Clustering-assignment-FML

Anurodh Singh

2023-11-10

#Interpretation

1.The equities analyst or investor is seeking guidance on identifying the best companies for investment in the pharmaceutical industry. 2.Through clustering techniques, the goal is to determine which clusters or groups of companies would be optimal for investment. 3.Among the clustering methods employed, the "K-means" method stands out for providing the best clusters and a cohesive representation in comparison to "DBSCAN" and "HIERARCHIAL" clsutering methods. 4.Cluster 4 and 3 from the K-means algorithm are identified as the most favorable due to appropriate values for market capitalization, return on equity, return on assets, asset turnover, estimated revenue growth, and net profit margin. 5.Negative or lower values for beta, price/earnings ratio, and leverage are also considered favorable. 6.Variables Analysis: a.Market Capitalization: Larger values may suggest stability or growth potential. b.Return on Equity (ROE): Positive values indicate profitability from shareholders' equity. c.Return on Assets (ROA): Positive values imply efficient asset utilization. d.Asset Turnover: Positive values indicate effective asset use for revenue generation. e.Estimated Revenue Growth: Positive values suggest potential revenue increase. f.Net Profit Margin: Positive values indicate a favorable percentage of revenue translating into profit. g.Beta: Negative values can provide risk diversification. h.Price/Earnings Ratio (P/E Ratio): Lower values might indicate a relatively cheaper stock. i.Leverage: Lower values suggest lower reliance on debt financing.

Cluster Evaluation: Using the elbow method with two clusters, 34.1% variance between clusters. Silhouette method with multiple clusters has a higher 60.9% variance between clusters. Higher variance between clusters suggests better separation, favoring method. Generally, a higher percentage of variance between clusters is desirable. Finally the silhouette method appears to offer better separation between clusters, as evidenced by the higher percentage of variance between clusters. This suggests that method may provide more distinct and well-defined clusters for investment decisions. The results of the clustering analysis reveal distinct patterns and characteristics within each cluster:

Cluster 1: Buy Cluster

Highest median for the "Hold" recommendation. Companies from Switzerland and the United States. Listed on the NYSE. Cluster 2: Sceptical Cluster

Even distribution across AMEX, NASDAQ, and NYSE. Distinct "Hold" and "Moderate Buy" medians. Companies from the United States and Germany. Cluster 3: Moderate Buy Cluster

Listed on the NYSE. Separate counts for the United States, Ireland, and France. Equal "Moderate Buy" and "Moderate Sell" medians. Cluster 4: Hold Cluster

Distributed throughout the United States and the United Kingdom. Listed shares with the same "Hold" and "Moderate Buy" medians. Cluster 5: High Hold Cluster

Only on the NYSE. Equally distributed in the US and Canada. Medians for "Hold" and "Moderate Buy." Pattern in Media Recommendation Variable:

"Hold" recommendation applies to Clusters 1 and 2. "Moderate Buy" recommendation for Clusters 3, 4, and 5.

Summary of Clusters:

Cluster 1: Buy Cluster Cluster 2: Sceptical Cluster Cluster 3: Moderate Buy Cluster Cluster 4: Hold Cluster Cluster 5: High Hold Cluster These clusters not only demonstrate geographical and stock exchange distribution but also showcase specific trends in media recommendations, providing valuable insights for investors and equities analysts.

#SUMMARY

1.Imported the given dataset. 2.Examined variable types, identifying numerical and categorical variables. 3.Scaled the data for convenience. 4.Utilized the cluster package, performing k-means clustering with k=2. 5.Obtained results with clusters of sizes 10, 11. 6.Used the elbow method to find the optimal k, which turned out to be 2. 7.Performed k-means clustering with k=2, but clusters were not satisfactory. 8.Employed the silhouette method, determining k=5 as the optimal value. 9.Obtained well-defined clusters using k=5 suggested by the silhouette method. 10.Analyzed cluster properties with k-means clustering and 5 clusters of sizes 3, 5, 3, 4, 6. 11.Examined output centers, the number of companies in each cluster, and identified the cluster of the 13th observation. 12.Visualized the output using the fviz_cluster function. 13.Applied DBSCAN cluster algorithm with random values for eps=30 and min. points=1. 14.Found 7 clusters and 0 noise points using Euclidean distances. 15.Plotted results and printed cluster details using fviz_cluster. 16.Conducted Hierarchical clustering, obtaining a dendrogram for the 21 companies based on variables, showing closeness. 17.Plotted a heatmap for hierarchical clustering. 18.Interpreted clusters concerning numerical variables (10 to 12) not used in forming clusters. 19.Provided a structured analysis of the pharmaceutical industry based on financial metrics, revealing patterns and insights. 20.Named each cluster appropriately using variables in the dataset, with following names : Cluster 1 :-Buy Cluster Cluster 2 :- Sceptical Cluster Cluster 3 :- Moderate Buy Cluster Cluster 4 :- Hold Cluster Cluster 5 :- High Hold Cluster

Q.1.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. Importing the given data

```
pharma_given <- read.csv("C:/Users/ASUS/Downloads/Pharmaceuticals.csv")
summary(pharma_given)
```

```
##     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(pharma_given)
```

