# Assingment 4 MACHINE LEARNING

```
library(readr)

library(caret)
```

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

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

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

```
library(cluster)

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(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.1.3
```

```
##
## Attaching package: 'gridExtra'
```

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

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.1
## v purrr   0.3.4
```

```
## Warning: package 'forcats' was built under R version 4.1.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x gridExtra::combine() masks dplyr::combine()
## x dplyr::filter()      masks stats::filter()
## x dplyr::lag()         masks stats::lag()
## x purrr::lift()        masks caret::lift()
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.1.3
```

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

```
setwd("C:/Users/nikes/Downloads/Machine Learning Assingment/Assingment 4")

Pharmaceuticals <- read_csv("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(Pharmaceuticals)
```

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

```
#'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.'
```

```
## [1] "Justify the various choices made in conducting the cluster analysis, such as \nweights for diffe
```

```
Pharmaceuticals <- read_csv("Pharmaceuticals.csv",col_types = cols(Market_Cap
= col_number(),Beta = col_number(), PE_Ratio = col_number(),ROE = col_number(),
ROA = col_number(), Asset_Turnover = col_number(), Leverage = col_number(),
Rev_Growth = col_number(), Net_Profit_Margin = col_number()))

PM_df <- data.frame(Pharmaceuticals[,3:10]) %>% na.omit(PM)

PM_Scale<- scale(PM_df)

