Machine learning Project

Varshini

2022-12-08

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(ISLR)
library(gridExtra)
```

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

```
library(cluster)
library(dplyr)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(dbscan)
library(factoextra)
```

```
Fuel_Dataset <- read.csv("~/Downloads/fuel.csv")

summary(Fuel_Dataset)
```

```
##      rowid         plant_id_eia   plant_id_eia_label report_date
```

```
##  Min.   :     1   Min.   :    3   Length:608565      Length:608565
##  1st Qu.:152142   1st Qu.: 2712   Class :character   Class :character
##  Median :304283   Median : 6155   Mode  :character   Mode  :character
##  Mean   :304283   Mean   :18290
##  3rd Qu.:456424   3rd Qu.:50707
##  Max.   :608565   Max.   :64020
##
##  contract_type_code contract_type_code_label contract_expiration_date
##  Length:608565      Length:608565            Length:608565
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##
##  energy_source_code energy_source_code_label fuel_type_code_pudl
##  Length:608565      Length:608565            Length:608565
##  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:608565      Min.   :   0     Min.   :   0       Length:608565
##  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   :391947   NA's   :391947
##  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:608565
##  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:608565                          Length:608565
##  Class :character                       Class :character
##  Mode  :character                       Mode  :character
##
##
##
```

```
##
##  secondary_transportation_mode_code_label natural_gas_transport_code
##  Length:608565                            Length:608565
##  Class :character                         Class :character
##  Mode  :character                         Mode  :character
##
##
##
##
##  natural_gas_delivery_contract_type_code moisture_content_pct
##  Length:608565                           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   :516589
##  chlorine_content_ppm data_maturity      data_maturity_label
##  Min.   :   0.0       Length:608565      Length:608565
##  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   :516589
```

```
#choosing the numerical variables and removing the Null Values from the dataset.
data1<-Fuel_Dataset[,c(11,15,16,20)]

#Checking NA
colMeans(is.na(data1))
```

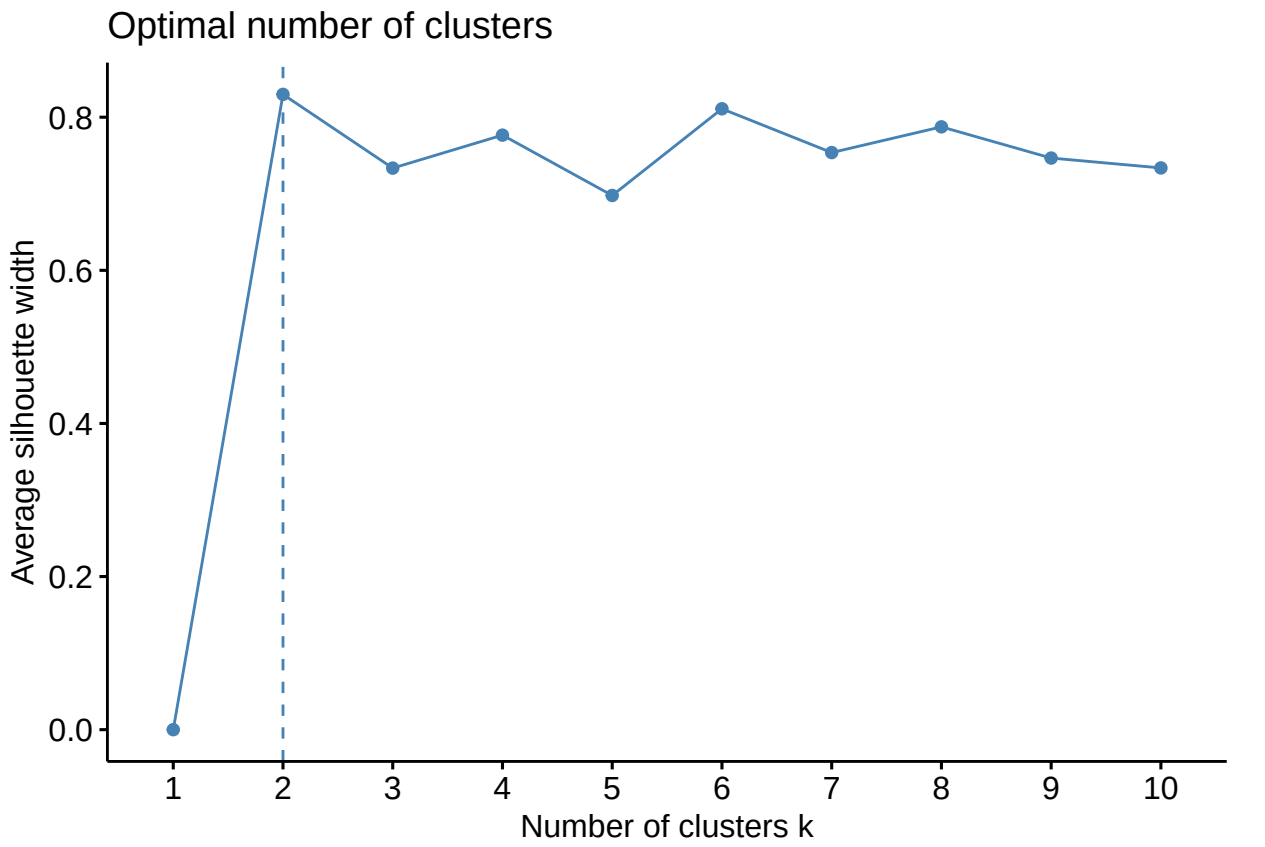
```
##     fuel_group_code fuel_received_units fuel_mmbtu_per_unit fuel_cost_per_mmbtu
##           0.0000000           0.0000000           0.0000000           0.3290363
```

