FML ASSIGNMENT 4

Sushma Palancha

2023-11-12

Reading the required libraries

```r
library(flexclust)
```

```
## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4
```

```r
library(cluster) #Generic Utility Functions
library(tidyverse) #Data manipulation
```

```
## -- 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   3.4.3     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
```

```r
library(factoextra) #Used for clustering algorithms and visualization
```

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

```r
library(FactoMineR)
library(ggcorrplot) #Visualizing a correlation matrix using ggplot2
```

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.

```
getwd()
```

```
## [1] "/Users/palanchasushma/Documents/FML/FML ASSIGNMENT 4"
```

```
pharma<- read.csv("Pharmaceuticals.csv") #Reading the Dataset
pharma1 <- pharma[ ,3:11] #Considering only numercial values i.e., 3-11 columns from csv file
head(pharma1)
```

```
##   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(pharma1)
```

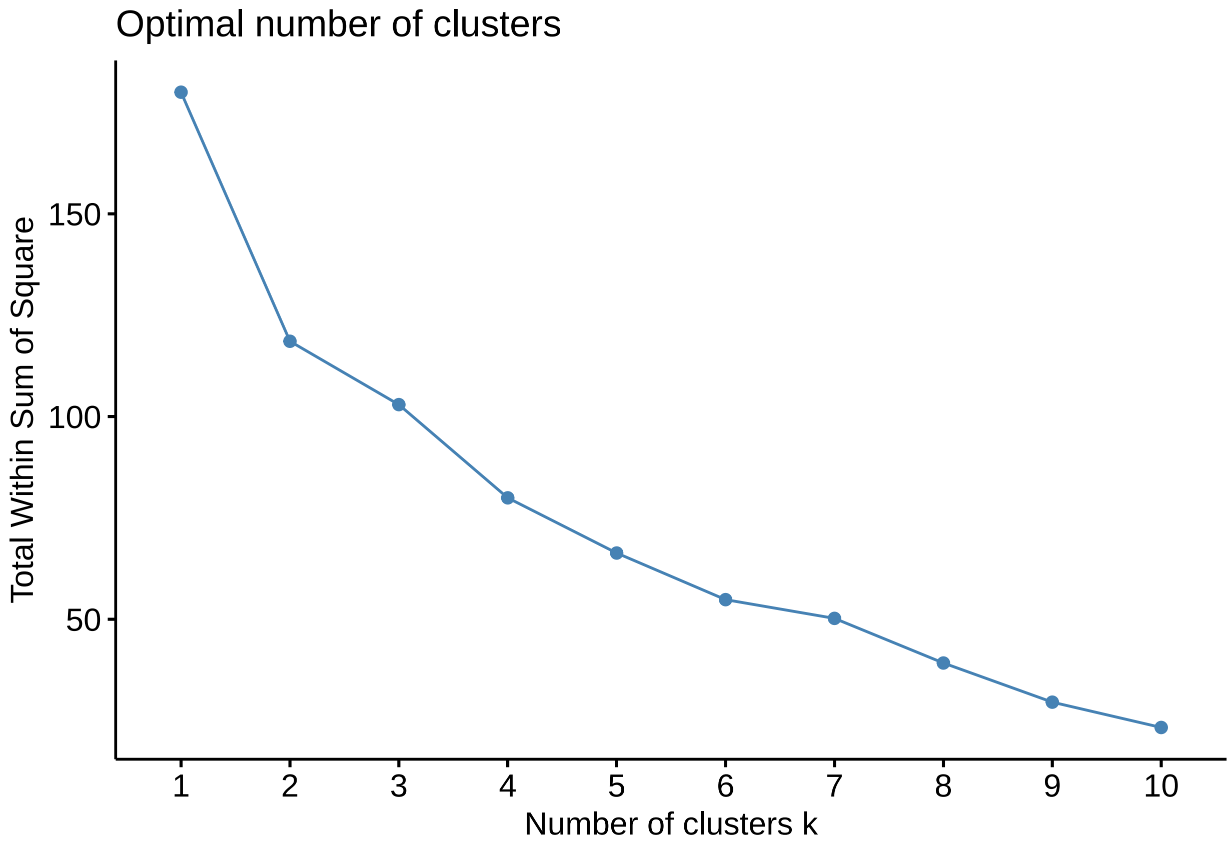
```
##    Market_Cap          Beta           PE_Ratio          ROE
##  Min.   :  0.41   Min.   :0.1800   Min.   : 3.60   Min.   : 3.9
##  1st Qu.:  6.30   1st Qu.:0.3500   1st Qu.:18.90   1st Qu.:14.9
##  Median : 48.19   Median :0.4600   Median :21.50   Median :22.6
##  Mean   : 57.65   Mean   :0.5257   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
```

Normalizing the data frame with scale method:

```
pharma2 <- scale(pharma1)
row.names(pharma2) <- pharma[,1]
distance <- get_dist(pharma2)
corr <- cor(pharma2)
```

Elbow Method to determine the number of clusters to do the cluster analysis:

