Assignment 4

Lukas van der Watt

11/7/2021

```r
library(tidyverse)  # used in data manipulation
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

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

```r
library(factoextra) # for clustering algorithms & visualization
```

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

```r
library(ISLR)
library(cluster)
library(dplyr)
set.seed(123)
```

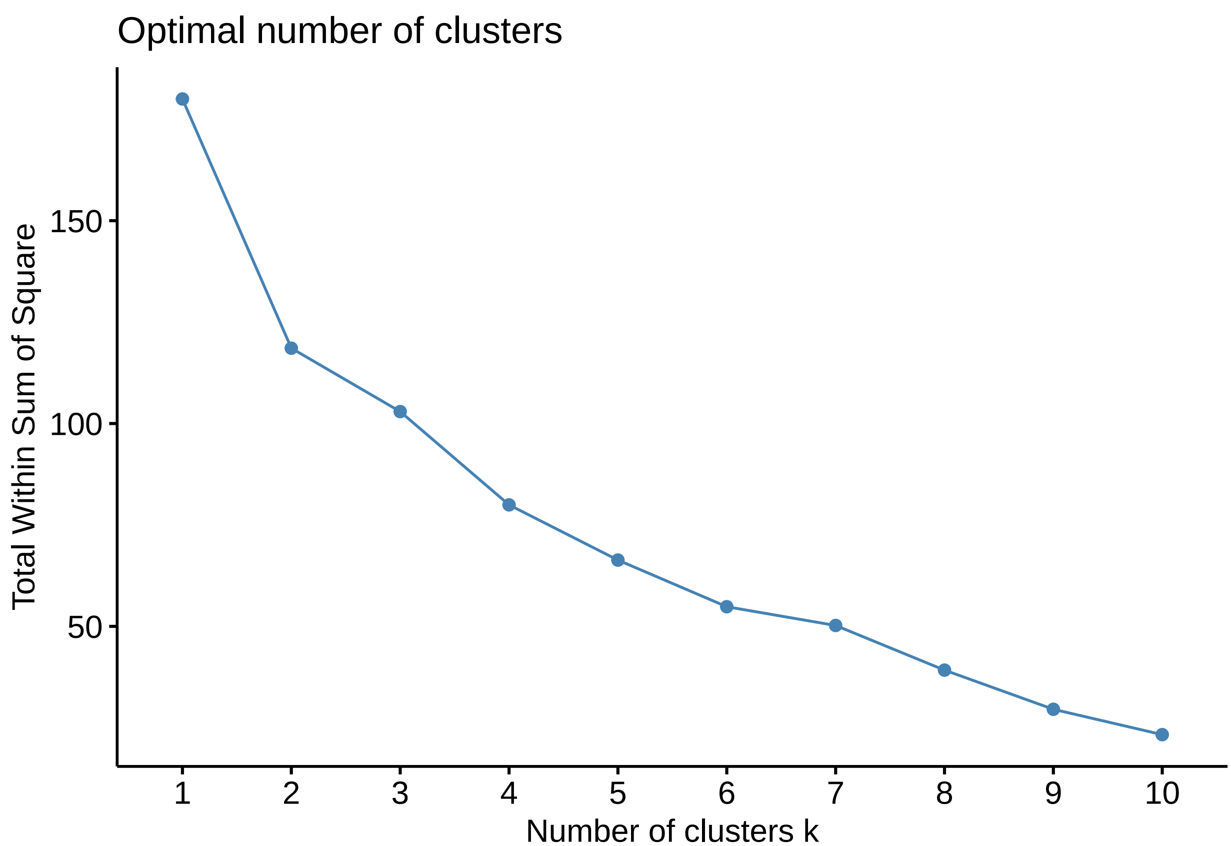
```r
Pharm <- read.csv('Pharmaceuticals.csv')
str(Pharm)
```

```
## 'data.frame':    21 obs. of  14 variables:
##  $ Symbol               : chr  "ABT" "AGN" "AHM" "AZN" ...
##  $ Name                 : chr  "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZeneca PLC
##  $ Market_Cap           : num  68.44 7.58 6.3 67.63 47.16 ...
##  $ Beta                 : num  0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...
##  $ PE_Ratio             : num  24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...
##  $ ROE                  : num  26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...
##  $ ROA                  : num  11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...
##  $ Asset_Turnover       : num  0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
##  $ Leverage             : num  0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...
##  $ Rev_Growth           : num  7.54 9.16 7.05 15 26.81 ...
##  $ Net_Profit_Margin    : num  16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...
##  $ Median_Recommendation: chr  "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sell" ...
##  $ Location             : chr  "US" "CANADA" "UK" "UK" ...
##  $ Exchange             : chr  "NYSE" "NYSE" "NYSE" "NYSE" ...
```

From the structure we see that variables 3 to 11 are numerical. We will use thes variables for the k means clustering analysis.

