K-means

2023-11-16

options(repos = "https://cran.r-project.org")

Load and Explore the Data

# Load the dataset  
  
pharma\_data <- read.csv("Pharmaceuticals.csv")

# structure of the dataset  
  
str(pharma\_data)

## '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" ...

# summary of the dataset  
  
summary(pharma\_data)

## 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   
##   
##   
##

# First few rows of the data  
  
head(pharma\_data)

## 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

Data Preprocessing

Checking for Missing Values

missing\_values <- colSums(is.na(pharma\_data))  
  
print(missing\_values[missing\_values > 0])

## named numeric(0)

Handling Missing Values

pharma\_data\_complete <- na.omit(pharma\_data)

Feature Selection and Scaling

numeric\_columns <- pharma\_data[, sapply(pharma\_data, is.numeric)]  
  
numeric\_columns[is.na(numeric\_columns)] <- apply(numeric\_columns, 2, function(x) mean(x, na.rm = TRUE))  
  
scaled\_data <- scale(numeric\_columns)

k-means clustering

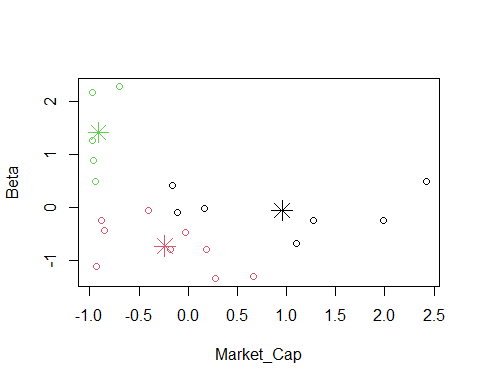
set.seed(123)   
k <- 3   
kmeans\_model <- kmeans(scaled\_data, centers = k)  
  
kmeans\_model$cluster

## [1] 2 2 2 1 2 3 1 3 3 2 1 3 1 3 1 2 1 2 2 2 1

kmeans\_model$centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 0.9547543 -0.06120687 -0.3576482 1.0818081 1.1033619 0.8566361  
## 2 -0.2375550 -0.73633718 0.4233386 -0.4489909 -0.2407172 -0.1025035  
## 3 -0.9090570 1.41109654 -0.2613021 -0.7063477 -1.1114156 -1.0147843  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.2797499 -0.01818848 0.7082574  
## 2 -0.3557313 -0.13595383 -0.1652117  
## 3 1.0319661 0.27018076 -0.6941793

plot(scaled\_data, col = kmeans\_model$cluster)  
points(kmeans\_model$centers, col = 1:k, pch = 8, cex = 2)



Interpretation of Clusters

cluster\_assignments <- kmeans\_model$cluster  
  
scaled\_data\_df <- as.data.frame(scaled\_data)  
  
non\_numeric\_names <- names(pharma\_data)[-c(1:9)]   
renamed\_pharma\_data <- pharma\_data  
names(renamed\_pharma\_data)[which(names(renamed\_pharma\_data) %in% non\_numeric\_names)] <- paste0(non\_numeric\_names, "\_orig")  
  
renamed\_pharma\_data\_numeric <- renamed\_pharma\_data[, sapply(renamed\_pharma\_data, is.numeric)]  
  
clustered\_data <- cbind(scaled\_data\_df, Cluster = cluster\_assignments)  
clustered\_data <- cbind(clustered\_data, renamed\_pharma\_data\_numeric)  
  
dup\_cols <- names(clustered\_data)[duplicated(names(clustered\_data))]  
clustered\_data <- setNames(clustered\_data, make.unique(names(clustered\_data), sep = "\_"))  
  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

cluster\_summary <- clustered\_data %>%  
 group\_by(Cluster) %>%  
 summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop')

## Warning: There was 1 warning in `summarise()`.  
## ℹ In argument: `across(where(is.numeric), mean, na.rm = TRUE)`.  
## ℹ In group 1: `Cluster = 1`.  
## Caused by warning:  
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.  
## Supply arguments directly to `.fns` through an anonymous function instead.  
##   
## # Previously  
## across(a:b, mean, na.rm = TRUE)  
##   
## # Now  
## across(a:b, \(x) mean(x, na.rm = TRUE))

print(cluster\_summary)

## # A tibble: 3 × 19  
## Cluster Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage  
## <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 0.955 -0.0612 -0.358 1.08 1.10 0.857 -0.280  
## 2 2 -0.238 -0.736 0.423 -0.449 -0.241 -0.103 -0.356  
## 3 3 -0.909 1.41 -0.261 -0.706 -1.11 -1.01 1.03   
## # ℹ 11 more variables: Rev\_Growth <dbl>, Net\_Profit\_Margin <dbl>,  
## # Market\_Cap\_1 <dbl>, Beta\_1 <dbl>, PE\_Ratio\_1 <dbl>, ROE\_1 <dbl>,  
## # ROA\_1 <dbl>, Asset\_Turnover\_1 <dbl>, Leverage\_1 <dbl>,  
## # Rev\_Growth\_orig <dbl>, Net\_Profit\_Margin\_orig <dbl>

