SDA Project

Group - 2

Santosh Chirag	S20170010134
Phani Kanala	S20170010063
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.

Data Description

Dataset : Air Quality dataset from UCI machine Learning Repository

Number of Instances: 9358 (hourly averaged responses of pollutants)

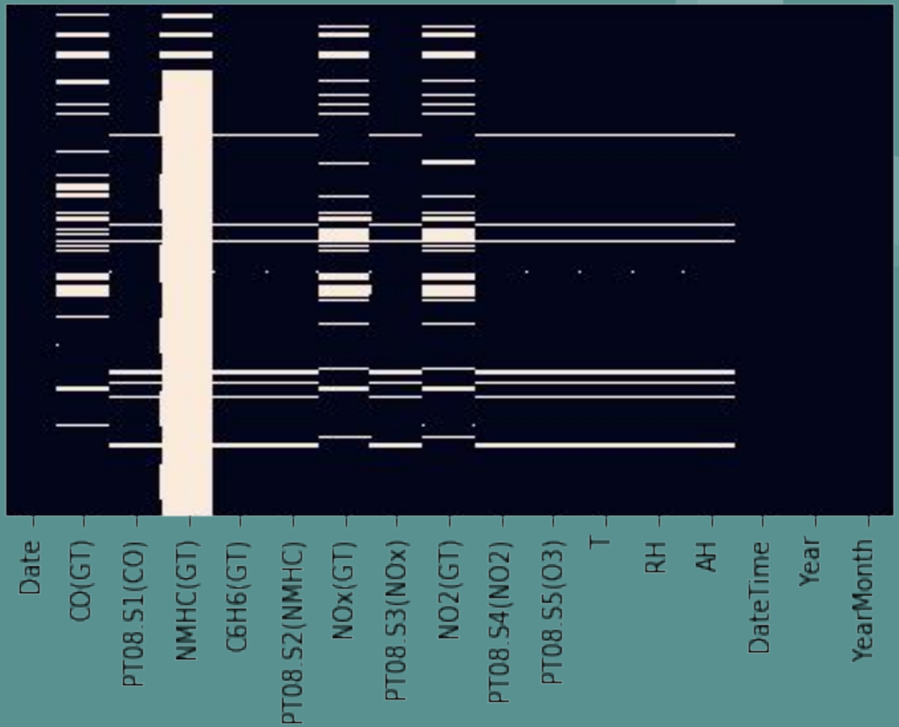
The dataset consists of following attributes: (Total :15)
Date, Time, CO(GT), PT08.S1(CO), NMHC(GT),C6H6(GT), PT08.S2(NHMC), NOx(GT), PT08.S3(NOx), NO2(GT), PT08.S4(NO2),PT08.S5(O3),T,RH and AH.

Dependent Variables: RH (Relative Humidity) and AH (Absolute Humidity)

Missing values are tagged by -200

Preprocessing

- Replaced missing values with NaN
- Below plot is heatmap of missing values of all variables



