FML\_Assignment\_4

Chakri

2023-10-29

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:
13. 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 the data of Pharmaceuticals  
p\_data <- read.csv("H:\\Kent Sem-1\\FML\\FML\_class\\Assignments\\Assignment4\\Pharmaceuticals.csv")  
head(p\_data) #printing the first 6 rows of the dataset

## 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

#loading the required libraries  
library(tidyverse) #data manipulation

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── 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(factoextra)#clustering algorithms and visualization

## 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

library(ISLR)  
library(cluster)

#gathering only the numerical variables that are from 3 to 11

df\_p.data <- p\_data[,c(3:11)]  
#summary of the data  
summary(df\_p.data)

## 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

#The above variables describe the financial metrics like investment potential, and evaluate between companies and to find out the risk factors that are affecting the investments in a particular company.

#we normalize the data such that each variable can have equal domination with one another

#scaling the data frame (Z-score)/ normalizing the data with scale function  
df\_phar.data <- scale(df\_p.data)

#By using Elbow method for the cluster analysis by determining the number of clusters.

fviz\_nbclust(df\_phar.data, kmeans, method = "wss")



#from the above graph we are unable to determine the number of clusters accurately for that we need to do by method called silhoulette.

fviz\_nbclust(df\_phar.data, kmeans, method = "silhouette")



#the above analysis measures that how similar of the object within its cluster compared to across the clusters. #from the above graph the most accurate number of clusters is 5

set.seed(96441)  
k.means <- kmeans(df\_phar.data, centers = 5, nstart = 25) #k=4, randomly pick the starting centeroids 25 times  
k.means$centers

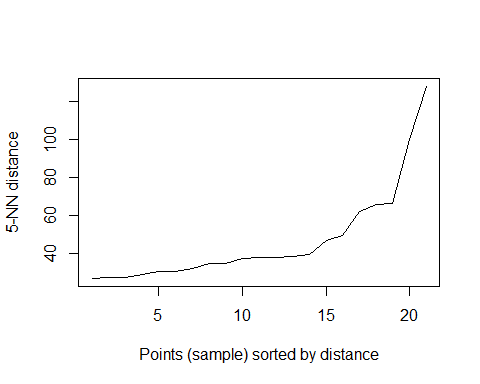
## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 2 1.69558112 -0.1780563 -0.19845823 1.2349879 1.3503431 1.1531640  
## 3 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 4 -0.43925134 -0.4701800 2.70002464 -0.8349525 -0.9234951 0.2306328  
## 5 -0.76022489 0.2796041 -0.47742380 -0.7438022 -0.8107428 -1.2684804  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.27449312 -0.7041516 0.556954446  
## 2 -0.46807818 0.4671788 0.591242521  
## 3 1.36644699 -0.6912914 -1.320000179  
## 4 -0.14170336 -0.1168459 -1.416514761  
## 5 0.06308085 1.5180158 -0.006893899

k.means

## K-means clustering with 5 clusters of sizes 8, 4, 3, 2, 4  
##   
## Cluster means:  
## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 2 1.69558112 -0.1780563 -0.19845823 1.2349879 1.3503431 1.1531640  
## 3 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 4 -0.43925134 -0.4701800 2.70002464 -0.8349525 -0.9234951 0.2306328  
## 5 -0.76022489 0.2796041 -0.47742380 -0.7438022 -0.8107428 -1.2684804  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.27449312 -0.7041516 0.556954446  
## 2 -0.46807818 0.4671788 0.591242521  
## 3 1.36644699 -0.6912914 -1.320000179  
## 4 -0.14170336 -0.1168459 -1.416514761  
## 5 0.06308085 1.5180158 -0.006893899  
##   
## Clustering vector:  
## [1] 1 4 1 1 5 3 1 3 5 1 2 3 2 5 2 1 2 4 1 5 1  
##   
## Within cluster sum of squares by cluster:  
## [1] 21.879320 9.284424 15.595925 2.803505 12.791257  
## (between\_SS / total\_SS = 65.4 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

#Db scan plot for

dbscan::kNNdistplot(df\_p.data, k=5)



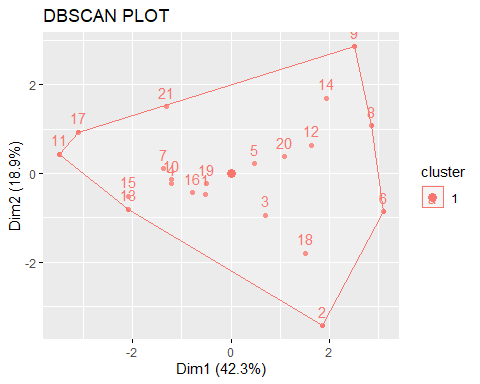
#take the package dbscan from the dbscan  
db <- dbscan::dbscan(df\_phar.data, eps = 40, minPts = 5) #perform clustering  
  
print(db)

