Statistical Data Analysis: Mainak Thakur **(Due: 05/12/20)**

Project Report

Name: PSVNB Shankar, PK Anand, S Hemanth, SVSK Rupesh

Roll: S20170010105, S20170020230, S20170010138, S20170020241 *Group:* 5

1 Introduction

The goal of this course project is to apply the concepts of Exploratory Data Analysis (EDA) and Modelling to practice on real-time datasets. The dataset that has been assigned to us is an Air Quality Dataset of an Italian city, this dataset is popularly known as Air Quality UCI dataset taken from UCI Machine Learning repository. EDA has been done to extract useful information and inference have been presented likewise. Linear Regression Model has been applied to predict concentration of the gases in the dataset. Lastly, few tests have been performed so as to measure the working of the model.

1.1 Dataset

The dataset is a multivariate time-series dataset that contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city. Data were recorded from March 2004 to February 2005 for a duration of one year. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons(NHMC), Benzene(C6H6), Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. Missing values are tagged with -200 value.

The detailed attribute information has been presented below:

- Date (DD/MM/YYYY)
- Time (HH.MM.SS)
- True hourly averaged concentration CO in mg/m^3 (reference analyzer)
- PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
- True hourly averaged overall Non Metanic HydroCarbons concentration in mg/m (reference analyzer)
- True hourly averaged Benzene concentration in mg/m^3 (reference analyzer)
- PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
- True hourly averaged NOx concentration in ppb (reference analyzer)
- PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
- True hourly averaged NO2 concentration in mg/m^3 (reference analyzer)
- PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)

- PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
- Temperature
- Relative Humidity
- Absolute Humidity

1.2 Approach and Work Presented

The dataset consists of 5 sensor readings, 5 Ground Truth concentration of 5 gases along with temperature, relative humidity and absolute humidity. Firstly, a detalied analysis of the both ground truth and sensor readings has been presented through different graphs which is also known as Exploratory Data Analysis. Since, regression is an associated task with the dataset, also it being a part of modelling, the concentration of the gases has been predicted using Linear Regression. Tests ranging from linearity, normality, multi-co linearity to homoscedacity tests and so on have been performed and presented. Finally, we also show that "" works well for this dataset showing better performance in prediction than linear regression.

2 Exploratory Data Analysis

EDA is an approach to extract useful information and conclusions from a data using visual methods. This not only gives us useful information but also an incentive and direction towards further analysis ie modeling. Before, we explore different kinds of visual representations of the data, we need to first clean the dataset. Later, different graphs along with their inferences has been presented. Data has been shown in Figure 1.

2.1 Data Wrangling

As the dataset description states that the missing values are substituted by the value -200, imputation is performed to replace those values as these -200 values will mess up any statistical measures and graphs/correlation.

Hence, the dummy value of -200 is replaced by NaN throughout the dataset.

After replacement, it has been found out that the NHMC(GT) column is missing many values, more than 85% of values are NaN. Thus this column has been removed from the dataset as these values are very less likely to be of any significant importance in this dataset.

In the remaining columns, the NaN values can be replaced by taking the mean or the median of the whole column, but this won't be an accurate and proper way to fill those values. Thus, filling the mean of that particular day in which day the value is NaN makes more sense and would be a proper way to impute the values. Hence, any NaN value has been replaced by the mean of the values of that day.

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
0	2004-03-10	18:00:00	2.6	1360.00	150	11.881723	1045.50	166.0	1056.25	113.0	1692.00	1267.50	13.60	48.875001	0.757754
1	2004-03-10	19:00:00	2.0	1292.25	112	9.397165	954.75	103.0	1173.75	92.0	1558.75	972.25	13.30	47.700000	0.725487
2	2004-03-10	20:00:00	2.2	1402.00	88	8.997817	939.25	131.0	1140.00	114.0	1554.50	1074.00	11.90	53.975000	0.750239
3	2004-03-10	21:00:00	2.2	1375.50	80	9.228796	948.25	172.0	1092.00	122.0	1583.75	1203.25	11.00	60.000000	0.786713
4	2004-03-10	22:00:00	1.6	1272.25	51	6.518224	835.50	131.0	1205.00	116.0	1490.00	1110.00	11.15	59.575001	0.788794

Figure 1: Data

	CO(GT)	PT08.S1(CO)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
count	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000
mean	2.082195	1102.604396	10.190299	942.422741	240.718147	832.618539	109.401453	1452.890358	1030.388426	18.315768	48.814853	1.017382
std	1.469801	219.599578	7.565011	269.583076	206.611257	255.704654	47.210774	347.427351	410.906048	8.822898	17.354492	0.404829
min	0.100000	647.250000	0.149048	383.250000	2.000000	322.000000	2.000000	551.000000	221.000000	-1.900000	9.175000	0.184679
25%	1.000000	938.250000	4.401596	732.500000	97.000000	654.500000	73.000000	1227.750000	726.000000	11.875000	35.425000	0.726213
50%	1.700000	1061.500000	8.276765	910.500000	174.000000	806.750000	102.000000	1459.750000	963.750000	17.575000	48.925001	0.987539
75%	2.800000	1237.250000	14.019301	1117.250000	318.000000	967.500000	137.000000	1676.750000	1286.500000	24.325000	61.875000	1.306671
max	11.900000	2039.750000	63.741476	2214.000000	1479.000000	2682.750000	339.700000	2775.000000	2522.750000	44.600000	88.725000	2.231036

Figure 2: Summary Statistics

2.2 Data Visualisation

The describe method is used to find the summary statistics for every column. Mean of values, standard deviation, minimum value, maximum value and 25%, 50% and 75% quartiles can be obtained. (Figure 2)

2.2.1 Trend of gases at different hours of the day

A rough trend of the gases throughout the different hours of the day has been shown in Figure 3 to 6, the hourly concentration of the gases has been shown via scatter plots. This hence gives an insight of how in a day concentration varies depending on which hour of the day it is. Along with the scatter plot, box plot showing summary statistics has also been shown. Outliers have also been detected in the data.

2.2.2 Density Heatmap

The given heat maps show the number of instances of a certain range of concentration in a particular month. These heat maps convey the general trend of the concentration of the gases over a period of time. This visual representation has been done for NOx and NO2 shown in Figure 7 and 8.

