FML_Assignment_4

2024-03-14

Cluster Analysis of Pharmaceutical Firms

Introduction

In this analysis, we perform cluster analysis on a dataset containing information about pharmaceutical firms. We focus on using numerical variables (1 to 9) to cluster the 21 firms. 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, are justified.

```
#Importing Required Packages
library(readr)
#Importing Data Set
data <- read_csv("/Users/meghana/Downloads/Pharmaceuticals.csv")
```

```
## Rows: 21 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Symbol, Name, Median_Recommendation, Location, Exchange
## dbl (9): Market_Cap, Beta, PE_Ratio, ROE, ROA, Asset_Turnover, Leverage, Rev...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Load necessary libraries

```
library("ggplot2")
library("factoextra")
```

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

```
library("flexclust")
```

```
## Loading required package: grid
```

```
## Loading required package: lattice
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```r
library("cluster")
library("tidyverse")
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v stringr   1.5.1
## v forcats   1.0.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library("cluster")
```

```r
# Removing null values in data (data cleaning)
Pharma_data =  na.omit(data)
Pharma_data
```

Question(A) 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.

```
## # A tibble: 21 x 14
##    Symbol Name     Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##    <chr>  <chr>         <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
##  1 ABT    Abbott ~      68.4   0.32     24.7  26.4  11.8            0.7     0.42
##  2 AGN    Allerga~       7.58  0.41     82.5  12.9   5.5            0.9     0.6
##  3 AHM    Amersha~       6.3   0.46     20.7  14.9   7.8            0.9     0.27
##  4 AZN    AstraZe~      67.6   0.52     21.5  27.4  15.4            0.9     0
##  5 AVE    Aventis       47.2   0.32     20.1  21.8   7.5            0.6     0.34
##  6 BAY    Bayer AG      16.9   1.11     27.9   3.9   1.4            0.6     0
##  7 BMY    Bristol~      51.3   0.5      13.9  34.8  15.1            0.9     0.57
##  8 CHTT   Chattem~       0.41  0.85     26    24.1   4.3            0.6     3.51
##  9 ELN    Elan Co~       0.78  1.08      3.6  15.1   5.1            0.3     1.07
## 10 LLY    Eli Lil~      73.8   0.18     27.9  31    13.5            0.6     0.53
## # i 11 more rows
## # i 5 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>,
## #   Median_Recommendation <chr>, Location <chr>, Exchange <chr>
```

```r
row.names <- Pharma_data[,1]
pharma_data1 <- Pharma_data[,3:11]#numerical variable from 3 to 11
head(pharma_data1)
```

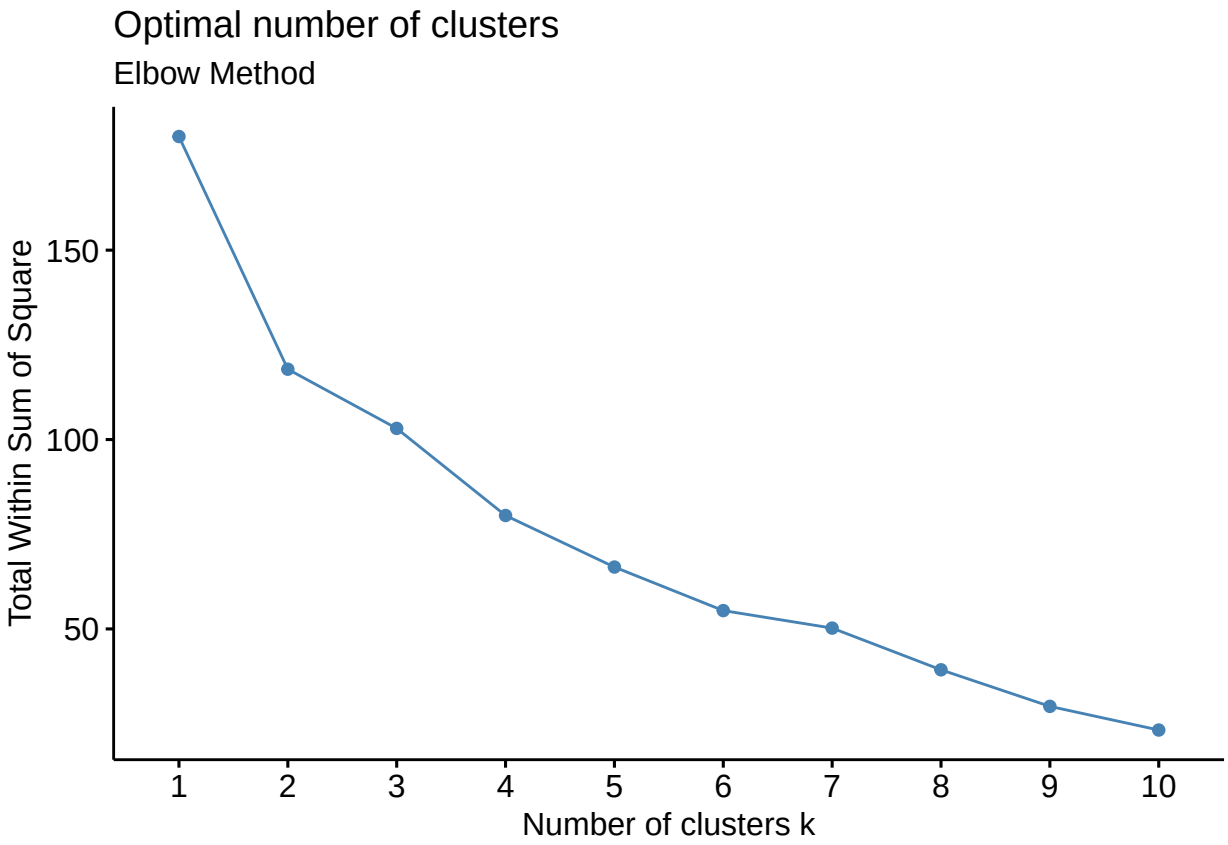
```
## # A tibble: 6 x 9
##   Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage Rev_Growth
##        <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>      <dbl>
```

