FML-Final Project

Chathurani Ekanayake

2023-05-07

install.packages(“missCforest”) install.packages(“fpc”) install.packages(“StatMatch”) install.packages(“magrittr”) install.packages(“missForest”)

library(dplyr) library(caret) library(missCforest) library(corrplot) library(factoextra) library(fpc) library(StatMatch) library(cluster) library(ggplot2) library(cowplot) library(magrittr) library(missForest) library(flexclust) library(ggcorrplot)

load Dataset

```
library(openxlsx)
Energy_Data <- read.xlsx("F:/1st sem/ML/Final Assi/fuel_receipts.xlsx", sheet = 1)
View(Energy_Data)
```

```
# Check the structure of the data set
str(Energy_Data)
```

```
## 'data.frame':    608564 obs. of  30 variables:
##  $ rowid                                  : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ plant_id_eia                           : num  3 3 3 7 7 7 7 8 8 8 ...
##  $ plant_id_eia_label                     : chr  "Barry" "Barry" "Barry" "Gadsden" ...
##  $ report_date                            : num  39448 39448 39448 39448 39448 ...
##  $ contract_type_code                     : chr  "C" "C" "C" "C" ...
##  $ contract_type_code_label               : chr  "C" "C" "C" "C" ...
##  $ contract_expiration_date               : num  39539 39539 NA 42339 39753 ...
##  $ energy_source_code                     : chr  "BIT" "BIT" "NG" "BIT" ...
##  $ energy_source_code_label               : chr  "BIT" "BIT" "NG" "BIT" ...
##  $ fuel_type_code_pudl                    : chr  "coal" "coal" "gas" "coal" ...
##  $ fuel_group_code                        : chr  "coal" "coal" "natural_gas" "coal" ...
##  $ mine_id_pudl                           : num  0 0 NA 1 2 3 NA 4 4 1 ...
##  $ mine_id_pudl_label                     : num  0 0 NA 1 2 3 NA 4 4 1 ...
##  $ supplier_name                          : chr  "interocean coal" "interocean coal" "bay ga
s pipeline" "alabama coal" ...
##  $ fuel_received_units                    : num  259412 52241 2783619 25397 764 ...
##  $ fuel_mmbtu_per_unit                    : num  23.1 22.8 1.04 24.61 24.45 ...
##  $ sulfur_content_pct                     : num  0.49 0.48 0 1.69 0.84 1.54 0 2.16 1.24 1.9
...
##  $ ash_content_pct                        : num  5.4 5.7 0 14.7 15.5 14.6 0 15.4 11.9 15.4
...
##  $ mercury_content_ppm                    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ fuel_cost_per_mmbtu                    : num  2.13 2.12 8.63 2.78 3.38 ...
##  $ primary_transportation_mode_code       : chr  "RV" "RV" "PL" "TR" ...
##  $ primary_transportation_mode_code_label : chr  "RV" "RV" "PL" "TR" ...
##  $ secondary_transportation_mode_code     : chr  NA NA NA NA ...
##  $ secondary_transportation_mode_code_label: chr  NA NA NA NA ...
##  $ natural_gas_transport_code             : chr  "firm" "firm" "firm" "firm" ...
##  $ natural_gas_delivery_contract_type_code : chr  NA NA NA NA ...
##  $ moisture_content_pct                   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ chlorine_content_ppm                   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ data_maturity                          : chr  "final" "final" "final" "final" ...
##  $ data_maturity_label                    : chr  "final" "final" "final" "final" ...
```

```
# Check the summary
summary(Energy_Data)
```

```
##      rowid          plant_id_eia   plant_id_eia_label  report_date
##  Min.   :     1   Min.   :    3   Length:608564      Min.   :39448
##  1st Qu.:152142   1st Qu.: 2712   Class :character   1st Qu.:40360
##  Median :304283   Median : 6155   Mode  :character   Median :41518
##  Mean   :304283   Mean   :18290                      Mean   :41707
##  3rd Qu.:456423   3rd Qu.:50707                      3rd Qu.:43040
##  Max.   :608564   Max.   :64020                      Max.   :44531
##
##  contract_type_code contract_type_code_label contract_expiration_date
##  Length:608564      Length:608564            Min.   :36526
##  Class :character   Class :character         1st Qu.:40878
##  Mode  :character   Mode  :character         Median :42156
##                                              Mean   :42556
##                                              3rd Qu.:43800
##                                              Max.   :73020
##                                              NA's   :344301
##  energy_source_code energy_source_code_label fuel_type_code_pudl
##  Length:608564      Length:608564            Length:608564
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##
##  fuel_group_code     mine_id_pudl    mine_id_pudl_label supplier_name
##  Length:608564      Min.   :   0     Min.   :   0       Length:608564
##  Class :character   1st Qu.:  42     1st Qu.:  42       Class :character
##  Mode  :character   Median : 972     Median : 972       Mode  :character
##                     Mean   :1577     Mean   :1577
##                     3rd Qu.:3121     3rd Qu.:3121
##                     Max.   :4562     Max.   :4562
##                     NA's   :391946   NA's   :391946
##  fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct
##  Min.   :       1    Min.   :   0.000    Min.   : 0.0000    Min.   : 0.000
##  1st Qu.:    3700    1st Qu.:   1.025    1st Qu.: 0.0000    1st Qu.: 0.000
##  Median :   21565    Median :   1.061    Median : 0.0000    Median : 0.000
##  Mean   :  242967    Mean   :   8.839    Mean   : 0.5145    Mean   : 3.606
##  3rd Qu.:  106164    3rd Qu.:  17.809    3rd Qu.: 0.4900    3rd Qu.: 5.800
##  Max.   :48159765    Max.   :1049.000    Max.   :11.0100    Max.   :72.200
##
##  mercury_content_ppm fuel_cost_per_mmbtu primary_transportation_mode_code
##  Min.   :0.00        Min.   :   -71.9    Length:608564
##  1st Qu.:0.00        1st Qu.:     2.3    Class :character
##  Median :0.00        Median :     3.3    Mode  :character
##  Mean   :0.01        Mean   :    14.2
##  3rd Qu.:0.00        3rd Qu.:     4.8
##  Max.   :1.82        Max.   :562572.2
##  NA's   :289482      NA's   :200240
##  primary_transportation_mode_code_label secondary_transportation_mode_code
##  Length:608564                          Length:608564
##  Class :character                       Class :character
##  Mode  :character                       Mode  :character
```

```
##
##
##
##
##  secondary_transportation_mode_code_label natural_gas_transport_code
##  Length:608564                            Length:608564
##  Class :character                         Class :character
##  Mode  :character                         Mode  :character
##
##
##
##
##  natural_gas_delivery_contract_type_code moisture_content_pct
##  Length:608564                           Min.   :  0.0
##  Class :character                        1st Qu.:  6.6
##  Mode  :character                        Median : 11.9
##                                          Mean   : 15.6
##                                          3rd Qu.: 26.8
##                                          Max.   :247.0
##                                          NA's   :516588
##  chlorine_content_ppm data_maturity      data_maturity_label
##  Min.   :   0.0       Length:608564      Length:608564
##  1st Qu.:   0.0       Class :character   Class :character
##  Median :   0.0       Mode  :character   Mode  :character
##  Mean   :  59.2
##  3rd Qu.:   0.0
##  Max.   :3747.0
##  NA's   :516588
```

```
# select relevant variables
Fuel_Data <- Energy_Data[,-c(1,3:5,7,9,10,12:14,21:30)]
```

```
# 01. Remove missing Values
colMeans(is.na(Fuel_Data))
```

```
##             plant_id_eia contract_type_code_label       energy_source_code
##             0.0000000000             0.0003910846             0.0000000000
##          fuel_group_code      fuel_received_units      fuel_mmbtu_per_unit
##             0.0000000000             0.0000000000             0.0000000000
##       sulfur_content_pct          ash_content_pct      mercury_content_ppm
##             0.0000000000             0.0000000000             0.4756804543
##      fuel_cost_per_mmbtu
##             0.3290368803
```

```
Fuel_Data$mercury_content_ppm[is.na(Fuel_Data$mercury_content_ppm)] <- median(Fuel_Data$mercury_
content_ppm,na.rm = T)
Fuel_Data$fuel_cost_per_mmbtu[is.na(Fuel_Data$fuel_cost_per_mmbtu)] <- median(Fuel_Data$fuel_cos
t_per_mmbtu,na.rm = T)
colMeans(is.na(Fuel_Data)) # remove all missing values
```

```
##             plant_id_eia contract_type_code_label       energy_source_code
##             0.0000000000             0.0003910846             0.0000000000
##          fuel_group_code      fuel_received_units      fuel_mmbtu_per_unit
##             0.0000000000             0.0000000000             0.0000000000
##       sulfur_content_pct          ash_content_pct      mercury_content_ppm
##             0.0000000000             0.0000000000             0.0000000000
##      fuel_cost_per_mmbtu
##             0.0000000000
```

```
# 02. Ensure variables in right attribute
summary(Fuel_Data)
```

```
##   plant_id_eia   contract_type_code_label energy_source_code fuel_group_code
##  Min.   :    3   Length:608564            Length:608564      Length:608564
##  1st Qu.: 2712   Class :character         Class :character   Class :character
##  Median : 6155   Mode  :character         Mode  :character   Mode  :character
##  Mean   :18290
##  3rd Qu.:50707
##  Max.   :64020
##  fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct
##  Min.   :       1    Min.   :   0.000    Min.   : 0.0000    Min.   : 0.000
##  1st Qu.:    3700    1st Qu.:   1.025    1st Qu.: 0.0000    1st Qu.: 0.000
##  Median :   21565    Median :   1.061    Median : 0.0000    Median : 0.000
##  Mean   :  242967    Mean   :   8.839    Mean   : 0.5145    Mean   : 3.606
##  3rd Qu.:  106164    3rd Qu.:  17.809    3rd Qu.: 0.4900    3rd Qu.: 5.800
##  Max.   :48159765    Max.   :1049.000    Max.   :11.0100    Max.   :72.200
##  mercury_content_ppm fuel_cost_per_mmbtu
##  Min.   :0.00000     Min.   :   -71.9
##  1st Qu.:0.00000     1st Qu.:     2.7
##  Median :0.00000     Median :     3.3
##  Mean   :0.00422     Mean   :    10.6
##  3rd Qu.:0.00000     3rd Qu.:     3.9
##  Max.   :1.82000     Max.   :562572.2
```

```
str(Fuel_Data)
```

```
## 'data.frame':    608564 obs. of  10 variables:
##  $ plant_id_eia            : num  3 3 3 7 7 7 7 8 8 8 ...
##  $ contract_type_code_label: chr  "C" "C" "C" "C" ...
##  $ energy_source_code      : chr  "BIT" "BIT" "NG" "BIT" ...
##  $ fuel_group_code         : chr  "coal" "coal" "natural_gas" "coal" ...
##  $ fuel_received_units     : num  259412 52241 2783619 25397 764 ...
##  $ fuel_mmbtu_per_unit     : num  23.1 22.8 1.04 24.61 24.45 ...
##  $ sulfur_content_pct      : num  0.49 0.48 0 1.69 0.84 1.54 0 2.16 1.24 1.9 ...
##  $ ash_content_pct         : num  5.4 5.7 0 14.7 15.5 14.6 0 15.4 11.9 15.4 ...
##  $ mercury_content_ppm     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ fuel_cost_per_mmbtu     : num  2.13 2.12 8.63 2.78 3.38 ...
```

