Machine Learning Assignment 4

Brian Yim

2025-10-26

####An equities analyst is studying the pharmaceutical industry and would like your help in exploring and understanding the financial data collected by her firm. Her main objective is to understand the structure of the pharmaceutical industry using some basic financial measures. Financial data gathered on 21 firms in the pharmaceutical industry are available in the file Pharmaceuticals.csv. For each firm, the following variables are recorded:

library(readr)  
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.4.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.4.3

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

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.4.3

## Warning: package 'dplyr' was built under R version 4.4.3

## Warning: package 'forcats' was built under R version 4.4.3

## Warning: package 'lubridate' was built under R version 4.4.2

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ stringr 1.5.1  
## ✔ forcats 1.0.0 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ 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(broom)

## Warning: package 'broom' was built under R version 4.4.2

library(clValid)

## Warning: package 'clValid' was built under R version 4.4.3

## Loading required package: cluster

library('fastDummies')

## Warning: package 'fastDummies' was built under R version 4.4.3

library(flexclust)

## Warning: package 'flexclust' was built under R version 4.4.3

library(dplyr)  
library(ggplot2)  
library(zoo)

## Warning: package 'zoo' was built under R version 4.4.3

##   
## Attaching package: 'zoo'  
##   
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(reshape2)

## Warning: package 'reshape2' was built under R version 4.4.2

##   
## Attaching package: 'reshape2'  
##   
## The following object is masked from 'package:tidyr':  
##   
## smiths

library(NbClust)

#reading the data

Pharma <- read.csv("C:/Users/hyim/OneDrive - Kent State University/Desktop/64060 Assignment/Pharmaceuticals.csv")  
View(Pharma)  
data.frame(colnames(Pharma))

## colnames.Pharma.  
## 1 Symbol  
## 2 Name  
## 3 Market\_Cap  
## 4 Beta  
## 5 PE\_Ratio  
## 6 ROE  
## 7 ROA  
## 8 Asset\_Turnover  
## 9 Leverage  
## 10 Rev\_Growth  
## 11 Net\_Profit\_Margin  
## 12 Median\_Recommendation  
## 13 Location  
## 14 Exchange

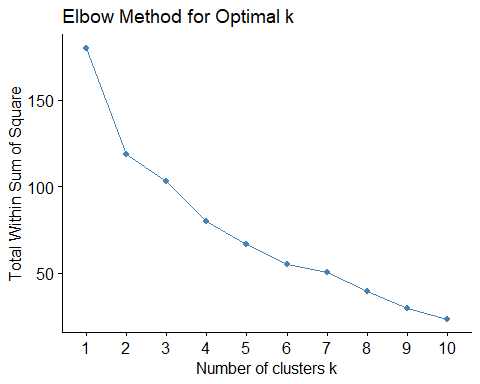
str(Pharma)

## 'data.frame': 21 obs. of 14 variables:  
## $ Symbol : chr "ABT" "AGN" "AHM" "AZN" ...  
## $ Name : chr "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZeneca PLC" ...  
## $ Market\_Cap : num 68.44 7.58 6.3 67.63 47.16 ...  
## $ Beta : num 0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...  
## $ PE\_Ratio : num 24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...  
## $ ROE : num 26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...  
## $ ROA : num 11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...  
## $ Asset\_Turnover : num 0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...  
## $ Leverage : num 0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...  
## $ Rev\_Growth : num 7.54 9.16 7.05 15 26.81 ...  
## $ Net\_Profit\_Margin : num 16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...  
## $ Median\_Recommendation: chr "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sell" ...  
## $ Location : chr "US" "CANADA" "UK" "UK" ...  
## $ Exchange : chr "NYSE" "NYSE" "NYSE" "NYSE" ...

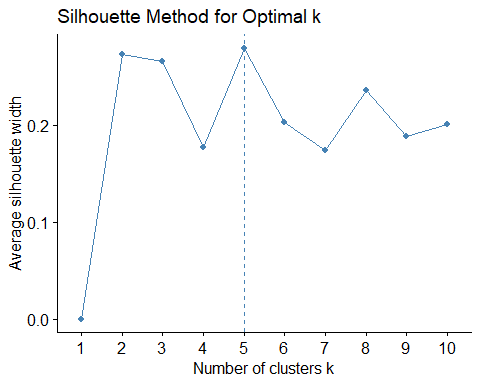
#Q.a. The objective of this analysis is to identify distinct groups of pharmaceutical firms that exhibit similar financial characteristics. The dataset consists of 21 firms and 12 variables, of which the first nine numerical indicators—Market Capitalization, Beta, Price/Earnings Ratio, Return on Equity (ROE), Return on Assets (ROA), Asset Turnover, Leverage, Estimated Revenue Growth, and Net Profit Margin—were selected for clustering. All numerical variables were standardized using z-scores to mitigate the effects of differing scales and units of measurement. This step ensures that variables measured in large magnitudes, such as Market Capitalization, do not disproportionately influence those expressed as ratios or percentages. Equal weighting was applied to all variables, as no theoretical or empirical justification existed for prioritizing one financial measure over another.Clustering was performed using the k-means algorithm on the standardized data. The k-means method partitions the observations into k clusters by minimizing the within-cluster sum of squares, producing compact and internally coherent groups. The optimal number of clusters was determined using the elbow method and silhouette width, and NbClust Package and all of the method indicated that a five-cluster solution provided the best balance between cluster separation and interpretability. The resulting clusters reveal meaningful distinctions among firms in terms of size, profitability, growth, and financial structure. Please see below for the solution for “a”

