BA64060_Assignment4

Ruthvick Bulagakula

2023-11-10

Calling Installed Libraries

```
library(cluster)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(flexclust)
```

```
## Warning: package 'flexclust' was built under R version 4.3.2

## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4
```

```
library(factoextra)
```

```
## Loading required package: ggplot2

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

```r
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.3.2
```

```r
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.3.2
```

```r
library(dbscan)
```

```
## Warning: package 'dbscan' was built under R version 4.3.2
```

```
##
## Attaching package: 'dbscan'
```

```
## The following object is masked from 'package:stats':
##
##     as.dendrogram
```

Reading Data

By using read.csv function we can read data in the csv file.

```r
data= read.csv("pharmaceuticals.csv")
filter_data = data[3:11]
head(filter_data)
```

```
##   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
## 1              16.1
## 2               5.5
## 3              11.2
## 4              18.0
## 5              12.9
## 6               2.6
```

Reason for Choosing Market_Cap, Beta, PE_Ratio, ROE, ROA, Asset_Turnover, Leverage, Rev_Growth, and Net_Profit_Margin

The selected variables Market_Cap, Beta, PE_Ratio, ROE, ROA, Asset_Turnover, Leverage, Rev_Growth, and Net_Profit_Margin are common financial metrics used to evaluate and compare the performance of companies. These variables collectively provide a comprehensive overview of a firm's financial health, profitability, and efficiency.

1. Market_Cap:

Ranges from 0.41 to 199.47. Indicates the overall size and valuation of the pharmaceutical firms.

2. Beta:

Ranges from 0.18 to 1.11. Measures the sensitivity of a firm's returns to market fluctuations.

3. PE_Ratio:

Ranges from 3.6 to 82.5. Represents the valuation of a firm's stock relative to its earnings.

4. ROE:

Ranges from 3.9 to 62.9. Indicates how effectively a firm utilizes shareholder equity to generate profit.

5. ROA:

Ranges from 0.3 to 1.1. Measures a firm's ability to generate profit from its assets.

6. Asset_Turnover:

Ranges from 0.5 to 1.1. Represents how efficiently a firm utilizes its assets to generate revenue.

7. Leverage:

Ranges from 0 to 3.51. Reflects the extent to which a firm uses debt to finance its operations.

8. Rev_Growth:

Ranges from -3.17 to 34.21. Indicates the percentage change in revenue over a specific period.

9.Net_Profit_Margin:

Ranges from 2.6 to 25.54. Represents the percentage of revenue that turns into profit.

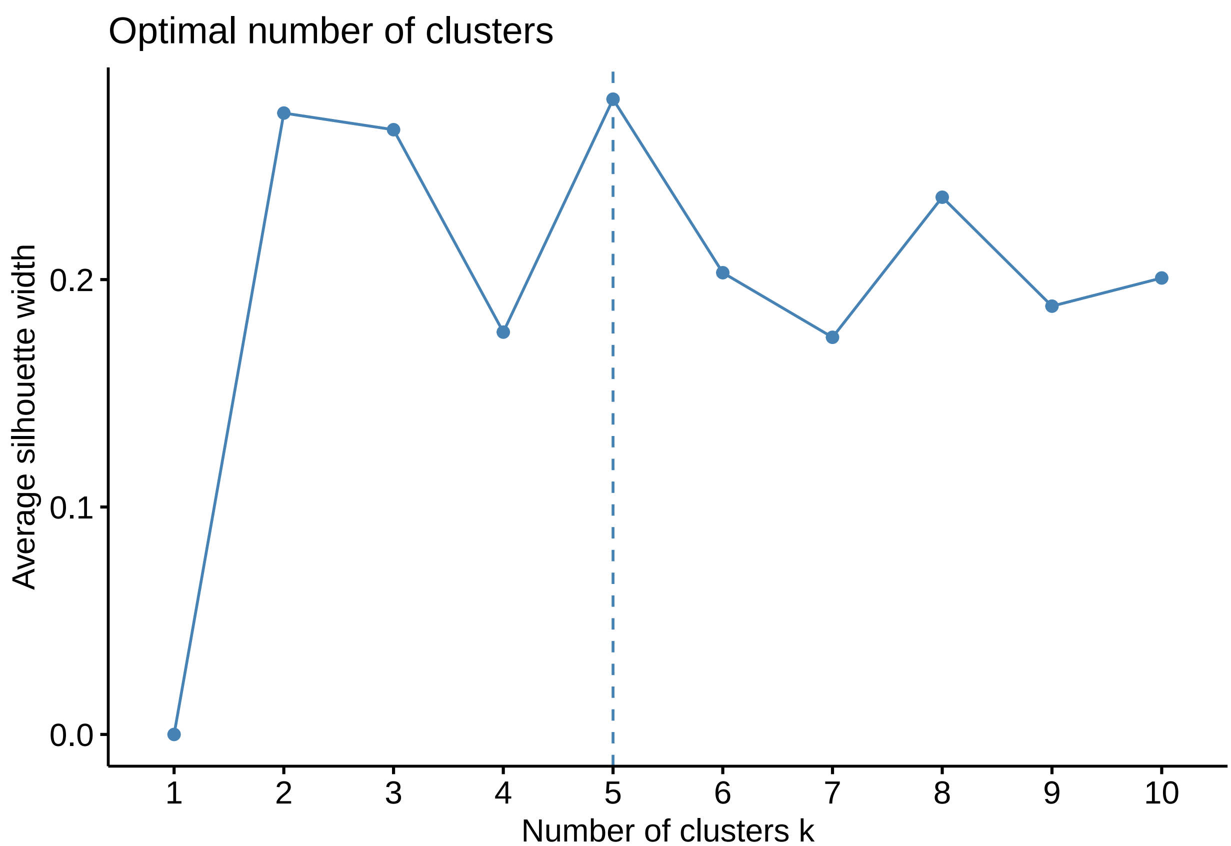
Normalising Data

