STATISTICS FOR DATASCIENCE PROJECT

Analysis of Regression on Air Quality Dataset

Team:

Santosh Chirag S20170010134
K Phani Sainath S20170010063
E Pruthvik Reddy S20170020203

Problem Statement:

The goal is to perform 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:

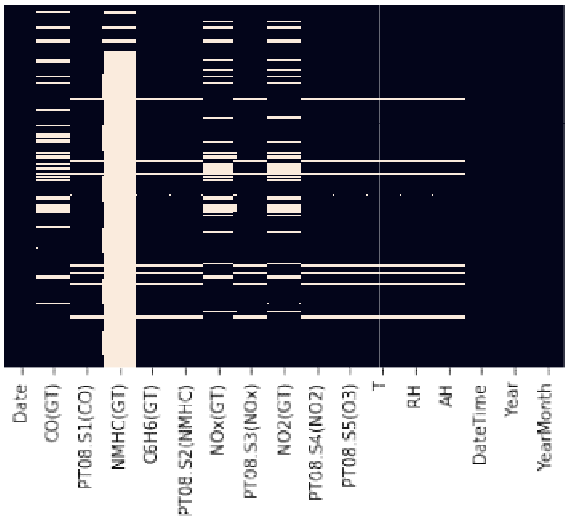
Time series analysis is a statistical technique that deals with time series data, or trend analysis. Time series data means that data is in a series of particular time periods or intervals.
In this report we get to know about the seasonal trends in the data and various tests to check the stationary and 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. It consists 9358 instances of hourly averaged responses from the years 2004 and 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.



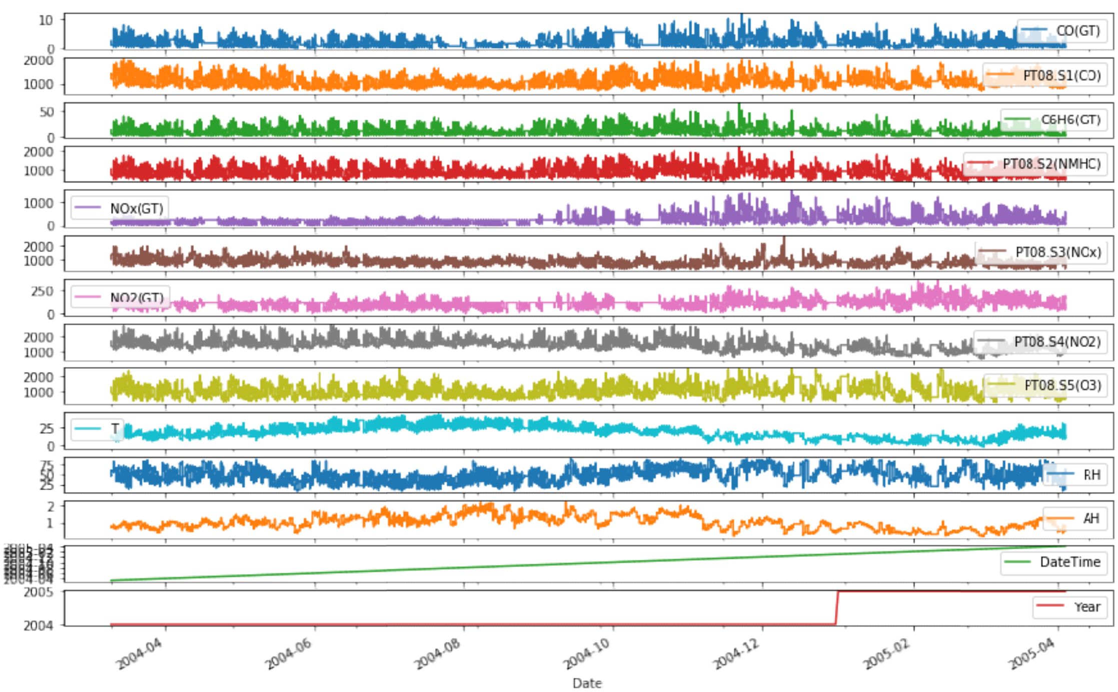
This is number of missing values after replacing -200 with Nan. There are some Nan values left even after imputing with mean because there is no data available for whole day so I filled those Nan values with previous values.

There is one feature NHMC which has 90% 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.

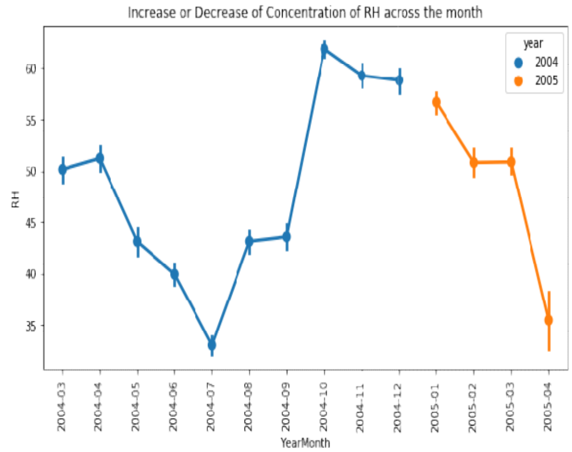
Outlier Detection:

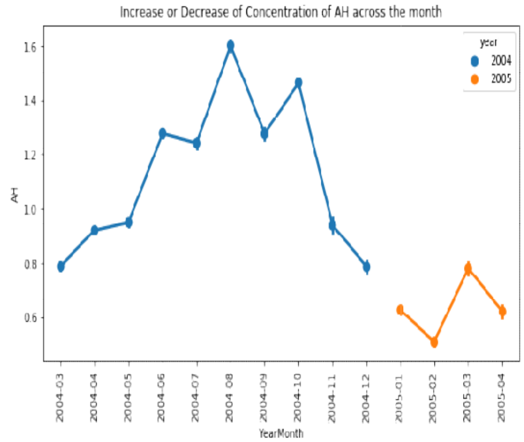
We use Z-score to detect and remove outliers