```
P_data <- Pharm[,3:11]
P_data <- scale(P_data) # We need to scale the data in order to have releavant numbers that are free of
distance <- get_dist(P_data)
#fviz_dist(distance)
```

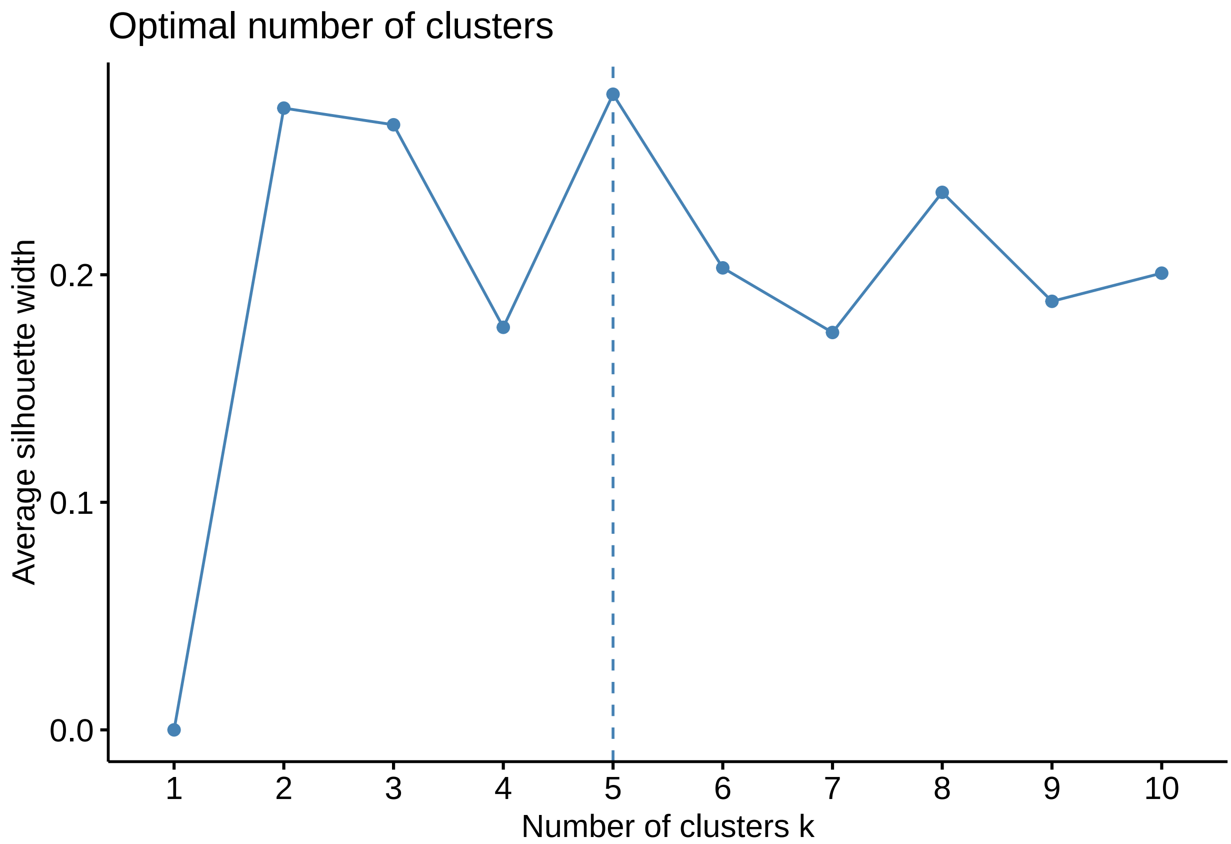
```
#Determining K value
fviz_nbclust(P_data,kmeans,method = "wss") # We first have to determine the optimal k value using the "
```




```
#(Within-cluster sum of squares - WSS ) measures "compactness" of clustersthe meaning the smaller the v
```

In the elbow chart we identify the "Elbow Point" which is the optimal number of clusters as k=5, because we can see that as the k value increases the sum of squares decreases at a smaller rate. (Slope is lesser than that of the first four k-values). Going beyond a k value of 5 (5 clusters) would bring less improvement to cluster homogeneity.

