The University of Texas at Dallas

Naveen Jindal School of Management

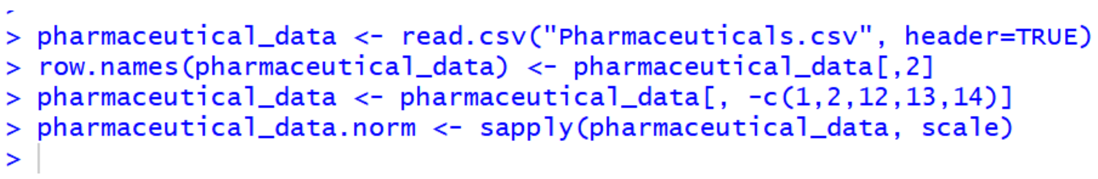
Pharmaceutical Industry Data Analysis

An equities analyst is studying the pharmaceutical industry and would like your help in exploring and understanding the financial data collected by her firm. Her main objective is to understand the structure of the pharmaceutical industry using some basic financial measures.

Financial data gathered on 21 firms in the pharmaceutical industry are available in the file Pharmaceuticals.xls.

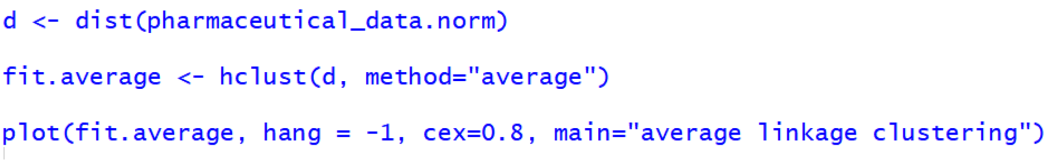
Name	Model Role	Data Type	Description
Symbol	Ignore	Categoric	Company stock symbol
Name	Ignore	Categoric	Company name
Market_Cap	Input	Numeric	Market capitalization (in billions of dollars)
Beta	Input	Numeric	Beta
PE_Ratio	Input	Numeric	Price to earnings ratio
ROE	Input	Numeric	Return on equity
ROA	Input	Numeric	Return on investment
Asset_Turnover	Input	Numeric	Asset turnover
Leverage	Input	Numeric	Leverage
Rev_Growth	Input	Numeric	Estimated revenue growth
Net_Profit	Input	Numeric	Net profit margin
Median_ Recommendation	Ignore	Categoric	Median recommendations (across major brokerages)
Location	Ignore	Categoric	Location of company headquarters
Exchange	Ignore	Categoric	Stock exchange on which the firm is listed

Load the Pharmaceuticals file in R. Assign the second column i.e. the pharmaceutical company to row names of the dataset. Remove all the text columns. This leaves us with 9 numeric columns. Normalize the column values.

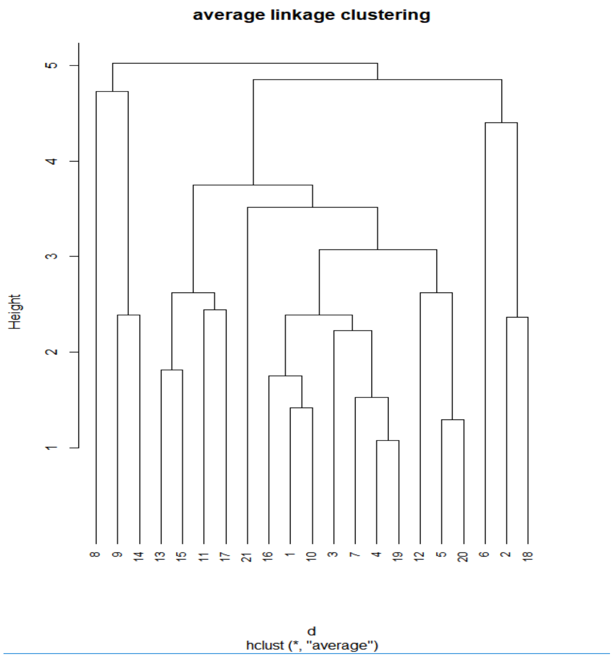


```
> pharmaceutical_data <- read.csv("Pharmaceuticals.csv", header=TRUE)
> row.names(pharmaceutical_data) <- pharmaceutical_data[,2]
> pharmaceutical_data <- pharmaceutical_data[, -c(1,2,12,13,14)]
> pharmaceutical_data.norm <- sapply(pharmaceutical_data, scale)
> 
```

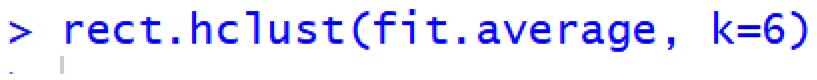
After normalizing the data, calculating the distance between the observations. Then using this distance to perform hierarchical clustering. After this I plotted the dendrogram.



