Assignment 4

Sandeep More

3/16/2021

Assignment 4 - K-Means clustering

Use k-Means for clustering in exploring and understand the structure of the pharmaceutical industry using given financial measures.

Data prep

Load given data.

Calculate Distance

```
#1. euclidean
euclidean.dist <- dist(data.df.norm, method="euclidean")
print(euclidean.dist)
```

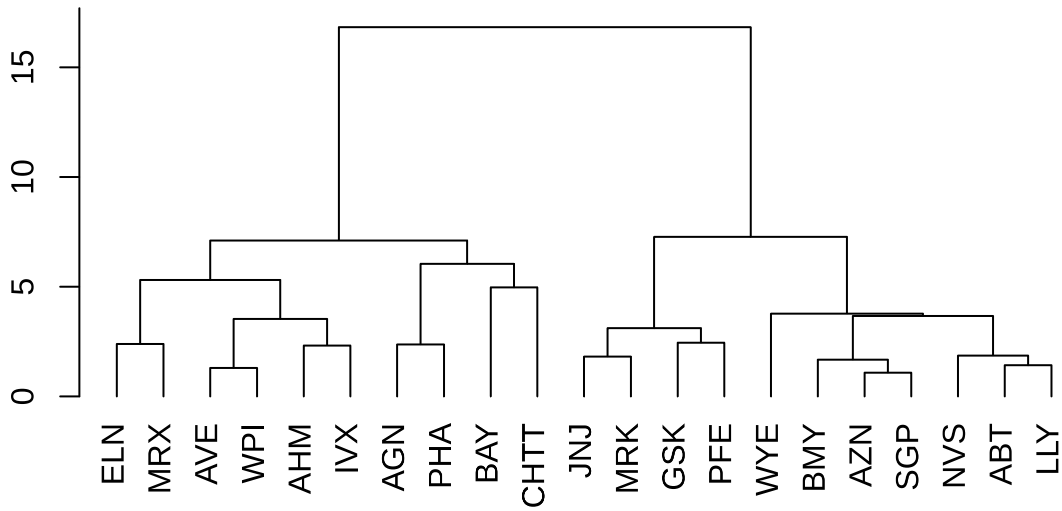
```
##           ABT      AGN      AHM      AZN      AVE      BAY      BMY     CHTT
## AGN  4.415575
## AHM  2.018793 3.945745
## AZN  1.669541 4.909566 2.364249
## AVE  2.111983 4.642699 2.487172 2.632282
## BAY  4.690231 4.853901 3.636353 5.065563 4.764654
## BMY  1.805543 5.419487 2.600986 1.572582 3.400602 5.273023
## CHTT 5.020726 5.612226 4.760341 5.719174 5.096246 4.969438 5.287400
## ELN  4.901141 6.695261 4.695844 4.974521 3.748778 4.608660 5.378092 4.675606
## LLY  1.422680 5.140253 3.238353 2.405951 2.910766 5.804419 2.189107 5.657801
## GSK  3.689906 6.747789 4.904614 2.957494 4.476690 7.546154 3.099023 7.080175
## IVX  2.624729 4.470028 2.316548 3.282195 2.386850 3.658011 3.279927 2.951511
## JNJ  2.333874 5.317942 3.593764 1.958326 3.640773 5.724303 2.511309 6.310233
## MRX  3.920297 5.479080 4.120549 4.269231 2.927258 4.848442 4.734766 4.786213
## MRK  2.680733 5.443918 3.361981 1.859280 3.472410 5.918477 2.432281 6.101541
## NVS  1.922731 5.468844 3.331743 3.056196 3.330879 5.331004 2.866126 6.063738
## PFE  3.887235 6.906828 5.268858 3.109413 4.495242 7.163993 3.666674 7.180257
## PHA  2.908982 2.367912 2.925627 3.715808 2.718441 3.955926 4.408645 5.000709
## SGP  1.312599 4.725384 1.704709 1.080519 2.464855 4.426418 1.478433 5.346513
## WPI  2.882610 5.007086 2.943946 3.414127 1.296549 5.055769 4.116074 5.540296
## WYE  3.038549 6.446458 4.185594 3.324966 4.254562 5.954379 2.269808 5.127981
##           ELN      LLY      GSK      IVX      JNJ      MRX      MRK      NVS
## AGN
## AHM
## AZN
## AVE
## BAY
```

```
## BMY
## CHTT
## ELN
## LLY  5.554227
## GSK  6.731204 3.631174
## IVX  3.115283 3.537378 5.276601
## JNJ  6.070533 2.722434 2.988672 4.354581
## MRX  2.389723 4.191466 6.187185 2.825394 5.306512
## MRK  5.921987 3.380695 2.218040 4.164267 1.814184 5.532520
## NVS  5.732322 1.577953 4.783039 3.899915 3.083678 4.478040 4.112418
## PFE  6.123133 3.783136 2.447177 5.356598 2.447341 5.518379 2.831329 4.536250
## PHA  5.007721 3.754900 5.773960 3.073579 4.112432 3.827019 4.448933 3.884035
## SGP  4.665611 2.205815 3.780283 2.763476 2.604437 3.907501 2.710607 2.542763
## WPI  3.756437 3.412378 5.437193 2.857109 4.591764 2.653341 4.569336 3.626404
## WYE  5.312455 2.747839 3.670720 3.719962 3.858028 4.709401 3.935039 3.525940
##           PFE      PHA      SGP      WPI
## AGN
## AHM
## AZN
## AVE
## BAY
## BMY
## CHTT
## ELN
## LLY
## GSK
## IVX
## JNJ
## MRX
## MRK
## NVS
## PFE
## PHA  5.587119
## SGP  3.955078 3.449579
## WPI  5.403128 3.172178 3.026610
## WYE  4.026095 5.286507 3.145472 4.922945
```

