FML_Clustering

2023-11-13

#We need the following packages to use the clustering needed for this assignment.

```
library(flexclust)
```

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

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

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

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

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

```
library(cluster)
```

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

```
library(tidyverse)
```

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

```
## -- 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
```

```r
library(FactoMineR)
```

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

```r
library(ggcorrplot)
```

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

```r
df_pharma <- read.csv("C:/Users/navat/Downloads/Pharmaceuticals.csv")
```

```r
summary(df_pharma)
```

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

```r
head(df_pharma)
```

```
##   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
```

Q1) 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.

Sol: As mentioned, lets only select the variables from 1 to 9.

```
df_pharma_1 <- df_pharma[3:11]
head(df_pharma_1)
```

```
##   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
```

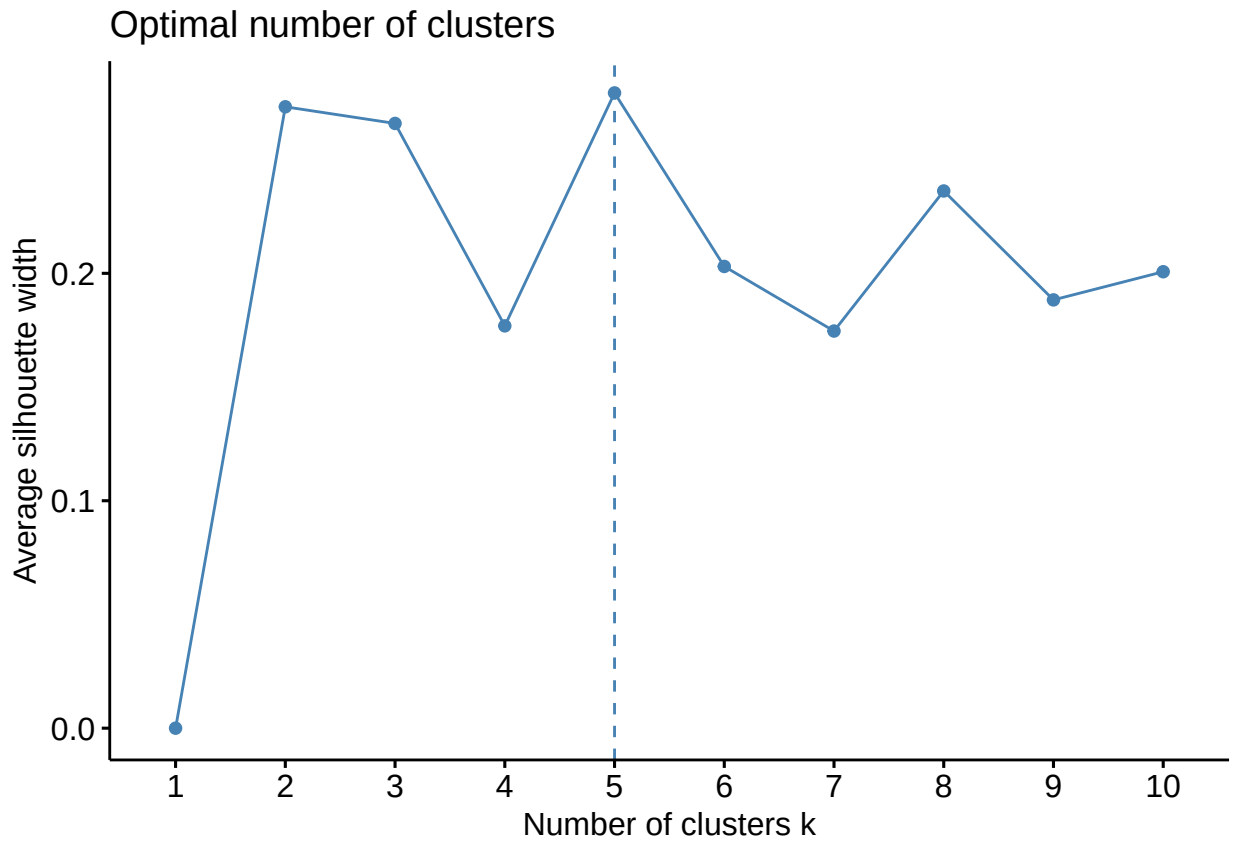
```
summary(df_pharma_1)
```

```
##    Market_Cap          Beta           PE_Ratio          ROE
##  Min.   :  0.41   Min.   :0.1800   Min.   : 3.60   Min.   : 3.9
##  1st Qu.:  6.30   1st Qu.:0.3500   1st Qu.:18.90   1st Qu.:14.9
##  Median : 48.19   Median :0.4600   Median :21.50   Median :22.6
##  Mean   : 57.65   Mean   :0.5257   Mean   :25.46   Mean   :25.8
##  3rd Qu.: 73.84   3rd Qu.:0.6500   3rd Qu.:27.90   3rd Qu.:31.0
##  Max.   :199.47   Max.   :1.1100   Max.   :82.50   Max.   :62.9
##       ROA        Asset_Turnover    Leverage        Rev_Growth
##  Min.   : 1.40   Min.   :0.3    Min.   :0.0000   Min.   :-3.17
##  1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600   1st Qu.: 6.38
##  Median :11.20   Median :0.6    Median :0.3400   Median : 9.37
##  Mean   :10.51   Mean   :0.7    Mean   :0.5857   Mean   :13.37
##  3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000   3rd Qu.:21.87
##  Max.   :20.30   Max.   :1.1    Max.   :3.5100   Max.   :34.21
##  Net_Profit_Margin
##  Min.   : 2.6
##  1st Qu.:11.2
##  Median :16.1
##  Mean   :15.7
```

```
##  3rd Qu.:21.1
##  Max.   :25.5
```

