# Assignment_4

Rishitha Reddy Muddasani

2023-03-18

Loading the data set and the Libraries

```
library(flexclust)
```

```
## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4
```

```
library(cluster)
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   1.2.1      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
```

```
library(FactoMineR)
library(tinytex)
library(ggcorrplot)

P_Data<-read.csv("~/Downloads/Pharmaceuticals.csv")
P_Data<-na.omit(P_Data)
```

*TASK 1* The 21 firms are grouped using the numerical variables (1–9).

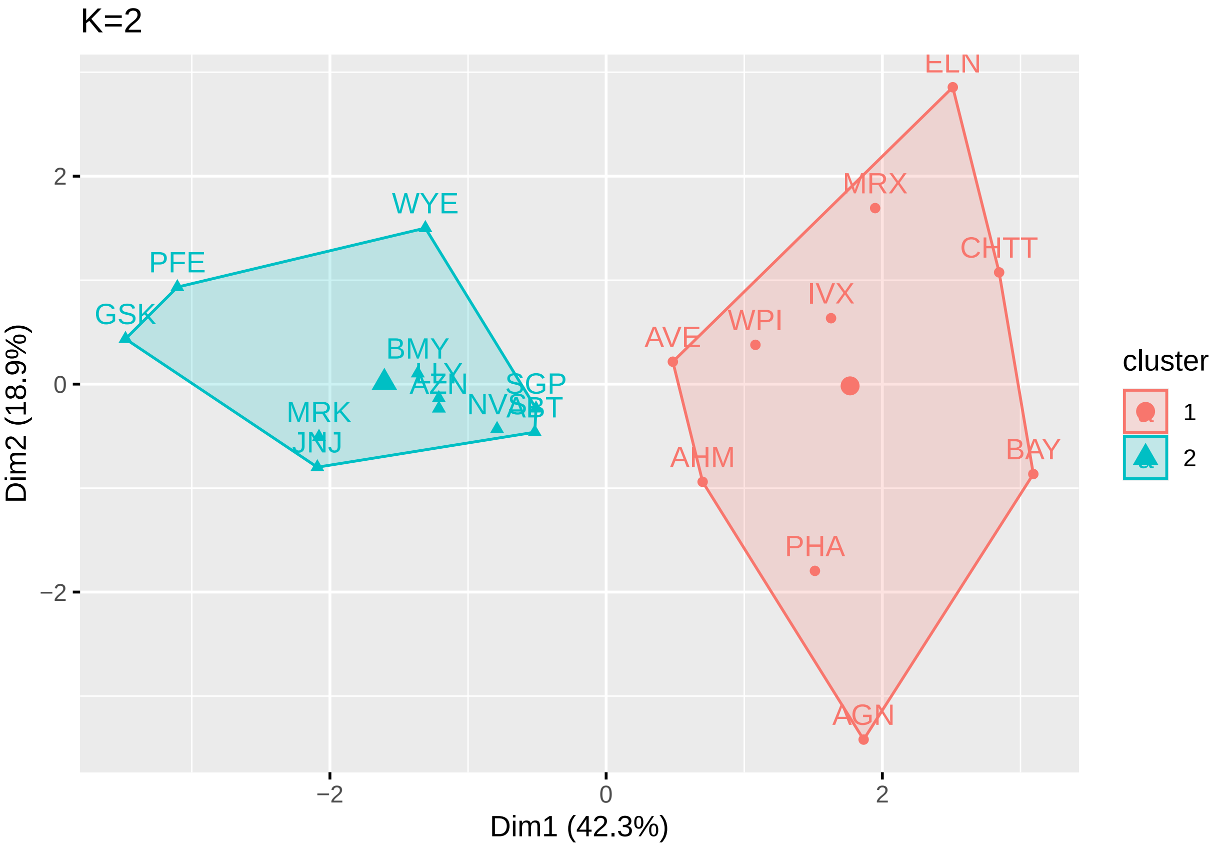
```
row.names(P_Data)<-P_Data[,1]
Clustering_dataset<-P_Data[,3:11]
```

