Assignment 4 - K Means Fundamentals of Machine Learning

2025-06-16

#Assignment 4 - K-Means Fundamentals of Machine Learning

##by David Wilkinson

#Load Libraries and Pharmaceutical Data

#data manipulation  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── 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

#clustering algorithms & visualization  
library(factoextra)

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

#load Pharmaceuticals data to dataframe  
pharma.df <- read.csv("C:\\Users\\david\\OneDrive\\Documents\\Kent State University MSBA\\Fundamentals of Machine Learning June 2025\\Pharmaceuticals.csv")  
pharma.df <- data.frame(pharma.df)  
#top 5 rows of pharma dataframe  
head(pharma.df)

## 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

#Select only numerical columns

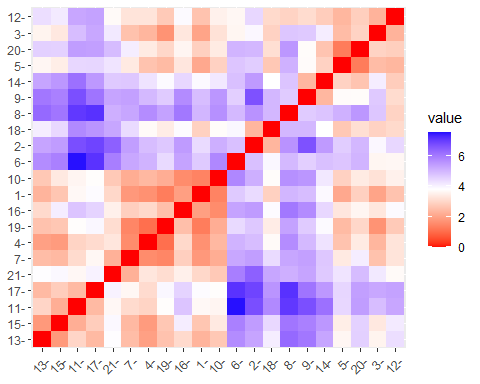
pharma\_numeric <- pharma.df[,c(3:11)]  
head(pharma\_numeric)

## 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

##Question 1

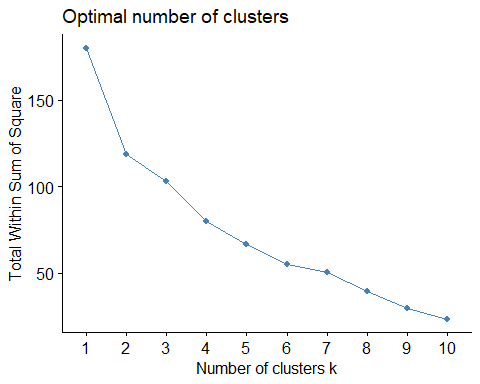
#Cluster 21 Companies Using Numerical Variables

pharma\_numeric.norm <- scale(pharma\_numeric)  
distance <- get\_dist(pharma\_numeric.norm)  
fviz\_dist(distance)

 The values in the above distance chart are properly scaled. As a result, they produce a symmetrical image representing the scaled distances between each numeric variable. The red diagonal line represents each points distance to itself, which is 0. They are mirror images of one another because the distance between 16 to 4 is the same value as the distance from 4 to 16. As a result, we technically only need the values on only one side of the diagonal line of 0 values.

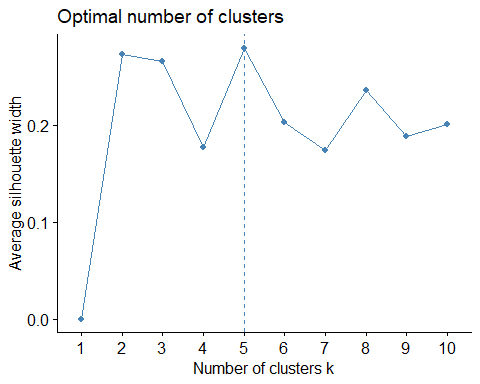
#Choosing options for K to test - wss method

#use elbow method to estimate the best number of clusters   
set.seed(89)  
fviz\_nbclust(pharma\_numeric.norm, kmeans, method = "wss")

 This method shows that the best value for k (barring any domain specific knowledge) is 5 or 6.

#Choosing options for K to test - silhouette method

#use the silhouette method to estimate the number of clusters to test   
set.seed(89)  
fviz\_nbclust(pharma\_numeric.norm, kmeans, method = "silhouette")

 The silhouette method shows that the optimal number of clusters (barring any domain specific knowledge) is 5.

