Assignment 4

Venkata Naga Siddartha Gutha

Loading libraries and data set

```
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   1.2.0      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)
```

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

```
pharmaceutical_data<-read.csv("C:/Users/sidda/Downloads/Pharmaceuticals.csv")
pharmaceutical_data<-na.omit(pharmaceutical_data)
```

Using the numerical variables (1 to 9) to cluster the 21 firms.

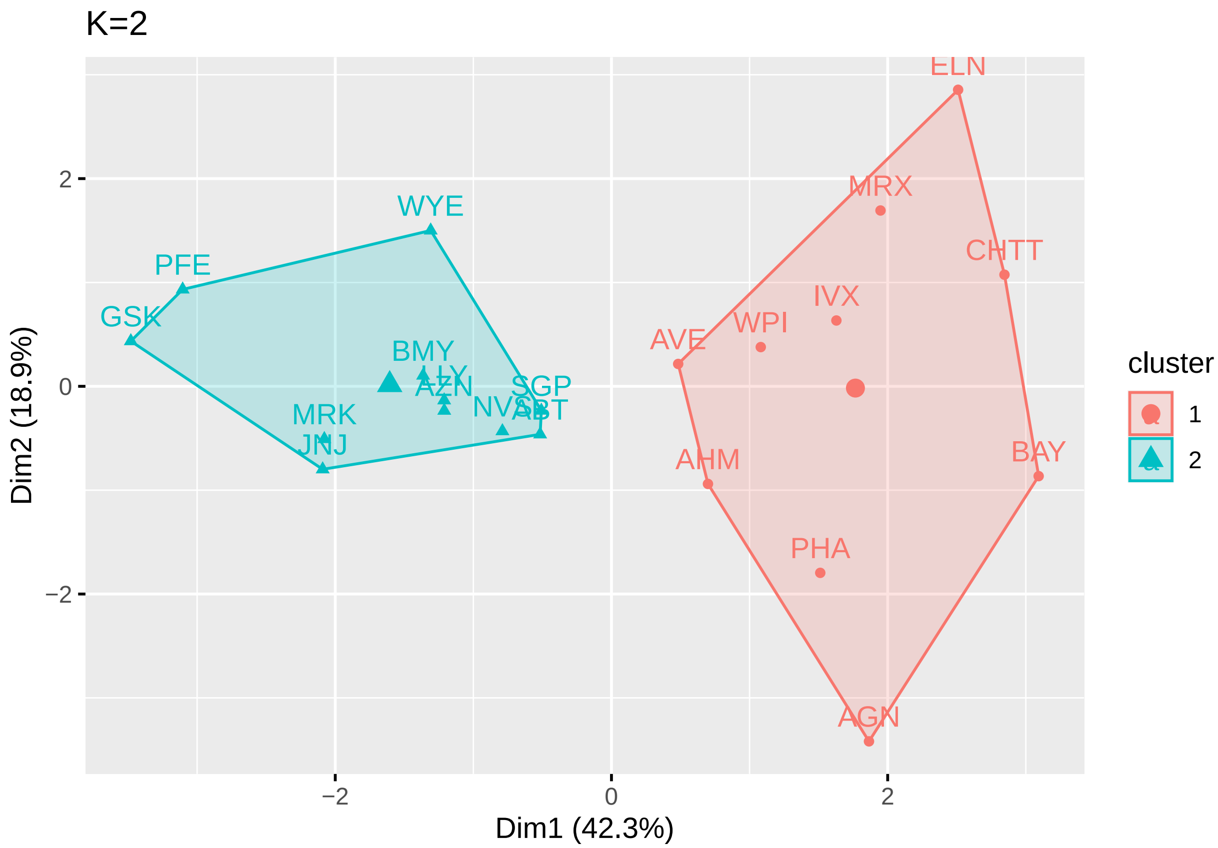
```
row.names(pharmaceutical_data)<-pharmaceutical_data[,1]
Clustering_dataset<-pharmaceutical_data[,3:11]
```

