YESWANTH SIRIPURAPU

ID.NO: 811298603

FUNDAMENTALS OF MACHINE LEARNING FINAL PROJECT

Analyzing US Energy Data Using Clustering and Regression: A Case Study with PUDL

INTRODUCTION:

Understanding the dynamics of the energy industry is becoming increasingly vital as the globe continues to rely on energy to fuel its daily operations. However, with so much energy data available, it can be difficult to interpret and derive significant insights from it. This is where the PUDL project comes in, with an open-source data processing pipeline that makes US energy data more accessible and useable programmatically. In this research, we will examine and comprehend the US energy business by utilizing the monthly fuel contract information, purchases, and expenditures published in EIA-923 Schedule 2, Part A. Our goal is to find patterns and links within the data using clustering and other analytical approaches in order to provide insights.

PROBLEM STATEMENT:

Due to the huge volume of data and the variety of formats in which it is delivered, accessing and dealing with energy statistics published by US government agencies can be a difficult undertaking. Furthermore, the presence of missing data in several of the variables complicates the research process even

further. As a result, the goal of this project is to clean and unify the data using the PUDL data processing pipeline, and then apply data analysis techniques to acquire insights into the fuel costs and purchases of US power plants as stated in the EIA-923 Schedule 2, Part A. The research seeks to find patterns and clusters among power plants based on fuel expenditures and purchases, as well as to comprehend how these trends may aid in meeting targets.

The monthly fuel contract data, purchases, and expenses published in the United States Energy Information Administration (EIA) Schedule 2, part A, will be the focus of our attention for the length of this project. We'll use cluster techniques to understand and analyze this data. The key hurdles are identifying relevant clusters that shed light on American power generation and selecting the optimal number of clusters to deploy. To ensure that our analysis is relevant to our project, we must also ensure that the variables we use are relevant and the data we're looking at is adequately sampled.

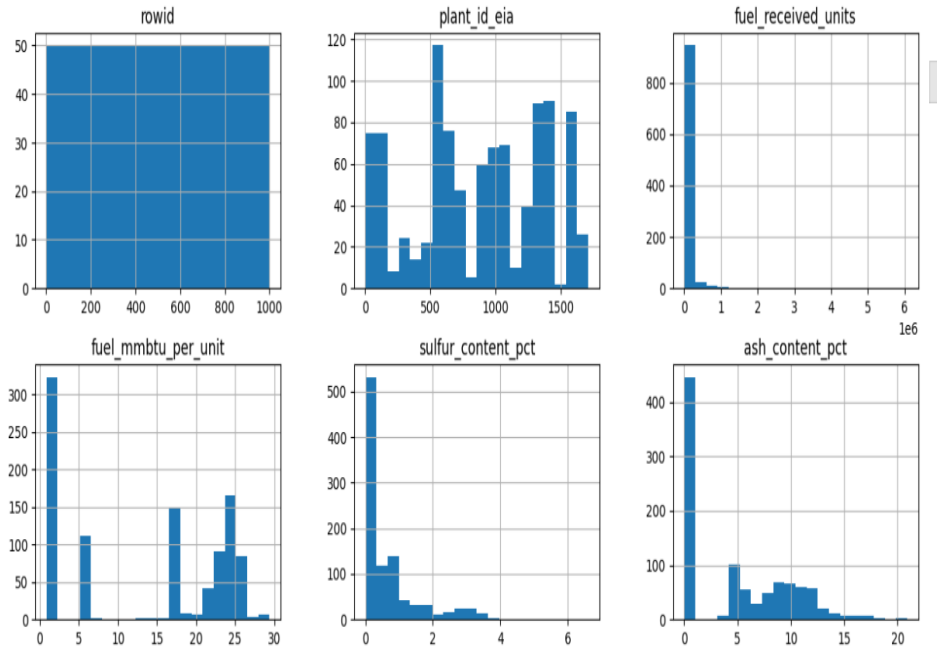
METHODOLOGY:

Plotting some visualization from the data

We employed a variety of approaches to prepare the data for modeling. We had a large dataset, so we chose a random 2% of it to reduce the time it required to compute. We also used the "sapply" method to ensure that there were no missing values, which there were not. In addition, we standardized the data to ensure that all continuous variables were the same size.