#Let's normalise the data. Using the k-means clustering algorithm, this scales the data which will ensure that all the variables contributes equally and also calculating the distance. For the optimal number of clusters, we have used the fviz_nbclust with the "silhouette" method, which calculates the silhouette scores for different clusters and helps us to identify the number to maximise the seperation between clusters.

```
df_pharma_2 <- scale(df_pharma_1)
row.names(df_pharma_2) <- df_pharma[,1]
dist <- get_dist(df_pharma_2)
corr <- cor(df_pharma_2)
fviz_nbclust(df_pharma_2,kmeans, method = "silhouette")
```



#After the application, we can find that the optimal number of clusters to be 5. So, using k = 5 and number of restarts are 25.

```
set.seed(69)
k_5 <- kmeans(df_pharma_2,centers = 5,nstart = 25)
print(k_5)
```

```
## K-means clustering with 5 clusters of sizes 4, 8, 2, 4, 3
##
## 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 -0.03142211 -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 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.46807818  0.4671788       0.591242521
## 2 -0.27449312 -0.7041516       0.556954446
## 3 -0.14170336 -0.1168459      -1.416514761
## 4  0.06308085  1.5180158      -0.006893899
## 5  1.36644699 -0.6912914      -1.320000179
##
## Clustering vector:
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    2    3    2    2    4    5    2    5    4    2    1    5    1    4    1    2
##  PFE  PHA  SGP  WPI  WYE
##    1    3    2    4    2
##
## Within cluster sum of squares by cluster:
## [1]  9.284424 21.879320  2.803505 12.791257 15.595925
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
k_5$centers
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 2 -0.03142211 -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 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.46807818  0.4671788       0.591242521
## 2 -0.27449312 -0.7041516       0.556954446
## 3 -0.14170336 -0.1168459      -1.416514761
## 4  0.06308085  1.5180158      -0.006893899
## 5  1.36644699 -0.6912914      -1.320000179
```

