Assignment_4

Krishna Kumar Tavva - 811283461

2023-03-16

install packages and call the libraries

```
#install.packages("factoextra")
#install.packages("pander")
#install.packages("cowplot")
library(tidyverse) #data manipulation
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(factoextra) #clustering algorithms & visualization
```

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

```
library(ISLR) #Statistical Learning
library(dplyr)
library(ggplot2)
library(pander)
library(cowplot)
```

Loading the data

```
Pharma <- read.csv("E:\\Fundamentals of Machine Learning\\Module 6\\Pharmaceuticals.csv")
summary(Pharma)
```

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

```
row.names(Pharma) <- Pharma[,1]
```

Looking for null values

```
any(is.na.data.frame(Pharma))
```

```
## [1] FALSE
```

a.Normalization and finding the optimal k by Elbow chart and the Silhouette Method

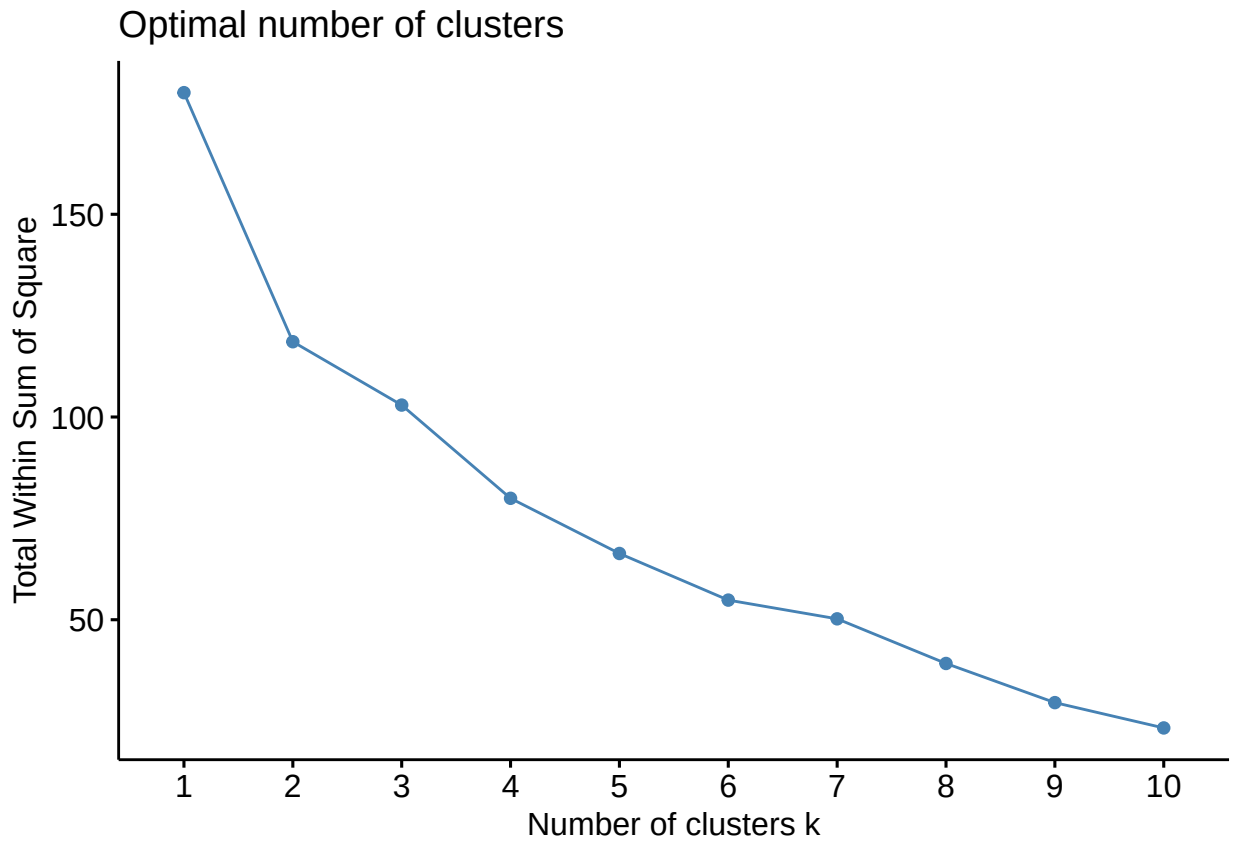
```
set.seed(1)
Pharma_norm <- scale(Pharma[,-c(1:2,12:14)])

pandoc.table(head(Pharma_norm),style="grid", split.tables = Inf)# top 6 Observation from pharma_Norm
```