#03. To ensure that both the data, and the analysis are unique to each student, randomly sample about 2% of your data using a random 4-digit number as the seed to sample the data. Use 75% of the sampled data as the training set, and the rest as the test set (if needed). This should yield a training set of about 9000 and a test of about 3000.

```
set.seed(1234)
```

```
# randomly sample about 2% of your data
sampled_data <- Fuel_Data[sample(nrow(Fuel_Data), size = round(nrow(Fuel_Data)*0.02)),]

# split the sampled data into training and test sets
train_index <- sample(seq_len(nrow(sampled_data)), size = round(0.75*nrow(sampled_data)))
train_data <- sampled_data[train_index, ]
test_data <- sampled_data[-train_index, ]
```

```
# normalize data
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(caret)
```

```
## Loading required package: ggplot2
```

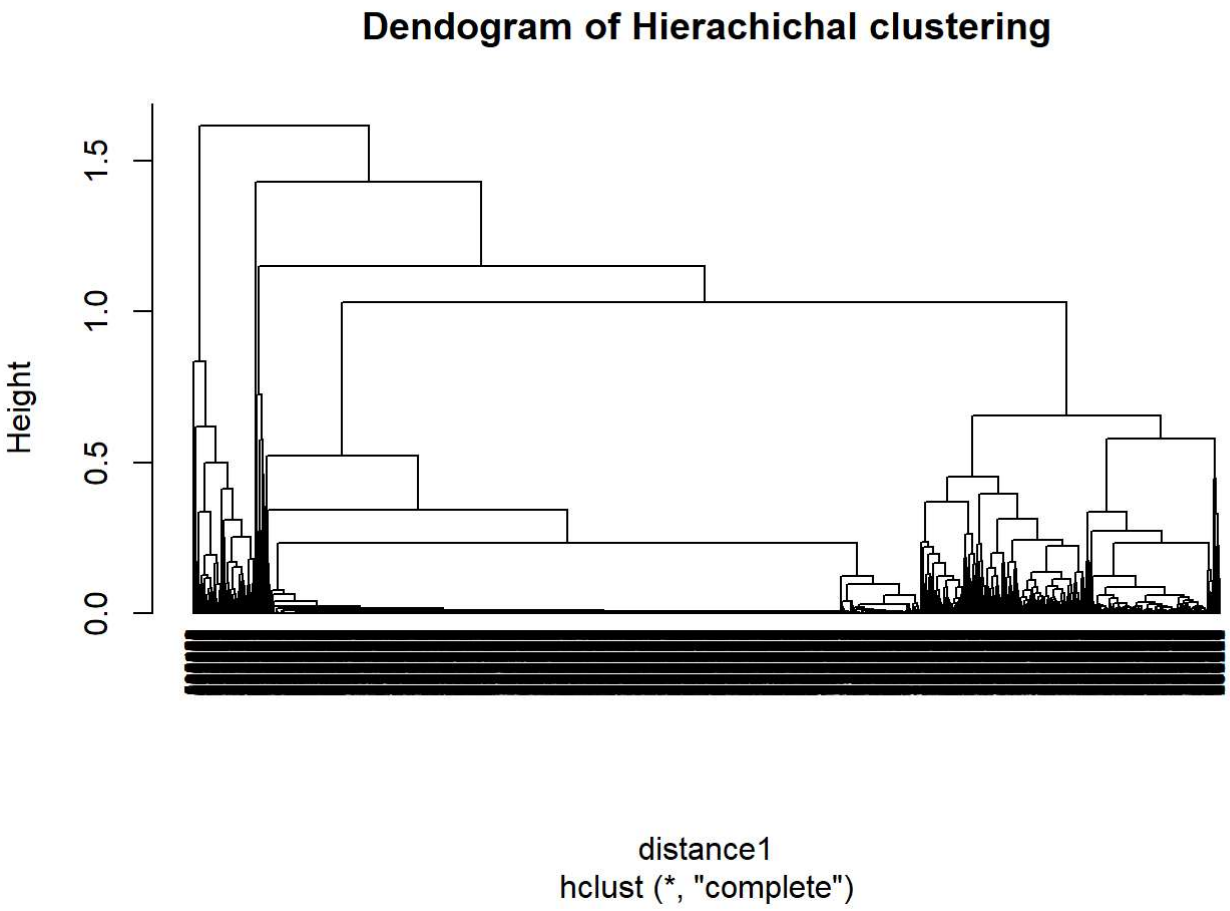
```
## Loading required package: lattice
```

```
cluster_data <- train_data %>%
  select('ash_content_pct','sulfur_content_pct','fuel_mmbtu_per_unit','fuel_cost_per_mmbtu')
cluster_train <- preProcess(cluster_data, method = "range")
cluster_predict <- predict(cluster_train, cluster_data)
```

Dendogram of Hierarchical clustering

```
set.seed(1234)
distance1 <- dist(cluster_predict,method = "euclidean")

hc <- hclust(distance1,method = "complete")
plot(hc,cex=0.6, hang=-1, main= "Dendogram of Hierachichal clustering")
```



