ST306 - Mini Project Report

S18822

January 30, 2024

1 Introduction

Air pollution, is one of the outstanding health concerns in today's world affecting both developed and developing countries at both social and economic level. It accounts for an estimated 6.7 million premature death annually worldwide and also a root cause for many diseases including heart disease, stroke, lung cancer, pneumonia, type 2 diabetes, neonatal disorders, mental health conditions and neurological impairment.

Pollution often takes the form of chemical composition of particulate matter (PM2.5 and PM10), nitrogen oxide (NO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3) where an air quality index (AQI) is a scale used to show how polluted the air is with these pollutants, along with the risks associated with each rating. These pollutants are released to atmosphere due to various activities such as domestic solid biomass energy use, exhaust and non–exhaust emissions from vehicles, industrial emissions, and burning of solid waste.

This project aims to analyze and discover the trends in air quality across London city.

1.1 Description of DataSet

Two datasets named "london local data 2022" [dataset 1] and "london local sites" [dataset 2] were provided for the project as csv files.

The first dataset contains hourly measurement of NO_2, NO, NO_x, O_3, SO_2,$PM10$, $PM2.5$ collected at 34 air monitoring sites located in London from 01/01/2022 to 01/01/2023. It had a size of 289,069 rows x 10 columns and contained following variables of importance.

Name	**Description**
site	name of the monitoring site the data was obtained from
code	A unique identifier for the monitoring site the data was obtained from.
date	The date and timestamp when the observation was made.
no	The measured value of the NO pollutant
nox	The measured value of the NO_x pollutant
no2	The measured value of the NO_2 pollutant
pm10	The measured value of the $PM10$ pollutant.
o3	The measured value of the O_3 pollutant.
pm2_5	The measured value of the $PM2.5$ pollutant.
so2	The measured value of the SO_2 pollutant.

Table 1: description of the fields for dataset 1

The details of each monitoring site are given in the second dataset which include the following variables.

Name	**Description**
site	name of the monitoring site
code	A unique identifier for the monitoring site .
Latitude	Latitude of the site
Longitude	Longitude of the site
parameter name	name of the substance measured

Table 2: description of the fields for dataset 2

It is noted that not all monitoring sites measured all pollutants.The following table summarises the no.of monitoring sites that measured each pollutant.

Pollutant	NOx	NOx	NO2	PM10	O3	PM2.5	SO2
No. of Monitoring Sites	33	33	33	23	3	3	1

Table 3: Number of monitoring sites considered for each pollutant measurement

For the ease of the analysis, the date column in the dataset 1 was divided into two columns using *mutate()* function to create two new variables for the date and time of the measurement .Thus the resulting dataset has 12 columns.

1.2 Air quality Assessment

Air quality is assessed based on a banding system which measures the levels of pollutants such as Ozone (O_3), Nitrogen dioxide (NO_2) and Particulate matter ($PM10$ and $PM2.5$) . The overall air quality index at any particular time is given as the maximum band for any pollutant.

UK air quality banding system issued by the Committee on Medical Effects of Air Pollutants (COMEAP) specifies the following values for the different bands of pollutants.

Band	**Index**	O_3	NO_2	$PM2.5$	$PM10$
Low	1-3	0-100	0-200	0-35	0-50
Moderate	4-6	101-160	201-400	36-53	51-75
High	7-9	161-240	401-600	54-70	76-100
Very High	10	241 or more	601 or more	71 or more	101 or more

Table 4: air quality

This index sets short-term thresholds for each of main pollutants measured in the UK and grades current air quality into four 'bands'; Low, Moderate, High and Very High. Each of these bands has separate 'index' levels to provide gradation between the bands.The index runs on a scale from 1-10

2 Literature Review

Ninsiima et al. (2023) assessed the spatio–temporal trends in air quality in Kampala City during January 2020 –June 2022 using PM2.5 concentrations.They used line graphs to visualize the trend of 24 hour average and hourly average PM 2.5 concentrations.The seasonal Mann–Kendall statistical test was applied to assess the significance of observed trends by quarterly periods.The findings in the study showed that air quality exceeds the WHO targeted safe level throughout the day, even during times of less traffic and economic activities in Kampala City.

