Mechine learning-Final project

Harish Kunaparaju

2022-11-30

#Clustring Algorithma & Visualization

```
library(ISLR)
library(pivottabler)
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(mice)

##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind

library(cluster)
```

#Importing dataset from the give data,Total number of Observations:608565 of 23 variables.

```
Project<-read.csv("fuelcost.csv")
str(Project)
```

```
## 'data.frame':    608565 obs. of  23 variables:
##  $ rowid                                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ plant_id_eia                         : int  3 3 3 7 7 7 7 8 8 8 ...
##  $ report_date                          : chr  "2008-01-01" "2008-01-01"
"2008-01-01" "2008-01-01" ...
##  $ contract_type_code                   : chr  "C" "C" "C" "C" ...
##  $ contract_expiration_date             : chr  "2008-04-01" "2008-04-01"
"" "2015-12-01" ...
##  $ energy_source_code                   : chr  "BIT" "BIT" "NG" "BIT" ..
.
##  $ fuel_type_code_pudl                  : chr  "coal" "coal" "gas" "coal
" ...
##  $ fuel_group_code                      : chr  "coal" "coal" "natural_ga
s" "coal" ...
##  $ mine_id_pudl                         : int  0 0 NA 1 2 3 NA 4 4 1 ...
##  $ supplier_name                        : chr  "interocean coal" "intero
cean coal" "bay gas pipeline" "alabama coal" ...
##  $ fuel_received_units                  : num  259412 52241 2783619 2539
7 764 ...
##  $ fuel_mmbtu_per_unit                  : num  23.1 22.8 1.04 24.61 24.4
5 ...
##  $ sulfur_content_pct                   : num  0.49 0.48 0 1.69 0.84 1.5
4 0 2.16 1.24 1.9 ...
##  $ ash_content_pct                      : num  5.4 5.7 0 14.7 15.5 14.6
0 15.4 11.9 15.4 ...
##  $ mercury_content_ppm                  : num  NA NA NA NA NA NA NA NA N
A NA ...
##  $ fuel_cost_per_mmbtu                  : num  2.13 2.12 8.63 2.78 3.38
...
##  $ primary_transportation_mode_code     : chr  "RV" "RV" "PL" "TR" ...
##  $ secondary_transportation_mode_code   : chr  "" "" "" "" ...
##  $ natural_gas_transport_code           : chr  "firm" "firm" "firm" "fir
m" ...
##  $ natural_gas_delivery_contract_type_code: chr  "" "" "" "" ...
##  $ moisture_content_pct                 : num  NA NA NA NA NA NA NA NA N
A NA ...
##  $ chlorine_content_ppm                 : num  NA NA NA NA NA NA NA NA N
A NA ...
##  $ data_maturity                        : chr  "final" "final" "final" "
final" ...
```

Removing Unwanted columns like Characters and id numbers from the given dataset.

#From the summary statistics it is observed that the fuel_mmbtu_units the maximum and minimum consumption are to be 11 and 0 respectively.And the other variable factor here is ash_content_pct where max and min values are 0 and 72 from the given data.

```
Fuelcost<- Project[,-c(1,2,3,4,5,8,9,15,17,18,19,20,21,22,23)]
head(Fuelcost)

##   energy_source_code fuel_type_code_pudl    supplier_name fuel_received_un
its
## 1                BIT                coal  interocean coal              259
412
## 2                BIT                coal  interocean coal               52
241
## 3                 NG                 gas bay gas pipeline             2783
619
## 4                BIT                coal     alabama coal               25
397
## 5                BIT                coal     d & e mining
764
## 6                BIT                coal     alabama coal
603
##   fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct fuel_cost_per_mmb
tu
## 1              23.100               0.49             5.4               2.1
35
## 2              22.800               0.48             5.7               2.1
15
## 3               1.039               0.00             0.0               8.6
31
## 4              24.610               1.69            14.7               2.7
76
## 5              24.446               0.84            15.5               3.3
81
## 6              24.577               1.54            14.6               2.1
99

summary(Fuelcost)

##  energy_source_code fuel_type_code_pudl supplier_name      fuel_received_u
nits
##  Length:608565      Length:608565       Length:608565      Min.   :
1
##  Class :character   Class :character    Class :character   1st Qu.:    370
0
##  Mode  :character   Mode  :character    Mode  :character   Median :   2156
5
##                                                            Mean   :  24296
7
##                                                            3rd Qu.:  10616
4
##                                                            Max.   :4815976
5
##
##  fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct  fuel_cost_per_mmb
```

```
tu
##  Min.   :   0.000    Min.   : 0.0000    Min.   : 0.000   Min.   :   -71.9
##  1st Qu.:   1.025    1st Qu.: 0.0000    1st Qu.: 0.000   1st Qu.:     2.3
##  Median :   1.061    Median : 0.0000    Median : 0.000   Median :     3.3
##  Mean   :   8.839    Mean   : 0.5145    Mean   : 3.606   Mean   :    14.2
##  3rd Qu.:  17.809    3rd Qu.: 0.4900    3rd Qu.: 5.800   3rd Qu.:     4.8
##  Max.   :1049.000    Max.   :11.0100    Max.   :72.200   Max.   :562572.2
##                                                          NA's   :200240
```

