1. **Description**

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.

1. **Key attributes of the database**

The database includes the following indicators:

**Key attributes of the database**

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

**3 Fuel Type Aggregation**

* We define the "Fuel Type" attribute of our database based on common fuel categories.

#### Prediction : Make two prediction 1) Primary Fuel 2) capacity\_mw

* **Find the dataset link below.**
* Downlaod Files:
* <https://github.com/wri/global-power-plant-database/blob/master/source_databases_csv/database_IND.csv>

***No Duplicate Entry Present in data.***

If we Check CSV file and look at dataset head, there are lot of data cleaning operation need to done before performing any EDA and ML modelling.At first sight we can come across following observation in CSV file:

* Lot of missing data in certain columns.
* Lot of Non relevant data like gppd\_idnr,url.
* and many more.

At end data need to clean and we will try to do some feature engineering afterwards to modify some columns.

**Start with looking at missing value.**

**Before checking null value and missing value imputation , first remove empty columns and non relevalent columns.**

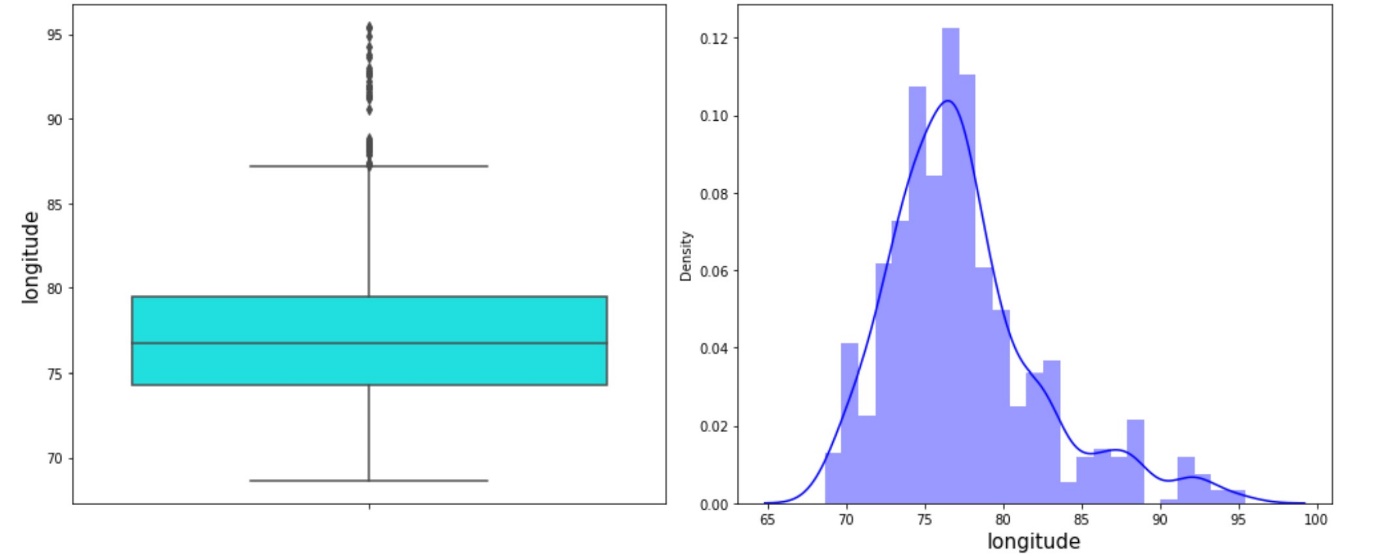
Columns we are going remove are :

