Assignment_4

Harshith Kumar Yadav Temura

2023-11-12

Required packages

```
library(flexclust)
```

```
## Warning: package 'flexclust' was built under R version 4.2.3

## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4
```

```
library(cluster)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'ggplot2' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'readr' was built under R version 4.2.3

## Warning: package 'purrr' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.3

## Warning: package 'forcats' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3
```

```
## -- 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
```

```r
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.2.3
```

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

```r
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.2.3
```

```r
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.2.3
```

1.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.

Loading the data

```r
pharma<- read.csv("C:\\Users\\CherRyY\\Downloads\\Pharmaceuticals.csv")

head(pharma)
```

```
##   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
```

Choosing columns 3 to 11 now, and putting the information in variable Info 1

```
pharma1 <- pharma[3:11]

head(pharma1)
```

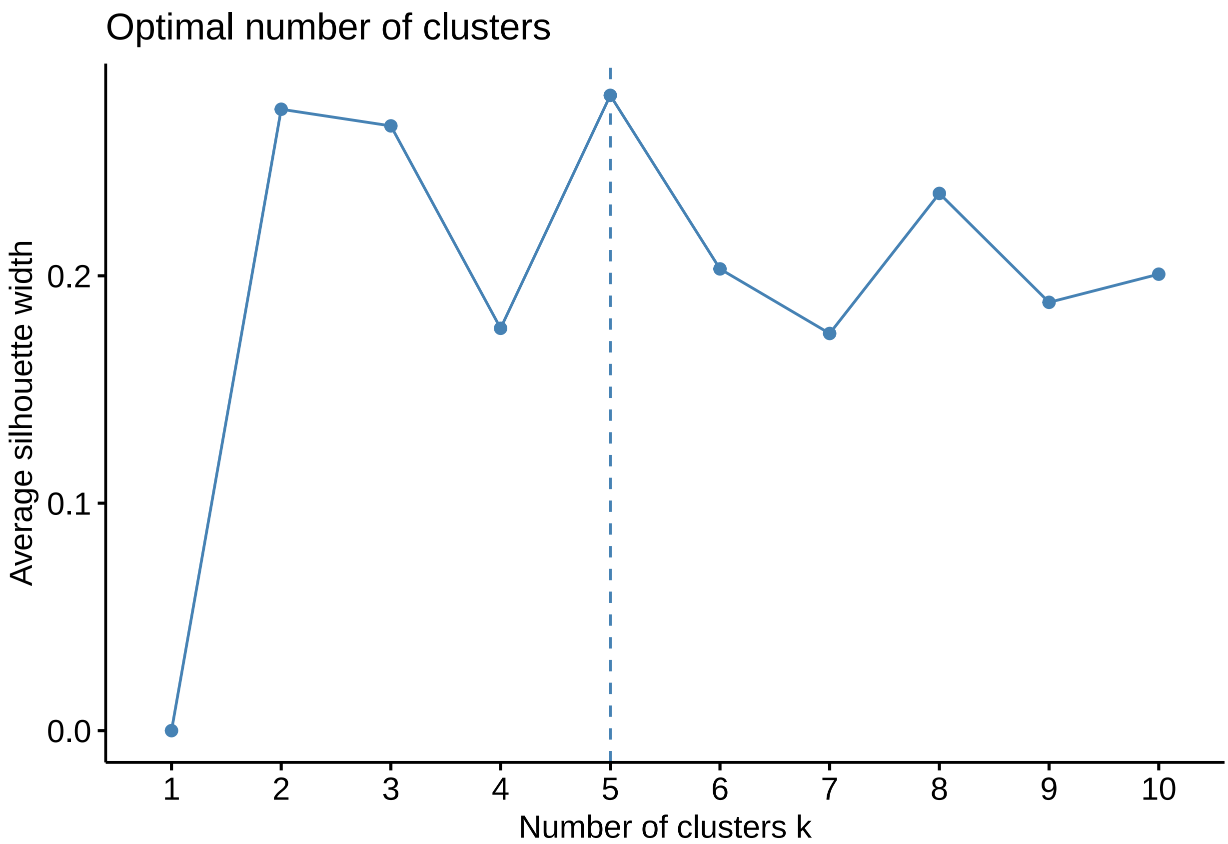
```
##   Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## 1      68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54
## 2       7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16
## 3       6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05
## 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
```

