FML FINAL PROJECT

Rishitha Reddy Muddasani

2023-05-06

#Loading the necessary Libraries for project

```
library(caret)
library(class)
library(tidyverse)
library(dlookr)
library(missRanger)
library(factoextra)
library(esquisse)
```

#1.Importing the dataset:

```
library(readr)
data <- read_csv("fuel.csv")
```

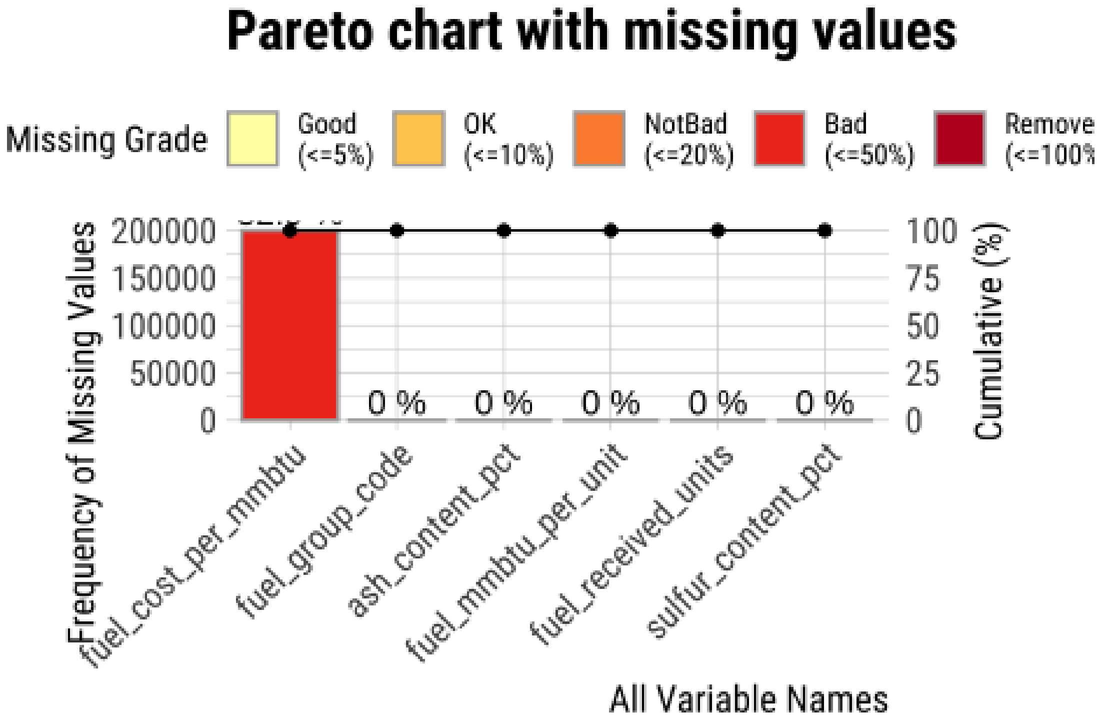
#2. Removing insignnificant variables and selecting main attributes for clustering to understand Power generation:

```
data_new<-data[,c(8,11:14,16)]
str(data_new)

## tibble [608,565 × 6] (S3: tbl_df/tbl/data.frame)
##  $ fuel_group_code    : chr [1:608565] "coal" "coal" "natural_gas" "coal"
...
##  $ fuel_received_units: num [1:608565] 259412 52241 2783619 25397 764 ...
##  $ fuel_mmbtu_per_unit: num [1:608565] 23.1 22.8 1.04 24.61 24.45 ...
##  $ sulfur_content_pct : num [1:608565] 0.49 0.48 0 1.69 0.84 1.54 0 2.16
1.24 1.9 ...
##  $ ash_content_pct    : num [1:608565] 5.4 5.7 0 14.7 15.5 14.6 0 15.4
11.9 15.4 ...
##  $ fuel_cost_per_mmbtu: num [1:608565] 2.13 2.12 8.63 2.78 3.38 ...
```

#3. Using the dlookr package, plotting the values that are missing from the above dataset to check for missing values: #dlookr package provides a visual representation of how many values are missing from every variable in percentages. This helps in understanding the dataset and determining if missing values should be imputed or eliminated.

```
plot_na_pareto(data_new)
```



#As can be seen from the visual plot, fuel_cost_per_mmbtu has missing values. fuel_cost_per_mmbtu is an essential predicting factor in determining heat generation and fuel source type. As a result, rather of fully eliminating the missing numbers, it is critical to impute them.

