Assignment_4

Bharadwaj

Loading the required libraries and data set

```r
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
```

```
pharmaceutical_0data<-read.csv("C:/Users/CherRyY/Documents/R/dataset/Pharmaceuticals.csv")
pharmaceutical_data<-na.omit(pharmaceutical_0data)
```

Here we'll be using the numerical variables (1 to 9) to cluster the 21 firms.

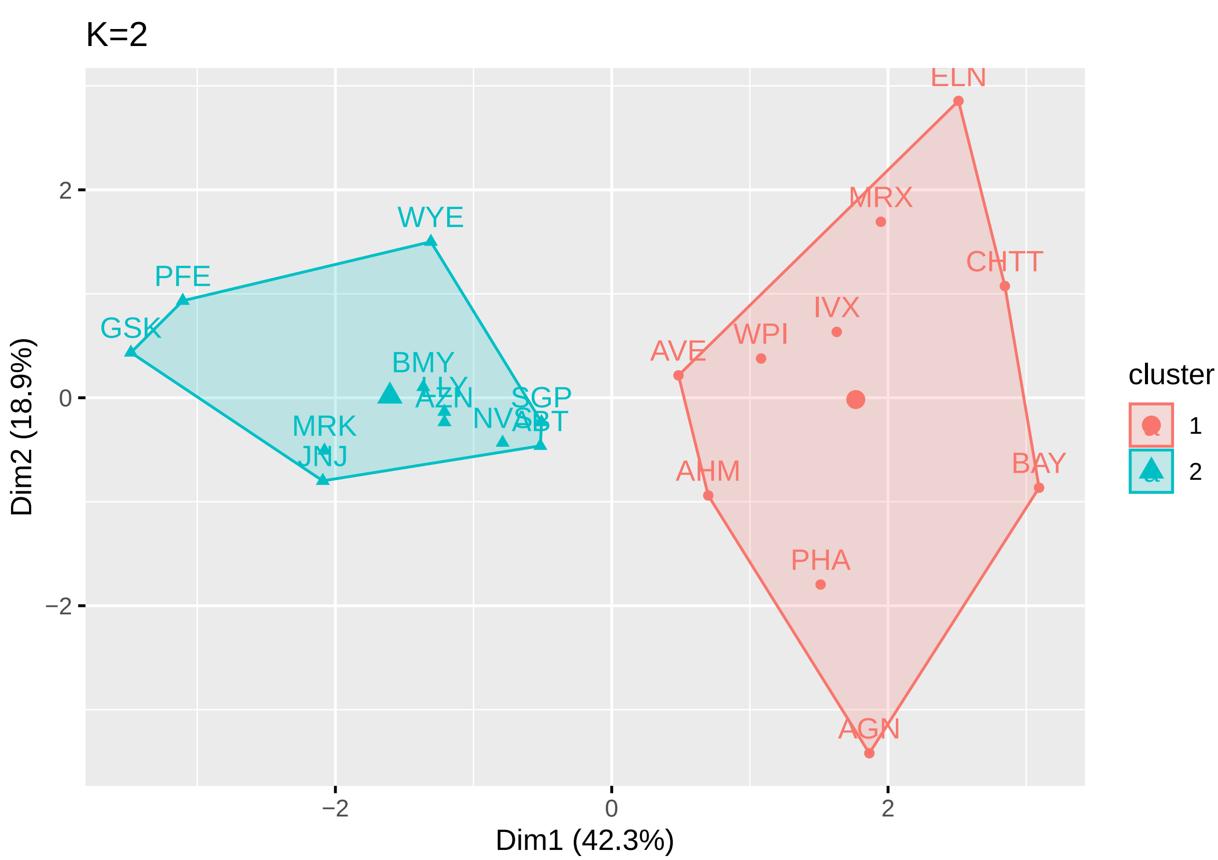
```
row.names(pharmaceutical_0data)<-pharmaceutical_0data[,1]
Clustering_data<-pharmaceutical_0data[,3:11]
```