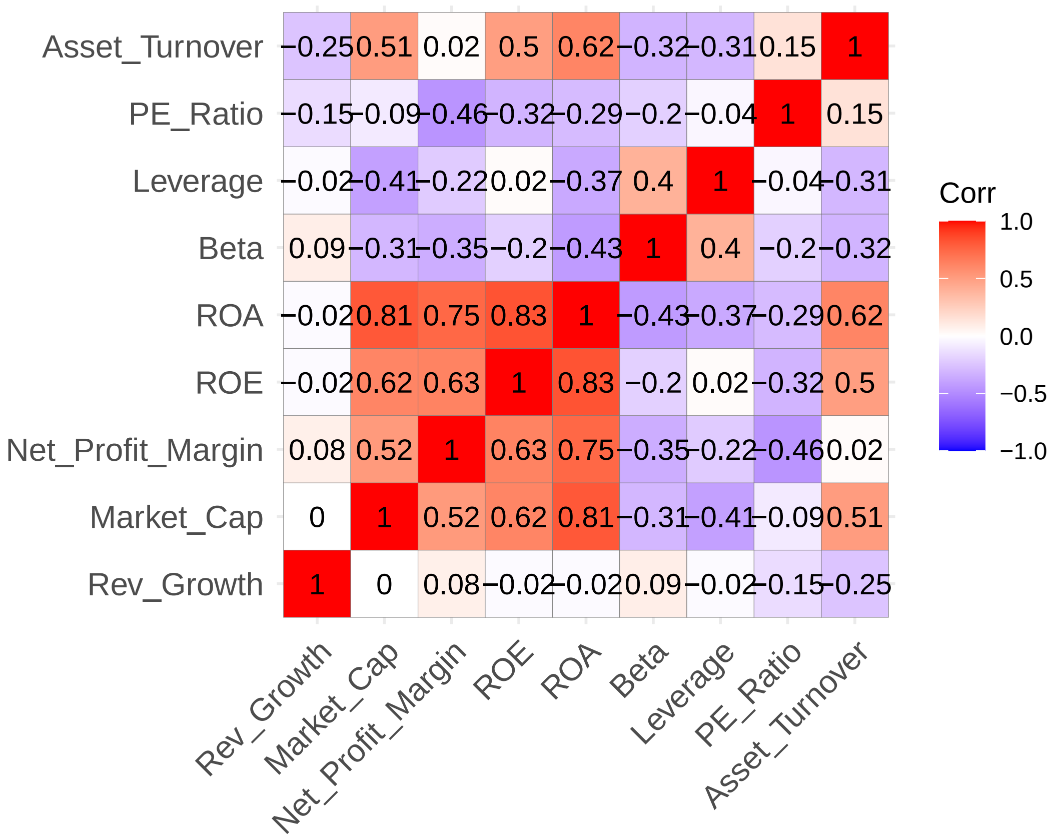
The different values allocated to each variable along the rows will be used to scale the data in pharma1 and the pharma updated dataframe. calculating the distance between data rows and visualizing the distance matrix using the get dist and fviz dist functions of the factoextra package

```
norm_data <- scale(pharma1)
row.names(norm_data) <- pharma[,1]
distance <- get_dist(norm_data)
corr <- cor(norm_data)
fviz_nbclust(norm_data,kmeans,method = "silhouette")
```



Make a correlation matrix and print it to see which variables are correlated.

```
corr <- cor(norm_data)
ggcorrplot(corr, outline.color = "grey50", lab = TRUE, hc.order = TRUE, type = "full")
```

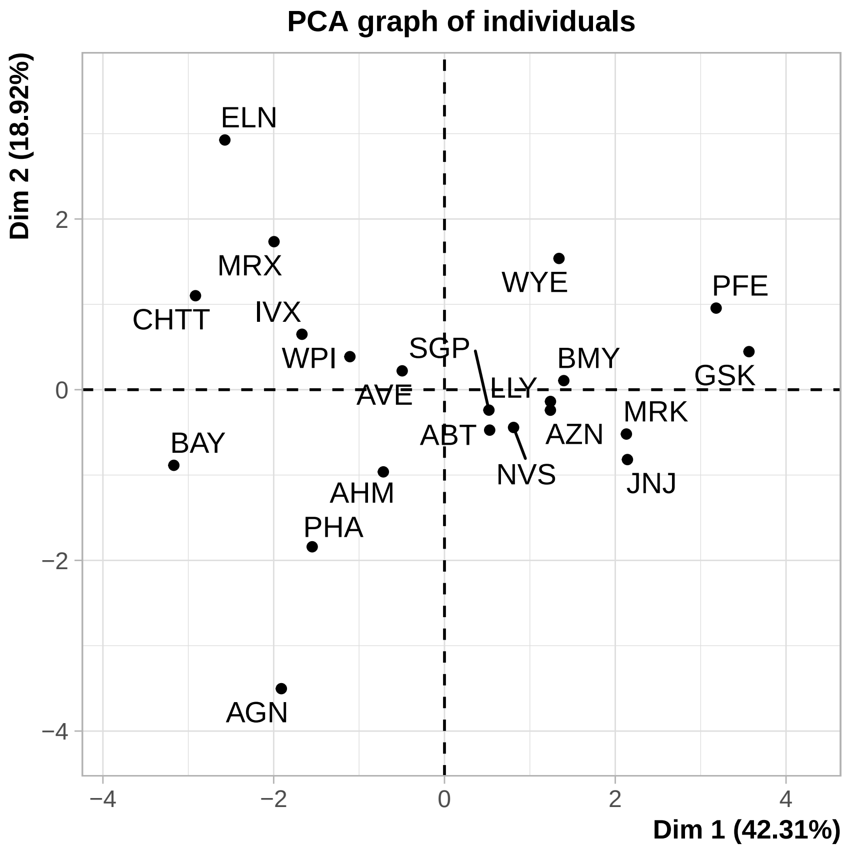


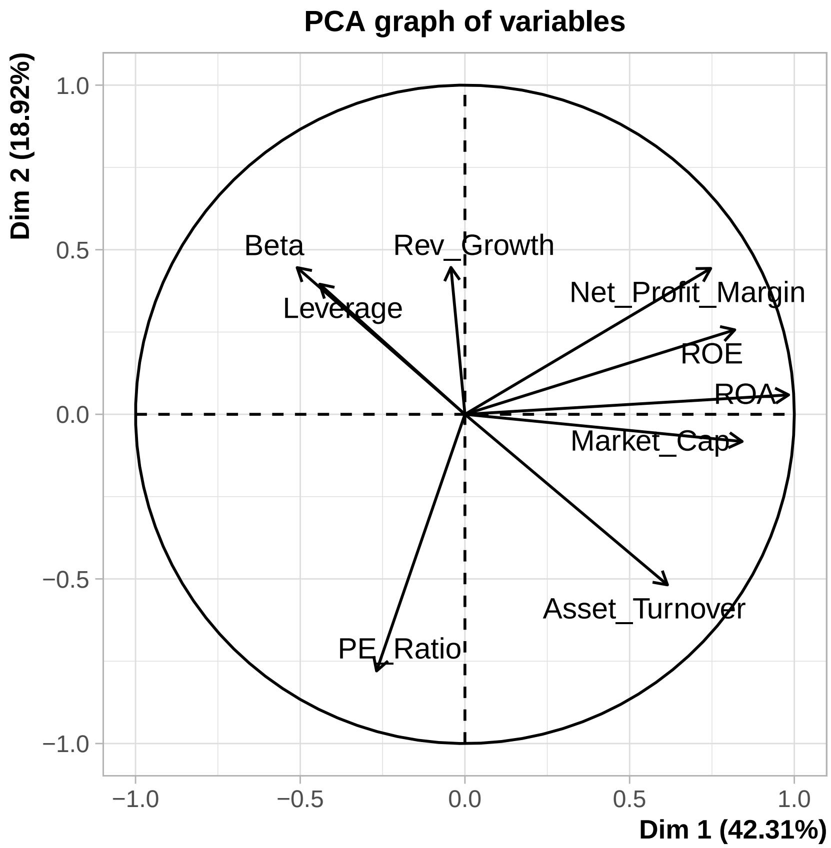
The ROA, ROE, Net Profit Margin, and Market Cap are all high, according to the Correlation Matrix

