ASSIGNMENT4

Nikitha Kotagiri

2023-11-13

SUMMARY:

The variable "niki" is filled with the Pharmaceuticals data set. We utilized the head function to determine whether or not the data set had been appropriately loaded.Let us now move on to the challenges that we must fix.

1: For cluster analysis, I used 9 numerical variables (3 to 11 columns) from the dataset.First, we obtained a summary of all 9 numerical variables that we are employing. Finding and displaying the distance between the rows matrix.Then we begin our clustering analysis with the wss and silhouette approaches.

2: In this case, I utilize the inside sum of squares and Silhouette techniques to determine the ideal number of clusters to construct.The wss and Silhouette techniques are used to determine the ideal amount of clusters to generate.

2.1: Within Sum of Squares: The graph resembles a human hand with a bend similar to our elbow.The precise place in the graph when their will be less decrease. Looking at the graph, we can see that the rate of decline in wss slows down around "k=2" (this would be the ideal answer).

The lower the Wss number, the tighter the clusters produced.The ideal wss value is 0.Furthermore, if identifying the best solution for specific data sets is challenging, we will use different approaches.

2.2: Silhouette method: We may discover the best option by looking for the peak of the graph when the silhouette coefficient is at its highest value. We can observe from our graph that the curve reached its maximum point at k = 5. This implies that "k=5" is the best answer for the pharma dataset.

If the silhouette distance is 1, the datapoints are appropriately allocated to the cluster; if it is -1, the datapoints are not properly assigned.

Sometimes the best solution comes from a combination of both strategies.Then you must follow the other ways, or we must pick which one to use depending on the findings of the cluster summary. Wss technique:

-Based on the Wss clustering study, which produced two clusters, we may deduce the following. Cluster 1: Profitable with a Moderate risk.The initial cluster discovered here has a high success rate, making it a good investment.The metrics listed below are used to Asset turnover, return on assets (ROA), return on expenses (ROE), and net profit margin are several ways to quantify success.This When the investment is large, cluster has a capital value of 73.84, a return on equity (ROE) of 31, and a return on investment (ROI) of 0.assets (ROA) of 15, which shows the profit expected from a company's high asset investments.In a In a similar vein, both net profit and asset turnover are high.The fact that the PE Ratio of the first company is lower than that of the second cluster implies In general, the beta value should be smaller than one, suggesting that the These businesses' variability is modest and lacks adequate variations. Furthermore, a company's "Leverage" value (the amount of cash borrowed for an investment) should be as low as possible. Because the market is constantly unpredictable, there is a possibility that the money borrowed for the investment would be lost although it was expected to provide gains. The leverage value in this case is 0.28, which is smaller than in the second cluster. "With a good investment, there should be very little chance of losing the entire amount invested," and enterprises in this cluster are reporting better success rates than those in the second cluster. Cluster 2: High risk, low profit. In this situation, the second cluster's performance measurements are inferior to those of the first. Its market capitalization is exceptionally low, 4.78 vs 73.84 in the first cluster, indicating that the firms listed in this cluster have a lower market share than the companies listed in the first cluster. Return on Equity (ROE), Return on Assets (ROA), Asset Turnover, and Net Profit Margin all experience drops in return on investment. The degree of hazard, which is reinforced by these enterprises' high leverage and beta values, suggesting a high degree of unpredictability and high borrowing rates as compared to the first cluster. In comparison, the PE Ratio is high. -> From the graph, we can see that the majority of pharmaceutical industry enterprises are headquartered in the United States, and we can observe a similar trend in clusters 1 and 2. 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 contain a higher proportion of enterprises headquartered in the United States.

Method of Silhouette: -

We may deduce the following from the Silhouette clustering study, which produced five clusters. Cluster 1: The First Cluster looks to be overhyped. The PE Ratio appears to be highly flexible, measuring the share price in proportion to the company's worth and indicating whether or not the stock is overpriced. Furthermore, this group has significant beta and leverage levels, indicating that there is associated risk. There must be a better investing opportunity than this for an investment.

Cluster 2:When it is concerned with providing returns on investment—basically, the value that any investor would want as a return on investment. There is also a significant amount of external borrowing and a reasonable amount of business variability (beta). Furthermore, its capital worth is the lowest of all the categories. Surprisingly, these companies also have the most income. This might be because the firms are young and need to establish themselves before moving into the market.

Cluster 3: The Destiny Class's third cluster consists of firms with a decent market capitalization, an acceptable PE ratio, and moderate degrees of risk (beta and leverage). Furthermore, it has assets with a lucrative propensity and higher returns on investment. Even if the capital value is smaller in comparison, it may still be a suitable investment option because the valuation may change or improve in the future.

Cluster 4: The Cluster is a very unpredictable cluster with greater beta (firm variability) and leverage (outside borrowings) values, indicating that these enterprises have a strong feeling of risk. Furthermore, due to its smaller market capitalization and net profit margin, it is less suitable for future investments.

Cluster 5:Anyone wishing to establish a lucrative pitch might consider investing in the Fourth Cluster. It has the "Highest Market Capital" of 153.245 in this cluster, 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 contrast to other companies in distinct clusters. It also has a "less leverage value," which indicates that little borrowed cash will be required for future investments, and a "decent beta value," which indicates that there will be less fluctuation and risk associated.A corporation having a greater capital ratio, moderate risk, and a positive cash flow. and having fewer obligations is a favorable option for investors.Companies in this cluster choose the best choice. The wss and silhouette clusters show a comparable degree of patterning toward the site.

