**Global Power Plant Project**

**Introduction**

Generation of electricity is crucial for a country’s economic and social development. There are many types of power plants that produce electrical energy based on the energy source from which it is harnessed such as the from fossil fuels that includes coal, oil, and natural gas. Low carbon sources as the nuclear power, and finally renewable resources such as solar, wind and hydro.

It is essential that sufficient energy is produced by exploiting these available sources of energy to meet the growing demands of a country. However, production of electricity involves expense of the equipment, infrastructure, transportation, and distribution cost. Additionally, with rising concerns over the depletion of energy resources, it is vital to ensure its sustainable use and also tap into renewable resources.

**Capacity**

Capacity is the amount of electricity a generator produces when it is operating at its maximal performance which is measured in megawatts. However, capacity is not synonymous to generation of electricity, rather it tells us how much load it is able to handle. Capacity of a power plant varies considerably based on the fuel type that is being used.

**Primary Fuel:**

Primary fuels are the energy resources which are extracted and converted into electricity. Coal, oil, natural gas which are exhaustive resources essentially covers the major demands. But it has several negative impacts such as production of carbon dioxide that is the largest contributor of global climate change. Hence renewable resources such as nuclear, solar, wind and hydro energy are also being tapped in to help compensate the energy requirements.

**Problem Statement:**

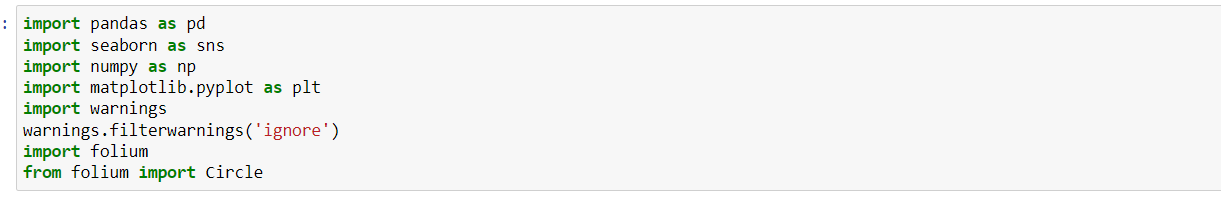
With the ever-growing demands of electricity and the rising concerns over depletion of non-renewable resources combined with the pollution caused by the fossil fuels, it is important that we come up with a solution to generate optimal capacity that is economically feasible and sustainable.

With the help of machine learning this is an achievable feat. Accurate prediction of capacity using a machine learning model will help decide if the demand supply ratio of electricity will be met. Accordingly, the professionals can take decisions regarding the investment that are to be made to support the expenses. A predictive machine learning model for primary fuels can help strategize ways in which fuels can be utilized such that it is sustainable, economically feasible and meet the energy demand of a country.

In this project, we have the dataset containing the features of different power plant based in India. We are supposed to solve a regression model that will help predict the capacity and a classification model that will predict the primary fuel used.

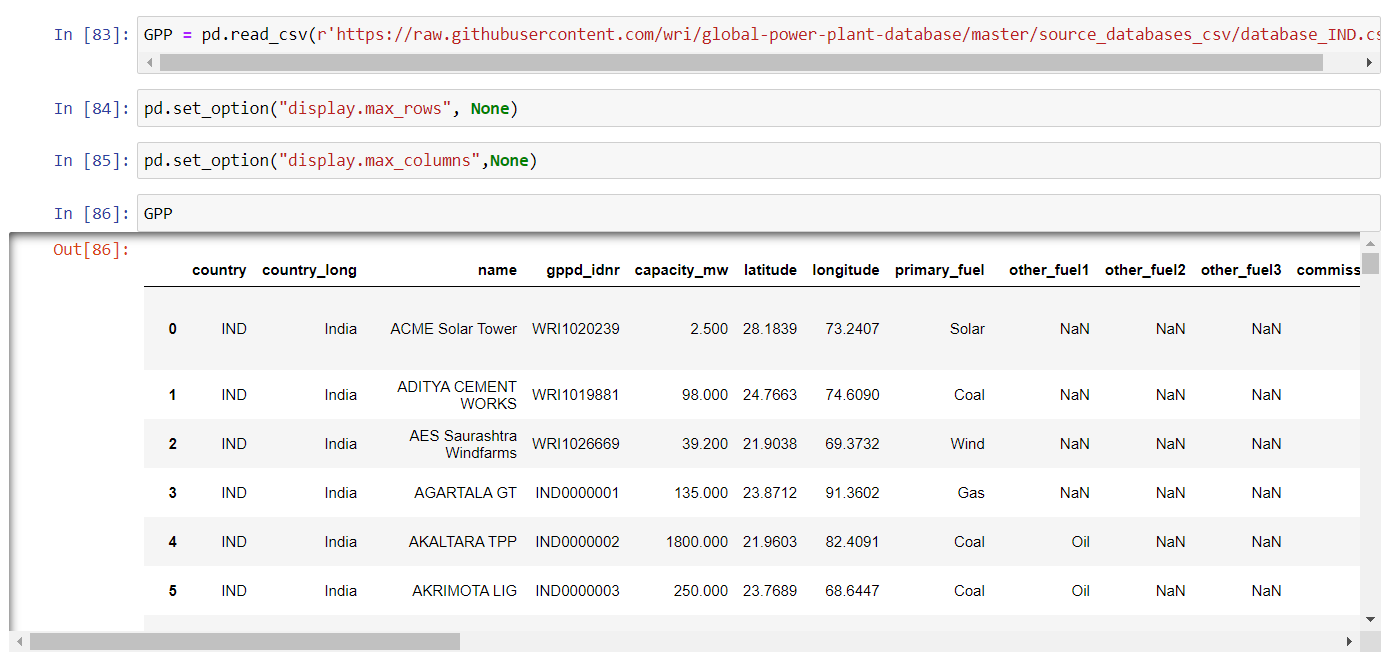
**Data Analysis**

We first need to import all necessary libraries required for data analysis



After importing all the necessary libraries that is requires for the analysis of the data, initial analysis was conducted to get a better idea of the data.