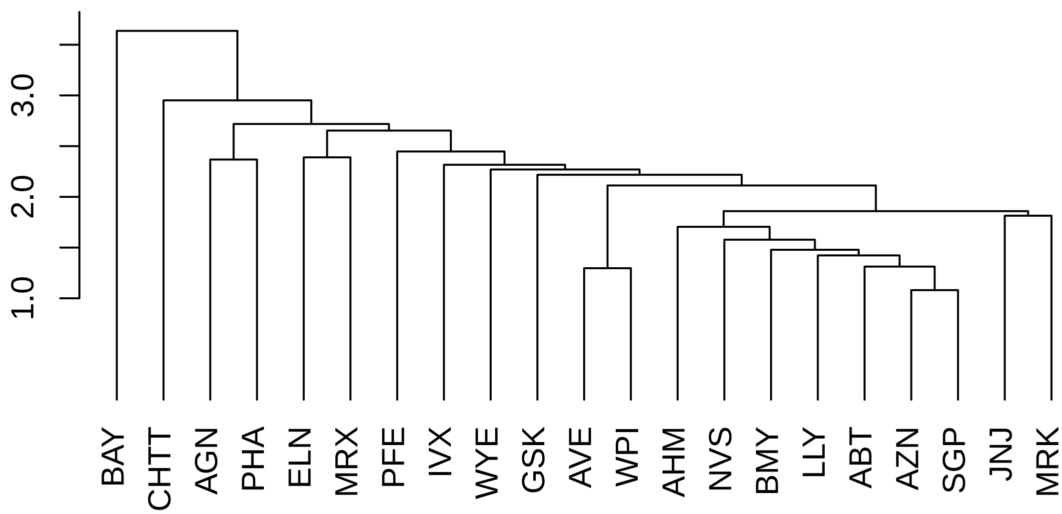
Clustering

Agglomerative Cluster

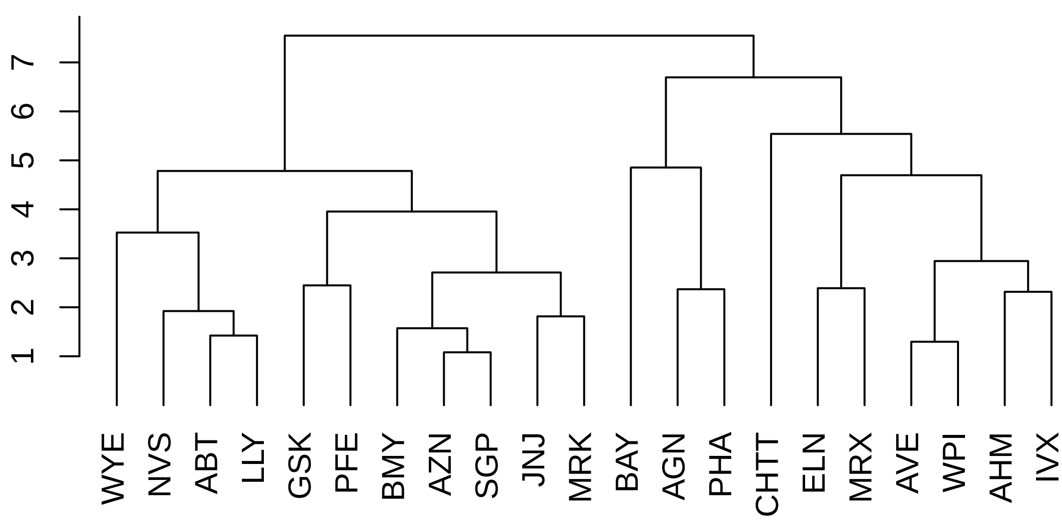
```r
#method could be "ward.D", "single", "complete", "average", "median", "centroid"
agglo.cluster.ward <- hclust(euclidean.dist, method = "ward.D")
plot(agglo.cluster.ward, hang = -1, ann=FALSE)
```



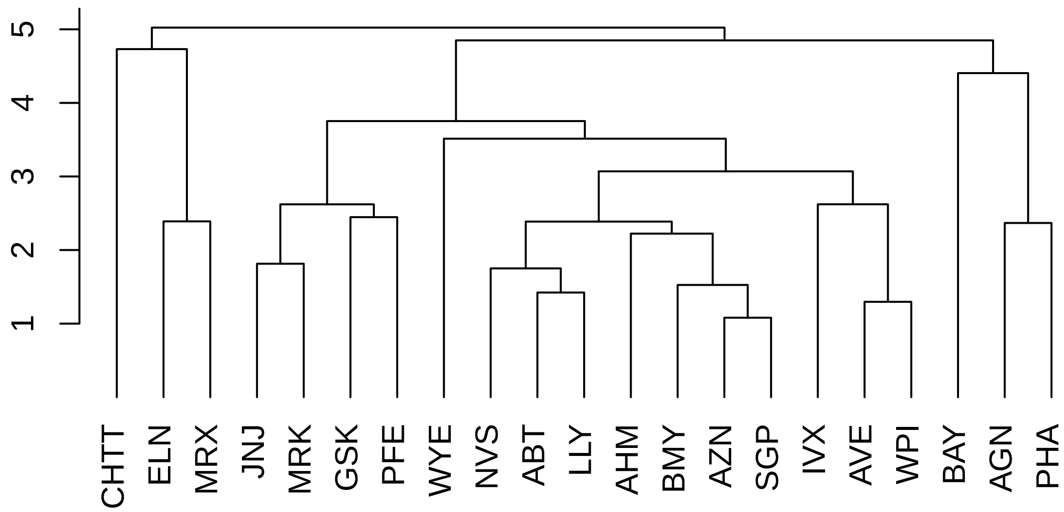

```r
agglo.cluster.single <- hclust(euclidean.dist, method = "single")
plot(agglo.cluster.single, hang = -1, ann=FALSE)
```



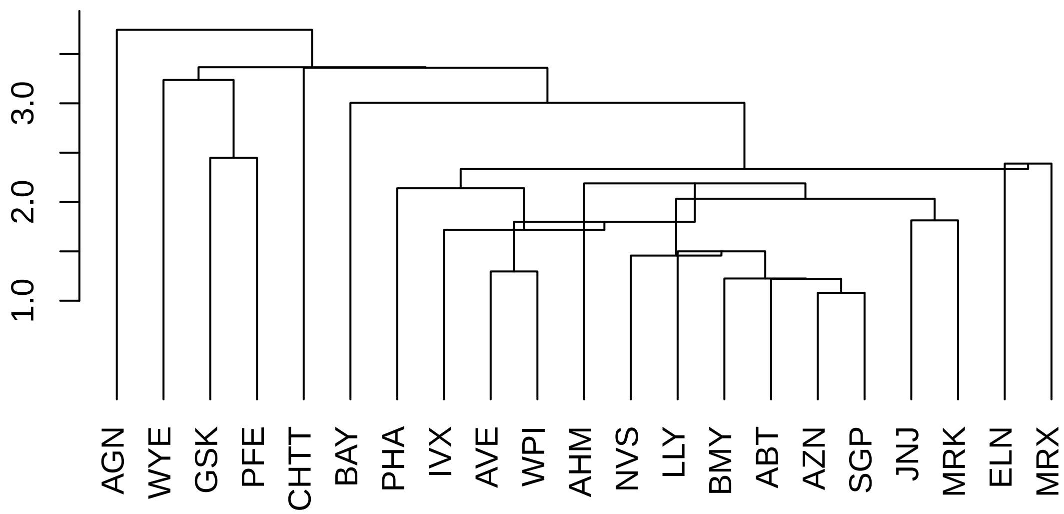

```r
agglo.cluster.complete <- hclust(euclidean.dist, method = "complete")
plot(agglo.cluster.complete, hang = -1, ann=FALSE)
```



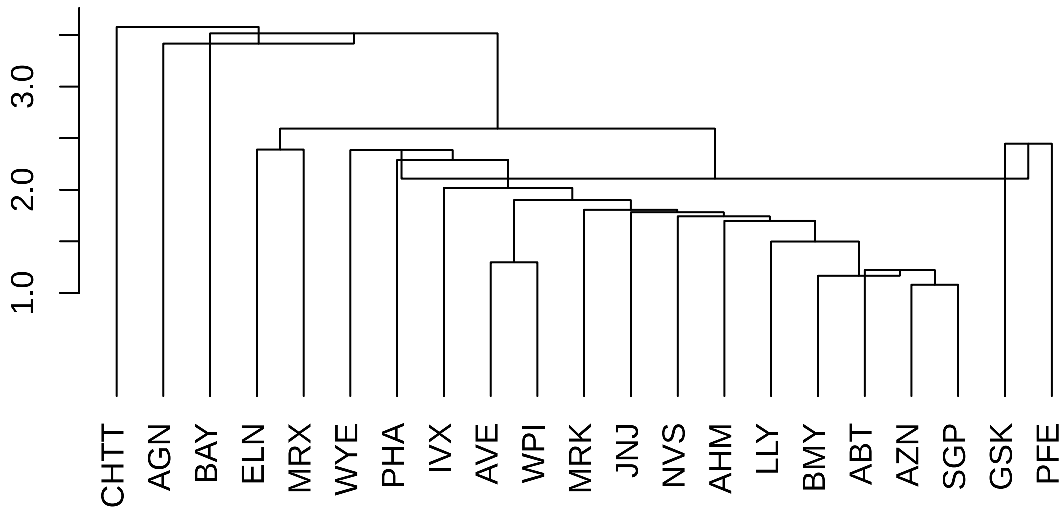
```r
agglo.cluster.average <- hclust(euclidean.dist, method = "average")
plot(agglo.cluster.average, hang = -1, ann=FALSE)
```




```r
agglo.cluster.median <- hclust(euclidean.dist, method = "median")
plot(agglo.cluster.median, hang = -1, ann=FALSE)
```




```r
agglo.cluster.centroid <- hclust(euclidean.dist, method = "centroid")
plot(agglo.cluster.centroid, hang = -1, ann=FALSE)
```



Experimenting with number of clusters

```
agglo.cluster.ward.cut <- cutree(agglo.cluster.ward, k = 5)
print(agglo.cluster.ward.cut)
```

