FUNDAMENTALS OF MACHINE LEARNING
FINAL PROJECT REPORT

- Shiva Chaitanya Goud Gadila

- **EXECUTIVE SUMMARY**:

One of the outcomes of a review of the fuel contract data, purchases, and costs included in EIA-923 Timetable 2, Section A is this report. According to the report, petrochemicals specifically refer to sulfur, debris, and mercury in the fuel industry. In order to advance the power age in the US, it is necessary to investigate environmentally safe elective options and bring in less ignitable fills, according to the analysis.

- **INTRODUCTION:**

The examination used a dataset connected with power age in the US, which was obtained from the "Public Utility Information Freedom (PUDL) project."It has 608,565 rows and 20 variables, but only seven were chosen for the analysis. Fuel Group Code, Fuel Received Units, Fuel MMBtu per Unit, Sulphur Content, Ash Content, Mercury Content, and Fuel Cost Per MMbtu are the selected variables.

- **GOAL**

Utilizing clustering analysis more effectively and extracting useful information from the dataset are the primary objectives of this project. By drawing meaningful conclusions from the data's significant characteristics and patterns, this can be accomplished. By examining the data from a variety of perspectives and employing a variety of strategies to extract as much information as possible, the ultimate objective is to enhance the clustering results. Thusly, we can acquire a more profound comprehension of the hidden cycles and factors that impact the information, and utilize this information to further develop navigation and key preparation in different fields like energy age and utilization, asset the board, and natural maintainability.

- **PROCESS**

The data analysis process involves various steps to extract meaningful insights from the data. In this particular case, several actions were taken to address the issues based on the dataset.
a. Choosing the necessary variables
b. Imputing the variables with NA values
c. Sampling and partitioning the data

d. Data normalization
e Using the silhouette method, finding the optimal "K"
f. Clustering is done with the K-means algorithm.

K-means:- It is a clustering algorithm that partitions data points into K clusters by minimizing the variance within each cluster. It is an unsupervised learning technique that is simple, flexible, efficient, and suitable for large datasets. The algorithm iteratively updates the means of each cluster and assigns data points to the nearest cluster based on their distances. The process continues until convergence is achieved or a maximum number of iterations is reached.

KNN:- K-closest neighbours (KNN) is a straightforward and usually utilized AI calculation for characterization and relapse. It works by predicting the label or value based on the most common or average value among the k data points that are closest to a given query point

PROBLEM STATEMENT:

1. Data analysis overall
2. Findings on how the observed ash content would affect the US power generation system given that it is relatively higher than the sulphur and mercury
3. The best models and approaches used to derive the data.

According to the data, fuel composition, particularly the presence of sulphur, ash, and mercury, significantly affects fuel consumption. Coal, natural gas, and petroleum were the three groups into which the different fuel types were divided, and they were all spoken on the same day.The composition of these fuels was found to be relatively consistent.

