <int>   <dbl> <dbl>   <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1            1   55.8  0.414    20.3  28.7 12.7    0.738   0.371    5.59   19.4
## 2            2   31.9  0.405    69.5  13.2  5.6    0.75    0.475   12.1     6.4
## 3            3   13.1  0.598    17.7  14.6  6.2    0.425   0.635   30.1    15.6
## 4            4    6.64 0.87     24.6  16.5  4.17   0.6     1.65     5.73    7.03
## 5            5  157.   0.48     22.2  44.4 17.7    0.95    0.22    18.5    19.6
## # ... with abbreviated variable names 1: Silhouette_Group, 2: Market_Cap,
## #   3: PE_Ratio, 4: Asset_Turnover, 5: Leverage, 6: Rev_Growth,
## #   7: Net_Profit_Margin
```

Following are the observations from each clusters based on the above output.

Cluster 1 The companies in this cluster has the lower Revenue growth than all the other companies.By examining the other variables, Net_profit_Margin of these companies are doing good and Leverage is also lower which means the companies in this cluster have less debts compared to the others in the clusters 2,3,and 4.

Cluster 2 The companies in this cluster have the least Net_Profit_Margin compares to all other companies in the other clusters and also it has the least Return on Equity(ROE) which indicates that companies in this cluster are very week in converting their equities into profits.In addition it has the highest Price Earning Ratio (PE_ratio) indicating that they may be overvalued and not gaining the profits.Beta value value is also low compared to others which means that these companies stocks are less volatile.

Cluster 3 The companies in this cluster have the highest Rvenue_Growth indicating that they are going in the right path for development but they are utilizing its assets to generate revenue as we can see the companies in this cluster have least Asset_Turnover ratio.However the PE_ratio is less compared to others which means these companies have the better earnings.

