FML_Project

Karthik Badiganti

2022-12-04

Loading Packages

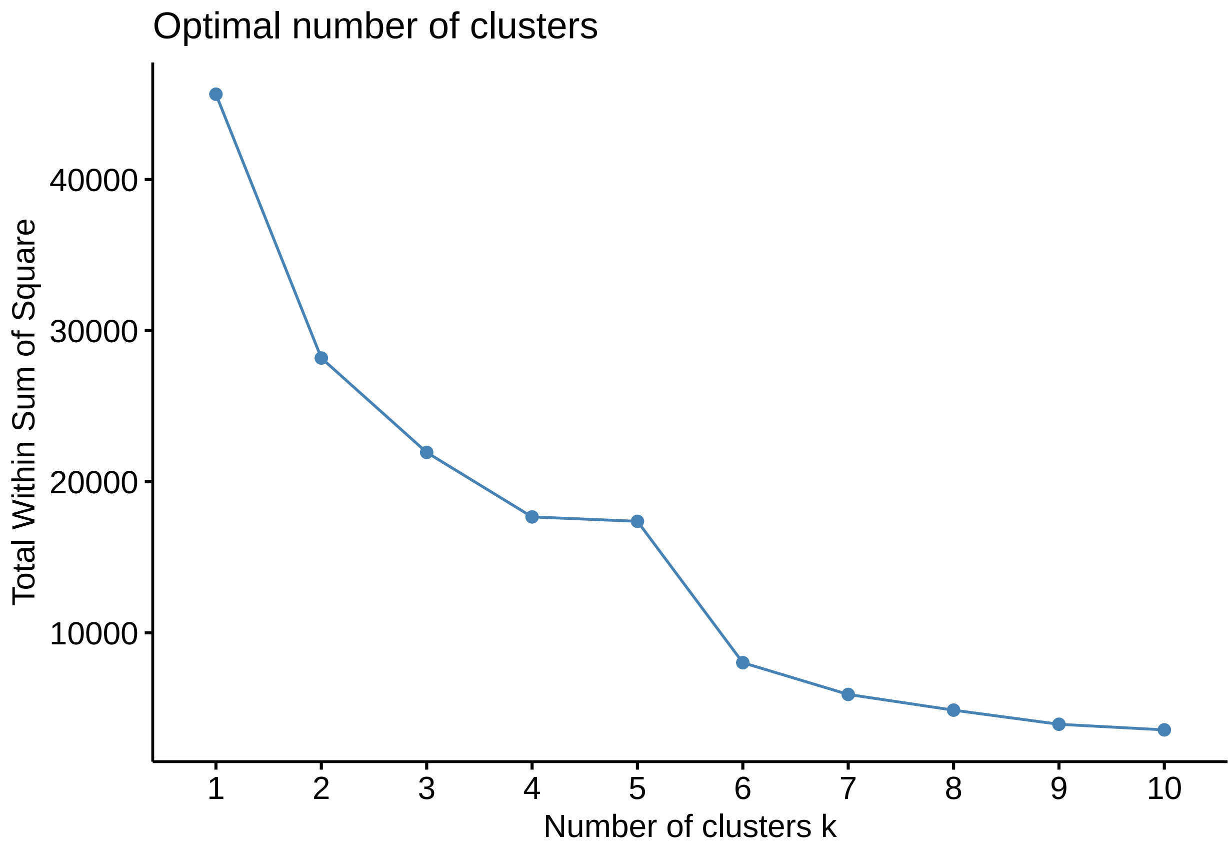
Importing & Cleaning Data

```
fuel <- read.csv("R Scripts/fuel_receipts_costs_eia923.csv",na.strings = "")



fuel<-fuel[,-c(3,7,12,13,19,21,22,23,24,25,26,27,28)]
```

