Kiran Kour

Professor Dr.Murali Shanker

Kent State University

MIS-64060:Fundamentals of Machine Learning

11 December 2022

This final exam aims to apply the appropriate machine learning technique to select the best segmentation and help understand power generation in the US.

Introduction

The ***PUDL Project*** is an open source data processing pipeline that makes US energy data easier to access and use programmatically.

Hundreds of gigabytes of valuable data are published by US government agencies, but it's often difficult to work with. PUDL takes the original spreadsheets, CSV files, and databases and turns them into a unified resource. This allows users to spend more time on novel analysis and less time on data preparation.

For this project, we will use one specific table, the monthly fuel contract information, purchases, and costs reported in EIA-923 Schedule 2, Part A, for our analysis. This table contains 608,565 rows and 20 variables, though note that several variables have significant missing values.

Random sampling of the dataset

```
library(dplyr)
set.seed(2222)
fuel_data <- sample_frac(fuel,0.02)
```

The imported data has been randomly sampled about 2%

Checking for the NA values

```
colMeans(is.na(fuel_data))

##                                    rowid
##                                0.0000000
##                             plant_id_eia
##                                0.0000000
##                       plant_id_eia_label
##                                0.0000000
##                              report_date
##                                0.0000000
##                       contract_type_code
##                                0.0000000
##                 contract_type_code_label
##                                0.0000000
##                 contract_expiration_date
##                                0.0000000
##                       energy_source_code
##                                0.0000000
##                 energy_source_code_label
##                                0.0000000
##                      fuel_type_code_pudl
##                                0.0000000
##                          fuel_group_code
##                                0.0000000
##                             mine_id_pudl
##                                0.6357736
##                       mine_id_pudl_label
##                                0.6357736
##                            supplier_name
##                                0.0000000
##                      fuel_received_units
##                                0.0000000
##                      fuel_mmbtu_per_unit
##                                0.0000000
##                       sulfur_content_pct
##                                0.0000000
##                          ash_content_pct
##                                0.0000000
##                      mercury_content_ppm
##                                0.4771177
##                      fuel_cost_per_mmbtu
##                                0.3268425
##         primary_transportation_mode_code
##                                0.0000000
##   primary_transportation_mode_code_label
##                                0.0000000
##       secondary_transportation_mode_code
##                                0.0000000
## secondary_transportation_mode_code_label
##                                0.0000000
##               natural_gas_transport_code
##                                0.0000000
##  natural_gas_delivery_contract_type_code
##                                0.0000000
##                     moisture_content_pct
##                                0.8458631
##                     chlorine_content_ppm
##                                0.8458631
##                            data_maturity
##                                 0.0000000
##                       data_maturity_label
##                                 0.0000000
```

We can see that 6 variables are having NA values.

Imputing the missing values of Fuel cost per mmbtu and Mercury content based on the mean of the columns

```
fuel_data <- fuel_data %>%
mutate_at(vars(fuel_cost_per_mmbtu,mercury_content_ppm), ~replace_na(.,
mean(., na.rm=TRUE)))
```

Using 75% of the sampled data as the training set, and the rest as the test set

```
set.seed(1234)
training_fuel  <- fuel_data %>% dplyr::sample_frac(0.75)


testing_fuel   <- dplyr::anti_join(fuel_data, training_fuel,
                                   by= "rowid")
```

Excluding the columns which will not be included in the analysis

```
fuel_data <- subset (fuel_data, select = -c(1:4,6:10,12:13,21:30))

training_fuel <- subset(training_fuel, select= -c(1:4,6:10,12:13,21:30))

testing_fuel <- subset(testing_fuel,select = -c(1:4,6:10,12:13,21:30))
```

Creating a new column named gases_emission which is an aggregate of sulphur,ash and mercury content

```
fuel_data$gases_emission=rowSums(cbind(fuel_data$sulfur_content_pct,fuel_data
$ash_content_pct,fuel_data$mercury_content_ppm ),na.rm=TRUE)


training_fuel$gases_emission=rowSums(cbind(training_fuel$sulfur_content_pct,t
raining_fuel$ash_content_pct,testing_fuel$mercury_content_ppm ),na.rm=TRUE)
```

```
testing_fuel$gases_emission=
rowSums(cbind(testing_fuel$sulfur_content_pct,testing_fuel$ash_content_pct,te
sting_fuel$mercury_content_ppm ),na.rm=TRUE)
```

