Final Project

Fundamentals of Machine Learning

Chase Holland

17 December 2023

Executive Summary:

This project covers the different aspects associated with fuel delivery to power plants. By using a data set which covers a number of variables with fuel delivery, we are able to analyze different aspects of the costs of energy generation in the United States along with its consumption. According to the data source, this is the most granular data available that describes the costs of fuel as well as the marginal cost of the generation of electricity. Closely tracking and analyzing the costs of energy generation can help the United States to find the best ways to create greater efficiencies and cut back on unnecessary consumption and costs.

By analyzing the data with concepts learned in machine learning, we are able to better understand the trends and relationships within the raw data. We are able to cluster data based on sulfur and ash content of fuel and we can see whether money is being spent on efficient fuel types or if it could be better relegated elsewhere.

Introduction:

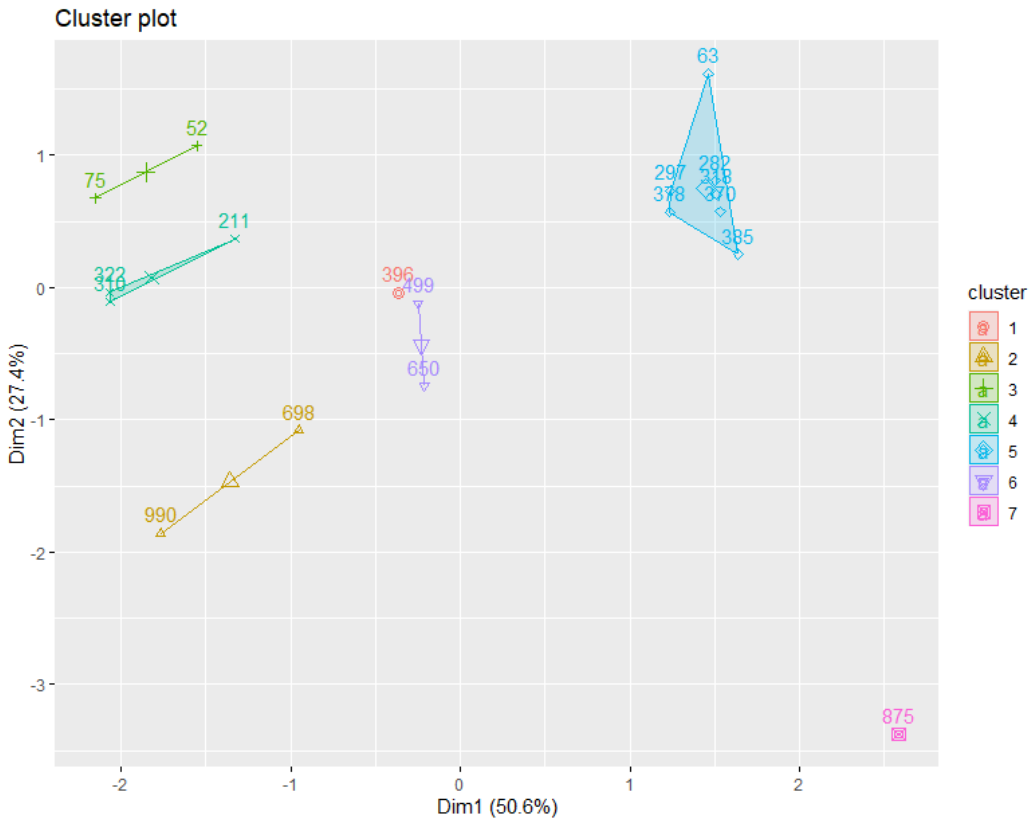
The data in this project comes from PUDL, which is an open-source data pipeline containing data points related to costs in energy delivery in the United States. There are over 600,000 rows in the data with 20 variables, including fuels units received, sulfur content, ash content, cost per mmbtu and primary form of transportation of the energy sources. I used a random sample of 2% of data using a random four-digit number, then splitting the data into training sets using 75% of the sampled data.

Problem Statement:

The code used is for determining two main things. First, we look at the clustering of data in order to understand what types of segmentations are involved with the raw data. Next, we look at how this data helps us understand power generation and where inefficiencies may lie.

Analysis and Discussion:

After sampling the data, segmenting into training and test sets and removing categorical variables, I searched for the appropriate K value and clustered the date to prepare it for interpretation. I was able to find seven main clusters for the data, mostly based on the sulfur and ash content of the raw source, as seen below:



Conclusion:

By using machine learning data analysis, we are better able to visualize and understand the usage and costs of energy and where inefficiencies may lie. By finding areas where less efficient fuels are incurring outsized costs, we are able to identify where federal policy may be able to address these issues and how they would be able to have the largest impact.