CEE322: Data Analytics for Urban Systems

Lab 1: Exploratory Data Analysis

Author: Davyd Tamrazov

Due date: Sept 28, 2020

Stanford University

1 Introduction

Emissions from the electric power generation significantly contribute to the reduced air quality in the industrialized areas, while also having a detrimental effect on the environment globally. The aim of this project is to perform exploratory data analysis for Environmental Protection Agency (EPA) focusing on the air contamination associated with electric power generation. The dataset used for this analysis contains characteristics and emission details of electric power plants in the US for 2010. As such, the following report aims to understand problematic areas in the plants' performance and emission generation in order to identify areas where EPA should focus to achieve future emission reduction. Additionally, the impact of the fuel type on the plant's performance is assessed to help determine the optimal allocation of resources between different technologies for the EPA.

2 Dataset

The dataset used contains 5393 entries corresponding to the plants with non-zero generation and/or heat input characteristic. Each entry contains a unique identification number, name, and geographical information of the plant as well as information related to plants performance. The key variables from the latter can be described with three categories: energy resources (see Table 1), energy generation (see Table 2), and emission (see Table 3).

1. **Energy Resources**

Variable Name	*Description*
Plant Combustion Status	Takes the value of 1 for full combustion plants, 0.5 for partially combustion plants (combustion power plant that contains non-combustion generators), and 0 for non-combustion plants.
Plant Primary Fuel Type	Identifies plant's primary fuel type based on the maximum heat input as one of the 43 fuel types.
Plant Primary Fuel Category	Categorizes the Plant Primary Fuel Type variable into Coal, Oil, Gas, Nuclear, Hydro, Biomass, Wind, Solar, Geothermal, Other Fossil, and Other Unknown/Purchased/Waste (referred to as Other).

Table 1: *Summary of the plant energy resources variables*

2. **Energy Generation**

Variable Name	*Units*	*Mean*	*St. Dev.*	*Min*	*Median*	*Max*
Nameplate Capacity	MW	207.9	456.8	1.000	28.80	6809
Annual Heat Input	MMBtu	5.255e+06	1.858e+07	0.000	7210	2.450e+08
Annual Net Generation	MWh	7.663e+05	2.398e+06	1.000	4.501e+04	3.120e+07

Table 2: *Summary of the key plant energy generation variables*

Note: An additional set of variables is included in the dataset corresponding to the net generation by each fuel type, in total constituting annual net generation by a plant.

3. **Emission**

Variable Name	*Units*	*Mean*	*St. Dev.*	*Min*	*Median*	*Max*
CO_2 Equivalent Annual Emission	Short Tons	4.738e+05	1.881e+06	0.000	77.60	2.527e+07
NO_X Annual Emission	Short Tons	427.9	1854	0.000	0.580	3.884e+04
SO_2 Annual Emission	Short Tons	1011	5435	0.000	0.020	1.130e+ 05

Table 3: *Summary of the key emission variables*

2.1 Pre-processing

Combustion plants alone account for 72.7% of the total annual net generation from all plants, meaning that they constitute the primary way that electricity was generated in the US in 2010. Furthermore, plants with partial combustion represent only 2% of all combustion plants, generating 2.9% of the combustion fueled electricity. Thus, it is of prime importance for EPA to analyze and understand the emissions associated with the largest source of electricity – fully combustion plants. As a result, for the following exploratory data analysis, plant with partial and no combustion are excluded from the dataset.

Considering combustion plants only, 91.6% of the plants produce between 75-100% of electricity by the primary fuel category. Hence, given a small fraction of plants having mixed fuel sources, it is reasonable to analyze the dataset by identifying each plant with its primary fuel category.

Further, in order to improve the quality of the data analysis, the following additional ratio-based parameters were computed:

- *Plant Capacity Factor* = (Annual Net Generation / (Plant Capacity × 8760)) [unitless]
- *Plant Nominal Heat Rate* = 1000 × (Annual Heat Input / Annual Net Generation) [MMBtu/kWh]
- *Annual Output Emission Rate for Pollutant (CO_2, SO_2, NO_X)* = 2000 × (Pollutant Annual Emission / Net Generation) [lbs/MWh]

Finally, the data was cleaned for inconsistencies and apparent outliers. As such, entries with the negative electricity generation value in any of the fuel categories and/or annual net generation exceeding plant capacity were removed from the dataset. This resulted in the loss of approximately 1% of the data, what is not critical given the purpose of the exploratory nature of the analysis.

