Analysis of Power Generation and Classification of Census Regions of United States Power Plants

Vedant Gannu

Data Analytics 4000 Level

April 27, 2022

**Abstract**

Power plants are an engineering marvel that are extremely vital to the general population as they are the primary generators of manmade electricity. According to the EIA [1], “about 4, 116 billion kilowatthours (kWh) or 4.12 trillion kWh” were generated by United States utility-scale electricity generation facilities in 2021. The term “utility scale” refers to electricity generation with at least 1 MW or megawatt of generating capacity. Around “61% of this electricity generation was from fossil fuels—coal, natural gas, petroleum, and other gases. About 19% was from nuclear energy, and about 20% was from renewable energy sources”[1]. The three major fossil fuels that contribute to this 61% as mentioned previously are Natural Gas, Coal, and Petroleum whereas the renewable sources are contributed by Wind, Hydropower, Solar (including photovoltaic and thermal), Biomass (Wood, Landfill Gas, Biogenic waste, other kinds of biomass), and Geothermal. With this incredible generation capacity of just the United States alone to be able to supply electricity to the continually growing electric demands of the general population and technology, the major overhead cost the planet pays for is in the form of CO2, NO2, and SO2 emissions into the atmosphere. In 2020 the EIA reported that “all energy sources resulted in the emission of 1.55 billion metric tons—1.71 billion short tons—of carbon dioxide (CO2)” [2] of which coal, natural gas, and petroleum fuels accounted for 99% of those emissions. Considering power plants are globally considered the powerhouses of electricity generation, analyzing the efficiency and capacity of net electricity generation due to their associated efficiency and consumption capacity is extremely important in determining if power plants are generating the true electricity amounts that they are expected to produce. One of the major goals of this project is to analyze the net electricity generation of United States power plants attributed to their overall fuel consumption and total fuel consumption towards electricity generation, and to see if there is regression can be used to model it. A good regression fit and level of fit should in turn correspond to a how well fuel consumption the ratio of fuel consumption relates to the net generation of electricity. The second goal of this project is to try and determine if certain power plant generation information such as net electricity generation, fuel consumed for electricity generation, fuel type, and even the year of the data can be used to classify where the specific power plant was surveyed. Hypothesis 1: There will be no multivariate fit for net generation vs total heat from fuel consumption and heat produced for consumption towards electricity generation due to the high variations in power generation by small and large power plants. Hypothesis 2: Classification of census region using net electricity generation, year, fuel type, and heat produced for consumption towards electricity generation will yield a positive result.

**Data Preparation and Exploratory Data Analytics**

Finding power plant generation data is not extremely difficult as Google Datasets, the UN, Data.gov, and HIFLD Open Data are great sources of data for infrastructure and sustainability data, access to in depth power plant data such as the amount of fuel used overall, amount used towards electricity generation, gross and net generation, and even amount resold to consumers is extremely limited and in most cases also proprietary. As a result the best set of datasets that were acquired for this project have come form the EIA (U.S Energy Information Administration). The specific source of the data that was used comes from the EIA’s Form EIA-923 with previous data from Form-906/920. The actual data is collected in the form of Excel sheets that store data in the following form: Schedule 2 - fuel receipts and costs, Schedules 3A & 5A - generator data, including generation, fuel consumption and stocks, Schedule 4 - fossil fuel stocks, Schedules 6 & 7 - non-utility source and disposition of electricity, Schedules 8A-F - environmental data. The specific dataset being used comes from Schedules 3A & 5A which are compiled and located in Sheet 1 of the excel sheet and can be found in the reference 3 URL. For the goal of analyzing how well total consumption of fuel and total consumption for electricity generation relates to the net generation output, this project has aggregated data from 4 years-2018, 2019, 2020, and 2021 in order to better assess if the year is a source of discrepancy in disturbing a continuous multivariate regression model. The Exploratory Data Analysis is as follows:

At first glance the Generation and Fuel Data looks to be very much numerical heavy, but close analysis and cross calculating the "Total Fuel Consumption Quantity", "Electric Fuel Consumption Quantity", "Total Fuel Consumption MMBtu", "Elec Fuel Consumption MMBtu" with columnar data from Generator Data, Coal Stocks, Oil Stocks, and PetCoke Stocks data shows that these total columns are essentially aggregations of the monthly data that is provided in Generation and Fuel. Due to the fact that each power plant’s Consumption Quantity columns are amounts attributed to their respective physical unit labels (megawatthours, barrels, mcf, short tons, etc.), this analysis chooses the Consumption MMBtu columns due to the standardization of units. MMBtu stands for Metric Million British Thermal Unit, is a unit used to measure heat content or energy value of quantities of substances. The first thing this analysis aims to look at is the distribution of net generation across the 4 years by fuel type. Of the 2 fuel type recording columns, the analysis has chosen AER Fuel Type since it is much broader and more grouped than Reported Fuel Type which is provided by power plants. Before proceeding, it is important to generate the visualizations with respect to classification of nonrenewable and renewable fuels. The chosen fuel types to focus the analysis on are ["SUN", "GEO", "HPS", "HYC", "MLP", "NUC", "ORW", "WND", "WWW"] for renewable and ["COL", "DFO", "NG", "OOG", "PC", "RFO", "WOC", "WOO", "OTH"] for nonrewable sources. Information on the specific fuel names for these abbreviations can be found in the datasets’ spreadsheets. Other sources such as ["OTH", "PUR", "MWH"] from the Reported Fuel Type column were discarded to avoid ambiguity in net generation analysis across multiple fuel sources and possible outliers less used fuel types. In addition to the fuel source selection, the data was filtered to only include rows where the Net Generation was nonzero and Elec Fuel Consumption MMBtu was 0 as they are instances of unclean data. When the data is visualized with no adherence to aggregating the rows by plant id, AER fuel type, and year (since the collected data spans over 4 years), Figures 1 and 2 result. The visualization shows heavy right skewing across all fuel types, and heavy presence of outliers. This is due to the initial data being collected and plotted for each individual generator per power plant. Before the data is combined for visualization of distributions of net generation across the different fuel types with each power plant’s amounts aggregated properly per fuel type, we can that the distributions of the net generation across the 4 years are not drastically different, almost signifying that net generation across for the 2,534 power plants sampled hasn’t changed too much. The coal fuel type shows that there was a bit of a decrease and then increase in generation, whereas natural gas had a constant distribution for 3 years and then sharp increase/skewing in 2021. One notable observation here is that renewable energy sources overall produce less than nonrenewable except for nuclear energy which topples coal, natural gas, and petroleum coke. Figures 3 and 4 are much better and correct representations of the net fuel generation distribution across multiple fuel types over the 4 years. There are minimal outstanding outliers in the visualization even though the data is still skewed right for many of the less used fuel types in both nonrenewable and renewable fuel types. Figures 5,6,7 and 7 were attempts at closely looking at the distributions of some of the less used fuel types for both types of fuel to see if their distributions are also somewhat constant. Even though the distributions are completely constant, attempting to fit a multivariate model is still a way to see if the fuel types and their heat output even matter when it comes to assessing their effect on net generation of power. Figures 8 and 9 are visualizations of the total fuel consumption in MMBtu across the various fuel types over the 4 years for both fuel types and Figures 10 and 11 are the boxplot distributions for fuel consumption towards electricity generation. In all of the visualizations, it appears that each of the columns follows a similar distribution and shape which makes these columns potentially good for regression since they follow in the same direction as the net generation. For outlier removal, it was decided that outliers should only be removed from the response variable (net generation) since removing extremely large values is little bit more realistic in terms of fitting a model that more favors the IQR rather than extremely large or small values coming from large and small power plants. Figures 12 and 13 are visualizations of the net generation with these outliers removed. Outlier were classified as values that are in not in the (Q3-Q1) \*0.5 and (Q3-Q1)\*1.5 range. Unfortunately, the distributions across the multiple years are not consistent anymore meaning that there is less confidence in producing a model that is applicable to multiple years.

Chart, box and whisker chart

Description automatically generated

Figure1

Chart, box and whisker chart

Description automatically generated

Figure 2

A picture containing chart

Description automatically generated

Figure 3

Chart, box and whisker chart

Description automatically generated

Figure 4

Chart

Description automatically generated

Figure 5

Chart, box and whisker chart

Description automatically generated

Figure 6

A picture containing application

Description automatically generated

Figure 7

A picture containing chart

Description automatically generated

Figure 8

Chart, box and whisker chart

Description automatically generated

Figure 9

Chart, histogram

Description automatically generated

Figure 10

Chart, box and whisker chart

Description automatically generated

Figure 11

Chart

Description automatically generated

Figure 12

Chart, box and whisker chart

Description automatically generated

Figure 13

**Model Development and Results**

The EDA process and reasoning for selection of a multivariate regression model is explained above. With the removed outliers for the net generation response variable, lets see how a model fit works for relating total heat consumption, total heat consumption for electricity, and net generation. The creation of the multivariable model uses a test train split with 33 percent of the data being used for the test set. Output 1 shows the results of the model fit where we see a adjusted r squared value 0.995 which is suspiciously large since there was so much variance and skewing in the data we saw in the EDA portion. The coefficients are also very small meaning that not the errors between the data points and estimation are very low indicating that not much correction needs to be added to the variables when predicting. Since the consumption fuel MMBtu for electricity and total consumption columns vary so closely together, a plot of net generation vs consumption fuel MMBtu for electricity was generated (Figure 14) which shows that there is actually very high correlation and association between the two variables with the outliers removed. This could be the reason for the very low coefficient and high adjusted r squared value