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

```
# Scaling the numerical variables
pharma1 <- pharma_given[ ,-c(1,2,12,13,14)]
pharma1
```

```
##    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
## 7       51.33 0.50     13.9 34.8 15.1            0.9     0.57       2.70
## 8        0.41 0.85     26.0 24.1  4.3            0.6     3.51       6.38
## 9        0.78 1.08      3.6 15.1  5.1            0.3     1.07      34.21
## 10      73.84 0.18     27.9 31.0 13.5            0.6     0.53       6.21
## 11     122.11 0.35     18.0 62.9 20.3            1.0     0.34      21.87
## 12       2.60 0.65     19.9 21.4  6.8            0.6     1.45      13.99
## 13     173.93 0.46     28.4 28.6 16.3            0.9     0.10       9.37
## 14       1.20 0.75     28.6 11.2  5.4            0.3     0.93      30.37
## 15     132.56 0.46     18.9 40.6 15.0            1.1     0.28      17.35
## 16      96.65 0.19     21.6 17.9 11.2            0.5     0.06      -2.69
## 17     199.47 0.65     23.6 45.6 19.2            0.8     0.16      25.54
## 18      56.24 0.40     56.5 13.5  5.7            0.6     0.35      15.00
## 19      34.10 0.51     18.9 22.6 13.3            0.8     0.00       8.56
## 20       3.26 0.24     18.4 10.2  6.8            0.5     0.20      29.18
## 21      48.19 0.63     13.1 54.9 13.4            0.6     1.12       0.36
##    Net_Profit_Margin
## 1               16.1
## 2                5.5
## 3               11.2
```

```
## 4               18.0
## 5               12.9
## 6                2.6
## 7               20.6
## 8                7.5
## 9               13.3
## 10              23.4
## 11              21.1
## 12              11.0
## 13              17.9
## 14              21.3
## 15              14.1
## 16              22.4
## 17              25.2
## 18               7.3
## 19              17.6
## 20              15.1
## 21              25.5
```

```
pharma_scaled <- scale(pharma1)
pharma_scaled
```

```
##       Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
##  [1,]  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
##  [2,] -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
##  [3,] -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312
##  [4,]  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312
##  [5,] -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656
##  [6,] -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656
##  [7,] -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498      0.9225312
##  [8,] -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918     -0.4612656
##  [9,] -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553     -1.8450624
## [10,]  0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770     -0.4612656
## [11,]  1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364      1.3837968
## [12,] -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905     -0.4612656
## [13,]  1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544      0.9225312
## [14,] -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792     -1.8450624
## [15,]  1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577      1.8450624
## [16,]  0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598     -0.9225312
## [17,]  2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239      0.4612656
## [18,] -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030     -0.4612656
## [19,] -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929      0.4612656
## [20,] -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905     -0.9225312
## [21,] -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849     -0.4612656
##          Leverage  Rev_Growth Net_Profit_Margin
##  [1,] -0.21209793 -0.52776752        0.06168225
##  [2,]  0.01828430 -0.38113909       -1.55366706
##  [3,] -0.40408312 -0.57211809       -0.68503583
##  [4,] -0.74965647  0.14744734        0.35122600
##  [5,] -0.31449003  1.21638667       -0.42597037
##  [6,] -0.74965647 -1.49714434       -1.99560225
##  [7,] -0.02011273 -0.96584257        0.74744375
##  [8,]  3.74279705 -0.63276071       -1.24888417
##  [9,]  0.61983791  1.88617085       -0.36501379
```

```
## [10,] -0.07130879 -0.64814764        1.17413980
## [11,] -0.31449003  0.76926048        0.82363947
## [12,]  1.10620040  0.05603085       -0.71551412
## [13,] -0.62166634 -0.36213170        0.33598685
## [14,]  0.44065173  1.53860717        0.85411776
## [15,] -0.39128411  0.36014907       -0.24310064
## [16,] -0.67286239 -1.45369888        1.02174835
## [17,] -0.54487226  1.10143723        1.44844440
## [18,] -0.30169102  0.14744734       -1.27936246
## [19,] -0.74965647 -0.43544591        0.29026942
## [20,] -0.49367621  1.43089863       -0.09070919
## [21,]  0.68383297 -1.17763919        1.49416183
## attr(,"scaled:center")
##        Market_Cap              Beta          PE_Ratio               ROE
##        57.6514286         0.5257143        25.4619048        25.7952381
##               ROA    Asset_Turnover          Leverage        Rev_Growth
##        10.5142857         0.7000000         0.5857143        13.3709524
## Net_Profit_Margin
##        15.6952381
## attr(,"scaled:scale")
##        Market_Cap              Beta          PE_Ratio               ROE
##        58.6029595         0.2567406        16.3102568        15.0849752
##               ROA    Asset_Turnover          Leverage        Rev_Growth
##         5.3213988         0.2167948         0.7813103        11.0483351
## Net_Profit_Margin
##         6.5620482
```

```r
set.seed(321)
# Choosing a clustering algorithm
library(cluster)

# Setting the number of clusters to 2
k <- 2

# Clustering the data using k-means clustering
kmeans_model <- kmeans(pharma_scaled, k)
kmeans_model
```

```
## K-means clustering with 2 clusters of sizes 10, 11
##
## Cluster means:
##   Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.7407208  0.3945061  0.3039863 -0.7222576 -0.9178575     -0.5073922
## 2  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656
##     Leverage Rev_Growth Net_Profit_Margin
## 1  0.3664175  0.3192379        -0.7505641
## 2 -0.3331068 -0.2902163         0.6823310
##
## Clustering vector:
##  [1] 2 1 1 2 1 1 2 1 1 2 2 1 2 1 2 2 2 1 2 1 2
##
## Within cluster sum of squares by cluster:
## [1] 75.26049 43.30886
##  (between_SS / total_SS =  34.1 %)
```

```
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
# Getting the cluster assignments
cluster_assignments <- kmeans_model$cluster

# Calculating the mean values of the numerical variables in each cluster
cluster_means <- aggregate(pharma_scaled, by = list(cluster_assignments), FUN = mean)

# Printing the cluster means
View(cluster_means)
```

```
library(tidyverse)  # for data manipulation
```

