

Assignment 4: Clustering

Akhil Seliveri

2023-11-11

Summary

Using numerical variables 1 through 9, the R code provided performs cluster analysis on the dataset of 21 pharmaceutical companies. Here is a summary and analysis of the code:

1. The first section of the code loads the required libraries and reads the Pharmaceuticals.csv file containing the dataset. 2. Data for Clustering Analysis:

2.1 Preprocessing:

The NA values are removed and numerical variables from (3-11) are used for clustering.

Z-score standardization is used to choose and standardize numerical variables, ranging from 1 to 9.

2.2 Choosing the Right Number of Clusters:

The ideal number of clusters ($k=5$) is found using the elbow and silhouette methods.

2.3 K-Means Clustering: $K=5$ is used for K-means clustering.

Cluster sizes and centroids are extracted.

fviz_cluster is used to visualize clusters.

Additionally, cluster plots are produced.

2.4 Additional Cluster Analysis:

Distance matrix and clustering are visualized.

The original data is fitted with cluster labels.

Mean values of numerical variables are calculated for each cluster.

3. Interpretation of Clusters:

Clusters are interpreted based on mean values of numerical variables. Characteristics of each cluster are described, such as high or low values for specific financial metrics.

4. Analysis of Categorical Variables (12 to 14):

Categorical variables (Location, Exchange, Median Recommendation) are explored with respect to cluster labels. Bar plots are created to visualize the distribution of locations, exchanges, and median recommendations within each cluster.

Summary, Based on categorical values:

Group-1: UK and US Stability

Firms in this group, located in the UK and US, prefer the NYSE for stability (largest participants) and have conservative median recommendations of Hold and Moderate Buy.

Group-2: North American Participation

Firms in this group, from Canada and the US, use the NYSE (more participants) and receive median recommendations of Median Buy and Hold.

Group-3: Global Presence with Varied Recommendations

Firms in this group, with locations in Switzerland, the UK, and the US, are globally present. They predominantly use NYSE and receive diverse median recommendations, including Hold, Moderate Buy, Relatively High for Moderate Sell, and Strong Buy.

Group-4: International Presence with Multiple Exchanges

Firms in this group, from Germany and the US, participate in multiple exchanges (NYSE, AMEX, NASDAQ) and receive recommendations to Hold or Buy Moderately.

Group-5: Diverse International Presence

Firms in this group, from France, Ireland, and the US, choose the NYSE (largest participants) and receive median recommendations of Median Buy and Sell.

5. Naming Clusters:

Based on the analysis of numerical variables, clusters are named:

Cluster 1: High Risk, Low Performance

Firms in this cluster have high financial risk (high beta and leverage) and exhibit lower financial performance metrics, including market capitalization, ROE, ROA, revenue growth, and net profit margin.

Cluster 2: Profitable and Stable

Firms in this cluster are characterized by high net profit margins and low beta, suggesting stability and profitability in their financial performance.

Cluster 3: Growth Opportunities

Firms in this cluster show potential for growth, with high revenue growth. They also have lower PE ratios and asset turnover, indicating a focus on long-term growth rather than short-term profitability.

Cluster 4: Large and Efficient

Firms in this cluster are larger in terms of market capitalization and exhibit high efficiency, as reflected in high ROE, ROA, and asset turnover.

Cluster 5: Premium Valuation

Firms in this cluster are characterized by a high PE ratio, suggesting that they are valued relatively higher compared to earnings.

Problem Statement:

An equities analyst is studying the pharmaceutical industry and would like your help in exploring and understanding the financial data collected by her firm. Her main objective is to understand the structure of the pharmaceutical industry using some basic financial measures. Financial data gathered on 21 firms in the pharmaceutical industry are available in the file Pharmaceuticals.csv. For each firm, the following variables are recorded:

1. Market capitalization (in billions of dollars)
2. Beta
3. Price/earnings ratio
4. Return on equity
5. Return on assets
6. Asset turnover
7. Leverage
8. Estimated revenue growth
9. Net profit margin
10. Median recommendation (across major brokerages)
11. Location of firm's headquarters
12. Stock exchange on which the firm is listed

Use cluster analysis to explore and analyze the given dataset as follows:

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.
2. 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)
3. Provide an appropriate name for each cluster using any or all of the variables in the dataset.

Answers:

First, loading the needed libraries

```
library(ggplot2)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WB
a
```

```
library(flexclust)
```

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

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

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

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

```
library(cluster)
library(tidyverse)
```

```
## — Attaching core tidyverse packages — tidyverse 2.0.0 —
## ✓ dplyr      1.1.3      ✓ 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 become errors
```

```
library(class)
library(e1071)
```

```
##
## Attaching package: 'e1071'
##
## The following object is masked from 'package:flexclust':
##
##      bclust
```

```
library(caret)
```

```
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##      lift
```

Reading the data

```
data <- read.csv("/Users/akhilchintu/Downloads/Pharmaceuticals.csv")
head(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

Questions

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 na values
data <- na.omit(data)
head(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

```
#Selecting the mentioned numerical variables:
data1<- data[,3:11]
head(data1)
```

