Assignment 4

2023-10-23

Install required packages

```
library(tidyverse)  # data manipulation
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# install.packages("factoextra")
library(factoextra)
```

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

```
library(ISLR)
library(httr)
library(flexclust)
```

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

```
library(caret)
```

```
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:httr':
##
##     progress
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
set.seed(123)
```

Import data set

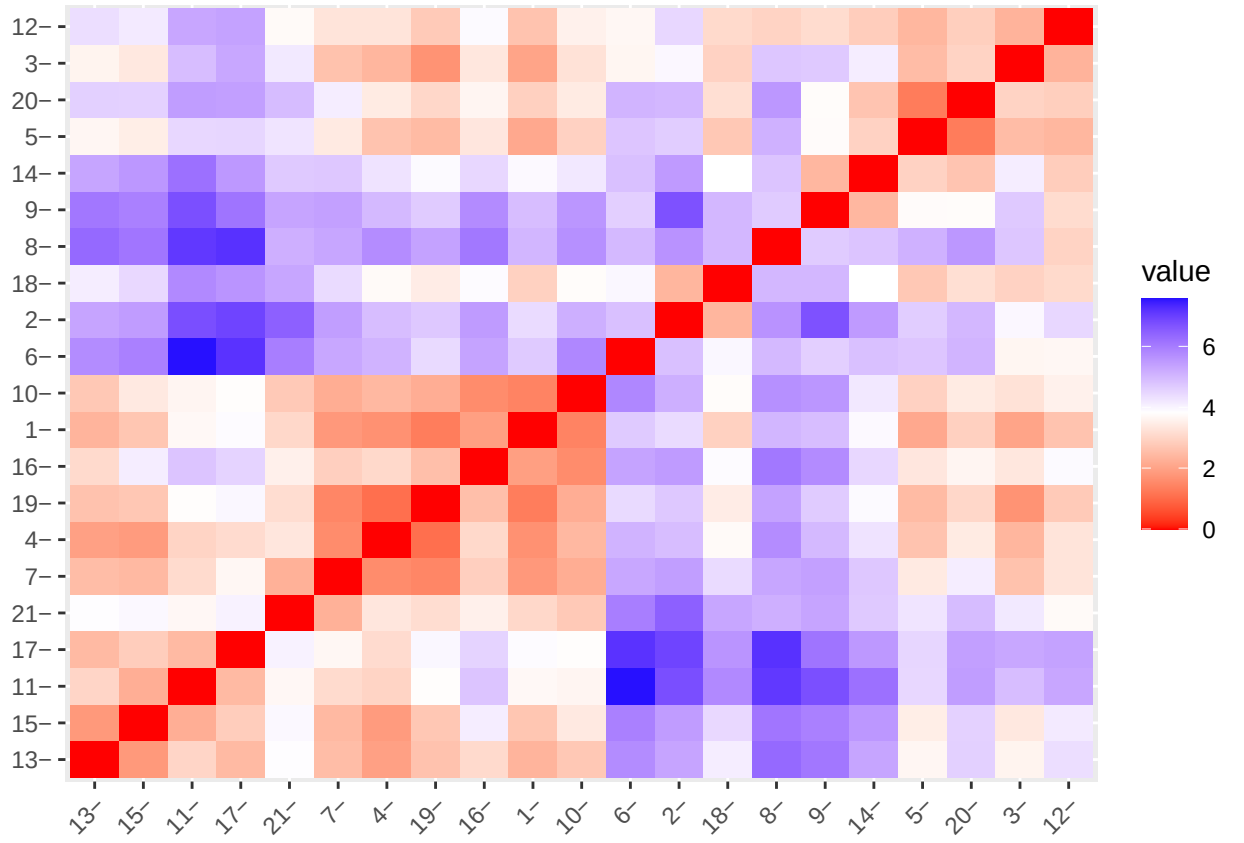
```r
library(readr)
Pharmaceuticals <- read_csv("Pharmaceuticals.csv")
```

```
## Rows: 21 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Symbol, Name, Median_Recommendation, Location, Exchange
