**Global Power Plant Data Analysis Using Python**

**using pandas, matplotlib, and seaborn**

**Introduction:**

The Global Power Plant Database is a comprehensive, open-source database of power plants around the world. It centralizes power plant data to make it easier to navigate, compare and draw insights for one’s own analysis. The database covers approximately 35,000 power plants from 167 countries and includes thermal plants (e.g. coal, gas, oil, nuclear, biomass, waste, geothermal) and renewables (e.g. hydro, wind, solar). Each power plant is geolocated, and entries contain plant capacity, generation, ownership, and fuel type information.

**Making two predictions**:

1. primary fuel (energy source used in primary electricity generation or export)
2. capacity megawatts(electrical generating capacity in megawatts)

Here I'm using a python library like **pandas**(a software library written for the Python programming language for data manipulation and analysis), **Matplotlib**(a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy), **Seaborn**(a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions)

Here’s the step-by-step outline of this project:

1. Defining the question
2. Collecting the data
3. Cleaning the data
4. Analysing the data
5. Visualizing data and sharing findings

**Defining the question**

The Global Power Plant Database is a comprehensive, open-source database of power plants around the world. It centralizes power plant data to make it easier to navigate, compare and draw insights for one’s own analysis.

Making two predictions:

1. primary fuel (energy source used in primary electricity generation or export)
2. capacity megawatts(electrical generating capacity in megawatts)

We have the database covers approximately 35,000 power plants from 167 countries and include thermal plants (e.g. coal, gas, oil, nuclear, biomass, waste, geothermal) and renewables (e.g. hydro, wind, solar).

### Collecting the data

I collect this data from GitHub repository of World Resources Institute

WRI is a global nonprofit organization that works with leaders in government, business and civil society to research, design, and carry out practical solutions that simultaneously improve people’s lives and ensure nature can thrive.

GitHub link: [global-power-plant-database/database\_IND.csv at master · wri/global-power-plant-database · GitHub](https://github.com/wri/global-power-plant-database/blob/master/source_databases_csv/database_IND.csv)

This data covers approximately 35,000 power plants from 167 countries and includes thermal plants (e.g. coal, gas, oil, nuclear, biomass, waste, geothermal) and renewables (e.g. hydro, wind, solar).

**Cleaning the data**

After collecting a data, the next step is to get it ready for analysis. This means cleaning, or ‘scrubbing’ it, and is crucial in making sure that you’re working with high quality data.

* Removing unwanted data points — extracting irrelevant observations that have no bearing on intended analysis.
* Bringing structure to data — fixing layout issues, which will help to map and manipulate this data more easily.
* Filling in major gaps — This data contains null values and I notice that important data are missing. Once we identified gaps, we can go about filling them.

During cleaning the data we have to do EDA(Exploratory Data Analysis (EDA) is an approach to analyse the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations). This helps identify initial trends and characteristics, and can even refine our hypothesis.

**Analysing the data**

We are doing Predictive analysis (Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning that analyse current and historical facts to make predictions about future or otherwise unknown events)

Making two predictions:

1. primary fuel (energy source used in primary electricity generation or export)
2. capacity megawatts(electrical generating capacity in megawatts)

First I describe this data set by using describe function we get to know that:

* The mean is more than the median in capacity mw.
* There is a large difference between the 75% percentile and max in
* capacity MW.
* These observations suggest that there are extreme outliers in these columns.
* The mean value is higher than 50% which shows there is a little skewness in present capacity MW.

-After this, I checked the correlation in the data set and also plot the heat map to visualize the correlation in between the datasets

-Observing box plot we can see the outliers present in the boxplot of electrical generating capacity in megawatts.

-After plotting distplot we can see the lump of a data at start and a long tail follows lump of data this is positive skew (Right-Skewed distribution)

-I plotted bar plot and observe that which year the high electricity generation in gigawatt-hours reported. We can observe that the highest electricity generation in 2015 and 2016.

I compare all the electricity generation in gigawatt-hours reported in a year with capacity mw by plotting the regression graph

### Visualizing data and Concluding Remarks

-We can observe that the from the count plot of primary fuel   
(energy source used in primary electricity generation or export),   
most of the energy source used in primary electricity generation is coal energy, hydro fuel energy and solar energy.

-We can predict based on these observations that there is a shortage of coal energy, because of excessive use.

- Also, we can observe most of the coal energy is used for electrical generating capacity in megawatts and other energies are hydro fuel energy, solar energy, and wind energy.

**Summary**

I work on Global Power Plant Data Analysis Using Python using pandas,matplotlib, and seaborn.

1. I define the questions to work on them and give a prediction
2. I collect all the data required for this project from World Resources Institute GitHub repository. (GitHub link: [global-power-plant-database/database\_IND.csv at master · wri/global-power-plant-database · GitHub](https://github.com/wri/global-power-plant-database/blob/master/source_databases_csv/database_IND.csv))
3. I loaded this data set in a python notebook and firstly started Cleaning the data and removed the unisensory null values.
4. After removing all unwanted data I start to Analyze the data by using different python library(pandas,matplotlib, and seaborn)
5. At last, I predict the results by Visualizing data sets.