ML Final Exam

Tongxiang Lu

2022-12-12

# Load the data.

fuel\_costs.df <- read.csv("fuel\_receipts\_costs\_eia923.csv")

# Set random four digit seed and split data.

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

set.seed(3272)  
fuel\_costs.df1<-createDataPartition(fuel\_costs.df$fuel\_type\_code\_pudl,p=0.02,list=F)

## Warning in createDataPartition(fuel\_costs.df$fuel\_type\_code\_pudl, p = 0.02, :  
## Some classes have a single record ( ) and these will be selected for the sample

Sampled\_Data = fuel\_costs.df[fuel\_costs.df1,]  
  
Train\_Index=createDataPartition(Sampled\_Data$fuel\_type\_code\_pudl, p=0.75, list=F)

## Warning in createDataPartition(Sampled\_Data$fuel\_type\_code\_pudl, p = 0.75, :  
## Some classes have a single record ( ) and these will be selected for the sample

Train\_Data=Sampled\_Data[Train\_Index,]  
Validation\_Data=Sampled\_Data[-Train\_Index,]

# Remove the missing value.

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

sapply(Train\_Data,function(x) sum(is.na(x)))

## rowid   
## 0   
## plant\_id\_eia   
## 0   
## plant\_id\_eia\_label   
## 0   
## report\_date   
## 0   
## contract\_type\_code   
## 0   
## contract\_type\_code\_label   
## 0   
## contract\_expiration\_date   
## 0   
## energy\_source\_code   
## 0   
## energy\_source\_code\_label   
## 0   
## fuel\_type\_code\_pudl   
## 0   
## fuel\_group\_code   
## 0   
## mine\_id\_pudl   
## 5884   
## mine\_id\_pudl\_label   
## 5884   
## supplier\_name   
## 0   
## fuel\_received\_units   
## 0   
## fuel\_mmbtu\_per\_unit   
## 0   
## sulfur\_content\_pct   
## 0   
## ash\_content\_pct   
## 0   
## mercury\_content\_ppm   
## 4290   
## fuel\_cost\_per\_mmbtu   
## 3084   
## primary\_transportation\_mode\_code   
## 0   
## primary\_transportation\_mode\_code\_label   
## 0   
## secondary\_transportation\_mode\_code   
## 0   
## secondary\_transportation\_mode\_code\_label   
## 0   
## natural\_gas\_transport\_code   
## 0   
## natural\_gas\_delivery\_contract\_type\_code   
## 0   
## moisture\_content\_pct   
## 7718   
## chlorine\_content\_ppm   
## 7718   
## data\_maturity   
## 0   
## data\_maturity\_label   
## 0

head(Train\_Data)

## rowid plant\_id\_eia plant\_id\_eia\_label report\_date contract\_type\_code  
## 23 23 26 E C Gaston 2008-01-01 C  
## 81 81 99 Frederickson 2008-01-01 S  
## 126 126 136 Seminole 2008-01-01 C  
## 250 250 535 McClellan 2008-01-01 S  
## 375 375 642 Scholz 2008-01-01 C  
## 397 397 667 Northside 2008-01-01 C  
## contract\_type\_code\_label contract\_expiration\_date energy\_source\_code  
## 23 C 2008-06-01 BIT  
## 81 S NG  
## 126 C 2012-12-01 BIT  
## 250 S NG  
## 375 C 2009-02-01 BIT  
## 397 C 2009-08-01 RFO  
## energy\_source\_code\_label fuel\_type\_code\_pudl fuel\_group\_code mine\_id\_pudl  
## 23 BIT coal coal 9  
## 81 NG gas natural\_gas NA  
## 126 BIT coal coal 26  
## 250 NG gas natural\_gas NA  
## 375 BIT coal coal 88  
## 397 RFO oil petroleum NA  
## mine\_id\_pudl\_label supplier\_name fuel\_received\_units fuel\_mmbtu\_per\_unit  
## 23 9 t c sales 27907 24.062  
## 81 NA conoco 229 1.030  
## 126 26 alliance coal 34199 24.434  
## 250 NA powerex 95 1.000  
## 375 88 alpha coal 10177 22.040  
## 397 NA bp 6430 6.485  
## sulfur\_content\_pct ash\_content\_pct mercury\_content\_ppm fuel\_cost\_per\_mmbtu  
## 23 1.86 15.0 NA 2.496  
## 81 0.00 0.0 NA 9.494  
## 126 2.98 8.5 NA 2.201  
## 250 0.00 0.0 NA 8.836  
## 375 0.80 9.7 NA 3.126  
## 397 1.44 0.0 NA 6.993  
## primary\_transportation\_mode\_code primary\_transportation\_mode\_code\_label  
## 23 RR RR  
## 81   
## 126 RR RR  
## 250   
## 375 RR RR  
## 397 WT WT  
## secondary\_transportation\_mode\_code secondary\_transportation\_mode\_code\_label  
## 23   
## 81   
## 126 RR RR  
## 250   
## 375   
## 397   
## natural\_gas\_transport\_code natural\_gas\_delivery\_contract\_type\_code  
## 23 firm   
## 81 firm   
## 126   
## 250 firm   
## 375 firm   
## 397   
## moisture\_content\_pct chlorine\_content\_ppm data\_maturity data\_maturity\_label  
## 23 NA NA final final  
## 81 NA NA final final  
## 126 NA NA final final  
## 250 NA NA final final  
## 375 NA NA final final  
## 397 NA NA final final

