DS 250: Data Analysis and Visualization

Solar Power Potential Estimation in India

Thummala Milind Kesar (11841160)
Pratik Sanjay Patil (11840850)
Gundu Shreya (11840530)
Ruchika Gaur(11840960)
Saksham Bhushan (11840980)

Background

- Geographical location.
- Annual radiation of 5000 trillion kWh.
- India is 3rd largest solar power industry.
- Targets getting achieved ahead of deadline.
- Target was raised to 100 GW of solar capacity by 2022, targeting an investment of US$100 billion.

Objective

- Identify potential districts.
    - Solar Radiation
    - Electricity Rates
    - Per capita consumption
- Study the variation of production of a solar power plant and correlate it with weather.
- Predict future production of a particular plant using time series model.

Data Collection

Three types of datasets: -

- Data includes the name of power plants, capacity, location (district and state), latitude, longitude, date of commissioning, average domestic electricity rates in Rs./KWh, per capita electricity consumption in KWh, monthly solar radiation, the monthly plane of the array, district-wise population, all sky insolation.
- Time series data for some power plants containing power production for every 15 min for an year (for example, 1st January 2014 to 31st December 2014).

- The columns include the date, maximum temperature, minimum temperature, the average temperature of that day, wind chill, heat index, precipitation,snow depth, wind speed, wind gust, visibility, cloud cover, relative humidity, and weather condition

Datasets-

Unnamed: 0	Location (District)	Installed Capacity (MW)	Latitude	Longitude	Average domestic electricity rates in Rs./KWh	Per capita electricity consumption in KWh	Solar Rad Monthly	Solar Rad Annual	poa monthly	allsky_insolatio
0	Aalo	0.05	28.170120	94.798230	4.00	703.0	[3.999358654022217, 3.6590754985809326, 3.6333...	4.204289	[123.98011779785156, 102.45411682128906, 112.6...	[3.36, 3.46, 3.3 3.31, 3.2
1	Abali	0.01	30.458560	78.250090	4.00	703.0	[5.87696647644043, 5.861727237701416, 6.584075...	5.881289	[182.1859588623047, 164.12835693359375, 204.10...	[4.68, 4.69, 4. 4.68, 4.4
3	Adilabad	228.00	19.666670	78.533330	9.50	1896.0	[6.148787975311279, 6.589134216308594, 6.79743...	5.831581	[190.6124267578125, 184.49575805664062, 210.72...	[5.17, 5.07, 5.1 5.02, 4.9

This is Data.csv file

Some Snapshots of Datasets

	Name	Date time	Maximum Temperature	Minimum Temperature	Temperature	Heat Index	Precipitation	Wind Speed	Visibility	Cloud Cover	Relative Humidity	Conditions
0	Jagalur, KA, India	01-01-2014	26.4	17.1	21.2	NaN	0.0	7.6	6.3	30.0	67.49	Partially cloudy
1	Jagalur, KA, India	01-02-2014	26.2	17.7	21.4	NaN	0.0	7.6	6.3	43.8	64.43	Partially cloudy
2	Jagalur, KA, India	01-03-2014	28.7	16.7	22.3	27.7	0.0	9.4	6.3	31.3	57.97	Partially cloudy
3	Jagalur, KA, India	01-04-2014	29.4	17.4	22.3	27.9	0.0	5.4	7.0	25.0	52.18	Clear
4	Jagalur, KA, India	01-05-2014	30.1	18.4	23.2	28.3	0.0	9.4	7.0	17.5	48.39	Clear
5	Jagalur, KA, India	01-06-2014	30.4	17.4	23.9	28.6	0.0	5.4	6.3	10.0	46.38	Clear
6	Jagalur, KA, India	01-07-2014	29.7	19.9	24.0	28.4	0.0	9.4	7.0	2.5	44.46	Clear
7	Jagalur, KA,	01-08-	29.1	17.4	22.8	28.0	0.0	5.4	7.0	0.0	59.30	Clear

This is weather_data.csv

	index	time	Values
0	0	01-01-2014 00:00	0.000
1	1	01-01-2014 01:00	0.000
2	2	01-01-2014 02:00	0.000
3	3	01-01-2014 03:00	0.000
4	4	01-01-2014 04:00	0.000
5	5	01-01-2014 05:00	0.000
6	6	01-01-2014 06:00	24.172
7	7	01-01-2014 07:00	76.397
8	8	01-01-2014 08:00	122.814
9	9	01-01-2014 09:00	154.061
10	10	01-01-2014 10:00	163.537
11	11	01-01-2014 11:00	161.615
12	12	01-01-2014 12:00	152.405
13	13	01-01-2014 13:00	131.113

This is jagalur_hourlyv3.csv

Data Cleaning

- Removal of data for the plants which were not district identified.
- Correction in the name of some districts, whose data were repeating due to some manual mistakes in the dataset.
- District-wise grouping of the data.
- The raw dataset was very inconsistent, as there were some repetitive data and for some rows, the data was missing. So, we cleaned the dataset, grouped the data on an hourly basis, and arranged it in a day by day hourly manner.

Columns

For DATASET 1

- *Name:* Name of the Solar Park or Name of the Company owning the Solar Plant.
- *Installed Capacity (MW):* Capacity of the solar plant in Megawatts.
- *Location (District):* Corresponds to the district in which the plant is located.
- *Latitude & Longitude:* Latitude and longitude of the district, obtained through Geocoder.
- *State:* State of the plant in which it is located.
- *Date of Commissioning:* Date of approval of the plant by the government.
- *Average domestic electricity rates in Rs./KWh:* Average domestic electricity per unit rate state-wise.
- *Per capita electricity consumption:* State-wise per capita consumption of electricity in (kWh per capita).

Columns Contd.

- *Solar Rad Monthly:* Month-wise solar radiation in $kWh/m^2/day$ for each location of the already known solar plants. An array of values representing monthly data.
- *Solar Rad Annual:* Average of solar radiation annually in $kWh/m^2/day$ for each location of the already known solar plants.
- *Poa monthly:* Monthly average of the Plane of Array Irradiance in kWh/m^2. Plane of Array Irradiance is the irradiation coming from the sun and also including the diffusion and reflection components of the irradiation. An array of values representing monthly data.

  (Dataset1)

Columns Contd.

- *All sky insolation*: The average amount of solar radiation incident on a horizontal surface at the surface of the earth under all-sky conditions with the direct radiation from the sun's beam blocked by a shadow band or tracking disk.

