FINAL PROJECT

Nikhil Reddy Addula

2022-12-08

```r
#loading library functions
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
```

```r
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v tibble  3.1.8     v purrr   0.3.4
## 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()
## x purrr::lift()   masks caret::lift()
```

```r
library(gridExtra)
```

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

```r
library(factoextra)
```

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

```r
library(ISLR)
library(flexclust)
```

```
## Loading required package: grid
## Loading required package: modeltools
## Loading required package: stats4
```

```
library(cluster)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

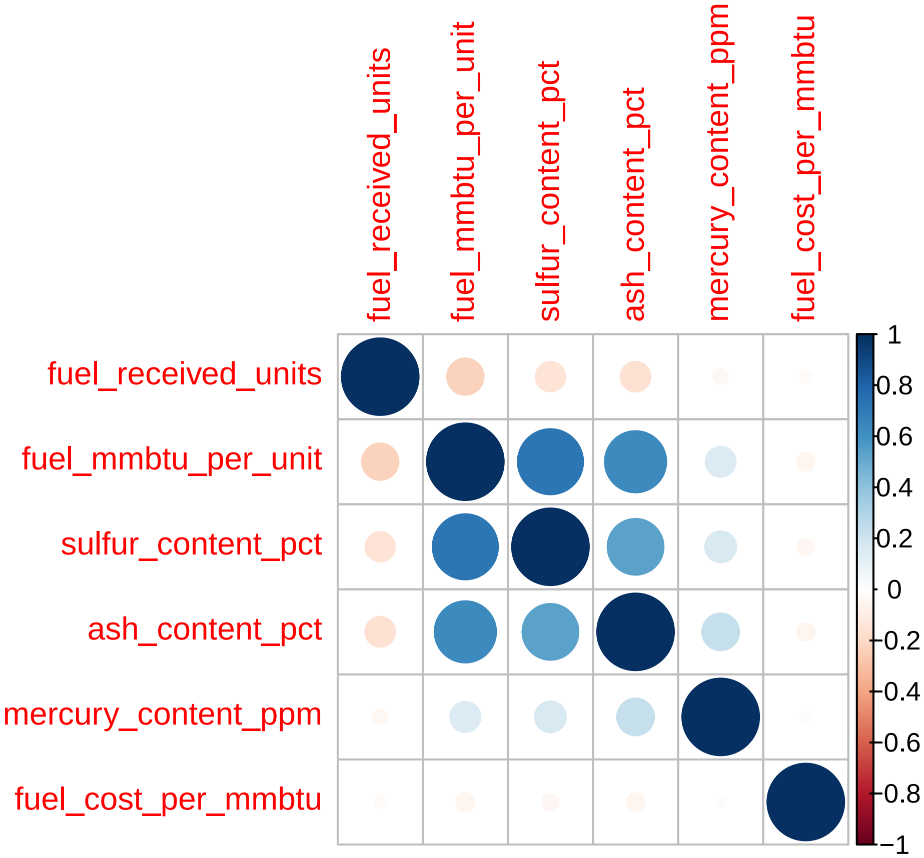
```
set.seed(1789)
#importing Data set and converting
getwd()
```

