FML assignment-4

Akshaya Mamidipalli

2024-03-17

library(flexclust)

## Warning: package 'flexclust' was built under R version 4.3.3

## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4

library(cluster)  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'ggplot2' was built under R version 4.3.2

## Warning: package 'forcats' was built under R version 4.3.3

## ── 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 3.4.4 ✔ 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)

## Warning: package 'factoextra' was built under R version 4.3.3

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

library(FactoMineR)

## Warning: package 'FactoMineR' was built under R version 4.3.3

library(ggcorrplot)

## Warning: package 'ggcorrplot' was built under R version 4.3.3

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.

data <- read.csv("C:/Users/mamid/Downloads/Pharmaceuticals.csv")

Numerical\_data <- data[3:11]

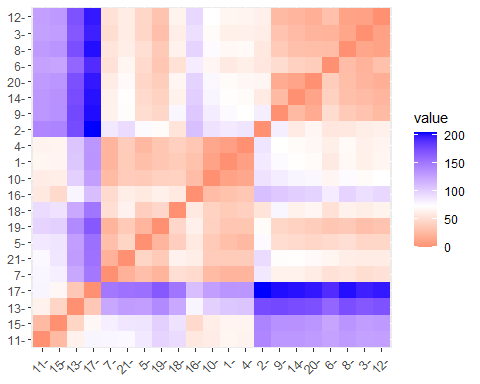
head(Numerical\_data)

## 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

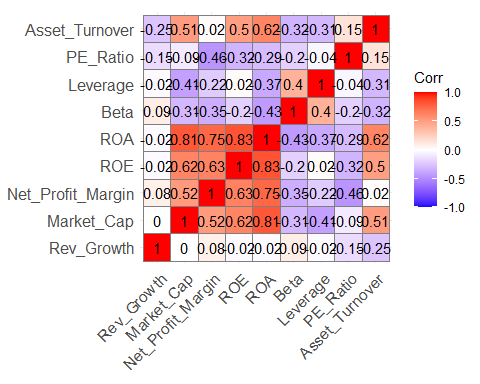
summary(Numerical\_data)

## 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

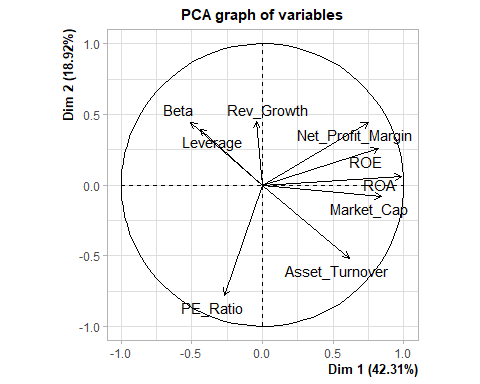
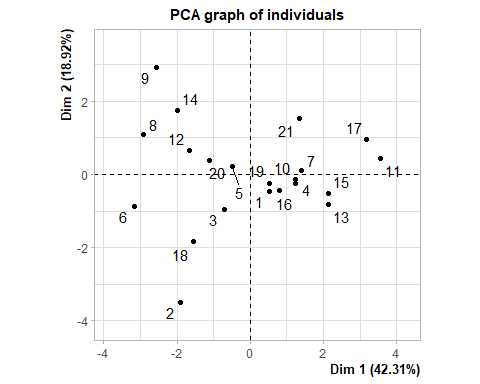
standardized\_data <- scale(Numerical\_data)  
distance <- get\_dist(Numerical\_data)  
fviz\_dist(distance)



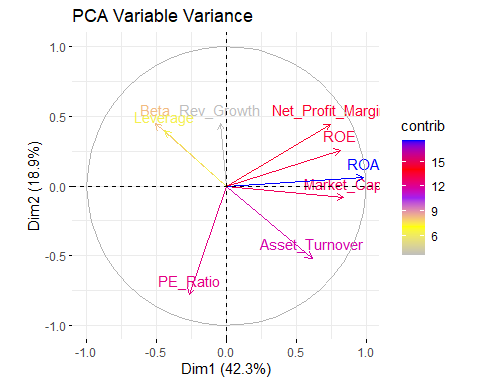
corr <- cor(standardized\_data)  
ggcorrplot(corr, outline.color = "grey50", lab = TRUE, hc.order = TRUE, type = "full")