#4. Using the missRanger package, Imputing missing values in fuel_cost_per_mmbtu: #Imputation is the process of replacing missing values with different values that helps to complete the dataset. Imputation can be accomplished in a variety of ways. The missRanger package imputes missing variable values using other variables as predictors. The process is continued until the error rate stops improving.

```
data_clean<- missRanger(data_new, formula = .~., num.trees = 100, seed = 3)

##
## Missing value imputation by random forests
##
##   Variables to impute:       fuel_group_code, fuel_cost_per_mmbtu
##   Variables used to impute:  fuel_group_code, fuel_received_units,
fuel_mmbtu_per_unit, sulfur_content_pct, ash_content_pct, fuel_cost_per_mmbtu
##
## iter 1
##   |
|                                                                          |   0%
|
|===================================                                      |  50%
```

```
|
|======================================================================| 100%
## iter 2
##    |
|                                                                      |   0%
|
|===================================                                   |  50%
|
|======================================================================| 100%
## iter 3
##    |
|                                                                      |   0%
|
|===================================                                   |  50%
|
|======================================================================| 100%
```

#5. Sampling data and splitting data: #The population dataset containing 608565 observations was sampled to a sample size of 2% by setting the seed value as (9596).

```
set.seed(9596)
sample_data <- data_clean[sample(nrow(data_clean), size = 12000, replace =
FALSE), ]
```

#6. Dataset has been divided into TRAINING (which consists of 75% of the data) and TEST SETS(remaining 25% of data) with respect to the fuel_cost_per_mmbtu. Since fuel_cost_per_mmbtu helps understand how the heat output of the obtained fuel units behaves, the fuel cost has been designated as an important factor in classifying the data.

```
train_index <- createDataPartition(sample_data$fuel_cost_per_mmbtu, p=0.75,
list = FALSE)
train_data<- sample_data[train_index,]
test_data<- sample_data[-train_index,]
```

#7.Subsetting numerical variables for the purpose of scaling and clustering:

```
#For the basis of clustering, the data set has been filtered to only
represent only numerical variables.

cluster_data <- train_data %>% select('fuel_received_units',
'fuel_mmbtu_per_unit', 'sulfur_content_pct', 'ash_content_pct',
'fuel_cost_per_mmbtu')

#Normalization of numerical values using center, scale. Center and scale was
used as the mean values to 0 and standard deviation to 1. This reduces the
impact of outliers in the data set as mean considers the lowest and  highest
values to calculate the average.

cluster_train <- preProcess(cluster_data, method = c("center", "scale"))
cluster_predict <- predict(cluster_train, cluster_data)
summary(cluster_predict)
```