```
##   Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## 1      68.44 0.32    24.7 26.4 11.8           0.7     0.42      7.54
## 2       7.58 0.41    82.5 12.9  5.5           0.9     0.60      9.16
## 3       6.30 0.46    20.7 14.9  7.8           0.9     0.27      7.05
## 4      67.63 0.52    21.5 27.4 15.4           0.9     0.00     15.00
## 5      47.16 0.32    20.1 21.8  7.5           0.6     0.34     26.81
## 6      16.90 1.11    27.9  3.9  1.4           0.6     0.00     -3.17
##   Net_Profit_Margin
## 1              16.1
## 2               5.5
## 3              11.2
## 4              18.0
## 5              12.9
## 6               2.6
```

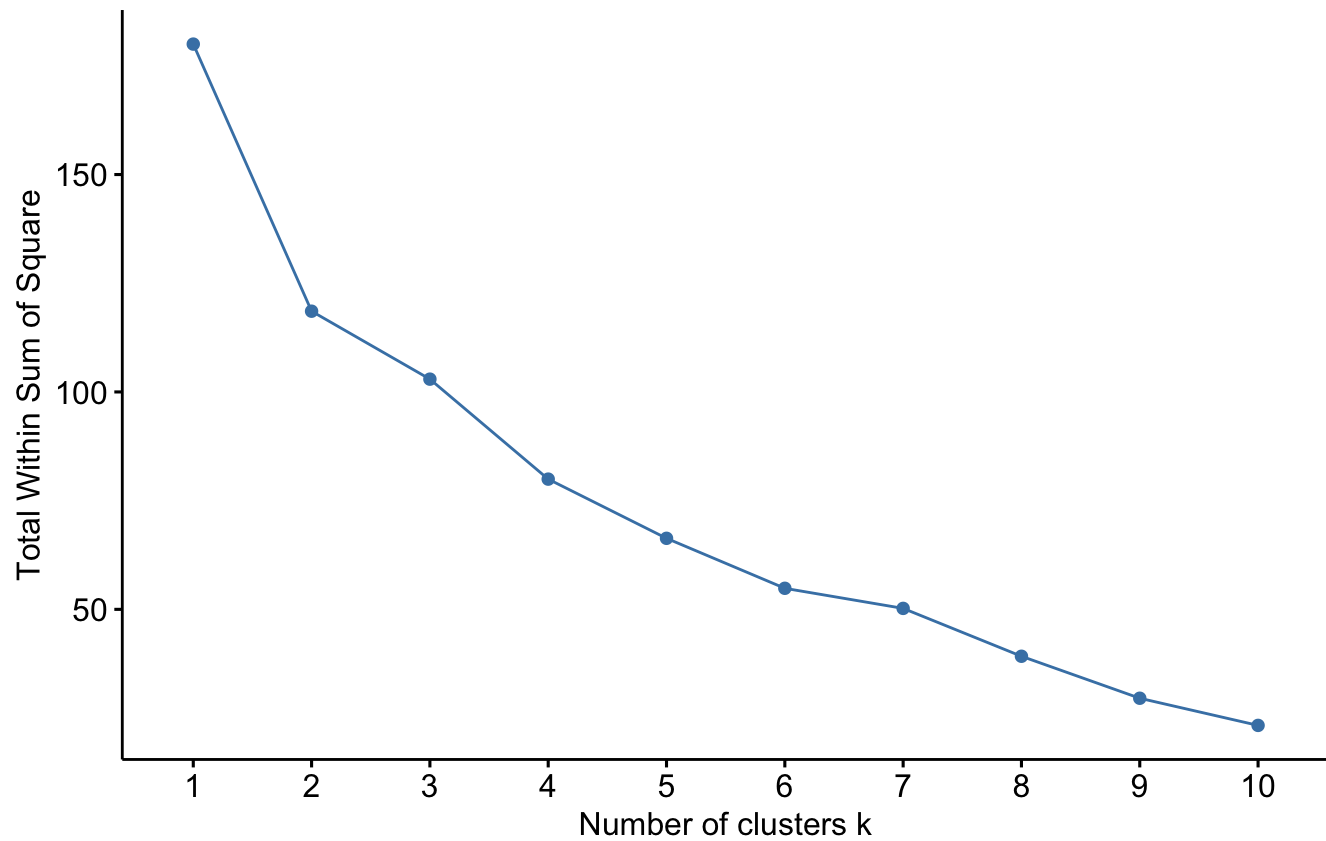
```
#Standardizing the numerical variables using z-score standardization:
data2<-scale(data1)
head(data2)
```

```
##   Market_Cap      Beta    PE_Ratio      ROE      ROA Asset_Turnover
## 1  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121 -5.121077e-16
## 2 -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871  9.225312e-01
## 3 -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700  9.225312e-01
## 4  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259  9.225312e-01
## 5 -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461 -4.612656e-01
## 6 -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612 -4.612656e-01
##   Leverage Rev_Growth Net_Profit_Margin
## 1 -0.2120979 -0.5277675      0.06168225
## 2  0.0182843 -0.3811391     -1.55366706
## 3 -0.4040831 -0.5721181     -0.68503583
## 4 -0.7496565  0.1474473      0.35122600
## 5 -0.3144900  1.2163867     -0.42597037
## 6 -0.7496565 -1.4971443     -1.99560225
```

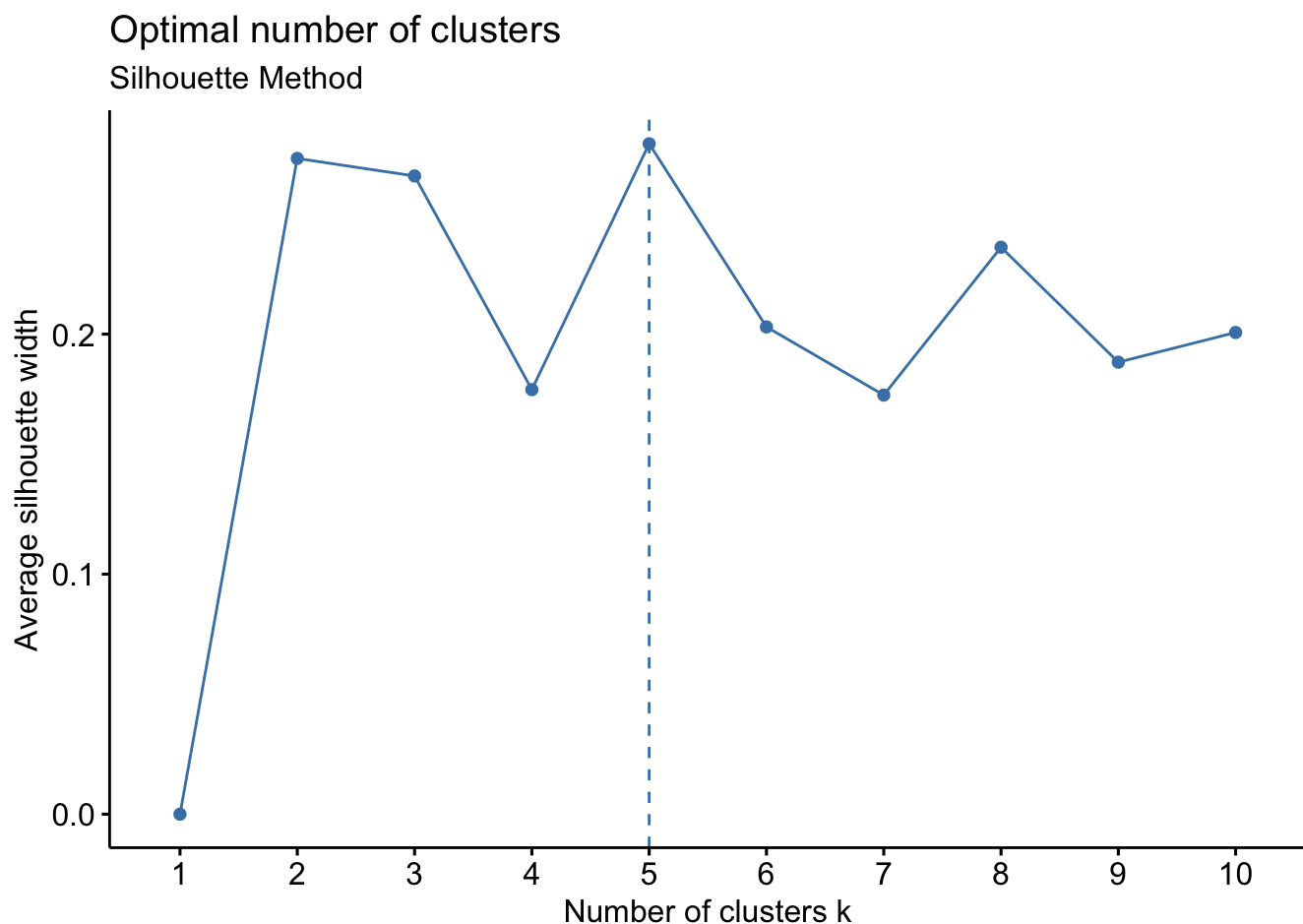
```
#To determine the number of clusters we can use elbow Method
fviz_nbclust(data2, kmeans, method = "wss") + labs(subtitle = "Elbow Method")
```

Optimal number of clusters

Elbow Method