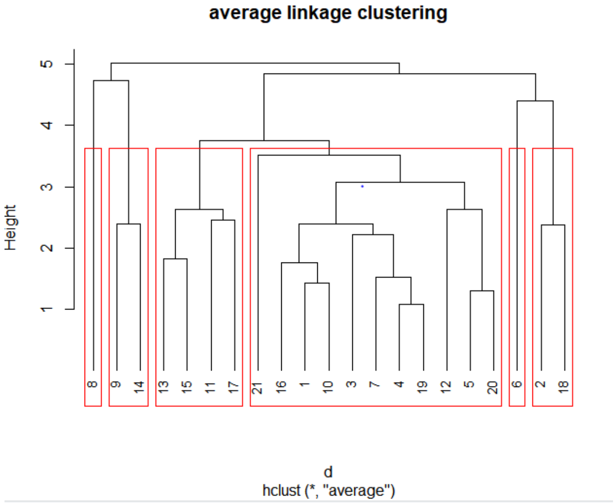
```
d <- dist(pharmaceutical_data.norm)

fit.average <- hclust(d, method="average")

plot(fit.average, hang = -1, cex=0.8, main="average linkage clustering")
```

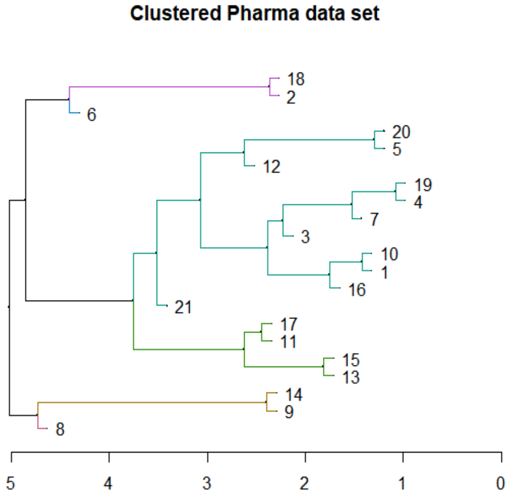


1.> I have taken the optimal number of clusters as per the dendrogram as 6. The height in a dendrogram represents the inverse of dissimilarity and we are calculating it using the average distance method. The more the height the lesser is the dissimilarity and vice- versa. We want to achieve maximum dissimilarity that is minimum height keeping the clusters meaningful. I decided to cut the tree for 6 clusters because the decrease in height from the hang for cluster 5 (Height= 4.4) to the hang for cluster 6 (Height= 3.7) is the maximum (the decrease is of 0.7). This means that the dissimilarity between the clusters increases a lot when we go from 5 clusters t0 6 clusters. So, I cut the tree at k=6 and for better visualization of the points in the six clusters I defined boundaries to distinguish between the hangs for different clusters and also color-coded the dendrogram for the six clusters.



```
> rect.hclust(fit.average, k=6)
```

