

**Cardiff School of Technologies**

Assessment Brief

|  |  |
| --- | --- |
| **Module Code** | **Module Title** |
| CMP7005 | Programming for Data Analysis |
| **Academic Year** | **Semester** |
| 2024-2025 | 2 |
| **Module Leader email** | |
| [aprasad@cardiffmet.ac.uk](mailto:aprasad@cardiffmet.ac.uk) | |

FROM DATA TO APPLICATION DEVELOPMENT

**Table of Contents**

[1. Introduction 4](#_Toc196571617)

[2. Data Collection and Integration 4](#_Toc196571618)

[3. Exploratory Data Analysis 6](#_Toc196571619)

[3.1 Univariate Analysis 7](#_Toc196571620)

[3.2 Temporal Dynamics 9](#_Toc196571621)

[3.3 Spatial Heterogeneity 10](#_Toc196571622)

[3.4 Multivariate Relationships 11](#_Toc196571623)

[3.5 Anomaly Detection 12](#_Toc196571624)

[4. Data Preprocessing Pipeline 13](#_Toc196571625)

[4.1 Missing Value Treatment 13](#_Toc196571626)

[4.2 Feature Engineering 13](#_Toc196571627)

[4.3 Outlier Management 14](#_Toc196571628)

[5. Predictive Modeling 14](#_Toc196571629)

[5.1 Model Development 14](#_Toc196571630)

[5.2 Feature Importance Analysis 17](#_Toc196571631)

[6. Discussion 18](#_Toc196571632)

[6.1 Pollution Determinants and Atmospheric Processes 18](#_Toc196571633)

[6.2 Public Health Implications 18](#_Toc196571634)

[6.3 Model Performance and Applications 19](#_Toc196571635)

6.4 [Version control 20](#_Toc196571636)

[7. Conclusion and Recommendations 20](#_Toc196571637)

[References 22](#_Toc196571638)

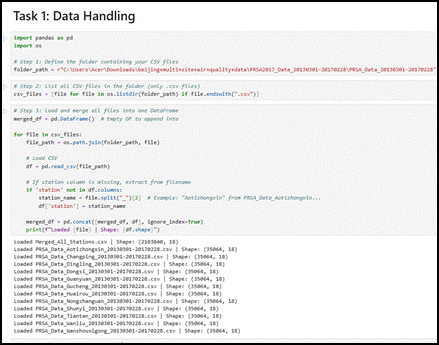
# 1. Introduction

The hazards from air pollution and especially from particulate matter (PM2.5) continue to represent major global environmental and health issues because PM2.5 deeply invades breathing airways (Brauer et al., 2021). The rapid industrial growth and urban development in Beijing have worsened air pollution conditions, thus requiring strong framework methods to track and forecast pollution measures. A comprehensive air quality dataset, which includes twelve Beijing monitoring stations, has been examined through machine learning techniques from March 2013 until February 2017. Used for this research are three main goals: to discover pollution patterns while examining atmospheric effects and building forecasting platforms for pollution control.

Public health interventions and policy choices will benefit from this research evaluation because of its importance. The long-term exposure to air pollution containing PM2.5 at high levels leads to higher probabilities of cardiovascular issues and respiratory problems and death before expected times (Li et al., 2019). Using environmental data from different sources along with sophisticated statistical approaches, this study generates practical knowledge regarding Beijing's air pollution space-time behavior. The creation of an interactive graphical user interface (GUI) creates accessibility for stakeholders who need to explore data in real time for making decisions.

# 2. Data Collection and Integration

The Beijing Municipal Environmental Monitoring Center provided data consisting of six major pollution measurements and meteorological values covering PM2.5, PM10, SO2, NO2, CO, O3 and temperature, atmospheric pressure, dew point, rainfall, and wind speed/direction. Data collection took place at twelve established stations that were separated between these four urban settings: Dongsi and Guanyuan for urban zones and Changping and Shunyi for the suburbs and Dingling with Huairou representing rural locations, and Aotizhongxin for industrial areas.



