Assignment_4

Tirumal Achina

2023-11-10

Summary:

The given dataset Pharmaceuticals is loaded into the variable “tiru”. We used head function to check whether the dataset has been correctly loaded or not.From hereon let us move on to the problems that we have to solve.

1: I have used first 9 numerical variables(3 to 11 columns) from the dataset to conduct cluster analysis.First we got the summary of all 9 numerical variables that we are using.

- Finding the distance between the rows matrix and visualizing it.Then we are staring our clustering analysis using wss and silhouette methods.

2: Here I use within sum of squares and Silhouette methods to find the optimal number of clusters to be formed.

- 2.1: Within Sum of Squares: The graph looks like a human hand having a bend like our elbow.The exact point at which their will be less decreasee in the graph. By having a glance at the graph we can see that their is slow down of rate of decrease in wss at “k=2”(this would be the optimal solution).
- Usually the lower the Wss value the tighter clusters are formed.The optimal value of wss is 0.Moreover it is difficult to identify optimal solution for some datasets then we shall opt to other methods.
- 2.2: Silhouette method: In this we can find the optimal solution by seeing the peak of the graph where the silhouette coefficient reaches its maximum value. From our graphh we can see that the graph reached its max point at k = 5. This suggests that the optimal solution for pharma dataset is “k=5”.
- If the value of silhouette distance is 1, then it means the datapoints are well assigned to the cluster and if its is -1 then datapoints are not assigned properly.
- Sometimes the optimal solution may be given different from both the methods.Then you have to follow the other methods or we have to decide which one choose based on the given results of by cluster summary.

Wss method:

-From Wss clustering analysis that formed two cluster we can interpret the following.

- Cluster_1: Reasonable profit margin at a Moderate risk

The high success rate of the first cluster found here makes it a wise investment.The following metrics are used to define success: asset turnover, return on assets (ROA), return on expenses (ROE), and net profit margin.This cluster has a capital value of 73.84, a return on equity (ROE) of 31 when the investment is high, and a return on assets (ROA) of 15, which represents the profit a company expects to make from its high asset investments.In a similar vein, net profit and asset turnover are both high.The fact that the PE Ratio is lower than that of the second cluster suggests that the company’s share price is evenly valued.The investment carries a low level of risk, as indicated by the “Beta” value of 0.46. Generally speaking, the beta value should be less than 1, indicating that the variability of these firms is moderate and does not exhibit sufficient fluctuations. Additionally, a company’s “Leverage” value—a measure of how much capital it has borrowed for an investment—should be as low as possible because the market is always unpredictable and there’s a chance that the money it borrowed for the investment could be lost while it was expected to yield profits. In this instance, the leverage value is 0.28, which is lower than in the second cluster. “There should be very little chance of losing the entire amount invested with a good investment,” and the businesses in this cluster are reporting higher success rates than those in the second cluster.

- Cluster_2: Low profit with high risk.

In this case, the second cluster’s performance metrics are subpar when compared to the first cluster’s. Its market capitalization is extremely low, 4.78 versus 73.84 in the first cluster, indicating that the companies listed in this cluster have a smaller market share than those in the first cluster. Return on Investment decreases in Return on Equity (ROE), Return on Assets (ROA), Asset Turnover, and Net Profit Margin. The degree of danger, which is emphasized by the high leverage and beta values in these companies, indicating a high degree of unpredictability and high borrowing rates in these businesses compared to the first cluster. In contrast, the PE Ratio is elevated.

-From the graph we can see that most of the pharmaceutical business firms are based in US and we can see a similar pattern in cluster 1 and cluster 2. 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.

Silhouette Method:

-From Silhouette clustering analysis that formed 5 clusters we can interpret the following.

- Cluster 1:

It appears that the First Cluster is being overhyped. The PE Ratio, which measures the share price in relation to the company’s value and indicates whether or not the stock is overvalued, appears to be quite variable. Additionally, this group has high beta and leverage values, indicating that there is consequent risk involved. There must be a better option than this for an investment.

- Cluster 2:

When it comes to offering returns on investment—basically, the value that any investor would look for as a return on investment. There is also a lot of external borrowing and a fair degree of firm variability (beta). In addition, its capital value is the lowest of all the groups. Surprisingly, these firms also have the highest revenue. This could be the case because the companies are relatively new and are settling in before venturing out into the market.

- Cluster 3:

The Third Cluster of the Destiny Class is a group of companies with a reasonable market capitalization, a fair PE ratio, and moderate levels of risk (beta and leverage). Additionally, it has assets with a profitable tendency and better returns relative to expenditure. Even though the capital value is lower when compared, it could still be a viable investment option because there’s a chance that the valuation will fluctuate or increase in the future.

- Cluster 4:

With higher beta (firm variability) and leverage (outside borrowings) values, the Cluster is a highly erratic cluster that suggests these businesses have a high sense of risk. Furthermore, the lower market capitalization and net profit margin render it less appropriate for potential investments.

- Cluster 5:

Anyone looking to set up a profitable pitch for themselves should consider investing in the Fourth Cluster. In this cluster,it has the “Highest Market Capital” of 153.245, the “Lofty ROE - Return on Expenditure of 43.10” & ROA - Return on Assets of 17.75”, the “Sky-Spiking Asset Turnover” of 0.95, and the “Net Profit Margin” of 19.5. This is in comparison to other firms across various clusters. In addition, it has a “less leverage value,” which denotes that there will be little borrowed capital needed for future investments, and a “decent beta value,” which suggests that there would be less variance and less risk involved.A company with a higher capital ratio, moderate risk, and fewer liabilities is a good choice for investors.The best option is made by the companies in this cluster.

- We can see a similar degree of pattern toward the location as seen in the wss and in the silhouette clusters. When compared to the other locations, every cluster in this one has a higher percentage of its locations in the “US.” Nonetheless, it’s intriguing to note that Cluster 4, the best cluster that accurately characterizes the domain, has a higher proportion of US-based businesses than non-US-based businesses.