Scaling the data according to requirement

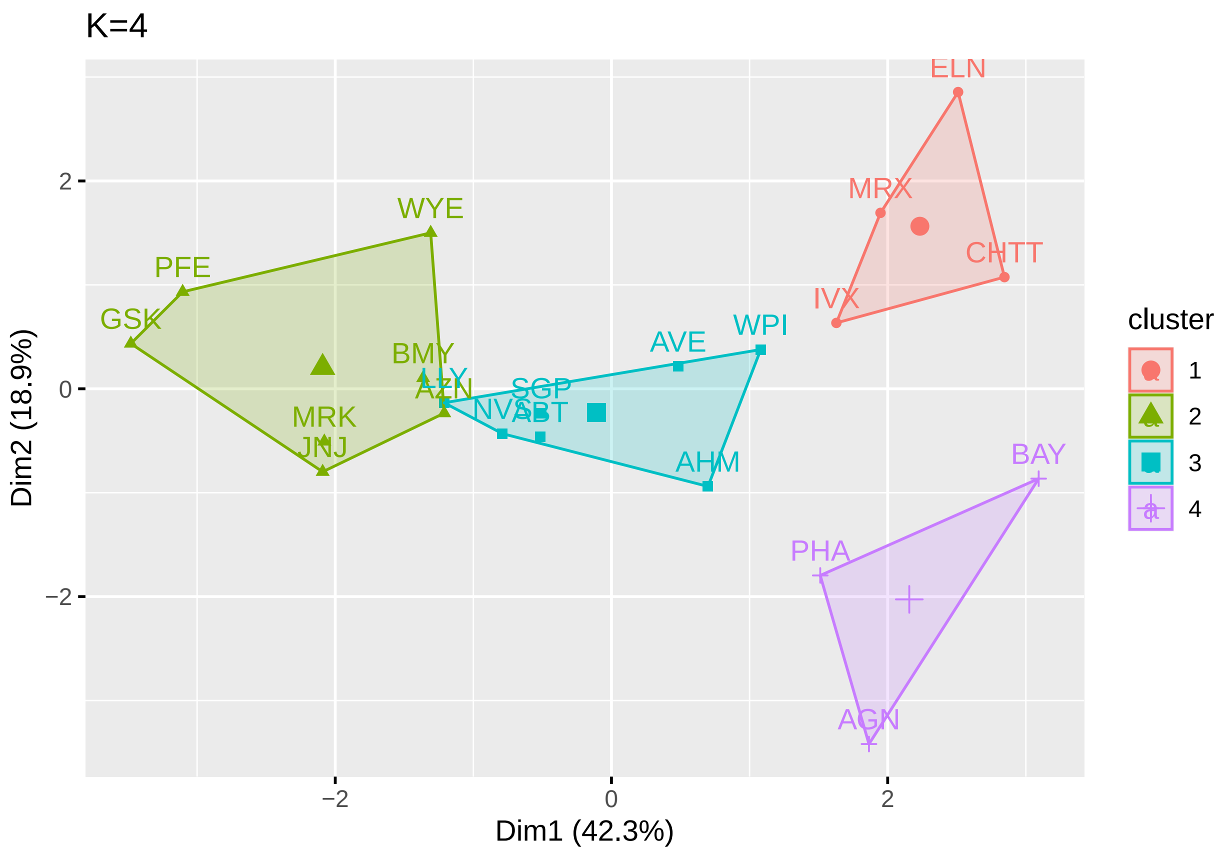
```
set.seed(120)
ScaledData<-scale(Clustering_data)
```

Performing Kmeans clustering for random K values(trail and error)

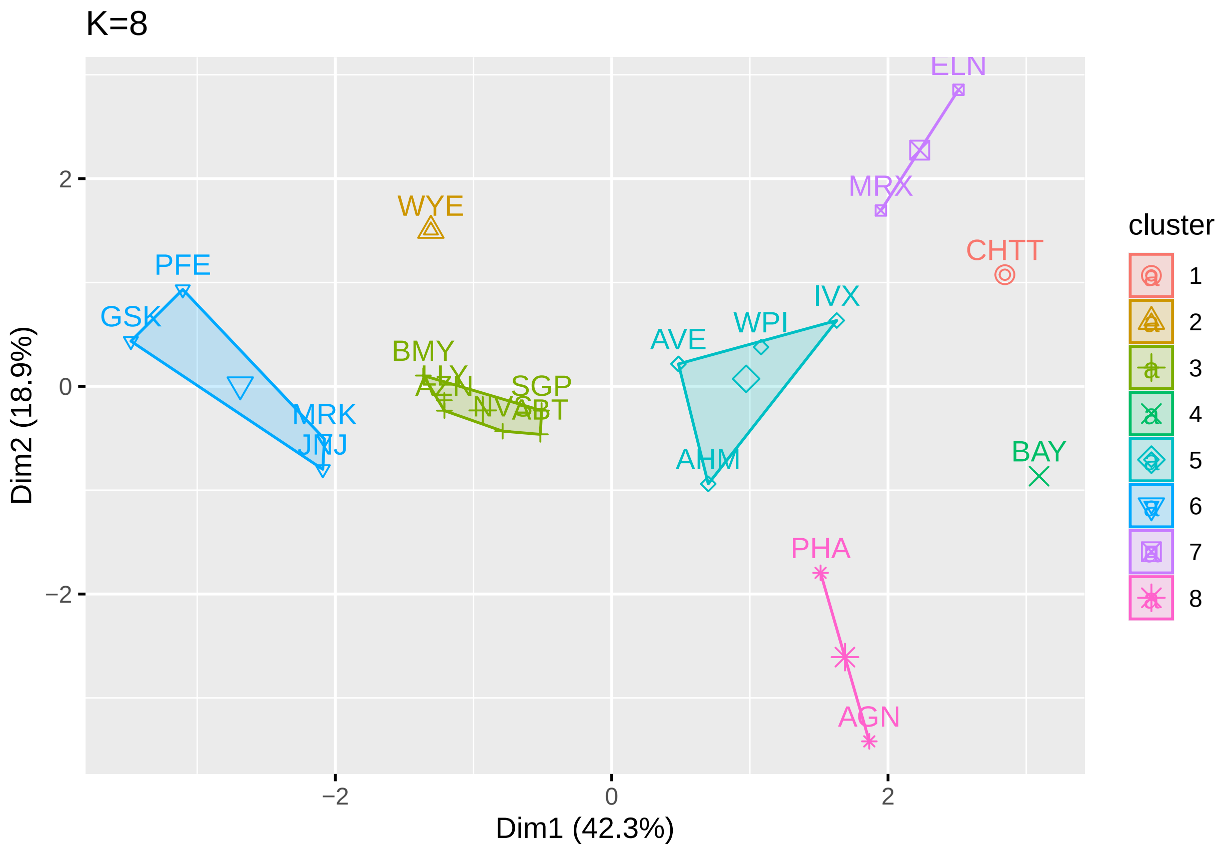
```
set.seed(143)
kmeans_2<-kmeans(ScaledData,centers = 2, nstart = 15)
kmeans_4<-kmeans(ScaledData,centers = 4, nstart = 15)
kmeans_8<-kmeans(ScaledData,centers = 8, nstart = 15)

plot_kmeans_2<-fviz_cluster(kmeans_2,data = ScaledData) + ggtitle("K=2")
plot_kmeans_4<-fviz_cluster(kmeans_4,data = ScaledData) + ggtitle("K=4")
plot_kmeans_8<-fviz_cluster(kmeans_8,data = ScaledData) + ggtitle("K=8")

plot_kmeans_2
```



```
plot_kmeans_4
```

