Data Analysis of Air Quality

Hinduja
hinduja.b16@iiits.in

Abstract:

A time series is a set of ordered observations on a quantitative characteristic of a phenomenon at equally spaced time points. One of the main goals of time series analysis is to forecast future values based on existing values.

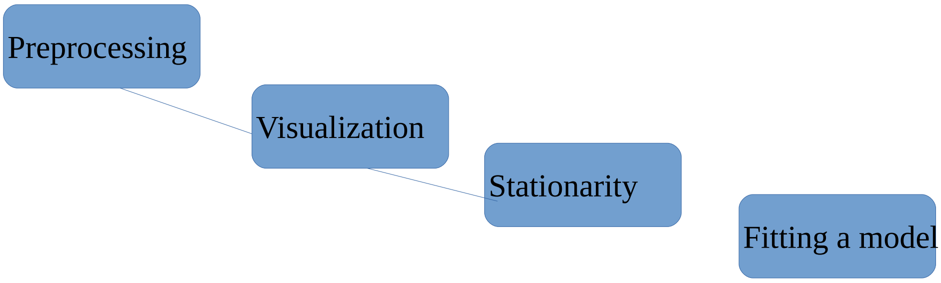
Dataset : AirQuality

Dependent variables: RelativeHumidity(RH) , AbsoluteHumidity(AH)

Problem Statement:

Our aim is to perform time series analysis on AirQuality dataset which consists of 15 features and 9358 instances and predict dependent variables (RH,AH) using ARIMA.

Methodology:



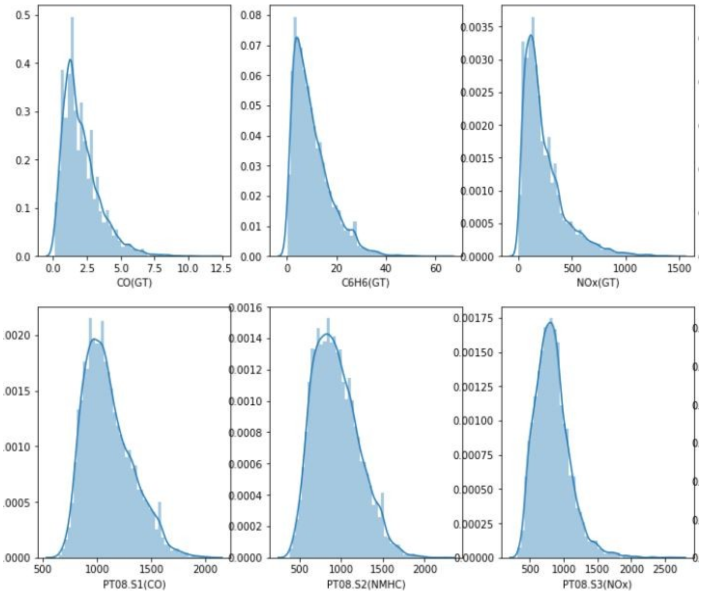
Preprocessing

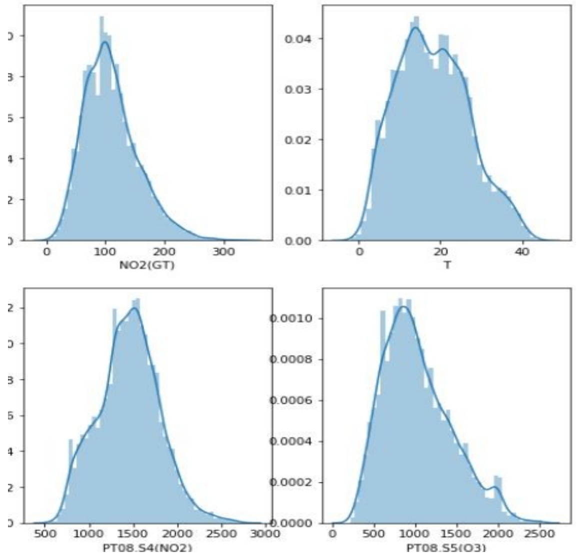
Visualization

Stationarity

Fitting a model

Plotting the variables:





Preprocessing:

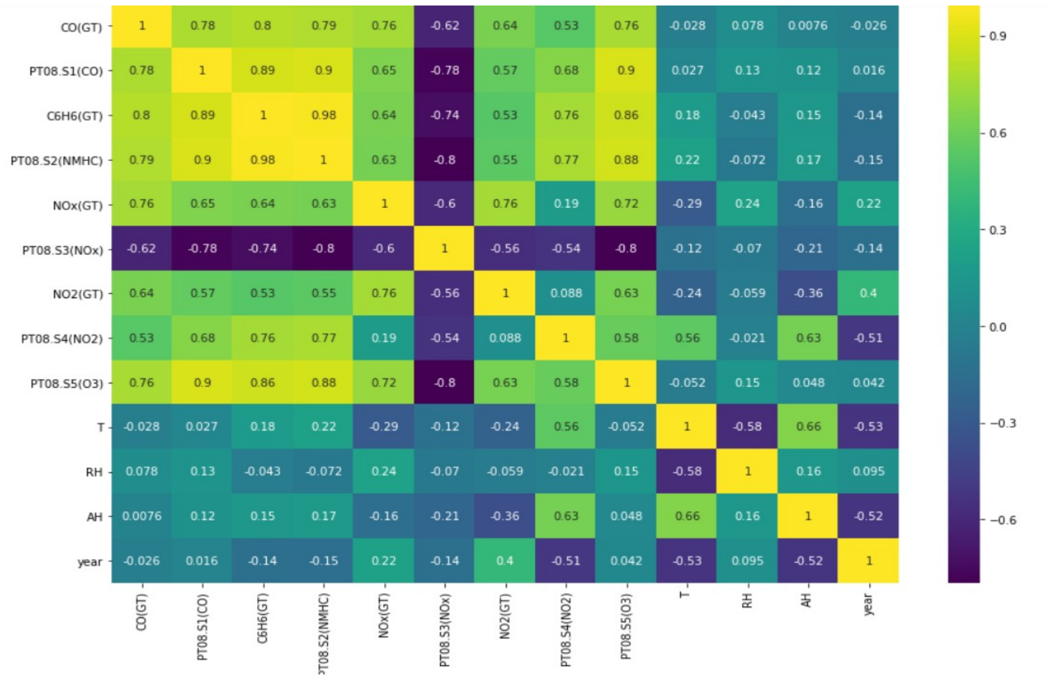
Prior to perform analysis on the data we need to remove null values. Null values are given by -200 in the dataset.There are still missing values because there is no data for whole day so the values are nan we can fill the values or delete the whole row here we are filling the null values with previous row.

Visualization:

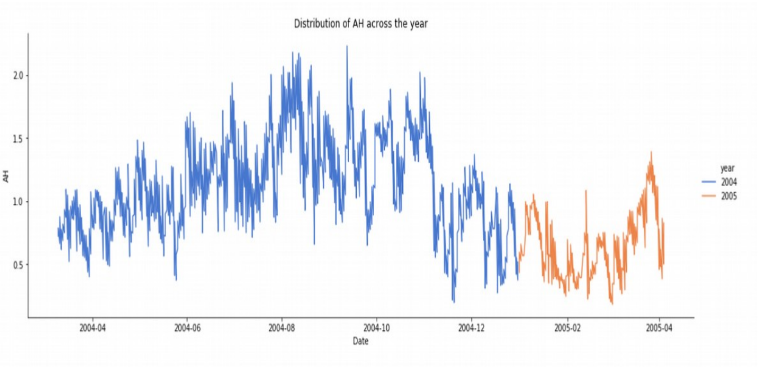
Correlation Matrix:

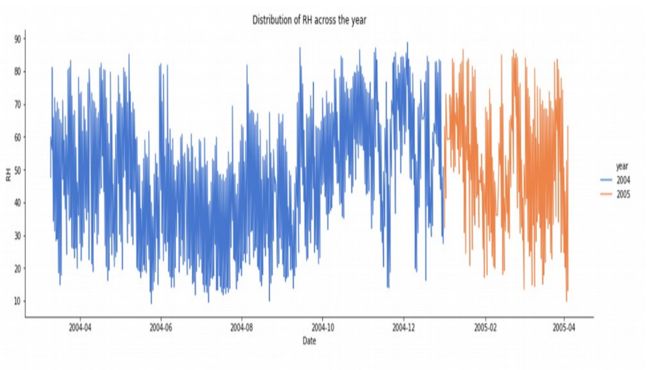
With the help of correlation matrix we can know the relation between any two variables.

