Assignment 4

Yashwanth K

2023-11-07

Summary:

1. • Absent Values Check: determining the percentage of missing values in each of our p_dataset's columns.

• Normalisation: Applying the scale function to normalise the p_data. Variables of the same scale are required for K-Means clustering.

• Determining Optimal K: The ideal number of clusters by employing the Silhouette and Elbow methods (wss). Whereas k = 5 is the ideal k value determined by the Silhouette approach and k = 2 by the WSS method

2.K-Means Clustering:

• Utilising the Within-Sum-of-Squares (WSS) approach, do K-Means clustering with k = 2. To prevent local minima, the nstart argument facilitates the algorithm's repeated execution with various beginning centroids. Sum of squares within clusters by clusters = 43.3, 75.2 & Between-Cluster Proportion(between_SS / total_SS) = 34.1%

• By applying the Silhouette approach to K-Means clustering with k = 5, a more intricate knowledge of the cluster structure may be obtained. Nstart is employed to increase the findings' robustness, much like the WSS approach does.The within-cluster sum of squares by cluster is 12.79, 2.8, 15.595925, 21.879320, and 9.284424. The between-cluster proportion is 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, sizes 11 and 10.

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

• WSS -Regarding where the pharmaceutical companies are located, Clusters 1 and 2 appear to follow a pattern. Across both clusters, "US" is the location for more than 50% of the businesses. This indicates that the United States of America comprises companies that are lucrative to invest in (Acceptable Profitability with Moderate Risk) and companies that don't make as much money as they could (Low Profitability with High Risk). However in comparison, it appears that a higher percentage of the businesses in Cluster 1, which performs better, are US-based.

• silhouette - We can notice a similar degree of pattern towards the site as shown in the wss in the silhouette clusters. Compared to the other locations, every cluster in this one has a higher percentage of its locations in the "US". However, it's intriguing to see that Cluster 4, the best cluster that accurately characterises the domain, has a higher proportion of US-based businesses than non-US-based businesses.

3.

WSS –

1. Reasonable Returns with a Moderate Risk
2. Low Earnings but High Risk:

Silhouette-

1. Group of High-Risk Investors
2. Overvalued and High-Risk Investment Group
3. Potentially Rich Opportunity Group
4. Outstanding Investment with Slighter Risk Group
5. A New Group

```
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 1.9.2     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")
```

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

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

#Loading and exploring the p_data

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

```
##   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(p_data)
```

```
##     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.

#Looking for na values

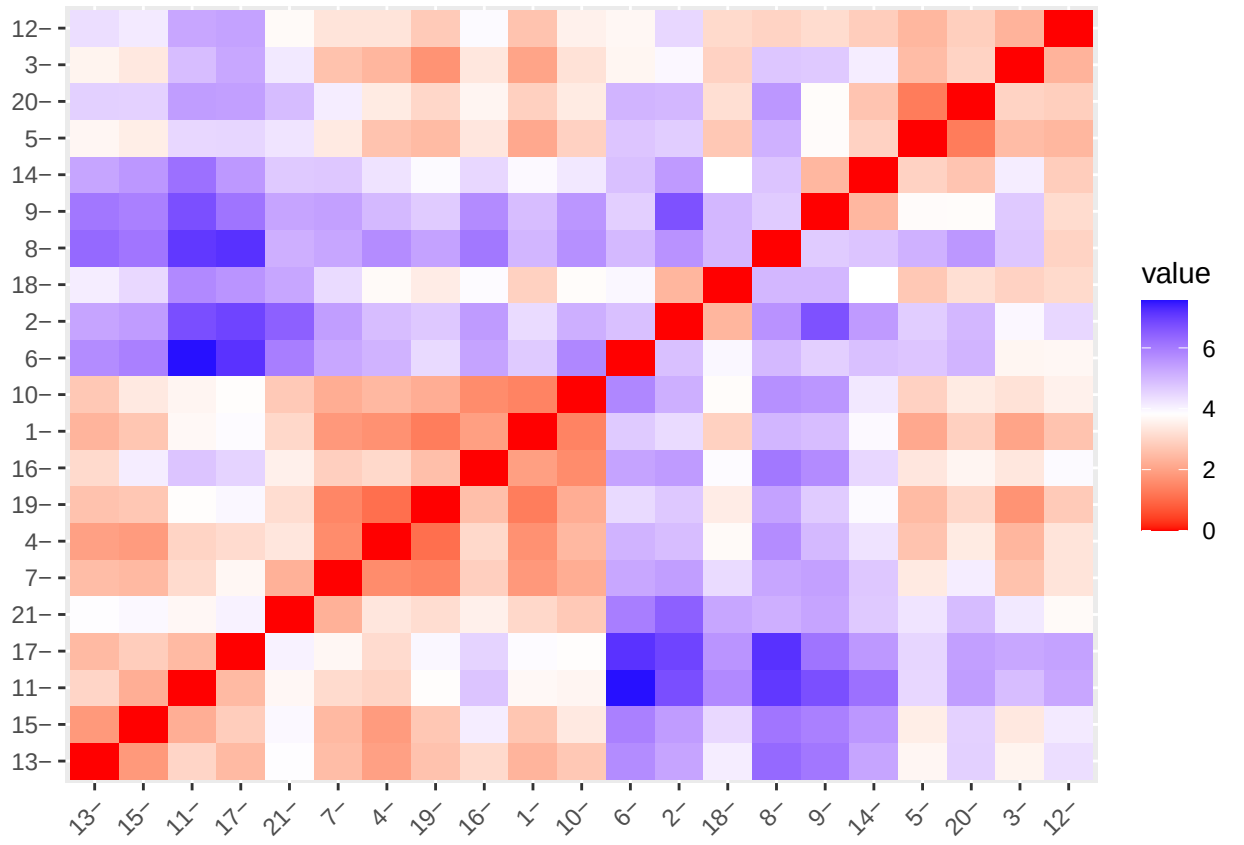
```
colMeans(is.na(p_data))
```

