K-Means Assignment

Setting default values to get a clean output

```
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)
```

Loading the required packages

```
library("tidyverse")
library("factoextra")
library("ggplot2")
library("dplyr")
```

Loading the data

```
pharma.df <- read.csv("Pharmaceuticals.csv")
```

Looking for na values

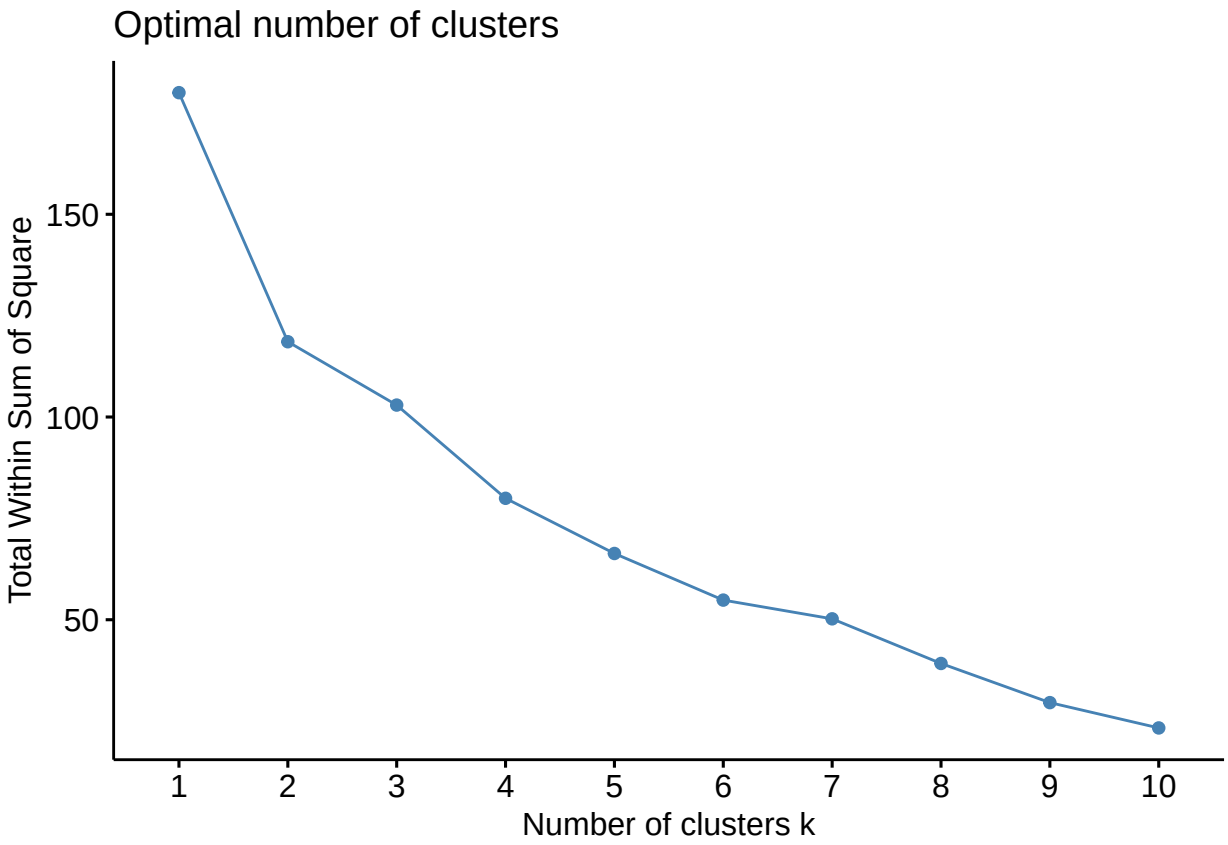
```
colMeans(is.na(pharma.df))
```

```
##                 Symbol                   Name             Market_Cap
##                      0                      0                      0
##                   Beta               PE_Ratio                    ROE
##                      0                      0                      0
##                    ROA         Asset_Turnover               Leverage
##                      0                      0                      0
##             Rev_Growth     Net_Profit_Margin Median_Recommendation
##                      0                      0                      0
##               Location               Exchange
##                      0                      0
```