Scaling the data

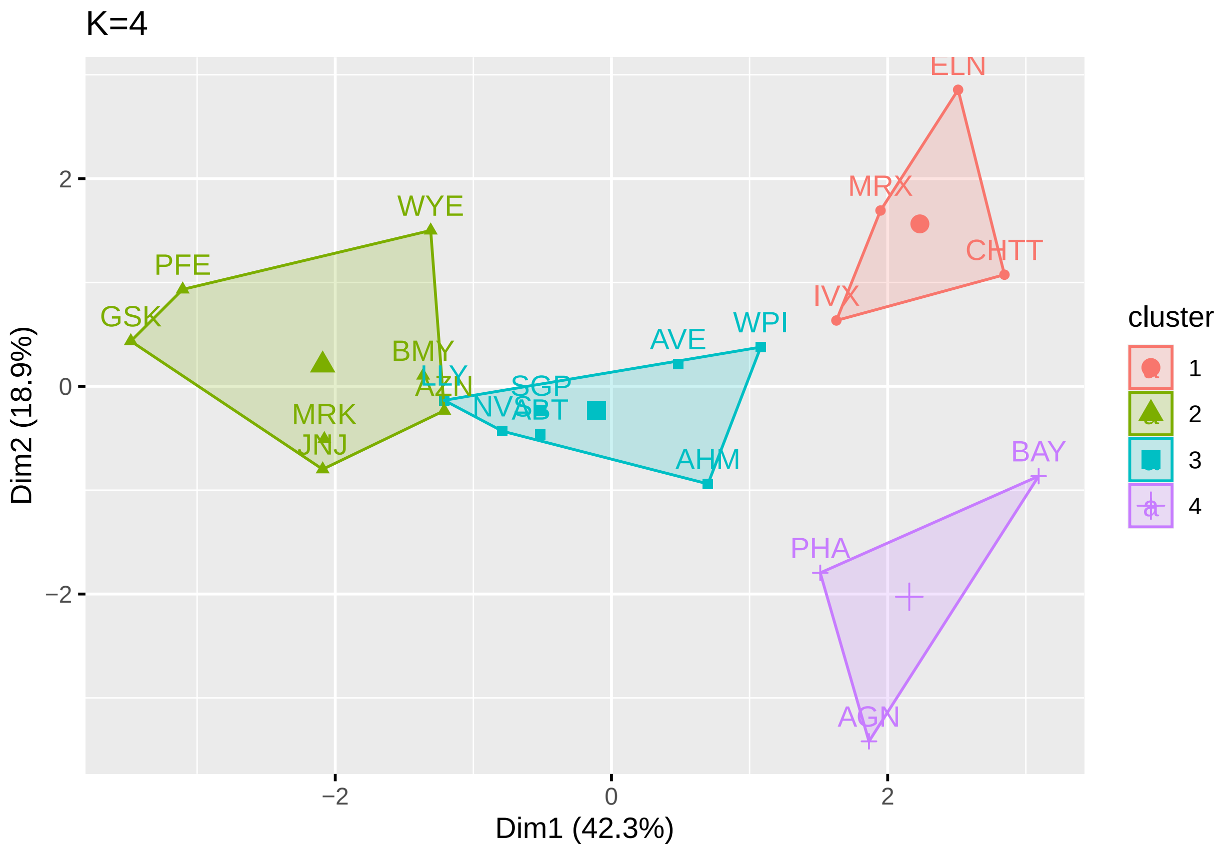
```
set.seed(143)
Scaled_data<-scale(Clustering_dataset)
```

Performing Kmeans for random K values

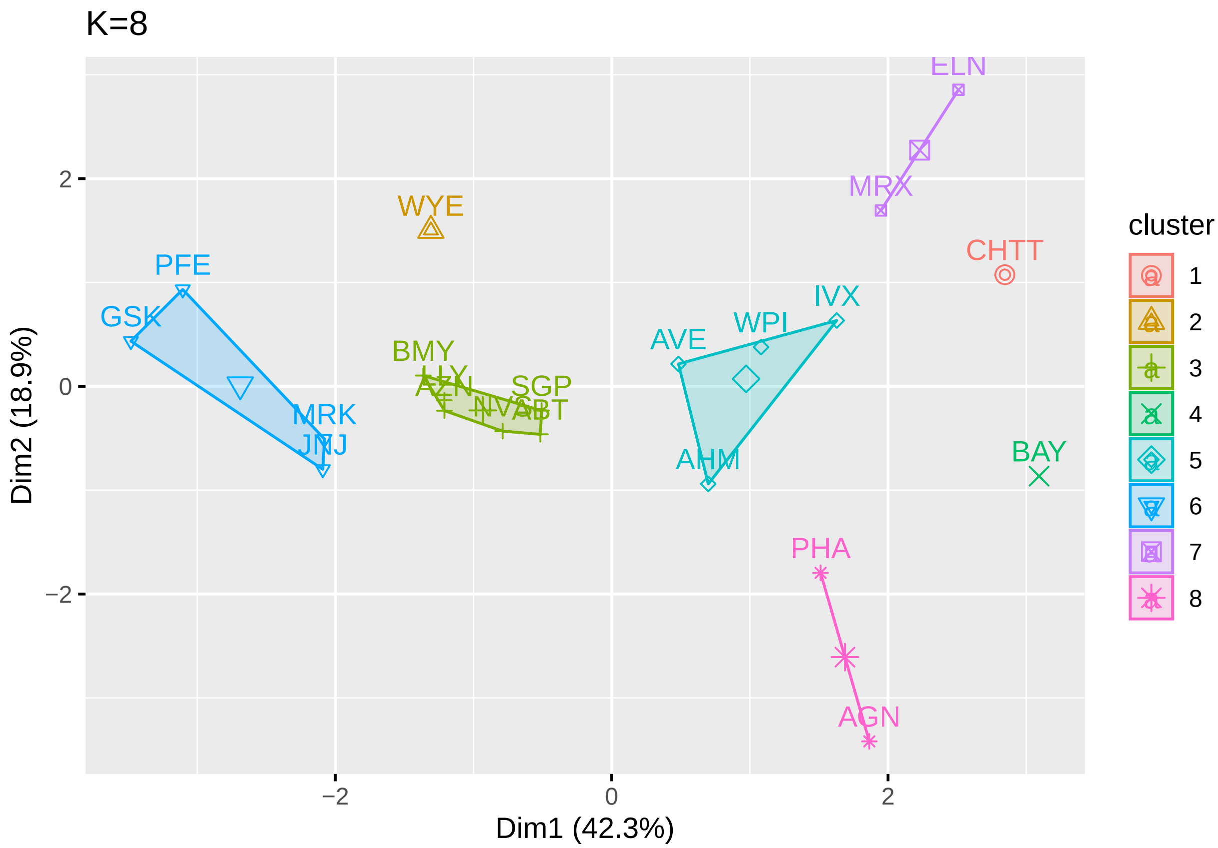
```
set.seed(143)
kmeans_2<-kmeans(Scaled_data,centers = 2, nstart = 15)
kmeans_4<-kmeans(Scaled_data,centers = 4, nstart = 15)
kmeans_8<-kmeans(Scaled_data,centers = 8, nstart = 15)

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

plot_kmeans_2
```




```
plot_kmeans_4
```

