# FML_Assignment_4

Sri Naga Dattu Gummadi

2023-11-10

---

**SUMMARY**

## Interpreting the clusters with respect to the numerical variables used in forming the clusters.

Cluster 1 - 2, 18 (lowest Beta,lowest Asset_Turnover, Highest PE Ratio).

Cluster 2 - 1,3,4,7,10,16,19,21 ( lowest Market_Cap,lowest Beta,lowest PE_Ratio,highest Leverage,highest Rev_Growth).

Cluster 3 - 5, 9, 14, 20 (lowest PE_Ratio,highest ROE,lowest ROA,lowest Net_Profit_Margin, highest Rev_Growth).

Cluster 4 - 11, 13, 15, 17 (Highest Market_Cap,ROE, ROA,Asset_Turnover Ratio and lowest Beta/PE Ratio).

Cluster 5 - 6, 8, 12 (lowest Rev_Growth,highest Beta and levearge,lowest Net_Profit_Margin).

## Ques) Is there a pattern in the clusters with respect to the numerical variables? (10 to 12) variables? (those n not utilized in the cluster formation).

As per the graphs below, the interpretation is as follows:

Cluster 1: has the same hold and moderate buy medians and is spread out across the US, UK, and listed in NYSE.

Cluster 2 : In this cluster, which displays distinct Hold, Moderate Buy, a little increased Moderate Sell, and Strong Buy medians, the Hold median is the highest. They are from the US, the UK, and Switzerland and are traded on the NYSE.

Cluster 3: Exclusively listed on the NYSE, evenly distributed across the US and Canada, with medians of Moderate Hold and Moderate Buy.

Cluster 4: listed on the NYSE, has distinct counts for France, Ireland, and the US, and has medians for buy and sell orders that are equally Moderate.

Cluster 5 : Listed on AMEX, NASDAQ, and NYSE stock exchanges, all have an equal distribution of companies, but there is a clear Hold and Moderate Buy median as well as a different count between the US and Germany.

# Question 3.

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

Based on certain criteria, preferably financial measures such as performance, potential for growth, or risk factors, Investors and analysts can use such clusters to make informed decisions about their investment strategies. We can name each cluster appropriately as follows:

Cluster 1 :- HOLD-BUY CLUSTER or Balanced Investment Cluster.

Cluster 2 :- HIGH HOLD CLUSTER or Robust Holding Cluster.

Cluster 3 :- HOLD-BUY CLUSTER or Balanced Investment Cluster.

Cluster 4 :- BUY-SELL CLUSTER or Dynamic Portfolio Cluster.

Cluster 5 :- HOLD CLUSTER or Stable Investment Cluster.

---

# Imporing the Pharmaceuticals dataset

```r
library(readr)
pharmacts <- read.csv("/Users/srinagadattugummadi/Downloads/Pharmaceuticals.csv")
summary(pharmacts)
```

```
##     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  
```

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

```r
str(pharmacts)
```

```
## '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" ...
```

# Loading the packages

```r
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
```

```r
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```r
library(factoextra)
```

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

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(cluster)
library(gridExtra)
```

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

---

## Question 1.

### Cluster the 21 companies using only the numerical variables (1–9). Justify the numerous decisions taken throughout the cluster analysis, including the weights assigned to various variables, the particular clustering algorithm(s) utilized, the number of clusters created, and so on.

### Removing the dataset's null values and choosing the monetary variables.

```r
colSums(is.na(pharmacts))
```

```
##          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 <- pharmacts[,1]
pharmac_cl <- pharmacts[,3:11]
head(pharmac_cl)
```

```
##   Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## 1      68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54
## 2       7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16
## 3       6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05
## 4      67.63 0.52     21.5 27.4 15.4            0.9     0.00      15.00
## 5      47.16 0.32     20.1 21.8  7.5            0.6     0.34      26.81
## 6      16.90 1.11     27.9  3.9  1.4            0.6     0.00      -3.17
##   Net_Profit_Margin
## 1              16.1
## 2               5.5
## 3              11.2
## 4              18.0
## 5              12.9
## 6               2.6
```

## scaling and normalization of the dataset.

Normalization of the numerical variables is essential to guarantee that each variable contributes proportionately to the clustering process. Normalizing these variables helps avoid one variable from predominating the clustering based only on its magnitude because they may have different units or scales.