```r
norm_data = scale(filter_data)
row.names(norm_data) = data[,1]
distance = get_dist(norm_data)
corr = cor(norm_data)
```

Reason for Normalization

Normalization of the numerical variables is crucial to ensure that each variable contributes proportionally to the clustering process. Since these variables may have different units or scales, normalizing them helps prevent one variable from dominating the clustering based on its magnitude. For example, Market_Cap is in the hundreds, while Beta is a fraction between 0 and 1.

```r
fviz_nbclust(norm_data, kmeans, method = "silhouette")
```



Reason for choosing 5 clusters

Silhouette analysis measures how similar an object is to its own cluster compared to other clusters. It provides a graphical representation of the quality of clusters for different values of k.

```
set.seed(1)
k5 = kmeans(norm_data, centers = 5, nstart = 25)
k5$centers
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 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 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
##      Leverage Rev_Growth Net_Profit_Margin
## 1  0.06308085  1.5180158      -0.006893899
## 2 -0.14170336 -0.1168459      -1.416514761
## 3 -0.27449312 -0.7041516       0.556954446
## 4  1.36644699 -0.6912914      -1.320000179
## 5 -0.46807818  0.4671788       0.591242521
```

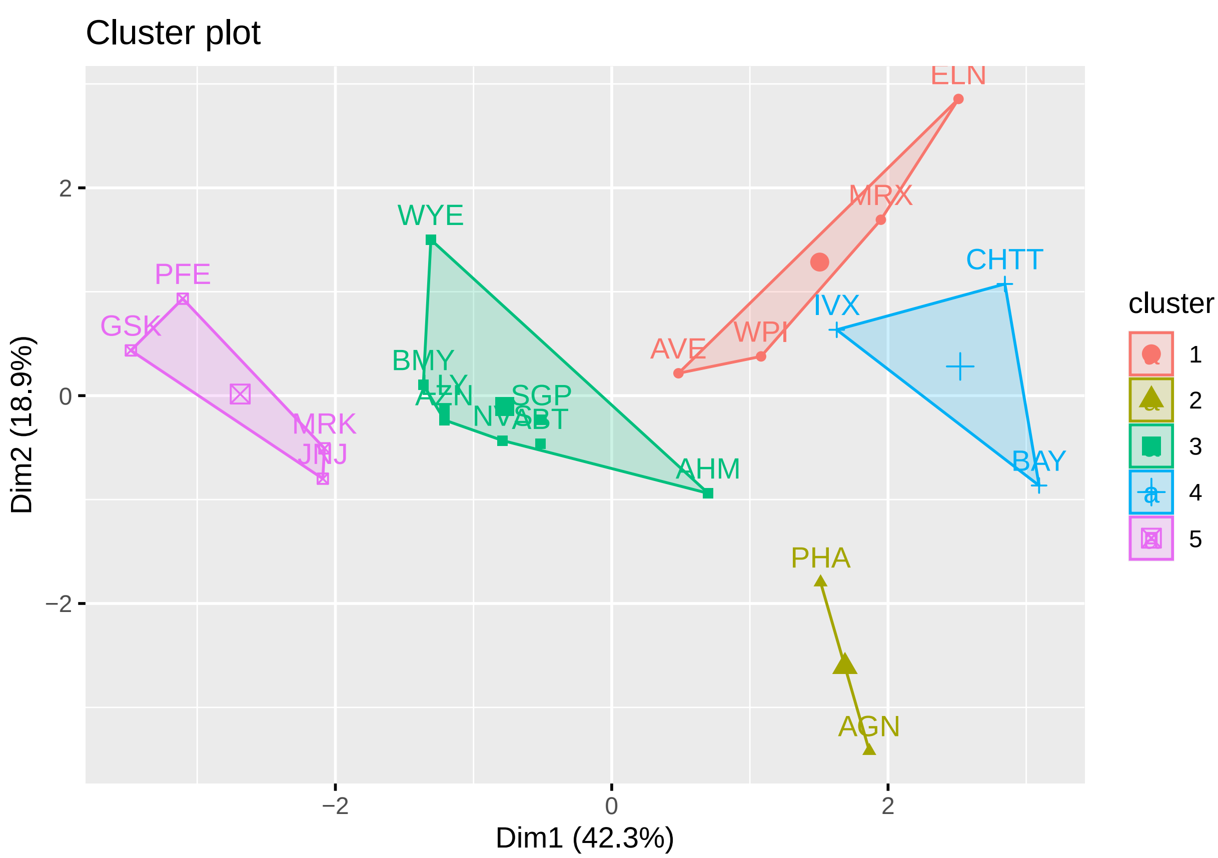
Reason for selecting K-means

Reason why i'm selecting K-means over DBSCAN is that, K-means is often used in exploratory data analysis to identify patterns and groupings within the data, K-means clustering can provide insights into the financial profiles of pharmaceutical firms. It may reveal groups of firms with similar financial characteristics, aiding in strategic decision-making or investment analysis, easy to interpret, and DBSCAN is effective for datasets with dense regions.

