IST652 Scripting for Data Analysis

December 9th, 2024

Air Quality Analysis and Forecasting

Final Project Report

Team Members

Tanu Rana

Rijul Chaturvedi

Ninad Bhikaje

Rushikesh Shinde

TABLE OF CONTENTS

ABSTRACT

The primary aim of this project is to analyze air quality patterns and forecast future levels using datasets from data.gov and nyc.gov. The study focuses on identifying the major pollutants affecting air quality in NYC, understanding regional and temporal variations, and predicting future air quality trends. Additionally, it investigates the health impacts of specific pollutants such as PM2.5. This report is intended for researchers, policymakers, environmental analysts, and public health officials who are interested in air quality management, urban planning, and mitigating health risks associated with pollution. The methodology involves:

- Data Cleaning and Preprocessing: Handling missing values, outlier detection, and consistent date-time formatting.
- Exploratory Data Analysis (EDA): Visualizing pollutant distributions and identifying seasonal trends.
- Correlation Analysis: Examining relationships between pollutants and external factors like weather, i.e Seasonal and Geographical Analysis
- Forecasting Models: Implementing machine learning models (e.g., Random Forest, XGBoost) to predict future pollutant levels.
- Health Impact Analysis: Correlating pollutant levels with health statistics (e.g., Cardiovascular Hospitalizations and Asthma rates).

The key findings from our results are that PM2.5 and NO2 are key pollutants influencing air quality in NYC. Seasonal trends were evident, with higher pollutant levels during certain periods such as winter. After hyperparameter tuning, the Gradient Boosting Regressor model achieved an R-squared score of 0.94, demonstrating strong predictive accuracy for PM2.5 levels. This project highlights critical insights into air quality dynamics in NYC, including the temporal and spatial distribution of pollutants and their health implications. The forecasting models provide a reliable tool for proactive air quality management and public health planning.

INTRODUCTION

Air quality has become a critical area of interest globally, particularly in urban environments like New York City (NYC). The increasing levels of pollutants such as PM2.5, NO2, and black carbon have raised significant concerns about their impact on public health and the environment. Understanding air quality patterns, identifying major pollutants, and predicting future trends are essential for effective urban planning, policy-making, and public health interventions.

The primary concern addressed in this study is the deteriorating air quality in NYC and its implications for public health. Pollutants such as PM2.5 and NO2 are known to cause respiratory and cardiovascular issues, especially among vulnerable populations like children and the elderly. Despite ongoing monitoring efforts, there is a pressing need to analyze temporal and spatial variations in air quality and establish predictive models to anticipate future pollution levels. This can help mitigate adverse health outcomes and guide regulatory measures.

Air pollution is a leading environmental risk factor for premature mortality worldwide. In NYC, poor air quality exacerbates chronic conditions such as asthma, particularly in underserved communities. Addressing this issue is vital for improving public health outcomes and achieving broader sustainability goals. Moreover, accurate forecasting of air quality can enable timely interventions, reduce healthcare costs, and enhance the overall quality of life for residents.

This paper leverages datasets from data.gov and nyc.gov to analyze historical air quality data, identify key pollutants, and understand their seasonal trends. Using advanced statistical methods and machine learning models like Gradient Boosting Regressor, the study aims to forecast future pollutant levels with high accuracy. The analysis also includes correlating pollutant concentrations with health outcomes to provide actionable insights into mitigating their impact.

LITERATURE REVIEW

Air quality analysis and forecasting are critical for understanding environmental and public health challenges, particularly in urban areas. The primary pollutants of interest such as PM2.5 are linked to adverse health outcomes, including respiratory and cardiovascular diseases. Monitoring and predicting air

quality trends enable policymakers and researchers to develop strategies to mitigate pollution and its impacts.