```
## 1      68.4   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     0.6        9.16
## 3       6.3   0.46     20.7  14.9   7.8           0.9     0.27       7.05
## 4      67.6   0.52     21.5  27.4  15.4           0.9     0          15
## 5      47.2   0.32     20.1  21.8   7.5           0.6     0.34       26.8
## 6      16.9   1.11     27.9   3.9   1.4           0.6     0          -3.17
## # i 1 more variable: Net_Profit_Margin <dbl>
```

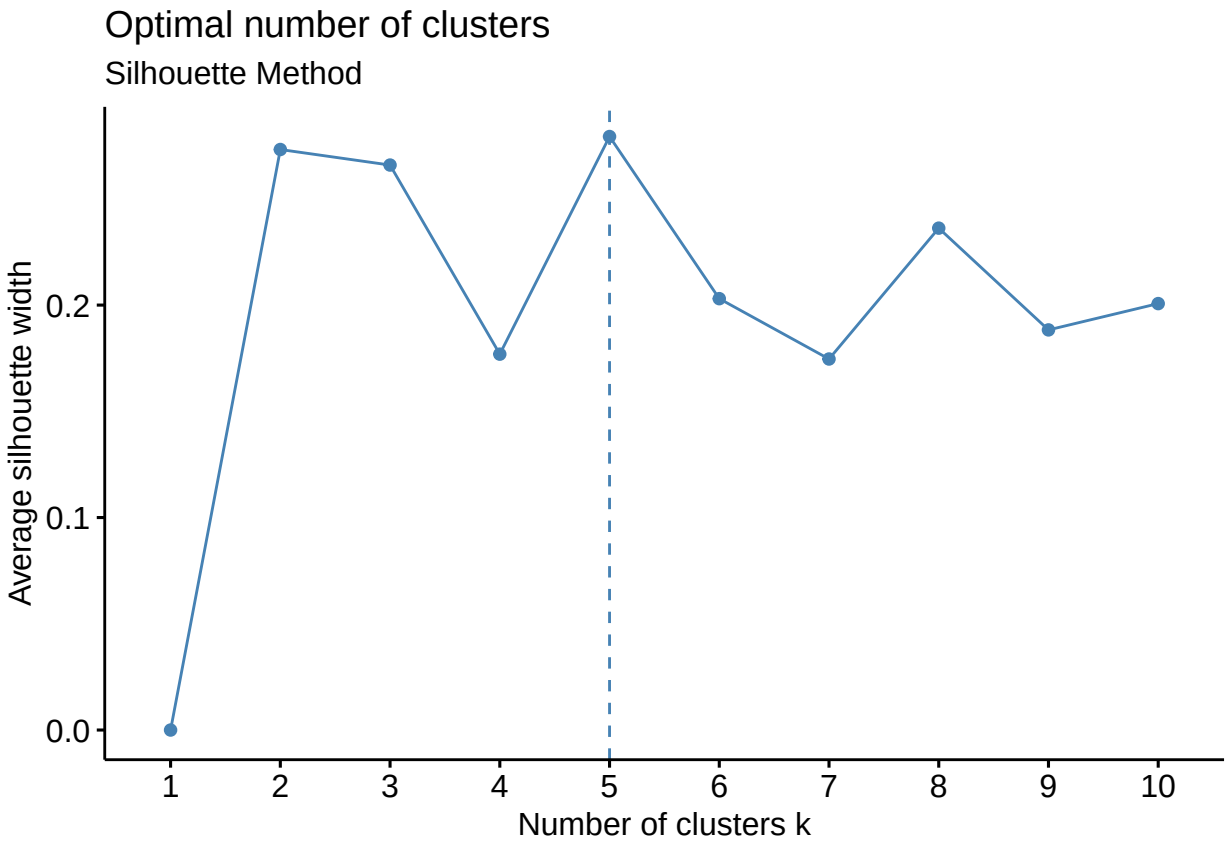
```
pharma_data2 <- scale(pharma_data1)
head(pharma_data2)
```