Preprocessing

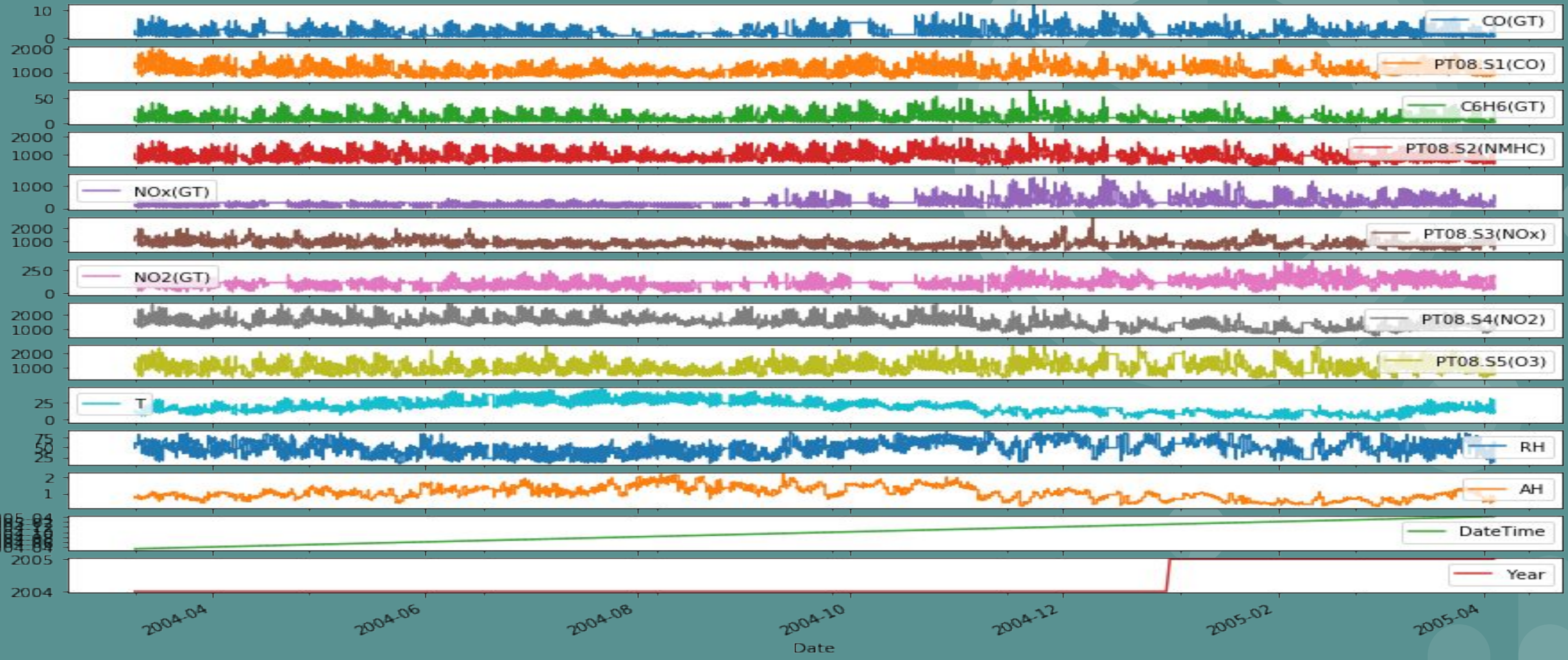
The figure indicates number of missing values after replacing -200 with Nan.

There are some Nan values left even after replacing with mean because there is no data available for whole day so we filled those Nan values with previous values.

There is one feature NHMC which has 90% of data composed of Nan. Replacing it with mean values is not a good idea as there are more Nan values.

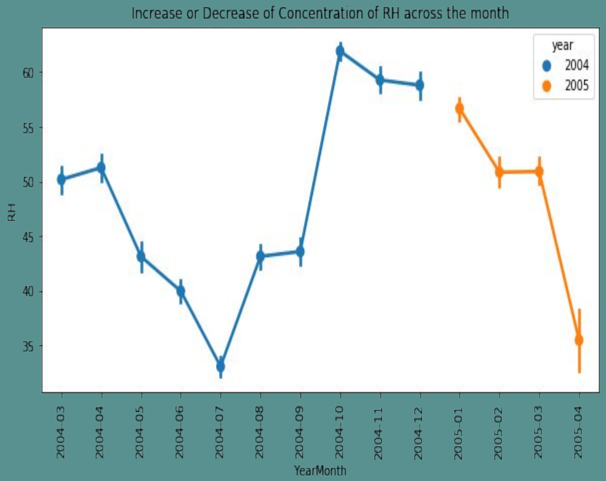
As it doesn’t provide much information that attribute was dropped.