```
##                 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

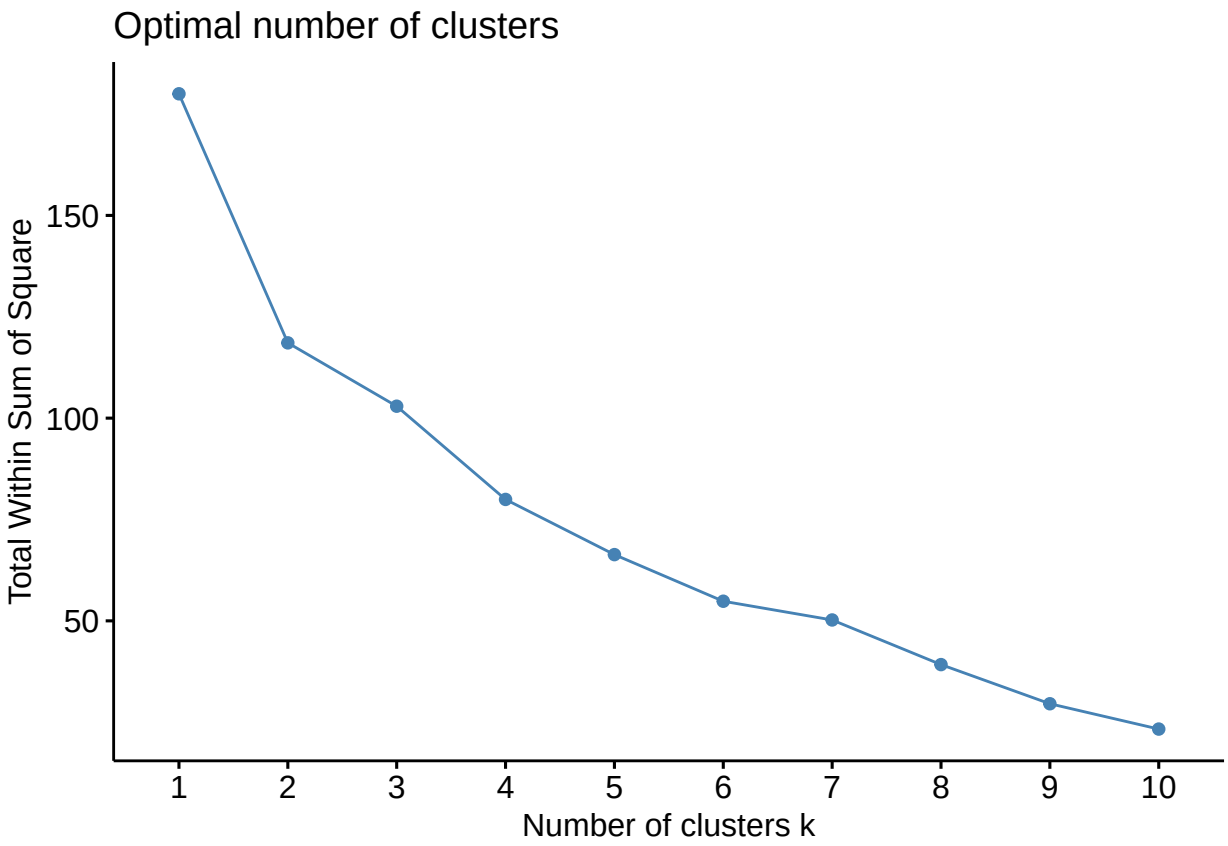
```
set.seed(123)
p_data.norm <- scale(p_data[,-c(1:2,12:14)])
```

```
Distance <- dist(p_data.norm, method = "euclidian")
fviz_dist(Distance)
```