#Test Each Number of Clusters - K = 5

set.seed(66)  
k5 <- kmeans(pharma\_numeric.norm, centers = 5, nstart = 100) # k = 5, number of restarts = 100  
# 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 -0.87051511 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

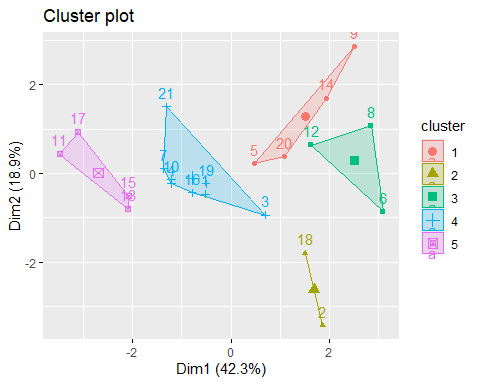
k5$size # Number of records in each cluster

## [1] 4 2 3 8 4

k5$cluster[21] # Identify the cluster of the 21 observation as an example

## [1] 4

fviz\_cluster(k5, data = pharma\_numeric.norm) # Visualize the output



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

#Calculate Distance Between Clusters for K =5

k5$withinss

## [1] 12.791257 2.803505 15.595925 21.879320 9.284424

dist(k5$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

#Test Each Number of Clusters - K = 6

set.seed(66)  
k6 <- kmeans(pharma\_numeric.norm, centers = 6, nstart = 100) # k = 6, number of restarts = 100  
# Visualize the output  
k6$centers # output the centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.43925134 -0.4701800 2.70002464 -0.8349525 -0.9234951 0.2306328  
## 2 -0.79605926 0.3205014 -0.45014035 -0.6533148 -0.7881923 -1.1070374  
## 3 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 4 1.69558112 -0.1780563 -0.19845823 1.2349879 1.3503431 1.1531640  
## 5 -0.97676686 1.2630872 0.03299122 -0.1123792 -1.1677918 -0.4612656  
## 6 -0.69538175 2.2757827 0.14948233 -1.4514600 -1.7127612 -0.4612656  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.1417034 -0.1168459 -1.4165148  
## 2 0.2717048 1.2256188 -0.1486179  
## 3 -0.2744931 -0.7041516 0.5569544  
## 4 -0.4680782 0.4671788 0.5912425  
## 5 3.7427970 -0.6327607 -1.2488842  
## 6 -0.7496565 -1.4971443 -1.9956023

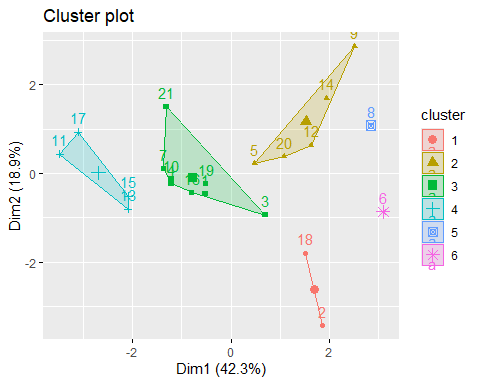
k6$size # Number of records in each cluster

## [1] 2 5 8 4 1 1

k6$cluster[21] # Identify the cluster of the 21 observation as an example

## [1] 3

fviz\_cluster(k6, data = pharma\_numeric.norm) # Visualize the output

 Based on K= 5 and K = 6, K = 5 is preferred, because it doesn’t classify multiple points that are grouped by themselves.

#Calculate Distances Between Clusters for K=6

k6$withinss

## [1] 2.803505 16.542597 21.879320 9.284424 0.000000 0.000000

dist(k6$centers)

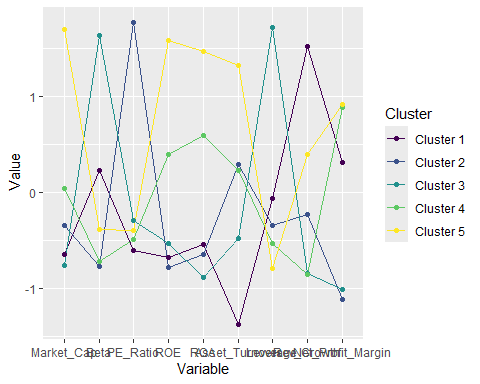
## 1 2 3 4 5  
## 2 4.012153   
## 3 4.045579 3.079758   
## 4 5.275301 4.630077 2.720924   
## 5 5.181731 4.326767 5.127326 6.508617   
## 6 4.266509 4.239064 4.795381 6.457144 4.969438

