FML\_Final exam

2023-12-07

#The dataset provided is comprised of a total of 12 variables, but for our specific research, we have focused on three key variables – sulfur content, ash content, and fuel cost. Upon conducting our analysis. #we considered both ward and complete linkages for our clustering method. After careful consideration of the results, we have ultimately chosen to use the ward.D linkage approach. This decision was based on evaluating the silhouette scores, which showed that the ward.D linkage yielded a higher evaluation score of .58, compared to the complete linkage method, which had a lower score of .41. Our decision to use the ward.D linkage was also supported by its ability to minimize the total within-cluster variance. This in turn allowed us to create more homogenous clusters with distinct differences between them, which is a key factor in producing ideal clusters.Upon careful examination of the dendograms, it was observed that ward linkage resulted in evenly sized clusters, while complete linkage produced disproportionately large clusters and numerous smaller ones. As such, it can be concluded that ward linkage is preferred due to its ability to generate homogeneous groups, supported by a higher silhouette score, making it the superior choice. In order to determine the optimum number of clusters to be formed, an elbow plot was utilized. After careful analysis, it was determined that the ideal number of clusters is 4. The elbow plot, depicting the within-groups sum of squares against the number of clusters, displayed a distinct elbow at 4 clusters (marked by a blue dotted line), indicating that increasing the number of clusters beyond 4 would not lead to a significant decrease in within-cluster variation. Additionally,We performed calculations to determine the silhoette scores for both ward and complete linkages at a K value of 4. The results were .58 and .41 respectively, indicating strong cluster cohesion and separation. Visually, the dendograms also displayed 4 distinct clusters, avoiding too many low-value clusters. The first cluster (n=34) had low sulfur and ash content, with fuel costs in the mid-range. The low sulfur content was due to a tight distribution, centered around -.4. Similarly, the ash content also had a low tight distribution, with values ranging from 0.5 to 1.0. The second cluster (n=8) had very high fuel costs, mid-range sulfur content, and low ash content.

library(hclust1d)

## Warning: package 'hclust1d' was built under R version 4.3.2

## Loaded hclust1d version 0.1.1

## To acknowledge my work please cite hclust1d in publications as:

## Nowakowski, Szymon (2023). hclust1d: Hierarchical Clustering of Univariate (1d) Data. https://cran.r-project.org/package=hclust1d

library(ggplot2)

#Importing the dataset  
fuel<- read.csv("fuel\_receipts1.csv")  
summary(fuel)

## X rowid plant\_id\_eia energy\_source\_code  
## Min. : 1.0 Min. : 1.0 Min. : 3.0 Length:756   
## 1st Qu.:189.8 1st Qu.: 221.8 1st Qu.: 527.0 Class :character   
## Median :378.5 Median : 450.5 Median : 728.0 Mode :character   
## Mean :378.5 Mean : 465.7 Mean : 808.7   
## 3rd Qu.:567.2 3rd Qu.: 704.2 3rd Qu.:1252.0   
## Max. :756.0 Max. :1000.0 Max. :1710.0   
## fuel\_type\_code\_pudl fuel\_group\_code supplier\_name fuel\_received\_units  
## Length:756 Length:756 Length:756 Min. : 1   
## Class :character Class :character Class :character 1st Qu.: 1300   
## Mode :character Mode :character Mode :character Median : 12570   
## Mean : 89258   
## 3rd Qu.: 42025   
## Max. :4823176   
## fuel\_mmbtu\_per\_unit sulfur\_content\_pct ash\_content\_pct fuel\_cost\_per\_mmbtu  
## Min. : 0.857 Min. :0.0000 Min. : 0.000 Min. : 0.343   
## 1st Qu.: 1.028 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.: 1.947   
## Median :17.141 Median :0.3200 Median : 5.000 Median : 3.082   
## Mean :13.349 Mean :0.7314 Mean : 4.995 Mean : 5.786   
## 3rd Qu.:23.870 3rd Qu.:0.9400 3rd Qu.: 9.600 3rd Qu.: 8.400   
## Max. :29.400 Max. :6.6100 Max. :20.900 Max. :29.514

