K-Means for Clustering

Dutt Thakkar

2023-03-19

#importing data

```
library(readr)
Pharm = read.csv("/Users/duttthakkar/Desktop/Pharm.csv")
df= Pharm
```

#viewing the summary of the dataset

```
summary(df)
```

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

#attaching required libraries

```
library(tinytex)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.0     ✔ purrr     1.0.1
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.2     ✔ tibble    3.1.8
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conf
licts to become errors
```

```
library(ISLR)
library(flexclust)
```

```
## Loading required package: grid
## Loading required package: lattice
## Loading required package: modeltools
## Loading required package: stats4
```

```
library(FactoMineR)
library(ggcorrplot)
library(ggplot2)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve
3WBa
```

Question A: 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.

#subsetting the data

```
names(df)
```

```
##  [1] "Symbol"                "Name"                  "Market_Cap"
##  [4] "Beta"                  "PE_Ratio"              "ROE"
##  [7] "ROA"                   "Asset_Turnover"        "Leverage"
## [10] "Rev_Growth"            "Net_Profit_Margin"     "Median_Recommendation"
## [13] "Location"              "Exchange"
```

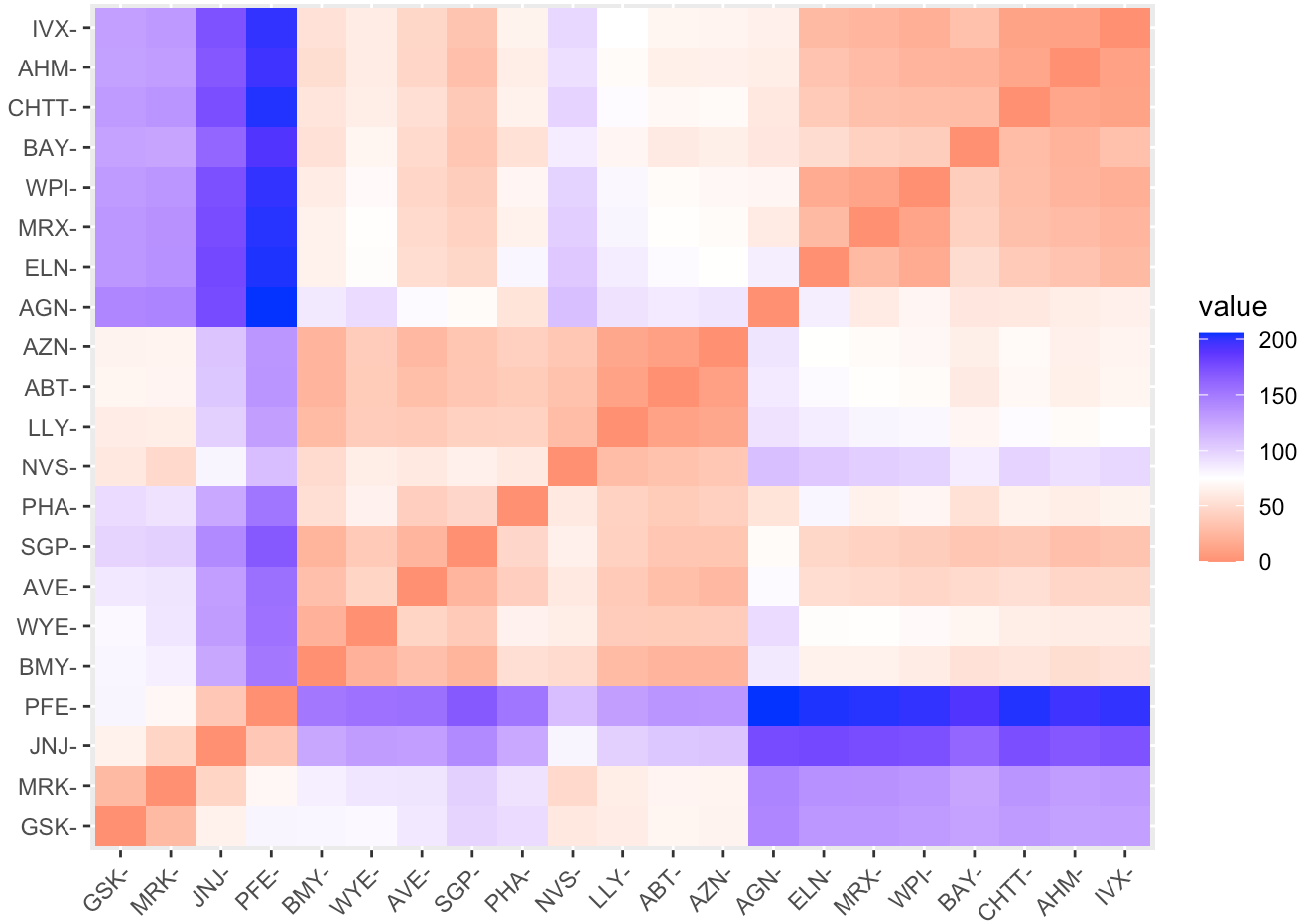
```
dataset=df[,c(1,3:11)]
row.names(dataset)=dataset[,1]
dataset=dataset[,-1]
head(dataset)
```