PM_Distance<- get_dist(PM_df)
```

```
kmeans_model <- kmeans(PM_Scale, centers = 3, nstart = 20)
str(kmeans_model)
```

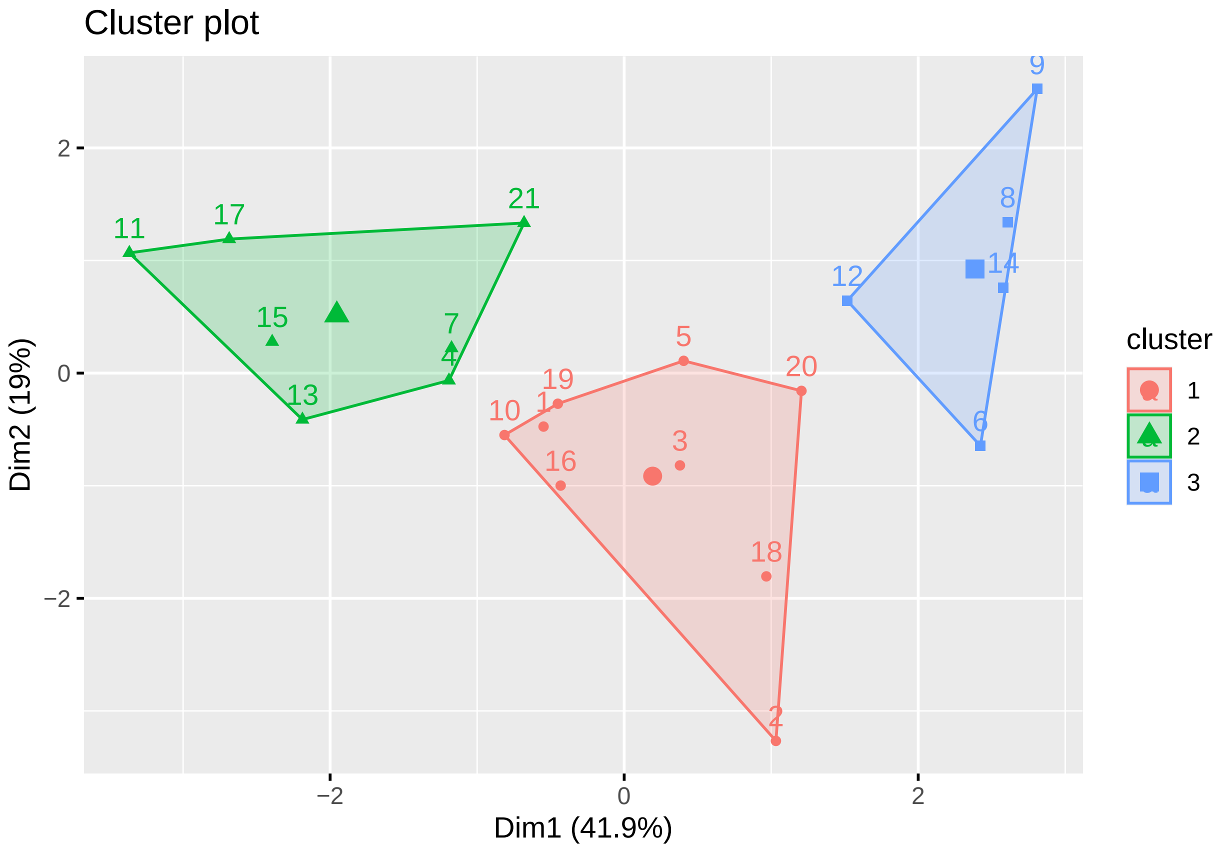
```
## List of 9
##  $ cluster     : Named int [1:21] 1 1 1 2 1 3 2 3 3 1 ...
##   ..- attr(*, "names")= chr [1:21] "1" "2" "3" "4" ...
##  $ centers     : num [1:3, 1:8] -0.2376 0.9548 -0.9091 -0.7363 -0.0612 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:3] "1" "2" "3"
##   .. ..$ : chr [1:8] "Market_Cap" "Beta" "PE_Ratio" "ROE" ...
##  $ totss       : num 160
##  $ withinss    : num [1:3] 35.3 22.9 27.4
##  $ tot.withinss: num 85.6
##  $ betweenss   : num 74.4
##  $ size        : int [1:3] 9 7 5
##  $ iter        : int 2
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
kmeans_model
```

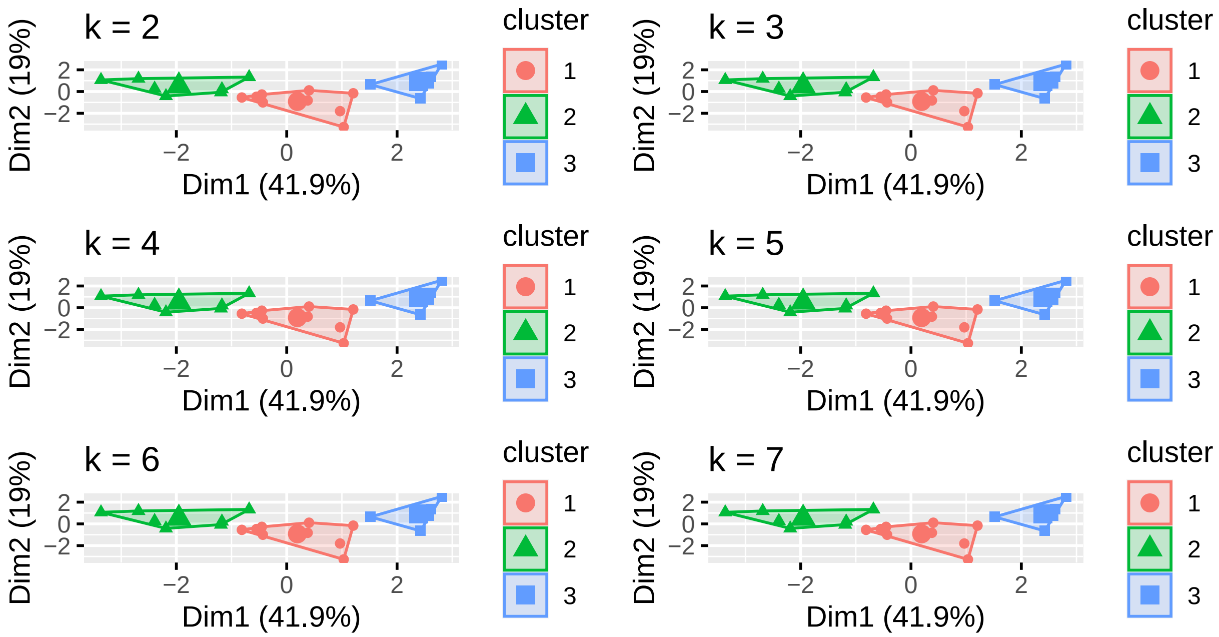
```
## K-means clustering with 3 clusters of sizes 9, 7, 5
```

