Assignment_4_FML

Aditya Ashish Kulkarni

2024-03-16

Directions

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. For each firm, the following variables are recorded:

1. Market capitalization (in billions of dollars) 2. Beta 3. Price/earnings ratio 4. Return on equity 5. Return on assets 6. Asset turnover 7. Leverage 8. Estimated revenue growth 9. Net profit margin 10. Median recommendation (across major brokerages) 11. Location of firm's headquarters 12. Stock exchange on which the firm is listed

---

Reafing CSV file

```
library(readr)
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:

A: 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.

Removing Non Numeric values...

```
#3rd column to 11th column has numeric values only
Pharma.df <- df[,c(3:11)]
```

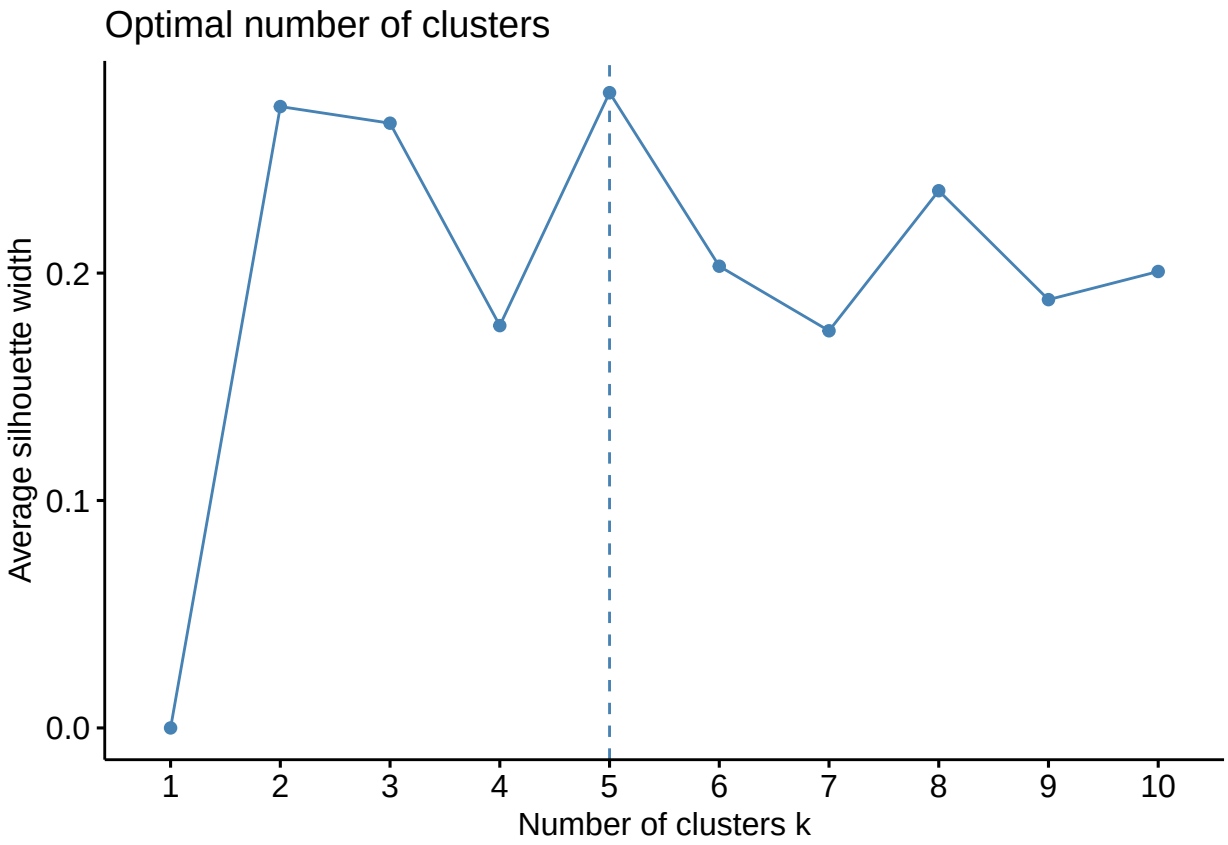
Need to Normalize the data

