Kmeans Clustering model

Shashidhar Reddy Boreddy

2022-11-01

```
#Importing Required Packages
library(readr)
#Importing Data Set
sb_pharma <- read_csv("C:/Users/shash/Dropbox/PC/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.
```

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

```
## Warning: package 'factoextra' was built under R version 4.2.2
```

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

```
library(flexclust)
```

```
## Warning: package 'flexclust' was built under R version 4.2.2
```

```
## 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 4.2.2
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v purrr   0.3.4      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#summary Of the Data
summary(sb_pharma)
```

```
##     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
##
##
##
```

```
#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.
S <- na.omit(sb_pharma)
S
```

```
## # A tibble: 21 x 14
##    Symbol Name         Marke~1  Beta PE_Ra~2   ROE   ROA Asset~3 Lever~4 Rev_G~5
##    <chr>  <chr>          <dbl> <dbl>   <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl>
##  1 ABT    Abbott Labo~   68.4   0.32    24.7  26.4  11.8     0.7    0.42    7.54
##  2 AGN    Allergan, I~    7.58  0.41    82.5  12.9   5.5     0.9    0.6     9.16
```

```
##  3 AHM    Amersham plc    6.3   0.46    20.7  14.9   7.8     0.9    0.27    7.05
##  4 AZN    AstraZeneca~   67.6   0.52    21.5  27.4  15.4     0.9    0       15
##  5 AVE    Aventis        47.2   0.32    20.1  21.8   7.5     0.6    0.34   26.8
##  6 BAY    Bayer AG       16.9   1.11    27.9   3.9   1.4     0.6    0      -3.17
##  7 BMY    Bristol-Mye~   51.3   0.5     13.9  34.8  15.1     0.9    0.57    2.7
##  8 CHTT   Chattem, Inc    0.41  0.85    26    24.1   4.3     0.6    3.51    6.38
##  9 ELN    Elan Corpor~    0.78  1.08     3.6  15.1   5.1     0.3    1.07   34.2
## 10 LLY    Eli Lilly a~   73.8   0.18    27.9  31    13.5     0.6    0.53    6.21
## # ... with 11 more rows, 4 more variables: Net_Profit_Margin <dbl>,
## #   Median_Recommendation <chr>, Location <chr>, Exchange <chr>, and
## #   abbreviated variable names 1: Market_Cap, 2: PE_Ratio, 3: Asset_Turnover,
## #   4: Leverage, 5: Rev_Growth
```

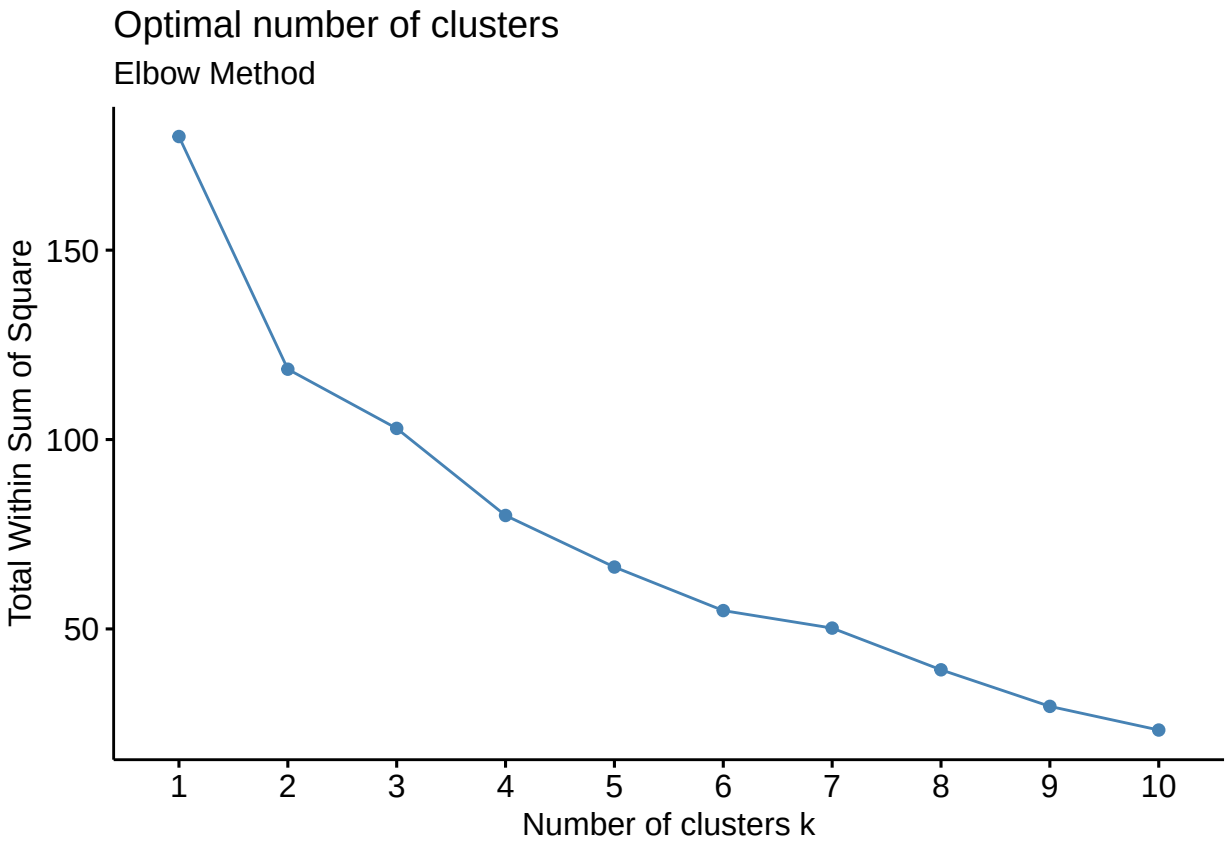
```
row_names <- S[,1]
sb_pharma1 <- S[,3:11]#numerical variable from 3 to 11
head(sb_pharma1)
```

```
## # A tibble: 6 x 9
##   Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage Rev_Gr~1 Net_P~2
##        <dbl> <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    16.1
## 2       7.58  0.41     82.5  12.9   5.5            0.9     0.6      9.16     5.5
## 3       6.3   0.46     20.7  14.9   7.8            0.9     0.27     7.05    11.2
## 4      67.6   0.52     21.5  27.4  15.4            0.9     0       15       18
## 5      47.2   0.32     20.1  21.8   7.5            0.6     0.34    26.8     12.9
## 6      16.9   1.11     27.9   3.9   1.4            0.6     0       -3.17     2.6
## # ... with abbreviated variable names 1: Rev_Growth, 2: Net_Profit_Margin
```

