Assignment 4

Bhavika

2024-03-13

```
library(readr)
library(tidyverse)
library(factoextra)
library(ISLR)
library(caret)
library(cluster)
```

```
pharmaceuticals_data<-read_csv("Pharmaceuticals.csv")
head(pharmaceuticals_data)
```

```
## # A tibble: 6 x 14
##   Symbol Name      Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##   <chr>  <chr>          <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
## 1 ABT    Abbott L~      68.4   0.32     24.7  26.4  11.8            0.7     0.42
## 2 AGN    Allergan~       7.58  0.41     82.5  12.9   5.5            0.9     0.6
## 3 AHM    Amersham~       6.3   0.46     20.7  14.9   7.8            0.9     0.27
## 4 AZN    AstraZen~      67.6   0.52     21.5  27.4  15.4            0.9     0
## 5 AVE    Aventis        47.2   0.32     20.1  21.8   7.5            0.6     0.34
## 6 BAY    Bayer AG       16.9   1.11     27.9   3.9   1.4            0.6     0
## # i 5 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>,
## #   Median_Recommendation <chr>, Location <chr>, Exchange <chr>
```

```
#Print Number of Columns
ncol(pharmaceuticals_data)
```

```
## [1] 14
```

```
#print number of rows
nrow(pharmaceuticals_data)
```

```
## [1] 21
```

```
#1.identify the numerical variables (1 to 9) to cluster the 21 firms
numerical_vars <- pharmaceuticals_data[, 3:11]
head(numerical_vars,11)
```

```
## # A tibble: 11 x 9
##    Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage Rev_Growth
##         <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>      <dbl>
```

```
## 1       68.4   0.32     24.7  26.4  11.8           0.7     0.42       7.54
## 2        7.58  0.41     82.5  12.9   5.5           0.9     0.6        9.16
## 3        6.3   0.46     20.7  14.9   7.8           0.9     0.27       7.05
## 4       67.6   0.52     21.5  27.4  15.4           0.9     0         15
## 5       47.2   0.32     20.1  21.8   7.5           0.6     0.34      26.8
## 6       16.9   1.11     27.9   3.9   1.4           0.6     0         -3.17
## 7       51.3   0.5      13.9  34.8  15.1           0.9     0.57       2.7
## 8        0.41  0.85     26    24.1   4.3           0.6     3.51       6.38
## 9        0.78  1.08      3.6  15.1   5.1           0.3     1.07      34.2
## 10      73.8   0.18     27.9  31    13.5           0.6     0.53       6.21
## 11     122.    0.35     18    62.9  20.3           1       0.34      21.9
## # i 1 more variable: Net_Profit_Margin <dbl>
```

```r
#Checking the % of missing values in each column

missing_values <- (colMeans(is.na(pharmaceuticals_data))*100)

head(missing_values,11)
```

```
##            Symbol              Name        Market_Cap              Beta
##                 0                 0                 0                 0
##          PE_Ratio               ROE               ROA    Asset_Turnover
##                 0                 0                 0                 0
##          Leverage        Rev_Growth Net_Profit_Margin
##                 0                 0                 0
```

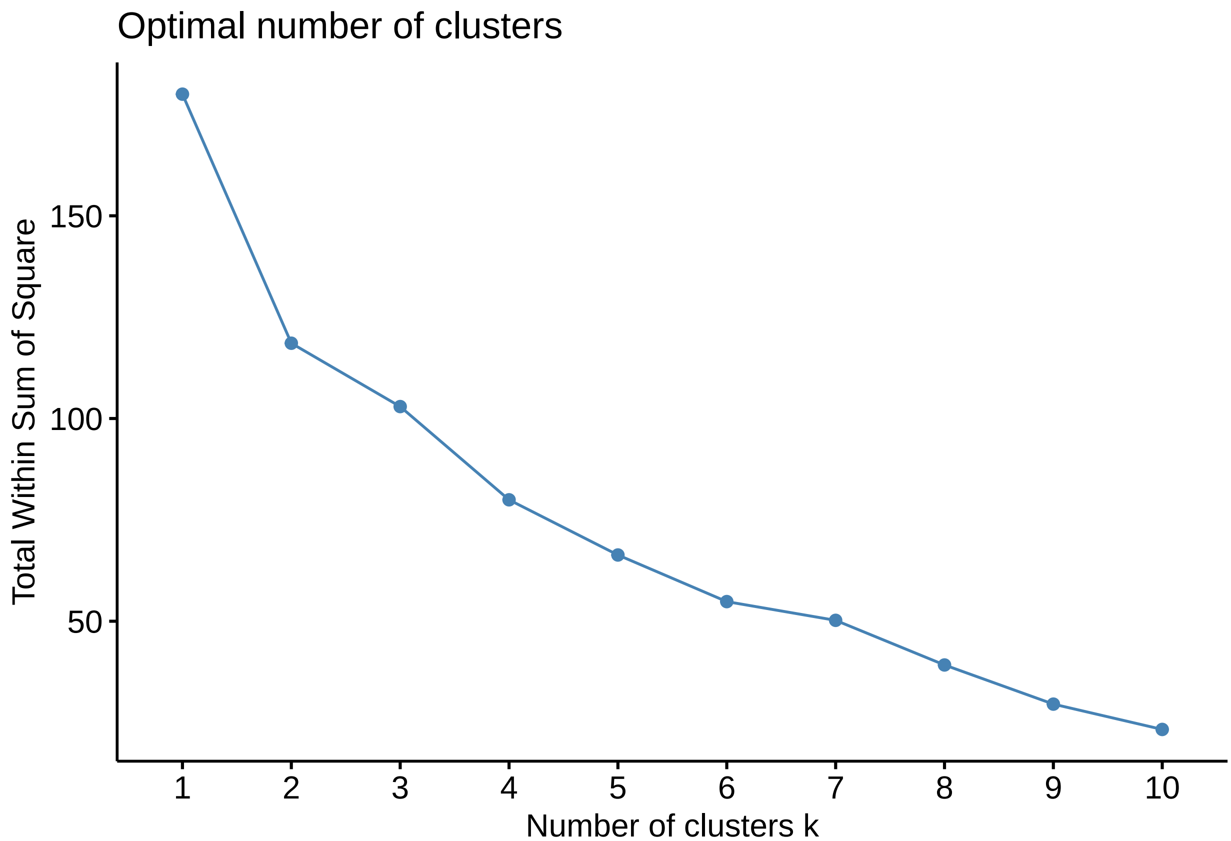
There are no missing values in the data.

#Normalizing the Data:

```r
preprocess_data <- preProcess(numerical_vars, method = c("center", "scale"))
normalized_data <- predict(preprocess_data, numerical_vars)
```

