# ML Assignment 4

Tejaswini Yeruva

2022-11-02

#Loading the Required packages

```
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)
library(tidyverse)
```

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

## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
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(FactoMineR)
```

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

```r
library(ggcorrplot)
```

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

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.

## Loading the data

```r
getwd()
```

```
## [1] "C:/Users/tejar/OneDrive/Desktop/ML Assignments"
```

```r
setwd("C:/Users/tejar/OneDrive/Documents")
Info<- read.csv("C:/Users/tejar/Downloads/Pharmaceuticals.csv")


## Now selecting columns from 3 to 11 and storing the data in variable Info1

Info1 <- Info[3:11]

## Using head function to display the first 6 rows of data

head(Info1)
```

```
##   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
```

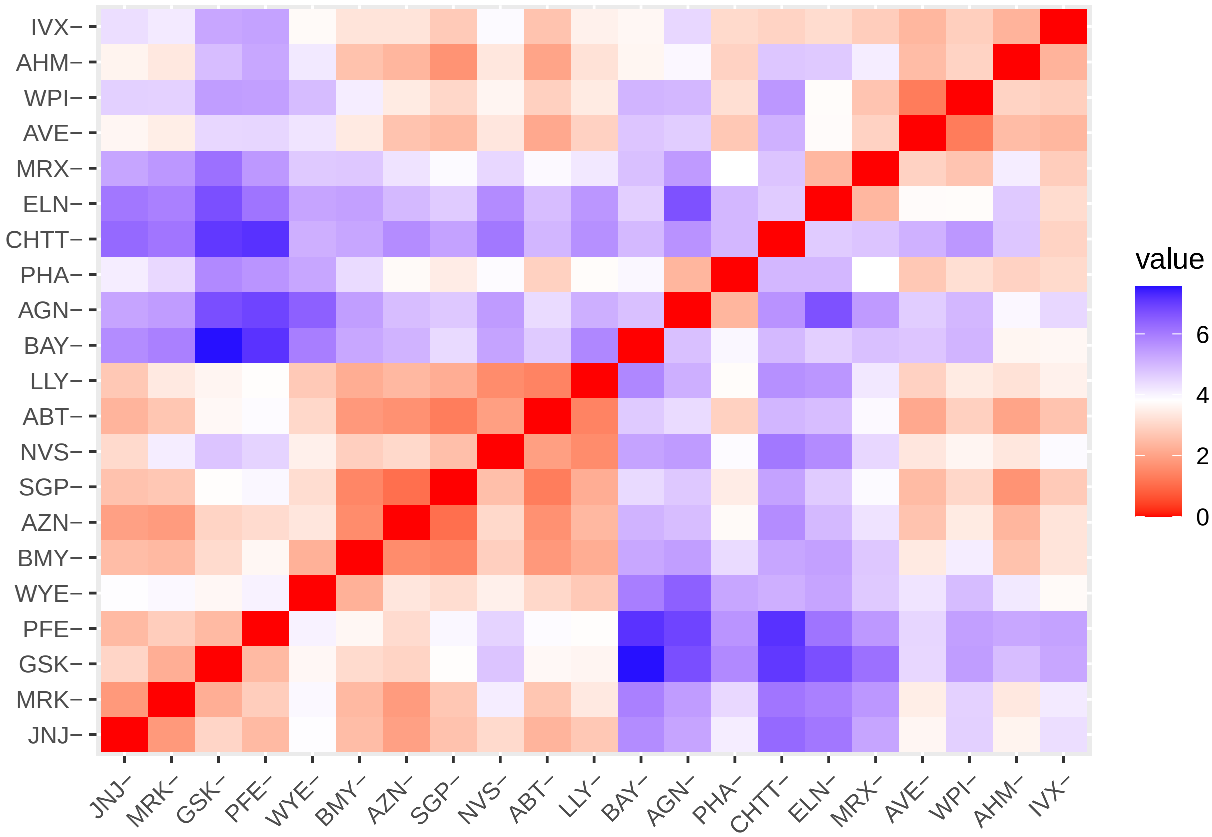
```r
summary(Info1)
```

```
##    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
```

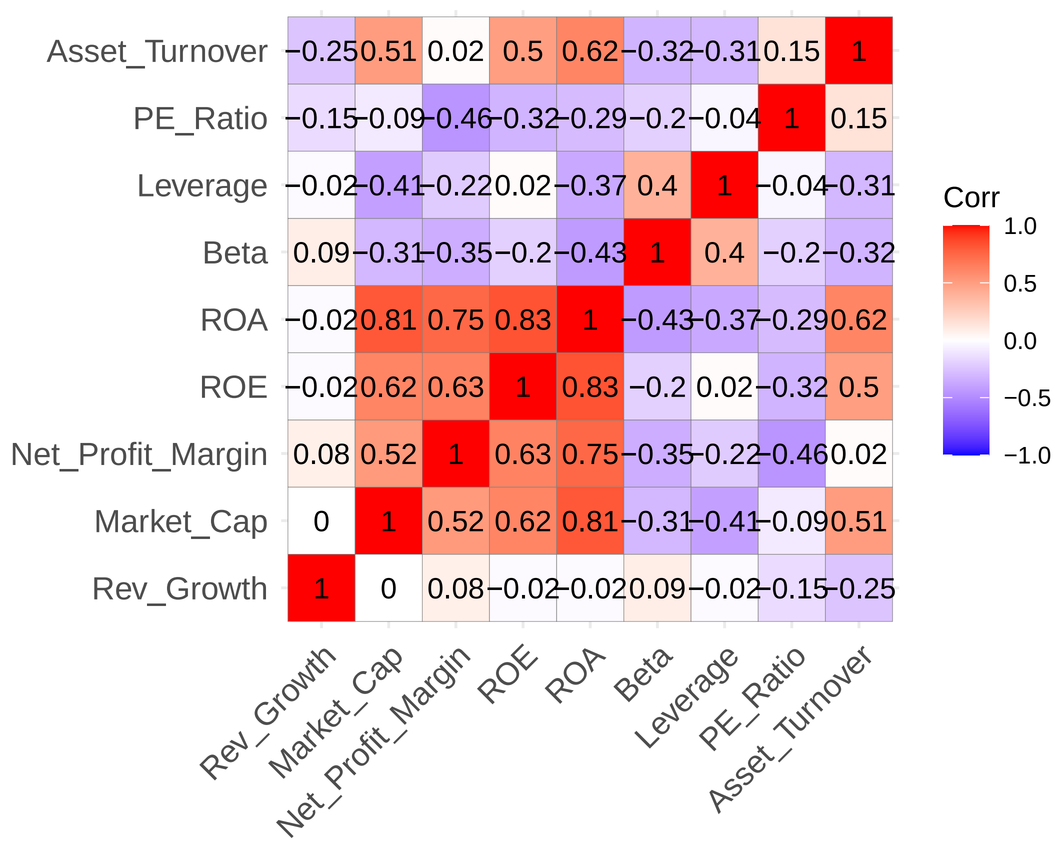
**All the variables are measured in different weights throughout the rows, and will scale the data in Info1 and save the scaled data in the Infoupdated dataframe. Calculating the distance between the rows of data and visualizing the distance matrix using get_dist and fviz_dist functions which are available in factoextra package.**

