Assignment 4 Clustering

YASHWANTH REDDY SADALA

2023-11-13

#Problem Statement: 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:

1.Market capitalization (in billions of dollars) 2.Beta 3.Price/earnings ratio 4.Return on equity 5.Return on assets 6.Asset turnover 7.Leverage 8.Estimated revenue growth 9.Net profit margin 10.Median recommendation (across major brokerages) 11.Location of firm's headquarters 12.Stock exchange on which the firm is listed

Use cluster analysis to explore and analyze the given dataset as follows:

1.Use 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. 2.Interpret the clusters with respect to the numerical variables used in forming the clusters. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters) 3.Provide an appropriate name for each cluster using any or all of the variables in the dataset.

#Running the necessary libraries

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## ── 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   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ 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 becom
e 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)
```

#Importing and reading the CSV file

```
Pharmaceuticals <- read.csv("C:/Users/yashw/FML/Pharmaceuticals.csv")
Pharma <- na.omit(Pharmaceuticals)
head(Pharmaceuticals)
```

```
##   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
```

```
dim(Pharmaceuticals)
```

```
## [1] 21 14
```

```
t(t(names(Pharmaceuticals)))
```

```
##       [,1]
##  [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"
```

#1. Use 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.

```
#using only the quantitative variables(1-9) to cluster the 21 firms
row.names(Pharma)<- Pharma[,1]
Pharma1<- Pharma[,3:11]

#Considering only numerical values 3-11 columns.
head(Pharma1)
```

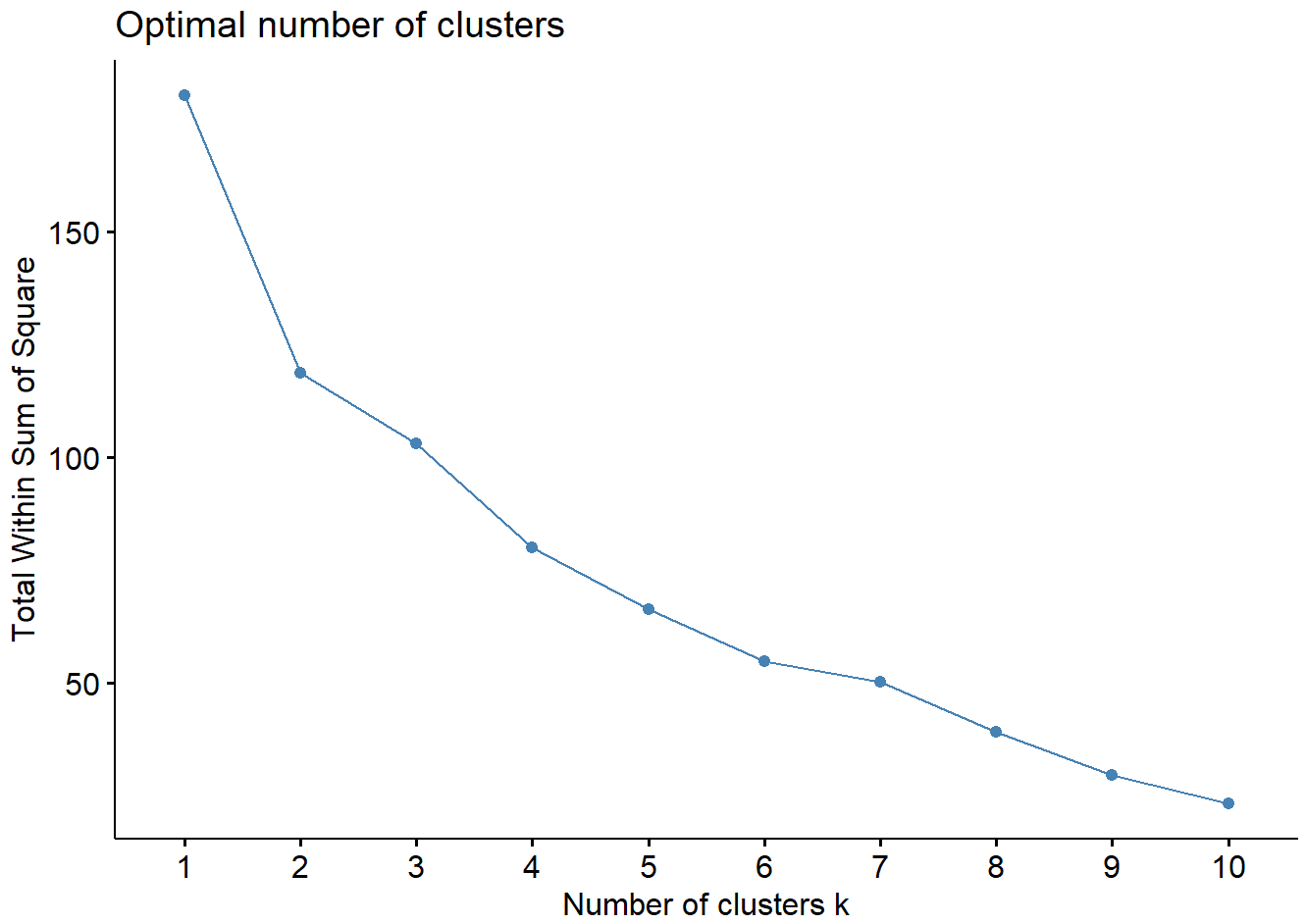
```
##     Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## ABT      68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54
## AGN       7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16
## AHM       6.30 0.46     20.7 14.9  7.8            0.9     0.27       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
```