Finding optimal K using wss method

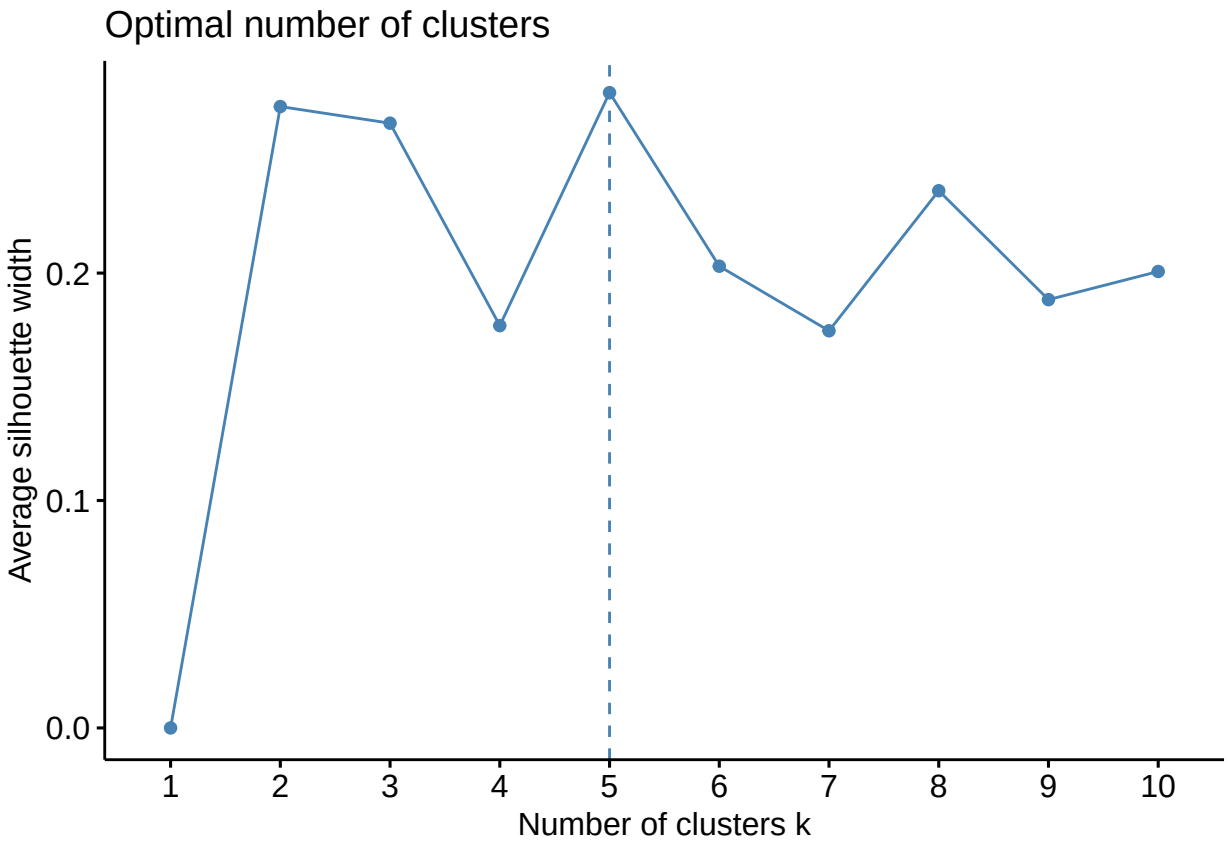
```
wss <- fviz_nbclust(p_data.norm,kmeans,method="wss")
wss
```



