FML_Project

Sandhya Cheepurupalli

2022-12-07

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.2
```

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

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

```
library(ISLR)
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(tidyr)
library("dbscan")
```

```
## Warning: package 'dbscan' was built under R version 4.2.2
```

```
##
## Attaching package: 'dbscan'
```

```
## The following object is masked from 'package:stats':
##
##     as.dendrogram
```

```
library('factoextra')
```

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

```
library('fpc')
```

```
## Warning: package 'fpc' was built under R version 4.2.2
```

```
##
## Attaching package: 'fpc'
```

```
## The following object is masked from 'package:dbscan':
##
##     dbscan
```

```
library(flexclust)
```

```
## Warning: package 'flexclust' was built under R version 4.2.2
```

```
## Loading required package: grid
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.2.2
```

```
##
## Attaching package: 'gridExtra'
```

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

```
setwd("C:\\Users\\sandh\\Downloads")
my.data<-read.csv("fuel_receipts_costs_eia923.csv",,na.strings = "")
```

```
my.data$report_date<-
  as.numeric(format(as.Date(my.data$report_date, format="%Y-%m-%d"),"%Y") )
mydata2<- select(my.data,-3,-7,-12,-13,-19,-20,-21,-22,-23,-24,-25,-26,-27,-28)
```

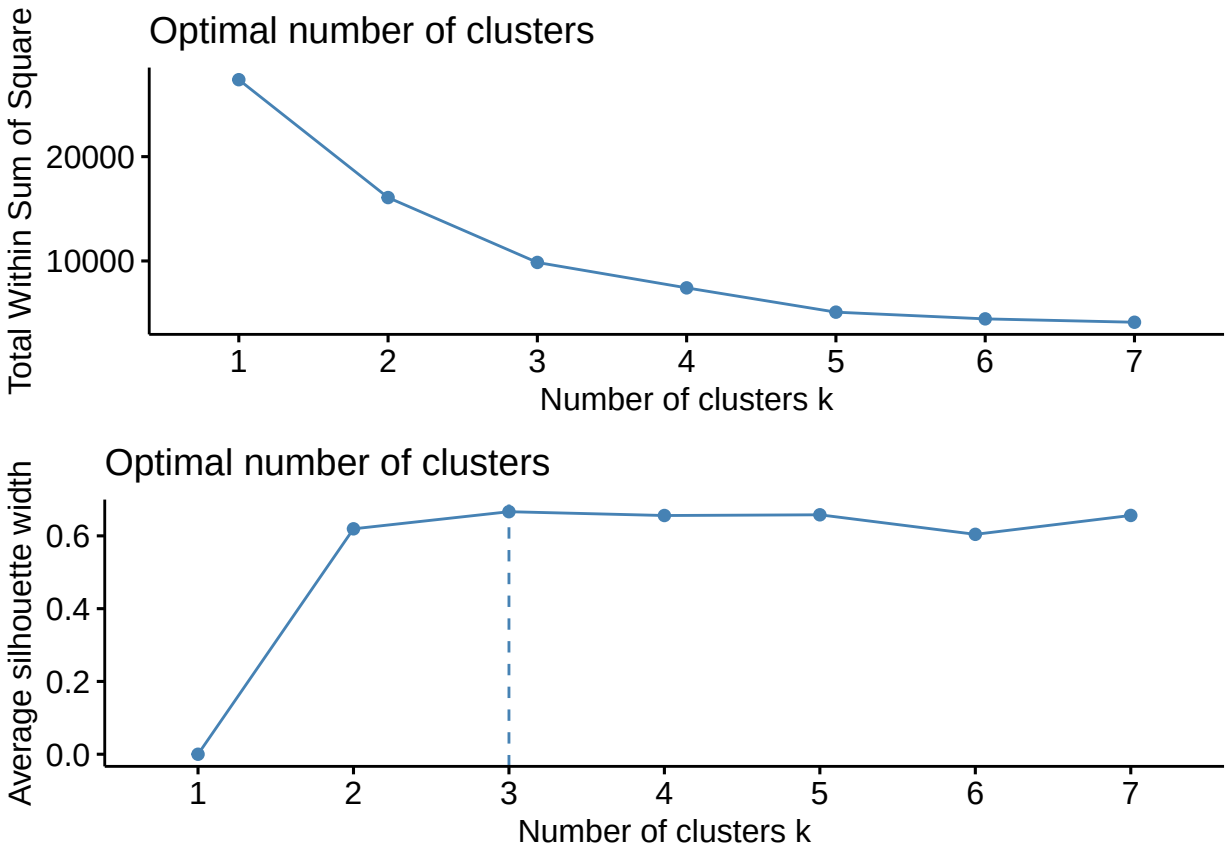
```
set.seed(2929)
rand_mydata2 <- mydata2%>%sample_frac(0.02)
rand_mydata2<-na.omit(rand_mydata2)
rand_mydata2<-rand_mydata2[,c(3,8,11,13,14)]
```