data scalability

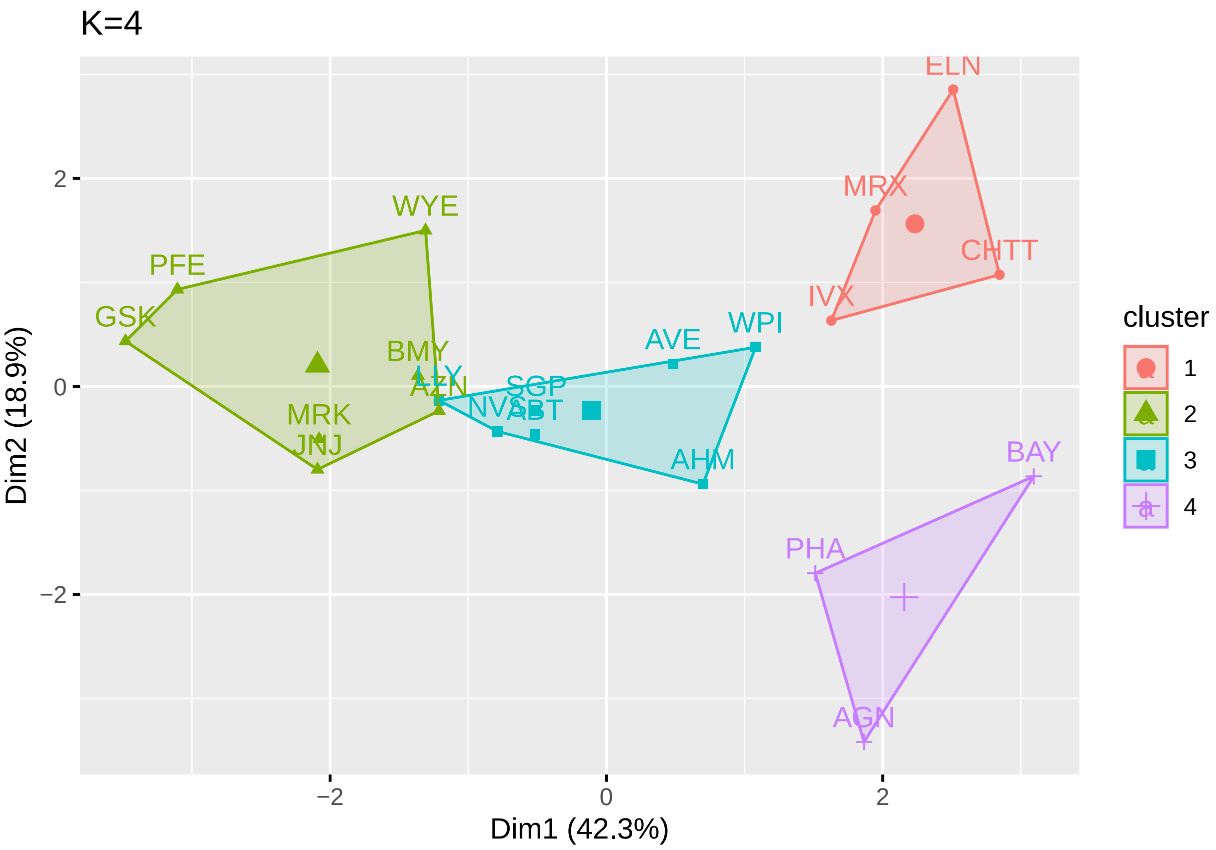
```
set.seed(143)
Scaled_data<-scale(Clustering_dataset)
```