```
##      Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## [1,]  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121  -5.121077e-16
## [2,] -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871   9.225312e-01
## [3,] -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700   9.225312e-01
## [4,]  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259   9.225312e-01
## [5,] -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461  -4.612656e-01
## [6,] -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612  -4.612656e-01
##        Leverage Rev_Growth Net_Profit_Margin
## [1,] -0.2120979 -0.5277675        0.06168225
## [2,]  0.0182843 -0.3811391       -1.55366706
## [3,] -0.4040831 -0.5721181       -0.68503583
## [4,] -0.7496565  0.1474473        0.35122600
## [5,] -0.3144900  1.2163867       -0.42597037
## [6,] -0.7496565 -1.4971443       -1.99560225
```

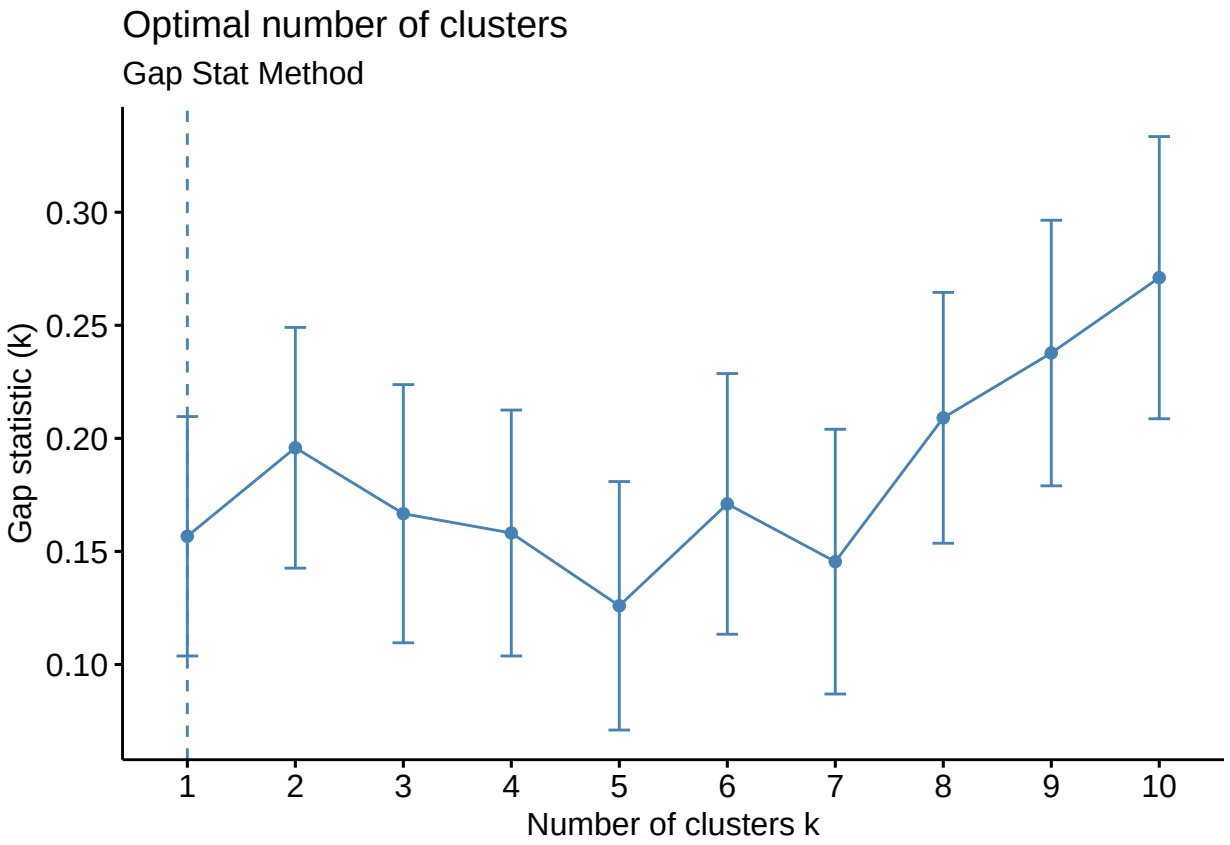
```
#Determination of Number of Clusters
#We determine the optimal number of clusters using different methods such as the Elbow Method, Silhouet
fviz_nbclust(pharma_data2, kmeans, method = "wss") +labs(subtitle = "Elbow Method")
```




