FML4

Vijay

2022-03-16

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.

#Loading packages  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.1.3

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.6 v dplyr 1.0.8  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1

## Warning: package 'readr' was built under R version 4.1.3

## Warning: package 'forcats' was built under R version 4.1.3

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(factoextra)

## Warning: package 'factoextra' was built under R version 4.1.3

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

library(flexclust)

## Warning: package 'flexclust' was built under R version 4.1.3

## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4

library(ggcorrplot)

## Warning: package 'ggcorrplot' was built under R version 4.1.3

library(FactoMineR)

## Warning: package 'FactoMineR' was built under R version 4.1.3

library(cluster)

Pharmdata <- read.csv("Pharmaceuticals.csv")  
head(Pharmdata)

## 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

Pharmdata1 <- Pharmdata[3:11]  
head(Pharmdata1)

## 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

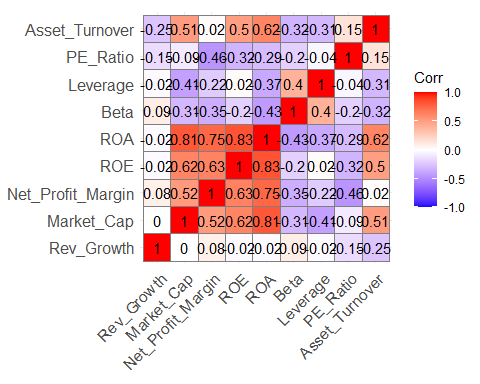
summary(Pharmdata1)

## 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

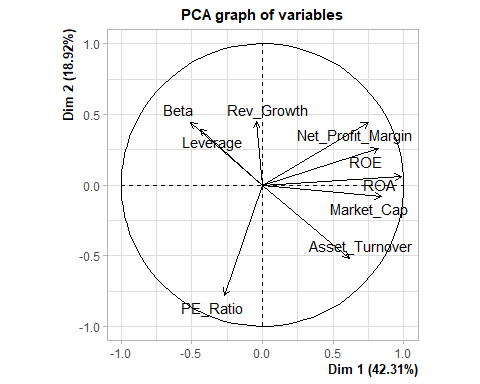
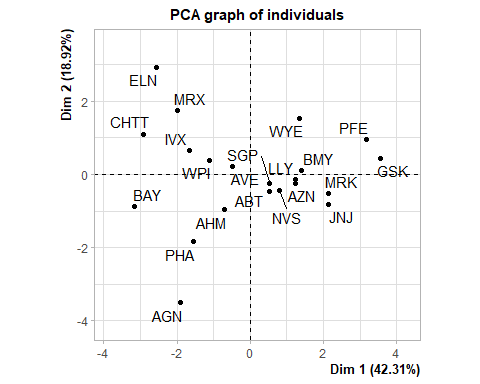
Pharmdata2 <- scale(Pharmdata1)  
row.names(Pharmdata2) <- Pharmdata[,1]  
distance <- get\_dist(Pharmdata2)  
fviz\_dist(distance)



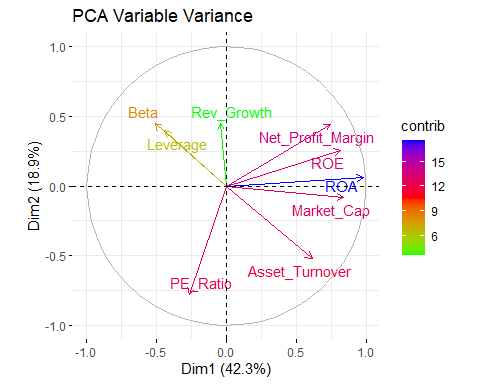
corr <- cor(Pharmdata2)  
ggcorrplot(corr, outline.color = "grey50", lab = TRUE, hc.order = TRUE, type = "full")



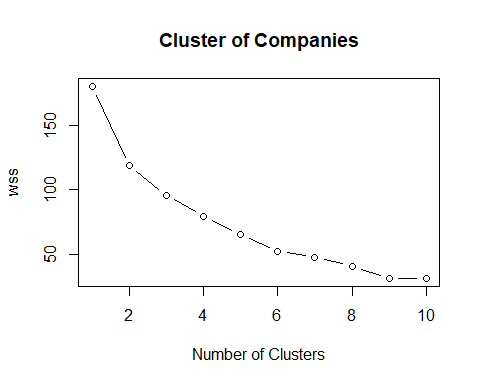
pca <- PCA(Pharmdata2)



var <- get\_pca\_var(pca)  
  
