STATISTICS FOR DATA SCIENCE

D PAWAN KALYAN REDDY
S20160010020

STUDY ON MULTI-LINEAR REGRESSION & TIME SERIES

ABSTRACT

In this study, The dataset that we used contains 9358 instances of hourly averaged responses and 15 variables are analyzed using Multiple linear regression and Time series. Among 15 variables, the two dependent variables are Relative humidity and Absolute humidity. A multiple linear regression analysis is carried out to predict the values of a dependent variable, Y, given a set of p explanatory variables. This is done by fitting a regression model, measuring the goodness of fit, testing the assumptions and finally doing model adequacy checking. By analyzing the correlation matrix and by Bartlett's Sphericity test, Principle Component Analysis(PCA) is applied.

A Time series is a sequence taken at successive equally spaced points in time. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.

PROBLEM STATEMENT

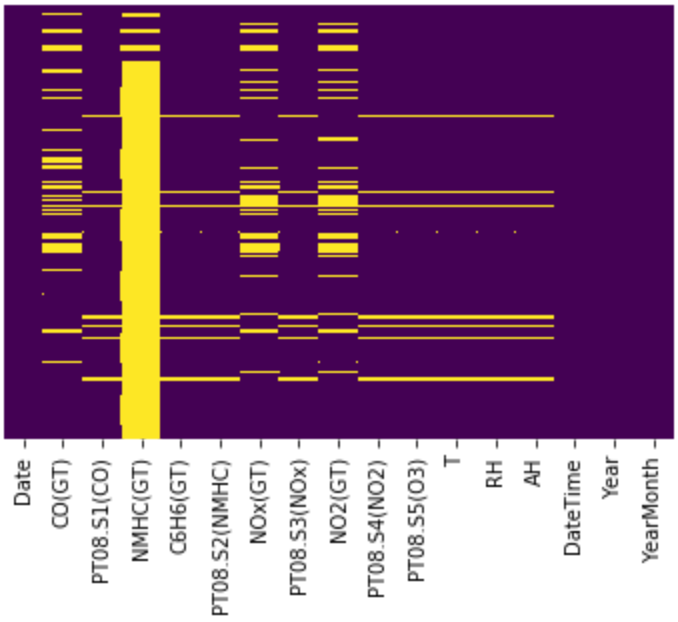
The goal is to perform multiple linear regression and time series analysis on the Air Quality Dataset(UCI) with 15 features collected at 9358 various instances of time and to predict the relative humidity and absolute humidity from the given time series.

METHODOLOGY

1. Pre Processing
2. Multiple Linear Regression
3. Time Series Visualization
4. Stationarity of the Series
5. Fitting a model to predict the dependent variables

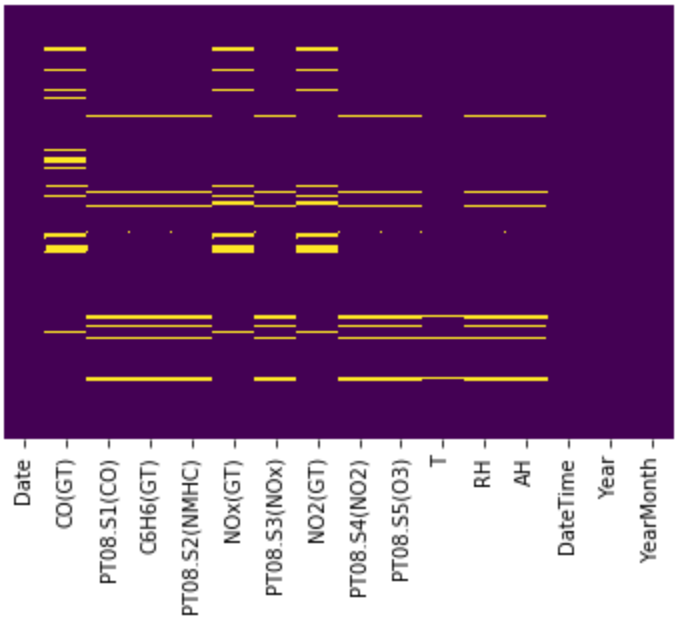
Pre-processing

We drop all the rows which have NaN values for all the columns and the two columns for which the values for all the rows are NaN. Now the data is free of all the null values, but there are more unknown values which are -200 in value, as given in the description of the dataset. These -200 values in the data should be imputed with appropriate values (or be removed). Firstly, let us replace the dummy value of -200 by NaN throughout the dataset. Now let us visualize the distribution of the null values using a heat map.



