**FROM DATA TO APPLICATION DEVELOPMENT**

**Table of Contents**

[Introduction 3](#_Toc196839368)

[Dataset Information 3](#_Toc196839369)

[Data Handling 3](#_Toc196839370)

[Exploratory Data Analysis (EDA) 5](#_Toc196839371)

[Fundamental data understanding 5](#_Toc196839372)

[Data preprocessing 8](#_Toc196839373)

[Statistics/computation-based analysis and Visualisation 10](#_Toc196839374)

[Model Building 17](#_Toc196839375)

[Application development 20](#_Toc196839376)

[Version Control 24](#_Toc196839377)

[Discussion 25](#_Toc196839378)

[Conclusion and recommendations 25](#_Toc196839379)

[Conclusion 25](#_Toc196839380)

[Recommendations 26](#_Toc196839381)

[References 27](#_Toc196839382)

# Introduction

PM2.5 also known as fine particulate matter is one of the problems in the environmental and health sector that is being experienced globally. PM2.5 are particles of up to 2.5 microns and are hazardous due to their ability to penetrate human respiratory systems and cause diseases which include respiratory diseases, cardiovascular diseases and may even cause death. The evaluation of PM2.5 measurement results is crucial in establishing the relationship between them and various meteorological and pollution variables. Consequently, the analysis made in this paper focused on the PM2.5 data that contains aspects such as temperature, wind speed, and other pollutants. In the EDA process, correlations of these variables with PM2.5 are established. Consequently, in order to predict the air quality, specifically, the PM2.5 level, it is recommended to use linear regression analysis and the random forest. In order to address these challenges, a complex multiple page graphical user interface is developed where the users can use to browse, analyze and even generate the trends in the dataset.

# Dataset Information

The information used in the context of this work is the hourly measurements of air quality, with a particular focus on the PM2.5 rate. This involve such factors as temperature (TEMP), pressure (PRES), relative humidity (DEWP), and wind speed (WSPM), and other pollutants such as PM 10, SO2, NO2, CO and O3. The other variable is for wind Direction (Wd) and datetime including year, month, day and hour. Handling of missing data were done using the ‘mode’ on the nominal ‘Wind Direction’ and the ‘median’ for the numerical values. It was then sorted by datetime to have the data presented in a chronological manner in this format. This dataset can be used effectively for both exploratory analysis as well as the prediction analysis for PM2.5 concentration along with other variables of the surroundings.

# Data Handling

The data comprises 11 files, the observation of air quality at different stations in China, for the period of March 1 2013, to February 28, 2017. All of these files contain hourly readings for PM2.5, PM10, SO2, NO2, CO, and O3 and environmental parameters of temperature, pressure, dew point, wind direction and wind speed, respectively. These datasets offer an extensive insight into the air quality in its present trends in pollution levels as well as its relation to various weather parameters.



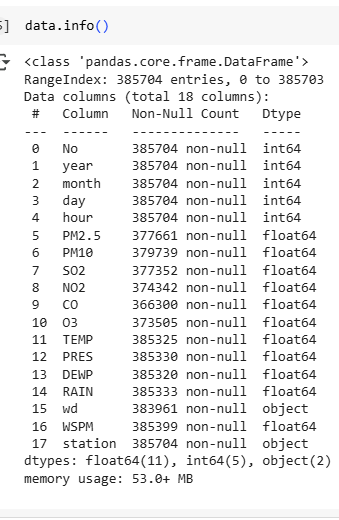
**Figure 1: Merging the datasets**

(Source: Google colab)

The code starts with the fact that there must be a folder with the CSV files and uses glob to get all the paths to the CSV files. It then utilises coded statements that read each CSV file into a DataFrame and save this in an array. Subsequently, all of the DataFrames are joined into one DataFrame using pd.concat. The shape of the resulting dataset is illustrated, and some of the first rows of the data are shown by using data.head(). This helped to combine the data from the many single files into a large and more manageable file to analyse.

# Exploratory Data Analysis (EDA)

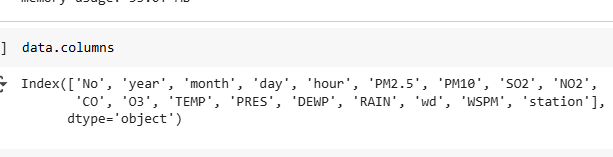
## Fundamental data understanding



**Figure 2: Information about the data**

(Source: Google colab)

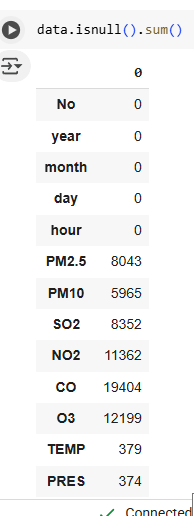
This dataset includes information on pollutants, weather parameters, and station characteristics with 385,704 entries and 18 columns. It has missing values for some columns, such as PM2.5, CO, and NO2. Data types are of three types, namely integer, floating point, and categorical (or object) type.



**Figure 3: Columns of the data**

(Source: Google colab)

The dataset includes 18 columns capturing various features such as date and time, air pollutants (e.g., PM2.5, SO2, NO2), meteorological data (e.g., TEMP, PRES, WSPM), wind direction (wd), and station details for each recorded observation.



**Figure 4: Null values of the dataset**

(Source: Google colab)

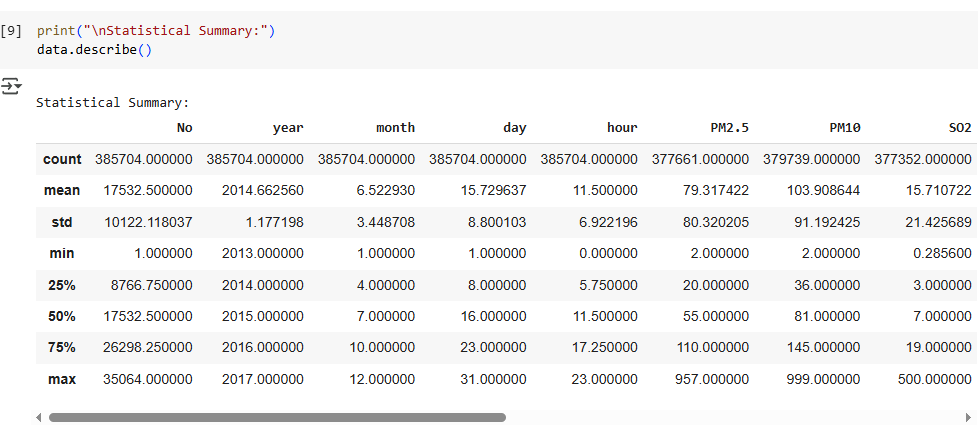
The dataset contains several missing values, notably in key pollutant columns: PM2.5 (8,043), CO (19,404), and O3 (12,199). Weather-related fields like TEMP, DEWP, and wind direction (wd) also have minor gaps that require preprocessing.



