**Analysis of Global Power Plant Database** **using Machine Learning**

(Article Written By: Trupti Badole)

****

Here we are going to do prediction of primary fuel & capacity of different power plants which are at different locations with the help of Machine Learning algorithms from Global Power Plant Database.

**Introduction**

Here we are going to do the complete analysis of Global Power Plant Database and will predict primary fuel & capacity of different power plants. We will cover all the aspects that we are going to use in ML model and projects we will also do the complete analysis using the Data visualization to model building and finding the main observations from the analysis which is going to help us a lot for prediction of the best results.

**We will cover:**

* *Part 1: Problem Definition*
* *Part 2: Data Analysis*
* *Part 3: EDA Concluding Remarks*
* *Part 4: Pre-Processing Pipeline*
* *Part 5: Building Machine Learning Models*
* *Part 6: Concluding Remarks*

1. ***Problem Definition***

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) from 2014 to 2018. Each power plant is geolocated and entries contain information on plant capacity, generation, ownership, and fuel type. It will be continuously updated as data becomes available.

1. ***Data Analysis***
2. We have copied the raw file from the GitHub link and created a csv file in our system and imported the data frame using pandas.



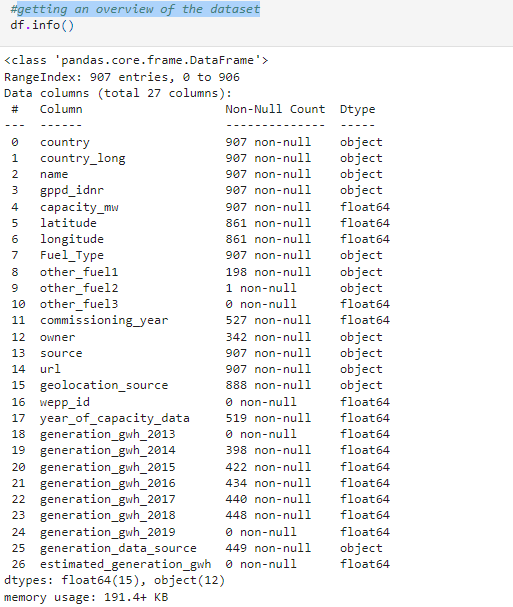
1. We can see that the dataset is comprised of 907 rows & 27 columns including two target variables capacity\_mw and primary\_fuel where we need to predict both capacity\_mw (Continuous Target Variable) and Fuel Type (Categorical Target Variable) on separate Regression and Classification Models.
2. We will define primary\_fuel as Fuel\_Type to understand the dataset in a better way.



1. ***EDA Concluding Remarks***

After loading the data, we will explore the data and will perform some basic analysis and visualizations.

* 1. First, we will get the overall overview of the dataset:



**Observation:** We can clearly see that we have null values and also object datatype, which we will handle separately by imputer and encoding techniques.

* 1. We will check data types of the dataset:



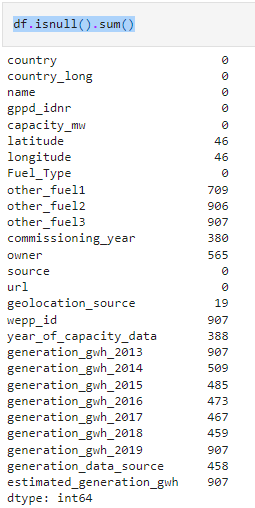
**Observation:** The features that needs encoding are country, country\_long, name, gppd\_idnr, Fuel\_Type, other\_fuel1, other\_fuel2, owner, source, url, geolocation\_source, generation\_data\_source as they are object data type and the ML model needs numeric datatype.

* 1. We will check number of unique values present in each column:



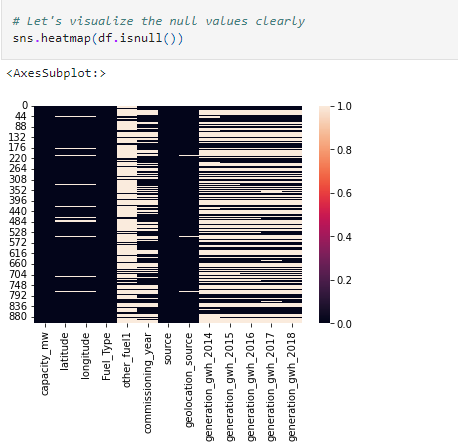
**Observation:** Here the columns country, country\_long, other\_fuel2, year\_of\_capacity\_data and generation\_data\_source have only one unique value. Also, other\_fuel3, wepp\_id, generation\_gwh\_2013, generation\_gwh\_2019, estimated\_generation\_gwh have no unique values which means they are filled with only NAN values. Since these columns have same entries throughout the dataset so we can drop these columns.