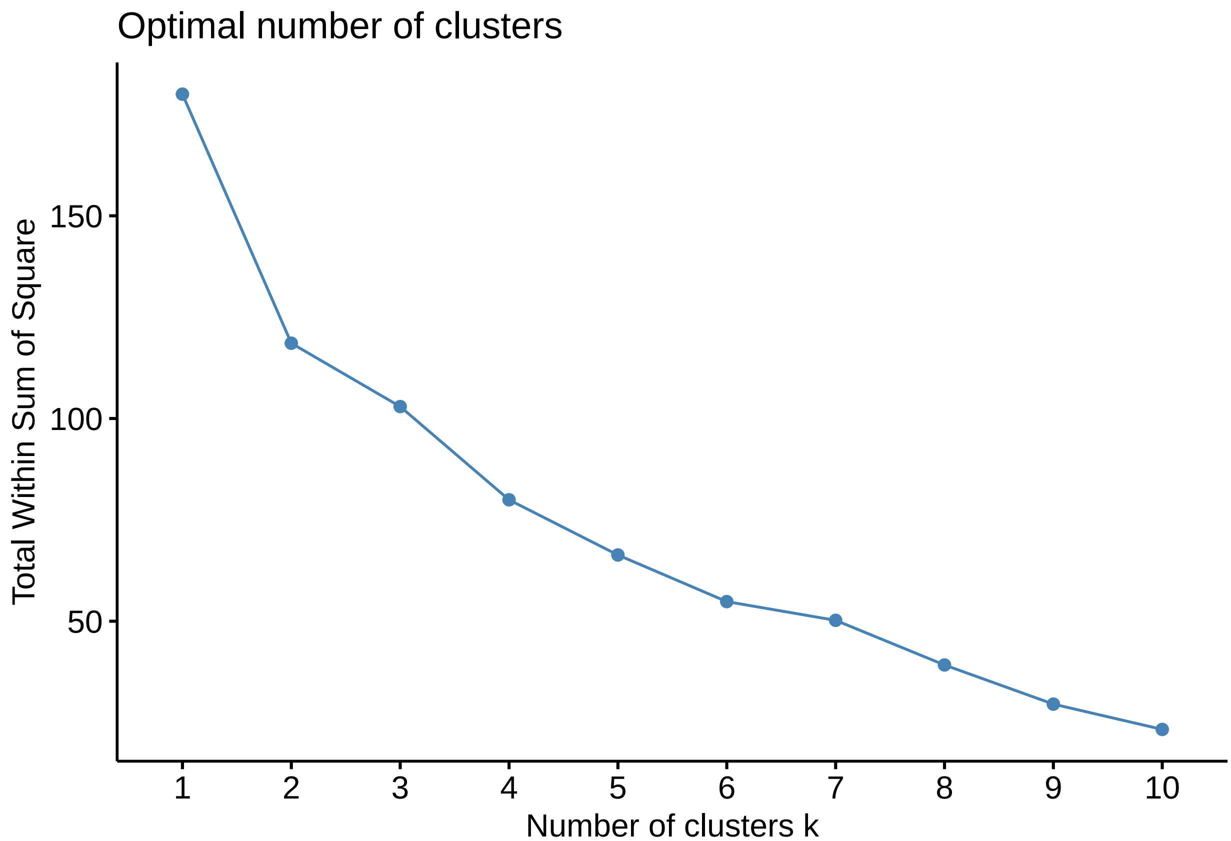



```
plot_kmeans_8
```

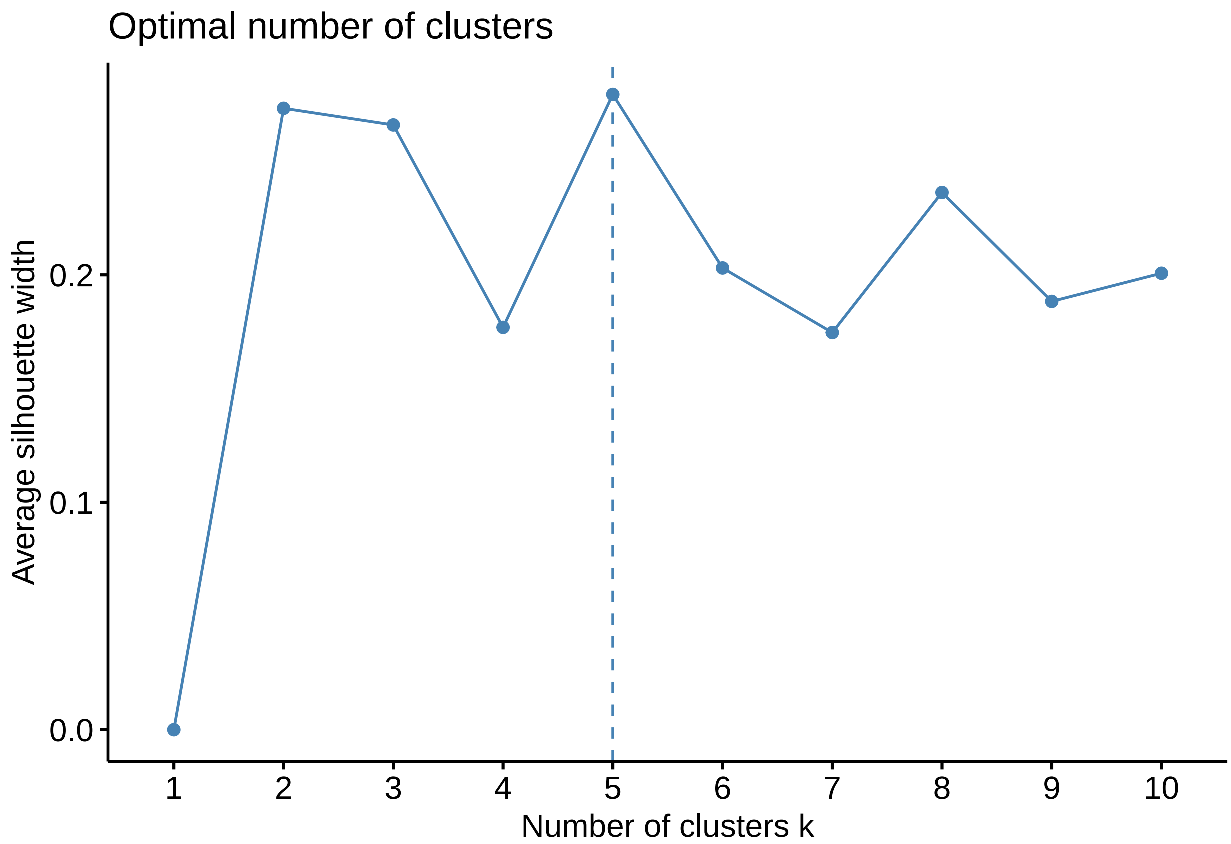


to find best K suitable for clustering, we'll use the WSS and silhouette method

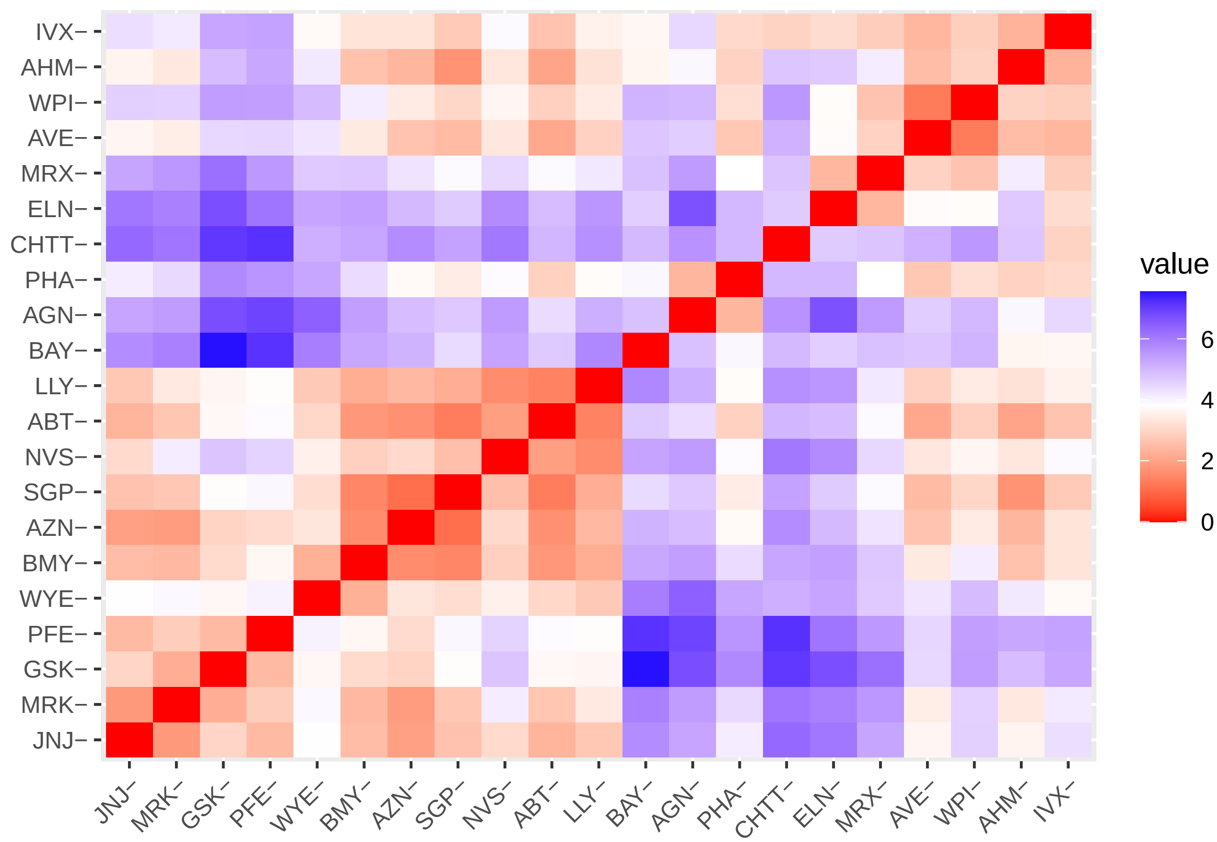
```
k_wss<-fviz_nbclust(ScaledData,kmeans,method="wss")
k_silhouette<-fviz_nbclust(ScaledData,kmeans,method="silhouette")
k_wss
```




```
k_silhouette
```




```
distance<-dist(ScaledData,metho='euclidean')
fviz_dist(distance)