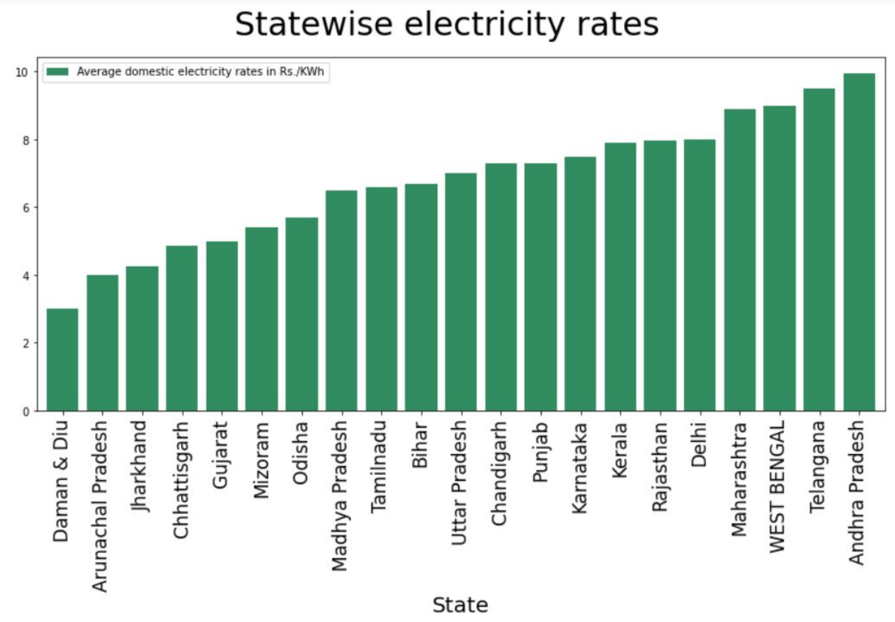
For DATASET 2

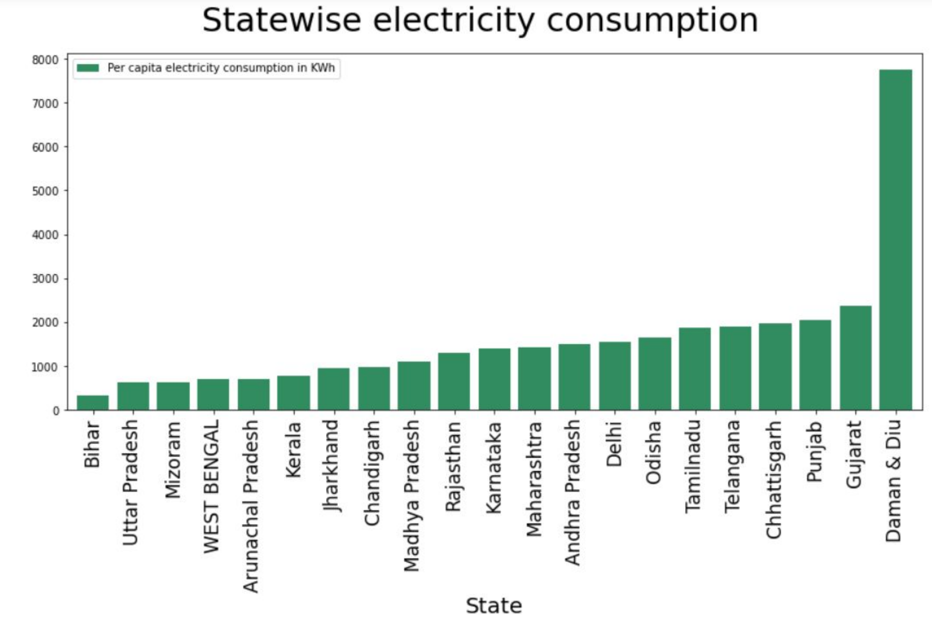
- *time:* Date and time(mm-dd-yyyy hh:mm) of data collected.
- *Values:* Solar energy produced in that hour (MW)