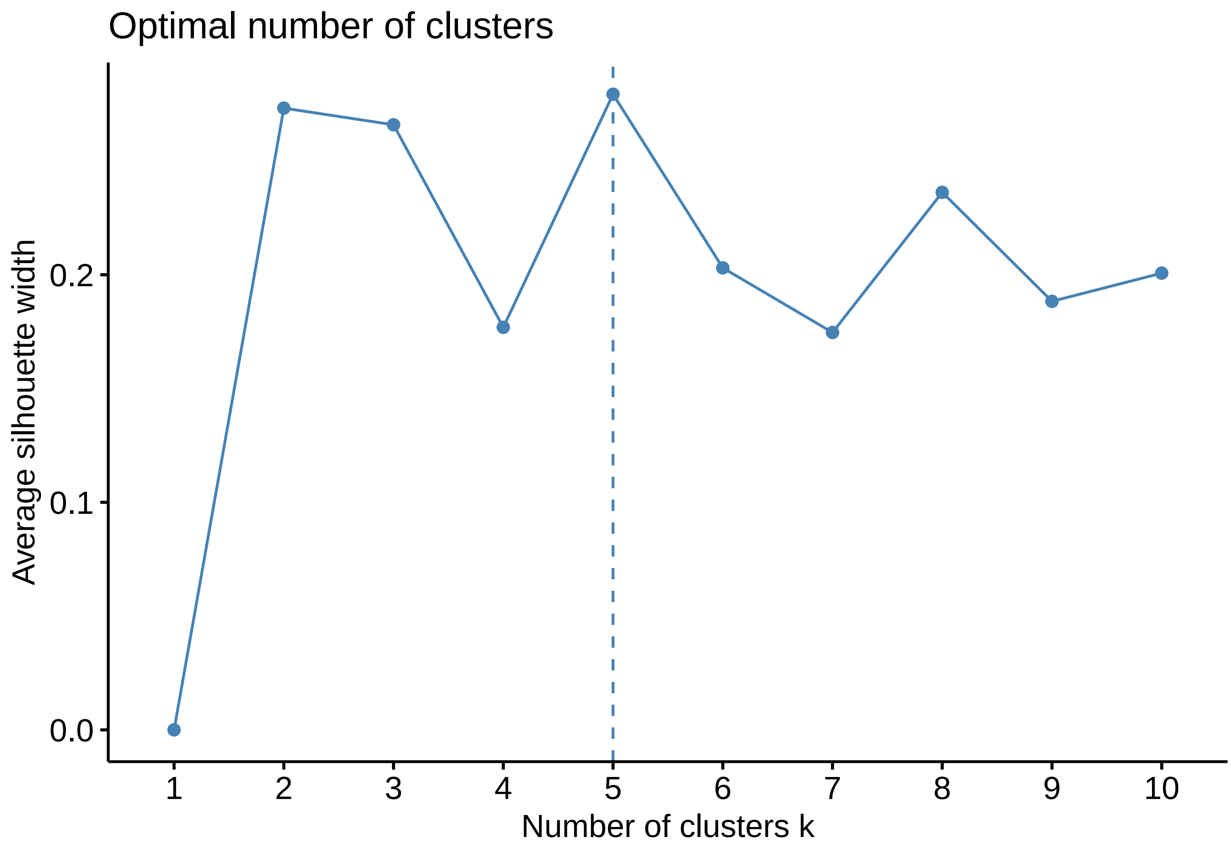
Checking the optimal number of Clusters using Elbow method and Silhouette Method:

```r
set.seed(7895)
fviz_nbclust(normalized_data,kmeans,method="wss")
```



In elbow method, Optimal value of k is 2.

```
fviz_nbclust(normalized_data,kmeans,method="silhouette")
```



In Silhouette Method, Optimal Value of k is 5.

As the optimal number of clusters obtained from both Elbow method and Silhouette method is different, we will run the knn model using both K values and based on the formation of clusters, we will decide which optimal K value is to considered for further analysis.

#Applying k-Means clustering:

```
k2<-kmeans(normalized_data,centers=2)
k2
```

```
## K-means clustering with 2 clusters of sizes 11, 10
##
## Cluster means:
##    Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656
## 2 -0.7407208  0.3945061  0.3039863 -0.7222576 -0.9178575     -0.5073922
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.3331068 -0.2902163         0.6823310
## 2  0.3664175  0.3192379        -0.7505641
##
## Clustering vector:
##  [1] 1 2 2 1 2 2 1 2 2 1 1 2 1 2 1 1 1 2 1 2 1
##
## Within cluster sum of squares by cluster:
## [1] 43.30886 75.26049
##  (between_SS / total_SS =  34.1 %)
```