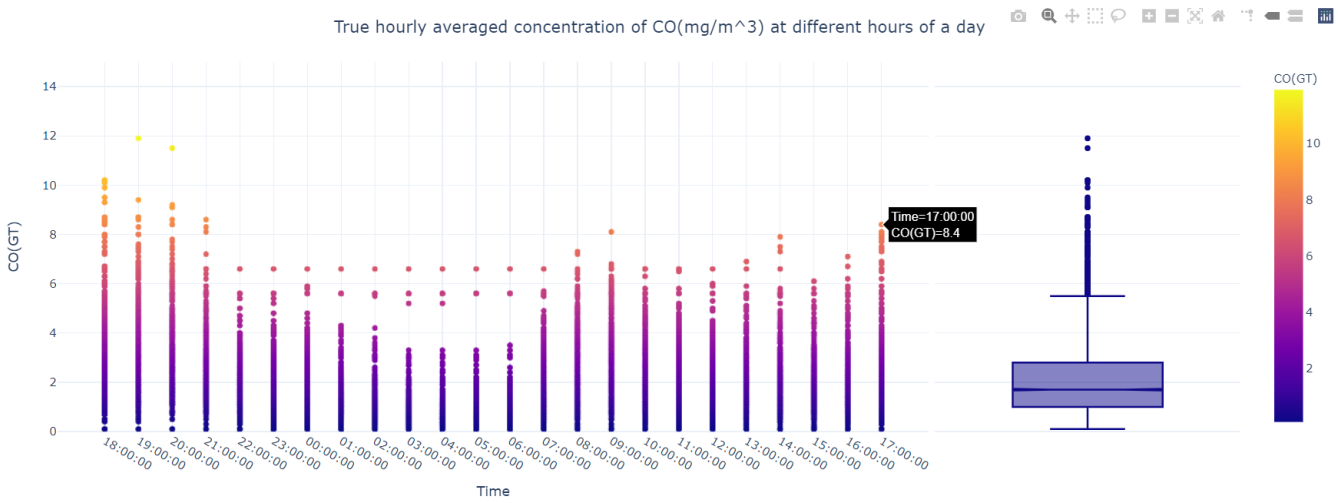

Figure 3

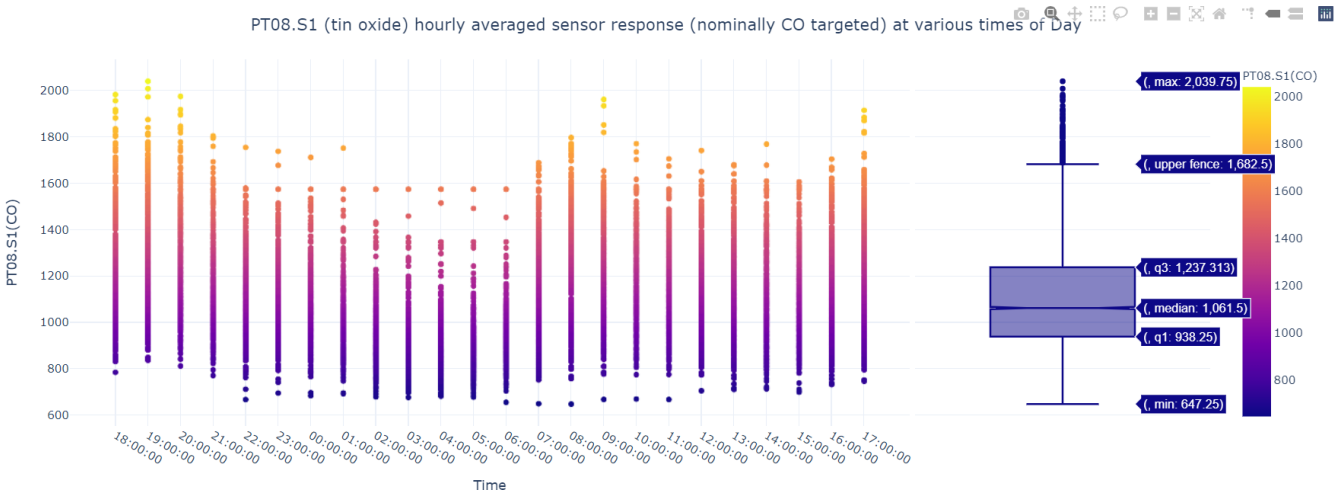

Figure 4

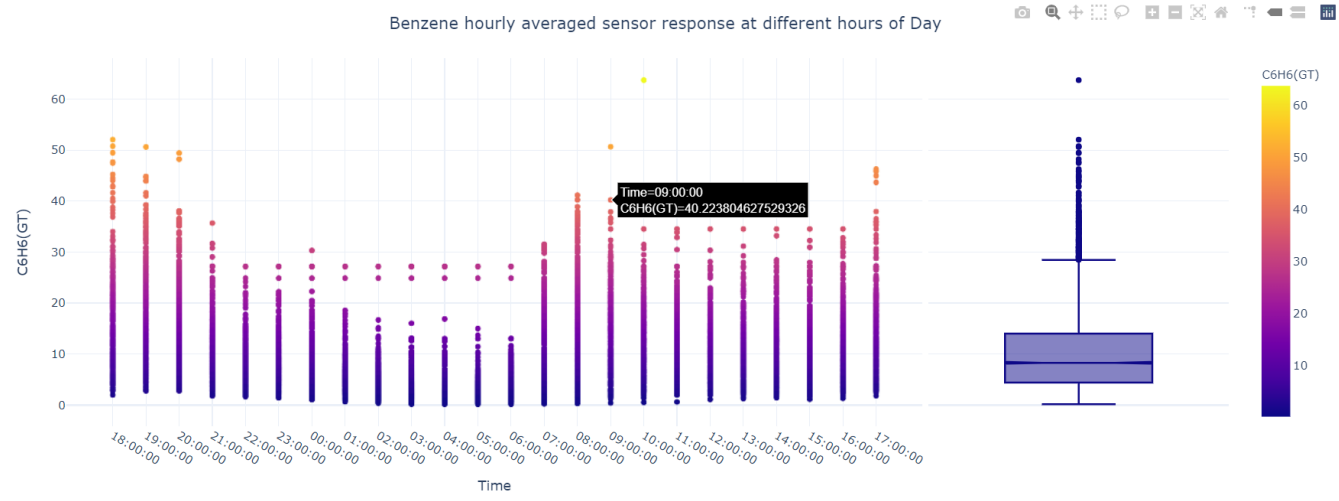

Figure 5

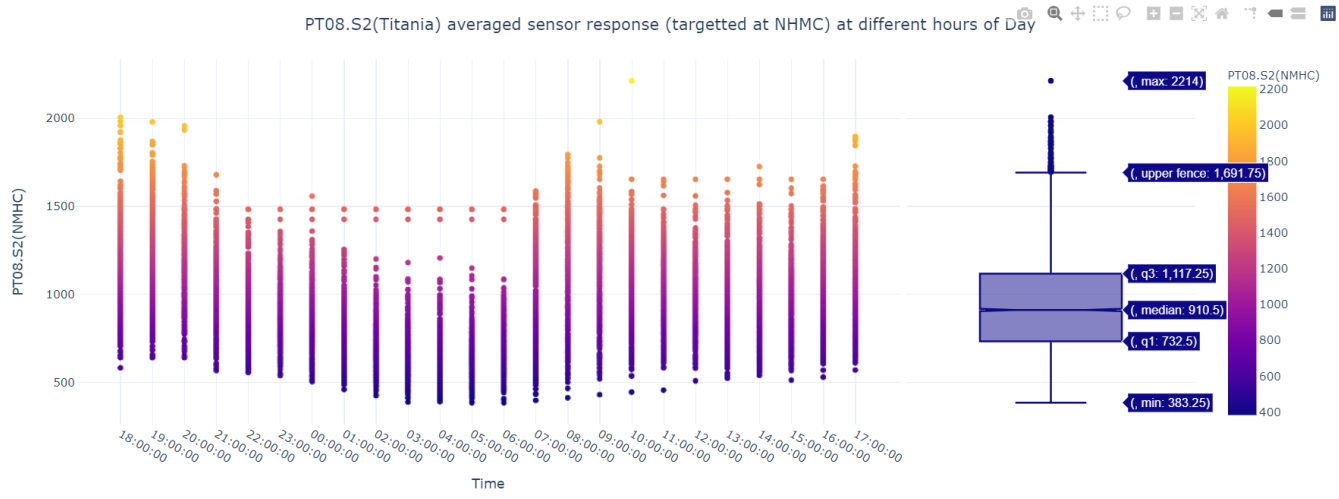

Figure 6

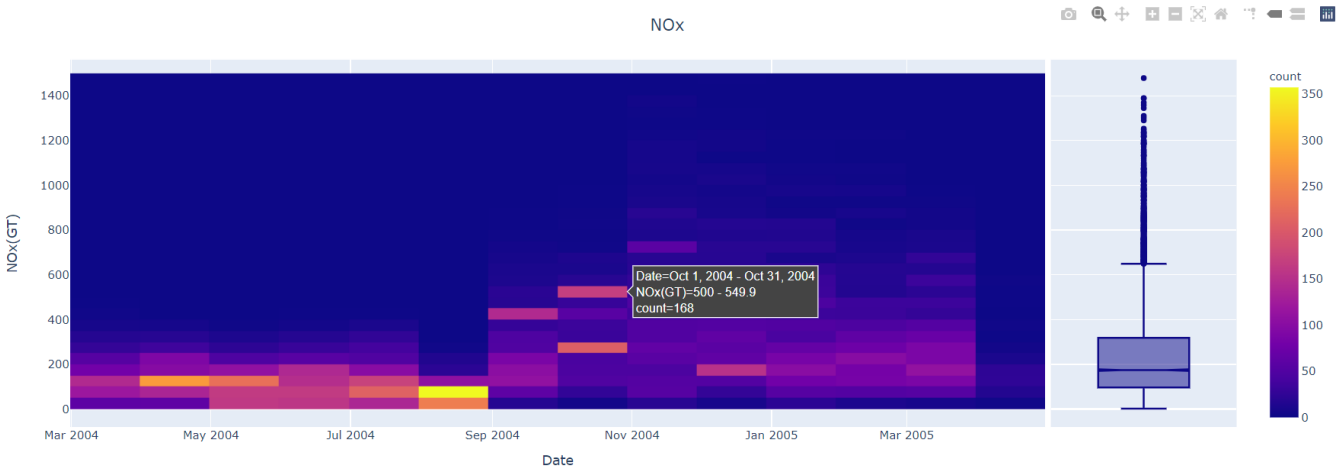

Figure 7: Density Heatmap for NOx gases

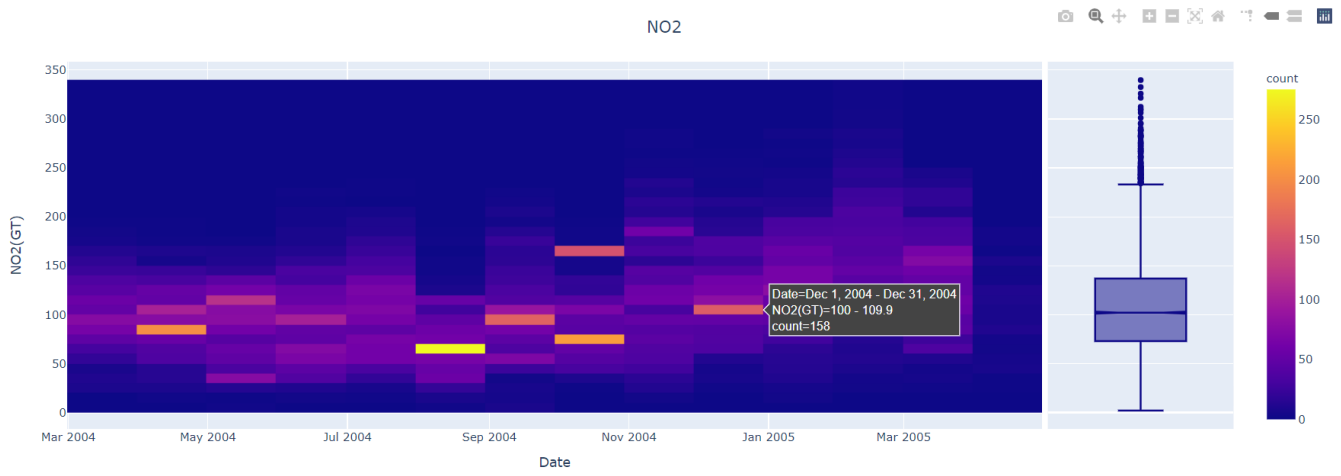

Figure 8: Density Heatmap for NO2

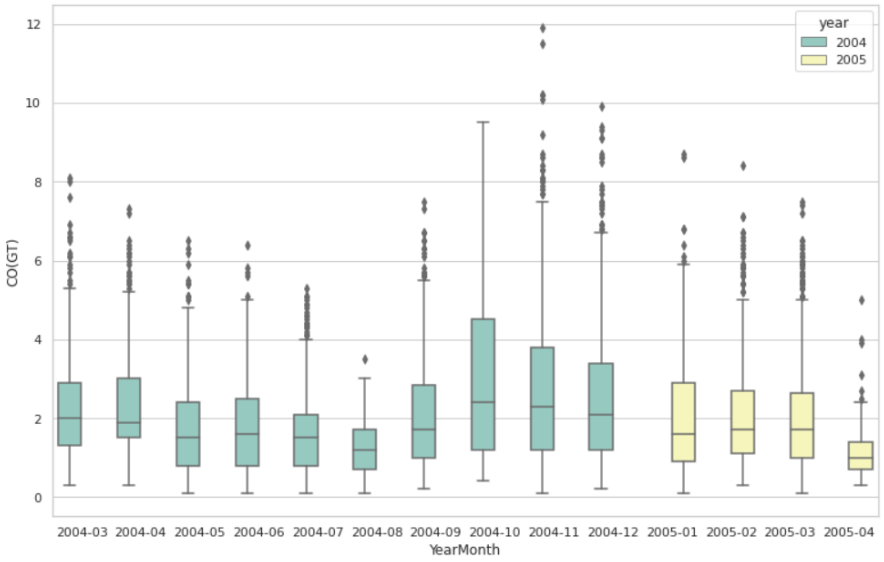

Figure 9: Carbon Monoxide (GT) vs YearMonth

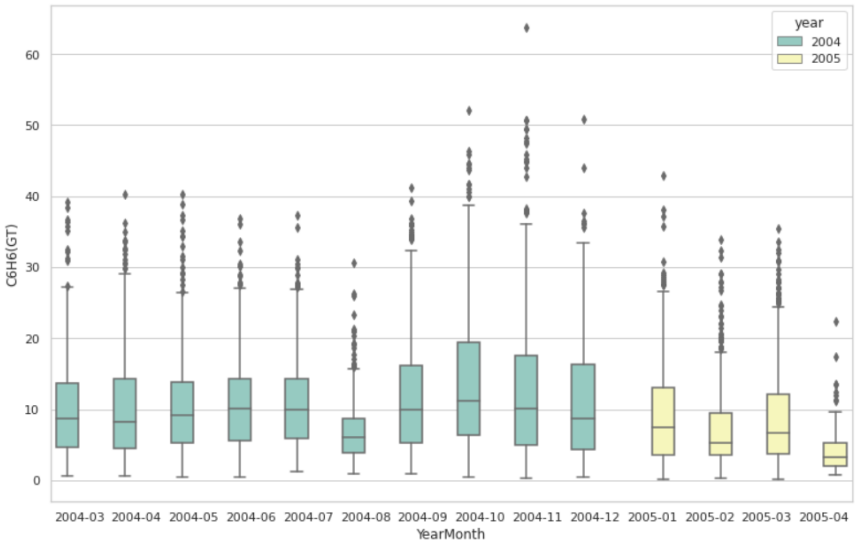

Figure 10: Benzene (GT) vs YearMonth

2.2.3 Box Plots

Boxplots show the visual representation of a five point summary, and also show the outliers, which are taken using the calculation < 1.5 times lower quartile and > 1.5 times the higher quartile. So the small stars above or below the boxes are outliers. This gives a quick insight into what are the potential row values or groups which have certain potential outliers, as they are significantly away from the quartile values(Lower and Higher). (Figure 9 to 16)

