Machine Learning Final Project

Dutt Thakkar

2023-05-07

#loading reqiured pacakges

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(ggcorrplot)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.0     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ lubridate 1.9.2     ✔ tibble    3.1.8
## ✔ purrr     1.0.1     ✔ tidyr     1.3.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ purrr::lift()   masks caret::lift()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conf
licts to become errors
```

```
library(tidyr)
library(dplyr)
library(e1071)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve
3WBa
```

```
library(cluster)
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
library(pander)
library(kernlab)
```

```
##
## Attaching package: 'kernlab'
##
## The following object is masked from 'package:purrr':
##
##     cross
##
## The following object is masked from 'package:ggplot2':
##
##     alpha
```

#importing dataset and viewing summary

```
Data_set <- read.csv("/Users/duttthakkar/Desktop/fuel_receipts_costs_eia923(1).csv")
summary(Data_set)
```

```
##      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 :304282   Median : 6155   Mode  :character   Mode  :character
##  Mean   :304282   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
```

#Gathering the percentages of all the null values from each column

```
fuel_data<-Data_set%>% replace(.=="", NA)
Null_values<-fuel_data%>%is.na()%>%colMeans()*100
Null_values
```

```
##                                    rowid
##                             0.000000e+00
##                             plant_id_eia
##                             0.000000e+00
##                       plant_id_eia_label
##                             1.834647e+00
##                              report_date
##                             0.000000e+00
##                       contract_type_code
##                             3.910846e-02
##                 contract_type_code_label
##                             3.910846e-02
##                 contract_expiration_date
##                             5.657597e+01
##                       energy_source_code
##                             0.000000e+00
##                 energy_source_code_label
##                             0.000000e+00
##                      fuel_type_code_pudl
##                             0.000000e+00
##                          fuel_group_code
##                             0.000000e+00
##                             mine_id_pudl
##                             6.440506e+01
##                       mine_id_pudl_label
##                             6.440506e+01
##                            supplier_name
##                             4.929638e-04
##                      fuel_received_units
##                             0.000000e+00
##                      fuel_mmbtu_per_unit
##                             0.000000e+00
##                       sulfur_content_pct
##                             0.000000e+00
##                          ash_content_pct
##                             0.000000e+00
##                      mercury_content_ppm
##                             4.756805e+01
##                      fuel_cost_per_mmbtu
##                             3.290369e+01
##         primary_transportation_mode_code
##                             9.562182e+00
##   primary_transportation_mode_code_label
##                             9.562182e+00
##       secondary_transportation_mode_code
##                             9.453336e+01
## secondary_transportation_mode_code_label
##                             9.453336e+01
##               natural_gas_transport_code
##                             4.398256e+01
##  natural_gas_delivery_contract_type_code
##                             7.298969e+01
##                     moisture_content_pct
##                             8.488639e+01
##                     chlorine_content_ppm
##                             8.488639e+01
##                            data_maturity
##                             0.000000e+00
##                      data_maturity_label
##                             0.000000e+00
```

#Removing all variables with null values having percentage more than 50 % and few other variables which doesn’t add much value to the analysis

```
fuel_data_1<- subset(fuel_data,select=c(rowid,plant_id_eia,fuel_received_units,fuel_mmbtu_per_unit,sulfur_content_pct,ash_content_pct,mercury_content_ppm,fuel_cost_per_mmbtu,contract_type_code,energy_source_code,fuel_type_code_pudl,fuel_group_code,supplier_name,primary_transportation_mode_code,plant_id_eia_label, natural_gas_transport_code,contract_type_code))
head(fuel_data_1)
```

```
##   rowid plant_id_eia fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct
## 1     1            3              259412              23.100               0.49
## 2     2            3               52241              22.800               0.48
## 3     3            3             2783619               1.039               0.00
## 4     4            7               25397              24.610               1.69
## 5     5            7                 764              24.446               0.84
## 6     6            7                 603              24.577               1.54
##   ash_content_pct mercury_content_ppm fuel_cost_per_mmbtu contract_type_code
## 1             5.4                  NA               2.135                  C
## 2             5.7                  NA               2.115                  C
## 3             0.0                  NA               8.631                  C
## 4            14.7                  NA               2.776                  C
## 5            15.5                  NA               3.381                  S
## 6            14.6                  NA               2.199                  S
##   energy_source_code fuel_type_code_pudl fuel_group_code    supplier_name
## 1                BIT                coal            coal  interocean coal
## 2                BIT                coal            coal  interocean coal
## 3                 NG                 gas     natural_gas bay gas pipeline
## 4                BIT                coal            coal     alabama coal
## 5                BIT                coal            coal     d & e mining
## 6                BIT                coal            coal     alabama coal
##   primary_transportation_mode_code plant_id_eia_label
## 1                               RV              Barry
## 2                               RV              Barry
## 3                               PL              Barry
## 4                               TR            Gadsden
## 5                               TR            Gadsden
## 6                               TR            Gadsden
##   natural_gas_transport_code contract_type_code.1
## 1                       firm                    C
## 2                       firm                    C
## 3                       firm                    C
## 4                       firm                    C
## 5                       firm                    S
## 6                       firm                    S
```

#Here we are sampling the 2% of fuel data:

```
set.seed(2299)

fuel_data_2<-sample_n(fuel_data_1,12000)
#Splitting the data into 75:25 test and train ratio

set.seed(2299)

sample<-createDataPartition(fuel_data_2$rowid,p=0.75, list=FALSE)

train<-fuel_data_2[sample,]

test<-fuel_data_2[-sample,]
```

#Combining the required Categorical and Numerical variables

```
data_<-train[,c(2,3,4,5,6,7,8,11)]
```

#Replacing the “NA” values with 0 for the calculations:

```
data_a<-data_%>% replace(.=="", NA)
head(data_a)
```

```
##   plant_id_eia fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct
## 1         7032                4149               1.078                  0
## 2        55230                3245               1.011                  0
## 3        54268              423130               0.959                  0
## 6         3443              113767               1.029                  0
## 8         1719                4172               1.005                  0
## 9         1556              193910               1.070                  0
##   ash_content_pct mercury_content_ppm fuel_cost_per_mmbtu fuel_type_code_pudl
## 1               0                   0               2.967                 gas
## 2               0                   0               2.268                 gas
## 3               0                   0                  NA                 gas
## 6               0                   0                  NA                 gas
## 8               0                  NA               7.411                 gas
## 9               0                  NA                  NA                 gas
```

```
data_a[is.na(data_a)] <- 0
```

#Assigning the dummy variables to the categorical variables fuel_type_code_pudl

```
coal <- ifelse(data_a$fuel_type_code_pudl=="coal" ,1,0)
gas <- ifelse(data_a$fuel_type_code_pudl=="gas" ,1,0)
oil <- ifelse(data_a$fuel_type_code_pudl=="oil" ,1,0)


fuel_data3<-cbind(data_a[,-c(8)], coal, gas, oil)
head(fuel_data3)
```

```
##   plant_id_eia fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct
## 1         7032                4149               1.078                  0
## 2        55230                3245               1.011                  0
## 3        54268              423130               0.959                  0
## 6         3443              113767               1.029                  0
## 8         1719                4172               1.005                  0
## 9         1556              193910               1.070                  0
##   ash_content_pct mercury_content_ppm fuel_cost_per_mmbtu coal gas oil
## 1               0                   0               2.967    0   1   0
## 2               0                   0               2.268    0   1   0
## 3               0                   0               0.000    0   1   0
## 6               0                   0               0.000    0   1   0
## 8               0                   0               7.411    0   1   0
## 9               0                   0               0.000    0   1   0
```

#Normalizing the Data

```
fuel_data4<-scale(fuel_data3)
```

Applying hierarchical clustering algorithm

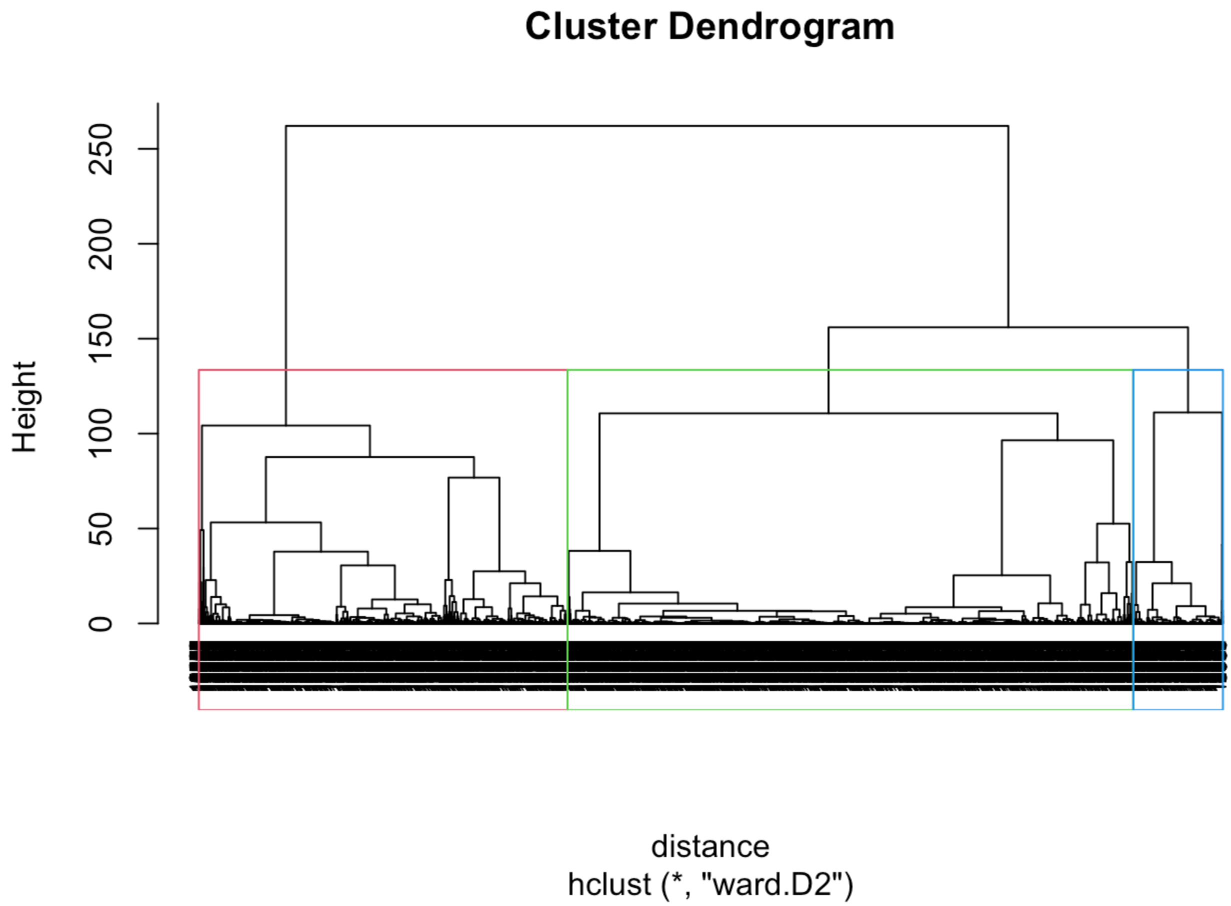
Creating the dissimilarity matrix for data set the through Euclidean distance

```
distance <- dist(fuel_data4, method = "euclidean")

# Hierarchical clustering using the Ward's method
cluster_fuel <- hclust(distance, method = "ward.D2" )
cluster_fuel
```

```
##
## Call:
## hclust(d = distance, method = "ward.D2")
##
## Cluster method   : ward.D2
## Distance         : euclidean
## Number of objects: 9000
```

#Because of Ward’s minimal variance, Ward’s distance is employed. the standard reduces the overall within-cluster variance

```
# Plotting the cluster Dendrogram

plot(cluster_fuel, cex = 0.6, hang = -1)
rect.hclust(cluster_fuel,k=3,border=2:5)
```



#Cut-off height = 140.Therefore number of clusters = 3. We select k value = 3 using the domain knowledge to determine the distribution of 3 fuel kinds in each cluster.

#cutting the dendrogram tree for k=3

```
group <- cutree(cluster_fuel, k = 3)
```

#Finding the number of members in each of the clusters.

