FML-Clustering Assignment 4

Dhanush Myneni

2023-11-12

**CODE FOR ASSIGNMENT 4-CLUSTERING ANALYSIS

#Summary The assignment 4 of the Machine learning course deals with the clustering analysis. For this assignment, we are using the pharmaceuticals dataset which is a brief description on 21 pharmaceutical firms and and provides details on 9 key performance indicators such as the market cap, ROA, ROE etc.

For the first question we have used the 9 numeric columns of data to do the clustering analysis.We are using the K means clustering algorithm. This algorithm scales the data meaning that, it pre-processes the data to ensure that alla the variables contribute equally,calculates the pairwise distances between observations.K-means clustering partitions data into k clusters by minimizing within-cluster variation.In order to find out the optimal number of clusters to be formed we have used fviz_nbclust function. Fviz_nbclust with the silhouette method is a helpful tool to determine the optimal number of clusters. It calculates silhouette scores for different cluster numbers and helps us to identify the number of clusters that maximizes the separation between clusters while minimizing the overlap. After doing so we found out that the optimal number of clusters to be formed are 5.We later have formed a clustering analysis for K value 7. However, within cluster sum of square value when K is 7 is 77.5% whereas it is 65.4% when the K value is 5. Lower WCSS values generally indicate better-defined clusters which is why we decided upon the optimal number of clusters to be 5 rather than 7. Since we are using the K means algorithm, it treats all the variables equally during the clustering process. The "centers" returned by the kmeans function represent the mean values for each variable within each cluster, and these means collectively define the centroids of the clusters.

For the second question, there are some patterns relating the non-numeric variables to the numeric variable based clusters. Cluster 1 and 3 seem to be more "moderate", while 4 and 5 are more "extreme" in some variables.

Clusters 1 and 3 seem more "moderate" in the sense that:

They have moderate growth (Cluster 1 has high valuation/profitability but lower growth, Cluster 3 has high PE but lower profitability).Their recommendations are more middle-of-the-road (hold/moderate buy rather than strong buy/sell).They are listed on major exchanges (NYSE) and located in mature markets (US/UK).

In contrast, Clusters 4 and 5 seem more "extreme": Cluster 4 has high growth but also high risk (high leverage, low PE, moderate sell recs).Cluster 5 has distressed stocks with low growth, high volatility (beta), and high leverage.

```
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)
```

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

```
library(tidyverse)
```

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

## Warning: package 'ggplot2' was built under R version 4.3.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## 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
```

```
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
```

```
setwd("C:/Users/dhanu/OneDrive/Desktop/FML")

Pharmaceuticals <- read.csv("C:/Users/dhanu/OneDrive/Desktop/FML/Pharmaceuticals.csv")
```

```r
summary(Pharmaceuticals)  #Provides a summary for the pharmaceuticals data.
```

```
##     Symbol              Name             Market_Cap          Beta
##  Length:21          Length:21          Min.   :  0.41   Min.   :0.1800
##  Class :character   Class :character   1st Qu.:  6.30   1st Qu.:0.3500
##  Mode  :character   Mode  :character   Median : 48.19   Median :0.4600
##                                        Mean   : 57.65   Mean   :0.5257
##                                        3rd Qu.: 73.84   3rd Qu.:0.6500
##                                        Max.   :199.47   Max.   :1.1100
##     PE_Ratio           ROE             ROA        Asset_Turnover    Leverage
##  Min.   : 3.60   Min.   : 3.9   Min.   : 1.40   Min.   :0.3    Min.   :0.0000
##  1st Qu.:18.90   1st Qu.:14.9   1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600
##  Median :21.50   Median :22.6   Median :11.20   Median :0.6    Median :0.3400
##  Mean   :25.46   Mean   :25.8   Mean   :10.51   Mean   :0.7    Mean   :0.5857
##  3rd Qu.:27.90   3rd Qu.:31.0   3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000
##  Max.   :82.50   Max.   :62.9   Max.   :20.30   Max.   :1.1    Max.   :3.5100
##    Rev_Growth    Net_Profit_Margin Median_Recommendation   Location
##  Min.   :-3.17   Min.   : 2.6      Length:21             Length:21
##  1st Qu.: 6.38   1st Qu.:11.2      Class :character      Class :character
##  Median : 9.37   Median :16.1      Mode  :character      Mode  :character
##  Mean   :13.37   Mean   :15.7
##  3rd Qu.:21.87   3rd Qu.:21.1
##  Max.   :34.21   Max.   :25.5
##    Exchange
##  Length:21
##  Class :character
##  Mode  :character
##
##
##
```

```r
head(Pharmaceuticals)     #Gives an insight on how the data looks like.
```

```
##   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
```

```r
Pharmaceuticals_1 <- Pharmaceuticals[3:11] #Selecting only the numerical variables 1 to 9
head(Pharmaceuticals_1)
```