The majority of the dataset is retained when using impute to replace missing data with substitute values. I selected the MICE program for the impute process since it effectively replaces missing values in datasets by examining data from other columns and provides the best prediction.

```
fuel_impute<-mice(Fuelcost,m=5,maxit=10,meth='pmm',seed=500)

##
##  iter imp variable
##   1   1  fuel_cost_per_mmbtu
##   1   2  fuel_cost_per_mmbtu
##   1   3  fuel_cost_per_mmbtu
##   1   4  fuel_cost_per_mmbtu
##   1   5  fuel_cost_per_mmbtu
##   2   1  fuel_cost_per_mmbtu
##   2   2  fuel_cost_per_mmbtu
##   2   3  fuel_cost_per_mmbtu
##   2   4  fuel_cost_per_mmbtu
##   2   5  fuel_cost_per_mmbtu
##   3   1  fuel_cost_per_mmbtu
##   3   2  fuel_cost_per_mmbtu
##   3   3  fuel_cost_per_mmbtu
##   3   4  fuel_cost_per_mmbtu
##   3   5  fuel_cost_per_mmbtu
##   4   1  fuel_cost_per_mmbtu
##   4   2  fuel_cost_per_mmbtu
##   4   3  fuel_cost_per_mmbtu
##   4   4  fuel_cost_per_mmbtu
##   4   5  fuel_cost_per_mmbtu
##   5   1  fuel_cost_per_mmbtu
##   5   2  fuel_cost_per_mmbtu
##   5   3  fuel_cost_per_mmbtu
##   5   4  fuel_cost_per_mmbtu
##   5   5  fuel_cost_per_mmbtu
##   6   1  fuel_cost_per_mmbtu
##   6   2  fuel_cost_per_mmbtu
##   6   3  fuel_cost_per_mmbtu
##   6   4  fuel_cost_per_mmbtu
```

```
##    6    5  fuel_cost_per_mmbtu
##    7    1  fuel_cost_per_mmbtu
##    7    2  fuel_cost_per_mmbtu
##    7    3  fuel_cost_per_mmbtu
##    7    4  fuel_cost_per_mmbtu
##    7    5  fuel_cost_per_mmbtu
##    8    1  fuel_cost_per_mmbtu
##    8    2  fuel_cost_per_mmbtu
##    8    3  fuel_cost_per_mmbtu
##    8    4  fuel_cost_per_mmbtu
##    8    5  fuel_cost_per_mmbtu
##    9    1  fuel_cost_per_mmbtu
##    9    2  fuel_cost_per_mmbtu
##    9    3  fuel_cost_per_mmbtu
##    9    4  fuel_cost_per_mmbtu
##    9    5  fuel_cost_per_mmbtu
##   10    1  fuel_cost_per_mmbtu
##   10    2  fuel_cost_per_mmbtu
##   10    3  fuel_cost_per_mmbtu
##   10    4  fuel_cost_per_mmbtu
##   10    5  fuel_cost_per_mmbtu

## Warning: Number of logged events: 3

com_fuelimp<- complete(fuel_impute,1)
```