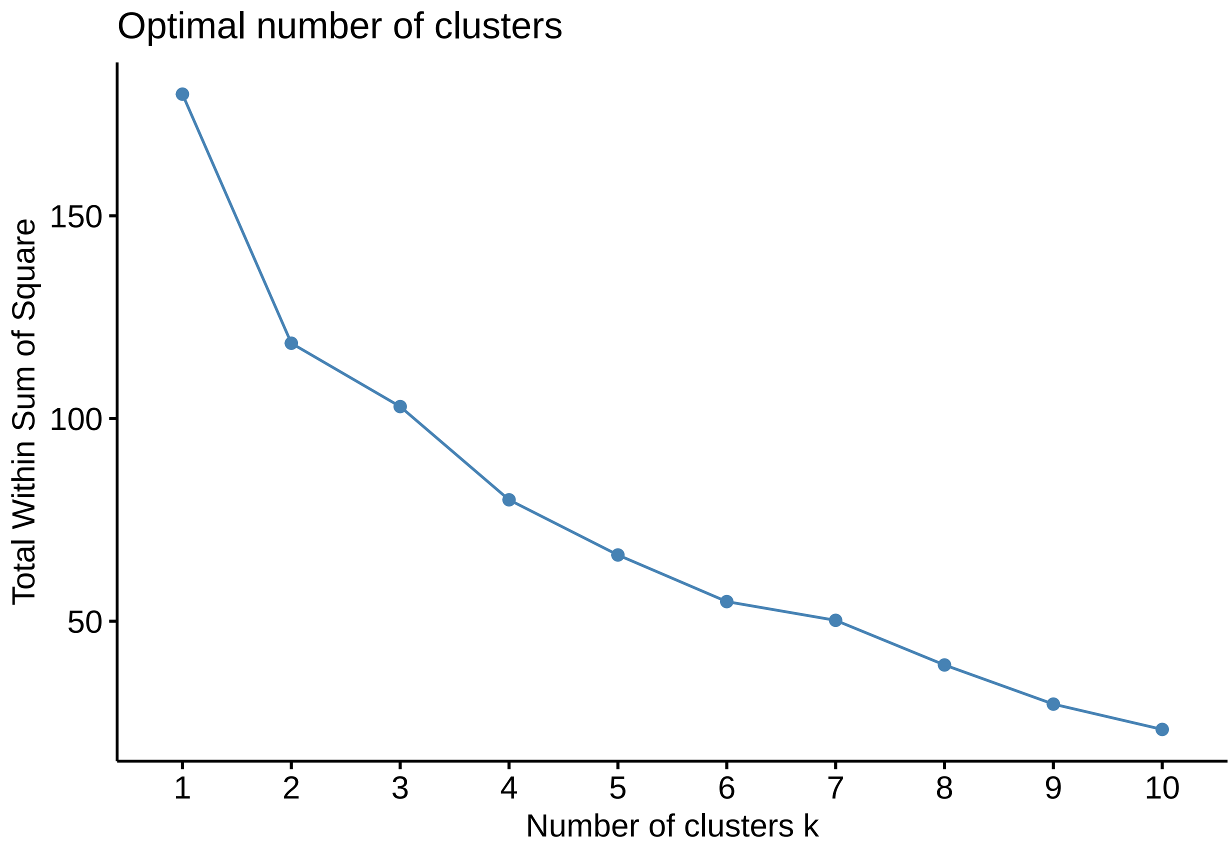



```
plot_kmeans_8
```

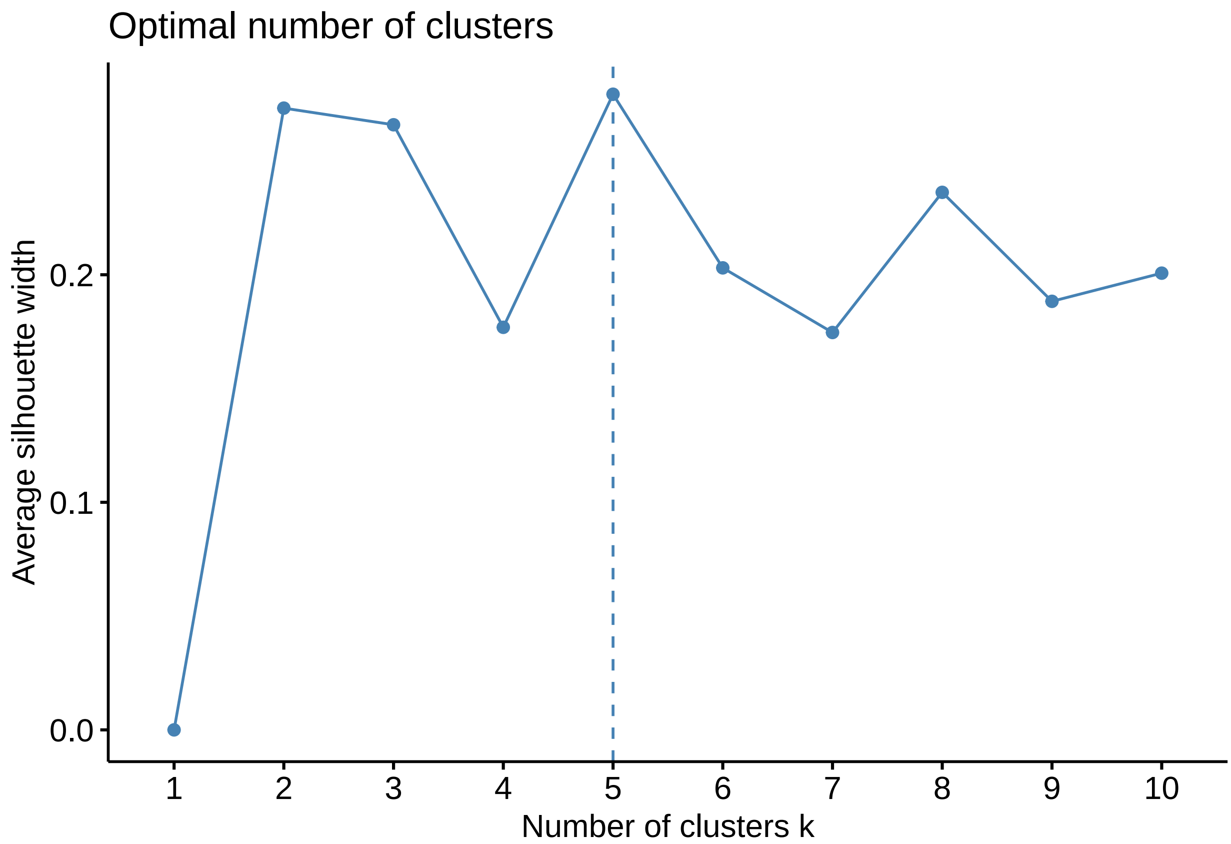


Using WSS and Silhouette to find best K suitable for clustering

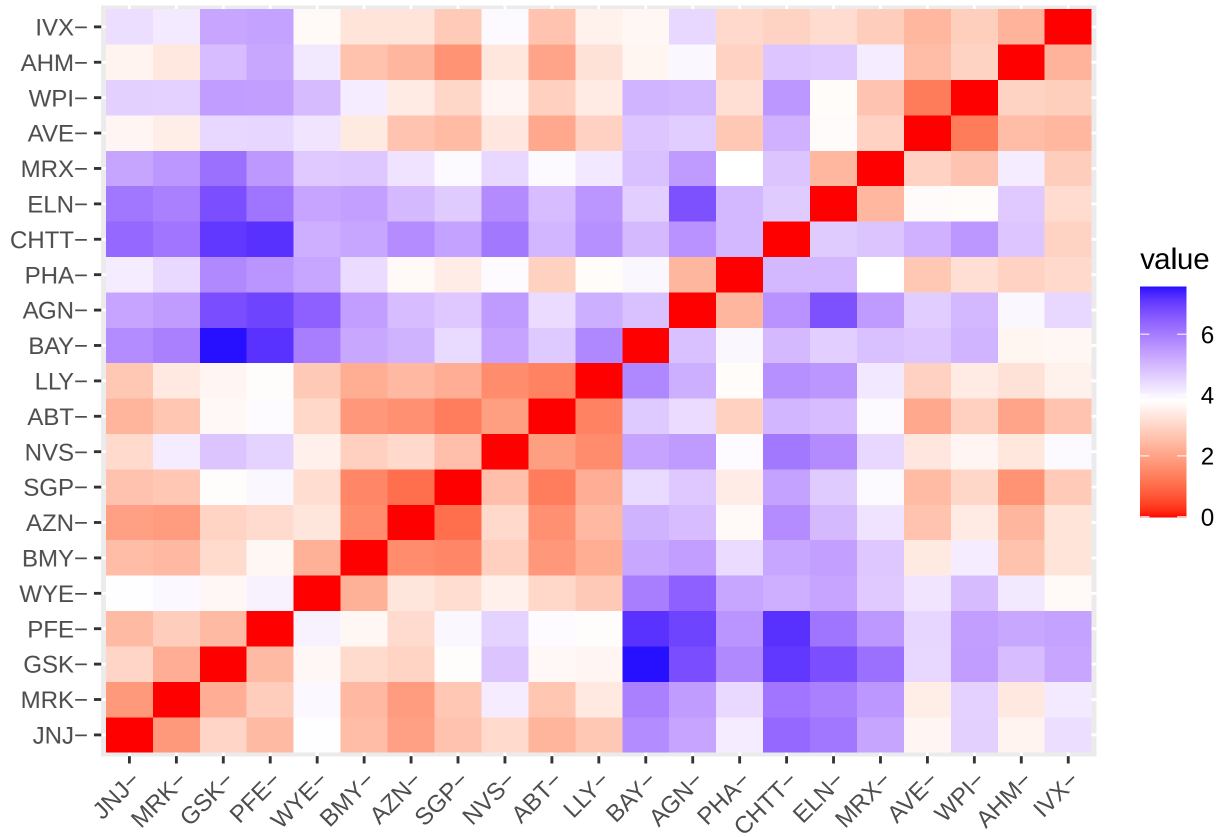
```
k_wss<-fviz_nbclust(Scaled_data,kmeans,method="wss")
k_silhouette<-fviz_nbclust(Scaled_data,kmeans,method="silhouette")
k_wss
```




```
k_silhouette
```




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



from WSS k is 2 and from silhouette k is 5. we are choosing 5 as this ensures that within sum of squires is low along with good separation within clusters

