Assignment_4

wliu16

2022-10-31

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.

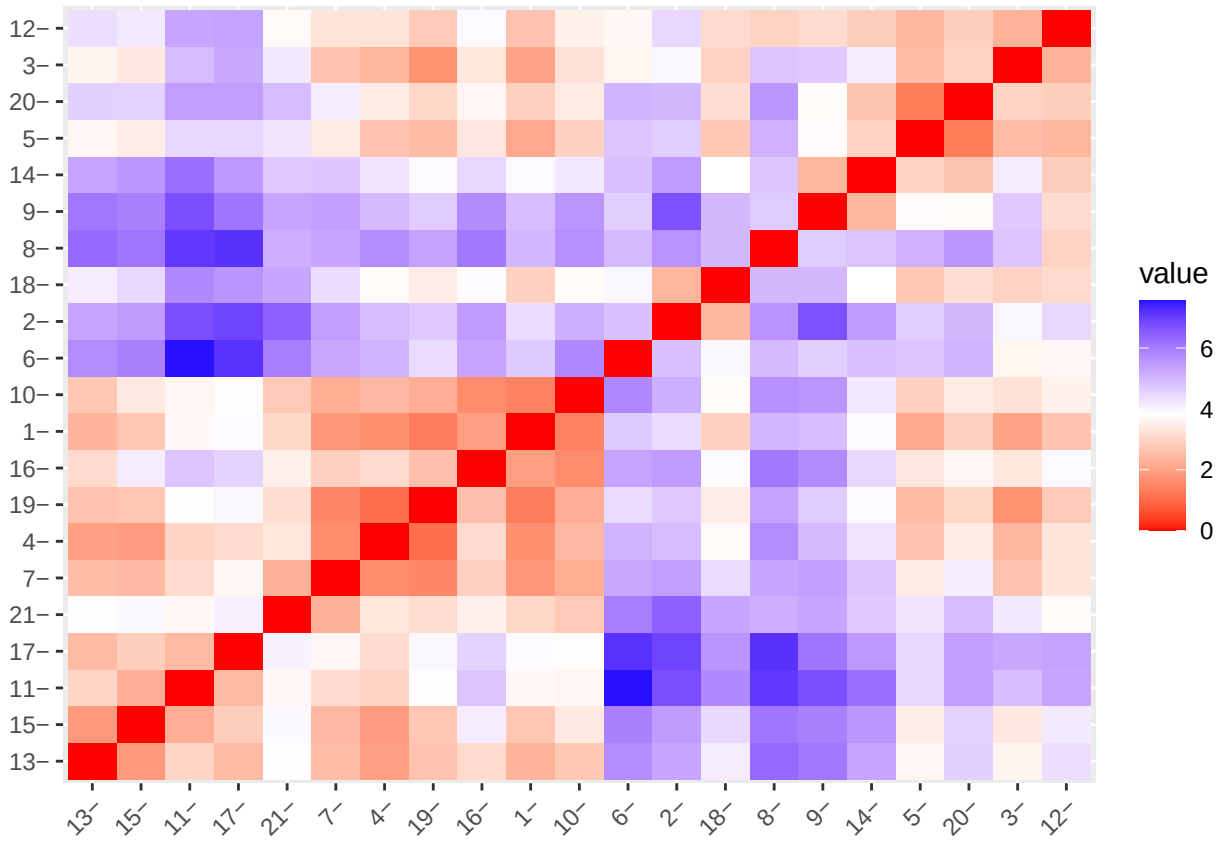
In the pharma problem, we use k-means algorithm to cluster the 21 firms into 5 clusters with no varying weights. We choose k=5 because it is the optimal k suggested by the silhouette method.

```
set.seed(123)
#scaling the dataframe (z-score)
ph_scaled <- scale(pharma[,3:11])
summary(ph_scaled)
```

