# Clustering method

Dev

2023-11-12

## Problem Statement

**An equities analyst is studying the pharmaceutical industry and would like your help in exploring and understanding the financial data collected by her firm. Her main objective is to understand the structure of the pharmaceutical industry using some basic financial measures. Financial data gathered on 21 firms in the pharmaceutical industry are available in the file Pharmaceuticals.csv Download Pharmaceuticals.csv. For each firm, the following variables are recorded:**

**Summary of answers:**

1. K-means is the most optimum method for clustering the given data.
2. Patters in variables not used in clustering are as follows:

- Cluster-1: Companies are listed in all three exchanges and do business in USA and Germany. Recommendation is to Hold more companies and Moderate buy on some.
- Cluster-2: All companies are listed in NYSE and do business in Canada and USA. Recommendation is to hold half and moderate buy half companies.
- Cluster-3: All companies are listed in NYSE and do business in Switzerland, UK and US. Recommendation is mixed but most are recommended to be hold.
- Cluster- 4: All companies are listed in NYSE and do business in UK and US. Recommendation is to Moderate buy half and hold half.
- Cluster-5: All companies are listed in NYSE and do business in France, Ireland and US. Recommendation is to Moderate buy half and Moderate sell half.

3. All five clusters can be named as follows:

- Cluster-1: Low cap Highly-Volatile companies.(Because of high Beta value and low profits and small market cap)
- Cluster-2: Small cap overpriced companies.(Because of higher PE ratio and smaller market cap)
- Cluster-3: Mid cap Profitable companies. (Most companies has above than average profits and have average market cap)
- Cluster-4: Large-cap Under-priced companies. (High Market cap and Lower than average PE ratio but all financials seem good)

- Cluster-5: Small cap Less-Profitable companies ( Smaller market cap and Profits are lower than average)

```
setwd("/Users/devmarwah/Downloads")
df=read.csv("Pharmaceuticals.csv")
summary(df)
```

```
    Symbol              Name             Market_Cap          Beta       
 Length:21          Length:21          Min.   :  0.41   Min.   :0.1800  
 Class :character   Class :character   1st Qu.:  6.30   1st Qu.:0.3500  
 Mode  :character   Mode  :character   Median : 48.19   Median :0.4600  
                                       Mean   : 57.65   Mean   :0.5257  
                                       3rd Qu.: 73.84   3rd Qu.:0.6500  
                                       Max.   :199.47   Max.   :1.1100  
    PE_Ratio          ROE             ROA        Asset_Turnover     Leverage     
 Min.   : 3.60   Min.   : 3.9   Min.   : 1.40   Min.   :0.3    Min.   :0.0000  
 1st Qu.:18.90   1st Qu.:14.9   1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600  
 Median :21.50   Median :22.6   Median :11.20   Median :0.6    Median :0.3400  
 Mean   :25.46   Mean   :25.8   Mean   :10.51   Mean   :0.7    Mean   :0.5857  
 3rd Qu.:27.90   3rd Qu.:31.0   3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000  
 Max.   :82.50   Max.   :62.9   Max.   :20.30   Max.   :1.1    Max.   :3.5100  
   Rev_Growth    Net_Profit_Margin Median_Recommendation   Location        
 Min.   :-3.17   Min.   : 2.6      Length:21             Length:21         
 1st Qu.: 6.38   1st Qu.:11.2      Class :character      Class :character  
 Median : 9.37   Median :16.1      Mode  :character      Mode  :character  
 Mean   :13.37   Mean   :15.7                                              
 3rd Qu.:21.87   3rd Qu.:21.1                                              
 Max.   :34.21   Max.   :25.5                                              
   Exchange        
 Length:21         
 Class :character  
 Mode  :character  
```

**Use cluster analysis to explore and analyze the given dataset as follows:**

**Q:Use only the numerical variables (1 to 9) to cluster the 21 firms. Justify the various choices made in conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on.**

Dropping non-numerical variables:

```
pharma.df=df[,c(3:11)] #Using basic code to remove non-numerical data
```

Normalising the data before analysing:

```
norm=preProcess(pharma.df,method = c("center","scale"))
pharma.norm.df=predict(norm,pharma.df)
```