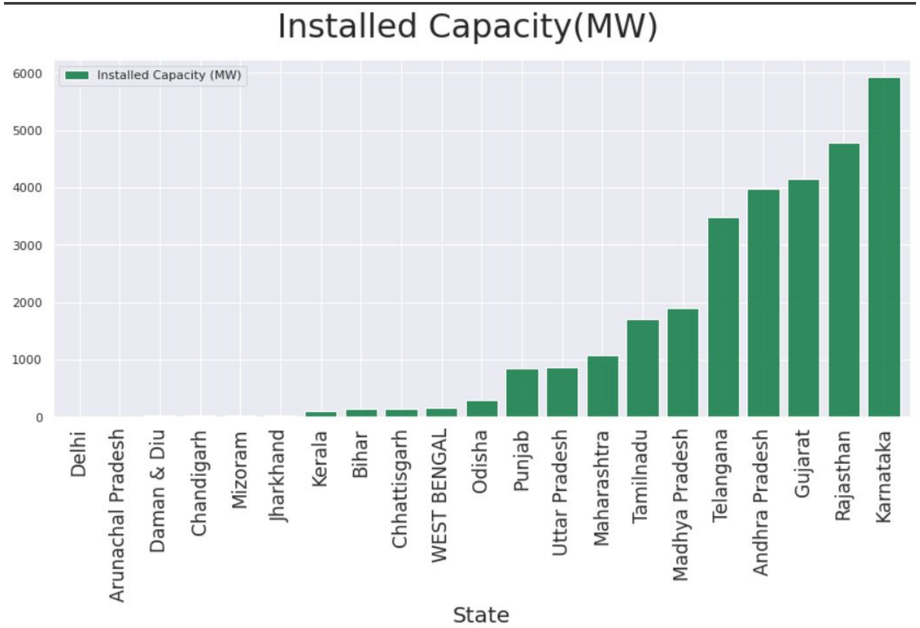
For DATASET 3

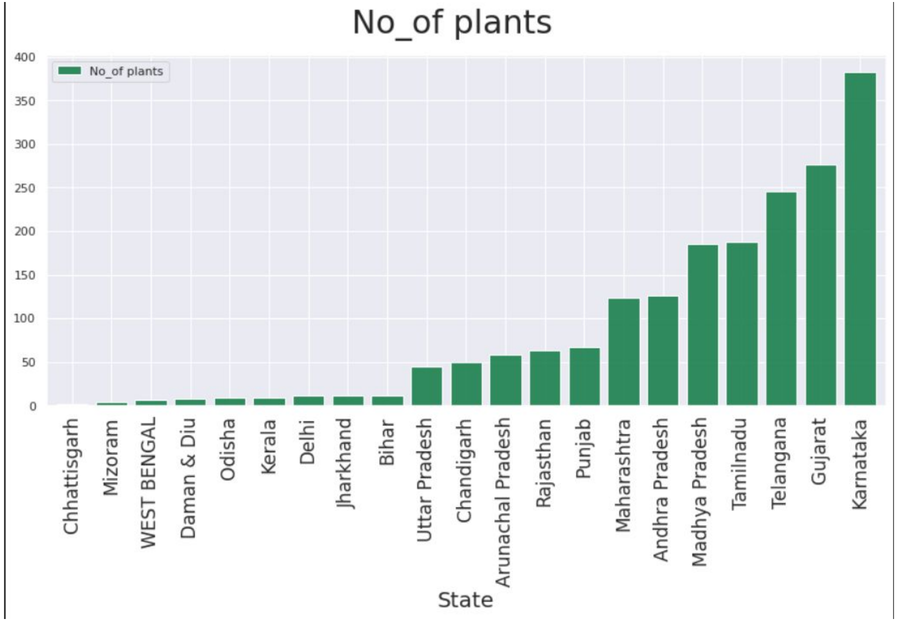
The columns include the date, max. temperature, min. temperature, the average temperature of that day, wind chill, heat index, precipitation, snow depth, wind speed, wind gust, visibility, cloud cover, relative humidity, and weather condition.

