* **Introduction**

The temperature plays an important role in our day-to-day life. It effects the weather which directly or indirectly has lasting effects on the environment and economy and diet of any country and the world. It is a marvel in this day and age that we can access it at our fingertips in real time. And with proper analysis, we can predict the temperature and other components that control the weather. So, in turn we can predict the weather by using various useful methods and creating algorithms and which is actually a great boon for any agrarian economy.

* **Problem Statement:**
* **Data Set Information:**

This data is for the purpose of bias correction of next-day maximum and minimum air temperatures forecast of the LDAPS model operated by the Korea Meteorological Administration over Seoul, South Korea. This data consists of summer data from 2013 to 2017. The input data is largely composed of the LDAPS model's next-day forecast data, in-situ maximum and minimum temperatures of present-day, and geographic auxiliary variables. There are two outputs (i.e. next-day maximum and minimum air temperatures) in this data. Hindcast validation was conducted for the period from 2015 to 2017.

* **Attribute Information:**

For more information, read [Cho et al, 2020].

1. **station - used weather station number: 1 to 25**
2. **Date - Present day: yyyy-mm-dd ('2013-06-30' to '2017-08-30')**
3. **Present\_Tmax - Maximum air temperature between 0 and 21 h on the present day (Â°C): 20 to 37.6**
4. **Present\_Tmin - Minimum air temperature between 0 and 21 h on the present day (Â°C): 11.3 to 29.9**
5. **LDAPS\_RHmin - LDAPS model forecast of next-day minimum relative humidity (%): 19.8 to 98.5**
6. **LDAPS\_RHmax - LDAPS model forecast of next-day maximum relative humidity (%): 58.9 to 100**
7. **LDAPS\_Tmax\_lapse - LDAPS model forecast of next-day maximum air temperature applied lapse rate (Â°C): 17.6 to 38.5**
8. **LDAPS\_Tmin\_lapse - LDAPS model forecast of next-day minimum air temperature applied lapse rate (Â°C): 14.3 to 29.6**
9. **LDAPS\_WS - LDAPS model forecast of next-day average wind speed (m/s): 2.9 to 21.9**
10. **LDAPS\_LH - LDAPS model forecast of next-day average latent heat flux (W/m2): -13.6 to 213.4**
11. **LDAPS\_CC1 - LDAPS model forecast of next-day 1st 6-hour split average cloud cover (0-5 h) (%): 0 to 0.97**
12. **LDAPS\_CC2 - LDAPS model forecast of next-day 2nd 6-hour split average cloud cover (6-11 h) (%): 0 to 0.97**
13. **LDAPS\_CC3 - LDAPS model forecast of next-day 3rd 6-hour split average cloud cover (12-17 h) (%): 0 to 0.98**
14. **LDAPS\_CC4 - LDAPS model forecast of next-day 4th 6-hour split average cloud cover (18-23 h) (%): 0 to 0.97**
15. **LDAPS\_PPT1 - LDAPS model forecast of next-day 1st 6-hour split average precipitation (0-5 h) (%): 0 to 23.7**
16. **LDAPS\_PPT2 - LDAPS model forecast of next-day 2nd 6-hour split average precipitation (6-11 h) (%): 0 to 21.6**
17. **LDAPS\_PPT3 - LDAPS model forecast of next-day 3rd 6-hour split average precipitation (12-17 h) (%): 0 to 15.8**
18. **LDAPS\_PPT4 - LDAPS model forecast of next-day 4th 6-hour split average precipitation (18-23 h) (%): 0 to 16.7**
19. **lat - Latitude (Â°): 37.456 to 37.645**
20. **lon - Longitude (Â°): 126.826 to 127.135**
21. **DEM - Elevation (m): 12.4 to 212.3**
22. **Slope - Slope (Â°): 0.1 to 5.2**
23. **Solar radiation - Daily incoming solar radiation (wh/m2): 4329.5 to 5992.9**
24. **Next\_Tmax - The next-day maximum air temperature (Â°C): 17.4 to 38.9**

* **Please note that there are two target variables here:**

1) Next\_Tmax: Next day maximum temperature

2) Next\_Tmin: Next day minimum temperature

**Analysing the dataset**

To understand each and every component of a dataset first we need to load the dataset then we can start analysing the data precisely and that can be done by using the pandas library along with other basic libraries like –

import pandas as pd

import numpy as np

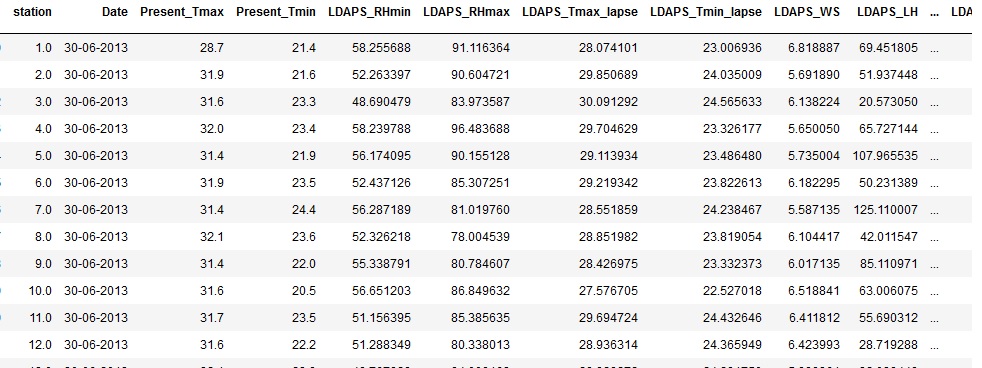
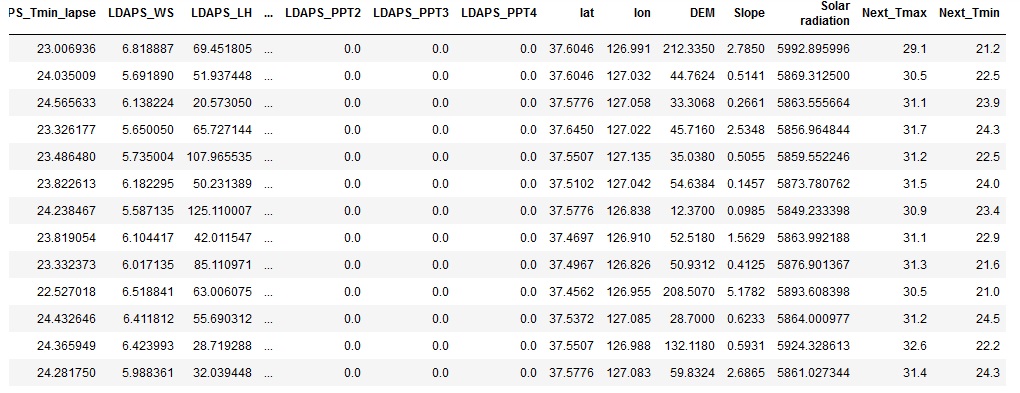
from sklearn import metrics

import seaborn as sns

import matplotlib.pyplot as plt

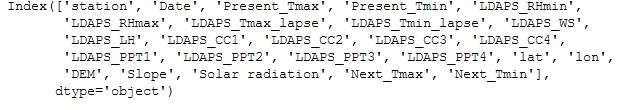
%matplotlib inline

import warnings

warnings.filterwarnings('ignore')

* **Observations:**
  + As we can see, this dataset is for the purpose of bias correction of next-day maximum and minimum air temperatures forecast of the LDAPS model operated by the Korea Meteorological Administration over Seoul, South Korea.
  + By reading the dataset we got to know that there are 25 columns in this dataset of which 23 columns are independent variables and there are 2 dependent variables, one is 'Next\_Tmax' and the other one is 'Next\_Tmin'.
  + As we need to predict the Next day maximum temperature and Next day minimum temperature ( both are continuous values), it is a regression problem.

The dataset is a good blend of numerical, categorical and nominal data. And all the continuous data are in different scales. The dataset contains 7752 rows and 25 columns.