```
#Silhouette method  
fviz_nbclust(data2, kmeans, method = "silhouette")+ labs(subtitle = "Silhouette Method")
```



#From the graph we can see that the 5 is number of clusters.

Since we've taken into account numerical variables, the optimal option in this case of utilizing financial metrics like market capitalization, price, earnings, etc. is kmeans clustering. In this case, the average silhouette method is used to determine the number of clusters, which equals '5' (as shown in the graph). To visualize and comprehend the properties and structure of the clusters that are formed, use `fviz_cluster`. Every dot on the output graph denotes a pharmaceutical company. The labels or colors designate which cluster each firm is assigned to.

```
#Kmeans clustering
set.seed(120)
kmeans <- kmeans(data2, centers = 5, nstart = 25)
#Centroids
kmeans$centers
```



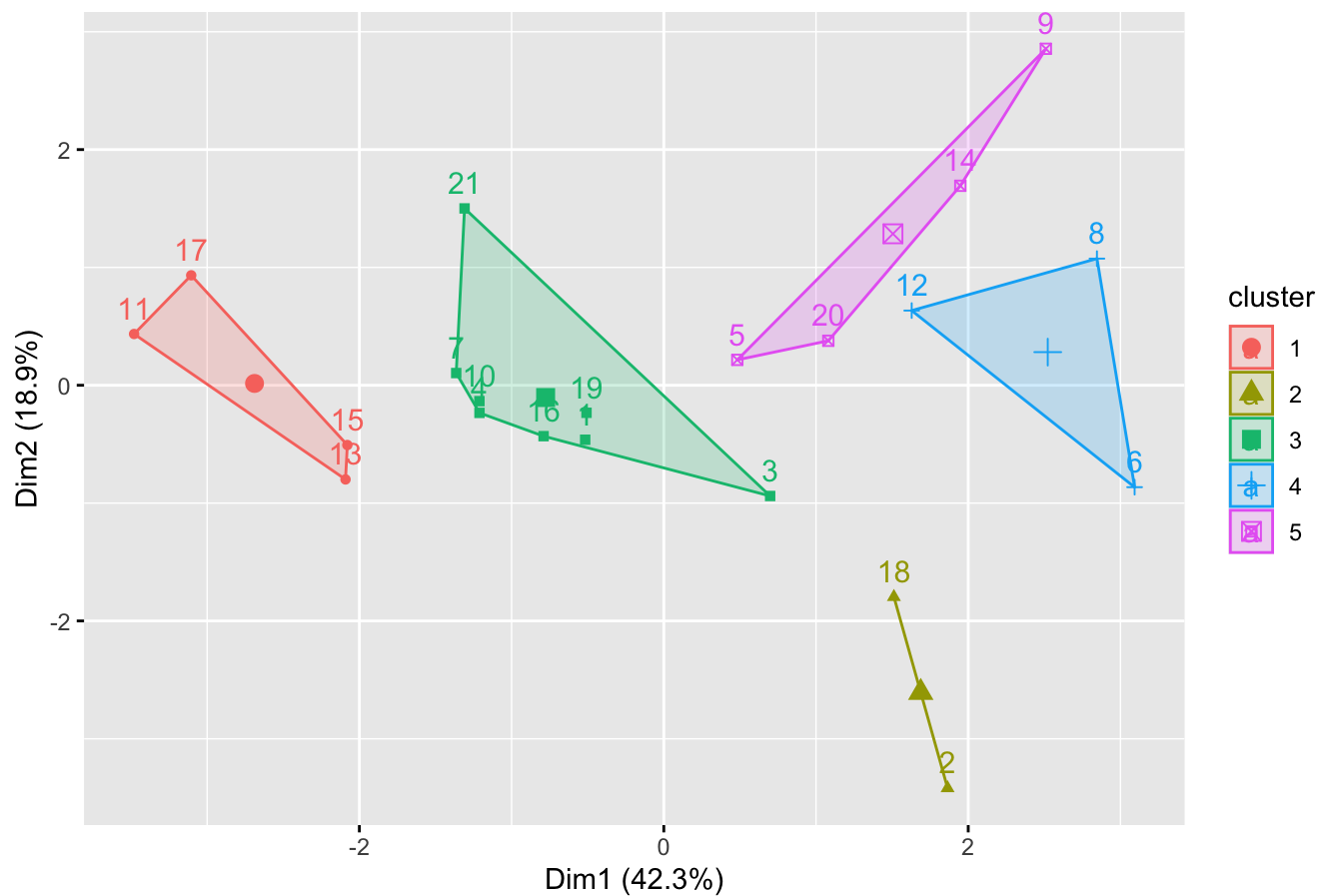
```
##      Market_Cap      Beta      PE_Ratio      ROE      ROA      Asset_Turnover
## 1  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431  1.1531640
## 2 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951  0.2306328
## 3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915  0.1729746
## 4 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656
## 5 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428 -1.2684804
##      Leverage      Rev_Growth      Net_Profit_Margin
## 1 -0.46807818  0.4671788      0.591242521
## 2 -0.14170336 -0.1168459     -1.416514761
## 3 -0.27449312 -0.7041516      0.556954446
## 4  1.36644699 -0.6912914     -1.320000179
## 5  0.06308085  1.5180158     -0.006893899
```