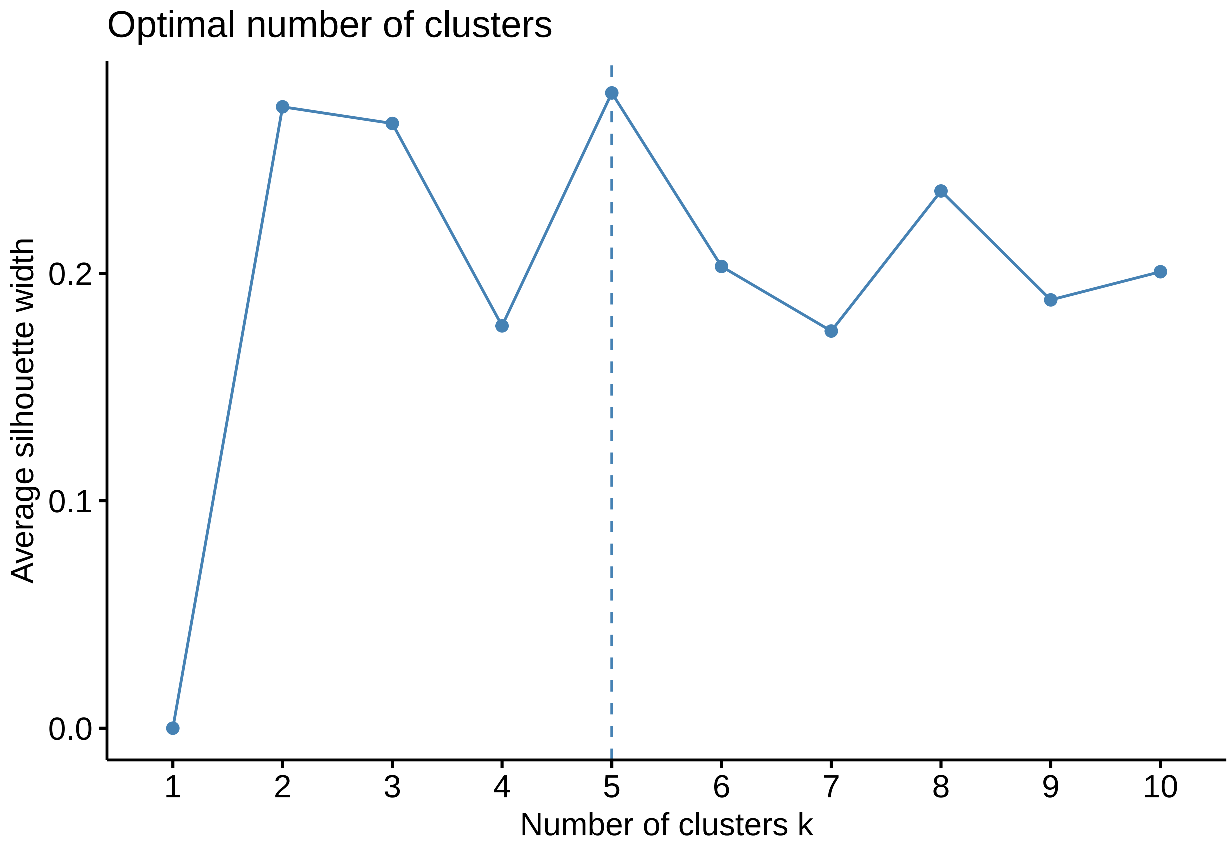
```
fviz_nbclust(pharma2, kmeans, method = "wss")
```



By seeing the above graph from Elbow method, Graph is not clear to choose k=2 or 3 or 4 or 5.

Silhouette method for determining no of clusters:

```
fviz_nbclust(pharma2, kmeans, method = "silhouette")
```



By seeing the graph from silhouette method, I can see sharp rise at k=5.So, considering the silhouette method.

