k-Means for clustering

Nikhil Reddy Addula

2022-11-05

```
#Importing the Dataset
library(readr)
PharmaC <- read_csv("~/Documents/assignments/FUNDAMENTALS ML/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.
```

```
summary(PharmaC)
```

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

```
str(PharmaC)
```

```
## spec_tbl_df [21 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Symbol               : chr [1:21] "ABT" "AGN" "AHM" "AZN" ...
##  $ Name                 : chr [1:21] "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZen
##  $ Market_Cap           : num [1:21] 68.44 7.58 6.3 67.63 47.16 ...
##  $ Beta                 : num [1:21] 0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...
##  $ PE_Ratio             : num [1:21] 24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...
##  $ ROE                  : num [1:21] 26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...
##  $ ROA                  : num [1:21] 11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...
##  $ Asset_Turnover       : num [1:21] 0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
##  $ Leverage             : num [1:21] 0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...
##  $ Rev_Growth           : num [1:21] 7.54 9.16 7.05 15 26.81 ...
##  $ Net_Profit_Margin    : num [1:21] 16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...
##  $ Median_Recommendation: chr [1:21] "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sell" ...
##  $ Location             : chr [1:21] "US" "CANADA" "UK" "UK" ...
##  $ Exchange             : chr [1:21] "NYSE" "NYSE" "NYSE" "NYSE" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Symbol = col_character(),
##   ..   Name = col_character(),
##   ..   Market_Cap = col_double(),
##   ..   Beta = col_double(),
##   ..   PE_Ratio = col_double(),
##   ..   ROE = col_double(),
##   ..   ROA = col_double(),
##   ..   Asset_Turnover = col_double(),
##   ..   Leverage = col_double(),
##   ..   Rev_Growth = col_double(),
##   ..   Net_Profit_Margin = col_double(),
##   ..   Median_Recommendation = col_character(),
##   ..   Location = col_character(),
##   ..   Exchange = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
#Loading the Packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

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

```
library(factoextra)
```

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

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v tibble  3.1.8     v stringr 1.4.1
## v tidyr   1.2.1     v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

```
library(cluster)
library(gridExtra)
```

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

```
#a. Cluster the 21 companies using only the numerical variables (1-9). Justify the numerous decisions t

#Removing the dataset's null values and choosing the monetary variables.
colSums(is.na(PharmaC))
```

```
##                Symbol                  Name            Market_Cap
##                     0                     0                     0
##                  Beta              PE_Ratio                   ROE
##                     0                     0                     0
##                   ROA        Asset_Turnover              Leverage
##                     0                     0                     0
##            Rev_Growth     Net_Profit_Margin Median_Recommendation
##                     0                     0                     0
##              Location              Exchange
##                     0                     0
```

