Assignment - 4

Vamshee Deepak Goud Katta

11/2/2021

Pharmaceutical Industry - K-Means Clustering

A. Inserting and Cleaning Data

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(factoextra)
```

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

```
library(ISLR)
library(flexclust)
```

```
## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4
```

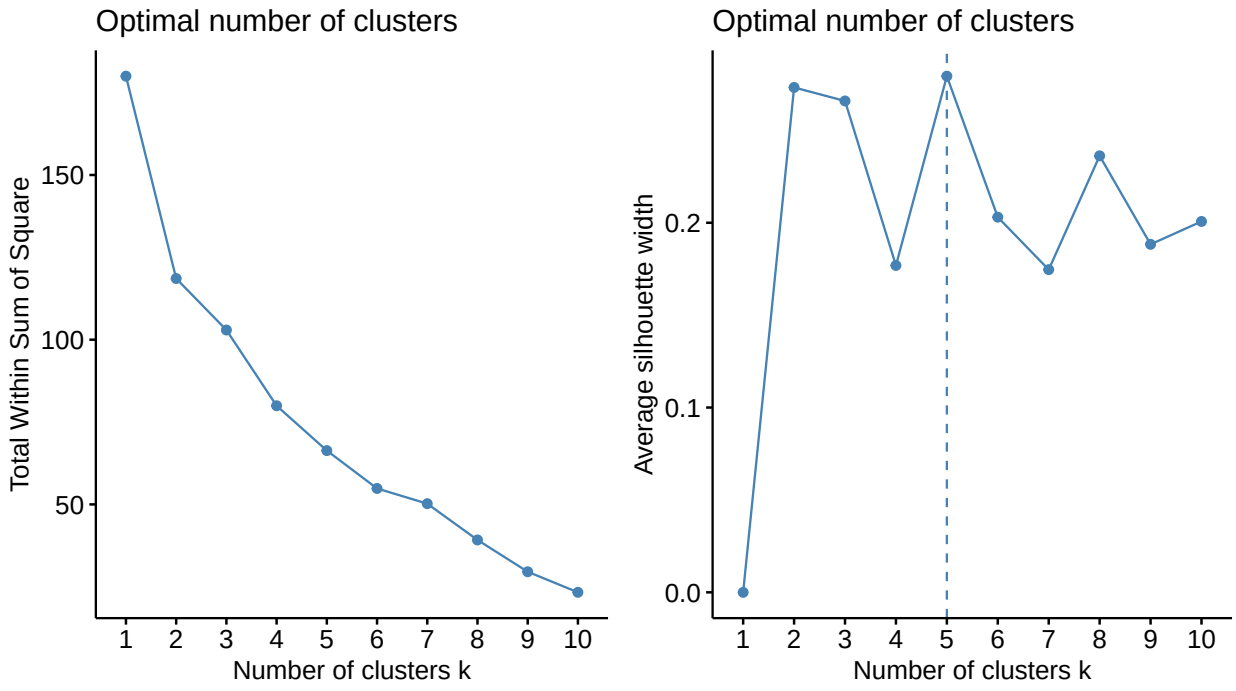
```
library(cowplot)
set.seed(123)

Pharma <- read.csv('Pharmaceuticals.csv')
Pharma1 <- Pharma[,3:11]
summary(Pharma1)
```

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

Scaling the dataframe (Z-score) and finding the optimal number of clusters

```
Pharma1 <- scale(Pharma1)
distance <- get_dist(Pharma1)
wss <- fviz_nbclust(Pharma1, kmeans, method = "wss")
silhouette <- fviz_nbclust(Pharma1, kmeans, method = "silhouette")
plot_grid(wss, silhouette)
```



Clustering the data and plotting the clusters