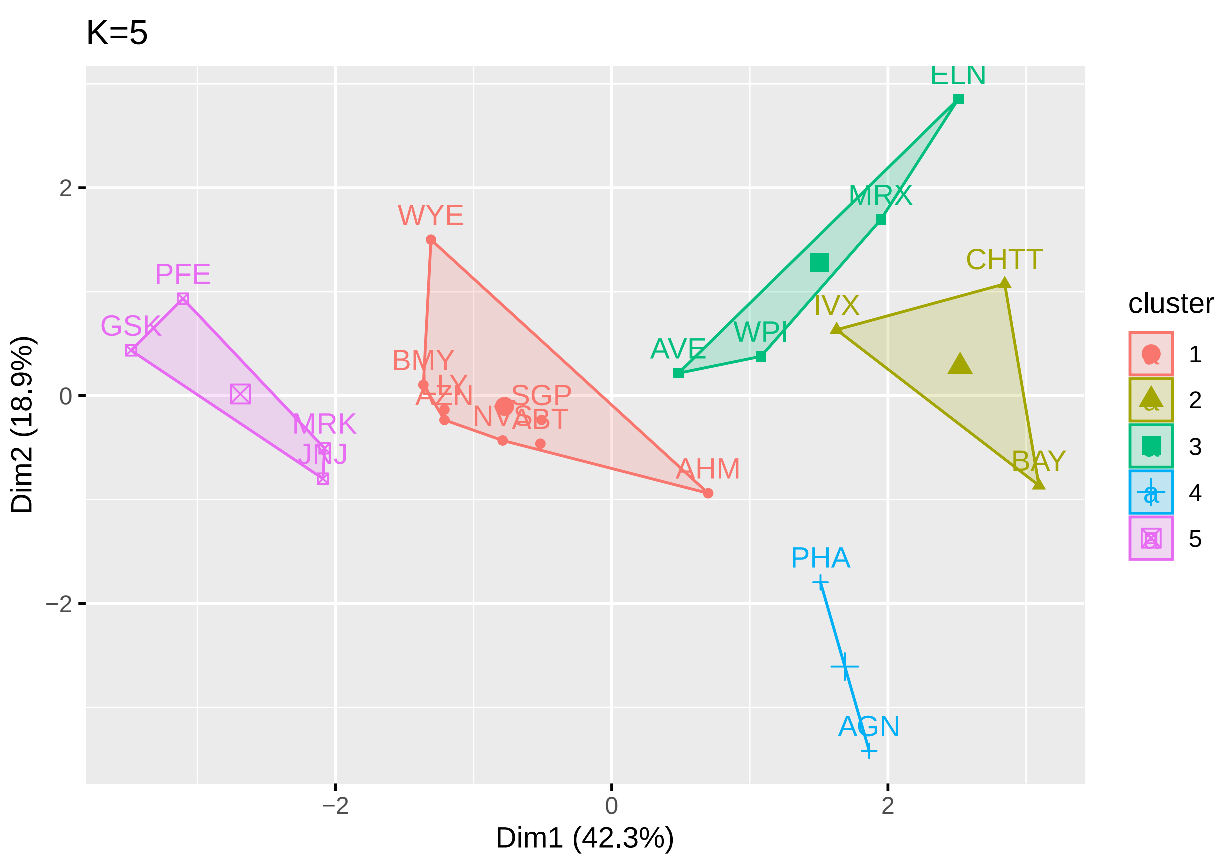
Performing Kmeans for suitable k

```
set.seed(143)
kmeans_5<-kmeans(Scaled_data,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 = Scaled_data) + ggtitle("K=5")
plot_kmeans_5
```




```
Clustering_dataset_1<-Clustering_dataset%>%
  mutate(Cluster_no=kmeans_5$cluster)%>%
  group_by(Cluster_no)%>%summarise_all('mean')
Clustering_dataset_1
```

```
## # A tibble: 5 x 10
##   Cluster_no Market_~1  Beta PE_Ra~2   ROE   ROA Asset~3 Lever~4 Rev_G~5 Net_P~6
##        <int>     <dbl> <dbl>   <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1          1      55.8 0.414    20.3  28.7 12.7    0.738   0.371    5.59    19.4
```

```
## 2          2      6.64 0.87      24.6  16.5  4.17   0.6     1.65     5.73    7.03
## 3          3     13.1  0.598     17.7  14.6  6.2    0.425   0.635   30.1    15.6
## 4          4     31.9  0.405     69.5  13.2  5.6    0.75    0.475   12.1     6.4
## 5          5    157.   0.48      22.2  44.4 17.7    0.95    0.22    18.5    19.6
## # ... with abbreviated variable names 1: Market_Cap, 2: PE_Ratio,
## #   3: Asset_Turnover, 4: Leverage, 5: Rev_Growth, 6: Net_Profit_Margin
```