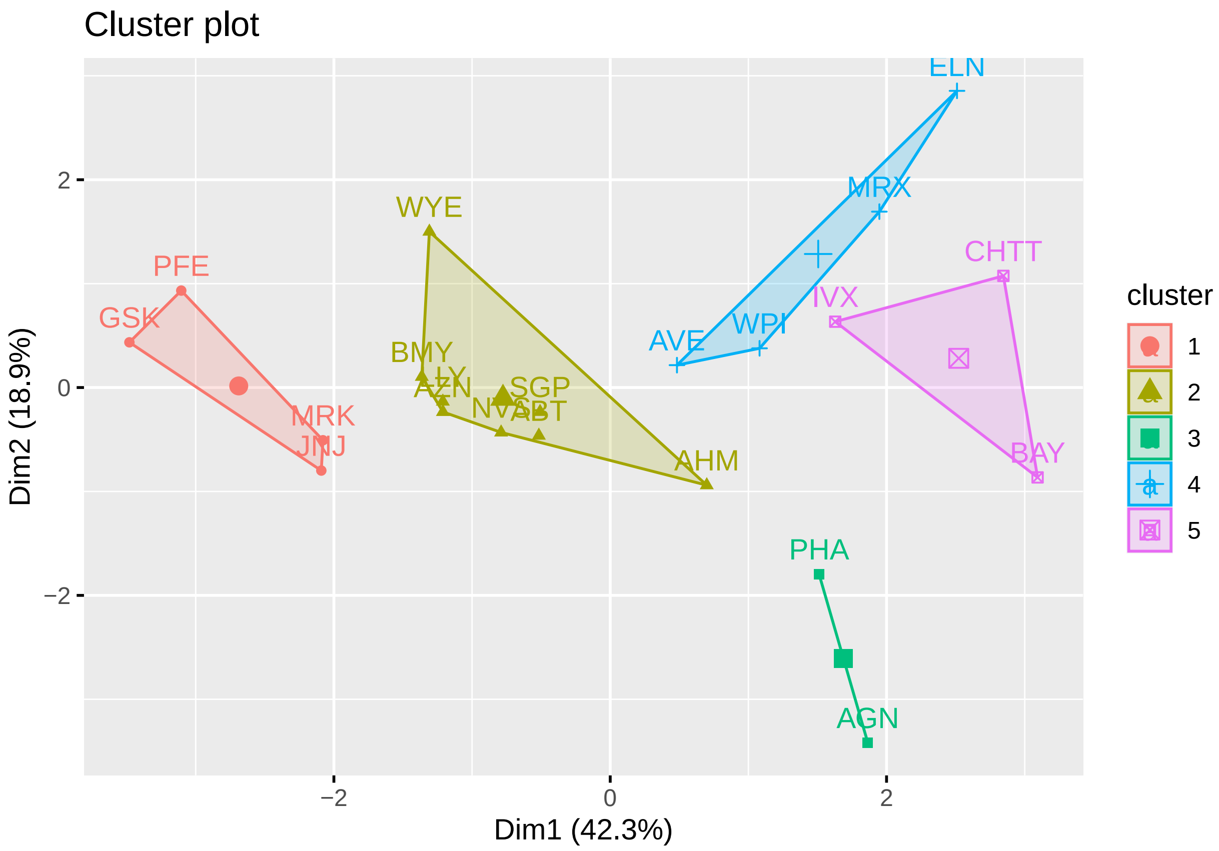
Selecting K-Means

```r
set.seed(69)
k5 <- kmeans(pharma2, centers = 5, nstart = 25) # k = 5, number of restarts = 25
#Visualizing the output
#Centroids
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
```

