

**CMP7005: Programming for Data Analysis**

**From Data to Application Development: PRAC1**

**Name: Salman Irfan Ahmed Qureshi**

**ID: 20313619**

**Table of Contents**

[Introduction 2](#_Toc197955232)

[Task 1: Data Acquisition and Preparation 3](#_Toc197955233)

[Dataset Description 3](#_Toc197955234)

[Site Selection Rationale 3](#_Toc197955235)

[Data Importation and Merging 3](#_Toc197955236)

[Initial Data Overview 4](#_Toc197955237)

[Task 2: Exploratory Data Analysis 5](#_Toc197955238)

[Task 2a: Fundamental Data Understanding 5](#_Toc197955239)

[Task 2b: Data Preprocessing 5](#_Toc197955240)

[Task 2c: Statistical Summary and Visualisation 6](#_Toc197955241)

[Task 3: Machine Learning Modelling 15](#_Toc197955242)

[Task 4: Application Development 18](#_Toc197955243)

[Task 5: Version Control 20](#_Toc197955244)

[Conclusion 21](#_Toc197955245)

[References 22](#_Toc197955246)

# Introduction

Air pollution, a harmful mix of chemicals and fine particles, poses major risks to public health and the environment. However, the long-term exposure increases the chances of developing serious illnesses such as stroke, heart disease, and lung cancer. In China’s capital, Beijing, pollution remains a significant challenge despite recent improvements. Therefore, accurately predicting air quality is essential for timely health warnings and effective policy action. This report aims to analyse air pollution data from two Beijing monitoring sites Changping and Huairou and develop a machine learning model to predict PM2.5 concentrations, a pollutant known for its severe health impact. Additionally, the project involves creating an interactive, user-friendly GUI to present findings clearly. The main objectives include acquiring and cleaning multi-site air quality data, conducting in-depth Exploratory Data Analysis (EDA) to identify trends and correlations, and implementing best practices for model building. These are scaling, encoding, and feature selection. The machine learning model will focus on predicting hourly PM2.5 levels using meteorological and pollutant data collected between March 2013 and February 2017. The GUI will display raw data, EDA insights, and prediction results across multiple pages. Git and GitHub will be used to manage the project workflow and code versioning. This report is structured into key sections, which are data acquisition and preparation, EDA, model development and evaluation, GUI implementation, version control, and a conclusion summarising outcomes and suggestions for future research.

# Task 1: Data Acquisition and Preparation

## Dataset Description

The data utilised in this study originates from a publicly available dataset comprising hourly air pollutant concentrations and meteorological readings from 12 nationally controlled air-quality monitoring sites across Beijing. This comprehensive dataset, covering the period from March 1st, 2013, to February 28th, 2017, provides a substantial four-year window for analysis (Kumar and Pande, 2023). Air quality measurements, including Particulate Matter (PM2.5, PM10), sulfur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), and ozone (O3), were sourced from the Beijing Municipal Environmental Monitoring Center. Corresponding meteorological data, such as wind speed (Wspd), rainfall (Rain), temperature (Temp), dew point (Dewp), and atmospheric pressure (Pre), were matched from the nearest weather stations managed by the China Meteorological Administration. This rich combination of pollutant and weather data allows for an in-depth investigation into the factors influencing air quality.

## Site Selection Rationale

The assessment brief recommended selecting sites representing diverse typologies, which are urban, suburban, rural, and industrial or hotspots for a comprehensive analysis. For this project, data from two specific monitoring sites were chosen. However, these are Changping and Huairou. The Changping station is typically representative of a suburban or developing urban fringe area. Hence, it experiences influences from both urban and local emission sources, which reflects a transition zone. However, the Huairou station is situated in a district known for its more rural characteristics and natural landscapes. This generally further removed from dense urban pollution, thus serving as a comparative background or cleaner air reference site (Gu *et al.*, 2022). This selection allows for an insightful comparative analysis of air quality dynamics between an area undergoing significant development. Moreover, a relatively less impacted region addressed the site types.

## Data Importation and Merging

The initial stage of data handling involved importing the selected datasets into the Python development environment using the pandas library. However, it is a powerful and widely used tool for data manipulation and analysis in Python. Specifically, the read\_csv() function was employed to load the data from “PRSA\_Data\_Changping\_20130301-20170228.csv” and “PRSA\_Data\_Huairou\_20130301-20170228.csv” into individual pandas DataFrames (Kirwa *et al.*, 2021). To facilitate a unified analysis across both selected sites, these two DataFrames were subsequently concatenated along the row axis using the pd.concat() function. The ignore\_index being True parameter was set during this concatenation to ensure a new, continuous numerical index for the combined DataFrame. Hence, it prevents potential issues arising from duplicated indices from the original files.

## Initial Data Overview

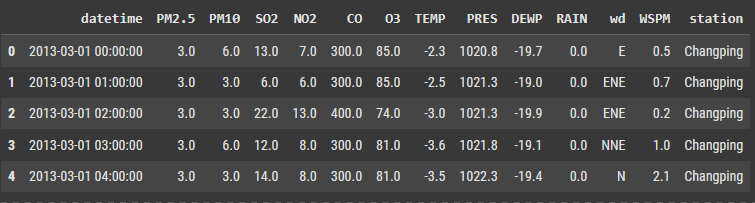


Figure 1: Data head

(Source: Self-Created)

Following the successful merging process, an initial inspection of the combined DataFrame was conducted. This step served to verify the correct amalgamation of data from both Changping and Huairou sites and to gain a preliminary understanding of the resulting data structure (Fig. 1). The head method in pandas was utilised to display the first few rows of the merged DataFrame. This confirmed the presence of all expected columns, including pollutant concentrations, meteorological variables, and station identifiers, and showcased the newly formed continuous index. This quick overview ensured that the data was loaded as anticipated and was structurally sound for the subsequent, more detailed stages of preprocessing and exploratory data analysis.

# Task 2: Exploratory Data Analysis

## Task 2a: Fundamental Data Understanding

The dataset includes thousands of records with several columns representing air pollutants and weather-related factors. A quick look at the structure indicates numeric and text-type columns. Most key pollutants like PM2.5, PM10, and CO show a wide range of values, with some very high spikes (Bekkar *et al.*, 2021). Some data is missing, but not too much. However, the impact is likely small. There are no repeated rows, which is great for analysis. A method was used to find very high or low values, which are outliers, and some pollutants like PM10 and CO showed more of these. Overall, the data looks mostly clean and ready for deeper study.

## Task 2b: Data Preprocessing

To prepare the data, rows with missing values were removed to keep things simple and clean. Enough data remained to continue analysis without problems. A new column was created to combine the year, month, day, and hour, which helps when looking at trends over time (Liao *et al.*, 2021). Some extra columns, such as row numbers and the separate date parts, were removed to reduce clutter. The columns were also reordered for better readability, putting the new date column at the front. The final cleaned data looks neat, with fewer rows but still enough information to carry out proper analysis and draw insights.