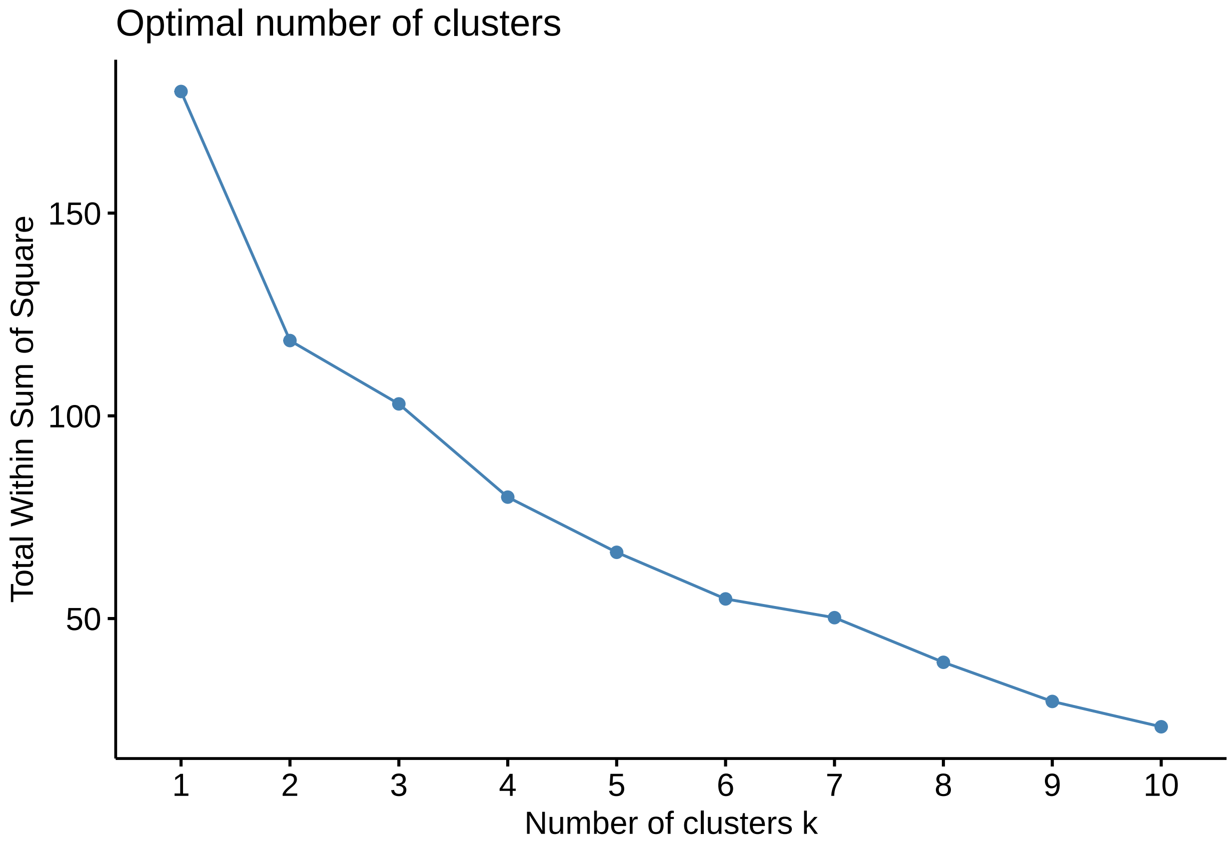
- **K-means:**

Now, using elbow method to determine the value of K.

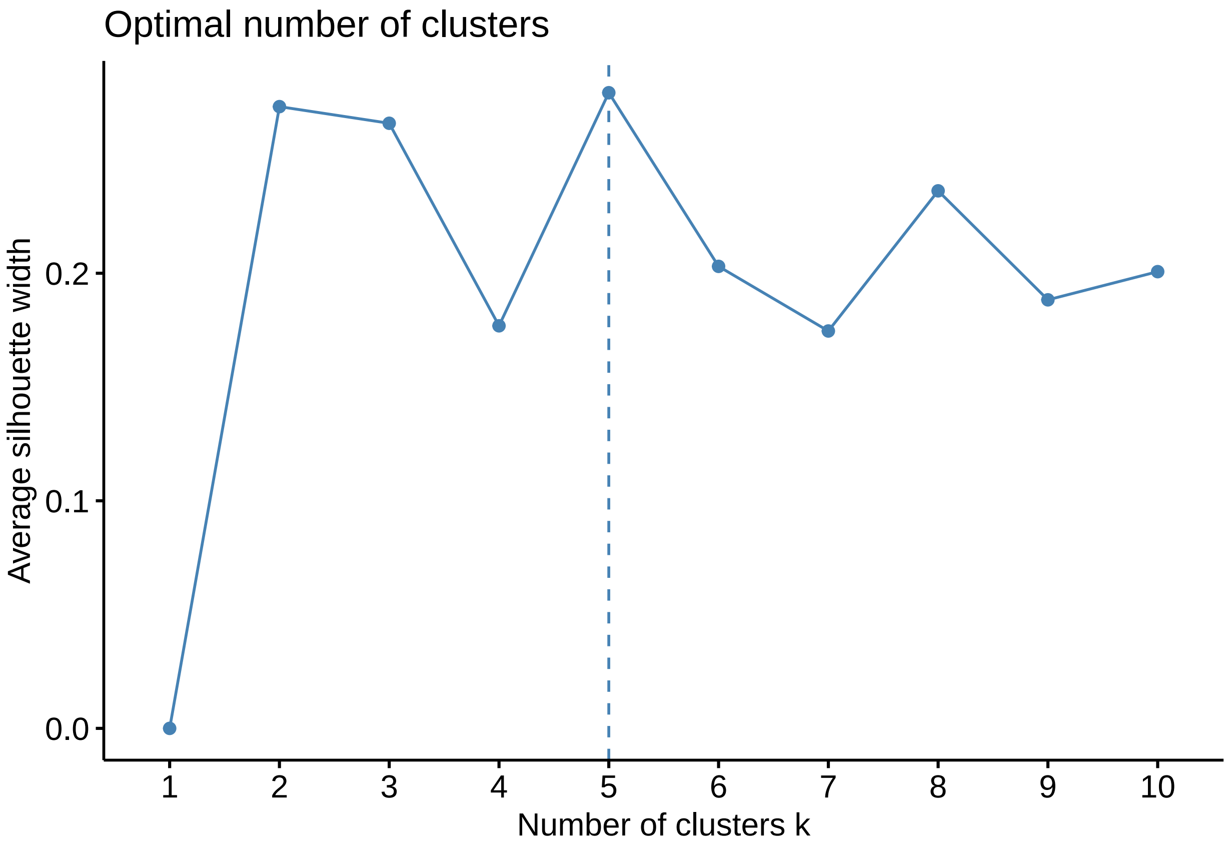
```
library(tidyverse)
library(factoextra)
```

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

```
fviz_nbclust(pharma.norm.df,kmeans,method = "wss")
```



```
fviz_nbclust(pharma.norm.df,kmeans,method="silhouette")
```



From the above graphs, we can determine that k=5 is the optimum value for k, keeping overfitting and bias into consideration.

