ML Assignment 4

Sean Bradford

```
rm(list=ls())

library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.1
```

```
## -- 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   1.2.0      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
```

```
## Warning: package 'tibble' was built under R version 4.2.1
```

```
## Warning: package 'readr' was built under R version 4.2.1
```

```
## Warning: package 'dplyr' was built under R version 4.2.1
```

```
## Warning: package 'stringr' was built under R version 4.2.1
```

```
## Warning: package 'forcats' was built under R version 4.2.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ISLR)
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.2.1
```

```
## 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 4.2.1
```

```
## Loading required package: grid
## Loading required package: lattice
## Loading required package: modeltools
## Loading required package: stats4
```

```
pharm=read.csv('C:\\Users\\Sean\\OneDrive\\Desktop\\Grad School\\Machine Learning\\Module 6 - K-mean Clu
head(pharm)
```

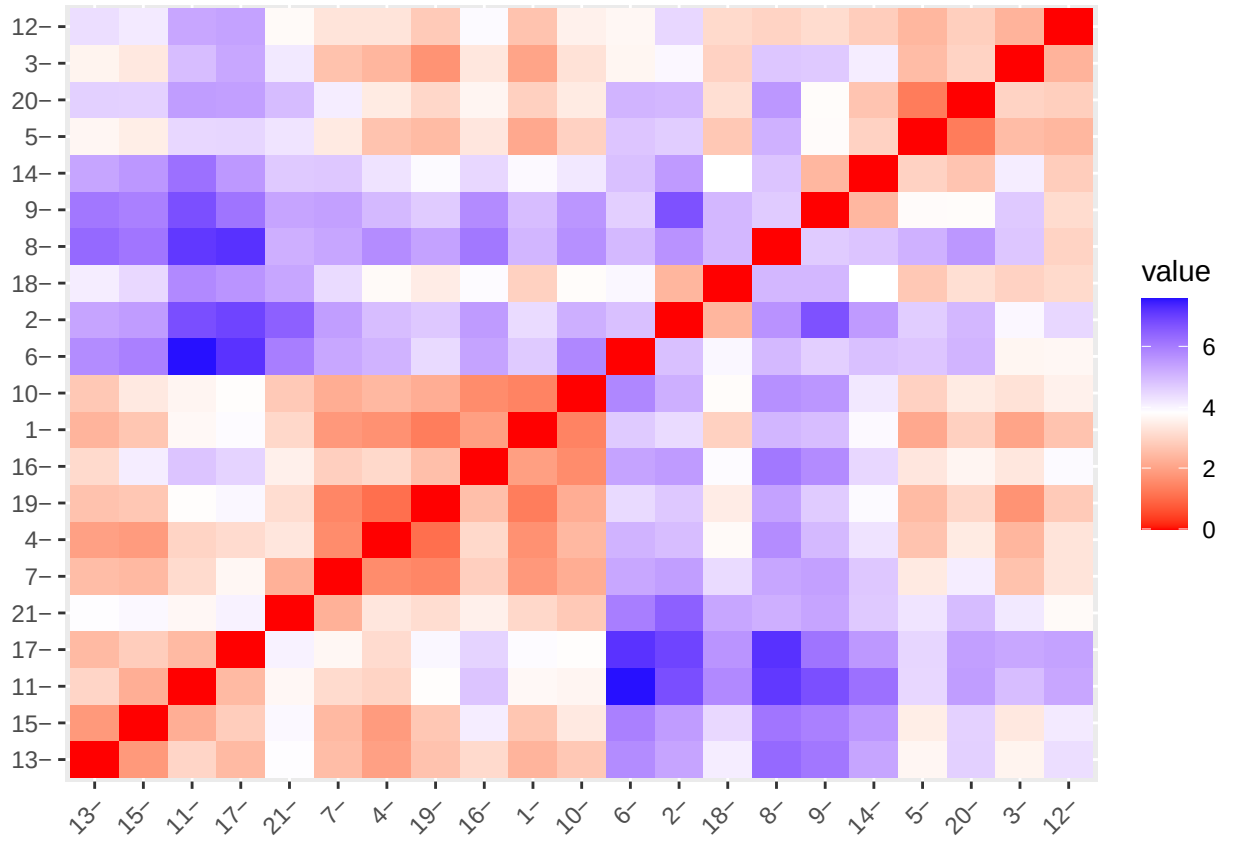
```
##   Symbol                Name Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover
## 1    ABT Abbott Laboratories      68.44 0.32     24.7 26.4 11.8            0.7
## 2    AGN      Allergan, Inc.       7.58 0.41     82.5 12.9  5.5            0.9
## 3    AHM        Amersham plc       6.30 0.46     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
```

