

Assignment 4 Machine Learning

```
library(caret) library(dcast) library(reshape) library(e1071) library(naivebayes)
library(klaR) library(bnclassify) library(rmarkdown) library(tinytex) library(cluster)
library(factoextra)
```

#Reading the Pharmaceutical data set

```
Myfile <- read.csv("Pharmaceuticals.csv")
head(Myfile)
```

```
##      Symbol      Name Market_Cap Beta PE_Ratio ROE ROA
Asset_Turnover
## 1    ABT Abbott Laboratories    68.44 0.32    24.7 26.4 11.8
0.7
## 2    AGN      Allergan, Inc.    7.58 0.41    82.5 12.9  5.5
0.9
## 3    AHM      Amersham plc    6.30 0.46    20.7 14.9  7.8
0.9
## 4    AZN      AstraZeneca PLC    67.63 0.52    21.5 27.4 15.4
0.9
## 5    AVE      Aventis    47.16 0.32    20.1 21.8  7.5
0.6
## 6    BAY      Bayer AG    16.90 1.11    27.9  3.9  1.4
0.6
```

```
##      Leverage Rev_Growth Net_Profit_Margin Median_Recommendation Location
Exchange
## 1    0.42      7.54      16.1      Moderate Buy      US
NYSE
## 2    0.60      9.16      5.5      Moderate Buy      CANADA
NYSE
## 3    0.27      7.05      11.2      Strong Buy      UK
NYSE
## 4    0.00      15.00      18.0      Moderate Sell      UK
NYSE
## 5    0.34      26.81      12.9      Moderate Buy      FRANCE
NYSE
## 6    0.00      -3.17      2.6      Hold      GERMANY
NYSE
```

```
X <- Myfile[,3:11]
head(X)
```

```
##      Market_Cap Beta PE_Ratio ROE ROA Asset_Turnover Leverage Rev_Growth
## 1    68.44 0.32    24.7 26.4 11.8    0.7    0.42    7.54
## 2    7.58 0.41    82.5 12.9  5.5    0.9    0.60    9.16
## 3    6.30 0.46    20.7 14.9  7.8    0.9    0.27    7.05
## 4    67.63 0.52    21.5 27.4 15.4    0.9    0.00    15.00
```

```
## 5      47.16 0.32      20.1 21.8  7.5      0.6      0.34      26.81
## 6      16.90 1.11      27.9  3.9  1.4      0.6      0.00      -3.17
## Net_Profit_Margin
## 1      16.1
## 2      5.5
## 3      11.2
## 4      18.0
## 5      12.9
## 6      2.6
```

#scale quantitative variables

```
scale <- scale(X)
```

```
head(scale)
```

```
##      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
```

##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.