As we can see the NHMC(GT) column is missing many values, more than 85% values are NaN. Thus we can remove this column from the dataset as these values are very less significant in this dataset. We can try to replace these NaN values by taking the mean or the median of the whole column, but this won't be an accurate and proper way to fill those values. Thus, filling the mean of that particular day in which day the value is NaN makes more sense and would be the proper way to impute the values. Hence, we group by the date and take the mean and replace any NaN value by the mean of the values of that day.

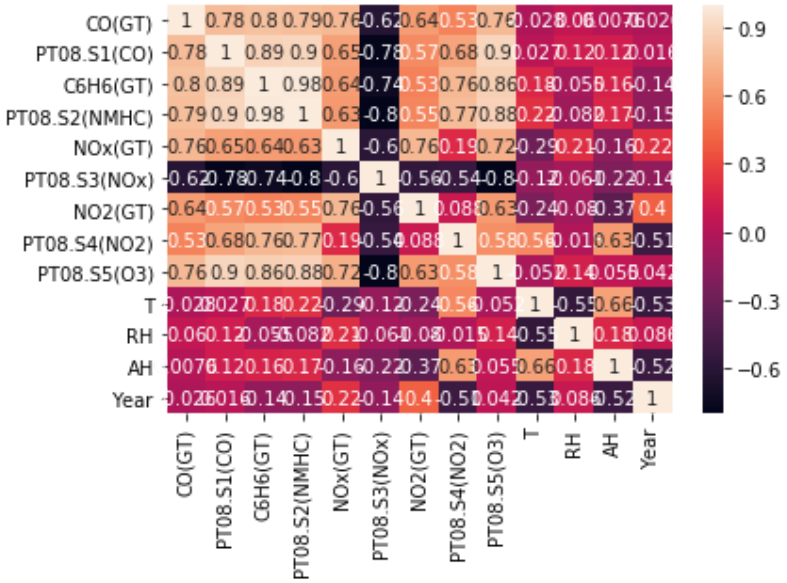
As we can see there are still some NaN values in the below figure, this is because for these values, and for their corresponding date all the values for that date are NaN, hence the mean is also NaN. In this case, We use forward fill here as even now taking the mean of the whole column(which has values for an entire year) does not make sense.



Now that the data is all cleaned and free of NaN values.

Multiple Linear Regression

Before starting our statistical analysis, we have to know the correlation between the variables.



Modeling

In our dataset, we have two dependent variables i.e., Relative humidity and Absolute humidity. We consider remaining variables as independent variables.

So, we need to do the analysis for two dependent variables once at a time. Hence, we consider Y[1] which has RH as a dependent variable and Y[2] which has AH as a dependent variable.

Model Adequacy Testing

Goodness of fit

The goodness of fit of a statistical model describes how well it fits a set of observations. R-squared and Adjusted R-squared are a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively.

Adjusted R-squared is used over R-squared to check the goodness of fit as R-squared has some problems.

For the above model, it is observed from the OLS summary that the

Y (Dependent variable)	R-squared	Adj. R-squared
RH	0.915	0.915
AH	0.885	0.885

Test of Individual parameters

1. After applying the regression model, we get the estimates of βj's that tell us the significance of each feature.
2. To check this significance, we test each parameter using the hypothesis testing with null Hypothesis : $\beta j = 0$; alternative hypothesis : $\beta j \neq 0$; with level of significance, $\alpha = 0.05$.

3. If the P-value > α, we fail to reject the null hypothesis. After testing the individual parameters ,no features are removed.

After this, The OLS method is again applied. The following table is the result of it

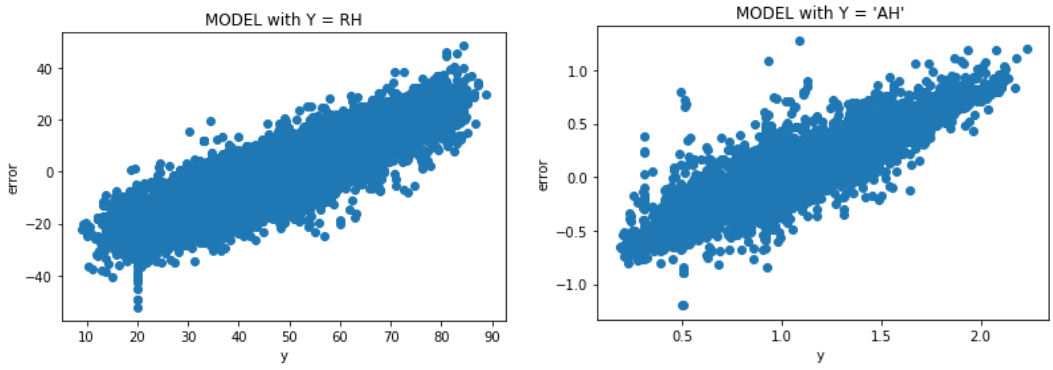
Y (Dependent variable)	R-squared	Adj. R-squared
RH	0.915	0.914
AH	0.885	0.885