Graphical user interface, text

Description automatically generated Text

Description automatically generated with medium confidence

Output 1

Chart, scatter chart

Description automatically generated

Figure 14

Some of the flaws in this analysis might be that there was a heavy imbalance in the amount of CHP (Combined Heat power plant) plants vs non CHP plants. According to the dataset, the total fuel consumption MMBtu is identical to the total fuel consumption for electricity MMBtu. This imbalance of class is the reason why this model fit so well, because the number of CHP plants is significantly less than the number of non CHP plants. In the future, stratified sampling is highly recommended be a better method for maintaining external class imbalance before doing regression, or exploring the use of another predictor variable such as gross generation or internal plant electricity usage will be better.

Now to look at the classification problem. This project aims to attempt to use net electricity generation, fuel consumed for electricity generation, fuel type, and even the year of the data can be used to classify where the specific power plant was surveyed. The census regions surveyed by the EIA were: NEW, MAT, SAT, ESC, WSC, ENC, WNC, MTN, PACC, PACN. Figures 16 and 17 are visualizations of how the census regions are clustered according to the net generation with AER fuel type code, and the consumption for electricity generation with AER fuel type. The reason for these visualizations is to see if the census regions are distinguishable with changing net generation and consumption for electricity generation values, across different fuel types since different fuel types and generation/consumptions values might be prominent in some places rather than others. The clusterings of the census regions are very tight due to the scale of the net generation column, but there seems to be sections in each fuel type where the census regions are minorly clustered closely and around each other for different net generation values. KNN might be useful in this case since it is based on nearest neighbors, but this analysis will look at RandomForest classification since the ensemble learning method might be better for sifting out noisy points for a particular fuel type, net generation value, and consumption for electricity generation values along with operating through the overlapping clusters. The training process for this involved Label Encoding the categorical variables (AER fuel type and census region), splitting the data into training and testing, running 5 split cross validation on the training set, and then predicting on the testing data set. Output 1 shows the fold F1 scores for random forest with 100 estimators along with the final testing score and Output 2 shows the fold F1 scores for random forest with 500 estimators with the year excluded from the feature list along with final testing score. Although F1 score is extremely low which doesn’t necessarily imply that census region is not distinguishable by the features selected, it does imply that random forest doesn’t work here for classification. One interesting this to note is that removing the year feature and made the F1 score a little higher, but change is so small that the change might even be due to random chance or variation in the data inside the folds. But overall the random forest classifier does not work in this scenario.

Chart, scatter chart

Description automatically generated

Figure 15

Chart, line chart

Description automatically generated

Figure 16





Output 2





Output 3

**Conclusion**

In a subsequent exploration, rather than trying to look at just fuel consumption across different fuel types to justify net generation, it might be better to aggregate power plant as a whole and assess if net generation has a relationship with varying resale to customer values, incoming and outgoing electricity, and even revenue. In addition, sampling with respect to class imbalance is also something that should taken into consideration as a way to prevent to assessment of misleading results. In addition, assessing other classification techniques such as KNN and clustering could be used to classify census regions besides just using random forest.

References:

[1] EIA. (n.d.). *Frequently asked questions (faqs) - U.S. energy information administration (EIA)*. Frequently Asked Questions (FAQs) - U.S. Energy Information Administration (EIA). Retrieved April 27, 2022, from https://www.eia.gov/tools/faqs/faq.php?id=427&t=3

[2] EIA. (n.d.). *Frequently asked questions (faqs) - U.S. energy information administration (EIA)*. Frequently Asked Questions (FAQs) - U.S. Energy Information Administration (EIA). Retrieved April 27, 2022, from https://www.eia.gov/tools/faqs/faq.php?id=74&t=11#:~:text=In%202020%2C%20total%20U.S.%20electricity,CO2%20emissions%20per%20kWh.

[3] EIA. (n.d.). *U.S. Energy Information Administration - EIA - independent statistics and analysis*. Form EIA-923 detailed data with previous form data (EIA-906/920). Retrieved April 27, 2022, from https://www.eia.gov/electricity/data/eia923/