Fundamentals of Machine Learning

Final project

By

Joshna Katta

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(caret)
```

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

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

```
library(ISLR)
library(factoextra)
```

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

```
library(ggplot2)
```

```
fuel_data<- read.csv("C:/Users/Joshn/Downloads/fuel_receipts_costs_eia923.csv
",na.strings="")

colSums(is.na(fuel_data))
```

```
##                                    rowid
##                                        0
##                              plant_id_eia
##                                        0
##                         plant_id_eia_label
##                                    11165
##                               report_date
##                                        0
##                         contract_type_code
##                                      238
```

```
##                  contract_type_code_label
##                                       238
##                  contract_expiration_date
##                                    344302
##                       energy_source_code
##                                         1
##                  energy_source_code_label
##                                         1
##                       fuel_type_code_pudl
##                                         1
##                           fuel_group_code
##                                         1
##                              mine_id_pudl
##                                    391947
##                        mine_id_pudl_label
##                                    391947
##                                supplier_name
##                                         3
##                       fuel_received_units
##                                         0
##                       fuel_mmbtu_per_unit
##                                         0
##                        sulfur_content_pct
##                                         0
##                           ash_content_pct
##                                         0
##                       mercury_content_ppm
##                                    289482
##                       fuel_cost_per_mmbtu
##                                    200240
##         primary_transportation_mode_code
##                                     58192
##   primary_transportation_mode_code_label
##                                     58192
##       secondary_transportation_mode_code
##                                    575297
## secondary_transportation_mode_code_label
##                                    575297
##                natural_gas_transport_code
##                                    267663
##  natural_gas_delivery_contract_type_code
##                                    444190
##                      moisture_content_pct
##                                    516589
##                      chlorine_content_ppm
##                                    516589
##                             data_maturity
##                                         0
##                       data_maturity_label
##                                         0
```

```
fuel_data<-fuel_data[,-c(3,7,12,13,19,20,21,22,23,24,25,26,27,28)]

colSums(is.na(fuel_data))
```

```
##                     rowid             plant_id_eia              report_date
##                         0                        0                        0
##       contract_type_code contract_type_code_label       energy_source_code
##                       238                      238                        1
## energy_source_code_label      fuel_type_code_pudl          fuel_group_code
##                         1                        1                        1
##            supplier_name      fuel_received_units      fuel_mmbtu_per_unit
##                         3                        0                        0
##       sulfur_content_pct          ash_content_pct             data_maturity
##                         0                        0                        0
##      data_maturity_label
##                         0
```

```
set.seed(2312)
data2<-fuel_data %>% sample_frac(0.02)
index<-createDataPartition(data2$rowid,p=0.75,list=FALSE)

data2_Train<-data2[index,]
data2_Validation<-data2[-index,]


