****

**National Level Online Hack-A-Thon**

**On**

**Sustainable Energy**

**DETAILED REPORT**

**SUBMITTED BY:**

**VIT/OW/55**

**TEAM MEMBERS:**

**Sarthak Kiran Sarode**

**(18bee1074)**

**Ashutosh Sanjay Lembhe**

**(18bee1025)**

**DATE: 08/01/2022**

**PROBLEM STATEMENT:**

**1. OLAP operation of the data in front end (dice, slice, roll up/ down, filter)**

**2. Ability to notify significant changes in the time series dataset imported in the tool**

**3. Ability to select from and to time stamp in the time series visualization and give a label or annotation**

**4. Annotation tool - Data labelling where the customer can import the data and the multiple columns render it in the chart where they can select from all and to frame and labelled the part and save in the database.**

**5. Exploratory data analysis- Where they can explore the data and find its relationship with the different parameters.**

**6. Prediction Analysis/ modelling - Where they can pass the data and application has to automatically select which model is best and show it’s all the model accuracy results.**

**7. Have a dashboard to display aggregated values.**

**8. Results should be in pictorial representation.**

**9. Data cleaning/Data sanitisation must be done (Should not have null values).**

**10. Working video of the application is expected.**

**SOFTWARES USED:**

**LIBRARIES USED:**

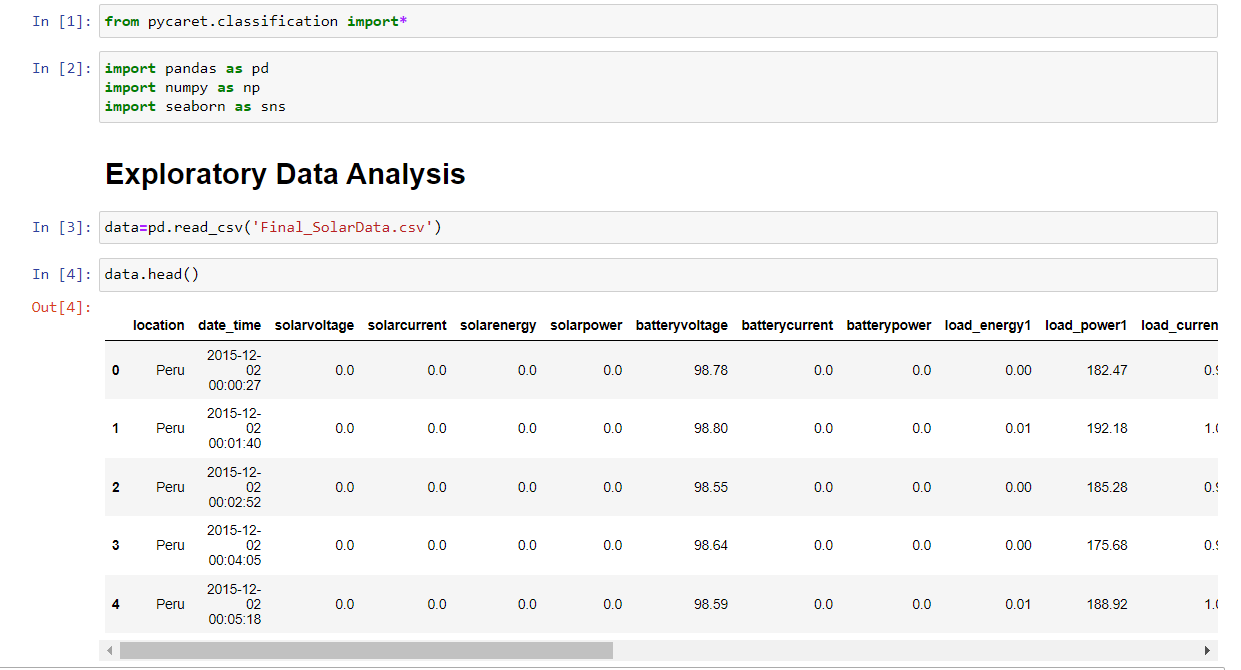
**EXPLORATORY ANALYSIS ON SOLAR DATASET**

**ENERGY EFFICIENCY PREDICTION BY AUTOML USING PYCART WITH A REGRESSION**

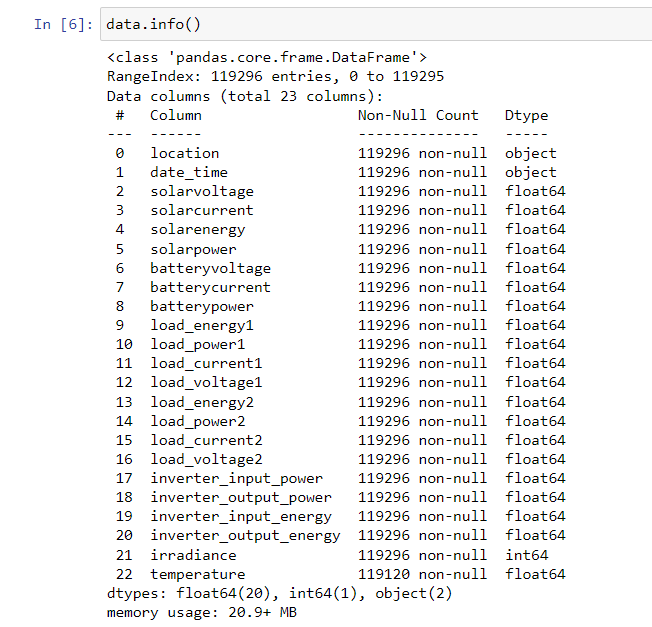
**(Sub problem statement 5, 9 covered)**

**Introduction:** We have taken the solar dataset and we are going to develop the regression machine learning model for the **temperature parameter**. As the solar panel’s output mainly depends upon the temperature.

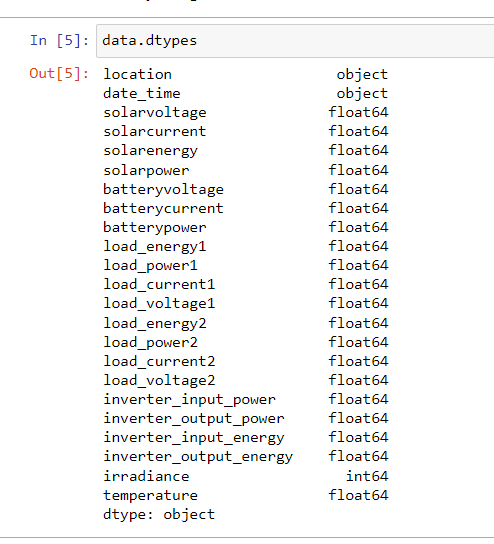
1) Importing the essential libraries and importing data set in python jupyter notebook.



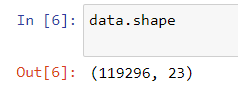
Basic Analysis of checking the datatypes and info of the dataframe



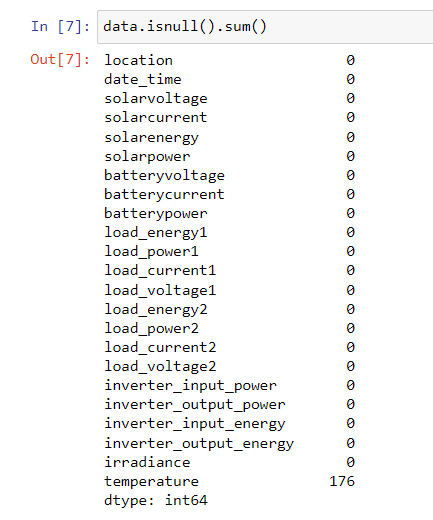
Here the data type of date\_time column is object we are supposed to convert it into standard format which is done ahead.

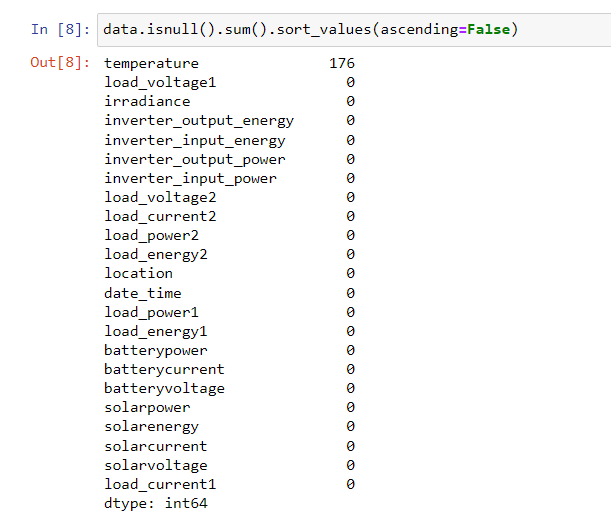


Shape of the dataframe before data cleaning i.e missing values cleaning

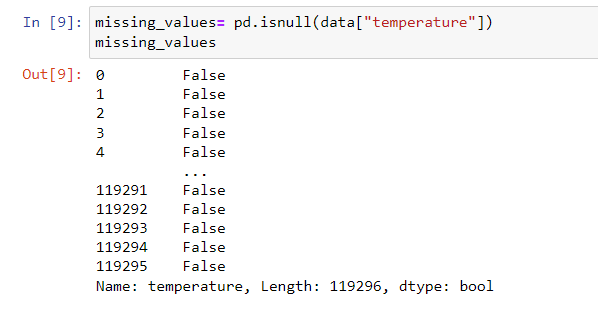


2) Missing values cleaning and Data Sanitization (9 th sub problem statement)

