Assignment_4

Balamanoj Reddy Kommareddy

2023-11-07

Summary:

1.Missing Values Check: determining the percentage of missing values in each column of our dataset.

• Normalization: Using the scale function to normalize the data. Variables on the same scale are required for K-Means clustering.

• Finding Optimal K: Using the Elbow technique (wss) and the Silhouette approach, determine the optimal number of clusters. Whereas the ideal k value for the wss method is k = 2, the optimal k value for the Silhouette approach is k = 5.

2.K-Means Clustering:

• The Within-Sum-of-Squares (WSS) approach is used for K-Means clustering with k = 2. The nstart parameter allows the algorithm to be performed numerous times with various initial centroids to prevent local minima. Within-cluster sum of squares by cluster = 43.3, 75.2, and between-cluster proportion (between_SS / total_SS) = 34.1%.

• The Silhouette approach is used for K-Means clustering with k = 5, providing a more detailed view of the cluster structure. nstart, like the WSS approach, is used to improve the robustness of the outcomes.Within-Cluster Sum of Squares (between_SS / total_SS = 65.4%): 12.79, 2.8, 15.595925, 21.879320, 9.284424 & Between-Cluster Proportion (between_SS / total_SS = 65.4%)

Cluster Plot Visualizations:

• Using the wss approach, a cluster plot for K-Means findings with k = 2 creates two clusters of size 11 and 10.

• Using the Silhouette approach, a cluster plot for K-Means findings with k = 5 generates 5 clusters of size 3, 2, 8, 4, and 4.

• WSS - Cluster 1 and Cluster 2 appear to follow a trend in terms of pharmaceutical firm location. “US” is the location of more than half of the enterprises in both clusters. This also implies that the United implies has enterprises that are both lucrative to invest in (Acceptable Profitability with Moderate Risk) and firms that are not profitable (Low Profitability with High Risk). However, the better performing cluster, Cluster 1, appears to have a higher proportion of enterprises based in the United States.

• silhouette - In the silhouette clusters, we can see the same level of pattern towards the place that we saw in the wss. Every cluster in here has a higher proportion of its locations in “US” than the other locations. However, it is important to see that the best cluster that truly characterizes the domain, Cluster 4, has a higher proportion of US enterprises and a lower proportion of non-US based companies.

3.

WSS –

1. Acceptable Profitability with Moderate Risk
2. Low Profitability with High Risk:

Silhouette-

1. Emerging Group
2. Overvalued and High-Risk Investment Group
3. High-Risk Investment
4. Promising Value opportunity Group
5. Prime Investment with Slighter Risk Group

```
#install.packages("factoextra")
library("tidyverse") #Loading library tidyverse for transforming data
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library("factoextra") #Loading factoextra library for extracting and visualizing the data
```

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

```
library("ggplot2")
library("dplyr")
library("esquisse")
```

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

#Loading and exploring the data

```
Pharmaceuticals <- read.csv("Pharmaceuticals.csv")
head(Pharmaceuticals)
```

```
##   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
```

```
summary(Pharmaceuticals)
```

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

#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.

Removing missing data and rescale variables for comparability before clustering data.

```
colMeans(is.na(Pharmaceuticals))
```

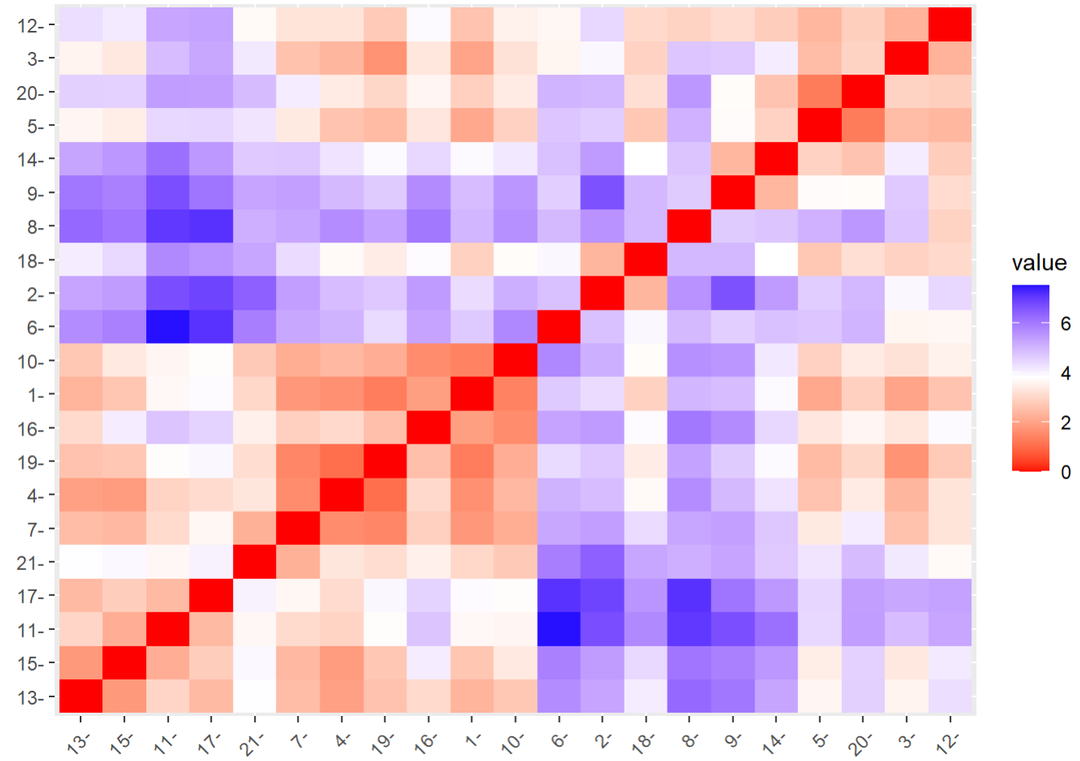
```
##               Symbol                 Name           Market_Cap
##                    0                    0                    0
##                 Beta             PE_Ratio                  ROE
##                    0                    0                    0
##                  ROA       Asset_Turnover             Leverage
##                    0                    0                    0
##           Rev_Growth     Net_Profit_Margin Median_Recommendation
##                    0                    0                    0
##             Location             Exchange
##                    0                    0
```

