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 pipeline" "alabama coal" ...  
## $ fuel\_received\_units : int 259412 52241 2783619 25397 764 603 2341 8869 75442 206741 ...  
## $ 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

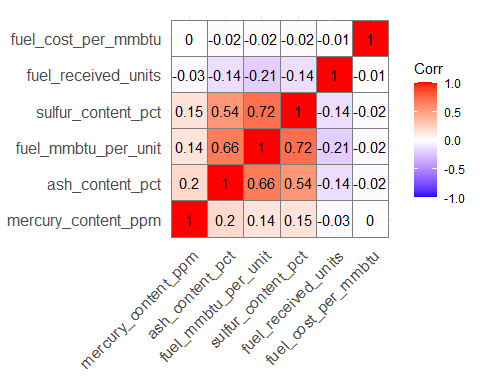
#Normalization  
Normalized\_Data <- scale(Fuel\_Data\_part[,-c(1:4)])  
Normalized\_Train <- scale(Train\_Data[,-c(1:4)])  
summary(Normalized\_Train)

## 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)

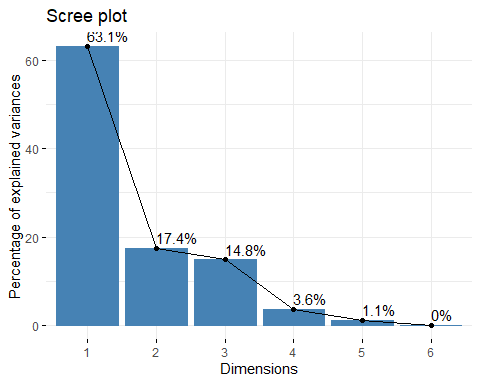
## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 0.7772648 0.4083527 0.3768495 0.18664026 0.10215463  
## Proportion of Variance 0.6305096 0.1740302 0.1482141 0.03635501 0.01089105  
## Cumulative Proportion 0.6305096 0.8045398 0.9527539 0.98910895 1.00000000  
## 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.*

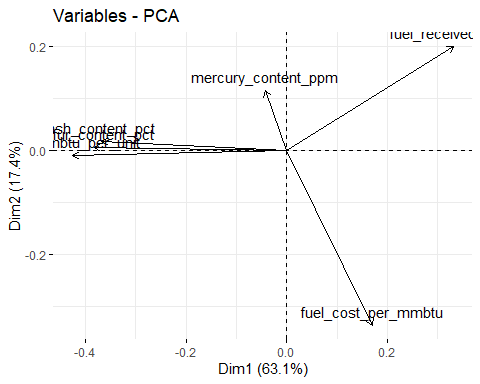
data.pca$loadings[, 1:3]

## 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)

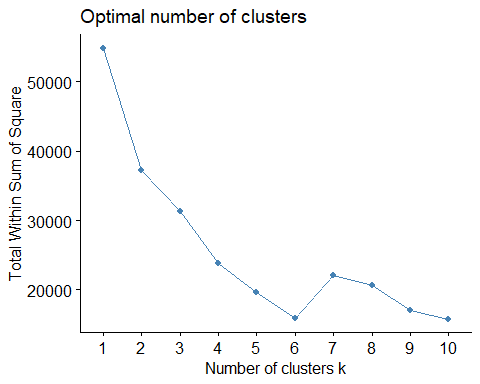


# Graph of the variables  
fviz\_pca\_var(data.pca, col.var = "black")

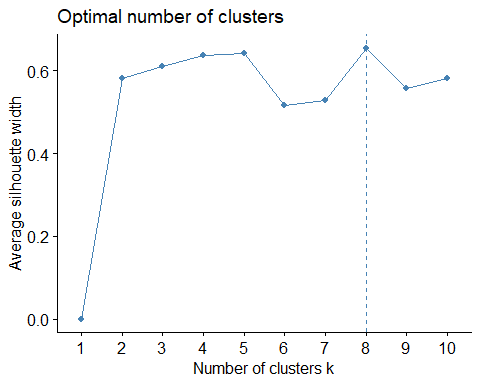
 *All the variables that are grouped together are positively correlated to each other. The higher the distance between the variable and the origin, the better represented that variable is. The variables that are negatively correlated are displayed to the opposite sides of the biplot’s origin.*

*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)

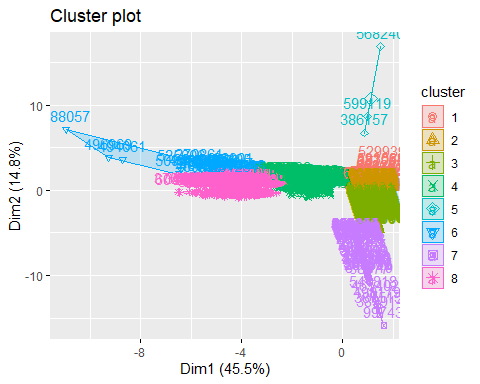
##   
##   
## +---------------------+---------------------+--------------------+-----------------+---------------------+---------------------+  
## | fuel\_received\_units | fuel\_mmbtu\_per\_unit | sulfur\_content\_pct | ash\_content\_pct | mercury\_content\_ppm | fuel\_cost\_per\_mmbtu |  
## +=====================+=====================+====================+=================+=====================+=====================+  
## | -0.3177 | -0.3088 | -0.3273 | -0.5447 | -0.1168 | 0.08907 |  
## +---------------------+---------------------+--------------------+-----------------+---------------------+---------------------+  
## | -0.1453 | -0.7948 | -0.5176 | -0.5454 | -0.1168 | -0.02138 |  
## +---------------------+---------------------+--------------------+-----------------+---------------------+---------------------+  
## | 2.58 | -0.7946 | -0.5176 | -0.5454 | -0.1168 | -0.02804 |  
## +---------------------+---------------------+--------------------+-----------------+---------------------+---------------------+  
## | -0.2462 | 1.185 | 0.07025 | 0.6753 | 0.1048 | -0.03446 |  
## +---------------------+---------------------+--------------------+-----------------+---------------------+---------------------+  
## | -0.326 | -0.7942 | -0.5176 | -0.5454 | -0.1168 | 49.88 |  
## +---------------------+---------------------+--------------------+-----------------+---------------------+---------------------+  
## | -0.2965 | 1.202 | 1.93 | 2.807 | 12.6 | -0.03051 |  
## +---------------------+---------------------+--------------------+-----------------+---------------------+---------------------+  
## | 7.955 | -0.8008 | -0.5176 | -0.5454 | -0.1168 | -0.0264 |  
## +---------------------+---------------------+--------------------+-----------------+---------------------+---------------------+  
## | -0.2772 | 1.43 | 2.281 | 1.418 | 0.03484 | -0.03386 |  
## +---------------------+---------------------+--------------------+-----------------+---------------------+---------------------+

