Assignment-4 Clustering

Vineeth Kadiyam

2023-11-12

```
library(readr)
Phaceut_RD <- read.csv("D:/Users/kadiyam/Documents/Pharmaceuticals.csv")
View(Phaceut_RD)
```

```
library(ggplot2)
library(factoextra)
```

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

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

```
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(cluster)
library(tidyverse)
```

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

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

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ stringr   1.5.0
## ✓ forcats   1.0.0     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
summary(Phaceut_RD)
```

```
##     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  
##                    
##                    
## 
```

```
#Task 1
#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.
R <- na.omit(Phaceut_RD)
R
```

	Sym…	Name	Market_Cap	B…	PE_Ratio	R…	R…	Asset
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
1	ABT	Abbott Laboratories	68.44	0.32	24.7	26.4	11.8	

	Sym…	Name	Market_Cap	B…	PE_Ratio	R…	R…	Asset
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
2	AGN	Allergan, Inc.	7.58	0.41	82.5	12.9	5.5	
3	AHM	Amersham plc	6.30	0.46	20.7	14.9	7.8	
4	AZN	AstraZeneca PLC	67.63	0.52	21.5	27.4	15.4	
5	AVE	Aventis	47.16	0.32	20.1	21.8	7.5	
6	BAY	Bayer AG	16.90	1.11	27.9	3.9	1.4	
7	BMY	Bristol-Myers Squibb Company	51.33	0.50	13.9	34.8	15.1	
8	CHTT	Chattem, Inc	0.41	0.85	26.0	24.1	4.3	
9	ELN	Elan Corporation, plc	0.78	1.08	3.6	15.1	5.1	
10	LLY	Eli Lilly and Company	73.84	0.18	27.9	31.0	13.5	

1-10 of 21 rows | 1-8 of 15 columns Previous 1 2 3 Next

