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.

**Data Processing 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. 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 with the exception of 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 net Figures 12 and 13 are visualizations of the net

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

Boxplots of net generation in megawatt hours of nonrenewable fuels sources for years 2018-2021

Chart, histogram

Description automatically generated

Boxplots of net generation in megawatt hours of nonrenewable fuels sources for years 2018-2021

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Instead of looking at distribution of net accumulated generation of each power plant, lets look at the accumulation for each power plant per fuel type since this gives a better view of outliers, spread, and significance of data.

Because there is no significant variation and changes in the boxplots between consecutive years, this might be an indication that we might be able to predict the net generation of power plants irrespective of year

References:

[1] <https://www.eia.gov/tools/faqs/faq.php?id=427&t=3>

[2] https://www.eia.gov/tools/faqs/faq.php?id=74&t=11#:~:text=In%202020%2C%20total%20U.S.%20electricity,CO2%20emissions%20per%20kWh.