```
#Size of each cluster
kmeans$size
```

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

```
#Visualizing the clusters
fviz_cluster(kmeans,data = data2)
```

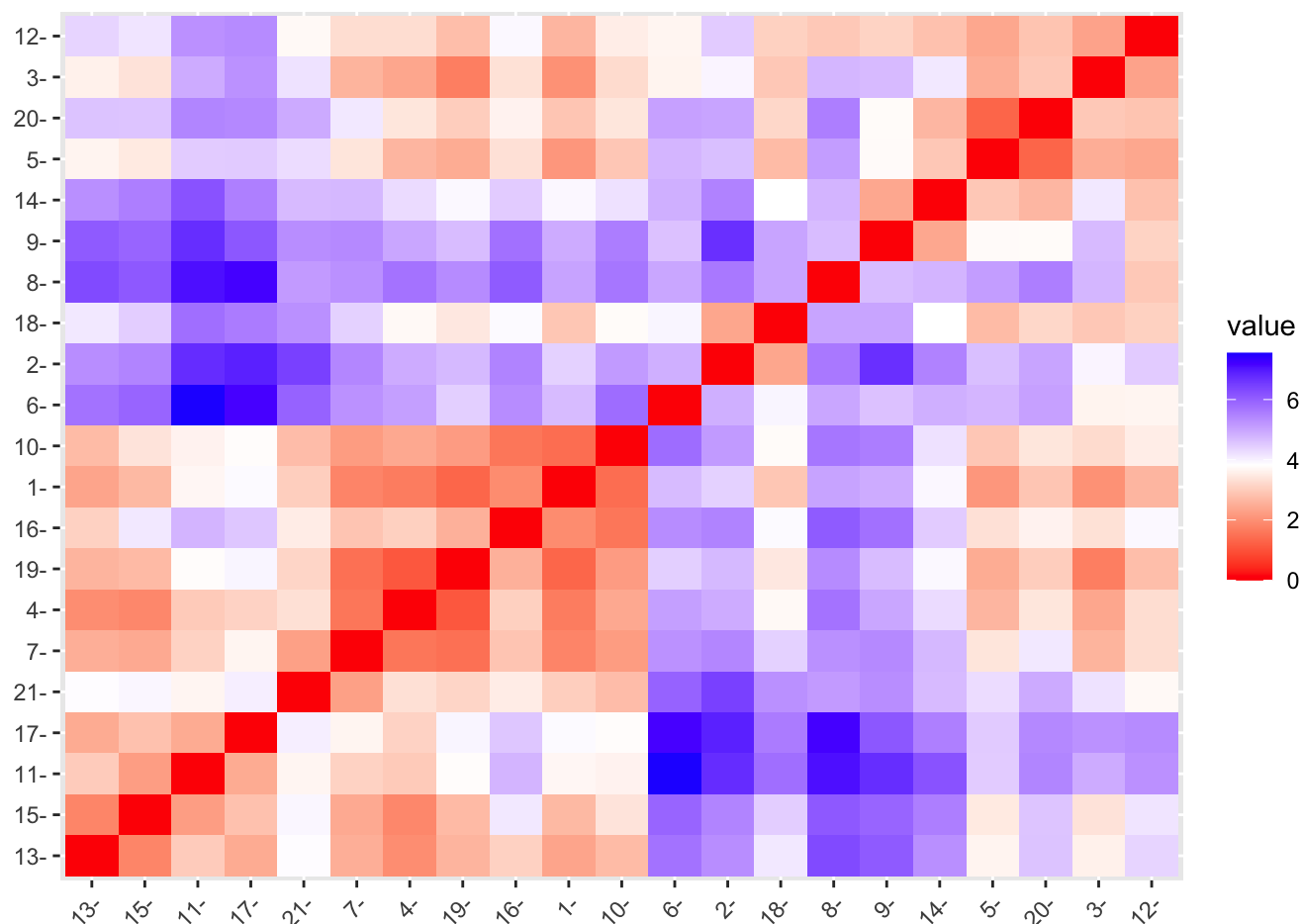
Cluster plot



```
kmeans
```

```
## K-means clustering with 5 clusters of sizes 4, 2, 8, 3, 4
##
## Cluster means:
##      Market_Cap      Beta      PE_Ratio      ROE      ROA Asset_Turnover
## 1  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431  1.1531640
## 2 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951  0.2306328
## 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 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428 -1.2684804
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.46807818  0.4671788      0.591242521
## 2 -0.14170336 -0.1168459      -1.416514761
## 3 -0.27449312 -0.7041516      0.556954446
## 4  1.36644699 -0.6912914      -1.320000179
## 5  0.06308085  1.5180158      -0.006893899
##
## Clustering vector:
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
##  3  2  3  3  5  4  3  4  5  3  1  4  1  5  1  3  1  2  3  5  3
##
## Within cluster sum of squares by cluster:
## [1]  9.284424  2.803505 21.879320 15.595925 12.791257
## (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
distance<- dist(data2, method = "euclidean")
fviz_dist(distance)
```