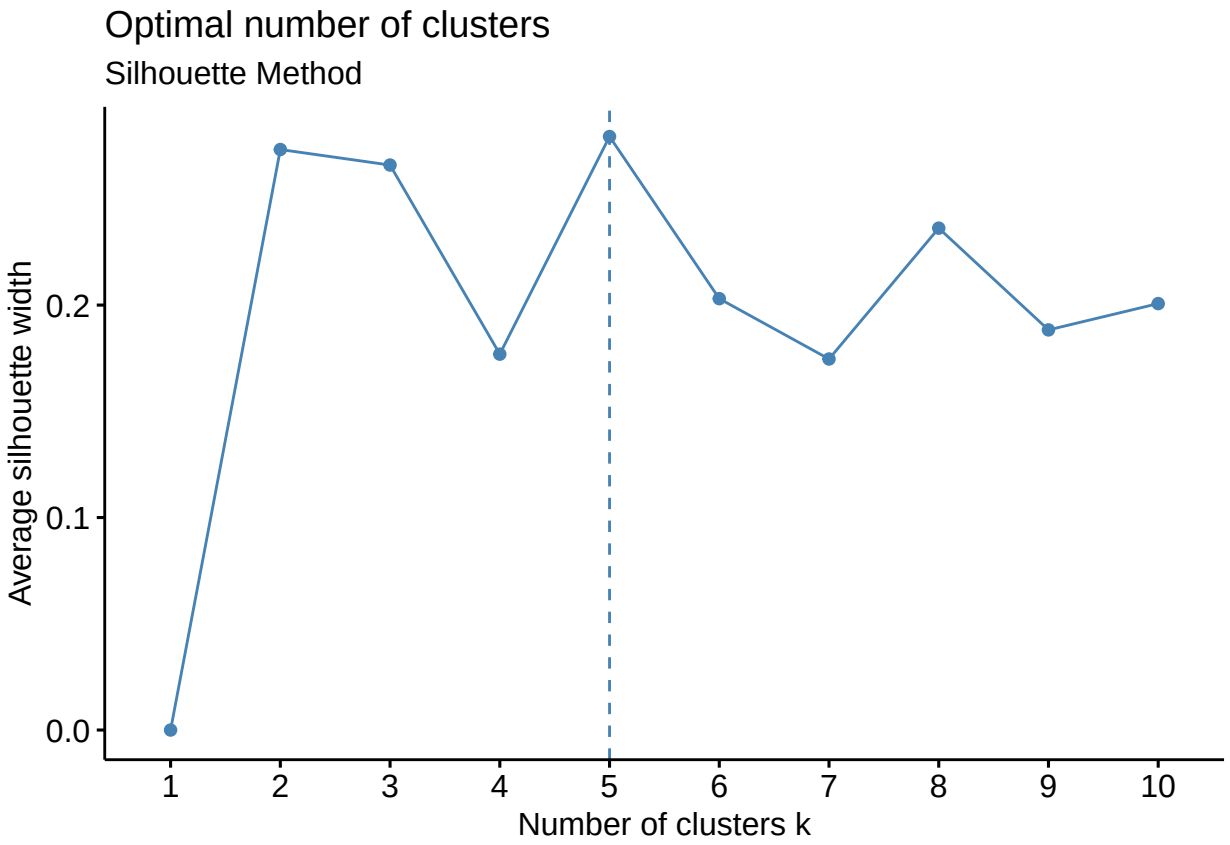
```
sb_pharma2 <- scale(sb_pharma1)
head(sb_pharma2)
```

```
##      Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## [1,]  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
## [2,] -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
## [3,] -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312
## [4,]  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312
## [5,] -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656
## [6,] -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656
##        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
```