Some visualizations on the collected data



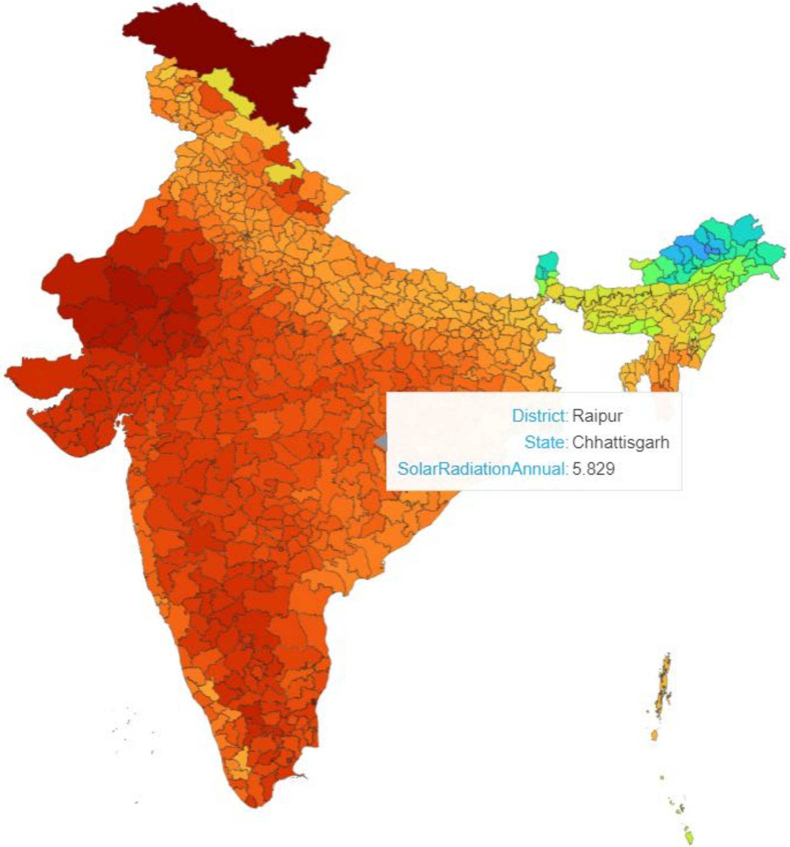




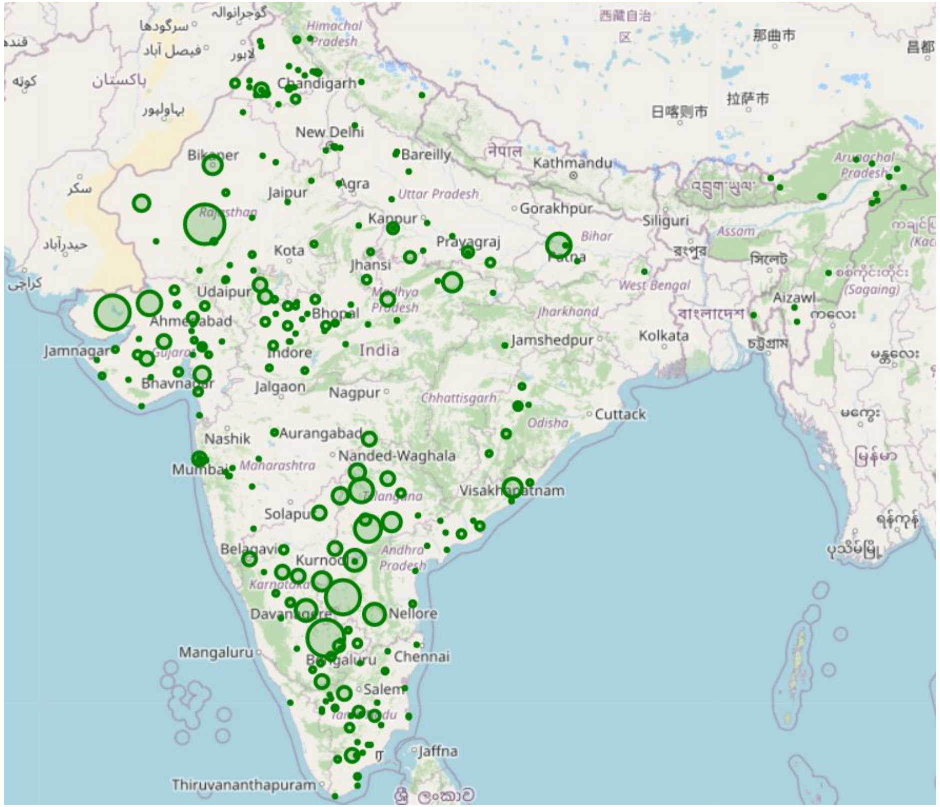


State-wise Installed capacity of Solar plants

State-wise number of Solar plants

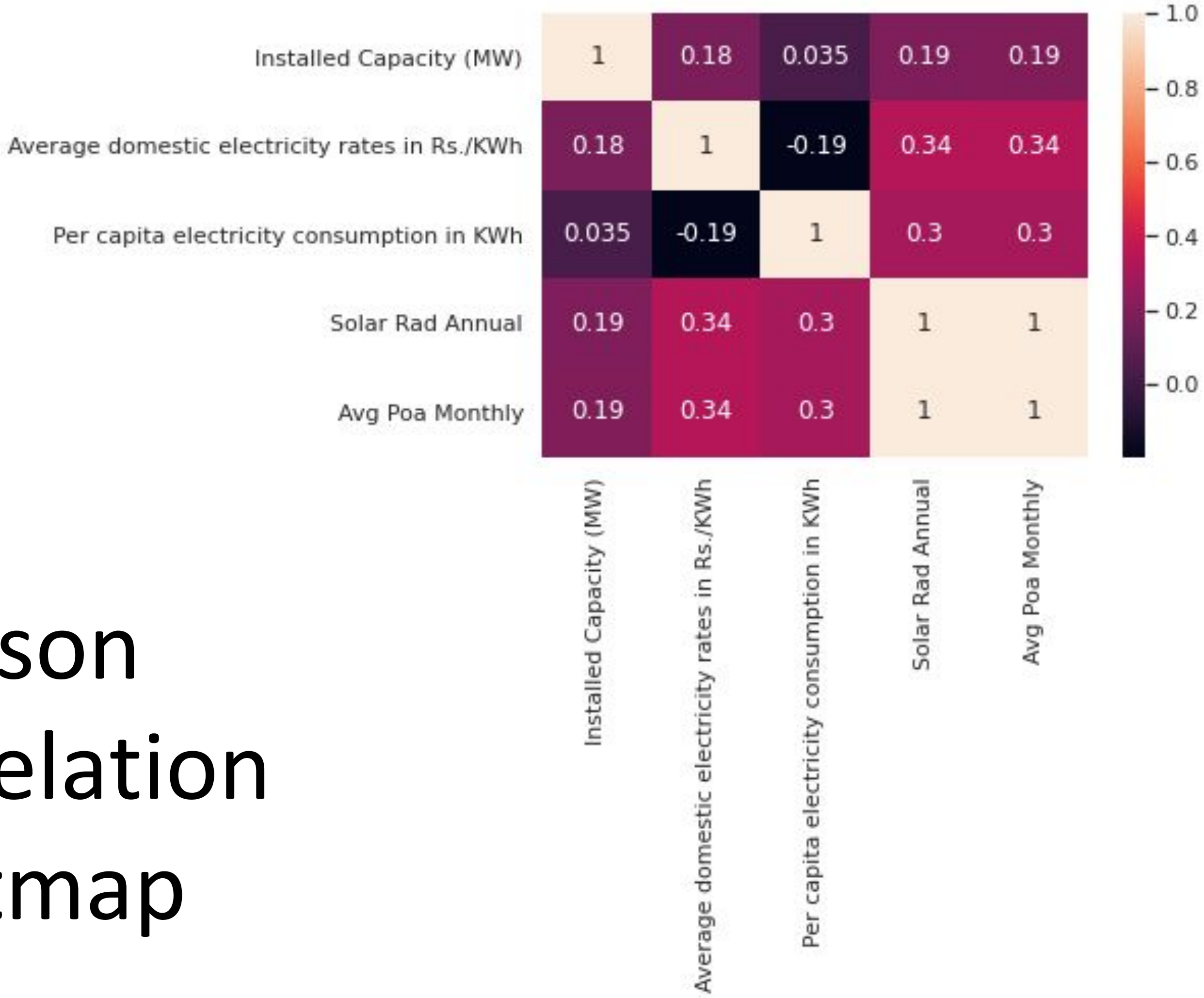


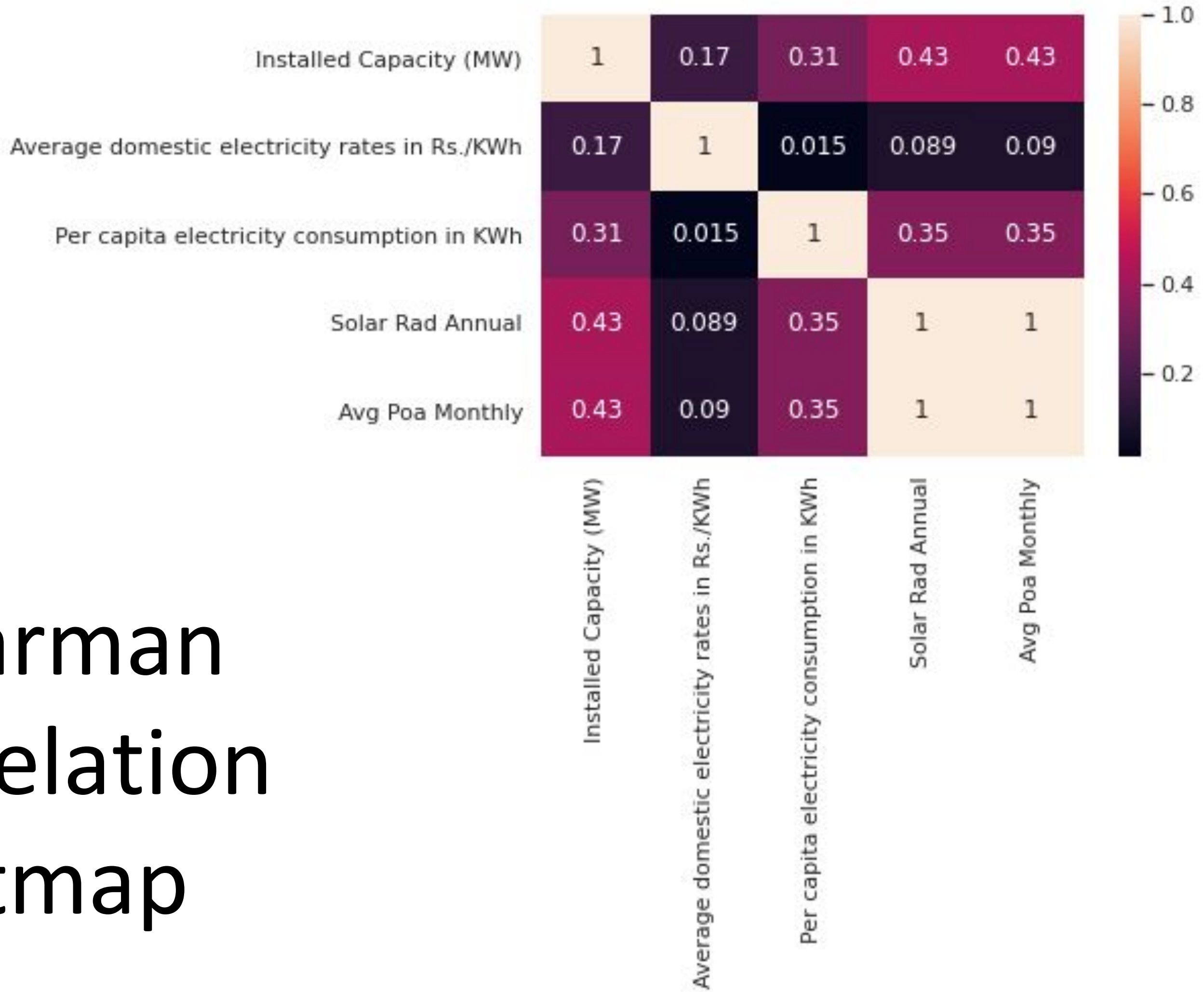
District-wise Solar Radiation in India



District-wise Installed capacity of Solar plants in India

Pearson Correlation Heatmap

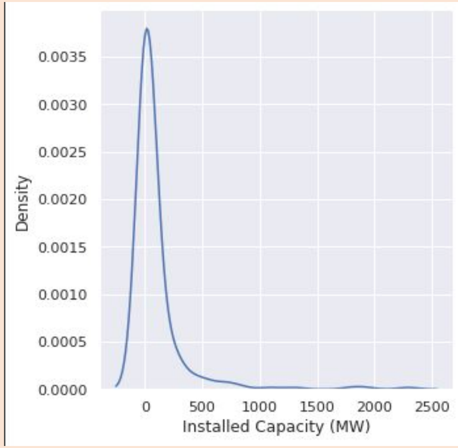


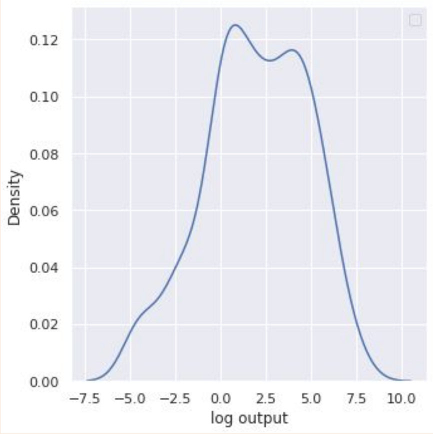


Spearman Correlation Heatmap

SOME STATISTICS

- The distribution of our dependent variable (Installed Capacity (MW)) is as follows

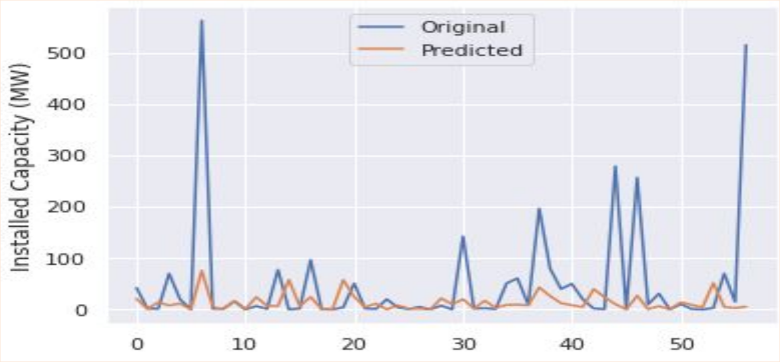




KstestResult(statistic=0.06510099672972569, pvalue=0.17386841155391364)

MODELLING - Regression

- MULTIVARIATE POLYNOMIAL REGRESSION
- Degree fitting best found to be 2
- Predicting the Installed Capacity(MW) using other features in the dataset



MODELLING - Regression

Generalized Linear Model

- Gaussian family with log link for sm.GLM (as our distribution was lognormal for the target variable)
- R2-score :- 0.296 (after 5 fold validation)

```
==============================================================================
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  163
Model:                            GLM   Df Residuals:                      157
Model Family:                Gaussian   Df Model:                            5
Link Function:                    log   Scale:                          44719.
Method:                          IRLS   Log-Likelihood:                -1101.0
Date:                Tue, 17 Nov 2020   Deviance:                   7.0208e+06
Time:                        02:27:22   Pearson chi2:                 7.02e+06
No. Iterations:                    44
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        -22.0273      9.175     -2.401      0.016     -40.009      -4.045
x1             0.2930      0.246      1.189      0.234      -0.190       0.776
x2             4.0222      2.883      1.395      0.163      -1.629       9.673
x3          1926.6959    738.359      2.609      0.009     479.539    3373.853