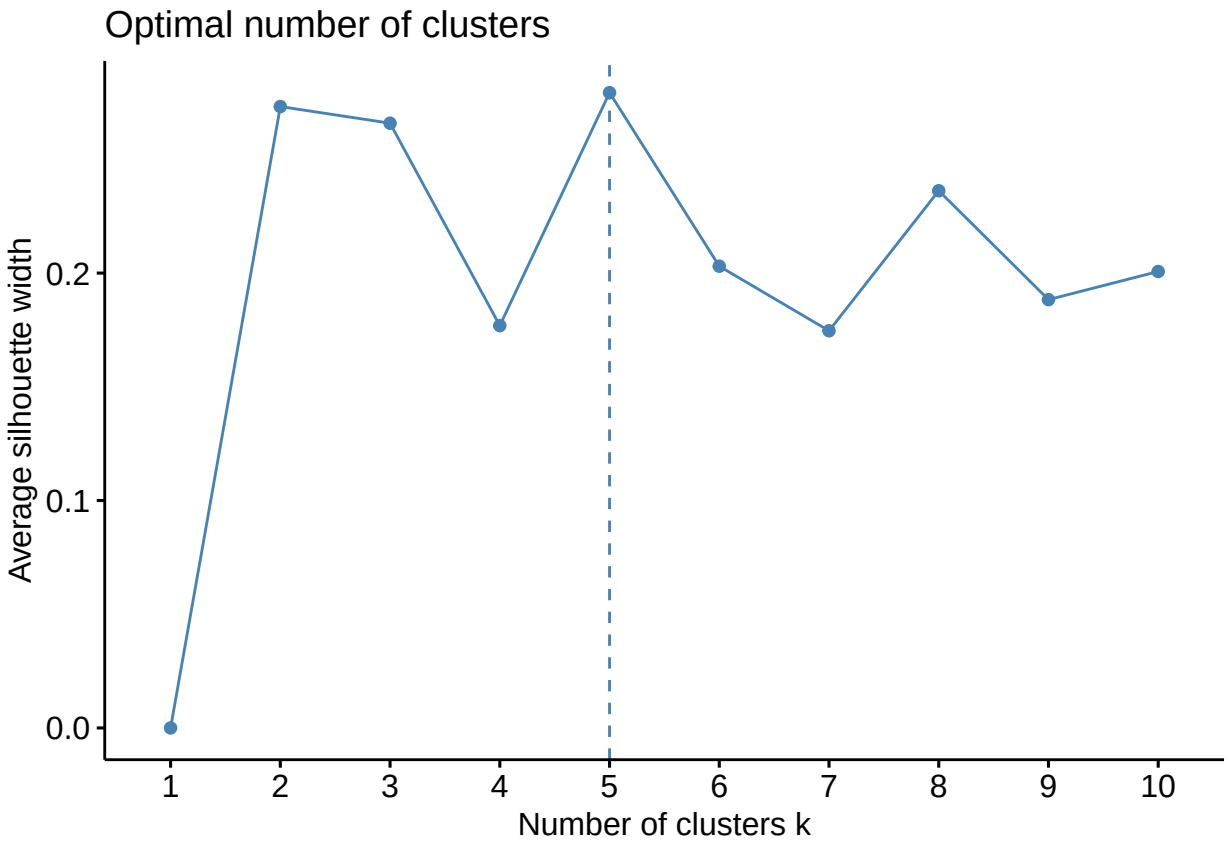
Normalization and finding the optimal k

```
pharma.df.norm <- scale(pharma.df[,-c(1:2,12:14)])

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




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



The optimal k thereby received using the wss method is k = 2 whereas by employing the silhouette method the optimal k received was k = 5.

Formulation of clusters using K-Means with k = 2 (WSS)