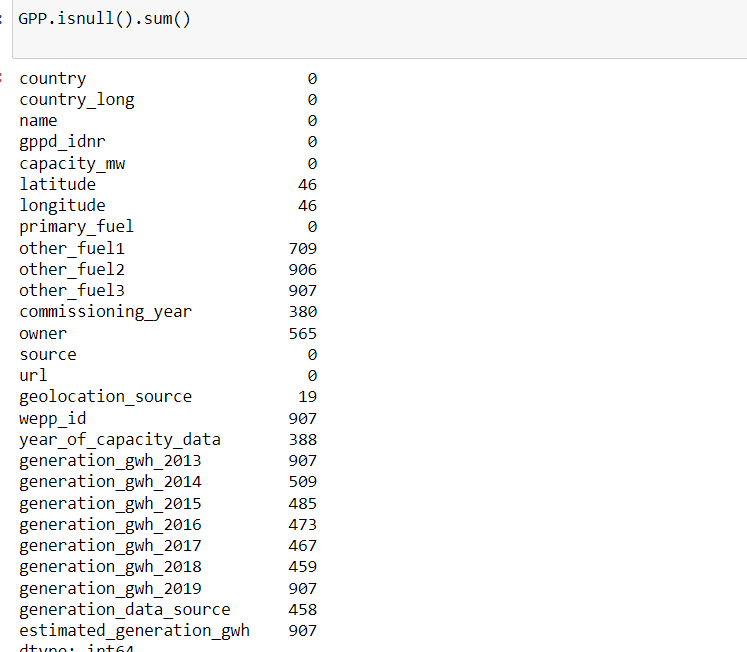
We first loaded the dataset with the help of pandas. Since I wanted to view the entire dataset, I changed pd.set\_options to None.



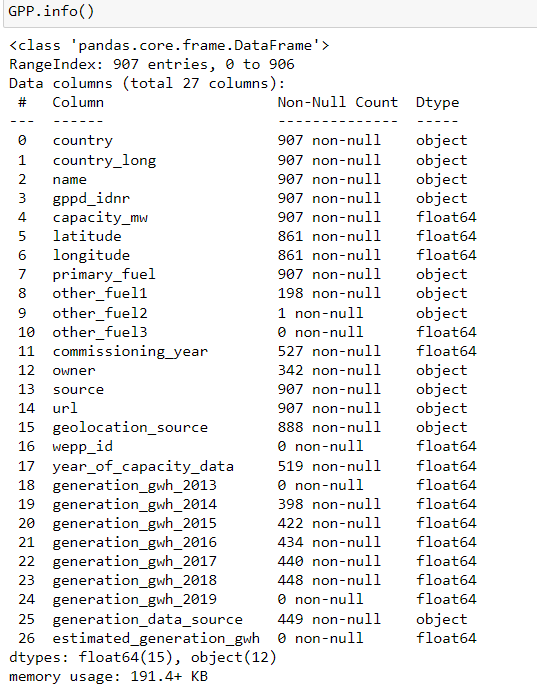
**Understanding the data**

****

There are 907 rows and 27 columns in the dataset.



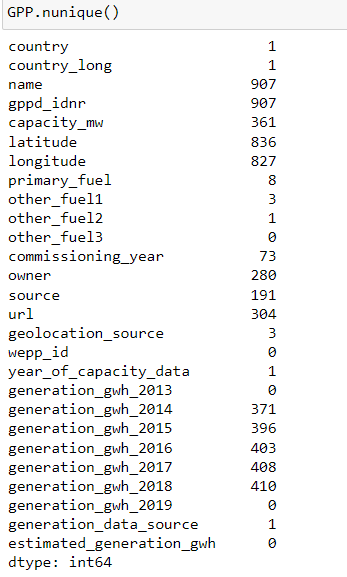
There were a lot of null values present in most of the columns in the data which needed to be handled. wepp\_id, estimated\_generation\_gwh, generation\_gwh\_2019, generation\_gwh\_2013, other\_fuel3 are fully null values. Other\_ fuel2 has only 1 non null value. These can be dropped. But before doing that, unnecessary column is required to be filtered out based on further analyses.

Further analyses were performed to check the dtype of each column

There were three types of data:

* object,
* float64

Object dtype will be encoded later on while feeding into the machine learning models as string type data cannot be interpreted by the models.



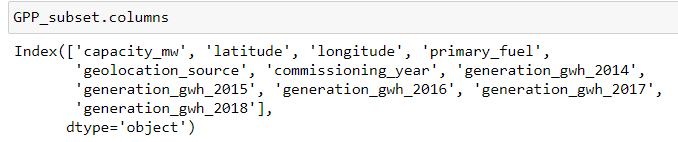
The number of unique values were checked. Three columns country, country\_long, year\_of\_capacity and generation data source contained only one value. Since it contains no variation, this data is not useful for prediction. name, gppd\_idnr contain all unique values which also does not help the analyses.

Also, from this analysis, we can assume that we only have the data from a single country which was found to be India.

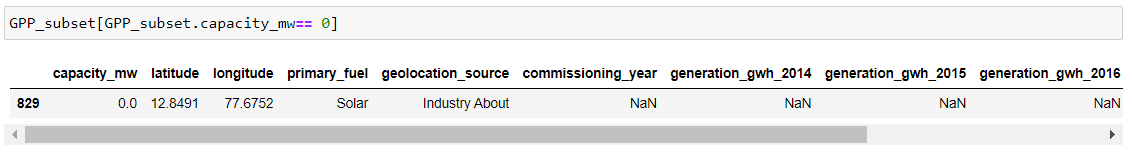


**Data Cleaning:**

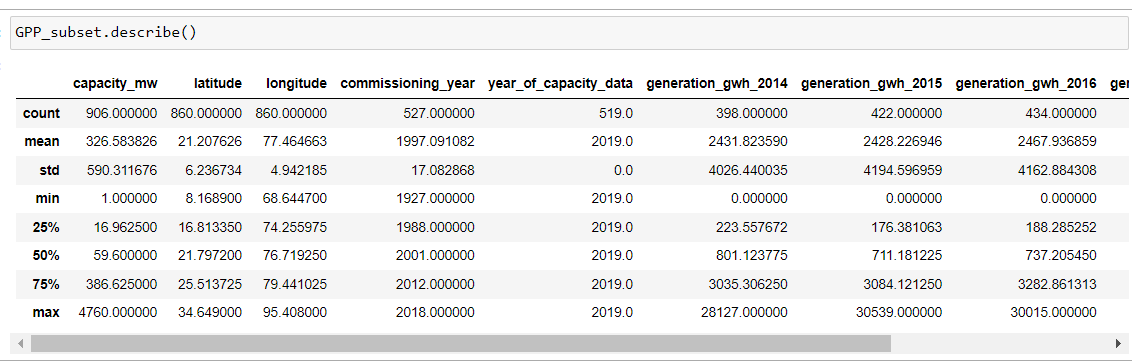
The columns unnecessary columns that we had found during the analyses were all filtered out giving us the remaining columns



A row was found to have 0 capacity, probably because it is non-functioning, hence this row was dropped.