Because sulfur, ash and mercury are now combined in a new column, I will be excluding the individual column of sulphur and mercury from the datasets.

```
fuel_data <- subset (fuel_data, select = -c(6:8))

training_fuel <- subset(training_fuel, select= -c(6:8))

testing_fuel <- subset(testing_fuel,select = -c(6:8))
```

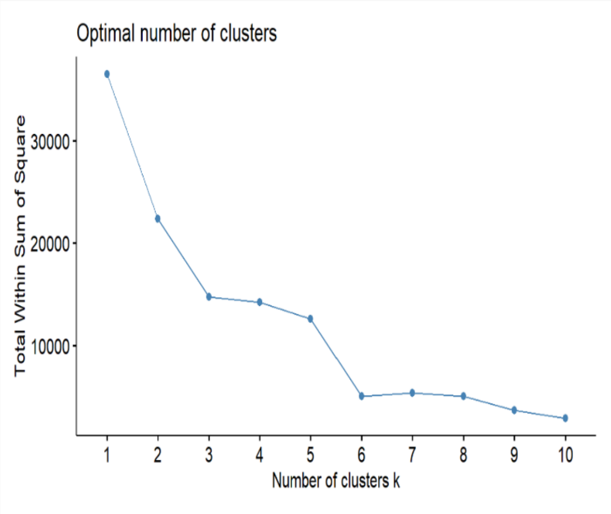
Normalization is an essential step in the data preparation. It will allow the dataset to have the same scale,and it will help to reduce the bias and its spread. I will be normalizing the training and testing fuel dataset.

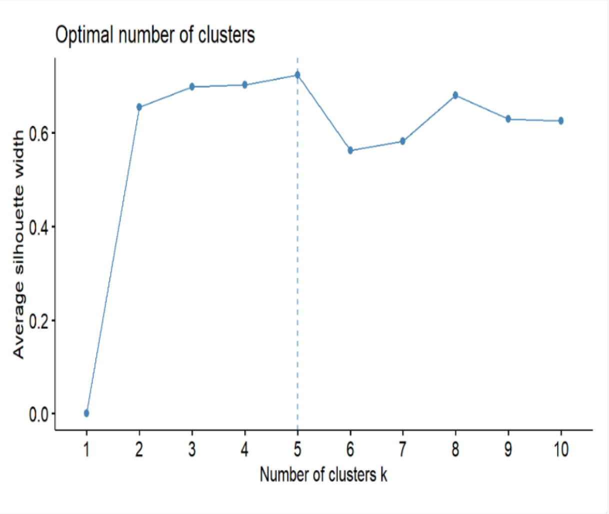
```
Train_norm <- scale(training_fuel[,-c(1:3)])
summary(Train_norm)

##  fuel_received_units fuel_mmbtu_per_unit fuel_cost_per_mmbtu
gases_emission
##  Min.   :-0.3365     Min.   :-0.9109     Min.   :-0.12250    Min.   :-
0.5828
##  1st Qu.:-0.3314     1st Qu.:-0.8091     1st Qu.:-0.07100    1st Qu.:-
0.5828
##  Median :-0.3060     Median :-0.8052     Median :-0.03175    Median :-
0.5816
##  Mean   : 0.0000     Mean   : 0.0000     Mean   : 0.00000    Mean   :
0.0000
##  3rd Qu.:-0.1935     3rd Qu.: 0.8991     3rd Qu.:-0.00500    3rd Qu.:
0.3326
##  Max.   :15.4387     Max.   : 2.2241     Max.   :87.21069    Max.   :
8.8411

Test_norm <- scale(testing_fuel[,-c(1:3)])
summary(Test_norm)

##  fuel_received_units fuel_mmbtu_per_unit fuel_cost_per_mmbtu
gases_emission
##  Min.   :-0.3302     Min.   :-0.9045     Min.   :-0.42326    Min.   :-
0.5796
##  1st Qu.:-0.3248     1st Qu.:-0.8110     1st Qu.:-0.23159    1st Qu.:-
0.5796
##  Median :-0.3003     Median :-0.8072     Median :-0.06647    Median :-
0.5784
##  Mean   : 0.0000     Mean   : 0.0000     Mean   : 0.00000    Mean   :
```

```
0.0000
##  3rd Qu.:-0.1898     3rd Qu.: 0.8878     3rd Qu.: 0.06703    3rd Qu.:
0.3210
##  Max.   :17.3547     Max.   : 2.0413     Max.   :48.31737    Max.   :
8.0910
```