-The other observations are-

- 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 3. Only mod purchase and hold information can be found in Clusters 1, 4, and 5.Both a moderate buy and moderate sell recommendation are present for Cluster 2.
- 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 3. Germany and the US are in Cluster 4. US and Canada are in Cluster 1. US and Britain are in Cluster 5. The US, France, and Ireland make up Cluster 2.
- There are 21 companies overall, divided into 1 Amex, 1 Nasdaq, and 19 NYSE. All three are in Cluster 4. clusters 1,2,3,5 all contains only NYSE.

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

Cluster 1: Non plus Organization(Hold)

Cluster 2: Reduced Compensation(Moderate)

Cluster 3: Destiny class(Moderate)

Cluster 4: Run-away investing(Hold)

Cluster 5: High Margins(Strong Buy)

Conclusion:

At the end every individual or an entity hopes to maximize their profit with minimal losses. They also look forward for the long-term run of the investment. From all analysis I can conclude that Cluster 5 is best to choose for investment. It has higher returns and has an long run. The other cluster I suggest is Cluster number 3. It has marginal profits that is associated with risk but has more chances to be in profit zone. The next clusters are not suggestable for any entity or venture capitalists, because it give up losses or no marginal profits are gained by investing with it.

```
library(factoextra)
```

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

```
## Loading required package: ggplot2
```

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

```
library(ggplot2)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ lubridate 1.9.2     ✔ tibble    3.2.1
## ✔ purrr     1.0.2     ✔ tidyr     1.3.0
```

```
## ── 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 becom
e errors
```

```
library(dplyr)
```

#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.

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

```
##   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
```

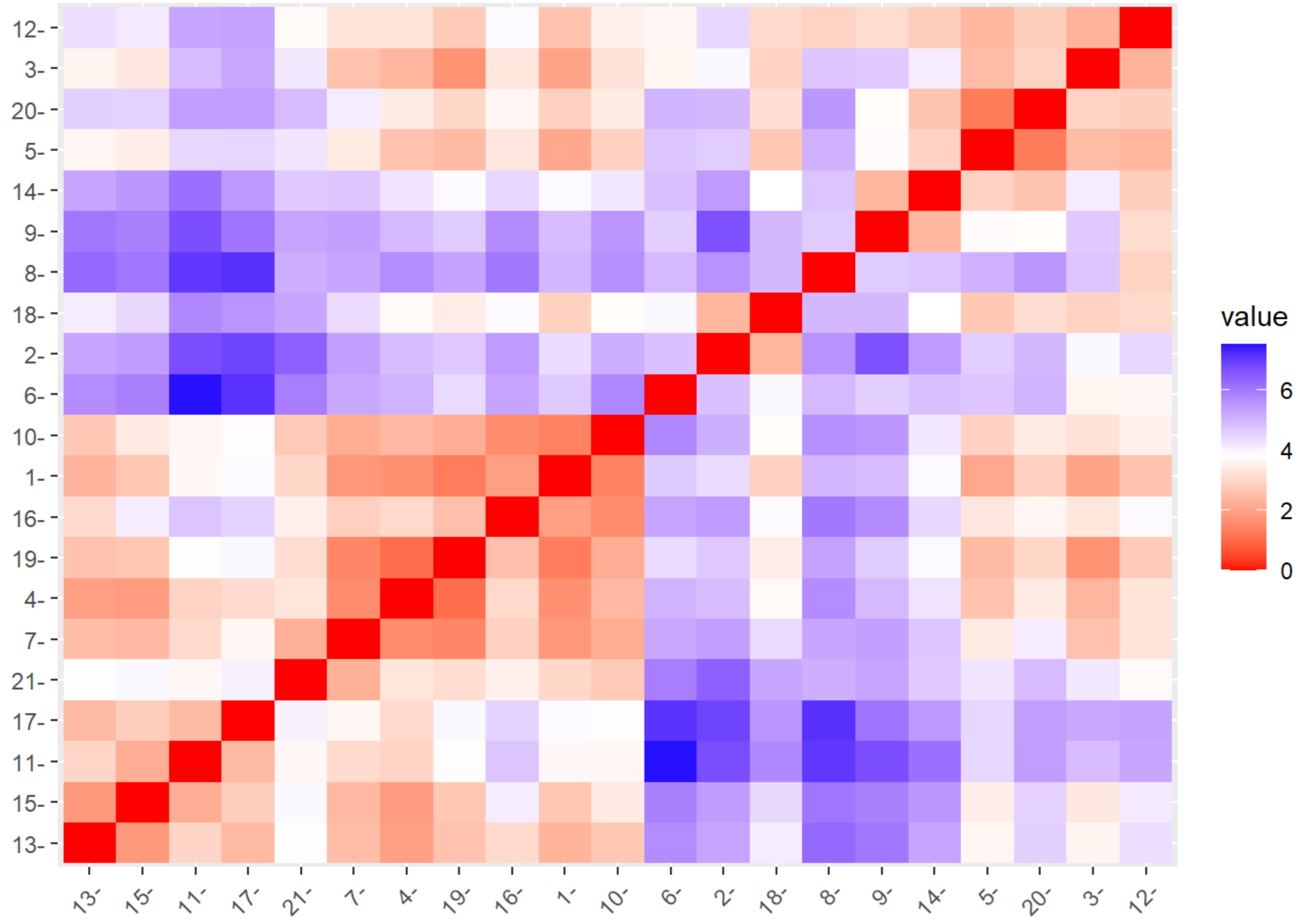
```
str(tiru)
```

```
## 'data.frame':    21 obs. of  14 variables:
##  $ Symbol               : chr  "ABT" "AGN" "AHM" "AZN" ...
##  $ Name                 : chr  "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZe
neca PLC" ...
##  $ 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" ...
```

```
tiru_new <- scale(tiru[,3:11])
summary(tiru_new)
```