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

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



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(p_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"
```

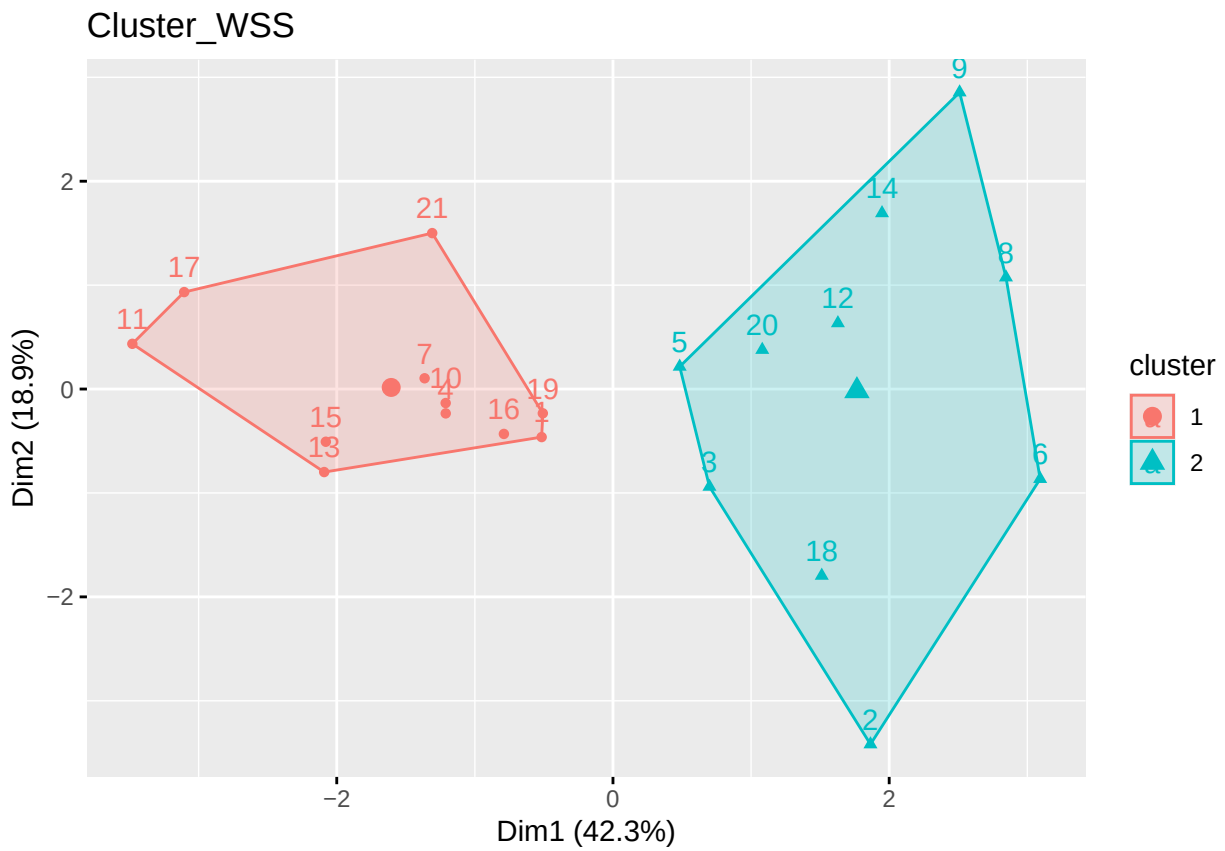
#Formulation of clusters using K-Means with k = 5 (Silhouette)

```
silhouette_kmeans <- kmeans(p_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)