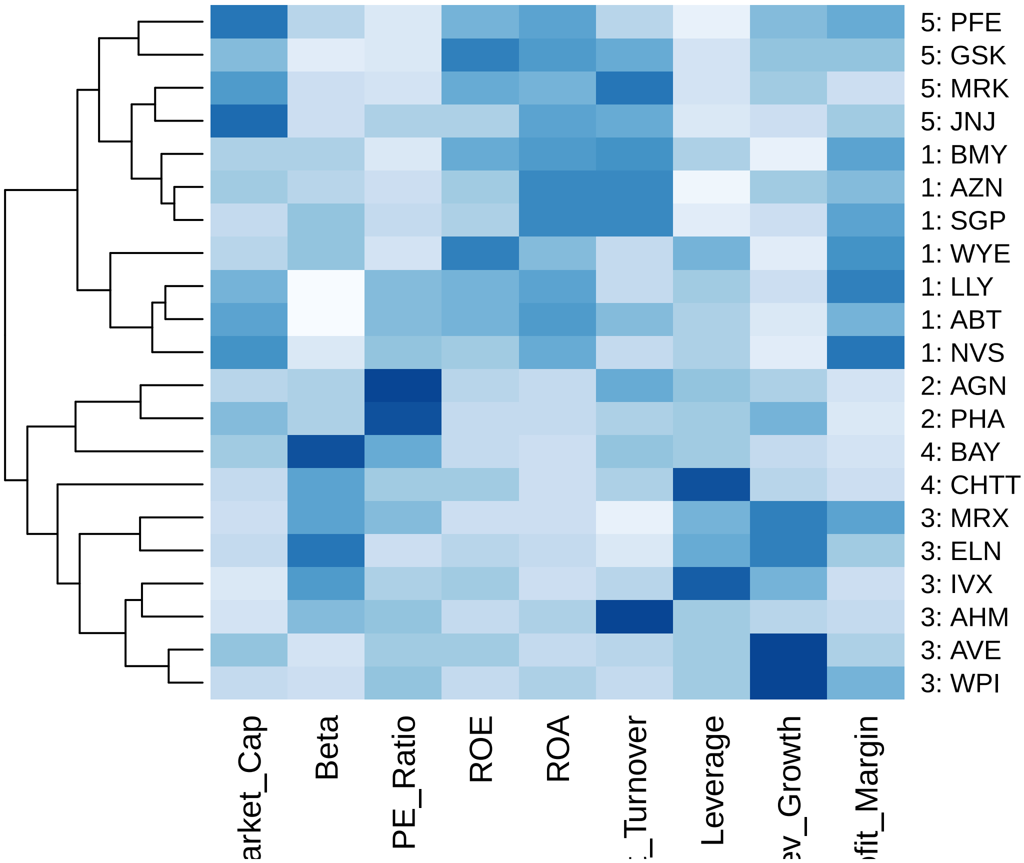
```
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    1    2    3    1    3    4    1    4    3    1    5    3    5    3    5    1
##  PFE  PHA  SGP  WPI  WYE
##    5    2    1    3    1
```

```
agglo.cluster.single.cut <- cutree(agglo.cluster.single, k = 2.5)
print(agglo.cluster.single.cut)
```

```
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    1    1    1    1    1    2    1    1    1    1    1    1    1    1    1    1
##  PFE  PHA  SGP  WPI  WYE
##    1    1    1    1    1
```

Heatmap

```
library(RColorBrewer)
# Make the labels as cluster membership (determined from cuttree) : row name
row.names(data.df.norm) <- paste(agglo.cluster.ward.cut, ": ", row.names(data.df), sep = "")

# plot
#color=rev(paste("gray", 1:99,sep = ""))
#color = terrain.colors(256)
color = colorRampPalette(brewer.pal(8, "Blues"))(25)
heatmap(as.matrix(data.df.norm), Colv = NA, hclustfun = hclust, col = color)
```

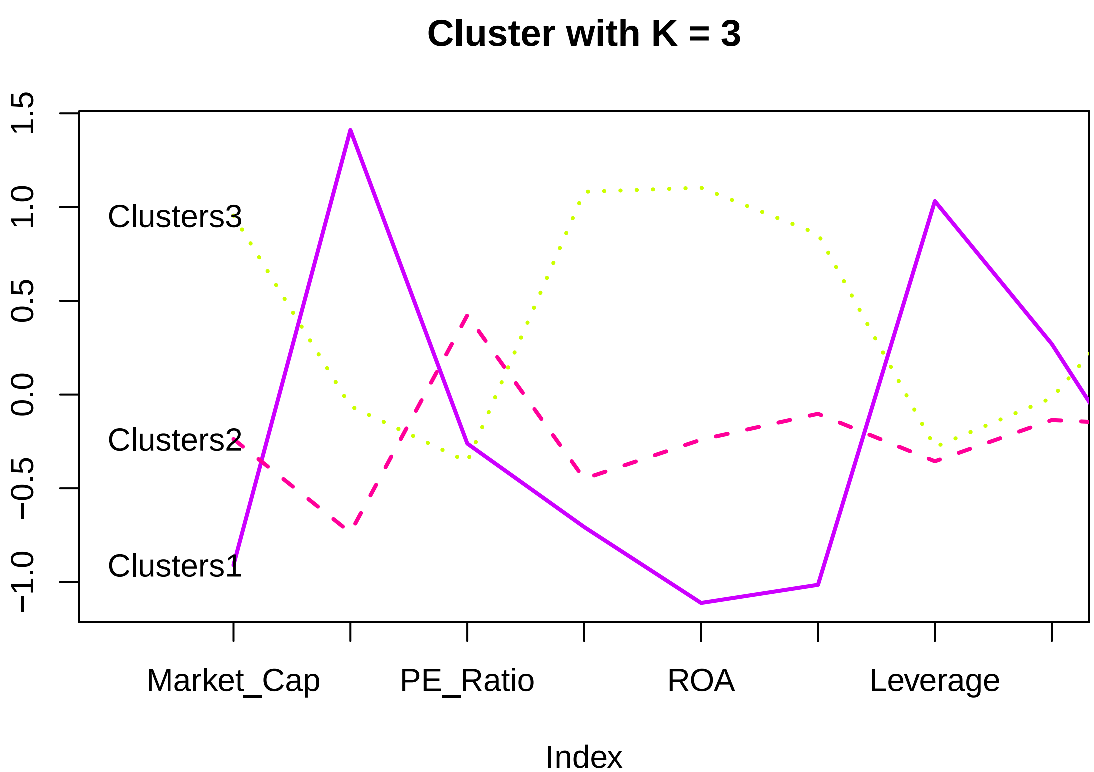


K-Means

```
#head(data.df.norm)
# run k-means
k <- 6

for(k in c(3,4,5,6,7)) {
    km <- kmeans(data.df.norm, k)
    print("-------------------------------------------------------------------------")
    # see cluster
    sprintf("K-means clusters with k = %s", k)
    print(km$cluster)
    # see centroids
    sprintf("K-means centroids for k = %s", k)
    print(km$centers)
    # see within cluster sum of squares
    sprintf("Within-cluster sum of squares for k = %s", k)
    print(km$withinss)

    # xaxt
    ## A character which specifies the x axis type. Specifying "n" suppresses plotting of the axis.
    ## The standard value is "s": for compatibility with S values "l" and "t" are     accepted  but are

    # type="l" is for lines
    plot(c(0), xaxt = 'n', ylab = "", type = "l", ylim = c(min(km$centers), max(km$centers)), xlim = c(
```