Data Analysis:



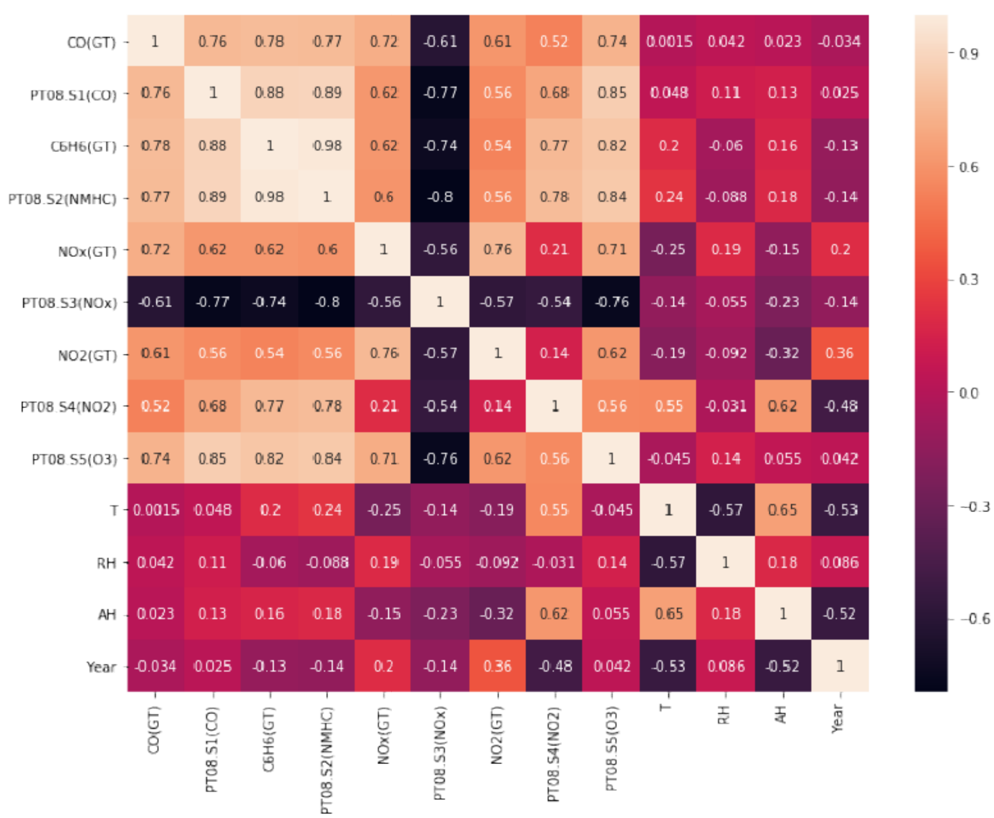
Above plot describes variations of all features across the time.





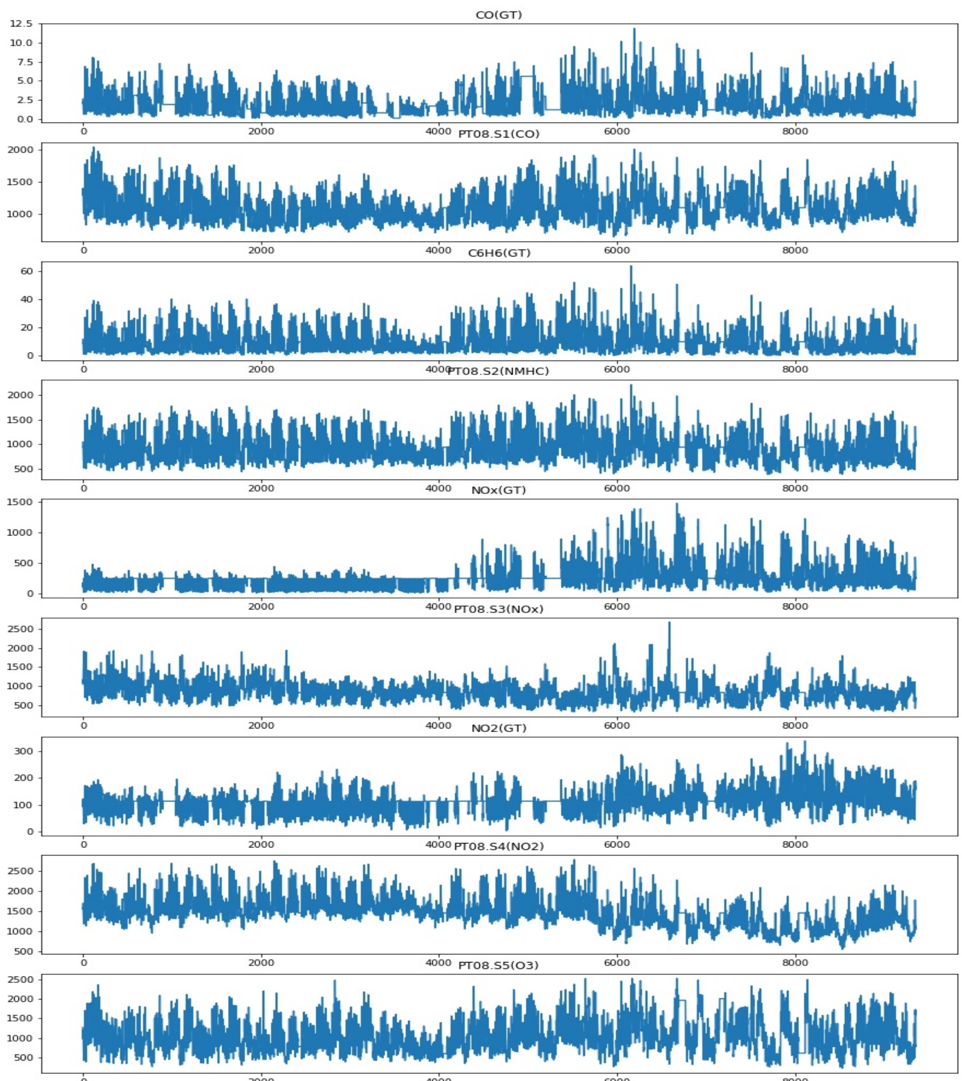
Above plots describes the variations of RH and AH along the year and increment or decrement of concentration across the months.