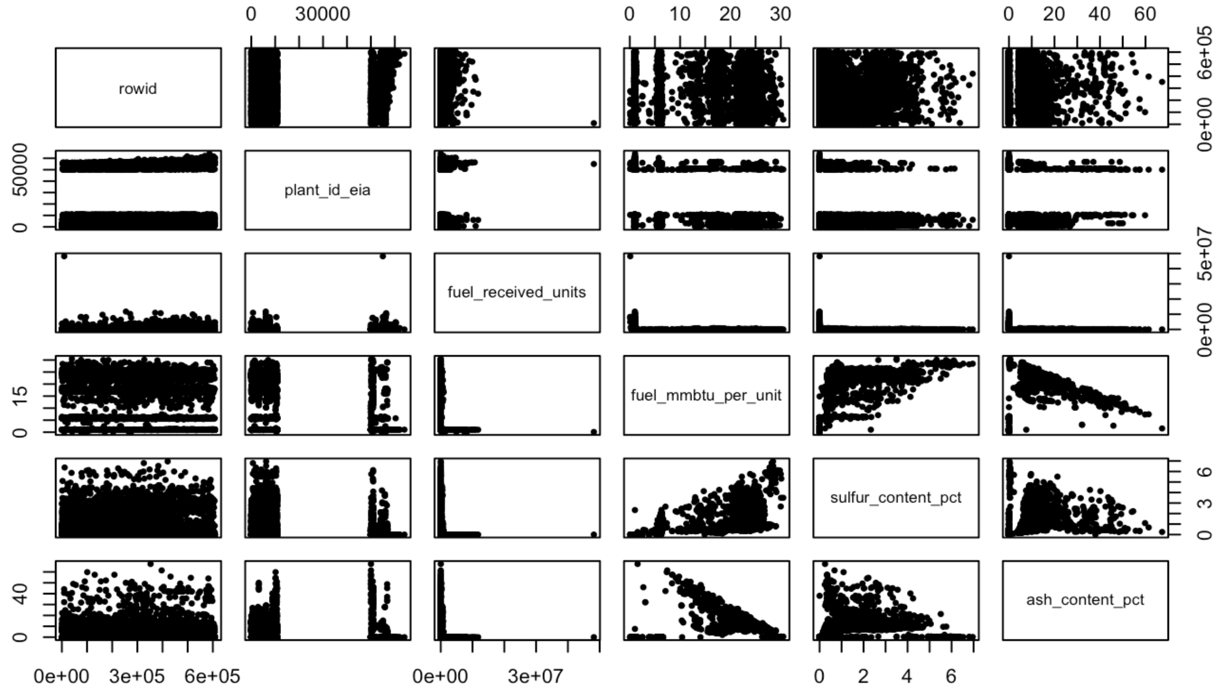
- The analysis of fuel receipts shows that the MMBtu per unit and fuel MMBtu cost have an inverse relationship.
- Out of 1001 rows, 562 are of coal which consists only of Sulphur and Ash, and has an average price of $2.5 per MMBtu and a fuel value of 24mmbtu per unit.
- There are 325 rows of natural gas that consist only of Sulphur with an average order of 1MMBtu and an average price of $7.5 per MMBtu. Natural gas has two modes of transportation- firm and interruptible, depending on the transfer location.
- Petroleum has 116 rows consisting only of Sulphur petrochemicals, with a fuel value of 4 MMBtu per unit and an average price of $18 per MMBtu.

It is Seen that in Coal the debris content is the most noteworthy with an avg of 15%. The debris content in fuel is ashfall is a worry which can cause blackouts or closures by obstructing generator air admissions and off-site power assets. Pulverized coal ash affects the blast furnace's heat balance by decreasing the carbon content and calorific value of the coal injected and requiring additional fluxes to remove this material as slag. This could result in an increase in fuel rate and a decrease in productivity. This infers that more coal will be bought for little use as it is yielding less efficiency.