**Figure 5: Data types**

(Source: Google colab)

The dataset includes 18 columns with various data types: 5 integer fields for temporal values, 11 float fields representing pollutant levels and weather conditions, and 2 object types for wind direction (wd) and station names, requiring appropriate encoding.

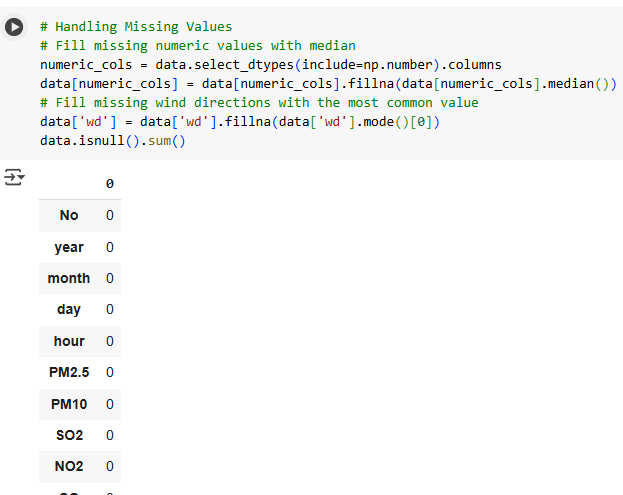


**Figure 6: Statistical summary**

(Source: Google colab)

The statistical summary reveals the central tendencies and spread of each numeric variable. PM2.5 values range from 2 to 957 µg/m³, while temperature spans from -19.9°C to 41.6°C. Wind speed varies up to 12.9 m/s, indicating significant meteorological variation.

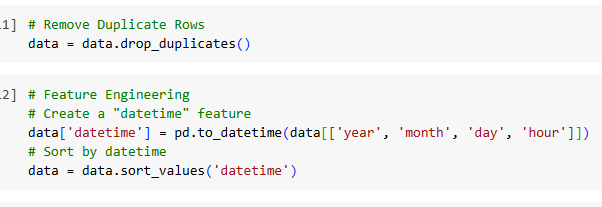
## Data preprocessing



**Figure 7: Handling missing values**

(Source: Google colab)

Missing values in the dataset were addressed using appropriate strategies. Numeric columns were filled with their respective median values to minimize distortion. For the categorical 'wd' (wind direction) column, the most frequent category was used. This ensured no missing values remained in the dataset.

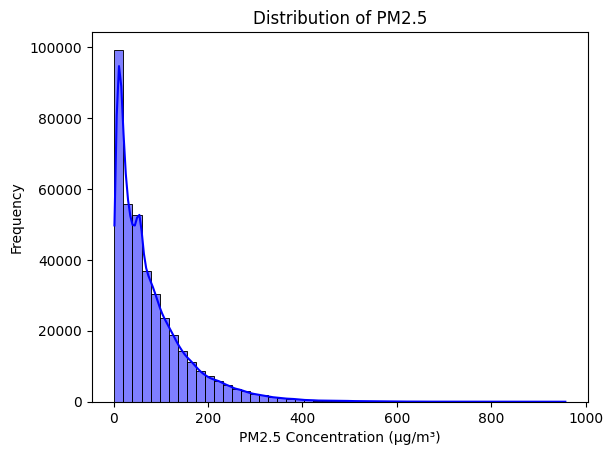


**Figure 8: Removing duplicates and feature engineering**

(Source: Google colab)

Duplicate records were removed to maintain data quality. A new “datetime” feature was created by combining the year, month, day, and hour columns, enabling time-based analysis. The dataset was then sorted by this datetime column to ensure chronological order for further time series processing.

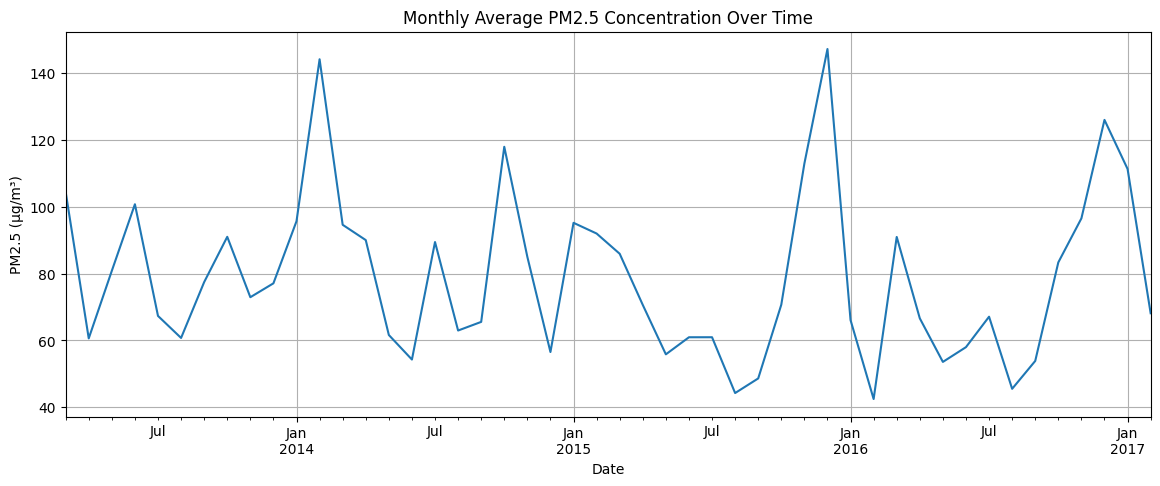
## Statistics/computation-based analysis and Visualisation



**Figure 9: Univariate Analysis: PM2.5 Distribution**

(Source: Google colab)

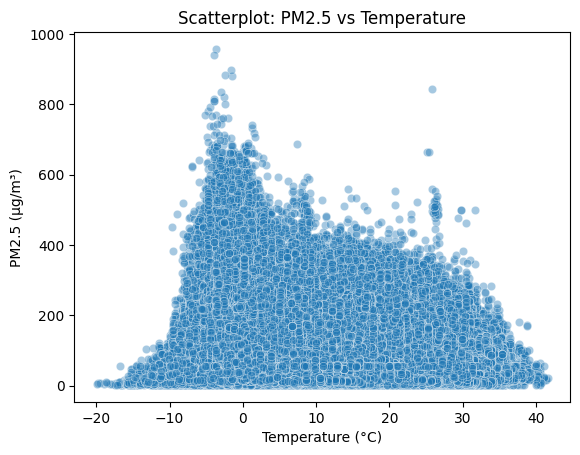
The histogram displaying the distribution of PM2.5 concentrations shows that the majority of the values fall between 0 and 200 µg/m³. This indicates relatively moderate pollution levels for most records, with fewer occurrences of extremely high PM2.5 concentrations in the dataset.