```
## [1] "/Users/nikhilreddya/Documents/assignments/FUNDAMENTALS ML/final"
```

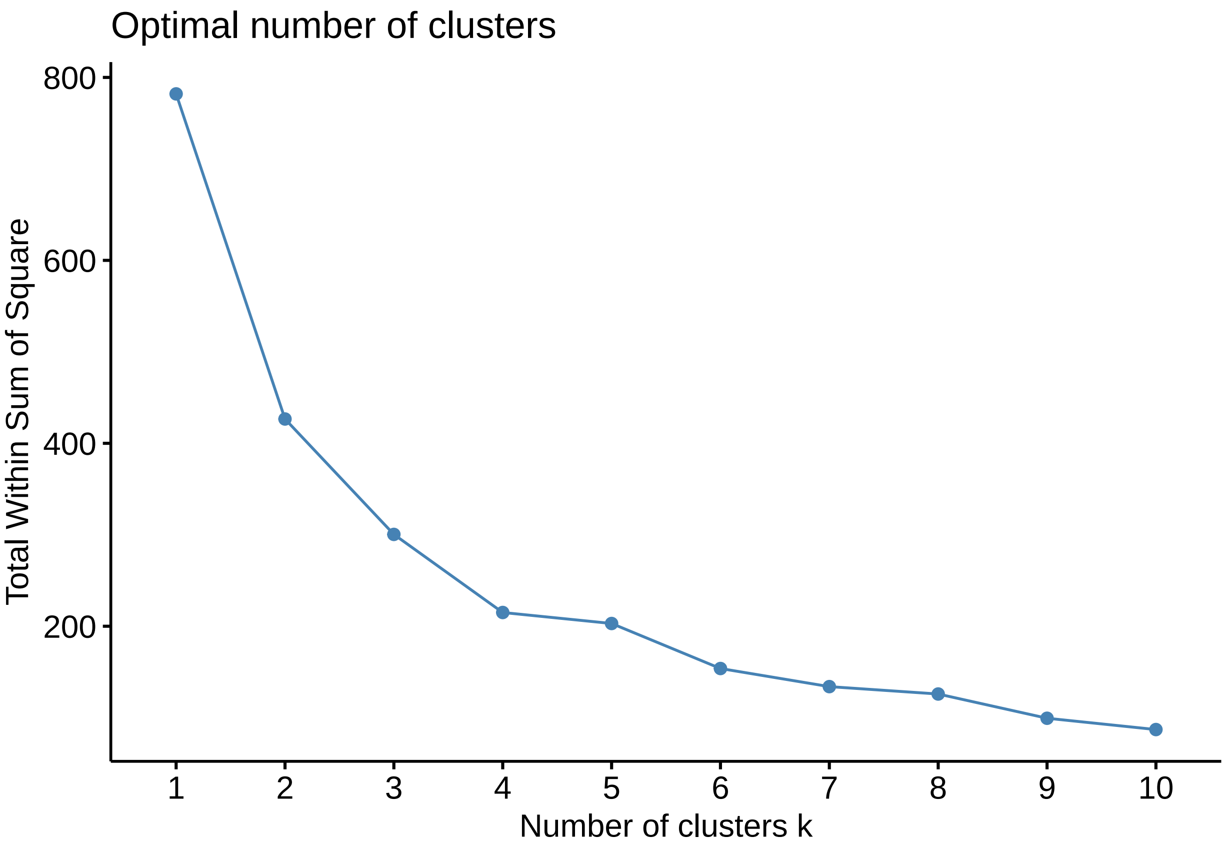
```
NR<-read.csv("~/Downloads/fuel_receipts_costs_eia923 (1).csv")

#Replace NA values with median
NR_1<-NR %>% replace(is.na(.), 0)
NR_2 <- NR_1%>% mutate(across(where(is.numeric), ~replace_na(., median(., na.rm=TRUE))))
```

```
#randomly sample about 2% of your data
Nr_model2<-NR_2%>%sample_frac(0.02)

#normalizing data using scale
norm_model<-preProcess(NR_2,method = c("scale"))
Nr_model2_normalized<-predict(norm_model,Nr_model2)

#75% of the sampled data as the training set
Index_t<-createDataPartition(Nr_model2$fuel_cost_per_mmbtu, p = 0.75,list = FALSE)
train<- Nr_model2_normalized[Index_t,]
test<-Nr_model2_normalized[-Index_t,]