	Market_Cap <dbl>	B... <dbl>	PE_Ratio <dbl>	ROE <dbl>	R... <dbl>	Asset_Turnover <dbl>	Leverage <dbl>	Rev_Growth <dbl>
ABT	68.44	0.32	24.7	26.4	11.8	0.7	0.42	7.54
AGN	7.58	0.41	82.5	12.9	5.5	0.9	0.60	9.16
AHM	6.30	0.46	20.7	14.9	7.8	0.9	0.27	7.05
AZN	67.63	0.52	21.5	27.4	15.4	0.9	0.00	15.00
AVE	47.16	0.32	20.1	21.8	7.5	0.6	0.34	26.81
BAY	16.90	1.11	27.9	3.9	1.4	0.6	0.00	-3.17

6 rows | 1-9 of 10 columns

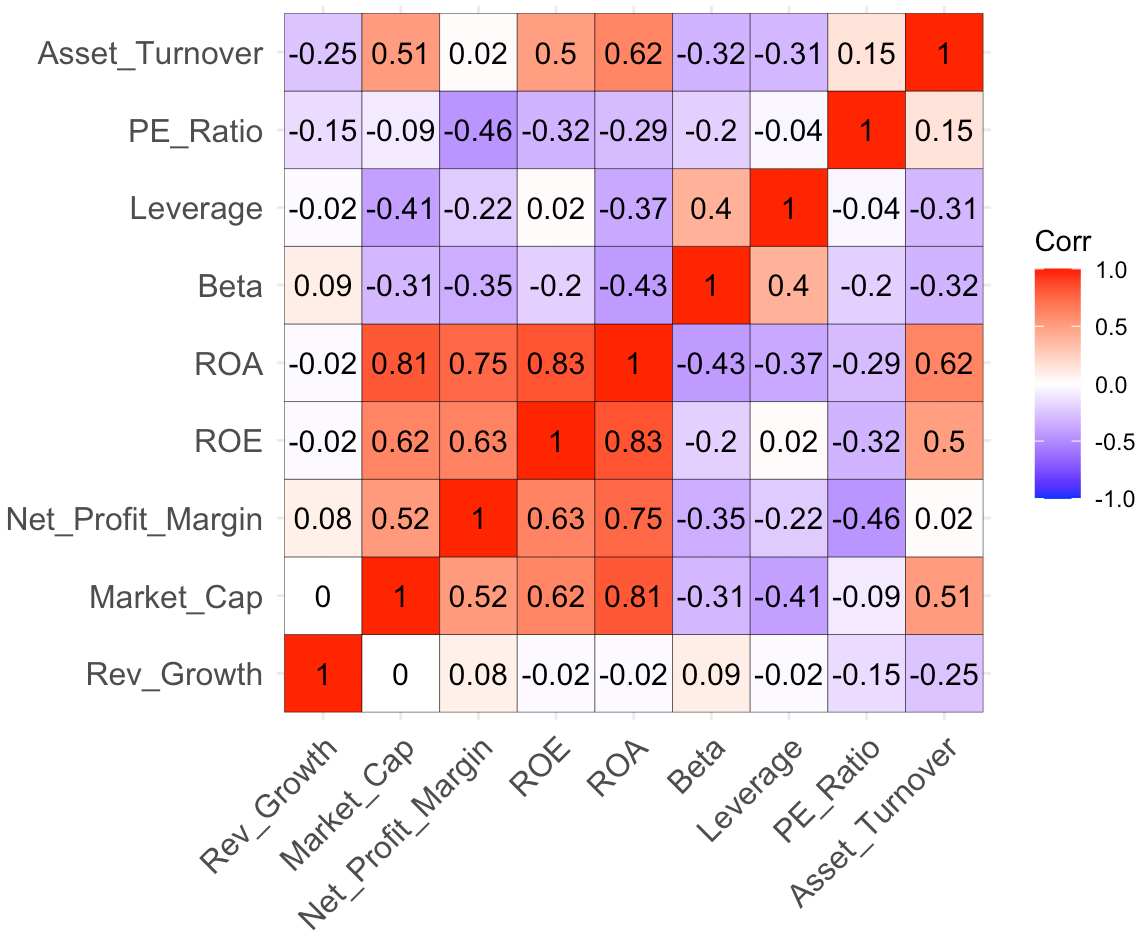
#normalizing the dataset

```
dataset2 = scale(dataset)
distance=get_dist(dataset)
fviz_dist(distance)
```