x4         -1922.1957    735.743     -2.613      0.009   -3364.225    -480.166
x5            18.6009      5.285      3.520      0.000       8.243      28.958
==============================================================================
```

MODELLING - Regression

Random Forest Regressor

- Using random forest regressor we were able to obtain an R2 score of 0.38 with 5 fold validation.
- This was our best regression metric, however, it was still pretty low.
- Hence we can see that regression techniques don’t really fit very well on our dataset, maybe because there are several more features involved in predicting Installed Capacity and more reliable data sources to get solar radiation.

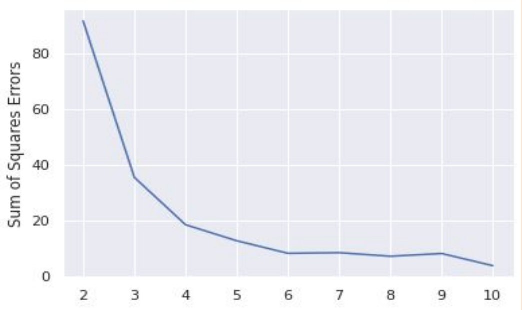
MODELLING - Classification Based Approach

- Used a classification based approach where we assigned labels to each district based on Installed Capacity (MW)
- Using these labels we trained a classification model to assign class for unknown districts.
- Intuitively the capacity of Districts was divided into small, medium and large plants.
- For assigning class labels we used two approaches k-means clustering and assigning the labels based on thresholds

MODELLING - Classification

K-means Clustering

- Found Optimal Clusters to be 3 - Elbow Method and Silhouette Score
- But one class had very less data points so we merged them to make two classes.
- Not suitable as number of classes was not equally distributed



MODELLING - Classification

Labels Based on threshold

- After trial and error we used the following thresholds

  We used these labels for classification algorithms for improved accuracy

Capacity	Label	Count
Capacity < 5 MW	0	124
5 MW =< Capacity < 100	1	101
100 =< Capacity	2	58

MODELLING - Classification Based Approach

LOGISTIC REGRESSION CV

- Using the labels assigned we created a multi-class classification model using LogisticRegressionCV with cross validation 5
- For districts with no plants, model will assign labels and based on the label we will tell the potential of the district
- For this, we obtained an accuracy of around 0.55.

