# K-means Clustering

***Loading required libraries:***

```
library(factoextra)
library(dplyr)
library(ggcorrplot)
```

***Reading a dataset:***

```
Pharmaceuticals<-read.csv("Pharmaceuticals.csv")
```

***Selecting Numerical values to cluster 21 firms:***

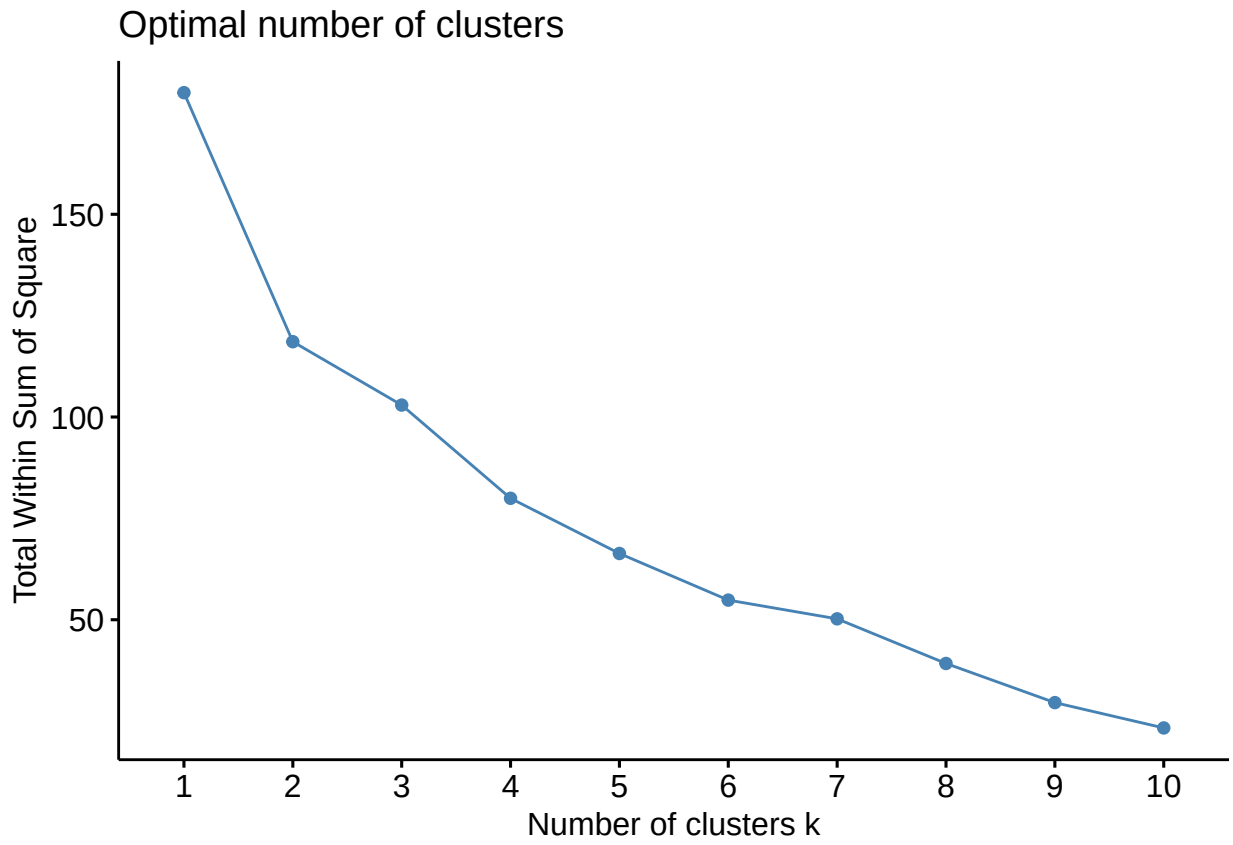
```
data1<-Pharmaceuticals[,(3:11)]
```

***Normalizing data:***

```
norm_data1<-scale(data1)
```

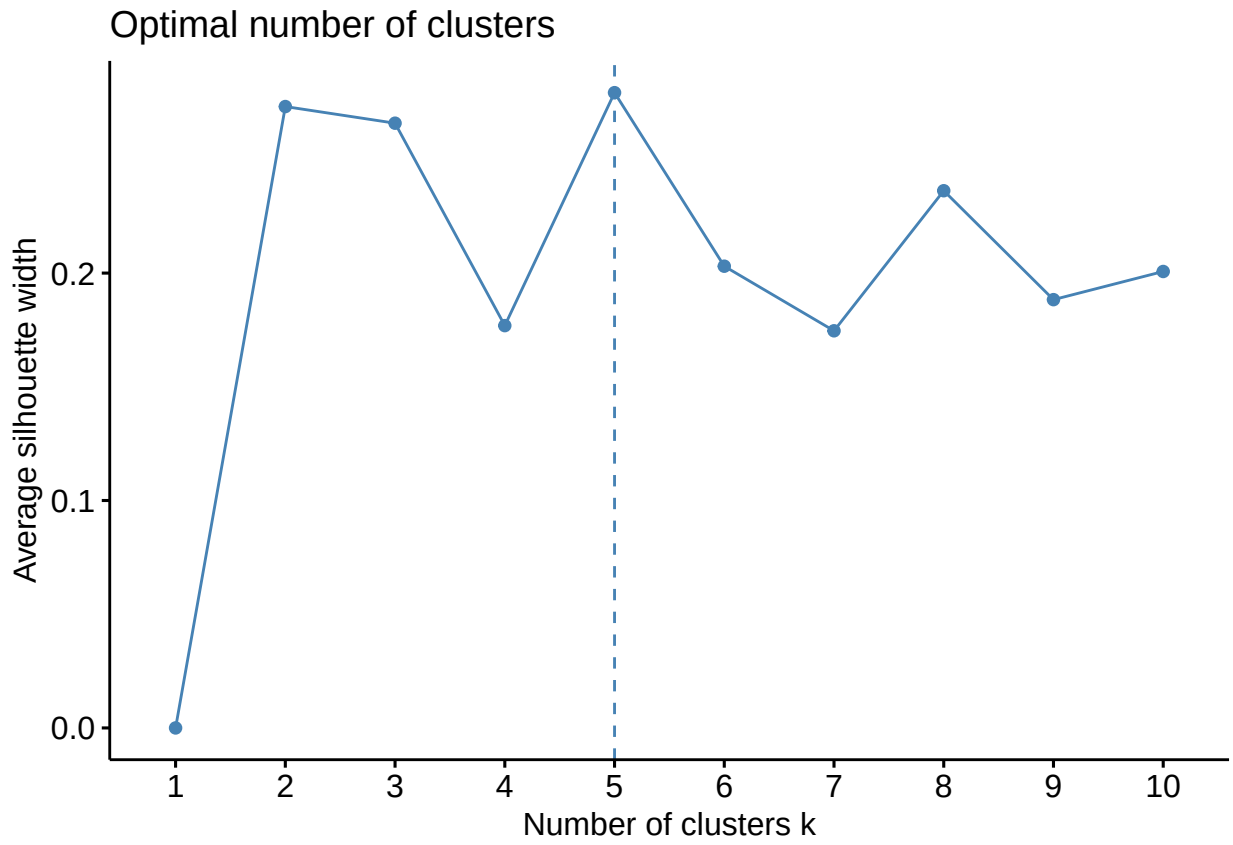
***Finding the optimal k value using Elbow method and silhouette method:***

```
set.seed(561)
fviz_nbclust(norm_data1,kmeans,method="wss")
```




*Optimal Value of k using elbow method is 2*

```
set.seed(351)
fviz_nbclust(norm_data1,kmeans,method="silhouette")
```




*Optimal Value of k using silhouette method is 5*

*As the optimal value of k obtained from Elbow method and Silhouette method is different, we will form clusters using both the optimal values and try to understand which optimal value of k forms better clusters*

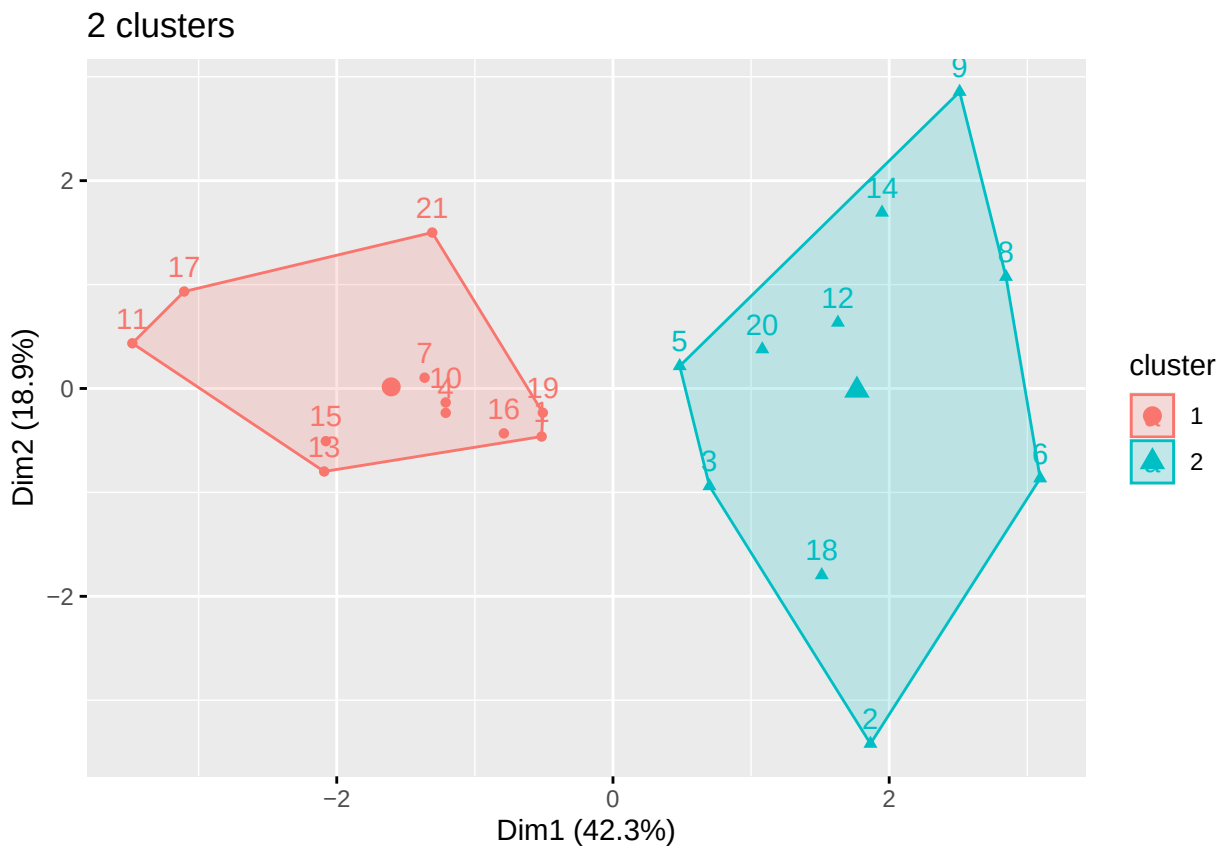
***Finding kmeans using k=2:***

```
k2<-kmeans(norm_data1,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"
```

```
#Plot of the clusters
fviz_cluster(k2,Pharmaceuticals[,(3:11)],main="2 clusters")
```




```
#Assigning the cluster to each firm using CBIND
data2<-cbind(data1,k2$cluster)
head(data2)
```

```
##   Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## 1      68.44 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     0.60       9.16
## 3       6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05
## 4      67.63 0.52     21.5 27.4 15.4            0.9     0.00      15.00
## 5      47.16 0.32     20.1 21.8  7.5            0.6     0.34      26.81
## 6      16.90 1.11     27.9  3.9  1.4            0.6     0.00      -3.17
##   Net_Profit_Margin k2$cluster
## 1              16.1          1
## 2               5.5          2
## 3              11.2          2
## 4              18.0          1
## 5              12.9          2
## 6               2.6          2
```