data2_Train<-na.omit(data2_Train)
data2_Validation<-na.omit(data2_Validation)
colSums(is.na(data2_Train))
```

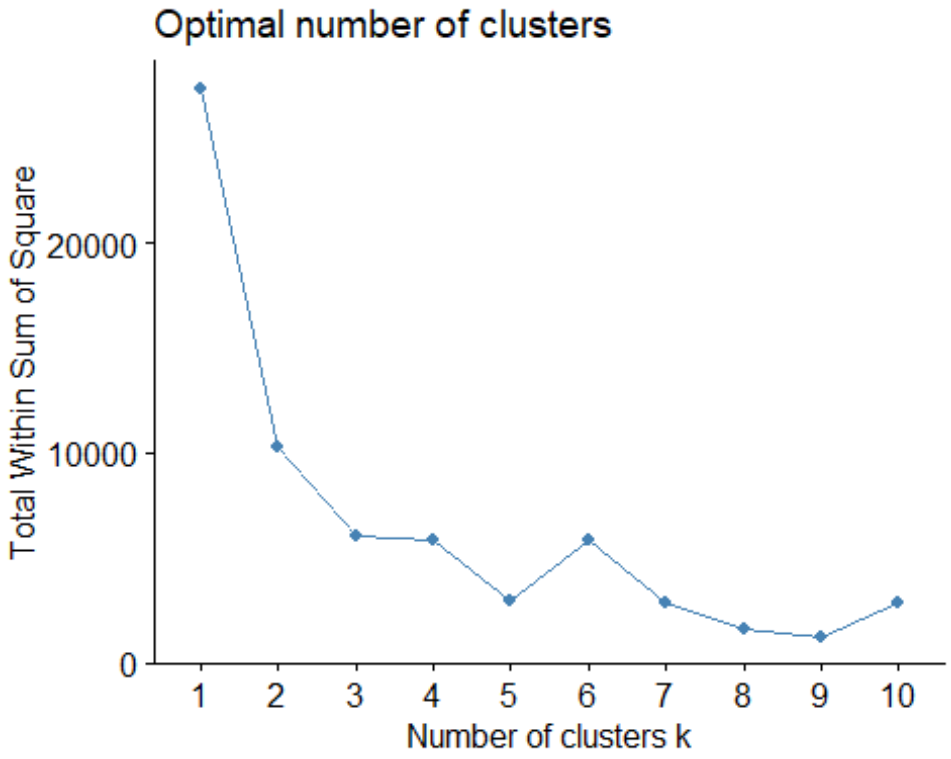
```
##                     rowid             plant_id_eia              report_date
##                         0                        0                        0
##       contract_type_code contract_type_code_label       energy_source_code
##                         0                        0                        0
## energy_source_code_label      fuel_type_code_pudl          fuel_group_code
##                         0                        0                        0
##            supplier_name      fuel_received_units      fuel_mmbtu_per_unit
##                         0                        0                        0
##       sulfur_content_pct          ash_content_pct             data_maturity
##                         0                        0                        0
##      data_maturity_label
##                         0
```

```
colSums(is.na(data2_Validation))
```

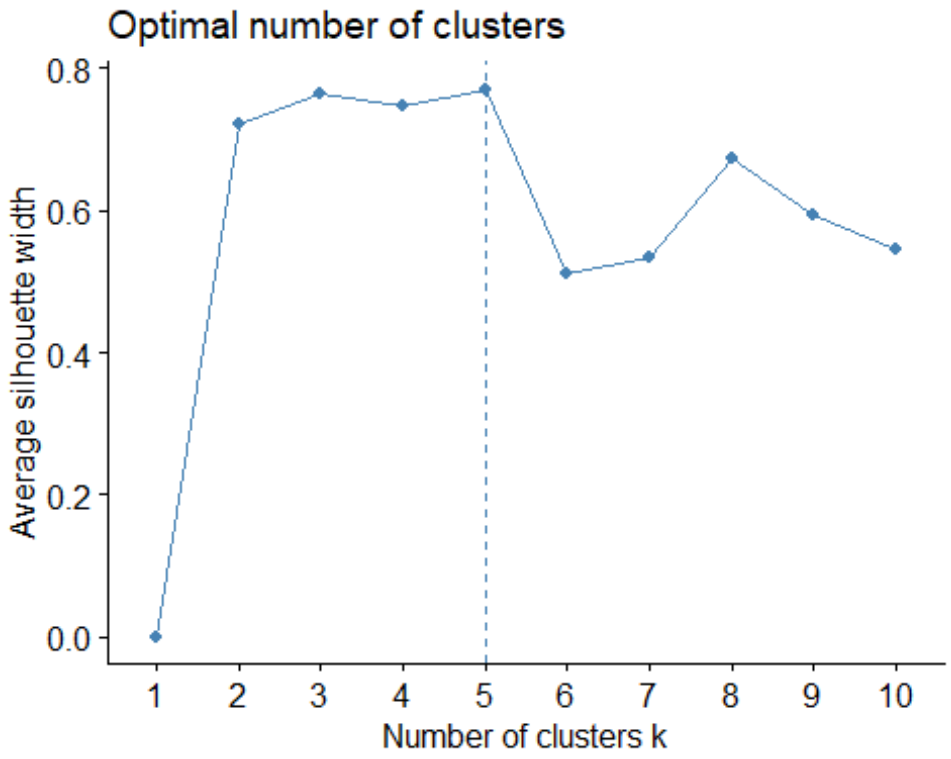
```
##                     rowid             plant_id_eia              report_date
##                         0                        0                        0
##       contract_type_code contract_type_code_label       energy_source_code
##                         0                        0                        0
## energy_source_code_label      fuel_type_code_pudl          fuel_group_code
##                         0                        0                        0
##            supplier_name      fuel_received_units      fuel_mmbtu_per_unit
```

