FML Project - Krishna Kumar Tavva - 811283461

2023-05-07

Setting default values to get a clean output

```
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)
rm(list = ls())
```

Loading the required packages

```
library("readr")
library("dplyr")
library("ISLR")
library("caret")
library("class")
library("ggplot2")
library("FactoMineR")
library("ggcorrplot")
library("corrr")
library("tidyverse")
library("esquisse")
library("gmodels")
library("factoextra")
library("fpc")
library("cluster")
library("pandoc")
library("pander")
```

```
setwd("E:/MSBA Github Repository/64060_ktavva/Final Project")
```

Loading the data

```
Fuel_Receipts <- read.csv("Fuel_Receipts.csv")

row.names(Fuel_Receipts) <- Fuel_Receipts[,1] #changing column name of Fuel data set to row name

head(Fuel_Receipts)
```

```
##   rowid plant_id_eia plant_id_eia_label report_date contract_type_code
## 1     1            3              Barry    01-01-08                  C
## 2     2            3              Barry    01-01-08                  C
## 3     3            3              Barry    01-01-08                  C
## 4     4            7            Gadsden    01-01-08                  C
## 5     5            7            Gadsden    01-01-08                  S
```

```
## 6     6            7            Gadsden    01-01-08                  S
##   contract_type_code_label contract_expiration_date energy_source_code
## 1                        C                 01-04-08               BIT
## 2                        C                 01-04-08               BIT
## 3                        C                                         NG
## 4                        C                 01-12-15               BIT
## 5                        S                 01-11-08               BIT
## 6                        S                 01-01-08               BIT
##   energy_source_code_label fuel_type_code_pudl fuel_group_code mine_id_pudl
## 1                      BIT               coal            coal            0
## 2                      BIT               coal            coal            0
## 3                       NG                gas     natural_gas           NA
## 4                      BIT               coal            coal            1
## 5                      BIT               coal            coal            2
## 6                      BIT               coal            coal            3
##   mine_id_pudl_label    supplier_name fuel_received_units fuel_mmbtu_per_unit
## 1                  0  interocean coal              259412              23.100
## 2                  0  interocean coal               52241              22.800
## 3                 NA bay gas pipeline             2783619               1.039
## 4                  1     alabama coal               25397              24.610
## 5                  2     d & e mining                 764              24.446
## 6                  3     alabama coal                 603              24.577
##   sulfur_content_pct ash_content_pct mercury_content_ppm fuel_cost_per_mmbtu
## 1               0.49             5.4                  NA               2.135
## 2               0.48             5.7                  NA               2.115
## 3               0.00             0.0                  NA               8.631
## 4               1.69            14.7                  NA               2.776
## 5               0.84            15.5                  NA               3.381
## 6               1.54            14.6                  NA               2.199
##   primary_transportation_mode_code primary_transportation_mode_code_label
## 1                               RV                                     RV
## 2                               RV                                     RV
## 3                               PL                                     PL
## 4                               TR                                     TR
## 5                               TR                                     TR
## 6                               TR                                     TR
##   secondary_transportation_mode_code secondary_transportation_mode_code_label
## 1
## 2
## 3
## 4
## 5
## 6
##   natural_gas_transport_code natural_gas_delivery_contract_type_code
## 1                       firm
## 2                       firm
## 3                       firm
## 4                       firm
## 5                       firm
## 6                       firm
##   moisture_content_pct chlorine_content_ppm data_maturity data_maturity_label
## 1                   NA                  NA         final               final
## 2                   NA                  NA         final               final
## 3                   NA                  NA         final               final
```

```
## 4                   NA                   NA         final               final
## 5                   NA                   NA         final               final
## 6                   NA                   NA         final               final
```

```
str(Fuel_Receipts)
```

```
## 'data.frame':    608564 obs. of  30 variables:
##  $ rowid                                  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ plant_id_eia                           : int  3 3 3 7 7 7 7 8 8 8 ...
##  $ plant_id_eia_label                     : chr  "Barry" "Barry" "Barry" "Gadsden" ...
##  $ report_date                            : chr  "01-01-08" "01-01-08" "01-01-08" "01-01-08" ...
##  $ contract_type_code                     : chr  "C" "C" "C" "C" ...
##  $ contract_type_code_label               : chr  "C" "C" "C" "C" ...
##  $ contract_expiration_date               : chr  "01-04-08" "01-04-08" "" "01-12-15" ...
##  $ 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                           : int  0 0 NA 1 2 3 NA 4 4 1 ...
##  $ mine_id_pudl_label                     : int  0 0 NA 1 2 3 NA 4 4 1 ...
##  $ supplier_name                          : chr  "interocean coal" "interocean coal" "bay gas pipeli
##  $ fuel_received_units                    : int  259412 52241 2783619 25397 764 603 2341 8869 75442
##  $ 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  "" "" "" "" ...
##  $ secondary_transportation_mode_code_label: chr  "" "" "" "" ...
##  $ natural_gas_transport_code             : chr  "firm" "firm" "firm" "firm" ...
##  $ natural_gas_delivery_contract_type_code : chr  "" "" "" "" ...
##  $ moisture_content_pct                   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ chlorine_content_ppm                   : int  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" ...
```

```
summary(Fuel_Receipts)
```

```
##      rowid          plant_id_eia   plant_id_eia_label report_date
##  Min.   :     1   Min.   :    3   Length:608564      Length:608564
##  1st Qu.:152142   1st Qu.: 2712   Class :character   Class :character
##  Median :304283   Median : 6155   Mode  :character   Mode  :character
##  Mean   :304283   Mean   :18290
##  3rd Qu.:456423   3rd Qu.:50707
##  Max.   :608564   Max.   :64020
##
##  contract_type_code contract_type_code_label contract_expiration_date
##  Length:608564      Length:608564            Length:608564
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
```

```
##
##
##
##  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
```

```
#Removing Unnecessary Variables
Fuel_Data <- Fuel_Receipts[,-c(1,3:5,7,9,10,12:14,21:30)]
```

Data Cleaning & Transformation

```
#Looking for missing Values
colMeans(is.na(Fuel_Data))
```

```
##             plant_id_eia contract_type_code_label       energy_source_code
##                0.0000000                0.0000000                0.0000000
##          fuel_group_code      fuel_received_units      fuel_mmbtu_per_unit
##                0.0000000                0.0000000                0.0000000
##       sulfur_content_pct          ash_content_pct      mercury_content_ppm
##                0.0000000                0.0000000                0.4756805
##      fuel_cost_per_mmbtu
##                0.3290369
```

```
#Treating the null values with median of the column.
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_cost_per_mmbtu, na.rm = T)
```

```
#Dropping all variables that have significant missing values
any(is.na.data.frame(Fuel_Data)) #checking the data after omitting null values
```

```
## [1] FALSE
```