```
row.names <- R[,1]
Phaceut1 <-  R[,3:11]
head(Phaceut1)
```

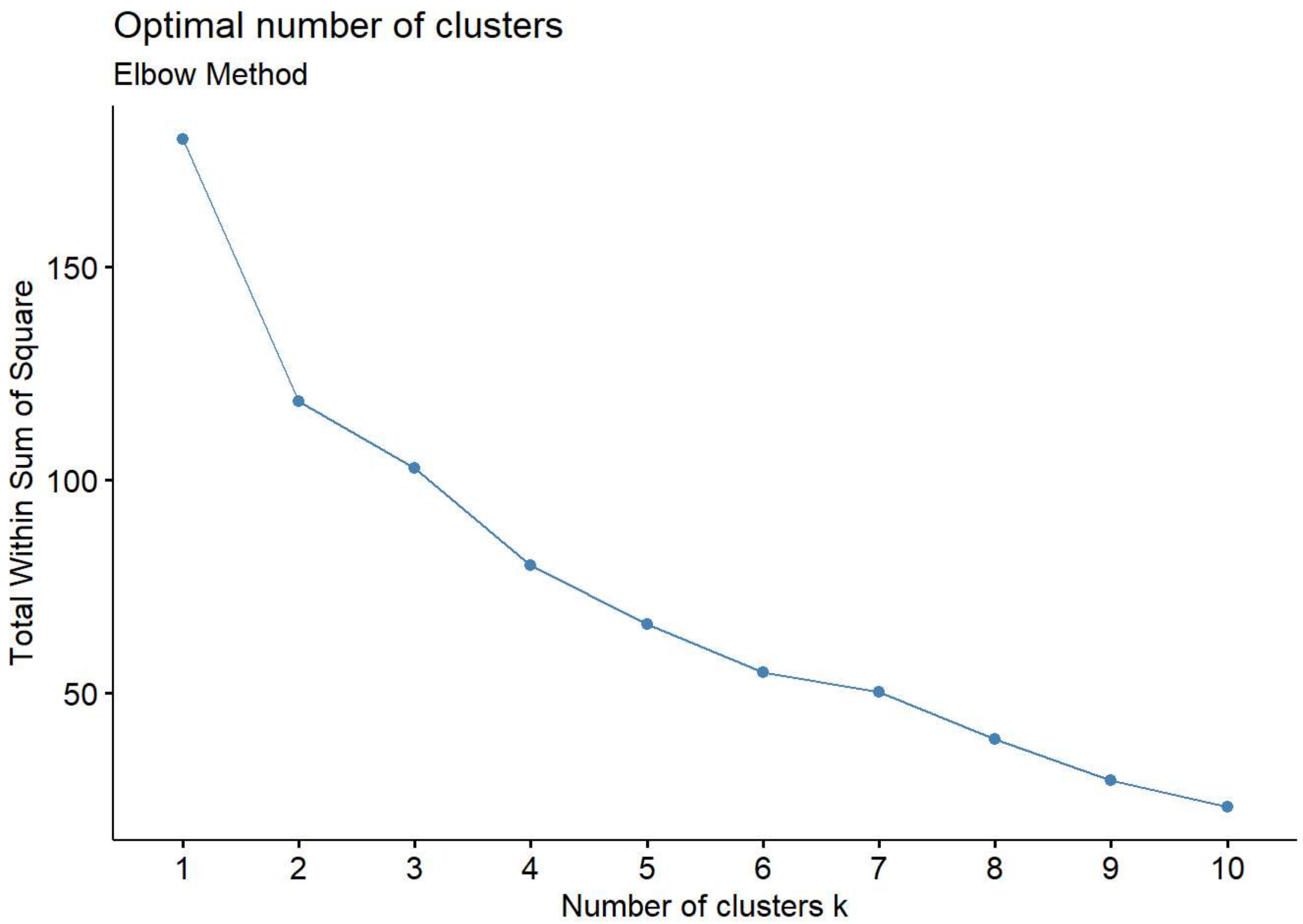
	Market_Cap	B…	PE_Ratio	R…	R…	Asset_Turnover	Leverage	Rev_Gro…	Net_Profit_Marg
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<db
1	68.44	0.32	24.7	26.4	11.8	0.7	0.42	7.54	16
2	7.58	0.41	82.5	12.9	5.5	0.9	0.60	9.16	5
3	6.30	0.46	20.7	14.9	7.8	0.9	0.27	7.05	11
4	67.63	0.52	21.5	27.4	15.4	0.9	0.00	15.00	18
5	47.16	0.32	20.1	21.8	7.5	0.6	0.34	26.81	12
6	16.90	1.11	27.9	3.9	1.4	0.6	0.00	-3.17	2

6 rows

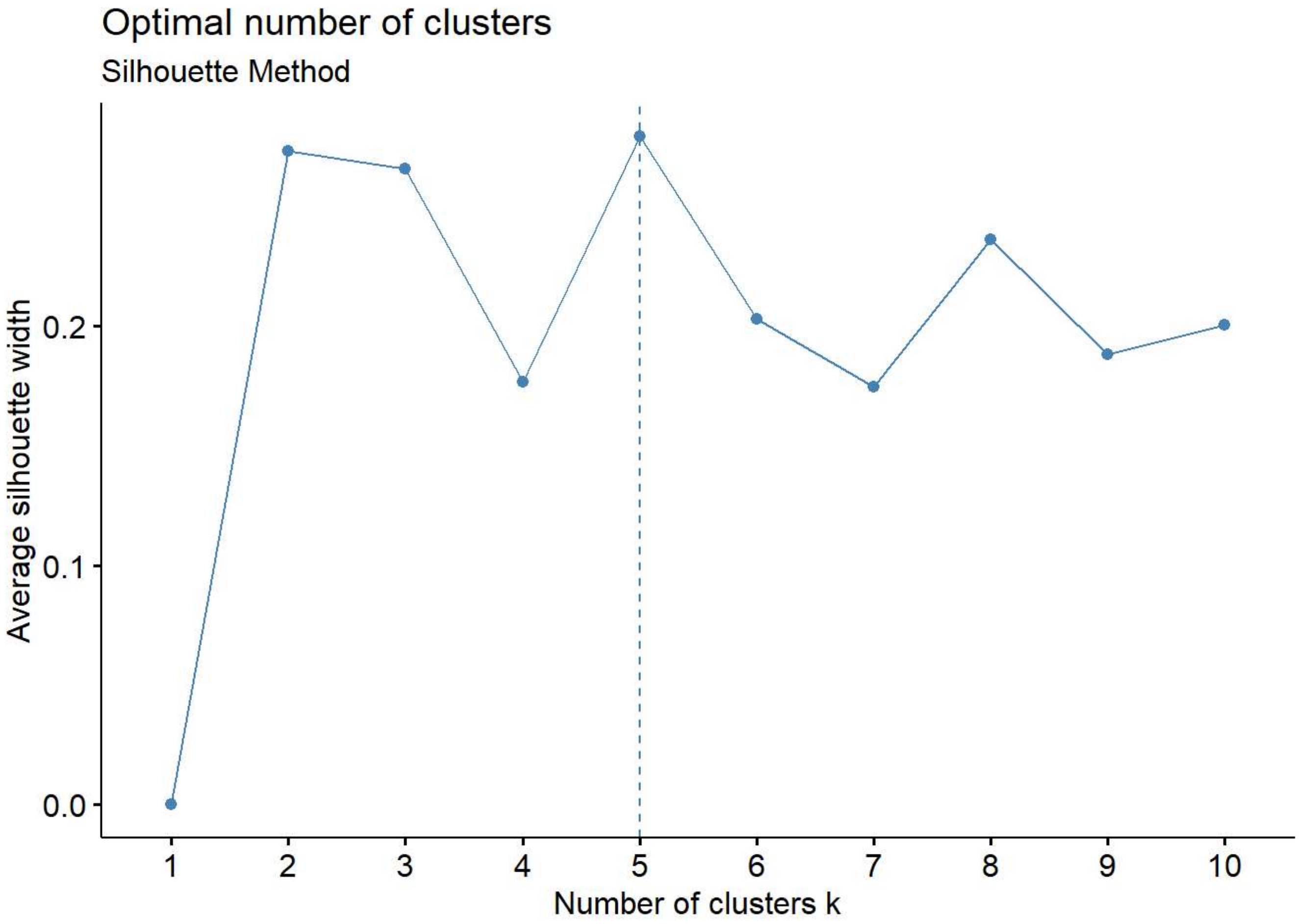
```
Phaceut2 <- scale(Phaceut1)
head(Phaceut2)
```