A. 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.

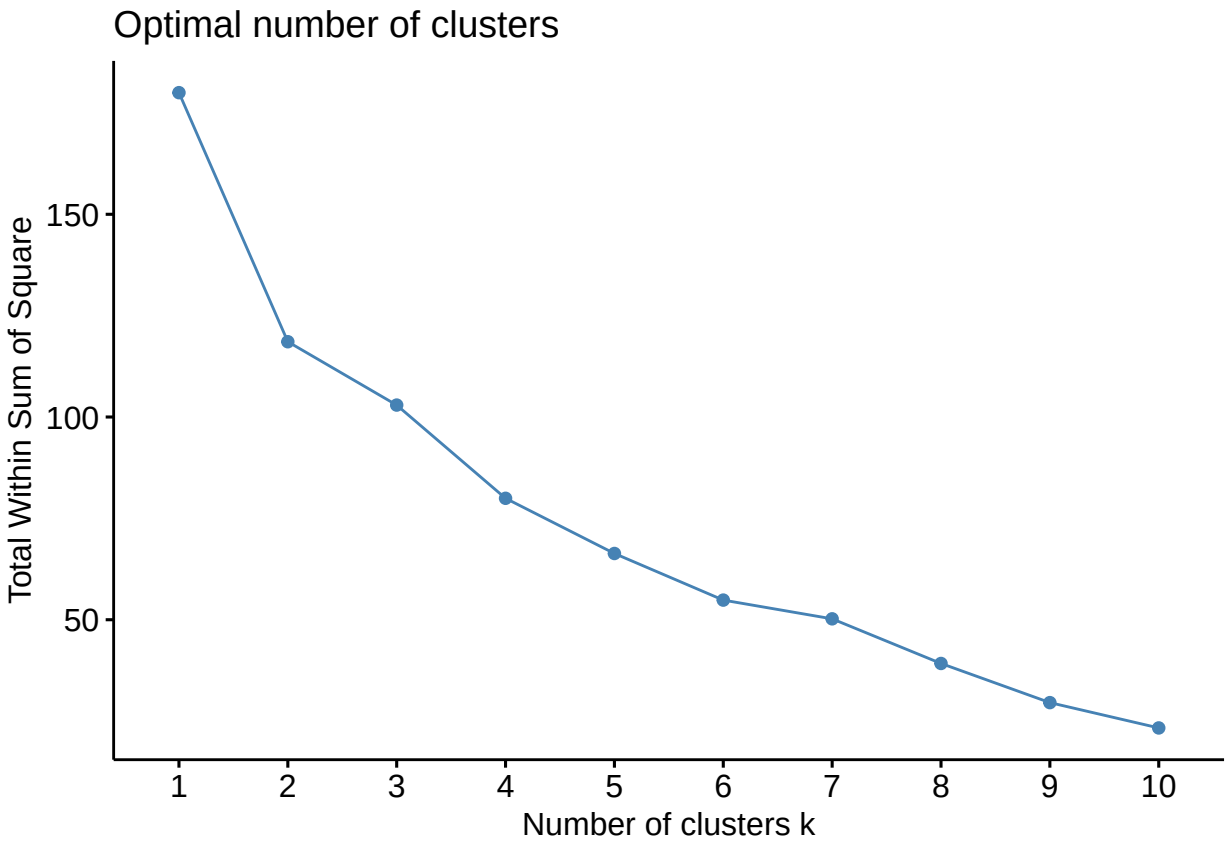
```
set.seed(111)
# Retain only quantitative variables from original df
dfp=data.frame(pharm[,c(3:11)])
summary(dfp)
```

```
##    Market_Cap          Beta           PE_Ratio          ROE
##  Min.   :  0.41   Min.   :0.1800   Min.   : 3.60   Min.   : 3.9
##  1st Qu.:  6.30   1st Qu.:0.3500   1st Qu.:18.90   1st Qu.:14.9
##  Median : 48.19   Median :0.4600   Median :21.50   Median :22.6
##  Mean   : 57.65   Mean   :0.5257   Mean   :25.46   Mean   :25.8
##  3rd Qu.: 73.84   3rd Qu.:0.6500   3rd Qu.:27.90   3rd Qu.:31.0
##  Max.   :199.47   Max.   :1.1100   Max.   :82.50   Max.   :62.9
##       ROA        Asset_Turnover    Leverage        Rev_Growth
##  Min.   : 1.40   Min.   :0.3    Min.   :0.0000   Min.   :-3.17
##  1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600   1st Qu.: 6.38
##  Median :11.20   Median :0.6    Median :0.3400   Median : 9.37
##  Mean   :10.51   Mean   :0.7    Mean   :0.5857   Mean   :13.37
##  3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000   3rd Qu.:21.87
##  Max.   :20.30   Max.   :1.1    Max.   :3.5100   Max.   :34.21
##  Net_Profit_Margin
##  Min.   : 2.6
##  1st Qu.:11.2
##  Median :16.1
##  Mean   :15.7
##  3rd Qu.:21.1
##  Max.   :25.5
```