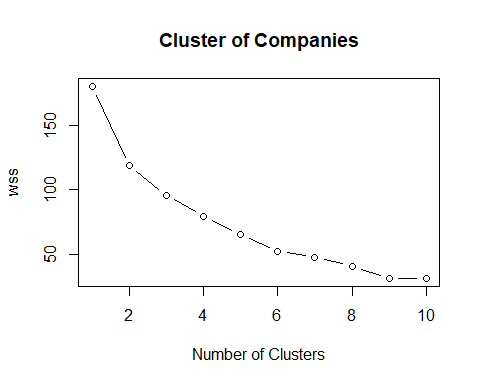
pca <- PCA(standardized\_data)



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



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



fviz\_nbclust(standardized\_data, kmeans, method = "silhouette")



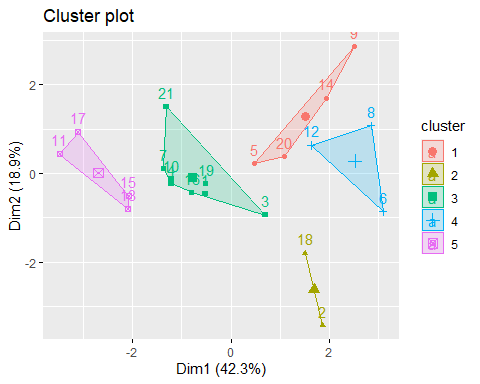
set.seed(1)  
k5 <- kmeans(standardized\_data, centers = 5, nstart = 20)   
# k = 5, number of restarts = 20  
k5$centers

## 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

## [1] 4 2 8 3 4

fviz\_cluster(k5, data = standardized\_data)

 #K-means clustering, using Manhattan Distance

set.seed(1)  
k51 <- kcca(standardized\_data, k=5, kccaFamily("kmedians"))  
k51

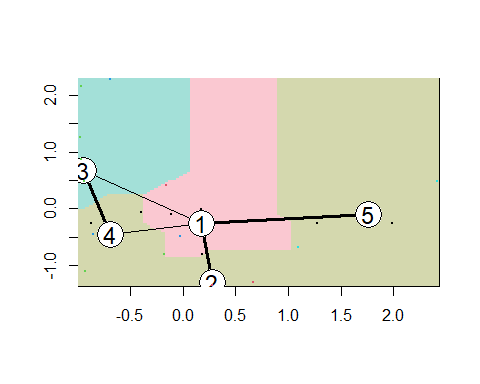
## kcca object of family 'kmedians'   
##   
## call:  
## kcca(x = standardized\_data, k = 5, family = kccaFamily("kmedians"))  
##   
## cluster sizes:  
##   
## 1 2 3 4 5   
## 7 3 6 3 2

#Using predict function

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

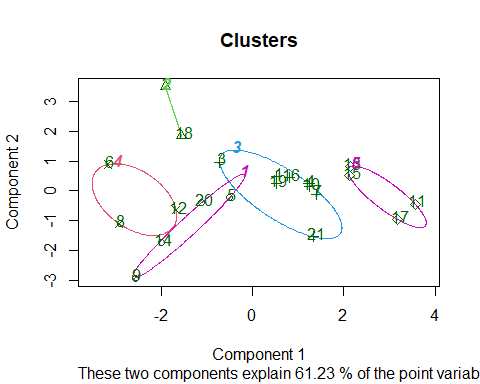
image(k51)  
points(standardized\_data, col=clusters\_index, pch=19, cex=0.3)

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

Numerical\_data %>% mutate(Cluster = k5$cluster) %>% group\_by(Cluster) %>%   
summarise\_all("mean")

