Assignment 4

Sai Prasad

2024-03-31

Here I am importing requrired libraries and dataset. And removing NA (missing) values

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.3.3
```

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

```
pharma_data<-read.csv("C://Users//desineni//Downloads//Pharmaceuticals (2).csv")
pharma_data<-na.omit(pharma_data)
```

Employing the numerical variables (from 1 to 9) to group the 21 companies into clusters

```
row.names(pharma_data)<-pharma_data[,1]
clustered_data<-pharma_data[,3:11]
```

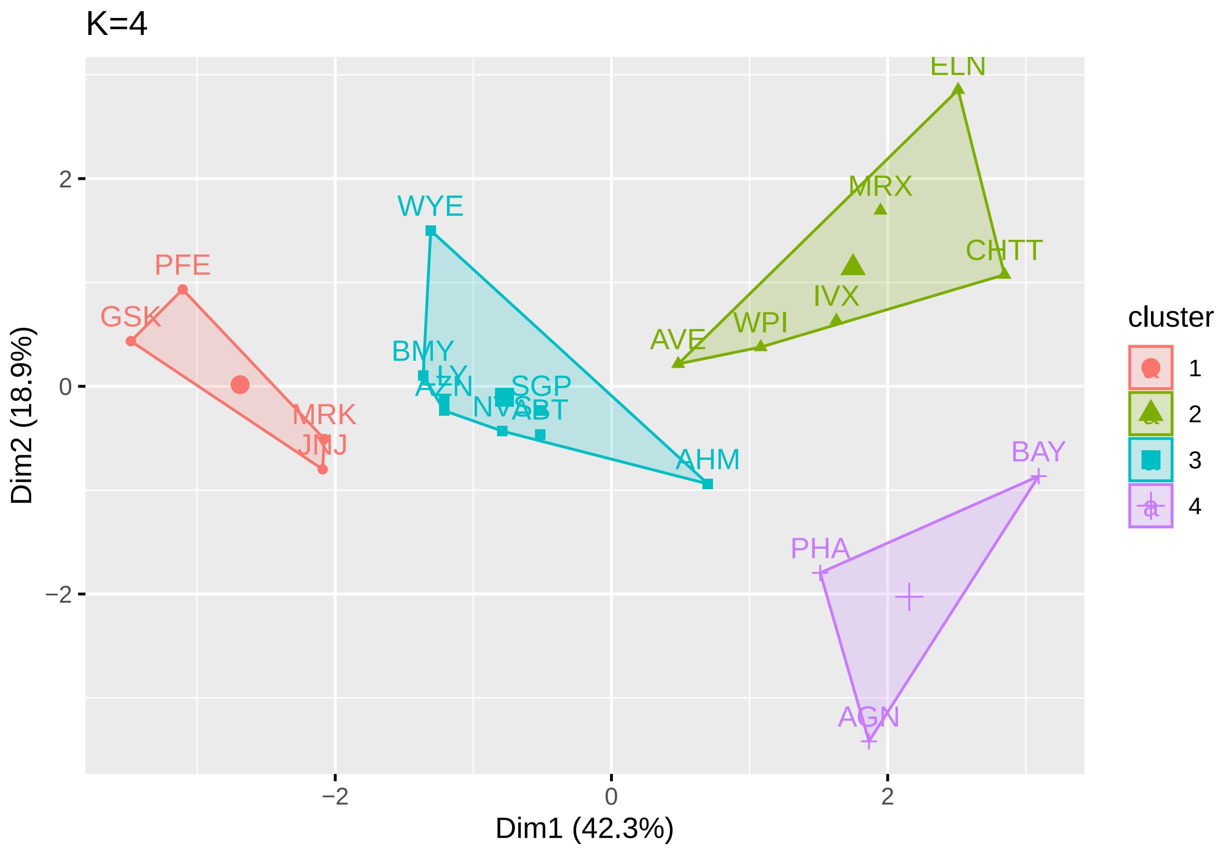
Here I am scaling the clutertered data

