Assignment_4 - Gabriel Arsego

Garsego

2025-10-21

A.

Selecting only the numerical variables:

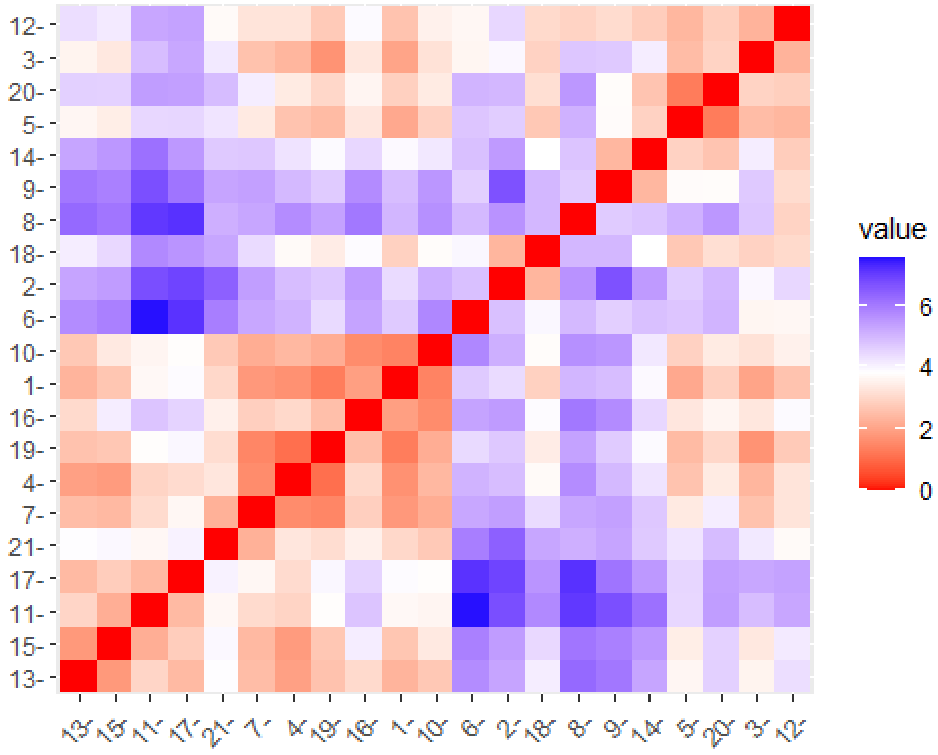
```r
pharma <- data[ ,c(3:11)]
```

Here I normalize the data using scale(). It is necessary to do that when doing a cluster analysis since we don't want a variable to weight more on the analysis on detriment of others. By doing that, all variables have the same scale which eliminates that problem:

```r
pharma_scale <- scale(pharma)
```

Using fviz_dist() to create a visual representation of the normalized distance between the observations. What I'm looking for in this graph is to identify possible groups and have an idea of what a good k value would be for this data. That is also what I'm trying to achieve with the following graphs after this:

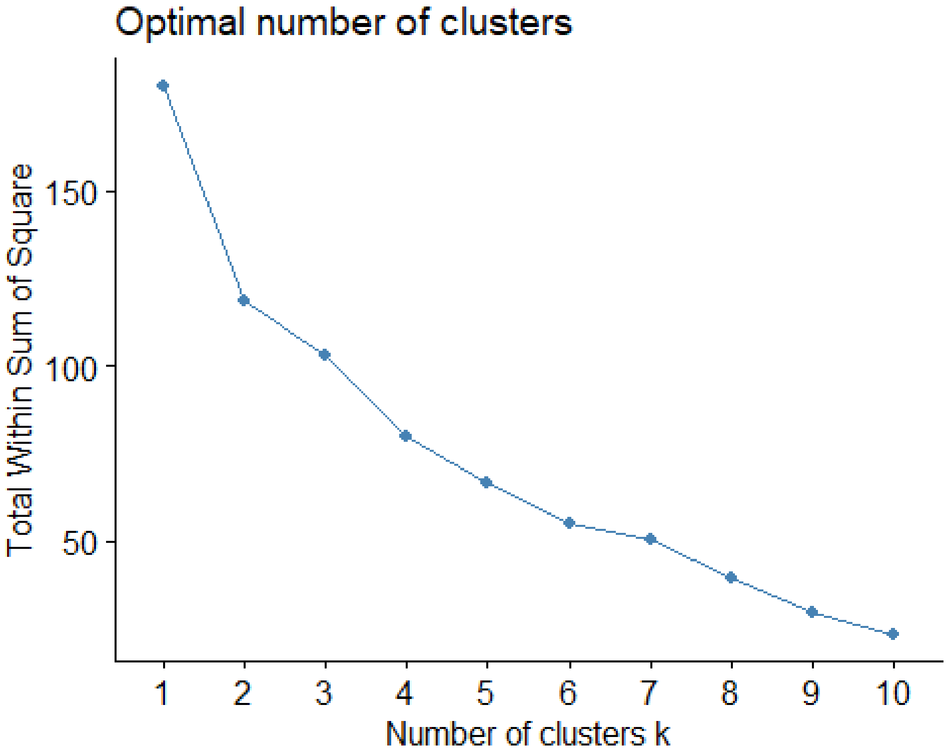
```r
distance <- get_dist(pharma_scale)
fviz_dist(distance)
```



```
#Considering the red areas of this graph, I could argue for k = 4 or k = 5.
More testing needs to be done.
```

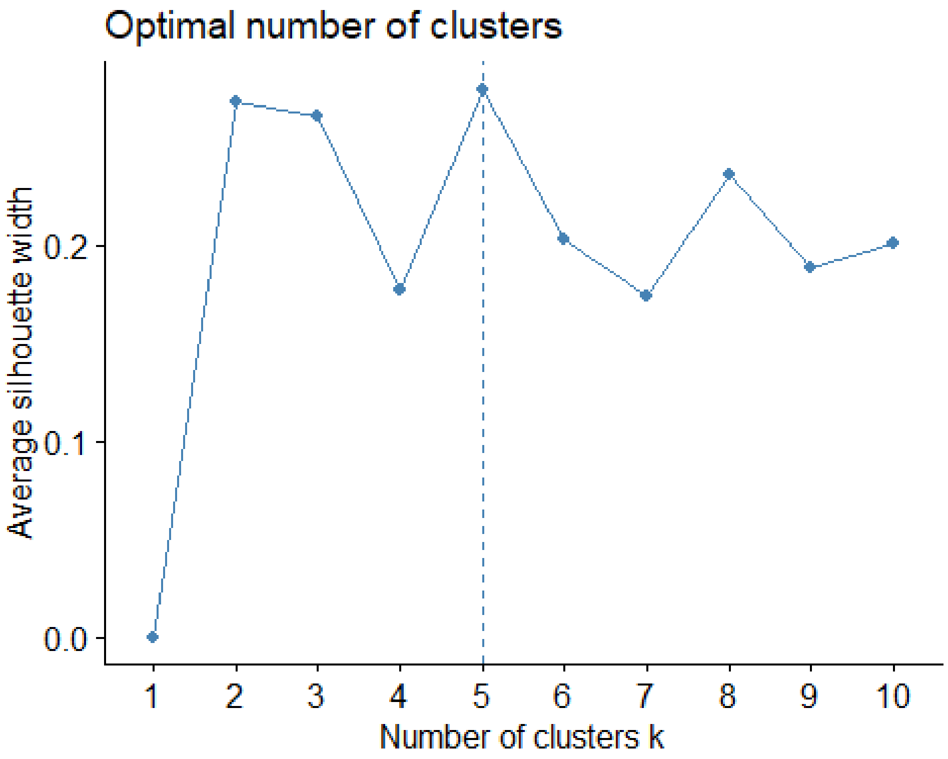
Using fviz_nbclust() to vizualise and determine the optimal number of clusters:

```
#First using the Within-Cluster Sum of Squares method:
fviz_nbclust(pharma_scale, kmeans, method = "wss")
```