```
rand_mydata2$fuel_type_code_pudl<-as.factor(rand_mydata2$fuel_type_code_pudl)
rand_mydata2$report_date<-as.factor(rand_mydata2$report_date)
```

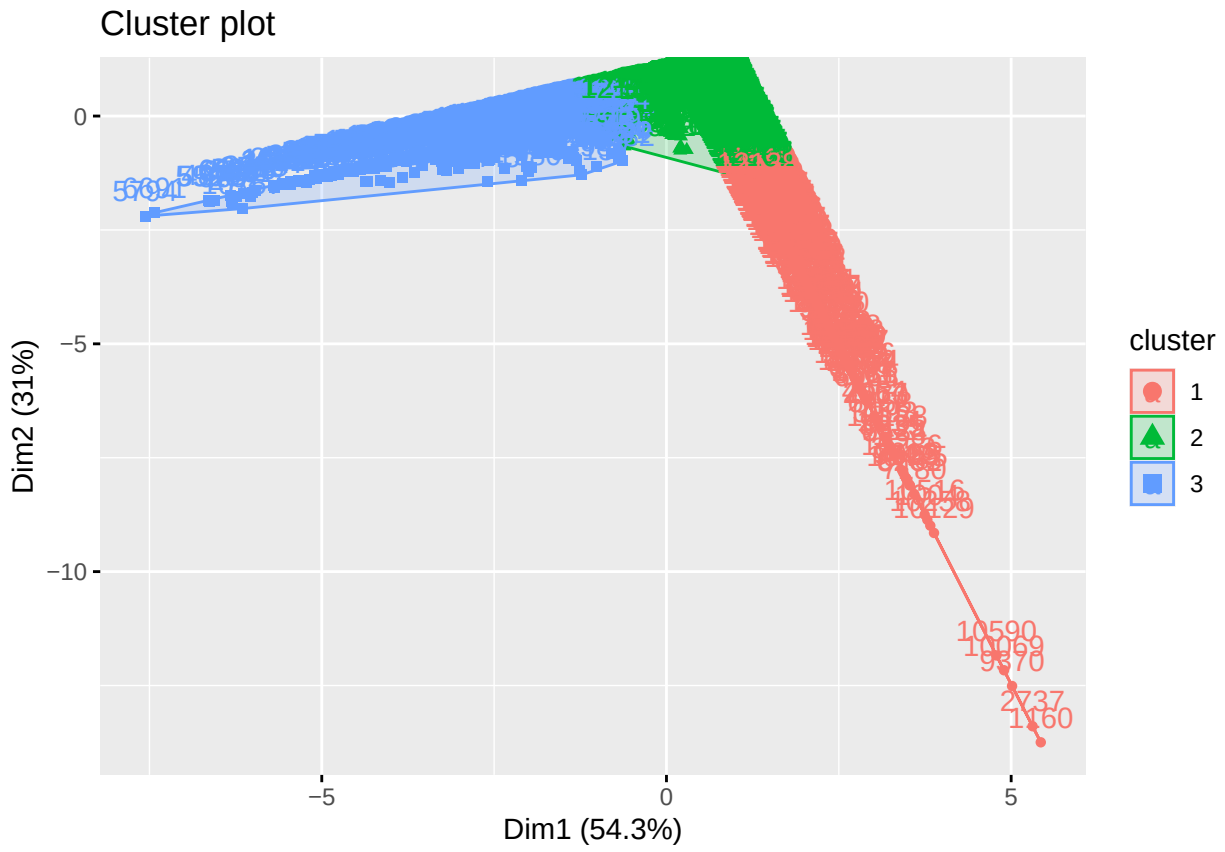
```
set.seed(2929)
Train_Index <-
  createDataPartition(rand_mydata2$fuel_type_code_pudl,p=0.75, list=FALSE)
Train_Data <- rand_mydata2[Train_Index,]
Test_Data <- rand_mydata2[-Train_Index,]
```

```
set.seed(123)
norm_model<-preProcess(Train_Data, method = c("center", "scale"))
train.norm <-predict(norm_model,Train_Data)
```

