# Assignment_4

chandu

2023-11-12

***Summary***

**For this project, we are using the pharmaceuticals dataset, which comprises a brief profile of 21 pharmaceutical firms as well as data on nine critical performance criteria such as market capitalization and return on equity.***

***We ran a clustering analysis on the nine numerical columns of data for the first question. We're using the K clustering approach. This method pre-processes the data to guarantee that each variable contributes equally before computing the pairwise distances between observations and scaling the data. To determine the optimal number of clusters, the "fviz_nbclust" function was used. The Fviz_nbclust silhouette approach is useful for calculating the appropriate number of clusters. It assists us in determining how many clusters will optimize separation while minimize overlap by calculating silhouette scores for different cluster sizes. As a consequence, we discovered that five clusters are the best number to build. We then conducted a clustering analysis for K = 7. The total of square values within the cluster at K = 7 is 77.5%, while it is 65.4% at K = 5. We discovered that the best number was actually five clusters, because more clearly described clusters are typically indicated by lower WCSS values.below 7. During the clustering phase, the K means technique that we are using considers all variables equally. The kmeans' "centers" represent the means of all the variables inside each cluster. The cluster centroids are determined by combining these methods.***

#Certain patterns connect the non-numerical variables.. #In the second question, the numerical factors were used to group the clusters. Clusters 1 and 3 appear to be more "moderate" in certain ways, whereas Clusters 4 and 5 appear to be more "extreme". It emerges. #Clusters 1 and 3 are "moderate" in the following ways: Their growth rate is slow. #Cluster 1 has a high valuation and profitability, but Cluster 3 has a high PE but a lower profit. However, slower expansion).Their recommendations are more in the center of buying and selling than on the road (buy moderately or keep).Their developed market bases include the United States and the United Kingdom, both of which are listed on major exchanges (NYSE).. #Clusters 4 and 5 appear to be more "extreme": Cluster 4 is quickly expanding while maintaining a high level of quality #risk (low PE, large leverage, and mild sell suggestions).In Cluster 5, distressed stocks have substantial poor growth, excessive beta volatility, and leverage

```
#Load the required libraries
library(tidyverse)
```

```
## -- 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   3.4.3     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)
```

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

```
library(dplyr)
set.seed(11)
```

```
#Loading the dataset
pharma_data <- read.csv("Pharmaceuticals (1).csv")
```

#Question 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.

```
# Set row names as the company name column, select columns required for clustering (3 through 11) and d
pharma_data<- pharma_data
row.names(pharma_data) <- pharma_data[,2]
pharma_data <- pharma_data[,c(3,4,5,6,7,8,9,10,11)]
row.names(pharma_data)
```

```
##  [1] "Abbott Laboratories"               "Allergan, Inc."
##  [3] "Amersham plc"                      "AstraZeneca PLC"
##  [5] "Aventis"                           "Bayer AG"
##  [7] "Bristol-Myers Squibb Company"      "Chattem, Inc"
##  [9] "Elan Corporation, plc"             "Eli Lilly and Company"
## [11] "GlaxoSmithKline plc"               "IVAX Corporation"
## [13] "Johnson & Johnson"                 "Medicis Pharmaceutical Corporation"
## [15] "Merck & Co., Inc."                 "Novartis AG"
## [17] "Pfizer Inc"                        "Pharmacia Corporation"
## [19] "Schering-Plough Corporation"       "Watson Pharmaceuticals, Inc."
## [21] "Wyeth"
```

```
summary(pharma_data)
```

```
##    Market_Cap          Beta           PE_Ratio          ROE
##  Min.   :  0.41   Min.   :0.1800   Min.   : 3.60   Min.   : 3.9
##  1st Qu.:  6.30   1st Qu.:0.3500   1st Qu.:18.90   1st Qu.:14.9
##  Median : 48.19   Median :0.4600   Median :21.50   Median :22.6
##  Mean   : 57.65   Mean   :0.5257   Mean   :25.46   Mean   :25.8
##  3rd Qu.: 73.84   3rd Qu.:0.6500   3rd Qu.:27.90   3rd Qu.:31.0
##  Max.   :199.47   Max.   :1.1100   Max.   :82.50   Max.   :62.9
##       ROA        Asset_Turnover    Leverage        Rev_Growth
##  Min.   : 1.40   Min.   :0.3    Min.   :0.0000   Min.   :-3.17
##  1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600   1st Qu.: 6.38
##  Median :11.20   Median :0.6    Median :0.3400   Median : 9.37
##  Mean   :10.51   Mean   :0.7    Mean   :0.5857   Mean   :13.37
##  3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000   3rd Qu.:21.87
##  Max.   :20.30   Max.   :1.1    Max.   :3.5100   Max.   :34.21
```

```
##  Net_Profit_Margin
##  Min.   : 2.6
##  1st Qu.:11.2
##  Median :16.1
##  Mean   :15.7
##  3rd Qu.:21.1
##  Max.   :25.5
```