Data Partition and Normalization

```
#Data Partition
set.seed(1234)
Data_Part <- createDataPartition(Fuel_Data$fuel_cost_per_mmbtu,
                                 p=0.02,list=F)
Fuel_Data_part <- Fuel_Data[Data_Part,]
Data_Part_Train <- createDataPartition(Fuel_Data_part$fuel_cost_per_mmbt,
```

```
                                        p=0.75,list=F)

Train_Data <- Fuel_Data_part[Data_Part_Train,]
Test_Data <- Fuel_Data_part[-Data_Part_Train,]
summary(Train_Data[,-c(1:4)])
```

```
##  fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct
##  Min.   :       1    Min.   : 0.077      Min.   :0.0000     Min.   : 0.000
##  1st Qu.:    3464    1st Qu.: 1.025      1st Qu.:0.0000     1st Qu.: 0.000
##  Median :   20758    Median : 1.060      Median :0.0000     Median : 0.000
##  Mean   :  242805    Mean   : 8.809      Mean   :0.5223     Mean   : 3.619
##  3rd Qu.:  103600    3rd Qu.:17.810      3rd Qu.:0.5000     3rd Qu.: 6.000
##  Max.   :12560185    Max.   :29.570      Max.   :7.8100     Max.   :69.300
##  mercury_content_ppm fuel_cost_per_mmbtu
##  Min.   :0.000000    Min.   :    0.000
##  1st Qu.:0.000000    1st Qu.:    2.747
##  Median :0.000000    Median :    3.276
##  Mean   :0.004328    Mean   :    7.803
##  3rd Qu.:0.000000    3rd Qu.:    3.948
##  Max.   :1.820000    Max.   :11750.256
```

```
#Normalization
Normalized_Data <- scale(Fuel_Data_part[,-c(1:4)])
Normalized_Train <- scale(Train_Data[,-c(1:4)])
summary(Normalized_Train)
```

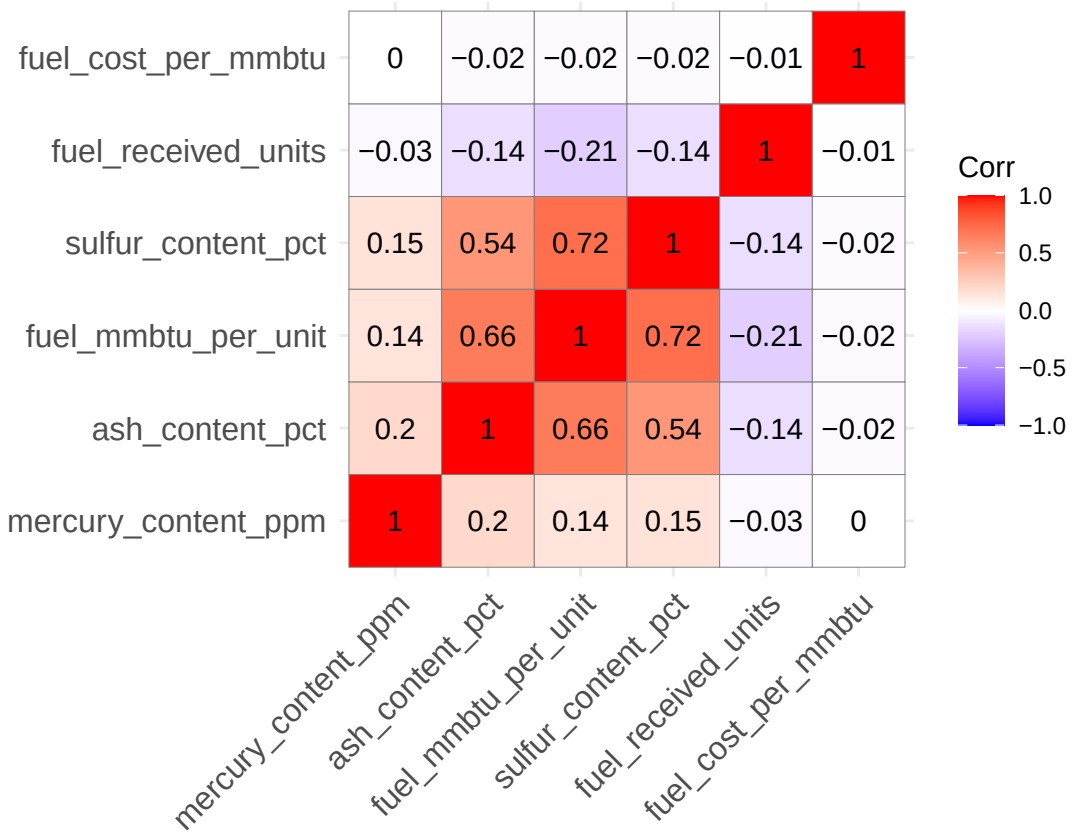
```
##  fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct
##  Min.   :-0.3261     Min.   :-0.8918     Min.   :-0.51756   Min.   :-0.5454
##  1st Qu.:-0.3215     1st Qu.:-0.7950     1st Qu.:-0.51756   1st Qu.:-0.5454
##  Median :-0.2982     Median :-0.7914     Median :-0.51756   Median :-0.5454
##  Mean   : 0.0000     Mean   : 0.0000     Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.:-0.1870     3rd Qu.: 0.9192     3rd Qu.:-0.02206   3rd Qu.: 0.3588
##  Max.   :16.5434     Max.   : 2.1203     Max.   : 7.22206   Max.   : 9.8981
##  mercury_content_ppm fuel_cost_per_mmbtu
##  Min.   :-0.1168     Min.   :-0.05216
##  1st Qu.:-0.1168     1st Qu.:-0.03379
##  Median :-0.1168     Median :-0.03026
##  Mean   : 0.0000     Mean   : 0.00000
##  3rd Qu.:-0.1168     3rd Qu.:-0.02577
##  Max.   :49.0101     Max.   :78.49493
```

```
Normalized_Test <- scale(Test_Data[,-c(1:4)])
summary(Normalized_Test)
```

```
##  fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct  ash_content_pct
##  Min.   :-0.3155     Min.   :-0.9054     Min.   :-0.526923   Min.   :-0.5663
##  1st Qu.:-0.3109     1st Qu.:-0.8100     1st Qu.:-0.526923   1st Qu.:-0.5663
##  Median :-0.2881     Median :-0.8062     Median :-0.526923   Median :-0.5663
##  Mean   : 0.0000     Mean   : 0.0000     Mean   : 0.000000   Mean   : 0.0000
##  3rd Qu.:-0.1705     3rd Qu.: 0.8975     3rd Qu.:-0.003089   3rd Qu.: 0.3913
##  Max.   :18.9828     Max.   : 2.0785     Max.   : 6.523881   Max.   : 8.9167
##  mercury_content_ppm fuel_cost_per_mmbtu
```

```
##  Min.   :-0.1412     Min.   :-0.21702
##  1st Qu.:-0.1412     1st Qu.:-0.11468
##  Median :-0.1412     Median :-0.08895
##  Mean   : 0.0000     Mean   : 0.00000
##  3rd Qu.:-0.1412     3rd Qu.:-0.05657
##  Max.   :20.3333     Max.   :45.56557
```

```
#Looking at the Correlation between Variables.
corr_matrix <- cor(Normalized_Data)
ggcorrplot(corr_matrix, outline.color = "grey50", lab = TRUE, hc.order = TRUE, type = "full")
```