```
set.seed(5097)
scaling_data<-scale(clustered_data)
```

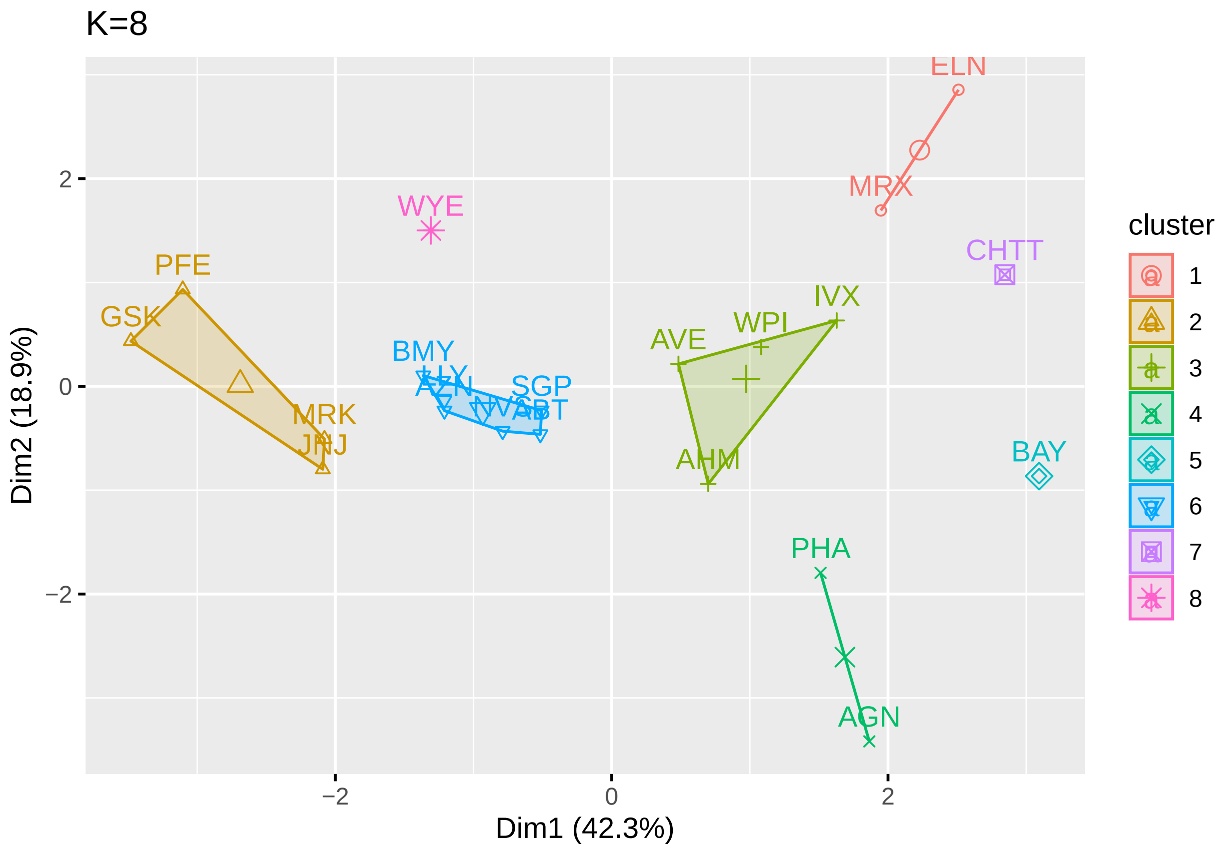
Here I was doing K-means clustering with randomly selected K values.

```
set.seed(5097)
k_mean1<-kmeans(scaling_data,centers = 2, nstart = 15)
k_mean4<-kmeans(scaling_data,centers = 4, nstart = 15)
k_mean8<-kmeans(scaling_data,centers = 8, nstart = 15)

plot_k_mean1<-fviz_cluster(k_mean1,data = scaling_data) + ggtitle("K=2")
plot_k_mean4<-fviz_cluster(k_mean4,data = scaling_data) + ggtitle("K=4")
plot_k_mean8<-fviz_cluster(k_mean8,data = scaling_data) + ggtitle("K=8")

