Assignment 4

Snehitha Anpur

2022-11-01

```
Pharmaceuticals_main = read.csv("D:\\MSBA\\rTutorial\\Rtutorial\\Pharmaceuticals.csv")

Pharmaceuticals= Pharmaceuticals_main[,3:11]

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

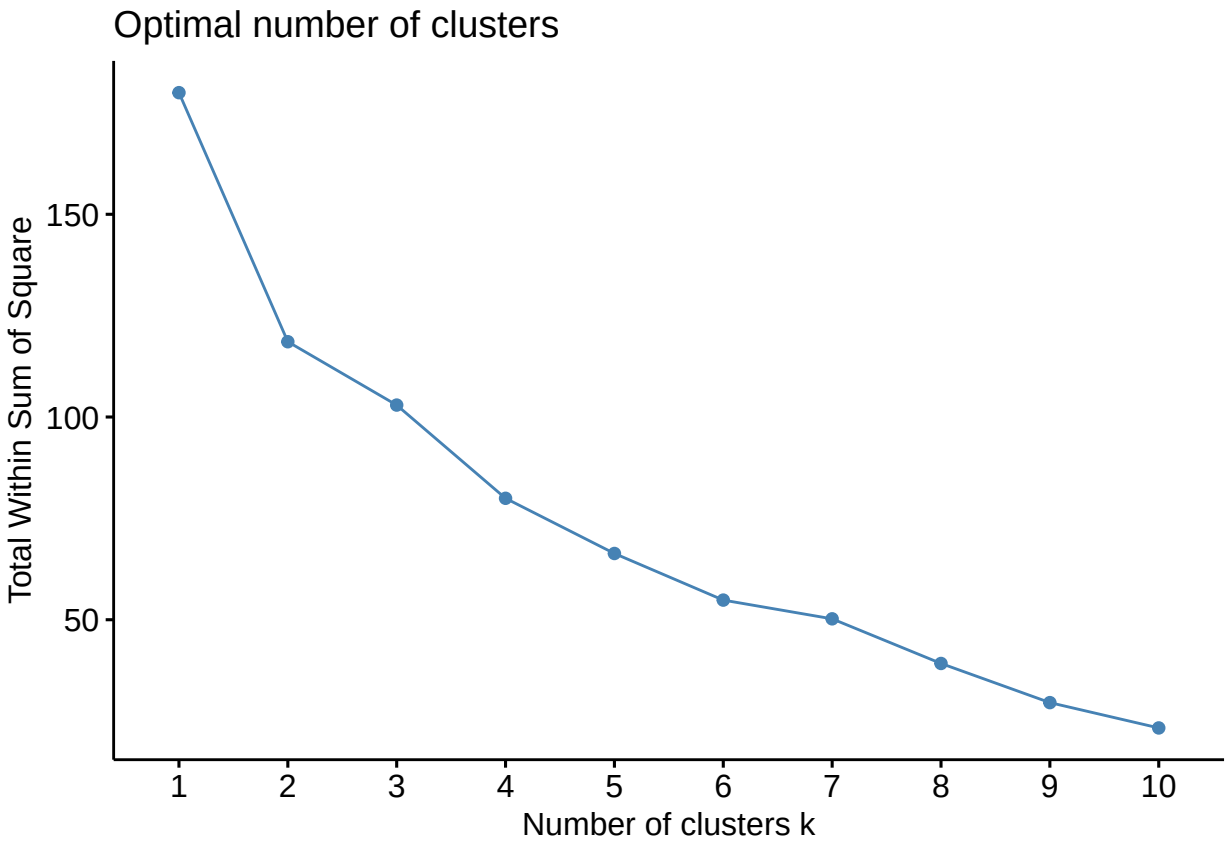
```
library(factoextra)
```

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

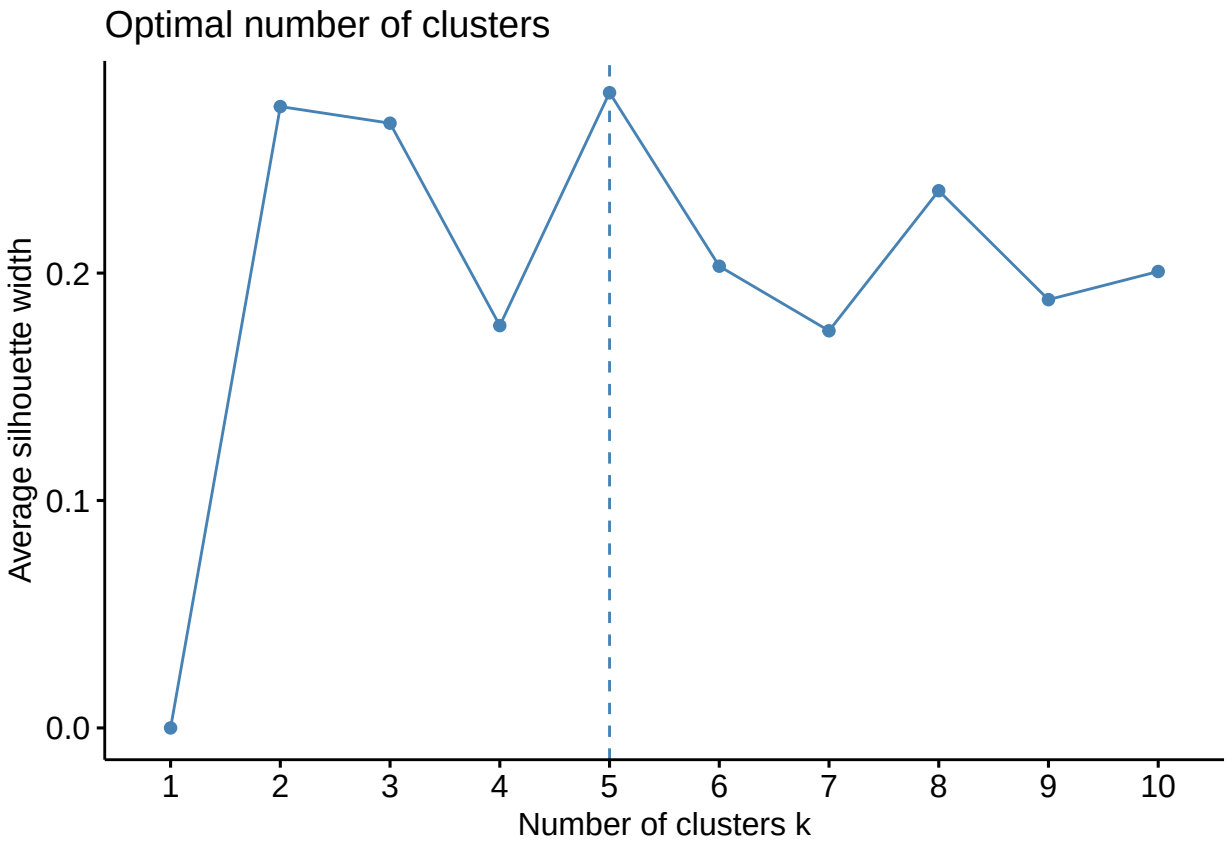
```
library(dplyr)