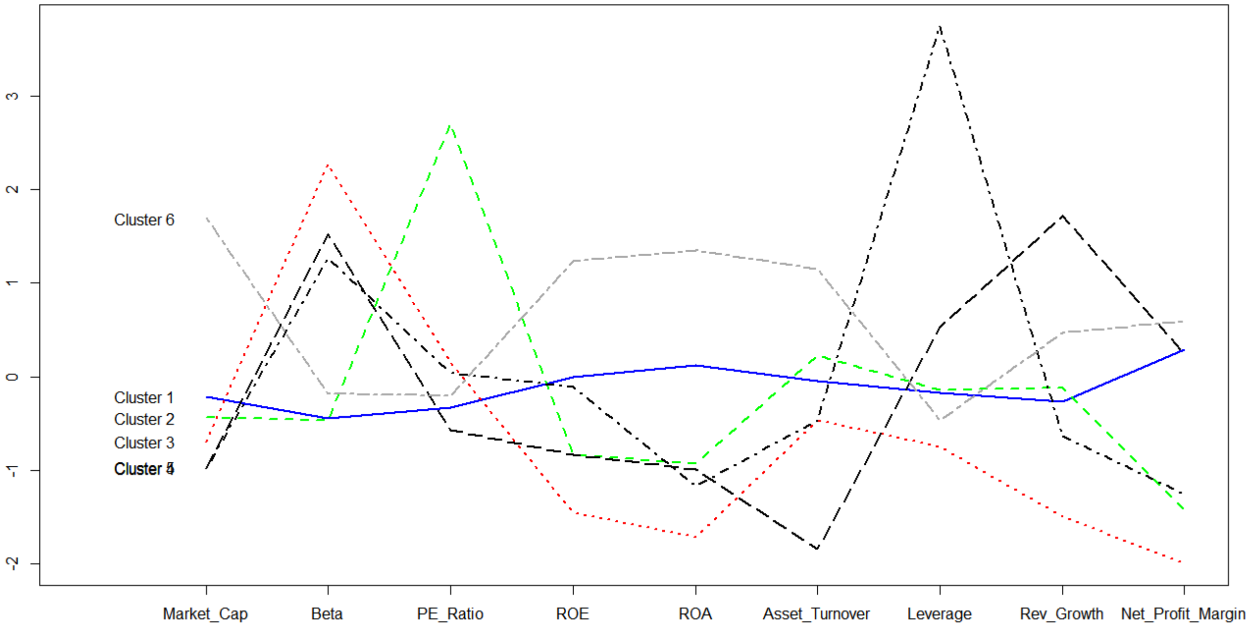




2.> For interpreting the clusters with respect to the variables used in their formation I need to create the profile plot of centroids. To do that I need to find the centroids first that I was not able to get from the hclust result. So, I created a function to get the centroids for each cluster. It uses the normalized data frame and the result of hclust (that gives us the row numbers for observations in a cluster) to calculate the cluster centroids.

After this, I create the profile plot of the cluster centroids.

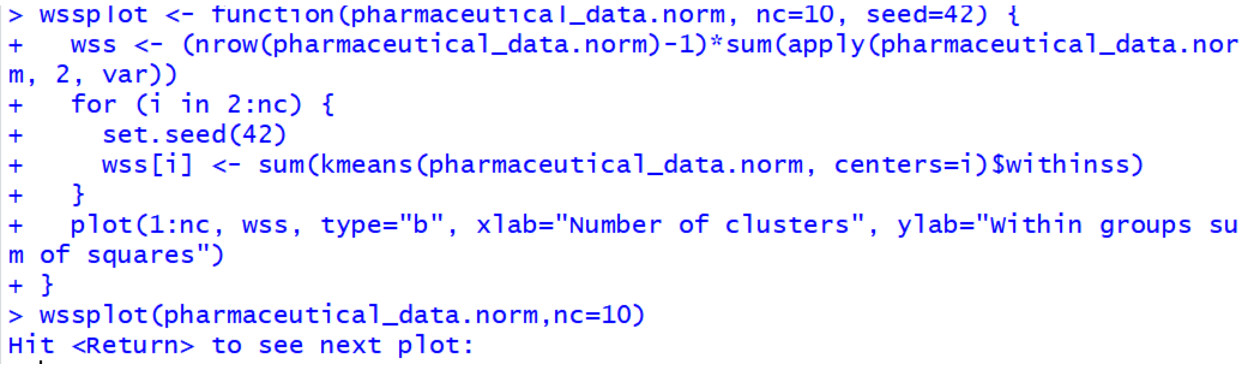
```
> plot(c(0), xaxt = 'n', ylab = "", type = "l", ylim = c(min(y), max(y)), xlim = c(0,9))
> # label x-axes
> axis(1, at = c(1:9), labels = names(pharmaceutical_data))
> # plot centroids
> for (i in c(1:6))
+   lines(y[i,],  lty = i, lwd = 2,
+         col = ifelse(i %in% c(1),"blue",
+                       (ifelse(i %in% c(2),"green",
+                               (ifelse(i %in% c(3),"red",
+                                       (ifelse(i %in% c(4,5),"black","dark grey"))))))))
>
> text(x = 0.5, y = pharma_centroids[, 1], labels = paste("Cluster", c(1:6)))
```