```
row.names <- PharmaC[,1]
PharmaC_data_n<- PharmaC[, 3:11]
head(PharmaC_data_n)
```

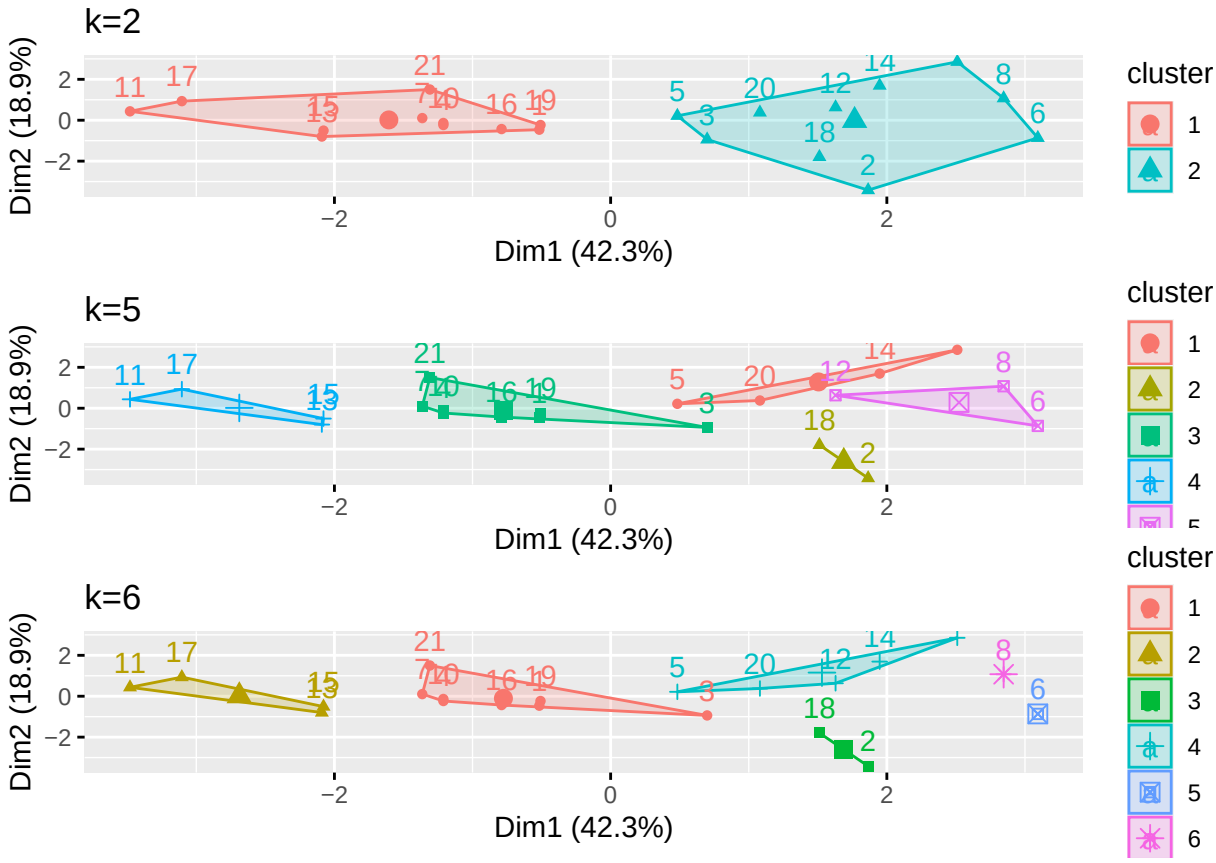
```
## # A tibble: 6 x 9
##   Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage Rev_Gr~1 Net_P~2
##        <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>    <dbl>   <dbl>
## 1      68.4   0.32     24.7  26.4  11.8            0.7     0.42     7.54    16.1
## 2       7.58  0.41     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
```

```r
# Scaling and Normalisation of dataset.
PharmaC_scale <- scale(PharmaC_data_n)
head(PharmaC_scale)
```

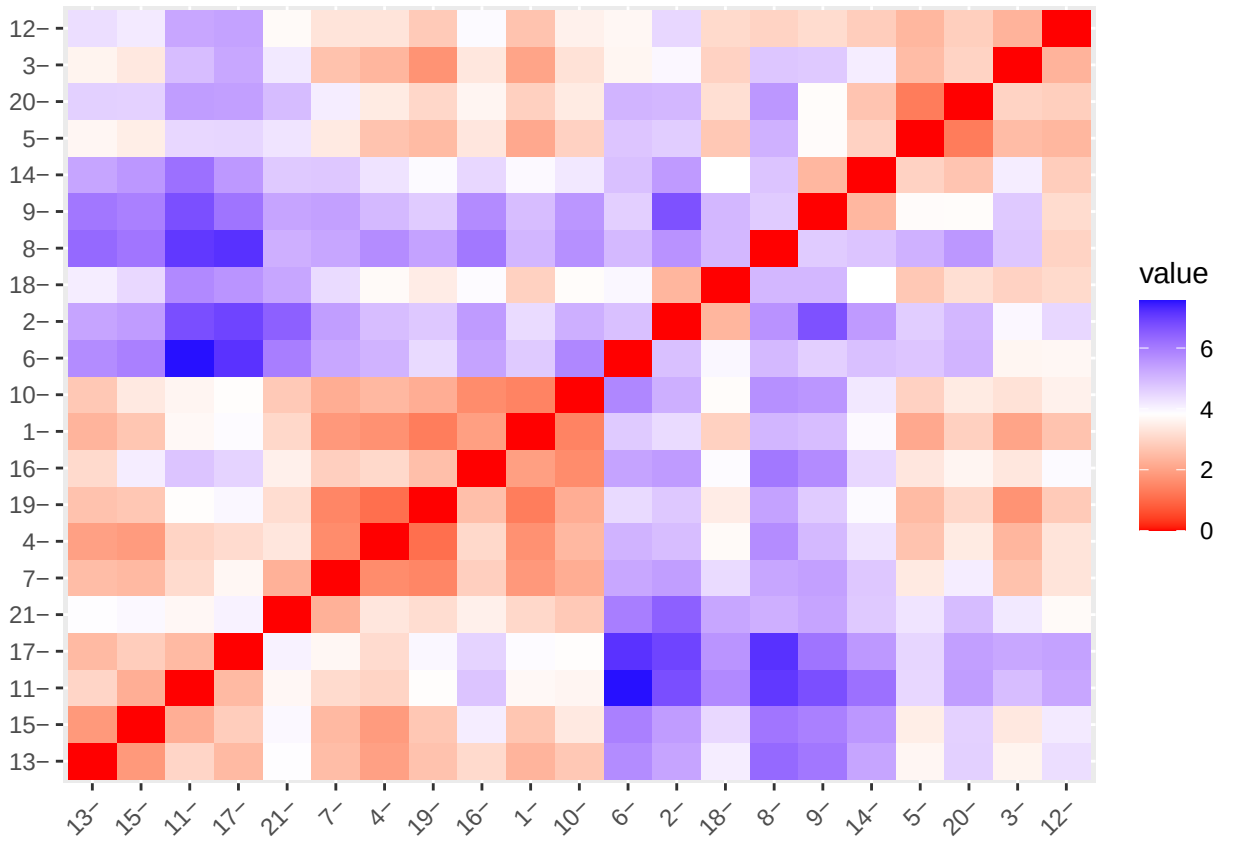
```
##      Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## [1,]  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121  -5.121077e-16
## [2,] -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871   9.225312e-01
## [3,] -0.8762600 -0.25595600 -0.29195768 -0.72225761 -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
```

