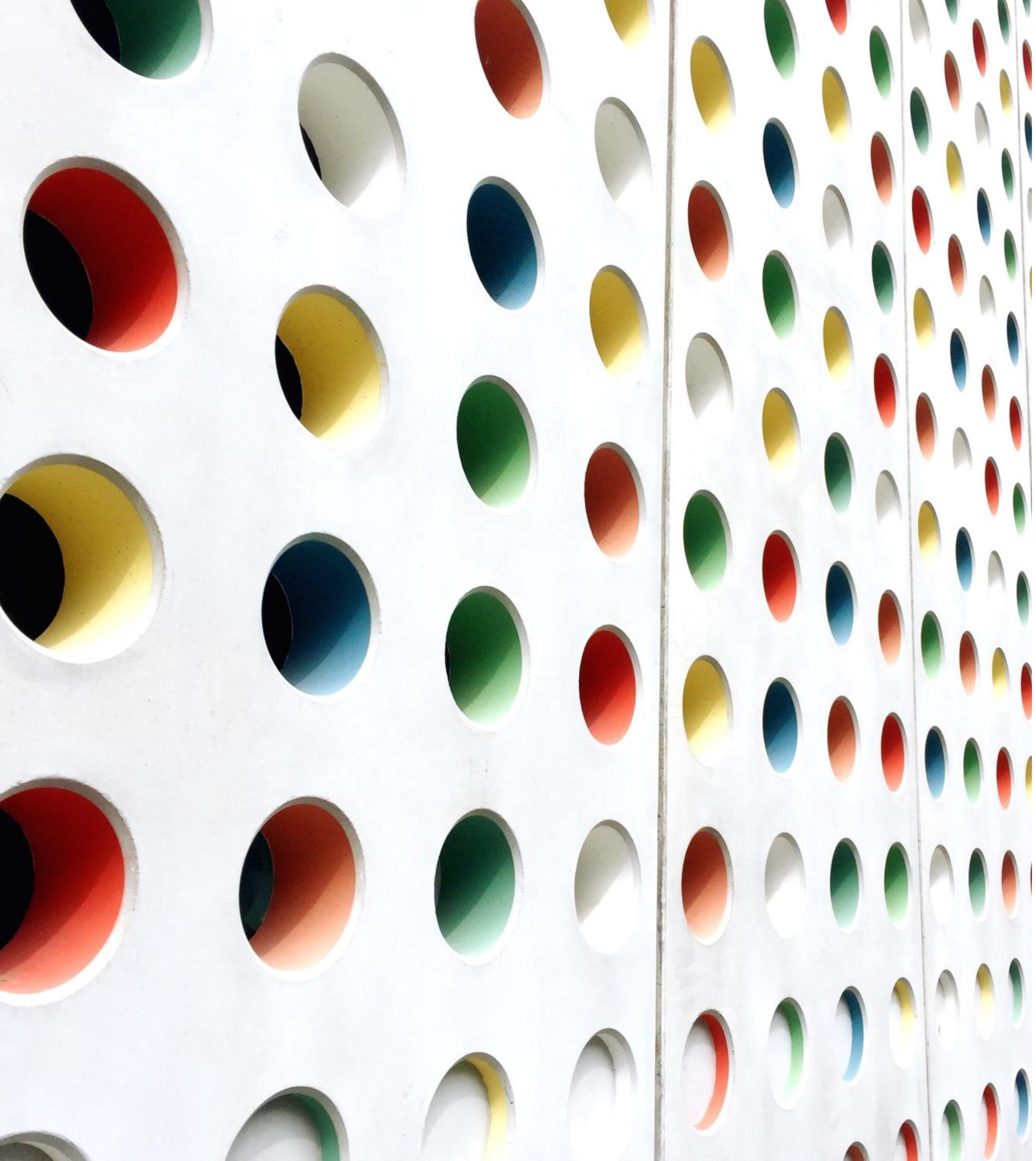


Air Quality - Forecast

SARA SANCHEZ

TIME SERIES ANALYSIS AND MODELING

Objective

To predict the Relative Humidity (RH) based on several features from a data set over the period March 2004 to February 2005.

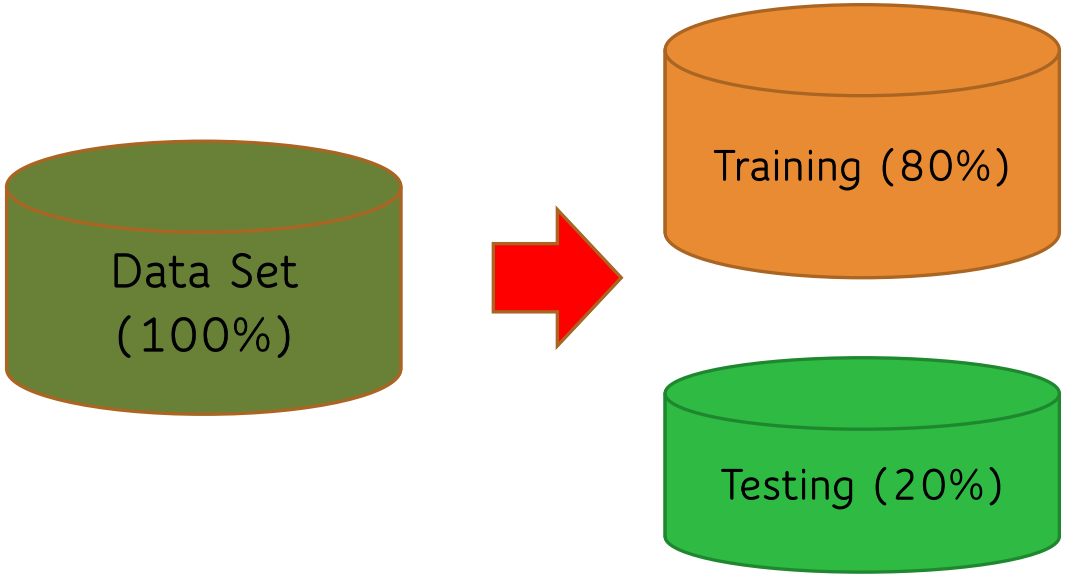
Data Set

- Repository: UCI Machine Learning
- 9,358 Instances
- Dependent variable: Relative Humidity (RH)

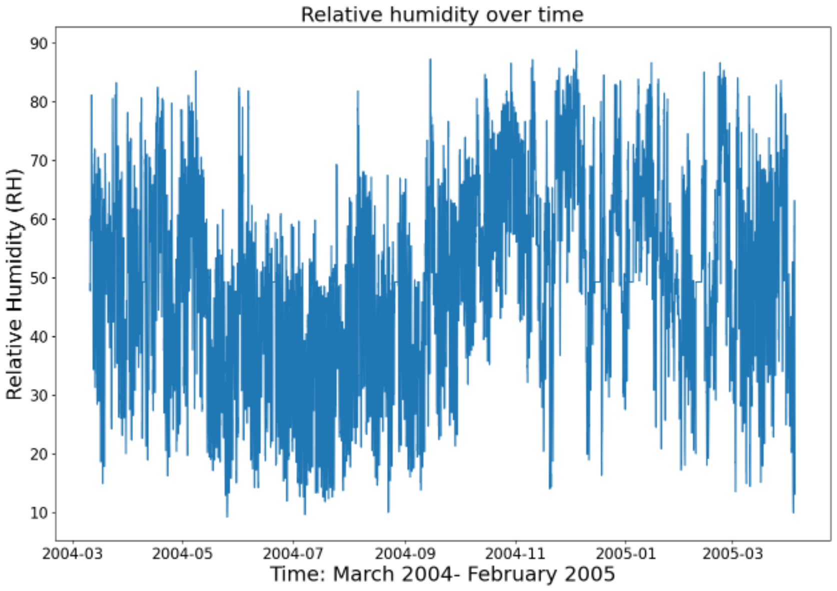
Data Preprocesing

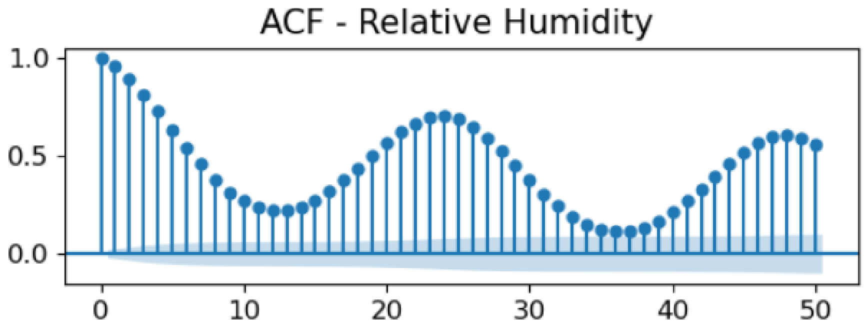
As parte of the data preprocesing:

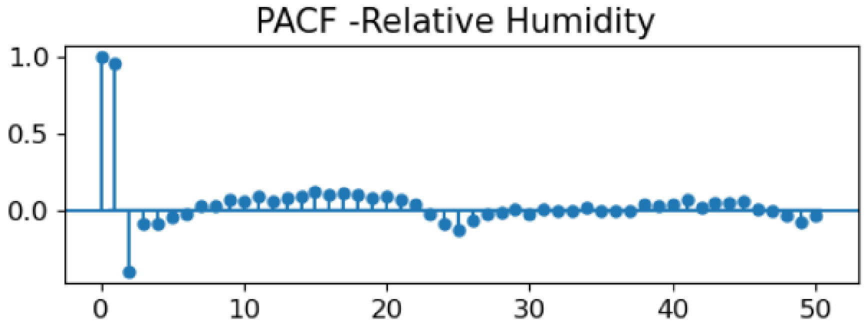
- Removing Unamed columns
- Manage the dates
- Changing some datatypes
- Handling Null values
- Handling NaN values



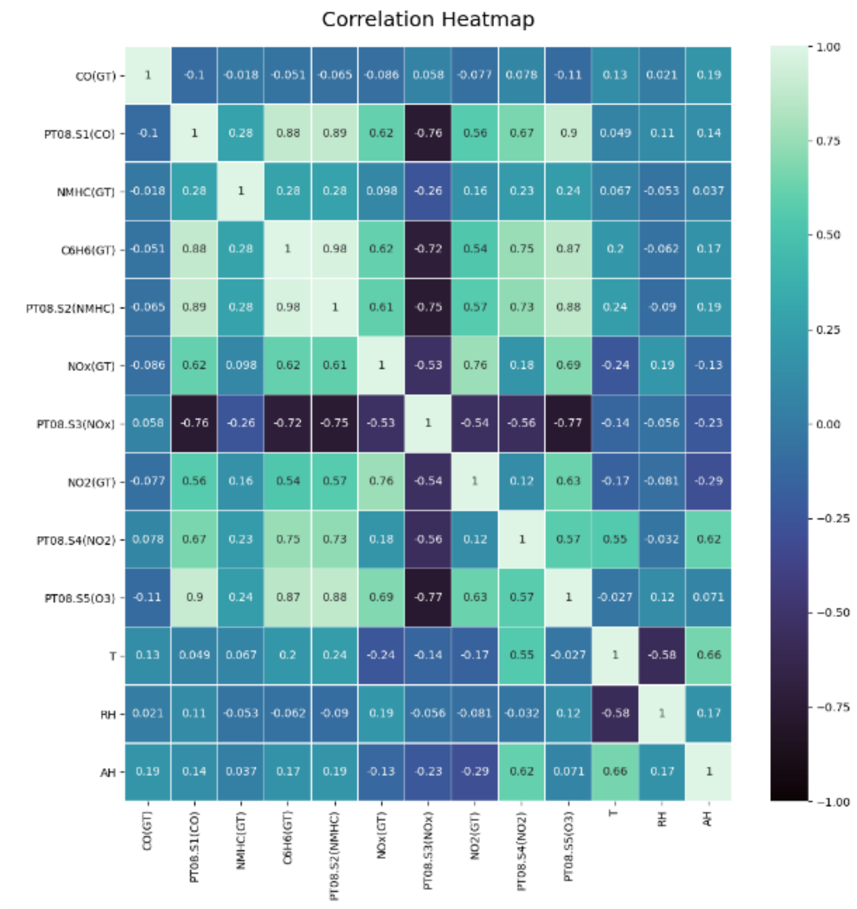
EDA


Relative humidity over time
Relative Humidity (RH)
90
80
70
60
50
40
30
20
10
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
Time: March 2004- February 2005

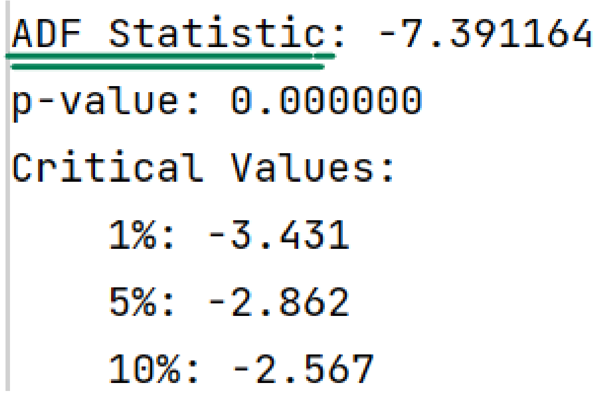

ACF - Relative Humidity
1.0
0.5
0.0
0
10
20
30
40
50


PACF -Relative Humidity
1.0
0.5
0.0
0
10
20
30
40
50

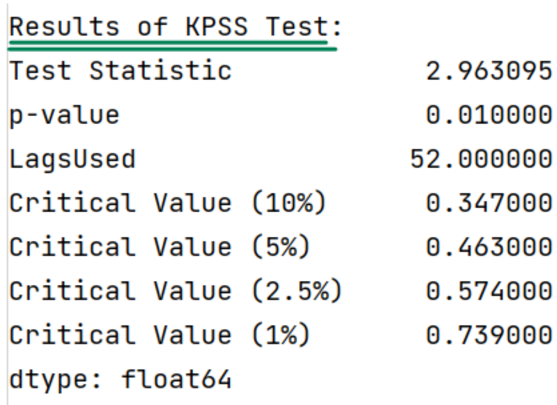
Correlation Heatmap



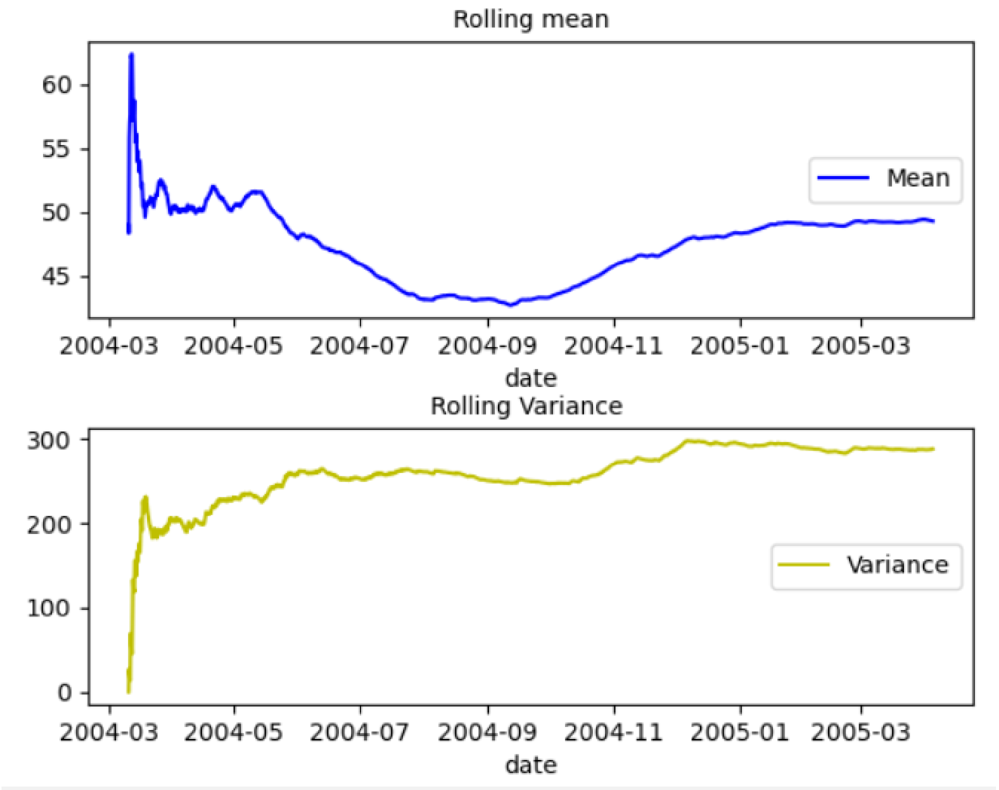
Checking - Stationality



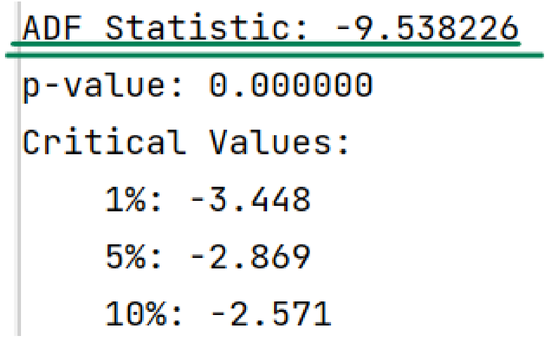
```
ADF Statistic: -7.391164
p-value: 0.000000
Critical Values:
	1%: -3.431
	5%: -2.862
	10%: -2.567
```



