Kiran Kour

Professor Murali Shanker

Kent State University

MIS-64060:Fundamentals of Machine Learning

11 December 2022

The objective of this final exam is to apply the appropriate machine learning technique to select the best segmentation and help understand power generation in the US.

**Introduction**

Warning: package 'factoextra' was built under R version 4.2.2

Warning: package 'flexclust' was built under R version 4.2.2

Warning: package 'gridExtra' was built under R version 4.2.2

Warning: package 'cowplot' was built under R version 4.2.2

Warning: package 'writexl' was built under R version 4.2.2

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. I

***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,training\_fuel$ash\_content\_pct,testing\_fuel$mercury\_content\_ppm ),na.rm=TRUE)

## Warning in cbind(training\_fuel$sulfur\_content\_pct,  
## training\_fuel$ash\_content\_pct, : number of rows of result is not a multiple of  
## vector length (arg 3)

testing\_fuel$gases\_emission= rowSums(cbind(testing\_fuel$sulfur\_content\_pct,testing\_fuel$ash\_content\_pct,testing\_fuel$mercury\_content\_ppm ),na.rm=TRUE)

*Because sulphur, 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.***

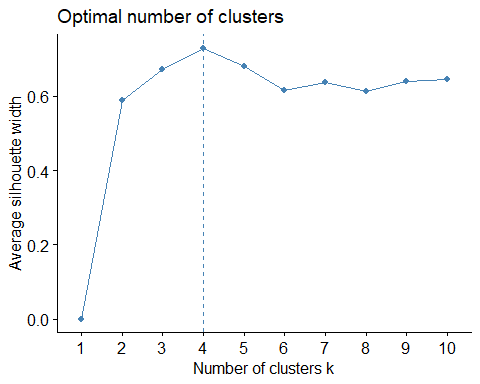
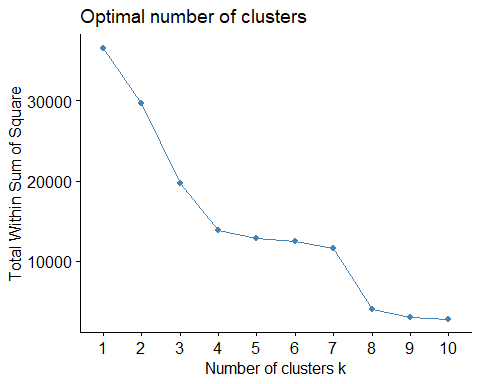
Train\_norm <- scale(training\_fuel[,-c(1:3)])  
summary(Train\_norm)

## 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   
## 5290 3403 435

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

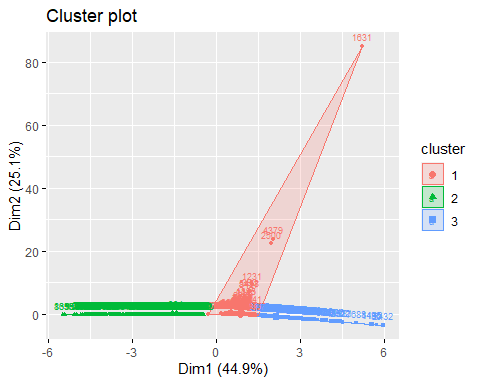
set.seed(1234)  
k5\_Sil <- kmeans(Train\_norm,centers= 5,nstart=25)  
table(k5\_Sil$cluster)

##   
## 1 2 3 4 5   
## 5069 98 1 3403 557

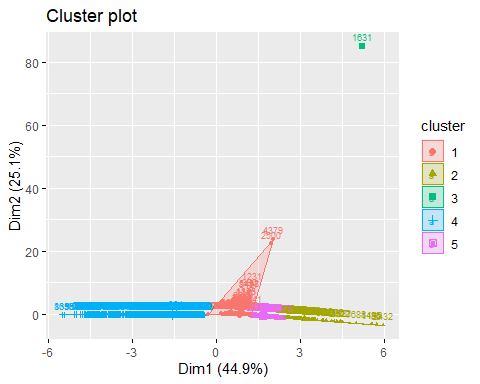
*This output shows 5 clusters of sizes 3218, 370, 3, 5394, 143.*