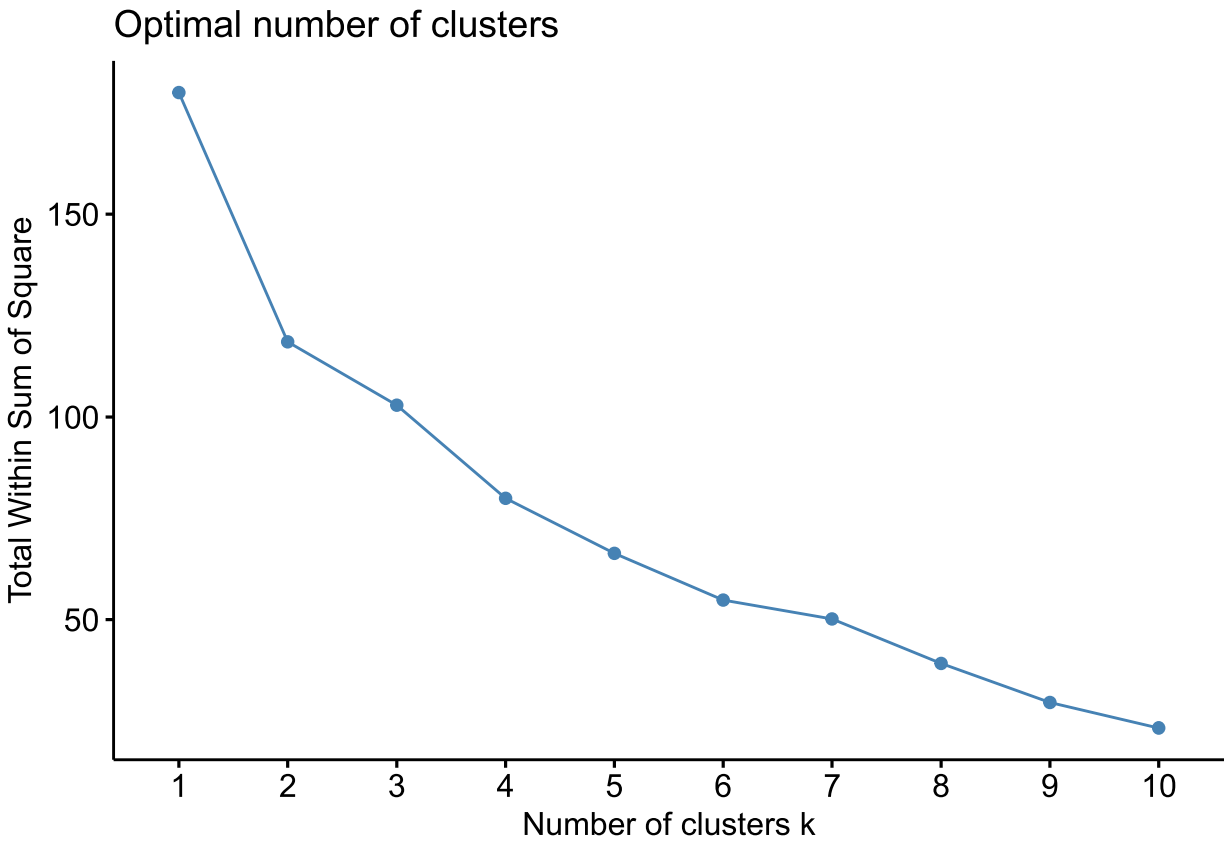
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

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

```
## Warning: package 'factoextra' was built under R version 4.3.2
```

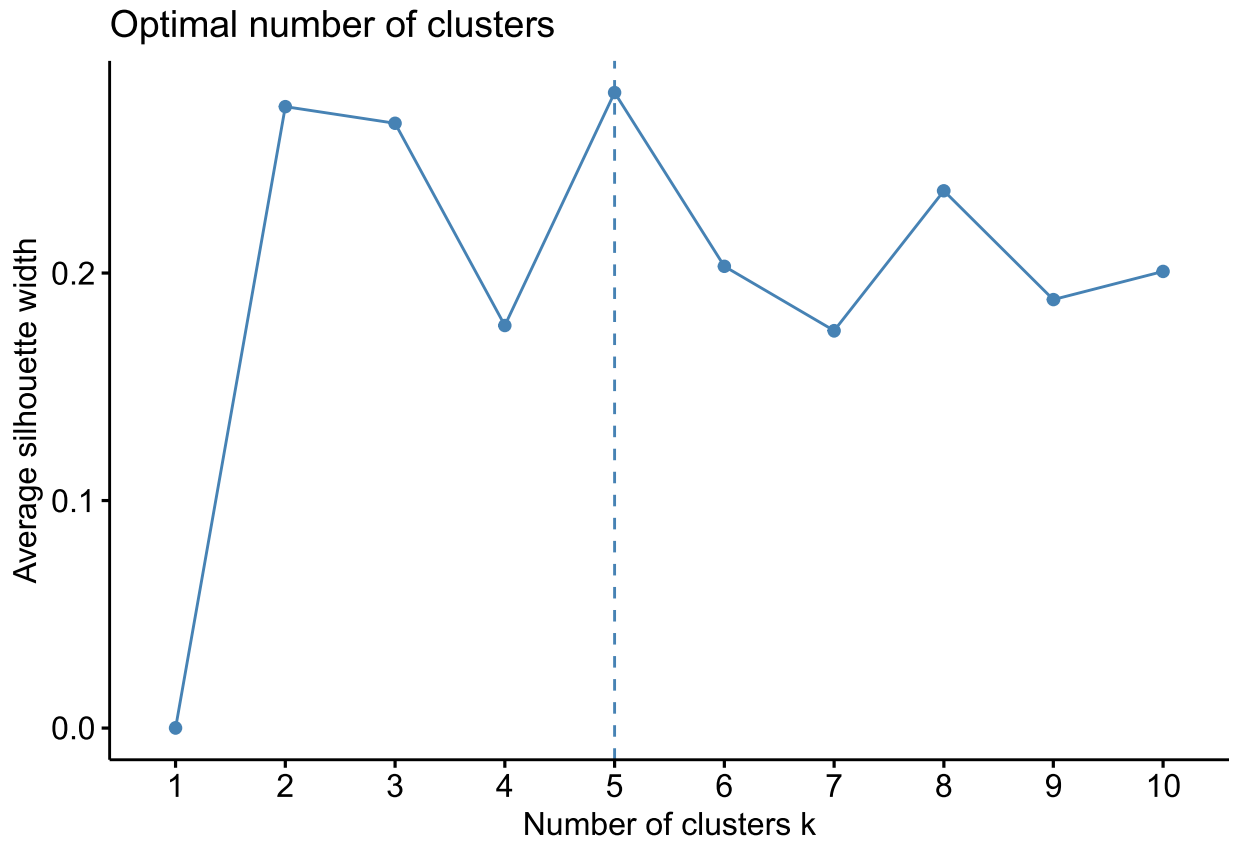
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(ISLR)