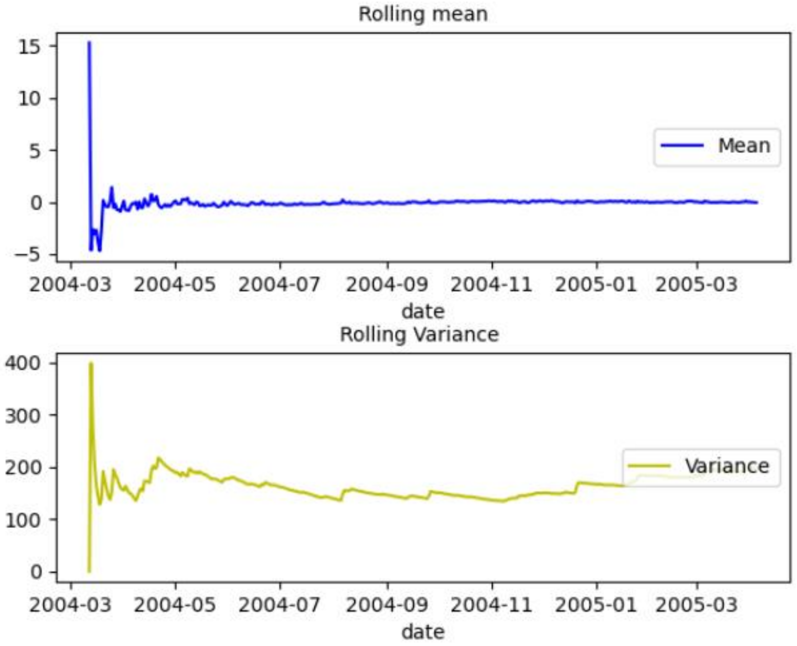
```
Results of KPSS Test:
Test Statistic            2.963095
p-value                   0.010000
LagsUsed                 52.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
dtype: float64
```

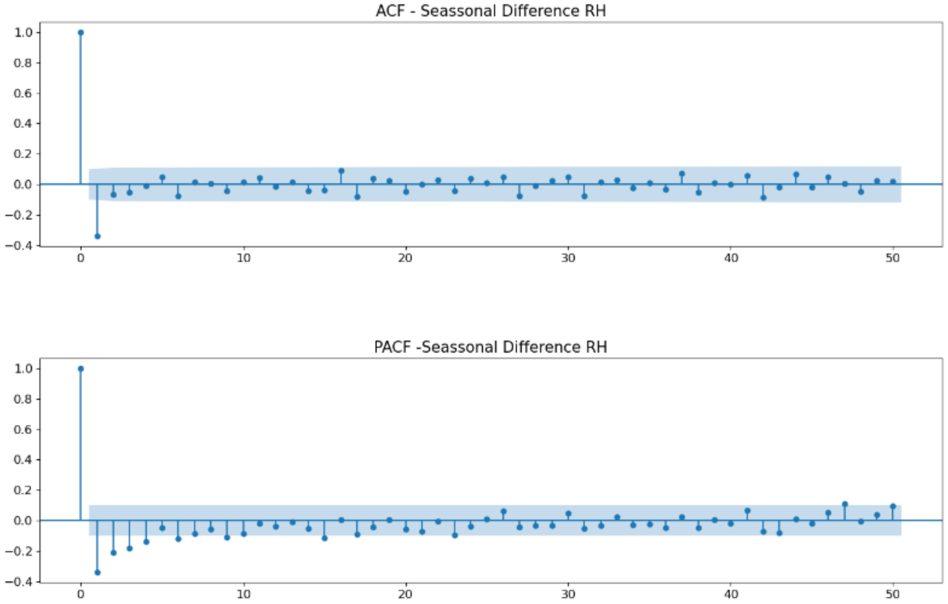


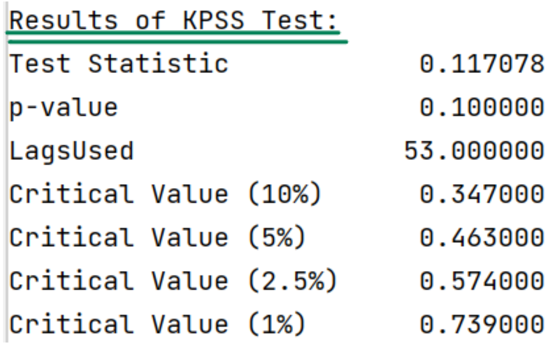
Seasonal difference – 24 periods



```
ADF Statistic: -9.538226
p-value: 0.000000
Critical Values:
	1%: -3.448
	5%: -2.869
	10%: -2.571
```

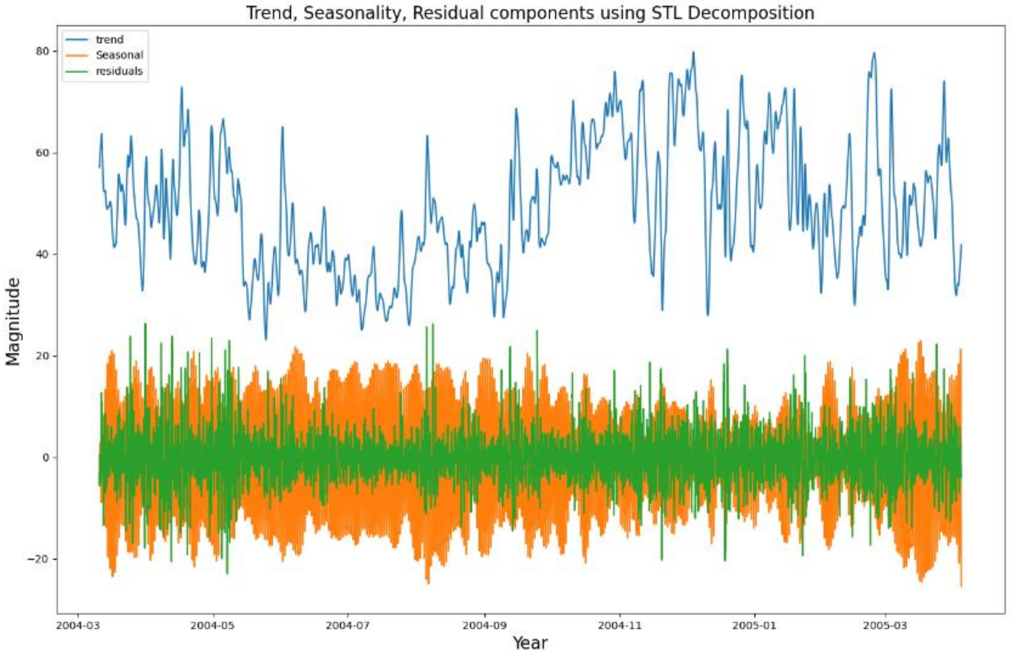


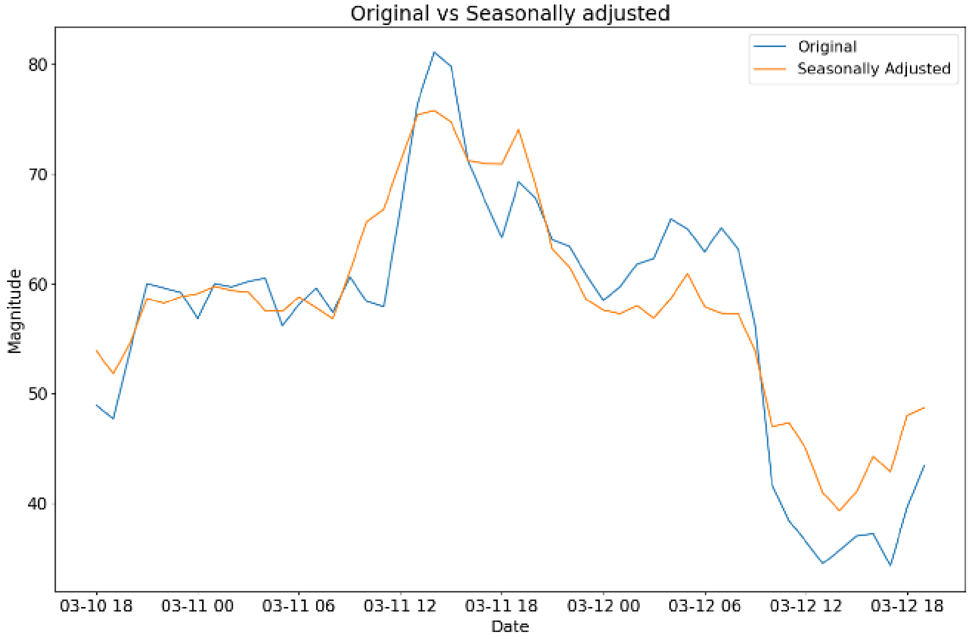


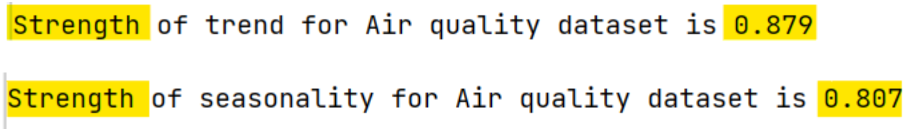


```
Results of KPSS Test:
Test Statistic              0.117078
p-value                     0.100000
LagsUsed                   53.000000
Critical Value (10%)        0.347000
Critical Value (5%)         0.463000
Critical Value (2.5%)       0.574000
Critical Value (1%)         0.739000
```