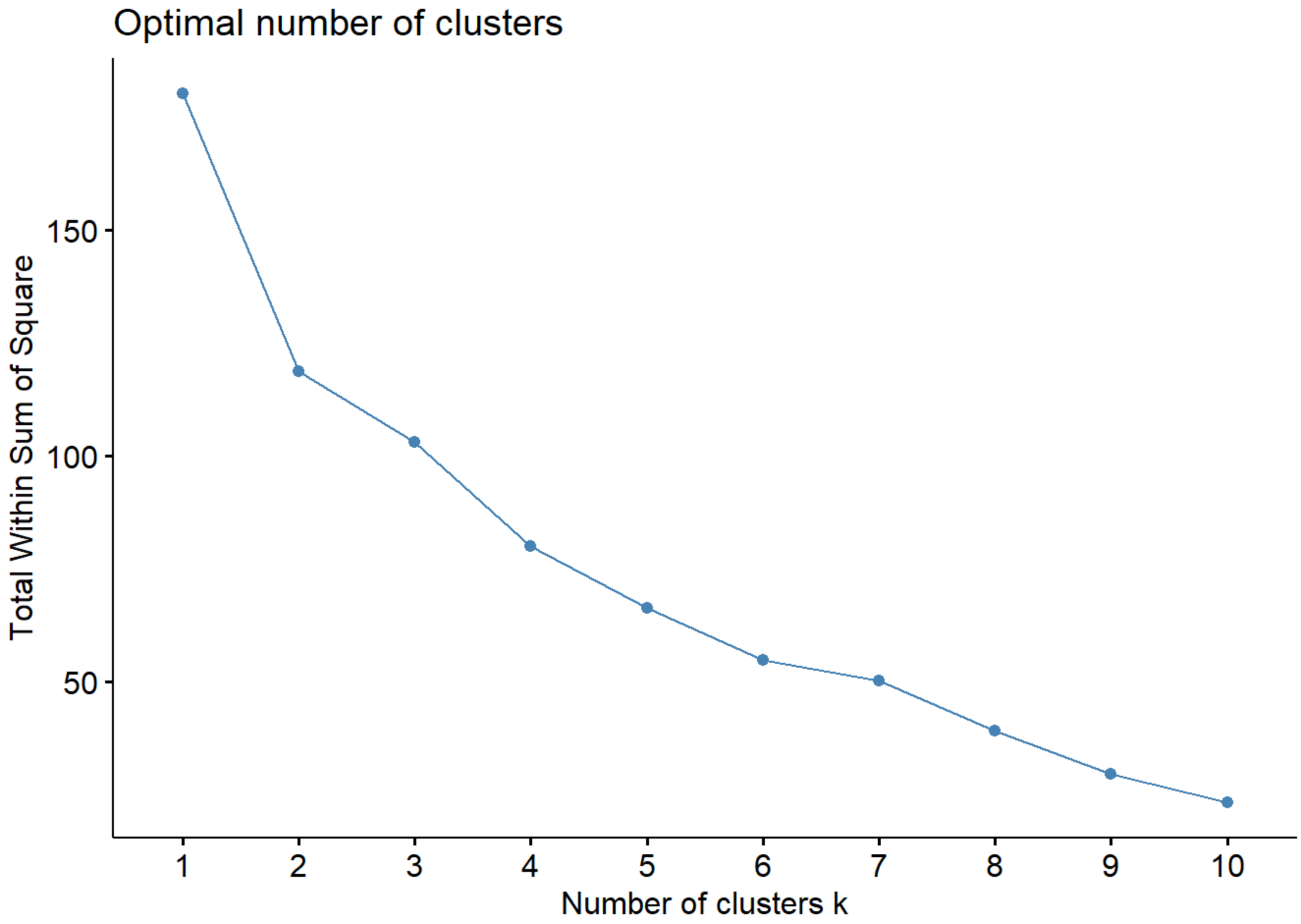
```
##    Market_Cap           Beta             PE_Ratio            ROE         
##  Min.   :-0.9768   Min.   :-1.3466   Min.   :-1.3404   Min.   :-1.4515  
##  1st Qu.:-0.8763   1st Qu.:-0.6844   1st Qu.:-0.4023   1st Qu.:-0.7223  
##  Median :-0.1614   Median :-0.2560   Median :-0.2429   Median :-0.2118  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.: 0.2762   3rd Qu.: 0.4841   3rd Qu.: 0.1495   3rd Qu.: 0.3450  
##  Max.   : 2.4200   Max.   : 2.2758   Max.   : 3.4971   Max.   : 2.4597  
##       ROA          Asset_Turnover       Leverage          Rev_Growth     
##  Min.   :-1.7128   Min.   :-1.8451   Min.   :-0.74966   Min.   :-1.4971  
##  1st Qu.:-0.9047   1st Qu.:-0.4613   1st Qu.:-0.54487   1st Qu.:-0.6328  
##  Median : 0.1289   Median :-0.4613   Median :-0.31449   Median :-0.3621  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000  
##  3rd Qu.: 0.8430   3rd Qu.: 0.9225   3rd Qu.: 0.01828   3rd Qu.: 0.7693  
##  Max.   : 1.8389   Max.   : 1.8451   Max.   : 3.74280   Max.   : 1.8862  
##  Net_Profit_Margin 
##  Min.   :-1.99560  
##  1st Qu.:-0.68504  
##  Median : 0.06168  
##  Mean   : 0.00000  
##  3rd Qu.: 0.82364  
##  Max.   : 1.49416
```