#using euclidean distance formula which is given by: d=√((x_2-x)1)^2+(y_2-y_1)^2)

```
Corr=cor(dataset2)
ggcorrplot(Corr,outline.color = "black",lab = TRUE,hc.order = TRUE,type = "full")
```

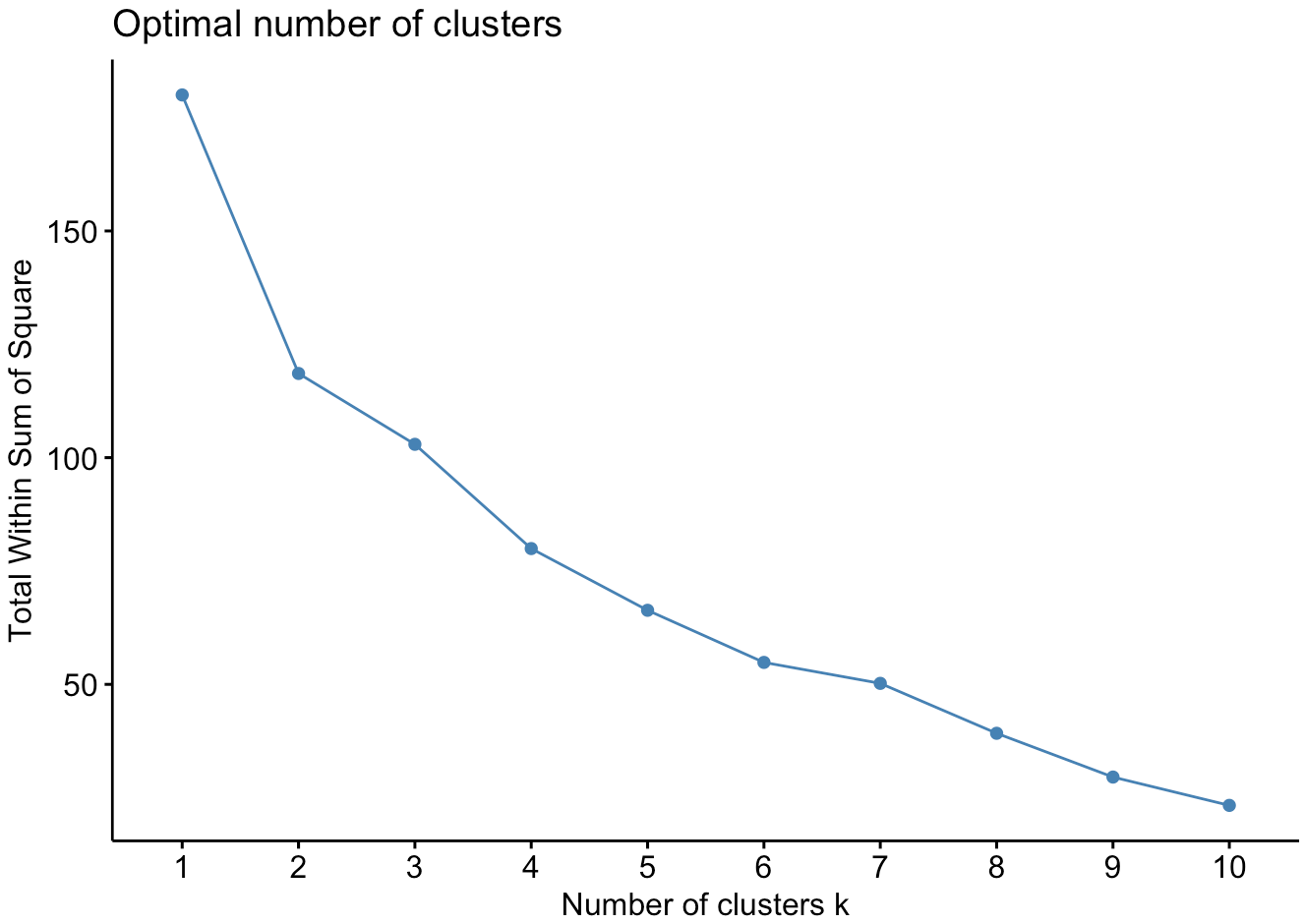


#installing factoextra and cluster to plot elbow chart and silhouette chart

```
library(cluster)
library(factoextra)
```

