**Global Power Plant Database**

# **Introduction**

An affordable, reliable, and environmentally sustainable power sector is central to modern society. Governments, utilities, and companies make decisions that both affect and depend on the power sector. For example, if governments apply a carbon price to electricity generation, it changes how plants run and which plants are built over time. On the other hand, each new plant affects the electricity generation mix, the reliability of the system, and system emissions.

Plants also have significant impact on climate change, through carbon dioxide (CO2) emissions; on water stress, through water withdrawal and consumption; and on-air quality, through sulphur oxides (Sox), nitrogen oxides (NOx), and particulate matter (PM) emissions.

Despite the importance of the power sector, there is no global, open-access database of power plants. Existing databases fail to be either truly comprehensive or fully open. Many countries do not report their power sector data at the plant level, and those that do vary wildly in what they report, how they report it, and how frequently they report.

The lack of reporting standards makes data gathering time intensive, as the data are in different formats and must be harmonized. This creates a barrier for conducting global and national analysis of the power sector.

The Global Power Plant Database is an open-source open-access dataset of grid-scale (1 MW and greater) electricity generating facilities operating across the world. he Database currently contains over 35000 power plants in 167 countries, representing about 80% of the world's capacity. Entries are at the facility level only, generally defined as a single transmission grid connection point. Generation unit-level information is not currently available. The methodology for the dataset creation is given in the World Resources Institute publication "A Global Database of Power Plants"

# **Problem Statement:**

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 information on plant capacity, generation, ownership, and fuel type. It will be continuously updated as data becomes available.

In this dataset we need to make two predictions -

* Fuel Type
* capacity

First, we will predict capacity prediction, since the target variable "capacity" has continuous data so it is a "Regression problem". After this will move to Fuel Type prediction which is "Classification problem".

# **Description of Dataset**

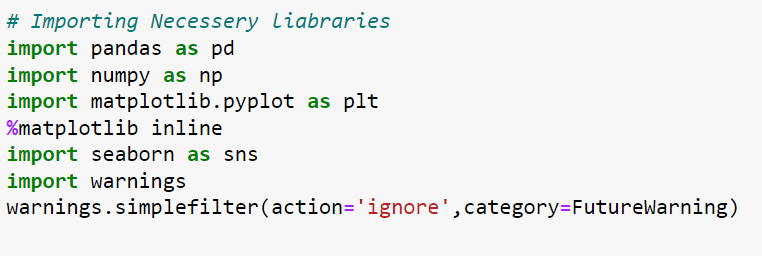
The database includes the following indicators:

* `country` (text): 3-character country code corresponding to the ISO 3166-1 alpha-3 specification [5]
* `country long` (text): longer form of the country designation
* `name` (text): Name or title of the power plant, generally in Romanized form
* `gppd\_idnr` (text): 10- or 12-character identifier for the power plant
* `capacity\_mw` (number): electrical generating capacity in megawatts
* `latitude` (number): geolocation in decimal degrees; WGS84 (EPSG:4326)
* `longitude` (number): geolocation in decimal degrees; WGS84 (EPSG:4326)
* `primary fuel` (text): energy source used in primary electricity generation or export
* `other\_fuel1` (text): energy source used in electricity generation or export
* `other\_fuel2` (text): energy source used in electricity generation or export
* `other\_fuel3` (text): energy source used in electricity generation or export
* `commissioning year` (number): year of plant operation, weighted by unit-capacity when data is available
* `owner` (text): majority shareholder of the power plant, generally in Romanized form
* `source` (text): entity reporting the data; could be an organization, report, or document, generally in Romanized form
* `URL` (text): web document corresponding to the `source` field
* `geolocation source` (text): attribution for geolocation information
* `wepp\_id` (text): a reference to a unique plant identifier in the widely-used PLATTS-WEPP database.
* `year\_of\_capacity\_data` (number): year the capacity information was reported
* `generation\_gwh\_2013` (number): electricity generation in gigawatt-hours reported for the year 2013
* `generation\_gwh\_2014` (number): electricity generation in gigawatt-hours reported for the year 2014
* `generation\_gwh\_2015` (number): electricity generation in gigawatt-hours reported for the year 2015
* `generation\_gwh\_2016` (number): electricity generation in gigawatt-hours reported for the year 2016
* `generation\_gwh\_2017` (number): electricity generation in gigawatt-hours reported for the year 2017
* `generation\_gwh\_2018` (number): electricity generation in gigawatt-hours reported for the year 2018
* `generation\_gwh\_2019` (number): electricity generation in gigawatt-hours reported for the year 2019
* `generation\_data\_source` (text): attribution for the reported generation information
* `estimated\_generation\_gwh\_2013` (number): estimated electricity generation in gigawatt-hours for the year 2013
* `estimated\_generation\_gwh\_2014` (number): estimated electricity generation in gigawatt-hours for the year 2014
* `estimated\_generation\_gwh\_2015` (number): estimated electricity generation in gigawatt-hours for the year 2015
* `estimated\_generation\_gwh\_2016` (number): estimated electricity generation in gigawatt-hours for the year 2016
* `estimated\_generation\_gwh\_2017` (number): estimated electricity generation in gigawatt-hours for the year 2017
* 'Estimated\_generation\_note\_2013` (text): label of the model/method used to estimate generation for the year 2013
* `estimated\_generation\_note\_2014` (text): label of the model/method used to estimate generation for the year 2014
* `estimated\_generation\_note\_2015` (text): label of the model/method used to estimate generation for the year 2015
* `estimated\_generation\_note\_2016` (text): label of the model/method used to estimate generation for the year 2016
* `estimated\_generation\_note\_2017` (text): label of the model/method used to estimate generation for the year 2017

Detailed analysis of Dataset was carried out with Python Jupiter notebook following important libraries were used for data analysis and model building –

