kmeans

joshna katta

2022-11-06

```
#loading required packages
library(flexclust)
```

```
## Warning: package 'flexclust' was built under R version 4.2.2
```

```
## Loading required package: grid
```

```
## Loading required package: lattice
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

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

```
## ── Attaching packages
## ─────────────────────────────────────
## tidyverse 1.3.2 ──
```

```
## ✓ ggplot2 3.3.6      ✓ purrr   0.3.4
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.0      ✓ stringr 1.4.1
## ✓ readr   2.1.2      ✓ forcats 0.5.2
## ── Conflicts ──────────────────────────────────────── tidyverse_conflict
s() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

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

```
## Warning: package 'factoextra' was built under R version 4.2.2
```

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

```
pharma.df = read.csv("C:/Users/Pavan Chaitanya/Downloads/Pharmaceuticals.csv"
)
colMeans(is.na(pharma.df))
```

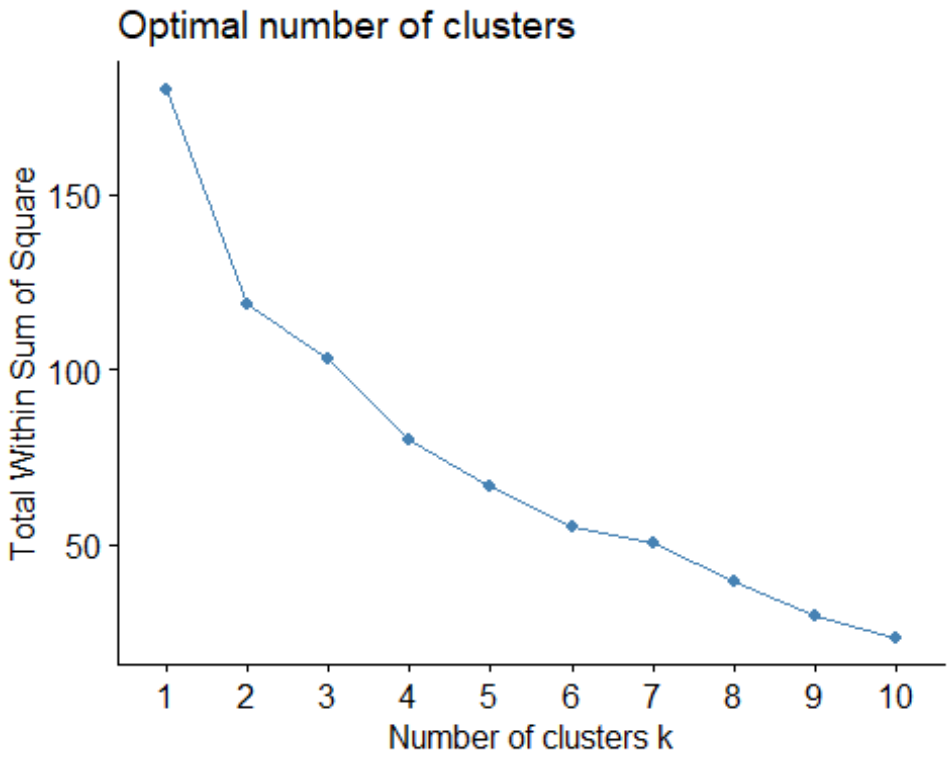
```
##                Symbol                  Name            Market_Cap
##                     0                     0                     0
```

```
##                   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
```

```
#normalizing the data
norm.pharma = scale(pharma.df[,-c(1:2,12:14)])
```