Performing z-score scaling Normalization

```
set.seed(1)
data.norm <- scale(Pharmaceuticals[,-c(1:2,12:14)])
```

calculating the distance of the scaled pharmaceuticals data

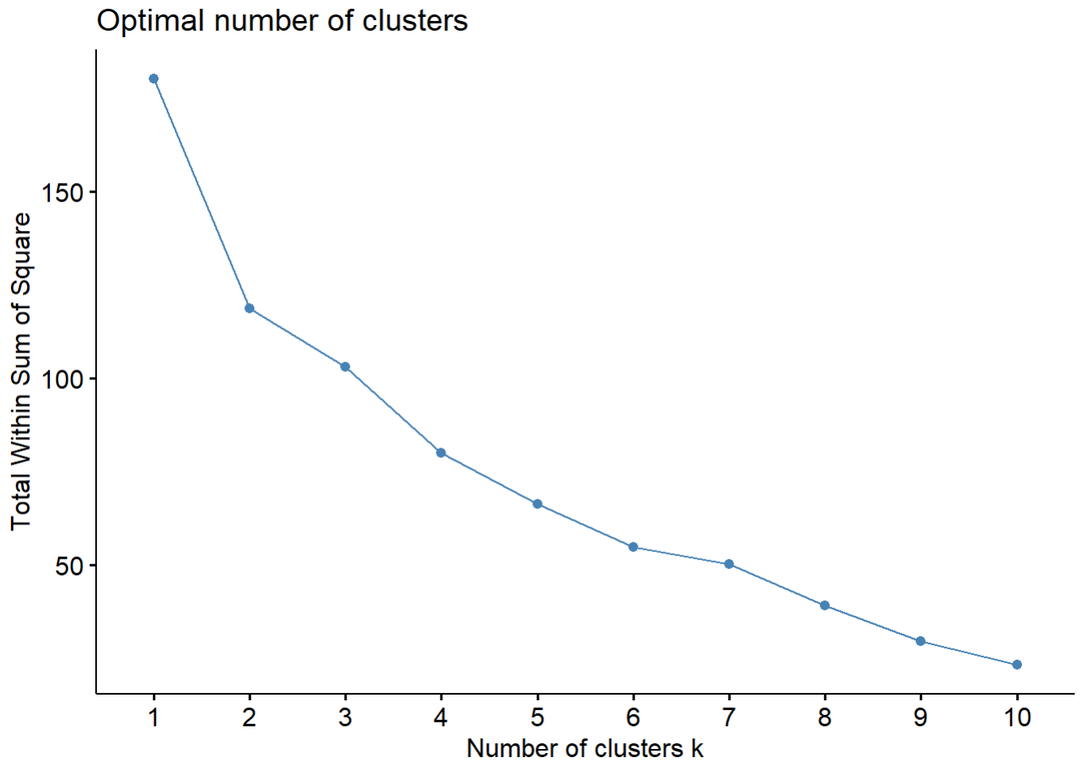
```
Distance <- dist(data.norm, method = "euclidian")
# visualizing the distance between rows of the distance matrix
fviz_dist(Distance)
```



Finding optimal K using wss method

Determining the no of clusters to do the cluster analysis using Elbow Method

```
wss <- fviz_nbclust(data.norm,kmeans,method="wss")
wss
```