```
#Normalizing data
Pharma2<-scale(Pharma1)
head(Pharma2)
```

```
##     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
```

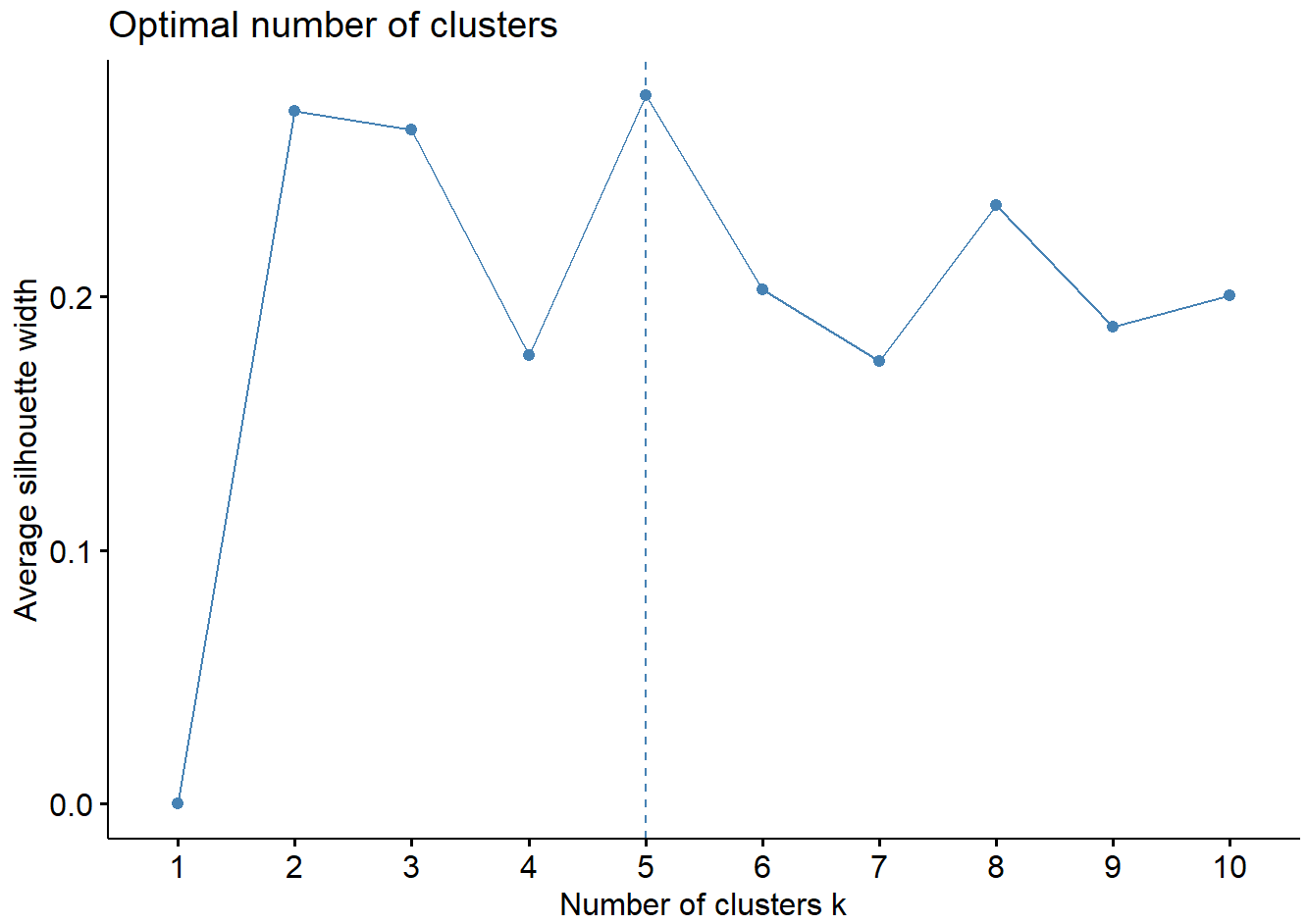
#Additionally, to calculate how many clusters to need for the Elbow Method cluster analysis

```
fviz_nbclust(Pharma2, kmeans, method = "wss")
```



#Looking at the Elbow method's above graph, we can see that it's unclear which of the values k=2, 3, 4, or 5 to select. The silhouette method for counting the number of clusters

```
fviz_nbclust(Pharma2, kmeans, method = "silhouette")
```



#Applying K-means clustering

