Untitled

Shreya Thodupunuri

2023-11-12

**SUMMARY:**

**Data Preparation and Clustering**

**Data Loading:** The dataset is loaded from a specified path, ensuring it's correctly imported for analysis.

**Library Import:** Essential libraries like ‘tidyverse’,’ factoextra’,’ dplyr’, ‘ggplot2’, and cluster are imported for data manipulation and cluster analysis.

**Preprocessing:** The script removes any missing data and selects numerical variables (columns 3 to 11) from the dataset, which are crucial for clustering.

**Cluster Analysis**

**Data Normalization:** The selected numerical variables are normalized to ensure uniformity in scale and distribution.

**Determining Cluster Number:**

The Elbow Method is initially used to find the optimal number of clusters, analyzing the within-cluster sum of squares.

The Silhouette Method is applied as a secondary measure, indicating that 5 clusters might be optimal.

**K-Means Clustering**: The k-means clustering algorithm is performed with 5 clusters, chosen based on the silhouette method results. The process includes setting a seed for reproducibility.

**Visualization:** Cluster centroids and distances are visualized to gain insights into the distribution and separation of the clusters.

**Interpretation and Insights**

**Cluster Characteristics:**

The mean values of each variable within each cluster are calculated to understand their central tendencies and defining features.

Clusters are interpreted based on these features, revealing distinct characteristics like growth expectations, market capitalization, asset turnover, and revenue growth.

**Non-Numerical Variable Analysis:** The script also considers how non-numerical variables (like recommendations) align with each cluster, providing a deeper understanding of the clusters in the context of market perceptions.

**Cluster Naming:**

The clusters are named based on their characteristics:

Cluster 1 - "Balanced Performers": Indicates stable and decent financial metrics.

Cluster 2 - "Steady Growth Contenders": Suggests consistent growth and stability.

Cluster 3 - "Dynamic Opportunity Firms": Reflects firms with varied investment opportunities.

Cluster 4 - "Stable Investment Picks": Implies firms with good stability and solid financial metrics.

Cluster 5 - "Long-term Value Holders": Denotes firms ideal for long-term holding due to their consistent revenue growth and high asset turnover.

# Loading the dataset  
Pharmaceuticals <- read.csv("C:/Users/shrey/OneDrive/Documents/Pharmaceuticals.csv")  
# Ensuring the file path is correct and the dataset is loaded properly.

# Reading required libraries for data manipulation and clustering  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ 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)

## 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(dplyr)  
library(ggplot2)  
library(cluster)  
# These libraries are essential for manipulating data and performing cluster analysis.

**Task1**

kUse 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. Removal of missing data helps in maintaining the accuracy of the cluster analysis. Prior to clustering data, remove the missing data and rescale variables for comparability.

# Removing missing data and selecting relevant variables for cluster analysis  
Pharma\_data <- na.omit(Pharmaceuticals)   
# Removing incomplete cases to maintain the accuracy of the cluster analysis.

#Taking the quantitative variables(1-9) to cluster the 21 firms  
  
row.names(Pharma\_data)<- Pharma\_data[,1]  
Pharma\_data1<- Pharma\_data[,3:11]# Considering only numercial values i.e., 3-11 columns from csv file  
head(Pharma\_data1)

## 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  
## AGN 7.58 0.41 82.5 12.9 5.5 0.9 0.60 9.16  
## 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  
## AVE 47.16 0.32 20.1 21.8 7.5 0.6 0.34 26.81  
## BAY 16.90 1.11 27.9 3.9 1.4 0.6 0.00 -3.17  
## Net\_Profit\_Margin  
## ABT 16.1  
## AGN 5.5  
## AHM 11.2  
## AZN 18.0  
## AVE 12.9  
## BAY 2.6

# Focusing on numerical variables (columns 3 to 11) as they are key for clustering.

#Normalizing the data frame with scale method  
  
Pharma\_data2<-scale(Pharma\_data1)  
head(Pharma\_data2)