Visualizing Clustering Results

```r
fviz_cluster(k5, data = pharma2)
```

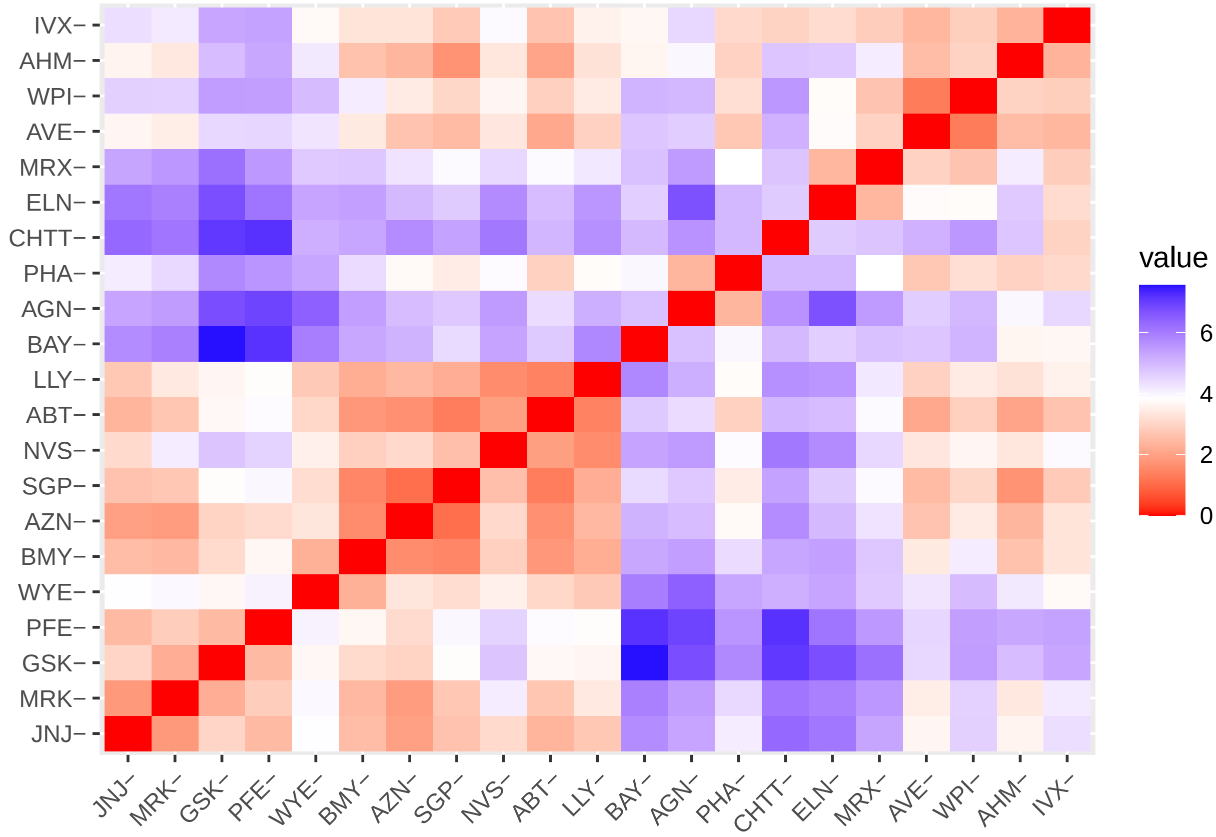


```r
k5
```

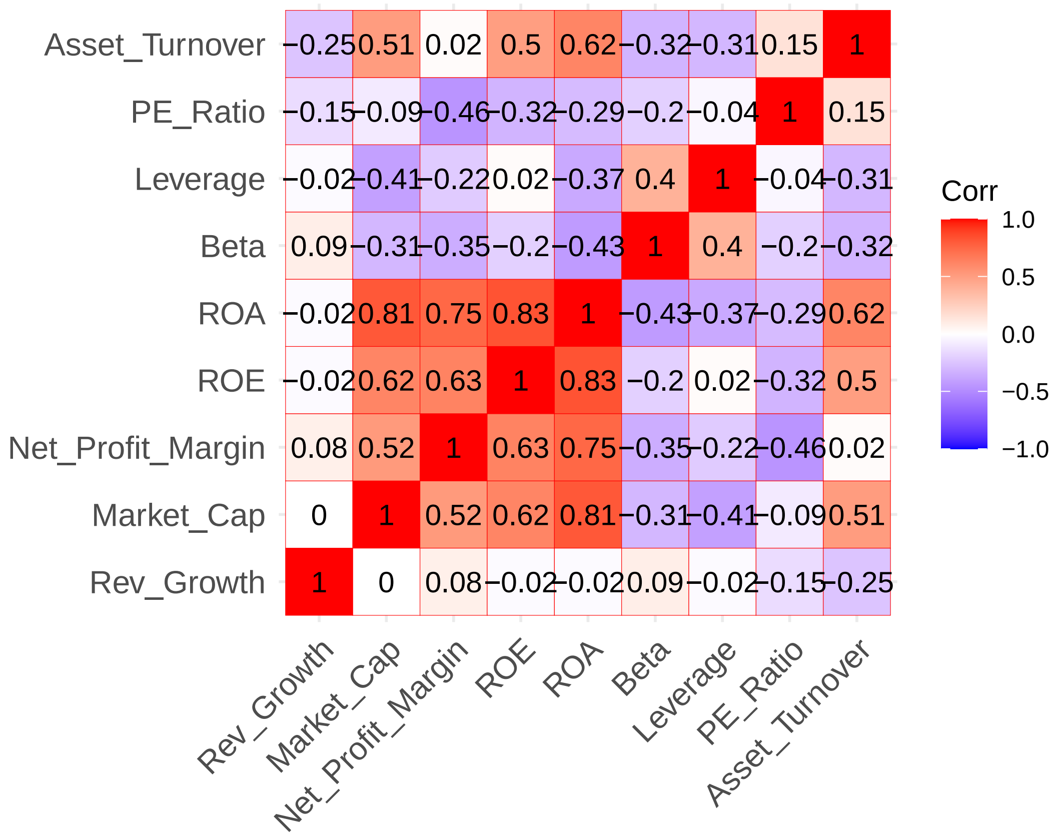
```
## K-means clustering with 5 clusters of sizes 4, 8, 2, 4, 3
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 2 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      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"
```

Distance Matrix Computation and Visualization

```
distance<- dist(pharma2, method = "euclidean")
fviz_dist(distance)
```



```
corr<-cor(pharma2)
ggcorrplot(corr,outline.color = "red",lab = TRUE,hc.order = TRUE,type = "full")
```



I can see there are 5 clusters and the center is defined after 25 restarts which is determined in kmeans.

```
#K-Means Cluster Analysis- Fit the data with 5 clusters
fit<-kmeans(pharma2,5)

#Finding the mean value of all quantitative variables for each cluster
aggregate(pharma2,by=list(fit$cluster),FUN=mean)
```