## Task 2c: Statistical Summary and Visualisation

***Univariate Analysis***

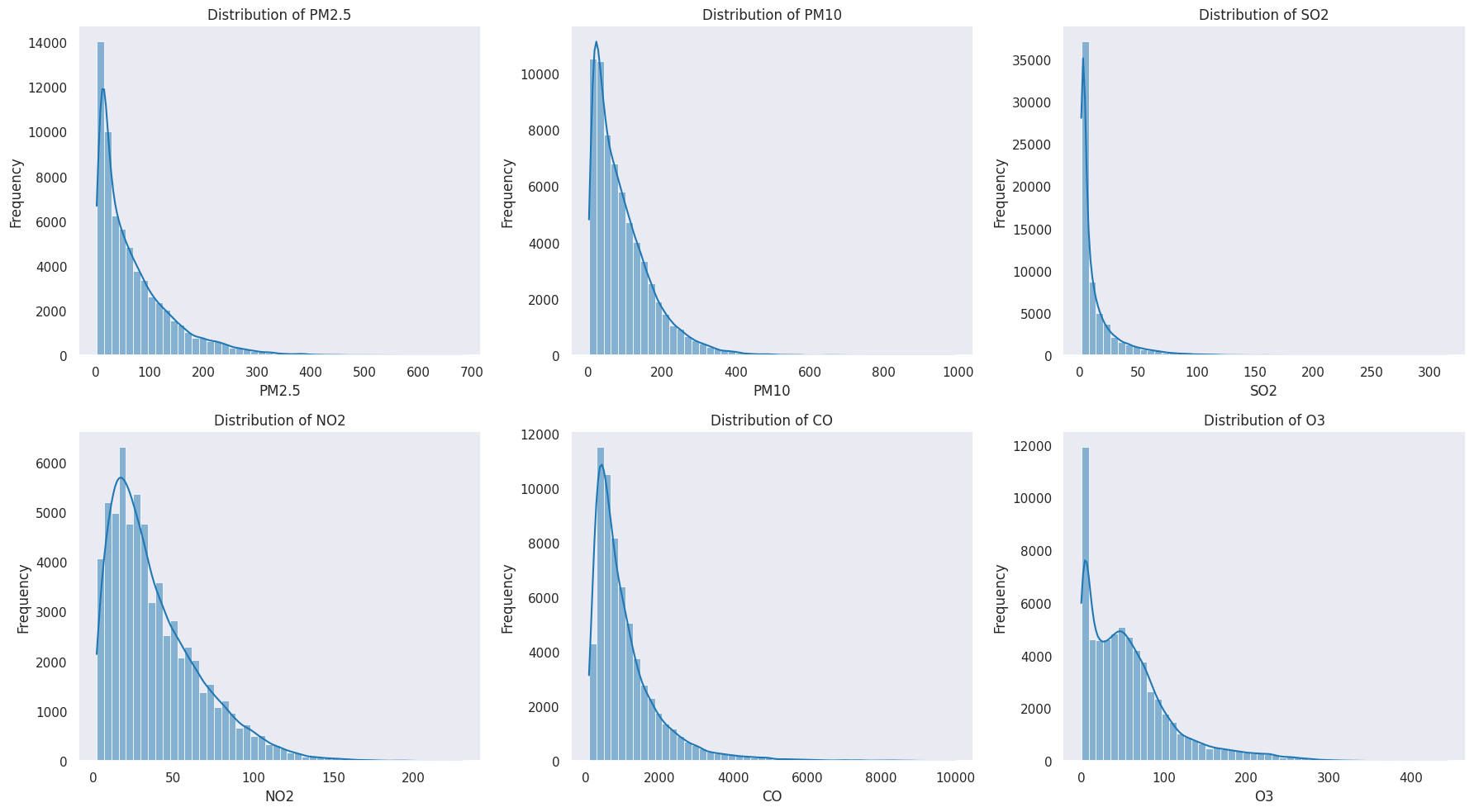


Figure 2: Pollutants distribution

(Source: Self-Created)

The histograms (Fig. 2) illustrate the distribution of six key air pollutants, which are PM2.5, PM10, SO2, NO2, CO, and O3. A general trend observed across most pollutants is right-skewness, indicating that higher pollutant concentrations are relatively rare. Pollutants such as PM2.5, PM10, SO2, and CO display a steep decline in their distributions, which suggests that lower values are more common and dominate the dataset. In contrast, NO2 exhibits a broader spread with a gradual peak between 20 and 30 units. Notably, ozone (O3) stands out with a bimodal distribution. One at low values and another in the 50-100 range. This implies variability in its formation and accumulation patterns.

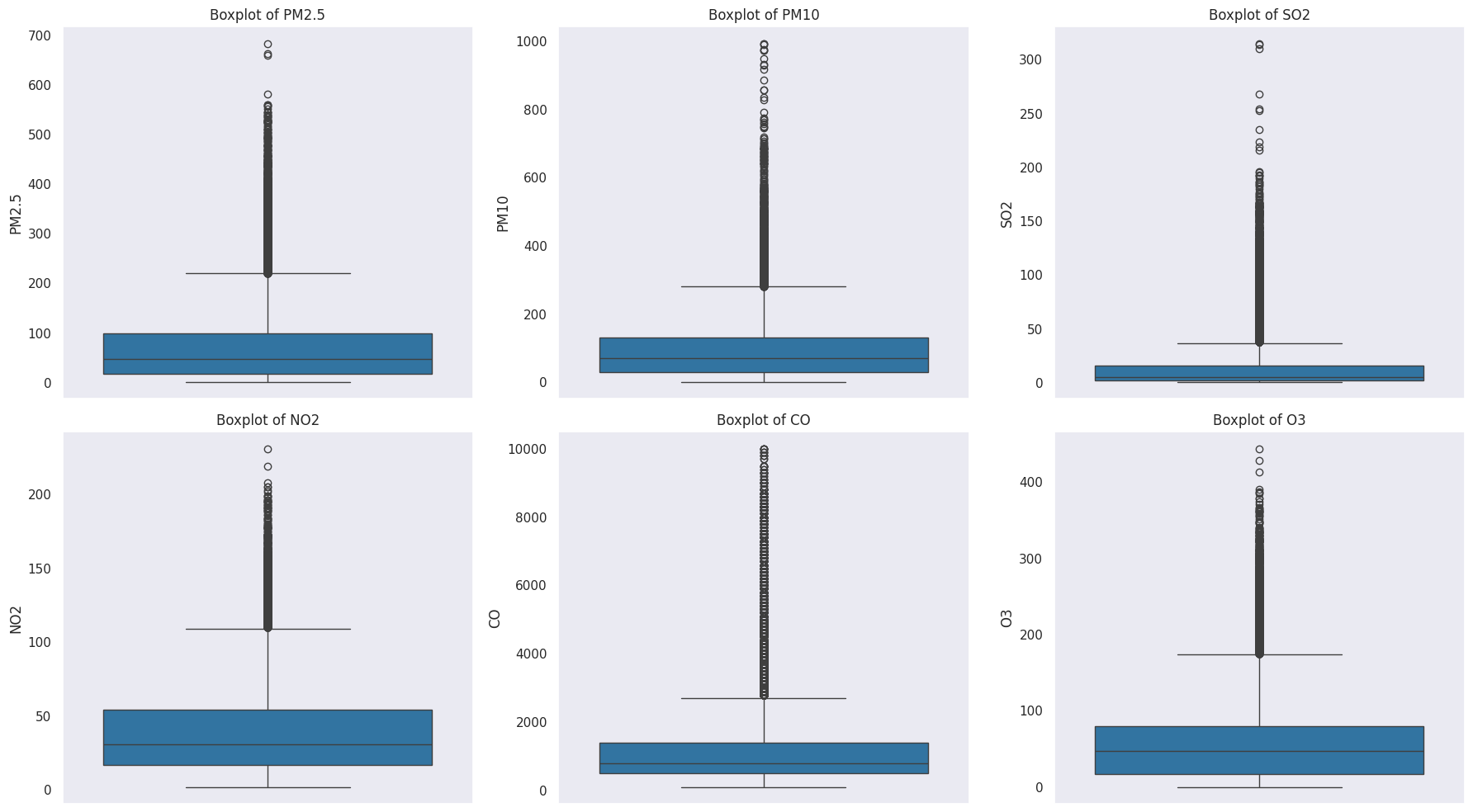


Figure 3: Pollutants box plot

(Source: Self-Created)

The box plots (Fig. 3) further support the right-skewness observed in the histograms. PM2.5, PM10, SO2, and CO all show positive skew. Several outliers are evident, especially for PM2.5 and PM10. The interquartile range (IQR) is widest for PM2.5 and PM10. In contrast, NO2 and O3 have medians situated closer to the centre of the boxes and relatively shorter whiskers. This suggests that NO2 and O3 exhibit more stable trends compared to other pollutants (Fig. 3).