#selecting/using the required columns for clustering
nr<-train[,c(15:20)]
```

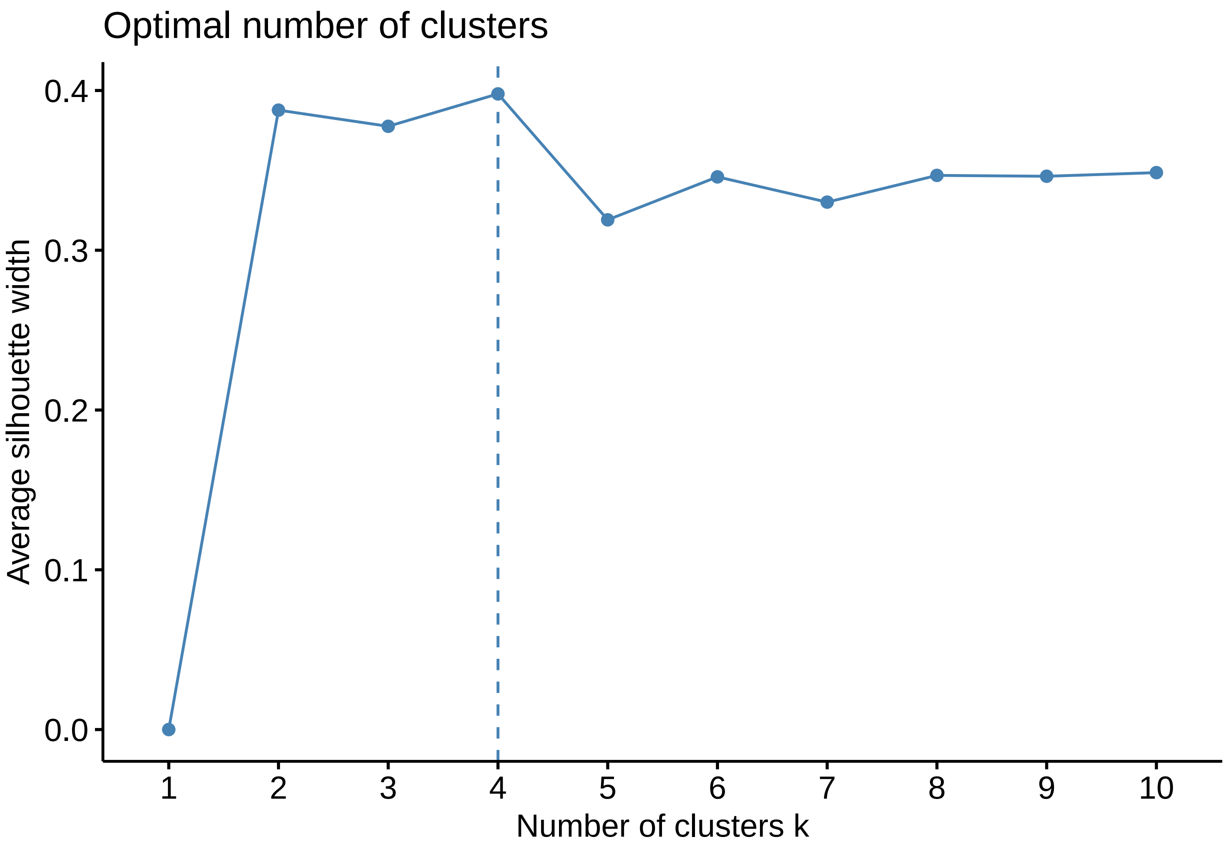
```
#correlation of columns selected
corrplot(cor(nr))
```



```
#using kmeans clustering with both the methods "WSS" & "silhouette" and getting the clusters points 'k'
set.seed(1789)
ANR<-Auto[,c(1,6)]
# Scaling the data frame (z-score)
ANR_1 <- scale(ANR)
fviz_nbclust(ANR_1, kmeans, method = "wss")
```



```
fviz_nbclust(ANR_1, kmeans, method = "silhouette")
```



```r
#After checking the above graph we found k=4
set.seed(1789)
anr <- kmeans(nr, centers = 4, nstart = 50)
shiloh_kmeans<- kmeans(nr,centers = 4,nstart = 50)
anr$centers # output the centers
```