```
table(group)
```

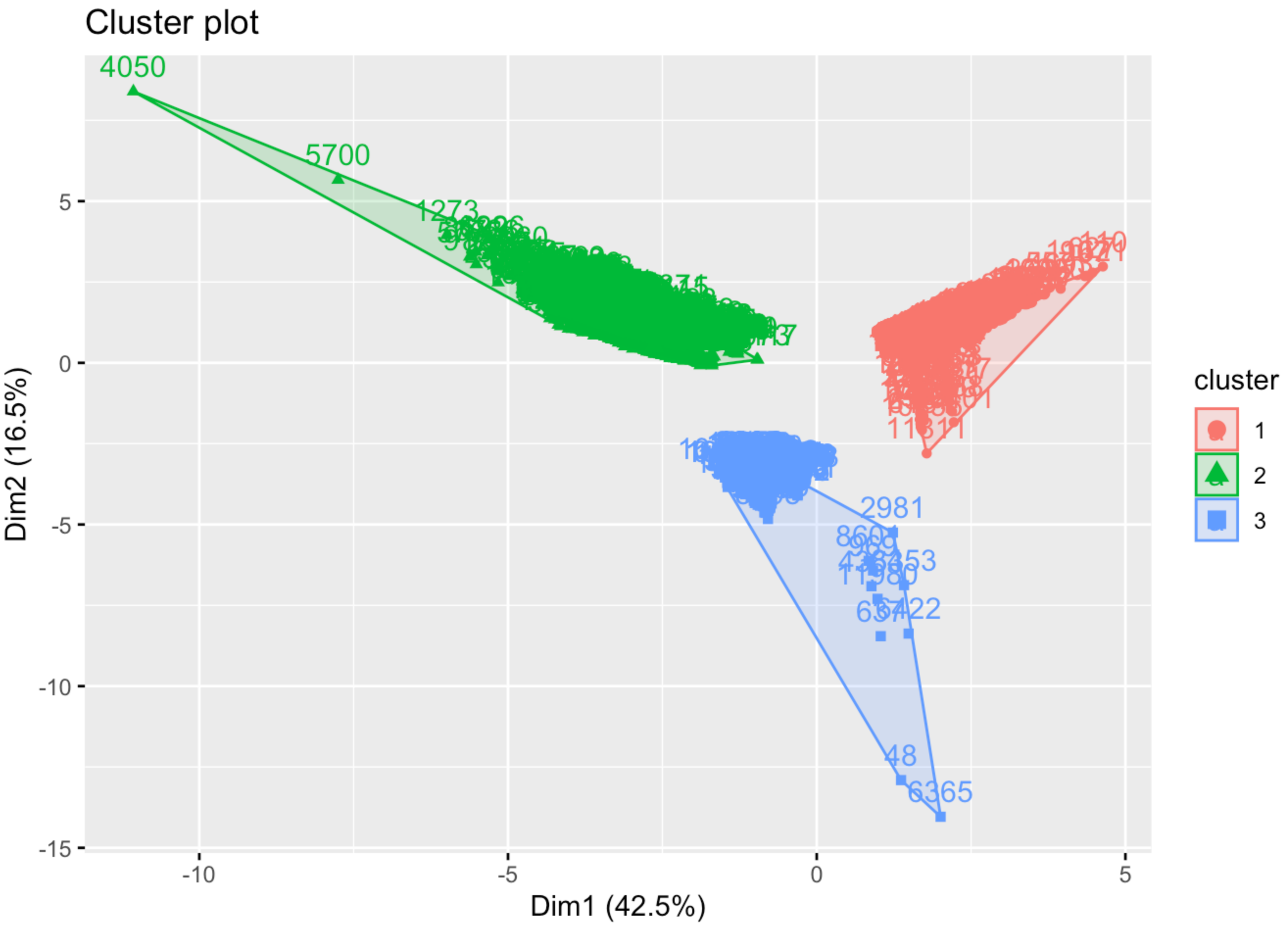
```
## group
##    1    2    3 
## 4972 3240  788
```

#Binding the clusters to main data

```
fuel_data5<- cbind(fuel_data3, clustering = group)
```

#Cluster visualization