The study leverages datasets from data.gov and nyc.gov to explore air quality dynamics in New York City (NYC). Prior research highlights the importance of identifying temporal and spatial patterns in pollutant levels, as well as their correlation with external factors like weather. Existing literature emphasizes the need for predictive modeling to anticipate future air quality levels, which can inform proactive interventions. Following is a few Hypotheses/Research Questions:

- What are the primary pollutants contributing to poor air quality in NYC?
- How do pollutant levels vary across different seasons and geographic regions?
- Can advanced machine learning models improve the accuracy of air quality forecasting?
- What is the relationship between pollutant levels and health outcomes?

The Design Goals - The project aims to:

- Analyze historical air quality data to identify key patterns.
- Develop predictive models using machine learning techniques such as Gradient Boosting Regressor.
- Correlate pollutant trends with health statistics to provide actionable insights.
- Create visualizations that effectively communicate findings to stakeholders.

METHODOLOGY

The study employs a data-driven approach combining exploratory data analysis (EDA), statistical correlation analysis, and machine learning-based predictive modeling to address the research questions. The methodology is designed to analyze historical air quality data, identify key pollutants, and forecast future trends while correlating pollutant levels with health outcomes.

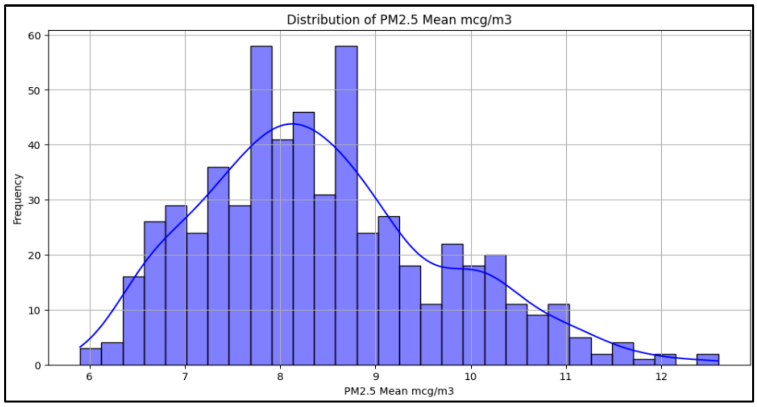
1) Data Collection and Preprocessing: The datasets were sourced from publicly available repositories such as data.gov and nyc.gov. These datasets include pollutant concentrations (e.g., PM2.5, NO2), over time from 2009 to 2022, weather variables, and health-related statistics.

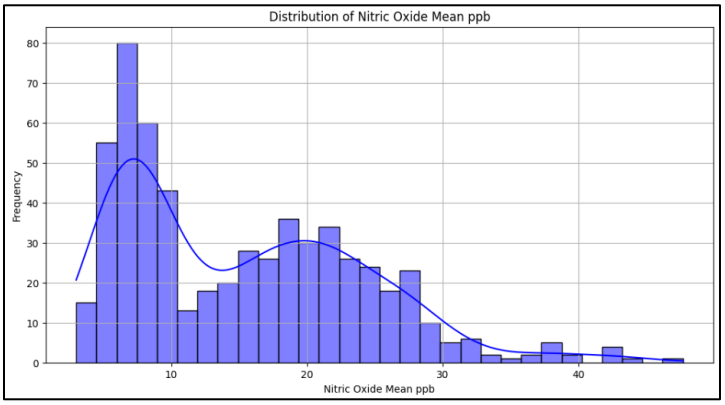
- Missing values in pollutant and health data were imputed using mean values to retain dataset integrity while minimizing information loss.