***Finding Mean within each cluster to interpret the clusters:***

```
mean_k2 <- data1 %>% 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>
```

**(b)Interpreting the clusters:**

*After thoroughly analysing both the clusters based on the averages calculated above, I would recommend to invest in cluster 2 for the following reasons:*

*>Firms in the second cluster has good market value indicating that these firms are well-established with generally longer history in businesses. It is safer to invest in such firms.*

*>The average P/E Ratio of the firms in second cluster falls under the current market range of 20-25 which is considered to be good.*

*>Return on equity is good for firms in second cluster as it has firms with average ROE of 30.42 which is more than the average of 15-20%.(as per industry standards, 15-20% is considered to be good ROE)*

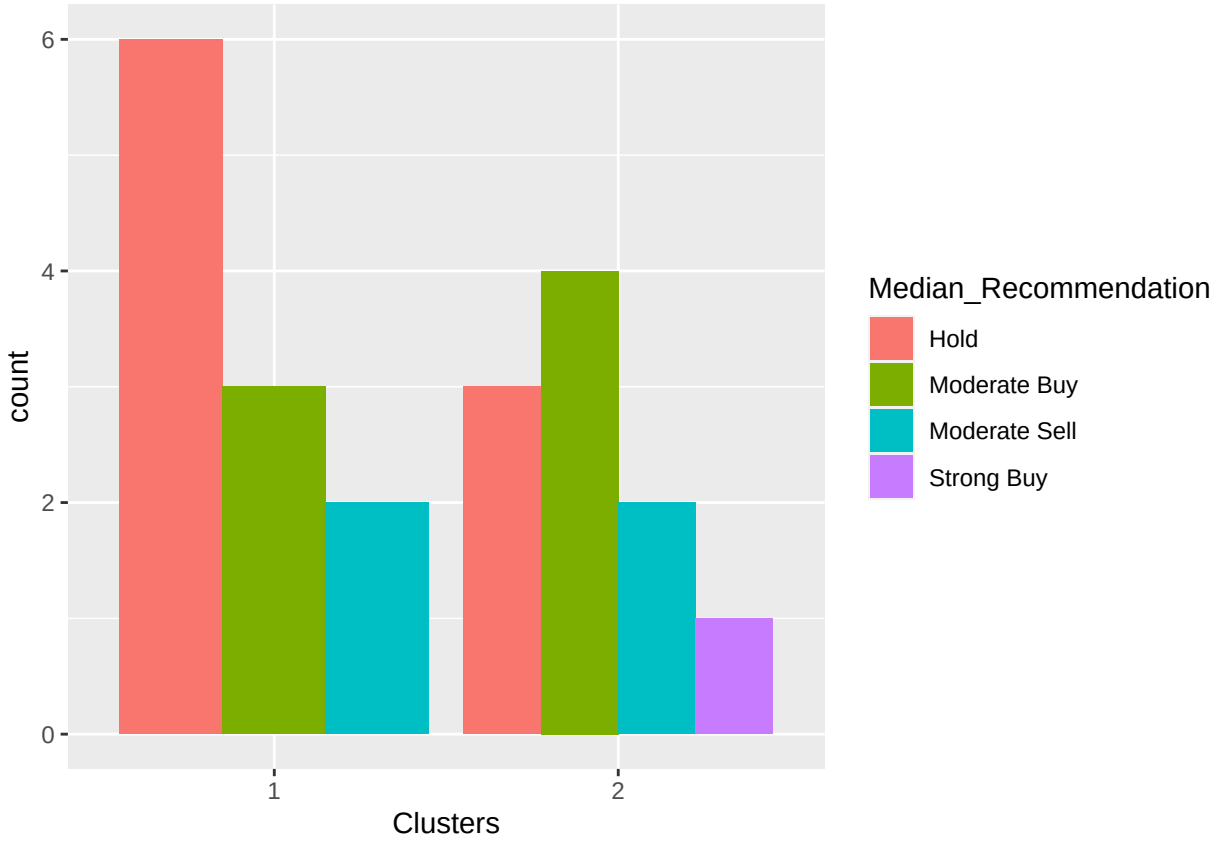
*>Comparitively,ROA in cluster 2 has better Return on Assets.(ROA more than 5% is considered to be good for investments)*