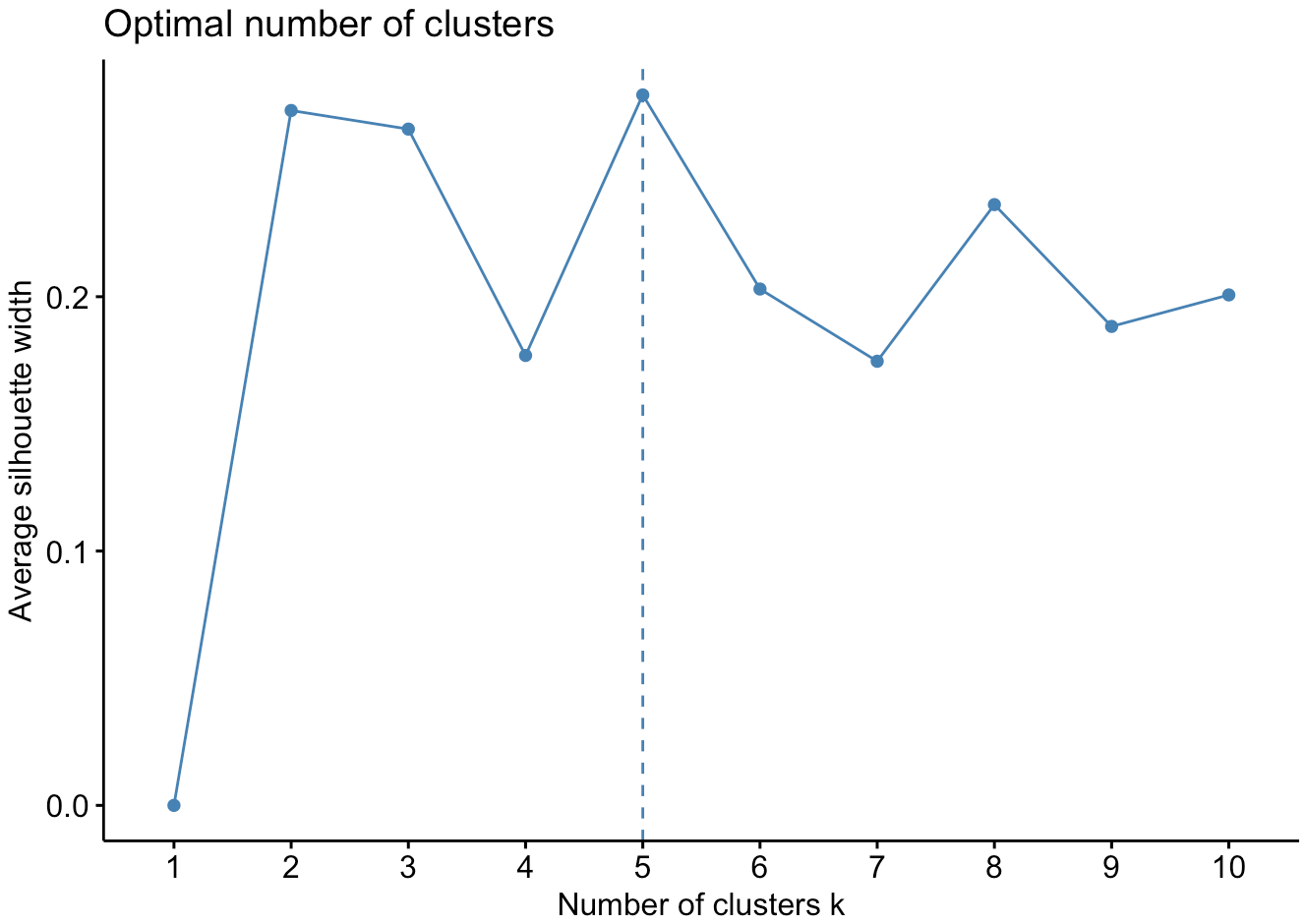
#Finding the number of clusters using elbow chart

```
set.seed(100)
fviz_nbclust(dataset2, kmeans, method = "wss")
```



#after looking at the elbow chart, it shows that the optimal number of clusters is 2 or 7 #Finding the number of clusters using silhouette method

```
fviz_nbclust(dataset2,kmeans,method = "silhouette")
```



#after looking at the elbow chart, it shows that the optimal number of clusters is 5. Therefore, we will try and find an optimal value between 2 and 7 per the results gathered from elbow and silhouette method respectively

```
k2<-kmeans(dataset2,centers =2,nstart=25)
k3<-kmeans(dataset2,centers =3,nstart=25)
k4<-kmeans(dataset2,centers =4,nstart=25)
k5<-kmeans(dataset2,centers =5,nstart=25)
k6<-kmeans(dataset2,centers =6,nstart=25)
k7<-kmeans(dataset2,centers =7,nstart=25)
p1<-fviz_cluster(k2,geom = "point", data=dataset2)+ggtitle("k=2")
p2<-fviz_cluster(k3,geom = "point", data=dataset2)+ggtitle("k=3")
p3<-fviz_cluster(k4,geom = "point", data=dataset2)+ggtitle("k=4")
p4<-fviz_cluster(k5,geom = "point", data=dataset2)+ggtitle("k=5")
p5<-fviz_cluster(k6,geom = "point", data=dataset2)+ggtitle("k=6")
p6<-fviz_cluster(k7,geom = "point", data=dataset2)+ggtitle("k=7")
```