```
##   Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
```

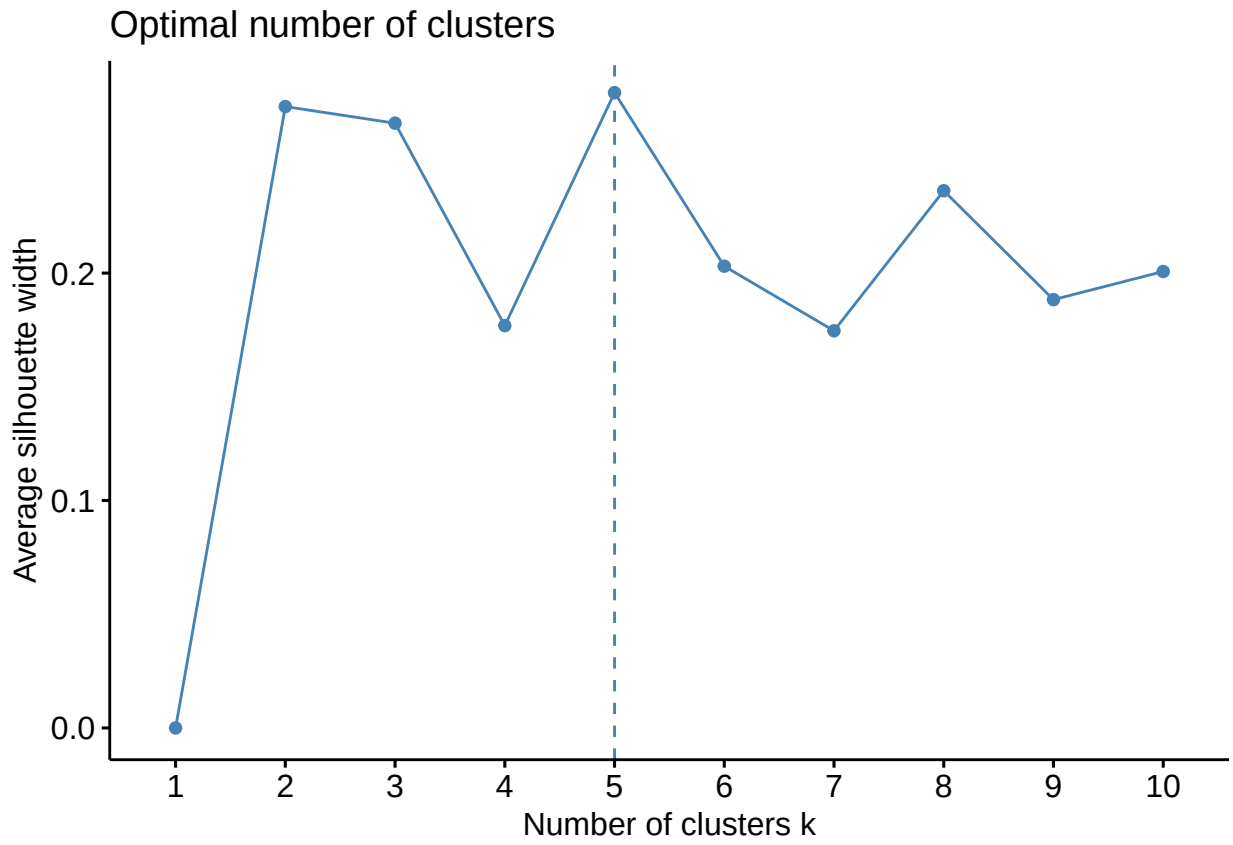
```
## 1      68.44 0.32     24.7 26.4 11.8           0.7     0.42       7.54
## 2       7.58 0.41     82.5 12.9  5.5           0.9     0.60       9.16
## 3       6.30 0.46     20.7 14.9  7.8           0.9     0.27       7.05
## 4      67.63 0.52     21.5 27.4 15.4           0.9     0.00      15.00
## 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
```

```
summary(Pharmaceuticals_1)
```

```
##    Market_Cap          Beta            PE_Ratio          ROE
##  Min.   :  0.41   Min.   :0.1800   Min.   : 3.60   Min.   : 3.9
##  1st Qu.:  6.30   1st Qu.:0.3500   1st Qu.:18.90   1st Qu.:14.9
##  Median : 48.19   Median :0.4600   Median :21.50   Median :22.6
##  Mean   : 57.65   Mean   :0.5257   Mean   :25.46   Mean   :25.8
##  3rd Qu.: 73.84   3rd Qu.:0.6500   3rd Qu.:27.90   3rd Qu.:31.0
##  Max.   :199.47   Max.   :1.1100   Max.   :82.50   Max.   :62.9
##       ROA        Asset_Turnover    Leverage        Rev_Growth
##  Min.   : 1.40   Min.   :0.3    Min.   :0.0000   Min.   :-3.17
##  1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600   1st Qu.: 6.38
##  Median :11.20   Median :0.6    Median :0.3400   Median : 9.37
##  Mean   :10.51   Mean   :0.7    Mean   :0.5857   Mean   :13.37
##  3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000   3rd Qu.:21.87
##  Max.   :20.30   Max.   :1.1    Max.   :3.5100   Max.   :34.21
##  Net_Profit_Margin
##  Min.   : 2.6
##  1st Qu.:11.2
##  Median :16.1
##  Mean   :15.7
##  3rd Qu.:21.1
##  Max.   :25.5
```

#Next we proceed onto normalizing the data.

```
Pharmaceuticals_2 <- scale(Pharmaceuticals_1)
row.names(Pharmaceuticals_2) <- Pharmaceuticals[,1]
distance <- get_dist(Pharmaceuticals_2)
Co_relation <- cor(Pharmaceuticals_2)
fviz_nbclust(Pharmaceuticals_2,kmeans,method = "silhouette")
```




```
set.seed(69)
K5 <- kmeans(Pharmaceuticals_2,centers = 5,nstart = 25) #k=5 and number of restarts are 25
print(K5)
```

```
## K-means clustering with 5 clusters of sizes 4, 8, 2, 4, 3
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 2 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 3 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 4 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 5 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.46807818  0.4671788       0.591242521
## 2 -0.27449312 -0.7041516       0.556954446
## 3 -0.14170336 -0.1168459      -1.416514761
## 4  0.06308085  1.5180158      -0.006893899
## 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
##    2    3    2    2    4    5    2    5    4    2    1    5    1    4    1    2
##  PFE  PHA  SGP  WPI  WYE
##    1    3    2    4    2
```

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

```
K5$centers
```

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

```
K5$size
```

```
## [1] 4 8 2 4 3
```

```
K7 <- kmeans(Pharmaceuticals_2,centers = 7,nstart = 25) #k=7 and number of restarts are 25
print(K7)
```