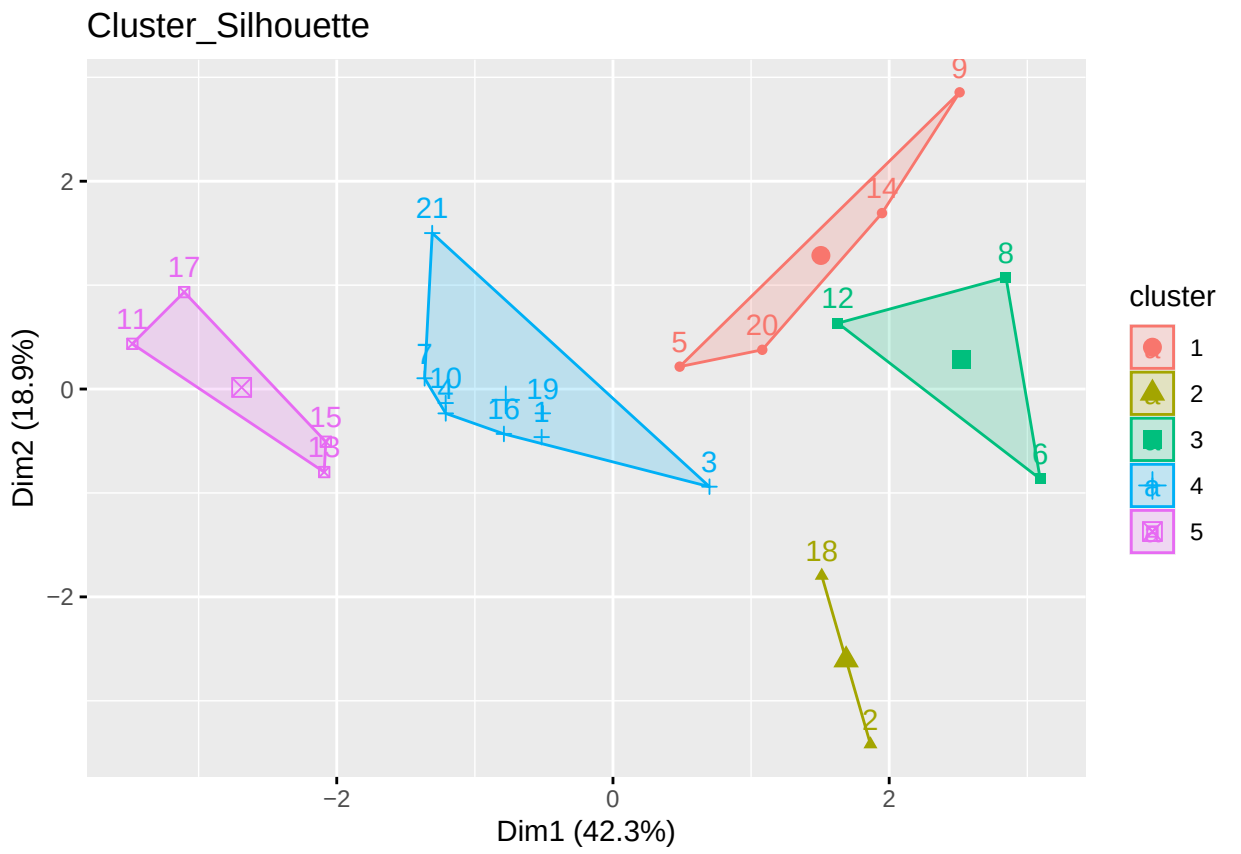
```
fviz_cluster(wss_kmeans,p_data[,-c(1:2,12:15)],main="Cluster_WSS")
```



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,p_data[,-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 p_data frame for analysis

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

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

#Aggregating the clusters to interpret the attributes - WSS

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

```
intial_silhouette <- aggregate(p_data.2[,-c(1:2,12:14)],by=list(p_data.2$clusters_silhouette),FUN="media
print(intial_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

```
recom_table1 <- table(p_data.1$cluster, p_data.1$Median_Recommendation)
names(dimnames(recom_table1)) <- c("Cluster", "Recommendation")
recom_table1 <- addmargins(recom_table1)
recom_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
```

There are 21 suggestions in total, consisting of 1 strong buy, 7 moderate buys, 9 holds, and 4 moderate sells. In Cluster 2, all four recommendations—including the opposing advise on buys and sells—are combined. Cluster 1 has just Buy Moderate, Sell Strong, and Hold Moderate

#median calculation - Silhouette

```
recom_table2 <- table(p_data.2$cluster, p_data.2$Median_Recommendation)
names(dimnames(recom_table2)) <- c("Cluster", "Recommendation")
recom_table2 <- addmargins(recom_table2)
recom_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
```

The overall amount of 21 suggestions is comprised of one strong buy, seven moderate buys, nine holds, and four moderate sells. Cluster 5 contains all four suggestions, including the opposing guidance on purchases

and sales. Clusters 1, 2, and 3 provide information only on mod purchase and hold matters.For Cluster 4, recommendations for both a moderate buy and a moderate sell are shown.

#Location of firm headquarter's breakdown of clusters based on the mergedp_data - wss

```
l_table <- table(p_data.1$cluster, p_data.1$Location)
names(dimnames(l_table)) <- c("Cluster", "Location")
l_table <- addmargins(l_table)
l_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 a total of 21 firms: 13 are located in the United States, 3 in the United Kingdom, and 1 in each of Canada, France, Germany, Ireland, and Switzerland. In Cluster 2, the US, UK, and Switzerland are all highlighted.Cluster 1 includes the US, Switzerland, and the UK. With the exception of Switzerland, all other nations are in Cluster 2.

