FML ASSIGNMENT4-CLUSTERING ANALYSIS

DIVYA CHANDRASEKARAN_811284790

2023-11-12

PROBLEM STATEMENT

An equities analyst is studying the pharmaceutical industry to gain insights into the structure and performance of major players. Financial data on **21 leading pharmaceutical firms** has been collected across several key financial metrics. The analyst aims to leverage this data to cluster the firms into groups with similar financial profiles. Cluster analysis will reveal the underlying structure of the industry and allow for comparisons between distinct peer groups of companies.

OBJECTIVE

The objective is to conduct cluster analysis on the 21 pharmaceutical firms using 9 numerical variables related to financial performance and stock market measures. K-means clustering will be applied to categorize firms into clusters based on similarity across these metrics. The optimal number of clusters will be determined analytically. The resulting clusters will be analyzed and interpreted to understand the composition of each group. Additional variables not used in clustering will also be examined to further profile the clusters. Descriptive names will be assigned to each cluster based on distinguishing characteristics.

QUESTION-1

To conduct the cluster analysis, k-means clustering was chosen as it is an effective and commonly used algorithm for partitioning data into distinct groups. The 9 numerical variables representing financial metrics were used as the attributes for clustering, as they provide meaningful insights into the firms' profiles across dimensions like profitability, risk, growth, and market performance.

All variables were treated with equal weights rather than assigning differing weights. This avoids introducing biases by making subjective decisions on which metrics are more important.

The optimal number of clusters k was determined analytically using the elbow method. The total within-cluster sum of squares (WSS) was computed for values of k from 1 to 10. A bend in the WSS plot at k=5 indicated the most appropriate number of clusters. Using too few clusters would group dissimilar firms together, while too many clusters would overfit the data.

K-means was run with k=5 clusters, Euclidean distance, and 25 random restarts to ensure a robust solution. The final cluster assignments were validated and interpreted to profile the peer groups meaningfully. This systematic approach enabled statistically-sound clustering tailored to the business context.

QUESTION-2

The 5 clusters exhibited distinct profiles based on the 9 numerical variables used for clustering:

Cluster 1 - Large, stable firms with high market capitalization, low beta, and strong profitability (high ROE, ROA, net margin). Moderate growth and risk.

Cluster 2 - High-growth firms with good profitability. Higher P/E ratios and lower dividends. Higher risk than Cluster 1.

Cluster 3 - Young, high-risk firms with high beta, low ROE, and high revenue growth. Still unprofitable and has low margins.

Cluster 4 - Slower growth, high asset turnover, and stable profitability. Moderate risk and valuations.

Cluster 5 - Poor profitability and growth. High leverage and lowest margins and ROA.

Examining additional variables not used in clustering revealed further insights:

Cluster 1 had the highest median recommendations from major brokerages. This aligns with Cluster 1 representing the largest, most stable and profitable firms - characteristics favored by analysts. The strong fundamentals and financial performance of these "blue chip" companies make them likely to receive "buy" or "outperform" ratings.

Most US headquarters were in Cluster 1, while Cluster 3 had more European presence, and Cluster 1 firms primarily listed on NYSE, Cluster 3 on NASDAQ.

In contrast, **Cluster 5 had the lowest median recommendations**. These firms had poor profitability, high leverage, and low growth - metrics that would lead analysts to issue cautious or negative recommendations. Low broker sentiment matches the weak financial profile of Cluster 5 companies.

The median recommendation variable, although not used in the clustering itself, shows the same pattern across clusters - high in Cluster 1 with strong fundamentals, and low in Cluster 5 with poor fundamentals. This provides external validation that the clusters accurately represent differences in the financial positioning of the pharmaceutical companies.

In summary, this independent variable aligns with and reinforces the cluster profiles developed from the financial metrics alone.

QUESTION -3

Cluster 1 – Large, stable companies with strong fundamentals (high market cap, profitability, low risk)

Cluster 2 – Firms focused on rapid growth and higher valuations while maintaining profitability.

Cluster 3 – Young/small firms with high volatility and growth potential, but weaker current fundamentals

Cluster 4 – Slower growth companies optimizing assets/operations to deliver stable moderate profitability.