Using WSS and Silhouette method to find the optimal K value

The optimal K values after employing the WSS method is 3 and Silhouette method is 5





```
set.seed(1234)
k3_WSS<- kmeans(Train_norm, centers=3, nstart = 25)
table(k3_WSS$cluster)

## 
##    1    2    3 
## 5397  370 3361
```

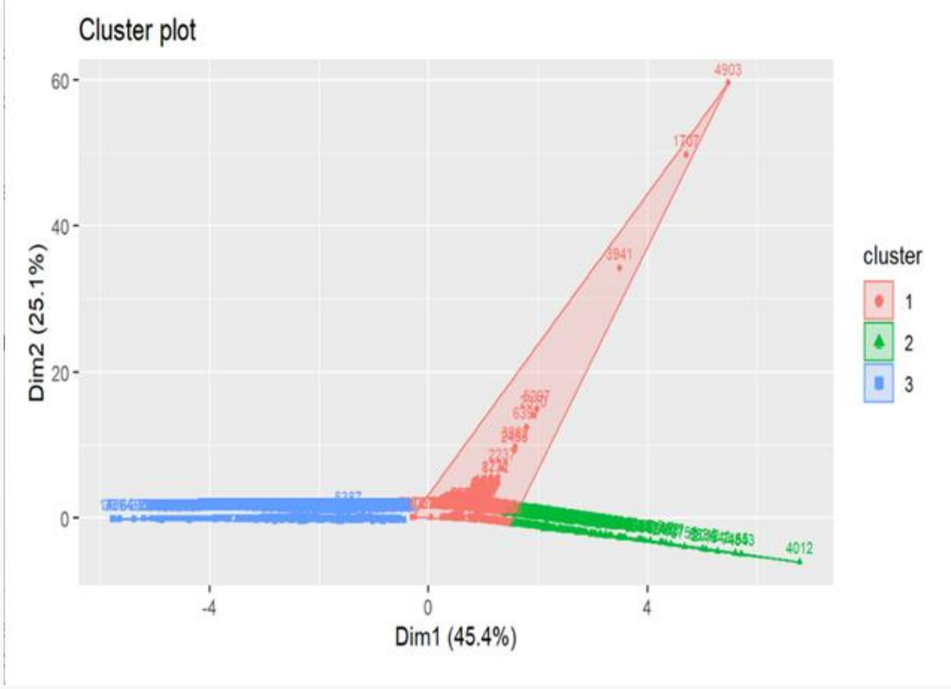
This output shows 3 clusters of sizes 5397, 370, 3361.

```
set.seed(1234)
k5_Sil <- kmeans(Train_norm,centers= 5,nstart=25)
table(k5_Sil$cluster)

## 
##    1    2    3    4    5 
## 3218  370    3  5394  143
```