Kmeans computation using random K values

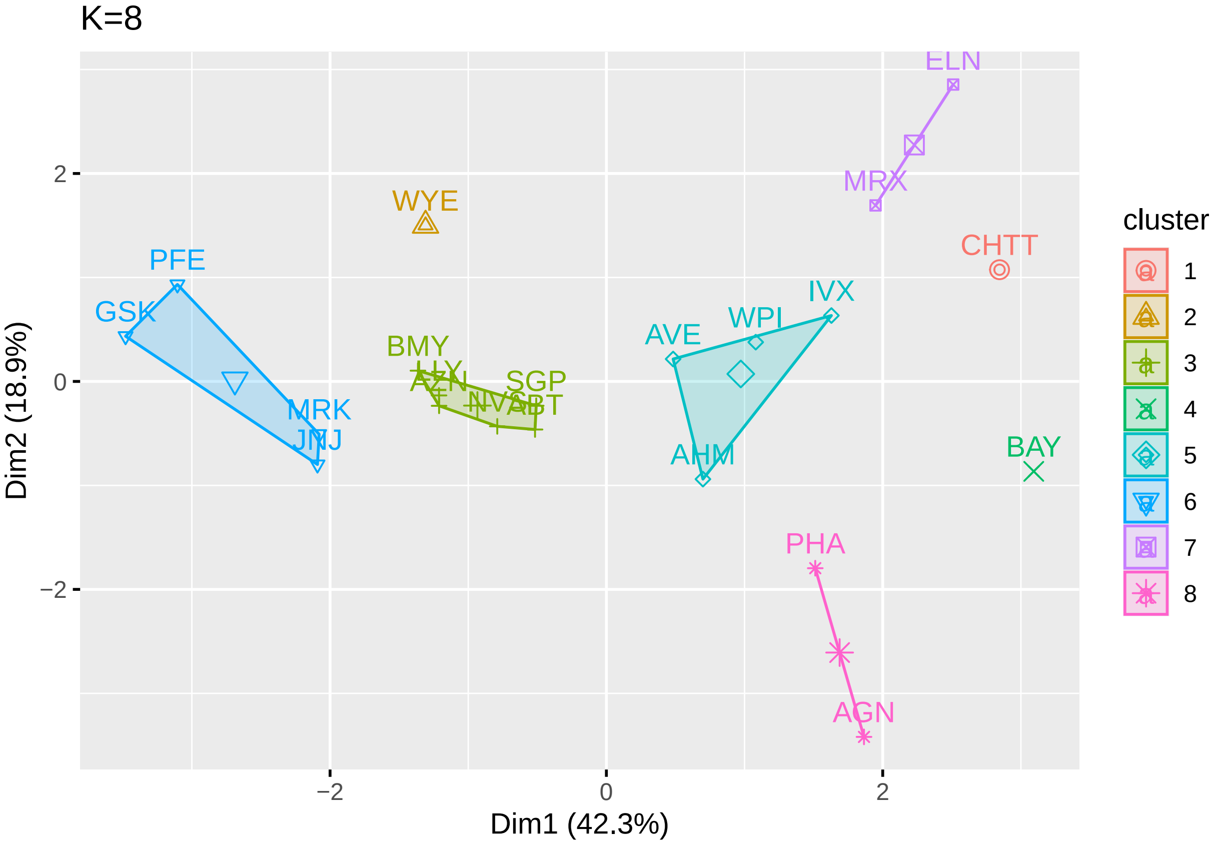
```r
set.seed(143)
kmeans_2_centers<-kmeans(Scaled_data,centers = 2, nstart = 15)
kmeans_4_centers<-kmeans(Scaled_data,centers = 4, nstart = 15)
kmeans_8_centers<-kmeans(Scaled_data,centers = 8, nstart = 15)
plot_kmeans_2_centers<-fviz_cluster(kmeans_2_centers,data = Scaled_data) + ggtitle("K=2")
plot_kmeans_4_centers<-fviz_cluster(kmeans_4_centers,data = Scaled_data) + ggtitle("K=4")
plot_kmeans_8_centers<-fviz_cluster(kmeans_8_centers,data = Scaled_data) + ggtitle("K=8")
plot_kmeans_2_centers
```



```r
plot_kmeans_4_centers
```