*>Firms in cluster 2 has good leverage value as the leverage value of less than 1 is considered to be good enough to invest as per industry standards.*

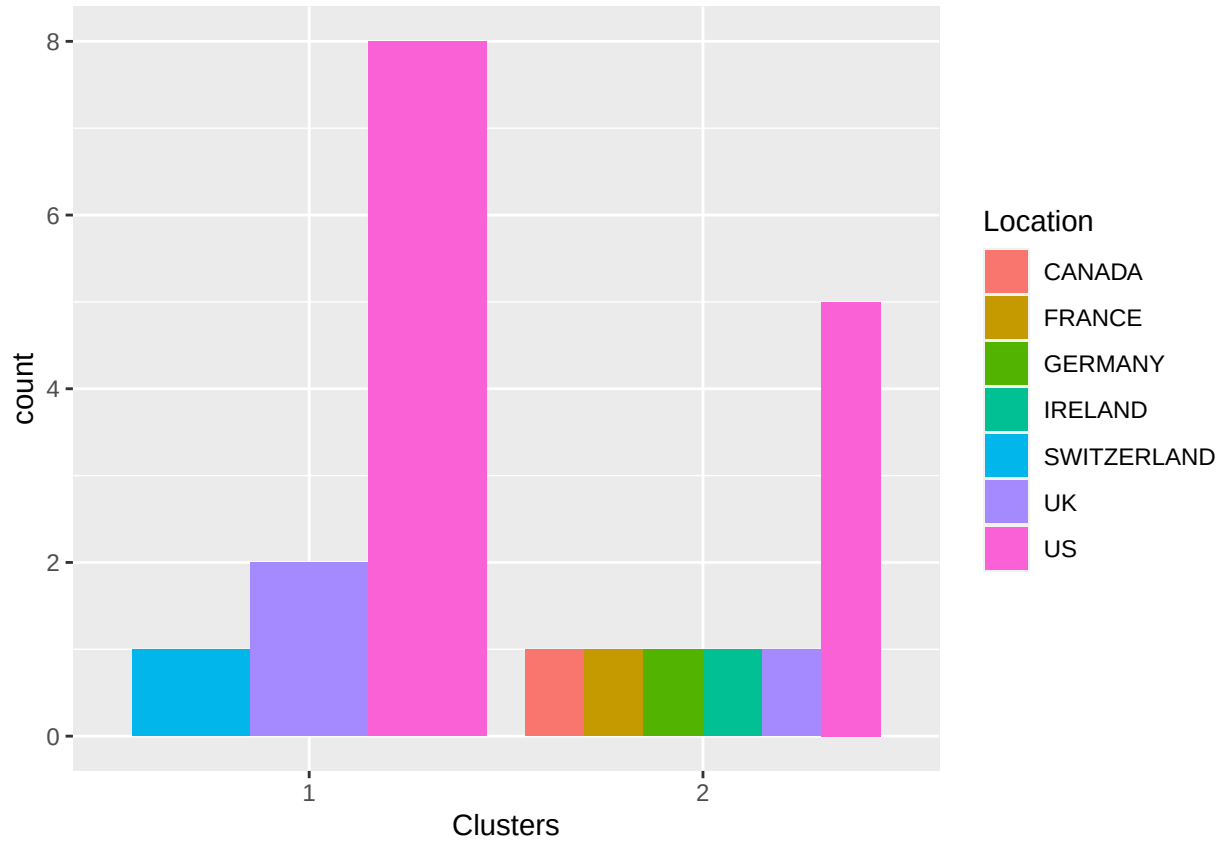
*>Cluster 2 has firms with good net profit margin which usually indicates how much of each dollar in revenue collected by a company translates into profit. In general, net profit margin above 10% indicates good and above 20% indicates excellent. Average Net Profit Margin in cluster 2 is 20.17%*