TEST OF ASSUMPTIONS

1. Test of linearity

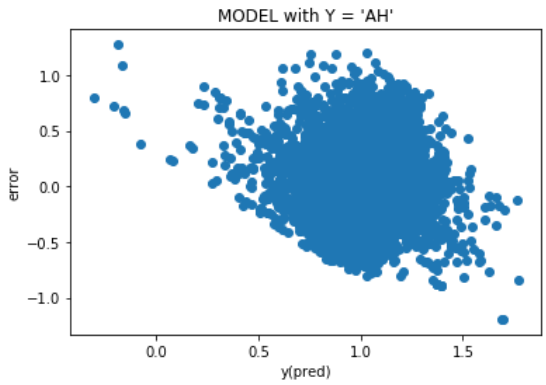
The most important assumption that we take while fitting a regression model is that the residuals are linear. We can test this by plotting the residuals versus actual prediction plot.

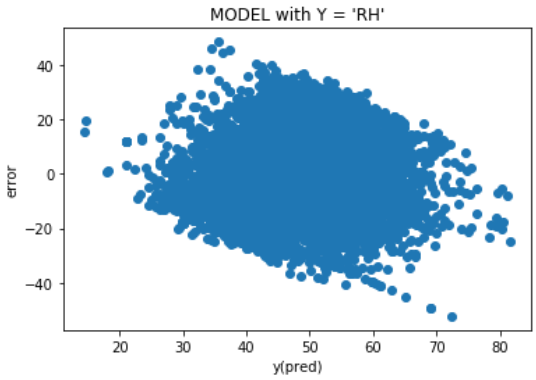
From the below plots, it is clear that both the models are linear.



2. Homoscedasticity

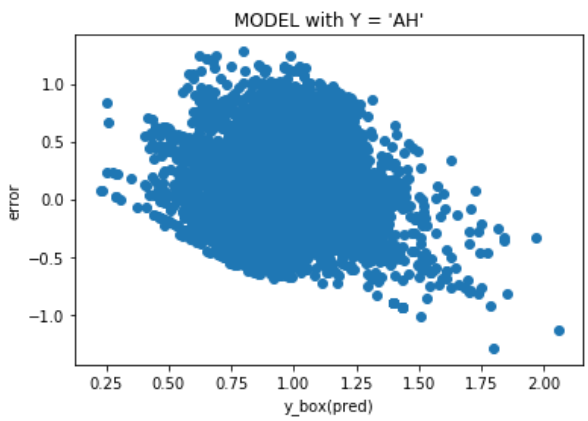
To check the homoscedasticity a graph is plotted with the residuals and the predicted values of training Y. The graphs have funnel shape which concludes the model has heteroscedasticity.

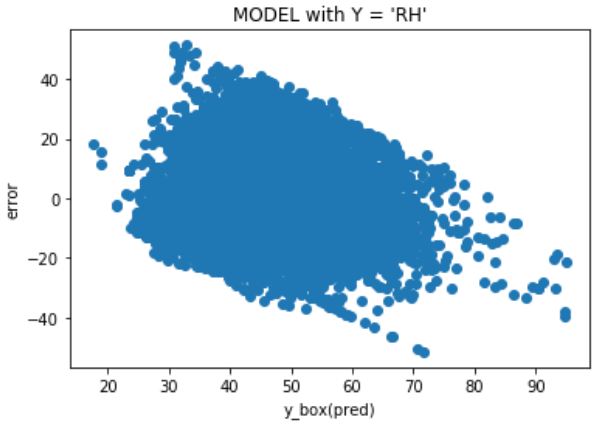




We have to transform the Y to another domain to overcome this problem. The method is BOX-COX method where the parameter for the transformation, λ is selected in such a way to reduce the Squared error. The λ obtained is 0.15, that is used to transform Y. After the transformation, the model is fit to the regression model and Adjusted R squared is checked.But we should transform the predicted Y values to the original domain.

Y (Dependent variable)	R-squared	Adj. R-squared
RH	0.983	0.983
AH	0.895	0.895





The above plots are like in between two parallel lines which states that the residuals are homoscedastic.

3. Uncorrelated Error

To test uncorrelated error we have to use Darwin - Watson test where the statistic is DW = 2(1-r) .

