Jocelyn Ohlemacher

BA 64060 Final Exam

10 Dec 2023

**Executive Summary:**

Data available in the Public Utility Data Liberation Project (PUDL) table EIA-293 was used to evaluate clustering and trends in fuel purchasing. In each of the three clustering methods used, clusters primarily formed around fuel type and subtype. Further clustering within the three fuel types (coal, gas, oil), clusters were again primarily driven by fuel subtype, though some smaller clusters appeared with extreme values such as exceptionally high fuel cost per MMBtu or units received. Among the larger clusters, there was overall stability in units received and fuel cost per MMBtu. Given the distinguishing attributes of fuel type observed during data exploration, it is no surprise that fuel type was the primary driver of cluster formation.

**Introduction:**

The data used in this analysis was obtained from the Public Utility Data Liberation (PUDL) Project, table eia923, “fuel receipt costs”. This data includes details on purchase information (cost per unit, quantity of units), fuel type purchased, attributes of the fuel type (sulfur and ash content), transportation information, and month reported.

For analysis, a random 4% sample of the data was obtained from the original dataset. In order to prepare the data for analysis, variables with a significant number of missing data were removed, variables that contained duplicate information were removed, and categorical variables were one-hot coded to make them suitable for analysis. Date reported field was converted to season (winter, spring, summer, fall) for purposes of temporal evaluation. Finally, cases with missing fields were removed, and the numerical variables were mean scaled.

**Problem statement:**

Various clustering methods will be used to evaluate patterns in the PUDL fuel receipts data. Specifically, to evaluate if fuel types tend to cluster together, if within fuel types there are trends or patterns that create separate segments, and what the driving factors of the clusters are. Clustering will also be used to evaluate seasonal/temporal patterns of fuel purchasing.

**Analysis and discussion:**

From initial data exploration, we see that fuel types (coal, gas, oil) are distinct considering all available numerical variables. We see fairly stable fuel cost by report month. Total units received by report month is also overall stable, though there is some increase from June through August, suggesting potential stockpiling for upcoming colder months.

In performing k-means clustering, we use the silhouette method to determine best k=2, where results show cluster 1 being entirely comprised of coal fuel types and cluster 2 of gas and oil together. Not surprisingly, we see that the coal cluster has higher MMBtu per unit and lower cost per unit. To further evaluate, we isolate each fuel type and repeat k-means clustering to evaluate trends within fuel type.

Within fuel type coal, we see petroleum coke stand alone as its own small cluster (cluster 7), and the remaining clusters are primarily driven by energy source code (sub-types of each fuel). For example, clusters 2, 3, 5, and 8 are primarily comprised of bituminous coal, and clusters 1, 6, and 9 of sub-bituminous coal. Further evaluating fuel sub-types, we see some seasonality trend in sub-bituminous coal and sulfur, ash, and MMBtu per unit trends in the bituminous coal. Cluster 4 is a cluster of mixed subtypes with exceptionally high ash content.

|  |  |  |
| --- | --- | --- |
| **Cluster** | **# Cases in Cluster** | **Description** |
| Cluster 1 | 958 | Sub-bituminous coal, spring and winter purchases |
| Cluster 2 | 467 | Bituminous coal, spot purchased |
| Cluster 3 | 840 | Bituminous coal, primarily transported via river, high sulfur content |
| Cluster 4 | 75 | Mixed coal types with very high ash content |
| Cluster 5 | 442 | Bituminous coal, primarily transported via railroad |
| Cluster 6 | 490 | Sub-bituminous coal, fall purchases |
| Cluster 7 | 60 | Petroleum Coke, high sulfur content |
| Cluster 8 | 1233 | Bituminous coal, lowest sulfur content of bit coal, highest ash content |
| Cluster 9 | 472 | Sub-bituminous coal, summer purchases |

Within fuel type oil, we again see clustering by sub-type. Unlike coal, we see more clustering around seasonality and contract type, particularly among distillate fuel oils. With the exception of residual fuel oils, the primary method of transportation for oil products is via truck.

