Assignment 04

Jovan Zivak

2025-10-26

```r
# Standard set-up: libraries and data import
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v lubridate 1.9.4     v tibble    3.3.0
## v purrr     1.1.0     v tidyr     1.3.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

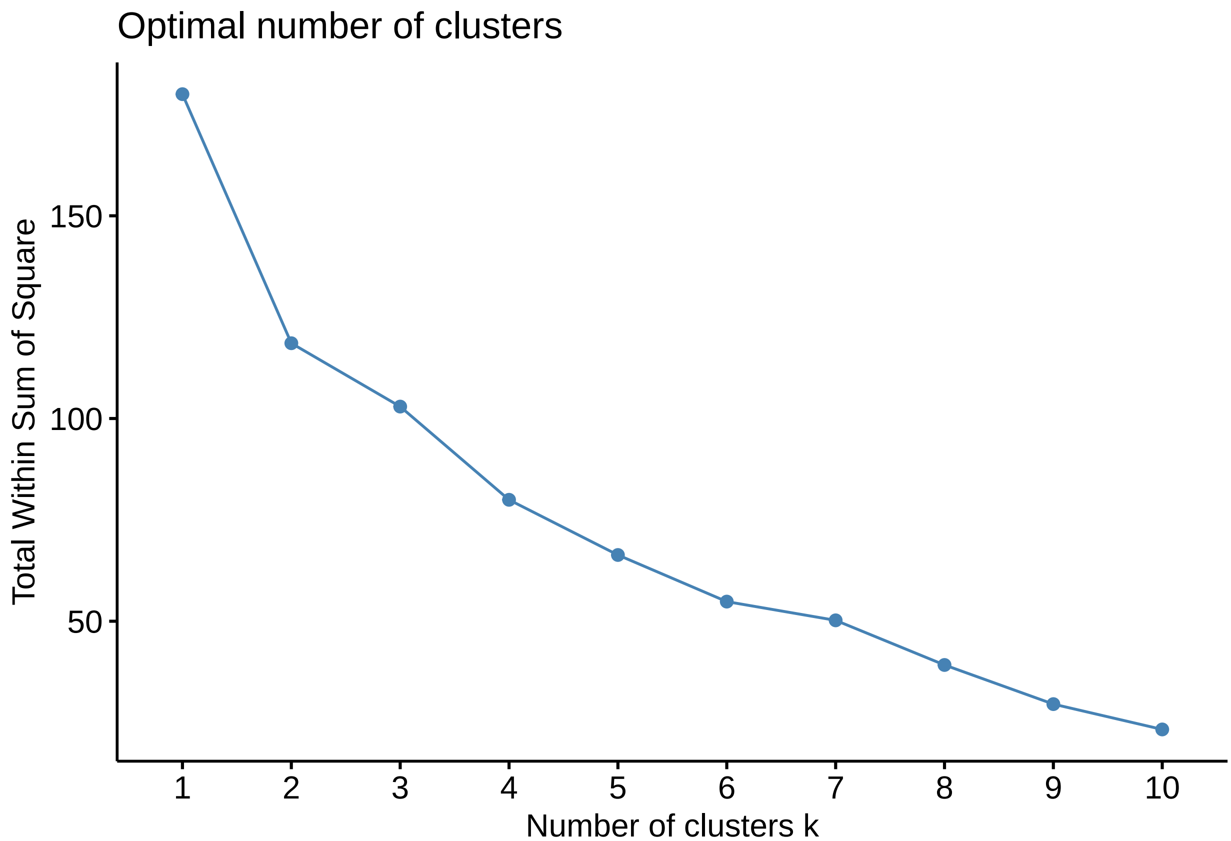
```r
library(factoextra)
```

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

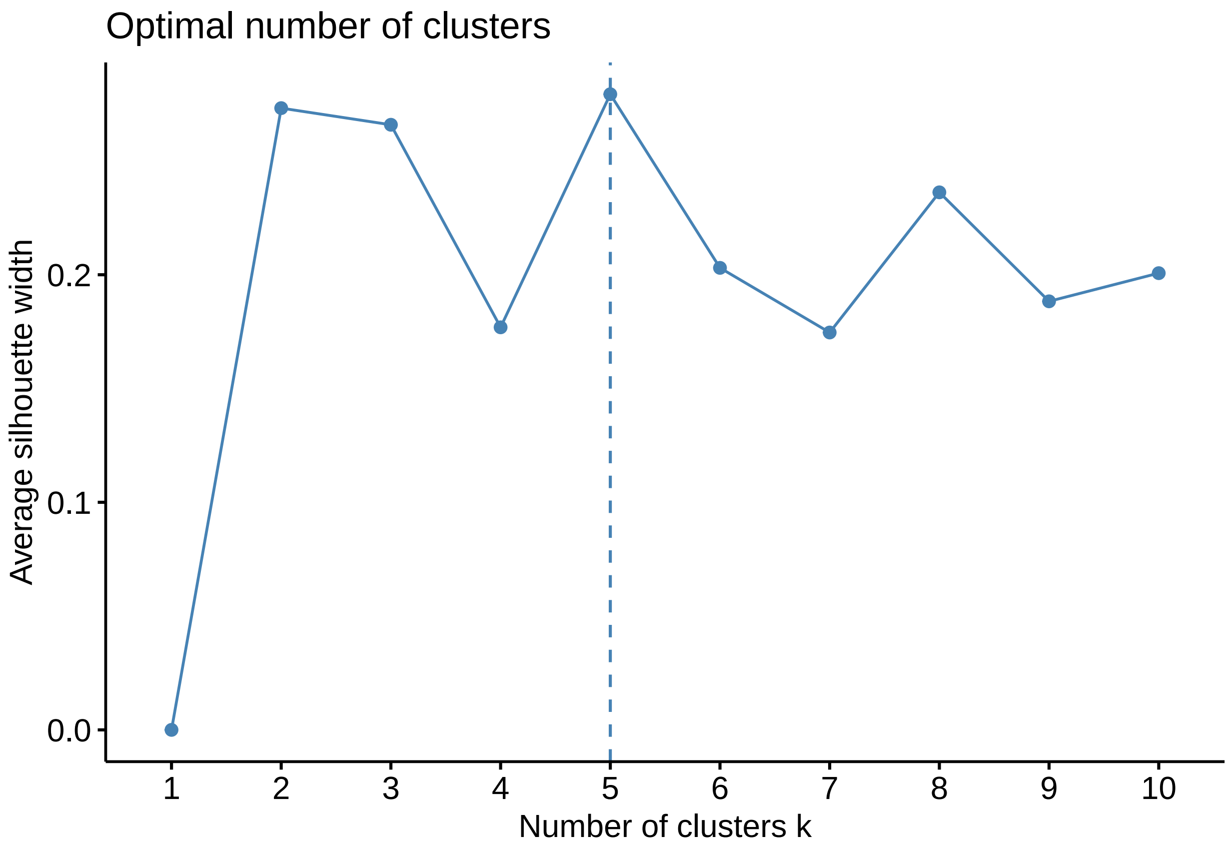
```r
library(flexclust)