```
fviz_nbclust(pharma_data2, kmeans, method = "silhouette") + labs(subtitle = "Silhouette Method")
```



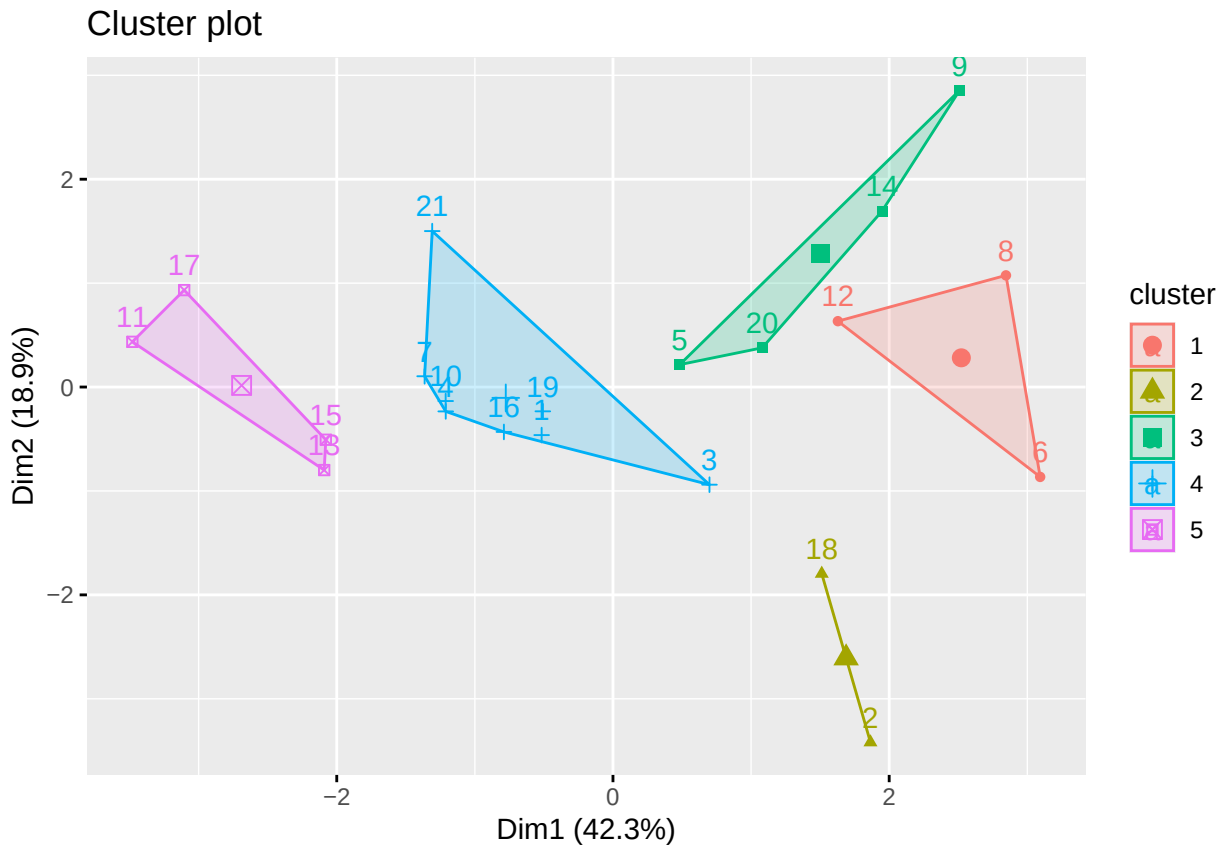

```
fviz_nbclust(pharma_data2, kmeans, method = "gap_stat") + labs(subtitle = "Gap Stat Method")
```




```
set.seed(64060)
k_5 <- kmeans(pharma_data2, centers = 5, nstart = 25)
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)
```




```
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:
##  [1] 4 2 4 4 3 1 4 1 3 4 5 1 5 3 5 4 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 = "euclidian")
#fvi_dist(distance)
```