***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)

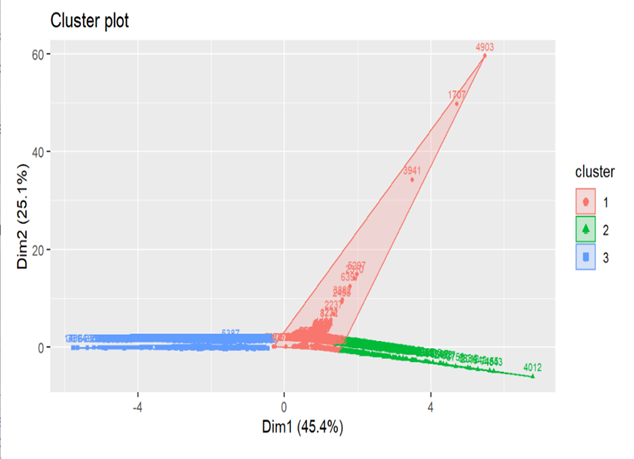
***Here, I will be using aggregate function to calculate the mean of each variables in the cluster to analyze the clusters as to which one will be optimal to consider.***

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 15504.5 1.030 6.185587  
## 2 2 21347.0 22.582 2.727000  
## 3 3 2396834.0 1.030 6.185587  
## gases\_emission Groups  
## 1 0.008453646 1  
## 2 10.450000000 2  
## 3 0.008453646 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 12584 1.030 6.185587  
## 2 2 4604376 1.029 6.185587  
## 3 3 53 1.019 4588.417000  
## 4 4 21347 22.582 2.727000  
## 5 5 1722421 1.030 6.185587  
## gases\_emission Groups  
## 1 0.008453646 1  
## 2 0.000000000 2  
## 3 0.000000000 3  
## 4 10.450000000 4  
## 5 0.000000000 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.*