```
## K-means clustering with 7 clusters of sizes 4, 2, 4, 1, 7, 2, 1
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431     1.15316401
## 2 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951     0.23063280
## 3 -0.73070420 -0.4214928 -0.34867046 -0.5780744 -0.6181243    -0.23063280
## 4 -0.69538175  2.2757827  0.14948233 -1.4514600 -1.7127612    -0.46126560
## 5  0.08926902 -0.4618336 -0.32086149  0.3260892  0.5396003     0.06589509
## 6 -0.96686975  1.5162611 -0.57398880 -0.8382671 -0.9892673    -1.84506242
## 7 -0.97676686  1.2630872  0.03299122 -0.1123792 -1.1677918    -0.46126560
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.46807818  0.4671788        0.5912425
## 2 -0.14170336 -0.1168459       -1.4165148
## 3 -0.02651224  0.5327995       -0.4793074
## 4 -0.74965647 -1.4971443       -1.9956023
## 5 -0.25598026 -0.7230135        0.7343816
## 6  0.53024482  1.7123890        0.2445520
## 7  3.74279705 -0.6327607       -1.2488842
##
```

```
## Clustering vector:
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    5    2    3    5    3    4    5    7    6    5    1    3    1    6    1    5
##  PFE  PHA  SGP  WPI  WYE
##    1    2    5    3    5
##
## Within cluster sum of squares by cluster:
## [1]  9.284424  2.803505  8.940101  0.000000 16.655937  2.855389  0.000000
##  (between_SS / total_SS =  77.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
K7$centers
```

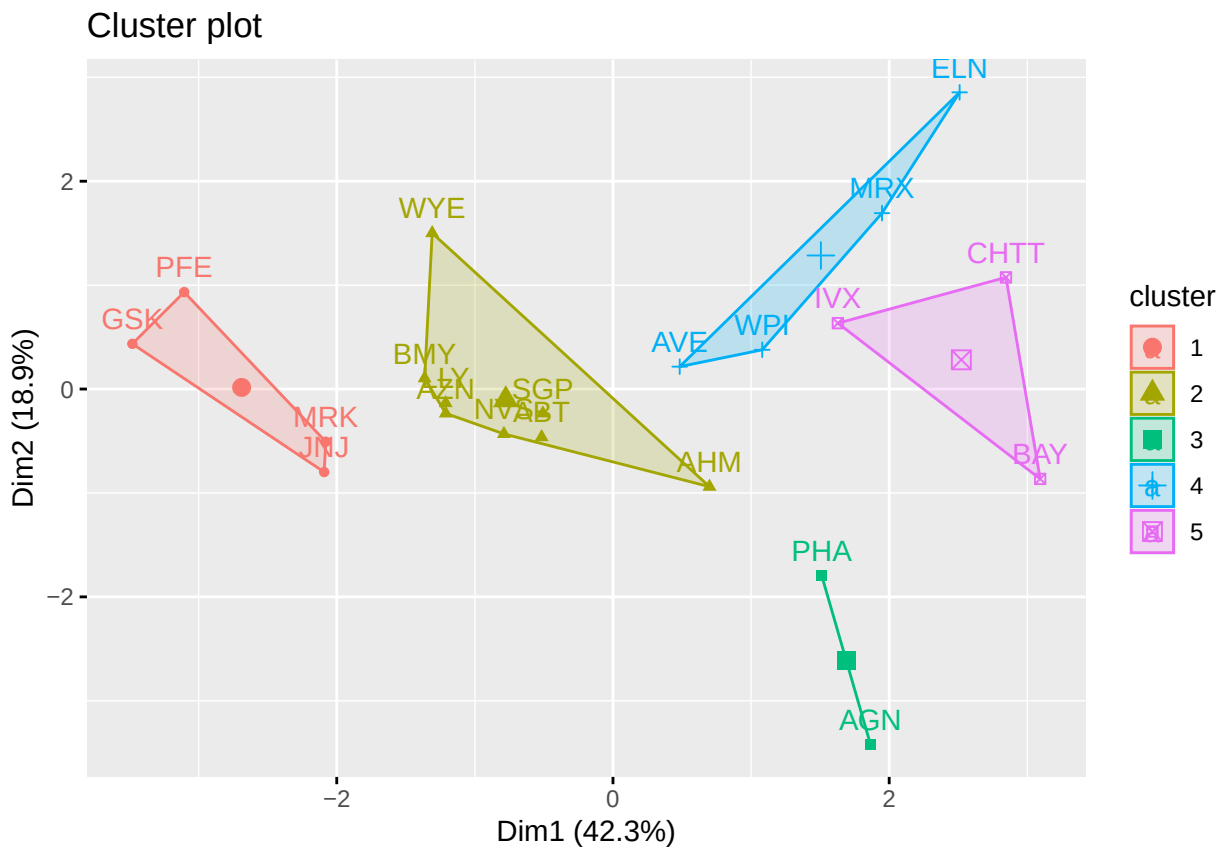
```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431     1.15316401
## 2 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951     0.23063280
## 3 -0.73070420 -0.4214928 -0.34867046 -0.5780744 -0.6181243    -0.23063280
## 4 -0.69538175  2.2757827  0.14948233 -1.4514600 -1.7127612    -0.46126560
## 5  0.08926902 -0.4618336 -0.32086149  0.3260892  0.5396003     0.06589509
## 6 -0.96686975  1.5162611 -0.57398880 -0.8382671 -0.9892673    -1.84506242
## 7 -0.97676686  1.2630872  0.03299122 -0.1123792 -1.1677918    -0.46126560
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.46807818  0.4671788         0.5912425
## 2 -0.14170336 -0.1168459        -1.4165148
## 3 -0.02651224  0.5327995        -0.4793074
## 4 -0.74965647 -1.4971443        -1.9956023
## 5 -0.25598026 -0.7230135         0.7343816
## 6  0.53024482  1.7123890         0.2445520
## 7  3.74279705 -0.6327607        -1.2488842
```