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## ABT 0.1840960 -0.80125356 -0.04671323 0.04009035 0.2416121 0.0000000  
## AGN -0.8544181 -0.45070513 3.49706911 -0.85483986 -0.9422871 0.9225312  
## AHM -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700 0.9225312  
## AZN 0.1702742 -0.02225704 -0.24290879 0.10638147 0.9181259 0.9225312  
## AVE -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461 -0.4612656  
## BAY -0.6953818 2.27578267 0.14948233 -1.45146000 -1.7127612 -0.4612656  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## ABT -0.2120979 -0.5277675 0.06168225  
## AGN 0.0182843 -0.3811391 -1.55366706  
## AHM -0.4040831 -0.5721181 -0.68503583  
## AZN -0.7496565 0.1474473 0.35122600  
## AVE -0.3144900 1.2163867 -0.42597037  
## BAY -0.7496565 -1.4971443 -1.99560225

# Scaling standardizes the data, making it suitable for clustering.

# Determining the number of clusters using Elbow Method  
  
fviz\_nbclust(Pharma\_data2, kmeans, method = "wss")



# The Elbow Method helps identify the optimal cluster count by analyzing within-cluster sum of squares.

##By seeing the above graph from Elbow method, Graph is not clear to choose k=2 or 3 or 4 or 5.  
# Using Silhouette method to determine the number of clusters  
  
fviz\_nbclust(Pharma\_data2, kmeans, method = "silhouette")



By seeing the graph from silhouette method, I can see sharp rise at k=5. So, considering the silhouette method.

Silhouette method assesses how well each object lies within its cluster, aiding in determining the best number of clusters.

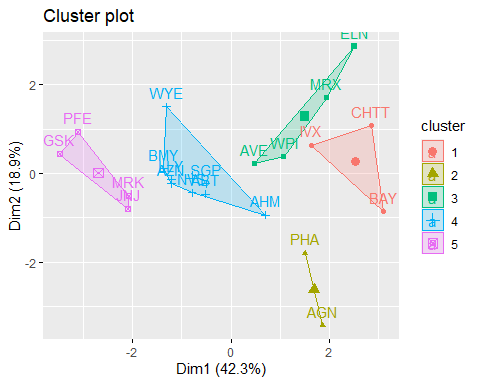
# Performing K-means clustering  
set.seed(64060)  
k\_5 <- kmeans(Pharma\_data2, centers=5, nstart=25)

K-means clustering is performed with 5 clusters, determined by previous methods.

# Visualizing cluster centroids and distances  
  
k\_5$centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 2 -0.43925134 -0.4701800 2.70002464 -0.8349525 -0.9234951 0.2306328  
## 3 -0.76022489 0.2796041 -0.47742380 -0.7438022 -0.8107428 -1.2684804  
## 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 1.36644699 -0.6912914 -1.320000179  
## 2 -0.14170336 -0.1168459 -1.416514761  
## 3 0.06308085 1.5180158 -0.006893899  
## 4 -0.27449312 -0.7041516 0.556954446  
## 5 -0.46807818 0.4671788 0.591242521

fviz\_cluster(k\_5,data = Pharma\_data2) # to Visualize the clusters



k\_5

## K-means clustering with 5 clusters of sizes 3, 2, 4, 8, 4  
##   
## Cluster means:  
## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 2 -0.43925134 -0.4701800 2.70002464 -0.8349525 -0.9234951 0.2306328  
## 3 -0.76022489 0.2796041 -0.47742380 -0.7438022 -0.8107428 -1.2684804  
## 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 1.36644699 -0.6912914 -1.320000179  
## 2 -0.14170336 -0.1168459 -1.416514761  
## 3 0.06308085 1.5180158 -0.006893899  
## 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 3 1 4 1 3 4 5 1 5 3 5 4   
## PFE PHA SGP WPI WYE   
## 5 2 4 3 4   
##   
## Within cluster sum of squares by cluster:  
## [1] 15.595925 2.803505 12.791257 21.879320 9.284424  
## (between\_SS / total\_SS = 65.4 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

distance<- dist(Pharma\_data2, method = "euclidean")  
fviz\_dist(distance)



# I can see there are 5 clusters and the center is defined after 25 restarts which is determined in kmeans.

Visualization aids in understanding the distribution and separation of clusters.