When compared to the other locations, this one's clusters have a larger percentage of their locations in the "US." - It's worth noting, however, that Cluster 4, the strongest cluster for correctly characterizing the domain, has a greater share of US-based enterprises than non-US-based businesses.Other observations include .There is one strong buy, seven moderate buys, nine holds, and four moderate sells for a total of 21 recommendations. Cluster combines all four suggestions, including opposing advice on buys and sells. Group 3. Clusters 1, 4, and 5 include just mod buy and hold information.Cluster 2 has both a moderate buy and a moderate sell recommendation. There are 21 businesses in all, with 13 in the United States, three in the United Kingdom, and one each in Canada, France, Germany, Ireland, and Switzerland. Cluster 3 includes the United States, the United Kingdom, and Switzerland. Germany and the United States are in Cluster 4. Cluster 1 includes the United States and Canada. Cluster 5 includes the United States and the United Kingdom. Cluster 2 consists of the United States, France, and Ireland.

There are 21 corporations in all, including 1 Amex, 1 Nasdaq, and 19 NYSE. Cluster 4 includes all three. Only NYSE is found in clusters 1,2,3,5.

3: Using any or all of the variables in the dataset, give each cluster a suitable name. Cluster 1:Non-plus Organization (Hold) Cluster 2: Moderate Compensation (Reduced) Cluster 3:Destiny class (Moderate) Cluster 4: Excessive investment (Hold) Cluster 5:High Margins (Strong Buy) Conclusion: Finally, every individual or business aspires to maximize their profit while incurring the fewest losses. They also anticipate the investment's long-term success. Based on my findings, Cluster 5 is the greatest option for investment. It provides larger rewards and a longer term. Cluster 3 is the other cluster I recommend. It has marginal gains that are risky but have a higher possibility of becoming profitable. The following clusters are not recommended for any company or venture capitalists since they incur losses or yield no marginal gains when invested in

```
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 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0

## -- 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 error
```

```
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.

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

```
##   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(niki)
```

```
## ’data.frame’:    21 obs. of  14 variables:
##  $ Symbol               : chr  "ABT" "AGN" "AHM" "AZN" ...
##  $ Name                 : chr  "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZeneca PL
##  $ Market_Cap           : num  68.44 7.58 6.3 67.63 47.16 ...
##  $ Beta                 : num  0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...
##  $ PE_Ratio             : num  24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...
##  $ ROE                  : num  26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...
##  $ ROA                  : num  11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...
##  $ Asset_Turnover       : num  0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
##  $ Leverage             : num  0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...
##  $ Rev_Growth           : num  7.54 9.16 7.05 15 26.81 ...
##  $ Net_Profit_Margin    : num  16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...
##  $ Median_Recommendation: chr  "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sell" ...
##  $ Location             : chr  "US" "CANADA" "UK" "UK" ...
##  $ Exchange             : chr  "NYSE" "NYSE" "NYSE" "NYSE" ...
```

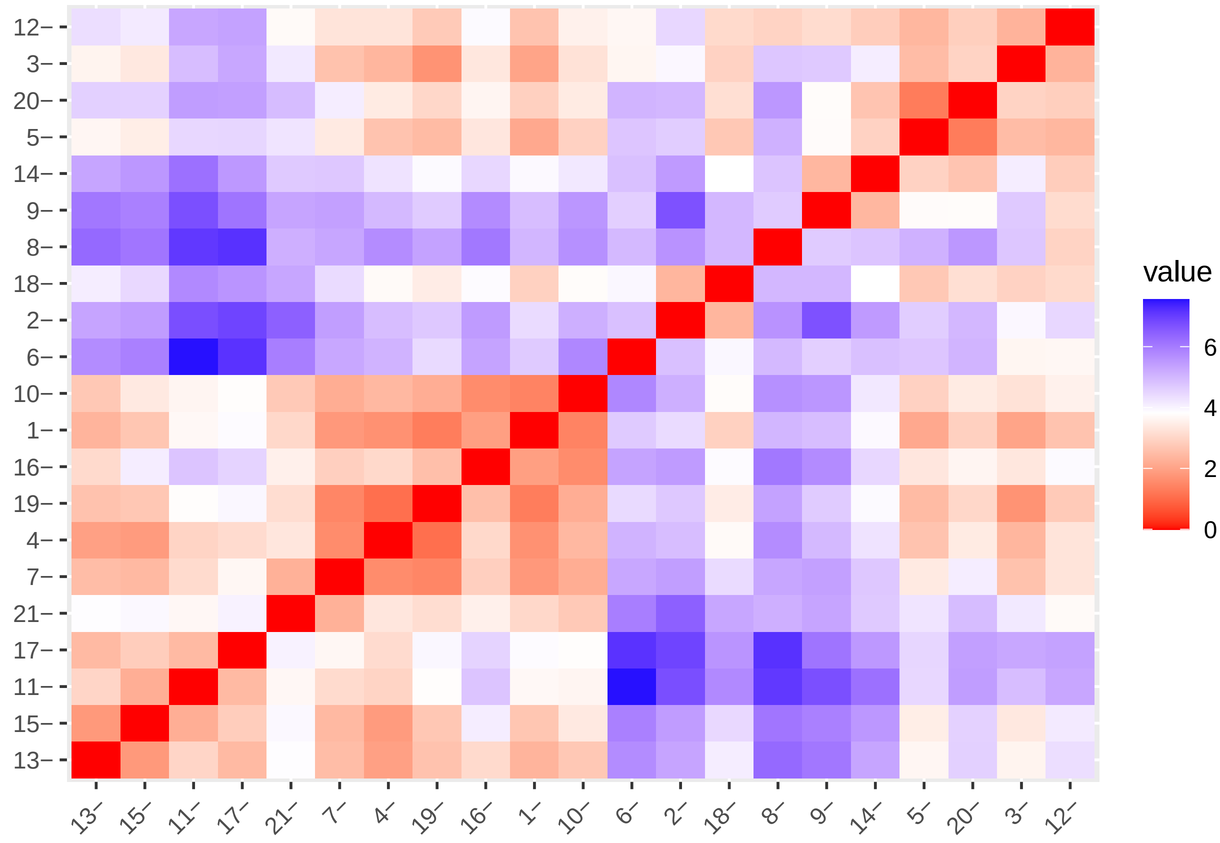
```
na.omit(niki)
```

