ENVIRONMENTAL CHALLENGES AND THEIR IMPACT ON SUSTAINABILITY AND PUBLIC HEALTH IN AN INDIAN MEGACITY

Ridhi Jain
bs23dmu050

DETAILED PROJECT REPORT

Supervised by
Dr. Suchismita Das

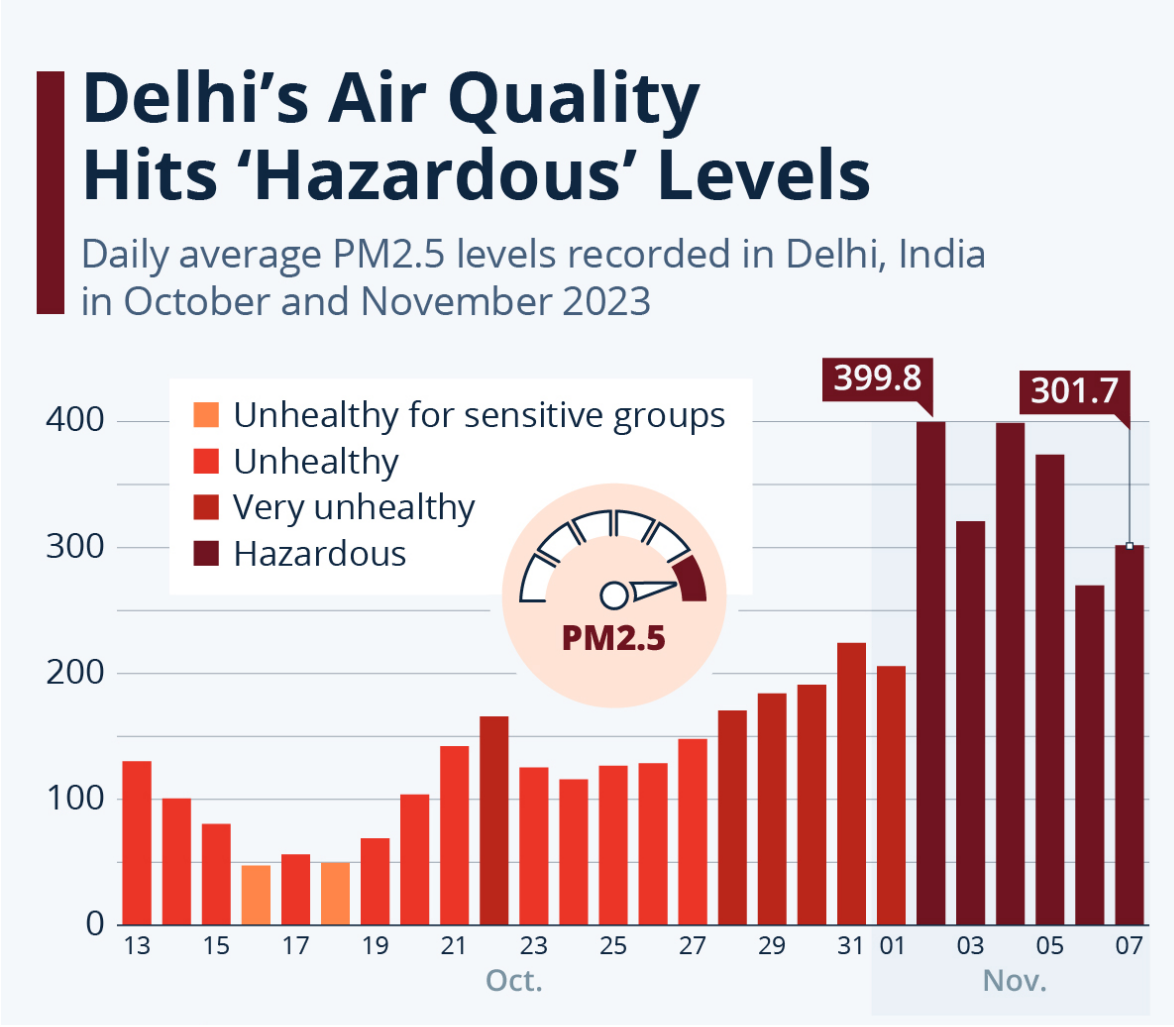


TABLE OF CONTENTS

Abstract

This Environmental Impact Assessment (EIA) analyses air quality trend Delhi using data from 2018 to 2023 obtained from a Kaggle dataset. The analysis employs descriptive statistics and line graphs to explore variations in pollutant concentrations (PM2.5, PM10, SO2, NO2, and O3). The discussion focuses on identifying potential factors influencing these trends, including industrial activity, traffic patterns, and air pollution control measures. The findings from this analysis will inform the EIA process by providing a baseline understanding of existing air quality challenges and potential project impacts in the chosen location.

Keywords: Air quality trends, EIA, Kaggle data, PM2.5, PM10, SO2, NO2, O3, Mumbai, Indore, Chennai, Delhi.

INTRODUCTION

India's rapid urbanization presents both opportunities and challenges. While cities drive economic growth, they also grapple with environmental issues that threaten sustainability and public health. Air Quality is a crucial aspect of environment health.

Pollutants like particulate matter (PM2.5 and PM10), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), and ozone (O_3) can cause respiratory problems, cardiovascular diseases, and other health issues. Understanding existing air quality trends is essential for assessing the potential impacts of the proposed project and developing effective mitigation strategies. This section of EIA focuses on analyses air quality trend Delhi using the data obtained from a Kaggle dataset.

Methodology

The air quality for this analysis was obtained from Kaggle dataset. This dataset covers the period from 2018 to 2023.

The pollutants and the attributes consist of:

1. PM 2.5
2. PM10
3. SO_2
4. NO_2
5. NOx
6. NH3
7. SO2
8. CO
9. Ozone

10. Benzene
11. Toluene
12. Relative Humidity
13. Wind Speed
14. Wind Direction
15. Solar Radiation
16. Barometric Pressure
17. Atmospheric Temperature
18. Rainfall

PM 2.5 is the dependent variable to be studied based on the input variables. The unit for calculating it is shown as ug/m3. Selenium's automation capabilities expedite data collection tasks, automating processes such as data extraction from the CPCB website. Comprehensive data pre-processing techniques, including addressing missing values, standardizing units, and identifying outliers, fortify the dataset's reliability. These steps are pivotal in generating trustworthy insights and minimizing inaccuracies in subsequent analyses. Employing a multiple linear regression model offers a quantitative lens into the relationships between air quality parameters and meteorological variables. This statistical framework empowers the identification of key determinants influencing air quality, enabling predictive modelling for future scenarios.