* Out of the 25 columns 23 columns are the features and these are as follows,

**station - used weather station number,**

**Date - Present day: yyyy-mm-dd ,**

**Present\_Tmax - Maximum air temperature between 0 and 21 h on the present day**

**Present\_Tmin - Minimum air temperature between 0 and 21 h on the present day**

**LDAPS\_RHmin - LDAPS model forecast of next-day minimum relative humidity**

**LDAPS\_RHmax - LDAPS model forecast of next-day maximum relative humidity**

**LDAPS\_Tmax\_lapse - LDAPS model forecast of next-day maximum air temperature applied lapse rate**

**LDAPS\_Tmin\_lapse - LDAPS model forecast of next-day minimum air temperature applied lapse rate**

**LDAPS\_WS - LDAPS model forecast of next-day average wind speed (m/s)**

**LDAPS\_LH - LDAPS model forecast of next-day average latent heat flux (W/m2)**

**LDAPS\_CC1 - LDAPS model forecast of next-day 1st 6-hour split average cloud cover (0-5 h)**

**LDAPS\_CC2 - LDAPS model forecast of next-day 2nd 6-hour split average cloud cover (6-11 h)**

**LDAPS\_CC3 - LDAPS model forecast of next-day 3rd 6-hour split average cloud cover (12-17 h)**

**LDAPS\_CC4 - LDAPS model forecast of next-day 4th 6-hour split average cloud cover (18-23 h)**

**LDAPS\_PPT1 - LDAPS model forecast of next-day 1st 6-hour split average precipitation (0-5 h)**

**LDAPS\_PPT2 - LDAPS model forecast of next-day 2nd 6-hour split average precipitation (6-11 h)**

**LDAPS\_PPT3 - LDAPS model forecast of next-day 3rd 6-hour split average precipitation (12-17 h)**

**LDAPS\_PPT4 - LDAPS model forecast of next-day 4th 6-hour split average precipitation (18-23 h)**

**lat - Latitude**

**lon - Longitude**

**DEM - Elevation**

**Slope**

**Solar radiation - Daily incoming solar radiation (wh/m2)**

* And the dependent variables are as follows,

**Next\_Tmax - The next-day maximum air temperature (Â°C)**

**Next\_Tmin - The next-day minimum air temperature (Â°C)**

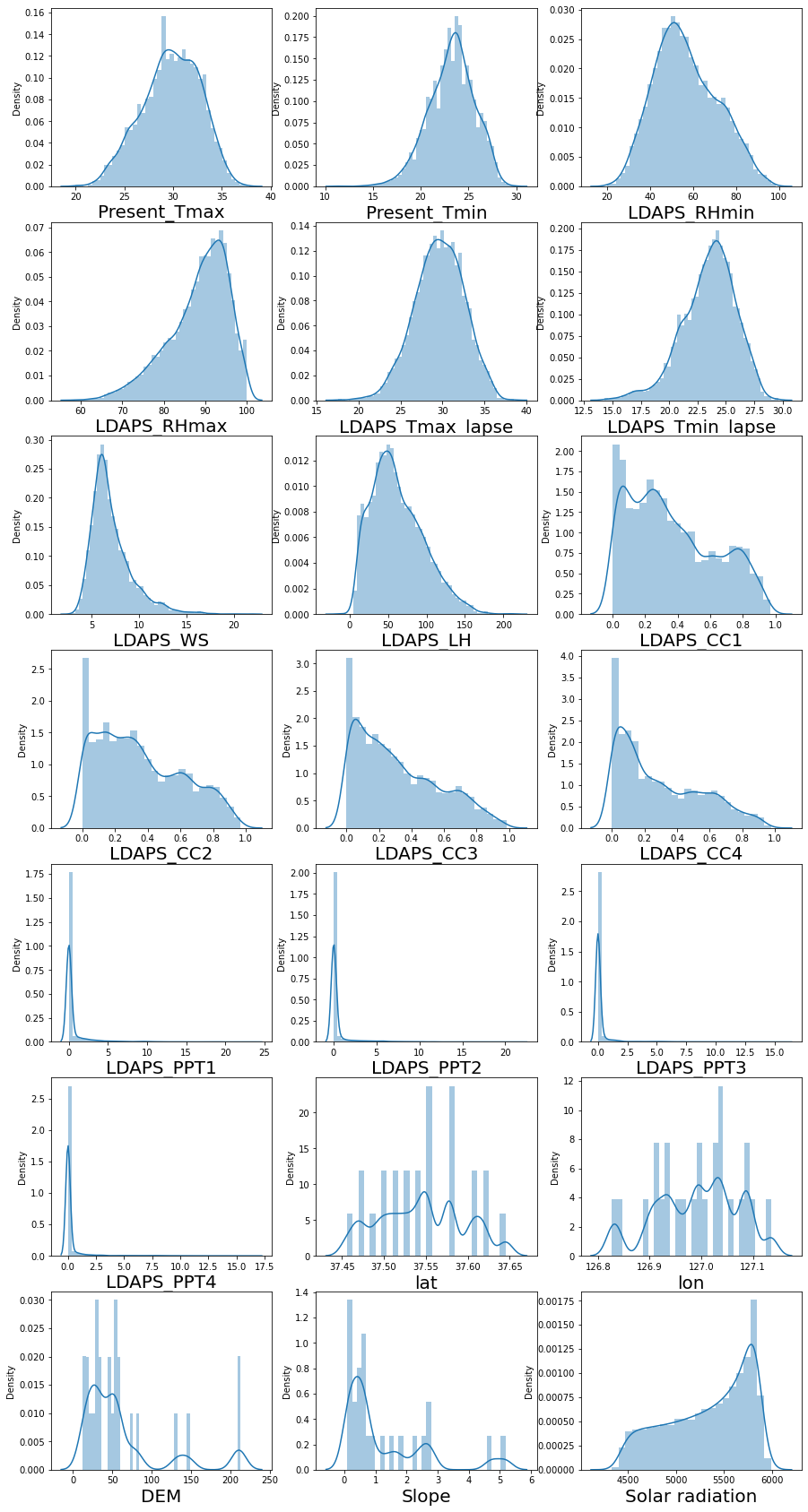
As we need to **predict two key elements one is highest temperature and other is lowest temperature.**

* **Exploratory Data Analysis:**

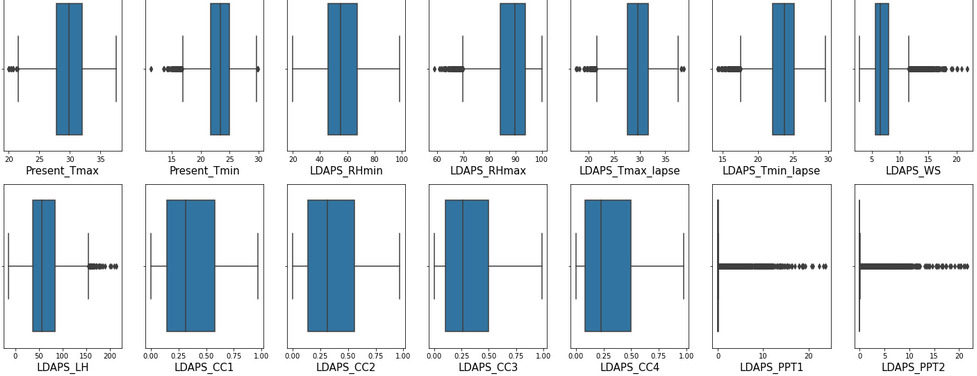
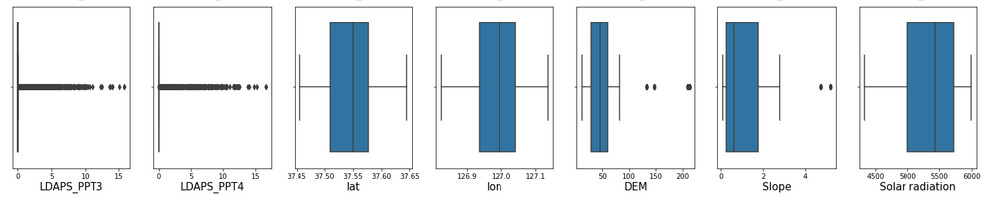
The exploratory data analysis (EDA) mainly consists of –

* **UNIVARIATE ANALYSIS** - It was mainly done to understand the distribution of the continuous data and also to get the count of the categorial data.

Some of the analysis done –

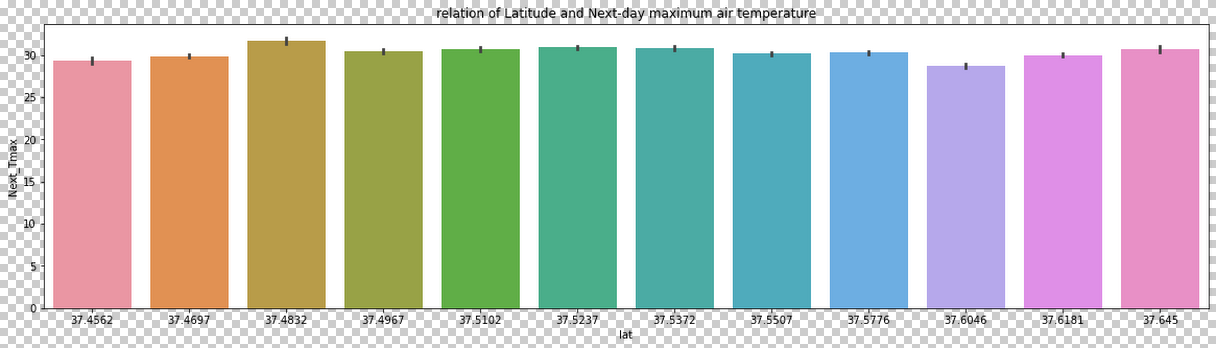


* The dependent variables showing more or less normal distribution.
* Next day lowest minimum temperature is 11.3, Next day highest minimum temperature is 29.8
* Next day lowest maximum temperature is 17.4, Next day highest maximum temperature is 38.9
* The independent variables, Present\_Tmin and Present\_Tmax are showing more or less normal distribution, a little left skewness is in present day minimum temperature.
* Present day lowest minimum temperature is 11.3, Present day highest minimum temperature is 29.9 .
* Present\_day\_lowest maximum temperature is 20.0, Present day highest maximum temperature is 37.6 .
* The independent variables, LDAPS\_Tmin\_lapse and LDAPS\_Tmax\_lapse are showing more or less normal distribution, a little left skewness is present in both variables.
* Next day lowest minimum air temperature applied lapse rate is 14.27264631, Next day highest minimum air temperature applied lapse rate is 29.61934244.
* Next-day lowest max air temperature applied lapse rate is 17.62495378, Next-day highest max air temperature applied lapse rate is 38.54225522 .
* The independent variables, LDAPS\_WS and LDAPS\_LH are showing more or less normal distribution, a right skewness is present in both variables.
* Next day minimum average wind speed is 2.882579625, Next day maximum average wind speed is 21.85762099 .
* Next day mininimum average latent heat flux is -13.60321209, Next day maximum average latent heat flux is 213.4140062 .
* The independent variables, LDAPS\_CC1 and LDAPS\_CC2 are showing more or less normal distribution, a little right skewness is present in both variables.
* Next day 1st 6-hour split minimum average cloud cover is 0.0, Next day 1st 6-hour split maximum average cloud cover is 0.967277328
* Next day 2nd 6-hour split minimum average cloud cover is 0.0, Next day 2nd 6-hour split maximum average cloud cover is 0.96835306.
* The independent variables, LDAPS\_PPT1 and LDAPS\_PPT2 both the variables are showing right skewness.
* Next day 1st 6-hour split minimum average precipitation is 0.0, Next day 1st 6-hour split maximum average precipitation is 23.70154408 .
* Next day 2nd 6-hour split minimum average precipitation is 0.0, Next day 2nd 6-hour split maximum average precipitation is 21.62166078 .
* The independent variables, Latitude and Longitude both the variables are showing more or less normal distribution.
* Minimum Latitude is 37.4562 and Maximum Latitude is 37.645.
* Minimum Longitude is 126.82 and Maximum Longitude is 127.135.
* **Outlier Detection**:

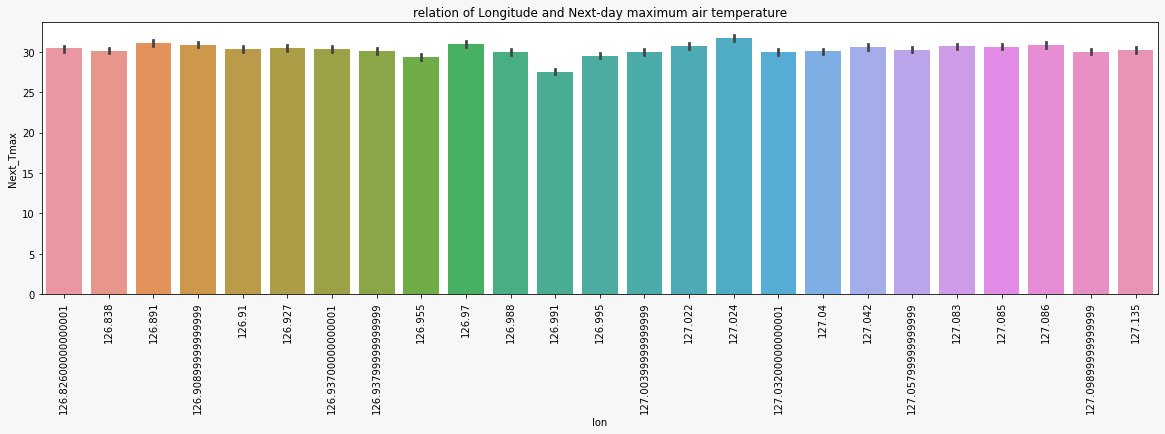


* There are outliers in most of the variables like, Present\_Tmax, Present\_Tmin, LDAPS\_RHmax, LDAPS\_Tmax\_lapse, LDAPS\_Tmin\_lapse, LDAPS\_WS, LDAPS\_LH, LDAPS\_PPT1, LDAPS\_PPT2, LDAPS\_PPT3, LDAPS\_PPT4 and DEM and Slope is having 2-3 outliers
* **BIVARIATE ANALYSIS**- This type of analysis allows to understand **relations** between **feature vs target** and **feature vs feature**.