Regarding the studies focusing on air pollution and health effects in London, we should mention the study of Gil-Alana, Yaya, and Carmona-González (2020) who examined the air quality in London by providing evidence of persistence, seasonality and time trends in various air quality pollutants using roadside and background standard air quality chemistry readings.The results of the study demonstrated a large degree of heterogeneity across pollutants and a persistent behavior based on a long memory pattern practically in all cases. Seasonality and decreasing linear trends are also found in some cases.

Sampath (2019) analyzed and compared the performance of classification and regression in the field of air quality prediction.He experimented with several models which can predict PM2.5 levels using present and historical pollution data in combination with predicted weather data and their performance was successfully evaluated.The exploratory data analysis and feature engineering methods implemented for the prediction models revealed interesting correlations between weather and pollution data.Moreover,different approaches to handle null values yielded varied performance from each of the models.The study concluded that classifier models perform better for air quality prediction than regression model.

3 Results and Discussion

Figure 1 demonstrates the distribution of missing values for the numerical variables of the data set 1.

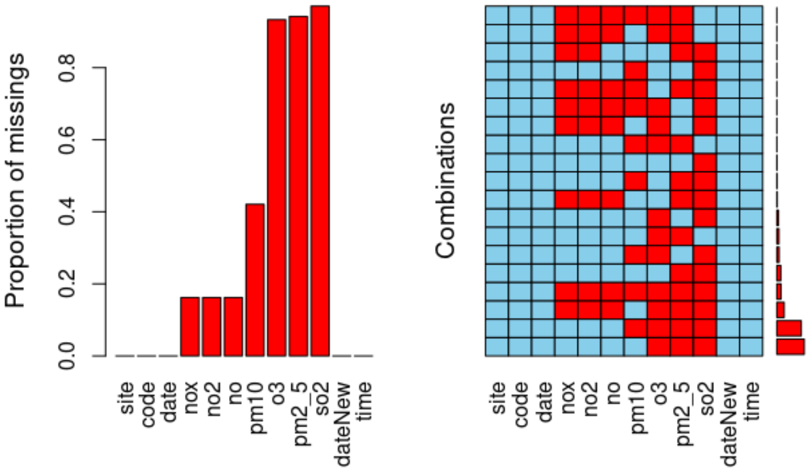

Figure 1: Proportion of missing values

First, let us investigate the **temporal behavior** of some air pollutants across London city.

Since the proportion of missing values for the parameters `nox`,`no2` and `no` is less than 20%, the missing values were removed to obtain the following time series visualizations for the 3 pollutants.