***(c)Interpreting the pattern in the clusters with respect to the categorical values:***

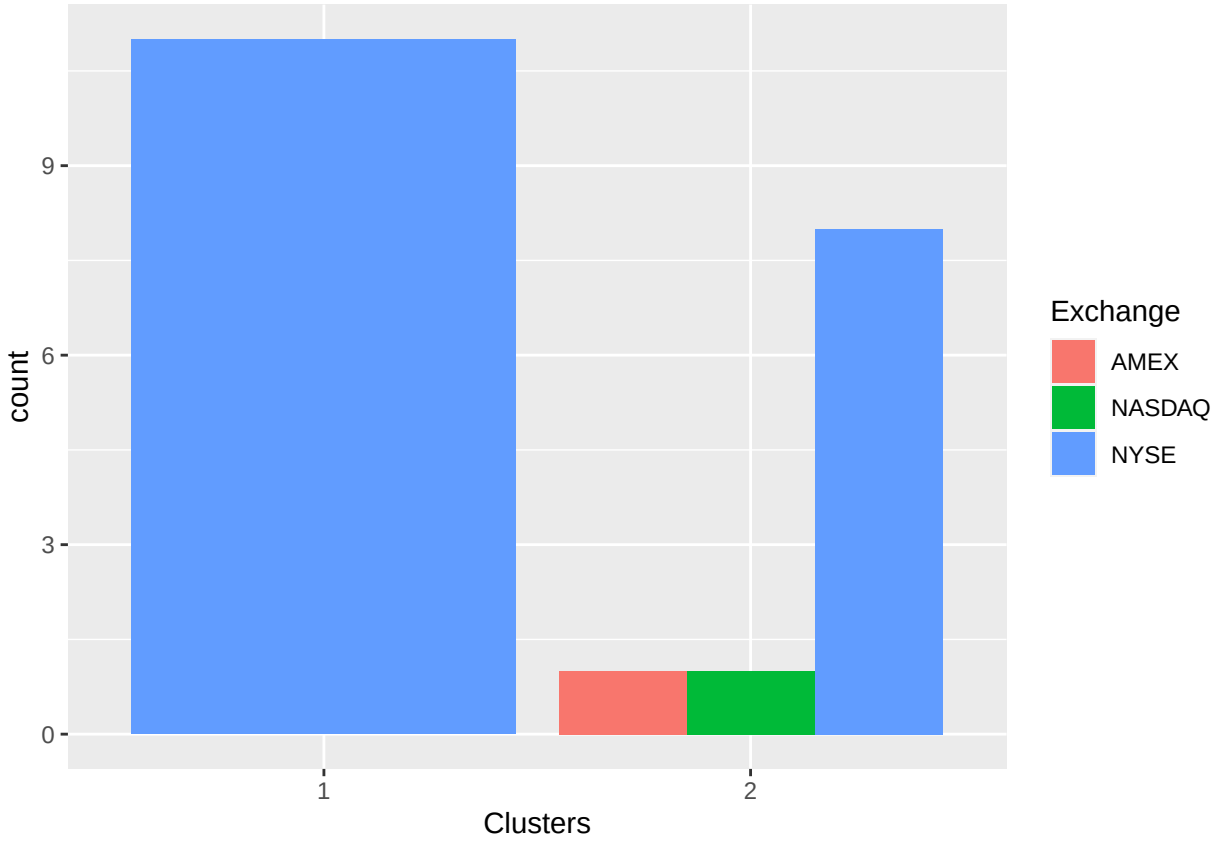
```
plot <- Pharmaceuticals[12:14] %>% mutate(Clusters=k2$cluster)
ggplot(plot, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position='dodge')+la
```




