Fundamentals of Machine Learning - Final Report

By: Jeremy Glasgow

Executive Summary

Power generation in the United States comes from three energy sources which are oil, natural gas, and coal. Oil as an energy source makes up approximately 9.3% of power generation in the United States with an average cost of $17.77 per one million British thermal units (MMBtu). Natural gas as an energy source makes up approximately 55.3% of all power generation in the United States with an average cost of $19.73 per MMBtu. Coal as an energy source makes up approximately 35.4% of all power generation in the United States with an average cost of $2.44 per MMBtu. Coal is significantly more cost efficient than natural gas and oil. Coal contains approximately 9.57% ash content which contains selenium and mercury, two major contributors to cancer, water and air pollution, and other health risks in humans.

After performing cluster analysis, the following conclusions have been found:
1. A decrease in sulfur content in oil will result in an increase in cost per MMbtu.
2. Changes in one energy source price will not affect others.
3. A decrease in ash content in coal will result in an increase in a higher quantity per MMbtu, but will not affect price per MMBtu.
4. Gas and coal both have seen an increase in quantity per MMBtu which results in an overall decrease in price for power generation.
5. Natural gas usage volume has increased while coal has decreased.

Introduction

One of the biggest social concerns globally is the effort to reduce pollution. As other energy sources for power generation have been introduced such as solar power and wind turbines, these energy sources fall short to oil, natural gas, and coal in adoption due to the higher costs, availability, and reliability.

The dataset used in this report is a collection of power plant deliveries from 2008 until 2022. This dataset is quite large and contains many missing values. According to the summary of the dataset, it indicates around ⅓ of all prices are redacted from public databases because this information is proprietary for companies. Due to the significant missing values for pricing, this report will use the fuel costs data available to provide statistics, predictions, and trends to provide insight into power generation in the United States.

The first operation I performed during the data cleaning phase, I gathered a list of all the labels used in the dataset and removed the redundant columns. Next, I removed all columns with 10,000 or more missing values. I also standardized the data types in the columns and converted recorded dates to individual columns for the month, day, and year. Cost of fuel will not be used as a variable for clustering due to the significant missing values.

Problem Statement

At first glance, the data does not show trends which makes it difficult to extract information from the dataset. In this report, I will explore historical usage, cost, and trends within power generation in the United States. Through advanced analysis, I will uncover pricing, composition, and variable relationship trends for the different energy sources.

Analysis and Discussion

I chose Guassian Mixture Models and K-Means as the two models to compare against each other for cluster analysis. I performed visual analysis using the elbow method to find the optimal number of clusters. For Gaussian Mixture Models, the metric produced is the Bayesian Information Criterion (BIC) score that evaluates model performance. A lower BIC score, relative to other BIC scores, is used to determine the best number of clusters. For K-Means, the metric produced is the Within-Cluster Sum of Squares (WCSS) score. A lower WCSS score, relative to the other WCSS scores is used to determine the best number of clusters. Silhouette Scores are a measure to determine how similar a datapoint is to its cluster. Pictured below in Figures 1, 2, and 3 are the results of cluster analysis using only numerical data.

