Assignment 4

Kerrie Mars

2022-11-03

setwd("/Users/Kyle Mars/Desktop/Kerrie/Machine Learning")  
  
library("tidyverse") #data manipulation

## Warning: package 'tidyverse' was built under R version 4.1.3

## -- 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 2.1.3 v forcats 0.5.2

## Warning: package 'ggplot2' was built under R version 4.1.3

## Warning: package 'tibble' was built under R version 4.1.3

## Warning: package 'tidyr' was built under R version 4.1.3

## Warning: package 'readr' was built under R version 4.1.3

## Warning: package 'dplyr' was built under R version 4.1.3

## Warning: package 'stringr' was built under R version 4.1.3

## Warning: package 'forcats' was built under R version 4.1.3

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library("factoextra") #clustering algorithms and visualization

## Warning: package 'factoextra' was built under R version 4.1.3

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

library("ISLR")

## Warning: package 'ISLR' was built under R version 4.1.3

library("flexclust")

## Warning: package 'flexclust' was built under R version 4.1.3

## Loading required package: grid  
## Loading required package: lattice  
## Loading required package: modeltools

## Warning: package 'modeltools' was built under R version 4.1.1

## Loading required package: stats4

#clear existing data in Environment  
rm(list=ls())  
  
set.seed(1)  
  
#open dataset  
pharmaceuticals <- read.csv("Pharmaceuticals.csv", header=TRUE)  
head(pharmaceuticals)

## 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

#remove non-numerical columns from the dataset  
pharmaceuticals <- pharmaceuticals[, -c(1,2,12,13,14)]  
#scale the data frame (z-score) / normalize data  
pharmaceuticals <- scale(pharmaceuticals)  
  
summary(pharmaceuticals)

## Market\_Cap Beta PE\_Ratio ROE   
## Min. :-0.9768 Min. :-1.3466 Min. :-1.3404 Min. :-1.4515   
## 1st Qu.:-0.8763 1st Qu.:-0.6844 1st Qu.:-0.4023 1st Qu.:-0.7223   
## Median :-0.1614 Median :-0.2560 Median :-0.2429 Median :-0.2118   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.2762 3rd Qu.: 0.4841 3rd Qu.: 0.1495 3rd Qu.: 0.3450   
## Max. : 2.4200 Max. : 2.2758 Max. : 3.4971 Max. : 2.4597   
## ROA Asset\_Turnover Leverage Rev\_Growth   
## Min. :-1.7128 Min. :-1.8451 Min. :-0.74966 Min. :-1.4971   
## 1st Qu.:-0.9047 1st Qu.:-0.4613 1st Qu.:-0.54487 1st Qu.:-0.6328   
## Median : 0.1289 Median :-0.4613 Median :-0.31449 Median :-0.3621   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.00000 Mean : 0.0000   
## 3rd Qu.: 0.8430 3rd Qu.: 0.9225 3rd Qu.: 0.01828 3rd Qu.: 0.7693   
## Max. : 1.8389 Max. : 1.8451 Max. : 3.74280 Max. : 1.8862   
## Net\_Profit\_Margin   
## Min. :-1.99560   
## 1st Qu.:-0.68504   
## Median : 0.06168   
## Mean : 0.00000   
## 3rd Qu.: 0.82364   
## Max. : 1.49416

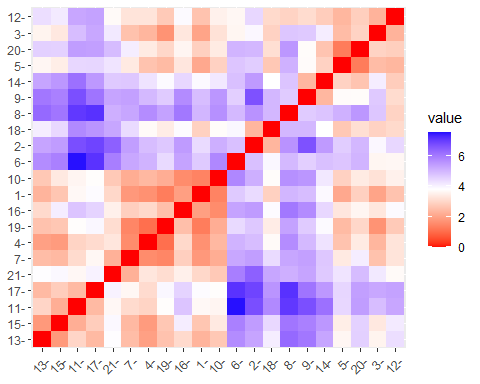
head(pharmaceuticals)