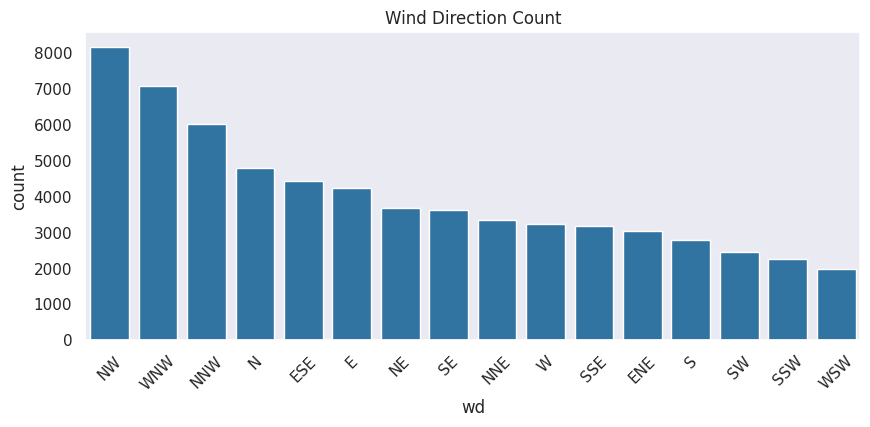


Figure 4: Countplot of wind direction

(Source: Self-Created)

Wind direction also plays a role in pollution patterns, as shown in the count plot (Fig. 4). The most frequent wind directions are NW, WNW, and NNW. However, the least common are from the south and southwest that are S, SW, SSW and WSW.

***Bivariate Analysis***

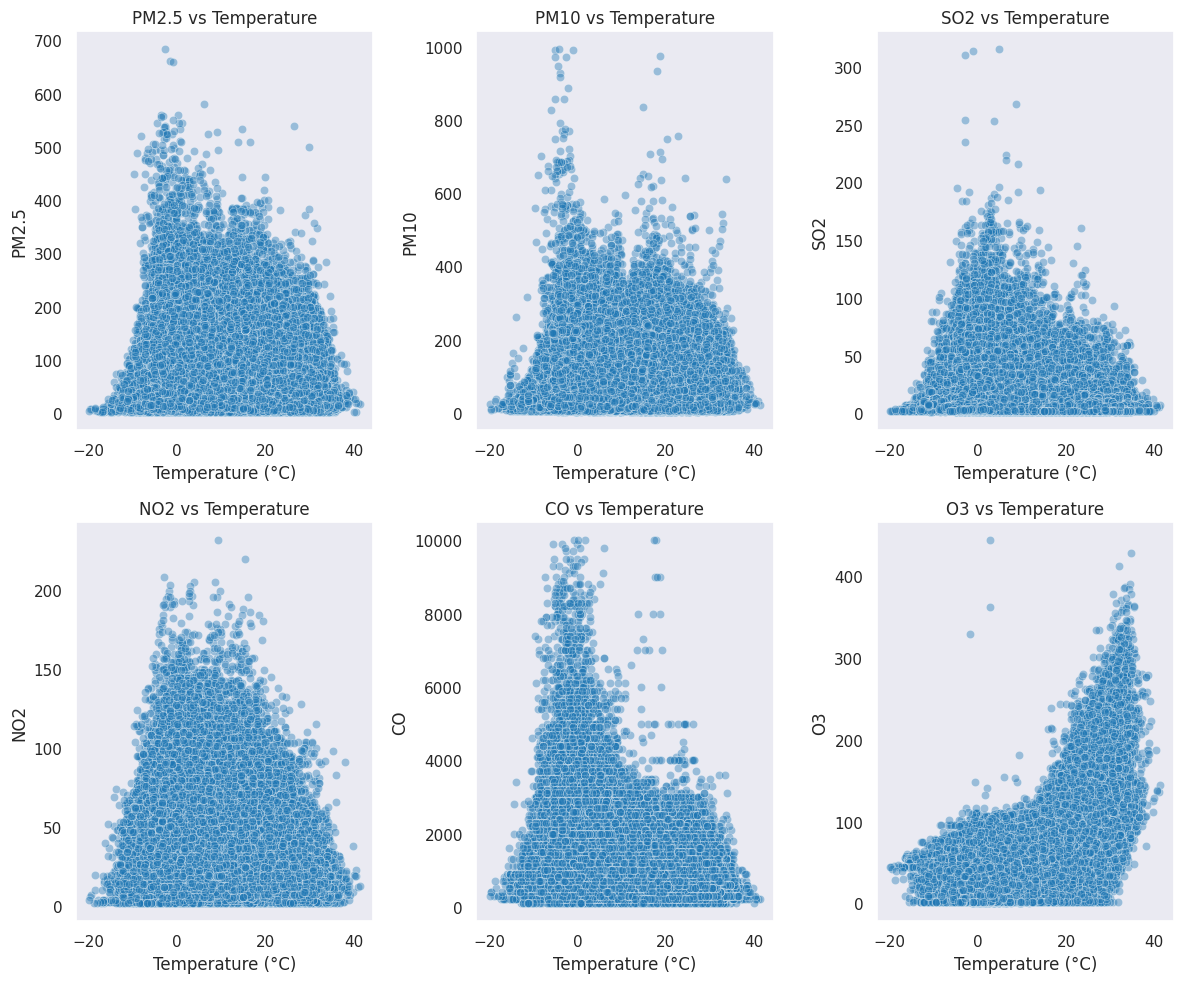


Figure 5: Pollutants vs. temperature

(Source: Self-Created)

When examining pollutant concentrations in relation to temperature (Fig. 5), a strong inverse relationship is apparent for PM2.5, PM10, NO2, SO2, and CO. These tend to decrease as temperature rises. These pollutants also form dense clusters at lower temperatures, particularly PM2.5, PM10, and CO. This indicates higher pollutant build-up in colder weather. O3 behaves differently. This points towards complex chemical reactions such as photochemical smog formation. PM10 and CO show a high degree of variability, implying they are influenced by several external environmental or human factors. NO2 and SO2, on the other hand, show a steadier and more gradual decline (Fig. 5).