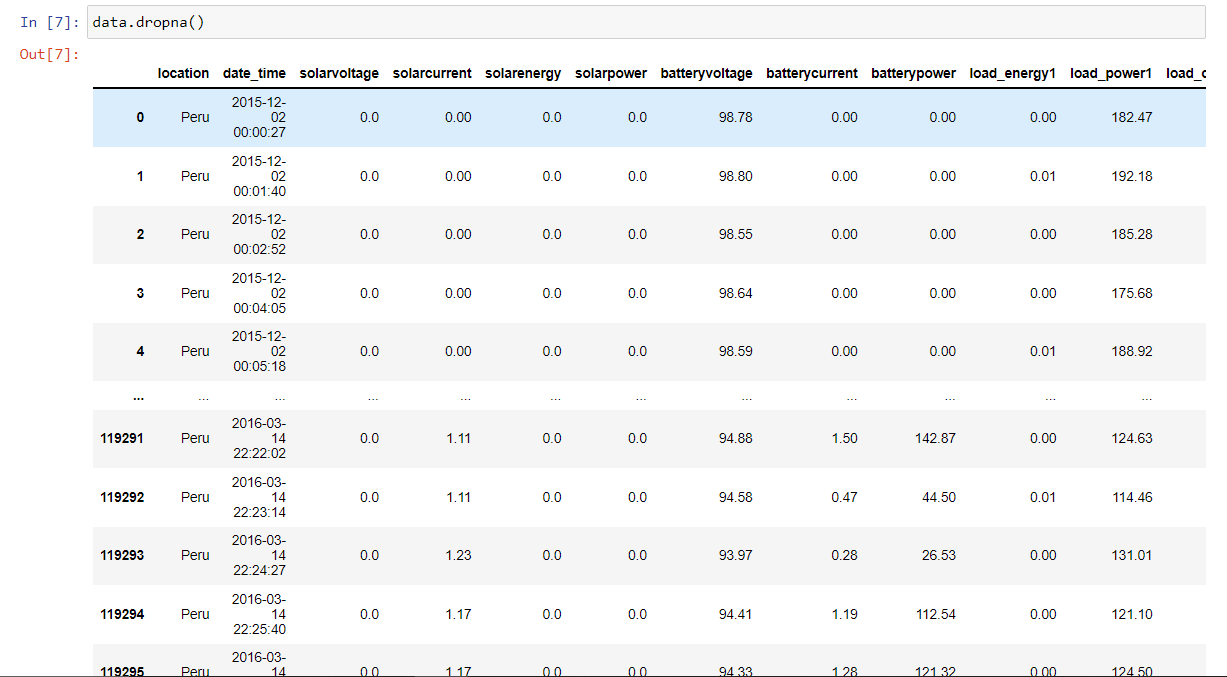




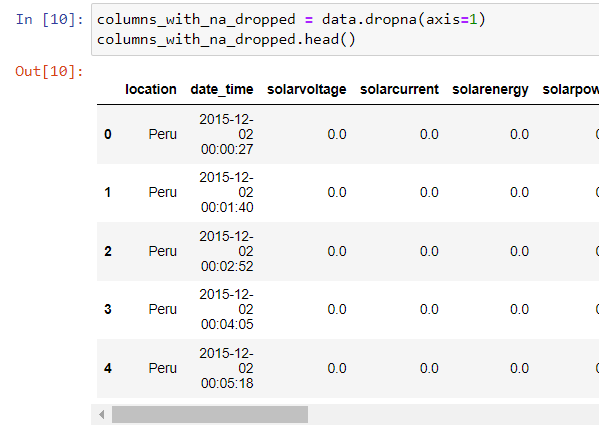
Temperature has some missing values. Checking which row of temperature has missing value.



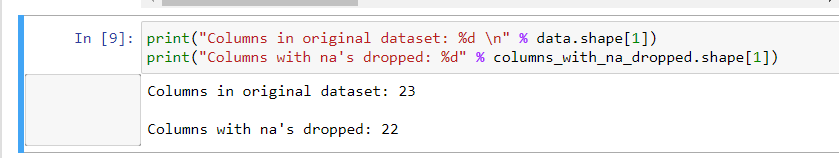
“False” means no missing values. Dropping rows having missing values.



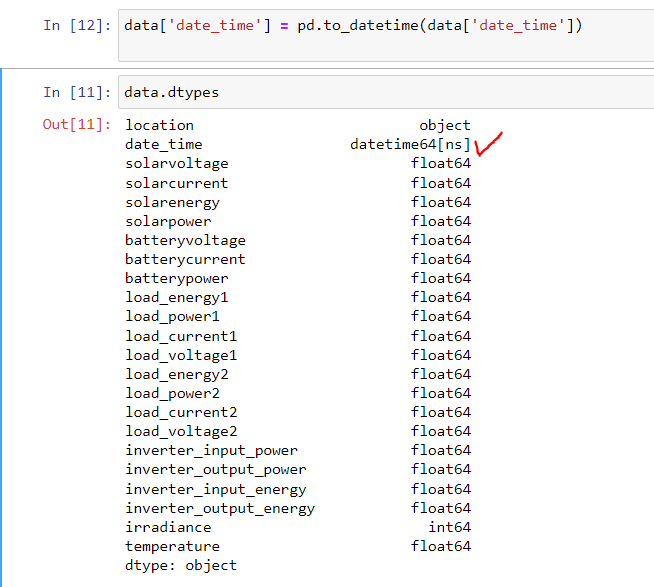
Deleting the columns which has missing values.



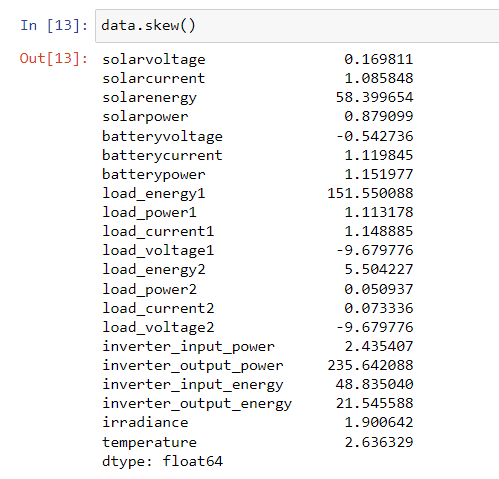
Final conclusion on cleaning the missing value.

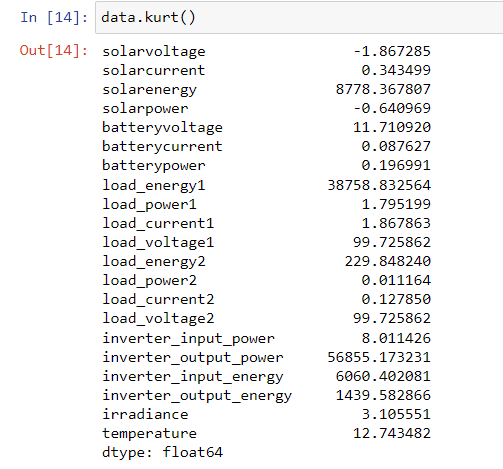


3) Date datatype converting to standard format.



4) Statistical Analysis of the dataset.Checking the skewness and the kurtosis of the dataset.



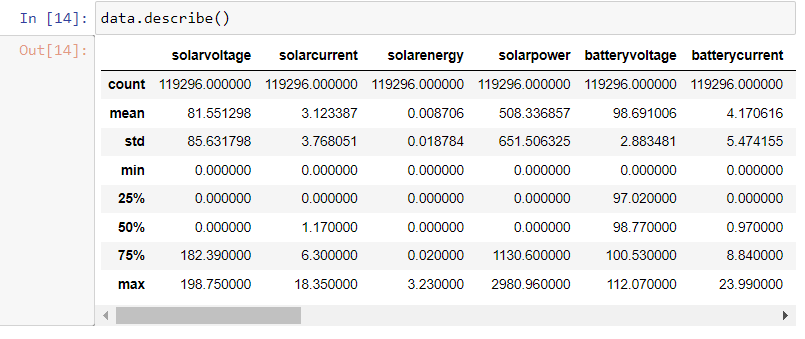


If skewness is less than -1 or greater than 1, the distribution is highly skewed. If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed. If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

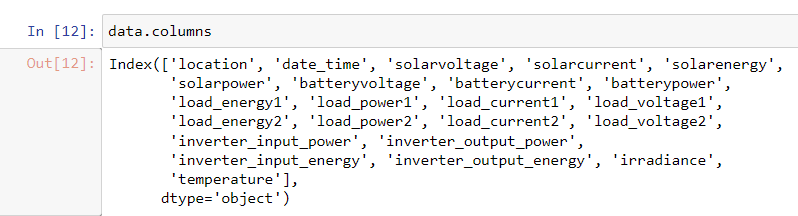
Low kurtosis in a data set is an indicator that data has light tails or lack of outliers. ... The peak is lower and broader than Mesokurtic, which means that data are light-tailed or lack of outliers. The reason for this is because the extreme values are less than that of the normal distribution.

**Anyways, we have omitted the skewness and outliers while developing the setup model for regression which will be explained later.**

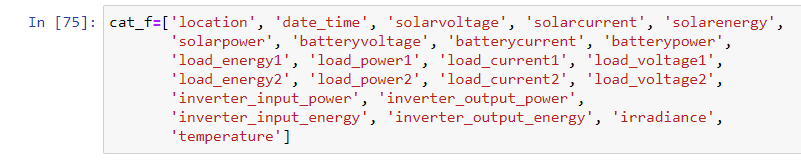
Checking the statistical data like mean, standard deviation of the entire dataset



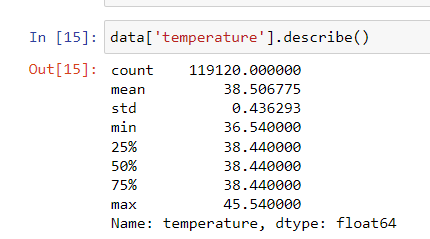
Printing all the columns name.

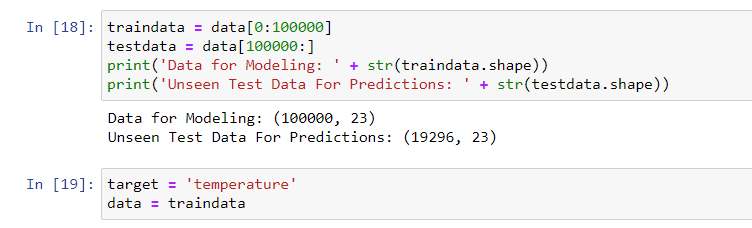


Storing all the columns inside the cat\_F variable which will be used later for the analysis.

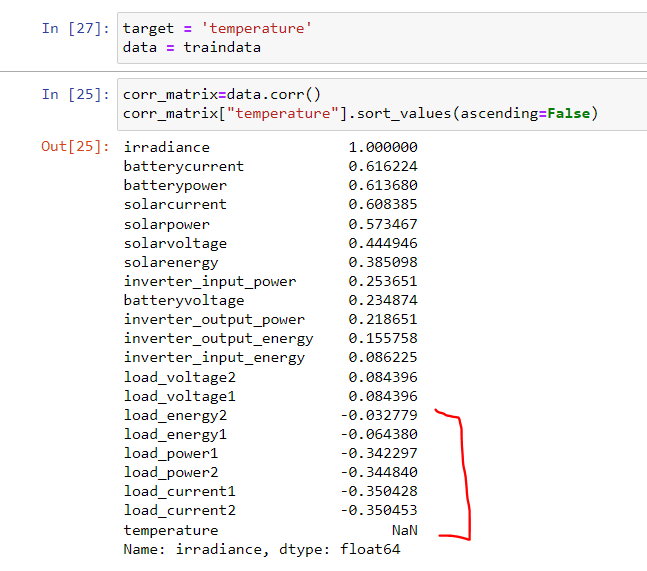


As the energy efficiency depends upon the temperature model and we are going to establish the regression relationship between temperature and other parameters**. Let’s train the data and before that split in the test and train data set.**





5) Checking the correlation of the temperature on the other parameters. The parameters which have negative correlation in heatmap will be ignore in the regression model development.