* 1. We will check if our dataset contains any null values or not.

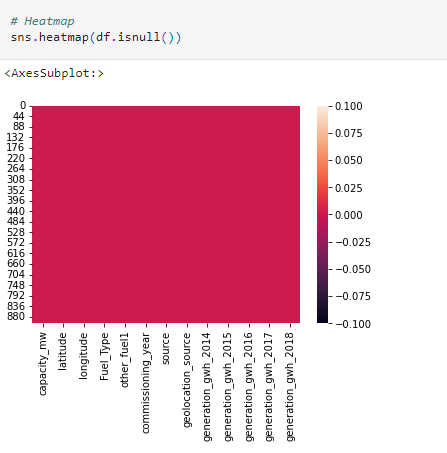


**Observation:** It can be clearly seen that we have high number of null values in the dataset.

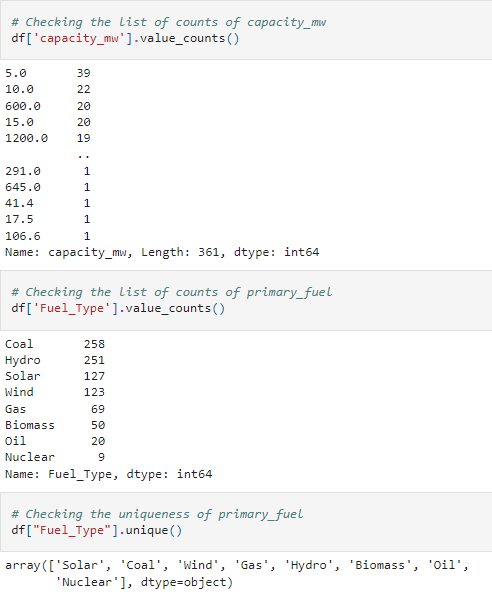
* 1. We will perform feature selection to remove the skewness & to treat null values: country, country\_long, other\_fuel2, year\_of\_capacity\_data, generation\_data\_source, other\_fuel3, wepp\_id, generation\_gwh\_2013, generation\_gwh\_2019, estimated\_generation\_gwh all are contained most of null values and there is no impact of this features on prediction hence we can drop these features. Also, the columns "name", "gppd\_idnr", "owner", "url" are all unique values and there is no impact of this feature on prediction hence we can drop these features as well.
  2. Even after treating the feature selection our dataset still contains some null values.



* 1. We will treat null values using imputation technique by calculating mean (of latitude) & mode (of other\_fuel1 & geolocation\_source) & will fill the null values.
  2. After performing imputation technique our does not contain any null value.

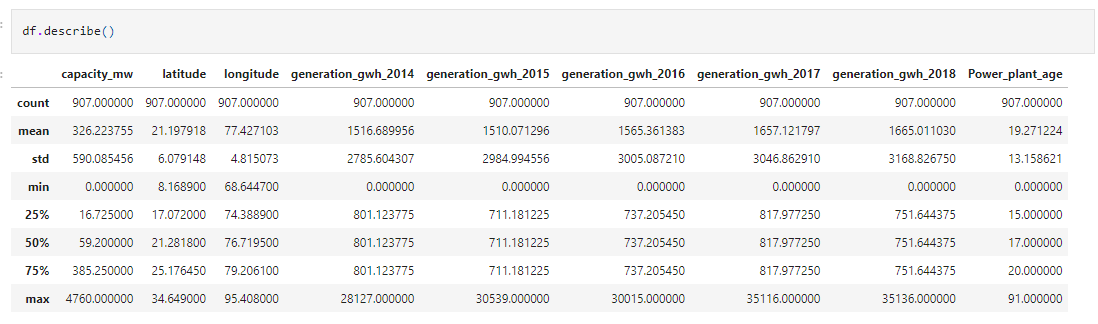


* 1. We will check list of counts of primary\_fuel & capacity\_mw & will also check uniqueness of primary\_fuel and we calculate age of power plant.



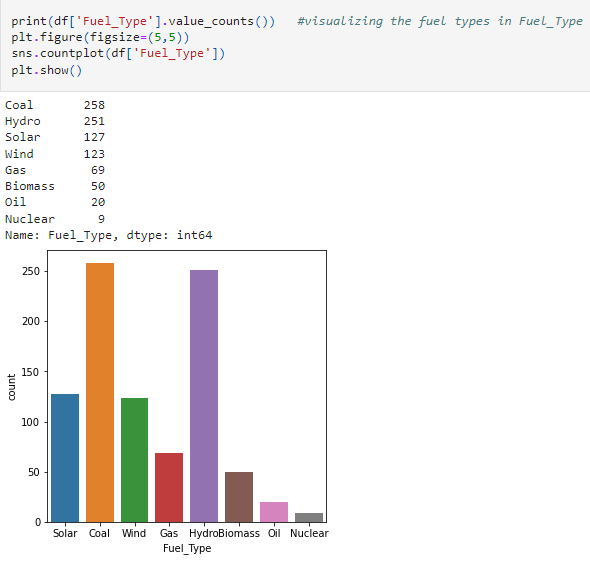
* 1. Statistical Summary:

We can observe that the count of the columns is same, which means the dataset is balanced. The minimum capacity of the power plant is zero and maximum in 4760 and there is huge difference in mean and standard deviation. From the difference between maximum and 75% percentile we can infer that there are huge outliers present in most of the columns, we will need to remove them using appropriate methods before building our model.