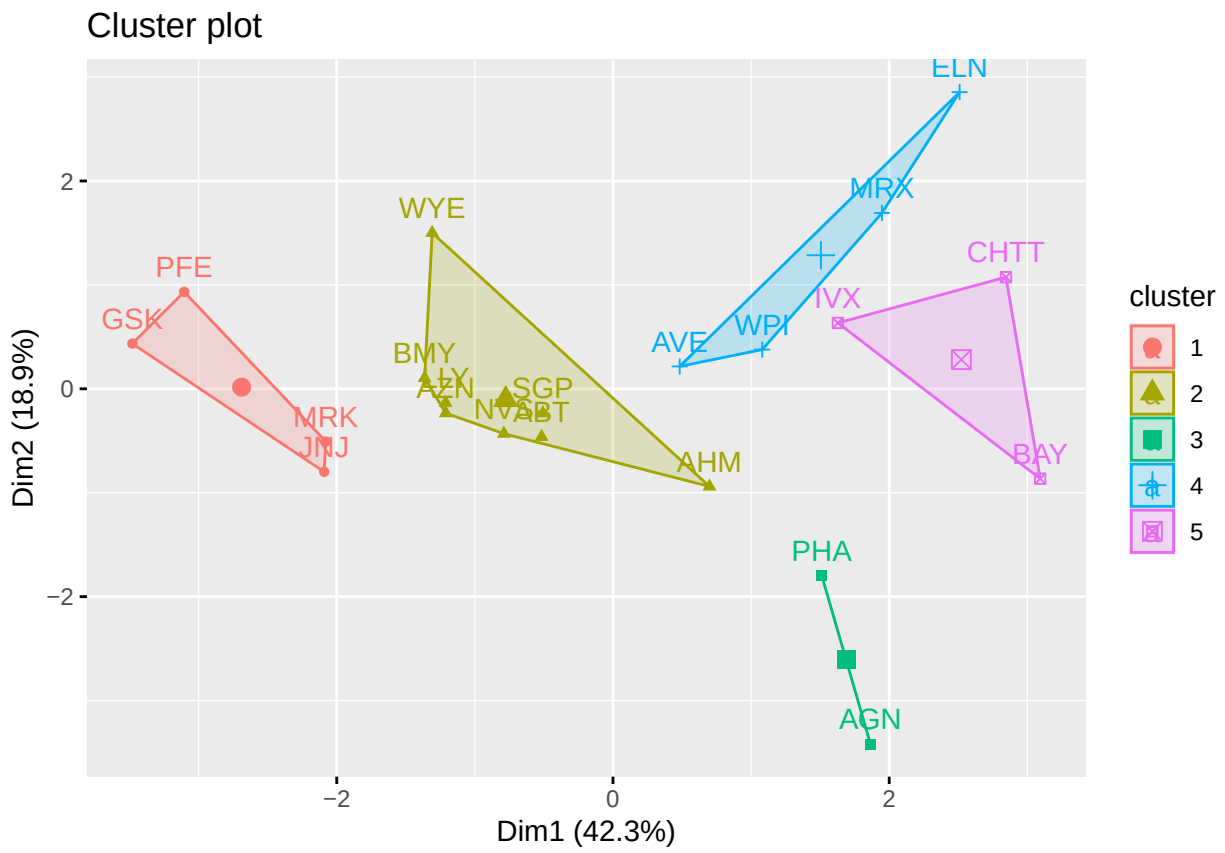
```
k_5$size
```

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

#We see that the within cluster sum of square value is 65.5% when k=5. The "centers" produced from the kmeans function provides us with the mean values of variables within clusters.

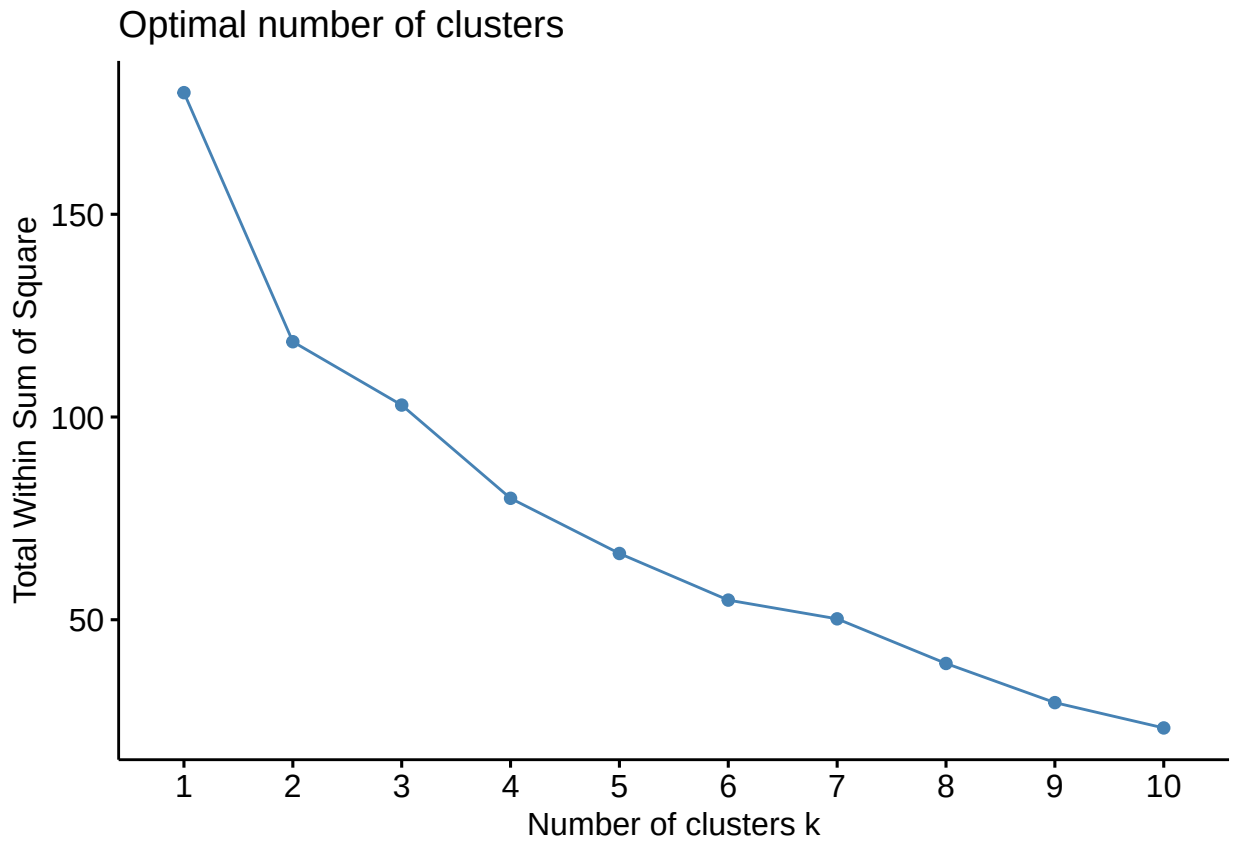
#When k=5, using the fviz_cluster to form the clusters.

```
fviz_cluster(k_5,data = df_pharma_2)
```