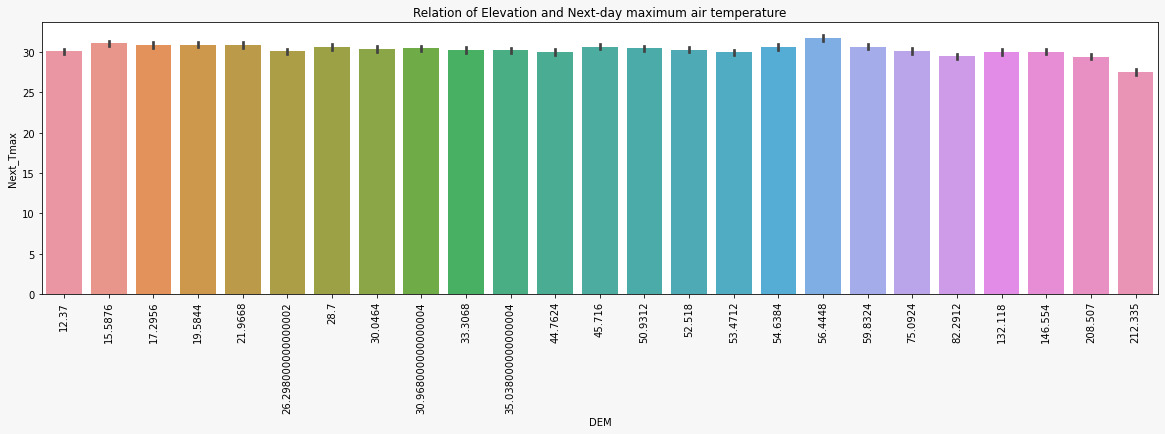
Some of the analysis done –



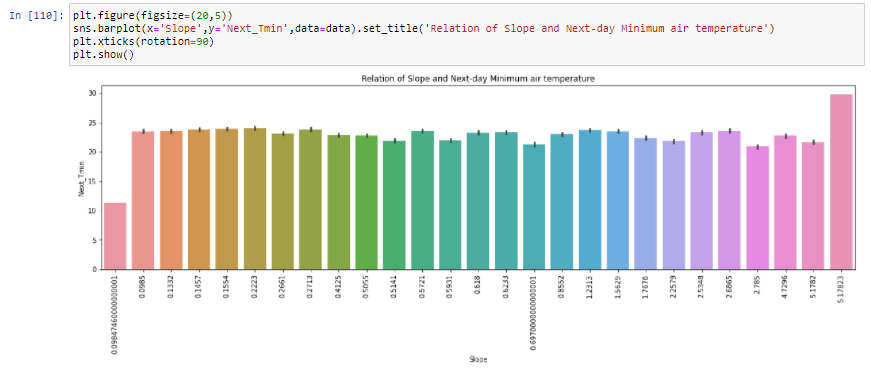
* In Latitude 37.4697 the next day's maximum temperature is highest.



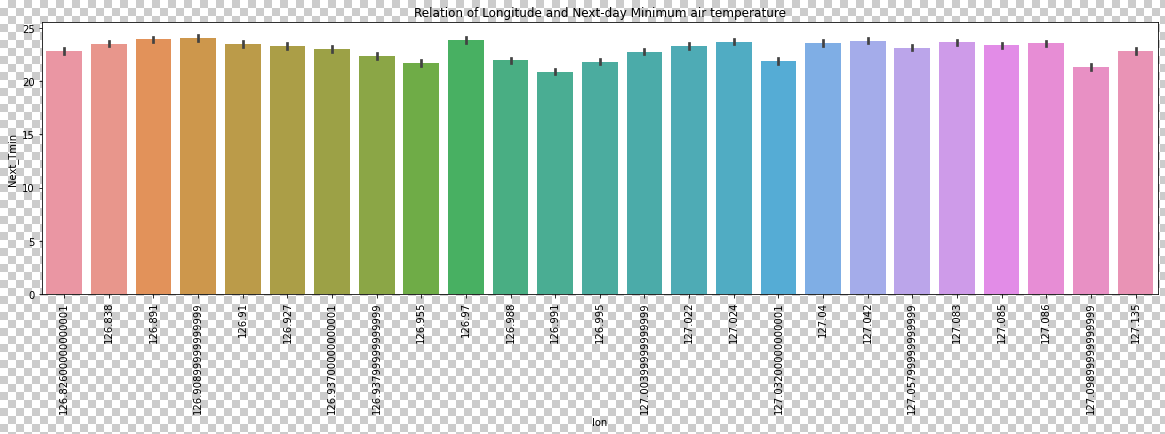
* Elevation 212.335 has the maximum temperature on next day.



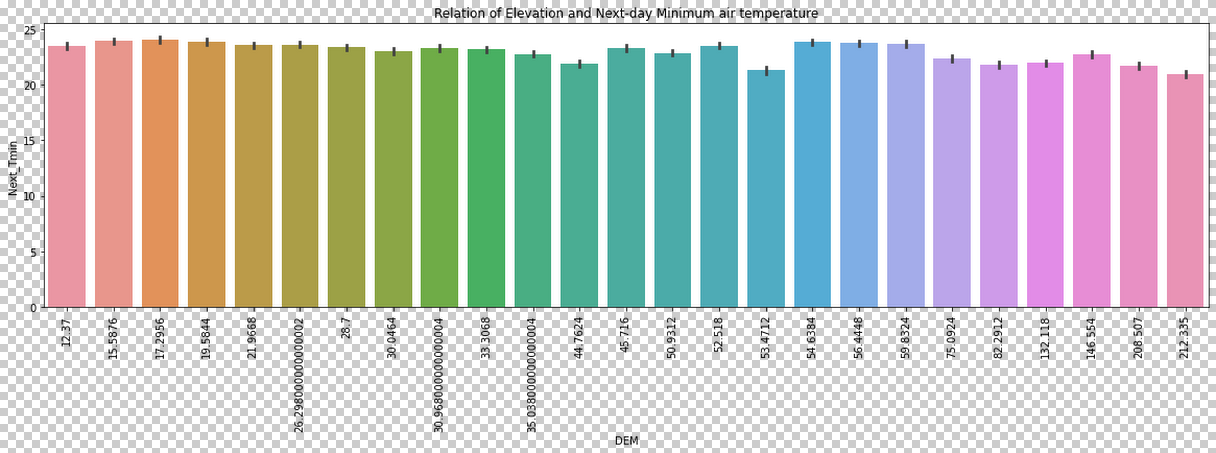
* In case of Elevation ( height above sea level ) 56.448 the next day's maximum temperature is highest.



* Slope 0.09847 has the minimum temperature on next day.

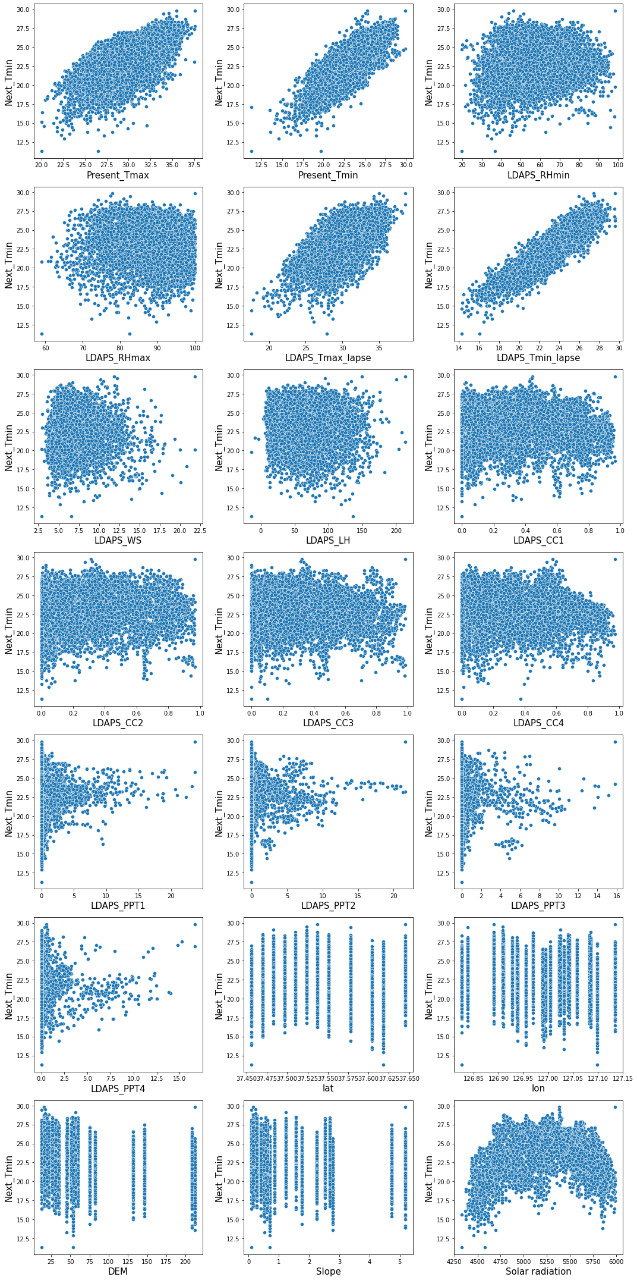


* Longitude 126.991 has the lowest minimum temperature on next day.



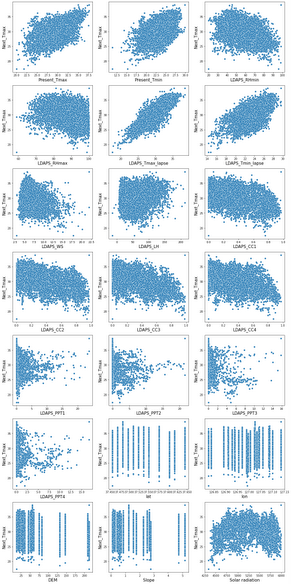
* Elevation 212.335 has the lowest minimum temperature on next day.

Now let’s check the relation between continuous features and next day min temperature.

+

* Next day's minimum temperature and present day's maximum temperature has a positive linear relation.
* Next day's minimum temperature and present day's minimum temperature has a positive linear relation.
* Next day's minimum temperature and next-day minimum relative humidity has a positive relation.
* Next day's minimum temperature and next-day maximum relative humidity has a positive relation.
* Next day's minimum temperature and next-day maximum air temperature applied lapse rate has a positive linear relation.
* Next day's minimum temperature and next-day minimum air temperature applied lapse rate has a positive linear relation.
* Next day's minimum temperature and next-day average wind speed has a scattered relation.
* Next day's minimum temperature and next-day average latent heat flux has a more or less scattered relation.
* Next day's minimum temperature and next-day 1st 6-hour split average cloud cover has a more or less scattered relation.
* Next day's minimum temperature and next-day 2nd 6-hour split average cloud cover has a more or less scattered relation.
* Next day's minimum temperature and next-day 3rd 6-hour split average cloud cover has a more or less scattered relation.
* Next day's minimum temperature and next-day 4th 6-hour split average cloud cover has a more or less scattered relation.
* The relation between Next day's minimum temperature and next-day 1st 6-hour split average precipitation is scattered.
* The relation between Next day's minimum temperature and next-day 2st 6-hour split average precipitation is somewhat scattered.
* The relation between Next day's minimum temperature and next-day 3st 6-hour split average precipitation is somewhat scattered.
* The relation between Next day's minimum temperature and next-day 4st 6-hour split average precipitation is somewhat scattered.
* Between Lattitude 37.600 - 37.625 next day's minimum temperature is showing more low.
* Between Longitude 127.10 next day's minimum temperature is showing more low.
* Between Elevation 50 next day's minimum temperature is showing more low.
* Between Slope 0 - 1 next day's minimum temperature is showing more low.
* Less solar radiation less temperature.

