**Executive Summary:**

In the US, around 1600 electricity utility companies generate and distribute electricity [(1)](#one). To understand the competition for entering the power generation market, the report's outcome helps an organization determine which fuel sources are efficient in terms of cost and sustain in the market for a more extended period. With the help of the current dataset, various methods used in Machine Learning have been used with the R studio to achieve this goal.

Power generation has been mainly dominated by coal and Natural gas over the years. Various types of coal and natural gas are received from different suppliers all over the US. It is essential to pick the right fuel source from suitable suppliers to maintain the optimal cost for electricity production.

**Introduction:**

The current dataset is PUDL data extracted from massive datasets published by the US government. The dataset contains various variables that show fuel receipts costs from 2008-2021.

* + 1. The following variables have been selected that help understands power generation in the US. The variables are as follows:
       1. fuel\_group\_code(Type: Char): Types of fuel used in power generation.
       2. fuel\_received\_units (Type: Num): Number of fuel units received.
       3. fuel\_mmbtu\_per\_unit (Type: Num): Amount of heat produced per unit.
       4. sulfur\_content\_pct (Type: Num): Sulfur generation of fuels burnt.
       5. ash\_content\_pct (Type: Num): Ash generation of fuels burnt.
       6. fuel\_cost\_per\_mmbtu (Type: Num): Cost of heat generation per unit.

**Problem Statement:**

1. What cluster has been formed based on the algorithm used?
2. Which cluster shows the highest amount of heat produced?
3. What clusters show minimal and high fuel costs?
4. Which clusters have the highest ash and sulfur output?

**Analysis: [set.seed(9596)]**

1. **What cluster has been formed based on the algorithm used?**

The k-means clustering algorithm was used to understand the variables and their behavior. According to the clustering algorithm, there are 6 clusters.

1. **Which cluster shows the minimal and highest amount of heat produced?**

Clusters 1, 3, and 4 have the highest heat produced. Even though Clusters are mainly dominated by coal, increased heat production is due to petroleum coke.

Clusters 2,5 and 6 have a minimal amount of heat produced. Natural\_gas dominate all these clusters.

1. **What clusters show minimal and high fuel costs?**

Clusters 1,3 and 4 have a very minimal cost for the high heat produced.

Cluster 5 has a minimal cost for the heat produced and has minimal heat built.

Cluster 6 has only 3 data points with the highest cost for the very minimal output of Heat. These clusters can be called outliers as their cost to heat output is very high.

1. **Which clusters have the highest ash and sulfur output?**

Both clusters 1,3, and 4 have high sulfur and ash outputs. This is because these groups have fuel sources such as coal and petroleum coke.

**Conclusion:**

This report is aimed to understand what factors can be used to enter the power generation market for a new organization.

The following are the assumptions based on the clusters formed:-

* Coal has been a widely used fuel source over the years as the groups are dominated mainly by coal.
* Coal produces the highest output of heat with minimal cost, but there’s additional output, i.e., sulfur and ash content, in addition to that. This could result in the organization spending additional cost to process and clean the final production for cleaner power generation.
* Natural gas is also a significant contributor to fuel sources over the years. There is minimal heat output for the minimal cost incurred.
* The clustering algorithm shows that there are various types of natural gas that are used in the production of heat. Clusters 2 and 6 show that high costs are incurred for less heat produced. This is because the types of natural gas used could be from various sources, which adds to the price.
* Clusters 2, 5, and 6 show a median value of 0 for sulfur and ash, which signifies that no additional costs are incurred while producing heat using Natural gas.

In conclusion, Natural gas, as per cluster 5, is an optimal recommendation for new organizations in the power generation market. In addition, specific types of natural gas that show the characteristics of cluster 4 concerning cost and heat output are more optimal than those compared to other fuel sources.

References:

1. [https://www.statista.com/topics/2597/electric-utilities/ - topicOverview](https://www.statista.com/topics/2597/electric-utilities/#topicOverview)