## dbl (9): Market_Cap, Beta, PE_Ratio, ROE, ROA, Asset_Turnover, Leverage, Rev...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

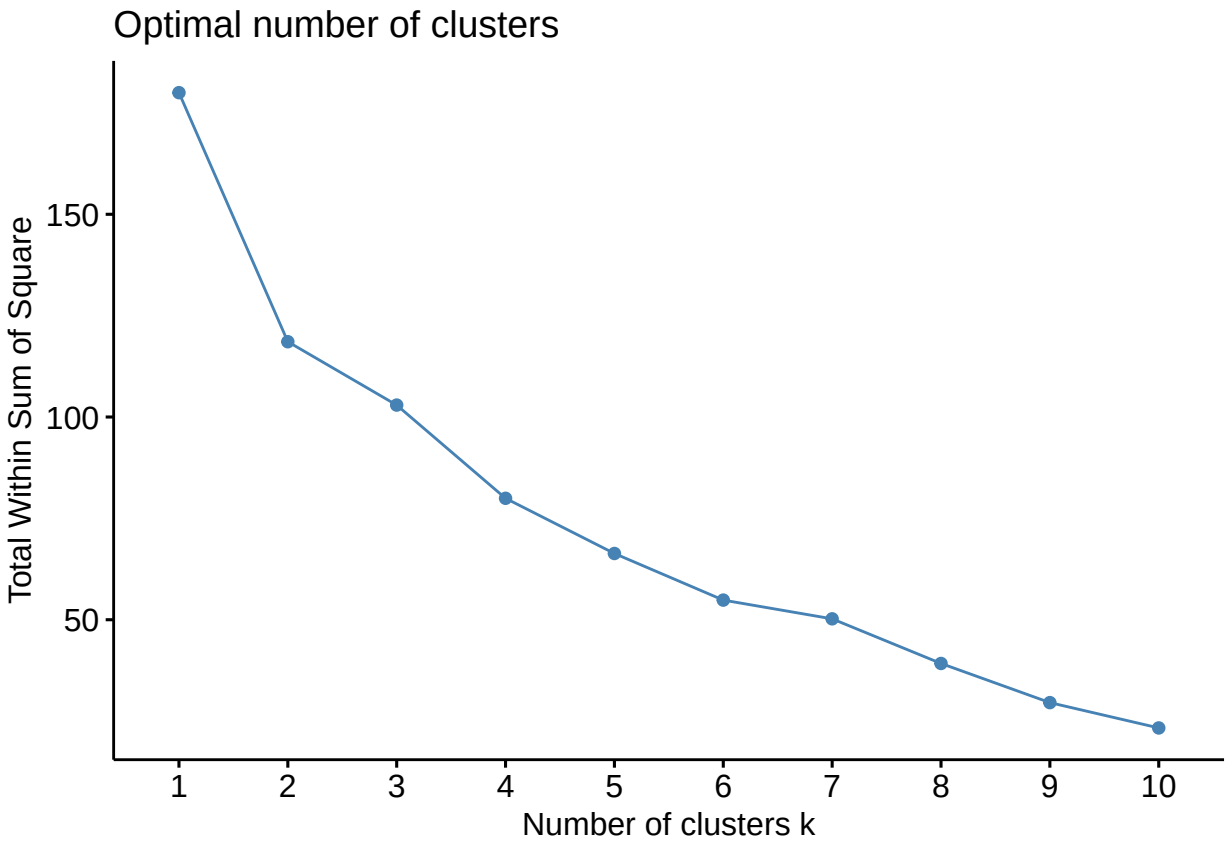
```r
View(Pharmaceuticals)
```

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.

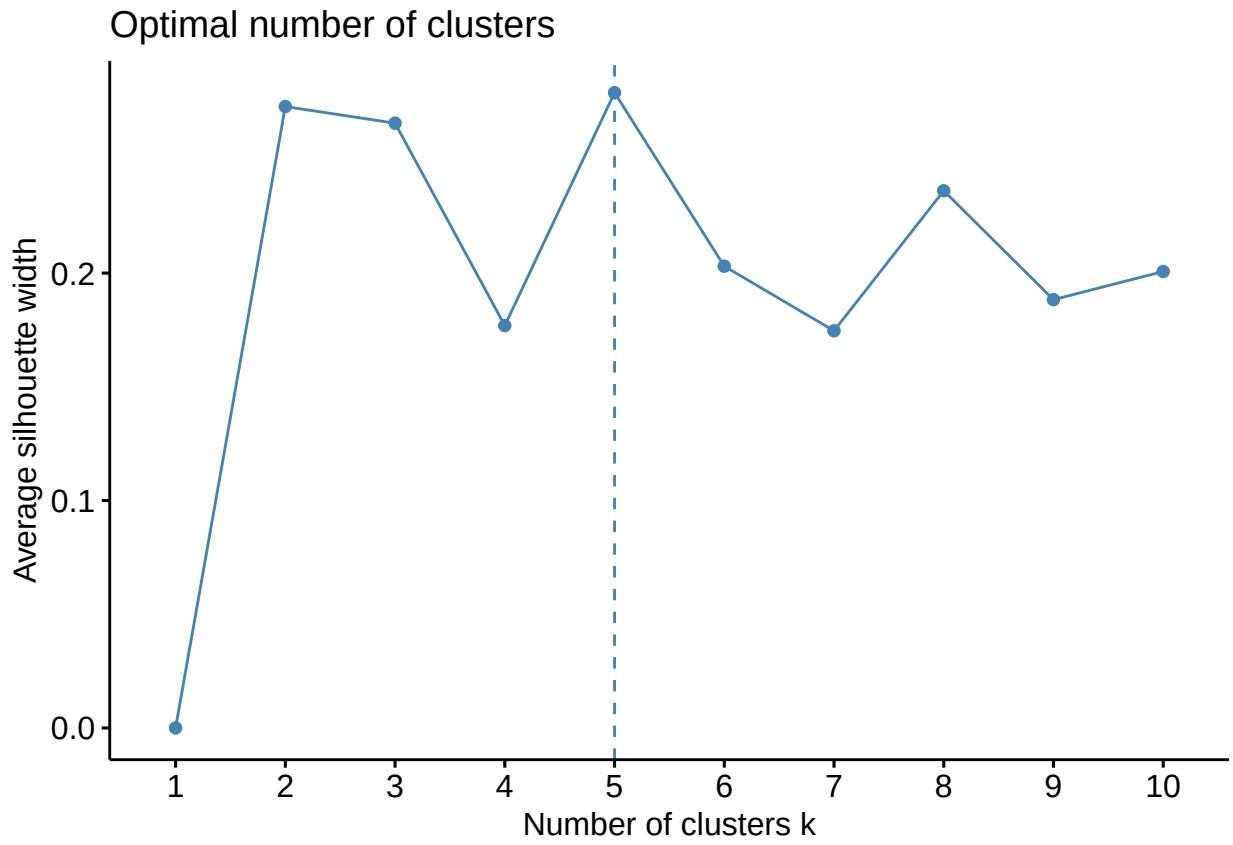
```r
# Select variables 1 to 9
Pharmaceuticals.data <- Pharmaceuticals[, -c(1,2,12,13,14)]