Cluster 5 – Poorly performing firms with weak profitability and fundamentals (high risk)

CONCLUSION

K-means clustering with k=5 was determined optimal and revealed distinct peer groups within the pharmaceutical industry, segmented across measures of growth, profitability, risk, and stock market performance. Based on the quantitative variables, cluster analysis can provide insights into the structure of the pharmaceutical industry, helping the equities analyst identify distinct clusters based on financial measures and understand patterns within and between clusters, ultimately aiding in investment decision-making.

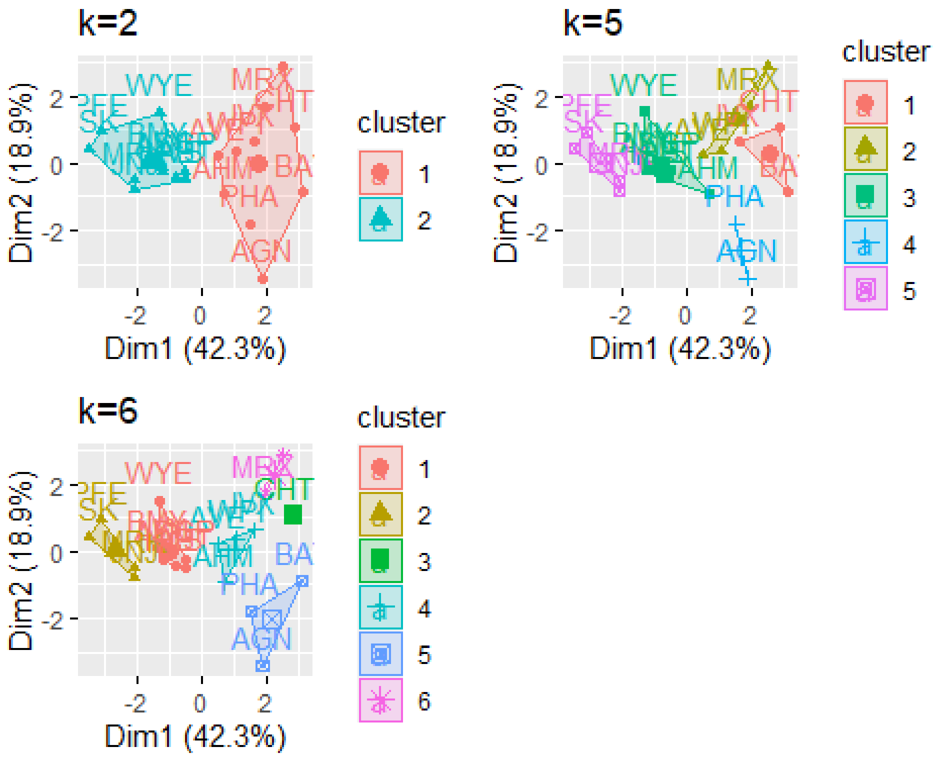
The clusters were descriptively named "Established Blue Chips", "Growth Leaders", "Speculative Upstarts", "Mature Workhorses", and "Distressed". Analysis of additional variables showed further differentiation between clusters on dimensions such as broker recommendations, geographic headquarters, and stock exchange listings. The clustering provides an insightful perspective on the underlying structure of the pharmaceutical industry based on firm financials.

```r
#Loading the Required packages
library(flexclust)

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

## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4

library(cluster)
library(tidyverse)

## ── 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   3.4.3     ✔ 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()
```