```
#using Pre process
normal=preProcess(Pharma.df, method = c("center", "scale"))
pharm.df.norm=predict(normal, Pharma.df)
```

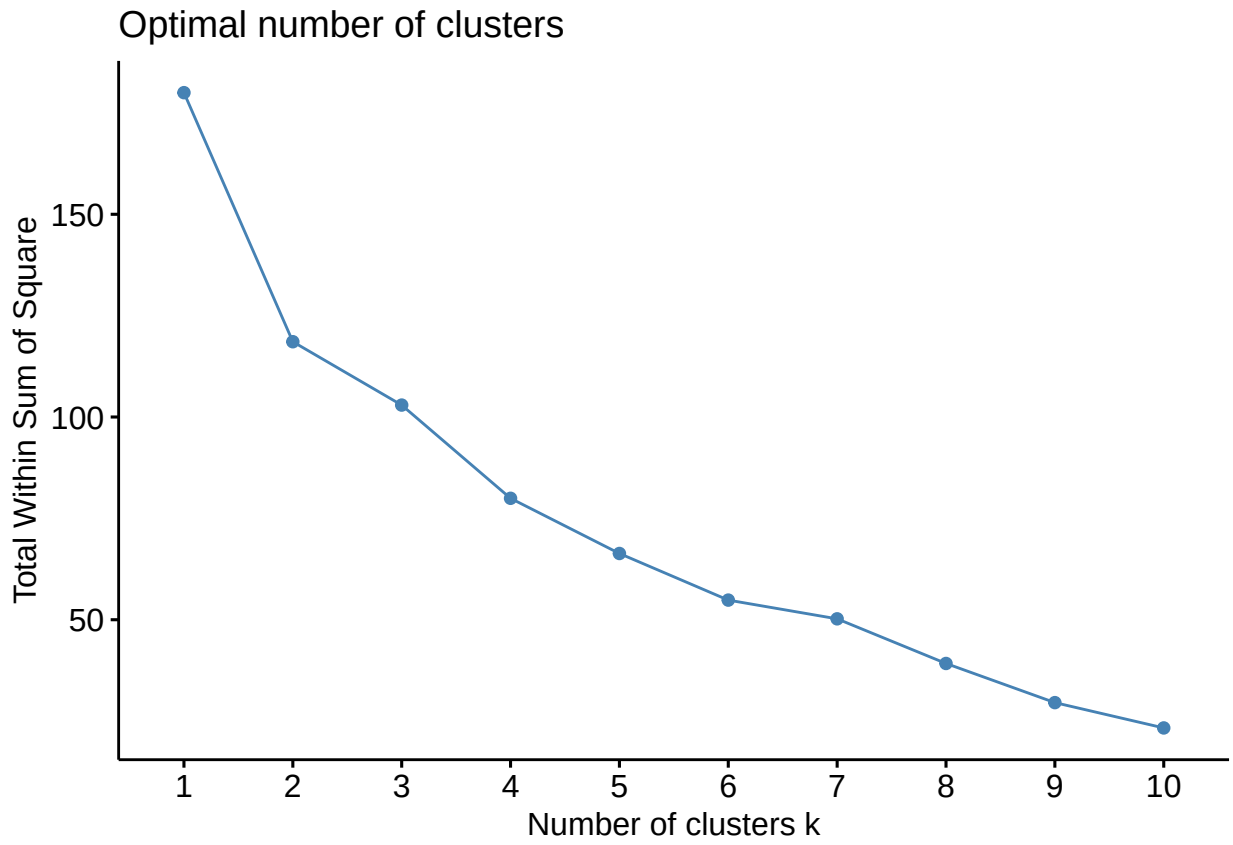
K Means Method

Using silhouette and wss method and Creating graph

```
#elbow method to calculate the value of K
fviz_nbclust(pharm.df.norm, kmeans, method = "silhouette")
```




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



From above two graphs, K=5 is the optimum value and is over fitting and bias into consideration...

So, Now we need to apply K-means Clustering method on this

```
#center wil be 5 and start 10
k=kmeans(pharm.df.norm,centers=5,nstart = 10)
```