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

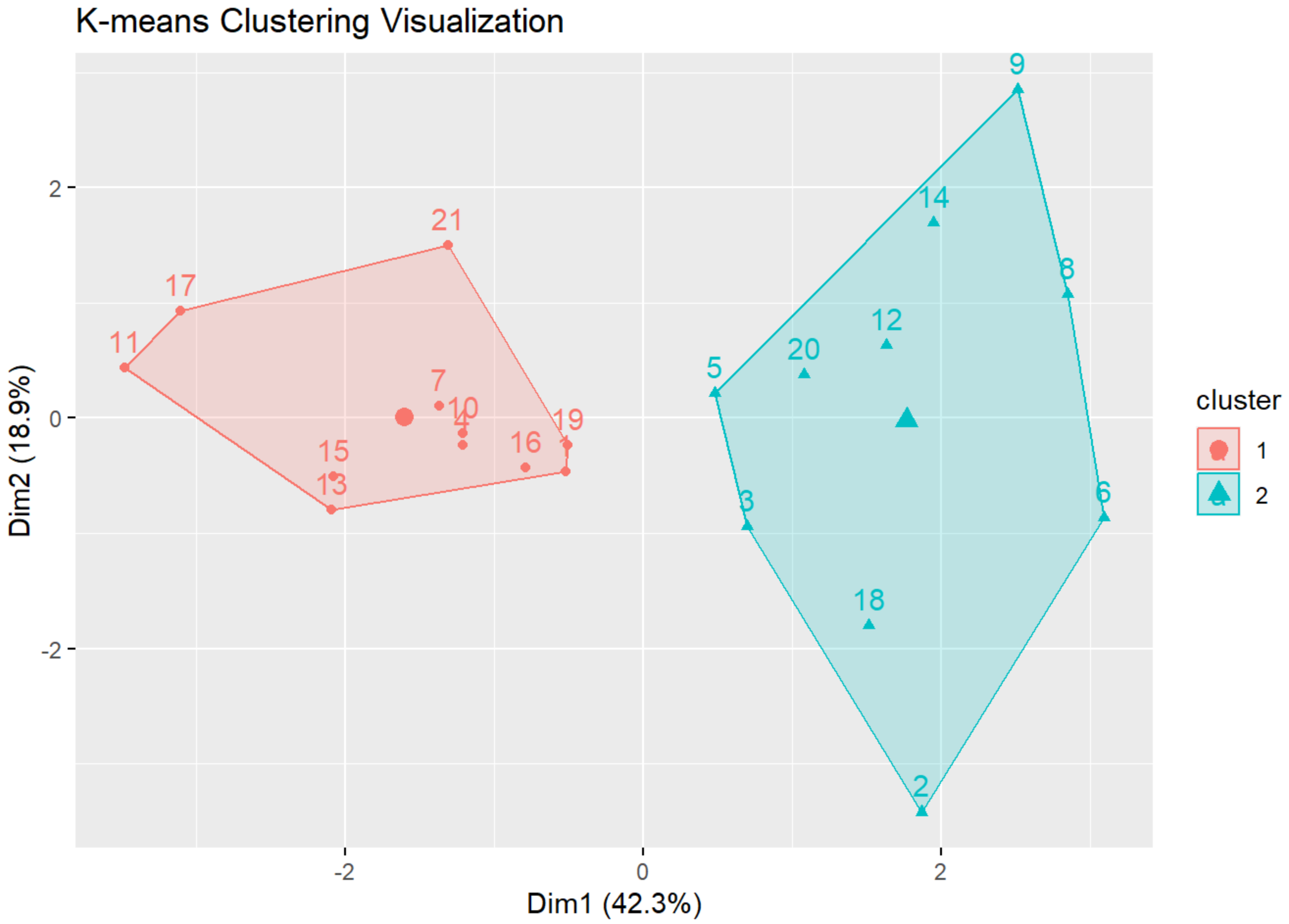


#Applying k_means clustering

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



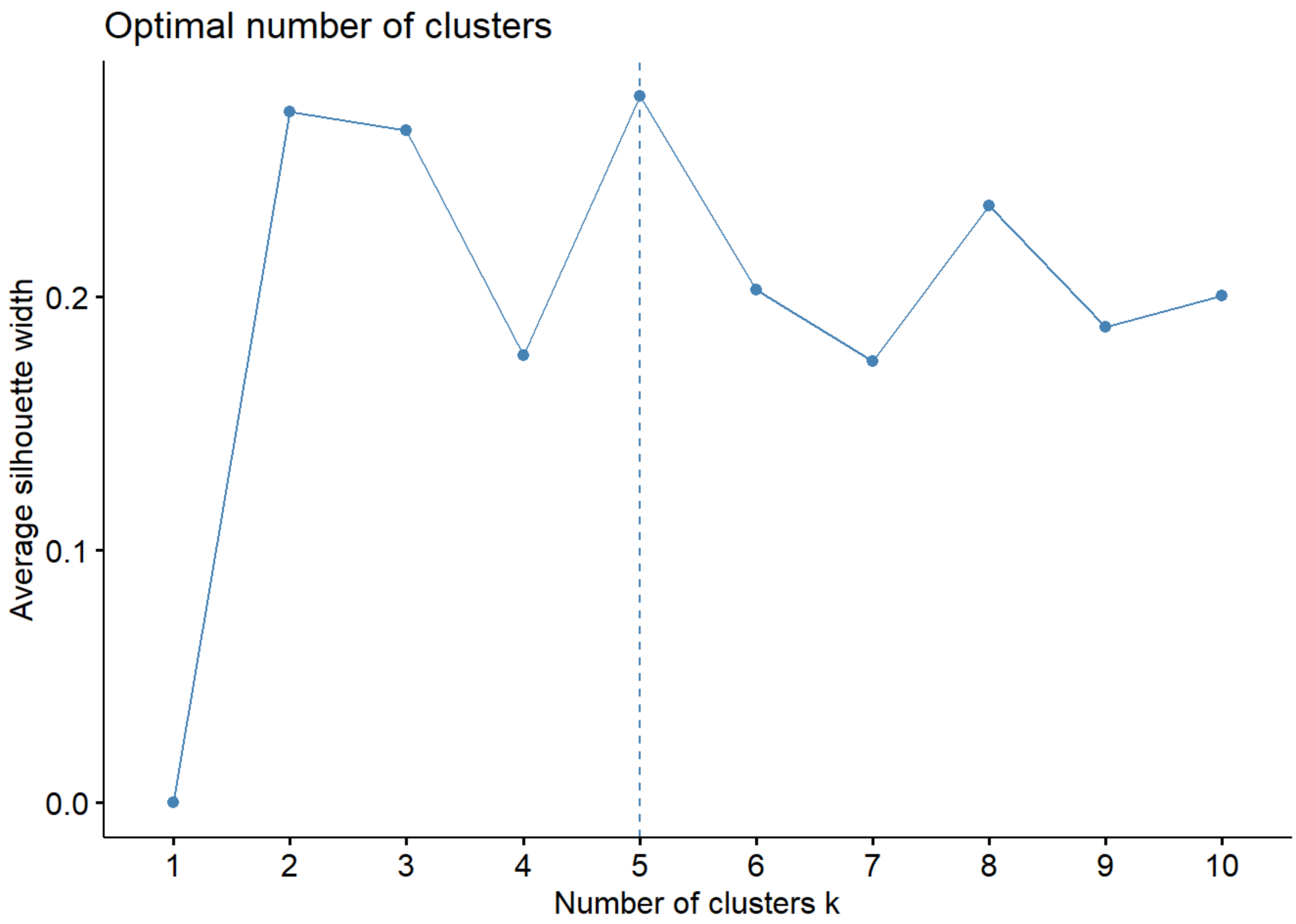
```
kmeans_wss <- kmeans(tiru_new, centers = 2, nstart = 10)

fviz_cluster(kmeans_wss, data = tiru_new) + ggtitle("K-means Clustering Visualization")
```