Sulphur_content and ash_content_pct are highly positively correlated with Fuel_mmbtu_per_unit. There no much significant negatively correlated fields.

```
data.pca <- princomp(corr_matrix)
summary(data.pca)
```

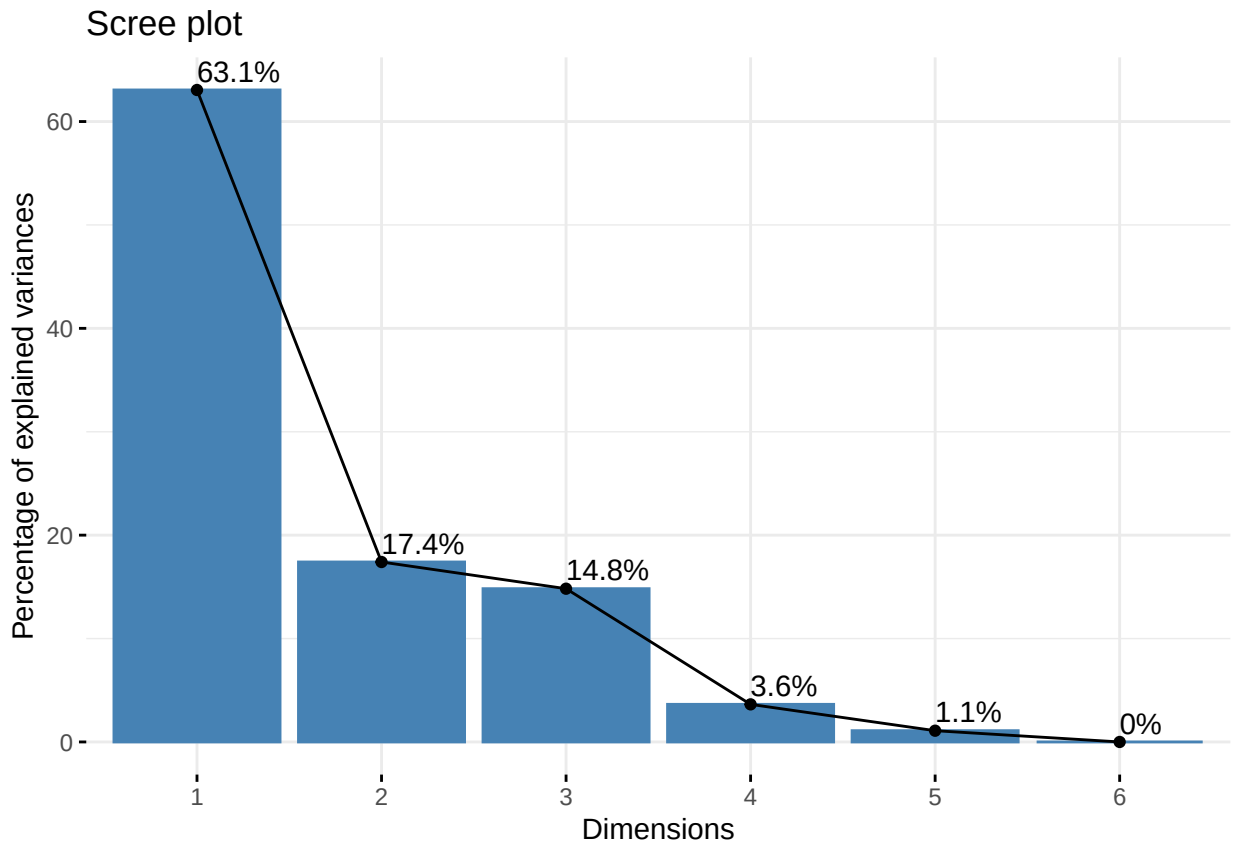
```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     0.7772648 0.4083527 0.3768495 0.18664026 0.10215463
## Proportion of Variance 0.6305096 0.1740302 0.1482141 0.03635501 0.01089105
## Cumulative Proportion  0.6305096 0.8045398 0.9527539 0.98910895 1.00000000
##                            Comp.6
## Standard deviation     8.791168e-09
## Proportion of Variance 8.065789e-17
## Cumulative Proportion  1.000000e+00
```

Six principal components have been generated (Comp.1 to Comp.6), which also correspond to the number of variables in the data. Each component explains a percentage of the total variance in the data set. In the Cumulative Proportion section, the first principal component explains almost 63% of the total variance. This implies that almost two-thirds of the data in the set of 6 variables can be represented by just the first principal component. The second one explains 17.4% of the total variance and the third one explains 14.8% of the total variance. The cumulative proportion of Comp.1, Comp.2 and Comp.3 explains nearly 95% of the total variance. This means that the first three principal components can accurately represent the data.