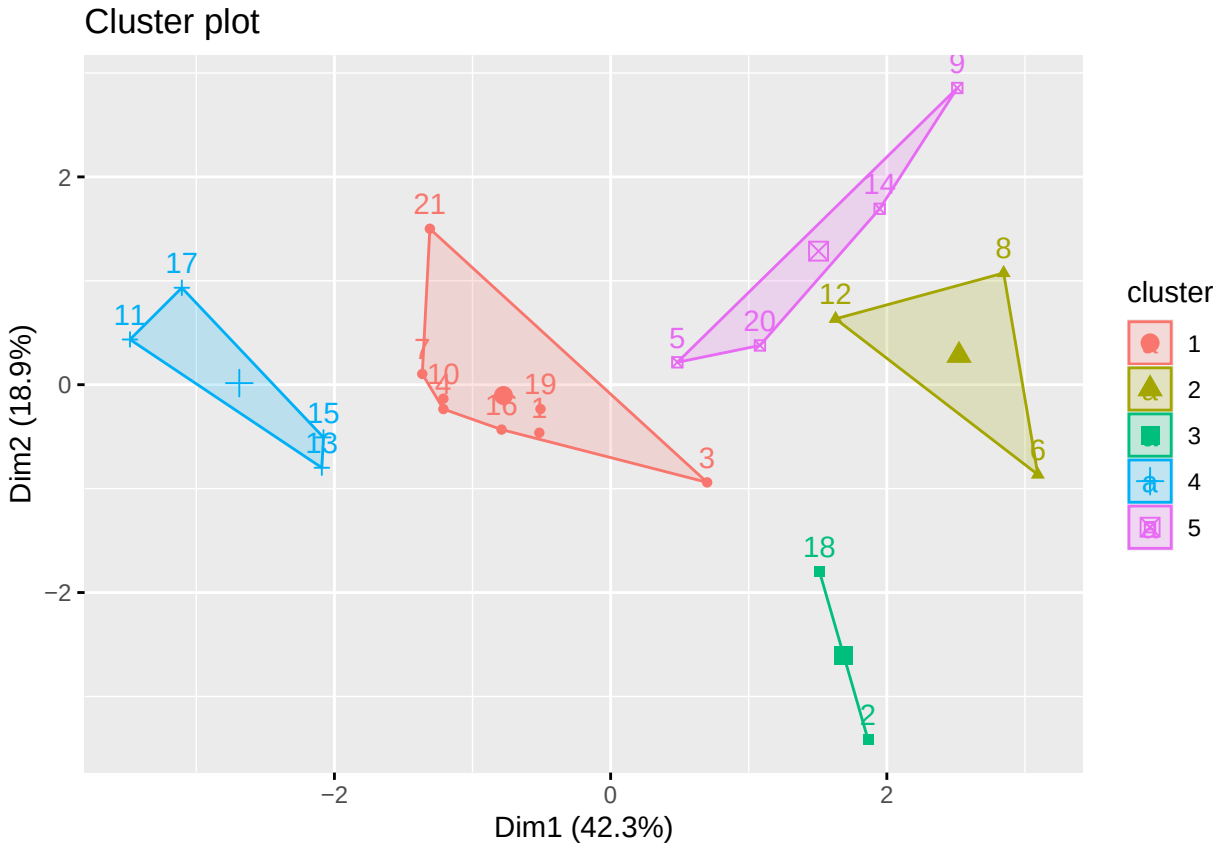
```
##    Market_Cap           Beta            PE_Ratio            ROE
##  Min.   :-0.9768   Min.   :-1.3466   Min.   :-1.3404   Min.   :-1.4515
##  1st Qu.:-0.8763   1st Qu.:-0.6844   1st Qu.:-0.4023   1st Qu.:-0.7223
##  Median :-0.1614   Median :-0.2560   Median :-0.2429   Median :-0.2118
##  Mean   : 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
```

```
#scaling the dataframe (range)
ph_range <- scale(pharma[,3:11])
#summary(ph_range), save for later
```

```
set.seed(123)
distance <- get_dist(ph_scaled)
fviz_dist(distance) #visualize distance between rows of the matrix
```




```
k1 <- kmeans(ph_scaled, centers = 5, nstart = 25)
fviz_cluster(k1, data = ph_scaled)
```




```
print(k1)
```

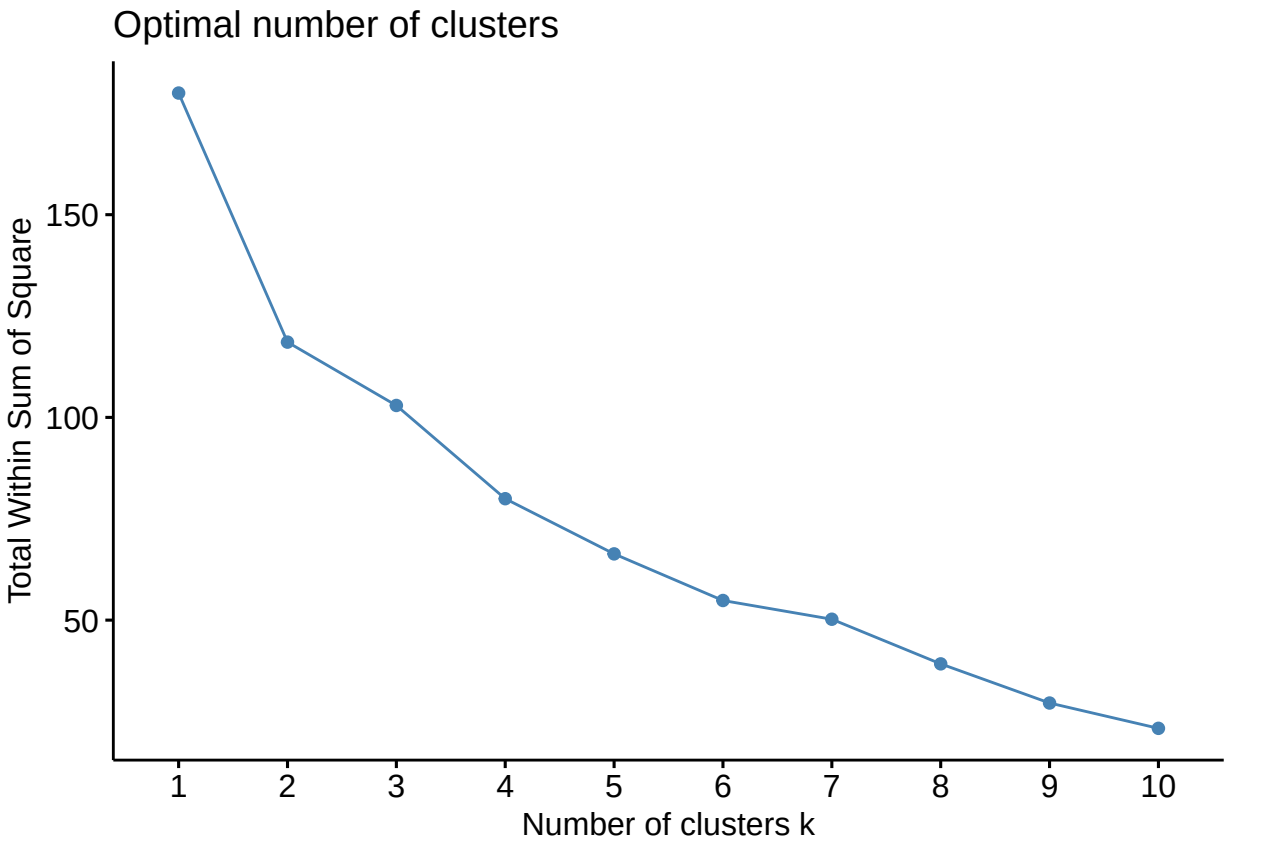
```
## K-means clustering with 5 clusters of sizes 8, 3, 2, 4, 4
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -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"
```