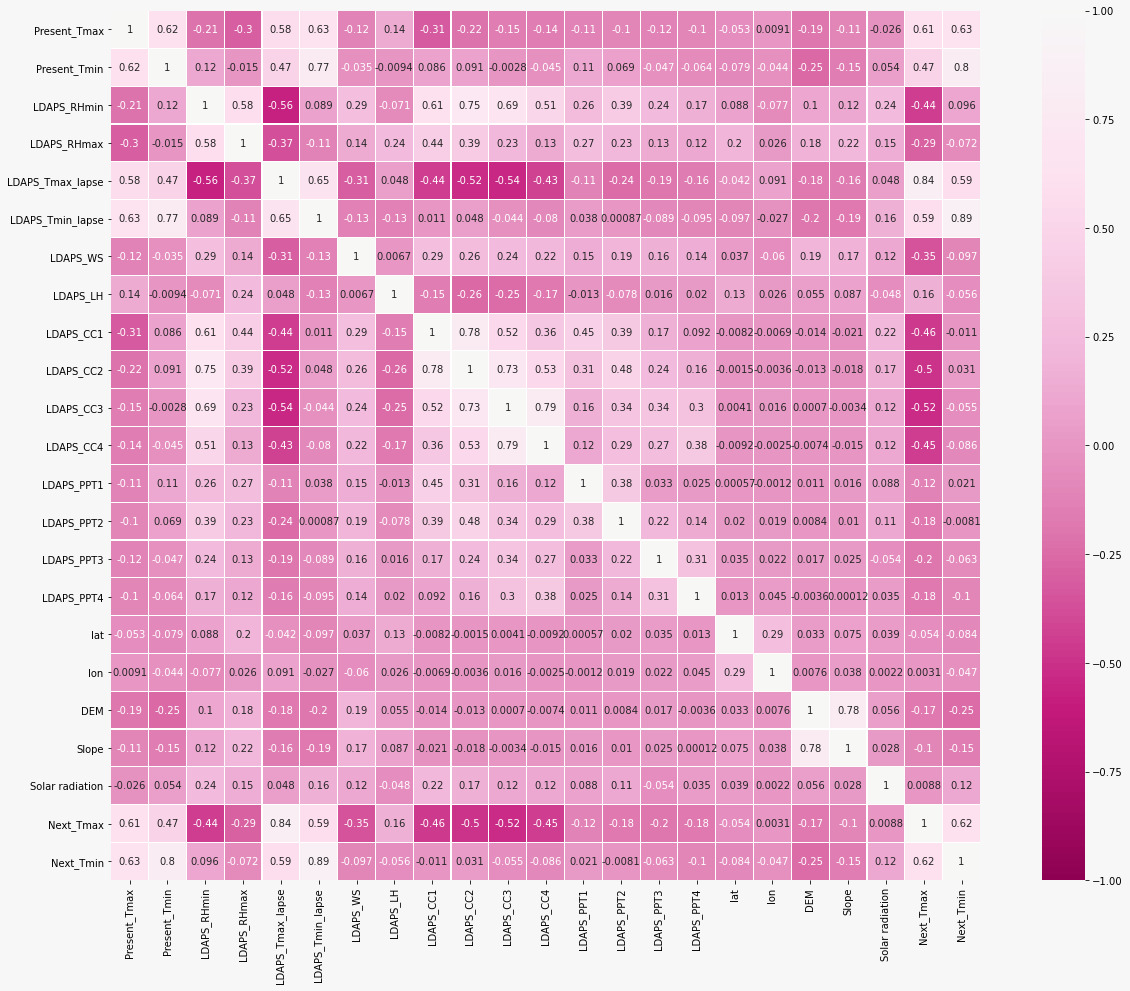
Now let’s check the relation between continuous features and next day maximum temperature.



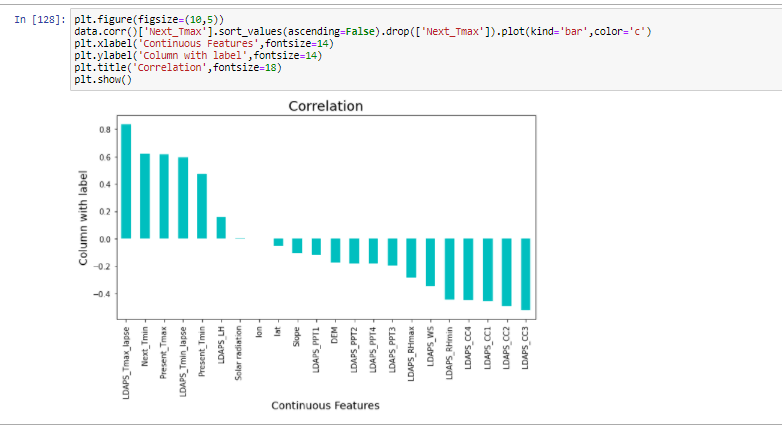
* Next day's maximum temperature and present day's maximum temperature has a positive linear relation.
* Next day's maximum temperature and present day's minimum temperature has a positive linear relation.
* Next day's maximum temperature and next-day minimum relative humidity has a negative linear relation.
* Next day's maximum temperature and next-day maximum relative humidity has a negative linear relation.
* Next day's maximum temperature and next-day maximum air temperature applied lapse rate has a positive linear relation.
* Next day's maximum temperature and next-day minimum air temperature applied lapse rate has a positive linear relation.
* Next day's maximum temperature and next-day average wind speed has a negative linear relation.
* Next day's maximum temperature and next-day average latent heat flux has a more or less positive linear relation.
* Next day's maximum temperature and next-day 1st 6-hour split average cloud cover has a more or less negative linear relation.
* Next day's maximum temperature and next-day 2nd 6-hour split average cloud cover has a more or less negative linear relation.
* Next day's maximum temperature and next-day 3rd 6-hour split average cloud cover has a more or less negative linear relation.
* Next day's maximum temperature and next-day 4th 6-hour split average cloud cover has a more or less negative linear relation.
* The relation between Next day's maximum temperature and next-day 1st 6-hour split average precipitation is somewhat scattered.
* The relation between Next day's maximum temperature and next-day 2st 6-hour split average precipitation is somewhat scattered.
* The relation between Next day's maximum temperature and next-day 3st 6-hour split average precipitation is somewhat scattered.
* The relation between Next day's maximum temperature and next-day 4st 6-hour split average precipitation is somewhat scattered.
* Between Latitude 35.475 - 35.500 next day's maximum temperature is showing little high.
* Between Longitude 127.00 - 127.05 next day's maximum temperature is showing little high.
* Between Elevation 25 - 50 next day's maximum temperature is showing little high.
* Between Slope 0 - 1 next day's maximum temperature is showing little high.
* More solar radiation more temperature.

**MULTIVARIATE ANALYSIS**— The analysis is done to mainly check the relation among different features that may lead to bias results. This mainly used to check **correlation** among different variables in the dataset and for checking outliers as well .

Some of the analysis done ---



* Next\_Tmax has a good correlation with LDAPS\_Tmax\_lapse(0.84).
* Next\_Tmin has a good correlation with present\_Tmin(0.8).
* The features are not much correlated with each other.



* Next\_Tmax has highest positive correlation with LDAPS\_Tmax\_lapse.
* Next\_Tmax has highest negative correlation with LDPAS\_CC3.
* **Remarks on Exploratory Data Analysis:**

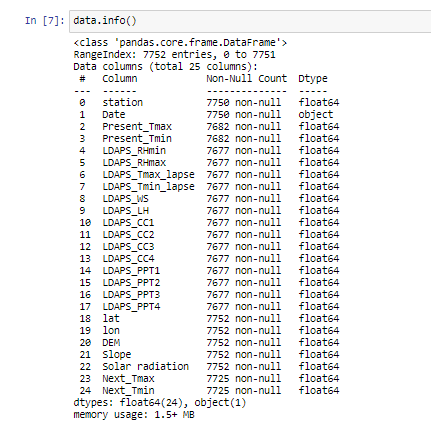
The EDA process helped shine light among the features given in the dataset. The process helped show the data distribution where it was understood that the features were a mixed bag showing a normal distribution, to slightly skewed to skewed distribution. The analysis also showed that the data certain factors like elevation, slope were directly related with temperature and also that both the factors like ‘Next\_Tmax’ , ‘Next\_Tmin’ are correlated with ‘LDAPS\_Tmax\_lapse’ and ‘present\_Tmin’ , respectively. The dataset was filled with outliers which helped to understand the skewness in the dataset. It also helped to understand which year and month had the highest temperature

**Data Preprocessing :**

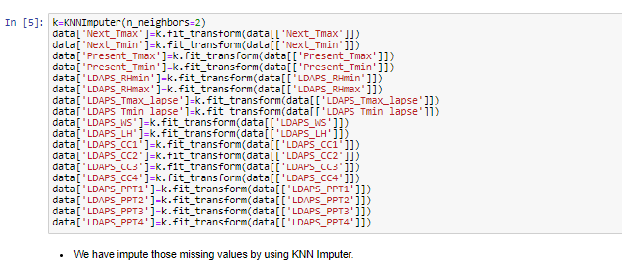
We need to import necessary libraries for preprocessing:

* The dataset had 7752 rows and 25 columns.
* The dataset had int64 (24 columns) and object (1 column) datatype.
* The dataset is having missing values 19 out of the 25 columns.
* The dataset also contains the time-date data in date format which had to be extracted.

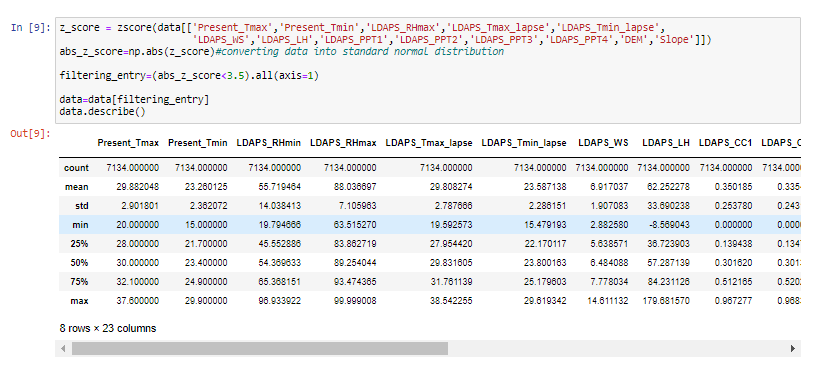
The following image shows the aforementioned information-

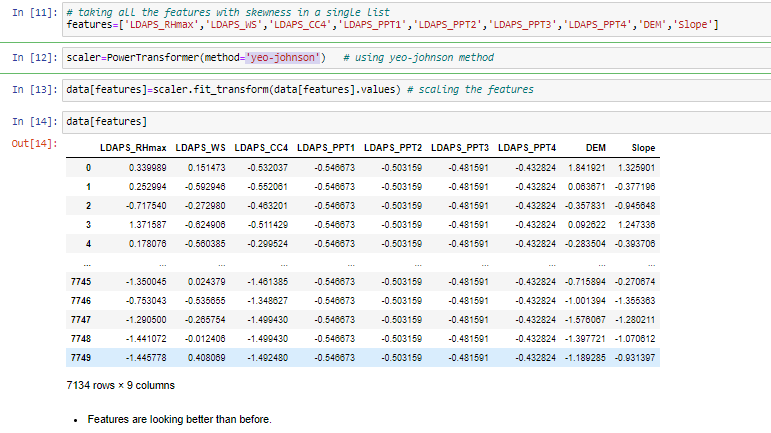


In order to predict the temperature with high accuracy the data had to be cleaned . 1. **Filling Missing values**: The missing values had to be replaced using the KNN Imputer .