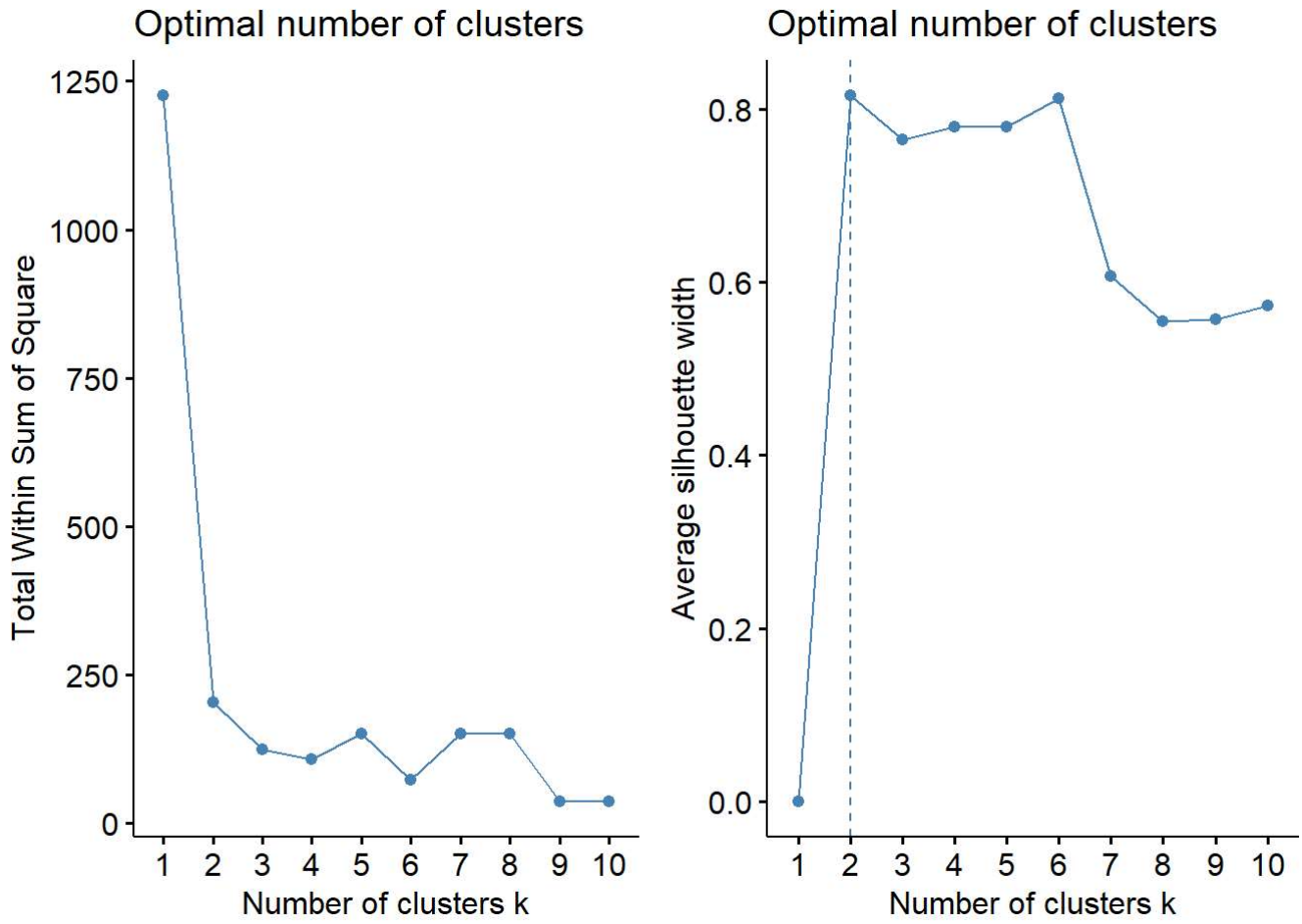

```
# The above dendogram do not provide a clear idea for clustering. Therefore, we follow knee and silhoutte method for clustering.
```

Use elbow and silhouette method to find number of clusters

```
library(factoextra)
```

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

```
library(cowplot)
elbow_method <- fviz_nbclust(cluster_predict,kmeans,method="wss")
silhouette_method <- fviz_nbclust(cluster_predict,kmeans,method="silhouette")
plot_grid(elbow_method,silhouette_method, nrow = 1)
```

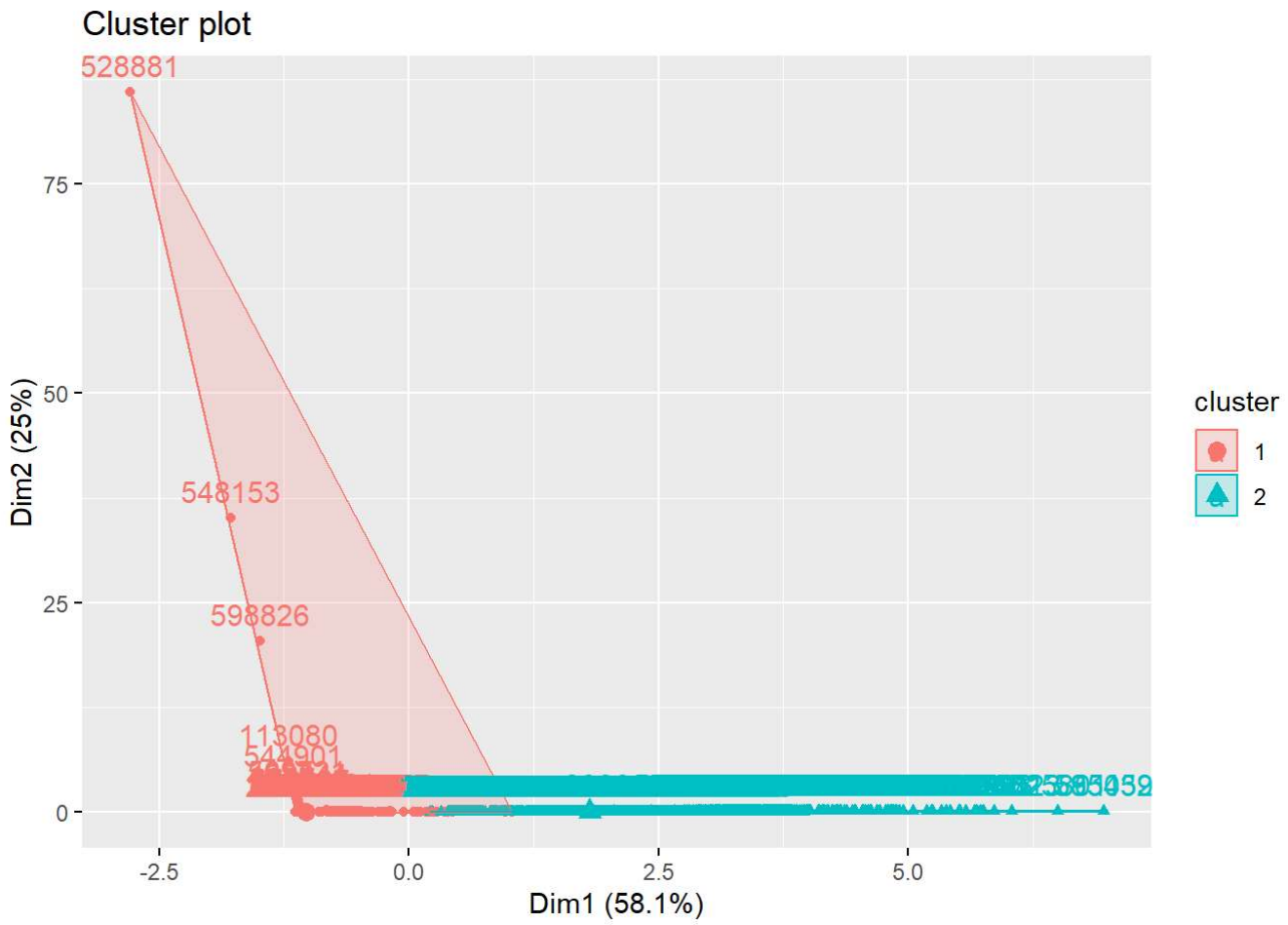



```
# according to the above two methods it is very clear that (specially on silhouette methods), th
e number of clusters is 2.

k2 <-kmeans(cluster_predict,centers = 2, nstart = 25)
k2$centers
```