```
##   Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## 1  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
## 2 -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
## 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
## 5 -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656
## 6 -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.2120979 -0.5277675        0.06168225
## 2  0.0182843 -0.3811391       -1.55366706
## 3 -0.4040831 -0.5721181       -0.68503583
## 4 -0.7496565  0.1474473        0.35122600
## 5 -0.3144900  1.2163867       -0.42597037
## 6 -0.7496565 -1.4971443       -1.99560225
```

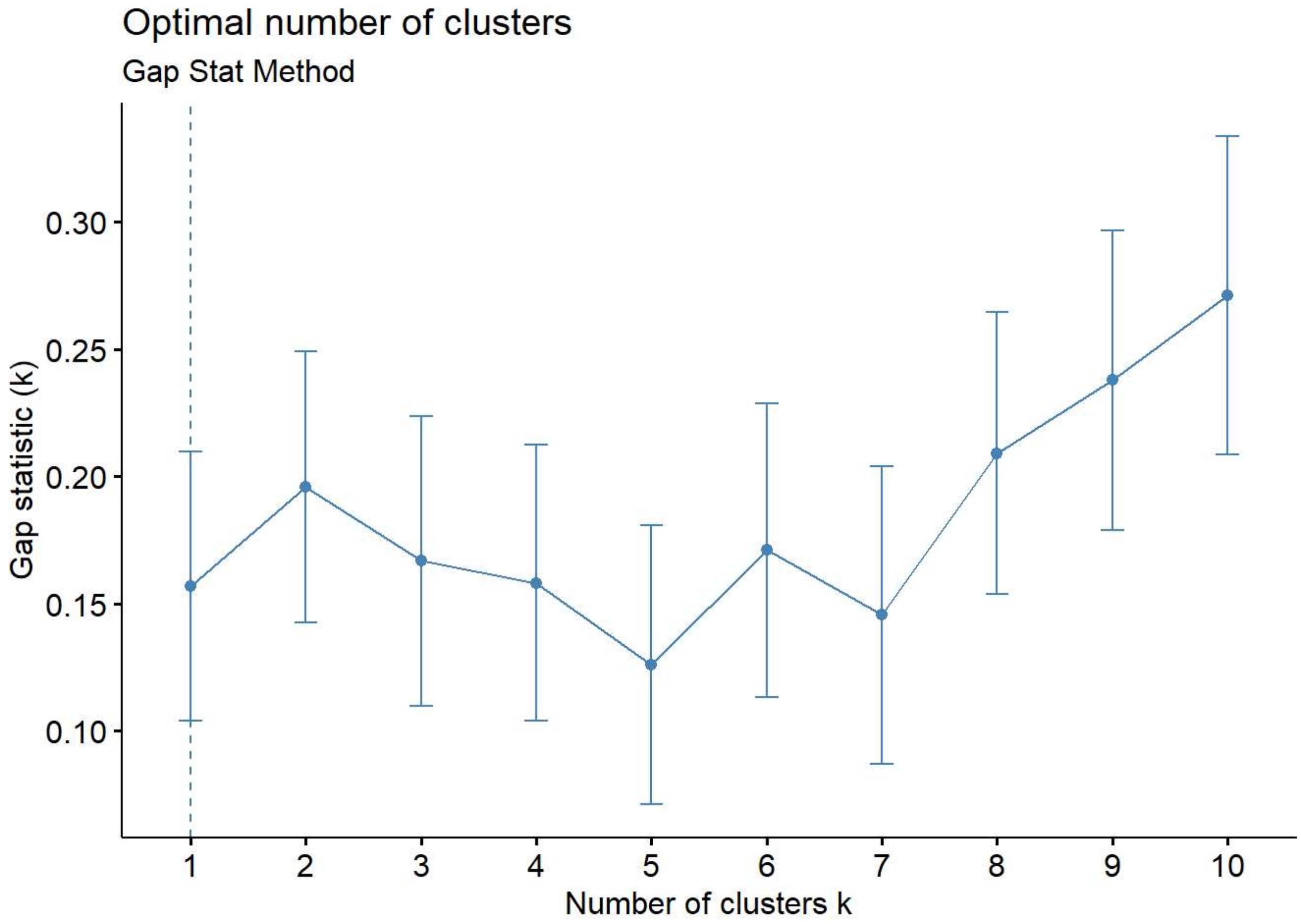
```
fviz_nbclust(Phaceut2, kmeans, method = "wss") +
  labs(subtitle = "Elbow Method")
```



```
fviz_nbclust(Phaceut2, kmeans, method = "silhouette") + labs(subtitle = "Silhouette Method")
```



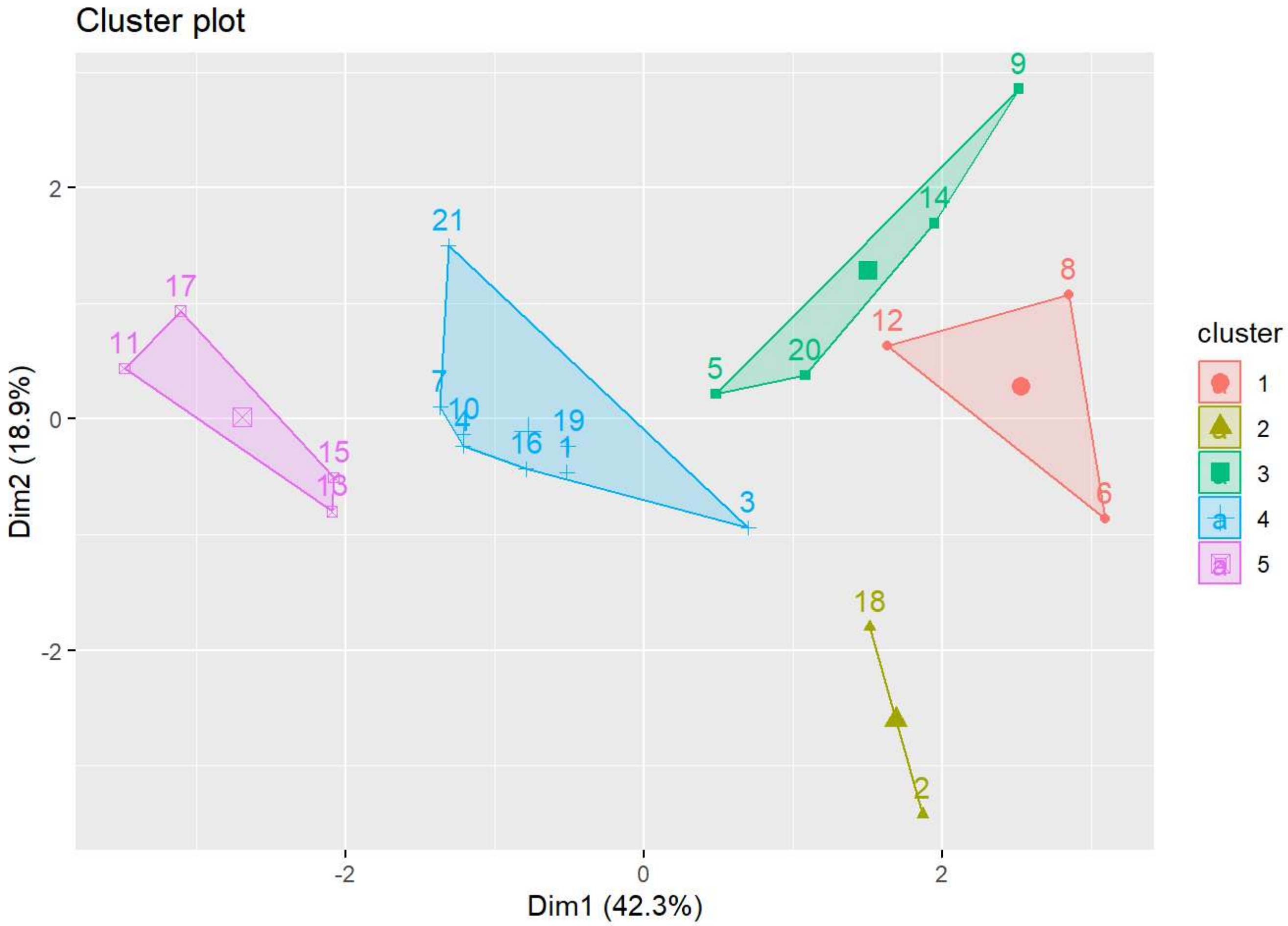
```
fviz_nbclust(Phaceut2, kmeans, method = "gap_stat") + labs(subtitle = "Gap Stat Method")
```



