**Exploratory Data Analysis to predict high temperatures Using Linear Regression**

**Synopsis:**

This analysis will try to answer the following question from a data source which collected weather details in Austin over 2 years:

* Across Austin, how the temperature varies and taking maximum temperature as input we will predict the maximum temperature

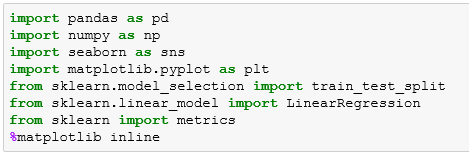
Data Source: Kaggle

Attributes: Date, TempHighF, TempLowF, TempAvgF, HumidityHighPercentF, HumidityLowPercentF, HumidityAvgPercentF etc.,

Size: 1321 X 21

**Exploratory Data Analysis:**

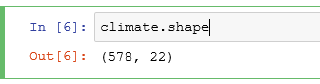
Importing all the required libraries



Reading data from CSV file using Pandas

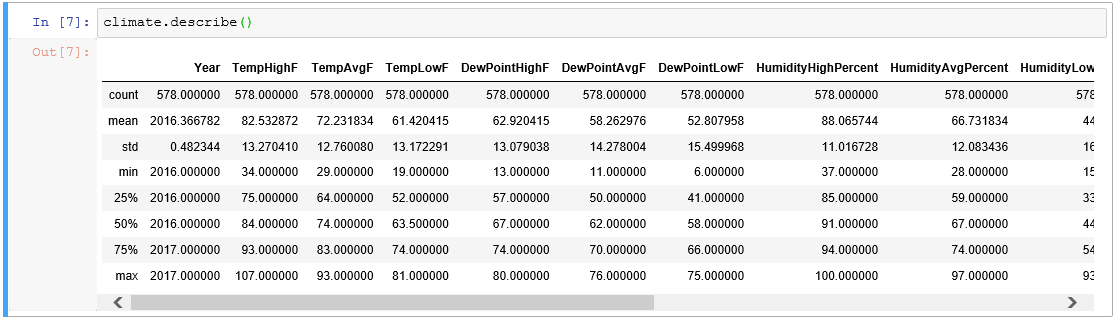


Exploring the dataset by checking the number of rows and columns

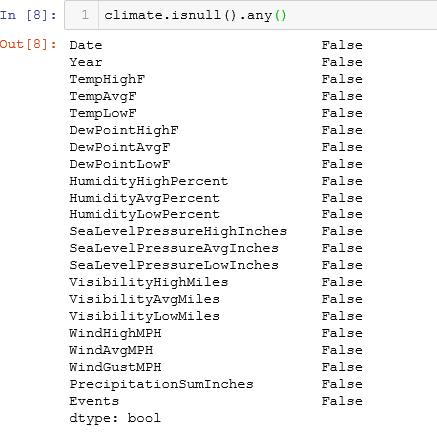


We have in total 578 Rows with 22 Columns

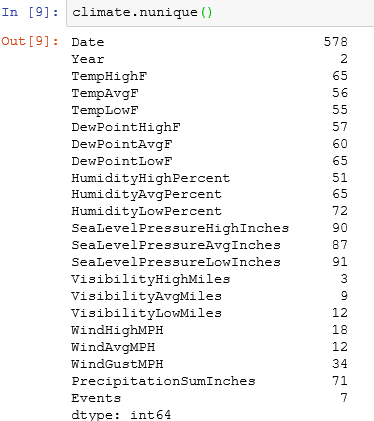
Checking the Statistical details of the data set



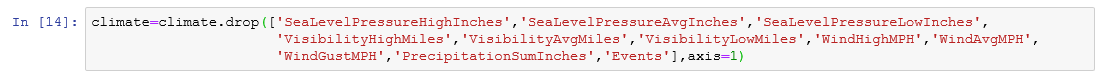
Checking for null values in the data, hopefully our data doesn’t have any null values



Checking the unique columns of the data set to perform the data analysis

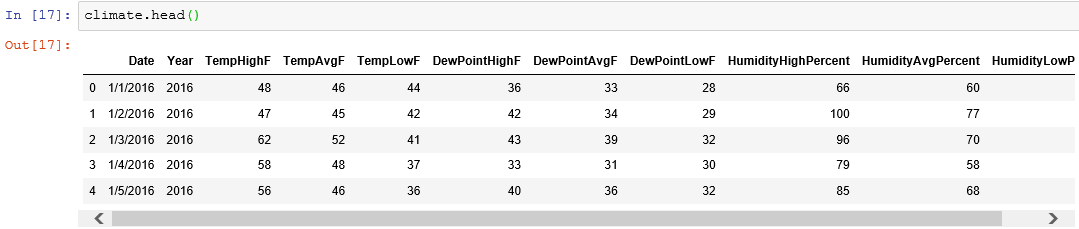


Cleaning the data where the redundant fields are not necessary and will not impact our analysis.