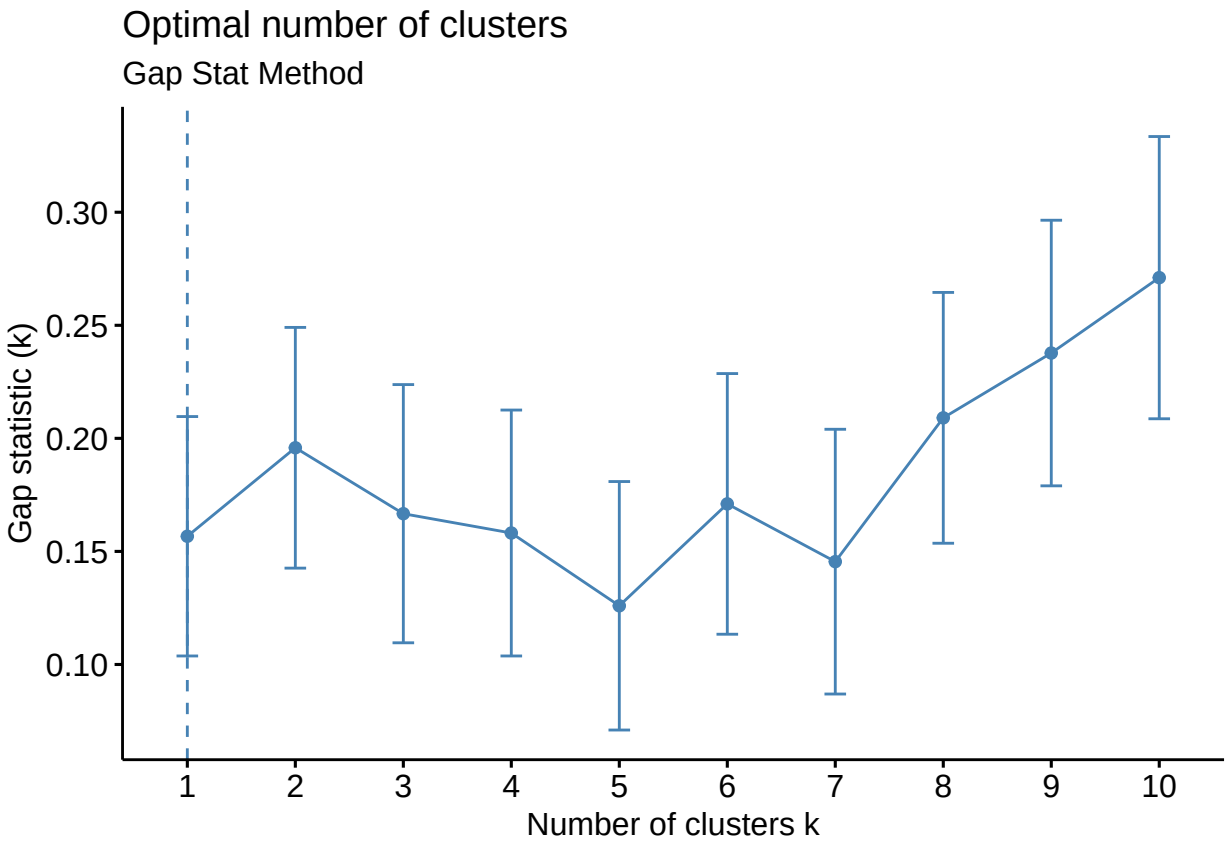
```
fviz_nbclust(sb_pharma2, kmeans, method = "wss") +labs(subtitle = "Elbow Method")
```




```
fviz_nbclust(sb_pharma2, kmeans, method = "silhouette") + labs(subtitle = "Silhouette Method")
```