```
##   fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct
## 1          0.18177742           0.4458332         0.07151265       0.1298978
## 2          3.95554628           0.1033619         0.00000000       0.0000000
## 3          0.05023818           2.3366447         1.90602700       1.8240368
## 4          0.03445104           2.0088691         2.14128102       3.5344152
##   mercury_content_ppm fuel_cost_per_mmbtu
## 1           0.0467504        0.0048009553
## 2           0.0000000        0.0014246208
## 3           0.1836899        0.0017970585
## 4          13.1404600        0.0000770555
```

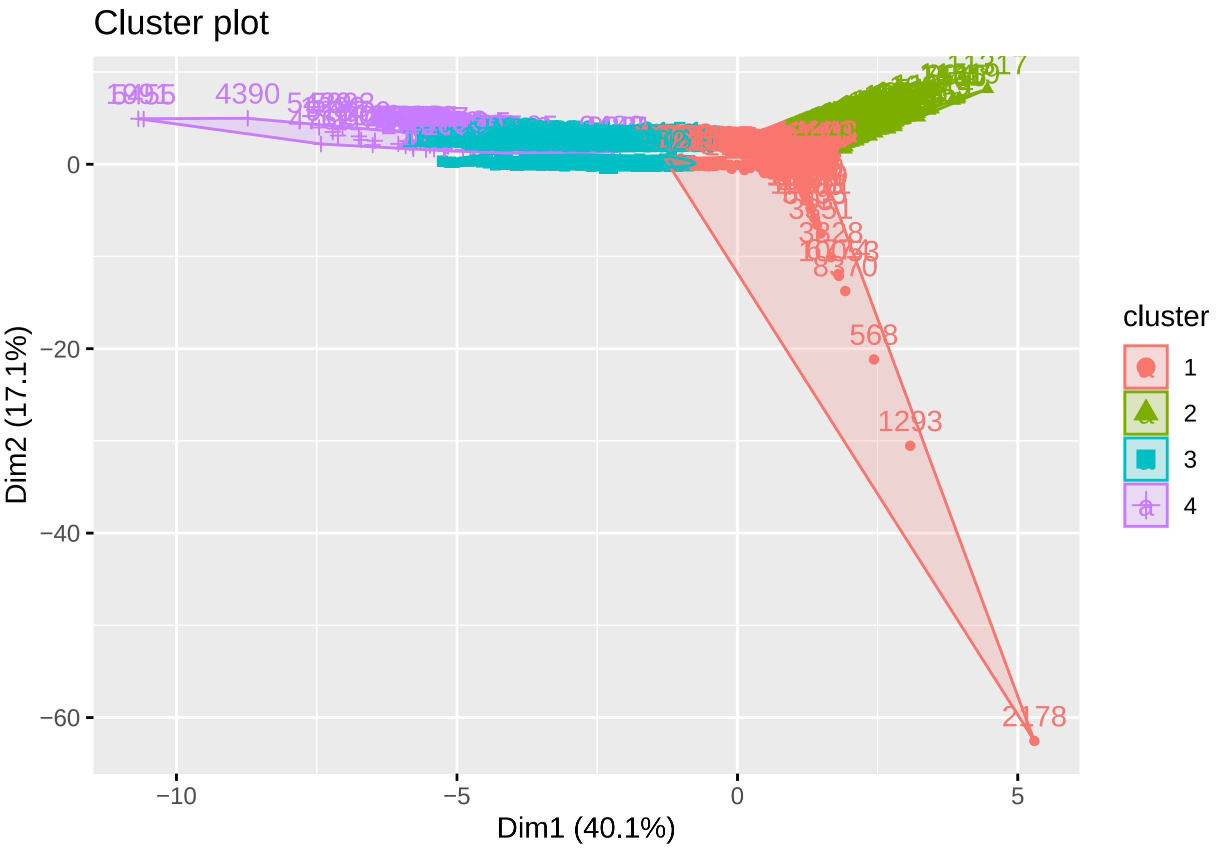
```r
anr$size # Number in each cluster
```

```
## [1] 6426  438 2214   51
```

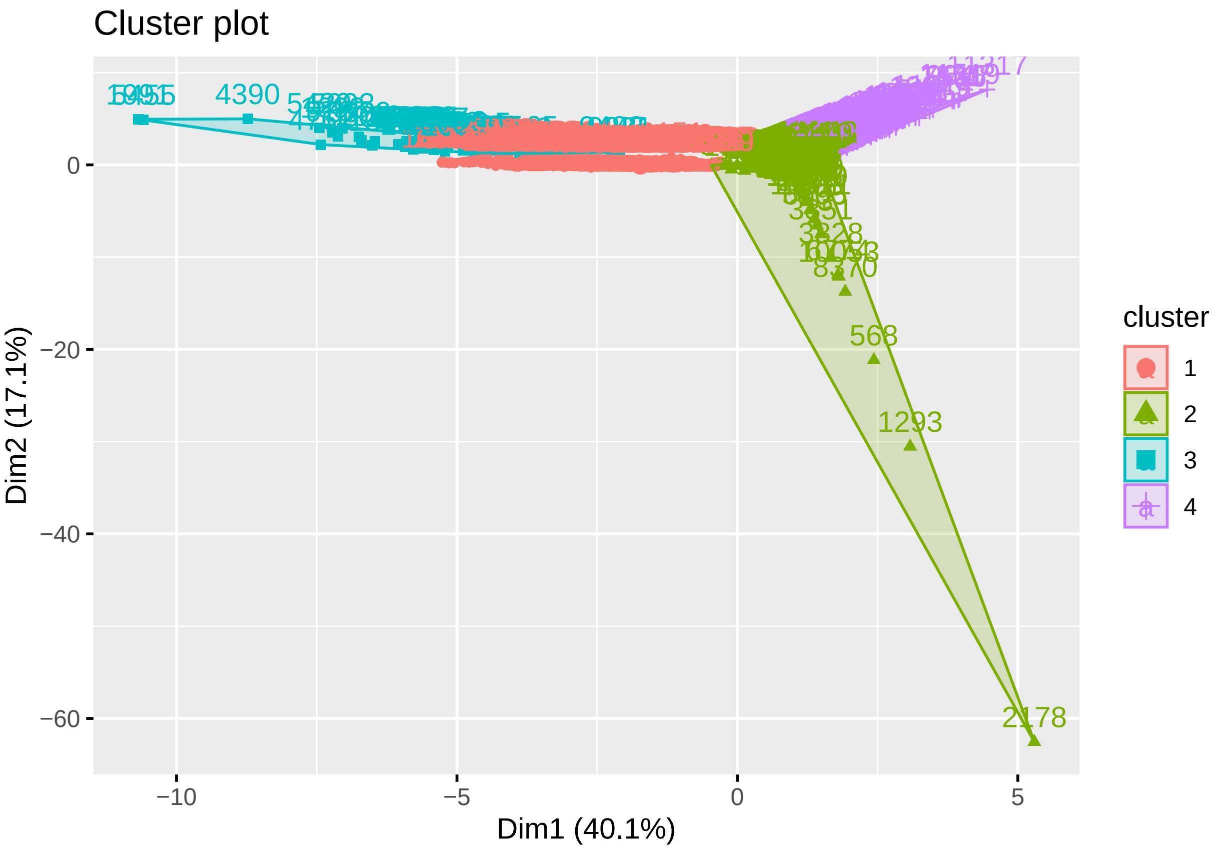
```r
anr$cluster["120"] # Identify the cluster of the 120th observation as an example
```