plot_k_mean1
```


```
plot_k_mean4
```

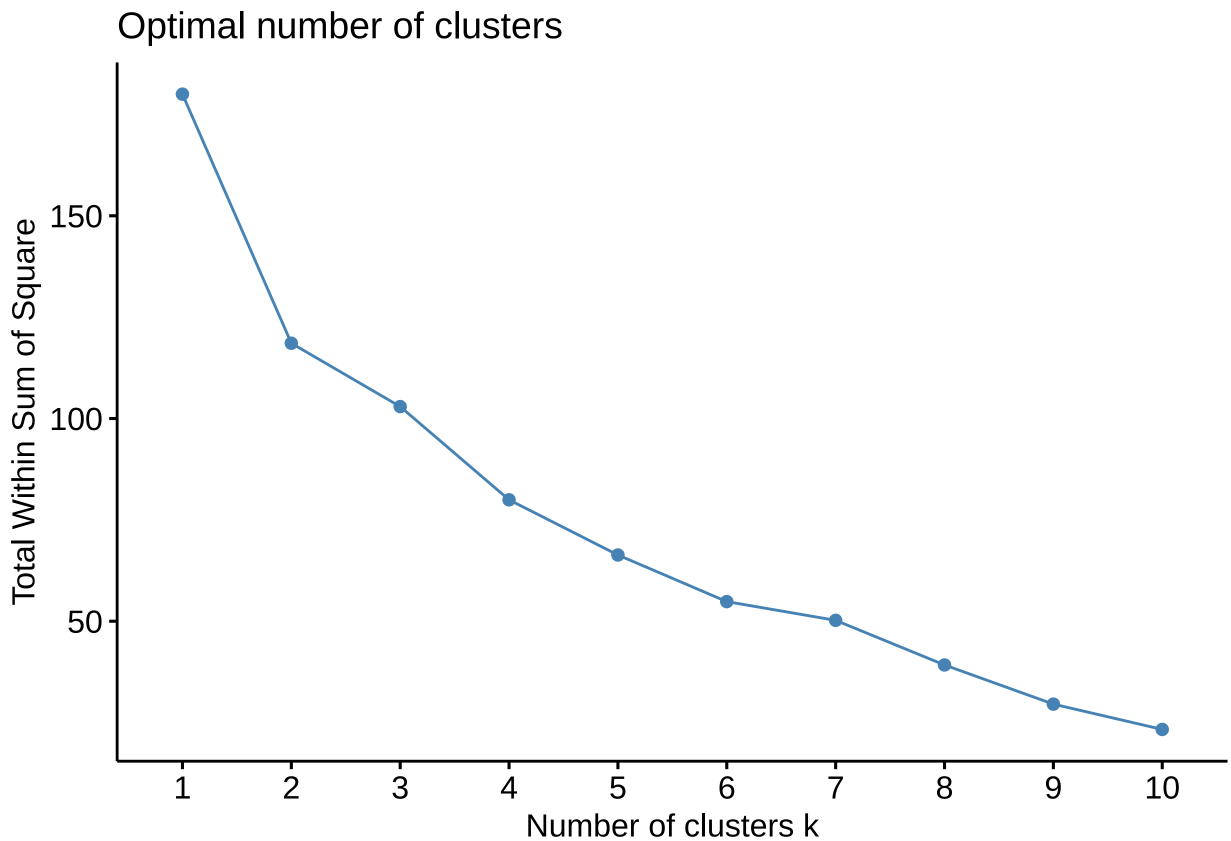

K=4
Dim2 (18.9%)
Dim1 (42.3%)
PFE
GSK
MRK
JNJ
WYE
BMY
AZN
NVS
SGP
ABT
AHM
ELN
MRX
CHTT
IVX
WPI
AVE
BAY
PHA
AGN
cluster
1
2
3
4