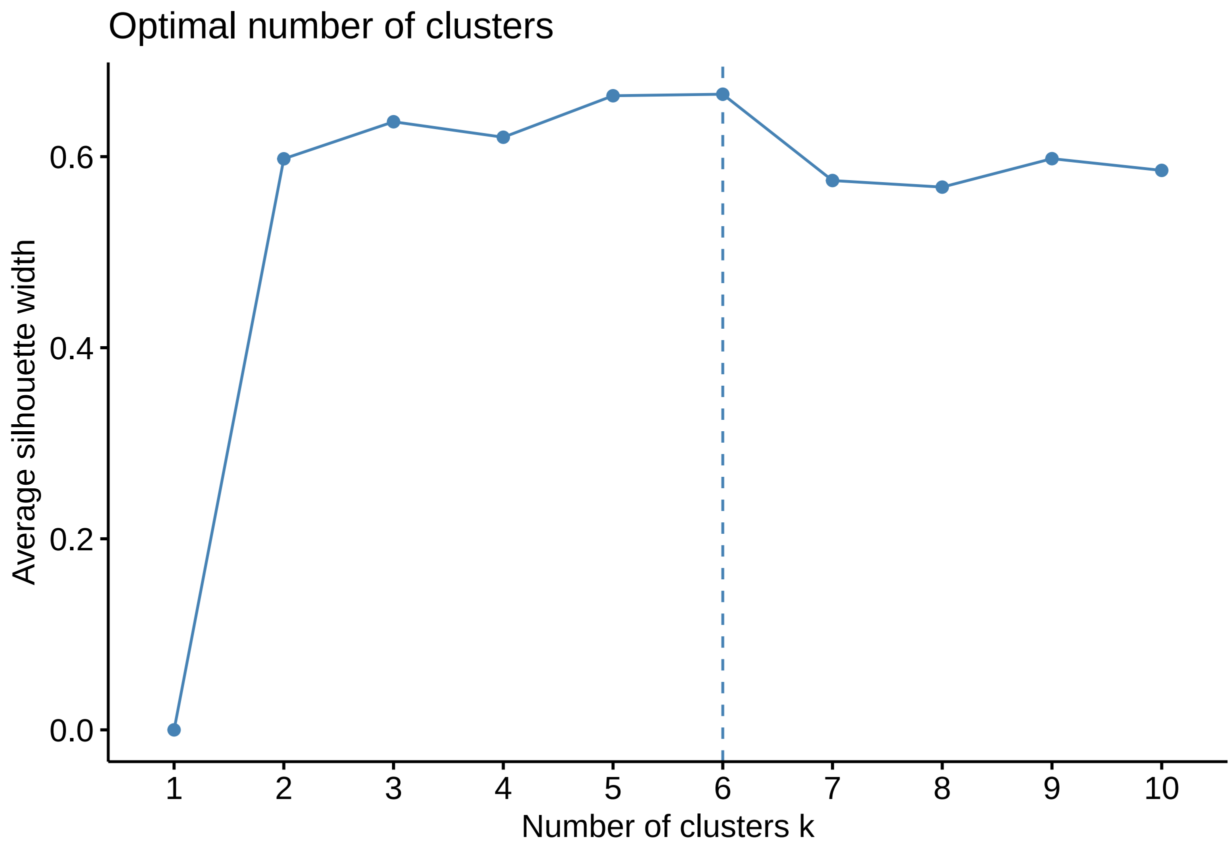
Sampling 2% percent data

```
set.seed(2121)
norm_model<-preProcess(fuel_3_Train, method = c('center','scale'))
fuel_3_Train_norm<-predict(norm_model,fuel_3_Train)

fuel_3_Validation_norm<-predict(norm_model,fuel_3_Validation)
```

```
set.seed(1212)
fviz_nbclust(fuel_3_Train_norm[-c(1,2)], kmeans, method = "wss")
```