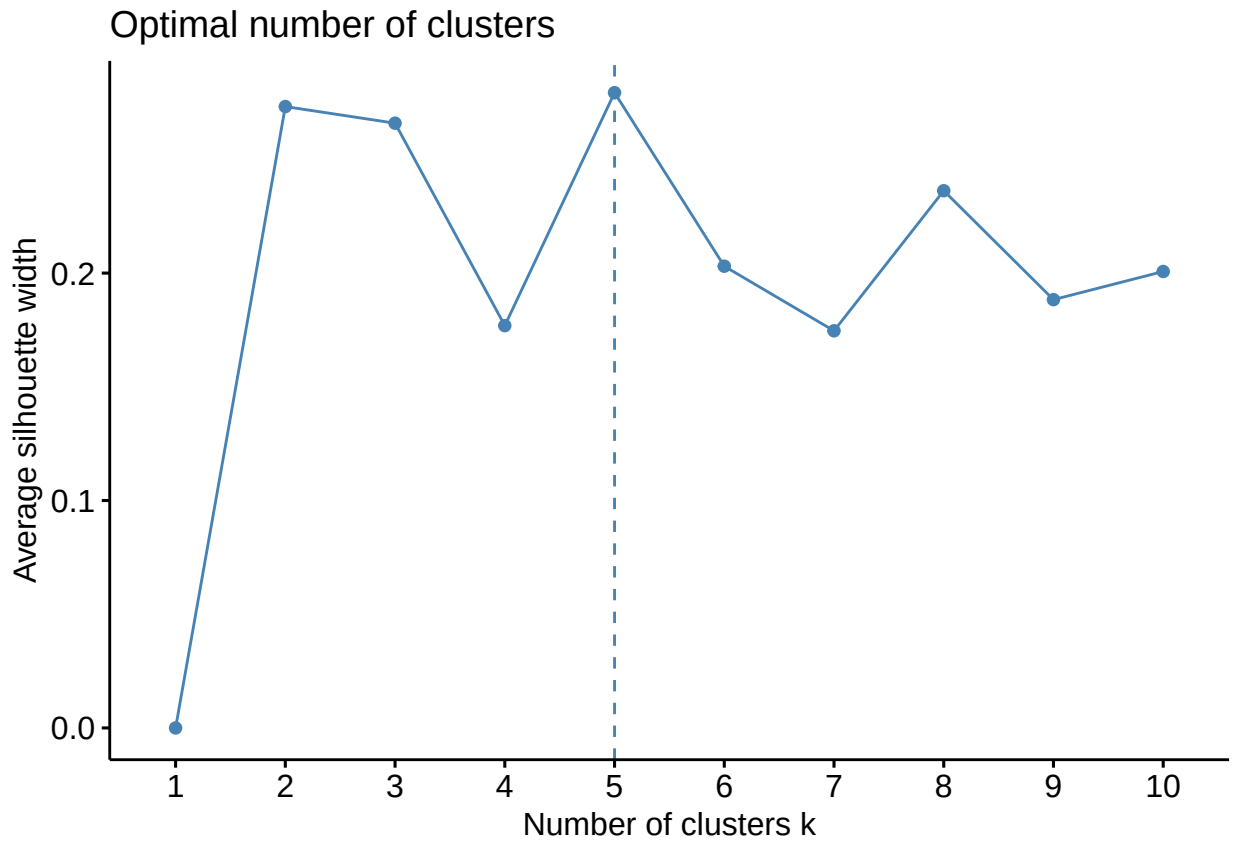
```
##
##
## +---------+------------+----------+----------+---------+---------+----------------+----------+-------
## |    | Market_Cap |   Beta   | PE_Ratio |   ROE   |   ROA   | Asset_Turnover | Leverage | Rev_G
## +=========+============+==========+==========+=========+=========+================+==========+=======
## | **ABT** |   0.1841   | -0.8013  | -0.04671 | 0.04009 | 0.2416  |       0        | -0.2121  |  -0.5
```

```
## +---------+------------+----------+----------+---------+---------+----------------+----------+-------
## | **AGN** |  -0.8544   | -0.4507  |  3.497   | -0.8548 | -0.9423 |     0.9225     | 0.01828  |  -0.38
## +---------+------------+----------+----------+---------+---------+----------------+----------+-------
## | **AHM** |  -0.8763   |  -0.256  |  -0.292  | -0.7223 | -0.5101 |     0.9225     | -0.4041  |  -0.57
## +---------+------------+----------+----------+---------+---------+----------------+----------+-------
## | **AZN** |   0.1703   | -0.02226 | -0.2429  | 0.1064  | 0.9181  |     0.9225     | -0.7497  |   0.14
## +---------+------------+----------+----------+---------+---------+----------------+----------+-------
## | **AVE** |   -0.179   | -0.8013  | -0.3287  | -0.2648 | -0.5664 |    -0.4613     | -0.3145  |   1.21
## +---------+------------+----------+----------+---------+---------+----------------+----------+-------
## | **BAY** |  -0.6954   |  2.276   |  0.1495  | -1.451  | -1.713  |    -0.4613     | -0.7497  |   -1.4
## +---------+------------+----------+----------+---------+---------+----------------+----------+-------
```

```
wss <- fviz_nbclust(Pharma_norm,kmeans,method="wss")
wss
```




```
silhouette <- fviz_nbclust(Pharma_norm,kmeans,method="silhouette")
silhouette
```



