Assignment_4

Ranadheer Gonuguntla

Loading the required libraries and data set

```
library(tidyverse)
```

```
## -- 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   3.4.3     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
```

```
library(factoextra)
```

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

```
pharmaceutical_0data<-read.csv("C:/Users/pc/Downloads/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]
```

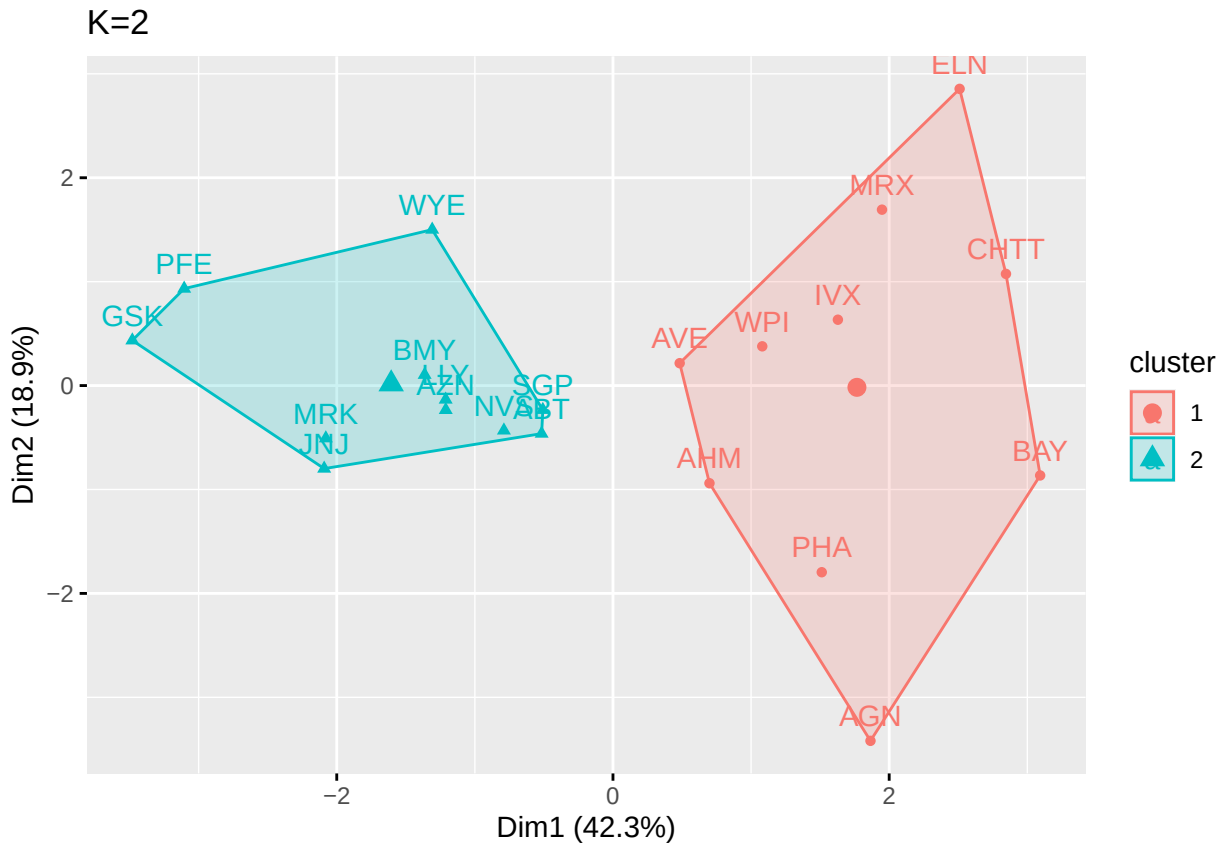
Scaling the data according to requirement

```
set.seed(143)
ScaledData<-scale(Clustering_data)
```