```r
n_data <- as.data.frame(scale(PharmaC_data_n))
```

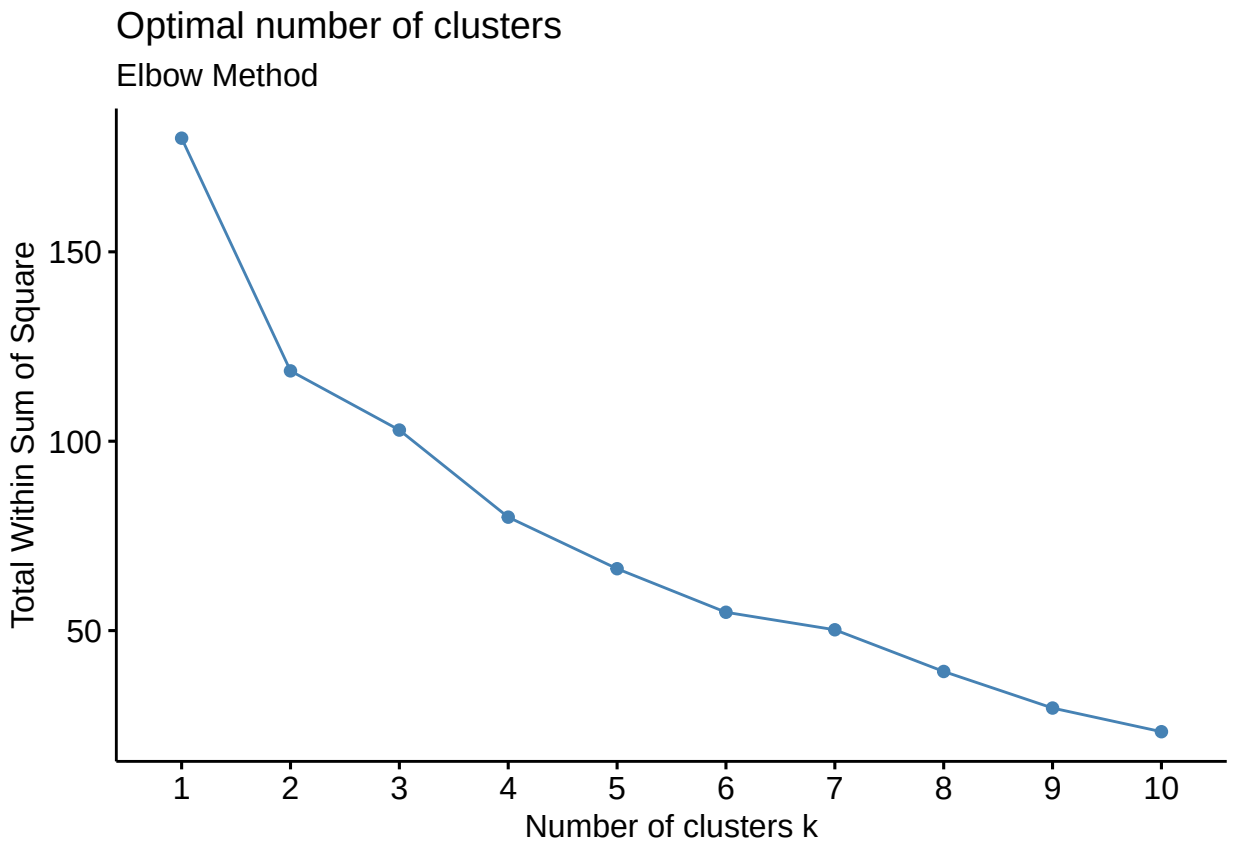
```r
# Calculate K-means clustering for various centers, use a variety of K values, and compare the results.
kmeans_1n <- kmeans(PharmaC_scale, centers = 2, nstart = 30)
kmeans_2n<- kmeans(PharmaC_scale, centers = 5, nstart = 30)
