# Fundamentals of machine learning final project

## Executive summary:

Fuel is one of the most used resources that is available. We know that burning fossil fuels can cause damage to the environment. This project mainly focused on the heat content and the ash content present in the fuel that we use every day for commuting. We used the k-means clustering algorithm to form the clusters and chose which cluster is best for our purposes. To determine the optimal number of clusters we used silhouette and elbow method and generated the clusters. From this project we can find out the cluster that is environmentally friendly. Cluster 4 is considered environmentally friendly with lower ash content. Cluster 4 is considered the best for fuel efficiency with the max heat content.

**Objective:** the main goal of this project is to use the K-means clustering algorithm and identify the importance of each cluster and find the optimal cluster that is most suitable for the environment.

**Problem statement:** the fuel that we use everyday contains a ton of harmful compounds. One of the main compounds is “ash content “.

What is the percentage of ash content and heat content present in the fuel types that are received monthly?

How does the ash content and heat content present in the fuels play a major role in deciding the type of fuel that is best.

## Process:

I downloaded the data of fuel costs reported from the website. And then selected the required attributes for the project. ( ash_content_pct, fuel_mmbtu_per_unit, fuel_received_units). This data had some missing values, so I tried to impute the values by calculating the median. I have used 75 % of the data for the training and the other 25% testing. Here I have used center and scale function for normalization. To find out the optimal number of clusters, I have used both elbow method and silhouette method. I got the optimal number of clusters as 4. Finally used group by function to make it easy for analysis.

## Deciding k value:

As mentioned above I have used both silhouette and elbow method to get an idea how to get the maximum information out of it. I formed the clusters using k value as 4 and it gave good results. The clusters that I got made a lot of sense and analysis of each cluster as to what each cluster is trying to say about the data. So, I decided on using ka value as 4.

The constant value or set seed value used for this project is 0088

## Interpretation:

From the data we have formed 4 clusters

Cluster 1- the avg units of fuel is 139642 .11 with the ash content of 1.8e-05 and heat content of 1.7.

Cluster 2-- the avg units of fuel is 36131 with the ash content of 3.7e+01 and heat content of 14.8.

Cluster 3-- the avg units of fuel is 2873502 with the ash content of 0 and heat content of 1.0.

Cluster 4- - the avg units of fuel is 47811 with the ash content of 8.3+00 and heat content of 21.5.

- If we must choose the best cluster considering the environment, we will choose cluster because it has 0 ash content.
- If we were to choose the one with most fuel efficiency, we would compromise on the environment and end up choosing cluster 4