## 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

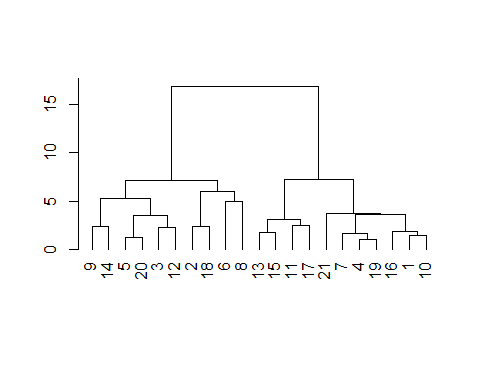
pharmaceuticals

## 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  
## [7,] -0.1078688 -0.10015669 -0.70887325 0.59693581 0.8617498 0.9225312  
## [8,] -0.9767669 1.26308721 0.03299122 -0.11237924 -1.1677918 -0.4612656  
## [9,] -0.9704532 2.15893320 -1.34037772 -0.70899938 -1.0174553 -1.8450624  
## [10,] 0.2762415 -1.34655112 0.14948233 0.34502953 0.5610770 -0.4612656  
## [11,] 1.0999201 -0.68440408 -0.45749769 2.45971647 1.8389364 1.3837968  
## [12,] -0.9393967 0.48409069 -0.34100657 -0.29136529 -0.6979905 -0.4612656  
## [13,] 1.9841758 -0.25595600 0.18013789 0.18593083 1.0872544 0.9225312  
## [14,] -0.9632863 0.87358895 0.19240011 -0.96753478 -0.9610792 -1.8450624  
## [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  
## [1,] -0.21209793 -0.52776752 0.06168225  
## [2,] 0.01828430 -0.38113909 -1.55366706  
## [3,] -0.40408312 -0.57211809 -0.68503583  
## [4,] -0.74965647 0.14744734 0.35122600  
## [5,] -0.31449003 1.21638667 -0.42597037  
## [6,] -0.74965647 -1.49714434 -1.99560225  
## [7,] -0.02011273 -0.96584257 0.74744375  
## [8,] 3.74279705 -0.63276071 -1.24888417  
## [9,] 0.61983791 1.88617085 -0.36501379  
## [10,] -0.07130879 -0.64814764 1.17413980  
## [11,] -0.31449003 0.76926048 0.82363947  
## [12,] 1.10620040 0.05603085 -0.71551412  
## [13,] -0.62166634 -0.36213170 0.33598685  
## [14,] 0.44065173 1.53860717 0.85411776  
## [15,] -0.39128411 0.36014907 -0.24310064  
## [16,] -0.67286239 -1.45369888 1.02174835  
## [17,] -0.54487226 1.10143723 1.44844440  
## [18,] -0.30169102 0.14744734 -1.27936246  
## [19,] -0.74965647 -0.43544591 0.29026942  
## [20,] -0.49367621 1.43089863 -0.09070919  
## [21,] 0.68383297 -1.17763919 1.49416183  
## attr(,"scaled:center")  
## Market\_Cap Beta PE\_Ratio ROE   
## 57.6514286 0.5257143 25.4619048 25.7952381   
## ROA Asset\_Turnover Leverage Rev\_Growth   
## 10.5142857 0.7000000 0.5857143 13.3709524   
## Net\_Profit\_Margin   
## 15.6952381   
## attr(,"scaled:scale")  
## Market\_Cap Beta PE\_Ratio ROE   
## 58.6029595 0.2567406 16.3102568 15.0849752   
## ROA Asset\_Turnover Leverage Rev\_Growth   
## 5.3213988 0.2167948 0.7813103 11.0483351   
## Net\_Profit\_Margin   
## 6.5620482

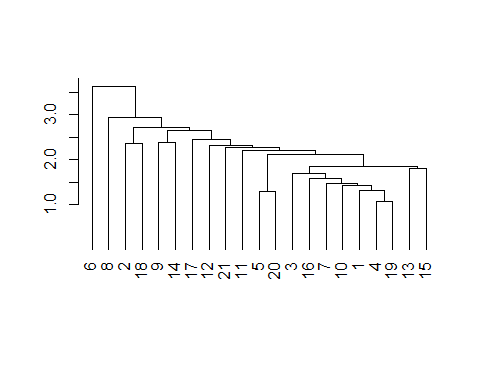
distance <- dist(pharmaceuticals, method = "euclidean")  
  
fviz\_dist(distance)