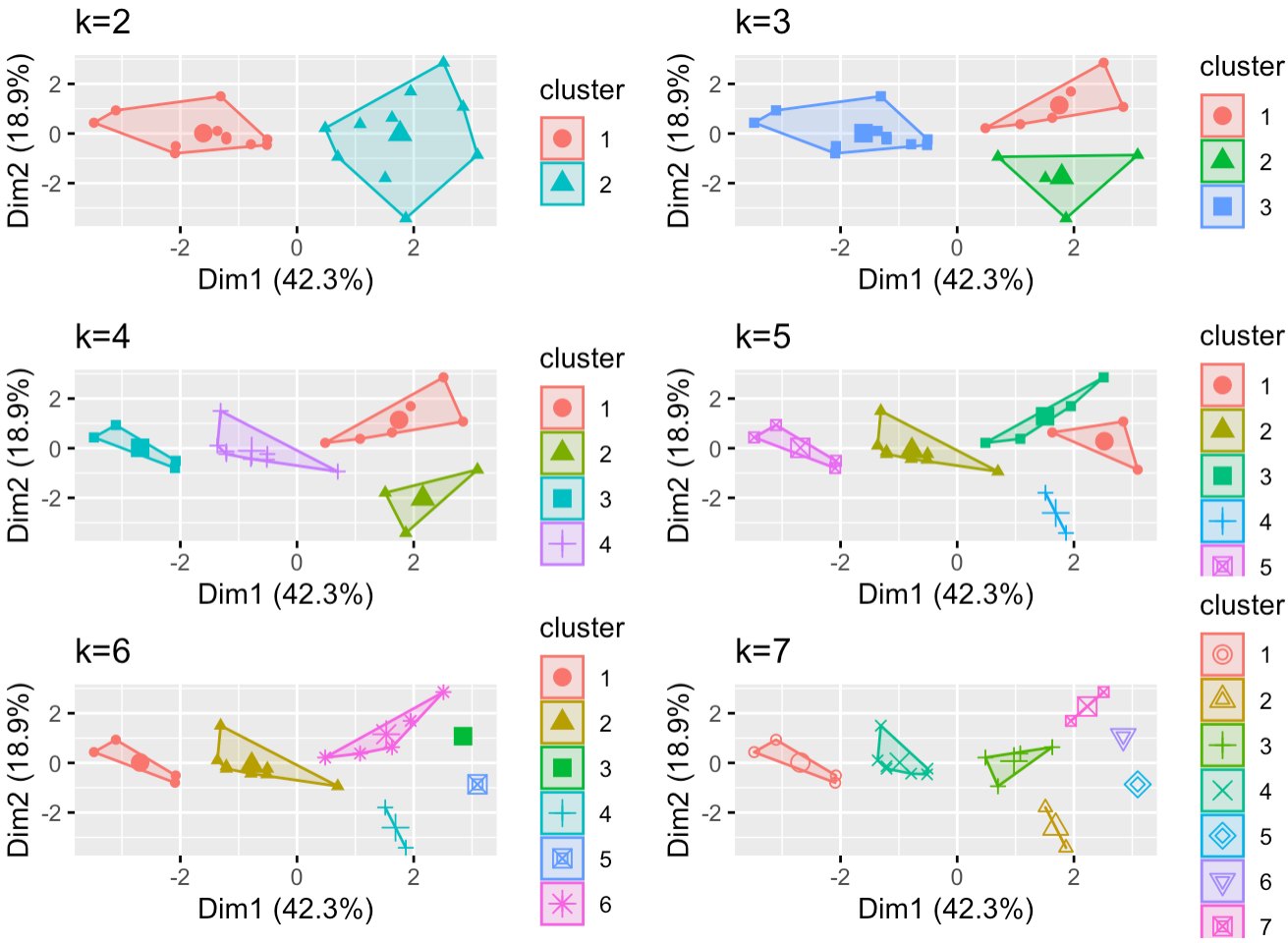
#attaching library gridExtra to combine the clusters

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
grid.arrange(p1,p2,p3,p4,p5,p6)
```



##Question B: Interpret the clusters with respect to the numerical variables used in forming the clusters. # after reviewing the clusters, K = 5 seems appropriate as per the grouping.

#using K=5 for the analysis

```
k5=kmeans(dataset2, centers = 5, nstart = 25)
k5$size
```

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

```
k5$cluster
```

```
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    2    3    2    2    4    1    2    1    4    2    5    1    5    4    5    2
##  PFE  PHA  SGP  WPI  WYE
##    5    3    2    4    2
```