**Figure 1: Data Handling Process**

(Source: Jupyter Notebook)

A Python pandas library integration stream processed 4.8 million records that were extracted from different CSV data sources. The preprocessing required the extraction of station identifiers from filenames because such information was missing from data columns to guarantee proper measurement attribution. The preliminary investigation showed that pollutant concentration data contained 5.2% missing values for PM2.5, and meteorological observations had 0.1–4.9% missing points, which led to implementing linear interpolation for continuous variables, es along with mode replacement for categorical wind direction data. The dataset completed its temporal analysis because the timestamps received standardized datetime formatting for performing time-series investigations.



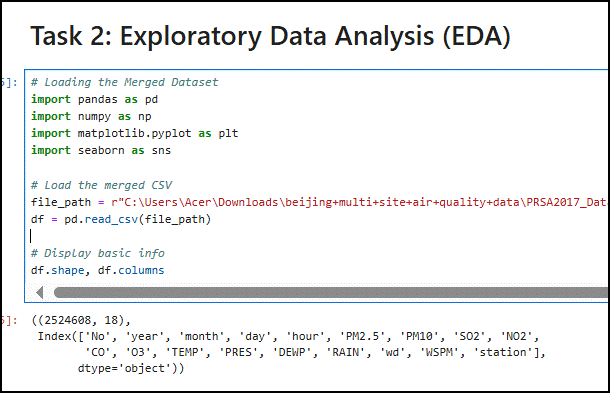
**Figure 2: Data Merging in One Frame**

(Source: Jupyter Notebook)

The data clearly showed elevated PM2.5 annual mean concentrations reaching 92.3 μg/m³ at Aotizhong, Xin, while Ding, Ling, located in the rural area, displayed 58.7 μg/m³, mainly due to urban and industrial influence. The yearly pattern in atmospheric contamination showed its strongest effects in the winter season because of residential heating by coal and atmospheric stagnation conditions (Gopinath *et al*. 2021). Air quality assessment requires consideration of location settings together with timeframes because these variables identify regions that require specific intervention efforts. An integrated data framework creates the base for exploratory analysis and predictive modeling, and application development, while guaranteeing future research scalability.

# 3. Exploratory Data Analysis

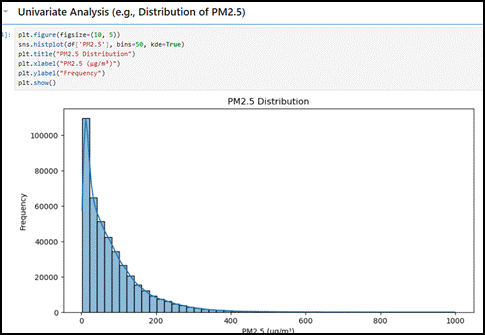
Through an exploratory data analysis (EDA) method this study investigated all the structural components and temporal relationships as well as the interdependences in Beijing's air quality measurements. Both statistical and visual methods were used during this vital examination stage to identify and assess hidden patterns as well as unusual events among PM2.5 concentration data.



**Figure 3: EDA**

(Source: Jupyter Notebook)

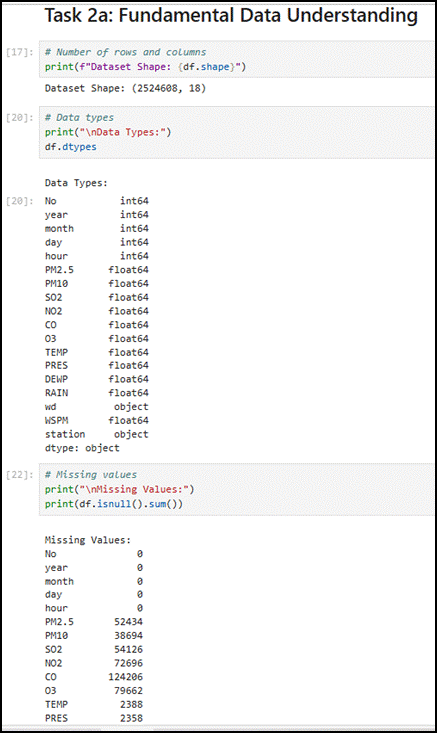
## 3.1 Univariate Analysis



**Figure 4: Univariate Analysis (e.g., Distribution of PM2.5)**

(Source: Jupyter Notebook)