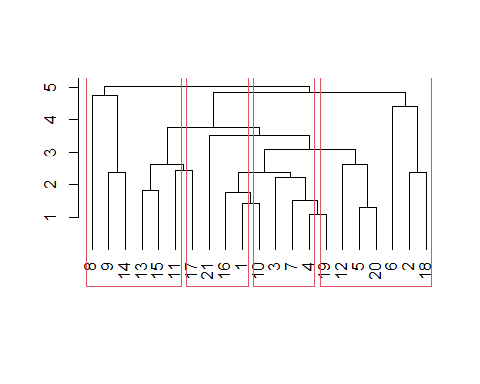
#in hclust() set argument method = to ward.D  
hc1 <- hclust(distance, method = "ward.D") # cluster analysis ward is most robust method  
plot(hc1, hang = -1, ann = FALSE)



#based on the above plot, there should be 4 clusters  
  
hc2 <- hclust(distance, method = "single")  
plot(hc2, hang = -1, ann = FALSE)



hc3 <- hclust(distance, method = "average")  
plot(hc3, hang = -1, ann = FALSE)  
  
rect.hclust(hc1, k=4)



#best option above  
  
#run kmeans algorithm using an inital value of k=4 to cluster the companies  
k4 <- kmeans(pharmaceuticals, centers = 4, nstart = 25) # k = 4, number of restarts = 25  
k4$centers #centroids / output the centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 1.69558112 -0.1780563 -0.1984582 1.2349879 1.3503431 1.153164e+00  
## 2 -0.03142211 -0.4360989 -0.3172485 0.1950459 0.4083915 1.729746e-01  
## 3 -0.82617719 0.4775991 -0.3696184 -0.5631589 -0.8514589 -9.994088e-01  
## 4 -0.52462814 0.4451409 1.8498439 -1.0404550 -1.1865838 1.480297e-16  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.4680782 0.4671788 0.5912425  
## 2 -0.2744931 -0.7041516 0.5569544  
## 3 0.8502201 0.9158889 -0.3319956  
## 4 -0.3443544 -0.5769454 -1.6095439

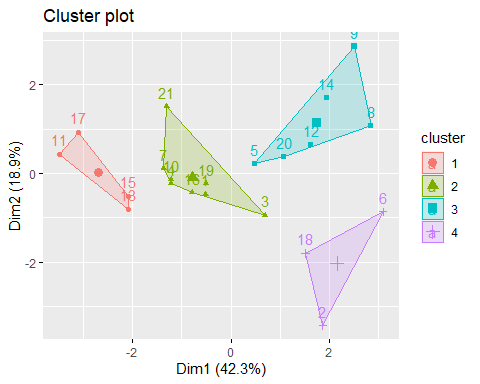
k4$size #size of each cluster

## [1] 4 8 6 3

k4$cluster[15] #cluster of 15th observations in data set as an example

## [1] 1

fviz\_cluster(k4, data = pharmaceuticals) #visualize the 4 clusters



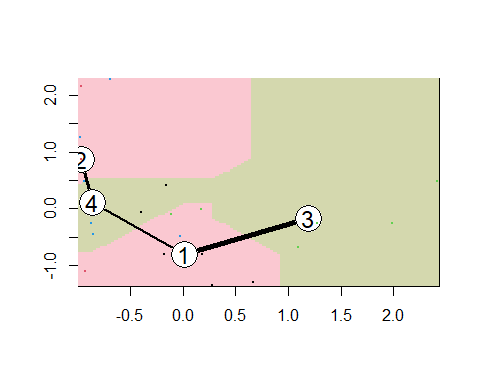
#Cluster using Manhattan distance  
#kmeans clustering, using Manhattan distance  
k4 = kcca(pharmaceuticals, k=4, kccaFamily("kmedians"))  
k4

## kcca object of family 'kmedians'   
##   
## call:  
## kcca(x = pharmaceuticals, k = 4, family = kccaFamily("kmedians"))  
##   
## cluster sizes:  
##   
## 1 2 3 4   
## 6 3 6 6

#Cluster 1 contains rows 11, 17, 13, 15  
#cluster 2 contains rows 21, 7, 10, 4, 16, 19, 3, 1  
#Cluster 3 contains rows 5, 20, 12, 14, 9, 8  
#Cluster 4 contains rows 18, 2, 6  
  
#apply the predict() function  
clusters\_index <- predict(k4)  
dist(k4@centers)