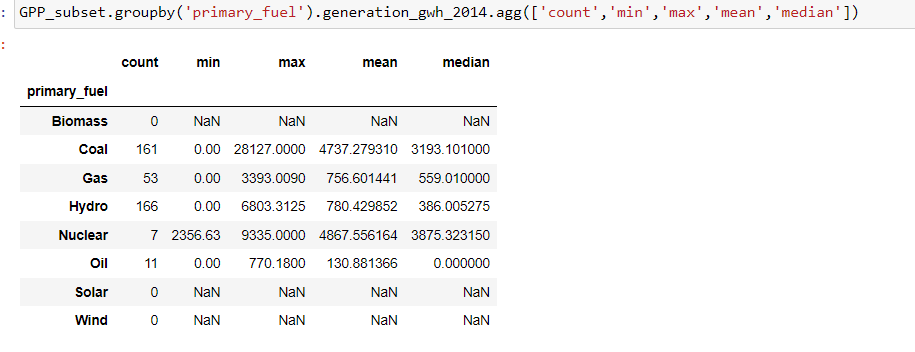
Missing values in the remaining columns were to be handled carefully. First the summary of the data was checked.

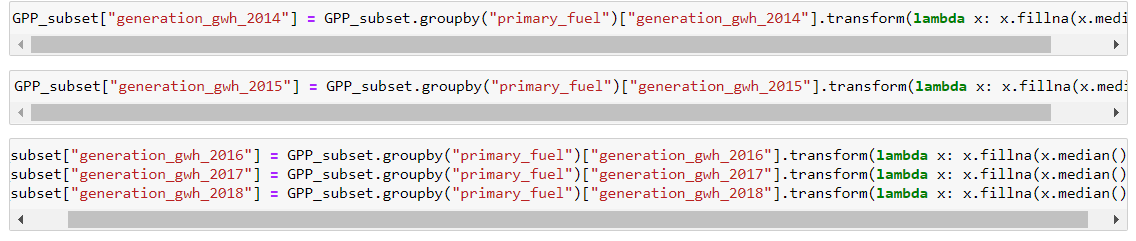


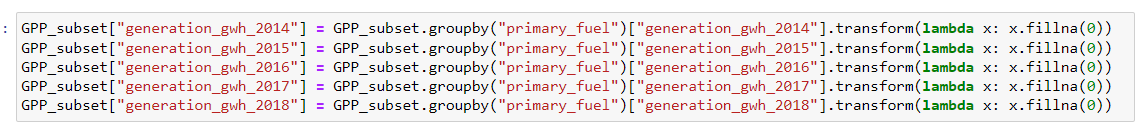
We can see most of the columns are skewed since the mean is greater than median.

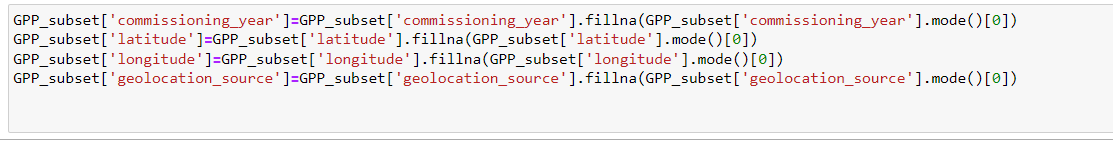
There are a lot of outliers in some of the columns due to the huge difference in 75% quantile and max value.

Groupby function was then used to see the generation of electricity in the year 2014,2015,2016, 2017 and 2018.



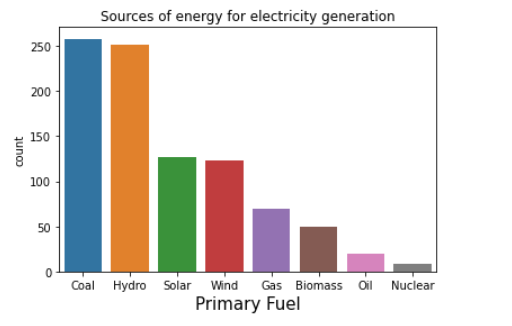
We see that for Biomass, Solar and Wind there are no available data. From the mean values it can be inferred that, the values differ greatly in all columns. Also, the mean is greater than the median for all fuels. This trend is seen in all other years. Based on this analyses, missing values were replaced by the median of electricity generated by each fuel type. However, since Biomass, Solar and Wind contained all null values, Nan was replaced by 0.

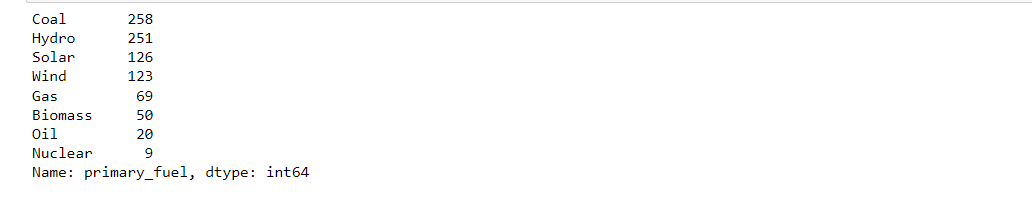


Rest of the missing values in columns which were categorical were replaced by the mode value.