```
fviz_nbclust(P_data,kmeans,method = "silhouette") # In this statement we are essentially doing the same
```



Both the Elbow chart and the Silhouette chart indicate the same results.

```
# K Is a hyperparamater calculated externally from the data. Note: A parameter is calculated from the d
K5 <- kmeans(P_data, centers = 5, nstart = 25) # using kmeans (euclidean distance) where k = 5 and the 
# Visualize the output
K5$centers # output shows the centroids of each cluster per column variable
```

```
##    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
```

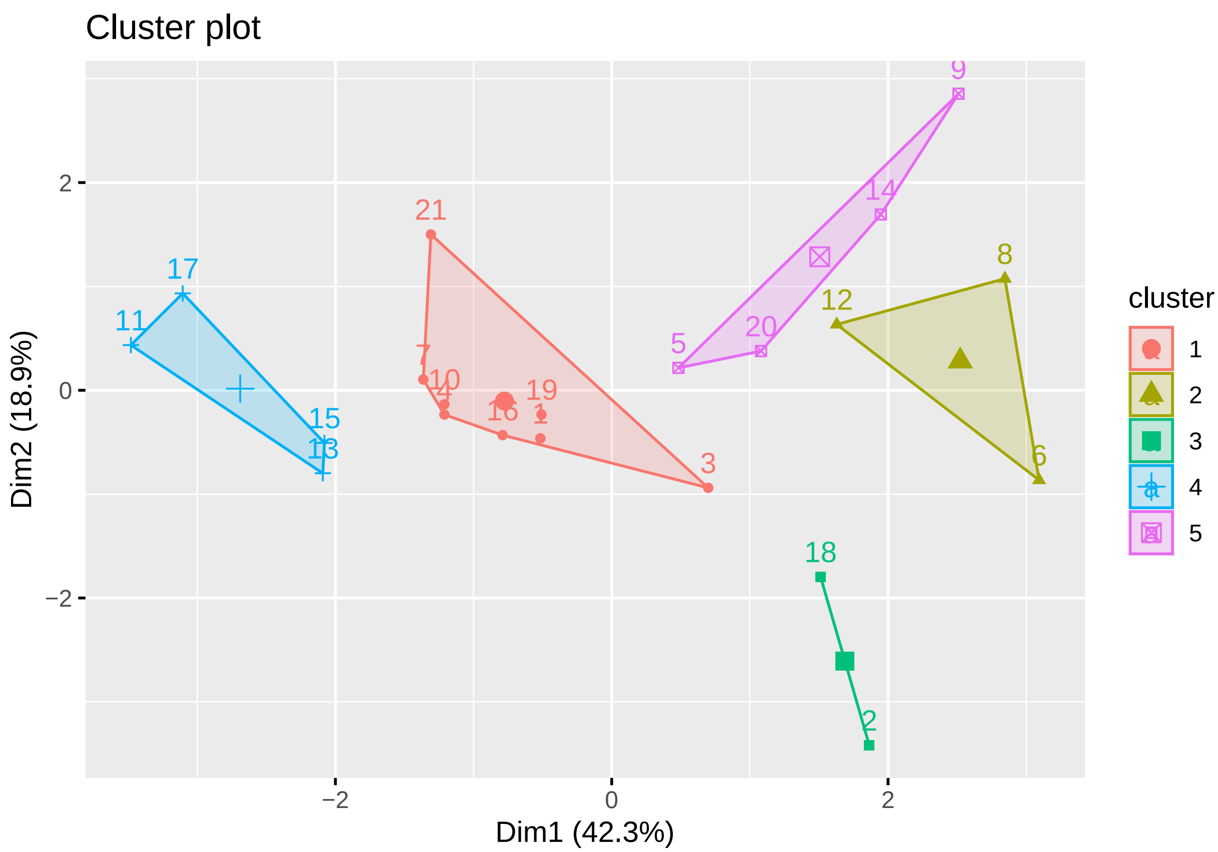
```
K5$size # This shows the number of observations/items in each cluster
```

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

