

# **Fundamentals of Machine Learning**

## **Final Project**

**MIS64060-001**

**Professor:** Dr. Murali Shanker

### **Exclusive summary**

Fossil Fuels such oil, natural gas, and coal are used to generate energy which release unburnt particles into the environment. The particles result in air pollution and cause respiratory diseases such as respiratory illness, lung damage, ozone (smog) effect, reduces the ability of blood to bring oxygen to the blood cells and tissues, liver, and kidney, etc., contributing to 7 million premature deaths worldwide every year. Also, there is a rapid increase in global temperature over past few years primarily due to the burning of fossil fuels So, the need for Identification of a fuel which is safe for the environment and humans has increased.

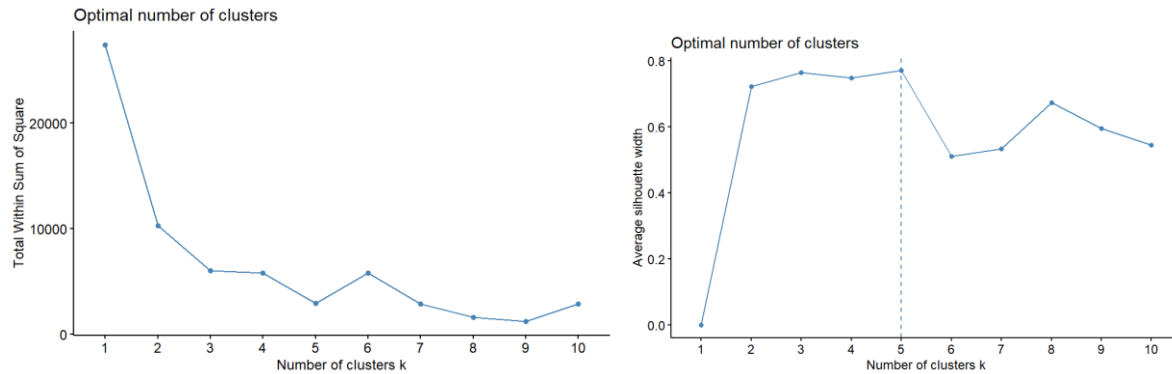
With the help of k-means clustering, and methods such as elbow method and Silhouette Methods natural gas and petroleum are identified as a group of fuels which omit low sulfur, ash, and heat. Since natural gas and petroleum are the two kinds of fuels which are less pollutant its better, we use these two fuels to generate power.

### **Problem Statement**

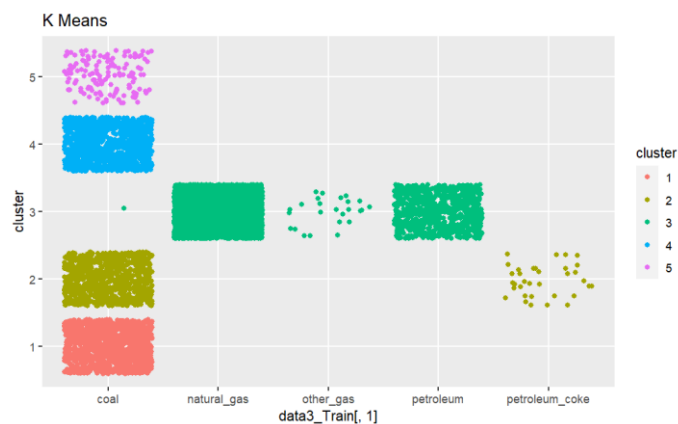
*The goal is to identify a fuel group which is safe for environment and public health.* Based on the data published by US government agencies

### **Technique**

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. the objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset A cluster refers to a collection of data points aggregated together because of certain similarities. When evaluating purchase behavior there were many ways that K means clustering was applied. Before running K-means, it is needed to determine the value of K. K stands for the number of clusters into which the data needed to be grouped. elbow method and silhouette method are most widely used for determining the value of K.



After running Elbow and Silhouette method, it can be observed that the value of K is equal which  $K=5$ . Hence,  $k=5$  is chosen to run k-means algorithm.



After running the k-means algorithm, from the above plot it is observed that

- Cluster 1: has coal with very low sulfur (1.10) and little ash (11.18) but high fuel MMBtu (24.10)
- Cluster 2: has coal and petroleum we have sulfur (3.20) ash (9.66) and fuel MMBtu (23.88)
- Cluster 3: has natural gas other gas and petroleum with very low sulfur ash and heat i.e., sulfur (0.02) Ash (0.00) fuel MMBtu (1.74)
- Cluster 4: has coal with sulfur (0.29) ash (5.30) and average heat (17.62)
- Cluster 5: has coal with sulfur (1.39) high ash (39.46) and average heat (13.69).

## Conclusion

- After analyzing the data, it can be concluded that cluster 3 produces less sulfur ash and heat. The type of fuels grouped in cluster 3 are natural gas, other gas, and petroleum.
- As sulfur, ash and heat produced by the fuel is harmful for the environment the lesser the substances are emitted the safer the environment and public health.

## Appendix

Data

[https://www.google.com/url?q=https://data.catalyst.coop/pudl/fuel\\_receipts\\_costs\\_eia923&sa=D&source=docs&ust=1670894432560193&usg=AOvVaw17hMbA90DYmJukRqxT8BaP](https://www.google.com/url?q=https://data.catalyst.coop/pudl/fuel_receipts_costs_eia923&sa=D&source=docs&ust=1670894432560193&usg=AOvVaw17hMbA90DYmJukRqxT8BaP)

Code

<https://github.com/k-joshna/FML-FINAL-PROJECT>

Katta joshna

[Jkatta1@kent.edu](mailto:Jkatta1@kent.edu)

811236826