```
fviz_cluster(list(data = fuel_data5, cluster = group))
```



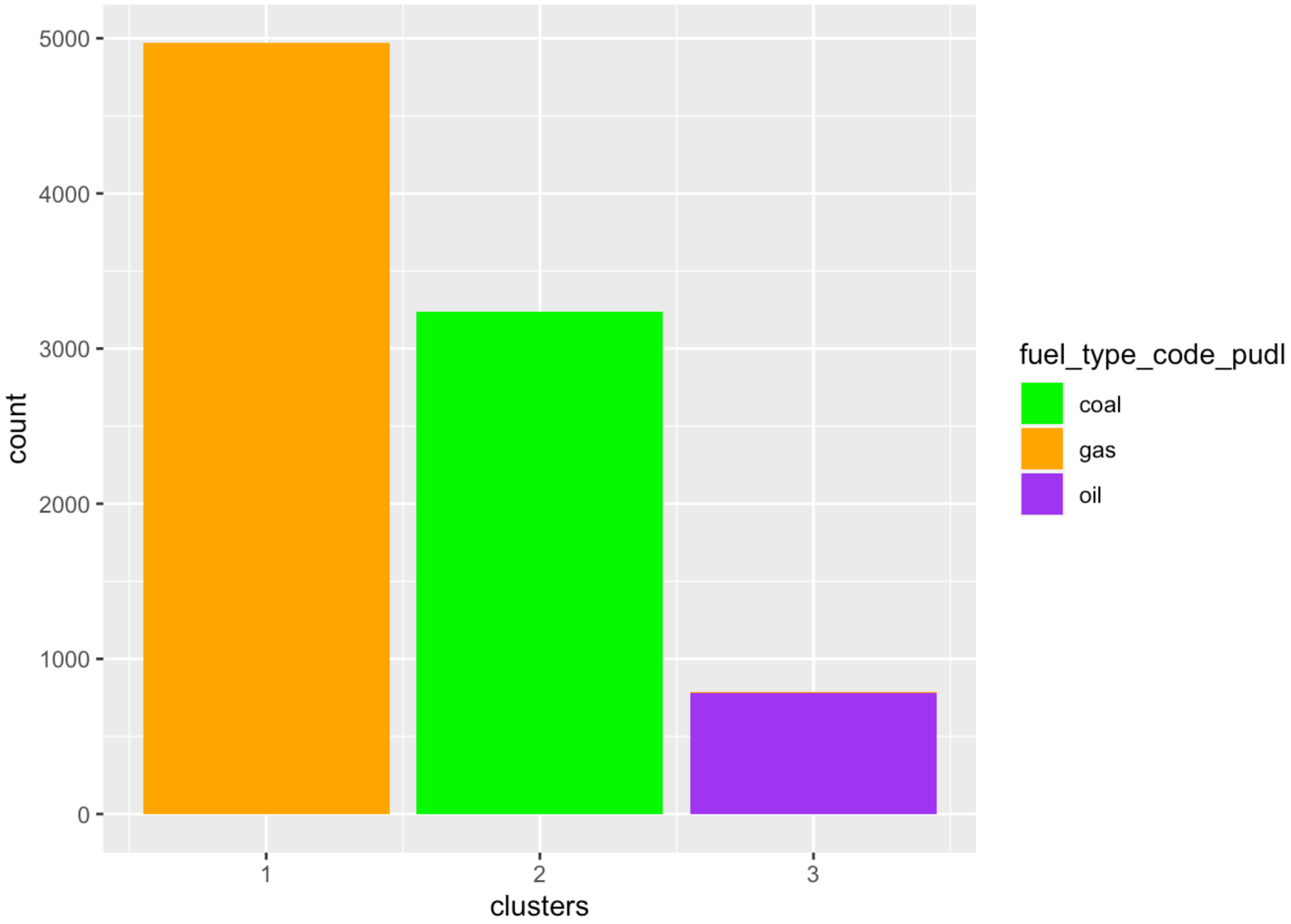
#The three clusters are named as Coal, Gas and Oil #Cluster 1= GAS #Cluster 2= COAL #Cluster 3= OIL

#Finding the mean of the required columns for interpretation

```
combined_data<-cbind(fuel_data5,train[,c(9,11,12,14,15,16)])

fuel<-combined_data %>% mutate(clusters=combined_data$clustering) %>% group_by(cluste
rs)

fuel_data5<-fuel[,c(2:11)]%>%group_by(clustering)%>%summarise_all("mean")
```

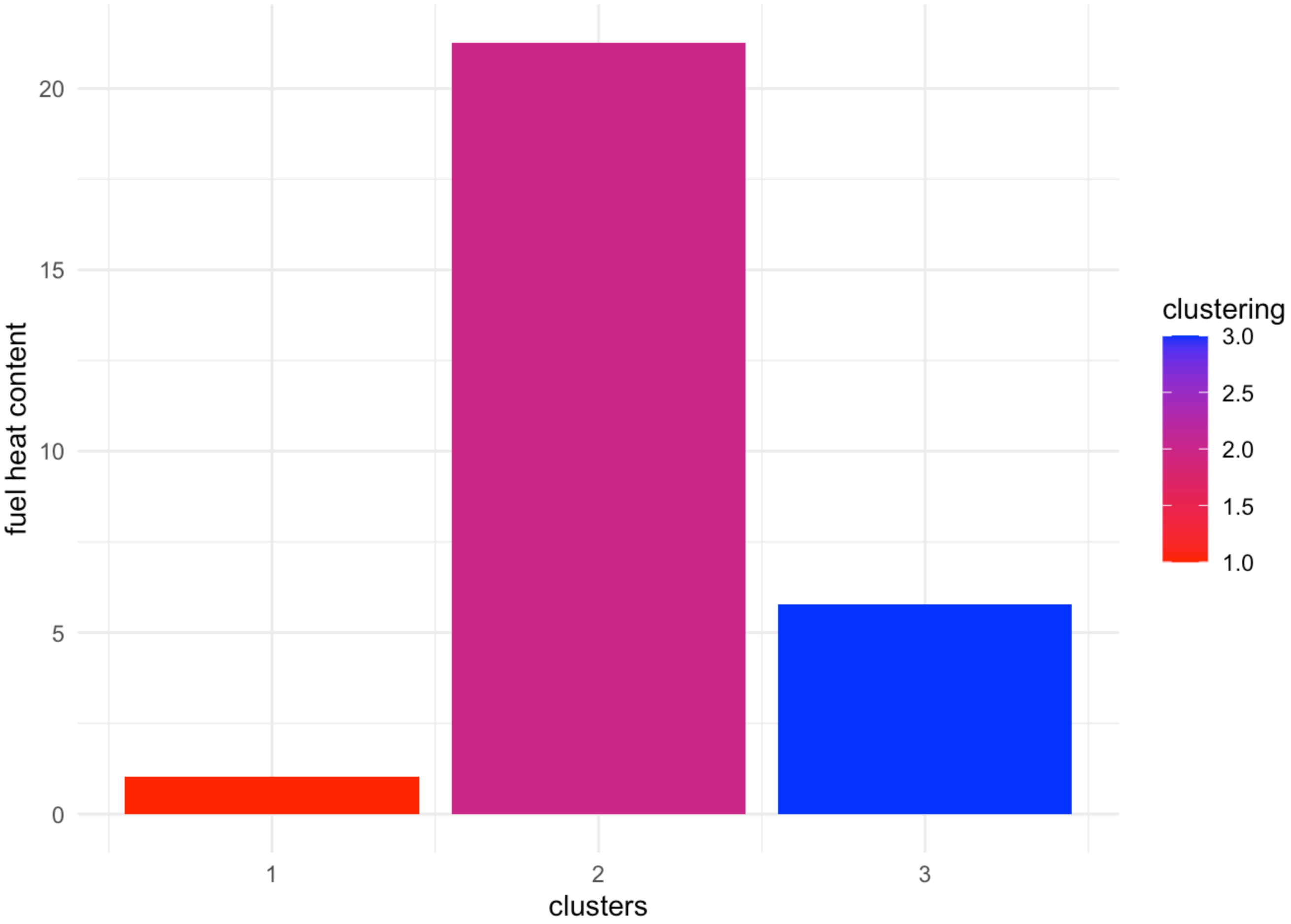
#Plotting clusters vs other variables #Cluster vs fuel type