```
pharmacts_scale <- scale(pharmac_cl)
head(pharmacts_scale)
```

```
##      Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## [1,]  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121  -5.121077e-16
## [2,] -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871   9.225312e-01
## [3,] -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700   9.225312e-01
## [4,]  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259   9.225312e-01
## [5,] -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461  -4.612656e-01
## [6,] -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612  -4.612656e-01
##        Leverage Rev_Growth Net_Profit_Margin
## [1,] -0.2120979 -0.5277675        0.06168225
## [2,]  0.0182843 -0.3811391       -1.55366706
## [3,] -0.4040831 -0.5721181       -0.68503583
## [4,] -0.7496565  0.1474473        0.35122600
## [5,] -0.3144900  1.2163867       -0.42597037
## [6,] -0.7496565 -1.4971443       -1.99560225
```

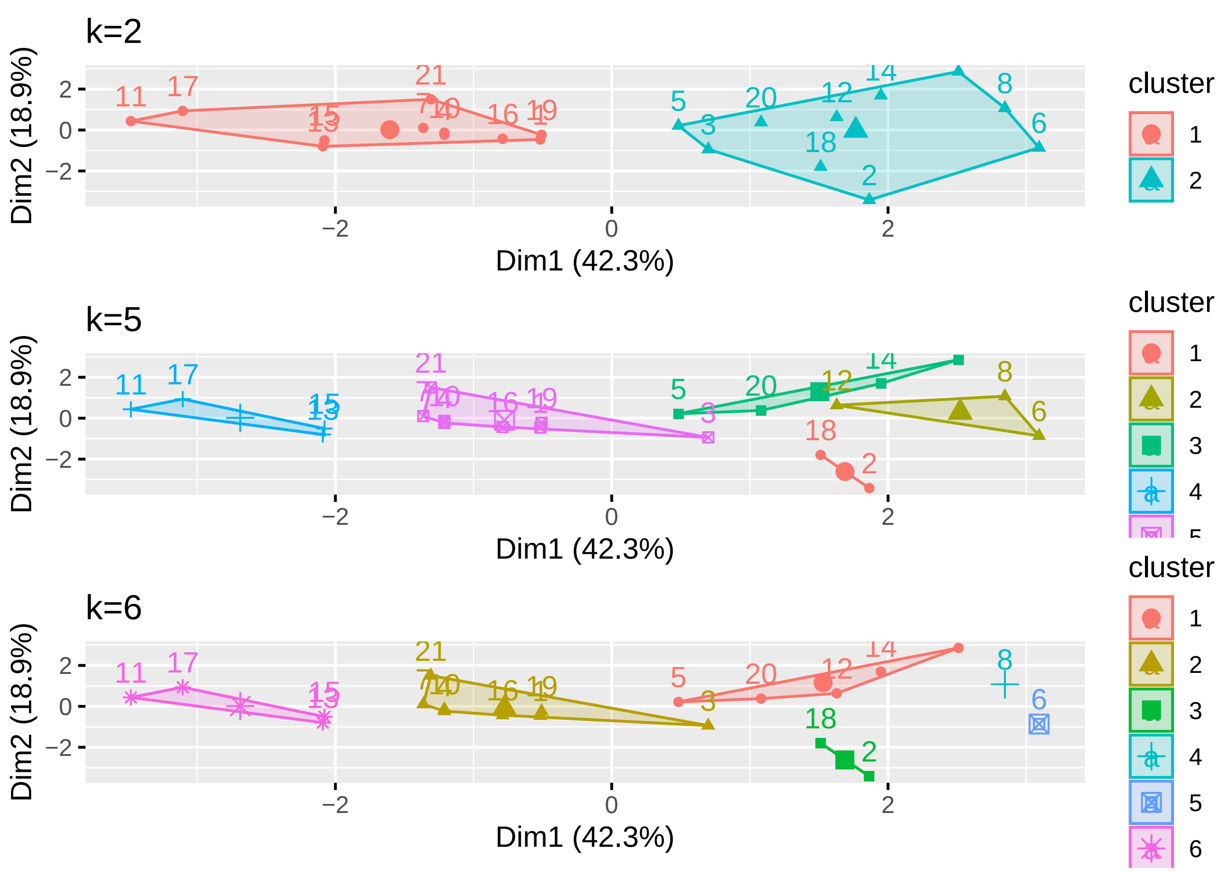
```
cl_data <- as.data.frame(scale(pharmac_cl))
```

## Calculating K-means clustering for various centers, use a variety of K values, and comparing the results.

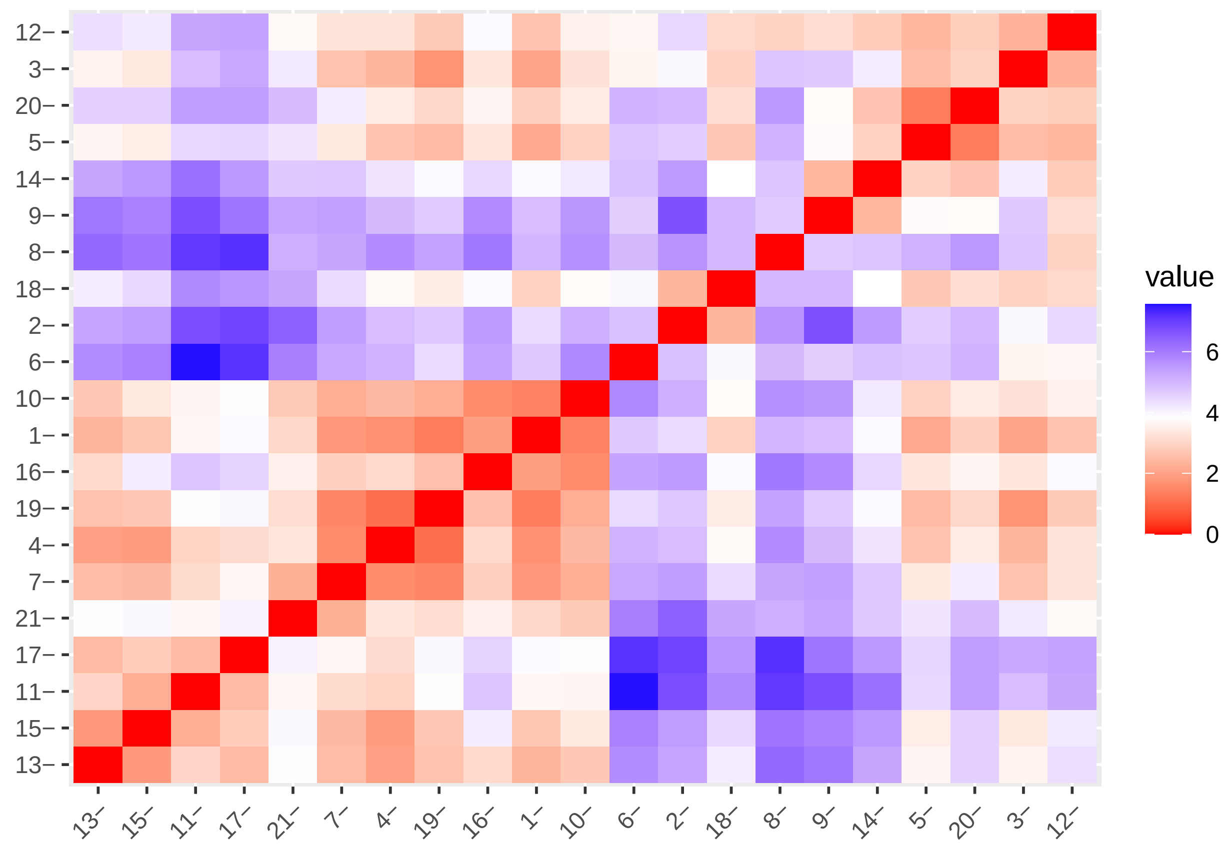
Here,Selecting and prefering K-means over DBSCAN because it's frequently used in exploratory data analysis to find patterns and groups in the data, and because K-means clustering can reveal information about the financial profiles of pharmaceutical companies. DBSCAN is useful for datasets with dense areas and can help with investment analysis and strategic decision-making by revealing groups of companies with comparable financial features. It is also easily interpreted.