## Reason for Choosing the 9 variable that is Market_Cap, Beta, PE_Ratio, ROE, ROA, Asset_Turnover, Leverage, Rev_Growth, and Net_Profit_Margin

#The range and importance of each variable can be used to support the variables I chose for my analysis, which was based on the summary with min and max values shown above. Let's examine each variable in turn and talk about why its minimum and maximum values could be used to choose it:

## i) Market_Cap (Market Capitalization): Min Value: 0.41, Max Value: 199.47

#Reasoning: A company's outstanding shares' total market value is represented by its market capitalization. Analyzing this variable might reveal information about the market positioning and scale of the companies, as evidenced by the vast range of values from 0.41 to 199.47, which represents varied company sizes.

## ii) Beta: Min Value: 0.18, Max Value: 1.11

#Reasoning: The volatility of a stock is gauged by its beta value in relation to the market. The range of 0.18 to 1.11 indicates that the companies' risk profiles vary from one another. Investors evaluating the risk-return trade-off in their investments need to understand beta.

## iii) PE_Ratio (Price-to-Earnings Ratio): Min Value: 3.60, Max Value: 82.50

#Reasoning: The PE ratio shows how much the market is willing to pay for a company's shares. It is possible to find both overvalued and undervalued stocks by evaluating this variable, since it shows a broad range of valuation values from 3.60 to 82.50.

## iv) ROE (Return on Equity): Min Value: 3.9, Max Value: 62.9

#Reasoning: ROE measures a company's ability to generate profit from shareholders' equity. The range from 3.9 to 62.9 suggests varying levels of profitability and efficiency in utilizing equity, making it an essential metric for assessing financial health.

## v) ROA (Return on Assets): Min Value: 1.40, Max Value: 20.30

ROA assesses how well a business makes money off of its assets. Different levels of asset efficiency are shown by the range from 1.40 to 20.30, which provides information about how successfully businesses turn their assets into profits.

### vi) Asset_Turnover: Min Value: 0.3, Max Value: 1.1

Asset turnover measures a company's ability to generate sales from its assets. The range from 0.3 to 1.1 indicates varying efficiency in utilizing assets to generate revenue, making it valuable for assessing operational efficiency.

### vii) Leverage: Min Value: 0.0000, Max Value: 3.5100

The amount of debt a firm utilizes as part of its capital structure is reflected in its leverage. Financial leverage levels vary, as seen by the range from 0.0000 to 3.5100, which provides information on risk and capital structure choices.

### viii) Rev_Growth (Revenue Growth): Min Value: -3.17, Max Value: 34.21

The percentage change in revenue is measured as revenue growth. Diverse degrees of revenue expansion or contraction are indicated by the wide range from -3.17 to 34.21, which sheds light on business dynamics.

### ix) Net_Profit_Margin: Min Value: 2.6, Max Value: 25.5

#The percentage of revenue that is converted into profit is known as the net profit margin. The range of 2.6 to 25.5 represents a range of profitability, which is important to consider when evaluating the financial health of a company.

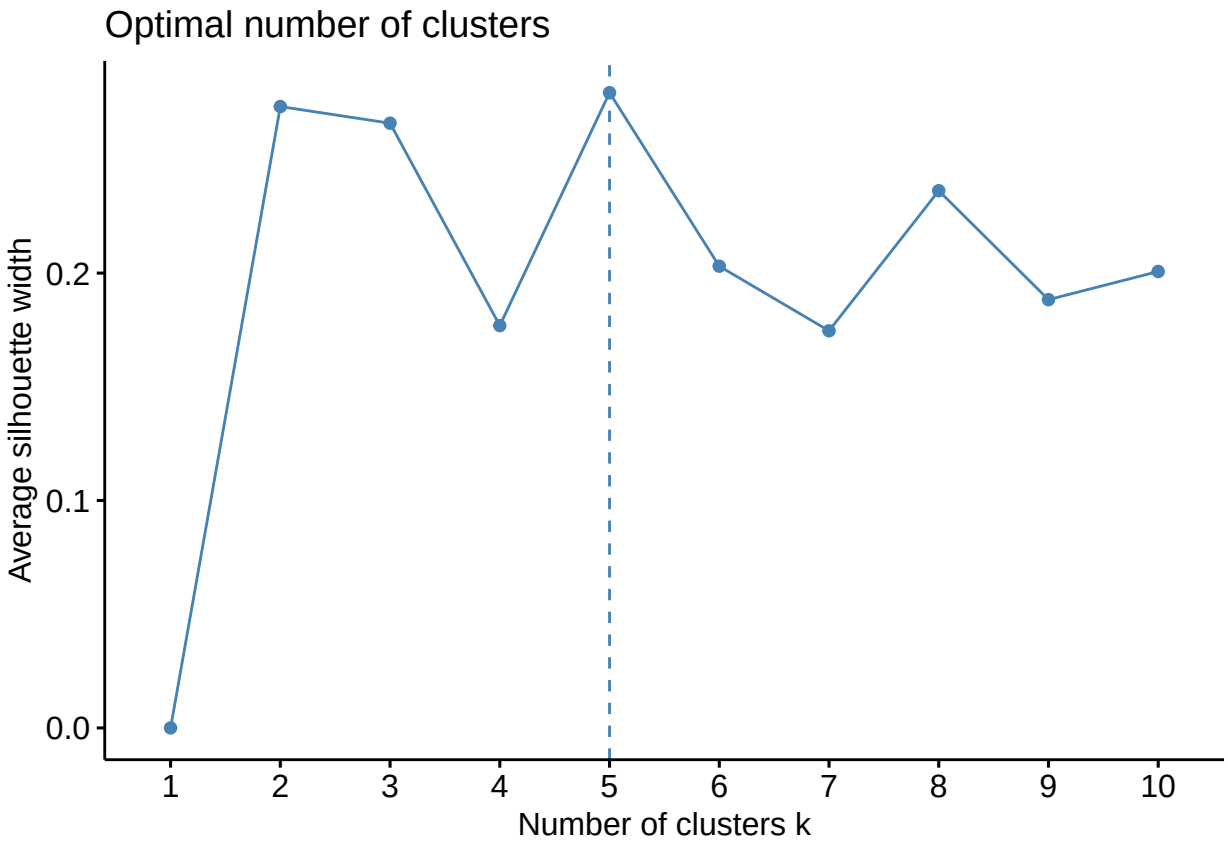
### Now let's Normalize the data

```
pharma_data <- scale(pharma_data)
row.names(pharma_data) <- pharma_data[,1]
dist <- get_dist(pharma_data)
corr <- cor(pharma_data)
```

### Reason for Normalization:

#To guarantee that each variable contributes proportionately to the clustering process, normalization of the numerical variables is essential. Normalizing these variables helps stop one variable from controlling the clustering based solely on their magnitude because they may have different scales or units. For instance, Beta is a fraction between 0 and 1, whereas Market Cap is in the hundreds.

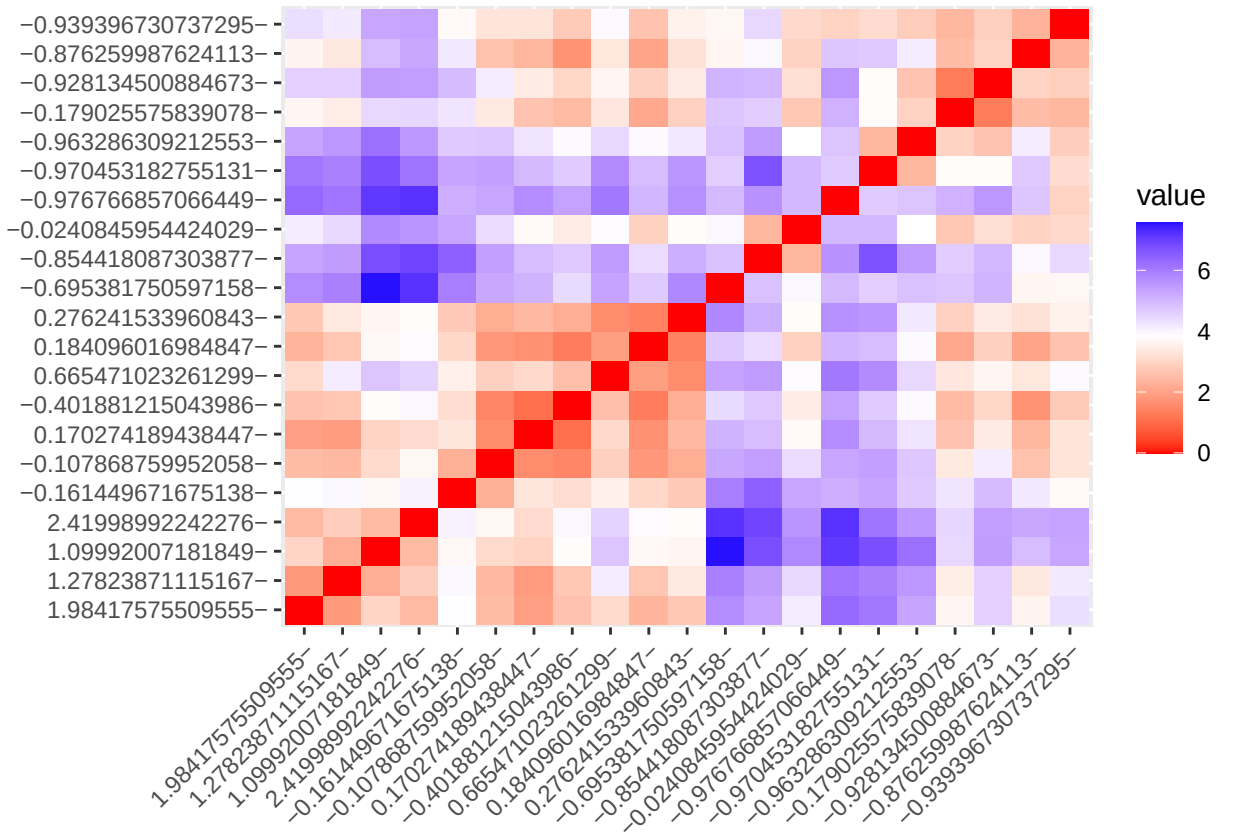
```
fviz_nbclust(pharma_data, kmeans, method = "silhouette")
```