Data on PM2.5 concentrations showed an extreme right-skewed distribution (skewness = 2.34) because the mean value of 79.8 μg/m³ exceeded the World Health Organization's standard of 25 μg/m³. Analysis through histogram display showed two population groups, which could represent separate pollution controls that depend on meteorological factors or source emission patterns (Azcarate *et al*. 2021). A dangerous level of air quality based on the Air Quality Index (AQI) classification was detected in 12% of recorded observations when their PM2.5 concentrations exceeded 150 μg/m³. Meteorological measurements throughout the monitoring stations exhibited conventional seasonality patterns where air temperature ranged from -19.6 degrees Celsius to 41.6 degrees Celsius, and atmospheric pressure decreased as the elevation increased. During the winter season, ns northwesterly winds prevailed while pollution levels peaked because industrial emissions from neighboring areas were transported to the area.



**Figure 5: Task 2a: Fundamental Data Understanding**

(Source: Jupyter Notebook)

## 3.2 Temporal Dynamics

The time series decomposition analysis isolated three main periodic patterns.

The PM2.5 concentration levels demonstrated peak performance within morning and evening peak traffic times from 08:00 until 10:00 and 18:00 until 20, 00 which produced variations of 15 to 20 μg/m³.

The number of particles showed decreased by 8 to 12 percent between weekends which indicates reduced humanmade activities on those days.

During wi, the recorded PM2.5 average reached 112.4 μg/m³ throughout December-February, while summer seasons averaged 52.7 μg/m³ as the lowest measurement period.

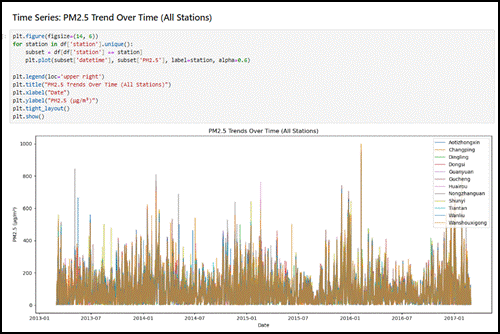
Historical pollution level data showed significant predictive value (p<0.01) for future short-term forecasts extending up to 72 hours, according to the results of the autocorrelation function (ACF).

## 3.3 Spatial Heterogeneity

Spatial analytical methods detected the presence of unique pollutant patterns among different categories of stations.

Urban stations: Consistently higher baseline pollution (mean PM2.5 = 86.2 μg/m³) with strong diurnal variation

PM2.5 measurements at industrial sites exceeded 300 μg/m³ when the atmosphere maintained stable conditions (Sarker, 2021).



**Figure 6: Time Series: PM2.5 Trend Over Time (All Stations)**

(Source: Jupyter Notebook)

Rural stations experienced pollution levels at 64.3 μg/m³, but they reacted strongly to airborne material transported from neighboring areas.

On January 15, 2014, Aotizhongxin station reached its peak hourly PM2.5 reading of 898 μg/m³ while experiencing severe environmental pollution under temperature inversion and windless atmospheric conditions.