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

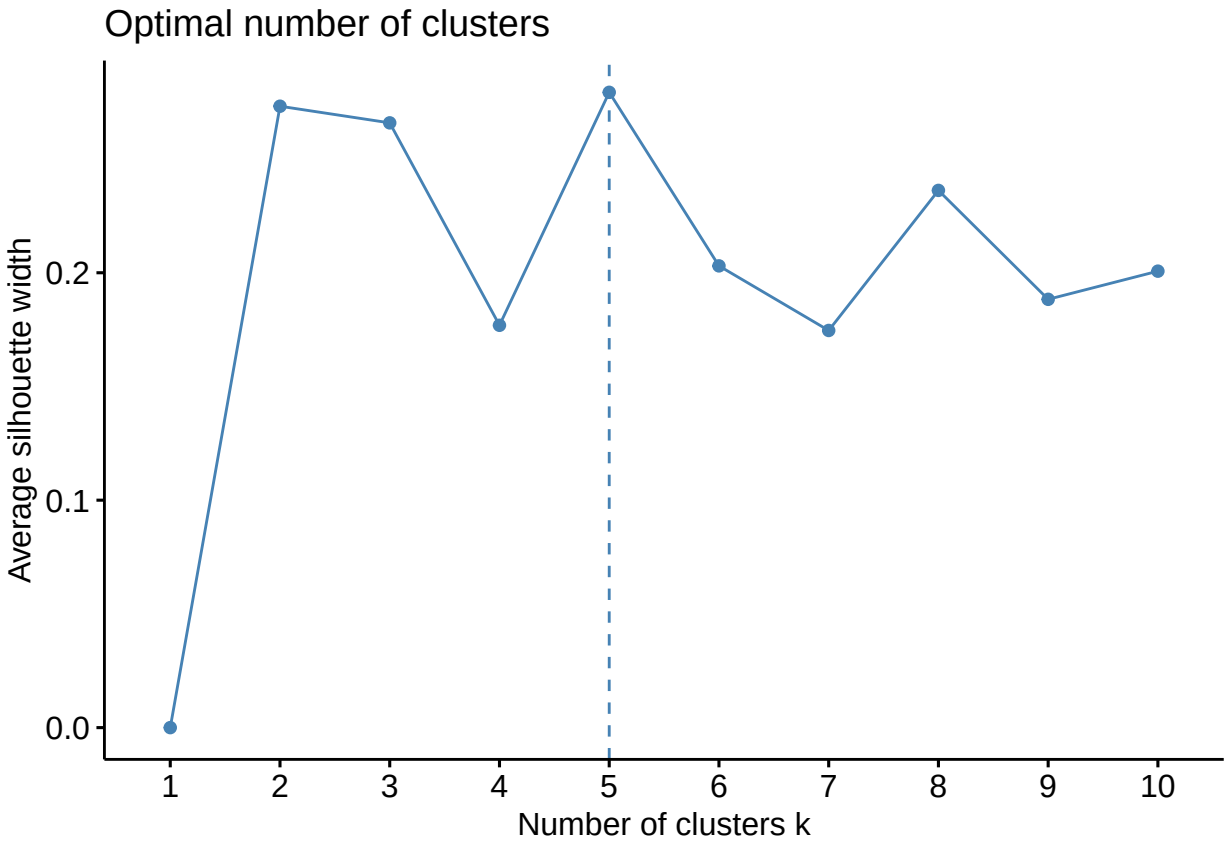
Comments The summary data gives us an overview of the 21 points in 9 numeric columns. Centroid points show the 5 centroid locations and each cluster has a size of 8, 3, 2, 4, 4.
The distance graph shows the distance between rows of the matrix. The darker purple shows the distance is the most and the red shows distance is 0 between same points.
Cluster 1 contains 8 companies including 1, 3, 4, 7, 10, 16, 19, 21
Cluster 2 contains 3 companies including 6, 8, 12
Cluster 3 contains 2 companies including 2, 18
Cluster 4 contains 4 companies including 11, 13, 15, 17
Cluster 5 contains 4 companies including 5, 9, 14, 20

```
fviz_nbclust(ph_scaled, kmeans, method = "wss")
```



Comments We don't see a clear elbow from the graph and it is quite ambiguous. The graph does not show the elbow/knee position and it flattens out more than once at k =4 and 6 respectively.

```
fviz_nbclust(ph_scaled, kmeans, method = "silhouette")
```



Comments It is clear from the silhouette that 5 is the optimal cluster answer.