```
k5 <- kmeans(Pharma1, centers = 5, nstart = 25)
k5$centers
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 3 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 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 -0.27449312 -0.7041516       0.556954446
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.14170336 -0.1168459      -1.416514761
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.06308085  1.5180158      -0.006893899
```

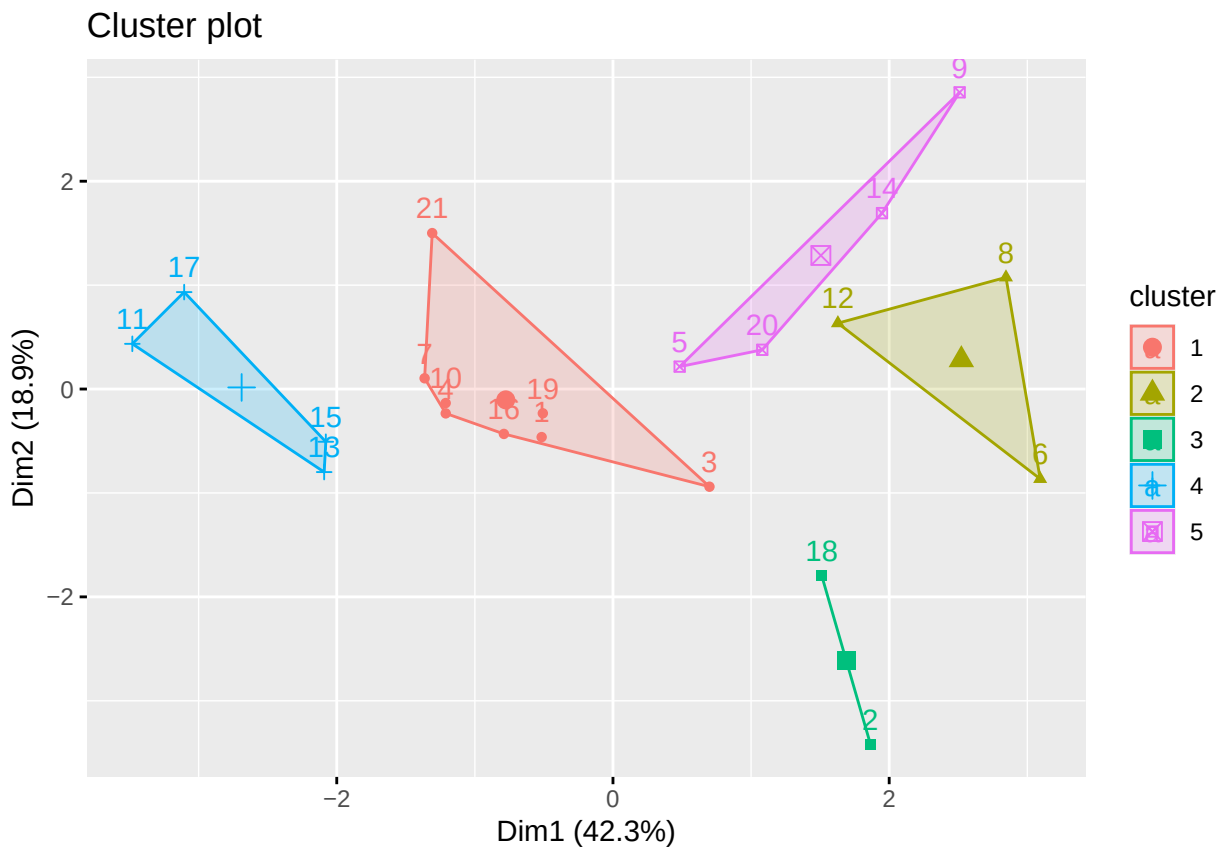
```
k5$size
```

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

```
k5$cluster[15]
```

```
## [1] 4
```

```
fviz_cluster(k5, data = Pharma1)
```



Number of clusters is determined using the Average silhouette method. The optimum number of clusters formed using this method is 5.

The data is clustered using K-means clustering algorithm.

