# Statistics for Data Science

## Study on Time Series Analysis

Sowmya Vasuki J
S20160010035

## Abstract:

Time Series is a sequence of well defined data points collected at particular time intervals over a period of time. Time Series Analysis is the use of statistical methods to analyse time series data and extract meaningful trends/characteristics from the data. Time Series Analysis helps us in understanding the underlying forces leading to the particular trend in the data and furthermore helps us forecast the data points by fitting appropriate models. The data is stationary as the test statistic(-7.3) is less than the critical value . The two dependant variables i.e., Relative humidity and Absolute humidity are predicted by ARIMA model.

## Problem Statement:

The goal is to perform 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. Visualization of the Time Series
3. Stationarity of the Series
4. Fitting a model to predict the dependant variables

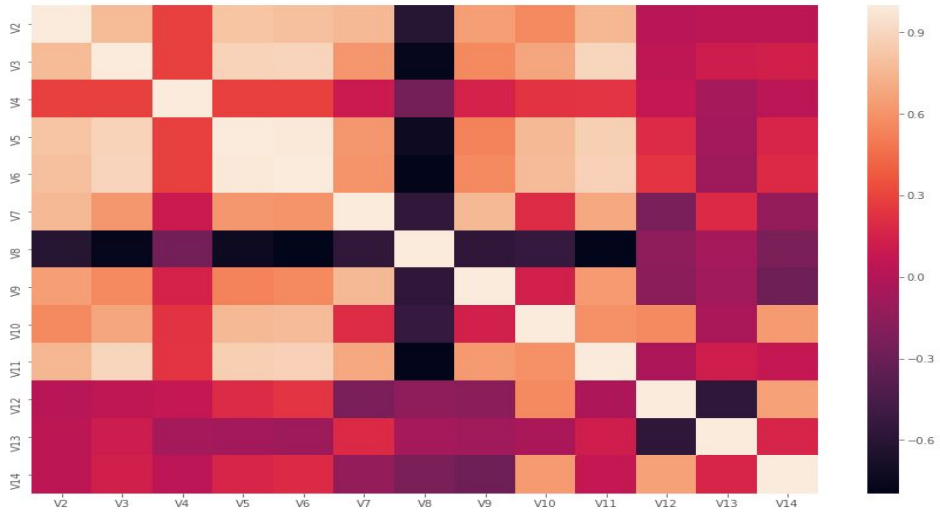
### Pre-Processing:

Before performing the analysis on the air quality data , the missing values in the data are removed. In the data the missing values are tagged with -200 value.These are replaced using the **.fillna()** function.