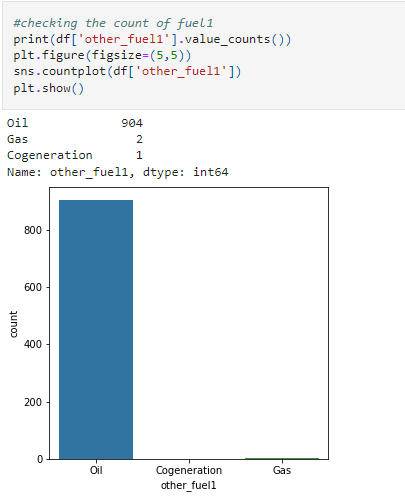


* 1. As our dataset contain categorial as well as numerical data, we will check the categorical & numerical columns.
  2. **Data Visualization:** To understand our data, we will look for correlations between features and the label. This can be important when choosing a model. E.g., if features and a label are linearly correlated, a linear model like Linear Regression can do well; if the relationship is very non-linear, more complex models such as Decision Trees can be better. We can use matplotlib & seaborn in visualization to view each of our predictors in relation to the label column as a scatter plot to see the correlation between the predictors and the label.

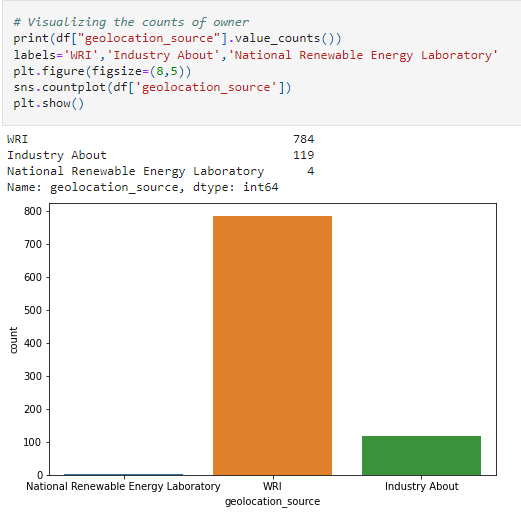
1. **Univariate Analysis: visualization using Categorical column**



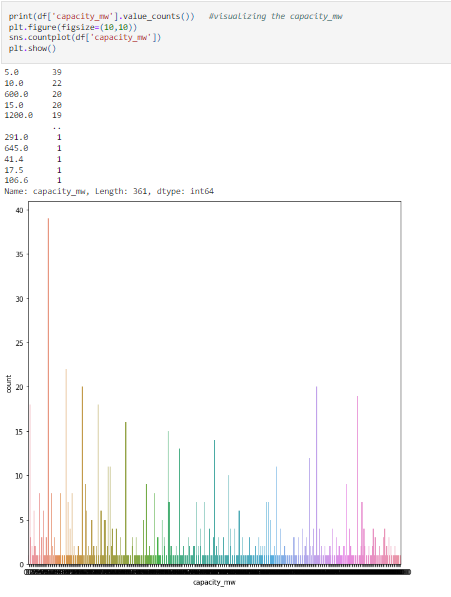
**Observation:** In the above count plot for "primary\_fuel" column, we can see that the highest number of values have been covered by coal and hydro fuel types then comes solar and wind. Finally, we see that gas, biomass, oil and nuclear have very low data counts.