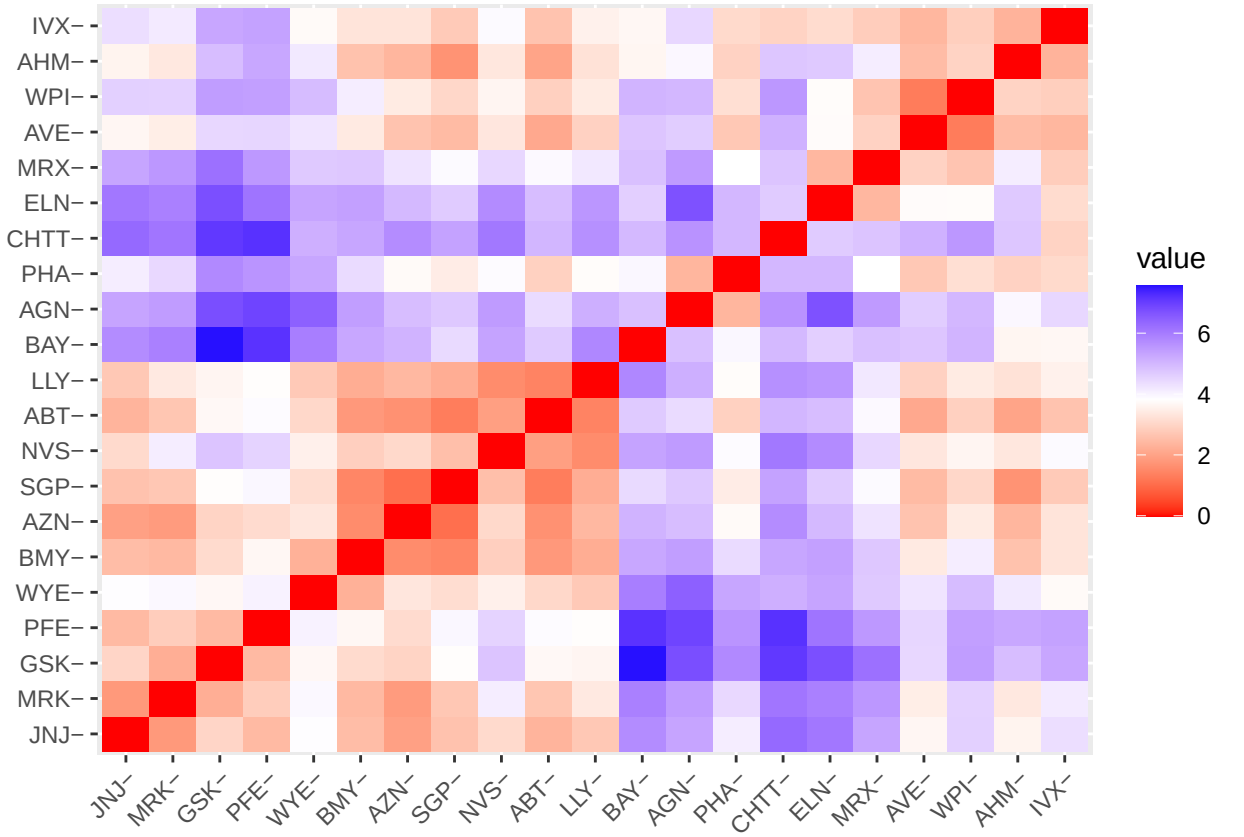
#Using the elbow method

```
fviz_nbclust(df_pharma_2,kmeans,method = "wss")
```



#Using the "euclidean" distance

```
dist <- dist(df_pharma_2,method="euclidean")
fviz_dist(dist)
```