Now, performing the k-means clustering:

```
k=kmeans(pharma.norm.df,centers=5,nstart = 10)
```

Following are the centers:

```
k$centers
```

```
   Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
1 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
2 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
4  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
5 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
     Leverage Rev_Growth Net_Profit_Margin
1  1.36644699 -0.6912914      -1.320000179
2 -0.14170336 -0.1168459      -1.416514761
3 -0.27449312 -0.7041516       0.556954446
4 -0.46807818  0.4671788       0.591242521
5  0.06308085  1.5180158      -0.006893899
```

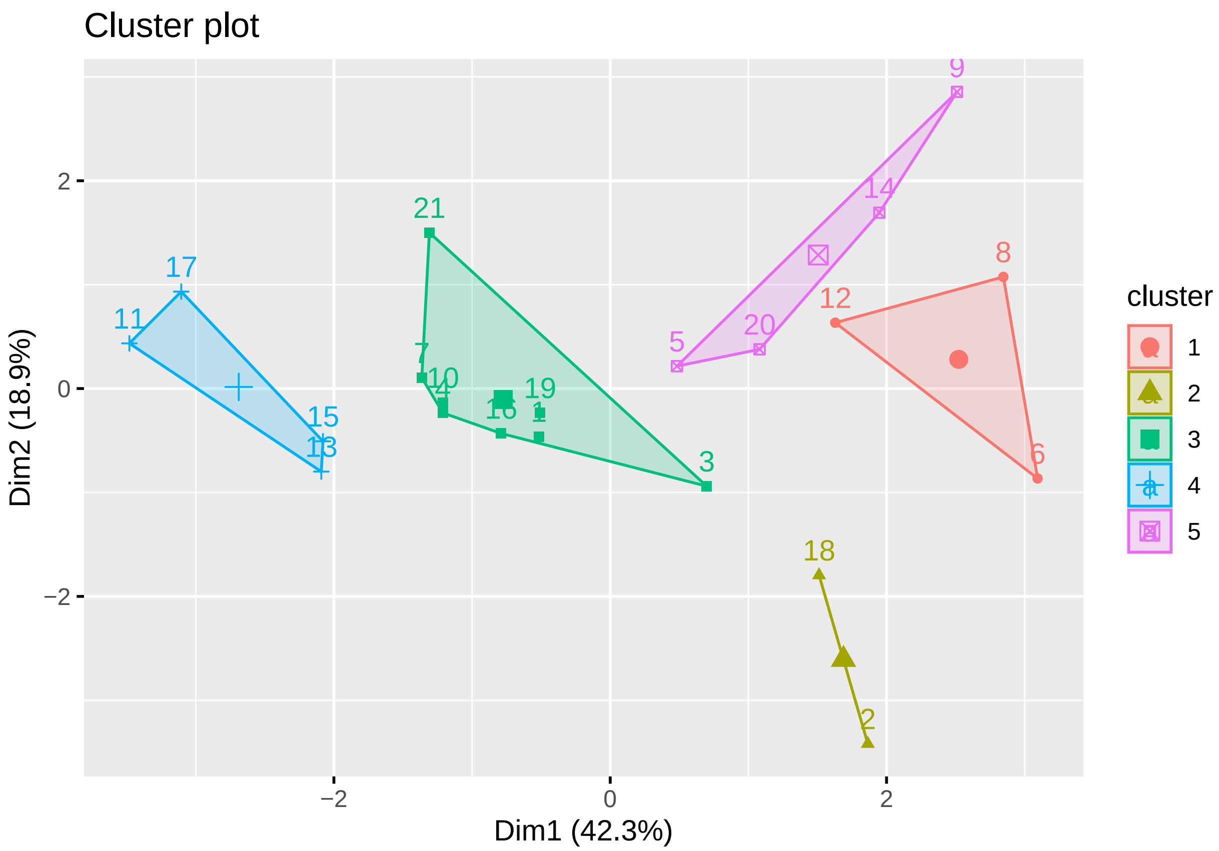
Following are sizes the clusters:

```r
k$size
```

```
[1] 3 2 8 4 4
```

Clustering can be visualized as:

```r
fviz_cluster(k,data = pharma.norm.df)
```



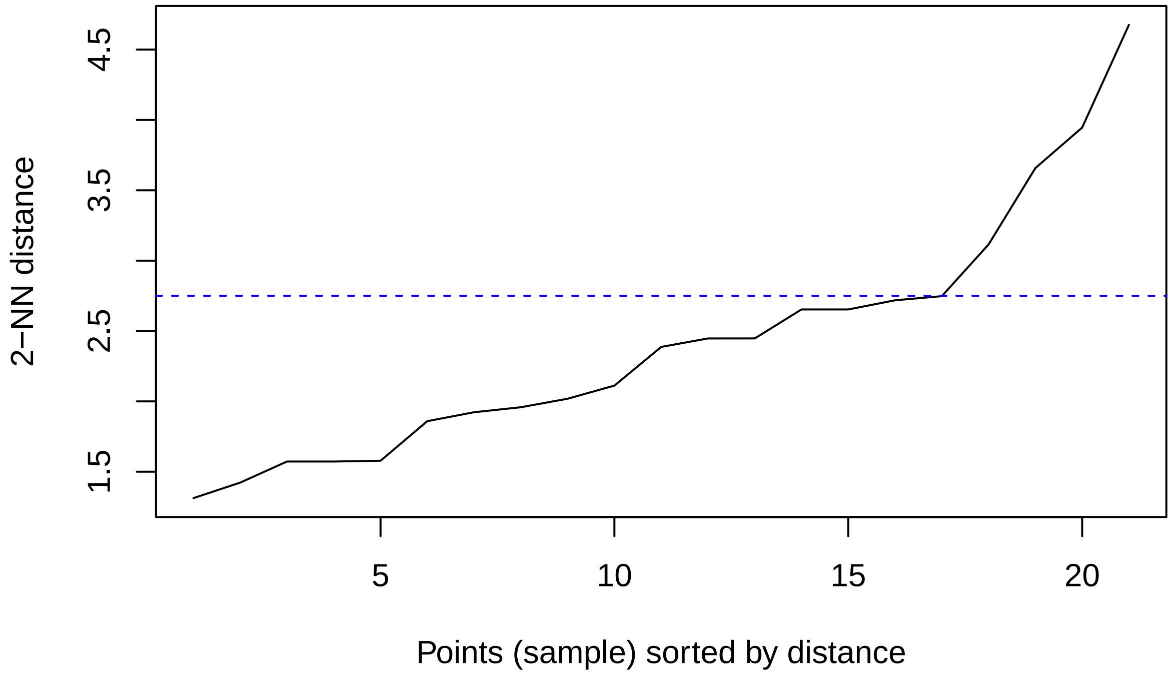
**Interpretation:**

K-means plot shows us an optimum picture of clusters. All points which are close to each other are in one cluster. Also, it's easy to study the structure of given using this kind of uniform cluster plot.