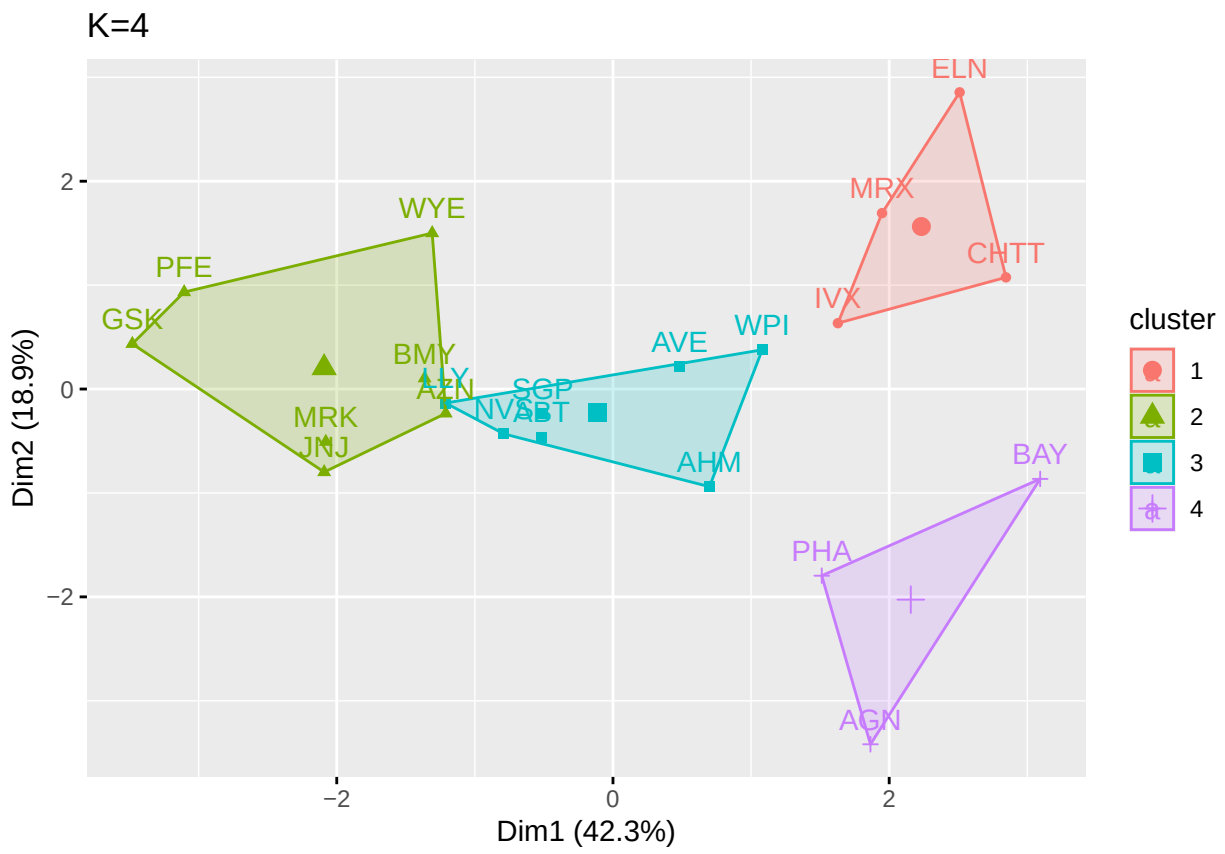
Performing Kmeans clustering for random K values(trail and error)

```
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
```