# Set only numerical variable for my train data analysis.

fuel\_costs.df1 <- Train\_Data[c(2,15,16,17,18,20)]  
head(fuel\_costs.df1)

## plant\_id\_eia fuel\_received\_units fuel\_mmbtu\_per\_unit sulfur\_content\_pct  
## 23 26 27907 24.062 1.86  
## 81 99 229 1.030 0.00  
## 126 136 34199 24.434 2.98  
## 250 535 95 1.000 0.00  
## 375 642 10177 22.040 0.80  
## 397 667 6430 6.485 1.44  
## ash\_content\_pct fuel\_cost\_per\_mmbtu  
## 23 15.0 2.496  
## 81 0.0 9.494  
## 126 8.5 2.201  
## 250 0.0 8.836  
## 375 9.7 3.126  
## 397 0.0 6.993

summary(fuel\_costs.df1)

## plant\_id\_eia fuel\_received\_units fuel\_mmbtu\_per\_unit sulfur\_content\_pct  
## Min. : 3 Min. : 1 Min. : 0.082 Min. :0.000   
## 1st Qu.: 2721 1st Qu.: 3427 1st Qu.: 1.025 1st Qu.:0.000   
## Median : 6181 Median : 20980 Median : 1.062 Median :0.000   
## Mean :18636 Mean : 233464 Mean : 8.845 Mean :0.521   
## 3rd Qu.:50776 3rd Qu.: 98826 3rd Qu.:17.800 3rd Qu.:0.500   
## Max. :63688 Max. :12597588 Max. :30.000 Max. :7.980   
##   
## ash\_content\_pct fuel\_cost\_per\_mmbtu  
## Min. : 0.00 Min. : 0.188   
## 1st Qu.: 0.00 1st Qu.: 2.258   
## Median : 0.00 Median : 3.256   
## Mean : 3.63 Mean : 9.702   
## 3rd Qu.: 5.80 3rd Qu.: 4.743   
## Max. :61.40 Max. :13464.320   
## NA's :3084

# Drop column 6 as it has significant number of missing data.

fuel\_costs.df2 <- fuel\_costs.df1[,-6]  
summary(fuel\_costs.df2)

## plant\_id\_eia fuel\_received\_units fuel\_mmbtu\_per\_unit sulfur\_content\_pct  
## Min. : 3 Min. : 1 Min. : 0.082 Min. :0.000   
## 1st Qu.: 2721 1st Qu.: 3427 1st Qu.: 1.025 1st Qu.:0.000   
## Median : 6181 Median : 20980 Median : 1.062 Median :0.000   
## Mean :18636 Mean : 233464 Mean : 8.845 Mean :0.521   
## 3rd Qu.:50776 3rd Qu.: 98826 3rd Qu.:17.800 3rd Qu.:0.500   
## Max. :63688 Max. :12597588 Max. :30.000 Max. :7.980   
## ash\_content\_pct  
## Min. : 0.00   
## 1st Qu.: 0.00   
## Median : 0.00   
## Mean : 3.63   
## 3rd Qu.: 5.80   
## Max. :61.40

# Set only numerical variable for my test data analysis.

fuel\_costs.df1\_valid <- Validation\_Data[c(2,15,16,17,18,20)]  
summary(fuel\_costs.df1\_valid)