* NumPy
* Pandas
* Matplotlib
* Seaborn

We are going to predict the Plant capacity at first in which is continuous data hence we are going to build regression model for it. Importing modules

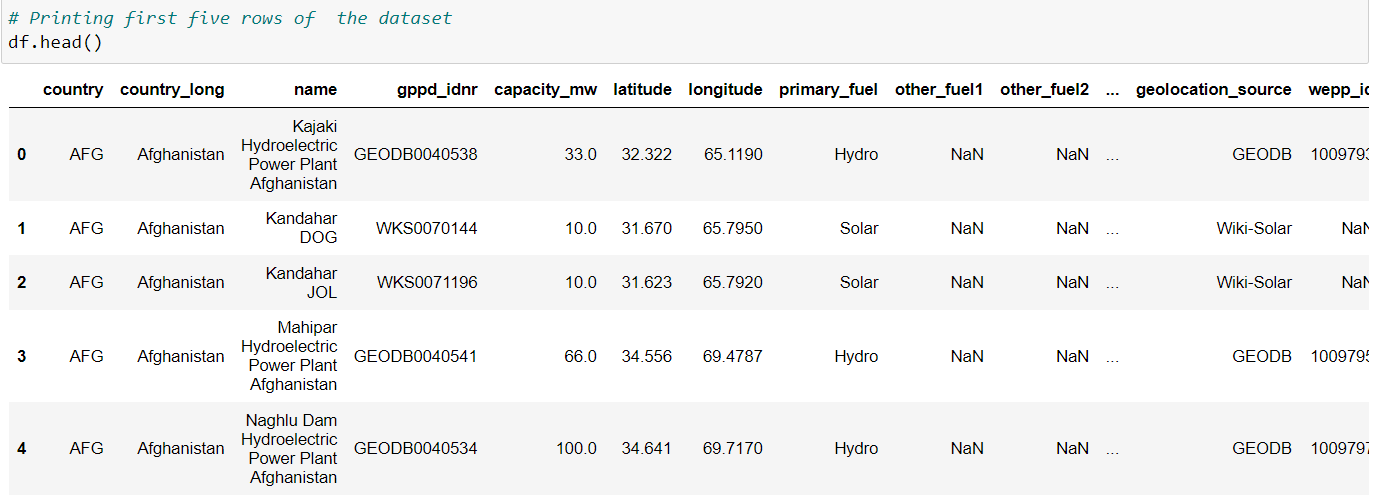


Let’s we take brief about these libraries; Pandas are used for data analysis NumPy is for n-dimensional array seaborn and matplotlib both have similar functionalities which are used for visualization.

Next step is to import dataset

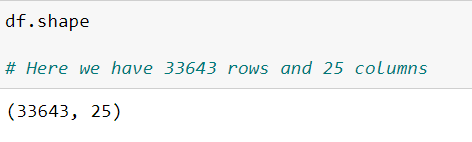
### 

### Study of Dataset



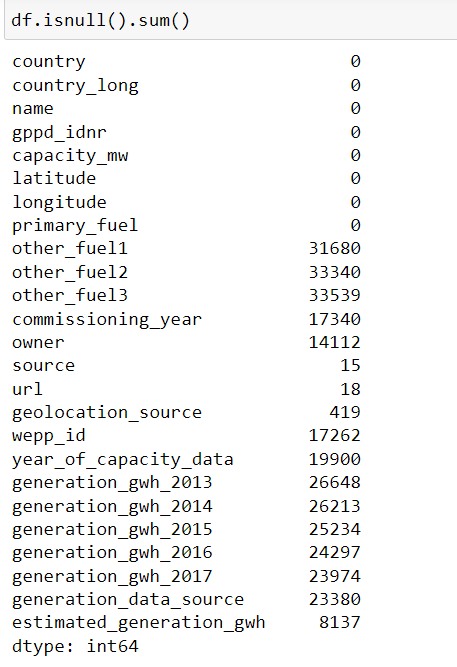
Printing the first five rows from the dataset gives us insight about the dataset, it consists both categorical and numerical columns and we need to do lot of pre-processing before visualization.

Printing the shape shows that our dataset consists of 33643 rows and 25 columns –



Our dataset consists of 33643 rows and 25 different columns.

Checking Null values from the dataset.



It was seen that our dataset has lot of null values which needs to be handled before going forward.

We will fill least null values because filling lost of null values may unbalance our dataset.

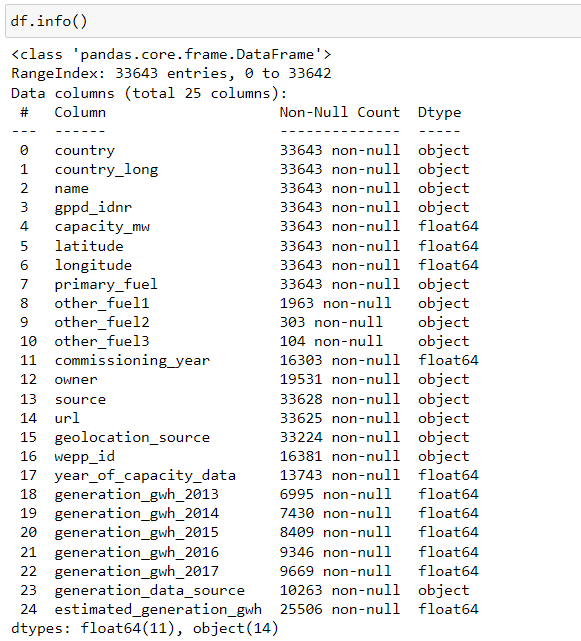
Columns containing more null values are –

* other\_fuel1
* other\_fuel2
* other\_fuel3
* generation\_hwh\_2013
* generation\_hwh\_2014
* generation\_hwh\_2015
* generation\_hwh\_2016
* generation\_hwh\_2017

Columns from which null values can be treated are –

* Estimated Generation
* Source
* Owner
* geolocation source

Printing datatypes of each column –



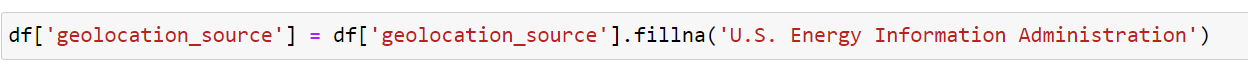
It was found that our dataset has total 11 numerical and 14 object columns.