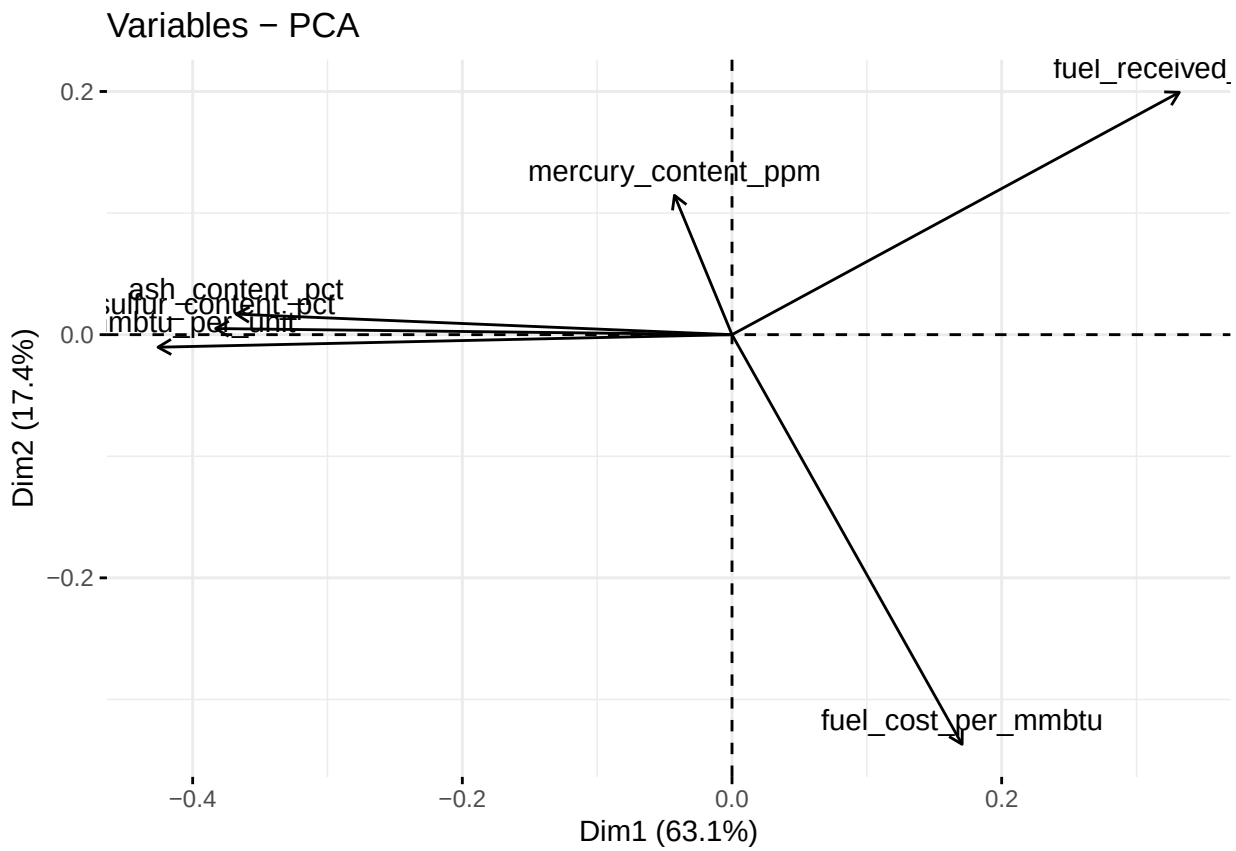
```
data.pca$loadings[, 1:3]
```

```
##                         Comp.1      Comp.2      Comp.3
## fuel_received_units  0.4268177  0.48805140  0.42234746
## fuel_mmbtu_per_unit -0.5476821 -0.02528103  0.16857375
## sulfur_content_pct  -0.4927337  0.01228943  0.18978069
## ash_content_pct     -0.4731486  0.04137523  0.06844589
## mercury_content_ppm -0.0548836  0.28057131 -0.86636862
## fuel_cost_per_mmbtu  0.2195511 -0.82497485 -0.04369632
```

```
fviz_eig(data.pca, addlabels = TRUE)
```



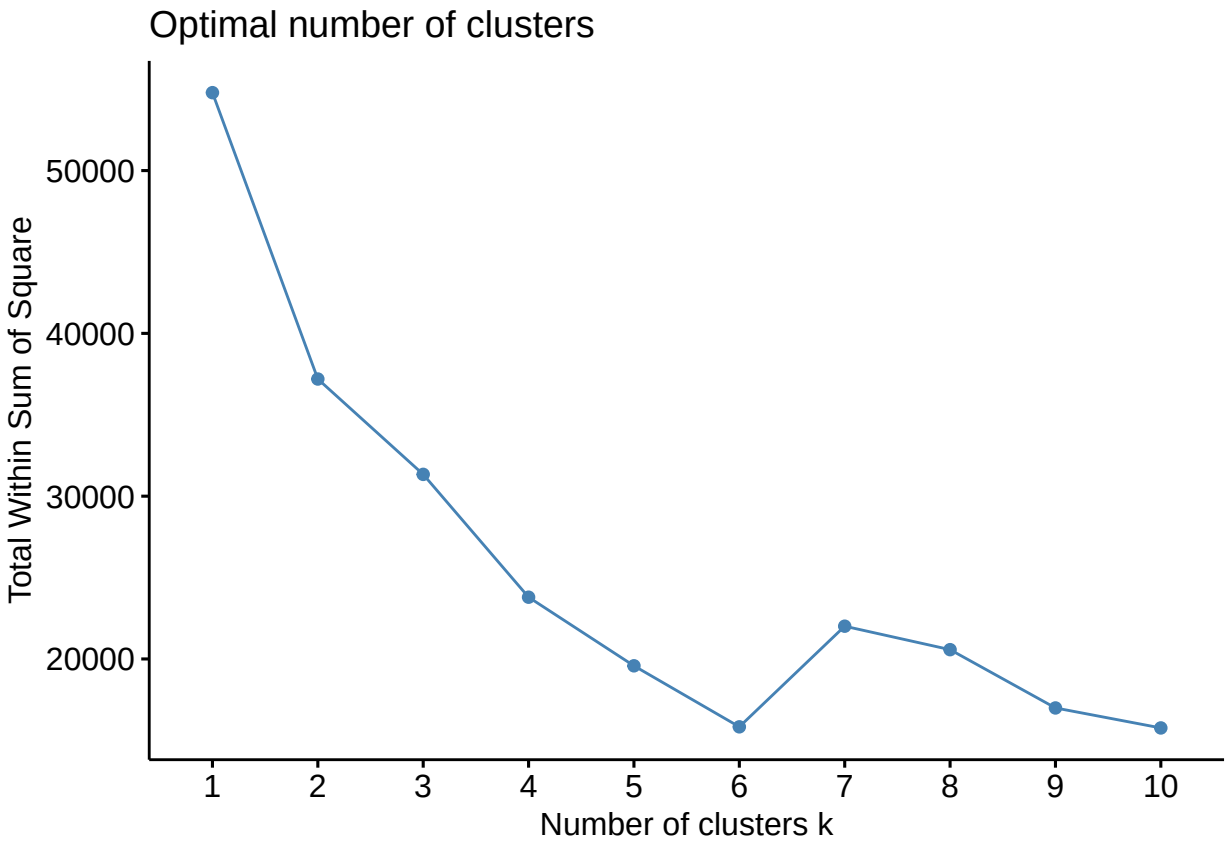

```
# Graph of the variables
fviz_pca_var(data.pca, col.var = "black")
```