#set the working data  
mydata<-Pharma [,c(3:11)] #select numerical variables (1-9)  
View(mydata)  
row.names(mydata)<-Pharma[,1]#set row names to the pharmaceutical column  
View(mydata)  
# Standardize data (important for fair comparison)  
pharma\_scaled <- scale(mydata)

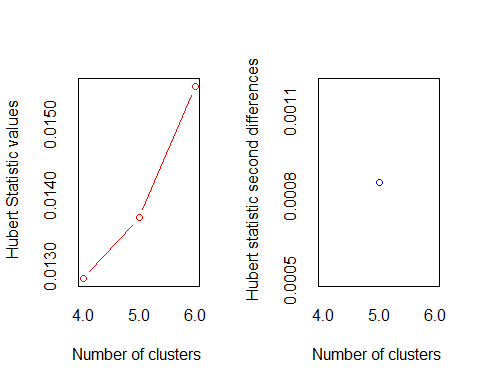
#Determine optimal number of clusters (k) using elbow and silhouette methods  
# Elbow method  
fviz\_nbclust(pharma\_scaled, kmeans, method = "wss") +   
 ggtitle("Elbow Method for Optimal k")



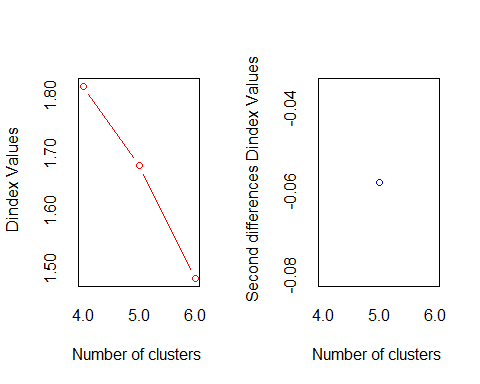
#Silhouette method  
fviz\_nbclust(pharma\_scaled, kmeans, method = "silhouette") +   
 ggtitle("Silhouette Method for Optimal k")



#NbClust Package for determining the best number of clusters  
K\_Suggestion<- NbClust(pharma\_scaled[,1:9], min.nc = 4, max.nc =6, method="kmeans")



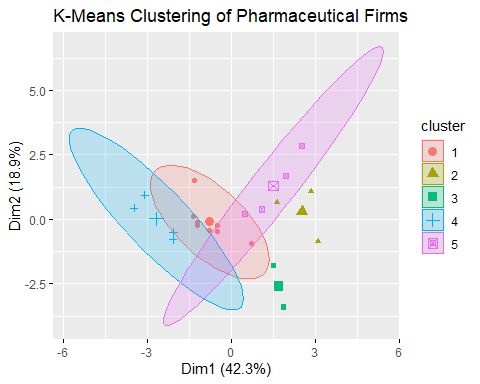
## \*\*\* : The Hubert index is a graphical method of determining the number of clusters.  
## In the plot of Hubert index, we seek a significant knee that corresponds to a   
## significant increase of the value of the measure i.e the significant peak in Hubert  
## index second differences plot.   
##



## \*\*\* : The D index is a graphical method of determining the number of clusters.   
## In the plot of D index, we seek a significant knee (the significant peak in Dindex  
## second differences plot) that corresponds to a significant increase of the value of  
## the measure.   
##   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*   
## \* Among all indices:   
## \* 7 proposed 4 as the best number of clusters   
## \* 8 proposed 5 as the best number of clusters   
## \* 8 proposed 6 as the best number of clusters   
##   
## \*\*\*\*\* Conclusion \*\*\*\*\*   
##   
## \* According to the majority rule, the best number of clusters is 5   
##   
##   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#Apply k-means clustering  
set.seed(123) # for reproducibility  
km <- kmeans(pharma\_scaled, centers = 5, nstart = 25)  
#Add cluster assignments to dataset  
mydata$Cluster <- factor(km$cluster)  
#Visualize clusters  
fviz\_cluster(km, data = pharma\_scaled, geom = "point", ellipse.type = "norm",  
 main = "K-Means Clustering of Pharmaceutical Firms")

## Too few points to calculate an ellipse  
## Too few points to calculate an ellipse



#Run k-means algorithm (k=5)

#run the k-means algorithm to cluster the companies (k=5) in order to evaluate the clustering quality  
k5<-kmeans(pharma\_scaled, centers=5, nstart=25) # test run with k=5, and 25 starts (re-run).   
k5

## K-means clustering with 5 clusters of sizes 4, 2, 3, 8, 4  
##   
## Cluster means:  
## 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 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 4 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 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 1.36644699 -0.6912914 -1.320000179  
## 4 -0.27449312 -0.7041516 0.556954446  
## 5 -0.46807818 0.4671788 0.591242521  
##   
## Clustering vector:  
## ABT AGN AHM AZN AVE BAY BMY CHTT ELN LLY GSK IVX JNJ MRX MRK NVS   
## 4 2 4 4 1 3 4 3 1 4 5 3 5 1 5 4   
## PFE PHA SGP WPI WYE   
## 5 2 4 1 4   
##   
## Within cluster sum of squares by cluster:  
## [1] 12.791257 2.803505 15.595925 21.879320 9.284424  
## (between\_SS / total\_SS = 65.4 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