```
K7$size
```

```
## [1] 4 2 4 1 7 2 1
```

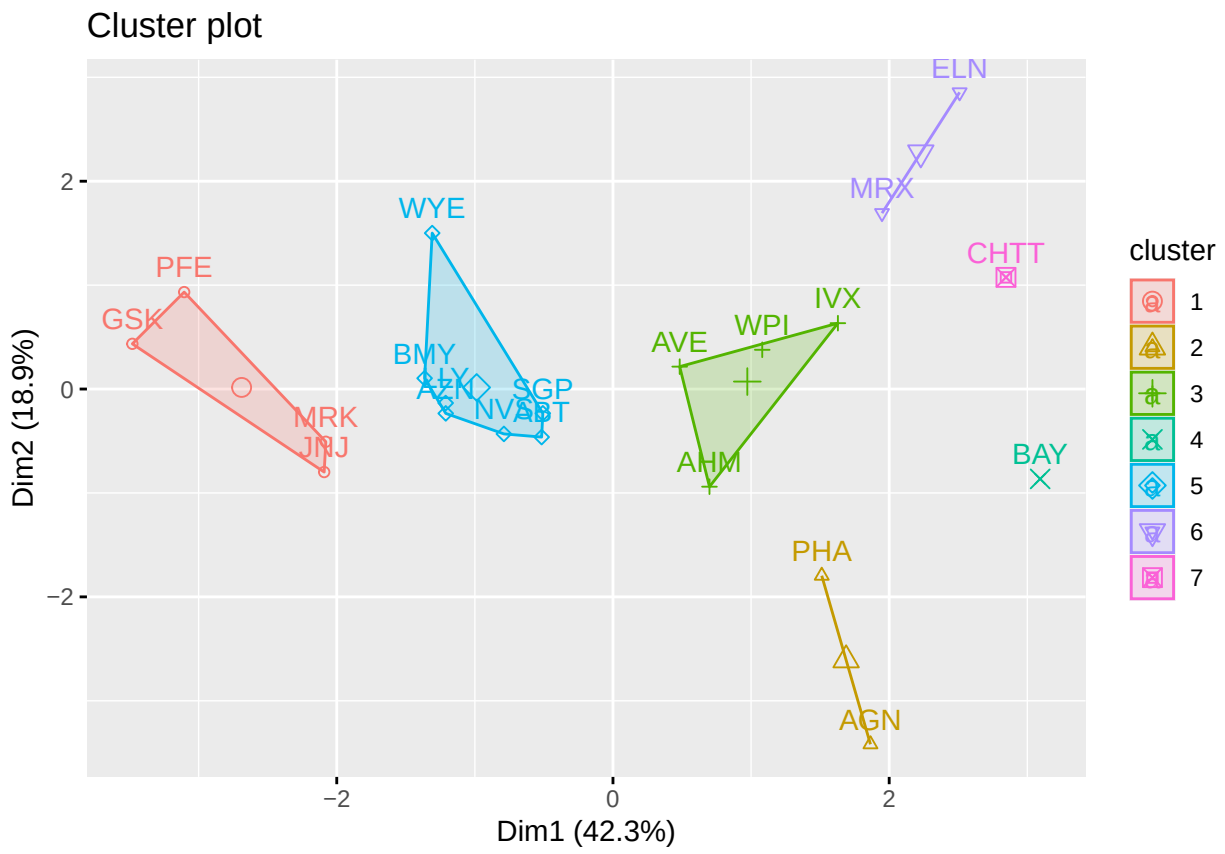
#Forming the clusters based on the value K=5.

```
fviz_cluster(K5, data = Pharmaceuticals_2)
```



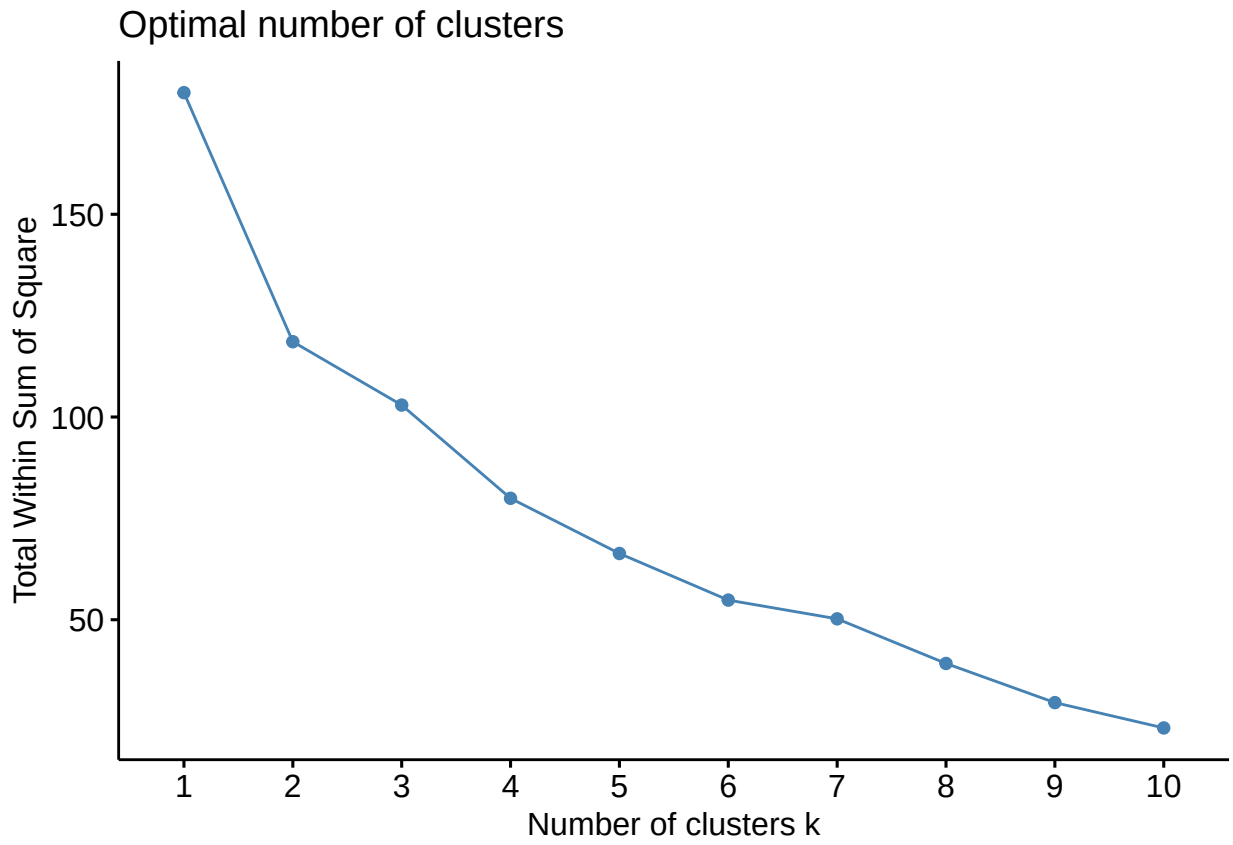
#Forming the clusters based on the value K=7