## DBSCAN clustering for 21 objects.  
## Parameters: eps = 40, minPts = 5  
## Using euclidean distances and borderpoints = TRUE  
## The clustering contains 1 cluster(s) and 0 noise points.  
##   
## 1   
## 21   
##   
## Available fields: cluster, eps, minPts, dist, borderPoints

#clusters for dbscan  
  
db$cluster

## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

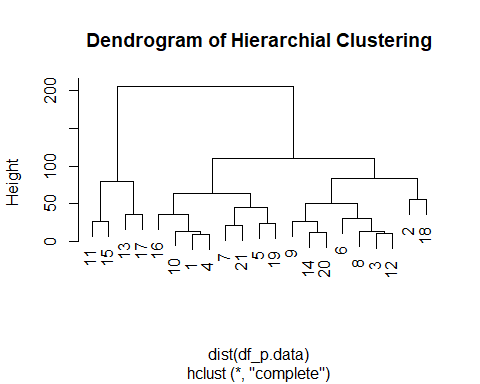
fviz\_cluster(db,df\_p.data)+ ggtitle("DBSCAN PLOT")

 #it has only one cluster and it is highly affect for interpreting the results

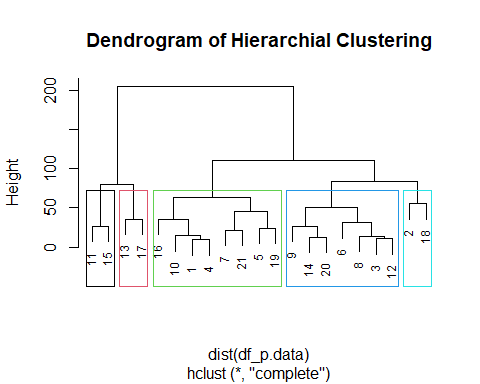
hc <- hclust(dist(df\_p.data), method = "complete")  
hc

##   
## Call:  
## hclust(d = dist(df\_p.data), method = "complete")  
##   
## Cluster method : complete   
## Distance : euclidean   
## Number of objects: 21

plot(hc, main = "Dendrogram of Hierarchial Clustering")



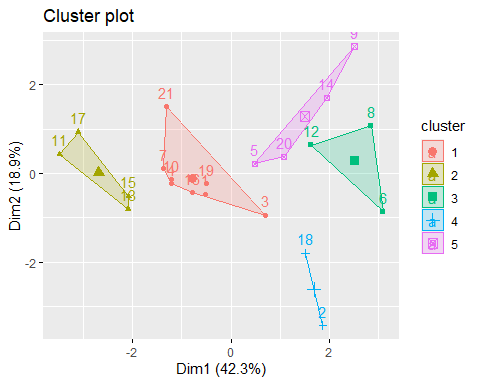
plot(hc,cex= 0.75, main="Dendrogram of Hierarchial Clustering")  
rect.hclust(hc, k=5, border = 1:5)



#From the above interpretation k-means will be used over the DBSCAN and hierarchial because it has more clusters and identified patterns/characteristics in variables and grouping the data and it provides in depth analysis of financial threats and profits that can develop the large, medium and small scale industries to easily interpret the data. DBSCAN will applicable only in collude points or in a crowd points where as in heirarchial few clusters has only few points and might be affect with other business data.

#TO visualize the five clusters

fviz\_cluster(k.means, df\_phar.data)

 #Cluster 1 represents large companies with strong profitability and high leverage #Cluster 2 represents medium-sized companies high profitability and low leverage #Cluster 3 represents medium-sized companies moderate profitability and low leverage #Cluster 4 represents small-sized companies with low profitability and leverage #Cluster 5 represents small-sized companies with moderate profitability and high leverage

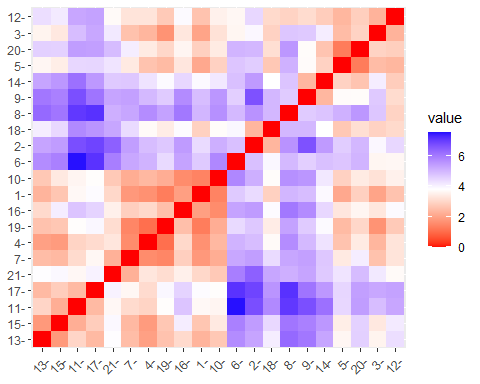
#interpretation of Clusters #Cluster 1- As it represents strong profit and leverage investors are more confident in these prospects where as the most investors will invest in these companies #CLuster 2- As it represents high profit and low leverage lower median recommendation than cluster 1 but they can be useful/potential in their growth #Cluster 3- Represents low profit and high revenue growth and moderate median recommendations might be the investors put an interest and excited to invest eventhough it is risk #Cluster 4- Represents low profit and leverage and lower median recommendation than cluster 2 but they might offer potential and growth in their fields #Cluster 5- Represents moderate profit and high leverage but it has highest revenue growth of these companies are good.