We randomly selected 2% of the data as a sample, storing 13000 observations in the sample data, using the seed 3333, a random 4-digit number,where doing the sampling with a precise and chosen data gives an accurate results and provides the correct set of findings in determining the clusters.We also want to set the seed so that we ensure reproducibility with this code:

```
set.seed(3333)
sampledata<-com_fuelimp[sample(nrow(com_fuelimp), size=13000), ]
```

set up data partition 75% of sampled data as the tarining set and reaminng 25% used as a test data.Here the data is divided into train and test where prediction is done with the help of test with the other selected data.

```
Train_index<-createDataPartition(sampledata$fuel_cost_per_mmbtu,p=.75,list=FA
LSE)
traning<-sampledata[Train_index,]
test<-sampledata[-Train_index,]
```

Normalize the data while removing unnecessary variables from the training data, (such as Energy source code, fuel type code, and supplier name), Because I am only accepting numbers here.

```
select_data<-traning[,-c(1,2,3)]

Nordata<-preProcess(select_data,method = c("center", "scale"))
Nor_Tdata<-predict(Nordata,select_data)
summary(Nor_Tdata)
```

```
##  fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct ash_content_pc
t
##  Min.   :-0.3352     Min.   :-0.8983     Min.   :-0.52274   Min.   :-0.551
0
##  1st Qu.:-0.3302     1st Qu.:-0.8028     1st Qu.:-0.52274   1st Qu.:-0.551
0
##  Median :-0.3059     Median :-0.7987     Median :-0.52274   Median :-0.551
0
##  Mean   : 0.0000     Mean   : 0.0000     Mean   : 0.00000   Mean   : 0.000
0
##  3rd Qu.:-0.1929     3rd Qu.: 0.9079     3rd Qu.:-0.02432   3rd Qu.: 0.362
3
##  Max.   :15.7791     Max.   : 2.1309     Max.   : 6.26581   Max.   : 9.031
2
##  fuel_cost_per_mmbtu
##  Min.   :-0.04790
##  1st Qu.:-0.03417
##  Median :-0.02859
##  Mean   : 0.00000
##  3rd Qu.:-0.02073
##  Max.   :83.91432
```