```
plot_k_mean8
```

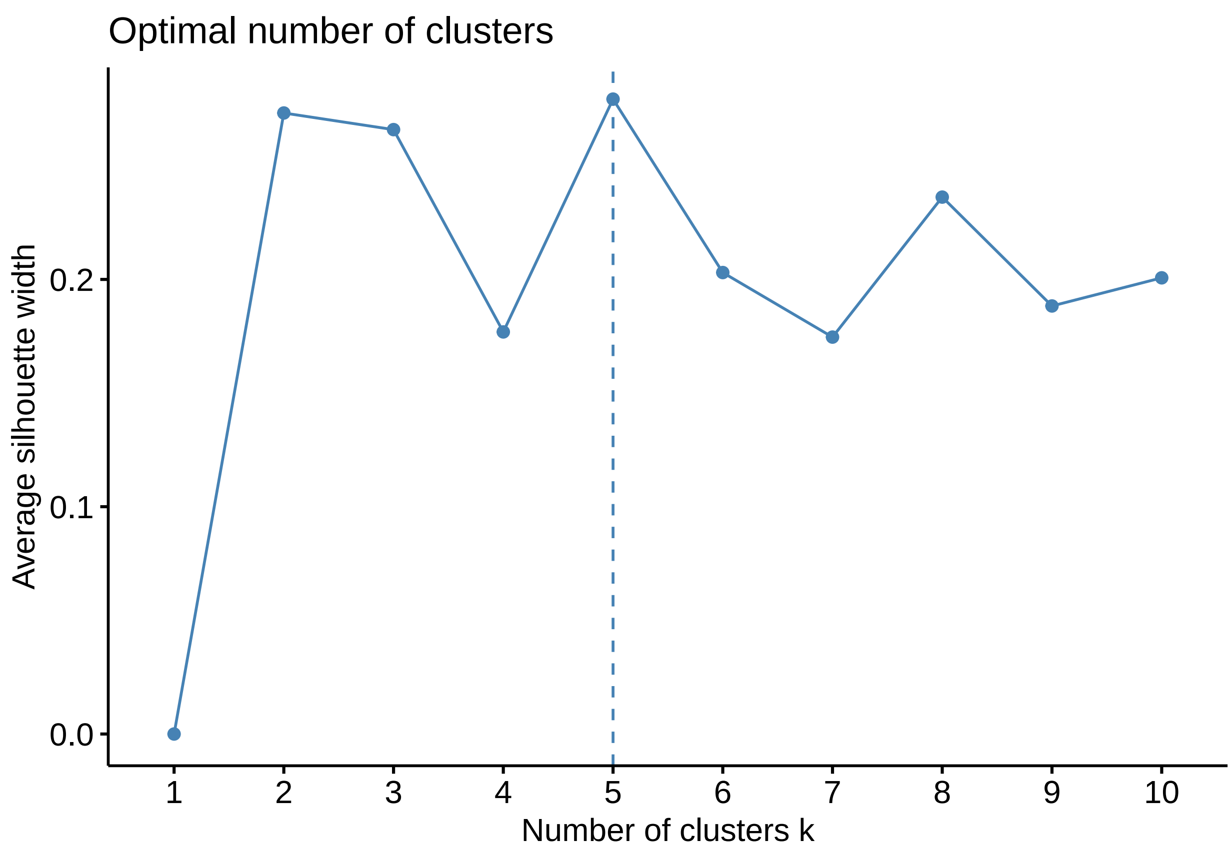


Employing WSS (Within-Cluster Sum of Square) and Silhouette scores to identify the optimal K value for clustering.

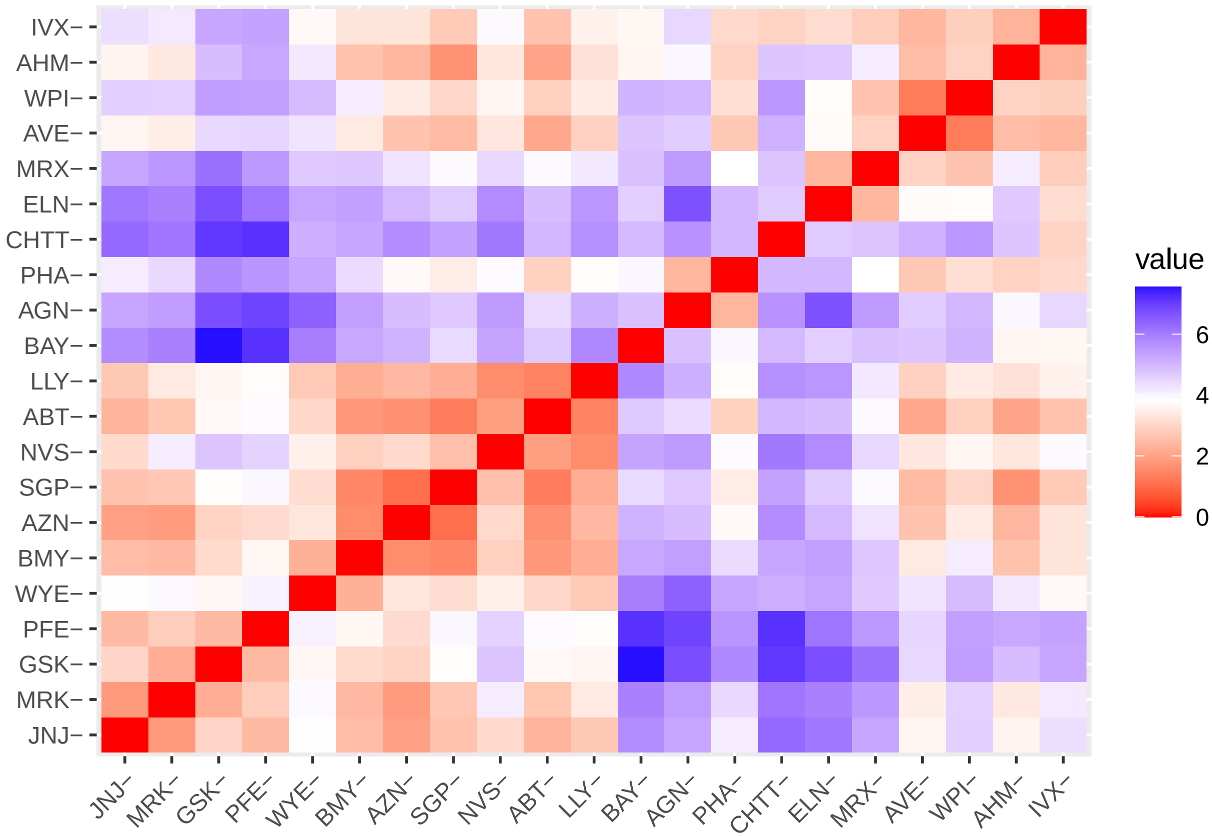
```
K_WSS<-fviz_nbclust(scaling_data,kmeans,method="wss")
K_Silhouette<-fviz_nbclust(scaling_data,kmeans,method="silhouette")
K_WSS
```