#Finding the optimal value of 'K' using elbow method

fviz_nbclust(pharma_scaled, kmeans, method = "wss")
```



We can see that the optimal value from the graph above is 2

```
#Now applying silhouette method to obtain the optimal value of 'K'
fviz_nbclust(pharma_scaled, kmeans, method = "silhouette")
```




```
set.seed(321)
#applying the clustering algorithm

# Setting the number of clusters to 5 as obtained by silhouette method
k <- 5

# Clustering the data using k-means clustering
kmeans_model_sil <- kmeans(pharma_scaled, k)
kmeans_model_sil
```

```
## K-means clustering with 5 clusters of sizes 3, 5, 3, 4, 6
##
## Cluster means:
##   Market_Cap       Beta   PE_Ratio         ROE        ROA Asset_Turnover
## 1 -0.5246281  0.4451409  1.8498439 -1.04045502 -1.1865838   1.480297e-16
## 2 -0.2063280 -0.2481660 -0.3385541 -0.03813318  0.4069821   6.457718e-01
## 3  0.2600876 -0.7493205 -0.2817392  0.58367759  0.4107405  -6.150208e-01
## 4  1.6955811 -0.1780563 -0.1984582  1.23498791  1.3503431   1.153164e+00
## 5 -0.8261772  0.4775991 -0.3696184 -0.56315890 -0.8514589  -9.994088e-01
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.34435439 -0.5769454        -1.6095439
## 2 -0.42712134 -0.4707453         0.1531171
## 3 -0.02011273 -1.0931619         1.2300167
## 4 -0.46807818  0.4671788         0.5912425
## 5  0.85022014  0.9158889        -0.3319956
```

```
##
## Clustering vector:
##  [1] 2 1 2 2 5 1 2 5 5 3 4 5 4 5 4 3 4 1 2 5 3
##
## Within cluster sum of squares by cluster:
## [1] 14.938904  6.586586  7.490937  9.284424 32.143356
##  (between_SS / total_SS =  60.9 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
cluster_assignments_sil <- kmeans_model_sil$cluster

# Calculating the mean values of the numerical variables in each cluster
cluster_means_sil <- aggregate(pharma_scaled, by = list(cluster_assignments_sil), FUN = mean)

# Printing the cluster means
print(cluster_means_sil)
```

```
##   Group.1 Market_Cap       Beta   PE_Ratio         ROE        ROA
## 1       1 -0.5246281  0.4451409  1.8498439 -1.04045502 -1.1865838
## 2       2 -0.2063280 -0.2481660 -0.3385541 -0.03813318  0.4069821
## 3       3  0.2600876 -0.7493205 -0.2817392  0.58367759  0.4107405
## 4       4  1.6955811 -0.1780563 -0.1984582  1.23498791  1.3503431
## 5       5 -0.8261772  0.4775991 -0.3696184 -0.56315890 -0.8514589
##   Asset_Turnover    Leverage Rev_Growth Net_Profit_Margin
## 1   1.480297e-16 -0.34435439 -0.5769454        -1.6095439
## 2   6.457718e-01 -0.42712134 -0.4707453         0.1531171
## 3  -6.150208e-01 -0.02011273 -1.0931619         1.2300167
## 4   1.153164e+00 -0.46807818  0.4671788         0.5912425
## 5  -9.994088e-01  0.85022014  0.9158889        -0.3319956
```