```
set.seed(64060)
k5 <- kmeans(Phaceut2, centers = 5, nstart = 25)
k5 $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 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 4 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
##      Leverage Rev_Growth Net_Profit_Margin
## 1  1.36644699 -0.6912914      -1.320000179
## 2 -0.14170336 -0.1168459      -1.416514761
## 3  0.06308085  1.5180158      -0.006893899
## 4 -0.27449312 -0.7041516       0.556954446
## 5 -0.46807818  0.4671788       0.591242521
```

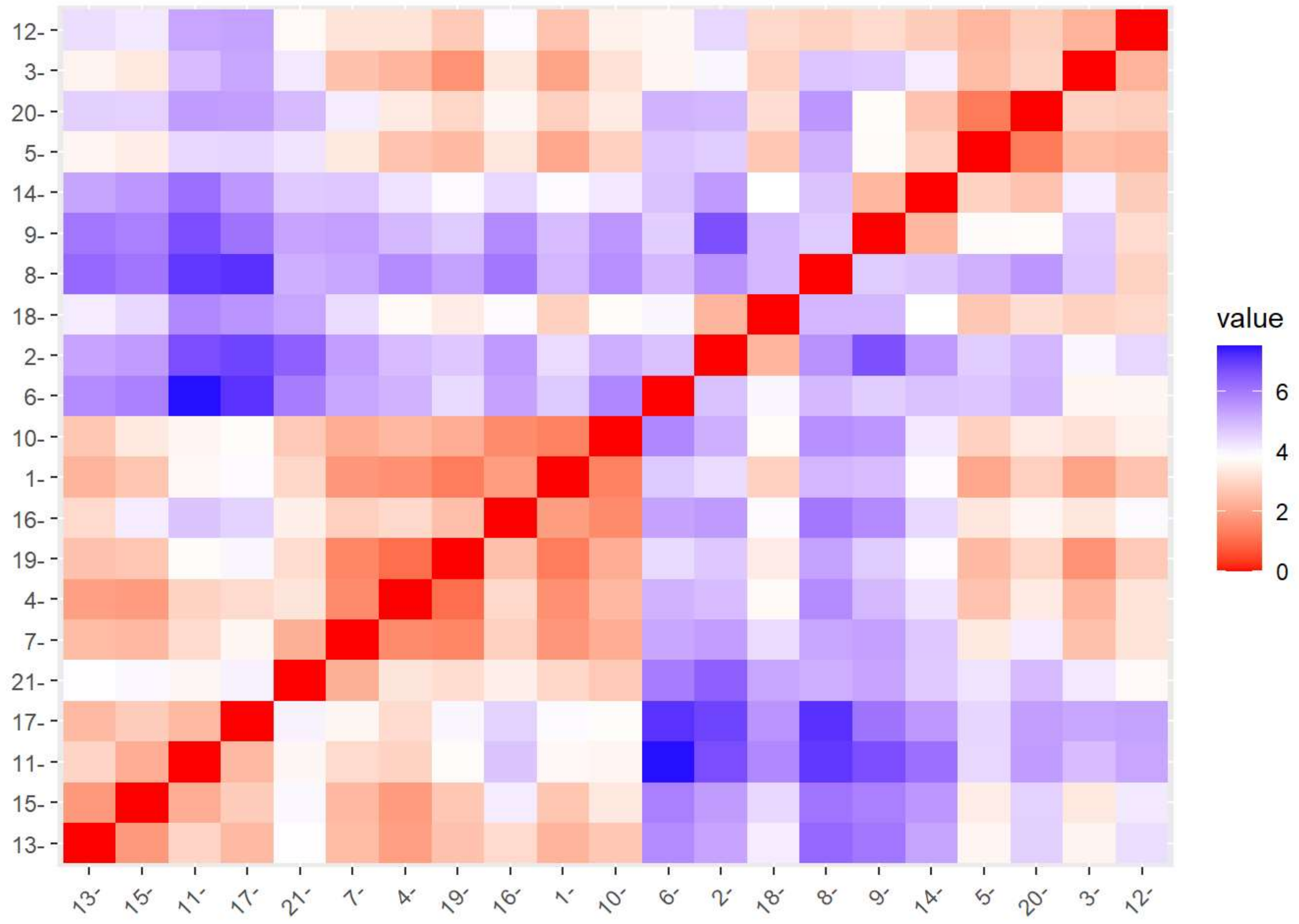
```
fviz_cluster(k5, data = Phaceut2)
```


Cluster plot
Dim2 (18.9%)
Dim1 (42.3%)
2
0
-2
-2
0
2
9
14
21
17
11
8
12
20
5
7
10
4
16
19
1
15
13
3
6
18
2
cluster
1
2
3
4
5