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




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




**Analysis based on the plots of categorical variables**

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

*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*

***(d) Appropriate names for clusters:***

*Cluster 1 : Bad Investment*

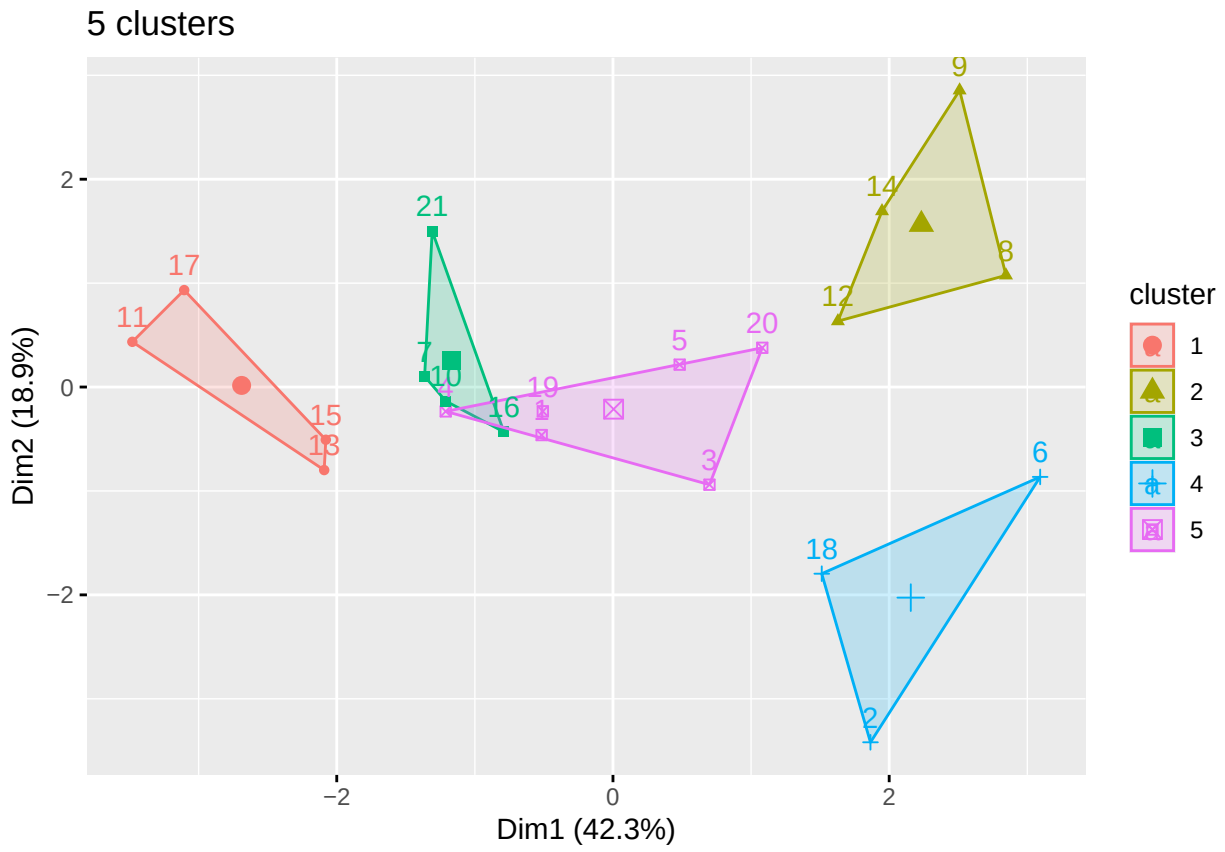
*Cluster 2: Good Investment*

***Finding kmeans using k=5***

```
k5<-kmeans(norm_data1,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"
```

```
#plot of the clusters
fviz_cluster(k5,Pharmaceuticals[,(3:11)],main="5 clusters")
```




```
#Assigning the cluster to each firm using CBIND
data3<-cbind(data1,k5$cluster)
head(data3)
```

```
##   Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## 1      68.44 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     0.60       9.16
## 3       6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05
## 4      67.63 0.52     21.5 27.4 15.4            0.9     0.00      15.00
## 5      47.16 0.32     20.1 21.8  7.5            0.6     0.34      26.81
## 6      16.90 1.11     27.9  3.9  1.4            0.6     0.00      -3.17
##   Net_Profit_Margin k5$cluster
## 1              16.1          5
## 2               5.5          4
## 3              11.2          5
## 4              18.0          5
## 5              12.9          5
## 6               2.6          4
```

***Finding Mean within each cluster to interpret the clusters:***

```
mean_k5 <- data1 %>% mutate(Cluster = k5$cluster) %>% group_by(Cluster) %>% summarise_all("mean")
mean_k5
```

```
## # A tibble: 5 x 10
```

```
##   Cluster Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##     <int>      <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
## 1       1     157.   0.48      22.2  44.4  17.7           0.95     0.22
## 2       2       1.25 0.832     19.5  18.0   5.4           0.45     1.74
## 3       3      67.5  0.375     19.1  34.6  13.3           0.65     0.57
## 4       4      26.9  0.64      55.6  10.1   4.2           0.7      0.317
## 5       5      37.8  0.395     20.7  20.6  10.4           0.733    0.205
## # i 2 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>
```

**{b)Analysing the clusters:**

*On carefully interpreting the average value os each variable in all clusters, I would recommend to invest in cluster 5 as it has the highest Market Capital value*

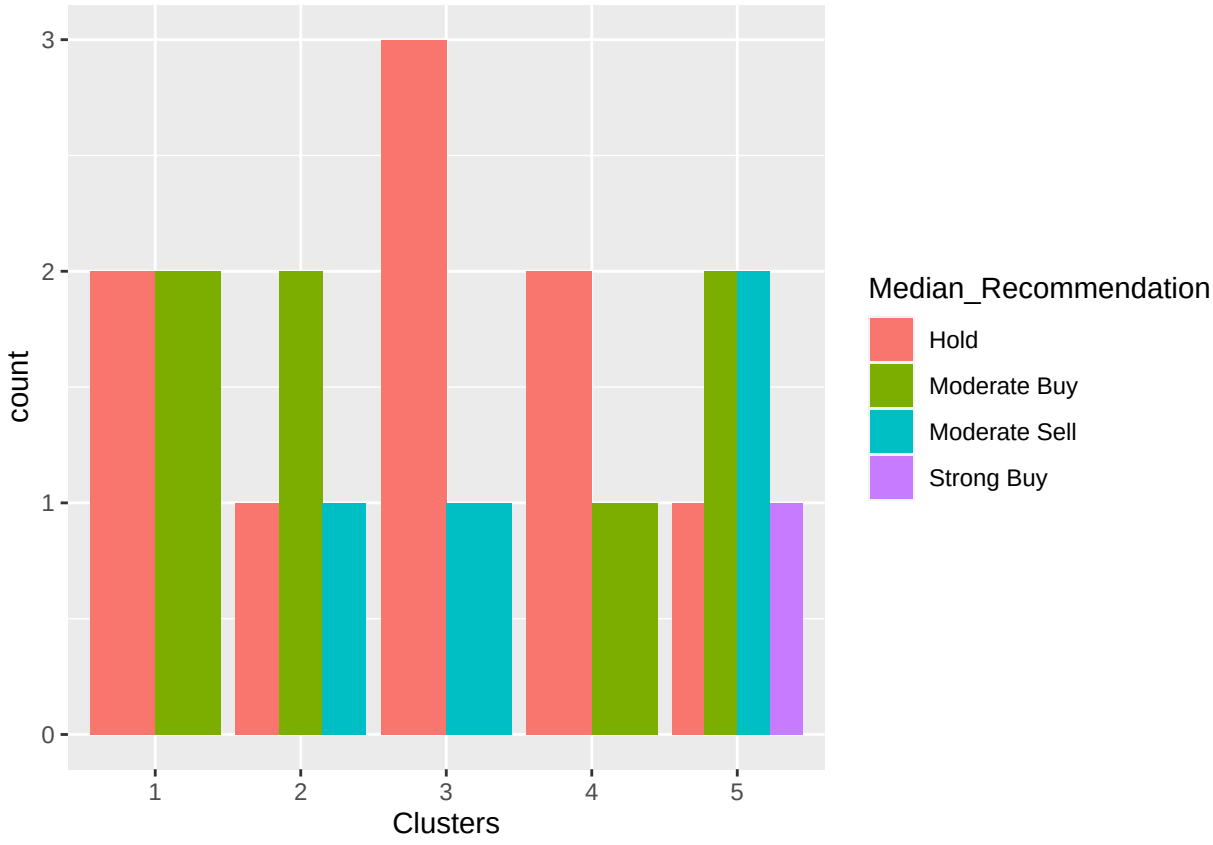
*Furthermore, it can observed that the cluster 5 has firms with good ROE,ROA,asset turnover*