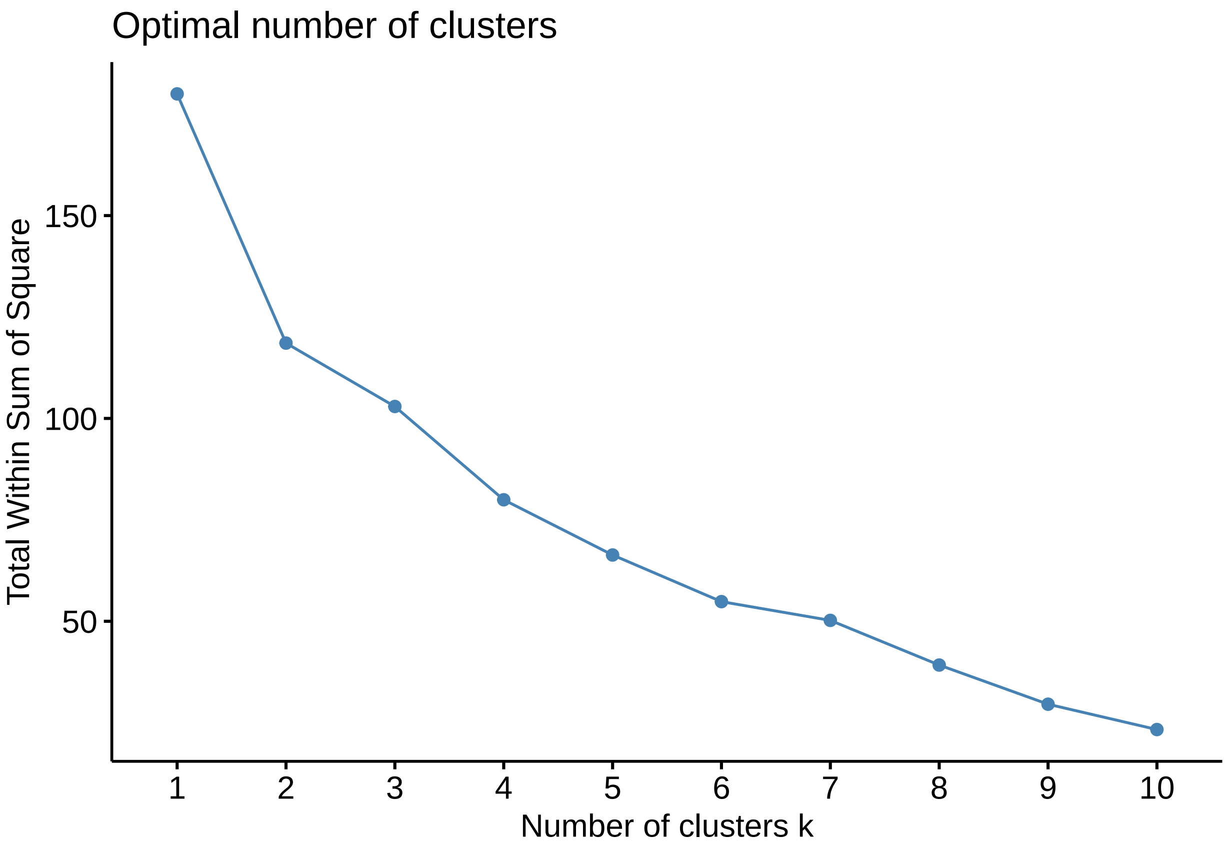


```
plot_kmeans_8_centers
```

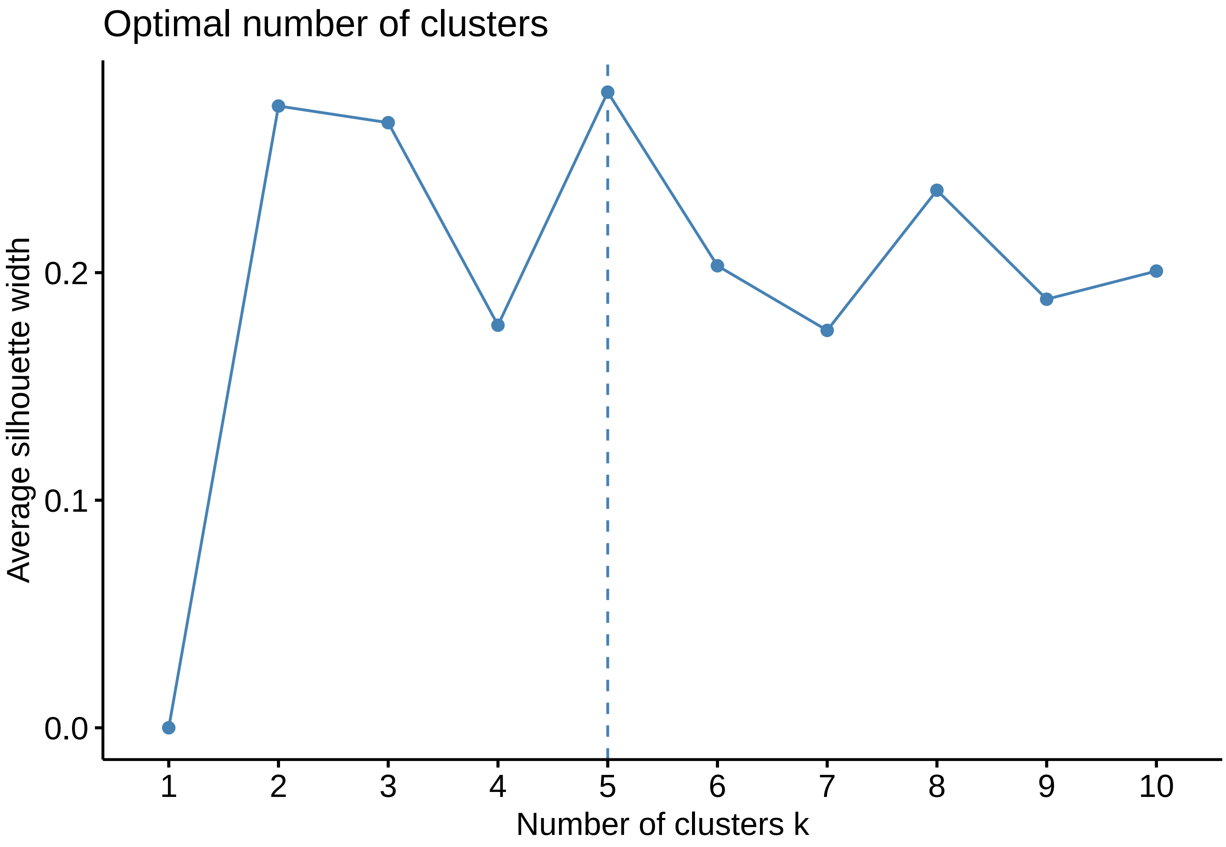


Finding the optimal K appropriate for clustering using WSS and Silhouette

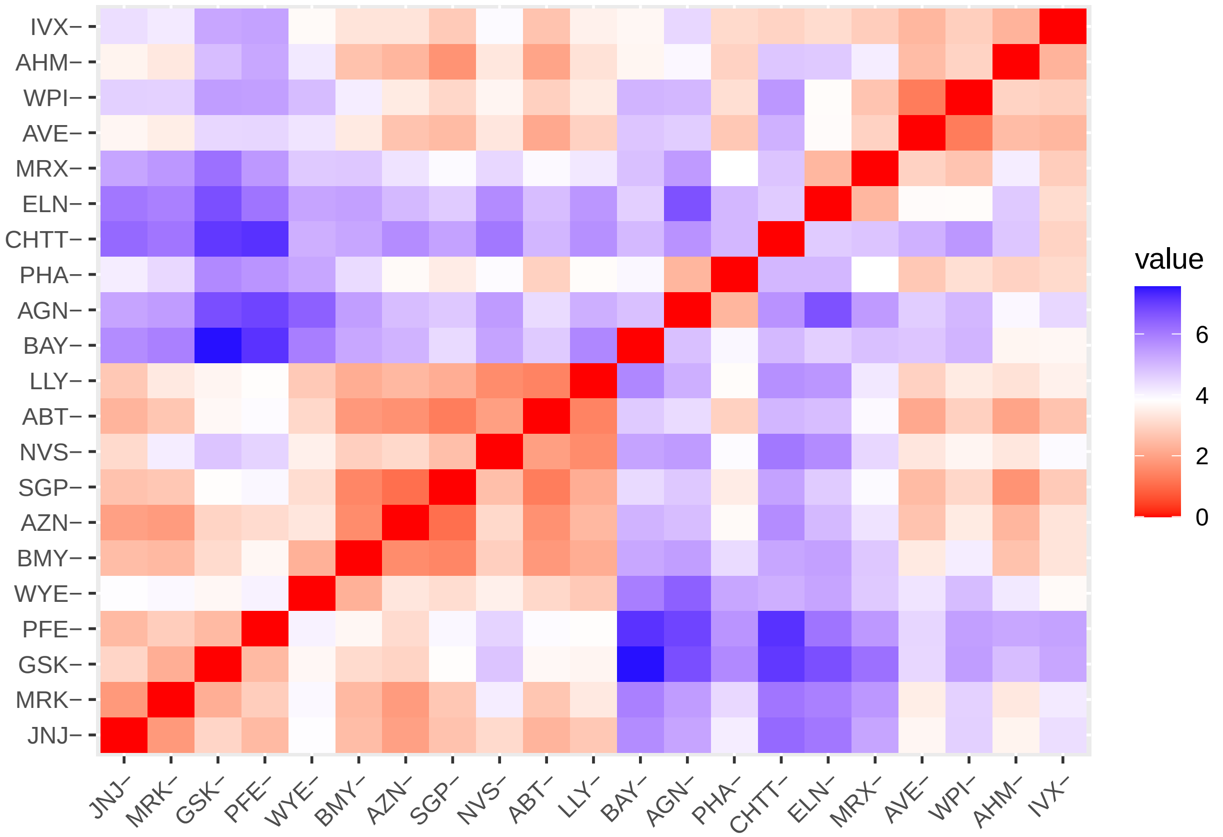
```
wss<-fviz_nbclust(Scaled_data,kmeans,method="wss")
silhouette<-fviz_nbclust(Scaled_data,kmeans,method="silhouette")
wss
```


Optimal number of clusters
Total Within Sum of Square
150
100
50
1
2
3
4
5
6
7
8
9
10
Number of clusters k

```
silhouette
```


Optimal number of clusters
Average silhouette width
0.2
0.1
0.0
1
2
3
4
5
6
7
8
9
10
Number of clusters k

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



k is 2 from WSS and 5 from silhouette. The number 5 ensures that the sum of squires inside each cluster is minimal and that there is considerable spacing between them.

*TASK 2*

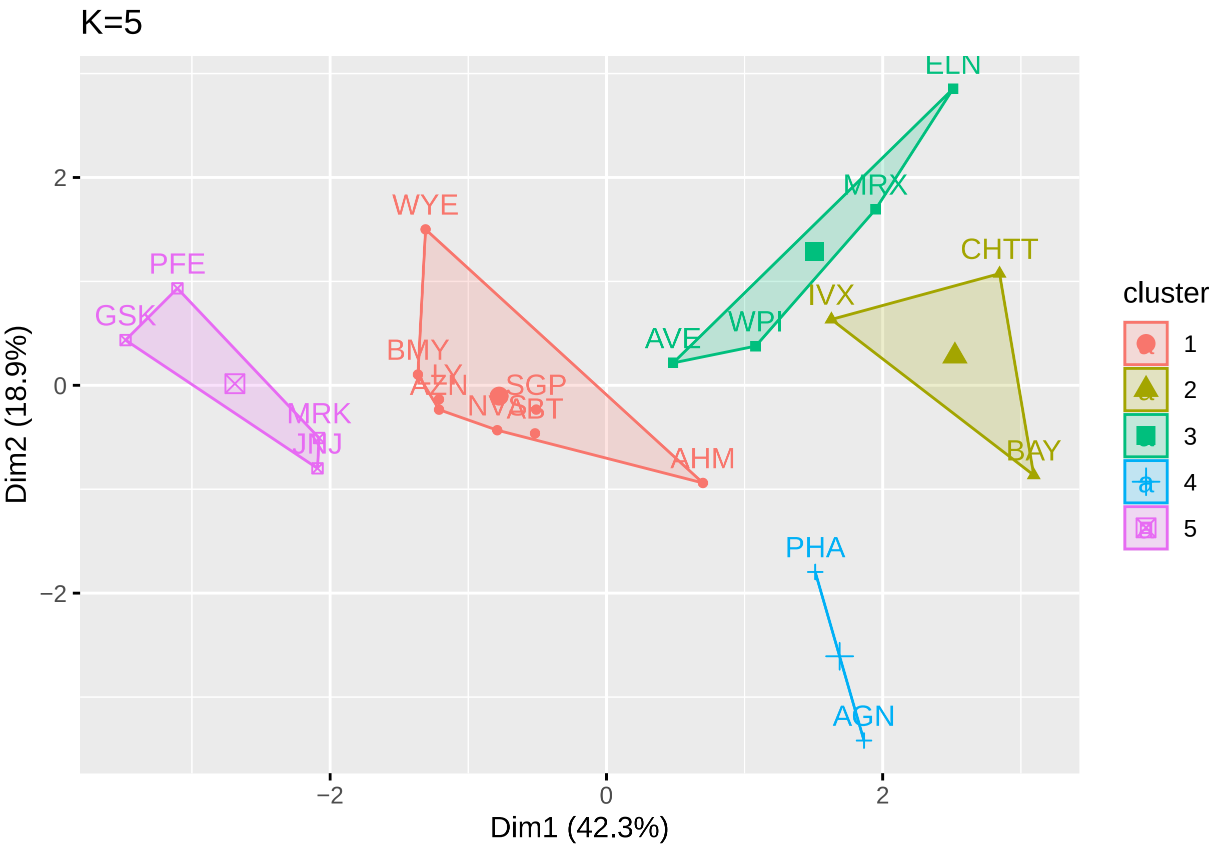
Using Kmeans to find an appropriate k

```
set.seed(143)
kmeans_5_centers<-kmeans(Scaled_data,centers = 5, nstart = 10)
kmeans_5_centers
```

```
## 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_centers<-fviz_cluster(kmeans_5_centers,data = Scaled_data) + ggtitle("K=5")
plot_kmeans_5_centers
```



```
Clustering_dataset_1<-Clustering_dataset%>% mutate(Cluster_no=kmeans_5_centers$cluster)%>% group_by(Clu
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
```

Following clusters have been created for companies:

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

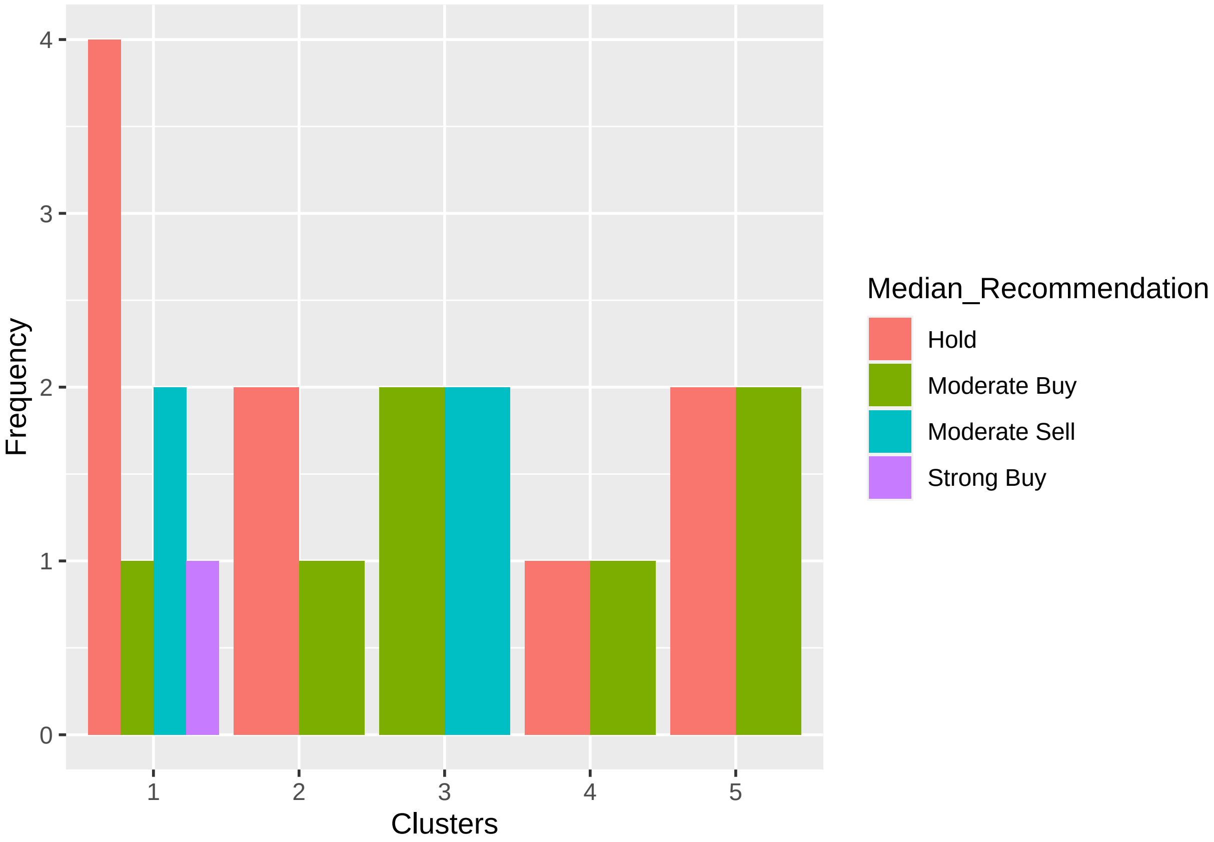
Cluster_5=GSK,JNJ,MRK,PFE

This can be inferred from the clusters that were generated.

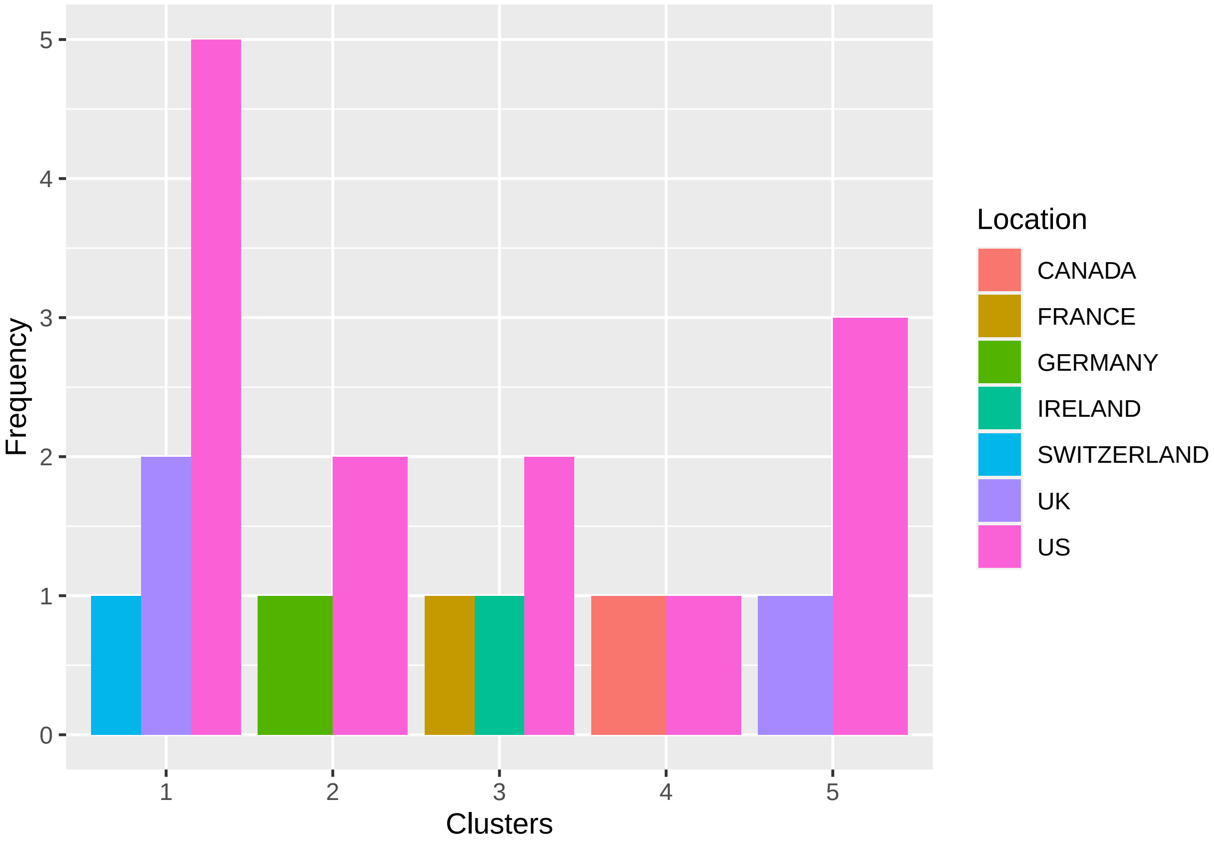
1. Cluster 1 contains a collection of businesses with a modest return on equity and return on investment.
2. Cluster 2 Companies have extremely low ROA, ROE, market capitalization, and asset turnover. This means that these businesses are exceedingly dangerous.
3. Similar to cluster 2, Cluster 3 features group corporations, but with slightly lower risk.
4. Companies in cluster 4 are more risky than those in cluster 2 because they have very good PE ratios but weak ROA and ROE.
5. Companies in Cluster 5 have excellent ROE, ROA, and market capitalization.