```
FITT <- kmeans(pharma_data2,5)
aggregate(pharma_data2,by = list(FITT$cluster), FUN = mean)
```

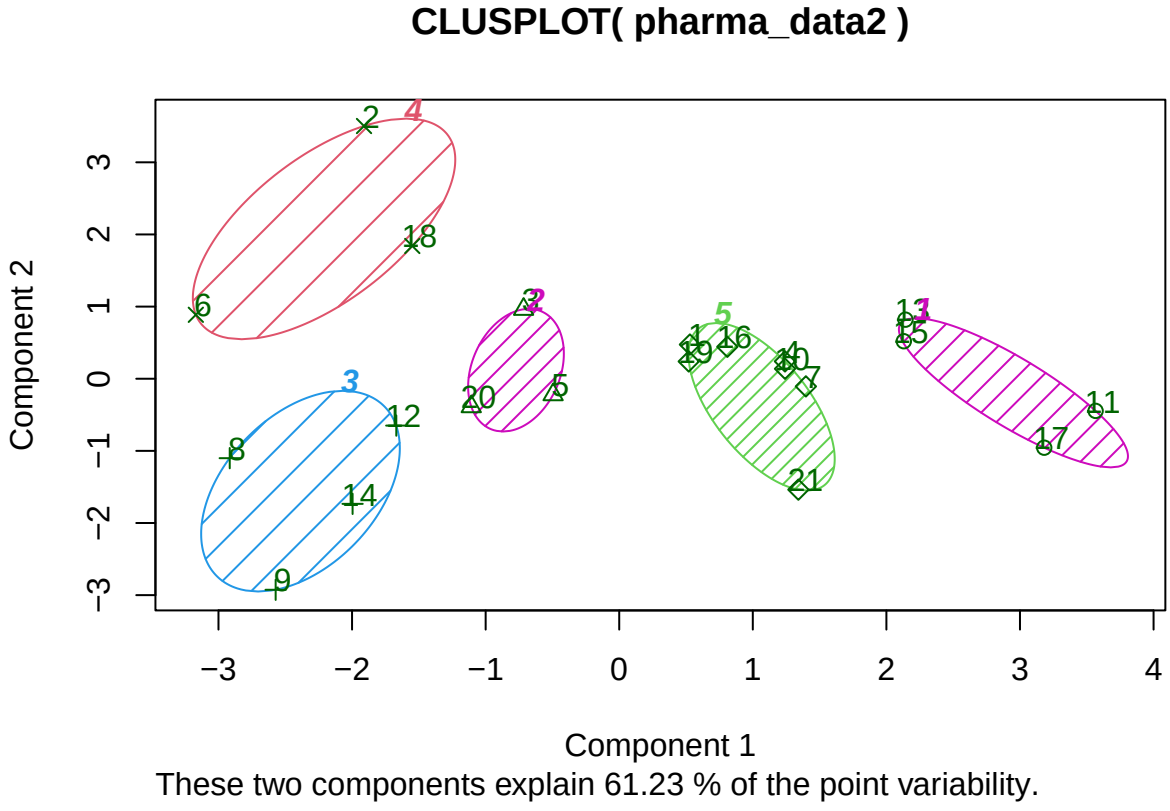
```
##   Group.1  Market_Cap       Beta   PE_Ratio        ROE        ROA
## 1       1  1.69558112 -0.1780563 -0.1984582  1.2349879  1.3503431
## 2       2 -0.66114002 -0.7233539 -0.3512251 -0.6736441 -0.5915022
## 3       3 -0.96247577  1.1949250 -0.3639982 -0.5200697 -0.9610792
## 4       4 -0.52462814  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  -3.330669e-16 -0.3443544 -0.5769454        -1.6095439
## 5   6.589509e-02 -0.2559803 -0.7230135         0.7343816
```

```
pharma_data3 <- data.frame(pharma_data2,FITT$cluster)
pharma_data3
```

```
##    Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## 1   0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121  -5.121077e-16
## 2  -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871   9.225312e-01
## 3  -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700   9.225312e-01
## 4   0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259   9.225312e-01
## 5  -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461  -4.612656e-01
## 6  -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612  -4.612656e-01
## 7  -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498   9.225312e-01
## 8  -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918  -4.612656e-01
## 9  -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553  -1.845062e+00
## 10  0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770  -4.612656e-01
## 11  1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364   1.383797e+00
## 12 -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905  -4.612656e-01
## 13  1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544   9.225312e-01
## 14 -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792  -1.845062e+00
## 15  1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577   1.845062e+00
## 16  0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598  -9.225312e-01
## 17  2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239   4.612656e-01
## 18 -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030  -4.612656e-01
## 19 -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929   4.612656e-01
## 20 -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905  -9.225312e-01
## 21 -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849  -4.612656e-01
##       Leverage  Rev_Growth Net_Profit_Margin FITT.cluster
## 1  -0.21209793 -0.52776752        0.06168225            5
## 2   0.01828430 -0.38113909       -1.55366706            4
## 3  -0.40408312 -0.57211809       -0.68503583            2
```

```
## 4  -0.74965647  0.14744734        0.35122600           5
## 5  -0.31449003  1.21638667       -0.42597037           2
## 6  -0.74965647 -1.49714434       -1.99560225           4
## 7  -0.02011273 -0.96584257        0.74744375           5
## 8   3.74279705 -0.63276071       -1.24888417           3
## 9   0.61983791  1.88617085       -0.36501379           3
## 10 -0.07130879 -0.64814764        1.17413980           5
## 11 -0.31449003  0.76926048        0.82363947           1
## 12  1.10620040  0.05603085       -0.71551412           3
## 13 -0.62166634 -0.36213170        0.33598685           1
## 14  0.44065173  1.53860717        0.85411776           3
## 15 -0.39128411  0.36014907       -0.24310064           1
## 16 -0.67286239 -1.45369888        1.02174835           5
## 17 -0.54487226  1.10143723        1.44844440           1
## 18 -0.30169102  0.14744734       -1.27936246           4
## 19 -0.74965647 -0.43544591        0.29026942           5
## 20 -0.49367621  1.43089863       -0.09070919           2
## 21  0.68383297 -1.17763919        1.49416183           5
```