1) If DW = 2, there is no correlation. 2) If Dw > 2, residuals have negative correlation

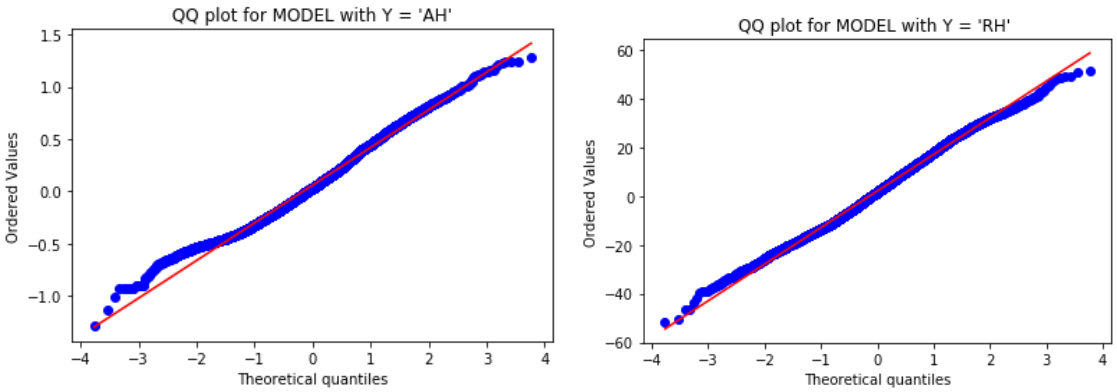
3) If DW < 2 , residuals have positive correlation

The DW values are 1.939643 and 2.010600 for Y = AH & RH respectively which are nearly equal to 2, that implies that residuals are uncorrelated.

4. Normality

The Normality test is done using Q-Q plot. A Q–Q plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

The residuals are plotted against these quantiles. The graphs shown below are linear without large deviation, using that we can state that the residuals are normally distributed.

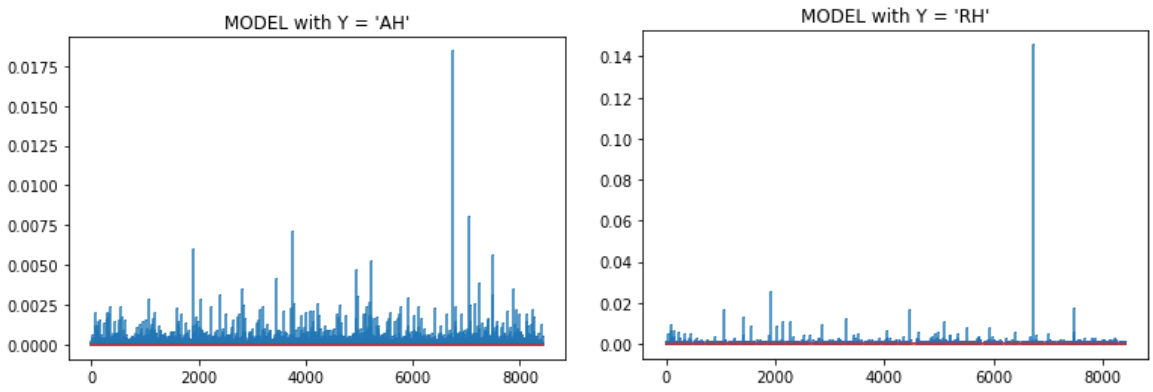


Model Diagnostics

In model diagnosis, we have to detect influential points. This can be done using Cook's distance criteria. Generally, if Cook's distance is greater than 1, it is said to be influential point.

In our analysis, there are no such points whose cook's distance is greater than 1.

Thus no influential points.



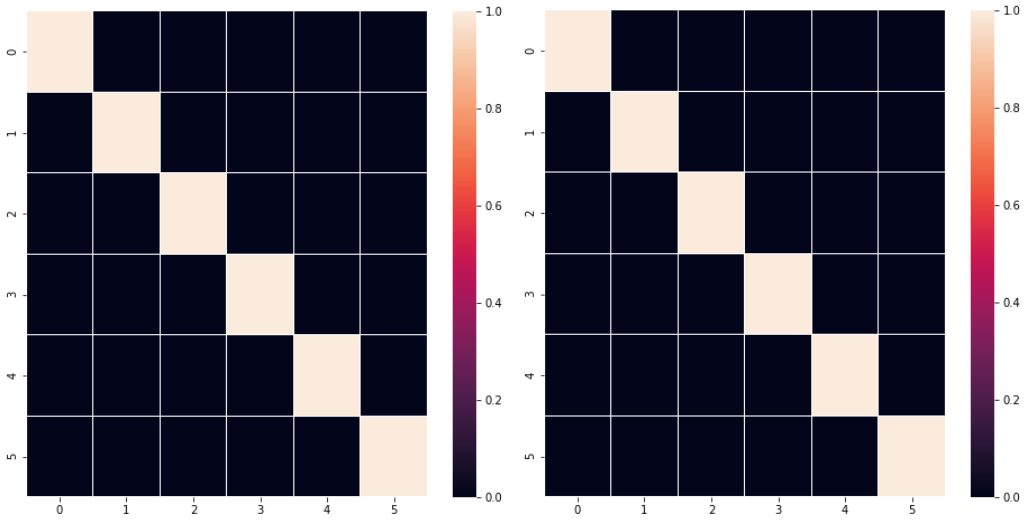
Principal Component Analysis(PCA)

Principal Component Analysis is used to dimension reduction technique. We obtain a set of Principal Components which summarize, as well as possible, the information available in the data. The factors are linear combinations of the original variables. The approach can handle only quantitative variables.

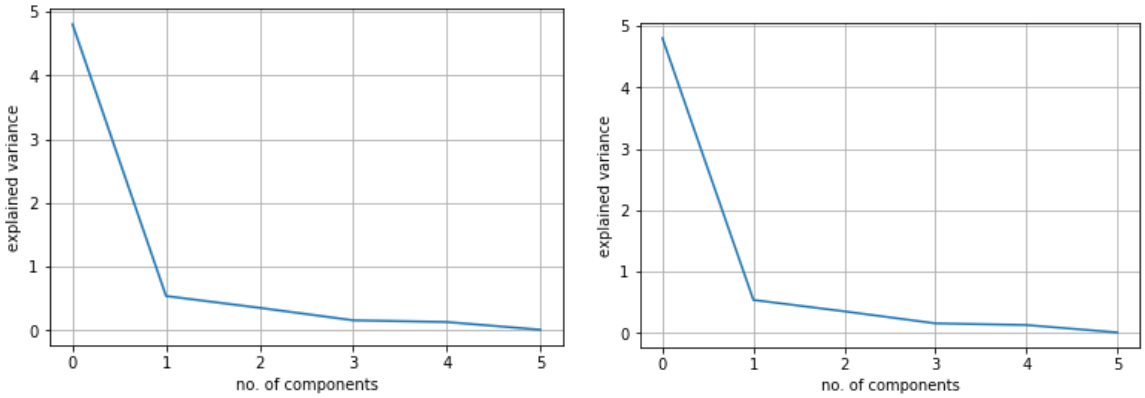
Implementation

We fit the PCA model to the dataset after standardizing it and graphs of explained variance and cumulative variance are plotted. According to the graphs, we select the number of components that are enough to explain most of the variance.

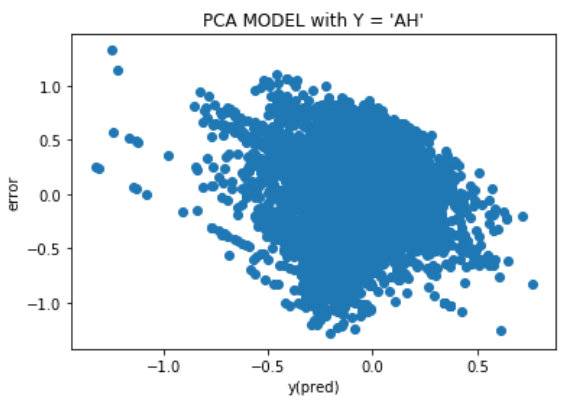
Correlation matrix is plotted to verify if the there are still correlated principal components and we can also notice that the extracted PC's are uncorrelated.

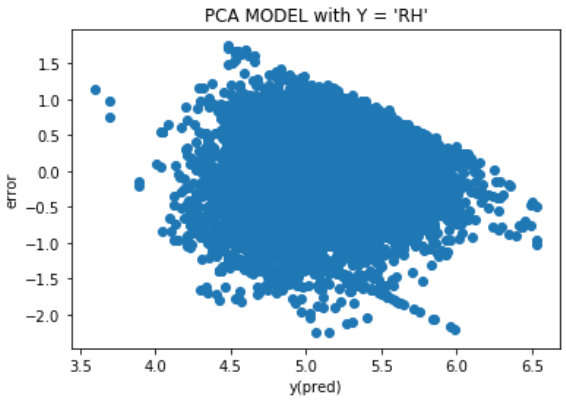


Then the explained variance and cumulative variance graphs are plotted.
For the Graph of explained variance, the number of PC to be used are selected in such a way that the components that explain too less variance are ignored. In this case the number of components derived are 6.

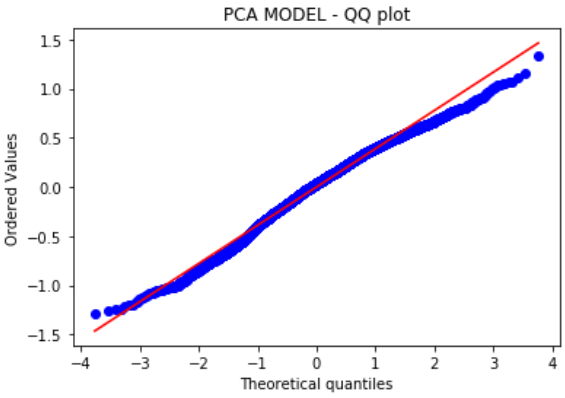


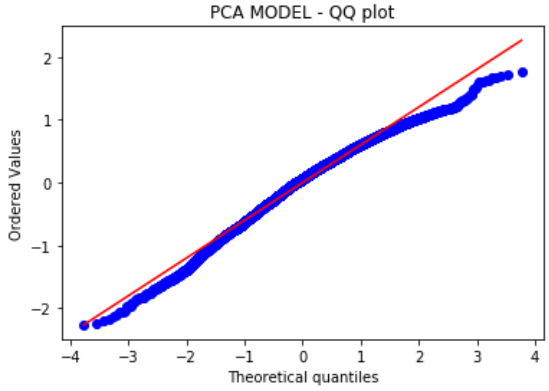
Graph is plotted between the residuals and predicted values of Y, which resulted in homoscedasticity.





The Q-Q plot for model with Y = RH & AH are also plotted and residuals followed normal distribution respectively.

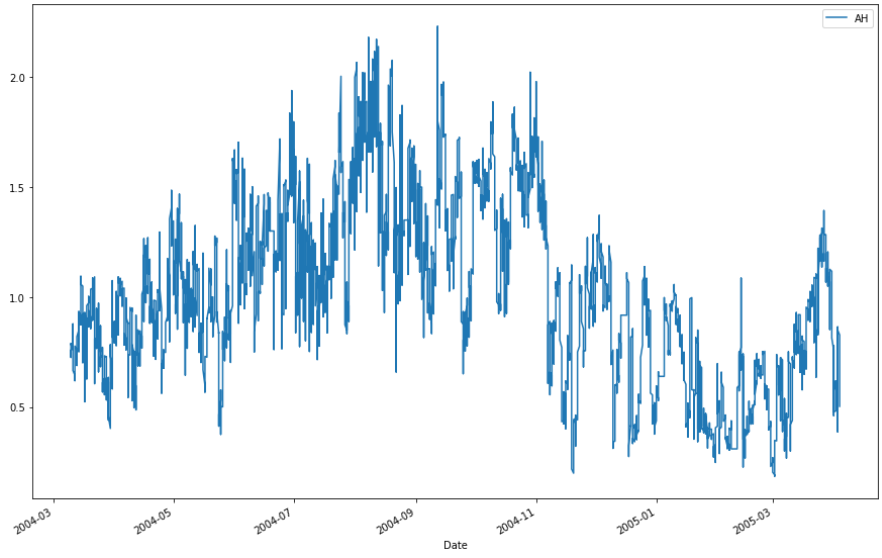


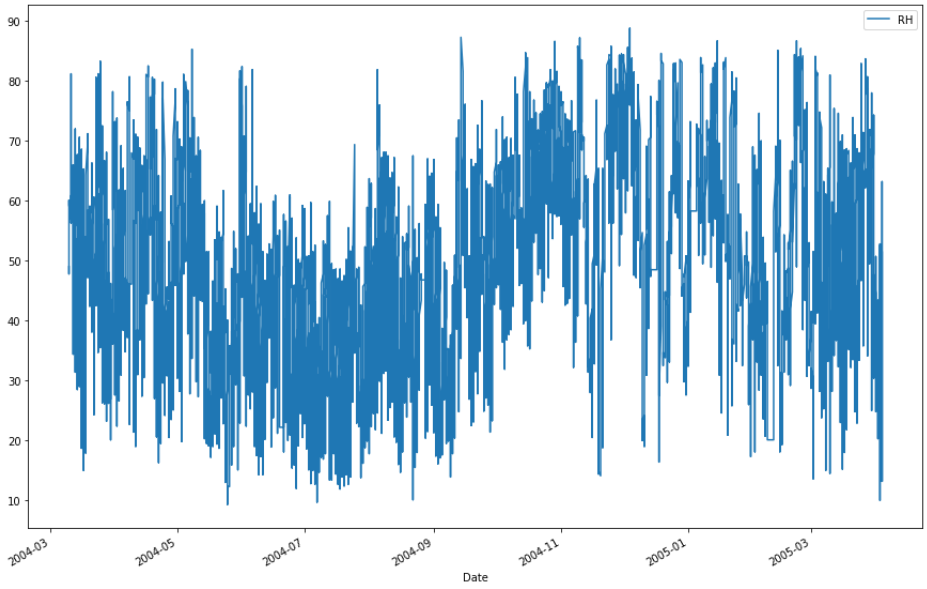


Finally, after the analysis it resulted that the regression model with 15 variables,after applying Box-Cox transformation has the best Adjusted R-Square of 0.983 & 0.898 respectively, for model with Y = RH & AH when compared to all other models implemented here. This model is also satisfying all the assumptions taken for fitting a regression model.

Time Series Visualization

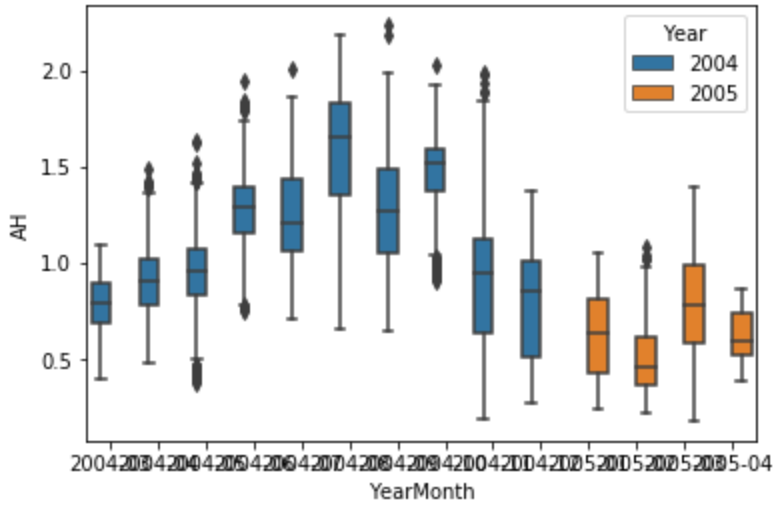
The plot between Relative humidity and Absolute humidity with Datetime as follows respectively.

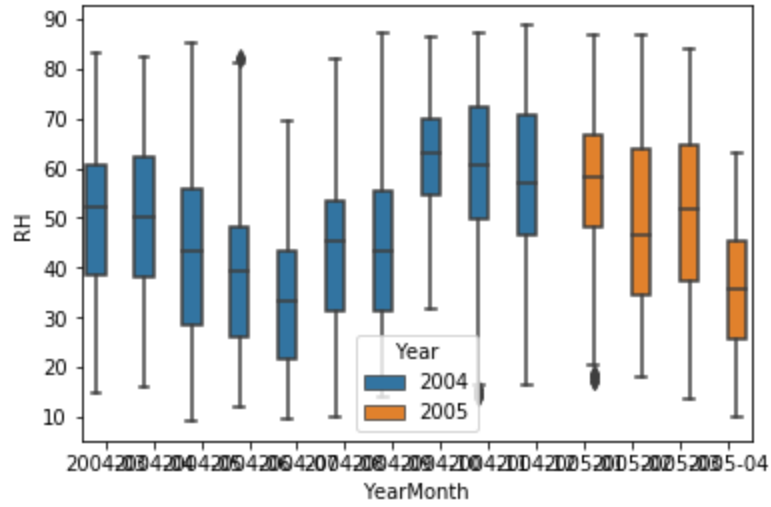




Boxplots:

The box plots of RH & AH with year are respectively as shown below.





Stationarity of the series:

Dickey-Fuller Test: Dickey-Fuller test is one of the statistical tests for checking stationarity. The null hypothesis is that the time series is **non-stationary**. The test results consist of a Test Statistic and some Critical Values for different confidence levels.

If the **'Test Statistic' < 'Critical Value'**, we can reject the null hypothesis and conclude that the series is stationary.

results of Dickey-Fuller Test:(Absolute Humidity)

Test Statistic -5.494518

p-value 0.000002
#Lags Used 25.000000
Number of Observations Used 9331.000000
Critical Value (1%) -3.431051
Critical Value (5%) -2.861850
Critical Value (10%) -2.566935
dtype: float64

results of dickey-fuller test for RH:
Test Statistic -9.995876e+00
p-value 1.940944e-17
#Lags used 3.800000e+01
Number of observations used 9.318000e+03
Critical Value (1%) -3.431052e+00
Critical Value (5%) -2.861850e+00
Critical Value (10%) -2.566935e+00
dtype: float64

For both cases, the test statistic is less than the critical value(1%).
Hence we reject the null hypothesis and state that the series is stationary.

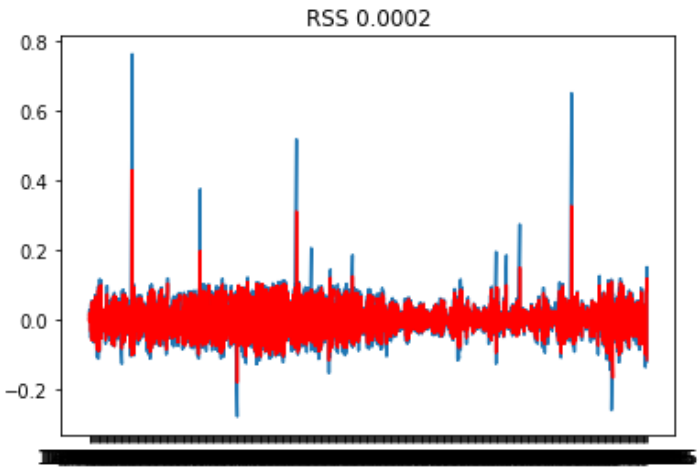
Fitting the model:

1. We have to use differencing on the response variables to fit the model. Differencing involves transforming the feature into logarithmic form, finding the shift and their difference. This difference is taken as our response variable and model if fit to find this difference.
2. The given data is not strictly stationary. we use the ARIMA model to forecast the data. The ARIMA predictors depend on (p,d,q) where
   1. p is the number of Auto Regressive terms
   2. q is the number of Moving Average terms
   3. D is the number of differences
3. The p, q, can be found using two plots namely Autocorrelation Plot(ACF) and Partial Autocorrelation plot(PACF)

For p= 3, q=4. Using these values we fit the ARIMA model and predict the difference.

The below plot shows the predicted values(Red) to the original values (Red) of the difference.

Relative Humidity log difference

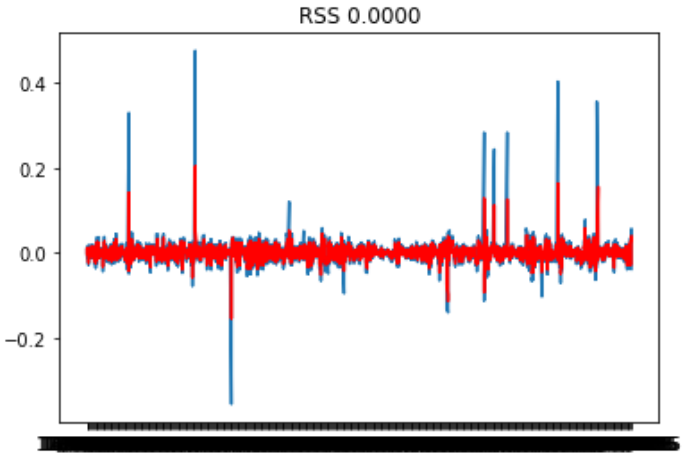


Predicted vs Actual values

For p= 2, q=2. Using these values we fit the ARIMA model and predict the difference.

The below plot shows the predicted values(Red) to the original values (Red) of the difference

Relative Humidity log difference



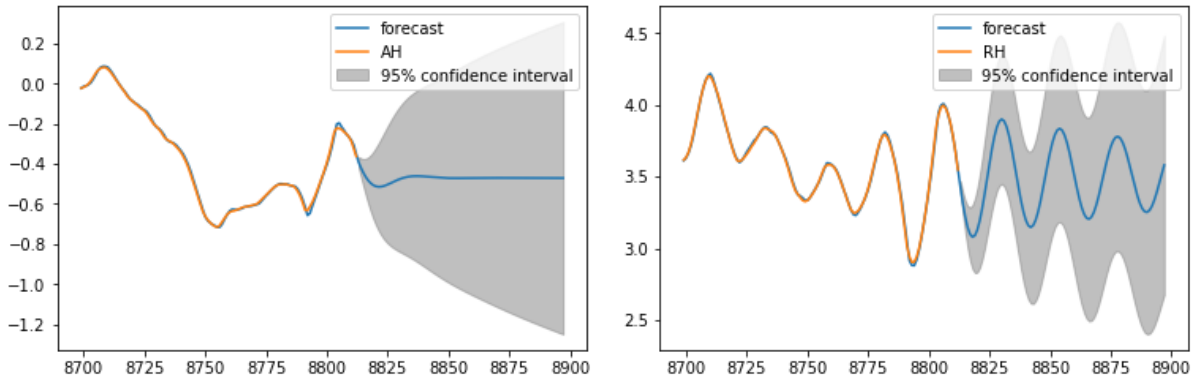
Predicted vs Actual values

ARIMA FORECAST

This helps us to predict values for next set of months. Here, we used this prediction to find the values of the next 10 years. The below code snippet allows us to predict the future values based on previous data using ARIMA.

results_ARIMA.plot_predict(8700,8898)

The plot for the forecast is as follow for AH & RH respectively.



Conclusion:

The given dataset with 15 variables was removed of missing values and was analyzed. It was a non-strict stationary time series. ARIMA forecasting was used to fit a model and predict the dependent variables i.e., Relative Humidity and Absolute Humidity and also predicted the future values using ARIMA.