kmeans_3n<- kmeans(PharmaC_scale, centers = 6, nstart = 30)
Plot_1r<-fviz_cluster(kmeans_1n, data = PharmaC_scale)+ggtitle("k=2")
Plot_2r<-fviz_cluster(kmeans_2n, data = PharmaC_scale)+ggtitle("k=5")
Plot_3r<-fviz_cluster(kmeans_3n, data = PharmaC_scale)+ggtitle("k=6")
grid.arrange(Plot_1r,Plot_2r,Plot_3r, nrow = 3)
```



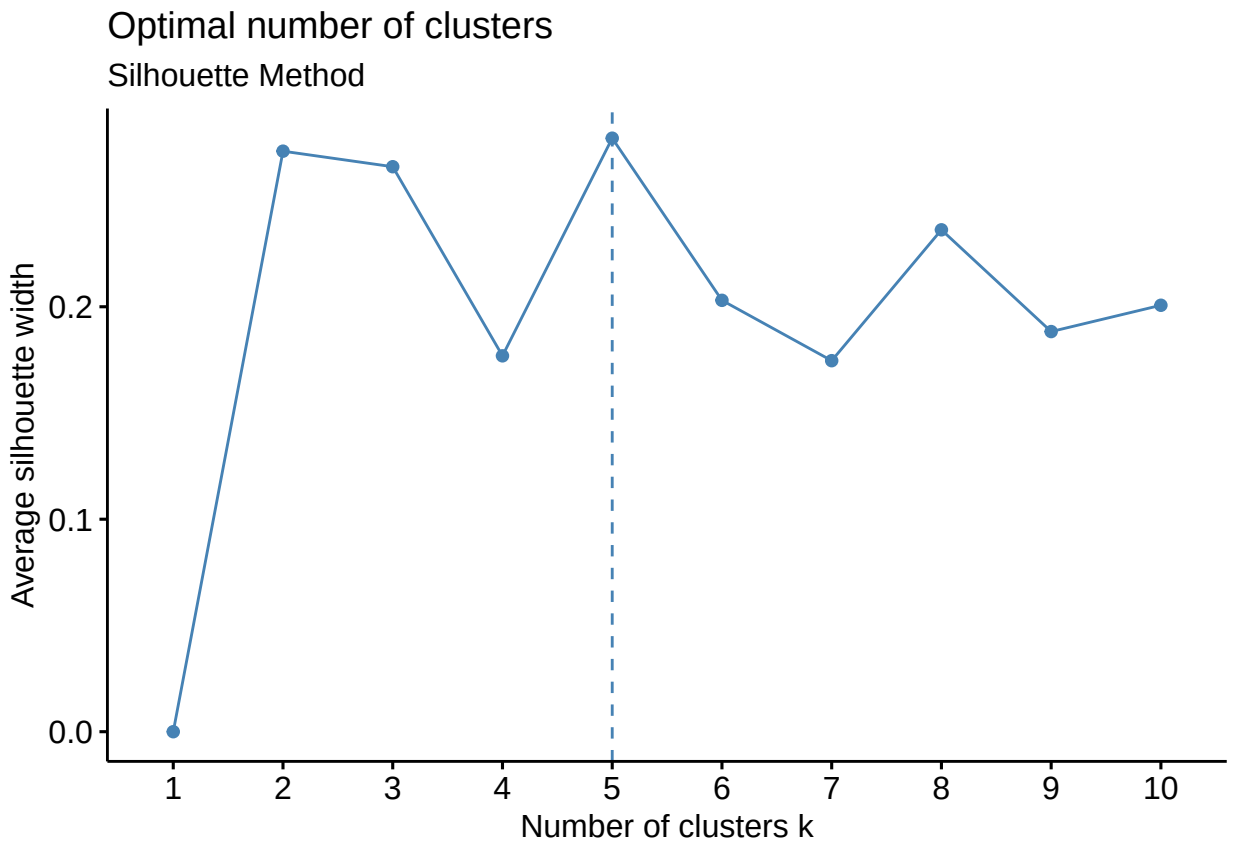