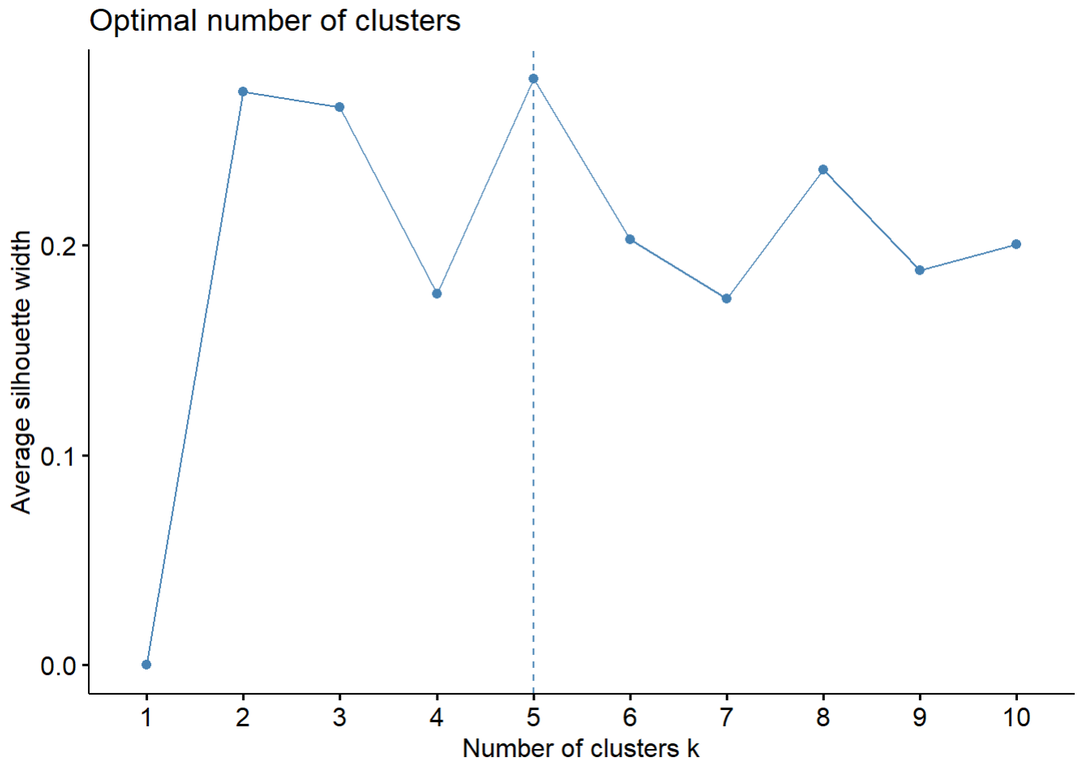


Here in this plot we can clearly see that the graph is forming an elbow shape at 2, The optimal number of clusters (k) determined through the Within-Sum-of-Squares (WSS) method is 2.

Finding optimal K using silhouette method

Using Silhouette method for determining no of clusters

```
silhouette_k <- fviz_nbclust(data.norm,kmeans,method="silhouette")
silhouette_k
```



The optimal number of clusters (k) determined through the silhouette method is 5.

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

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

```
wss_kmeans <- kmeans(data.norm,centers = 2,nstart=25)
wss_kmeans
```