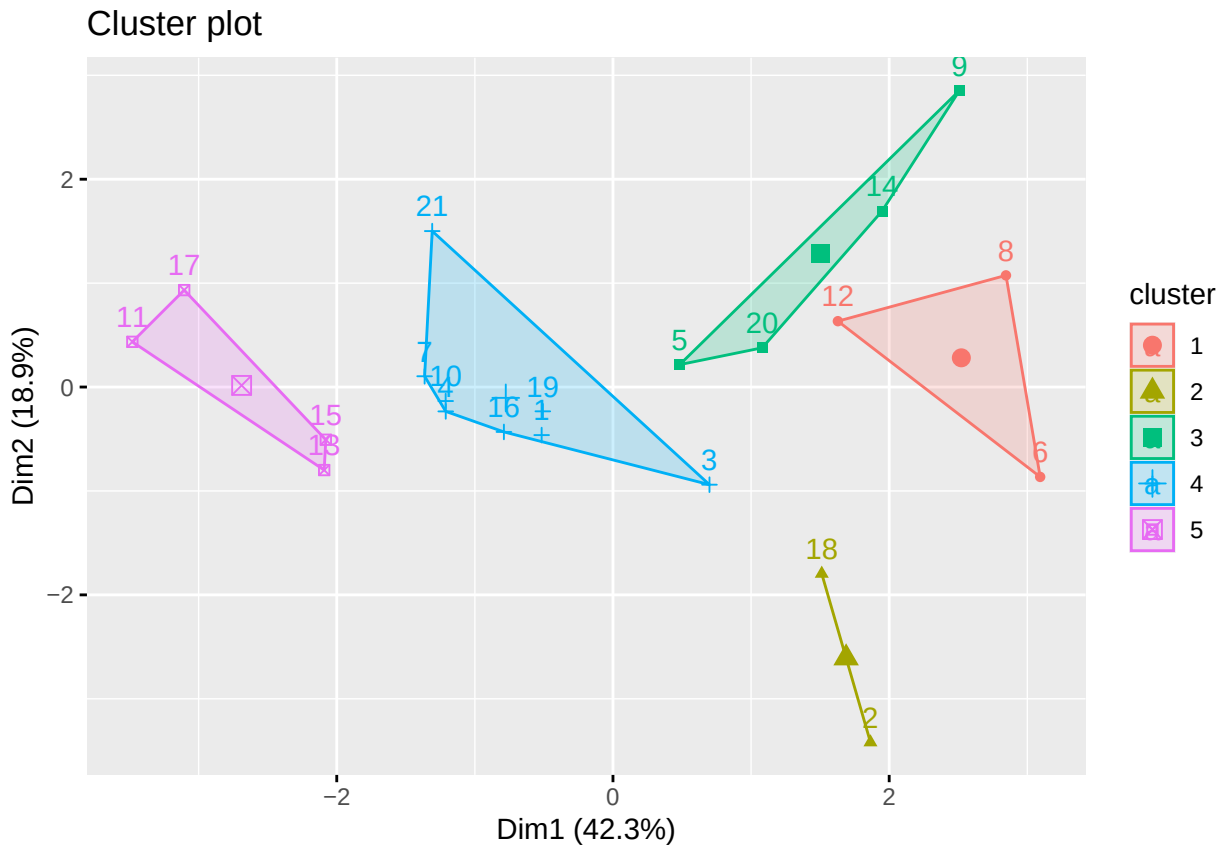

```
fviz_nbclust(sb_pharma2, kmeans, method = "gap_stat") + labs(subtitle = "Gap Stat Method")
```




```
set.seed(64060)
k_5 <- kmeans(sb_pharma2, 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 = sb_pharma2)
```