3 Analysis

In order to identify areas for improvement in combustion plants, it is first important to understand contribution of each fuel category to the electricity generation as well as the resulting emissions. The total amounts and percentages of the total for each combustion fuel category are summarized in Table 4 below.

Fuel Category	*Net Generated Electricity*		*CO_2 Emission*		*SO_2 Emission*		*NO_X Emission*	
	MWh	%	lbs	%	lbs	%	lbs	%
Oil	2.14e+07	0.74	3.96e+10	0.79	1.08e+08	1.02	7.41e+07	1.64
Gas	9.61e+08	33.30	9.23e+11	18.48	1.68e+08	1.59	4.61e+08	10.23
Coal	1.83e+09	63.32	3.97e+12	79.55	1.00e+10	94.93	3.82e+09	84.63
Other Fossil	4.45e+06	0.15	4.40e+09	0.09	2.15e+06	0.02	4.10e+06	0.09
Biomass	6.71e+07	2.33	5.42e+10	1.08	2.57e+08	2.44	1.54e+08	3.41
Other	4.70e+06	0.16	1.27e+08	0.00	1.22e+04	0.00	1.59e+05	0.00

Table 4: *Contributions of each fuel category to the net electricity generation and pollutant emission*

In this table, it is shown that 97% of the net electricity production by combustion fuel is generated by gas and coal fuel categories. On the other hand, oil fuel appears to contribute to a very small fraction of the generated electricity, as it is likely to be only utilized in exceptional circumstances, primarily due to its high price. Similarly, the large part of the CO_2 and NO_X pollutants result from the gas and coal plants, roughly corresponding to the fraction of the generated electricity. However, coal fuel appears to have a significantly larger contribution when compared to the amount of generated electricity. Likewise, SO_2 pollutants mainly result from coal combustion with the rest of fuel categories having an insignificant effect on the total amount emitted. From this table, it can be concluded that coal and oil constitute a larger percentage of emissions than that of the generated electricity, making them the least sustainable fuel categories. This hypothesis can be confirmed by analyzing pollutant emission rates, illustrated on Figures 1 and 2.

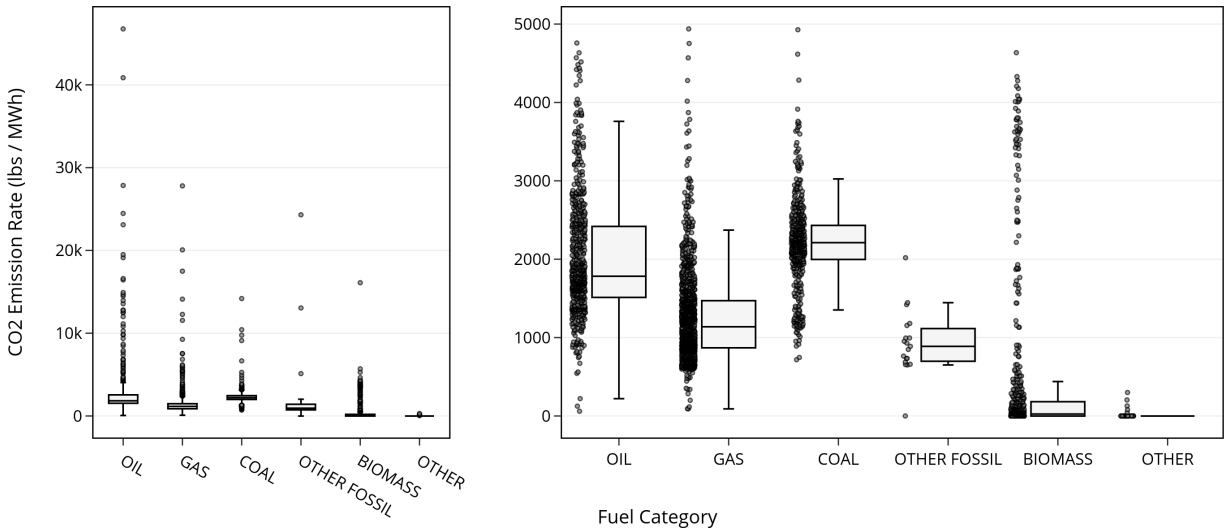

Figure 1: *CO_2 emission rates for combustion fuel categories*

From the left plot on Figure 1, it is evident that there is very large variability in CO_2 emission rates for every fuel category. However, limiting the range to between 0 and 5000 lbs/MMWh reveals that oil and coal plants indeed have noticeably higher CO_2 rates. On the other hand, biomass plants have a significantly lower variability of the CO_2 emissions, with the majority actually being CO_2 emission-free, making biomass a more sustainable choice of combustion fuel in this regard.