```
Infoupdated <- scale(Info1)
row.names(Infoupdated) <- Info[,1]
distance <- get_dist(Infoupdated)
fviz_dist(distance)
```




**Now creating the correlation Matrix and printing to check the correlation among major variables**

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

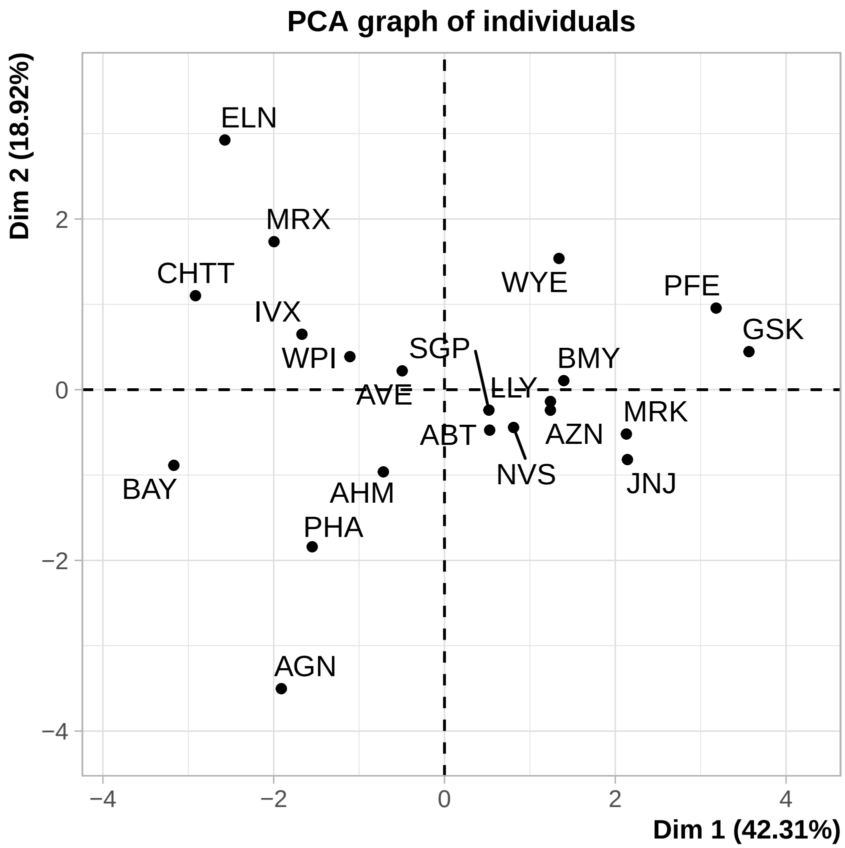


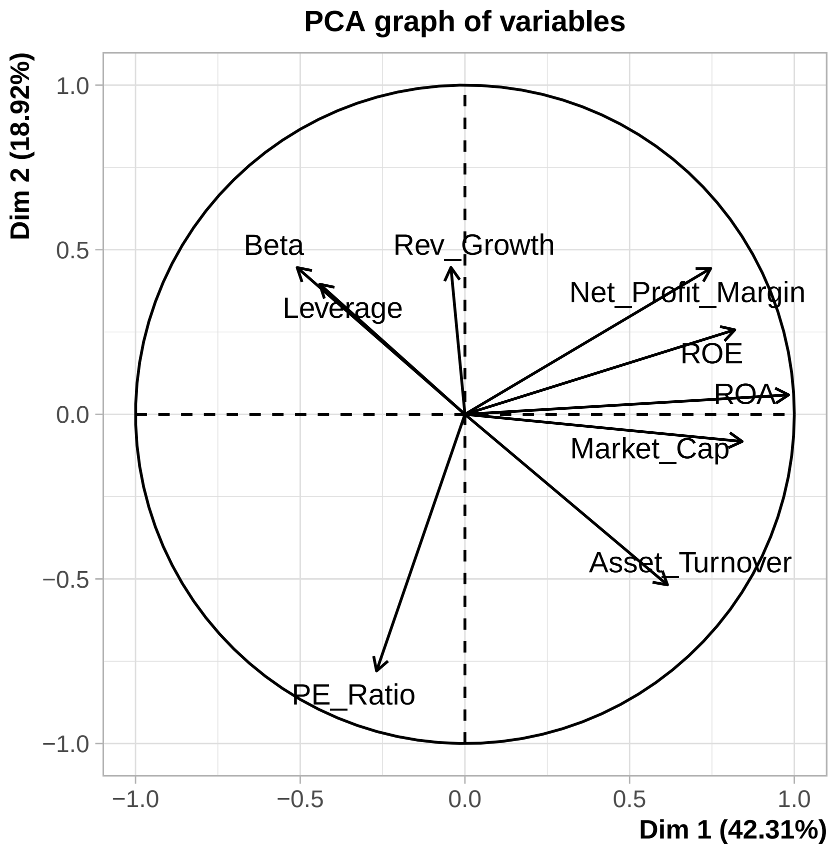

**The Correlation Matrix reveals that ROA, ROE, Net Profit Margin, and Market Cap is high.**

**By using Principal Component Analysis will be finding out weightage of major variables in the data set.**

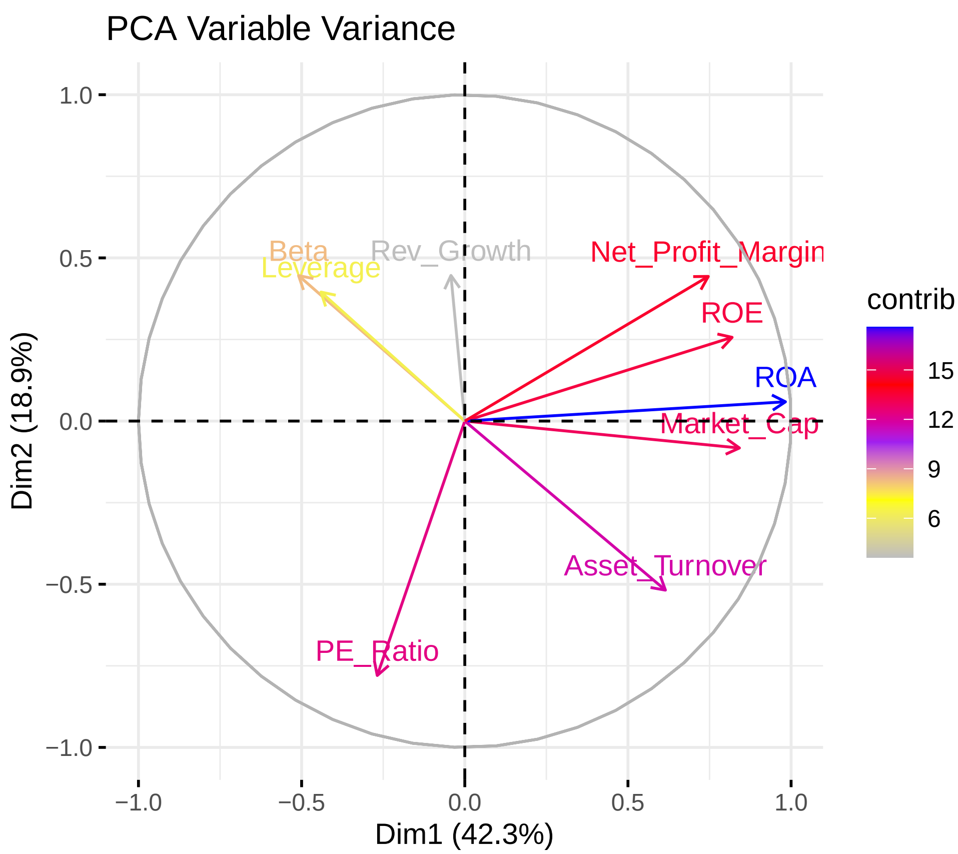
**assuming best number of cluster is 5.**