**Figure 10: Time Series Line Plot: PM2.5 Over Time**

(Source: Google colab)

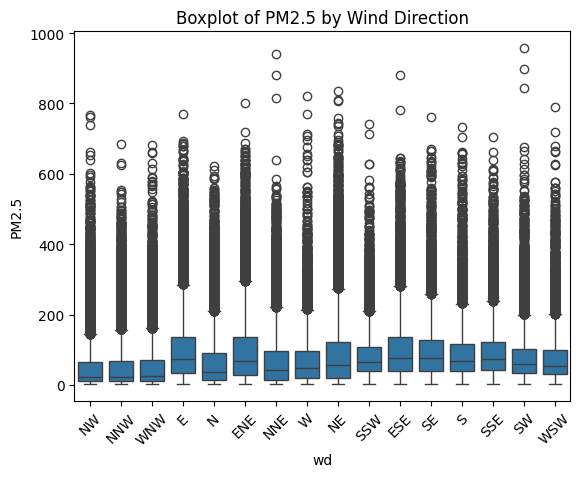
The time series line plot of monthly average PM2.5 concentrations reveals noticeable peaks during January of 2014, 2015, 2016, and 2017. These spikes suggest higher pollution levels in the winter months, likely due to increased heating activities and stagnant atmospheric conditions.



**Figure 11: Bivariate Analysis: PM2.5 vs Temperature**

(Source: Google colab)

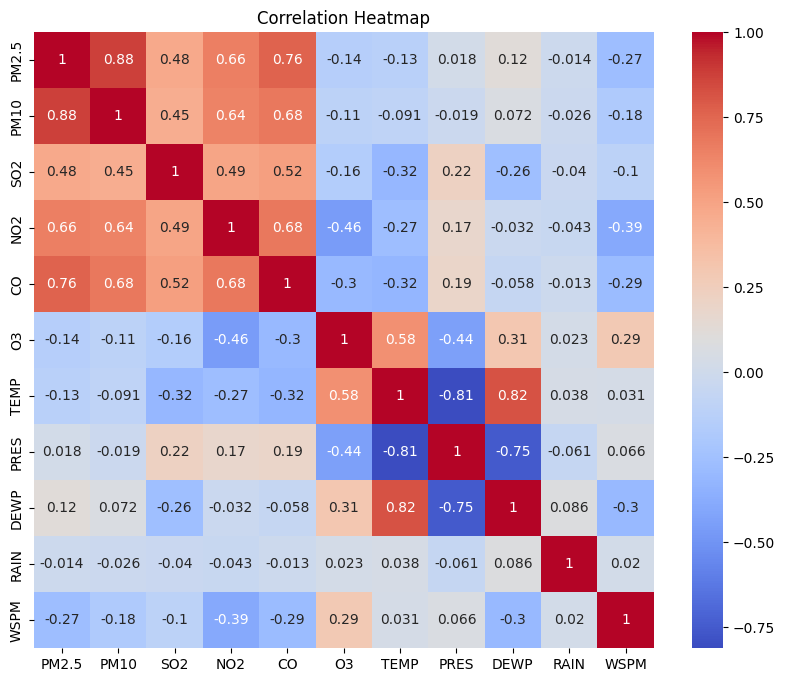
The scatterplot reveals a negative correlation between temperature and PM2.5 concentration. Higher PM2.5 levels are concentrated at lower temperatures, especially below 10°C, indicating winter-related pollution. As temperature rises above 20°C, PM2.5 levels decrease, suggesting better air quality in warmer conditions.



**Figure 12: Boxplot: PM2.5 by Wind Direction**

(Source: Google colab)

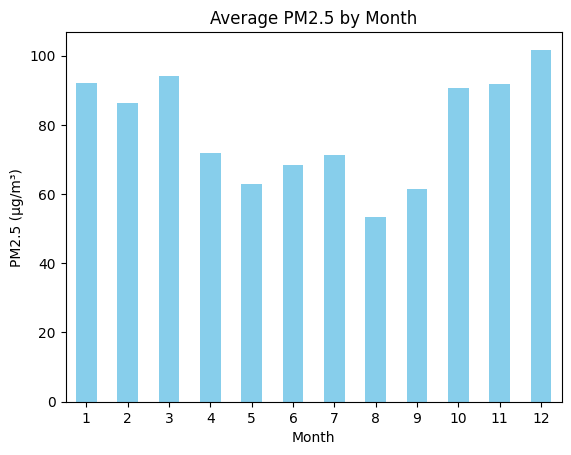
The boxplot shows that wind direction influences PM2.5 levels, with higher median concentrations observed for NE, NNE, E, and ENE. These directions also have greater variability and extreme outliers, suggesting pollution sources or dispersion patterns linked to specific wind directions.



**Figure 13: Heatmap: Correlation Matrix**

(Source: Google colab)

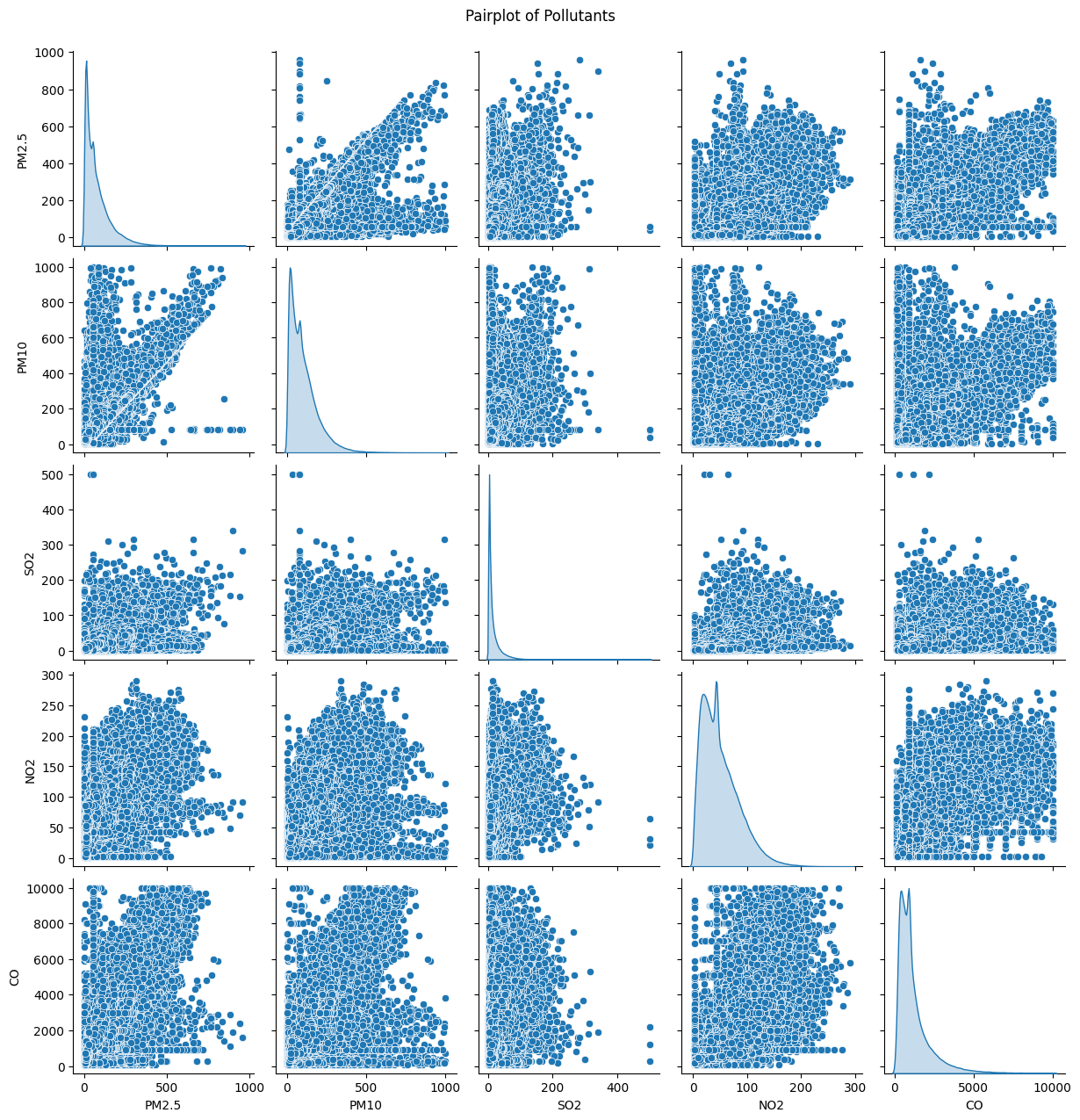
The heatmap shows strong positive correlations between pollutants like PM2.5, PM10, CO, and NO2, indicating common sources. Temperature inversely correlates with several pollutants, suggesting worse air quality in colder conditions. Wind speed mildly disperses pollutants, and ozone rises with temperature.