Cluster 4 The companies in this cluster have highest leverage which indicates that these companies are using higher debts to finance its operations.The Beta value is high indicating that the stock is more volatile.The companies in this cluster have lowest Market capital,ROA,Revenue Growth and Net Profit Margin which indicates these companies are facing the high financial and competition problems.

CLuster 5 The Companies in this cluster have the highest Market capital,ROA,ROE,Asset Turnover,Revenue growth and Net Profit Margin and have the less leverage compared to other companies in the all other clusters which indicates that these companies are performing well with very less debts. This cluster has the best performing companies among all the clusters.

Question: c

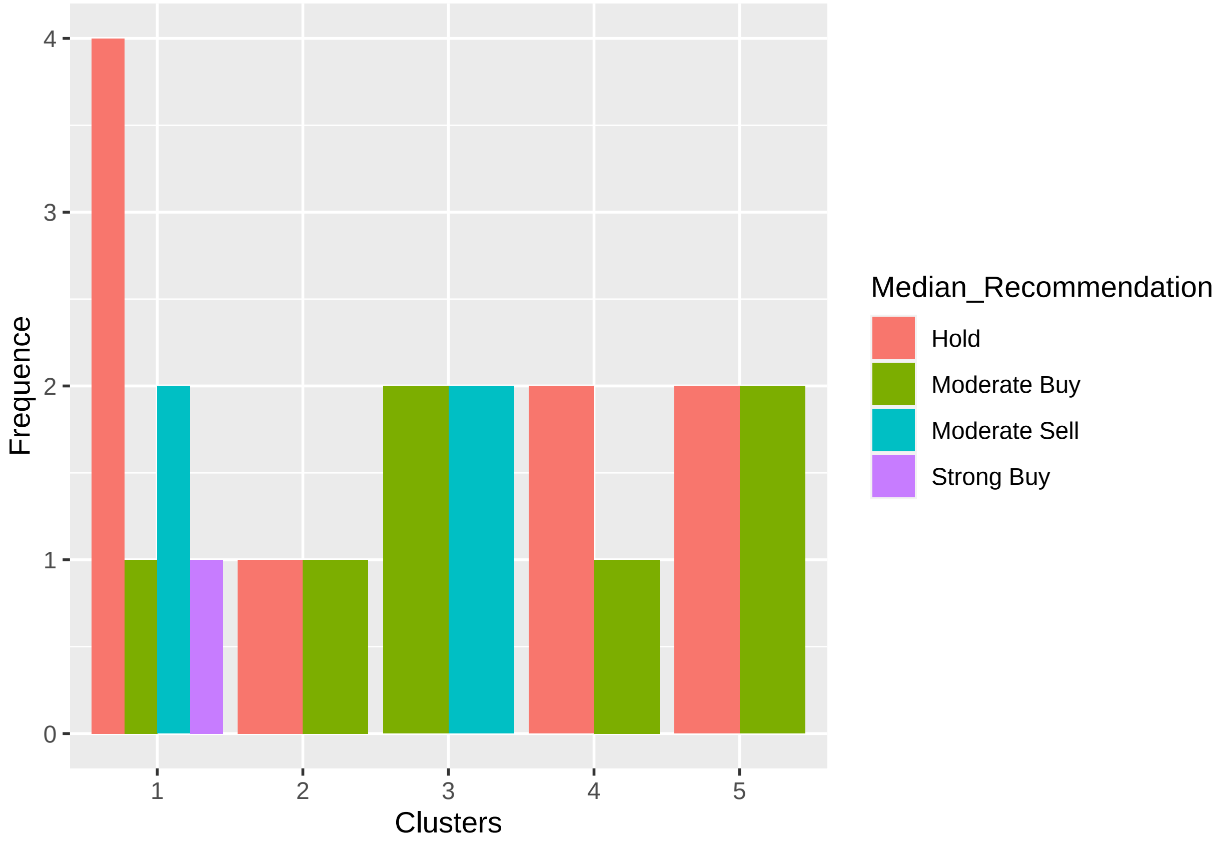
#Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

```
Pharmaceutical_Pattern<- Pharmaceuticals_Data %>% select(c(12,13,14)) %>% mutate(Cluster=Silhouette_k5$
print(Pharmaceutical_Pattern)
```

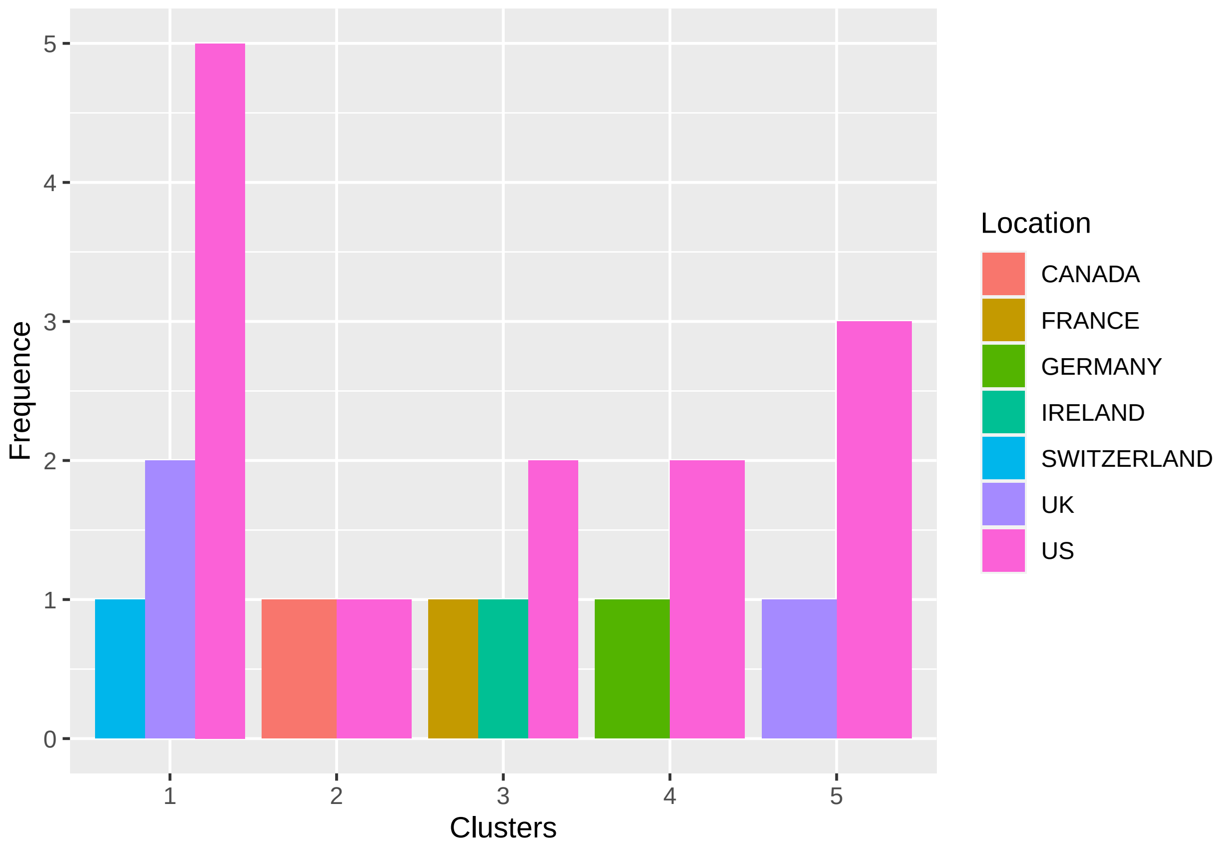
```
##    Median_Recommendation    Location Exchange Cluster
## 1           Moderate Buy          US     NYSE       1
## 2           Moderate Buy      CANADA     NYSE       2
## 3             Strong Buy          UK     NYSE       1
## 4          Moderate Sell          UK     NYSE       1
## 5           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
## 11                  Hold          UK     NYSE       5
## 12                  Hold          US     AMEX       4
## 13          Moderate Buy          US     NYSE       5
## 14          Moderate Buy          US     NYSE       3
## 15                  Hold          US     NYSE       5
## 16                  Hold SWITZERLAND     NYSE       1
## 17          Moderate Buy          US     NYSE       5
## 18                  Hold          US     NYSE       2
## 19                  Hold          US     NYSE       1
## 20         Moderate Sell          US     NYSE       3
## 21                  Hold          US     NYSE       1
```

#Identifying if there is any trends in the data and by utilizing the barcharts we will visualize the distribution of bussiness group by clusters.