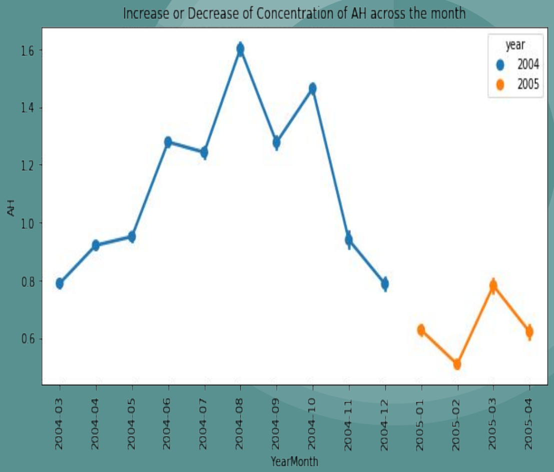
Data Analysis



Above plot describes variations of all features across the time.

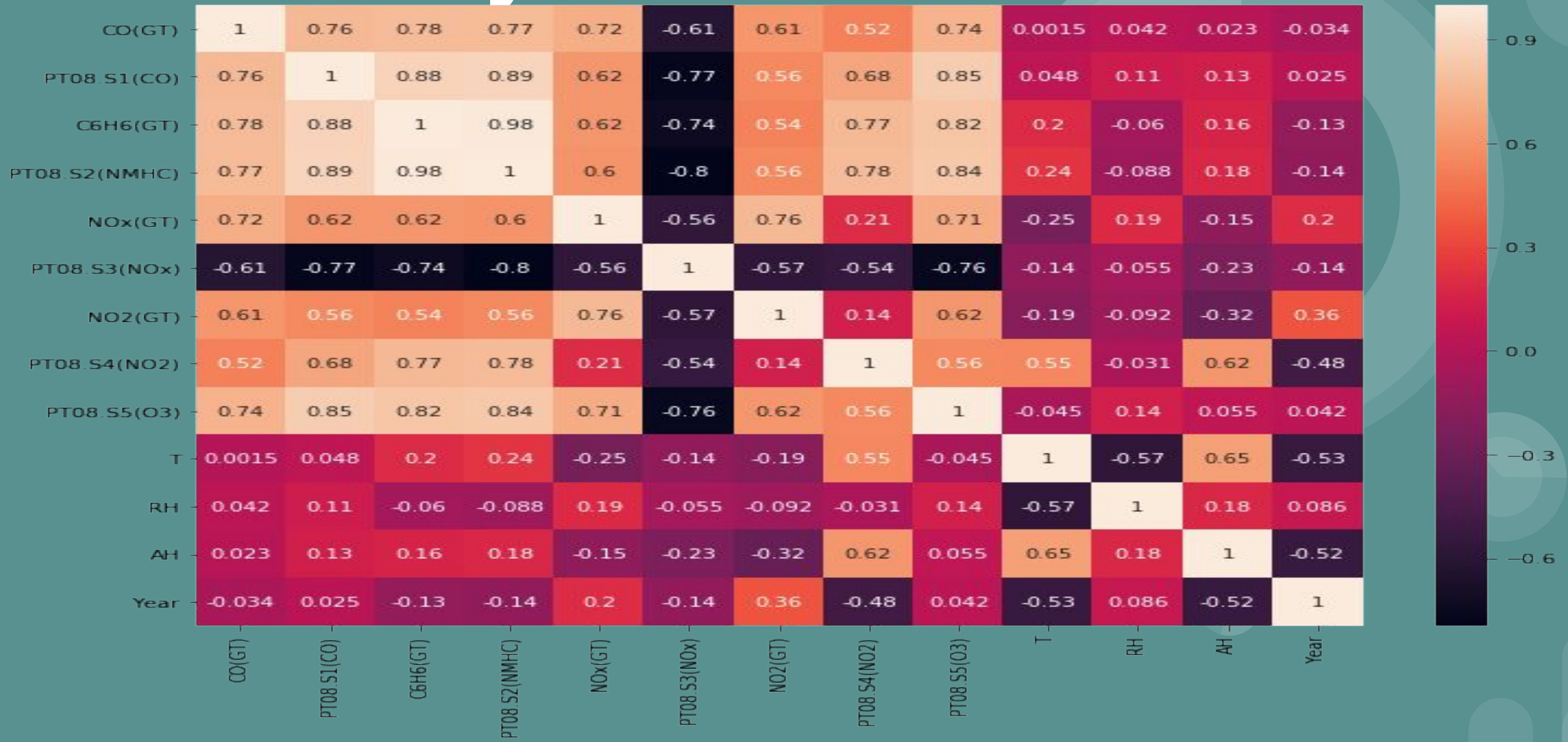
Data Analysis





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

Data Analysis



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

Check for Stationarity

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

Check for Stationarity

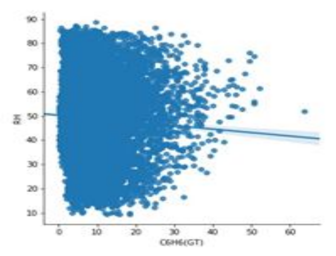
Results of Dickey-Fuller Test:

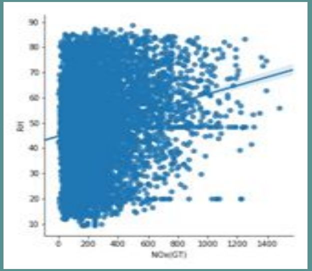
Feature	P_value
RH	1.219023e-10
AH	0.000014
CO (GT)	5.412775e-16