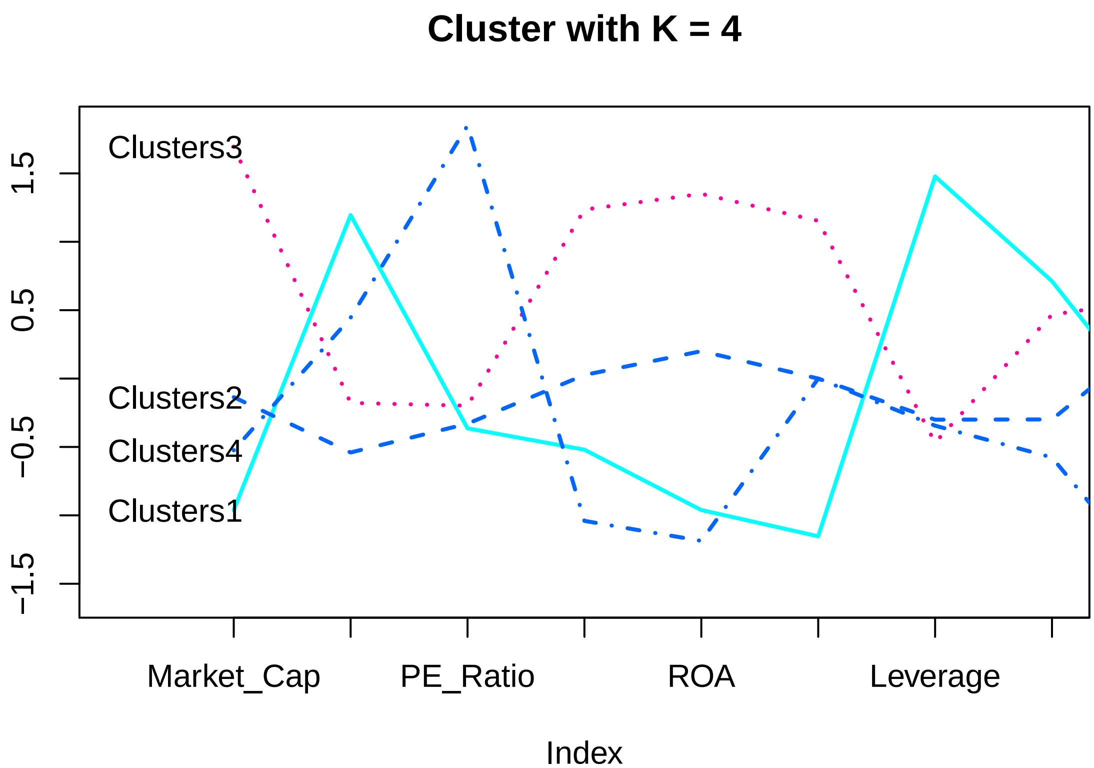
```
    #label x-axis
    axis(1, at = c(1:9), labels = colnames(data.df))

    # plot centroids
    for(i in 1:k)
      lines(km$centers[i,], lty = i, lwd = 2, col = sample(rainbow(10)))

    # name the clusters
    text(x = 0.5, y = km$centers[,1], labels = paste0("Clusters", c(1:k)))
}
```

```
## [1] "---------------------------------------------------------------------------"
##  1: ABT  2: AGN  3: AHM  1: AZN  3: AVE  4: BAY  1: BMY 4: CHTT  3: ELN  1: LLY
##       2       2       2       3       2       1       3       1       1       2
##  5: GSK  3: IVX  5: JNJ  3: MRX  5: MRK  1: NVS  5: PFE  2: PHA  1: SGP  3: WPI
##       3       1       3       1       3       2       3       2       2       2
##  1: WYE
##       3
##    Market_Cap        Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.9090570  1.41109654 -0.2613021 -0.7063477 -1.1114156     -1.0147843
## 2 -0.2375550 -0.73633718  0.4233386 -0.4489909 -0.2407172     -0.1025035
## 3  0.9547543 -0.06120687 -0.3576482  1.0818081  1.1033619      0.8566361
##     Leverage  Rev_Growth Net_Profit_Margin
## 1  1.0319661  0.27018076        -0.6941793
## 2 -0.3557313 -0.13595383        -0.1652117
## 3 -0.2797499 -0.01818848         0.7082574
## [1] 31.94053 42.25037 25.26414
```



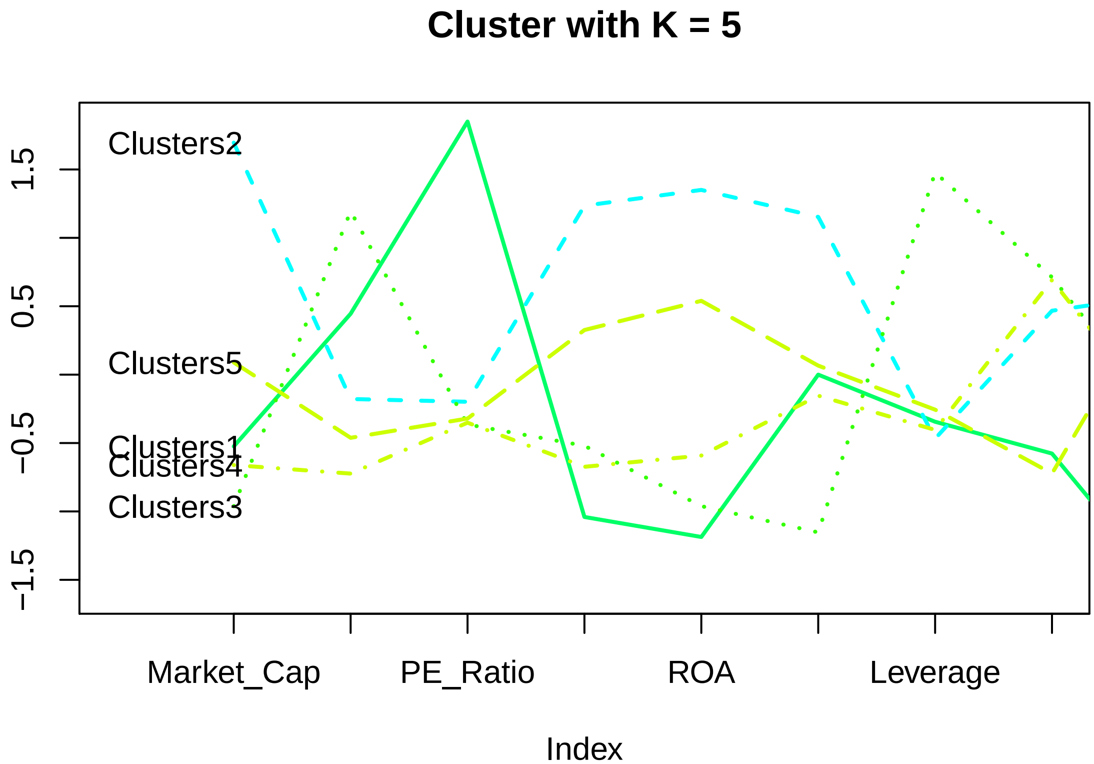

```
## [1] "---------------------------------------------------------------------------"
##  1: ABT  2: AGN  3: AHM  1: AZN  3: AVE  4: BAY  1: BMY 4: CHTT  3: ELN  1: LLY
```

```
##       2       4       2       2       2       4       2       1       1       2
##  5: GSK  3: IVX  5: JNJ  3: MRX  5: MRK  1: NVS  5: PFE  2: PHA  1: SGP  3: WPI
##       3       1       3       1       3       2       3       4       2       2
##  1: WYE
##       2
##    Market_Cap       Beta   PE_Ratio         ROE        ROA Asset_Turnover
## 1 -0.9624758  1.1949250 -0.3639982 -0.52006967 -0.9610792  -1.153164e+00
## 2 -0.1358537 -0.5402897 -0.3299706  0.02616921  0.2002696   1.443290e-16
## 3  1.6955811 -0.1780563 -0.1984582  1.23498791  1.3503431   1.153164e+00
## 4 -0.5246281  0.4451409  1.8498439 -1.04045502 -1.1865838   1.480297e-16
##     Leverage Rev_Growth Net_Profit_Margin
## 1  1.4773718  0.7120120        -0.3688236
## 2 -0.3004111 -0.2985927         0.3938956
## 3 -0.4680782  0.4671788         0.5912425
## 4 -0.3443544 -0.5769454        -1.6095439
## [1] 19.219788 35.336469  9.284424 14.938904
```