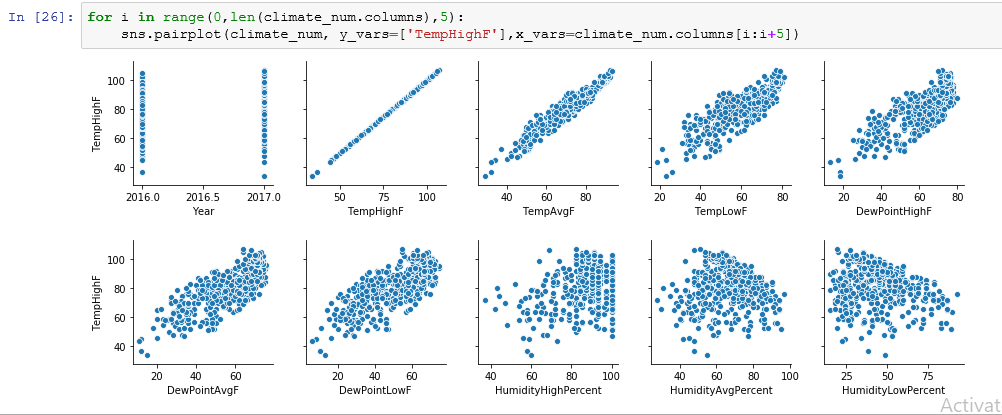


Again checking the shape of our data set and checking the first 5 rows and columns



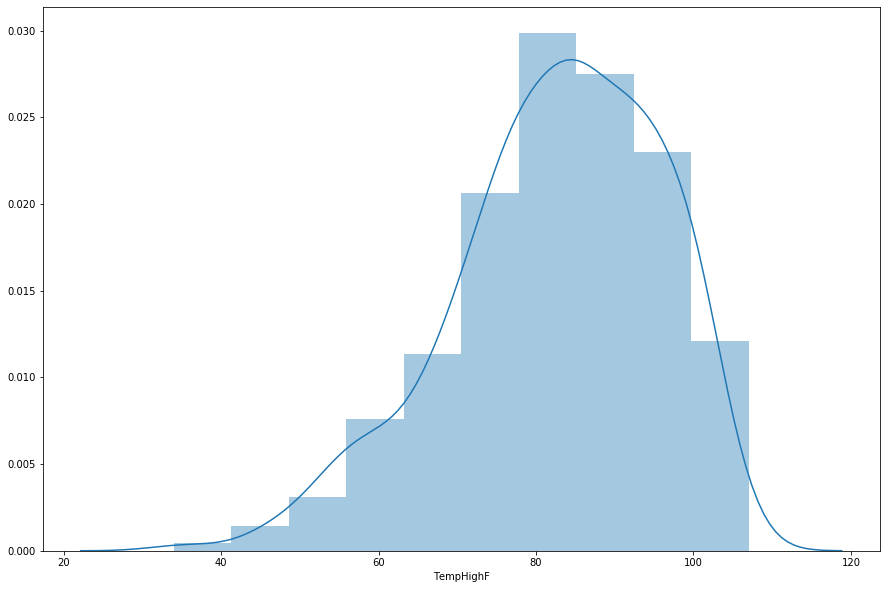


Using Seaborn, plotting a pair plot to understand the relation between the fields in a wider scope.

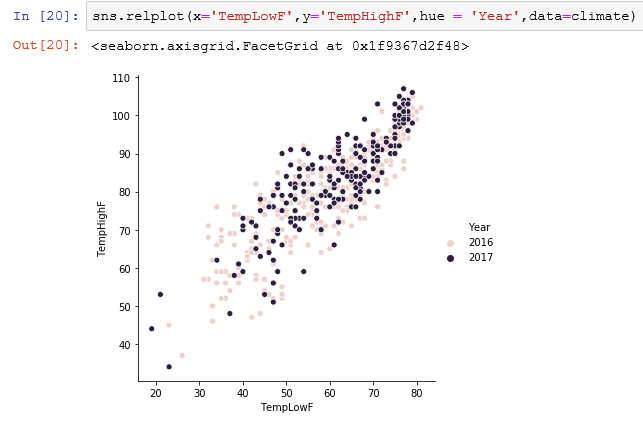


After observing the pair plot, it was evident that most of the data from our data set are having a linear relationship, but before going to perform cross validation and in order to conclude the relation we will use few more visualizations.

Using a distplot to check the maximum temperature, as we can observe from the below plot the maximum temperature lies between 80 to 100.

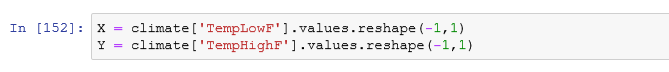


Using relational plot between the TempLowF and TempHighF values with hue as the year we analyzed the patterns.



Linear Regression Prediction:

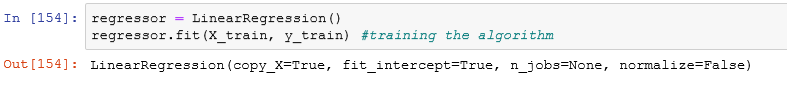
We are dividing the data into “attributes” and “labels”. We want to predict the Maximum Temperature depending upon the Minimum Temperature recorded. Therefore, our attribute set will consist of “TempLowF” column and which is store in X variable, and the label will be the “TempHighF” which is stored in y variable.



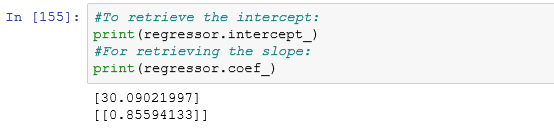
The data has been split with 80% as training set and 20% of the data to test set.



Calling fit method to along with the training data



The linear regression found the best value for the intercept and slope, which results in a line that best fits the data.

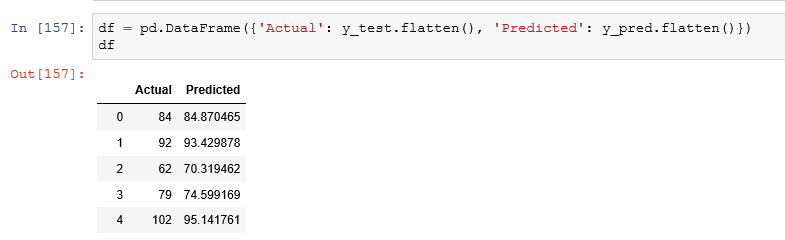


It is clear from the intercept and slope that for every unit change in minimum temperature, the change in the maximum temperature is about 0.85%

Let us make some predictions using our test data and check how accurately the algorithm is predicting the temperatures.



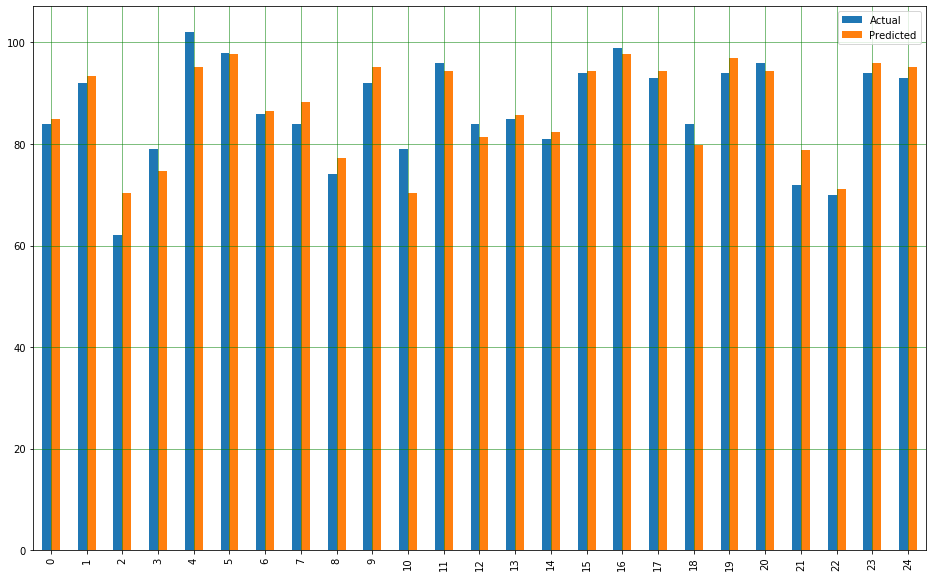
Let us compare the actual output values for X\_test with the predicted values.



For representation purpose taking first 25 records from the result and plotting a bar graph for better visualization

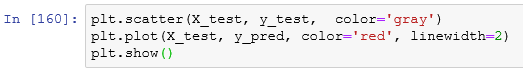


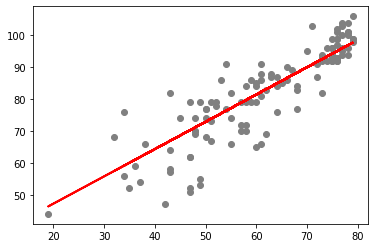
Though our model is not very precise, the predicted percentages are close to the actual ones



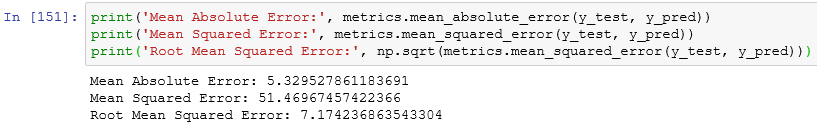
Bar graph showing the comparison of Actual and Predicted Values

Scatter plot with test data and a predicted straight line





Calculating the Root Mean Square Error (RMSE) using Scikit-Learn library



**Conclusion**:

The root mean squared error is 7.17, slightly less than 10% of the mean value of all the temperatures i.e., 72.23. This means that our algorithm was not accurate but can still make reasonably good predictions.