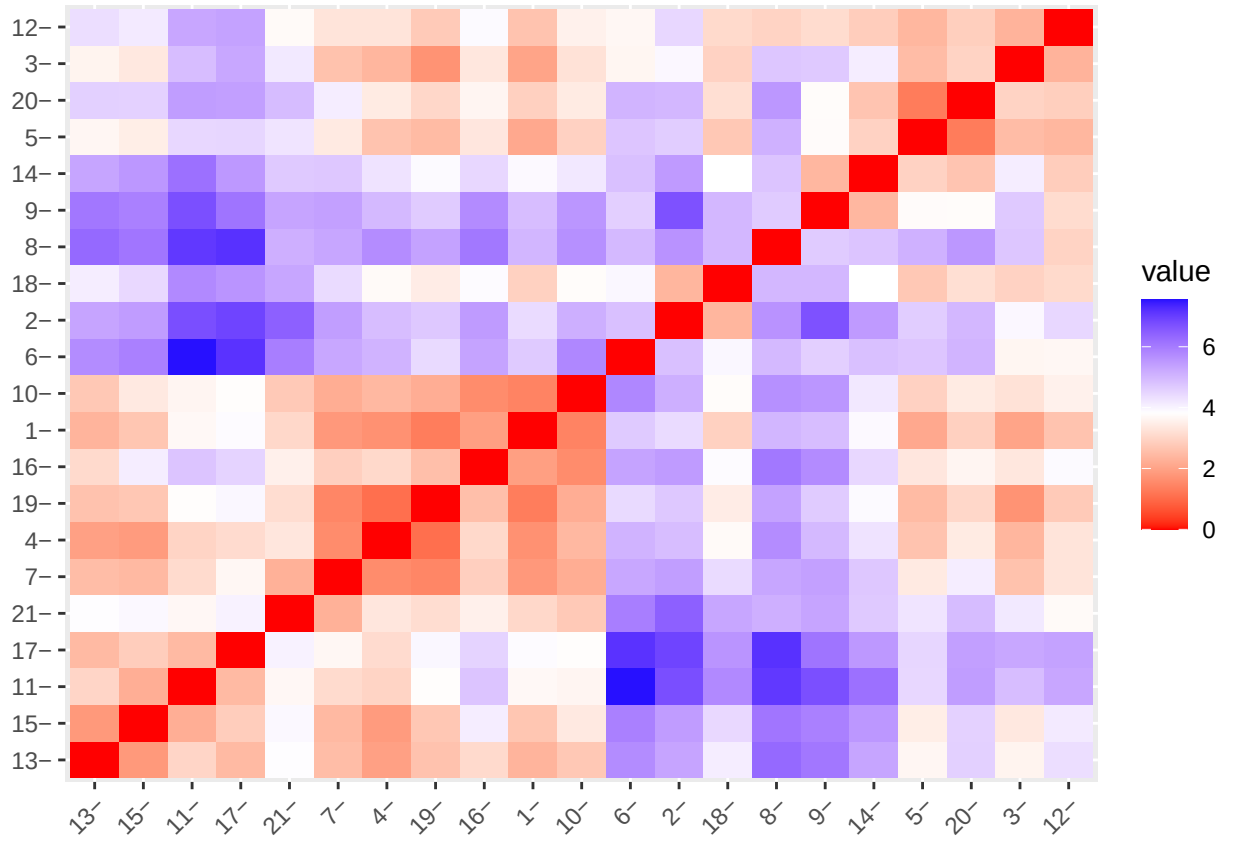



```
k_5
```

```
## 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"
```

```
Dist <- dist(sb_pharma2, method = "euclidian")
fviz_dist(Dist)
```



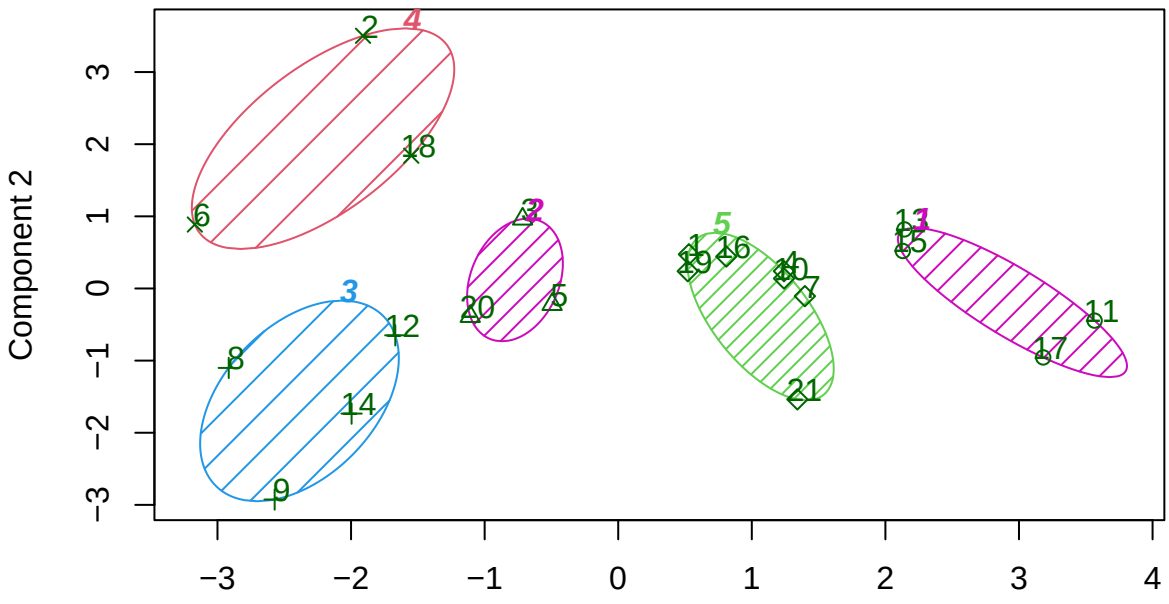

```
FITT <- kmeans(sb_pharma2,5)
aggregate(sb_pharma2,by = list(FITT$cluster), FUN = mean)
```

```
##   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   1.480297e-16 -0.3443544 -0.5769454        -1.6095439
## 5   6.589509e-02 -0.2559803 -0.7230135         0.7343816
```

```
sb_pharma3 <- data.frame(sb_pharma2,FITT$cluster)
sb_pharma3
```

```
##    Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## 1   0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
## 2  -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
## 3  -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312
## 4   0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312
## 5  -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656
## 6  -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656
## 7  -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498      0.9225312
## 8  -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918     -0.4612656
## 9  -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553     -1.8450624
## 10  0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770     -0.4612656
## 11  1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364      1.3837968
## 12 -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905     -0.4612656
## 13  1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544      0.9225312
## 14 -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792     -1.8450624
## 15  1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577      1.8450624
## 16  0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598     -0.9225312
## 17  2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239      0.4612656
## 18 -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030     -0.4612656
## 19 -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929      0.4612656
## 20 -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905     -0.9225312
## 21 -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849     -0.4612656
##       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
```

```
library(cluster)
clusplot(sb_pharma2,FITT$cluster, color = TRUE, shade = TRUE,
 labels = 2,
 lines = 0)
```




```
#QUestion (B)
#Interpret the clusters with respect to the numerical variables used in forming
#the clusters.
aggregate(sb_pharma2, 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   1.480297e-16 -0.3443544 -0.5769454        -1.6095439
## 5   6.589509e-02 -0.2559803 -0.7230135         0.7343816
```