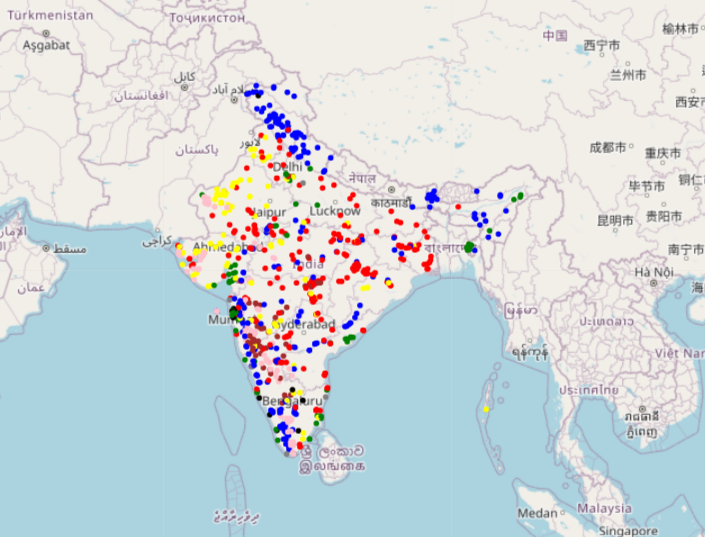
The dataset was now free of null values.

**Data Visualization**



 Primary fuels consisted of Coal, hydro, solar, wind, gas, biomass, oil and nuclear. It can be seen that coal and hydro are the fuels run by in majority of the power plant where as nuclear power plants are the least.

This column will be the label for the Classification model. Hence it will be required to be balanced.

A folium map was generated to visualize the distribution of the powerplants India.

Red: Coal

Blue: Hydro

Yellow: Solar

Green: Gas

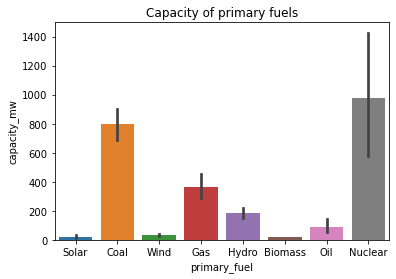
Biomass: Brown

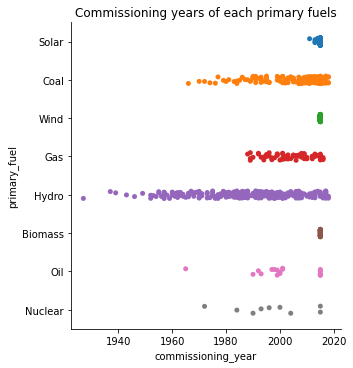
Pink: Wind

Black: Oil

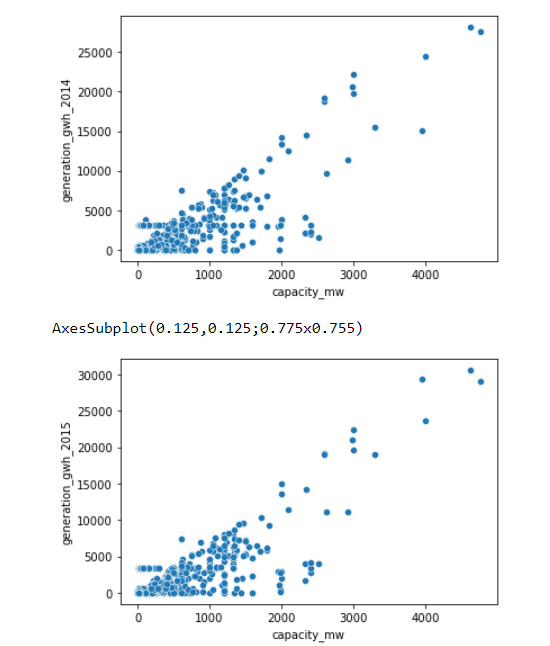
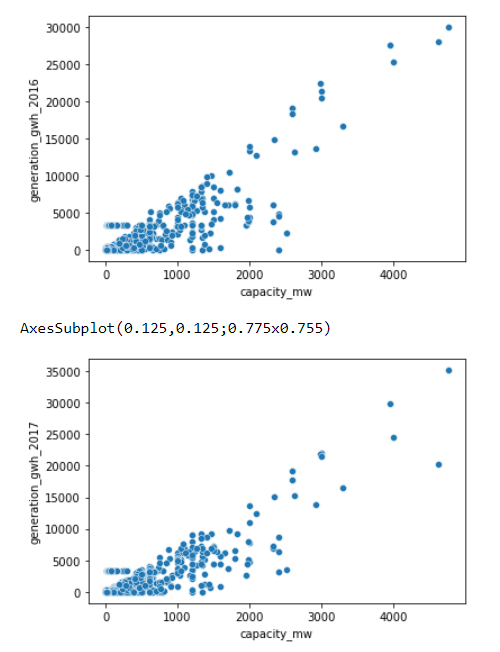
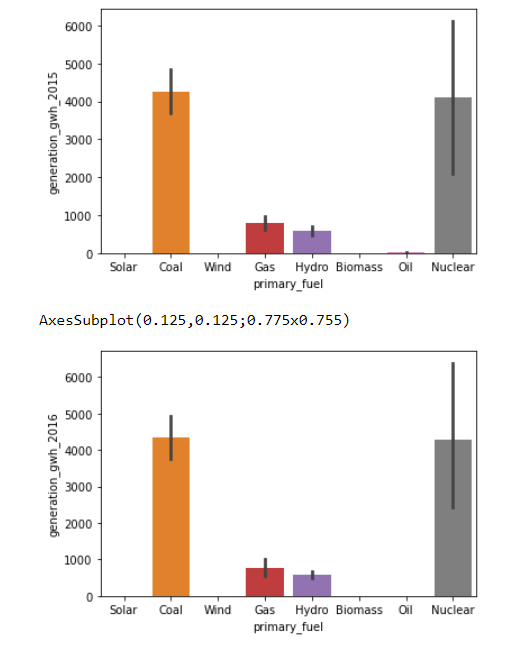
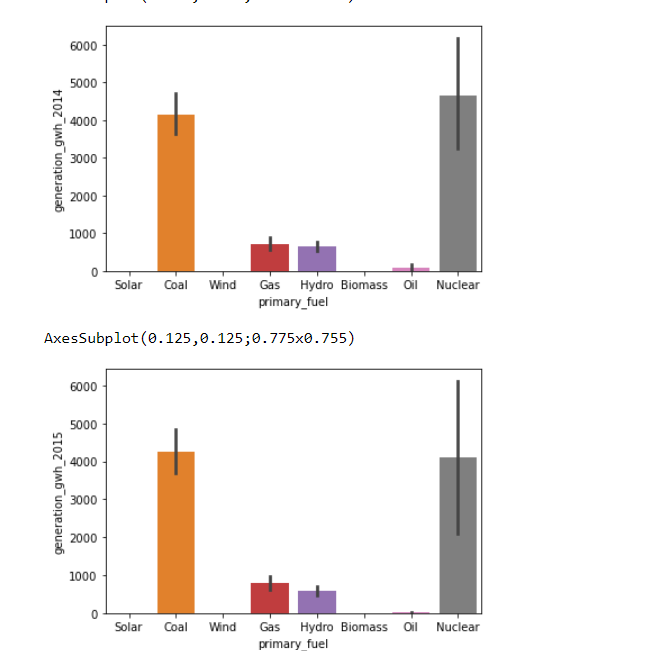
Gray: Nuclear

Here we can see which power plants are found predominantly is each region. For eg we can see solar power plants more in the west side of India. Wind power powerplants are seen more in the South-west side of India. Oil based power plant can be found to be located somewhere near Maharashtra.

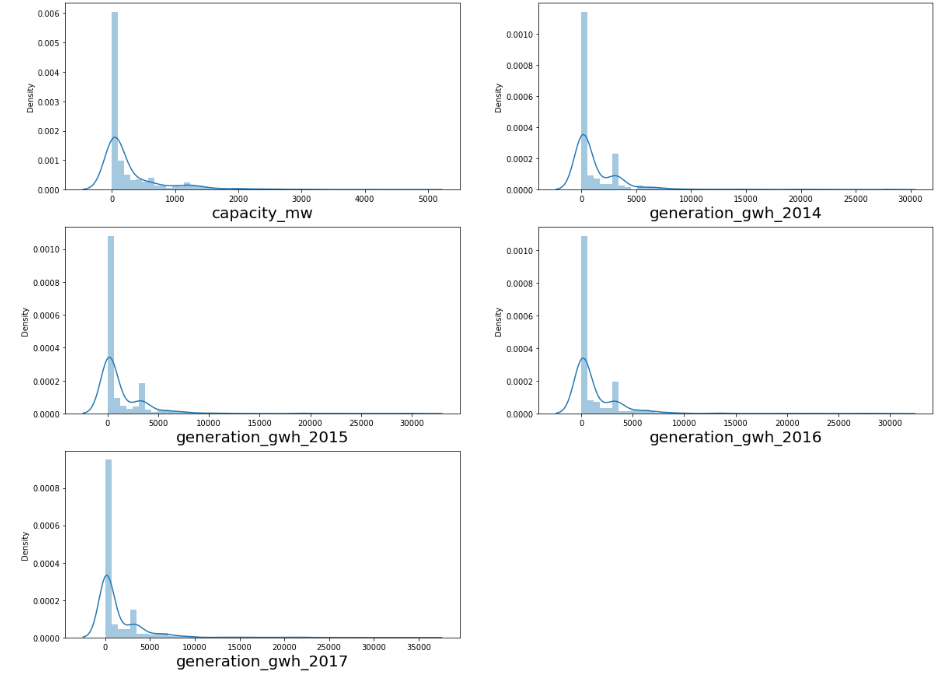


We see that nuclear energy has the highest capacity followed by coal, gas, hydro and oil. solar wind and biomass are found to have least capacity.

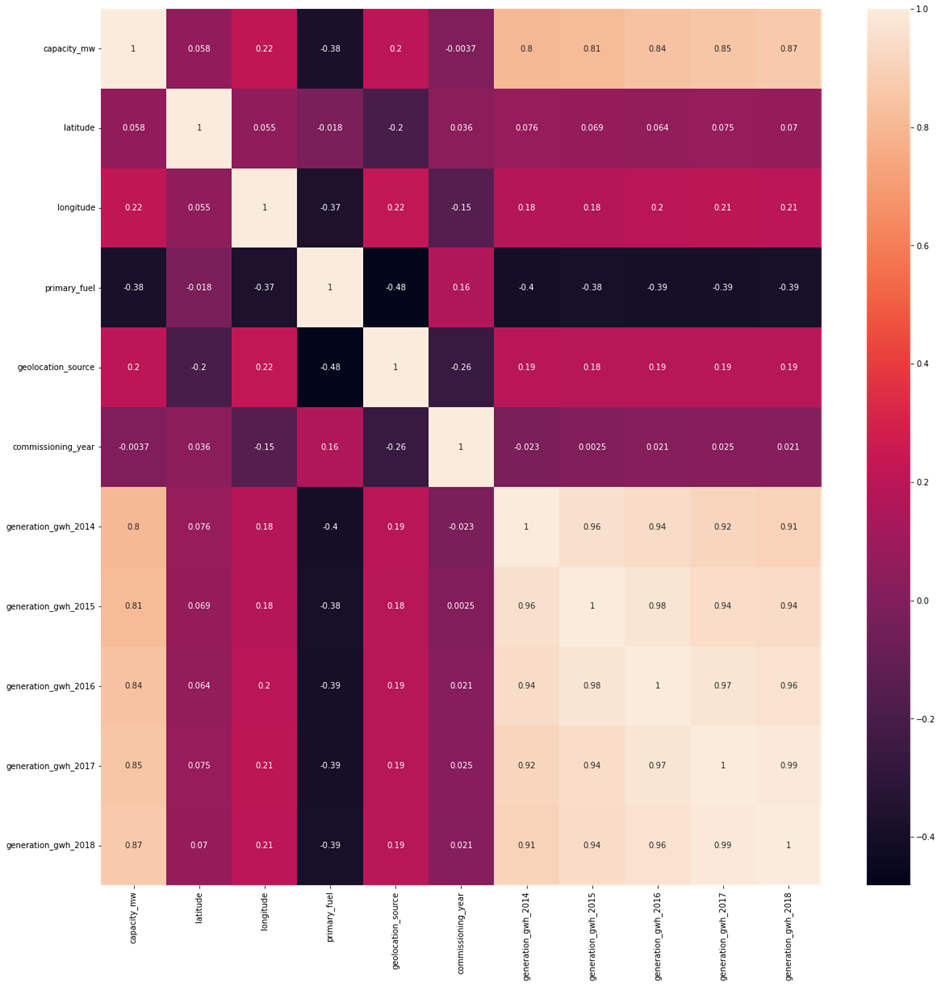
Authorisation of hydro plants dated back to 1940s while the more recent ones are solar, wind and biomass.