DATA UNDERSTANDING

Sample Data:

The sample data is shown below:

	From Date	To Date	PM2.5 (ug/m3)	PM10 (ug/m3)	NO (ug/m3)	NO2 (ug/m3)	NOx (ppb)	NH3 (ug/m3)	SO2 (ug/m3)	CO (ug/m3)	...	Benzene (ug/m3)	Toluene (ug/m3)	Temp (degree C)	RH (%)	WS (m/s)	WD (degree)	SR (W/mt2)
0	2018-02-01 10:00:00	2018-02-01 11:00:00	351.0	632.00	19.70	43.93	38.62	24.76	25.24	0.59	...	NaN	NaN	NaN	34.80	1.14	287.00	403.80
1	2018-02-01 11:00:00	2018-02-01 12:00:00	299.0	541.00	17.81	44.46	62.12	40.20	22.90	1.24	...	NaN	NaN	NaN	32.83	1.57	300.83	466.67
2	2018-02-01 12:00:00	2018-02-01 13:00:00	262.0	510.00	11.83	35.40	47.28	28.56	10.26	0.82	...	NaN	NaN	NaN	30.00	1.71	296.83	502.42
3	2018-02-01 13:00:00	2018-02-01 14:00:00	152.0	410.00	6.20	23.92	30.15	27.02	8.41	0.61	...	NaN	NaN	NaN	29.25	1.88	296.08	492.58
4	2018-02-01 14:00:00	2018-02-01 15:00:00	108.0	305.00	6.92	26.38	33.34	23.09	7.33	0.57	...	NaN	NaN	NaN	28.00	1.97	306.50	416.92
5	2018-02-01 15:00:00	2018-02-01 16:00:00	83.0	278.00	7.75	29.44	37.33	21.83	6.83	0.63	...	NaN	NaN	NaN	29.83	2.06	311.92	287.67

- Top 5 observations from a dataset of 45230 observations.
- The data appears to be comprehensive, the observations represent air quality and meteorological data.
- The dataset spans multiple dates and times.
- The data seems to be interval data of continuous type according to the NOIR classification.
- The dataset appears to be unbiased, as it covers a wide range of dates over a 6 year period and it measured every hour.

Key Statistics for data:

From the below table output:

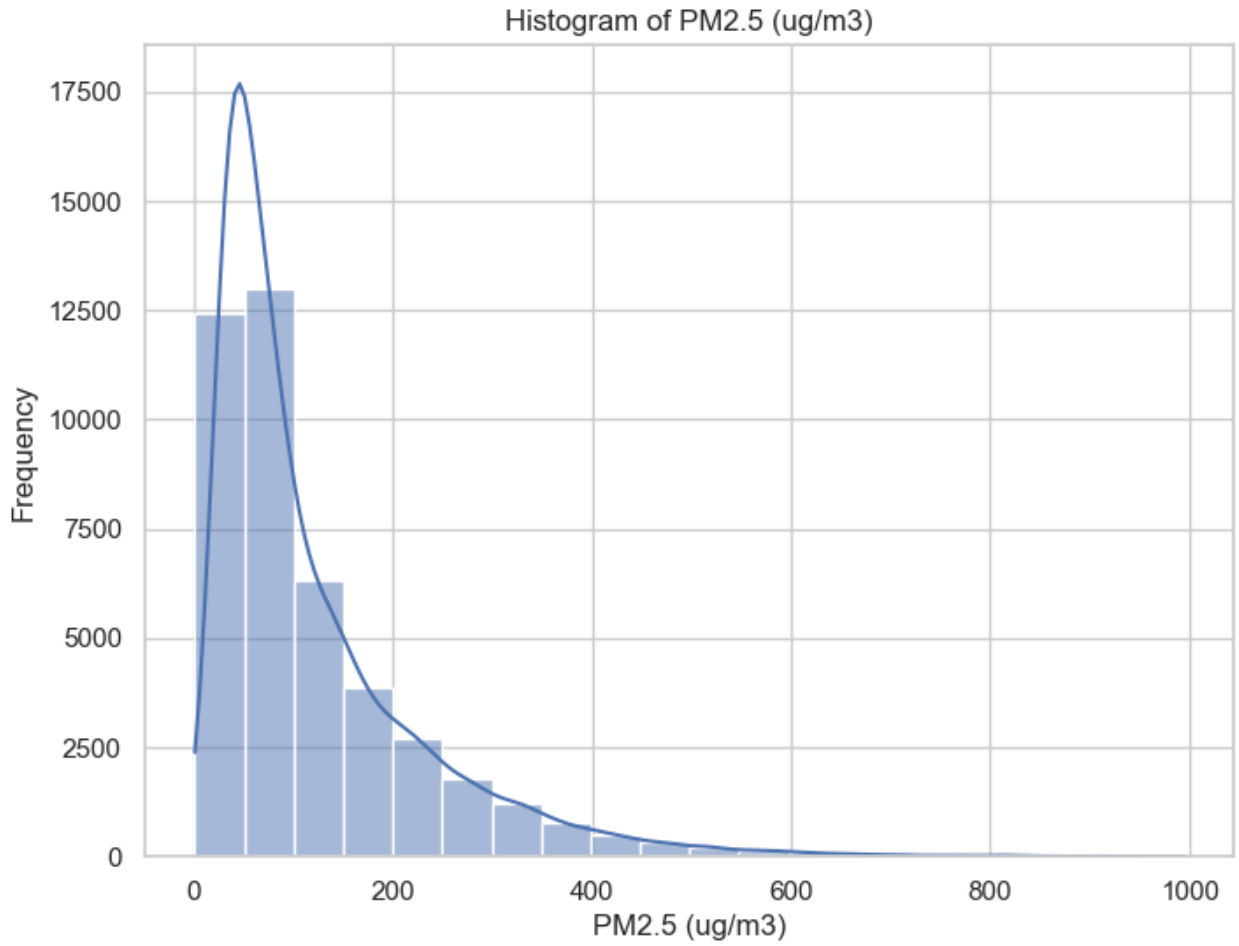
Several observations and deductions can be made:

- The range of values varies significantly across different variables. For example, PM2.5 (ug/m3)" ranges from 1.00 to 995.00, while "RF (mm)" ranges from 0.00 to 36.33.
- Variables like "PM2.5 (ug/m3)" and "PM10 (ug/m3)" have relatively high standard deviations compared to their means, indicating a wide spread of data points around the mean.
- Other variables like "RH (%)" and "WS (m/s)" have lower standard deviations relative to their means
- Variables like "NO (ug/m3)", "NO2 (ug/m3)", "NOx (ppb)", "NH3 (ug/m3)", "SO2 (ug/m3)", "CO (ug/m3)", "Ozone (ug/m3)", "Benzene (ug/m3)", and "Toluene (ug/m3)" have maximum values that are significantly higher than their 75th percentile values, indicating the presence of outliers.
- Similarly, variables "RF (mm)" have a maximum value much higher than its 75th percentile value, suggesting the presence of extreme outliers.
- Temp has all null values

	count	mean	std	min	25%	50%	75%	max
PM2.5 (ug/m3)	43448.0	121.619797	111.309322	1.00	46.50	81.50	158.750	995.00
PM10 (ug/m3)	41396.0	266.544094	170.477198	2.00	143.00	224.75	349.000	1000.00
NO (ug/m3)	43515.0	39.718316	57.508985	0.10	8.88	18.07	44.675	496.90
NO2 (ug/m3)	43512.0	42.572547	31.364454	0.10	22.68	35.99	54.170	471.70
NOx (ppb)	43595.0	63.028696	60.000517	0.00	27.70	42.90	74.885	499.60
NH3 (ug/m3)	42782.0	53.749006	26.783044	0.10	35.20	47.73	67.000	390.77
SO2 (ug/m3)	43436.0	13.690811	10.460615	0.10	5.70	11.01	18.930	164.35
CO (ug/m3)	42139.0	1.472046	1.093601	0.00	0.78	1.20	1.850	9.90
Ozone (ug/m3)	42773.0	21.543758	26.362688	0.10	5.80	10.18	25.440	199.70
Benzene (ug/m3)	41264.0	2.937612	4.061452	0.00	0.50	1.68	3.855	283.77
Toluene (ug/m3)	41940.0	27.763495	47.258623	0.00	3.65	9.85	31.970	494.50
Temp (degree C)	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
RH (%)	43870.0	56.604065	21.311419	0.40	40.58	56.60	73.500	99.90
WS (m/s)	43302.0	1.743398	1.046827	0.15	1.10	1.56	2.150	40.83
WD (degree)	43268.0	209.653383	101.634576	1.33	115.08	236.92	308.070	359.10
SR (W/mt2)	43781.0	118.626442	184.294845	0.15	4.27	11.85	169.750	823.85
BP (mmHg)	43749.0	985.122934	6.992150	724.00	979.43	984.98	991.250	1001.95
AT (degree C)	43841.0	25.845346	8.036744	5.17	19.57	26.90	31.800	51.17
RF (mm)	40510.0	0.019975	0.393567	0.00	0.00	0.00	0.000	36.33

Histograms and Boxplots:

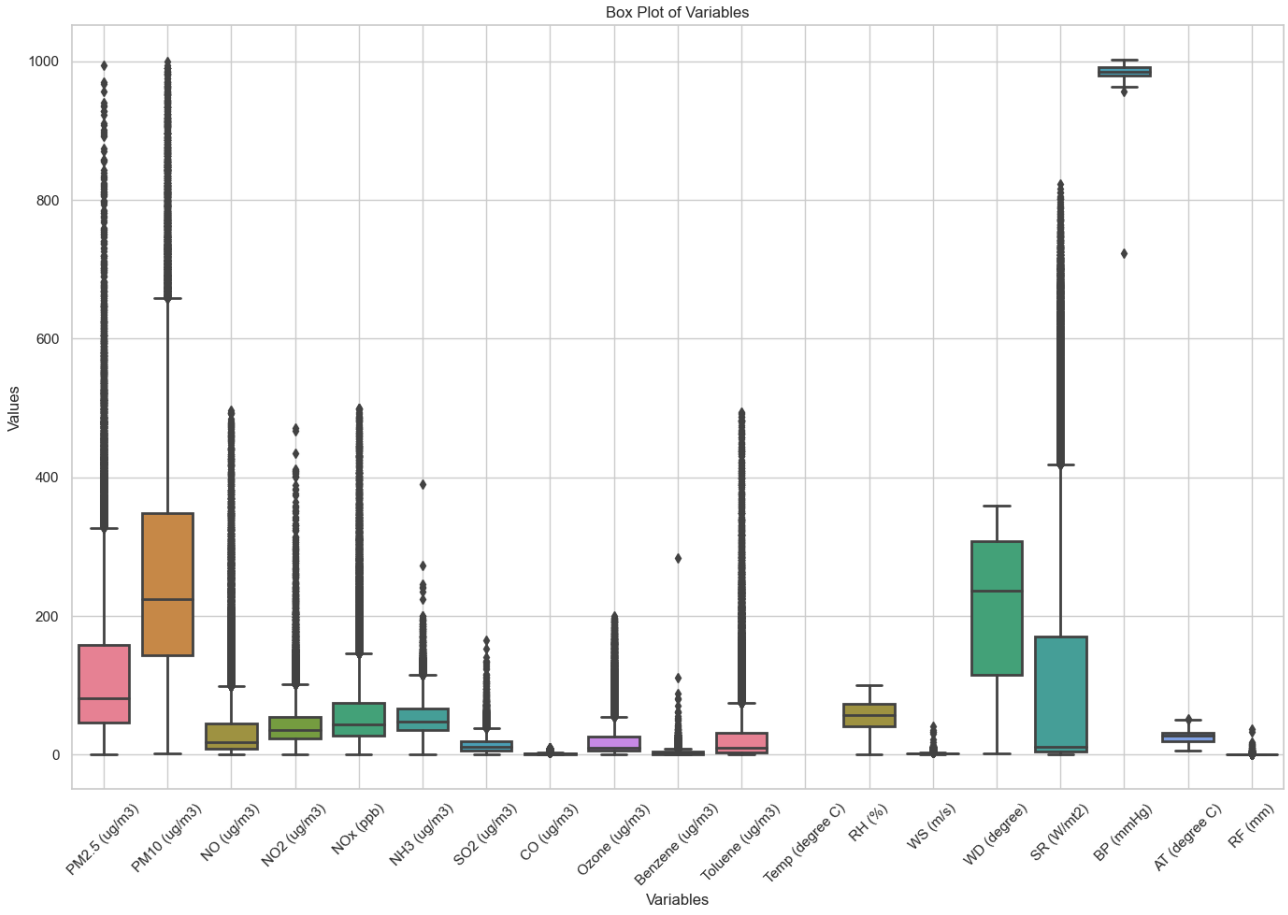
To understand the distribution and characteristics of the data we plot a histogram and boxplot.