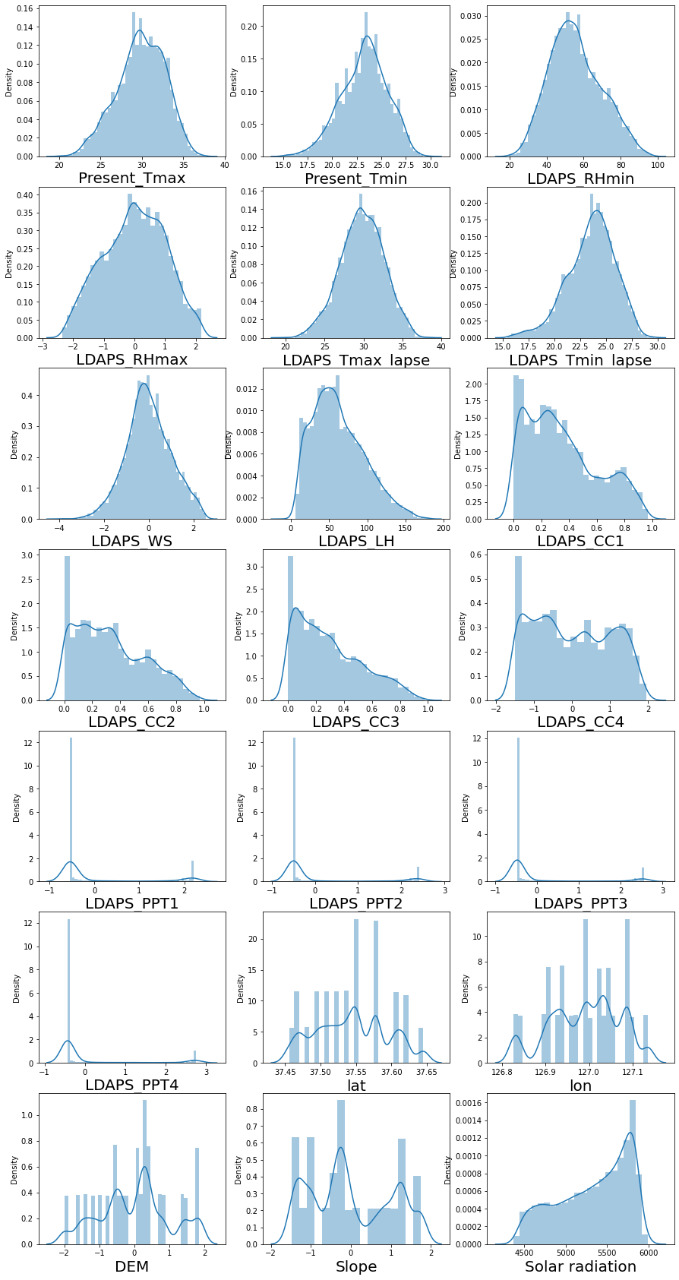
2. **Removal of Outliers and Skewness**:The outliers had to be moved using the ZScore and PowerTransformer('yeo-johnson')





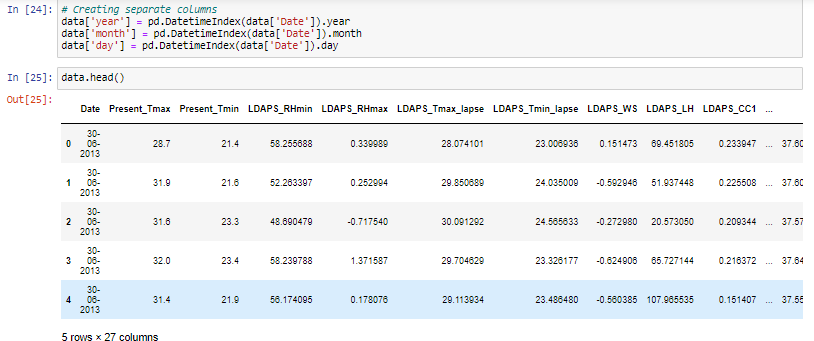
The power transformer was used to reduce the skewness of the residual features.

For checking whether the data distribution has improved from before, it’s always good to check the distribution after the outlier removal.

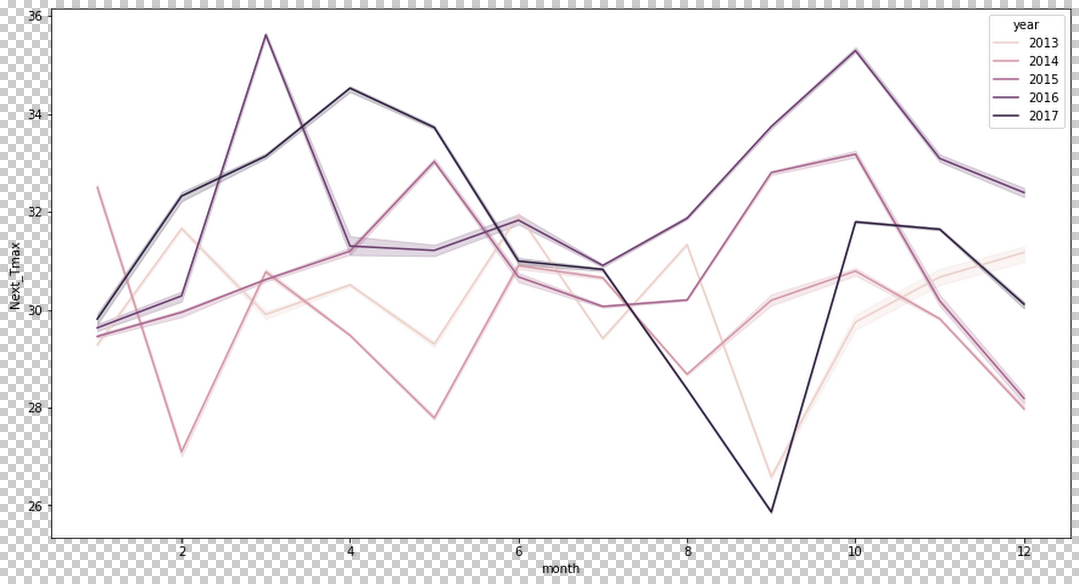


**Feature Engineering** :

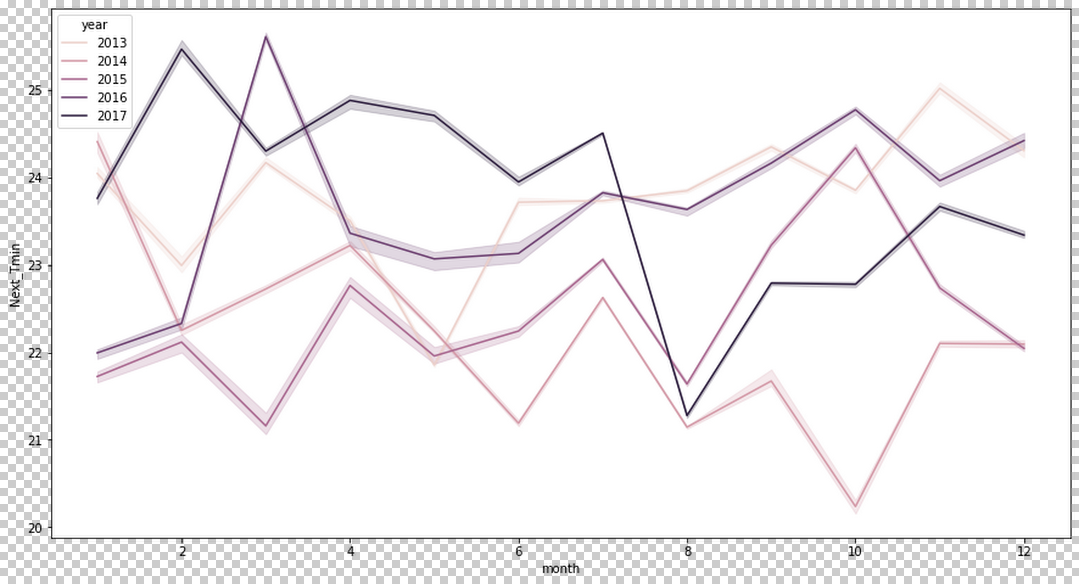
The dataset contained date format which is an integral part of the dataset as it provides information related to day- day , month-month and year-year of the temperature dataset.



* Separate columns have been obtained like day, month and year which are crucial to the analysis process.



* In April 2017 the temperature seems highest.
* In march 2016 the temperature seems highest.
* In October 2015 the temperature seems highest.
* In June 2014 the temperature seems highest.
* In June 2013 the temperature seems highest.



* In August 2017 the temperature seems lowest.
* In February 2016 the temperature seems lowest.
* In march 2015 the temperature seems lowest.
* In October 2014 the temperature seems lowest.
* In may 2013 the temperature seems lowest.

Part of the cleaning process also contained removal of unnecessary data and dropping columns like the ‘date’. As the correlation values are low it can be understood that other columns need not be dropped. As this unique dataset has two columns of target variables so in each case one of it had to be dropped while working with the other. While working with ‘Next\_Tmax’ ,’Next\_Tmin’ had to be dropped and vice-versa.

* **Building Machine Learning Models**

The interesting thing about the dataset is that it has **2 columns of target data**. Namely:

* **Next\_Tmax: Next day maximum temperature**
* **Next\_Tmin: Next day minimum temperature**

Model training and creation and to done for both the target variables, or else both highest and lowest temperature cannot be predicted.

For achieving model creation the following libraries were used-

* **from sklearn.model\_selection import train\_test\_split,GridSearchCV**
* **from sklearn.linear\_model import LinearRegression**
* **from sklearn.neighbors import KNeighborsRegressor**
* **import xgboost as xgb**
* **from sklearn.ensemble import RandomForestRegressor**
* **from sklearn.ensemble import GradientBoostingRegressor**
* **from sklearn.linear\_model import Ridge,Lasso,RidgeCV,LassoCV**
* **from sklearn.model\_selection import cross\_val\_score**

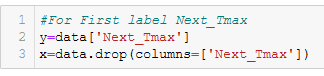
And Evaluation will be done with –

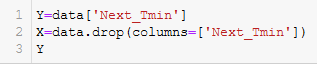
* **R2 score**
* **Mean\_absolute\_error**
* **Mean\_squared\_error**
* **Root mean squared error**
* **Cross Validation**

Let’s start the building process:

**Dataset Division:**

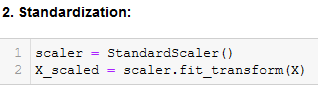
Now as we have got 2 labels in the dataset, we are going to split it 2 times. For prediction of minimum temperature we will use X,Y(upper case) and for maximum temperature we will use x,y(lower case).





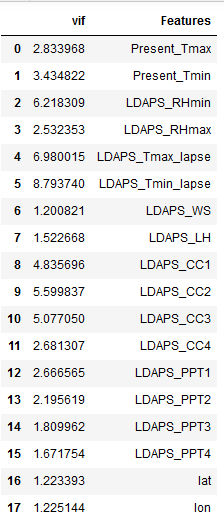
Then we can standardize the x\_scaled variable and X\_scaled variable separately and prediction must be done separately. And for now onwards we will do all the steps 2 times one for the target variable maximum temperature and other is for minimum temperature.

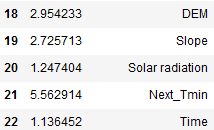
* **Standardization:**

****

* **VIF (Varience inflation factor):**

Before proceeding into the model development, it is very important that data is standardized and checked for corelation. The correlation matrix does provide a overview but it is always better to double the data and it can be done using the Variance Inflation Factor (VIF). Generally, if the VIF scores are more than 5, it means there is correlation among the data. The higher the vif scores higher the corelation and probably it means a few of the columns have to be dropped or else it will result in faulty data. Although VIF is very simple to use it must be that the data should be standardized so that the variance in reduced and outliers removed as the standard deviation of the data is corrected to 1 and the mean to 0.

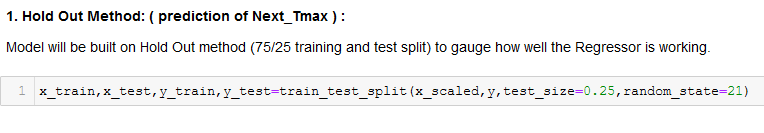




* **As it can be seen that the VIF scores low and it means less multi-collinearity. Before applying the VIF the data was standardized using a standard scaler which helps in the removal of outliers so that data is without much noise and clean.**
* **Hold Out Method:**

The model was split using the train-test-split using Hold Out method (75/25 training and test split) to gauge how well the Regressor is working.