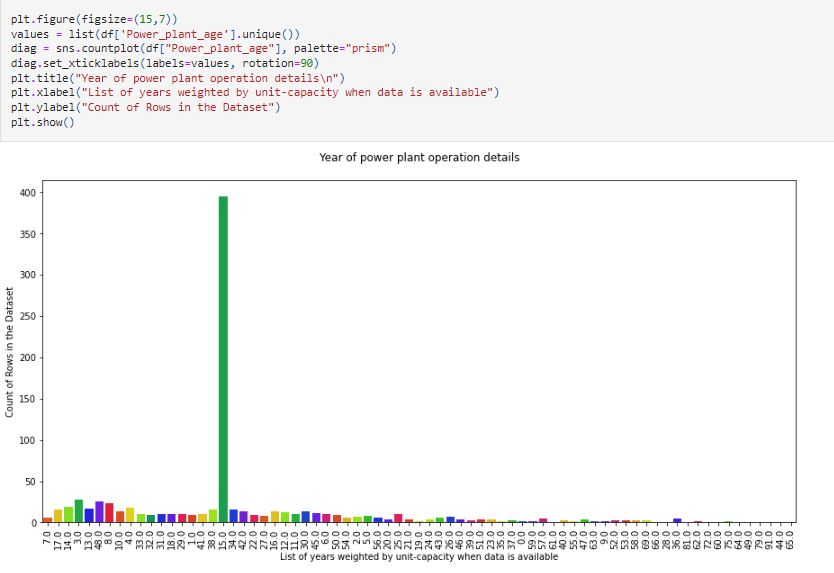
**Observation:** It can be observed that 'other\_fuel1' type has 3 unique types namely: 'Oil', 'Cogeneration other fuel', 'Gas'. And it is clearly seen that oil is the max used fuel type.



**Observation:** Here it can be seen that the count of WRI is the max, which means that the max information is shared by this source.

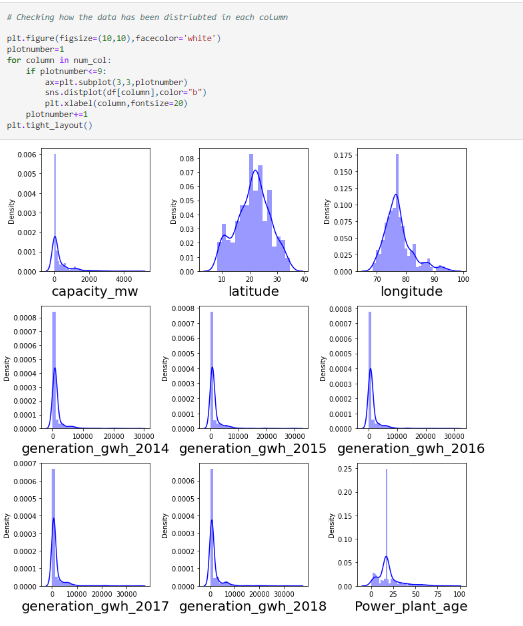


**Observation:** Here it can be seen the counts with respect to capacity\_mw.



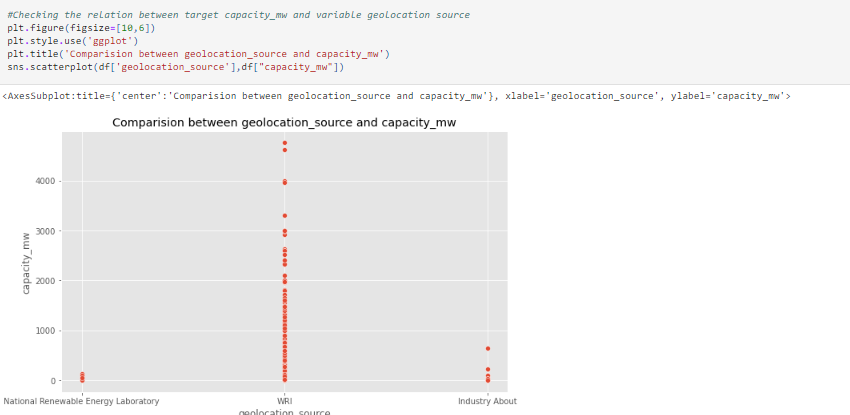
**Observation:** In the above count plot we can see the list of years as to when the power plant data was made available. Since we had missing values in the "commissioning\_year" column we replaced it with the median wherein the year "15" covered the most rows in our dataset compared to all the other years.

**Distribution of the Dataset: To check skewness present in the dataset if any**

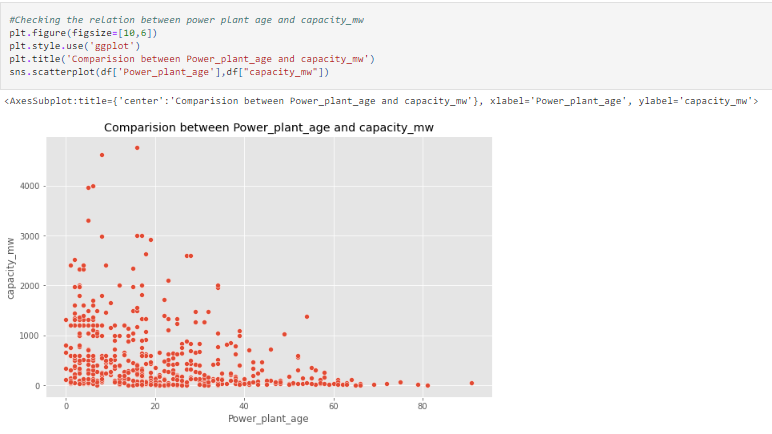


**Observation:** Here in the plots, we can see that the data is not normally distributed. Outliers and skewness are present, which needs to be treated.