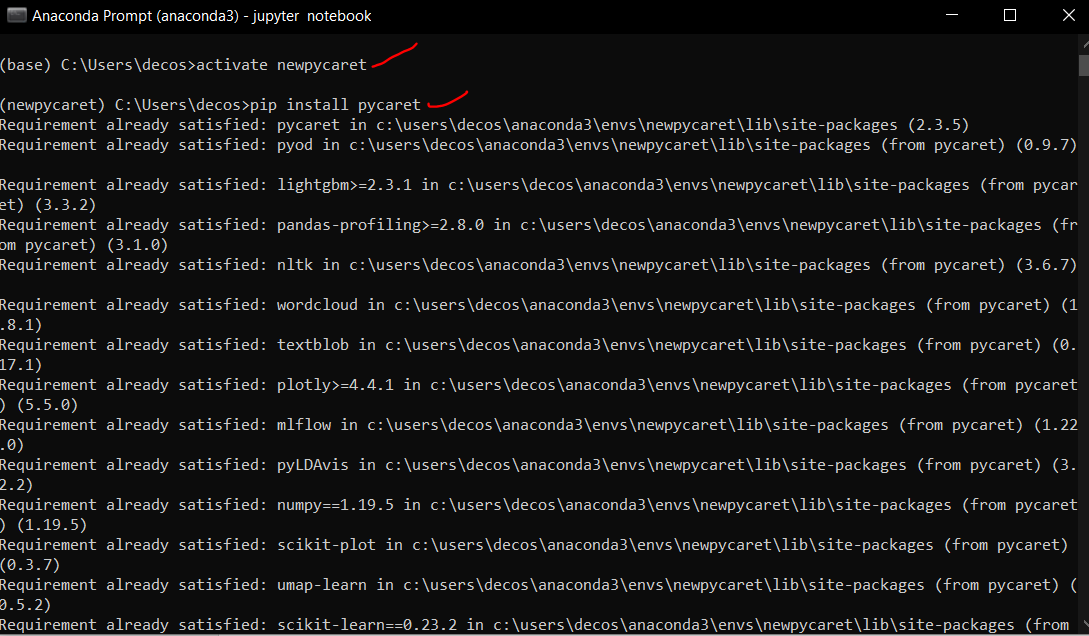


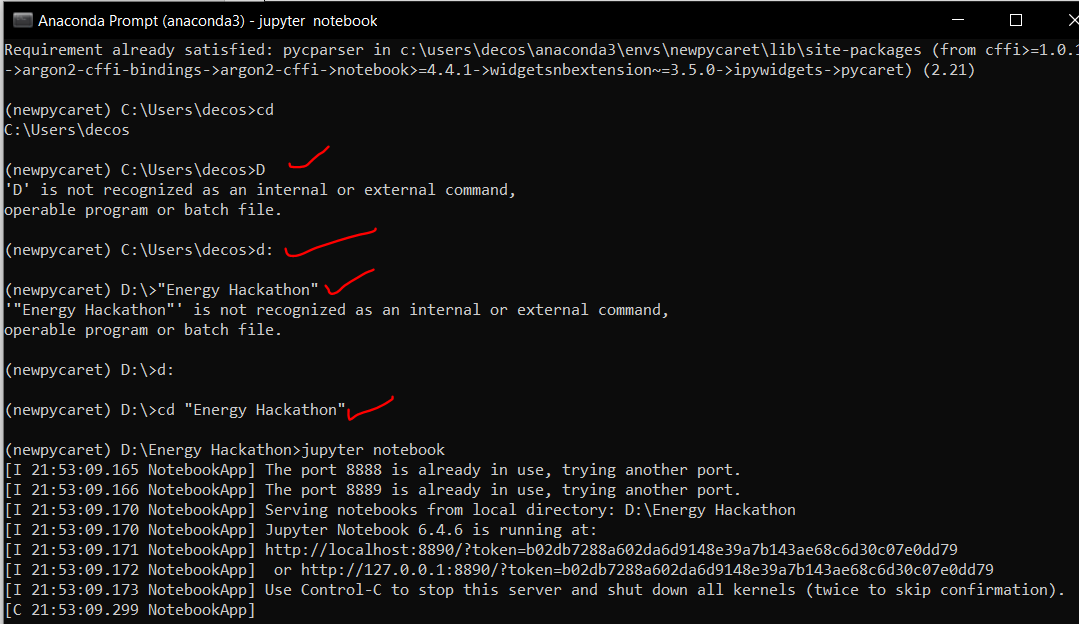
**PREDICTIVE ANALYSIS ON SOLAR DATASET**

**ENERGY EFFICIENCY PREDICTION BY AUTOML USING PYCART WITH A REGRESSION**

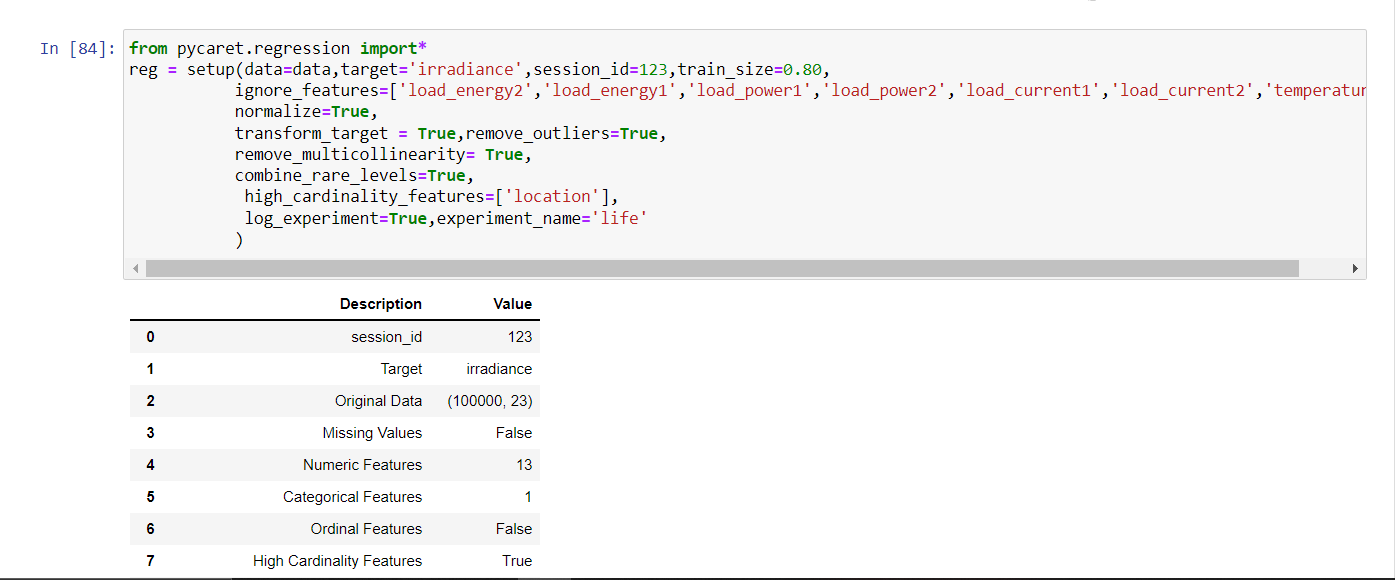
**(Sub problem statement 6,8 covered)**

6) **Setting up the environment:** Setting up an environment in Caret to run regression models hassle-free. Installing the picrate library in anaconda prompt in D drive folder and then Jupiter notebook is launched from there.

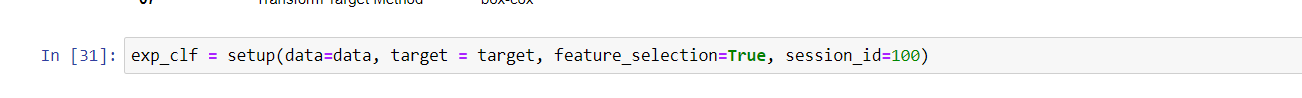
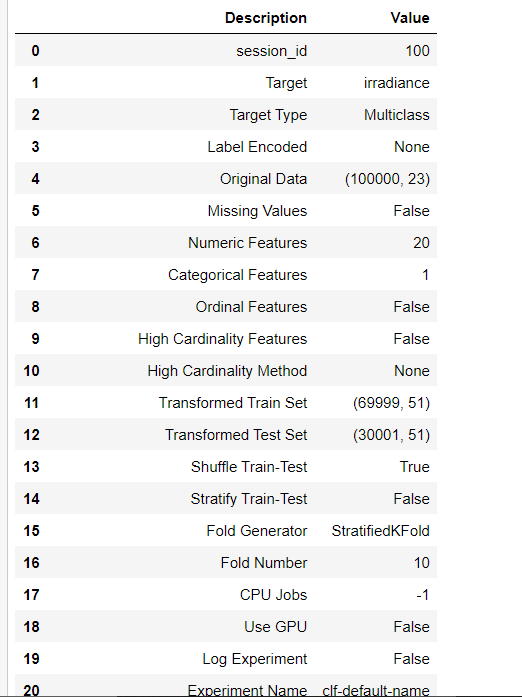
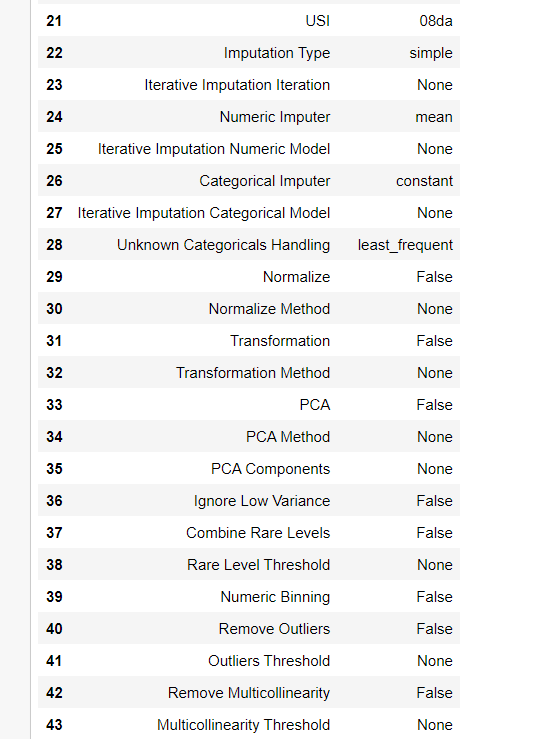