```
k5$size
```

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

```
fviz_cluster(k5, data = norm_data)
```



Appropriate Name:

Cluster 1 - Profitable Ventures

Cluster 2 - Risk-Reward Seekers

Cluster 3 - Stable Giants

Cluster 4 - Beta Boosted Enterprises

Cluster 5 - Market Dominators

Interpretation of Clusters based on Variables used forming Clusters:

Cluster 1: AVE, WPI, MRX, ELN exhibit moderate values across Market_Cap, Beta, PE_Ratio, ROE, ROA, Asset_Turnover, Leverage, Rev_Growth, and Net_Profit_Margin.

Cluster 2: PHA, AGN exhibit lower Market_Cap, Beta, and PE_Ratio.

Cluster 3: WYE, BMY,AZN, SGP, AHM, LLY,NVS, ABT exhibit higher Market_Cap, Beta, PE_Ratio, Rev_Growth, and Net_Profit_Margin compared to other clusters.

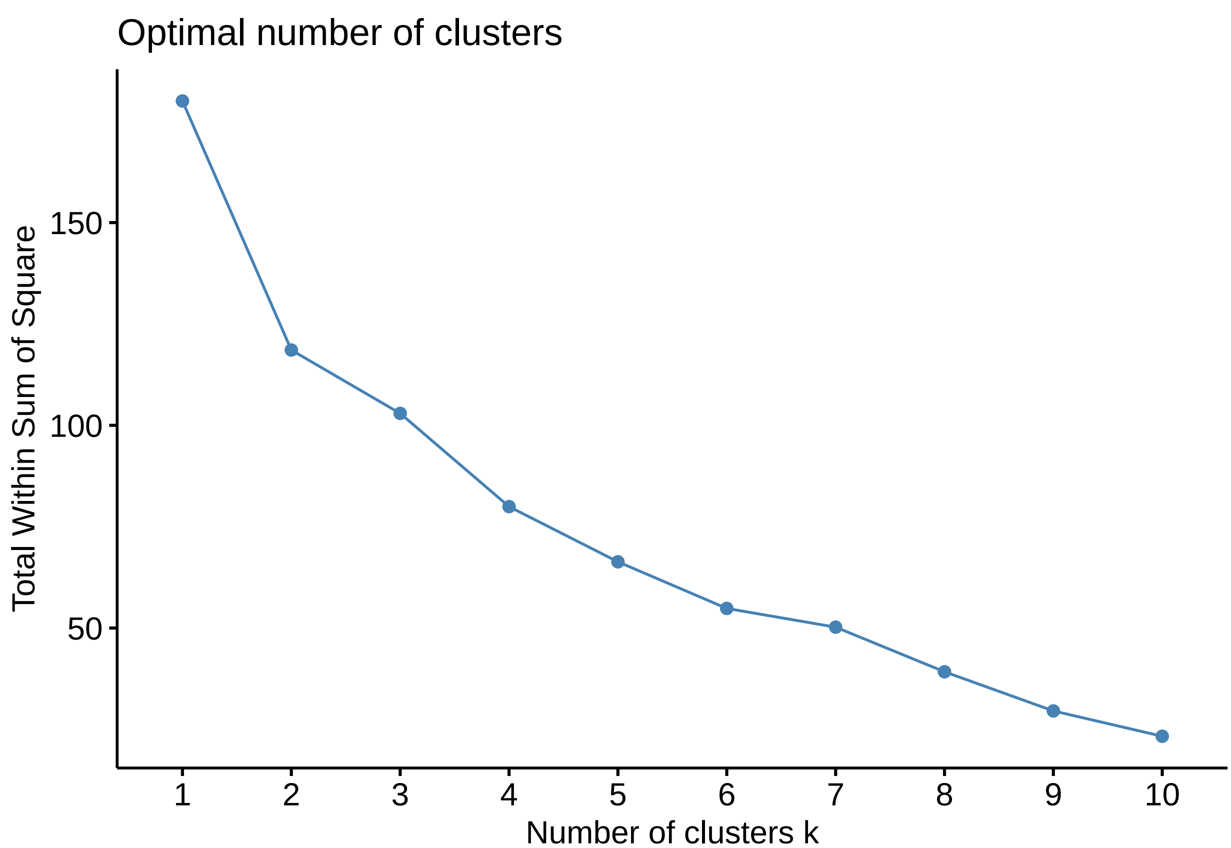
Cluster 4: IVX, CHTT, BAY exhibit lower Market_Cap and PE_Ratio.