```
##    Symbol                               Name Market_Cap Beta PE_Ratio  ROE  ROA
## 1     ABT                Abbott Laboratories      68.44 0.32     24.7 26.4 11.8
## 2     AGN                     Allergan, Inc.       7.58 0.41     82.5 12.9  5.5
## 3     AHM                       Amersham plc       6.30 0.46     20.7 14.9  7.8
## 4     AZN                    AstraZeneca PLC      67.63 0.52     21.5 27.4 15.4
## 5     AVE                            Aventis      47.16 0.32     20.1 21.8  7.5
## 6     BAY                           Bayer AG      16.90 1.11     27.9  3.9  1.4
## 7     BMY       Bristol-Myers Squibb Company      51.33 0.50     13.9 34.8 15.1
## 8    CHTT                       Chattem, Inc       0.41 0.85     26.0 24.1  4.3
## 9     ELN              Elan Corporation, plc       0.78 1.08      3.6 15.1  5.1
## 10    LLY              Eli Lilly and Company      73.84 0.18     27.9 31.0 13.5
## 11    GSK                GlaxoSmithKline plc     122.11 0.35     18.0 62.9 20.3
## 12    IVX                   IVAX Corporation       2.60 0.65     19.9 21.4  6.8
## 13    JNJ                  Johnson & Johnson     173.93 0.46     28.4 28.6 16.3
## 14    MRX Medicis Pharmaceutical Corporation       1.20 0.75     28.6 11.2  5.4
## 15    MRK                  Merck & Co., Inc.     132.56 0.46     18.9 40.6 15.0
## 16    NVS                        Novartis AG      96.65 0.19     21.6 17.9 11.2
## 17    PFE                         Pfizer Inc     199.47 0.65     23.6 45.6 19.2
## 18    PHA              Pharmacia Corporation      56.24 0.40     56.5 13.5  5.7
## 19    SGP        Schering-Plough Corporation      34.10 0.51     18.9 22.6 13.3
## 20    WPI       Watson Pharmaceuticals, Inc.       3.26 0.24     18.4 10.2  6.8
## 21    WYE                              Wyeth      48.19 0.63     13.1 54.9 13.4
##    Asset_Turnover Leverage Rev_Growth Net_Profit_Margin Median_Recommendation
## 1             0.7     0.42       7.54              16.1          Moderate Buy
## 2             0.9     0.60       9.16               5.5          Moderate Buy
## 3             0.9     0.27       7.05              11.2            Strong Buy
## 4             0.9     0.00      15.00              18.0         Moderate Sell
## 5             0.6     0.34      26.81              12.9          Moderate Buy
## 6             0.6     0.00      -3.17               2.6                  Hold
## 7             0.9     0.57       2.70              20.6         Moderate Sell
## 8             0.6     3.51       6.38               7.5          Moderate Buy
## 9             0.3     1.07      34.21              13.3         Moderate Sell
## 10            0.6     0.53       6.21              23.4                  Hold
## 11            1.0     0.34      21.87              21.1                  Hold
## 12            0.6     1.45      13.99              11.0                  Hold
## 13            0.9     0.10       9.37              17.9          Moderate Buy
## 14            0.3     0.93      30.37              21.3          Moderate Buy
## 15            1.1     0.28      17.35              14.1                  Hold
## 16            0.5     0.06      -2.69              22.4                  Hold
## 17            0.8     0.16      25.54              25.2          Moderate Buy
## 18            0.6     0.35      15.00               7.3                  Hold
## 19            0.8     0.00       8.56              17.6                  Hold
## 20            0.5     0.20      29.18              15.1         Moderate Sell
## 21            0.6     1.12       0.36              25.5                  Hold
##       Location Exchange
## 1           US     NYSE
## 2       CANADA     NYSE
## 3           UK     NYSE
## 4           UK     NYSE
## 5       FRANCE     NYSE
## 6      GERMANY     NYSE
## 7           US     NYSE
## 8           US   NASDAQ
## 9      IRELAND     NYSE
## 10          US     NYSE
## 11          UK     NYSE
## 12          US     AMEX
## 13          US     NYSE
## 14          US     NYSE
## 15          US     NYSE
## 16 SWITZERLAND     NYSE
## 17          US     NYSE
## 18          US     NYSE
## 19          US     NYSE
## 20          US     NYSE
## 21          US     NYSE
```