```
#Fitting the data with clusters
```

```
fitting<-kmeans(data2,5)
```

```
#Finding the mean value of all quantitative variables for each cluster
```

```
aggregate(data2,by=list(fitting$cluster),FUN=mean)
```

```
##      Group.1  Market_Cap      Beta  PE_Ratio      ROE      ROA
## 1          1 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478
## 2          2  0.08926902 -0.4618336 -0.32086149  0.3260892  0.5396003
## 3          3 -0.96686975  1.5162611 -0.57398880 -0.8382671 -0.9892673
## 4          4  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431
## 5          5 -0.57238455 -0.6220844  0.86927480 -0.7381675 -0.7242993
##      Asset_Turnover  Leverage Rev_Growth Net_Profit_Margin
## 1 -4.612656e-01  1.3664470 -0.6912914      -1.3200002
## 2  6.589509e-02 -0.2559803 -0.7230135      0.7343816
## 3 -1.845062e+00  0.5302448  1.7123890      0.2445520
## 4  1.153164e+00 -0.4680782  0.4671788      0.5912425
## 5 -2.442491e-16 -0.2991312  0.3682951      -0.8069490
```

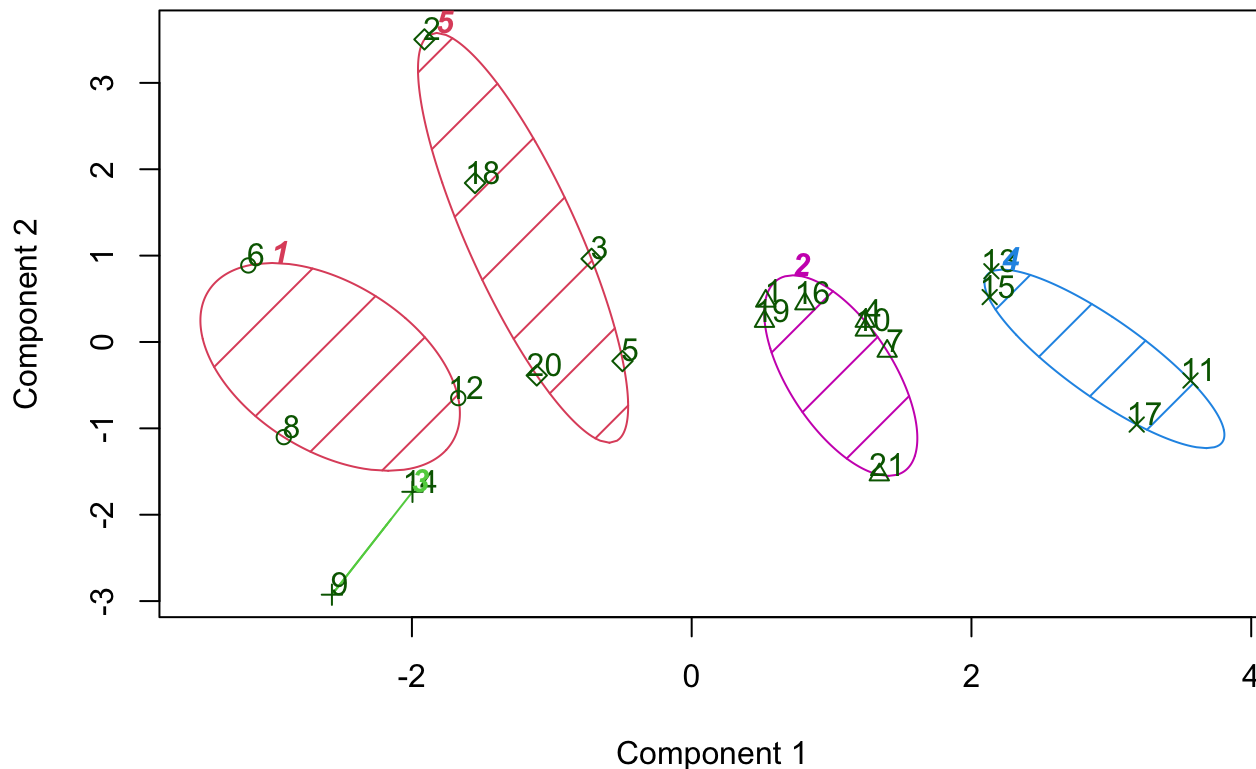
```
data3<-data.frame(data2,fitting$cluster)
```

```
#To view the clusters plot
```

```
library(cluster)
```

```
clusplot(data2,fitting$cluster, color = TRUE, shade = TRUE,
          labels = 2,
          lines = 0)
```

CLUSPLOT(data2)



These two components explain 61.23 % of the point variability.

2. 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)

```
aggregate(data2, by = list(fitting$cluster), FUN = mean)
```

##	Group.1	Market_Cap	Beta	PE_Ratio	ROE	ROA
## 1	1	-0.87051511	1.3409869	-0.05284434	-0.6184015	-1.1928478
## 2	2	0.08926902	-0.4618336	-0.32086149	0.3260892	0.5396003
## 3	3	-0.96686975	1.5162611	-0.57398880	-0.8382671	-0.9892673
## 4	4	1.69558112	-0.1780563	-0.19845823	1.2349879	1.3503431
## 5	5	-0.57238455	-0.6220844	0.86927480	-0.7381675	-0.7242993

##	Asset_Turnover	Leverage	Rev_Growth	Net_Profit_Margin
## 1	-4.612656e-01	1.3664470	-0.6912914	-1.3200002
## 2	6.589509e-02	-0.2559803	-0.7230135	0.7343816
## 3	-1.845062e+00	0.5302448	1.7123890	0.2445520
## 4	1.153164e+00	-0.4680782	0.4671788	0.5912425
## 5	-2.442491e-16	-0.2991312	0.3682951	-0.8069490