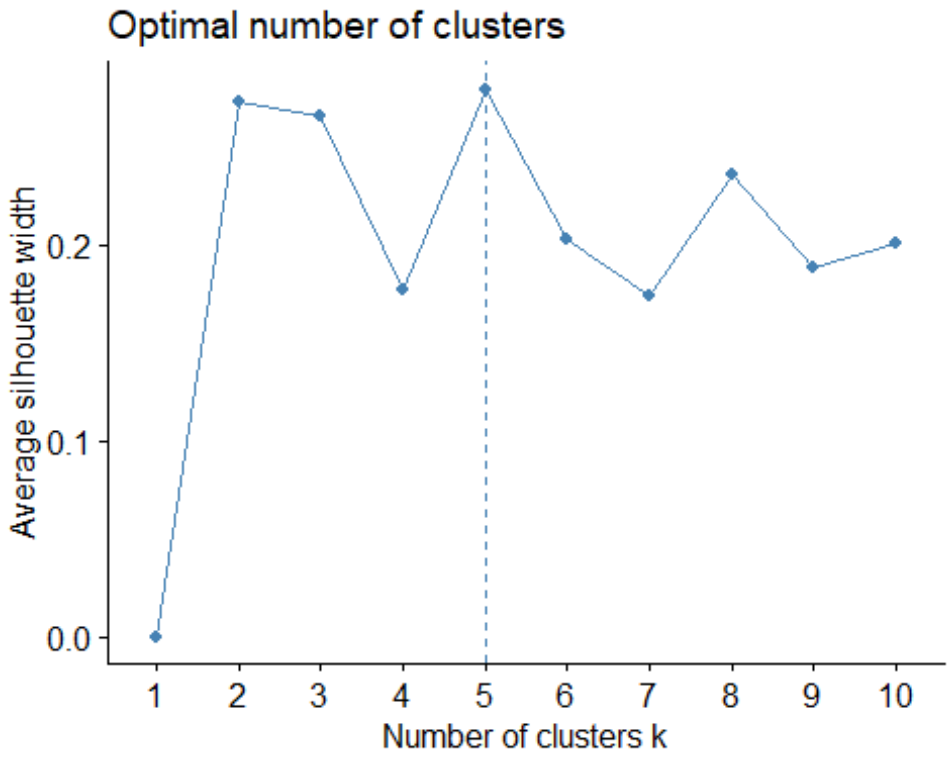
#using wss method finding the optimal k value

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



#finding the optimal k value from the silhouette method

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



here we got two different k values are from wss method is k= 2 and silhouette method was k = 5

```
#formulating the kmeans with wss
wss_kmeans = kmeans(norm.pharma,centers = 2,nstart=50)
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.withi
```

```
nss"
## [6] "betweenss"    "size"         "iter"         "ifault"

silhouette_kmeans =kmeans(norm.pharma,centers = 5,nstart = 50)
silhouette_kmeans

## K-means clustering with 5 clusters of sizes 3, 4, 8, 2, 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  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 4 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 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 -0.46807818  0.4671788       0.591242521
## 3 -0.27449312 -0.7041516       0.556954446
## 4 -0.14170336 -0.1168459      -1.416514761
## 5  0.06308085  1.5180158      -0.006893899
##
## Clustering vector:
##  [1] 3 4 3 3 5 1 3 1 5 3 2 1 2 5 2 3 2 4 3 5 3
##
## Within cluster sum of squares by cluster:
## [1] 15.595925  9.284424 21.879320  2.803505 12.791257
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withi
nss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

therefore by performing the wss method wet get clusters of size 11 and 10. therefore by performing the silhousette method wet got 5 clusters of sizes are 8,3,2,4,4

cluster plot for wss

```
fviz_cluster(wss_kmeans,data = norm.pharma)
```


```
fviz_cluster(silhouette_kmeans,data = norm.pharma)
```

5 clusters have been noticed from the above. The symbols/shapes in each cluster are 'centroids' of that specific cluster. Nstart value 25 and above is defined as no other centroid can be taken into

consideration until new data is being added. (b) Interpret the clusters with respect to the numerical variables used in forming the clusters.

```
#interpretation
silhouette_clusters = silhouette_kmeans$cluster
silhouette_cluster = as.data.frame(silhouette_clusters)
silhouette_2 = cbind(pharma.df,silhouette_clusters)
cluster_mean = silhouette_2 %>% group_by(silhouette_clusters) %>%
summarise_all("mean")

cluster_mean
```

```
## # A tibble: 5 × 15
##   silho…¹ Symbol  Name Marke…²  Beta PE_Ra…³   ROE   ROA Asset…⁴ Lever…⁵ R
ev_G…⁶
##     <int>  <dbl> <dbl>   <dbl> <dbl>   <dbl> <dbl> <dbl>   <dbl>   <dbl>
<dbl>
## 1       1     NA    NA    6.64 0.87     24.6  16.5  4.17   0.6     1.65
5.73
## 2       2     NA    NA  157.   0.48     22.2  44.4 17.7    0.95    0.22
18.5
## 3       3     NA    NA   55.8  0.414    20.3  28.7 12.7    0.738   0.371
5.59
## 4       4     NA    NA   31.9  0.405    69.5  13.2  5.6    0.75    0.475
12.1
## 5       5     NA    NA   13.1  0.598    17.7  14.6  6.2    0.425   0.635
30.1
## # … with 4 more variables: Net_Profit_Margin <dbl>,
## #   Median_Recommendation <dbl>, Location <dbl>, Exchange <dbl>, and
## #   abbreviated variable names ¹silhouette_clusters, ²Market_Cap, ³PE_Rati
o,
## #   ⁴Asset_Turnover, ⁵Leverage, ⁶Rev_Growth
```

#c.pattern with variables 10 to 12.