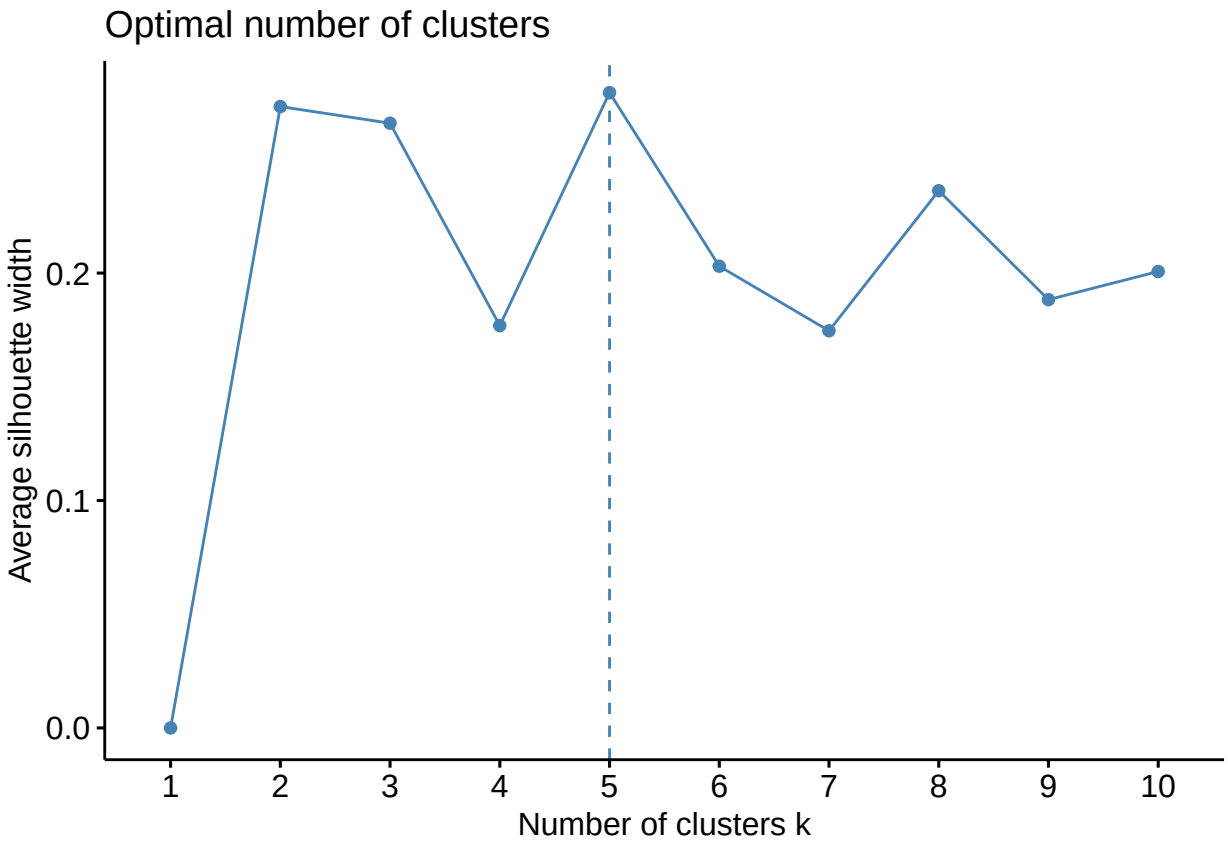
```
# Scaling the data (z-score)
dfp=scale(dfp)
dis=get_dist(dfp)
fviz_dist(dis)
```




```
# Determining k
fviz_nbclust(dfp,kmeans,"wss")
```




```
fviz_nbclust(dfp,kmeans,"silhouette")
```