```
niki_new <- scale(niki[,3:11])
summary(niki_new)
```

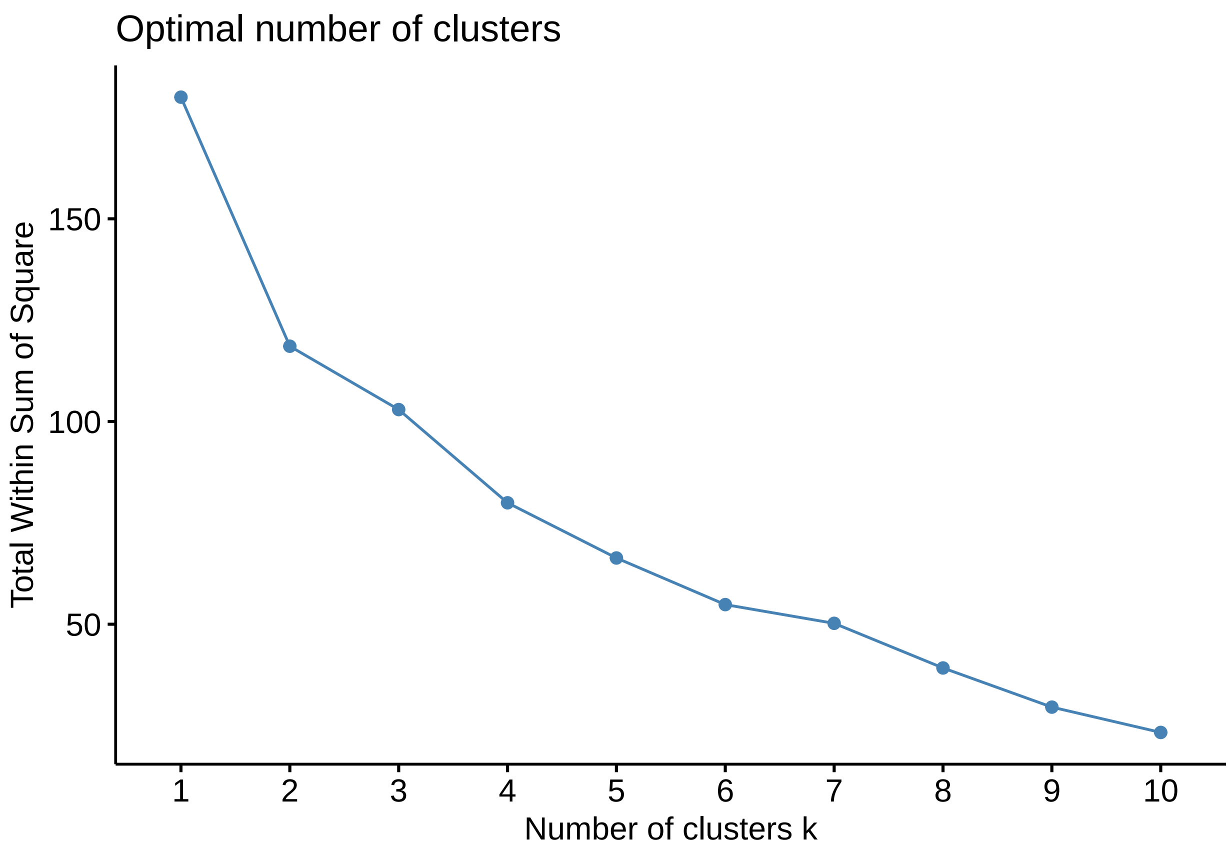
```
##    Market_Cap          Beta            PE_Ratio            ROE
##  Min.   :-0.9768  Min.   :-1.3466  Min.   :-1.3404  Min.   :-1.4515
##  1st Qu.:-0.8763  1st Qu.:-0.6844  1st Qu.:-0.4023  1st Qu.:-0.7223
##  Median :-0.1614  Median :-0.2560  Median :-0.2429  Median :-0.2118
##  Mean   : 0.0000  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(niki_new, method = "euclidian")
fviz_dist(Distance)
```

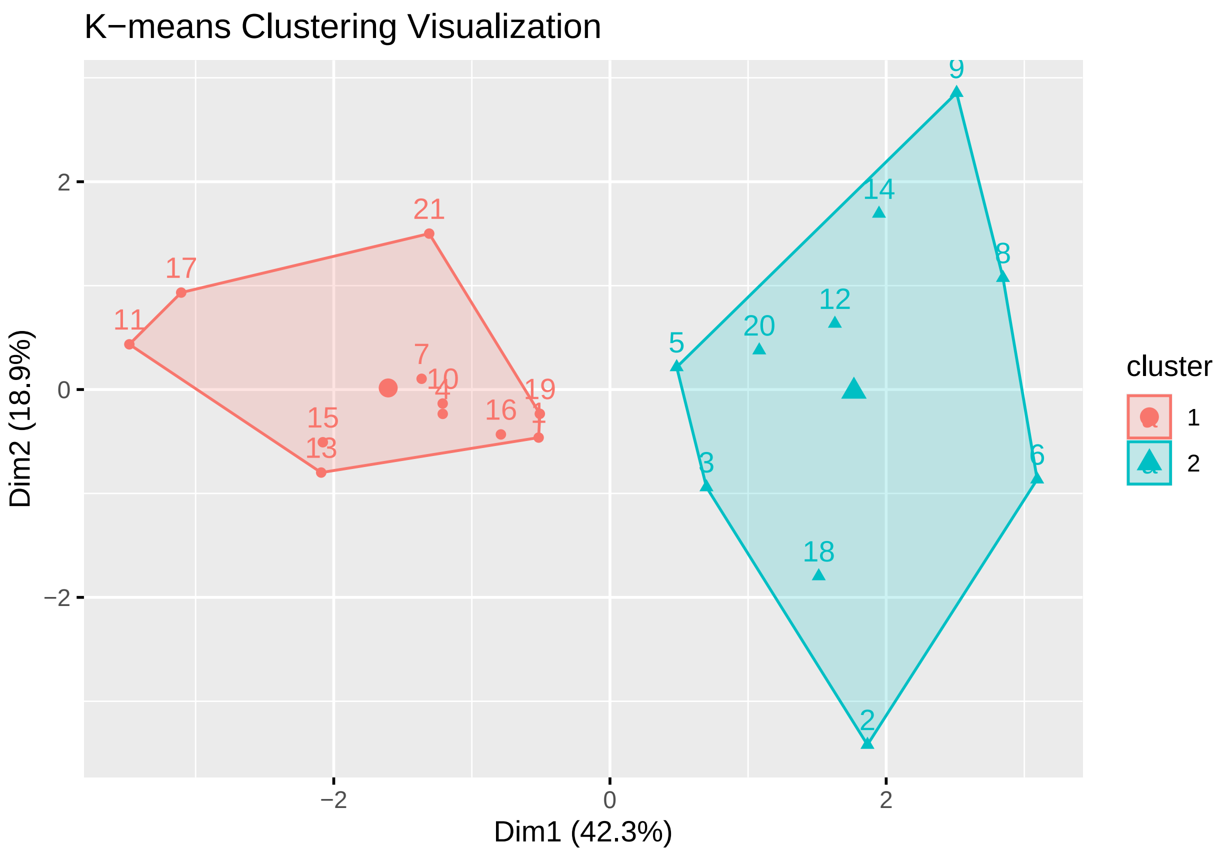


#Applying k_means clustering for the question.

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



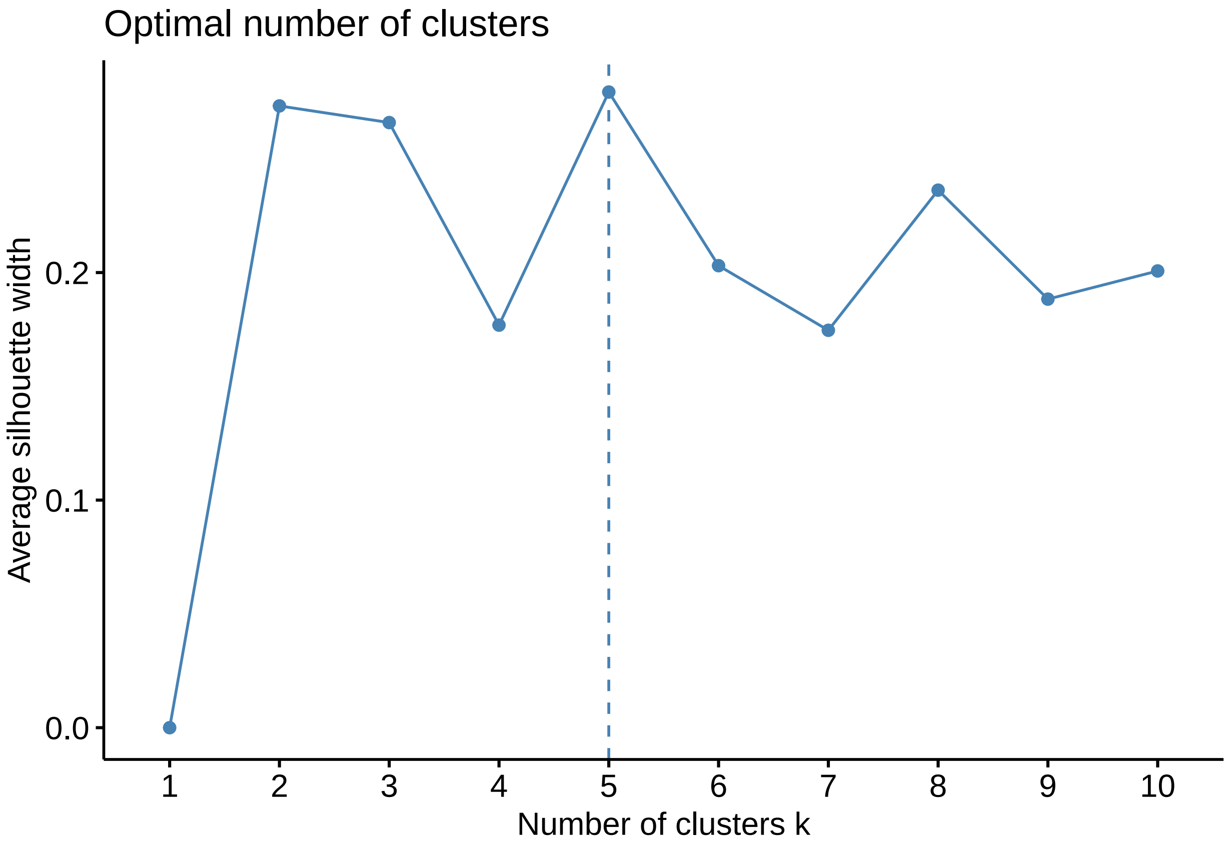
```r
kmeans_ab <- kmeans(niki_new, centers = 2, nstart = 20)