# **Pre-processing**

We are dropping following columns as they are irrelevant of building the model for to both of the target variables also, they are containing lot of null values-

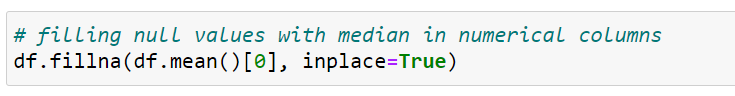
* country\_long
* name
* gppd\_idnr
* other\_fuel1
* other\_fuel2
* other\_fuel3
* url
* wepp\_id
* commissioning\_year
* owner
* year\_of\_capacity\_data
* generation\_data\_source

We will fill null values from source and geolocation columns with most repeated value (mode) of the column.

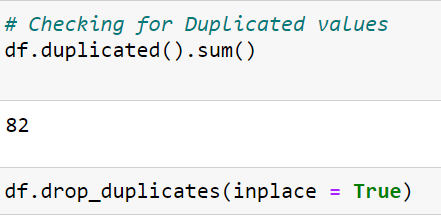




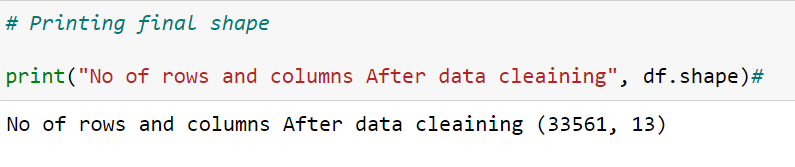
We are filling null values from the numerical columns with the mean of the dataset –



Also, we found that our dataset has 82 duplicated values we are dropping them for better processing



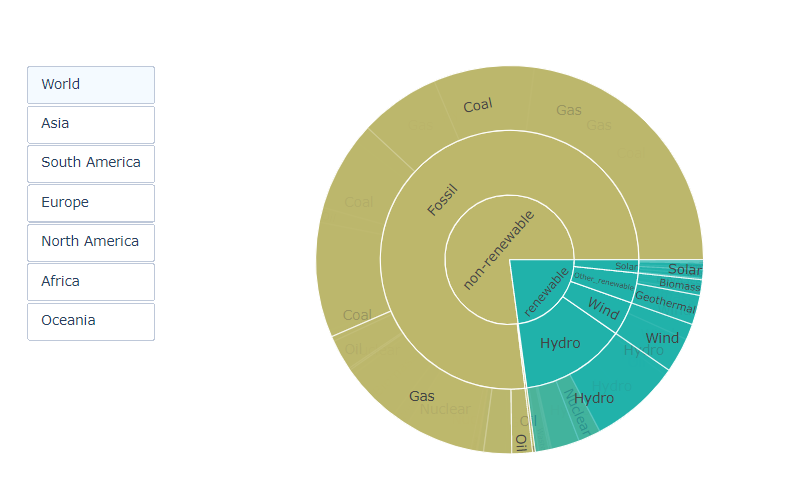
Printing Final shape after Pre-processing –



After pre-processing we have 33561 rows and 13 columns.