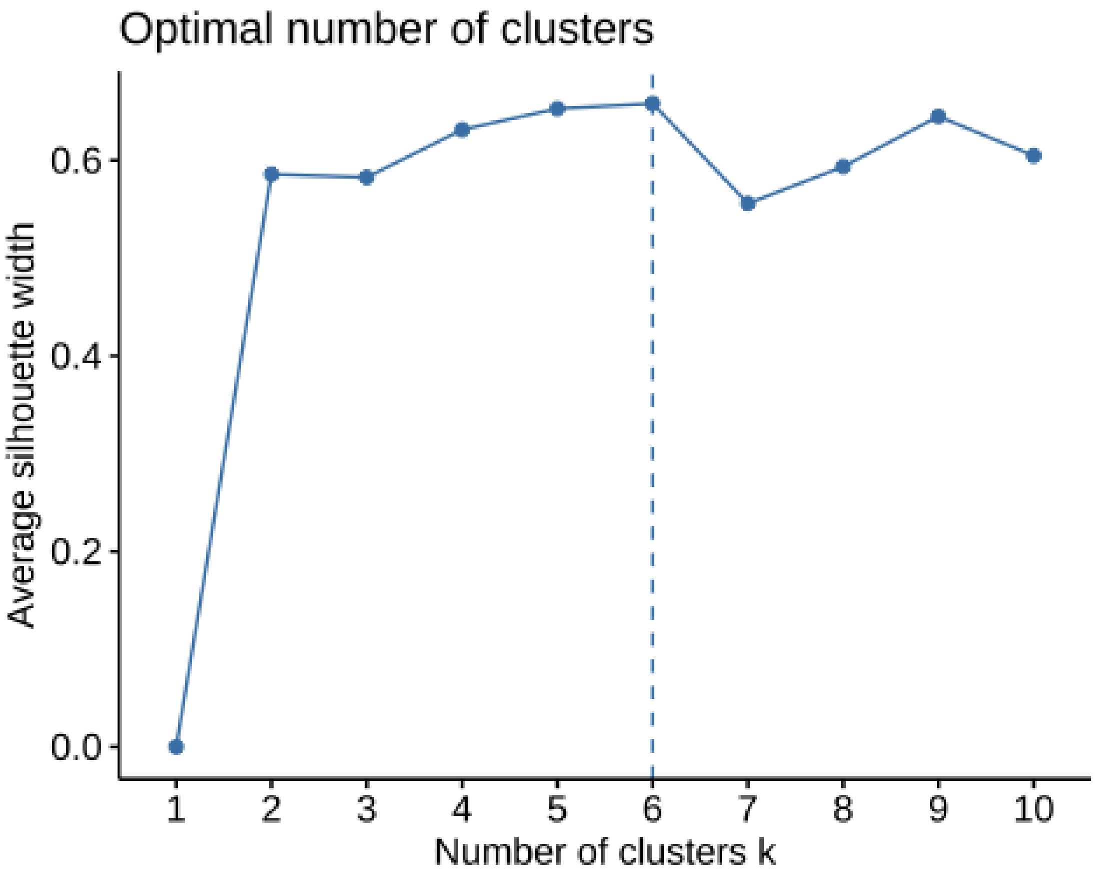
```
##  fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct 
ash_content_pct
##  Min.   :-0.3261     Min.   :-0.9012     Min.   :-0.51701   Min.   :-
0.5462
##  1st Qu.:-0.3215     1st Qu.:-0.7986     1st Qu.:-0.51701   1st Qu.:-
0.5462
##  Median :-0.2986     Median :-0.7946     Median :-0.51701   Median :-
0.5462
##  Mean   : 0.0000     Mean   : 0.0000     Mean   : 0.00000   Mean   : 
0.0000
##  3rd Qu.:-0.1852     3rd Qu.: 0.9175     3rd Qu.:-0.01405   3rd Qu.: 
0.3406
##  Max.   :19.1545     Max.   : 2.1656     Max.   : 7.18832   Max.   : 
9.7440
##  fuel_cost_per_mmbtu
##  Min.   :-0.24533   
##  1st Qu.:-0.15366   
##  Median :-0.10032   
##  Mean   : 0.00000   
##  3rd Qu.:-0.02945   
##  Max.   :63.18793
```

#8. Using the Silhouette approach to locate the optimum clustering centers: #Clustering is the classification of similar objects into one category. The K-means clustering technique clusters the groups using the K value, where each k value denotes what group represents based on the data set's centers and how various data points behave around these centers. As a result, it is critical to determine the value of k.

#Silhoutte method is one such approach for determining the value of k. The silhouette approach defines cluster values based on how data points behave inside their own clusters and how each cluster differs from others.

#Understanding the Business objective: The dataset is categorised based on fuel_cost_per_mmbtu; silhouette assists in understanding how the data points in each cluster behave in terms of cost within each cluster and how they differ compared to other clusters. This allows us to examine each cluster based on heat production, sulfur, and ash content, which aids in identifying the best cluster.

```
fviz_nbclust(cluster_predict, kmeans, method = "silhouette")
```