```
#following 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
```

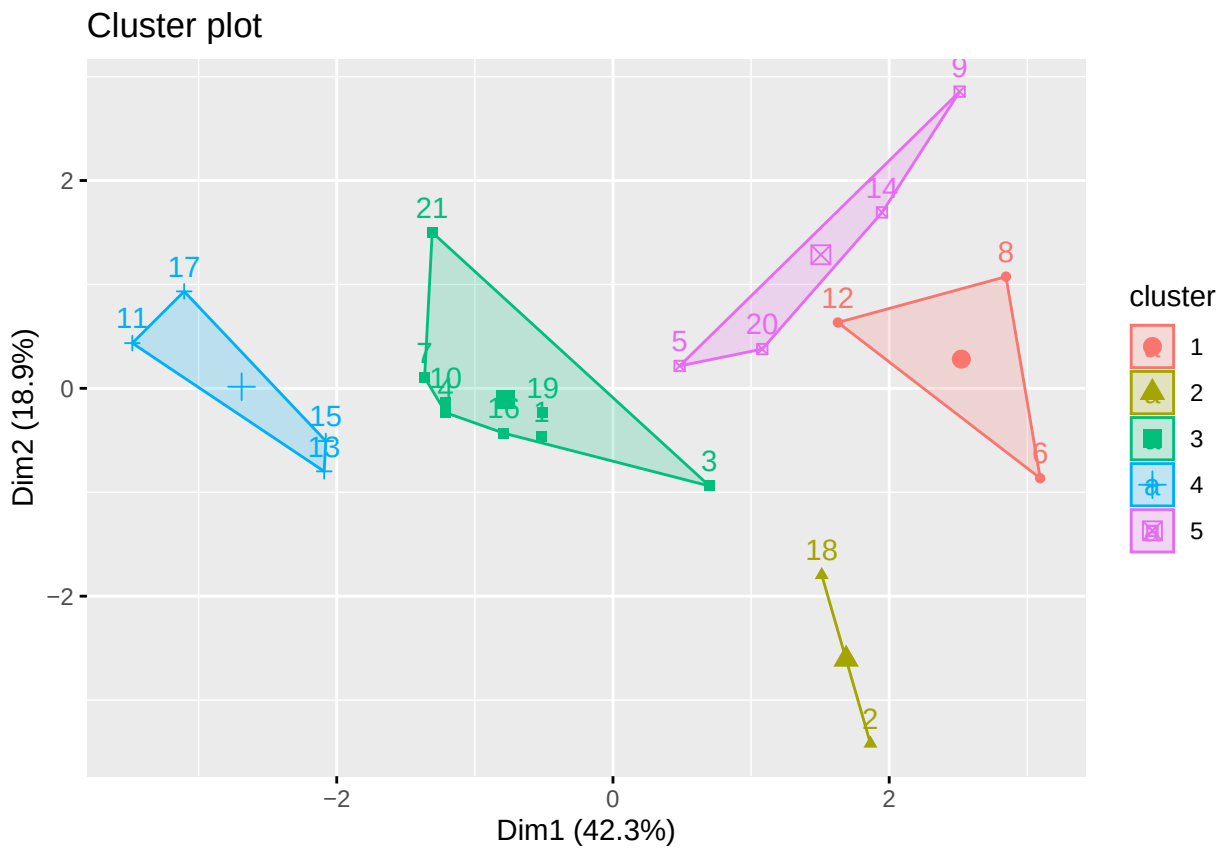
```
#calculating size of cluster