```
K_Silhouette
```




```r
dist<-dist(scaling_data,metho='euclidean')
fviz_dist(dist)
```



Based on the WSS, the optimal number of clusters (k) is 2, whereas the Silhouette score suggests k is 5. We are opting for k=5 as it guarantees a lower within-cluster sum of squares while also ensuring satisfactory separation between clusters.

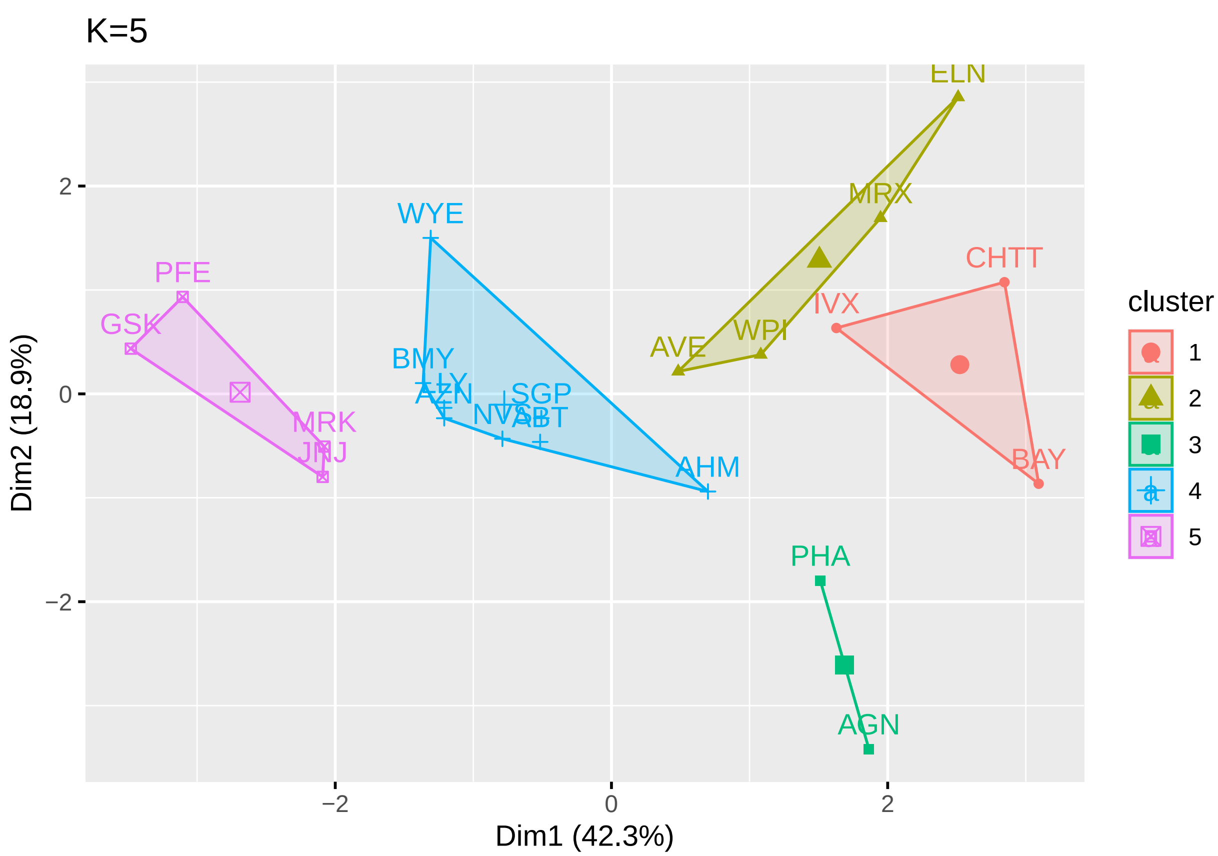
Here I was doing K Means to the required K