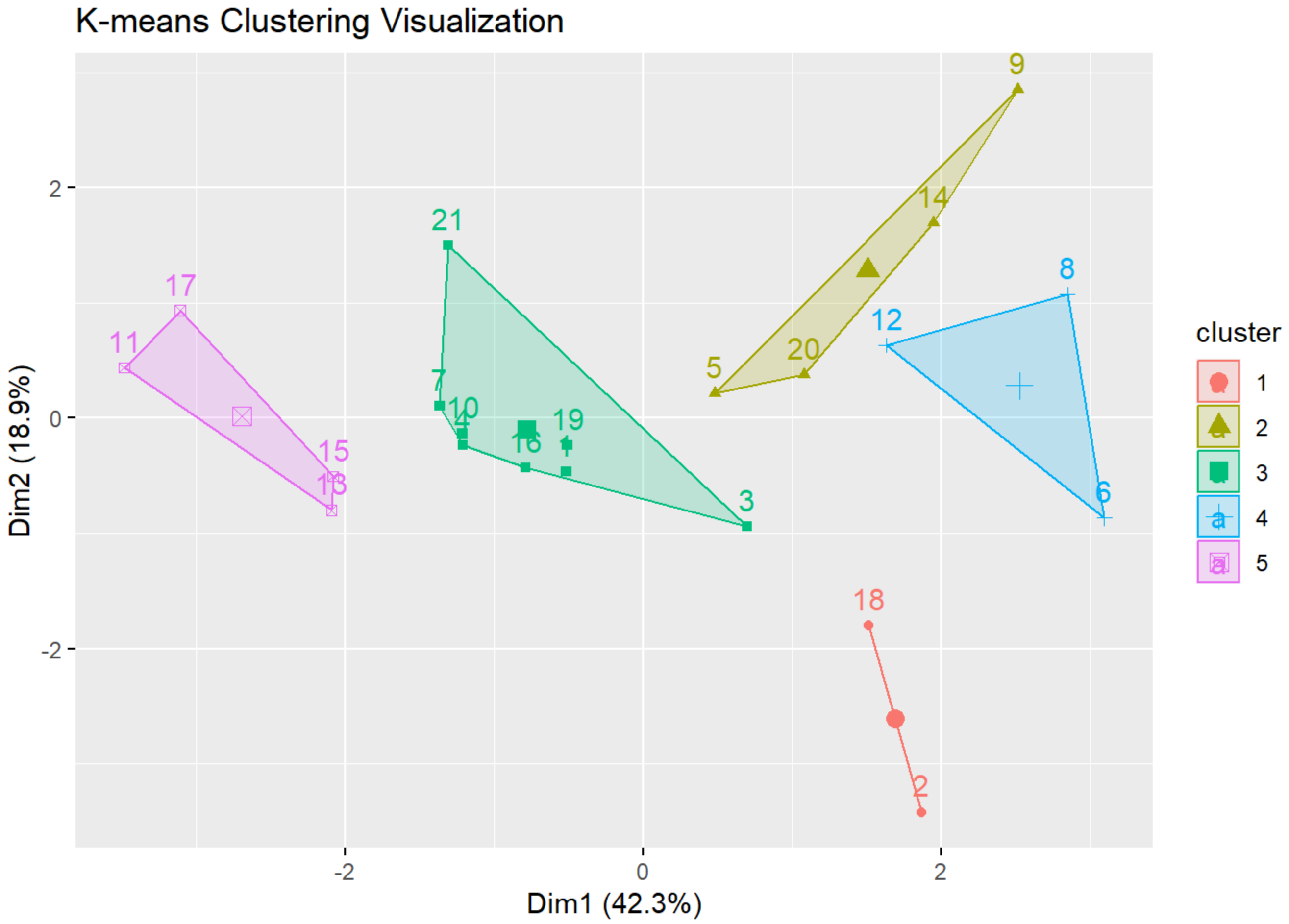
```
print(kmeans_wss)
```

```
## 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"
```