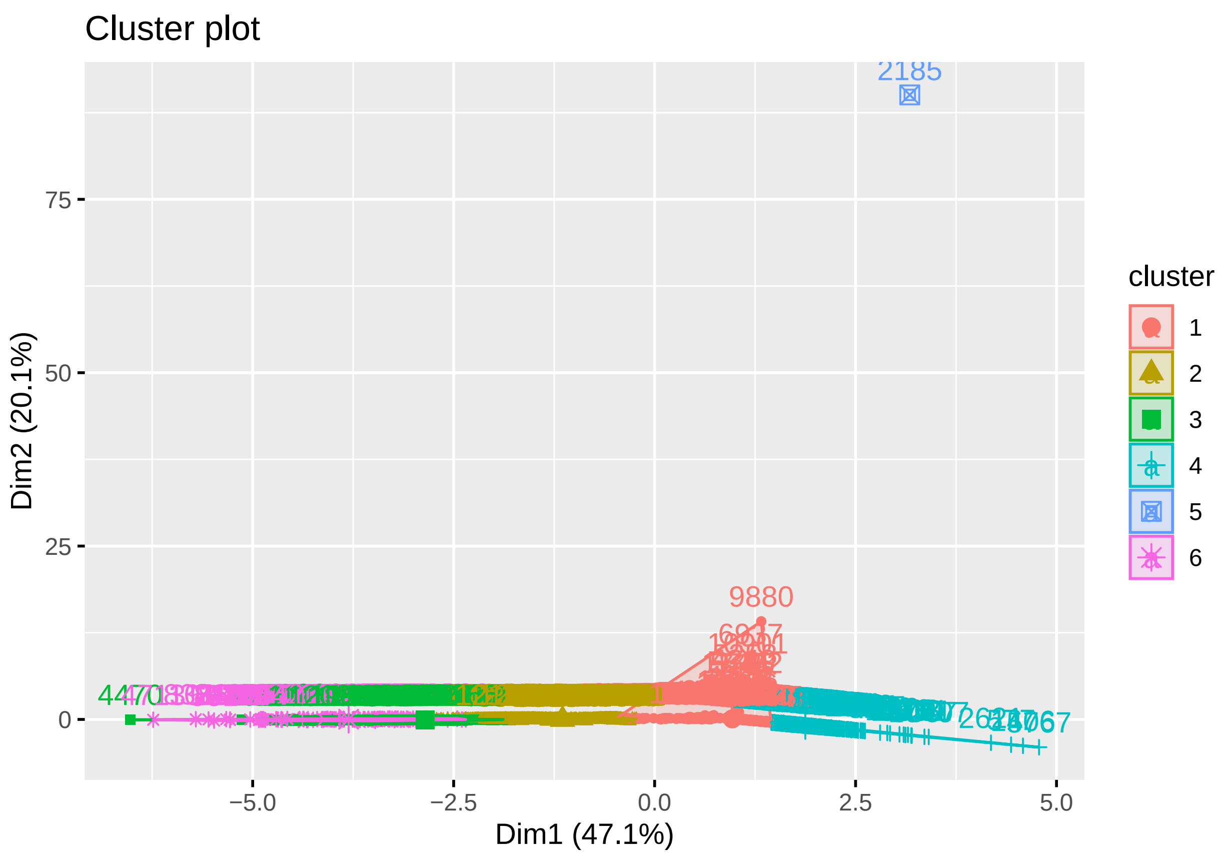

```
fviz_nbclust(fuel_3_Train_norm[-c(1,2)], kmeans, method = "silhouette")
```