```
wss_kmeans <- kmeans(pharma.df.norm,centers = 2,nstart=25)
wss_kmeans
```

```
## K-means clustering with 2 clusters of sizes 11, 10
##
## Cluster means:
##   Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656
## 2 -0.7407208  0.3945061  0.3039863 -0.7222576 -0.9178575     -0.5073922
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.3331068 -0.2902163         0.6823310
## 2  0.3664175  0.3192379        -0.7505641
##
## Clustering vector:
##  [1] 1 2 2 1 2 2 1 2 2 1 1 2 1 2 1 1 1 2 1 2 1
##
## Within cluster sum of squares by cluster:
## [1] 43.30886 75.26049
##  (between_SS / total_SS =  34.1 %)
##
## Available components:
```

```
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

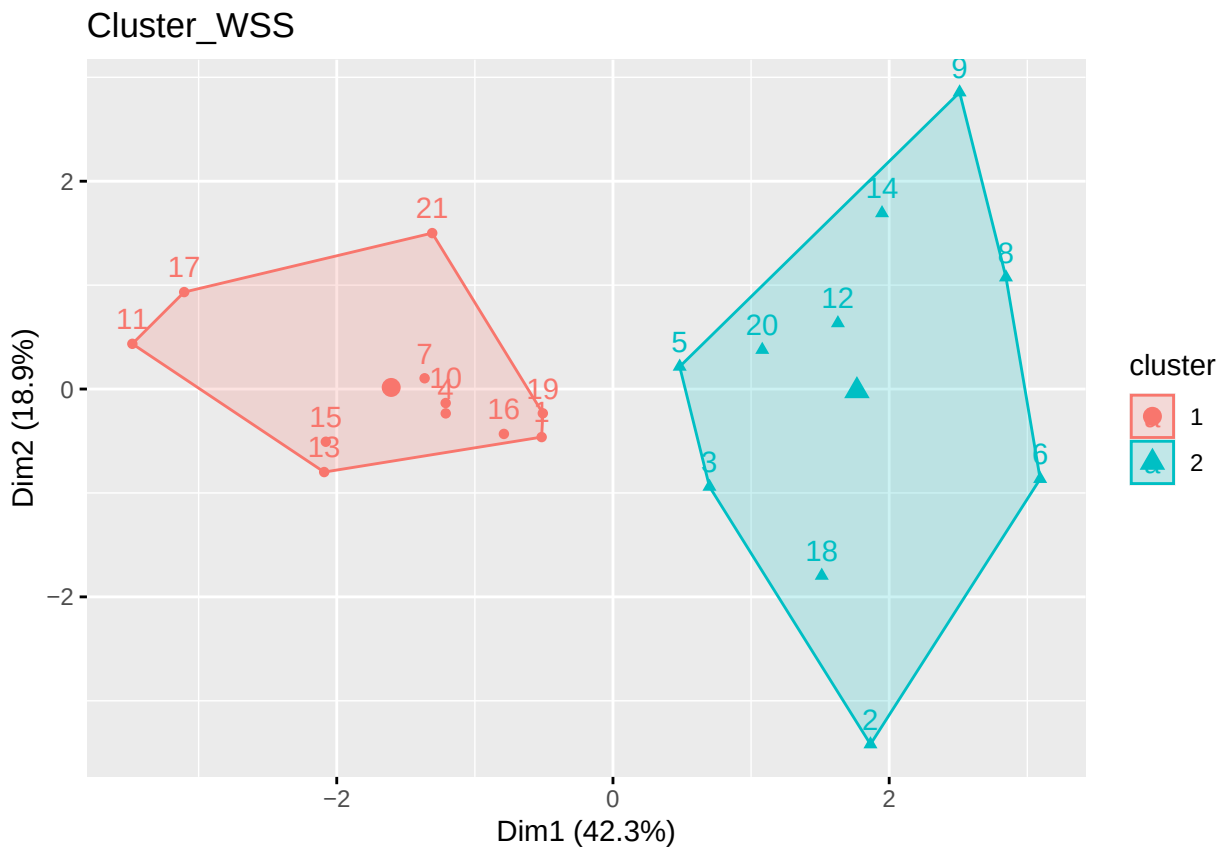