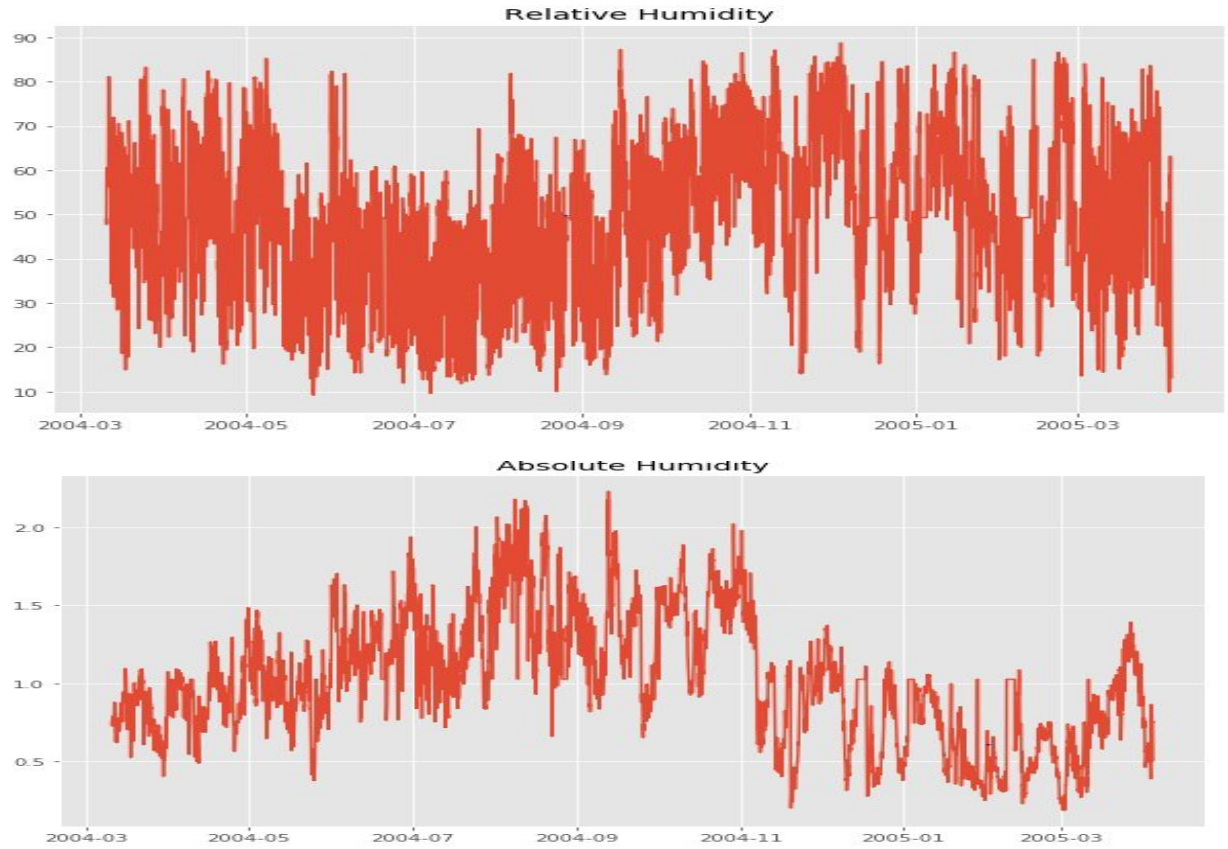
Date time indexing is set to the data.

### Visualization of the Time Series:

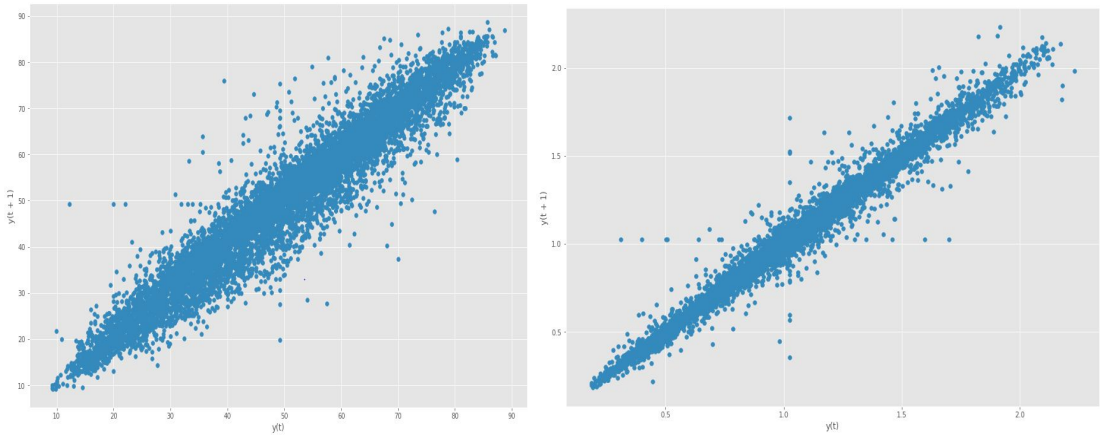
Before applying the statistical methods, we need to know the correlations between the variables.The following correlation heat map serves this purpose.