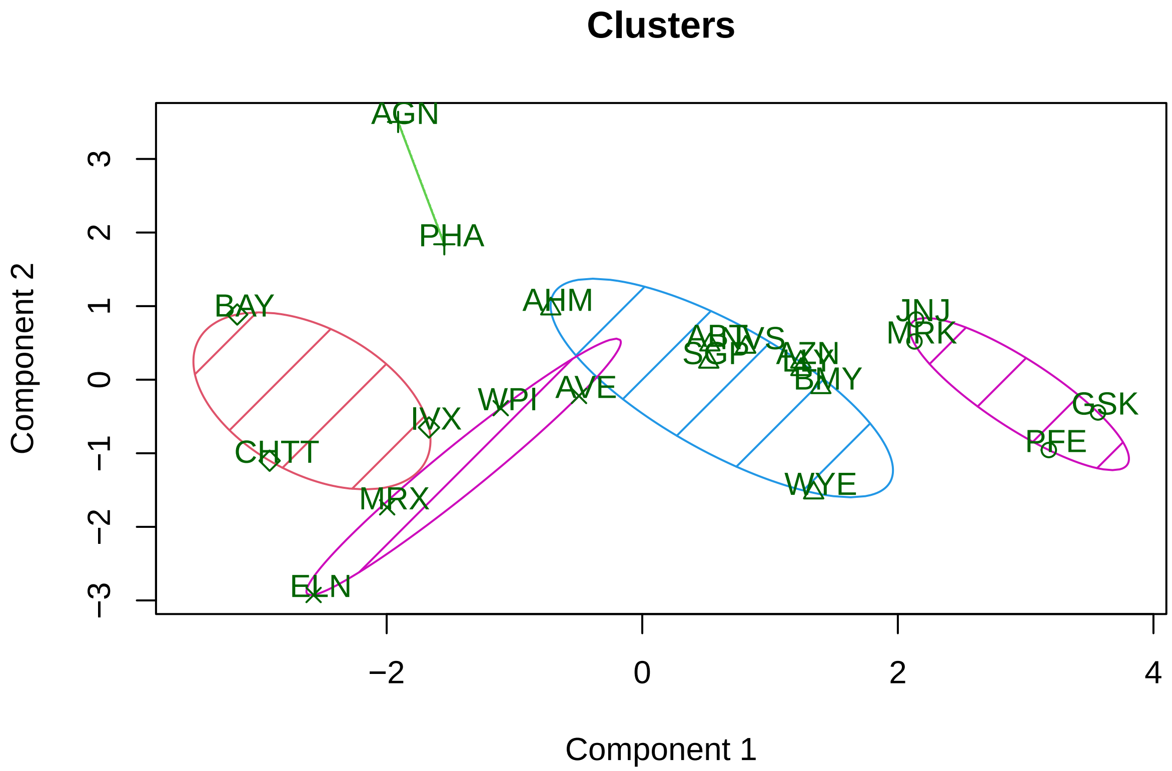
```
##   Group.1  Market_Cap       Beta   PE_Ratio        ROE        ROA
## 1       1  0.07576815 -0.5350964 -0.3338536  0.3627072  0.4765127
## 2       2 -0.90905697  1.4110965 -0.2613021 -0.7063477 -1.1114156
## 3       3 -0.57238455 -0.6220844  0.8692748 -0.7381675 -0.7242993
## 4       4  1.14422955 -0.1780563 -0.1550295  0.4245789  0.9494460
## 5       5  1.75995500 -0.1001567 -0.2858266  1.8862982  1.7355802
##   Asset_Turnover   Leverage  Rev_Growth Net_Profit_Margin
## 1  -7.687760e-02 -0.1737009 -0.86809028         0.7982409
## 2  -1.014784e+00  1.0319661  0.27018076        -0.6941793
## 3  -2.442491e-16 -0.2991312  0.36829509        -0.8069490
## 4   1.230042e+00 -0.5875356  0.04848824         0.1480374
## 5   9.225312e-01 -0.4296811  0.93534886         1.1360419
```

```
pharma3<-data.frame(pharma2,fit$cluster)
pharma3
```

```
##       Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## ABT   0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121  -5.121077e-16
## AGN  -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871   9.225312e-01
## AHM  -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700   9.225312e-01
## AZN   0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259   9.225312e-01
## AVE  -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461  -4.612656e-01
## BAY  -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612  -4.612656e-01
## BMY  -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498   9.225312e-01
## CHTT -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918  -4.612656e-01
## ELN  -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553  -1.845062e+00
## LLY   0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770  -4.612656e-01
## GSK   1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364   1.383797e+00
## IVX  -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905  -4.612656e-01
## JNJ   1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544   9.225312e-01
## MRX  -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792  -1.845062e+00
## MRK   1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577   1.845062e+00
## NVS   0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598  -9.225312e-01
## PFE   2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239   4.612656e-01
## PHA  -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030  -4.612656e-01
## SGP  -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929   4.612656e-01
## WPI  -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905  -9.225312e-01
## WYE  -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849  -4.612656e-01
##         Leverage  Rev_Growth Net_Profit_Margin fit.cluster
## ABT  -0.21209793 -0.52776752        0.06168225           1
## AGN   0.01828430 -0.38113909       -1.55366706           3
## AHM  -0.40408312 -0.57211809       -0.68503583           3
## AZN  -0.74965647  0.14744734        0.35122600           4
## AVE  -0.31449003  1.21638667       -0.42597037           3
## BAY  -0.74965647 -1.49714434       -1.99560225           2
## BMY  -0.02011273 -0.96584257        0.74744375           1
## CHTT  3.74279705 -0.63276071       -1.24888417           2
## ELN   0.61983791  1.88617085       -0.36501379           2
## LLY  -0.07130879 -0.64814764        1.17413980           1
## GSK  -0.31449003  0.76926048        0.82363947           5
## IVX   1.10620040  0.05603085       -0.71551412           2
## JNJ  -0.62166634 -0.36213170        0.33598685           4
## MRX   0.44065173  1.53860717        0.85411776           2
## MRK  -0.39128411  0.36014907       -0.24310064           4
## NVS  -0.67286239 -1.45369888        1.02174835           1
## PFE  -0.54487226  1.10143723        1.44844440           5
## PHA  -0.30169102  0.14744734       -1.27936246           3
## SGP  -0.74965647 -0.43544591        0.29026942           1
## WPI  -0.49367621  1.43089863       -0.09070919           3
## WYE   0.68383297 -1.17763919        1.49416183           1
```

```
view(pharma3)
```

```
clusplot(pharma2,k5$cluster, main="Clusters" ,shade=TRUE ,color = TRUE, labels = 3,lines = 0)
```



These two components explain 61.23 % of the point variability.