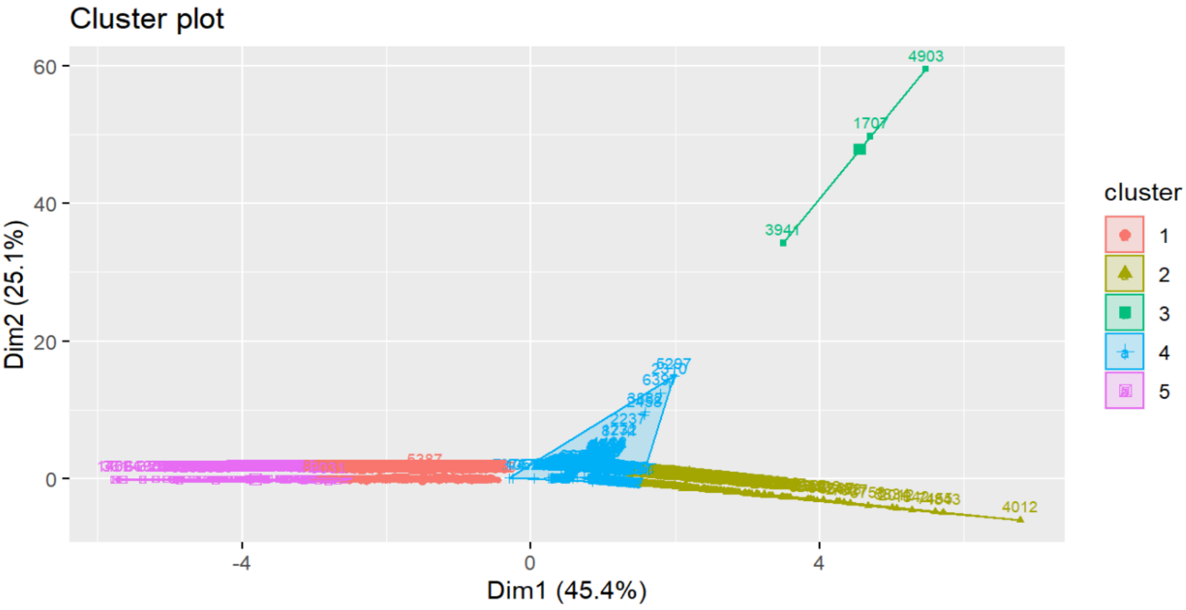
This output shows 5 clusters of sizes 3218, 370, 3, 5394, 143.

Visualizing the clusters formed by WSS method and Silhouette method

```
fviz_cluster(k3_WSS, data = Train_norm, pointsize = 1, labelsize = 7)
```



```
fviz_cluster(k5_Sil, data = Train_norm, pointsize = 1, labelsize = 7)
```



Transforming the data to further analyze using the unnormalized data

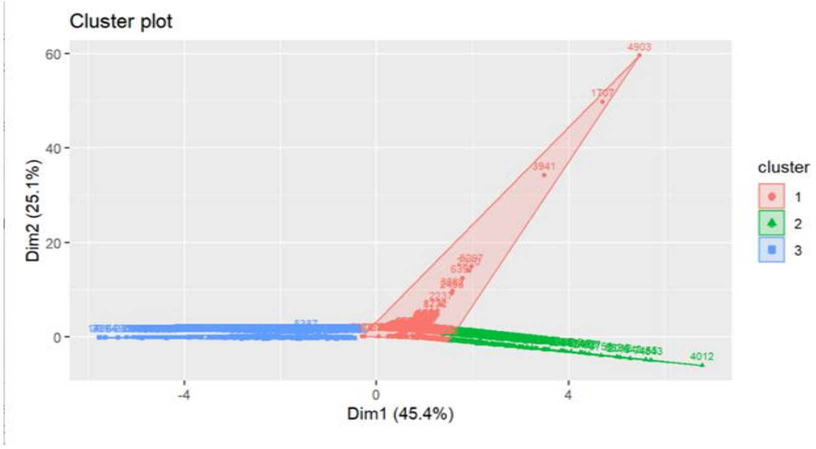
```
#Data Transformation for WSS

set.seed(1234)
fuel_data1 <- cbind(training_fuel[,-c(1:3)], k3_WSS$cluster)
colnames(fuel_data1) <- c("fuel_received_units",
"fuel_mmbtu_per_unit","fuel_cost_per_mmbtu","gases_emission","Groups")
fuel_data1$Groups <- as.numeric(fuel_data1$Groups)
```