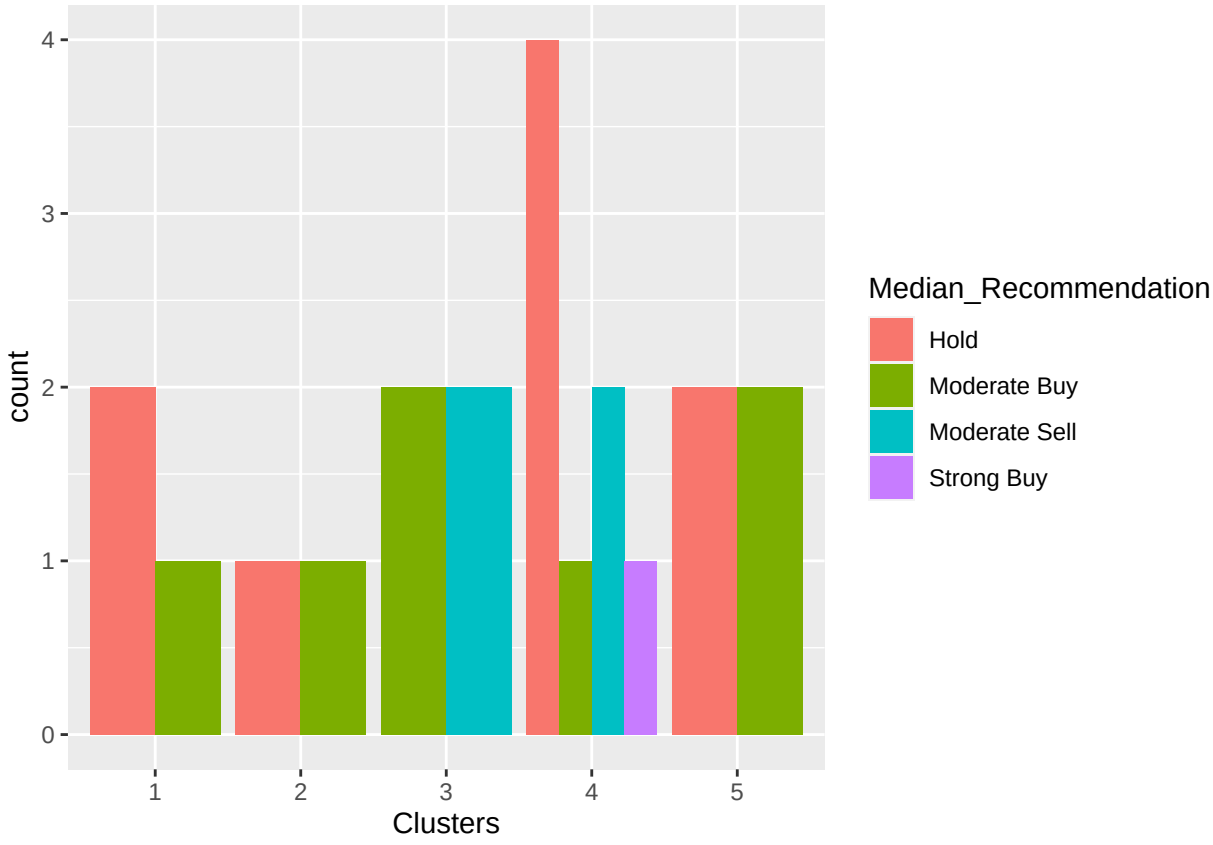
```
Sb_Pharmacy <- data.frame(sb_pharma2,k_5$cluster)
Sb_Pharmacy
```

```
##    Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## 1   0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
## 2  -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
```