#Question 2 - K =5

library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

centroids <- data.frame(k5$centers)   
centroids['Cluster'] = paste('Cluster', seq(1:5))  
ggparcoord(centroids, columns = 1:9,   
 groupColumn ='Cluster',   
 showPoints = TRUE) +   
 scale\_color\_viridis\_d() + labs(x ='Variable', y =' Value')

 Cluster 1: This relatively small cluster has a high PE\_Ratio, Revenue Growth, but very low Asset Turnover Cluster 2: This very small cluster has a very high PE\_Ratio and mild Asset\_Turnover Cluster 3: This medium sized cluster has a high Beta and high Leverage.  
Cluster 4: This large cluster has a very low profit margin, but medium ROE, ROA, and Asset Turnover. Cluster 5: This small cluster has very high revenue growth and very lower asset\_turnover.

#Question 3 - Variables 10 - 12

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

centroids <- data.frame(k5$centers)   
centroids['Cluster'] = paste('Cluster', seq(1:5))  
centroids['Cluster']

## Cluster  
## 1 Cluster 1  
## 2 Cluster 2  
## 3 Cluster 3  
## 4 Cluster 4  
## 5 Cluster 5

summary(k5$centers)

## Market\_Cap Beta PE\_Ratio ROE   
## Min. :-0.87052 Min. :-0.4702 Min. :-0.47742 Min. :-0.8350   
## 1st Qu.:-0.76022 1st Qu.:-0.4361 1st Qu.:-0.31725 1st Qu.:-0.7438   
## Median :-0.43925 Median :-0.1781 Median :-0.19846 Median :-0.6184   
## Mean :-0.08117 Mean : 0.1073 Mean : 0.33081 Mean :-0.1534   
## 3rd Qu.:-0.03142 3rd Qu.: 0.2796 3rd Qu.:-0.05284 3rd Qu.: 0.1950   
## Max. : 1.69558 Max. : 1.3410 Max. : 2.70002 Max. : 1.2350   
## ROA Asset\_Turnover Leverage Rev\_Growth   
## Min. :-1.1928 Min. :-1.26848 Min. :-0.46808 Min. :-0.70415   
## 1st Qu.:-0.9235 1st Qu.:-0.46127 1st Qu.:-0.27449 1st Qu.:-0.69129   
## Median :-0.8107 Median : 0.17297 Median :-0.14170 Median :-0.11685   
## Mean :-0.2337 Mean :-0.03459 Mean : 0.10905 Mean : 0.09458   
## 3rd Qu.: 0.4084 3rd Qu.: 0.23063 3rd Qu.: 0.06308 3rd Qu.: 0.46718   
## Max. : 1.3503 Max. : 1.15316 Max. : 1.36645 Max. : 1.51802   
## Net\_Profit\_Margin   
## Min. :-1.416515   
## 1st Qu.:-1.320000   
## Median :-0.006894   
## Mean :-0.319042   
## 3rd Qu.: 0.556954   
## Max. : 0.591243

#Question 4 - Cluster Names

#small\_PERatio\_RevenueGrowth - Cluster 1: This relatively small cluster has a high PE\_Ratio, Revenue Growth, but very low Asset Turnover  
#small\_PERatio\_Asset - Cluster 2: This very small cluster has a very high PE\_Ratio and mild Asset\_Turnover  
#medium\_Beta\_Leverage - Cluster 3: This medium sized cluster has a high Beta and high Leverage.   
#large\_lowprofit - Cluster 4: This large cluster has a very low profit margin, but medium ROE, ROA, and Asset Turnover.  
#small\_high revenue - Cluster 5: This small cluster has very high revenue growth and very lower asset\_turnover.