```
clusplot(pharma_data2,FITT$cluster, color = TRUE, shade = TRUE,
 labels = 2,
 lines = 0)
```



```
aggregate(pharma_data2, by = list(FITT$cluster), FUN = mean)
```

Question(B) Interpret the clusters with respect to the numerical variables used in forming the clusters.

```
##   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  -3.330669e-16 -0.3443544 -0.5769454        -1.6095439
## 5   6.589509e-02 -0.2559803 -0.7230135         0.7343816
```

```
Pharmacy <- data.frame(pharma_data2,k_5$cluster)
Pharmacy
```

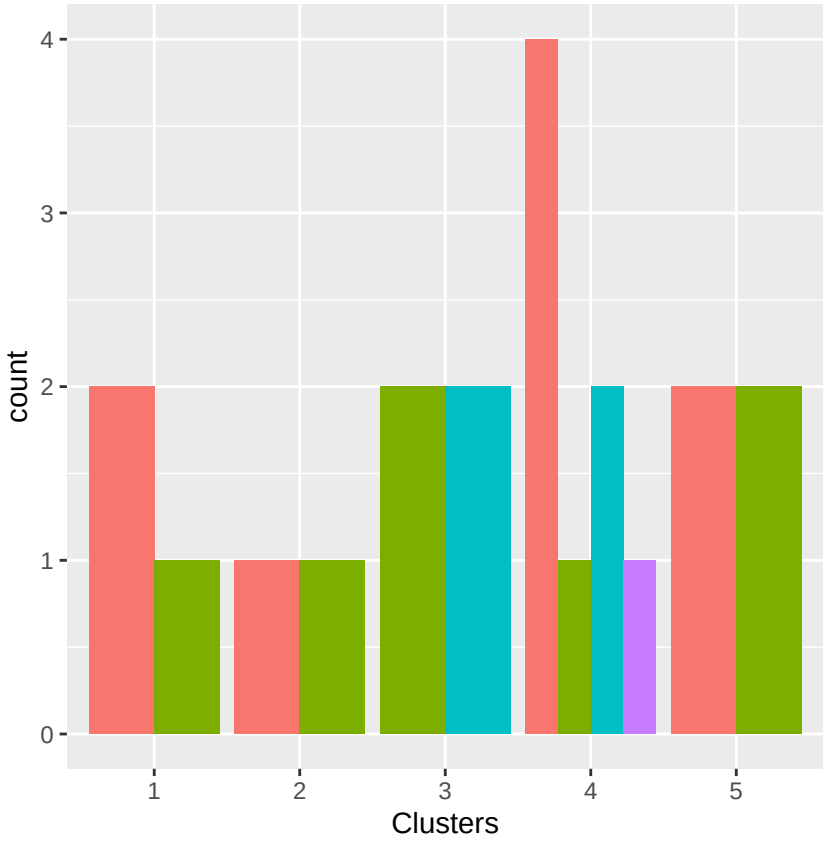
```
##    Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## 1   0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121  -5.121077e-16
## 2  -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871   9.225312e-01
## 3  -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700   9.225312e-01
## 4   0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259   9.225312e-01
## 5  -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461  -4.612656e-01
## 6  -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612  -4.612656e-01
## 7  -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498   9.225312e-01
## 8  -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918  -4.612656e-01
## 9  -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553  -1.845062e+00
## 10  0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770  -4.612656e-01
## 11  1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364   1.383797e+00
## 12 -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905  -4.612656e-01
## 13  1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544   9.225312e-01
## 14 -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792  -1.845062e+00
## 15  1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577   1.845062e+00
## 16  0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598  -9.225312e-01
## 17  2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239   4.612656e-01
## 18 -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030  -4.612656e-01
## 19 -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929   4.612656e-01
## 20 -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905  -9.225312e-01
## 21 -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849  -4.612656e-01
##       Leverage  Rev_Growth Net_Profit_Margin k_5.cluster
## 1  -0.21209793 -0.52776752        0.06168225           4
## 2   0.01828430 -0.38113909       -1.55366706           2
## 3  -0.40408312 -0.57211809       -0.68503583           4
## 4  -0.74965647  0.14744734        0.35122600           4
## 5  -0.31449003  1.21638667       -0.42597037           3
## 6  -0.74965647 -1.49714434       -1.99560225           1
## 7  -0.02011273 -0.96584257        0.74744375           4
```