```
fviz_cluster(K7, data = Pharmaceuticals_2)
```



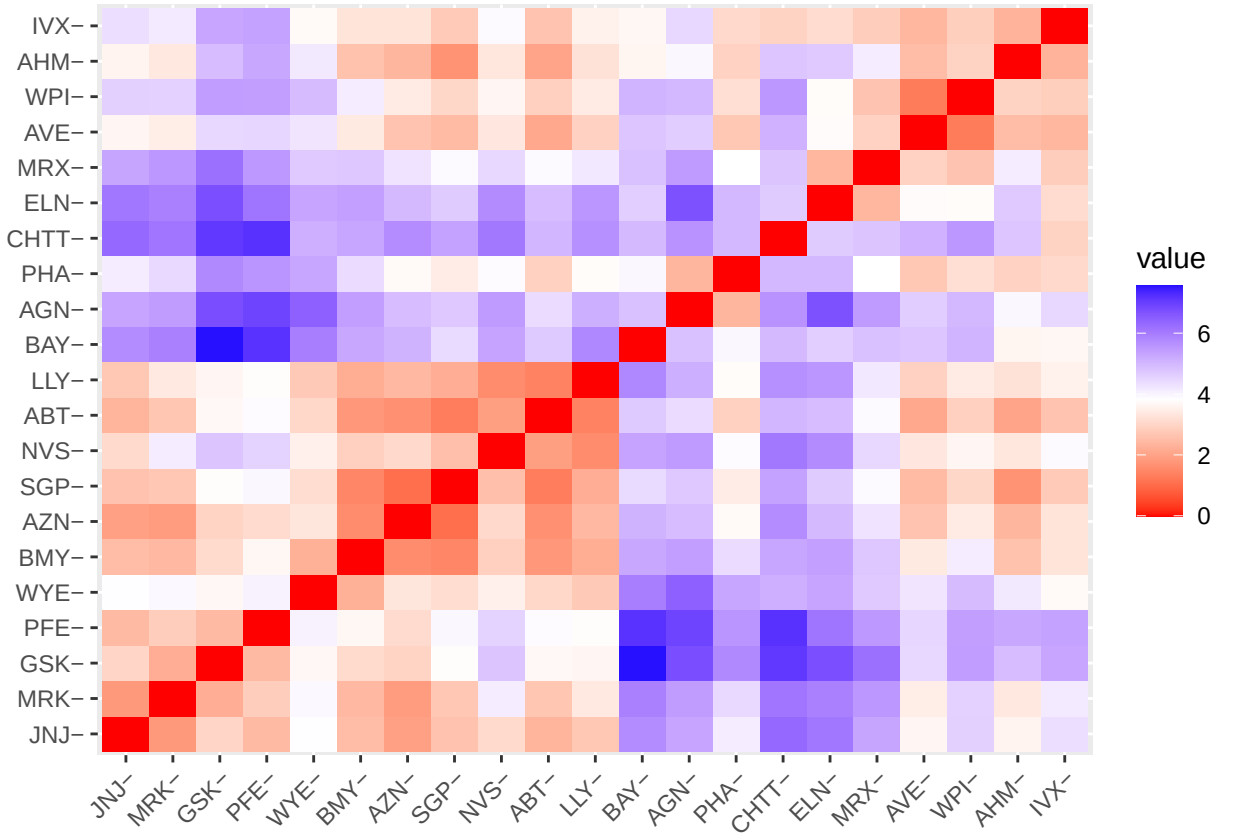
#Finding the optimal number of clusters using the elbow method.

```
fviz_nbclust(Pharmaceuticals_2,kmeans,method = "wss")
```



#Euclidean

```
distance<- dist(Pharmaceuticals_2, method = "euclidean")
fviz_dist(distance)
```



#Manhattan

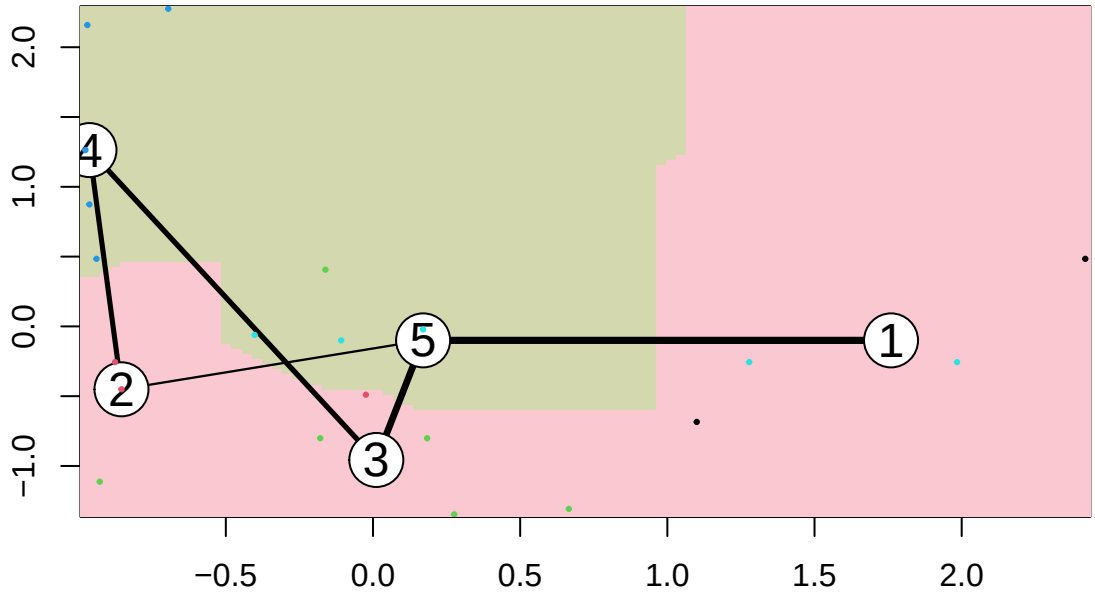
```
set.seed(69)