Formulation of clusters using K-Means with k = 5 (Silhouette)

```
silhouette_kmeans <- kmeans(pharma.df.norm,centers=5,nstart=25)
silhouette_kmeans
```

```
## K-means clustering with 5 clusters of sizes 3, 2, 8, 4, 4
##
## Cluster means:
##    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
##
## Clustering vector:
##  [1] 3 2 3 3 5 1 3 1 5 3 4 1 4 5 4 3 4 2 3 5 3
##
## Within cluster sum of squares by cluster:
## [1] 15.595925  2.803505 21.879320  9.284424 12.791257
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Cluster Plot (WSS)

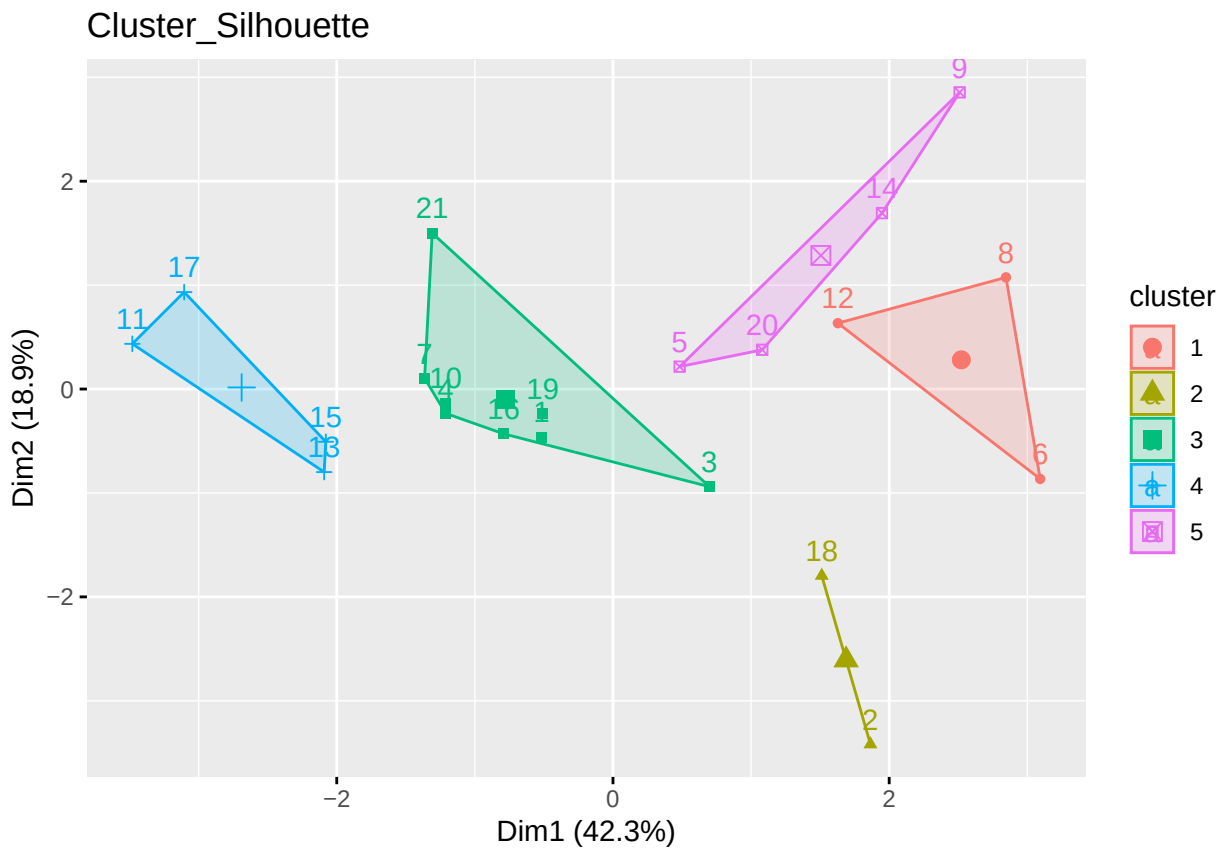
```
fviz_cluster(wss_kmeans,pharma.df[,-c(1:2,12:15)],main="Cluster_WSS")
```



By employing the WSS Method we get 2 clusters of size 11 and 10.

Cluster Plot (Silhouette)