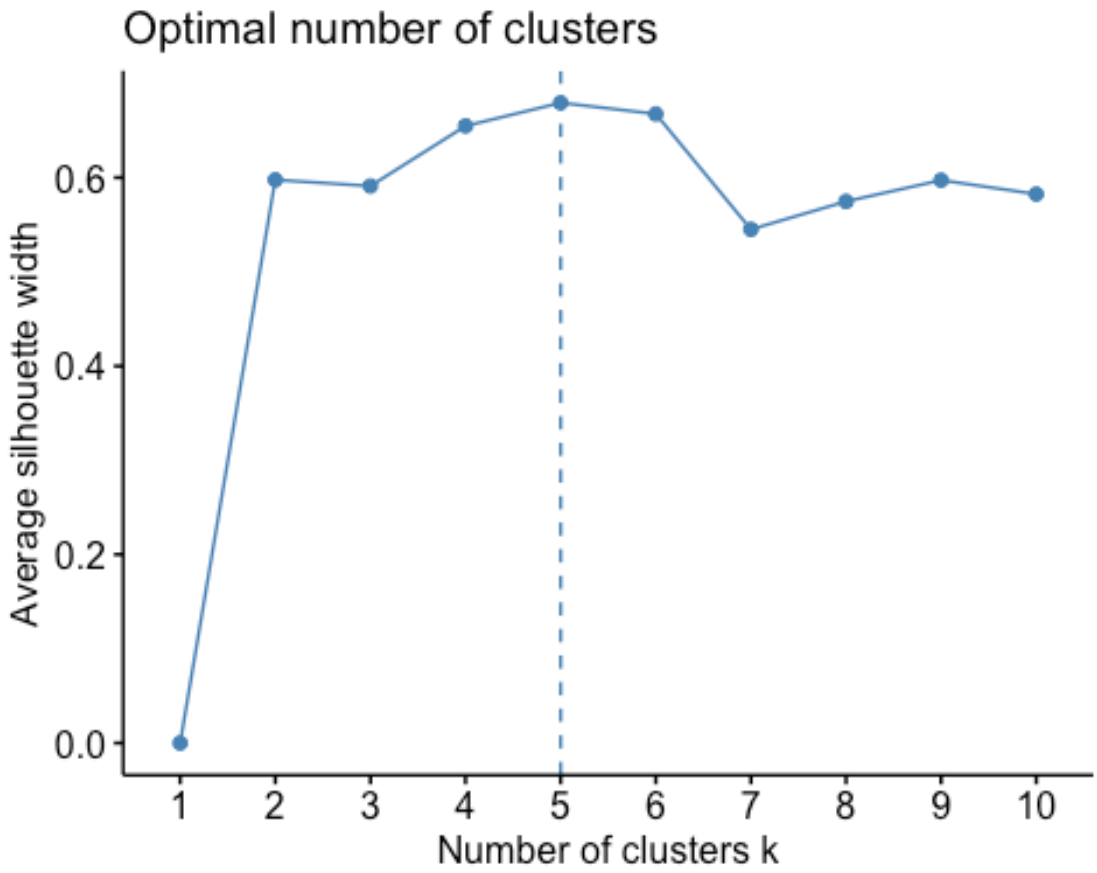
#Here The Silhouette Method is used to know how each member fits within its cluster by calculating its silhouette value.The silhouette value is a measure of how similar an observation is to its assigned cluster (cohesion) compared to the other clusters (separation). These values range from -1 (poor match within its assigned cluster) to +1 (perfect match within its assigned cluster).Silhoutte method represented with distnace to the cluster centroid insted of the average distance of all other data points in cluster In Business point of you silhoutte method can give

#K Mean Clustering- i used k mean clustering to generate groups with similar characteristics and used large data scale the number of groups is represented by k,and i used Silhouette method to get optimal numbers OF clusters 'K' The optimal number of clusters K=5.

```
library(factoextra) # Determining and visualizing the optimal number of clust
ers.
```

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

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



We will just scale the data, make 5 clusters (our optimal number), and set nstart to 25 for simplicity.The centers argument describes the number of clusters we want, while the nstart argument describes a starting point for the algorithm. (Here it was specified for precise reproducibility, different starting points typically have minimal impact on the results)

```
Fculter<-kmeans(Nor_Tdata,centers = 5,nstart = 25)
Fculter$centers

##   fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct ash_content_p
ct
## 1          -0.2914786           1.3892789          2.2348420       1.50161
83
## 2           3.7705209          -0.8043433         -0.5227415      -0.55095
42
## 3          -0.3351192          -0.8025946         -0.5227415      -0.55095
42
## 4          -0.1245781          -0.7303014         -0.4959499      -0.55095
42
## 5          -0.2636423           1.1782631          0.0754072       0.62506
```

```
44
##   fuel_cost_per_mmbtu
## 1        -0.031865311
## 2        -0.029885148
## 3        50.396153647
## 4        -0.003559909
## 5        -0.032975820
```

#Thus, silhouettes can be used to assess individual observations, or the average silhouette can be used to assess the choice of k. which gives k = 5 the optimal number of cluster that can be formed is to be 5 clusters.

Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

In this the silhouette analysis is used to choose an optimal value for n_clusters. Where we have found the optimal number of clusters formed are 5.

The silhouette plot shows that the n_clusters value of 3, 4 are a good pick for the given data due to the presence of clusters with below average silhouette scores and also due to wide fluctuations in the size of the silhouette plots. Silhouette analysis is more ambivalent in deciding between 3 and 4.