Principal component analysis will be used to determine the relative importance of each of the key variables in the data collection.

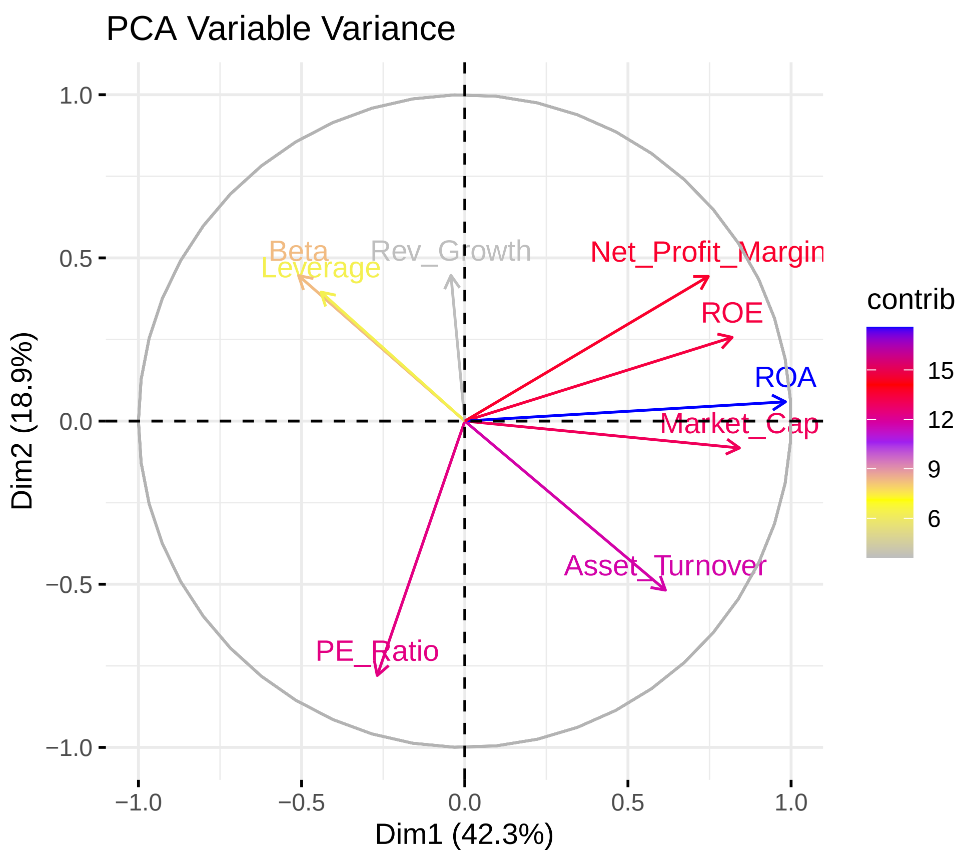
`assuming the optimal cluster size is 5`

```
pca <- PCA(norm_data)
```


PCA graph of individuals
Dim 2 (18.92%)
Dim 1 (42.31%)
ELN
MRX
CHTT
IVX
WPI
SGP
AVE
WYE
PFE
BMY
GSK
LLY
MRK
ABT
AZN
BAY
NVS
JNJ
AHM
PHA
AGN
2
0
−2
−4
−4
−2
0
2
4



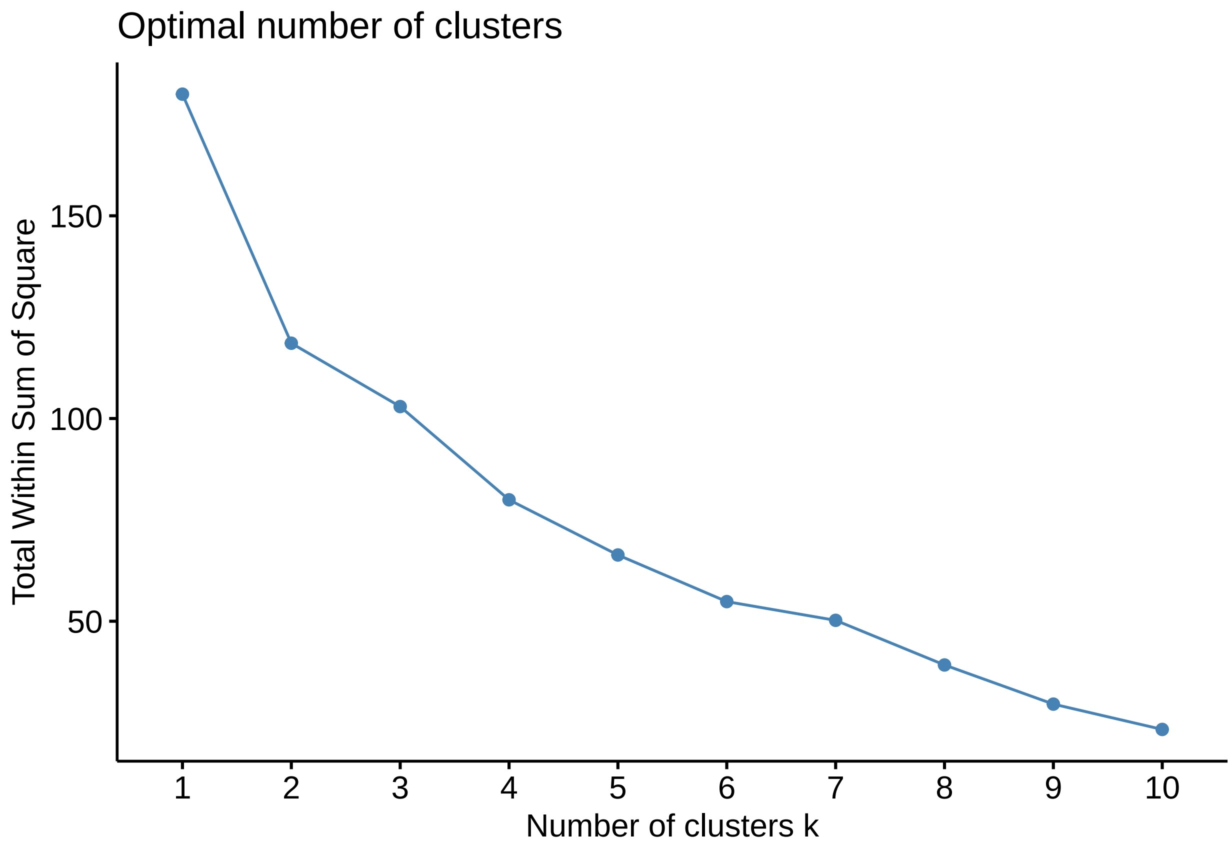

```
var <- get_pca_var(pca)
fviz_pca_var(pca, col.var="contrib",
             gradient.cols = c("grey","yellow","purple","red","blue"),ggrepel = TRUE ) + labs( title =
```