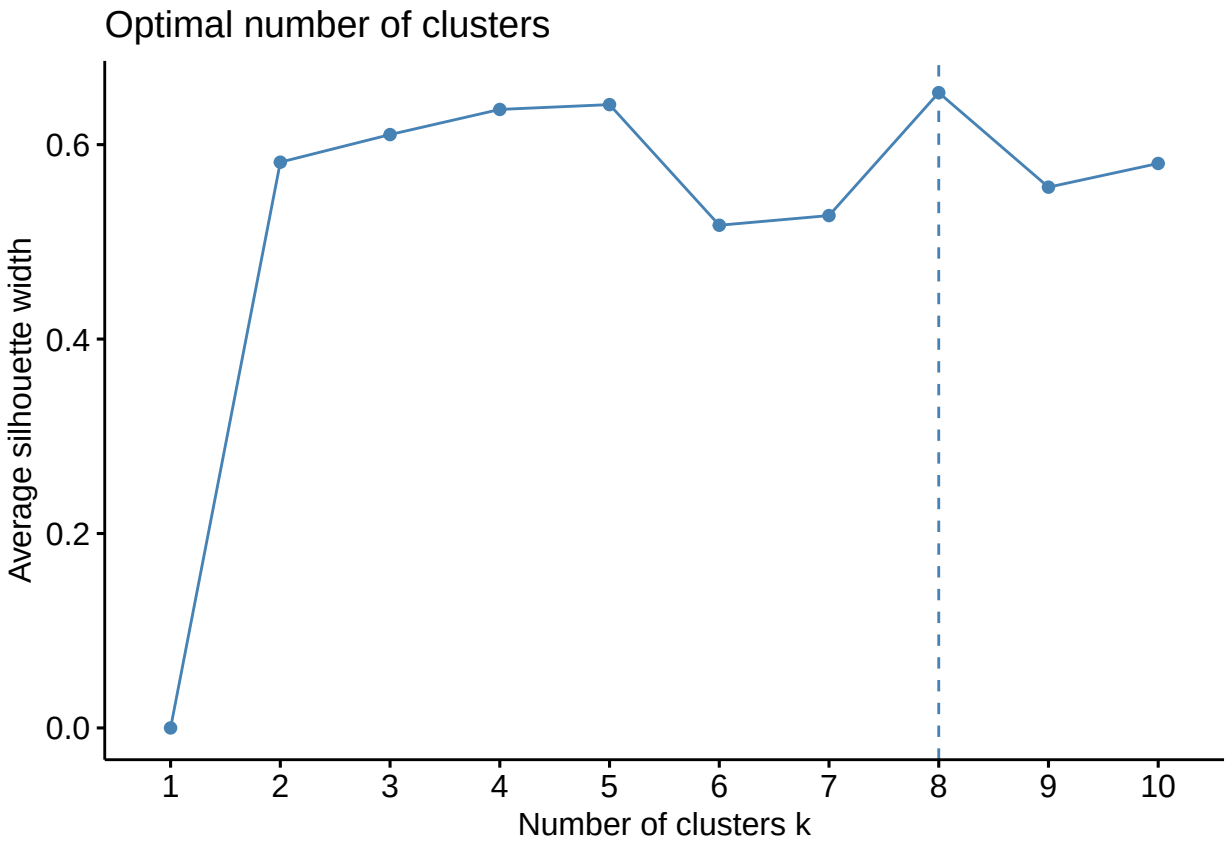
All the variables that are grouped together are positively correlated to each other. The higher the distance between the variable and the origin, the better represented that variable is. The variables that are negatively correlated are displayed to the opposite sides of the biplot's origin.

Finding the Optimal K

```
#Elbow Method
Elbow_method <- fviz_nbclust(Normalized_Train,kmeans,method="wss")
Elbow_method
```




```
#Silhouette Method
Silhouette <- fviz_nbclust(Normalized_Train,kmeans,method="silhouette")
Silhouette
```



The optimal value of k can be considered as k = 8 by using "Silhouette Method" as it is clear compared to Elbow Method.

Formulation of clusters with K=8

```
#Using K Means -Silhouette
kmeans_clust <- kmeans(Normalized_Train,centers = 8,nstart=25)
pandoc.table(kmeans_clust$centers,style="grid", split.tables = Inf)
```

```
##
##
## +--------------------+---------------------+--------------------+-----------------+-----------------
## | fuel_received_units | fuel_mmbtu_per_unit | sulfur_content_pct | ash_content_pct | mercury_content_
## +====================+=====================+====================+=================+=================
## |       -0.3177       |       -0.3088       |      -0.3273       |     -0.5447     |       -0.1168
## +--------------------+---------------------+--------------------+-----------------+-----------------
## |       -0.1453       |       -0.7948       |      -0.5176       |     -0.5454     |       -0.1168
## +--------------------+---------------------+--------------------+-----------------+-----------------
## |        2.58         |       -0.7946       |      -0.5176       |     -0.5454     |       -0.1168
## +--------------------+---------------------+--------------------+-----------------+-----------------
## |       -0.2462       |        1.185        |      0.07025       |     0.6753      |       0.1048
## +--------------------+---------------------+--------------------+-----------------+-----------------
## |       -0.326        |       -0.7942       |      -0.5176       |     -0.5454     |       -0.1168
## +--------------------+---------------------+--------------------+-----------------+-----------------
## |       -0.2965       |        1.202        |        1.93        |      2.807      |        12.6
```

```
## +--------------------+--------------------+--------------------+--------------------+--------------------
## |       7.955        |      -0.8008       |      -0.5176       |      -0.5454       |       -0.1168
## +--------------------+--------------------+--------------------+--------------------+--------------------
## |      -0.2772       |        1.43        |       2.281        |       1.418        |       0.03484
## +--------------------+--------------------+--------------------+--------------------+--------------------
```

```
kmeans_clust$size
```

```
## [1]  830 4430  486 2130    3   33   64 1156
```