k5.1 = kcca(Pharmaceuticals_2, k=5, kccaFamily("kmedians"))
k5.1
```

```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = Pharmaceuticals_2, k = 5, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
## 1 2 3 4 5
## 2 3 6 5 5
```

```
Clusters_Index <- predict(k5.1)
dist(k5.1@centers)
```

```
##          1        2        3        4
## 2 5.796625
## 3 3.847926 3.569392
## 4 5.559563 3.121363 3.249042
## 5 2.925045 3.649894 1.859338 3.521639
```

```
image(k5.1)
points(Pharmaceuticals_2, col=Clusters_Index, pch=19, cex=0.3)
```



#Question2 Interpret the clusters with respect to the numerical variables used in forming the clusters.

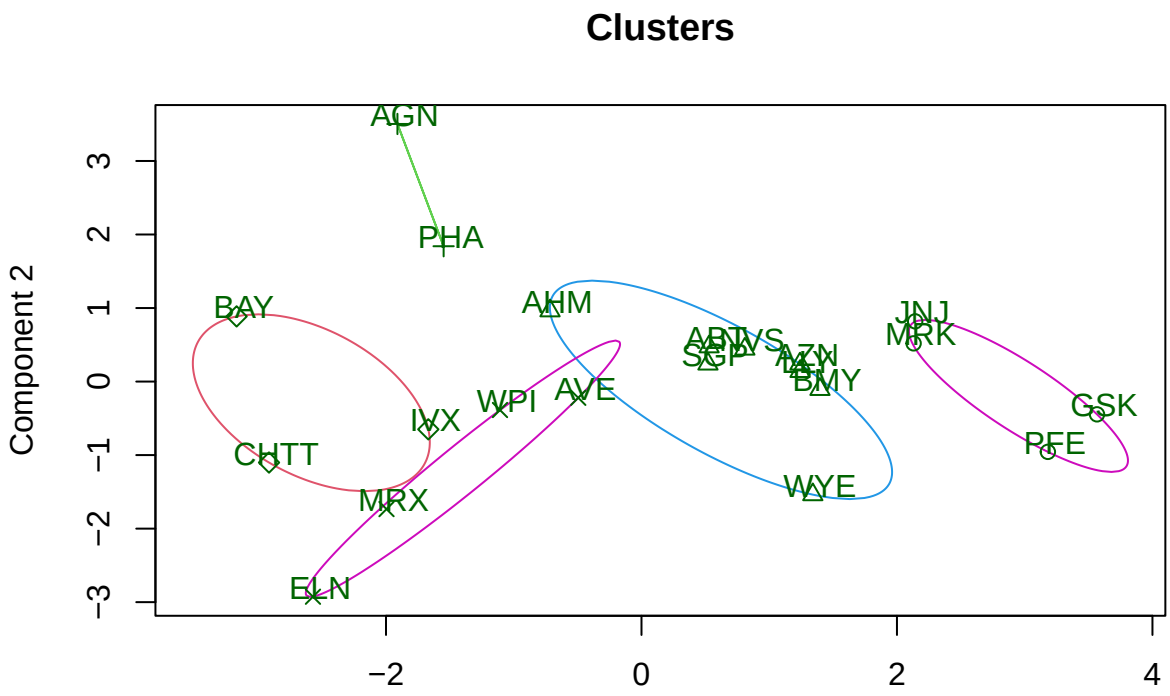
```
#Cluster 1 consists of stocks with high market valuation, profitability (ROE, ROA) and low leverage.
#Cluster 2 consists of stocks with moderate market valuation, profitability and leverage.
#Cluster 3 consists of stocks with high PE Ratio, lowest Asset Turnover and moderate leverage.
#Cluster 4 consists of stocks with lowest PE Ratio, highest Rev Growth, moderate ROE, ROA and high leve
#Cluster 5 consists of stocks with lowest Rev Growth, highest Beta and leverage
```

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

```
Pharmaceuticals_1 %>% mutate(Cluster = K5$cluster) %>% group_by(Cluster) %>%
summarise_all("mean")
```