We may deduce from PCA Variable Variance that ROA, ROE, Net Profit Margin, Market Cap, and Asset Turnover contribute more than 61% to the two PCA components/dimensions, using the elbow technique to determine the optimal customer base.

```
set.seed(10)

wss <- vector()
for(i in 1:10) wss[i] <- sum(kmeans(norm_data,i)$withinss)
fviz_nbclust(norm_data, kmeans, method = "wss")
```




```
wss
```

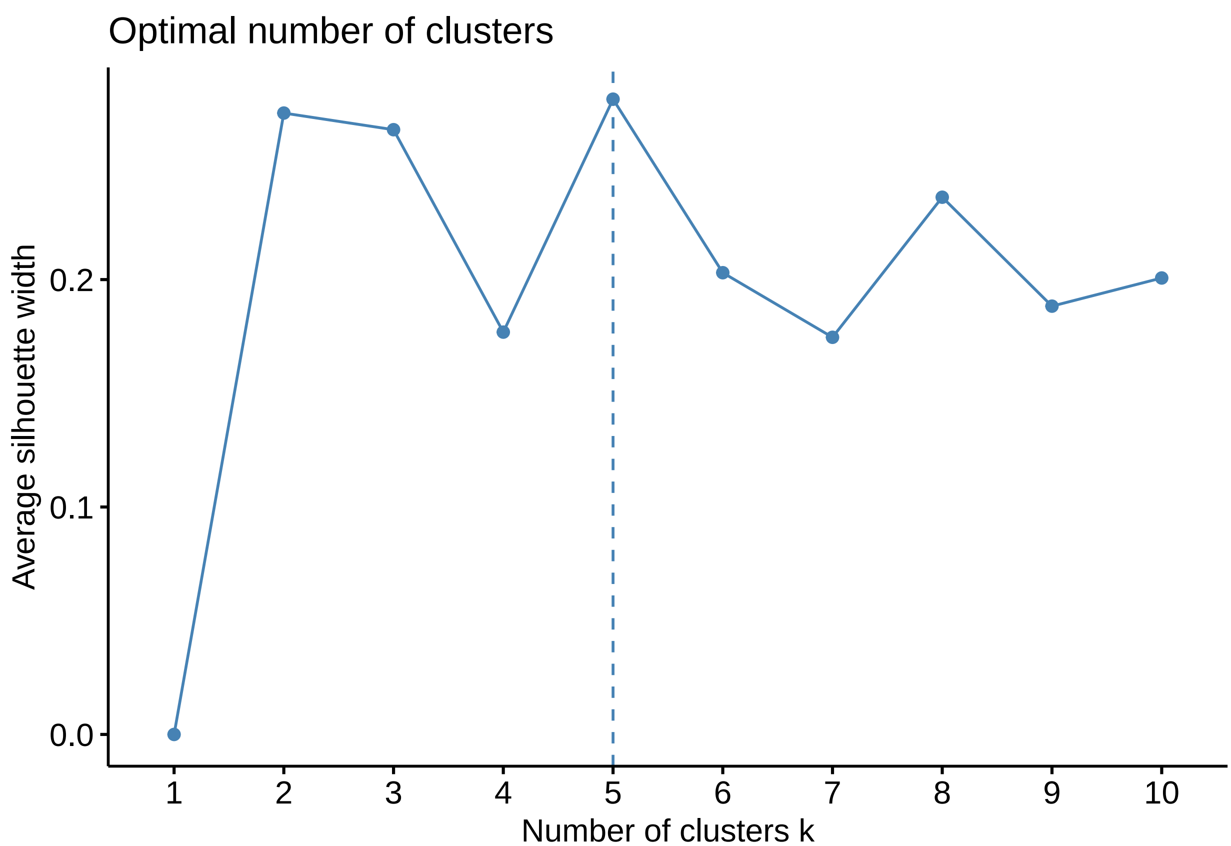
```
##  [1] 180.00000 118.56934  95.99420  79.21748  65.61035  52.67476  47.66961
##  [8]  41.12605  31.81763  31.57252
```

As expected, number 5 is the ideal cluster..

Determining the optimal cluster size.