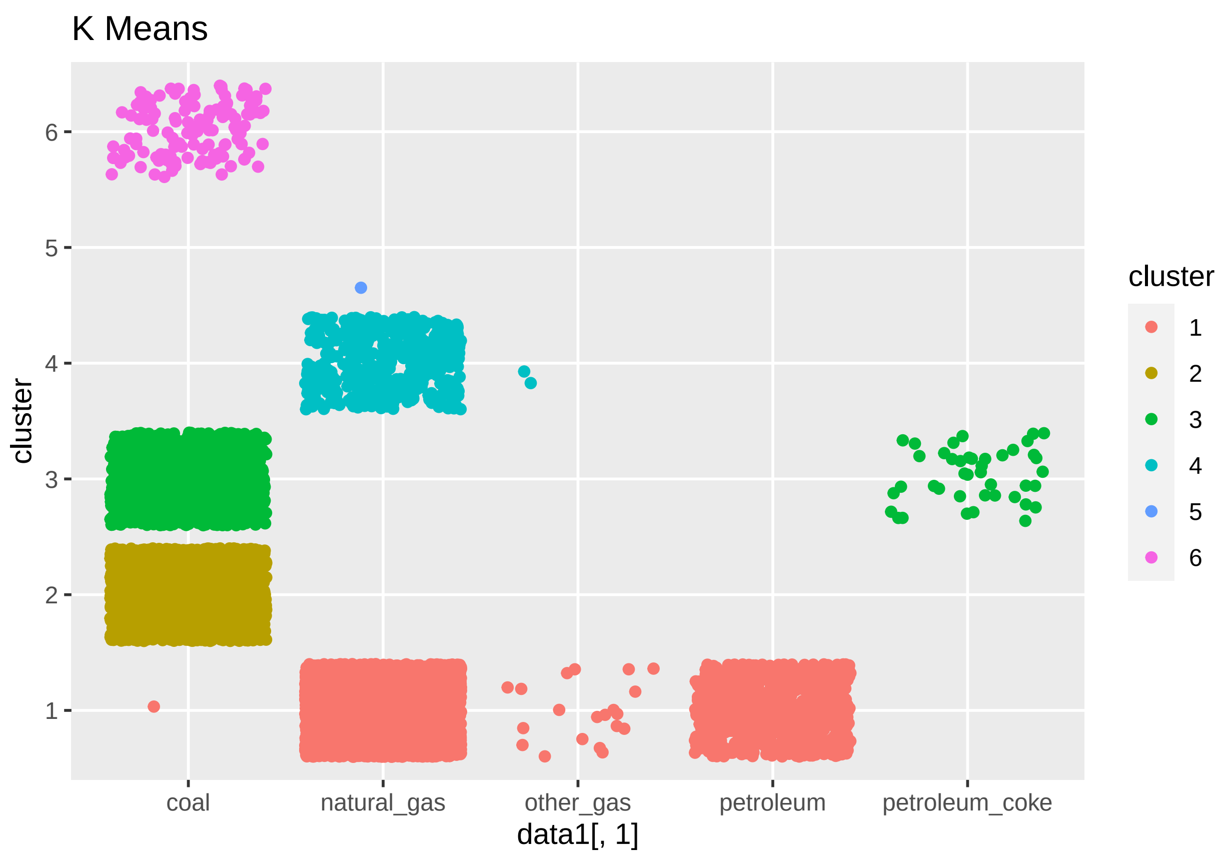

```
k4 <- kmeans(fuel_3_Train_norm[-c(1,2)], centers = 6, nstart = 25)
k4$centers
```

```
##   fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct
## 1          -0.1473333          -0.7233862         -0.4904550      -0.5503783
## 2          -0.2617929           1.1814553          0.0647893       0.6474282
## 3          -0.2909246           1.5239258          2.4133925       1.0048613
## 4           3.6910678          -0.7994052         -0.5186813      -0.5503783
## 5          -0.3435425          -0.7961202         -0.5186813      -0.5503783
## 6          -0.3126849           0.4338855          0.7641239       5.6987376
##   fuel_cost_per_mmbtu
## 1          0.01064469
## 2         -0.04436896
## 3         -0.04404701
## 4         -0.03037173
## 5         92.68072689
## 6          0.05410513
```

```
fviz_cluster(k4, data = fuel_3_Train[-c(1,2)])
```




```r
data1 <- bind_cols(fuel_3_Train, cluster = factor(k4$cluster))

# make a table to confirm it gives the same results as the original code

# using ggplot, make a point plot with "jitter" so each point is visible
# x-axis is species, y-axis is cluster, also coloured according to cluster
ggplot(data1) +
  geom_point(mapping = aes(x=data1[,1], y = cluster, colour = cluster),
             position = "jitter") +
  labs(title = "K Means")
```




```
s1<-data1%>%group_by(cluster)%>%summarise(avg_sulphur=mean(sulfur_content_pct),
                                          avg_ash=mean(ash_content_pct),
                                      avg_units=mean(fuel_received_units),
                                      avg_mmbtu=mean(fuel_mmbtu_per_unit),
                                      avg_cost=mean((fuel_cost_per_mmbtu)),
                                      supplier_count=n())%>%
  arrange(supplier_count)

s2<-data1%>%group_by(fuel_group_code)%>%summarise(avg_sulphur=mean(sulfur_content_pct),
                                                  avg_ash=mean(ash_content_pct),
                                      avg_units=mean(fuel_received_units),
                                      avg_mmbtu=mean(fuel_mmbtu_per_unit),
                                      avg_cost=mean(fuel_cost_per_mmbtu),
                                      supplier_count=n())%>%
  arrange(supplier_count)
s3<-data1%>%filter(fuel_group_code=='coal')%>%group_by(cluster)%>%
  summarise(avg_sulphur=mean(sulfur_content_pct),avg_ash=mean(ash_content_pct),
                                      avg_units=mean(fuel_received_units),
            avg_mmbtu=mean(fuel_mmbtu_per_unit),
                                      avg_cost=mean((fuel_cost_per_mmbtu)),