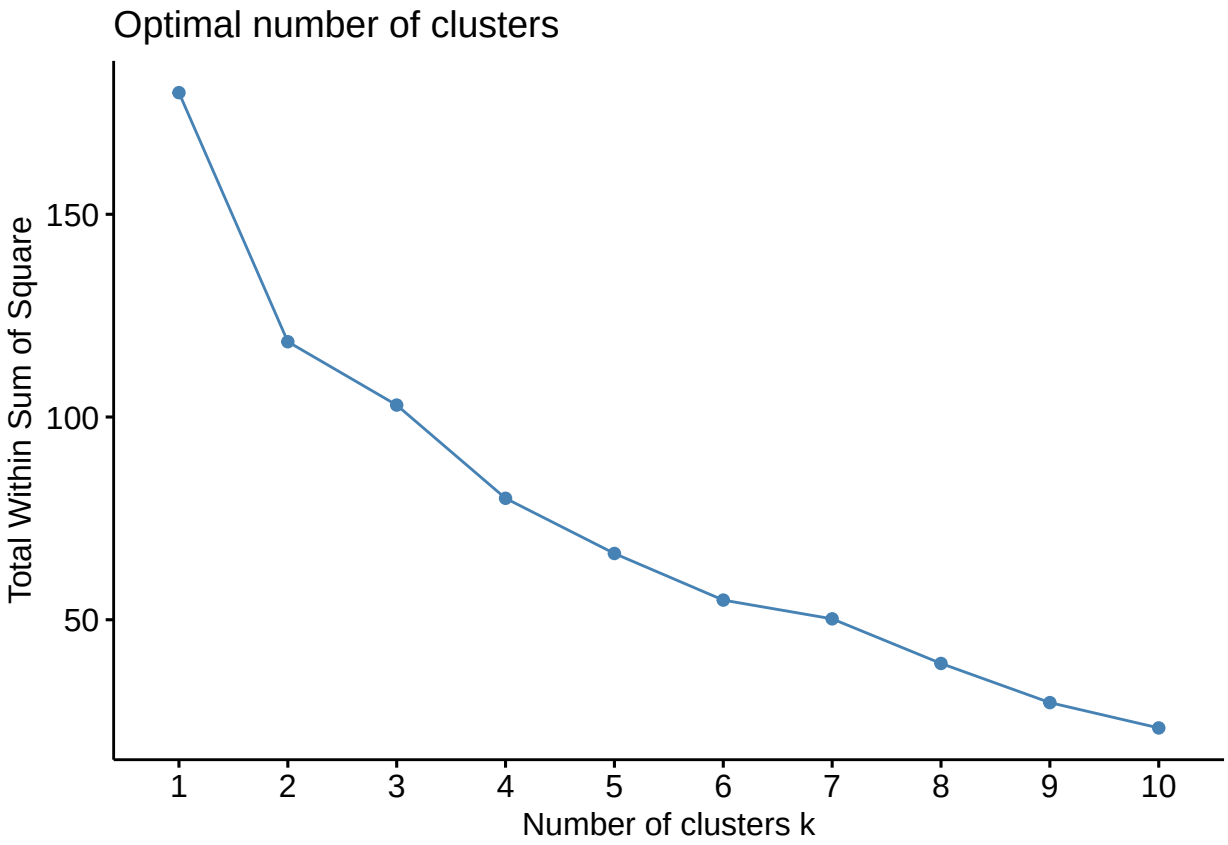

```
plot_kmeans_4
```



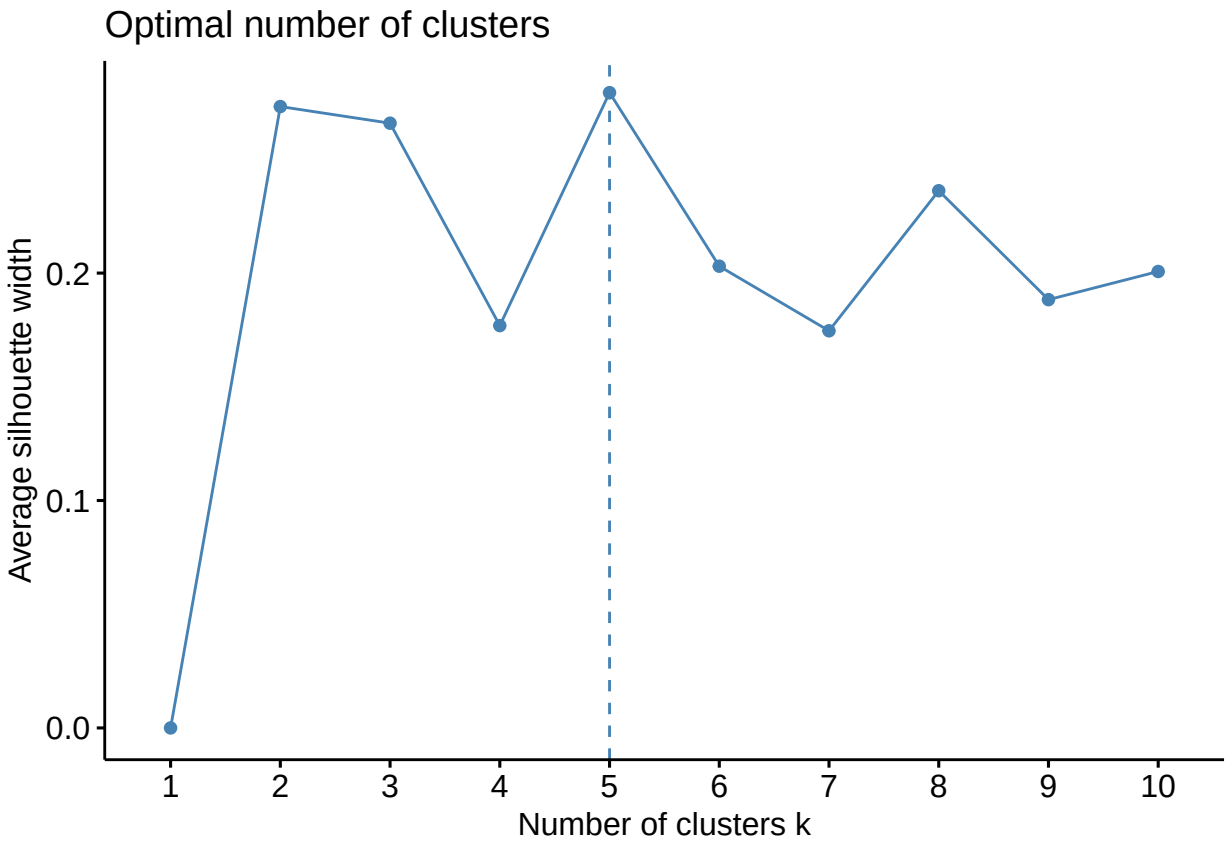

```
plot_kmeans_8
```

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