```
## 120
##   1
```

```
fviz_cluster(anr, data = nr) # Visualize the output
```



```
fviz_cluster(shiloh_kmeans,data = nr)
```



```
#finding the mean of clusters k=4
train$cluster<-anr$cluster
train%>%group_by(cluster)%>%summarise(avg_mmbtu=mean(fuel_mmbtu_per_unit),avg_fuel_recived =mean(fuel_r
```

```
## # A tibble: 4 x 6
##   cluster avg_mmbtu avg_fuel_recived avg_sulphur_content avg_ash_content avg_c~1
##     <int>     <dbl>            <dbl>               <dbl>           <dbl>   <dbl>
## 1       1     0.446           0.182               0.0715           0.130 4.80e-3
## 2       2     0.103           3.96                0                0     1.42e-3
## 3       3     2.34            0.0502              1.91             1.82  1.80e-3
## 4       4     2.01            0.0345              2.14             3.53  7.71e-5
## # ... with abbreviated variable name 1: avg_cost_perunit
```

Other Distances

```
set.seed(1789)
#kmeans clustering, using manhattan distance
k4 = kcca(nr, k=4, kccaFamily("kmedians"))
k4
```

```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = nr, k = 4, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
```

```
##    1    2    3    4
##   55 5772 1546 1756
```

```
# predict() function
clusters_index <- predict(k4)
dist(k4@centers)
```

```
##           1         2         3
## 2 12.057370
## 3 11.646943  1.884454
## 4 11.407657  3.593256  2.196640
```

As for the Sulphur ,ash & mercury content are less than 0.002 m they can be neglected # Cluster 1

This cluster recieves fuel of 0.18177742 .

As they are receiving low fuel,sulphur & ash their heat content in fuel(fuel_mmbtu) is also low (0.4458332).

The fuel cost per mmbtu is higher(0.0048009553) than all the 4 clusters formed.

Due to the high cost of fuel per mmbtu, this Cluster is not a favoured one to suggest to the US government.

Cluster 2

This cluster receives fuel of 3.95554628 which is high than all the clusters.

Their heat content in the fuel is very very low of 0.1033619 comapared to all the 4 clsuters.

The fuel cost per mmbtu is lower(0.0014246208) than all the 4 clusters formed.

This cluster is also not a preferred one to recommend for us Government because of fuel mmbtu per unit.

Cluster 3

This cluster receives fuel of 0.05023818 which is minimal.

Their heat content in the fuel is 2.3366447 which is good to the fuel recieves compared to other 3 clsuters.

The fuel cost per mmbtu is also very good(0.0017970585) to fuel recieved and the heat content.

This Cluster is the one that the US Government should be recommended since it takes all the variables, including (fuel recieved,heat content,fuel cost per mmbtu.