# Refitting K-means for a clearer interpretation  
#K-Means Cluster Analysis- Fit the data with 5 clusters  
  
fit<-kmeans(Pharma\_data2,5)

# Analyzing the mean values of each variable within each cluster  
  
aggregate(Pharma\_data2,by=list(fit$cluster),FUN=mean)

## Group.1 Market\_Cap Beta PE\_Ratio ROE ROA  
## 1 1 1.69558112 -0.1780563 -0.1984582 1.2349879 1.3503431  
## 2 2 -0.66114002 -0.7233539 -0.3512251 -0.6736441 -0.5915022  
## 3 3 -0.96247577 1.1949250 -0.3639982 -0.5200697 -0.9610792  
## 4 4 -0.52462814 0.4451409 1.8498439 -1.0404550 -1.1865838  
## 5 5 0.08926902 -0.4618336 -0.3208615 0.3260892 0.5396003  
## Asset\_Turnover Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 1.153164e+00 -0.4680782 0.4671788 0.5912425  
## 2 -1.537552e-01 -0.4040831 0.6917224 -0.4005718  
## 3 -1.153164e+00 1.4773718 0.7120120 -0.3688236  
## 4 1.480297e-16 -0.3443544 -0.5769454 -1.6095439  
## 5 6.589509e-02 -0.2559803 -0.7230135 0.7343816

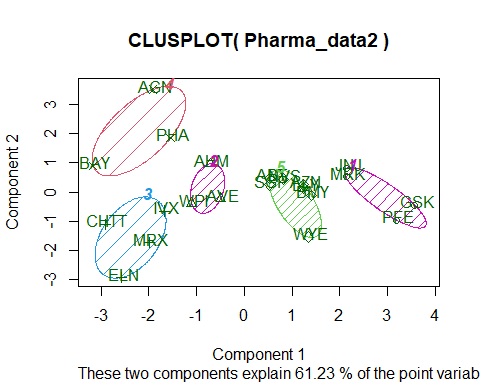
Pharma\_data3<-data.frame(Pharma\_data2,fit$cluster)  
Pharma\_data3

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## ABT 0.1840960 -0.80125356 -0.04671323 0.04009035 0.2416121 0.0000000  
## AGN -0.8544181 -0.45070513 3.49706911 -0.85483986 -0.9422871 0.9225312  
## AHM -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700 0.9225312  
## AZN 0.1702742 -0.02225704 -0.24290879 0.10638147 0.9181259 0.9225312  
## AVE -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461 -0.4612656  
## BAY -0.6953818 2.27578267 0.14948233 -1.45146000 -1.7127612 -0.4612656  
## BMY -0.1078688 -0.10015669 -0.70887325 0.59693581 0.8617498 0.9225312  
## CHTT -0.9767669 1.26308721 0.03299122 -0.11237924 -1.1677918 -0.4612656  
## ELN -0.9704532 2.15893320 -1.34037772 -0.70899938 -1.0174553 -1.8450624  
## LLY 0.2762415 -1.34655112 0.14948233 0.34502953 0.5610770 -0.4612656  
## GSK 1.0999201 -0.68440408 -0.45749769 2.45971647 1.8389364 1.3837968  
## IVX -0.9393967 0.48409069 -0.34100657 -0.29136529 -0.6979905 -0.4612656  
## JNJ 1.9841758 -0.25595600 0.18013789 0.18593083 1.0872544 0.9225312  
## MRX -0.9632863 0.87358895 0.19240011 -0.96753478 -0.9610792 -1.8450624  
## MRK 1.2782387 -0.25595600 -0.40231769 0.98142435 0.8429577 1.8450624  
## NVS 0.6654710 -1.30760129 -0.23677768 -0.52338423 0.1288598 -0.9225312  
## PFE 2.4199899 0.48409069 -0.11415545 1.31287998 1.6322239 0.4612656  
## PHA -0.0240846 -0.48965495 1.90298017 -0.81506519 -0.9047030 -0.4612656  
## SGP -0.4018812 -0.06120687 -0.40231769 -0.21181593 0.5234929 0.4612656  
## WPI -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905 -0.9225312  
## WYE -0.1614497 0.40619104 -0.75792214 1.92938746 0.5422849 -0.4612656  
## Leverage Rev\_Growth Net\_Profit\_Margin fit.cluster  
## ABT -0.21209793 -0.52776752 0.06168225 5  
## AGN 0.01828430 -0.38113909 -1.55366706 4  
## AHM -0.40408312 -0.57211809 -0.68503583 2  
## AZN -0.74965647 0.14744734 0.35122600 5  
## AVE -0.31449003 1.21638667 -0.42597037 2  
## BAY -0.74965647 -1.49714434 -1.99560225 4  
## BMY -0.02011273 -0.96584257 0.74744375 5  
## CHTT 3.74279705 -0.63276071 -1.24888417 3  
## ELN 0.61983791 1.88617085 -0.36501379 3  
## LLY -0.07130879 -0.64814764 1.17413980 5  
## GSK -0.31449003 0.76926048 0.82363947 1  
## IVX 1.10620040 0.05603085 -0.71551412 3  
## JNJ -0.62166634 -0.36213170 0.33598685 1  
## MRX 0.44065173 1.53860717 0.85411776 3  
## MRK -0.39128411 0.36014907 -0.24310064 1  
## NVS -0.67286239 -1.45369888 1.02174835 5  
## PFE -0.54487226 1.10143723 1.44844440 1  
## PHA -0.30169102 0.14744734 -1.27936246 4  
## SGP -0.74965647 -0.43544591 0.29026942 5  
## WPI -0.49367621 1.43089863 -0.09070919 2  
## WYE 0.68383297 -1.17763919 1.49416183 5