```
library(hrbrthemes)
```

```
## Warning: package 'hrbrthemes' was built under R version 4.2.2
```

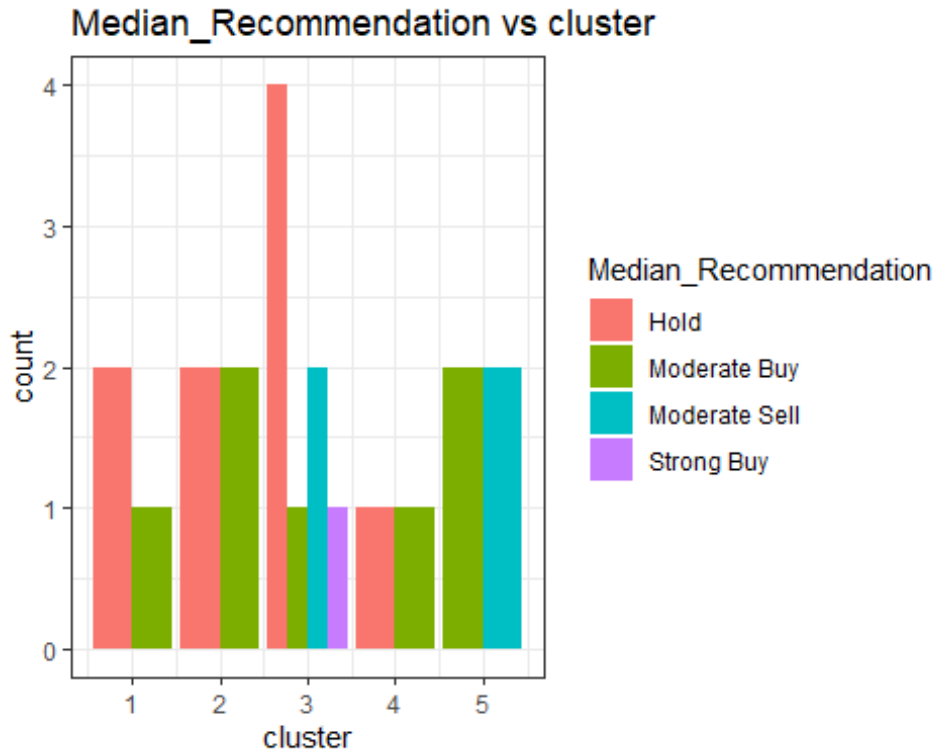
```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use th
ese themes.
```

```
##       Please use hrbrthemes::import_roboto_condensed() to install Roboto C
ondensed and
```

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

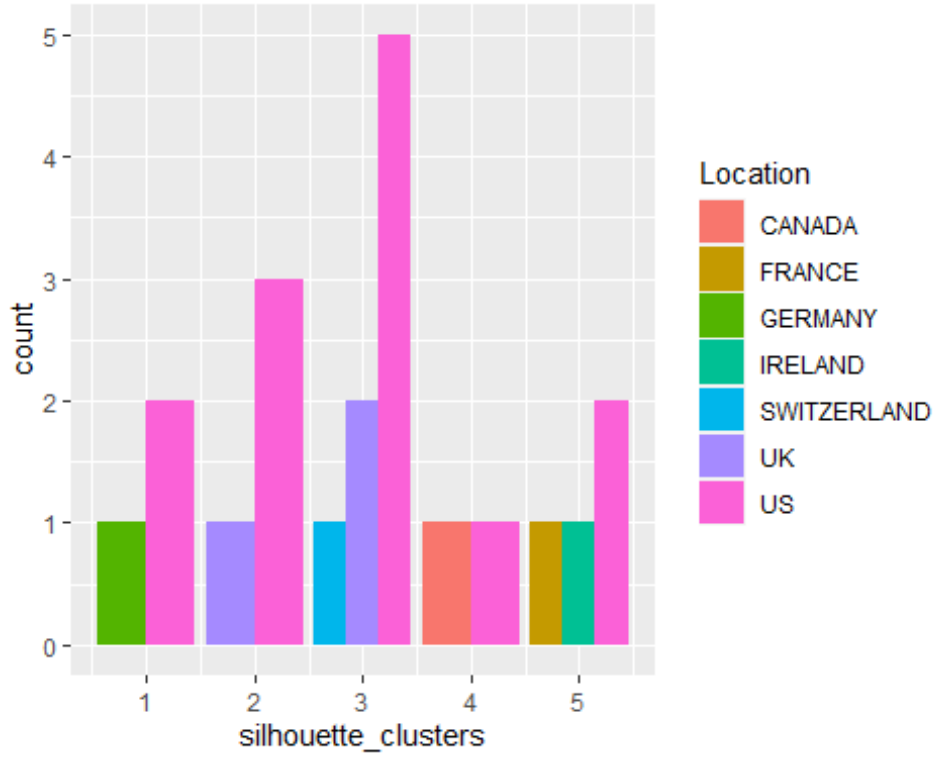
```
#median_recommendation vs cluster
pharma2 = pharma.df[,c(12,13,14)]
pharma3 =cbind(pharma2,silhouette_cluster)
```