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

```
#let's look at the mean value from actual data by clusters
aggregate(pharma[3:11], by=list(cluster=k1$cluster), mean)
```

```
##   cluster Market_Cap    Beta PE_Ratio      ROE       ROA Asset_Turnover
## 1       1  55.810000 0.41375  20.2875 28.73750 12.687500         0.7375
## 2       2   6.636667 0.87000  24.6000 16.46667  4.166667         0.6000
## 3       3  31.910000 0.40500  69.5000 13.20000  5.600000         0.7500
## 4       4 157.017500 0.48000  22.2250 44.42500 17.700000         0.9500
## 5       5  13.100000 0.59750  17.6750 14.57500  6.200000         0.4250
##   Leverage Rev_Growth Net_Profit_Margin
## 1 0.371250   5.591250         19.350000
## 2 1.653333   5.733333          7.033333
## 3 0.475000  12.080000          6.400000
## 4 0.220000  18.532500         19.575000
## 5 0.635000  30.142500         15.650000
```

```
dd <- cbind(pharma, cluster = k1$cluster)
#tibble(dd)
```

```
#Here's a more detailed quantitative breakdown by cluster
by(dd, factor(dd$cluster), summary)
```

```
## factor(dd$cluster): 1
##     Symbol              Name             Market_Cap          Beta
##  Length:8           Length:8           Min.   : 6.30   Min.   :0.1800
```

```
##  Class :character   Class :character   1st Qu.:44.67   1st Qu.:0.2875
##  Mode  :character   Mode  :character   Median :59.48   Median :0.4800
##                                        Mean   :55.81   Mean   :0.4138
##                                        3rd Qu.:69.79   3rd Qu.:0.5125
##                                        Max.   :96.65   Max.   :0.6300
##     PE_Ratio          ROE             ROA        Asset_Turnover
##  Min.   :13.10   Min.   :14.90   Min.   : 7.80   Min.   :0.5000
##  1st Qu.:17.65   1st Qu.:21.43   1st Qu.:11.65   1st Qu.:0.6000
##  Median :21.10   Median :26.90   Median :13.35   Median :0.7500
##  Mean   :20.29   Mean   :28.74   Mean   :12.69   Mean   :0.7375
##  3rd Qu.:22.38   3rd Qu.:31.95   3rd Qu.:13.90   3rd Qu.:0.9000
##  Max.   :27.90   Max.   :54.90   Max.   :15.40   Max.   :0.9000
##     Leverage        Rev_Growth     Net_Profit_Margin Median_Recommendation
##  Min.   :0.0000   Min.   :-2.690   Min.   :11.20     Length:8
##  1st Qu.:0.0450   1st Qu.: 2.115   1st Qu.:17.23     Class :character
##  Median :0.3450   Median : 6.630   Median :19.30     Mode  :character
##  Mean   :0.3713   Mean   : 5.591   Mean   :19.35
##  3rd Qu.:0.5400   3rd Qu.: 7.795   3rd Qu.:22.65
##  Max.   :1.1200   Max.   :15.000   Max.   :25.50
##    Location           Exchange            cluster
##  Length:8           Length:8           Min.   :1
##  Class :character   Class :character   1st Qu.:1
##  Mode  :character   Mode  :character   Median :1
##                                        Mean   :1
##                                        3rd Qu.:1
##                                        Max.   :1
## ------------------------------------------------------------
## factor(dd$cluster): 2
##     Symbol              Name              Market_Cap          Beta
##  Length:3           Length:3           Min.   : 0.410   Min.   :0.65
##  Class :character   Class :character   1st Qu.: 1.505   1st Qu.:0.75
##  Mode  :character   Mode  :character   Median : 2.600   Median :0.85
##                                        Mean   : 6.637   Mean   :0.87
##                                        3rd Qu.: 9.750   3rd Qu.:0.98
##                                        Max.   :16.900   Max.   :1.11
##     PE_Ratio          ROE             ROA        Asset_Turnover    Leverage
##  Min.   :19.90   Min.   : 3.90   Min.   :1.400   Min.   :0.6    Min.   :0.000
##  1st Qu.:22.95   1st Qu.:12.65   1st Qu.:2.850   1st Qu.:0.6    1st Qu.:0.725
##  Median :26.00   Median :21.40   Median :4.300   Median :0.6    Median :1.450
##  Mean   :24.60   Mean   :16.47   Mean   :4.167   Mean   :0.6    Mean   :1.653
##  3rd Qu.:26.95   3rd Qu.:22.75   3rd Qu.:5.550   3rd Qu.:0.6    3rd Qu.:2.480
##  Max.   :27.90   Max.   :24.10   Max.   :6.800   Max.   :0.6    Max.   :3.510
##    Rev_Growth     Net_Profit_Margin Median_Recommendation   Location
##  Min.   :-3.170   Min.   : 2.600    Length:3              Length:3
##  1st Qu.: 1.605   1st Qu.: 5.050    Class :character      Class :character
##  Median : 6.380   Median : 7.500    Mode  :character      Mode  :character
##  Mean   : 5.733   Mean   : 7.033
##  3rd Qu.:10.185   3rd Qu.: 9.250
##  Max.   :13.990   Max.   :11.000
##    Exchange            cluster
##  Length:3           Min.   :2
##  Class :character   1st Qu.:2
##  Mode  :character   Median :2
##                     Mean   :2
```

```