```
## # A tibble: 5 x 10
##   Cluster Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##     <int>      <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
## 1       1     157.   0.48      22.2  44.4 17.7           0.95     0.22
## 2       2      55.8  0.414     20.3  28.7 12.7           0.738    0.371
## 3       3      31.9  0.405     69.5  13.2  5.6           0.75     0.475
## 4       4      13.1  0.598     17.7  14.6  6.2           0.425    0.635
## 5       5       6.64 0.87      24.6  16.5  4.17          0.6      1.65
## # i 2 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>
```

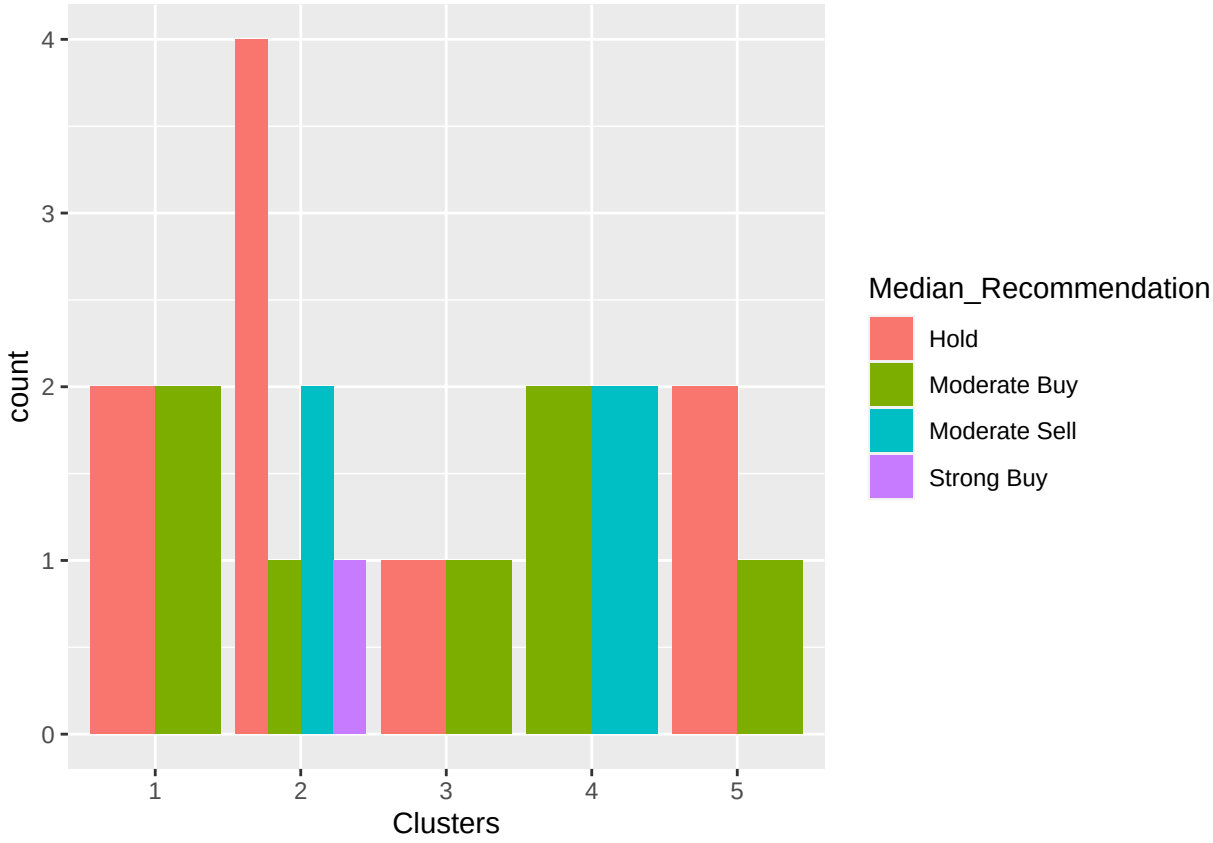
```
clusplot(Pharmaceuticals_2,K5$cluster,main="Clusters",color = TRUE,labels = 3, lines = 0)
```



These two components explain 61.23 % of the point variability.

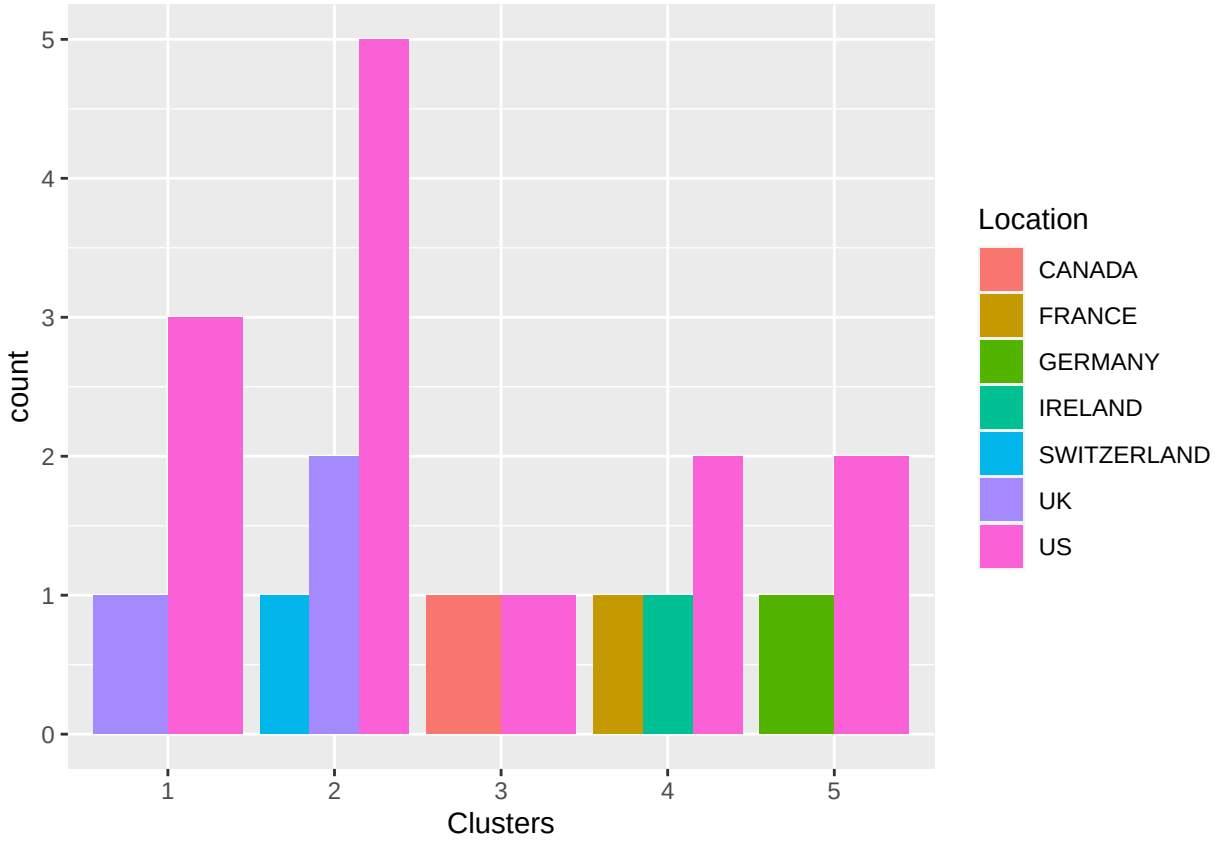
#For the non-numeric value **median recommendation**