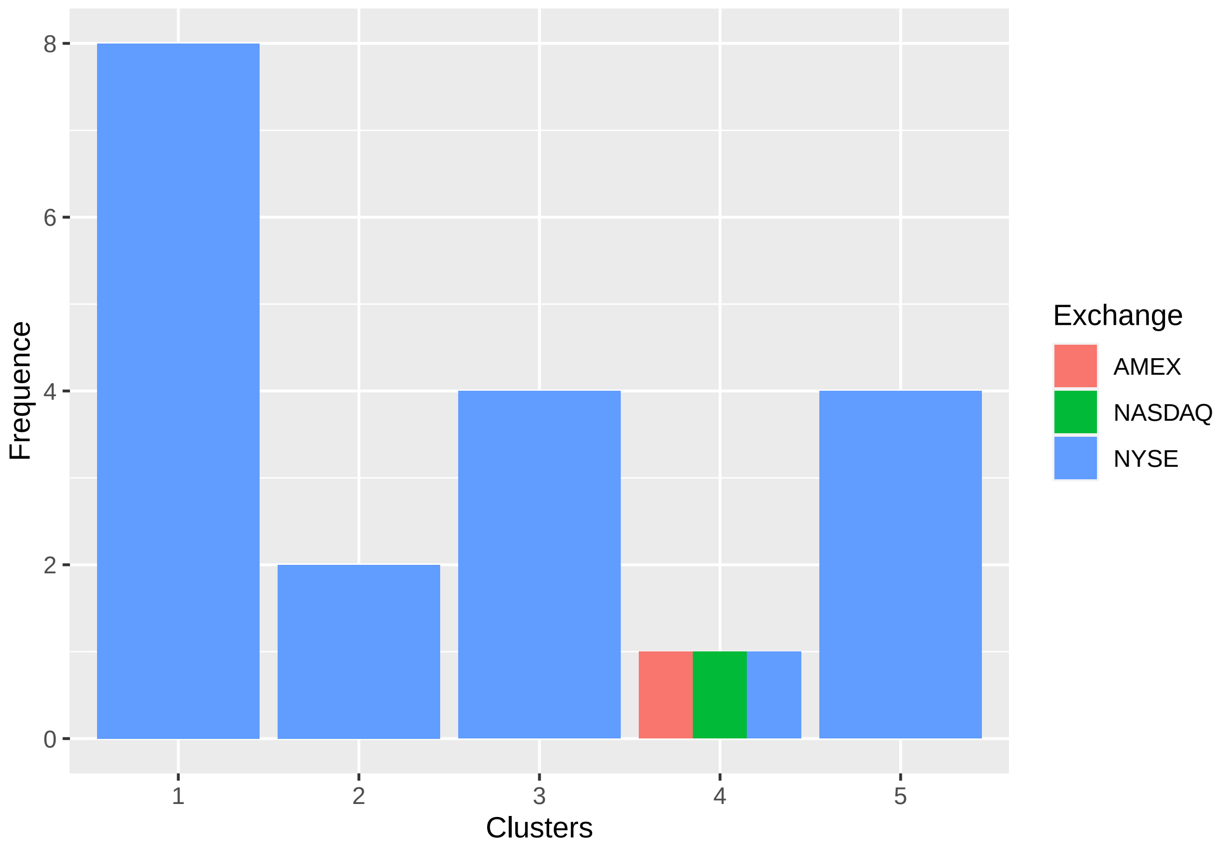
```
Median_Recom <- ggplot(Pharmaceutical_Pattern, mapping = aes(factor(Cluster), fill=Median_Recommendatio
  geom_bar(position = 'dodge') + labs(x='Clusters', y='Frequence')
plot(Median_Recom)
```



```
Location <- ggplot(Pharmaceutical_Pattern, mapping = aes(factor(Cluster), fill=Location)) + geom_bar(po
plot(Location)
```



```
Exchange <- ggplot(Pharmaceutical_Pattern, mapping = aes(factor(Cluster), fill=Exchange)) +
geom_bar(position = 'dodge') + labs(x='Clusters', y='Frequence')
plot(Exchange)
```



##From above plots there are no any patterns seen in the clusters with respect to the variables those are not used to form the clusters, but there are some observations that can be made.

Cluster-1 Cluster 1 grouped all the New York Stock Exchange(NYSE) listed companies predominantly in North America(United States and Canada). Based on the recommendations, most of them are either hold or moderate sell, which implies they might have low opportunities for growth.

Cluster-2 All companies in cluster 2 are NYSE listed and located in North America(United States and Canada). This is a safe investment cluster with just two companies, with one company's median recommendation as a moderate buy and the other as a hold.

Cluster-3 All companies in the cluster are NYSE listed and located in the United States or European countries(France and Ireland). Median recommendations in the cluster reflect that the investments in these companies are for growth investments with a possible balanced risk involved, as the ratio of moderate buy and moderate sell is the same.

Cluster-4 All the exchanges available in the United States listed companies are in cluster 4. With the recommendations hold and moderate buy, this could be considered a low-risk investment. Companies are located in Germany and the United States.

Cluster-5 All companies are NYSE listed and from the United States and the UK. Companies' median recommendations are with an equal proportion of hold and moderate buy, implying that the companies are low risk.