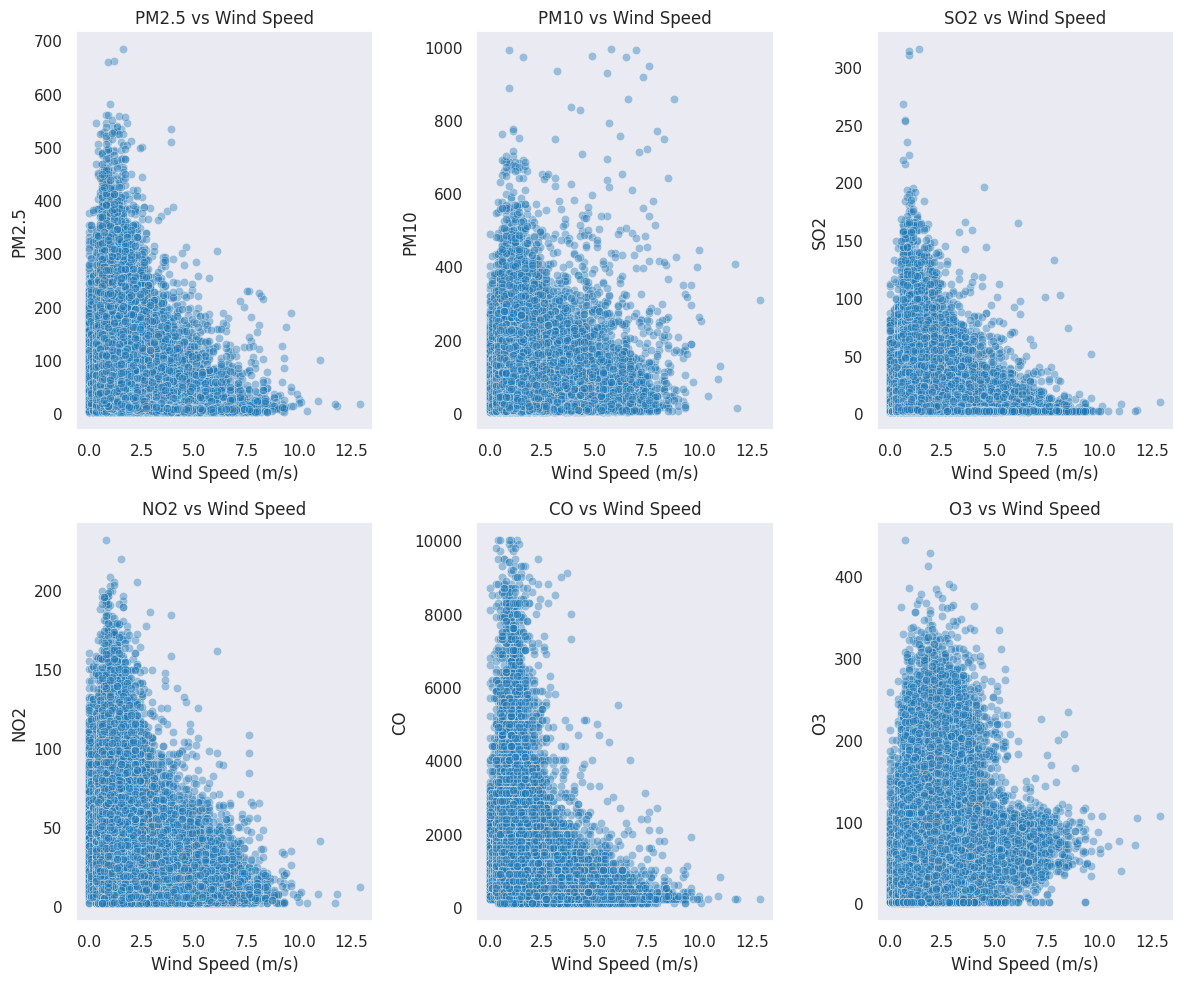


Figure 6: Pollutants vs. wind speed

(Source: Self-Created)

A similar trend is observed in the analysis of pollutants against wind speed (Fig. 6). PM2.5, PM10, NO2, SO2, and CO generally decrease as wind speed increases. Again, O3 is an exception, showing a different pattern likely due to its unique formation mechanisms. Dense clustering at lower wind speeds such as for PM2.5, PM10, and CO. Nonlinear declines for NO2 and SO2 reinforce the observation that these pollutants react more steadily to changes in environmental variables. PM10 and CO continue to demonstrate high variability, likely due to mixed pollution sources or environmental influences (Fig. 6).

***Multivariate Analysis***

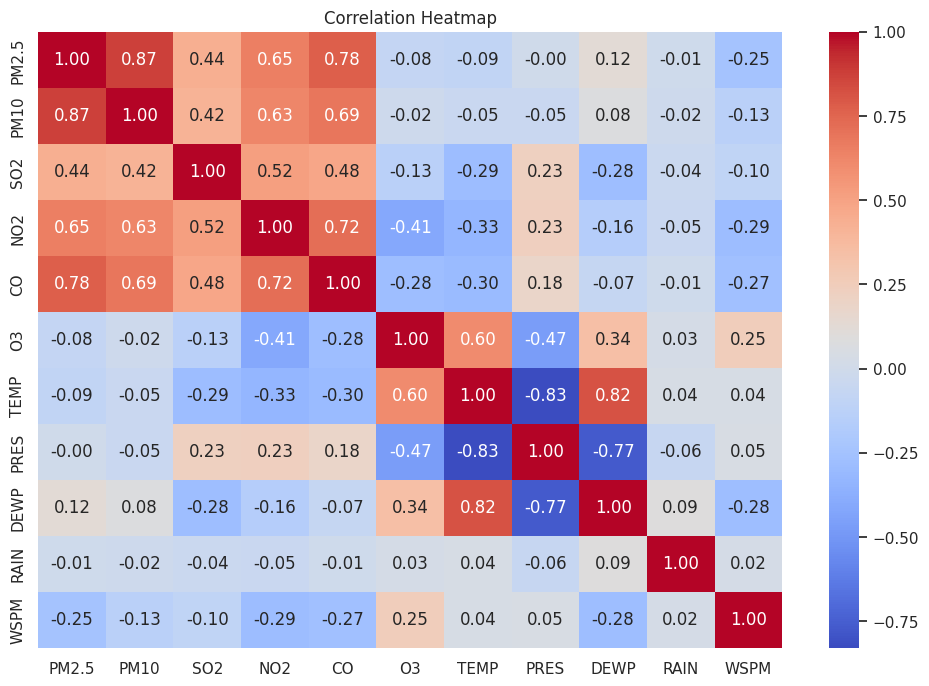


Figure 7: Correlation heatmap

(Source: Self-Created)

The correlation heatmap (Fig. 7) provides a deeper insight into interrelationships among the pollutants and meteorological variables. PM2.5, PM10, SO2, NO2, and CO exhibit a strong negative correlation with temperature. This affirms the earlier observation that these pollutant levels tend to decrease as temperatures rise. In contrast, O3 shows a positive correlation with temperature. PM2.5 and PM10 are highly correlated with each other, likely indicating shared emission sources such as dust or combustion by-products. CO and NO2 also display a strong correlation, suggesting common origins like vehicular emissions. SO2 shows a moderate correlation with PM2.5 and PM10 (Fig. 7).

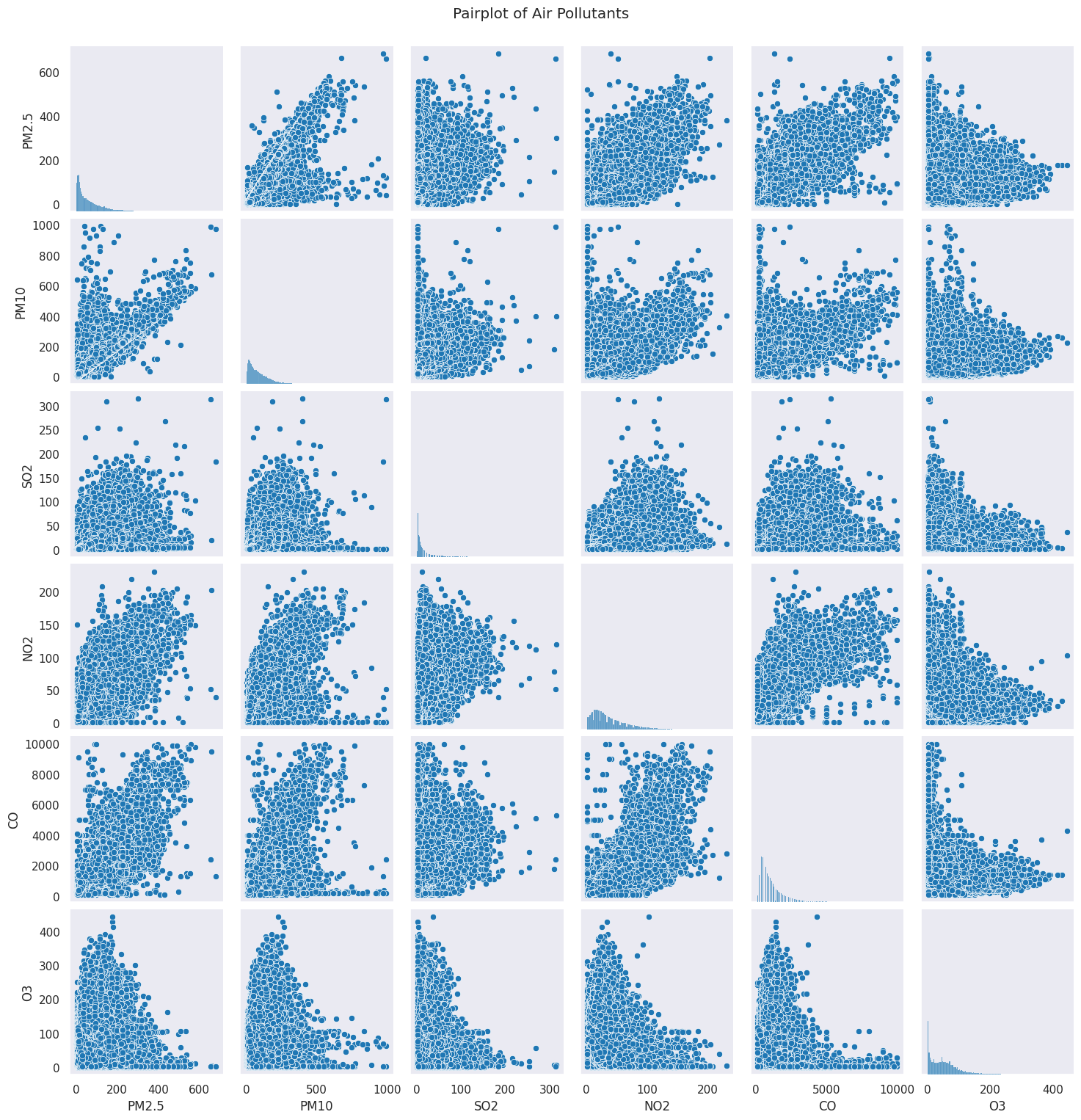


Figure 8: Pairplot of pollutants

(Source: Self-Created)