```
set.seed(150)
clus5<- kmeans(Pharma2,centers=5,nstart = 25)
```

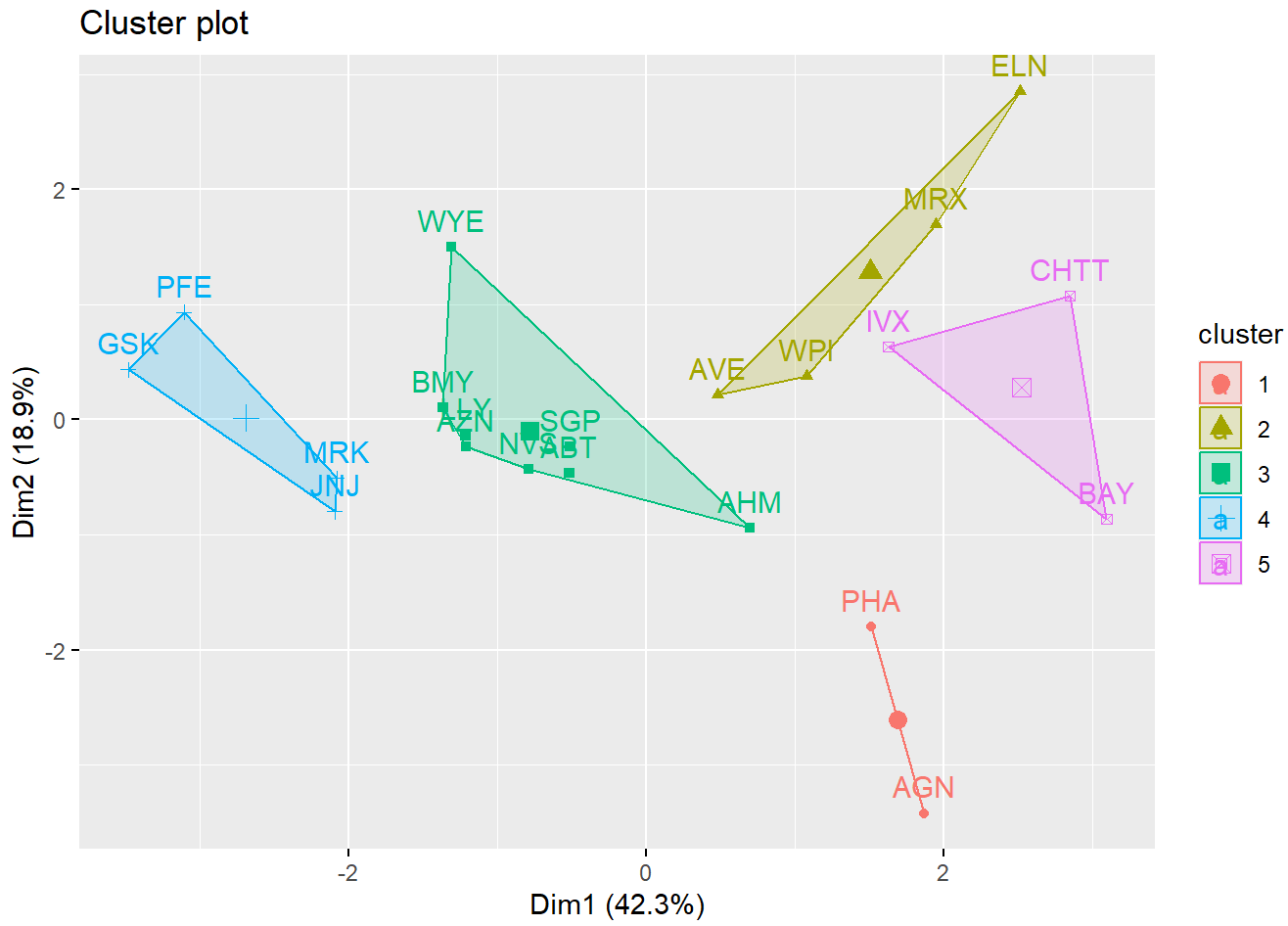
#Visualizing the output