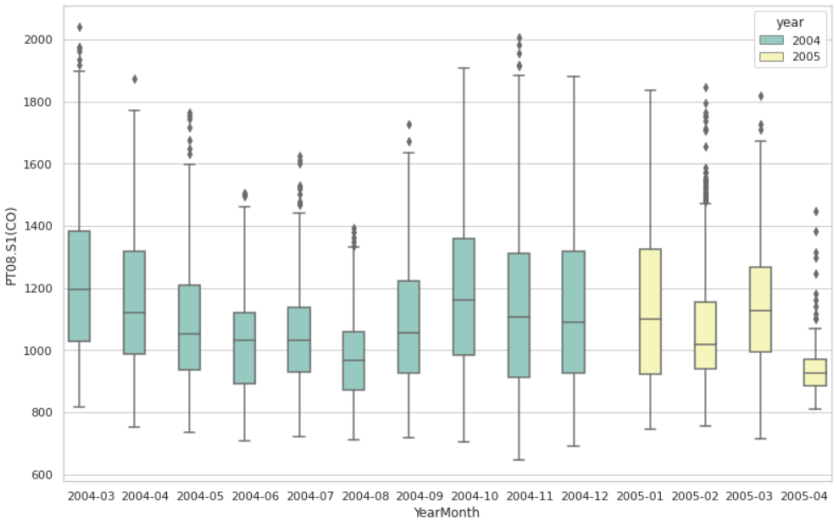

Figure 11: PT08.S1(CO) vs YearMonth

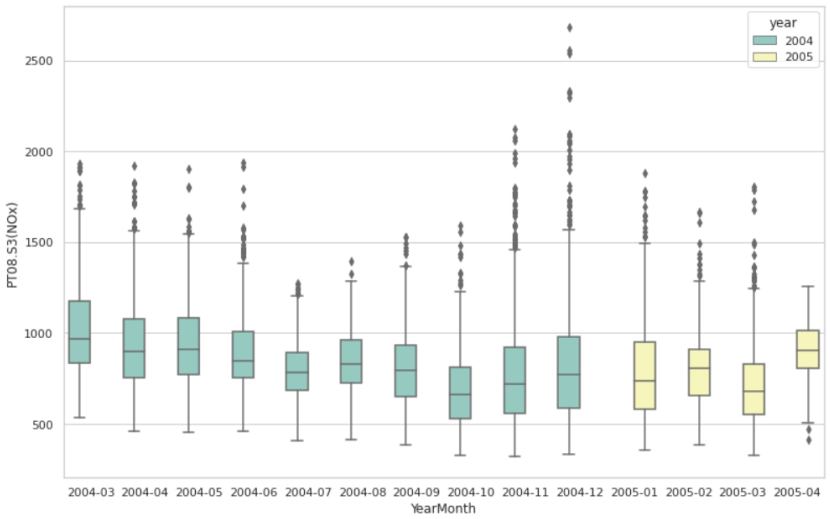

Figure 12: PT08.S3(NOx) vs YearMonth

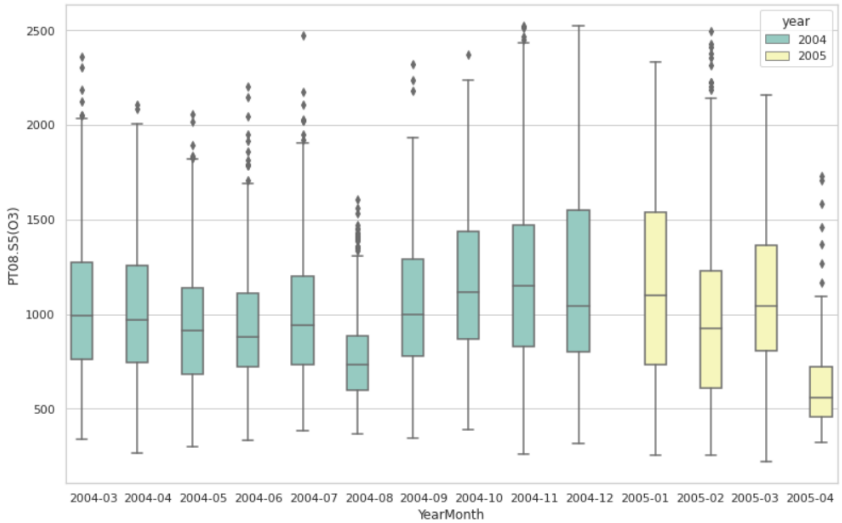

Figure 13: PT08.S5(O3) vs YearMonth

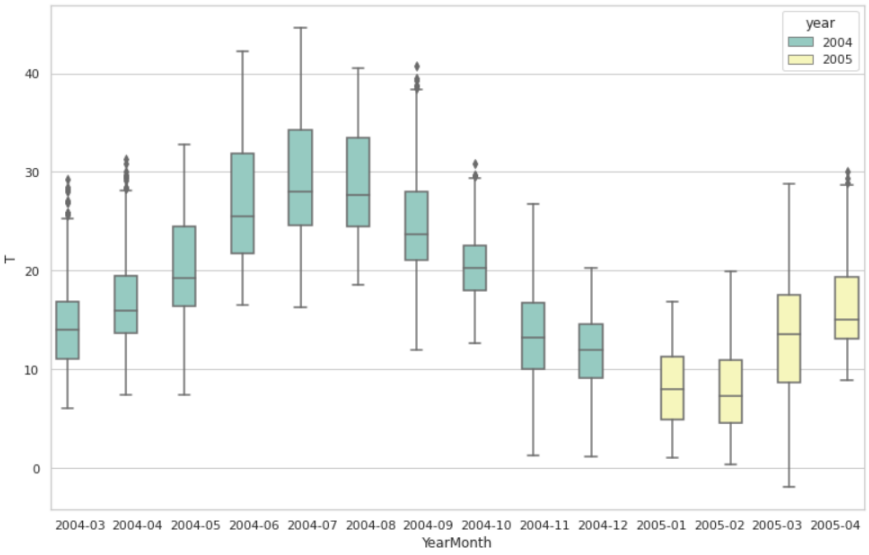

Figure 14: Temperature vs YearMonth

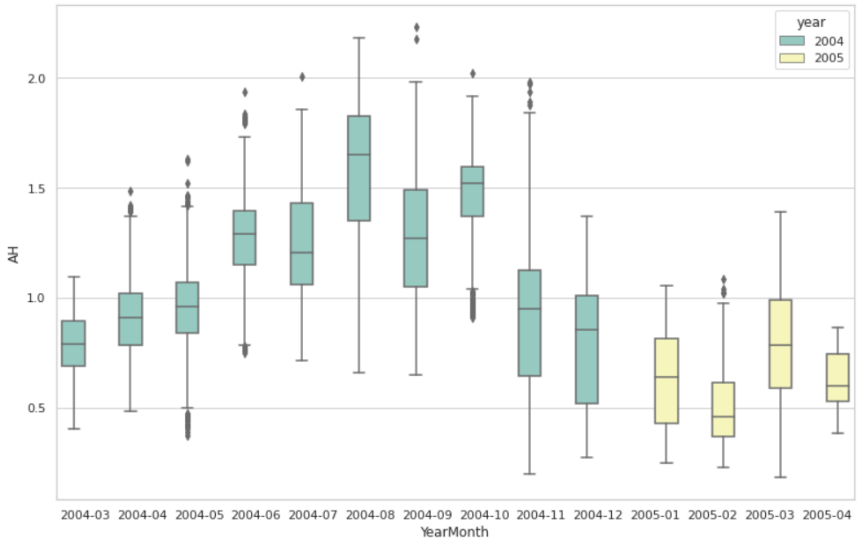

Figure 15: Absolute Humidity vs YearMonth

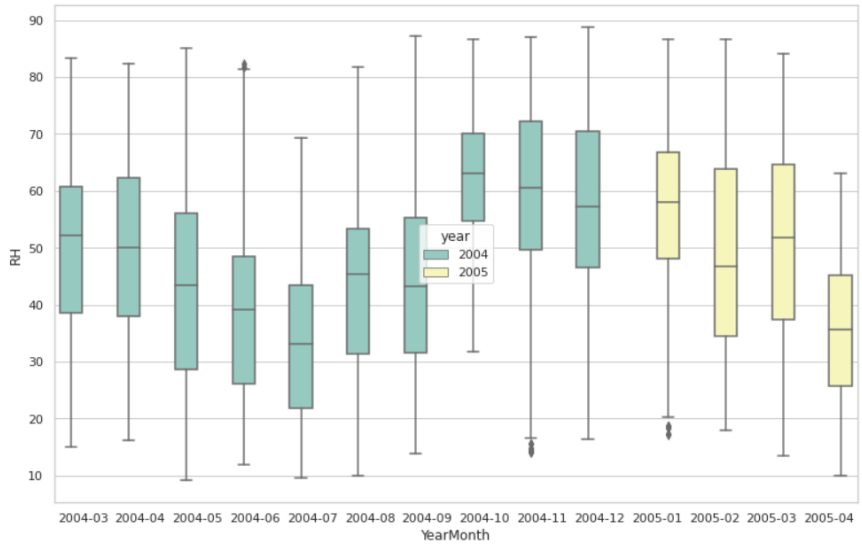

Figure 16: Relative Humidity vs YearMonth

2.2.4 Distribution

Distributions of columns have been obtained using distplot. Distplot provides a quick way to look at the univariate distribution. Distplots of some columns have been shown in Figure 17 to 22:

2.2.5 Correlation

From the figure (23) ie the correlation heatmap, it has been noticed that there is a very significant correlation between the gases among themselves. Also, NOx(GT) and NO2(GT) are also highly correlated as expected because the NO2(GT) is a subset value of the NOx(GT) values. The temperature(T), RH and AH values show a somewhat significant correlation around 0.60 amongst themselves but are very weakly correlated with the gas concentration attributes. Hence in a model approximating the gas concentrations, T, RH or AH values might not contribute well to the model estimation. The value of correlation in the correlation matrix above 0.6-0.7 is generally considered a decent correlation. The correlations greater than 0.8 are good, and the ones with values greater than 0.9 are very well correlated.