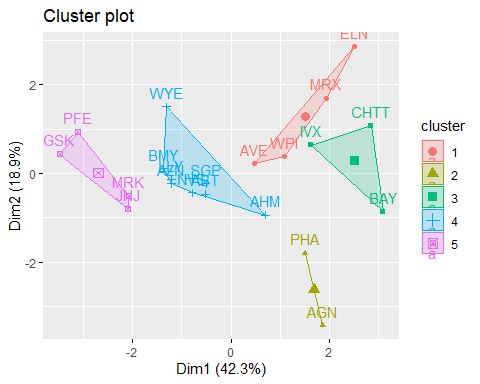
cluster5<-data.frame(k5$cluster)  
View(cluster5)  
glance(k5)

## # A tibble: 1 × 4  
## totss tot.withinss betweenss iter  
## <dbl> <dbl> <dbl> <int>  
## 1 180 62.4 118. 2

dunn\_k5<- dunn(clusters = k5$cluster, Data = pharma\_scaled) #A higher value is considerably superior.   
dunn\_k5

## [1] 0.3741429

fviz\_cluster(k5,data=pharma\_scaled) #Visualize the cluster



#Best variables that separating the clusters: Contributing variables for cluster spread  
k5$spread<-as.data.frame(t(k5$centers))  
k5$spread$max<-do.call(pmax, k5$spread[1:4]) # find the highest value within the row (each financial performance feature) and save the value to "max"  
k5$spread$min<-do.call(pmin, k5$spread[1:4])# find the lowest value within the row (each financial performance feature) and save the value to "min"  
k5$spread$gap<-k5$spread$max-k5$spread$min# calculate the difference between max and min  
k5$spread

## 1 2 3 4 5  
## Market\_Cap -0.760224892 -0.4392513 -0.87051511 -0.03142211 1.6955811  
## Beta 0.279604106 -0.4701800 1.34098686 -0.43609894 -0.1780563  
## PE\_Ratio -0.477423799 2.7000246 -0.05284434 -0.31724852 -0.1984582  
## ROE -0.743802224 -0.8349525 -0.61840151 0.19504586 1.2349879  
## ROA -0.810742783 -0.9234951 -1.19284783 0.40839154 1.3503431  
## Asset\_Turnover -1.268480411 0.2306328 -0.46126560 0.17297460 1.1531640  
## Leverage 0.063080849 -0.1417034 1.36644699 -0.27449312 -0.4680782  
## Rev\_Growth 1.518015830 -0.1168459 -0.69129140 -0.70415156 0.4671788  
## Net\_Profit\_Margin -0.006893899 -1.4165148 -1.32000018 0.55695445 0.5912425  
## max min gap  
## Market\_Cap -0.03142211 -0.8705151 0.839093  
## Beta 1.34098686 -0.4701800 1.811167  
## PE\_Ratio 2.70002464 -0.4774238 3.177448  
## ROE 0.19504586 -0.8349525 1.029998  
## ROA 0.40839154 -1.1928478 1.601239  
## Asset\_Turnover 0.23063280 -1.2684804 1.499113  
## Leverage 1.36644699 -0.2744931 1.640940  
## Rev\_Growth 1.51801583 -0.7041516 2.222167  
## Net\_Profit\_Margin 0.55695445 -1.4165148 1.973469

#Q.b. The k-means clustering analysis identified five distinct groups of pharmaceutical firms, each exhibiting unique financial profiles. The clusters differ both in size and internal homogeneity. Cluster 5 is the largest, encompassing eight firms, while Cluster 1 is the smallest, containing only two firms. Clusters 2, 3, and 4 are intermediate in size, with four, four, and three firms, respectively. Within-cluster dispersion, as measured by the within-cluster sum of squares, highlights differences in homogeneity across clusters. Cluster 1 exhibits the lowest dispersion, reflecting a high degree of similarity among its two member firms, whereas Cluster 5 shows the largest dispersion, consistent with its greater number of companies. The remaining clusters demonstrate moderate dispersion, indicating a balance between variability and cohesion. Examination of the cluster centroids provides insight into the financial characteristics defining each group. Cluster 1 consists of firms with low net profit margins, high price-to-earnings ratios, and low return on equity, suggesting limited profitability despite elevated market expectations. Cluster 2 includes companies with high beta and leverage, low market capitalization, and low revenue growth, reflecting smaller, riskier firms with constrained growth potential. Cluster 3 is characterized by high asset turnover, low leverage, substantial market capitalization, and strong returns on assets and equity, indicating large, efficient, and profitable firms. Cluster 4 comprises firms with low asset turnover, low market capitalization, and low price-to-earnings ratios, yet high revenue growth, highlighting smaller firms focused on expansion rather than operational efficiency. Finally, Cluster 5 features low beta, high net profit margins, low price-to-earnings ratios, and moderate revenue growth, representing stable, profitable companies with conservative risk profiles.