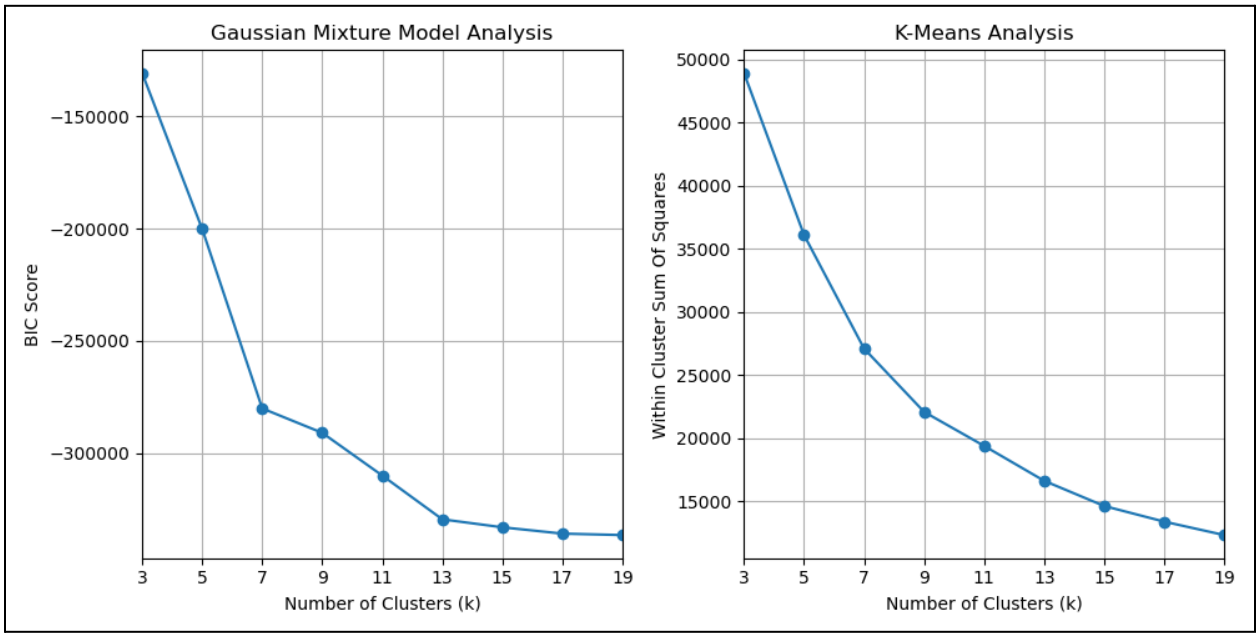

Figure 1 - Elbow Method visual analysis before adding dummy variables.

	Number of Clusters	Silhouette Score
0	3	0.285854
1	5	0.233369
2	7	0.271011
3	9	0.294631
4	11	0.296635
5	13	0.304448
6	15	0.313043
7	17	0.314175
8	19	0.320413

Figure 2 - K-Means Silhouette Scores before adding dummy variables.

	Number of Clusters	Silhouette Score
0	3	0.178342
1	5	0.168614
2	7	-0.011571
3	9	-0.059658
4	11	-0.072650
5	13	0.061389
6	15	0.025475
7	17	-0.013424
8	19	-0.029514

Figure 3 - Gaussian Mixture Models Silhouette Scores before adding dummy variables.

These metrics found in Figures 1, 2, and 3 were not very strong so I decided to include categorical data. After looking through the dataset, I included fuel types and the source of energy as variables represented by dummy variables. Below in Figures 4, 5, 6, and 7 show analysis after including categorical data.