Similarly, Figure 2 illustrates emission rates of each of the combustion fuel category for the pollutants related to health concerns – SO_2 and NO_X. Note that 1.2% of data constituting apparent outliers were ignored for the purpose of clarity of this plot. There, it is clear that emission rates are extremely variable between different plants, with coal and oil fuel categories having the largest variability and emission rates in both pollutant classes. Oppositely, plants using gas and biomass fuels display much more consistent and lower SO_2 and NO_X emission rates, with the gas fuel having close to zero SO_2 emissions.

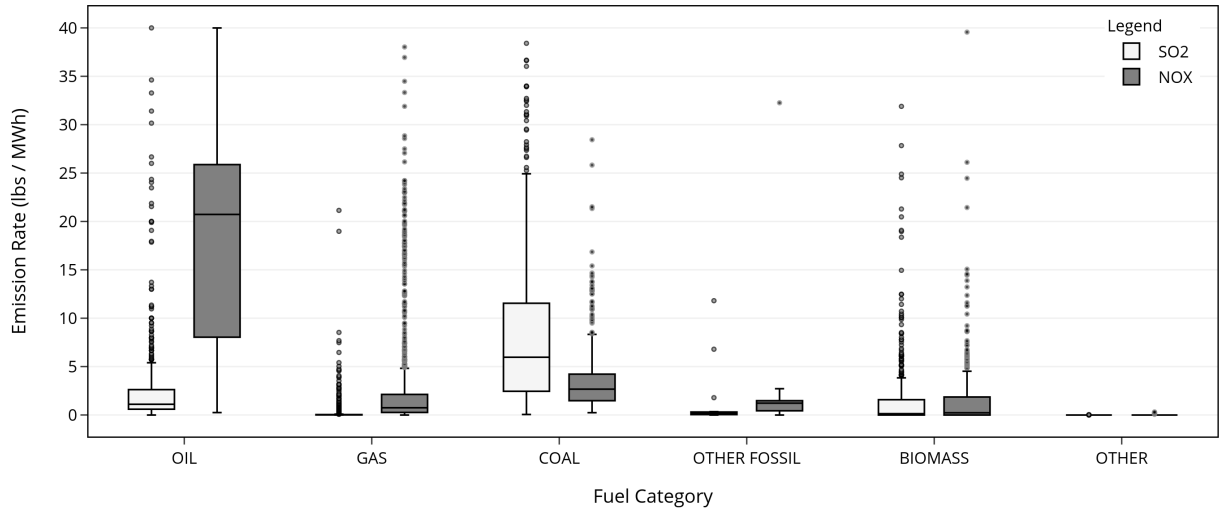

Figure 2: *SO_2 and NO_X emission rates for combustion fuel categories*

Based on the above analysis, relatively low CO_2 emission rate as well as insignificant SO_2 and NO_X emission rates, gas fuel constitutes a much more sustainable alternative to coal and oil fuels. Similarly, biomass fuel generally outperforms other fuels, including gas, in the pollutant emission rates. Thus, it is important for EPA to concentrate on transferring electricity generation from coal and oil plants to gas and biomass, to the extent possible.

In order to identify opportunities for expanding and limiting certain technologies, it is necessary to look at the utilization rate of plants using each of the fuel categories, illustrated on Figure 3. This plot shows that, similarly to the previous metrics, there is a large variability in the plant capacity factor. However, median of the distribution for each fuel category shows that the majority of coal and biomass plants are used to around 60% capacity, while gas plants are generally significantly underused, on average having utilization of 10-20%. As a result, this creates a great potential for expansion of biomass and, in particular, gas fueled plants, making them the primary options for outsourcing electricity generation from the coal plants. On the other hand, as previously assumed,

oil plants are generally not utilized to a great capacity with the plant capacity factor being well below 5% for the majority of cases, suggesting that this fuel category should not be a source of concern for the EPA. Hence, EPA should focus on allocating more resources towards the decommissioning or transformation of coal plants to prioritize utilization of gas and biomass fuels.