Here, I will be using aggregate function to calculate the mean of each variables in the cluster to analyze the clusters as to which one will be optimal to consider.

```
set.seed(1234)
fuel_data1_Median<- aggregate(fuel_data1,by=list(k3_WSS$cluster),FUN=median)
fuel_data1_Median

##   Group.1 fuel_received_units fuel_mmbtu_per_unit fuel_cost_per_mmbtu
## 1       1               16654               1.030            6.905031
## 2       2             2586365               1.029            6.905031
## 3       3               22033              22.600            2.705000
##   gases_emission Groups
## 1    0.008215271      1
## 2    0.008215271      2
## 3   10.360000000      3
```

The highest units received are by Cluster 2, followed by Cluster 3, and lowest by Cluster 1. The value of heat content of fuels, the average cost of fuels per MMBtu of heat content, and the content of sulfur, ash, and mercury (gases emission) are the same in Clusters 1 and 2 and the highest in Cluster 3.

I will not continue my analysis based on the clusters formed by WSS. The clusters formed by the WSS method will not help me answer the questions raised to understand power generation in the US. From the Cluster plot below:



We can see that ***Cluster 1 has outliers*** which means they are away from other data values and hence ***disturb the overall distribution of the dataset.*** This is usually assumed as an abnormal distribution of the data values. Therefore, these outliers in the Cluster 1 can ***distort predictions and affect the accuracy of this cluster.*** Moving toward ***Cluster 2*** shows us that this cluster ***also has an outlier***, so it would not make sense to consider this cluster for our segmentation.

Cluster 3, on the other hand, ***is an expensive cluster both in cost and from the environmental point of view (High content of Sulfur, Ash, and Mercury).*** It will not help me justify the reasons to understand the best segmentation.

Transforming the data to further analyze using the unnormalized data

```
#Data Transformation for Silhoutte

set.seed(1234)
fuel_data2 <- cbind(training_fuel[,-c(1:3)], k5_Sil$cluster)
colnames(fuel_data2) <- c("fuel_received_units",
"fuel_mmbtu_per_unit","fuel_cost_per_mmbtu","gases_emission","Groups")
fuel_data2$Groups <- as.numeric(fuel_data2$Groups)
```

Using the same aggregate function to calculate the mean of the numerical variables in the clusters formed by the silhouette method which will be used to further analyze the clusters

```
set.seed(1234)
fuel_data2_Median<- aggregate(fuel_data2,by=list(k5_Sil$cluster),FUN=median)
fuel_data2_Median
```

```
##   Group.1 fuel_received_units fuel_mmbtu_per_unit fuel_cost_per_mmbtu
## 1       1             23170.5              22.865            2.633000
## 2       2           2586365.0               1.029            6.905031
## 3       3                10.0               1.041         1699.808000
## 4       4             16719.0               1.030            6.905031
## 5       5              2950.0              14.100            6.905031
##   gases_emission Groups
## 1    10.140000000      1
## 2     0.008215271      2
## 3     0.008215271      3
## 4     0.008215271      4
## 5    38.718215271      5
```

Let us first interpret the fuel received units by each cluster. The maximum number of units received are by **Cluster 2- The Expensive Power**, which comprised only Natural gas and other gas. **Clusters 3- The Outliers*** and **5- Black Diamond** are at a lower level with just 10 and 2950 units, respectively. **Clusters 1- The Thermogenic Plants** and **4- Clean Power Future** received 23170.5 and 16719 units of fuel, respectively.

The Heat Content of the fuels is maximum for Cluster 1- The Thermogenic Plants (Coal and Petroleum coke) and second highest in Cluster 5- Black Diamond (Coal). In contrast, Cluster 2- The Expensive Power (Natural gas and other gas), Cluster 3- The Outliers (Natural gas), and Cluster 4- Clean Power Future (Natural gas, Coal, Petroleum, and Other gas) have somewhat the same heat content.