set.seed(100)
```

Question a:

```
#Normalizing the numerical variables
Norm_Pharmaceuticals = scale(Pharmaceuticals)

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




```
fviz_nbclust(Norm_Pharmaceuticals, kmeans, method = "silhouette")
```



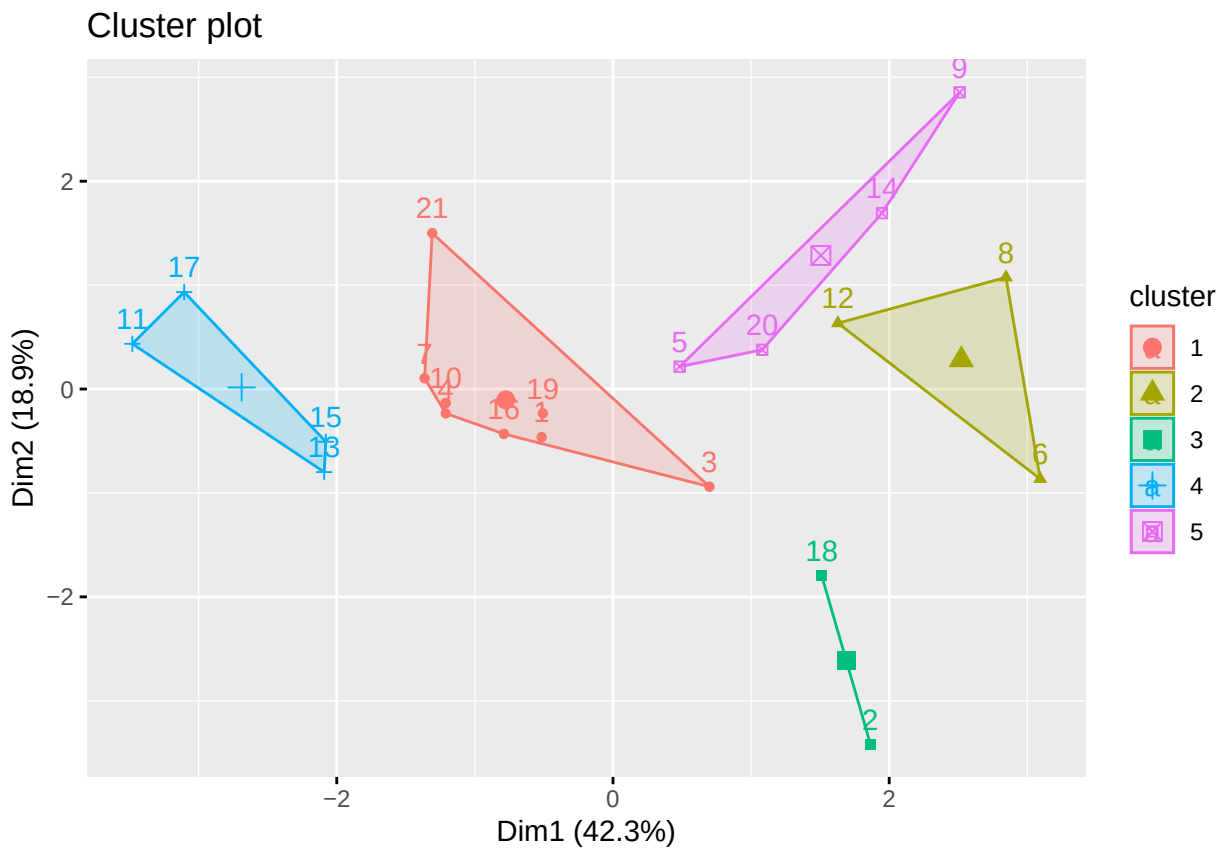
a)

The weight of variables effect the clusters as the distribution of data points on the scale is based on the weight of the each variable. As a result,the distance between the data points will be effected and hence the clusters.