# Scale the data
Pharmaceuticals.scaled<-scale(Pharmaceuticals.data)
distance <-get_dist(Pharmaceuticals.scaled)
fviz_dist(distance)
```




```
# Determine k
fviz_nbclust(Pharmaceuticals.scaled, kmeans, method = "wss") # using Elbow Method
```




```
fviz_nbclust(Pharmaceuticals.scaled, kmeans, method = "silhouette") # Using silhouette method
```



From analyzing the graphs we can see the best value of k is 5. Adding more or having less clusters than 5 will bring less improvement to cluster homogeneity.

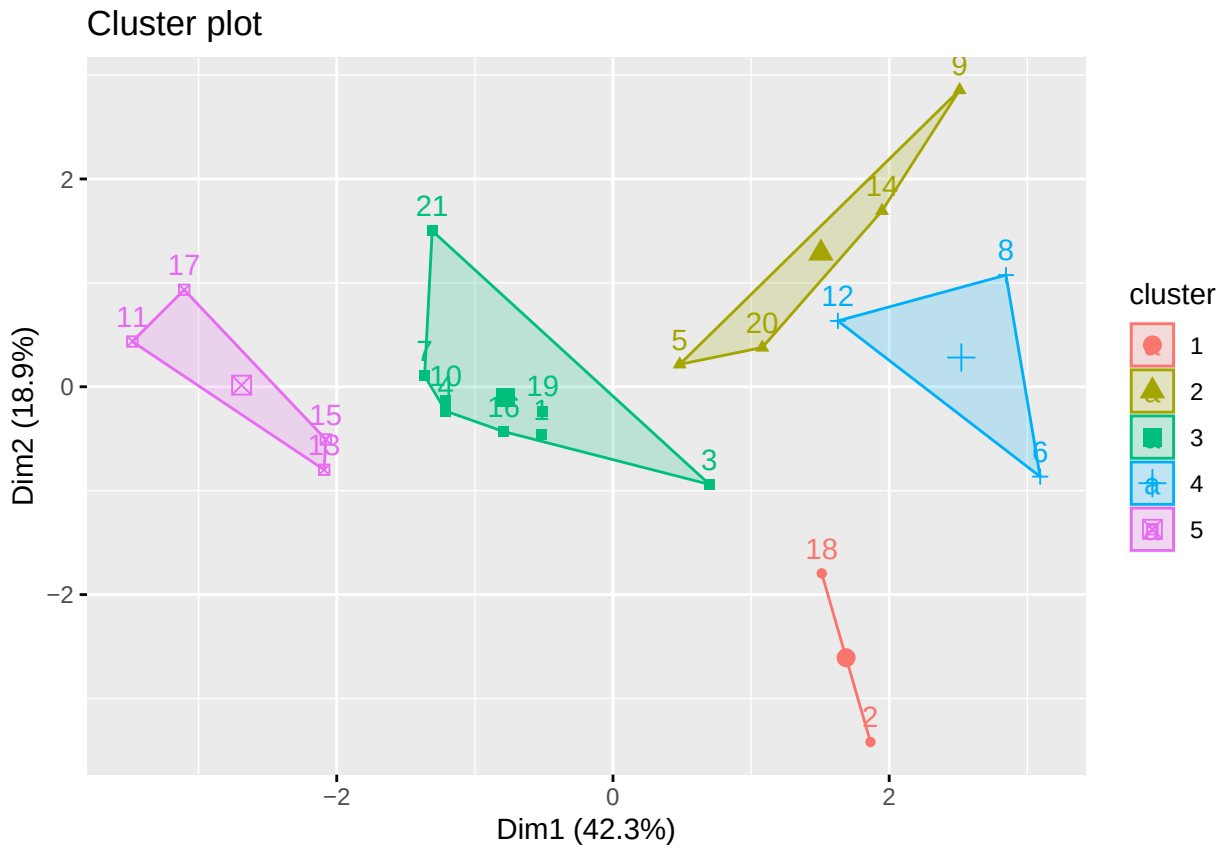
```
# Cluster analysis - we chose k-means since we know the number of clusters that are best for the analys
# Numer of clusters formed = 5. We can find by using an elbow chart and the Silhouette Method
# By default we use Euclidean distance
k5 <- kmeans(Pharmaceuticals.scaled, centers = 5, nstart = 25)

# Visualize the output
k5$centers
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 2 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 4 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.14170336 -0.1168459      -1.416514761
## 2  0.06308085  1.5180158      -0.006893899
## 3 -0.27449312 -0.7041516       0.556954446
## 4  1.36644699 -0.6912914      -1.320000179
## 5 -0.46807818  0.4671788       0.591242521
```

```
k5$size
```

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

```
fviz_cluster(k5, data = Pharmaceuticals.scaled)
```



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

```
# Print the mean value of the variables by cluster
Pharmaceuticals.data %>%
  mutate(Cluster = k5$cluster) %>%
  group_by(Cluster) %>%
  summarise_all("mean")
```