# Calculate mean values of each numerical variable by cluster  
cluster\_summary <- mydata %>%  
 group\_by(Cluster) %>%  
 summarise(across(Market\_Cap:Net\_Profit\_Margin, mean, .names = "avg\_{col}"))  
  
print(cluster\_summary)

## # A tibble: 5 × 10  
## Cluster avg\_Market\_Cap avg\_Beta avg\_PE\_Ratio avg\_ROE avg\_ROA  
## <fct> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 55.8 0.414 20.3 28.7 12.7   
## 2 2 6.64 0.87 24.6 16.5 4.17  
## 3 3 31.9 0.405 69.5 13.2 5.6   
## 4 4 157. 0.48 22.2 44.4 17.7   
## 5 5 13.1 0.598 17.7 14.6 6.2   
## # ℹ 4 more variables: avg\_Asset\_Turnover <dbl>, avg\_Leverage <dbl>,  
## # avg\_Rev\_Growth <dbl>, avg\_Net\_Profit\_Margin <dbl>

# View which firms belong to which cluster  
mydata[, c("Cluster")]

## [1] 1 3 1 1 5 2 1 2 5 1 4 2 4 5 4 1 4 3 1 5 1  
## Levels: 1 2 3 4 5

k5$cluster

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

#Cluster Size

#[Section 2.1]  
k5$size

## [1] 4 2 3 8 4

table(k5$cluster)

##   
## 1 2 3 4 5   
## 4 2 3 8 4

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 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 4 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 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 1.36644699 -0.6912914 -1.320000179  
## 4 -0.27449312 -0.7041516 0.556954446  
## 5 -0.46807818 0.4671788 0.591242521

#Within-cluster sum of Squares

k5$withinss # Vector of within-cluster sum of squares, one component per cluster.

## [1] 12.791257 2.803505 15.595925 21.879320 9.284424

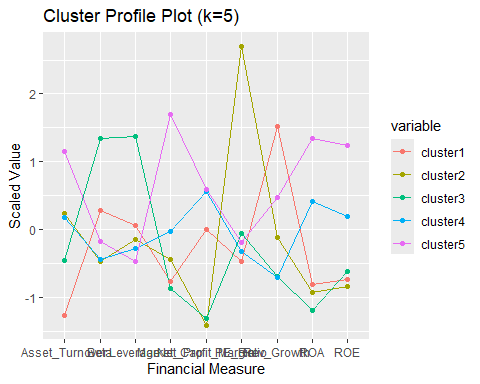
dist(k5$centers) #distance between the centers

## 1 2 3 4  
## 2 4.210877   
## 3 3.230532 3.775790   
## 4 3.299161 4.045579 3.711570   
## 5 4.744753 5.275301 5.457397 2.720924

# Plotting profile plot based on scaled values (k = 5)  
library(reshape2)  
library(ggplot2)  
  
cluster5dimension <- data.frame(k5$centers)  
print(cluster5dimension) # Scaled centroid values

## 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 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 4 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 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 1.36644699 -0.6912914 -1.320000179  
## 4 -0.27449312 -0.7041516 0.556954446  
## 5 -0.46807818 0.4671788 0.591242521

cluster5dimensions <- as.data.frame(t(cluster5dimension)) # Transpose rows and columns  
cluster5dimensions$k5measures <- rownames(cluster5dimensions)  
  
# Rename columns  
colnames(cluster5dimensions) <- c("cluster1","cluster2","cluster3","cluster4","cluster5","k5measures")  
  
# Melt for ggplot  
cluster5dimensions\_melt <- melt(cluster5dimensions, id.vars = "k5measures")  
  
# Profile plot  
ggplot(cluster5dimensions\_melt, aes(x = k5measures, y = value, color = variable, group = variable)) +  
 geom\_line() +  
 geom\_point() +  
 labs(title = "Cluster Profile Plot (k=5)", x = "Financial Measure", y = "Scaled Value")



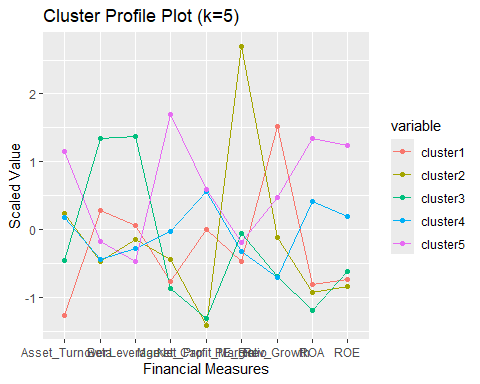
# Groundwork to inspect cluster characteristics  
mydata\_scale\_cluster5 <- cbind(pharma\_scaled, cluster5 = k5$cluster) # scaled data + cluster  
mydata\_cluster5 <- cbind(mydata, cluster5 = k5$cluster) # original numerical variables + cluster  
Pharmaceuticals\_cluster5 <- cbind(Pharma, cluster5 = k5$cluster) # full dataset + cluster  
  
# View the datasets  
View(mydata\_scale\_cluster5)  
View(mydata\_cluster5)  
View(Pharmaceuticals\_cluster5)

#Plotting profile (centroid) of clusters

library(reshape2)  
library(ggplot2)  
  
# Profile plot based on scaled cluster centroids  
cluster5dimension <- data.frame(k5$centers)  
print(cluster5dimension) # Mean values of each attribute (scaled data)