*TASK 3*

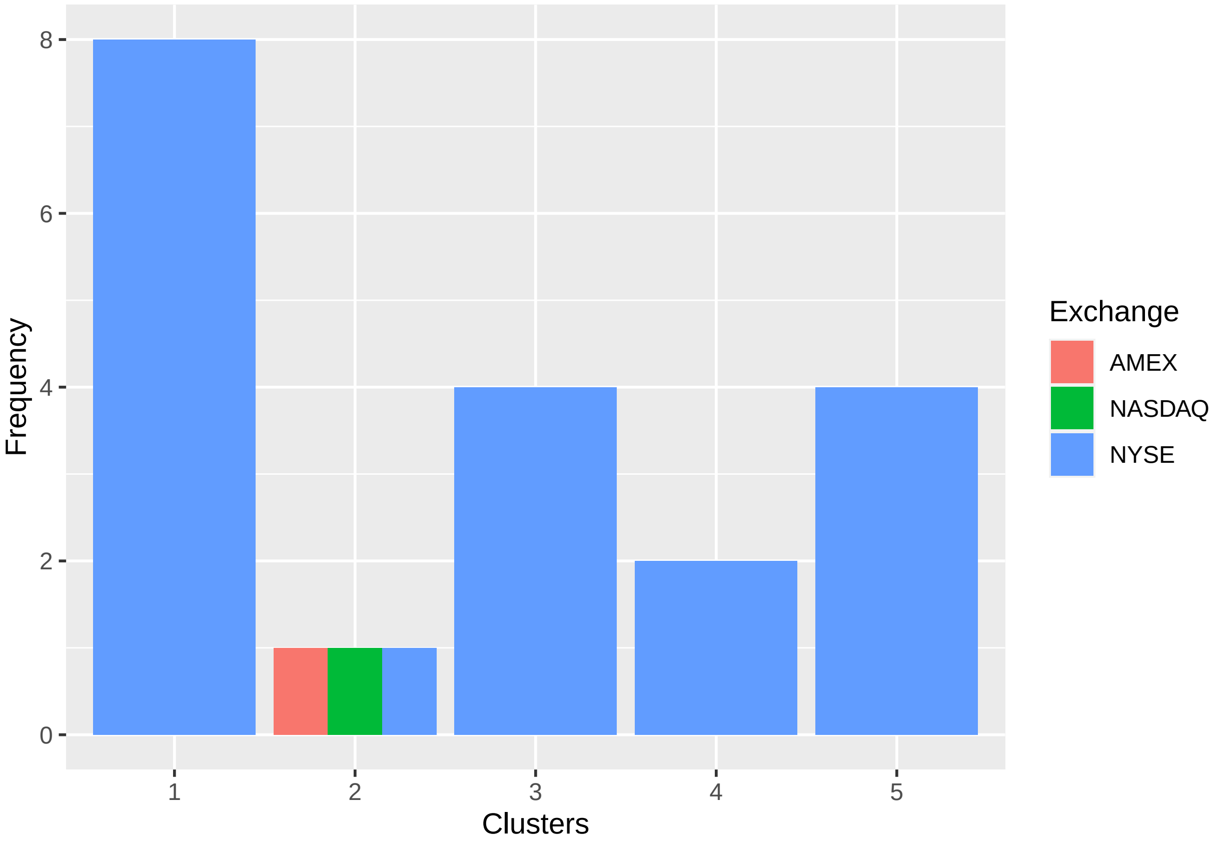
```
#Is there a pattern in the clusters with respect to the numerical
#variables (10 to 12)? (those \n #not used in forming the clusters)
Clustering_datase_2<- P_Data[,12:14] %>% mutate(Clusters=kmeans_5_centers$cluster)
ggplot(Clustering_datase_2, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(posi
```



```
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
```



Clusters and the variable Median Recommendation exhibit a pattern, as can be observed. Similar to what the second cluster shows between moderate buy and hold, the third cluster recommends between moderate purchase and moderate sell. The majority of pharmaceutical businesses are based in the US, as can be seen from the location graph, although there isn't much of a pattern there. With the exception of the bulk of companies being listed on NYSE, there is no discernible relationship between clusters and exchanges.

*TASK 4* - Naming clusters:

[It is done based on the 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