**Report on Clustering of Energy Usage in the US**

**Executive summary:**

The data analysis reveals that the power plants can be clustered into two main groups based on their primary fuel source - natural gas and coal. Natural Gas Cluster, or Cluster 1, consists of power plants with lower levels of ash and sulfur content, lower fuel costs, and predominantly natural gas as the primary fuel source. Coal Cluster, or Cluster 2, contains power plants with higher levels of ash and sulfur content, higher fuel costs, and predominantly coal as the primary fuel source. The analysis also shows that the clustering algorithm effectively grouped the data points based on their similarity, with a high silhouette coefficient indicating well-separated clusters. Moreover, based on the ANOVA analysis, plant\_id\_eia has a significant impact on predicting fuel cost per unit.

Overall, the findings suggest that natural gas is a relatively cleaner and cheaper fuel source compared to coal. Hence, policymakers and energy planners could use this analysis to encourage the adoption of natural gas in power generation. However, further analysis is required to determine the best set of variables to predict fuel cost per unit accurately.

**Problem Statement:**

Data Consumption matters both to the economy and environment of the country. It is important to monitor the usage of energy resources and their effects on the environment. This project aims to analyze energy usage in the United States and cluster the data into different groups to find useful insights and submit them to the country’s energy management board.

Objective:

the main goal of this project is to use the K-means clustering algorithm to identify each cluster’s importance and find the optimal cluster that is most suitable for the environment.

Description of the process done:

From the US energy data set of PUDL, I considered ten variables for analysis. They are plan id eia, contract type code label, energy source code, fuel group code, fuel received units, fuel mmbtu per unit, sulfur content pct, mercury content ppm, and fuel cost per mmtbu.

Then I took 2% of the entire data as a sample set to do analysis. There were significant missing values in Fuel cost, So I imputed those missing values with the median value of the data. As the fuel type is a categorical one it is converted into three different variables using the dummy variable method. The data set is then partitioned into a train set and test set with 75% and 25% of sample data. This train set is used to cluster the data. WSS and Silhouette methods are used to get an idea of which K to use. I choose the K value as 2 for clustering the data as this has produced clear clusters with good separation from other clusters also with this k value two clusters are formed with the same fuel type in each cluster which is also different from another cluster. This type of clustering helped to interpret the data better and to find useful insights. Hence, I chose K=2. I used the K means method to cluster the data as this has produced better results compared to other methods. Clusters formed from this method are used to interpret the data.

**Analysis and Findings:**

* The silhouette plot indicates that the clustering algorithm has effectively grouped the data points based on their similarity, with well-separated clusters and properly assigned data points.
* The table and plot show that there are two clusters in the data, one dominated by power plants that use natural gas as their primary fuel source (Cluster 1), and the other dominated by power plants that use coal as their primary fuel source (Cluster 2). Cluster 1 has lower fuel cost and sulfur and ash content, while Cluster 2 has higher fuel cost and sulfur and ash content.
* The plot also shows that the cost per mmtbu is higher in Cluster 1 than in Cluster 2.
* The ash content is an important factor that distinguishes the two clusters, with Cluster 1 having a lower range of ash content and Cluster 2 having a wider range of ash content.
* The dendrogram plot helps in determining an appropriate value for the eps parameter in the DBSCAN algorithm, which shows multiple peaks in the distribution, indicating multiple clusters in the data.
* The ANOVA analysis suggests that only the plant\_id\_eia variable has a significant impact on predicting fuel\_cost\_per\_mmbtu.

Overall, the analysis suggests that there are distinct differences between power plants that use natural gas and those that use coal as their primary fuel source, with natural gas plants having lower costs and less sulfur and ash content. These findings can be useful in informing decisions related to energy production and policy.

**Interpretation:**

* The clustering algorithm effectively grouped the power plants based on their similarity, with well-separated clusters and high silhouette coefficients.
* Cluster 1 is primarily composed of power plants that use natural gas as their primary fuel source, with lower levels of ash and sulfur content and lower fuel costs. Therefore, a possible name for Cluster 1 could be “Natural Gas Cluster”.
* Cluster 2 is primarily composed of power plants that use coal as their primary fuel source, with higher levels of ash and sulfur content and higher fuel costs. Therefore, a possible name for Cluster 2 could be “Coal Cluster”.
* The ANOVA analysis showed that only the plant\_id\_eia variable had a p-value less than 0.05, indicating its statistical significance in predicting fuel\_cost\_per\_mmbtu.
* The distribution plot of the data points can help in determining an appropriate value for the eps parameter in the DBSCAN algorithm, which shows multiple peaks but does not necessarily correspond to the number of clusters in the data. Further analysis is needed to identify the best variables to predict fuel\_cost\_per\_mmbtu.
* The table shows various characteristics of each power plant, such as the amount of fuel used per unit, sulfur and ash content, and fuel cost per unit of energy, which can be used to compare and contrast the power plants.

**Conclusion**

The clustering algorithm effectively grouped the data points based on their similarity, with well-separated clusters and properly assigned data points.

Cluster 1 is primarily composed of power plants that use natural gas as their primary fuel source, with lower levels of ash and sulfur content and lower fuel costs. Therefore, a possible name for Cluster 1 could be “Natural Gas Cluster”.

Cluster 2 is primarily composed of power plants that use coal as their primary fuel source, with higher levels of ash and sulfur content and higher fuel costs. Therefore, a possible name for Cluster 2 could be “Coal Cluster”.

The ANOVA analysis showed that only the plant\_id\_eia variable had a significant impact on deciding fuel\_cost\_per\_mmbtu, although the contract\_type\_code\_label variable was close to being statistically significant.

The eps parameter in the DBSCAN algorithm can be determined by using the knee method or elbow method to identify a threshold distance that separates the clusters from the noise.

Further analysis may be required to determine the best variables to predict fuel\_cost\_per\_mmbtu.