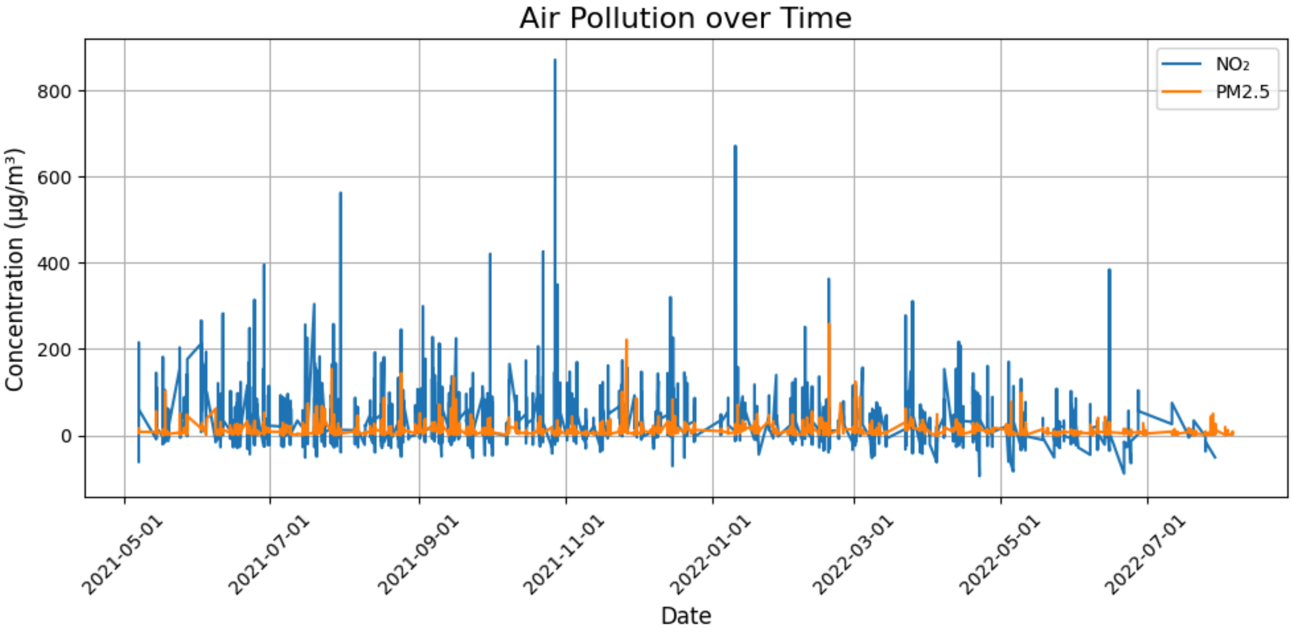
Google Air View Dataset Overview

The original Google Air View dataset provided by Dublin City Council consists of over 5 million records of **air quality** data collected at high spatial and temporal resolution across the city. This raw dataset includes a wide range of atmospheric pollutants such as nitrogen dioxide (NO_2), particulate matter (PM2.5 and other PM channels), carbon monoxide (CO), carbon dioxide (CO_2), and ozone (O_3).

After preliminary exploration, **NO_2** and **PM2.5** were selected as focal indicators due to their public health relevance and relatively more consistent coverage. However, NO_2 was ultimately excluded from the final analysis as it contained a large number of missing values and erratic behaviour, including negative values and extreme spikes. In contrast, PM2.5 had more reliable coverage and was retained for all subsequent analyses and modelling.



Hourly concentrations of NO_2 and PM2.5 over time. While PM2.5 displays relatively stable trends, NO_2 exhibits extreme spikes and variability, contributing to its exclusion from further analysis.

The dataset required substantial refinement due to missing and noisy values. It underwent essential preprocessing to address quality issues and enhance consistency. This ensured the data was structured appropriately for analysis and integration with contextual variables relevant to air pollution in Dublin. However, several aspects of the data collection process limit the scope and representativeness of the dataset, such as:

- **Temporal Coverage Limitation:** Data was collected on weekdays, predominantly between 9 AM and 5 PM, excluding evenings, nights, and weekends when air quality conditions might differ.

- **Lack of Contextual Information:** The dataset lacks details on land use (e.g., residential vs. industrial areas) and air quality standards, limiting interpretation of health or policy relevance.