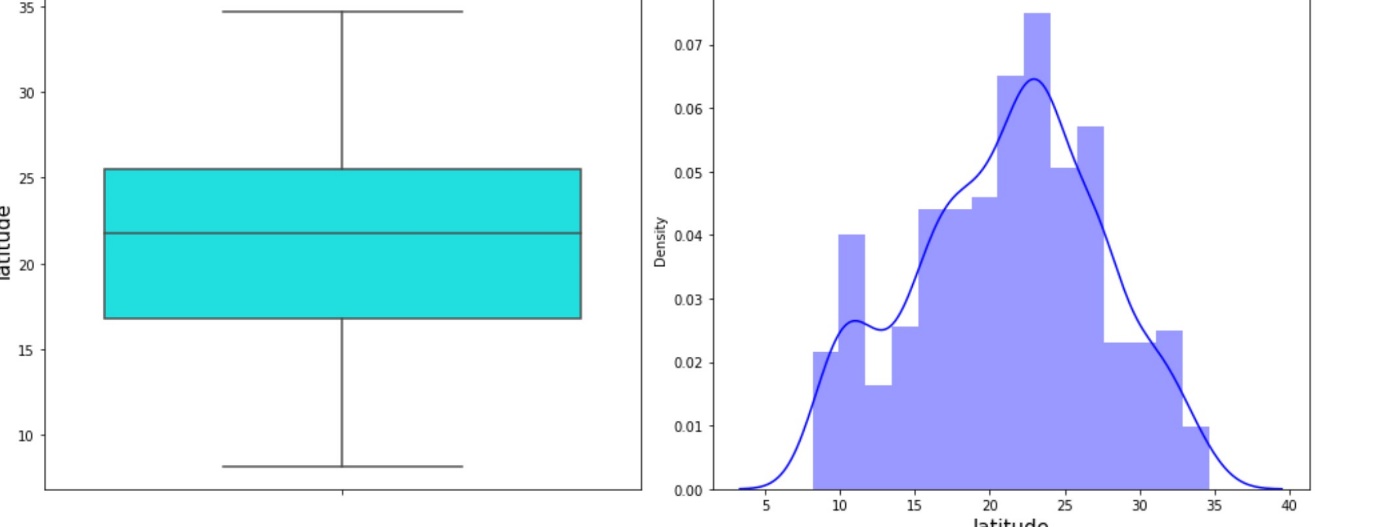
* estimated\_generation\_gwh - Empty
* wepp\_id - Empty
* other\_fuel2 - 98% data missing
* other\_fuel3 - Empty
* owner - More than 60 % data missing
* year\_of\_capacity\_data - Missing data with single unique value
* country - non relevalent info
* country\_long - non relevalent info
* gppd\_idnr -non relevalent info
* url - No missing value but of no use

**Proposed Strategy to Handle Missing data :**

* As Geolocation source is categorical data we can impute it with mode of category.
* longitude and latitude can be impute with mean or median of longitude and latitude. This imputation will not distrub statstical balance of data as mean will be same at the end.
* As commissioning year for most of industrial powerplant is missing(40%) after checking correlation we will decide to keep or drop this features.
* In generation\_data\_source 50% data is missing and it doesnot have any importance in our analysis. So it is better to drop this feature.
* There are 5 Different columns of GenerationGW-Hours for year 2013 to 2017. Its dive into it further
  + It is important feature in for coal and hydro powerplant.
  + It is natural to have missing data in this category. As Oil,Gas based Powerplant operated in intermitant periodic way and some renewable powerplant like wind,tide are operated seasonaly.
  + Some new powerplant commission between 2013 and 2018. For these powerplant some data will definitely available.
  + We cannot do any mean or median imputation here as different powerplant have different generation capacity & Generation per year depend on runtime of powerplant.
  + We all know old powerplant normally kept off unless more demand of generation required. Reason to kept is low efficiency & high operating cost.
  + We can neglect real value data for such important feature. We will keep this feature along with missing value and perform further investigation.
* Other\_fuel1 is another feature of some importance with missing value. Lets dive into it :
  + Not every powerplant build to work with alternate fuel.
  + Idea of other fuel is totally irrelevant to renewable energy source based powerplant like solar,wind,hydro.
  + First talk about powerplant for which concept of other fuel is applicable. We can impute them based domain knowledge.
    - Alternate Fuel of Coal based powerplant mostly is Oil or cogeneration.
    - Alternate Fuel of Industrial Oil based powerplant is Gas.
    - Alternate Fuel of Industrial Gas based powerplant is Oil.
    - We cannot define any alternative fuel for nuclear powerplant as it sole based on plant design & so many option.
  + For Renewable energy source based powerplant no alternate fuel needed. As it is categorical feature we can impute these powerplant with "Not Applicable". At end we are going Encoding these labels, 'Not Applicable' will be just one other additional label in encoding. Making no alternation on final result.