```
k5
```

```
## K-means clustering with 5 clusters of sizes 3, 2, 4, 8, 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 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 4 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
##      Leverage Rev_Growth Net_Profit_Margin
## 1  1.36644699 -0.6912914      -1.320000179
## 2 -0.14170336 -0.1168459      -1.416514761
## 3  0.06308085  1.5180158      -0.006893899
## 4 -0.27449312 -0.7041516       0.556954446
## 5 -0.46807818  0.4671788       0.591242521
## 
## Clustering vector:
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 
##  4  2  4  4  3  1  4  1  3  4  5  1  5  3  5  4  5  2  4  3  4 
## 
## Within cluster sum of squares by cluster:
## [1] 15.595925  2.803505 12.791257 21.879320  9.284424
##  (between_SS / total_SS =  65.4 %)
## 
## Available components:
## 
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
Distance <- dist(Phaceut2, method = "euclidian")
fviz_dist(Distance)
```



```
Fitting <- kmeans(Phaceut2,5)
aggregate(Phaceut2,by = list(Fitting$cluster), FUN = mean)
```

Grou… <int>	Market_Cap <dbl>	Beta <dbl>	PE_Ratio <dbl>	ROE <dbl>	ROA <dbl>	Asset_Turnover <dbl>	Leve
1	1.69558112	-0.1780563	-0.1984582	1.2349879	1.3503431	1.153164e+00	-0.468
2	-0.66114002	-0.7233539	-0.3512251	-0.6736441	-0.5915022	-1.537552e-01	-0.404
3	-0.96247577	1.1949250	-0.3639982	-0.5200697	-0.9610792	-1.153164e+00	1.477
4	-0.52462814	0.4451409	1.8498439	-1.0404550	-1.1865838	1.480297e-16	-0.344
5	0.08926902	-0.4618336	-0.3208615	0.3260892	0.5396003	6.589509e-02	-0.255

5 rows | 1-8 of 10 columns

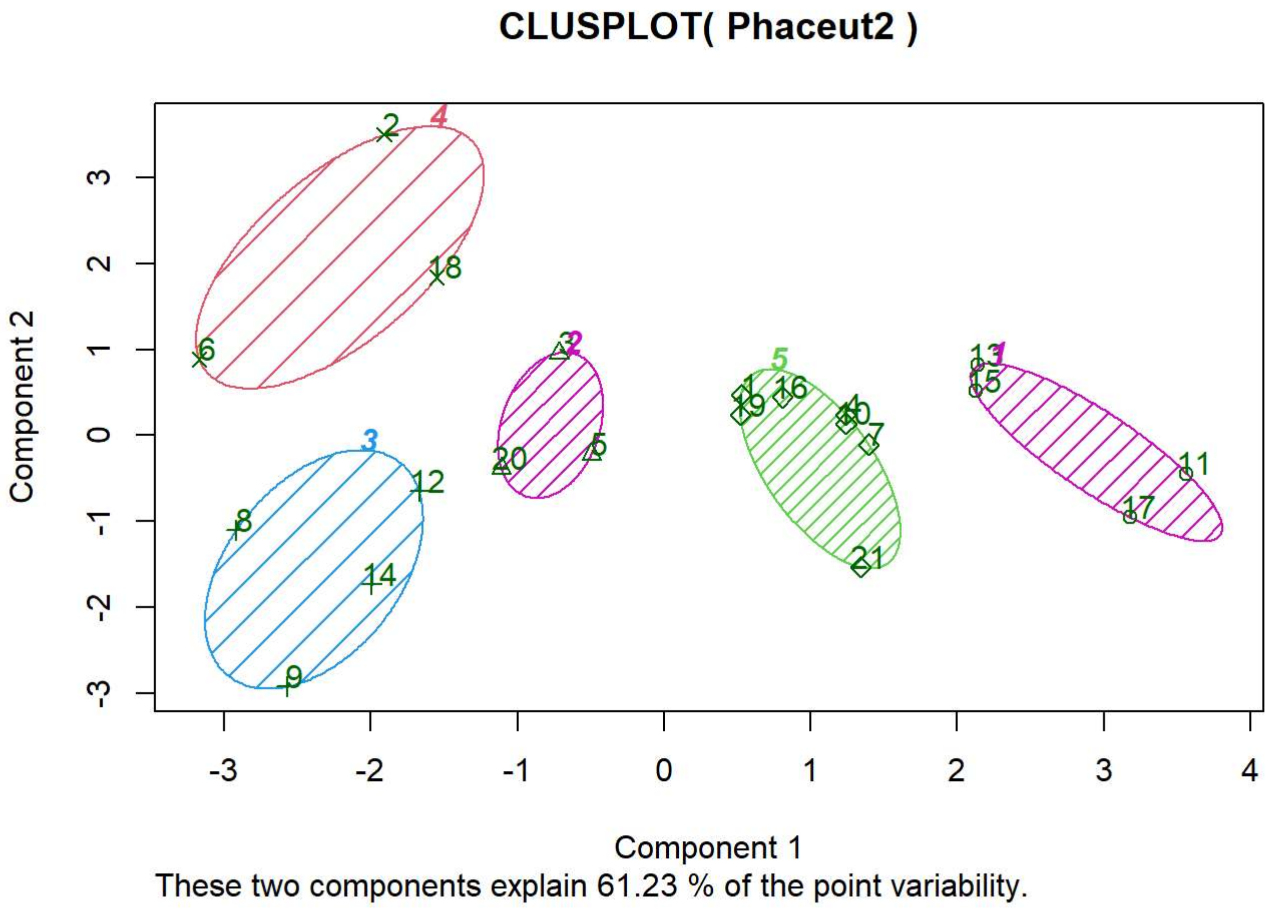
◀ ▶