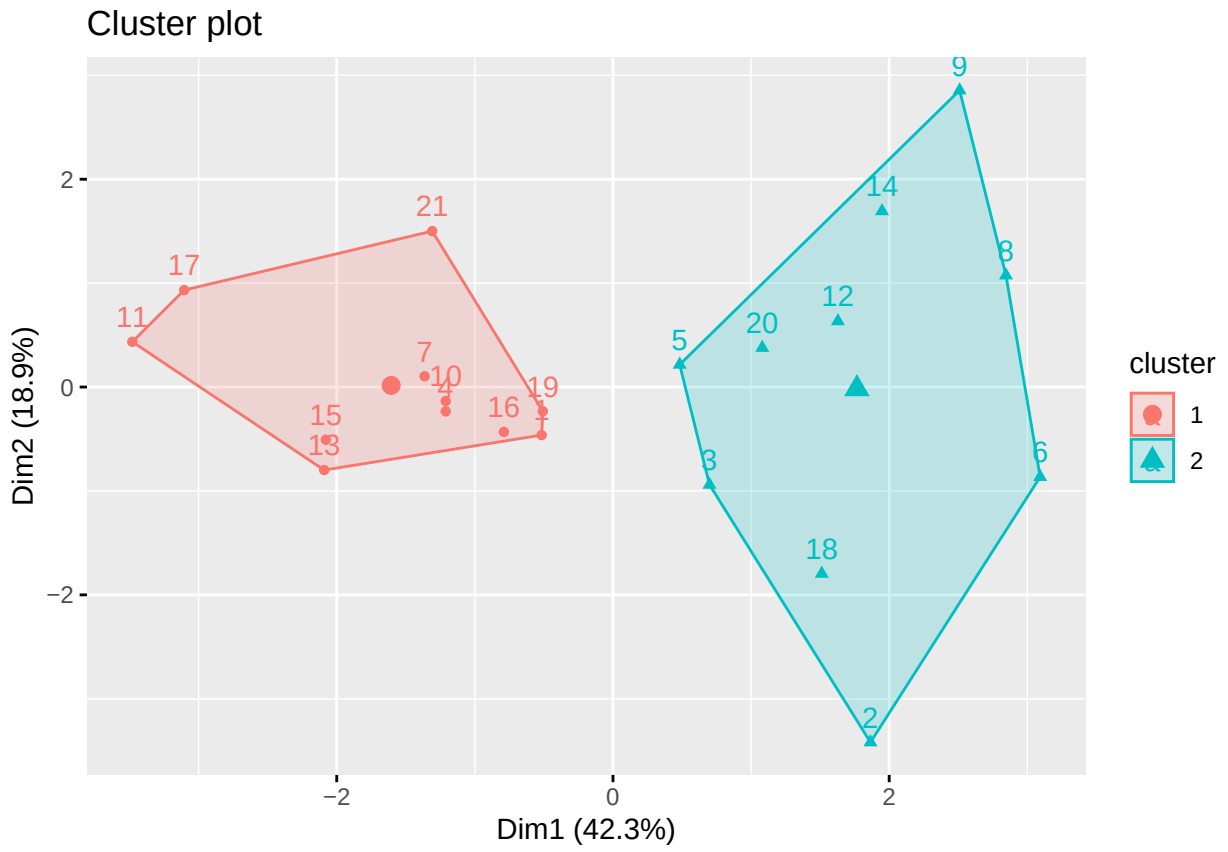
As part of Cluster Analysis, I have executed two clustering algorithms on the Pharmaceuticals data listed below: - Elbow Method - Silhouette Method

Using Elbow method, the optimal count of clusters (k) is 2. Whereas, the silhouette method gave a result of 5 clusters.

```
Sil_k5 = kmeans(Norm_Pharmaceuticals, centers=5,nstart=50)

fviz_cluster(Sil_k5,data=Norm_Pharmaceuticals)
```




```
Elb_k2 = kmeans(Norm_Pharmaceuticals, centers=2,nstart=50)

fviz_cluster(Elb_k2,data=Norm_Pharmaceuticals)
```




```
Sil_group=Sil_k5$cluster
Sil_k5$withinss
```

```
## [1] 21.879320 15.595925  2.803505  9.284424 12.791257
```

```
Sil_k5$tot.withinss
```

```
## [1] 62.35443
```

```
Elb_k2$withinss
```

```
## [1] 43.30886 75.26049
```

```
Elb_k2$tot.withinss
```

```
## [1] 118.5693
```

In order to select the right k value for this data set, I observed that the total sum of squares within the cluster for Silhouette method is 62.35 (Sil_k5$tot.withinss) which is less than 118.56 (Elb_k2tot.withinss) the value I got for Elbow method. As the sum of squares within the cluster is less which leads to homogeneous clusters, I prefer to choose Silhouette method for this assignment. Hence, the optimal k value is 5.

