**Executive Summary:**

In the United States, around 1600 power utility firms generate and distribute electricity. To fully understand the competition for joining the power generating industry, the report's findings assist a business in determining which fuel sources are cost-effective and will remain in the market for a longer length of time. To achieve this purpose, with the help of the current dataset, multiple machine learning algorithms have been applied with the R studio.

Over the years, coal and natural gas have dominated power generation. Various types of coal and natural gas are acquired from various providers around the United States. To maintain electricity production costs as low as possible, it is critical to select the best fuel source from the best providers.

**Introduction:**

The present dataset is PUDL data taken from huge databases made available by the US government. The dataset includes several variables that represent the costs of fuel receipts from 2008 to the present.

1. The following factors have been chosen for better understanding of electricity generation in the United States. The following are the variables:

* fuel\_group\_code(Type: Char): Types of fuel used in power generation.
* fuel\_received\_units (Type: Num): Number of fuel units received.
* fuel\_mmbtu\_per\_unit (Type: Num): Amount of heat produced per unit.
* sulfur\_content\_pct (Type: Num): Sulfur generation of fuels burnt.
* ash\_content\_pct (Type: Num): Ash generation of fuels burnt.
* 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 technique was used to analyze the variables and their behavior. There are six clusters based on the clustering algorithm.

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

Clusters 2, 4, and 6 have the highest heat produced. Even though Clusters are mainly dominated by coal, increased heat production is due to petroleum coke.

Clusters 1,3 and 5 have a minimal amount of heat produced. Natural\_gas dominate all these clusters.

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

Clusters 2,4 and 6 have a very minimal cost for the high heat produced.

Cluster 3 has a minimal cost for the heat produced and has minimal heat built.

Cluster 1 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 2 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 popular fuel source for many years since 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 1 and 5 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 1, 3, and 5 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, according to cluster 3, is an optimal recommendation for new firms in the power production sector. Furthermore, some types of natural gas that exhibit cluster 6 features in terms of cost and heat output are more ideal than other fuel sources.

**References:**

[https://www.statista.com/topics/2597/electric-utilities/ - topicOverview](https://www.statista.com/topics/2597/electric-utilities/%20-%20topicOverview)