```
clus5$centers
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 2 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 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 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.14170336 -0.1168459      -1.416514761
## 2  0.06308085  1.5180158      -0.006893899
## 3 -0.27449312 -0.7041516       0.556954446
## 4 -0.46807818  0.4671788       0.591242521
## 5  1.36644699 -0.6912914      -1.320000179
```

#Visualizing the clusters

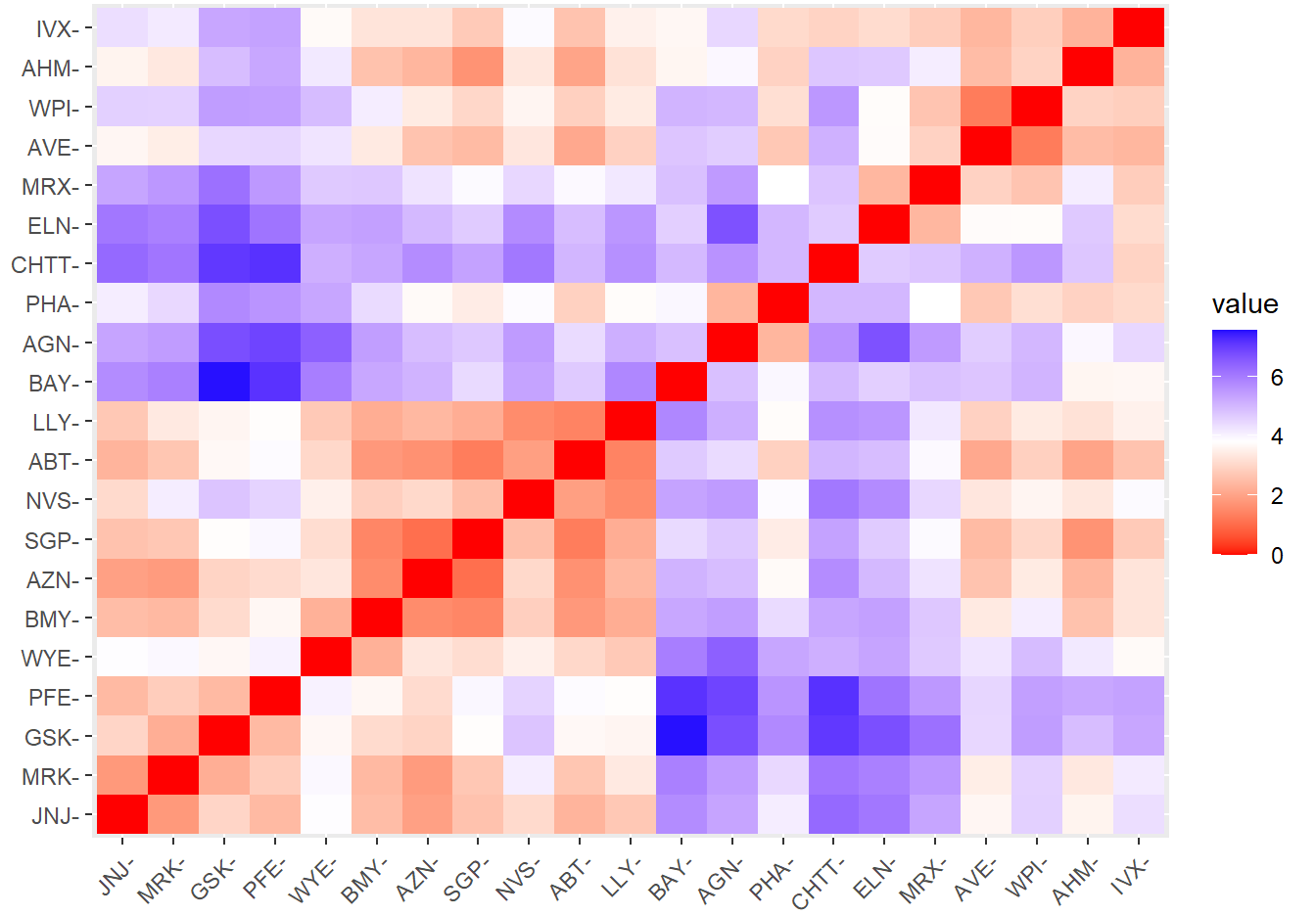
```
fviz_cluster(clus5,data = Pharma2)
```


Cluster plot
ELN
MRX
CHTT
IVX
WPI
AVE
WYE
PFE
GSK
BMY
AZN
SGP
NVS
ABT
MRK
JNJ
AHM
BAY
PHA
AGN
2
0
-2
Dim2 (18.9%)
-2
0
2
Dim1 (42.3%)
cluster
1
2
3
4
5

```
clus5
```

```
## K-means clustering with 5 clusters of sizes 2, 4, 8, 4, 3
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 2 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 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 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.14170336 -0.1168459      -1.416514761
## 2  0.06308085  1.5180158      -0.006893899
## 3 -0.27449312 -0.7041516       0.556954446
## 4 -0.46807818  0.4671788       0.591242521
## 5  1.36644699 -0.6912914      -1.320000179
##
## Clustering vector:
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    3    1    3    3    2    5    3    5    2    3    4    5    4    2    4    3
##  PFE  PHA  SGP  WPI  WYE
##    4    1    3    2    3
##
## Within cluster sum of squares by cluster:
## [1]  2.803505 12.791257 21.879320  9.284424 15.595925
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
distance<- dist(Pharma2, method = "euclidean")
fviz_dist(distance)
```



#There are five clusters, as can be seen, and the center is established after 25 restarts, as calculated using the k-means algorithm. K - Cluster Analysis of Means Fit five clusters to the data.