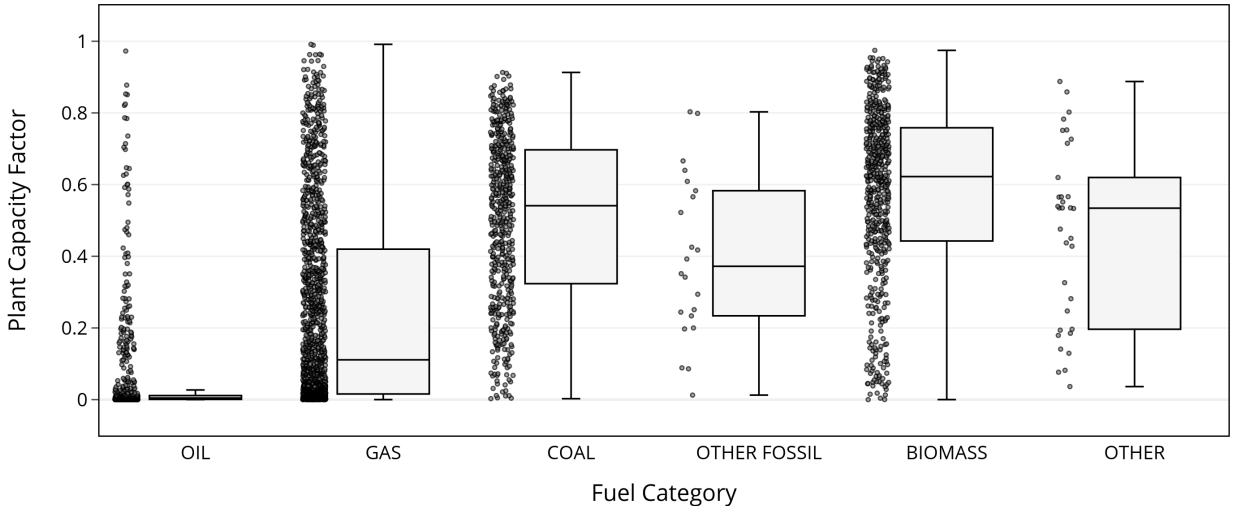

Figure 3: *Plant capacity factor box plot for each combustion fuel category*

Next, as a part of the exploratory data analysis, emissions were related to the plant characteristics. In particular, strong correlation was found between the efficiency of a plant, represented with heat rate, and CO_2 emissions. As shown on Figure 4, there is a strong positive correlation and an almost perfect clear linear relationship between the CO_2 emission rate and heat rate for oil, gas, and coal fuels, showing that the efficiency characteristic of these plants can be solely used to predict CO_2 emission rates. Moreover, the gradient of the line visibly varies between fuels, confirming previously identified differences in emission rates.

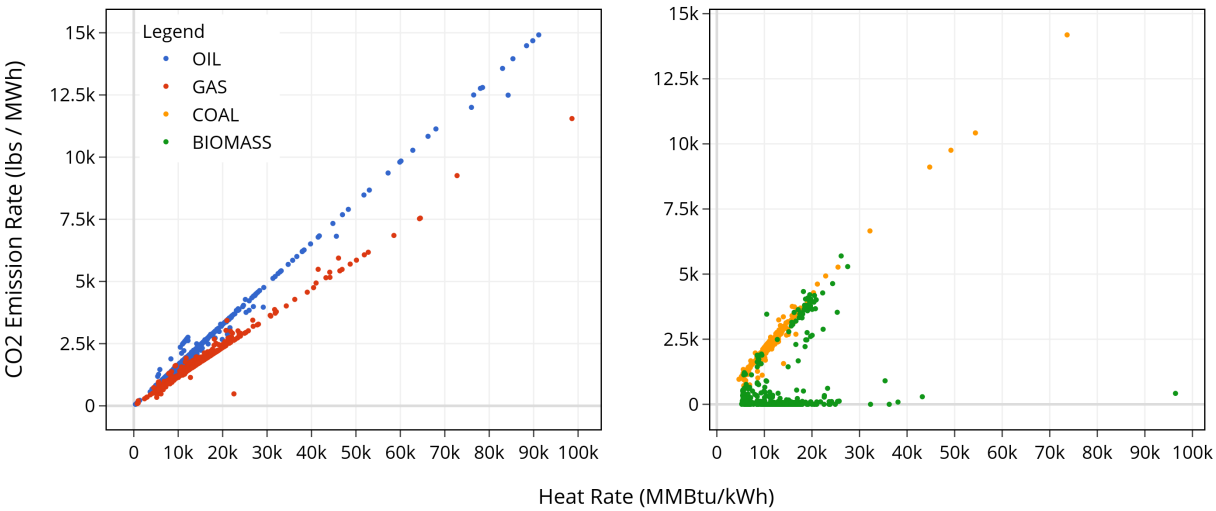

Figure 4: *CO_2 emission rate vs plant's heat rate scatter (0.8% of data constituting apparent outliers are ignored for the purpose of clarity of this plot)*