#Location of firm headquarter's breakdown of clusters based on the mergedp_data - Silhouette

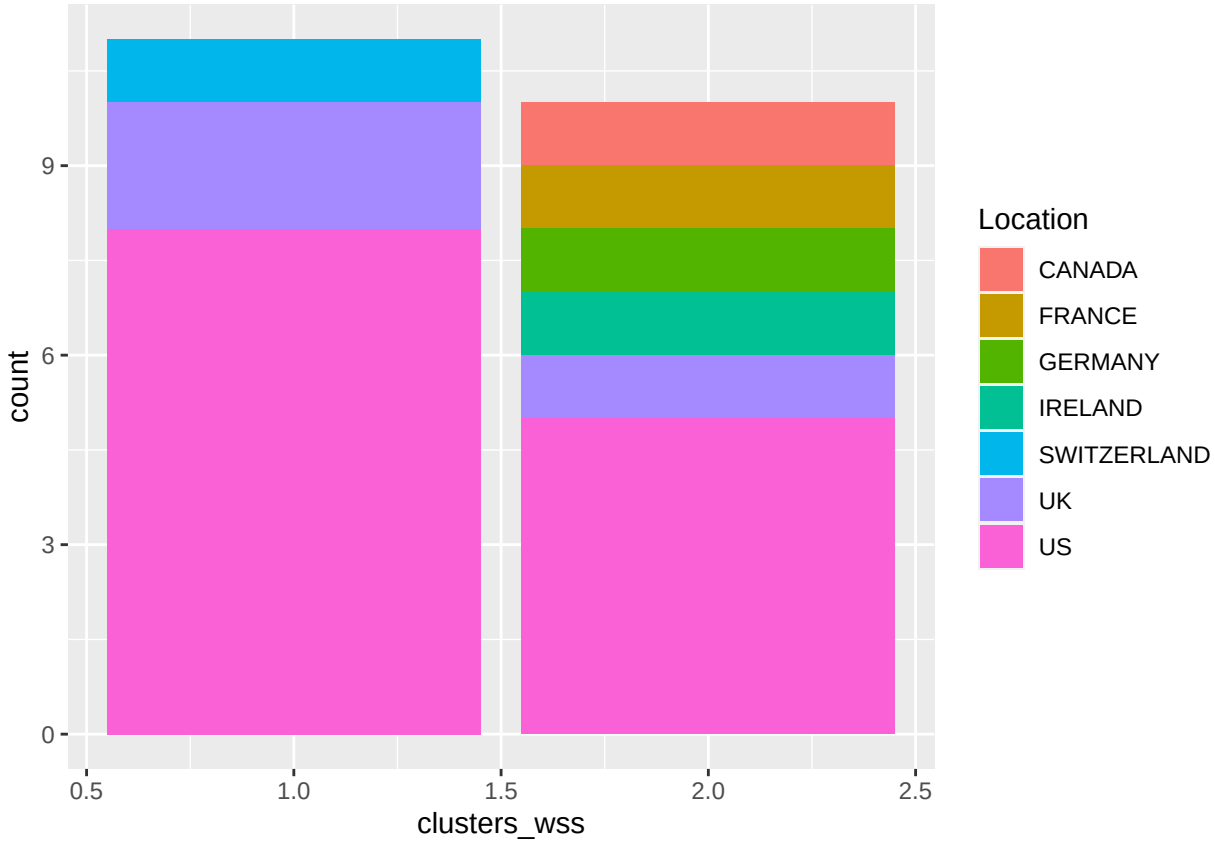
```
l_table <- table(p_data.2$cluster, p_data.2$Location)
names(dimnames(l_table)) <- c("Cluster", "Location")
l_table <- addmargins(l_table)
l_table
```

```
##         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
```

A total of 21 businesses are involved, comprising 13 in the United States, 3 in the United Kingdom, and 1 in each of Canada, France, Germany, Ireland, and Switzerland. Cluster 5 includes the US, UK, and Switzerland. Cluster 2 includes the US and Germany. Cluster 1 includes the US and Canada. Britain and the US are in Cluster 3. Cluster 4 consists of the US, France, and Ireland.