```
data4 <- data.frame(data2, kmeans$cluster)
data4
```

##	Market_Cap	Beta	PE_Ratio	ROE	ROA	Asset_Turnover
## 1	0.1840960	-0.80125356	-0.04671323	0.04009035	0.2416121	-5.121077e-16
## 2	-0.8544181	-0.45070513	3.49706911	-0.85483986	-0.9422871	9.225312e-01
## 3	-0.8762600	-0.25595600	-0.29195768	-0.72225761	-0.5100700	9.225312e-01
## 4	0.1702742	-0.02225704	-0.24290879	0.10638147	0.9181259	9.225312e-01
## 5	-0.1790256	-0.80125356	-0.32874435	-0.26484883	-0.5664461	-4.612656e-01
## 6	-0.6953818	2.27578267	0.14948233	-1.45146000	-1.7127612	-4.612656e-01
## 7	-0.1078688	-0.10015669	-0.70887325	0.59693581	0.8617498	9.225312e-01
## 8	-0.9767669	1.26308721	0.03299122	-0.11237924	-1.1677918	-4.612656e-01
## 9	-0.9704532	2.15893320	-1.34037772	-0.70899938	-1.0174553	-1.845062e+00
## 10	0.2762415	-1.34655112	0.14948233	0.34502953	0.5610770	-4.612656e-01
## 11	1.0999201	-0.68440408	-0.45749769	2.45971647	1.8389364	1.383797e+00
## 12	-0.9393967	0.48409069	-0.34100657	-0.29136529	-0.6979905	-4.612656e-01
## 13	1.9841758	-0.25595600	0.18013789	0.18593083	1.0872544	9.225312e-01
## 14	-0.9632863	0.87358895	0.19240011	-0.96753478	-0.9610792	-1.845062e+00
## 15	1.2782387	-0.25595600	-0.40231769	0.98142435	0.8429577	1.845062e+00
## 16	0.6654710	-1.30760129	-0.23677768	-0.52338423	0.1288598	-9.225312e-01
## 17	2.4199899	0.48409069	-0.11415545	1.31287998	1.6322239	4.612656e-01
## 18	-0.0240846	-0.48965495	1.90298017	-0.81506519	-0.9047030	-4.612656e-01
## 19	-0.4018812	-0.06120687	-0.40231769	-0.21181593	0.5234929	4.612656e-01
## 20	-0.9281345	-1.11285216	-0.43297324	-1.03382590	-0.6979905	-9.225312e-01
## 21	-0.1614497	0.40619104	-0.75792214	1.92938746	0.5422849	-4.612656e-01
##	Leverage	Rev_Growth	Net_Profit_Margin	kmeans.cluster		