- **DBSCAN:**

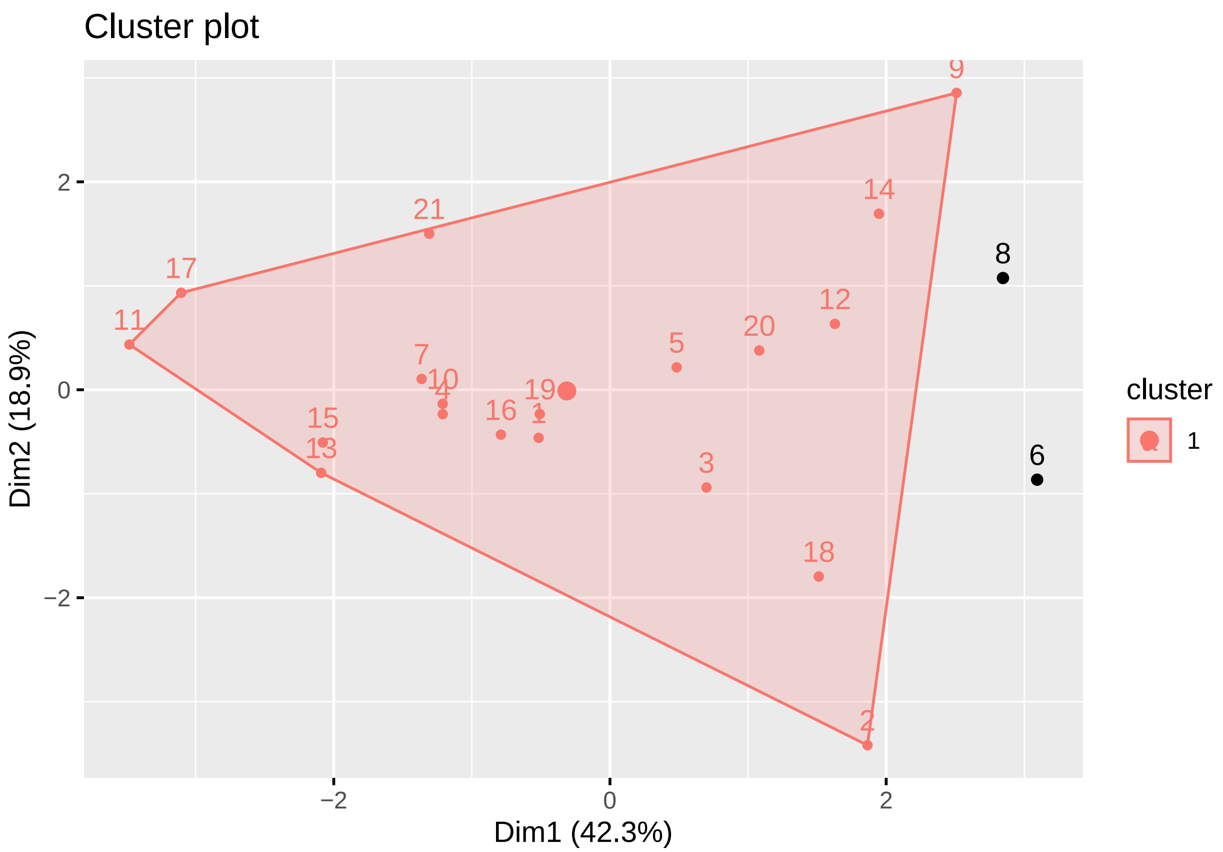
Trying to find optimum value of eps using k=5:

```r
dbscan::kNNdistplot(pharma.norm.df,k=2)
abline(h=2.75,lty="dashed",col="blue")
```



Hence, value of eps=2.75 (knee-point) is optimum and we can construct a dbscan model on this.

```
db= dbscan::dbscan(pharma.norm.df,eps=2.75,minPts = 2)
fviz_cluster(db,pharma.norm.df)
```