Manhattan Method

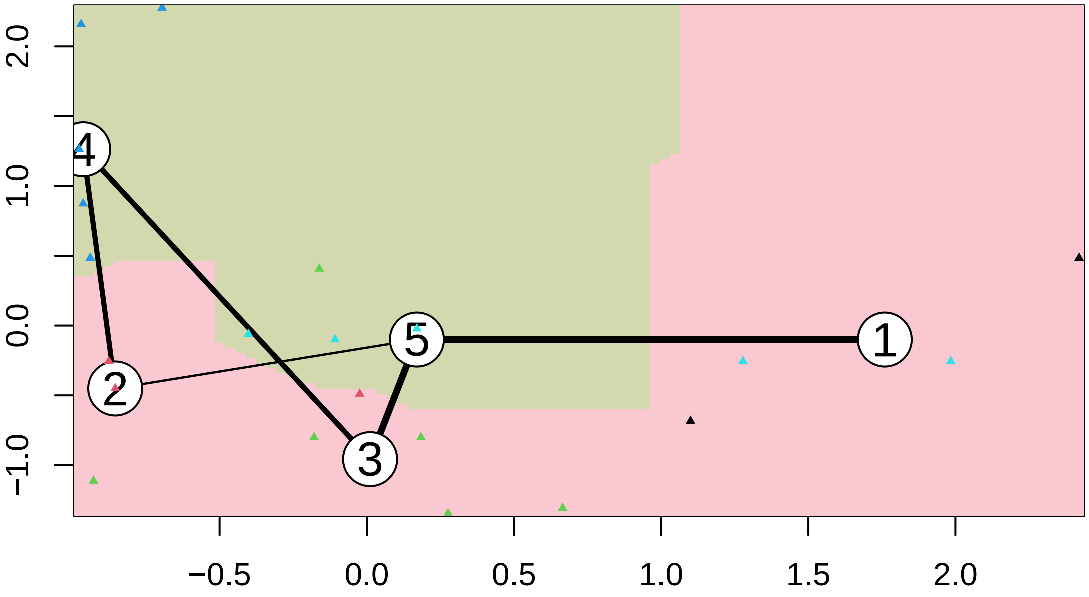
```
set.seed(69)
k51 = kcca(pharma2, k=5, kccaFamily("kmedians"))
k51
```

```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = pharma2, k = 5, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
## 1 2 3 4 5
## 2 3 6 5 5
```

```
clusters_index <- predict(k51)
dist(k51@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(k51)
points(pharma2, col=clusters_index,  pch=17, cex=0.5)
```



QuestionB:#Interpret the clusters with respect to the numerical variables

#used in forming the clusters.

```
pharma1 %>% 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>
```

Cluster 1:High mean values in certain variables suggest a specific profile for Cluster 1.

Cluster 2:Unique characteristics are indicated by mean values in Cluster 2.

Cluster 3:Patterns in mean values differentiate Cluster 3 from others.

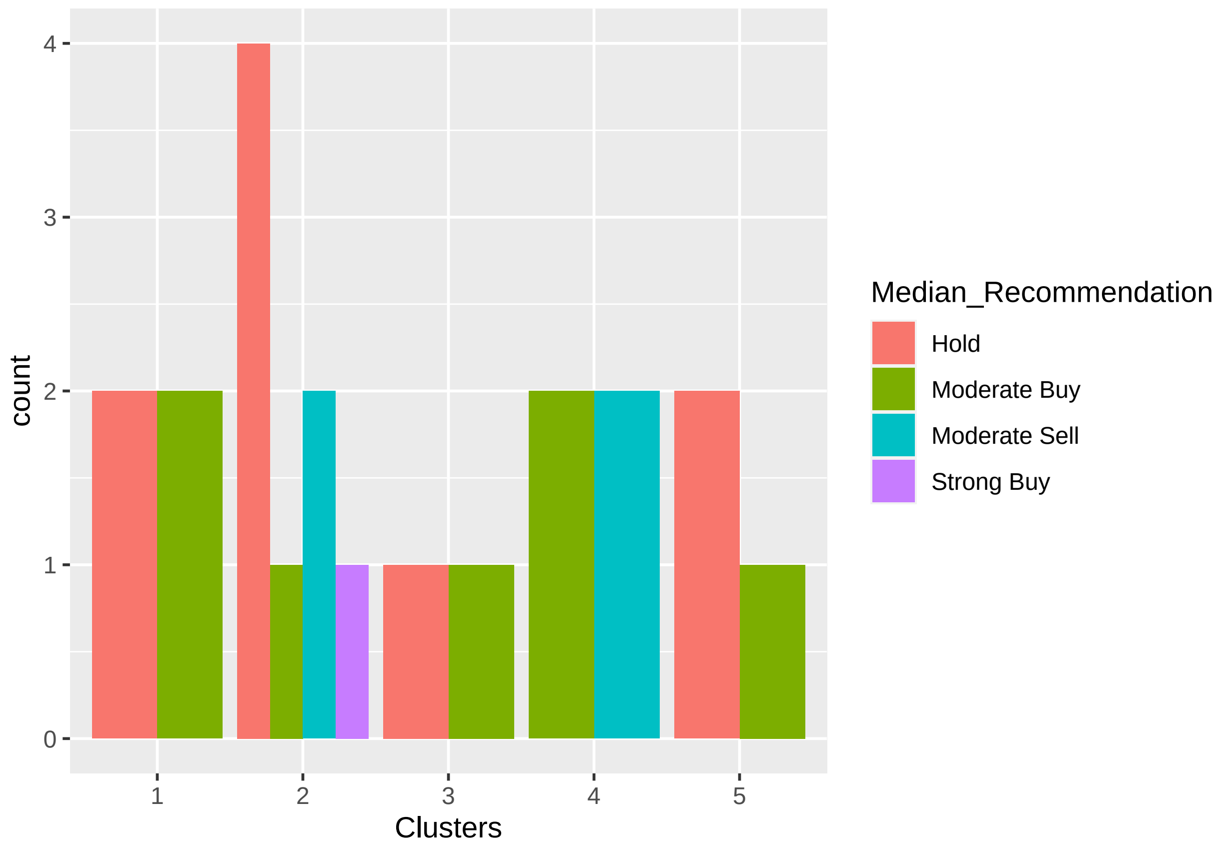
Cluster 4:Distinct attributes are reflected in the mean values of Cluster 4.