Time Series Decomposition

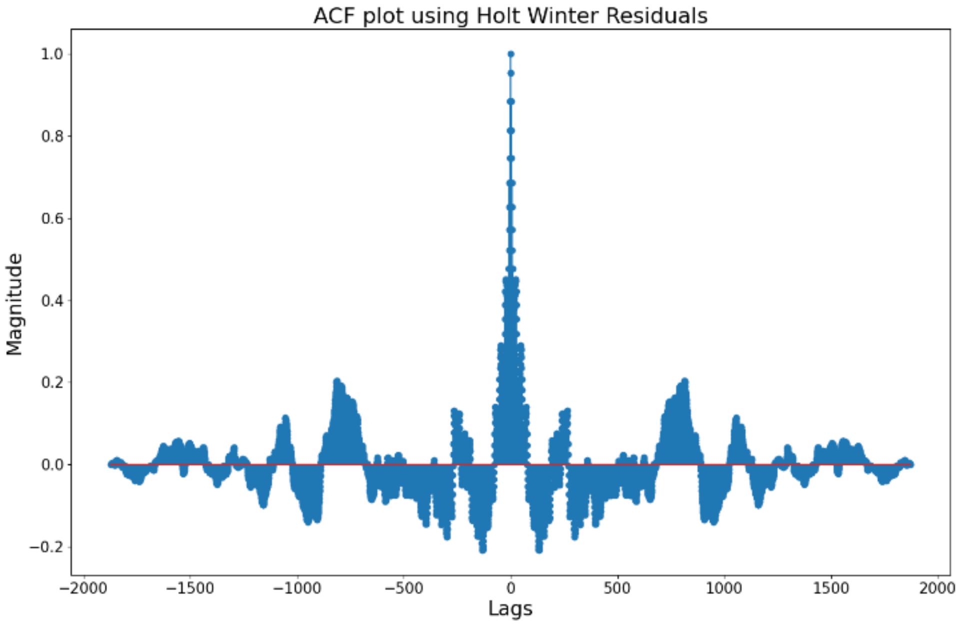

Trend, Seasonality, Residual components using STL Decomposition

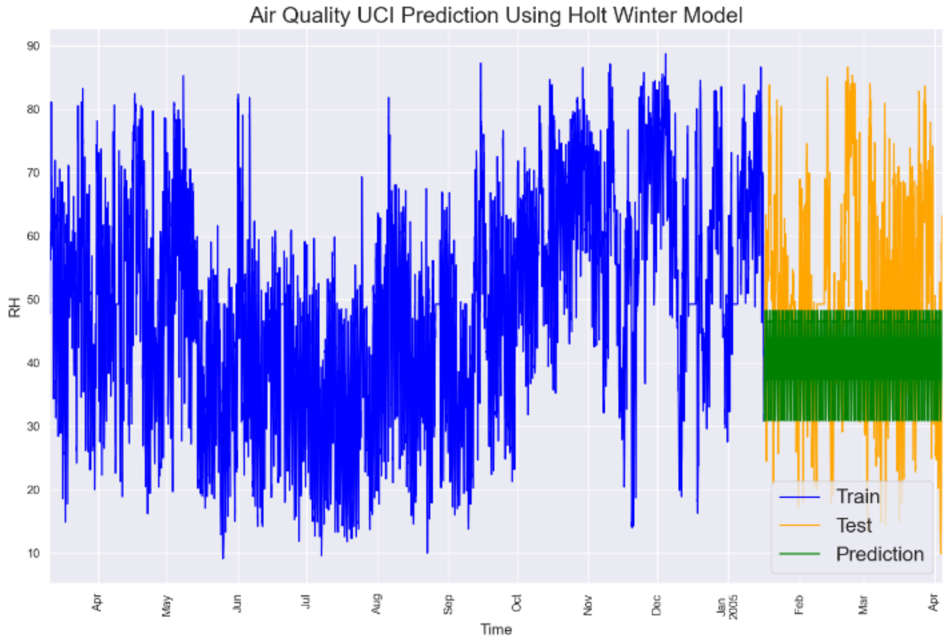

Original vs Seasonally adjusted



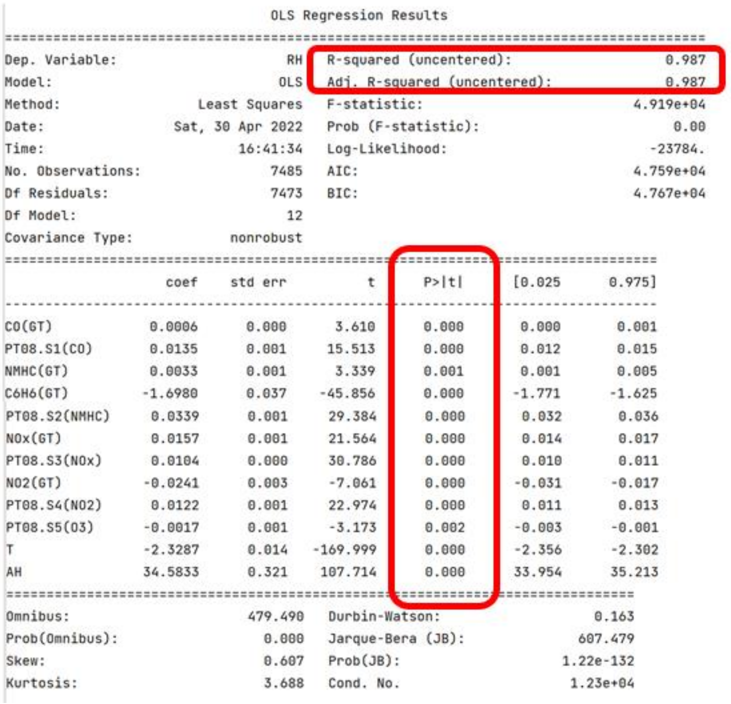
```
Strength of trend for Air quality dataset is 0.879

Strength of seasonality for Air quality dataset is 0.807
```

Holt-Winter

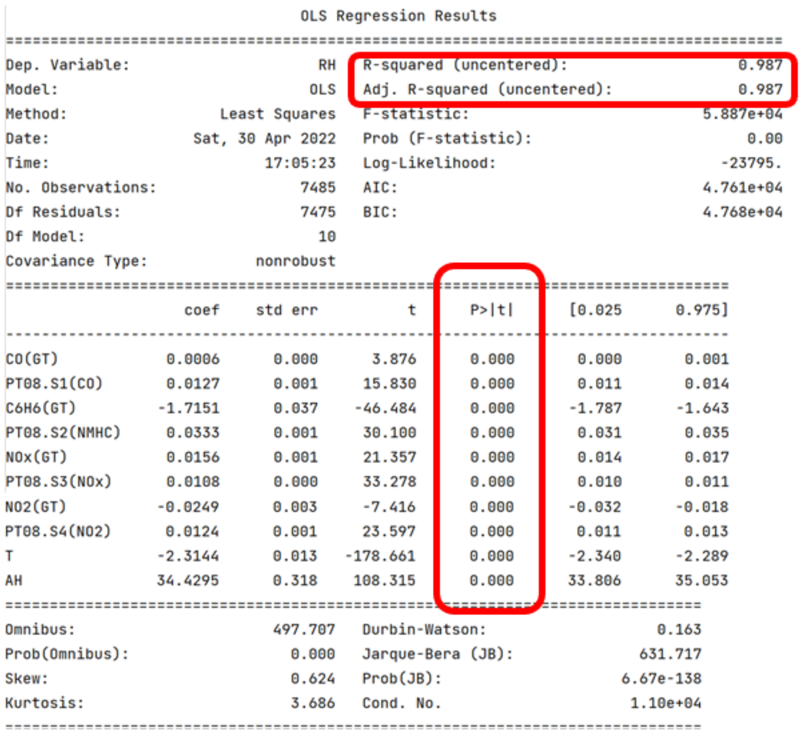

ACF plot using Holt Winter Residuals
Magnitude
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−2000
−1500
−1000
−500
0
500
1000
1500
2000
Lags


Air Quality UCI Prediction Using Holt Winter Model
RH
90
80
70
60
50
40
30
20
10
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
2005
Feb
Mar
Apr
Time
Train
Test
Prediction