## plant\_id\_eia fuel\_received\_units fuel\_mmbtu\_per\_unit sulfur\_content\_pct  
## Min. : 3 Min. : 1 Min. : 0.283 Min. :0.0000   
## 1st Qu.: 2721 1st Qu.: 3742 1st Qu.: 1.025 1st Qu.:0.0000   
## Median : 6139 Median : 20798 Median : 1.061 Median :0.0000   
## Mean :18274 Mean : 238712 Mean : 8.862 Mean :0.5138   
## 3rd Qu.:50498 3rd Qu.: 107664 3rd Qu.:17.773 3rd Qu.:0.4700   
## Max. :63688 Max. :11237212 Max. :29.220 Max. :7.2400   
##   
## ash\_content\_pct fuel\_cost\_per\_mmbtu  
## Min. : 0.000 Min. : -6.310   
## 1st Qu.: 0.000 1st Qu.: 2.272   
## Median : 0.000 Median : 3.245   
## Mean : 3.601 Mean : 6.609   
## 3rd Qu.: 5.800 3rd Qu.: 4.829   
## Max. :60.700 Max. :1939.507   
## NA's :1040

# Drop column 6 as it has significant number of missing data.

fuel\_costs.df2\_valid <- fuel\_costs.df1\_valid[,-6]  
summary(fuel\_costs.df2\_valid)

## plant\_id\_eia fuel\_received\_units fuel\_mmbtu\_per\_unit sulfur\_content\_pct  
## Min. : 3 Min. : 1 Min. : 0.283 Min. :0.0000   
## 1st Qu.: 2721 1st Qu.: 3742 1st Qu.: 1.025 1st Qu.:0.0000   
## Median : 6139 Median : 20798 Median : 1.061 Median :0.0000   
## Mean :18274 Mean : 238712 Mean : 8.862 Mean :0.5138   
## 3rd Qu.:50498 3rd Qu.: 107664 3rd Qu.:17.773 3rd Qu.:0.4700   
## Max. :63688 Max. :11237212 Max. :29.220 Max. :7.2400   
## ash\_content\_pct   
## Min. : 0.000   
## 1st Qu.: 0.000   
## Median : 0.000   
## Mean : 3.601   
## 3rd Qu.: 5.800   
## Max. :60.700

# Normalize data.

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ tibble 3.1.7 ✔ purrr 0.3.4  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.2  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ purrr::lift() masks caret::lift()

fuel\_costs.df3 <- scale(fuel\_costs.df2[,1:5])  
head(fuel\_costs.df3)

## plant\_id\_eia fuel\_received\_units fuel\_mmbtu\_per\_unit sulfur\_content\_pct  
## 23 -0.8133595 -0.2906048 1.5500835 1.3213046  
## 81 -0.8101690 -0.3297344 -0.7961466 -0.5140515  
## 126 -0.8085519 -0.2817096 1.5879785 2.4264653  
## 250 -0.7911134 -0.3299238 -0.7992027 -0.5140515  
## 375 -0.7864369 -0.3156705 1.3441059 0.2753490  
## 397 -0.7853443 -0.3209678 -0.2404552 0.9068694  
## ash\_content\_pct  
## 23 1.7094611  
## 81 -0.5458617  
## 126 0.7321545  
## 250 -0.5458617  
## 375 0.9125804  
## 397 -0.5458617

fuel\_costs.df3\_valid<-scale(fuel\_costs.df2\_valid[,1:5])  
head(fuel\_costs.df3\_valid)

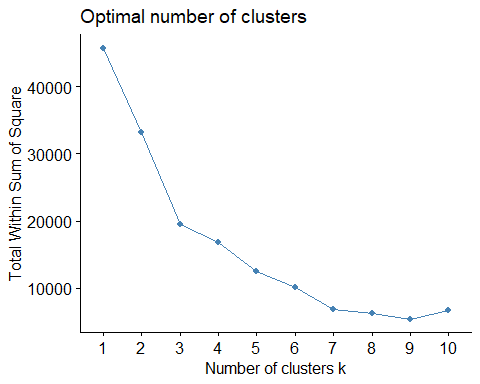
## plant\_id\_eia fuel\_received\_units fuel\_mmbtu\_per\_unit sulfur\_content\_pct  
## 37 -0.8025552 -0.3479648 1.5627664 -0.1140933  
## 155 -0.7945856 -0.3508313 -0.7960437 -0.5151702  
## 310 -0.7797911 -0.2741325 1.6471149 0.5476835  
## 422 -0.7736708 -0.2865769 1.6908135 0.4173335  
## 1750 -0.6741604 -0.3499650 -0.7980762 -0.5151702  
## 2072 -0.6584853 -0.3343688 1.6196762 0.2970105  
## ash\_content\_pct  
## 37 0.7278437  
## 155 -0.5461576  
## 310 1.0766774  
## 422 0.6368436  
## 1750 -0.5461576  
## 2072 1.1676775

# Trying to find the optimal k

library(factoextra)