#Using the "Manhattan" distance

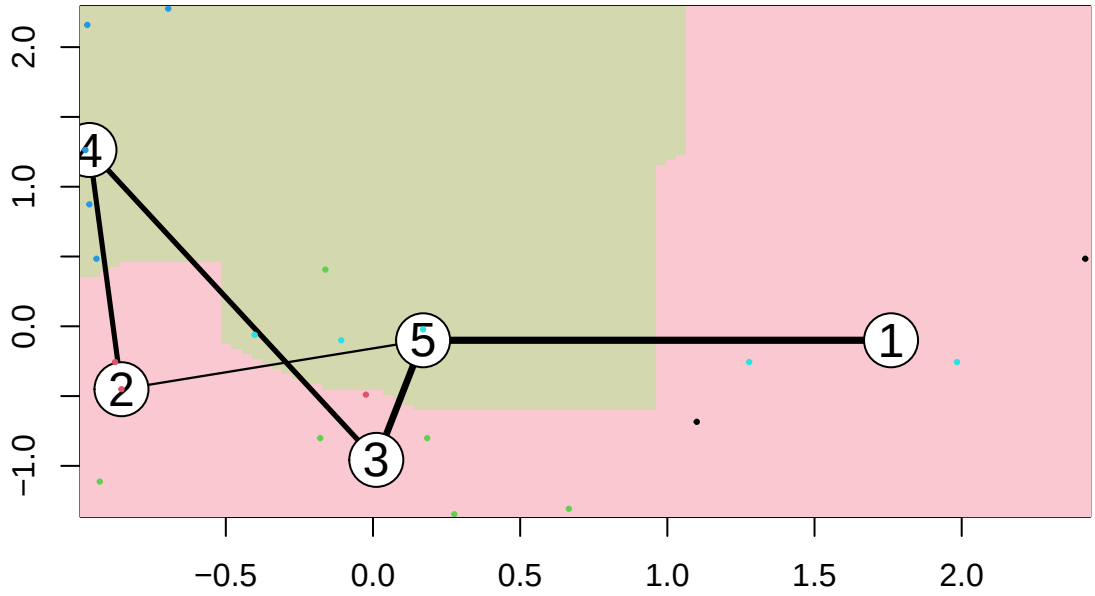
```
set.seed(69)

k_5.1 = kcca(df_pharma_2,k=5,kccaFamily("kmedians"))
k_5.1
```

```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = df_pharma_2, k = 5, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
## 1 2 3 4 5
## 2 3 6 5 5
```

```
Cluster_in <- predict(k_5.1)
dist(k_5.1@centers)
```

```
##          1        2        3        4
## 2 5.796625
## 3 3.847926 3.569392
## 4 5.559563 3.121363 3.249042
## 5 2.925045 3.649894 1.859338 3.521639
```

```
image(k_5.1)
points(df_pharma_2,col=Cluster_in,pch=19,cex=0.3)
```



Q2) 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)?

Sol: Here, we can see some patterns between the non-numeric variables to the numeric variable based clusters. For, #Cluster1: These stocks has higher market cap, ROE and ROA comparing to other clusters but low leverage. #Cluster2: These stocks are moderate in market cap, ROE, ROA and leverage. #Cluster3: These stocks are moderate again where the recommendations are more halfway, not complete. #Cluster4: The summary of cluster 4 is more in negatives, with second lowest market cap, ROE, ROA, lowest PE_ratio, Asset_turnover but highest Revenue growth. #Cluster5: These stocks have the highest leverage but lowest market cap.