From the following histogram of PM2.5 ($\mu g/m^3$) we can observe:

- The histogram appears to be right-skewed
- Peak of the histogram, which represents the most common PM2.5 concentration, appears to be around 50 $\mu g/m^3$.

The box plot below shows how each attribute is classified and how much outliers are present.



In our dataset, we observe a limited number of outliers across the variables. Upon closer inspection, it becomes evident that removing outliers for any particular parameter might disrupt the observed correlations with the output variables (PM2.5 concentration). Therefore, no outlier removal was performed, and the dataset was left unchanged to maintain the integrity of the data analysis.

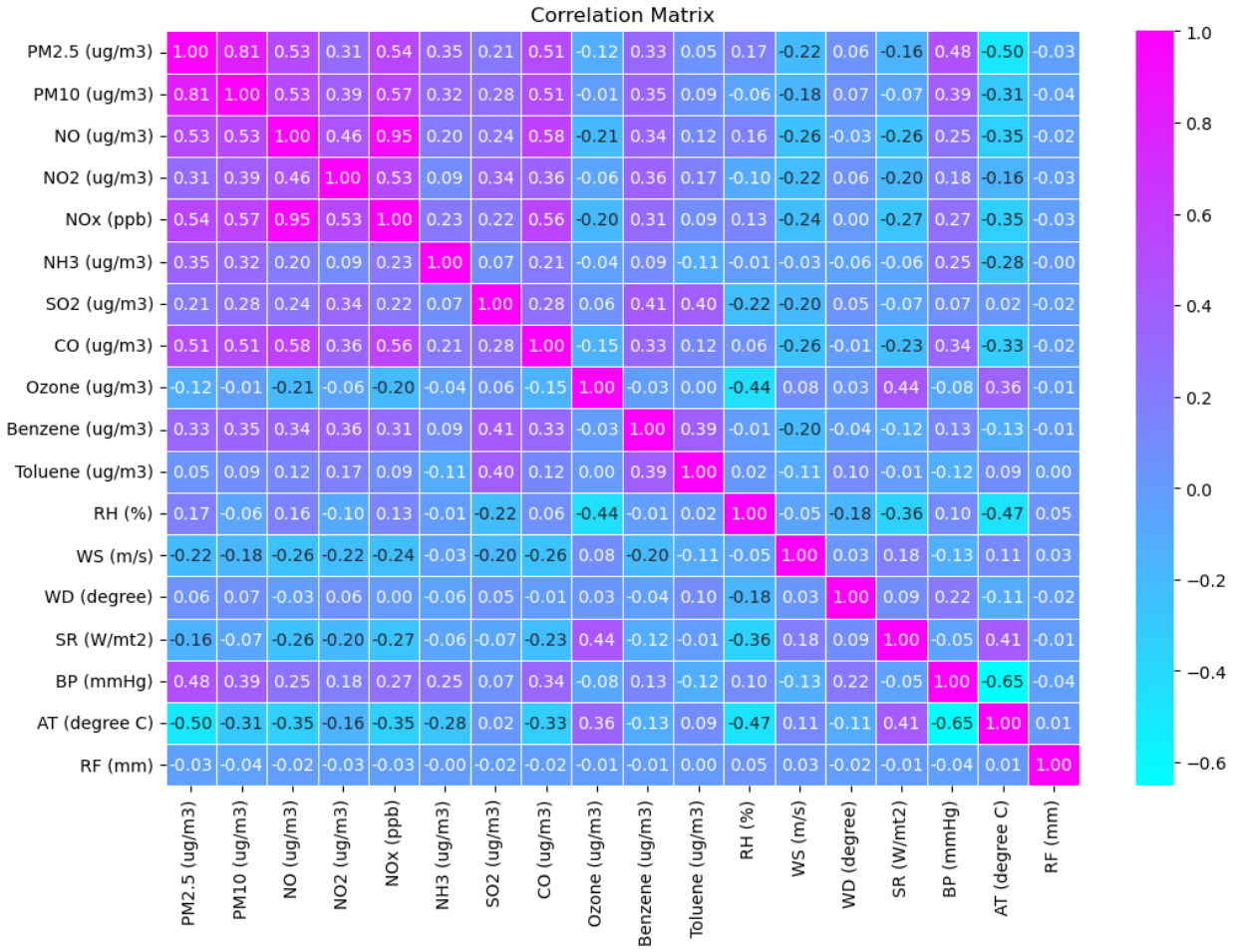
Correlation matrix and Scatterplot:

- Correlation coefficient table shows correlation coefficients between variables. It is a measure of how closely two variables are related.
- The coefficient is denoted by $r_{X,Y}$, can be calculated by:

$$r_{X,Y} = \frac{Cov(X,Y)}{\sigma_{X,Y}}$$

- $r_{X,Y}$ provided a measure of linear relationship between X and Y.
- It is a measure of degree of relationship
- Correlation Matrix is a $n \times n$ matrix (square, upper triangular or lower triangular consisting of total of 'n' attributes) of correlation coefficient between all variables of the data taken 2 at a time.
- The scatterplot matrix allows us to visualize the relationships between pairs of variables.
- By examining the scatterplots, we can also assess the strength and direction of the correlation between variables.

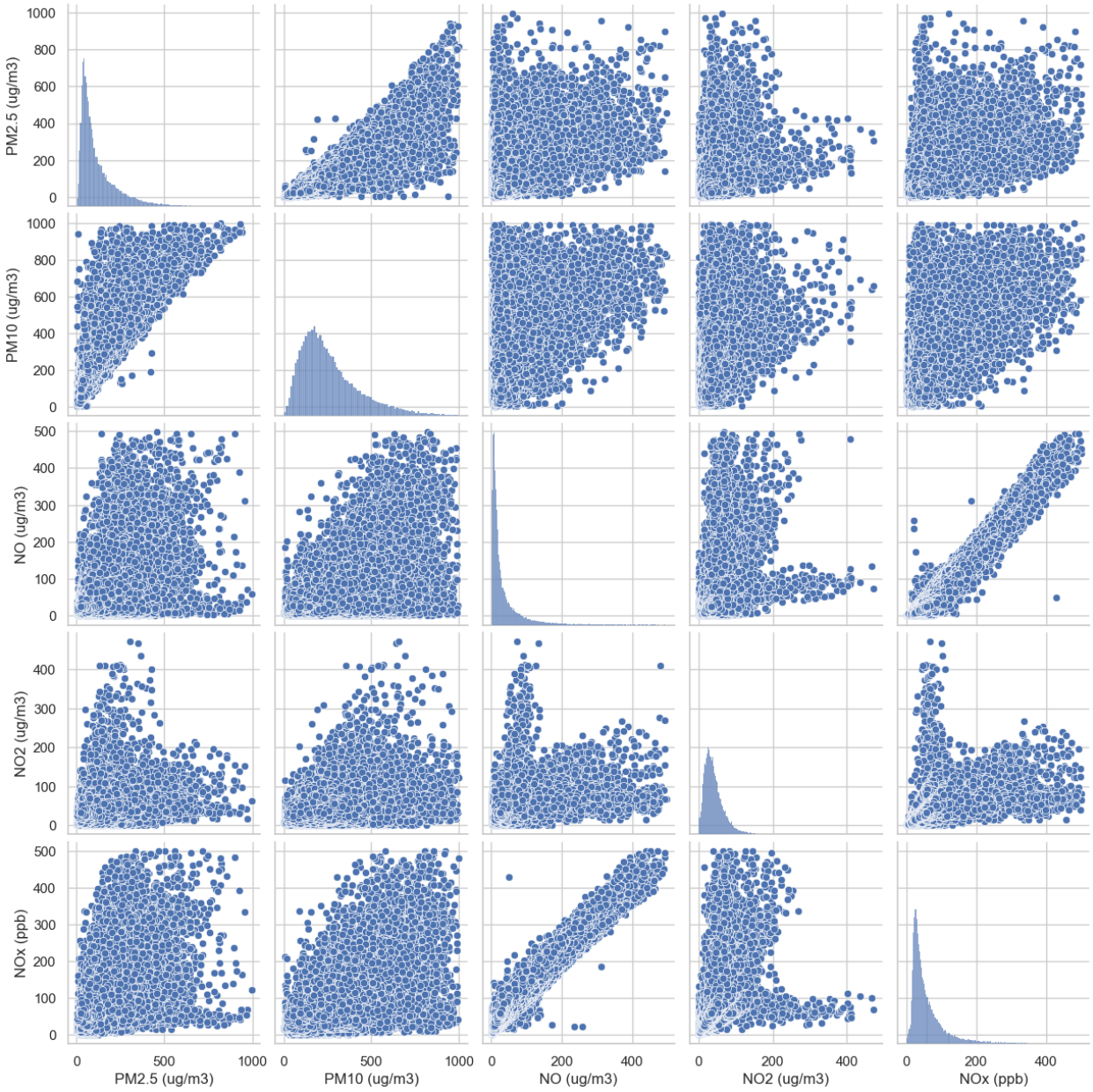
Correlation Matrix:



Scatterplot:

- The scatterplot below provides a visual summary of the relationships between the variables in the dataset, helping to identify patterns and correlations between air quality parameters.
- PM2.5 and PM10 have a correlation coefficient of approximately 0.81, indicating a strong positive linear relationship.
- Toluene and RF (Rainfall) have a correlation coefficient of approximately -0.03, indicating a weak negative linear relationship.

- Scatterplot that appear as a straight line indicate a perfect linear relationship, while scattered points suggest a weaker relationship or no relationship at all.



SPLITTING DATA FOR TRAINING AND TESTING

The next step of our analysis should be to split into training and testing the data.

```
: from sklearn.model_selection import train_test_split

# Define your features (X) and target variable (y)
X = df.drop(columns=['PM2.5 (ug/m3)'])  # Features
y = df['PM2.5 (ug/m3)']  # Target variable

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

- Data is split in 80:20 ratio, 80% of the data will be used for training the data and remaining 30% will be used for testing the data.
- Training data will be used to train the regression model.
- In the end we will use the testing data for predictions and accuracy of the data

REGRESSION ANALYSIS

Now that we have visualised the data and seen their respective correlations, we take this data for the regression model

- A multiple linear regression model uses the formula:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_{k-1} + \varepsilon$$

- For hypothesis testing and the setting of confidence limits, we also assume that ε is normally distributed.
- Hence, $\sum \varepsilon = 0$.
- The linearity of the model above is defined with respect to the regression coefficients.

X variables β_1, β_2, etc ... in the test as follows:

1. PM10

2. SO2

3. NO2

4. NOx

5. NH3

6. CO

7. Ozone

8. Benzene

9.Toluene

10. Relative Humidity

11. Wind Speed

12. Wind Direction

13. Solar Radiation

14. Barometric Pressure

15. Atmospheric Temperature

16. Rainfall

Y variable for the model is:

1. PM2.5

Regression Model – Trial 1:

Fitting Multiple Linear Regression with all X variables included:

OLS Regression Results

Dep. Variable:	PM2.5 (ug/m3)	R-squared:	0.756
Model:	OLS	Adj. R-squared:	0.756
Method:	Least Squares	F-statistic:	8251.
Date:	Tue, 16 Apr 2024	Prob (F-statistic):	0.00
Time:	21:11:17	Log-Likelihood:	-2.4448e+05
No. Observations:	45230	AIC:	4.890e+05
Df Residuals:	45212	BIC:	4.892e+05
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-860.9400	55.995	-15.375	0.000	-970.692	-751.188
PM10 (ug/m3)	0.4549	0.002	211.314	0.000	0.451	0.459
NO (ug/m3)	0.0023	0.015	0.152	0.879	-0.028	0.032
NO2 (ug/m3)	-0.1296	0.011	-11.991	0.000	-0.151	-0.108
NOx (ppb)	0.0608	0.015	4.026	0.000	0.031	0.090
NH3 (ug/m3)	0.2324	0.011	21.546	0.000	0.211	0.254
SO2 (ug/m3)	0.3480	0.031	11.327	0.000	0.288	0.408
CO (ug/m3)	5.3284	0.322	16.551	0.000	4.697	5.959
Ozone (ug/m3)	0.0796	0.012	6.716	0.000	0.056	0.103
Benzene (ug/m3)	1.1315	0.081	13.927	0.000	0.972	1.291
Toluene (ug/m3)	-0.0698	0.007	-10.423	0.000	-0.083	-0.057
RH (%)	0.8345	0.017	50.467	0.000	0.802	0.867
WS (m/s)	-4.7907	0.265	-18.050	0.000	-5.311	-4.270
WD (degree)	0.0168	0.003	6.035	0.000	0.011	0.022
SR (W/mt2)	0.0189	0.002	10.681	0.000	0.015	0.022
BP (mmHg)	0.8558	0.056	15.274	0.000	0.746	0.966
AT (degree C)	-2.0482	0.056	-36.265	0.000	-2.159	-1.937
RF (mm)	-1.9823	0.682	-2.906	0.004	-3.319	-0.645

Omnibus:	29031.346	Durbin-Watson:	0.307
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1245313.454
Skew:	2.503	Prob(JB):	0.00
Kurtosis:	28.214	Cond. No.	2.34e+05

This model has a R2 of 75.6%, which indicates a good model but can be further improved.

From the above summary of the model, the Multiple Regression Equation to be:

$$PM2.5 = -860.9400 + 0.4549*PM10 + 0.0023*NO - 0.1296*NO_2 + 0.0608*NO_x + 0.2324*NH_3 + 0.3480*SO_2 + 5.3284*CO + 0.0796*Ozone + 1.1315*Benzene - 0.0698*Toluene + 0.8345*RH - 4.7907*WS + 0.0168*WD + 0.0189*SR + 0.8558*BP - 2.0482*AT - 1.9823*RF$$

Multiple Regression Equation with all the variables

Since the dimensionality is very high and there are many variables in one equation, we can perform 'P-Test'.

Criteria:

We perform *Anova* on the model to find p-value for each variable. The variables with p value < 0.05 hold good for predicting the output, hence we keep them. We remove the insignificant variables from the equation.

Anova - for Model 1:

```
# Define a significance threshold
alpha = 0.05

# Extract p-values for each coefficient
p_values = model.pvalues

# Identify significant variables based on the significance threshold
significant_variables = p_values[p_values < alpha]

# Display significant variables
print("Significant Variables:")
print(significant_variables)
```

```
Significant Variables:
const                1.142205e-96
PM10 (ug/m3)         0.000000e+00
NO2 (ug/m3)          1.400950e-40
NOx (ppb)            1.499346e-04
NH3 (ug/m3)         5.217498e-111
SO2 (ug/m3)          9.254869e-16
CO (ug/m3)           1.693053e-51
Ozone (ug/m3)        1.026339e-20
Benzene (ug/m3)      1.460160e-28
RH (%)               0.000000e+00
WS (m/s)             3.767142e-62
BP (mmHg)            5.356132e-98
AT (degree C)       2.511699e-280
RF (mm)              4.503465e-03
dtype: float64
```

- The *Anova* table shows the p-value for each variable
- Based on the table above, we can say that since variables [PM10, NO_2, NO_x, NH_3, SO_2, CO, Ozone, Benzene, RH, WS, BP, AT, RF] have p-values<0.05, apart from these variables, we can remove other ones.
- We remove variables that has no effect on the y variable so based on that we are rejecting the null hypothesis and hence disregarding that variable.

Let us proceed further to improve the model in the following steps:

Model 2: (After removing the insignificant attributes)

X variables $\beta 1$, $\beta 2$ $etc.$ … in the test are as follows:

1. PM10
2. SO2
3. NO2
4. NOx
5. NH3
6. CO
7. Ozone
8. Benzene
9. ~~Toluene~~
10. Relative Humidity
11. Wind Speed
12. ~~Wind Direction~~
13. ~~Solar Radiation~~

14. Barometric Pressure

15. Atmospheric Temperature

16. Rainfall

Y variable for the model is:

1. PM2.5

Regression Model – Trial 2:

Fitting Multiple Linear Regression with updated X variables:

OLS Regression Results

Dep. Variable:	PM2.5 (ug/m3)	R-squared (uncentered):	0.890
Model:	OLS	Adj. R-squared (uncentered):	0.890
Method:	Least Squares	F-statistic:	2.806e+04
Date:	Thu, 18 Apr 2024	Prob (F-statistic):	0.00
Time:	09:49:18	Log-Likelihood:	-2.4482e+05
No. Observations:	45230	AIC:	4.897e+05
Df Residuals:	45217	BIC:	4.898e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PM10 (ug/m3)	0.4637	0.002	217.462	0.000	0.460	0.468
NO2 (ug/m3)	-0.1454	0.011	-13.806	0.000	-0.166	-0.125
NOx (ppb)	0.0534	0.006	8.427	0.000	0.041	0.066
NH3 (ug/m3)	0.2367	0.011	22.191	0.000	0.216	0.258
SO2 (ug/m3)	0.2223	0.029	7.595	0.000	0.165	0.280
CO (ug/m3)	5.4058	0.316	17.106	0.000	4.786	6.025
Ozone (ug/m3)	0.1333	0.011	11.628	0.000	0.111	0.156
Benzene (ug/m3)	0.8200	0.077	10.601	0.000	0.668	0.972
RH (%)	0.6989	0.016	45.082	0.000	0.668	0.729
WS (m/s)	-4.7854	0.264	-18.134	0.000	-5.303	-4.268
BP (mmHg)	0.0090	0.002	4.076	0.000	0.005	0.013
AT (degree C)	-2.6379	0.041	-64.667	0.000	-2.718	-2.558
RF (mm)	-2.2281	0.687	-3.243	0.001	-3.575	-0.882

Omnibus:	28540.981	Durbin-Watson:	0.302
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1199207.553
Skew:	2.445	Prob(JB):	0.00
Kurtosis:	27.747	Cond. No.	2.77e+03

From the above summary for our model, we derive the new Multiple Regression equation to be:

PM2.5 = -860.9400 + 0.4637PM10 - 0.1454NO2 +

0.0534NOx + 0.2367NH3 + 0.2223SO2 + 5.4058CO +

0.1333Ozone + 0.8200Benzene + 0.6989RH - 4.7854WS

New Multiple Regression Equation

Anova for Model 2:

```
const: F-score = 10717.842447482715, p-value = 8.798261602659546e-97
PM10 (ug/m3): F-score = 10717.842447482715, p-value = 0.0
NO2 (ug/m3): F-score = 10717.842447482715, p-value = 2.3555533575613842e-43
NOx (ppb): F-score = 10717.842447482715, p-value = 4.9855722100882794e-23
NH3 (ug/m3): F-score = 10717.842447482715, p-value = 2.3255439379758647e-111
SO2 (ug/m3): F-score = 10717.842447482715, p-value = 5.14798528274281e-16
CO (ug/m3): F-score = 10717.842447482715, p-value = 1.2295575638987135e-53
Ozone (ug/m3): F-score = 10717.842447482715, p-value = 9.232998818454881e-21
Benzene (ug/m3): F-score = 10717.842447482715, p-value = 1.8259533752731865e-29
RH (%): F-score = 10717.842447482715, p-value = 0.0
WS (m/s): F-score = 10717.842447482715, p-value = 8.873593050460108e-63
BP (mmHg): F-score = 10717.842447482715, p-value = 3.719178064109909e-98
AT (degree C): F-score = 10717.842447482715, p-value = 3.9826021037002665e-281
RF (mm): F-score = 10717.842447482715, p-value = 0.0045484502975368084
```

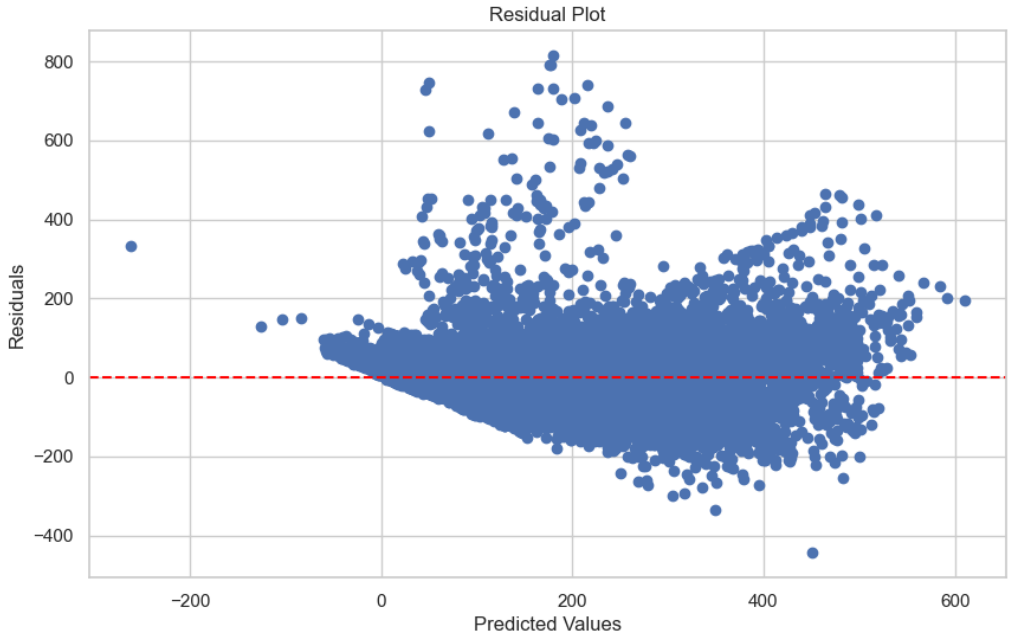
⇒ As seen, the p-value<0.05 for the above variables indicating that Model-2 holds good for predicting the output

⇒ Hence, the above results convey that all the p-values are significant.

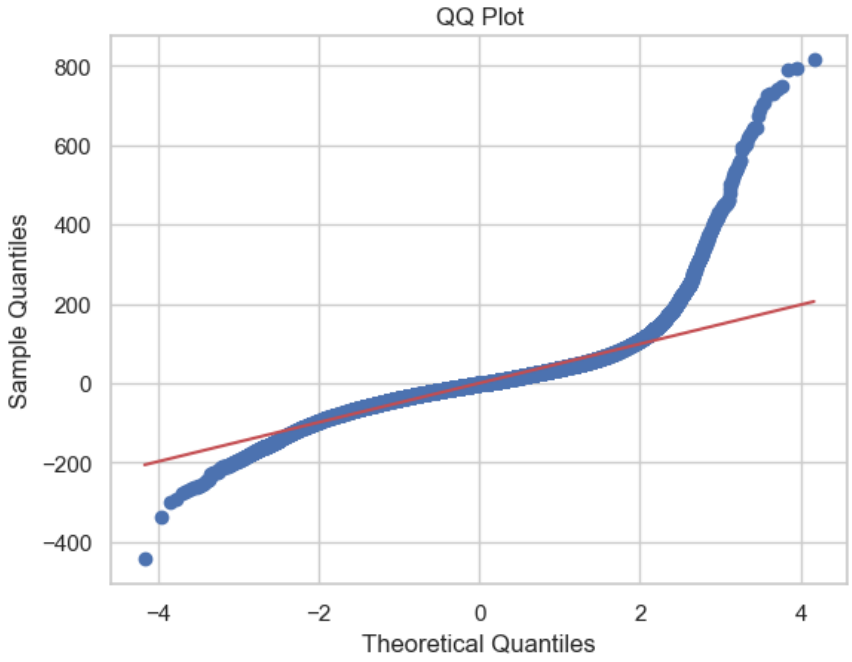
Furthermore, now that our final model is ready with the final Regression Equation, we can do 'Residual Analysis', which is draw the QQ plots and Residual Plots.

RESIDUAL ANALYSIS

Residual Plot:



Normal Q-Q Plot:



Inference:

- ⇒ As observed our data has a skewness of 2.445, we are going to assume that are data is normally distributed.
- ⇒ Q-Q plot is over-dispersed data as it appears as flipped S shape.
- ⇒ The extended lower tail and upper tail of the data's distribution is relative to a normal distribution.
- ⇒ The variability in the response variable is greater than what would be expected under a normal distribution assumption.

HYPOTHESIS TESTING ON MODEL 2 (FINAL MODEL) OUTCOME:

Null Hypothesis:

$$H0: \beta 0 = \beta 1 = \cdots = \beta k{-}1 = 0$$
$$H1: \beta j \neq 0, \; for \; atleast \; one \; j$$

```
import statsmodels.api as sm

# Define the dependent variable (y) and independent variables (X)
y = df['PM2.5 (ug/m3)']
X = df[['PM10 (ug/m3)', 'NO2 (ug/m3)', 'NOx (ppb)', 'NH3 (ug/m3)', 'SO2 (ug/m3)',
        'CO (ug/m3)', 'Ozone (ug/m3)', 'Benzene (ug/m3)', 'RH (%)', 'WS (m/s)',
        'BP (mmHg)', 'AT (degree C)', 'RF (mm)']]

# Add a constant to the independent variables matrix
X = sm.add_constant(X)

# Fit the OLS (Ordinary Least Squares) model
model = sm.OLS(y, X).fit()

# Perform hypothesis testing for each coefficient
hypothesis_tests = model.t_test(np.eye(len(model.params)))

# Print the results
print(hypothesis_tests)
```

```
                             Test for Constraints
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c0         -1114.7686     53.270    -20.927      0.000   -1219.178   -1010.359
c1             0.4565      0.002    212.353      0.000       0.452       0.461
c2            -0.1449      0.010    -13.820      0.000      -0.165      -0.124
c3             0.0625      0.006      9.887      0.000       0.050       0.075
c4             0.2387      0.011     22.486      0.000       0.218       0.260
c5             0.2364      0.029      8.111      0.000       0.179       0.293
c6             4.8714      0.316     15.439      0.000       4.253       5.490
c7             0.1073      0.011      9.349      0.000       0.085       0.130
c8             0.8686      0.077     11.279      0.000       0.718       1.020
c9             0.7759      0.016     48.921      0.000       0.745       0.807
c10           -4.4091      0.263    -16.749      0.000      -4.925      -3.893
c11            1.1189      0.053     21.079      0.000       1.015       1.223
c12           -1.9190      0.053    -36.084      0.000      -2.023      -1.815
c13           -1.9404      0.684     -2.838      0.005      -3.281      -0.600
==============================================================================
```

Based on the results of the hypothesis testing at a 95% confidence level, all coefficients in the regression model are found to be statistically significant. This implies that the predictor variables included in the model have a significant effect on the dependent variable, PM2.5 concentration. The p-values associated with each coefficient are all below the threshold of 0.05, indicating strong evidence against the null hypothesis for each variable.

PREDICTION FOR TEST DATA

1) Using test data to make predictions:

```
# Initialize the random forest regressor
rf_regressor = RandomForestRegressor()

# Train the model
rf_regressor.fit(X_train, y_train)

# Predict missing values for test set
predicted_values = rf_regressor.predict(X_test)
```

- These lines of the Python code predicts the value for testing data, which was divided in 80:20 ratio before staring regression analysis
- It fits test into the final model (Model 2) and stores the predictions as a variable *'predicted_values'*.

2) Model Evaluating Results

```
# Calculate evaluation metrics
mse_test = mean_squared_error(y_test, predicted_values)
mae_test = mean_absolute_error(y_test, predicted_values)
r2_test = r2_score(y_test, predicted_values)

# Print evaluation metrics
print("Mean Squared Error (Test):", mse_test)
print("Mean Absolute Error (Test):", mae_test)
print("R-squared (Test):", r2_test)
```

```
Mean Squared Error (Test): 1027.5075656398074
Mean Absolute Error (Test): 19.30340609034635
R-squared (Test): 0.9136225548493486
```

These results indicate the performance of our model in predicting the concentration of air pollutants. The MSE and MAE values provide insights into the average squared and absolute differences between our model's predictions and actual values, respectively. A lower MSE and MAE indicate better performance. Additionally, the R-square value represents the proportion of variance in the dependent variable that is explained by our model. A higher R-squared value closer to 1 indicates a better fit of the model to the data.

Based on these evaluation metrics, our model demonstrates effective predictive capability for air pollutant concentrations based on the provided meteorological and environmental variables.

Visualising the performance

In our analysis, we utilised a performance metrics plot to comprehensively evaluate the performance of our predictive model on the testing data. The plot served as a crucial tool in understanding the effectiveness of our model in predicting air quality parameters.

The performance metrics plot provided a vivid comparison of key evaluation metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared. By presenting these metrics on a logarithm scale, the plot enabled us to discern the magnitude of errors and assess the overall quality of the model's predictions. These metrics indicate a

relatively high level of accuracy and goodness of fit, suggesting that our model performs well in predicting air quality parameters.

The performance metrics plot not only provided a comprehensive overview of model performance but also facilitated the identification of potential areas for further refinement. By monitoring trends in the metrics over time or across different iterations of the model, we can iteratively improve our predictive model to better meet our objectives.

Overall, the performance metrics plot played a pivotal role in the analysis, enabling us to gauge the effectiveness of our predictive model and make informed decisions for further model enhancement.

CONCLUSION

In the context of this analysis, exploring alternative regression models offers promising avenues for improving predictive accuracy and gaining deeper insights into air quality dynamics. Among these alternatives, Gradient Boosting Regression stands out for its ability to incrementally enhance predictive performance and capture intricate relationships within the data. By considering algorithms like XGBoost or LightGBM, we can unlock the potential to achieve superior results while navigating complex datasets.

Additionally, the utilization of Neural Network Regression presents an intriguing opportunity to delve into nonlinear relationships and nuanced patterns embedded within air quality data. Leveraging deep learning architectures such as feedforward neural networks or LSTM networks can empower us to extract meaningful insights and make more accurate predictions, particularly in scenarios characterized by intricate data structures.

Furthermore, the incorporation of Time Series Forecasting Models offers a robust framework for capturing temporal dependencies and seasonality patterns inherent in air quality data. Techniques such as ARIMA, LSTM, or Prophet hold promise in uncovering temporal dynamics and enabling more accurate forecasting, thereby enhancing our understanding of air quality fluctuations over time.

In parallel, the adoption of Bayesian Regression models holds significant potential in quantifying uncertainty and furnishing interpretable results. By embracing Bayesian regression techniques, we can navigate the complexities of uncertainty inherent in air quality data and make more informed decisions with confidence.

Finally, the exploration of Ensemble Regression Techniques offers a comprehensive approach to model building, leveraging the strengths of multiple regression models to create robust and reliable predictive models. Techniques such as stacking or blending provide avenues for amalgamating diverse modelling approaches and harnessing their collective predictive power, ultimately enhancing the overall performance and reliability of our predictive models.

REFERENCES:

1. Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu ." Detection and Prediction of Air Pollution using Machine Learning Models". International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4 – May 2018
2. Central Pollution Control Board. (n.d.). Air Pollution. Retrieved from https://cpcb.nic.in/air-pollution/
3. Dutta A, Jinsart W. Air Quality, Atmospheric Variables and Spread of COVID-19 in Delhi (India): An Analysis. Aerosol Air Qual. Res. 2021b. 21: 200417. https://doi.org/10.4209/aaqr.2020.07.0417.
4. Sharma AK, Baliyan P, Kumar P. Air pollution and public health: the challenges for Delhi, India. Rev Environ. Health. 2018. 28; 33(1): 77–86. https://doi.org/10.1515/reveh-2017-0032
5. Guttikunda, S. K., & Gurjar, B. R. (2012). Role of meteorology in seasonality of air pollution in megacity Delhi, India. Environmental Monitoring and Assessment, 184, 3199–3211. https://doi.org/10.1007/s10661-011-2182-8