Cluster 5: GSK, PFE, MRK, JNJ exhibit higher values across Market_Cap, Beta, PE_Ratio, ROE, ROA, Asset_Turnover, Rev_Growth, and Net_Profit_Margin.

Elbow

The elbow method is used to determine the optimal number of clusters (k) in a k-means clustering analysis

```
fviz_nbclust(norm_data, kmeans, method = "wss")
```



The choice of k=5 suggests that there is a reasonable balance between model complexity and the ability to explain variance in the data. Beyond k=5, the reduction in within-cluster sum of squares becomes less significant, indicating diminishing returns in terms of explaining the variance.

Manhattan

It measures the sum of absolute differences between coordinates

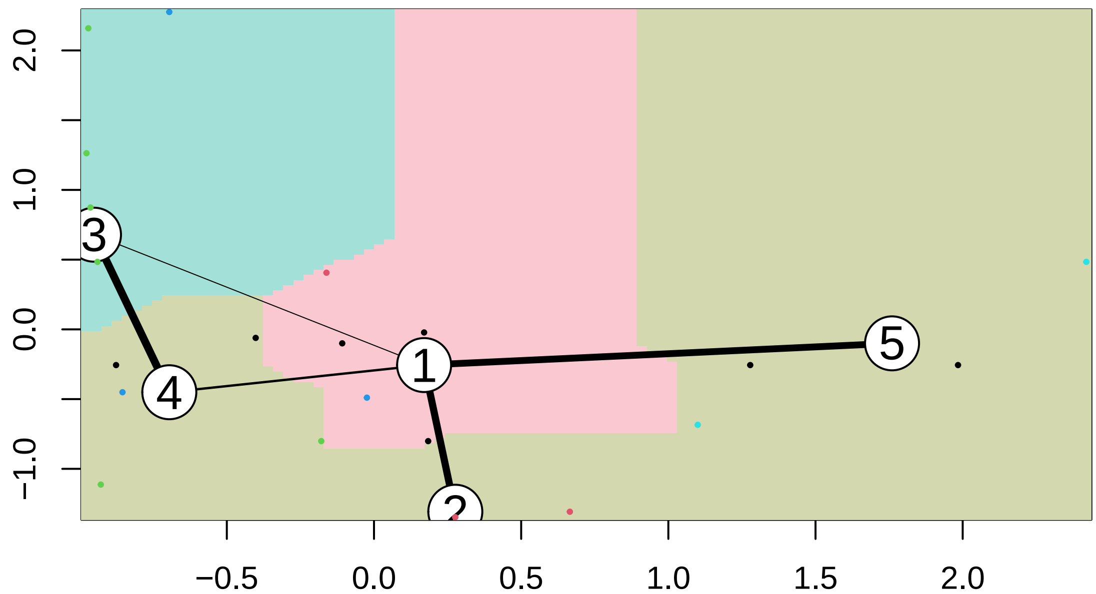
```
set.seed(1)
k51 = kcca(norm_data, k=5, kccaFamily("kmedians"))
k51
```