```
# Visualizing the output

kmeans_model_sil$centers # output the centers
```

```
##   Market_Cap       Beta   PE_Ratio         ROE        ROA Asset_Turnover
## 1 -0.5246281  0.4451409  1.8498439 -1.04045502 -1.1865838   1.480297e-16
## 2 -0.2063280 -0.2481660 -0.3385541 -0.03813318  0.4069821   6.457718e-01
## 3  0.2600876 -0.7493205 -0.2817392  0.58367759  0.4107405  -6.150208e-01
## 4  1.6955811 -0.1780563 -0.1984582  1.23498791  1.3503431   1.153164e+00
## 5 -0.8261772  0.4775991 -0.3696184 -0.56315890 -0.8514589  -9.994088e-01
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.34435439 -0.5769454        -1.6095439
## 2 -0.42712134 -0.4707453         0.1531171
## 3 -0.02011273 -1.0931619         1.2300167
## 4 -0.46807818  0.4671788         0.5912425
## 5  0.85022014  0.9158889        -0.3319956
```

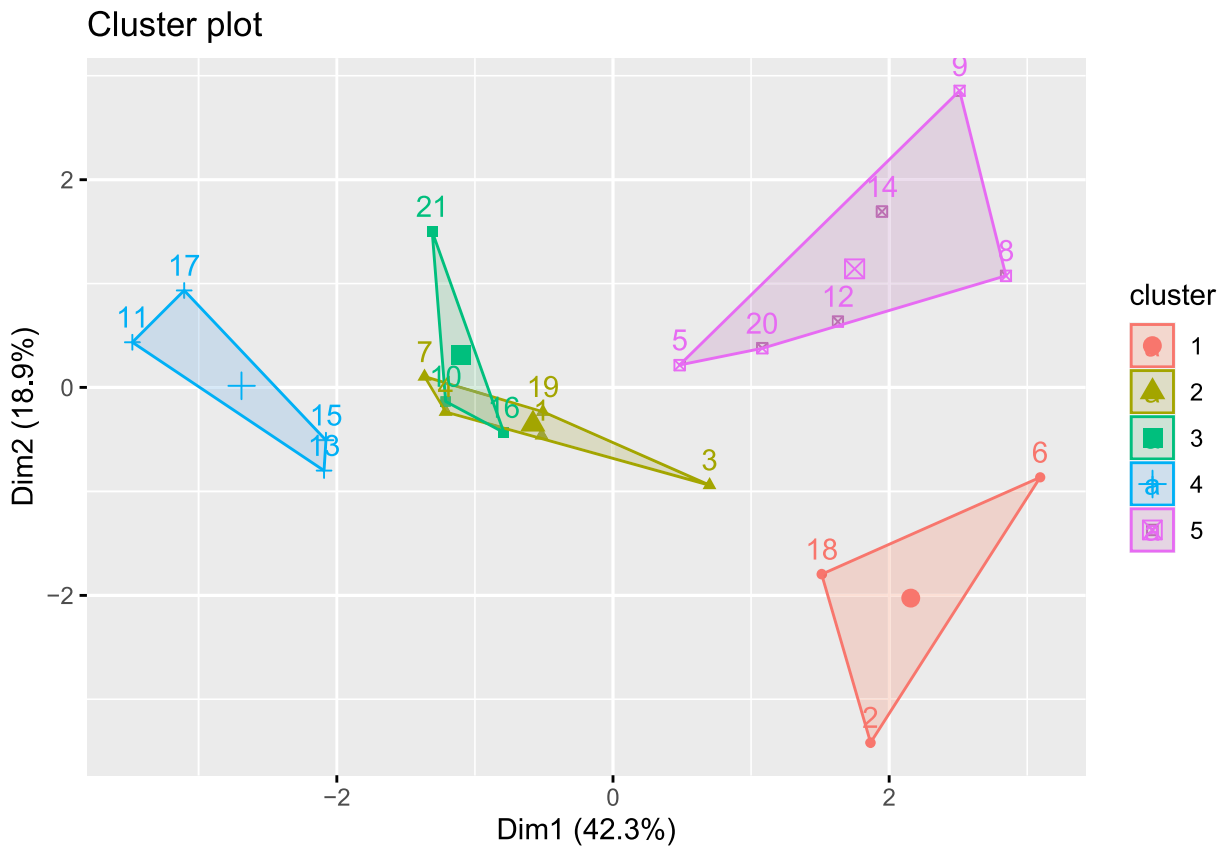
```
kmeans_model_sil$size # Number of companies in each cluster
```

```
## [1] 3 5 3 4 6
```

```
kmeans_model_sil$cluster[13] # Identify the cluster of the 13th observation as an example
```

```
## [1] 4
```

```
fviz_cluster(kmeans_model_sil, data = pharma_scaled) # Visualize the output
```



DBSCAN

```
library("dbscan")
```

```
## Warning: package 'dbscan' was built under R version 4.3.2
```

```
##
## Attaching package: 'dbscan'
```

```
## The following object is masked from 'package:stats':
##
##     as.dendrogram
```

```r
library("fpc")
```

```
## Warning: package 'fpc' was built under R version 4.3.2
```

```
##
## Attaching package: 'fpc'
```

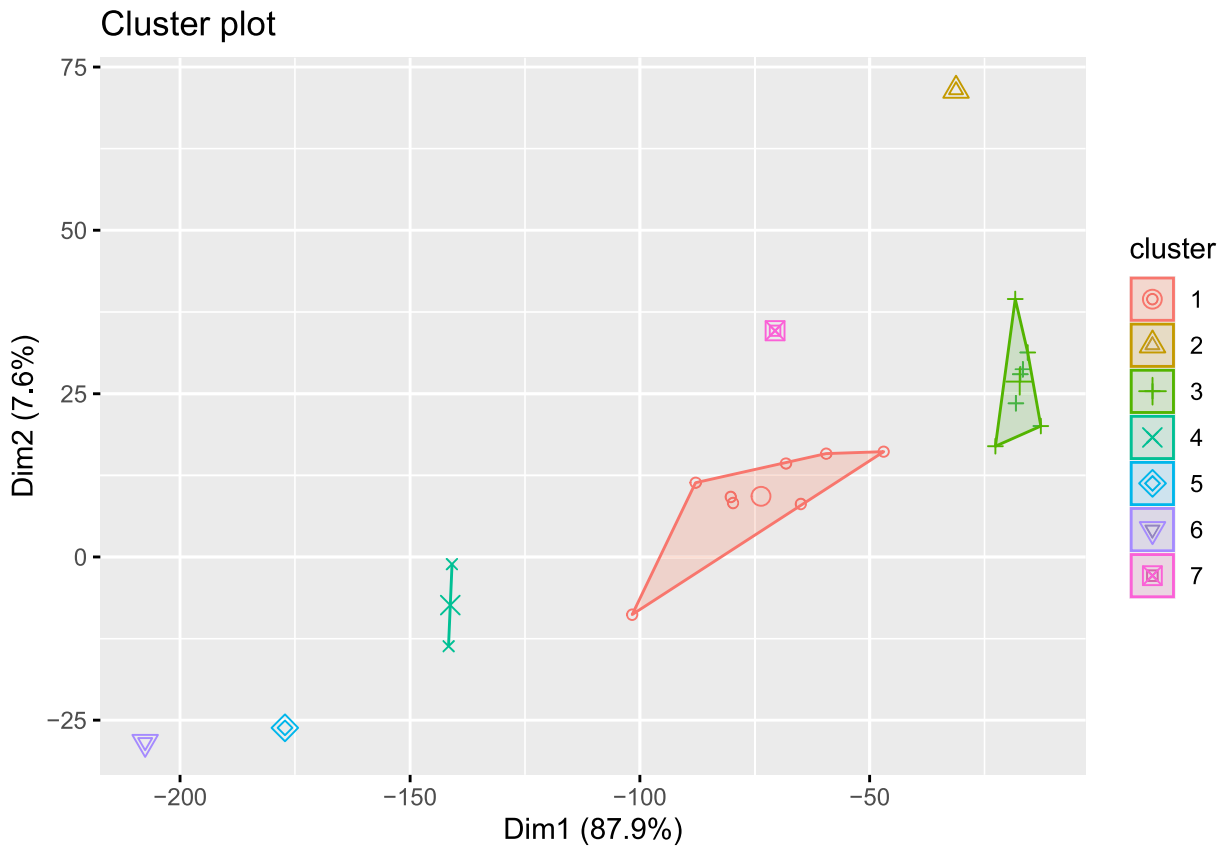
```
## The following object is masked from 'package:dbscan':
##
##     dbscan
```