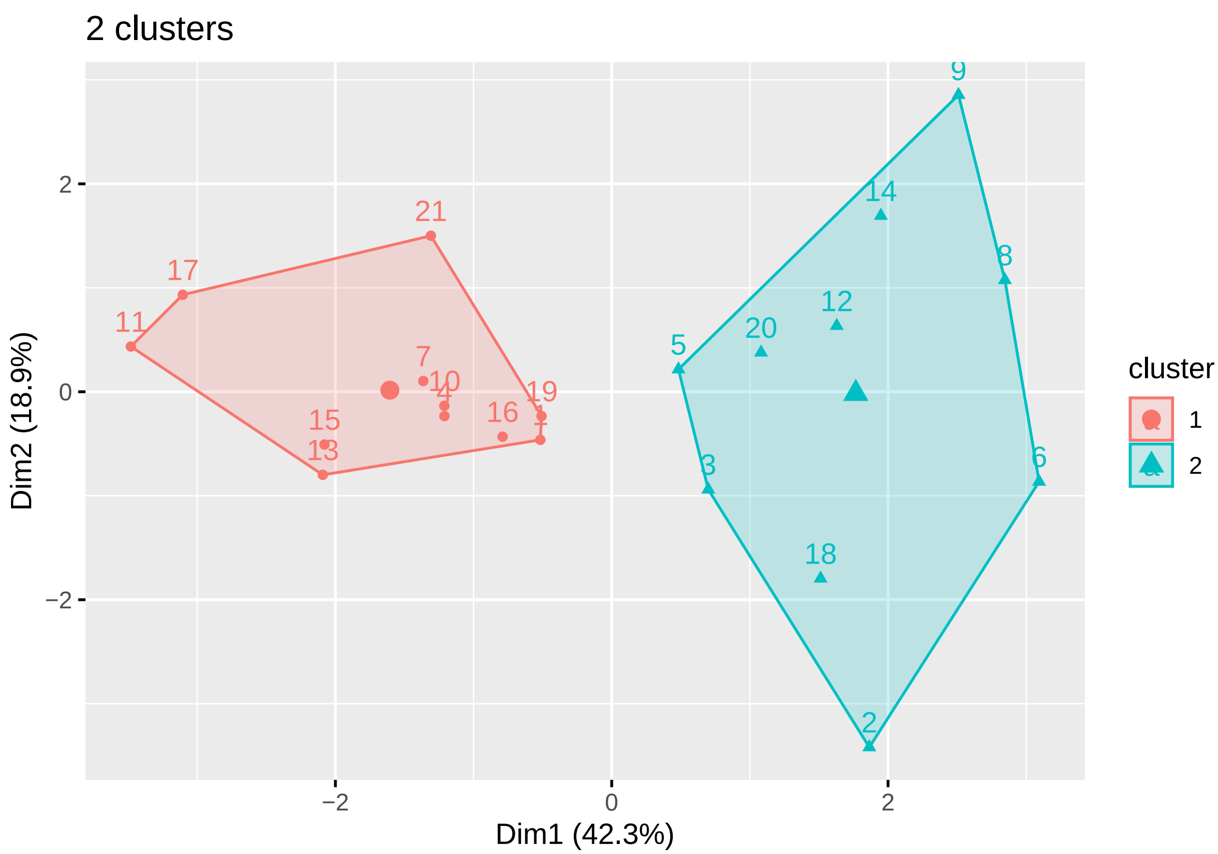
```
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
k5<-kmeans(normalized_data,centers=5)
k5
```