```
## # A tibble: 5 x 10
##   Cluster Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##     <int>      <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
## 1       1      31.9  0.405     69.5  13.2  5.6           0.75     0.475
## 2       2      13.1  0.598     17.7  14.6  6.2           0.425    0.635
## 3       3      55.8  0.414     20.3  28.7 12.7           0.738    0.371
## 4       4       6.64 0.87      24.6  16.5  4.17          0.6      1.65
## 5       5     157.   0.48      22.2  44.4 17.7           0.95     0.22
## # i 2 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>
```

```
# Print all data adding a cluster column. To visualize the cluster assigned to each company.
full.data <- cbind(Pharmaceuticals, cluster = k5$cluster)
tibble(full.data)
```

```
## # A tibble: 21 x 15
```

```
##    Symbol Name     Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##    <chr>  <chr>         <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
##  1 ABT    Abbott ~      68.4   0.32     24.7  26.4  11.8            0.7     0.42
##  2 AGN    Allerga~       7.58  0.41     82.5  12.9   5.5            0.9     0.6
##  3 AHM    Amersha~       6.3   0.46     20.7  14.9   7.8            0.9     0.27
##  4 AZN    AstraZe~      67.6   0.52     21.5  27.4  15.4            0.9     0
##  5 AVE    Aventis       47.2   0.32     20.1  21.8   7.5            0.6     0.34
##  6 BAY    Bayer AG      16.9   1.11     27.9   3.9   1.4            0.6     0
##  7 BMY    Bristol~      51.3   0.5      13.9  34.8  15.1            0.9     0.57
##  8 CHTT   Chattem~       0.41  0.85     26    24.1   4.3            0.6     3.51
##  9 ELN    Elan Co~       0.78  1.08      3.6  15.1   5.1            0.3     1.07
## 10 LLY    Eli Lil~      73.8   0.18     27.9  31    13.5            0.6     0.53
## # i 11 more rows
## # i 6 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>,
## #   Median_Recommendation <chr>, Location <chr>, Exchange <chr>, cluster <int>
```

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

```
# Breakdown clusters by median recommendation
Recomendation <- table(k5$cluster, Pharmaceuticals$Median_Recommendation)
names(dimnames(Recomendation)) <- c("Cluster", "Recommendation")
Recomendation <- addmargins(Recomendation)
Recomendation
```

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

```
# Breakdown cluster by the location of the firm's headquarters
Location.firm <- table(k5$cluster, Pharmaceuticals$Location)
names(dimnames(Location.firm)) <- c("Cluster", "Location")
Location.firm <- addmargins(Location.firm)
Location.firm
```

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

```r
# Breakdown clusters by the stock exchange on which the firm is listed
Stock.Exchange <- table(k5$cluster, Pharmaceuticals$Exchange)
names(dimnames(Stock.Exchange)) <- c("Cluster", "Stock Exchange")
Stock.Exchange <- addmargins(Stock.Exchange)
Stock.Exchange
```

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

```r
# Create a new data set to include the cluster column
Pharma.Cluster<-Pharmaceuticals
Pharma.Cluster$Cluster <- as.factor(k5$cluster)

# To create a mode table we define the mode function
mode_stat <- function(x) {
  tbl <- table(x)
  name <- names(tbl)[which.max(tbl)]
  if (is.null(name)) {
    return(NA)
  } else {
    return(name)
  }
}

pattern.table <- Pharma.Cluster[,c(12:15)]
pattern.table <- aggregate(pattern.table[-4], pattern.table[4], mode_stat)

print(pattern.table)
```

```
##   Cluster Median_Recommendation Location Exchange
## 1       1                 Hold   CANADA     NYSE
## 2       2          Moderate Buy       US     NYSE
## 3       3                 Hold       US     NYSE
## 4       4                 Hold       US     AMEX
## 5       5                 Hold       US     NYSE
```

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

Cluster 1 - Moderate Risk High PE Ratio - Moderate asset turnover, low profit margin, recommendation to hold Cluster 2 - High Company Growth High Risk - moderate net profit margin, high revenue growth, moderate recommendation to buy Cluster 3 - Moderate Risk High Profitability - high net profit margin, recommended to hold Cluster 4 - High Risk Low Profitability - low ROA, high leverage, recommended to hold Cluster 5 - Stable High Profit - high asset turnover, recommended to hold