## 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 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 4 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 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 1.36644699 -0.6912914 -1.320000179  
## 4 -0.27449312 -0.7041516 0.556954446  
## 5 -0.46807818 0.4671788 0.591242521

cluster5dimensions <- as.data.frame(t(cluster5dimension)) # Transpose  
cluster5dimensions$k5measures <- rownames(cluster5dimensions)  
  
# Rename columns  
colnames(cluster5dimensions) <- c("cluster1","cluster2","cluster3","cluster4","cluster5","k5measures")  
  
# Melt for ggplot  
cluster5dimensions\_melt <- melt(cluster5dimensions, id.vars = "k5measures")  
  
# Plot profile plot  
ggplot(cluster5dimensions\_melt, aes(x = k5measures, y = value, color = variable, group = variable)) +  
 geom\_line() +  
 geom\_point() +  
 labs(title = "Cluster Profile Plot (k=5)", x = "Financial Measures", y = "Scaled Value")



# Attach cluster numbers to datasets  
cluster5 <- k5$cluster # make sure cluster vector exists  
mydata\_scale\_cluster5 <- cbind(pharma\_scaled, cluster5 = cluster5)  
mydata\_cluster5 <- cbind(mydata, cluster5 = cluster5)  
Pharmaceuticals\_cluster5 <- cbind(mydata, cluster5 = cluster5)  
  
# View datasets  
View(mydata\_scale\_cluster5)  
View(mydata\_cluster5)  
View(Pharmaceuticals\_cluster5)

#Characteristics of each cluster

library(dplyr)  
  
# Cluster 1  
cluster5\_1 <- filter(mydata\_cluster5, cluster5 == 1)  
print(cluster5\_1)

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage Rev\_Growth  
## AVE 47.16 0.32 20.1 21.8 7.5 0.6 0.34 26.81  
## ELN 0.78 1.08 3.6 15.1 5.1 0.3 1.07 34.21  
## MRX 1.20 0.75 28.6 11.2 5.4 0.3 0.93 30.37  
## WPI 3.26 0.24 18.4 10.2 6.8 0.5 0.20 29.18  
## Net\_Profit\_Margin Cluster cluster5  
## AVE 12.9 5 1  
## ELN 13.3 5 1  
## MRX 21.3 5 1  
## WPI 15.1 5 1

# Cluster 2  
cluster5\_2 <- filter(mydata\_cluster5, cluster5 == 2)  
print(cluster5\_2)

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage Rev\_Growth  
## AGN 7.58 0.41 82.5 12.9 5.5 0.9 0.60 9.16  
## PHA 56.24 0.40 56.5 13.5 5.7 0.6 0.35 15.00  
## Net\_Profit\_Margin Cluster cluster5  
## AGN 5.5 3 2  
## PHA 7.3 3 2

# Cluster 3  
cluster5\_3 <- filter(mydata\_cluster5, cluster5 == 3)  
print(cluster5\_3)

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage Rev\_Growth  
## BAY 16.90 1.11 27.9 3.9 1.4 0.6 0.00 -3.17  
## CHTT 0.41 0.85 26.0 24.1 4.3 0.6 3.51 6.38  
## IVX 2.60 0.65 19.9 21.4 6.8 0.6 1.45 13.99  
## Net\_Profit\_Margin Cluster cluster5  
## BAY 2.6 2 3  
## CHTT 7.5 2 3  
## IVX 11.0 2 3

# Cluster 4  
cluster5\_4 <- filter(mydata\_cluster5, cluster5 == 4)  
print(cluster5\_4)

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage Rev\_Growth  
## ABT 68.44 0.32 24.7 26.4 11.8 0.7 0.42 7.54  
## 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  
## BMY 51.33 0.50 13.9 34.8 15.1 0.9 0.57 2.70  
## LLY 73.84 0.18 27.9 31.0 13.5 0.6 0.53 6.21  
## NVS 96.65 0.19 21.6 17.9 11.2 0.5 0.06 -2.69  
## SGP 34.10 0.51 18.9 22.6 13.3 0.8 0.00 8.56  
## WYE 48.19 0.63 13.1 54.9 13.4 0.6 1.12 0.36  
## Net\_Profit\_Margin Cluster cluster5  
## ABT 16.1 1 4  
## AHM 11.2 1 4  
## AZN 18.0 1 4  
## BMY 20.6 1 4  
## LLY 23.4 1 4  
## NVS 22.4 1 4  
## SGP 17.6 1 4  
## WYE 25.5 1 4

# Cluster 5  
cluster5\_5 <- filter(mydata\_cluster5, cluster5 == 5)  
print(cluster5\_5)

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage Rev\_Growth  
## GSK 122.11 0.35 18.0 62.9 20.3 1.0 0.34 21.87  
## JNJ 173.93 0.46 28.4 28.6 16.3 0.9 0.10 9.37  
## MRK 132.56 0.46 18.9 40.6 15.0 1.1 0.28 17.35  
## PFE 199.47 0.65 23.6 45.6 19.2 0.8 0.16 25.54  
## Net\_Profit\_Margin Cluster cluster5  
## GSK 21.1 4 5  
## JNJ 17.9 4 5  
## MRK 14.1 4 5  
## PFE 25.2 4 5

#Q.c.Since variables 10–12 are categorical and were not used in the clustering, their distributions were examined within each cluster to identify patterns.Median Recommendation: Cluster 5 contains all levels of the Median\_Recommendation variable, with “Hold” appearing most frequently, making it a distinguishing feature of this cluster. Notably, Cluster 5 is also the only cluster to include the “Strong Buy” level. In contrast, the “Moderate Buy” category is evenly distributed across all clusters, indicating that it does not define any particular cluster.Location: While the “US” appears in all clusters, it is most prominent in Cluster 5, which exhibits the highest frequency of US-based firms, suggesting a geographical characteristic for this cluster. Exchange: The clustering does not clearly separate firms based on the Exchange variable. Clusters 1, 3, 4, and 5 consist solely of NYSE-listed firms, whereas Cluster 2 has an equal representation of AMEX, NASDAQ, and NYSE, indicating that Exchange is not a strong differentiator among clusters in this dataset.