```
##                          0                          0                          0
##       sulfur_content_pct          ash_content_pct            data_maturity
##                          0                          0                          0
##      data_maturity_label
##                          0
```

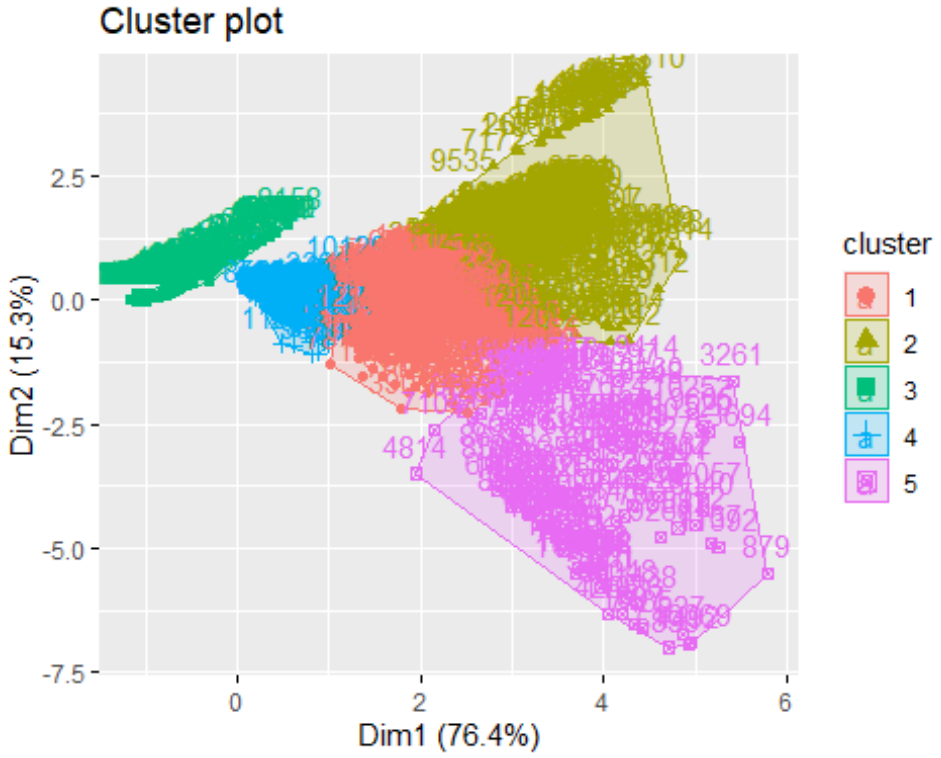
```
data3_Train<-data2_Train[,c(9,12,13,14)]
data3_Validation<-data2_Validation[,c(9,12,13,14)]

set.seed(111)
norm_model<-preProcess(data3_Train, method = c("center", "scale"))

data3_train_norm<-predict(norm_model,data3_Train)
data3_Validation_norm<-predict(norm_model,data3_Validation)

set.seed(1254)