```
k5$centers
```

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

```
k5$withinss
```

```
## [1] 15.595925 21.879320  2.803505 12.791257  9.284424
```

```
fviz_cluster(k5, data = dataset2)
```

#Interpretation of the clusters #The entire data is divided into 5 different clusters: Cluster #5 have the 4

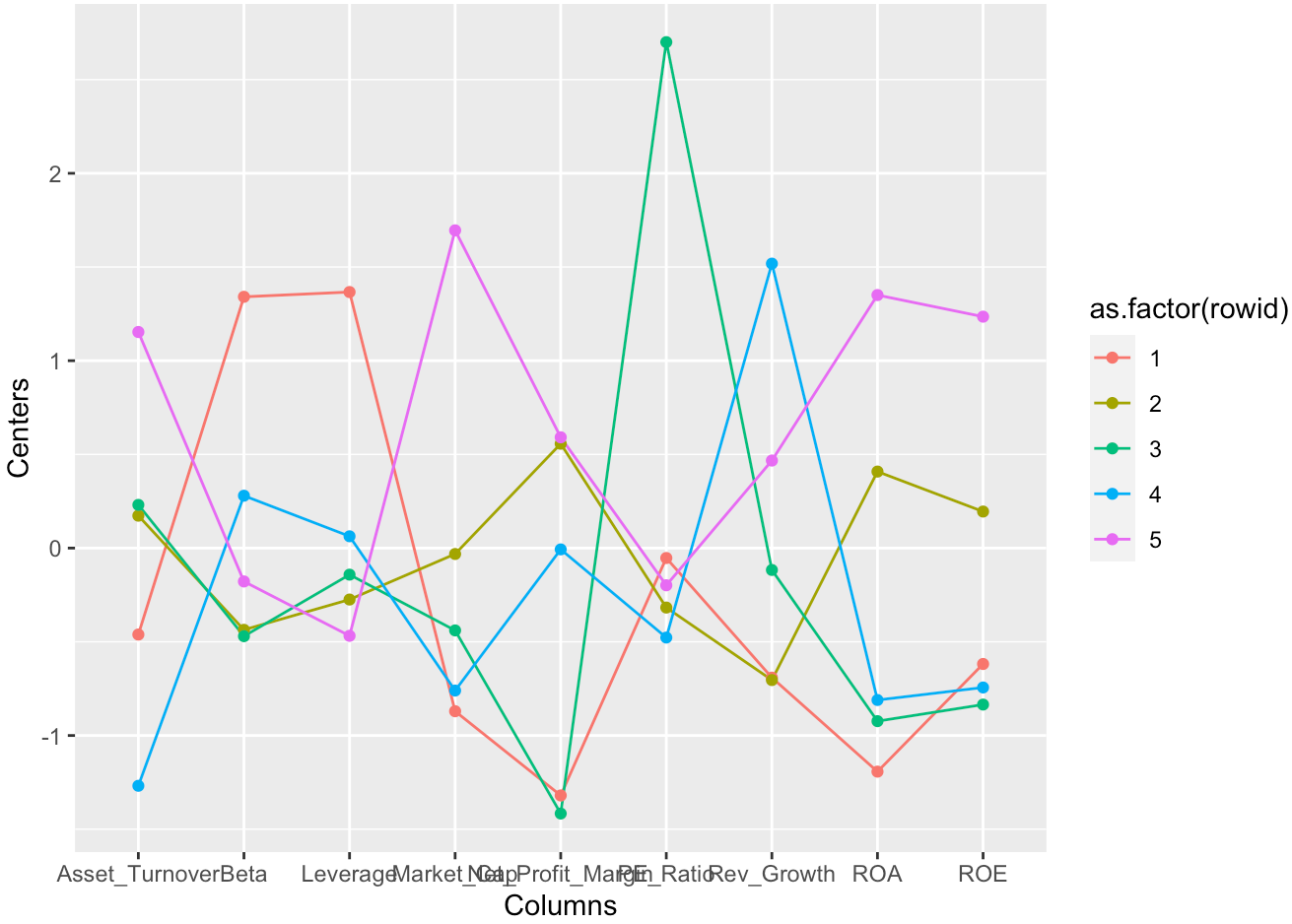
companies and their net profit margin is high as well as their asset turnover making them a credible institutions. On the other hand, cluster #1 has 3 companies and their net profit margin is -1.32 as they are more dependent on levraging (1.36) and less on equities (-0.61)

#plotting graphs of data grouped in clusters

```
Centroid <- data.frame(k5$centers) %>% rowid_to_column() %>% gather('Columns', 'Cente
rs', -1)
print(Centroid)
```