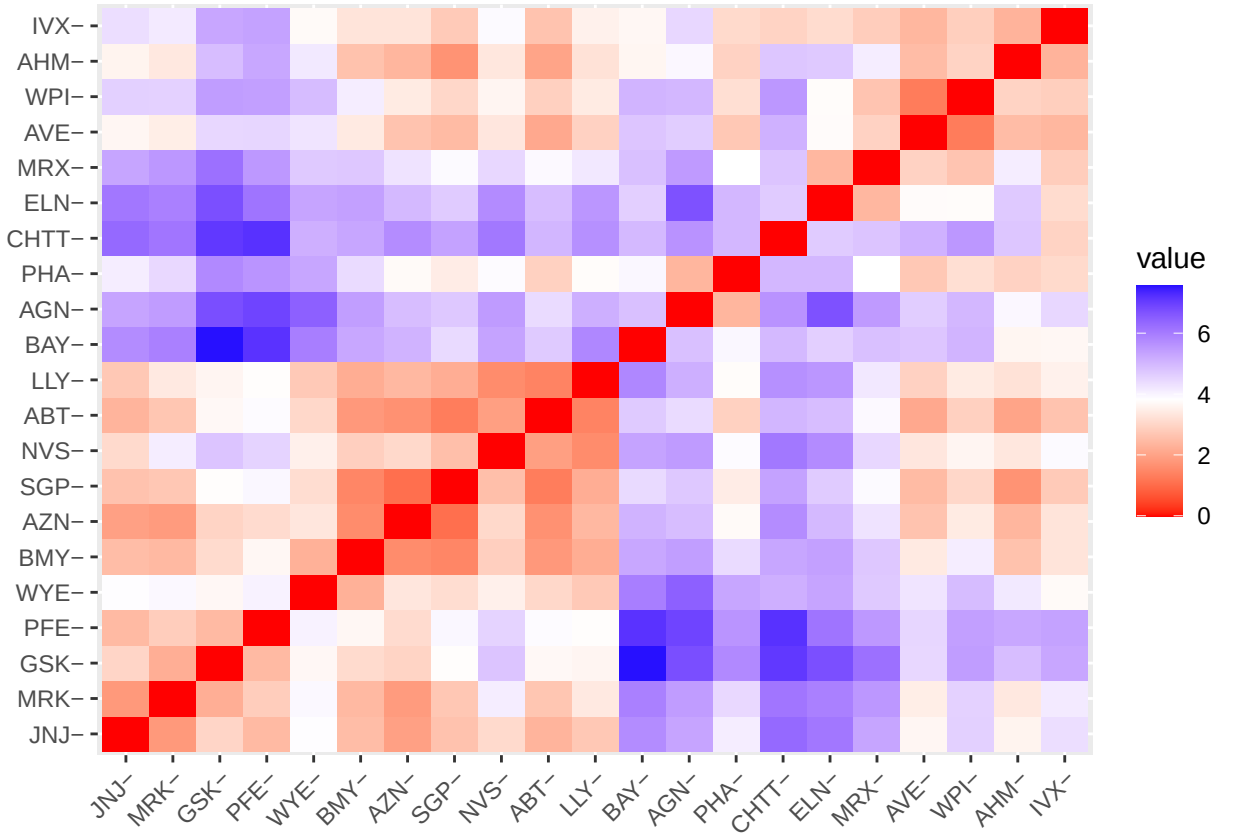
```
k_wss<-fviz_nbclust(ScaledData,kmeans,method="wss")
k_silhouette<-fviz_nbclust(ScaledData,kmeans,method="silhouette")
k_wss
```




```
k_silhouette
```