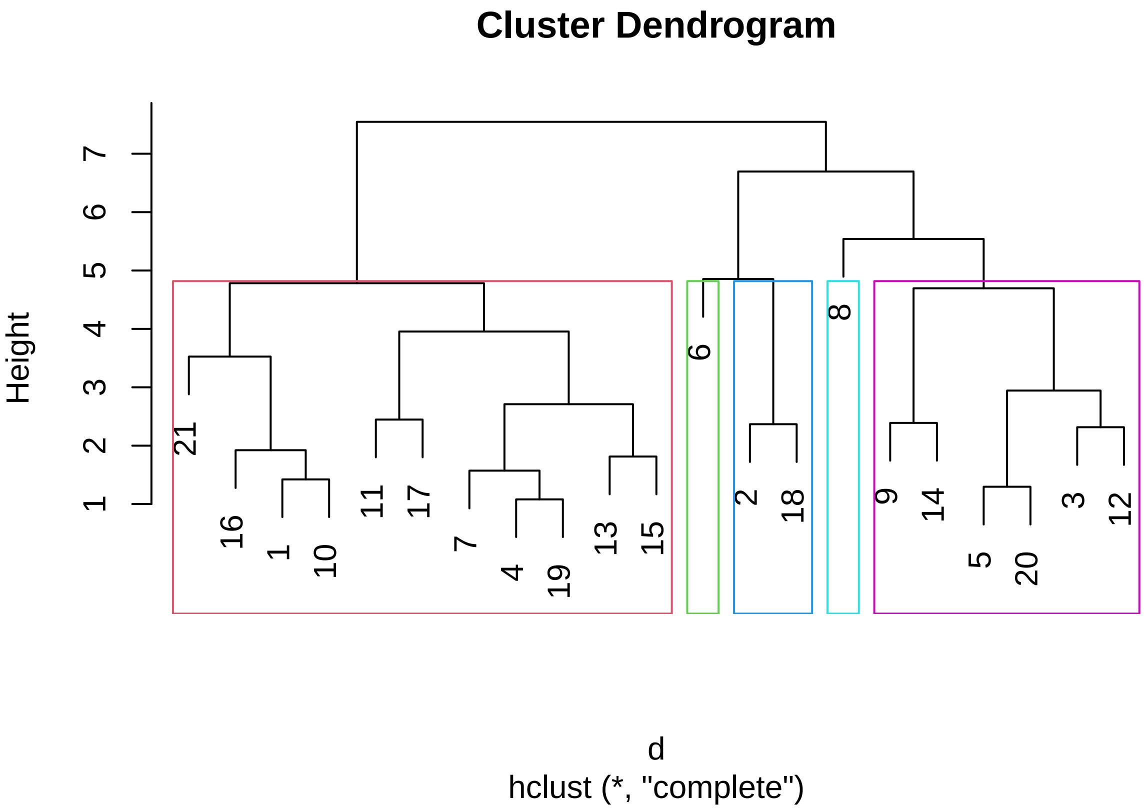
**Interpretation:**

DBSCAN method seems to be a wrong method for clustering the given data because it keeps almost all of the data points in a single cluster. Also, if a smaller value of eps is chosen then it keeps most of the points as outliers. Hence, DBSCAN is not an optimum method for given data.

- **Hierarchical Method:**

Applying hclust directly

```
d=dist(pharma.norm.df,method = "euclidean")
hc=hclust(d,method = "complete")
plot(hc)
rect.hclust(hc,k=5,border = 2:7)
```



**Interpretations:**

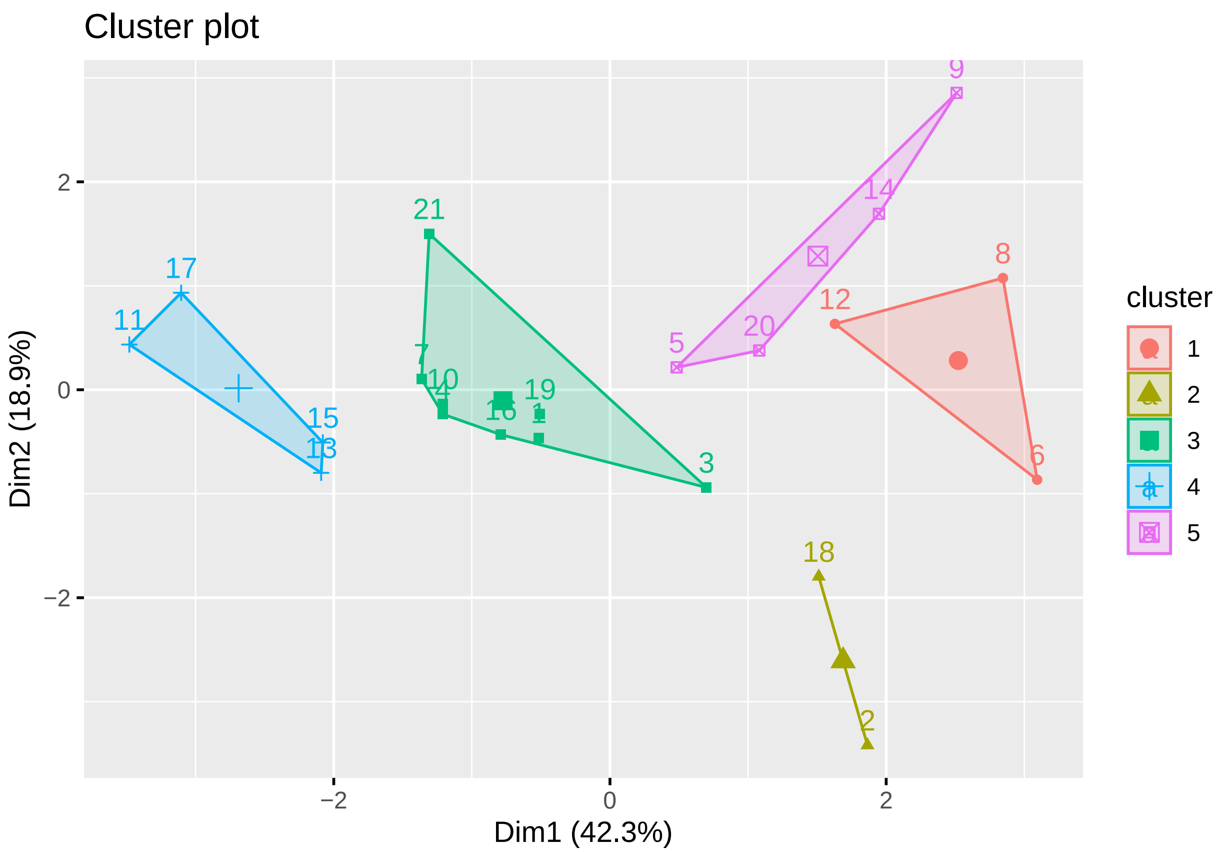
Hierarchical though gives us a seemingly nice picture of clusters, when we try to make 5 clusters like k-means with it, it presents some outliers which are indeed close to some other points when compared to k-means plot and instead should be considered in a cluster with other points. Also, there is no hierarchy in financial data and hence doing a hierarchical clustering doesn't make any sense .This method too seems to be lacking behind of k-means method in organizing the data into clusters.