```
Phaceut3 <- data.frame(Phaceut2,Fitting$cluster)
Phaceut3
```

	Market_Cap <dbl>	Beta <dbl>	PE_Ratio <dbl>	ROE <dbl>	ROA <dbl>	Asset_Turnover <dbl>	Leverag <db
1	0.1840960	-0.80125356	-0.04671323	0.04009035	0.2416121	0.0000000	-0.212097

	Market_Cap <dbl>	Beta <dbl>	PE_Ratio <dbl>	ROE <dbl>	ROA <dbl>	Asset_Turnover <dbl>	Leverag <db
2	-0.8544181	-0.45070513	3.49706911	-0.85483986	-0.9422871	0.9225312	0.018284
3	-0.8762600	-0.25595600	-0.29195768	-0.72225761	-0.5100700	0.9225312	-0.404083
4	0.1702742	-0.02225704	-0.24290879	0.10638147	0.9181259	0.9225312	-0.749656
5	-0.1790256	-0.80125356	-0.32874435	-0.26484883	-0.5664461	-0.4612656	-0.314490
6	-0.6953818	2.27578267	0.14948233	-1.45146000	-1.7127612	-0.4612656	-0.749656
7	-0.1078688	-0.10015669	-0.70887325	0.59693581	0.8617498	0.9225312	-0.020112
8	-0.9767669	1.26308721	0.03299122	-0.11237924	-1.1677918	-0.4612656	3.742797
9	-0.9704532	2.15893320	-1.34037772	-0.70899938	-1.0174553	-1.8450624	0.619837
10	0.2762415	-1.34655112	0.14948233	0.34502953	0.5610770	-0.4612656	-0.071308

```
library(cluster)
clusplot(Phaceut2,Fitting$cluster, color = TRUE, shade = TRUE,
         labels = 2,
         lines = 0)
```



```
#Task 2
#Interpret the clusters with respect to the numerical variables used in forming the clusters. Is
there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those \n #n
ot used in forming the clusters)