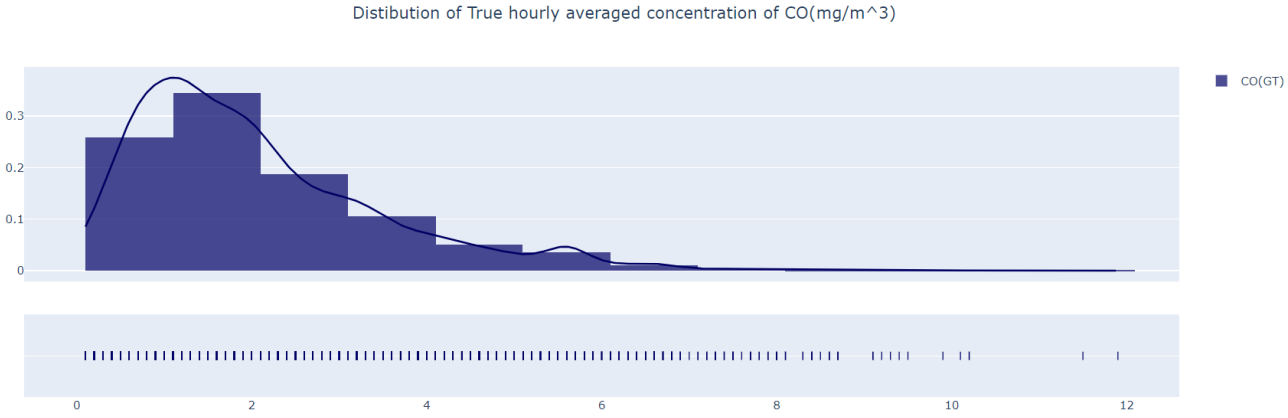

Figure 17

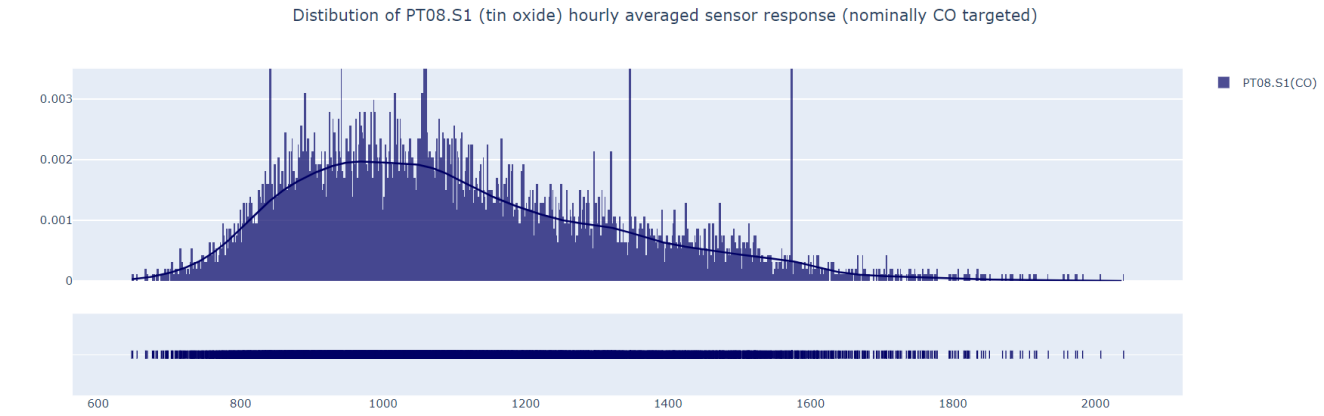

Figure 18

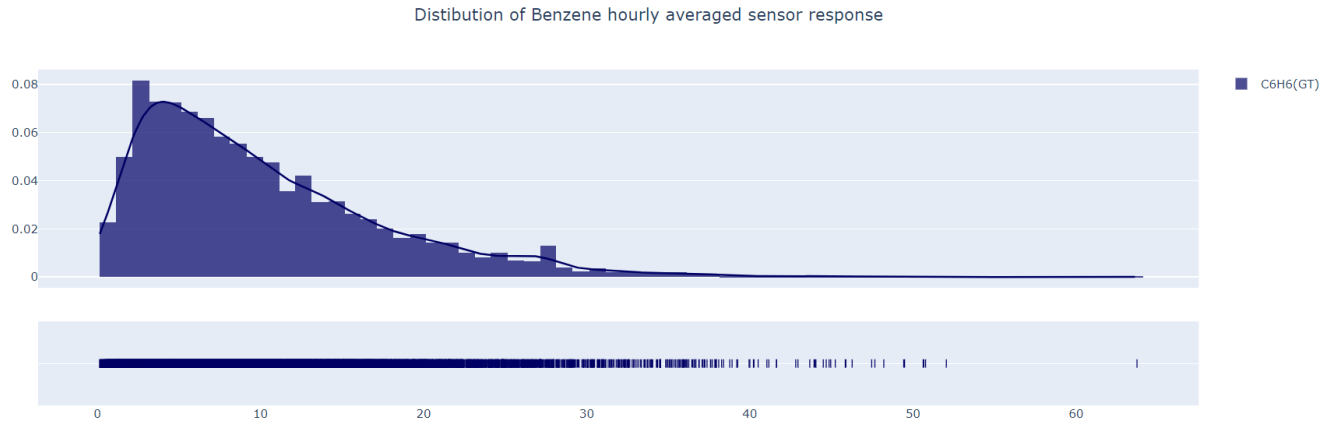

Figure 19

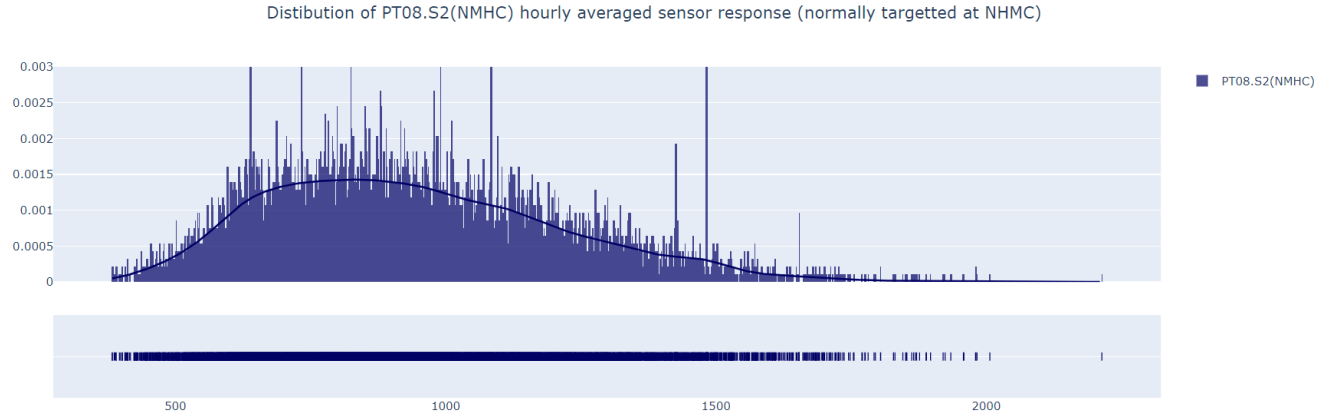

Figure 20

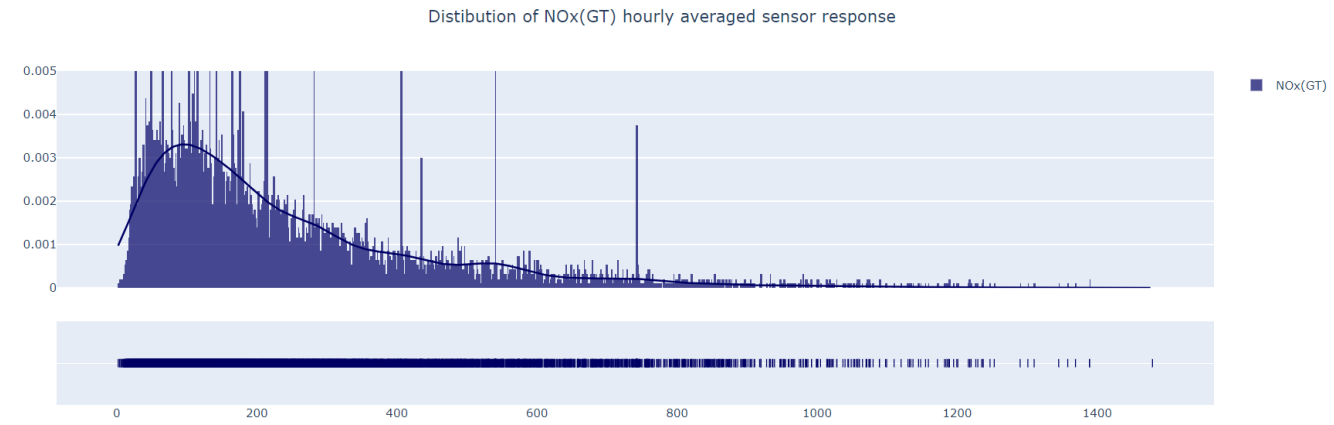

Figure 21

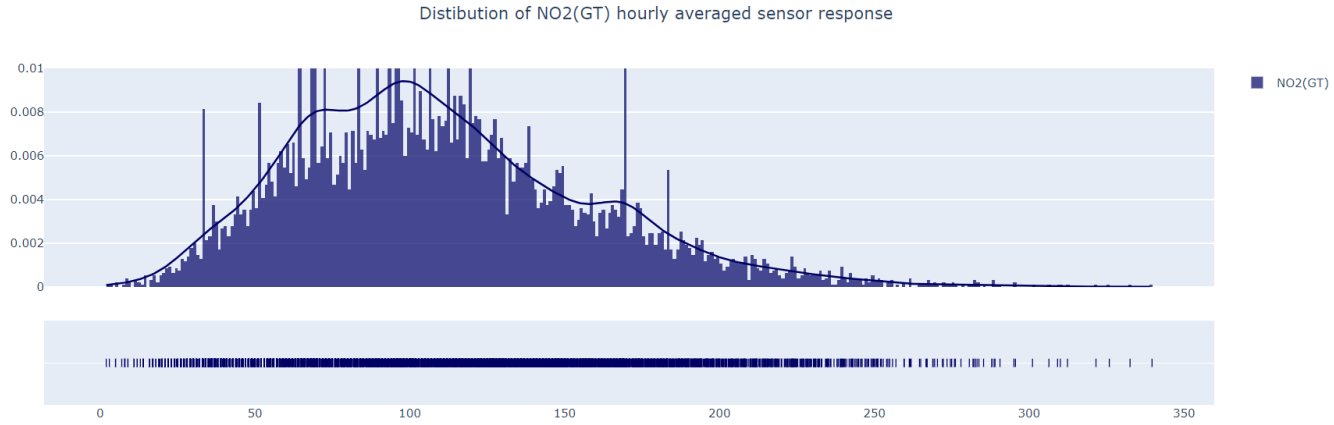

Figure 22

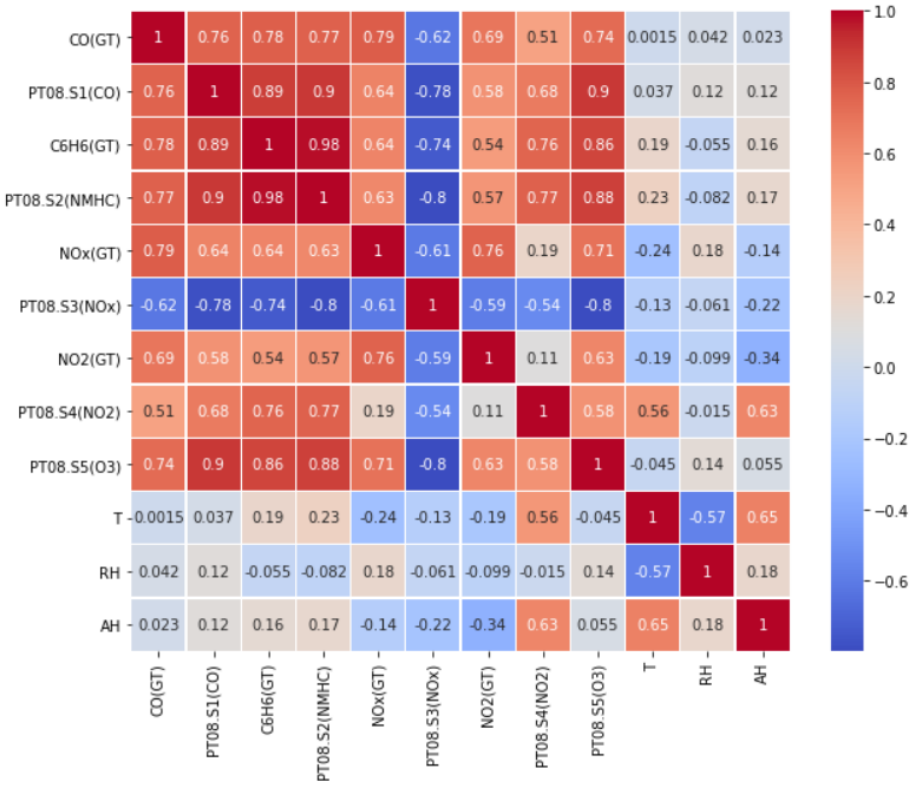

Figure 23: Correlation

2.2.6 Relative Humidity, Temperature and Absolute Humidity

Growth or Decline of Relative Humidity, Temperature and Absolute Humidity can be visualised using the plots below.

- From the figure 24, it can be inferred that there has been an increasing trend in relative humidity from July 2004 to October 2004, from where it started declining again. But, the relative humidity in the first part of 2005 is comparatively higher than that of in early 2004.
- Figure 25 says that there isn't any particular increasing or decreasing trend in the absolute humidity in 2004. However, mid 2004 had relatively higher AH than the early and late 2004 periods. Absolute Humidity in 2005 can be seen to be comparatively lesser than in 2004.
- Plots 26 to 28 showing trends of temperature, relative humidity and absolute humidity have also been shown.
- From figure 29 we can infer that with increase in temperature, there has been a decrease in relative humidity. It should be checked whether this phenomenon is indeed due to correlation.
- Joint distribution plots of temperature and relative humidity have also been shown.

