Final

2025-07-02

library(dplyr)

##   
## Attaching package: 'dplyr'

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

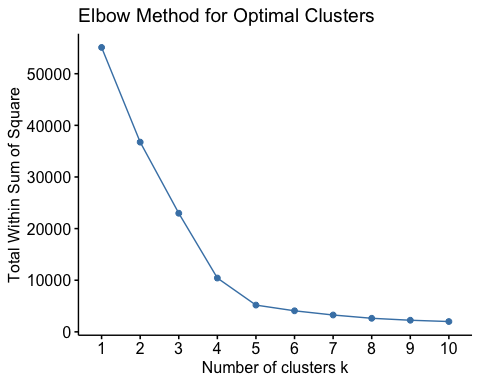
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
library(cluster)  
library(factoextra)

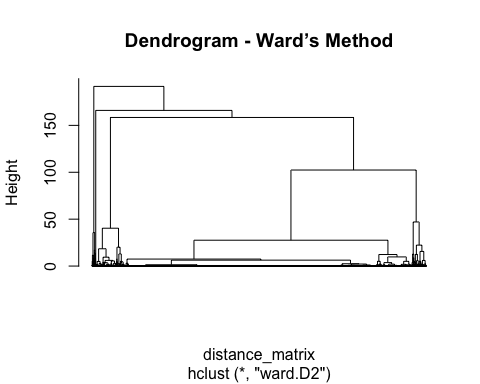
## Loading required package: ggplot2

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

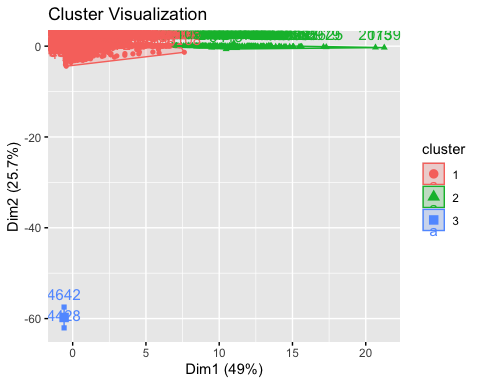
library(ggplot2)  
  
# Loading the data  
file\_path <- "EIA923\_Schedules\_2\_3\_4\_5\_M\_12\_2023\_Final\_5.csv" # Update with your actual file name  
raw\_data <- read.csv(file\_path, stringsAsFactors = FALSE, check.names = FALSE)  
  
# Cleaning and selecting data  
ytd\_data <- raw\_data %>%  
 select(  
 Total\_Fuel\_Consumption\_Quantity = `Total\_Fuel\_Consumption\_Quantity`,  
 Electric\_Fuel\_Consumption\_Quantity = `Electric\_Fuel\_Consumption\_Quantity`,  
 Total\_Fuel\_Consumption\_MMBtu = `Total\_Fuel\_Consumption\_MMBtu`,  
 Elec\_Fuel\_Consumption\_MMBtu = `Elec\_Fuel\_Consumption\_MMBtu`,  
 Net\_Generation = `Net\_Generation`,  
 YEAR = `YEAR`  
 )  
  
ytd\_data <- ytd\_data %>%  
 mutate(  
 across(  
 c(  
 Total\_Fuel\_Consumption\_Quantity,  
 Electric\_Fuel\_Consumption\_Quantity,  
 Total\_Fuel\_Consumption\_MMBtu,  
 Elec\_Fuel\_Consumption\_MMBtu,  
 Net\_Generation  
 ),  
 ~ as.numeric(gsub(",", "", .))  
 )  
 )  
  
ytd\_data <- ytd\_data %>%  
 drop\_na() %>%  
 filter(Net\_Generation > 0, Total\_Fuel\_Consumption\_MMBtu > 0)  
  
# Creating variables  
ytd\_data <- ytd\_data %>%  
 mutate(  
 Fuel\_Efficiency\_MMBtu\_per\_MWh = Total\_Fuel\_Consumption\_MMBtu / Net\_Generation,  
 Elec\_Fuel\_Share = Elec\_Fuel\_Consumption\_MMBtu / Total\_Fuel\_Consumption\_MMBtu  
 ) %>%  
 filter\_all(all\_vars(!is.infinite(.)))  
  
# clustering  
clustering\_vars <- ytd\_data %>%  
 select(  
 Total\_Fuel\_Consumption\_MMBtu,  
 Net\_Generation,  
 Fuel\_Efficiency\_MMBtu\_per\_MWh,  
 Elec\_Fuel\_Share  
 )  
  
# scaling data  
clustering\_scaled <- scale(clustering\_vars)  
  
# determining K  
fviz\_nbclust(clustering\_scaled, FUN = hcut, method = "wss") +  
 ggtitle("Elbow Method for Optimal Clusters")



distance\_matrix <- dist(clustering\_scaled)  
hc\_ward <- hclust(distance\_matrix, method = "ward.D2")  
plot(hc\_ward, labels = FALSE, hang = -1, main = "Dendrogram - Ward’s Method")



# clustering  
k <- 3 # Adjust as needed based on elbow plot  
clusters <- cutree(hc\_ward, k = k)  
  
  
clustered\_data <- ytd\_data %>%  
 mutate(Cluster = as.factor(clusters))  
  
# Viewing the clusters  
fviz\_cluster(list(data = clustering\_scaled, cluster = clusters)) +  
 ggtitle("Cluster Visualization")



# Cluster summary  
cluster\_summary <- clustered\_data %>%  
 group\_by(Cluster) %>%  
 summarise(  
 Avg\_Total\_MMBtu = mean(Total\_Fuel\_Consumption\_MMBtu),  
 Avg\_Net\_Generation = mean(Net\_Generation),  
 Avg\_Efficiency = mean(Fuel\_Efficiency\_MMBtu\_per\_MWh),  
 Avg\_Elec\_Share = mean(Elec\_Fuel\_Share),  
 Count = n()  
 )  
  
print(cluster\_summary)

## # A tibble: 3 × 6  
## Cluster Avg\_Total\_MMBtu Avg\_Net\_Generation Avg\_Efficiency Avg\_Elec\_Share Count  
## <fct> <dbl> <dbl> <dbl> <dbl> <int>  
## 1 1 1633546. 202619. 13.8 0.936 13622  
## 2 2 87043455. 8555703. 10.3 0.998 150  
## 3 3 2248606. 4 559364. 0.500 2