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

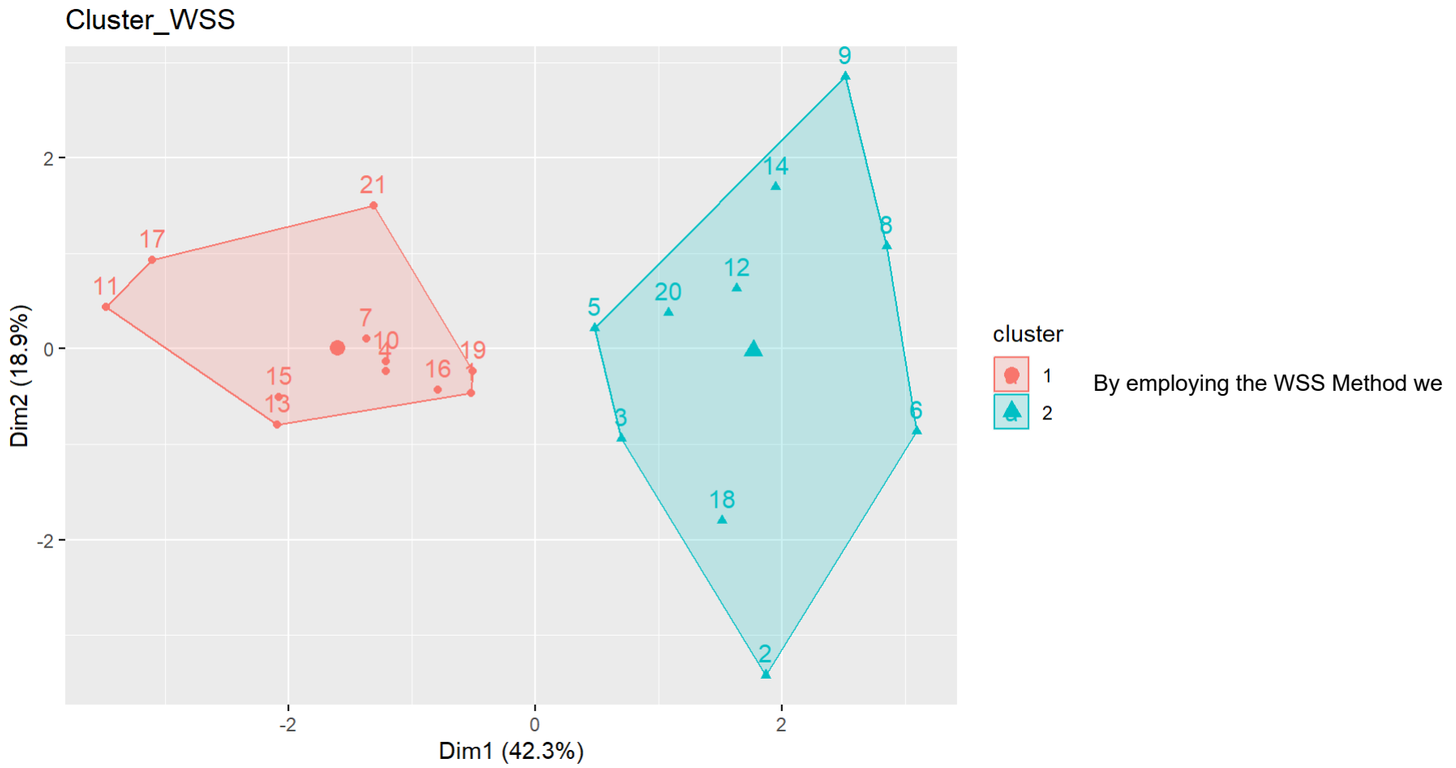
Formulation of clusters using K-Means with k = 5 (Silhouette)

```
silhouette_kmeans <- kmeans(data.norm,centers=5,nstart=25)
silhouette_kmeans
```

```
## K-means clustering with 5 clusters of sizes 4, 2, 3, 8, 4
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 2 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 3 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 4 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
##      Leverage Rev_Growth Net_Profit_Margin
## 1  0.06308085  1.5180158      -0.006893899
## 2 -0.14170336 -0.1168459      -1.416514761
## 3  1.36644699 -0.6912914      -1.320000179
## 4 -0.27449312 -0.7041516       0.556954446
## 5 -0.46807818  0.4671788       0.591242521
##
## Clustering vector:
##  [1] 4 2 4 4 1 3 4 3 1 4 5 3 5 1 5 4 5 2 4 1 4
##
## Within cluster sum of squares by cluster:
## [1] 12.791257  2.803505 15.595925 21.879320  9.284424
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Cluster Plot Visualizations for k=2 (WSS)

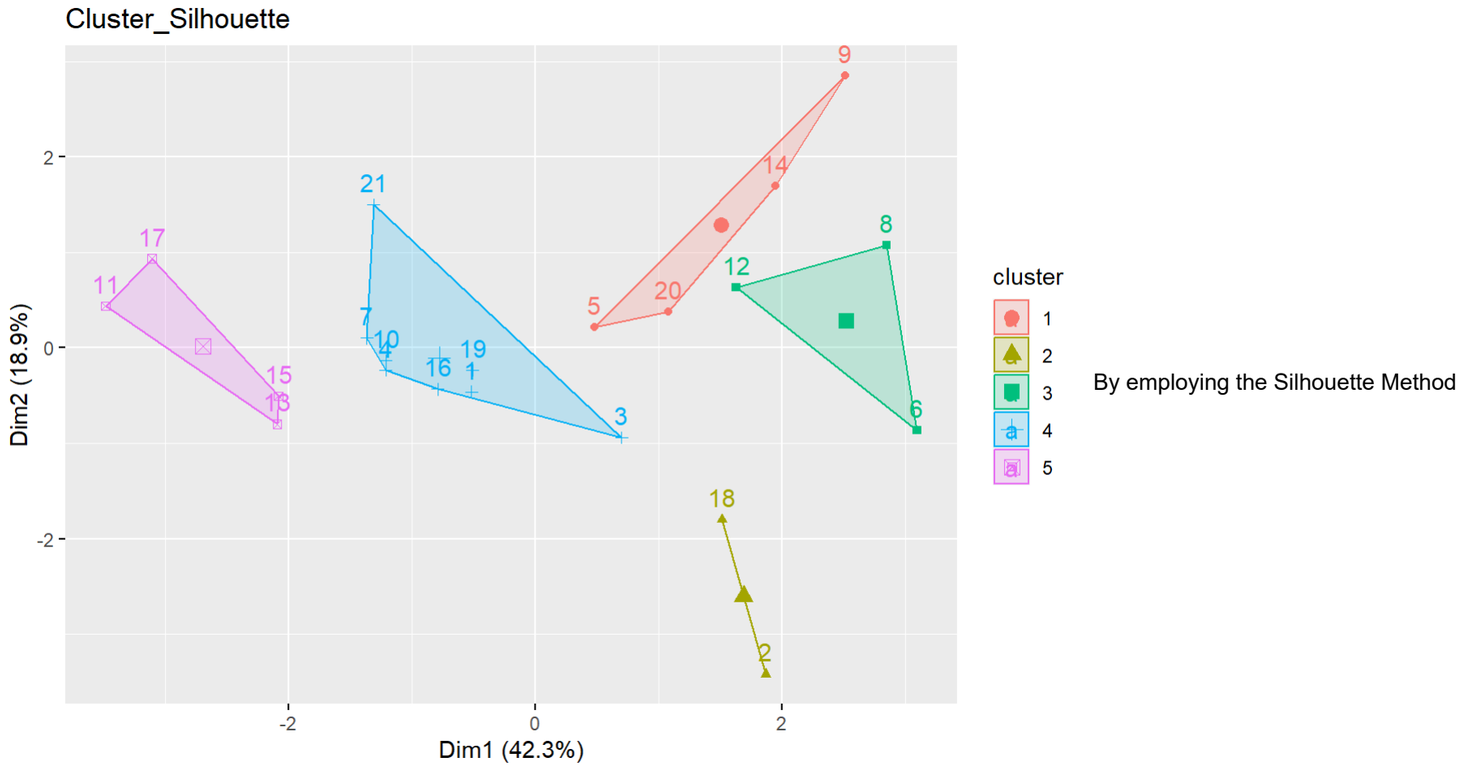
```
fviz_cluster(wss_kmeans,Pharmaceuticals[,-c(1:2,12:15)],main="Cluster_WSS")
```



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

Cluster Plot Visualizations for k=5 (Silhouette)