```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = norm_data, k = 5, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
## 1 2 3 4 5
## 7 3 6 3 2
```

```
clusters_index = predict(k51)
dist(k51@centers)
```

```
##          1        2        3        4
## 2 2.150651
## 3 3.513242 4.146567
## 4 3.878726 4.246051 3.388339
## 5 3.018500 3.737739 5.124420 6.043691
```

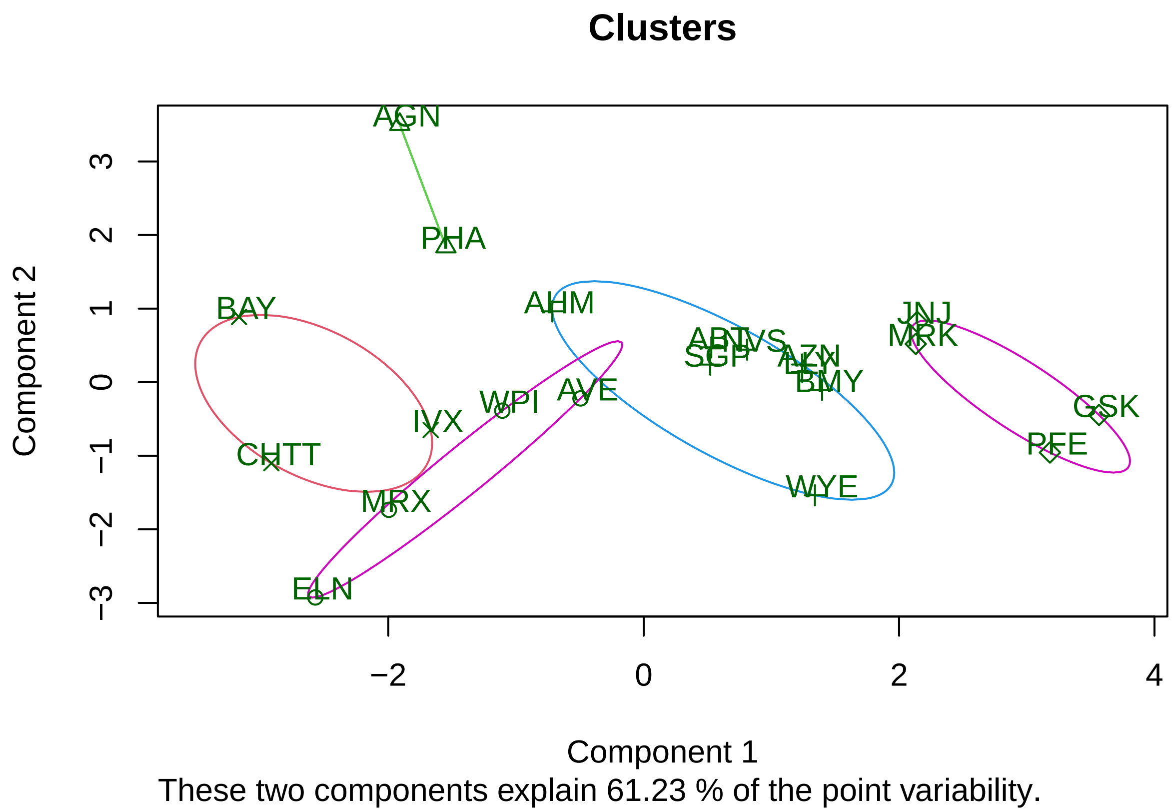
```
image(k51)
points(norm_data, col=clusters_index, pch=19, cex=0.3)
```



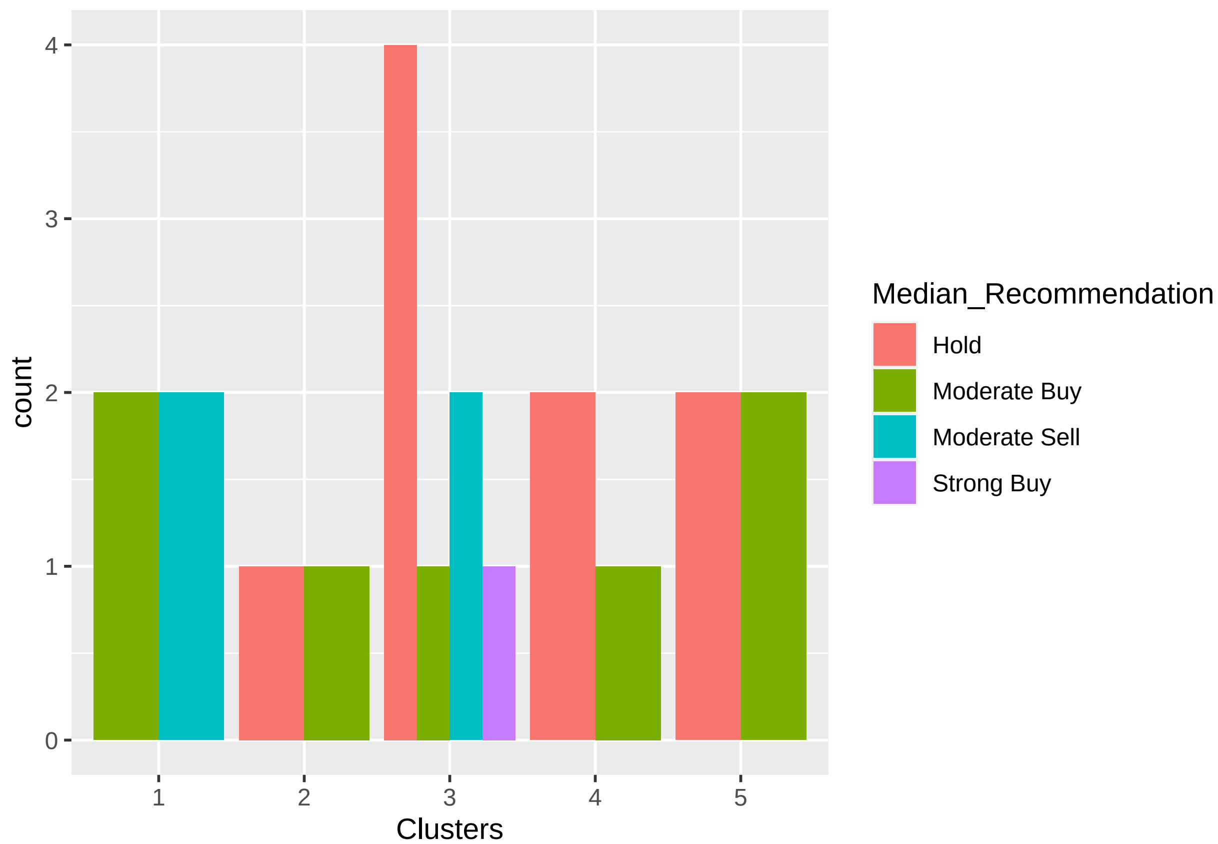

```r
filter_data %>% mutate(Cluster = k5$cluster) %>% group_by(Cluster) %>% summarise_all("mean")
```

