**Project Report**

**Executive Summary:**

The power generation in the United States is significantly based on Coal, Natural Gas and Oil. Historically, the Coal based plants highly contributed to the electricity demands of the consumers as Coal is easily available at cheaper price and also the heat content is high generating large amount of power. These plants emit hazardous contents along with Green House Gases that contaminate the environment. However, these emissions can be marginally reduced by utilizing cleaner variants of Coal. In the recent times, the rise of Natural Gas and Oil power plants was seen in the U.S. due to the less polluting nature. Nevertheless, the Gas and Oil are highly expensive and the heat content is lower which makes it a better option for producing power under emission limits.

**Recommendations:**

As per United Nations, “Global addiction to fossil fuels must end and renewables revolution jumpstarted” - fossil fuels are better sources of energy but they are non-renewable. Over dependency on fossil fuels could cause an imbalance to the eco system in the long run. Alternatively, the advancements of technology have enabled mankind to invent renewable resources like wind energy, solar energy, hydro power etc., to diversify the energy demands of the world and conserve the fossil fuels.

**Introduction:**

In the United States, the major source for the power generation is from fossil fuels - coal, natural gas and petroleum. As per U.S. Energy Information Administration (EIA), in 2021 fossil fuels electricity generation was about 61%, nuclear energy was about 19% and the rest 20% was from renewable energy sources. The data related to power generation across the United States is collected, analyzed, and published by EIA. This data is used to formulate policies based on the economy, help the public understand energy conservation and protect the environment.

As part of the PUDL project, the fuel receipts and costs data for power plants is refined and made available for the users for novel analysis. The dataset for this project contains monthly data which is a total of 608,565 records and 24 variables over a period from 01-01-2008 to 12-01-2021 in a CSV file.

**Data Cleaning:**

The columns with more than 50% missing values have been dropped from the dataset to obtain meaningful insights from the data. A random sample of 2% data was extracted which resulted in 12,172 records.

Upon further review of the data the below columns were dropped:

* “data\_maturity” column had a constant value “final”
* “row\_id” column was an identifier
* “fuel\_type\_code\_pudl” had similar information as “fuel\_group\_code”

The remaining missing values in the dataset were imputed using the missForest package. missForest package utilizes the random forest technique to impute the missing data.

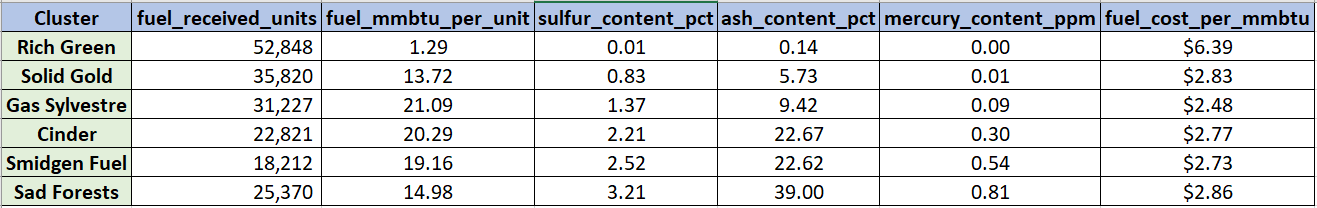
The fuel\_mmbtu\_per\_unit is correlated with sulfur\_content\_pct and ash\_content\_pct with correlation values of 0.72 and 0.63 respectively. Hence, having the sulfur\_content\_pct and ash\_content\_pct variables would be redundant, which were dropped from the dataset.

**Problem Statement:**

* Which power plants are generating power with less environmental impact?
* Which group of suppliers are providing the better-quality fuel resources at cheaper prices?
* Which power plants (coal, natural gas or petroleum) are on the raise and which are declining?
* What type of coal plants can be promoted for the reduction of CO2 emissions?
* What type of contract is better suited for economical fuel purchases?

**Analysis and Discussion:**

Hierarchical clustering was used for data analysis from which 6 clusters were obtained. Below is the analysis from the clusters:



**Rich Green** cluster has Natural Gas and Petroleum power plants mostly. The Sulphur (0.01), Ash (0.04) and Mercury (0.00) content is very low in these fuels causing less pollution due to which the environmental impact is minimal. These fuels are expensive as their demand in the market is higher but the supply is less. In addition, many of the fuel purchases are spot which could be another reason for high fuel costs.