**Missing value Impuatation**

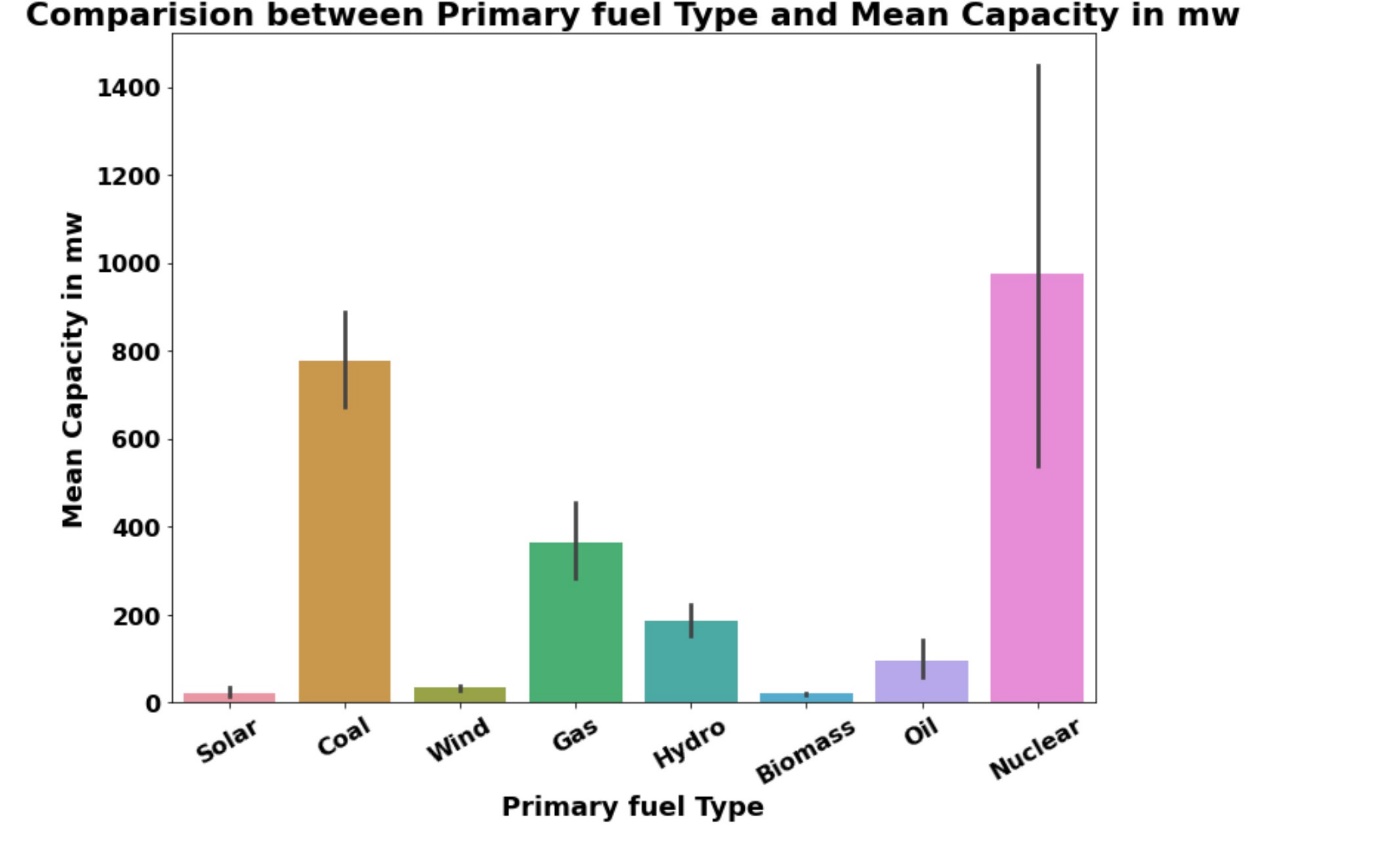
****

* **Based on presense of outliers we will impute longitude we with median**
* 

**Observation:**

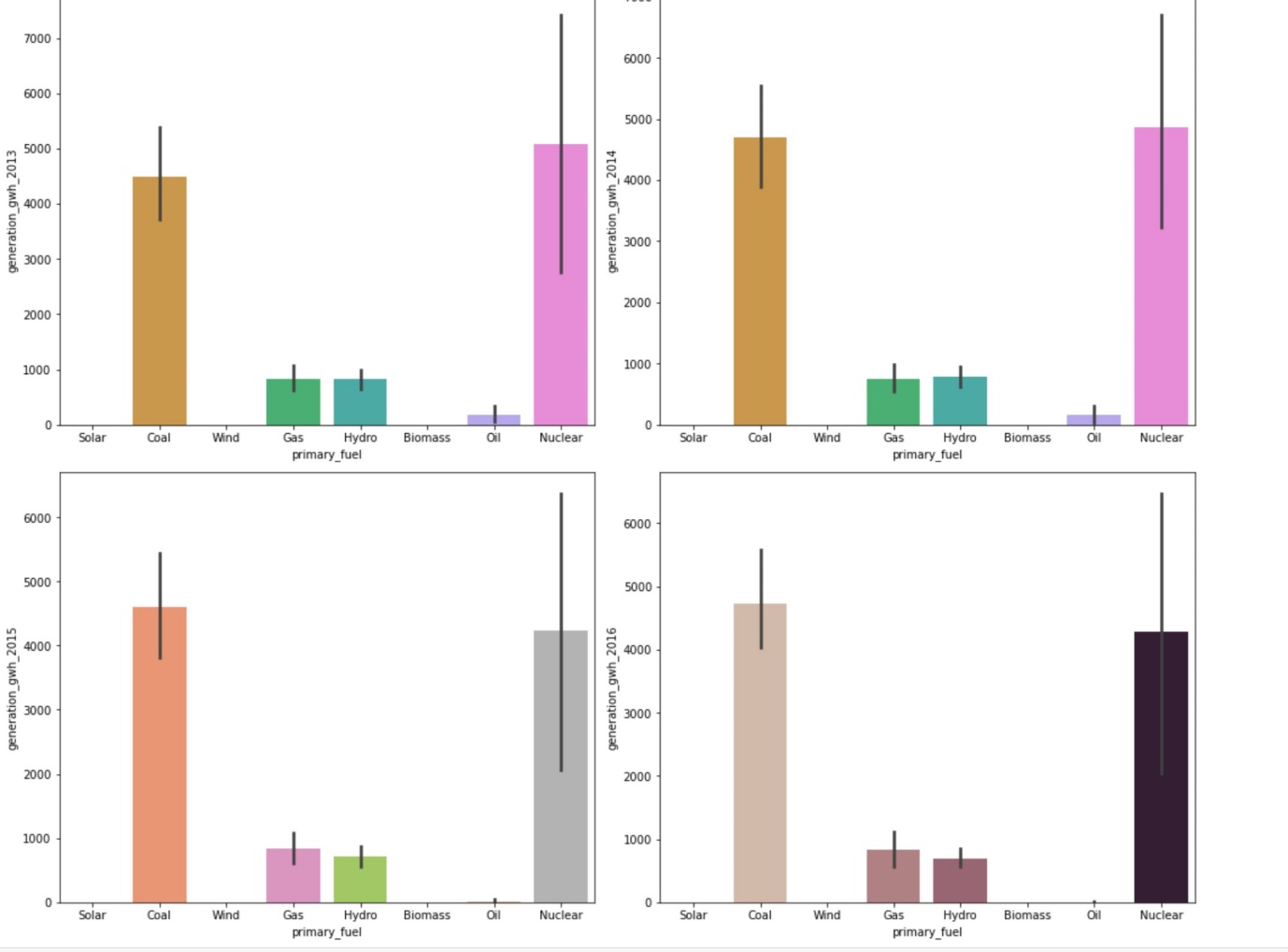
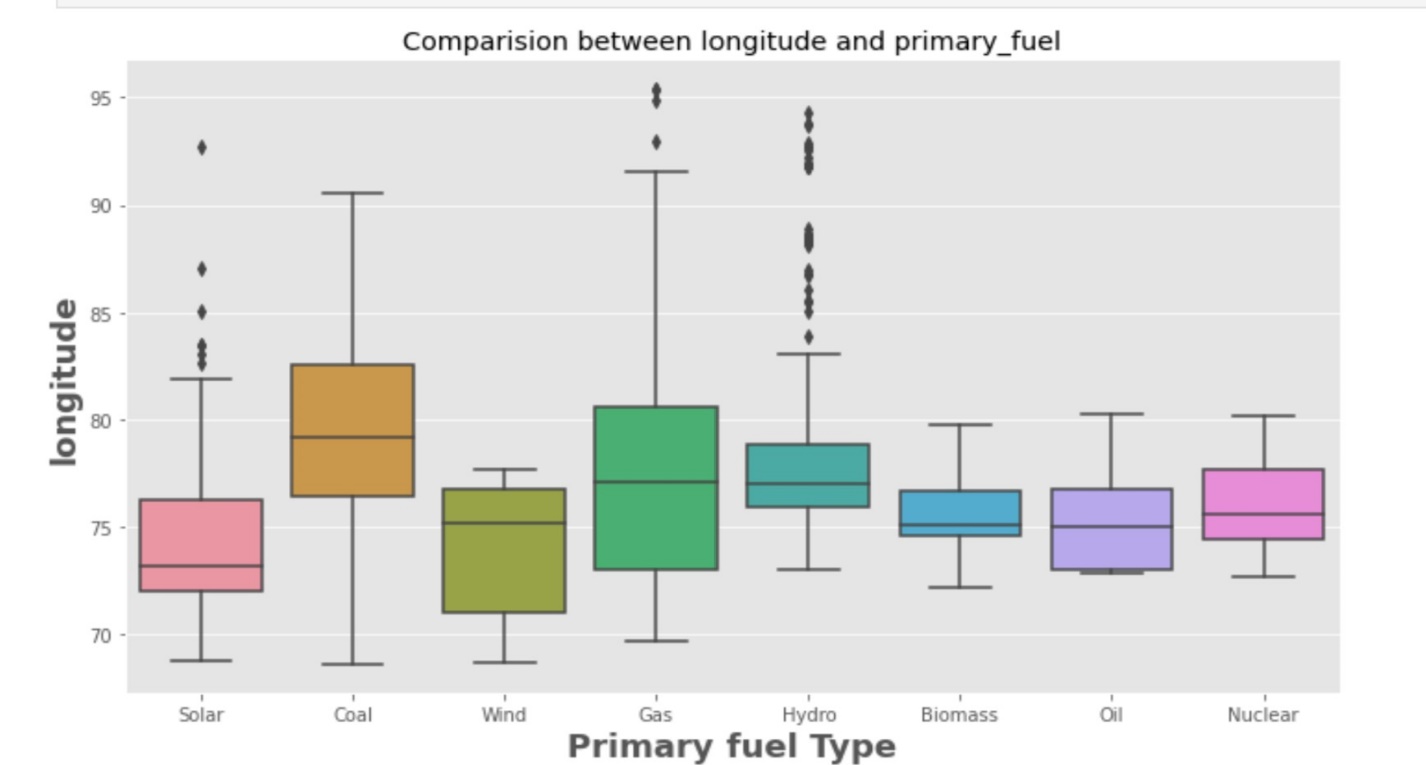
* Bigest Powerplant has power generation capacity of 4760 MW.
* In each Generation-GWHours columns Mean is greater than Median.
* Powerplants are located in latitude range of 8.1689 to 34.6490 while longitudal range is 68.64 to 95.4080.
* Oldest powerplant commission date back to 1927 and most recent powerplant is build in 2018. We will check in which timeframe most of powerplants are commission.
* Median of capacity MW is 60 MW. This suggest that there are lot of small capacity powerplant in dataset.