```
##    rowid           Columns      Centers
## 1      1        Market_Cap -0.870515113
## 2      2        Market_Cap -0.031422109
## 3      3        Market_Cap -0.439251341
## 4      4        Market_Cap -0.760224892
## 5      5        Market_Cap  1.695581115
## 6      1              Beta  1.340986857
## 7      2              Beta -0.436098941
## 8      3              Beta -0.470180039
## 9      4              Beta  0.279604106
## 10     5              Beta -0.178056346
## 11     1          PE_Ratio -0.052844340
## 12     2          PE_Ratio -0.317248516
## 13     3          PE_Ratio  2.700024643
## 14     4          PE_Ratio -0.477423799
## 15     5          PE_Ratio -0.198458234
## 16     1               ROE -0.618401510
## 17     2               ROE  0.195045857
## 18     3               ROE -0.834952524
## 19     4               ROE -0.743802224
## 20     5               ROE  1.234987906
## 21     1               ROA -1.192847826
## 22     2               ROA  0.408391543
## 23     3               ROA -0.923495091
## 24     4               ROA -0.810742783
## 25     5               ROA  1.350343113
## 26     1    Asset_Turnover -0.461265604
## 27     2    Asset_Turnover  0.172974602
## 28     3    Asset_Turnover  0.230632802
## 29     4    Asset_Turnover -1.268480411
## 30     5    Asset_Turnover  1.153164010
## 31     1          Leverage  1.366446992
## 32     2          Leverage -0.274493115
## 33     3          Leverage -0.141703357
## 34     4          Leverage  0.063080849
## 35     5          Leverage -0.468078185
## 36     1        Rev_Growth -0.691291399
## 37     2        Rev_Growth -0.704151557
## 38     3        Rev_Growth -0.116845875
## 39     4        Rev_Growth  1.518015830
## 40     5        Rev_Growth  0.467178770
## 41     1 Net_Profit_Margin -1.320000179
## 42     2 Net_Profit_Margin  0.556954446
## 43     3 Net_Profit_Margin -1.416514761
## 44     4 Net_Profit_Margin -0.006893899
## 45     5 Net_Profit_Margin  0.591242521
```

```
ggplot(Centroid, aes(x = Columns, y = Centers, color = as.factor(rowid))) + geom_line
(aes(group = as.factor(rowid))) + geom_point()
```



#Question C: Is there a pattern in the clusters with respect to the numerical variables (10 to 12)?

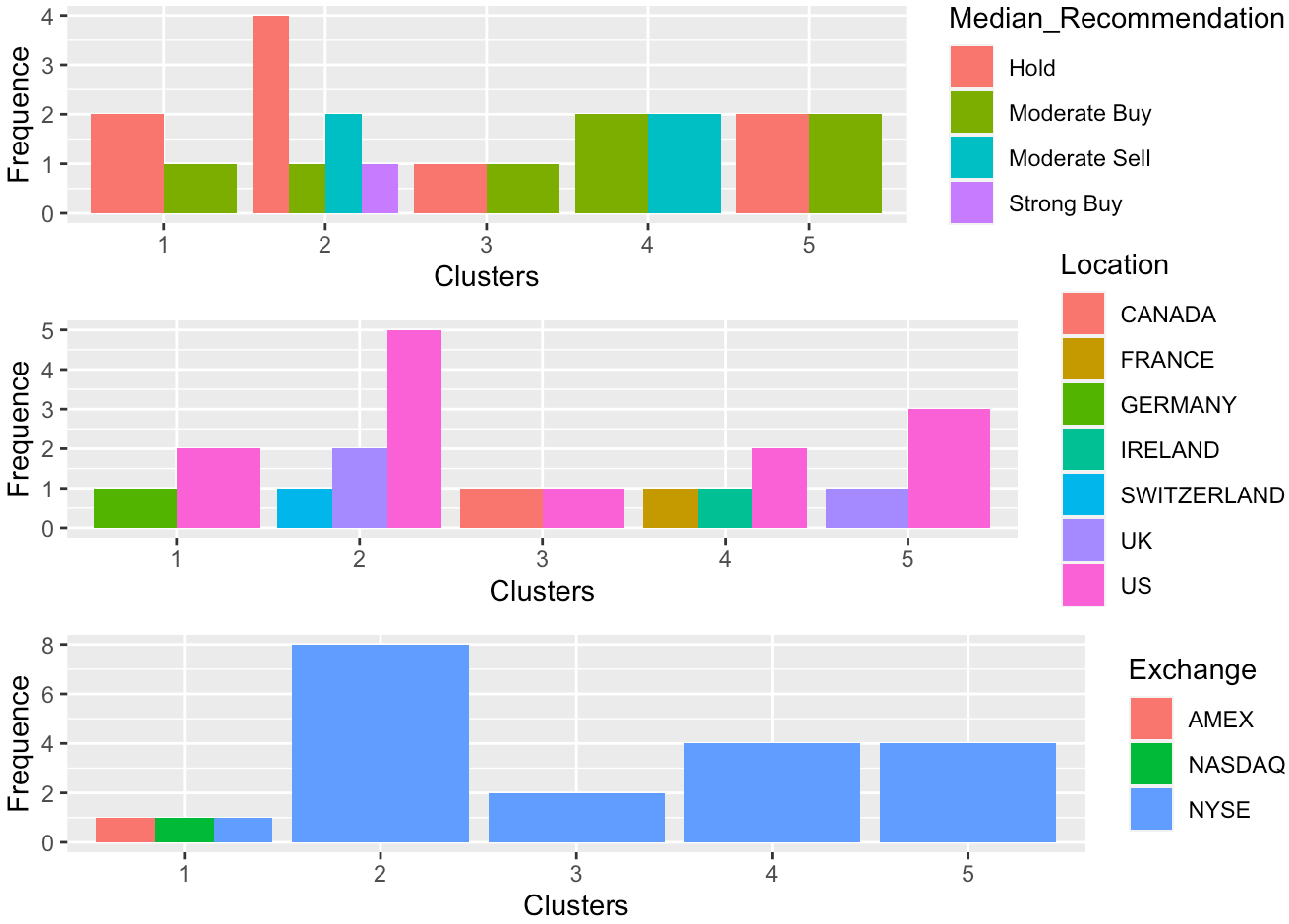
#considering the last three variables; median_recommendation, location, and exchange