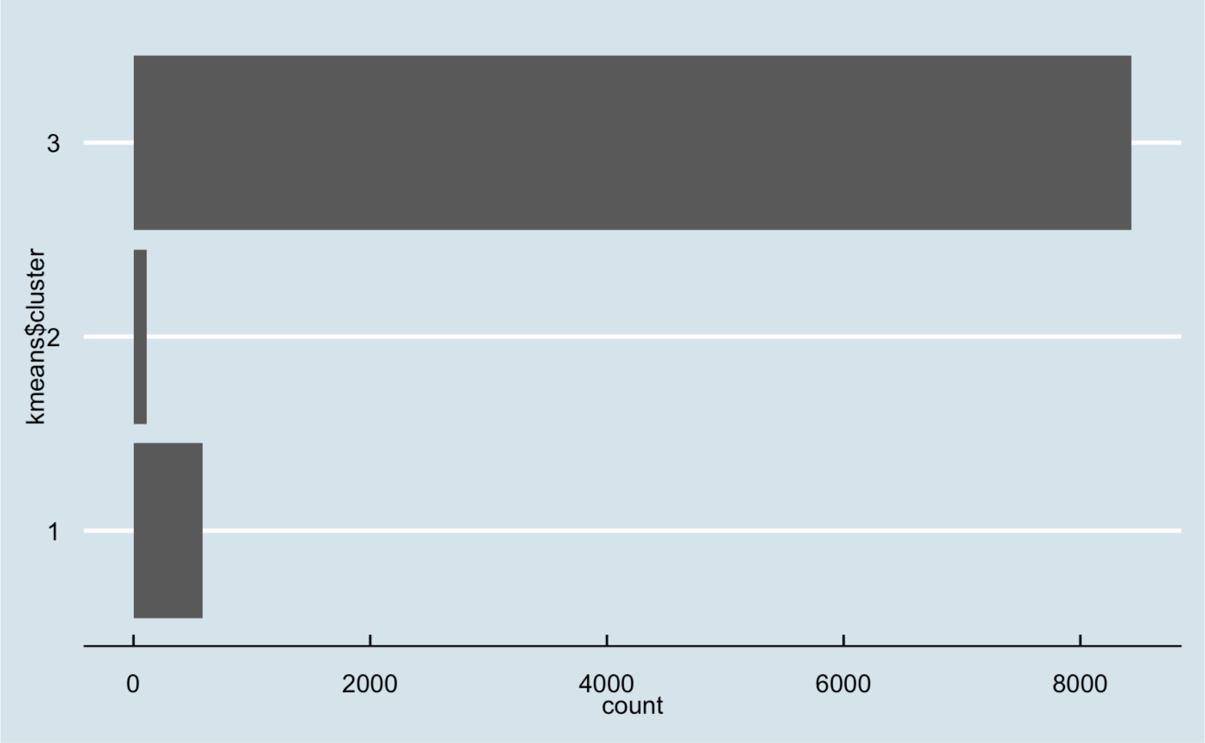
When coal and natural gas are burned, mercury is naturally released into the air and into the atmosphere. Elevated degrees of Sulfur are normal in all fills as it is expected in motor oils that forestalled the development of Sulfur Oxide covered leaves behind a defensive layer to decrease the degree of destructive wear

- **ANALYSIS AND DISCUSSION:**

The fluctuating ratio of fuel imports to consumption could be the result of shifting fuel demand or prices. Ash content affects combustion by influencing the formation of fuel clusters. I was able to fully comprehend the patterns of fuel consumption by utilizing z-score normalization and theory-based analysis to comprehend how fuel use affects the environment as well as imports and exports.