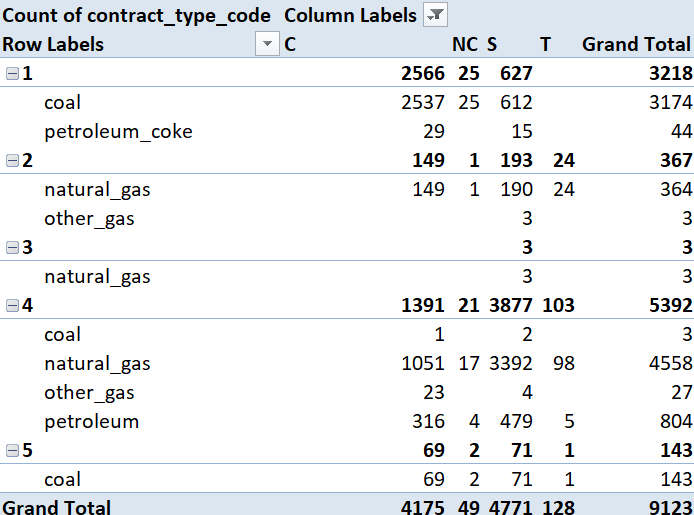
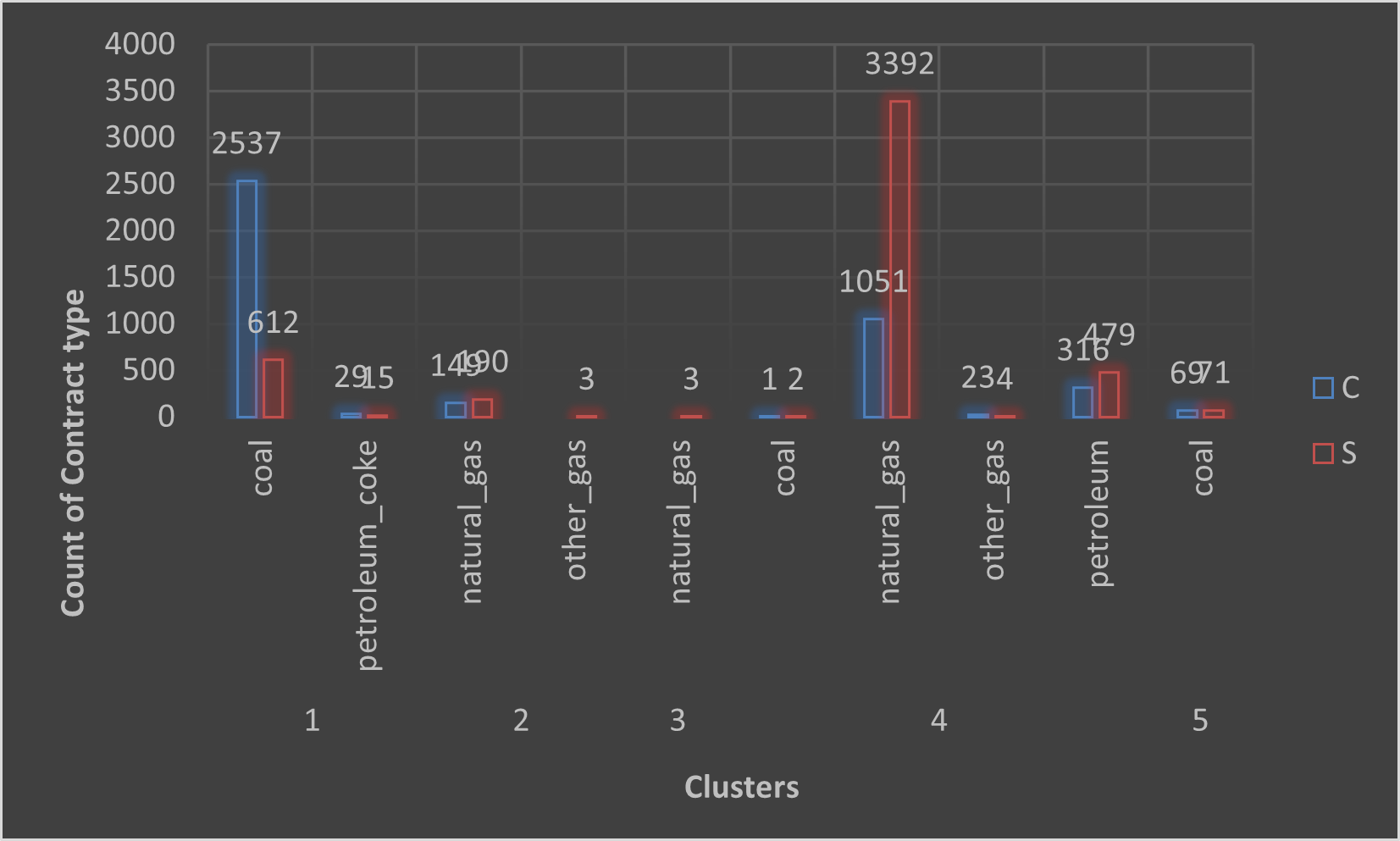
*Coming next to the Average cost per MMBtu of heat content, we can see that Cluster 3- The Outliers, comprising only Natural gas, has the maximum Cost per MMBtu of heat content; comparatively, Cluster 1- The Thermogenic Plants containing Coal and Petroleum Coke, has less cost per MMBtu of heat content 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***