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



```
kmeans_silhou <- kmeans(tiru_new, centers = 5, nstart = 25)

fviz_cluster(kmeans_silhou, data = tiru_new) + ggtitle("K-means Clustering Visualization")
```



```
print(kmeans_silhou)
```

```
## K-means clustering with 5 clusters of sizes 2, 4, 8, 3, 4
## 
## Cluster means:
##    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
## 
## Clustering vector:
##  [1] 3 1 3 3 2 4 3 4 2 3 5 4 5 2 5 3 5 1 3 2 3
## 
## Within cluster sum of squares by cluster:
## [1]  2.803505 12.791257 21.879320 15.595925  9.284424
##  (between_SS / total_SS =  65.4 %)
## 
## Available components:
## 
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

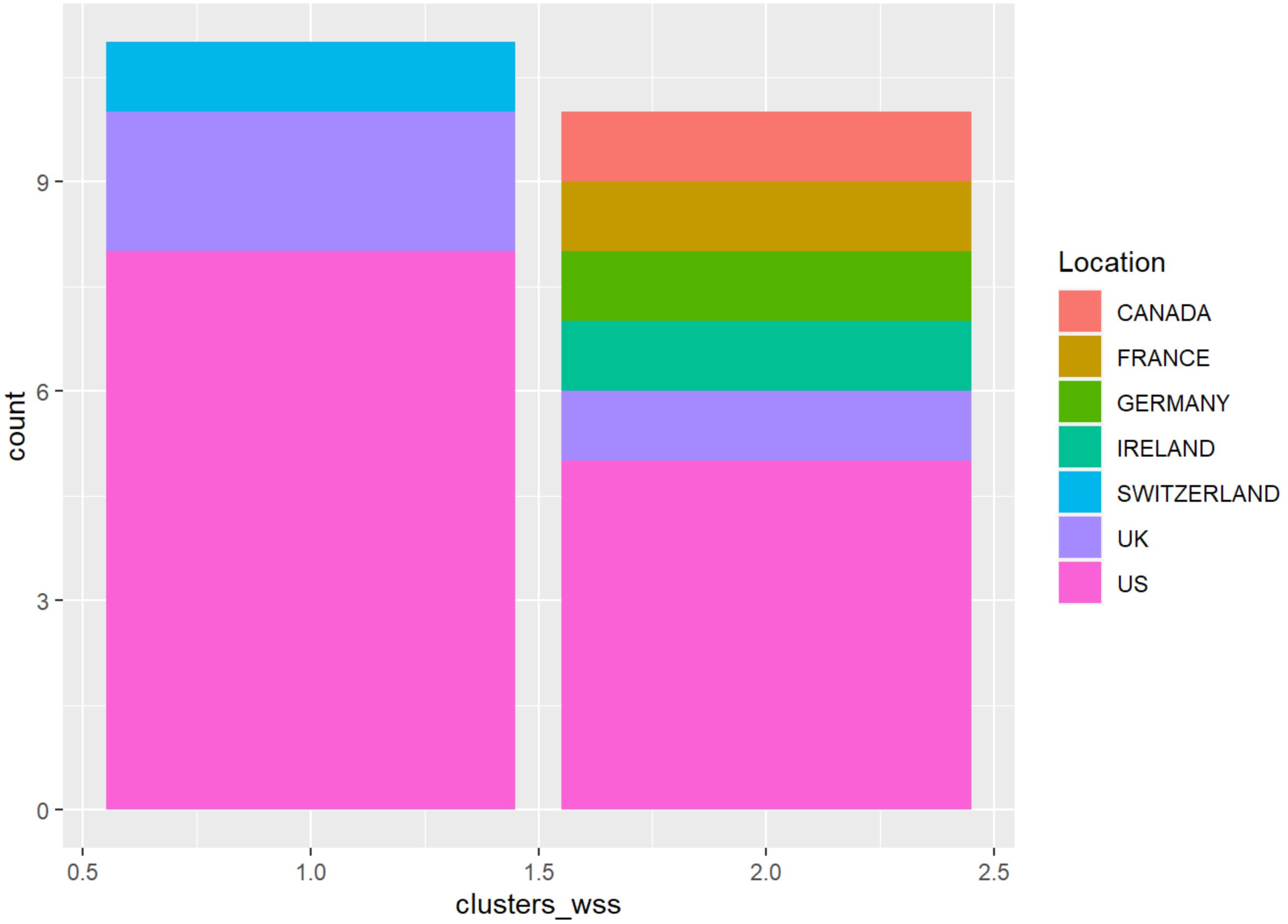
#Interpret the clusters with respect to the numerical variables used in forming the clusters. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

```
clusters_wss <- kmeans_wss$cluster
clusters_silhouette <- kmeans_silhou$cluster

temp_data_1 <- cbind(tiru,clusters_wss)
temp_data_2 <- cbind(tiru,clusters_silhouette)
```