#The optimal k thereby received using the wss method is k = 2 whereas by employing the silhouette method the optimal k received was k = 5.

Formulation of clusters using K-Means with k = 2 (WSS)

```
wss_kmeans <- kmeans(Pharma_norm,centers = 2,nstart=25)
pandoc.table(wss_kmeans$centers,style="grid", split.tables = Inf)
```

```
##
##
## +------------+---------+----------+---------+---------+----------------+----------+------------+-----
## | Market_Cap |  Beta   | PE_Ratio |   ROE   |   ROA   | Asset_Turnover | Leverage | Rev_Growth | Net_
## +============+=========+==========+=========+=========+================+==========+============+=====
## |   0.6734   | -0.3586 | -0.2764  | 0.6566  | 0.8344  |     0.4613     | -0.3331  |  -0.2902   |
## +------------+---------+----------+---------+---------+----------------+----------+------------+-----
## |  -0.7407   | 0.3945  |  0.304   | -0.7223 | -0.9179 |    -0.5074     |  0.3664  |   0.3192   |
## +------------+---------+----------+---------+---------+----------------+----------+------------+-----
```

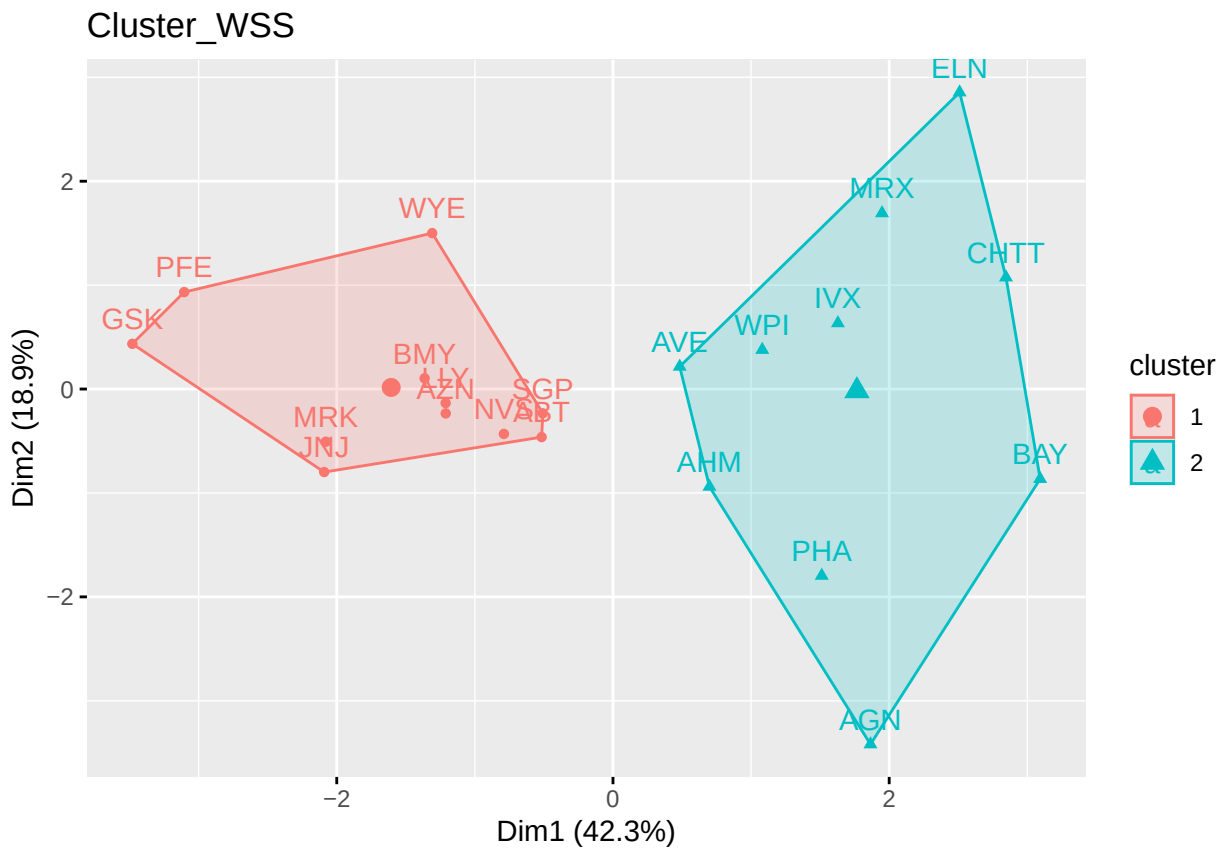
Formulation of clusters using K-Means with k = 5 (Silhouette)

```
silhouette_kmeans <- kmeans(Pharma_norm,centers=5,nstart=25)
pandoc.table(silhouette_kmeans$centers,style="grid", split.tables = Inf)
```

```
##
##
## +------------+---------+----------+---------+---------+----------------+----------+------------+-----
## | Market_Cap |  Beta   | PE_Ratio |   ROE   |   ROA   | Asset_Turnover | Leverage | Rev_Growth | Net_