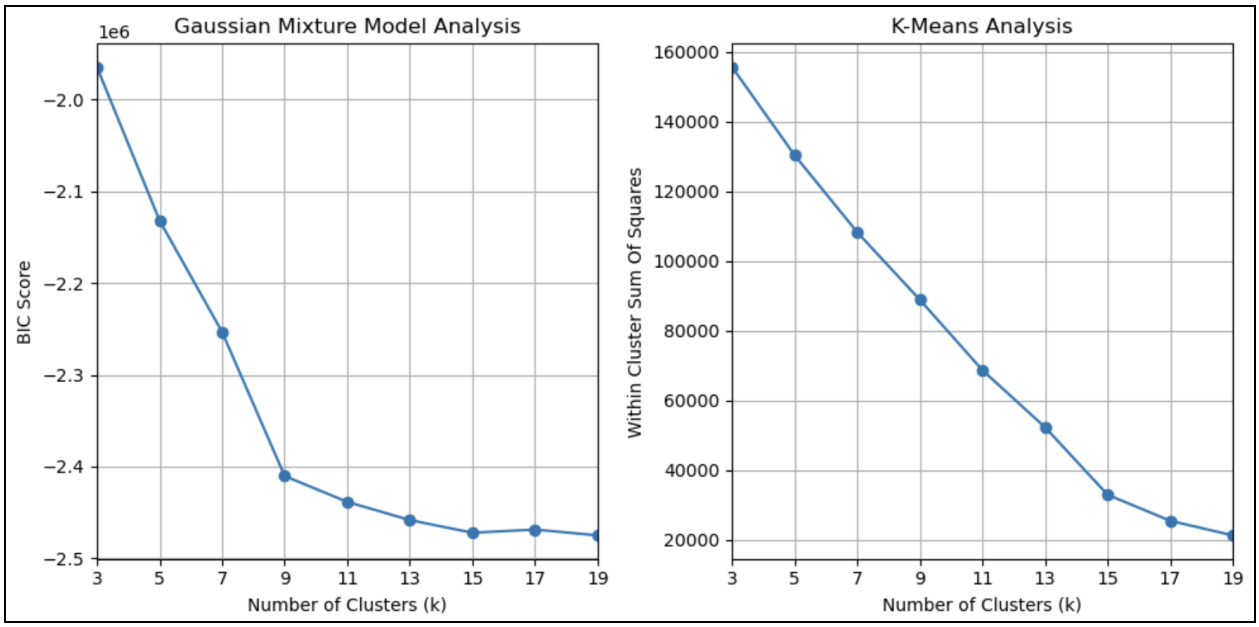

Figure 4 - Elbow Method visual analysis after adding dummy variables.

	Number of Clusters	Silhouette Score
0	3	0.449510
1	5	0.484278
2	7	0.496255
3	9	0.506451
4	11	0.533518
5	13	0.426265
6	15	0.434723
7	17	0.422172
8	19	0.402348

Figure 5 - K-Means Silhouette Scores before adding dummy variables.

	Number of Clusters	Silhouette Score
0	3	0.449510
1	5	0.298597
2	7	0.347805
3	9	0.536204
4	11	0.535432
5	13	0.541555
6	15	0.278165
7	17	0.216106
8	19	0.257746

Figure 6 - Gaussian Mixture Models Silhouette Scores after adding dummy variables.

	Number of Clusters	Silhouette Score
0	3	0.449510
1	5	0.491211
2	7	0.507518
3	9	0.522891
4	11	0.540344
5	13	0.542597
6	15	0.415348
7	17	0.416032
8	19	0.373708

Figure 7 - Agglomerative Clustering Silhouette Scores after adding dummy variables.

After adding in the categorical variables, the silhouette scores were much better for both models. I also added Agglomerative Clustering as an additional model to perform clustering analysis with. Figure 7 shows the silhouette scores for agglomerative clustering. I selected K-Means and 11 clusters for segmentation and interpretation of the data because the silhouette score was the highest for this value.