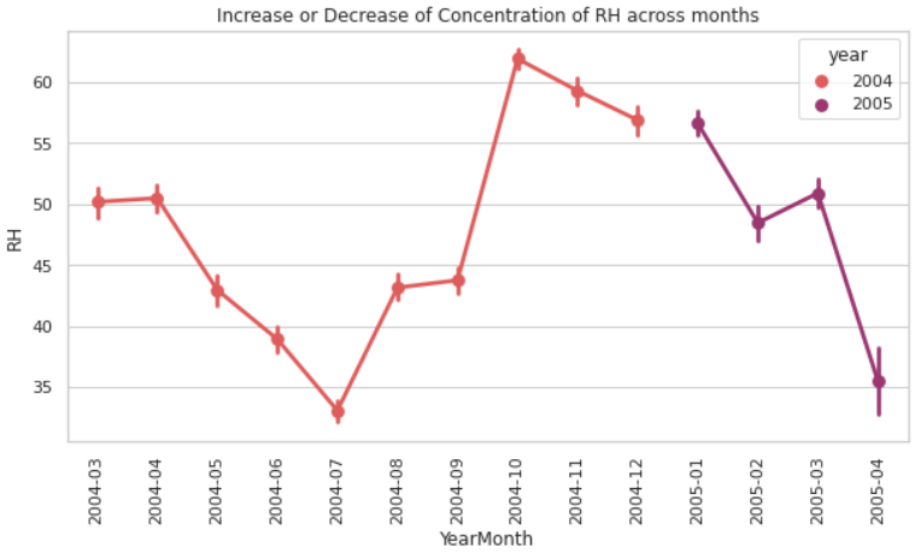

Figure 24

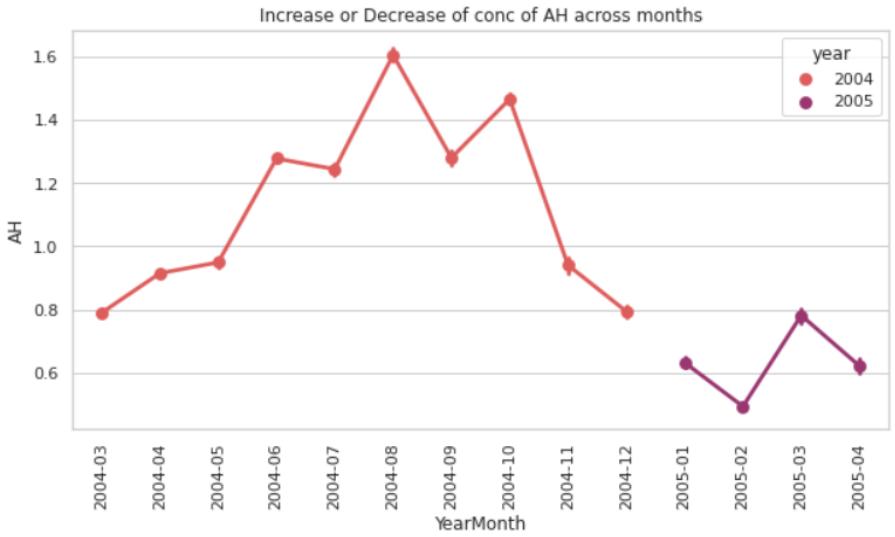

Figure 25

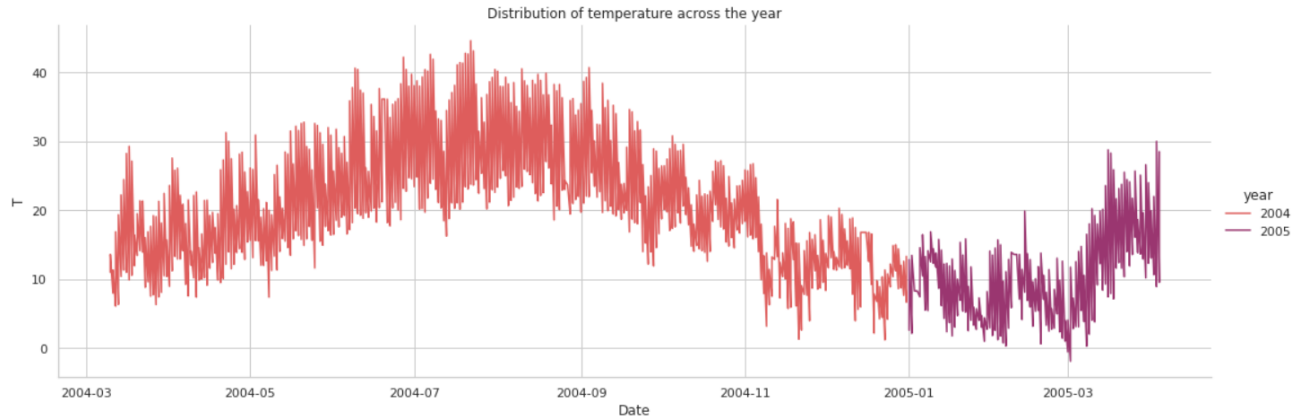

Figure 26

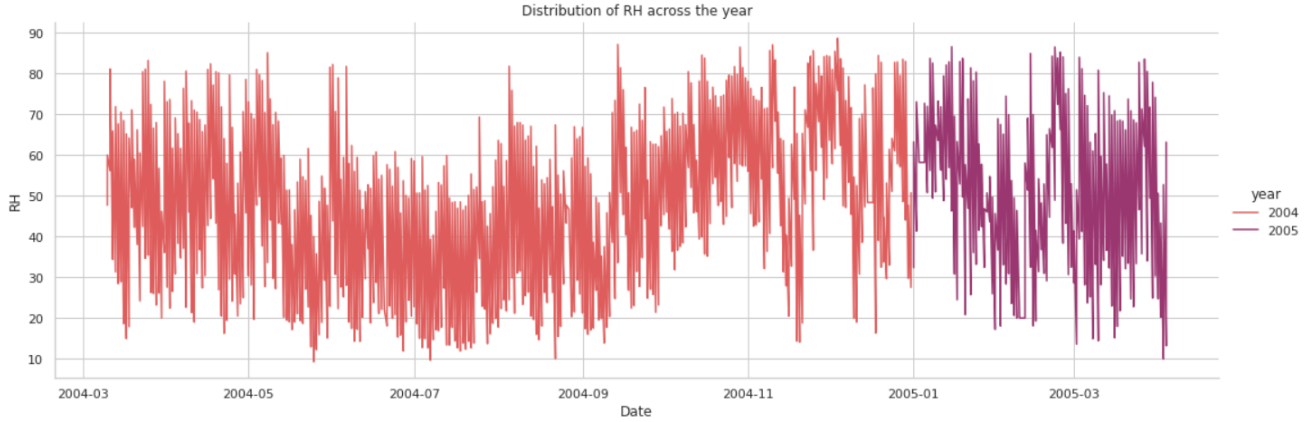

Figure 27

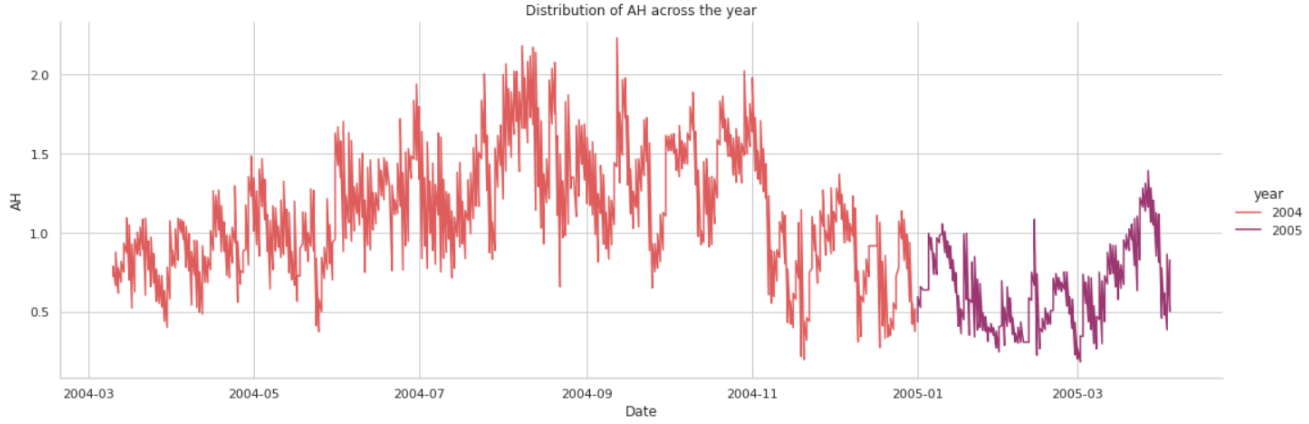

Figure 28

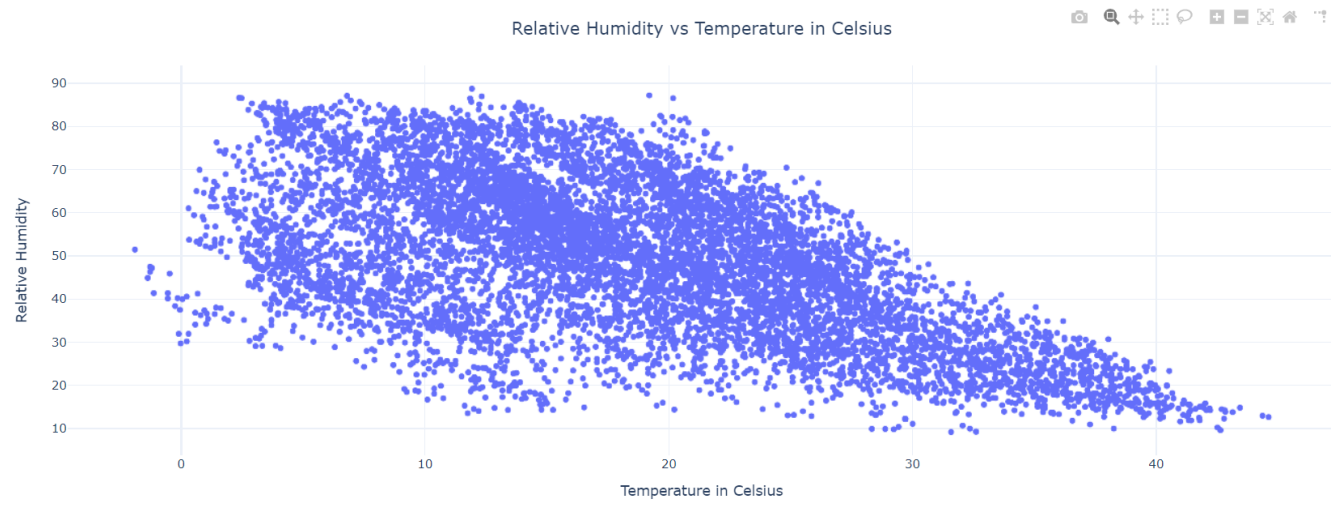

Figure 29

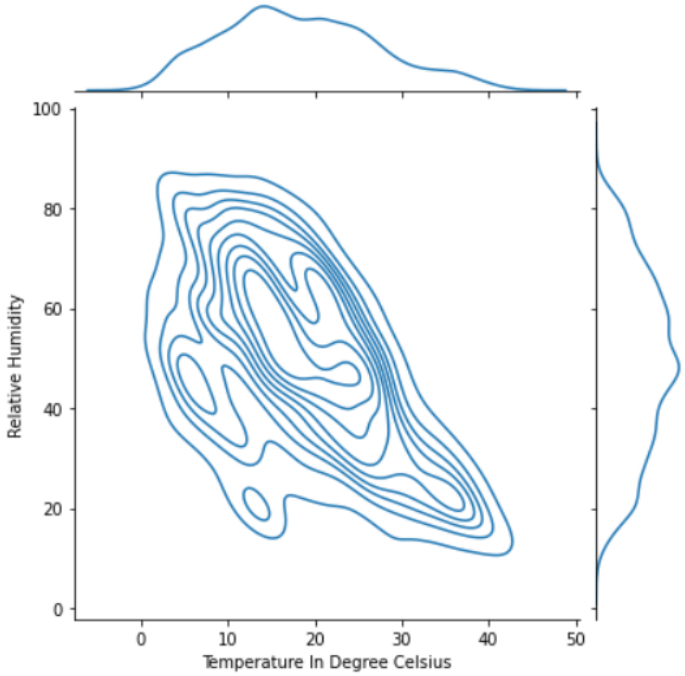

Figure 30

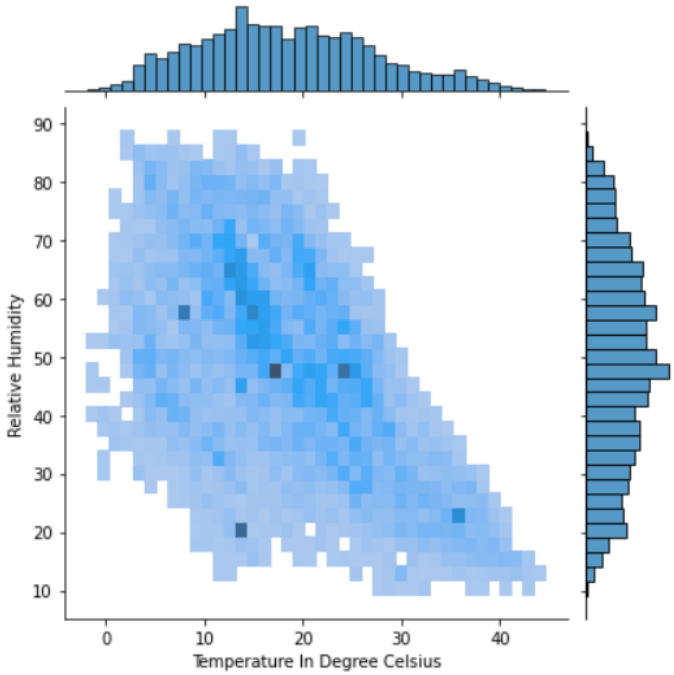

Figure 31

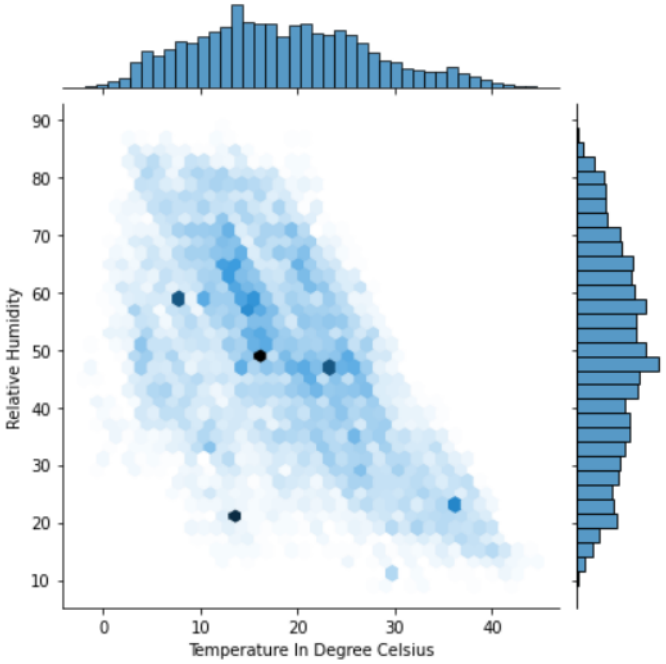

Figure 32

3 Modelling

3.1 Regression Analysis

3.1.1 Stationarity Check

In a weaker sense, a time-series $Y_t(t = 1, 2...)$ data is said to be stationary if its statistical properties do not vary with time (expectation, variance, autocorrelation). Stationarity tests allow verifying whether a series is stationary or not. Stationarity can be tested by plotting Rolling statistics or Dickey-Fuller Test.

Unit Root (Dickey-fuller) And Stationarity Tests On Time Series:

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 stationary or not.

Null hypothesis (H0) - Series possesses Unit root and hence is not stationary.
Alternate hypothesis (H1) - Does not possess Unit root and hence is stationary.

Results of Dickey-Fuller test have been shown in Table 1. From the results obtained, as the computed p-value is lower than the significance level $\alpha = 0.05$, we reject the null hypothesis H0 which implies that the series is stationary.

Data	P-value
RH	$1.219023e-10$
AH	0.000014
T	0.019787
CO(GT)	$5.412775e-16$
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)	$2.985511e-11$

Table 1

3.1.2 Normality Check Using Q-Q plots

The linear regression assumes that the response variable is normally distributed. This assumption can best be checked with a histogram or a Q-Q-Plot graphically or normal test. (Figure 33 to 35)

3.1.3 Multi Collinearity Test

Linear regression assumes that there is little or no multi collinearity in the data. Multi collinearity occurs when the independent variables are too highly correlated with each other. Multi collinearity may be tested based on various criteria.

- Correlation matrix – While computing the matrix of Pearson's Bivariate Correlation among all independent variables the correlation coefficients need to be smaller than 1. (Figure 36)
- Variance Inflation Factor (VIF) – the variance inflation factor of the linear regression is defined as $VIF = 1/T$. With $VIF > 5$ there is an indication that multicollinearity may be present; with $VIF > 10$ there is certainly multicollinearity among the variables.

Results:

Data	VIF
RH	7.67720
AH	9.96032
T	13.3785
PT08.S4(NO2)	8.91573
PT08.S5(O3)	4.98130
PT08.S2(NMHC)	14.0364
PT08.S3(NOx)	4.12303
PT08.S1(CO)	7.25182

From the figure 36, it can be clearly seen that there is multicollinearity among the variables. To mitigate this, either independent variables can be dropped with $VIF > 10$ or PCA can be used.