Here I am using Euclidean distance for measuring the distance between the data points over manhattan distance as this would the distance as the absolute difference between the data points

Question b:

```
Sil_group = as.data.frame(Sil_group)

Sil_Pharmaceuticals=cbind(Pharmaceuticals,Sil_group)

Cluster_mean= Sil_Pharmaceuticals %>% group_by(Sil_group) %>% summarise_all("mean")
Cluster_mean
```

```
## # A tibble: 5 x 10
##   Sil_group Market_Cap  Beta PE_Ra~1   ROE   ROA Asset~2 Lever~3 Rev_G~4 Net_P~5
##       <int>      <dbl> <dbl>   <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1         1      55.8  0.414    20.3  28.7 12.7    0.738   0.371    5.59   19.4
## 2         2       6.64 0.87     24.6  16.5  4.17   0.6     1.65     5.73    7.03
## 3         3      31.9  0.405    69.5  13.2  5.6    0.75    0.475   12.1     6.4
## 4         4     157.   0.48     22.2  44.4 17.7    0.95    0.22    18.5    19.6
## 5         5      13.1  0.598    17.7  14.6  6.2    0.425   0.635   30.1    15.6
## # ... with abbreviated variable names 1: PE_Ratio, 2: Asset_Turnover,
## #   3: Leverage, 4: Rev_Growth, 5: Net_Profit_Margin
```

Cluster 1

This cluster has the lower leverage, which indicates that these companies have fewer debts when compared to the companies in other clusters. This cluster has the least revenue growth among all the clusters but the net profit margin is higher for these companies. By examining the other variables, the companies in this cluster are performing better than Clusters 2,3 and 5

Cluster 2

The mean beta value for this cluster is higher when compared to other clusters. This indicates that the stocks of the companies in this cluster are more volatile. This cluster has the highest mean leverage, which indicates that the debt is higher for these companies. The companies in this cluster have less Market Capital, ROA, Revenue Growth, and Net Profit Margin. This indicates that these companies need to develop financially.

Cluster 3

The companies in this cluster have least net profit margin. Also, this cluster has the least Return on Equity (ROE) which says that the companies are weak at converting equities into profits. In addition, this cluster has the highest Price Earnings Ratio which indicates that the companies are not gaining profits. Though, profits are declining we can see that these companies' stocks are less volatile as this cluster has the least beta value.

Cluster 4

This cluster has the companies with the highest net profit margin, market capital, return on assets (ROA), return on equity (ROE), and Asset turnover. This cluster has the least mean leverage value which informs that these companies have fewer debts when compared to the shareholder's equity. Thus, this cluster has the best-performing companies when compared to other clusters.