## 1	-0.21209793	-0.52776752	0.06168225	3		
## 2	0.01828430	-0.38113909	-1.55366706	2		
## 3	-0.40408312	-0.57211809	-0.68503583	3		
## 4	-0.74965647	0.14744734	0.35122600	3		
## 5	-0.31449003	1.21638667	-0.42597037	5		
## 6	-0.74965647	-1.49714434	-1.99560225	4		
## 7	-0.02011273	-0.96584257	0.74744375	3		
## 8	3.74279705	-0.63276071	-1.24888417	4		
## 9	0.61983791	1.88617085	-0.36501379	5		
## 10	-0.07130879	-0.64814764	1.17413980	3		
## 11	-0.31449003	0.76926048	0.82363947	1		
## 12	1.10620040	0.05603085	-0.71551412	4		
## 13	-0.62166634	-0.36213170	0.33598685	1		
## 14	0.44065173	1.53860717	0.85411776	5		
## 15	-0.39128411	0.36014907	-0.24310064	1		
## 16	-0.67286239	-1.45369888	1.02174835	3		
## 17	-0.54487226	1.10143723	1.44844440	1		
## 18	-0.30169102	0.14744734	-1.27936246	2		
## 19	-0.74965647	-0.43544591	0.29026942	3		
## 20	-0.49367621	1.43089863	-0.09070919	5		
## 21	0.68383297	-1.17763919	1.49416183	3		

Cluster:1 with Firm no.: 6, 8, 12 has high Beta,Leverage and lowest Market_Cap,ROE,ROA,Leverage,Rev_Growth,Net_Profit_Margin

Cluster:2 with Firm no.: 1,9,16,4,10,7,21 has high Net_Profit_Margin and low Beta.

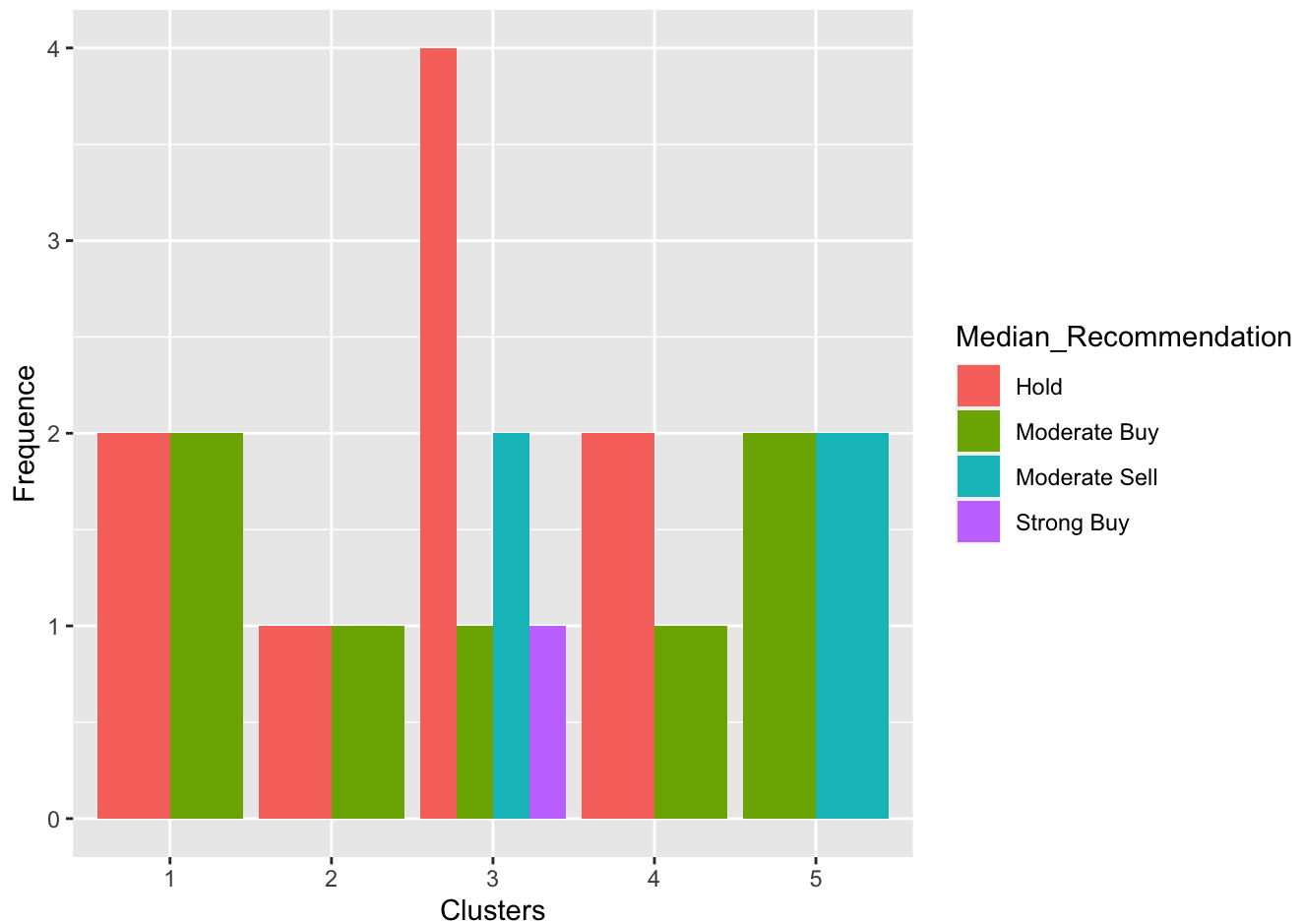
Cluster:3 with Firm no.: 9,14 has high Rev_Growth and low PE_Ratio, Asset_Turnover.

Cluster:4 with Firm no.: 13,15,11,17 has high Market_Cap, ROE, ROA,Asset_Turnover

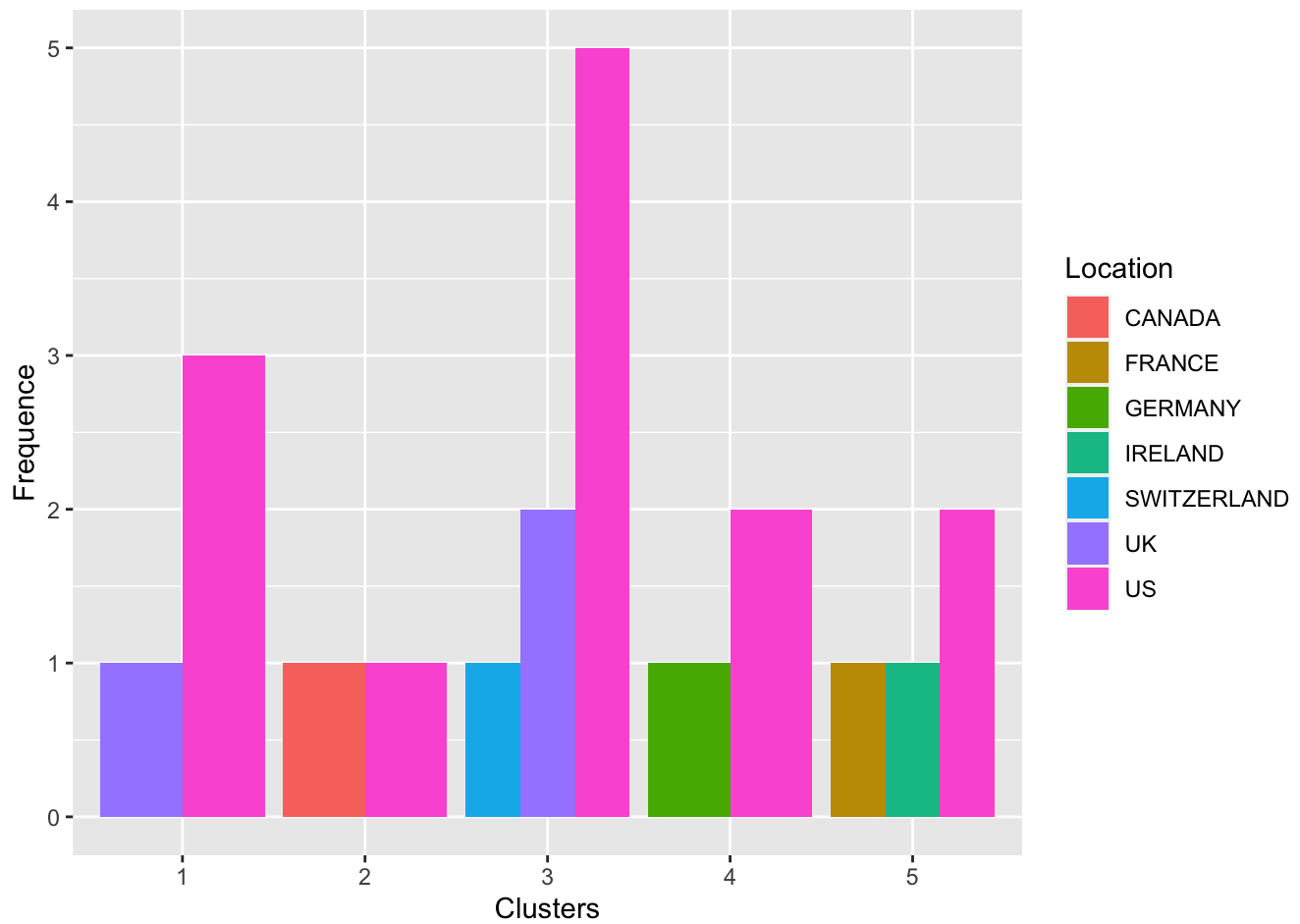
Cluster:5 with Firm no.: 2,18,3,20,5 has high PE_Ratio.

Part-2: Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

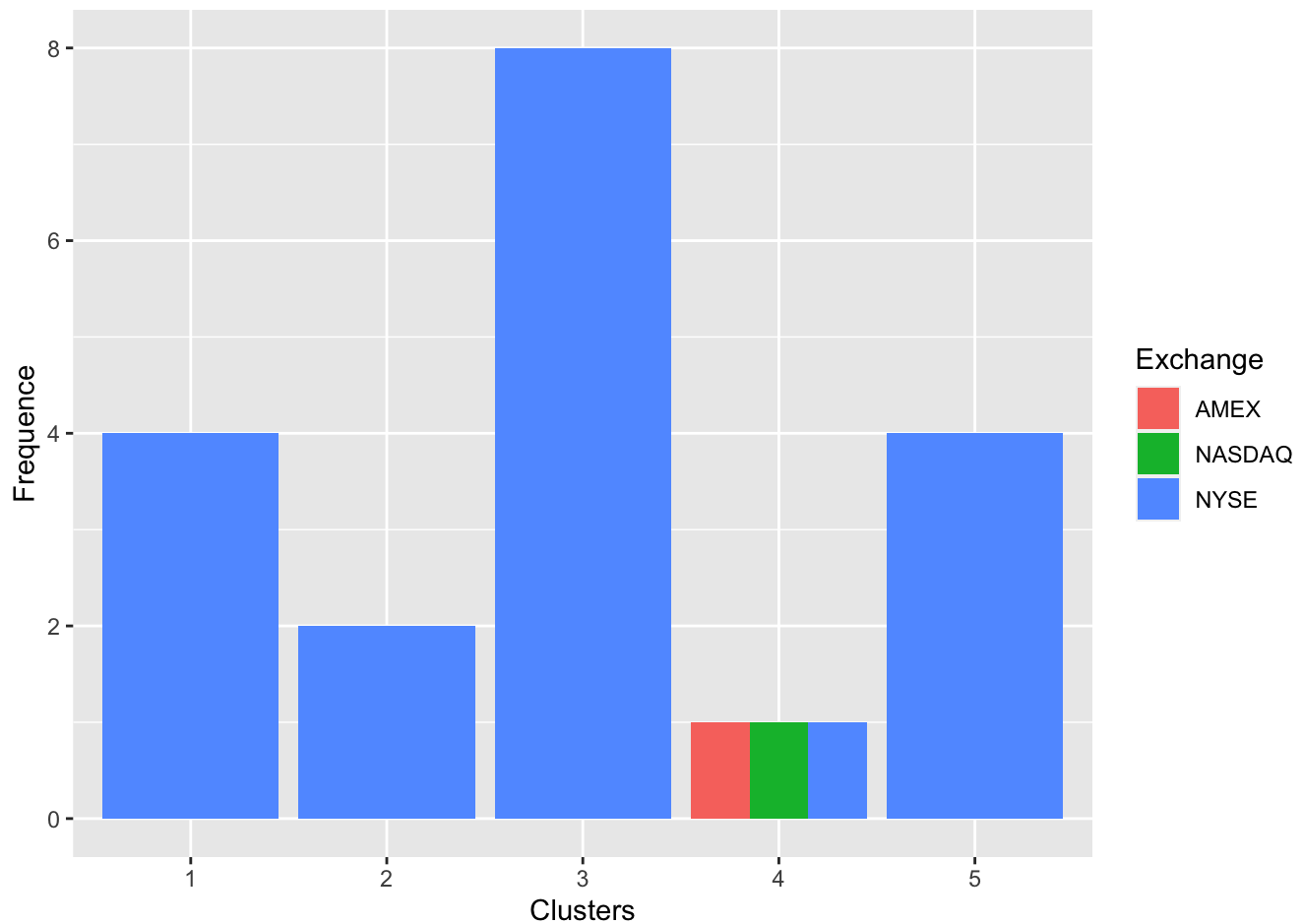
```
data5 <- data[12:14] %>% mutate(Clusters=kmeans$cluster)
ggplot(data5, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position='dodge')+labs(x = 'Clusters',y = 'Frequency')
```