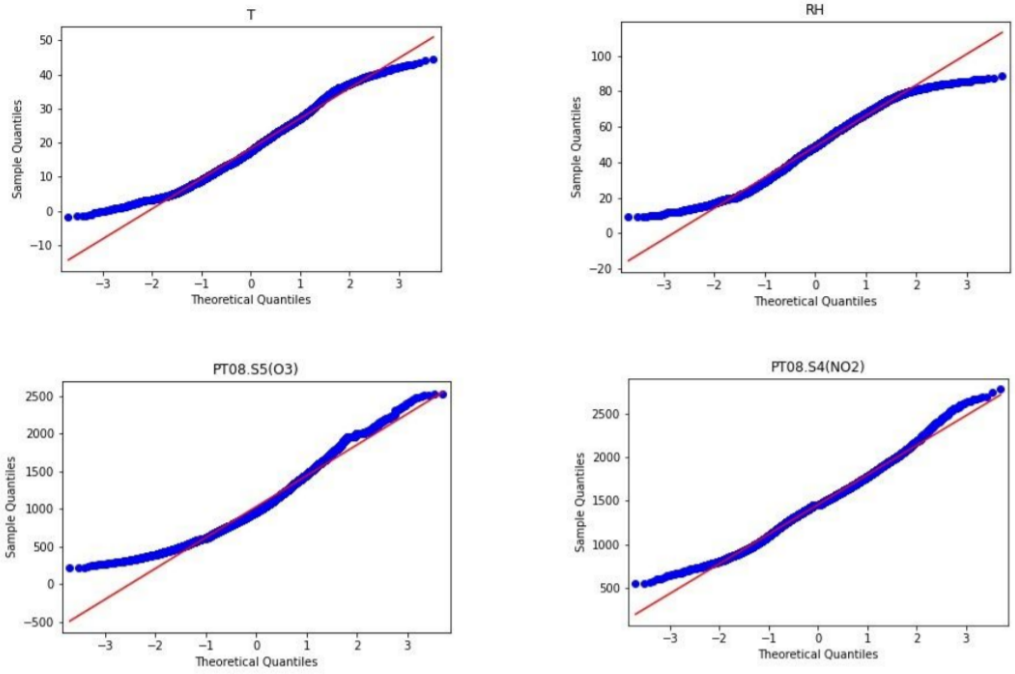

Figure 33

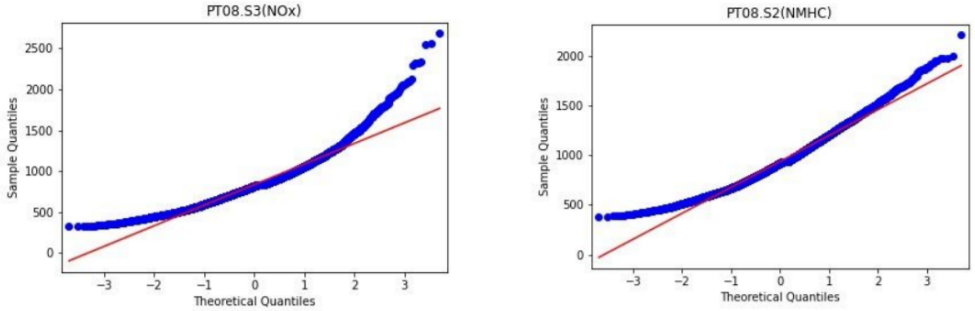

Figure 34

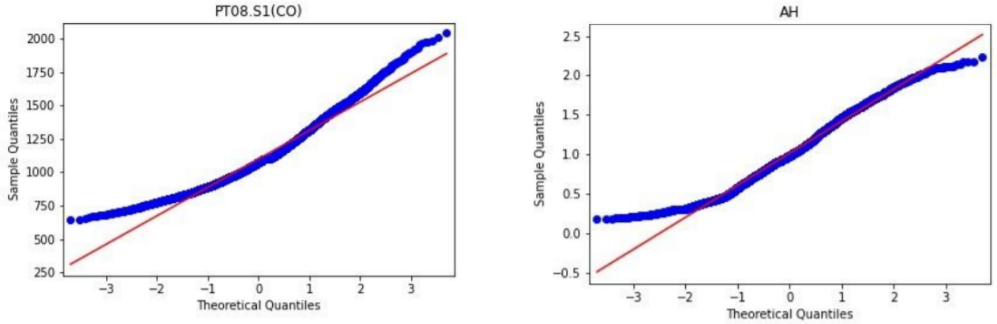

Figure 35

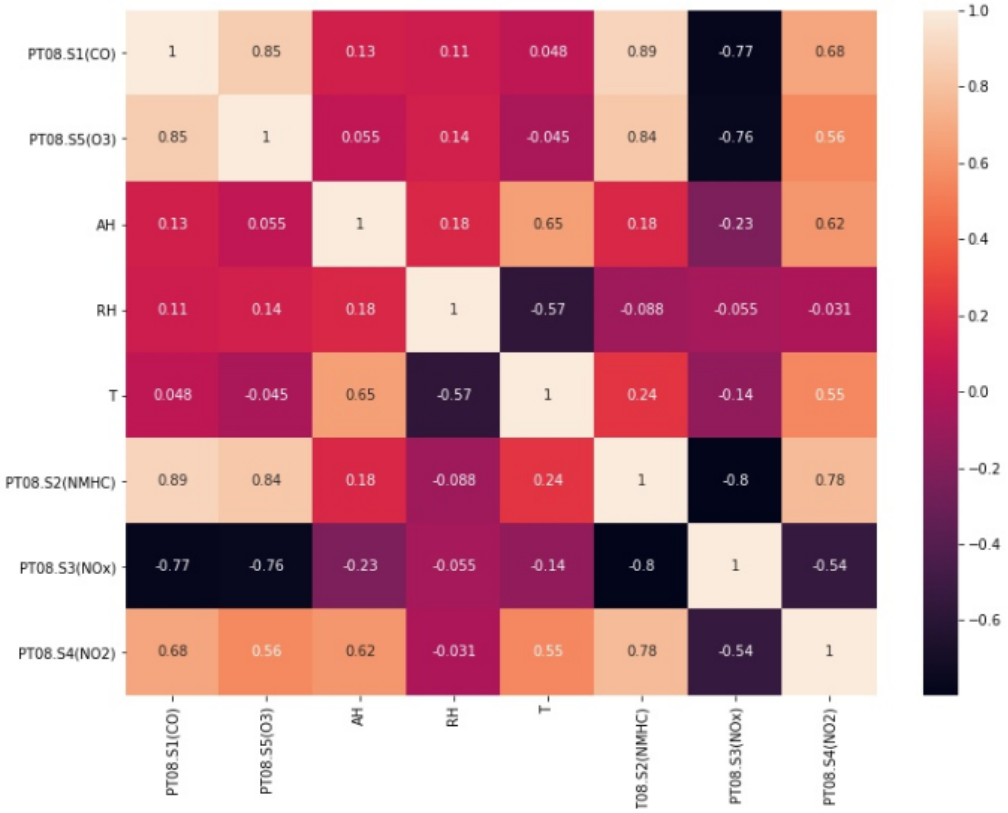
PT08.S1(CO)
PT08.S5(O3)
AH
RH
T
PT08.S2(NMHC)
PT08.S3(NOx)
PT08.S4(NO2)
1 0.85 0.13 0.11 0.048 0.89 -0.77 0.68
0.85 1 0.055 0.14 -0.045 0.84 -0.76 0.56
0.13 0.055 1 0.18 0.65 0.18 -0.23 0.62
0.11 0.14 0.18 1 -0.57 -0.088 -0.055 -0.031
0.048 -0.045 0.65 -0.57 1 0.24 -0.14 0.55
0.89 0.84 0.18 -0.088 0.24 1 -0.8 0.78
-0.77 -0.76 -0.23 -0.055 -0.14 -0.8 1 -0.54
0.68 0.56 0.62 -0.031 0.55 0.78 -0.54 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Figure 36

3.2 Regression Analysis with Ordinary Least Squares

The Data set consists of four ground truths, so the analysis and result have been done on four models (Gases). Initially, the data set has been split into Training and Testing set with a 90 - 10 split randomly and later scaled using MinMaxScaler.

Linear Regression using all Regressors:

Initially a regression model is fit without any tuning so as to compare with the results after tuning.

3.2.1 Auto-Correlation Test (OLS Regression)

Durbin Watson Statistic : The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals from a statistical regression analysis. The Durbin-Watson statistic will always have a value between 0 and 4. A value of 2.0 means that there is no autocorrelation detected in the sample. Values from 0 to less than 2 indicate positive autocorrelation and values from from 2 to 4 indicate negative autocorrelation.

- CO(GT) :
  $R^2 = 0.645$
  Durbin Watson Test = 2.005
- C6H6(GT):
  $R^2 = 0.983$
  Durbin Watson Test = 2.005
- NOx(GT):
  $R^2 = 0.672$
  Durbin Watson Test = 2.033
- NO2(GT):
  $R^2 = 0.604$
  Durbin Watson Test = 2.016

From the above results, it can be concluded that there is very little auto-correlation, as the Durbin-Watson test values are around 2.

3.2.2 Homoscedasticity Test:

The last assumption of the linear regression analysis is homoscedasticity. The residuals plot is a good way to check for homoscedasticity. (Figure 37 to 38)

From the plots, it can be said that there is a funnel-shaped structure that suggests heteroscedasticity. This may be mitigated by transforming dependent variables (Box-Cox, yeo-Johnson).

Results :

- Model1 - CO(GT) :
  $R^2 = 0.645$
  Train MSE = 0.760
  Test MSE = 0.7614

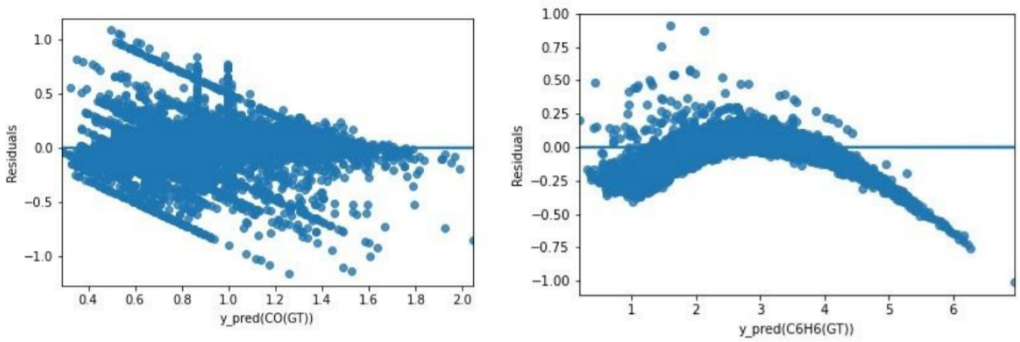

Figure 37

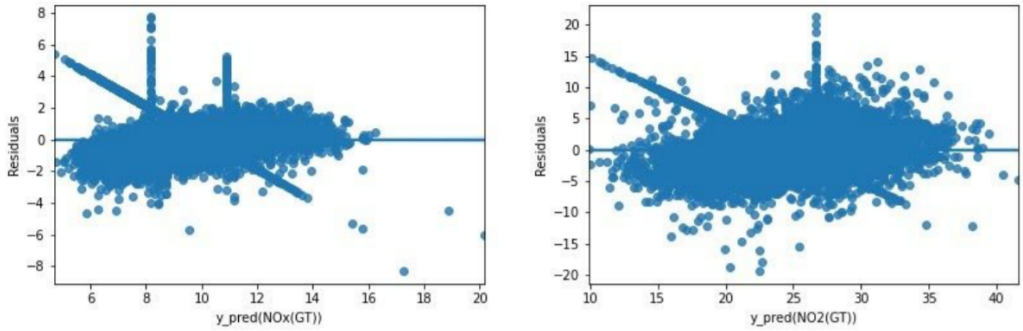

Figure 38

- Model2 - C6H6(GT):
  $R^2 = 0.974$
  Train MSE = 1.376
  Test MSE = 1.613
- Model3 - NOx(GT):
  $R^2 = 0.642$
  Train MSE = 13425.64
  Test MSE = 12302.075
- Model4 - NO2(GT):
  $R^2 = 0.598$
  Train MSE = 775.84
  Test MSE = 756.18

3.2.3 Interpretation of regression coefficients using p-value:

H0 = coefficient is equal to zero (no effect)
H1 = coefficient not equal to zero

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that null hypothesis can be rejected. In other words, a predictor that has a low p-value is likely to be a meaningful addition to the model because changes in the predictor's value are related to changes in the response variable.

Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

OLS Regression Results

Dep. Variable:	CO(GT)	R-squared:	0.645
Model:	OLS	Adj. R-squared:	0.645
Method:	Least Squares	F-statistic:	1911.
Date:	Sat, 05 Dec 2020	Prob (F-statistic):	0.00
Time:	13:14:46	Log-Likelihood:	-10795.
No. Observations:	8421	AIC:	2.161e+04
Df Residuals:	8412	BIC:	2.167e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.7852	0.112	-7.013	0.000	-1.005	-0.566
PT08.S1(CO)	2.4343	0.167	14.555	0.000	2.106	2.762
PT08.S5(O3)	1.4090	0.118	11.911	0.000	1.177	1.641
AH	0.3092	0.151	2.043	0.041	0.012	0.606
RH	-0.0908	0.121	-0.753	0.452	-0.327	0.146
T	-0.7307	0.183	-3.995	0.000	-1.089	-0.372
PT08.S2(NMHC)	6.3651	0.249	25.566	0.000	5.877	6.853
PT08.S3(NOx)	1.8317	0.181	10.123	0.000	1.477	2.186
PT08.S4(NO2)	-1.2628	0.185	-6.817	0.000	-1.626	-0.900

Omnibus:	3006.655	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	39508.012
Skew:	1.338	Prob(JB):	0.00
Kurtosis:	13.268	Cond. No.	48.8

Figure 39

- Here it can be noticed that p-values are < 0.05 resulting in rejecting the null hypothesis indicating that coefficients are not equal to zero.
- Similarly, the process is done for the rest of the models resulting in dropping of some of the independent variables.

Final Set of Independent Variables per model:

Model1 (CO)	'PT08.S2(NMHC)', 'AH', 'PT08.S5(O3)', 'PT08.S1(CO)', 'PT08.S3(NOx)', 'T', 'PT08.S4(NO2)'
Model 2 (C6H6)	'PT08.S2(NMHC)', 'AH', 'PT08.S1(CO)', 'PT08.S3(NOx)', 'T', 'PT08.S4(NO2)'
Model 3 (NOx)	'PT08.S2(NMHC)', 'PT08.S5(O3)', 'PT08.S3(NOx)', 'AH', 'RH', 'PT08.S4(NO2)'
Model 4 (NO2)	'PT08.S2(NMHC)', 'AH', 'PT08.S5(O3)', 'PT08.S1(CO)', 'PT08.S3(NOx)', 'RH', 'PT08.S4(NO2)'

3.2.4 Yeo-Johnson Transform

Applying Yeo-Johnson Transform on Dependent variable, following are the findings. (Figure 40 to 44)

Results :

- Model1 - CO(GT) :
  $R^2 = 0.645$
  Train MSE = 0.816
  Test MSE = 0.767
- Model2 - C6H6(GT):
  $R^2 = 0.988$
  Train MSE = 2.713
  Test MSE = 3.556