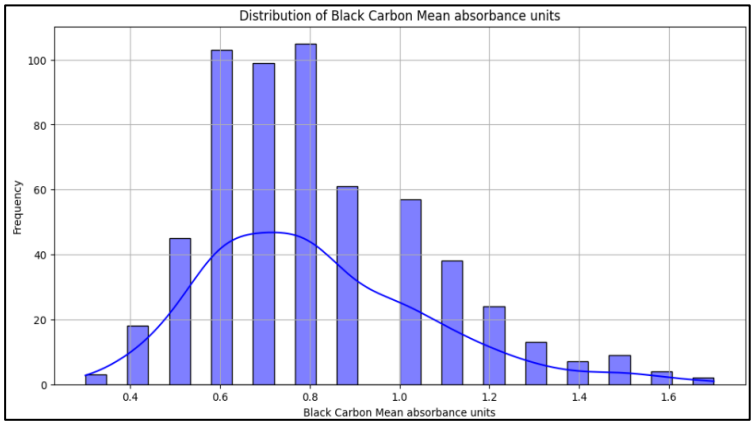
- Outliers in pollutant readings were identified and addressed to ensure robust analysis.
- TimePeriod fields were split into StartYear, EndYear, and Season columns for better temporal analysis.
- Feature engineering: Dummy variables were created for categorical features like seasons to enable machine learning models to process them effectively.

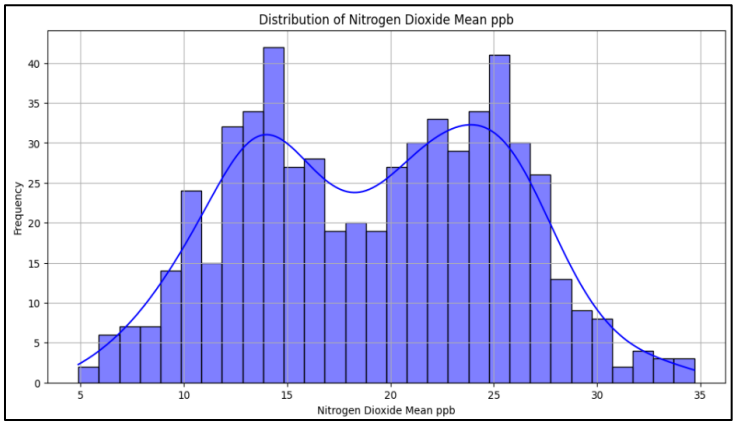
2) Exploratory Data Analysis (EDA)