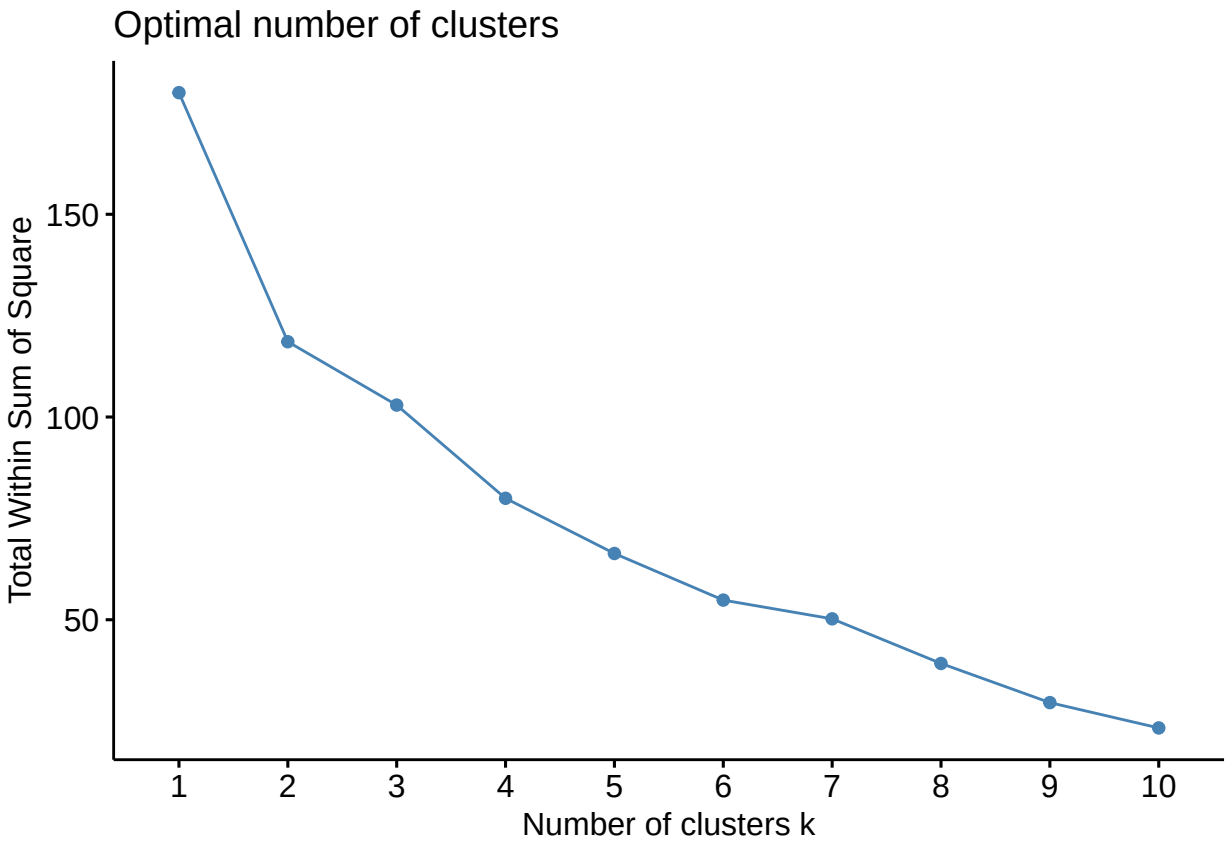

### Explanation for selecting 5 clusters

The net profit margin is the proportion of revenue that is turned into profit. A range of profitability, which is crucial to take into account when assessing a company's financial health, is between 2.6 and 25.5.

```
# I'm going to create my first k-means clustering algorithm using the Euclidean distance since it is th
pharma_data <- scale(pharma_data)
distance <- get_dist(pharma_data)
fviz_dist(distance)
```