fviz_cluster(kmeans_ab, data = niki_new) + ggtitle("K-means Clustering Visualization")
```



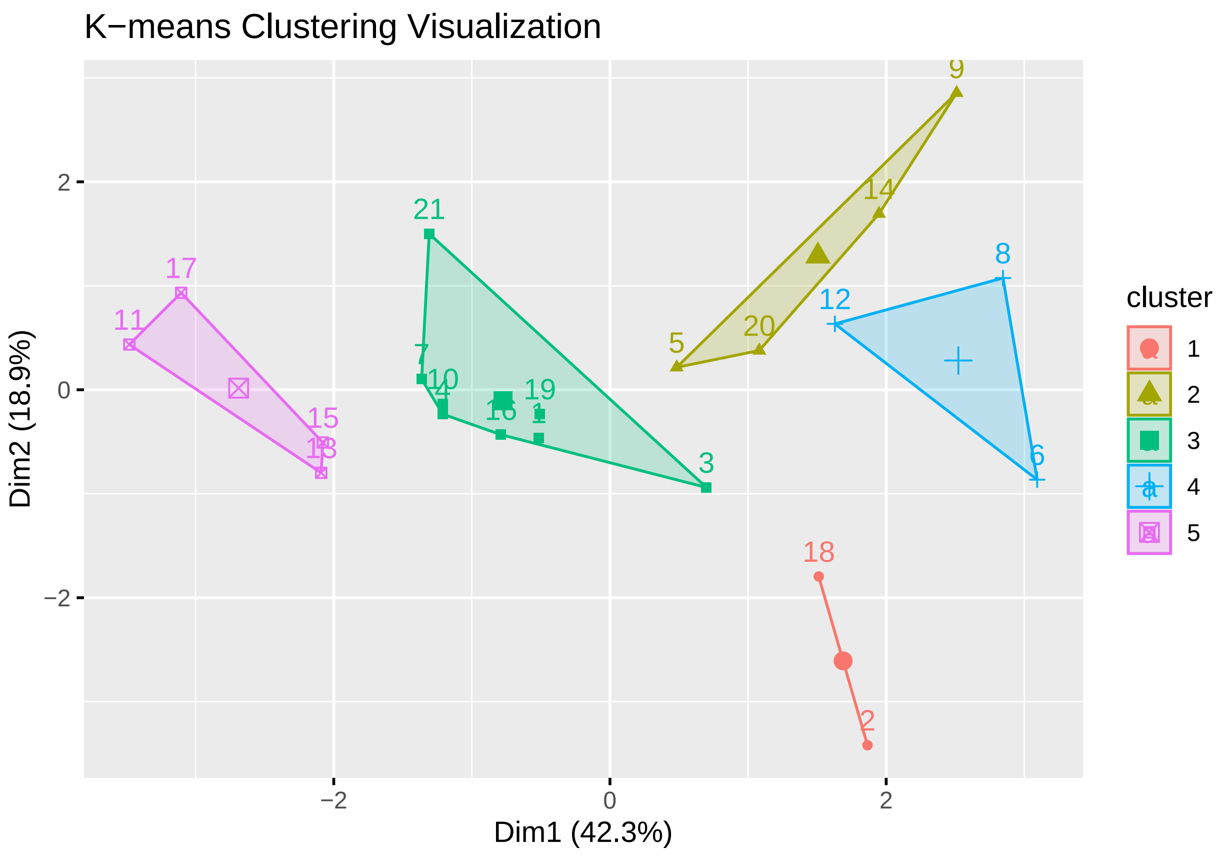
```r
print(kmeans_ab)
```