```
ggplot(fuel, aes(x = clusters, fill = fuel_type_code_pudl)) +
  geom_bar() +
  scale_fill_manual(values = c("green", "orange", "purple"))
```



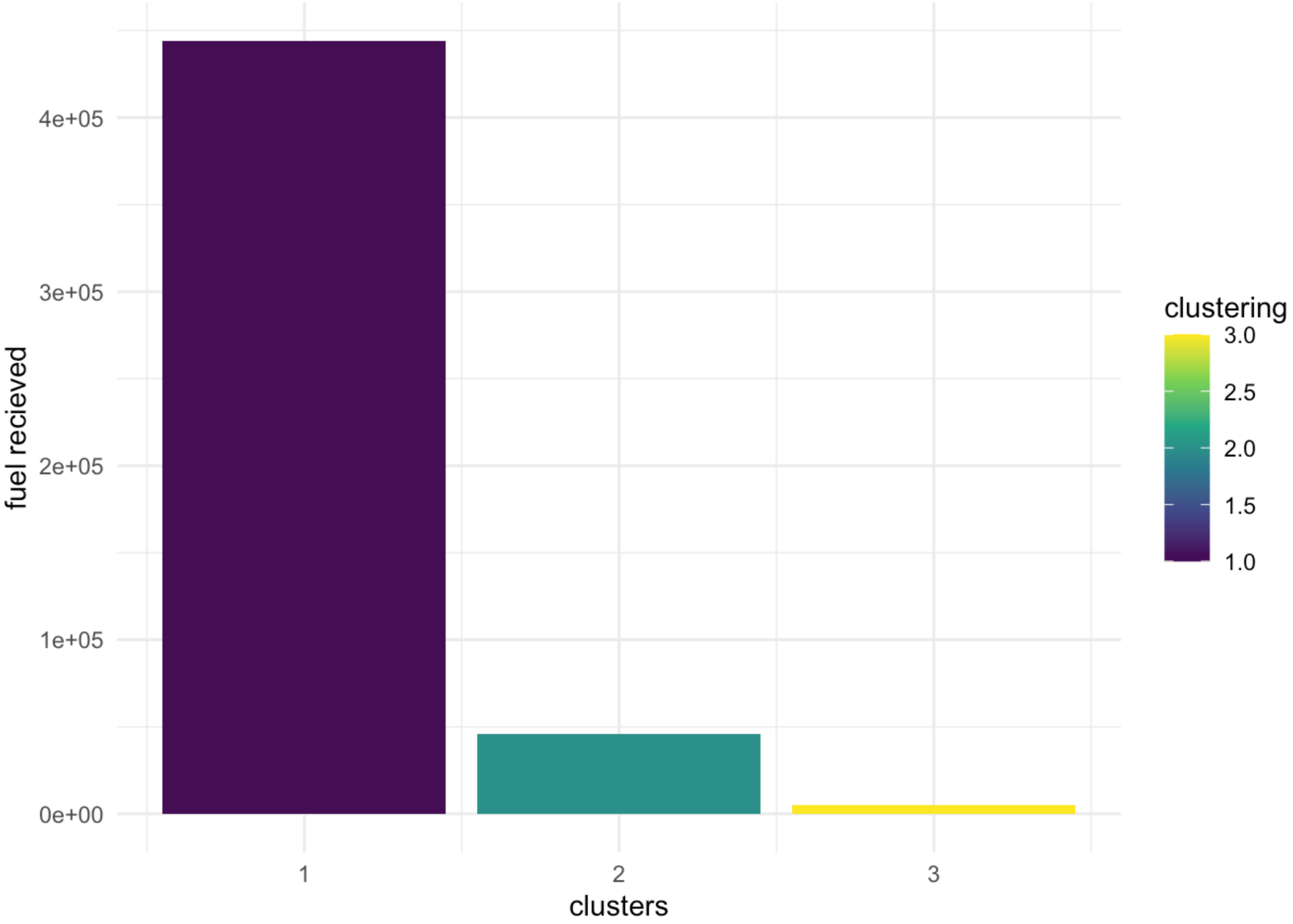
#cluster vs heat content #Clustering for heat content in the fuel