We can observe that all throughout the years Nuclear and Coal are generating more electricity as in comparison to other fuels.

We can observe linear relationship between capacity and generation.



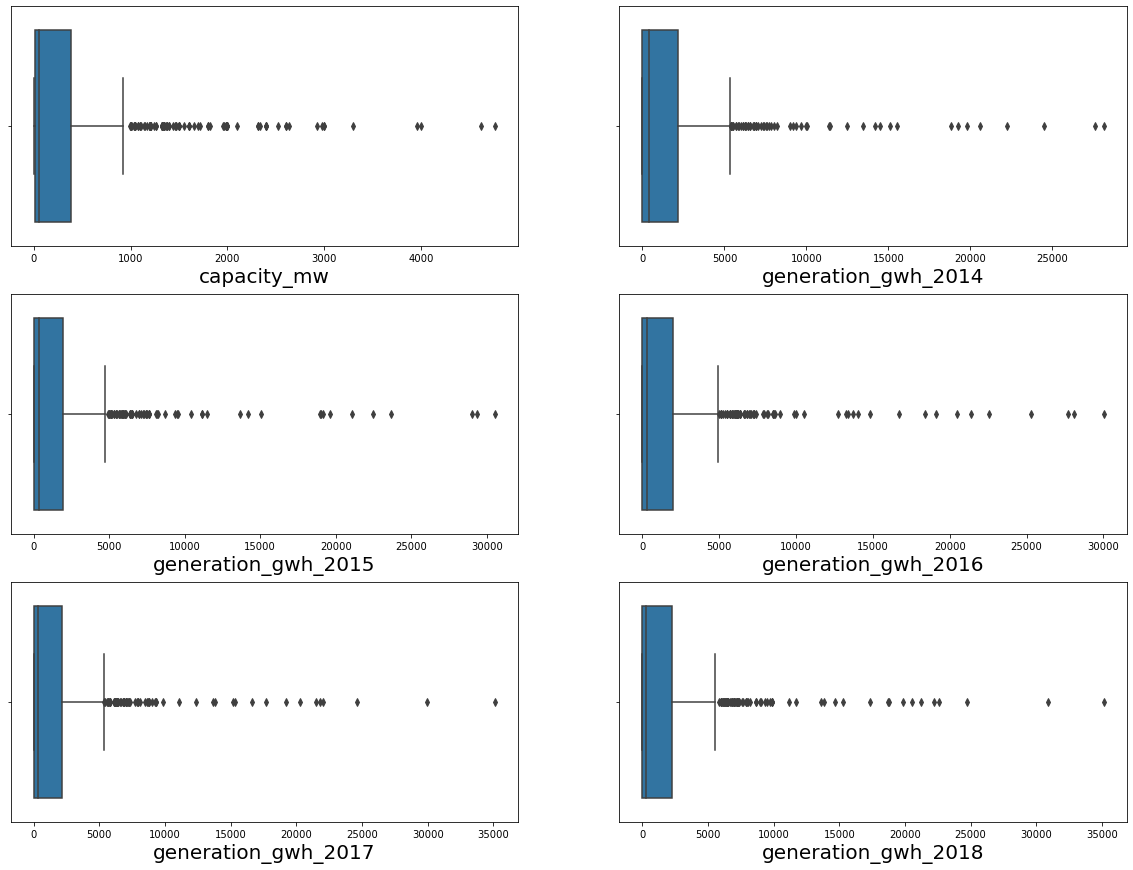
Distribution of numerical columns showed that all the data is highly skewed to the right.



We can visualise the correlation between features and label

- Between feature we can see generation in each year seem highly correlated with each other, which is expected.

- capacity\_mw is seen to be highly correlated with generation data of each year



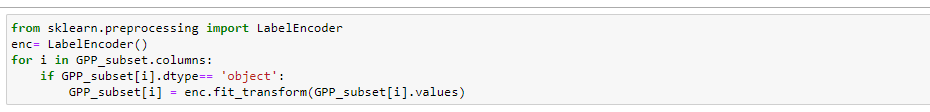
We followed it up with checking the outliers in all the numerical columns. However, these are not random errors but correct values.

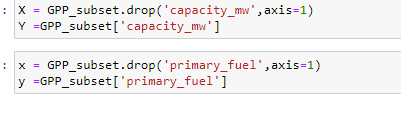
**EDA Concluding Remark**

During the EDA, many missing values were found in column. We conducted analysis to see how imputation can be done to clean the data. Further, unnecessary columns were found and dropped from the dataset. We proceeded to data visualization, where we found the relationship between features using scatterplot and bar plots. We found primary fuels generated maximum capacity and that capacity and electricity generation were linearly related. A map was also generated to visualize the distribution of powerplants throughout India. We also saw the correlation between the feature and label.