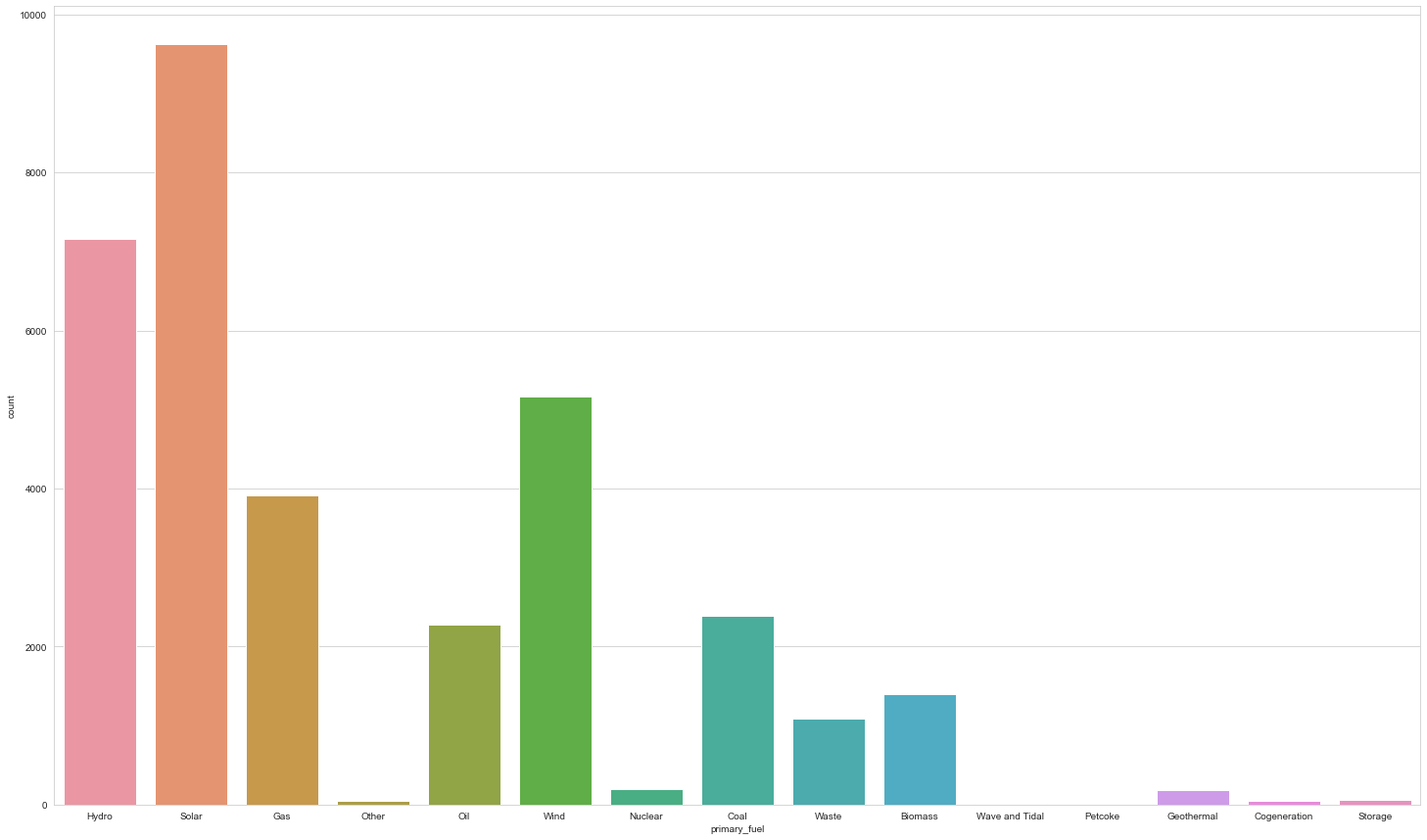
# **Data-Visualization –**

**The sunburst chart below shows what fuels we use to generate electricity in each continent.**



* Around 25% of power comes from renewable fuels worldwide.
* South America is blessed with abundant hydro resources.
* Nuclear power has a considerable share in Europe.

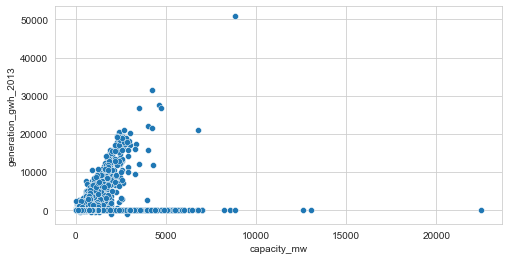
### Primary Fuel

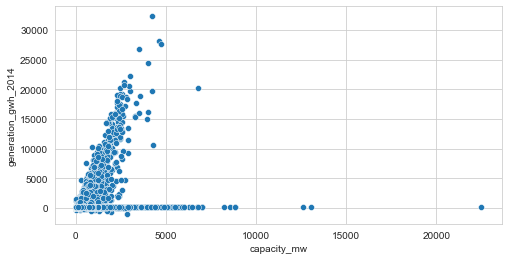


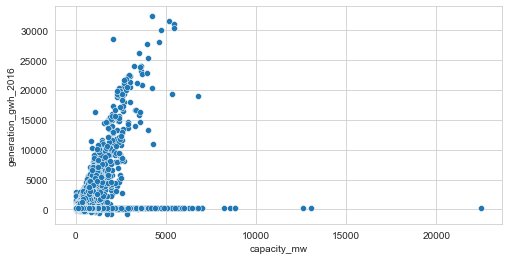
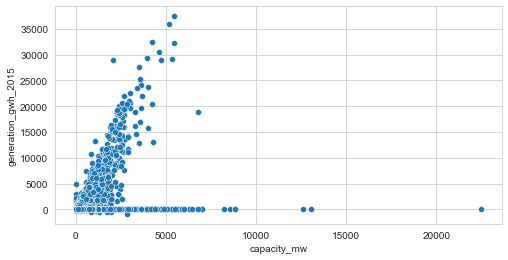
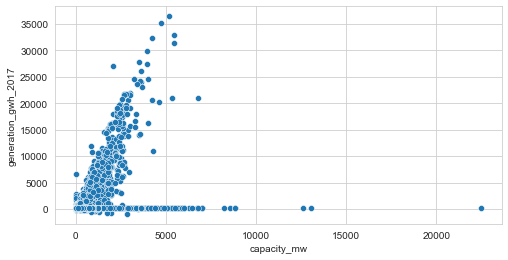
Count plot plotted for the Primary fuel shows that there are 16 different types of fuels used by the all over 35000 plants. We have following finding from the above plot most popular of them are – Solar, Hydro, wind, gas, oil, coal, waste, Biomass, Geothermal etc

Solar is most used primary fuel in all fuel types and it is followed by Hydro . Wind and gas are also popular fuel types.

### Power Generation and Capacity –







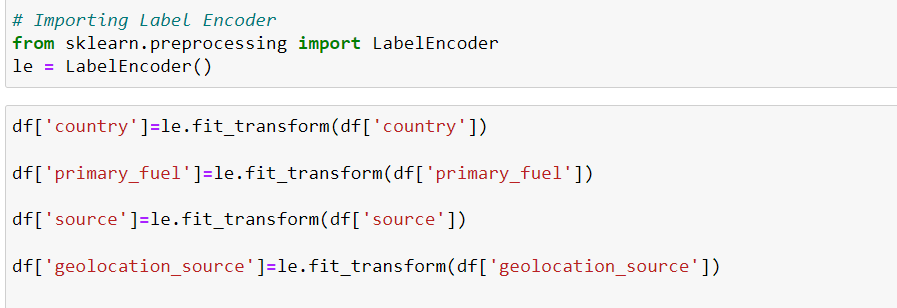
Above plots shows that linear regression exists between power generation and capacity of the plant.