```
## 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(niki_new, kmeans, method = "silhouette")
```



```
kmeans_silh <- kmeans(niki_new, centers = 5, nstart = 25)

fviz_cluster(kmeans_silh, data = niki_new) + ggtitle("K-means Clustering Visualization")
```



```r
print(kmeans_silh)
```

```
## 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"
```

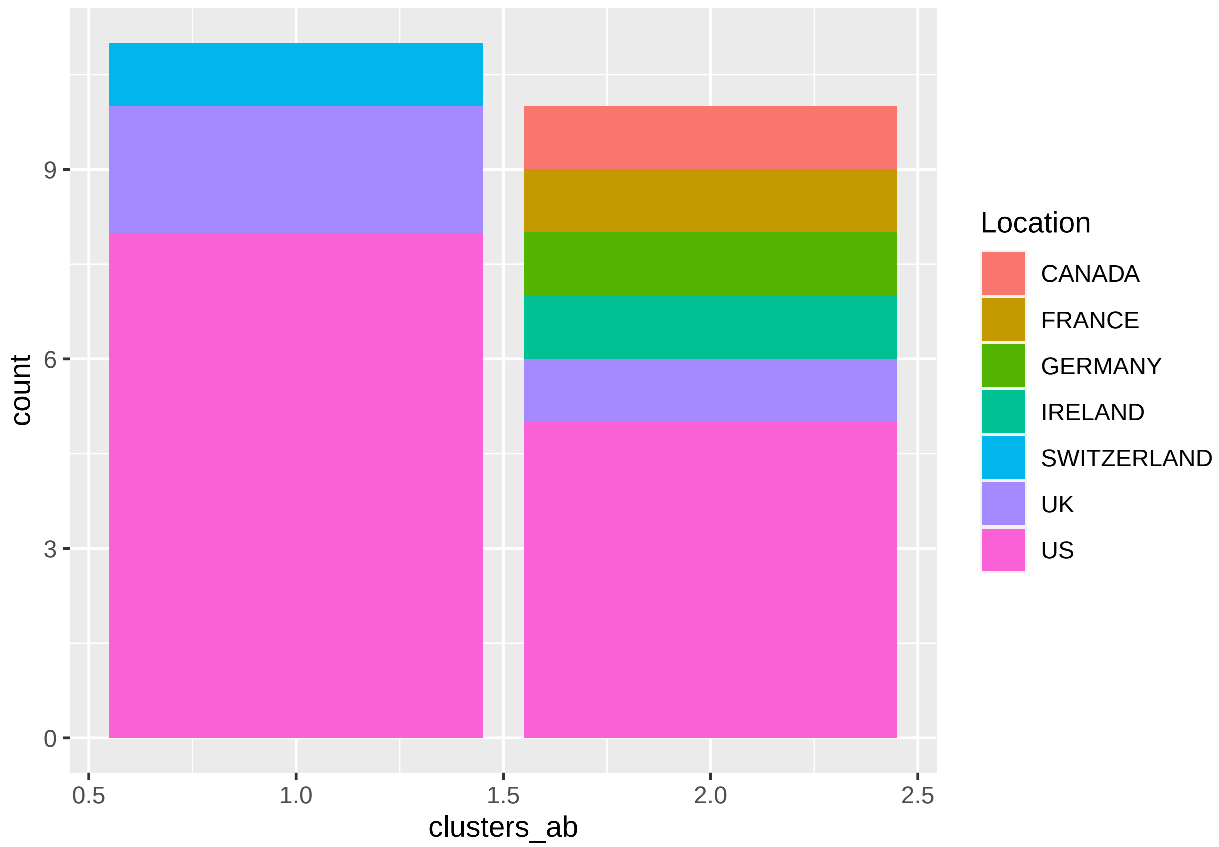
#Interpret the clusters with respect to the numerical variables used in forming the clusters. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

```
clusters_ab <- kmeans_ab$cluster
clusters_silh <- kmeans_silh$cluster

temp_data_11 <- cbind(niki,clusters_ab)
temp_data_22 <- cbind(niki,clusters_silh)
```