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

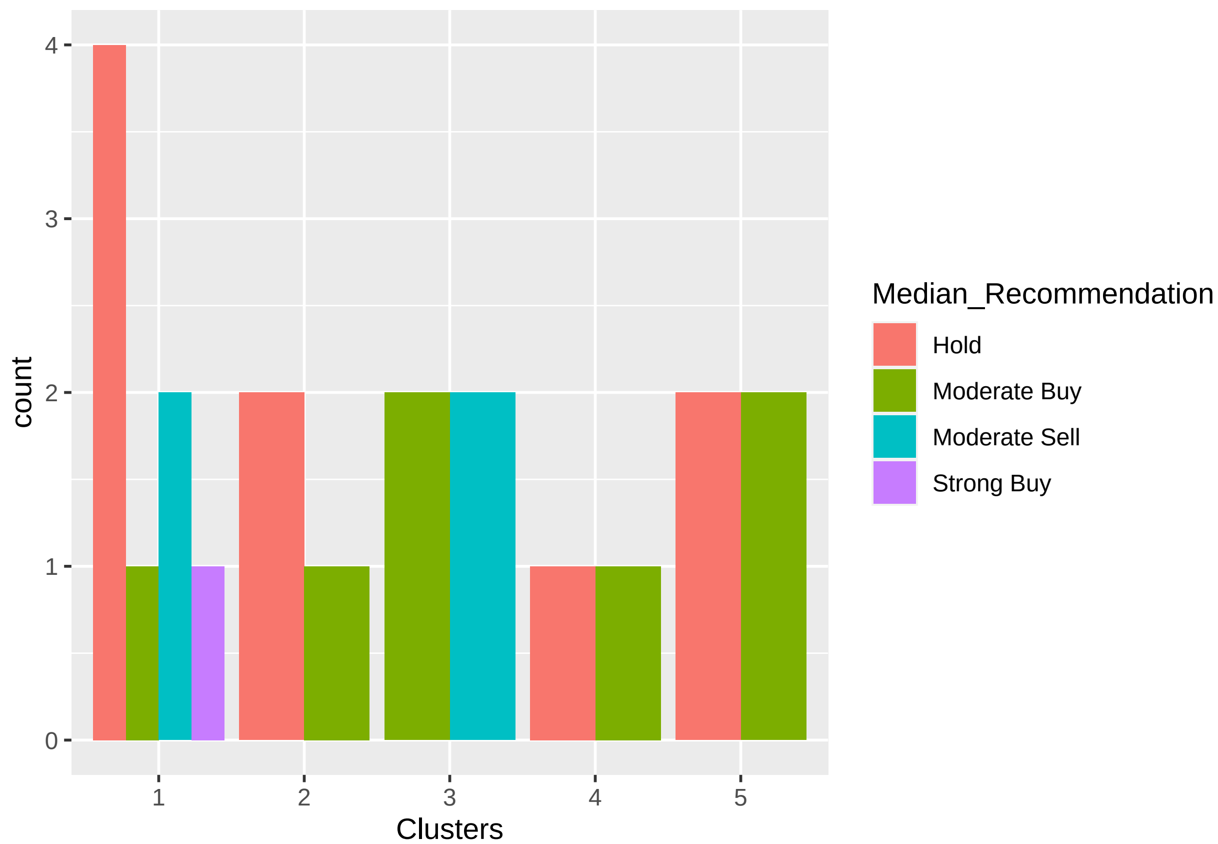
Cluster_4=AGN,PHA

Cluster_5=GSK,JNJ,MRK,PFE

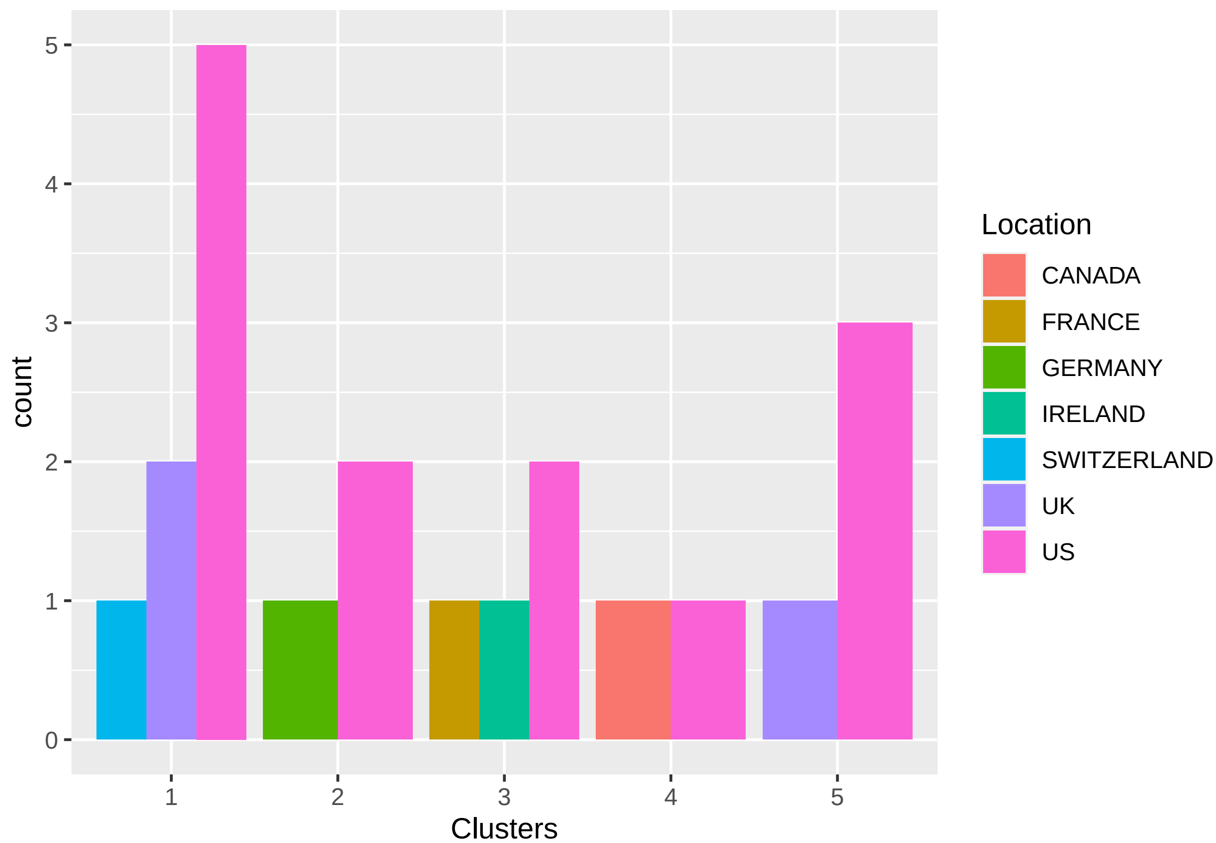
From the clusters formed it can be understood that