```
ggplot(data5, mapping = aes(factor(Clusters),fill = Location))+
  geom_bar(position = 'dodge')+labs(x = 'Clusters',y = 'Frequency')
```



```
ggplot(data5, mapping = aes(factor(Clusters), fill = Exchange)) + geom_bar(position = 'dodge') +  
  labs(x = 'Clusters', y = 'Frequency')
```



Interpretation based on above graphs:

Group-1: The first cluster comprises two locations: the UK and the US. The firms in this cluster have only the NYSE exchange, which has the largest participants. The firms in this cluster have hold and moderate buy median recommendations.

Group-2: Only the NYSE exchange, which has comparatively more participants, is used by the firms in the second cluster. The firms in this cluster have median buy and hold recommendations, and the cluster is located in both Canada and the US.

Group-3: The third cluster's firms are located in Switzerland, the UK, and the high-frequency US. Its only exchange, the NYSE, has the most participants out of all the clusters. The firms' median recommendations are the highest for hold, moderate buy, relatively high for moderate sell, and strong buy.

Group-4: The fourth cluster, which consists of two locations, comprises firms with NYSE, AMEX, and NASDAQ exchanges. The firms in this cluster are recommended to hold or buy moderately. German additionally US.

Group-5: The fifth cluster comprises three locations: France, Ireland, and the US. The firms in this cluster have median buy and sell recommendations, and the only exchange they participate in is the NYSE, which has the largest participants.

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

Based on the analysis of numerical variables each cluster can be named as following:

Cluster 1: High Risk, Low Performance

Cluster 2: Profitable and Stable

Cluster 3: Growth Opportunities

Cluster 4: Large and Efficient

Cluster 5: Premium Valuation