```
int_wss <- aggregate(temp_data_1[,-c(1:2,12:14)],by = list(temp_data_1$clusters_wss),FUN="media
n")
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
```

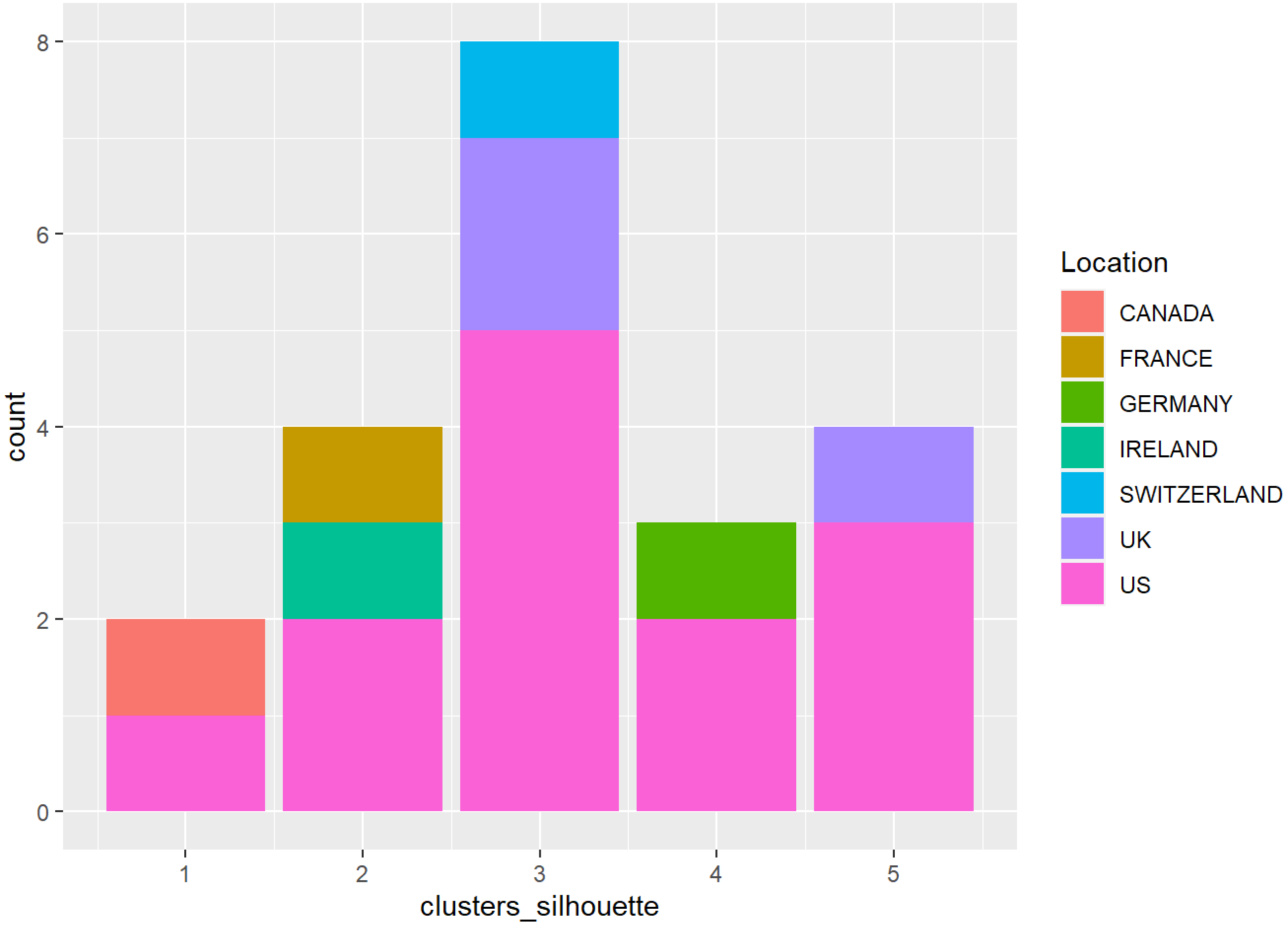
```
#pattern in catergorical variables
ggplot(temp_data_1,aes(x=clusters_wss,fill=Location)) + geom_bar()
```



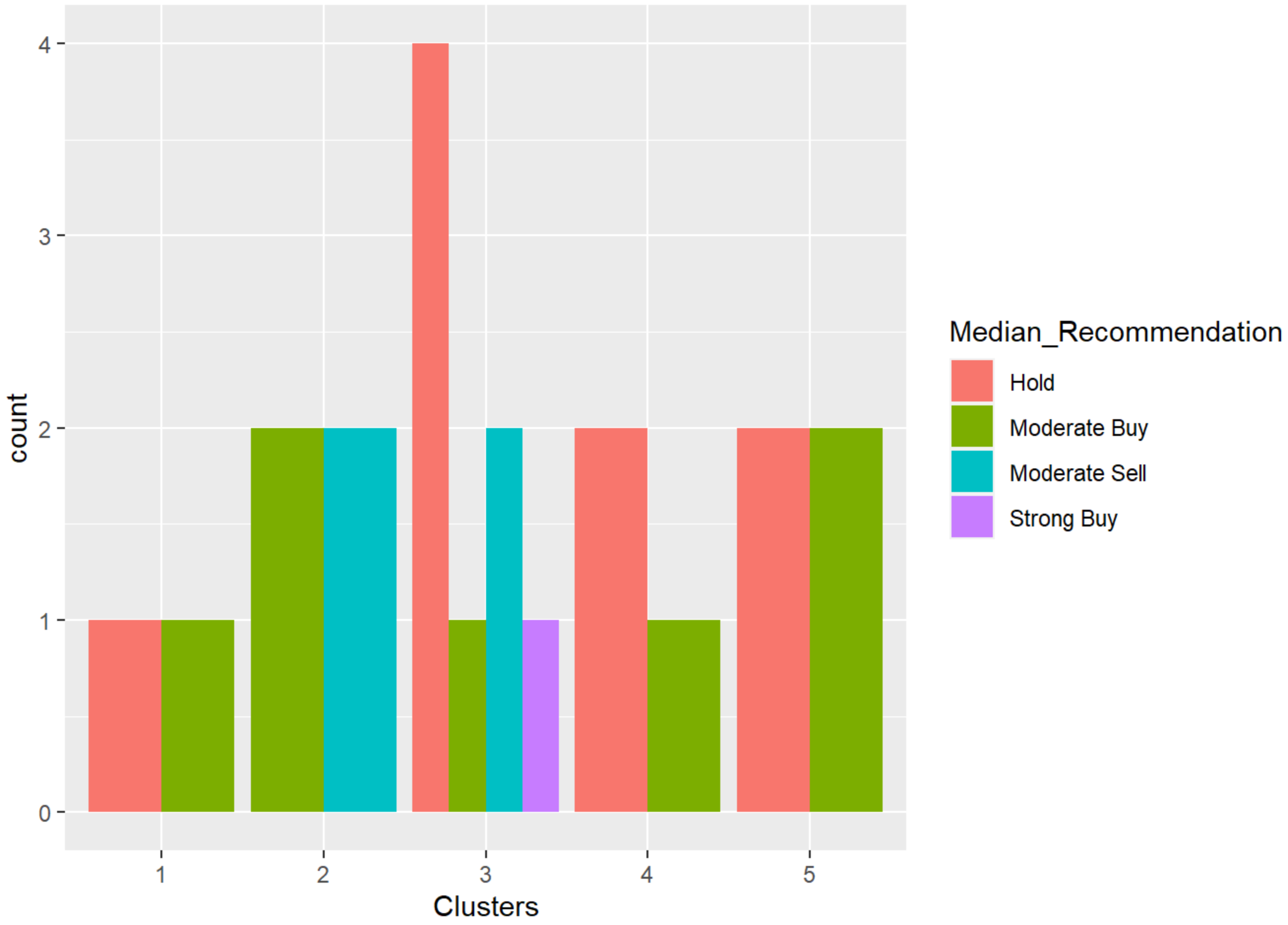
```
int_silhou <- aggregate(temp_data_2[,-c(1:2,12:14)],by=list(temp_data_2$clusters_silhouette),FUN
="median")
print(int_silhou[,-1])
```

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