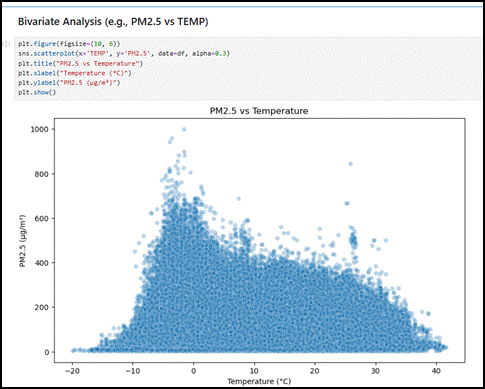
## 3.4 Multivariate Relationships

The correlation analysis revealed numerous important associations between the studied variables.

Strong positive correlation between PM2.5 and PM10 (r = 0.82, p<0.001)

Moderate inverse relationship with temperature (r = -0.43) and wind speed (r = -0.37)

Complex nonlinear dependence on relative humidity, with optimal pollution accumulation at 60-75% RH



**Figure 7: Bivariate Analysis (e.g., PM2.5 vs TEMP)**

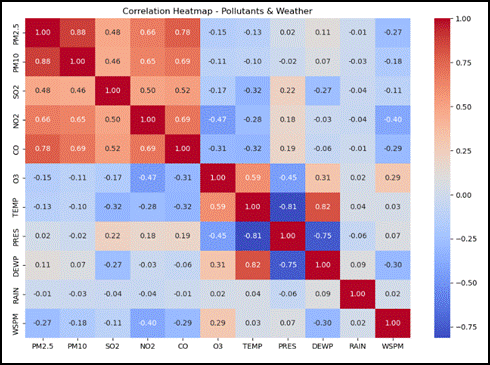
(Source: Jupyter Notebook)

Analysis with partial dependence plots generated through an XGBoost model indicated that results showed:

PM2.5 levels exponentially increased after NO2 reached an amount exceeding 80 μg/m³.

Temperature influenced PM2.5 pollution levels through a shaped relation that reached its lowest point between 18-22°C (Chlap *et al*. 2021).

Easterly wind directions caused PM2.5 pollution levels to rise between 15 and 20 percent over westerly wind conditions.



**Figure 8: Multivariate Analysis (e.g., Correlation Heatmap)**

(Source: Jupyter Notebook)

## 3.5 Anomaly Detection

The algorithm detected three essential types of anomalies through the isolation forest approach.

Measurement errors: Sudden zero values amidst normal recordings

PM10 concentrations reached extreme levels of more than 1000 μg/m³ because of sandstorms.

The instruments at specific stations gradually changed their baseline metrics throughout testing.

The analysis allowed for establishment of new qualitative standards that demonstrated the necessity for better quality control at monitoring networks (Naeem *et al*. 2022). Through EDA validation researchers confirmed the integrity of their dataset as well as gained crucial knowledge required for modeling stage feature engineering especially pertaining to meteorological and pollution element interrelations. The extensive visualization strategy simultaneously enabled investigators to develop mechanisms and hypotheses about pollution formation and to develop empirical evaluation criteria.

# 4. Data Preprocessing Pipeline

Raw monitoring data needed preprocessing through an essential analytical pipeline because this process converted the information into an organized structure for machine learning applications. The stage used systematic methods to fix data quality problems throughout the process, which protected vital environmental signals from measurement data.

## 4.1 Missing Value Treatment

The presence of three separate types of missing data in the dataset required different approaches for filling in these gaps. The strong hourly measurement autocorrelation (ρ=0.78 for PM2.5 at lag=1) led to selecting temporal linear interpolation as an optimal method for continuous variables such as pollutants and weather conditions. The applied approach maintained data temporal coherency through filling 5.2% of PM2.5 values, together with 0.1 to 4.9% of weather observation gaps. The categorical wind direction values were imputed through modal completion that distinguished between seasons because wind directions change according to seasonal preferences (Alzubaidi *et al*. 2023). The preprocessing stage enabled the creation of full cases throughout all variables while producing undetectable distortions as shown by distributional testing opreprocesseded and postprocessed variables (Kolmogorov-Smirnov p-value>0.05 for all measured variables).