```
kmeans_1cl <- kmeans(pharmacts_scale, centers = 2, nstart = 30)
kmeans_2cl<- kmeans(pharmacts_scale, centers = 5, nstart = 30)
kmeans_3cl<- kmeans(pharmacts_scale, centers = 6, nstart = 30)

Plot_1r<-fviz_cluster(kmeans_1cl, data = pharmacts_scale)+ggtitle("k=2")
Plot_2r<-fviz_cluster(kmeans_2cl, data = pharmacts_scale)+ggtitle("k=5")
Plot_3r<-fviz_cluster(kmeans_3cl, data = pharmacts_scale)+ggtitle("k=6")
grid.arrange(Plot_1r,Plot_2r,Plot_3r, nrow = 3)
```




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

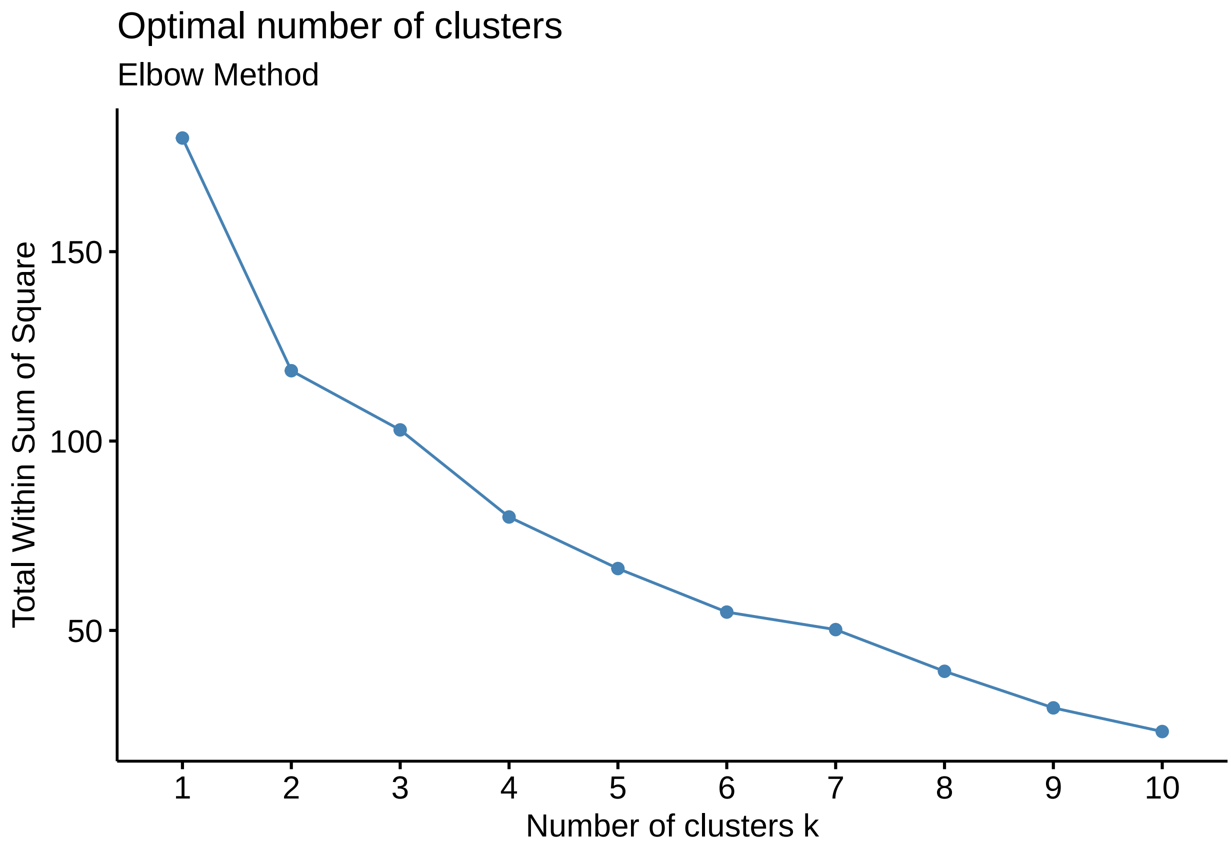



## Estimating the number of clusters.

## Elbow Method is used in scaling the data to determine the K value.

The elbow method is used to determine the optimal number of clusters (k) in a k-means clustering analysis.