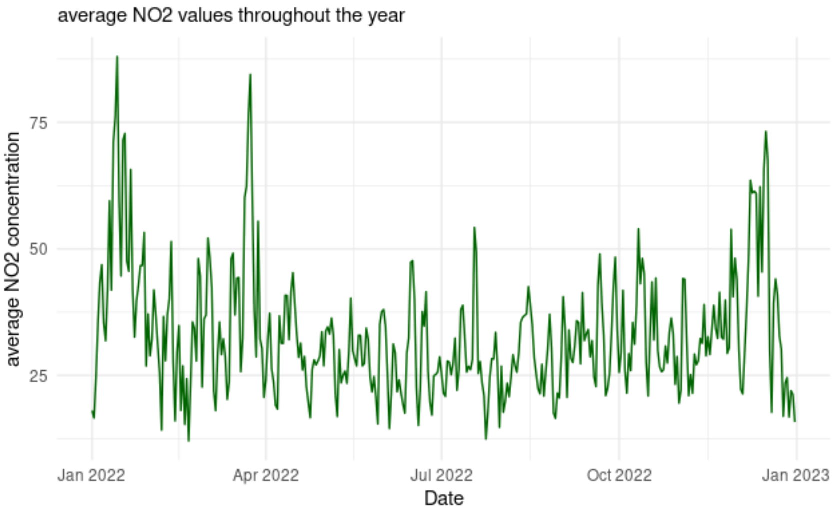

Figure 2: Average NO_2 distribution

The trend of average NO_2 concentration in London city over the year doesn't indicate any significant increase or decrease, rather goes relatively stable overtime. The highest NO_2 concentration is recorded in month of January (2022-01-14) and some significant spikes are recorded during the end of the year (in December) as well. Moreover,all values are below 200 thus signifying that the air quality with respect to NO_2 is at a satisfactory level.

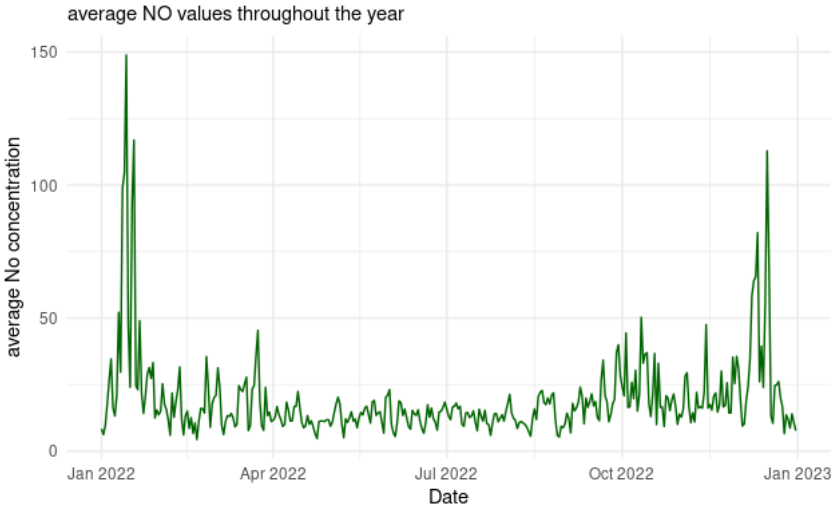

Figure 3: Average NO distribution

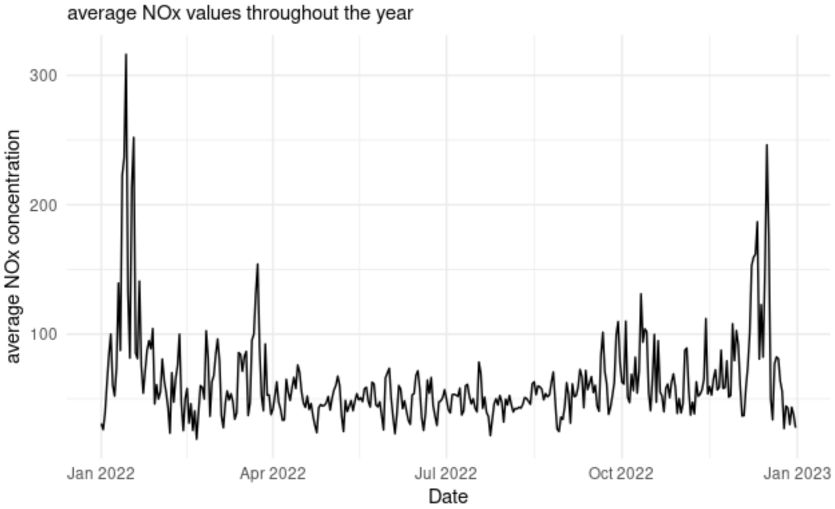

Figure 4: Average NO_x distribution

For the above plots in Figure 3 and 4 ,the highest concentration of the pollutant can be detected in the month of January(2022-01-14) and few more high concentrations in December.

From figure 1, we can see that columns like `O3`, `so2` and `pm25` have over 80% missing values. This could be because all monitoring sites did not measure all pollutants of interest. In this analysis, for these pollutants,only the relevant sites are considered and missing values are imputed using interpolation.