**But before that we need to standardize the model.We have done hold out method 2 times because there is 2 labels.**

****

****

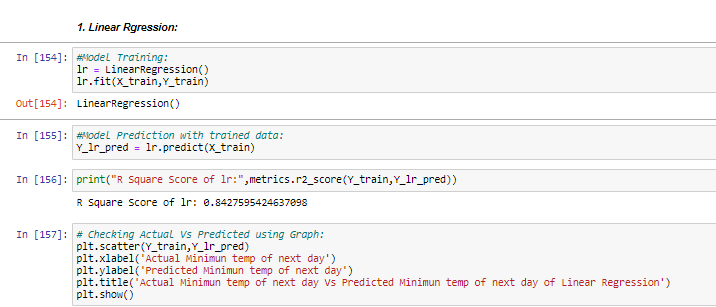
* **Model Creation:**

For building a model that can be trained to give the best results five different models were used. As this target columns are continuous and numerical it is can be seen that it is a regression problem and five such regression models (regressors) were created of which the one with the lowest bias and best result was selected. They are as follows-

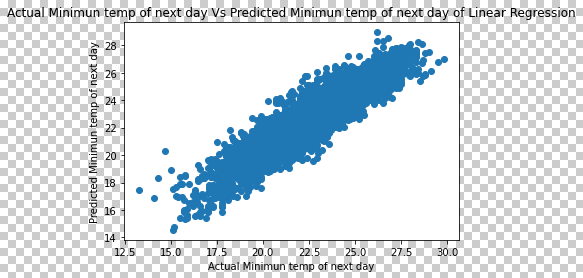
1. **LinearRegression**
2. **KNeiborsRegressor**
3. **RandomForestRegressor**
4. **Gradient Boosting Regressor**
5. **XGBRegressor**

**Linear Regression model** for minimum temperature using ‘Next\_Tmin’ as the target variable.

**Model creation**:

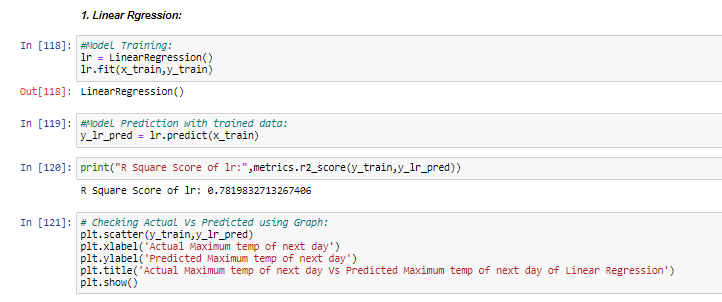


* Plotting the model in through visualizations to check the linearity of the data by matching the actual temperate of the day with the predicted data

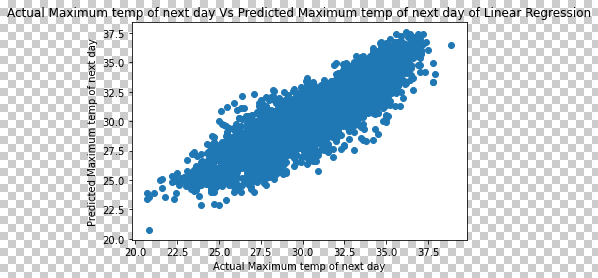


**Linear Regression model** for maximum temperature using ‘Next\_Tmax’ as the target variable.

**Model creation:**



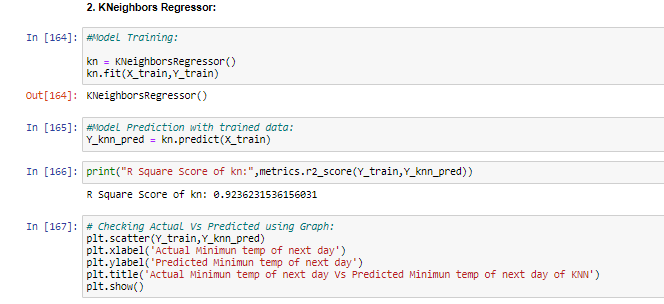
* Plotting the model in through visualizations to check the linearity of the data by matching the actual temperate of the day with the predicted data-:



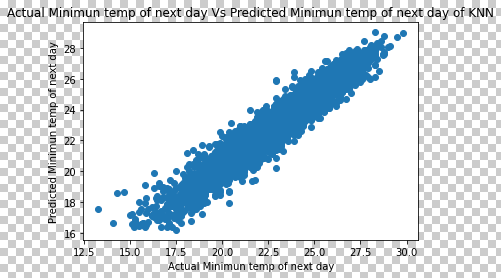
For both cases it can be seen that the movements of points after plotting is in a linear fashion but at the same across the diagonal of the plot the points are spread over on both sides.

**KNeighbor regressor model** for minimum temperature using ‘Next\_Tmin’ as the target variable.

**Model creation:**

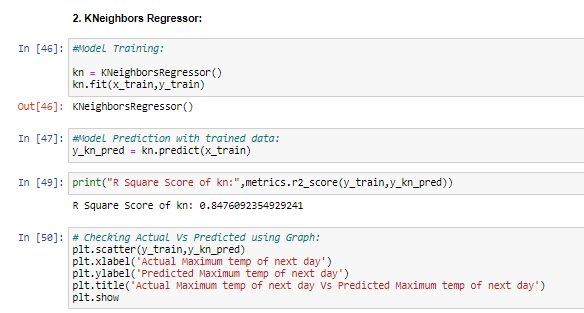


* Plotting the model in through visualizations to check the linearity of the data by matching the actual temperate of the day with the predicted data :



**KNeighbor regressor model** for maximum temperature using ‘Next\_Tmax’ as the target variable.

**Model creation:**



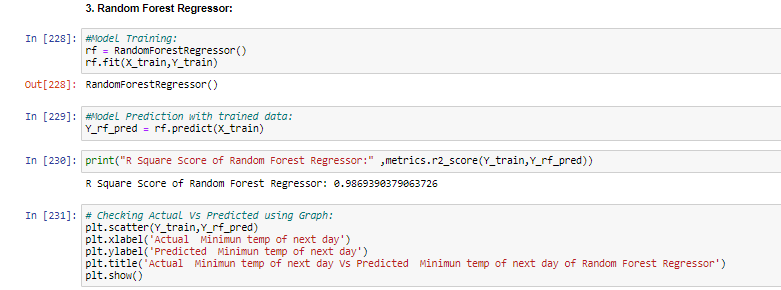
* Plotting the model in through visualizations to check the linearity of the data by matching the actual temperate of the day with the predicted data:



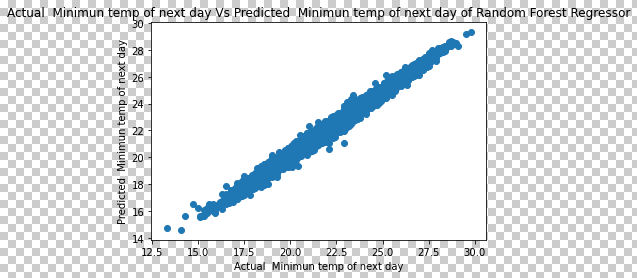
For both cases it can be seen that the movements of points after plotting is in a linear fashion but at the same across the diagonal of the plot the points are spread over on both sides.

**Random forest regressor model** for minimum temperature using ‘Next\_Tmin’ as the target variable.

**Model creation :**

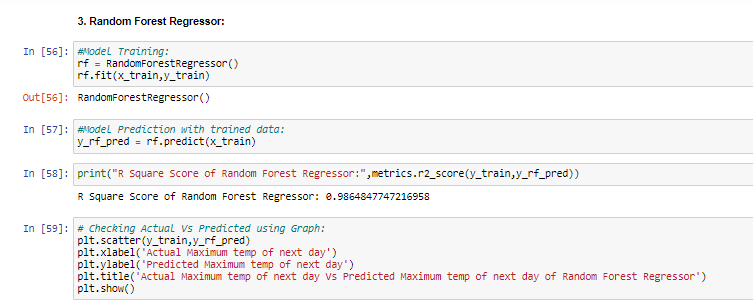


* Plotting the model in through visualizations to check the linearity of the data by matching the actual temperate of the day with the predicted data:

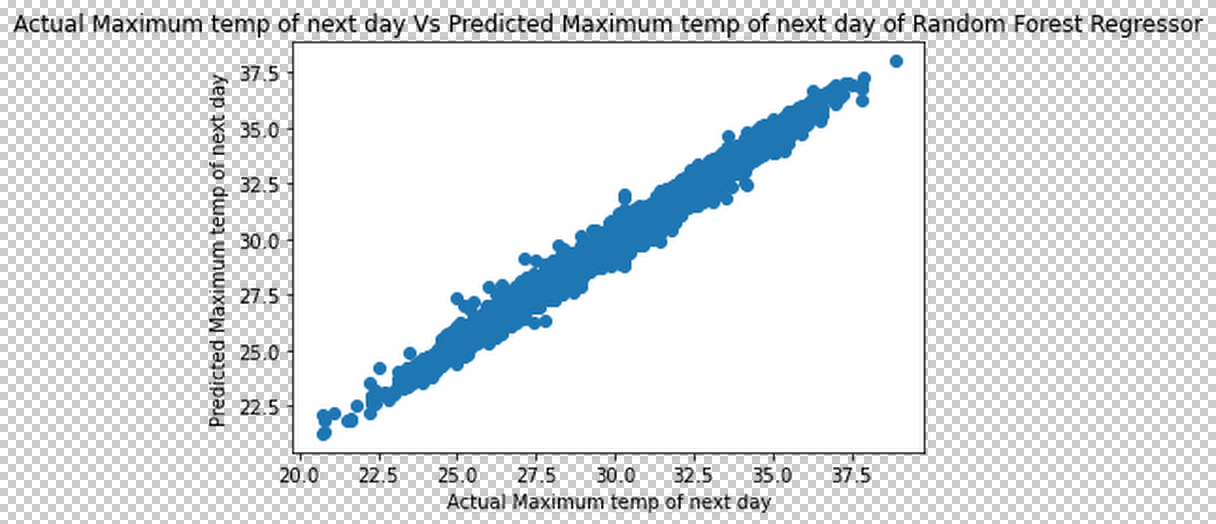


**Random forest regressor model** for maximum temperature using ‘Next\_Tmax’ as the target variable.

**Model creation :**



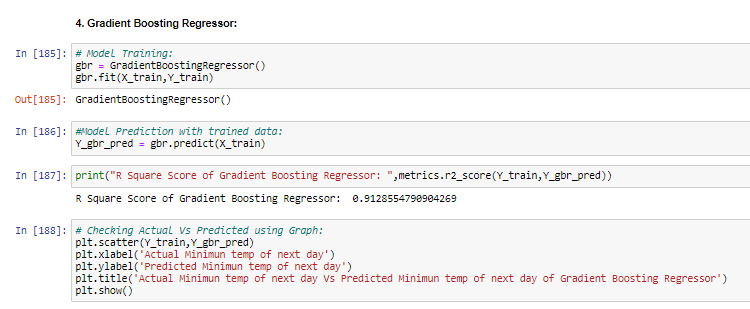
* Plotting the model in through visualizations to check the linearity of the data by matching the actual temperate of the day with the predicted data-:



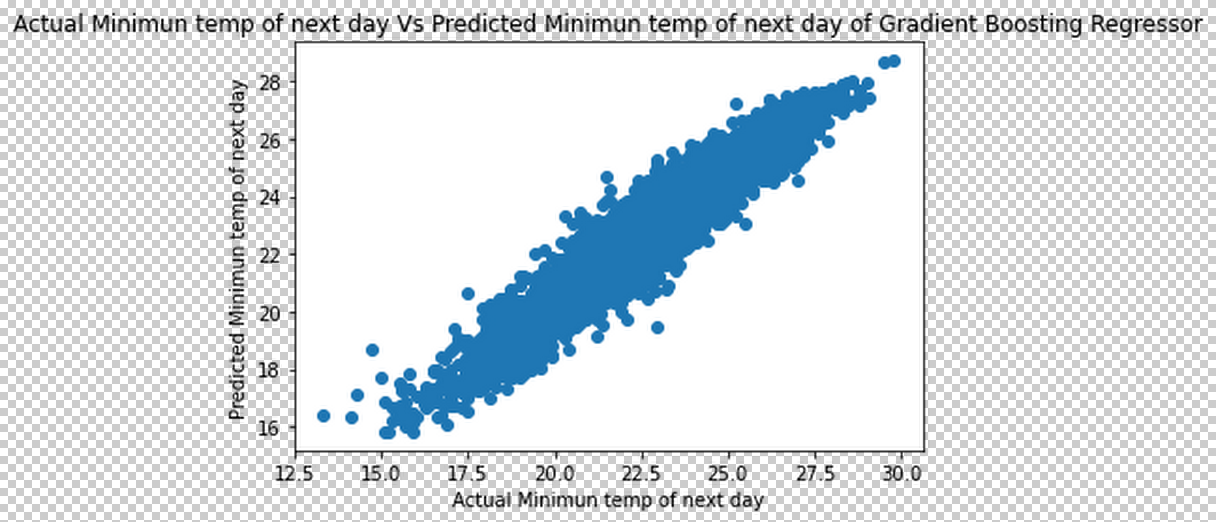
For both cases it can be seen that the movements of points after plotting is in a linear fashion but at the same across the diagonal of the plot the points are less spread over on both sides than the previous two models. This might hint on the fact that random forest regressor is by far the best model suited for predicting the temperature.

**Gradient Boosting regressor** model for minimum temperature using ‘Next\_Tmin’ as the target variable.

**Model creation:**

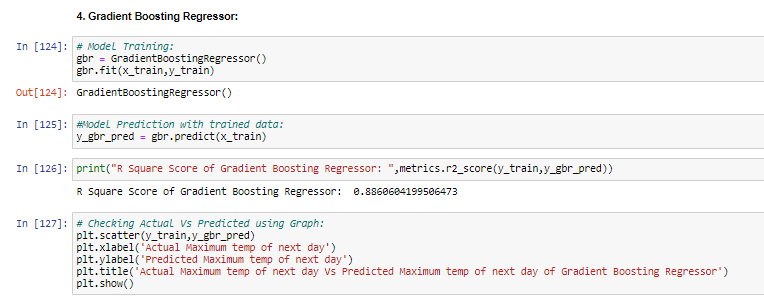


* Plotting the model in through visualizations to check the linearity of the data by matching the actual temperate of the day with the predicted data:

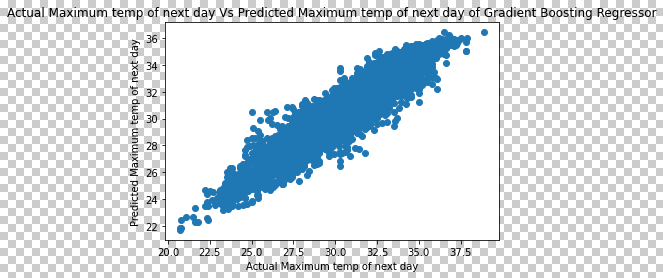


**Gradient Boosting regressor model** for maximum temperature using ‘Next\_Tmax’ as the target variable.

**Model creation**:



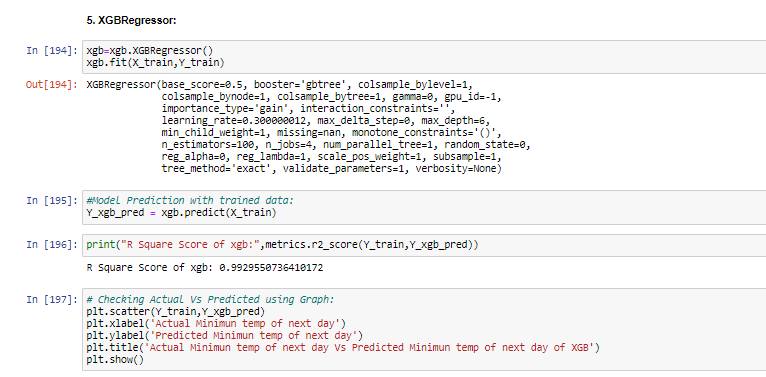
* Plotting the model in through visualizations to check the linearity of the data by matching the actual temperate of the day with the predicted data:



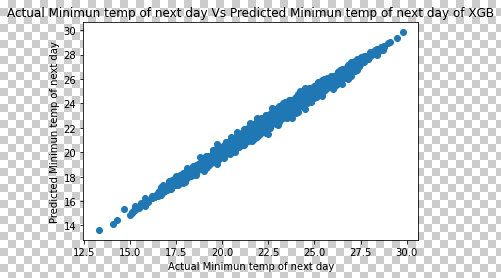
For both cases it can be seen that the movements of points after plotting is in a linear fashion but at the same across the diagonal of the plot the points are spread over on both sides.

**XGB regressor model** for minimum temperature using ‘Next\_Tmin’ as the target variable.

**Model creation:**

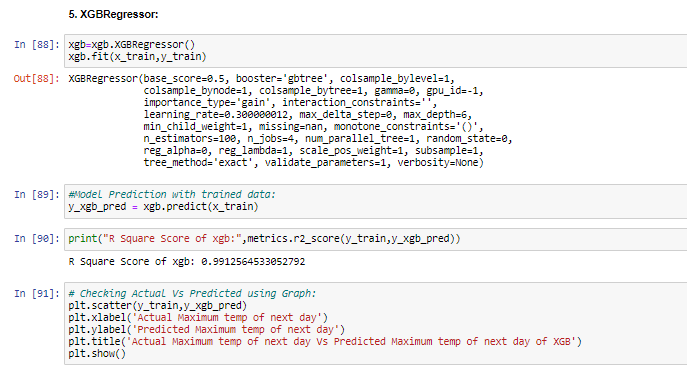


* Plotting the model in through visualizations to check the linearity of the data by matching the actual temperate of the day with the predicted data-:

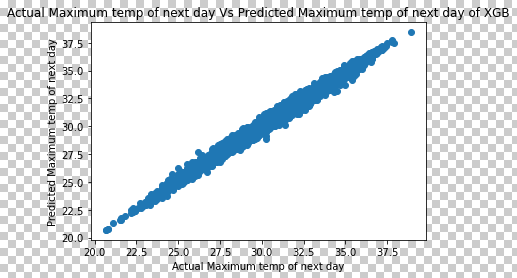


**XGB regressor model** for maximum temperature using ‘Next\_Tmax’ as the target variable.

**Model creation :**



* Plotting the model in through visualizations to check the linearity of the data by matching the actual temperate of the day with the predicted data-:



For both cases it can be seen that the movements of points after plotting is in a linear fashion but at the same across the diagonal of the plot the points are least spread over on both sides. Making the XGB regressor the best among all the five models for which the dataset was trained.

But as good data scientist we must er on the side of caution and therefore check the accuracy of these models using various accuracy metrics available to us via the sklearn library. The models must be doubled checked to make sure that there is less bias, no overfitting or underfitting so that the best results can be obtained. Checking the model accuracy gets precedence because the data that will be predicted using model will be utilised in various industries for various results so creating a faulty model will render the whole process of model creation redundant.

There are ways through which the accuracy, biasness, over/underfitting can be checked and corrected so much so that predictions are in same line with the actual values. Such tools are –

* **Regularization** - This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. Thereby reducing the noise in the dataset , basically reducing the chance of overfitting. It uses two types of techniques namely –
* Ridge regularization
* Lasso regularization

The regularizations done in this dataset have a score of –

* Ridge regularization – 0.835 (for ‘Next\_Tmin’)
* Lasso regularization -- 0.835(for ‘Next\_Tmin’)
* Ridge regularization – 0.0.766 (for ‘Next\_Tmax’)
* Lasso regularization -- 0.766(for ‘Next\_Tmax’)

For further assessment of models, the followings are used –

* **R Square Score**-The score goes on to give the accuracy score.
* **mean absolute error**- Helps to understand the variation.
* **mean squared error**- Helps to understand how close the regression line is to the set of points.
* **Root mean squared error**: Helps to score the average magnitude of error.
* **For Linear Regression scores stands as follows**-
  + **for ‘Next\_Tmin’**

Accuracy of LR - 0.8352945560410863

MAE of LR - 0.7595701436479201

MSE of LR - 0.9429025068825235

RMSE of LR - 0.971031671410631

* + **for ‘Next\_Tmax’**

Accuracy of LR - 0.7660684436689984

MAE of LR - 1.1082196967681546

MSE of LR - 2.088637831776274

RMSE of LR - 1.4452120369607617

* **For KNeighbors Regression scores stands as follows-**
  + **for ‘Next\_Tmin’:**

Accuracy of KNN - 0.883083112962173

MAE of KNN - 0.6153720696010564

MSE of KNN - 0.6693235101103757

RMSE of KNN - 0.8181219408562367

* + **for ‘Next\_Tmax’**

Accuracy of KNN - 0.7558618033095934

MAE of KNN - 1.0171607093618935

MSE of KNN - 1.0171607093618935

RMSE of KNN - 1.0171607093618935

* **For RandomForest Regression scores stands as follows-**
  + **for ‘Next\_Tmin’**

Accuracy of RF - 0.9031254203100196

MAE of RF - 0.5596701315541236

MSE of RF - 0.5626196932415777

RMSE of RF - 0.750079791249956

* + **for ‘Next\_Tmax’**

Accuracy of RF -  0.9023560622975415

MAE of RF -  0.6880023190677287

MSE of RF - 0.8718055209293417

RMSE of RF - 0.9337052644862519

* **For Gradient Boosting Regression scores stands as follows**-
  + **for ‘Next\_Tmin’**

Accuracy of GBR - 0.8899452148543628

MAE of GBR - 0.6135125851409379

MSE of GBR - 0.6300394832979826

RMSE of GBR - 0.7937502650695512

* + **for ‘Next\_Tmax’**

Accuracy of GRB -  0.8564748837163055

MAE of GRB -  0.862905120487228

MSE of GRB - 1.2814516877579805

RMSE of GRB - 1.1320122295090191

* **For XGB Regression scores stands as follows**-
  + **for ‘Next\_Tmin’**

Accuracy of XGB - 0.9326111638012836

MAE of XGB - 0.45483750547881024

MSE of XGB - 0.3857862925497224

RMSE of XGB - 0.6211169717128348

* + **for ‘Next\_Tmax’**

Accuracy of XGB -  0.8564748837163055

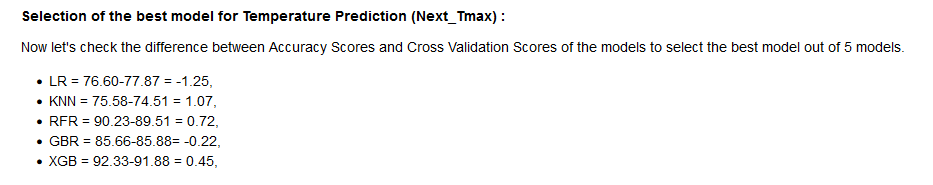
MAE of XGB -  0.862905120487228

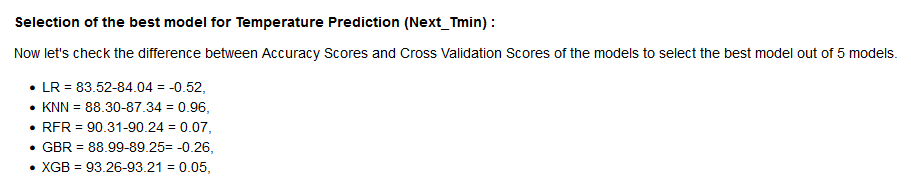
MSE of XGB - 1.2814516877579805

RMSE of XGB- 1.1320122295090191

In comparison to the scores of all the models created and trained, XGB stands out with not only the highest accuracy score but it also has the least errors. So, it is safe to say that **XGB has outperformed** all the others and it can presumed that it might be the best model to be used for predicting the data. But it can still be made better or at least try to make it a bit better by reducing overfitting. It can be done by the process of cross-validation it might help increase the scores of all the models after which the best performing model will used for hyper-parameter tunning. In this way it will allow for the regressor models to be checked multiple times for discrepancies. The cross validation and hyperparameter tunning will be done for both the target columns.