Feature Selection

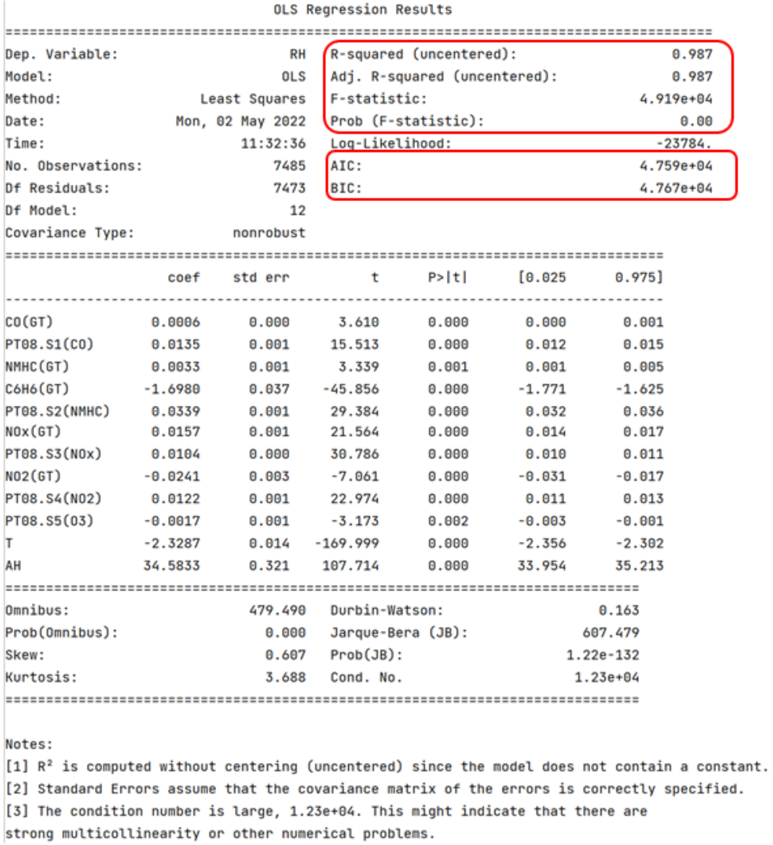


```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                     RH   R-squared (uncentered):                   0.987
Model:                            OLS   Adj. R-squared (uncentered):              0.987
Method:                 Least Squares   F-statistic:                          4.919e+04
Date:                Sat, 30 Apr 2022   Prob (F-statistic):                        0.00
Time:                        16:41:34   Log-Likelihood:                         -23784.
No. Observations:                7485   AIC:                                  4.759e+04
Df Residuals:                    7473   BIC:                                  4.767e+04
Df Model:                          12
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
CO(GT)           0.0006      0.000      3.610      0.000       0.000       0.001
PT08.S1(CO)      0.0135      0.001     15.513      0.000       0.012       0.015
NMHC(GT)         0.0033      0.001      3.339      0.001       0.001       0.005
C6H6(GT)        -1.6980      0.037    -45.856      0.000      -1.771      -1.625
PT08.S2(NMHC)    0.0339      0.001     29.384      0.000       0.032       0.036
NOx(GT)          0.0157      0.001     21.564      0.000       0.014       0.017
PT08.S3(NOx)     0.0104      0.000     30.786      0.000       0.010       0.011
NO2(GT)         -0.0241      0.003     -7.061      0.000      -0.031      -0.017
PT08.S4(NO2)     0.0122      0.001     22.974      0.000       0.011       0.013
PT08.S5(O3)     -0.0017      0.001     -3.173      0.002      -0.003      -0.001
T               -2.3287      0.014   -169.999      0.000      -2.356      -2.302
AH              34.5833      0.321    107.714      0.000      33.954      35.213
==============================================================================
Omnibus:                      479.490   Durbin-Watson:                   0.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              607.479
Skew:                           0.607   Prob(JB):                    1.22e-132
Kurtosis:                       3.688   Cond. No.                     1.23e+04
```

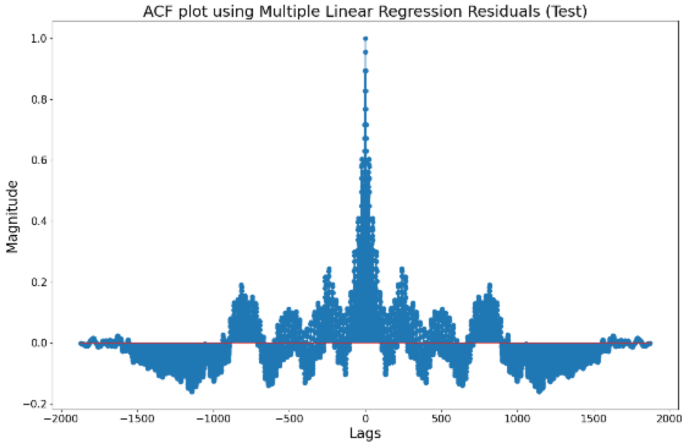


```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                     RH   R-squared (uncentered):                   0.987
Model:                            OLS   Adj. R-squared (uncentered):              0.987
Method:                 Least Squares   F-statistic:                          5.887e+04
Date:                Sat, 30 Apr 2022   Prob (F-statistic):                        0.00
Time:                        17:05:23   Log-Likelihood:                         -23795.
No. Observations:                7485   AIC:                                  4.761e+04
Df Residuals:                    7475   BIC:                                  4.768e+04
Df Model:                          10
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
CO(GT)           0.0006      0.000      3.876      0.000       0.000       0.001
PT08.S1(CO)      0.0127      0.001     15.830      0.000       0.011       0.014
C6H6(GT)        -1.7151      0.037    -46.484      0.000      -1.787      -1.643
PT08.S2(NMHC)    0.0333      0.001     30.100      0.000       0.031       0.035
NOx(GT)          0.0156      0.001     21.357      0.000       0.014       0.017
PT08.S3(NOx)     0.0108      0.000     33.278      0.000       0.010       0.011
NO2(GT)         -0.0249      0.003     -7.416      0.000      -0.032      -0.018
PT08.S4(NO2)     0.0124      0.001     23.597      0.000       0.011       0.013
T               -2.3144      0.013   -178.661      0.000      -2.340      -2.289
AH              34.4295      0.318    108.315      0.000      33.806      35.053
==============================================================================
Omnibus:                      497.707   Durbin-Watson:                   0.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              631.717
Skew:                           0.624   Prob(JB):                    6.67e-138
Kurtosis:                       3.686   Cond. No.                     1.10e+04
==============================================================================
```