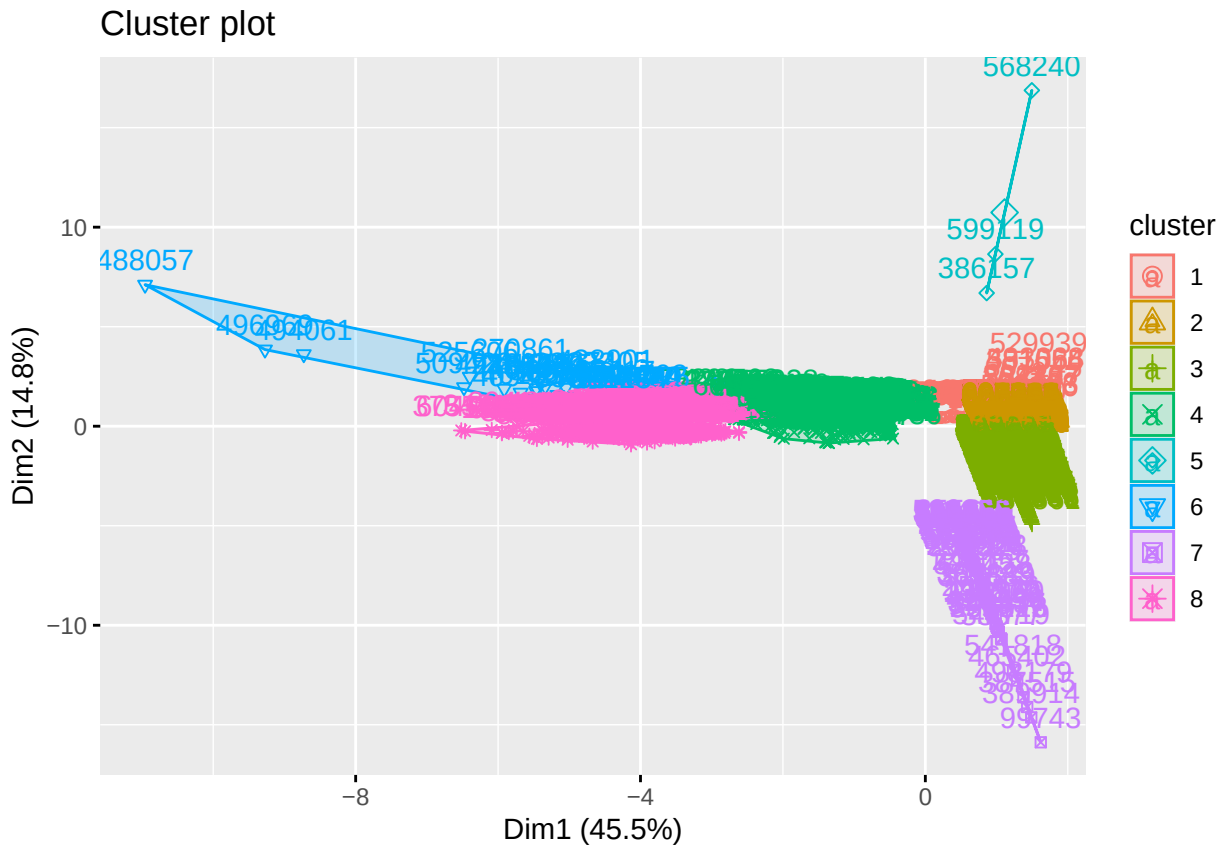
By employing the Silhouette Method we get 5 clusters of size 89,3,665,2916,6952,159,1340 and 50. Out of all, Cluster 4 has more number of observations.

Whereas, "silhouette" as a method of finding optimal k gives the analyst/user a wider scope to understand the problem.

Thus, we say that by proceeding with k=8 we can ideally have a wider vision to look and also understand about the power generation in the US..

```
cluster <- kmeans_clust$cluster
kmean_clustering <- cbind(Train_Data, cluster)

plot.cluster <- fviz_cluster(kmeans_clust, kmean_clustering[,-c(1:4)])
plot.cluster
```



```r
fuel_median <- kmean_clustering %>% group_by(cluster) %>%
summarise(median_cost = median(fuel_cost_per_mmbtu),
median_mmbtu = median(fuel_mmbtu_per_unit),
median_received_units = median(fuel_received_units),
median_sulfur = median(sulfur_content_pct)*0.01,
median_ash = median(ash_content_pct)*0.1,
median_mercury = median(mercury_content_ppm)*0.001)
fuel_median
```

```
## # A tibble: 8 x 7
##   cluster median_cost median_mmbtu median_received_units median_sulfur
##     <int>       <dbl>        <dbl>                 <dbl>         <dbl>
## 1       1       11.6          5.8                   828         0
## 2       2        3.28         1.03                22154.        0
## 3       3        3.28         1.03              2069810         0
## 4       4        2.64        18.0                 25185         0.0044
## 5       5     6010.           1.04                   27         0
## 6       6        3.28        22.2                  9532         0.0245
## 7       7        3.28         1.03              5364812         0
## 8       8        2.92        23.4                 18815         0.0283
## # i 2 more variables: median_ash <dbl>, median_mercury <dbl>
```