```
aggregate(Pharma1,by=list(k5$cluster),FUN=mean)
```

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

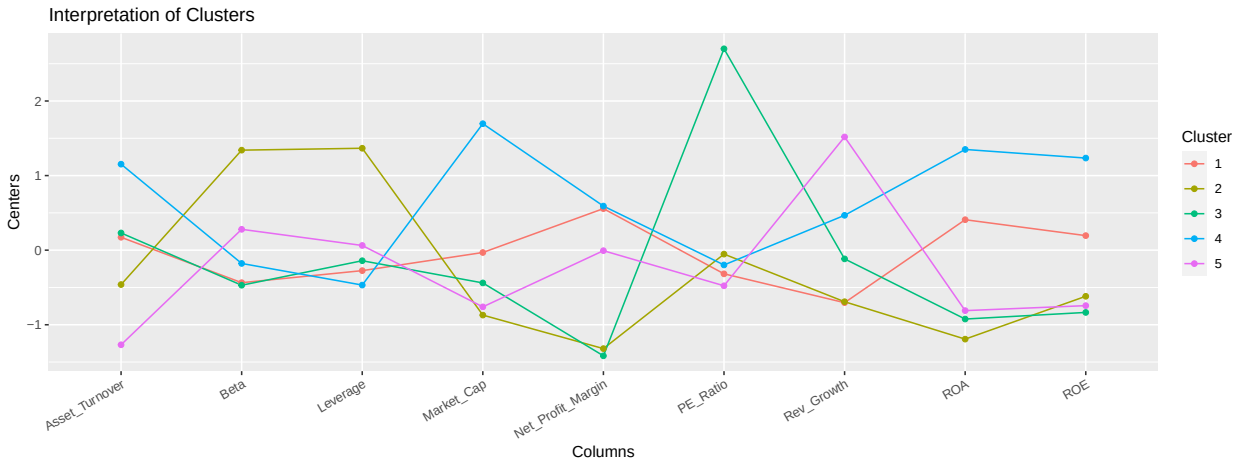
Assigning the cluster numbers to the companies

```
Pharma2 <- data.frame(Pharma$Name, k5$cluster)
Pharma2
```

```
##                           Pharma.Name k5.cluster
## 1                  Abbott Laboratories          1
## 2                       Allergan, Inc.          3
## 3                         Amersham plc          1
## 4                       AstraZeneca PLC          1
## 5                                Aventis          5
## 6                              Bayer AG          2
## 7          Bristol-Myers Squibb Company          1
## 8                          Chattem, Inc          2
## 9                 Elan Corporation, plc          5
## 10                Eli Lilly and Company          1
## 11                  GlaxoSmithKline plc          4
## 12                     IVAX Corporation          2
## 13                    Johnson & Johnson          4
## 14 Medicis Pharmaceutical Corporation          5
## 15                    Merck & Co., Inc.          4
## 16                          Novartis AG          1
## 17                           Pfizer Inc          4
## 18                Pharmacia Corporation          3
## 19         Schering-Plough Corporation          1
## 20        Watson Pharmaceuticals, Inc.          5
## 21                                Wyeth          1
```

B. Interpreting the clusters with respect to the numerical variables