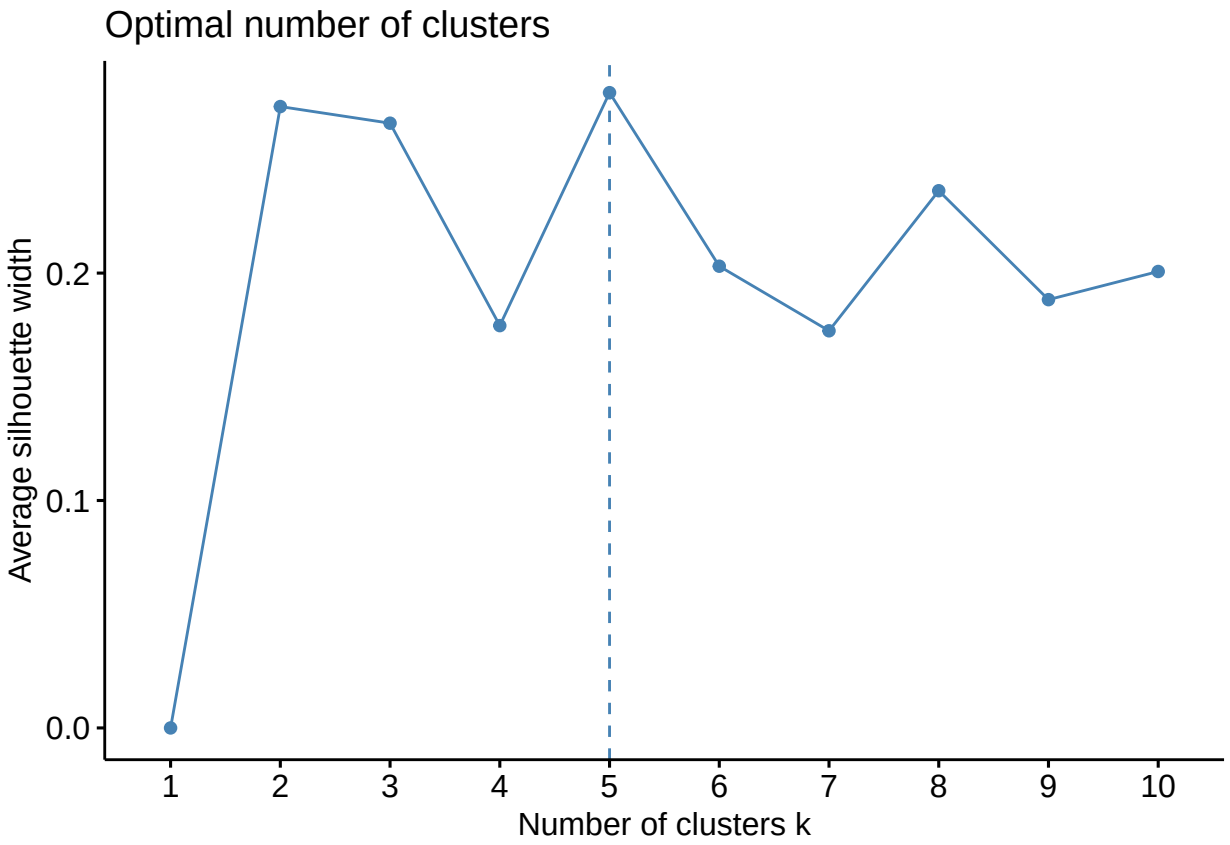



```
# Determine the best value for k using an "elbow" chart
fviz_nbclust(pharma_data, kmeans, method = "wss")
```