**Figure 14: Bar Chart: Average PM2.5 by Month**

(Source: Google colab)

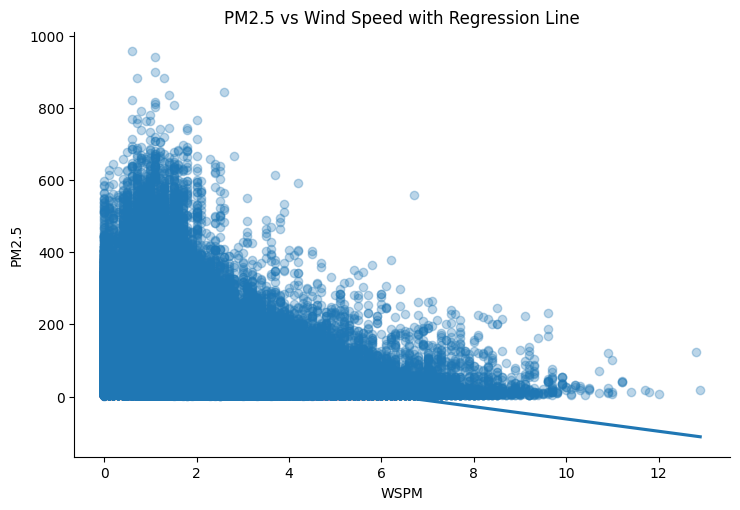
The bar chart illustrates the average PM2.5 concentration by month, with December (Month 12) showing the highest levels. This suggests that PM2.5 concentrations are typically elevated during the winter months, potentially due to seasonal pollution sources or weather patterns.



**Figure 15: Multivariate Analysis: Pairplot of Pollutants**

(Source: Google colab)

The pairplot provides insights into the relationships between PM2.5, PM10, SO2, NO2, and CO. PM2.5 and PM10 are strongly correlated, while other pollutants show more varied and less linear relationships. Right-skewed distributions and outliers suggest variability in pollution levels across observations.



**Figure 16: Wind Speed vs PM2.5: Scatter Plot with Regression Line**

(Source: Google colab)

The scatter plot with a regression line shows a clear negative correlation between wind speed and PM2.5 concentration. Higher wind speeds generally lead to lower PM2.5 levels, as wind helps disperse pollutants. However, some outliers and variability at low wind speeds suggest other influencing factors.

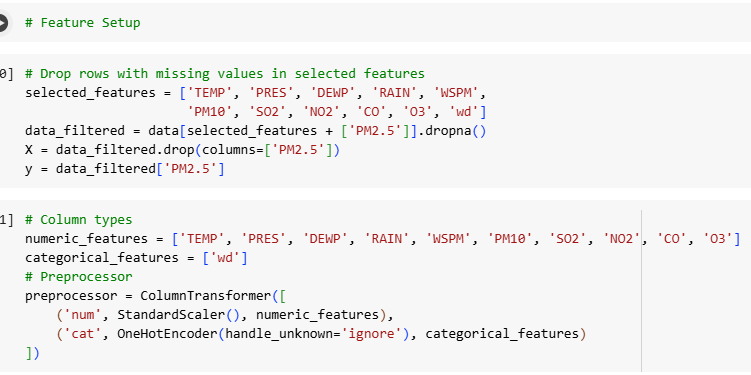


**Figure 17: Statistical summary**

(Source: Google colab)

The custom statistical summary provides a detailed overview of various pollutants and meteorological variables. Key highlights include high skewness and kurtosis in pollutants like PM2.5, CO, and NO2, indicating a right-skewed distribution with potential extreme values or outliers.

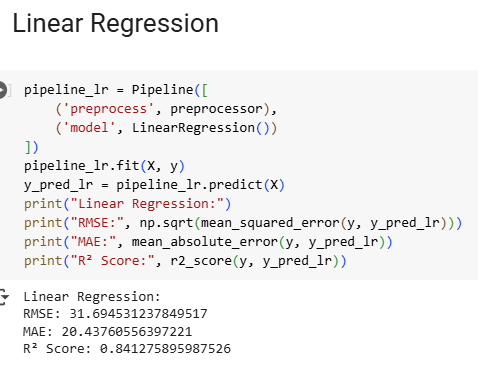
# Model Building



**Figure 18: Feature setup**

(Source: Google colab)

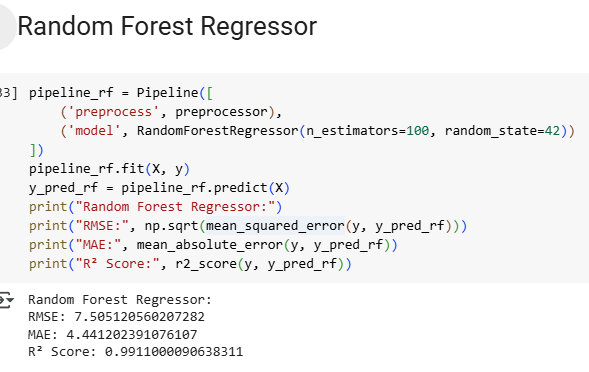
In this task, the goal is to prepare the data for building a model to predict PM2.5 concentrations based on various environmental factors. First, rows with missing values in the selected features are dropped, and the relevant columns for prediction are chosen. The data is then split into features (X) and target (y), where the target is the PM2.5 concentration. The preprocessing step involves scaling numeric features using StandardScaler for normalization, and encoding categorical features (wind direction, wd) using OneHotEncoder to convert them into a suitable format for model training. This preprocessed data will be used to train machine learning models for PM2.5 prediction.