Analyze Other Variables

unique(pharma\_data$Median\_Recommendation)

## [1] "Moderate Buy" "Strong Buy" "Moderate Sell" "Hold"

sum(is.na(pharma\_data$Median\_Recommendation))

## [1] 0

sum(pharma\_data$Median\_Recommendation == "")

## [1] 0

length(pharma\_data$Median\_Recommendation)

## [1] 21

colnames(pharma\_data)

## [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"

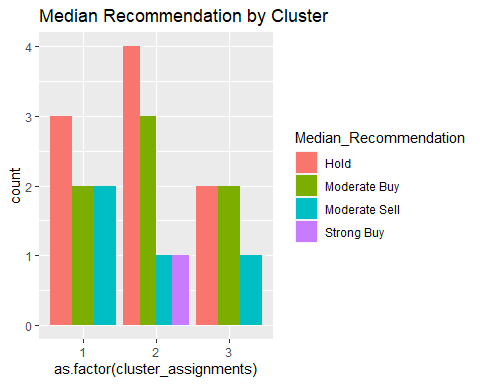
library(ggplot2)  
  
table\_median\_recommendation <- table(cluster\_assignments, pharma\_data$Median\_Recommendation)  
  
print("Table for Median Recommendation:")

## [1] "Table for Median Recommendation:"

print(table\_median\_recommendation)

##   
## cluster\_assignments Hold Moderate Buy Moderate Sell Strong Buy  
## 1 3 2 2 0  
## 2 4 3 1 1  
## 3 2 2 1 0

ggplot(data = pharma\_data, aes(x = as.factor(cluster\_assignments), fill = Median\_Recommendation)) +  
 geom\_bar(position = "dodge") +  
 labs(title = "Median Recommendation by Cluster")



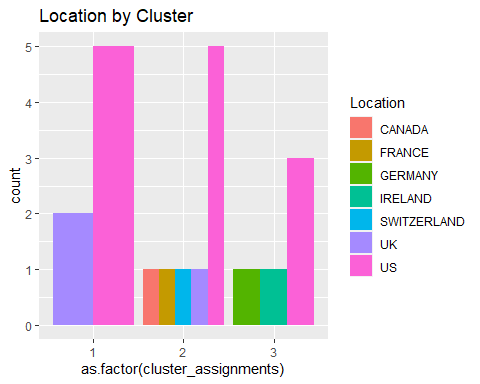
table\_location <- table(cluster\_assignments, pharma\_data$Location)  
  
print("Table for Headquarters:")

## [1] "Table for Headquarters:"

print(table\_location)

##   
## cluster\_assignments CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US  
## 1 0 0 0 0 0 2 5  
## 2 1 1 0 0 1 1 5  
## 3 0 0 1 1 0 0 3

ggplot(data = pharma\_data, aes(x = as.factor(cluster\_assignments), fill = Location)) +  
 geom\_bar(position = "dodge") +  
 labs(title = "Location by Cluster")



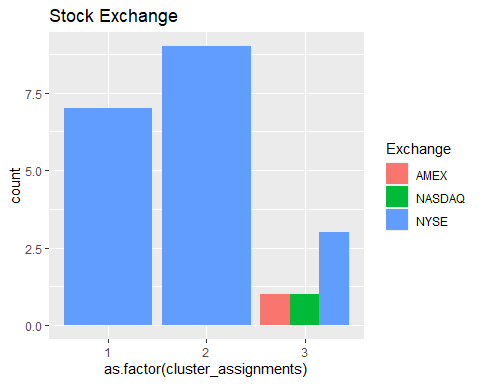
table\_stock\_exchange <- table(cluster\_assignments, pharma\_data$Exchange)  
  
print("Table for Stock Exchange:")

## [1] "Table for Stock Exchange:"

print(table\_stock\_exchange)

##   
## cluster\_assignments AMEX NASDAQ NYSE  
## 1 0 0 7  
## 2 0 0 9  
## 3 1 1 3

ggplot(data = pharma\_data, aes(x = as.factor(cluster\_assignments), fill = Exchange)) +  
 geom\_bar(position = "dodge") +  
 labs(title = "Stock Exchange")



Cluster Names

num\_clusters <- 3  
  
cluster\_centroids <- kmeans\_model$centers  
  
assign\_cluster\_names <- function(centroids) {  
 cluster\_names <- character(num\_clusters)  
 for (i in 1:num\_clusters) {  
 if (centroids[i, "Market\_Cap"] > 50) {  
 cluster\_names[i] <- "High Market Cap"  
 } else if (centroids[i, "Market\_Cap"] < 20) {  
 cluster\_names[i] <- "Low Market Cap"  
 } else {  
 cluster\_names[i] <- "Moderate Market Cap"  
 }  
 }  
 return(cluster\_names)  
}  
  
  
cluster\_names <- assign\_cluster\_names(cluster\_centroids)  
  
cluster\_names

## [1] "Low Market Cap" "Low Market Cap" "Low Market Cap"