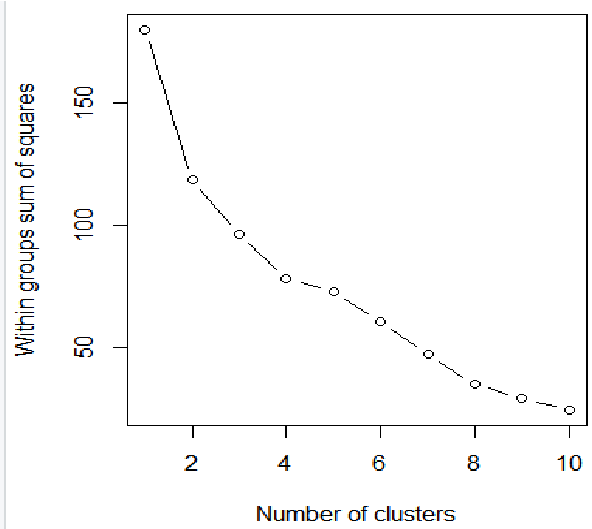
As you can see from the above profile plot:
Cluster 6 can be called as 'Big doing great' as it has the highest market cap with high Asset turnover, low Beta(risk) and a profit margin greater than all other clusters.
Cluster 4 can be called 'Recovering fast' as it has a low market cap, high Beta(risk), lowest asset turnover, good profit and highest revenue growth.
Cluster 5 can be called as 'Recovering slow' as it has a low market cap, high Beta, average asset turnover, below average revenue growth and below average net profit
Cluster 3 can be called as 'High risk no recovery' as it has the highest Beta, low market cap, low ROE, low ROA, least revenue growth and least net profit margin.
Cluster 1 can be called as 'Stable going good' as it has the lease Beta(risk), high ROE, high ROA, good revenue growth and high Net Profit Margin.
Cluster 2 can be called as 'Stable best buy' as it has the least Beta just like Cluster 1 , has an average asset turnover, average revenue growth and has the highest PE Ratio among all the clusters which is the factor used to select which stocks to buy. PE Ratio is the ratio of current Stock market price to the earning per share.

3.> The clusters formed seem reasonable to an extent. Cluster 5 could be a part of high risk low recovery except one or two factors.

4.> Using k-means next.

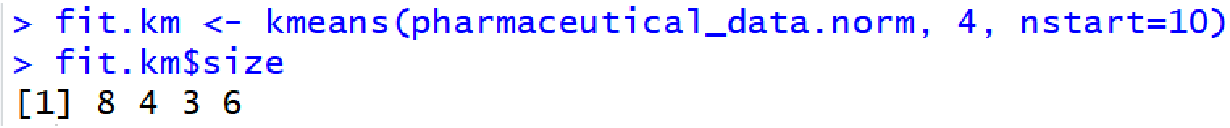


```
> wssplot <- function(pharmaceutical_data.norm, nc=10, seed=42) {
+   wss <- (nrow(pharmaceutical_data.norm)-1)*sum(apply(pharmaceutical_data.nor
m, 2, var))
+   for (i in 2:nc) {
+     set.seed(42)
+     wss[i] <- sum(kmeans(pharmaceutical_data.norm, centers=i)$withinss)
+   }
+   plot(1:nc, wss, type="b", xlab="Number of clusters", ylab="Within groups su
m of squares")
+ }
> wssplot(pharmaceutical_data.norm,nc=10)
Hit <Return> to see next plot:
```



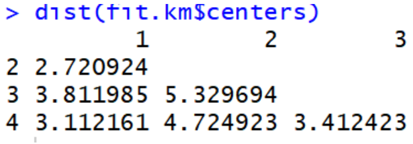
Looking at the within group sum of squares vs the number of clusters graph. The within group distances decrease less from # of clusters =4 to # of clusters =5.

Therefore, taking 4 as the number of clusters to perform k-means.



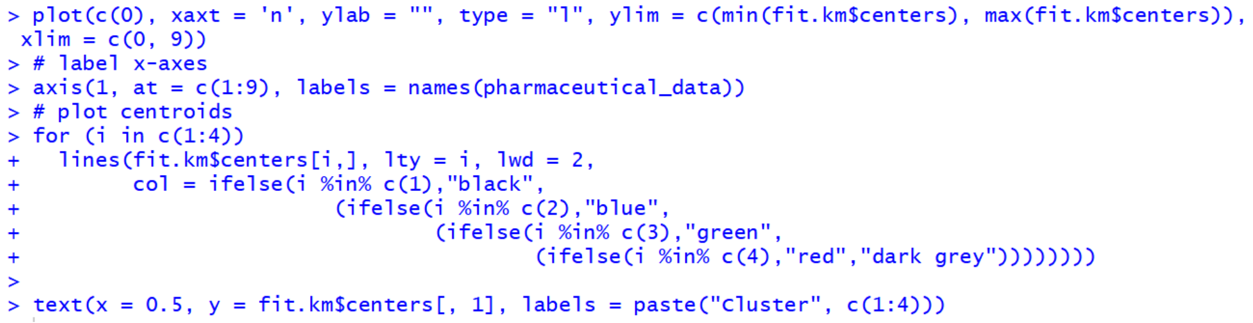
```
> fit.km <- kmeans(pharmaceutical_data.norm, 4, nstart=10)
> fit.km$size
[1] 8 4 3 6
```