We can visualize these clusters using fviz_cluster, which shows the clusters (which are by default created using all columns of fuel costs using the first two principle components to define the X-Y coordinates of each observation.

```
fviz_cluster(Fculter,data = select_data)
```

#

```
f_cluster<- Fculter$cluster
fcluster<-cbind(traning[,-c(1,2,3)], f_cluster)
head(fcluster)
```

```
##        fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct
## 143555            11375658               1.019                0.00
## 191262                6358               1.003                0.00
## 494117                3574              16.050                2.16
## 73154                84039              22.780                0.40
## 43961                30424              23.251                2.69
## 146842               31608              24.732                1.47
##        ash_content_pct fuel_cost_per_mmbtu f_cluster
## 143555            0.00                6.048         2
## 191262            0.00                4.385         4
## 494117           35.70                5.448         1
## 73154            10.97                2.838         5
## 43961             8.00                1.826         1
## 146842           11.40                3.531         5
```

Here, I'm using aggregate data, which is easily helpful for statistical analysis, making it simple to locate important information for business analysis.

```
aggregate(fcluster,by=list(Fculter$cluster),FUN="mean")
```

```
##   Group.1 fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct
## 1       1        3.248068e+04           22.675201         2.76629744
## 2       2        3.048034e+06            1.009396         0.00000000
## 3       3        8.266667e+01            1.026667         0.00000000
## 4       4        1.563845e+05            1.740686         0.02687632
## 5       5        5.314581e+04           20.591056         0.60003888
##   ash_content_pct fuel_cost_per_mmbtu f_cluster
## 1       13.709280            2.692749         1
## 2        0.000000            3.030022         2
## 3        0.000000         8591.895333         3
## 4        0.000000            7.513895         4
## 5        7.854713            2.503600         5
```

Now we can start interpreting the cluster results:

Cluster 1: 1.It looks to be a higher fuel_mmbtu_per_unit and high with respect to ash_content_pct and good with sulfur_content_pct (2.76)approximately

Cluster 2: It represents least in sulfur_content_pct, ash_content_pct, and maintains above average value with fuel_cost_per_mmbtu.

Cluster 3 is dominant in the fuel_received_units, very highly influenced with "fuel_cost_per_mmbtu"

Cluster 4 is next in place with fuel_mmbtu_per_unit

Cluster 5 might be either the fuel_mmbtu_per_unit and fuel_received_units are optimum.

In order to better understand this, let's look at Clusters 1 and 5. As the fuel mmbtu per unit is used more, the fuel received unit will rise. I'd like to share a few reasons why this is happening. First, according to a recent report by Americangeoscience, there are three types of fossil fuels that are used more frequently in the USA: Natural gas (32%), oil (28%) and coal (17.8%).I want to talk about natural gas here. Electricity in the United States in 2019 consumes about 31% of all natural gas, and other businesses besides electricity also utilize it for operations. This is the key factor driving up natural gas use.When compared to other fuels, natural gas is less expensive and more readily available, which is why industries would prefer it.Another benefit of using natural gas is that it does not cause pollution. Compared to other fuels, natural gas is the most environmentally friendly since it produces more energy with less pollution.so moreover industries can save more money.

Add the coloumns names using Cbind

```
new_data<- cbind(fcluster, traning$energy_source_code, traning$fuel_type_code
_pudl, traning$supplier_name)
head(new_data)

##        fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct
## 143555            11375658               1.019               0.00
## 191262                6358               1.003               0.00
## 494117                3574              16.050               2.16
## 73154                84039              22.780               0.40
## 43961                30424              23.251               2.69
## 146842               31608              24.732               1.47
##        ash_content_pct fuel_cost_per_mmbtu f_cluster traning$energy_source
_code
## 143555            0.00               6.048         2
NG
## 191262            0.00               4.385         4
NG
## 494117           35.70               5.448         1
WC
## 73154            10.97               2.838         5
BIT
## 43961             8.00               1.826         1
BIT
## 146842           11.40               3.531         5
BIT
##        traning$fuel_type_code_pudl traning$supplier_name
## 143555                         gas           florida gas
## 191262                         gas           ameren cips
## 494117                        coal           enersystems
## 73154                         coal         mountain coal
## 43961                         coal         alliance coal
## 146842                        coal      nally & hamilton
```