T	0.019787
NO2 (GT)	7.786800e-13
PT08.S4 (NO2)	3.185933e-08
PT08.S5 (O3)	2.251934e-19
C6H6 (GT)	3.127256e-18
PT08.S2 (NMHC)	1.779690e-18
PT08.S3 (NOx)	5.035225e-19
PT08.S1 (CO)	8.914162e-17
NOx (GT)	2.985511e-11

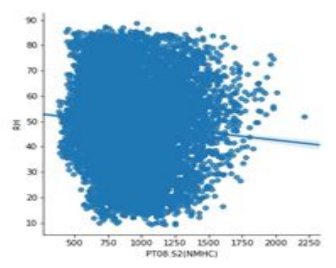
From the above results the test statistic < critical value(alpha = 0.05), which implies that the series is stationary

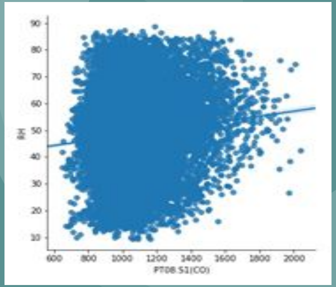
Linearity Test:

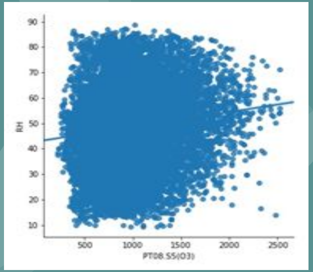
Linear regression needs the relationship between the independent and dependent variables to be linear.