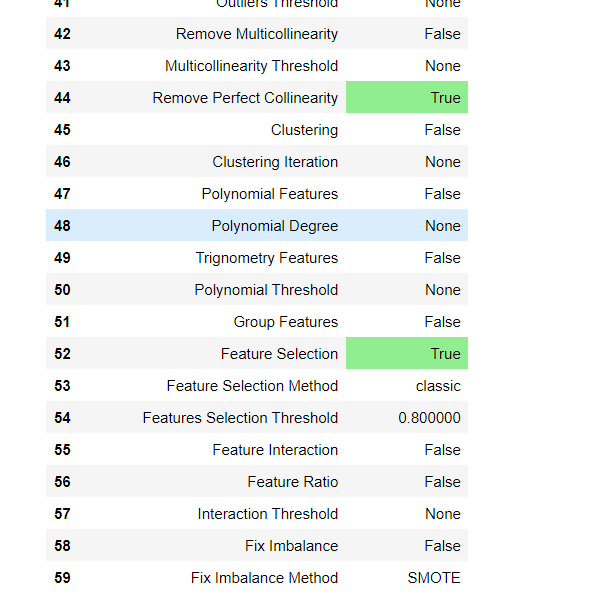


7) Setup of the Regression model



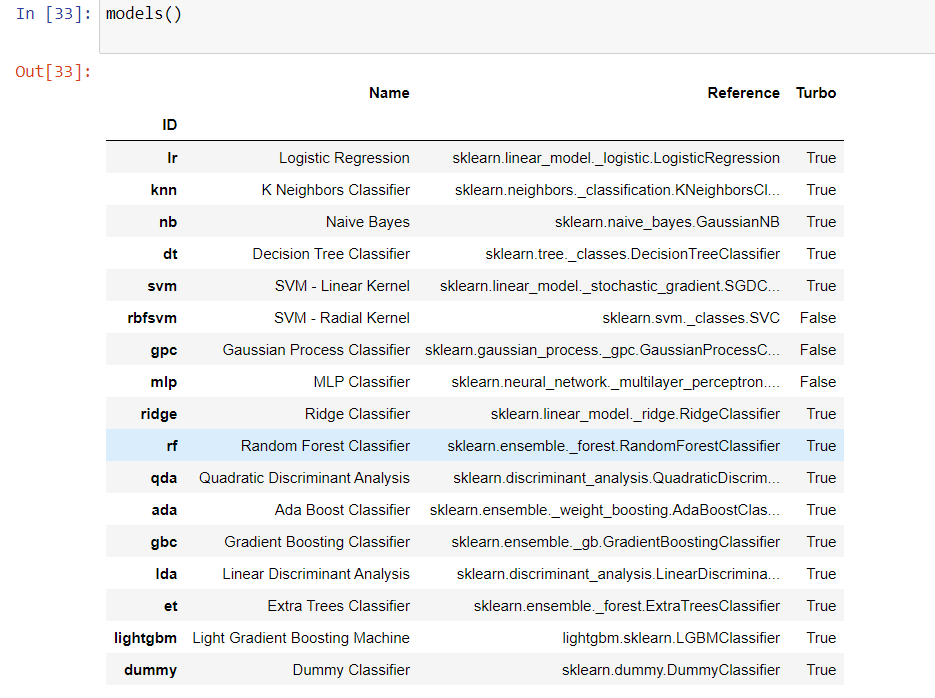
We have removed the **outliers and skewness** here. The columns which have less correlation are also ignored.



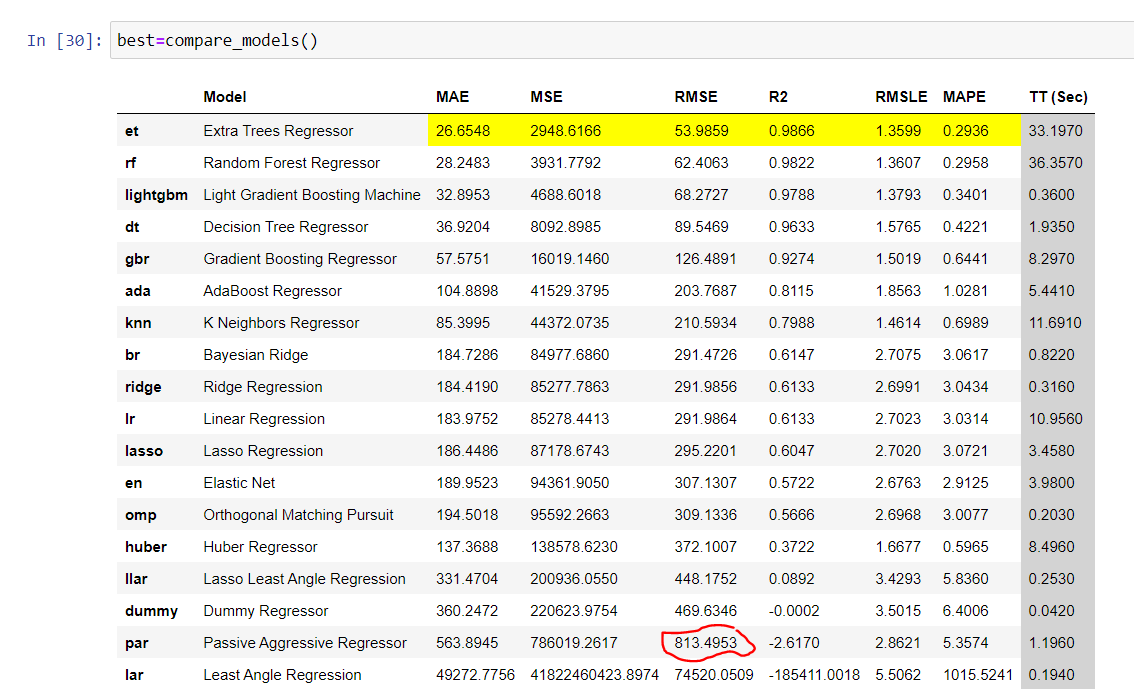


8) Comparing all models and picking up the best one which has maximum **R2 score.**

Checking all the machine learning models available for the regression.



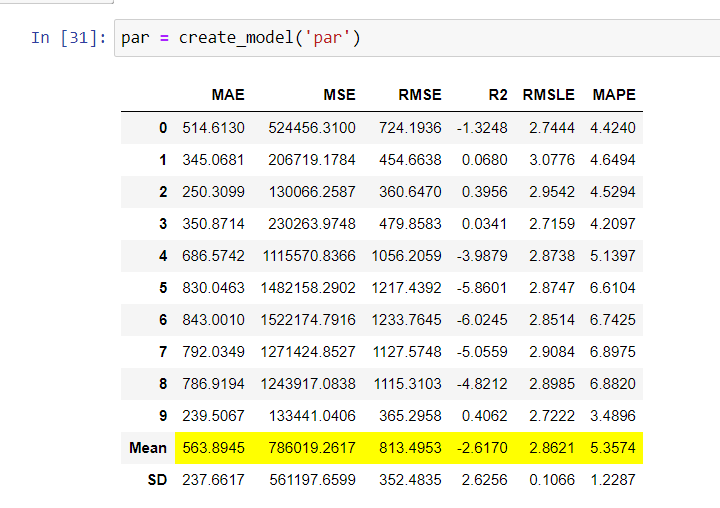
Comparing the models and selecting the best



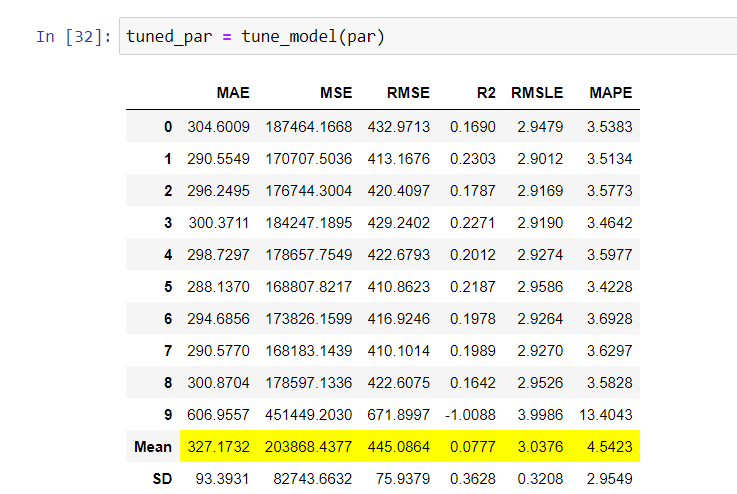
We select par model for its significant RMSE value.