##                     3rd Qu.:2
##                     Max.   :2
## ------------------------------------------------------------
## factor(dd$cluster): 3
##     Symbol              Name             Market_Cap         Beta
##  Length:2           Length:2           Min.   : 7.58   Min.   :0.4000
##  Class :character   Class :character   1st Qu.:19.75   1st Qu.:0.4025
##  Mode  :character   Mode  :character   Median :31.91   Median :0.4050
##                                        Mean   :31.91   Mean   :0.4050
##                                        3rd Qu.:44.08   3rd Qu.:0.4075
##                                        Max.   :56.24   Max.   :0.4100
##     PE_Ratio          ROE             ROA       Asset_Turnover     Leverage
##  Min.   :56.5   Min.   :12.90   Min.   :5.50   Min.   :0.600   Min.   :0.3500
##  1st Qu.:63.0   1st Qu.:13.05   1st Qu.:5.55   1st Qu.:0.675   1st Qu.:0.4125
##  Median :69.5   Median :13.20   Median :5.60   Median :0.750   Median :0.4750
##  Mean   :69.5   Mean   :13.20   Mean   :5.60   Mean   :0.750   Mean   :0.4750
##  3rd Qu.:76.0   3rd Qu.:13.35   3rd Qu.:5.65   3rd Qu.:0.825   3rd Qu.:0.5375
##  Max.   :82.5   Max.   :13.50   Max.   :5.70   Max.   :0.900   Max.   :0.6000
##    Rev_Growth    Net_Profit_Margin Median_Recommendation   Location
##  Min.   : 9.16   Min.   :5.50      Length:2              Length:2
##  1st Qu.:10.62   1st Qu.:5.95      Class :character      Class :character
##  Median :12.08   Median :6.40      Mode  :character      Mode  :character
##  Mean   :12.08   Mean   :6.40
##  3rd Qu.:13.54   3rd Qu.:6.85
##  Max.   :15.00   Max.   :7.30
##    Exchange            cluster
##  Length:2           Min.   :3
##  Class :character   1st Qu.:3
##  Mode  :character   Median :3
##                     Mean   :3
##                     3rd Qu.:3
##                     Max.   :3
## ------------------------------------------------------------
## factor(dd$cluster): 4
##     Symbol              Name             Market_Cap         Beta
##  Length:4           Length:4           Min.   :122.1   Min.   :0.3500
##  Class :character   Class :character   1st Qu.:129.9   1st Qu.:0.4325
##  Mode  :character   Mode  :character   Median :153.2   Median :0.4600
##                                        Mean   :157.0   Mean   :0.4800
##                                        3rd Qu.:180.3   3rd Qu.:0.5075
##                                        Max.   :199.5   Max.   :0.6500
##     PE_Ratio          ROE             ROA        Asset_Turnover
##  Min.   :18.00   Min.   :28.60   Min.   :15.00   Min.   :0.800
##  1st Qu.:18.68   1st Qu.:37.60   1st Qu.:15.97   1st Qu.:0.875
##  Median :21.25   Median :43.10   Median :17.75   Median :0.950
##  Mean   :22.23   Mean   :44.42   Mean   :17.70   Mean   :0.950
##  3rd Qu.:24.80   3rd Qu.:49.92   3rd Qu.:19.48   3rd Qu.:1.025
##  Max.   :28.40   Max.   :62.90   Max.   :20.30   Max.   :1.100
##     Leverage       Rev_Growth    Net_Profit_Margin Median_Recommendation
##  Min.   :0.100   Min.   : 9.37   Min.   :14.10     Length:4
##  1st Qu.:0.145   1st Qu.:15.36   1st Qu.:16.95     Class :character
##  Median :0.220   Median :19.61   Median :19.50     Mode  :character
##  Mean   :0.220   Mean   :18.53   Mean   :19.57
##  3rd Qu.:0.295   3rd Qu.:22.79   3rd Qu.:22.12
```

```
##  Max.   :0.340   Max.   :25.54   Max.   :25.20
##    Location           Exchange            cluster
##  Length:4           Length:4           Min.   :4
##  Class :character   Class :character   1st Qu.:4
##  Mode  :character   Mode  :character   Median :4
##                                        Mean   :4
##                                        3rd Qu.:4
##                                        Max.   :4
## ------------------------------------------------------------
## factor(dd$cluster): 5
##     Symbol              Name             Market_Cap          Beta
##  Length:4           Length:4           Min.   : 0.780   Min.   :0.2400
##  Class :character   Class :character   1st Qu.: 1.095   1st Qu.:0.3000
##  Mode  :character   Mode  :character   Median : 2.230   Median :0.5350
##                                        Mean   :13.100   Mean   :0.5975
##                                        3rd Qu.:14.235   3rd Qu.:0.8325
##                                        Max.   :47.160   Max.   :1.0800
##     PE_Ratio          ROE              ROA        Asset_Turnover
##  Min.   : 3.60   Min.   :10.20   Min.   :5.100   Min.   :0.300
##  1st Qu.:14.70   1st Qu.:10.95   1st Qu.:5.325   1st Qu.:0.300
##  Median :19.25   Median :13.15   Median :6.100   Median :0.400
##  Mean   :17.68   Mean   :14.57   Mean   :6.200   Mean   :0.425
##  3rd Qu.:22.23   3rd Qu.:16.77   3rd Qu.:6.975   3rd Qu.:0.525
##  Max.   :28.60   Max.   :21.80   Max.   :7.500   Max.   :0.600
##     Leverage       Rev_Growth    Net_Profit_Margin Median_Recommendation
##  Min.   :0.200   Min.   :26.81   Min.   :12.90     Length:4
##  1st Qu.:0.305   1st Qu.:28.59   1st Qu.:13.20     Class :character
##  Median :0.635   Median :29.77   Median :14.20     Mode  :character
##  Mean   :0.635   Mean   :30.14   Mean   :15.65
##  3rd Qu.:0.965   3rd Qu.:31.33   3rd Qu.:16.65
##  Max.   :1.070   Max.   :34.21   Max.   :21.30
##    Location           Exchange            cluster
##  Length:4           Length:4           Min.   :5
##  Class :character   Class :character   1st Qu.:5
##  Mode  :character   Mode  :character   Median :5
##                                        Mean   :5
##                                        3rd Qu.:5
##                                        Max.   :5
```

```r
#Median recommendation by cluster
table_rec <- table(dd$cluster, dd$Median_Recommendation)
names(dimnames(table_rec)) <- c("Cluster", "Recommendation")
table_rec <- addmargins(table_rec)
table_rec
```

```
##        Recommendation
## Cluster Hold Moderate Buy Moderate Sell Strong Buy Sum
##     1      4           1             2          1   8
##     2      2           1             0          0   3
##     3      1           1             0          0   2
##     4      2           2             0          0   4
##     5      0           2             2          0   4
##     Sum    9           7             4          1  21
```