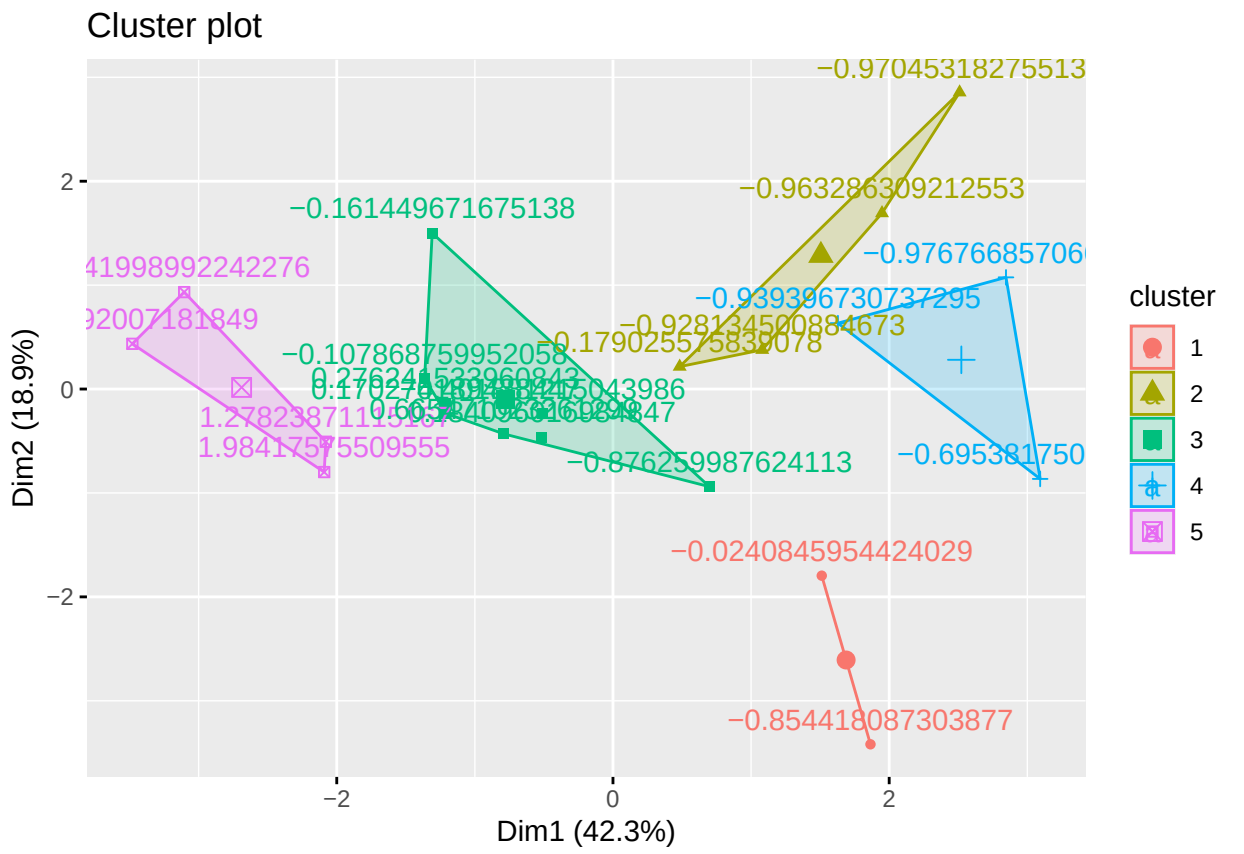

# The output above displays that around 5 - 6 is the ideal value for k (slope stops being as steep)

```
#Determine the best value for k using the Silhouette Method; compare to "elbow" chart results
fviz_nbclust(pharma_data, kmeans, method = "silhouette")
```




#The silhouette chart displays that 5 is the ideal value for k. I'm selecting k = 5 given both charts show it as an optimal value. #Run k-means using k = 5, number of restarts = 25

```
k5 <- kmeans(pharma_data, centers = 5, nstart = 25)
fviz_cluster(k5, data = pharma_data)
```




```
k5$cluster
```

```
##   0.184096016984847  -0.854418087303877  -0.876259987624113   0.170274189438447
##                   3                   1                   3                   3
##  -0.179025575839078  -0.695381750597158  -0.107868759952058  -0.976766857066449
##                   2                   4                   3                   4
##  -0.970453182755131   0.276241533960843   1.09992007181849  -0.939396730737295
##                   2                   3                   5                   4
##    1.98417575509555  -0.963286309212553   1.27823871115167   0.665471023261299
##                   5                   2                   5                   3
##    2.41998992242276 -0.0240845954424029  -0.401881215043986  -0.928134500884673
##                   5                   1                   3                   2
##  -0.161449671675138
##                   3
```

```
# Display the centroids
k5$centers
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 2 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 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 -0.14170336 -0.1168459      -1.416514761
## 2  0.06308085  1.5180158      -0.006893899
## 3 -0.27449312 -0.7041516       0.556954446
## 4  1.36644699 -0.6912914      -1.320000179
## 5 -0.46807818  0.4671788       0.591242521
```

```
# Display the size of each clusters
k5$size
```

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

**In financial analysis, particularly when evaluating a firm, these variables are often categorized into groups of financial ratios. For instance, ratios such as Return on Assets (ROA) and Net Profit Margin fall under the classification of profitability ratios since they derive from analogous figures extracted from a balance sheet or income statement, potentially leading to correlations between them.**

**summary statistics reveals there may be outliers in the dataset. Given the Euclidean distance is sensitive to outliers and ignores correlation, I'm going to cluster the data again using another distance to see the output. I've chosen the Manhattan Distance for this exercise.**