```
## 8   3.74279705 -0.63276071       -1.24888417          1
## 9   0.61983791  1.88617085       -0.36501379          3
## 10 -0.07130879 -0.64814764        1.17413980          4
## 11 -0.31449003  0.76926048        0.82363947          5
## 12  1.10620040  0.05603085       -0.71551412          1
## 13 -0.62166634 -0.36213170        0.33598685          5
## 14  0.44065173  1.53860717        0.85411776          3
## 15 -0.39128411  0.36014907       -0.24310064          5
## 16 -0.67286239 -1.45369888        1.02174835          4
## 17 -0.54487226  1.10143723        1.44844440          5
## 18 -0.30169102  0.14744734       -1.27936246          2
## 19 -0.74965647 -0.43544591        0.29026942          4
## 20 -0.49367621  1.43089863       -0.09070919          3
## 21  0.68383297 -1.17763919        1.49416183          4
```

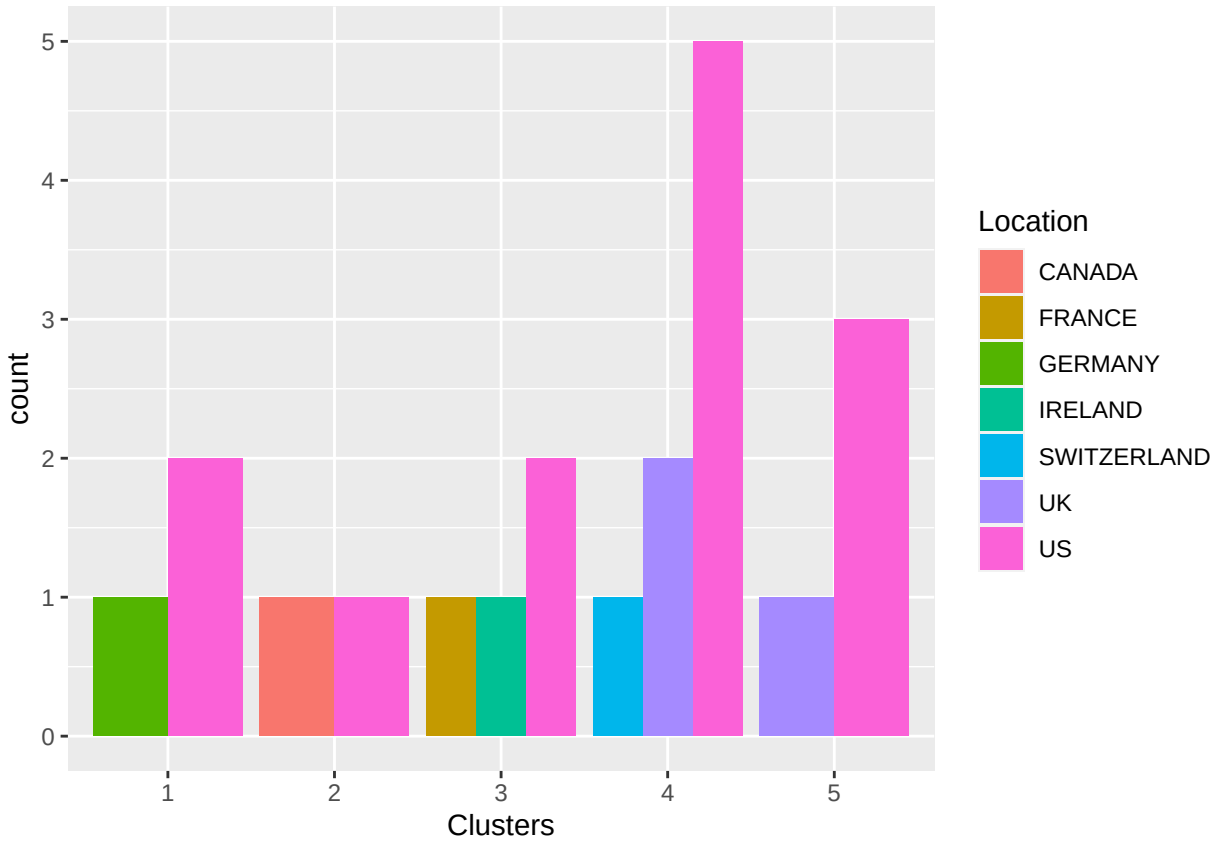
```
# CLuster 1: JNJ, MRK, GSK, PFE (lowest beta/PE ratio and highest market cap)
#Cluster 2: AHM, WPI, AVE (lowest PE/Asset Turnover Ratio and highest revenue growth)
# Cluster 3: CHTT, IVX, MRX, ELN (lowest Net Profit Margin, PE ratio, and Marke#Cluster, and highest be
#Cluster 4: AGN, BAY, PHA (lowest leverage/asset turnover and highest PE ratio)
#BT, WYE, AZN, SGP, BMY, NVS, LLY are in Cluster 5 (highest net profit margin and #lowest leverage).
```