The pair plot (Fig. 8) visually reinforces the relationships explored in the heatmap and scatter plots. PM2.5, PM10, NO2, SO2, and CO all show inverse relationships with temperature, with dense clustering at lower temperature ranges. O3 again deviates, rising with temperature and peaking in the mid-range before tapering off. NO2 and SO2 reflect stable, gradual trends, while PM10 and CO reveal widespread scattering (Fig. 8).

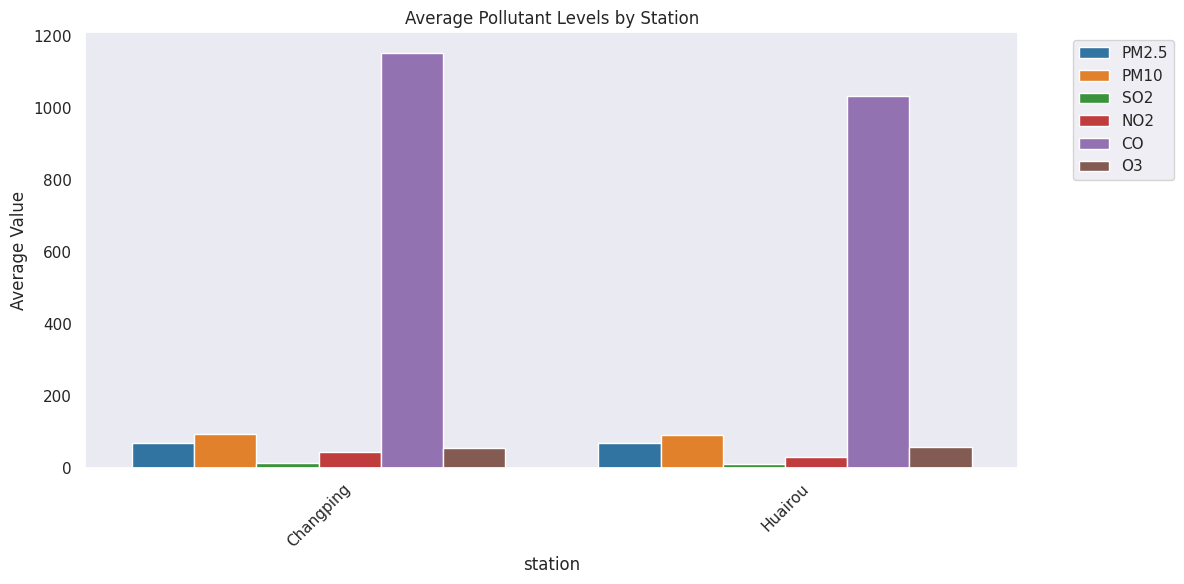


Figure 9: Average pollutant levels per station

(Source: Self-Created)

An analysis of average pollutant levels across monitoring stations (Fig. 9) shows that pollutant concentration at the Changping and Huairou stations are generally similar, which is suggesting uniformity in pollution levels across different locations within the study area (Fig. 9).

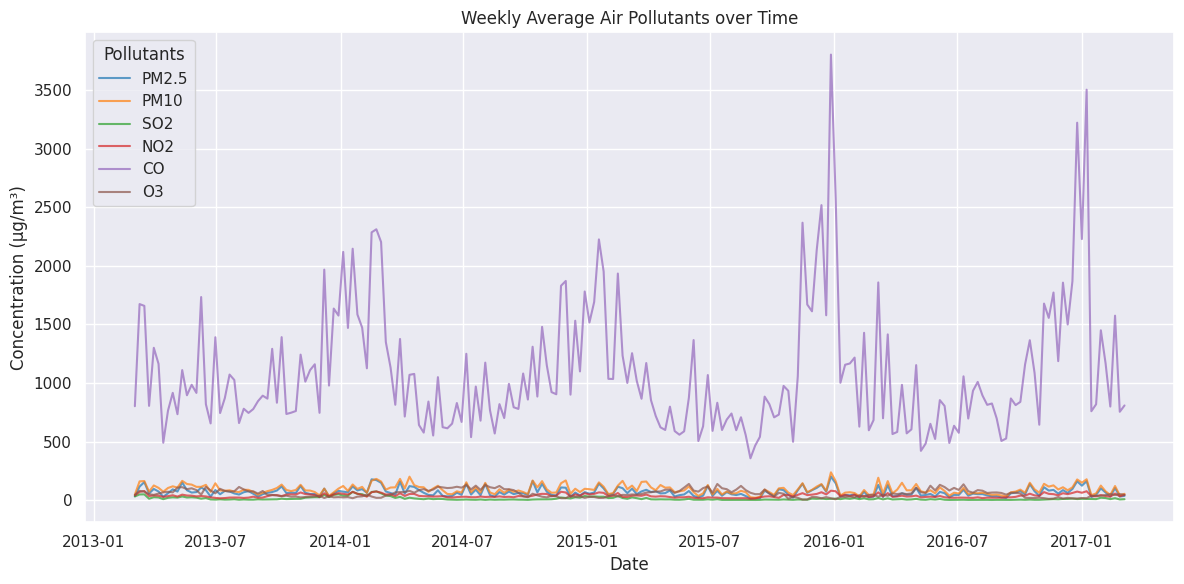


Figure 10: Weekly average pollutants over time

(Source: Self-Created)

Finally, the weekly average pollutants plotted over time (Fig. 10) further corroborate earlier findings. PM2.5, PM10, NO2, SO2, and CO show a downward trend with rising temperatures. O3 continues to rise with temperature. However, the tight clusters for PM2.5, PM10, and CO at lower temperatures reinforce their tendency to build up under cold weather. NO2 and SO2 follow gradual, predictable patterns, while PM10 and CO remain highly variable, reflecting sensitivity to various environmental and anthropogenic factors (Fig. 10).

# Task 3: Machine Learning Modelling

The primary objective of the modelling phase was to develop a predictive model capable of estimating hourly PM2.5 concentrations using other available atmospheric and pollutant data. PM2.5 was selected as the target variable due to its well-documented significant adverse impacts on human health, making its accurate prediction crucial for public health advisories and pollution mitigation strategies (Subramaniam *et al.*, 2022). The insights gained from forecasting this specific pollutant can be instrumental in understanding and managing air quality. For the development of the predictive model, all available columns in the pre-processed dataset were initially considered as potential features, except for the 'PM2.5' column itself. The 'datetime' column was excluded from the feature set because the chosen Random Forest Regressor. It treats each data point independently.

However, time-based features could be engineered, as for this iteration, a simpler feature set was adopted. Prior to model training, several preparation steps were undertaken to ensure data quality and compatibility with the machine learning algorithm. Although a dropna() operation was performed earlier. Following imputation, numerical features were scaled using StandardScaler (Tsokov *et al.*, 2022). This process standardises features by removing the mean and scaling to unit variance, which can contribute to more stable and faster convergence. Categorical features, specifically 'wd' (wind direction) and 'station', were converted into numerical representations using LabelEncoder. This method assigns a unique integer to each category within these features. LabelEncoder was chosen for its simplicity.