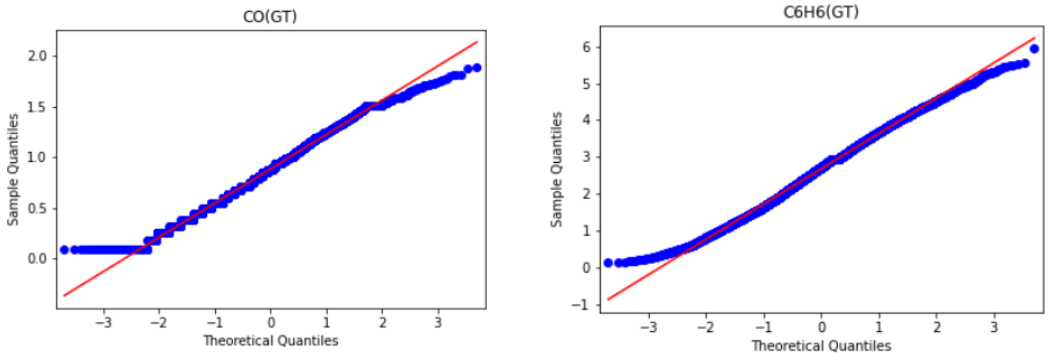

Figure 40: Q-Q plots of dependent variables after transformation

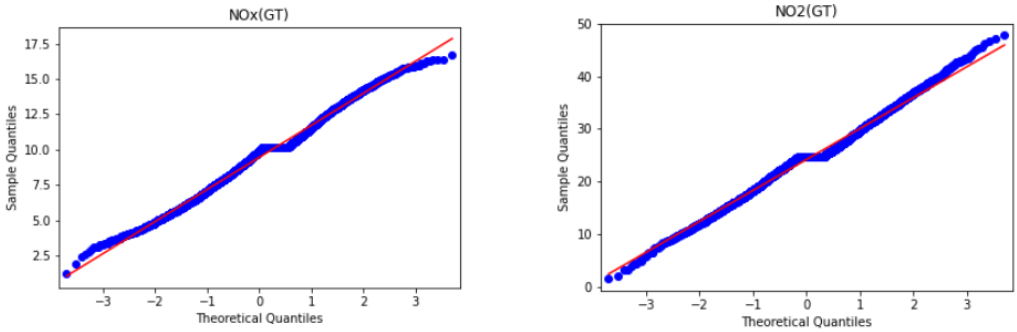

Figure 41: Q-Q plots of dependent variables after transformation

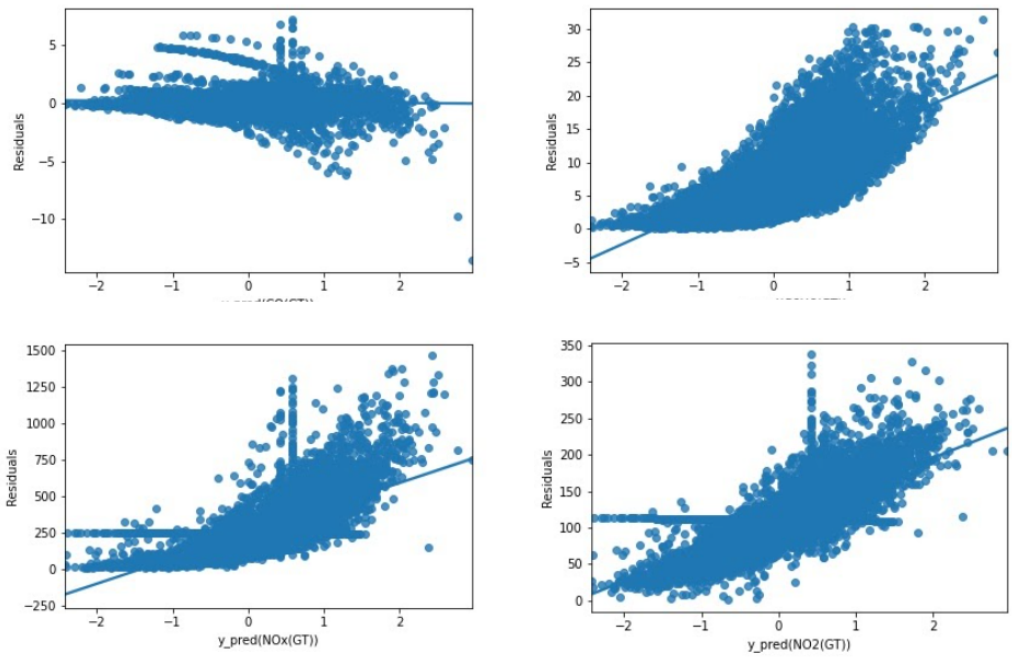

Figure 42: Residuals vs fitted values

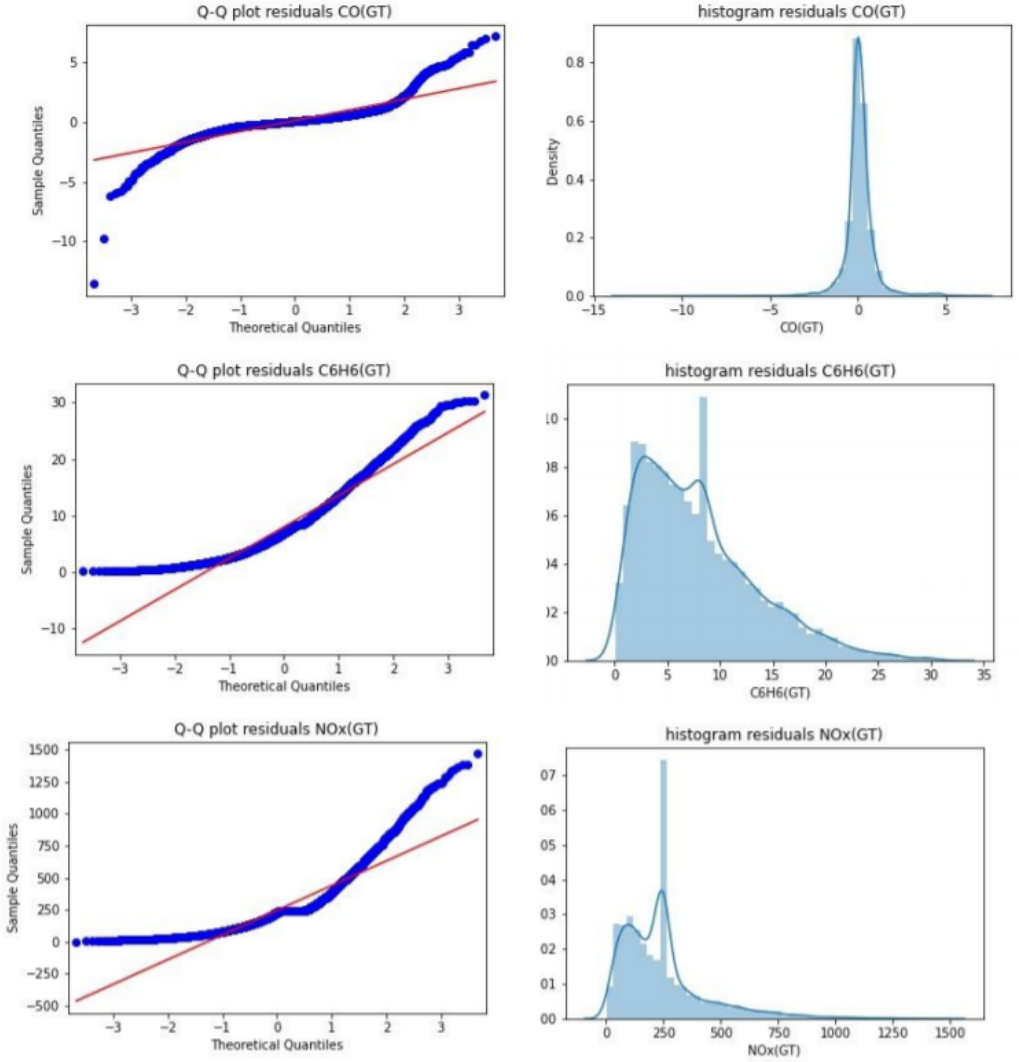

Figure 43: Q-Q and histogram plots of residuals

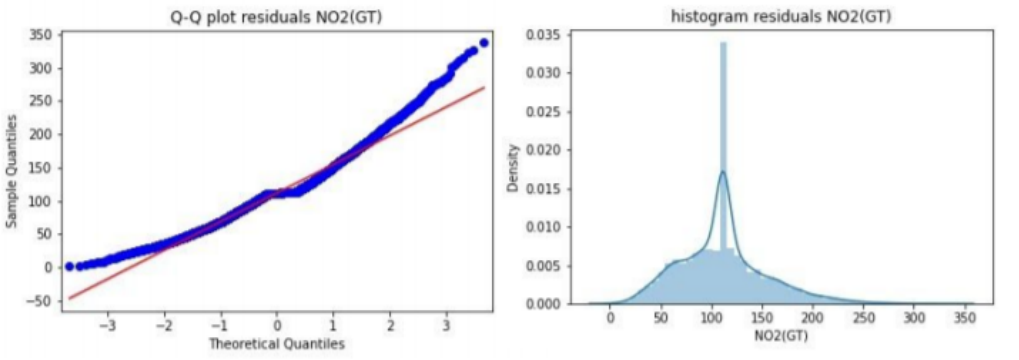

Figure 44: Q-Q and histogram plots of residuals

- Model3 - NOx(GT):
  $R^2 = 0.642$
  Train MSE = 12090
  Test MSE = 9644
- Model4 - NO2(GT):
  $R^2 = 0.598$
  Train MSE = 750
  Test MSE = 730

Conclusion: Even after transforming the dependent variables there is still evidence that Heteroscedasticity exists with an increased R-sqaured value . Different transformations such as log , squared root , Box-cox , yeo-johnson were applied but Heteroscedastic nature of the data could not be removed completely. Hence, further analysis has been done without any tranformations. Another observation is that there is no significant change in R-value or MSE after dropping the insignificant independent variables.

3.2.5 Outlier Analysis

Dealing with Outliers: In regression analysis, an outlier is an observation for which the residual is large in magnitude compared to other observations in the data set. The detection of outliers and influential points is an important step because they can have a strong influence on the least squares line. Cook's distance can be used to identify the outliers in the similar way Leverage Plot can be used to identify the leverage points. (Figure 45 to 46).

Results:

- Model1 - CO(GT) :
  $R^2 = 0.657$
  Train MSE = 0.663
  Test MSE = 0.706
- Model2 - C6H6(GT):
  $R^2 = 0.979$
  Train MSE = 0.982
  Test MSE = 1.703
- Model3 - NOx(GT):
  $R^2 = 0.703$
  Train MSE = 9458.286
  Test MSE = 12281.561
- Model4 - NO2(GT):
  $R^2 = 0.611$
  Train MSE = 736.799
  Test MSE = 761.346

Conclusion: It is evident from the results that there is an increase in the R-squared value in all of the models after the removal of outliers and also slight decrease in MSE values. Better model has been achieved when outliers were removed than in the earlier case where transformation was applied to independent variables.