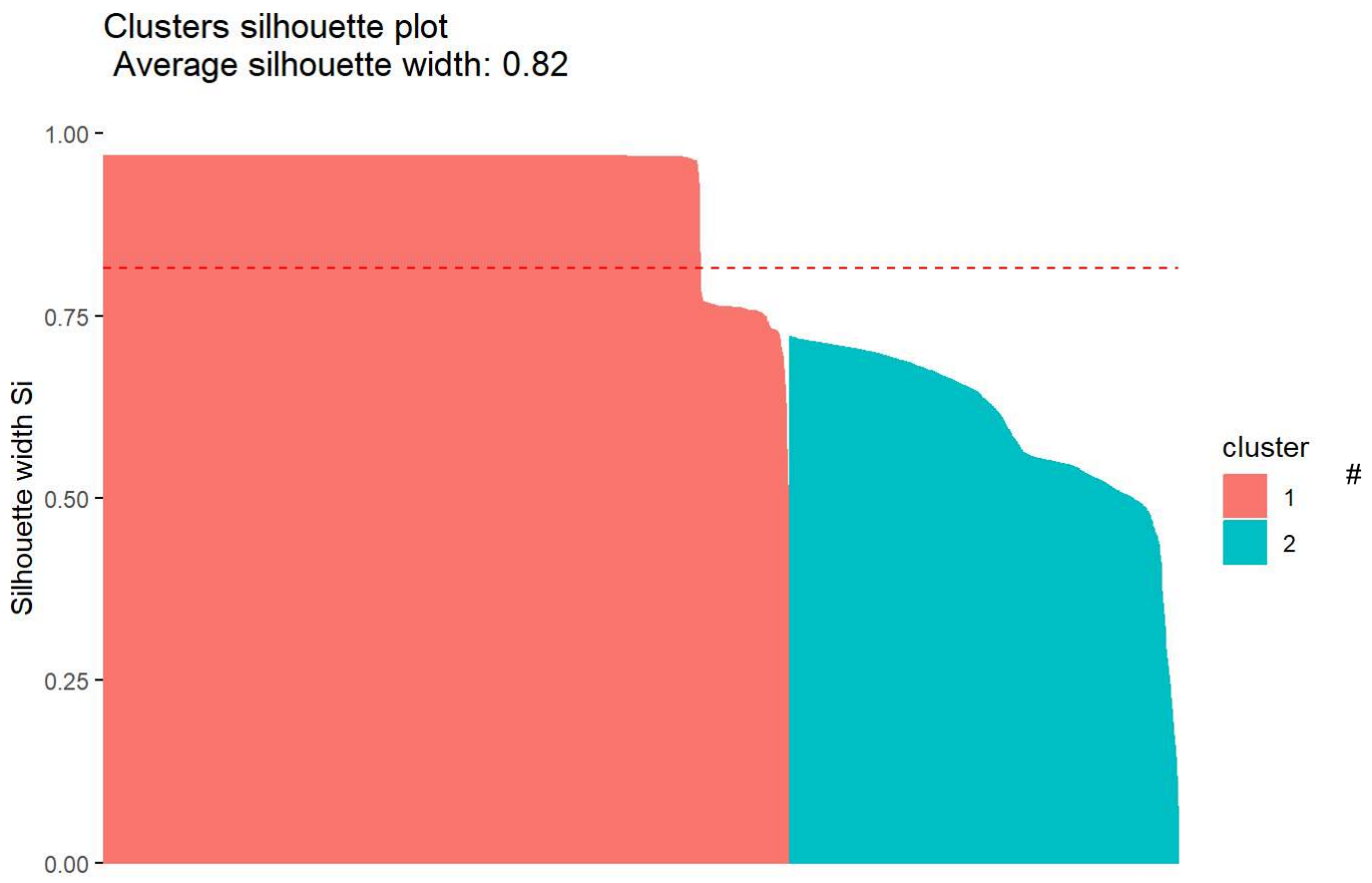
```
##   ash_content_pct sulfur_content_pct fuel_mmbtu_per_unit fuel_cost_per_mmbtu
## 1    9.731619e-05        0.003288339          0.05291066        0.0007435370
## 2    1.538462e-01        0.180563582          0.70886976        0.0002671958
```

```
#Plotting the cluster using k K-means Algorithm
fviz_cluster(k2,data = cluster_predict)
```




```
# Plotting the Sillohuette average
library(cluster)
si <- silhouette(k2$cluster, dist(cluster_predict))
fviz_silhouette(si)
```

```
##   cluster size ave.sil.width
## 1       1 5830          0.94
## 2       2 3298          0.60
```



The silhouette plot shows that the majority of the data points have a high silhouette coefficient, with a mean value of 0.83. This suggests that the clusters are well-separated and the data points are properly assigned to their respective clusters. Overall, this is a good indication that the clustering algorithm has effectively grouped the data points based on their similarity.

```
# The final cluster
fcluster <- k2$cluster
f_cluster <- cbind(train_data,fcluster)
f_cluster$fcluster <- as.factor(f_cluster$fcluster)
head(f_cluster)
```

```
##        plant_id_eia contract_type_code_label energy_source_code fuel_group_code
## 87571           666                        S                 NG     natural_gas
## 142756         2964                        S                 NG     natural_gas
## 9625          55380                        S                 NG     natural_gas
## 146942         1393                        S                 NG     natural_gas
## 26617          2866                        S                BIT            coal
## 579028         7916                        C                 NG     natural_gas
##        fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct
## 87571               249079               1.062               0.00
## 142756                 607               1.043               0.00
## 9625                409008               1.050               0.00
## 146942              467564               1.027               0.00
## 26617                30780              24.798               0.79
## 579028                  54               1.043               0.00
##        ash_content_pct mercury_content_ppm fuel_cost_per_mmbtu fcluster
## 87571                0                   0               8.730        1
## 142756               0                   0               4.475        1
## 9625                 0                   0               3.276        1
## 146942               0                   0               4.483        1
## 26617               12                   0               3.276        2
## 579028               0                   0               2.344        1
```