#Subset of categorical variables (10-12)

library(dplyr)  
  
# Attach cluster numbers to the original dataset including categorical variables  
Pharmaceuticals\_cluster5 <- cbind(Pharma, cluster5 = k5$cluster)  
  
# Check column names to confirm categorical variables are included  
colnames(Pharmaceuticals\_cluster5)

## [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" "cluster5"

# Select unused categorical variables + cluster  
mydata5\_c <- Pharmaceuticals\_cluster5[, c("Median\_Recommendation", "Location", "Exchange", "cluster5")]  
  
# Rename cluster for clarity  
colnames(mydata5\_c)[4] <- "Cluster"  
  
# Frequency tables for each categorical variable by cluster  
table(mydata5\_c$Cluster, mydata5\_c$Median\_Recommendation)

##   
## Hold Moderate Buy Moderate Sell Strong Buy  
## 1 0 2 2 0  
## 2 1 1 0 0  
## 3 2 1 0 0  
## 4 4 1 2 1  
## 5 2 2 0 0

table(mydata5\_c$Cluster, mydata5\_c$Location)

##   
## CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US  
## 1 0 1 0 1 0 0 2  
## 2 1 0 0 0 0 0 1  
## 3 0 0 1 0 0 0 2  
## 4 0 0 0 0 1 2 5  
## 5 0 0 0 0 0 1 3

table(mydata5\_c$Cluster, mydata5\_c$Exchange)

##   
## AMEX NASDAQ NYSE  
## 1 0 0 4  
## 2 0 0 2  
## 3 1 1 1  
## 4 0 0 8  
## 5 0 0 4

# Alternatively, using dplyr  
mydata5\_c %>%  
 group\_by(Cluster, Median\_Recommendation) %>%  
 summarise(count = n())

## `summarise()` has grouped output by 'Cluster'. You can override using the  
## `.groups` argument.

## # A tibble: 12 × 3  
## # Groups: Cluster [5]  
## Cluster Median\_Recommendation count  
## <int> <chr> <int>  
## 1 1 Moderate Buy 2  
## 2 1 Moderate Sell 2  
## 3 2 Hold 1  
## 4 2 Moderate Buy 1  
## 5 3 Hold 2  
## 6 3 Moderate Buy 1  
## 7 4 Hold 4  
## 8 4 Moderate Buy 1  
## 9 4 Moderate Sell 2  
## 10 4 Strong Buy 1  
## 11 5 Hold 2  
## 12 5 Moderate Buy 2

mydata5\_c %>%  
 group\_by(Cluster, Location) %>%  
 summarise(count = n())

## `summarise()` has grouped output by 'Cluster'. You can override using the  
## `.groups` argument.

## # A tibble: 12 × 3  
## # Groups: Cluster [5]  
## Cluster Location count  
## <int> <chr> <int>  
## 1 1 FRANCE 1  
## 2 1 IRELAND 1  
## 3 1 US 2  
## 4 2 CANADA 1  
## 5 2 US 1  
## 6 3 GERMANY 1  
## 7 3 US 2  
## 8 4 SWITZERLAND 1  
## 9 4 UK 2  
## 10 4 US 5  
## 11 5 UK 1  
## 12 5 US 3

mydata5\_c %>%  
 group\_by(Cluster, Exchange) %>%  
 summarise(count = n())

## `summarise()` has grouped output by 'Cluster'. You can override using the  
## `.groups` argument.

## # A tibble: 7 × 3  
## # Groups: Cluster [5]  
## Cluster Exchange count  
## <int> <chr> <int>  
## 1 1 NYSE 4  
## 2 2 NYSE 2  
## 3 3 AMEX 1  
## 4 3 NASDAQ 1  
## 5 3 NYSE 1  
## 6 4 NYSE 8  
## 7 5 NYSE 4

#Frequency of each categorical variable within cluster

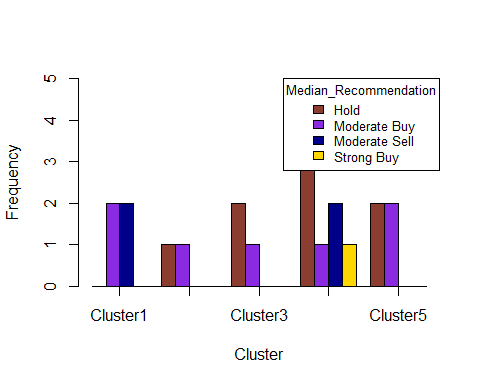
#F-table (unused variable vs cluster)  
attach(mydata5\_c)