```
pca <- PCA(Infoupdated)
```


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

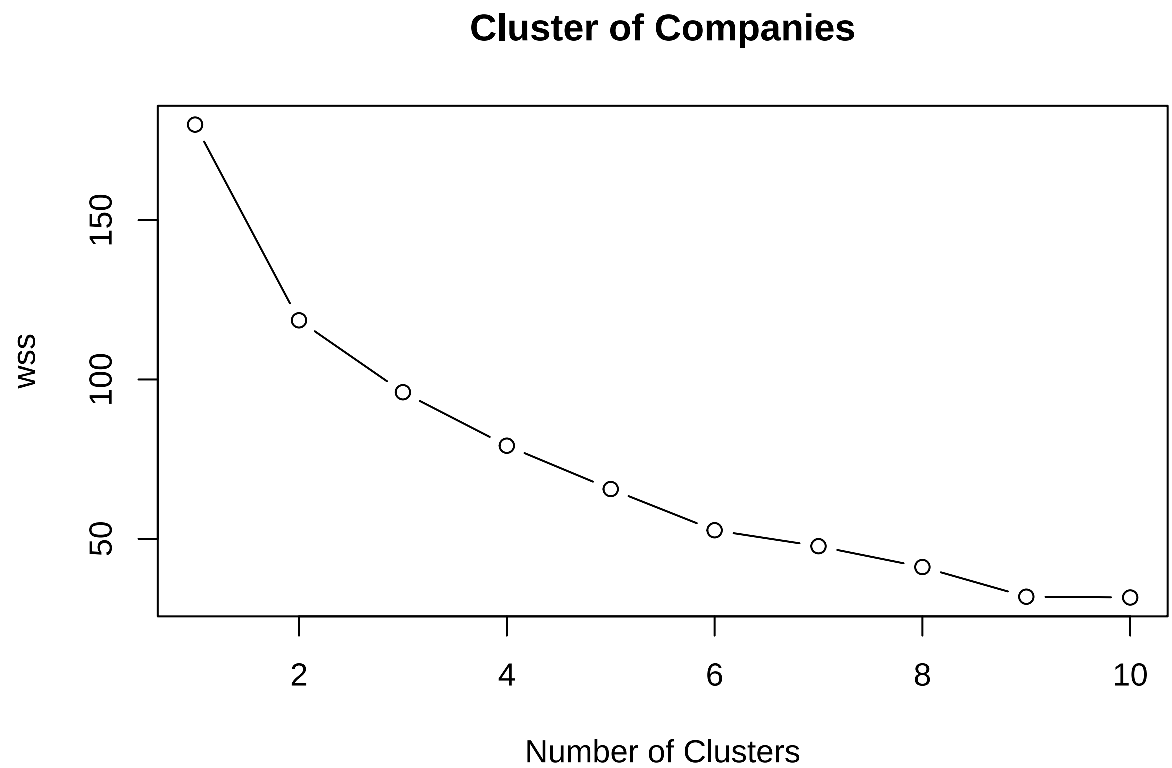