```r
db <- dbscan::dbscan(pharma1, eps = 30 , minPts =1) #performing clustering

print(db) #printing cluster details
```

```
## DBSCAN clustering for 21 objects.
## Parameters: eps = 30, minPts = 1
## Using euclidean distances and borderpoints = TRUE
## The clustering contains 7 cluster(s) and 0 noise points.
##
## 1 2 3 4 5 6 7
## 8 1 7 2 1 1 1
##
## Available fields: cluster, eps, minPts, dist, borderPoints
```

```r
fviz_cluster(db, pharma1,   stand = FALSE, frame = FALSE, geom = "point")
```

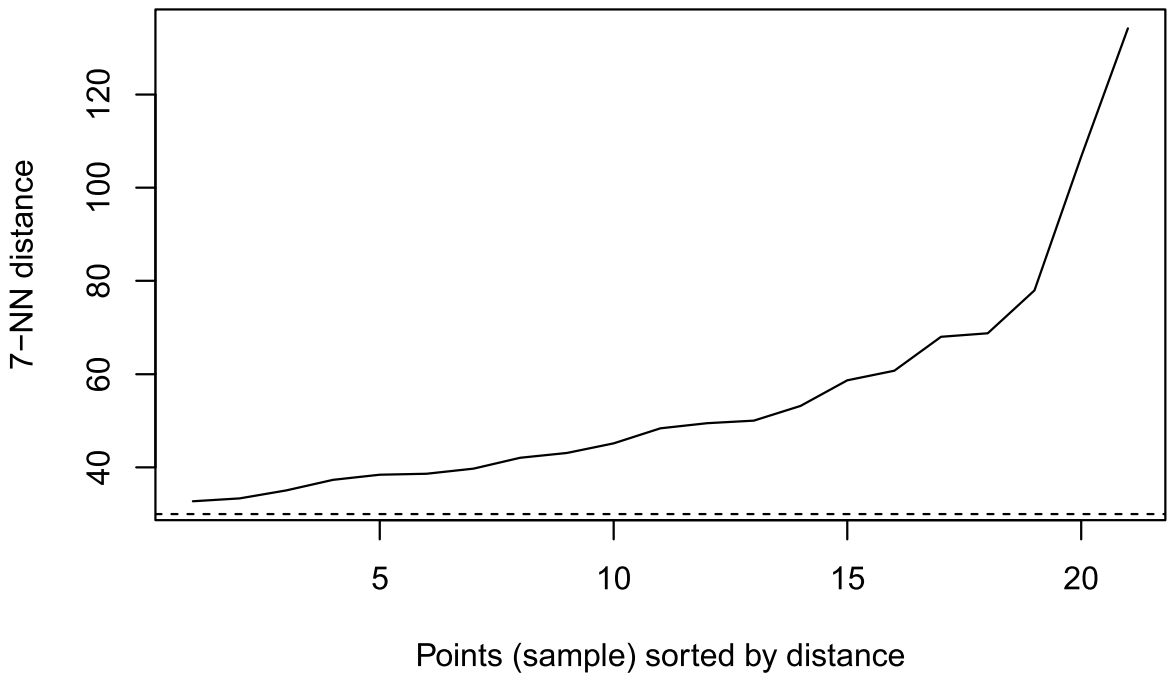
```
## Warning: argument frame is deprecated; please use ellipse instead.
```




```
df <- pharma1[, 1:9]

set.seed(321)
db <- fpc::dbscan(df, eps = 30, MinPts = 1) # DBSCAN using fpc package

print(db) # showing clusters' details
```

```
## dbscan Pts=21 MinPts=1 eps=30
##        1 2 3 4 5 6 7
## seed   8 1 7 2 1 1 1
## total  8 1 7 2 1 1 1
```