**Figure 19: Linear regression**

(Source: Google colab)

The Linear Regression model shows a strong performance in predicting PM2.5 concentrations. The Root Mean Squared Error (RMSE) of 31.69 indicates that, on average, the model's predictions deviate from the actual values by about 31.69 µg/m³. MAE also gives an idea of the average error of 20.44 of the model, suggesting that the model is normally off by about 20.44 µg/m³. The value of R² of 0.84 means the proposed model accounts for 84% of the variance for PM2.5, which is generally a good result, meaning that the model fits quite well and is accurate in predicting new data.

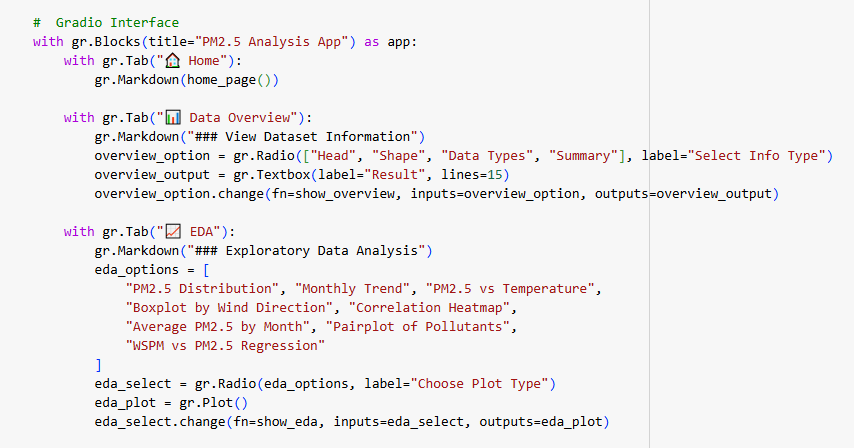


**Figure 20: Random forest regression**

(Source: Google colab)

The evaluation of the dependency predictors shows that the Random Forest Regressor yields higher mean accuracy and coefficient of determination values than the Linear Regression model in estimating the PM2.5 concentrations. In this case, with an RMSE of 7.51, the predictions depart much less from the actual values by their average, thus proving to be more accurate. The MAE of 4.44 reinforces the results supporting its reliability and indicates that the prediction average error margin is approximately 4.44 µg/m³. Thus, the Random Forest model has a good and reliable fit to the data with an R² score of 0.99, which means that the model explains 99% of the variance in PM2.5. This also implies that the model is very good in the determination of the probabilities as a relationship between the features and the PM2.5.

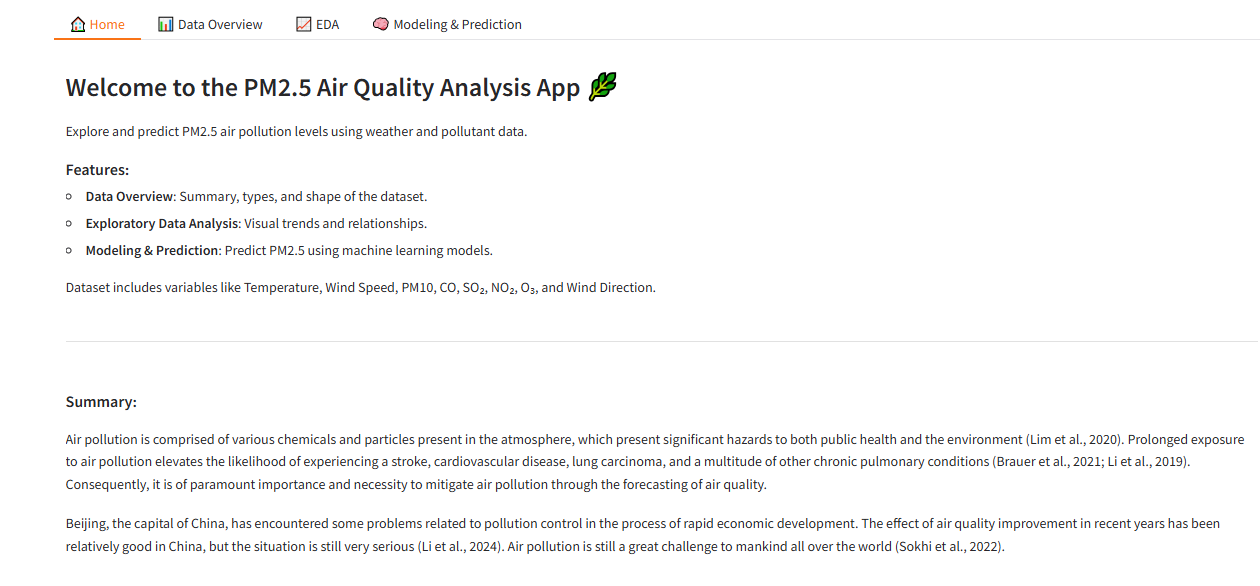
# Application development



**Figure 21: Gradio interface creation**

(Source: Google colab)

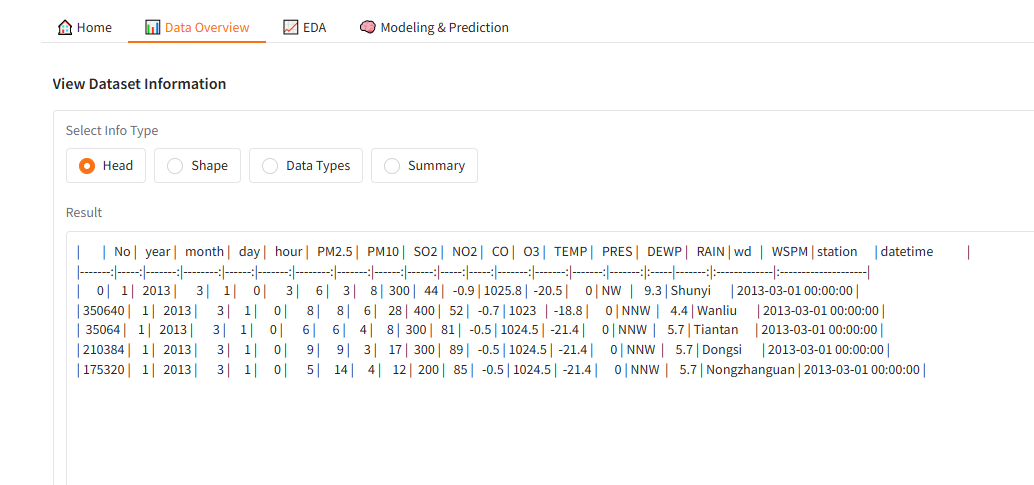
The code given here describes how to build a Gradio application for ML models to predict the PM2.5 levels. The application enables the user to look up weather and pollutant data to be able to make a prediction of the level of air pollution. The different parts of the interface are data preview, EDA and a model prediction section where the user can choose whether to use a linear regression model or a random forest model. The mobility of the application involves a database containing weather and air quality pollutants, and it is easy to use to make predictions about PM2.5 concentration.