```
##
## Cluster means:
##   Market_Cap        Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.2375550 -0.73633718  0.4233386 -0.4489909 -0.2407172     -0.1025035
## 2  0.9547543 -0.06120687 -0.3576482  1.0818081  1.1033619      0.8566361
## 3 -0.9090570  1.41109654 -0.2613021 -0.7063477 -1.1114156     -1.0147843
##     Leverage  Rev_Growth
## 1 -0.3557313 -0.13595383
## 2 -0.2797499 -0.01818848
## 3  1.0319661  0.27018076
##
## Clustering vector:
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
##  1  1  1  2  1  3  2  3  3  1  2  3  2  3  2  1  2  1  1  1  2
##
## Within cluster sum of squares by cluster:
## [1] 35.27579 22.91263 27.43310
##  (between_SS / total_SS =  46.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

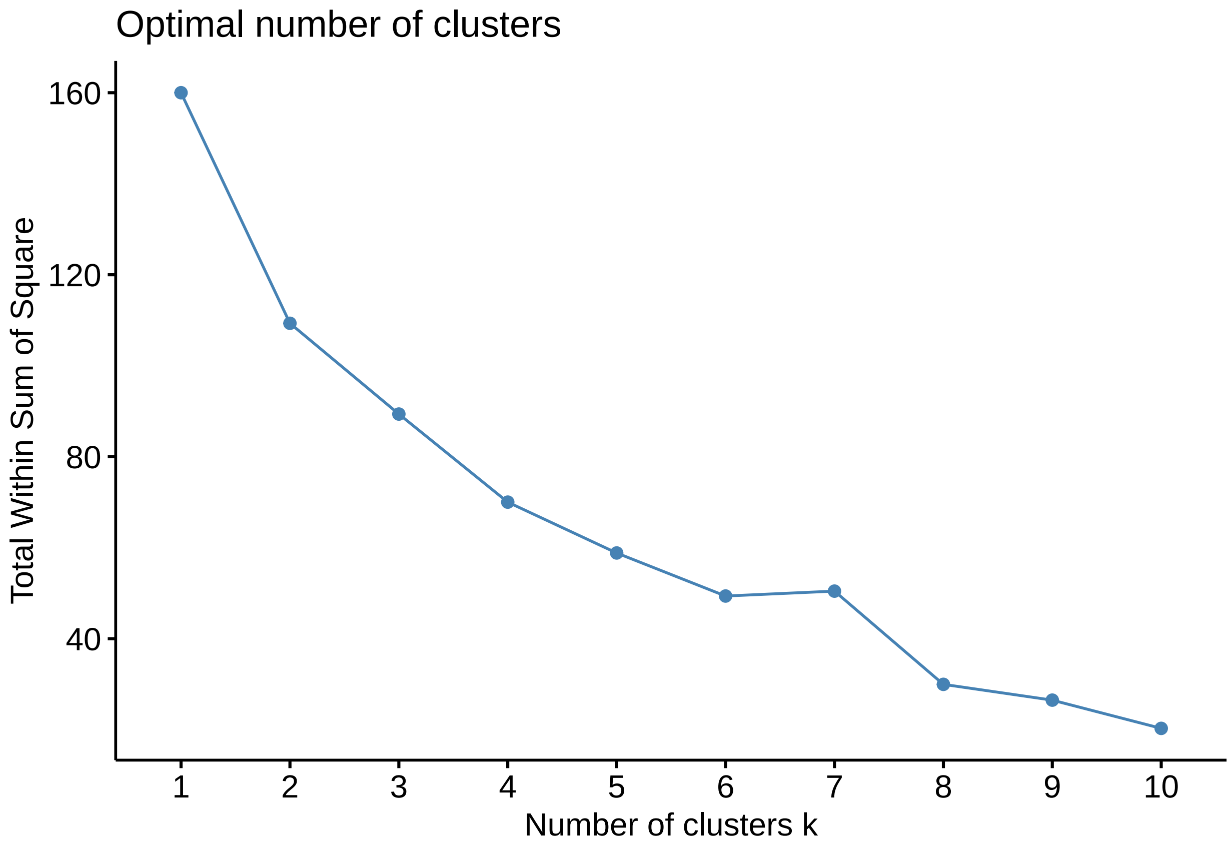
```
fviz_cluster(kmeans_model, data = PM_df)
```



```
# Defining optimal Value for K (Hyperparameter)

PM1 <- fviz_cluster(kmeans_model,  geom = "point",  data = PM_Scale) +ggtitle("k = 2")
PM2 <- fviz_cluster(kmeans_model, geom = "point",  data = PM_Scale) + ggtitle("k = 3")
PM3 <- fviz_cluster(kmeans_model, geom = "point",  data = PM_Scale) + ggtitle("k = 4")
PM4 <- fviz_cluster(kmeans_model, geom = "point",  data = PM_Scale) + ggtitle("k = 5")
PM5 <- fviz_cluster(kmeans_model, geom = "point",  data = PM_Scale) + ggtitle("k = 6")
PM6 <- fviz_cluster(kmeans_model, geom = "point",  data = PM_Scale) + ggtitle("k = 7")

grid.arrange(PM1, PM2, PM3, PM4, PM5, PM6, nrow = 4)
```