```
fviz_cluster(silhouette_kmeans,pharma.df[,-c(1:2,12:15)],main="Cluster_Silhouette")
```



By employing the Silhouette Method we get 5 clusters of size 3, 2, 8, 4 and 4.

Binding the cluster assignment to the original data frame for analysis

```
clusters_wss <- wss_kmeans$cluster
clusters_silhouette <- silhouette_kmeans$cluster

pharma.df.1 <- cbind(pharma.df,clusters_wss)
pharma.df.2 <- cbind(pharma.df,clusters_silhouette)
```

Aggregating the clusters to interpret the attributes - WSS

```
int_wss <- aggregate(pharma.df.1[,-c(1:2,12:14)],by=list(pharma.df.1$clusters_wss),FUN="median")
print(int_wss[,-1])
```

```
##   Market_Cap  Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## 1      73.84 0.460    21.50 31.0 15.0            0.8    0.280      8.560
## 2       4.78 0.555    23.35 14.2  5.6            0.6    0.475     14.495
##   Net_Profit_Margin clusters_wss
## 1              20.6            1
## 2              11.1            2
```

Interpretation:

Note: The interpretation is solely based on the financial attributes of the given firms in each of the clusters, the interpretation obtained would be thereby helping an individual to take a decision regarding choosing an

cluster among the two to invest in order to gain profits.

Acceptable Profitability with Moderate Risk

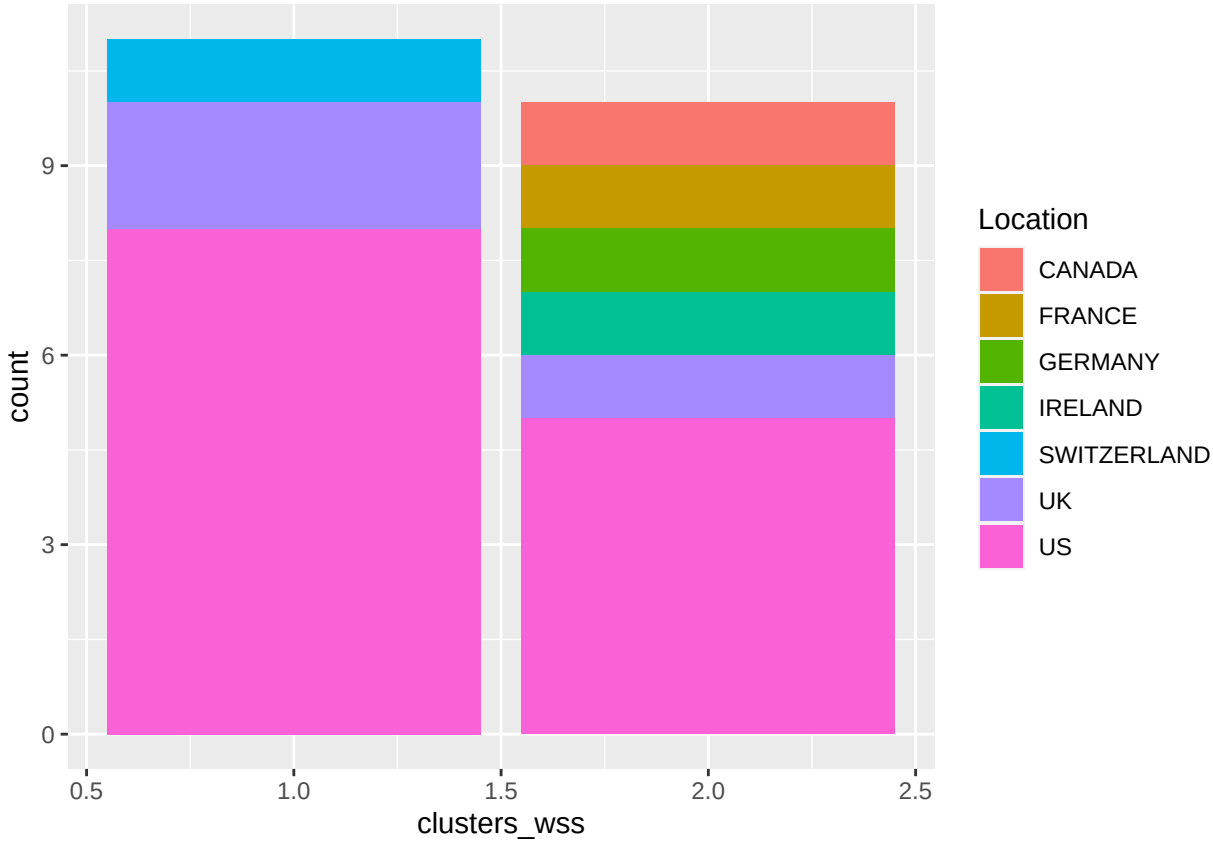
The First cluster here obtained is a good investment because of the high probability of success. The success is defined here with the help of the attributes "Market Capital", ROE - Return on Expenditure, ROA - Return on Assets, Asset Turnover and Net Profit Margin. The capital value in this cluster is 73.84, ROE which basically helps us know the returns on the money we put in as investment is high i.e. 31 and that of ROA which is the returns a firm expects to receive on the money they invest on the assets is also high i.e. 15. Similarly the turnover on the assets and net profit is high as well.The PE Ratio is less with that of the second cluster indicating that the company is properly valued without any disparity in it's share prices. The level of risk in this investment is low which is called out by the "Beta" value, generally beta value should be lower than 1 