```
fit<-kmeans(Pharma2,5)
```

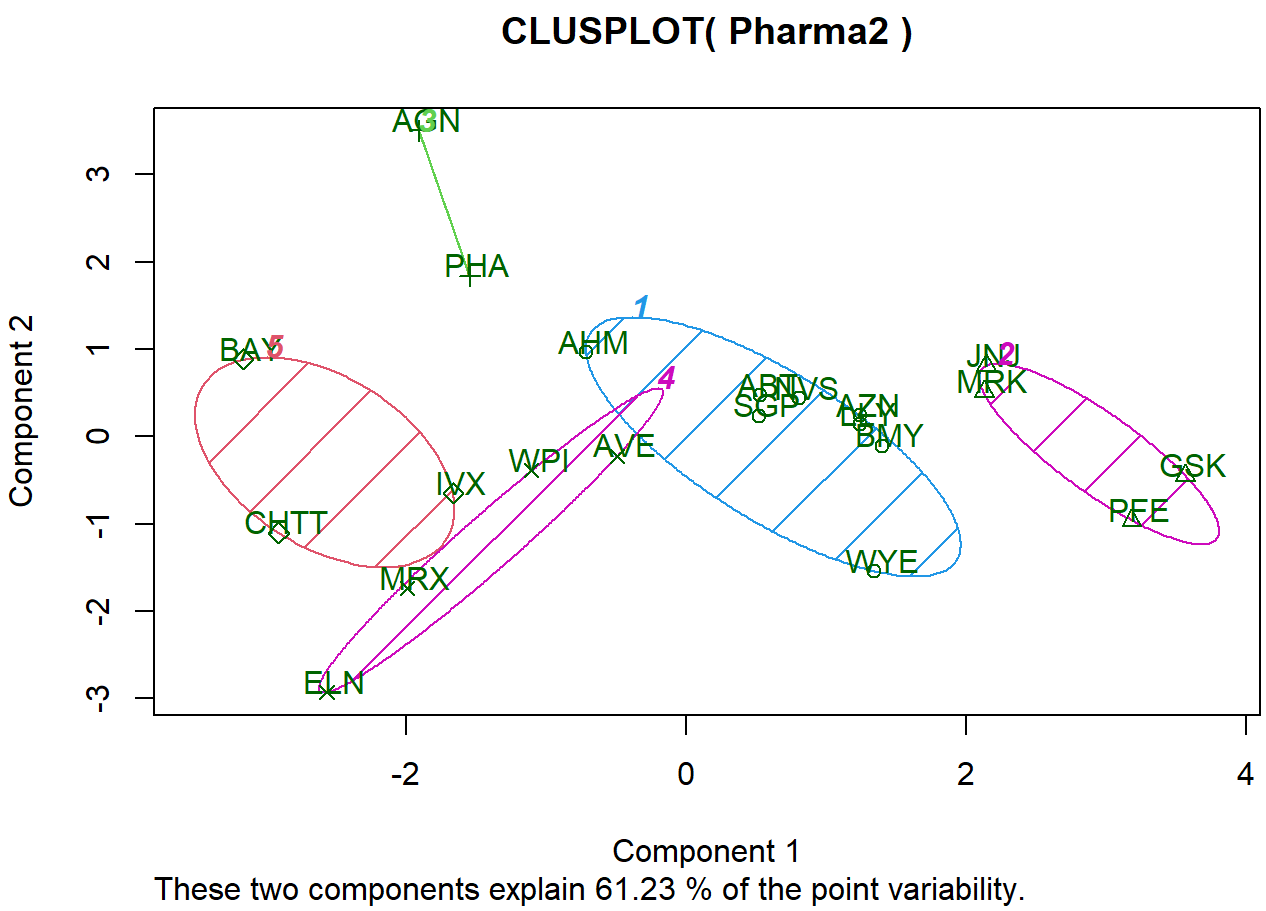
#Determining each cluster's mean value for every quantitative variable

```
aggregate(Pharma2,by=list(fit$cluster),FUN=mean)
```

```
##   Group.1  Market_Cap       Beta    PE_Ratio        ROE        ROA
## 1       1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915
## 2       2  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431
## 3       3 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951
## 4       4 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428
## 5       5 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478
##   Asset_Turnover    Leverage Rev_Growth Net_Profit_Margin
## 1      0.1729746 -0.27449312 -0.7041516       0.556954446
## 2      1.1531640 -0.46807818  0.4671788       0.591242521
## 3      0.2306328 -0.14170336 -0.1168459      -1.416514761
## 4     -1.2684804  0.06308085  1.5180158      -0.006893899
## 5     -0.4612656  1.36644699 -0.6912914      -1.320000179
```

#To see the layout of clusters