**Run k-means again using k = 5 (based on previous "elbow" and silhouette methods) using the Manhattan Distance.**

```
library(flexclust)
```

```
## Loading required package: grid
```

```
## Loading required package: lattice
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

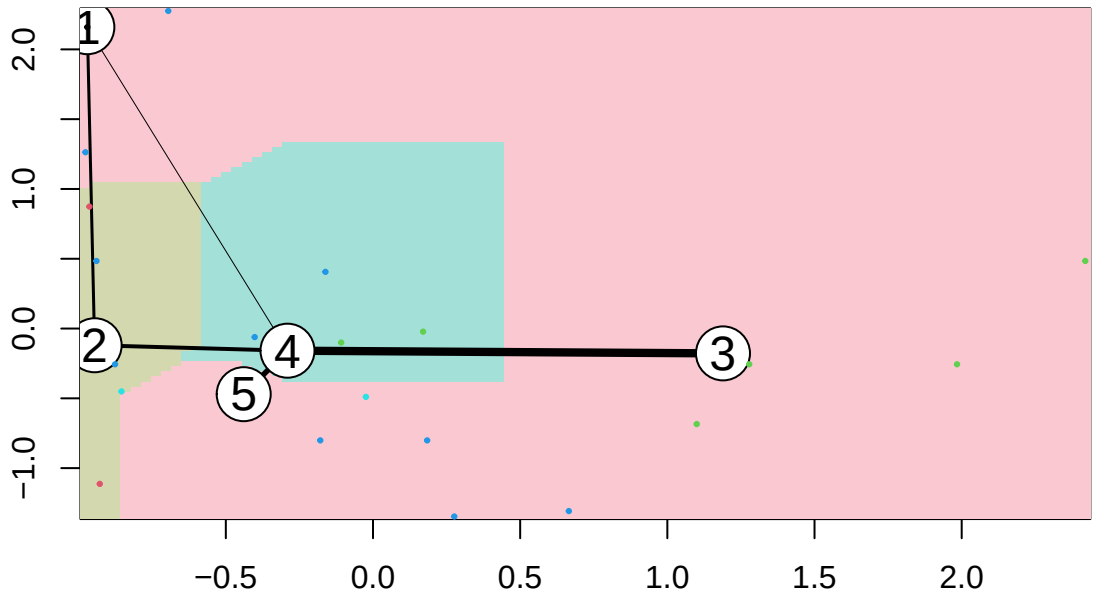
```
set.seed(101)
km5 = kcca(pharma_data, k=5, kccaFamily("kmedians"))
km5
```

```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = pharma_data, k = 5, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
##  1  2  3  4  5
##  1  2  6 10  2
```

```r
# Apply predict function
clusters_index <- predict(km5)
dist(km5@centers)
```

```
##          1        2        3        4
## 2 2.854951
## 3 5.461711 4.268466
## 4 4.100941 2.649994 2.810260
## 5 5.792248 4.097340 4.775806 3.471731
```