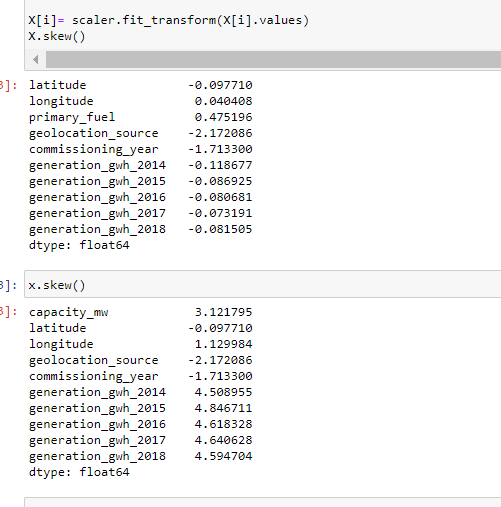
**Data Pre-processing Pipeline**

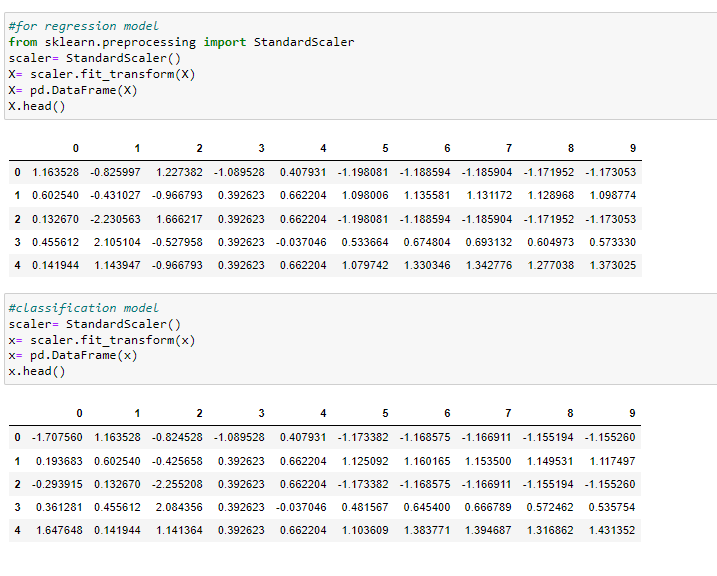
Data pre-processing is one of the most important steps where we process the data which is optimum to be interpreted and deduce predictions using the machine learning algorithms. The steps done in the current project are:

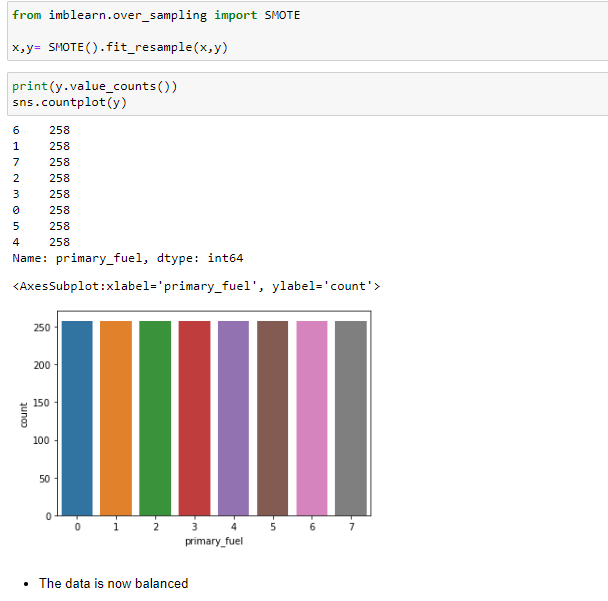
* Data cleaning was completed during the EDA.
* Object dtypes were encoded to numbers so that it can be interpreted by the machine learning model. Label encoding technique was used for encoding the data.
* Splitting data into X and Y for regression model where Y contained the label capacity\_mw
* Splitting data into x and y for classification model where y contained primary\_fuels



* Power transformation was used in both the data to remove skewness from necessary columns, this also ensures there is no data loss.



* Scaling of data was done using standard scaler so that no data bias is produced.
* Since primary\_fuels was imbalanced, oversampling technique was used for balancing it.



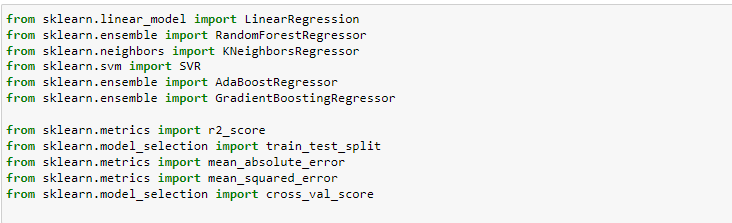
Now we are ready to build our Models.

**Model Building**

**Regression Model:**

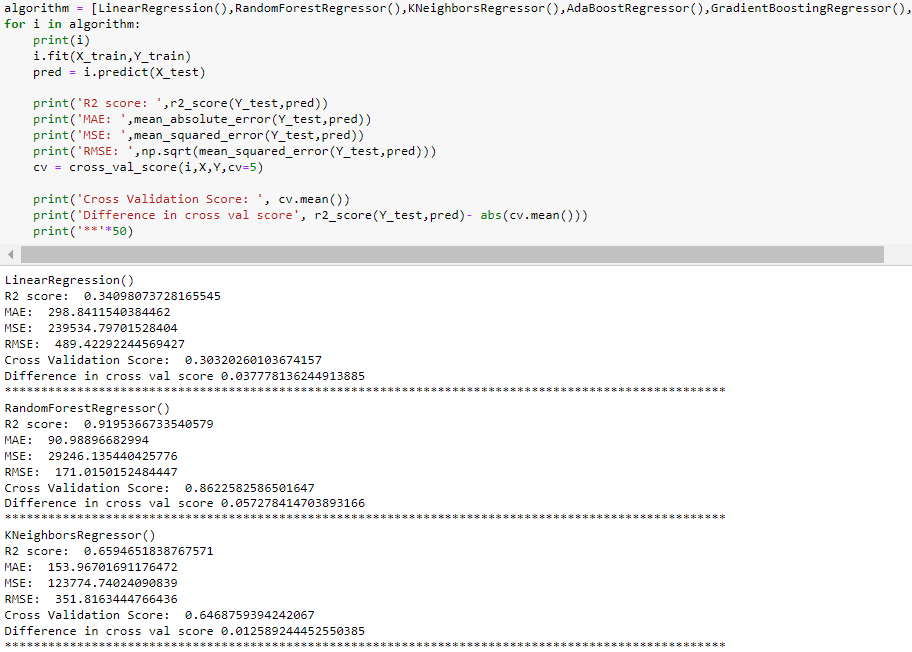
The data set was first split into train and test, where the models will first learn from the train data and will test the model’s efficiency using test data.

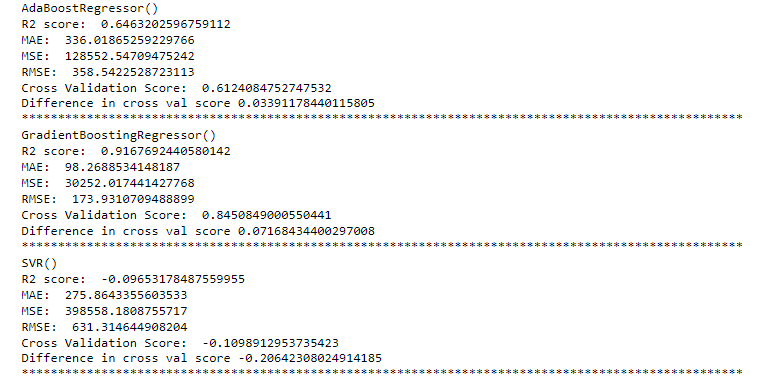
We then imported necessary algorithms and libraries which will test the r2 score, error, and cross validation score to see if the models are overfitting/underfitting.