Heat map of co-relation between variables



Distribution of dependent variables across the year

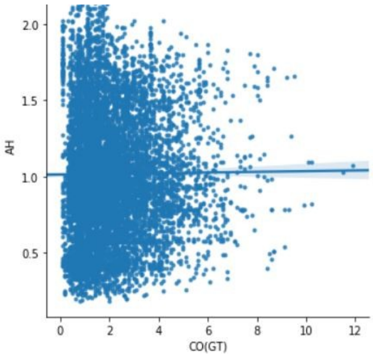


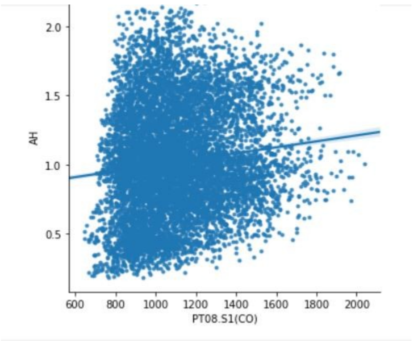


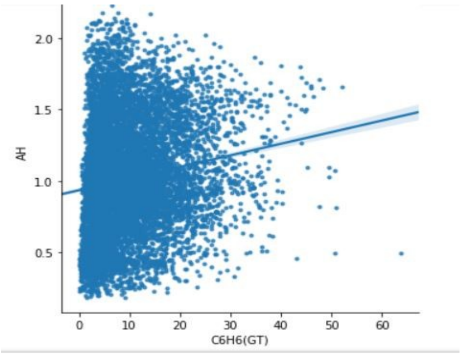
Assumptions of MLR :

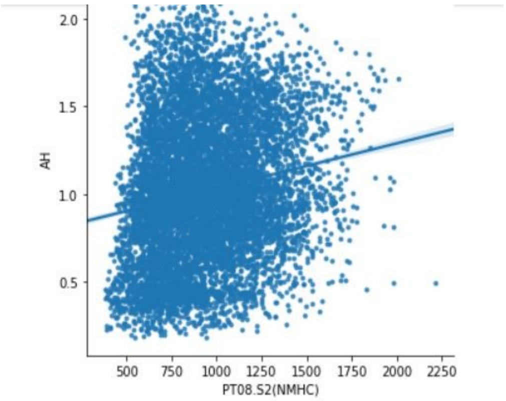
1)Linearity
2)Normality

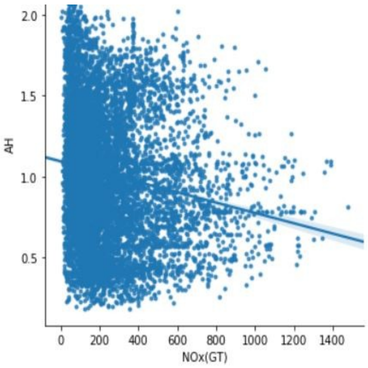
3)Homoskedasticity
4)No Multicollinearity