**Figure 22: Gradio interface**

(Source: Google colab)

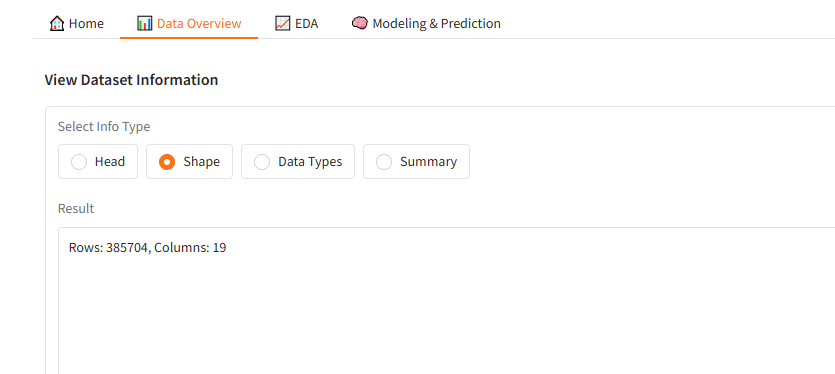
The gradio developed interface offers different options to users to visualise air pollution data, as well as the prediction of PM2.5 levels by using machine learning. It involves the use of data summaries, EDA plots and model predictions in forecasting the quality of air.



**Figure 23: Gradio interface dataset overview**

(Source: Google colab)

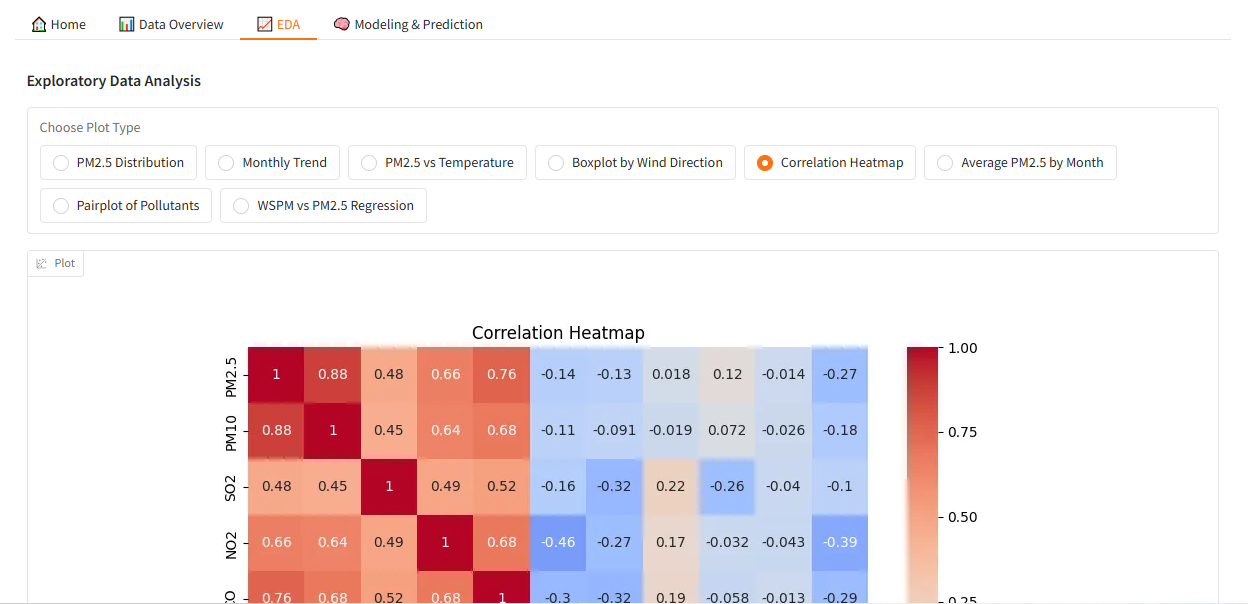
The Gradio interface also provides approximately all general information about the dataset of the features and their types aims to maximize the user experience.



**Figure 24: Gradio interface dataset shape**

(Source: Google colab)

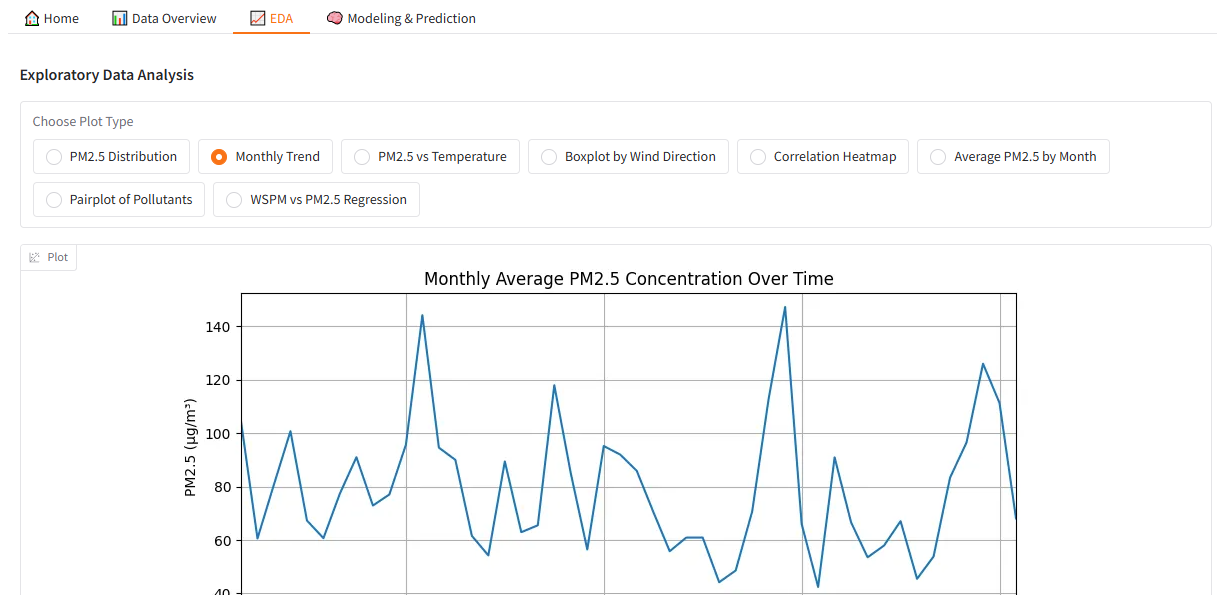
The interface in Gradio also allows showing the shape of the given dataset and the total number of rows and columns is shown to provide the user with an idea of the dataset size and the way it is arranged.



