Jacob Aylward

Dr. Wu

Fundamentals of Machine Learning

29 June 2025

Final Exam Report

Executive Summary:

Based on the sample dataset created and questions chosen to be answered it can be concluded that certain fuel types will be more frequently received despite other forms being more commonly used. Additionally, specific forms of fuel will be more commonly used by a large margin such as natural gas. In terms of power generation in the US this sample dataset shows that certain forms of fuel will require more units to be used to meet the same demand as compared to other fuels. Additionally, factors such as location and climate will affect the units needed for each type of fuel. Also, despite not being used to answer specific questions, the presence of mercury and chlorine in each type of fuel should be considered when determining the best kind to use. This consideration can also impact the overall demand for each type of fuel and how commonly they are used. Although cost will play a factor in determining what fuel type may be used for some consumers the overall quality and health should also play a part in the decision making process.

Introduction:

Due to the many variables and rows of data for this given dataset a sample of data was created to answer some questions regarding different fuel types used and the relationship between some variables. For the sample dataset, it was created using 2% of the full dataset and had some variables removed. The removed variables were chosen based on the amount of “null” values listed as they could result in inaccurate answers. There were five variables chosen to be removed that included “mercury_content_ppm”, “natrual_gas_delivery_contract_type_code”, “moisture_content_pct”, “chlorine_content_ppm”, and “mine_id_msha” as they were all deemed unnecessary for answering the chosen questions and due to the large amount of “null” values each variable contained. To ensure uniqueness in the 2% sample data a random four digit seed number was decided upon with it being “set.seed(1738)”.

Problem Statement:

After reviewing the variables of the sample data, the questions chosen to be answered revolve around each type of fuel that is used. These questions will aim to determine what kind of fuel type is most commonly used and if the average amount of fuel units ordered for each type of fuel supports the answer to the kind that is most used. Based on those answers we will make educated assumptions based on the difference in values for the most common type used and average number of units ordered for each type. Additionally, a cluster analysis graph will be used to compare the relationship between the variables “fuel_received_units” and “fuel_mmbtu_per_unit” with the intention of seeing if the clusters support the answers to the starting two questions. This cluster chart will also allow us to learn which orders gain the most value for their units of fuel based on the amount of fuel in each unit. The questions aimed to be answered will be given below:

- From the 2% sample of data what is the most common fuel type used?
- Does the average amount of fuel received units for each type support the kind that is most commonly used? If not, what are some assumptions that can be made?
- What are the clusters like for “fuel_received_units” and “fuel_mmbtu_per_unit” and

what can this tell us based on the units received and fuel in each unit?

Analysis & Discussions:

In order to answer the questions given above the first task was to find how often each type of fuel was used based on the 2% sample data. From this data it was determined that natural gas was the most common fuel type with a usage of 7,946 of around 13,000 different orders. When comparing natural gas usage to the other forms of fuel it can be seen that natural gas was used by a large margin based on the second most common fuel type used being coal with a usage of 4,720. From there the usage values of the remaining three fuel types was 1,151, 58, and 50 which better shows how much natural gas was the preferred fuel type. From there the next question to answer was finding the average of units received for each type of fuel. Based on the information found the highest average of fuel units received was “other gas” with 1,735,262.89 units being received overall based on all orders that used “other gas”. Due to the average units received not supporting natural gas being the common form of fuel used the cluster chart for question 3 will give insight as to why. For the cluster chart the goal is to see the relationship of the fuel units received and the fuel in each unit based on the amount received. To determine the optimal number of clusters I used k-means and implemented both the “silhouette” and “wss” methods. Through both of these methods the optimal number clusters found was k = 4. The results showed 4 different sized clusters with many of the orders showing less than 5 fuel units received and the amount of fuel per unit varying heavily from -1 to 2. Based on the clusters it was determined that many of the “other gas” orders likely had low values of fuel per unit resulting in more units needing to be ordered to meet demand and thus raising the average to be the highest. For natural gas it was determined that most of their orders were in clusters 1 and 4 where there were less than 5 units per order and the fuel per unit varying heavily. This would support natural gas being the second most received in terms of average but also being the most common form used.

Conclusions:

Based on the sample data given and questions asked it’s important to note that due to the small sample size the results could be different for other samples. Regardless, in terms of the whole data set it can be concluded that the amount of units received will be heavily based on the amount of fuel per unit which can lead to more units being ordered to certain forms of fuel resulting in the average being higher than hypothesized. Additionally, these orders can vary on each individual state and the seasons that the orders occur in. Certain fuels may be preferred in specific climates resulting in some sample dataset being skewed towards one type. Overall this dataset is very helpful in understanding the value of different fuel types and the demand that may occur for each type of fuel depending on a number of factors.