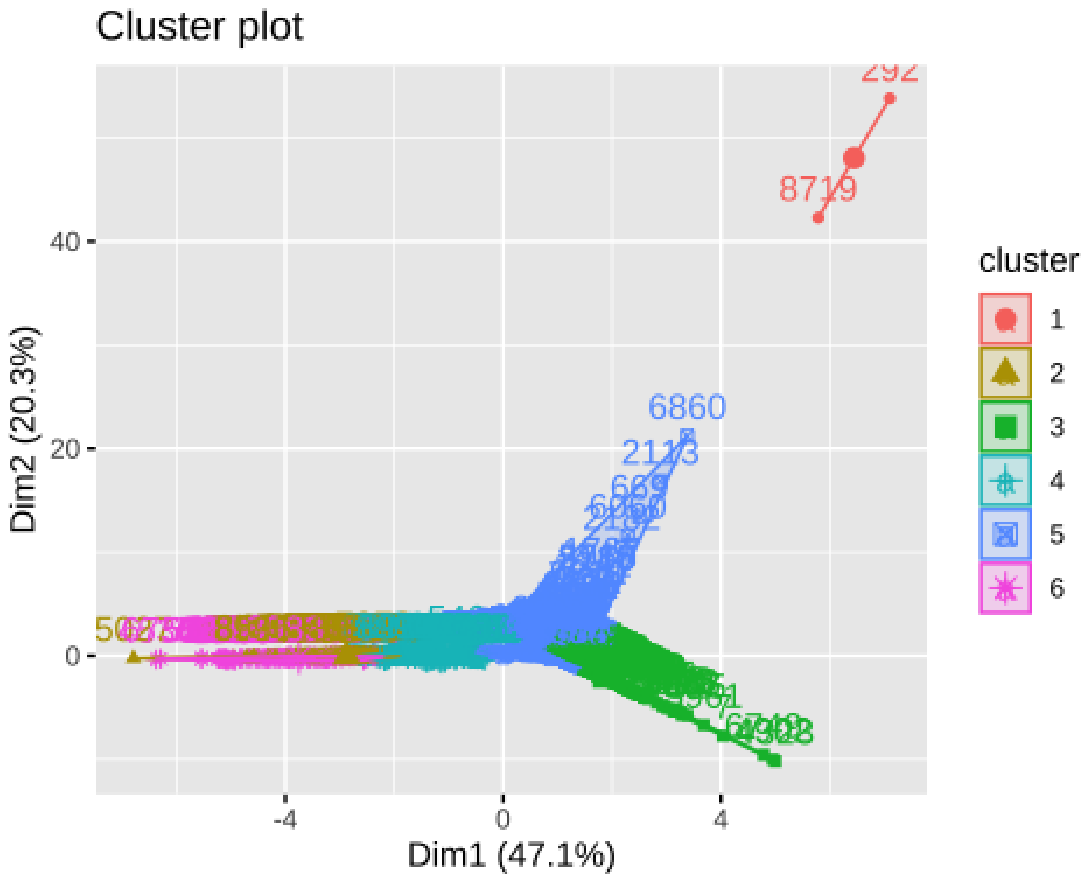


#9. Predicting clusters using 'K-Means' based on centers shown from silhouette method: # We've previously calculated the centers = 6 using silhouette method.

```
set.seed(9596)
kmeans_data <- kmeans(cluster_predict, centers = 6, nstart = 25)
```

#10.Plotting of clusters based on clusters formed with the numerical dataset:

```
fviz_cluster(kmeans_data, data= cluster_data)
```



#11. Binding the generated clusters to the initial numeric variables dataset: # Binding the values of the clusters to the original data set helps us identify where all data points belong to distinct clusters.

```
cluster_group<- kmeans_data$cluster
group_cluster <- cbind(cluster_data, cluster_group)
```

#12. Evaluating the middlemost value of every single cluster, i.e. the cluster median: # The aggregate function-Median assists us in determining the middle most value of each cluster.

```
aggregate(group_cluster,by=list(group_cluster$cluster_group), FUN="median")

##   Group.1 fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct
## 1       1               253.0               1.037              0.000
## 2       2             18225.5              23.686              2.890
## 3       3           2582356.0               1.030              0.000
## 4       4             25731.0              17.980              0.430
## 5       5             14162.0               1.030              0.000
## 6       6              3693.0              13.455              0.665
##   ash_content_pct fuel_cost_per_mmbtu cluster_group
## 1             0.0         1429.200000             1
## 2             8.8            2.340500             2
## 3             0.0            5.039403             3
## 4             6.1            2.238000             4
## 5             0.0            5.119532             5
## 6            40.0            2.155229             6
```

#Cluster 1: This cluster is a pattern since the median value of heat output is low but the cost is relatively high.