```r
var <- get_pca_var(pca)
fviz_pca_var(pca, col.var="contrib",
             gradient.cols = c("grey","yellow","purple","red","blue"),ggrepel = TRUE ) + labs( title =
```




**We can deduce from PCA Variable Variance that ROA,ROE, Net Profit Margin, Market Cap, and Asset Turnover contribute over 61% to the two PCA components/dimensions (Variables)and using elbow method to find optimal number of customers.**

```
set.seed(10)
wss <- vector()
for(i in 1:10) wss[i] <- sum(kmeans(Infoupdated,i)$withinss)
plot(1:10, wss , type = "b" , main = paste('Cluster of Companies') , xlab = "Number of Clusters", ylab=
```




```
wss
```

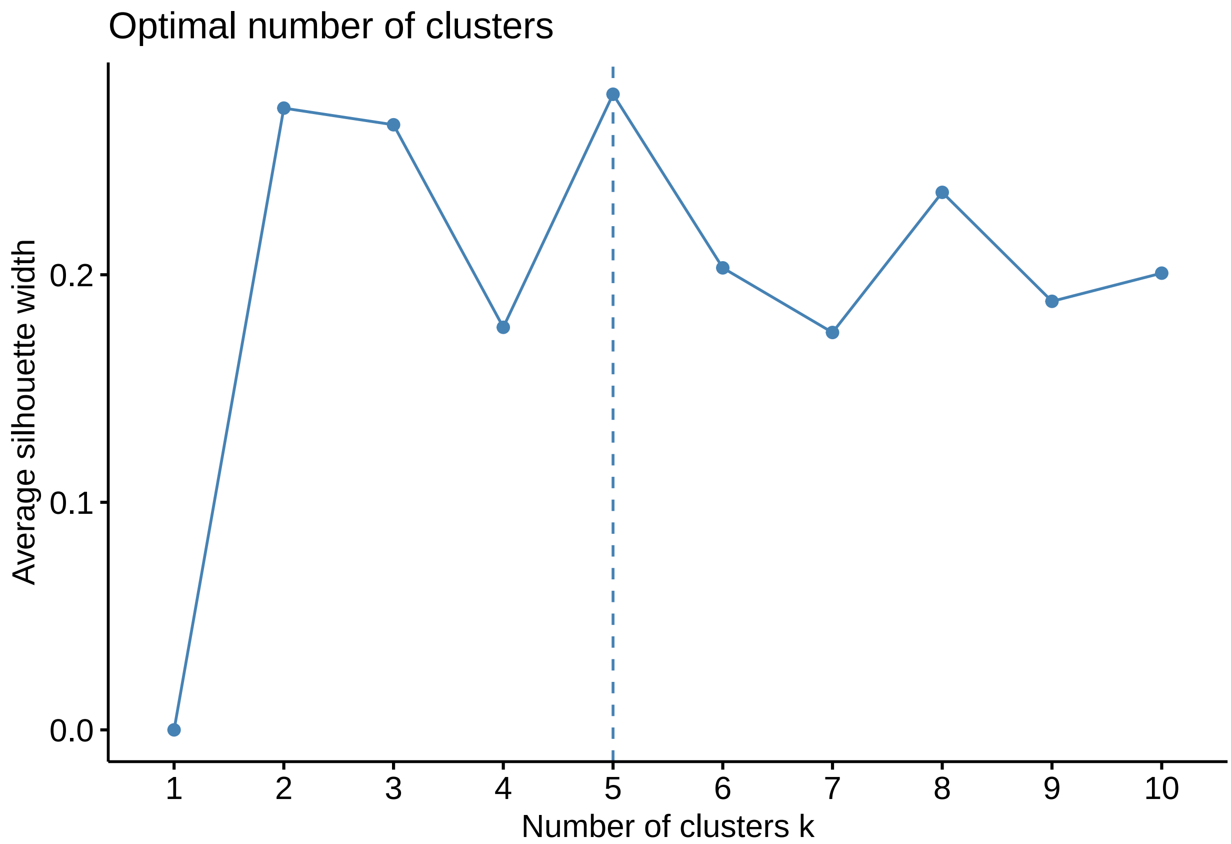
```
##  [1] 180.00000 118.56934  95.99420  79.21748  65.61035  52.67476  47.66961
##  [8]  41.12605  31.81763  31.57252
```

**Got the same number as assumed, optimal cluster is at 5.**

**Silhouette Method**

**Finding best number of clusters.**

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




**This also shows the idealnumber of clusters is 5. Using k-means algorithm to cluster with 5.**

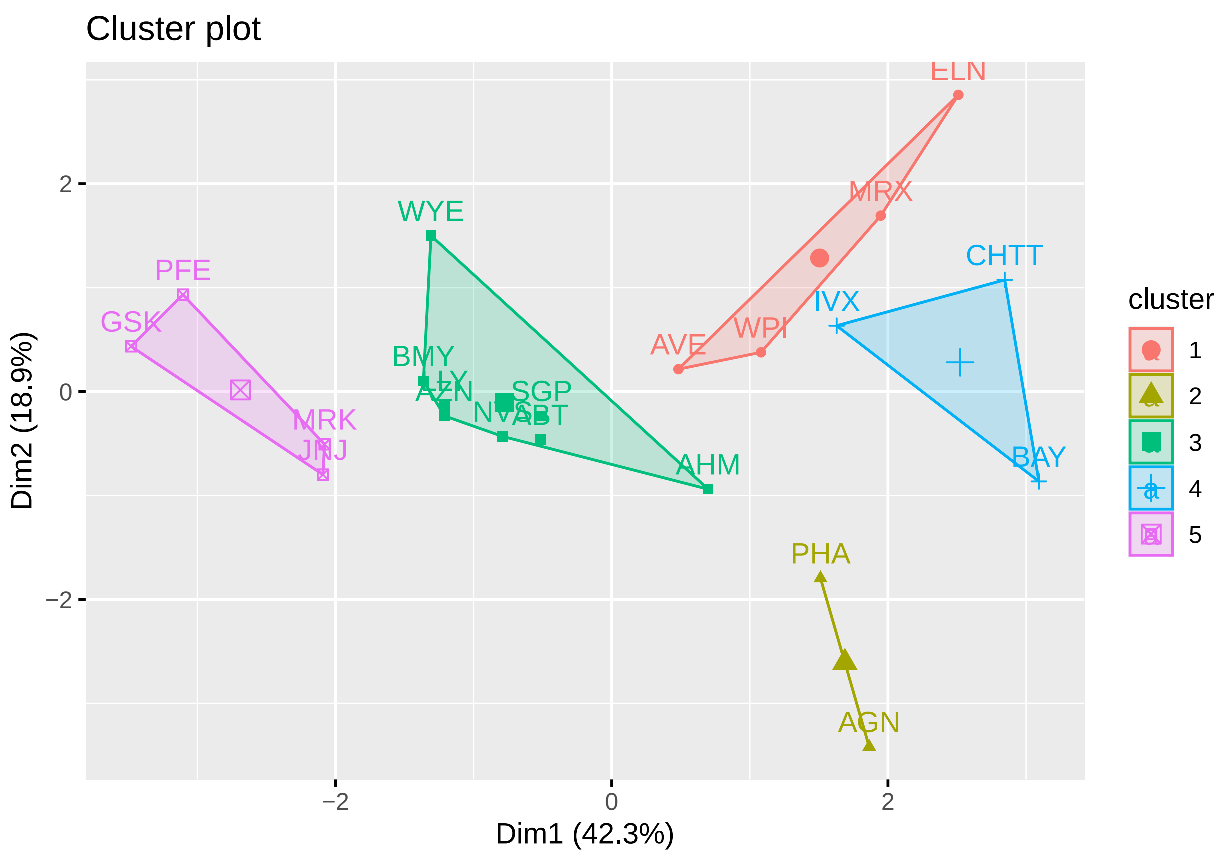
```
set.seed(1)
k5 <- kmeans(Infoupdated, centers = 5, nstart = 25) # k = 5, number of restarts = 25
k5$centers
```