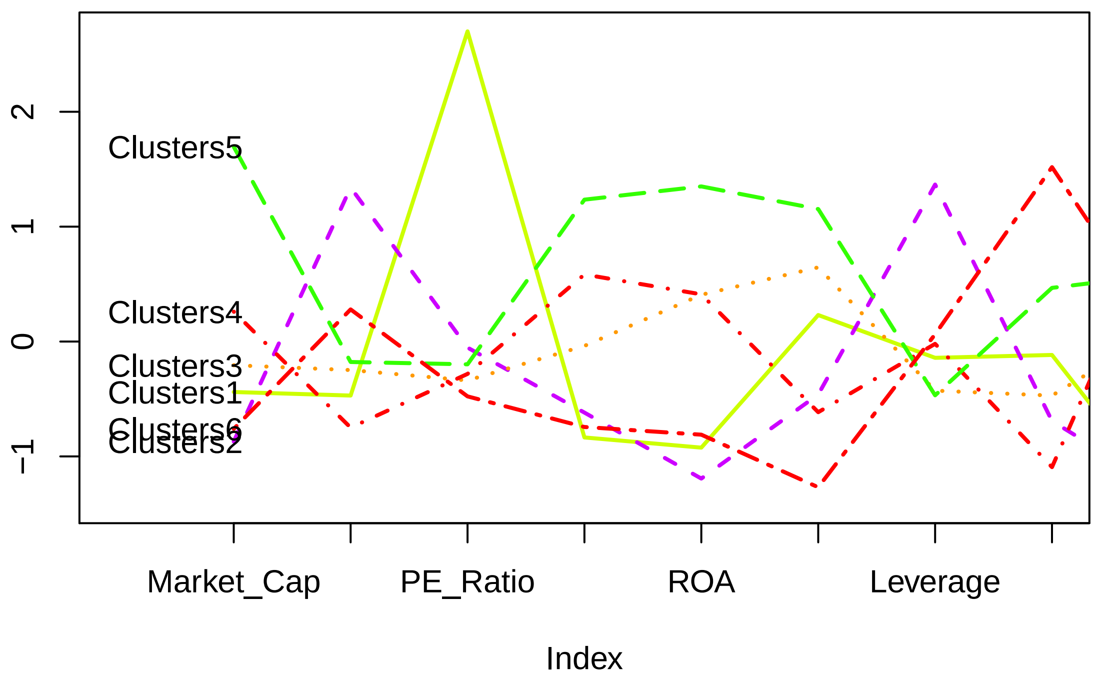

```
## [1] "------------------------------------------------------------------------"
##  1: ABT  2: AGN  3: AHM  1: AZN  3: AVE  4: BAY  1: BMY 4: CHTT  3: ELN  1: LLY
##       5       1       4       5       4       1       5       3       3       5
##  5: GSK  3: IVX  5: JNJ  3: MRX  5: MRK  1: NVS  5: PFE  2: PHA  1: SGP  3: WPI
##       2       3       2       3       2       5       2       1       5       4
##  1: WYE
##       5
##     Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.52462814  0.4451409  1.8498439 -1.0404550 -1.1865838   1.480297e-16
## 2  1.69558112 -0.1780563 -0.1984582  1.2349879  1.3503431   1.153164e+00
## 3 -0.96247577  1.1949250 -0.3639982 -0.5200697 -0.9610792  -1.153164e+00
## 4 -0.66114002 -0.7233539 -0.3512251 -0.6736441 -0.5915022  -1.537552e-01
## 5  0.08926902 -0.4618336 -0.3208615  0.3260892  0.5396003   6.589509e-02
```

```
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.3443544 -0.5769454        -1.6095439
## 2 -0.4680782  0.4671788         0.5912425
## 3  1.4773718  0.7120120        -0.3688236
## 4 -0.4040831  0.6917224        -0.4005718
## 5 -0.2559803 -0.7230135         0.7343816
## [1] 14.938904  9.284424 19.219788  5.511294 16.655937
```

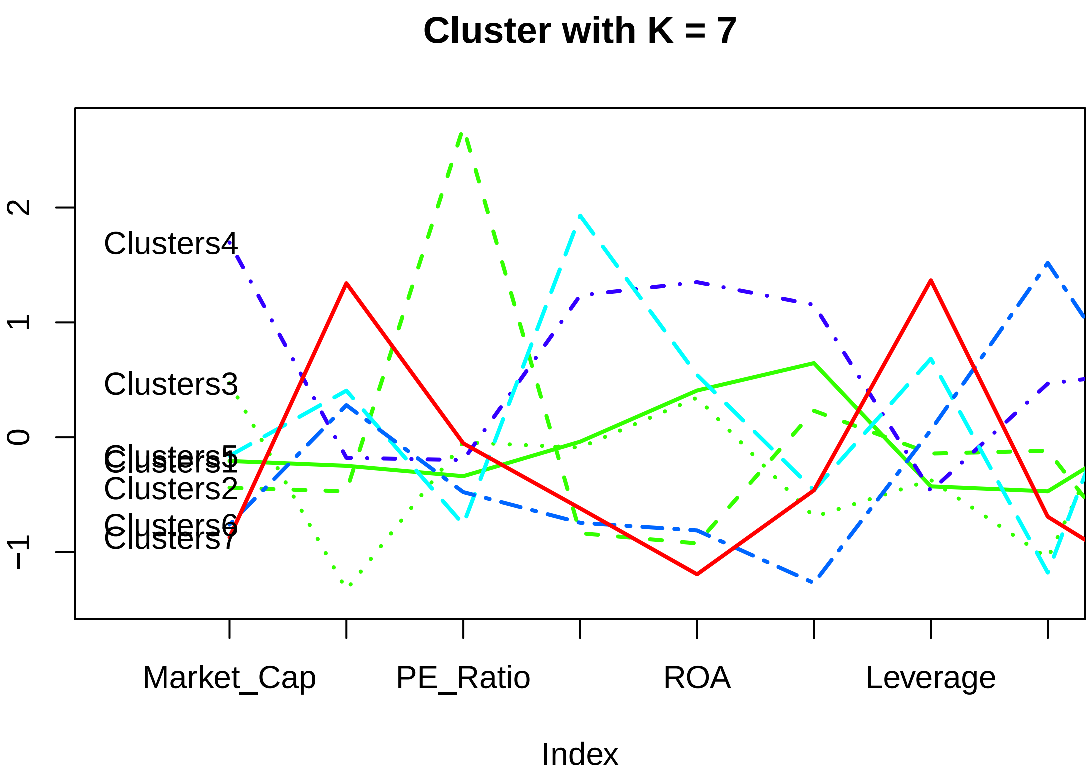



```
## [1] "----------------------------------------------------------------------"
##  1: ABT  2: AGN  3: AHM  1: AZN  3: AVE  4: BAY  1: BMY 4: CHTT  3: ELN  1: LLY
##       3       1       3       3       6       2       3       2       6       4
##  5: GSK  3: IVX  5: JNJ  3: MRX  5: MRK  1: NVS  5: PFE  2: PHA  1: SGP  3: WPI
##       5       2       5       6       5       4       5       1       3       6
##  1: WYE
##       4
##   Market_Cap       Beta   PE_Ratio         ROE        ROA Asset_Turnover
## 1 -0.4392513 -0.4701800  2.70002464 -0.83495252 -0.9234951      0.2306328
## 2 -0.8705151  1.3409869 -0.05284434 -0.61840151 -1.1928478     -0.4612656
## 3 -0.2063280 -0.2481660 -0.33855413 -0.03813318  0.4069821      0.6457718
## 4  0.2600876 -0.7493205 -0.28173916  0.58367759  0.4107405     -0.6150208
## 5  1.6955811 -0.1780563 -0.19845823  1.23498791  1.3503431      1.1531640
## 6 -0.7602249  0.2796041 -0.47742380 -0.74380222 -0.8107428     -1.2684804
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.14170336 -0.1168459      -1.416514761
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.42712134 -0.4707453       0.153117118
## 4 -0.02011273 -1.0931619       1.230016660
## 5 -0.46807818  0.4671788       0.591242521
## 6  0.06308085  1.5180158      -0.006893899
## [1]  2.803505 15.595925  6.586586  7.490937  9.284424 12.791257
```