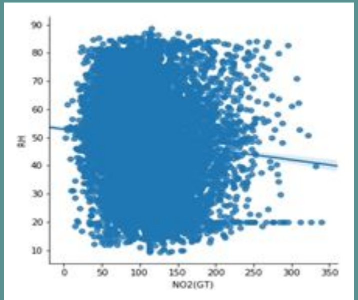


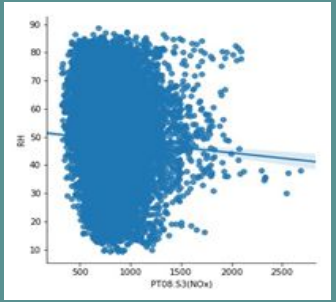


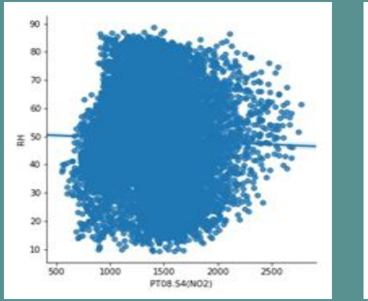


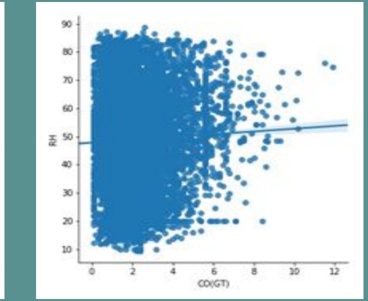








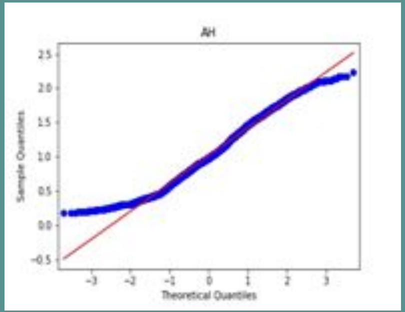


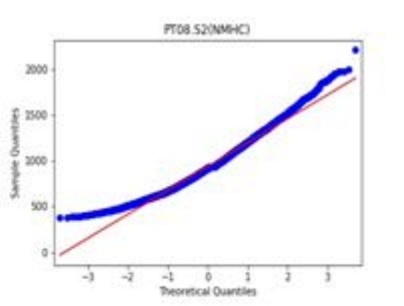


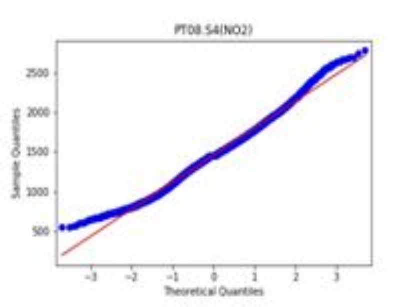
Above graphs prove that there is a linear relationship between independant variables and target variables.

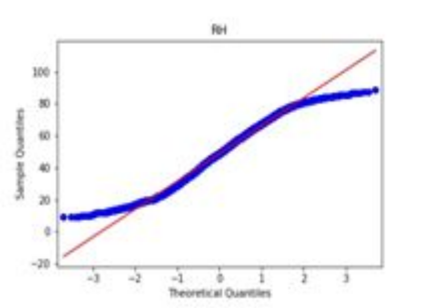
Normality Test:-

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.









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.

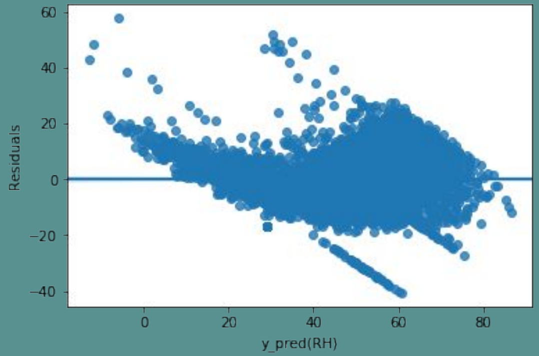
We use normaltest from scipy package to test the normality.

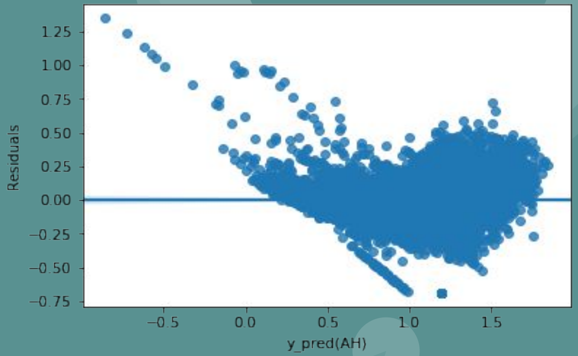
Feature	P value
CO (GT)	0.0
T	8.877198092891439e-64
NO2 (GT)	8.382026215464602e-183
PT08.S4 (NO2)	3.1093274946139267e-18
PT08.S5 (O3)	4.489891015439565e-125
C6H6 (GT)	0.0
PT08.S2 (NMHC)	5.154651033295069e-100
PT08.S3 (NOx)	0.0
PT08.S1 (CO)	1.7886360129305134e-173
NOx (GT)	0.0
RH	7.022315598570015e-178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

Homoscadasticity:

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’.