```
fviz_cluster(silhouette_kmeans,Pharmaceuticals[,-c(1:2,12:15)],main="Cluster_Silhouette")
```



By employing the Silhouette Method

we get 5 clusters of size 3, 2, 8, 4 and 4.

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

Binding the cluster assignment to the original data frame for analysis

```
clusters_wss <- wss_kmeans$cluster
clusters_silhouette <- silhouette_kmeans$cluster

data.1 <- cbind(Pharmaceuticals,clusters_wss)
data.2 <- cbind(Pharmaceuticals,clusters_silhouette)
```

Aggregating the clusters to interpret the attributes - WSS

```
int_wss <- aggregate(data.1[,-c(1:2,12:14)],by=list(data.1$clusters_wss),FUN="median")
print(int_wss[,-1])
```

```
##   Market_Cap  Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## 1      73.84 0.460    21.50 31.0 15.0            0.8    0.280      8.560
## 2       4.78 0.555    23.35 14.2  5.6            0.6    0.475     14.495
##   Net_Profit_Margin clusters_wss
## 1              20.6            1
## 2              11.1            2
```

Aggregating the clusters to interpret the attributes - Silhouette

```
int_silhouette <- aggregate(data.2[,-c(1:2,12:14)],by=list(data.2$clusters_silhouette),FUN="median")
print(int_silhouette[,-1])
```

```
##   Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage Rev_Growth
## 1      2.230 0.535    19.25 13.15  6.10           0.40    0.635     29.775
## 2     31.910 0.405    69.50 13.20  5.60           0.75    0.475     12.080
## 3      2.600 0.850    26.00 21.40  4.30           0.60    1.450      6.380
## 4     59.480 0.480    21.10 26.90 13.35           0.75    0.345      6.630
## 5    153.245 0.460    21.25 43.10 17.75           0.95    0.220     19.610
##   Net_Profit_Margin clusters_silhouette
## 1              14.2                   1
## 2               6.4                   2
## 3               7.5                   3
## 4              19.3                   4
## 5              19.5                   5
```

median calculation - WSS

```
recommend_table1 <- table(data.1$cluster, data.1$Median_Recommendation)
names(dimnames(recommend_table1)) <- c("Cluster", "Recommendation")
recommend_table1 <- addmargins(recommend_table1)
recommend_table1
```