*Moreover, the average P/E ratio of firms in cluster 5 is comparitively good to invest in those firms*

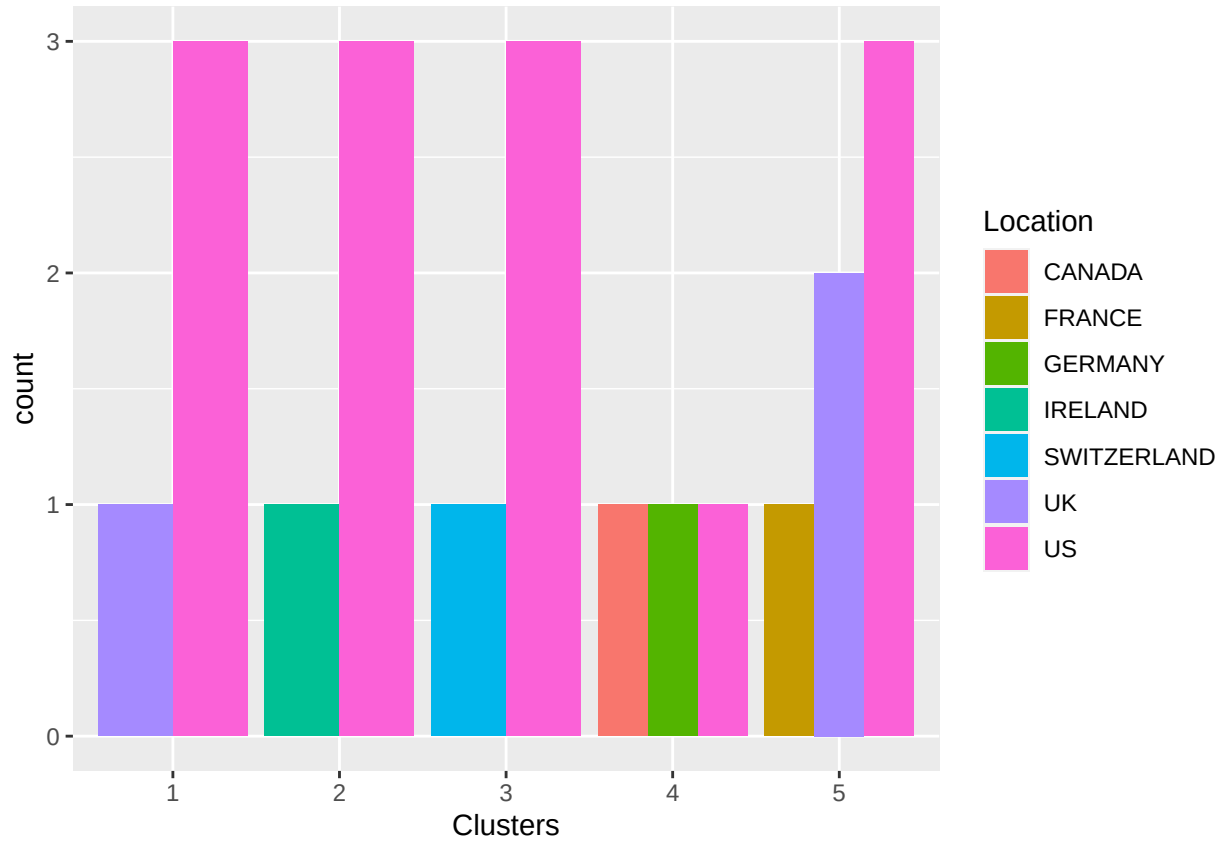
*Firms of cluster 5 also has the highest Net Profit Margin*

