Air Quality
Analysis & Prediction

1. Introduction

The world we live in is being rapidly automated and emerging technologies like Cloud, Internet of Things, and so forth are being continuously integrated into concepts such as Smart Cities to provide a high level of comfort to the residents with minimum human intervention [10]. A major challenge faced by corporations of developed cities is to control and regulate air quality. With the advent of modern air quality monitoring and pollution control systems, a novel prediction framework aids the process of finding effective solutions to complex problems.

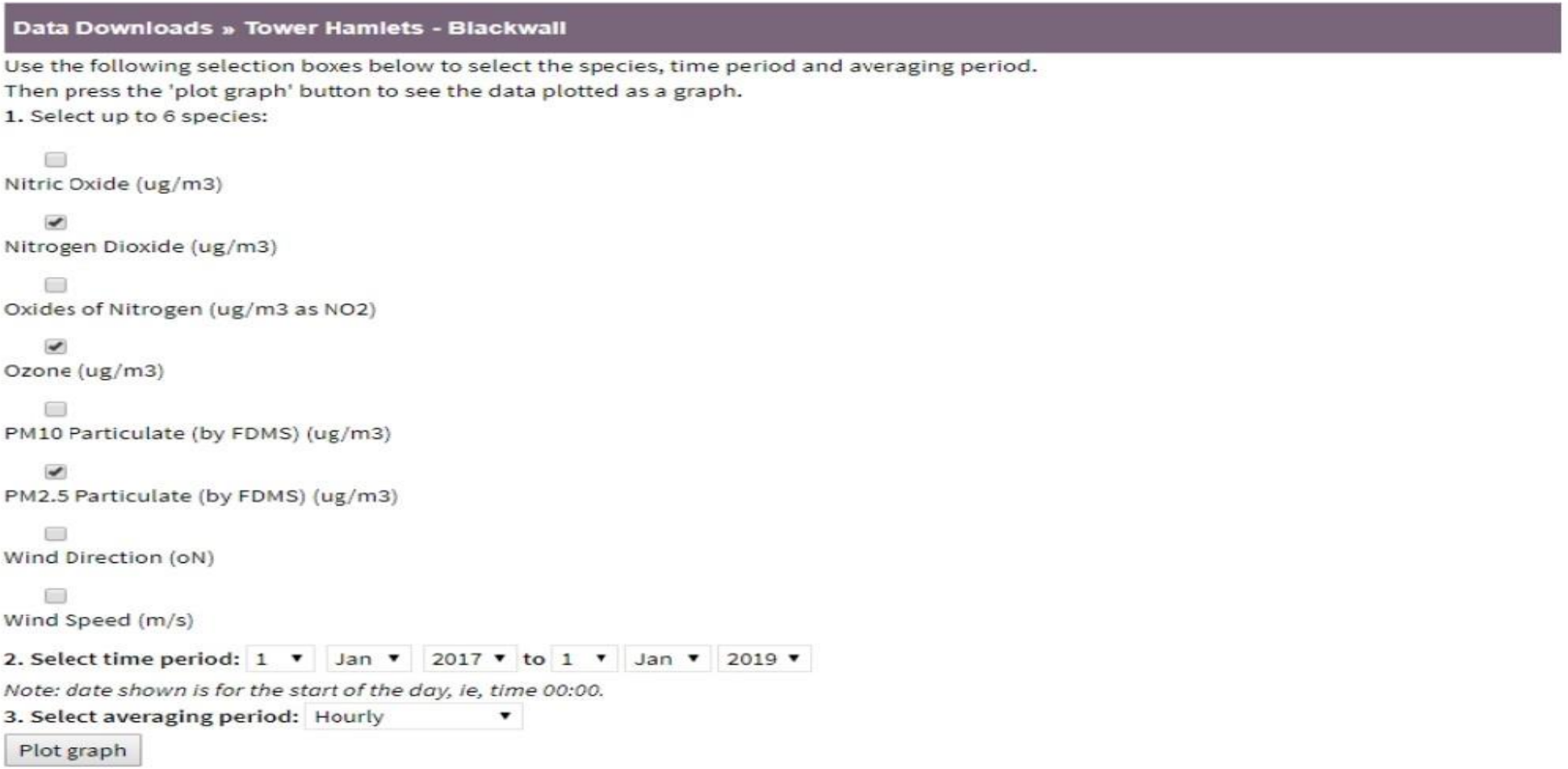
This project aims to predict the air quality band for PM2.5 using present and historical pollution data in combination with predicted weather data which is readily available. To solve this problem, firstly, exploratory data analysis will be conducted on available weather and pollution datasets to discover the correlation between different features. After employing suitable data cleaning and feature engineering methods based on the observations made, the feasibility of using different machine learning techniques such as classification and regression models will be analysed.