```
centers <- data.frame(k5$centers) %>%
rowid_to_column() %>%
gather('Columns', 'Centers', -1)
ggplot(centers, aes(x = Columns, y = Centers, color = as.factor(rowid))) +
geom_line(aes(group = as.factor(rowid))) + geom_point() +
labs(color = "Cluster", title = 'Interpretation of Clusters') +
theme(axis.text.x = element_text(angle = 30, hjust = 1, vjust = 1))
```



Based on the above analysis, the formed clusters can be interpreted as follows;

Cluster-1: The companies in cluster 1 have high Revenue Growth but very low Asset Turnover. They have moderate Beta, Leverage, Net Profit Margin and low Market Cap, ROA and ROE.

Cluster-2: The companies in cluster 2 have high Beta and Leverage but very low Net Profit Margin and ROA. They fare moderately in PE Ration but have less than moderate Asset Turnover, Market Cap, Revenue Growth and ROE.

Cluster-3: The companies in cluster 3 have the highest PE Ration but incur the least Net Profit Margin. They fare moderately in Revenue Growth and Leverage but have low Market Cap, ROA and ROE.

Cluster-4: The companies in cluster 4 have high Market Cap, Asset Turnover, ROA and ROE. They fare over moderate values in Net Profit Margin and Revenue Growth and less than moderate in Beta, Leverage and PE Ratio.

Cluster-5: The companies in cluster 5 fare moderately in Asset Turnover, Leverage, Market Cap, PE Ration and ROE. They have less than moderate Beta and Revenue Growth but fare over the moderate values in Net Profit Margin, ROA and ROE.

C. Plotting the clusters w.r.t. the variables not used in the data

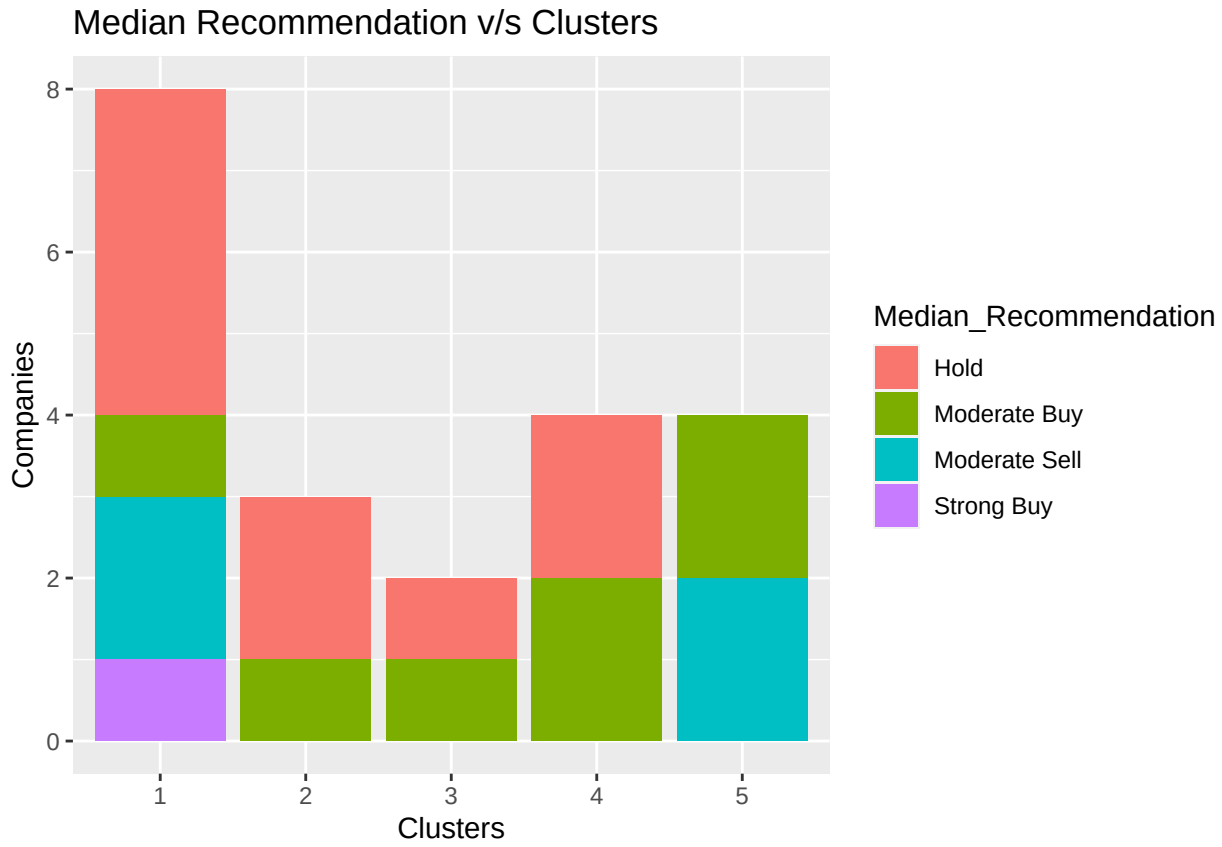
```
(Recommendations <-  Pharma %>%
  select(c("Median_Recommendation","Location","Exchange")) %>%
  mutate(cluster = k5$cluster) %>%
  arrange(desc(cluster)))
```

```
##    Median_Recommendation    Location Exchange cluster
## 1           Moderate Buy      FRANCE     NYSE       5
## 2          Moderate Sell     IRELAND     NYSE       5
## 3           Moderate Buy          US     NYSE       5
## 4          Moderate Sell          US     NYSE       5
## 5                   Hold          UK     NYSE       4
## 6           Moderate Buy          US     NYSE       4
## 7                   Hold          US     NYSE       4
## 8           Moderate Buy          US     NYSE       4
## 9           Moderate Buy      CANADA     NYSE       3
```