## 1 2 3  
## 2 3.721628   
## 3 2.367299 4.820685   
## 4 2.762659 2.931609 4.056830

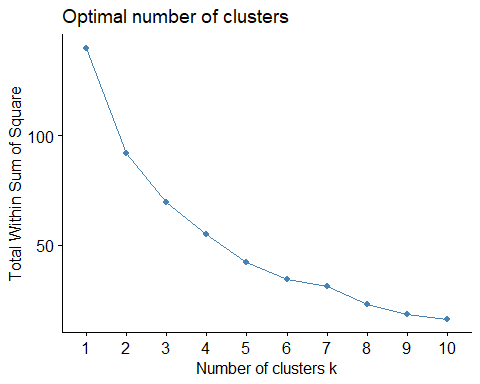
image(k4)  
points(pharmaceuticals, col=clusters\_index, pch=19, cex=0.3)



#compute Silhouette method  
fviz\_nbclust(pharmaceuticals, kmeans, method = "silhouette")



#this option shows that the optimal number of clusters is 5  
  
#determine k  
#use an "elbow chart"  
set.seed(1)  
  
pharmaceuticals <- pharmaceuticals[, -c(1,2,12,13,14)]  
#scale the data frame  
pharmaceuticals <- scale(pharmaceuticals)  
fviz\_nbclust(pharmaceuticals, kmeans, method = "wss")



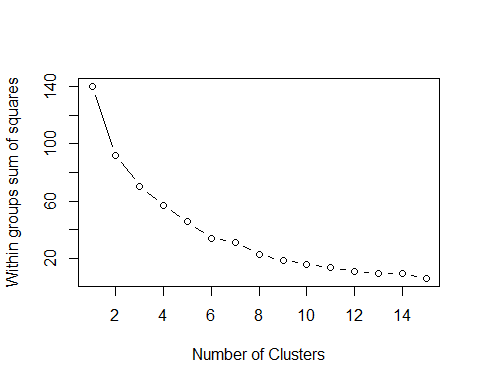
#the turning point of the data here is when k=5 where the curve is changing direction to more horizontal  
  
wss <- (nrow(pharmaceuticals)-1)\*sum(apply(pharmaceuticals,2,var))  
wss

## [1] 140

for(i in 2:15) wss[i] <- sum(kmeans(pharmaceuticals, centers=i)$withinss)  
wss

## [1] 140.000000 92.015114 70.122301 56.812869 45.726400 34.116465  
## [7] 31.314874 23.075057 18.514711 16.097883 13.639863 10.883256  
## [13] 9.686472 9.362772 6.062430

plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares")



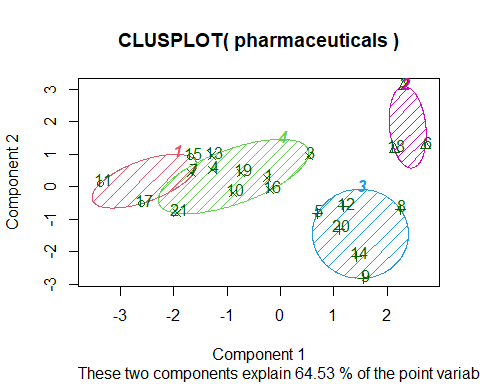
#this variation shows that the number of clusters should be 5 as is the initial turning point on the graph  
  
fit <- kmeans(pharmaceuticals, 4)  
aggregate(pharmaceuticals, by=list(fit$cluster), FUN=mean)

## Group.1 PE\_Ratio ROE ROA Asset\_Turnover Leverage Rev\_Growth  
## 1 1 -0.3246569 1.5846736 1.4380394 1.230042e+00 -0.4168821 0.7436156  
## 2 2 1.8498439 -1.0404550 -1.1865838 -7.401487e-17 -0.3443544 -0.5769454  
## 3 3 -0.3696184 -0.5631589 -0.8514589 -9.994088e-01 0.8502201 0.9158889  
## 4 4 -0.2619834 0.1940331 0.4838207 2.562587e-01 -0.3130679 -0.6661494  
## Net\_Profit\_Margin  
## 1 0.6763277  
## 2 -1.6095439  
## 3 -0.3319956  
## 4 0.5324025