***(c)Interpreting the pattern in the clusters with respect to the categorical values:***

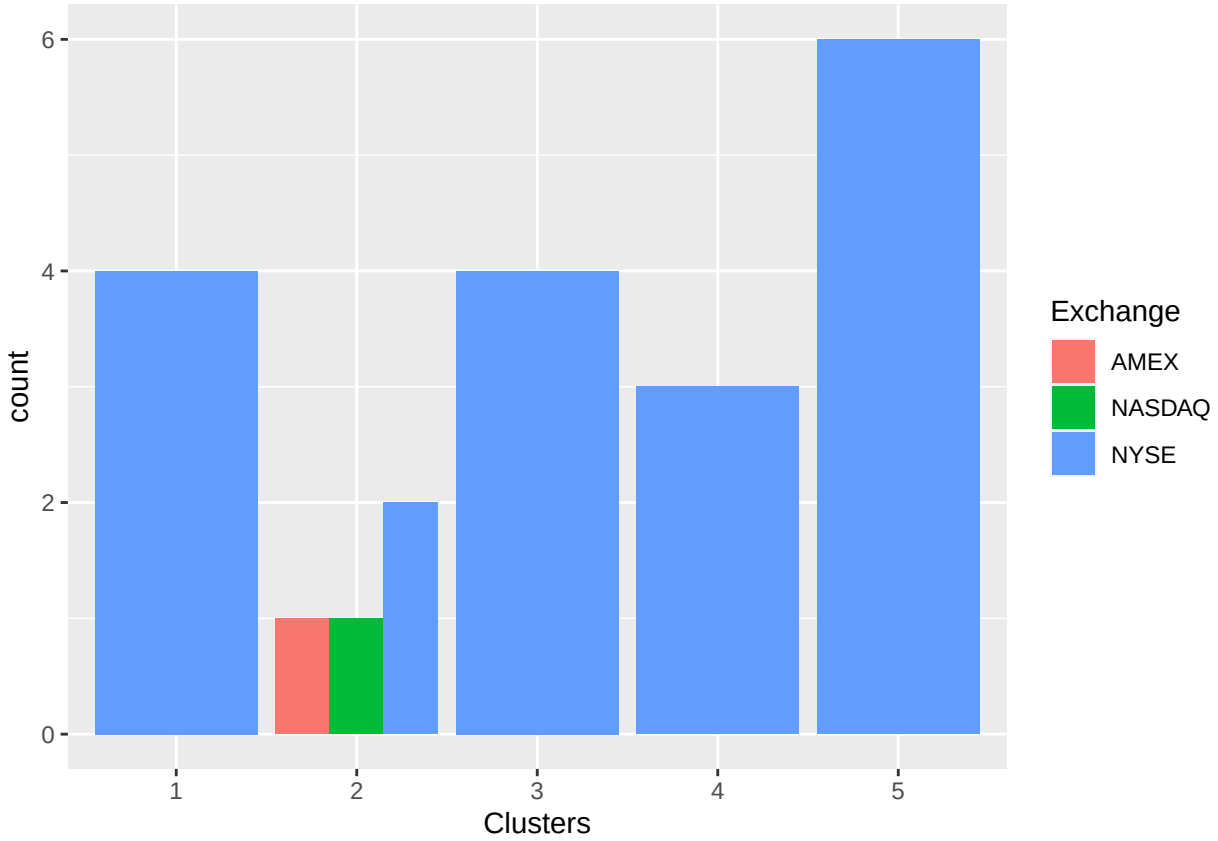
```
plots <- Pharmaceuticals[12:14] %>% mutate(Clusters=k5$cluster)
ggplot(plots, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position='dodge')+l
```




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




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




**Analysis**

*Median_Recommendation: It can be observed that in clusters 1,4 and 5- the firms are under 'Hold' recommendation. Few firms in all clusters has 'Moderate buy' recommendation.*

*Location: All clusters has firms from the US.*

*Exchange: Firms in clusters 1,2,3 and 5 are listed under NYSE whereas firms in cluster 4 are listed under AMEX,NYSE and NASDAQ.*

***(d)Appropriate names for clusters***

*Cluster 1: Safe Investment- This cluster has firms with high market capital,low beta value, good Return on Equity, Assets and good Net Profit Margin. Hence, it is safe to invest in this cluster*

*Cluster 2:Bad Investment- Firms in this cluster has low Market Capital indicating that these firms are relatively new in the market. They have bad leverage,high beta value. Hence, it would be a bad choice to invest in firms of this cluster*

*Cluster 3: Risky investment- Firms of this cluster has high market capital but low Net Profit Margin which means that these firms have overvalued stocks. The stock price of these firms can go down quickly so it would be a risk to invest in these firms*

*Cluster 4: Good Investment- This cluster has firms with almost same value of market capital and net profit margin. These firms has good P/E ratio, ROE and ROA. Hence, it would be a wise choice to invest in these firms*

*Cluster 5: Recommended Investment- As the name suggests, it is highly recommended to invest in this cluster as it has the highest Market Capital,Net Profit Margin,ROE and Asset turnover*

***Conclusion***

*After finding k-Means using optimal values of k obtained from both Elbow method and Silhouette Method, I would prefer to perform K-Means using k=5 based on silhouette method as the 5 clusters formed with this optimal value provides better insights on 21 firms. These 5 clusters gives individuals clear idea about all sorts of investments, hence categorizing their list of firms into safe,unsafe,risky,good and bad investments.*