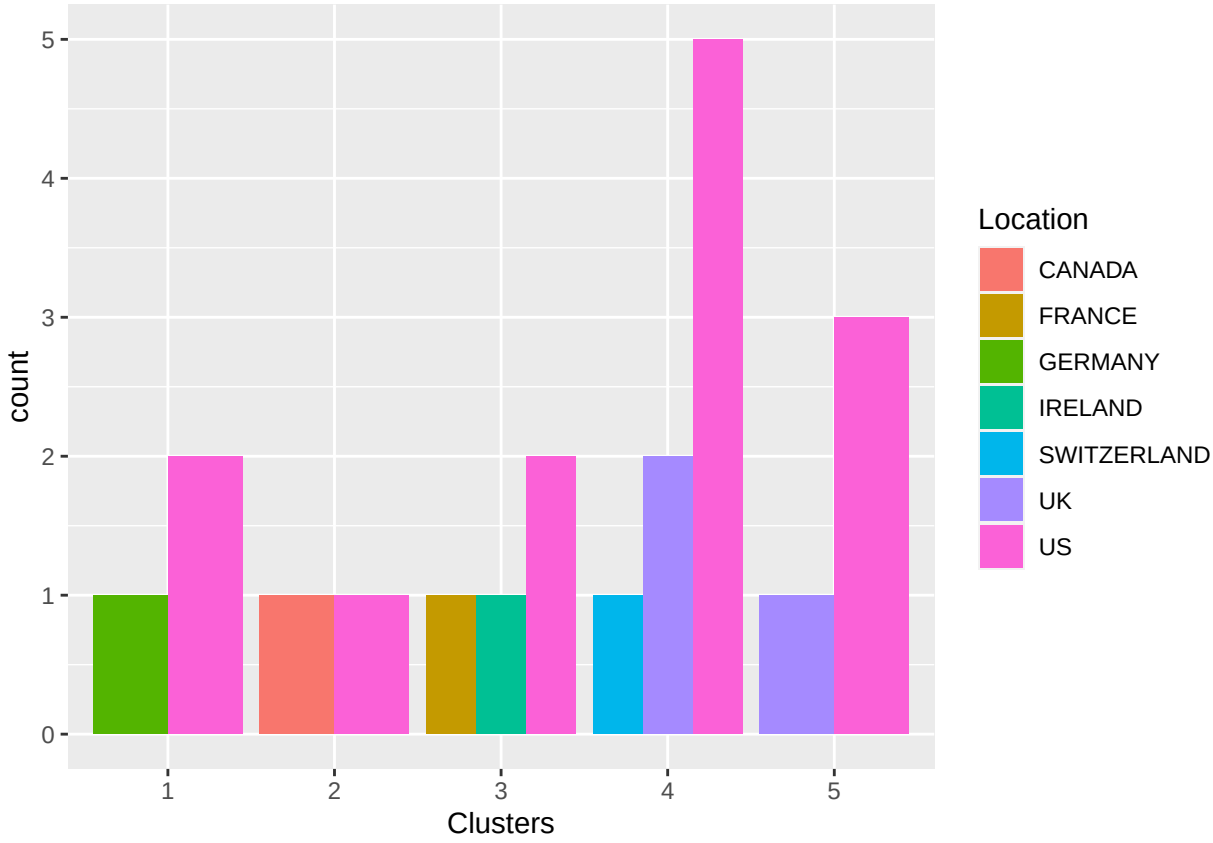
```
## 3  -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312
## 4   0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312
## 5  -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656
## 6  -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656
## 7  -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498      0.9225312
## 8  -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918     -0.4612656
## 9  -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553     -1.8450624
## 10  0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770     -0.4612656
## 11  1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364      1.3837968
## 12 -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905     -0.4612656
## 13  1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544      0.9225312
## 14 -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792     -1.8450624
## 15  1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577      1.8450624
## 16  0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598     -0.9225312
## 17  2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239      0.4612656
## 18 -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030     -0.4612656
## 19 -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929      0.4612656
## 20 -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905     -0.9225312
## 21 -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849     -0.4612656
##       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
#(Highest Market_Cap and lowest Beta/PE Ratio)
#Cluster 2:- AHM, WPI, AVE
#(Highest Revenue Growth and lowest PE/Asset Turnover Ratio)
#Cluster 3:- CHTT, IVX, MRX, ELN
#(Highest Beta/leverage/Asset Turnover Ratio and lowest Net_Profit_Margin, PE ratio and Marke#Cluster)
#Cluster 4:- AGN,BAY, PHA
#(Highest PE ratio and lowest Leverage/Asset_Turnover)
#Cluster 5:- ABT, WYE, AZN, SGP, BMY, NVS, LLY
# (Highest Net_Proft_Margin and lowest Leverage)
```