```
dis<- dist(PharmaC_scale, method = "euclidean")
fviz_dist(dis)
```




```
# Estimating the  number of clusters
# Elbow Method is used in scaling the data to determine the value of k
fviz_nbclust(n_data, FUNcluster = kmeans, method = "wss") + labs(subtitle = "Elbow Method")
```




```
# Silhouette Method is used in scaling the data to determine the number of clusters
fviz_nbclust(n_data,FUNcluster = kmeans,method = "silhouette")+labs(subtitle="Silhouette Method")
```

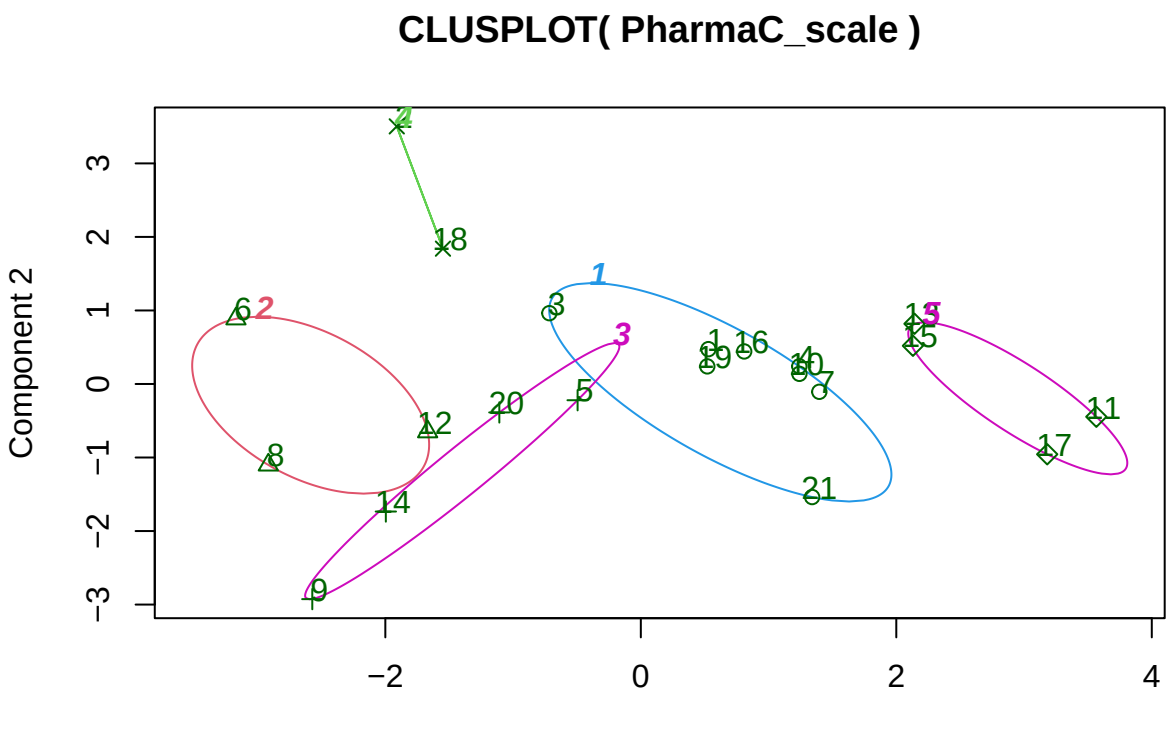