Silhouette*

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



This indicates that the ideal number of clusters is 5. forming five clusters with the k-means algorithm

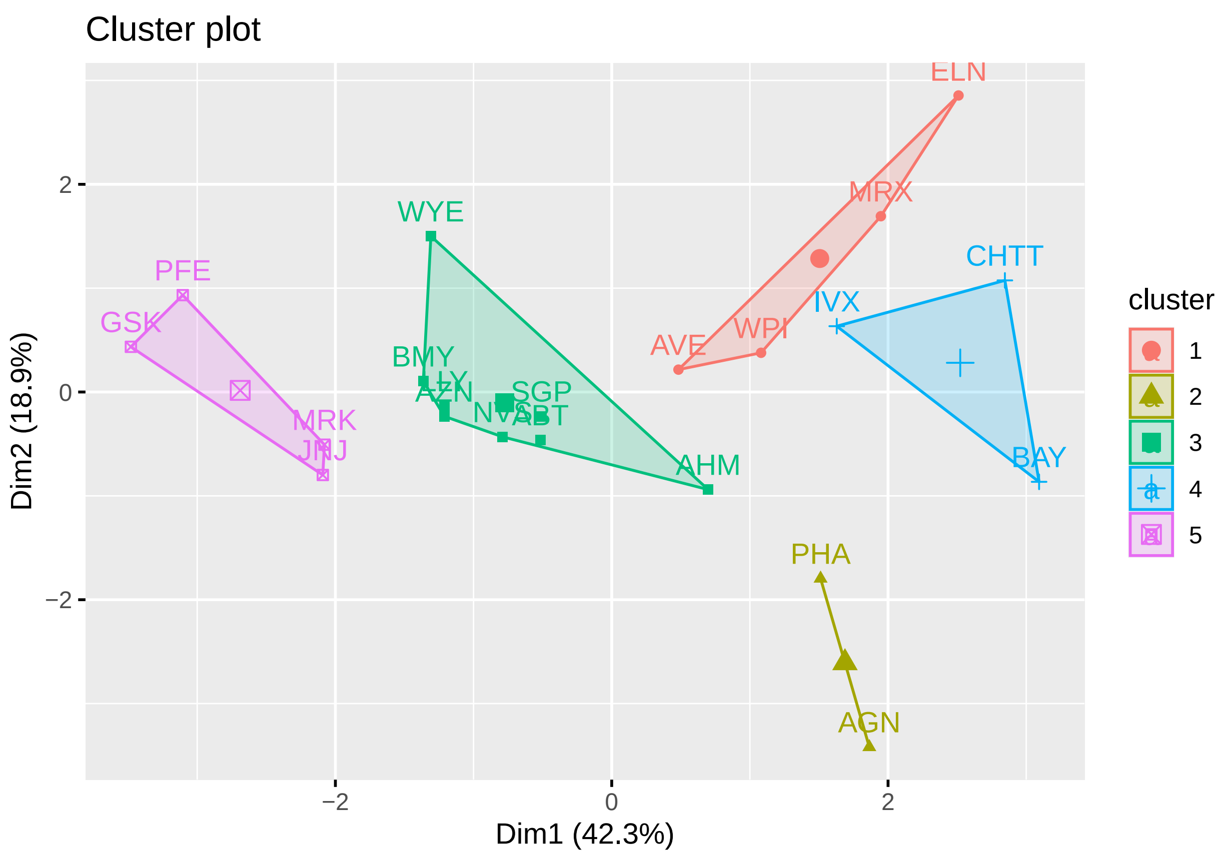
```
set.seed(1)
k5 <- kmeans(norm_data, centers = 5, nstart = 31) # k = 5, number of restarts = 31
k5$centers
```

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

```
k5$size
```

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

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




```r
set.seed(15)
k51 = kcca(norm_data, k=5, kccaFamily("kmedians"))
k51
```

Manhattan Distance when using Kmeans Clustering.

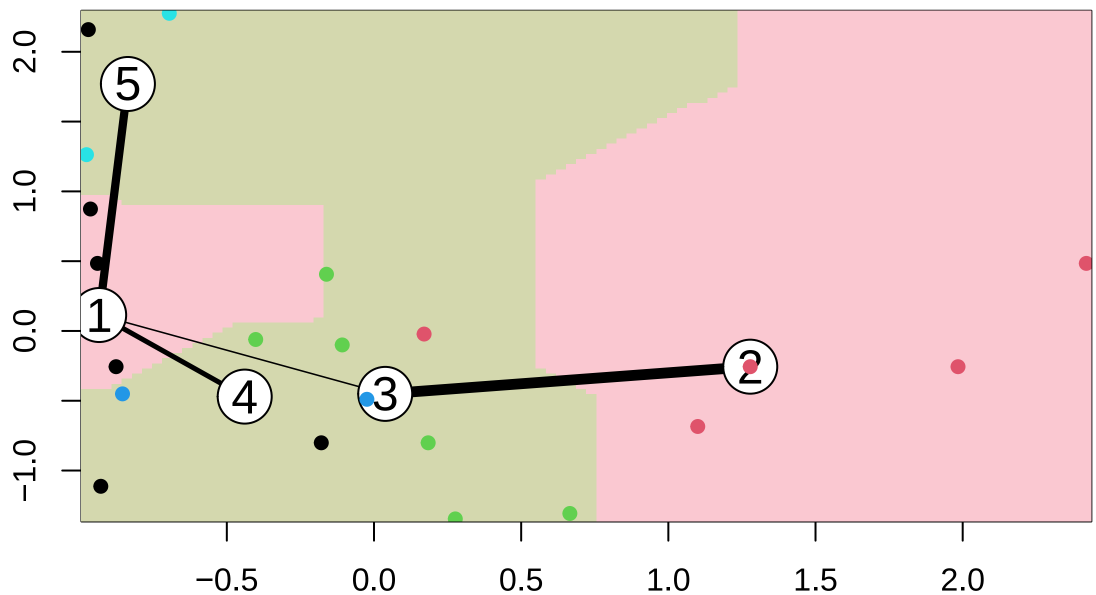
```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = norm_data, k = 5, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
## 1 2 3 4 5
## 6 5 6 2 2
```

```r
clusters_index <- predict(k51)
dist(k51@centers)
```