Below plot is the heatmap of correlation between all the variables.

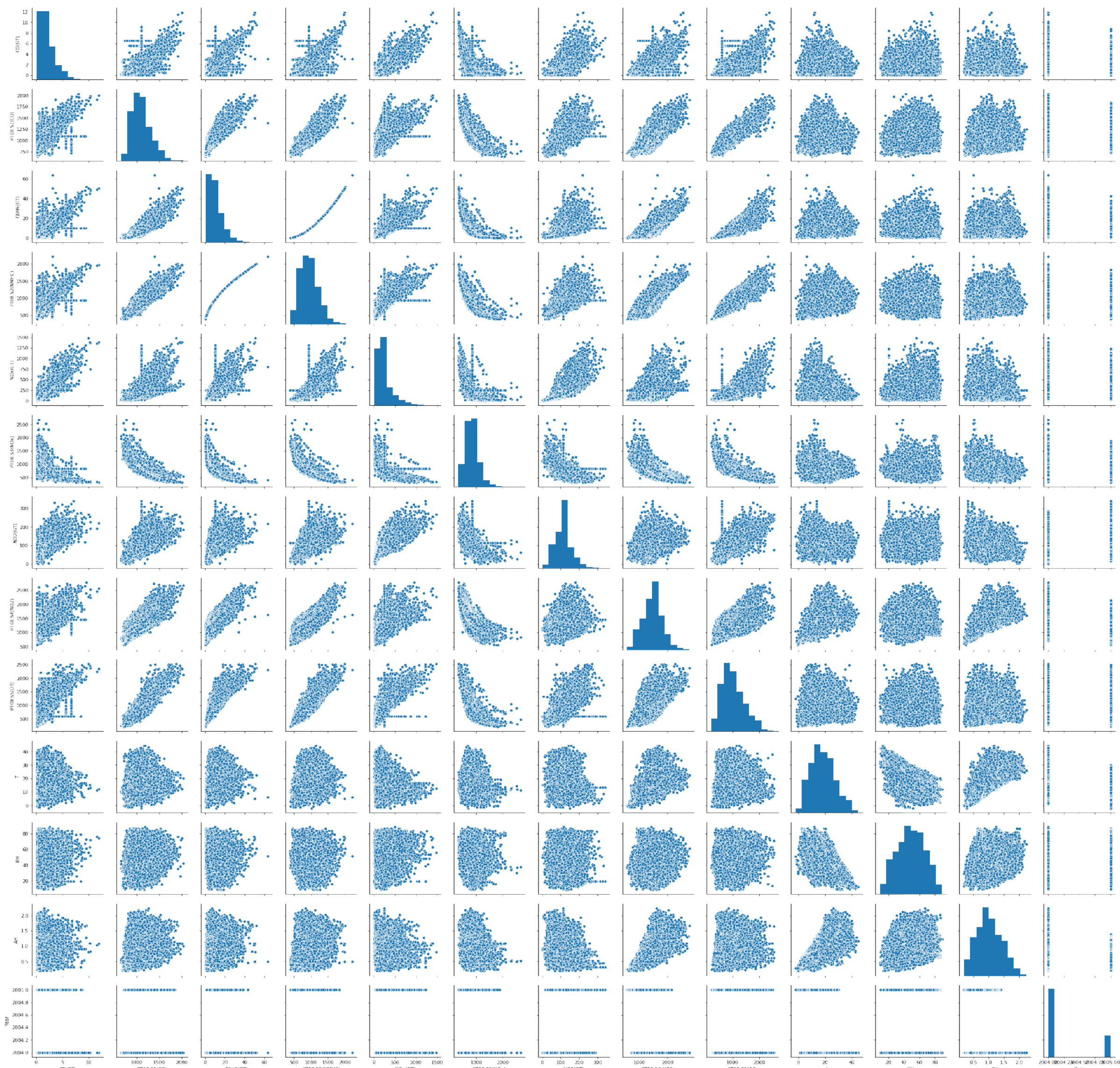


Check for stationarity:

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.



From the above plot, we can't decide the stationary of the data. To confirm stationary, we use Dickey Fuller Test for all the variables.



Above graph is the pair plot(scatter plot of all variables). Leaving the diagonal rows(as those scatter plots are plots between the variable and itself) and last row and last column of the graph(it is the Year variable). As it is evident from the graph the number of outliers are quite less, we assume that the outliers don't affect the model very much.