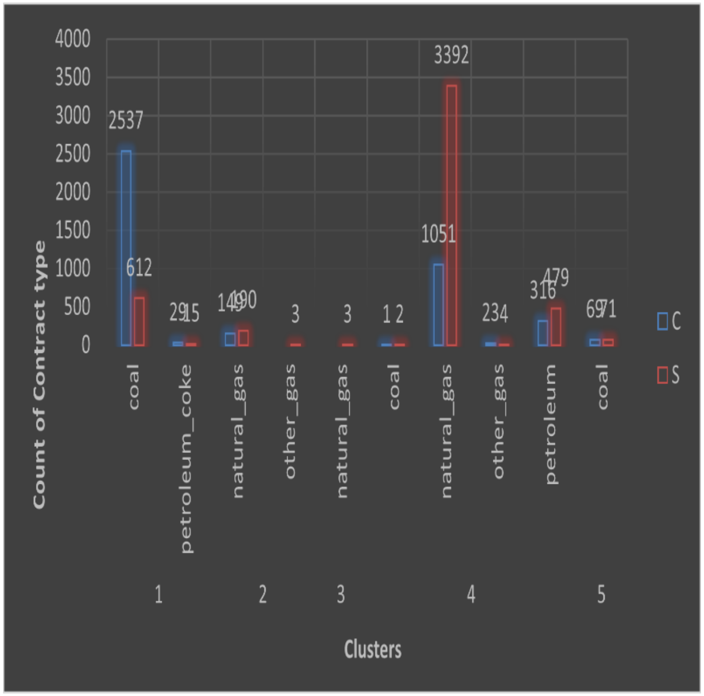
Coming next to **the Average cost per MMBtu of heat content**, we can see that Cluster 3- The Outliers, comprising only Natural gas, has the maximum Cost per MMBtu of heat content; comparatively, Cluster 1- The Thermogenic Plants containing Coal and Petroleum Coke, has less cost per MMBtu of heat content as compared to all the Clusters. Rest three clusters- Cluster 2- The Expensive Power, 4- Clean Power Future, and 5- Black Diamond have the exact average heat content cost per MMBtu.

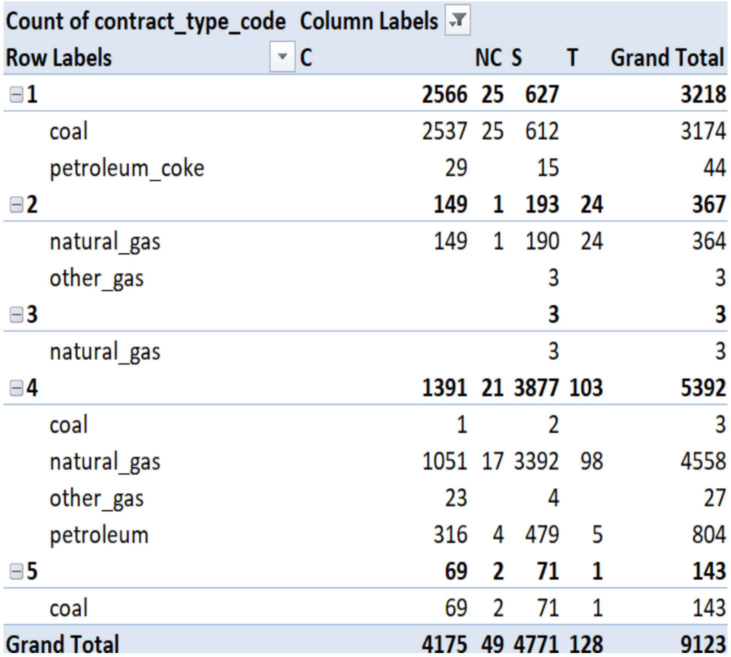
Analyzing from the table, the sulfur, mercury, and ash content **(gases emissions**) is the maximum for Cluster 5- Black Diamond. Minimum content can be seen in Cluster 1- The Thermogenic Plants, whereas Clusters 2- The Expensive Power, 4- Clean Power Future, and 3- The Outliers have the same sulfur, ash, and mercury content which is lowest compared to Clusters 5- Black Diamond and 1- The Thermogenic Plants.