```
set.seed(5097)
k_mean5<-kmeans(scaling_data,centers = 5, nstart = 10)
k_mean5
```

```
## K-means clustering with 5 clusters of sizes 3, 4, 2, 8, 4
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 2 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 3 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 4 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
##      Leverage Rev_Growth Net_Profit_Margin
## 1  1.36644699 -0.6912914      -1.320000179
## 2  0.06308085  1.5180158      -0.006893899
## 3 -0.14170336 -0.1168459      -1.416514761
## 4 -0.27449312 -0.7041516       0.556954446
## 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
##    4    3    4    4    2    1    4    1    2    4    5    1    5    2    5    4
##  PFE  PHA  SGP  WPI  WYE
##    5    3    4    2    4
##
## Within cluster sum of squares by cluster:
## [1] 15.595925 12.791257  2.803505 21.879320  9.284424
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
plot_kmeans5<-fviz_cluster(k_mean5,data = scaling_data) + ggtitle("K=5")
plot_kmeans5
```




```
clustering_data1<-clustered_data%>%
  mutate(Cluster_no=k_mean5$cluster)%>%
  group_by(Cluster_no)%>%summarise_all('mean')
clustering_data1
```

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

Below companies are grouped into following clusters:

Cluster_1= BAY,CHTT,IVX

Cluster_2= AVE,ELN,MRX,WPI

Cluster_3=AGN,PHA

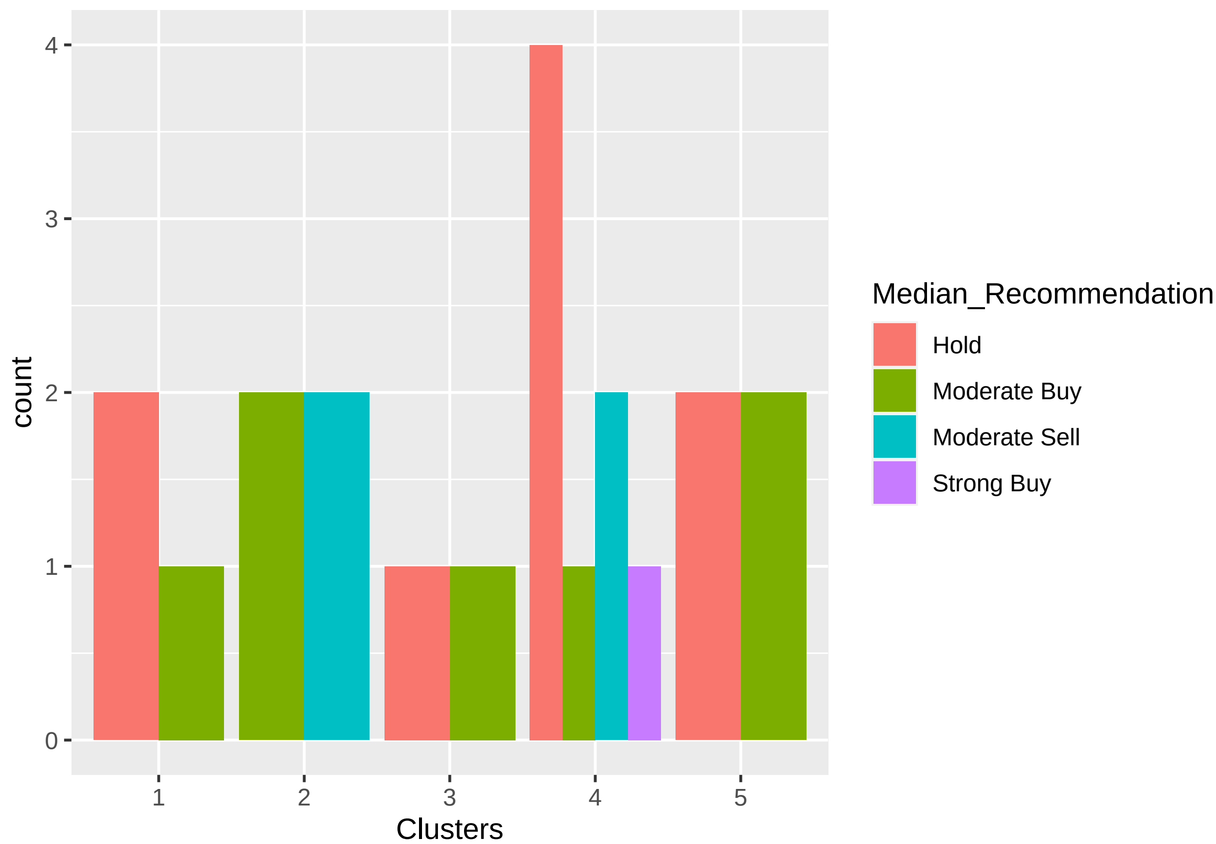
Cluster_4= ABT,AHM,AZN,BMY,LLY,NVS,SGP,WYE

Cluster_5=GSK,JNJ,PFE,MRK

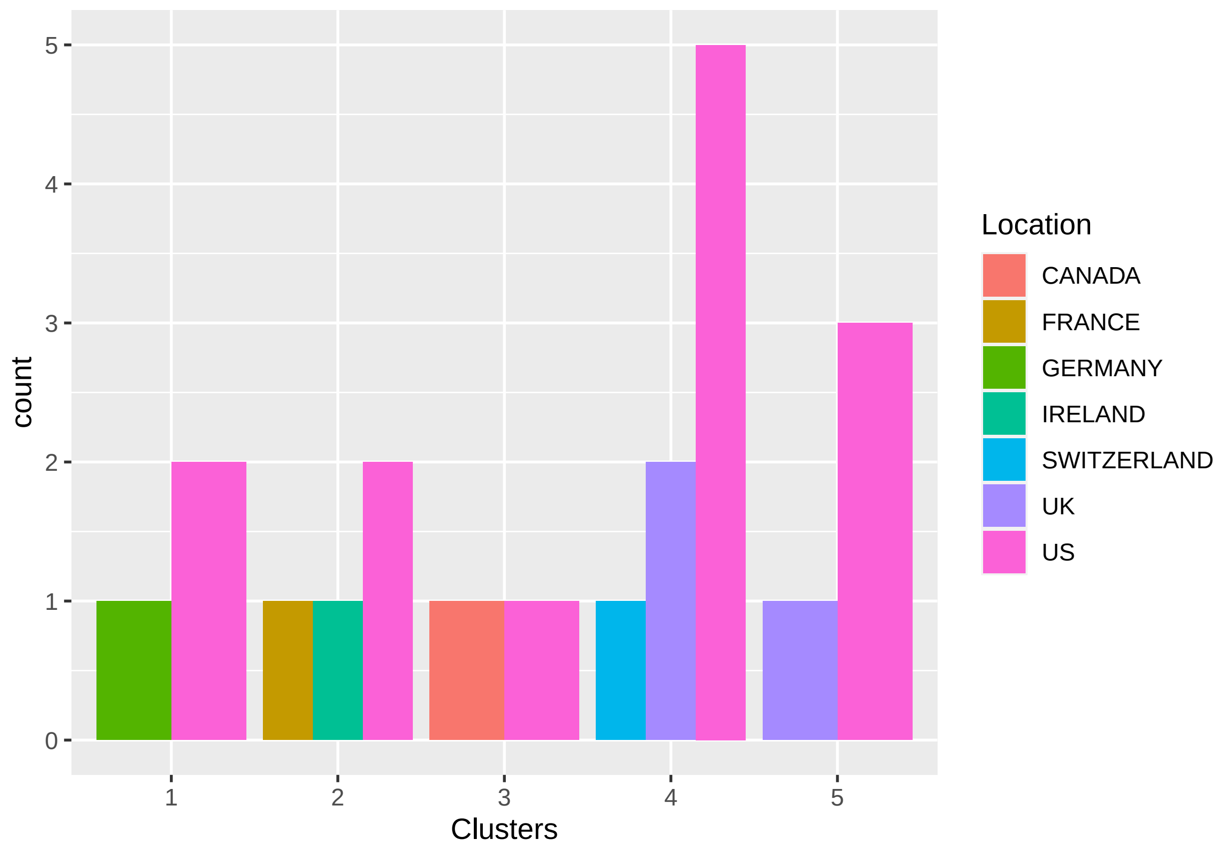
From the clusters formed it can be understood that

1. Cluster_1 includes companies with extremely poor ROA, ROE, market capitalization, and asset turnover, indicating a high level of risk associated with these firms.
2. Cluster_2 has companies collect firms resembling those in cluster_1, but with a bit less risk involved.
3. Cluster_3 has companies possess an excellent PE_ratio but suffer from very poor ROA and ROE, making them riskier than those in cluster_1.
4. Cluster_4 has collection of businesses with moderate return on equity and return on investment.
5. Cluster_5 has companies exhibiting excellent market capitalization, ROE, and ROA

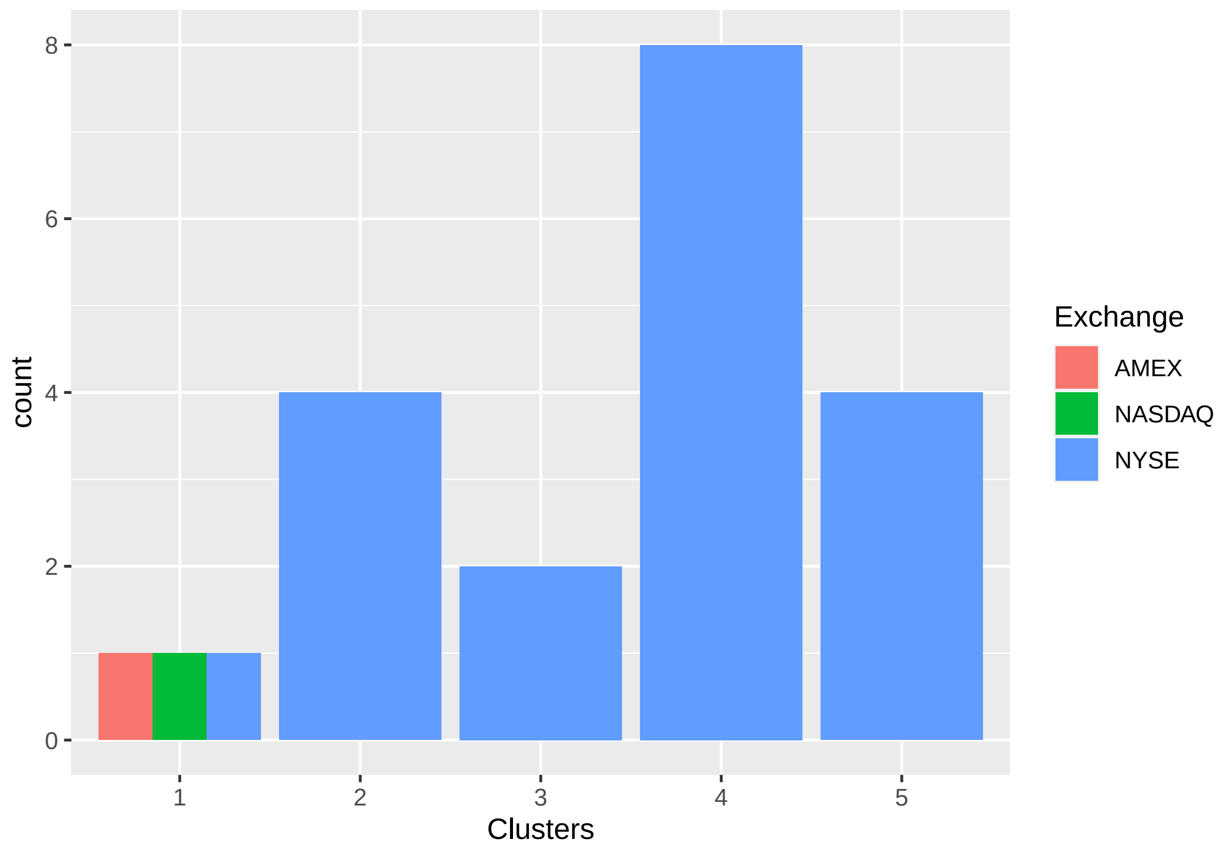
```r
clustering_dataset2<- pharma_data[,12:14] %>% mutate(Clusters=k_mean5$cluster)
ggplot(clustering_dataset2, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(posi
```




```
ggplot(clustering_dataset2, mapping = aes(factor(Clusters),fill = Location))+geom_bar(position = 'dodge
```




```
ggplot(clustering_dataset2, mapping = aes(factor(Clusters),fill = Exchange))+geom_bar(position = 'dodge
```



It's observable that there's a trend between clusters and the Median Recommendation variable. For instance, the first cluster implies a recommendation ranging from hold to moderate buy, while the second cluster leans towards a moderate buy to moderate sell suggestion. The location graph indicates that a majority of the pharmaceutical companies are based in the US, and there doesn't appear to be a significant pattern beyond that. The clusters do not exhibit a distinct pattern in relation to the stock exchange, aside from the observation that the bulk of the companies are traded on the NYSE.

Naming clusters:

[Based on the companies listed for each cluster, which seem to represent pharmaceutical firms]

Cluster 1: Innovative Biotech Pioneers.

Cluster 2: Specialty Pharma Developers.

Cluster 3:Focused Healthcare Duo.

Cluster 4: Diversified Healthcare Giants.

Cluster 5: Global Healthcare Leaders.