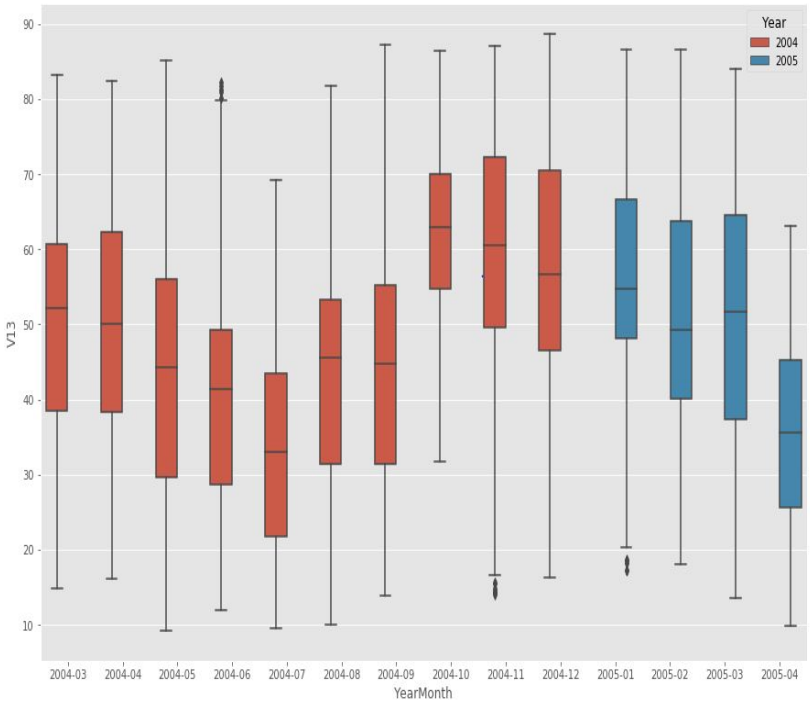

Plotting the Relative humidity and Absolute humidity (dependant variables)with datetime



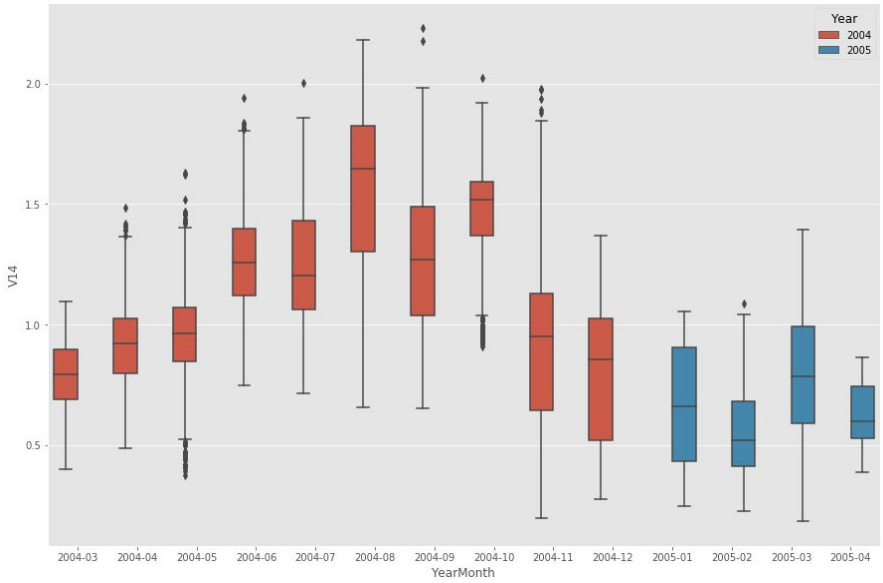
**Lag plot** is used to find whether the data is random.From the below plots it can be inferred that the data is not random.The lag plots for both the dependant variables are:




**Box plots:**




Box plot : Relative Humidity vs YearMonth




Box plot : Absolute Humidity vs YearMonth

## Stationarity of the series:

**Dickey-Fuller Test**: This is one of the statistical tests for checking stationarity. Here the null hypothesis is that the time series is non-stationary. The test results comprise of a Test Statistic and some Critical Values for difference confidence levels. If the 'Test Statistic' is less than the 'Critical Value', we can reject the null hypothesis and say that the series is stationary.