**9) Create model:**

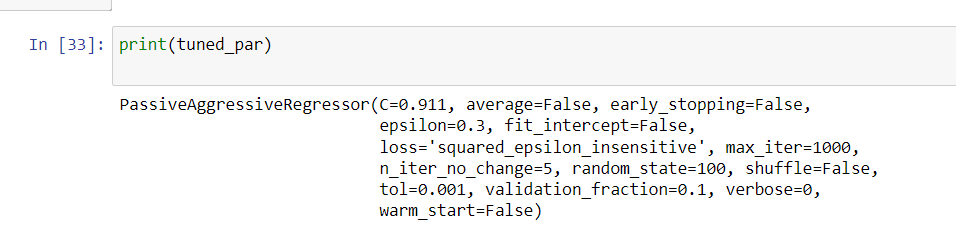
We selected par model of regression

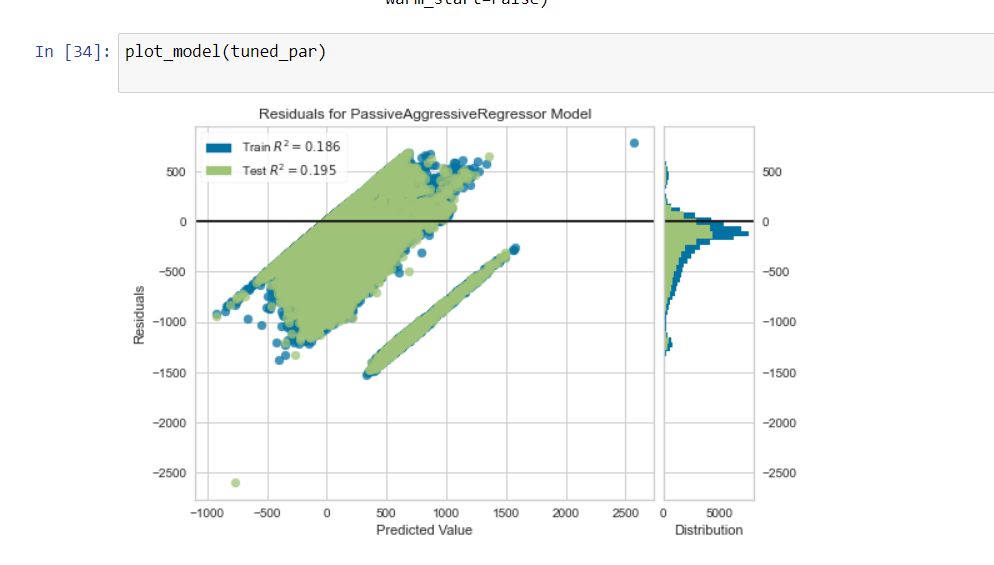


**10) Tune Model**: Automatically tuning the hyperparameters of a regression model.

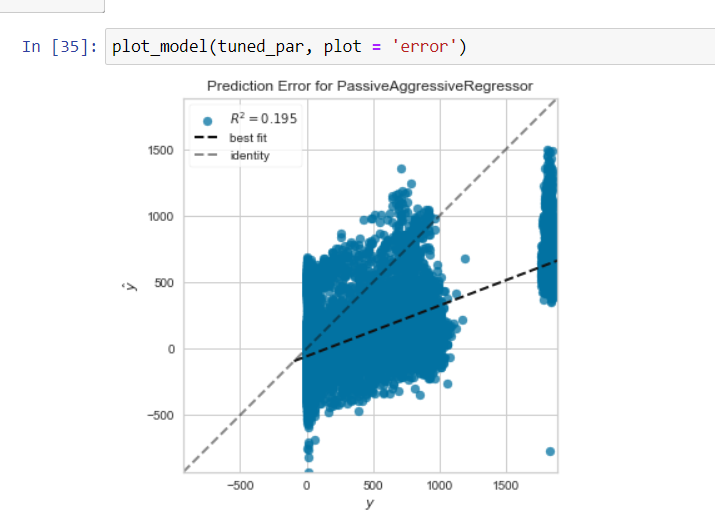


**11) Plot Model & Results** : plotting the performance of various models.

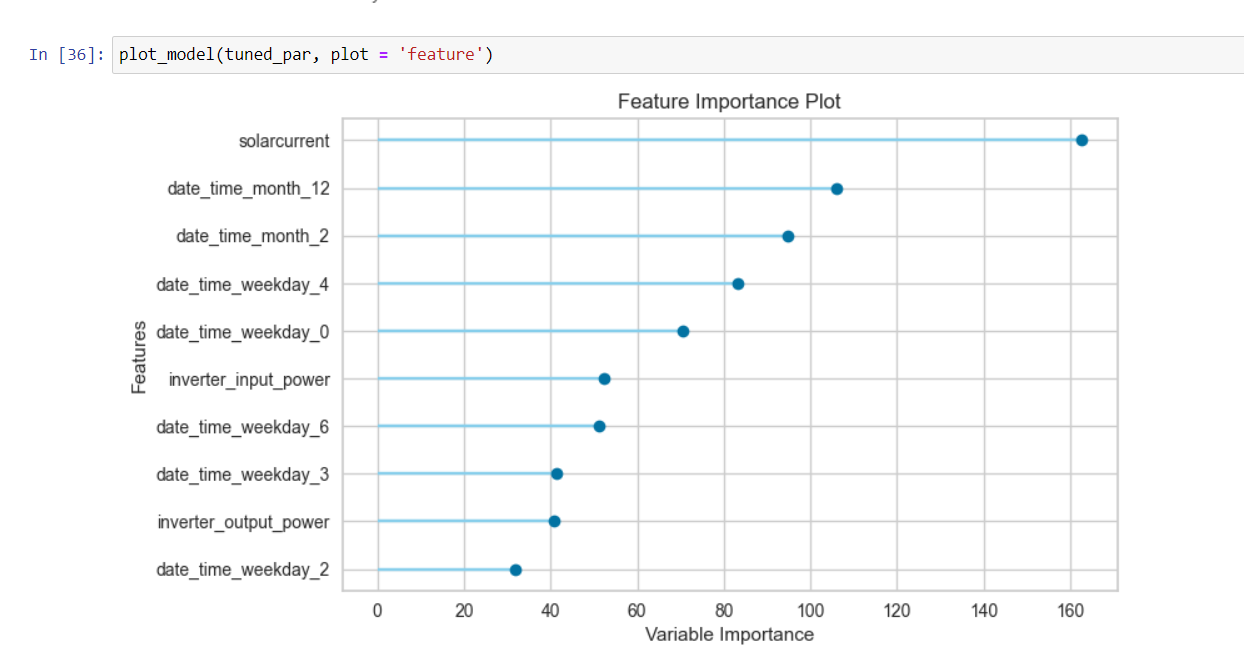


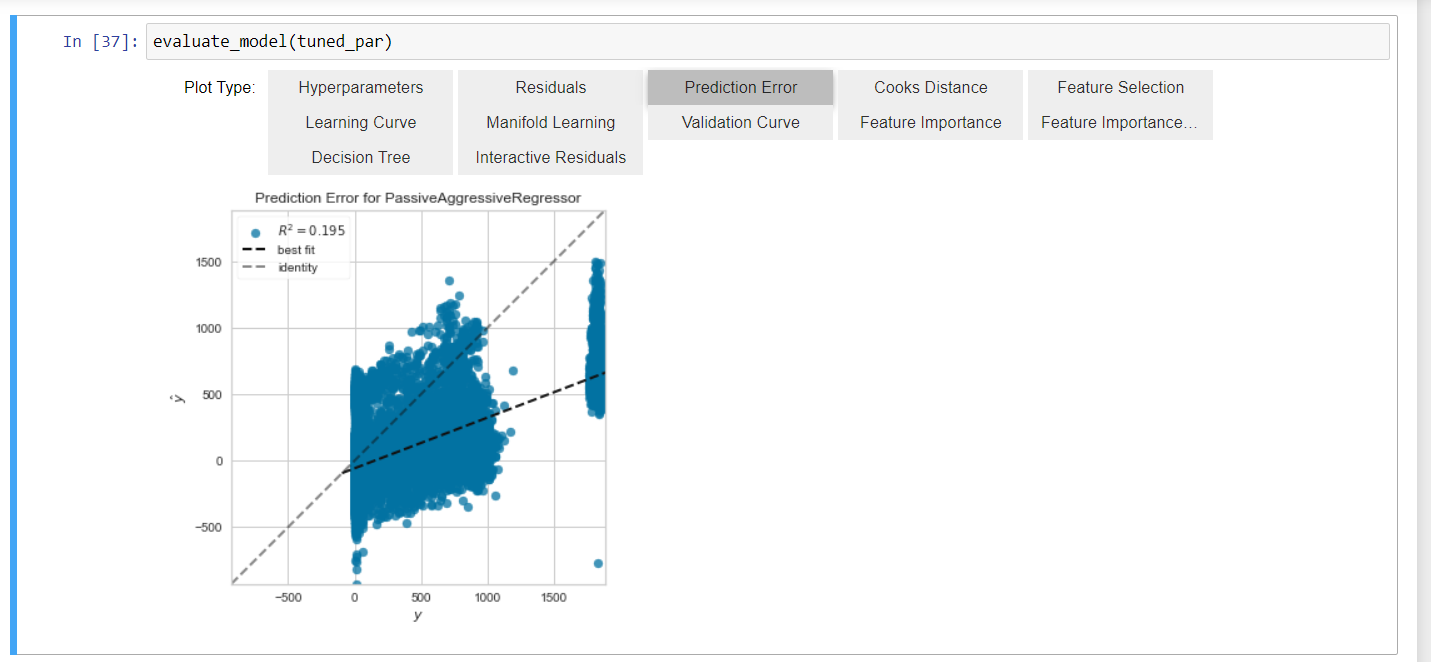


Prediction error plot:

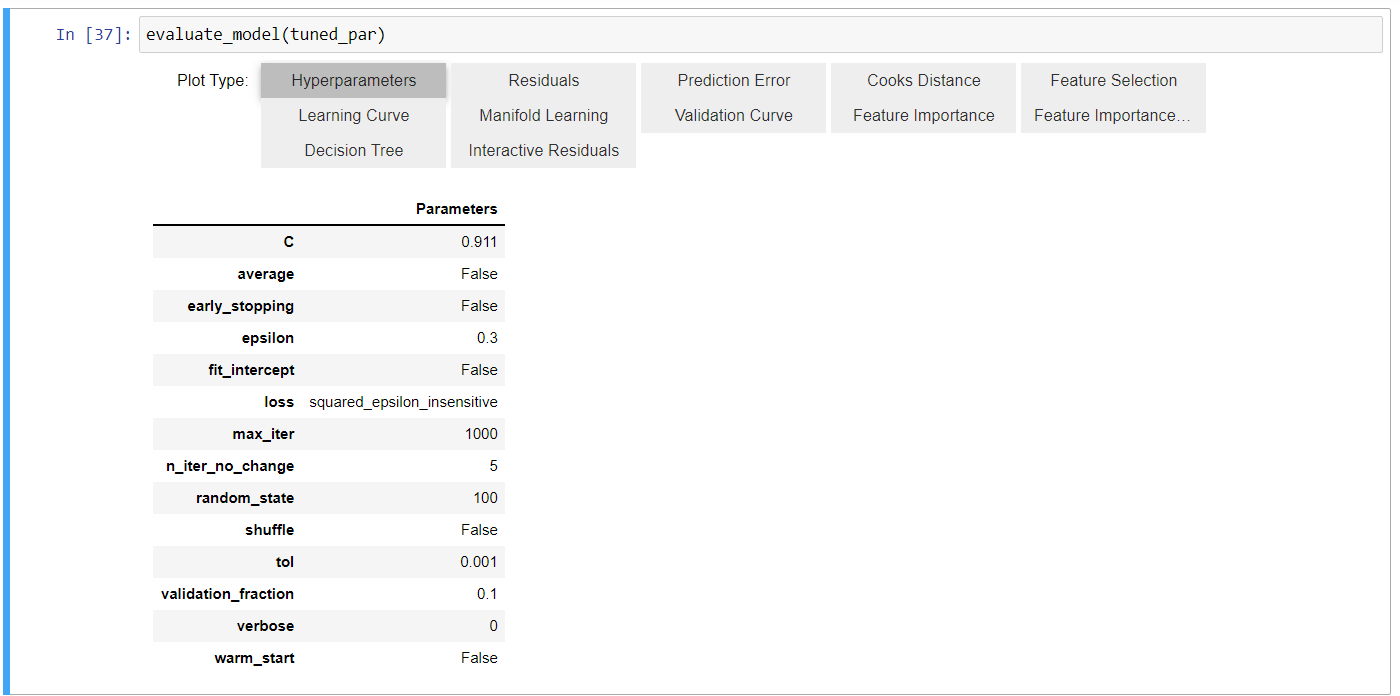


Feature importance plot:

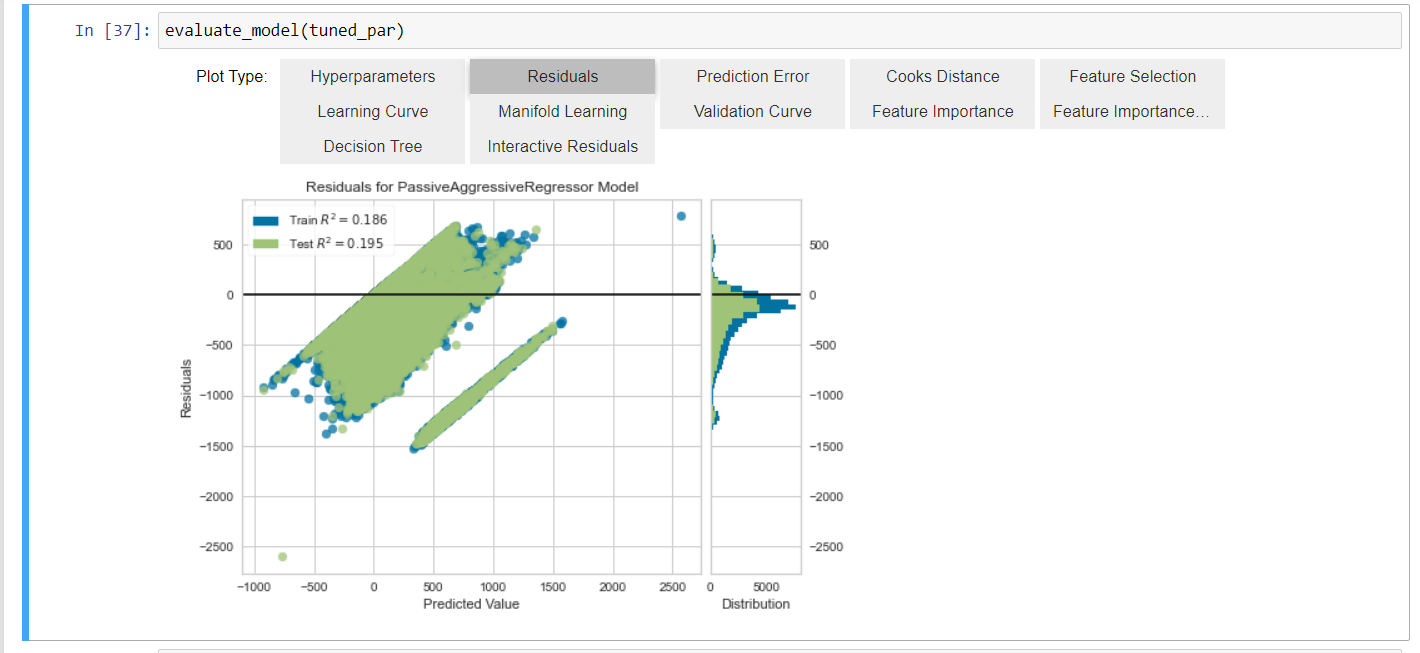




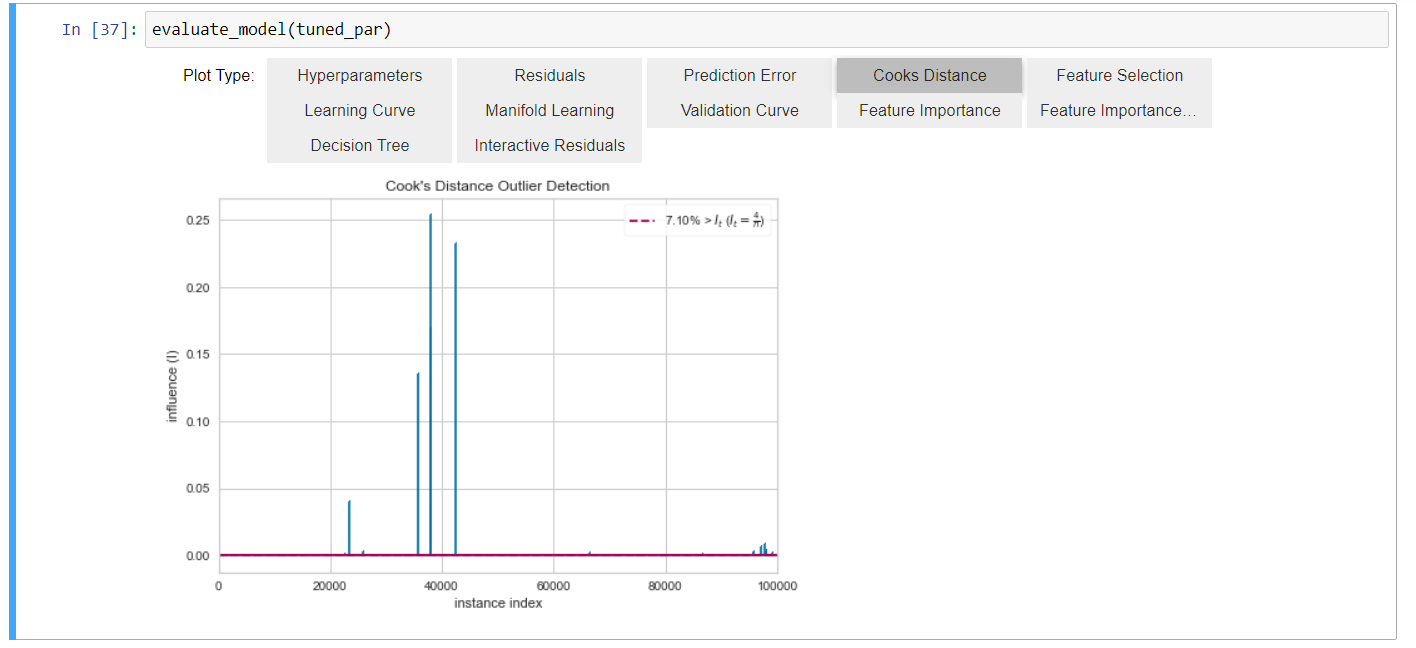
Hyperparameters:



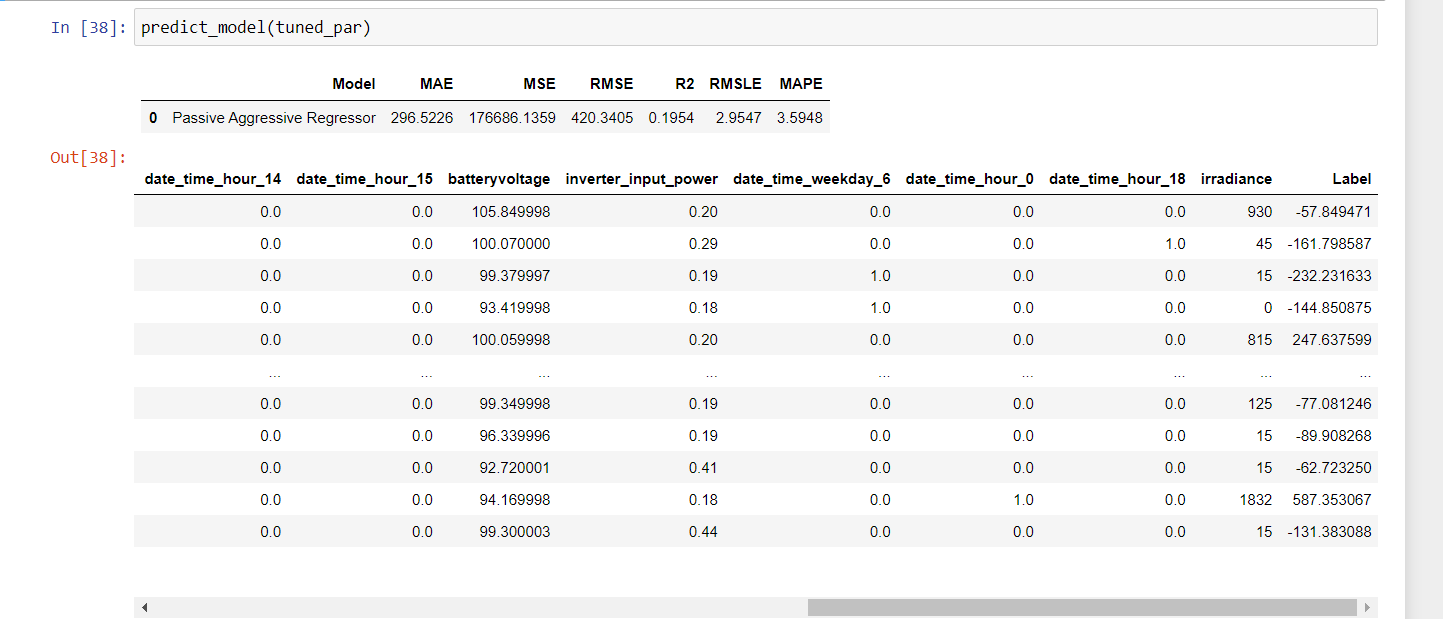
Residuals:

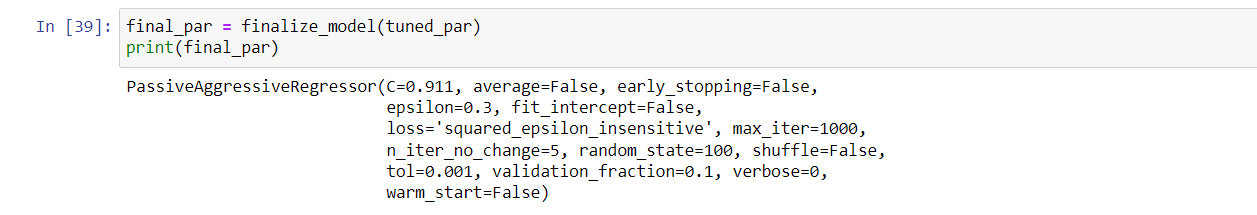


Cooks distance:

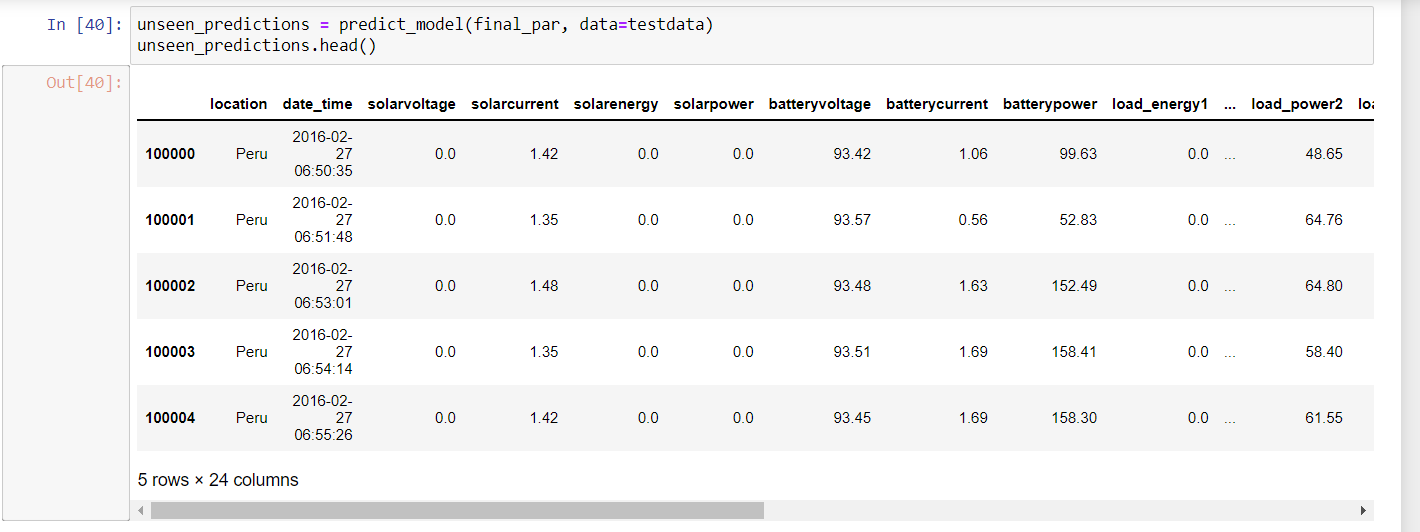


**11) Finalize model**: How to select and finalize the best model at the end of the experiment.

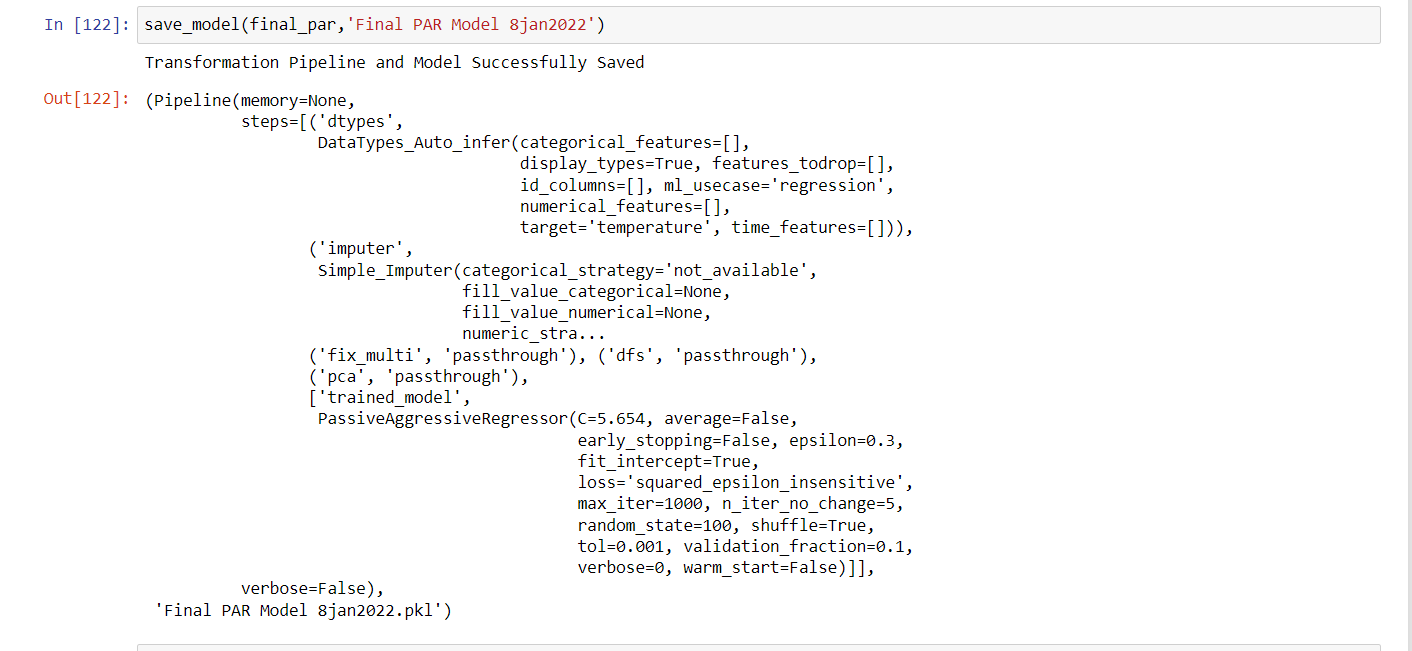


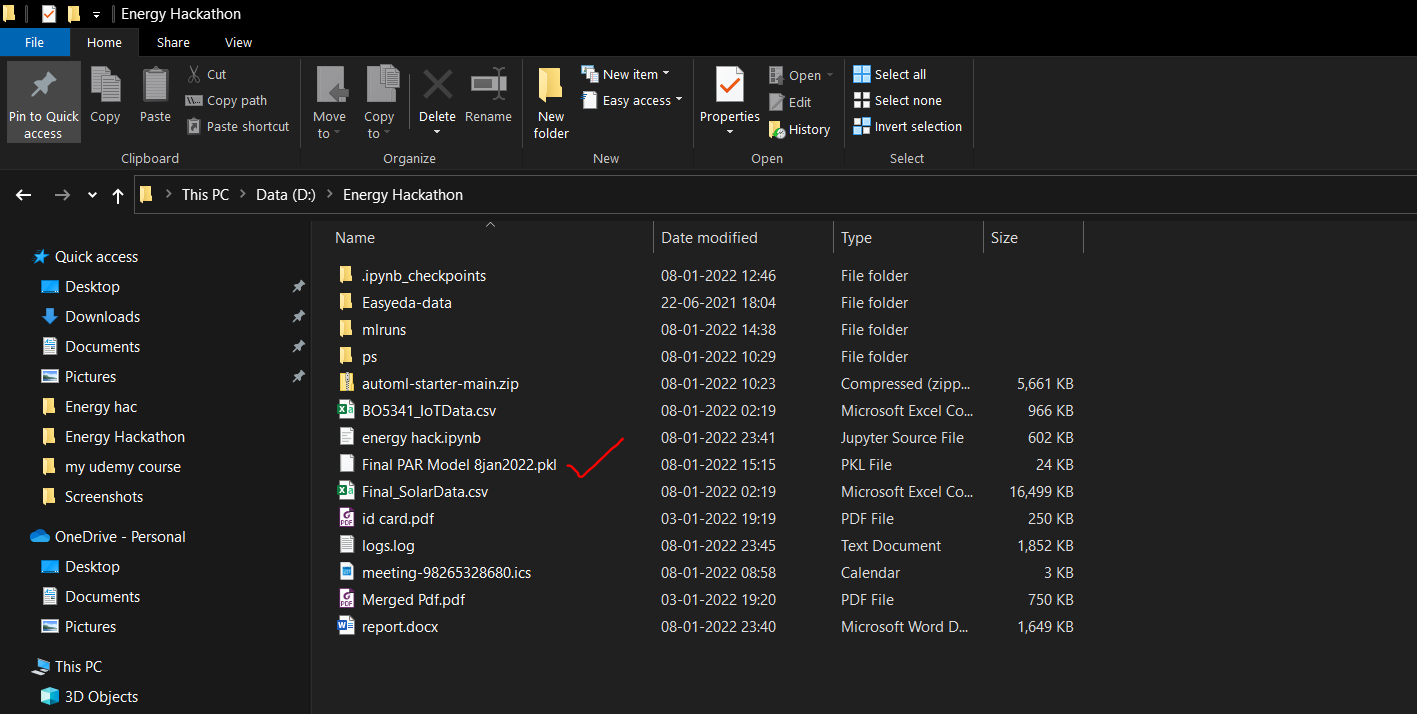


**12) Predict Model**: making predictions on new data.



**13) Save Model**: Saving the model for future use.

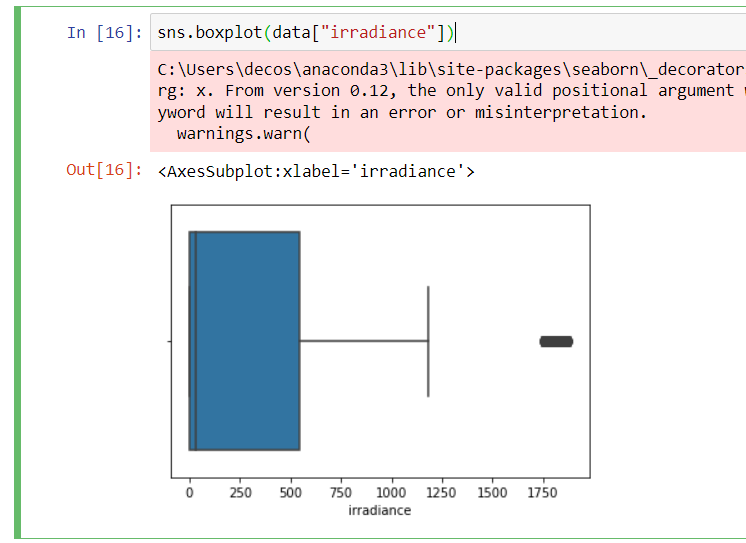


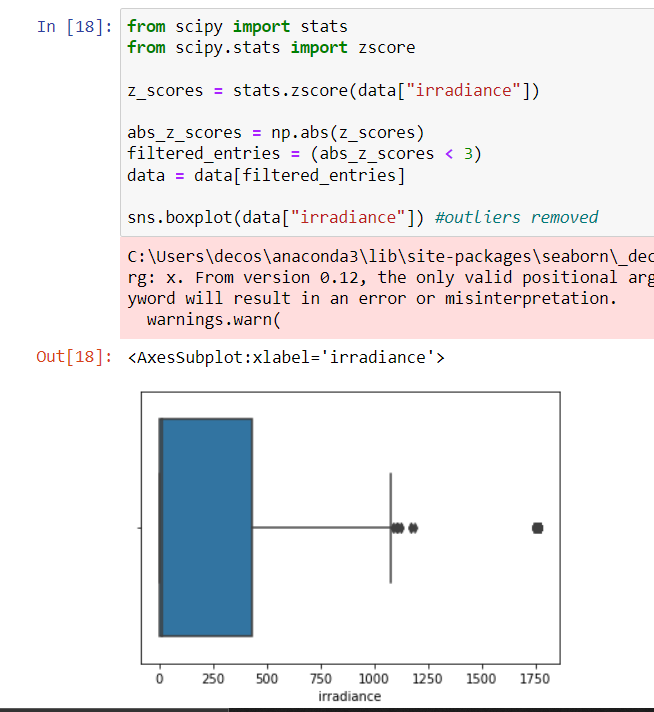


**TIME SERIES DATA VISUALIZATION AND DATE ANNOTATIONS**

**(sub problem statement 2,3 covered)**

Outlier in the irradiance and clearing it.