```
##         Recommendation
## Cluster Hold Moderate Buy Moderate Sell Strong Buy Sum
##     1      6            3             2          0  11
##     2      3            4             2          1  10
##     Sum    9            7             4          1  21
```

One strong buy, seven moderate buys, nine holds, and four moderate sells make the total number of 21 recommendations. All four recommendations, including the opposite advice on buys and sells, are mixed together in Cluster 2. Only Hold Moderate, Buy Moderate & Sell Strong are found in cluster 1.

median calculation - Silhouette

```
recommend_table2 <- table(data.2$cluster, data.2$Median_Recommendation)
names(dimnames(recommend_table2)) <- c("Cluster", "Recommendation")
recommend_table2 <- addmargins(recommend_table2)
recommend_table2
```

```
##        Recommendation
## Cluster Hold Moderate Buy Moderate Sell Strong Buy Sum
##     1      0            2             2          0   4
##     2      1            1             0          0   2
##     3      2            1             0          0   3
##     4      4            1             2          1   8
##     5      2            2             0          0   4
##     Sum    9            7             4          1  21
```

One strong buy, seven moderate buys, nine holds, and four moderate sells make the total number of 21 recommendations. All four recommendations, including the opposite advice on buys and sells, are mixed together in Cluster 5. Only mod purchase and hold information can be found in Clusters 1, 2, and 3.Both a moderate buy and moderate sell recommendation are present for Cluster 4.

Location of firm headquarter’s breakdown of clusters based on the mergeddata - wss

```
location_table <- table(data.1$cluster, data.1$Location)
names(dimnames(location_table)) <- c("Cluster", "Location")
location_table <- addmargins(location_table)
location_table
```

```
##        Location
## Cluster CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US Sum
##     1        0      0       0       0           1  2  8  11
##     2        1      1       1       1           0  1  5  10
##     Sum      1      1       1       1           1  3 13  21
```

There are 21 firms in all, with 13 in the US, 3 in the UK, and 1 each in Canada, France, Germany, Ireland, and Switzerland. US, UK, and Switzerland are all featured in Cluster 2.Switzerland, Uk And Us are in Cluster 1. Expect Switzerland Remaining All Countries are in Cluster 2.

Location of firm headquarter’s breakdown of clusters based on the mergeddata - Silhouette

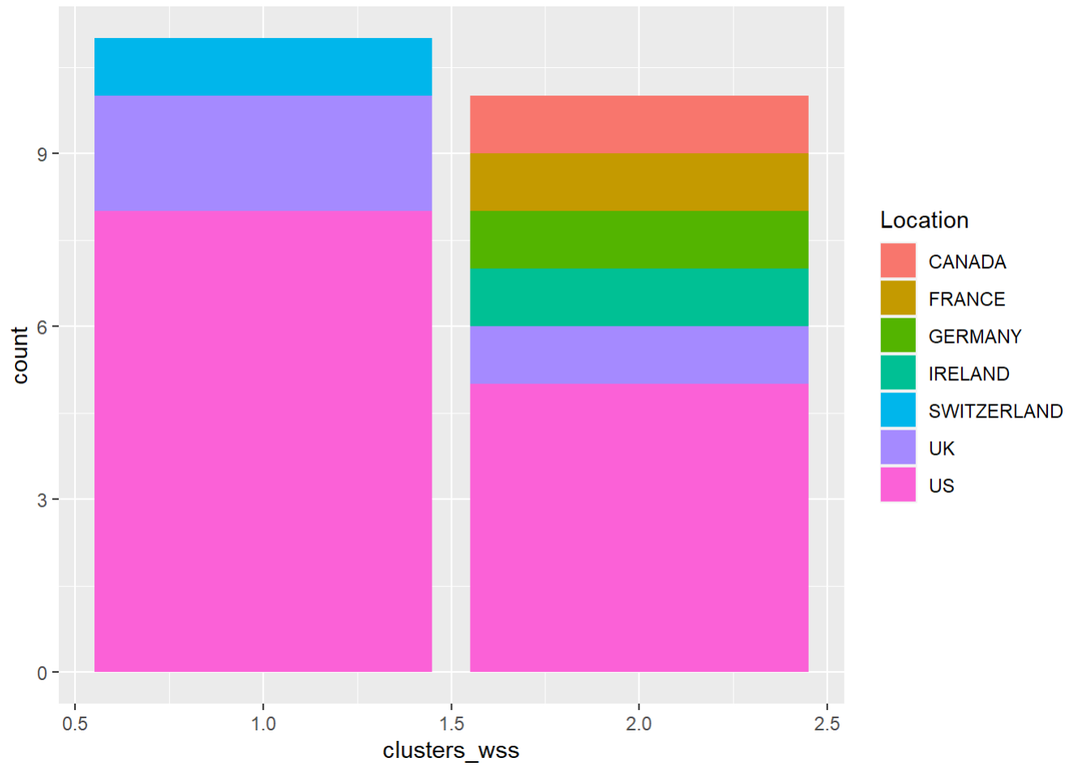
```
location_table <- table(data.2$cluster, data.2$Location)
names(dimnames(location_table)) <- c("Cluster", "Location")
location_table <- addmargins(location_table)
location_table
```