Interpreting and visualizing a pattern in the clusters with respect to the Categorical variables

```
set.seed(1234)
fuel_data2 <- cbind(training_fuel, k5_Sil$cluster)
colnames(fuel_data2) <-
c("contract_type_code","fuel_group_code","supplier_name","fuel_received_units
", "fuel_mmbtu_per_unit", "fuel_cost_per_mmbtu", "gases_emission","Groups")
fuel_data2$Groups <- as.numeric(fuel_data2$Groups)
```

Now, these interpretations were based on the Numerical Variables in the dataset; I will now analyze the clusters based on the categorical variables:

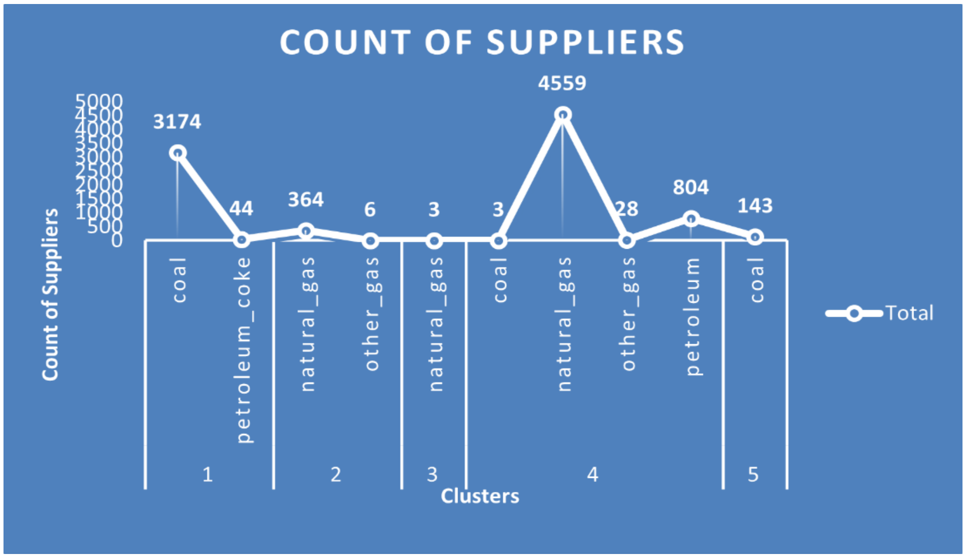




Count of contract_type_code	Column Labels				
Row Labels	C	NC	S	T	Grand Total
⊟1	2566	25	627		3218
coal	2537	25	612		3174
petroleum_coke	29		15		44
⊟2	149	1	193	24	367
natural_gas	149	1	190	24	364
other_gas			3		3
⊟3			3		3
natural_gas			3		3
⊟4	1391	21	3877	103	5392
coal	1		2		3
natural_gas	1051	17	3392	98	4558
other_gas	23		4		27
petroleum	316	4	479	5	804
⊟5	69	2	71	1	143
coal	69	2	71	1	143
Grand Total	4175	49	4771	128	9123

These fuels were purchased on some contract types, namely Contract (C), Spot Purchase (S), New Contract (NC), and Tolling Agreement (T). Analyzing figure 15 tells us which fuels were purchased under contract and spot purchase. The highest count of purchases under Contract and spot purchases are in Cluster 4- Clean Power Future, with Natural gas being purchased maximum times. The second highest purchasing Cluster is Cluster 1- The Thermogenic Plants, with coal being the most increased purchasing fuel within this cluster.

The Contract table also gives us an understanding of purchasing power between these clusters. Cluster 4- Clean Power Future is the highest purchasing cluster, followed by Cluster 1- The Thermogenic Plants. Cluster 3- The Outliers only have a purchase count of 3 under spot purchase. Cluster 2- The Expensive Cluster and Cluster 5- Black Diamond is also comparatively less than Cluster 1- The Thermogenic Plants and 4- Clean Power Future, with 367 and 143 purchase counts, respectively.



Let’s have a look at the supplier counts. The graph shows that Cluster 4- Clean Power Future has the maximum number of fuel suppliers (5394), the highest for natural gas. Coming next is Cluster 1- The Thermogenic Plants, with 3218 suppliers. Cluster 3- The Outliers has just three suppliers, Cluster 2- The Expensive Power has 370 suppliers, and Cluster 5- Black Diamond accounts for 143 suppliers.

Names of the Clusters are as follows:

Clusters 1- The Thermogenic Plants

Cluster 2- The Expensive Power

Cluster 3- The Outliers

Cluster 4- Clean Power Future

Cluster 5- Black Diamond