```
Pharmaceuticals_3 <- Pharmaceuticals[12:14] %>% mutate(Clusters=K5$cluster)
ggplot(Pharmaceuticals_3, mapping = aes(factor(Clusters), fill
=Median_Recommendation))+geom_bar(position='dodge')+labs(x ='Clusters')
```



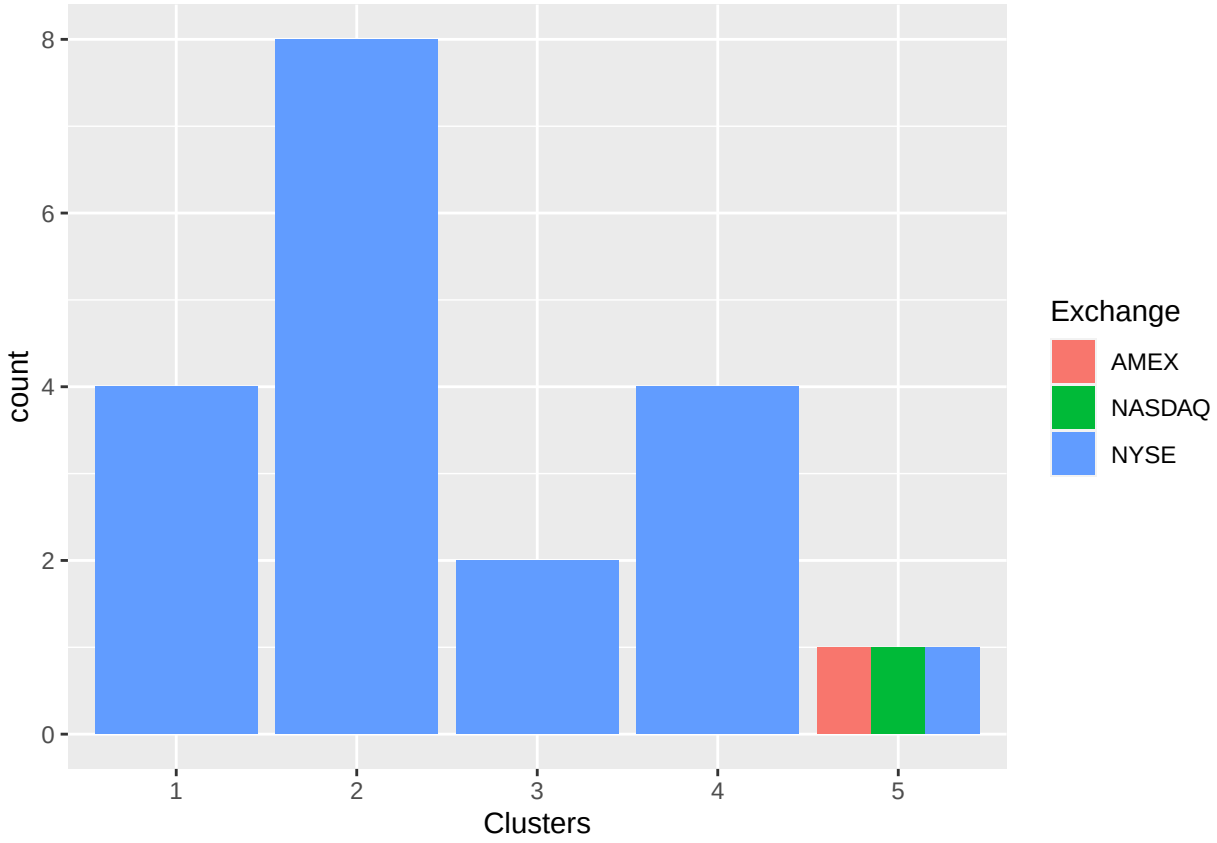
#For the non numeric value **location**

```
ggplot(Pharmaceuticals_3, mapping = aes(factor(Clusters),fill =
Location))+geom_bar(position = 'dodge')+labs(x ='Clusters')
```



#For the non numeric value **Exchange**

```
ggplot(Pharmaceuticals_3, mapping = aes(factor(Clusters),fill =
Exchange))+geom_bar(position = 'dodge')+labs(x ='Clusters')
```




```
#By looking at the graphs

#Cluster1 has a hold and a moderate buy recommendation and are situated in the UK and the US and are li
#Cluster2 has all of the recommendations that is hold, moderate buy, moderate sell and strong buy. The
#Cluster3 has again hold and moderate buy recommendation, listed on the NYSE and have their presence in
#Cluster4 has moderate sell and moderate buy recommendation listed on the NYSE and have their presence
#Cluster5 has Hold and and moderate buy recommendation and listed on all the three stock exchanges and
```

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

```
#Cluster 1: Stable large Caps- because they have high market valuation, profitability and low leverage.
#Cluster 2: Moderate Growth- because they have moderate valuation, profitability and leverage.
#Cluster 3: High PE value Traps- because they have high PE ratio, low asset turnover and moderate lever
#Cluster 4: High growth, High risk-because of their high revenue growth and moderate ROE/ROA.
#Cluster 5: Distressed Cyclical Stocks-because of their lowest revenue growth.
```