2. Description of Dataset

The dataset we have for this project was created by joining historical air pollution and weather datasets obtained from two different sources. The steps that were undertaken to obtain these datasets and creating the final dataset are detailed below:

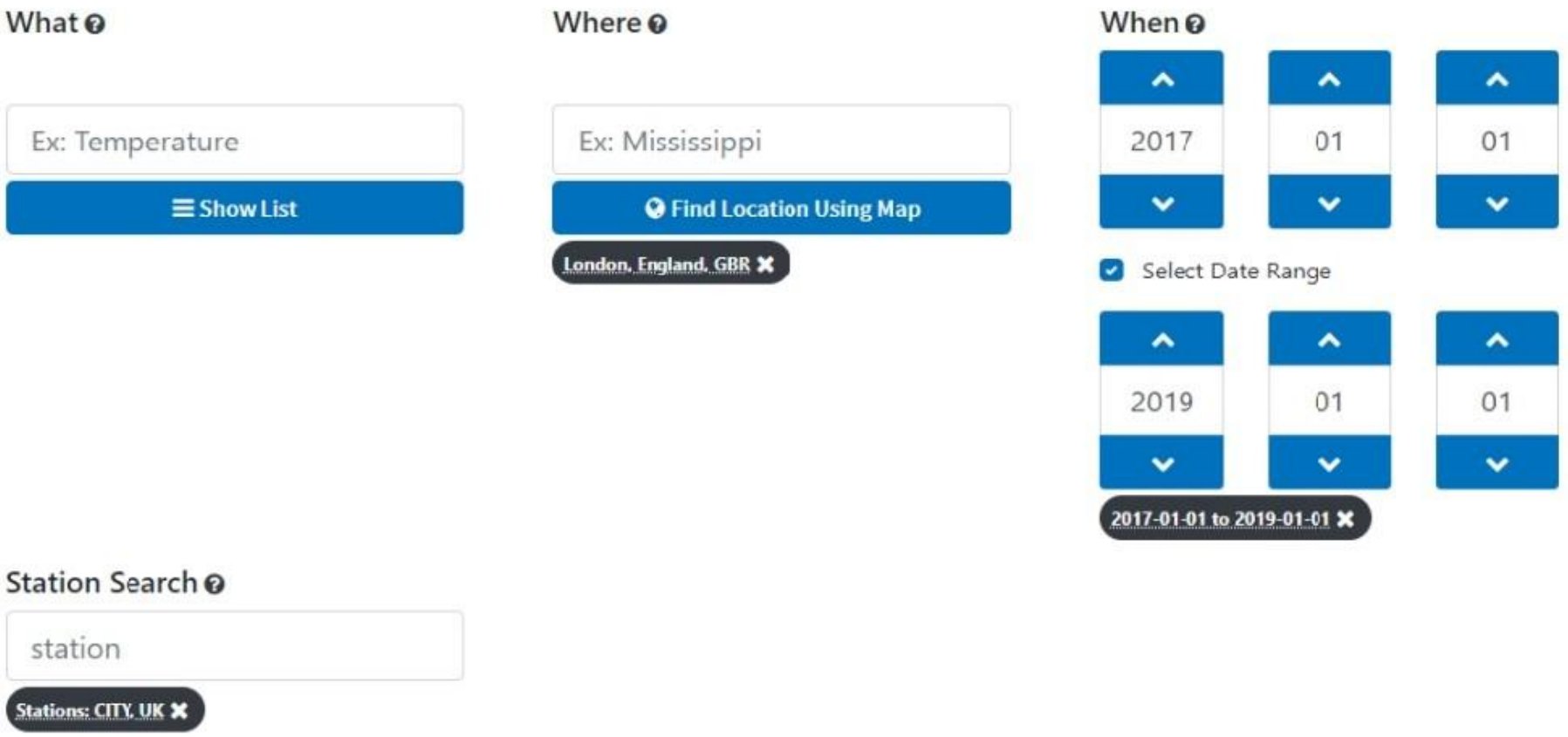
The air pollution data was obtained from the London Air, the website of the London Air Quality Network (LAQN), which monitors air pollution in London and South East England. The LAQN was formed in 1993 to coordinate and improve air pollution monitoring in London. The website provides publicly available datasets that contain independent scientific measurements of various pollutants obtained from over 121 active monitoring sites [2]. The London Air website provides a data download tool which allows the user to download either data for one site or data for one species for up to six sites.