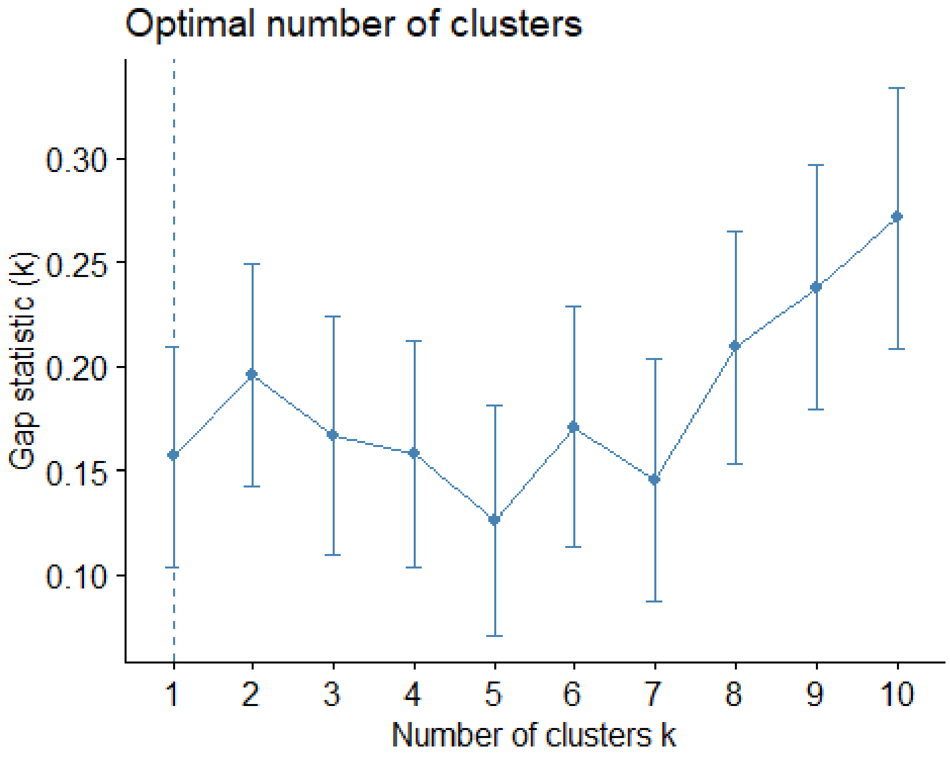
```
#This is also known as the "elbow" method, since we look for the point where
the drop slows down forming an elbow-like structure in the graph.
#In this graph specifically the conclusions we get are not precise enough,
since the drop in wss doesn't slow down in a easily visible way, not forming
an "elbow" structure. I could argue for k = 6, but I need to use more methods
to make it more consistent.
```

```r
#Then using the Silhouette method
fviz_nbclust(pharma_scale, kmeans, method = "silhouette")
```