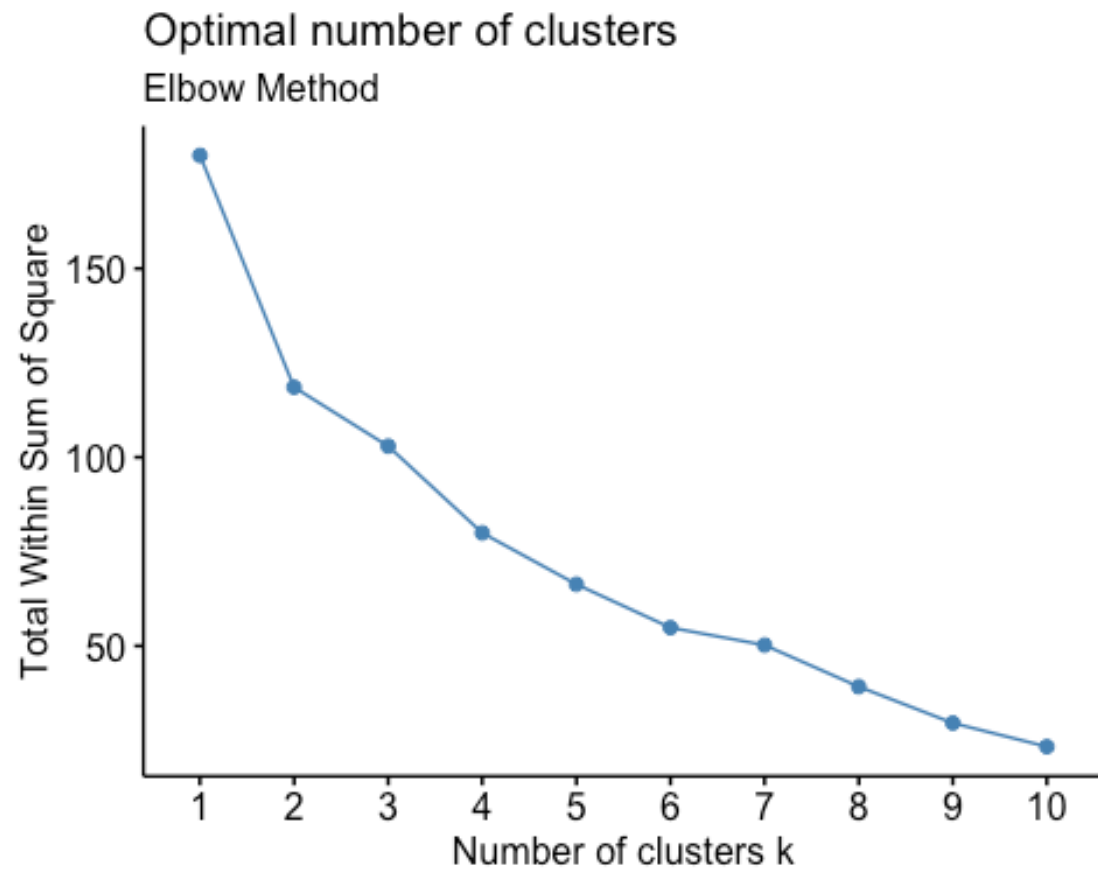
```
library(factoextra)
```

```
## Loading required package: ggplot2
```

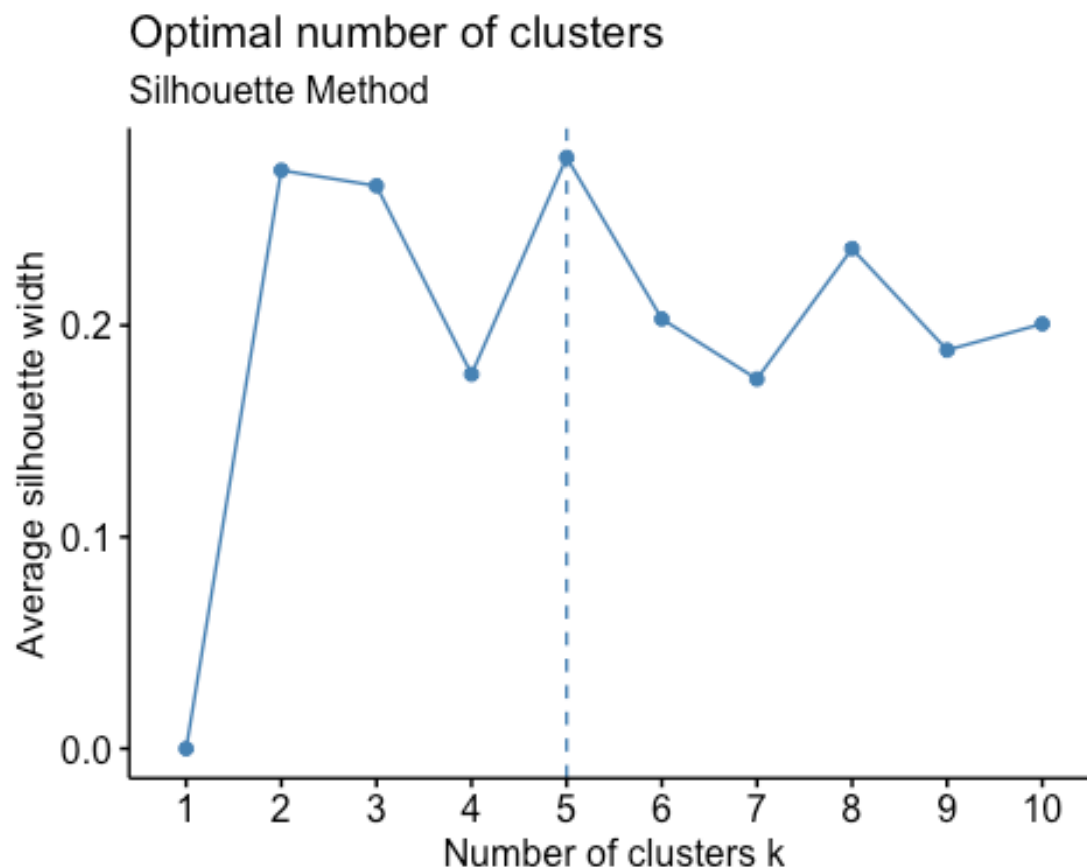
```
## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa
```