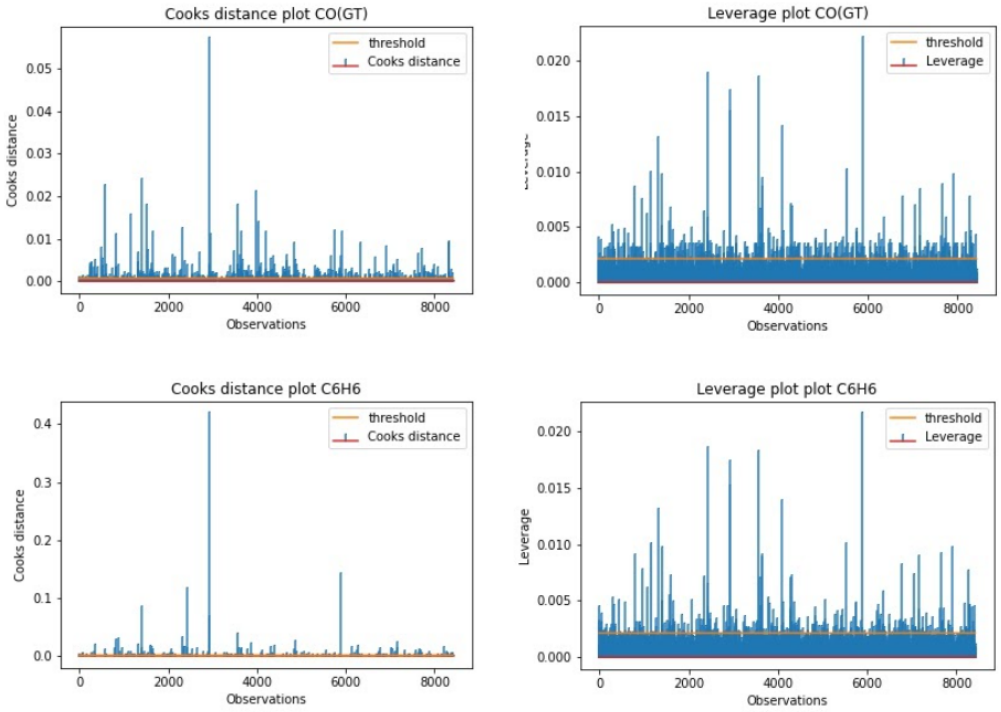

Figure 45: Cooks distance and leverage points

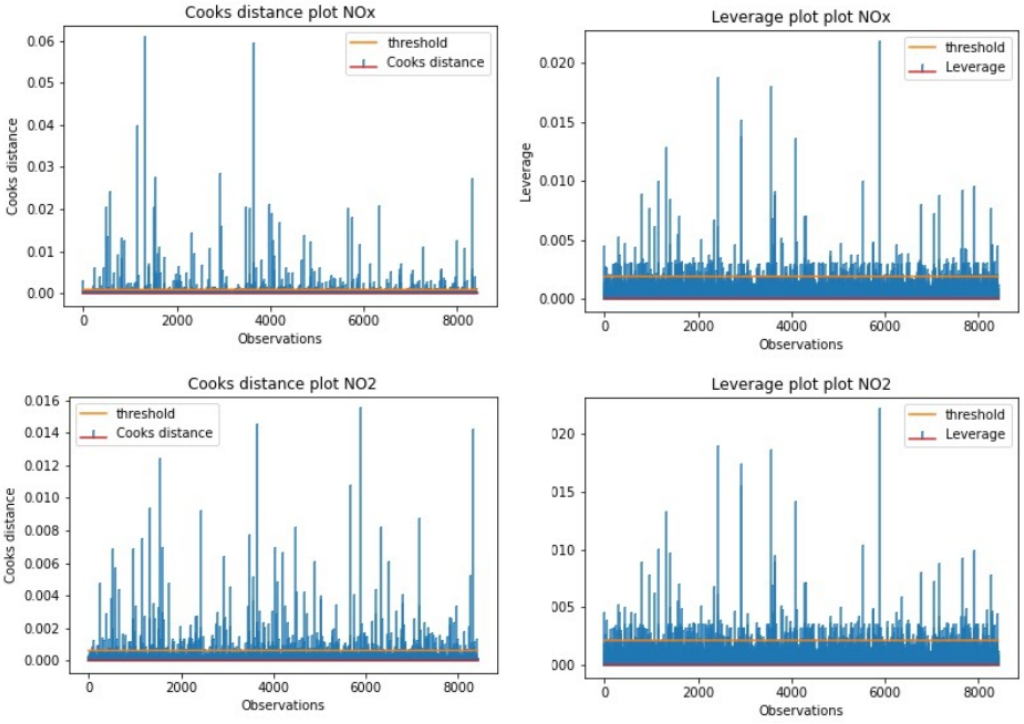

Figure 46: Cooks distance and leverage points

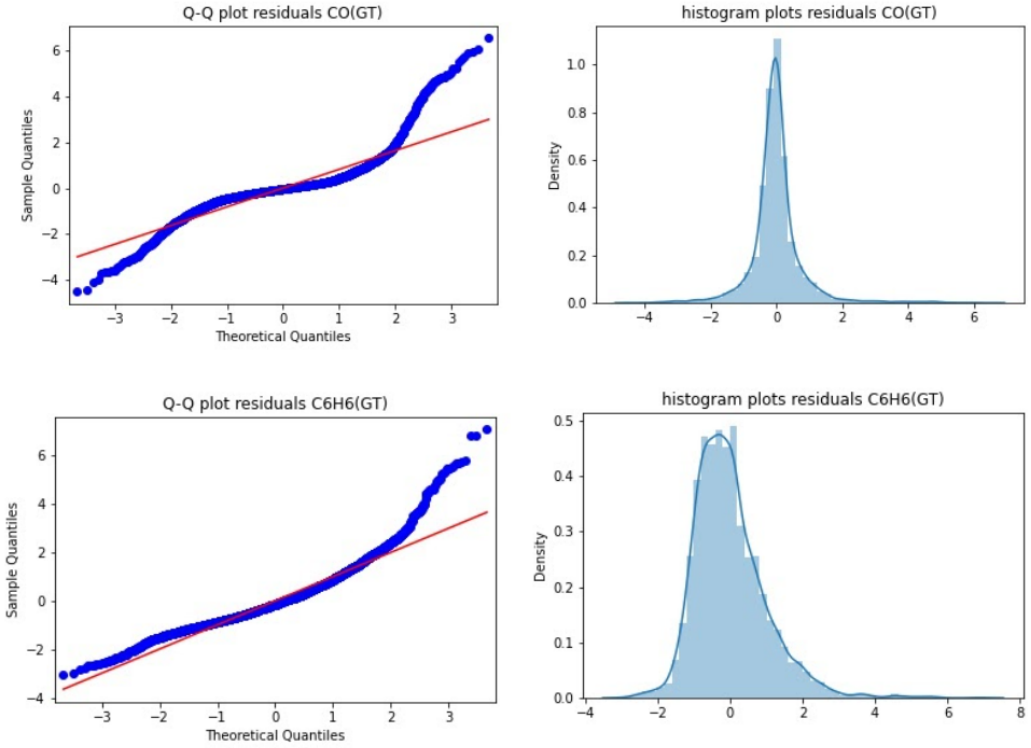

Figure 47: Q-Q and Histogram plots of residuals

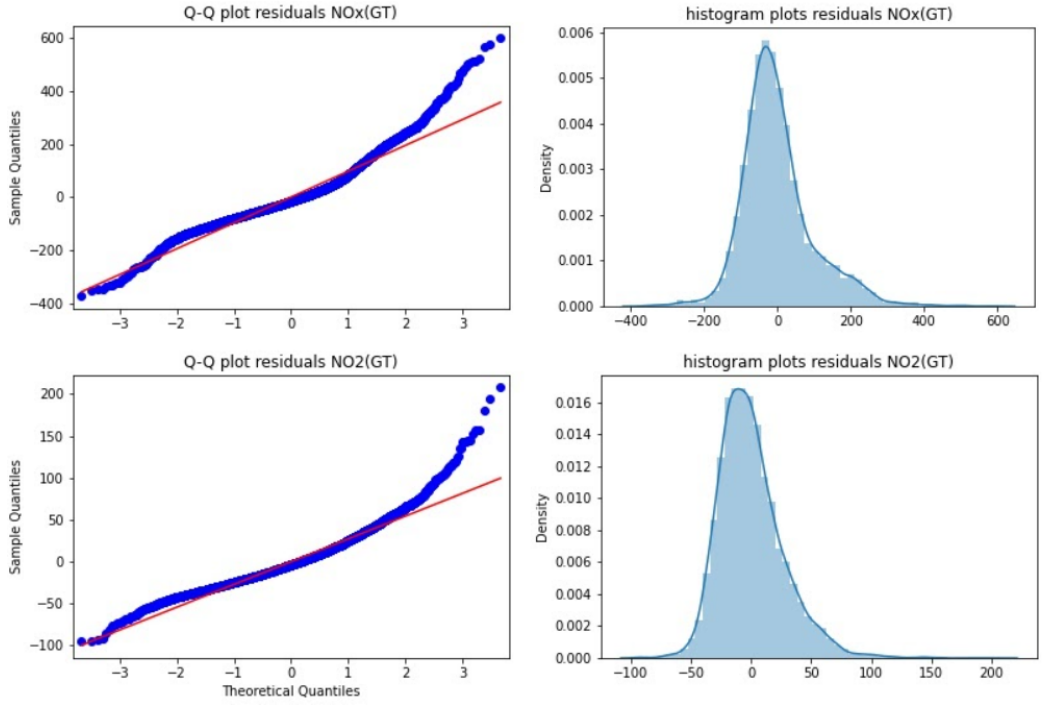

Figure 48: Q-Q and Histogram plots of residuals

3.2.6 Bartlett Sphericity Test :

The Bartlett's test checks if the observed correlation matrix R diverges significantly from the identity matrix (theoretical matrix under H0: the variables are orthogonal). The PCA can perform a compression of the available information only if we reject the null hypothesis.

Model No	Chi-sqrd	p-value
Model 1	79841.13536975828	0.0
Model 2	61581.60716784686	0.0
Model 3	42273.71105205295	0.0
Model 4	58594.539570906345	0.0

Note: The Bartlett's test has a strong drawback. It tends to be always statistically significant when the number of instances 'n' increases. Some references advise to use this test only if the ratio 'n:p' (number of instances divided by the number of variables) is lower than 5. So, KMO Measure of Sampling Adequacy has been done.

KMO Measure of Sampling Adequacy (MSA):

The KMO index has the same goal. It checks if we can factorize efficiently the original variables. But it is based on another idea. If the KMO index is high (1), the PCA can act efficiently; if KMO is low (0), the PCA is not relevant.

Model No	KMO index
Model 1	0.7073
Model 2	0.5608
Model 3	0.5491
Model 4	0.6732

Conclusion: From the above test it can be inferred that the KMO index are rather "mediocre" but close to 1. It can be concluded that PCS can be performed on this dataset.

3.2.7 Principal Component Analysis

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. Although this dataset is not relatively large, from the above test it can be found that PCA can still be applied to this dataset. 95% and 99% variance has been used as metric for performing PCA on the data and the results indicate that 3-4 components are required enough to explain 95% variance and around 5-6 components for explain 99% variance.

PCA Results:

Model no - Explained Var 95	$comp_1$	$comp_2$	$comp_3$	$comp_4$
Model 1	0.52901411	0.32988929	0.07553164	0.03059596
Model 2	0.43533069	0.30738307	0.21080658	-
Model 3	0.43387183	0.29360702	0.21954737	0.03181841
Model 4	0.47196006	0.26237942	0.20521324	0.02854562

- Model1 - CO(GT) :
  $R^2 = 0.635$
  Train MSE = 0.705
  Test MSE = 0.811

- Model2 - C6H6(GT):
  $R^2 = 0.925$
  Train MSE = 3.457
  Test MSE = 4.851
- Model3 - NOx(GT):
  $R^2 = 0.630$
  Train MSE = 11797.458
  Test MSE = 14686.159
- Model4 - NO2(GT):
  $R^2 = 0.597$
  Train MSE = 762.881
  Test MSE = 778.337

Model no - Explained Var 99	$comp_1$	$comp_2$	$comp_3$	$comp_4$	$comp_5$	$comp_6$
Model 1	0.5290	0.3298	0.0755	0.0305	0.0185	0.0120
Model 2	0.4353	0.3073	0.2108	0.0247	0.0107	0.0072
Model 3	0.4338	0.2936	0.2195	0.0318	0.0167	-
Model 4	0.4716	0.2623	0.2052	0.0285	0.0171	0.0108

- Model1 - CO(GT) :
  $R^2 = 0.636$
  Train MSE = 0.702
  Test MSE = 0.810
- Model2 - C6H6(GT):
  $R^2 = 0.943$
  Train MSE = 2.626
  Test MSE = 3.805
- Model3 - NOx(GT):
  $R^2 = 0.630$
  Train MSE = 11786.458
  Test MSE = 14691.159
- Model4 - NO2(GT):
  $R^2 = 0.611$
  Train MSE = 736.881
  Test MSE = 762.337

Conclusion : After performing PCA it can be seen that there is a decrease in R-squared values and there is no significant reduction in dimensionality.

- The residual Q-Q plots and histograms plots show that the residual are close to normal as per the assumptions of regression. (Figure 49 to 50)
- There is no significant autocorrelation among residuals.
- There is a slight evidence of multicollinearity but was later taken care by using PCA.
- The homoscedasticity assumption has been violated even after using various transformations.

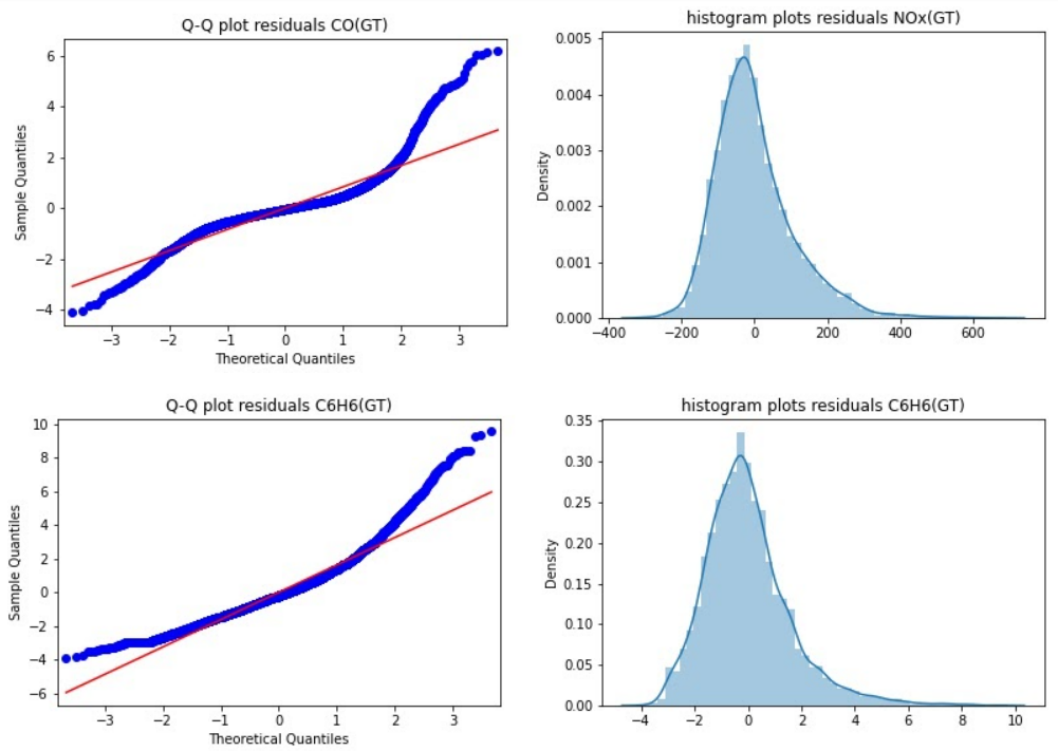

Figure 49: Q-Q and Histogram plots of residuals

- So from the R-squared and MSE results it can concluded that linear regression is a not good fit for Gases NOx and NO2.
- For the Gases CO and C6H6, the results were good but with the violation of assumption.
- So, non-linear regressors - SVM and Random Forest have been used as models to see if they fit better for the data.

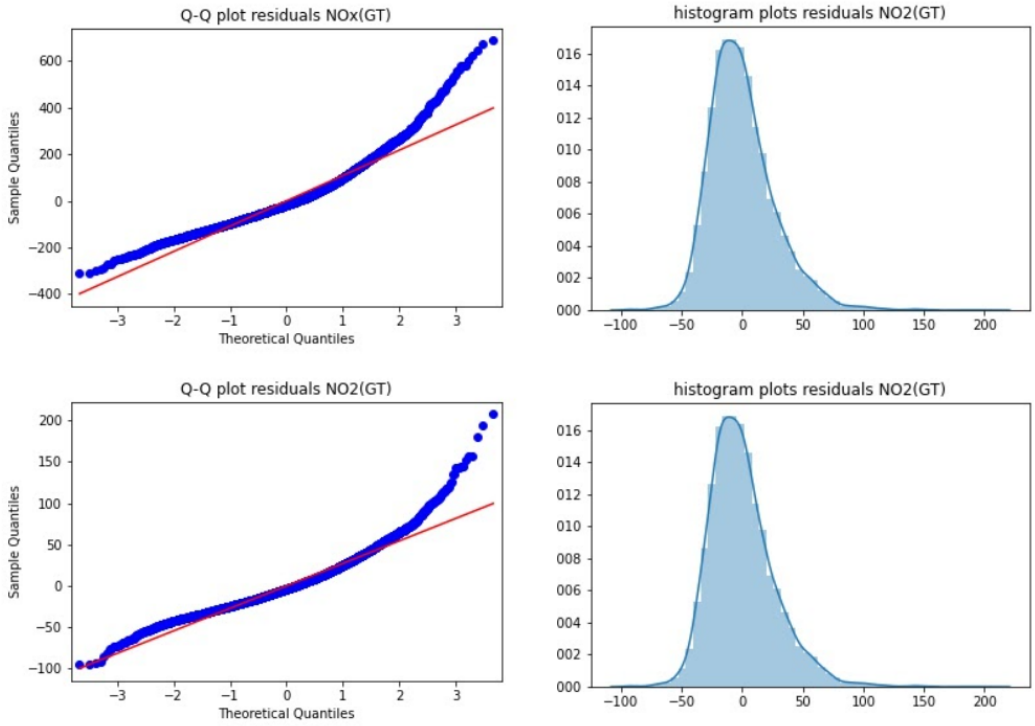

Figure 50: Q-Q and Histogram plots of residuals

3.3 Comparison Analysis of Linear Regression with non-linear Regressors

Non- linear regression models namely SVR (Support Vector Regression) and Random Forest Regression have been used.

Support Vector Regression:

SVR model has been applied on all 4 gases using various kernels. RBF kernel performed better than all other kernels. But the R^2 value is lesser for these models than linear regression models.

Random Forest Regression:

Random forest regression has also been performed on all 4 gases. These models have been run for 100 iterations. These models performed better than Linear Regression and Support Vector Regression.

CO - Model	Train R^2	Test R^2	Train MSE	Test MSE
Linear Regression	0.69	0.74	0.50	0.47
SVR	0.69	0.73	0.51	0.49
Random Forest	**0.90**	**0.79**	**0.15**	**0.39**

NOx - Model	Train R^2	Test R^2	Train MSE	Test MSE
Linear Regression	0.72	0.75	10709	9232
SVR	0.53	0.55	17601	16391
Random Forest	**0.89**	**0.79**	**4226**	**7717**

NO2 - Model	Train R^2	Test R^2	Train MSE	Test MSE
Linear Regression	0.60	0.62	771	735
SVR	0.59	0.6	794	761
Random Forest	**0.90**	**0.72**	**183**	**527**

C6H6 - Model	Train R^2	Test R^2	Train MSE	Test MSE
Linear Regression	0.999	0.999	0.03	0.02
SVR	0.994	0.995	0.26	0.25
Random Forest	**0.999**	**0.999**	**0.003**	**0.009**

Conclusion: It can be clearly seen that non-linear regression model Random forest outperforms Linear Regression for this dataset.