Cluster with K = 6




```
## [1] "-----------------------------------------------------------------------"
##   1: ABT   2: AGN   3: AHM   1: AZN   3: AVE   4: BAY   1: BMY 4: CHTT   3: ELN   1: LLY
##        1        2        1        1        6        7        1        7        6        3
##   5: GSK   3: IVX   5: JNJ   3: MRX   5: MRK   1: NVS   5: PFE   2: PHA   1: SGP   3: WPI
##        4        7        4        6        4        3        4        2        1        6
##   1: WYE
##        5
##    Market_Cap       Beta    PE_Ratio         ROE        ROA Asset_Turnover
## 1 -0.2063280 -0.2481660 -0.33855413 -0.03813318  0.4069821      0.6457718
## 2 -0.4392513 -0.4701800  2.70002464 -0.83495252 -0.9234951      0.2306328
## 3  0.4708563 -1.3270762 -0.04364767 -0.08917735  0.3449684     -0.6918984
## 4  1.6955811 -0.1780563 -0.19845823  1.23498791  1.3503431      1.1531640
## 5 -0.1614497  0.4061910 -0.75792214  1.92938746  0.5422849     -0.4612656
## 6 -0.7602249  0.2796041 -0.47742380 -0.74380222 -0.8107428     -1.2684804
## 7 -0.8705151  1.3409869 -0.05284434 -0.61840151 -1.1928478     -0.4612656
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.42712134 -0.4707453       0.153117118
## 2 -0.14170336 -0.1168459      -1.416514761
## 3 -0.37208559 -1.0509233       1.097944074
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.68383297 -1.1776392       1.494161830
## 6  0.06308085  1.5180158      -0.006893899
## 7  1.36644699 -0.6912914      -1.320000179
## [1]  6.586586  2.803505  1.244968  9.284424  0.000000 12.791257 15.595925
```



Solutions

Following are solutions

Problem a, b

Statement

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

Answer

Here we attempted two different clustering types *Note*: Due to the stochastic nature of K-Means the cluster numbers might not align with the order the clusters are displayed. This does not affect the analysis but the naming (such as Cluster 1, Cluster 2) could be a bit misleading. Cluster are named properly at the end of the report.

- Agglomerative clustering
- K-means which are discussed below

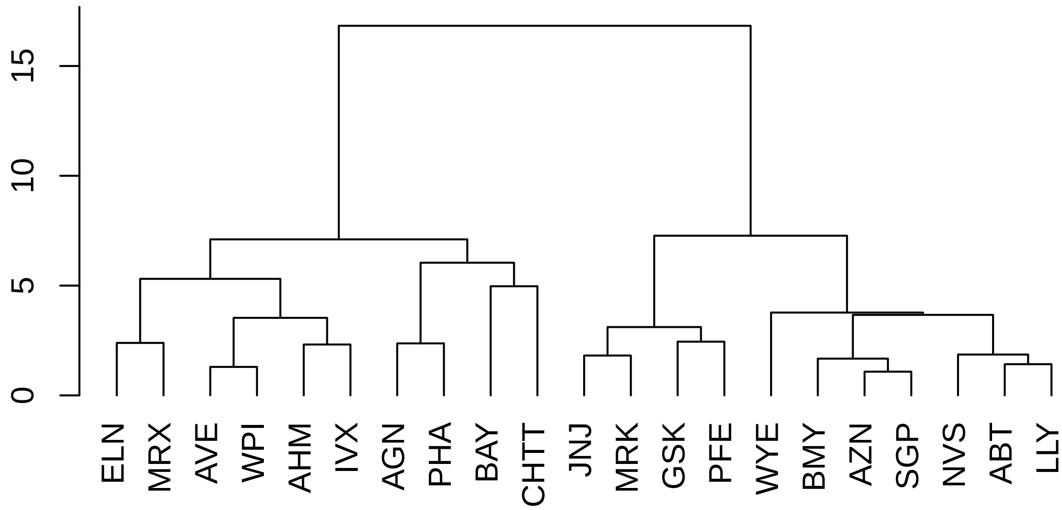
Agglomerative clustering Agglomerative clustering is examples are shown in the first half of the report where we use different distance measures to cluster. The distance measures used were

- Single Linkage
- Complete Linkage
- Average Linkage
- Centroid Linkage

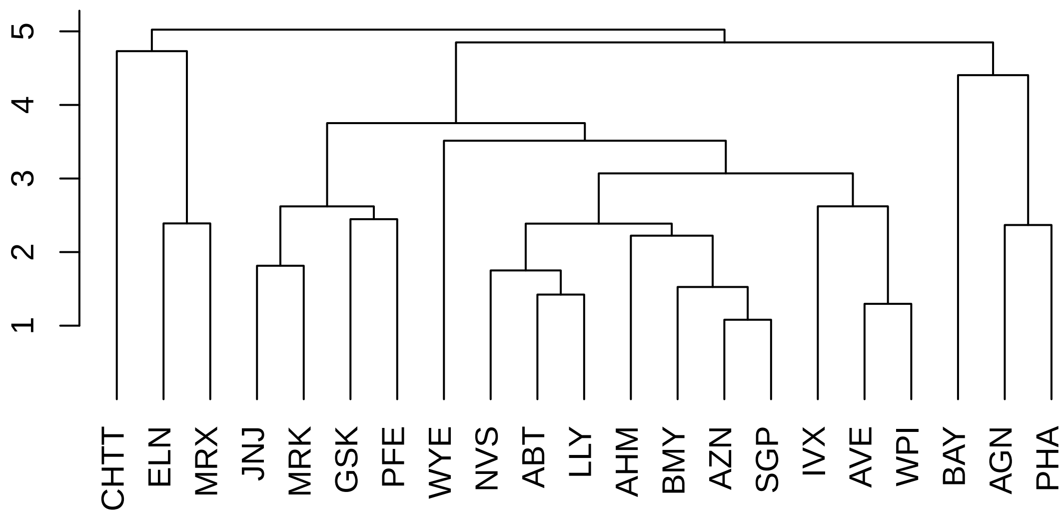
- Ward's method

Using the above methods we get clusters of various sizes ranging from 3 - 5. The cluster generated by Ward's and Average look promising

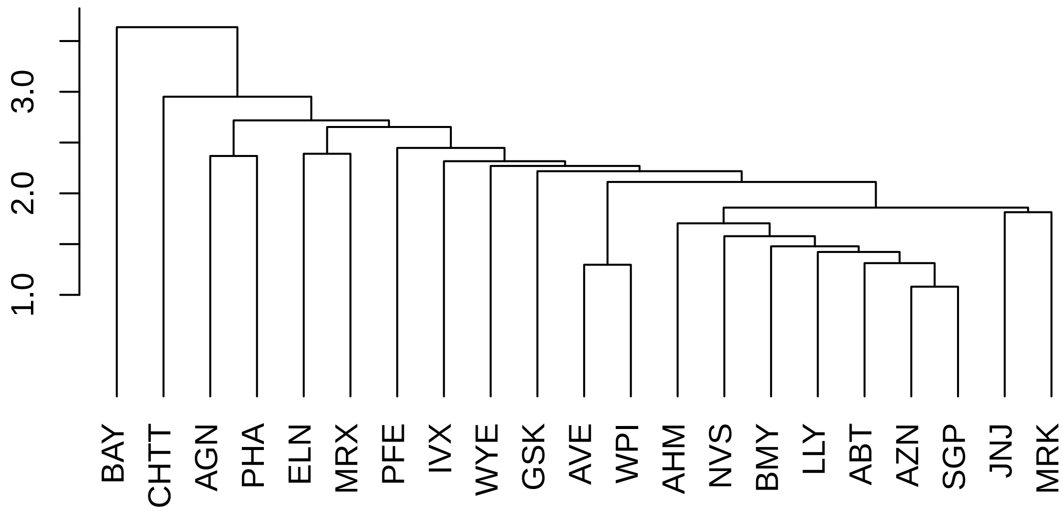
```
plot(agglo.cluster.ward, hang = -1, ann=FALSE)
```




```
plot(agglo.cluster.average, hang = -1, ann=FALSE)
```




```
plot(agglo.cluster.single, hang = -1, ann=FALSE)
```