The Random Forest Regressor was selected as the primary algorithm for this prediction task. This ensemble learning method is well-suited for this type of problem. However, this is due to its ability to capture complex non-linear relationships between features and the target variable. Hence, it is capable to provide feature importance scores. This aids in model interpretability (Heydari *et al.*, 2022). It generally shows strong performance across a variety of datasets. However, it also lacks the need of hyperparameter tuning. Moreover, other regression models such as Linear Regression that was imported or more complex models. These are such as Gradient Boosting Regressors could have been considered. Here, Random Forest offered a good balance of performance, ease and interpretable.

The dataset was partitioned into training and testing sets to evaluate the model's performance on unseen data. Here, 80% of the data allocated for training and the remaining 20% for testing. A random state of 42 was used to ensure reproducibility of this split. The Random Forest Regressor was then trained using the fit method on the prepared training data (X\_train, y\_train). Hence, the model was implemented with its default hyperparameter settings as provided by the scikit-learn library (Muthukumar *et al.*, 2021). However, assess the predictive capability of the trained Random Forest model. Here, two standard regression metrics were employed. These are Mean Squared Error (MSE) and R-squared (R²). An R2 score closer to 1 indicates that the model explains a large portion of the variability in the target. However, evaluation with the test set, the model achieved an MSE of approximately 0.0718 and an R² score of approximately 0.9291. An R² of 0.9291 is a strong result, suggesting that the model can explain about 92.91% of the variance in PM2.5 concentrations based on the selected features. However, it indicates a high degree of accuracy and predictive power on the unseen test data. The low MSE further supports the model's good performance.

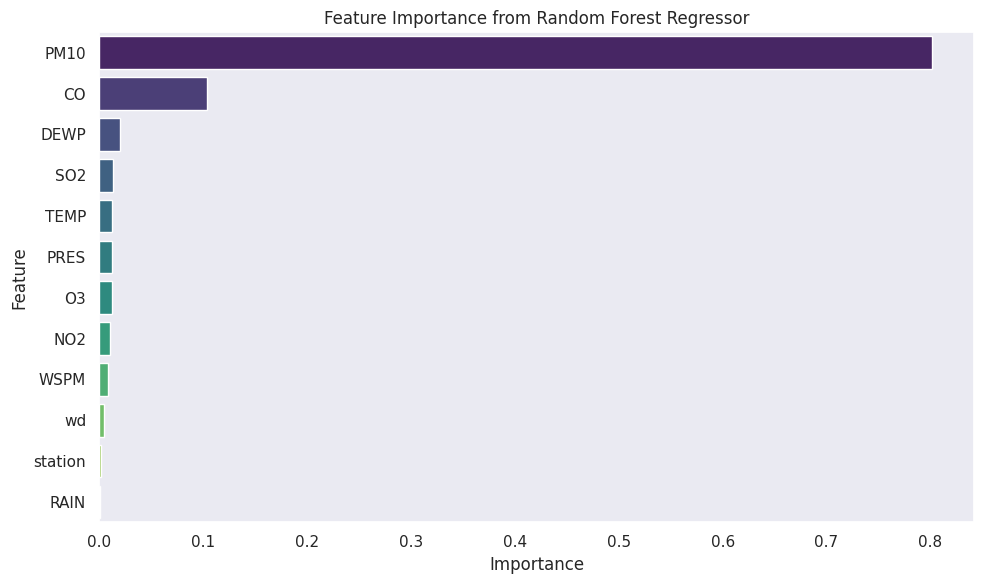


Figure 11: Feature importance plot

(Source: Self-Created)

An analysis of feature importance is conducted. Here, it is derived from the trained Random Forest model. Moreover, it is provided valuable insights into the factors most influential in predicting PM2.5 concentrations. The results showed that meteorological conditions such as temperature. This played a dominant role, strongly influencing variations in PM2.5, PM10, and NO2. Ozone (O3) levels were also shown to be highly sensitive to temperature. However, this shows the significance of photochemical reactions in its formation and its indirect link to particulate matter (Luo and Gong, 2023). However, PM10 was a very significant predictor for PM2.5. Moreover, pollutants such as carbon monoxide (CO) and nitrogen dioxide (NO2) also demonstrated considerable importance. Here, it contributed to PM2.5. Furthermore, sulfur dioxide (SO2) showed relatively lower importance compared to other pollutants. These insights highlight the complex interplay of meteorological factors and co-pollutants in determining PM2.5 levels.

# Task 4: Application Development



Figure 12: Home page

(Source: Self-Created)

The application was developed as a multi-page Graphical User Interface (GUI) using streamlit to offer an intuitive, interactive way for users to engage with the project’s data and results. The home page (Fig. 12) serves as the entry point, providing headings.

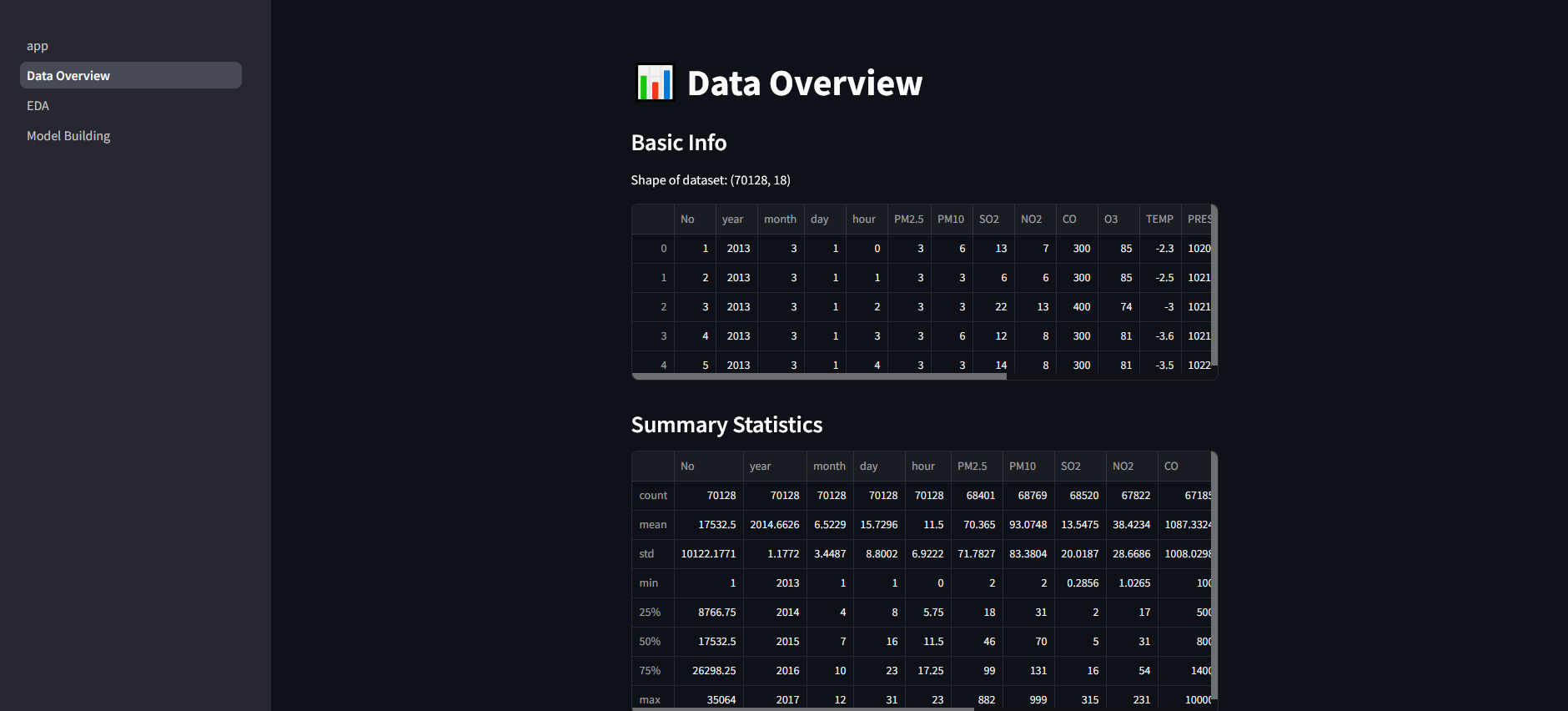


Figure 13: Data overview

(Source: Self-Created)

The data overview page (Fig. 13) presents key statistics on data, missing value summaries, and dataset structure in a user-friendly format.