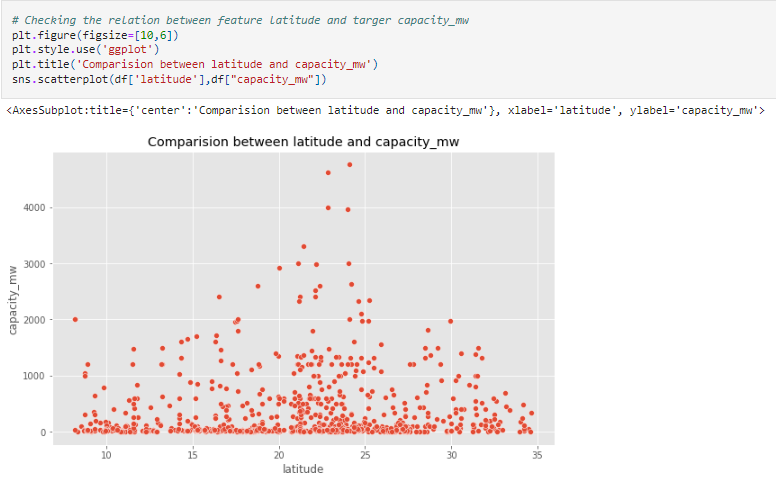
1. **Bivariate Analysis:** Correlation between features and target 'Capacity\_mw'



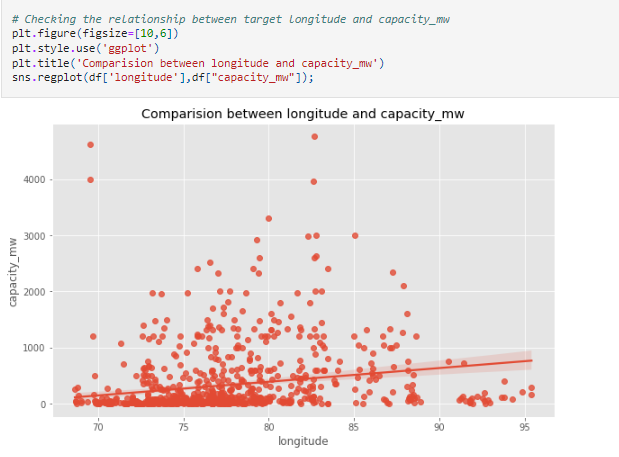
**Observation:** Here also we can see that WRI 'geolocation\_source' plays a major role in the power plant to increase the capacity of electricity generation.



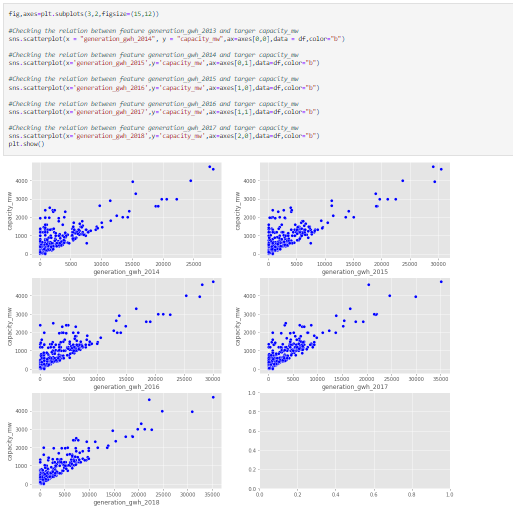
**Observation:** Here we can see a negative correlation between power\_plant\_age & capacity of\_mw. Which means that older the power plant lowers the electricity generation capacity.