```
Pharma <- data[12:14] %>% mutate(Clusters=k_5$cluster)
ggplot(Pharma, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position='dodge')
```

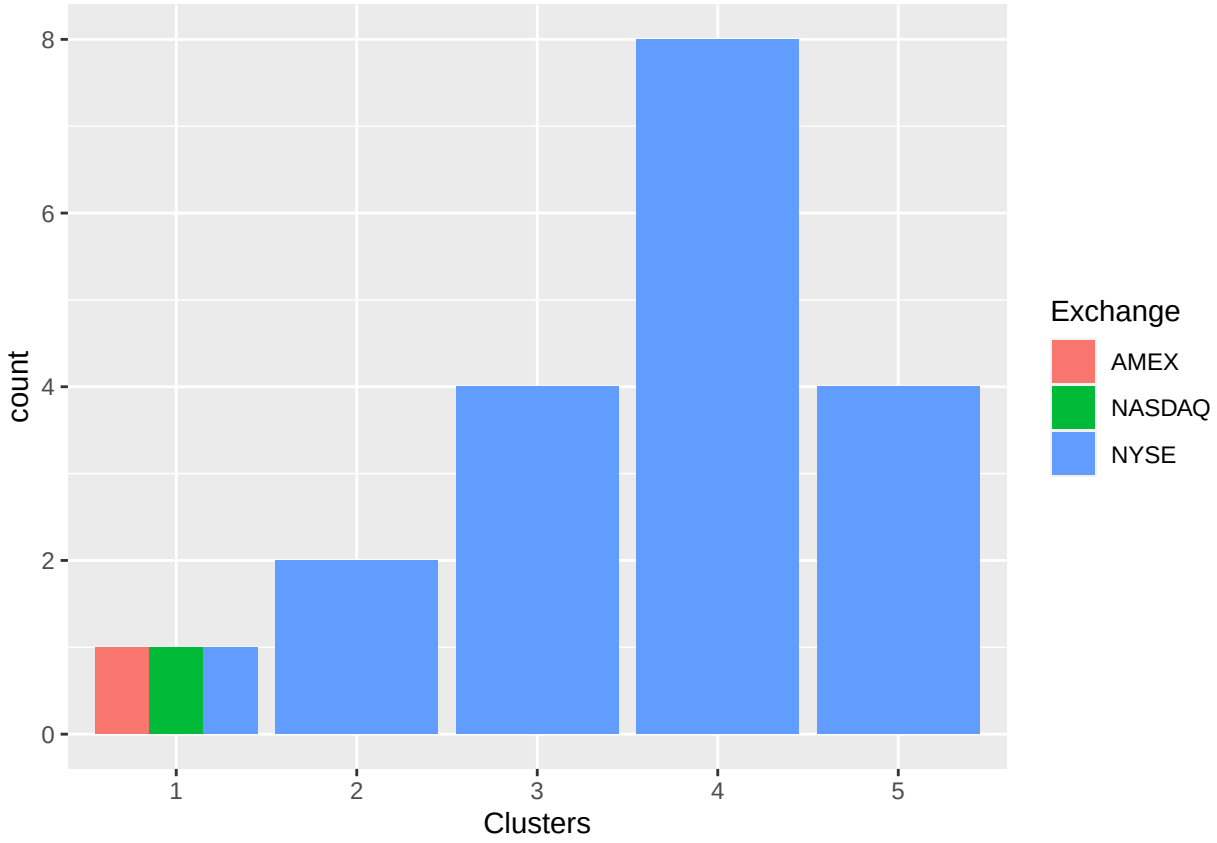
Question(C) Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)




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




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



Interpretation :

The clusters from the graphs above show a slight pattern that we can observe #### The businesses in cluster 1 are evenly distributed across the AMEX, NASDAQ, and NYSE, but it has distinct Hold and Moderate Buy medians and a different count from the US and Germany.

The medians for holds and moderate purchases are distributed similarly in Cluster 2.

The NYSE lists stocks from both the US and Canada.

Cluster 3 differs from Cluster in count, but its Moderate Buy and Sell medians are comparable.

The NYSE lists France, Ireland, and the US.

Hold, Moderate Buy, Moderate Sell, and Strong Buy options are available in Cluster 4.

The hold's median is the highest. They are listed on the NYSE and are citizens of the United States, the United Kingdom, and Switzerland.

Cluster 5 is spread across

countries, including the US and the UK, and is listed on the NYSE. It also has the same hold and median purchase values.

```
#Naming clusters
#After performing cluster analysis on the pharmaceutical firms dataset,Assigning  descriptive names to

#Cluster 1 :- Buy Cluster
#Cluster 2 :- Sceptical Cluster
#Cluster 3 :- Moderate Buy Cluster
#Cluster 4 :- Hold Cluster
#Cluster 5 :- High Hold Cluster
```

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