```
##    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 = Infoupdated)
```




### kmeans clustering, using Manhattan Distance

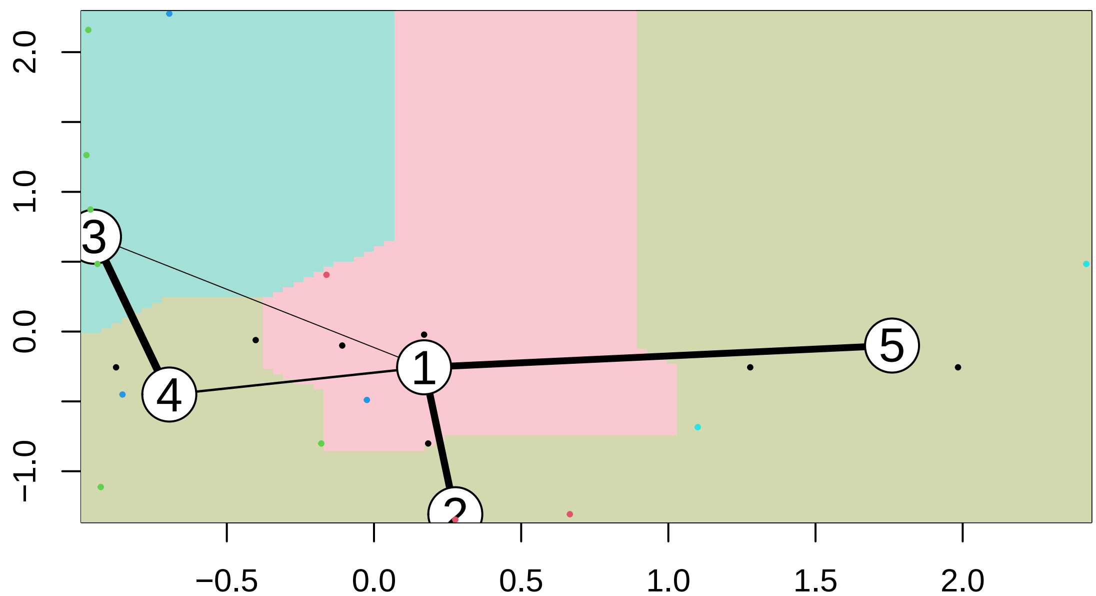
```r
set.seed(1)
k51 = kcca(Infoupdated, k=5, kccaFamily("kmedians"))
k51
```