## +============+=========+==========+=========+=========+================+==========+============+=====
## |  -0.7602   | 0.2796  | -0.4774  | -0.7438 | -0.8107 |     -1.268     | 0.06308  |   1.518    |
## +------------+---------+----------+---------+---------+----------------+----------+------------+-----
## |  -0.4393   | -0.4702 |   2.7    | -0.835  | -0.9235 |     0.2306     | -0.1417  |  -0.1168   |
## +------------+---------+----------+---------+---------+----------------+----------+------------+-----
## |  -0.8705   |  1.341  | -0.05284 | -0.6184 | -1.193  |    -0.4613     |  1.366   |  -0.6913   |
## +------------+---------+----------+---------+---------+----------------+----------+------------+-----
## |  -0.03142  | -0.4361 | -0.3172  |  0.195  | 0.4084  |     0.173      | -0.2745  |  -0.7042   |
## +------------+---------+----------+---------+---------+----------------+----------+------------+-----
## |   1.696    | -0.1781 | -0.1985  |  1.235  |  1.35   |     1.153      | -0.4681  |   0.4672   |
## +------------+---------+----------+---------+---------+----------------+----------+------------+-----
```

Cluster Plot (Elbow)

```
fviz_cluster(wss_kmeans,data=Pharma_norm,main="Cluster_WSS")
```



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

```
wss_kmeans$size #Size of the cluster
```

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

```
#By employing the WSS Method we get 2 clusters of size 11 and 10.

wss_kmeans$withinss #Total within-cluster sum of squares
```

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

```
wss_kmeans$cluster[19]
```