Ward's method is better suited because it accounts for loss of information during clustering. Using Ward's method we see clusters size of 4 with with the utilities clustered as

- {ELN, MRX, AVE, WPI, AHM, IVX}
- {AGN, PHA, BAY, CHTT}
- {JNJ, MRK, GSK, PFE}
- {WYE, BMY, AZN, SGP, NVS, ABT, LLY}

Ward's method is more spread out with meaningful clusters unline others (single, average linkage) We can see, looking at the categorical variables not used in clustering (Median Recommendation, Geography, Exchange) that the clusters are roughly clustered around Median Recommendation + Location

```
print(agglo.cluster.ward.cut)
```

```
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    1    2    3    1    3    4    1    4    3    1    5    3    5    3    5    1
##  PFE  PHA  SGP  WPI  WYE
##    5    2    1    3    1
```

```
agglo.cluster.single.cut <- cutree(agglo.cluster.single, k = 4)
print(agglo.cluster.single.cut)
```

```
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    1    2    1    1    1    3    1    4    1    1    1    1    1    1    1    1
##  PFE  PHA  SGP  WPI  WYE
##    1    2    1    1    1
```

```
agglo.cluster.average.cut <- cutree(agglo.cluster.average, k = 4)
print(agglo.cluster.average.cut)
```

```
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    1    2    1    1    1    2    1    3    4    1    1    1    1    4    1    1
##  PFE  PHA  SGP  WPI  WYE
##    1    2    1    1    1
```

```
cluster.1 <- data[data$Symbol %in% c("ELN","MRX", "AVE", "WPI", "AHM", "IVX"), ]
cluster.1
```

```
## # A tibble: 6 x 14
##   Symbol Name  Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##   <chr>  <chr>      <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
## 1 AHM    Amer~        6.3  0.46     20.7  14.9   7.8            0.9     0.27
```

```
## 2 AVE    Aven~      47.2   0.32     20.1  21.8   7.5            0.6     0.34
## 3 ELN    Elan~       0.78  1.08      3.6  15.1   5.1            0.3     1.07
## 4 IVX    IVAX~       2.6   0.65     19.9  21.4   6.8            0.6     1.45
## 5 MRX    Medi~       1.2   0.75     28.6  11.2   5.4            0.3     0.93
## 6 WPI    Wats~       3.26  0.24     18.4  10.2   6.8            0.5     0.2
## # ... with 5 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>,
## #   Median_Recommendation <chr>, Location <chr>, Exchange <chr>
```

```r
cluster.2 <- data[data$Symbol %in% c("AGN","PHA", "BAY", "CHTT"), ]
cluster.2
```

```
## # A tibble: 4 x 14
##   Symbol Name  Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##   <chr>  <chr>      <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
## 1 AGN    Alle~       7.58  0.41     82.5  12.9   5.5            0.9     0.6
## 2 BAY    Baye~      16.9   1.11     27.9   3.9   1.4            0.6     0
## 3 CHTT   Chat~       0.41  0.85     26    24.1   4.3            0.6     3.51
## 4 PHA    Phar~      56.2   0.4      56.5  13.5   5.7            0.6     0.35
## # ... with 5 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>,
## #   Median_Recommendation <chr>, Location <chr>, Exchange <chr>
```

```r
cluster.3 <- data[data$Symbol %in% c("JNJ","MRK", "GSK", "PFE"), ]
cluster.3
```

```
## # A tibble: 4 x 14
##   Symbol Name  Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##   <chr>  <chr>      <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
## 1 GSK    Glax~       122.  0.35     18    62.9  20.3            1        0.34
## 2 JNJ    John~       174.  0.46     28.4  28.6  16.3            0.9      0.1
## 3 MRK    Merc~       133.  0.46     18.9  40.6  15              1.1      0.28
## 4 PFE    Pfiz~       199.  0.65     23.6  45.6  19.2            0.8      0.16
## # ... with 5 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>,
## #   Median_Recommendation <chr>, Location <chr>, Exchange <chr>
```

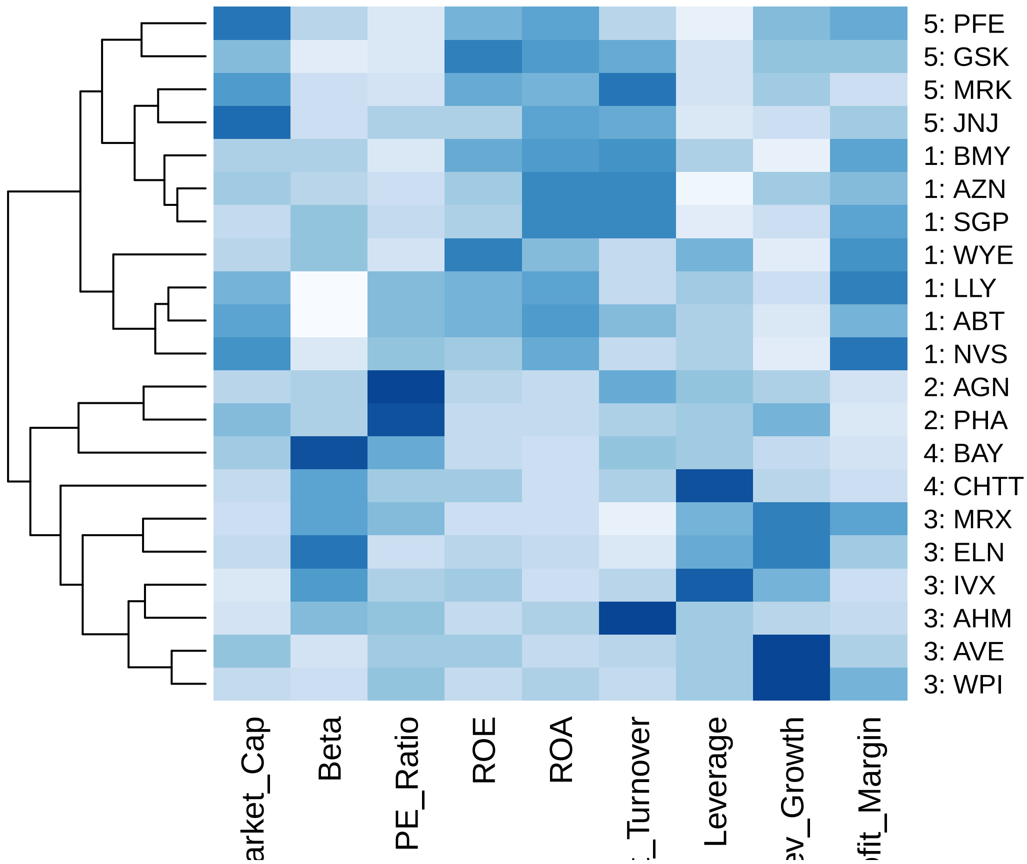
```r
cluster.4 <- data[data$Symbol %in% c("WYE","BMY", "AZN", "SGP", "NVS", "ABT", "LLY"), ]
cluster.4
```

```
## # A tibble: 7 x 14
##   Symbol Name  Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##   <chr>  <chr>      <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
## 1 ABT    Abbo~       68.4  0.32     24.7  26.4  11.8            0.7     0.42
## 2 AZN    Astr~       67.6  0.52     21.5  27.4  15.4            0.9     0
## 3 BMY    Bris~       51.3  0.5      13.9  34.8  15.1            0.9     0.570
## 4 LLY    Eli ~       73.8  0.18     27.9  31    13.5            0.6     0.53
## 5 NVS    Nova~       96.6  0.19     21.6  17.9  11.2            0.5     0.06
## 6 SGP    Sche~       34.1  0.51     18.9  22.6  13.3            0.8     0
## 7 WYE    Wyeth       48.2  0.63     13.1  54.9  13.4            0.6     1.12
## # ... with 5 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>,
## #   Median_Recommendation <chr>, Location <chr>, Exchange <chr>
```

The following heatmap of summary statistics shows 4 clusters is also lines up well with Ward's method and strengthens our choice of using Ward's method. We can see