**Figure 25: Gradio interface showing visualisation**

(Source: Google colab)

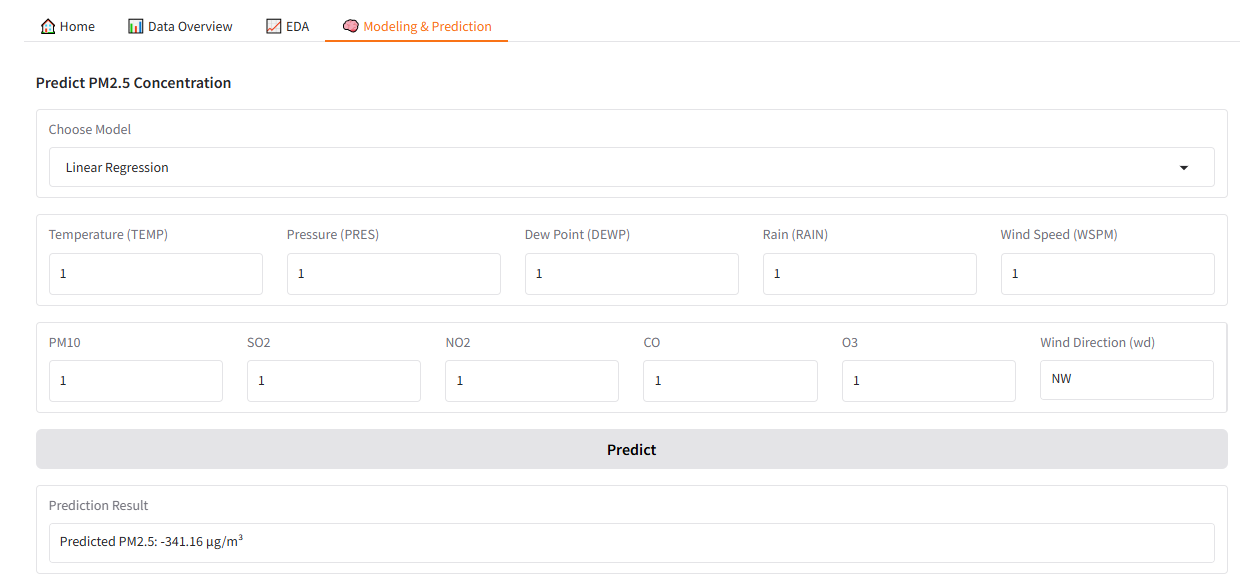
The Gradio interface also provides the features such as distributions or relationship that exists within the data set where the user can visually analyze the data set for further deeper analysis or interpretation.



**Figure 26: Gradio interface showing visualisation**

(Source: Google colab)

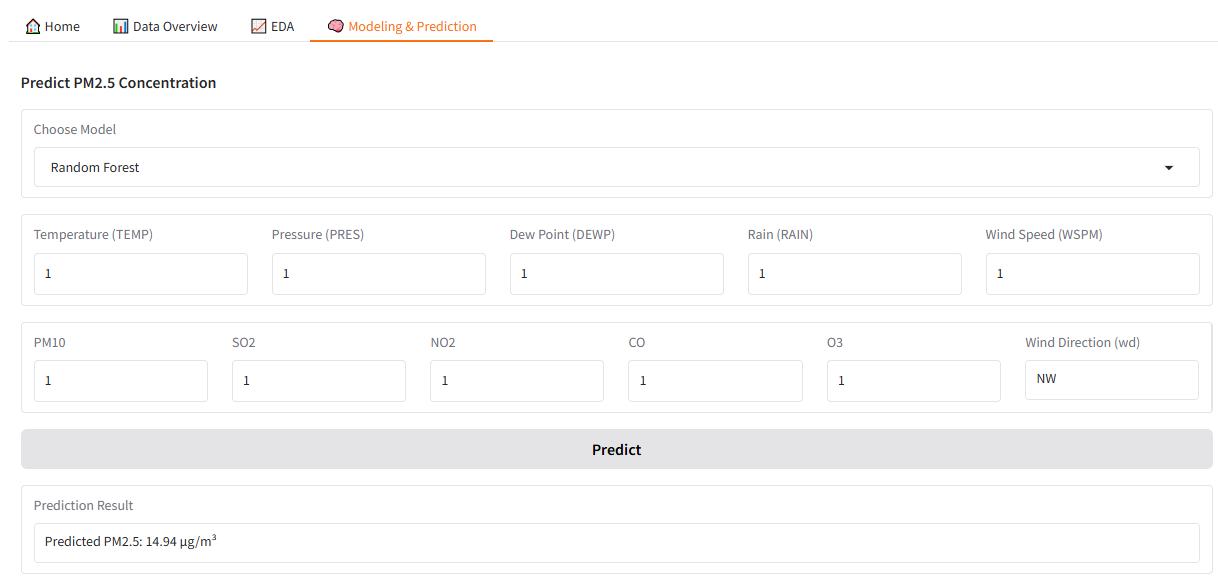
Another type of Gradio visualisation is the ones that pertain to Relations, which can be seen as a line plot that gives the users an insight into how two features are related in the dataset.



**Figure 27: Gradio interface showing linear regression prediction**

(Source: Google colab)

The Gradio interface provides dynamic predictions regarding the PM2.5 level based on configurable variable values to explain linear regression predictions in a more interactive model for air quality.