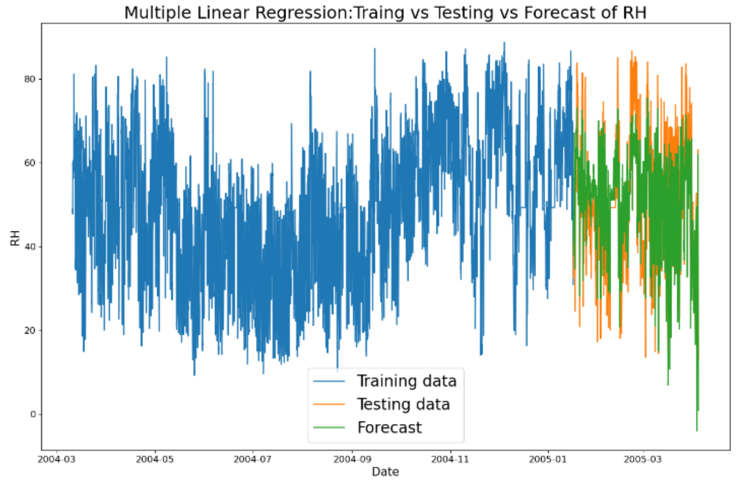
Multiple Linear Regression



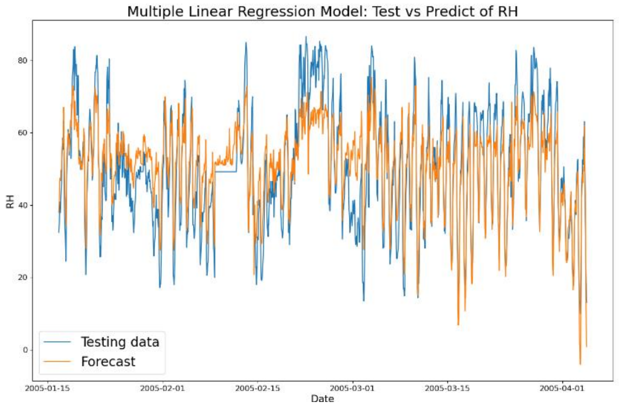
```
                                 OLS Regression Results
========================================================================================
Dep. Variable:                       RH   R-squared (uncentered):                    0.987
Model:                              OLS   Adj. R-squared (uncentered):               0.987
Method:                   Least Squares   F-statistic:                           4.919e+04
Date:                  Mon, 02 May 2022   Prob (F-statistic):                         0.00
Time:                          11:32:36   Log-Likelihood:                          -23784.
No. Observations:                  7485   AIC:                                   4.759e+04
Df Residuals:                      7473   BIC:                                   4.767e+04
Df Model:                            12
Covariance Type:              nonrobust
==================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
CO(GT)            0.0006      0.000      3.610      0.000       0.000       0.001
PT08.S1(CO)       0.0135      0.001     15.513      0.000       0.012       0.015
NMHC(GT)          0.0033      0.001      3.339      0.001       0.001       0.005
C6H6(GT)         -1.6980      0.037    -45.856      0.000      -1.771      -1.625
PT08.S2(NMHC)     0.0339      0.001     29.384      0.000       0.032       0.036
NOx(GT)           0.0157      0.001     21.564      0.000       0.014       0.017
PT08.S3(NOx)      0.0104      0.000     30.786      0.000       0.010       0.011
NO2(GT)          -0.0241      0.003     -7.061      0.000      -0.031      -0.017
PT08.S4(NO2)      0.0122      0.001     22.974      0.000       0.011       0.013
PT08.S5(O3)      -0.0017      0.001     -3.173      0.002      -0.003      -0.001
T                -2.3287      0.014   -169.999      0.000      -2.356      -2.302
AH               34.5833      0.321    107.714      0.000      33.954      35.213
===============================================================================
Omnibus:                        479.490   Durbin-Watson:                    0.163
Prob(Omnibus):                    0.000   Jarque-Bera (JB):               607.479
Skew:                             0.607   Prob(JB):                     1.22e-132
Kurtosis:                         3.688   Cond. No.                      1.23e+04
===============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 1.23e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```



ACF plot using Multiple Linear Regression Residuals (Test)

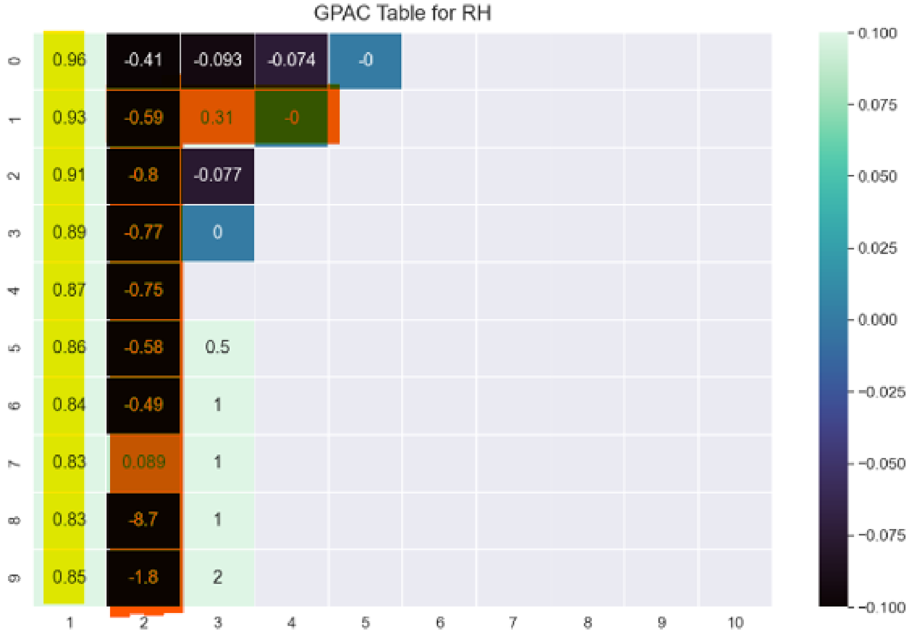


Multiple Linear Regression:Traing vs Testing vs Forecast of RH



Multiple Linear Regression Model: Test vs Predict of RH

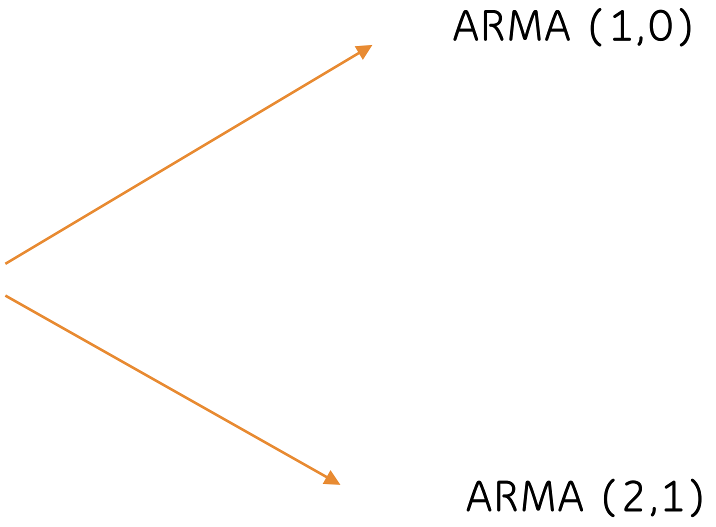
ARMA - Models



GPAC Table for RH

	1	2	3	4	5	6	7	8	9	10
0	0.96	-0.41	-0.093	-0.074	-0					
1	0.93	-0.59	0.31	-0						
2	0.91	-0.8	-0.077							
3	0.89	-0.77	0							
4	0.87	-0.75								
5	0.86	-0.58	0.5							
6	0.84	-0.49	1							
7	0.83	0.089	1							
8	0.83	-8.7	1							
9	0.85	-1.8	2							

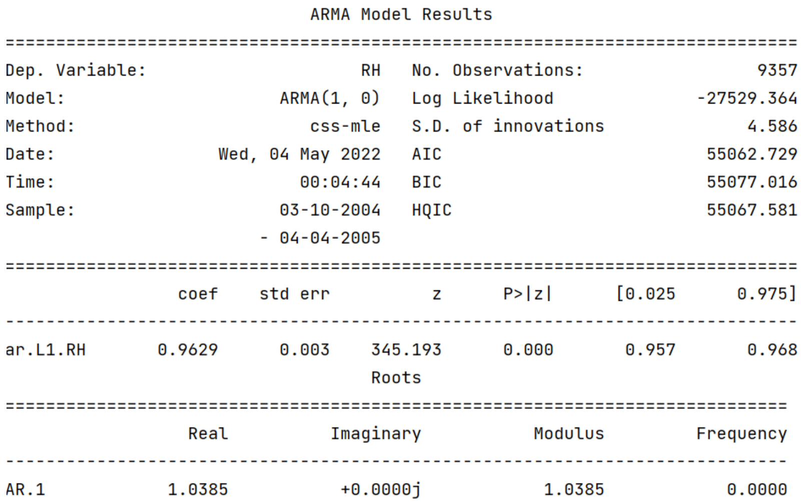
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100



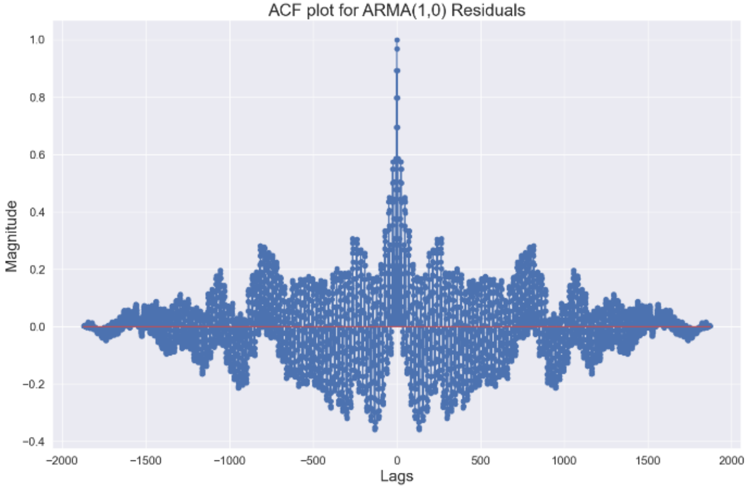
ARMA (1,0)

ARMA (2,1)

ARMA (1,0)

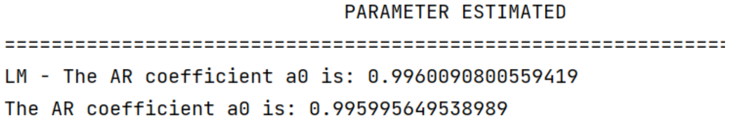