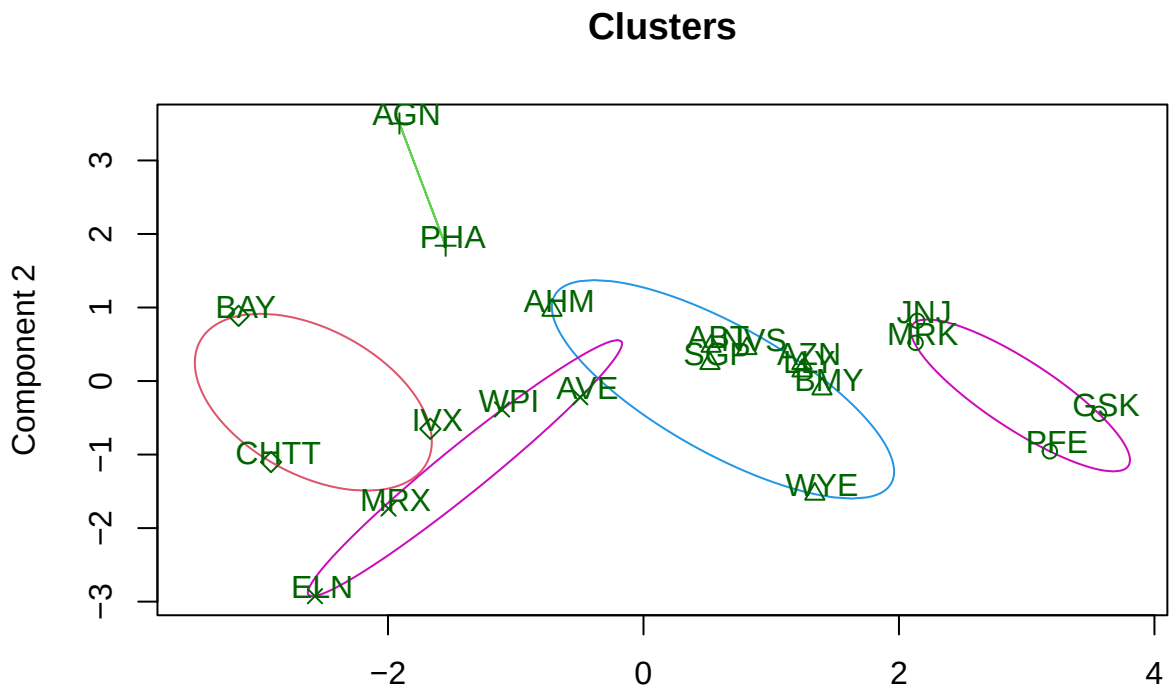
#Finding out the pattern between the numerical variables to non-numeric (10 to 12)

```
df_pharma_1%>%mutate(Cluster=k_5$cluster)%>%group_by(Cluster)%>%
summarise_all("mean")
```

```
## # A tibble: 5 x 10
##   Cluster Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##     <int>      <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
## 1       1     157.   0.48      22.2  44.4 17.7           0.95     0.22
## 2       2      55.8  0.414     20.3  28.7 12.7           0.738    0.371
## 3       3      31.9  0.405     69.5  13.2  5.6           0.75     0.475
## 4       4      13.1  0.598     17.7  14.6  6.2           0.425    0.635
```

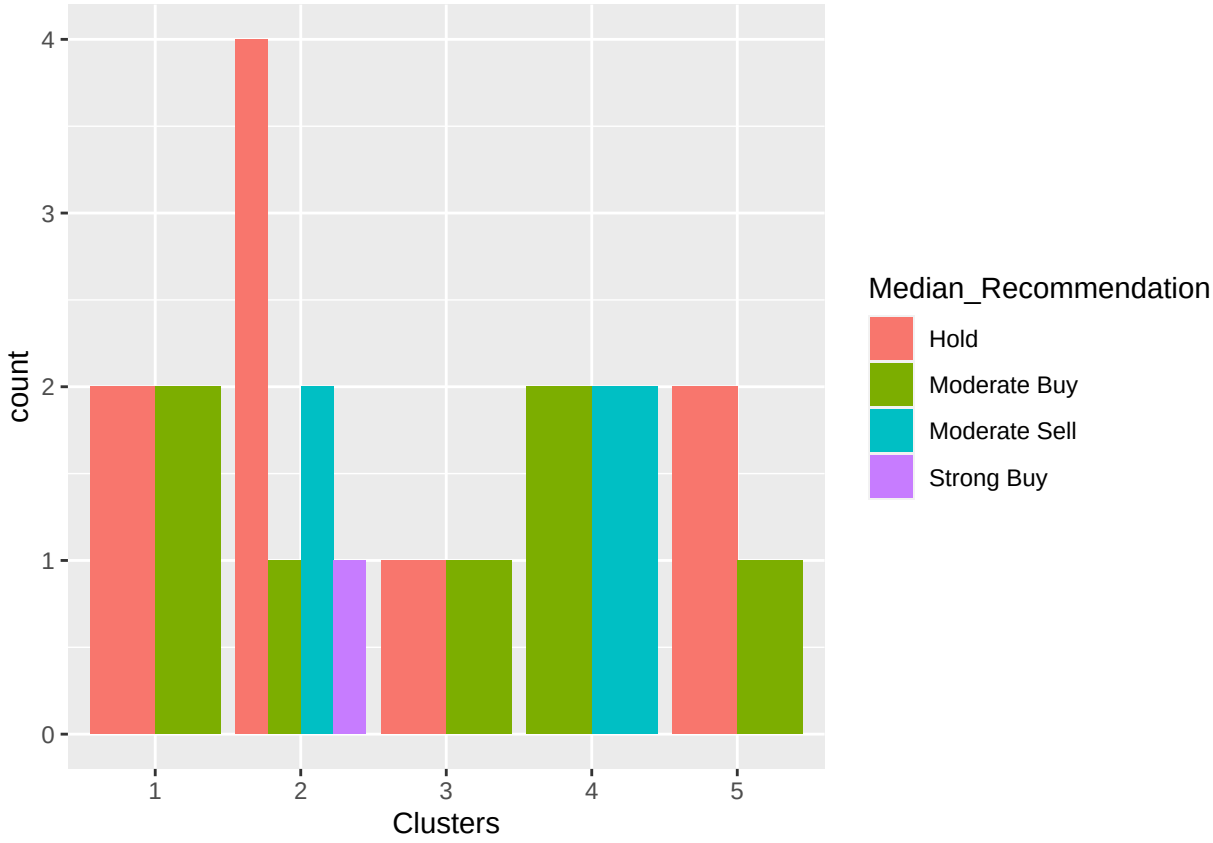
```
## 5       5       6.64 0.87      24.6  16.5  4.17          0.6       1.65
## # i 2 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>
```

```
clusplot(df_pharma_2,k_5$cluster,main="Clusters",color = TRUE, labels = 3,lines = 0)
```