Comments From the results, we can't determine a clear cut relationship between cluster~Median_Recommendation. A total of 21 recommendation is split into 1 strong buy, 7 moderate buy, 9 hold and 4 moderate sell.

Cluster 1 has a mix of all four recommendations which includes opposite rec on buy and sells. Cluster 2, 3 and 4 contain only mod. buy and hold information. Cluster 5 has both moderate buy and moderate sell recommendation.

```
#Location breakdown by cluster
table_loc <- table(dd$cluster, dd$Location)
names(dimnames(table_loc)) <- c("Cluster", "Location")
table_loc <- addmargins(table_loc)
table_loc
```

```
##        Location
## Cluster CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US Sum
##     1        0      0       0       0           1  2  5   8
##     2        0      0       1       0           0  0  2   3
##     3        1      0       0       0           0  0  1   2
##     4        0      0       0       0           0  1  3   4
##     5        0      1       0       1           0  0  2   4
##     Sum      1      1       1       1           1  3 13  21
```

Comments From the results, we can't determine any relationship between cluster~Location. A total of 21 companies is split into 13 US, 3 UK and 1 for Canana, France, Germany, Ireland and Switzerland each.

Cluster 1 has a mix of US, UK, Switzerland. Cluster 2 has US and Germany. Cluster 3 has US and Canada. Cluster 4 contains US and UK. Cluster 5 has US, France and Ireland.

```
#Exchange breakdown by cluster
table_ex <- table(dd$cluster, dd$Exchange)
names(dimnames(table_ex)) <- c("Cluster", "Exchange")
table_ex <- addmargins(table_ex)
table_ex
```

```
##        Exchange
## Cluster AMEX NASDAQ NYSE Sum
##     1      0      0    8   8
##     2      1      1    1   3
##     3      0      0    2   2
##     4      0      0    4   4
##     5      0      0    4   4
##     Sum    1      1   19  21
```

Comments From the results, we can't determine any relationship between cluster~Exchange. A total of 21 companies is split into 1 Amex, 1 Nasdaq, and 19 NYSE.

Cluster 1 has only NYSE. Cluster 2 has all three. Cluster 3 is only NYSE. Cluster 4 is only NYSE. Cluster 5 is only NYSE. Basically all clusters except cluster 2 is listed in NYSE exclusively

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

Cluster 1: Low_Revenue_Growth- Mix Recommendation- Mostly US comps- All NYSE

Cluster 2: Small Market Cap- Low RoA - Hold or Buy - US comps - Mix exchanges

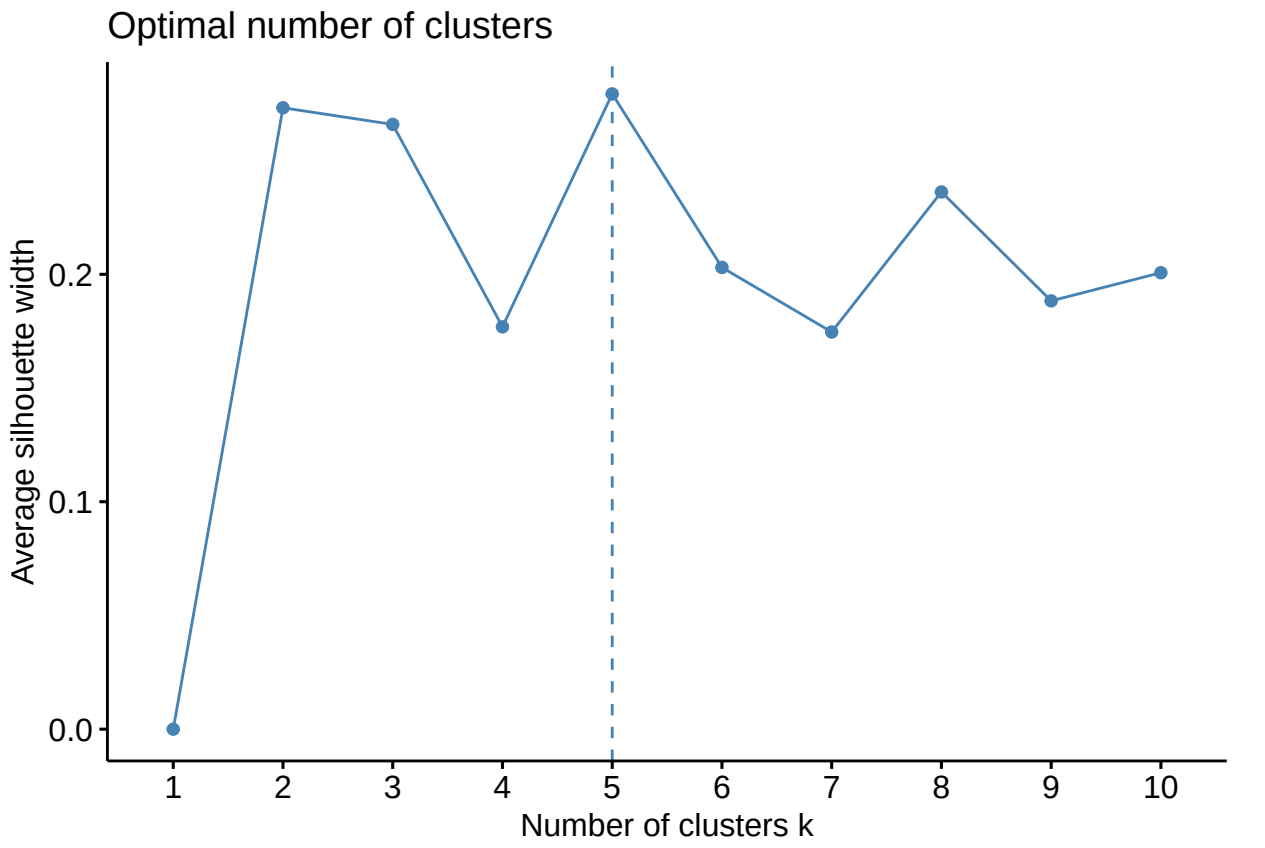
Cluster 3: Low Net_Profit_Margin-High PE ratio- Hold or Buy - NAM comps - NYSE