Both charts indicate that 5 is the ideal number of clusters.

```
# k means clustering (Euclidean distance)
k.5=kmeans(dfp,centers = 5, nstart=25)
k.5
```

```
## K-means clustering with 5 clusters of sizes 8, 3, 2, 4, 4
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 3 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 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 -0.27449312 -0.7041516       0.556954446
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.14170336 -0.1168459      -1.416514761
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.06308085  1.5180158      -0.006893899
##
## Clustering vector:
##  [1] 1 3 1 1 5 2 1 2 5 1 4 2 4 5 4 1 4 3 1 5 1
##
## Within cluster sum of squares by cluster:
```

```
## [1] 21.879320 15.595925  2.803505  9.284424 12.791257
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
k.5$centers
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 3 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 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 -0.27449312 -0.7041516       0.556954446
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.14170336 -0.1168459      -1.416514761
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.06308085  1.5180158      -0.006893899
```

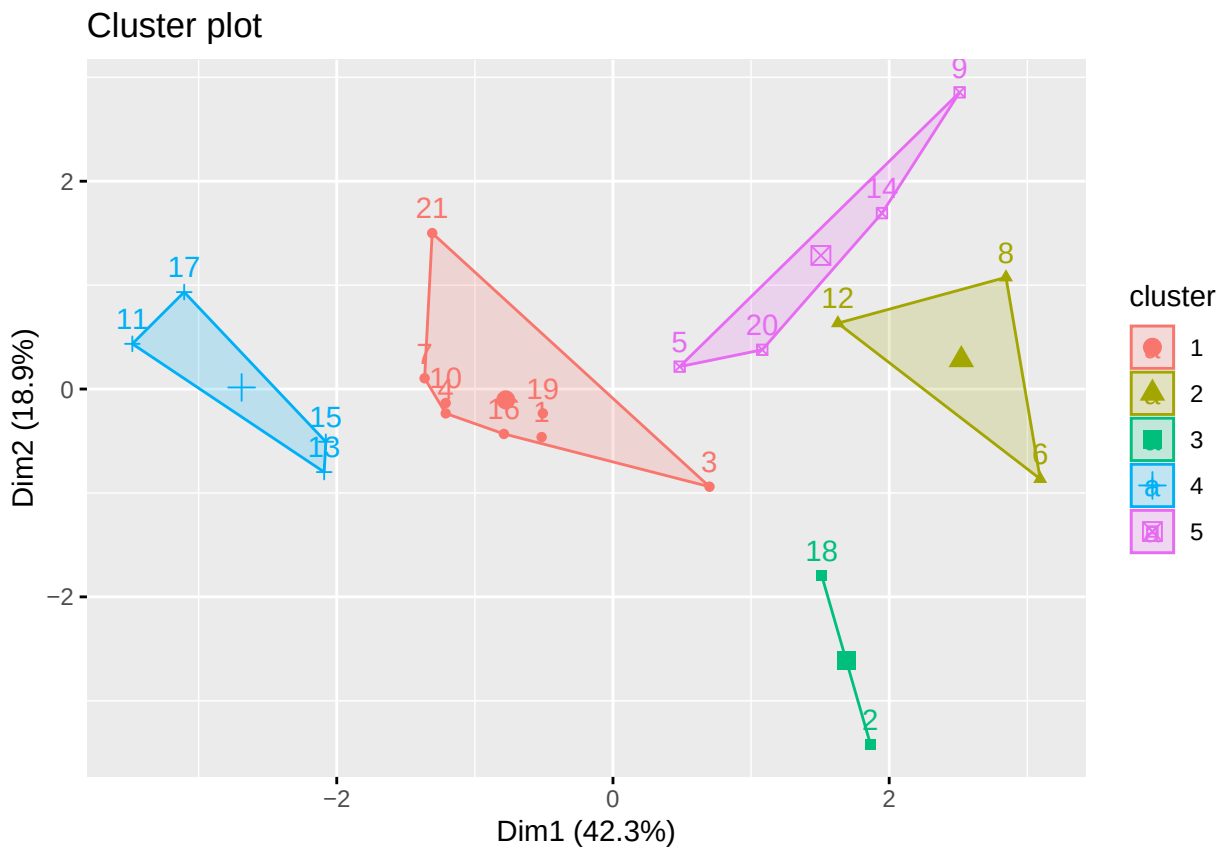
```
k.5$size
```

```
## [1] 8 3 2 4 4
```

```
k.5$cluster[6]
```

```
## [1] 2
```

```
fviz_cluster(k.5,data=dfp)
```



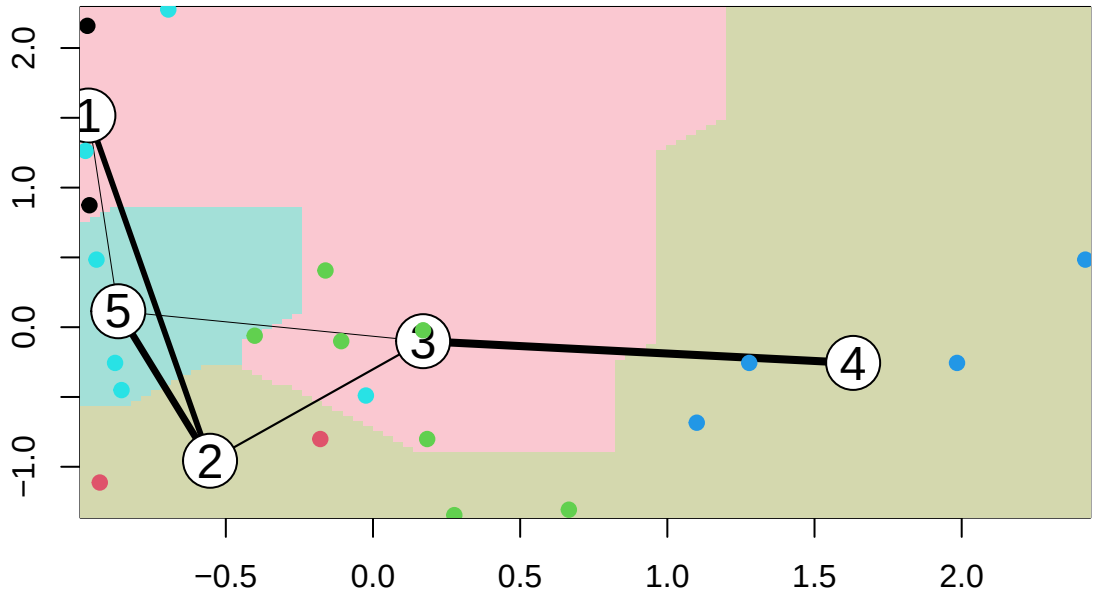

```
# k means clustering using Manhattan distance for comparison
set.seed(111)
kc.5=kcca(dfp, k=5, kccaFamily("kmedians"))
kc.5
```

```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = dfp, k = 5, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
## 1 2 3 4 5
## 2 2 7 4 6
```

```
clusters_index=predict(kc.5)
dist(kc.5@centers)
```

```
##          1        2        3        4
## 2 3.015849
## 3 4.127213 2.939894
## 4 5.555697 4.142701 2.608581
## 5 3.444192 2.437429 2.904788 4.751071
```

```
image(kc.5)
points(dfp,col=clusters_index,pch=19)
```