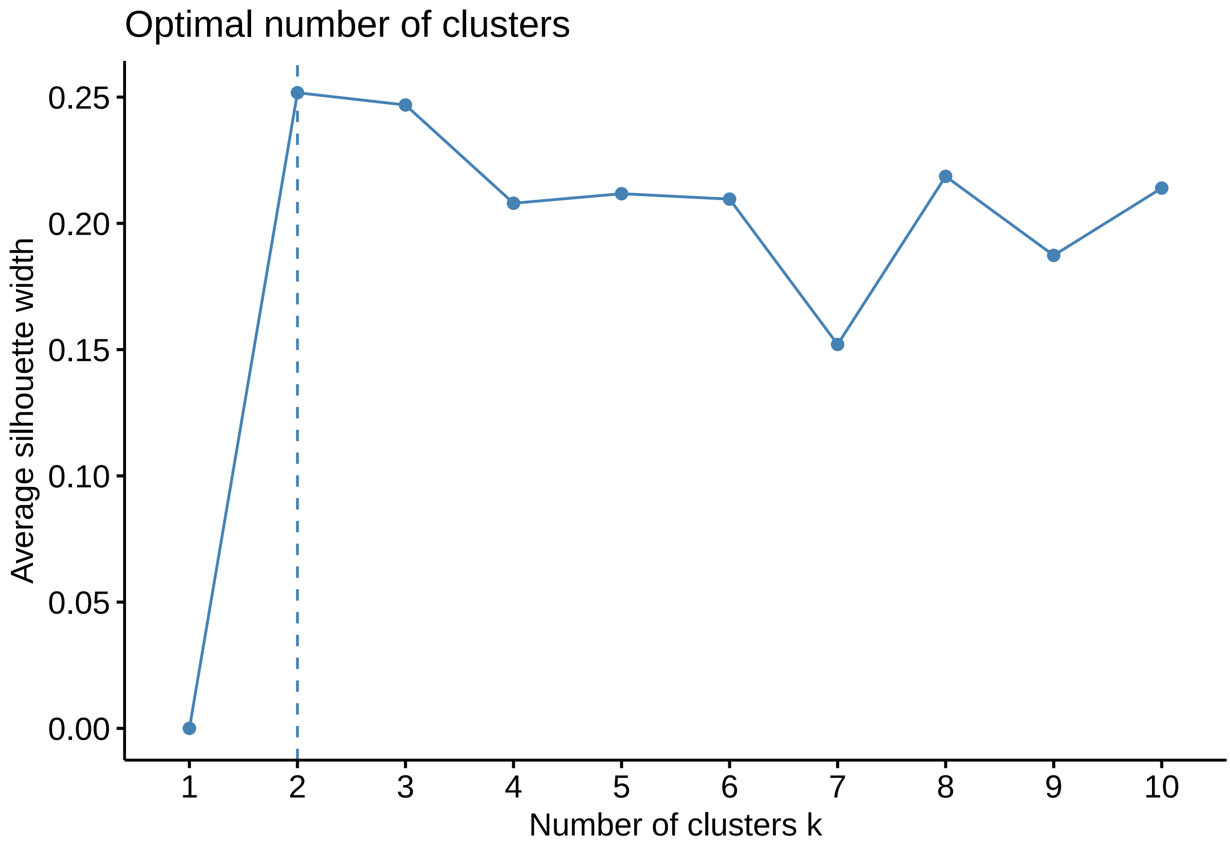


```
#Specifying K number of cluster through;

#Elbow method

set.seed(100)

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



#The Elbow Method calculates the Within-Cluster-Sum of Squared Errors (WSS) for various k values and selects the k for which WSS begins to diminish first. This is seen as an elbow in the WSS-versus-k figure.Within-Cluster-Sum of Squared Errors appears to be a complicated formula.We can choose K as 5 or 6 from the Elbow method provided by the figure.