Cluster 5:Specific patterns in mean values define the characteristics of Cluster 5.

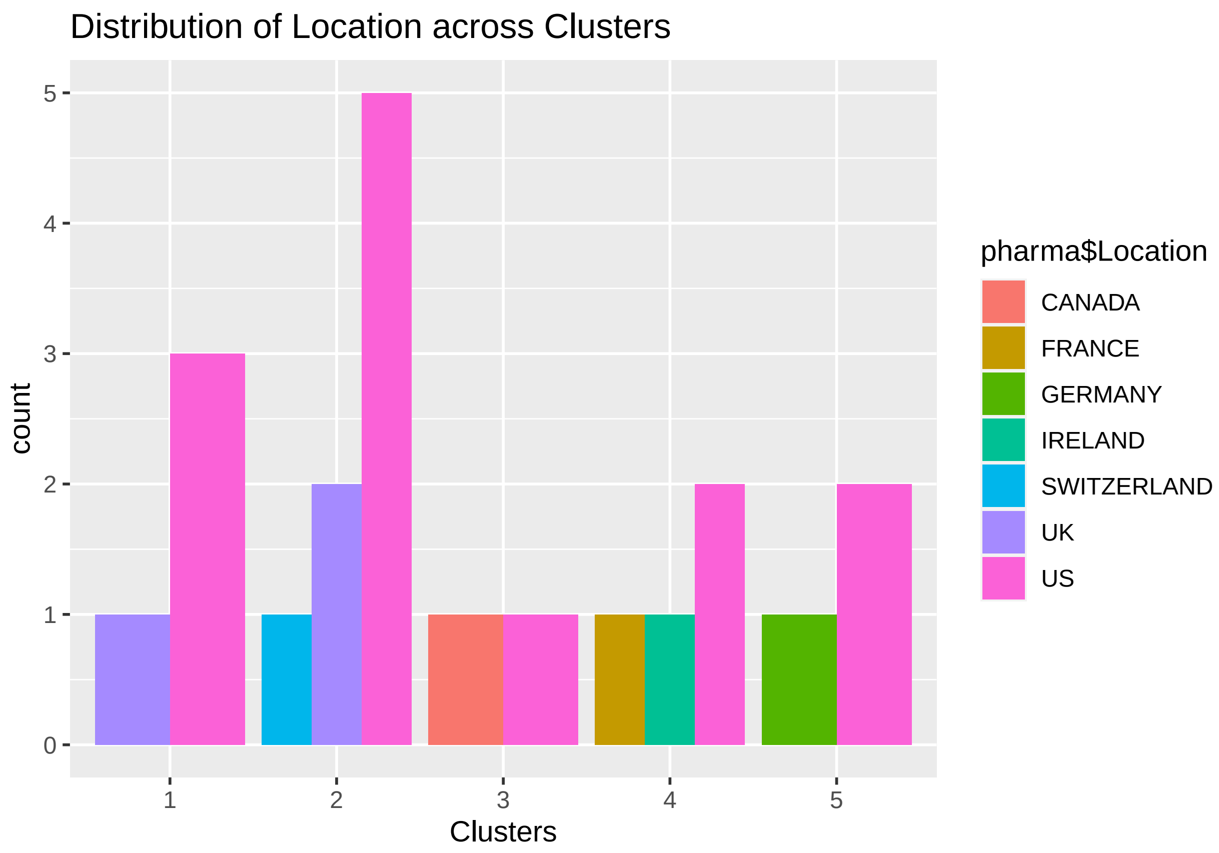
Is there a pattern in the clusters with respect to the numerical variables

#(10 to 12)? (those not used in forming the clusters)