*Results of Dickey-Fuller Test:*(Relative Humidity)
Test Statistic -7.391164e+00
p-value 7.993979e-11
#Lags Used 3.800000e+01
Number of Observations Used 9.318000e+03
Critical Value (5%) -2.861850e+00
Critical Value (1%) -3.431052e+00
Critical Value (10%) -2.566935e+00

*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 (5%) -2.861850
Critical Value (1%) -3.431051

Critical Value (10%) -2.566935
dtype: float64

Here it can be seen that the the test statistic is less than the critical value(1%).
Hence we reject the null hypothesis and the series is stationary.

## Fitting the model:

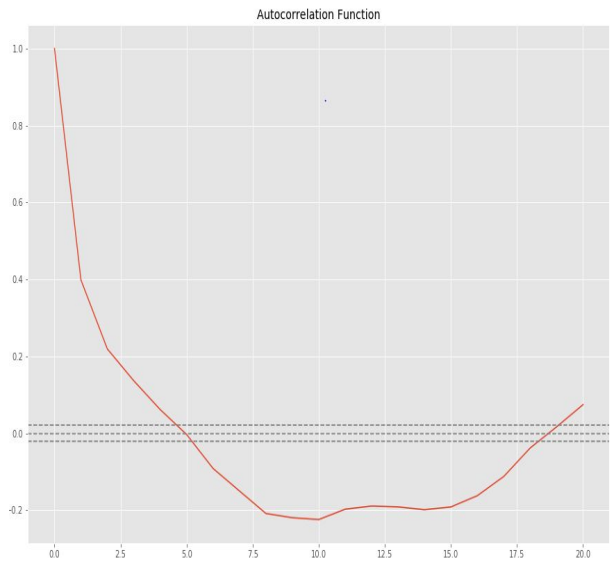
Before fitting the model, we use differencing on the response variables to fit the model. Differencing involves transforming the feature into logarithmic form, finding the shift and their difference.This difference is taken as our response variable and model if fit to find this difference.

The given data is not strictly stationary ( no interdependancy between variables) we use the ARIMA model to forecast the data.The ARIMA predictors depend on (p,d,q) where

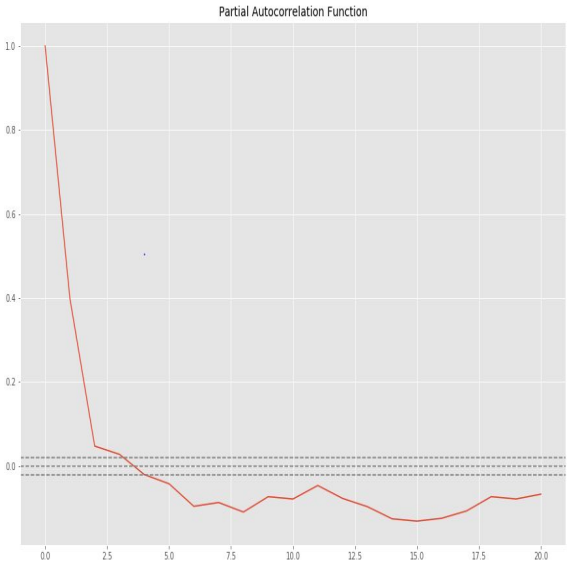
1. p is the number of AR(auto regressive) terms
2. q is the number of MA(moving average) terms
3. D is the number of differences

To find the values of p, q, we need two plots namely Autocorrelation Plot(ACF) and Partial Autocorrelation plot(PACF)
p – The lag value where the PACF chart crosses the upper confidence interval for the first time.
q – The lag value where the ACF chart crosses the upper confidence interval for the first time.