Cluster 4: High Market Cap - High RoE - High RoA- High Asset Turnover- High NetProfitMargin - Hold or Buy- US comps - NYSE

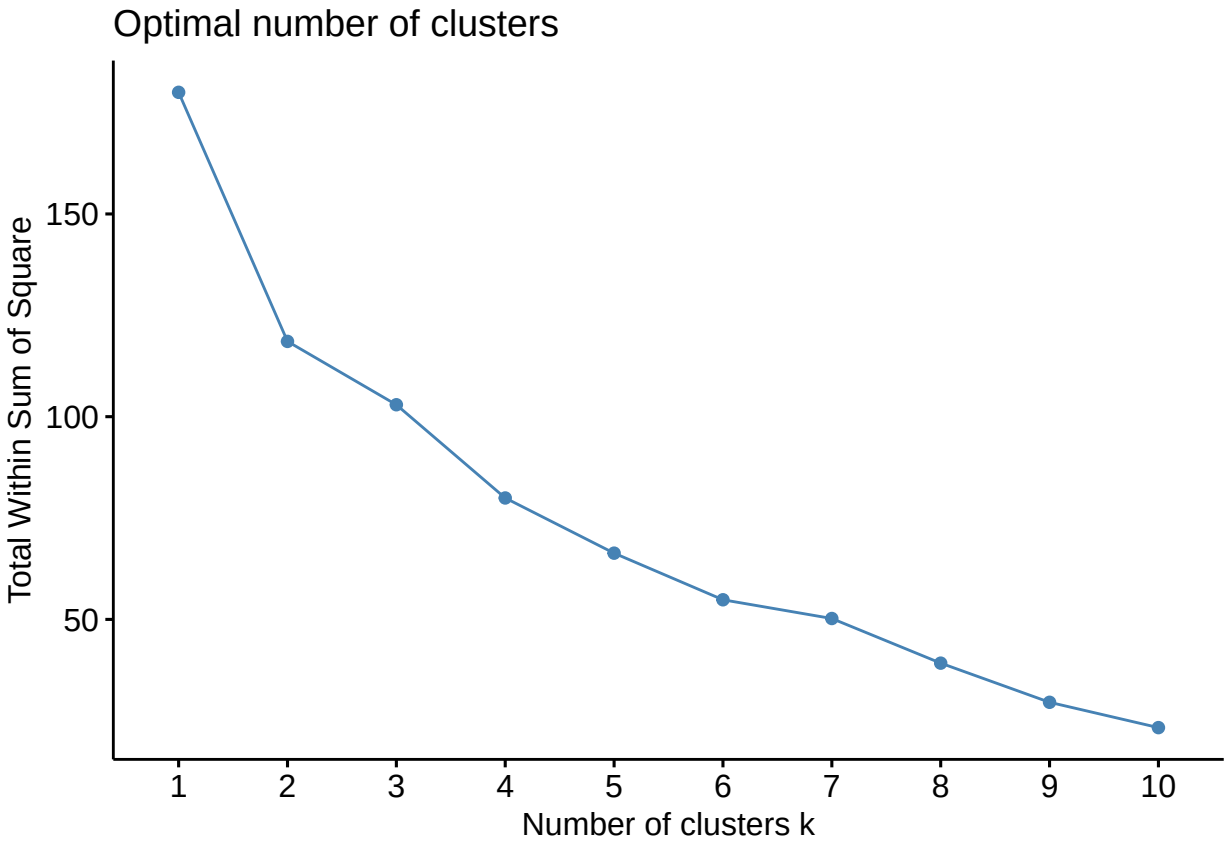
Cluster 5: Low PE ratio-Low RoE-Low Asset Turnover- High revenue growth - mix recommendation - US or European - NYSE

Exploring the alternatives:

```
fviz_nbclust(ph_range, FUN = kmeans, method = "silhouette")
```




```
fviz_nbclust(ph_range, kmeans, method = "wss")
```



We also run test exploring the optimal k through range normalization. The optimal k is 2 from silhouette and 6 from elbow (not clear). Since the k from range normalization is not as ideal, we will stay with z-score normailization data.