```
fviz_nbclust(cl_data, FUNcluster = kmeans, method = "wss") + labs(subtitle = "Elbow Method")
```

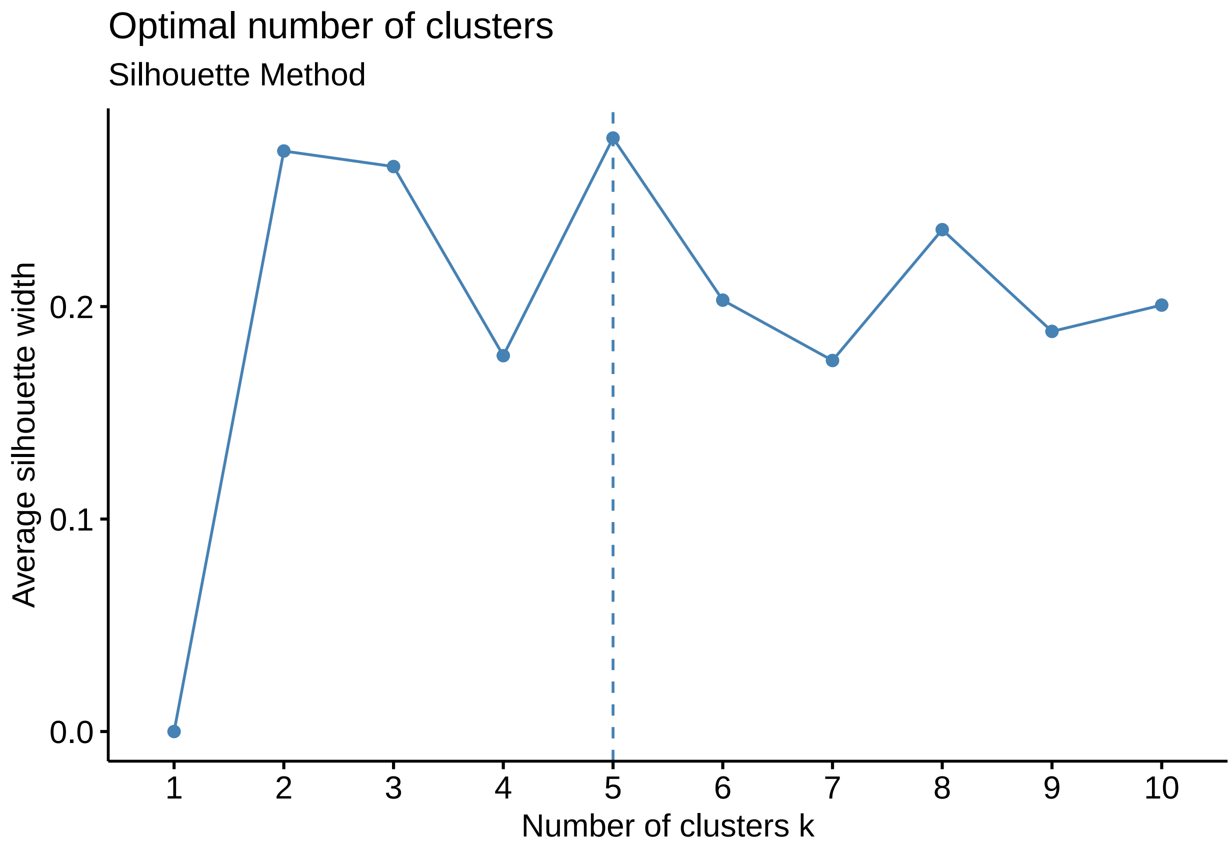



## It is evident that the output above displays that around 5 - 6 is the ideal value for k (slope stops being as steep)

## The Silhouette Method is used in scaling the data to determine the number of clusters.

Reason : The silhouette analysis calculates an object's degree of similarity to its own cluster in relation to other clusters. For various values of k, it offers a graphical depiction of the quality of clusters.

```
fviz_nbclust(cl_data,FUNcluster = kmeans,method = "silhouette")+labs(subtitle="Silhouette Method")
```



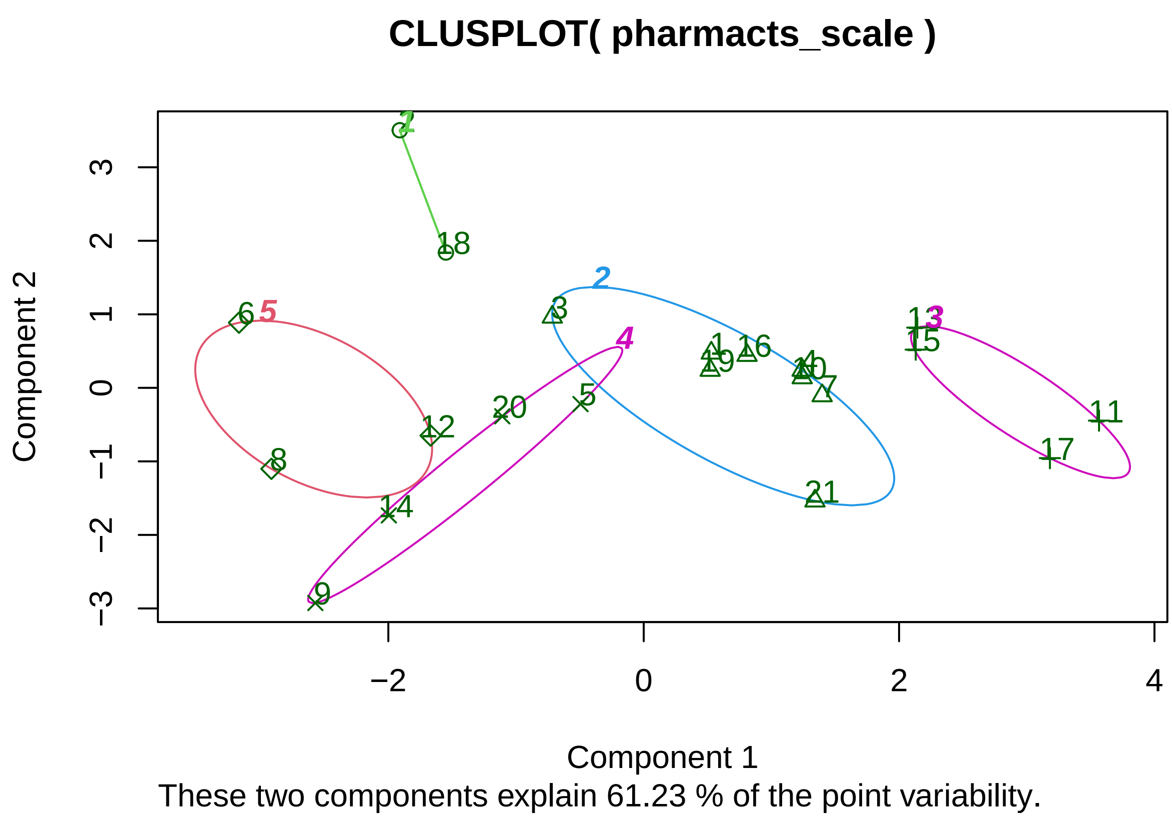

### Final analysis and Extracting results using 5 clusters and Visualising the results.