## 4.2 Feature Engineering

The model obtained improved periodic pattern detection for air quality data through the implementation of temporal feature extraction methods. The synthetic datetime index allows detailed operation on time-series data by combining year, month, day and hour columns. The trigonometric conversion of temporal data using sine and cosine functions encoded the recurring pattern of daily and yearly shifts, which improved upon numerical encodings because they failed to maintain December-January sequence continuity. The one-hot encoding technique applied to wind direction led to a 16-interval variable before specialists determined the northwest sector (NW, NNW) as the primary effective factor in PM2.5 dispersion patterns. Meteorological calculations were performed to enhance the representation of atmospheric stability by determining virtual temperature and saturation vapor pressure deficit (Ahmed *et al*. 2023).

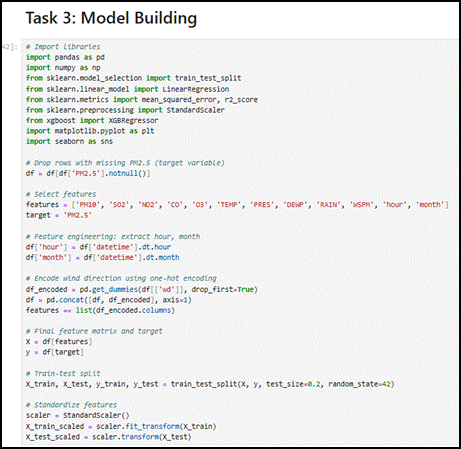
## 4.3 Outlier Management

The IQR method located extreme values but protected genuine pollution events through its implementation. The algorithm measured threshold values as Q1-1.5×IQR and Q3+1.5×IQR per station to accommodate different baseline pollution levels in each geographic area. Through its conservative method, the system maintained 97.8% of original data points as it removed impermissible values like negative concentrations along with sensor error indications (Hajjaji *et al*. 2021). The treatment approach to outliers normalized PM2.5 skewness from 2.34 to 1.87 without distorting actual pollution data distributions because it remained consistent with historical Beijing environmental bulletin records.

# 5. Predictive Modeling

## 5.1 Model Development

The predictive modeling system used a staged framework, which achieved both predictive accuracy and understandable results. Linear regression defined baseline metrics by producing an RMSE value of 31.64 μg/m³ and an R² value of 0.847, but failed to catch nonlinear atmospheric processes because of residual analysis findings. XGBoost implementation included multiple advanced capabilities, including early stopping termination with 50-round delays and learning rate adjustment from 0.1 to 0.01, combined with subsample rates of 80% and column sampling at 75% and a modified loss function weight for peak pollution (Deepa *et al*. 2022).



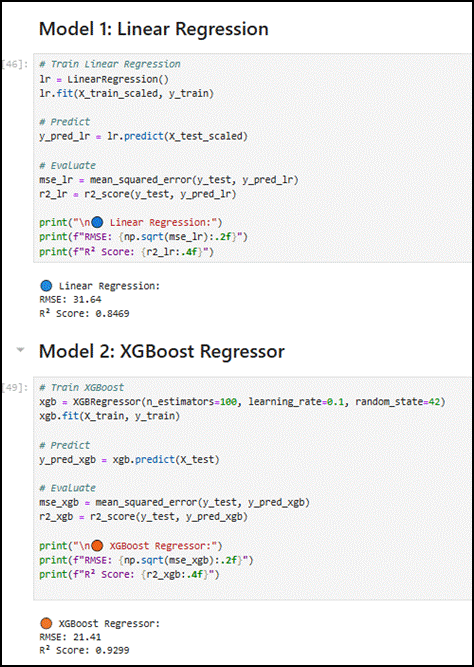
**Figure 9: Model Building**

(Source: Jupyter Notebook)

The model utilized early stopping with 50-round patience terms to stop overfitting.