in this case it is 0.46 which refers that the variability in these firms would be moderate not having enough of fluctuations.Also the "Leverage" value, which refers to a firm having borrowed capital for an investment should be as less as possible because market is always unpredictable and there would be possibilities of a firm loosing the money which they have borrowed for an investment expecting profits in return. Here the leverage value is 0.28 which is comparatively less to the second cluster. "With a good investment there should be very little chance of losing the total amount invested" and the group of firms in this cluster are expressing higher success rate when compared to that with the second cluster.

Low Profitability with High Risk

Here, the second cluster is having poor performance metrics when compared with that with the first cluster, the market capital is very less i.e. 4.78 with that of 73.84 in first cluster, this shows the firms listed in this cluster are having less market share when compared to that with the first cluster. Return on Expenditure (ROE), Return on Assets (ROA), Asset Turnover, Net Profit Margin is less as well. The level of risk which is called out by the Beta and Leverage value is high in these firms which means that there is high variability and high borrowings in these firms with contrast to that with the first cluster. Comparatively the PE Ratio is high as well stating that the company's share value is overvalued, making it as a negative mark to these companies.

Pattern in the categorical variables

```
ggplot(pharma.df.1,aes(x=clusters_wss,fill=Location)) + geom_bar()
```



Cluster 1 and Cluster 2 seems to have a pattern with respect to the location of the pharmaceutical firms. More than 50% of the firms across both the clusters have "US" as their location. This also states that US has firms which are both profitable to invest (Acceptable Profitability with Moderate Risk) as well as firms which don't yield that good profits (Low Profitability with High Risk). But comparatively the better performing cluster i.e. Cluster 1 seems to have a greater ratio of companies based in US.

Aggregating the clusters to interpret the attributes - Silhouette