```
clusplot(Pharma2,fit$cluster,color =
           TRUE,shade = TRUE,labels = 2,lines = 0)
```



#2. Interpret the clusters in light of the numerical variables that were utilised to create them.

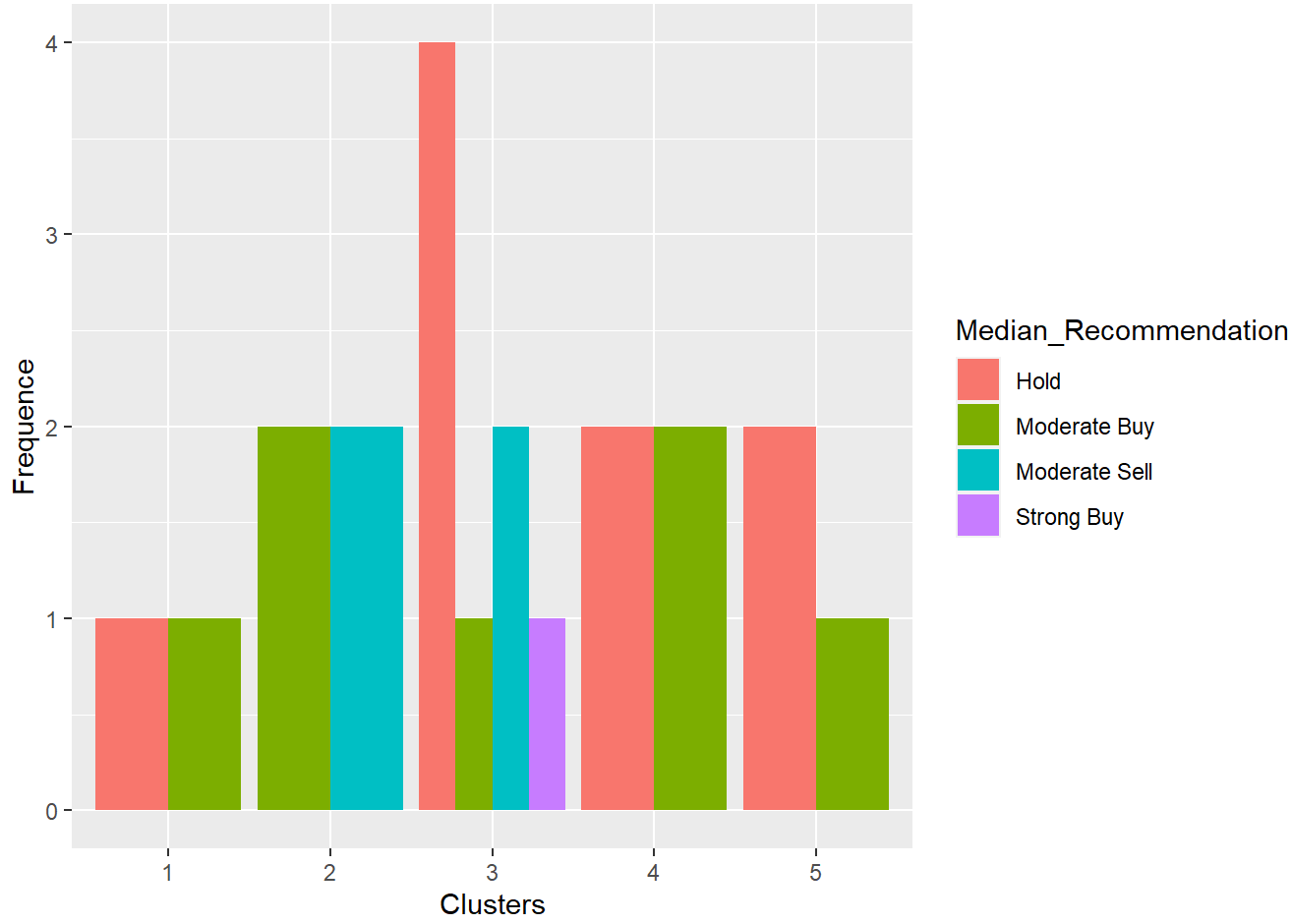
```
Pharmacluster <- Pharma[,c(12,13,14)]%>% mutate(clusters = clus5$cluster)%>% arrange(clusters, a
scending = TRUE)
Pharmacluster
```