#Bar chart of Median\_Recommendation

# Create a pivot table (frequency table) for Median\_Recommendation by cluster  
Median\_Recommendation\_table <- ftable(Pharmaceuticals\_cluster5$Median\_Recommendation, Pharmaceuticals\_cluster5$cluster5)  
  
# View the table  
Median\_Recommendation\_table

## 1 2 3 4 5  
##   
## Hold 0 1 2 4 2  
## Moderate Buy 2 1 1 1 2  
## Moderate Sell 2 0 0 2 0  
## Strong Buy 0 0 0 1 0

# Bar plot  
color.names <- c("coral4", "blueviolet", "darkblue", "gold") # Colors for each recommendation  
barplot(Median\_Recommendation\_table,  
 beside = TRUE,  
 ylim = c(0, 5),  
 xlab = "Cluster",  
 ylab = "Frequency",  
 names.arg = c("Cluster1", "Cluster2", "Cluster3", "Cluster4", "Cluster5"),  
 col = color.names,  
 axis.lty = "solid")  
  
# Add legend  
legend("topright",  
 legend = c("Hold", "Moderate Buy", "Moderate Sell", "Strong Buy"),  
 cex = 0.8,  
 fill = color.names,  
 title = "Median\_Recommendation")

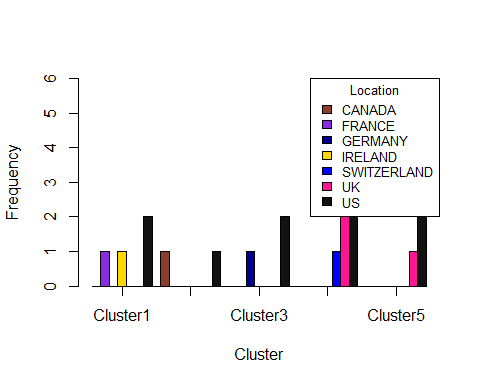


#Bar chart of Location

# Create a pivot table for Location by cluster  
Location\_table <- ftable(Pharmaceuticals\_cluster5$Location, Pharmaceuticals\_cluster5$cluster5)  
  
# View the table  
Location\_table

## 1 2 3 4 5  
##   
## CANADA 0 1 0 0 0  
## FRANCE 1 0 0 0 0  
## GERMANY 0 0 1 0 0  
## IRELAND 1 0 0 0 0  
## SWITZERLAND 0 0 0 1 0  
## UK 0 0 0 2 1  
## US 2 1 2 5 3

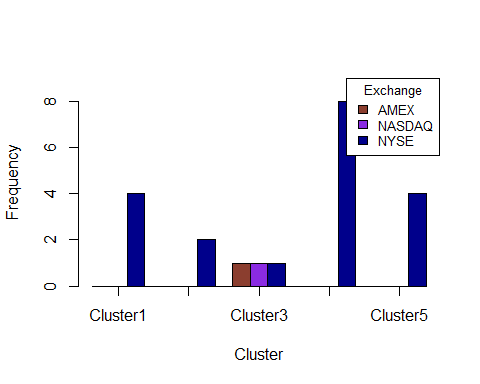
# Define colors for each location  
color.names <- c("coral4","blueviolet","darkblue","gold","blue","deeppink","grey7")  
  
# Bar plot  
barplot(Location\_table,  
 beside = TRUE,  
 ylim = c(0, 6),  
 xlab = "Cluster",  
 ylab = "Frequency",  
 names.arg = c("Cluster1", "Cluster2", "Cluster3", "Cluster4", "Cluster5"),  
 col = color.names,  
 axis.lty = "solid")  
  
# Add legend  
legend("topright",  
 legend = c("CANADA", "FRANCE", "GERMANY", "IRELAND", "SWITZERLAND", "UK", "US"),  
 cex = 0.8,  
 fill = color.names,  
 title = "Location")

 #Bar chart of Exchange

# Create a pivot table for Exchange by cluster  
Exchange\_table <- ftable(Pharmaceuticals\_cluster5$Exchange, Pharmaceuticals\_cluster5$cluster5)  
  