k$size
```

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

Visualizing Cluster

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



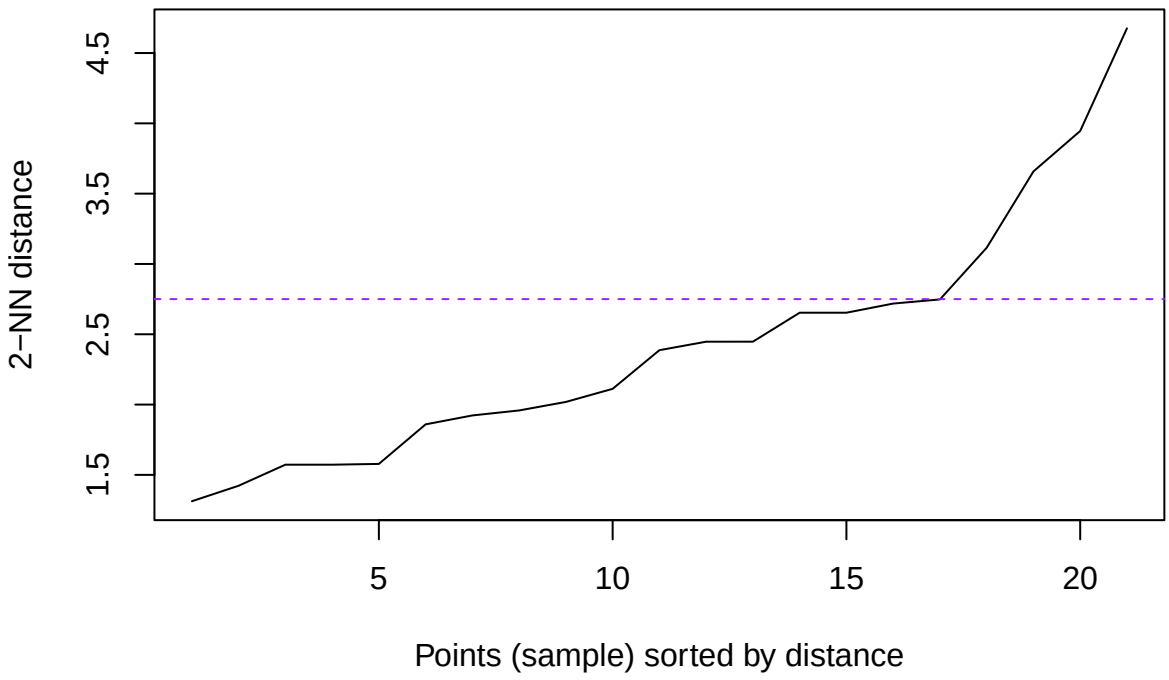
Interpretation:

- The K-means plot provides the best possible image of the clusters. Every point that is near another point is part of the same cluster. Furthermore, the structure of the supplied may be easily studied using this type of uniform cluster plot.

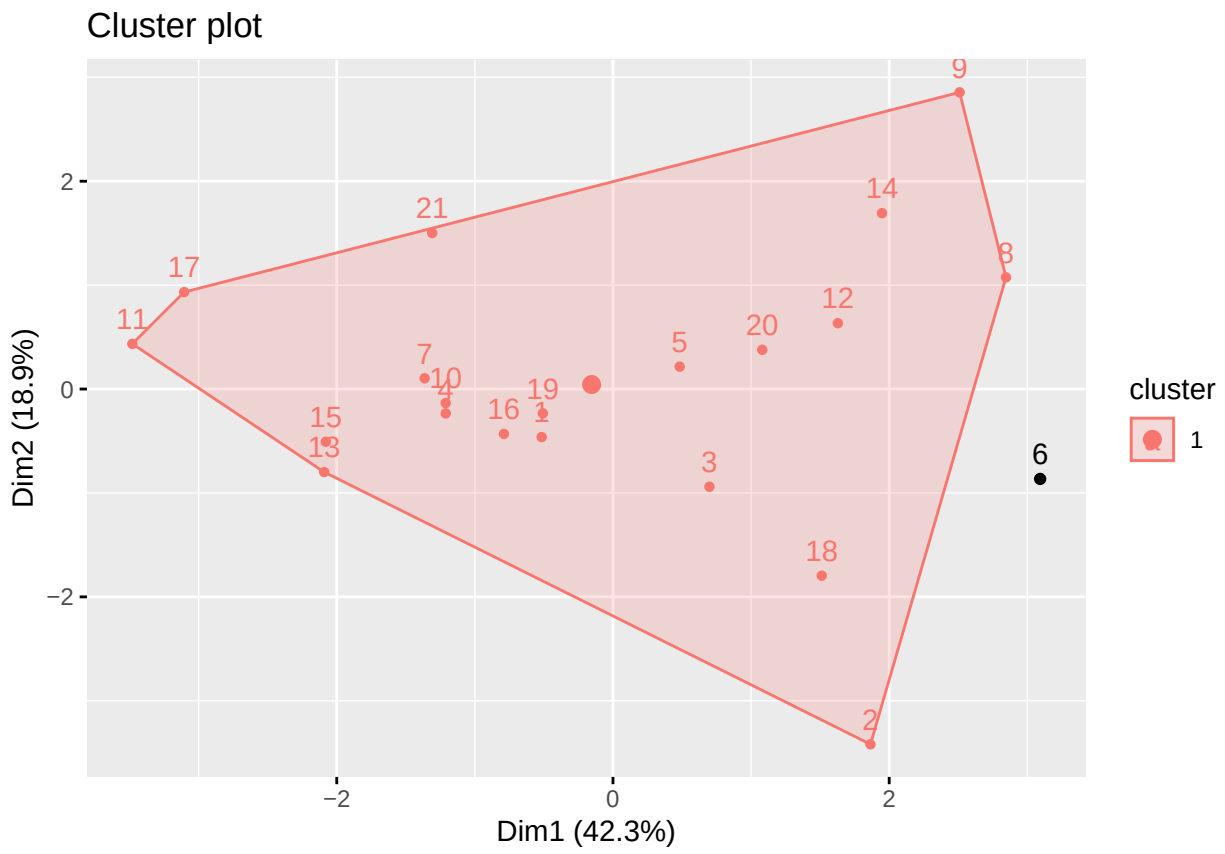
DBSCAN Method

Finding Optimum value of eps when K=5

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