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

```
# Centroids for Euclidean distance model
k.5$centers
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 3 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 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 -0.27449312 -0.7041516       0.556954446
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.14170336 -0.1168459      -1.416514761
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.06308085  1.5180158      -0.006893899
```

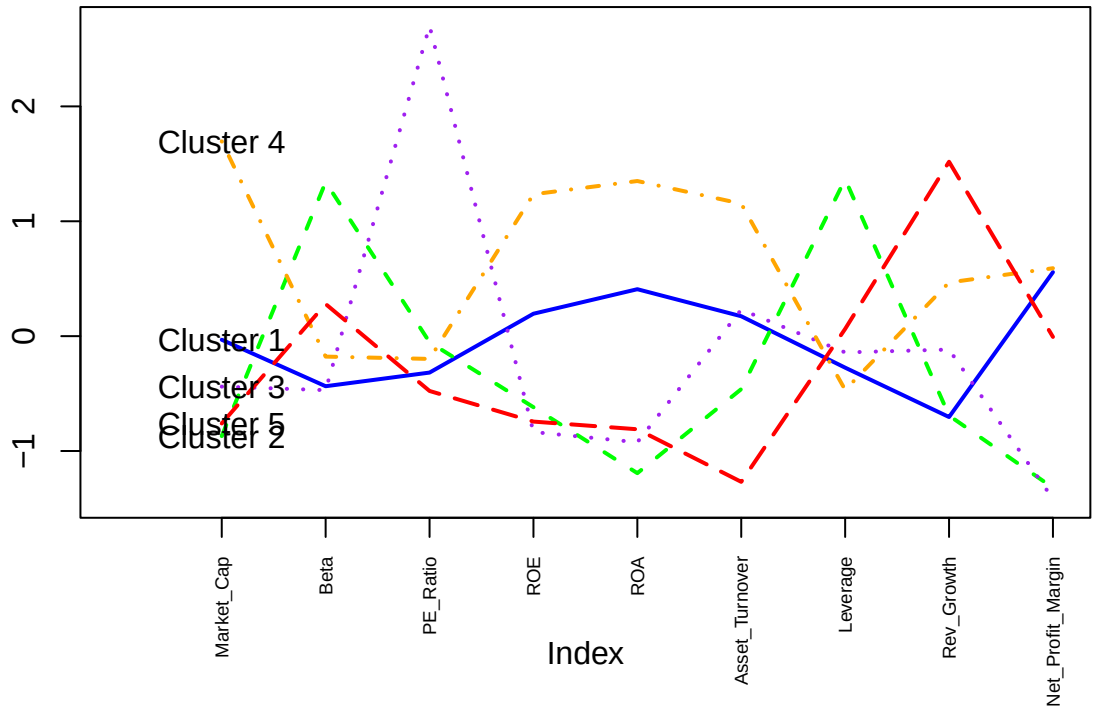
```
# Create profile plot of cluster centroids

dfp1=data.frame(pharm[,c(3:11)])
```

```
plot(c(0),xaxt='n',ylab="", type="l",
     ylim=c(min(k.5$centers),max(k.5$centers)),xlim=c(0,9))
axis(1,at=seq(1:9),labels = names(dfp1),cex.axis=0.55,las=3)

for (i in c(1:5))
  lines(k.5$centers[i,],lty=i,lwd=2,col= if(i %in% 1) {
    col="blue"
  } else if(i %in% 2) {
    col="green"
  } else if(i %in% 3) {
    col="purple"
  } else if(i %in% 4) {
    col="orange"
  } else {
    col="red"
  }
  )

text(x=1,y=k.5$centers[,1],labels=paste("Cluster",c(1:5)))
```