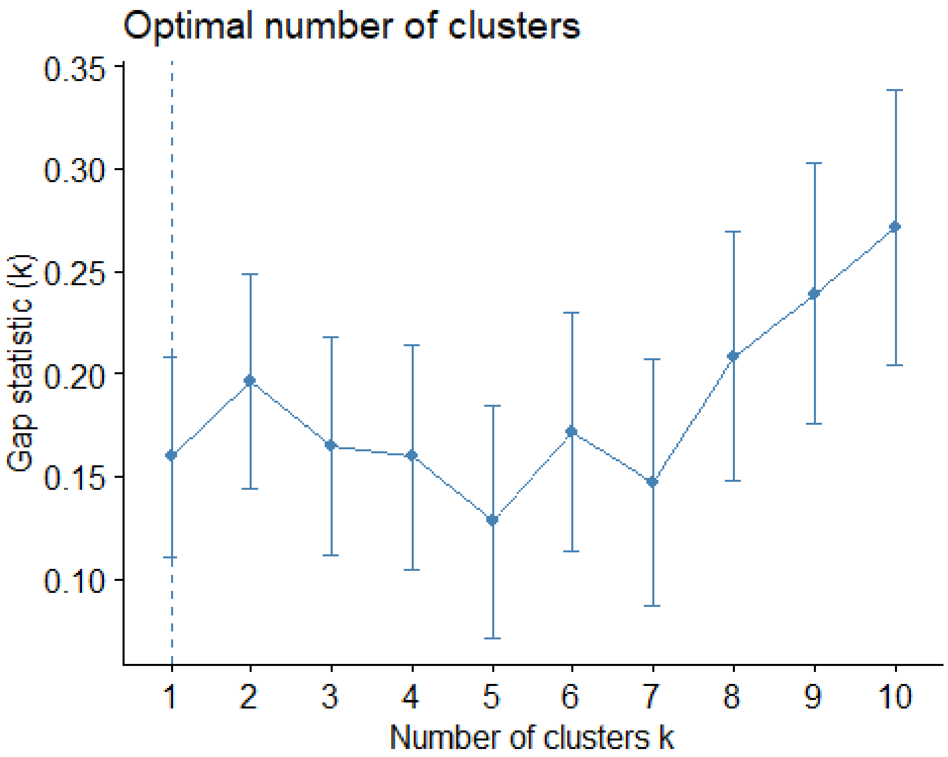
```r
#Here the graph indicates 5 as the optimal number of clusters, which would be
between 4 and 6 which we got from the previous graphs.
```

```r
fviz_nbclust(pharma_scale, kmeans, method = "gap")
```



```r
#When using the gap method, it suggests a k value of 1, which completely
disagrees with the previous methods. Considering that the gap method requires
more stability or more information, the k = 1 value is probably due to
uncertainty.
```

```
fviz_nbclust(pharma_scale, kmeans, method = "gap", nboot = 1000)
```



```
#Here I tried increasing the number of bootstrap reference samples to see if
it would change the result, but it stayed the same.
```

Since all the previous methods suggested a range from 4 to 6, I decided to go with k = 5 for clustering the data.

```
k5 <- kmeans(pharma_scale, centers = 5, nstart = 25)
```

This shows the size of each cluster:

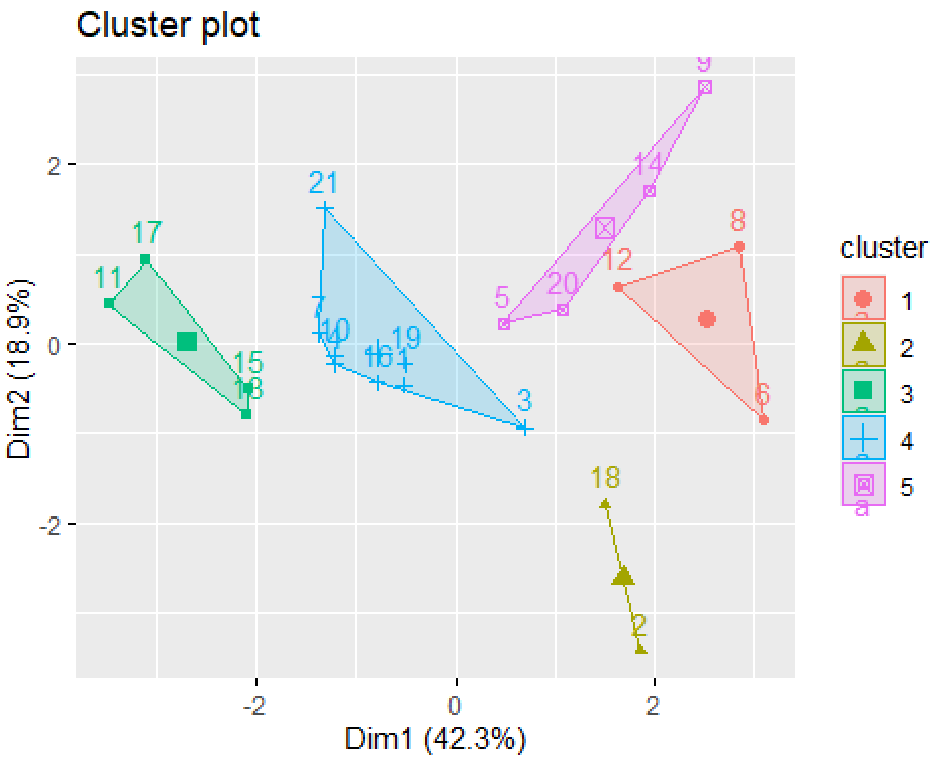
```
k5$size

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

B.

When analyzing the following Cluster plot, it is fair to say that 5 was a good choice for k.