**Justification of choices made:**

1. Since there is no indication in question regarding weightage of variables, we have normalized data and hence given equal weightage to all variables.
2. After trying out different clustering methods, I figured out that k-means method is the best method for given data because it gives us a better clustered picture of data with all points which are closer to each other are in same cluster.
3. We have used methods like elbow-method and distplot to find the values of k and eps in order to determine the number of clusters we need. Also, in hierarchical plot we kept number of clusters as 5 just so that we can compare it with K-means clustering.

**Q:Interpret the clusters with respect to the numerical variables used in forming the clusters. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)**

```
fviz_cluster(k,data = pharma.norm.df)
```



- With respect to numerical values used in the clustering, the numerical values of points in same k-means clusters are close to each other as compared points in different groups. We can view these clusters to interpret better.

Cluster-1:

```
pharma.norm.df[c(6,8,12),]
```

```
   Market_Cap      Beta    PE_Ratio        ROE        ROA Asset_Turnover
6  -0.6953818 2.2757827  0.14948233 -1.4514600 -1.7127612     -0.4612656
8  -0.9767669 1.2630872  0.03299122 -0.1123792 -1.1677918     -0.4612656
12 -0.9393967 0.4840907 -0.34100657 -0.2913653 -0.6979905     -0.4612656
     Leverage  Rev_Growth Net_Profit_Margin
6  -0.7496565 -1.49714434        -1.9956023
8   3.7427970 -0.63276071        -1.2488842
12  1.1062004  0.05603085        -0.7155141
```

This cluster has high Beta values, average PE-ratio and all other variables are below average values.However, leverage and Rev_growth have mixed values.

Cluster-2:

```
pharma.norm.df[c(2,18),]
```

```
   Market_Cap       Beta PE_Ratio        ROE        ROA Asset_Turnover
2  -0.8544181 -0.4507051 3.497069 -0.8548399 -0.9422871      0.9225312
```

```
18 -0.0240846 -0.4896550 1.902980 -0.8150652 -0.9047030     -0.4612656
     Leverage Rev_Growth Net_Profit_Margin
2   0.0182843 -0.3811391         -1.553667
18 -0.3016910  0.1474473         -1.279362
```

This cluster has a high PE ratio but all other variables are lower than average. However, this cluster also has mixed values for leverage and rev_growth.

Cluster-3:

```
pharma.norm.df[c(1,3,4,7,10,16,19,21),]
```

```
   Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
1   0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
3  -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312
4   0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312
7  -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498      0.9225312
10  0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770     -0.4612656
16  0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598     -0.9225312
19 -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929      0.4612656
21 -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849     -0.4612656
      Leverage Rev_Growth Net_Profit_Margin
1  -0.21209793 -0.5277675        0.06168225
3  -0.40408312 -0.5721181       -0.68503583
4  -0.74965647  0.1474473        0.35122600
7  -0.02011273 -0.9658426        0.74744375
10 -0.07130879 -0.6481476        1.17413980
16 -0.67286239 -1.4536989        1.02174835
19 -0.74965647 -0.4354459        0.29026942
21  0.68383297 -1.1776392        1.49416183
```

This cluster has very high Net_profit values. However, other variables have mixed values.

Cluster-4:

```
pharma.norm.df[c(11,17,15,13),]
```

```
   Market_Cap       Beta   PE_Ratio       ROE       ROA Asset_Turnover
11   1.099920 -0.6844041 -0.4574977 2.4597165 1.8389364      1.3837968
17   2.419990  0.4840907 -0.1141555 1.3128800 1.6322239      0.4612656
15   1.278239 -0.2559560 -0.4023177 0.9814243 0.8429577      1.8450624
13   1.984176 -0.2559560  0.1801379 0.1859308 1.0872544      0.9225312
     Leverage Rev_Growth Net_Profit_Margin
11 -0.3144900  0.7692605         0.8236395
17 -0.5448723  1.1014372         1.4484444
15 -0.3912841  0.3601491        -0.2431006
13 -0.6216663 -0.3621317         0.3359869
```

All companies in this cluster have high values. Besides, Beta, PE ratio and Leverage.