```
## # 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      13.1  0.598     17.7  14.6  6.2           0.425    0.635
## 2       2      31.9  0.405     69.5  13.2  5.6           0.75     0.475
## 3       3      55.8  0.414     20.3  28.7 12.7           0.738    0.371
## 4       4       6.64 0.87      24.6  16.5  4.17          0.6      1.65
## 5       5     157.   0.48      22.2  44.4 17.7           0.95     0.22
## # i 2 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>
```

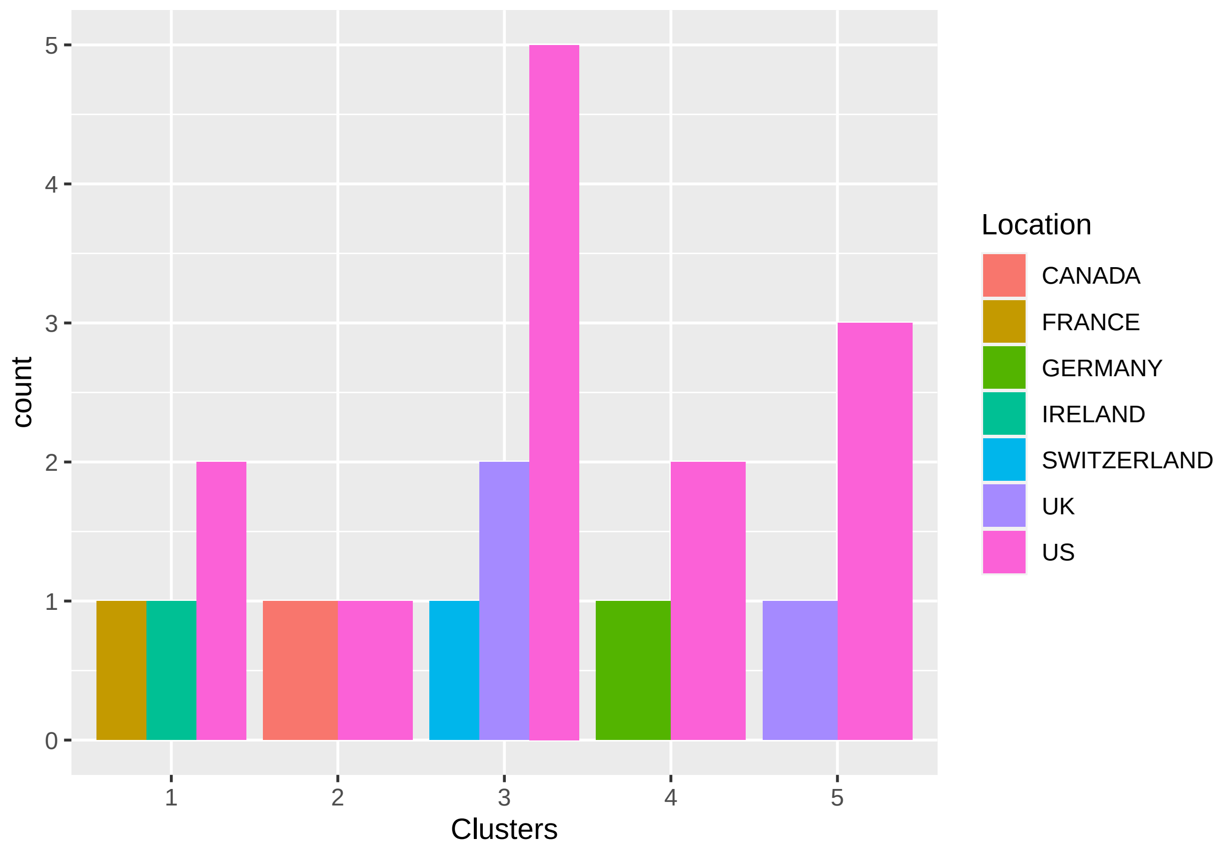
```r
clusplot(norm_data,k5$cluster, main="Clusters",color = TRUE, labels = 3,lines = 0)
```



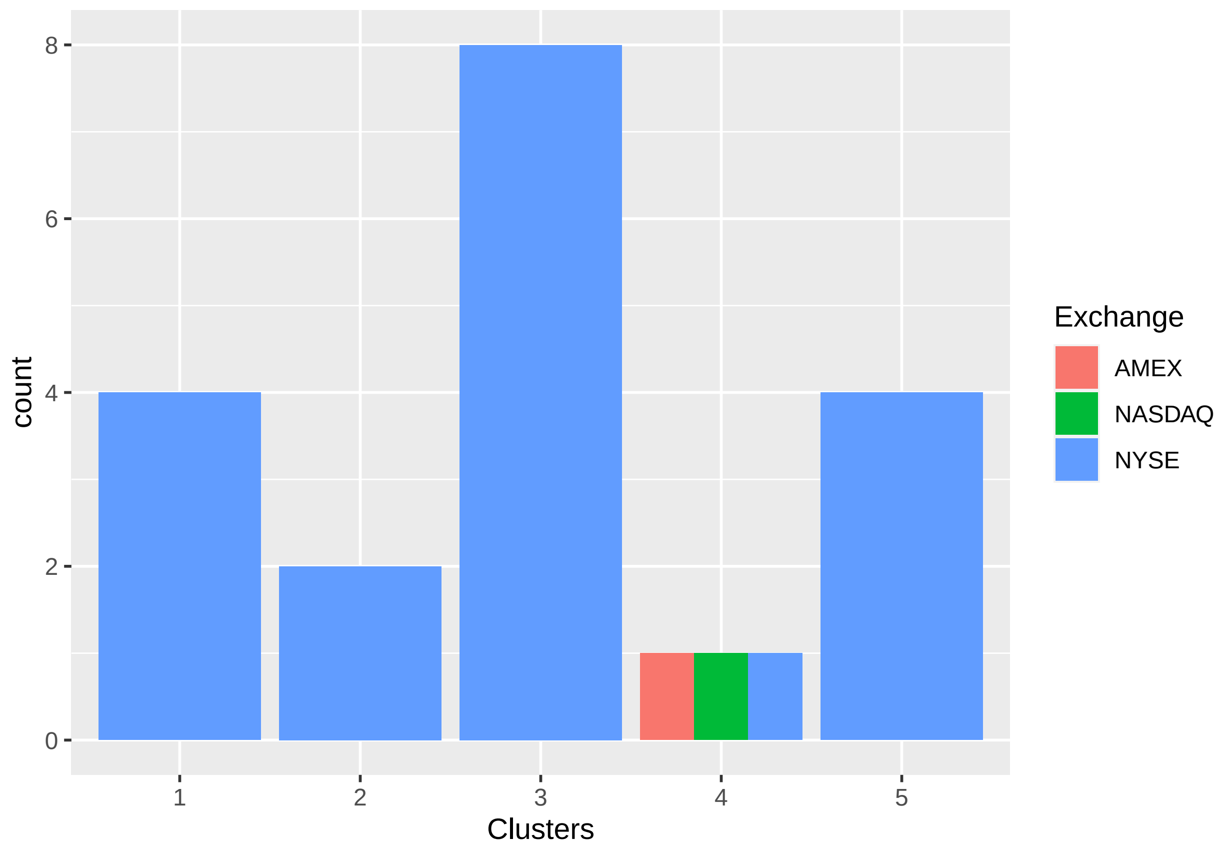