```
fviz_cluster(k5, data = pharma_scale)
```



Here I’m showing k5$cluster which is the the vector of cluster labels created when using kmeans().

```
k5$cluster

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

pharma$cluster <- k5$cluster
#For easier visualization I added "k5$cluster" to the pharma data to create
an extra column called cluster.
table(pharma$cluster)

##
## 1 2 3 4 5
## 3 2 4 8 4

#By using table() I added the new variable "pharma$cluster" to the pharma
dataset and grouped rows by pharma$cluster, so that it has 5 rows as the
number of clusters, and shows the means of each column for each cluster.
```

```
round(aggregate(pharma, list(cluster = pharma$cluster), mean), 2)
```

```
##   cluster Market_Cap Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
## 1       1       6.64 0.87    24.60 16.47  4.17           0.60     1.65
## 2       2      31.91 0.41    69.50 13.20  5.60           0.75     0.48
## 3       3     157.02 0.48    22.23 44.42 17.70           0.95     0.22
## 4       4      55.81 0.41    20.29 28.74 12.69           0.74     0.37
## 5       5      13.10 0.60    17.68 14.57  6.20           0.42     0.64
##   Rev_Growth Net_Profit_Margin cluster
## 1       5.73              7.03       1
## 2      12.08              6.40       2
## 3      18.53             19.58       3
## 4       5.59             19.35       4
## 5      30.14             15.65       5
```

This table showing the means for each variable of each cluster gives us a good understanding about the clustering process. There are clear differences between each cluster when it comes to Market Capitalization and that is also true for most of the other variables.

With ROE and ROA there isn't a big difference from clusters 1, 4 and 5, but if we pair that with other variables like Net Profit Margin and Revenue Growth, we can see patterns forming and those can provide a good idea about the type of pharmaceutical companies we are looking at.

For Cluster 1 we see the smaller companies, that have the lowest Market Capitalization, they also offer a good return on investment while also being the riskier ones to invest in considering their Beta and Leverage variables.

The companies with the highest Price/Earnings Ratio are in Cluster 2, so even though their Market Capitalization is the third lowest one, they are perceived as high growth companies or companies that offer lower risk on investment. They also have the third largest Revenue Growth out of all the clusters.

On Cluster 3 we see a huge spike in Market Capitalization with a mean difference of over 100 billion dollars in comparison with Cluster 4, which has the second largest Market Capitalization. These are somewhat safe companies to invest in, and they are well established in the market.

Cluster 4 has the lowest Revenue Growth of all other clusters. They are behind the smallest companies, from Cluster 1 while being almost 50 billion dollars larger on Market Capitalization on average.

Cluster 5 has the largest Revenue Growth while being second lowest on Market Capitalization. They also have a good Net Profit Margin when compared to other clusters.

C.

When looking at the Exchange variable there is almost no variation since only 2 of the companies don't have their shares listed on NYSE. When it comes to Location most of the companies are in the US so most clusters have a maximum of one company from a different location, except Cluster 5 which has 2 companies. Therefore, there aren't many conclusions we could draw from Location.

In the Median Recommendation variable is where we see more changes. For Cluster 1 we see that the recommendation is to "Hold" or "Moderately buy", which makes sense with our previous analysis where we have Cluster 1 as having smaller companies that are riskier to invest in, but that could offer good potential for growth. For Cluster 2 and 3 the recommendations are similar to Cluster 1.

On Cluster 4 we see the biggest change in pattern and a large variation within the cluster, leading to not much information. Some companies are a "Strong Buy", and some are a "Moderate Sell". What could be said is that for half of the companies in Cluster 4 investors should "Hold" their shares. Cluster 5 has 2 companies with "Moderate Buy" and 2 with "Moderate Sell" which don't tell us much either if we are considering clusters.

D.

Cluster 1: Small and Risky

Cluster 2: Low Risk

Cluster 3: Big Pharma

Cluster 4: Lowest Growth

Cluster 5: Highest Growth