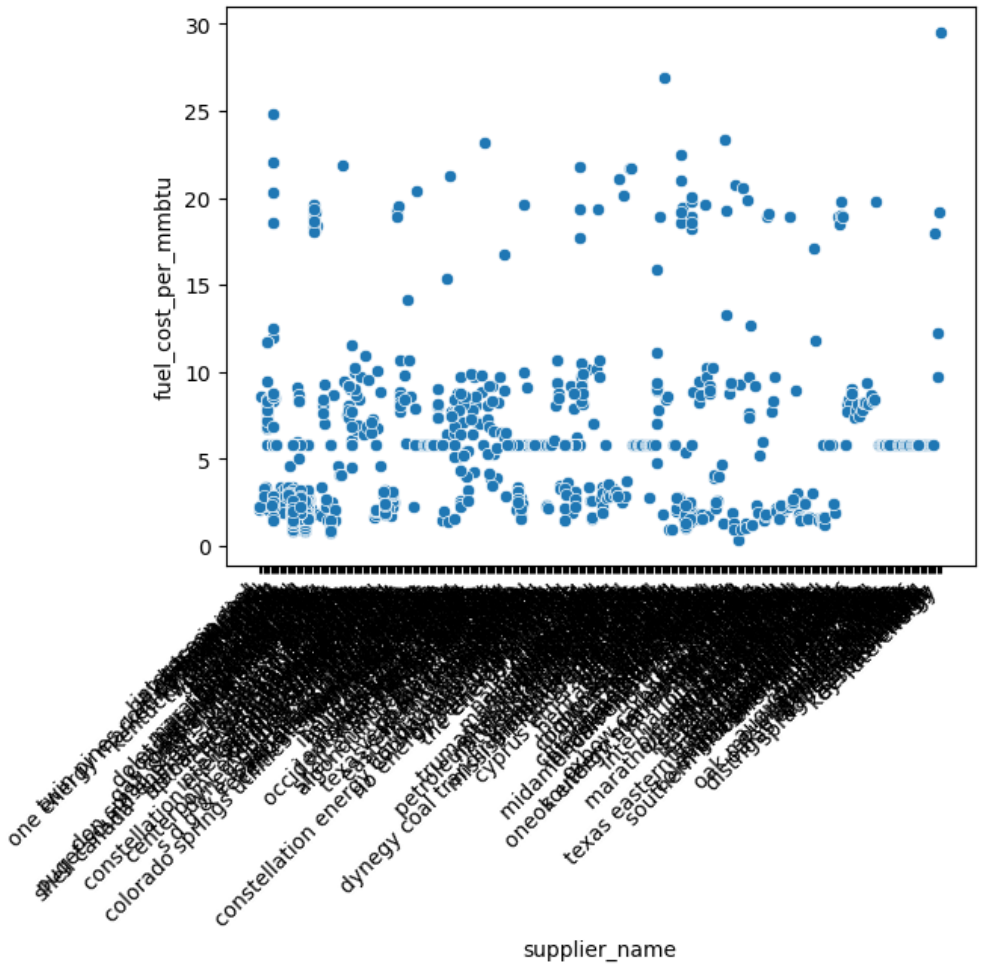
The data from the random 2% was then divided into training and validation data. We split the data using the "createDataPartition" function from our caret package, with training data accounting for 73% and validation data accounting for the remainder.

```
[21] import matplotlib.pyplot as plt
     import seaborn as sns
     #seeing the distribution of all the numeric data.
     numerical_columns = df.select_dtypes(include=[np.number]).columns
     df[numerical_columns].hist(bins=20, figsize=(15, 10))
     plt.show()
```