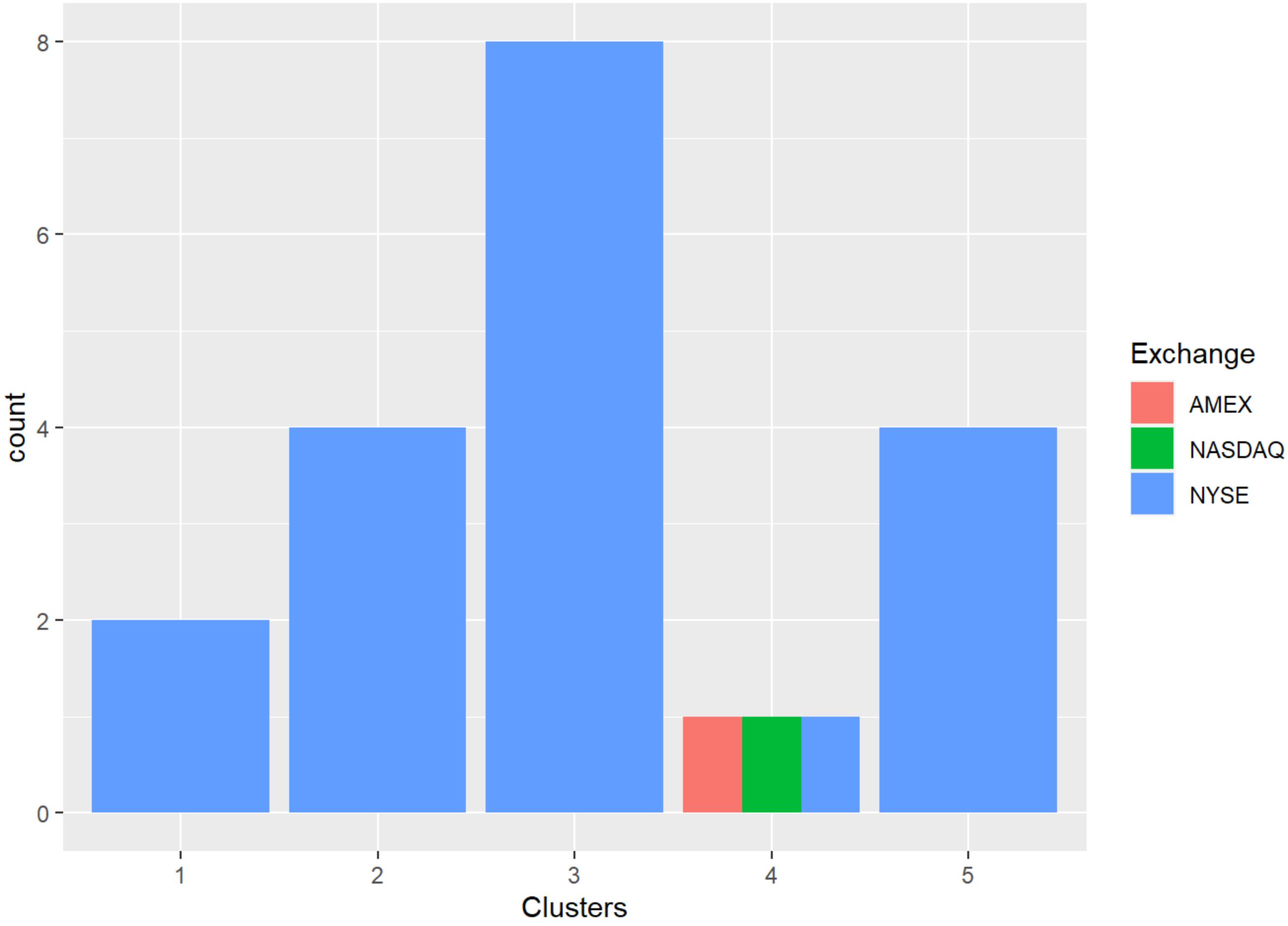
```
ggplot(temp_data_2,aes(x=clusters_silhouette, fill = Location)) + geom_bar()
```



```
temp_data_4 <- tiru[12:14] %>% mutate(Clusters=kmeans_silhou$cluster)

ggplot(temp_data_4, mapping=aes(factor(Clusters),fill=Median_Recommendation))+geom_bar(position
='dodge')+labs(x ='Clusters')
```



```
ggplot(temp_data_4, mapping = aes(factor(Clusters),fill = Exchange))+geom_bar(position ='dodge')
+labs(x ='Clusters')
```


count
8
6
4
2
0
1
2
3
4
5
Clusters
Exchange
AMEX
NASDAQ
NYSE