Cluster 5

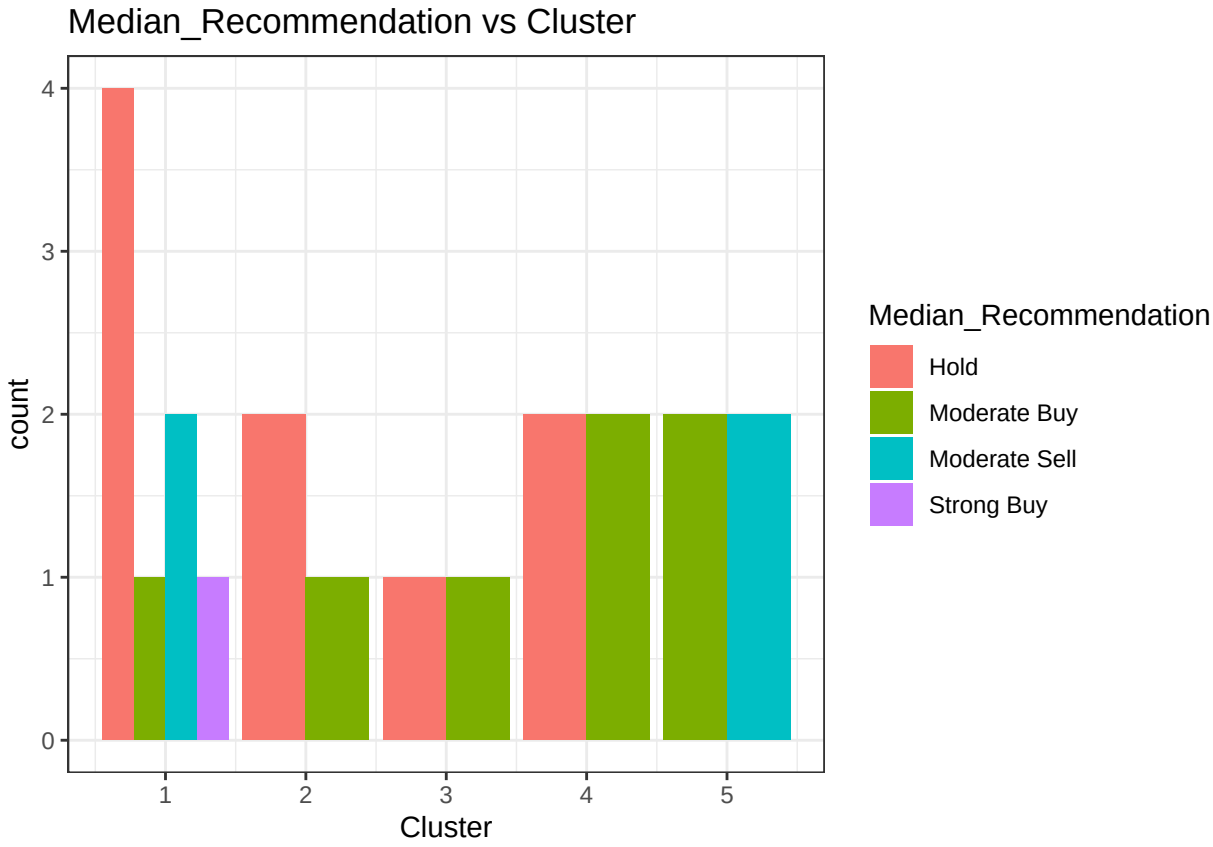
The companies in this cluster have high revenue growth which indicates that the companies are proceeding in the right direction of development. Ideally, the companies should utilize their assets to increase revenue which results in a higher asset turnover ratio. However, this cluster has the lowest asset turnover ratio. This cluster of companies has the lowest Price Earnings Ratio, which indicates that these companies have better earnings.