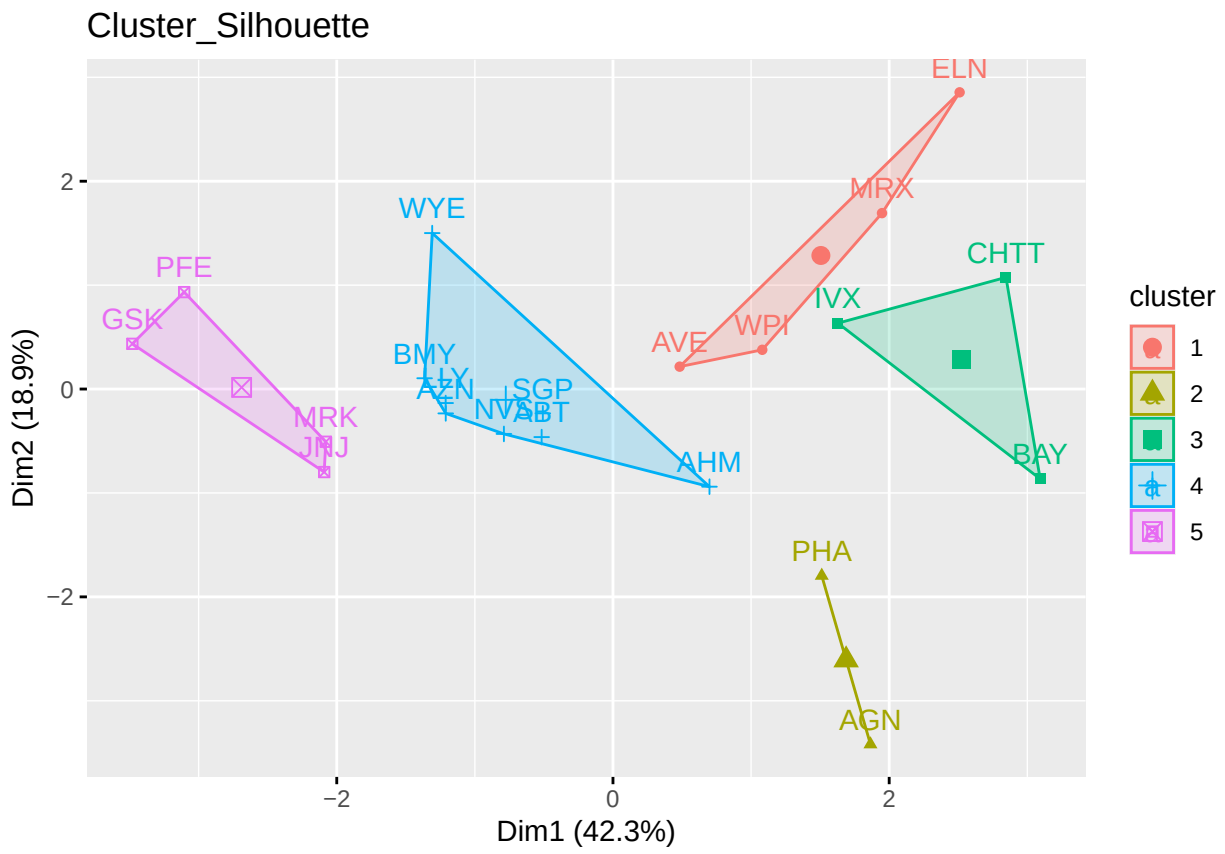
```
## SGP
##   1
```

```
paste("Observation 19th is SGP and it belongs to cluster", wss_kmeans$cluster[19])
```

```
## [1] "Observation 19th is SGP and it belongs to cluster 1"
```

Cluster Plot (Silhouette)

```
fviz_cluster(silhouette_kmeans,data=Pharma_norm,main="Cluster_Silhouette")
```




```
silhouette_kmeans$size #Size of the cluster
```

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

#By employing the Silhouette Method we get 5 clusters of size 4, 2, 3, 8 and 4. Out of all, Cluster 3 has more number of observations.