PM2.5 refers to fine particles with a diameter of 2.5 micrometers or smaller. It is a critical measure of air pollution as these particles can penetrate deep into the respiratory system.

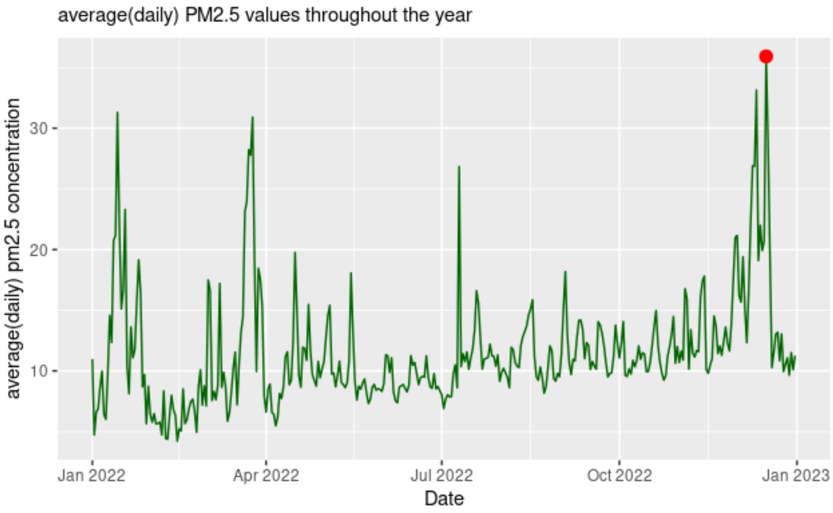

Figure 5: Average $PM2.5$ distribution

2022-12-16 records the highest average $PM2.5$ concentration for the year which pushes the level of concern to moderate which is still acceptable.However,adults and children with lung problems and heart problems might be at risk when engaging in strenuous physical activity,particularly outdoors.Moreover a slight increasing trend can also be observed.