|  |  |  |
| --- | --- | --- |
| **Cluster** | **# Cases in Cluster** | **Description** |
| Cluster 1 | 128 | Distillate fuel oils, spot & summer purchase |
| Cluster 2 | 87 | Distillate fuel oils, contract, fall purchase |
| Cluster 3 | 161 | Distillate fuel oils, spot & winter purchase |
| Cluster 4 | 122 | Distillate fuel oils, spot & fall purchase |
| Cluster 5 | 81 | Distillate fuel oils contract, summer purchase |
| Cluster 6 | 62 | Residual fuel oil, primary transport via plane |
| Cluster 7 | 10 | Distillate fuel oils, contract, winter purchase |
| Cluster 8 | 159 | Distillate fuel oils, spot & spring purchase |
| Cluster 9 | 103 | Distillate fuel oils, contract, fall purchase |

Within fuel type gas, we see that there are no prominent subtypes and that the primary method of transportation for all clusters is pipeline. We see a spot purchase which occurs in each season, and small clusters with exceptionally high units received or exceptionally high cost per unit.

|  |  |  |
| --- | --- | --- |
| **Cluster** | **# Cases in Cluster** | **Description** |
| Cluster 1 | 61 | Highest number of units received per purchase |
| Cluster 2 | 1269 | Spring, spot purchases |
| Cluster 3 | 1499 | Contract, low units received, consistent across seasons |
| Cluster 4 | 1307 | Summer, spot purchases |
| Cluster 5 | 420 | Second highest number of units received per purchase |
| Cluster 6 | 1224 | Fall spot purchases |
| Cluster 7 | 5 | Exceptionally high fuel cost per unit, fall purchases |
| Cluster 8 | 26 | Second highest fuel cost per unit, no seasonality |
| Cluster 9 | 1339 | Winter spot purchases |

Following k-means, Dbscan clustering was performed to evaluate if different segmentation would appear based on density and if any outliers become apparent. Again, clusters are formed primarily around fuel type/subtypes. Of the three natural gas clusters, one contains the majority of natural gas purchases, while clusters five and six contain few purchases with high units received and high fuel cost, respectively.

|  |  |  |
| --- | --- | --- |
| **Cluster** | **# Cases in Cluster** | **Description** |
| Cluster 1 | 7128 | Natural gas |
| Cluster 2 | 4976 | Coal |
| Cluster 3 | 1009 | Petroleum |
| Cluster 4 | 59 | Petroleum coke |
| Cluster 5 | 6 | Natural gas |
| Cluster 6 | 5 | Natural gas |

The cases identified as outliers primarily consist of very high cost and relatively high units received natural gas purchases.

Using hierarchical clustering, we again see that clusters form largely based on fuel types and subtypes. Seasonality does not appear to play a substantial role in cluster formation using hierarchical clustering method.

|  |  |
| --- | --- |
| **Cluster** | **Description** |
| Cluster 1 | Natural gas |
| Cluster 2 | Coal |
| Cluster 3 | Coal |
| Cluster 4 | Petroleum oil |
| Cluster 5 | Natural gas |
| Cluster 6 | Natural gas |

Further narrowing in by fuel type coal, we do not see petroleum coke become its own cluster as it did in other clustering mechanisms. We see three clusters form, fuel MMBtu per unit, ash and sulfur content are the primary attributes of the clusters.

|  |  |
| --- | --- |
| **Cluster** | **Description** |
| Cluster 1 | High ash content, high MMBtu per unit |
| Cluster 2 | Low sulfur and ash content, low MMBtu per unit |
| Cluster 3 | High sulfur content, high MMBtu per unit |

Evaluating gas purchases specifically, we see that two cluster form, a very small cluster with extremely high cost per MMBtu, and a much larger cluster with more moderate attributes.

Finally, evaluating oil, we see three clusters, where again we see that the attributes of the subtype are relevant.

|  |  |
| --- | --- |
| **Cluster** | **Description** |
| Cluster 1 | Distillate fuel oil, spot purchase, transport via truck |
| Cluster 2 | Mix of distillate fuel and residual fuel oils, mix transport methods |
| Cluster 3 | Distillate fuel oil, contract purchase, primary transport via truck |

**Conclusion:**

In evaluating the US fuel purchase data sourced from PUDL via clustering methods, it was found that fuel types and sub-type were the primary factor in cluster creation. Each of the three clustering methods used produced similar results – clusters by fuel type, and smaller clusters which identified groups with extreme values, such as extremely high units received or cost per MMBtu. In this data, it was assumed that report month corresponded directly with purchase month. There were no assumptions regarding time of use for the fuel purchased.