*Relative Humidity:*

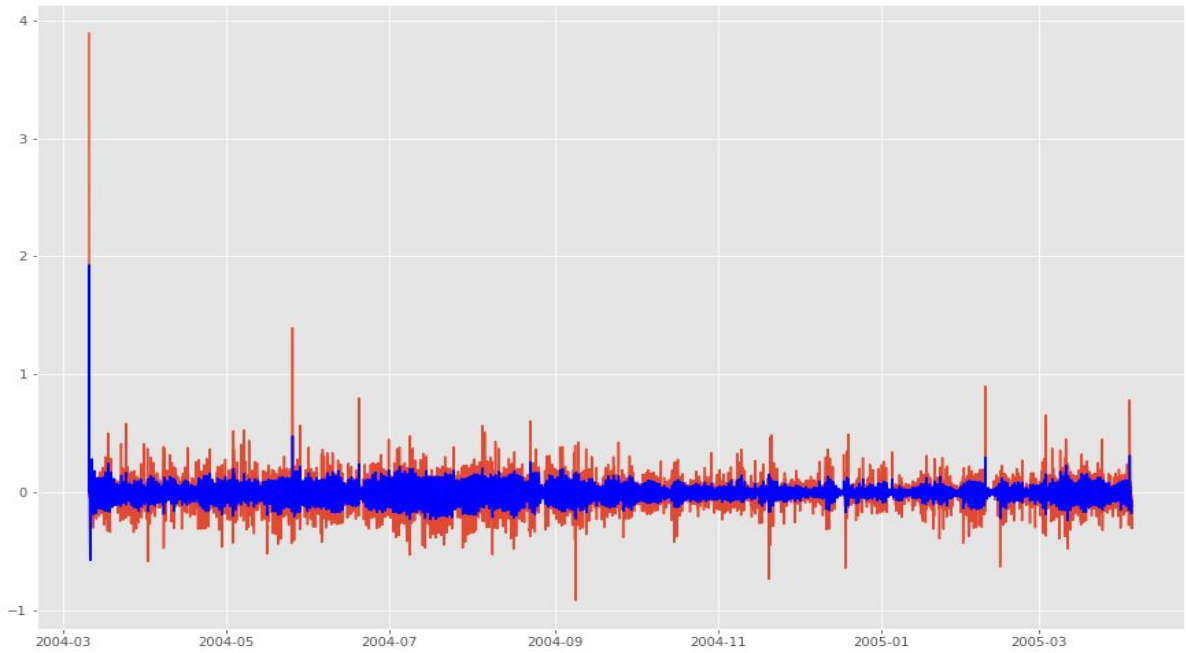



Autocorrelation Function



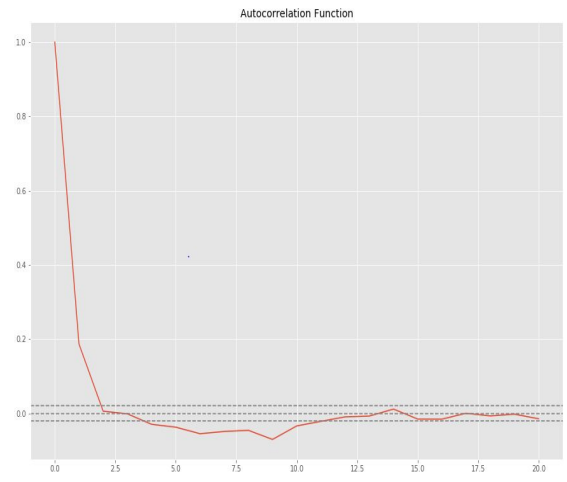

Partial Autocorrelation Function

From the above plots it can be inferred that p= 3, q=4. Using these values we fit the ARIMA model and predict the difference.The below plot shows the predicted values(Blue) to the original values (Red) of the difference.

