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://goo.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))

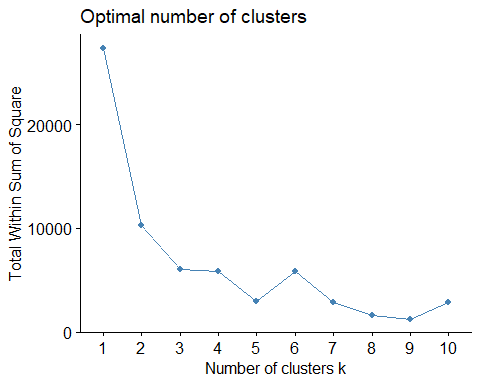
## rowid plant\_id\_eia report\_date   
## 0 0 0   
## contract\_type\_code contract\_type\_code\_label energy\_source\_code   
## 0 0 0   
## energy\_source\_code\_label fuel\_type\_code\_pudl fuel\_group\_code   
## 0 0 0   
## supplier\_name fuel\_received\_units fuel\_mmbtu\_per\_unit   
## 0 0 0   
## sulfur\_content\_pct ash\_content\_pct data\_maturity   
## 0 0 0   
## data\_maturity\_label   
## 0

colSums(is.na(data2\_Validation))

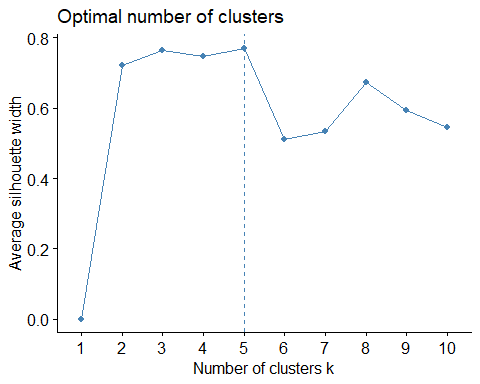
## rowid plant\_id\_eia report\_date   
## 0 0 0   
## contract\_type\_code contract\_type\_code\_label energy\_source\_code   
## 0 0 0   
## energy\_source\_code\_label fuel\_type\_code\_pudl fuel\_group\_code   
## 0 0 0   
## supplier\_name fuel\_received\_units fuel\_mmbtu\_per\_unit   
## 0 0 0   
## 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)

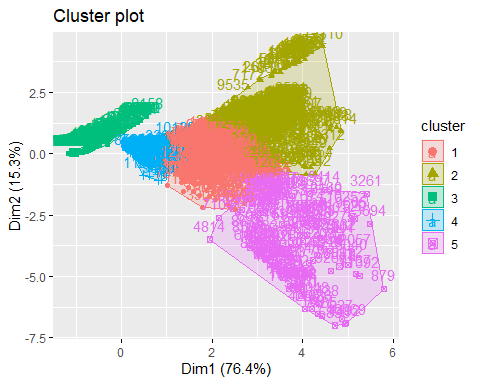
set.seed(1254)  
  
fviz\_nbclust(data3\_train\_norm[-1],kmeans,method='wss')



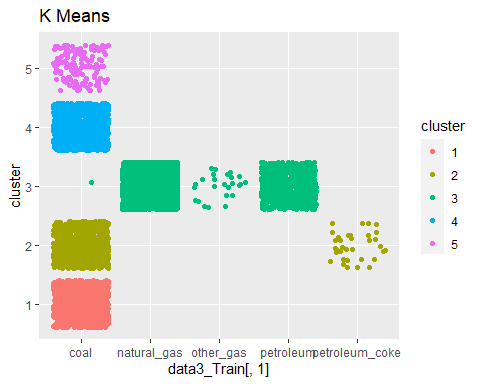
fviz\_nbclust(data3\_train\_norm[-1],kmeans,method='silhouette')



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\_mmbtu\_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\_mmbtu\_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