aggregate(Phaceut2, by = list(Fitting$cluster), FUN = mean)
```

Grou… <int>	Market_Cap <dbl>	Beta <dbl>	PE_Ratio <dbl>	ROE <dbl>	ROA <dbl>	Asset_Turnover <dbl>	Leve
1	1.69558112	-0.1780563	-0.1984582	1.2349879	1.3503431	1.153164e+00	-0.468
2	-0.66114002	-0.7233539	-0.3512251	-0.6736441	-0.5915022	-1.537552e-01	-0.404
3	-0.96247577	1.1949250	-0.3639982	-0.5200697	-0.9610792	-1.153164e+00	1.477
4	-0.52462814	0.4451409	1.8498439	-1.0404550	-1.1865838	1.480297e-16	-0.344
5	0.08926902	-0.4618336	-0.3208615	0.3260892	0.5396003	6.589509e-02	-0.255

5 rows | 1-8 of 10 columns

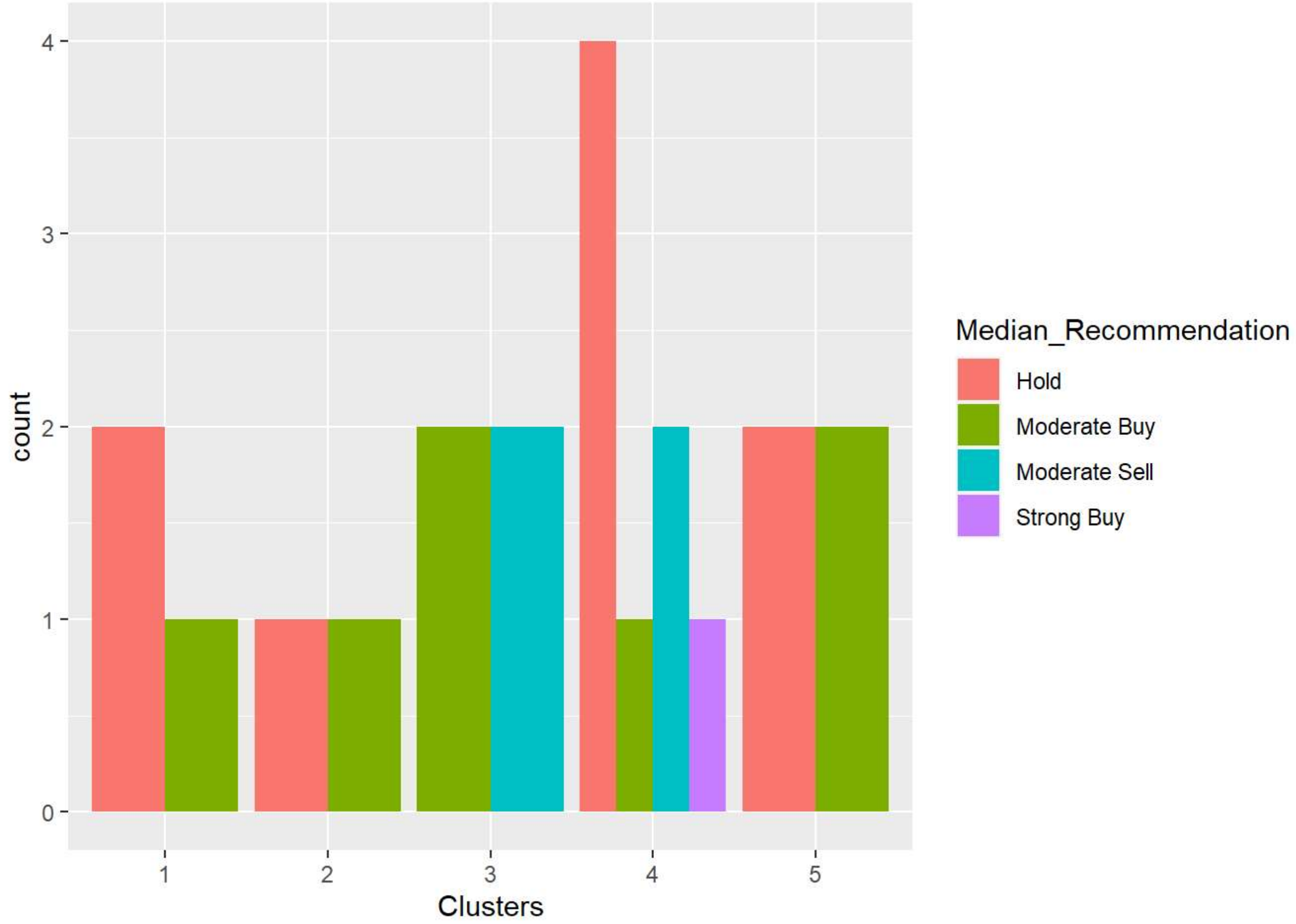
```
Pharmacies <- data.frame(Phaceut2,k5$cluster)
Pharmacies
```

	Market_Cap <dbl>	Beta <dbl>	PE_Ratio <dbl>	ROE <dbl>	ROA <dbl>	Asset_Turnover <dbl>	Leverag <db
1	0.1840960	-0.80125356	-0.04671323	0.04009035	0.2416121	0.0000000	-0.212097
2	-0.8544181	-0.45070513	3.49706911	-0.85483986	-0.9422871	0.9225312	0.018284
3	-0.8762600	-0.25595600	-0.29195768	-0.72225761	-0.5100700	0.9225312	-0.404083
4	0.1702742	-0.02225704	-0.24290879	0.10638147	0.9181259	0.9225312	-0.749656
5	-0.1790256	-0.80125356	-0.32874435	-0.26484883	-0.5664461	-0.4612656	-0.314490
6	-0.6953818	2.27578267	0.14948233	-1.45146000	-1.7127612	-0.4612656	-0.749656
7	-0.1078688	-0.10015669	-0.70887325	0.59693581	0.8617498	0.9225312	-0.020112
8	-0.9767669	1.26308721	0.03299122	-0.11237924	-1.1677918	-0.4612656	3.742797
9	-0.9704532	2.15893320	-1.34037772	-0.70899938	-1.0174553	-1.8450624	0.619837
10	0.2762415	-1.34655112	0.14948233	0.34502953	0.5610770	-0.4612656	-0.071308

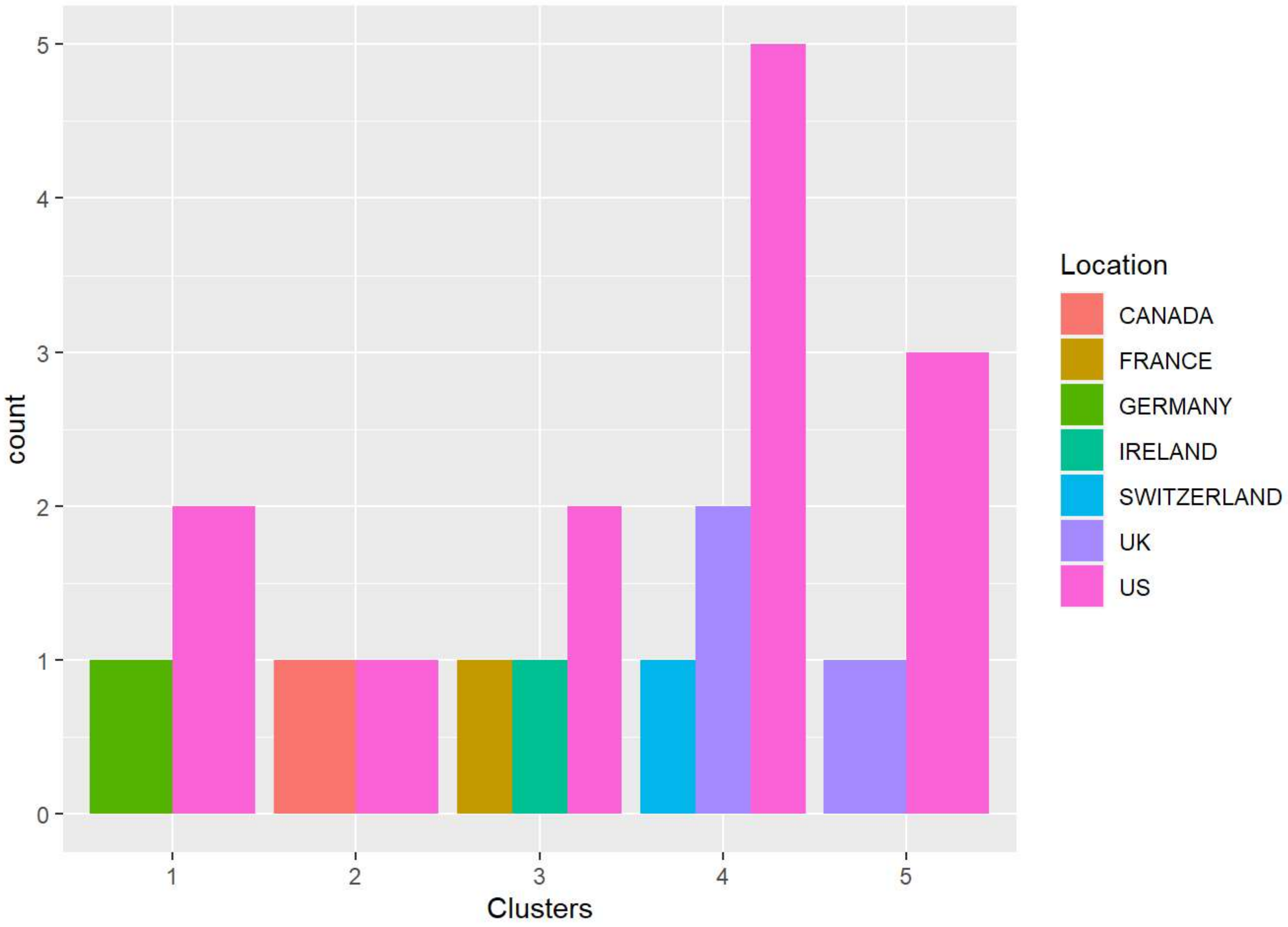
1-10 of 21 rows | 1-8 of 11 columns Previous **1** 2 3 Next