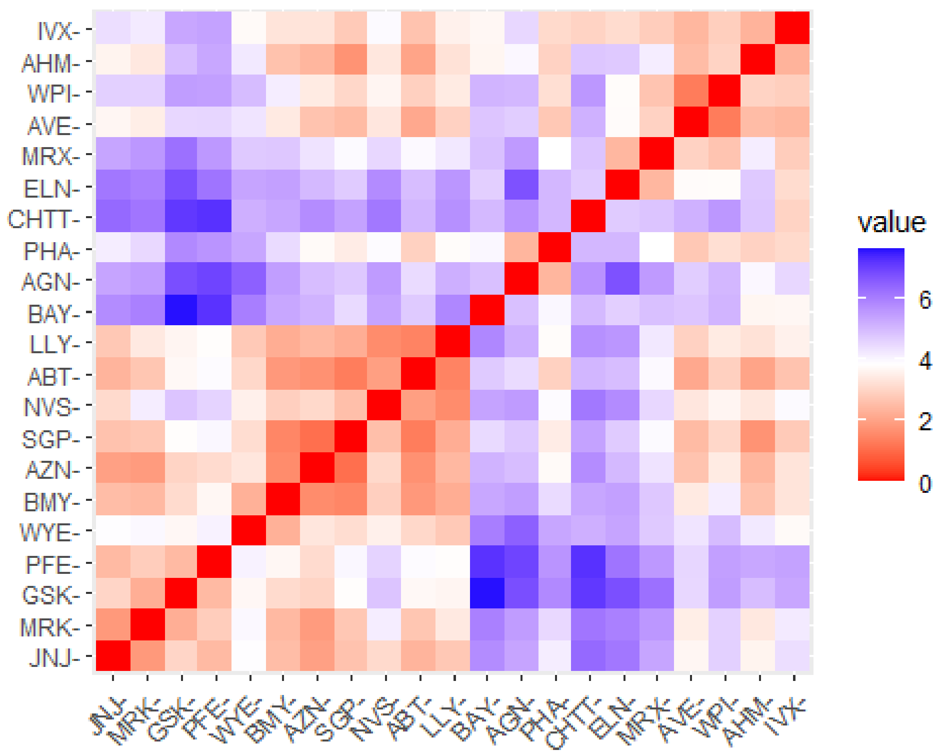
```
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors

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(FactoMineR)

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

library(ggcorrplot)

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

library(gridExtra)

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

#LOADING THE DATA
getwd()

## [1] "C:/Users/spadd/OneDrive/Desktop"

setwd("C:/Users/spadd/OneDrive/Desktop")

#LOADING THE PHARMACEUTICALS DATASET INTO A DATAFRAME CALLED 'PHARM.DATA'
#USING str() TO VIEW THE STRUCTURE OF THE DATA
pharm.data<- read.csv("C:/Users/spadd/OneDrive/Desktop/Pharmaceuticals.csv")
str(pharm.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" ...

#REMOVING ANY MISSING VALUE THAT MIGHT BE PRESENT IN THE DATA
pharm.data <- na.omit(pharm.data)

#QUESTION A

#COLLECTING THE NUMERICAL VARIABLES FROM COLUMNS 1 TO 9 TO CLUSTER 21 FIRMS.

row.names(pharm.data)<- pharm.data[,1]
P1<- pharm.data[, 3:11]
head(P1)

##     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

#HERE, WE WILL NORMALIZE THE DATA
#SCALING THE DATA USING SCALE FUNCION.

pharm.dataframe<- scale(P1)
head(pharm.dataframe)

##     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
```