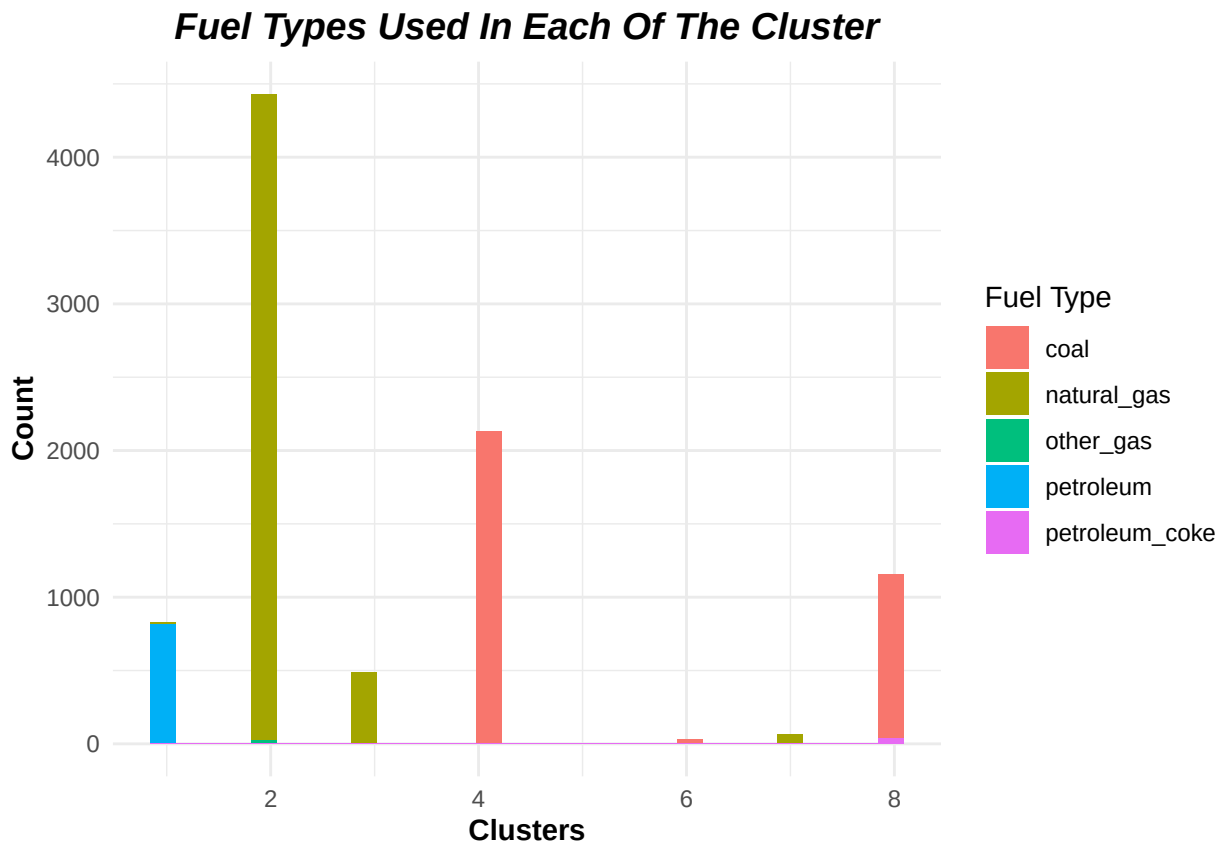
```r
fuel_clustering <- kmean_clustering %>% select(fuel_group_code, cluster) %>%
group_by(fuel_group_code, cluster) %>% count() %>% arrange(cluster)
fuel_clustering
```

```
## # A tibble: 13 x 3
## # Groups:   fuel_group_code, cluster [13]
##    fuel_group_code cluster     n
##    <chr>             <int> <int>
##  1 coal                  1     2
##  2 natural_gas           1    11
##  3 petroleum             1   817
##  4 natural_gas           2  4403
##  5 other_gas             2    27
##  6 natural_gas           3   486
##  7 coal                  4  2130
##  8 natural_gas           5     3
##  9 coal                  6    33
## 10 natural_gas           7    60
## 11 other_gas             7     4
## 12 coal                  8  1117
## 13 petroleum_coke        8    39
```

```r
Fuel_Plot <- ggplot(kmean_clustering) +
  aes(x = cluster, fill = fuel_group_code) +
  geom_histogram(bins = 30L) +
  scale_fill_hue(direction = 1) +
  labs(
    x = "Clusters",
    y = "Count",
    title = "Fuel Types Used In Each Of The Cluster",
```

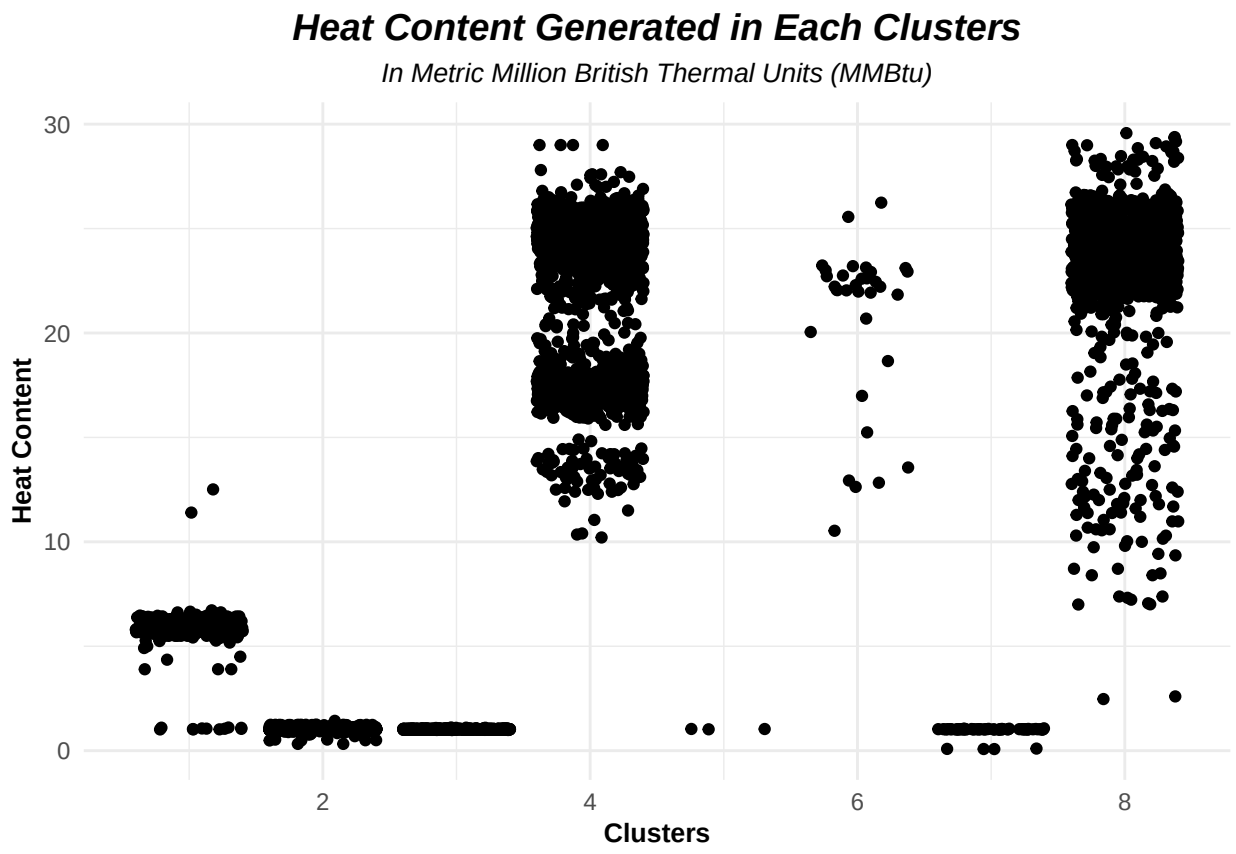
```
    fill = "Fuel Type"
) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14L,
    face = "bold.italic",
    hjust = 0.5),
    axis.title.y = element_text(face = "bold"),
    axis.title.x = element_text(face = "bold")
)


Fuel_Plot
```




```
Heat_Content_Plot <- ggplot(kmean_clustering) +
  aes(x = cluster, y = fuel_mmbtu_per_unit) +
  geom_jitter(size = 1.5) +
  labs(x = "Clusters",
    y = "Heat Content",
    title = "Heat Content Generated in Each Clusters",
    subtitle = "In Metric Million British Thermal Units (MMBtu)") +
  theme_minimal() +
  theme(plot.title = element_text(size = 14L,
    face = "bold.italic",
    hjust = 0.5),
```

```
    plot.subtitle = element_text(size = 10L,
    face = "italic",
    hjust = 0.5),
    axis.title.y = element_text(size = 10L,
    face = "bold"),
    axis.title.x = element_text(size = 10L,
    face = "bold"))
Heat_Content_Plot
```