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

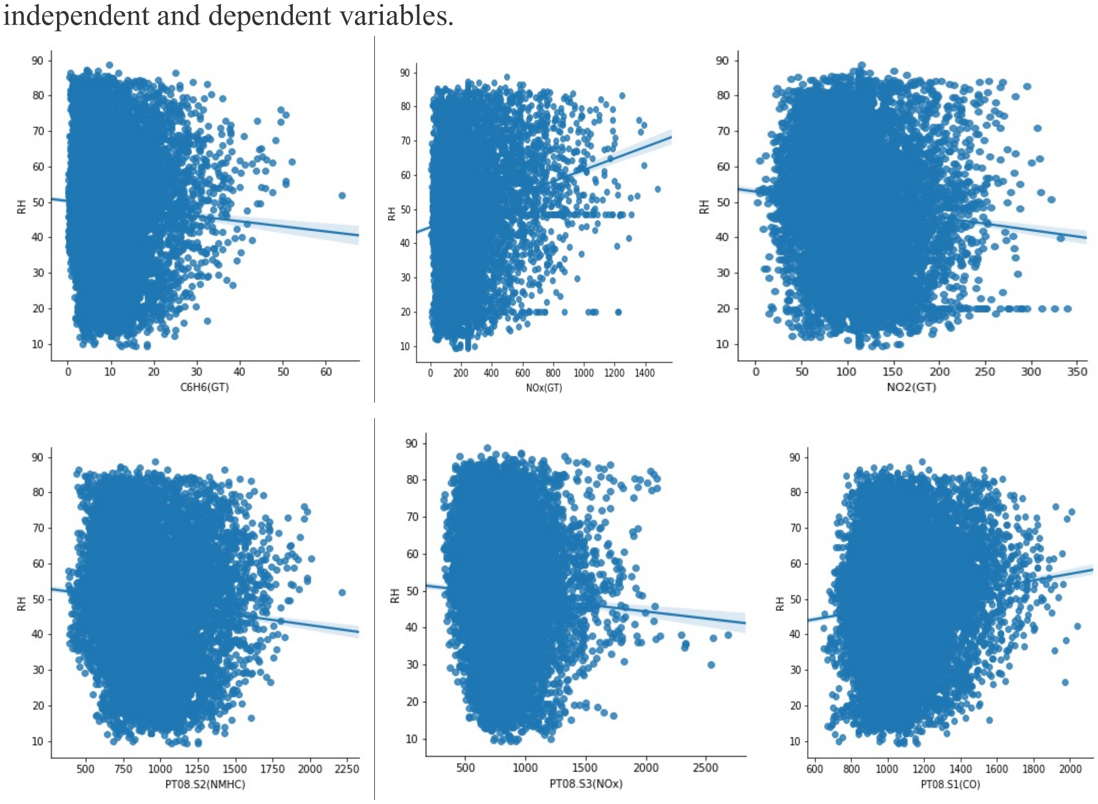
```
Results of Dickey-Fuller Test:
```

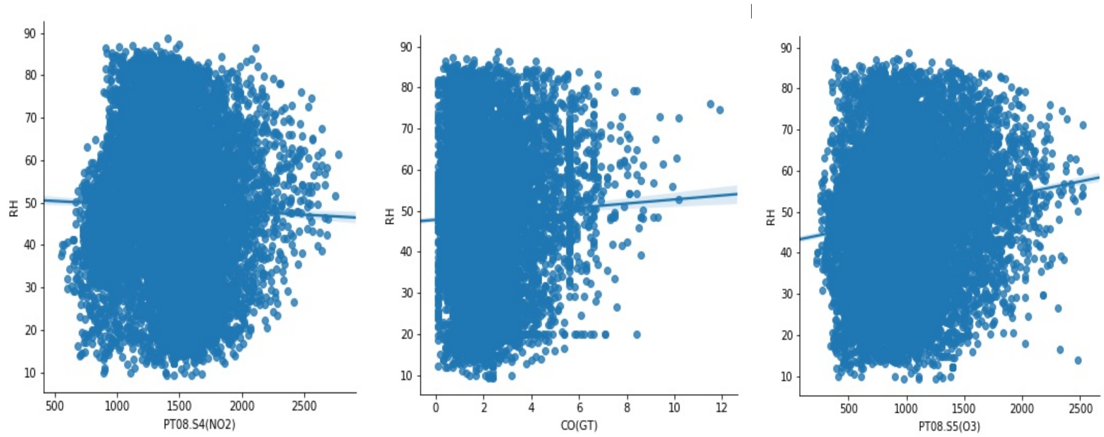
Feature	P_value
RH	1.219023e-10
AH	0.000014
CO (GT)	5.412775e-16
T	0.019787
NO2 (GT)	7.786800e-13
PT08.S4 (NO2)	3.185933e-08
PT08.S5 (O3)	2.251934e-19
C6H6 (GT)	3.127256e-18
PT08.S2 (NMHC)	1.779690e-18
PT08.S3 (NOx)	5.035225e-19
PT08.S1 (CO)	8.914162e-17
NOx (GT)	2.985511e-11

From the above results the test statistic < critical value, which implies that the series is stationary

Linearity Test:

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.