Question:D

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

#Naming the clusters based on the variables in the dataset.

Cluster 1 - "Stable Growth-profitable Companies".

Based on the stable and normal Financial Metrics in the variables indicates that the comapanies in this cluster are stable and they have less leverage and good Net profit margin which means they are performing effectively.

Cluster 2 - "Over Valued-Least Profitability Companies".

The cluster of companies in question has the lowest return on equity (ROE) among all clusters, suggesting weakness in converting equity into profits. Additionally, the cluster exhibits the highest price-earnings ratio (PE_ratio), which may imply that the companies are overvalued and not generating expected profits.

cluster 3 - "Better Earning-Low Risk Companies".

The companies in this cluster demonstrates the highest revenue growth, indicating that they are progressing well in terms of development.The companies in this cluster are utilizing their assets less efficiently, as evidenced by their low asset turnover ratio. Despite this, the companies in the cluster exhibit a lower price-earnings ratio (PE_ratio), which may suggest that they have better earnings.

cluster 4 - "High Debt-Risky Companies".

In this cluster the companies have the highest leverage, indicating that they are utilizing more debt and high beta value indicating that their stocks are more volatile.These companies have lowest market capitalization, return on assets (ROA), Revenue_growth, and Net_profit_margin suggesting that they have a higher risk of financial challenges.

cluster 5 - "High Performing and Financially strong Companies".

This cluster have the comapnies that demonstrates exceptional performance, as it has the highest market capitalization, return on assets (ROA), return on equity (ROE), asset turnover, revenue growth, and net profit margin.This cluster exhibit less leverage, indicating that they are performing well while maintaining low levels of debt and are financially strong.