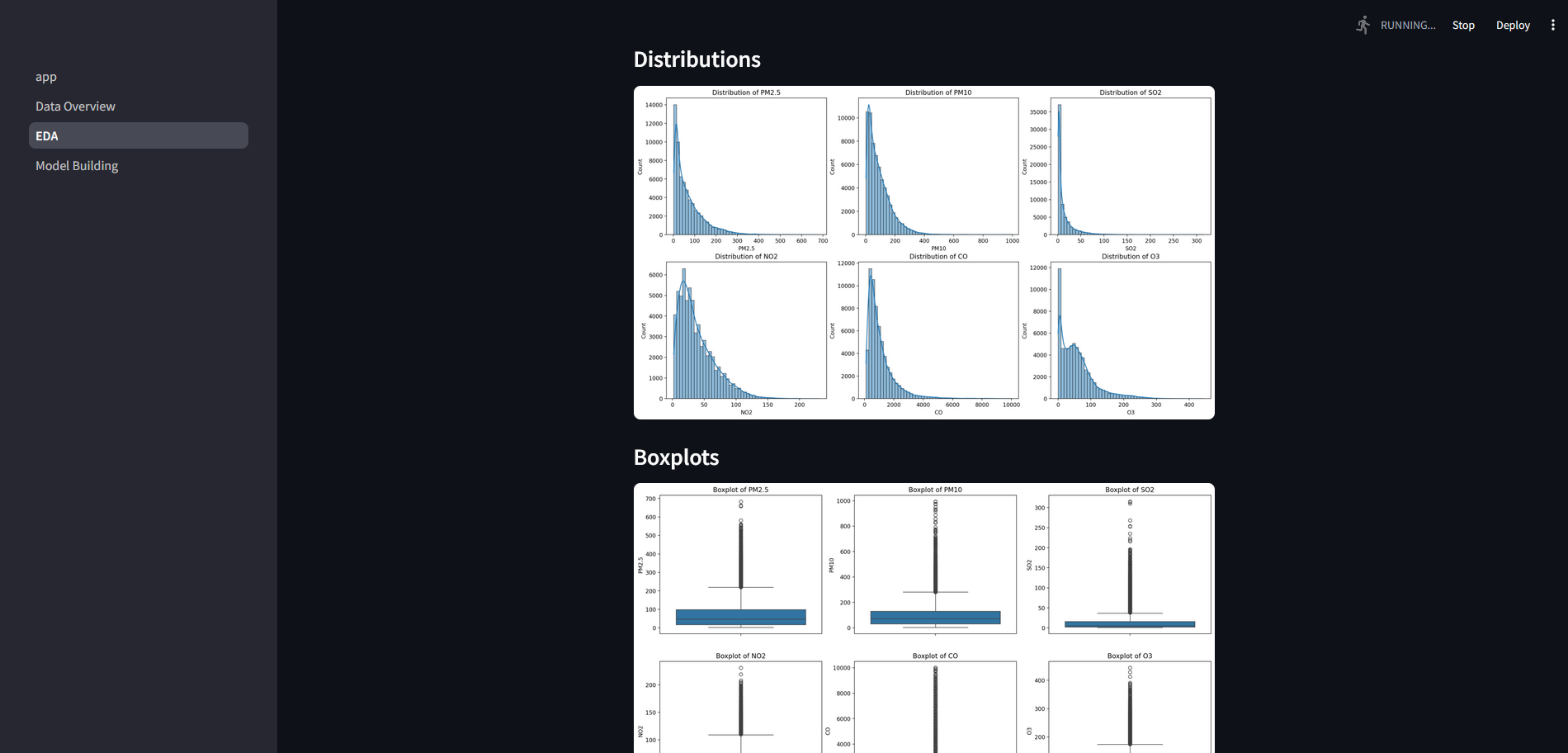


Figure 14: EDA page

(Source: Self-Created)

The EDA page (Fig. 14) allows users to explore visual insights such as histograms and boxplots for pollutants along with other plots.

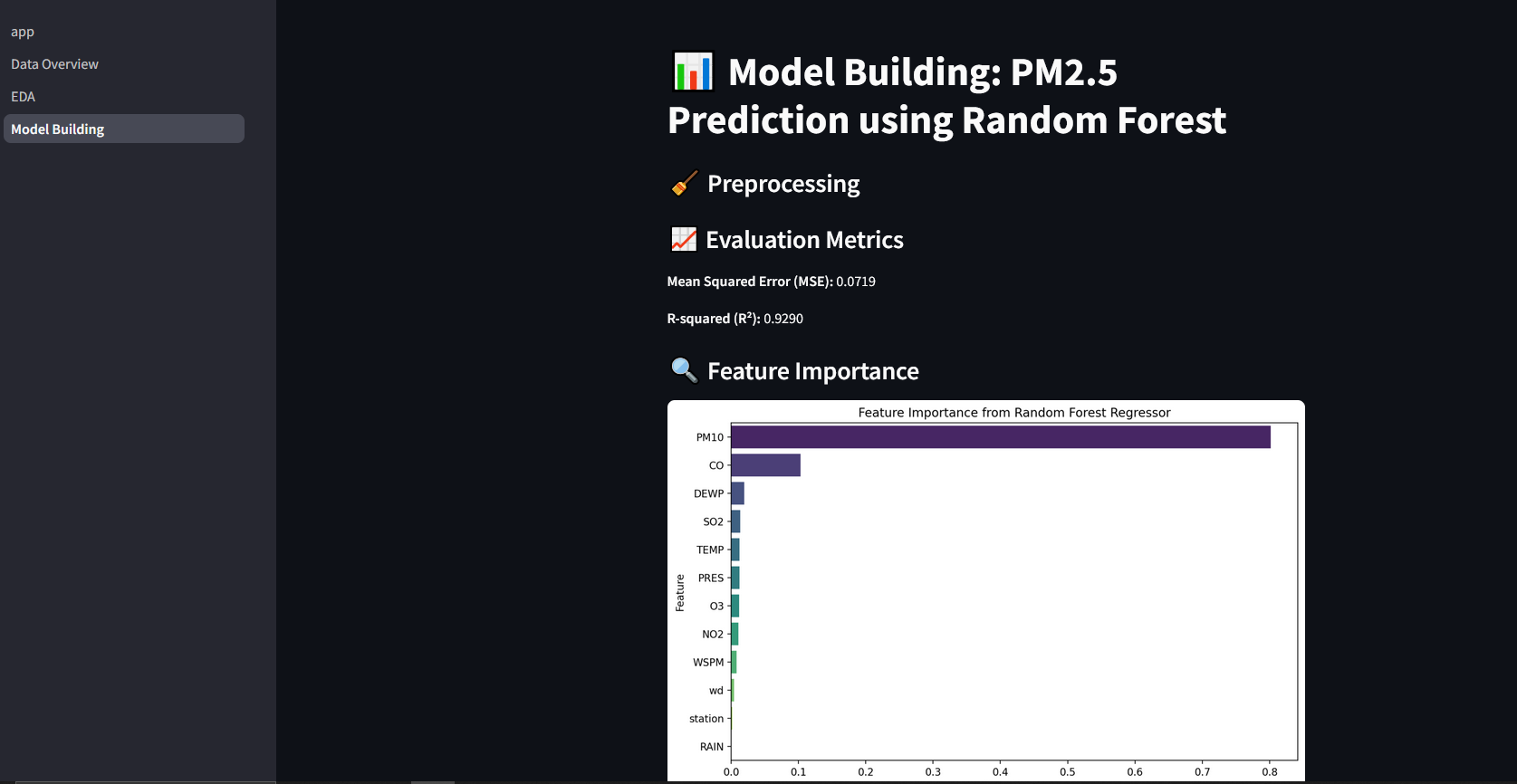


Figure 15: Model building page

(Source: Self-Created)

The model building page (Fig. 15) demonstrates the machine learning workflow used to predict PM2.5 levels such as performance metrics and feature importance plot. The entire interface was designed to guide both technical and non-technical users through the air quality analysis and prediction process, ensuring accessibility and clarity at each stage.