- Histograms or Density plots for each pollutant

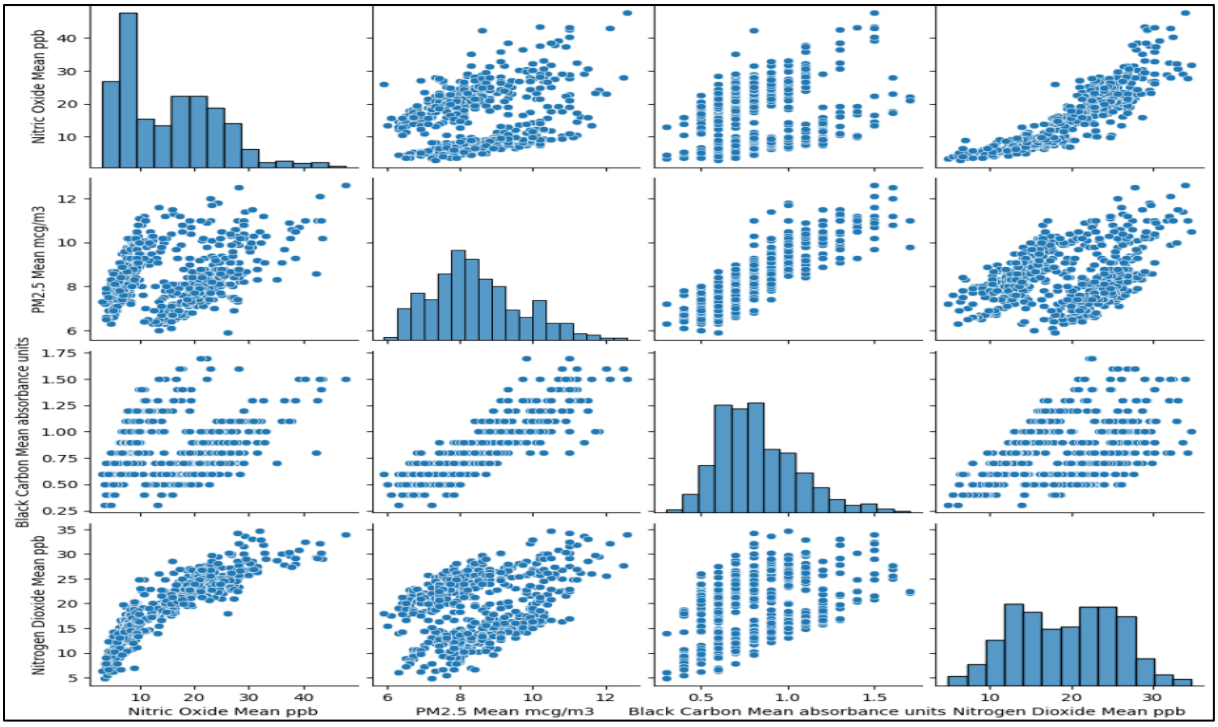




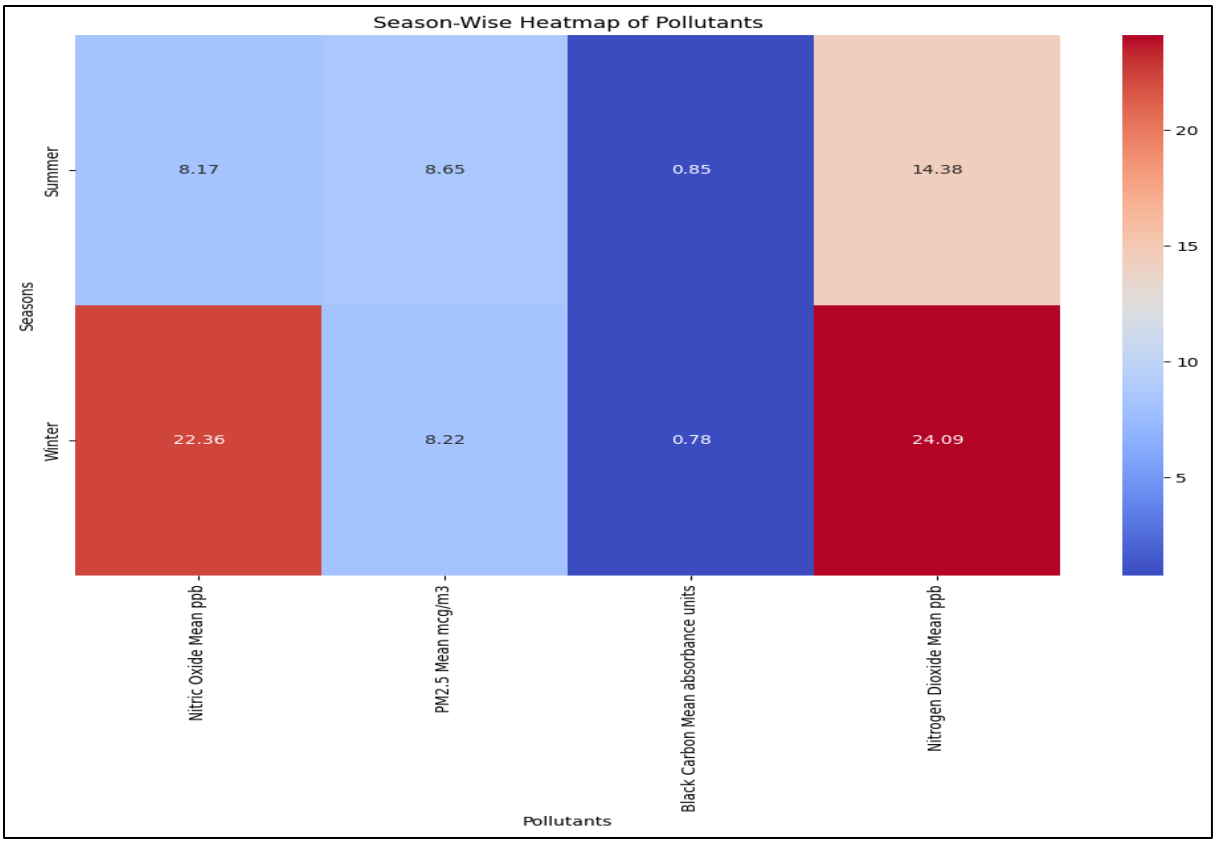