```
#
db= dbscan::dbscan(pharm.df.norm,eps=3.5,minPts = 2)
fviz_cluster(db,pharm.df.norm)
```



Interpretation

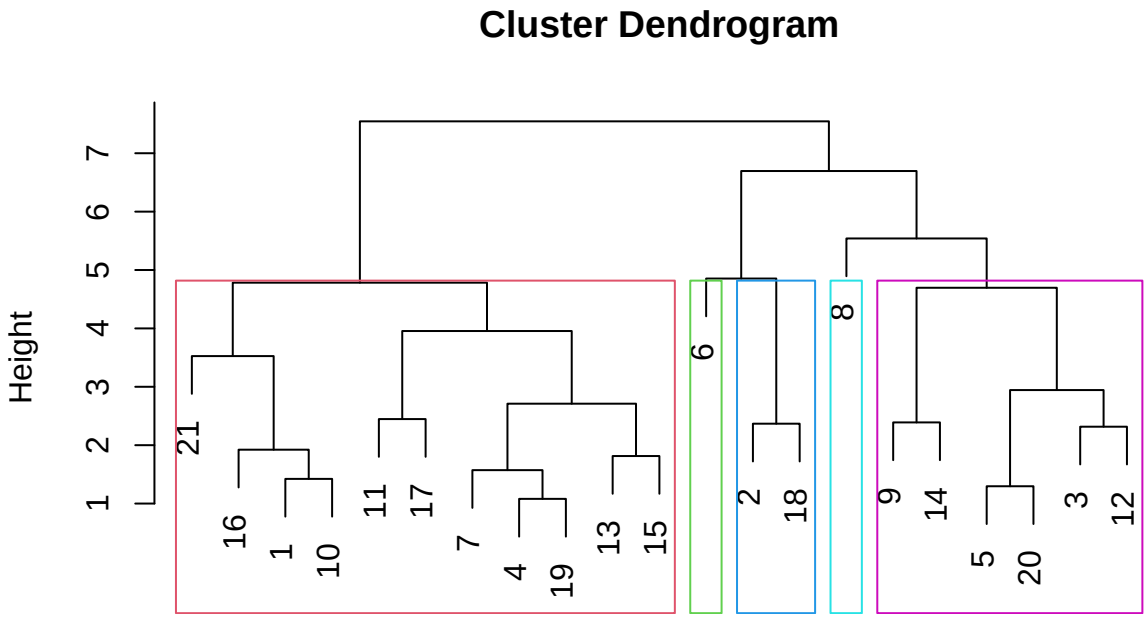
Since the DBSCAN approach retains nearly all of the data points in a single cluster, it appears to be an incorrect method for clustering the provided data. Additionally, the majority of the points remain outliers if a lower value of eps is used. DBSCAN is therefore not the best technique for the available data.

Note: Here, we can also plot the hierarchical cluster graph to determine the interpretations strongly...

Hierarchical Method

```
#Applying hclut directly on cluster
#Using euclidean

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



Interpretations:

Although hierarchical clustering appears to provide a lovely depiction of clusters, when we attempt to create five clusters similar to k-means using it, it exhibits some outliers that, when compared to the k-means plot, are actually near to some other points and should be included in a cluster with other points. Furthermore, financial data lacks structure, therefore conducting a hierarchical clustering makes little sense.In terms of grouping the data into clusters, this approach appears to lag behind the k-means method as well.

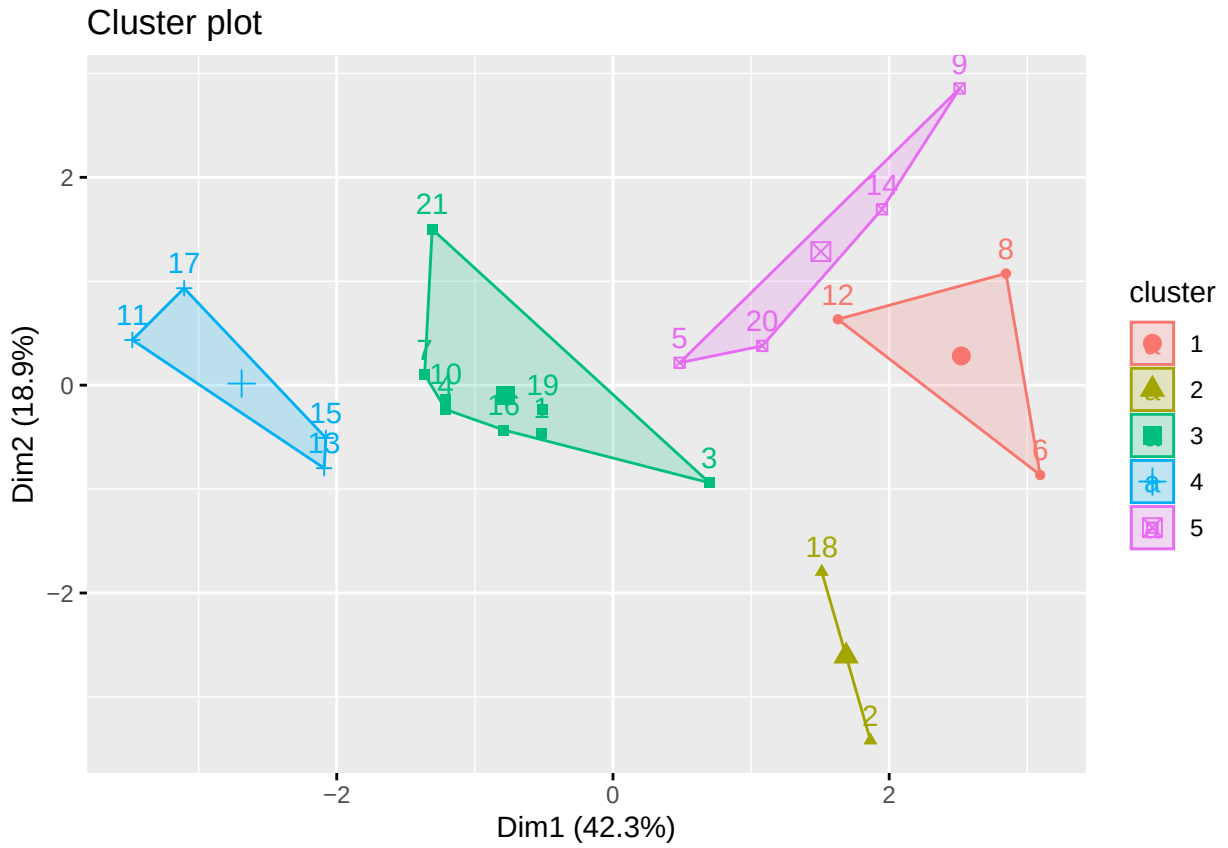
Justification of choices:

1. We have normalized the data and assigned equal weight to each variable since the weighting of the variables is not in doubt.
2. After experimenting with several clustering techniques, I discovered that the k-means algorithm yields the best results for the given data since it produces a more accurately grouped image where all of the dots that are closest to one another are in the same cluster.
3. To determine the required number of clusters, we have employed techniques such as the elbow-method and distplot to figure out the values of k and eps. In order to compare the hierarchical plot with K-means clustering, we also maintained the number of clusters at five.

B: Interpret the clusters with respect to the numerical variables used in forming the clusters.

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

```
#forming Cluster Plot
fviz_cluster(k,data = pharm.df.norm)
```



The numerical values of the points in the same k-means clusters are closer to one another than the numerical values of the points in separate groups, with regard to the numerical values employed in the clustering. In order to better understand, we may view these clusters.

Cluster 1:

```
pharm.df.norm[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
```

The average PE-ratio and high beta values of this cluster are accompanied by below-average values for all other variables.Leverage and Rev_growth, however, have different values.

Cluster 2:

```
pharm.df.norm[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
```

The PE ratio of this cluster is high, while all other variables are below average. Nevertheless, leverage and rev_growth have varied values in this cluster as well.

Cluster 3:

```
pharm.df.norm[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
```

The Net_profit values of this cluster are extremely high. Other variables, nevertheless, have inconsistent values.

Cluster 4:

```
pharm.df.norm[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
```

Every business inside this cluster has a high value. In addition, leverage, beta, and PE ratio.

Cluster 5:

```
pharm.df.norm[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
```

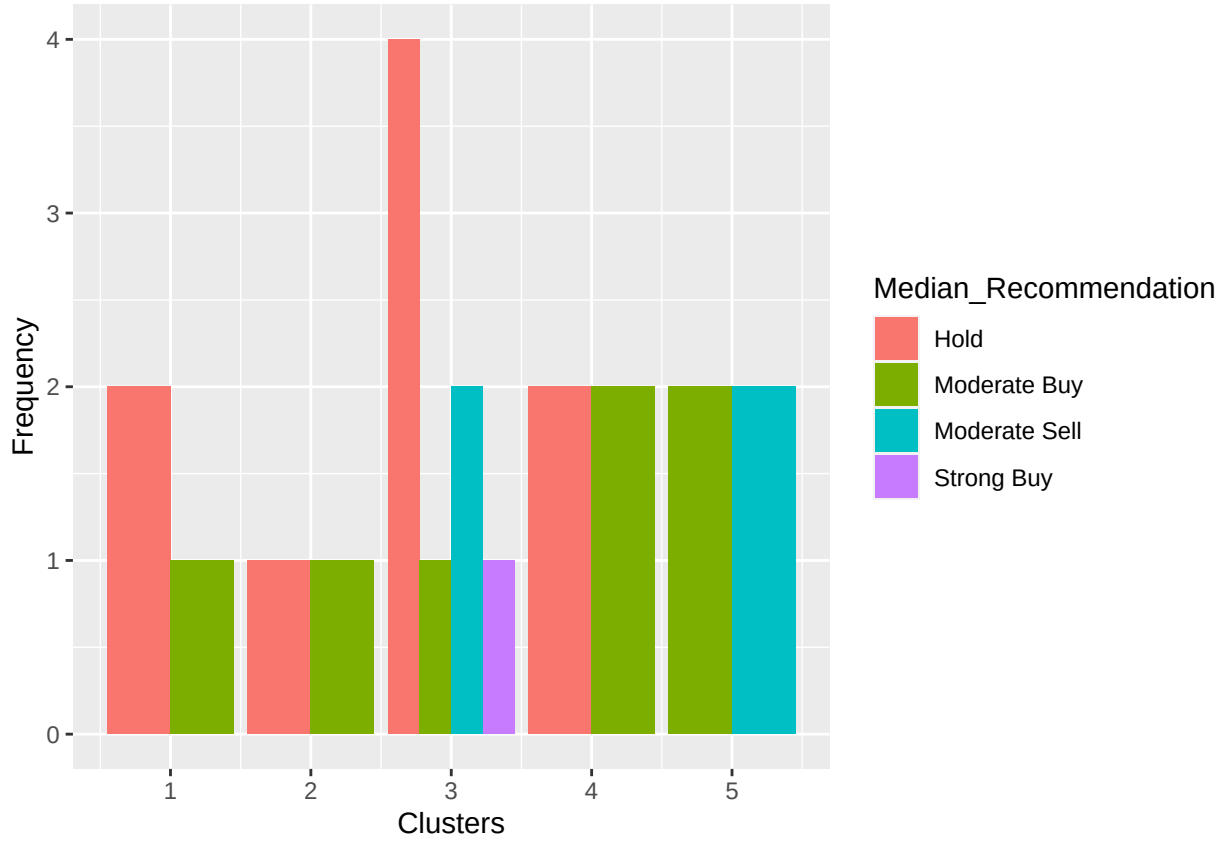
The market_Cap, PE-ratio, ROE, ROA, and net profit for this cluster are all low, but the values for the other metrics are all mixed.