```



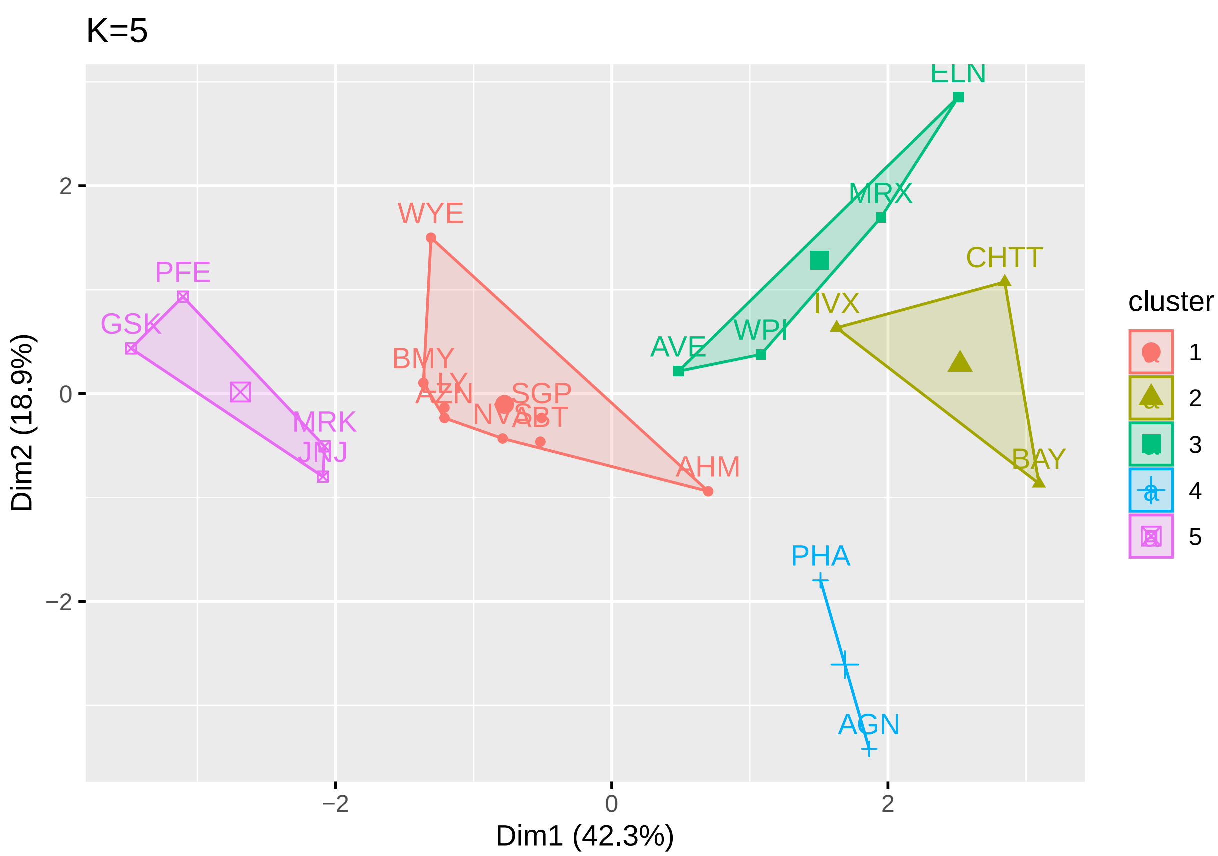
Given a K value of 2 from the within-cluster sum of squares (wss) and a K value of 5 from the silhouette analysis, the choice is to opt for K equals 5. This decision is motivated by the aim to achieve a low sum of squares while ensuring effective separation within the clusters. Now, the next step is to conduct a K-means analysis to determine the optimal K for the given data.

```
set.seed(143)
kmeans_5<-kmeans(ScaledData,centers = 5, nstart = 10)
kmeans_5
```

```
## 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:
```

```
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    1    4    1    1    3    2    1    2    3    1    5    2    5    3    5    1
##  PFE  PHA  SGP  WPI  WYE
##    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"
```

```r
plot_kmeans_5<-fviz_cluster(kmeans_5,data = ScaledData) + ggtitle("K=5")
plot_kmeans_5
```



