FINAL PROJECT

Elmy Luka

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
```

```r
Fuel_Data <- read.csv("/Users/ELMYLUKA/Desktop/MS BA/Fundamentals Of Machine Learning/Final_Project/Fuel
```

```r
#choosing the 4 numerical variables from the dataset and removing the null values.
data_1<-Fuel_Data[,c(10,15,16,20)]

#Checking NA
colMeans(is.na(data_1))
```

```
## fuel_type_code_pudl fuel_received_units fuel_mmbtu_per_unit fuel_cost_per_mmbtu
##           0.0000000           0.0000000           0.0000000           0.3290363
```

```r
#Removing missing values using imputation  for fuel_cost_per_mmbtu
data_1$fuel_cost_per_mmbtu [is.na(data_1$fuel_cost_per_mmbtu )]<-
  median(data_1$fuel_cost_per_mmbtu , na.rm = T)

nrow(data_1)
```

```
## [1] 608565
```

```r
#DATA PARTITION
#2% of the entire data set is considered and out of which the data has been split to  9000 train sets a

set.seed(1111)
#Trainset
data_1_partition <- createDataPartition(data_1$fuel_cost_per_mmbtu ,p=.015, list = FALSE)
Train <- data_1[data_1_partition,]
Exc_Data <- data_1[-data_1_partition,]

#Testset
data_2_partition <- createDataPartition(Exc_Data$fuel_cost_per_mmbtu,p=0.005,list=F)
Test <- Exc_Data[data_2_partition,]
Exc.Data.1 <- Exc_Data[-data_2_partition,]
```

```r
#Data Normalization
#(min-max normalization)

norm_data <- preProcess(Train[,-1],
                method=c("center","scale"))
train_norm <-predict(norm_data,Train)
test_norm <-predict(norm_data,Test)

nrow(train_norm)
```

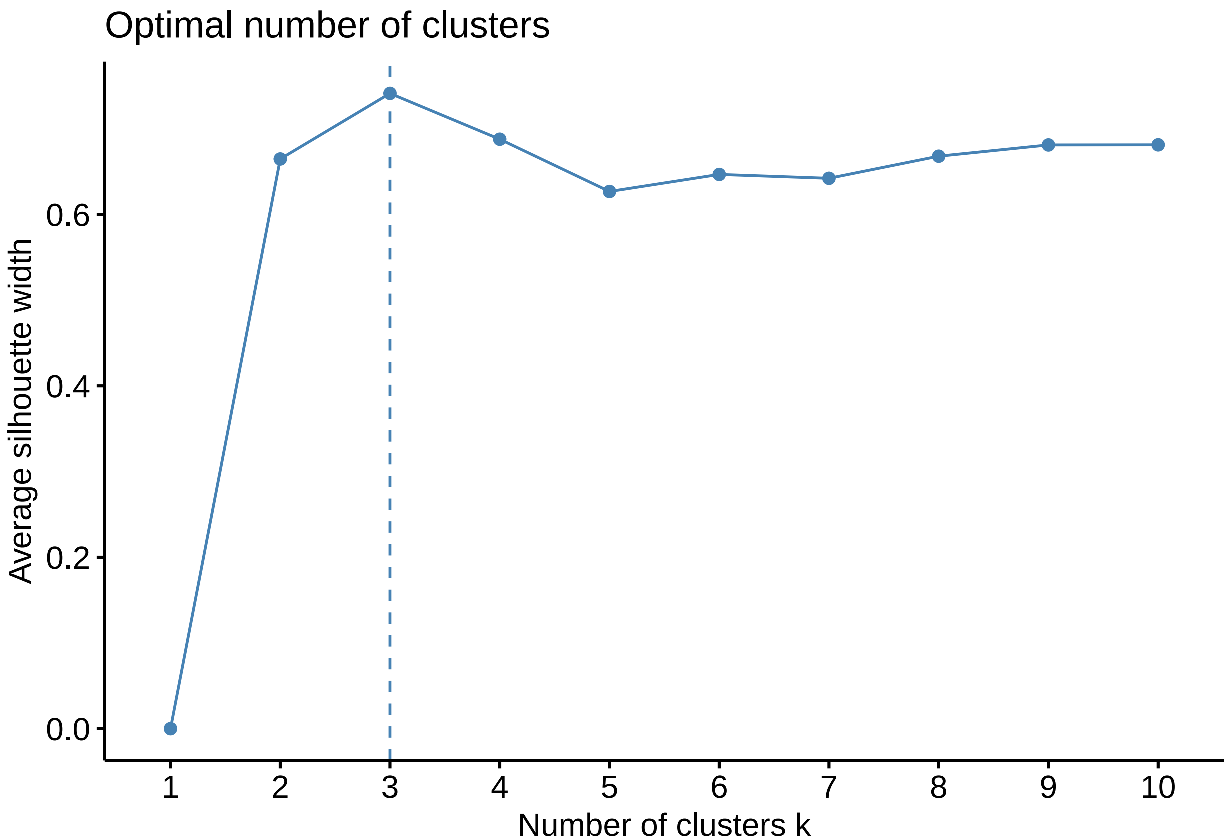
```
## [1] 9130
```