MODELLING - Classification Based Approach

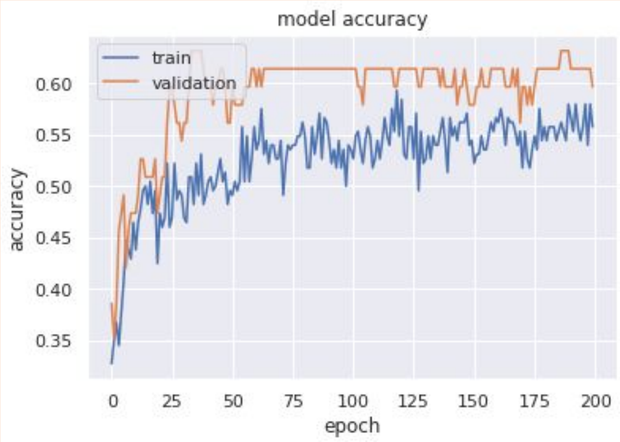
Simple Neural Network for classification

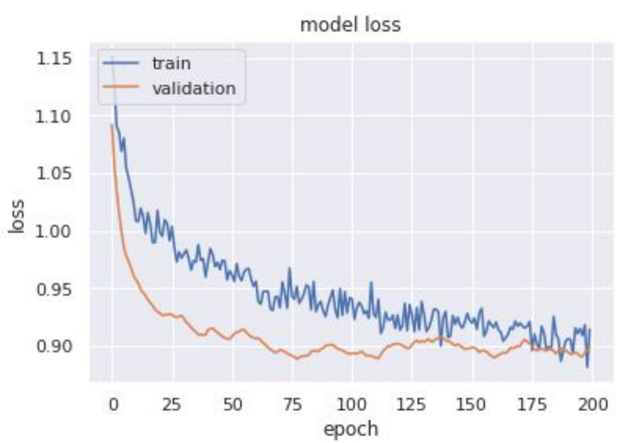
- Model Architecture was as follows
- We used "relu" activation and for the last layer softmax. We used Dropout to tackle overfitting which we first encountered. We used the optimizer as Adam and the loss function as categorical cross-entropy.

```
model = Sequential()
model.add(Dense(16, input_dim=5, activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(12, activation='relu'))
model.add(Dense(3, activation='softmax'))
```

MODELLING - Classification Based Approach

Accuracy and Loss :





MODELLING - Classification Based Approach

- Metrics for the neural network

Accuracy is: 66.66666666666666

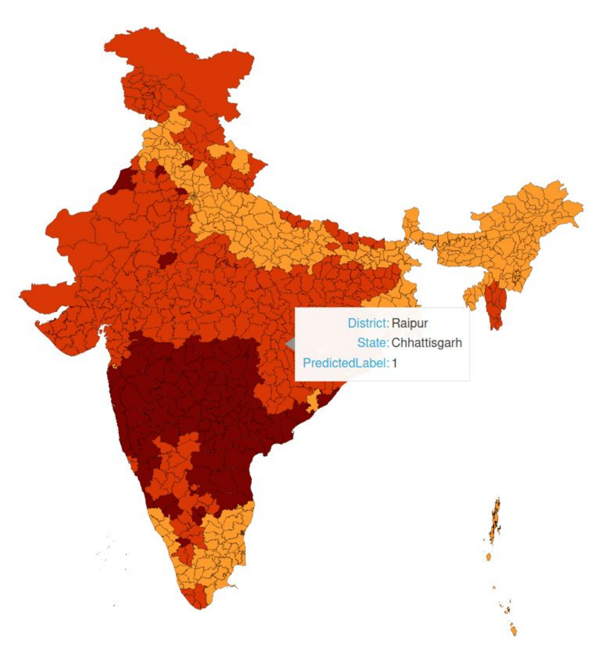
Micro F1 score is: 0.6666666666666666

Macro F1 score is: 0.6075036075036075

- We used this model as the final model to assign labels to districts

Predicted Label for Districts

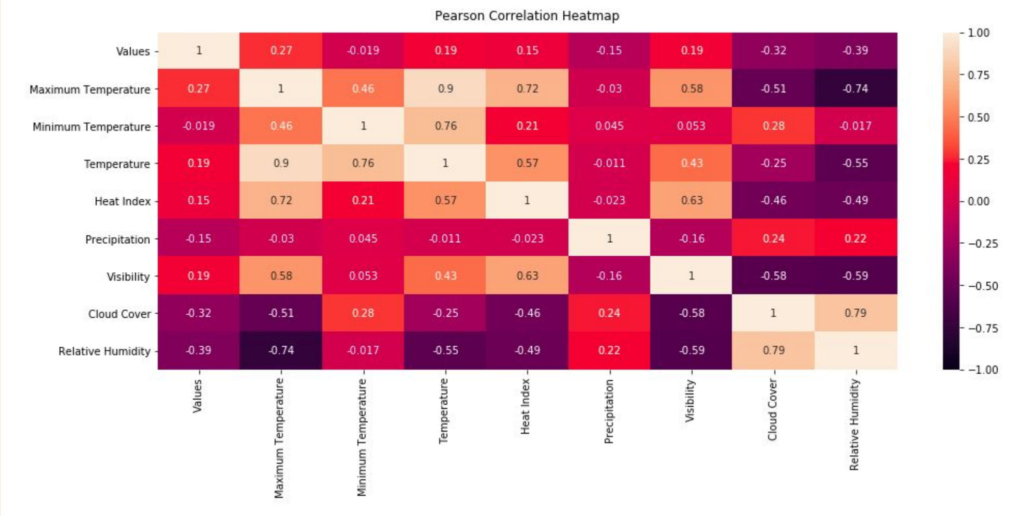
- Label 0: Orange (Less than 5 MW)
- Label 1: Red (Between 5 MW to 100 MW)
- Label 2: Dark red (Greater than 100 MW)