Figure 1: Pollution Data Collection



- In this project, we have chosen to obtain the air pollution data for three specific species, namely Nitrogen Dioxide (NO2), Ozone (O3) and PM2.5Particulate Matter, from the Tower Hamlets monitoring station in Central London. The time period chosen was from January 1, 2017 to January 1, 2019. The sampling rate was chosen to be hourly which resulted in a total of 17520 samples for each pollutant during the chosen time period.
- The weather data was obtained from the Integrated Surface Global Hourly Dataset hosted by the National Oceanic and Atmospheric Administration (NOAA). The Integrated Surface Dataset is composed of worldwide surface weather observations from over 35,000 stations.
- The dataset was available to download as a CSV file.The Integrated Surface Dataset offers a data search tool that allows the user to choose the observed species, location and time period. In this project, we have chosen to collect temperature, humidity, wind speed and direction since these are expected to demonstrate a large correlation to air pollution. The dataset was available to download as a CSV file. The monitoring site was chosen to be City, London according to the geographical proximity to the Tower Hamlets station chosen in the air pollution dataset.

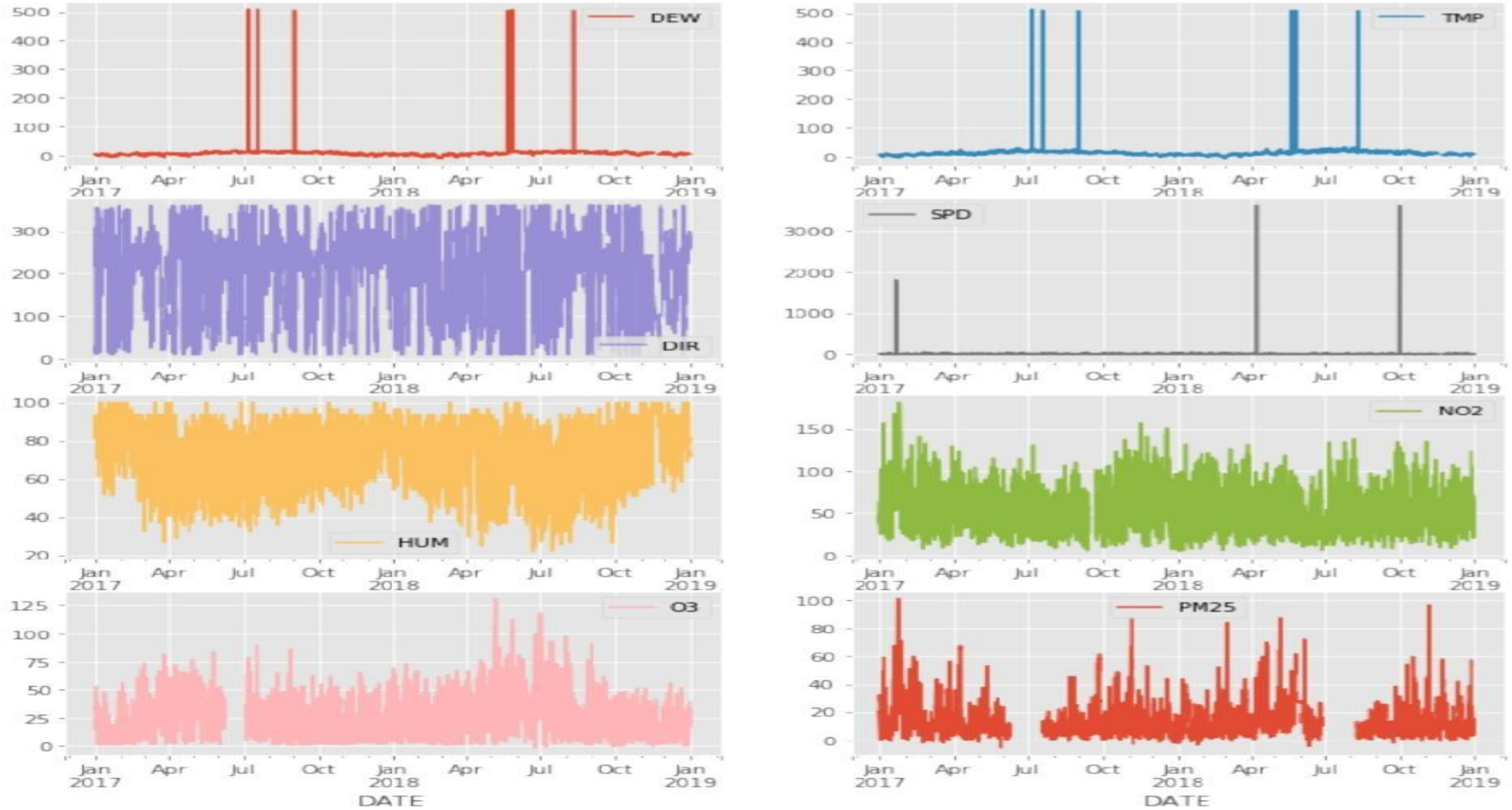
Figure 2: Weather Data Collection



3. Data Visualization

- The time series visualization of the final dataset was plotted to construct initial impressions about the collected data.It is clear from the initial visualizations that many features of the dataset contain null values and outliers. Given the scientific nature of the dataset, it is not ideal to handle these issues generically by employing statistical methods. The outliers and null values in the dataset have to be handled in context with available scientific information, and the methodological approach that was undertaken to perform these operations is detailed in the next section