```
#Silhouette method

set.seed(100)

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

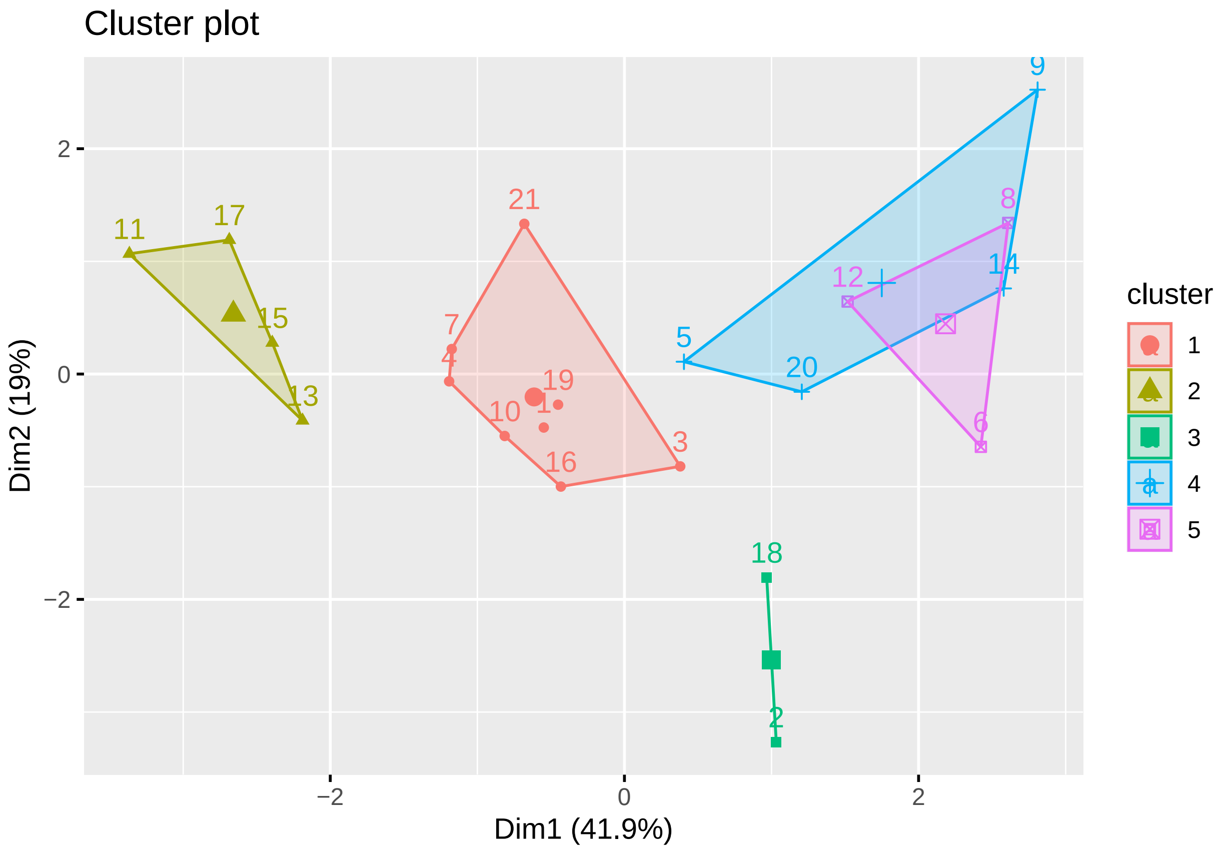


When compared to the Elbow approach, silhouette analysis may be utilized to investigate the separation distance between the generated clusters and can be deemed a better method. The separation distance between the generated clusters can be studied using silhouette analysis. The silhouette plot shows how close each point in one cluster is to points in neighboring clusters, and so allows you to visually examine factors like cluster count. here, it shows optimal number cluster as 2.

```
#Choosing final results for K

set.seed(100)

final_model <- kmeans(PM_Scale, 5, nstart = 20)

print(final_model)
```

```