```
##        Location
## Cluster CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US Sum
##     1        0      1       0       1           0  0  2   4
##     2        1      0       0       0           0  0  1   2
##     3        0      0       1       0           0  0  2   3
##     4        0      0       0       0           1  2  5   8
##     5        0      0       0       0           0  1  3   4
##     Sum      1      1       1       1           1  3 13  21
```

There are 21 firms in all, with 13 in the US, 3 in the UK, and 1 each in Canada, France, Germany, Ireland, and Switzerland. US, UK, and Switzerland are all featured in Cluster 5. Germany and the US are in Cluster 2. US and Canada are in Cluster 1. US and Britain are in Cluster 3. The US, France, and Ireland make up Cluster 4.

Pattern in the categorical variables - wss

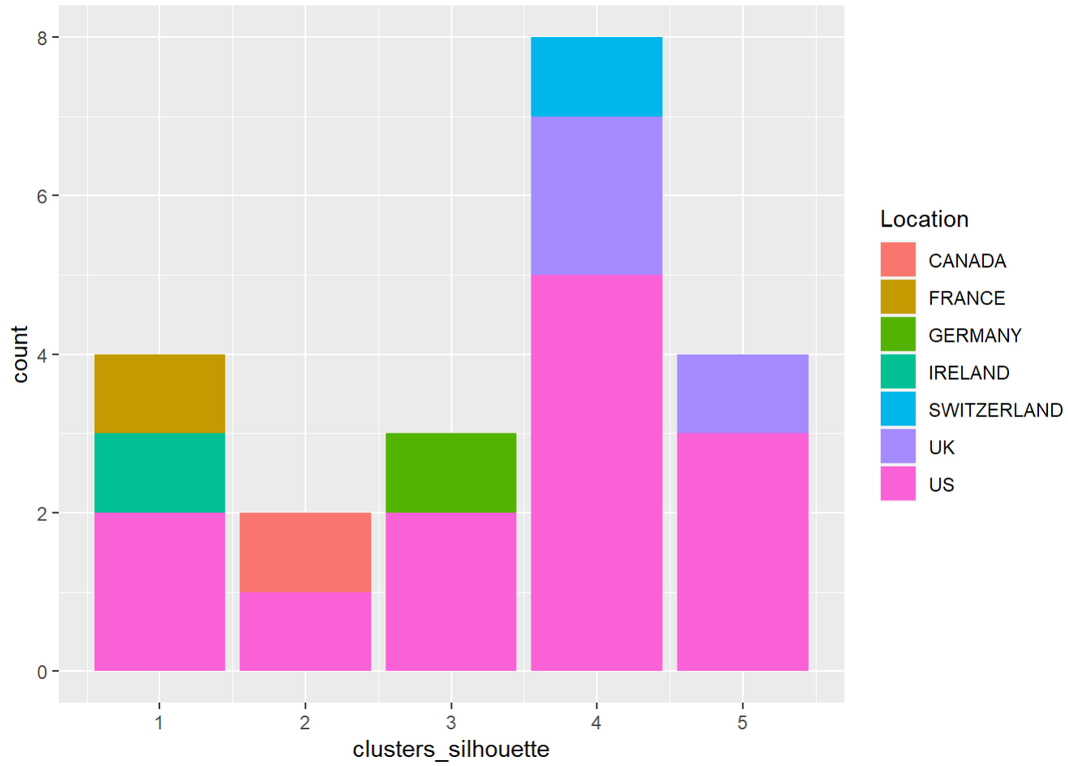
```
ggplot(data.1,aes(x=clusters_wss,fill=Location)) + geom_bar()
```



Cluster 1 and Cluster 2 seems to have a pattern with respect to the location of the pharmaceutical firms. More than 50% of the firms across both the clusters have “US” as their location. This also states that US has firms which are both profitable to invest (Acceptable Profitability with Moderate Risk) as well as firms which don’t yield that good profits (Low Profitability with High Risk). But comparatively the better performing cluster i.e. Cluster 1 seems to have a greater ratio of companies based in US.

Pattern in the categorical variables - silhouette