Using predict function.

```
##          1        2        3        4
## 2 3.945545
## 3 3.168054 2.377053
## 4 3.724526 4.795056 4.301987
## 5 3.578425 5.494529 4.448919 4.043870
```

```
image(k51)
points(norm_data, col=clusters_index, pch=19, cex=0.9)
```



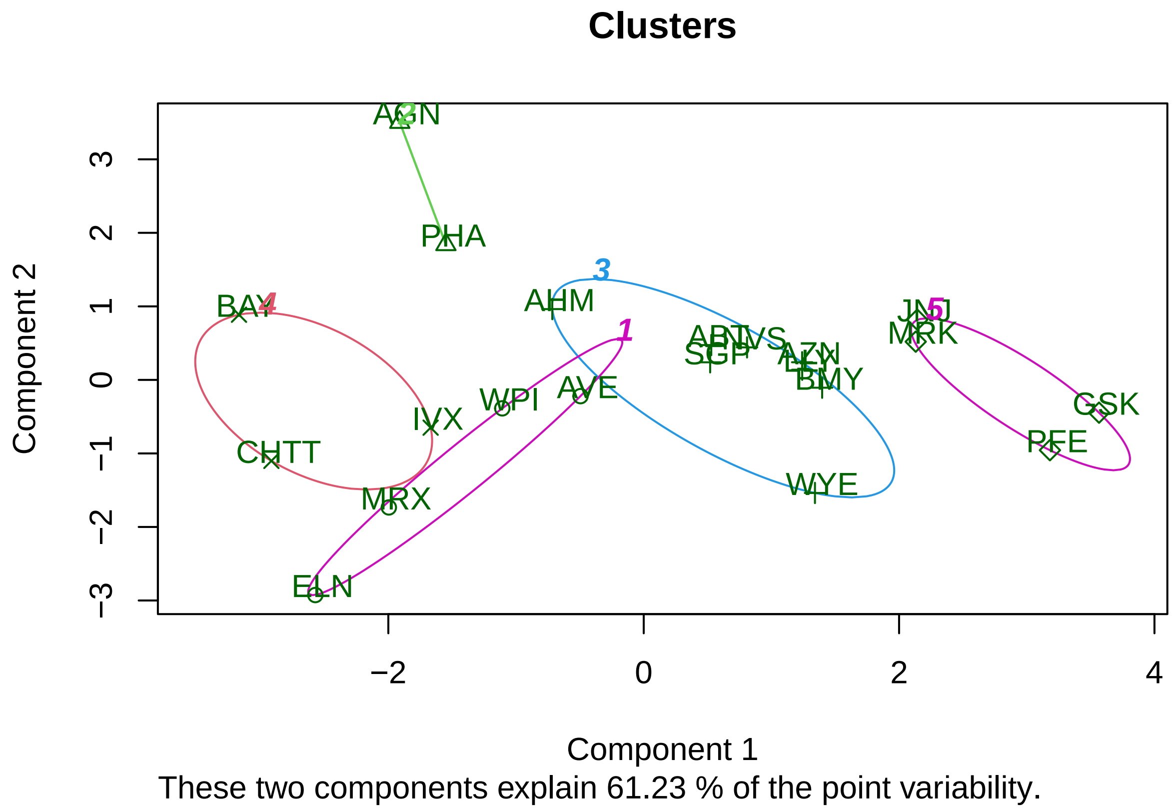
2.Interpret the clusters with respect to the numerical variables used in forming the clusters Using Kmeans method to calculate Mean.

```
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      13.1  0.598     17.7  14.6  6.2           0.425    0.635
## 2       2      31.9  0.405     69.5  13.2  5.6           0.75     0.475
## 3       3      55.8  0.414     20.3  28.7 12.7           0.738    0.371
## 4       4       6.64 0.87      24.6  16.5  4.17          0.6      1.65
```

```
## 5        5      157.    0.48       22.2   44.4 17.7            0.95     0.22
## # i 2 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>
```

```r
clusplot(norm_data,k5$cluster, main="Clusters",color = TRUE, labels = 2,lines = 0)
```



Businesses are divided into the following distinct cluster:

- Cluster 1: ELN, MRX, WPI and AVE-
- Cluster 2: AGN and PHA-
- Cluster 3: AHM,WYE,BMY,AZN, LLY, ABT, NVS and SGP-
- Cluster 4: BAY, CHTT and IVX-
- Cluster 5: JNJ, MRK, PFE and GSK-

The following can be obtained from the cluster variables' means:

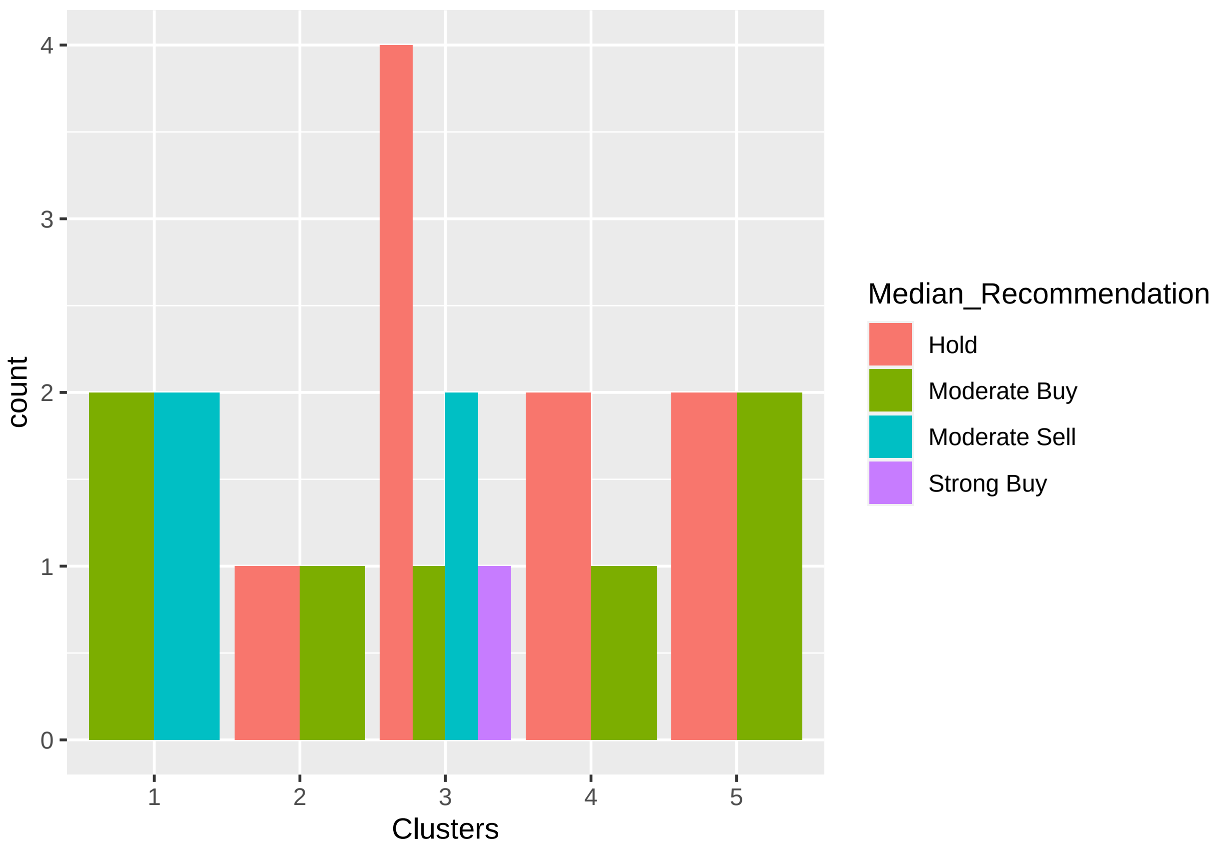
- With the fastest sales growth, the lowest PE ratio, and the largest net profit margin, Cluster 1 leads the pack. It can be purchased or held in reserve.-
- Cluster 2 PE ratio is very high.-
- Cluster 3 has a medium risk.-
- Group 4 Its extremely high risk, extremely high leverage, and weak Net Profit margin make it exceedingly dangerous to purchase, even with its great PE ratio. Revenue growth is likewise quite low.-

- Strong market capitalization, return on investment, return on assets, return on asset turnover, and return on net profit margin characterize Cluster 5. A low price-to-earnings ratio suggests that the company is reasonably valued and can be purchased and held. An 18.5% increase in revenue is also advantageous.-

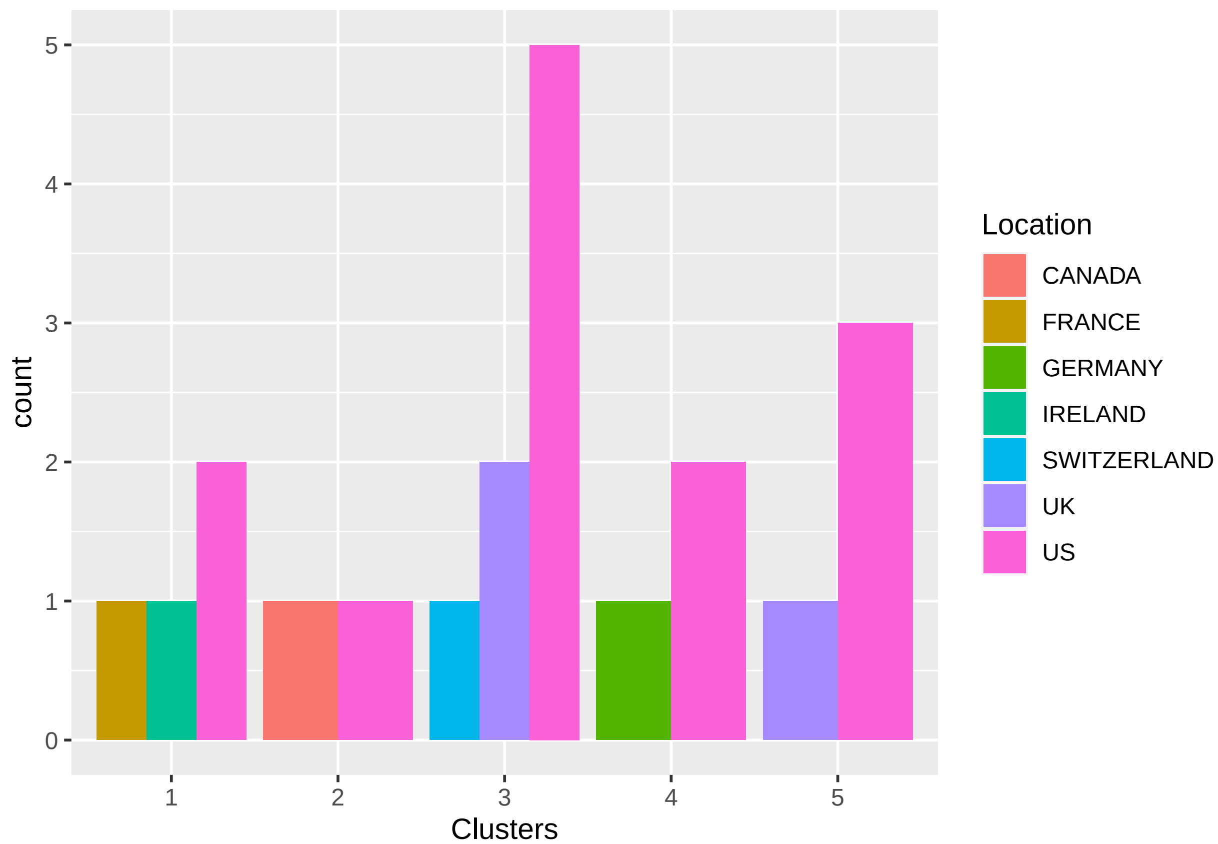
2B Is there a pattern in the clusters with respect to the numerical variables (10 to 12)?

```
By comparing clusters to the variables, we can visualize patterns.
```