Also We found that all plants have power generation capacity between 0 to 5000 MW.

It was seen that from year 2013 to 2014 we see a sudden rise in average generation of the power plants. After year 2014 we see small but constant rise in the in the power generation of the plants.

We can also see we have some outliers which needs to be handled before model building.

Before going forward, we need to we need to encode the categorical columns into numerical columns in order to model to understand the data –

Encoding columns with label encoder –

We are Encoding following columns into numerical values –

* Country
* Primary fuel
* Source
* Geolocation\_source

Label Encoder assign a unique integer based on alphabetical ordering.

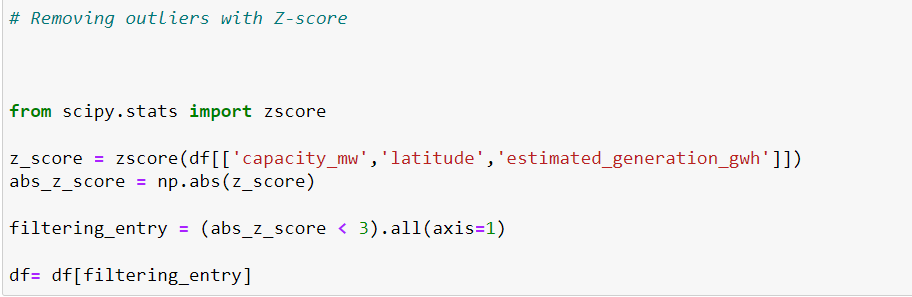
### Distplots for columns

Distplots plotted for the columns shows that Capacities are distributed between 0 to 1000 mw,

Latitude is distributed between -50 to +50 and longitude is distributed between -150 to +150.

Also, our dataset has skewness and it needs to be removed before moving forward.

Removing outliers with Z-Score –



We are using Z-score technique to remove the outliers

Z score is also called standard score. This score helps to understand if a data value is greater or smaller than mean and how far away it is from the mean. More specifically, Z score tells how many standard deviations away a data point is from the mean.

***Z score = (x -mean) / std. deviation***

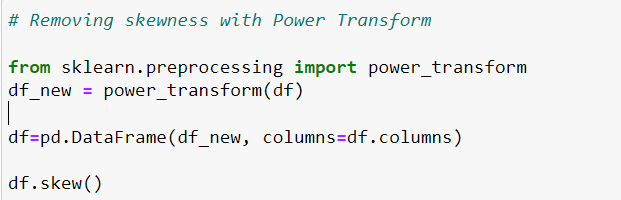
If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier.

After removing the outliers we have 32535 rows left In the dataset.

Data lost – 3.05 %

### Removing Skewness with Power Transform –

The power transformation is chosen so that the sampling distribution of the transformed test statistic is less skewed. Scale invariant sufficient conditions on the cumulants of the statistic are given which guarantee the reduction of skewness.

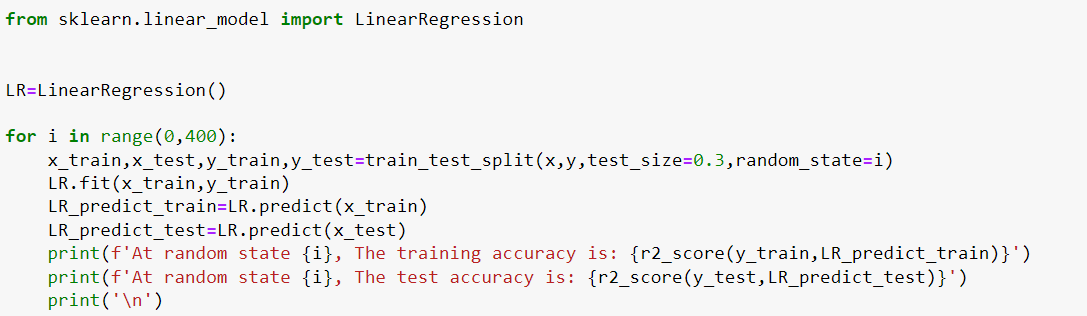


### Building Regression Model by taking Capacity as Target

Splitting data into columns and Lebel(target) –

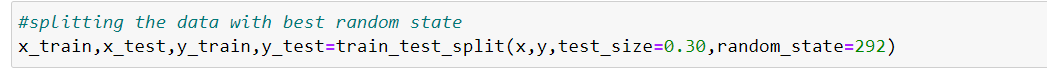


Now we need to split data into training and testing set, we will check best random state to split the data with checking R2 score for first 400 random states with split size 0.3 .