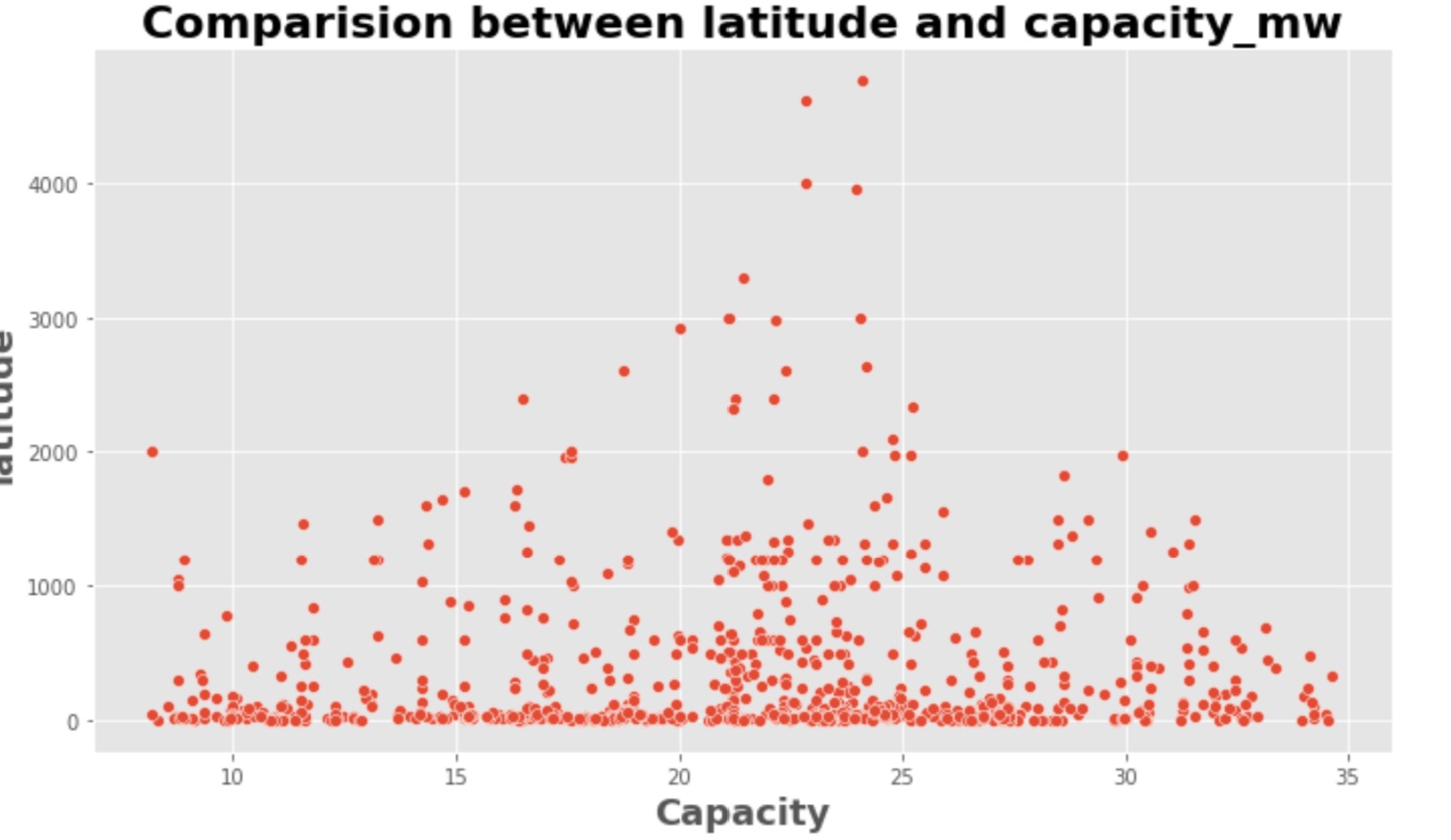
***We have two different Target feature for regression and classification model. LetStart exploring both Target Feature***

