Report for US power generation

Introduction

This project uses the Public Utility Data Liberation (PUDL hereafter) dataset to cluster some of the numerical variables. The project uses a specific table of the dataset, the monthly fuel contract information, purchases, and costs reported of the US that makes US energy data easier to access and use programmatically. The data originally contained 608,565 rows and 20 observations with several missing values.

Problem Statement

For this final project, I am requested to solve two problems for US power generation (data refers to ***Monthly fuel contract information, purchases, and costs reported in EIA-923 Schedule 2, Part A)***. The first problem for me is to select the best segmentation and describe the clusters. For the second problem, I need to help the US understand their power generation based on my segmentation.

Analysis and Discussion

Using R programming language, first, I set a random 4-digit number 3272 as the seed and randomly sampled 2 % of my data. Second, I removed a significant number of missing data. Third, I dropped the categorical variables and keeped only numerical variables. Fourth, I used 75% of the sampled data as the training set, and the rest as the test set. Fifth, I normalized each training and testing

dataset. Sixth, I used both WSS and Shiloutte methods to find the optimal K for K-means. I choose shiloutte over wss, as shiloutte method produces clustering very clearly. Last but not least, I use cluster.k1$centers to interpret my findings.

Executive Summary

After my final analysis, I got seven Clusters. Cluster 1 indicates surplus inflow of pure fuel. Cluster 2 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. Cluster 3 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. Cluster 4 is the highest fuel received cluster with low ash content. Cluster 5 has a higher heat value of the fuel together with moderately high sulfur content and ash content. Cluster 6 is a cluster with everything low. Cluster 7 can be named a larger volume of fuel than average with very high sulfur content in it.

Conclusions

To sum up, I recommend US power generation to bring proper policy to curb the impurity in the fuel for cluster 3. Contrary to cluster 3, I consider cluster 4 is the best segmentation, since it can help US power generation use the incentive policy to further encourage collecting good fuel with less ash content. As cluster 5 has more fuel with more efficiency and releases more energy, it can be treated as a second option. From the perspective of emission control and the longevity of automobiles, the US power generation must keep an eye on and formulate policy to reduce the sulfur content cluster 7.