The Global Power Plant Database Project

**Introduction**  
The database of power plant is a very large list that contains all details from the three main countries involved with power plants, including the USA, Australia (AUS), and India. The database included facilities in detail of a power plant such as power generation capacity, fuel type, geographical location, and data on generation. This dataset is essential to monitor the energy production by the world because this gives insights into fuel and distribution and predicts the future trends of energy.  
  
1. Problem *Definition*

The two major questions of the Global Power Plant Database are:

1. Primary Fuel Type Prediction From a given plant location, capacity, and commissioning year, can we predict the primary fuel type of a power plant?

2. Power Plant Capacity With the attributes of the plant, such as whether it uses coal, what area or country, in what year it was commissioned, among other attributes available in the dataset, can one determine the electrical megawatt capacity of a power plant?

It can also be utilized for more detailed analyses in the energy infrastructure, facilitates future on energy projects, and sustainable policies concerning energy.

2. Data Analysis

Summary of Data  
The dataset has the following key attributes:  
•    Country: ISO 3166-1 alpha-3 code, e.g. USA, AUS, IND  
•    Country Long: Full country name  
•    Name: Name of the power plant  
•    GPPD IDNR: ID number of the power plant  
•    Capacity MW: Power installed in Megawatts  
•    Latitude: Geolocation (latitude) decimal degrees  
•    Longitude: Geolocation (longitude) decimal degrees  
•    Principal Fuel: Type of fuel that the power station uses as primary fuel to generate electricity  
•    Auxiliary Fuels: Type of fuel(s) which are also used  
•  Year of commissioning: Year the power plant was commissioned  
•  Owner: Primary owner of the plant  
•  Source: Source of the data reported  
•  URL: web link of the source  
•  Generation Data: Volume of electricity produced monthly and yearly from 2013 to 2019.  
  
**Exploration of Preliminary Data**

The number, quality, and general features of the dataset would be explored. Therefore, there is a need to; Identify missing values, get insights into the distribution of the data as well as its relationships.

**General Findings**

• Missing Values: There are only a few columns having missing values in the dataset such as state/province and capacity.  
• Fuel Types: Diversified fuel types have always appeared in the dataset including coal, natural gas, and renewables.  
• Capacity Range: The value for the capacity ranges from small plants to larger facilities with an incredibly wide range.  
•Generation Data: Both reported and estimated data are available regarding generation; however, the number of details varies.

3.EDA Concluding Remarks

Result Summary  
•Distribution Fuel Type: Coal and natural gas make up much of the total primary sources globally, but there is growing renewal source capacity now available like wind and solar.  
•Capacity by Fuel Type: Renewable energy-based power plants have much smaller capacities that only reflect their short history as well as the relatively immature technology.  
•Geographical Distribution: Power plants are so unevenly distributed with such a huge concentration in North America, Europe, and parts of Asia. The geographical distribution reflects regionalization of energy demand and infrastructural development.

**Visualizations**

Visualizations were designed to accompany our arguments:  
•Fuel Type Distribution Pie chart showing percentage composition of each fuel consumed across the world  
•Capacity by Region Bar chart showing cumulative capacity of all continents  
•Generational Trends Line graphs of the trends in the patterns of change in power generation capacity over time

4. Preprocessing Pipeline

Handling Missing Values  
Missing values were imputed into the pipeline one at a time or a blend using the following  
• Capacity: The missing values were replaced using the median capacity of similar characteristic features of plants, such as same type of fuel

**Data Transformation**

• Encoding of categorical variables like one-hot encoding, e.g.: fuel type, country, etc.  
Normalization: capacity column normalized with the uniform scaling of all the data

**Data Splitting**

The data was split into a training set and the test set where the model would be trained and its correctness would be tested over it

5. Developing Machine Learning Models

***Model Selection***

The two models developed below was to predict:  
1. Primary Fuel Type Prediction:  
 Logistic Regression: In case of multi-class classification, it is mostly predicting the primary fuel type  
 Random Forest Classifier: An ensemble model primarily meant for better data and interactions' categorization.  
2. Capacity prediction:  
 Linear Regression: Simple model used to predict continuous values  
 Random Forest Regressor: This is an ensemble method that captures much more complex relationships in features and capacity

**Models' Training and evaluation**

Primary Fuel Type Prediction:

Capacity Prediction:  
•Fuel Type Classification: Using the Random Forest Classifier, one can classify power plant fuel types with good precision and classification metrics.  
•Capacity Classification: The Random Forest Regressor is pretty good at predicting capacity, but the acceptable mean squared error presented nice precision.

6. Conclusion

The outcome of the analysis of the Global Power Plant Database showed some interesting results:  
• Fuel Source Trends: Renewable energy sources hold a majority but are in a rising trend. Thus, there indicates a 'step trend' toward a more sustainable energy source.  
• Capacity Distribution: The fossil fuel plants are generally coupled with higher capacities whereas renewals at the hands of smaller capacities suggest the recent rise and technological updating of renewals.  
• Geographical Insights: The power plant maps also denote the regions that are in the process of developing energy infrastructure, and those lagging in the development.

**Future Work**

There is more work along these lines:  
• More Data: Include even more granular data on environmental and operational data for deeper analysis  
• Real-time analysis: Include real-time functionality to conduct an analysis of the data for insights in real-time and insightful decision-making process