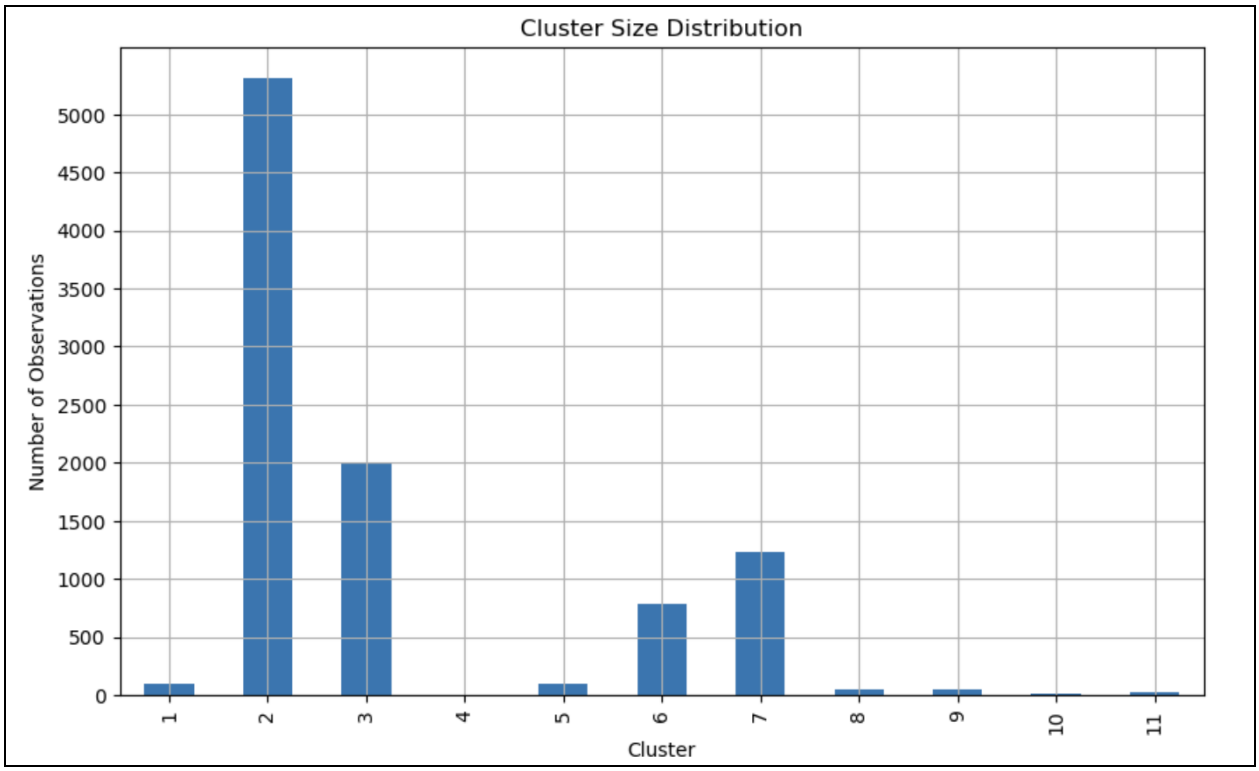

Figure 8 - K-Means cluster distribution using 11 clusters.

	Cluster Label	Average Sulfur	Average Ash	Most Common Fuel Type
0	1	2.42	31.10	coal
1	2	0.00	0.00	gas
2	3	1.87	10.51	coal
3	5	0.49	0.00	oil
4	6	0.08	0.00	oil
5	7	0.30	5.34	coal
6	8	5.15	0.36	coal
7	9	0.73	11.27	coal
8	11	0.00	0.00	gas

Figure 9 - Average ash and sulfur content, with the most common fuel type found in each cluster.

	Cluster Label	Most Common Fuel Type	Fuel Per Unit	Cost Per Unit	Fuel Received	Total Cost
0	1	coal	17.58	2.33	32386.50	4288.89
1	2	gas	1.03	19.73	278665.31	5331135.92
2	3	coal	23.84	2.85	30690.62	3671.62
3	5	oil	6.33	13.28	76797.94	161031.44
4	6	oil	5.79	18.07	3669.89	11458.12
5	7	coal	17.52	1.87	69557.65	7438.69
6	8	coal	28.44	2.01	28280.85	2000.91
7	9	coal	13.57	2.20	222281.17	35953.34
8	11	gas	0.77	5.61	70639.50	517684.64

Figure 10 - Fuel Costs associated with the most common fuel types in each cluster.

After segmenting the data into clusters, Figures 9 and 10 show ash and sulfur content, fuel costs, fuel usage, and total costs associated with the clusters.

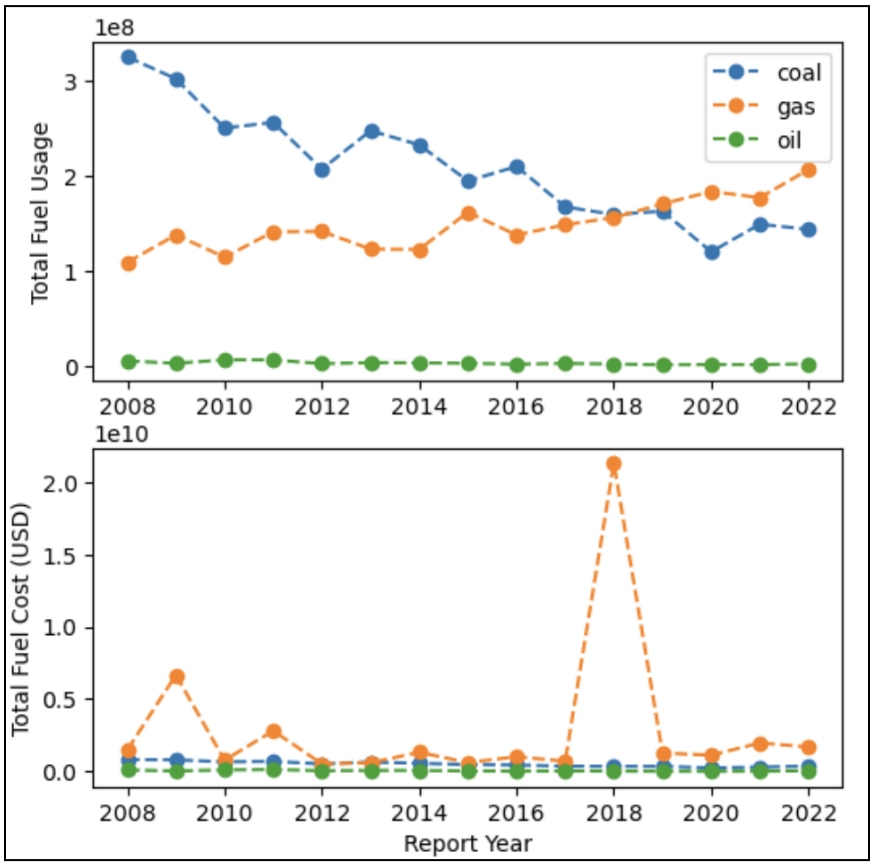

Figure 11 - Historical fuel usage and fuel costs for each fuel type.