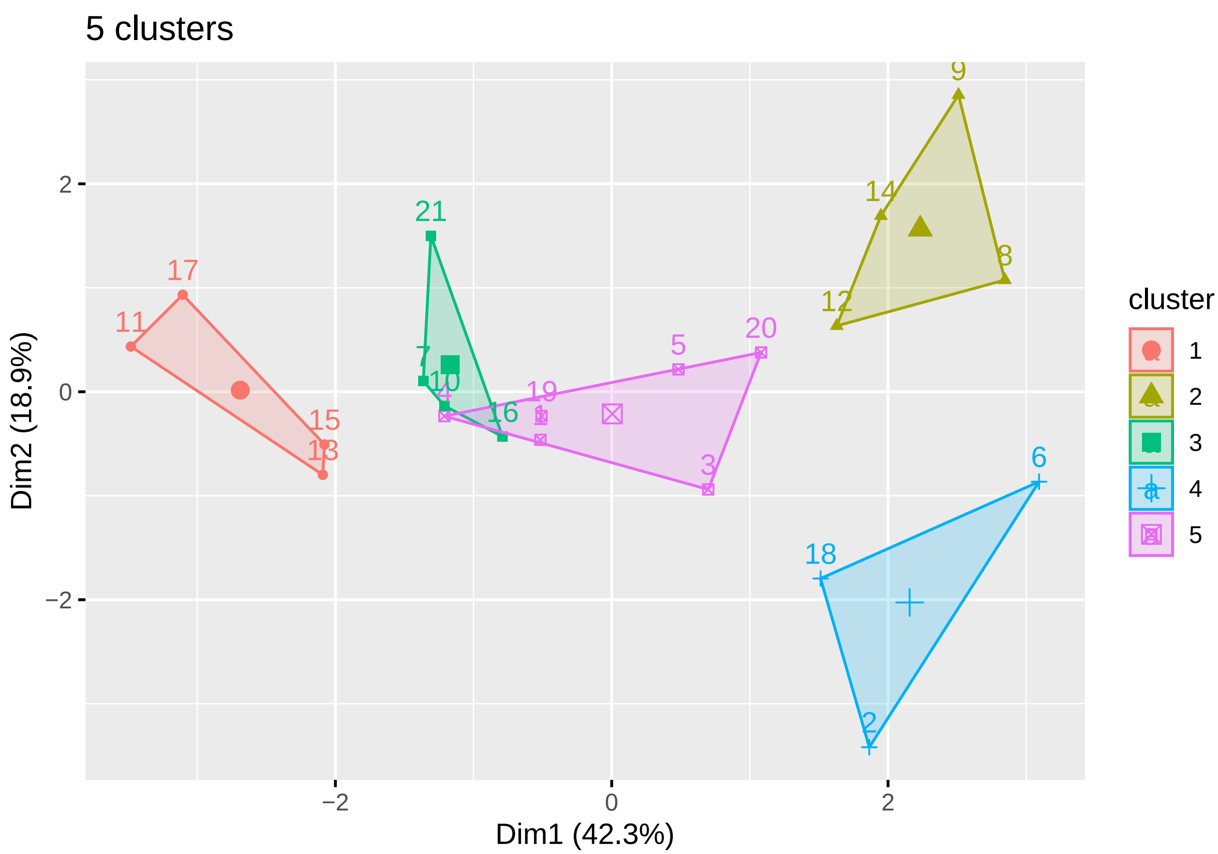
```
## K-means clustering with 5 clusters of sizes 4, 4, 4, 3, 6
##
## Cluster means:
##    Market_Cap       Beta   PE_Ratio        ROE         ROA Asset_Turnover
## 1  1.6955811 -0.1780563 -0.1984582  1.2349879  1.35034311   1.153164e+00
## 2 -0.9624758  1.1949250 -0.3639982 -0.5200697 -0.96107919  -1.153164e+00
## 3  0.1680985 -0.5870295 -0.3885227  0.5869921  0.52349286  -2.306328e-01
## 4 -0.5246281  0.4451409  1.8498439 -1.0404550 -1.18658381   1.480297e-16
## 5 -0.3384885 -0.5091299 -0.2909358 -0.3477127 -0.01521261   1.537552e-01
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.46807818  0.4671788        0.59124252
## 2  1.47737177  0.7120120       -0.36882358
## 3 -0.02011273 -1.0613321        1.10937343
## 4 -0.34435439 -0.5769454       -1.60954392
## 5 -0.48727670  0.2099002       -0.08308962
##
## Clustering vector:
##  [1] 5 4 5 5 5 4 3 2 2 3 1 2 1 2 1 3 1 4 5 5 3
##
## Within cluster sum of squares by cluster:
## [1]  9.284424 19.219788 10.157927 14.938904 13.562315
##  (between_SS / total_SS =  62.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

#Plotting the clusters:

```
fviz_cluster(k2,pharmaceuticals_data[,(3:11)],main="2 clusters")
```




```
fviz_cluster(k5,pharmaceuticals_data[,(3:11)],main="5 clusters")
```



When clusters with k=5 is plotted, clusters are overlapping.

On the other hand, the 2 clusters formed in the first plot are away from each other and also has divided all 21 firms into 2 groups. Hence, considering k=2 as the optimal number of clusters.

```
#Assigning the cluster to each firm using CBIND
New_data<-cbind(numerical_vars,k2$cluster)
View(New_data)
```

Finding Mean within each cluster to interpret the clusters:

```
mean_k2 <- numerical_vars %>% mutate(Cluster = k2$cluster) %>% group_by(Cluster) %>% summarise_all("mean
mean_k2
```