We will use Linear Regression to check the best R2 score

With checking the all 400 random states we got 292 as best random state –

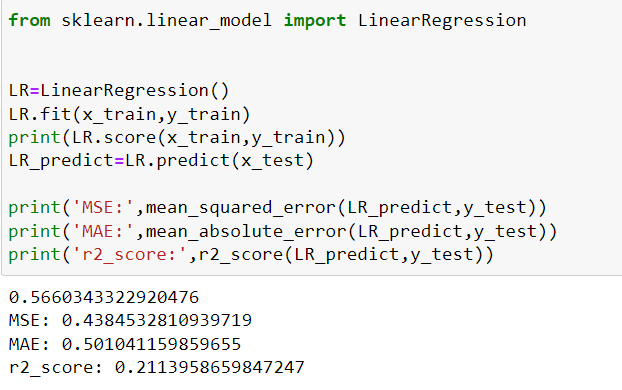
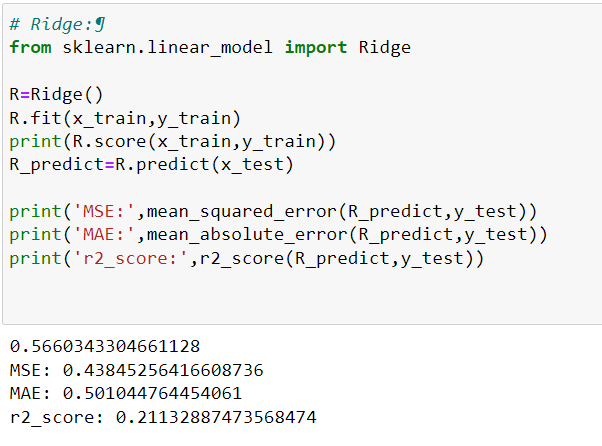
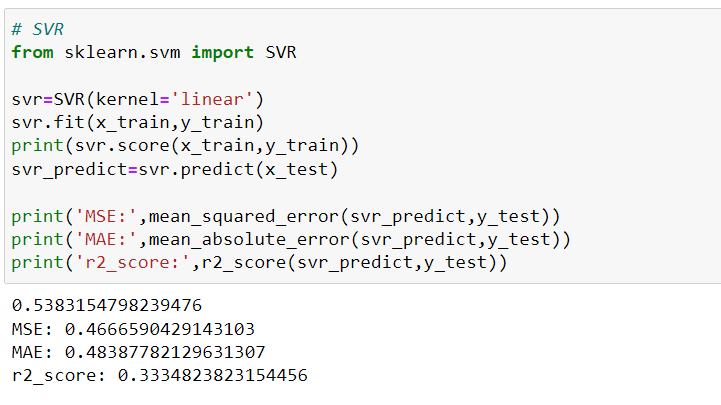
Splitting data into testing and training data with best random state



## Model Building

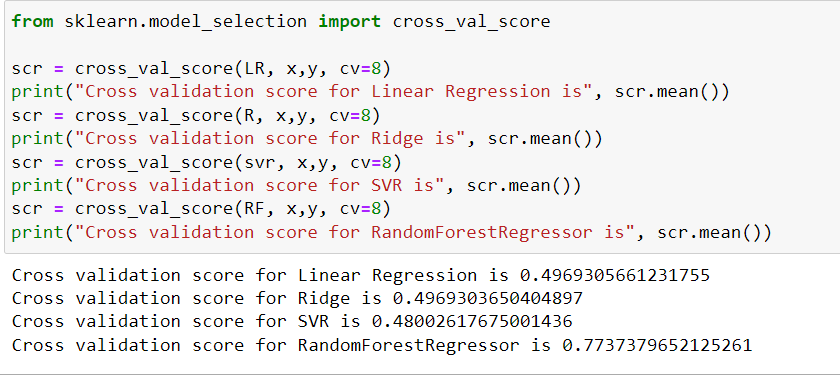
We have checked following regression models for our training and testing data –

* Linear Regression
* Ridge
* Support Vector Regressor
* Random Forest Regressor



We can see from above four models Random Forest model is Preforming better – It has best testing accuracy(99%) and 0.93 as best R2 score.

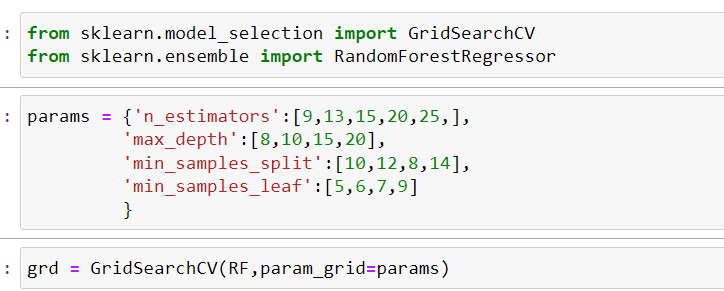
As we got good accuracy and good R2 score with Random forest Regressor but when we fit data into model it is possible that our model may underfit or overfit the data , we need to confirm it with cross validation –

Performing Cross validation for all four models –

With cross validation we confirmed that Random Forest is best performing model with cross validation score 0.78.

We can further improve this score by tunning parameters for model –

### Tunning Parameters with GridSearchCV -



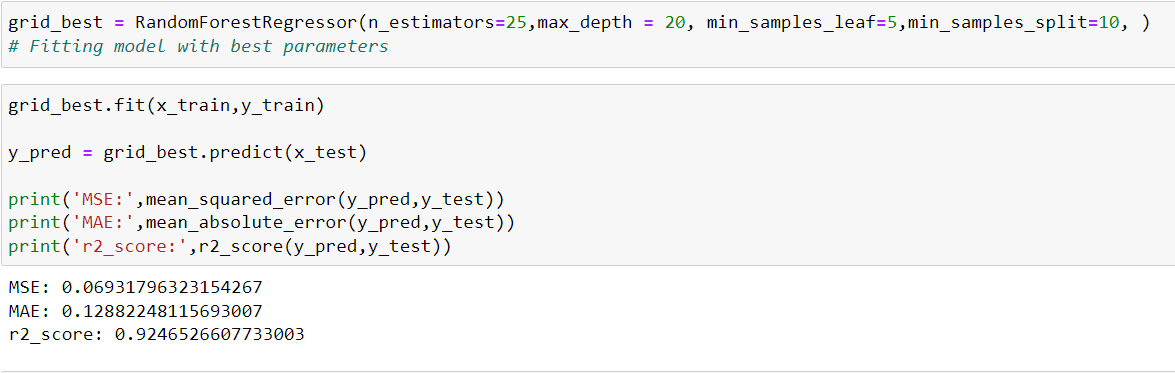
With will use tune following paramaters for random forest regressor –

* n\_estimators
* max\_depth
* min\_samples\_split
* min\_samples\_leaf

After tunning the parameters we got following best parameters for RF model –

Best Params= {'max\_depth': 20, 'min\_samples\_leaf': 5, 'min\_samples\_split': 10, 'n\_estimators': 25}

Model Building with best Parameters –



We can see we got good R2 score and least Mean Squared errors with our Grid best estimators model.