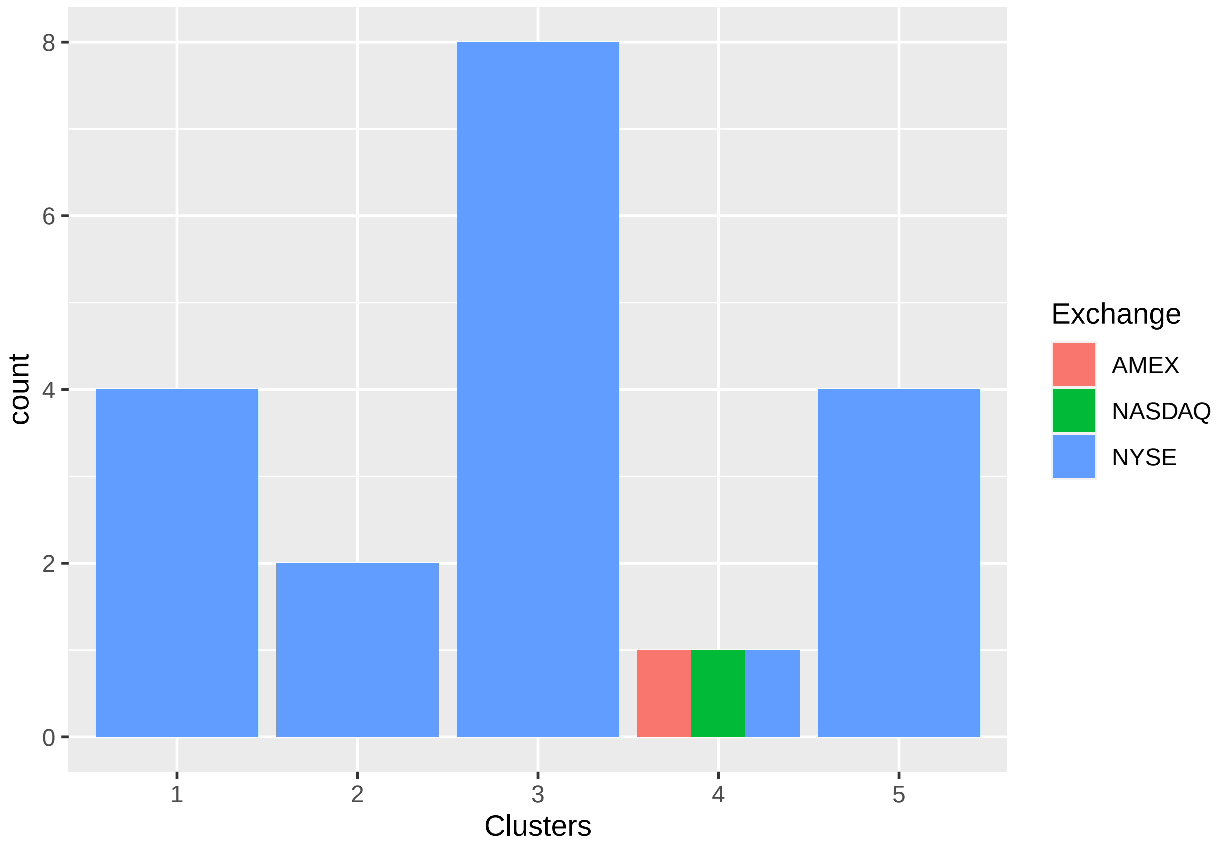
```
Info_2 <- pharma[12:14] %>% mutate(Clusters=k5$cluster)
ggplot(Info_2, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position='dodge')+
```




```
ggplot(Info_2, mapping = aes(factor(Clusters),fill = Location))+geom_bar(position = 'dodge')+labs(x ='Cl
```




```
ggplot(Info_2, mapping = aes(factor(Clusters),fill = Exchange))+geom_bar(position = 'dodge')+labs(x ='Cl
```




```
The variable grouped together, The median recommendations show a pattern.

The most of the clusters/companies are listed on the NYSE and are based in the United States, but other
```

3. Provide an appropriate name for each cluster using any or all of the variables in the data set.

```
Here, I've taken into account Market Cap, Beta, PE Ratio, ROE, ROA, and Asset Turnover when naming the c
```

Cluster 1: Profitable Giants

- noticed for having a large market capitalization, a low beta, a low PE ratio, a strong return on assets, ROE, and ROA. These organizations stand in for powerful, lucrative industry titans.-

Cluster 2: High Beta, High Risk Players

- Cluster 2 denotes businesses with higher risk levels and is identified by heightened Beta and PE Ratio. Due to potential overvaluation and increasing market sensitivity, investors should proceed with caution.-

Cluster 3: Balanced Performers

- Cluster 3 represents businesses in a moderate-risk category by balancing Market Cap, Beta, and PE Ratio. These well-balanced performers show promise and stability.-

Cluster 4: High Risk, Low Efficiency

- Entities in Cluster 4 suffer very high risk despite having a great PE Ratio; low efficiency is demonstrated by low ROE, ROA, and asset turnover. This cluster is thought to be less effective and high-risk.-

Cluster 5: Efficient Powerhouses

- Cluster 5 presents companies with a modestly valued PE Ratio along with strong efficiency measures, such as high ROE, ROA, and asset turnover. These effective workhorses are desirable for acquisition as well as retention.-