Now we need to make BarPlot to check patterns in variables which we didn't use in clustering

Comparing recommended clusters

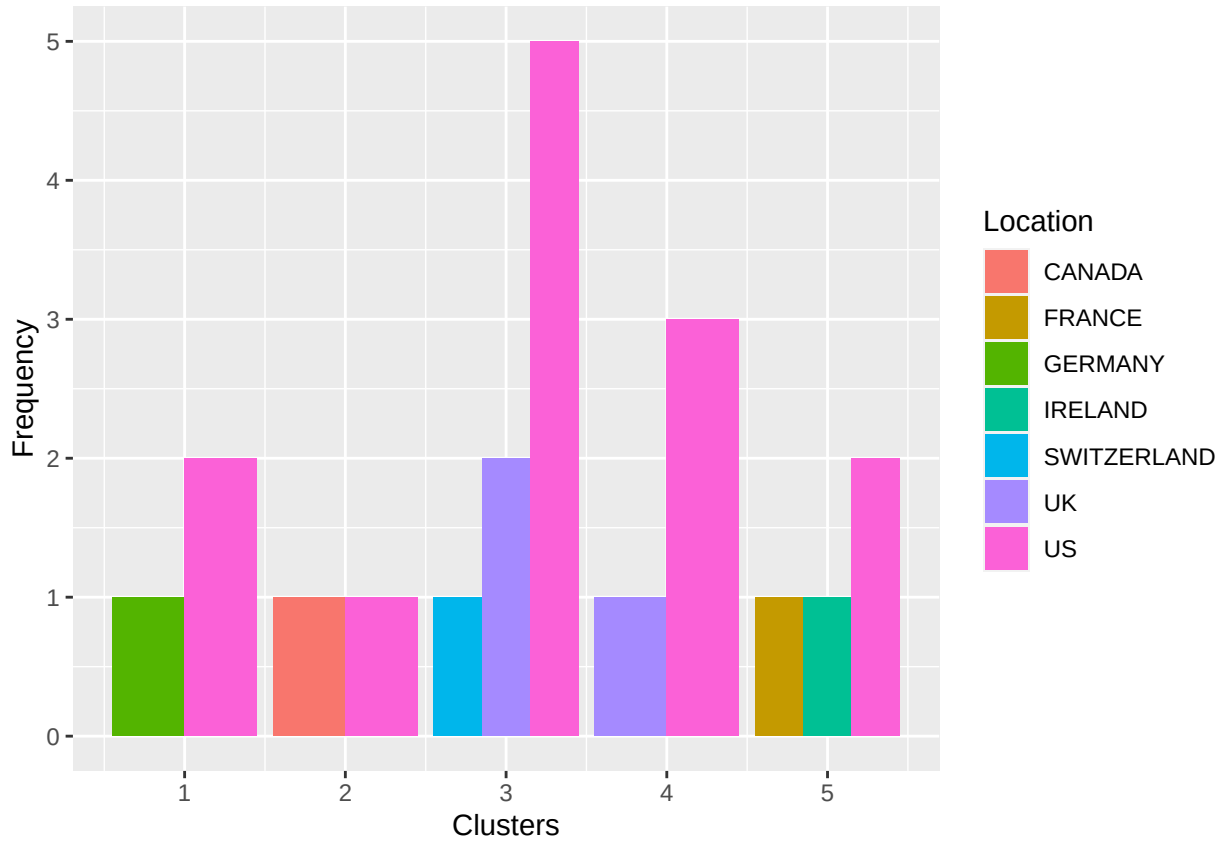
```
df.1= df %>%
  select(c(1,12,13,14)) %>%
  mutate(cluster=k$cluster)

ggplot(df.1,mapping = aes(cluster,fill=Median_Recommendation))+ geom_bar(position = 'dodge') + labs(x='
```