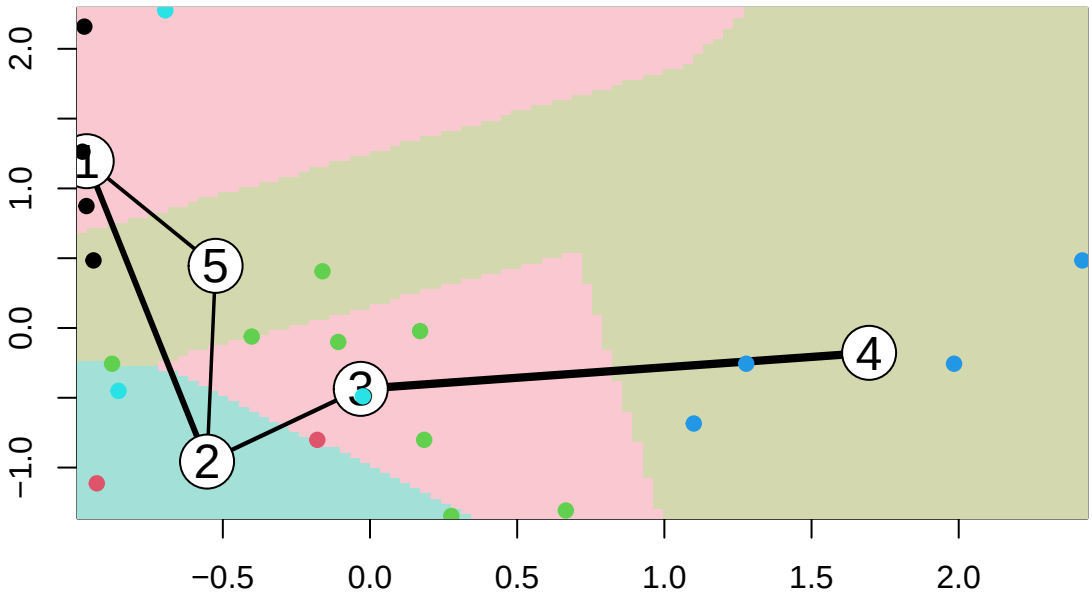
```
set.seed(111)
k2 = kcca(ph_scaled, k=5, kccaFamily("kmeans"))
k2
```

```
## kcca object of family 'kmeans'
##
## call:
## kcca(x = ph_scaled, k = 5, family = kccaFamily("kmeans"))
##
## cluster sizes:
##
## 1 2 3 4 5
## 4 2 8 4 3
```

```
clusters(k2) #cluster membership
```

```
##  [1] 3 5 3 3 2 5 3 1 1 3 4 1 4 1 4 3 4 5 3 2 3
```

```
#Apply the predict() function
clusters_index <- predict(k2)
image(k2)
points(ph_scaled, col=clusters_index, pch=19, cex=1.0)
```



Here we use kcca algorithm instead of kmeans from base R to run kmeans cluster on k =5. The clustering has the same size but different assignment between points compared to base R method. The clustering graph shows the clustering isn't clean cut as we want esp between cluster 1, 3 and 5.

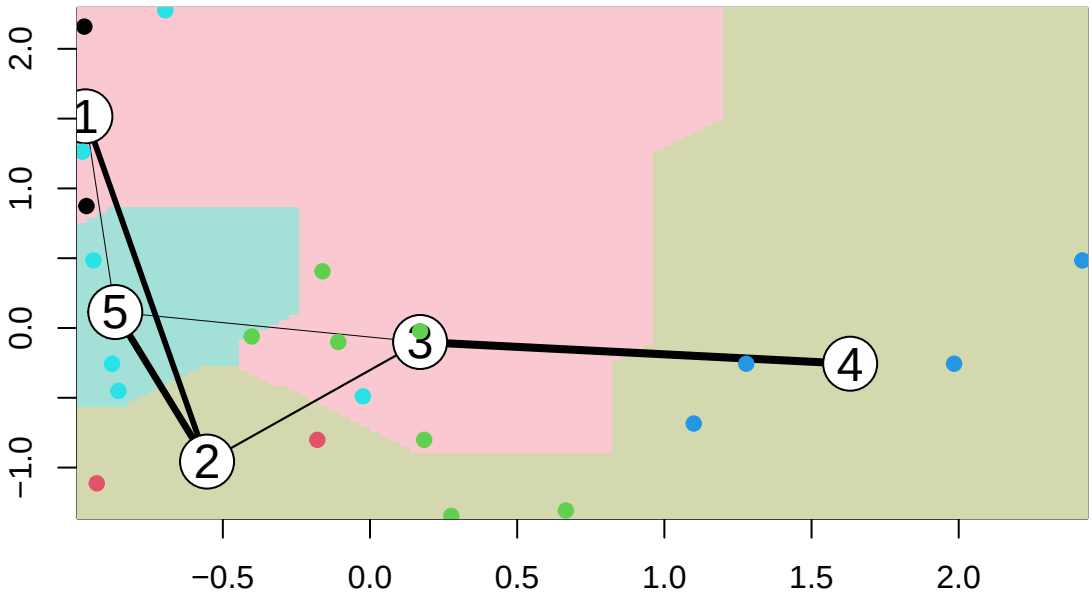
```
set.seed(111)
k2 = kcca(ph_scaled, k=5, kccaFamily("kmedians"))
k2
```

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

```
clusters(k2) #cluster membership
```

```
##  [1] 3 5 5 3 2 5 3 5 1 3 4 5 4 1 4 3 4 5 3 2 3
```

```
#Apply the predict() function
clusters_index <- predict(k2)
image(k2)
points(ph_scaled, col=clusters_index, pch=19, cex=1.0)
```



If we switch to kmedian from kmeans in kcca, the size of the five clusters are 2, 2, 7, 4, 6. Still, the clustering isn't as clean cut. We are exploaring the additional to see if there are better methods or k we can use to improve the visual cluster but it is not clear that a better cluster exists.