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



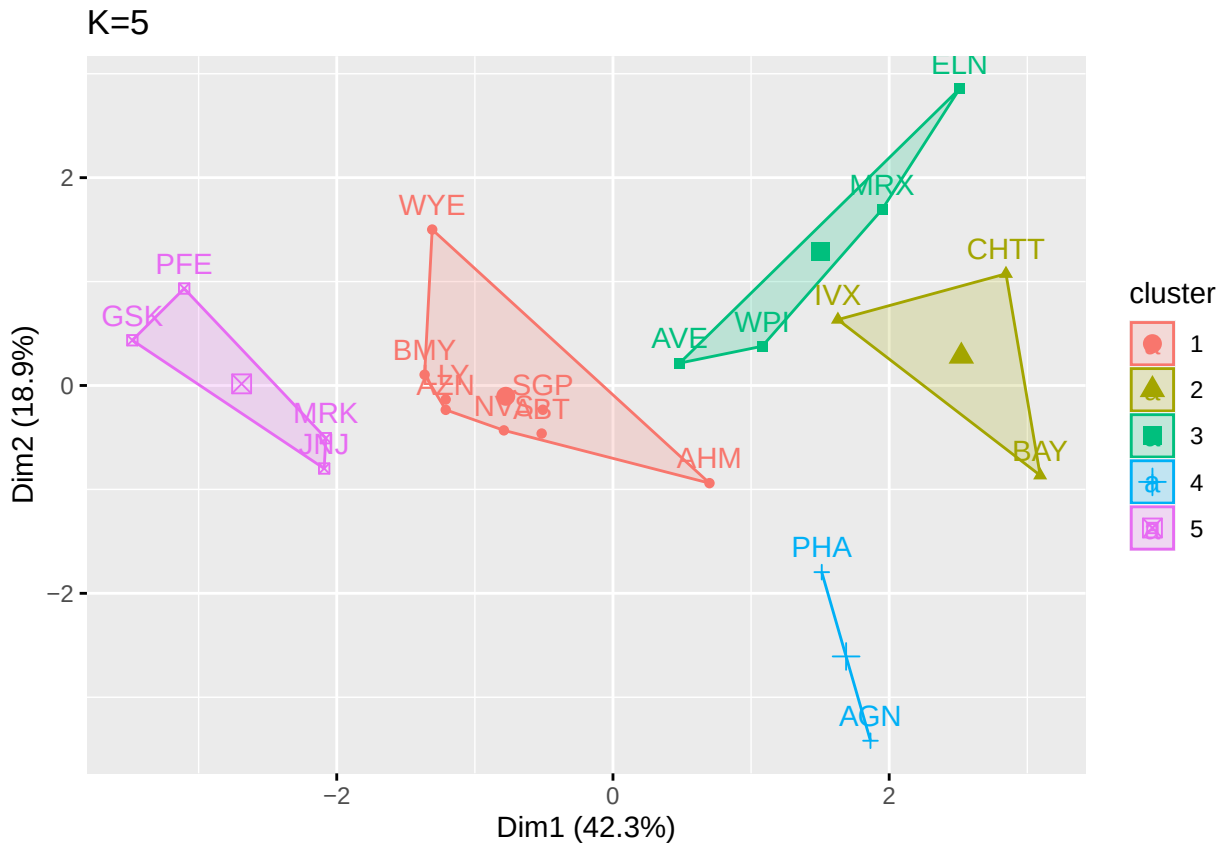
From wss K value is 2 and from silhouette k value is 5. we will be choosing 5 because this will ensure that the sum of squares is low along with proper separation within the clusters. Now lets Perform Kmeans to find out suitable k

```
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"
```

