Assignment 4

Jacob Aylward

2025-06-16

```
library(readr)
Pharmaceuticals <- read_csv("C:/Users/jacob/Downloads/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.
```

```
View(Pharmaceuticals)
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ purrr     1.0.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(factoextra)
```

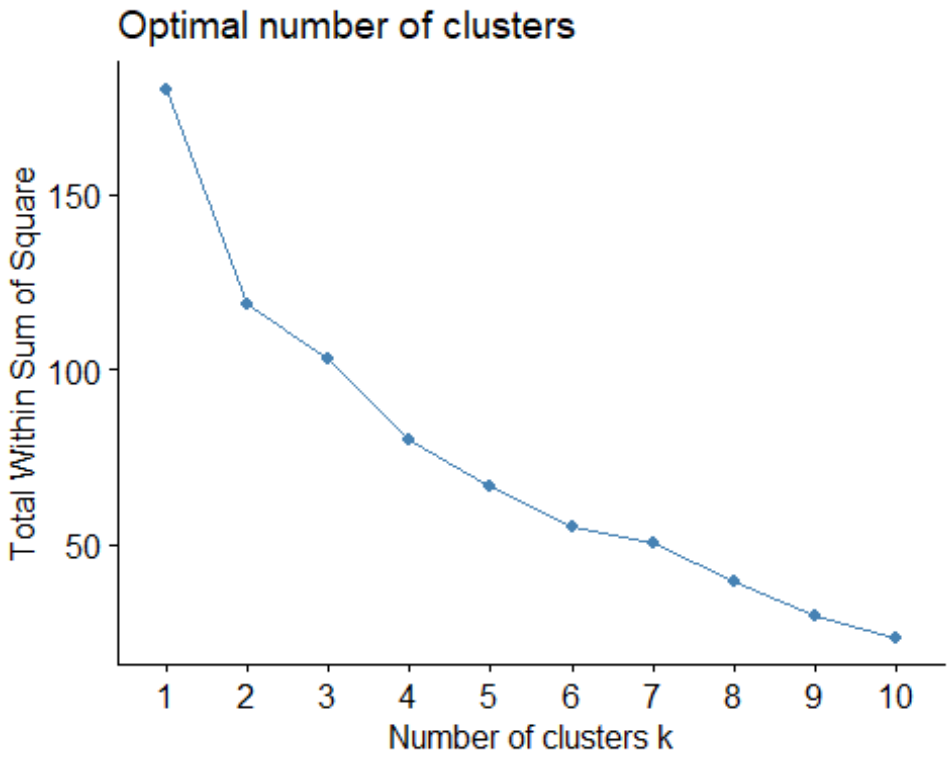
```
## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa
```

```
set.seed(123)
PH<- Pharmaceuticals[,c(3:11)]
summary(PH)
```

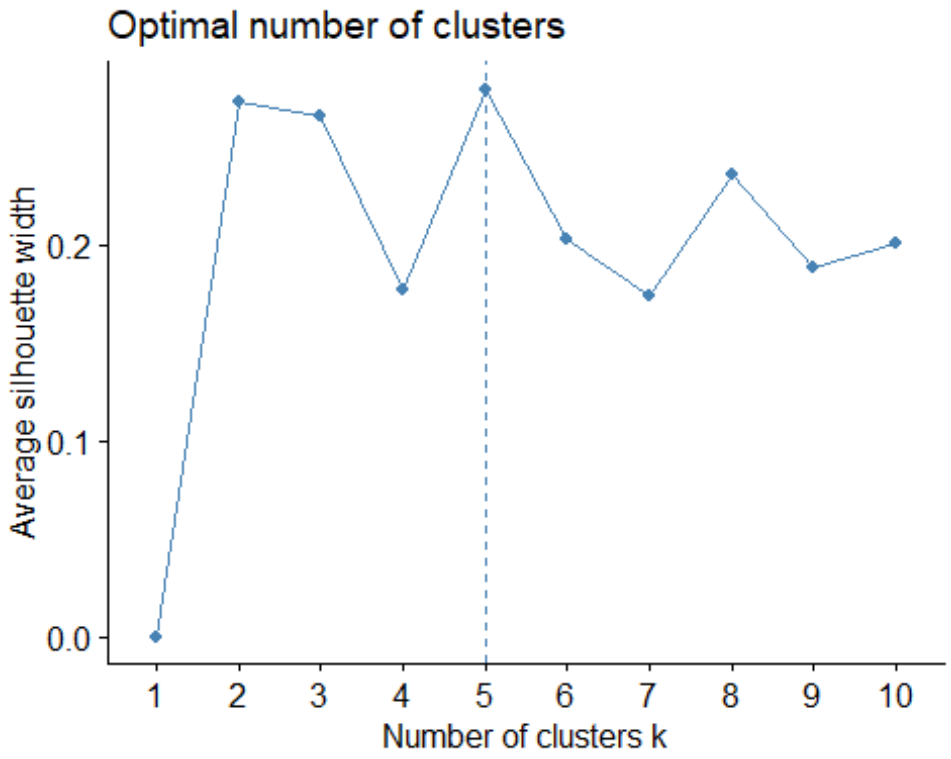
```
##    Market_Cap          Beta            PE_Ratio          ROE
##  Min.   :  0.41   Min.   :0.1800   Min.   : 3.60   Min.   : 3.9
##  1st Qu.:  6.30   1st Qu.:0.3500   1st Qu.:18.90   1st Qu.:14.9
```

```
##  Median : 48.19   Median :0.4600   Median :21.50   Median :22.6
##  Mean   : 57.65   Mean   :0.5257   Mean   :25.46   Mean   :25.8
##  3rd Qu.: 73.84   3rd Qu.:0.6500   3rd Qu.:27.90   3rd Qu.:31.0
##  Max.   :199.47   Max.   :1.1100   Max.   :82.50   Max.   :62.9
##       ROA        Asset_Turnover    Leverage        Rev_Growth
##  Min.   : 1.40   Min.   :0.3    Min.   :0.0000   Min.   :-3.17
##  1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600   1st Qu.: 6.38
##  Median :11.20   Median :0.6    Median :0.3400   Median : 9.37
##  Mean   :10.51   Mean   :0.7    Mean   :0.5857   Mean   :13.37
##  3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000   3rd Qu.:21.87
##  Max.   :20.30   Max.   :1.1    Max.   :3.5100   Max.   :34.21
##  Net_Profit_Margin
##  Min.   : 2.6
##  1st Qu.:11.2
##  Median :16.1
##  Mean   :15.7
##  3rd Qu.:21.1
##  Max.   :25.5
```