#K-means cluster analysis–> by Fitting the data with 5 clusters

fit <- kmeans(df\_phar.data, 5)  
#finding the mean value of all numerical variables for each of the five clusters  
aggregate(df\_p.data, by=list(fit$cluster),FUN=mean)

## Group.1 Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage  
## 1 1 72.89333 0.3333333 20.86667 34.600 12.700 0.5666667 0.5700  
## 2 2 28.56000 0.3425000 44.37500 14.600 6.375 0.6500000 0.3725  
## 3 3 157.01750 0.4800000 22.22500 44.425 17.700 0.9500000 0.2200  
## 4 4 45.56000 0.4620000 19.94000 25.220 12.680 0.8400000 0.2520  
## 5 5 4.37800 0.8880000 21.20000 15.140 4.600 0.4800000 1.3920  
## Rev\_Growth Net\_Profit\_Margin  
## 1 1.293333 23.76667  
## 2 20.037500 10.20000  
## 3 18.532500 19.57500  
## 4 8.170000 16.70000  
## 5 16.356000 11.14000

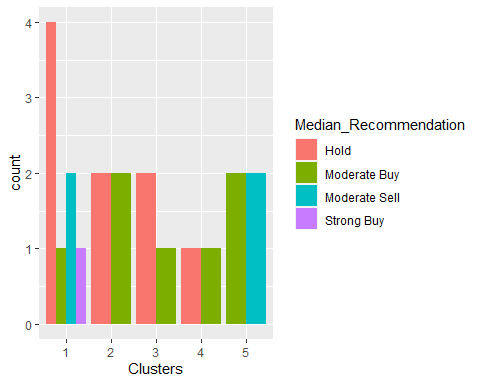
#here the dist is to compute the distance matrix by taking a data frame as an input  
#the below distance is called euclidean distance  
dist\_df <- dist(df\_phar.data, method = "euclidean")  
#computing and visualizing the distance matrix  
fviz\_dist(dist\_df)



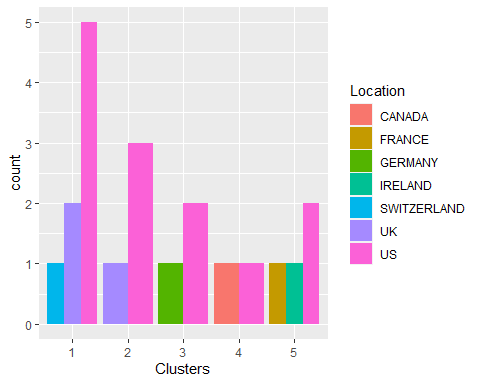
#the above graph indicates that there are two varaibles that are x values indicates market capitalization and y values indicate median recommendation. so, therefore the median recommendation is depends on the markent capitalization. #And the above graph or distance matrix is highly correlated because the hi #the above matrix shows that distance between each pair of rows in the dataframe. the darkest color which is above 6 indicates the greater distance between the two rows that means for higher the market capitalization the less the median recommendation and vice versa

1. 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)

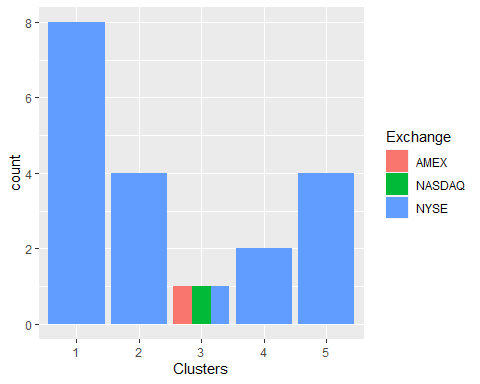
rem\_data <- p\_data[12:14]  
rem\_data$Clusters <- k.means$cluster  
ggplot(rem\_data, aes(factor(Clusters), fill= Median\_Recommendation))+ geom\_bar(position='dodge')+labs(x='Clusters')

 #from the above graph of the median recommendation there is a high possibility that the hold ones are in cluster one and as well has strong buy which the investors of large group will come into this cluster #from the second cluster the hold and moderate buy seems to be equal that meand it is moderate profit with leverage #from the third cluster there are both hold which is to be expected some growth and high risk and the investors will give less interest in it #from the fourth cluster there are hold and moderate profit and which is very less compared to other clusters #from the fifth cluster there are moderate profit and high leverage

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

 #from the all the clusters united states is the number one to spend or to get the most profit and high leverage in every cluster #as for the germany and switzerland it has low profit and negative revenue growth compared to other countries #Ireland has the highest revenue growth that is highest from all over the country with moderate profit #where as france has high revenue but low profit compared to other countries

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

 #from the above graph NYSE exchange is the highest stock markets and it is well known in all over the countries #where as AMEX and NASDAQ is only included in third cluster that means germany is the one using this exchange #interpretation of clusters based on the remaining categorical varibles from 10 to 12

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

#Cluster-1 Profitable companies #Cluster-2 High growth and potential companies #Cluster-3 Low Revenue companies #Cluster-4 Value companies #Cluster-5 Growth and Leveraged companies