Cluster-5:

```
pharma.norm.df[c(5,9,14,20),]
```

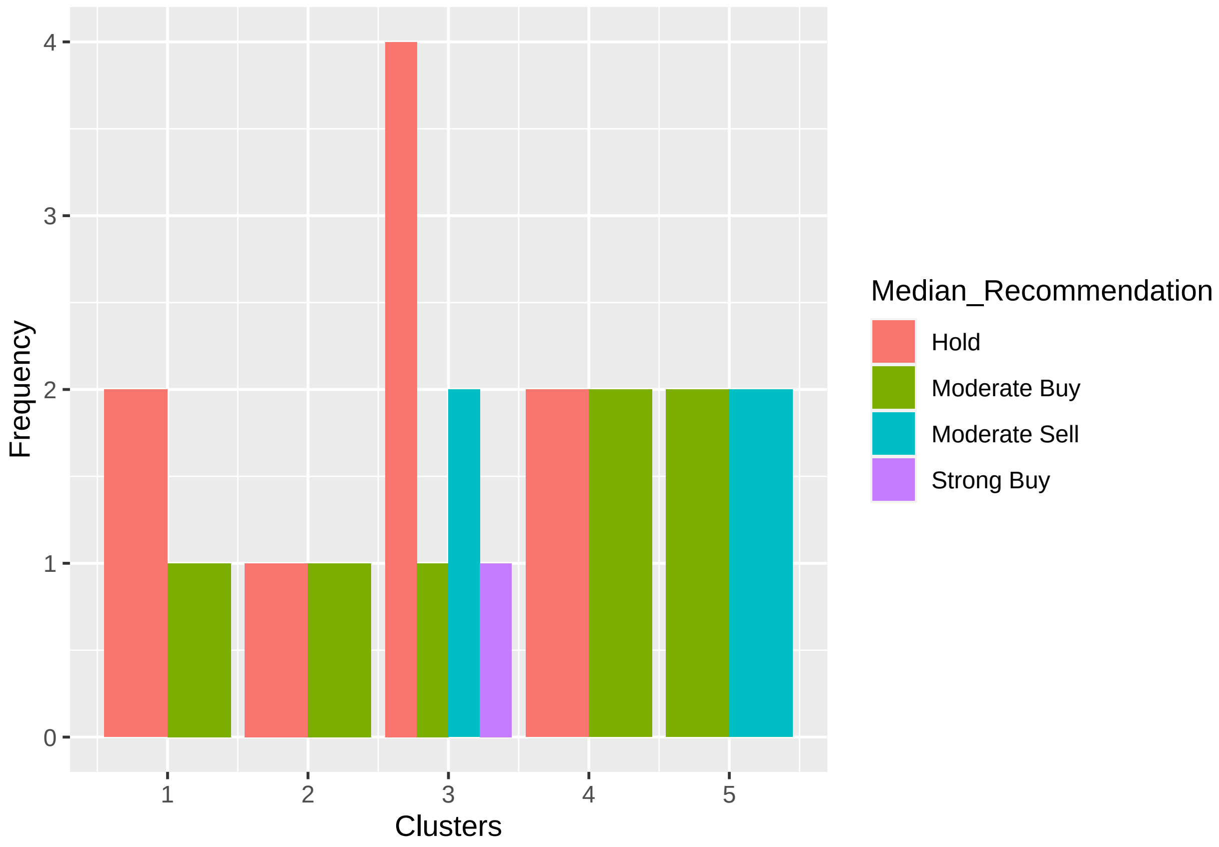
```
    Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
5  -0.1790256 -0.8012536 -0.3287443 -0.2648488 -0.5664461     -0.4612656
9  -0.9704532  2.1589332 -1.3403777 -0.7089994 -1.0174553     -1.8450624
14 -0.9632863  0.8735889  0.1924001 -0.9675348 -0.9610792     -1.8450624
20 -0.9281345 -1.1128522 -0.4329732 -1.0338259 -0.6979905     -0.9225312
     Leverage Rev_Growth Net_Profit_Margin
5  -0.3144900   1.216387       -0.42597037
9   0.6198379   1.886171       -0.36501379
14  0.4406517   1.538607        0.85411776
20 -0.4936762   1.430899       -0.09070919
```

This cluster has low values for Marketcap, PE-ratio, ROE, ROA and Net_Profit but all other values have mixed values.

- Making barplots to check patters in variables we did not use in clustering.

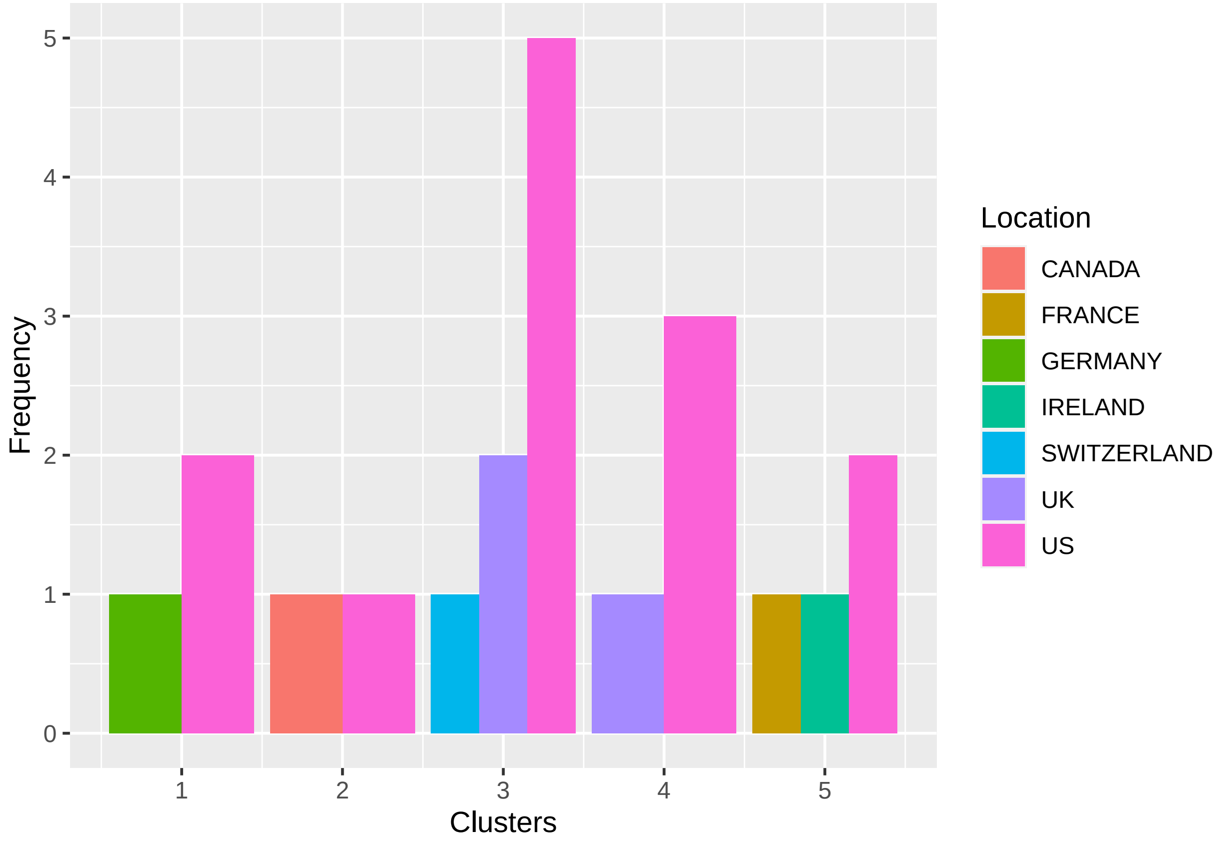
Comparing recommendation of clusters-

```
df.2=df %>%
  select(c(1,12,13,14)) %>%
  mutate(cluster=k$cluster)
ggplot(df.2,mapping = aes(cluster,fill=Median_Recommendation))+
  geom_bar(position = 'dodge') +
  labs(x='Clusters',y='Frequency')
```