```
                              ARMA Model Results
==============================================================================
Dep. Variable:                     RH   No. Observations:                 9357
Model:                     ARMA(1, 0)   Log Likelihood              -27529.364
Method:                       css-mle   S.D. of innovations              4.586
Date:                Wed, 04 May 2022   AIC                          55062.729
Time:                        00:04:44   BIC                          55077.016
Sample:                    03-10-2004   HQIC                         55067.581
                         - 04-04-2005
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1.RH       0.9629      0.003    345.193      0.000       0.957       0.968
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.0385           +0.0000j            1.0385            0.0000
```

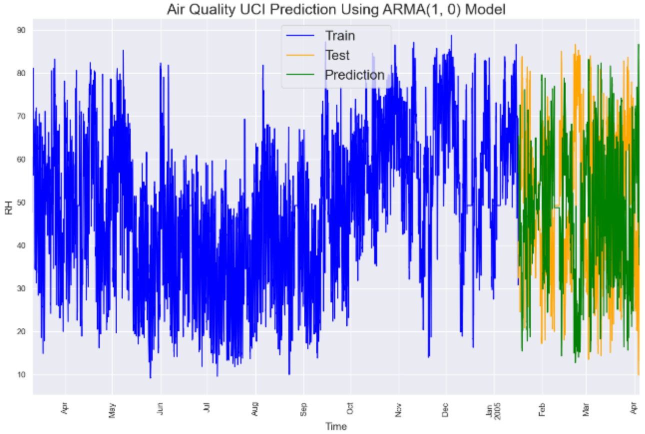


ACF plot for ARMA(1,0) Residuals

LM – Parameter Estimation

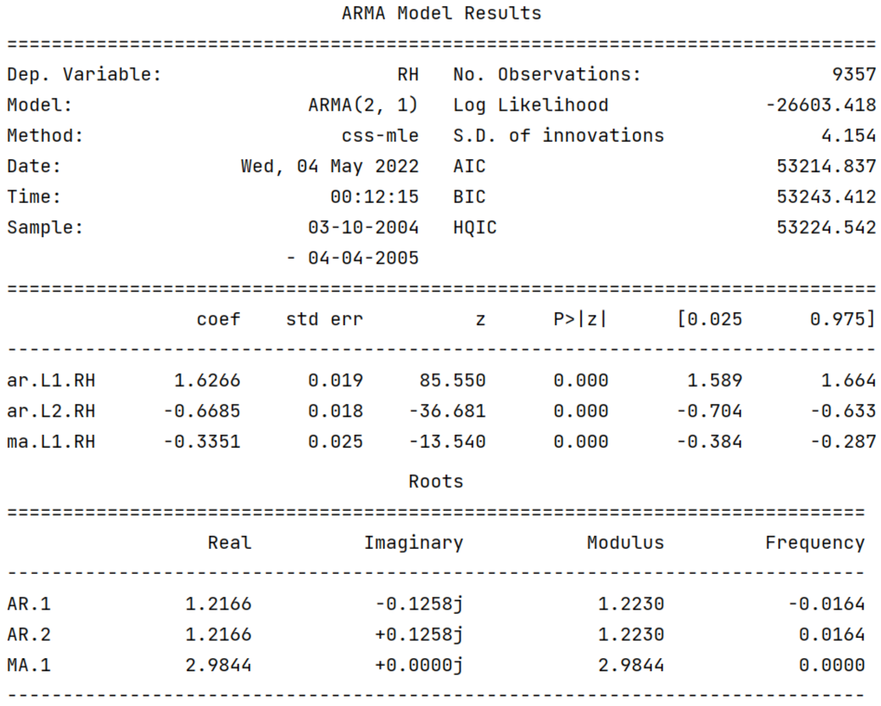


```
                    PARAMETER ESTIMATED
===============================================================
LM - The AR coefficient a0 is: 0.9960090800559419
The AR coefficient a0 is: 0.995995649538989
```

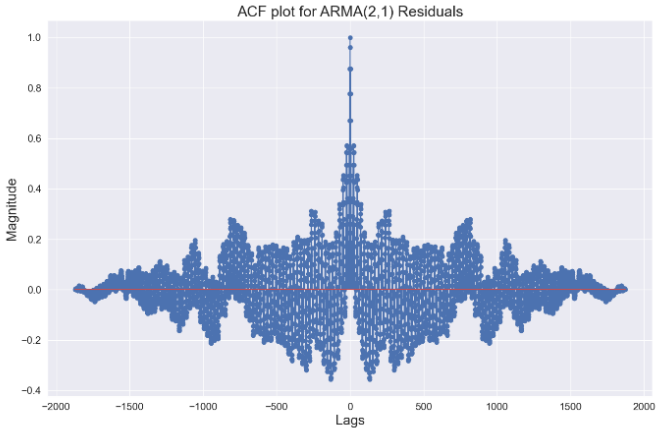


Air Quality UCI Prediction Using ARMA(1, 0) Model

ARMA (2,1)

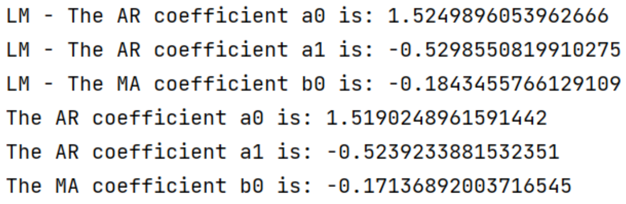


```
                              ARMA Model Results
==============================================================================
Dep. Variable:                     RH   No. Observations:                 9357
Model:                     ARMA(2, 1)   Log Likelihood              -26603.418
Method:                       css-mle   S.D. of innovations              4.154
Date:                Wed, 04 May 2022   AIC                          53214.837
Time:                        00:12:15   BIC                          53243.412
Sample:                    03-10-2004   HQIC                         53224.542
                         - 04-04-2005
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1.RH       1.6266      0.019     85.550      0.000       1.589       1.664
ar.L2.RH      -0.6685      0.018    -36.681      0.000      -0.704      -0.633
ma.L1.RH      -0.3351      0.025    -13.540      0.000      -0.384      -0.287
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.2166           -0.1258j            1.2230           -0.0164
AR.2            1.2166           +0.1258j            1.2230            0.0164
MA.1            2.9844           +0.0000j            2.9844            0.0000
-----------------------------------------------------------------------------
```

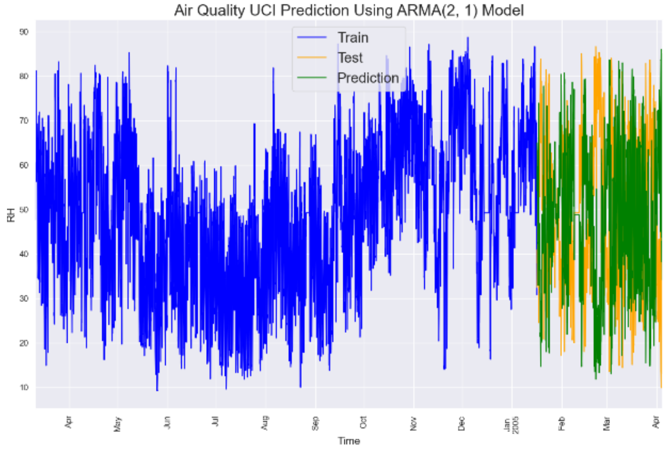


ACF plot for ARMA(2,1) Residuals

LM – Parameter Estimation

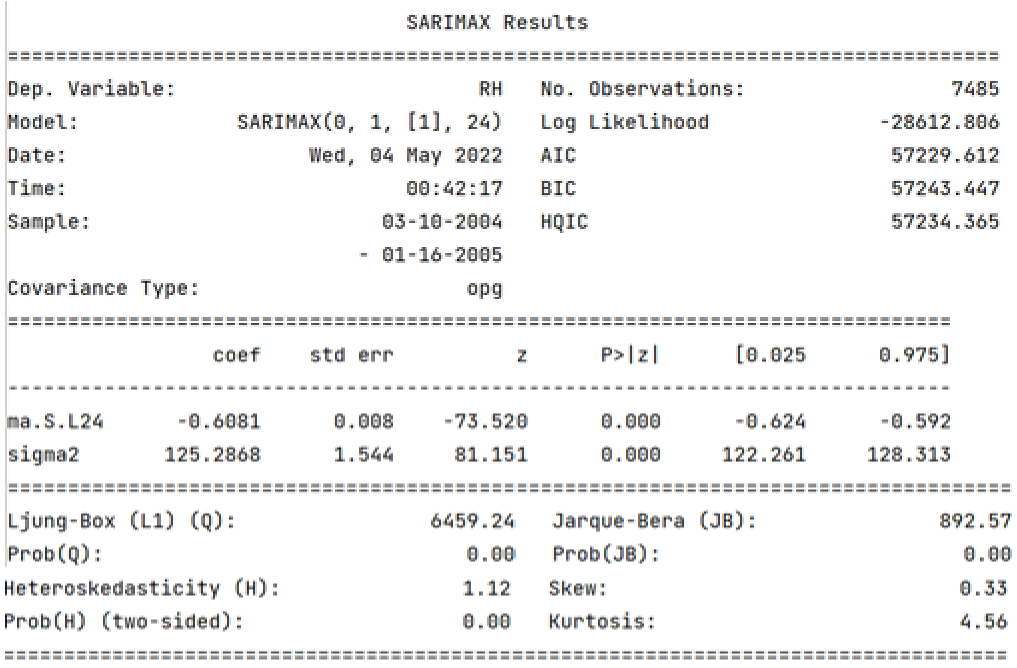