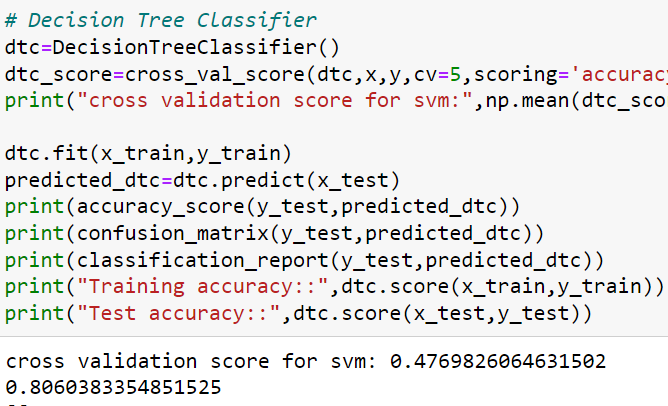
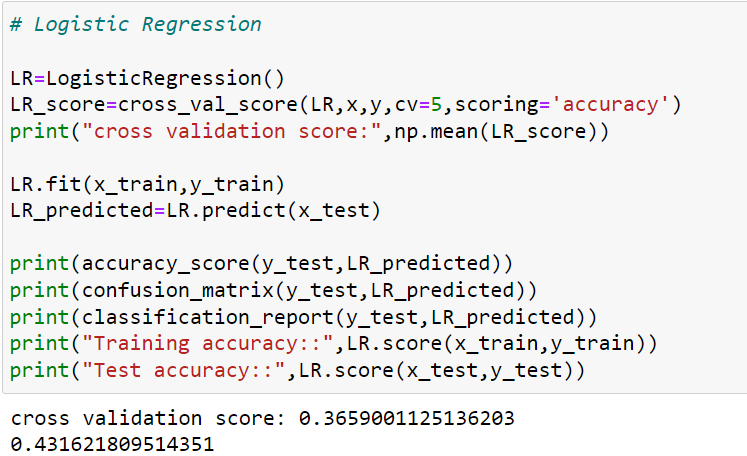
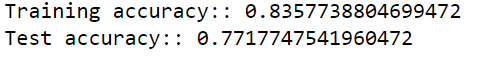
Checking Predicted Vs Actual with scatter plot .

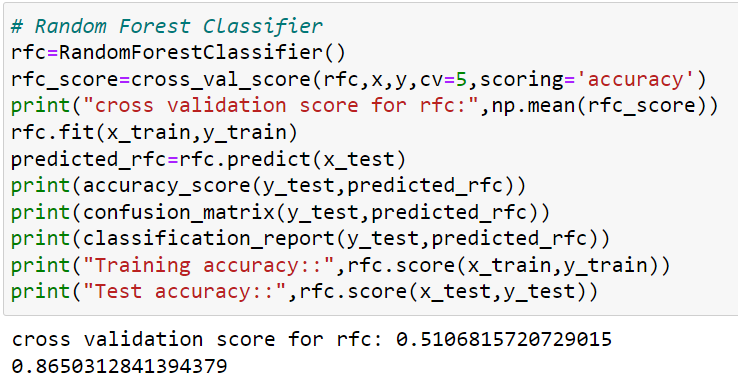
### 

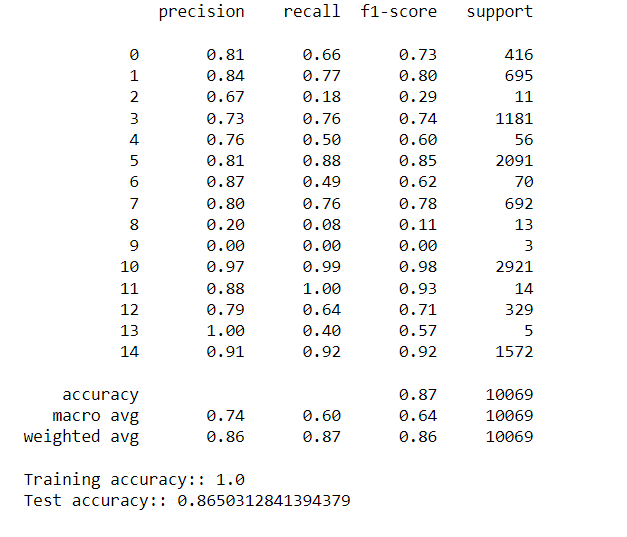
### Building Classification Model by taking Primary Fuel as Target –

For building the Classification model , we have checked following models –

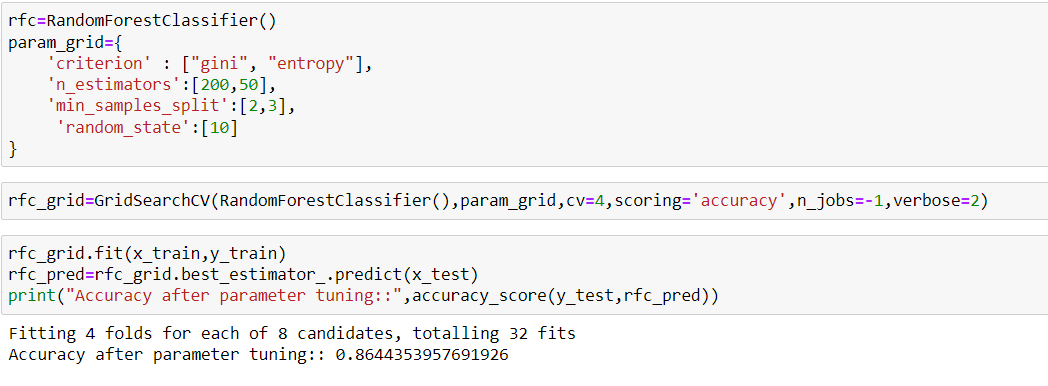
* Logistic Regression
* Decision Tree Classifier
* K Neighbors Classifier
* Random Forest Classifier
* AdaBoost Classifier
* Bagging Classifier
* Gradient Boosting Classifier