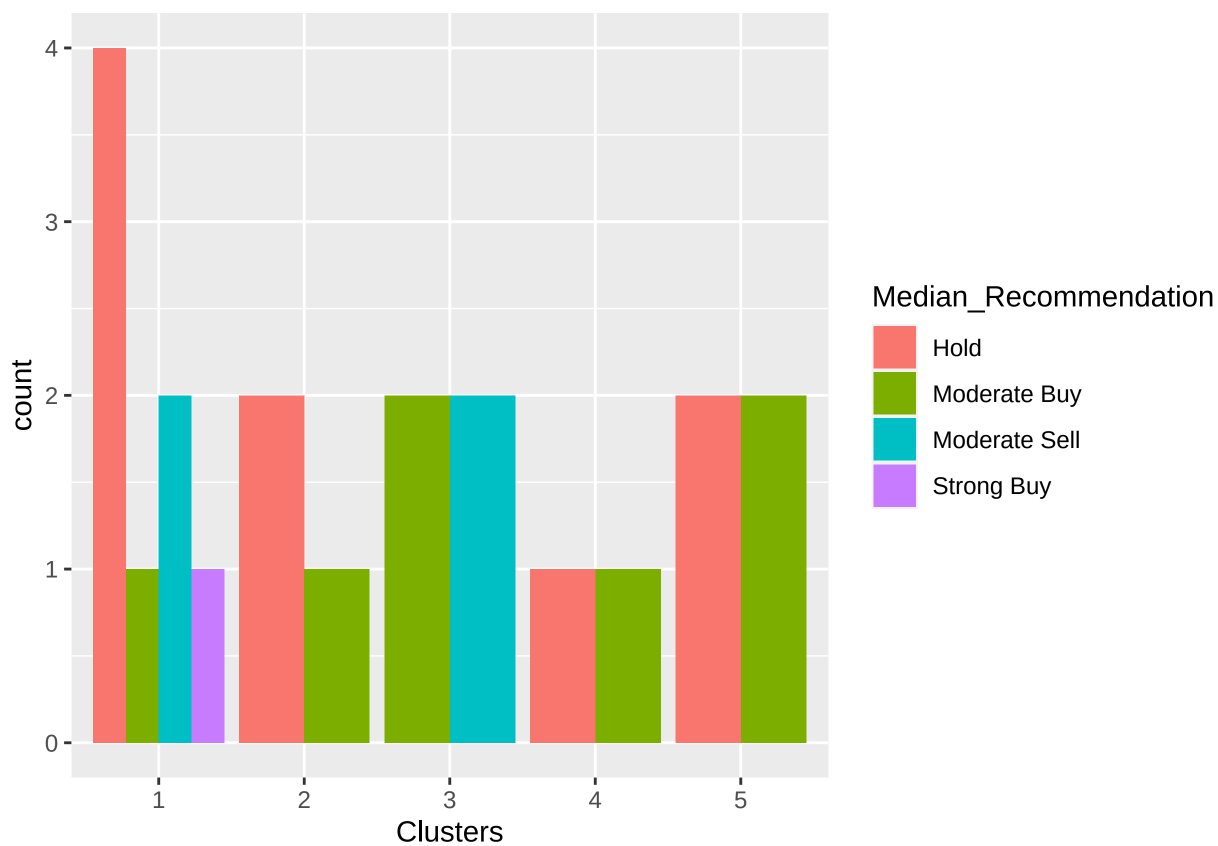

```r
Clustering_data_1<-Clustering_data%>%
  mutate(Cluster_no=kmeans_5$cluster)%>%
  group_by(Cluster_no)%>%summarise_all('mean')
Clustering_data_1
```

```
## # A tibble: 5 x 10
##    Cluster_no Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##         <int>      <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
```