```
##      Median_Recommendation    Location Exchange clusters
## AGN           Moderate Buy      CANADA     NYSE        1
## PHA                   Hold          US     NYSE        1
## AVE           Moderate Buy      FRANCE     NYSE        2
## ELN          Moderate Sell     IRELAND     NYSE        2
## MRX           Moderate Buy          US     NYSE        2
## WPI          Moderate Sell          US     NYSE        2
## ABT           Moderate Buy          US     NYSE        3
## AHM             Strong Buy          UK     NYSE        3
## AZN          Moderate Sell          UK     NYSE        3
## BMY          Moderate Sell          US     NYSE        3
## LLY                   Hold          US     NYSE        3
## NVS                   Hold SWITZERLAND     NYSE        3
## SGP                   Hold          US     NYSE        3
## WYE                   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
## BAY                   Hold     GERMANY     NYSE        5
## CHTT          Moderate Buy          US   NASDAQ        5
## IVX                   Hold          US     AMEX        5
```

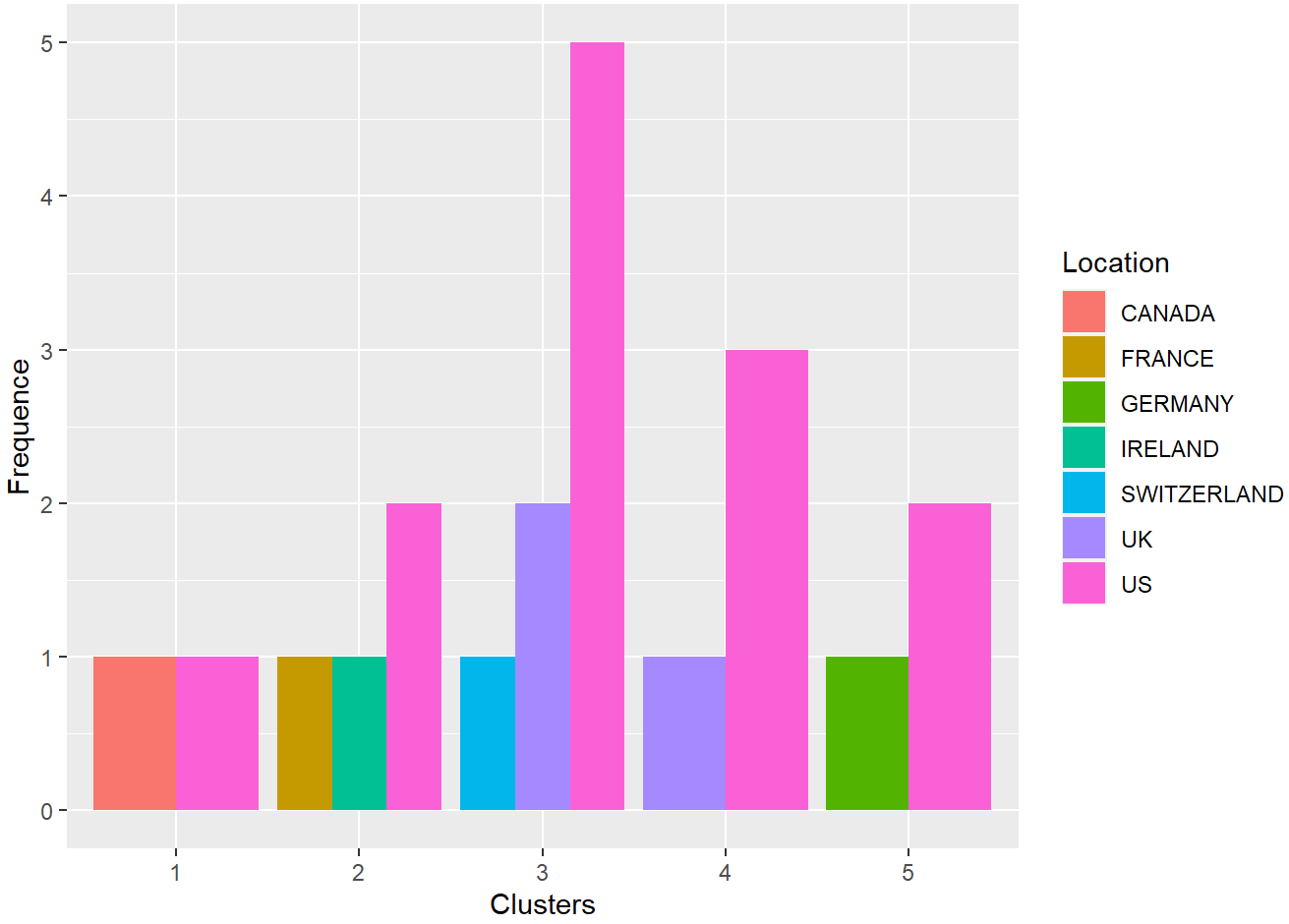
#Cluster 1 - AHM, WYE,BMY,AZN,ABV,SGP,LLY,NVS (low Market_Cap,low Beta,low PE_Ratio,high Leverage,high Rev_Growth.) #Cluster 2 - GSK,JNJ,MRK and PFE (High Market_Cap,ROE, ROA,Asset_Turnover Ratio and low Beta/PE Ratio) #Cluster 3 - AGN,PHA (low Asset_Turnover, High PE Ratio) #Cluster 4 - ELN,MRX,WPI,AVE (low PE_Ratio,high ROE,low ROA,low Net_Profit_Margin, high Rev_Growth) #Cluster 5 - BAY,CHTT,IVX (low Rev_Growth,high Beta and levearge,low Net_Profit_Margin)