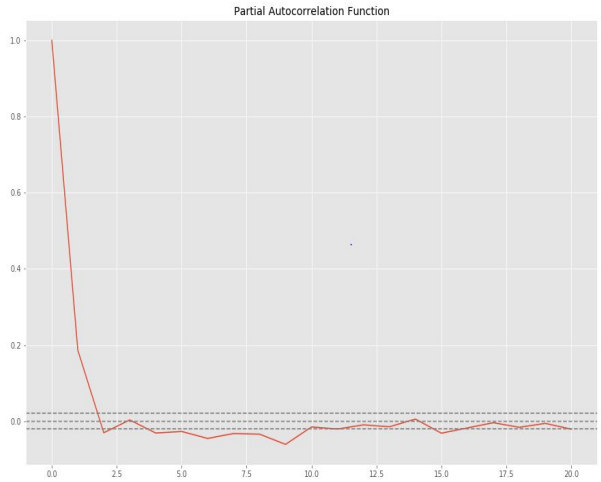



Relative humidity log Difference: Predicted vs Actual values

*Absolute Humidity:*

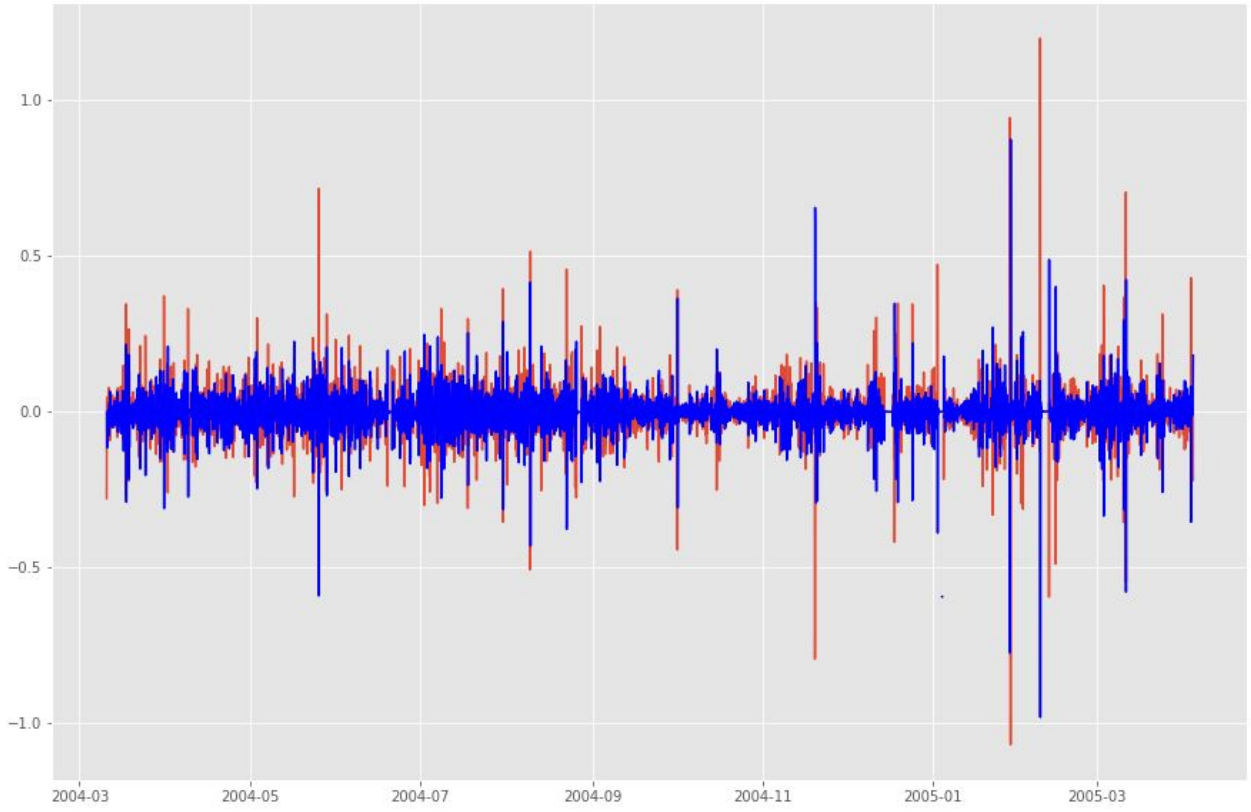



Autocorrelation Function




Partial Autocorrelation Function

From the above plots it can be inferred that p= 2, q=2. Using these values we fit the ARIMA model and predict the difference.The below plot shows the predicted values(Blue) to the original values (Red) of the difference.




Absolute humidity log Difference: Predicted vs Actual values

## Conclusion:

The Air quality data with 15 variables was removed of missing values and was analysed. It was a non strict stationary time series. ARIMA forecasting was used to fit a model and predict the dependant variables i.e., Relative Humidity and Absolute Humidity.