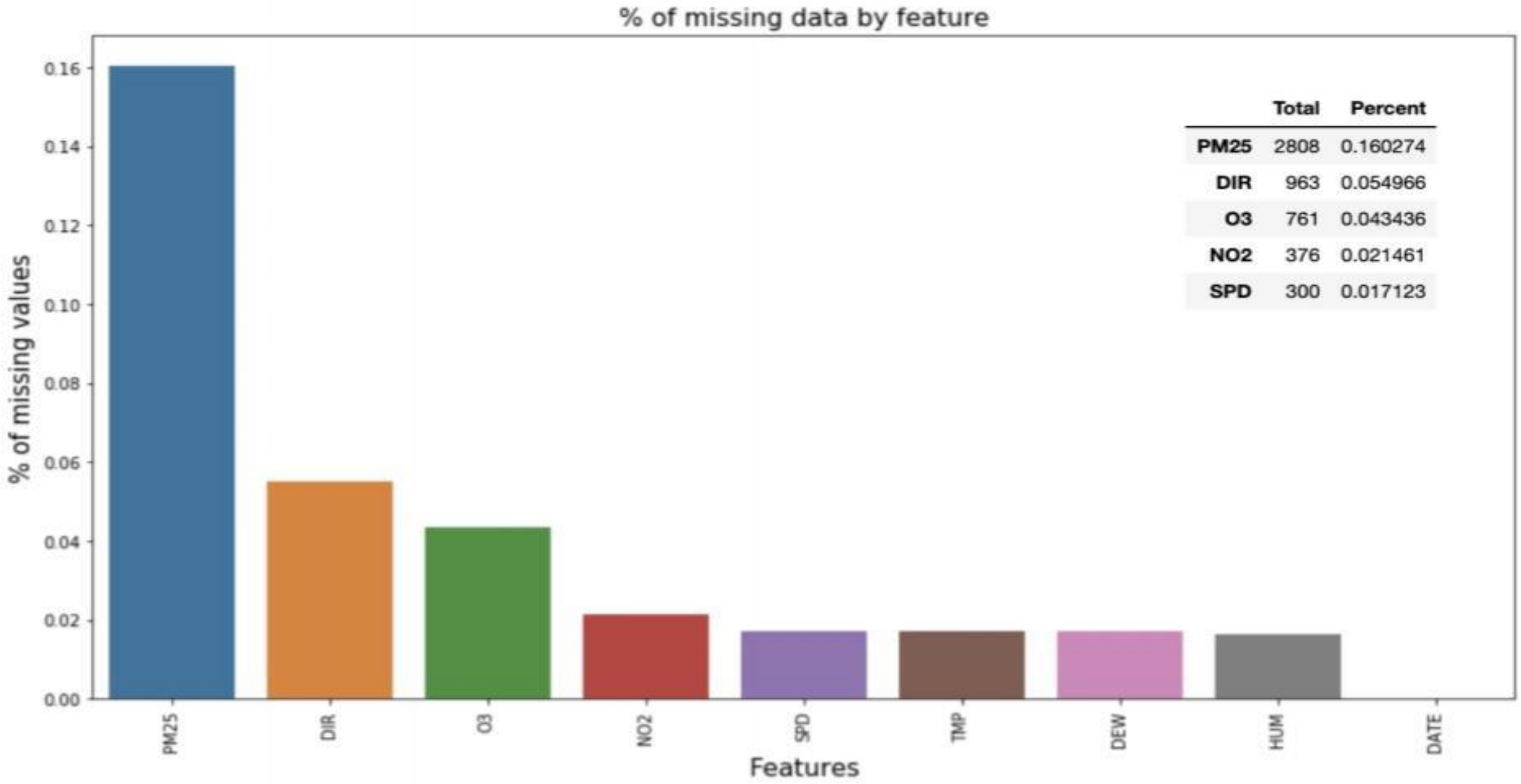
Figure 3: Timeseries visualisation



4. Data Cleaning

- This section deals with handling of missing null values. Different approaches were implemented in order to handle and impute the missing values. This is a crucial part of pre-processing as it could lead to wrong prediction and classification of any model being used in the future. The fundamental approach is to understand the reason of missing values and observe the distribution of null values.
- From the previous time series visualization and exploring the percentage of missing data by feature, it is obvious that PM2.5 has a large amount of missing data in account of no observations being made in the late summer months. Wind direction has the next higher percentage of missing values on account of the conversion of '999' missing values from Section 2.3. Ozone has no observations throughout the month of July, 2017 as apparent from the time series visualization. Several imputation methods were tested and the advantages & disadvantages of each of these methods were extensively studied.

Figure 4: Number of null values for different features



	Total	Percent
PM25	2808	0.160274
DIR	963	0.054966
O3	761	0.043436
NO2	376	0.021461
SPD	300	0.017123