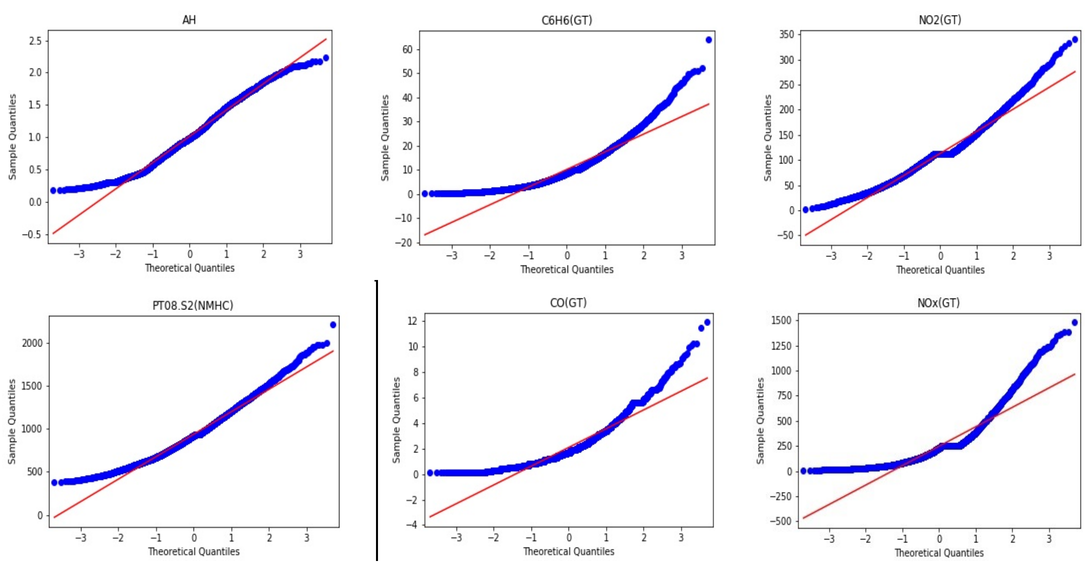


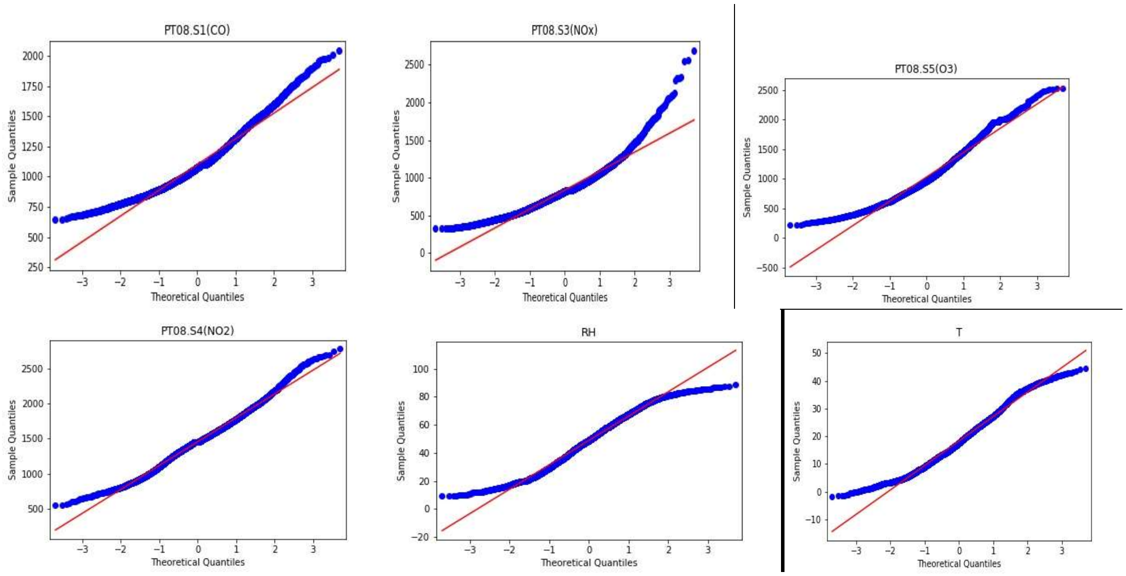


From the above graphs, we can say that there is no non-linear relationship between features and target variables.