```
#Plotting the knee method graph for DBSCAN method
dbscan::kNNdistplot(df, k =  7)
abline(h = 30, lty = 2)
```



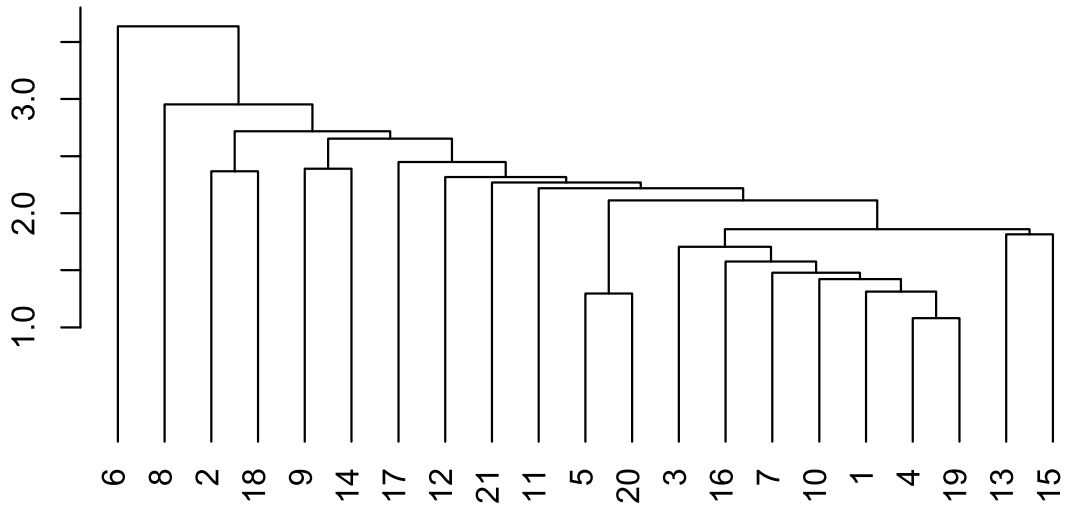
Hierarchical

Compute Euclidean distance

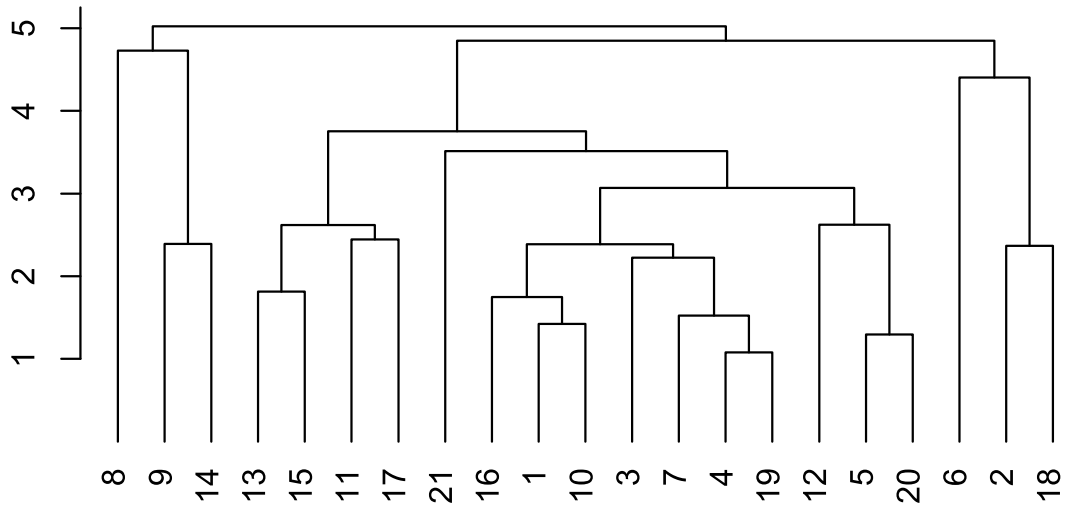
```
# (to compute other distance measures, change the value in method = )
d <- dist(pharma1, method = "euclidean")
d.norm <- dist(pharma1[,c(8,9)], method = "euclidean")
```

```
# normalizing input variables
Sorted.data.norm <- sapply(pharma1, scale)

# adding row names: utilities
row.names(Sorted.data.norm) <- row.names(pharma1)

# computing normalized distance based on variables ROA and Asset Turnover
d.norm <- dist(Sorted.data.norm[,c(8,9)], method = "euclidean")
```

```
# computing normalized distance based on all 9 variables
d.norm <- dist(Sorted.data.norm, method = "euclidean")

# in hclust() set argument method =
# to "ward.D", "single", "complete", "average", "median", or "centroid"
hc1 <- hclust(d.norm, method = "single")
plot(hc1, hang = -1, ann = FALSE)
```




```
hc2 <- hclust(d.norm, method = "average")
plot(hc2, hang = -1, ann = FALSE)
```



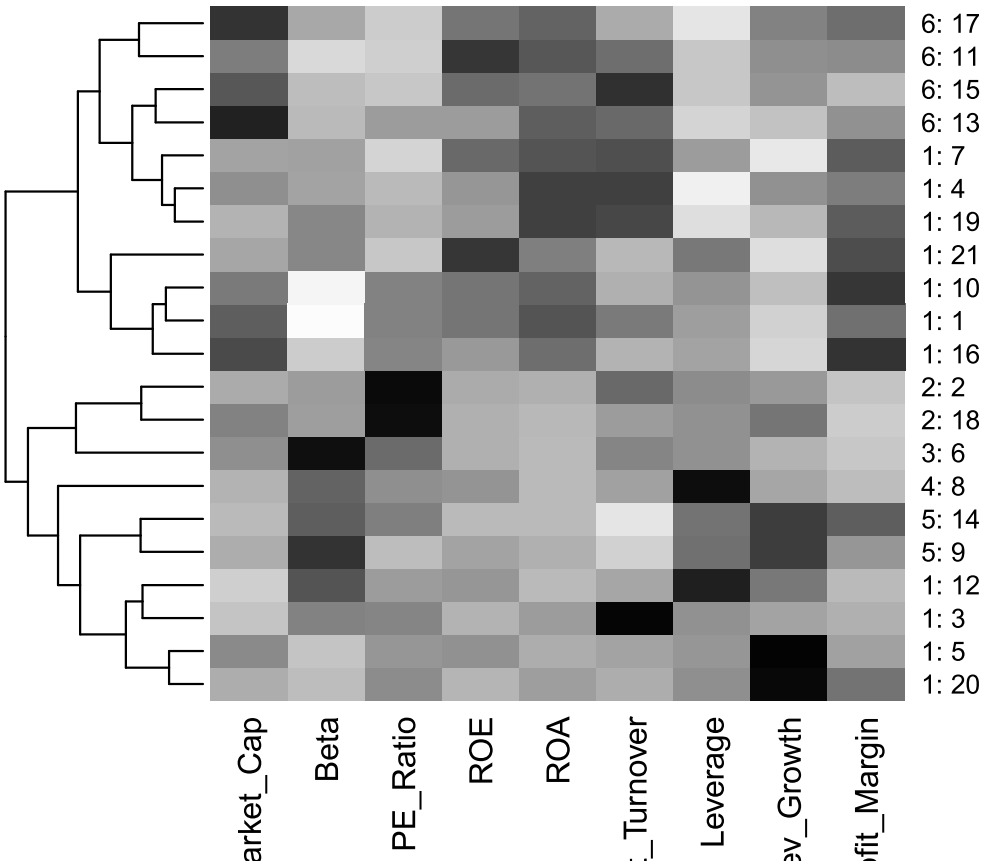