```
ggplot(fuel_data5, aes(x=clustering, y=fuel_mmbtu_per_unit,fill=clustering)) + geom_bar(stat="identity") +
labs(x="clusters", y="fuel heat content")+scale_fill_gradient(low = "red", high = "blue") + theme_minimal()
```



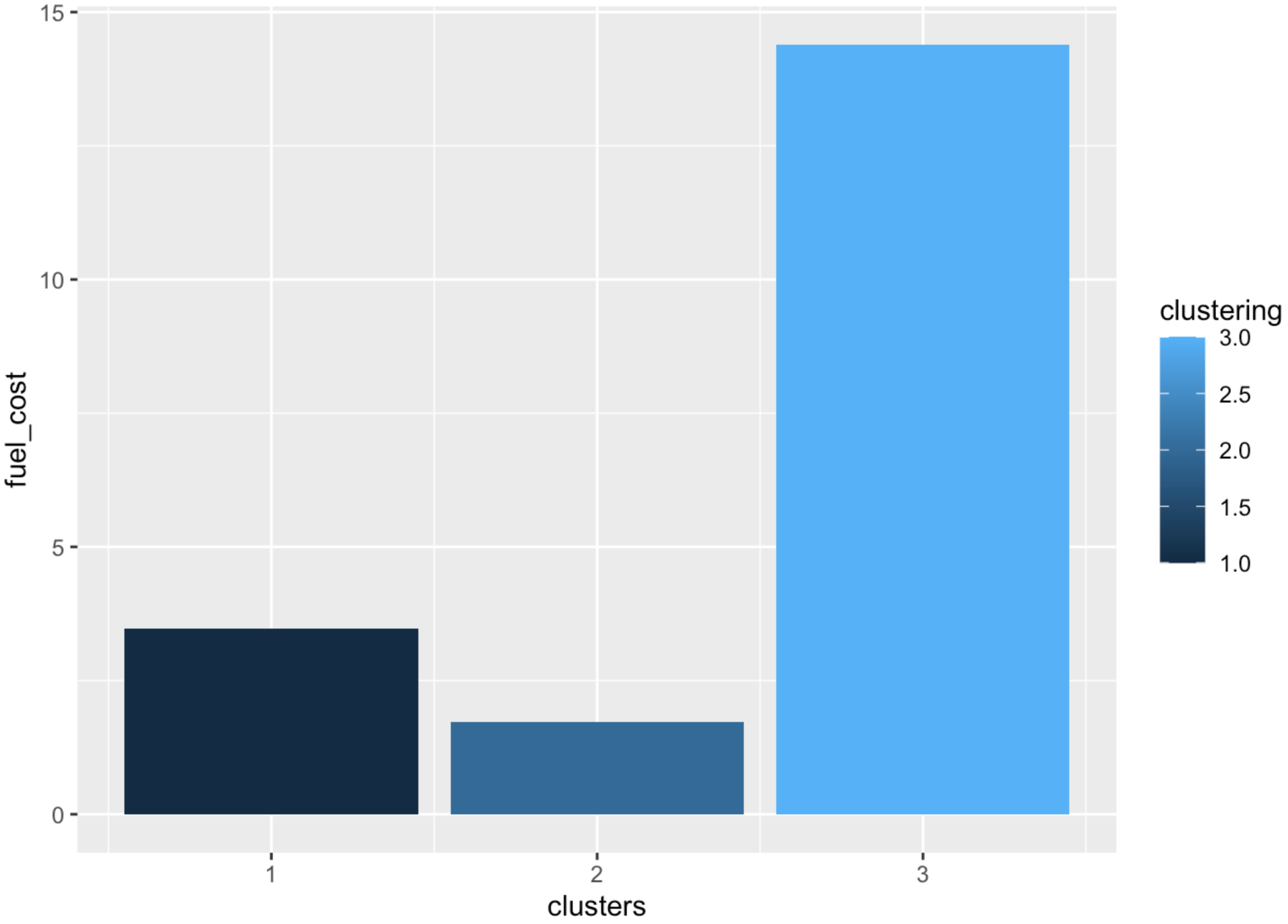
#Cluster vs fuel received #Clustering for Fuel received

```
ggplot(fuel_data5, aes(x=clustering, y=fuel_received_units,fill=clustering)) + geom_b
ar(stat="identity") +
labs(x="clusters", y="fuel recieved")+scale_fill_viridis_c(option = "viridis", direct
ion = 1) +
theme_minimal()
```



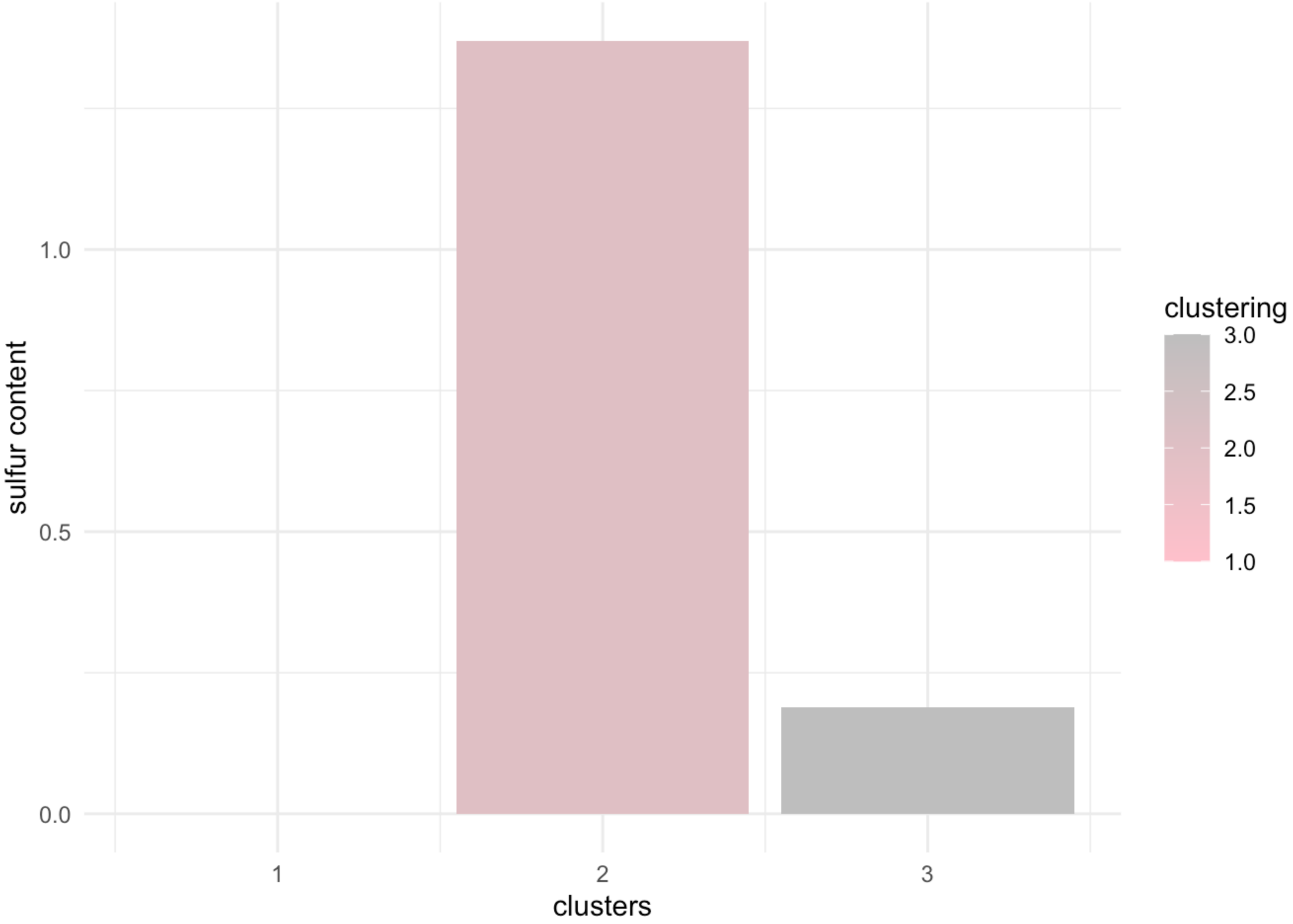
#Cluster vs fuel cost #Clustering for fuel cost