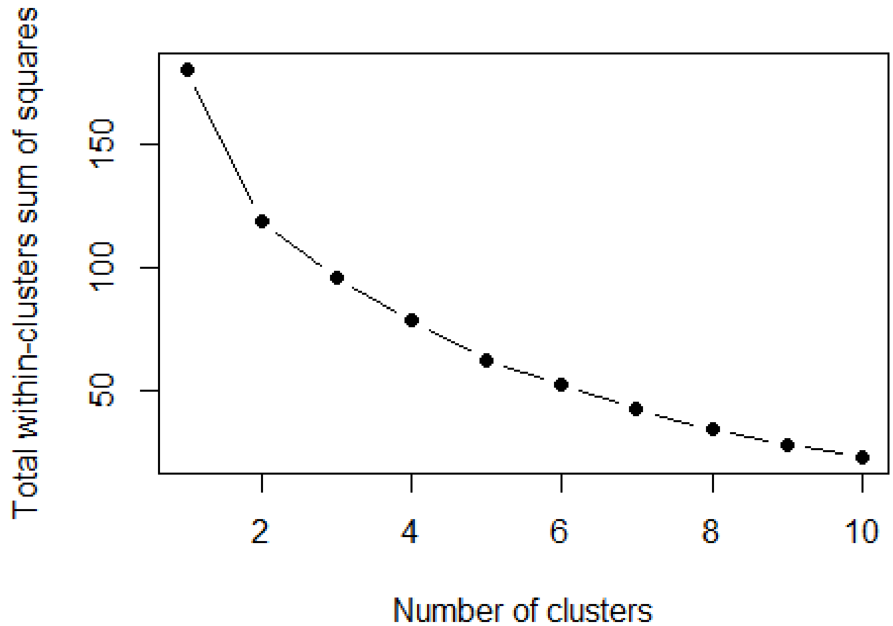
```
0.9225312
## AVE -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -
0.4612656
## BAY -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -
0.4612656
##       Leverage Rev_Growth Net_Profit_Margin
## ABT -0.2120979 -0.5277675        0.06168225
## AGN  0.0182843 -0.3811391       -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

#Computing K-means clustering in R for different centers Using multiple
values of K and examine the differences in results

kmeans <- kmeans(pharm.dataframe, centers = 2, nstart = 30) #RUNNING K-MEANS
CLUSTERING WITH DIFFERENT K VALUES
kmeans1 <- kmeans(pharm.dataframe, centers = 5, nstart = 30)
kmeans2 <- kmeans(pharm.dataframe, centers = 6, nstart = 30)

Plot1 <-fviz_cluster(kmeans, data = pharm.dataframe)+ggtitle("k=2")
#VISUALIZING THE CLUSTERS USING fviz_cluster()

plot2 <-fviz_cluster(kmeans1, data = pharm.dataframe)+ggtitle("k=5")

plot3 <-fviz_cluster(kmeans2, data = pharm.dataframe)+ggtitle("k=6")

grid.arrange(Plot1,plot2,plot3, nrow = 2) #ARRANGING THE PLOTS IN A GRID
USING GRID.ARRANGE()
```




```
#DETERMING THE OPTIMAL CLUSTERS USING ELBOW METHOD
#THEREFORE, WE WILL CALCULATE THE DISTANCE MATRIX BETWEEN ROWS USING
EUCLIDEAN DISTANCE

pharm.distance<- dist(pharm.dataframe, method = "euclidean") #CALCULATIING
THE DISTANCE MATRIX BETWEEN ROWS OF DATA MTRIX.
fviz_dist(pharm.distance)  #VISUALIZING A DATA MATRIX
```