```
LM - The AR coefficient a0 is: 1.5249896053962666
LM - The AR coefficient a1 is: -0.5298550819910275
LM - The MA coefficient b0 is: -0.1843455766129109
The AR coefficient a0 is: 1.5190248961591442
The AR coefficient a1 is: -0.5239233881532351
The MA coefficient b0 is: -0.17136892003716545
```

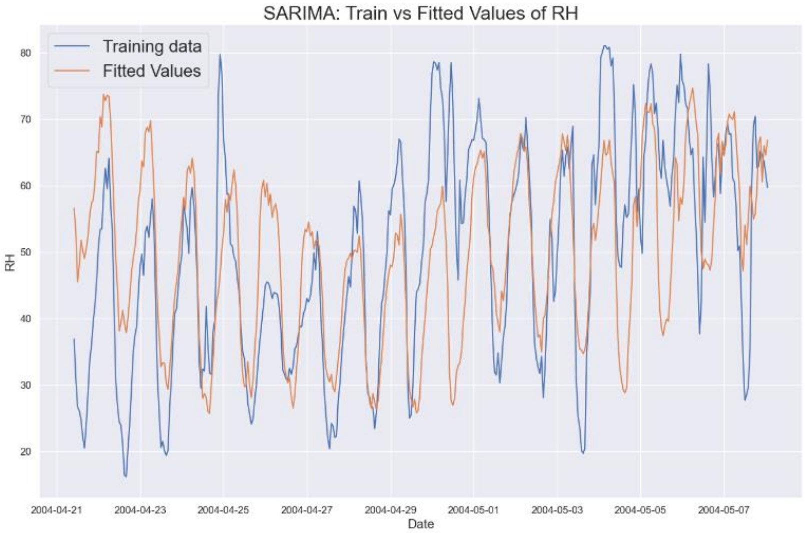


Air Quality UCI Prediction Using ARMA(2, 1) Model

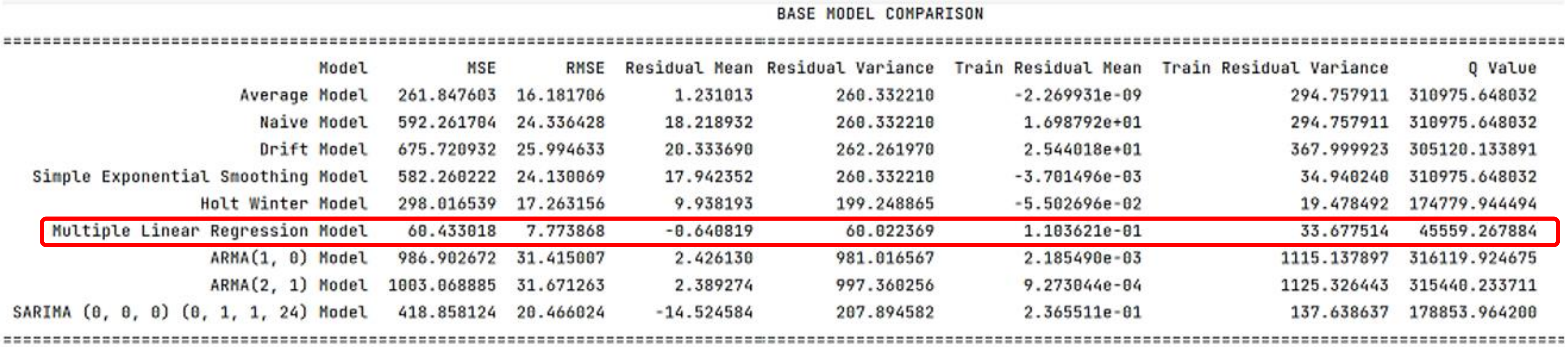
SARIMA (0,0,0) x(0,1,1,24)



```
                                     SARIMAX Results
==================================================================================
Dep. Variable:                            RH   No. Observations:             7485
Model:                SARIMAX(0, 1, [1], 24)   Log Likelihood          -28612.806
Date:                       Wed, 04 May 2022   AIC                      57229.612
Time:                               00:42:17   BIC                      57243.447
Sample:                           03-10-2004   HQIC                     57234.365
                                - 01-16-2005
Covariance Type:                         opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.S.L24      -0.6081      0.008    -73.520      0.000      -0.624      -0.592
sigma2       125.2868      1.544     81.151      0.000     122.261     128.313
===================================================================================
Ljung-Box (L1) (Q):                6459.24   Jarque-Bera (JB):                892.57
Prob(Q):                              0.00   Prob(JB):                          0.00
Heteroskedasticity (H):               1.12   Skew:                              0.33
Prob(H) (two-sided):                  0.00   Kurtosis:                          4.56
===================================================================================
```

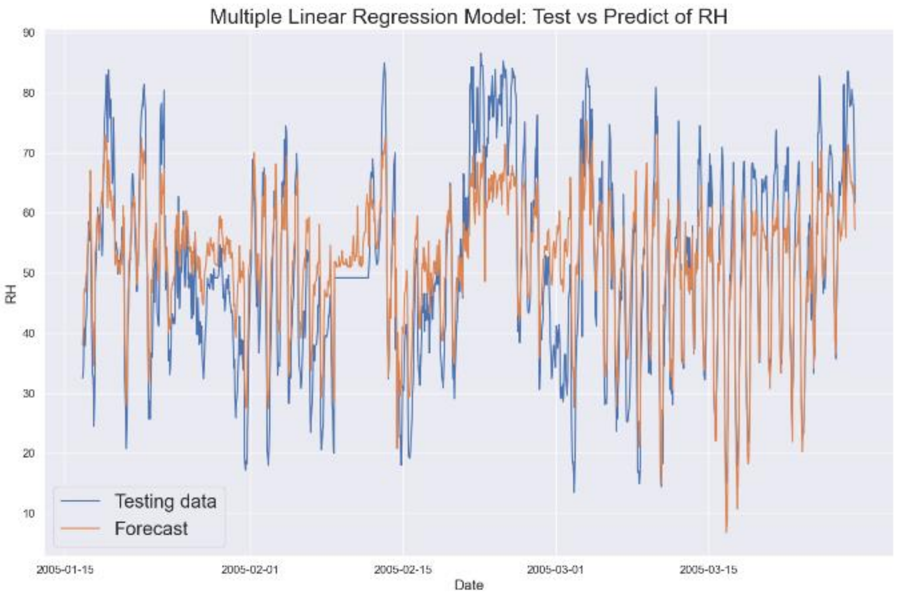


Model's Comparison



```
                                                                              BASE MODEL COMPARISON
==============================================================================================================================================================
                                Model          MSE       RMSE  Residual Mean  Residual Variance  Train Residual Mean  Train Residual Variance        Q Value
                        Average Model   261.847603  16.181706       1.231013         260.332210        -2.269931e-09               294.757911  310975.648032
                          Naive Model   592.261704  24.336428      18.218932         260.332210         1.698792e+01               294.757911  310975.648032
                          Drift Model   675.720932  25.994633      20.333690         262.261970         2.544018e+01               367.999923  305120.133891
   Simple Exponential Smoothing Model   582.260222  24.130069      17.942352         260.332210        -3.701496e-03                34.940240  310975.648032
                    Holt Winter Model   298.016539  17.263156       9.938193         199.248865        -5.502696e-02                19.478492  174779.944494
     Multiple Linear Regression Model    60.433018   7.773868      -0.640819          60.022369         1.103621e-01                33.677514   45559.267884
                     ARMA(1, 0) Model   986.902672  31.415007       2.426130         981.016567         2.185490e-03              1115.137897  316119.924675
                     ARMA(2, 1) Model  1003.068885  31.671263       2.389274         997.360256         9.273044e-04              1125.326443  315440.233711
 SARIMA (0, 0, 0) (0, 1, 1, 24) Model   418.858124  20.466024     -14.524584         207.894582         2.365511e-01               137.638637  178853.964200
==============================================================================================================================================================
```

Final Model – Multiple Linear Regression



Final Equation:

Y= $0.0006 * CD(GT) + 0.0135 * PT08.S1(CO) + 0.0033 * NMHC(GT) - 1.6980 * C6H6(G) + 0.0339 * PT08.S2(NMHC) + 0.0157 * NOx(GT) + 0.0104 * PT08.S3(N0x) - 0.0241 * NO2(GT) + 0.0122 * PT08.S4(NO2) - 0.0017 * PT08.S4(NO2) - 2.3287 * T + 34.5833 * AH$

Thanks!!!!!!