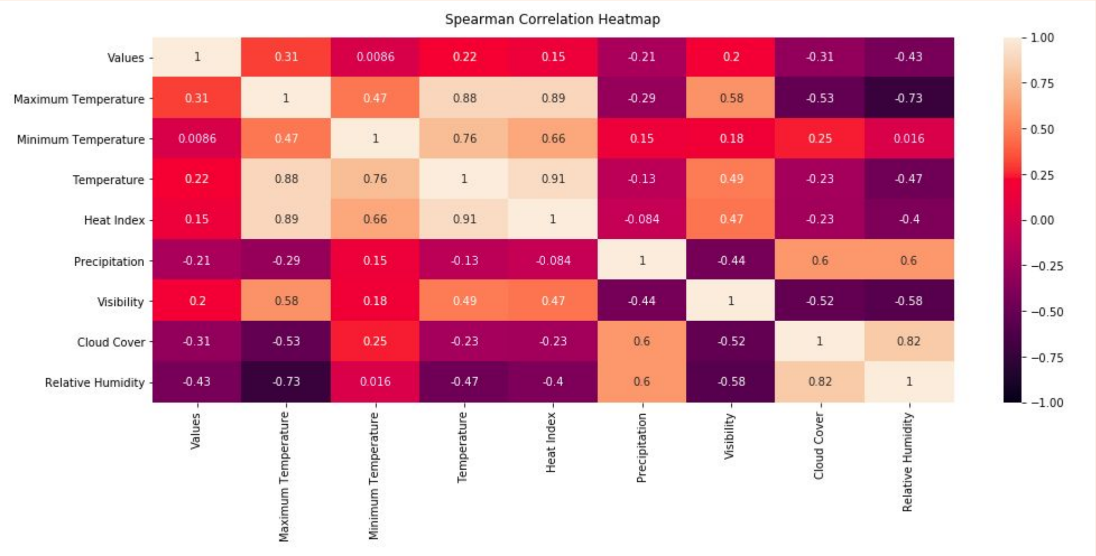


Weather Correlation

Correlated the day-wise production of one particular solar power plant (Dataset 3) with the weather.

Pearson Correlation Heatmap





Spearman Correlation Heatmap

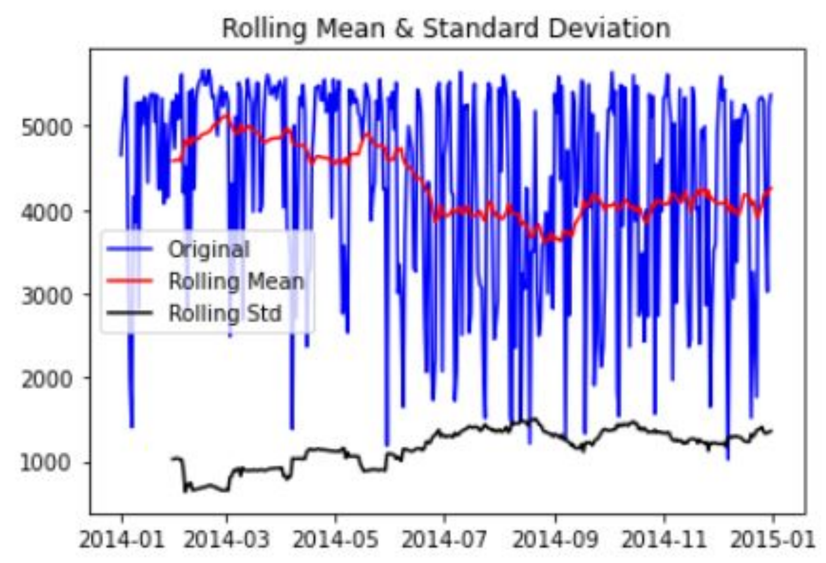
ARIMA model

Before we can build a model, we must ensure that the time series is stationary. We used the two primary ways to determine whether a given time series is stationary.

1. **Rolling Statistics**
2. **Augmented Dickey-Fuller Test**

ARIMA model

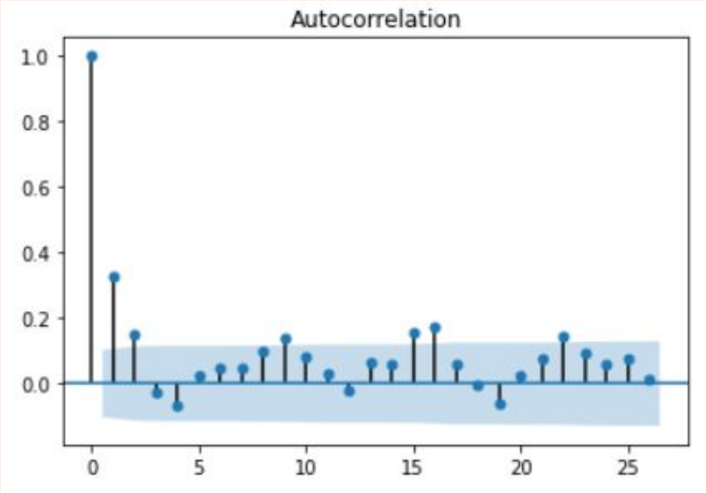
We can see very low p-value, therefore it will be safe to say that the time series is stationary and hence it is relevant to apply ARIMA model.

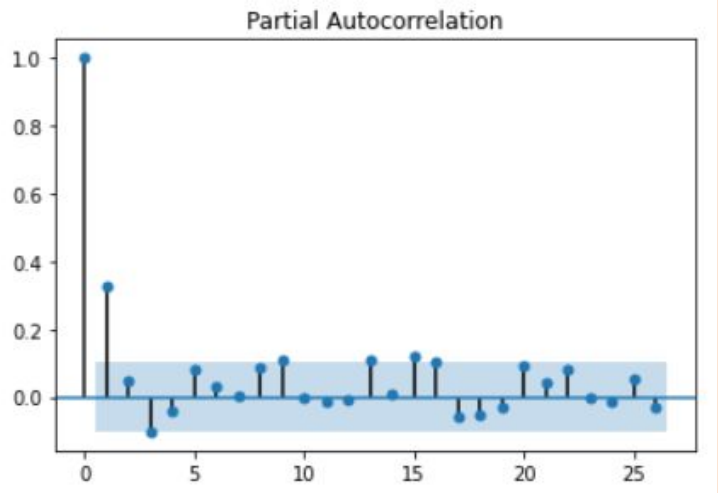



```
ADF Statistic: -13.548132607769483
p-value: 2.4388770970713854e-25
Critical Values:
	1%: -3.4484434475193777
	5%: -2.869513170510808
	10%: -2.571017574266393
```