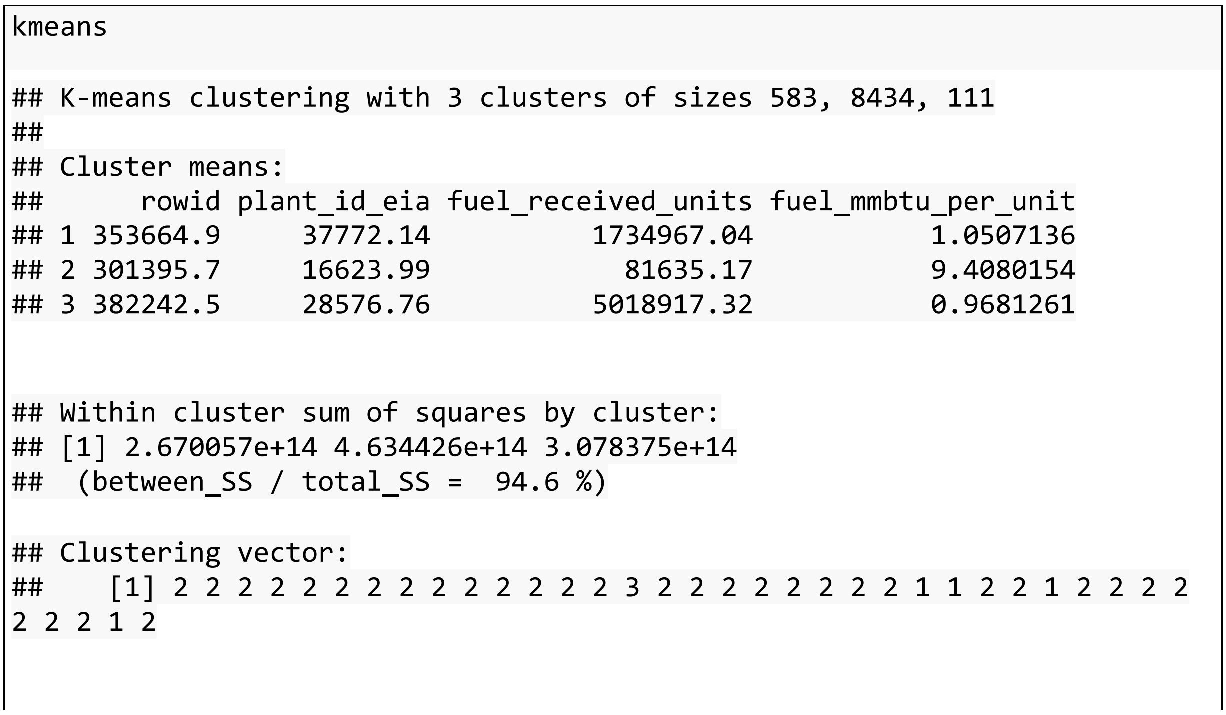


I analyzed the fuel data using the ash content per one million British Thermal Units. High ash content indicates less volatile matter. Ash content had the highest numerical value among the variables analyzed, while sulphur and mercury had the lowest. The data was divided into three groups based on the variation in ash content values.



The Ash(8234), Sulphur(854) & Mercury(1234) composition.

The Clustering of the data through K means strategy



```
kmeans

## K-means clustering with 3 clusters of sizes 583, 8434, 111
##
## Cluster means:
##      rowid plant_id_eia fuel_received_units fuel_mmbtu_per_unit
## 1 353664.9     37772.14          1734967.04           1.0507136
## 2 301395.7     16623.99            81635.17           9.4080154
## 3 382242.5     28576.76          5018917.32           0.9681261


## Within cluster sum of squares by cluster:
## [1] 2.670057e+14 4.634426e+14 3.078375e+14
##  (between_SS / total_SS =  94.6 %)

## Clustering vector:
##    [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 1 1 2 2 1 2 2 2 2
2 2 2 1 2
```

```
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.wi
thinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

#Best Segmentation of K

```
#Identify numeric columns in the train data
numValues <- sapply(train, is.numeric)

# Step 2: Perform k-means clustering on the numeric train data, with 3 clu
sters
kmeans <- kmeans(train[, numValues], centers = 3)

kmeans

##  K-means  clustering  with  3  clusters  of  sizes  583,  8434,  111
##
##                              Cluster                            means:
##            rowid  plant_id_eia  fuel_received_units  fuel_mmbtu_per_unit
## 1 353664.9      37772.14             1734967.04              1.0507136
## 2 301395.7      16623.99               81635.17              9.4080154
## 3 382242.5      28576.76             5018917.32              0.9681261
##                              sulfur_content_pct          ash_content_pct
##   1                           0.002521441                  0.02504288
##   2                           0.558086317                  3.79857482
##   3                           0.000000000                  0.00000000
##
##                            Clustering                           vector:
##    [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 1 1 2 2 1 2 2 2 2 2
2 2 1 2
 ## Within cluster sum of squares by cluster:
## [1] 2.670057e+14 4.634426e+14 3.078375e+14
##  (between_SS / total_SS =  94.6 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.wi
thinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Conclusion:

For US Public Utility Districts (PUDL), it is possible to manage power generation in a way that is both economically and environmentally sustainable. By optimizing the use of natural gas or petroleum in place of coal, this can be accomplished.

Utilizing flammable gas or petrol produces less ozone depleting substance discharges contrasted with coal, along these lines advancing natural preservation.

Coal also produces ash when burned, which can reduce the efficiency of power generation plants, leading to increased fuel usage.

By minimizing the use of coal and maximizing the use of natural gas or petroleum, the productivity of power generation plants can be enhanced, resulting in lower fuel consumption and reduced costs for US PUDL.

Shifting to cleaner and more efficient fuels can improve the overall sustainability of power generation while maintaining cost-effectiveness.

Thus, optimizing fuel consumption can be an effective way to promote environmental conservation and economic sustainability in power generation.