```
ggplot(fuel_data5, aes(x=clustering, y = fuel_cost_per_mmbtu,fill=clustering)) + geom
_bar(stat="identity") +
labs(x="clusters", y="fuel_cost")
```



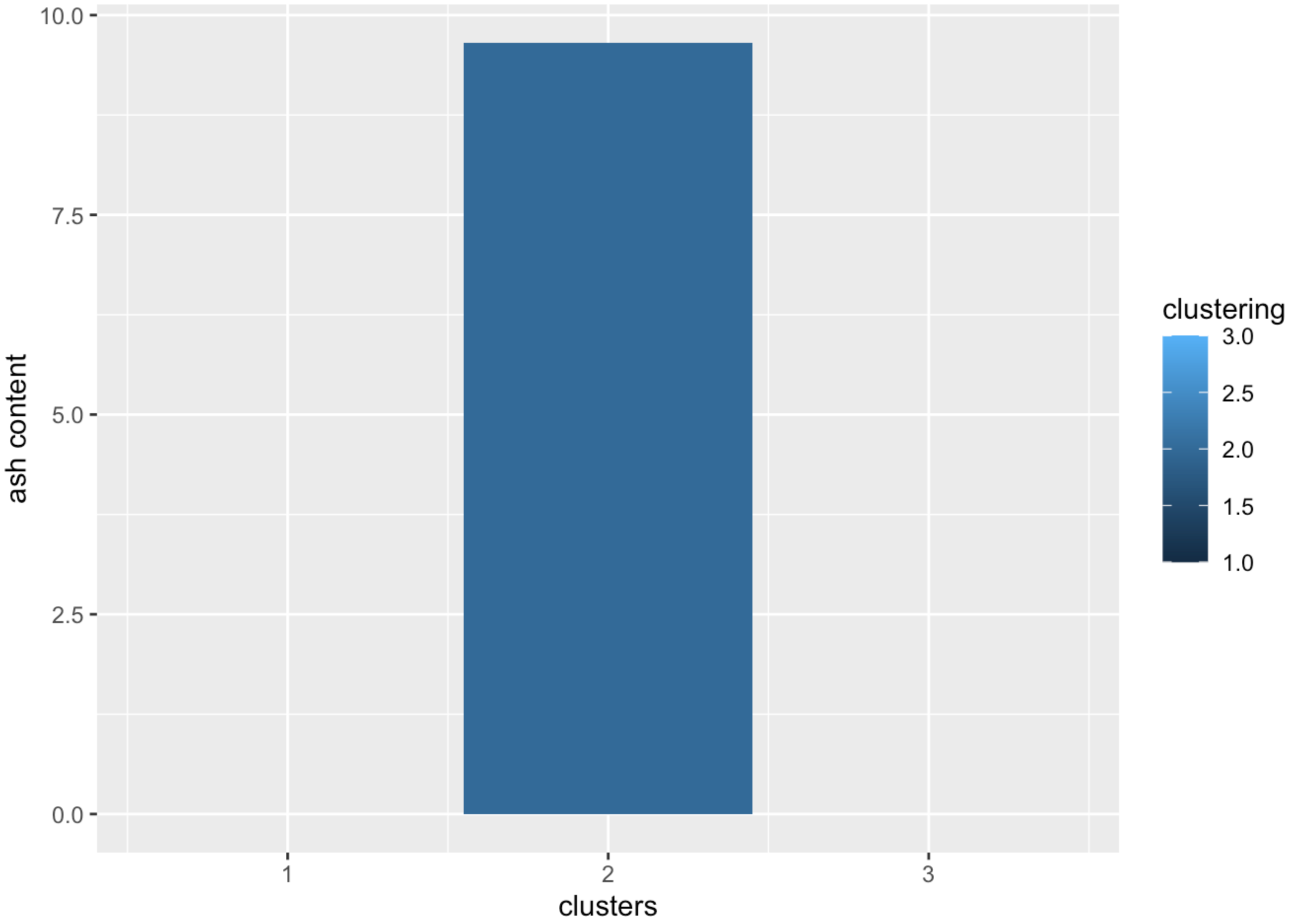
#Cluster vs sulfur content #Clustering for Sulfur Content.

```
ggplot(fuel_data5, aes(x=clustering, y=sulfur_content_pct,fill=clustering)) + geom_bar(stat="identity") +
labs(x="clusters", y="sulfur content")+scale_fill_gradient(low = "pink", high = "grey") + theme_minimal()
```



#Cluster vs ash content #Clustering for Ash Content