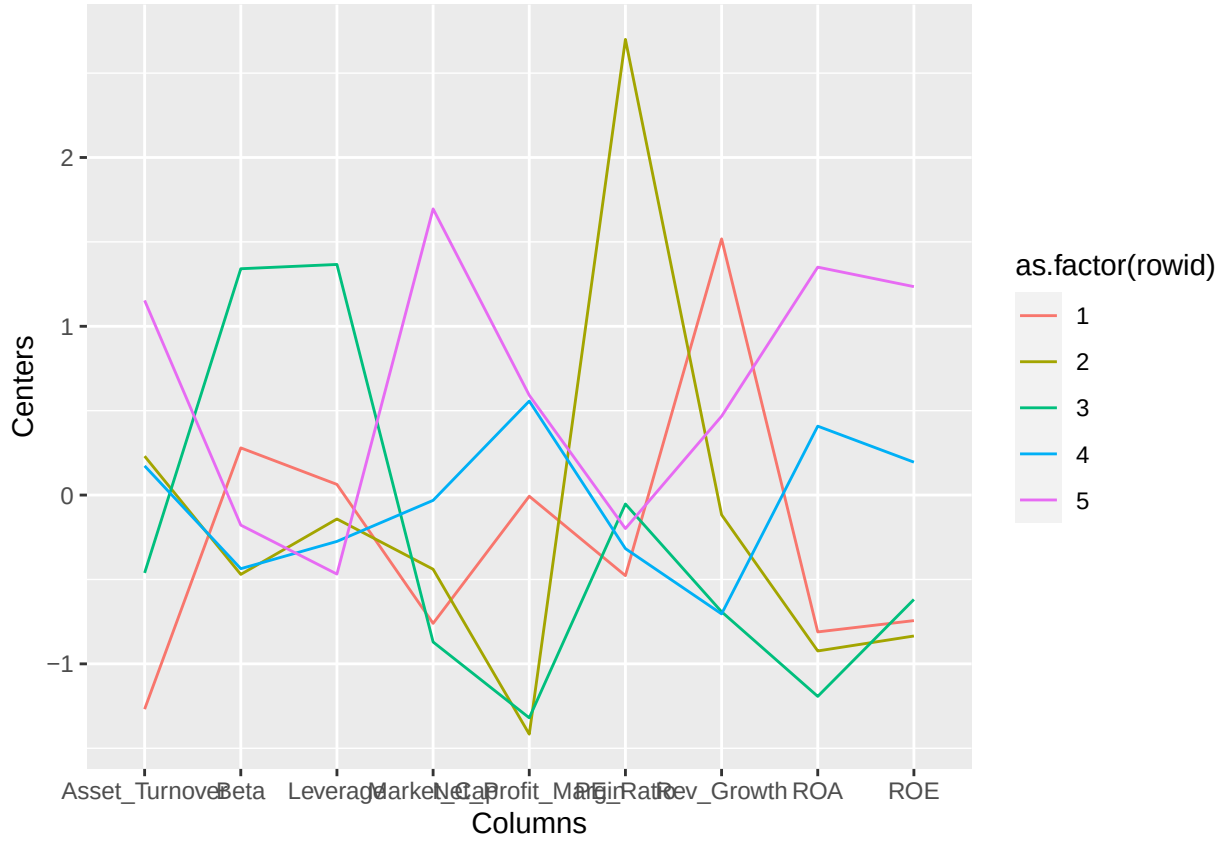
#graphical plotting of data grouped in clusters

```
Centroid_1 <- data.frame(silhouette_kmeans$centers) %>% rowid_to_column() %>%
  gather('Columns', 'Centers', 2:10)
print(Centroid_1)
```

```
##    rowid          Columns      Centers
## 1      1       Market_Cap -0.760224892
## 2      2       Market_Cap -0.439251341
## 3      3       Market_Cap -0.870515113
## 4      4       Market_Cap -0.031422109
## 5      5       Market_Cap  1.695581115
## 6      1             Beta  0.279604106
## 7      2             Beta -0.470180039
## 8      3             Beta  1.340986857
## 9      4             Beta -0.436098941
## 10     5             Beta -0.178056346
## 11     1         PE_Ratio -0.477423799
```

```
## 12     2          PE_Ratio  2.700024643
## 13     3          PE_Ratio -0.052844340
## 14     4          PE_Ratio -0.317248516
## 15     5          PE_Ratio -0.198458234
## 16     1               ROE -0.743802224
## 17     2               ROE -0.834952524
## 18     3               ROE -0.618401510
## 19     4               ROE  0.195045857
## 20     5               ROE  1.234987906
## 21     1               ROA -0.810742783
## 22     2               ROA -0.923495091
## 23     3               ROA -1.192847826
## 24     4               ROA  0.408391543
## 25     5               ROA  1.350343113
## 26     1    Asset_Turnover -1.268480411
## 27     2    Asset_Turnover  0.230632802
## 28     3    Asset_Turnover -0.461265604
## 29     4    Asset_Turnover  0.172974602
## 30     5    Asset_Turnover  1.153164010
## 31     1          Leverage  0.063080849
## 32     2          Leverage -0.141703357
## 33     3          Leverage  1.366446992
## 34     4          Leverage -0.274493115
## 35     5          Leverage -0.468078185
## 36     1        Rev_Growth  1.518015830
## 37     2        Rev_Growth -0.116845875
## 38     3        Rev_Growth -0.691291399
## 39     4        Rev_Growth -0.704151557
## 40     5        Rev_Growth  0.467178770
## 41     1 Net_Profit_Margin -0.006893899
## 42     2 Net_Profit_Margin -1.416514761
## 43     3 Net_Profit_Margin -1.320000179
## 44     4 Net_Profit_Margin  0.556954446
## 45     5 Net_Profit_Margin  0.591242521
```

```
ggplot(Centroid_1, aes(x = Columns, y = Centers, color = as.factor(rowid))) +
  geom_line(aes(group = as.factor(rowid)))
```