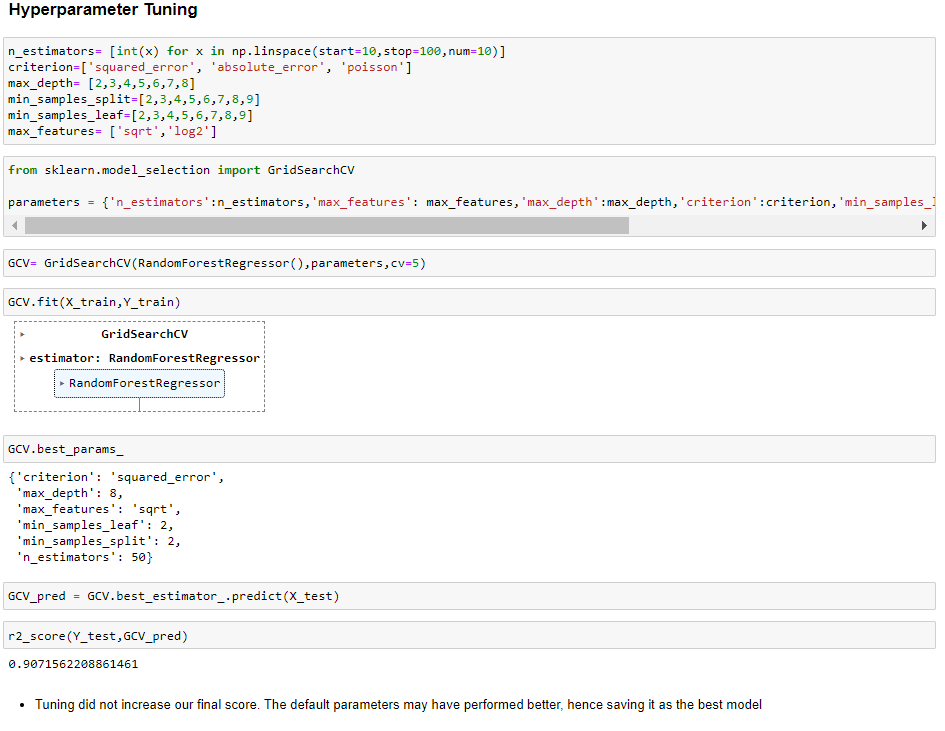
For Regression models, the best random state was found as 71.

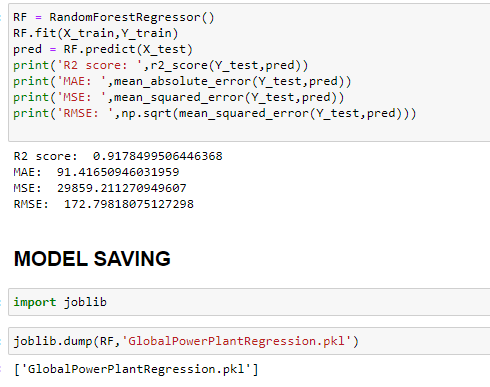
The dataset was tested using Linear Regression, RandomForestRegressor, KNeighborsRegressor, AdaBoostRegressor, GradientBoostingRegressor, and SVR. The r2 score was found to be best in RandomForestRegressor (91.95%) with a when cross validated score of 86.22%





Hyperparameter tuning was performed but it did not improve the score hence the earlier model was saved.





**Classification Model**

For classification model, the best random state was found as 121. We trained and tested the models using :

LogisticRegression(),

RandomForestClassifier(),

DecisionTreeClassifier(),

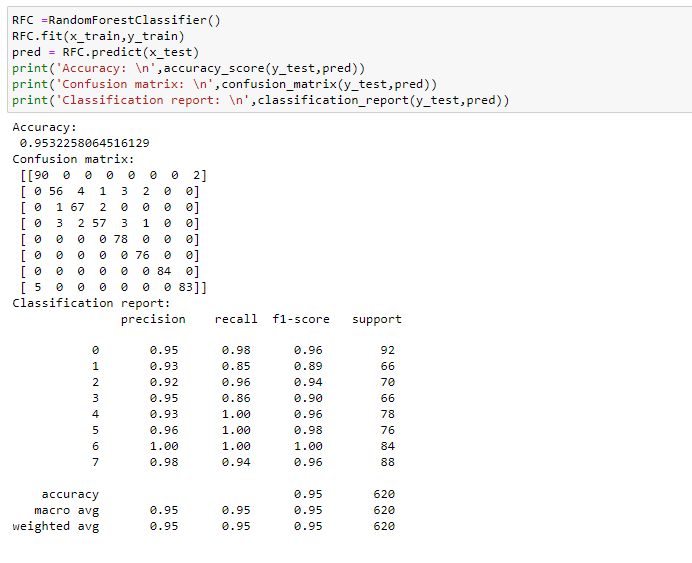
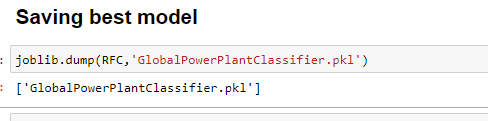
KNeighborsClassifier(),

AdaBoostClassifier(),

GradientBoostingClassifier(),

SVC()

RandomForestClassifier ()was found to have the best accuracy of 95.56 and cross validation score 93.80. This will be saved as our best model



**Conclusion**:

In this project, our objective was to build a regression machine learning model which predicted the capacity and a classification model that predicted the primary fuels used in the power plants. The most crucial part in the project was handling the missing values and replacing them with a more appropriate value. After conducting EDA and Data Pre-processing we proceeded to build a RandomForestRegressor model which predicted with an r2 score of 91.95%. A classification model was build using RandomForestClassifier() which predicted the primary fuels with an accuracy of 95.56%.

Using such predictive model will help analyse the capacity of the power plants and ensure demand supply ratio of electricity. It will also help estimate if the primary fuels being used is economically feasible and environmentally sustainable