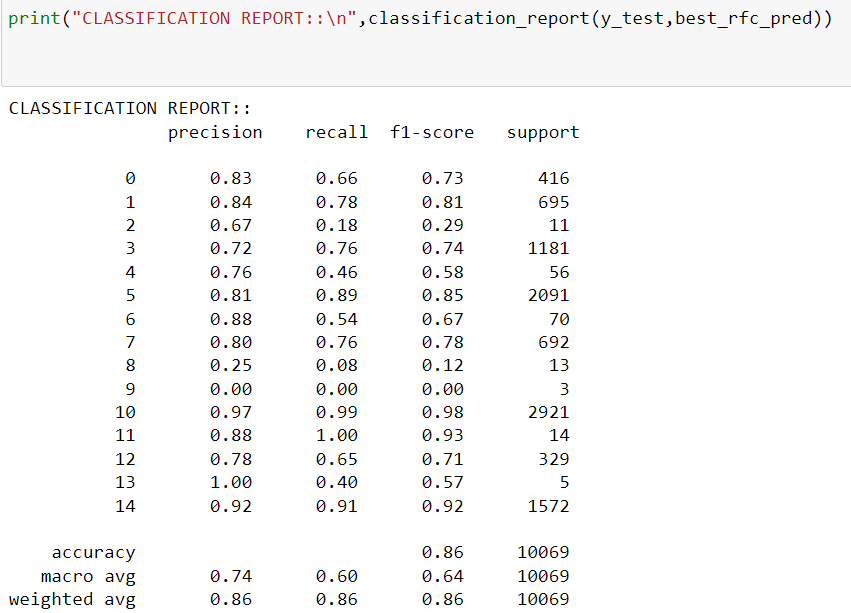


After evaluating all models, we found that Our Random Forest model is performing better than the others.

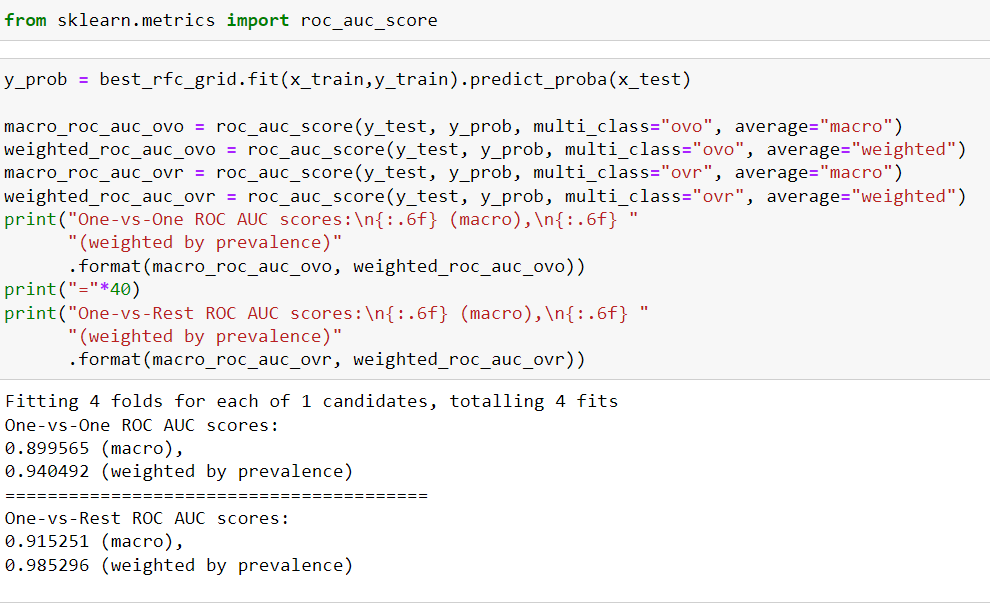
Hyper-Parameter tunning for Random Forest Classifier –

After Hyper-Parameter Tunning we got 86% accuracy for our model –

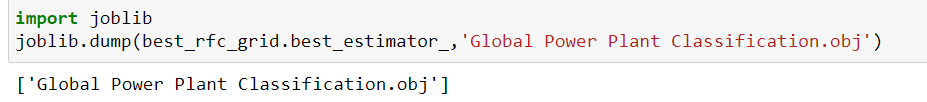
Printing Classification Report with Grid Search Best Parameters –



Printing Roc-AUC scores for the best parameters tunned model –



Saving the best grid-best estimator model –



# Conclusion

The benefits and costs of power plants, including their environmental impacts, depend on their technology and on how much electricity each plant actually generates. However, plant-level generation data are not reported in most countries. This technical note documents methods to estimate the annual electricity generation of power plants for the Global Power Plant Database. We use distinct estimation models for different fuel types, including wind, solar, hydropower (hydro), and gas power plants. The methodology combines statistical regression with machine learning techniques. Explanatory variables include plant-level characteristics such as plant size and fuel type, and country-level characteristics, such as country- and fuel-specific average generation per megawatt of installed capacity. We show that fuel-specific models can provide more accurate results for wind, solar, and hydro plants. Estimations for natural gas plants also improve, but the error remains high, especially for smaller plants.

Also, with visualization we found that nearly 25% of power comes from renewable fuels worldwide and yet the graph tends to increase in the future. South America is blessed with abundant hydro resources. Whereas Europe has great share in nuclear power plants. We don’t have more data collected from Asian countries future scope for project should be collecting data from more Asian and Smaller countries and removing errors from Estimation of natural gas plants.