```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = Infoupdated, k = 5, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
## 1 2 3 4 5
## 7 3 6 3 2
```

```r
#Using predict function.
clusters_index <- predict(k51)
dist(k51@centers)
```

```
##          1        2        3        4
## 2 2.150651
## 3 3.513242 4.146567
## 4 3.878726 4.246051 3.388339
## 5 3.018500 3.737739 5.124420 6.043691
```

```
image(k51)
points(Infoupdated, col=clusters_index, pch=19, cex=0.3)
```

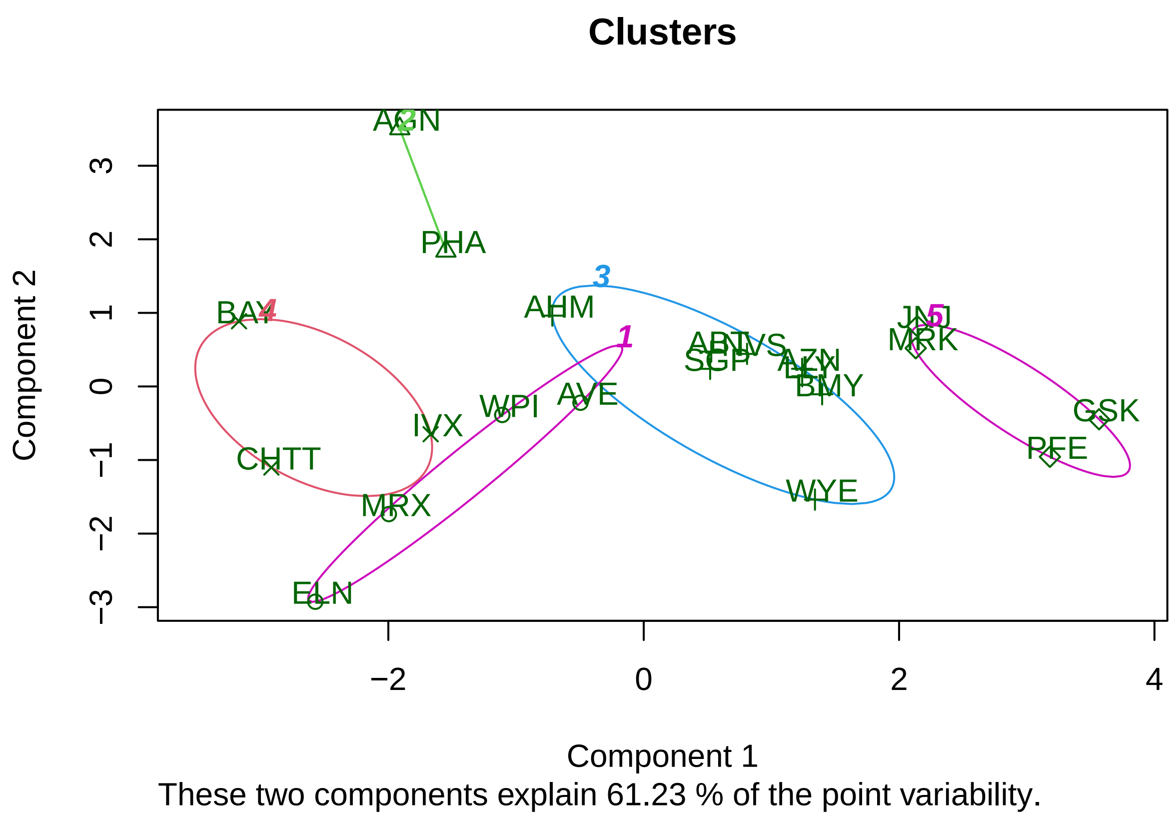



b.Interpret the clusters with respect to the numerical variables used in forming the clusters Using Kmeans method to calculate Mean.

```
Info1 %>% mutate(Cluster = k5$cluster) %>% group_by(Cluster) %>% summarise_all("mean")
```