```
pattern <- df%>% select(c(12,13,14)) %>% mutate(Cluster = k5$cluster)
print(pattern)
```

```
##    Median_Recommendation    Location Exchange Cluster
## 1           Moderate Buy          US     NYSE       2
## 2           Moderate Buy      CANADA     NYSE       3
## 3             Strong Buy          UK     NYSE       2
## 4          Moderate Sell          UK     NYSE       2
## 5           Moderate Buy      FRANCE     NYSE       4
## 6                   Hold     GERMANY     NYSE       1
## 7          Moderate Sell          US     NYSE       2
## 8           Moderate Buy          US   NASDAQ       1
## 9          Moderate Sell     IRELAND     NYSE       4
## 10                  Hold          US     NYSE       2
## 11                  Hold          UK     NYSE       5
## 12                  Hold          US     AMEX       1
## 13          Moderate Buy          US     NYSE       5
## 14          Moderate Buy          US     NYSE       4
## 15                  Hold          US     NYSE       5
## 16                  Hold SWITZERLAND     NYSE       2
## 17          Moderate Buy          US     NYSE       5
## 18                  Hold          US     NYSE       3
## 19                  Hold          US     NYSE       2
## 20         Moderate Sell          US     NYSE       4
## 21                  Hold          US     NYSE       2
```

#identifying if there are any trends

```
Median_Recommenation <- ggplot(pattern, mapping = aes(factor(Cluster), fill=Median_Re
commendation)) + geom_bar(position = 'dodge') + labs(x='Clusters', y='Frequence')
Location <- ggplot(pattern, mapping = aes(factor(Cluster), fill=Location)) + geom_bar
(position = 'dodge') + labs(x='Clusters', y='Frequence')
Exchange <- ggplot(pattern, mapping = aes(factor(Cluster), fill=Exchange)) + geom_bar
(position = 'dodge') + labs(x='Clusters', y='Frequence')
grid.arrange(Median_Recommenation,Location,Exchange)
```




```
#Cluster1 has majority of the companies from the US listed equally in AMEX, NASDAQ, a
nd NYSE. This segment contains low-risk companies as their holding rate is higher tha
n the buying rates.
#Cluster2 has majority of the companies the US followed by UK and Switzerland. All th
e companies are listed in the NYSE. These companies are moderately low-risk companies
as their holding rate is still higher but also shows adequate selling.
#Cluster3 has companies from Canada and US, listed in NYSE. These companies demonstra
tes some growth potential as equal number of holding and buying rates.
#Cluster 4 has companies from France, Germany and US investing at NYSE.These companie
s shows the most risky activities as they have equal buying and selling rates. On the
contrary, this shows that as they take the risk, they have higher potential of growt
h.
#Cluster 5 has companies US and UK again listed in NYSE. These companies practice the
safest among all the clusters. Their holding and buying rates are equal but slightly
higher that cluster#3. These are the most profitable companies.
```

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

```
#Cluster 1: Low-risk companies (well-ordered)
#Cluster 2: Growing companies
#Cluster 3: high-risk companies
#Cluster 4: risky-companies
#Cluster 5: Stable companies
#These titles has been given after comparing the five different clusters and their K-
center values (Market_Cap, Beta, PE_Ratio, ROE, ROA, Asset_Turnover, Leverage, Rev_Gr
owth, Net_Profit_Margin ).
```