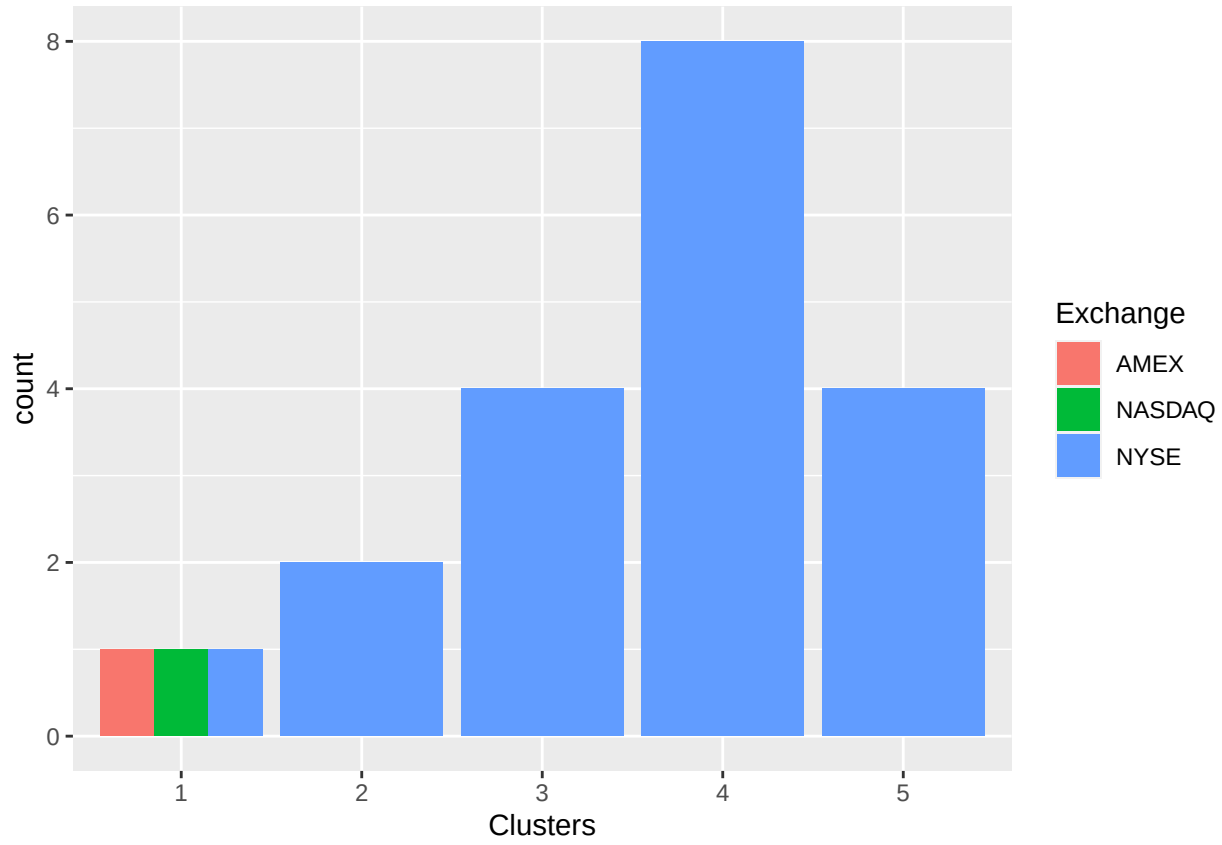
```r
#Question(C)
#Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? #(those not use
PH <- sb_pharma[12:14] %>% mutate(Clusters=k_5$cluster)
ggplot(PH, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position='dodge')+labs
```




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




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




```
#We can see a minor pattern in the clusters from the graphs above
#The cluster 1 has distinct Hold and Moderate Buy medians, as well as a different count from the US and
#Hold and Moderate buy medians are similarly distributed in Cluster 2
#the United States and Canada are listed on the NYSE..
#Cluster 3 has similar Moderate Buy and Sell medians, but a different count from Cluster
#France, Ireland, and the United States are all listed on the NYSE
#Cluster 4 offers Hold, Moderate Buy, Moderate Sell, and Strong Buy options.
#The median for the hold is the highest. They are from the United States, the United Kingdom, and Switz
# Cluster 5 has the same hold and mdoerate purchase medians and is spread in
#countries UK and US and is also listed in NYSE
```

```
#Question(D)
#Provide an appropriate name for each cluster using any or all of the variables in the dataset.
#Cluster 1 :- Buy Cluster
#Cluster 2 :- Sceptical Cluster
#Cluster 3 :- Moderate Buy Cluster
#Cluster 4 :- Hold Cluster
#Cluster 5 :- High Hold Cluster
```