#Using the elbow method and silhouette method to cluster

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



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



Looking at the plots produced above in the silhouette method you can see that 5 clusters are appropriate. I choose to use the silhouette method because the k value is more clear at 5.

#Using the silhouette method to produce a cluster plot

```
set.seed(10)
```

```
kmeans5 <- kmeans(scale,centers=5,nstart=25)
```

```
kmeans5$centers
```

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

#5 clusters, data must be fitted to

```
fitkmeans<- kmeans(scale,5)
```

```
library(cluster)
aggregate(scale,by=list(fitkmeans$cluster), FUN=mean)
```

```
##   Group.1 Market_Cap      Beta    PE_Ratio      ROE      ROA
## 1      1  0.6733825 -0.3586419 -0.27635122  0.6565978  0.8344159
## 2      2 -0.9767669  1.2630872  0.03299122 -0.1123792 -1.1677918
## 3      3 -0.5246281  0.4451409  1.84984387 -1.0404550 -1.1865838
## 4      4 -0.7307042 -0.4214928 -0.34867046 -0.5780744 -0.6181243
## 5      5 -0.9668697  1.5162611 -0.57398880 -0.8382671 -0.9892673
##   Asset_Turnover    Leverage Rev_Growth Net_Profit_Margin
## 1  4.612656e-01 -0.33310678 -0.2902163      0.6823310
## 2 -4.612656e-01  3.74279705 -0.6327607     -1.2488842
## 3  1.480297e-16 -0.34435439 -0.5769454     -1.6095439
## 4 -2.306328e-01 -0.02651224  0.5327995     -0.4793074
## 5 -1.845062e+00  0.53024482  1.7123890      0.2445520
```

```
Y <- data.frame(scale,fitkmeans$cluster)
Y
```

```
##   Market_Cap      Beta    PE_Ratio      ROE      ROA
## 1  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121
## 2 -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871
## 3 -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700
## 4  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259
## 5 -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461
## 6 -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612
## 7 -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498
## 8 -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918
## 9 -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553
## 10 0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770
## 11 1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364
## 12 -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905
## 13 1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544
## 14 -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792
## 15 1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577
```