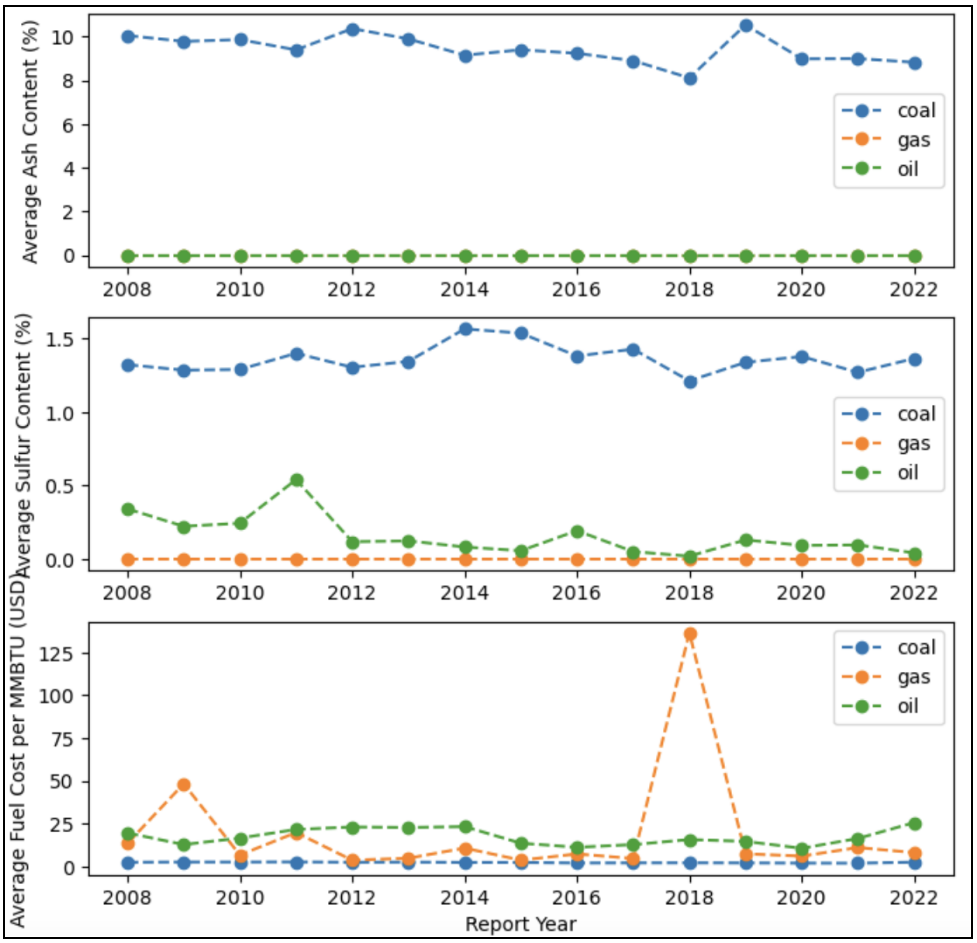

Figure 12 - Historical sulfur content, ash content, and prices for each fuel type.

The historical data in Figures 11 and 12 show a decrease in coal usage and an increase in oil for power generation. Coal contains the highest ash and sulfur contents which are harmful to the environment and cause pollution. This clearly depicts that decreasing coal usage is a part of the ongoing pollution reduction and sustainability efforts.

Conclusions

The prices of oil, natural gas, and coal have remained very consistent. Before performing cluster analysis, it seems the average quantity per MMBtu of natural gas and oil has remained consistent, while coal has decreased between 2008 and 2022. After performing cluster analysis and looking deeper into the historical trends, the data clearly shows the decrease in coal usage and increase in natural gas. Coal contains sulfur and ash which are both pollutants, which validates the global push to use sustainable practices for power generation in the United States.