```
## # A tibble: 5 x 10
##   Cluster Market_Cap  Beta PE_Ratio   ROE   ROA Asset_~1 Lever~2 Rev_G~3 Net_P~4
##     <int>      <dbl> <dbl>    <dbl> <dbl> <dbl>    <dbl>   <dbl>   <dbl>   <dbl>
## 1       1      13.1  0.598     17.7  14.6  6.2     0.425   0.635   30.1    15.6
## 2       2      31.9  0.405     69.5  13.2  5.6     0.75    0.475   12.1     6.4
## 3       3      55.8  0.414     20.3  28.7 12.7     0.738   0.371    5.59   19.4
## 4       4       6.64 0.87      24.6  16.5  4.17    0.6     1.65     5.73    7.03
## 5       5     157.   0.48      22.2  44.4 17.7     0.95    0.22    18.5    19.6
## # ... with abbreviated variable names 1: Asset_Turnover, 2: Leverage,
## #   3: Rev_Growth, 4: Net_Profit_Margin
```

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




Companies are categorized into different clusters as follows:

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

From the means of the cluster variables, it can be derived as follow:

Cluster 1 has the fastest revenue growth, the highest Net Profit Margin, and the lowest PE ratio. It can be purchased or held in reserve..

Cluster 2 PE ratio is very high

Cluster 3 has average risk

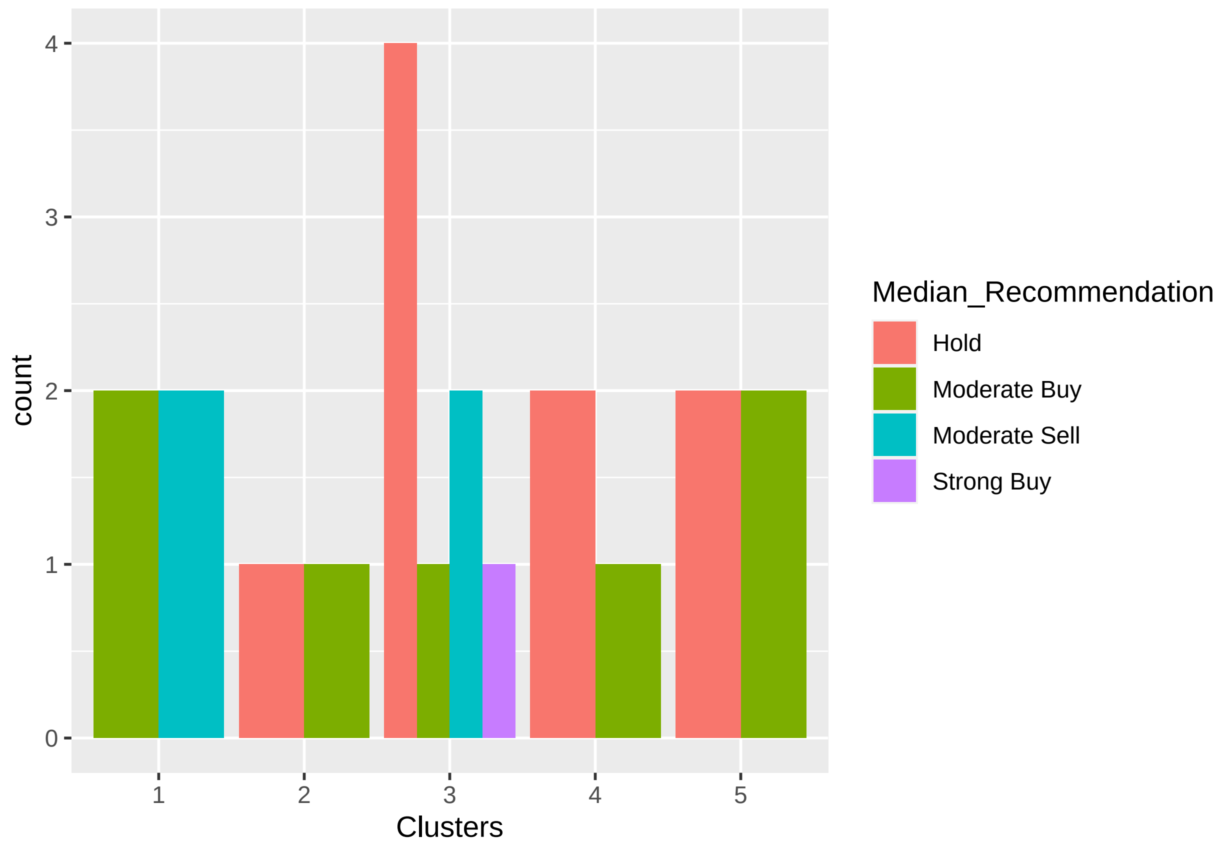
Cluster 4 Though it has a good PE ratio, it carries a very high risk , very very high leverage and low Net Profit margin , making it very risky to own. Revenue growth is also very low.