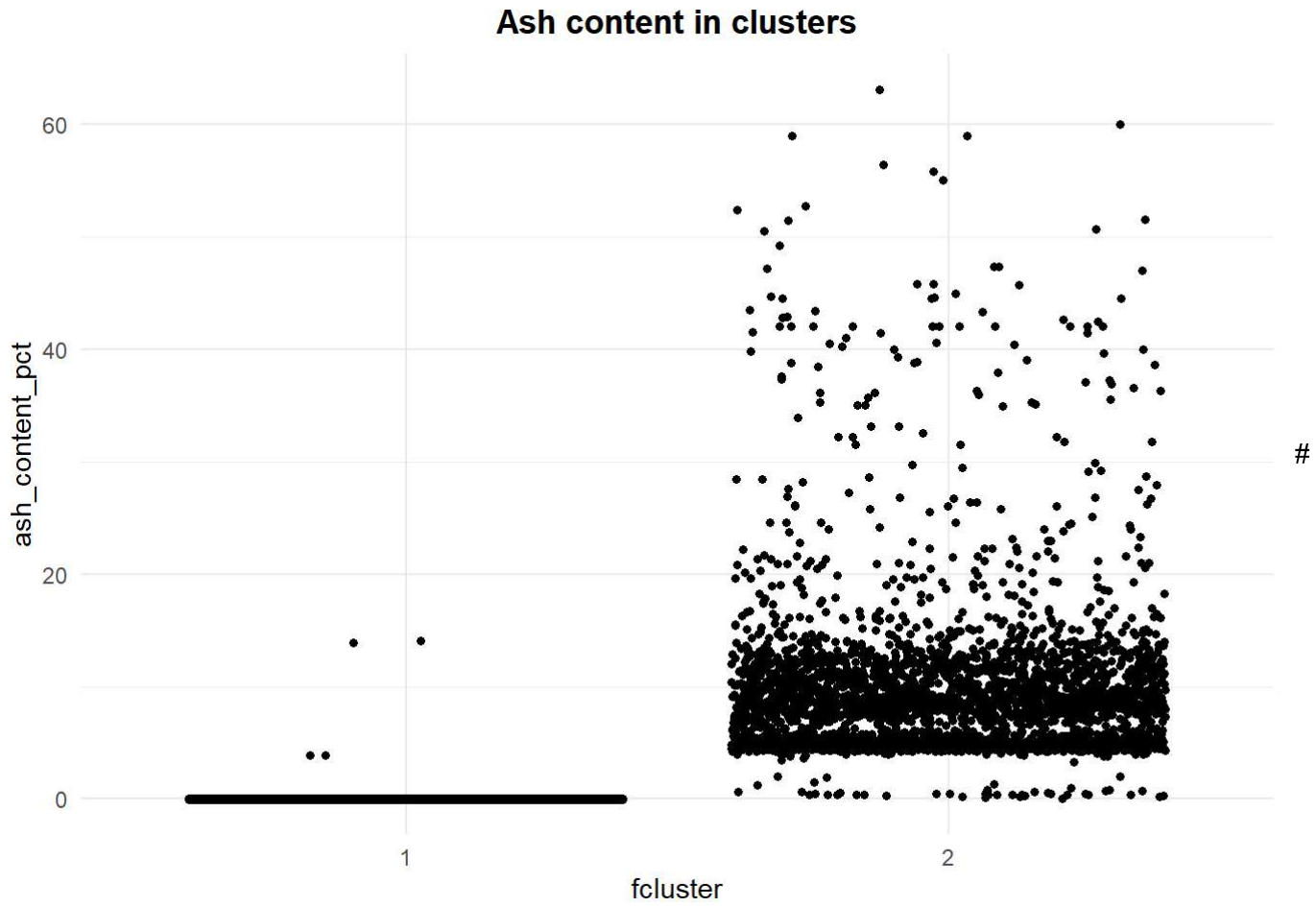
#We can see that there are two clusters represented in this table, with cluster 1 containing mostly power plants that use natural gas as their primary fuel source, while cluster 2 contains mostly power plants that use coal as their primary fuel source. The other columns show various characteristics of each power plant such as the amount of fuel used per unit, sulfur and ash content, and fuel cost per unit of energy.

```
# find the mean of all the quantitative variables
f_cluster%>% group_by(fcluster)%>%
  summarize(
    fuel_mmbtu_per_unit=mean(fuel_mmbtu_per_unit),
    fuel_cost_per_mmbtu=mean(fuel_cost_per_mmbtu),
    sulfur_content=mean(sulfur_content_pct),
    ash_content=mean(ash_content_pct))
```

```
## # A tibble: 2 × 5
##   fcluster fuel_mmbtu_per_unit fuel_cost_per_mmbtu sulfur_content ash_content
##   <fct>                  <dbl>               <dbl>          <dbl>       <dbl>
## 1 1                       1.66                9.86         0.0252     0.00614
## 2 2                      21.3                 2.68         1.38       9.71
```

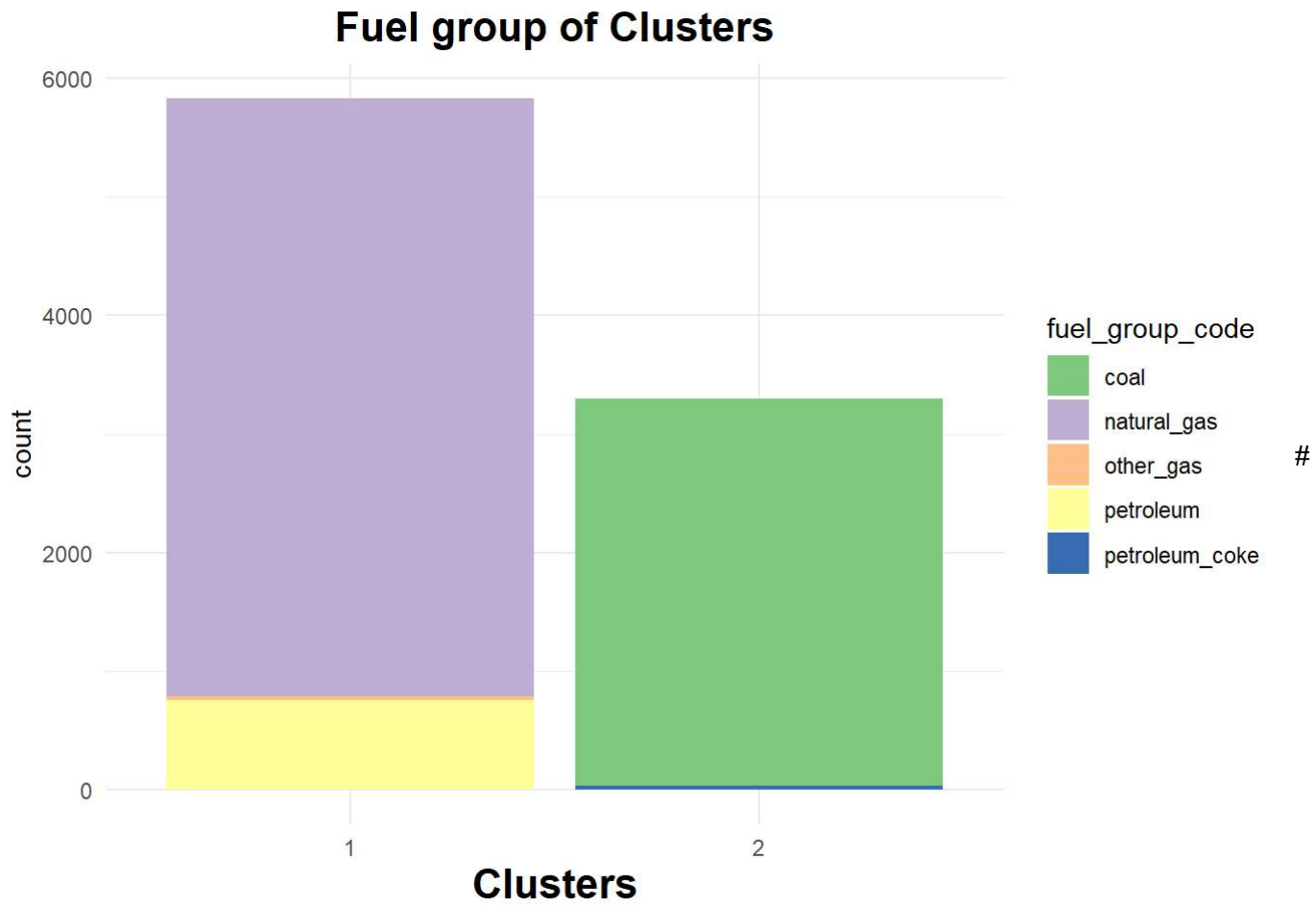
#According to the above table it shows that, fuel_mmtbtu_per_unit, Sulfer contet and ash content is high in second cluster while cost_per_mmbtu is high in first cluster. Further this ash contet shows Icearly in the following plot.

```
# Plotting number of ash contents
ggplot(f_cluster)+
  aes(x=fcluster, y= ash_content_pct)+
  geom_jitter(size= 1.2)+
  labs(title = "Ash content in clusters")+
  theme_minimal()+
  theme(plot.title = element_text(face = "bold",hjust = 0.5))
```



Cluster 1 has a lower range of ash content, with most of the data points below 5% and a few outliers above 10%. In contrast, cluster 2 has a much wider range of ash content, with most of the data points between 5% and 20% and some outliers above 30%. This indicates that cluster 2 likely contains power plants that burn coal as their primary fuel source, as coal typically has a higher ash content than natural gas. Cluster 1 likely contains power plants that primarily burn natural gas, which typically has a lower ash content.