```
ggplot(data.2,aes(x=clusters_silhouette,fill=Location)) + geom_bar()
```



In the silhouette clusters we get to see the see the similar level of pattern towards to the location as observed in the wss. Every cluster in here as more of it's locations in "US" when compared to that with the other locations. But it seems interesting to observe that the best cluster which defines the domain with true sense i.e. Cluster 4 has a greater ratio of US companies with a lesser ratio of Non - US based companies.

*Note: The patterns therefore obtained in each of the clustering methods are generic, this is mostly because of the less amount of data which didn't give any further scope to visualize the categorical attributes.

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

Interpretation:(WSS)

Note: The interpretation is solely based on the financial characteristics of the specified firms in each of the clusters; as a result, the interpretation obtained would assist a person in deciding which of the two clusters to invest in order to benefit.

A. Acceptable Profitability with Moderate Risk:

The first cluster acquired here is an excellent investment due to its high likelihood of success. Success is measured using the criteria "Market Capital", ROE - Return on Expenditure, ROA - Return on Assets, Asset Turnover, and Net Profit Margin. The capital value in this cluster is 73.84, ROE, which indicates the returns on investment, is high (31), and ROA, which indicates the returns a corporation expects to earn on assets, is also high (15). Likewise, asset turnover and net profit are high.The PE Ratio is lower in the second cluster, indicating that the company is fairly valued with no disparities in share prices.

The level of risk in this investment is low which is called out by the "Beta" value, generally beta value should be lower than 1 in this case it is 0.46 which refers that the variability in these firms would be moderate not having enough of fluctuations.Also the "Leverage" value, which refers to a firm having borrowed capital for an investment should be as less as possible because market is always unpredictable and there would be possibilities of a firm loosing the money which they have borrowed for an investment expecting profits in return. Here the leverage value is 0.28 which is comparatively less to the second cluster. "With a good investment there should be very little chance of losing the total amount invested" and the group of firms in this cluster are expressing higher success rate when compared to that with the second cluster.

B. Low Profitability with High Risk:

When compared to the first cluster, the second cluster has poor performance metrics; the market capital is very low, at 4.78, compared to 73.84 in the first cluster, indicating that the firms listed in this cluster have a lower market share. Return on Expenditure (ROE), Return on Assets (ROA), Asset Turnover, and Net Profit Margin all have decreased. In these firms, the amount of risk indicated by the Beta and Leverage values is high, implying that there is significant variability and borrowing in comparison to the first cluster.

Interpretation:(silhouette)

A. Emerging Group

The First Cluster struggles to provide returns on expenditure, which is essentially the value that any investor would seek as a return on investment. External borrowings are also high, with a high degree of variability in the firms (beta). It also has the lowest capital value among all groups, and it is amusing to note that the revenue across these firms is also the highest. This could be because the firms were founded recently and are settling in to begin their journey in the market.

B. Overvalued and High-Risk Investment Group

The Second Cluster is most likely similar to the “High-Risk Investment Group”. It appears to have a high degree of variability in its PE Ratio, which is the ratio of share price to company value, indicating that it is likely overvalued. The beta and leverage values are also high, indicating that there is additional risk in this group. This cannot be a wise investment decision.

C. High-Risk Investment Group

The third Cluster is a highly volatile cluster with higher beta (firm variability) and leverage (outside borrowings) values, indicating that these firms have a high sense of risk. Furthermore, the market capitalization and net profit margin are lower, making it less suitable for any potential investments.

D. Promising Value opportunity Group

The fourth Cluster can be defined as a group of firms with viable market capital that are properly valued (PE Ratio) and involve moderate risk (Beta and Leverage). It also has higher returns on investment and assets with a profitable tendency.Despite the fact that the capital value is lower when compared to the fourth cluster, there is a possibility that the valuation will change/rise in the future.

E. Prime Investment with Slighter Risk Group

The Fifth Cluster is a good source of investment for any discrete individual who want to set a beneficial pitch for him/her. Here in this cluster as we see when compared to other firms across various clusters, the fourth cluster is having the “Highest Market Capital” of “153.245”, “Lofty ROE - Return on Expenditure of”43.10” & ROA - Return on Assets of “17.75”, “Sky-Spiking Asset Turnover” of “0.95” and “Net Profit Margin” of “19.5”. It also has a “decent beta value” - indicating that the variance would be less and no much of risk would be involved and not only that it has “less leverage value” - which refers stating that the borrowed capital for future investments is small. PE Ratio is less indicating that the price to earnings ratio (share price to company value) is manageable indicating that the company is properly valued. If anyone wants to invest in a company which has a higher capital ratio and moderate risk with fewer liabilities then the firms which are part of this cluster make the best choice.

Conclusion:

Any investment can be divided into three categories based on three criteria: security, income, and capital growth. Every investor must choose an appropriate combination of these three factors.

The “profit to loss ratio” is always a constraint on investment; every individual would want to maximize their profit while incurring the least amount of loss or incurring no loss at all. In this case, the supplied data set’s cluster titled “Prime Investment with Slighter Risk” demonstrates all of these characteristics. Based on my research and interpretation, I believe this is the best cluster to invest in because there is less risk and more earnings.

Note: The reason for choosing a cluster from the silhouette approach is that it helps in better defining the domain, which can be used by anyone to make an informed decision about their investment choices.