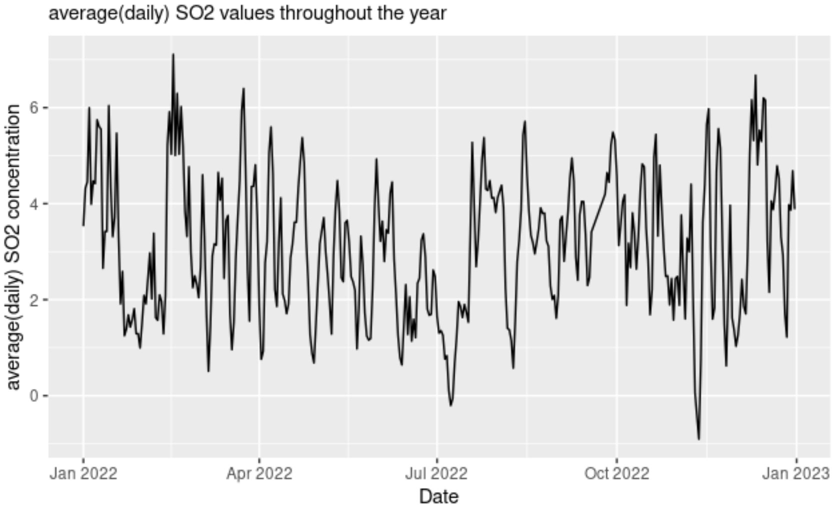

Figure 6: Average SO_2 distribution

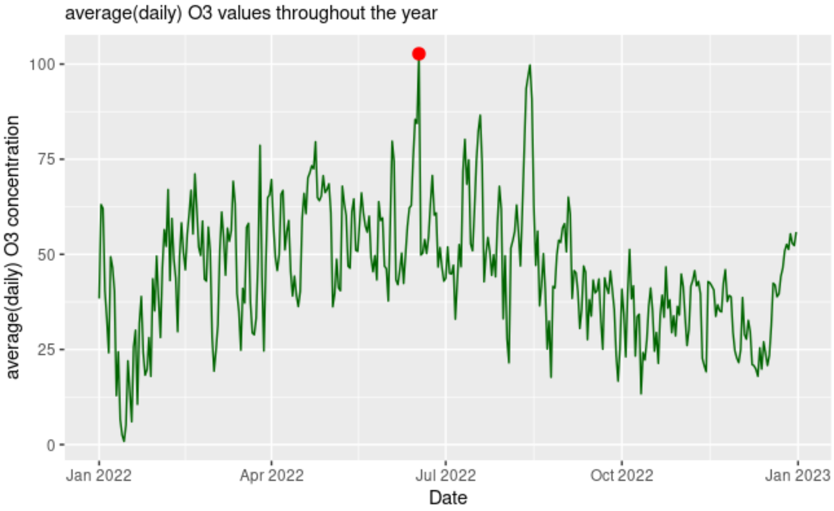

Figure 7: Average O_3 distribution

2022-06-17 records the highest O_3 concentration which falls to the moderate index of air quality risk where Adults and children with lung problems, and heart condition might be at risk. Moreover a non linear trend in O_3 can be observed.

Now, let us take a look at most polluted sites with respect to some pollutants

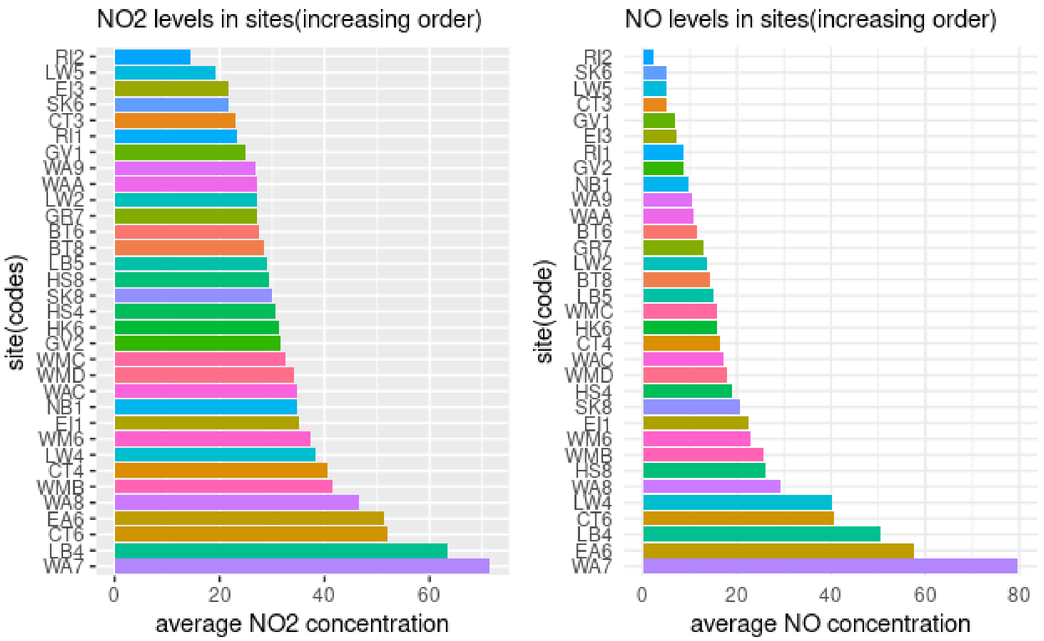

Figure 8: site-wide NO_2 and NO pollution

Wandsworth - Putney High Street(WA7) and Lambeth - Brixton Road(LB4) are top 2 polluted sites with respect to NO_2 where as WA7 and Ealing - Hanger Lane Gyratory(EA6) are the most polluted sites with respect to NO.
It is noted that Wandsworth - Putney High Street has recorded the highest concentration for both NO_2 and NO.