# Displaying the first few rows of the data  
head(fuel)

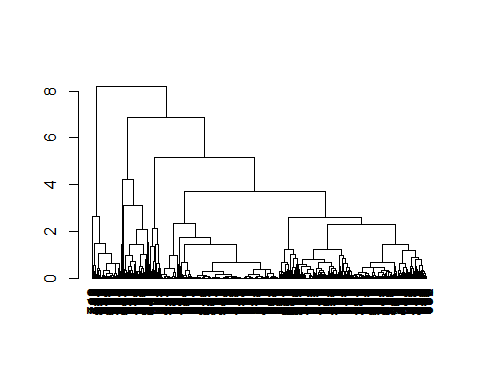
## X rowid plant\_id\_eia energy\_source\_code fuel\_type\_code\_pudl fuel\_group\_code  
## 1 1 1 3 BIT coal coal  
## 2 2 2 3 BIT coal coal  
## 3 3 3 3 NG gas natural\_gas  
## 4 4 4 7 BIT coal coal  
## 5 5 5 7 BIT coal coal  
## 6 6 6 7 BIT coal coal  
## supplier\_name fuel\_received\_units fuel\_mmbtu\_per\_unit sulfur\_content\_pct  
## 1 interocean coal 259412 23.100 0.49  
## 2 interocean coal 52241 22.800 0.48  
## 3 bay gas pipeline 2783619 1.039 0.00  
## 4 alabama coal 25397 24.610 1.69  
## 5 d & e mining 764 24.446 0.84  
## 6 alabama coal 603 24.577 1.54  
## ash\_content\_pct fuel\_cost\_per\_mmbtu  
## 1 5.4 2.135  
## 2 5.7 2.115  
## 3 0.0 8.631  
## 4 14.7 2.776  
## 5 15.5 3.381  
## 6 14.6 2.199

# From the data taking only the required variables sulfur\_content,ash\_content and fuel\_content  
fuel\_1 <- fuel[,c(10:12)]  
fuel\_2<-scale(fuel\_1)

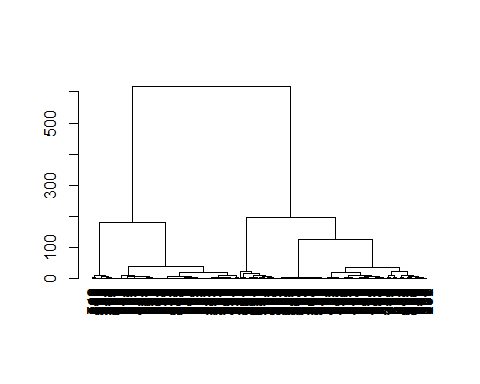
Hierarchical\_clustering <- hclust(dist(fuel\_2), method = "complete") #using the complete linkage method.

Hierarchical\_clustering\_1 <- hclust(dist(fuel\_2), method = "ward.D") # using an other Ward.D linkage

dendrogram <- as.dendrogram(Hierarchical\_clustering)  
plot(dendrogram)



dendrogram <- as.dendrogram(Hierarchical\_clustering\_1)  
plot(dendrogram)



clusters <- cutree(Hierarchical\_clustering, h=2) #cuttree at the height of 2.  
clusters <- cutree(Hierarchical\_clustering, h=5) #cuttree at the height of 5.

# Calculate total within sum of squares for different number of clusters  
within\_sumof\_squares <- vector()   
for (i in 1:10) within\_sumof\_squares[i] <- sum(kmeans(fuel\_2, centers=i)$within\_sumof\_squares)  
  
# Plot elbow curve   
plot(1:10, within\_sumof\_squares, type="b", xlab="Number of Clusters", ylab="Within sum of squares")  
  
# Add blue dotted line at optimal k  
abline(v=4, lty=2, col="blue")