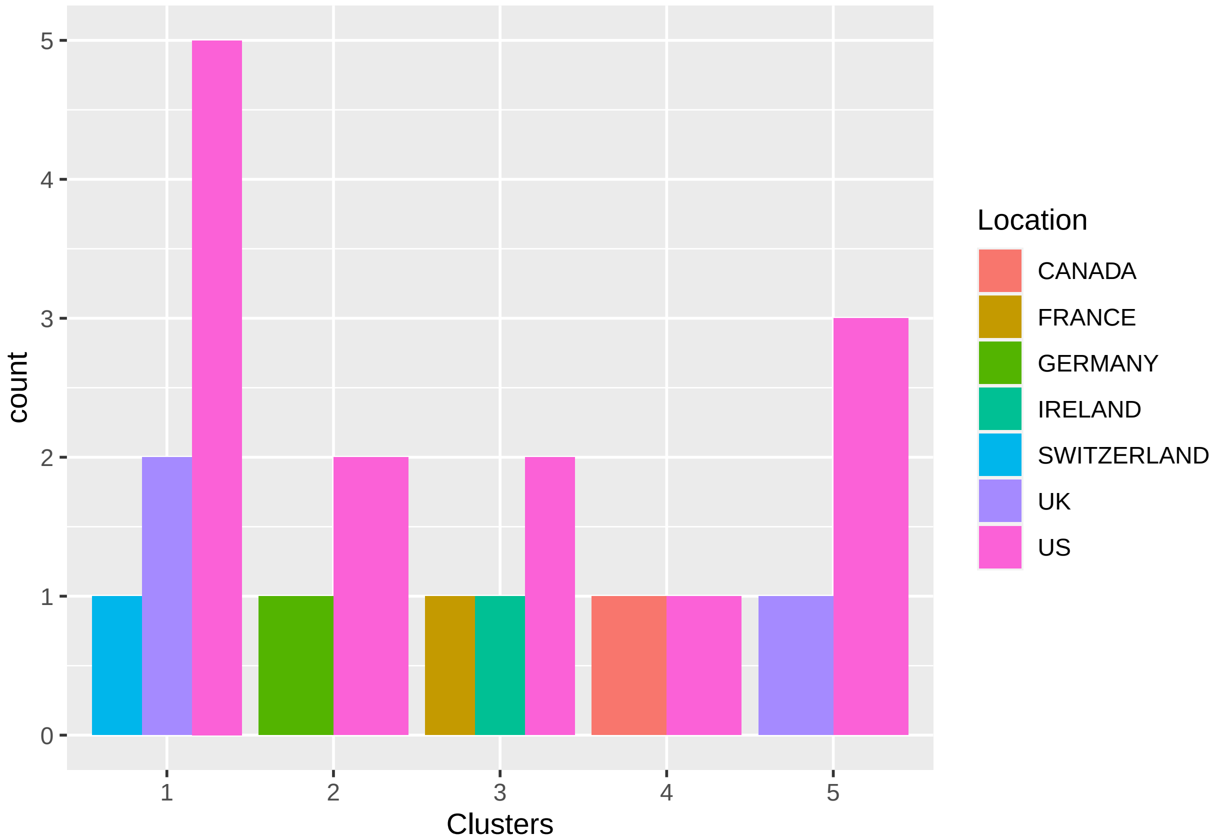
```
## 1          1      55.8  0.414     20.3  28.7 12.7          0.738    0.371
## 2          2       6.64 0.87      24.6  16.5  4.17         0.6      1.65
## 3          3      13.1  0.598     17.7  14.6  6.2          0.425    0.635
## 4          4      31.9  0.405     69.5  13.2  5.6          0.75     0.475
## 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>
```

Companies are grouped into following clusters: Cluster_1= ABT,AHM,AZN,BMY,LLY,NVS,SGP,WYE Cluster_2= BAY,CHTT,IVX Cluster_3=AVE,ELN,MRX,WPI Cluster_4=AGN,PHA Cluster_5=GSK,JNJ,MRK,PFE From the above clusters 1.Cluster_1 comprizes companies with moderate returns on Equity and Investment 2.Cluster_2 Consists of Companies with Poor returns on Assets(ROA), Return on Equity (ROE), Low market Capitalization, and weak Asset turnover. This suggests that these Companies are Highly Risky 3.Cluster_3 Includes Companies Similar to those in cluster 2 but with Slightly lower levels of risk 4.Cluster_4 Contains companies with very high price to earnings (P/E) ratios but extremely poor ROA and ROE, making them even riskier than those in cluster 2 5.Cluster_5 is made up of companies with Excellent market capitalization, ROE and ROA