c.To find a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

```
Pharma_Pattern <- Pharma %>% select(c(12,13,14)) %>% mutate(Cluster = silhouette_kmeans$cluster)
print(Pharma_Pattern) #The remaining three categories are Stock Exchange, Location, and Median
```

```
##      Median_Recommendation    Location Exchange Cluster
## ABT           Moderate Buy          US     NYSE       4
## AGN           Moderate Buy      CANADA     NYSE       2
## AHM             Strong Buy          UK     NYSE       4
## AZN          Moderate Sell          UK     NYSE       4
## AVE           Moderate Buy      FRANCE     NYSE       1
## BAY                   Hold     GERMANY     NYSE       3
## BMY          Moderate Sell          US     NYSE       4
## CHTT          Moderate Buy          US   NASDAQ       3
## ELN          Moderate Sell     IRELAND     NYSE       1
## LLY                   Hold          US     NYSE       4
## GSK                   Hold          UK     NYSE       5
## IVX                   Hold          US     AMEX       3
## JNJ           Moderate Buy          US     NYSE       5
## MRX           Moderate Buy          US     NYSE       1
## MRK                   Hold          US     NYSE       5
```

```
## NVS                    Hold SWITZERLAND     NYSE       4
## PFE            Moderate Buy           US     NYSE       5
## PHA                    Hold           US     NYSE       2
## SGP                    Hold           US     NYSE       4
## WPI           Moderate Sell           US     NYSE       1
## WYE                    Hold           US     NYSE       4
```

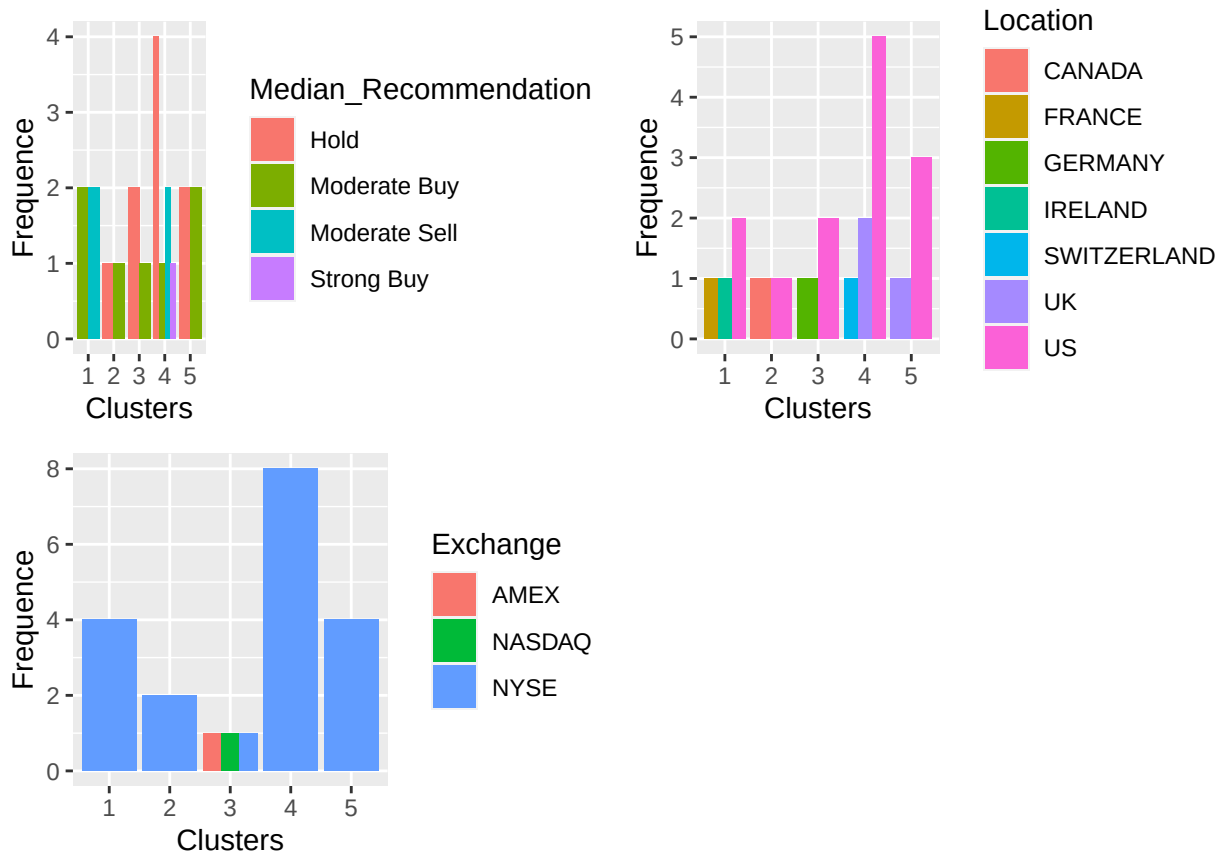
To visualize the distribution of businesses grouped by clusters and to identify any trends in the data, utilizing bar charts