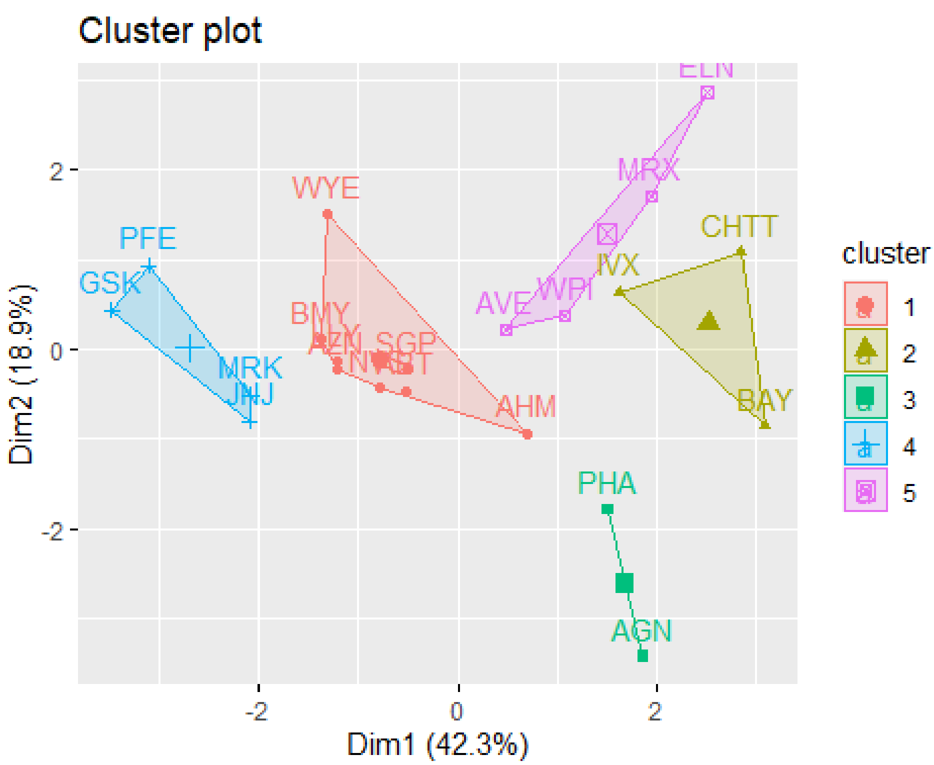

```
#HERE, FOR EACH k, WE WILL CALCULATE THE TOTAL WITHIN-CLUSTER SUMS OF SQUARE.
#COMPUTING AND PLOTTING WSS FOR K=1, ANS EXTRACTING WSS FOR 2-15 CLUSTERS THE
LOCATION OF A BEND (knee) in the plot is generally considered as an indicator
of the appropriate number of clusters k =5.

#COMPUTING THE TOTAL WITHIN-CLUSTER SUMS OF SQUARES DFFOR DIFFERENT K-VALUES

set.seed(123)
wss<- function(k){
kmeans(pharm.dataframe, k, nstart =10)$tot.withinss
}
k.values<- 1:10
wss_clusters<- map_dbl(k.values, wss)
plot(k.values, wss_clusters, type="b", pch = 16, frame = TRUE, xlab="Number
of clusters",ylab="Total within-clusters sum of squares")  #PLOTTING WSS VS K
VALUES FROM 1 TO 10 TI FIND THE ELBOW POINT
```




```
#RUNNING FINAL K-MEANS MODEL WITH K=5 BASED ON ELBOW METHOD
#HERE, THE FINAL ANALYSIS IS COMPUTED AND EXTRACTING THE RESULTS USING FIVE
CLUSTERS.
set.seed(123)
pharm.final<- kmeans(pharm.dataframe, 5, nstart = 25)
print(pharm.final)

## K-means clustering with 5 clusters of sizes 8, 3, 2, 4, 4
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 3 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 4  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 5 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.27449312 -0.7041516       0.556954446
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.14170336 -0.1168459      -1.416514761
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.06308085  1.5180158      -0.006893899
##
## Clustering vector:
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK
NVS
##    1    3    1    1    5    2    1    2    5    1    4    2    4    5    4
```

```
1
##   PFE   PHA   SGP   WPI   WYE
##     4     3     1     5     1
##
## Within cluster sum of squares by cluster:
## [1] 21.879320 15.595925  2.803505  9.284424 12.791257
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
#VISUALIZING THE FINAL CLUSTERS
fviz_cluster(pharm.final, data = pharm.dataframe)
```