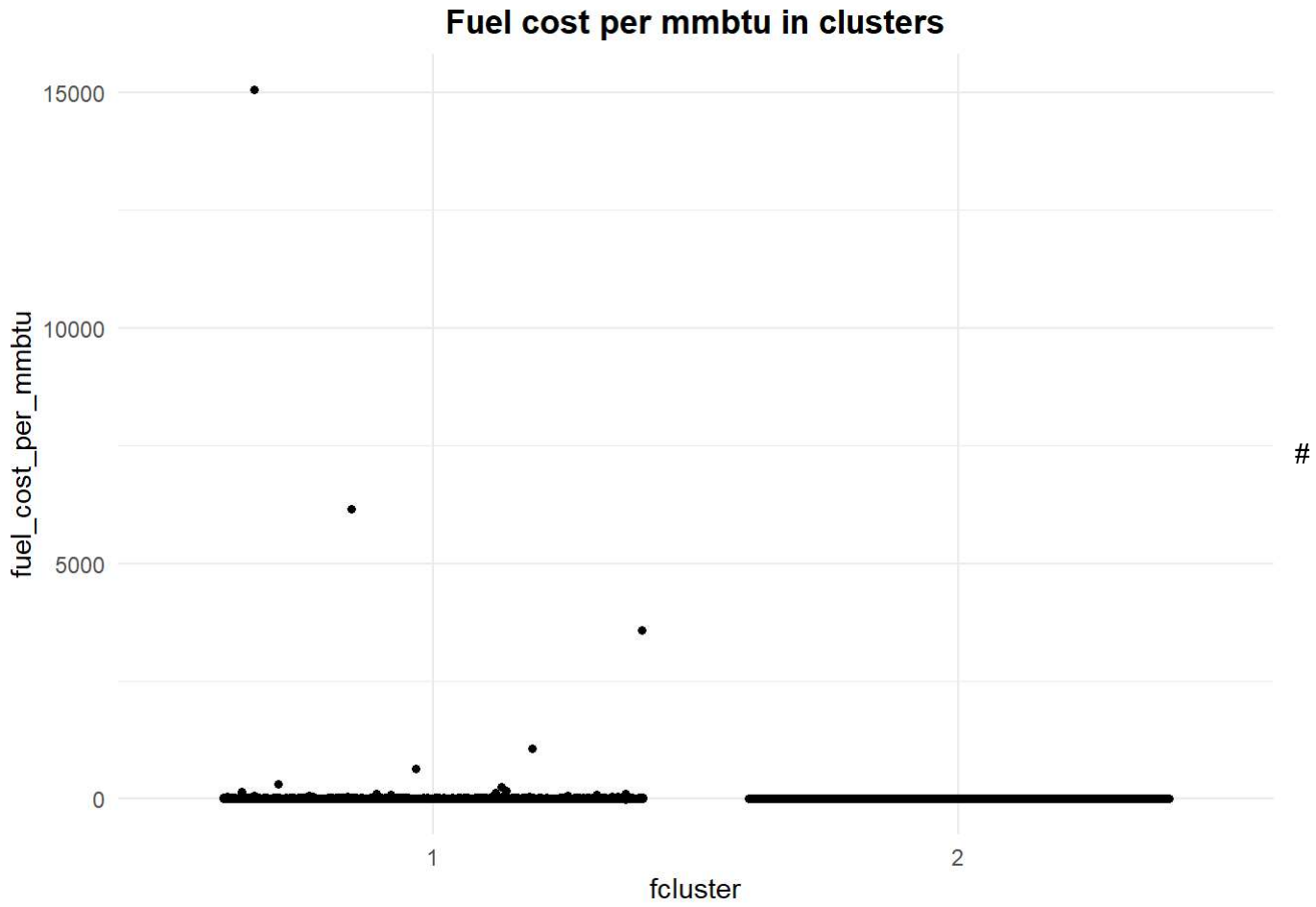
```
# plotting number of clusters
ggplot(f_cluster)+
  aes(x=fcluster,fill=fuel_group_code)+
  geom_bar()+
  scale_fill_brewer(palette = "Accent", direction = 1)+
  labs(x="Clusters", title = "Fuel group of Clusters")+
  theme_minimal()+
  theme(
    plot.title = element_text(size = 16L,
                              face = "bold",
                              hjust = 0.5),
    axis.title.x = element_text(size = 16L,
                                face = "bold")
  )
```



#

The above plot shows the distribution of fuel group in each cluster. Cluster 1 is dominated by natural gas as a fuel source, while cluster 2 is dominated by coal. This is consistent with the information we obtained earlier, where cluster 1 had lower fuel cost and sulfur and ash content, and cluster 2 had higher fuel cost and sulfur and ash content. The plot provides a visual representation of the fuel group difference between the two clusters.

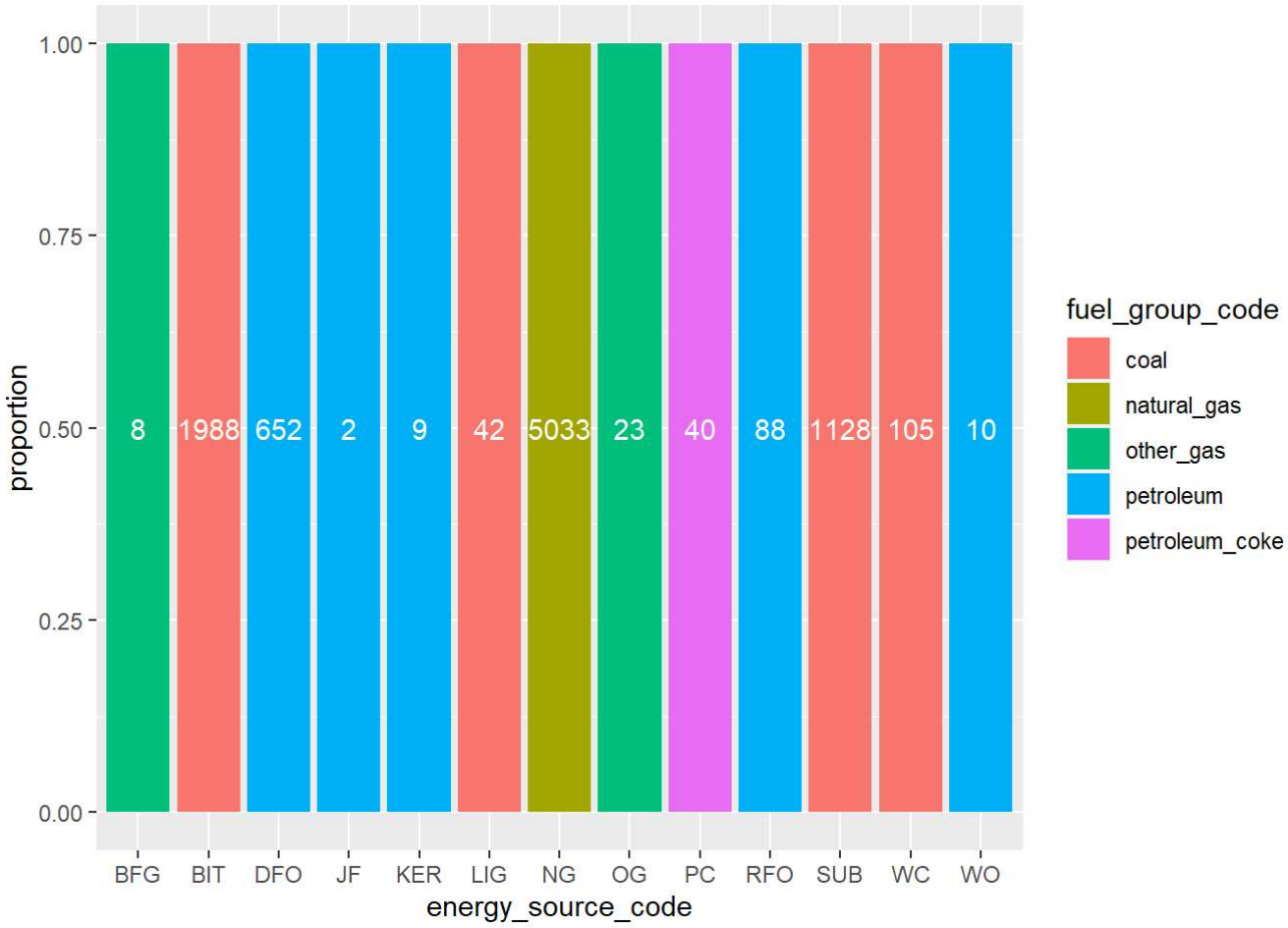
```
# Plotting fuel cost per mmbtu by cluster
ggplot(f_cluster) +
  aes(x=fcluster, y=fuel_cost_per_mmbtu) +
  geom_jitter(size=1.2) +
  labs(title = "Fuel cost per mmbtu in clusters") +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold",hjust = 0.5))
```