```
set.seed(555)
final_Cl<- kmeans(pharmacts_scale, 5, nstart = 25)
print(final_Cl)
```

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

```
##
## Clustering vector:
##  [1] 2 1 2 2 4 5 2 5 4 2 3 5 3 4 3 2 3 1 2 4 2
##
## Within cluster sum of squares by cluster:
## [1]  2.803505 21.879320  9.284424 12.791257 15.595925
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
clusplot(pharmacts_scale,final_Cl$cluster, color = TRUE, labels = 2,lines = 0)
```



**Question 2.**

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

**Cluster 1 - 2, 18 (lowest Beta,lowest Asset_Turnover, Highest PE Ratio)**

**Cluster 2 - 1,3,4,7,10,16,19,21 ( lowest Market_Cap,lowest Beta,lowest PE_Ratio,highest Leverage,highest Rev_Growth.)**

**Cluster 3 - 5, 9, 14, 20 (lowest PE_Ratio,highest ROE,lowest ROA,lowest Net_Profit_Margin, highest Rev_Growth)**

**Cluster 4 - 11, 13, 15, 17 (Highest Market_Cap,ROE, ROA,Asset_Turnover Ratio and lowest Beta/PE Ratio)**

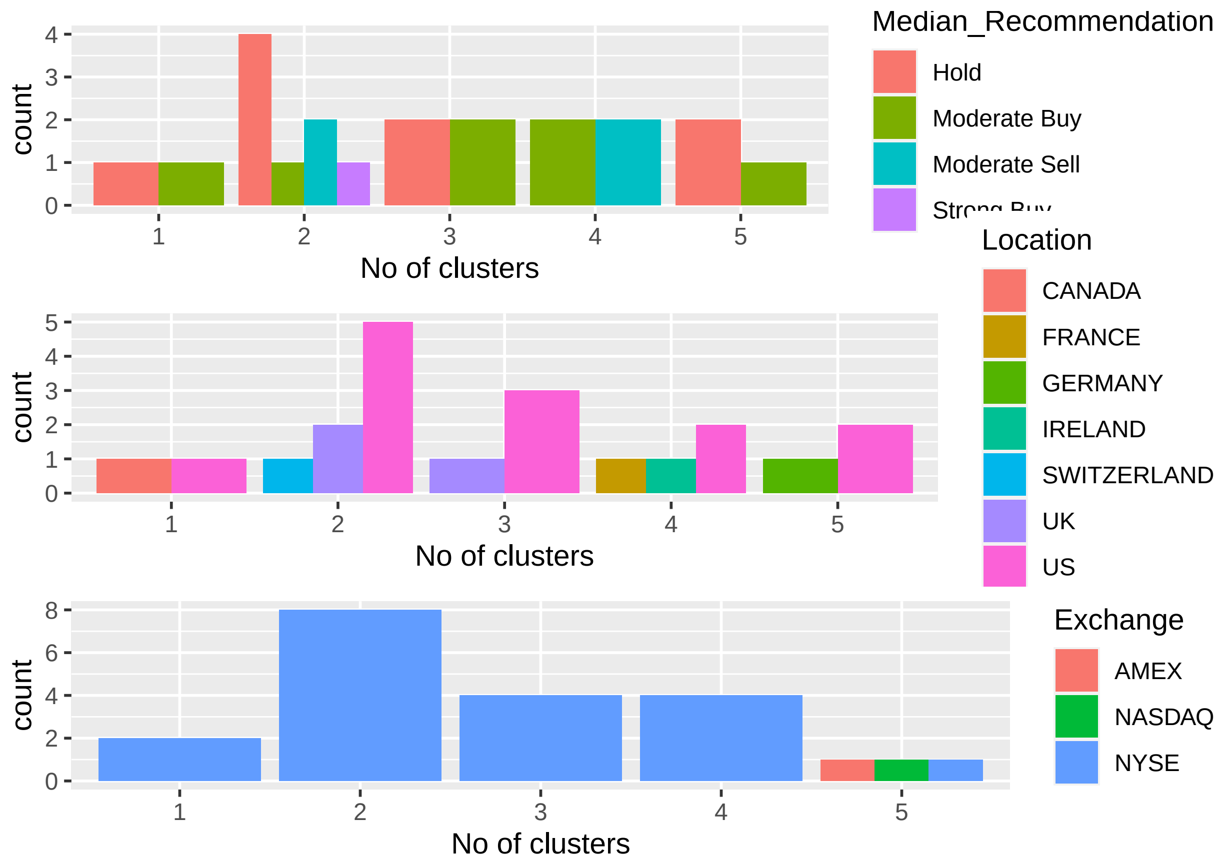
**Cluster 5 - 6, 8, 12 (lowest Rev_Growth,highest Beta and leveарge,lowest Net_Profit_Margin)**