* 

**Above result is obvious as only 9 Nuclear powerplant so its mean Capacity is higher**

**.**





**Observation:**

* There is a positive linear relationship between the capacity and the electricity generation reported for the years mentioned.
* The plot shows the electricity generation reported for the years 2014,2015,2016,2017 have high capacity of above 1000mw. Also the power generation growth is more than 5000gwh in all the years.
* As the electricity generation growth increases, the capacity of plant also increases

## . **Logistics Regression Model**

* **Confusion matrix of Logistics Regression :**
* [[ 0 0 0 0 0 0 0 9]
* [ 0 48 0 19 0 0 7 2]
* [ 0 5 0 8 0 0 0 6]
* [ 0 14 0 50 0 0 5 6]
* [ 0 0 0 2 0 0 0 0]
* [ 0 0 0 4 0 0 0 1]
* [ 0 0 0 5 0 0 26 3]
* [ 0 2 0 3 0 0 9 29]]
* **classification Report of Logistics Regression**
* precision recall f1-score support
* 0 0.00 0.00 0.00 9
* 1 0.70 0.63 0.66 76
* 2 0.00 0.00 0.00 19
* 3 0.55 0.67 0.60 75
* 4 0.00 0.00 0.00 2
* 5 0.00 0.00 0.00 5
* 6 0.55 0.76 0.64 34
* 7 0.52 0.67 0.59 43
* accuracy 0.58 263
* macro avg 0.29 0.34 0.31 263
* weighted avg 0.51 0.58 0.54 263

**Decision Tree Classifier**

**classification Report of DecisionTreeClassifier**

precision recall f1-score support

0 0.70 0.78 0.74 9

1 0.78 0.53 0.63 76

2 0.24 0.32 0.27 19

3 0.72 0.80 0.76 75

4 0.00 0.00 0.00 2

5 0.12 0.20 0.15 5

6 0.70 0.76 0.73 34

7 0.70 0.74 0.72 43

accuracy 0.65 263

macro avg 0.50 0.52 0.50 263

weighted avg 0.68 0.65 0.66 263

* **Cross Validation Score LogisticRegression() :**
* Score : [0.53409091 0.48 0.52571429 0.53142857 0.52 ]
* Mean Score : 0.5182467532467533
* Std deviation : 0.019730431593058796
* ============================================================================================================
* **Cross Validation Score DecisionTreeClassifier() :**
* Score : [0.69318182 0.66285714 0.74857143 0.69714286 0.69142857]
* Mean Score : 0.6986363636363636
* Std deviation : 0.02777624253173901
* ============================================================================================================
* **Cross Validation Score RandomForestClassifier() :**
* Score : [0.72727273 0.77714286 0.78857143 0.77142857 0.76 ]
* Mean Score : 0.7648831168831169
* Std deviation : 0.020941186949287257
* ============================================================================================================
* **Cross Validation Score ExtraTreesClassifier() :**
* Score : [0.73863636 0.76 0.81714286 0.76571429 0.78857143]
* Mean Score : 0.774012987012987
* Std deviation : 0.02679131071012648

**Hyper Parameter Tuning : GridSearchCV**

gridSearchCV(estimator=RandomForestClassifier(),

param\_grid={'bootstrap': [True], 'criterion': ['gini', 'entropy'],

'max\_depth': [5, 10, 20, 40, 50, 60],

'max\_features': ['auto', 'log2'],

'n\_estimators': [5, 10, 15, 25, 50, 60, 70]},

verbose=5)

**Final Classification Model**

**Accuracy Score :**

0.7604562737642585