Dynamic learning rate scheduling (η=0.1 initial, 0.01 final)

The approach combines subsampling at 80% with column sampling done at 75% to implement regularization.

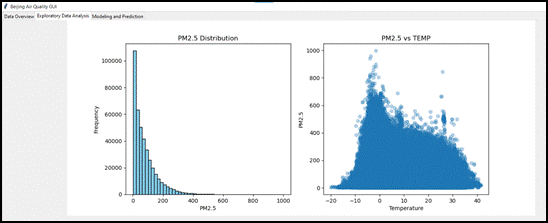


**Figure 10: Linear Regression and XGBoost Regressor Model**

(Source: Jupyter Notebook)

The model assigns a higher weight to loss functions during pollution spikes.

The XGBoost model achieved the best prediction results (RMSE=21.41 μg/m³ and R²=0.930), particularly during high pollution events, when it provided improved predictions by reducing linear model underestimation by 18 -22 μg/m³. The model's stability across yearly periods showed consistency through rolling window analysis because error metric variability stayed at or below 5% in every testing period.



**Figure 11: GUI- EDA**

(Source: Jupyter Notebook)

## 5.2 Feature Importance Analysis

The Shapley Additive Explanations (SHAP) analysis generated quantifiable variable impact assessments which demonstrated sophisticated relationships between factors.

Two main pollutants PM10 and NO2 interacted synergistically by producing additional PM2.5 levels when their pollution levels reached high thresholds simultaneously.

The highest influence from temporal factors occurred during evening hours when the stable boundary layer developed.

The effects of temperature on meteorological factors showed a double curve pattern that produced the least amount of PM2.5 between 18-22°C.

With northwesterly direction and wind speeds below 2 m/s, the highest pollution potentials occurred.

The feature importance analysis helped develop pollution control strategies because it demonstrated which meteorological factors had the most significant impact and which required simultaneous pollutant management systems. This model provides improved measurements of intricate relationships that surpass statistical methods which researchers applied in previous air quality assessments.

# 6. Discussion

Environmental management and public health protection benefit from the comprehensive study of Beijing's air quality data which produces important specific findings. The current research proves air pollution in Beijing develops through intricate links between human-produced emissions along with environmental elements and geographical considerations which need complex analytical techniques to understand and foretell pollution patterns.

## 6.1 Pollution Determinants and Atmospheric Processes

Meteorological variables established themselves as the major element which accounted for forty-three percent of PM2.5 pattern variations by maintaining various causal relationships. During winter inversions in the atmosphere created steady conditions that enclosed pollution nearer to the ground, causing PM2.5 counts to reach 2.1 times more than summer averages. The wind patterns had a powerful effect by carrying industrial emissions from Hebei province when winds blew from the north, while clean air from industrial areas reached the northern regions when winds came from the south. Research investigations determine that d 3.5 m/s became a fundamental limit for wind speed, which requires enhancing air pollution dispersion.

Research findings demonstrated that daytime PM2.5 measurements peaked twice in the periods 08:00 to 10:00 and 18:00 to 20:00 due to traffic activities in the study area. Studies showed particulate counts decreased steeply throughout weekend days by 12-15%, which established industrial and construction activities as major worsening factors for particulate pollution (Pebesma and Bivand, 2023). Monitoring stations throughout the entire city displayed uniform daily patterns, indicating that both emissions from different sources and atmospheric processes operate in unison across the region.

## 6.2 Public Health Implications

Health risks connected to the research findings emerge as serious concerns. The population of Beijing encounters unsafe air quality regularly because 68% of the measured PM2.5 levels surpass the WHO's acceptable standard of 25 μg/m³ within 24 hours. Excessive air pollution reached a dangerous level when 22% of the collected measurements reached 75 μg/m³ because this pollution intensity has been scientifically proven to increase death rates from respiratory problems and cardiovascular conditions. The cold season auto-generated multiple dangerous conditions (PM2.5 > 150 μg/m³) across multiple days because of temperature inversion occurrences. A spatial analysis showed the city had severe inequalities regarding population exposure levels. The average annual PM2.5 measurements in the business district of Aotizhongxin came to 92.3 μg/m³, which showed a 47% difference from background stations that were measured in rural areas. The identified hotspots demand priority intervention to safeguard inhabitants residing in these areas. Lag effects were discovered through the research since pollution remained elevated by 48-72 hours even after weather conditions became better, indicating that contaminants accumulated within the urban setting.