Describing the cluster

Cluster 1: This cluster predominantly utilizes petroleum and natural gas as fuel sources for power generation. The median cost for generating power units with a heat content of 5.8 million metric British thermal units (MMBtu) is $11.6. This cluster generates a total of 828 units, and no impurities of Sulphur, mercury, or ash content are detected.

Cluster 2: Natural gas and other gases are the primary fuel sources in this cluster. The median cost for generating power units with a heat content of 1.0270 MMBtu is $3.276. This cluster generates approximately 22,154.5 units of power, and there are no impurities of Sulphur, mercury, or ash content.

Cluster 3: This cluster also relies on natural gas as its fuel source. The median cost for 1.0300 natural gas units is $3.276. It generates a significant amount of power with 2,069,810.0 units, which is the second highest among all clusters. No impurities are observed in this cluster.

Cluster 4: The fuel source in this cluster is coal. The price for generating power units with a heat content of 18.0220 MMBtu is $2.643. This cluster generates a total of 25,185.0 units of power. However, it exhibits impurities of ash exceeding the permissible levels at 0.65 parts per million (ppm) and Sulphur at 0.0044 ppm.

Cluster 5: Natural gas is once again the fuel source in this cluster. The median cost for generating power units, which includes extreme outliers, is $6010.289 for a heat content of 1.0370. This cluster generates a minimal amount of power, with only 27.0 units, and no impurities are present.

Cluster 6: Coal serves as the fuel source in this cluster. The price for generating power units with a heat content of 22.2140 MMBtu is $3.276. This cluster generates a total of 9,532.0 units of power. However, it surpasses the permissible levels of ash impurities at 2.02 ppm and Sulphur at 0.0245 ppm. Additionally, there is a presence of mercury impurities at 0.00038.

Cluster 7: Among all the clusters, this cluster generates the highest number of gas units for power generation, amounting to 5,364,812.0 units. The fuel sources used are natural gas and other gases. The price for generating power units with a heat content of 1.0255 MMBtu is $3.276. No impurities are detected in this cluster.

Cluster 8: This cluster utilizes coal and petroleum coke as fuel sources. The price for generating power units with a heat content of 23.3515 MMBtu is $2.916. It generates a total of 18,815.0 units of power. However, the cluster exhibits impurities of ash surpassing the permissible levels at 0.91 ppm, and Sulphur levels exceeding the permissible limit at 0.0283 ppm.

```
#Extra Credit - building the model

model<-lm(Fuel_Data_part$fuel_cost_per_mmbtu~.,data=Fuel_Data_part[,-c(2:4)])

summary(model)
```

```
##
## Call:
## lm(formula = Fuel_Data_part$fuel_cost_per_mmbtu ~ ., data = Fuel_Data_part[,
##     -c(2:4)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##   -11.8    -6.3    -3.0     0.0 11737.4
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.366e+00  2.200e+00   4.257 2.09e-05 ***
## plant_id_eia         6.588e-05  5.704e-05   1.155   0.2481
## fuel_received_units -2.789e-06  1.627e-06  -1.715   0.0865 .
## fuel_mmbtu_per_unit -2.793e-01  2.056e-01  -1.359   0.1743
## sulfur_content_pct   2.687e-01  1.724e+00   0.156   0.8762
## ash_content_pct     -1.108e-01  2.434e-01  -0.455   0.6490
## mercury_content_ppm -8.738e-01  3.346e+01  -0.026   0.9792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 129.9 on 12167 degrees of freedom
## Multiple R-squared:  0.0009032,  Adjusted R-squared:  0.0004106
## F-statistic: 1.833 on 6 and 12167 DF,  p-value: 0.08848
```

This shows that by choosing variables with significant relationship and cluster information leads to better prediction. Here we can see that our p value is greater than 5 % and r squared value is too low. This means variables we considered doesn't account to the variability for fuel cost per mmbtu variable.

```
#Finding variable importance

varImp(model)
```

```
##                       Overall
## plant_id_eia        1.1549331
## fuel_received_units 1.7145515
## fuel_mmbtu_per_unit 1.3586670
## sulfur_content_pct  0.1557953
## ash_content_pct     0.4551717
## mercury_content_ppm 0.0261177
```

```
#Running the multiple linear regression model using just two variables which have greater statistical s
model_1 <- lm(fuel_cost_per_mmbtu~fuel_mmbtu_per_unit+fuel_received_units,data=Train_Data[,-c(2:4)])
summary(model_1)
```

```
##
## Call:
## lm(formula = fuel_cost_per_mmbtu ~ fuel_mmbtu_per_unit + fuel_received_units,
##     data = Train_Data[, -c(2:4)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##   -11.7    -7.8    -4.0     0.1 11738.5
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.224e+01  2.263e+00   5.410 6.47e-08 ***
## fuel_mmbtu_per_unit -4.210e-01  1.636e-01  -2.573   0.0101 *
## fuel_received_units -3.002e-06  2.151e-06  -1.396   0.1629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 149.6 on 9129 degrees of freedom
## Multiple R-squared:  0.0008073,  Adjusted R-squared:  0.0005884
## F-statistic: 3.688 on 2 and 9129 DF,  p-value: 0.02507
```

This shows that by choosing variables with significant relationship and cluster information leads to better prediction. Here we can see that our p value is greater than 1 % and r squared value is too low. This means variables we considered doesn't account to the variability for fuel cost per mmbtu variable.

```
model_predict <-predict(model_1,Test_Data,type="response")

summary(model_predict)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -32.321   3.849   9.787   7.723  11.717  12.022
```

```
Test_Predict <- cbind(Test_Data,model_predict)
```

#The predicted value seems far away from the actual values, this can be referred by looking at the "Test_Predict" data frame.