fviz\_pca\_var(pca, col.var="contrib",  
 gradient.cols = c("green", "red", "blue"),  
 repel = TRUE   
 ) +   
 labs( title = "PCA Variable Variance")



set.seed(10)  
wss <- vector()  
for(i in 1:10) wss[i] <- sum(kmeans(Pharmdata2,i)$withinss)   
plot(1:10, wss , type = "b" , main = paste('Cluster of Companies') , xlab = "Number of Clusters", ylab="wss")



wss

## [1] 180.00000 118.56934 95.99420 79.21748 65.61035 52.67476 47.66961  
## [8] 41.12605 31.81763 31.57252

## Silhouette Method

fviz\_nbclust(Pharmdata2, kmeans, method = "silhouette")



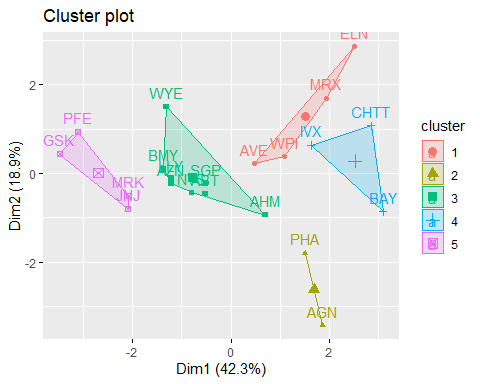
set.seed(1)  
k5 <- kmeans(Pharmdata2, centers = 5, nstart = 25) # k = 5, number of restarts = 25  
# Visualize the output  
k5$centers # output the centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.76022489 0.2796041 -0.47742380 -0.7438022 -0.8107428 -1.2684804  
## 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 1.69558112 -0.1780563 -0.19845823 1.2349879 1.3503431 1.1531640  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 0.06308085 1.5180158 -0.006893899  
## 2 -0.14170336 -0.1168459 -1.416514761  
## 3 -0.27449312 -0.7041516 0.556954446  
## 4 1.36644699 -0.6912914 -1.320000179  
## 5 -0.46807818 0.4671788 0.591242521

k5$size # Number of companies in each cluster

## [1] 4 2 8 3 4

fviz\_cluster(k5, data = Pharmdata2) # Visualize the output



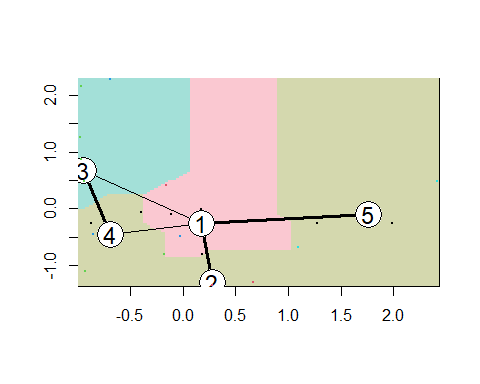
set.seed(1)  
k51 = kcca(Pharmdata2, k=5, kccaFamily("kmedians"))  
k51

## kcca object of family 'kmedians'   
##   
## call:  
## kcca(x = Pharmdata2, k = 5, family = kccaFamily("kmedians"))  
##   
## cluster sizes:  
##   
## 1 2 3 4 5   
## 7 3 6 3 2

clusters\_index <- predict(k51)  
dist(k51@centers)

## 1 2 3 4  
## 2 2.150651   
## 3 3.513242 4.146567   
## 4 3.878726 4.246051 3.388339   
## 5 3.018500 3.737739 5.124420 6.043691

image(k51)  
points(Pharmdata2, col=clusters\_index, pch=19, cex=0.3)



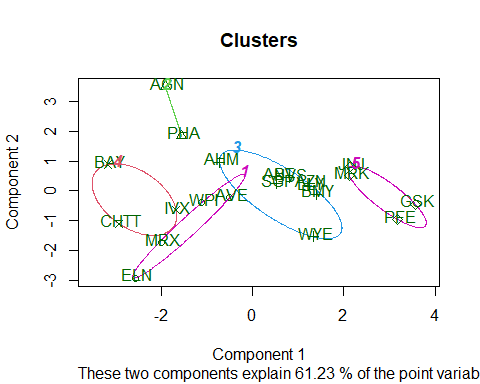
b.Interpret the clusters with respect to the numerical variables used in forming the clusters