Let's take a closer look at the daily $PM10$ levels of Wandsworth - Putney High Street

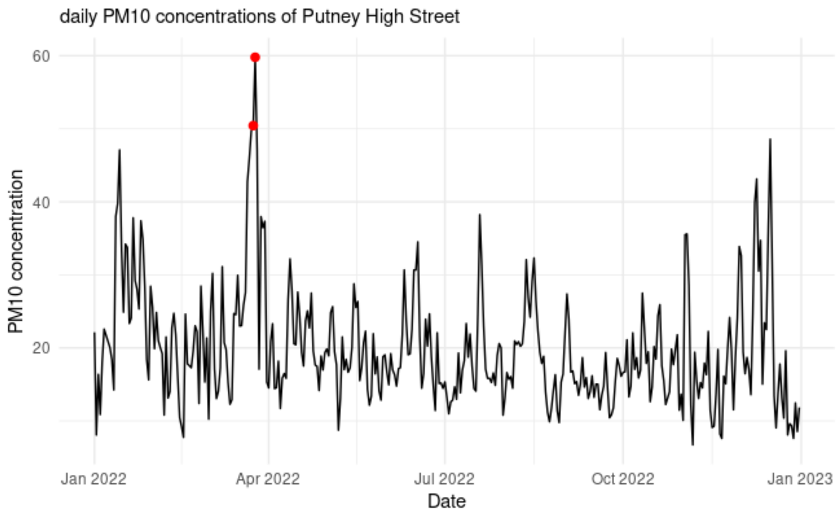

Figure 9: daily $PM10$ levels of Putney High Street

The plot indicates two time points which signifies a moderate the level of concern for air quality in March.

The following plots shows the most pollutant site with respect to the pollutants $PM2.5$ and O_3

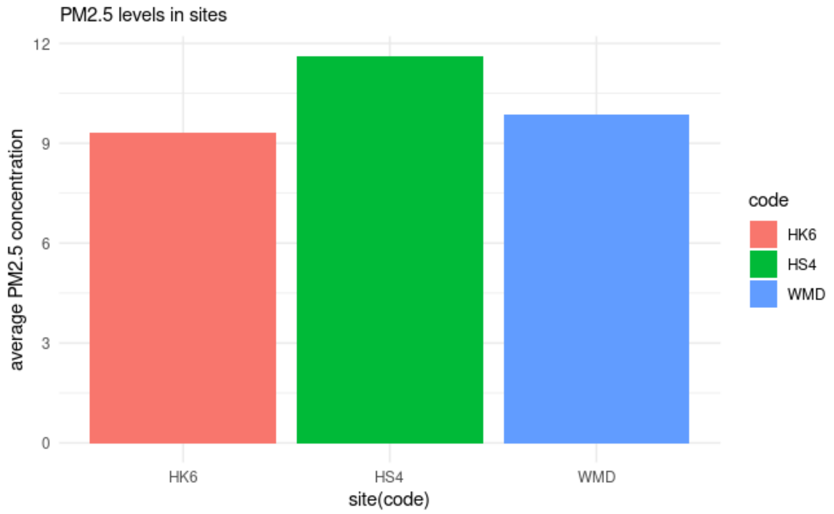

Figure 10: site-wide $PM2.5$ pollution

Out of the three sites which measured the $PM2.5$ levels in lodon city,Hounslow Chiswick (HS4) site records the highest pollution with respect to $PM2.5$