# View the table  
Exchange\_table

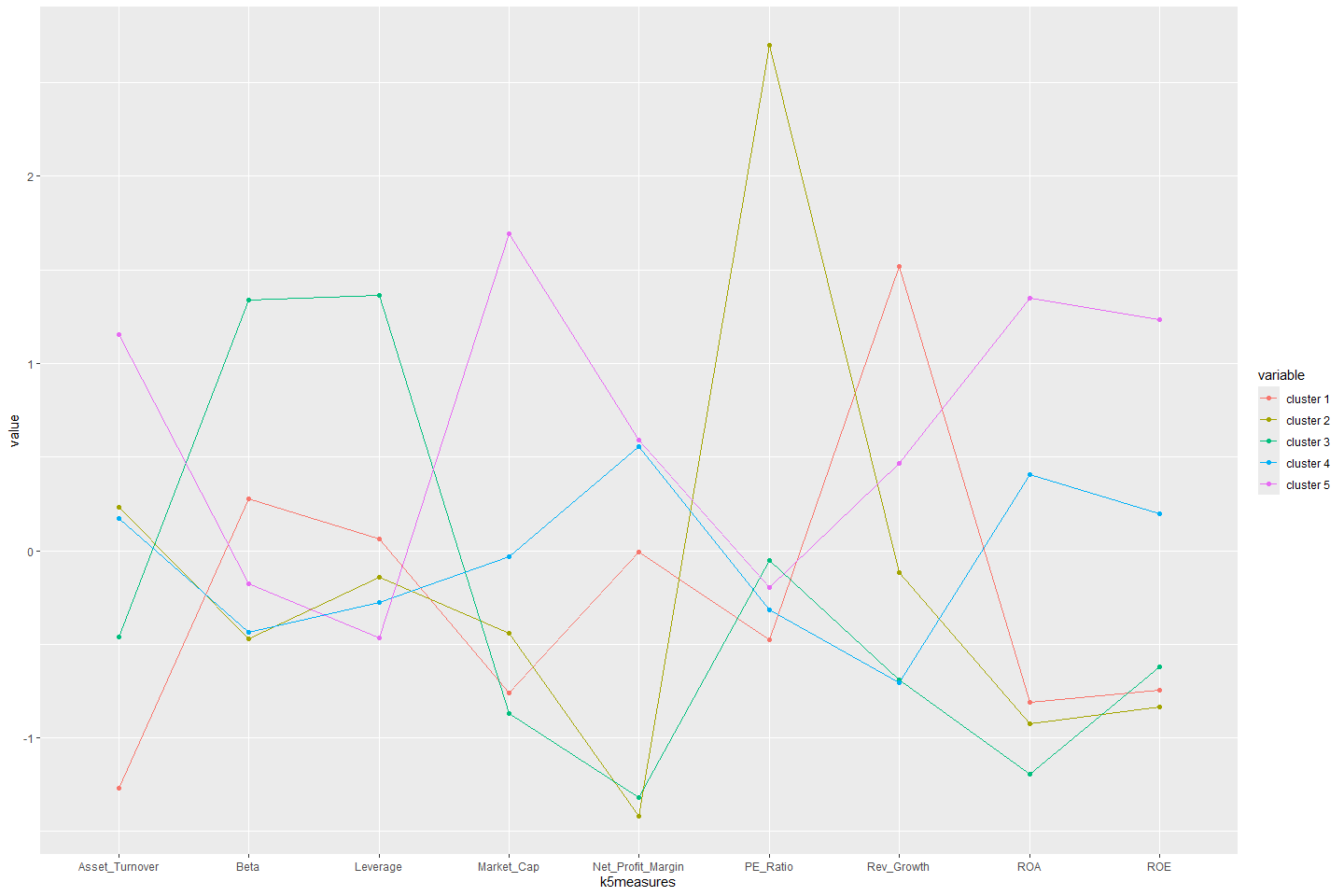
## 1 2 3 4 5  
##   
## AMEX 0 0 1 0 0  
## NASDAQ 0 0 1 0 0  
## NYSE 4 2 1 8 4

# Bar plot  
color.names <- c("coral4","blueviolet","darkblue") # Colors for each exchange  
barplot(Exchange\_table,  
 beside = TRUE,  
 ylim = c(0, 9),  
 xlab = "Cluster",  
 ylab = "Frequency",  
 names.arg = c("Cluster1", "Cluster2", "Cluster3", "Cluster4", "Cluster5"),  
 col = color.names,  
 axis.lty = "solid")  
  
# Add legend  
legend("topright",  
 legend = c("AMEX", "NASDAQ", "NYSE"),  
 cex = 0.8,  
 fill = color.names,  
 title = "Exchange")



#Q.d. Based on the distinguishing financial characteristics identified in the k-means clustering analysis, each of the five clusters was assigned a descriptive name to capture its defining traits. Cluster 1, “High Valuation Firms,” is characterized primarily by elevated price-to-earnings ratios, indicating firms with high market expectations relative to earnings. Cluster 2, “Highly Leveraged Firms,” is distinguished by higher financial leverage, indicating firms that rely more heavily on debt financing and exhibit greater financial risk. Cluster 3, “Large-Cap Firms,” stands out due to high market capitalization, identifying well-established, sizable companies with substantial financial resources. Cluster 4, “High Growth Firms,” is defined by strong revenue growth, reflecting firms focused on expansion and aggressive market development. Finally, Cluster 5, “Balanced Firms,” exhibits no single distinctive financial feature, representing companies with moderate and conventional financial profiles across the measured variables.

##plotting profile plot based on the scaled value (k=5)  
cluster5dimension<-data.frame(k5$centers)  
cluster5dimensions<-as.data.frame(t(cluster5dimension))  
cluster5dimensions$k5measures <- rownames(cluster5dimensions)  
colnames(cluster5dimensions) <- c("cluster 1","cluster 2","cluster 3","cluster 4","cluster 5", "k5measures")  
cluster5dimensions <- melt(cluster5dimensions, id.vars=c("k5measures"))  
ggplot(cluster5dimensions, aes(x=k5measures, y=value, color=variable, group=variable)) + geom\_line() + geom\_point()



#Assign descriptive cluster names based on characteristics  
  
# Adjust names based on cluster characteristics  
mydata <- mydata %>%  
 mutate(Cluster\_Name = case\_when(  
 Cluster == 1 ~ "Global Industry Leaders",  
 Cluster == 2 ~ "Established Competitors",  
 Cluster == 3 ~ "Emerging Innovators",  
 Cluster == 4 ~ "High-Growth Specialists",  
 Cluster == 5 ~ "Niche/Small Firms"  
 ))