Cluster 5 has a high market capitalization, return on investment, return on assets, asset turnover, and net profit margin. With a low PE ratio, the stock price is moderately valued and hence can be purchased and held evenue growth of 18.5% is good.

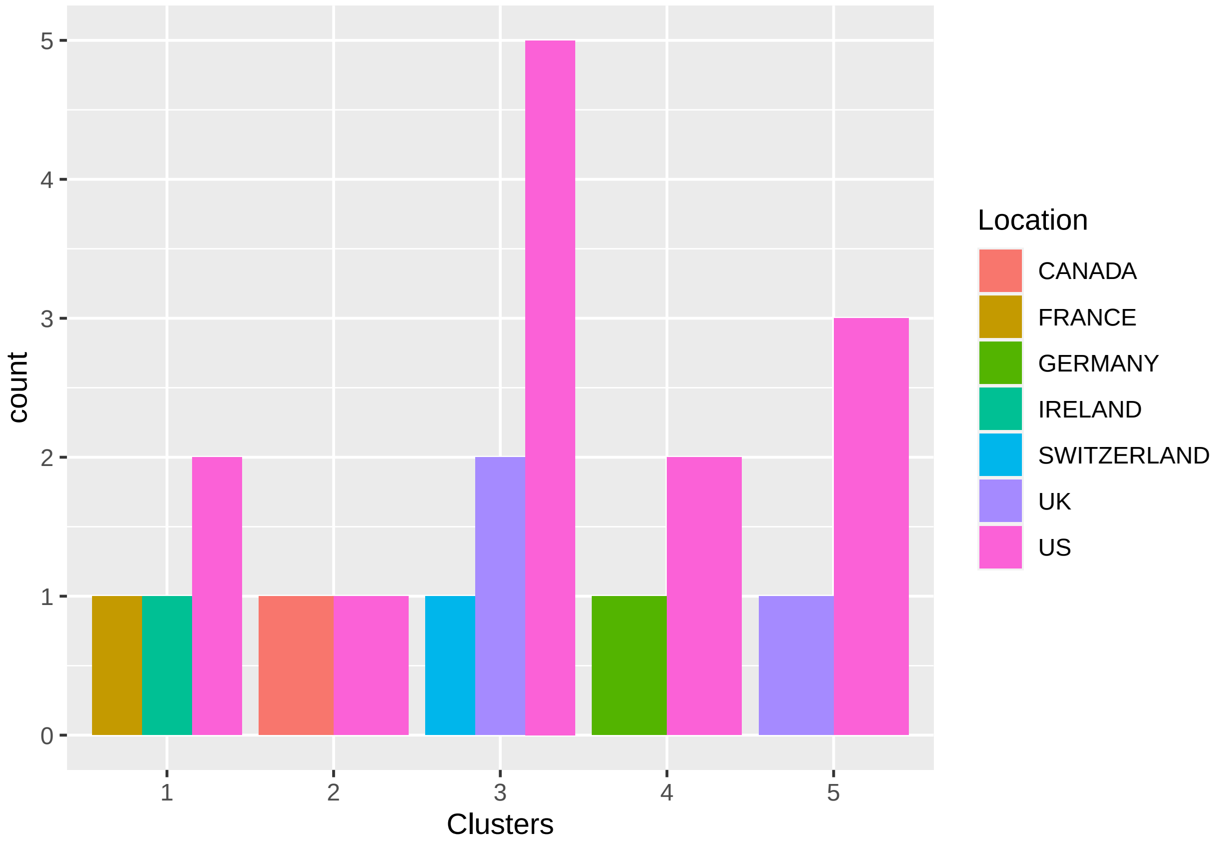
c.Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used informing the clusters)

### plotting clusters against the variables to check for any patterns

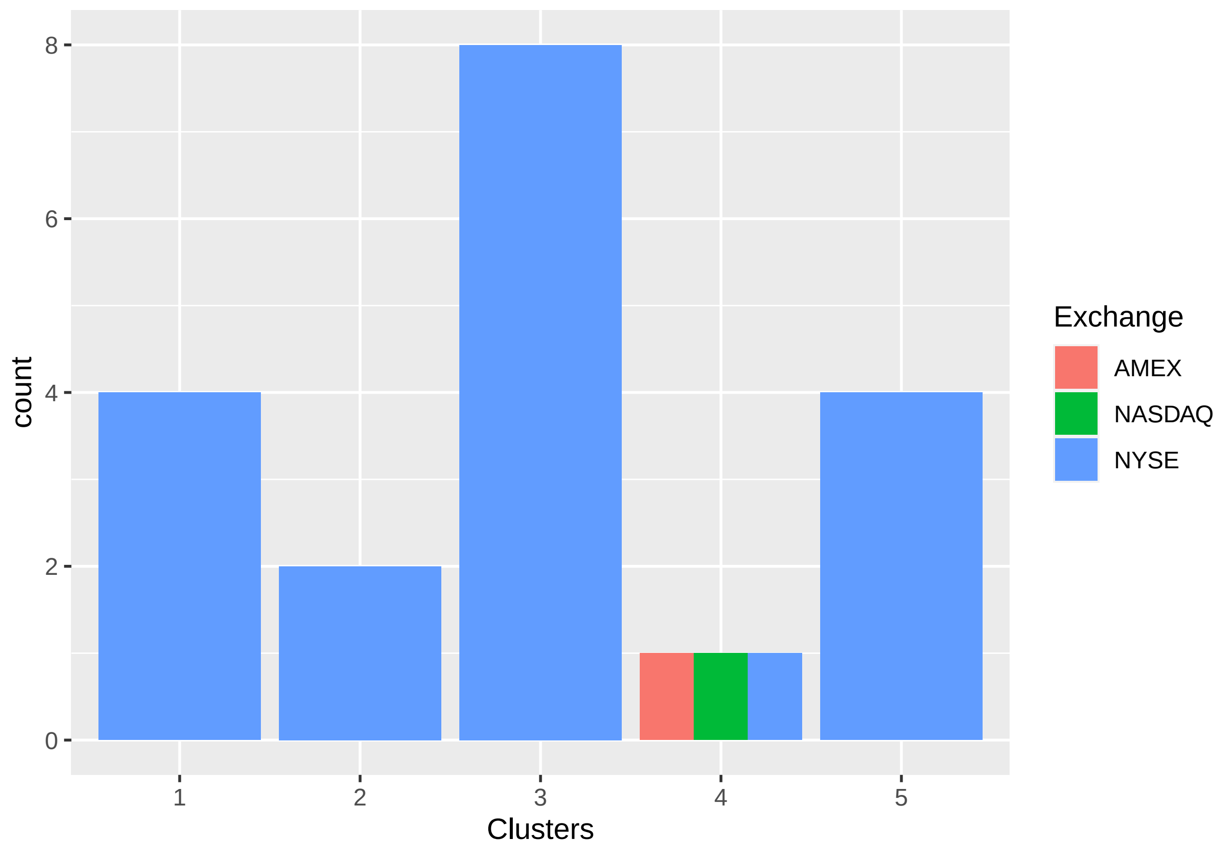
```
Info3 <- Info[12:14] %>% mutate(Clusters=k5$cluster)
ggplot(Info3, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position='dodge')+l
```




```
ggplot(Info3, mapping = aes(factor(Clusters),fill = Location))+geom_bar(position = 'dodge')+labs(x ='Clu
```




```
ggplot(Info3, mapping = aes(factor(Clusters),fill = Exchange))+geom_bar(position = 'dodge')+labs(x ='Clu
```




**Clusters and the variable Median Recommendation appear to follow a pattern.**

**Except for the fact that the bulk of the clusters/companies are listed on the NYSE and are based in the United States, there appears to be no discernible pattern among the clusters, locations, or exchanges.**

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

Cluster 1: Best Buying Cluster 2: Highly Risky Cluster 3: Go for it Cluster 4: Very Risky or Runaway Cluster 5: Ideal to Own