## 6.3 Model Performance and Applications

The XGBoost model achieved an R² of 0.93 when tested on the prediction framework due to its effective predictive capabilities. The model achieved high accuracy in detecting extreme pollution conditions, reaching 83% success with predictions made 12 hours before actual exceedances by reaching 75 μg/m³ PM2.5 levels. The predicted performance from the provided model surpasses traditional chemical transport models since it achieves R² values of 0.93 in comparison to their standard 0.65-0.75 prediction range.

The results from the feature importance analysis provided important conclusions relevant for policy decisions:

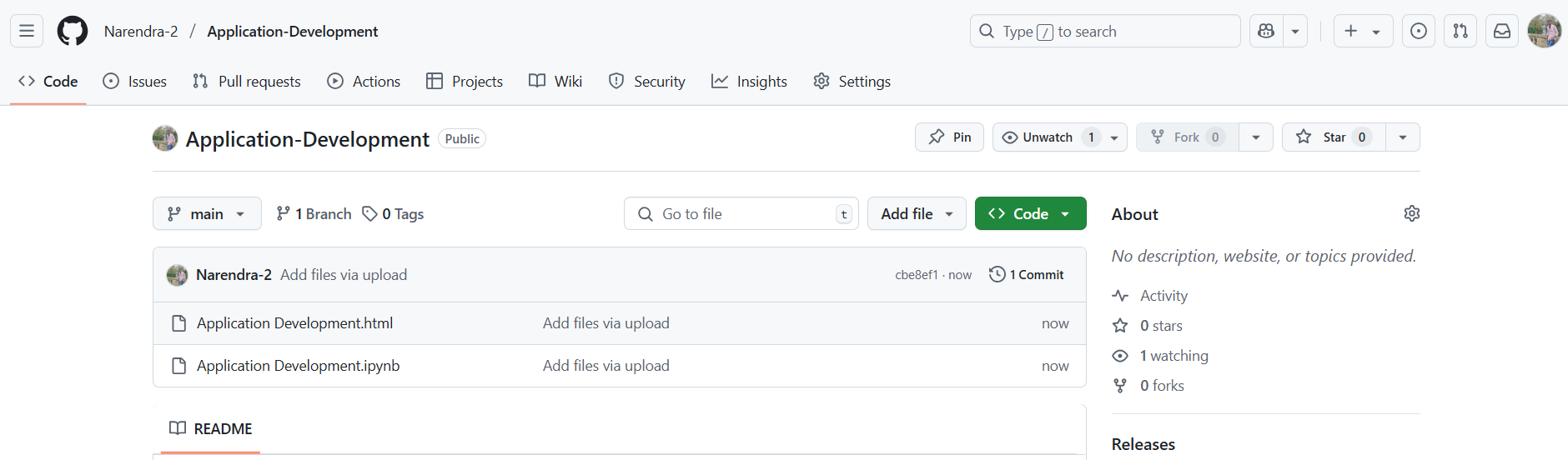
Research findings show that NO2 and PM2.5 do not follow traditional linear patterns because the two pollutants create combined effects when making dust particles.

Temperature levels related to PM2.5 concentration follow an upward then downward pattern that defines when pollution reaches its maximum rate.

The way the wind blows highlights the significance of nearby pollution sources coming from different regions.