**Solid Gold** cluster has BIT coal suppliers that provide better quality coal at an optimal cost of $2.91 per unit which produces less Ash content (10.24), Sulphur content (1.85) and Mercury content (0.01). Also, the Natural Gas suppliers in this cluster sell the fuel at cheaper prices of $3.15 per unit which is almost 50% less than that of Rich Green cluster ($6.2 per unit). Similar pattern can also be observed for Petroleum suppliers in this cluster where the purchase price is $3.16 per unit which is 67% lesser than the price of Rich Green cluster ($9 per unit). Hence, this can be considered as best cluster.

**Gas Sylvestre** cluster has a combination of Bituminous (BIT), Subbituminous (SUB), Lignite (LIG) and Waste Coal (WC) types of coals. These power plants generate more power as the heat content is highest (21.08) among all the clusters. SUB and LIG plants occupy 41% share and have a high emission factor of CO2 as 98 (Kilograms of CO2 Per Million Btu) leading to excessive air pollution. Furthermore, BIT and WC can be promoted for lesser CO2 emissions.

**Sad Forests** cluster has the highest emissions of Sulphur (3.2), Ash (39) and Mercury (0.81) though these plants are using BIT coal making this cluster the highest polluting plants. The reason for this could be the low-quality coal from the suppliers of BIT coal at the cheapest price of $2.74 per unit.

**Findings:**

* Natural Gas and Oil power plants are increasing in United States while the coal power plants are declining
* Natural Gas and Petroleum have less heat content and produce less energy which indicates that these plants generate less power. Hence, more fuel units are required to generate power
* For Coal, the cost of fuel is lowest but the heat content and emissions are high
* The Power plants’ management should go for long-term contracts where the suppliers will offer lower prices due to stability

**Conclusion:**

From the study of Power Generation in the United States, it is evident that all segments of electricity systems can affect the planet as most of the power is produced from fossil fuels. Upon research, it is known that renewable resources could potentially be alternatives for fossil fuels which are eco-friendly and significantly diversify the demand for electricity. Notable renewable resources are solar energy, wind energy, hydroelectric power etc., which are on the rise in the United States as per EIA.

**References:**

<https://www.eia.gov/electricity/annual/html/epa_a_03.html>

<https://news.un.org/en/story/2022/09/1126931>

https://www.eia.gov/energyexplained/electricity/electricity-in-the-us.php

**Appendix:**

**Model Selection**

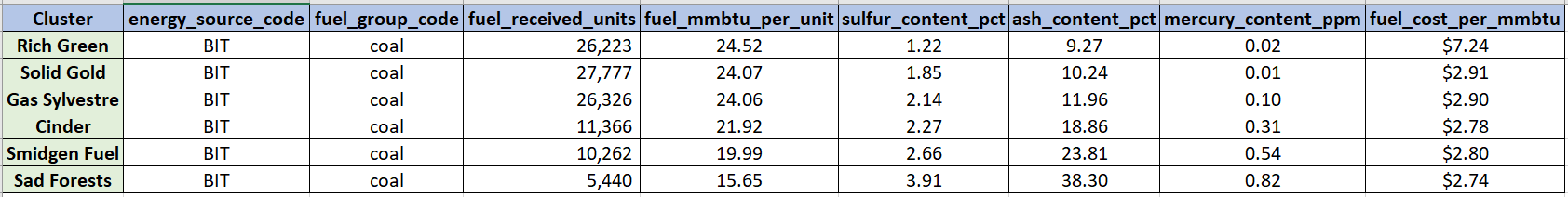
Hierarchical Clustering is chosen for over DB Scan and K-Means Clustering is due to following reasons

K-Means: The clusters formed in this type had high sum of squares between the datapoints, hence heterogeneity is more for these clusters.

DB-Scan: A large group of data was segmented under one group where the interpretation was not clear.

**Cluster Analysis**

BIT Coal distribution in clusters:



All the clusters except Cinder and Smidgen Fuel were mentioned in the above report. Below are the findings for these clusters.

**Cinder:**

This cluster has the power plants that are based on “Coal.” The primary transport of around 90% this cluster is Truck (TR). The power plants in this cluster are declining in the number of purchases in the recent years which indicates that these plants have reduced the power generation. Most of them are contract purchases

**Smidgen Fuel:**

The power plants in this cluster have lowest fuel received units. The suppliers in this cluster sell Waste coal at the cheapest price of $2.37 per unit.