## K-means clustering with 5 clusters of sizes 8, 4, 2, 4, 3
##
## 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  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 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
## 1 -0.27449312 -0.7041516
## 2 -0.46807818  0.4671788
## 3 -0.14170336 -0.1168459
## 4  0.06308085  1.5180158
## 5  1.36644699 -0.6912914
##
## Clustering vector:
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
##  1  3  1  1  4  5  1  5  4  1  2  5  2  4  2  1  2  3  1  4  1
##
## Within cluster sum of squares by cluster:
## [1] 18.466446  7.734337  2.765883 11.739016 14.769026
##  (between_SS / total_SS =  65.3 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

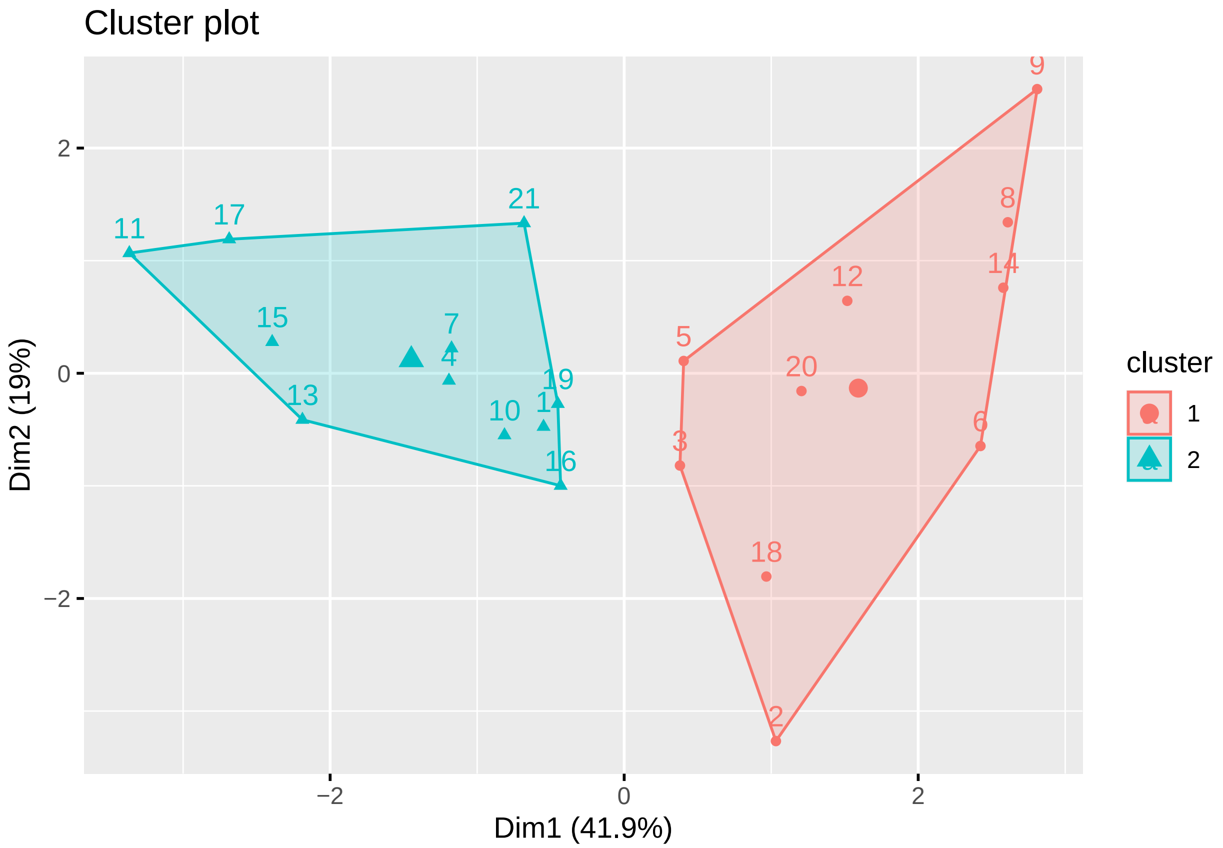
```
fviz_cluster(final_model, data = PM_Scale)
```



```
#Silhouette method Clusters

set.seed(100)

final_model2 <- kmeans(PM_Scale, 2, nstart = 20)

print(final_model2)
```

```
## K-means clustering with 2 clusters of sizes 10, 11
##
## Cluster means:
##   Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.7407208  0.3945061  0.3039863 -0.7222576 -0.9178575     -0.5073922
## 2  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656
##     Leverage Rev_Growth
## 1  0.3664175  0.3192379
## 2 -0.3331068 -0.2902163
##
## Clustering vector:
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
##  2  1  1  2  1  1  2  1  1  2  2  1  2  1  2  2  2  1  2  1  2
##
## Within cluster sum of squares by cluster:
## [1] 69.26750 40.05664
##  (between_SS / total_SS =  31.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
fviz_cluster(final_model2, data = PM_Scale)
```



B. Interpret the clusters with respect to the numerical variables used in forming the clusters.

Elbow method is an empirical method for determining the best k value. It selects a set of values and selects the best among them. It calculates the average distance and the sum of the squares of the spots. K-means clustering attempts to arrange items that are similar in nature into clusters.

It compares the objects and divides them into clusters based on their similarity. The “elbow” (the point of inflection on the curve) is the best value of k if the line chart resembles an arm. The “arm” can move up or down, but if there is a significant inflection point, its a positive sign that the underlying model fits well at that point, which in this case is toward the bottom i.e. 5.

#K-means clustering with 5 clusters of sizes 8, 4, 2, 4, 3

Silhouette method values are little different. The silhouette graph is drawn using the elbow approach, which additionally picks up the range of k values. Every points silhouette coefficient is calculated. It computes the average distance between points in its cluster a I and the average distance between points in its next closest cluster b. (i). The clusters are quite dense and well divided with a silhouette score of 1. Clusters that have a score of 0 are overlapping. A score of less than 0 indicates that data in clusters may be erroneous or incorrect. The silhouette plots can be used to determine the best K value. We have see 2 clusters from the graph in this method.

#K-means clustering with 2 clusters of sizes 10, 11

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

There seems to be the pattern regarding numeric variables medain_recommendation ,location and exchange as we see both are assingned to different clusters through elbow and silhouette method. Clusters are created by linking data points based on their distance from one another. Varying clusters form at different distances and can be depicted using a dendrogram, which explains why theyre also known as “hierarchical clustering.”Clustering approaches simply attempt to arrange comparable patterns into clusters whose members are more similar to one another (as measured by some distance metric) than members of other clusters. There is no way of knowing ahead of time which patterns belong to which groups, or even how many groups are acceptable.

’D. Provide an appropriate name for each cluster using any or all of the variables in the dataset.

Elbow Method

#-Cluster 1. Strong buy and hold cluster.

#-Cluster 2. Medium buy and strong hold cluster

#-Cluster 3. Weak buy cluster

#-Cluster 4. Medium buy cluster

#-Cluster 5. Weak buy Cluster.

Silhouette Method

#-Cluster 1. Medium buy and hold cluster.

#-Cluster 2. Strong buy and strong hold cluster