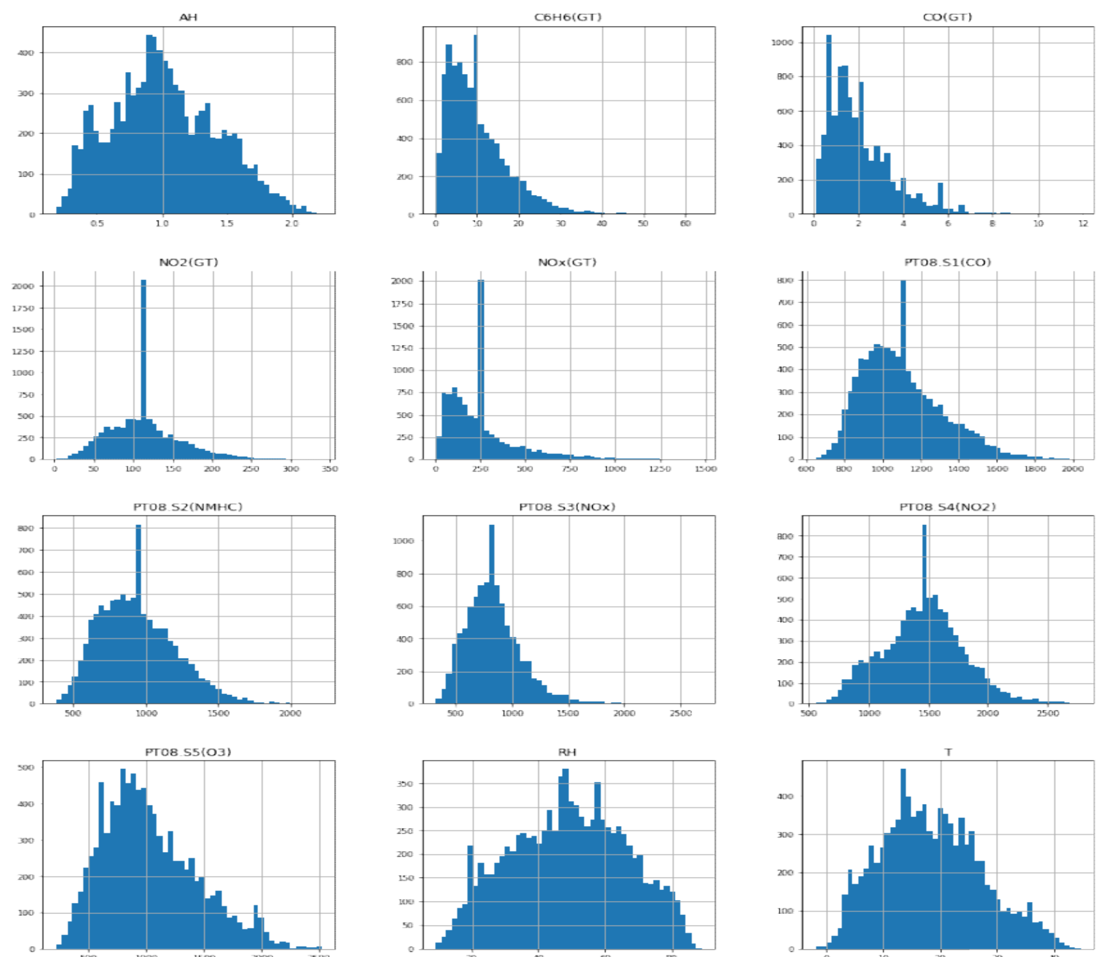
Normality Test:-

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 graphically or normal test.





Histogram Plot of the features:



Some of the features are normal as it can be clearly seen from either of the graphs.

For remaining features, we can perform the stats.normaltest and find the corresponding p value.

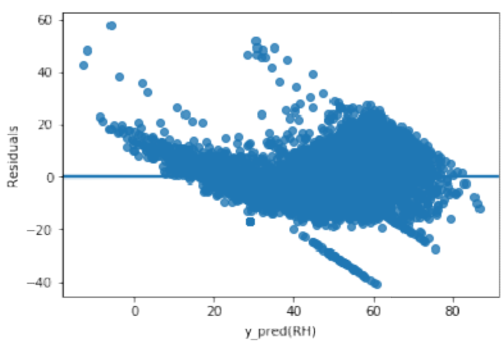
Normal Test:

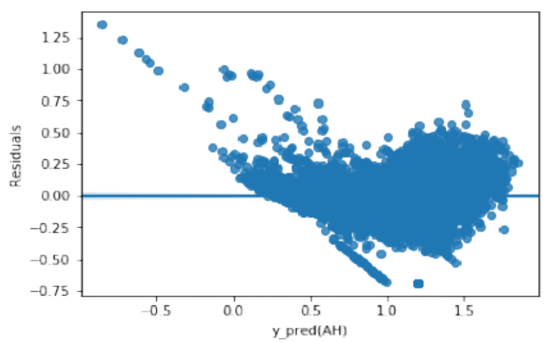
Feature	P value
CO (GT)	0.0
T	8.877198092891439e-64
NO2 (GT)	8.382026215464602e-183
PT08.S4 (NO2)	3.1093274946139267e-18
PT08.S5 (O3)	4.489891015439565e-125
C6H6 (GT)	0.0
PT08.S2 (NMHC)	5.154651033295069e-100
PT08.S3 (NOx)	0.0
PT08.S1 (CO)	1.7886360129305134e-173
NOx (GT)	0.0
RH	7.022315598570015e-178
AH	2.1014760517340424e-74

From the table, we can see that all the p values are almost equal to 0 i.e., negligible value indicating that the data is 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,meaningthat 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.

Ordinary Least Squares(OLS):

Applying the ols regression from statsmodels.api, we get the following results for both the target variables ‘RH’ and ‘AH’.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     RH   R-squared:                       0.736
Model:                            OLS   Adj. R-squared:                  0.736
Method:                 Least Squares   F-statistic:                     1824.
Date:                Thu, 28 Nov 2019   Prob (F-statistic):               0.00
Time:                        02:06:01   Log-Likelihood:                -23628.
No. Observations:                6549   AIC:                         4.728e+04
Df Residuals:                    6538   BIC:                         4.735e+04
Df Model:                          10
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const            92.9243      2.574     36.107      0.000      87.879      97.969
CO(GT)           -0.6787      0.143     -4.740      0.000      -0.959      -0.398
T                -1.7483      0.019    -89.939      0.000      -1.786      -1.710
NO2(GT)          -0.1604      0.004    -36.348      0.000      -0.169      -0.152
PT08.S4(NO2)      0.0526      0.001     67.446      0.000       0.051       0.054
PT08.S5(O3)       0.0009      0.001      1.399      0.162      -0.000       0.002
C6H6(GT)         -0.4566      0.098     -4.656      0.000      -0.649      -0.264
PT08.S2(NMHC)    -0.0654      0.003    -20.903      0.000      -0.072      -0.059
PT08.S3(NOx)     -0.0304      0.001    -35.133      0.000      -0.032      -0.029
PT08.S1(CO)       0.0099      0.001      6.985      0.000       0.007       0.013
NOx(GT)           0.0422      0.001     36.377      0.000       0.040       0.044
==============================================================================
Omnibus:                      527.376   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1892.550
Skew:                           0.360   Prob(JB):                         0.00
Kurtosis:                       5.533   Cond. No.                     5.82e+04
==============================================================================
```

The value of R-squared is 0.736. The Durbin-Watson statistic is 1.993 indicating that there is very less positive auto correlation