#Clusters 2 and Cluster 4 and cluster 6: display a high median value of fuel_mmbtu_per_unit with a lower median value of fuel_cost_per_mmbtu, indicating that this cluster generates high heat at a low cost. It also contains a substantial quantity of sulfur and ash.

#Cluster 3 and #Cluster 5: The median values of both clusters reveal little heat output and expense spent. The sulfur and ash production values are displayed as zero.

#13. To understand the clusters, bind the final cluster to every fuel_group_code: # It lets determine where all of the data points in the clustered data are grouped in terms of the fuel sources utilized.

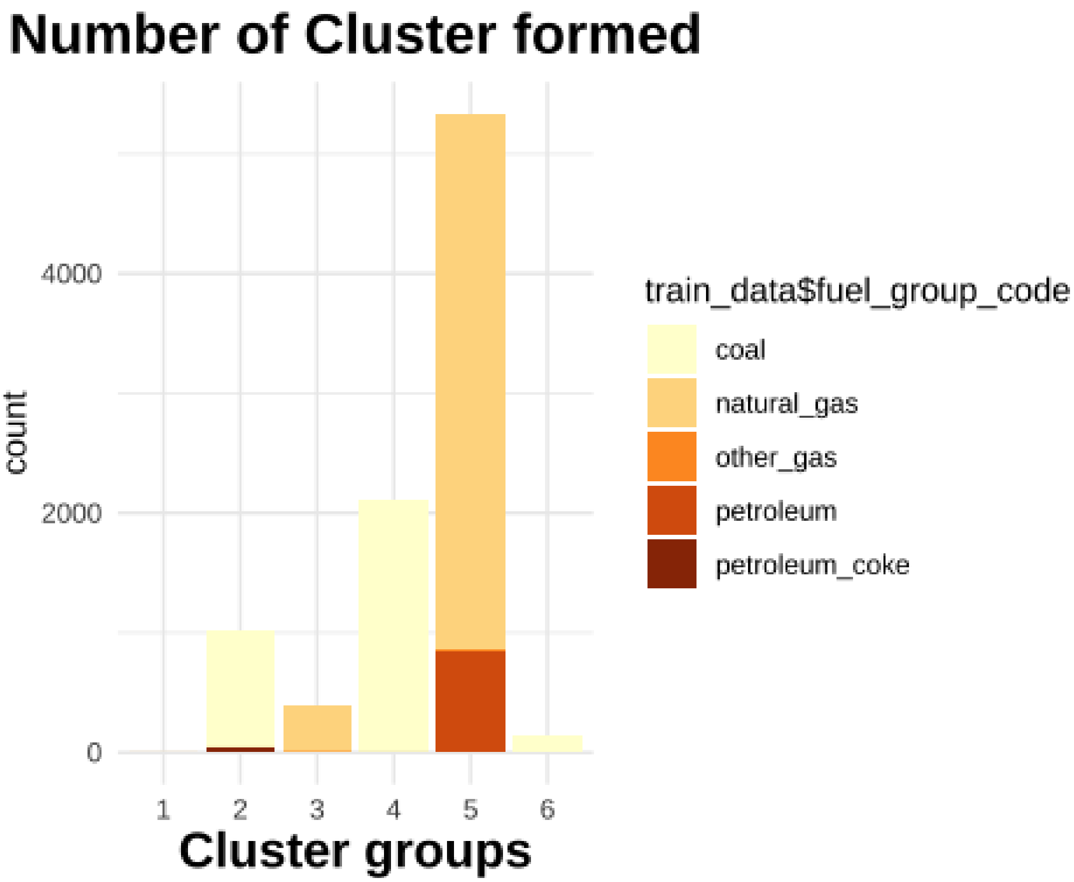
```
group_cluster$cluster_group <- as.factor(group_cluster$cluster_group)
final_cluster<- cbind(group_cluster, train_data$fuel_group_code)
head(final_cluster)

##   fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct
ash_content_pct
## 1                8569              26.048               1.81
7.5
## 2               43236               1.020               0.00
0.0
## 3              105552               1.012               0.00
0.0
## 4               16334               1.017               0.00
0.0
## 5                 893               5.712               0.00
0.0
## 6                2591               5.750               0.25
0.0
##   fuel_cost_per_mmbtu cluster_group train_data$fuel_group_code
## 1               2.873             2                       coal
## 2               2.780             5                natural_gas
## 3               6.469             5                natural_gas
## 4               8.147             5                natural_gas
## 5              17.632             5                  petroleum
## 6              16.829             5                  petroleum
```