            supplier_count=n())%>%arrange(supplier_count)

s1
```

```
## # A tibble: 6 x 7
##   cluster avg_sulphur avg_ash avg_units avg_mmbtu avg_cost supplier_count
```

```
##   <fct>          <dbl>   <dbl>     <dbl>     <dbl>    <dbl>          <int>
## 1 5             0        0            1      1.06  8234.              1
## 2 6             1.31    40.8      21226.    13.1     11.5             126
## 3 4             0        0      2775108.     1.03     3.97            460
## 4 3             2.99    10.1      36193.    23.8      2.76           1085
## 5 2             0.595    7.81     56230.    20.5      2.73           2129
## 6 1             0.0288   0       134959.     1.77     7.61           5329
```

```
s2
```

```
## # A tibble: 5 x 7
##   fuel_group_code avg_sulphur avg_ash avg_units avg_mmbtu avg_cost supplier_co~1
##   <fct>                 <dbl>   <dbl>     <dbl>     <dbl>    <dbl>         <int>
## 1 other_gas             0       0       659113.     0.869     4.54            22
## 2 petroleum_coke        5.46    0.466    20966.    28.2       2.24            42
## 3 petroleum             0.186   0         5104.     5.83     14.9            823
## 4 coal                  1.35    9.93     48735.    21.2       3.08          3299
## 5 natural_gas           0       0       399887.     1.03      7.74          4944
## # ... with abbreviated variable name 1: supplier_count
```

```
s3
```

```
## # A tibble: 4 x 7
##   cluster avg_sulphur avg_ash avg_units avg_mmbtu avg_cost supplier_count
##   <fct>         <dbl>   <dbl>     <dbl>     <dbl>    <dbl>          <int>
## 1 1             0.4      0          258      12.5     2.65              1
## 2 6             1.31    40.8      21226.     13.1    11.5             126
## 3 3             2.89    10.5      36806.     23.7     2.78           1043
## 4 2             0.595    7.81     56230.     20.5     2.73           2129
```

Fuel Cost Prediction

Building regression models

```
lm1<-lm(fuel_cost_per_mmbtu~.,data=fuel_3_Train)
print(paste("R square of the model before adding clustering information is",
            summary(lm1)$r.squared))
```

```
## [1] "R square of the model before adding clustering information is 0.47582508078956"
```

```
lm2<-lm(fuel_cost_per_mmbtu~.,data=data1)
print(paste("R square of the model after adding clustering information is",
            summary(lm2)$r.squared))
```

```
## [1] "R square of the model after adding clustering information is 0.946604421898915"
```

It can be observed that r.square value before adding clustering information is 47.5% after adding clustering 94.66%