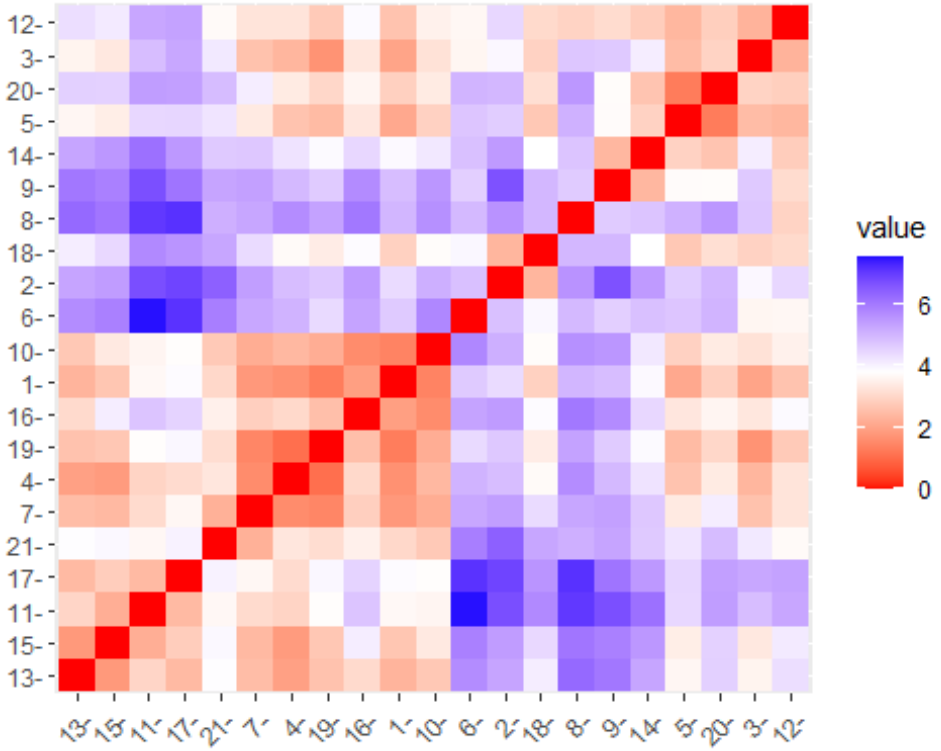
```
set.seed(123)
PH<- Pharmaceuticals[,c(3:11)]
PH <- scale(PH)
fviz_nbclust(PH, kmeans, method = "wss") # Determining the optimal number for
K
```




```
fviz_nbclust(PH, kmeans, method = "silhouette") # # Determining the optimal
number for K
```




```
PH <- scale(PH)
distance <- get_dist(PH)
fviz_dist(distance)
```



```r
k5 <- kmeans(PH, centers = 5, nstart = 25) # chose K=5 based on the
fviz_nbclust graph above

centers <- k5$centers

k5$size

## [1] 2 4 8 3 4

k5$cluster[120]

## [1] NA

fviz_cluster(k5, data = PH)
```

A. To properly cluster the 21 firms based on the variables given the clustering algorithim used was the Euclidean distance as it is the defualt distance metric used and becuase of the simplicity it offers. Additionally, since there were no major outliers in these variables there was no concern for senesitivity issues to them. Based on the fviz_nbclust graph above this is how the amount of clusters (k=5) was determined. Variables that likely had the most impact include market cap and revenue growth.

B. Cluster 4 has the most variability within it due to having 8 firms as part of it with ranging values in variables such as market cap or net profit margin. Due to cluster 1 having only two firms as part of it there is a variable that has them seperated in terms of distance from the other variables that are closer to the middle of the plotting graph.

C. One pattern that was observed with respect to the variables not used in forming the clusters was that each cluster has at least 1 firm that is not in the U.S. In cluster 1 there is a firm in Canada, cluster 2 has a firm in Germany, cluster 3 has a firm in Ireland and in France, cluster 4 has two firms in the UK and one in Switzerland, and cluster 5 has a firm in the UK.

D. Cluster 1: Low ROA
Cluster 2: Low Market Capitalization
Cluster 3: High Estimated Revenue Growth
Cluster 4: High Net Profit Margin
Cluster 5: High Market Capitalization