# Task 5: Version Control

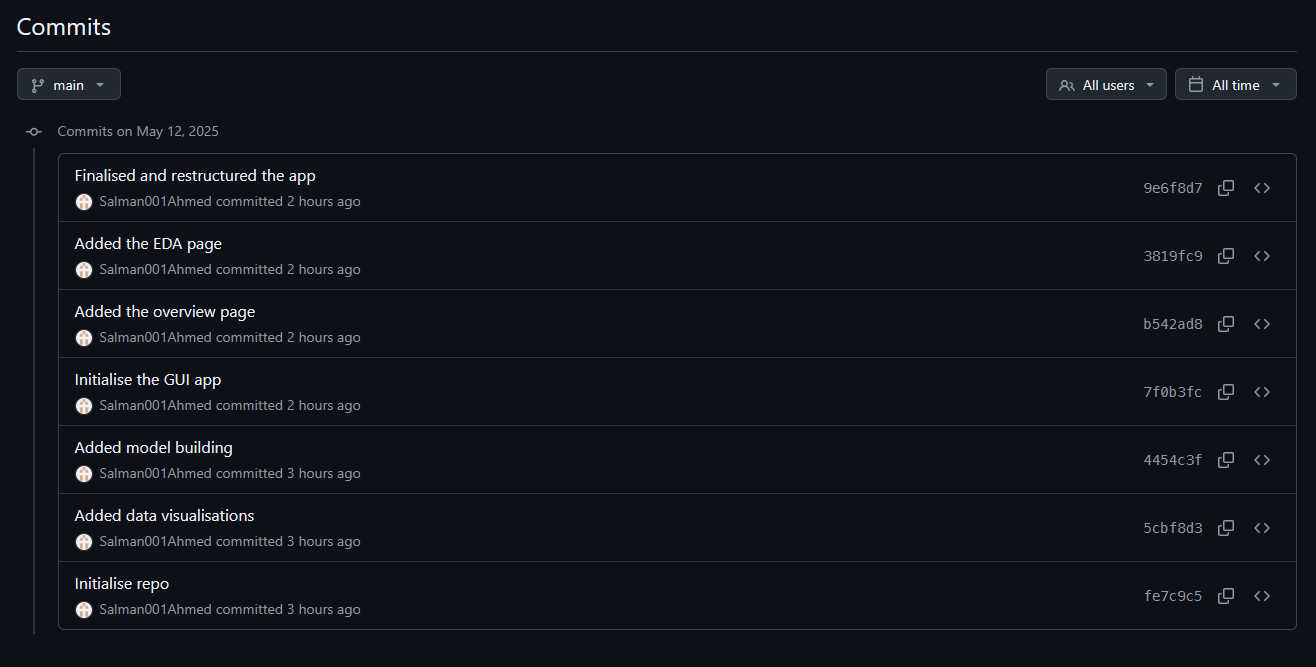


Figure 16: Commit history showing git usage

(Source: Self-Created)

Git version control was used to manage and track all stages of the project’s development. This ensured systematic progress and collaboration. The commit history (Fig. 16) showed consistent updates. This covered data cleaning, EDA implementation, model development, GUI construction, and documentation. However, commit was annotated with meaningful messages to maintain clarity and facilitate rollbacks if needed (Crystal‐Ornelas *et al.*, 2021). Therefore, using GitHub as the hosting platform allowed for centralised code management and backup. Here, Version control also helped in maintaining code integrity, which improves productivity. Hence, it ensures that changes could be traced throughout the lifecycle of the project. This approach ensured a professional and organised development process.

# Conclusion

This project successfully showed the challenge of analysing and forecasting air quality in Beijing by processing and merging multi-site data. Hence, it conducted a thorough Exploratory Data Analysis. Furthermore, it developed a detailed machine learning model. The Random Forest Regressor achieved high accuracy and R² of 0.9291 in predicting PM2.5 concentrations. Hence, it identified key influencing factors such as temperature and co-pollutants like PM10 and CO. Furthermore, a functional multipage GUI application was developed to facilitate data exploration and model interaction, and the project was managed using version control. While the model demonstrated strong predictive power, limitations include the use of data from only two sites and default model hyperparameters. Future work could involve incorporating data from a wider array of site typologies, exploring advanced time-series models, undertaking rigorous hyperparameter optimisation, and expanding the GUI's capabilities to offer more sophisticated analytical tools and real-time prediction features, thereby enhancing its practical utility.

# References

Bekkar, A., Hssina, B., Douzi, S. and Douzi, K. (2021). Air-pollution prediction in smart city, deep learning approach. *Journal of big Data*, *8*, pp.1-21. https://doi.org/10.1186/s40537-021-00548-1.

Crystal‐Ornelas, R., Varadharajan, C., Bond‐Lamberty, B., Boye, K., Burrus, M., Cholia, S., Crow, M., Damerow, J., Devarakonda, R., Ely, K.S. and Goldman, A. (2021). A guide to using GitHub for developing and versioning data standards and reporting formats. *Earth and Space Science*, *8*(8), p.e2021EA001797. https://doi.org/10.1029/2021EA001797.

Gu, Y., Li, B. and Meng, Q. (2022). Hybrid interpretable predictive machine learning model for air pollution prediction. *Neurocomputing*, *468*, pp.123-136. https://doi.org/10.1016/j.neucom.2021.09.051.

Heydari, A., Majidi Nezhad, M., Astiaso Garcia, D., Keynia, F. and De Santoli, L. (2022). Air pollution forecasting application based on deep learning model and optimization algorithm. *Clean Technologies and Environmental Policy*, pp.1-15. https://doi.org/10.1007/s10098-021-02080-5.

Kirwa, K., Szpiro, A.A., Sheppard, L., Sampson, P.D., Wang, M., Keller, J.P., Young, M.T., Kim, S.Y., Larson, T.V. and Kaufman, J.D. (2021). Fine-scale air pollution models for epidemiologic research: insights from approaches developed in the multi-ethnic study of atherosclerosis and air pollution (MESA air). *Current environmental health reports*, *8*(2), pp.113-126. https://doi.org/10.1007/s40572-021-00310-y.

Kumar, K. and Pande, B.P. (2023). Air pollution prediction with machine learning: a case study of Indian cities. *International Journal of Environmental Science and Technology*, *20*(5), pp.5333-5348. https://doi.org/10.1007/s13762-022-04241-5.

Liao, K., Huang, X., Dang, H., Ren, Y., Zuo, S. and Duan, C. (2021). Statistical approaches for forecasting primary air pollutants: a review. *Atmosphere*, *12*(6), p.686. https://doi.org/10.3390/atmos12060686.

Luo, J. and Gong, Y. (2023). Air pollutant prediction based on ARIMA-WOA-LSTM model. *Atmospheric Pollution Research*, *14*(6), p.101761. https://doi.org/10.1016/j.apr.2023.101761.

Muthukumar, P., Cocom, E., Nagrecha, K., Comer, D., Burga, I., Taub, J., Calvert, C.F., Holm, J. and Pourhomayoun, M. (2021). Predicting PM2. 5 atmospheric air pollution using deep learning with meteorological data and ground-based observations and remote-sensing satellite big data. *Air Quality, Atmosphere & Health*, pp.1-14. https://doi.org/10.1007/s11869-021-01126-3.

Subramaniam, S., Raju, N., Ganesan, A., Rajavel, N., Chenniappan, M., Prakash, C., Pramanik, A., Basak, A.K. and Dixit, S. (2022). Artificial intelligence technologies for forecasting air pollution and human health: a narrative review. *Sustainability*, *14*(16), p.9951. https://doi.org/10.3390/su14169951.

Tsokov, S., Lazarova, M. and Aleksieva-Petrova, A. (2022). A hybrid spatiotemporal deep model based on CNN and LSTM for air pollution prediction. *Sustainability*, *14*(9), p.5104. https://doi.org/10.3390/su14095104.