#14.Visual presentation of number of clusters formed showed in form of ggplot2:

```
#esquisser()

ggplot(final_cluster) +
  aes(x = cluster_group, fill = `train_data$fuel_group_code`) +
  geom_bar() +
  scale_fill_brewer(palette = "YlOrBr", direction = 1) +
  labs(
    x = "Cluster groups",
    title = "Number of Cluster formed"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 18L,
    face = "bold",
    hjust = 0.5),
    axis.title.x = element_text(size = 16L,
    face = "bold")
  )
```



#15. The final dataset was filtered in order to determine what each cluster represents: # We can see from the silhouette that each cluster has been categorised based on the similarity of their data points. As a result, filtering and understanding a few data points might assist us in determining the cluster's general behavior. This might be utilized to identify the best cluster for our business target.

#a. Cluster 1 includes only three data points, indicating that it contains outliers since the heat production is minimal and the cost output is extremely high.

```
cluster1<-final_cluster %>% select(fuel_mmbtu_per_unit,fuel_cost_per_mmbtu,
cluster_group) %>% group_by(train_data$fuel_group_code) %>%
arrange(desc(fuel_mmbtu_per_unit)) %>% filter(cluster_group == 1) %>% head()
cluster1

## # A tibble: 2 × 4
## # Groups:   train_data$fuel_group_code [1]
##   fuel_mmbtu_per_unit fuel_cost_per_mmbtu cluster_group
train_data$fuel_group_…¹
##                 <dbl>               <dbl> <fct>         <chr>
## 1                1.04               1601. 1             natural_gas
## 2                1.03               1258. 1             natural_gas
## # … with abbreviated variable name ¹`train_data$fuel_group_code`
```

#b.From the below representation, Although cluster 2 offers a high heat production at a minimal cost, both coal and petroleum coke release sulfur and ash.

```
cluster_imp<-final_cluster %>%
select(fuel_mmbtu_per_unit,fuel_cost_per_mmbtu, sulfur_content_pct,
ash_content_pct , cluster_group, `train_data$fuel_group_code`) %>%
group_by(train_data$fuel_group_code) %>% arrange(desc(sulfur_content_pct))
%>% head()
cluster_imp

## # A tibble: 6 × 6
## # Groups:   train_data$fuel_group_code [2]
##   fuel_mmbtu_per_unit fuel_cost_per_mmbtu sulfur_conte…¹ ash_c…² clust…³
train…⁴
##                 <dbl>               <dbl>          <dbl>   <dbl> <fct>
<chr>
## 1                20.1               1.87            7.66    28.4 2
coal
## 2                28.6               2.13            6.93     0   2
petrol…
## 3                29.3               0.888           6.8      2.2 2
petrol…
## 4                29.8               0.954           6.5      1.7 2
petrol…
## 5                29.0               2.53            6.39     0.2 2
petrol…
## 6                27.5               1.32            6.23     0.7 2
petrol…
## # … with abbreviated variable names ¹sulfur_content_pct, ²ash_content_pct,
## #   ³cluster_group, ⁴`train_data$fuel_group_code`
```

#c.Since we now understood that the median values of Cluster 3 contain zero sulfur and ash emission, we may proceed from there. Although one data point displays a significant cost, we can see that their heat and cost are modest. This might be due to the presence of outliers in this cluster.

```
cluster3<-final_cluster %>% select(fuel_mmbtu_per_unit,fuel_cost_per_mmbtu,
cluster_group, `train_data$fuel_group_code`) %>%
filter(train_data$fuel_group_code =='natural_gas')
%>%arrange(desc(fuel_mmbtu_per_unit)) %>% filter(cluster_group == 2) %>%
head()
cluster3
```

```
## [1] fuel_mmbtu_per_unit        fuel_cost_per_mmbtu
## [3] cluster_group              train_data$fuel_group_code
## <0 rows> (or 0-length row.names)
```