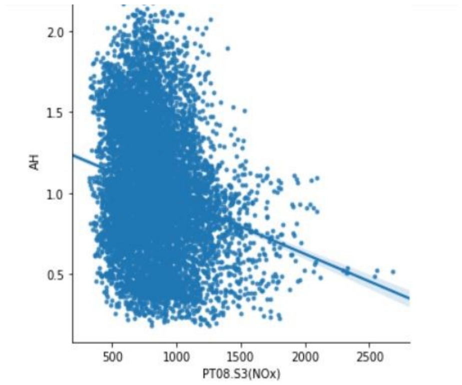


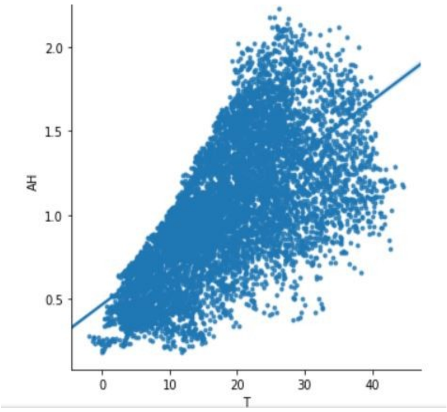


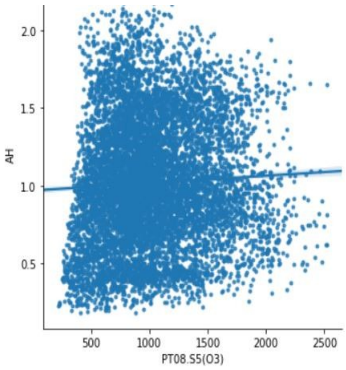


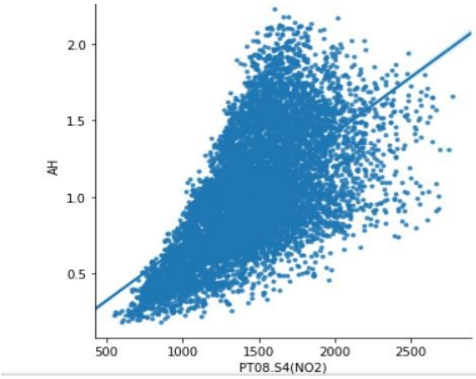


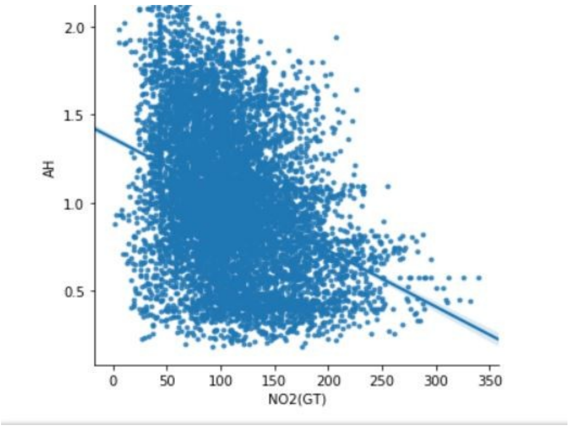


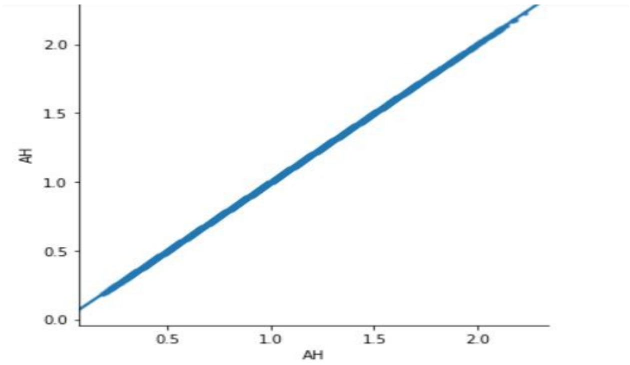






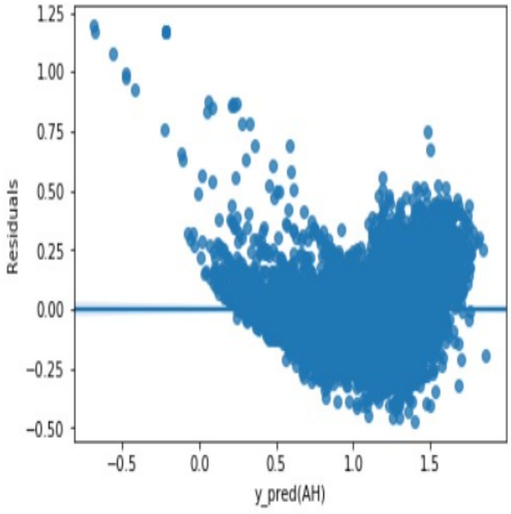




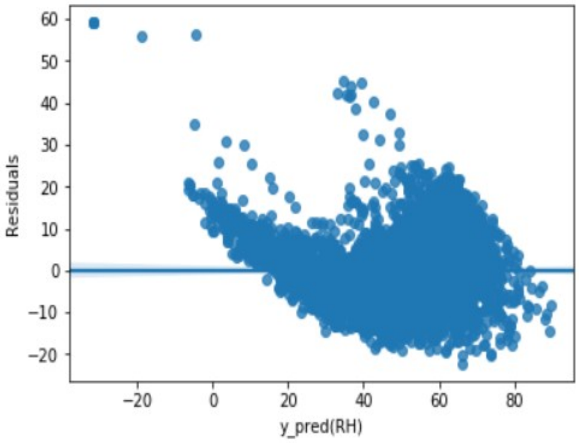


From these scatterplots we got the data as linear

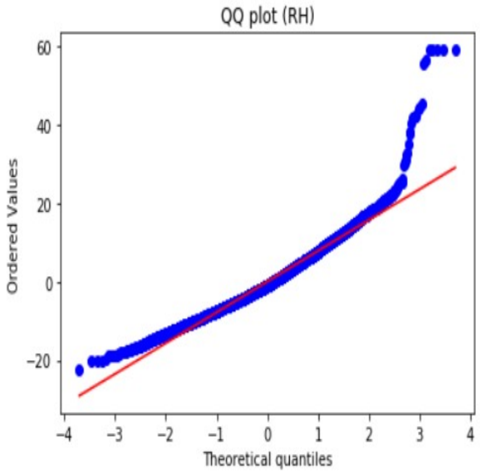
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     RH   R-squared:                       0.782
Model:                            OLS   Adj. R-squared:                  0.782
Method:                 Least Squares   F-statistic:                     2346.
Date:                Wed, 28 Nov 2018   Prob (F-statistic):               0.00
Time:                        14:41:48   Log-Likelihood:                -22932.
No. Observations:                6549   AIC:                         4.589e+04
Df Residuals:                    6538   BIC:                         4.596e+04
Df Model:                          10