Checking if any cluster is a striking outlier.

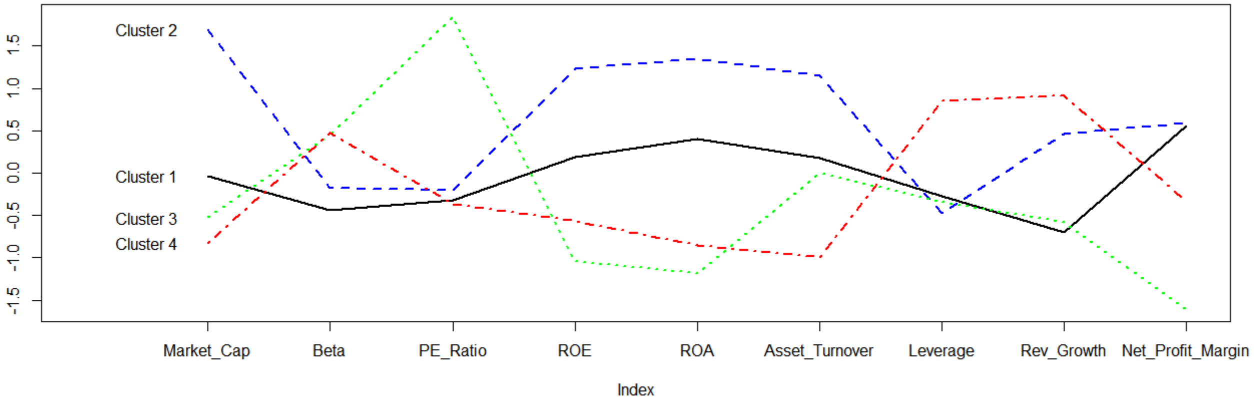


```
> dist(fit.km$centers)
         1        2        3
2 2.720924                  
3 3.811985 5.329694         
4 3.112161 4.724923 3.412423
```

Create the profile plot of centroids to check the characteristics of the clusters and whether they bring out meaning.



```
> plot(c(0), xaxt = 'n', ylab = "", type = "l", ylim = c(min(fit.km$centers), max(fit.km$centers)),
 xlim = c(0, 9))
> # label x-axes
> axis(1, at = c(1:9), labels = names(pharmaceutical_data))
> # plot centroids
> for (i in c(1:4))
+   lines(fit.km$centers[i,], lty = i, lwd = 2,
+         col = ifelse(i %in% c(1),"black",
+                      (ifelse(i %in% c(2),"blue",
+                              (ifelse(i %in% c(3),"green",
+                                      (ifelse(i %in% c(4),"red","dark grey")))))))
>
> text(x = 0.5, y = fit.km$centers[, 1], labels = paste("Cluster", c(1:4)))
```



Cluster 2 can be called as 'Big doing great' as it has the highest market cap with high Asset turnover, high ROA and ROE, low leverage, low Beta(risk) and a profit margin greater than all other clusters.
Cluster 4 can be called as 'Average risk high revenue' as it has highest Beta(risk), lowest market cap, highest leverage, lowest asset turnover but highest revenue.
Cluster 3 can be called as 'Average risk best buy' as it has risk as high as cluster 4 but the PE Ratio is the highest making it the one whose stock should ideally be bought.
Cluster 1 can be called as 'Stable going good' as it has the least Beta(risk), high ROE, high ROA and high net profit.

Although these clusters bring out some meaning, it is averaging on some factors making it hard to bring out some of the other things that we found in case of 6 clusters using hierarchical clustering. For example, Cluster 2 (Stable Best buy) in case of hierarchical clustering gave us a clear cluster having very low risk and very high PE Ratio giving us the best cluster from an investment standpoint.