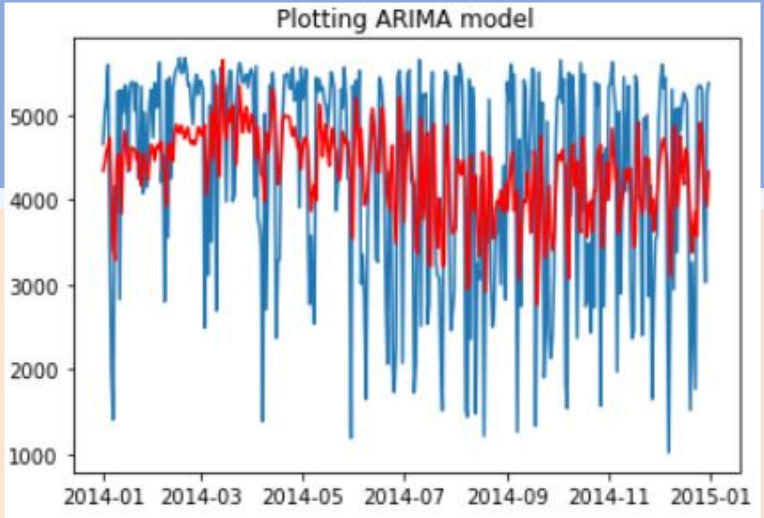
ARIMA model

Now we use ACF and PACF to figure out the best order of the ARIMA model.





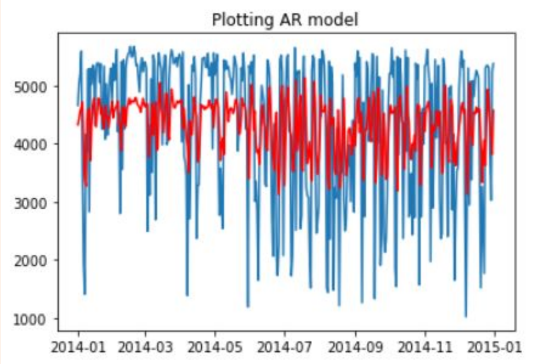
ARIMA model

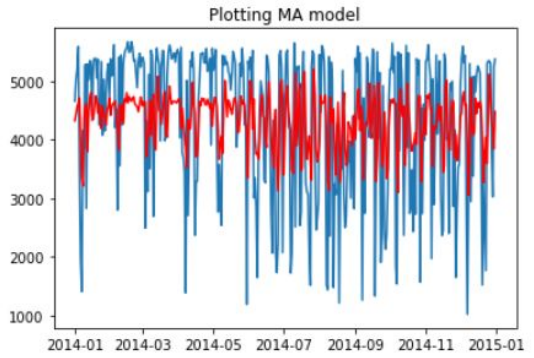


We get the value of p = 6, d = 0, q = 5

ARIMA model is the combination of **Auto Regression (AR)** and **Moving Average (MA)**

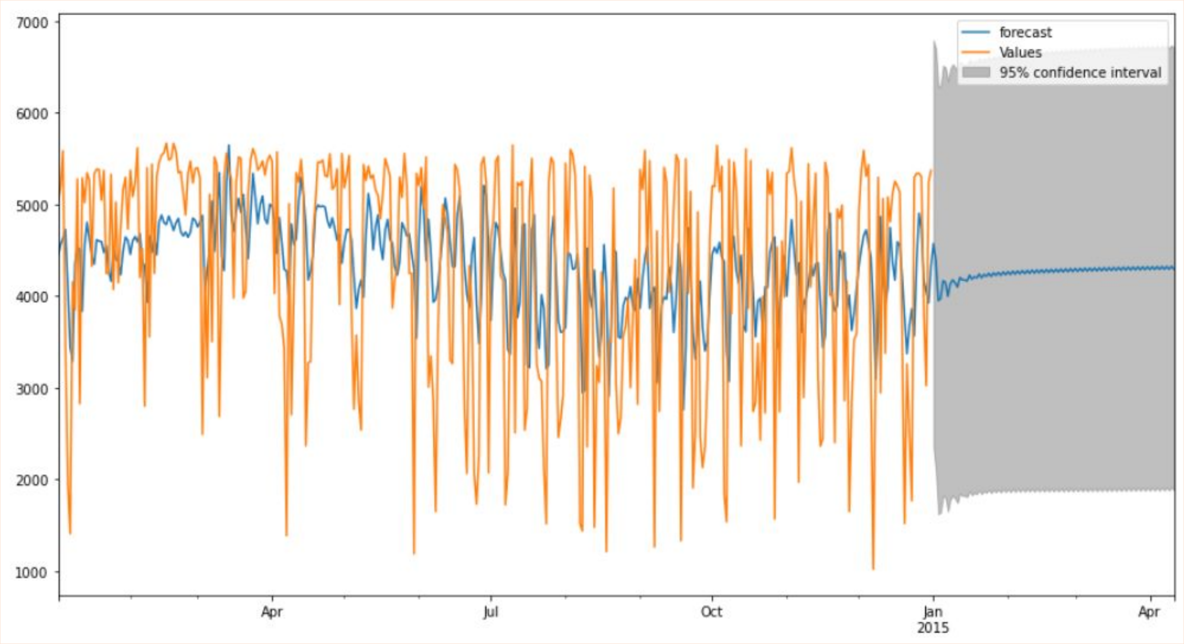
Note: The red line is predicted and the blue line is the actual value.





ARIMA model

The final output of the ARIMA model prediction for the next 100 days
i.e. 1st January 2015 to 10th April 2015



Final Results

- Making precise predictions for the value of Installed Capacity (MW) using regression techniques is not suitable.
- Classifiers to assign labels to the district works better.
- Performed correlation with the weather for day-wise production.
- Made simple ARIMA model to estimate the production.

Future Scopes

- Collecting more data for adding more input features for improving the prediction of Installed Capacity (MW) in the regression models
- Optimizing the classifiers further for improved results
- Trying out more Time Series models for better estimation of the day-wise production of solar power plant
- Collecting and analyzing production data from more power plants to generalize the time series model.

Thank You