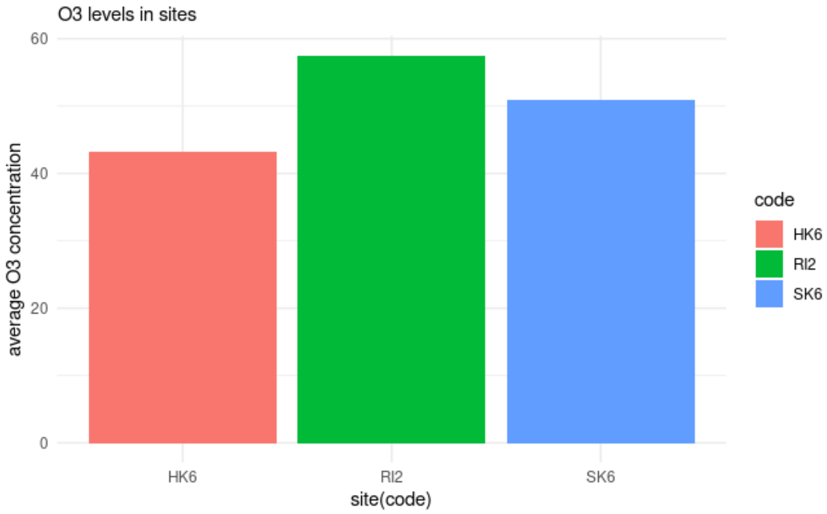

Figure 11: site-wide O_3 pollution

Out of the three sites which measured the O_3 levels in Lodon city,Richmond Upon Thames - Barnes Wetlands (RI2) site records the highest pollution with respect to O_3

The following figure visualize the correlation between pollutants of interest in London city

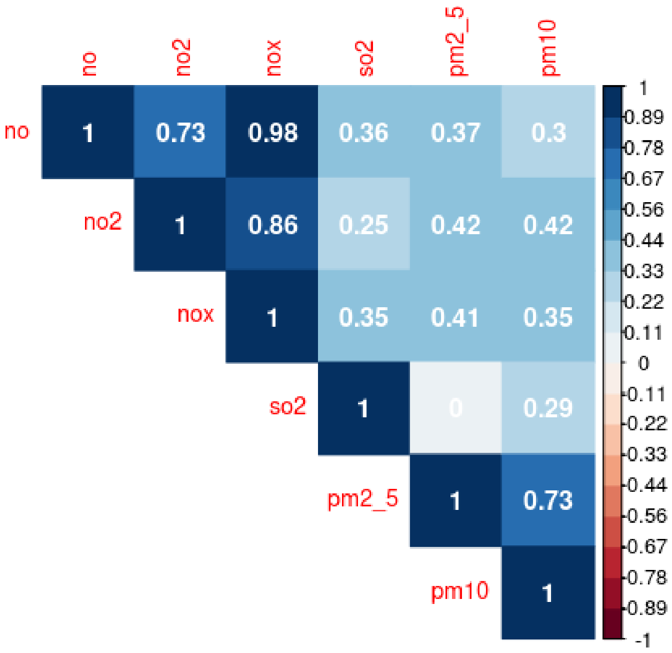

Figure 12: correlation matrix of pollutants

we can observe that there is a strong positive correlation (Correlation Coefficient greater than 0.8) between the pairs NO_x , NO and NO_x , NO_2.This could be because they have the same sources of pollution which is primarily from combustion of fuel. Moreover, similar patterns observed for time series visualizations for these pollutants in figure 3 and figure 4 might also be explained by this strong correlation.

4 conclusion

when analyzing the temporal behavior of the pollutants, we observed that majority of high concentrations are reported in the days of months in January, March and December.
According to $IQAir$ sources,most of London's air pollution comes from road transport, as well as domestic and commercial heating systems.Moreover the largest contributor of $PM2.5$ in cities is estimated to come from wood and coal heating. London's winter season typically happens November through February .During this time we could expect a surge in the use of domestic heating systems and hence a high emission of pollutants to the atmosphere.Therefore, this might be a reason for such patterns of high levels of pollutants across the city of London.

Furthermore,we observed that Wandsworth - Putney High Street recorded the highest concentration for both NO_2 and NO.
One reason for this could be the bus fleet travelling along Putney High Street.As a result, it remains very congested with a stop-start traffic flow, which is affected by vehicles delivering to businesses in the High Street.Moreover, a study commissioned by the Wandsworth Council showed that the buses were responsible for over two thirds of oxides of nitrogen pollution.