```
OLS Regression Results
==============================================================================
Dep. Variable:                     AH   R-squared:                       0.798
Model:                            OLS   Adj. R-squared:                  0.798
Method:                 Least Squares   F-statistic:                     2581.
Date:                Thu, 28 Nov 2019   Prob (F-statistic):               0.00
Time:                        02:06:17   Log-Likelihood:                 1854.9
No. Observations:                6549   AIC:                            -3688.
Df Residuals:                    6538   BIC:                            -3613.
Df Model:                          10
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const             1.6518      0.053     31.425      0.000       1.549       1.755
CO(GT)           -0.0119      0.003     -4.086      0.000      -0.018      -0.006
T                 0.0150      0.000     37.834      0.000       0.014       0.016
NO2(GT)          -0.0032   9.01e-05    -35.624      0.000      -0.003      -0.003
PT08.S4(NO2)      0.0012   1.59e-05     74.450      0.000       0.001       0.001
PT08.S5(O3)    1.538e-05   1.28e-05      1.204      0.229   -9.66e-06    4.04e-05
C6H6(GT)          0.0079      0.002      3.959      0.000       0.004       0.012
PT08.S2(NMHC)    -0.0019   6.39e-05    -29.676      0.000      -0.002      -0.002
PT08.S3(NOx)     -0.0009   1.77e-05    -48.164      0.000      -0.001      -0.001
PT08.S1(CO)   -4.621e-05    2.9e-05     -1.595      0.111      -0.000    1.06e-05
NOx(GT)           0.0008   2.37e-05     33.584      0.000       0.001       0.001
==============================================================================
Omnibus:                      724.529   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2793.300
Skew:                           0.510   Prob(JB):                         0.00
```

```
Kurtosis:                       6.033   Cond. No.                       5.82e+04
==============================================================================
```

Similarly, Durbin-Watson statistic is 2.006 which means there is negligible –ve auto correlation.

R-squared value is 0.798.

Linear Regression

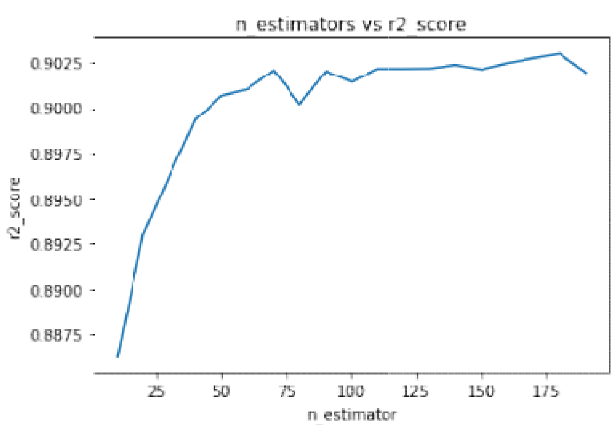
We apply Linear Regression to our data by training the model using train data and predicting the test data set.

Co-efficients of the linear regression after training the model:

```
NO2(GT)                      -1.60427075e-01
PT08.S2(NMHC)                -6.53760284e-02
PT08.S4(NO2)                  5.26389015e-02
PT08.S1(CO)                   9.90993150e-03
NOx(GT)                       4.21687016e-02
PT08.S5(O3)                   8.74899456e-04
T                            -1.74831852e+00
CO(GT)                       -6.78736334e-01
PT08.S3(NOx)                 -3.03696531e-02
C6H6(GT)                     -4.56647206e-01
R2_score for the above model is 0.7587165188730305
```

Random Forest Regression

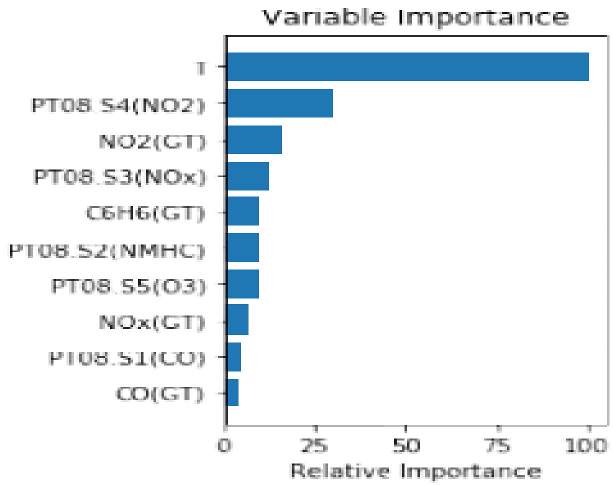
After pre processing and exploring the data, we apply RandomForestRegressor to the data with varying n_estimators from 10 to 190 with a increment of 10.



From the above graph, we can conclude that n_estimators = 180 gives the best r2_score.
Importances of the feature columns:

```
#NO2(GT)              0.0787668829454153
#PT08.S2(NMHC)        0.04617378205324278
#PT08.S4(NO2)         0.14925517187506443
PT08.S1(CO)           0.019729577642109388
NOx(GT)               0.03046353526095148
#PT08.S5(O3)          0.04569835497602727
```

```
#T                      0.4989338535487103
CO(GT)                  0.017958792496346597
#PT08.S3(NOx)           0.06407520224000192
#C6H6(GT)               0.04894484696213085
```



Factor Analysis:

Factor Analysis (FA) is an exploratory data analysis method used to search influential underlying factors or latent variables from a set of observed variables. It extracts maximum common variance from all variables and puts them into a common score.

Adequacy Test

Bartlett's Test

Kaiser-Meyer-Olkin Test

Bartlett's test of sphericity checks whether or not the observed variables inter correlate at all using the observed correlation matrix against the identity matrix. If the test found statistically insignificant, you should not employ a factor analysis.

Result for Bartlett sphericity test:

```
121299.59066771198 0.0
```

In this Bartlett 's test, the p-value is 0. The test was statistically significant, indicating that the observed correlation matrix is not an identity matrix.

Kaiser-Meyer-Olkin (KMO) Test measures the suitability of data for factor analysis. It determines the adequacy for each observed variable and for the complete model. KMO estimates the proportion of variance among all the observed variable. Lower proportion id more suitable for factor analysis. KMO values range between 0 and 1. Value of KMO less than 0.6 is considered inadequate.

Result:

```
0.865224518839246
```

The overall KMO for our data is 0.8652, which is considerably good for the above test.

Next ,we perform the Factor Analysis for the data.

After performing the Factor Analysis, we get the cumulative variance as follows

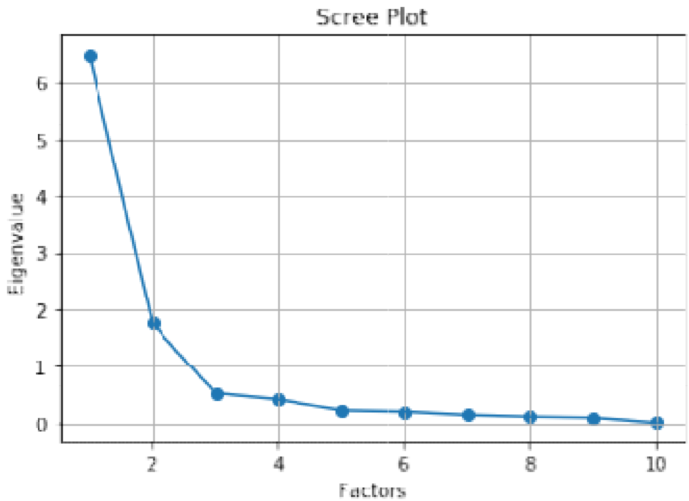
Cumulative Var	0.455848	0.572268	0.784539

Eigen values after performing factor analysis:

	Original_Eigenvalues
0	6.482814
1	1.760071
2	0.526873
3	0.415380
4	0.227362
5	0.206325
6	0.150146
7	0.118467
8	0.101563
9	0.010999

Cumulative variance of the eigen values:

```
0.64828141 0.82428849 0.87697581 0.91851384 0.91961372 0.94234993
 0.96298246 0.97799705 0.98815333 1.
```



Clearly, the four factors explain approximately 91% of the variance. Therefore, the number of factors will be equal to 4 in our case.

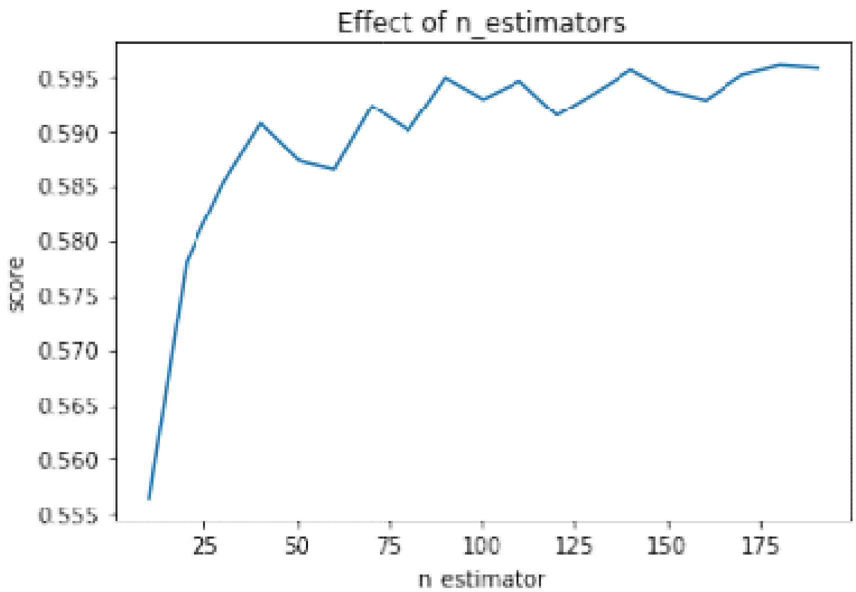
Factor loadings of the data:

	Factor1	Factor2	Factor3	Factor4
NO2(GT)	0.245677	-0.102703	0.762700	0.212014
PT08.S2(NMHC)	0.867775	0.171821	0.421193	0.168409
PT08.S4(NO2)	0.770448	0.446725	0.043618	0.091084
PT08.S1(CO)	0.777560	-0.009076	0.426697	0.264350
NOx(GT)	0.290956	-0.154306	0.884593	0.034160
PT08.S5(O3)	0.661551	-0.065486	0.543830	0.284721
T	0.159564	0.969817	-0.169311	0.018212
CO(GT)	0.584358	0.010877	0.608930	0.043435
PT08.S3(NOx)	-0.572996	-0.120531	-0.446290	-0.525115
C6H6(GT)	0.893076	0.127648	0.424018	0.036974

Principal Component Analysis (PCA):

Here we will explore the most important method of Feature Extraction which is Principal Component Analysis and will use this method to reduce the features and use the output for modeling.

After reducing the factors, and applying RandomForestRegression with varying n_estimators, we get the R-2 value as follows:


Effect of n_estimators
score
0.595
0.590
0.585
0.580
0.575
0.570
0.565
0.560
0.555
25
50
75
100
125
150
175
n_estimator