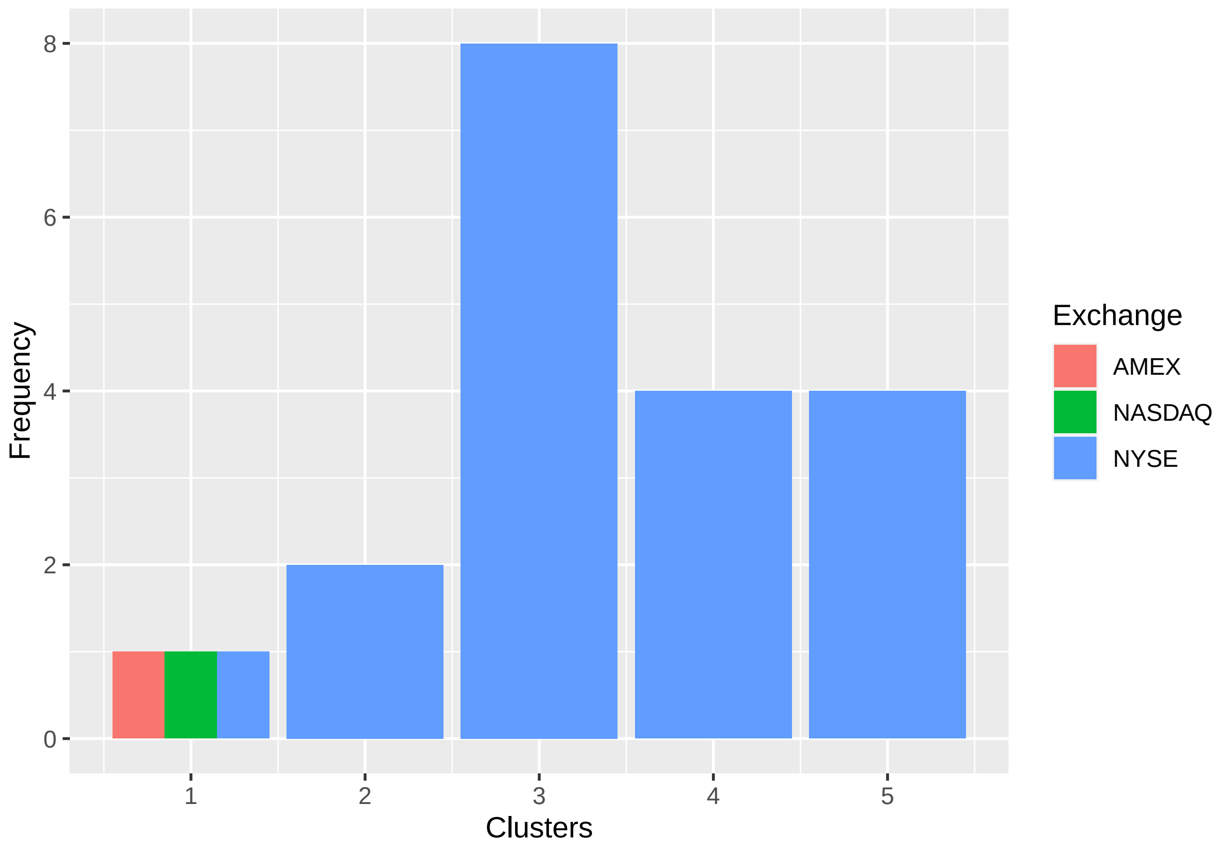
Comparing countries of clusters-

```
ggplot(df.2,mapping = aes(cluster,fill=Location))+
  geom_bar(position = 'dodge') +
  labs(x='Clusters',y='Frequency')
```



Comparing stock-exchanges of clusters-

```
ggplot(df.2,mapping = aes(cluster,fill=Exchange))+
  geom_bar(position = 'dodge') +
  labs(x='Clusters',y='Frequency')
```



**Interpretation:**

- Cluster-1: Companies are listed in all three exchanges and do business in USA and Germany. Recommendation is to Hold more companies and Moderate buy on some.
- Cluster-2: All companies are listed in NYSE and do business in Canada and USA. Recommendation is to hold half and moderate buy half companies.
- Cluster-3: All companies are listed in NYSE and do business in Switzerland, UK and US. Recommendation is mixed but most are recommended to be hold.
- Cluster- 4: All companies are listed in NYSE and do business in UK and US. Recommendation is to Moderate buy half and hold half.
- Cluster-5: All companies are listed in NYSE and do business in France, Ireland and US. Recommendation is to Moderate buy half and Moderate sell half.

**Q:Provide an appropriate name for each cluster using any or all of the variables in the dataset.**

- Cluster-1: Low cap Highly-Volatile companies.(Because of high Beta value and low profits and small market cap)
- Cluster-2: Small cap overpriced companies.(Because of higher PE ratio and smaller market cap)
- Cluster-3: Mid cap Profitable companies. (Most companies has above than average profits and have average market cap)
- Cluster-4: Large-cap Under-priced companies. (High Market cap and Lower than average PE ratio but all financials seem good)

- Cluster-5: Small cap Less-Profitable companies ( Smaller market cap and Profits are lower than average)