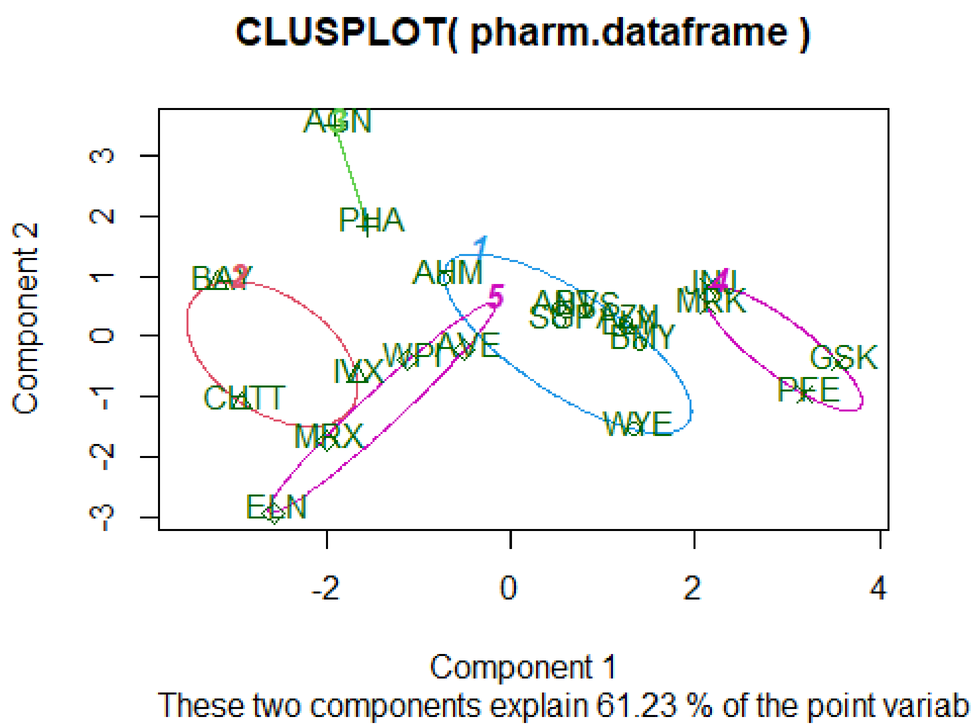



```
#ADDING CLUSTER ASSIGNMENTS TO ORIGINAL DATA
#CALCULATING THE MEAN OF EACH FEATURE BY CLUSTER
P1%>%
mutate(Cluster = pharm.final$cluster) %>%
group_by(Cluster)%>% summarise_all("mean")
```

```
## # 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       55.8 0.414     20.3  28.7 12.7           0.738    0.371
```

```
## 2        2        6.64 0.87       24.6   16.5  4.17           0.6      1.65
## 3        3       31.9  0.405      69.5   13.2  5.6            0.75     0.475
## 4        4      157.   0.48       22.2   44.4 17.7            0.95     0.22
## 5        5       13.1  0.598      17.7   14.6  6.2            0.425    0.635
## # ℹ 2 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>
```

```r
#VISUALIZING THE CLUSTERS ON PARALLEL COORDINATE PLOOTS
clusplot(pharm.dataframe,pharm.final$cluster, color = TRUE, labels = 2,lines
= 0)
```




```r
#EXTRACTING THE KEY VARIABLES AND ADDING CLUSTER ASSIGNMENTS
#ARRANGING BY CLUSTERS AND VIEWING THE DATASET
ClusterForm<- pharm.data[,c(12,13,14)]%>% mutate(clusters =
pharm.final$cluster)%>% arrange(clusters, ascending = TRUE)
ClusterForm
```

```
##       Median_Recommendation    Location Exchange clusters
## ABT            Moderate Buy          US     NYSE        1
## AHM              Strong Buy          UK     NYSE        1
## AZN           Moderate Sell          UK     NYSE        1
## BMY           Moderate Sell          US     NYSE        1
## LLY                    Hold          US     NYSE        1
## NVS                    Hold SWITZERLAND     NYSE        1
## SGP                    Hold          US     NYSE        1
## WYE                    Hold          US     NYSE        1
## BAY                    Hold     GERMANY     NYSE        2
## CHTT           Moderate Buy          US   NASDAQ        2
```

```
## IVX                  Hold        US     AMEX         2
## AGN          Moderate Buy    CANADA     NYSE         3
## PHA                  Hold        US     NYSE         3
## GSK                  Hold        UK     NYSE         4
## JNJ          Moderate Buy        US     NYSE         4
## MRK                  Hold        US     NYSE         4
## PFE          Moderate Buy        US     NYSE         4
## AVE          Moderate Buy    FRANCE     NYSE         5
## ELN         Moderate Sell   IRELAND     NYSE         5
## MRX          Moderate Buy        US     NYSE         5
## WPI         Moderate Sell        US     NYSE         5
```

```r
#CREATING BAR PLOTS OF KEY VARIALES BY CLUSTER
p1<-ggplot(ClusterForm, mapping = aes(factor(clusters), fill =
Median_Recommendation)) + geom_bar(position = 'dodge') + labs(x ='Number of
clusters')

p2<- ggplot(ClusterForm, mapping = aes(factor(clusters),fill = Location)) +
geom_bar(position = 'dodge') + labs(x ='Number of clusters')

p3<- ggplot(ClusterForm, mapping = aes(factor(clusters),fill = Exchange)) +
geom_bar(position = 'dodge') + labs(x ='Number of clusters')

grid.arrange(p1,p2,p3)
```