```
elbow<-fviz_nbclust(train.norm[,-c(1,2)], kmeans, method = "wss",k.max=7)
silhouette<-
  fviz_nbclust(train.norm[,-c(1,2)], kmeans, method = "silhouette",k.max=7)
grid.arrange(elbow,silhouette)
```

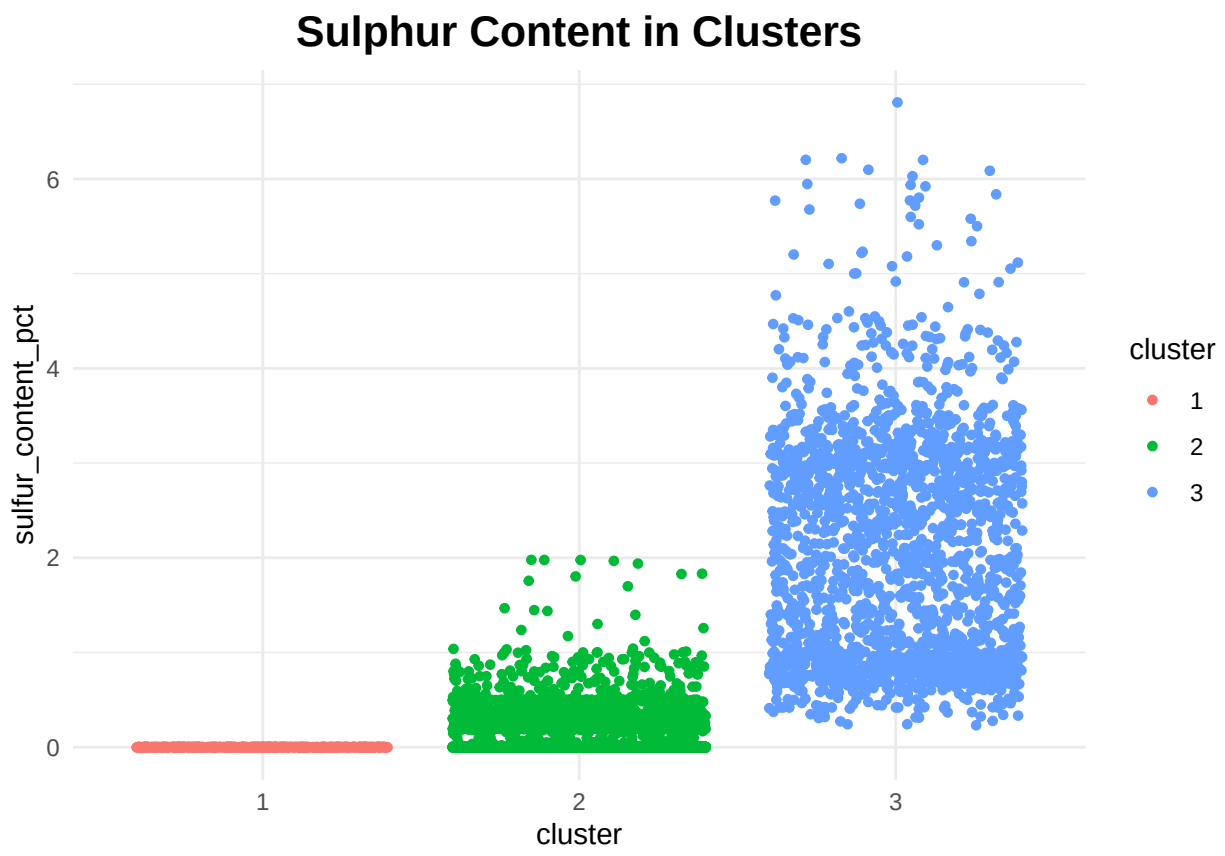



```
set.seed(245)
Kmeans1 <- kmeans(train.norm[,-c(1,2)], centers = 3, nstart = 25)
k2 <- fviz_cluster(Kmeans1, data = train.norm[,-c(1,2)])
k2
```




```
Train_Data$cluster<-as.factor(Kmeans1$cluster)

#Train_Data<-Train_Data%>%
  #mutate(cluster=case_when((cluster=='1')~
                             #'Environmental way of filling pocket',

                           #(cluster=='2')~'Balancing between environment and pocket',

                           #(cluster=='3')~'Environmental balance compromised'))

s1<-Train_Data%>%group_by(cluster)%>%
  summarise(avg_sulphur=mean(sulfur_content_pct),

                                        avg_ash=mean(ash_content_pct),

            avg_fuel_units=mean(fuel_received_units))
s2<-Train_Data%>%group_by(report_date)%>%
  summarise(avg_sulphur=mean(sulfur_content_pct),
                                        avg_ash=mean(ash_content_pct),

                                avg_fuel_units=mean(fuel_received_units))
s3<-Train_Data%>%group_by(cluster)%>%summarise(avg_sulphur=sum(sulfur_content_pct),
                                        avg_ash=sum(ash_content_pct),
                                        avg_fuel_units=sum(fuel_received_units))
```

```
ggplot(Train_Data) +
  aes(x = cluster, y = sulfur_content_pct, colour = cluster) +
  geom_jitter(size = 1.2) +
  scale_color_hue(direction = 1) +
  labs(title = "Sulphur Content in Clusters") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16L,
    face = "bold",
    hjust = 0.5)
  )
```




```
ggplot(Train_Data) +
  aes(
    x = cluster,
    y = ash_content_pct,
    fill = cluster,
    colour = cluster
  ) +
  geom_jitter(size = 1.2) +
  scale_fill_viridis_d(option = "plasma", direction = 1) +
  scale_color_viridis_d(option = "plasma", direction = 1) +
  labs(title = "Ash Content in Clusters") +
  theme_minimal() +
  theme(
```

```
    plot.title = element_text(size = 15L,
    face = "bold",
    hjust = 0.5)
)
```