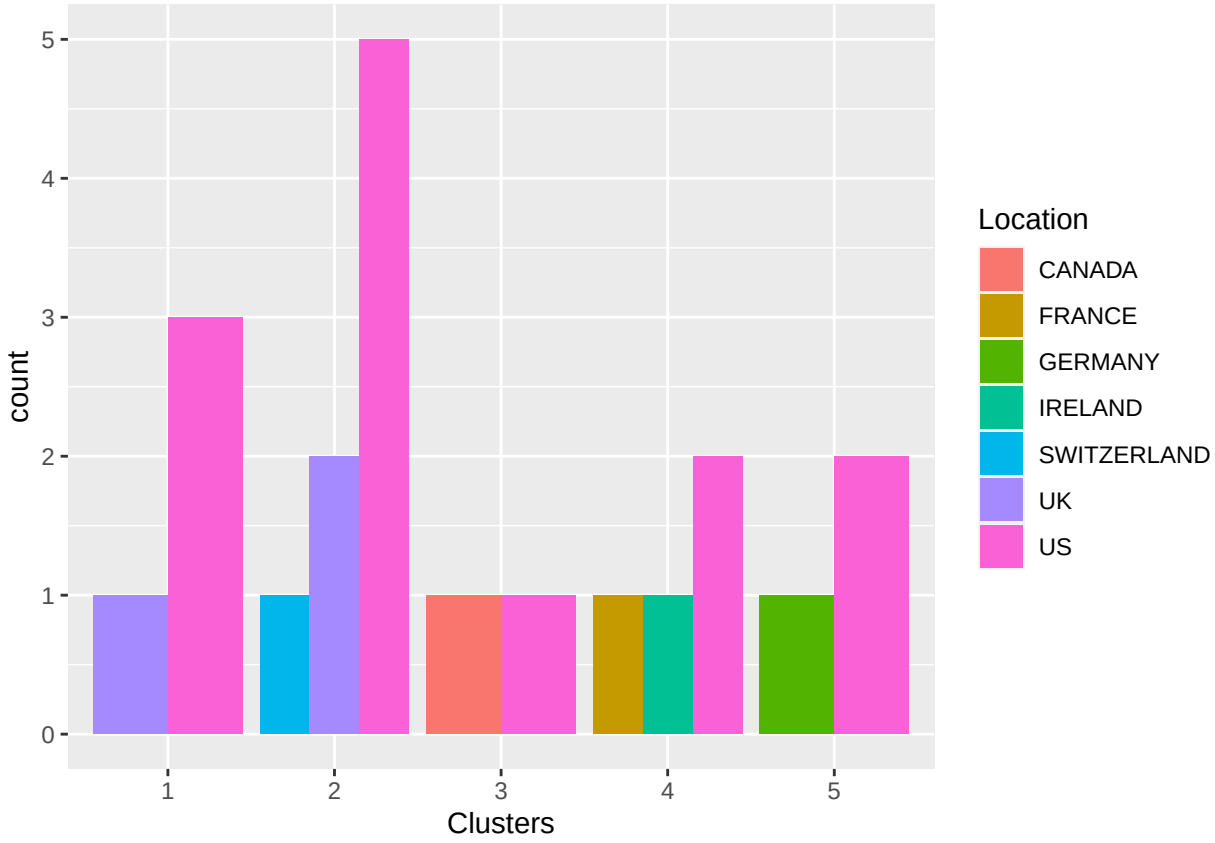
#Median recommendation for non-numeric value

```
df_pharma_3 <- df_pharma[12:14] %>% mutate(Clusters=k_5$cluster)
ggplot(df_pharma_3, mapping = aes(factor(Clusters), fill
=Median_Recommendation))+geom_bar(position='dodge')+labs(x ='Clusters')
```