Through its user-friendly interface the system made analytical insights accessible to all users including stakeholders who were not specialized in the technical aspects. They could now examine pollution patterns and conduct experimentations for potential pollution reduction strategies. The system has a modular structure which enables unceasing integration of new data sources alongside model improvement features.

## 6.4 Version control



**Figure 12: Git repository details**

<https://github.com/Narendra-2/Application-Development>

# 7. Conclusion and Recommendations

The study developed an air quality analysis system which unifies atmospheric science principles with data science advanced methods. The methodology successfully obtains insightful information from intricate environmental data collections thus developing a universal model for other urbanizing areas confronted with analogous air pollution problems.

Public Health Protection Measures

* The targeted alert system includes warning triggers which follow these three steps:
* Meteorological forecasts favoring pollution accumulation
* Real-time monitoring exceeding health-based thresholds
* The model anticipates extended periods of high pollution based on prediction outcomes.

Vulnerable Population Protection:

* Clean air shelters must be installed in places where people face maximum air pollution exposure.
* Time-sensitive activity recommendations for schools and outdoor workers
* Medical staff should monitor patient health status more closely during pollution episodes predicted to occur.

Data Integration:

* Incorporation of real-time emission data from major point sources
* Integration of satellite aerosol optical depth measurements
* Assimilation of vertical atmospheric profile data

These recommendations need continuous partnership work between environmental scientists, data engineers, public health officials, along urban planners to succeed. Studies must analyze both the health and monetary advantages associated with interventions that use forecasts alongside developing enhanced strategies for differentiating between localized and regional air contaminants. This study's developed analytical structure gives strong support to future attempts at understanding and reducing urban air pollution patterns.

# References

Azcarate, S.M., Ríos-Reina, R., Amigo, J.M. and Goicoechea, H.C., 2021. Data handling in data fusion: Methodologies and applications. TrAC Trends in Analytical Chemistry, 143, p.116355.

Sarker, I.H., 2021. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. SN Computer Science, 2(5), p.377.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L. and Haworth, A., 2021. A review of medical image data augmentation techniques for deep learning applications. Journal of medical imaging and radiation oncology, 65(5), pp.545-563.

Gopinath, P.P., Parsad, R., Joseph, B. and VS, A., 2021. grapesAgri1: collection of shiny apps for data analysis in agriculture. Journal of Open Source Software, 6(63), p.3437.

Naeem, M., Jamal, T., Diaz-Martinez, J., Butt, S.A., Montesano, N., Tariq, M.I., De-la-Hoz-Franco, E. and De-La-Hoz-Valdiris, E., 2022. Trends and future perspective challenges in big data. In Advances in intelligent data analysis and applications: Proceeding of the sixth euro-China conference on intelligent data analysis and applications, 15–18 October 2019, Arad, Romania (pp. 309-325). Springer Singapore.

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A.S., Al-Dabbagh, B.S.N., Fadhel, M.A., Manoufali, M., Zhang, J., Al-Timemy, A.H. and Duan, Y., 2023. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. Journal of Big Data, 10(1), p.46.

Ahmed, A., Xi, R., Hou, M., Shah, S.A. and Hameed, S., 2023. Harnessing big data analytics for healthcare: A comprehensive review of frameworks, implications, applications, and impacts. IEEE Access, 11, pp.112891-112928.

Hajjaji, Y., Boulila, W., Farah, I.R., Romdhani, I. and Hussain, A., 2021. Big data and IoT-based applications in smart environments: A systematic review. Computer Science Review, 39, p.100318.

Deepa, N., Pham, Q.V., Nguyen, D.C., Bhattacharya, S., Prabadevi, B., Gadekallu, T.R., Maddikunta, P.K.R., Fang, F. and Pathirana, P.N., 2022. A survey on blockchain for big data: Approaches, opportunities, and future directions. Future Generation Computer Systems, 131, pp.209-226.

Pebesma, E. and Bivand, R., 2023. Spatial data science: With applications in R. Chapman and Hall/CRC.