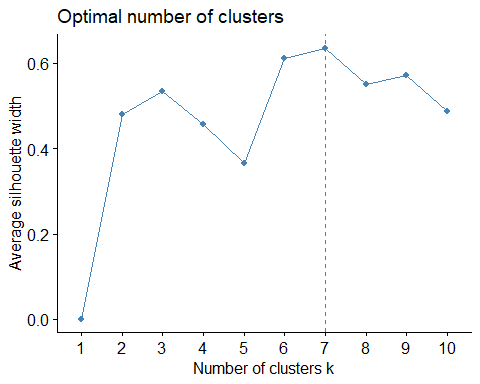
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

wss <- fviz\_nbclust(fuel\_costs.df3,kmeans,method="wss")  
wss



# It is very ambiguous to find the optimal K.

silhouete <- fviz\_nbclust(fuel\_costs.df3,kmeans,method="silhouette")  
silhouete

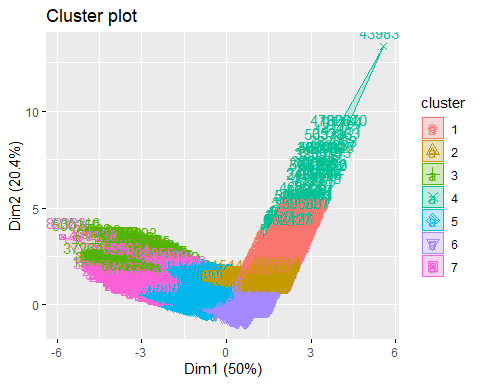


# I find my optimal K=7 by silhouete method.

# By WSS and silhouette method comparasion, I find silouette method more precise,

so I choose running kmeans k=7.

cluster.kmean <- kmeans(fuel\_costs.df3,centers=7,nstart=25)  
cluster.k1<-kmeans(fuel\_costs.df3,centers = 7, nstart = 25)  
fviz\_cluster(cluster.k1, data = fuel\_costs.df3)



cluster.k1$size

## [1] 398 2015 157 66 2126 3350 1020

cluster.k1$centers

## plant\_id\_eia fuel\_received\_units fuel\_mmbtu\_per\_unit sulfur\_content\_pct  
## 1 1.1545003 2.80208717 -0.7962344 -0.51405145  
## 2 1.5873003 -0.07744249 -0.7353519 -0.48028647  
## 3 0.5675617 -0.30984285 0.5236354 0.81410226  
## 4 0.4233654 7.93294810 -0.8024069 -0.51405145  
## 5 -0.5417620 -0.25902963 1.1991241 0.08384699  
## 6 -0.6182863 -0.17738288 -0.6995372 -0.48700565  
## 7 -0.5410837 -0.28351340 1.5328385 2.48205223  
## ash\_content\_pct  
## 1 -0.5458617  
## 2 -0.5328640  
## 3 5.2649225  
## 4 -0.5458617  
## 5 0.6476524  
## 6 -0.5458617  
## 7 0.9334664

# 

# Interpretation

Cluster 1

This cluster receives a very high volume of fuel（big positive numbers) with low ash and sulfur(negative numbers), which means this cluster accumulated pure fuel. Thus, I conclude this cluster with surplus inflow of pure fuel. Additionally, the positive number of plant indicates that the plant with larger identification number is relevant to the large volume of fuel.

Cluster 2

This cluster is identified as everything low with low quantity of fuel received, low heat content of the fuel, low sulfur content percentage and low ash content as well. Both the negative and positive things are low here.

Cluster 3

This cluster is basically characterized with the highest ash content in the fuel, in which the most impure energy is located with high heat content of the fuel, high sulfur content and high ash content percentage. So, I recommend US power generation to bring proper policy to curb the impurity in the fuel.

Cluster 4

This is the highest fuel received cluster with low ash content. Such a large volume of relatively purer fuel! Contrary to cluster 3, US power generation can use the incentive policy to further encourage collecting good fuel with less ash content.

Cluster 5

Cluster 5 has a higher heat value of the fuel together with moderately high sulfur content and ash content. My conclusion is cluster 5 has more fuel with more efficiency and releases more energy.

Cluster 6

This is the cluster with everything low. The difference between cluster 2 and cluster 6 is that this cluster has the plant with smaller identification numbers than cluster 2. Contrary to cluster 1, this cluster shows that the plants with lower identification number are associated with low volume of fuel.

Cluster 7

Cluster 7 can be named a larger volume of fuel than average with very high sulfur content in it. From the perspective of emission control and the longevity of automobiles, this is the cluster the US power generation must keep an eye on and formulate policy to reduce the sulfur content.