```
## 10                  Hold          US     NYSE       3
## 11                  Hold     GERMANY     NYSE       2
## 12          Moderate Buy          US   NASDAQ       2
## 13                  Hold          US     AMEX       2
## 14          Moderate Buy          US     NYSE       1
## 15            Strong Buy          UK     NYSE       1
## 16         Moderate Sell          UK     NYSE       1
## 17         Moderate Sell          US     NYSE       1
## 18                  Hold          US     NYSE       1
## 19                  Hold SWITZERLAND     NYSE       1
## 20                  Hold          US     NYSE       1
## 21                  Hold          US     NYSE       1
```

Plotting Median Recommendation v/s Clusters

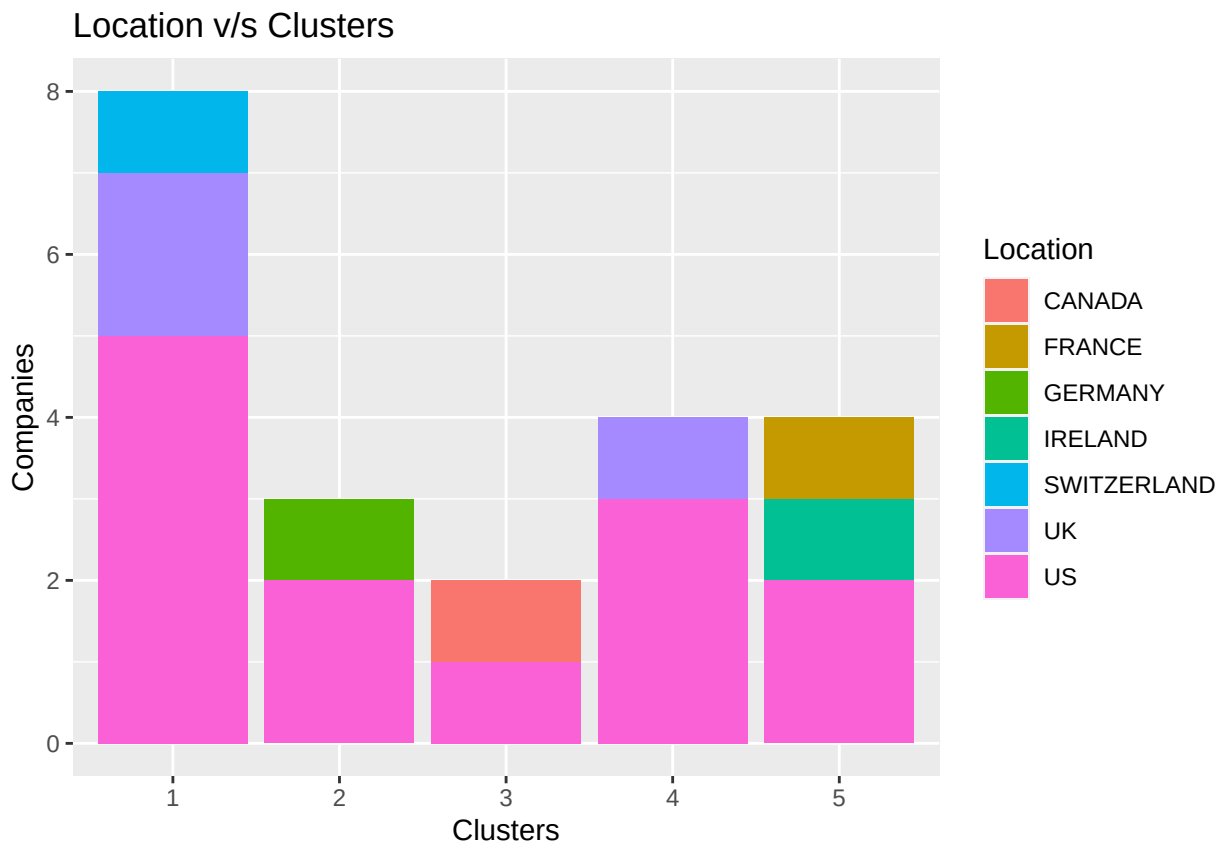
```
ggplot(Recommendations, aes(fill = Median_Recommendation, x = as.factor(cluster))) +
geom_bar(position = 'stack') + labs(x="Clusters", y="Companies",
title = "Median Recommendation v/s Clusters")
```



Cluster 1 has mixed recommendations with high Hold and low Buy recommendations with moderate sell recommendation. Cluster 5 has moderate buy/sell recommendations. Clusters 2, 3 and 4 have hold/moderate buy recommendations.