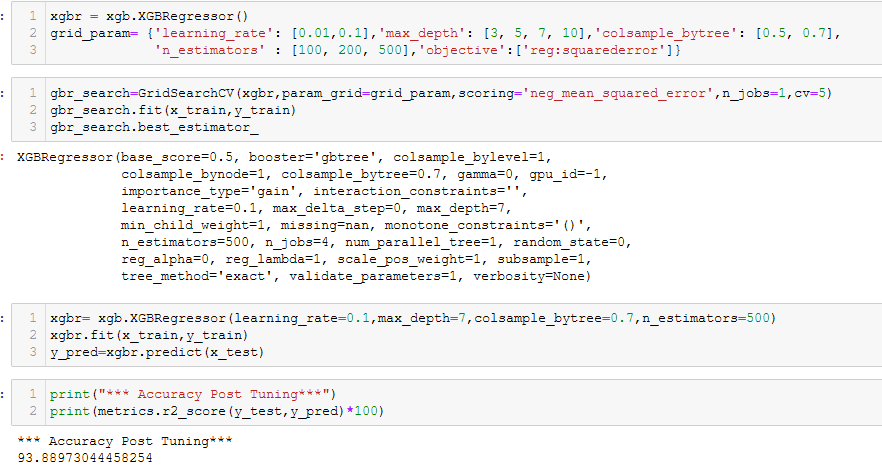
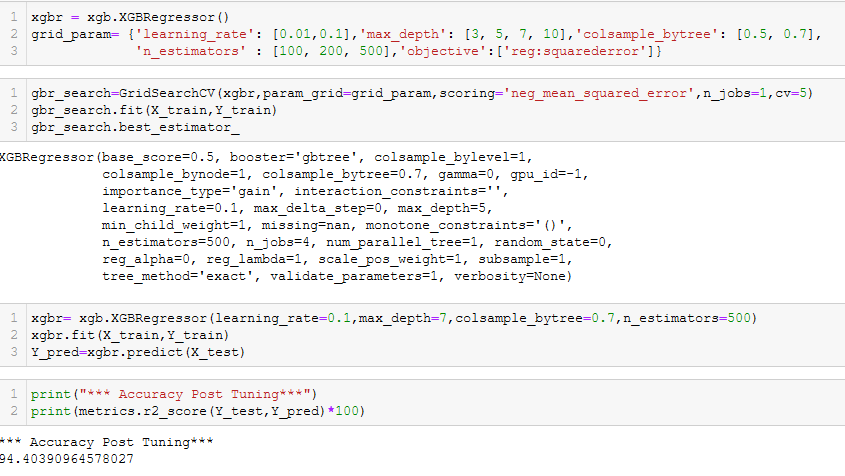
So the **Cross Validation** (CV) takes into account 5 different scenarios and gives a mean of the five scores obtained. The mean scores obtained from the cross validation of the models:-

* + **CV for ‘Next\_Tmin’**
* Cross validation score for linear regression- 84.0401
* Cross validation score for KNeighbour regression- 87.3420
* Cross validation score for Random forest regression- 90.217
* Cross validation score for Gradient Boosting regression- 89.253
* Cross validation score for XGB regression- 93.213
  + **CV for ‘Next\_Tmax’**
* Cross validation score for linear regression- 77.879
* Cross validation score for KNeighbour regression- 74.516
* Cross validation score for Random forest regression- 89.511
* Cross validation score for Gradient Boosting regression- 85.880
* Cross validation score for XGB regression- 91.887



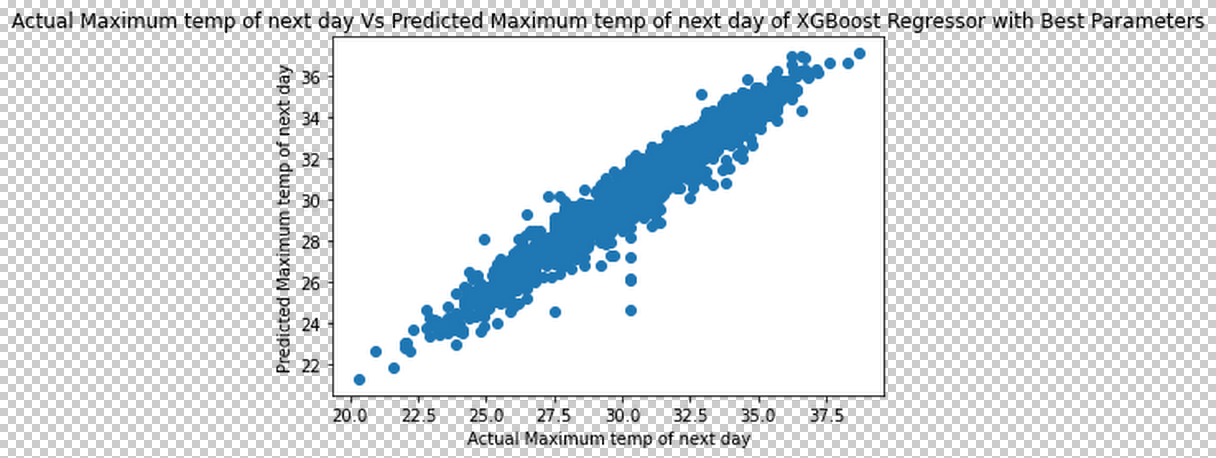
**Observation:**

Overall the result of all models are good. As most of the Accuracy of the Models are between 83-93%. As the difference between CV score and accuracy score of XGB is least and as it's having the best accuracy 93%, we are going to tune it's parameters to obtain best result. Even though the scores have not increased significantly it is safe to assume that XGB regressor is the best performing model for both the target variables and thus perfect for hyper parameter tunning.

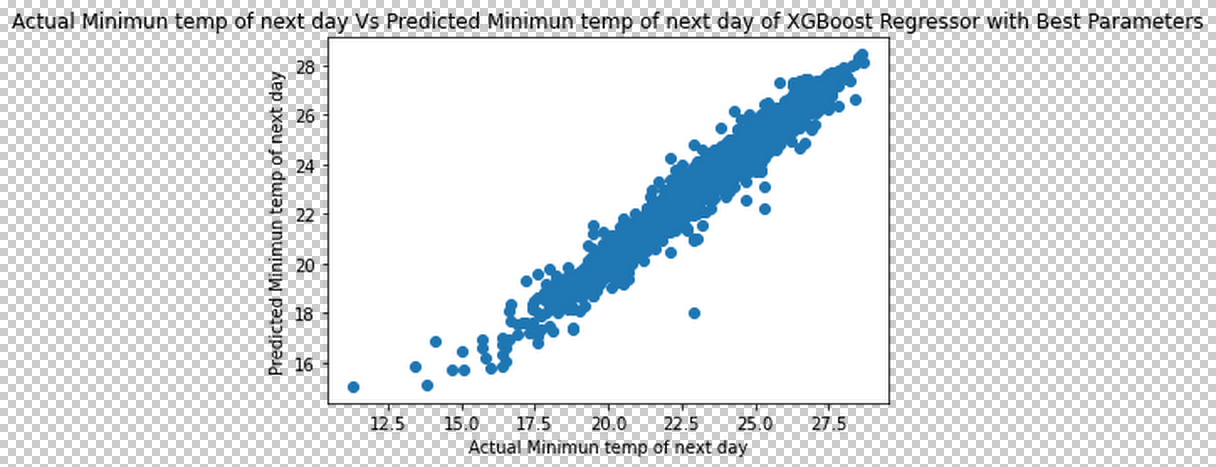
**Hyperparameter Tuning:** Now we will use Grid Search Technique to tune the hyperparameter of XGBRegressor for best accuracy for both the predictions. ****

On hyper-parameter tunning the data the accuracy post tunning further increases. The hyperparameter score – **94.40 % (for ‘Next\_Tmin)** and **93.889 % (for ’Next\_Tmax’)**. Its visualization also suggests the same.

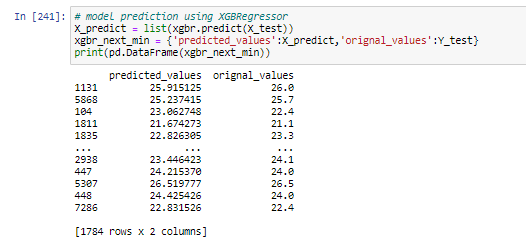
For predicting maximum temperature using hyper-tunned XGB-



For predicting minimum temperature using hyper-tunned XGB-



So the model XGB regressor has to save for future prediction of the dataset. But before finally saving model it had to be tested. More like a dry run to check whether the model is predicting correctly.



It shows XGB is predicting the minimum temperature with high accuracy (94%). The model can now be saved using joblib library.

**Concluding Remarks:**

Weather forecasting or meteorology is a essential thing for day to day life actions. It helps people to know what will be the next day's atmospheric condition. And to predict the next day's Temperature we need to understand and analyse several aspects of nature like Humidity, Precipitation, Cloud cover, Wind speed, latitude and longitude of that area, solar radiation and most importantly the previous year’s data of the same area. We have analysed those aspects and we have come to the conclusion that,

1. Next day maximum temperature has highest positive correlation with LDAPS\_Tmax\_lapse( next-day maximum air temperature applied lapse rate), Nextday's maximum temperature has highest negative correlation with LDPAS\_CC3(next-day 3rd 6-hour split average cloud cover (12-17 h)).
2. Next day's minimum temperature has highest positive correlation with LDAPS\_Tmin\_lapse(next-day maximum air temperature applied lapse rate)), Next day's minimum temperature has highest negative correlation with DEM (Elevation). It means lapse rate, cloud cover and Elevation have very strong effect on temperature of the area.

After analysing those points, we have made 5 models and predicted both minimum temperature of next day and maximum temperature of next day. out of those 5 models XGBRegressor has given the best accuracy for both the labels (minimum temperature (93%) and maximum temperature (93%)).

Forecasting the temperature is essential to understanding the weather and as an extension be able to forecast it. In the future weather mankind will be able to control the weather still remains a large question but certainly being able to predict it with utmost accuracy can be said as the next best thing. As whether plays a looming effect in our lives it perhaps most important and definitely a great feat to be able to predict is and use it to our advantage. Which makes machine learning a great weapon that can used. It would definitely be interest to observe how the rise in temperature or drop in temperature affects different aspects in the nature starting from environmental, economic, political, behavioural, etc.