Despite its level of detail, the Air View dataset alone does not provide enough contextual information to fully understand or predict changes in air quality. The absence of environmental factors such as weather conditions makes it difficult to explain variations in pollution levels. These limitations guided the decision to **incorporate an external weather dataset**, allowing examination of how variables like temperature, wind, humidity, and rainfall interact with pollution readings.

Air Pollution Factors

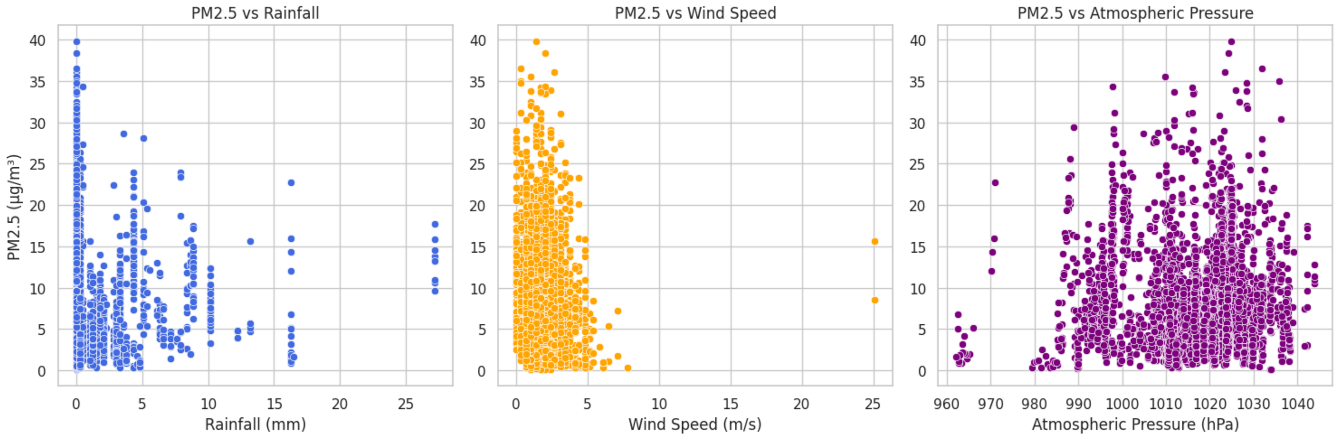
Weather was selected as the key factor in the analysis because it directly influences how air pollutants behave in the atmosphere. To complement the pollution data, weather information was incorporated from **WOW-IE** (Weather Observations Website Ireland), a platform managed by **Met Éireann** in partnership with the global WOW network, and selected the ABC Weather Station, close to Croke Park in Dublin city.

To prepare the weather data for analysis, the following steps were performed:

- Reducing the granularity by keeping only the first observation of each minute, as weather does not typically vary meaningfully at the second level;
- Filtering the air pollution dataset to include only records within a 2 km radius of the selected weather station using a Euclidean distance approximation;
- Merging both datasets on minute-level timestamps (through datetime variable), allowing for a ±6-minute tolerance to account for minor timing mismatches;
- Removing extreme outliers to improve model reliability.

This cleaning and preprocessing pipeline resulted in a final merged dataset with 3,986 high-quality observations, suitable for analysis and modeling. The retained key weather variables relevant to air pollution analysis include: **temperature**, **relative humidity**, **wind speed**, **rainfall**, and **atmospheric pressure**.

To better understand how environmental conditions influence air quality, the relationship between PM2.5 concentrations and key weather factors (plots below) were visualised. Higher PM2.5 levels were generally associated with low rainfall, low wind speed, and high atmospheric pressure, suggesting that **dry, calm, and stable weather conditions** contribute to local pollution buildup, while **rain and wind help disperse or remove particles,** "cleaning” the air.



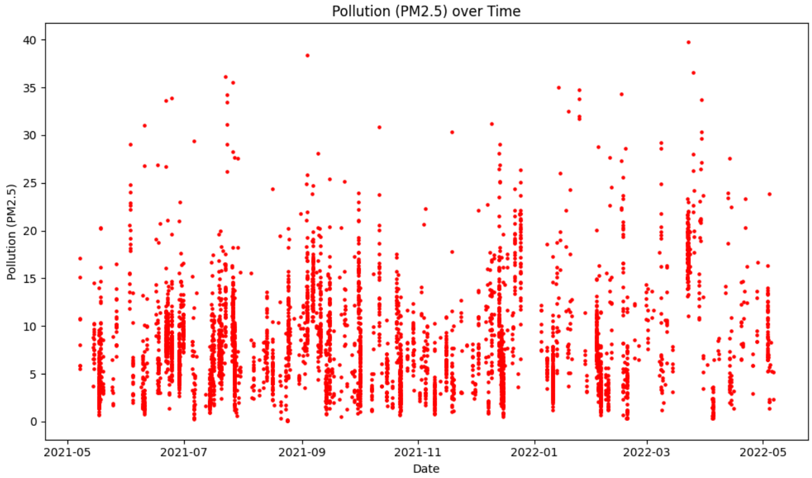
Selected scatter plots reveal how weather conditions relate to air quality.

To control for variations in traffic-related emissions, time-level proxies such as **rush hour** indicators, **school holiday** flags, and **day-of-week** effects were included.

The analysis focused on a limited set of variables from a single weather station, which may not fully capture local variations across the city. Future work could improve precision by incorporating high-resolution spatial data, satellite imagery,or modeling pollutant dispersion under varying meteorological conditions could improve the precision and policy relevance of the findings.

Modelling Approach

Before diving into the modelling approach, the evolution of PM2.5 pollution levels over time was examined. The figure on the right illustrates clear fluctuations throughout the year, highlighting daily and seasonal patterns likely influenced by weather conditions and human activity.

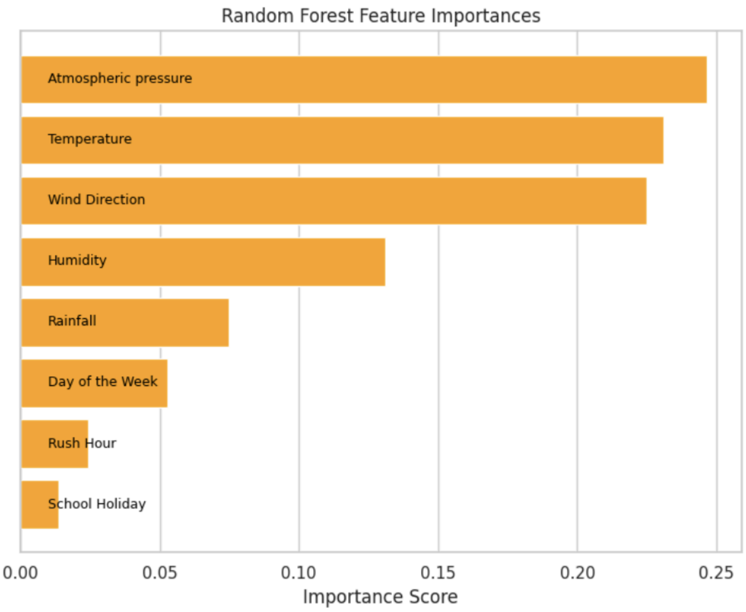


To explore how weather and time-based factors predict air pollution levels, a Random Forest Regressor was applied using PM2.5 as the target variable. The model was trained on 80% of the data and validated on the remaining 20%. The trained model was subsequently applied to the full dataset for analysis. It incorporated variables like temperature, humidity, rainfall, wind, atmospheric pressure, and also indicators for rush hours, school holidays and weekdays. Model performance metrics showed:

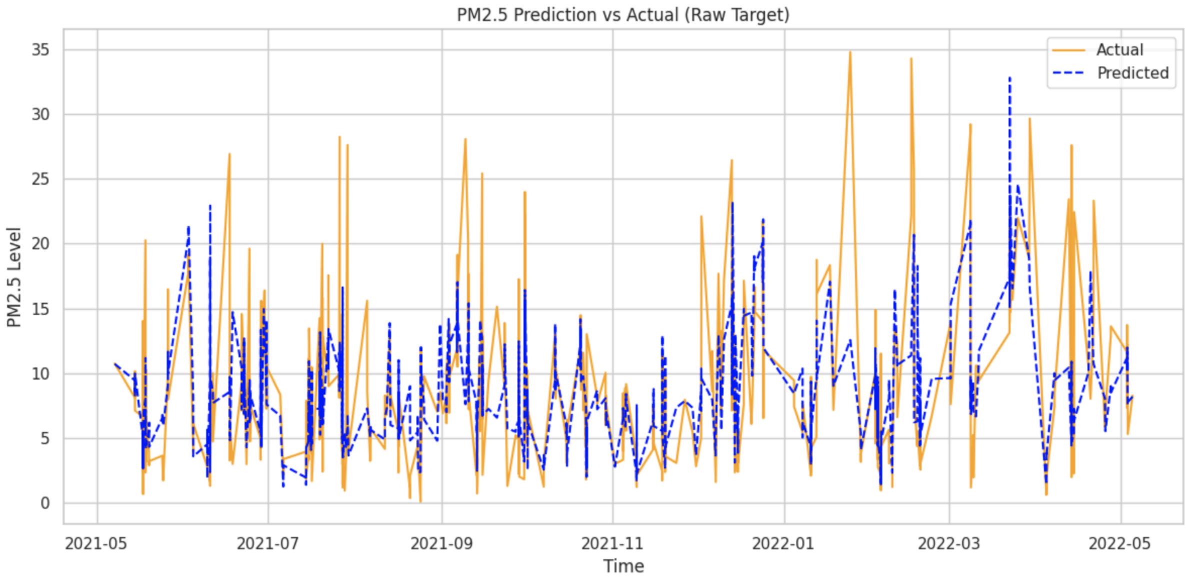
- **Mean Absolute Error (MAE):** 2.60 µg/m³ indicating that on average, the model's predictions were very close to the observed PM2.5 values.
- **Mean Squared Error (MSE):** 15.95, reflecting a low overall level of prediction error and confirming the model's stability despite occasional outliers.
- **Root Mean Squared Error (RMSE):** 4 µg/m³, meaning the model's average prediction error was relatively low and within a reasonable range.
- **R-squared score:** 0.49, meaning the model explained nearly 49% of the variance in PM2.5 levels, which suggests moderate predictive power.

To better understand which variables influenced the model's predictions most, the **feature importances** from the Random Forest algorithm were examined. The top predictors of PM2.5 levels are:

- Atmospheric pressure
- Temperature
- Wind direction



The predicted PM2.5 values against actual levels in the test set were visualized. This plot illustrates how closely the Random Forest model's predictions align with the actual PM2.5 levels recorded between mid-2021 and mid-2022. Overall, the predicted values (blue dashed line) follow the same general trends and seasonal patterns as the actual observations (yellow line), capturing periods of rising and falling pollution reasonably well.



Comparison of actual and predicted PM2.5 pollution levels over time using a Random Forest model.

The model tends to slightly underestimate extreme peaks and smooth out sharp fluctuations, which is common in tree-based models when dealing with highly variable environmental data. However, the model maintains a decent level of accuracy, particularly in moderate pollution ranges.

Given its ability to capture key temporal patterns and pollution shifts, this Random Forest model also holds potential as a short-term **forecasting tool**. It could be used to **predict PM2.5 levels for the following day(s)**, supporting the implementation of preventative measures when poor air quality is expected based on forecasted weather and time-related indicators.

Policy Recommendations

The model results highlighted higher pollution levels during **dry, high atmospheric pressure** and **calm wind conditions** days. While weather cannot be altered, public behavior can be influenced. Hence, It is recommended that DCC implement a **Smart Mobility & Awareness Program** consisting of:

- **Forecast-Based Alerts:** Use weather forecasts to identify high-risk pollution days (dry, calm, high-pressure) and issue real-time alerts via DCC apps, signs, and social media.

- **Eco-Incentives for Clean Travel:** Encourage public transport with app-based rewards (e.g., “Today is a high-pollution day. Use public transport and get 20% off your fare.”).

- **Boost Green Mobility Options:** Increase bus frequency during risk periods, expand protected cycling lanes, and support e-bike and electric vehicle adoption.

- **Discourage Car Use on Risk Days:** Apply flexible parking restrictions, carpool incentives, or temporary car-free zones during forecasted pollution spikes.

This solution aims to reduce emissions on pollution-prone days by shifting how people move around the city, using both nudges and infrastructure support.