#Cluster 2 has the highest PE\_ratio, lowest is cluster 3  
#Cluster 1 has the highest ROE, lowest is cluster 2  
#Cluster 1 has the highest ROA, lowest is cluster 2  
#Cluster 4 has the highest Asset\_Turnover, lowest is cluster 2  
#Cluster 3 has the highest leverage, lowest is cluster 1  
#Cluster 3 has the highest Rev\_Growth, lowest is cluster 4  
#Cluster 1 has the highest Net\_Profit\_Margin, lowest is cluster 2  
  
pharmaceuticals1 <- data.frame(pharmaceuticals, fit$cluster)  
pharmaceuticals1

## PE\_Ratio ROE ROA Asset\_Turnover Leverage Rev\_Growth  
## 1 -0.04671323 0.04009035 0.2416121 -2.008975e-16 -0.21209793 -0.52776752  
## 2 3.49706911 -0.85483986 -0.9422871 9.225312e-01 0.01828430 -0.38113909  
## 3 -0.29195768 -0.72225761 -0.5100700 9.225312e-01 -0.40408312 -0.57211809  
## 4 -0.24290879 0.10638147 0.9181259 9.225312e-01 -0.74965647 0.14744734  
## 5 -0.32874435 -0.26484883 -0.5664461 -4.612656e-01 -0.31449003 1.21638667  
## 6 0.14948233 -1.45146000 -1.7127612 -4.612656e-01 -0.74965647 -1.49714434  
## 7 -0.70887325 0.59693581 0.8617498 9.225312e-01 -0.02011273 -0.96584257  
## 8 0.03299122 -0.11237924 -1.1677918 -4.612656e-01 3.74279705 -0.63276071  
## 9 -1.34037772 -0.70899938 -1.0174553 -1.845062e+00 0.61983791 1.88617085  
## 10 0.14948233 0.34502953 0.5610770 -4.612656e-01 -0.07130879 -0.64814764  
## 11 -0.45749769 2.45971647 1.8389364 1.383797e+00 -0.31449003 0.76926048  
## 12 -0.34100657 -0.29136529 -0.6979905 -4.612656e-01 1.10620040 0.05603085  
## 13 0.18013789 0.18593083 1.0872544 9.225312e-01 -0.62166634 -0.36213170  
## 14 0.19240011 -0.96753478 -0.9610792 -1.845062e+00 0.44065173 1.53860717  
## 15 -0.40231769 0.98142435 0.8429577 1.845062e+00 -0.39128411 0.36014907  
## 16 -0.23677768 -0.52338423 0.1288598 -9.225312e-01 -0.67286239 -1.45369888  
## 17 -0.11415545 1.31287998 1.6322239 4.612656e-01 -0.54487226 1.10143723  
## 18 1.90298017 -0.81506519 -0.9047030 -4.612656e-01 -0.30169102 0.14744734  
## 19 -0.40231769 -0.21181593 0.5234929 4.612656e-01 -0.74965647 -0.43544591  
## 20 -0.43297324 -1.03382590 -0.6979905 -9.225312e-01 -0.49367621 1.43089863  
## 21 -0.75792214 1.92938746 0.5422849 -4.612656e-01 0.68383297 -1.17763919  
## Net\_Profit\_Margin fit.cluster  
## 1 0.06168225 4  
## 2 -1.55366706 2  
## 3 -0.68503583 4  
## 4 0.35122600 4  
## 5 -0.42597037 3  
## 6 -1.99560225 2  
## 7 0.74744375 4  
## 8 -1.24888417 3  
## 9 -0.36501379 3  
## 10 1.17413980 4  
## 11 0.82363947 1  
## 12 -0.71551412 3  
## 13 0.33598685 4  
## 14 0.85411776 3  
## 15 -0.24310064 1  
## 16 1.02174835 4  
## 17 1.44844440 1  
## 18 -1.27936246 2  
## 19 0.29026942 4  
## 20 -0.09070919 3  
## 21 1.49416183 4

library(cluster)  
clusplot(pharmaceuticals, fit$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)



#this visual also displays 4 clusters  
  
wssplot <- function(pharmaceuticals)  
  
#Our data shows that the optimal number of clusters is 4   
#graph displays two major groupings, and 4 minor groupings or clusters of pharmaceutical companies   
#Additionally the average height of these clusters is 4 or 5 on the graph  
#goal is to use the minimum height that still keeps the clusters meaningful more dissimilarity when the value is larger than 5  
  
#cluster 1 is only US pharmaceuticals  
#cluster 2 has the most Holds  
#cluster 3 is the only one with non NYSE  
#cluster 4 contains a mix of locations  
#overall, there is really not a pattern with the non numerical values in   
#cluster formation  
  
#Cluster 1 Highest ROE  
#Cluster 2 Lowest on most variables  
#Cluster 3 Highest Leverage  
#Cluster 4 Middle on most variables  
  
summary(pharmaceuticals)