For RH variable some of the results:

R-squared:	0.736
Durbin-Watson	1.993

For RH variable some of the results:

R-squared:	0.798
Durbin-Watson	2.006

Linear Regression

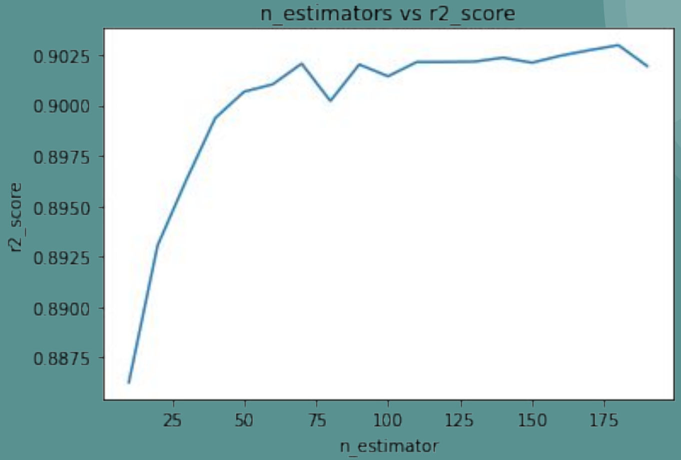
Linear regression is a statistical approach for modelling relationship between a dependent variable with a given set of independent variables.

After apply Linear Regression to our data by training the model using train data and predicting the test data set,we obtain the following results:

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.

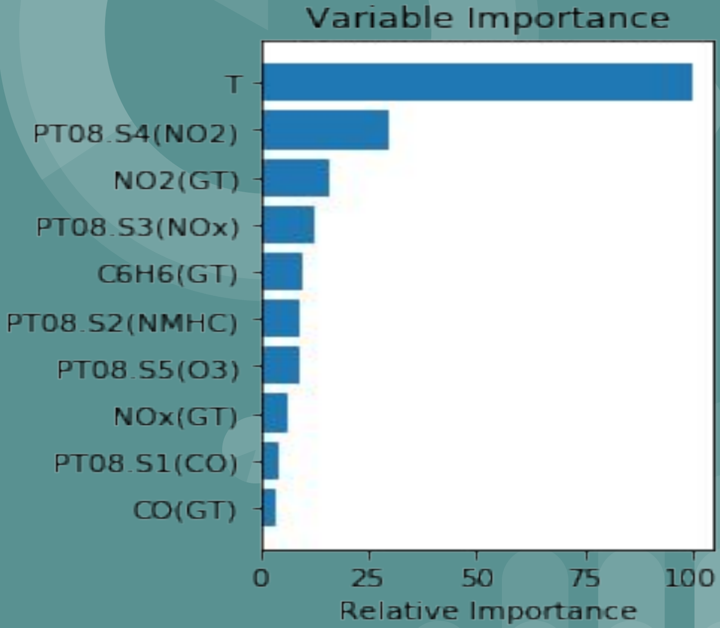


From the above graph, we can conclude that n_estimators = 180 gives the best r2_score.

Random Forest Regression

Importances of the feature columns:

#NO2(GT)	0.0787668829454153
#PT08.S2(NMHC)	0.04617378205324278
#PT08.S4(NO2)	0.1492551718750644З
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 Tests

- Bartlett's Test
- Kaiser-Meyer-Olkin Test

BARTLETT'S ADEQUACY 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.Hence we can proceed with Factor Analysis.

kaiser-Meyer-Olkin Test

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.

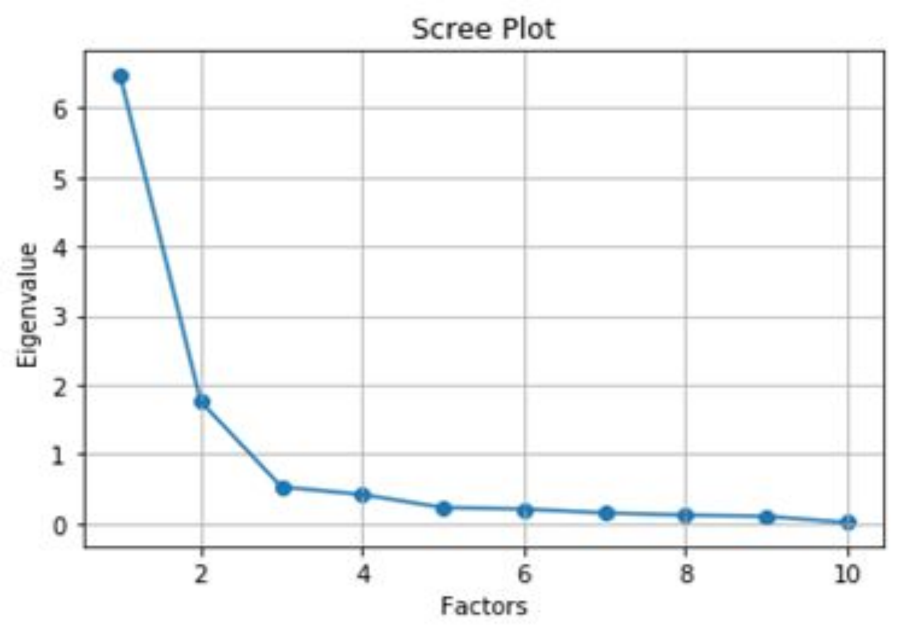
Result:

0.8652245188839246

The overall KMO for our data is 0.8652, which is considerably good for the above test.Hence,we can proceed with Factor Analysis.

Next ,we perform the Factor Analysis for the data.

Scree Plot



Factor Loadings of the Data

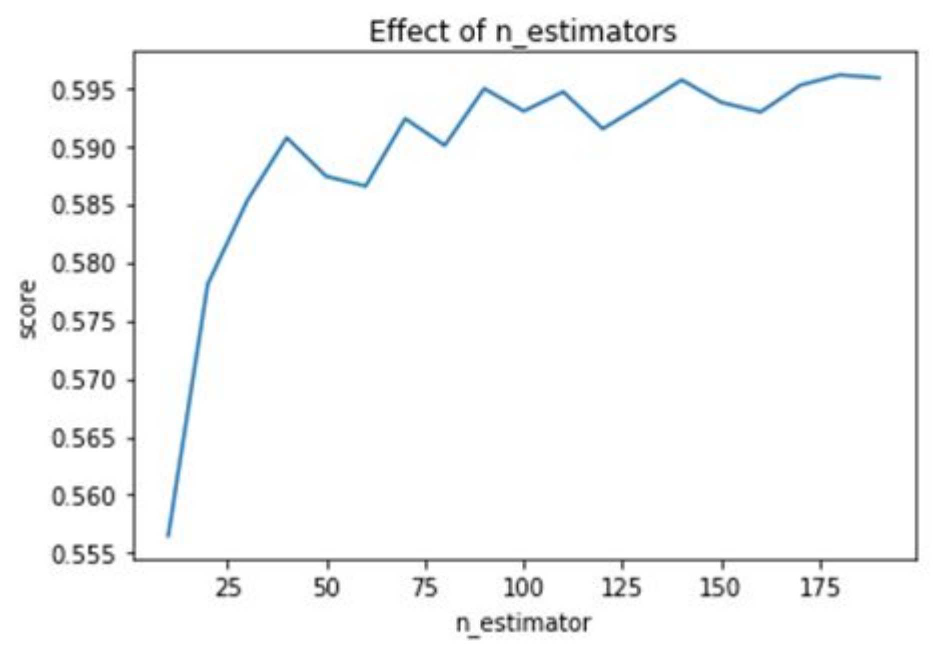
	Factor1	Factor2	Factor3	Factor4
NO2(GT)	0.245677	-0.102703	0.762700	0.212014
PT08.S2(NMHC)	0.867775	0.171821	0.421193	0.168409
PT08.S4(NO2)	0.770448	0.446725	0.043618	0.091084
PT08.S1(CO)	0.777560	-0.009076	0.426697	0.264350
NOx(GT)	0.290956	-0.154306	0.884593	0.034160
PT08.S5(O3)	0.661551	-0.065486	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

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:

REGRESSION AFTER APPLYING PCA


Effect of n_estimators
score
n_estimator