```
memb <- cutree(hc1, k = 6)
memb
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
##  1  2  1  1  1  3  1  4  5  1  1  1  1  5  1  1  6  2  1  1  1
```

```
memb <- cutree(hc2, k = 6)
memb
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
##  1  2  1  1  1  3  1  4  5  1  6  1  6  5  6  1  6  2  1  1  1
```

```
# setting labels as cluster membership and utility name
row.names(Sorted.data.norm) <- paste(memb, ": ", row.names(pharma1), sep = "")

# plotting heatmap
# rev() reverses the color mapping to large = dark
heatmap(as.matrix(Sorted.data.norm), Colv = NA, hclustfun = hclust,
        col=rev(paste("gray",1:99,sep="")))
```



Q.2.Interpret the clusters with respect to the numerical variables used in forming the clusters. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

```
library(dplyr)
set.seed(3)
p <- pharma_given[,c(12,13,14)]%>% mutate(clusters = kmeans_model_sil$cluster)
View(p)
```

```
# Loading required libraries
library(ggplot2)
library(gridExtra)
```

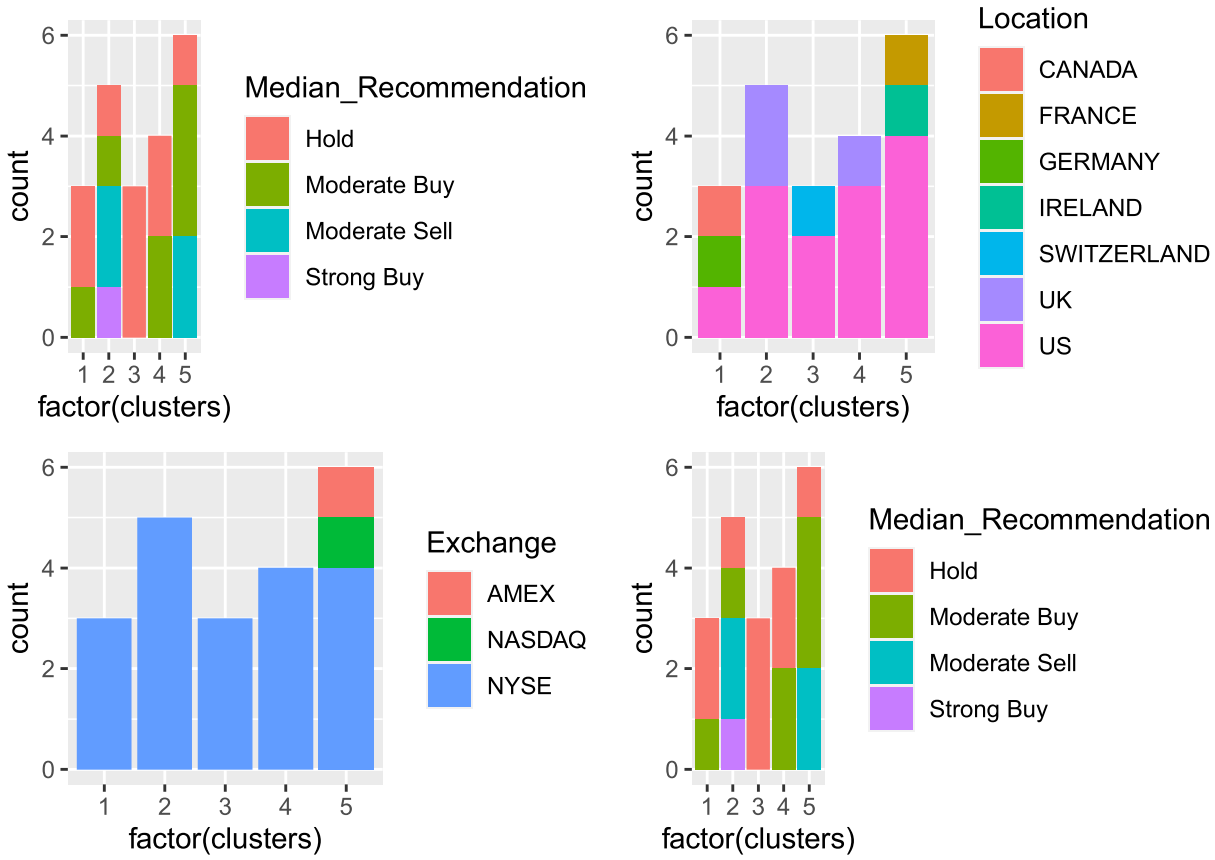
```
## Warning: package 'gridExtra' was built under R version 4.3.2
```

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

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

```
# Assuming "positive" is a valid geom_bar argument, otherwise replace it with the correct geom_bar func

plot1 <- ggplot(p, mapping = aes(factor(clusters), fill = Median_Recommendation)) + geom_bar(position =
```

```
plot2 <- ggplot(p, mapping = aes(factor(clusters), fill = Location)) + geom_bar(position = "stack")
plot3 <- ggplot(p, mapping = aes(factor(clusters), fill = Exchange)) + geom_bar(position = "stack")
plot4 <- ggplot(p, mapping = aes(factor(clusters), fill = Median_Recommendation)) + geom_bar(position =

# Combining the plots using grid.arrange
grid.arrange(plot1, plot2, plot3, plot4, ncol = 2)
```



Q.3.Provide an appropriate name for each cluster using any or all of the variables in the dataset.

```
#1 Cluster: In this cluster, which also has medians for Hold, Moderate Buy, Moderate Sell, and Strong B

#2 Cluster: Despite the fact that the companies are evenly distributed across the AMEX, NASDAQ, and NYS

#3 Cluster: listed on the NYSE, with separate counts for the United States, Ireland, and France, and mo

#4, Cluster: distributed throughout the United States and the United Kingdom and listed in, shares the

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

#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.
```

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