```
## # A tibble: 2 x 10
##   Cluster Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##     <int>      <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
## 1       1       97.1 0.434     21.0  35.7 15.0            0.8     0.325
## 2       2       14.2 0.627     30.4  14.9  5.63           0.59    0.872
## # i 2 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>
```

Question 2: Interpreting the clusters as per the numerical variables:

Based on the two clusters formed, Market cap of companies in the first cluster is ranging between 34 billion dollars to 199 billion dollars, whereas companies in cluster 2 has an average market cap of 14 billion dollars. This indicates that the companies in first cluster are well-established, and according to market cap, it will be safer to invest in cluster 1 companies.

When PE ratio of companies in both clusters are analyzed, cluster 1 has a better PE ratio with an average of 20.95 compared with an average of 30.42 of companies in second cluster.

When Return of Equity (ROE) and Return on Assets (ROA), companies in cluster 1 has better averages then companies in second cluster.

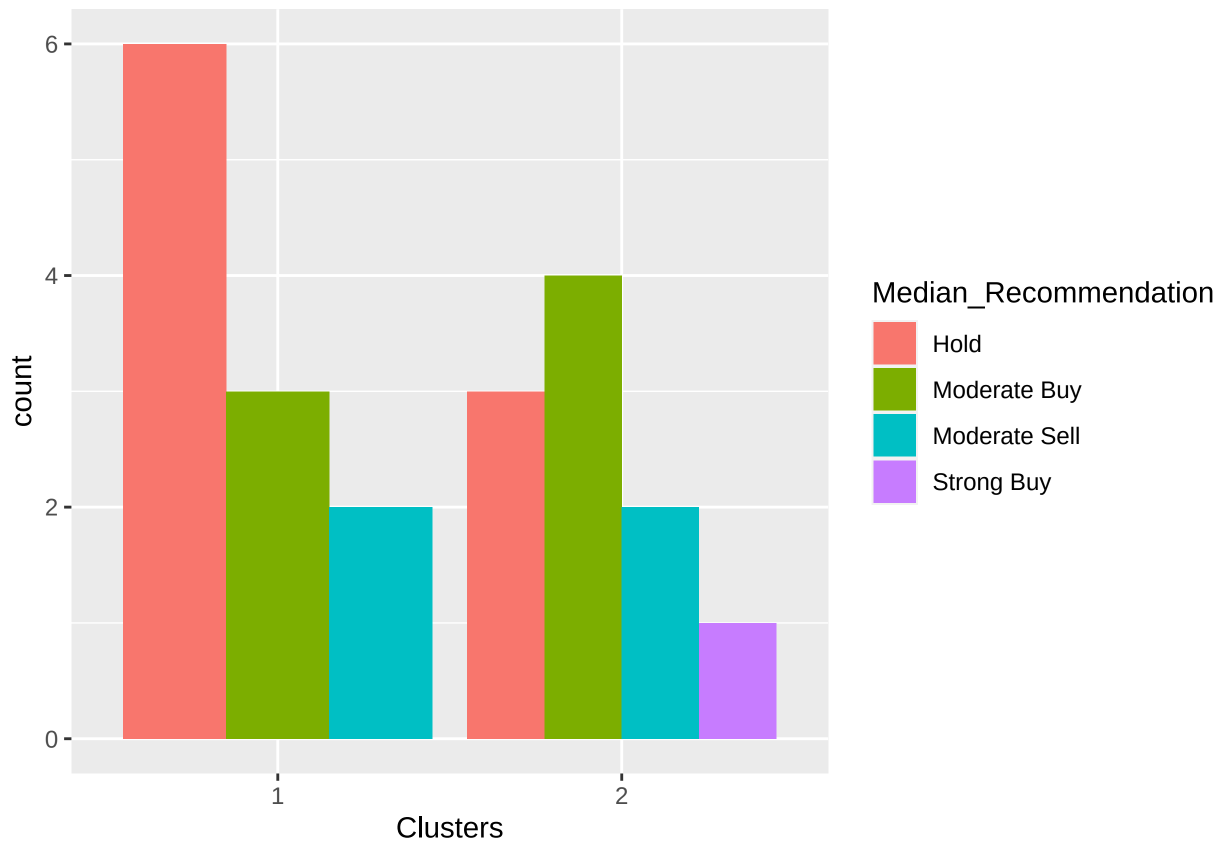
Surprisingly, companies in second cluster has better average of Revenue growth in comparision to companies in first luster. This could be possible as companies with small market cap seems to grow faster.

Net Profit Margin of companies in cluster 1 is twice than compared to cluster 2. This indicates that the first cluster companies are more profitable and has succesful businesses than second cluster companies.

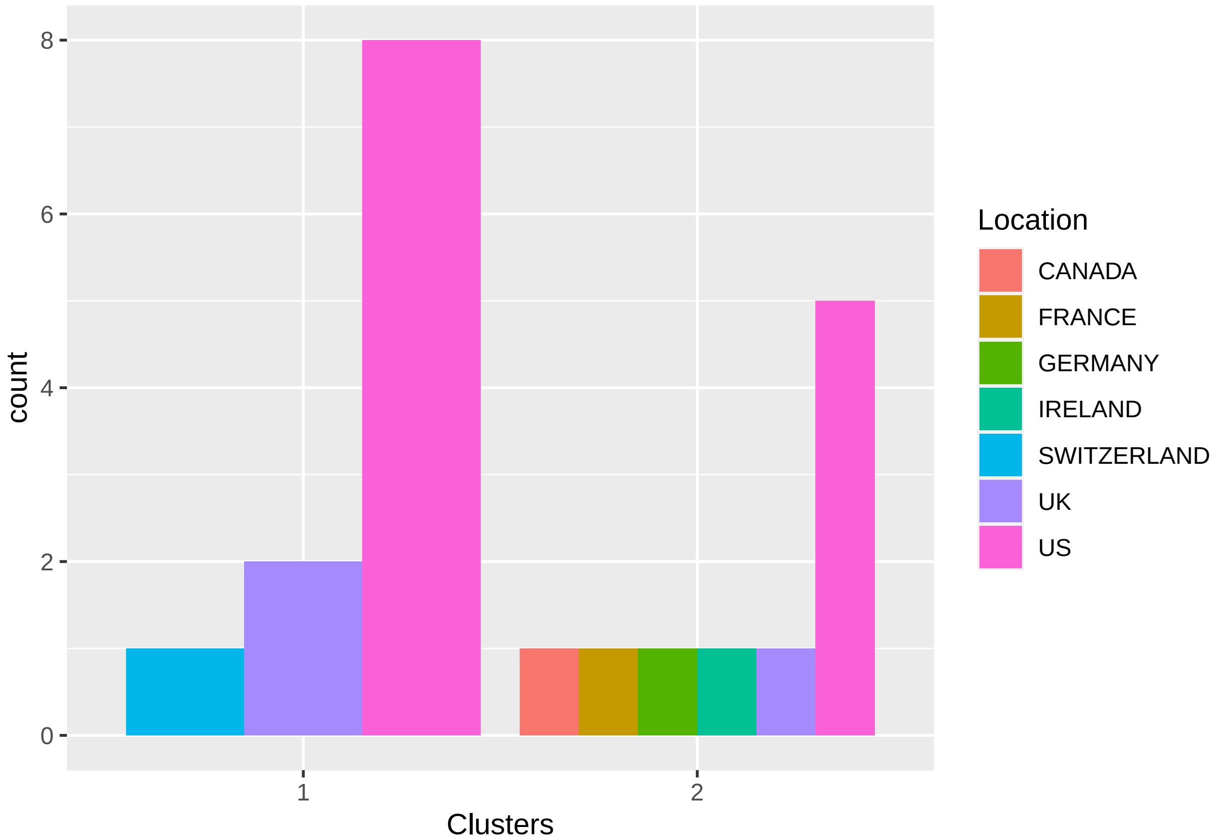
In overall comparision, I would recommend to invest in first cluster as those companies have bigger market cap, and better PE_RATIO,ROE,ROA,Asset Turnover and Net Profit Margin.

Question-3: Finding if there is a pattern with respect to categorical and Numerical variables:

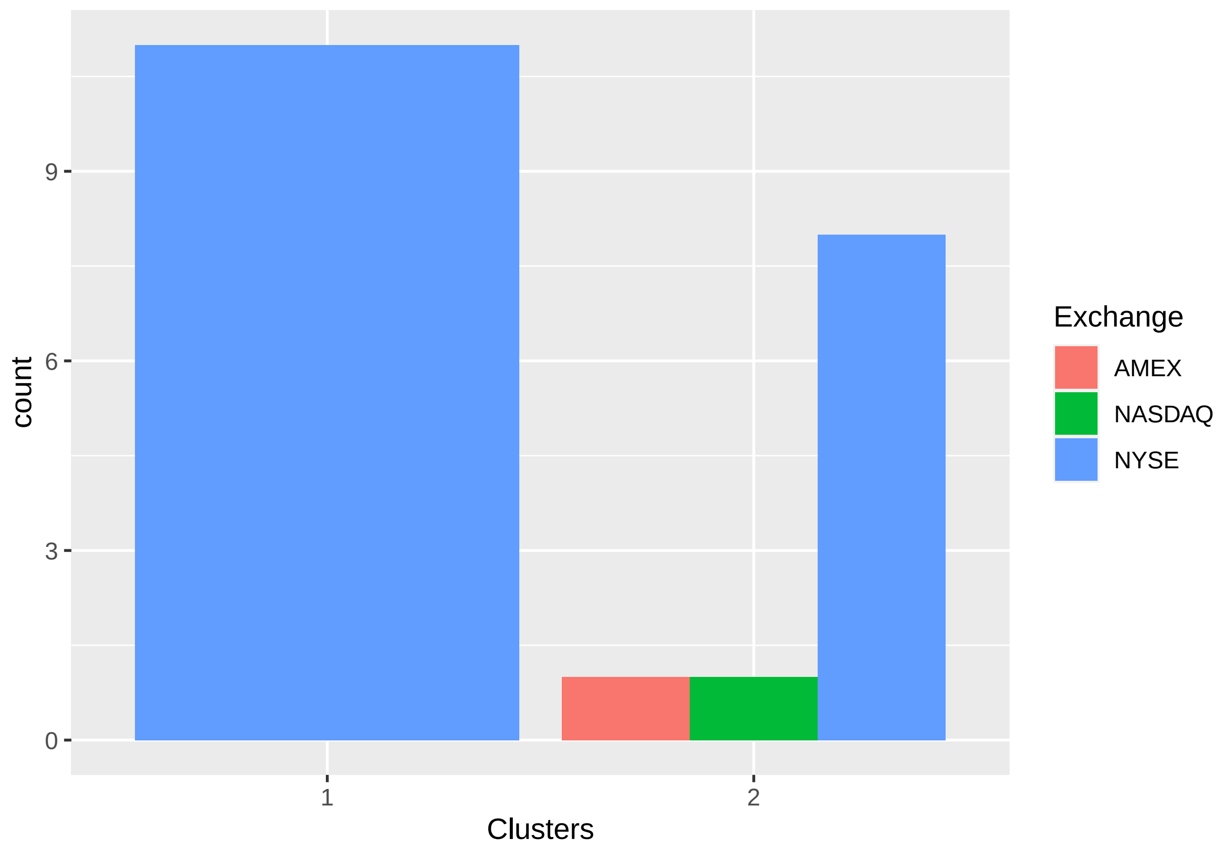
```
plot <- pharmaceuticals_data[12:14] %>% mutate(Clusters=k2$cluster)


library(ggplot2)

library(esquisse)

ggplot(plot, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position='dodge')+la
```



```
ggplot(plot, mapping = aes(factor(Clusters),fill = Location))+geom_bar(position = 'dodge')+labs(x ='Clu
```




```
ggplot(plot, mapping = aes(factor(Clusters),fill = Exchange))+geom_bar(position = 'dodge')+labs(x ='Clu
```



Analysis based on the plots of categorical variables

Median Recommendation: It can be observed from above plots that majority of the firms in cluster 1 are under "Hold" recommendation whereas firms in cluster 2 are under "Modern buy" recommendation

3 firms in cluster 1 are in Moderate buy recommendation and 2 firms are in moderate sell recommendation.

In cluster 2, 3 firms are in Hold, 2 are in Moderate sell and 1 is in strong buy.

Location: Highest number of firms in both the clusters are from the US

Exchange: Majority of the firms in both clusters are listed under NYSE, in fact, all firms in cluster 1 are listed under NYSE

QUESTION 4: Names of the Clusters:

Cluster 1: Low Risk Investments.

Cluster 2: High Risk Investments.