```
pcn_cluster <- pharmacts[,c(12,13,14)]%>% mutate(clusters = final_Cl$cluster)%>% arrange(clusters, ascen
pcn_cluster
```

```
##    Median_Recommendation    Location Exchange clusters
## 1            Moderate Buy      CANADA     NYSE        1
## 2                    Hold          US     NYSE        1
## 3            Moderate Buy          US     NYSE        2
## 4              Strong Buy          UK     NYSE        2
## 5           Moderate Sell          UK     NYSE        2
## 6           Moderate Sell          US     NYSE        2
## 7                    Hold          US     NYSE        2
## 8                    Hold SWITZERLAND     NYSE        2
## 9                    Hold          US     NYSE        2
## 10                   Hold          US     NYSE        2
## 11                   Hold          UK     NYSE        3
## 12           Moderate Buy          US     NYSE        3
## 13                   Hold          US     NYSE        3
## 14           Moderate Buy          US     NYSE        3
## 15           Moderate Buy      FRANCE     NYSE        4
## 16          Moderate Sell     IRELAND     NYSE        4
## 17           Moderate Buy          US     NYSE        4
## 18          Moderate Sell          US     NYSE        4
## 19                   Hold     GERMANY     NYSE        5
## 20           Moderate Buy          US   NASDAQ        5
## 21                   Hold          US     AMEX        5
```

## Ques) Is there a pattern in the clusters with respect to the numerical variables? (10 to 12) variables? (those n not utilized in the cluster formation).

```
plot1_nrc<-ggplot(pcn_cluster, mapping = aes(factor(clusters), fill=Median_Recommendation))+geom_bar(pos
plot2_nrc<- ggplot(pcn_cluster, mapping = aes(factor(clusters),fill = Location))+geom_bar(position = 'd
plot3_nrc<- ggplot(pcn_cluster, mapping = aes(factor(clusters),fill = Exchange))+geom_bar(position = 'd
grid.arrange(plot1_nrc, plot2_nrc, plot3_nrc)
```




## As per the graphs, the interpretation is as follows:

**# Cluster 1: has the same hold and moderate buy medians and is spread out across the US, UK, and listed in NYSE.**

**# Cluster 2 : In this cluster, which displays distinct Hold, Moderate Buy, a little increased Moderate Sell, and Strong Buy medians, the Hold median is the highest. They are from the US, the UK, and Switzerland and are traded on the NYSE.**

**# Cluster 3: Exclusively listed on the NYSE, evenly distributed across the US and Canada, with medians of Moderate Hold and Moderate Buy.**

**# Cluster 4: listed on the NYSE, has distinct counts for France, Ireland, and the US, and has medians for buy and sell orders that are equally Moderate.**

**# Cluster 5 : Listed on AMEX, NASDAQ, and NYSE stock exchanges, all have an equal distribution of companies, but there is a clear Hold and Moderate Buy median as well as a different count between the US and Germany.**

**With respect to the median Recommendation Variable ,the clusters follow a particular pattern:**

**Cluster 2 and Cluster 5 has Hold Recommendation.**

**Cluster 1, Cluster 3 and Cluster 4 has moderate buy Recommendation.**

---

**Question 3.**

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

Based on certain criteria, preferably financial measures such as performance, potential for growth, or risk factors, Investors and analysts can use such clusters to make informed decisions about their investment strategies. We can name each cluster appropriately as follows:

Cluster 1 :- HOLD-BUY CLUSTER or Balanced Investment Cluster.

Cluster 2 :- HIGH HOLD CLUSTER or Robust Holding Cluster.

Cluster 3 :- HOLD-BUY CLUSTER or Balanced Investment Cluster.

Cluster 4 :- BUY-SELL CLUSTER or Dynamic Portfolio Cluster.

Cluster 5 :- HOLD CLUSTER or Stable Investment Cluster.