- Cluster 1 is characterized by high net profit margin and cluster 2 lack thereof.
- Cluster 3 is characterized by high revenue growth
- Cluster 4 is characterized by high market cap
- We see some corelation between ROE and ROA

```
library(RColorBrewer)
# Make the labels as cluster membership (determined from cuttree) : row name
row.names(data.df.norm) <- paste(agglo.cluster.ward.cut, ": ", row.names(data.df), sep = "")

# plot
#color=rev(paste("gray", 1:99,sep = ""))
#color = terrain.colors(256)
color = colorRampPalette(brewer.pal(8, "Blues"))(25)
heatmap(as.matrix(data.df.norm), Colv = NA, hclustfun = hclust, col = color)
```



K-Means clustering

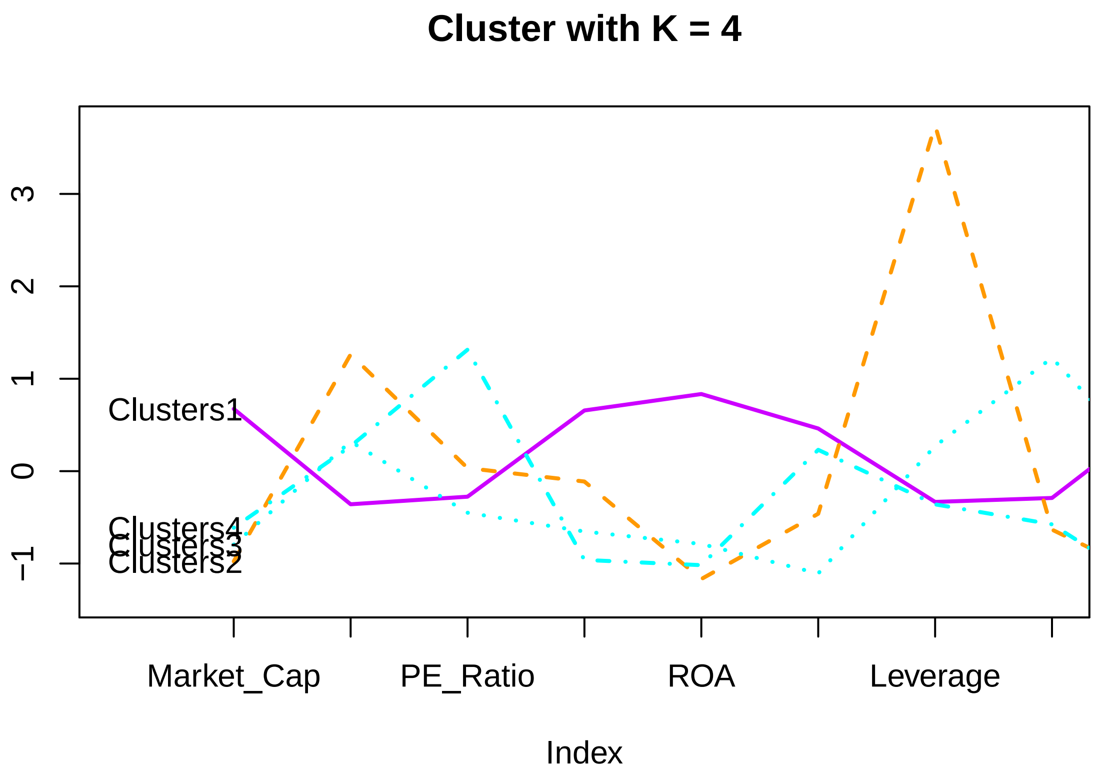
```
k = 4
# see cluster
km <- kmeans(data.df.norm, k)
```

Comparing it with K-Means cluster we can see that K=4 is the most optional value (considering different K values earlier in the report), going by the profile plot we can make some observations

- Cluster 1 is characterized by low Beta and PE ratio and high ROE, ROA and Asset Turnover
- Cluster 2 is characterized by high Beta and low ROE unlike Cluster 1
- Cluster 3 is characterized by high PE Ratio and low ROE and ROA
- Cluster 4 is characterized by average values for all variables

```
plot(c(0), xaxt = 'n', ylab = "", type = "l", ylim = c(min(km$centers), max(km$centers)), xlim = c(0, 8

    #label x-axis
    axis(1, at = c(1:9), labels = colnames(data.df))
```

```r
# plot centroids
for(i in 1:k)
  lines(km$centers[i,], lty = i, lwd = 2, col = sample(rainbow(10)))

# name the clusters
text(x = 0.5, y = km$centers[,1], labels = paste0("Clusters", c(1:k)))
```



We can see that Cluster 1 has lowest (9.2) within cluster dispersion and Cluster 2 has the highest (31.9) - Cluater labels may vary based on document generation

```r
sprintf("Within-cluster sum of squares for k = %s", k)
```

```
## [1] "Within-cluster sum of squares for k = 4"
```

```r
print(km$withinss)
```

```
## [1] 43.30886  0.00000 16.54260 20.54199
```

looking at the distances between clusters measured, we don't see any obvious outliers. We see that Cluster 1 and Cluster 3 are closely related and Cluster 1 and Cluster 2 are most distant.

```r
k = 4
# see cluster
km <- kmeans(data.df.norm, k)
sprintf("K-means clusters with k = %s", k)
```

```
## [1] "K-means clusters with k = 4"
```

```r
print(km$cluster)
```

```
##  1: ABT  2: AGN  3: AHM  1: AZN  3: AVE  4: BAY  1: BMY 4: CHTT  3: ELN  1: LLY
##       4        2        4        4        3        2        4        3        3        4
##  5: GSK  3: IVX  5: JNJ  3: MRX  5: MRK  1: NVS  5: PFE  2: PHA  1: SGP  3: WPI
```

```
##       1       3       1       3       1       4       1       2       4       3
##  1: WYE
##       4
```

```r
# see centroids
sprintf("K-means centroids for k = %s", k)
```

```
## [1] "K-means centroids for k = 4"
```

```r
print(km$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 -0.52462814  0.4451409  1.8498439 -1.0404550 -1.1865838   1.480297e-16
## 3 -0.82617719  0.4775991 -0.3696184 -0.5631589 -0.8514589  -9.994088e-01
## 4 -0.03142211 -0.4360989 -0.3172485  0.1950459  0.4083915   1.729746e-01
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.4680782  0.4671788         0.5912425
## 2 -0.3443544 -0.5769454        -1.6095439
## 3  0.8502201  0.9158889        -0.3319956
## 4 -0.2744931 -0.7041516         0.5569544
```

Summary

Overall summary, looking at both the clustering mechanisms, we prefer to use Hierarchical Clustering with Ward's method because of better cluster splits. Clusters formed using this Ward's method and K-Means with cutoff = 4 and K = 4 are very similar in composition. This observation strengthens our choice of using K = 4. Net Profit margin, ROE and ROA are good indiactors of cluster splits.

Problem c

c. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters) Solution: Looking at the categorical variables not used in clustering (Median Recommendation, Geography, Exchange) that the clusters are roughly clustered around Median Recommendation + Location. There is no strong correlation though. The clusters were mostly split on Net Profit margin, ROE / ROA, Beta variables.

Problem d

d. Provide an appropriate name for each cluster using any or all of the variables in the dataset. Solution: From our observations we can names the clusters as follows:

High growth cluster

- ELN - Elan Corporation, plc
- MRX - Medicis Pharmaceutical Corporation
- AVE - Aventis
- WPI - Watson Pharmaceuticals, Inc.
- AHM - Amersham plc
- IVX - IVAX Corporation

Low Profit & High PE ratio

- AGN - Allergan, Inc.
- PHA - Pharmacia Corporation
- BAY - Bayer AG
- CHTT - Chattem, Inc

High market cap

- JNJ - Johnson & Johnson
- MRK - Merck & Co., Inc.
- GSK - GlaxoSmithKline plc
- PFE - Pfizer Inc

High profit margin

- WYE - Wyeth
- BMY - Bristol-Myers Squibb Company
- AZN - AstraZeneca PLC
- SGP - Schering-Plough Corporation
- NVS - Novartis AG
- ABT - Abbott Laboratories
- LLY - Eli Lilly and Company