```
#Imputting NA
data1$fuel_cost_per_mmbtu [is.na(data1$fuel_cost_per_mmbtu )]<-
  median(data1$fuel_cost_per_mmbtu , na.rm = T)
```

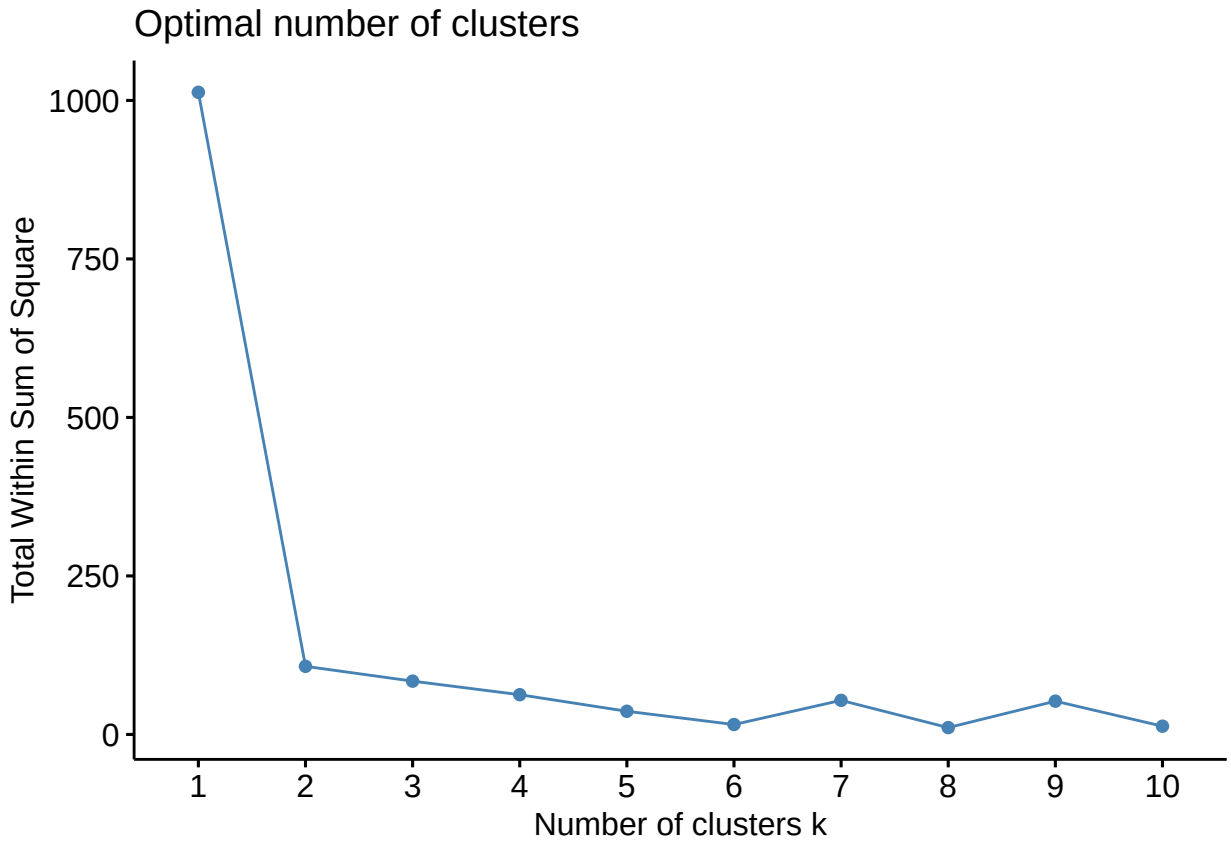
```
set.seed(1234)
data1_partition <- createDataPartition(data1$fuel_cost_per_mmbtu ,p=.015, list = FALSE)
Train <- data1[data1_partition,]
Exc_Data <- data1[-data1_partition,]

data2_partition <- createDataPartition(Exc_Data$fuel_cost_per_mmbtu,p=0.005,list=F)
Test <- Exc_Data[data2_partition,]
Exc.Data.1 <- Exc_Data[-data2_partition,]
```