```
int_ab <- aggregate(temp_data_11[,-c(1:2,12:14)],by = list(temp_data_11$clusters_ab),FUN="median")
print(int_ab[,-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_ab
## 1              20.6           1
## 2              11.1           2
```

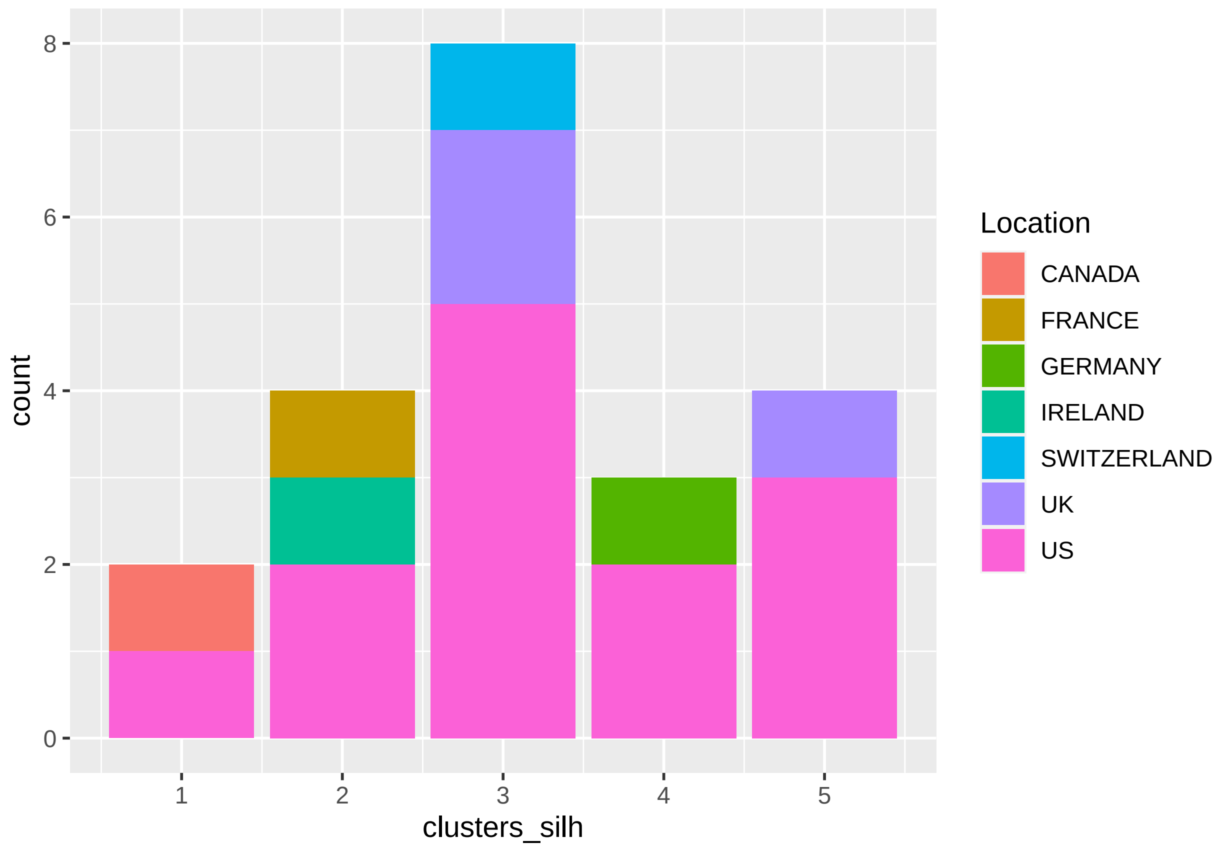
```
#pattern in categorical variables
ggplot(temp_data_11,aes(x=clusters_ab,fill=Location)) + geom_bar()
```



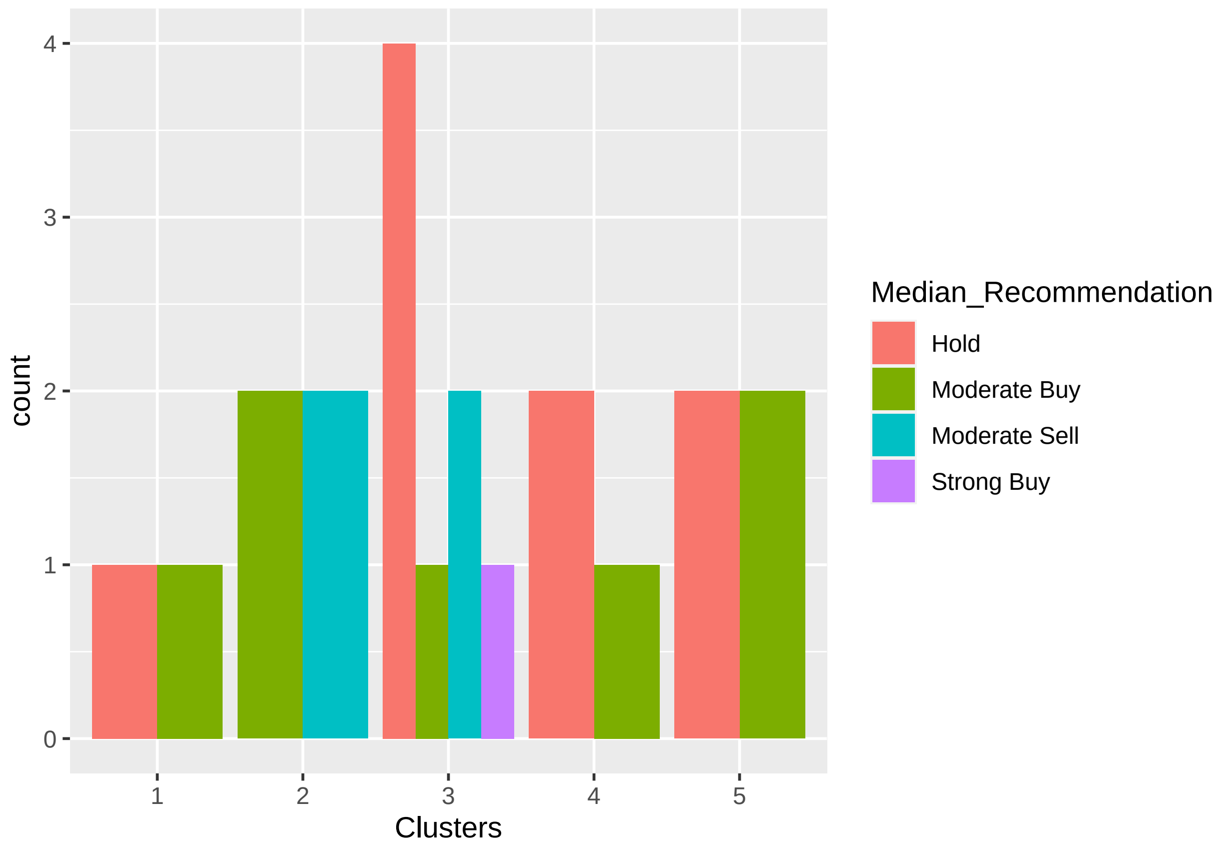
```
int_silh <- aggregate(temp_data_22[,-c(1:2,12:14)],by=list(temp_data_22$clusters_silh),FUN="median")
print(int_silh[,-1])
```