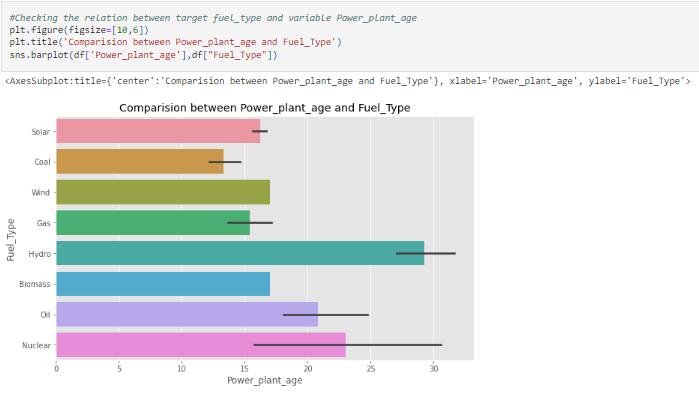
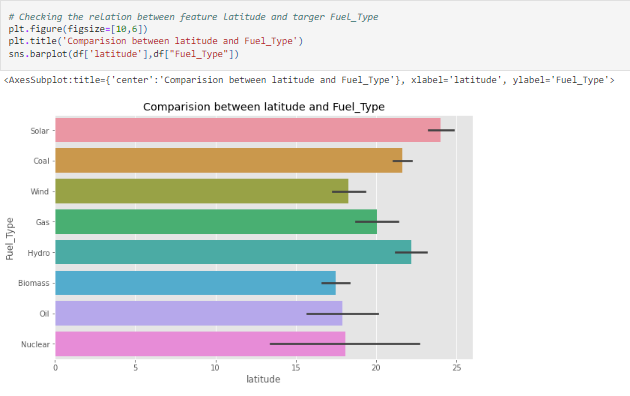
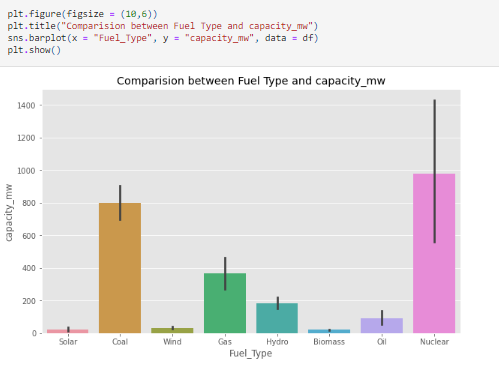
**Observation:** Here this feature does not show any linear relationship between latitude & capacity\_mw.

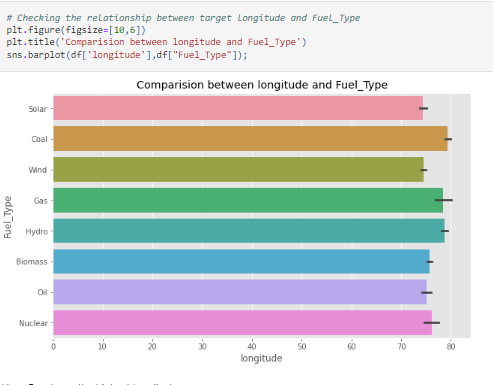


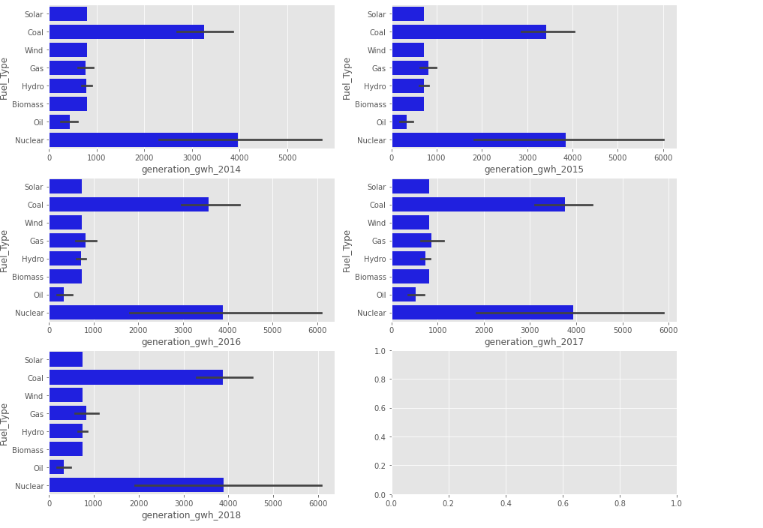
**Observation:** This feature also does not show any linear relationship between longitude & capacity\_mw.



**Observation:** This feature shows a positive correlation. Here the electricity generation reported for the years has capacity above 1000 mw also as the generation growth increases, the capacity of plant is also increasing moderately.

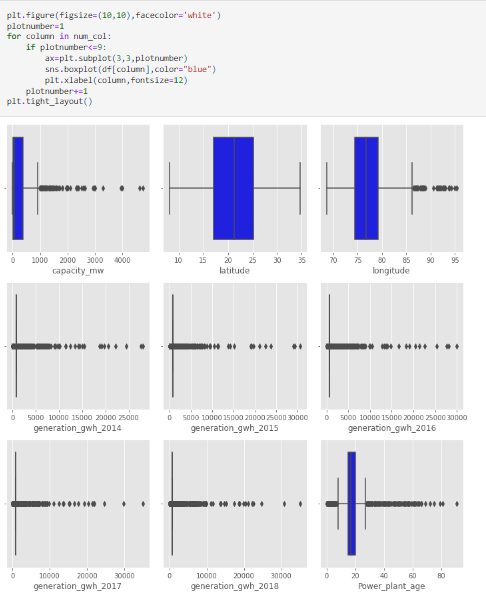
**Checking Correlation between features and target & checking relationship between both target:**