Question C:

```
library(ggplot2)
library(hrbrthemes)
```

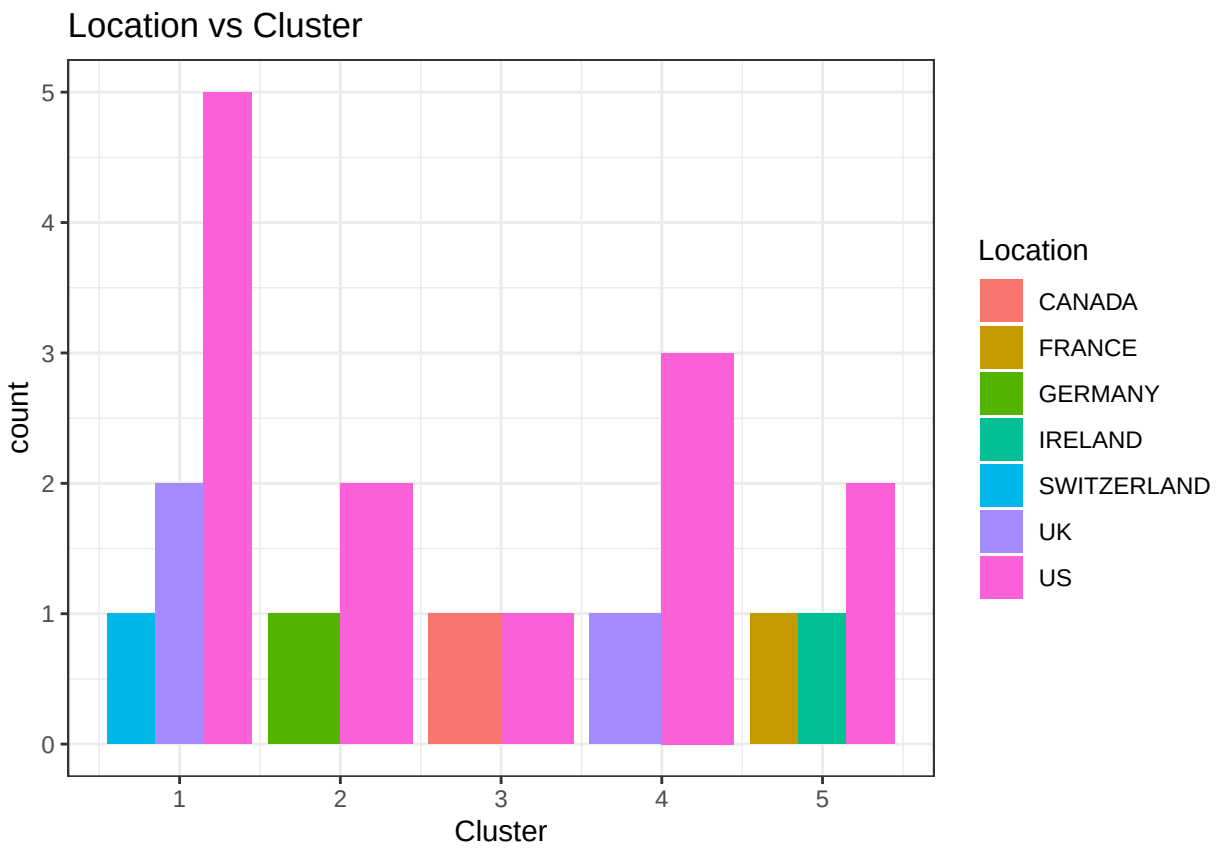
```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.

##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed and

##       if Arial Narrow is not on your system, please see https://bit.ly/arialnarrow
```

```
Pharma_categorical= Pharmaceuticals_main[,12:14]

Cluster_Pharma_cat = cbind(Pharma_categorical,Sil_group)

ggplot(Cluster_Pharma_cat, aes(x = Sil_group, fill = Median_Recommendation)) +
    geom_bar(position = "dodge") +
    labs(
      title = "Median_Recommendation vs Cluster",
      x = "Cluster"
      ) +
    theme_bw()
```

