# Report on Clustering of Energy Usage in the US

**Executive summary**:

Data from PUDL is used to do analysis on energy usage in the United States. The k-Means clustering method is used for analysing the data. From the results of clustering, it is found that coal as a fuel has produced the maximum amount of heat and that too at the least cost per average heat produced compared to gas and oil. At the same time, coal is the fuel type with the maximum amount of sulphur and ash content which when burnt is very harmful to the environment and to all living organisms. Gas has zero contents of sulphur and ash and it is also a fuel with good availability. From the results, it is evident that the energy management board has to choose coal as fuel in order to produce a large amount of heat economically. If the aim of the board is to produce heat and not cause any harm to nature and life on earth then gas is the best fuel to use.

**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. The aim of this project is to analyse energy usage in the United States and to cluster the data into different groups in order to find useful insights and submit them to the country's energy management board.

**Description of the process done:**

From the US energy data set of PUDL, I considered six variables for analysis. They are Fuel type, Fuel received, Heat produced, Cost per heat produced, Ash, and Sulphur contents. 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 3 for clustering the data as this has produced clear clusters with good separation from other clusters also with this k value three 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 choose K=3.

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:**

Let us understand the attributes of each cluster. Cluster 1 has fuel-type gas, Cluster 2 has fuel-type coal, and Cluster 3 has fuel-type oil.

The following are the findings from the analysis:

- Maximum heat is produced in Cluster 2, whereas Clusters 1 and 3 have produced almost the same level of heat.
- Average cost incurred to produce the average amount of heat is least in Cluster 2 and highest in Cluster 3.
- Sulphur and Ash contents are very high in Cluster 2. Cluster 1 has zero Sulphur and Ash content.
- Cluster 1 has received the highest number of fuel units compared to other clusters. Implying that availability of gas is better compared to other fuels.

**Interpretation:**

From the above findings, it is clear that Cluster 2 which has Coal as fuel has produced a high amount of heat and that too at the least cost. But cluster 2 also has high contents of sulphur and ash. This means that when the fuel in Cluster 2 that is coal is used it releases harmful gases. These gases are the reasons for environmental pollution and are also hazardous to life on earth. Based on these attributes this cluster is named an Efficient Cluster.

Cluster 1 which has Gas as fuel has produced less heat compared to Cluster 1. Most importantly this cluster has zero sulphur and ash contents in it. This implies that cluster 2 is environment-friendly and does no harm to life on earth. Even the cost of this fuel is not extremely high compared to coal and the availability of gas is also very high. This cluster is named Sustainable Cluster.

Cluster 3 is named Hopeless Cluster as this cluster has oil as fuel which produces less heat and the cost of fuel is also very high compared to other types of fuel. It also has sulphur content. Hence, this Cluster is neither economical to use nor environment-friendly.

**Conclusion:**

It can be concluded that if the goal of energy management is to produce a high amount of heat at the least cost, then Cluster 2 is the best option to choose. If energy management aims to produce heat in such a way that it is eco-friendly then Cluster 1 is most suitable. Cluster 3 is not suggested to use unless there is a shortage of coal and gas as it is very expensive.

The random four-digit code used is 5555.

**Appendix:**

https://data.catalyst.coop/pudl/fuel_receipts_costs_eia923