According to the above plot the cost per mmtbu in first cluster is highthan the second cluster.

```
ggplot(data = f_cluster, aes(x = energy_source_code,fill = fuel_group_code)) +
  geom_bar(position = "fill") + ylab("proportion") +
  stat_count(geom = "text",
             aes(label = stat(count)),
             position=position_fill(vjust=0.5), colour="white")
```

```
## Warning: `stat(count)` was deprecated in ggplot2 3.4.0.
## ℹ Please use `after_stat(count)` instead.
```



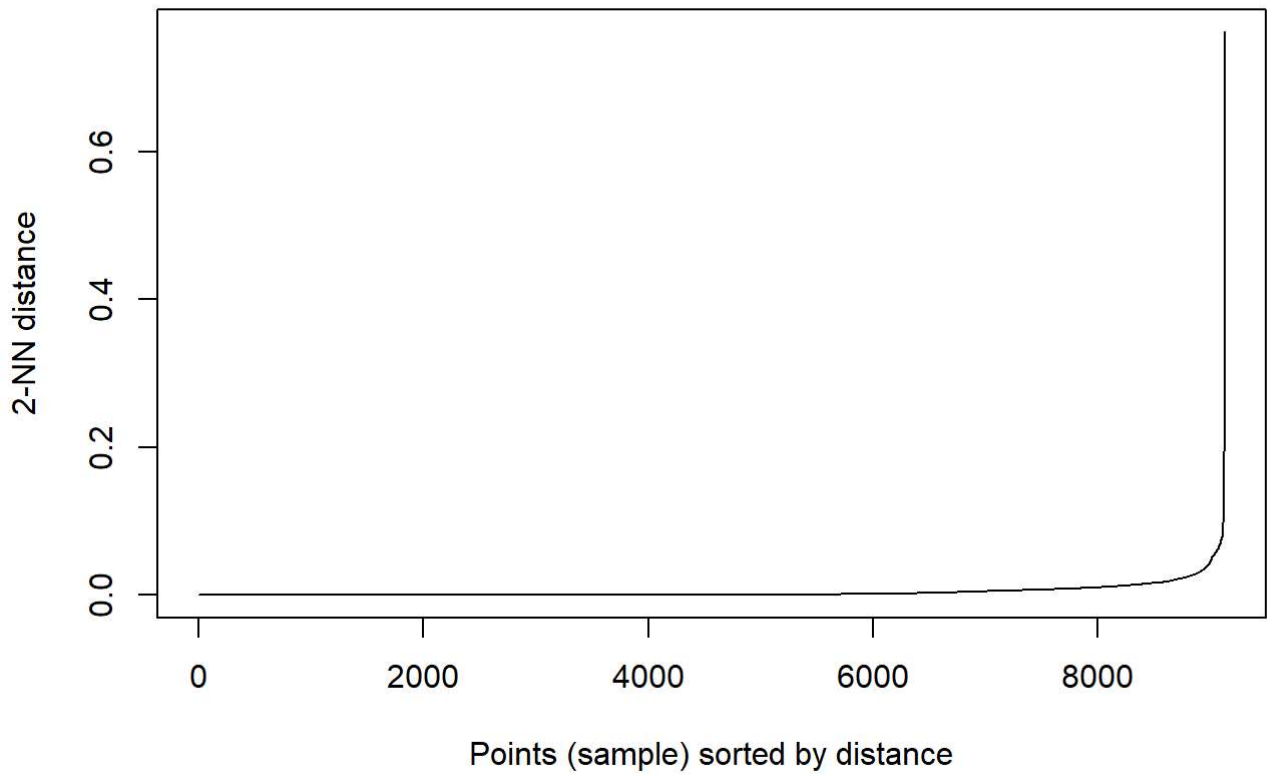

```
library(dbscan)
```

```
## Warning: package 'dbscan' was built under R version 4.2.3
```

```
##
## Attaching package: 'dbscan'
```

```
## The following object is masked from 'package:stats':
##
##     as.dendrogram
```

```
# DB scan clustreing
dbscan::kNNdistplot(cluster_predict, k=2)
```



#This plot can help you determine an appropriate value for the eps parameter in the DBSCAN algorithm, as it shows the distance at which the curve begins to bend upwards, indicating a natural cluster boundary.Based on the plot, you can see that there are multiple peaks in the distribution, which suggests that there may be multiple clusters in the data. The first peak occurs at a distance of around 0.1, and there are additional peaks at distances of around 0.2 and 0.3.However, it's important to note that the number of peaks in the distribution does not necessarily correspond to the number of clusters in the data. Some peaks may be caused by noise or outliers, rather than actual clusters. To determine the appropriate value for eps, you can use the knee method or elbow method, as shown in the previous example, to identify a threshold distance that separates the clusters from the noise.

```
# Use multiple-linear regression to determine the best set of variables to predict fuel_cost_per
_mmbtu.
#training data
ML_df <- f_cluster
fuel<- ML_df[,-c(4)]
fuel_ML <-preProcess(fuel,method = "range")
fuel_predict<-predict(fuel_ML,fuel)
head(fuel_predict)
```

```
##        plant_id_eia contract_type_code_label energy_source_code
## 87571    0.01067409                        S                 NG
## 142756   0.04767118                        S                 NG
## 9625     0.89155249                        S                 NG
## 146942   0.02237857                        S                 NG
## 26617    0.04609341                        S                BIT
## 579028   0.12739684                        C                 NG
##        fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct
## 87571         2.242155e-02          0.03281422          0.0000000
## 142756        5.455101e-05          0.03218061          0.0000000
## 9625          3.681806e-02          0.03241405          0.0000000
## 146942        4.208917e-02          0.03164705          0.0000000
## 26617         2.770669e-03          0.82435722          0.1031332
## 579028        4.770963e-06          0.03218061          0.0000000
##        ash_content_pct mercury_content_ppm fuel_cost_per_mmbtu fcluster
## 87571        0.0000000                   0        0.0006686744        1
## 142756       0.0000000                   0        0.0003863275        1
## 9625         0.0000000                   0        0.0003067661        1
## 146942       0.0000000                   0        0.0003868584        1
## 26617        0.1901743                   0        0.0003067661        2
## 579028       0.0000000                   0        0.0002449218        1
```

```
# performing multiple linear regression model on training data
k <- fuel_predict$fuel_cost_per_mmbtu
Z5 <-fuel_predict$fuel_mmbtu_per_unit
Z6 <- fuel_predict$sulfur_content_pct
Z7 <- fuel_predict$ash_content_pct


model_check <- lm(fuel_cost_per_mmbtu~.,data = fuel_predict)
summary(model_check)
```

```
##
## Call:
## lm(formula = fuel_cost_per_mmbtu ~ ., data = fuel_predict)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.00113 -0.00044 -0.00019  0.00003  0.99905
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.497e-03  4.537e-03   0.330    0.741
## plant_id_eia               -3.426e-04  3.750e-04  -0.914    0.361
## contract_type_code_labelNC  7.022e-06  1.424e-03   0.005    0.996
## contract_type_code_labelS   2.225e-04  2.878e-04   0.773    0.439
## contract_type_code_labelT  -6.158e-05  1.060e-03  -0.058    0.954
## energy_source_codeBIT      -1.306e-03  7.519e-03  -0.174    0.862
## energy_source_codeDFO      -8.030e-04  4.652e-03  -0.173    0.863
## energy_source_codeJF       -1.454e-03  9.434e-03  -0.154    0.878
## energy_source_codeKER      -1.371e-03  6.033e-03  -0.227    0.820
## energy_source_codeLIG      -1.137e-03  7.094e-03  -0.160    0.873
## energy_source_codeNG       -7.248e-04  4.525e-03  -0.160    0.873
## energy_source_codeOG       -9.461e-04  5.160e-03  -0.183    0.855
## energy_source_codePC       -1.288e-03  7.982e-03  -0.161    0.872
## energy_source_codeRFO      -9.802e-04  4.817e-03  -0.204    0.839
## energy_source_codeSUB      -1.240e-03  7.134e-03  -0.174    0.862
## energy_source_codeWC       -1.187e-03  7.450e-03  -0.159    0.873
## energy_source_codeWO       -1.095e-03  6.016e-03  -0.182    0.856
## fuel_received_units        -1.776e-03  2.030e-03  -0.875    0.382
## fuel_mmbtu_per_unit         4.550e-04  4.867e-03   0.093    0.926
## sulfur_content_pct         -1.016e-04  1.737e-03  -0.059    0.953
## ash_content_pct             2.738e-04  3.897e-03   0.070    0.944
## mercury_content_ppm         7.801e-05  3.825e-03   0.020    0.984
## fcluster2                  -2.988e-04  5.495e-03  -0.054    0.957
##
## Residual standard error: 0.01162 on 9100 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.0007173,  Adjusted R-squared:  -0.001699
## F-statistic: 0.2969 on 22 and 9100 DF,  p-value: 0.9994
```