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const            84.1017      2.319     36.265      0.000      79.556      88.648
CO(GT)           -1.4691      0.154     -9.555      0.000      -1.771      -1.168
PT08.S1(CO)       0.0056      0.001      4.190      0.000       0.003       0.008
C6H6(GT)         -0.5240      0.087     -6.006      0.000      -0.695      -0.353
PT08.S2(NMHC)    -0.0667      0.003    -23.774      0.000      -0.072      -0.061
NOx(GT)           0.0507      0.001     47.473      0.000       0.049       0.053
PT08.S3(NOx)     -0.0258      0.001    -32.719      0.000      -0.027      -0.024
NO2(GT)          -0.1291      0.004    -33.240      0.000      -0.137      -0.121
PT08.S4(NO2)      0.0585      0.001     78.358      0.000       0.057       0.060
PT08.S5(O3)       0.0014      0.001      2.057      0.040     6.7e-05       0.003
T                -1.7983      0.018    -98.246      0.000      -1.834      -1.762
==============================================================================
Omnibus:                     1356.184   Durbin-Watson:                   2.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4915.984
Skew:                           1.008   Prob(JB):                         0.00
Kurtosis:                       6.735   Cond. No.                     5.84e+04
                                ==============================================
```

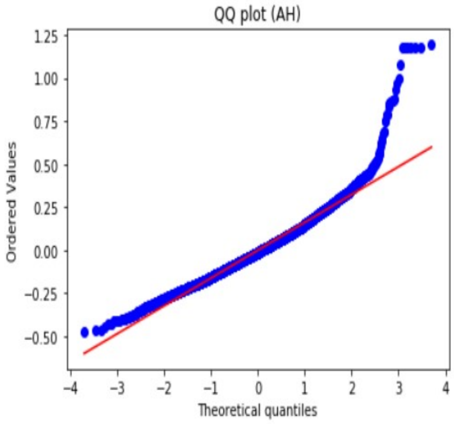


Residuals plot with dependent variables:

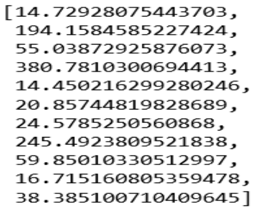


From the below Q-Q plot we can see that the data followed normality



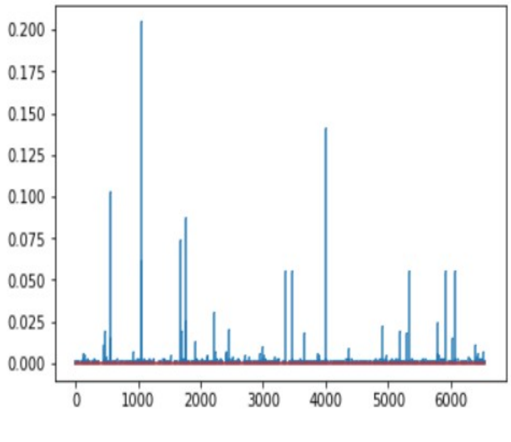


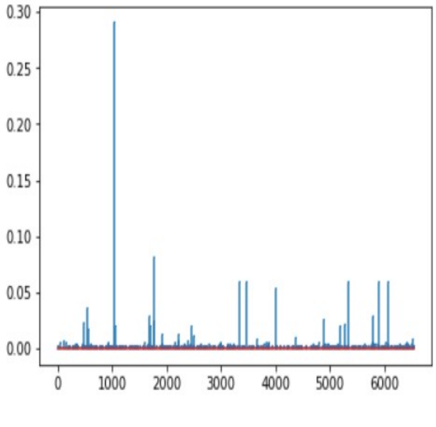
Checking Multicollinearity with VIF



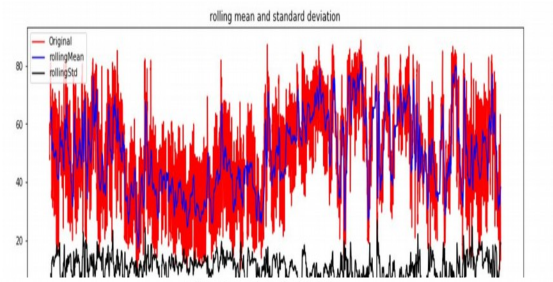
```
[14.72928075443703,
 194.1584585227424,
 55.03872925876073,
 380.7810300694413,
 14.450216299280246,
 20.85744819828689,
 24.5785250560868,
 245.4923809521838,
 59.85010330512997,
 16.715160805359478,
 38.385100710409645]
```

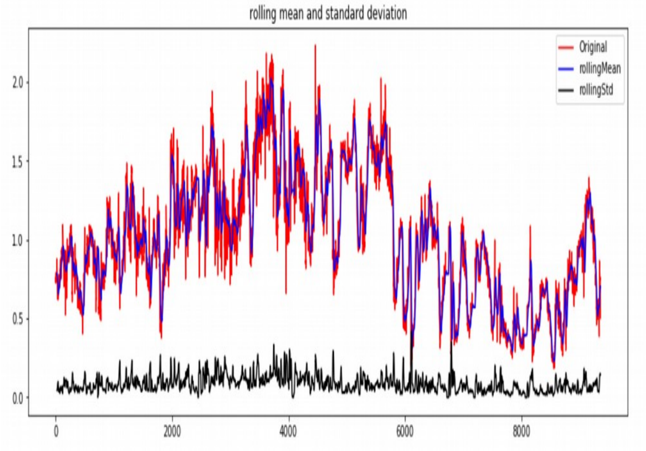
Checking for influential points

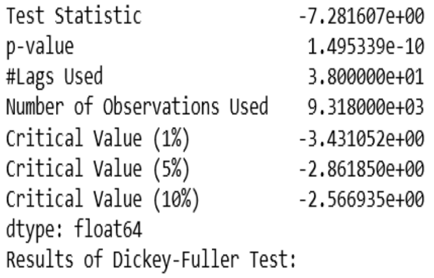