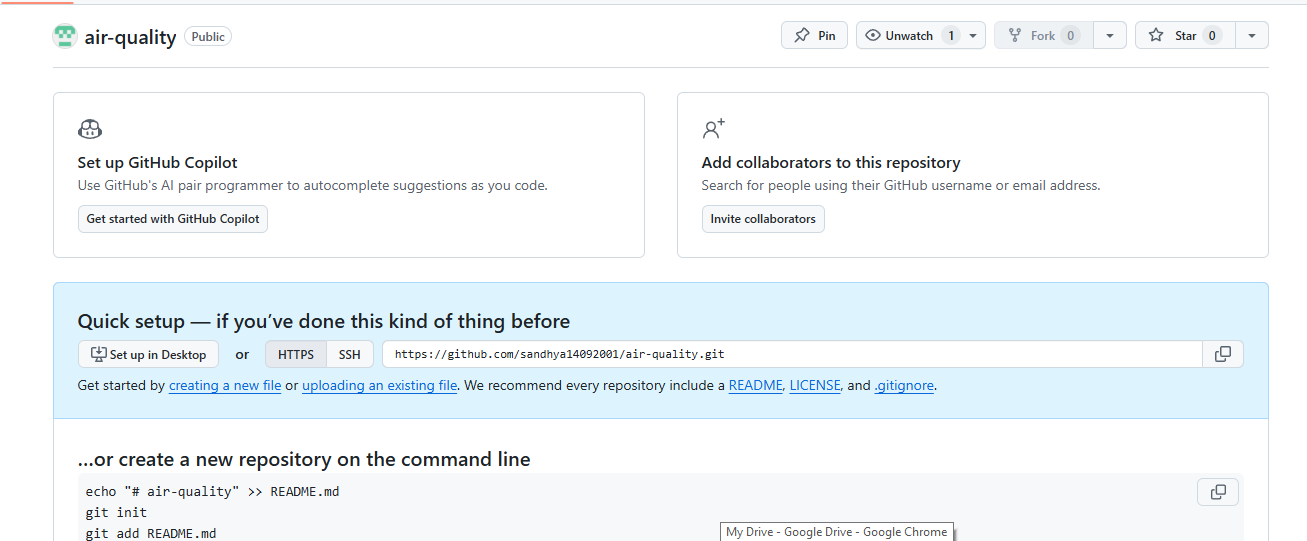


**Figure 28: Gradio interface showing Random Forest prediction**

(Source: Google colab)

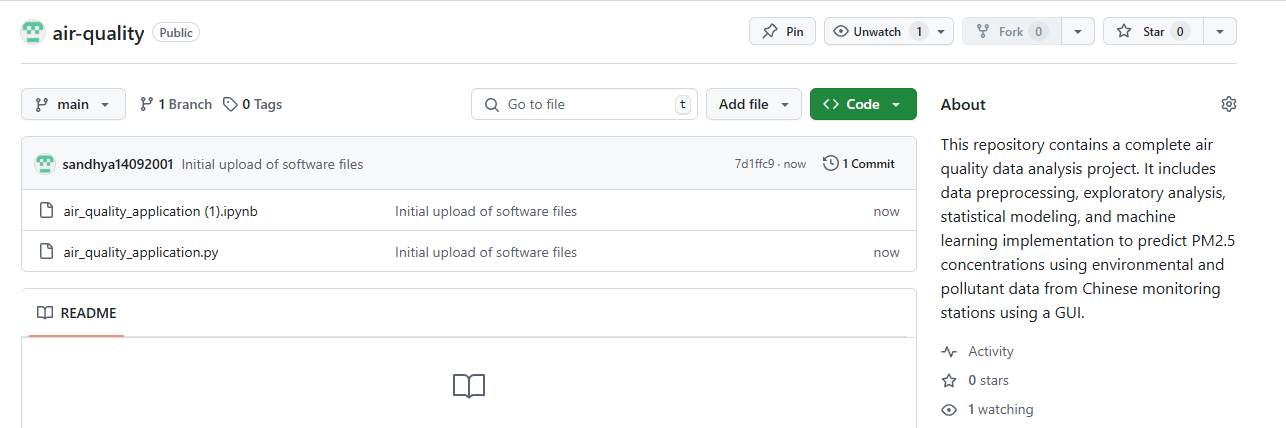
The Gradio interface enables presenting the decision made with the help of the Random Forest model, and thus presents an alternative to choosing an ML algorithm to predict PM2.5 levels using selected attributes.

# Version Control



**Figure 29: Github repository**

(Source: Self-created)



**Figure 30: Adding files into Github repository**

(Source: Self-created)

Version control has been executed through creating a GitHub repository and regularly committing the project’s progress. The first commit included creating a basic layout of the project and loading the dataset, data cleaning and exploratory data analysis (EDA). Further commits showed the inclusion of linear regression and the random forest model, as well as the incorporation of the evaluation parameters. Further, commits were made on improving the UI where the Gradio platform was integrated to allow users to peek at summaries and trends and also make predictions. Further commits documented enhancements in the model quality, say, the hyperparameter adjustments. The benefit of committing is that it provides a record of the progression of the project within the documentation trails, which can easily be accessed from the changes made from time to time.

# Discussion

The results show a systematic approach to analysing the data collected from different stations concerning air quality in China. The dataset was accumulated effectively and combined into one structure based on hourly values of pollutant and meteorology parameters (Nath *et al.* 2021). In order to get a better understanding of the dataset, various EDA procedures were conducted on the dataset, including missing values, statistical distribution, and the relation between pollutants and the weather conditions.

In the preprocessing stage, null values have been handled and removed, as well as the construction of some additional feature variables, including a datetime variable to support time series analysis. This was succeeded by simple distributions and simple associations that pointed out trends of, for instance, PM2.5 variation with months and the inverse proportional relationship between temperature and pollutants. This was also evident in the heatmap and pairplots, as they clearly showed interconnections of pollutants and how pollution is dispersed depending on the wind speed and direction of different sources (Kothandaraman *et al.* 2022). Model construction included building linear regression and random forest regression models with inputs of environmental factors to predict PM2.5 levels. The linear regression model produced an R² of 0.84, suggesting a good fit, and the random forest increases the effectiveness of the classification. The study conducted overall gives crucial information concerning the general trend of air quality and the effects of weather factors on pollutants.

# Conclusion and recommendations

## Conclusion

Thus, in this paper, the analysis of the air quality dataset has allowed to study the findings of the connections between pollutants and meteorological factors for different monitoring stations in China. Some of the insights that came out from Data preprocessing and EDA include the seasonal variation of pollution and the correlation between PM2.5 and weather factors, including temperature and wind speed. The results based on using linear regression and random forest also proved high accuracy of the models for PM2.5 concentration prediction, with deeper predictive capability of the random forest. It helps expand the knowledge of air quality changes and develop further prospects of air pollution prognosis.

## Recommendations

According to the findings, it is suggested to increase attention to the problem of air quality in areas with high rate fluctuations in the background of the winter period. Improvement in data collected, especially on the missing values of key pollutants, will advance the prediction models further. Further, including more of the meteorological fields, like humidity or pressure, into the model should enhance the results (Kumar and Pande, 2023). The integration of real-time data in the models will help improve the forecasting of the models through updates. Thus, the correlations between the developed PM concentrations and wind direction might produce pollution control measures that are more specific to several pollution sources associated with wind direction.

# References

Kothandaraman, D., Praveena, N., Varadarajkumar, K., Madhav Rao, B., Dhabliya, D., Satla, S. and Abera, W., 2022. Intelligent forecasting of air quality and pollution prediction using machine learning. Adsorption Science & Technology, 2022, p.5086622.

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.

Nath, P., Saha, P., Middya, A.I. and Roy, S., 2021. Long-term time-series pollution forecast using statistical and deep learning methods. Neural Computing and Applications, pp.1-20.