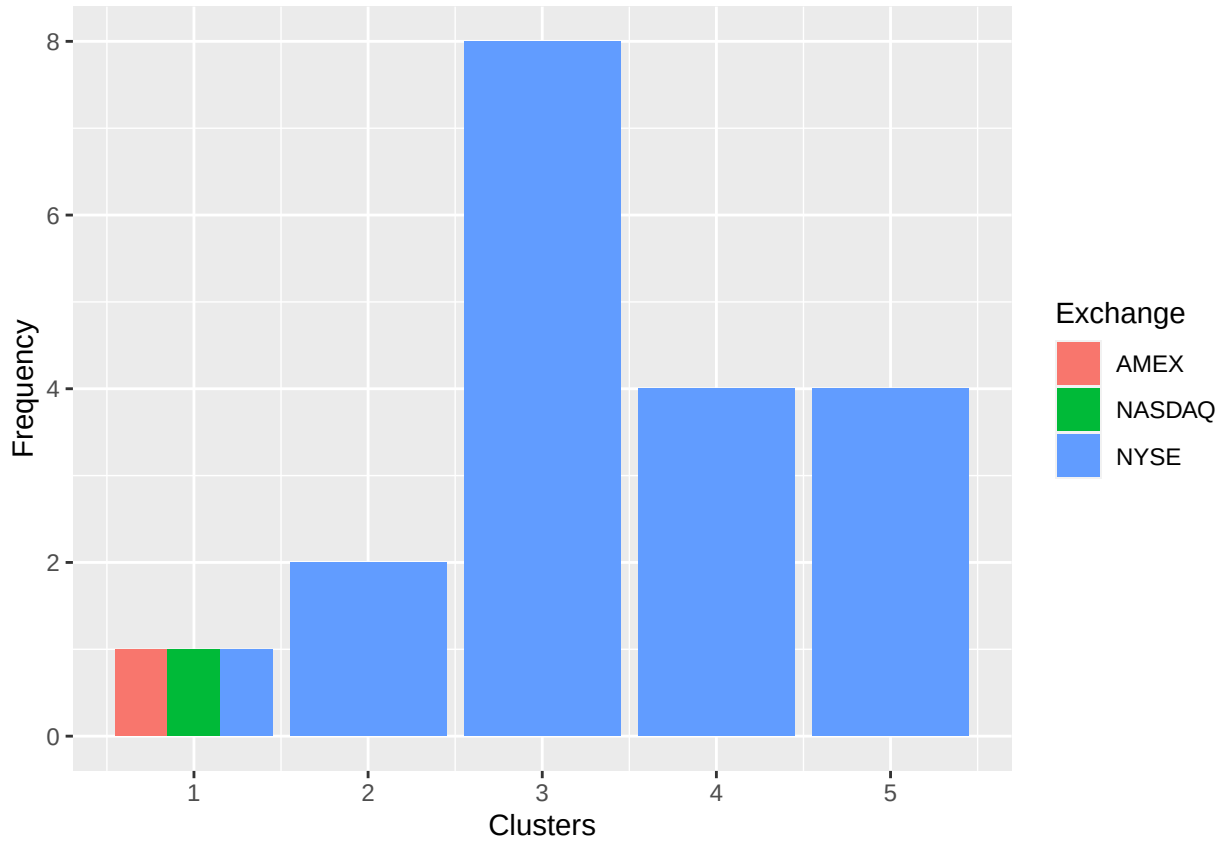
Now Let's compare countries of cluster

```
ggplot(df.1,mapping = aes(cluster,fill=Location))+
  geom_bar(position = 'dodge') +
  labs(x='Clusters',y='Frequency')
```



let's Compare Stock-Exchange of cluster

```
ggplot(df.1,mapping = aes(cluster,fill=Exchange))+
  geom_bar(position = 'dodge') +
  labs(x='Clusters',y='Frequency')
```



Interpretation from above clusters and data:

- **Cluster 1:** Businesses with listings on all three exchanges conduct business in Germany and the United States. Holding more firms and making moderate purchases of some is advised.
- **Cluster 2:** Every company operates in Canada and the USA and is listed on the NYSE. Holding half the firms and buying the other half in moderation is advised.
- **Cluster 3:** All businesses operate in the US, UK, and Switzerland and are listed on the NYSE. There are differing recommendations, but the majority should be held.
- **Cluster- 4:** Every company operates in the US and the UK and is listed on the NYSE. It is advised to purchase half and hold the other half moderately.
- **Cluster 5:** All businesses operate in France, Ireland, and the US and are listed on the NYSE. It is advised to purchase half and sell half moderately.

D. Provide an appropriate name for each cluster using any or all of the variables in the data set.

- Cluster 1: Low ceiling Very erratic businesses.(Due to the modest market capitalization, minimal profitability, and high beta value.)
- Cluster 2: Overpriced small-cap firms.(Due to a lower market capitalization and a higher PE ratio.)

- Cluster 3: The middle cap profitable businesses. (The majority of corporations have average market caps and above-average earnings.)
- Cluster 4: Under-priced, large-cap enterprises. (All financials appear strong, despite a high market value and a lower than normal PE ratio.)
- Cluster 5: Less profitable small-cap firms (profits are less than average and the market cap is less).