library(readxl)
Pharmaceuticals <- read_excel("C:/Users/jovan/Downloads/Pharmaceuticals.xlsx")
#View(Pharmaceuticals)

PharmaClusterData <- Pharmaceuticals[, c(3:11)]
```

```r
# Data normalization
PharmaScaled <- scale(PharmaClusterData)
```

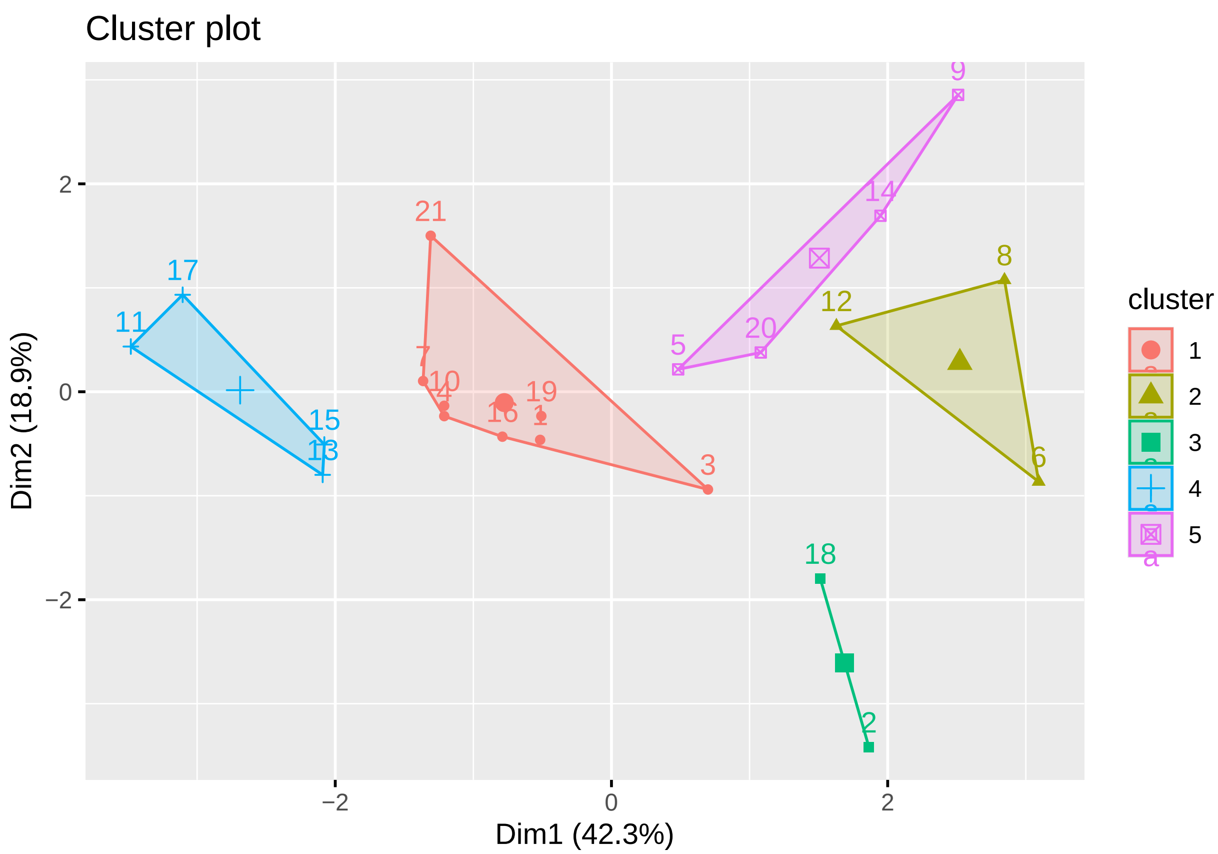
```r
# Cluster optimization via wss method
fviz_nbclust(PharmaScaled, kmeans, method = "wss")
```




```
# Cluster optimization via silhouette method for comparison
fviz_nbclust(PharmaScaled, kmeans, method = "silhouette")
```




```
set.seed(123)
kmeans_result <- kmeans(PharmaScaled, centers = 5, nstart = 25)
fviz_cluster(kmeans_result, data = PharmaScaled)
```




```
# Responses and justification

#A) Z-score was used to normalize data due to the variation in value
# multiple orders of magnitude. After running both WSS and silhouette,
# 5 clusters were determined to be the optimal quantity. K-means itself was
# used because of the relative regularity of the data, falling into distinct
# clusters.

#B) The five-cluster solution reveals clear segmentation among the 21
# pharmaceutical firms. Each polygon in the two-dimensional projection
# represents a financially distinct group. Cluster 4 (blue) contains
# closely related, low-risk firms with similar balance-sheet strength,
# while Cluster 2 (olive) stands out for its high-growth or high-profit
# profile. Cluster 3 (teal) is composed of a few small, specialized companies
# with distinctive financial ratios, and Cluster 1 (red) represents
# mid-performing firms with average profitability and size. Cluster 5 (purple)
# displays greater internal variability, suggesting firms with mixed strategies
# or transitional performance levels.
```

```
#C) Patterns were evident when comparing the clusters against qualitative firm
# characteristics. The most profitable and high-growth clusters
# (e.g., Cluster 2) tended to include firms with Buy or Strong Buy
# recommendations and were primarily listed on the NYSE, while smaller or
# lower-performing clusters had more diverse recommendations and exchange
# listings. This suggests that the financial performance-based clustering is
# consistent with analyst sentiment and firm size as reflected in exchange and
```

```
# location.
```

```
#D) Cluster 1: Average Performers
#   Cluster 2: Fast-Growing Companies
#   Cluster 3: Large, Stable Firms
#   Cluster 4: Highly-Borrowing Firms
#   Cluster 5: Smaller or Unpredictable Firms
```