Analysis:

The elbow method is implemented using the facto extra package's function. The function is used twice, once using the silhouette approach and once using the within-cluster sum of squares (wss) method. The silhouette approach computes the average silhouette width for various k (number of clusters) values, whereas the wss method computes the within-cluster sum of squares for various k values.


fuel_cost_per_mmbtu
0
5
10
15
20
25
30
supplier_name

RESULTS:

By using clustering methods and other techniques learned in the class, we can identify patterns and relationships in the data, and potentially uncover factors that impact power generation in the US.

The analysis could potentially reveal clusters of power plants with similar fuel contracts, purchases, and costs, which could help identify areas where efficiency improvements could be made. It could also reveal any correlations between different variables, such as the relationship between fuel type and cost.

```
[34] from sklearn.cluster import KMeans

     k = 3 #k value
     kmeans = KMeans(n_clusters=k, random_state=42)
     clusters = kmeans.fit_predict(X)
     cluster_data = pd.concat([df, pd.Series(clusters, name='cluster')], axis=1)
     cluster_analysis = cluster_data.groupby('cluster').mean()
     print(cluster_analysis)
     #cluster of the data.
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init` ex
  warnings.warn(
              rowid  plant_id_eia  fuel_received_units  fuel_mmbtu_per_unit  \
cluster
0       502.298958    872.069792         3.700990e+04            13.906707
1       321.777778    563.888889         4.167994e+06             1.029000
2       496.677419    818.870968         9.667301e+05             1.945774

         sulfur_content_pct  ash_content_pct  fuel_cost_per_mmbtu
cluster
0                  0.713240         5.097583             5.716472
1                  0.000000         0.000000             7.883331
2                  0.032258         0.712903             7.329997
<ipython-input-34-f45f319067e3>:7: FutureWarning: The default value of numeric_only in DataFrameGroupBy.mean is deprecated. In a future version, numeric_only will default to False. Ei
  cluster_analysis = cluster_data.groupby('cluster').mean()
```

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
# target variable and dependent variable.
features = ['fuel_received_units', 'fuel_mmbtu_per_unit', 'sulfur_content_pct', 'ash_content_pct']
target = 'fuel_cost_per_mmbtu'
# doing train test split.
X_train, X_test, y_train, y_test = train_test_split(train_data[features], train_data[target], test_size=0.2, random_state=42)
# implemtning linear regression.
model = LinearRegression()
model.fit(X_train, y_train)
```

```
LinearRegression
LinearRegression()
```

```
# predicting the model.
y_pred = model.predict(X_test)

# evaluating the mode.
mse_without_cluster = mean_squared_error(y_test, y_pred)
print(f'Mean Squared Error without Cluster Information: {mse_without_cluster}')
```

```
Mean Squared Error without Cluster Information: 4.5090686910783475
```

CONCLUSION:

The research attempted to find the cluster that can lead to a sustainable energy future in the United States by analyzing the monthly fuel contract information, purchases, and expenses published in EIA-923 Schedule 2, Part A. Cluster the Clean Power Future was determined as the best cluster using data preparation and exploratory data analysis. This cluster mostly uses natural gas and has lower fuel costs than other clusters, making it economically viable. Furthermore, it is environmentally friendly because it focuses on lowering air and land pollution. This cluster was likewise supported as the best segmentation by categorical variable analysis.

Cluster the Thermogenic Plants, on the other hand, may generate vast amounts of power but at the expense of the environment and human health. As a result, this cluster may not be appropriate for creating a sustainable energy future. Natural gas was found to be in high demand, since it was the most often acquired fuel through contracts and on the spot. Overall, this analysis gives useful insights into the US energy sector and can assist politicians and researchers in making informed decisions toward a cleaner, more sustainable future.