```
#CLuster 1:- JNJ, MRK, GSK, PFE
#Cluster 1: Highest Market_Cap and lowest Beta/PE Ratio
#Cluster 2:- AHM, WPI, AVE
#Cluster 2: Highest Revenue Growth and lowest PE/Asset Turnover Ratio
#Cluster 3:- CHTT, IVX, MRX, ELN
#Cluster 3: Highest Beta/leverage/Asset Turnover Ratio and lowest
#Net_Profit_Margin, PE ratio and Marke#Cluster
#Cluster 4:- BAY, PHA,AGN
#Cluster 4: Highest PE ratio and lowest Leverage/Asset_Turnover
#Cluster 5:- ABT, WYE, AZN, SGP, BMY, NVS, LLY
#Cluster 5: Highest Net_Proft_Margin and lowest Leverage
```

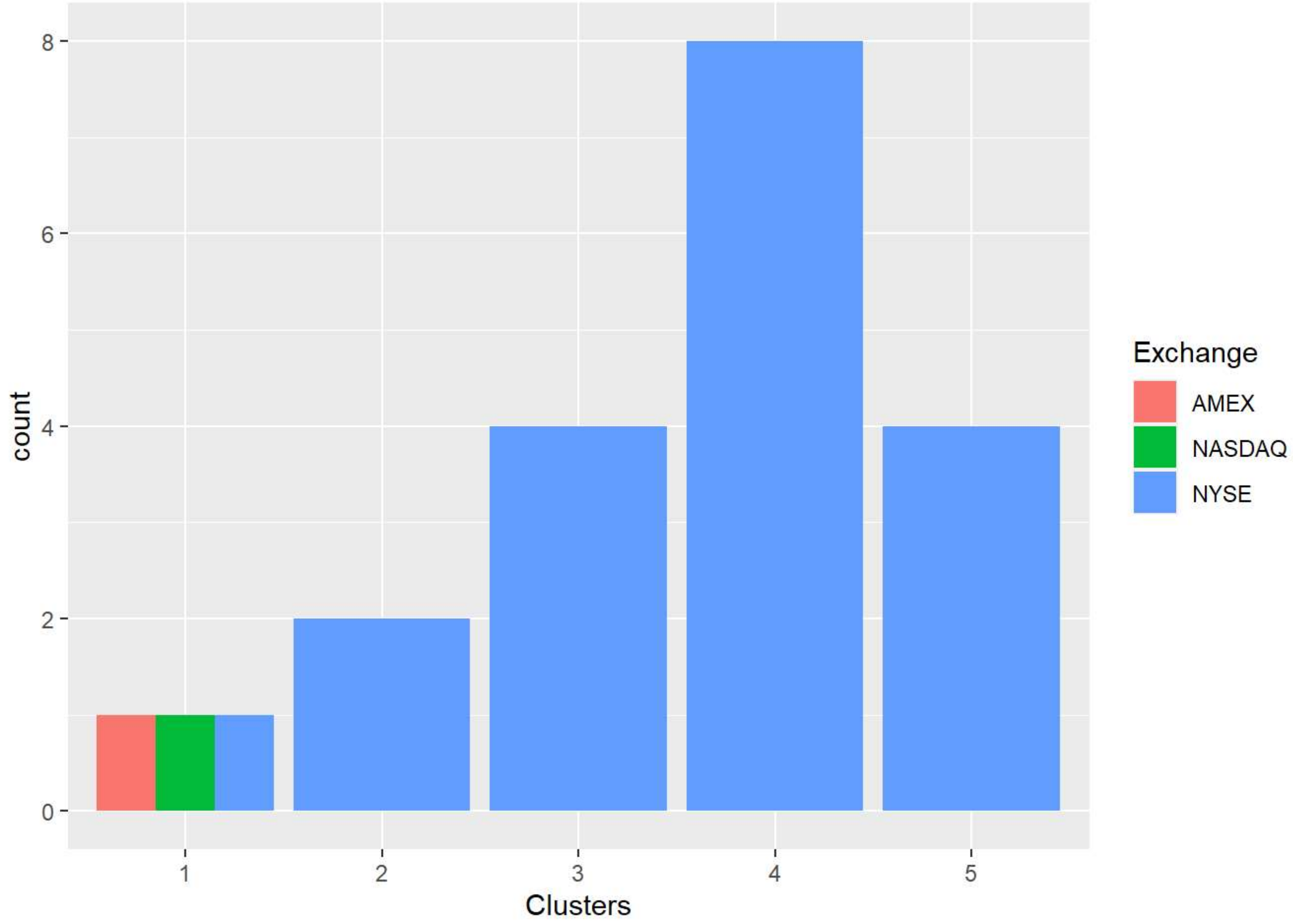
```
RD <- Phaceut_RD[12:14] %>% mutate(Clusters=k5$cluster)
ggplot(RD, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position='dodge')+labs(x ='Clusters')
```



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



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



```
#The graphs above show that there is a faint pattern in the clusters.

#Considering the fact that Cluster 1 has a distinct Hold and Moderate Buy median, a different co
unt from the US and Germany, and a different nation count, the firms are evenly distributed thro
ughout AMEX,NASDAQ, and NYSE.

#The cluster 2 is only listed on the NYSE, has equal Hold and Moderate Buy
#medians, and is evenly divided across the US and Canada.

#The Cluster 3 has trading on the NYSE and has equal Moderate Buy and Sell medians, as well as a
distinct count from France, Ireland, and the United States.

#Cluster 4 has the highest Hold median, followed by Moderate Buy, Strong Buy, and Hold medians.
They are from the United States, the United Kingdom, and Switzerland, and they are traded on the
New York Stock Exchange.

#The Cluster 5 is spread out throughout the United States and the United Kingdom, has the same h
old and moderate buy medians, and is also traded on the NYSE.
```

```
#TASK 3
#Provide an appropriate name for each cluster using any or all of the variables in the dataset.

#Cluster 1 :- Buy Cluster
#Cluster 2 :- Sceptical Cluster
#Cluster 3 :- Moderate Buy Cluster
#Cluster 4 :- Hold Cluster
#Cluster 5 :- High Hold Cluster
```