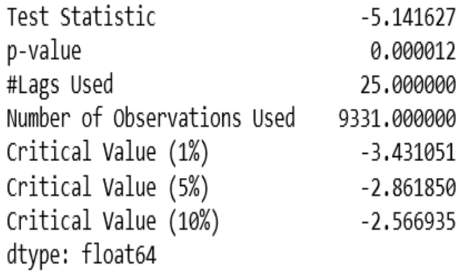
Since the cooks distance are less than one there are no influential points



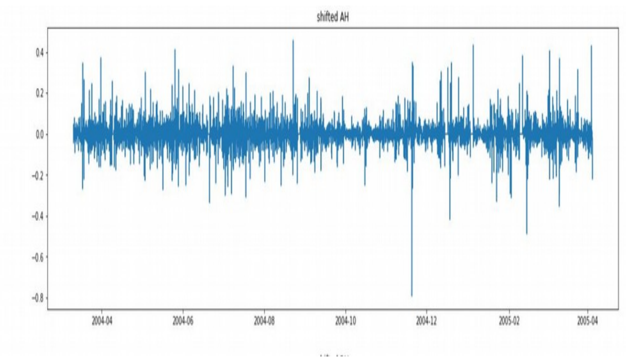


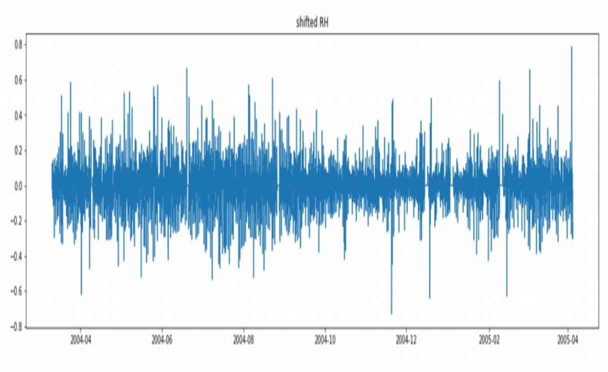


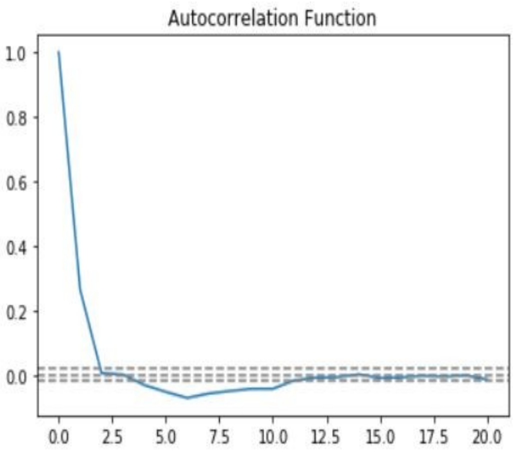
```
Test Statistic                  -7.281607e+00
p-value                          1.495339e-10
#Lags Used                       3.800000e+01
Number of Observations Used      9.318000e+03
Critical Value (1%)             -3.431052e+00
Critical Value (5%)             -2.861850e+00
Critical Value (10%)            -2.566935e+00
dtype: float64
Results of Dickey-Fuller Test:
```

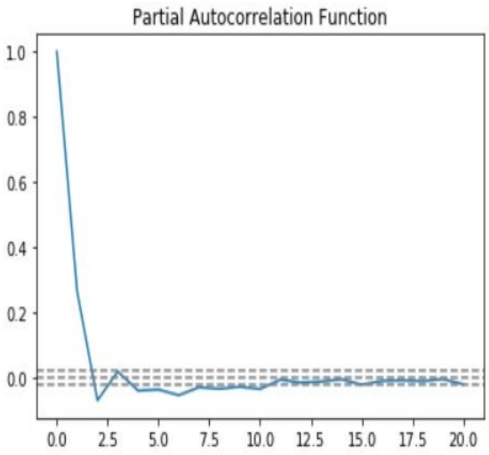


```
Test Statistic                   -5.141627
p-value                           0.000012
#Lags Used                       25.000000
Number of Observations Used    9331.000000
Critical Value (1%)              -3.431051
Critical Value (5%)              -2.861850
Critical Value (10%)             -2.566935
dtype: float64
```


shifted AH


shifted RH


Autocorrelation Function


Partial Autocorrelation Function