#Do the clusters exhibit any patterns in relation to the numerical variables (10–12)? (those not utilized in cluster formation)

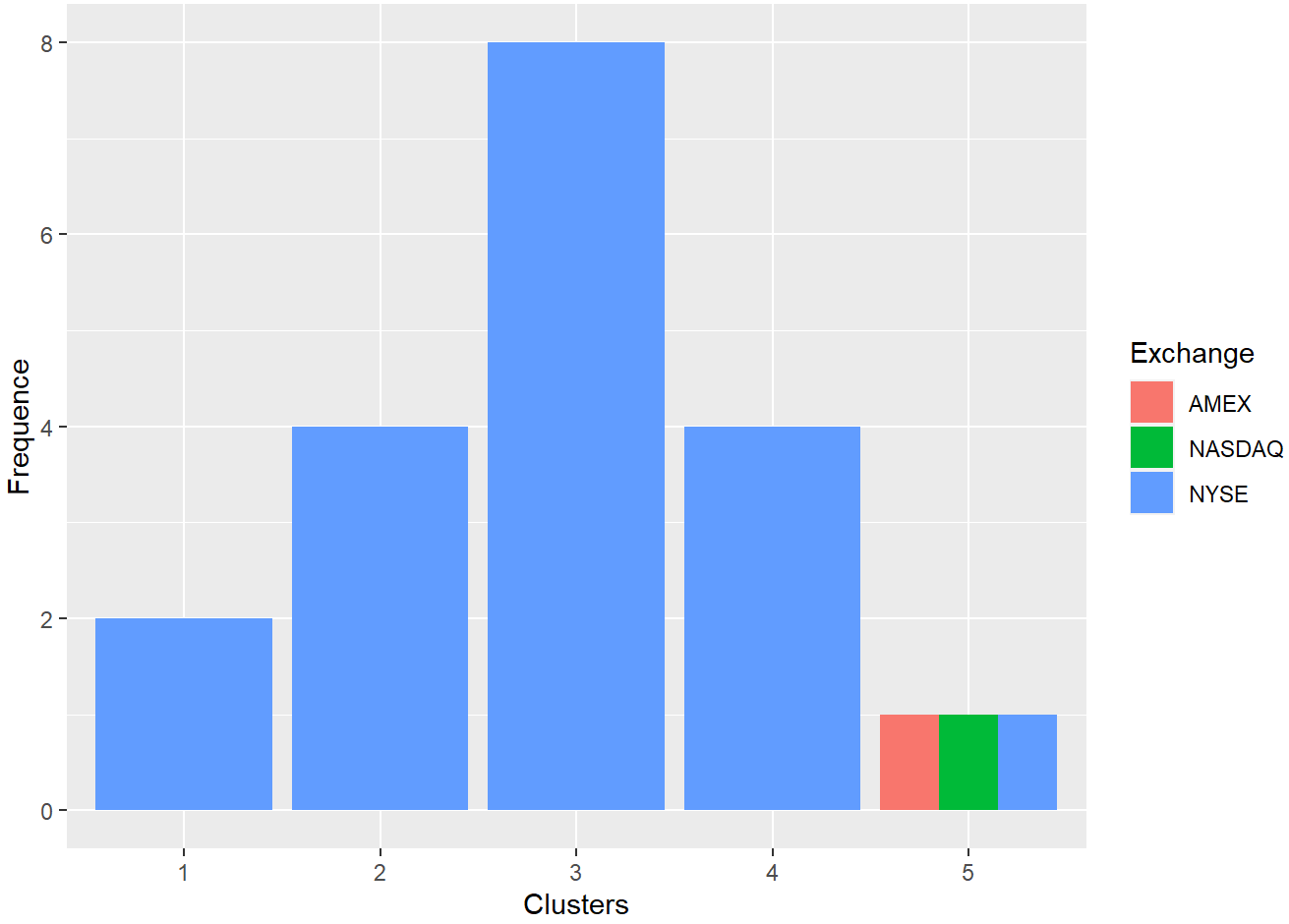
```
Pharma <- Pharmaceuticals[12:14] %>% mutate(Clusters=clus5$cluster)
ggplot(Pharma, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position
='dodge')+labs(x ='Clusters',y ='Frequence')
```




```
ggplot(Pharma, mapping = aes(factor(Clusters),fill = Location))+
  geom_bar(position = 'dodge')+labs(x ='Clusters',y = 'Frequence')
```




```
ggplot(Pharma, mapping = aes(factor(Clusters),fill = Exchange))+geom_bar(position = 'dodge')+
  labs(x ='Clusters',y = 'Frequence')
```



#As can be seen from the graph above, cluster 1 has a moderate level of leverage and little profit. The graph establishes a moderate purchase and hold level. #Because cluster 2 has an equal amount of moderate purchase and sell, there may be room for growth in these companies given their significant revenue growth. #Cluster 3 has a high hold rate and a sizable profit margin, which will entice investors to purchase more of this cluster. #Cluster 4 exhibits both large profit margins and a significant degree of market capitalization. It reflects the company's great potential and has a degree of purchase and hold equal to it. #The graph in cluster 5 illustrates how high debt causes high leverage.

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

#Cluster 1: Perfect Asset #Cluster 2: Investment over the long run #Cluster 3: Dangerous Risk #Cluster 4: potential Growth #Cluster 5: Investment over the short term