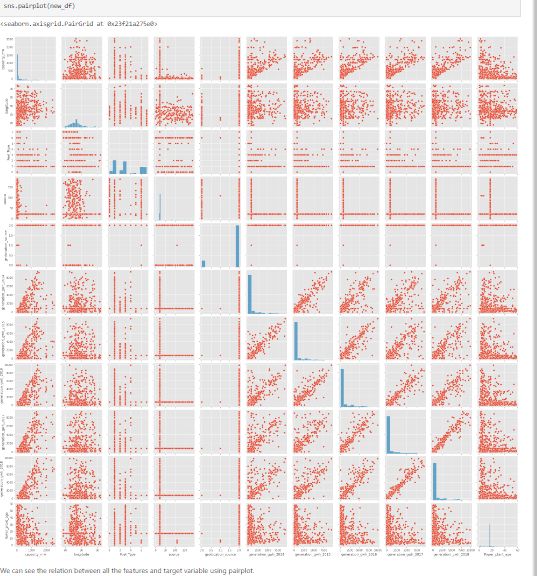
**Observation:** As per analysis we have observed that,older power plants use Hydro as energy source, followed by oil & the newer power plants are using more of Coal, Solar and Gas. Also, as per the analysis we have found that, most used energy source in all the years is nuclear followed by coal which contain high capacity respectively.

j) We will identify the outliers present the dataset or not.



# Observation: In the boxplot we can notice the outliers present in all the columns except latitude. Even target column has outliers but no need to remove it. Let's remove outliers using Zscore method.

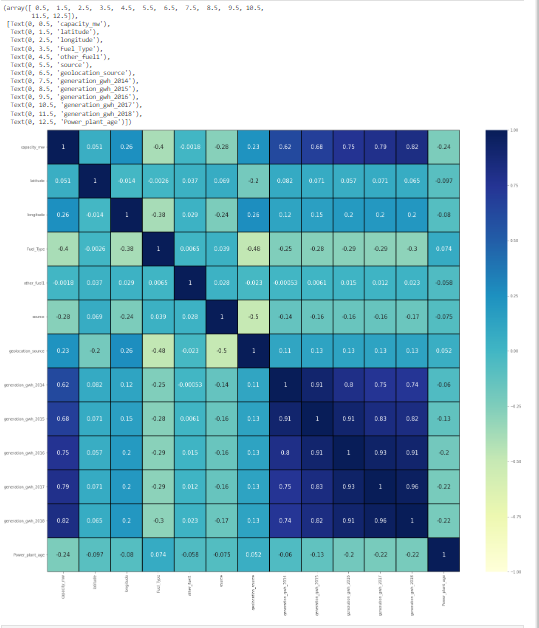
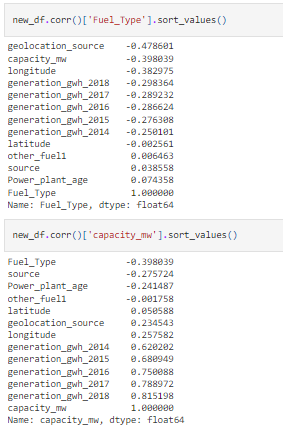
1. **Multivariate Analysis: Using Pairplot**



***4) Pre-Processing Pipeline:***

We have many steps included in pre- processing like Data cleaning, Data reduction, Data integration etc.

1. Label Encoding: We will perform label encoding technique to encode the categorical columns.
2. We will perform the Z score method to remove the outlier and will create new dataframe.
3. After removing the outlier, we will check the percentage of data loss. We have 6.17% data loss which is acceptable.
4. We will check correlation between the target variable & features of new dataframe.



**Observation:** The label capacity\_mw is highly positively correlated with the features generation\_gwh\_2017, generation\_gwh\_2016, generation\_gwh\_2015, generation\_gwh\_2014, generation\_gwh\_2013. And the label is negatively correlated with the features Fuel\_Type, source and Power\_plant\_age. The columns other\_fuel1 and latitude have no relation with the label, so we can drop them.

The label Fuel\_Type is less correlated with Power\_plant\_age and source. The label is negatively correlated with geolocation\_source, longitude, capacity\_mw, and all generation\_gwh years.

From the heat map we can notice most of the features are highly correlated with each other which leads to multicollinearity problem. So will try to solve this problem by Checking VIF value before building our models.