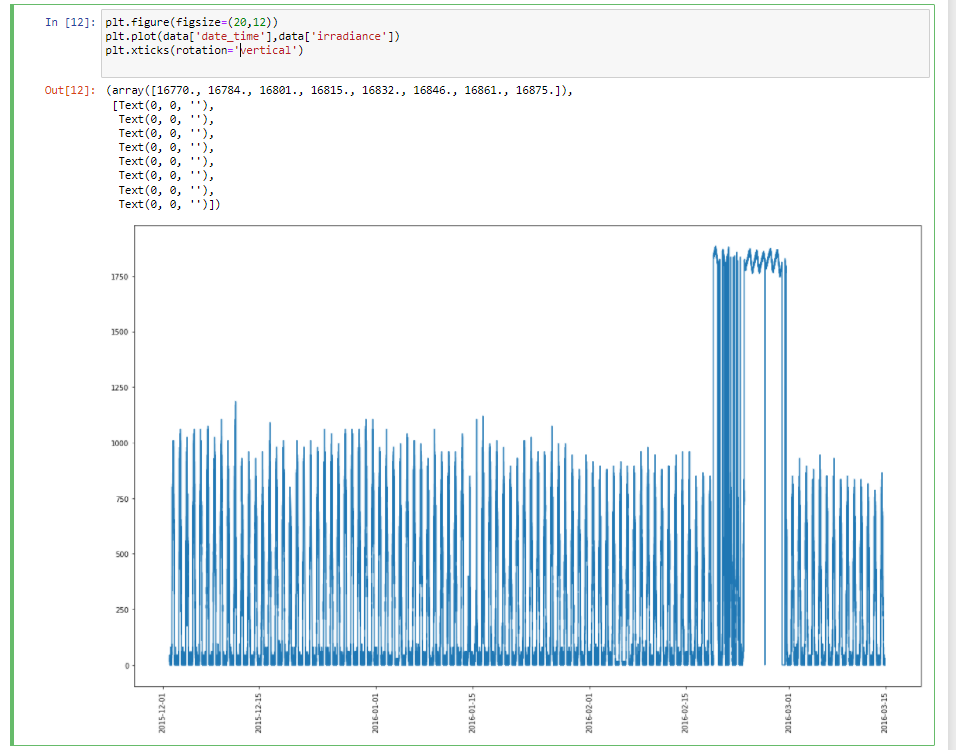




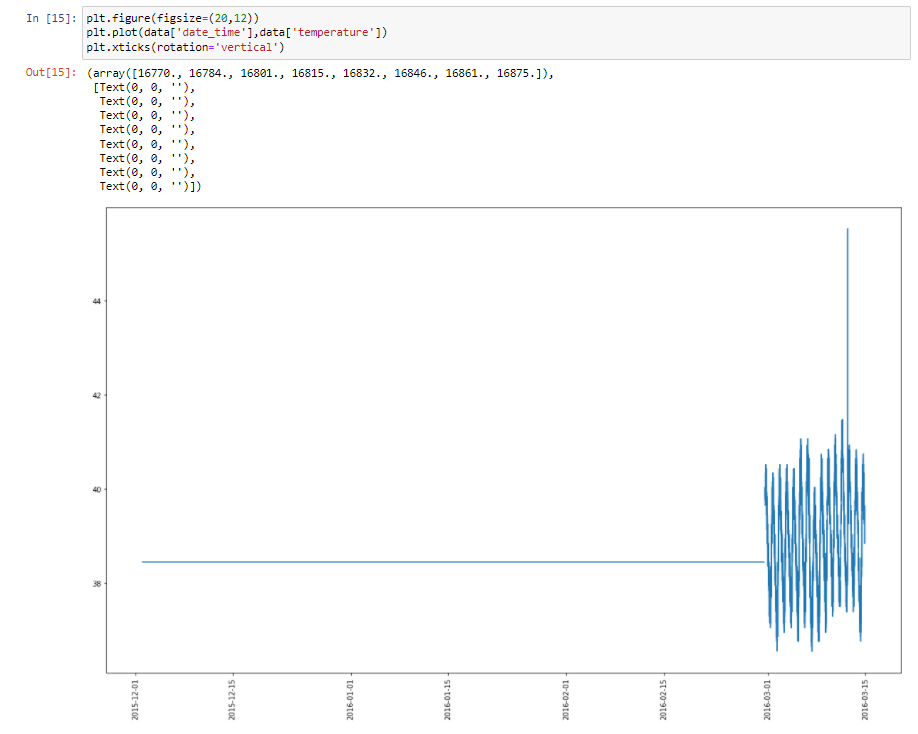
Time series of batteryvoltage



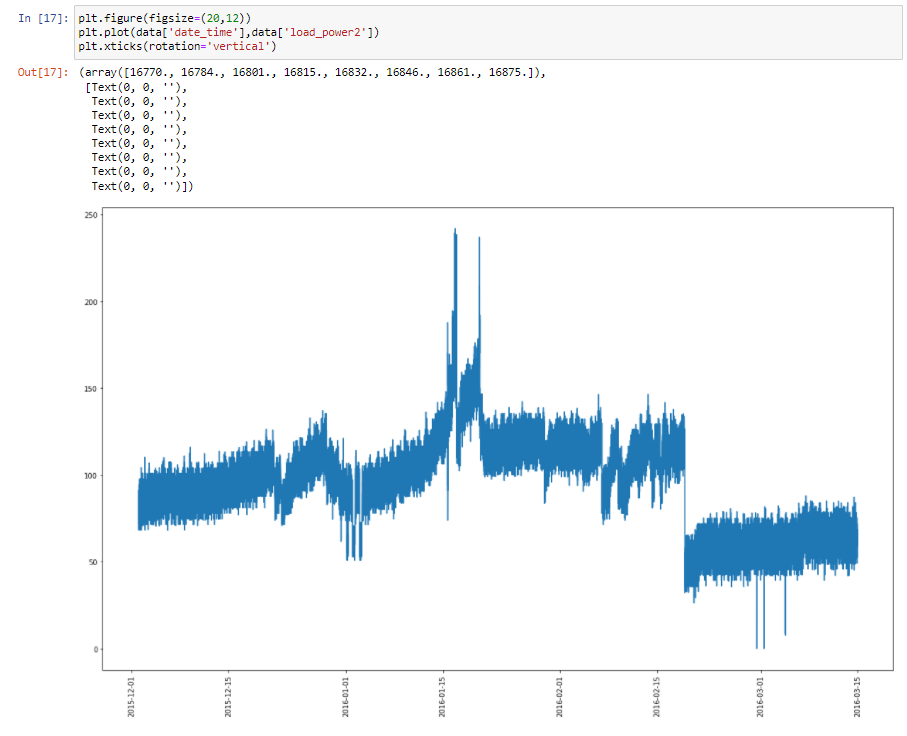
Time series of irradiance



Time series of Temperature



Time series of load\_power 2



**DASHBORAD CREATION AND RESULT VISUALIZATION**

**(Sub problem statements 7 & 10 covered in the video submitted)**

[**http://localhost:8892/voila/render/OneDrive/Desktop/Energy%20hac/Descriptive\_Analysis.ipynb**](http://localhost:8892/voila/render/OneDrive/Desktop/Energy%20hac/Descriptive_Analysis.ipynb)

The Dashboard is created by using **Voila library** in python which converts python jupyter notebook in to html page and then the results can be visualized on the web browser.

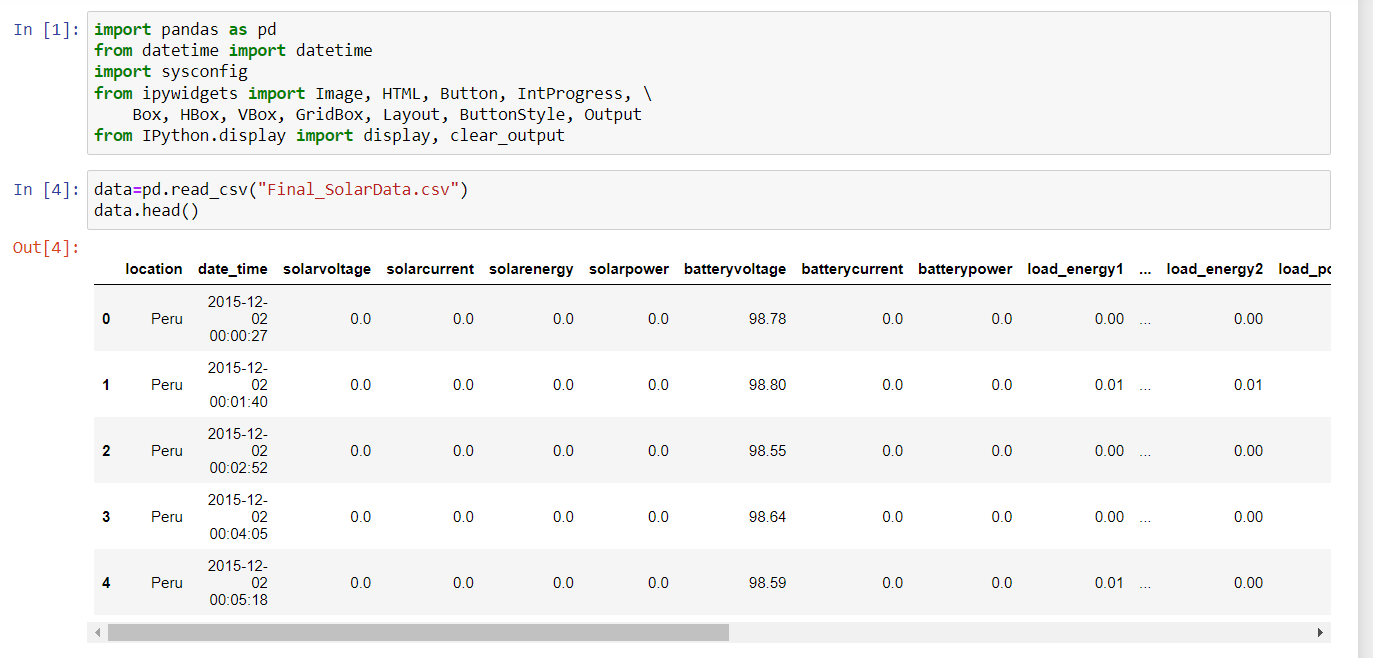
This thing is demonstrated in the submitted video.

**DATA ANNOTATION AND DATA SAVING IN DATABSE**

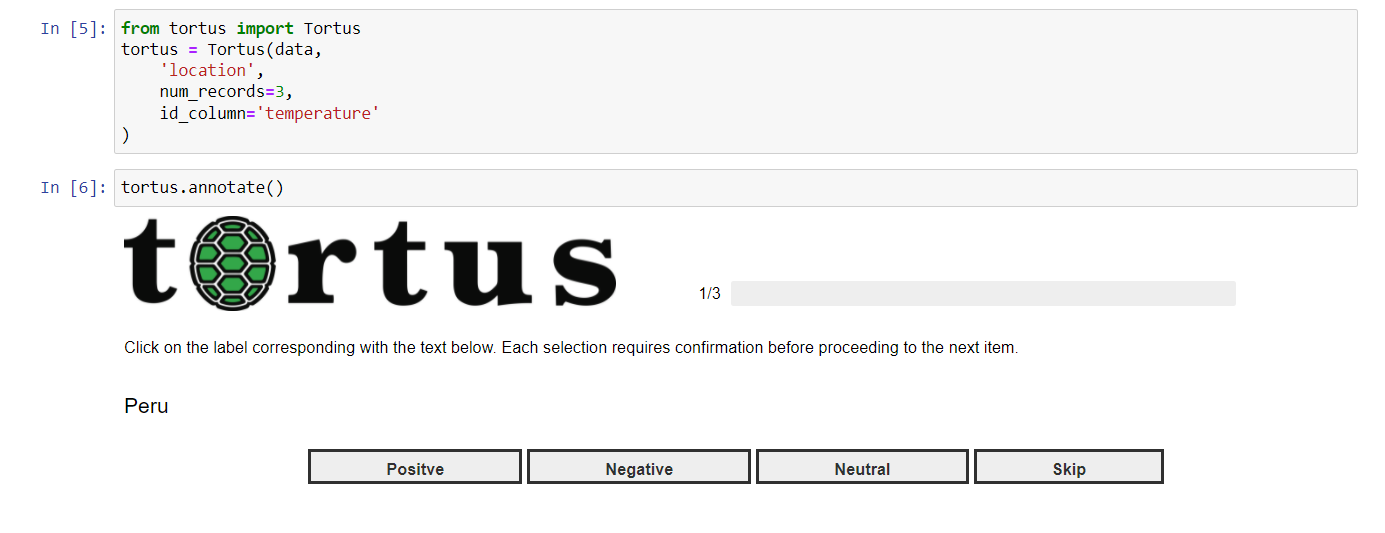
**(Sub problem statement 4 covered)**

Tortus library is used in python to make the data labelling in python.

Libraries installation and dataset importing.



Output:



Annotated and labelled data saved:

