Prediction of Nitrogen Oxides level in Air

By

Akshay Kumar Yadav (20MCB1015)

Ezhil Oviya (20MCB1003)

Vishnu Shashank (20MCB1009)

A project report submitted to

Dr. G. BHARADWAJA KUMAR

COMPUTER SCIENCE IN BIG DATA

in partial fulfilment of the requirements for the course of

CSE 5007 –EXPLORATORY DATA ANALYSIS

in

M.Tech. ELECTRONICS AND COMPUTER ENGINEERING

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Vandalur – Kelambakkam Road

Chennai – 600127

NOVEMBER 2020

ABSTRACT

There are many poisonous gases such as CO, NOx, titania, sulphur dioxide, nitrous oxides, methane etc in the polluted air which is harmful not to only humans but also animals and plan lives based on this references Air Pollution Effects on health.

This project aims to predict the level of Nitrogen Oxides in Air. The data was acquired from the UCI Machine Learning Repository. You can find all the information about Dataset from here Air Quality Dataset. It's hourly data.

Built a prediction model using linear regression with data cleaning, and exploratory data analysis with visualization.

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Dr. G. Bharadwaja Kumar,** Senior Professor, School of Computer Science Engineering, for his consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

We express our thanks to our **Head of The Dept Dr. Janaki Meena (for M.Tech – Big Data)** for her support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the course.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

AKSHAY **OVIYA** **VISHNU**

TABLE OF CONTENTS

INTRODUCTION

As the largest growing industrial nation, India is producing record amount of pollutants specifically Co2, Nox, NO2 etc and other harmful aerial contaminants. Air quality of a particular state or a country is a measure on the effect of pollutants on the respected regions, as per the Indian air quality standard pollutants are indexed in terms of their scale, these air quality indexes indicates the levels of major pollutants on the atmosphere. There are various atmospheric gases which causes pollution on our environment. Each pollution has individual index and scales at different levels. And finding which gas is more likely to harm for the environment .

Steps for Proposed Methodology:

- **DATA CLEANING**
- **PERFORMING ANALYSIS**
- **TRAINING THE LINEAR REGRESSION MODEL**
- **MODEL EVALUATION**
- **PREDICTION MODEL**

DATASET:

Air Quality Data Set is used

Attribute information:

Date(DD/MM/YYYY)

Time(HH.MM.SS)

True hourly averaged concentration CO in mg/m^3

PT08.S1 hourly averaged sensor response

True hourly averaged overall Non-Metallic Hydro Carbons

True hourly averaged Benzene concentration in microg/m^3

PT08.S2 (titania) hourly averaged sensor response

True hourly averaged NOx concentration in ppb

PT08.S3 hourly averaged sensor response

True hourly averaged NO2 concentration in microg/m^3

PT08.S4 hourly averaged sensor response

PT08.S5 hourly averaged sensor response

Temperature

RH Relative Humidity

AH Absolute Humidity

Methods Used:

Numpy : Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.
Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data.

Pandas : Pandas is an open-source library that is built on top of NumPy library. It is a Python package that offers various data structures and operations for manipulating numerical data and time series. It is mainly popular for importing and analyzing data much easier. Pandas is fast and it has high-performance & productivity for users.

Matplotlib : Matplotlib is an visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Seaborn : Seaborn is an visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas. Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, So that we can switch between different visual representations for same variables for better understanding of dataset.

Heat Map : A heat map (or heatmap) is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.

Correlation Coefficient With Heatmap : Correlation is a measure of the linear relationship of 2 or more variables. Through correlation, we can predict one variable from the other. The logic behind using correlation for feature selection is that the good variables are highly correlated with the target. Furthermore, variables should be correlated with the target but should be uncorrelated among themselves.

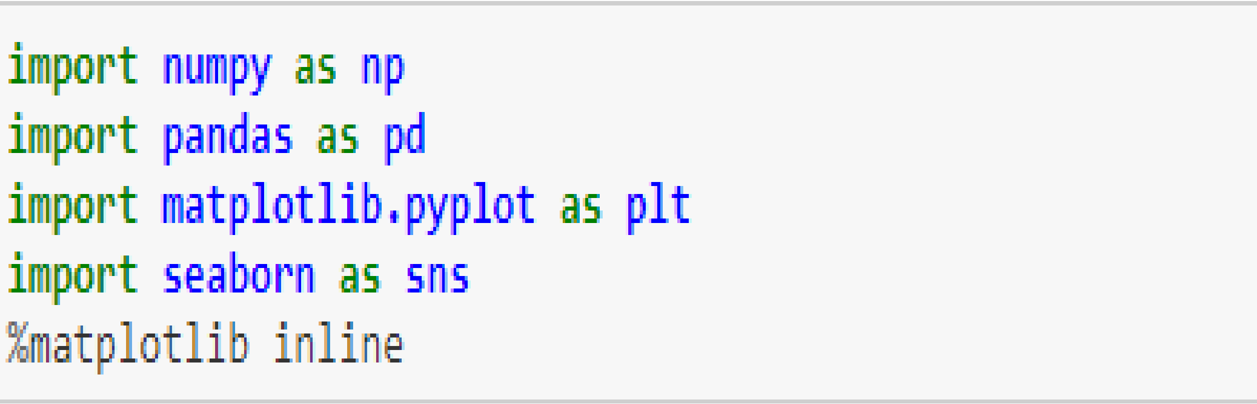
If two variables are correlated, we can predict one from the other. Therefore, if two features are correlated, the model only really needs one of them, as the second one does not add additional information. We will use the Pearson Correlation here.
A heat map (or heatmap) is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.

Linear regression : Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Scatter Plot And Linear Regression : A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

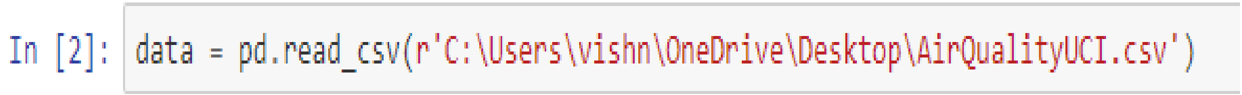
Exploratory Analysis:

Importing libraries:



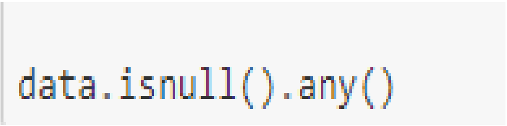
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Reading a file from the data set:

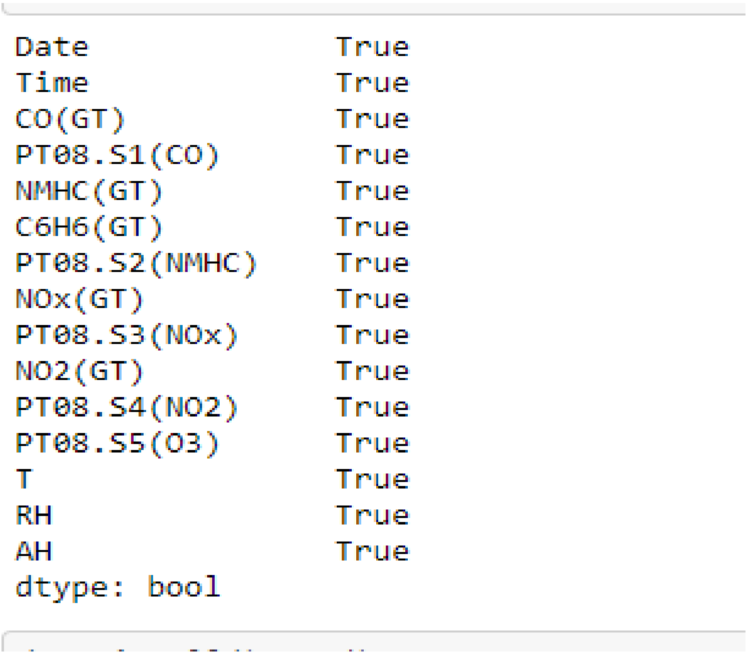


```
In [2]: data = pd.read_csv(r'C:\Users\vishn\OneDrive\Desktop\AirQualityUCI.csv')
```

This Gives All the Null Values in Data:

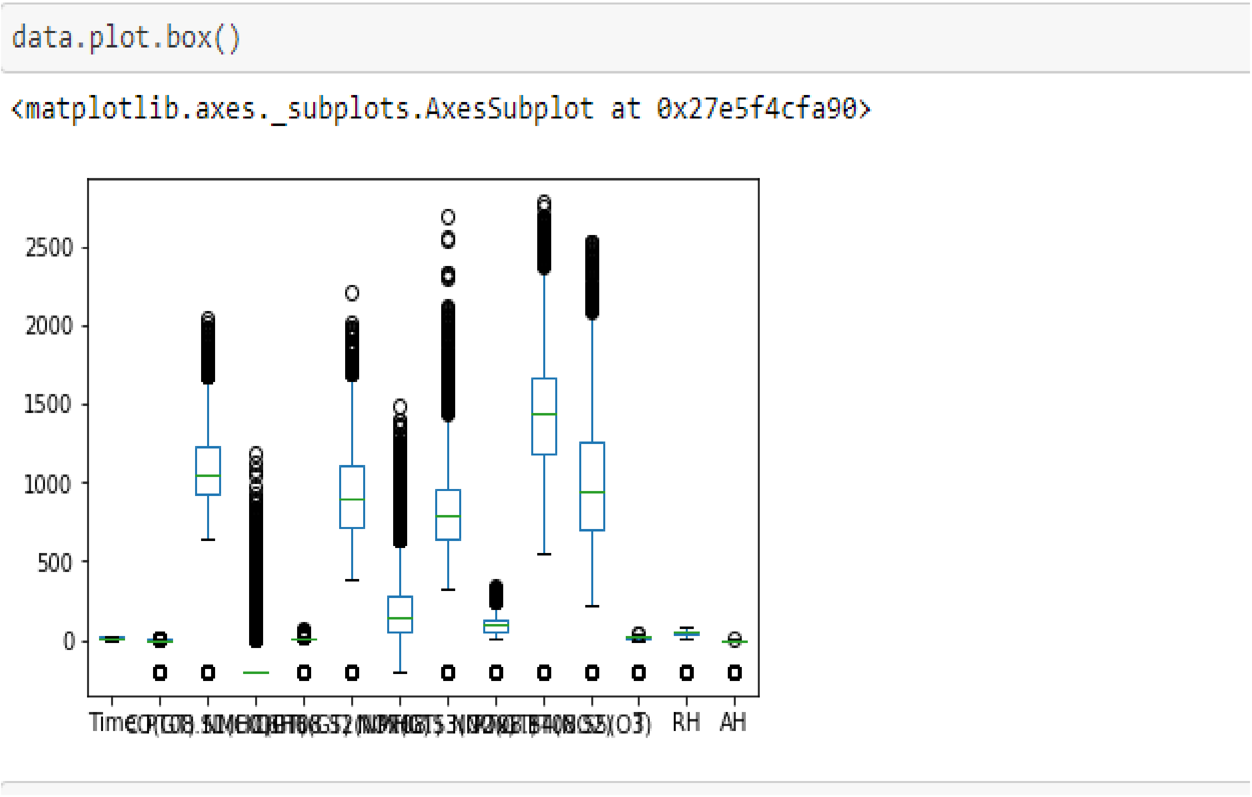


```
data.isnull().any()
```



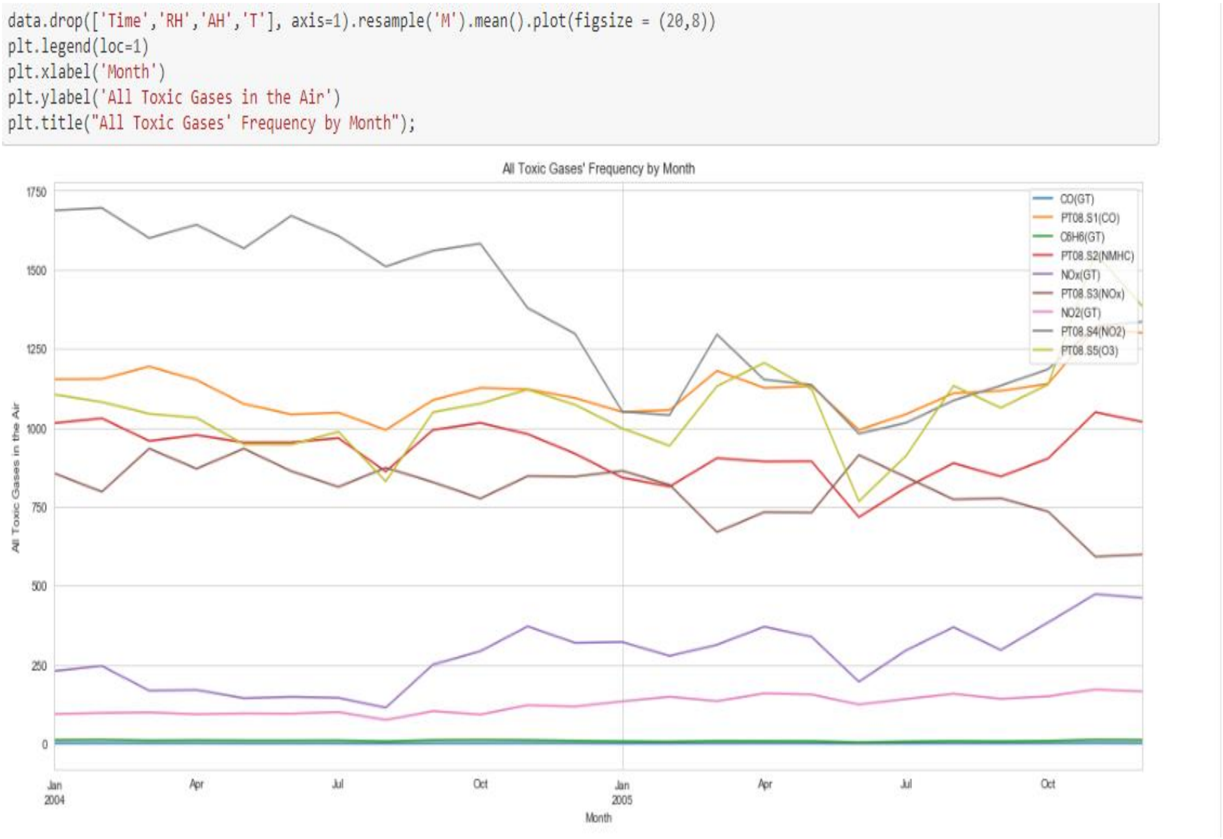
```
Date              True
Time              True
CO(GT)            True
PT08.S1(CO)       True
NMHC(GT)          True
C6H6(GT)          True
PT08.S2(NMHC)     True
NOx(GT)           True
PT08.S3(NOx)      True
NO2(GT)           True
PT08.S4(NO2)      True
PT08.S5(O3)       True
T                 True
RH                True
AH                True
dtype: bool
```

A boxplot is a graph that gives you a good indication of how the values in the data are spread out:



```
data.plot.box()
```

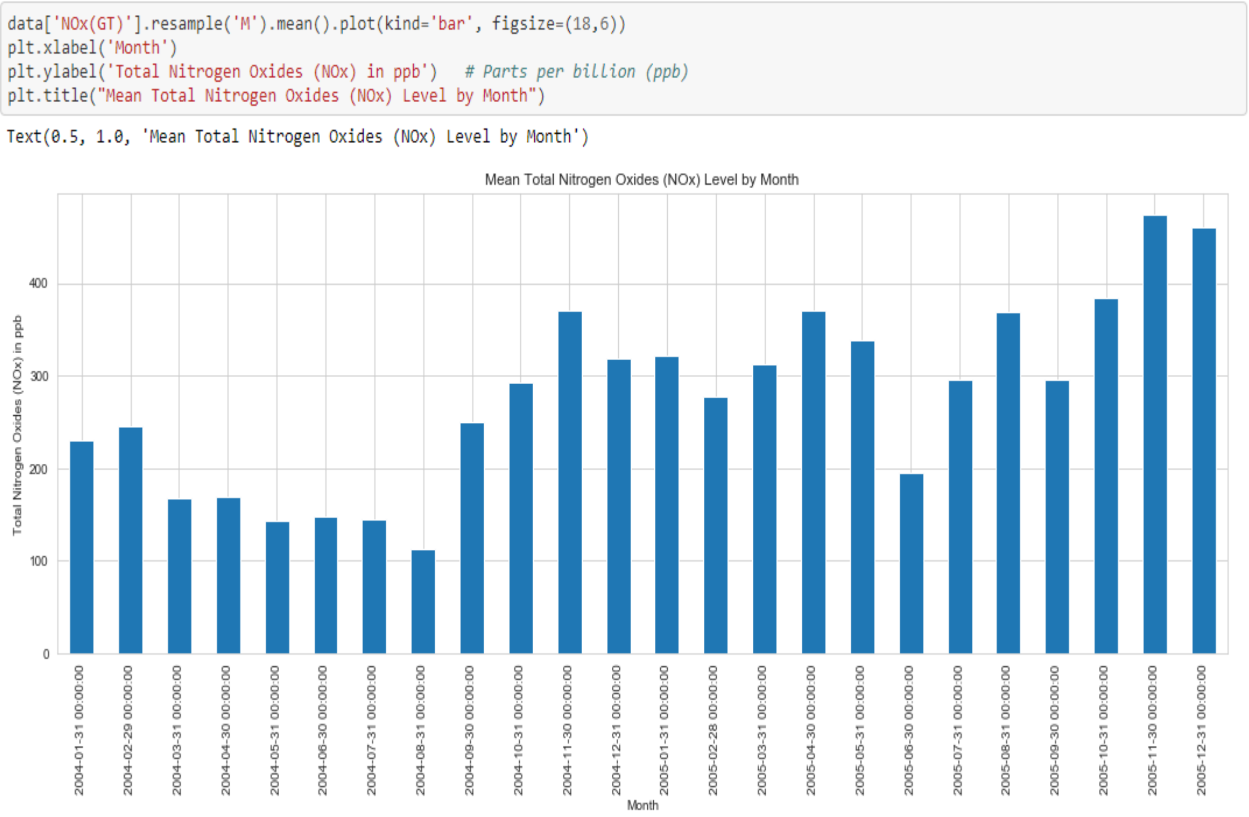
```
<matplotlib.axes._subplots.AxesSubplot at 0x27e5f4cfa90>
```



```
data.drop(['Time','RH','AH','T'], axis=1).resample('M').mean().plot(figsize = (20,8))
plt.legend(loc=1)
plt.xlabel('Month')
plt.ylabel('All Toxic Gases in the Air')
plt.title("All Toxic Gases' Frequency by Month");
```

- In the above graph, you can see the frequency of all toxics that is usually in polluted air. The Brown line shows Nitrogen Oxides (NOx) and Yellow line shows NO2 which is part of NOx. It is a mixture of gases are composed of nitrogen and oxygen. Two of the most toxicologically significant compounds are nitric oxide (NO) and nitrogen dioxide (NO2). We chose **Nitrogen Oxide (NOx)** because these are one of the most dangerous forms of air pollution and are most relevant for air pollution. However, there are many others ways to measure air pollution, including PM10 (particulate matter around between 2.5 and 10 microns in diameter), carbon monoxide, sulfur dioxide, nitrogen dioxide, ozone (O3), etc.

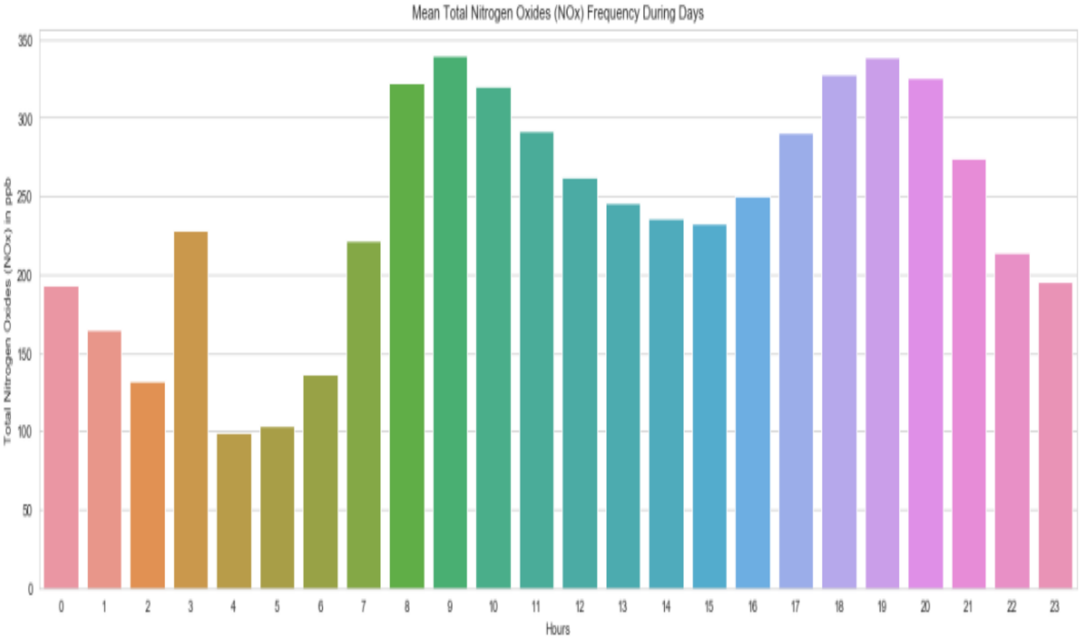
- NOx is produced from the reaction of nitrogen and oxygen gases in the air during combustion, especially at high temperatures. In areas of high motor vehicle traffic, such as in large cities, the amount of nitrogen oxides emitted into the atmosphere as air pollution can be significant.



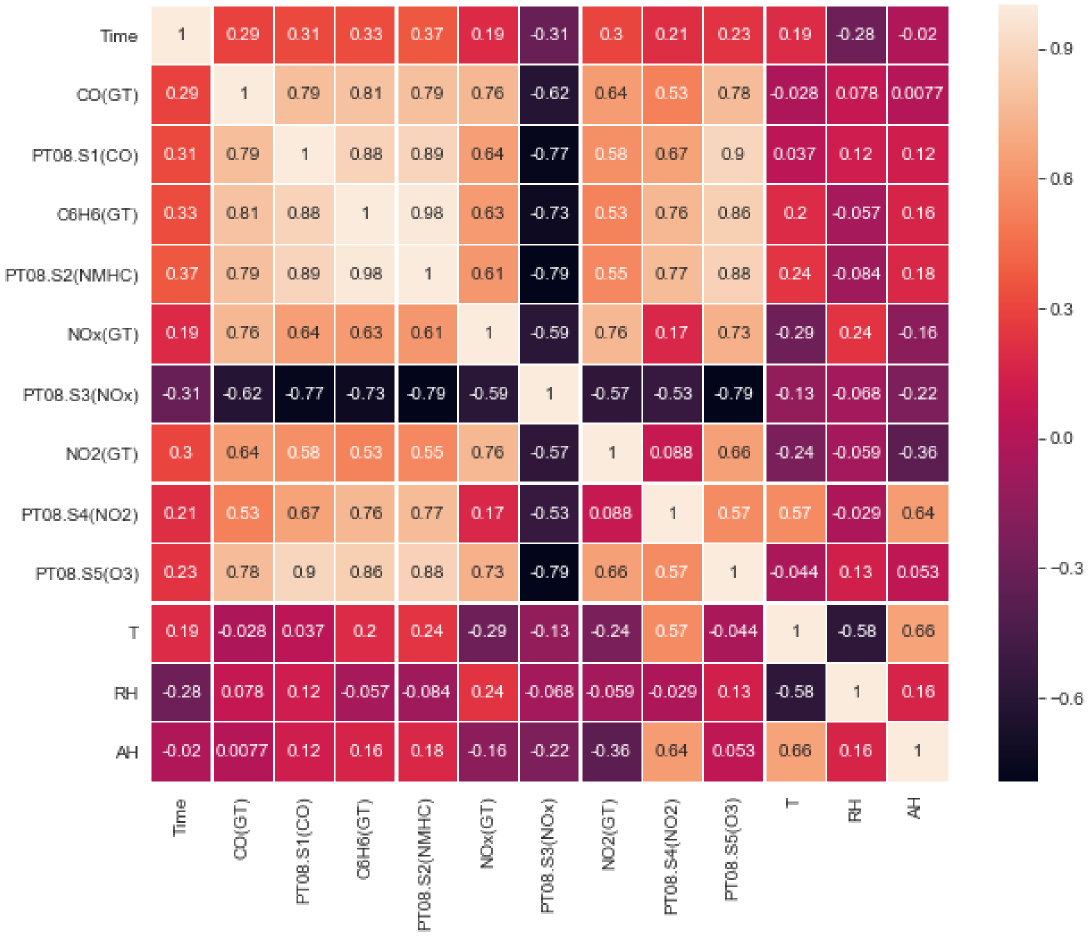
```
data['NOx(GT)'].resample('M').mean().plot(kind='bar', figsize=(18,6))
plt.xlabel('Month')
plt.ylabel('Total Nitrogen Oxides (NOx) in ppb')    # Parts per billion (ppb)
plt.title("Mean Total Nitrogen Oxides (NOx) Level by Month")
```

```
Text(0.5, 1.0, 'Mean Total Nitrogen Oxides (NOx) Level by Month')
```

In the above graph, we can see that frequency of Oxides of Nitrogen is increasing with little changes.



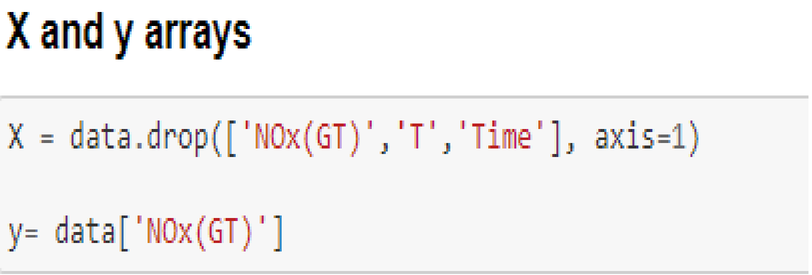
Here, the graph shows an average of Oxides of Nitrogen level with hours. It seems during the day, its level is high compared to night because of high use of transportations, phones, other electronics etc.



	Time	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
Time	1	0.29	0.31	0.33	0.37	0.19	-0.31	0.3	0.21	0.23	0.19	-0.28	-0.02
CO(GT)	0.29	1	0.79	0.81	0.79	0.76	-0.62	0.64	0.53	0.78	-0.028	0.078	0.0077
PT08.S1(CO)	0.31	0.79	1	0.88	0.89	0.64	-0.77	0.58	0.67	0.9	0.037	0.12	0.12
C6H6(GT)	0.33	0.81	0.88	1	0.98	0.63	-0.73	0.53	0.76	0.86	0.2	-0.057	0.16
PT08.S2(NMHC)	0.37	0.79	0.89	0.98	1	0.61	-0.79	0.55	0.77	0.88	0.24	-0.084	0.18
NOx(GT)	0.19	0.76	0.64	0.63	0.61	1	-0.59	0.76	0.17	0.73	-0.29	0.24	-0.16
PT08.S3(NOx)	-0.31	-0.62	-0.77	-0.73	-0.79	-0.59	1	-0.57	-0.53	-0.79	-0.13	-0.068	-0.22
NO2(GT)	0.3	0.64	0.58	0.53	0.55	0.76	-0.57	1	0.088	0.66	-0.24	-0.059	-0.36
PT08.S4(NO2)	0.21	0.53	0.67	0.76	0.77	0.17	-0.53	0.088	1	0.57	0.57	-0.029	0.64
PT08.S5(O3)	0.23	0.78	0.9	0.86	0.88	0.73	-0.79	0.66	0.57	1	-0.044	0.13	0.053
T	0.19	-0.028	0.037	0.2	0.24	-0.29	-0.13	-0.24	0.57	-0.044	1	-0.58	0.66
RH	-0.28	0.078	0.12	-0.057	-0.084	0.24	-0.068	-0.059	-0.029	0.13	-0.58	1	0.16
AH	-0.02	0.0077	0.12	0.16	0.18	-0.16	-0.22	-0.36	0.64	0.053	0.66	0.16	1

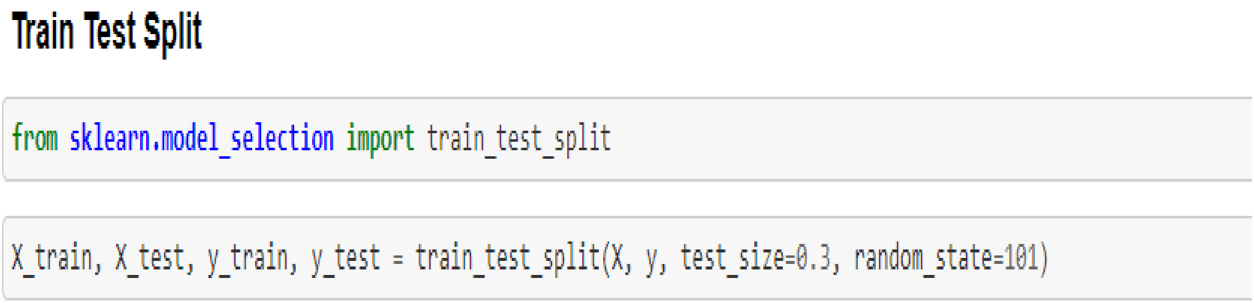
Basically, Heatmap shows 2d data in a graphical format. Each data value represents in a matrix and it has a special colour

Training a Linear Regression Model:



X and y arrays

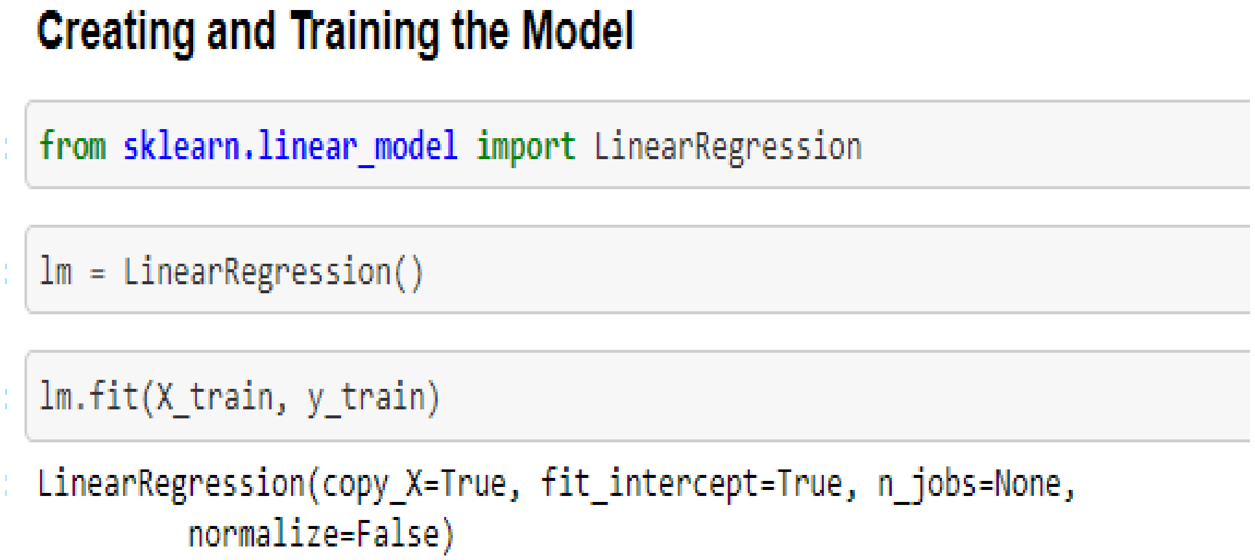
```
X = data.drop(['NOx(GT)','T','Time'], axis=1)

y= data['NOx(GT)']
```



Train Test Split

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
```



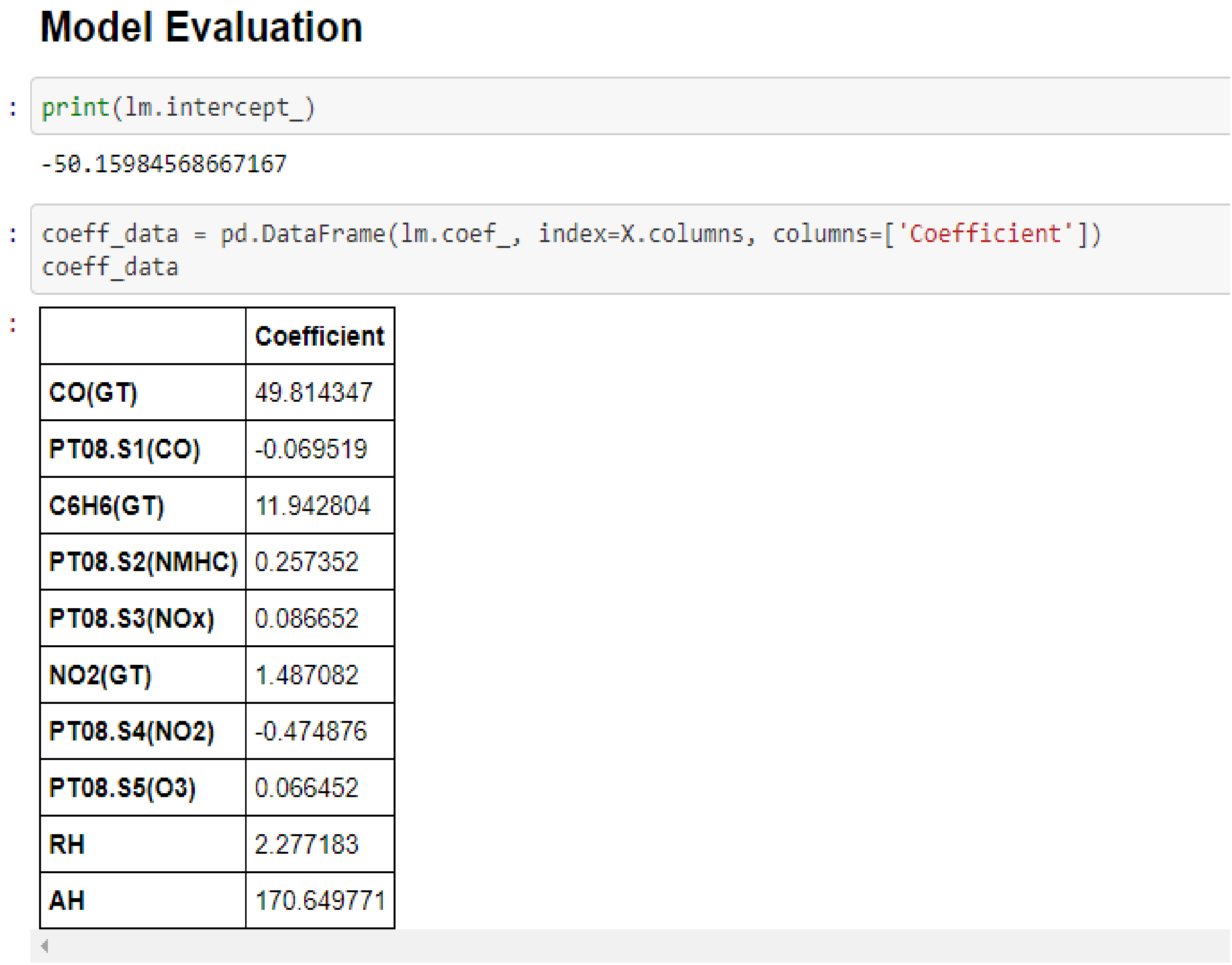
Creating and Training the Model

```
from sklearn.linear_model import LinearRegression
```

```
lm = LinearRegression()
```

```
lm.fit(X_train, y_train)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
         normalize=False)
```



Model Evaluation

```
: print(lm.intercept_)
```

```
-50.15984568667167
```

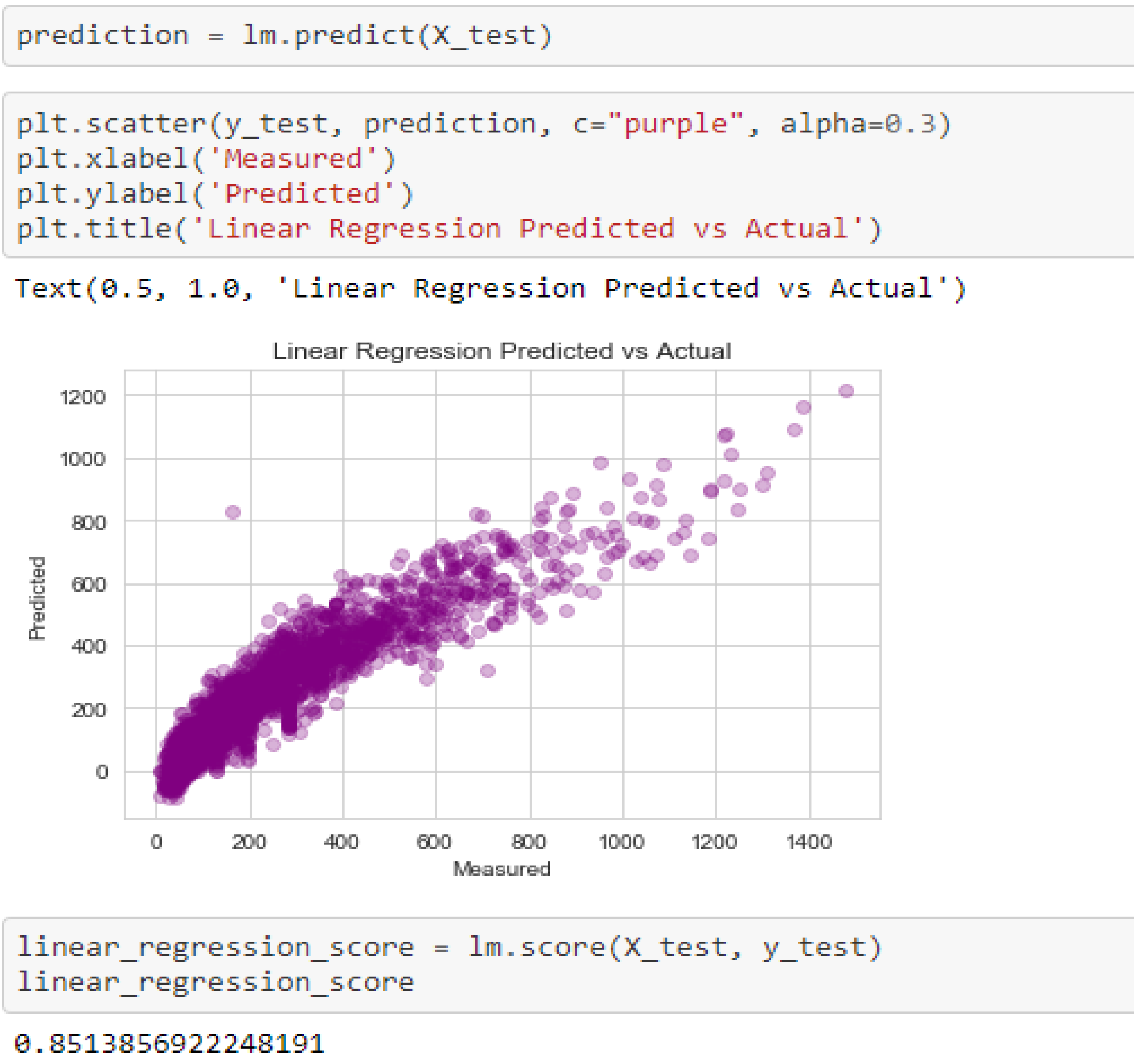
```
: coeff_data = pd.DataFrame(lm.coef_, index=X.columns, columns=['Coefficient'])
  coeff_data
```

:

	Coefficient
CO(GT)	49.814347
PT08.S1(CO)	-0.069519
C6H6(GT)	11.942804
PT08.S2(NMHC)	0.257352
PT08.S3(NOx)	0.086652
NO2(GT)	1.487082
PT08.S4(NO2)	-0.474876
PT08.S5(O3)	0.066452
RH	2.277183
AH	170.649771

From above coefficient values, we can say: if 1 unit increases in Benzene (C6H6), NOx increases by 12.66. Same as, if 1 unit increases in Nitrogen Dioxide(NO2) and Relative Humidity(RH), Oxides of Nitrogen will increase by 1.32 points and 2.49 points, respectively.

Prediction Model:



```
prediction = lm.predict(X_test)
```

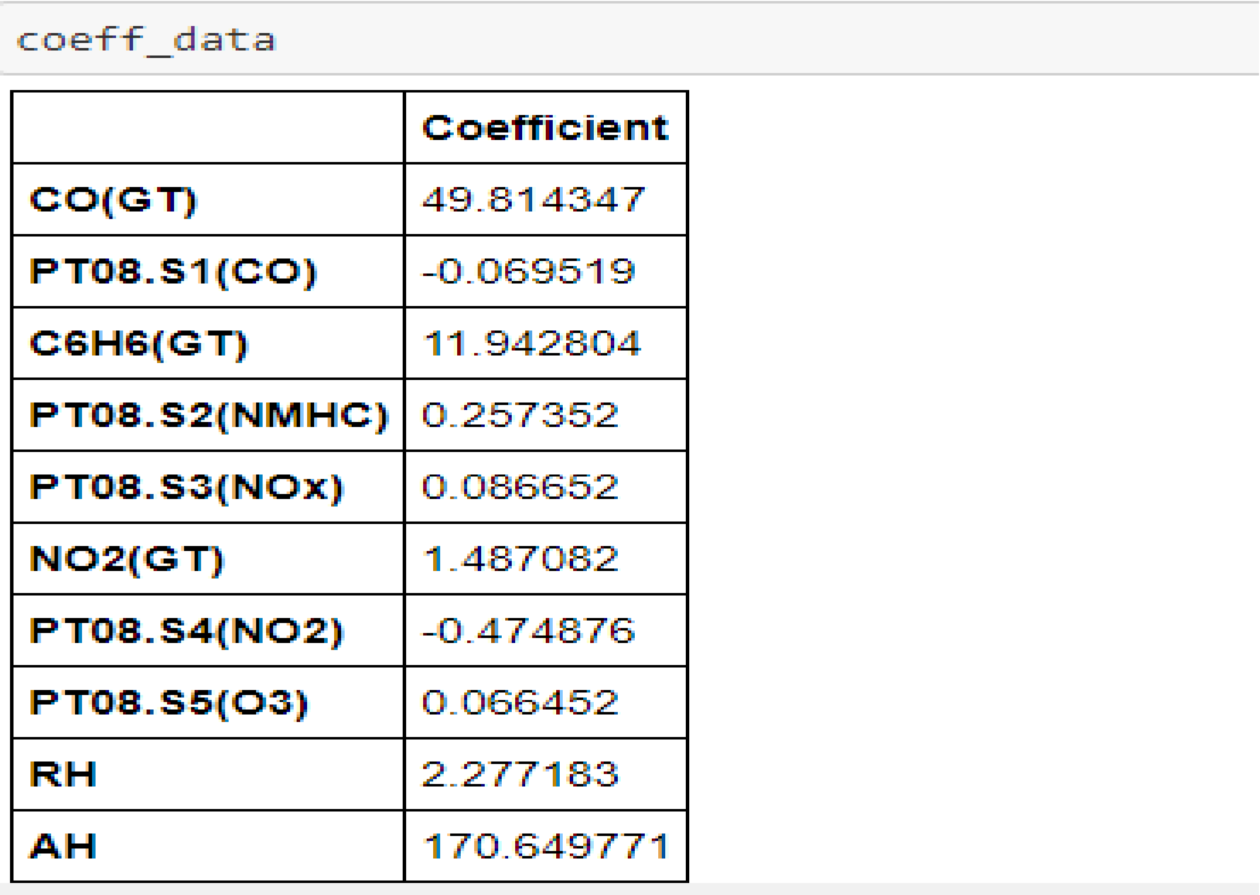
```
plt.scatter(y_test, prediction, c="purple", alpha=0.3)
plt.xlabel('Measured')
plt.ylabel('Predicted')
plt.title('Linear Regression Predicted vs Actual')
```

```
Text(0.5, 1.0, 'Linear Regression Predicted vs Actual')
```

```
linear_regression_score = lm.score(X_test, y_test)
linear_regression_score
```

```
0.8513856922248191
```

Prediction Score is good which is 85.



```
coeff_data
```

	Coefficient
CO(GT)	49.814347
PT08.S1(CO)	-0.069519
C6H6(GT)	11.942804
PT08.S2(NMHC)	0.257352
PT08.S3(NOx)	0.086652
NO2(GT)	1.487082
PT08.S4(NO2)	-0.474876
PT08.S5(O3)	0.066452
RH	2.277183
AH	170.649771

If we hold all other variables constant and 1 point increases in CO(GT), NOx will increase by 49.81. Similarly, If we hold all other variables constant and 1 point increases in NO2(GT), NOx will increase by 1.48. and, If we hold all other variables constant and 1 point increases in C6H6(GT), NOx will increase by 11.94.

Conclusion:

For this Air quality data analysis, we saw that NOx's ppb are increasing due to the air pollution causing factors as mentioned above and badly affects our health and environment and also some initiatives should be taken like Re-Circulating flue gas which a waste gas produced at the power station, Water Injection and Water Emulsion, in which water is added to reduce temperature of the combustion before it becomes too dangerous for us.

References:

- https://www.pollutiononline.com/doc/nox-emission-reduction-strategies-0001
- https://www.marineinsight.com/tech/10-technologiesmethods-for-controlling-nox-sox-emissions-from-ships/