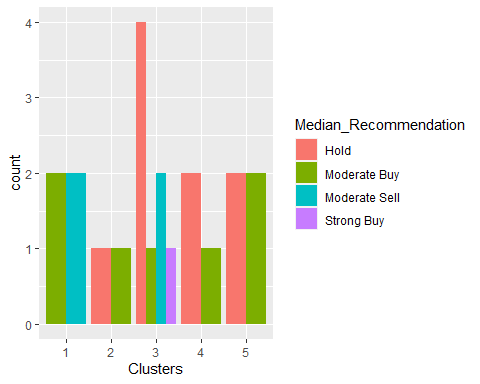
## # A tibble: 5 × 10  
## Cluster Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage  
## <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 13.1 0.598 17.7 14.6 6.2 0.425 0.635  
## 2 2 31.9 0.405 69.5 13.2 5.6 0.75 0.475  
## 3 3 55.8 0.414 20.3 28.7 12.7 0.738 0.371  
## 4 4 6.64 0.87 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   
## # ℹ 2 more variables: Rev\_Growth <dbl>, Net\_Profit\_Margin <dbl>

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

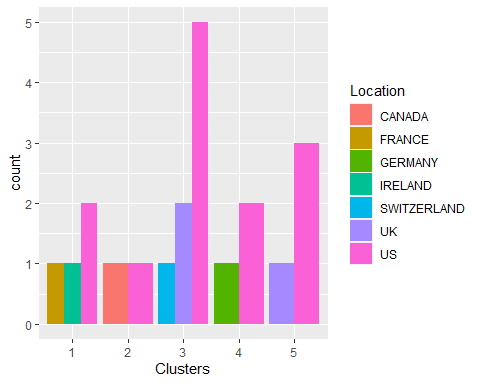
 Cluster 1: AVE, WPI, MRX, and ELN Group 2: PHA and AGN Cluster 3: AZN, LLY, ABT, NVS, SGP, AHM, WYE, BMY, and AZN Cluster 4: IVX, BAY, and CHTT Cluster 5: PFE, GSK, MRK, and JNJ We can infer that Cluster 1 has the highest Net Profit Margin, the lowest PE ratio, and the fastest revenue growth based on the means of the cluster variables. It is available for purchase or reserve.PE ratio for Cluster 2 is very high. Cluster 3’s risk is average. Cluster 4: Despite having a good PE ratio, this stock is extremely risky to own due to its high risk, extremely high leverage, and low net profit margin.. Revenue growth is also very low.Cluster 5 has a high market capitalization, return on investment, return on assets, asset turnover, and net profit margin. With a low PE ratio, the stock price is moderately valued and hence can be purchased and held evenue 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)

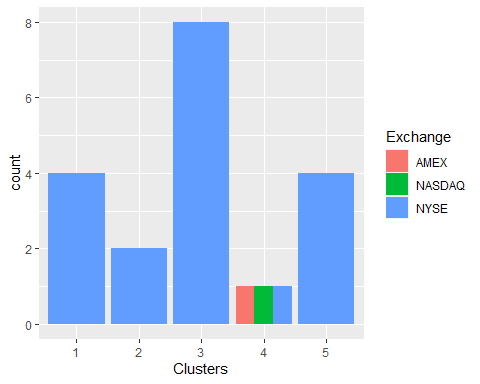
Info <- data[12:14] %>% mutate(Clusters=k5$cluster)  
ggplot(Info, mapping = aes(factor(Clusters), fill =Median\_Recommendation))+geom\_bar(position='dodge')+labs(x ='Clusters')



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



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

 There seems to be a pattern with clusters and the variable Median Recommendation. There doesn’t seem to be any discernible pattern among the clusters, locations, or exchanges, with the exception of the majority of the clusters/companies being based in the United States and listed on the NYSE.

##d.Provide an appropriate name for each cluster using any or all of the variables in the dataset. Cluster 1: Best Buying Cluster 2: Highly Risky Cluster 3: Go for it Cluster 4: Very Risky or Runaway Cluster 5: Ideal to Own