1. Cluster_1 has group of companies with moderate return on equity and return on investment
2. Cluster_2 contains companies with very bad ROA,ROE, market capitalization and asset turnover. this implies that these companies are very risky
3. Cluster_3 has group companies similar to cluster_2 but with little less risk involved
4. Cluster_4 companies has very good PE_ratio but very poor ROA,ROE which is more riskier that cluster_2
5. Cluster_5 has companies with very good market capitalization, ROE and ROA

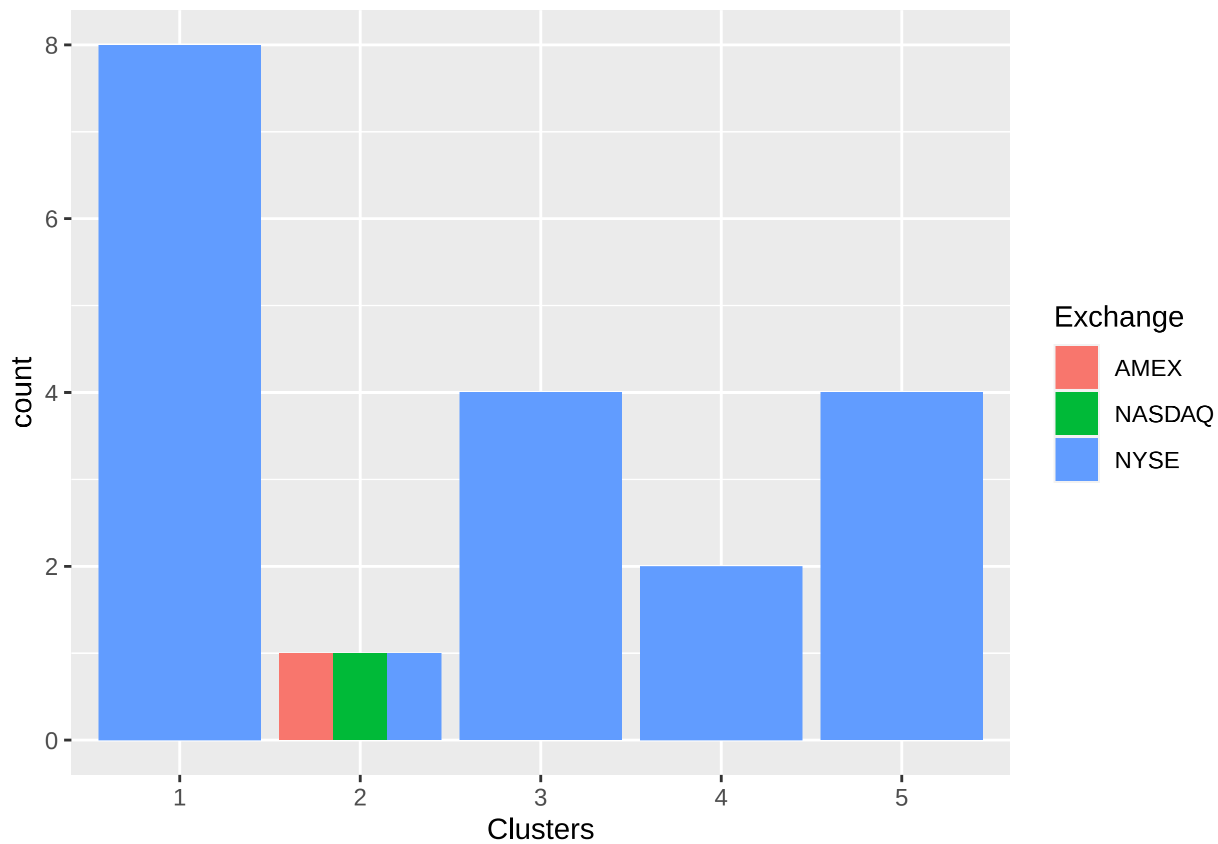
```
Clustering_datase_2<- pharmaceutical_data[,12:14] %>% mutate(Clusters=kmeans_5$cluster)
ggplot(Clustering_datase_2, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(posi
```




```
ggplot(Clustering_datase_2, mapping = aes(factor(Clusters),fill = Location))+geom_bar(position = 'dodge
```




```
ggplot(Clustering_datase_2, mapping = aes(factor(Clusters),fill = Exchange))+geom_bar(position = 'dodge
```



It can be seen that there is a pattern in clusters and the variable Median Recommendation. Like the 2nd cluster suggests between hold and moderate buy,3rd cluster suggests to moderate buy to moderate sell. From the location graph it can be noticed that most of the pharmaceutical companies are US based and there is no much pattern in it. There is no noticeable pattern between clusters and exchange except the fact that majority of companies are listed on NYSE.

Naming clusters:

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

Cluster 1: Large-Thousands

Cluster 2: Extra Small-Penny

Cluster 3: Small- Dollars

Cluster 4: Medium-Hundreds

Cluster 5: Extra Large-Millions