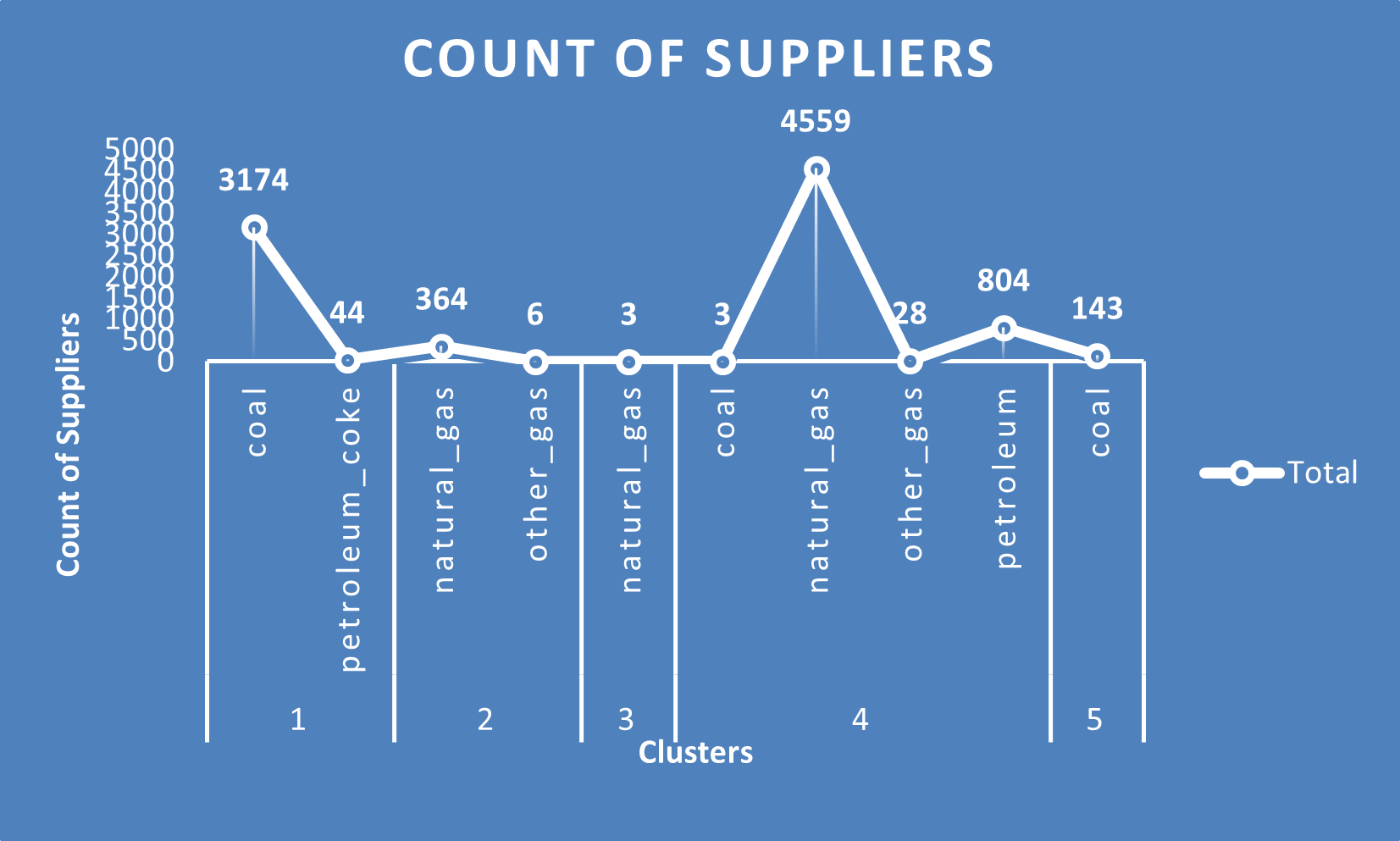
set.seed(1234)  
fuel\_data2 <- cbind(training\_fuel, k5\_Sil$cluster)  
colnames(fuel\_data2) <- c("contract\_type\_code","fuel\_group\_code","supplier\_name","fuel\_received\_units", "fuel\_mmbtu\_per\_unit", "fuel\_cost\_per\_mmbtu", "gases\_emission","Groups")  
fuel\_data2$Groups <- as.numeric(fuel\_data2$Groups)

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



*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.*