```r
clust_data = data[12:14] %>% mutate(Clusters=k5$cluster)
ggplot(clust_data, mapping = aes(factor(Clusters),
fill =Median_Recommendation))+geom_bar(position='dodge')+labs(x ='Clusters')
```




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




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



Interpretation of Clusters based on Variables 10 to 12:

Cluster 1:

Median Recommendation: Cluster 1 has moderate buy and moderate sell.

Location: Cluster 1 has three Locations,in which US is the highest.

Exchange: Cluster 1 has only one exchange that is NYSE.

Cluster 2:

Median Recommendation: Cluster 2 has low hold and low buy

Location: Cluster 2 has only two locations(US and Canada) and evenly distributed .

Exchange: Cluster 2 has only one exchange that is NYSE.

Cluster 3:

Median Recommendation: Cluster 3 very strong hold and high moderate sell

Location: Cluster 3 has three locations, US has more numbers, then UK and Switzerland

Exchange: Cluster 3 has only one exchange, that is NYSE which is very high in numbers.

Cluster 4:

Median Recommendation: Cluster 4 has strong hold and low buy.

Location: Cluster 4 has two locations in which US is high compared to Germany.

Exchange: Cluster 4 has three exchanges(AMEX, NASDAQ, NYSE) and all of them are evenly distributed.

Cluster 5:

Median Recommendation: Cluster 5 has high hold and high buy.

Location: Cluster 5 has two locations in which US is in large number compared to UK which is very less.

Exchange: Cluster 5 has only one exchange that is NYSE.