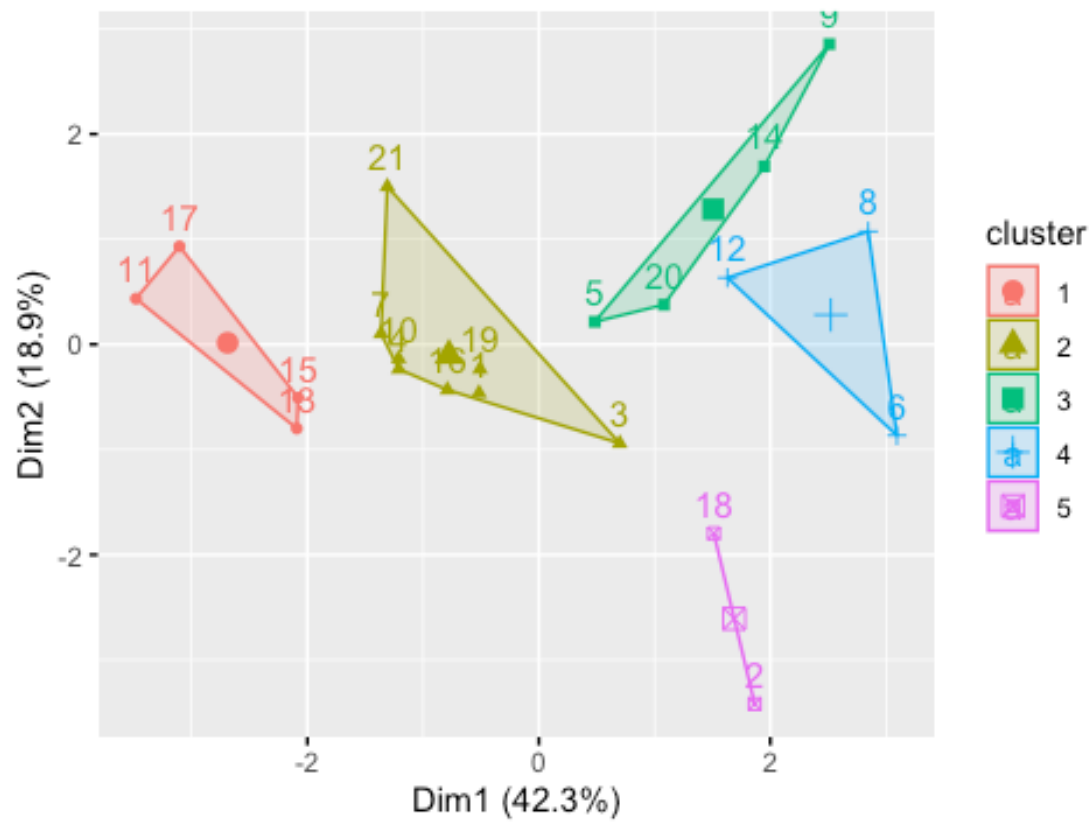
```

1.8450624
## 16  0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598    -
0.9225312
## 17  2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239
0.4612656
## 18 -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030    -
0.4612656
## 19 -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929
0.4612656
## 20 -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905    -
0.9225312
## 21 -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849    -
0.4612656
##      Leverage  Rev_Growth Net_Profit_Margin fitkmeans.cluster
## 1  -0.21209793 -0.52776752      0.06168225      1
## 2   0.01828430 -0.38113909     -1.55366706      3
## 3  -0.40408312 -0.57211809     -0.68503583      4
## 4  -0.74965647  0.14744734      0.35122600      1
## 5  -0.31449003  1.21638667     -0.42597037      4
## 6  -0.74965647 -1.49714434     -1.99560225      3
## 7  -0.02011273 -0.96584257      0.74744375      1
## 8   3.74279705 -0.63276071     -1.24888417      2
## 9   0.61983791  1.88617085     -0.36501379      5
## 10 -0.07130879 -0.64814764      1.17413980      1
## 11 -0.31449003  0.76926048      0.82363947      1
## 12  1.10620040  0.05603085     -0.71551412      4
## 13 -0.62166634 -0.36213170      0.33598685      1
## 14  0.44065173  1.53860717      0.85411776      5
## 15 -0.39128411  0.36014907     -0.24310064      1
## 16 -0.67286239 -1.45369888      1.02174835      1
## 17 -0.54487226  1.10143723      1.44844440      1
## 18 -0.30169102  0.14744734     -1.27936246      3
## 19 -0.74965647 -0.43544591      0.29026942      1
## 20 -0.49367621  1.43089863     -0.09070919      4
## 21  0.68383297 -1.17763919      1.49416183      1

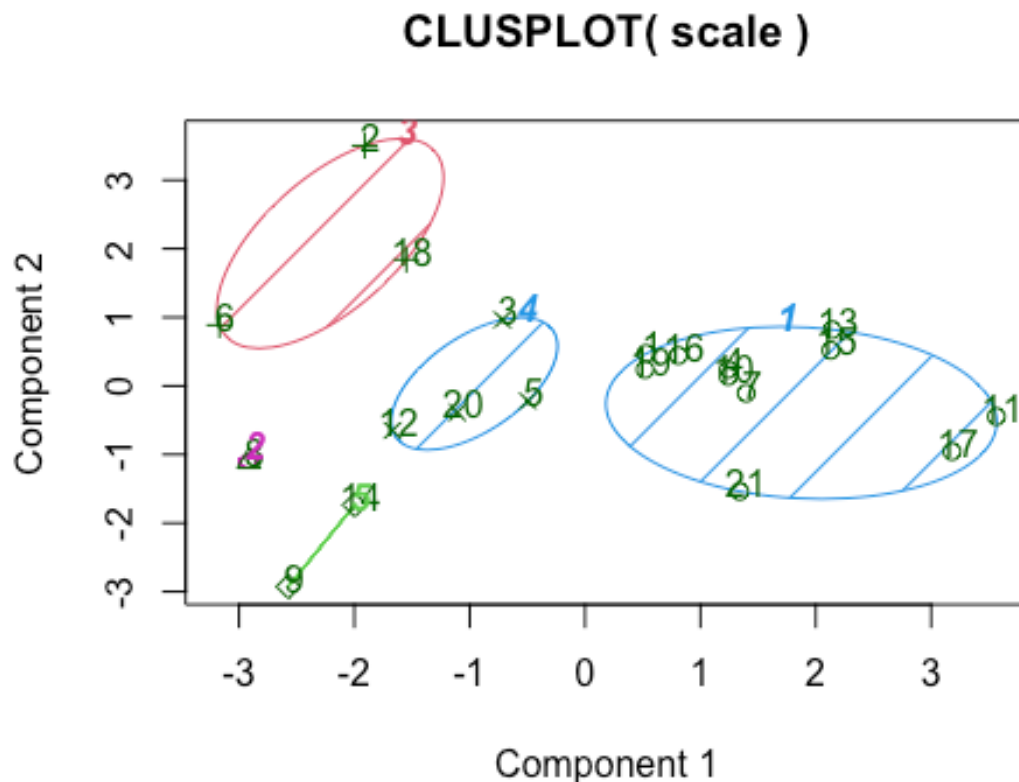
```

```
fviz_cluster(kmeans5, data = scale)
```