```
int_silhouette <- aggregate(pharma.df.2[,-c(1:2,12:14)],by=list(pharma.df.2$clusters_silhouette),FUN="m
print(int_silhouette[,-1])
```

```
##   Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage Rev_Growth
## 1      2.600 0.850    26.00 21.40  4.30           0.60    1.450      6.380
## 2     31.910 0.405    69.50 13.20  5.60           0.75    0.475     12.080
## 3     59.480 0.480    21.10 26.90 13.35           0.75    0.345      6.630
## 4    153.245 0.460    21.25 43.10 17.75           0.95    0.220     19.610
## 5      2.230 0.535    19.25 13.15  6.10           0.40    0.635     29.775
##   Net_Profit_Margin clusters_silhouette
## 1               7.5                   1
## 2               6.4                   2
## 3              19.3                   3
## 4              19.5                   4
## 5              14.2                   5
```

Interpretation:

Squandering Investment Group

The First Cluster is a highly fluctuative cluster with higher beta (variability in the firm) and leverage (outside borrowings) values indicating that there is high sense of risk in these firms. Also, the market capital & net profit margin are less making it less suitable for any possible investments.

Nonplussed Investment Group

The Second Cluster is likely mostly similar to that of the "Squandering Investment Group". It seems to have a lot of variability in it's PE Ratio which is the share price to the company value stating that it is likely overvalued. The beta and leverage values are also high stating that there is subsequent risk involved in this group. This cannot be a good choice for a better investment.

Fortune Class

Third Cluster can be considered as a set of firms with feasible market capital which are properly valued (PE Ratio) with middling risk involved (Beta and Leverage). It also has better returns over the expenditure and assets with a lucrative tendency.It can be a possible source of investment although the capital value is less when distinguished with the fourth cluster, there might be chances of the valuation to change/rise in the future.

Exorbitant Viability with Slighter Risk

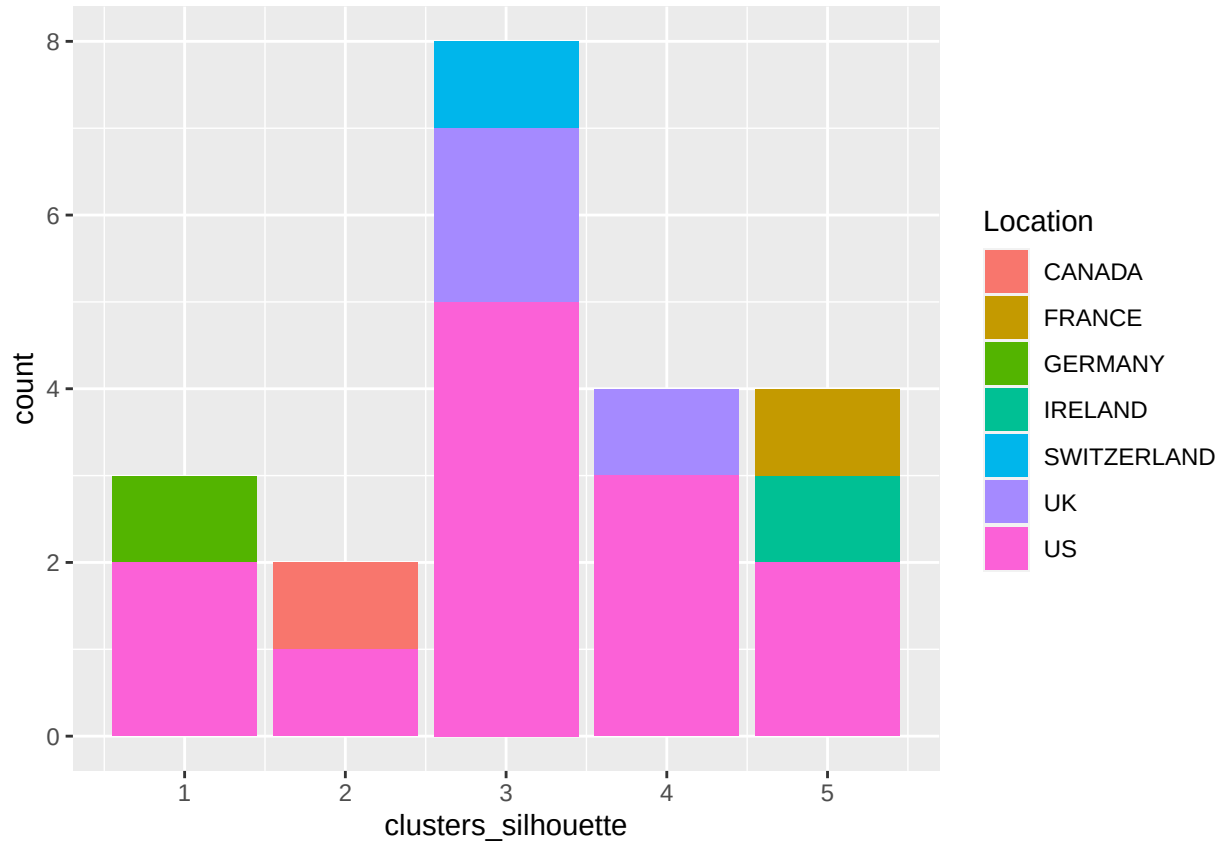
The Fourth Cluster is a good source of investment for any discrete individual who want to set a beneficial pitch for him/her. Here in this cluster as we see when compared to other firms across various clusters, the fourth cluster is having the "Highest Market Capital" of "153.245", "Lofty ROE - Return on Expenditure of"43.10" & ROA - Return on Assets of "17.75", "Sky-Spiking Asset Turnover" of "0.95" and "Net Profit Margin" of "19.5". It also has a "decent beta value" - indicating that the variance would be less