**US Power Generation**

**Fall-2022 FML Final Project Niharika Kolliboyana**

**Executive Summary**

## 

## This document is intended to analyze the dataset provided by the Public Utility Data Preparation (PUDL) and use any of the Machine Learning models, here we’ve used K-Means clustering algorithm to segment the data and use the segmented data to give some insights or provide some recommendations about power generation in the US. The primary aim is to determinate the type of fuel that is both cost-efficient and environmentally friendly.

**Problem**

One of the major problems with generating power by burning fossil fuels is that it releases harmful gases, and other greenhouse gases into the air. Fossil fuels, which include coal, petroleum, oil, and gas, are by far the biggest cause of climate change, contributing more than 75% of all greenhouse gas emissions and almost 90% of all carbon dioxide emissions.

However, burning them causes climate change and produces pollutants that increase the risk of early mortality, heart attacks, stroke, respiratory problems, asthma, and absence from work and school. It has also been connected to Alzheimer's disease and autism spectrum disorder.

The primary objective of this project is to interpret the data and find assorted solutions to mitigate the impact and help US Power Generation in understanding the benefits of the usage of a particular fossil fuel that could be used for power generation.

**Technique**

To conduct the study and create the model for this project, we needed the historical data from PUDL. The dataset consists of the monthly fuel contract information, purchases, and costs reported in EIA-923 Schedule 2, Part A. Variable values having missing values have been imputed and categorical variables are converted to numerical. The dataset's attributes considered are plant\_id\_eia, fuel\_received\_units, fuel\_mmbtu\_per\_unit, sulfur\_content\_pct, ash\_content\_pct,mercury\_content\_ppm,fuel\_cost\_per\_mmbtu,fuel\_group\_code reported in EIA-923 Schedule 2, Part A for our analysis.

Later, considered only 2% of the data randomly. Then divided the dataset into training and validation sets by 83% (10000) and 17% (2000) respectively.

For clustering huge data that would take an impractically long time with similar approaches, K Means is the quickest partitional method. The K-Means clustering algorithm was performed on the training data set, for determining the number of clusters to take part in segmentation, WSS method and Silhouette methods were used, from which the optimal k value generated was 3. The silhouette score was maximum at **k=3**. The basic idea behind k-means is to define k clusters such that total**within-cluster variation (or error) is minimum.** As a result, 3 clusters have been divided based on their similarities.

**Conclusions**

## From the below figure it can be said that, although cluster 1 which has a fuel group as coal is very inexpensive whilst compared to the other two groups, it has contaminants like ash, Sulphur and mercury that cause greenhouse gases. Our focus is to determine the type of fuel which is both environmentally friendly and cost friendly. Well, in that case, cluster 2 which has a fuel group as natural gas has many approving attributes such as reasonable cost concerning the fuel units it is receiving and has practically zero impurities. From this, Natural Gas can be suggested to US Power Generation.

Table

Description automatically generated

**Reference**

## 1)[https://www.jmp.com/en\_us/statistics-knowledge-portal/what-is-multiple-regression/variable- selection](https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-multiple-regression/variable-%20%20%20%20%20%20selection).

2) <https://towardsdatascience.com/multiple-linear-regression-model-using-python-machine-learning-d00c78f1172a>

3) <https://enjoymachinelearning.com/blog/k-means-accuracy-python-silhouette/>

4) <https://stats.stackexchange.com/questions/23472/how-to-decide-on-the-correct-number-of-clusters>

5) <https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb>