STATISTICS FOR DATASCIENCE ASSIGNMENT

Analysis of Regression on Air quality Dataset

Name: Upendra Reddy
Roll no: S20160010038

Problem Statement:

The goal is to perform linear regression and time series analysis on the UCI Air quality dataset which contains 15 features and 9358 instances of hourly averaged responses from chemical sensors embedded in an Air Quality Chemical Multi sensor Device.

Abstract:

Time series analysis is a statistical technique that deals with time series data, or trend analysis. Time series data means that data is in a series of particular time periods or intervals. In this report we get to know about the seasonal trends in the data and various tests to check the stationarity and other regression assumptions. Time Series Analysis helps us in understanding the trends in the data which is useful predicting the outputs by fitting appropriate models.

Data:

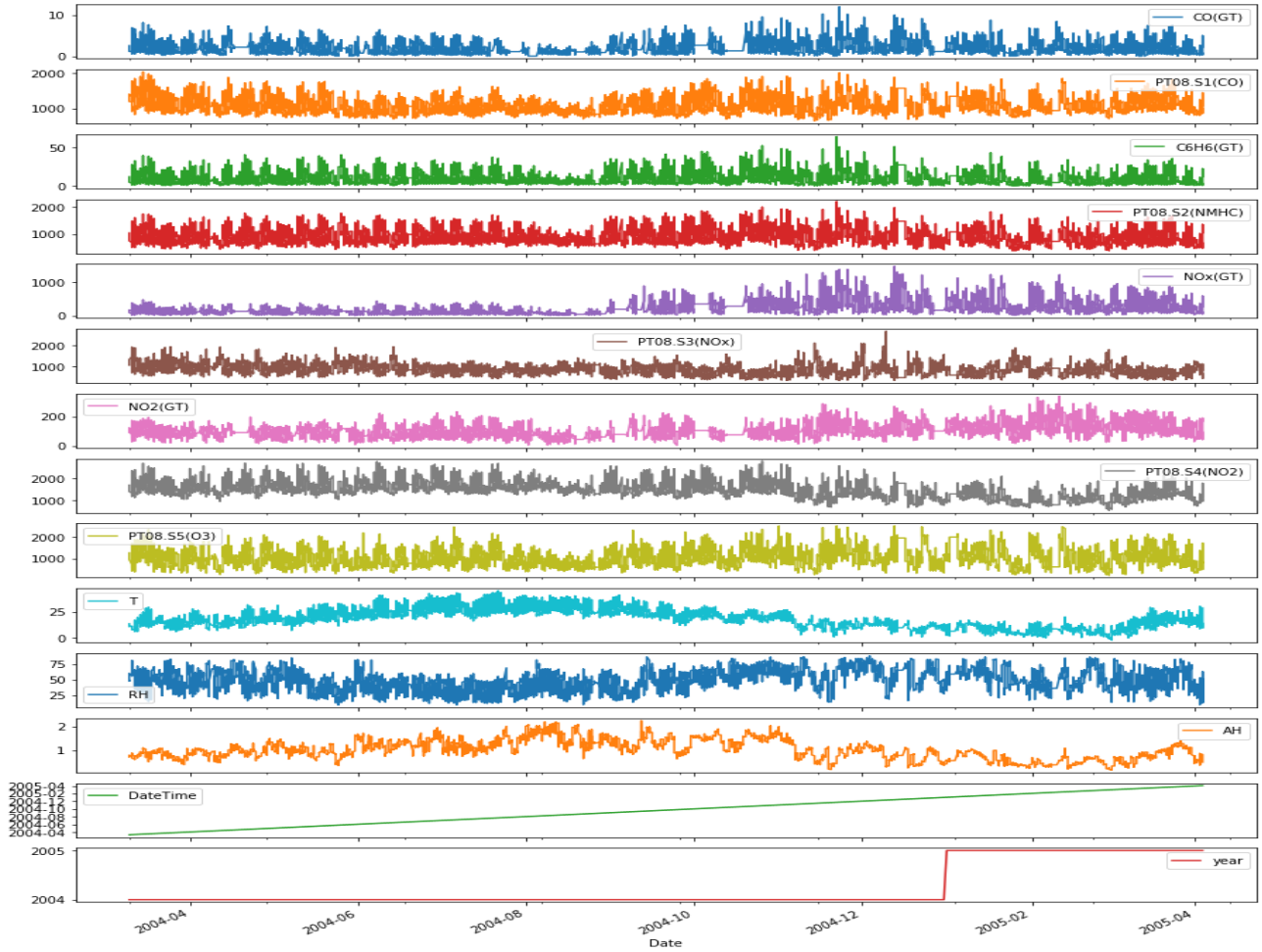
The dataset we used is Air Quality dataset from UCI machine Learning Repository. The dataset contains 9358 instances of hourly averaged responses spreading from March 2004 to February 2005. There are 15 features which contributes to 5 metal oxide chemical sensor readings. The dataset consists of following features: `Date, Time, CO(GT), PT08.S1(CO), NMHC(GT),C6H6(GT), PT08.S2(NHMC), NOx (GT), PT08.S3(NOx), NO2(GT), PT08.S4(NO2),PT08.S5(O3),T,RH, AH`. Here RH, AH are our dependent variables and remaining are our independent variables.

Preprocessing:

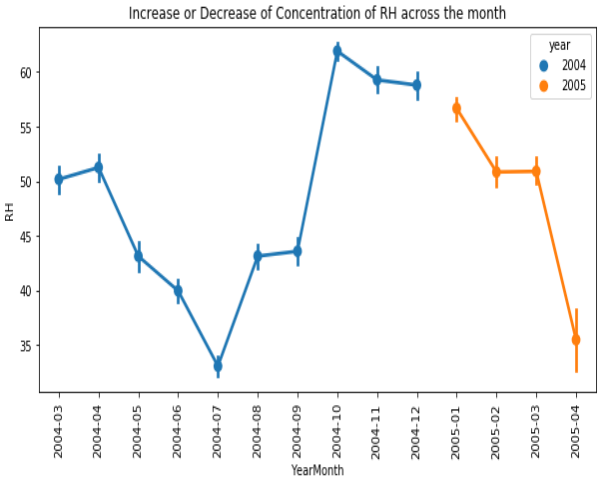
The dataset contains 9358 rows and there are some values that are tagged as -200 which means there is no data available for that values. I replaced -200 values with Nan and then filled those Nan with the mean values of that day. There are some Nan values left even after imputing with mean because there is no data available for whole day so I filled those Nan values with previous values. There is one feature NHMC which has 83% of data composed of Nan and imputing with mean values is not a good idea as there are more Nan values, So as the feature doesn't provide much information I removed that column.

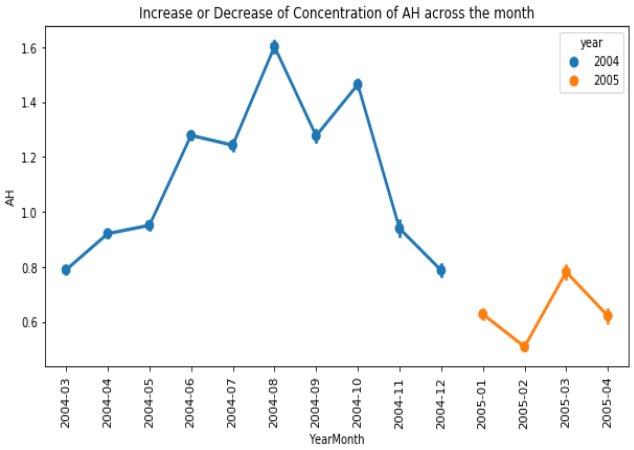
Data Analysis:

A brief look into the dataset shows that features increases or decreases with time and the there may seasonality in the data which may have impact on dependent variable. Below plots describes the variation of features across the time.



Now coming to the dependent variables the below plots describes the variations of RH and AH along the year and increment or decrement of concentration across the months.



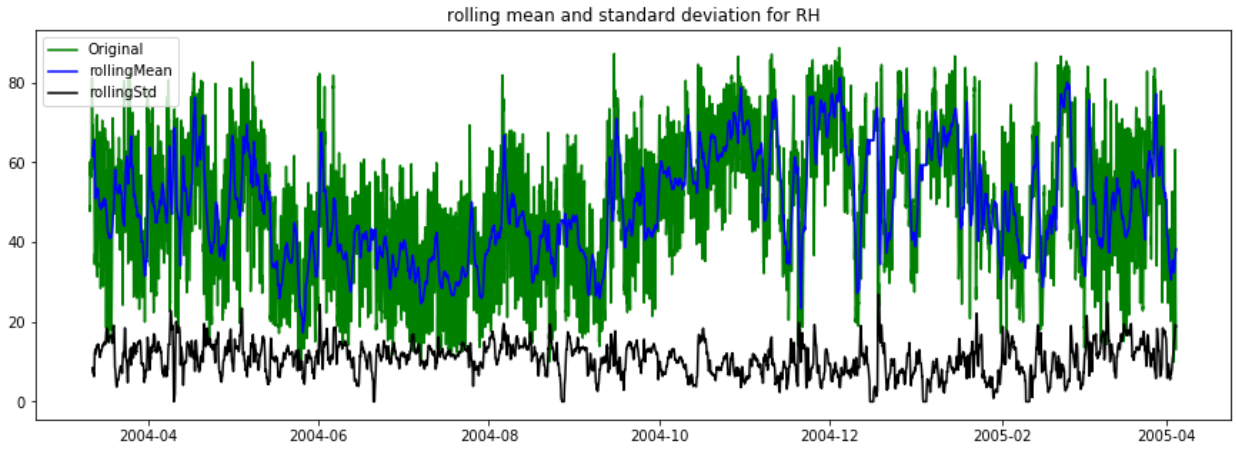


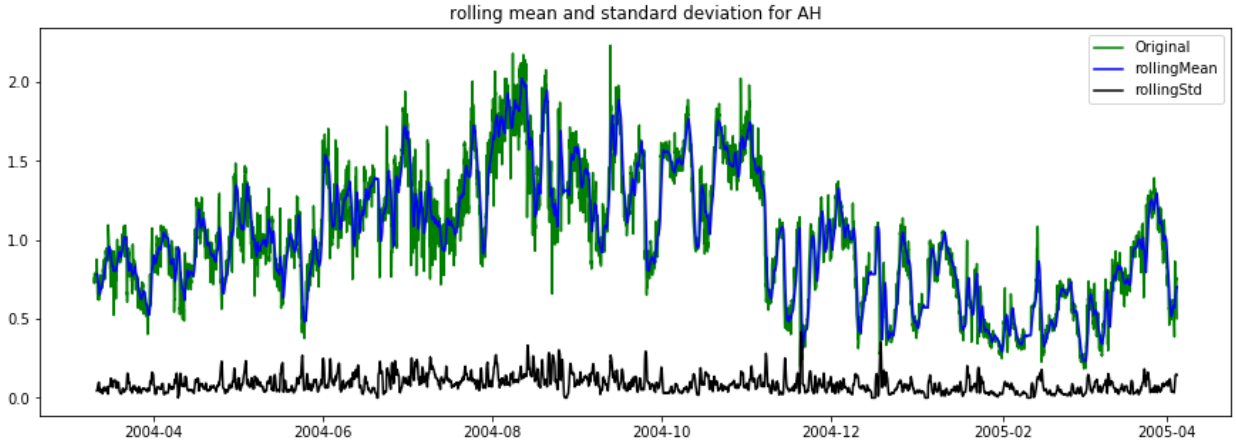
From the above plots we can that there is no incremental or decremental trend over time. There no obvious outliers present in the data. And the variability among the data is constant time. The Plots suggest that there is moderate correlation among the variables.

Check for stationarity:

Many Time series models assume that data is stationary and there are several methods to check stationarity. A time series is stationary if has constant mean, constant variance over time and auto correlation does not depend on time. Stationarity can be tested by Plotting Rolling statistics or Dickey – Fuller Test.

Plotting Rolling Statistics:





From the plots we can see that there is no gradual increment or decrement in data with time which confirms that data is stationary, the rolling mean and standard deviation is calculated for window period of 24 hours. But we can't confirm stationarity by only checking the plot we have to do Dickey-fuller Test.

Augmented Dickey-Fuller Test:

The Dickey Fuller test is one of the most popular statistical tests. It can be used to determine the presence of unit root in the series, and hence help us understand if the series is stationary or not. The null and alternate hypothesis of this test are:

Null Hypothesis: The series has a unit root (value of a =1)
Alternate Hypothesis: The series has no unit root.

If we fail to reject the null hypothesis, we can say that the series is non-stationary. This means that the series can be linear or difference stationary

```
Results of Dickey-Fuller Test(RH):
Test Statistic                -7.281607e+00
p-value                        1.495339e-10
#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
```

```
Results of Dickey-Fuller Test(AH):
Test Statistic                   -5.141627
p-value                           0.000012
#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
```

From the above results the test statistic < critical value, which implies that the series is stationary. This confirms our original observation which we initially saw in the plotting rolling statistics test.

Test of Assumptions:

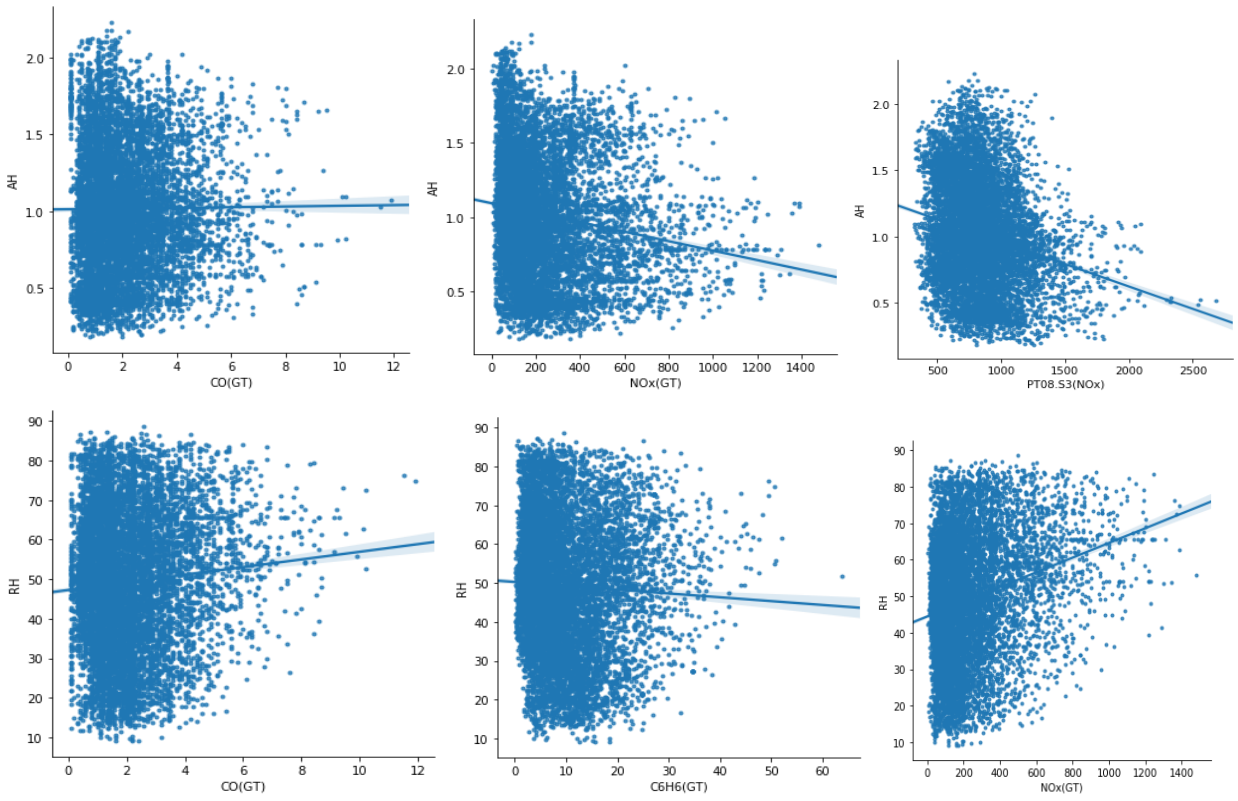
There are four principal assumptions which justify the use of linear regression models for purposes of inference or prediction:

1) linearity and additivity
2) homoscedasticity
3) multi collinearity
4) normality of error distribution

The dataset was split into training 70% and testing 30% and the assumptions of regressions are tested.

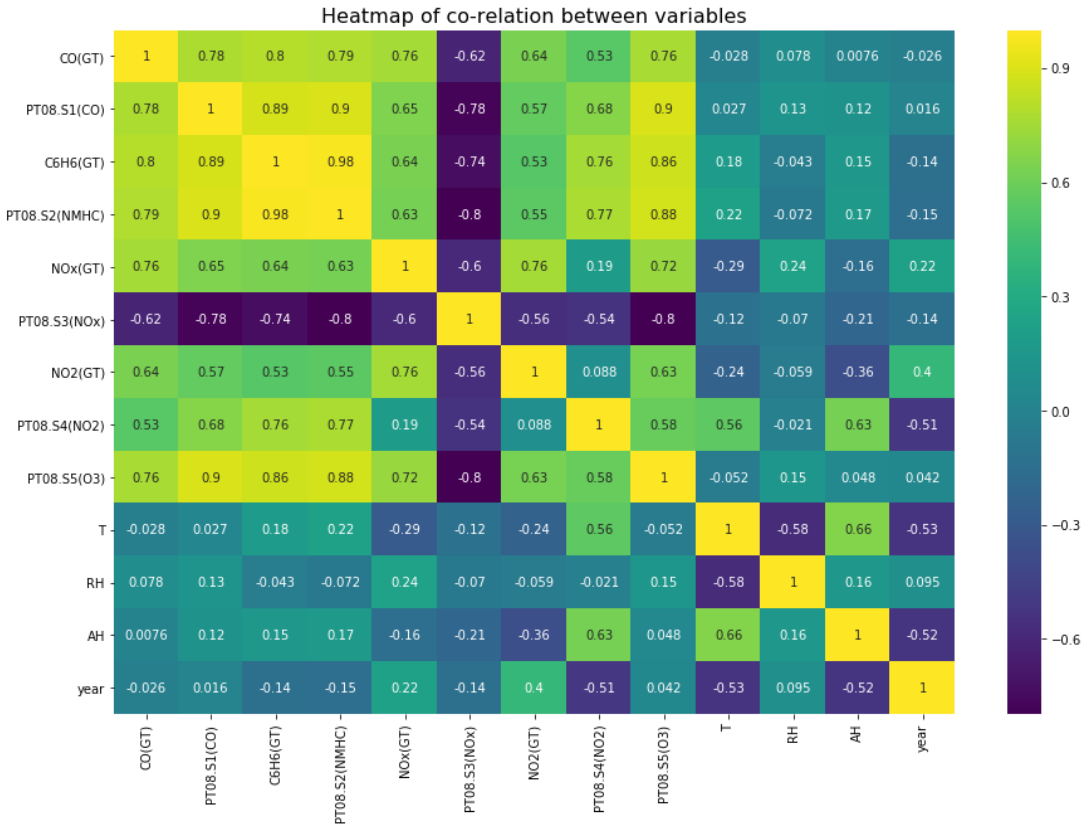
Linearity and additivity:

linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since linear regression is sensitive to outlier effects. The linearity assumption can be tested with scatter plots for some independent and dependent variables.



Multi Collinearity:

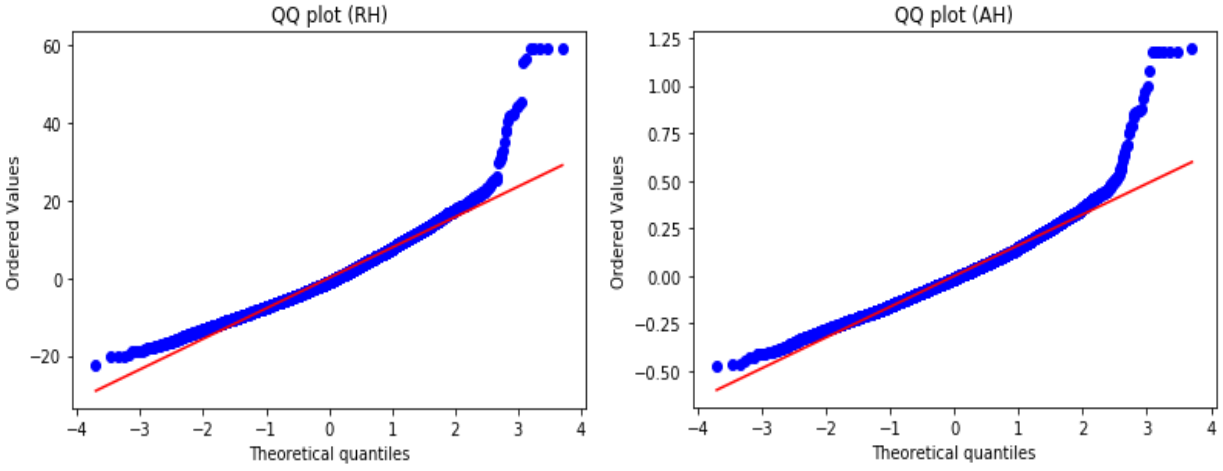
Multi collinearity can tested by using covariance matrix. when computing the matrix of Pearson's Bivariate Correlation among all independent variables the magnitude of correlation coefficients need to be smaller than 0.8. And the correlation between two dependent variables RH and AH shows that there is 0.16 correlation among each other.



The above covariance heat map shows that there is covariance between some independent variables. Since the data is time series it is normal that there will be high covariance among the independent variables.

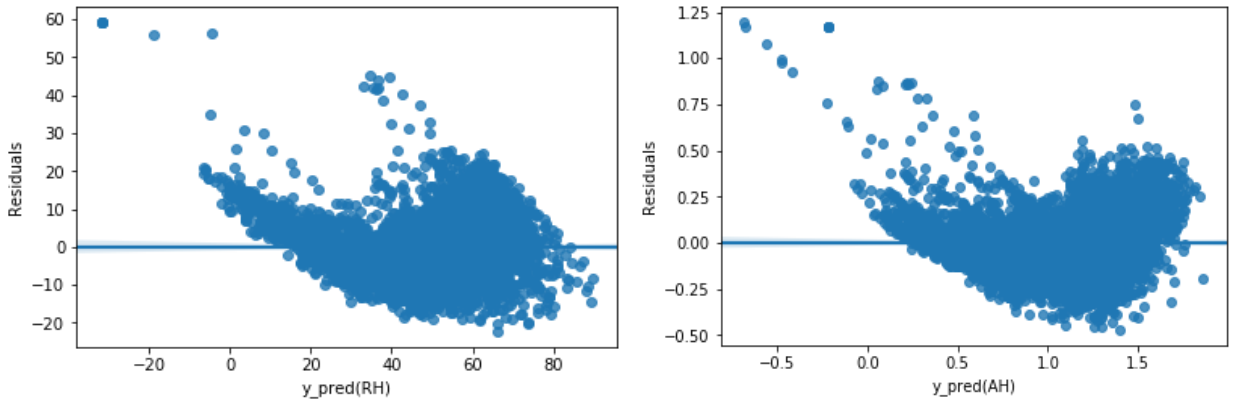
Normality:

The linear regression analysis needs all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot. When the data is not normally distributed a non-linear might fix this issue. The below plots show that it is almost linear thus all variables are multivariate normal.



Homoscedasticity:

The last assumption of the linear regression analysis is homoscedasticty. The residual plot is good way to check whether the data are homoscedastic,meaning that the residuals are equal across the regression line.

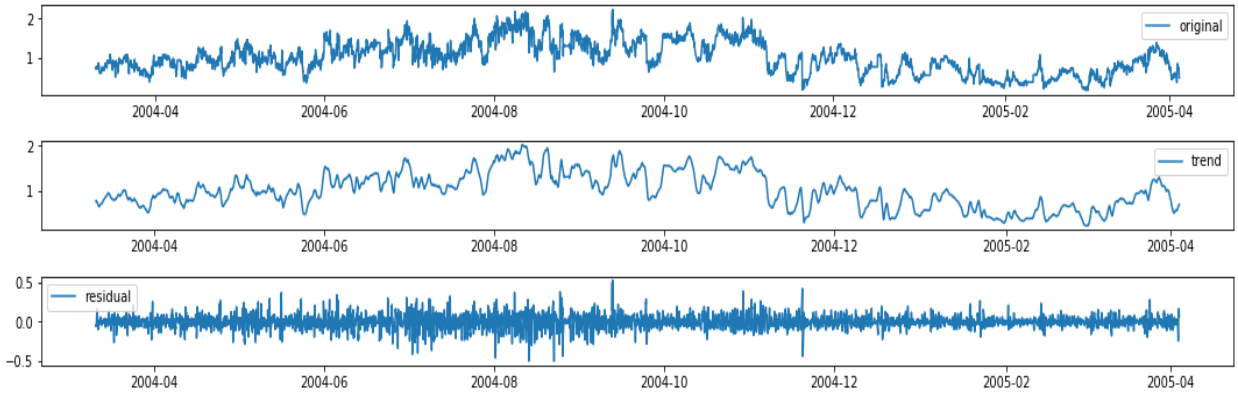


From above plots we can say that there is funnel shape structure forming in the plots which suggests that it is heteroscedastic. We can use Box-cox method as a remedy for heteroscedasity.

Fitting a Regression Model:

Though we can use linear regression model for predicting we are using ARIMA model as we are more interested in predicting future and ARIMA model does a good job in capturing the previous trend and thus helps in predicting future. ARIMA is a popular and widely used statistical method for time series forecasting.

Now checking the decomposition graph of AH we can get to know about trend in the data



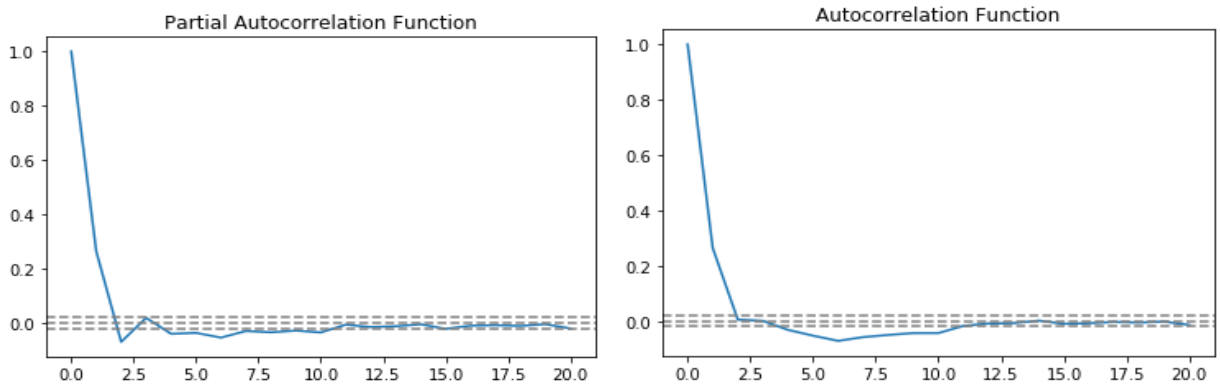
Here trend is random throught the year and nonlinear, there are no linear increments or decrements which shows that data is stationary and we can use ARIMA model.

ARIMA stands for Auto Regressive Integrated Moving Average. A standard notation is used of ARIMA(p, d, q) and The parameters of the ARIMA model are

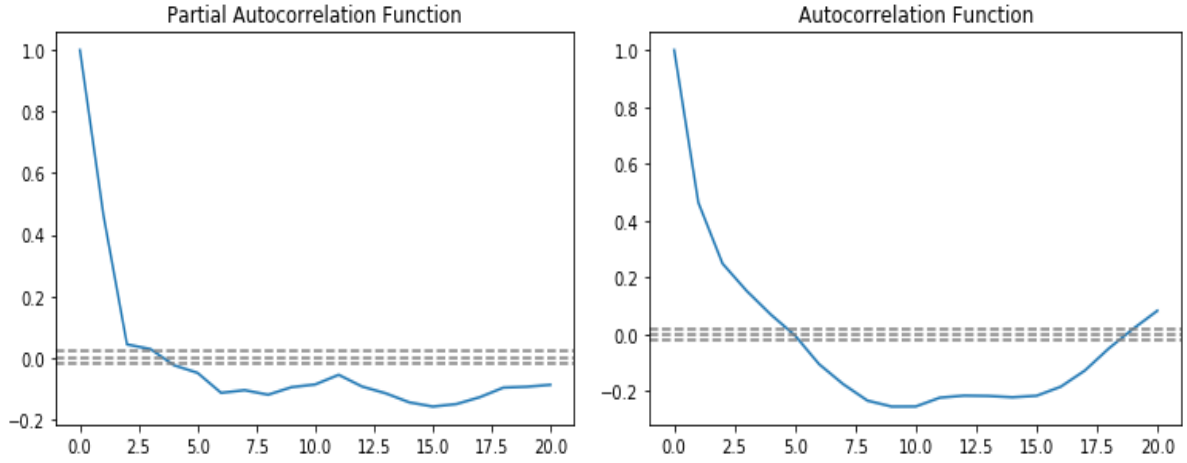
- **p**: The number of lag observations included in the model
- **d**: The number of times that the raw observations are differenced
- **q**: The size of the moving average window, also called the order of moving average.

To calculate p, q we need Auto Correlation plot and partial auto Correlation plot for both RH and AH variables.

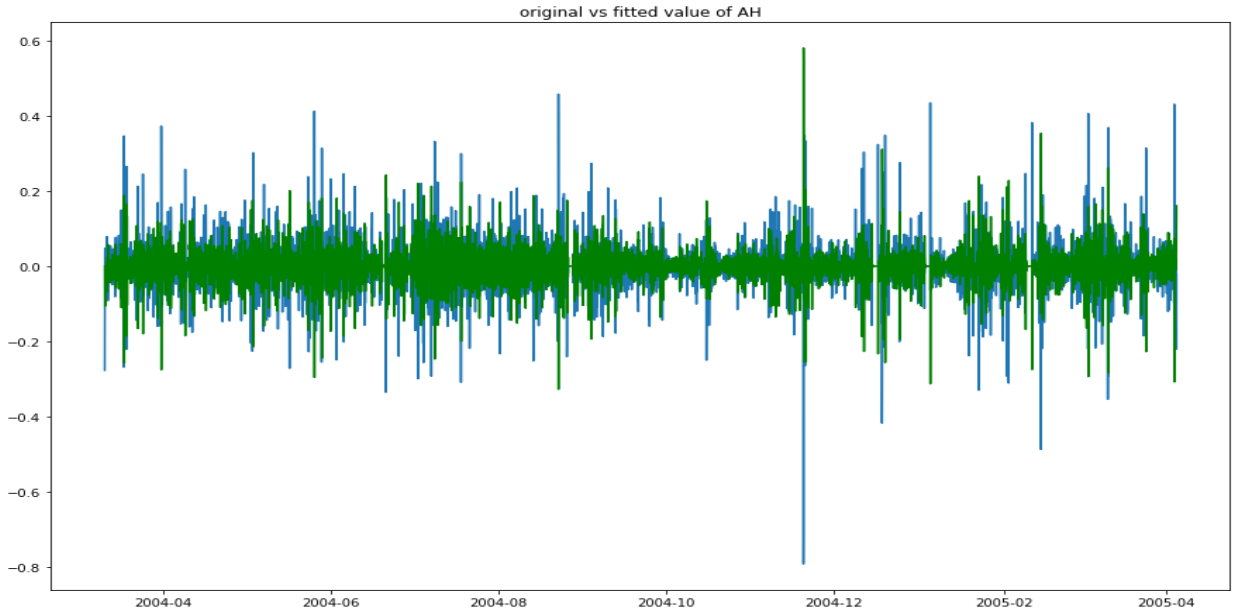
For AH variable:



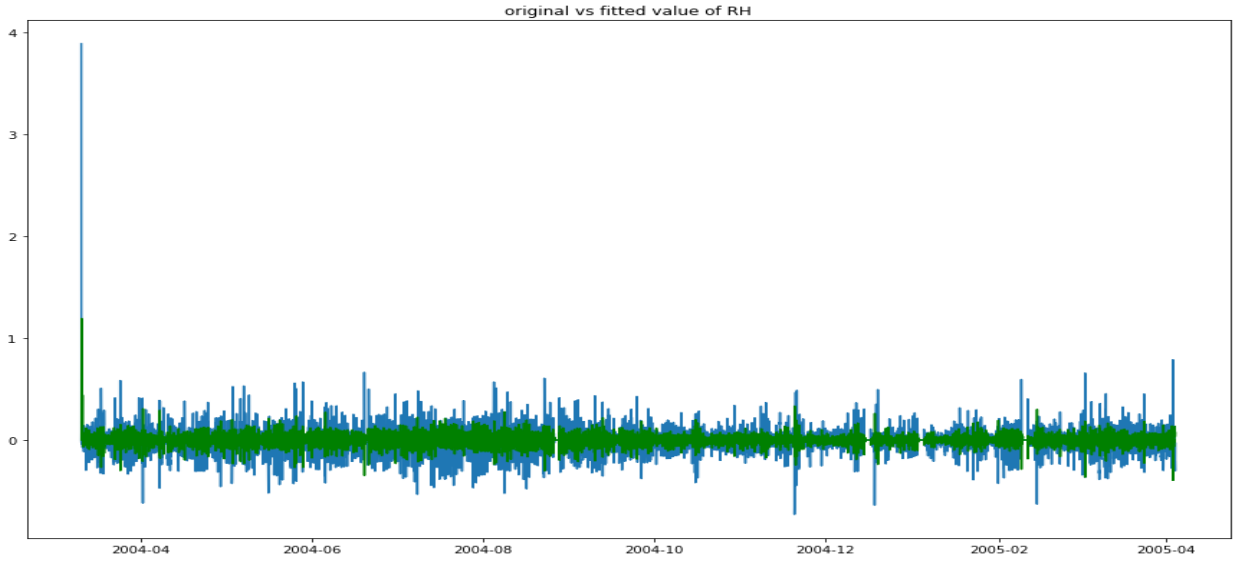
For RH variable.



From the plots we can get the values p= 2 and q = 4 for AH variable so the predicted vs original plot will be.



From the plots we can get the values p= 3 and q = 5 for AH variable so the predicted vs original plot will be.



Conclusion:

The Air quality data was cleaned and doesn't contain any null values. The data is stationary and ARIMA model was fitted for target variables RH and AH to give future predictions.