and no much of risk would be involved and not only that it has "less leverage value" - which refers stating that the borrowed capital for future investments is small. PE Ratio is less indicating that the price to earnings ratio (share price to company value) is manageable indicating that the company is properly valued. If anyone wants to invest in a company which has a higher capital ratio and moderate risk with fewer liabilities then the firms which are part of this cluster make the best choice.

Less Remunerative Clump

The Fifth Cluster is stammering when it comes to providing returns on the expenditure which is basically the value which any investor would seek as a return over investment. External borrowings are high as well including good amount of variability in the firms (beta). It also has least capital value across all the groups and shockingly it is amusing to see that the revenue across these firms are highest as well. This might be possibly because the firms might have originated recently and are stabilizing to start their journey in the market.

Pattern in the categorical variables

```
ggplot(pharma.df.2,aes(x=clusters_silhouette,fill=Location)) + geom_bar()
```



In the silhouette clusters we get to see the see the similar level of pattern towards to the location as observed in the wss. Every cluster in here as more of it's locations in "US" when compared to that with the other locations. But it seems interesting to observe that the best cluster which defines the domain with true sense i.e. Cluster 4 has a greater ratio of US companies with a lesser ratio of Non - US based companies.

Note: The patterns therefore obtained in each of the clustering methods are generic, this is mostly because of the less amount of data which didn't give any further scope to visualize the categorical attributes.

Conclusion:

Any investment can be characterized by three factors: safety, income, and capital growth. Every investor has to pick an appropriate mix of these three factors.

Investment is always bounded to "profit to loss ratio", any given individual would want to maximize their profit with less amount of loss or no loss. Here, from the given data set the cluster named "Exorbitant Viability with Slighter Risk" displays all such characteristics. From the analysis and interpretation done I think this can be the best cluster to choose for an investment given which there is less probability of risk and higher profits.

Note: The reason for choosing a cluster from the silhouette method is because it is helping in defining the domain in a better way which can be used by any individuals to make a profitable decision towards their investment choices.