```r
image(km5)
points(pharma_data, col= clusters_index, pch= 19, cex=0.3)
```

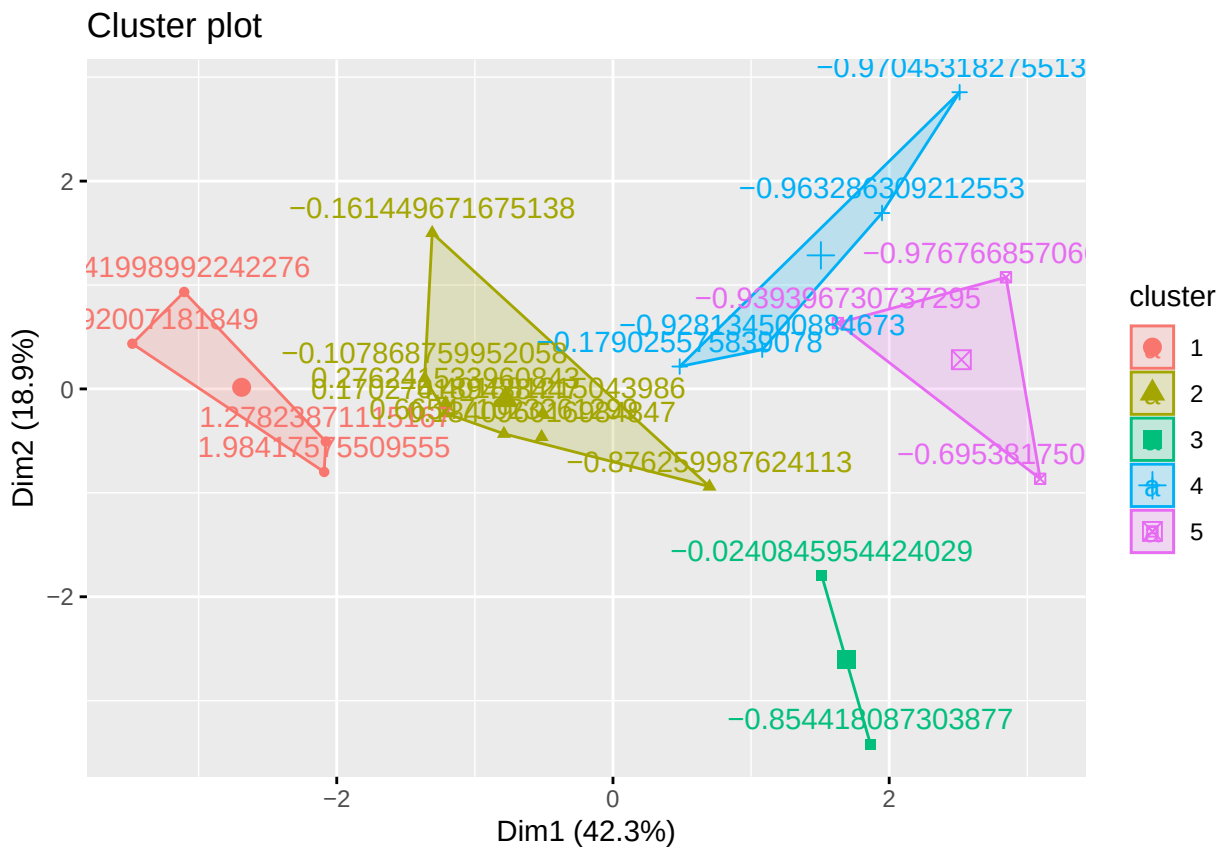



##The utilization of the k-means algorithm with the Manhattan distance distinctly yields a divergent clustering outcome, with the majority of the dataset now grouped into two clusters, in contrast to a singular cluster when utilizing the Euclidean distance. Furthermore, noteworthy is the observation that one of the clusters in this result encompasses only a single data point.

##Upon inspecting the clustering diagram above, it is evident that the data points exhibit less tight cohesion within the clusters compared to when the Euclidean distance was employed. Moreover, the identification of one cluster containing only a single data point raises questions about the suitability of the chosen number of clusters when utilizing the Manhattan distance. However, for the purposes of this analysis, I will proceed with k = 5, as determined to be optimal through the elbow and Silhouette methods. Despite both Euclidean and Manhattan distance methods being considered, I lean towards the Euclidean distance approach as the clusters appear more compact when utilizing the "optimal" value of k = 5.

**Display k-means algorithm using k = 5 and Euclidean distance again.**

```
k5 <- kmeans(pharma_data, centers = 5, nstart = 25)
fviz_cluster(k5, data = pharma_data)
```




```
k5$cluster
```

```
##   0.184096016984847  -0.854418087303877  -0.876259987624113   0.170274189438447
##                   2                   3                   2                   2
##  -0.179025575839078  -0.695381750597158  -0.107868759952058  -0.976766857066449
##                   4                   5                   2                   5
##  -0.970453182755131   0.276241533960843   1.09992007181849  -0.939396730737295
##                   4                   2                   1                   5
##    1.98417575509555  -0.963286309212553   1.27823871115167   0.665471023261299
##                   1                   4                   1                   2
##    2.41998992242276 -0.0240845954424029  -0.401881215043986  -0.928134500884673
##                   1                   3                   2                   4
##  -0.161449671675138
##                   2
```

```
clusterindex = predict(km5)

dist(km5@centers)
```

```
##          1        2        3        4
## 2 2.854951
## 3 5.461711 4.268466
## 4 4.100941 2.649994 2.810260
## 5 5.792248 4.097340 4.775806 3.471731
```

##Question 2- Interpret the clusters with respect to the numerical variables used in forming the clusters.

***ANS/:Cluster 1 is characterized by high market capital, high ROE, high ROA, and high asset turnover.***

***Cluster 2 is the largest cluster and is characterized by average market capital, beta, price/earnings ratio, average to above-average ROA, and above average net profit margin. It also represents a relatively wide range of ROE, leverage, and estimated revenue growth values.***

***Cluster 3 is characterized by similar beta values, high price/earnings ratio, and low ROE ROA, net profit margin.***

***Cluster 4 is characterized by below average ROE, ROA, and asset turnover with high estimated revenue growth.***

***Cluster 5 is characterized by low market capital, ROA, asset turnover, estimated revenue growth, net profit margin. It also represents high beta and extreme leverage (high or low).***

---

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

***ANS/:Cluster 1 : Hold/Buy these prominent companies (high ROE & ROA)***

***Cluster: The average mix with high net profit margin***

***Cluster: High price/earnings; low ROE, ROA, & net profit margin***

***Cluster 4: Global mix with low ROE, ROA, & asset turnover but high estimated revenue growth.***

***Cluster 5 : Unique stock exchange mix with mostly low variables; extreme beta & leverage.***