```
ggplot(fuel_data5, aes(x=clustering, y=ash_content_pct,fill=clustering)) + geom_bar(s
tat="identity") +
labs(x="clusters", y="ash content")
```



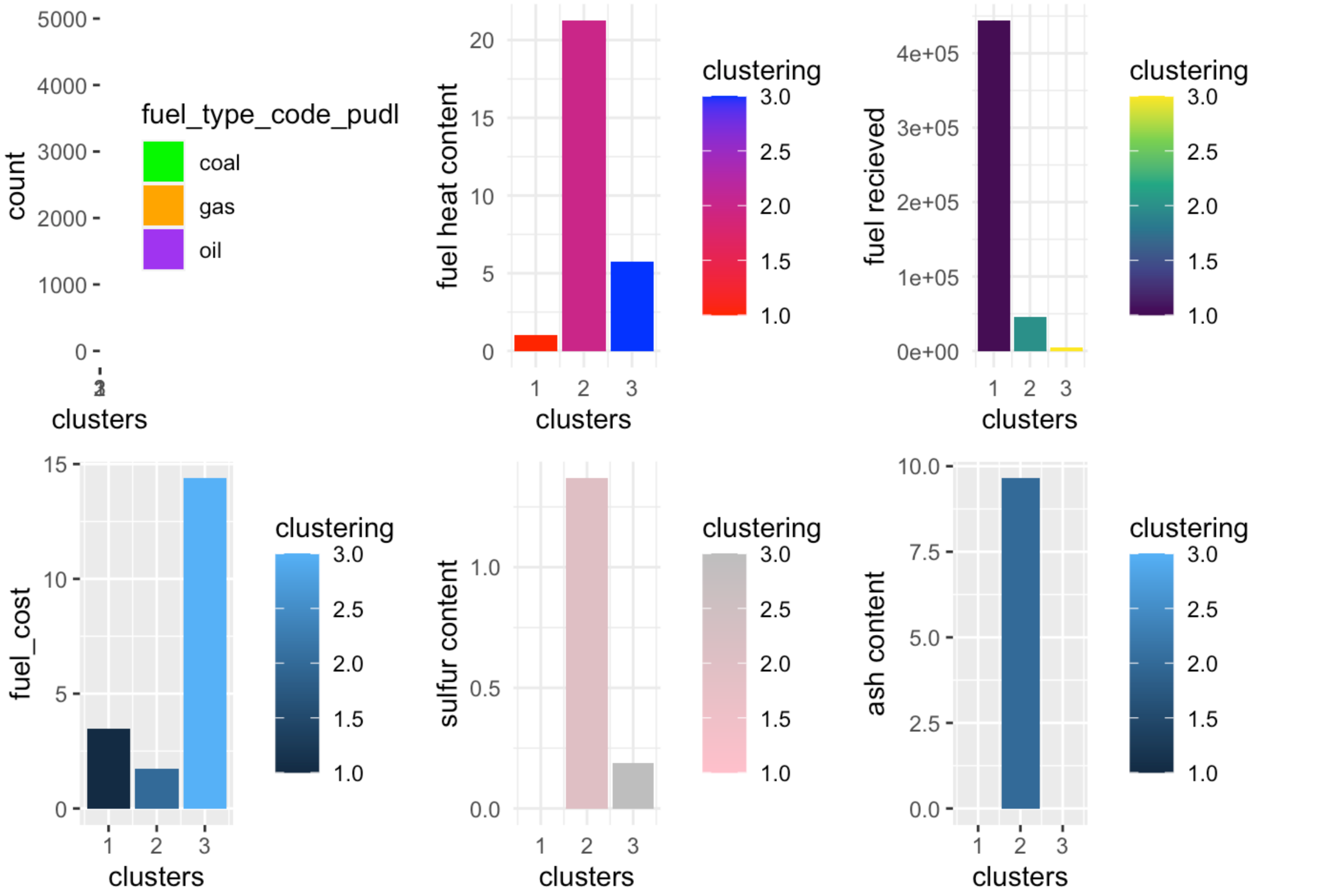
#Combining plots

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
# Define each plot separately
p1 <- ggplot(fuel, aes(x = clusters, fill = fuel_type_code_pudl)) +
  geom_bar() +
  scale_fill_manual(values = c("green", "orange", "purple"))

p2 <- ggplot(fuel_data5, aes(x=clustering, y=fuel_mmbtu_per_unit,fill=clustering)) +
geom_bar(stat="identity") +
  labs(x="clusters", y="fuel heat content")+scale_fill_gradient(low = "red", high = "blue") + theme_minimal()

p3 <- ggplot(fuel_data5, aes(x=clustering, y=fuel_received_units,fill=clustering)) +
geom_bar(stat="identity") +
  labs(x="clusters", y="fuel recieved")+scale_fill_viridis_c(option = "viridis", direction = 1) +
  theme_minimal()

p4 <- ggplot(fuel_data5, aes(x=clustering, y=fuel_cost_per_mmbtu,fill=clustering)) +
geom_bar(stat="identity") +
  labs(x="clusters", y="fuel_cost")

p5 <- ggplot(fuel_data5, aes(x=clustering, y=sulfur_content_pct,fill=clustering)) + geom_bar(stat="identity") +
  labs(x="clusters", y="sulfur content")+scale_fill_gradient(low = "pink", high = "grey") + theme_minimal()

p6 <- ggplot(fuel_data5, aes(x=clustering, y=ash_content_pct,fill=clustering)) + geom_bar(stat="identity") +
  labs(x="clusters", y="ash content")

# Combine the plots using grid.arrange()
grid.arrange(p1, p2, p3, p4, p5, p6, ncol = 3)
```



Conclusion of clusters

#Cluster 1 - Gas - Our sample contains Cluster 1 data, which make up about 55% of the data. Gas is the main fuel type in this cluster. With 411384 units, gas has the highest average number of units received compared to coal and oil. Materials like ash, and sulphur are absent from gas. Each MMBtu of gasoline costs 4.50 USD.

#Cluster 2 - Coal - 3215 observations make up Cluster 2, representing 35.72% of the data in our sample. The fuel type used in this cluster is coal. 47862 units are often obtained in terms of coal units. The fuel has an average heat content of 21.32, which is higher than the heat contents of the other two fuels. The typical sulfur and ash concentrations in coal fuel are 1.38 percent and 10 percent, respectively. With an average fuel price of 1.70 USD per MMBtu, coal energy is less expensive than gas and oil.

#Cluster 3 - Oil - Cluster 3 only accounts for 8.9% of the data in our sample. Oil is the fuel that is used. We received 6628 units of gasoline in total. Fuel has a heat content that is 5.83 units higher than gas. The sulphur content of this type of gasoline is extremely low at 0.19%. The price of fuel is $17 per MMBtu.