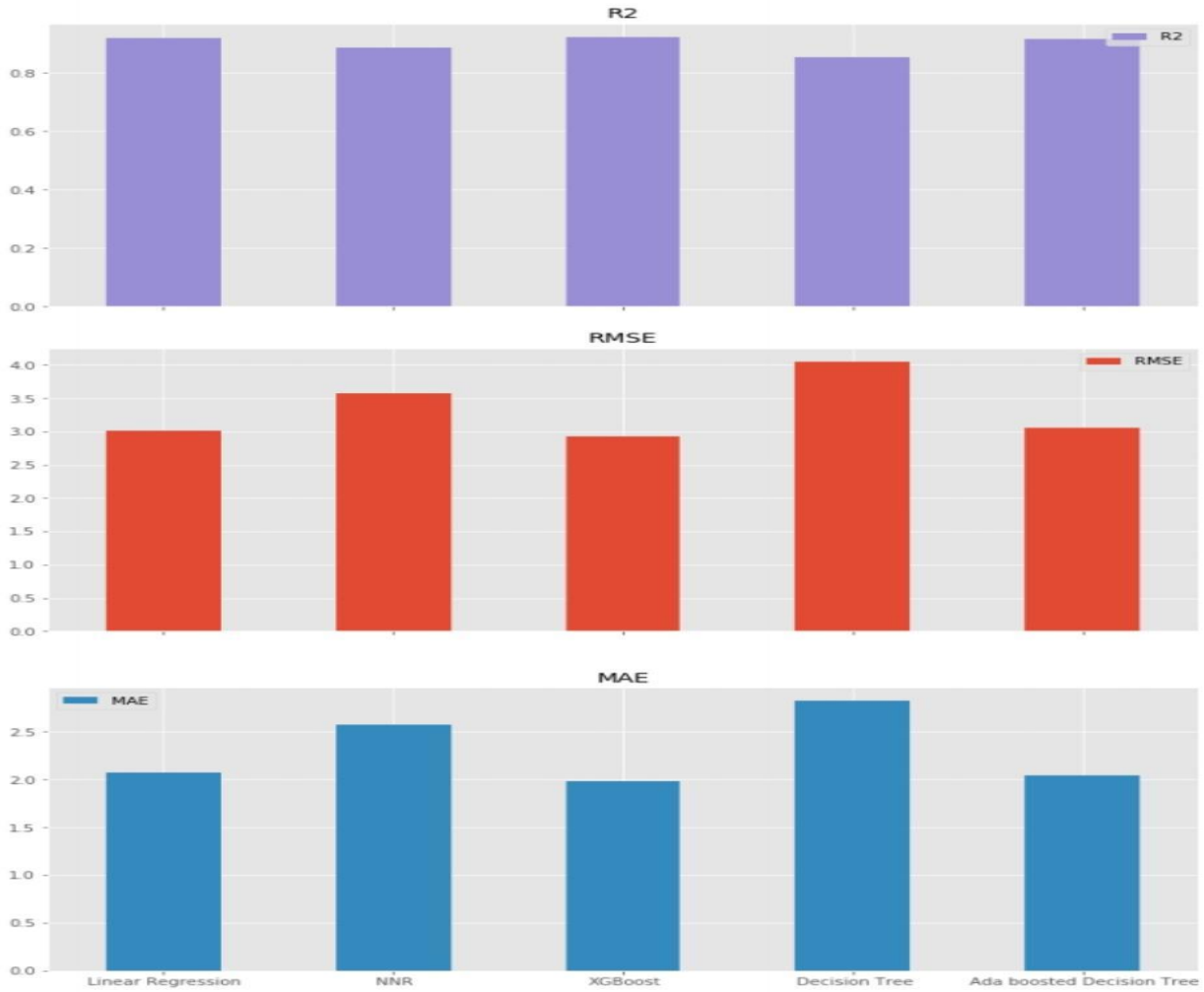
5. Prediction Models

- Before the final dataset is fed into the prediction model, some preparation has to be done. This includes deleting all rows containing null values, which was decided to be the best approach to deal with missing values within the context of the dataset used. The final dataset before dropping null values has a total of 17520 records, which is an hourly observation for each of the feature for the time period of two year. There are 53 features in total after the feature engineering steps detailed in the previous section. The date and day of week features are to be dropped because the relevant date features such as hour, month and numerical values for day of week have already been extracted.

6. Performance Comparisons

- A performance comparison of the different regression models tested in this project is made and their performance metrics are presented in the form of bar plots. It can be seen that, there is no drastic difference in the performance of the different models and all the models performs more or less equally with good numbers. Although XGBoost and Ada Boosted Decision tree had a close call, XGBoost topped the table with a lesser RMSE and better score

Figure 5: Performance comparison of different models



7. CONCLUSION

- Throughout this project, several models which can predict Pm2.5 levels and classify them into different pollution bands were experimented 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. We obtained several notable outcomes from the predictive models that are worth being discussed.
- Different approaches to handle null values yielded varied performance from each of the models, however simply dropping the records that had null values seemed to be the best approach. Between obtaining the AQI by predicting the PM2.5 values and using a classifier to predict the AQI band straight away, the classifier seemed to perform better. A regression model could be used for applications in data analytics, but it is concluded that classifier models perform better for air quality prediction.