#d.Cluster 4 represents that coal is a major source of heat provided at a low cost.

```
cluster4<-final_cluster %>% select(fuel_mmbtu_per_unit,fuel_cost_per_mmbtu,
cluster_group) %>% group_by(train_data$fuel_group_code) %>%
arrange(desc(fuel_mmbtu_per_unit)) %>% filter(cluster_group == 6)
cluster4
```

```
## # A tibble: 136 × 4
## # Groups:   train_data$fuel_group_code [1]
##    fuel_mmbtu_per_unit fuel_cost_per_mmbtu cluster_group
train_data$fuel_group…¹
##                  <dbl>               <dbl> <fct>         <chr>
##  1                20.5                2.55 6             coal
##  2                20.0                3.40 6             coal
##  3                19.9                2.17 6             coal
##  4                19.7                1.52 6             coal
##  5                19.7                2.7  6             coal
##  6                19.6                2.05 6             coal
##  7                19.0                2.15 6             coal
##  8                19.0                2.13 6             coal
##  9                18.8                1.98 6             coal
## 10                18.7                2.89 6             coal
## # … with 126 more rows, and abbreviated variable name
## #   ¹`train_data$fuel_group_code`
```

#e.This cluster displays uniform characteristics with low heat and cost, and all data points in this cluster are expressed by natural gas. This might be referred to as an ideal cluster for recommending current company problems.

```
cluster5<-final_cluster %>% select(fuel_mmbtu_per_unit,fuel_cost_per_mmbtu,
cluster_group) %>% group_by(train_data$fuel_group_code) %>%
arrange(desc(fuel_mmbtu_per_unit)) %>% filter(cluster_group == 3) %>% head()
cluster5
```

```
## # A tibble: 6 × 4
## # Groups:   train_data$fuel_group_code [1]
##   fuel_mmbtu_per_unit fuel_cost_per_mmbtu cluster_group
train_data$fuel_group_…¹
##                 <dbl>               <dbl> <fct>         <chr>
## 1                1.10                4.33 3             natural_gas
## 2                1.1                 5.56 3            natural_gas
## 3                1.09                3.69 3            natural_gas
## 4                1.09                2.50 3            natural_gas
## 5                1.09                5.59 3            natural_gas
## 6                1.09                3.02 3            natural_gas
## # … with abbreviated variable name ¹`train_data$fuel_group_code`
```

#This cluster is demonstrates that petroleum has a high heat production at a low cost.

```
cluster5<-final_cluster %>% select(fuel_mmbtu_per_unit,fuel_cost_per_mmbtu,
cluster_group) %>% group_by(train_data$fuel_group_code) %>%
arrange(desc(fuel_mmbtu_per_unit)) %>% filter(cluster_group == 5) %>% head()
cluster5

## # A tibble: 6 × 4
## # Groups:   train_data$fuel_group_code [1]
##   fuel_mmbtu_per_unit fuel_cost_per_mmbtu cluster_group
train_data$fuel_group_…¹
##                 <dbl>               <dbl> <fct>         <chr>
## 1                6.60                5.71 5             petroleum
## 2                6.55               15.4  5             petroleum
## 3                6.53               15.6  5             petroleum
## 4                6.49                9.25 5             petroleum
## 5                6.47                8.99 5             petroleum
## 6                6.44               12.0  5             petroleum
## # … with abbreviated variable name ¹`train_data$fuel_group_code`
```

#f.This cluster is similar to Cluster 4 in that it is dominated by coal, which has a high heat production and minimal costs.

```
cluster6<-final_cluster %>% select(fuel_mmbtu_per_unit,fuel_cost_per_mmbtu,
cluster_group) %>% group_by(train_data$fuel_group_code) %>%
arrange(desc(fuel_mmbtu_per_unit)) %>% filter(cluster_group == 4) %>% head()
cluster6

## # A tibble: 6 × 4
## # Groups:   train_data$fuel_group_code [2]
##   fuel_mmbtu_per_unit fuel_cost_per_mmbtu cluster_group
train_data$fuel_group_…¹
##                 <dbl>               <dbl> <fct>         <chr>
## 1                29                  3.75 4             coal
## 2                29                  3.34 4             coal
## 3                28.8                3.42 4             petroleum_coke
## 4                28.0                4.22 4             coal
## 5                27.9                3.53 4             coal
## 6                27.8                4.25 4             coal
## # … with abbreviated variable name ¹`train_data$fuel_group_code`
```