```
ggplot(pharma3,aes(x=silhouette_clusters,fill= Median_Recommendation))+geom_b
ar(position = "dodge")+labs(title = "Median_Recommendation vs cluster",
    x ="cluster"
) +
  theme_bw()
```



location versus cluster

```
ggplot(pharma3,aes(x=silhouette_clusters,fill = Location))+ geom_bar(position
= "dodge")
```



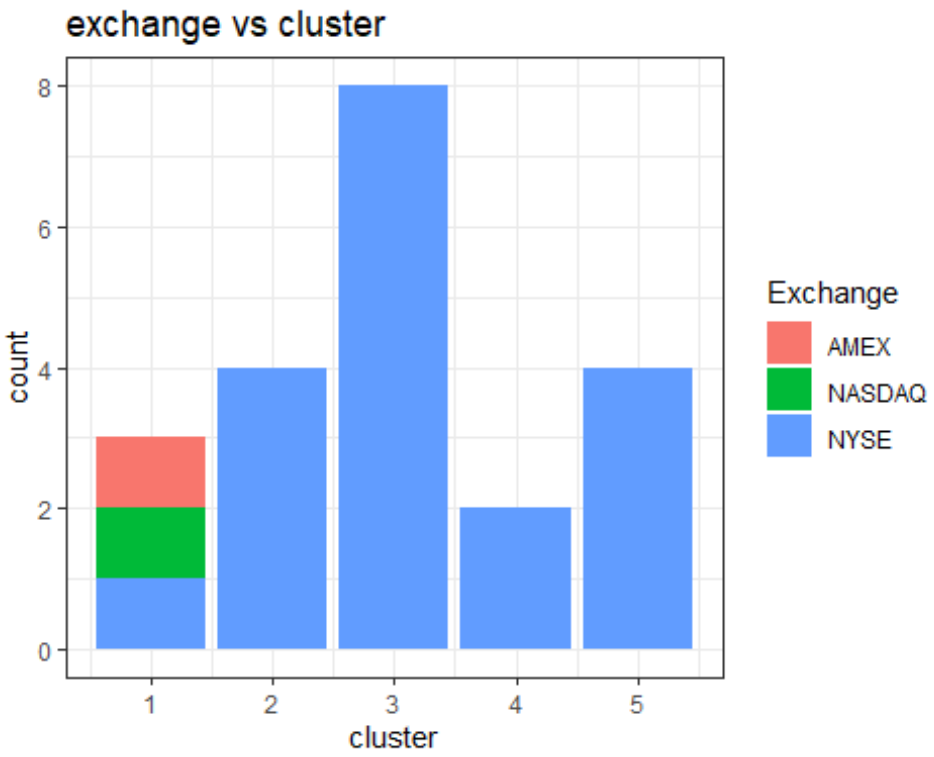

```
labs(
  title = "location vs cluster",
  x= "cluster"
)+
  theme_bw()
```

```
## NULL
```

cluster versus exchange

```
ggplot(pharma3,aes(x=silhouette_clusters, fill = Exchange,))+
geom_bar(postion = "dodge")+
labs(
title ="exchange vs cluster",
x="cluster"
)+
theme_bw()
```

```
## Warning: Ignoring unknown parameters: postion
```



#D)

1. cluster 1 is “Poorly Performing Pharma”, has low performance across all features and extremely high BETA and Leverage values.

2.cluster 2 “Overpriced Pharma”, with a high PE ratio

3.cluster 3 “Currently Profitable Pharma,” which has the lowest revenue growth but a solid net profit margin.

4.”Big Pharma” is in Cluster 4, and it has high market capitalization, ROE, ROA, asset turnover, and net profit margin.

5.The Sil Cluster 5 with the highest Rev Growth is “Future Potential Pharma.”