#Calculating Mean of all variables for every cluster and plotting them

Pharmdata1 %>% mutate(Cluster = k5$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 13.1 0.598 17.7 14.6 6.2 0.425 0.635  
## 2 2 31.9 0.405 69.5 13.2 5.6 0.75 0.475  
## 3 3 55.8 0.414 20.3 28.7 12.7 0.738 0.371  
## 4 4 6.64 0.87 24.6 16.5 4.17 0.6 1.65   
## 5 5 157. 0.48 22.2 44.4 17.7 0.95 0.22   
## # ... with 2 more variables: Rev\_Growth <dbl>, Net\_Profit\_Margin <dbl>

clusplot(Pharmdata2,k5$cluster, main="Clusters",color = TRUE, labels = 2,lines = 0)



Cluster 1: ELN, MRX, WPI and AVE

Cluster 2: AGN and PHA

Cluster 3: AHM,WYE,BMY,AZN, LLY, ABT, NVS and SGP

Cluster 4: BAY, CHTT and IVX

Cluster 5: JNJ, MRK, PFE and GSK

Cluster 1 has got highest revenue growth , very good Net profit Margin and leverage with lowest PE ratio. It can be bought or hold.

Cluster 2 PE ratio is very high , inferring that investors are expecting high growth , however, growth rate is only 12% and Net profit Margin is also low , making it overvalued and may not be a good choice overall.

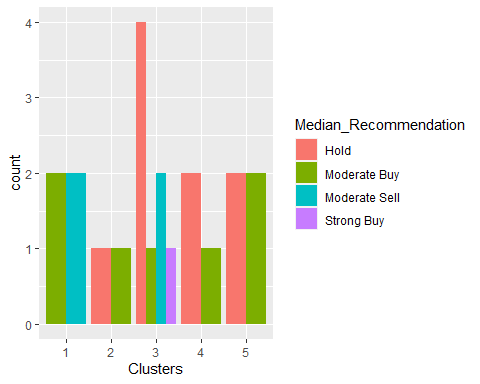
Cluster 3 has average risk (Beta) and relatively high Market Cap, ROE, ROA, Asset Turnover and Net Profit Margin ,high leverage.Attractive (relatively low) PE ratio indicates that the stock price is moderately valued hence can be bought and hold , making it ideal to own.

Cluster 4 Though it has a good PE ratio, it carries a very high risk , very very high leverage and low Net Profit margin , making it very risky to own. Revenue growth is also very low.

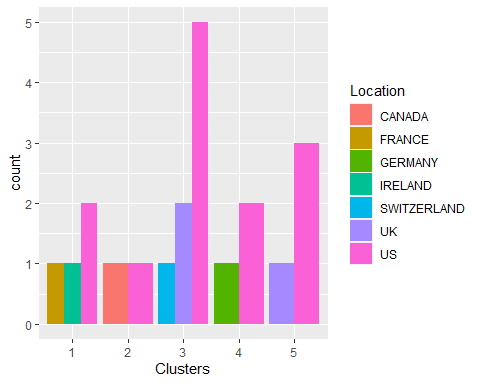
Cluster 5 is great with High Market Cap, ROE, ROA, Asset Turnover and Net Profit Margin. With a relatively low PE ratio the stock price is moderately valued, hence can be bought and hold.Further , revenue growth of 18.5% is good.

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

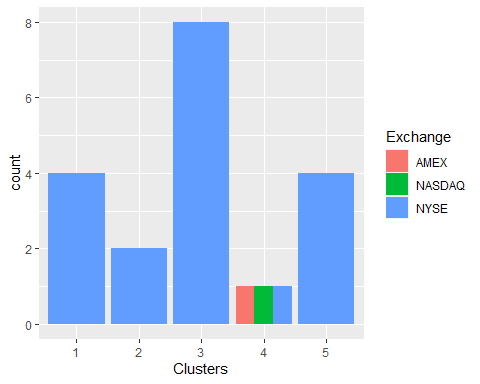
Pharmdata3 <- Pharmdata[12:14] %>% mutate(Clusters=k5$cluster)  
ggplot(Pharmdata3, mapping = aes(factor(Clusters), fill =Median\_Recommendation))+geom\_bar(position='dodge')+labs(x ='Clusters')



ggplot(Pharmdata3, mapping = aes(factor(Clusters),fill = Location))+geom\_bar(position = 'dodge')+labs(x ='Clusters')



ggplot(Pharmdata3, mapping = aes(factor(Clusters),fill = Exchange))+geom\_bar(position = 'dodge')+labs(x ='Clusters')



d.Provide an appropriate name for each cluster using any or all of the variables in the dataset. Cluster 1: Good to buy or to hold Cluster 2: Risk better to sell Cluster 3: Take chance to buy or to hold Cluster 4: Highly Risky better to sell Cluster 5: Best time to buy or to hold