```
# Final analysis and Extracting results using 5 clusters and Visualize the results
set.seed(300)
final_C<- kmeans(PharmaC_scale, 5, nstart = 25)
print(final_C)
```

```
## K-means clustering with 5 clusters of sizes 8, 3, 4, 2, 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 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 4 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.27449312 -0.7041516       0.556954446
## 2  1.36644699 -0.6912914      -1.320000179
## 3  0.06308085  1.5180158      -0.006893899
## 4 -0.14170336 -0.1168459      -1.416514761
## 5 -0.46807818  0.4671788       0.591242521
##
## Clustering vector:
##  [1] 1 4 1 1 3 2 1 2 3 1 5 2 5 3 5 1 5 4 1 3 1
##
## Within cluster sum of squares by cluster:
## [1] 21.879320 15.595925 12.791257  2.803505  9.284424
##  (between_SS / total_SS =  65.4 %)
##
```

```
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
clusplot(PharmaC_scale,final_C$cluster, color = TRUE, labels = 2,lines = 0)
```

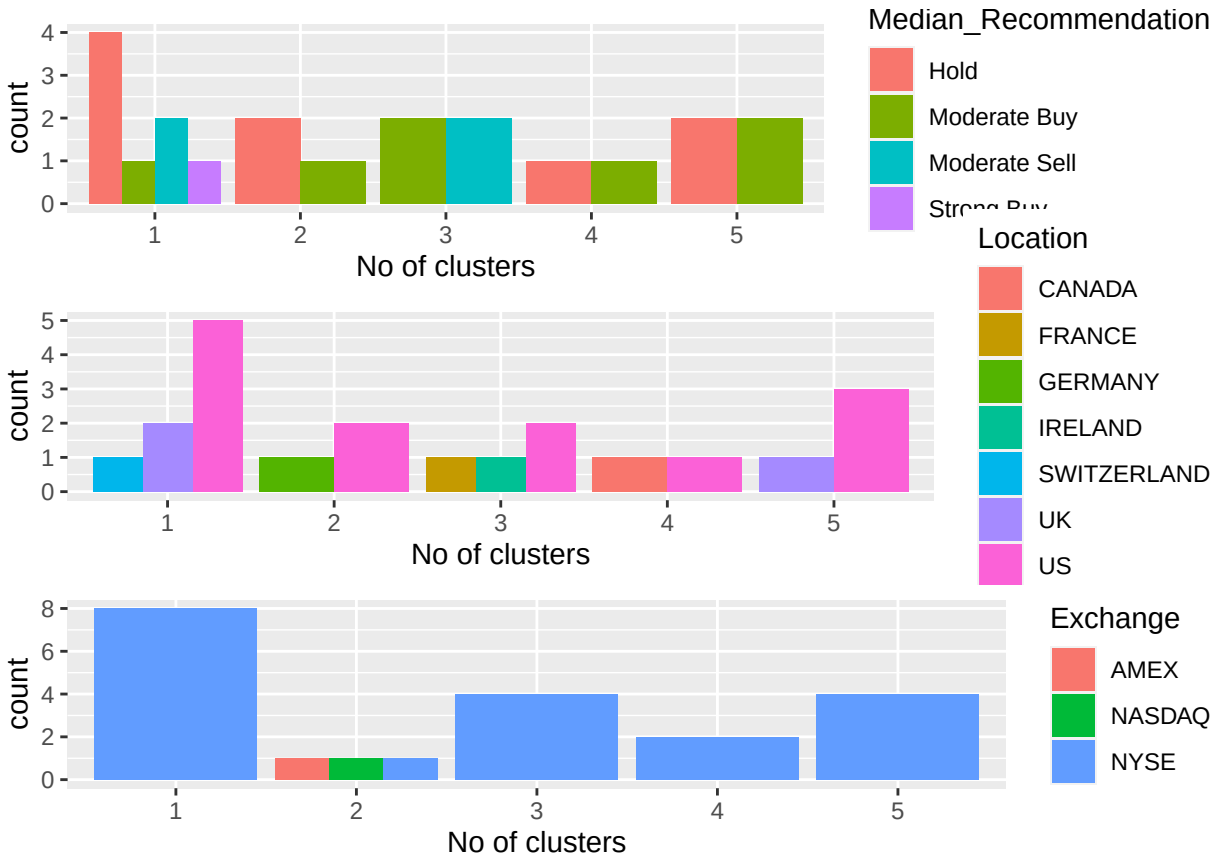



```
#b) Interpret the clusters with respect to the numerical variables used in forming the clusters.