#Pattern in the categorical variables - wss

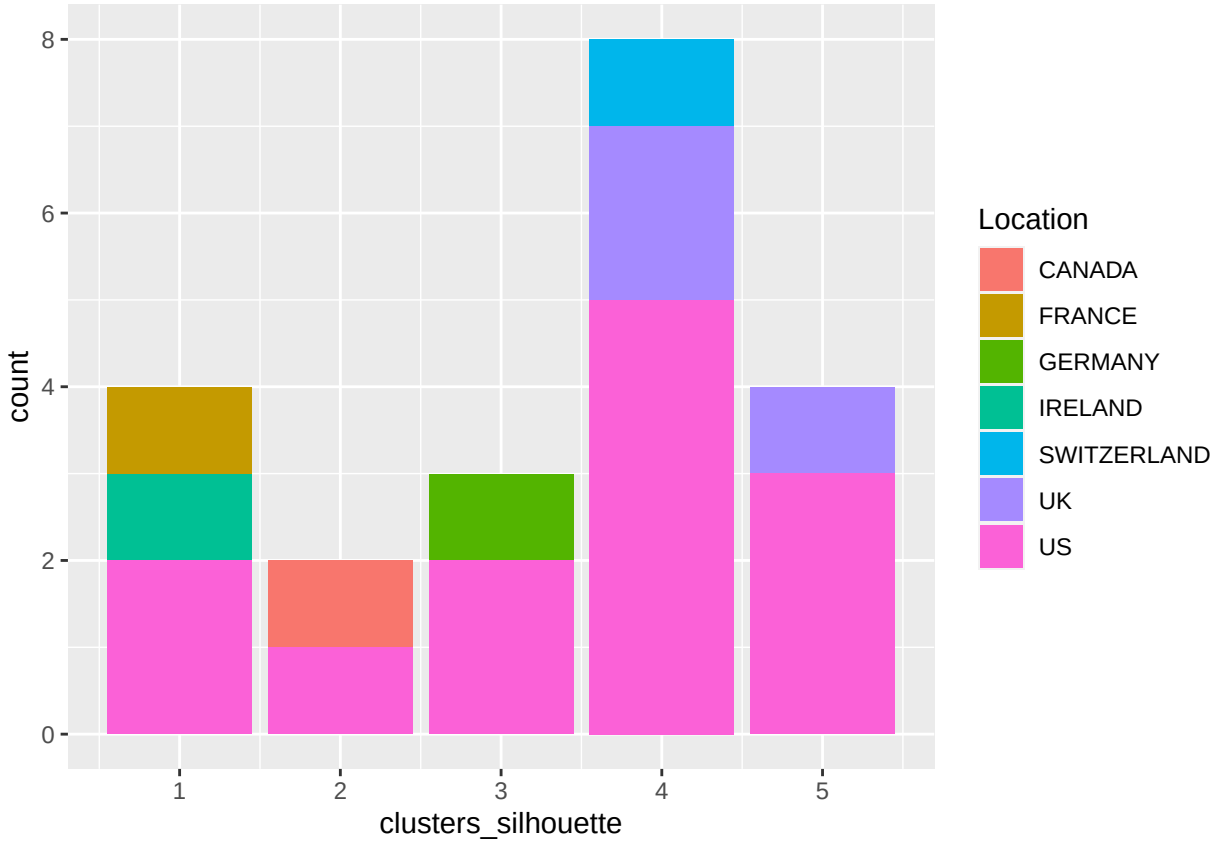
```
ggplot(p_data.1,aes(x=clusters_wss,fill=Location)) + geom_bar()
```



In terms of the pharmaceutical companies' locations, Clusters 1 and 2 appear to follow a pattern. In both clusters, "US" is the location for more than 50% of the businesses. This also says that the US has businesses that are lucrative to invest in (Acceptable Profitability with Moderate Risk) and businesses that don't make much money (Low Profitability with High Risk). However, in comparison, Cluster 1, which is the higher performing cluster, appears to contain a higher percentage of US-based businesses.

#Pattern in the categorical variables - silhouette

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



We can notice a comparable degree of pattern towards the site as shown in the wss in the silhouette clusters. In comparison to the other locations, every cluster in this one has a higher percentage of its locations in the "US". Nonetheless, it's intriguing to see that Cluster 4, the strongest cluster for accurately defining the domain, has a higher proportion of US-based businesses than non-US-based businesses.

*Note: As a result, the patterns found in each clustering approach are general. This is mostly due to the little amount of p_data, which left no room for further category attribute visualisation.

#summarizing the stock exchange values for each cluster - wss

```
exchange_table <- table(p_data.1$cluster, p_data.1$Exchange)
names(dimnames(exchange_table)) <- c("Cluster", "Exchange")
exchange_table <- addmargins(exchange_table)
exchange_table
```

```
##         Exchange
## Cluster AMEX NASDAQ NYSE Sum
##     1      0      0   11  11
##     2      1      1    8  10
##     Sum    1      1   19  21
```

There are 21 companies overall, divided into 1 Amex, 1 Nasdaq, and 19 NYSE. Cluster 1 just has the NYSE. All three are in Cluster 2.