Characteristics of each cluster:

Cluster 1 (blue): Lower Beta; Higher Net Proft Margin Cluster 2 (green): Lower Market Cap and ROA; Higher Beta and Leverage Cluster 3 (purple): Higher P/E Ratio; Lower ROE and Net Profit Margin Cluster 4 (orange): Higher Market Cap, ROE, ROA & Asset Turnover; Lower leverage Cluster 5 (red): Higher Revenue Growth; Lower Asset Turnover

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

```
# Seperate clusters to analyze unused variables
c1=pharm[c(1,4,7,10,21,16,19,3),c(12:14)]
c2=pharm[c(12,8,6),c(12:14)]
c3=pharm[c(2,18),c(12:14)]
c4=pharm[c(11,17,13,15),c(12:14)]
c5=pharm[c(5,20,14,9),c(12:14)]
```

```
c1
```

```
##    Median_Recommendation    Location Exchange
## 1           Moderate Buy          US     NYSE
## 4          Moderate Sell          UK     NYSE
## 7          Moderate Sell          US     NYSE
## 10                 Hold          US     NYSE
## 21                 Hold          US     NYSE
## 16                 Hold SWITZERLAND     NYSE
## 19                 Hold          US     NYSE
## 3             Strong Buy          UK     NYSE
```

```
c2
```

```
##    Median_Recommendation Location Exchange
## 12                 Hold       US     AMEX
## 8           Moderate Buy       US   NASDAQ
## 6                  Hold  GERMANY     NYSE
```

```
c3
```

```
##    Median_Recommendation Location Exchange
## 2           Moderate Buy   CANADA     NYSE
## 18                 Hold       US     NYSE
```

```
c4
```

```
##    Median_Recommendation Location Exchange
## 11                 Hold       UK     NYSE
## 17          Moderate Buy       US     NYSE
## 13          Moderate Buy       US     NYSE
## 15                 Hold       US     NYSE
```

```
c5
```

```
##    Median_Recommendation Location Exchange
## 5           Moderate Buy   FRANCE     NYSE
## 20         Moderate Sell       US     NYSE
## 14          Moderate Buy       US     NYSE
## 9          Moderate Sell  IRELAND     NYSE
```

For Median Recommendation variable:

Cluster 1: Mainly Hold Cluster 2: Mainly Hold Cluster 3: 1 mod buy and 1 hold Cluster 4: 2 mod buy and 2 hold Cluster 5: 2 mod buy and 2 mod sell

Clusters 1 and 2 have similar Revenue Growth and P/E Ratio. There appears to be a pattern between those two factors and receiving a Hold recommendation.

Clusters 3-5 are top 3 in Revenue Growth and half the recommendations for each cluster are Moderate Buy. Pattern showing that pharm companies with higher revenue growth are more likely to receive buy recommendations.

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

Cluster 1: High Net Profit Margin Cluster 2: High Beta & Leverage; Low ROA & Market Cap Cluster 3: Very High P/E Ratio; Very low Net Profit Margin Cluster 4: High ROE,ROA,Asset Turnover; Low Leverage Cluster 5: High Revenue Growth; Very Low Asset Turnover