On the other hand, biomass appears to be split into two clusters: one displays a weak positive correlation, while the other is uncorrelated with the heat rate. This suggests that there may another variable, not included in the dataset, that affects CO_2 of the biomass plants and further exploration has to be performed to understand emissions from the biomass fuels.

Similarly, as illustrated on Figure 5, there is less correlation between the NO_X emission rates and the plant's heat rate, with the only exception being oil plants. It may also be observed that gas emission rates appear to have multiple weak linear trends, implying a potential third variable that would help to fully characterize NO_X emission rates by the heat rate. Conversely, coal and biomass plants show no clear correlation with this plant's characteristic, meaning that, as previously noted, more features are required to be able to describe plant's emissions accurately.

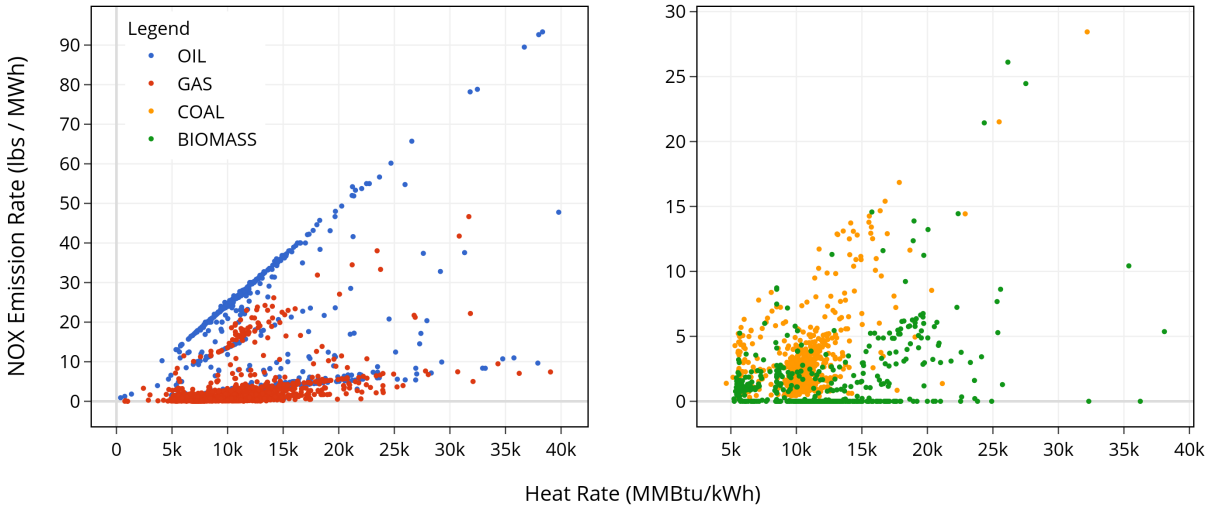

Figure 5: *NO_X emission rate vs plant's heat rate scatter (2.3% of data constituting apparent outliers are ignored for the purpose of clarity of this plot)*

These findings show that efficiency of the plant is an important characteristic that should be considered when allocating funds to and regulating various technologies. In particular, in cases where outsourcing electricity generation from coal and oil plants to more sustainable alternatives is not possible, investing into the improvement of the efficiency of used technologies and machinery at the plant can have a significant effect on its CO_2 and NO_X emissions.

4 Conclusion

In conclusion, with the performed exploratory data analysis, it was shown that coal is both the most utilized and the least efficient combustion fuel category, requiring immediate action. It was also identified that gas is a more sustainable alternative, while gas plants are generally considerably underused with a large potential for increasing utilization rates. Thus, one of the EPA's objective is to allocate more funds to outsourcing electricity generation from coal to gas and biomass plants. Finally, a strong positive correlation between plant's emissions and heat rate, suggesting that EPA should regulate and invest in improving the efficiency of the operating plants to reduce emissions.