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

Location_0 <- ggplot(Pharma_Pattern, mapping = aes(factor(Cluster), fill=Location)) +
  geom_bar(position = 'dodge') + labs(x='Clusters', y='Frequence')

Exchange_0 <- ggplot(Pharma_Pattern, mapping = aes(factor(Cluster), fill=Exchange)) +
  geom_bar(position = 'dodge') + labs(x='Clusters', y='Frequence')

plot_grid(Median_Recom, Location_0, Exchange_0)
```



#The clustering analysis suggests that the companies in each cluster have similar characteristics in terms

#Cluster -1 has companies from various locations listed on the NYSE, and they have a moderate buy or sell recommendation.

#Cluster -2 has a mix of American and Canadian companies listed on the NYSE, and they have a moderate buy or sell recommendation.

#Cluster -3 has companies from Germany and the USA listed on stock exchange markets other than NYSE (AMEX, NASDAQ) have a hold and moderate buy recommendation .

#Cluster -4 is dominated by American-based companies listed on the New York Stock Exchange, and they have a hold, moderate sell, strong buy and moderate buy recommendation.

#Cluster -5 has companies from the UK and USA, and they have a hold and moderate buy recommendation.

d.Naming for each cluster using the variables in the dataset.

#Based on the entire analysis and looking at the characterstics of the clusters, 21 pharmaceutical industries can be categorized into 5 different groups:

#Cluster 1 - "Growth oriented - Low risky companies": A company with low asset turnover and high revenue growth may indicate that the company has significant growth potential but is not yet operating at optimal efficiency. Investors should consider the company's industry and competitive landscape, as well as its ability to sustain high revenue growth over the long term. It's also important to evaluate the company's profitability, as high revenue growth may not necessarily lead to higher profits if the company is not utilizing its assets efficiently.Also,these are the companies from various locations listed on the NYSE, and they have a moderate buy or sell recommendation, suggesting that they may have some growth potential.

#Cluster 2 - "Overpriced - Risky companies": since it has high price-to-earnings (PE) ratio and a low net profit margin means that the market is valuing the company's stock at a premium compared to its current earnings, even though the company's net profit margin is relatively low. which means investors are willing to pay a high price for each dollar of earnings the company generates, despite the fact that the company is not generating a high level of profit compared to its revenue.Such companies can be risky, as they may not be able to meet the market's expectations and may experience a decline in stock price in the future.

#Cluster 3 - "Debt-ridden - very risky companies": Companies with high leverage and low net profit margin & ROA may indicate that the company is taking on a significant amount of debt to finance its operations, while not generating a sufficient level of profitability or returns on assets. This can be a concerning signal for investors, as the company may struggle to meet its debt obligations and may experience financial distress in the long term.Also,listed on stock exchange markets other than NYSE (AMEX and NASDAQ), and they have a hold or moderate buy recommendation.

#Cluster 4 - "Stable - efficient companies": company with normal levels across financial metrics can be considered that the company is operating efficiently and effectively within its industry and competitive landscape. Also it is dominated by American-based companies listed on the New York Stock Exchange, and they have a spread advice to keep their stock, suggesting that they are stable and relatively low-risk investments.

#Cluster 5 - "Established - profitable companies": Companies with high market capitalization are typically large and well-established companies that have a significant market presence and a strong financial position. High market capitalization means that the company has a large number of outstanding shares and a high stock price, resulting in a high valuation.Also,they have a partially hold and buy recommendation for their stocks listed on the NYSE.