#Cluster 1 - 1,3,4,7,10,16,19,21 ( lowest Market_Cap,lowest Beta,lowest PE_Ratio,highest Leverage,highe

#Cluster 2 - 6, 8, 12 (lowest Rev_Growth,highest Beta and leveargе,lowest Net_Profit_Margin)

#Cluster 3 - 5, 9, 14, 20 (lowest PE_Ratio,highest ROE,lowest ROA,lowest Net_Profit_Margin, highest Rev

#Cluster 4 - 2, 18 (lowest Beta,lowest Asset_Turnover, Highest PE Ratio)

#Cluster 5 - 11, 13, 15, 17 (Highest Market_Cap,ROE, ROA,Asset_Turnover Ratio and lowest Beta/PE Ratio)

PC_Cluster <- PharmaC[,c(12,13,14)]%>% mutate(clusters = final_C$cluster)%>% arrange(clusters, ascendin
PC_Cluster
```

```
## # A tibble: 21 x 4
##    Median_Recommendation Location    Exchange clusters
##    <chr>                 <chr>       <chr>       <int>
##  1 Moderate Buy          US          NYSE            1
##  2 Strong Buy            UK          NYSE            1
##  3 Moderate Sell         UK          NYSE            1
##  4 Moderate Sell         US          NYSE            1
```

```
##  5 Hold                  US          NYSE            1
##  6 Hold                  SWITZERLAND NYSE            1
##  7 Hold                  US          NYSE            1
##  8 Hold                  US          NYSE            1
##  9 Hold                  GERMANY     NYSE            2
## 10 Moderate Buy          US          NASDAQ          2
## # ... with 11 more rows
```

```
#Task3
#In terms of the numerical, are there any clusters that exhibit a pattern. (10 to 12) variables? (those


plot1_nr<-ggplot(PC_Cluster, mapping = aes(factor(clusters), fill=Median_Recommendation))+geom_bar(posi
plot2_nr<- ggplot(PC_Cluster, mapping = aes(factor(clusters),fill = Location))+geom_bar(position = 'dod
plot3_nr<- ggplot(PC_Cluster, mapping = aes(factor(clusters),fill = Exchange))+geom_bar(position = 'dod
grid.arrange(plot1_nr, plot2_nr, plot3_nr)
```



#As per graph:-

#Cluster 1 :In this cluster, which also includes distinct Hold, Moderate Buy, Moderate Sell, and Strong Buy medians, the Hold median is the highest. They are from the US, the UK, and Switzerland and are traded on the NYSE.

#Cluster 2:AMEX, NASDAQ, and NYSE all have an equal distribution of companies, but there is a clear Hold and Moderate Buy median as well as a different count between the US and Germany.

#Cluster 3: listed on the NYSE, has distinct counts for France, Ireland, and the US, and has medians for buy and sell orders that are equally moderate.

#Cluster 4: has the same hold and moderate buy medians and is spread out across the US, UK, and listed in.

#Cluster 5: #exclusively listed on the NYSE, evenly distributed across the US and Canada, with medians of Hold and Moderate Buy.

#With respect to media Recommendation Variable ,the clusters follow a particular pattern: #Cluster 1 and Cluster 2 has Hold Recommendation. #Cluster 3, Cluster 4and Cluster 5 has moderate buy Recommendation.

```
# (d) Give each cluster a suitable name using any or all of the dataset's variables.

#Cluster 1 :-  HIGH HOLD CLUSTER
#Cluster 2 :- HOLD CLUSTER
#Cluster 3 :- BUY-SELL CLUSTER
#Cluster 4 :- HOLD-BUY CLUSTER
#Cluster 5 :- HOLD-BUY CLUSTER
```