Plotting Location v/s Clusters

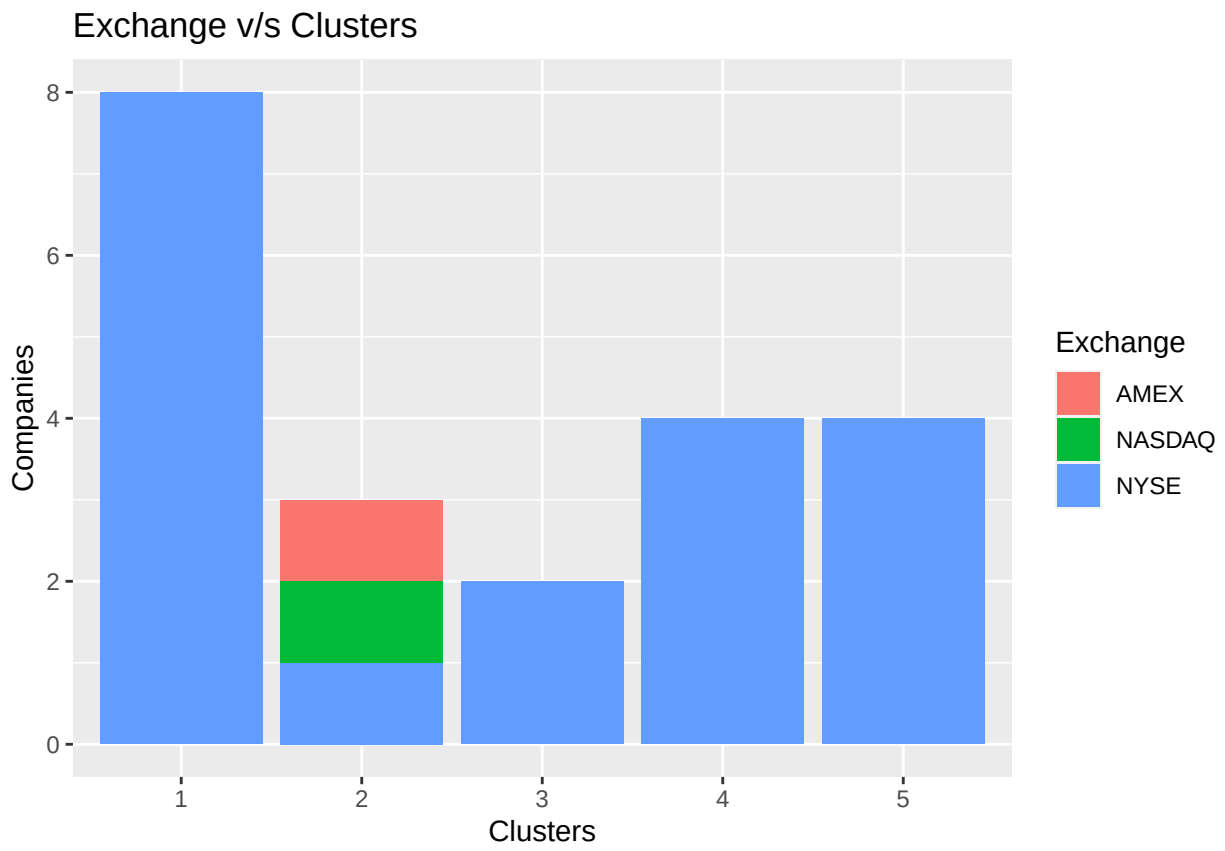
```
ggplot(Recommendations, aes(fill = Location, x = as.factor(cluster))) +
geom_bar(position = 'stack') + labs(x="Clusters", y="Companies",
title = "Location v/s Clusters")
```



All the clusters have companies from US. Cluster 5 has companies from France and Ireland. Cluster 3 has companies from Canada. Cluster 2 has companies from Germany. Clusters 1 and 4 have companies from UK. Cluster 1 also has companies from Switzerland.

Plotting Exchange v/s Clusters

```
ggplot(Recommendations, aes(fill = Exchange, x = as.factor(cluster))) +
geom_bar(position = 'stack') + labs(x="Clusters", y="Companies",
title = "Exchange v/s Clusters")
```



Clusters 1, 3, 4 and 5 have firms with stock exchange listed in NYSE. Cluster 2 has companies with stock exchange listed in AMEX, NASDAQ and NYSE.

D. Naming the clusters using variables in the dataset.

Cluster 1: High Hold/Sell Cluster

Cluster 2: High Hold/Buy Cluster

Cluster 3: Hold/Buy Cluster

Cluster 4: Hold/Buy Cluster

Cluster 5: Buy/Sell Cluster