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



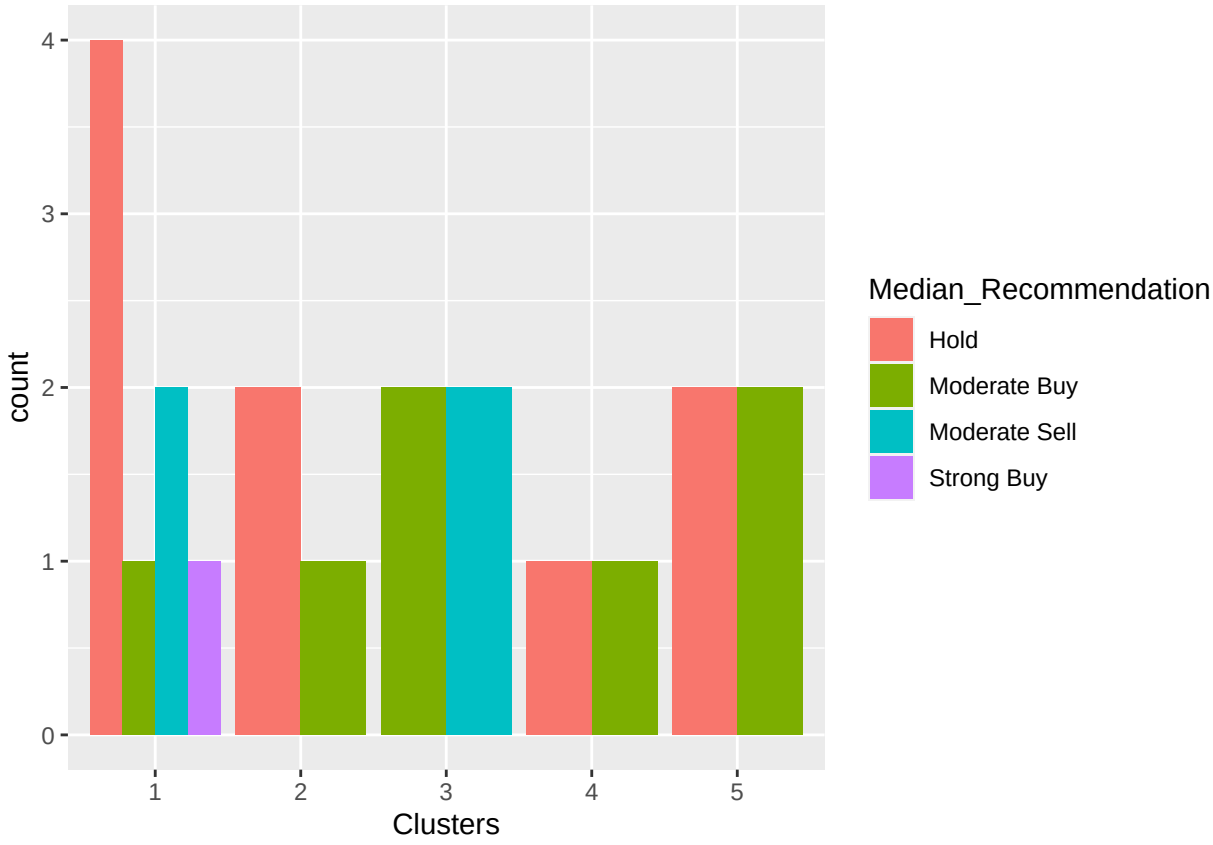

```
Clustering_data_1<-Clustering_data%>%
  mutate(Cluster_no=kmeans_5$cluster)%>%
  group_by(Cluster_no)%>%summarise_all('mean')
Clustering_data_1
```

```
## # 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
```