```
##   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_silh
## 1               6.4             1
## 2              14.2             2
## 3              19.3             3
## 4               7.5             4
## 5              19.5             5
```

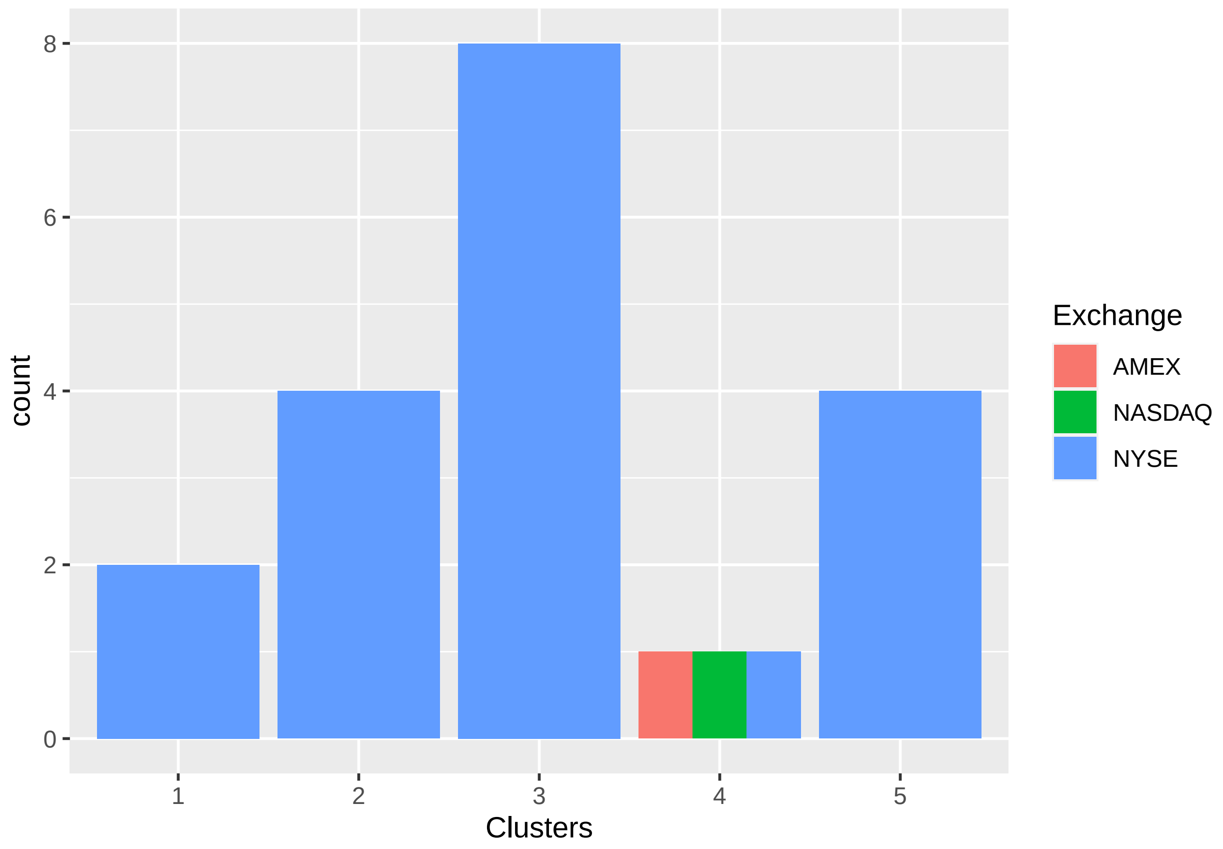
```
ggplot(temp_data_22,aes(x=clusters_silh, fill = Location)) + geom_bar()
```



```
temp_data_3 <- niki[12:14] %>% mutate(Clusters=kmeans_silh$cluster)

ggplot(temp_data_3, mapping=aes(factor(Clusters),fill=Median_Recommendation))+geom_bar(position='dodge')
```



```
ggplot(temp_data_3, mapping = aes(factor(Clusters),fill = Exchange))+geom_bar(position ='dodge')+labs(x
```


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