Final Exam

2025-07-01

# Load required packages  
if (!require("cluster")) install.packages("cluster", dependencies = TRUE)

## Loading required package: cluster

## Warning: package 'cluster' was built under R version 4.4.3

if (!require("factoextra")) install.packages("factoextra", dependencies = TRUE)

## Loading required package: factoextra

## Warning: package 'factoextra' was built under R version 4.4.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.4.3

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

if (!require("readr")) install.packages("readr", dependencies = TRUE)

## Loading required package: readr

if (!require("dplyr")) install.packages("dplyr", dependencies = TRUE)

## Loading required package: dplyr

## Warning: package 'dplyr' was built under R version 4.4.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

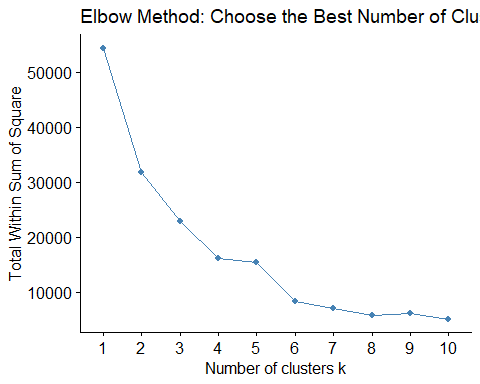
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(cluster)  
library(factoextra)  
library(readr)  
library(dplyr)  
  
# Load the CSV  
fuel\_df <- read\_csv("C:/Users/arkha/Downloads/core\_eia923\_\_monthly\_fuel\_receipts\_costs.csv")

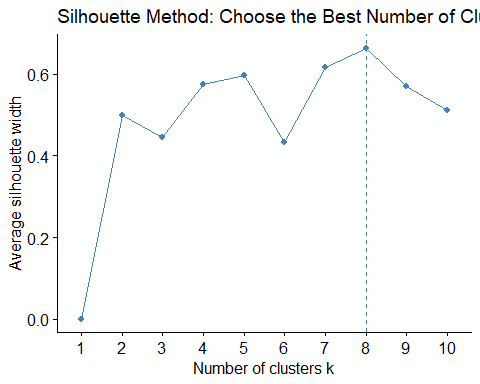
## Rows: 724337 Columns: 22

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (15): contract\_type\_code, energy\_source\_code, fuel\_type\_code\_pudl, fuel...  
## dbl (5): plant\_id\_eia, fuel\_received\_units, fuel\_mmbtu\_per\_unit, sulfur\_co...  
## dttm (2): report\_date, contract\_expiration\_date  
##   
## 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.

# Drop mostly empty or irrelevant columns based on the sample  
fuel\_df\_cleaned <- fuel\_df %>%  
 select(  
 -mercury\_content\_ppm,  
 -secondary\_transportation\_mode\_code,  
 -natural\_gas\_transport\_code,  
 -natural\_gas\_delivery\_contract\_type\_code,  
 -moisture\_content\_pct,  
 -chlorine\_content\_ppm  
 )  
  
# Convert date and character fields  
fuel\_df\_cleaned <- fuel\_df\_cleaned %>%  
 mutate(  
 report\_date = as.Date(report\_date),  
 contract\_expiration\_date = as.Date(contract\_expiration\_date)  
 ) %>%  
 mutate(across(where(is.character), as.factor))  
  
# Sample 2% of the data for performance  
set.seed(1234)  
fuel\_sample <- fuel\_df\_cleaned %>% sample\_frac(0.02)  
  
# Split into training and testing sets (75% training)  
train\_indices <- sample(seq\_len(nrow(fuel\_sample)), size = floor(0.75 \* nrow(fuel\_sample)))  
train\_data <- fuel\_sample[train\_indices, ]  
test\_data <- fuel\_sample[-train\_indices, ]  
  
# Extract numeric columns, remove rows with NA  
train\_numeric <- train\_data %>%  
 select(where(is.numeric)) %>%  
 na.omit()  
  
# Scale numeric data  
train\_scaled <- scale(train\_numeric)  
  
# --- Choose Optimal k ---  
# Elbow method  
fviz\_nbclust(train\_scaled, kmeans, method = "wss") +  
 labs(title = "Elbow Method: Choose the Best Number of Clusters")



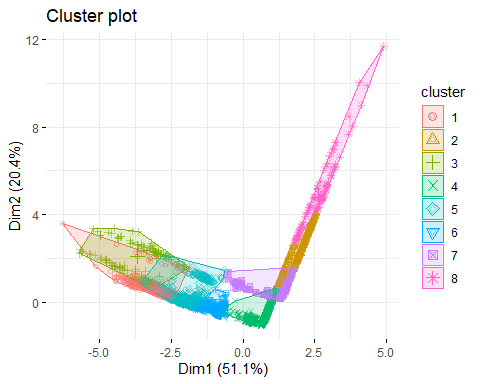
# Silhouette method  
fviz\_nbclust(train\_scaled, kmeans, method = "silhouette") +  
 labs(title = "Silhouette Method: Choose the Best Number of Clusters")



# --- Fit K-means clustering ---  
set.seed(1234)  
k <- 8 # Update this based on previous plot results  
kmeans\_result <- kmeans(train\_scaled, centers = k, nstart = 25)  
  
# Assign cluster labels back to data  
train\_clustered <- train\_data %>%  
 filter(complete.cases(select(., where(is.numeric)))) %>%  
 mutate(cluster = as.factor(kmeans\_result$cluster))  
  
# Cluster summary  
cluster\_summary <- train\_clustered %>%  
 group\_by(cluster) %>%  
 summarise(across(where(is.numeric), list(mean = mean, sd = sd), .names = "{col}\_{fn}"))  
  
print(cluster\_summary)

## # A tibble: 8 × 11  
## cluster plant\_id\_eia\_mean plant\_id\_eia\_sd fuel\_received\_units\_mean  
## <fct> <dbl> <dbl> <dbl>  
## 1 1 5800. 9062. 37478.  
## 2 2 44144. 21249. 2307727.  
## 3 3 34711. 20826. 29383.  
## 4 4 4450. 3027. 123610.  
## 5 5 8026. 13499. 33346.  
## 6 6 3963. 2740. 66382.  
## 7 7 55201. 2019. 194303.  
## 8 8 28394. 26599. 5323978.  
## # ℹ 7 more variables: fuel\_received\_units\_sd <dbl>,  
## # fuel\_mmbtu\_per\_unit\_mean <dbl>, fuel\_mmbtu\_per\_unit\_sd <dbl>,  
## # sulfur\_content\_pct\_mean <dbl>, sulfur\_content\_pct\_sd <dbl>,  
## # ash\_content\_pct\_mean <dbl>, ash\_content\_pct\_sd <dbl>

# --- Cluster plot (cleaned) ---  
fviz\_cluster(kmeans\_result, data = train\_scaled,  
 ellipse.type = "convex",  
 show.clust.cent = TRUE,  
 geom = "point",  
 pointsize = 1.5,  
 alpha = 0.5,  
 repel = TRUE,  
 labelsize = 0,  
 ggtheme = theme\_minimal())

  
  
  
We analyzed monthly fuel receipts and costs data from the EIA923 dataset to uncover patterns among U.S. power plants. After cleaning the data by removing columns with too many missing values and converting character fields to factors, we randomly sampled 2% of the dataset for efficiency and focused on numeric variables. We scaled the data and used K-means clustering to group similar facilities. In order to find the best number of clusters, we utilized both the Elbow Method and the Silhouette Method. Conversely, Elbow method suggested a k value around 6, and the Silhouette Method, who's k value peaked at 8. Based on this, we chose k = 8 and found that each cluster reflected meaningful differences—some included very large plants receiving millions of fuel units, while others represented smaller, more consistent operations. Our findings show that clustering can reveal useful insights about how fuel is used across the country, helping inform decisions about energy planning, efficiency, and policy.