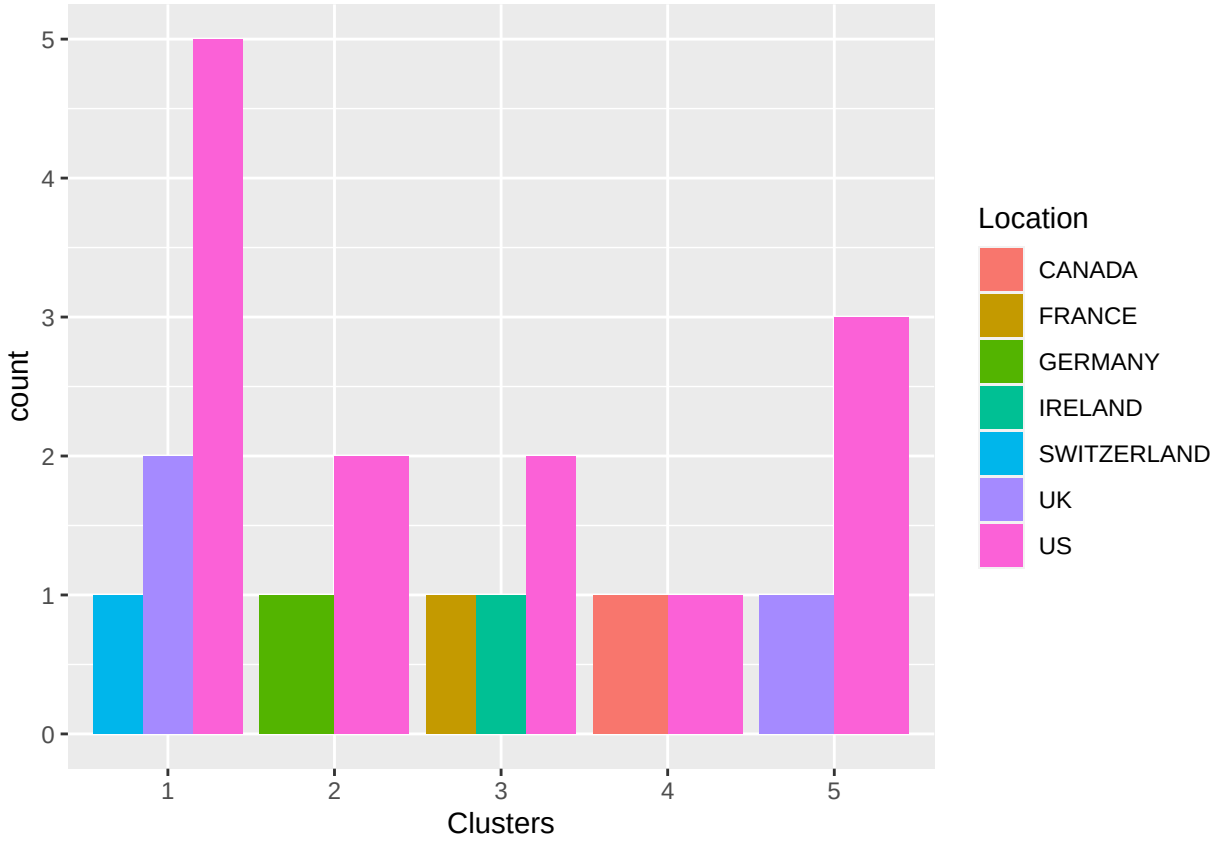
```
## 2          2       6.64 0.87      24.6  16.5  4.17          0.6      1.65
## 3          3      13.1  0.598     17.7  14.6  6.2           0.425    0.635
## 4          4      31.9  0.405     69.5  13.2  5.6           0.75     0.475
## 5          5     157.   0.48      22.2  44.4 17.7           0.95     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

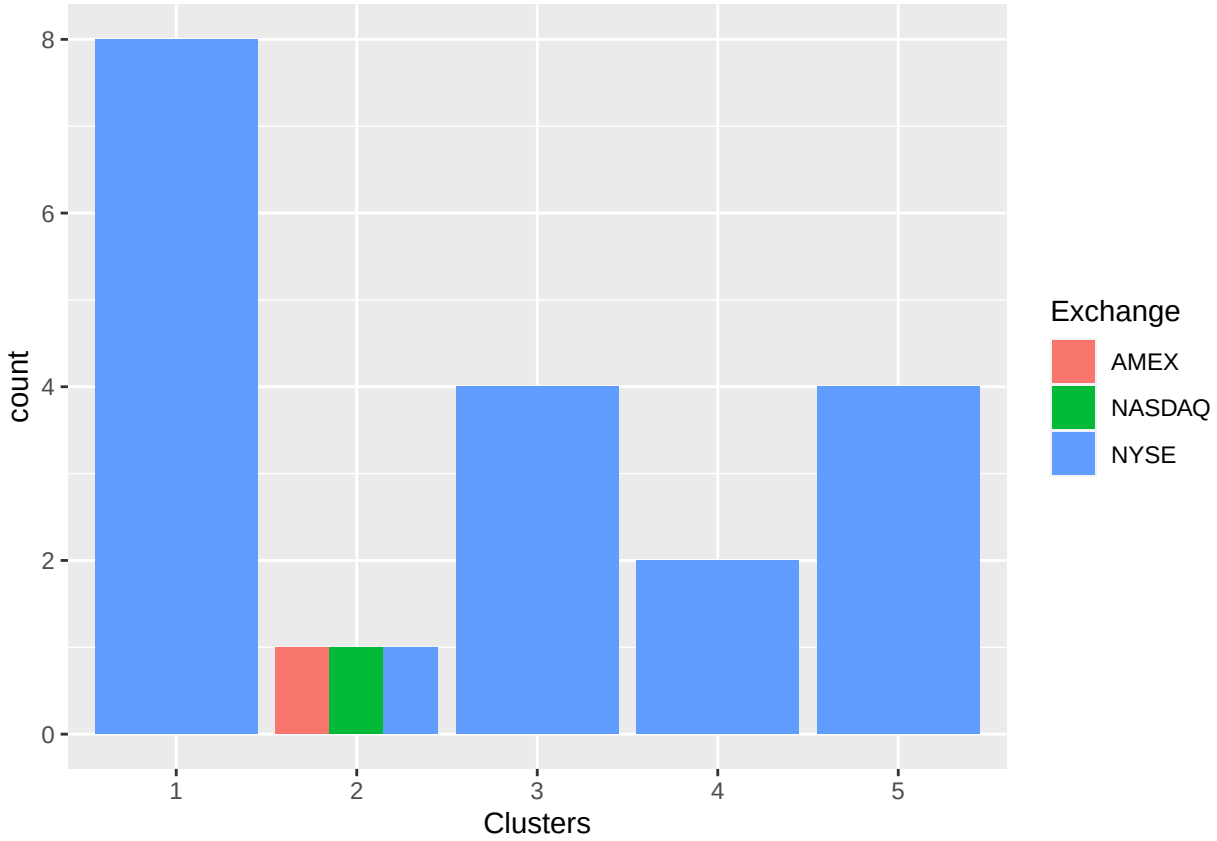
```
Clustering_data_2<- pharmaceutical_data[,12:14] %>% mutate(Clusters=kmeans_5$cluster)
ggplot(Clustering_data_2, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(positi
```



```
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')+
```



Observing the data, a clear pattern emerges in the relationship between the clusters and the 'Median Recommendation' variable. Cluster 2 tends to suggest a recommendation ranging from 'hold' to 'moderate buy,' while Cluster 3 leans toward 'moderate buy' to 'moderate sell.' When examining the location graph, it becomes evident that a significant proportion of pharmaceutical companies are based in the United States, and there isn't a conspicuous pattern in their distribution. However, there's no discernible pattern between clusters and the stock exchange, except for the fact that the majority of companies are listed on the New York Stock Exchange (NYSE). We can assign names to these clusters based on the combined criteria of net market capitalization and return on assets: Naming clusters:

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

Cluster 1:High_Million

Cluster 2: Extra Tiny_penny

Cluster 3: Dim Sums of Money

Cluster 4: Mid-Hundreds

Cluster 5: Extra Large-Millions