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



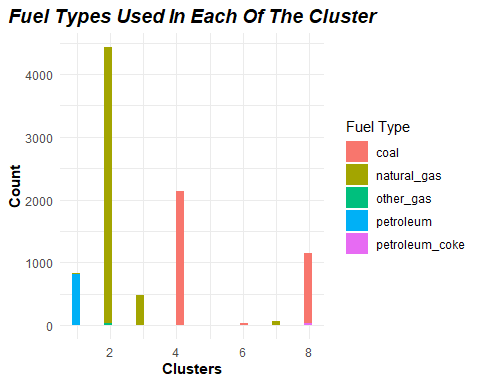
fuel\_median <- kmean\_clustering %>% group\_by(cluster) %>%  
summarise(median\_cost = median(fuel\_cost\_per\_mmbtu),  
median\_mmbtu = median(fuel\_mmbtu\_per\_unit),  
median\_received\_units = median(fuel\_received\_units),  
median\_sulfur = median(sulfur\_content\_pct)\*0.01,  
median\_ash = median(ash\_content\_pct)\*0.1,  
median\_mercury = median(mercury\_content\_ppm)\*0.001)  
fuel\_median

## # A tibble: 8 × 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>

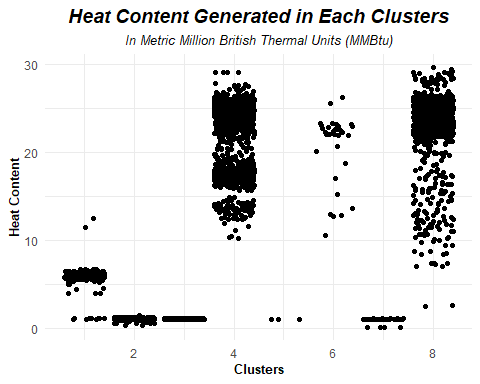
fuel\_clustering <- kmean\_clustering %>% select(fuel\_group\_code, cluster) %>%   
group\_by(fuel\_group\_code, cluster) %>% count() %>% arrange(cluster)  
fuel\_clustering

## # A tibble: 13 × 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

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 – The name states out the fuel used to generate power is Petroleum and natural gas, where the medial cost to produce power units of 5.8 MMBtu of heat content is 11.6$. The number of units thus received to produce power in this cluster is 828 units with no impurities of sulphur, mercury and ash content in it.*

*Cluster - 2 – This cluster has natural gas, and other gas. The median cost of generating power units of 1.0270 MMBtu of heat content in this cluster is 3.276$. The medial amount of units received and thereby used for power generation is 22154.5 units with no impurities of sulphur, mercury and ash content in it.*

*Cluster - 3 – This cluster has again natural gas. The fuel cost variable where the median price for 1.0300 natural gas units is pitched out to be 3.276$. The fuel units thus utilized to generate power is 2069810.0 units, the second highest out of all. There aren’t any impurities seen in this cluster.*

*Cluster - 4 – The fuel source used in this cluster is coal. The price induced to generate power units of 18.0220 MMBtu of heat content is 2.643$. The fuel units thus utilized to generate power is 25185.0 units. Lastly, the impurities of ash are greater than the permissible levels of 0.2 ppm i.e. 0.65 ppm and sulphur 0.0044 ppm.*

*Cluster - 5 – The fuel source used in this cluster is again natural gas. The median cost, which are extreme outliers, induced in generating power is 6010.289$ for 1.0370 of heat content. The fuel units thus utilized to generate power is 27.0 units with no impurities.*

*Cluster - 6 – The fuel source used in this cluster is coal. The price induced to generate power units of 22.2140 MMBtu of heat content is 3.276$. The fuel units thus utilized to generate power is 9532.0 units. Lastly, the impurities of ash are greater than the permissible levels of 0.2 ppm i.e. 2.02 ppm and sulphur 0.0245 ppm, greater than permissible 0.005 ppm. Out of all this is the only group which contains mercury impurities i.e. 0.00038*

*Cluster - 7 – Among all the clusters thereby formed this cluster has the highest amount of gas units being received for producing power i.e. 5364812.0 units with zero (0) impurities. The fuel source used in this cluster is natural gas and other gas. The price induced to generate power units of 1.0255 MMBtu of heat content is 3.276$.*

*Cluster - 8 – The fuel source used in this cluster is coal and petroleum coke. The price induced to generate power units of 23.3515 MMBtu of heat content is 2.916$. The fuel units thus utilized to generate power is 18815.0 units. Lastly, the impurities of ash are greater than the permissible levels of 0.2 ppm i.e. 0.91 ppm and sulphur also greater than permissible 0.005 ppm 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 significance when compared to other variables  
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.