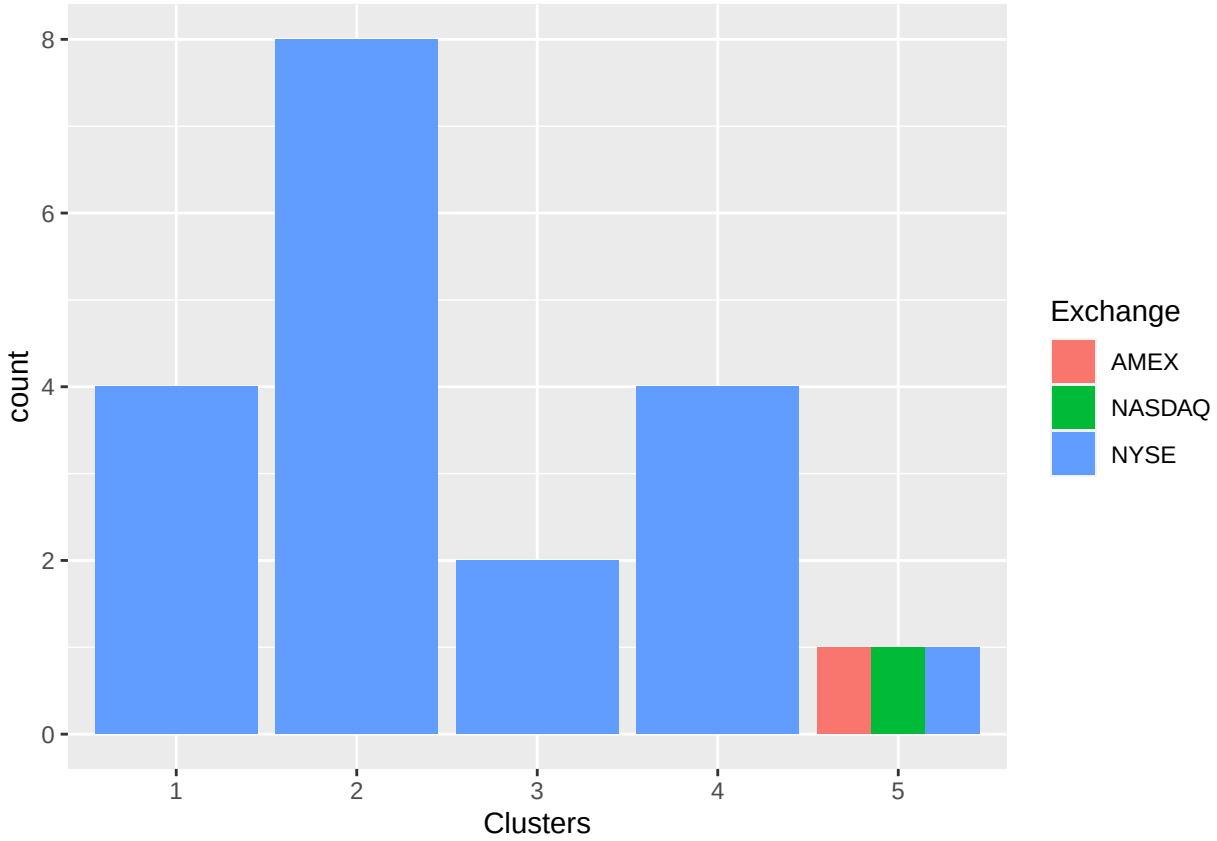
#Location for non-numeric value

```
ggplot(df_pharma_3, mapping = aes(factor(Clusters),fill =
Location))+geom_bar(position = 'dodge')+labs(x ='Clusters')
```



#Exchange for non-numeric value

```
ggplot(df_pharma_3, mapping = aes(factor(Clusters),fill =
Exchange))+geom_bar(position = 'dodge')+labs(x ='Clusters')
```



#By looking at all the graphs, we can interpret the following:

#Cluster1: It has a hold and a moderate buy recommendation, situated in UK and US and are traded on NYSE. #Cluster2: It is the only cluster with all the recommendations, in US, UK and Switzerland and are traded on NYSE. #Cluster3: It is similar to cluster 1 in recommendation but are less in count, functions in canada and US, and are traded on NYSE likewise clusters 1-4. #Cluster4: It is moderate in buy and sell for recommendations, functions in France, germany and US. #Cluster 5: It is the only cluster which is traded on all the markets(AMEX,NASDAQ,NYSE) and are present in Germany and US.

Q3) Provide an appropriate name for each cluster using any or all of the variables in the dataset.

Sol:

#For, #Cluster1: Topper with high scores since this has high market valuation, profit and low leverage. #Cluster2: Medium scaler since these are moderate in all the sections. #Cluster3: Moderate growth, since they are moderate in all the sections likewise cluster 2. #Cluster4: Risk bull, since it has high risk and also higher revenue growth. #Cluster5: Big market hero, since they are in all the exchange markets.