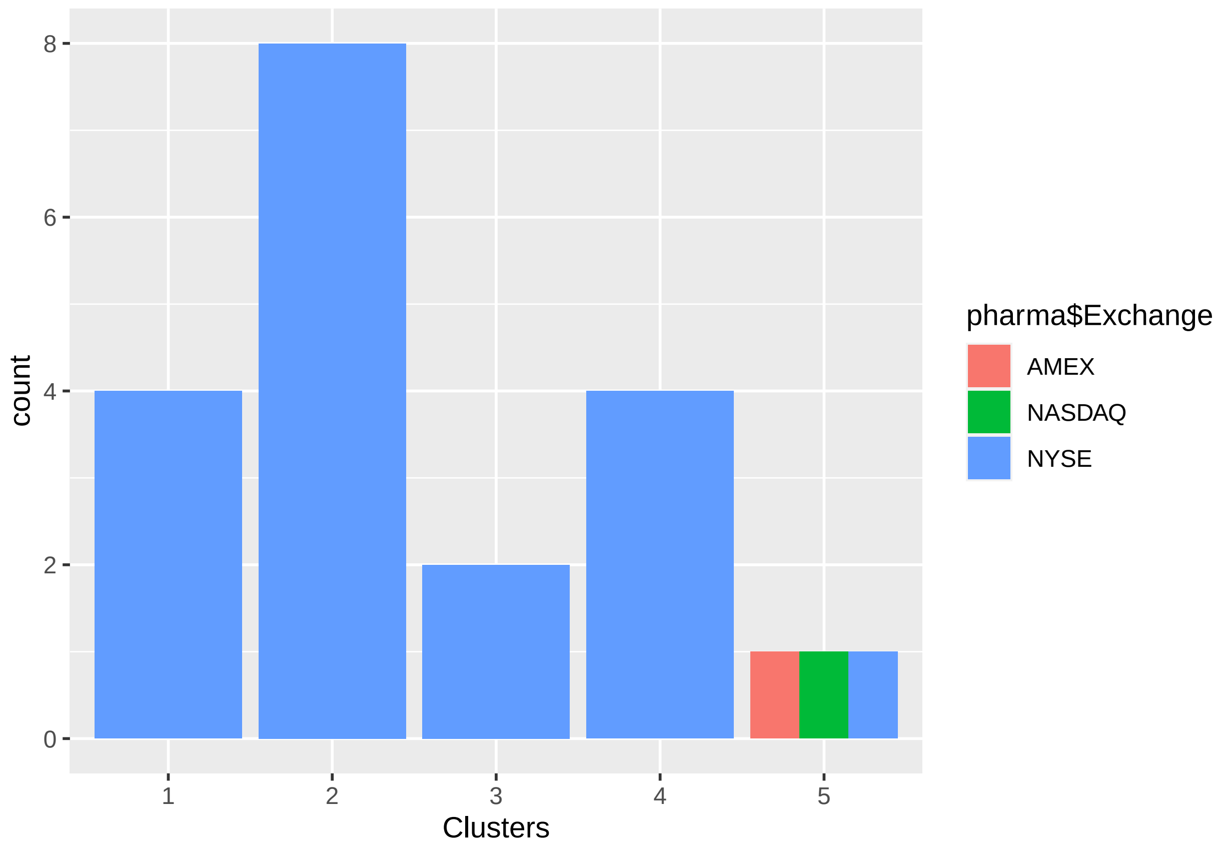
```r
pharma3 <- pharma[10:12] %>% mutate(Clusters=k5$cluster)
ggplot(pharma3, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position='dodge'
```



```r
ggplot(pharma3, aes(x = factor(Clusters), fill =pharma$Location)) +
  geom_bar(position = 'dodge') +
  labs(x = 'Clusters') +
  ggtitle('Distribution of Location across Clusters')
```



```
ggplot(pharma3, mapping = aes(factor(Clusters),fill=pharma$Exchange))+geom_bar(position = 'dodge')+labs
```


count
8
6
4
2
0
1
2
3
4
5
Clusters
pharma$Exchange
AMEX
NASDAQ
NYSE

Cluster 1: It has the highest PE_Ratio and needs to be held as per the media recommendations.

Cluster 2: It has the highest market_Cap and has Good Leverage value. And they can be moderately recommended.

Cluster 3: It has lowest asset_turnover,and lowest beta. But media recommendations are highly positive.

Cluster 4: The leverage ratio is high, they are moderately recommended.

Cluster 5: They have lowest revenue growth, highest assest turnover and highest net profit margin.

They are recommended to be held for longer time.

Question C:#Using any or all of the variables in the dataset, give each cluster a suitable name.

Cluster 1: Balanced Performers: The name of this cluster implies that the companies within it have respectable and consistent financial indicators. It suggests a well-rounded performance in all areas of finances.

Cluster 2-Steady Growing Contenders: As suggested by their name, these businesses exhibit steady development, which makes them a moderately risk-free but dependable choice for holding or investing. It exhibits both stability and room for expansion.

Cluster 3: Dynamic Opportunity Firms: As suggested by the name, companies in this cluster may offer a variety of investment options, which are marked by higher risk (sell) as well as possible growth (buy). It alludes to performance dynamism and variety.

Cluster 4-Stable Investment Picks: This moniker highlights companies that exhibit strong financial performance and stability, which makes them desirable for long-term investment and purchase.

Cluster 5: Long-term Value Holders: As implied by the name, companies in this cluster are good investments since they have the ability to generate long-term value. They are probably distinguished by high asset turnover and moderate but steady revenue growth.