```r
nrow(test_norm)
```

```
## [1] 3000
```

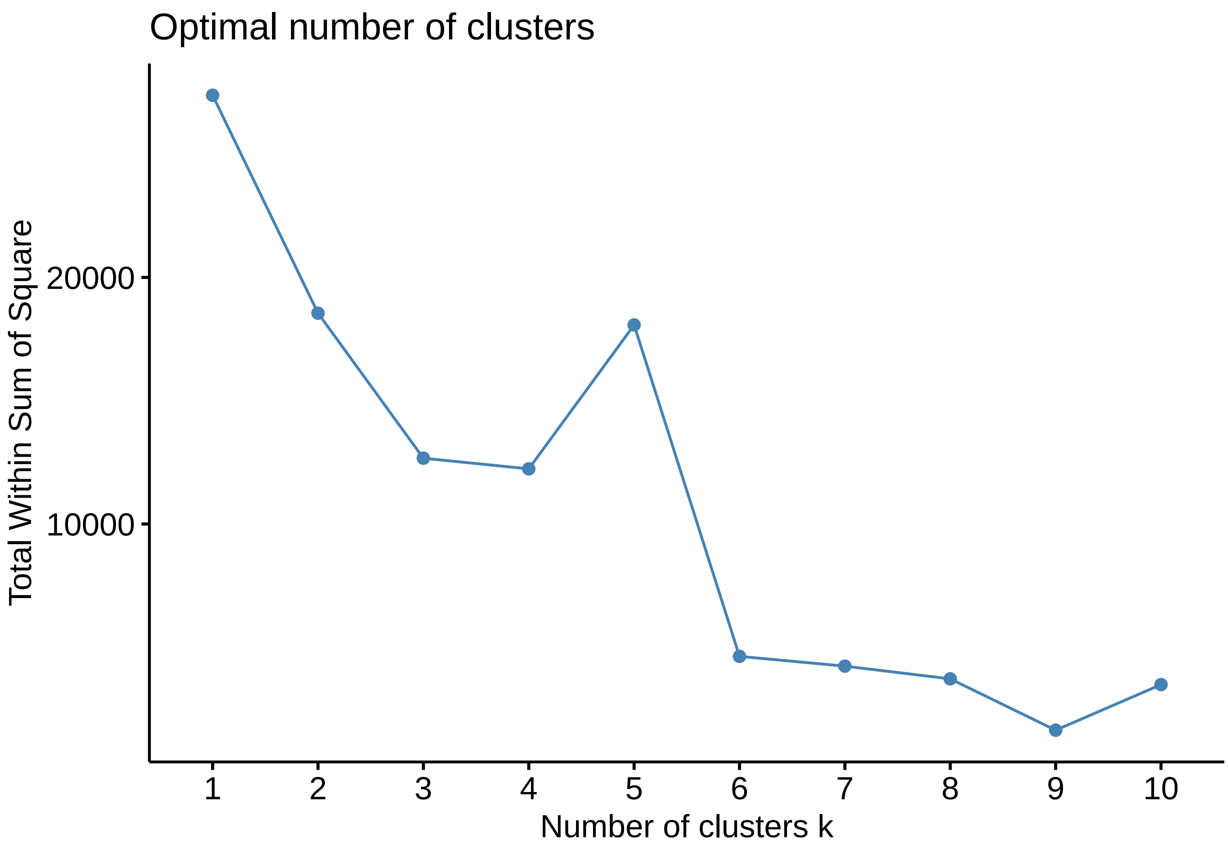
#kmeans clustering using the silhouette method.

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



##kmeans clustering using the wss method.

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

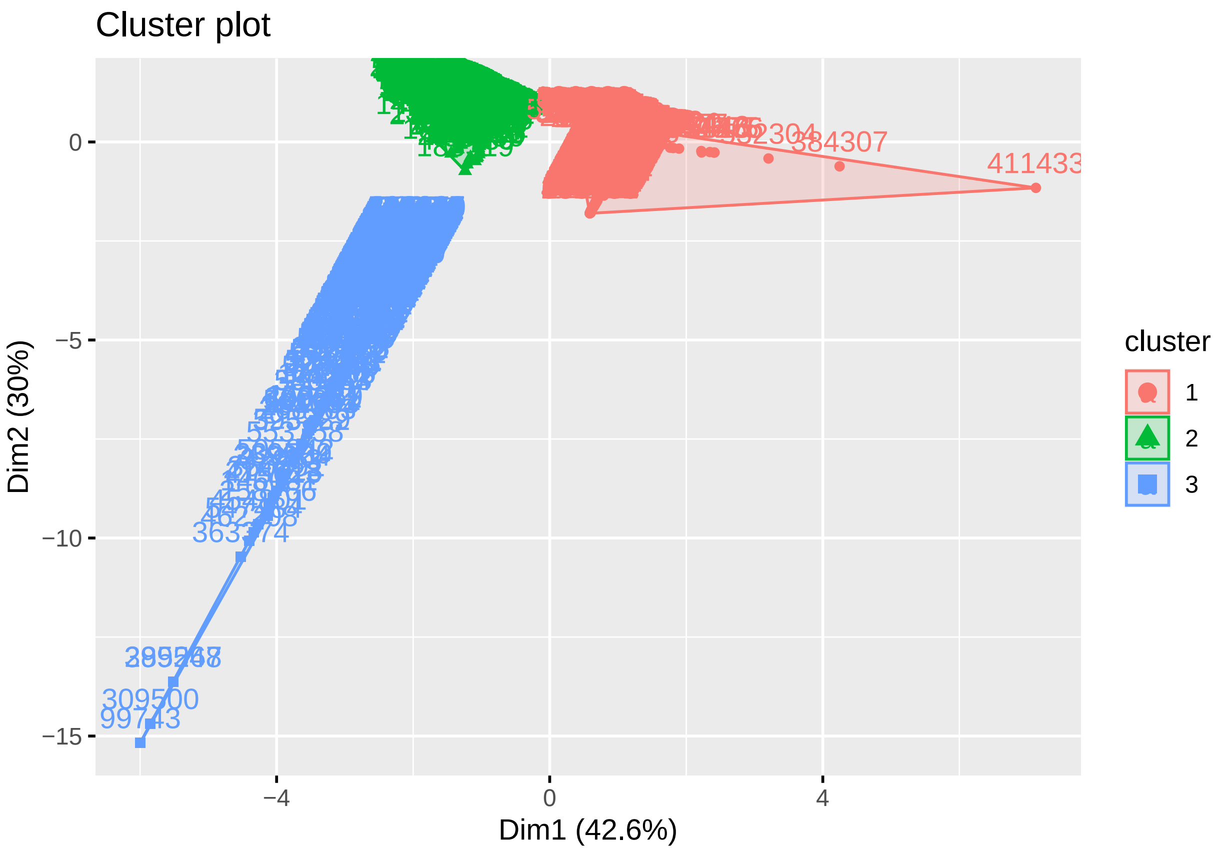



```r
#Plotting

set.seed(2222)
kmeans.df <- kmeans(train_norm[,-1], centers = 3, nstart = 25)
cluster <- kmeans.df$cluster

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

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




```r
#Using Group by to identify and summarize the clusters where a certain amount of each of the variables
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: 3 x 4
##   cluster median_units median_cost median_mmbtu
##     <int>        <dbl>       <dbl>        <dbl>
## 1       1        14188        3.28         1.03
## 2       2        21412        2.74        22.7
## 3       3     2446618.        3.28         1.03
```

```r
#identifying the natural resources that each of the clusters contain.
kmeans.df.1 %>% select(fuel_type_code_pudl,cluster) %>% group_by(cluster,fuel_type_code_pudl) %>% count
```

```
## # A tibble: 5 x 3
## # Groups:   cluster, fuel_type_code_pudl [5]
##   cluster fuel_type_code_pudl     n
##     <int> <chr>               <int>
## 1       1 coal                   45
## 2       1 gas                  4598
## 3       1 oil                   776
## 4       2 coal                 3275
## 5       3 gas                   436
```