```
K5$cluster[c(1,2,3,19,20,21)] # This can identify in which cluster an observation belongs in.
```

```
## [1] 1 3 1 1 5 1
```

```
fviz_cluster(K5, data = P_data) # Visualize the output
```



In this elbow chart we see the justification of the 5 clusters and how they are formed with the various observations. For example We can easily see to which cluster each observation belongs to. In the case of the *3rd cluster for example the 2nd and 18th observation is clustered together. By looking at the centroids we can also see a trend of the observations where market growth is smaller in general having a negative value with the exception of one observation. This forms a cluster of values below the 0 axis.Similarly most other variable centroids can be looked at to find a pattern which will attribute to the cluster distribution. For example we can see that cluster 4 has a low Market_Cap, Beta and PE_Ratio. Cluster 1 has a low Market_Cap, PE_Ratio,ROE,ROA and Asset_Turnover. Cluster 5 has a high market_cap, ROE,ROA and Asset_Turnover.

```
k <- kmeans(P_data,centers=5)
```

```
#Part C
aggregate(P_data,list(k$cluster),FUN = mean)
```

```
##   Group.1  Market_Cap      Beta    PE_Ratio        ROE        ROA
## 1       1 -0.97676686  1.2630872  0.03299122 -0.1123792 -1.1677918
## 2       2 -0.79605926  0.3205014 -0.45014035 -0.6533148 -0.7881923
```

```
## 3       3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915
## 4       4 -0.52462814  0.4451409  1.84984387 -1.0404550 -1.1865838
## 5       5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431
##   Asset_Turnover   Leverage Rev_Growth Net_Profit_Margin
## 1  -4.612656e-01  3.7427970 -0.6327607        -1.2488842
## 2  -1.107037e+00  0.2717048  1.2256188        -0.1486179
## 3   1.729746e-01 -0.2744931 -0.7041516         0.5569544
## 4   1.480297e-16 -0.3443544 -0.5769454        -1.6095439
## 5   1.153164e+00 -0.4680782  0.4671788         0.5912425
```

```
Pharm_with_clusters <- mutate(Pharm,(k$cluster))
head(Pharm_with_clusters)
```

```
##   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
##   (k$cluster)
## 1           3
## 2           4
## 3           3
## 4           3
## 5           2
## 6           4
```

```
LA <- Pharm_with_clusters %>% select(,(c(12,13,14,15)))
colnames(LA) <- c('Median_Recommendation' , 'Location' , 'Exchange', 'Cluster')
LA <- LA[order(LA$Cluster),]
LA
```

```
##    Median_Recommendation    Location Exchange Cluster
## 8           Moderate Buy          US   NASDAQ       1
## 5           Moderate Buy      FRANCE     NYSE       2
## 9          Moderate Sell     IRELAND     NYSE       2
## 12                  Hold          US     AMEX       2
## 14          Moderate Buy          US     NYSE       2
## 20         Moderate Sell          US     NYSE       2
## 1           Moderate Buy          US     NYSE       3
## 3             Strong Buy          UK     NYSE       3
## 4          Moderate Sell          UK     NYSE       3
## 7          Moderate Sell          US     NYSE       3
## 10                  Hold          US     NYSE       3
## 16                  Hold SWITZERLAND     NYSE       3
```

```
## 19            Hold      US    NYSE      3
## 21            Hold      US    NYSE      3
## 2     Moderate Buy  CANADA    NYSE      4
## 6             Hold GERMANY    NYSE      4
## 18            Hold      US    NYSE      4
## 11            Hold      UK    NYSE      5
## 13    Moderate Buy      US    NYSE      5
## 15            Hold      US    NYSE      5
## 17    Moderate Buy      US    NYSE      5
```

```
#This new data frame shows the cluster added as a column to the original data.
```

PART D we can name the Clusters based on the information they contain. Cluster 1 : New York Stock Exchange for US, UK and Switzerland Cluster 2 : New York Stock Exchange for US and UK only Cluster 3 : Diverse Exchange (Most diverse exchange in the data) Cluster 4 : New York Stock Exchange for US, UK and France Cluster 5 : New York Stock Exchange for US, Canada and Germany