```
#Data Normalization
norm_data <- preProcess(Train[,-1],
                        method=c("range"))
train_norm <-predict(norm_data,Train)
test_norm <-predict(norm_data,Test)
```

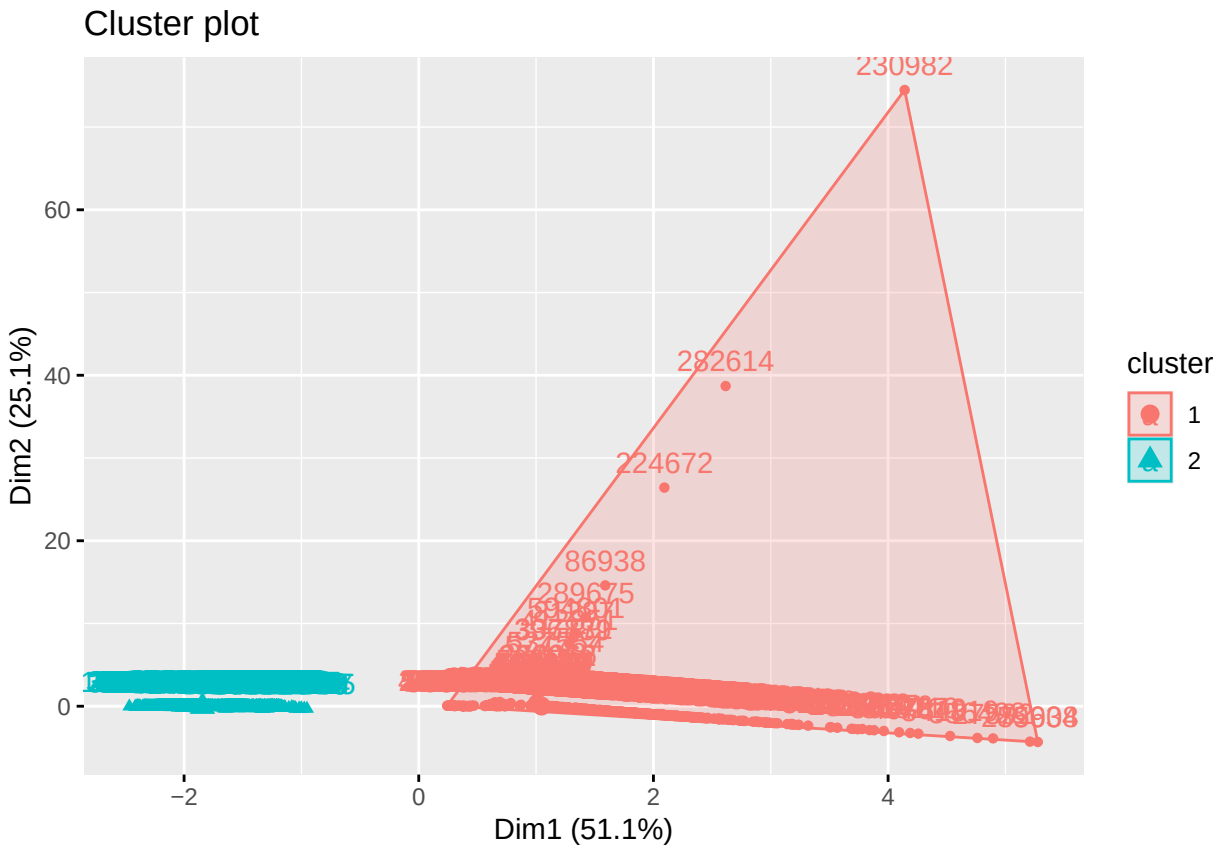
```
fviz_nbclust(train_norm[,-1],kmeans,method="silhouette")
```




```
fviz_nbclust(train_norm[,-1],kmeans,method="wss")
```




```
# Using several values of K, computing K-means clustering for various centers,
#and comparing the results

kmeans.df <- kmeans(train_norm[,-1], centers = 2, nstart = 25)
cluster <- kmeans.df$cluster

kmeans.df.1 <- cbind(Train,cluster)

plot.cluster <- fviz_cluster(kmeans.df,kmeans.df.1[,-1])
plot.cluster
```




```
kmeans.df.1%>%group_by(cluster)%>%
  summarize(median_units=median(fuel_received_units),
            median_cost=median(fuel_cost_per_mmbtu),
            median_mmbtu=median(fuel_mmbtu_per_unit))
```

```
## # A tibble: 2 x 4
##   cluster median_units median_cost median_mmbtu
##     <int>        <int>       <dbl>        <dbl>
## 1       1        20000        3.28         1.03
## 2       2        21009        2.72        22.6
```

```
kmeans.df.1 %>% select(fuel_group_code,cluster)%>%
  group_by(cluster,fuel_group_code) %>% count()
```

```
## # A tibble: 6 x 3
## # Groups:   cluster, fuel_group_code [6]
##   cluster fuel_group_code     n
##     <int> <chr>           <int>
## 1       1 coal               33
## 2       1 natural_gas      4949
## 3       1 other_gas          31
## 4       1 petroleum         810
## 5       2 coal             3271
## 6       2 petroleum_coke     36
```