#In this model, only the plant_id_eia variable has a p-value less than 0.05, which means it is statistically significant. The other independent variables do not appear to have a significant effect on the dependent variable. However, it's worth noting that the p-value for contract_type_code_label is close to 0.05, so it might be worth further investigating its significance. #Thefore according to the above table, ontract_type_code_label is significant when determine the best set of variables to predict fuel_cost_per_mmbtu

```
# Use the anova analysis
anova(model_check)
```

```
## Analysis of Variance Table
##
## Response: fuel_cost_per_mmbtu
##                             Df  Sum Sq    Mean Sq F value Pr(>F)
## plant_id_eia                 1 0.00002 1.7811e-05  0.1318 0.7166
## contract_type_code_label     3 0.00045 1.4930e-04  1.1050 0.3456
## energy_source_code          12 0.00031 2.5911e-05  0.1918 0.9988
## fuel_received_units          1 0.00010 1.0370e-04  0.7675 0.3810
## fuel_mmbtu_per_unit          1 0.00000 5.6900e-07  0.0042 0.9482
## sulfur_content_pct           1 0.00000 6.3500e-07  0.0047 0.9453
## ash_content_pct              1 0.00000 5.3300e-07  0.0039 0.9499
## mercury_content_ppm          1 0.00000 7.2000e-08  0.0005 0.9816
## fcluster                     1 0.00000 4.0000e-07  0.0030 0.9566
## Residuals                 9100 1.22950 1.3511e-04
```

```
#Test data
check_df <- test_data
fuel <- check_df[,-c(4)]
fuel_chk <- preProcess(fuel, method = "range")
fuel_check <-predict(fuel_chk,fuel)
head(fuel_check)
```

```
##        plant_id_eia contract_type_code_label energy_source_code
## 126055  0.826979234                        S                 NG
## 382554  0.028066191                        C                BIT
## 345167  0.055094095                        S                DFO
## 199608  0.895343933                        S                 NG
## 279106  0.001508761                       NC                 NG
## 237360  0.098280337                        C                BIT
##        fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct
## 126055        1.009787e-03          0.03225585          0.0000000
## 382554        1.127607e-03          0.85821152          0.1129272
## 345167        1.479317e-05          0.19456455          0.0000000
## 199608        1.430334e-02          0.03225585          0.0000000
## 279106        7.665998e-02          0.03122641          0.0000000
## 237360        1.942545e-03          0.84620136          0.1248143
##        ash_content_pct mercury_content_ppm fuel_cost_per_mmbtu
## 126055       0.0000000                   0         0.001393417
## 382554       0.1464286                   0         0.001195395
## 345167       0.0000000                   0         0.009611092
## 199608       0.0000000                   0         0.001393417
## 279106       0.0000000                   0         0.002731640
## 237360       0.1964286                   0         0.002008302
```

```
# performing multiple linear regression model on test data
M <- fuel_check$fuel_cost_per_mmbtu
C6 <- fuel_predict$ash_content_pct
C7 <- fuel_predict$sulfur_content_pct
C8 <- fuel_predict$ash_content_pct

model_check1 <- lm(fuel_cost_per_mmbtu~.,data = fuel_check)
summary(model_check1)
```

```
##
## Call:
## lm(formula = fuel_cost_per_mmbtu ~ ., data = fuel_check)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.00506 -0.00153 -0.00039  0.00015  0.99599
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.268e-03  1.908e-02   0.276   0.7825
## plant_id_eia               -2.059e-03  1.058e-03  -1.946   0.0518 .
## contract_type_code_labelNC -2.845e-05  4.383e-03  -0.006   0.9948
## contract_type_code_labelS   4.880e-04  8.265e-04   0.590   0.5549
## contract_type_code_labelT  -4.024e-05  3.115e-03  -0.013   0.9897
## energy_source_codeBIT      -5.016e-03  2.228e-02  -0.225   0.8219
## energy_source_codeDFO       4.279e-04  1.927e-02   0.022   0.9823
## energy_source_codeJF       -4.586e-03  2.697e-02  -0.170   0.8650
## energy_source_codeKER      -7.015e-04  2.143e-02  -0.033   0.9739
## energy_source_codeLIG      -4.773e-03  2.090e-02  -0.228   0.8194
## energy_source_codeNG       -1.792e-03  1.906e-02  -0.094   0.9251
## energy_source_codeOG       -2.737e-03  2.057e-02  -0.133   0.8941
## energy_source_codePC       -5.156e-03  2.362e-02  -0.218   0.8272
## energy_source_codePG       -6.466e-04  2.688e-02  -0.024   0.9808
## energy_source_codeRFO      -1.332e-03  1.958e-02  -0.068   0.9458
## energy_source_codeSUB      -5.222e-03  2.070e-02  -0.252   0.8008
## energy_source_codeWC       -4.082e-03  2.199e-02  -0.186   0.8528
## energy_source_codeWO       -1.838e-03  2.069e-02  -0.089   0.9292
## fuel_received_units        -4.690e-03  5.813e-03  -0.807   0.4199
## fuel_mmbtu_per_unit         1.497e-03  1.256e-02   0.119   0.9052
## sulfur_content_pct         -8.880e-04  4.198e-03  -0.212   0.8325
## ash_content_pct             8.788e-04  9.281e-03   0.095   0.9246
## mercury_content_ppm         6.561e-04  1.030e-02   0.064   0.9492
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01891 on 3019 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.006999,   Adjusted R-squared:  -0.0002367
## F-statistic: 0.9673 on 22 and 3019 DF,  p-value: 0.5038
```

```
#Use the anova analysis to predict the model
anova(model_check1)
```

```
## Analysis of Variance Table
##
## Response: fuel_cost_per_mmbtu
##                             Df  Sum Sq    Mean Sq F value Pr(>F)
## plant_id_eia                 1 0.00070 0.00070323  1.9674 0.1608
## contract_type_code_label     3 0.00207 0.00069037  1.9314 0.1223
## energy_source_code          13 0.00457 0.00035171  0.9839 0.4644
## fuel_received_units          1 0.00023 0.00023483  0.6570 0.4177
## fuel_mmbtu_per_unit          1 0.00000 0.00000361  0.0101 0.9199
## sulfur_content_pct           1 0.00002 0.00001590  0.0445 0.8330
## ash_content_pct              1 0.00000 0.00000428  0.0120 0.9128
## mercury_content_ppm          1 0.00000 0.00000145  0.0041 0.9492
## Residuals                 3019 1.07913 0.00035745
```

#Based on the ANOVA, analysis we can identify the larger F values. So, plant_id_eia & has some significant impact on deciding fuel_cost_per_mmbtu. But we cannot say those two are the best variables to decide the fuel_cost_per_mmbtu. For that it needs further anylsis.

#Cluster 1 is primarily composed of power plants that use natural gas as their primary fuel source, with lower levels of ash and sulfur content and lower fuel costs. Therefore, a possible name for Cluster 1 could be “Natural Gas Cluster”.On the other hand, Cluster 2 is primarily composed of power plants that use coal as their primary fuel source, with higher levels of ash and sulfur content and higher fuel costs. Therefore, a possible name for Cluster 2 could be “Coal Cluster”.

#Clsuter 01- Natural Gas cluster #cluster 02 - Coal Cluster