Cluster plot



```
#visualize the cluster
clusplot(scale,fitkmeans$cluster, color = TRUE, shade = TRUE, labels= 2,
lines = 0)
```



These two components explain 61.23 % of the point vari

##2 Interpret the clusters with respect to the numerical variables used in forming the clusters.

#Cluster 1: rows 1, 4, 7, 10, 11, 13, 15, 16, 17, 19, 21

#Cluster 2: row 8

#Cluster 3: rows 2, 6, 18

#Cluster 4: rows 3, 5, 12, 20

#Cluster 5: rows 9, 14

Analyzing the clusters in respect to the mean

#Cluster 1 has the highest market cap, highest ROE, highest ROA, highest asset turnover, and the highest net profit margin.

#Cluster 2 has the lowest market cap, lowest asset turnover, highest Leverage, and the lowest rev growth.

#Cluster 3 has the highest PE ratio, lowest ROE, lowest ROA, lowest Leverage, and the lowest net profit margin.

#Cluster 4 has the lowest beta.

#Cluster 5 has the highest beta, lowest PE ratio, and the highest rev growth.

##3 Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

#Using the clusters and the .csv file I can conclude that:

#Cluster 1 has a majority hold recommendation.

#Cluster 2 has majority moderate buy.

#Cluster 3 has majority hold.

#Cluster 4 has equal on strong buy, moderate buy, hold, moderate sell.

#Cluster 5 has equal moderate sell and moderate buy.

#The pattern that I see in respect to the numerical variables is that Cluster 1 and Cluster 3 have a majority of hold recommendation. Cluster 2 and 5 have a majority of the moderate buy.

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

Cluster 1 new name = High Market cap, ROE, ROA, asset turnover, and net profit margin.

Cluster 2 new name= Lowest market cap, asset turnover, rev growth and highest Leverage.

Cluster 3 new name= highest PE ratio, Lowest ROE, ROA, Leverage, and net profit margin

Cluster 4 new name= Lowest beta.

Cluster 5 new name= highest rev growth, beta and Lowest PE ratio.