Also, the features other\_fuel1 and latitude have very very less correlation with both the labels. Hence after checking VIF, we can think about dropping these 2 columns

1. We will visualize the correlation between label & features using bar plot.



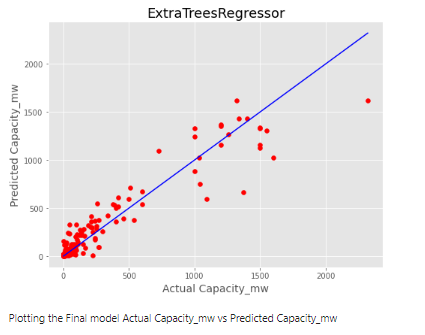
1. We will check the multicollinearity using VIF technique: other\_fuel1 has the highest VIF FACTOR so we will drop this column. latitude has the lowest contribution compared to both the targets so we will drop latitude column. After dropping these two columns we will perform feature selection technique.

**5.*****Building Machine Learning Models:*** Here we will perform ML for two target variables.

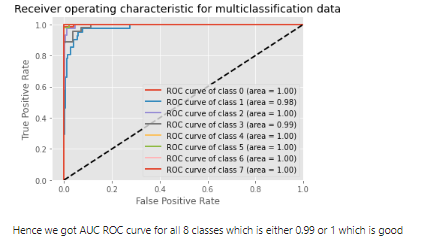
i) Predicting “Capacity\_mw” Target

ii) Predicting "Fuel\_Type" Target

1. **Predicting “Capacity\_mw” Target:**
2. First, we will split new dataset into features & target.
3. We will check & remove skewness from new dataset using PowerTransformer method ('yeo-johnson')
4. We will perform StandardScaler method to scale the dataset.
5. We will check multicollinearity with VIF technique
6. We will find best random state for .20 test size. We got best r2 score as 88.81% at best random state 185.
7. We will perform Train\_Test\_Split and we will create a function to run the regressor.
8. We will use different ML regression algorithm. Linear Regression, L1 Lasso Regression, L2 Ridge Regression, Elastic Net, Support Vector Regression, Decision Tree Regression, Random Forest Regression, k Neighbors Regression, SGD Regression, Ada Boost Regression, Extra Trees Regression.
9. Comparing all the regression algorithms the Extra Trees Regressor gives the best results since the R2 Score (88.47%) - Cross Validation Score (79.20%) are closest along with higher Cross Validation Score and the highest R2 score comparing all the models.
10. We will perform Hyper Parameter Tuning technique by using Extra Trees Regression. While tunning the model we will get the parameter keys. Then we will create the parameter list to pass into GridSearchCV. We will run the GCV for ETR. Then we will train the ETR model with GCV. We will get the best parameter for ETR & we will get the best R2 score 89.29% for our best ETR model.
11. We will use the mean\_squared\_error, mean\_absolute\_error and we will plot the Final Model for Actual Capacity\_mw vs Predicted Capacity\_mw.



1. We will save our best regression model using pickle.
2. **Predicting "Fuel\_Type" Target:**
3. First, we will split new dataset into features & target.
4. We will check & remove skewness from new dataset using PowerTransformer method ('yeo-johnson')
5. We will perform StandardScaler method to scale the dataset.
6. We will check multicollinearity with VIF technique. We can see that the target Fuel\_Type has multiple classes in the mode of energy source, hence we can see that this is a multi-classification problem. As the data between the classes are not balanced with 1 having 238 counts and 4 having only 9 counts, we have to do SMOTE oversampling of the data.
7. We will find best random state for .20 test size. We got best r2 score as 95.01% at best random state 153.
8. We will perform Train\_Test\_Split and we will create a function to run the Classification Models.
9. We will use different ML classification algorithm. Logistic Classification, Naïve Bayes, SVC Classifier, Decision Tree Classifier, k Neighbors Classifier, SGD Classifier, Random Forest Classifier, Extra Trees Classifier, Ada Boost Classifier, Gradient Boosting Classifier.
10. Comparing all the Classification algorithms the Extra Trees Classifier gives the best results since the R2 Score (93.43%) - Cross Validation Score (92.85%) are closest along with higher Cross Validation Score and the highest R2 score comparing all the models.
11. We will perform Hyper Parameter Tuning technique by using Extra Trees Classification model. While tunning the model we will get the parameter keys. Then we will create the parameter list to pass into GridSearchCV. We will run the GCV for ETR. Then we will train the ETR model with GCV. We will get the best parameter for ETR & we will get the best R2 score 93.96% for our best ETR model.
12. We will binarize output of the classification data & we will plot AUC ROC Curve for all the classes.



1. We will save our best regression model using pickle

**6)  *Concluding Remarks:*** After predicting the conclusion predicted values & original values of target variables are almost same. Here we had performed Regression as well as classification model of ML to predict two target variables. After performing ML algorithm, we found that Extra Trees Regression & Extra Trees Classification are best fit models for “Capacity\_mw” & “Fuel Type” target variable respectively with best R2 score 89.29% & 93.96 % respectively after tunning the model.