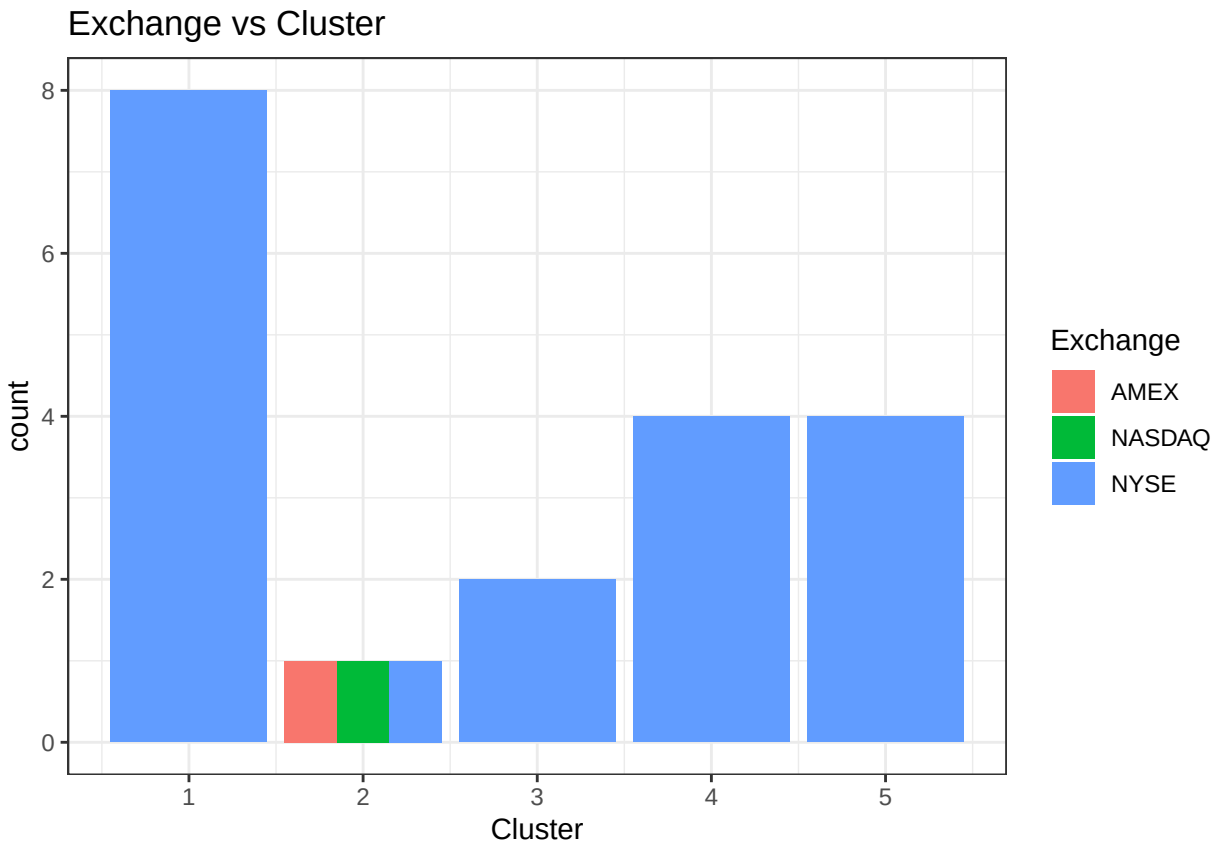



```
ggplot(Cluster_Pharma_cat, aes(x = Sil_group, fill = Location)) +
    geom_bar(position = "dodge") +
```

```
    labs(
      title = "Location vs Cluster",
      x = "Cluster"
      ) +
    theme_bw()
```




```
ggplot(Cluster_Pharma_cat, aes(x = Sil_group, fill = Exchange)) +
    geom_bar(position = "dodge") +
    labs(
      title = "Exchange vs Cluster",
      x = "Cluster"
      ) +
    theme_bw()
```



As pattern is a recognizable sequence, I don't see any specific patterns with respect to the categorical variables as these are not included when forming the cluster. However, I do see there are some observations that can be made from the plots.

1. In the Median Recommendation plot, I see that Cluster 1 has many "Hold" recommendations and Only Strong Buy is from this cluster. Moderate Buy is distributed in all the clusters.
2. From the Location vs Cluster Plot, I see that all the clusters have US based companies. Whereas, other distinct locations are distributed across all clusters
3. From the Exchange plot it can be seen that 19 out of 21 companies are listed on NYSE. This variable does not provide any pattern on the clusters. However, it was observed that Non-US countries are listed only on NYSE.Only Cluster 2 has all types of exchange

Question d

Naming the clusters

Cluster 1 - Developing Companies

Cluster 2 - High debt Companies

Cluster 3 - Low-profit Companies

Cluster 4 - Best Performing Companies

Cluster 5 - Better Earning Companies