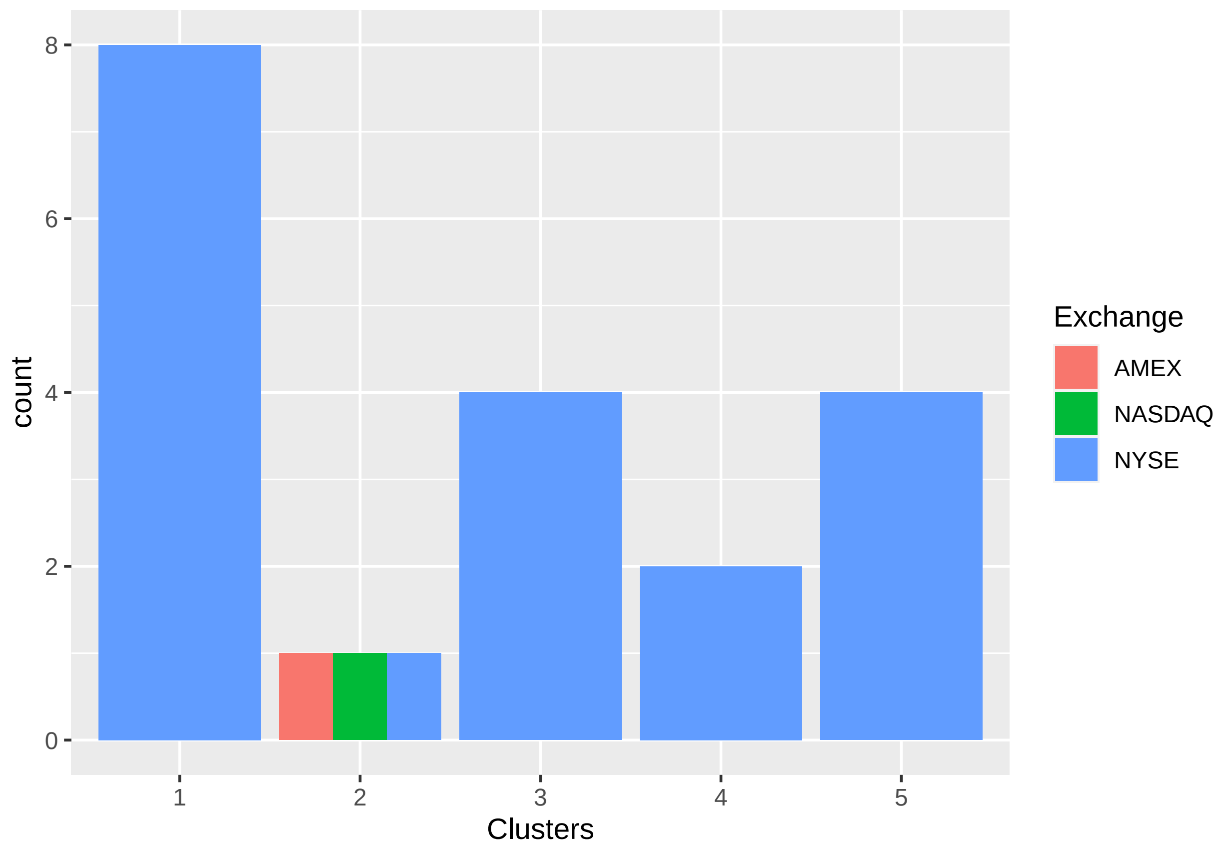
```r
Clustering_data_2<- pharmaceutical_data[,12:14] %>% mutate(Clusters=kmeans_5$cluster)
ggplot(Clustering_data_2, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(positi
```




```r
ggplot(Clustering_data_2, mapping = aes(factor(Clusters),fill = Location))+geom_bar(position = 'dodge')
```




```
ggplot(Clustering_data_2, mapping = aes(factor(Clusters),fill = Exchange))+geom_bar(position = 'dodge')+
```



Upon scrutinizing the data, a discernible pattern emerges in the relationship between clusters and the 'Median Recommendation' variable. Cluster 3 tends to incline towards ratings ranging from "moderate buy" to "moderate sell," while Cluster 2 tends to endorse recommendations within the spectrum of "hold" to "moderate buy."

A more in-depth analysis of the geographical distribution of pharmaceutical companies reveals a concentration in the United States. However, there is no clear trend in their distribution, except for the noteworthy observation that a majority of these companies are listed on the New York Stock Exchange (NYSE). Notably, there is no evident correlation between clusters and their presence in specific stock markets.

By considering both return on assets and net market capitalization, we can assign names to these clusters, thereby providing a more comprehensive labeling of the groupings.

[It is done based net Market capitalization(size) and Return on Assets(money)]

Cluster 1: High_Million

Cluster 2: Additional Tiny Penny

Cluster 3: Small Amounts of Money

Cluster 4: The Mid-Hundreds

Cluster 5: Extremely Large-Millions