#summarizing the stock exchange values for each cluster - silhouette

```r
exchange_table <- table(p_data.2$cluster, p_data.2$Exchange)
names(dimnames(exchange_table)) <- c("Cluster", "Exchange")
exchange_table <- addmargins(exchange_table)
exchange_table
```

```
##         Exchange
## Cluster AMEX NASDAQ NYSE Sum
##     1      0      0    4   4
##     2      0      0    2   2
##     3      1      1    1   3
##     4      0      0    8   8
##     5      0      0    4   4
##     Sum    1      1   19  21
```

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

Interpretation:(WSS)

Note: The interpretation is exclusively based on the financial attributes of the specified firms in each of the clusters; the interpretation obtained would therefore assist a person in making a decision about which of the two clusters to invest in order to benefit.

A) Acceptable Profitability with Moderate Risk:

Given the high likelihood of success, this initial cluster purchase is a great financial decision. Here, performance is evaluated using the following criteria: "Market Capital," "ROE," "Return on Expenditure," "ROA," "Asset Turnover," and "Net Profit Margin." The cluster under consideration exhibits a capital value of 73.84, a high return on equity (ROE) of (31), and an expectation of high returns on assets (15) from the firm. In a similar vein, net profit and asset turnover are also high.When compared to the second cluster, the PE Ratio is lower, meaning that the company's share price is evenly valued.

This investment has a low degree of risk, as indicated by the "Beta" value of 0.46. Generally speaking, a beta value of less than one indicates that the variability of these enterprises is mild, meaning there aren't enough variations.Furthermore, because the market is always unpredictable and there is a chance that a company would lose the money it has borrowed for an investment with the expectation of making profits, the "Leverage" value—which indicates how much a corporation has borrowed capital for an investment—should be as low as feasible. In this case, the leverage value is 0.28, which is lower than in the second cluster. "With a good investment there should be very little chance of losing the total amount invested"

B) Low Profitability with High Risk:

In this instance, the second cluster performs poorly in comparison to the first cluster. The market capitalization of the listed companies in this cluster is quite low, at 4.78, whereas it was 73.84 in the first cluster. Net profit margin, asset turnover, return on expenditure (ROE), and return on assets (ROA) are all lower. In compared to the first cluster, these businesses appear to have far more fluctuation and borrowing due to the high levels of risk suggested by their Beta and Leverage scores.

Interpretation:(silhouette)

A) Group of High-Risk Investors

When it comes to offering returns on expenditure—basically, the value that any investor would hope to receive as a return on investment—the First Cluster falters. There is also a lot of external borrowing and a fair degree of business variability (beta). In addition, its capital worth is the lowest of all the groupings. Surprisingly, these enterprises also have the largest income. This might be the case since the companies are very new and are settling in before venturing out into the market.

B) Overvalued and High-Risk Investment Group

The PE Ratio, or the ratio of share price to firm value, appears to be very variable for the Second Cluster, suggesting that it is probably overvalued. There is consequent risk associated in this group, as shown by the high beta and leverage ratings. There must be a better option for an investment than this.

C) Potentially Rich Opportunity Group

The Third Cluster exhibits significant volatility, as seen by its greater beta (firm variability) and leverage (outside borrowings) levels, which suggest a high level of risk in these companies. Additionally, it is less suited for any potential investments due to its lower market capital and net profit margin.

D) Outstanding Investment with Slighter Risk Group

The Fourth Cluster is a group of companies with a manageable market capitalization, a fair PE ratio, and moderately risky operations (beta and leverage). Better returns on investment and assets with a profitable propensity are other features of it.When compared to the fourth cluster, it has a lower capital value, but it may still be a viable investment option because there's always a risk that the valuation will increase or decrease in the future.

E) New group

The Fith Cluster is an excellent source of investment for any independent individual looking to establish an advantageous pitch for himself/herself. When compared to other firms in different clusters, the fourth cluster has 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," which means that the variation will be lower and there will be less danger, and it has a "less leverage value," which means that the borrowed money for future investments will be minimal. -The PE Ratio is lower, suggesting that the price to earnings ratio (share price to business value) is controllable and the company is appropriately priced. If you want to invest in a company that has a greater capital ratio, moderate risk, and fewer liabilities, the firms in this cluster are the best alternative.

Conclusion:

Three factors—safety, income, and capital growth—can be used to categorise every investment. All investors need to do is choose a reasonable mix of these three components.

The "profit to loss ratio" constantly places restrictions on investments; the goal of any particular person is to maximise profits while minimising losses or experiencing no losses at all. In this instance, all of these characteristics are shown by the cluster named "Prime Investment with Slighter Risk" from the provided p_data set. I think this is the best cluster to choose for an investment since there is less risk and more profit based on the analysis and study.

Note: The rationale behind choosing a cluster from the silhouette technique is that it helps define the domain more clearly, which anybody can use to make better investment decisions.