fviz_nbclust(data3_train_norm[-1],kmeans,method='wss')
```



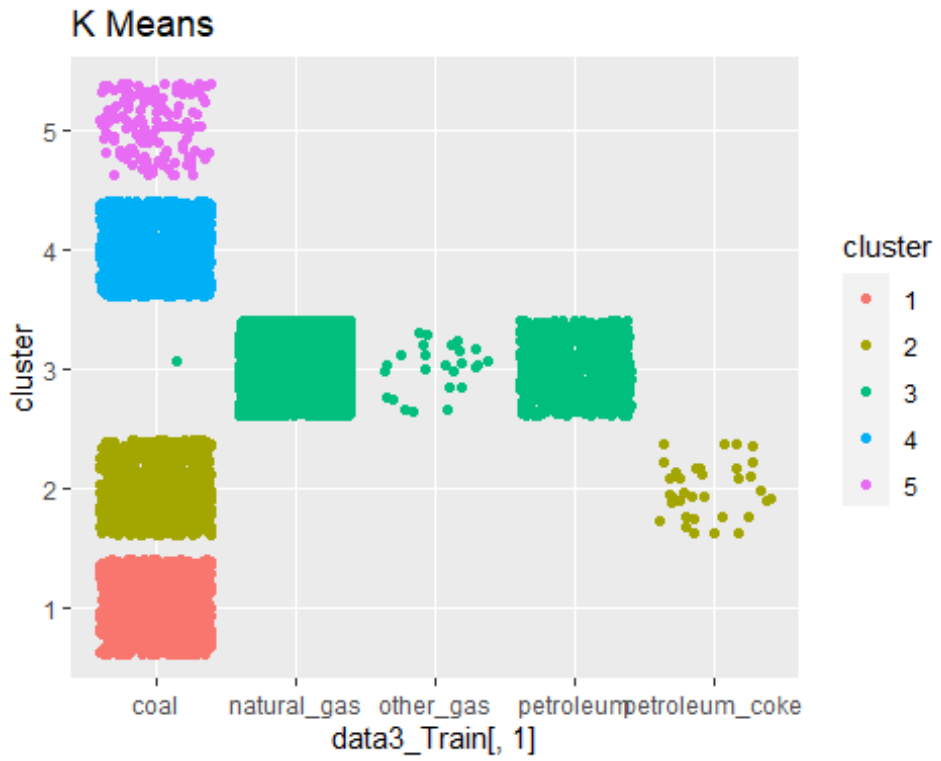

```
fviz_nbclust(data3_train_norm[-1],kmeans,method='silhouette')
```




```
set.seed(5467)
km<-kmeans(data3_train_norm[-1],centers=5,nstart = 25)
fviz_cluster(km,data=data3_train_norm[-1])
```



```
data3_Train$cluster<-as.factor(km$cluster)

ggplot(data3_Train)+geom_point(mapping=aes(x=data3_Train[,1],y=cluster,colour
=cluster),position='jitter')+labs(title='K Means')
```




```
data3_Train%>%group_by(fuel_group_code)%>%summarise(mean_sulphur=mean(sulfur_
content_pct),
                                                    mean_ash=mean(ash_content_pct),
                                                    avg_fuel_mmbtu_unit=mean(fuel_mmb
tu_per_unit))%>%arrange(mean_sulphur)

## # A tibble: 5 × 4
##   fuel_group_code mean_sulphur mean_ash avg_fuel_mmbtu_unit
##   <chr>                  <dbl>    <dbl>               <dbl>
## 1 natural_gas            0        0                    1.03
## 2 other_gas              0        0                    0.856
## 3 petroleum              0.183    0                    5.81
## 4 coal                   1.35    10.0                 21.3
## 5 petroleum_coke         5.43     0.435               28.2

data3_Train%>%group_by(cluster)%>%summarise(mean_sulphur=mean(sulfur_content_
pct),
                                                    mean_ash=mean(ash_content_pct),
                                                    mean_fuel_mmbtu_unit=mean(fuel_mm
btu_per_unit))%>%
  arrange(mean_ash)
```

```
## # A tibble: 5 × 4
##   cluster mean_sulphur mean_ash mean_fuel_mmbtu_unit
##   <fct>          <dbl>    <dbl>                <dbl>
## 1 3             0.0272     0                    1.74
## 2 4             0.299      5.30                17.6
## 3 2             3.21       9.66                23.9
## 4 1             1.11      11.2                 24.1
## 5 5             1.40      39.5                 13.7
```