View(Pharma\_data3)

Aggregate functions reveal the central tendencies of each cluster, highlighting their defining characteristics.

# Cluster plot visualization  
  
clusplot(Pharma\_data2,fit$cluster,color = TRUE,shade = TRUE,labels = 2,lines = 0)



Clusplot visually represents the clustering, showing the grouping and outliers if any.

**Task 2**

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

By noticing the mean values of all quantitative variables for each cluster

Cluster\_1 - AGN, PHA, BAY - Suggests higher growth expectations or overvaluation.

Cluster\_2 - JNJ, MRK, GSK, PFE - High Market Cap and Leverage: Indicates large, established companies.

Cluster\_3 - AHM, AVE, WPI - Low Asset Turnover and Beta: Represents conservative, stable firms.

Cluster\_4 - IVX, MRX, ELN, CHTT - Low Market Capital but High Revenue Growth: Reflects emerging growth companies.

Cluster\_5 - ABT, NVS, AZN, LLY, BMY, WYE, SGP - Low Revenue Growth, High Asset Turnover, and Net Profit Margin: Signifies efficient, profitably run firms.

**Task 3**

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

For cluster 1: It has the highest PE\_Ratio and needs to be held as per the media recommendations.

For cluster 2: It has the highest market\_Cap and has Good Leverage value. And they can be moderately recommended.

For cluster 3: It has lowest asset\_turnover,and lowest beta. But media recommendations are highly positive.

For cluster 4: The leverage ratio is high, they are moderately recommended.

For Cluster 5: They have lowest revenue growth, highest assest turnover and highest net profit margin.

They are recommended to be held for longer time.

**Task 4**

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

Cluster 1 - Balanced Performers: This name suggests that firms in this cluster have stable and decent financial metrics. It implies a balanced performance across various financial aspects.

Cluster 2 - Steady Growth Contenders: This name indicates that companies in this cluster demonstrate consistent growth, making them a moderate but reliable option for investment or holding. It reflects both stability and potential for growth.

Cluster 3 - Dynamic Opportunity Firms: This name implies that firms in this cluster might present varied investment opportunities, characterized by both potential growth (buy) and higher risk (sell). It suggests dynamism and variability in performance.

Cluster 4 - Stable Investment Picks: This name reflects firms with good stability and solid financial metrics, making them attractive for buying and long-term investment.

Cluster 5 - Long-term Value Holders: This name suggests that firms in this cluster are ideal for holding due to their potential to provide long-term value, likely characterized by lower but consistent revenue growth and high asset turnover.

**Conclusion**: This analysis effectively segments the pharmaceutical firms into five distinct clusters, each with unique financial characteristics. The use of both the Elbow and Silhouette methods provides a robust approach to determining the optimal number of clusters, and the visualization tools aid in the interpretation of the results. The analysis offers valuable insights into the firm’s market positions and operational efficiencies, useful for investment decisions or strategic industry analysis.