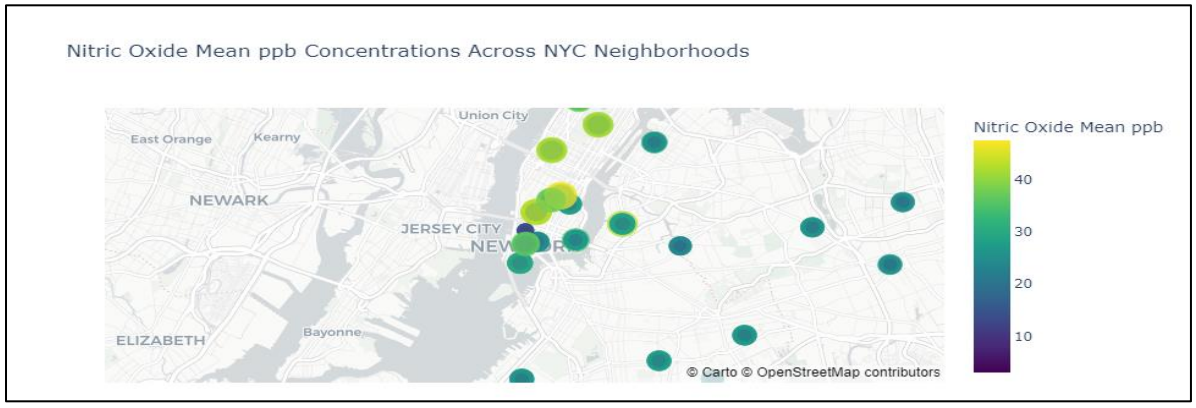
- EDA of Air Quality Data - Pollutants' Interaction

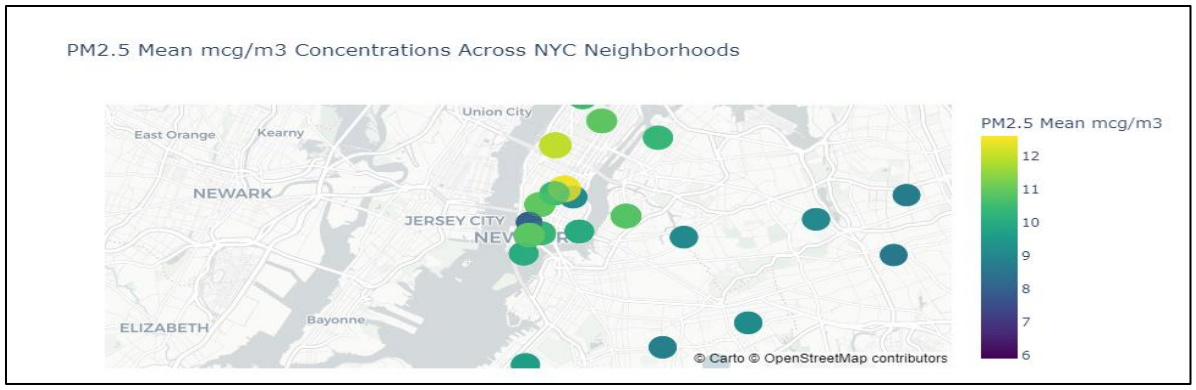


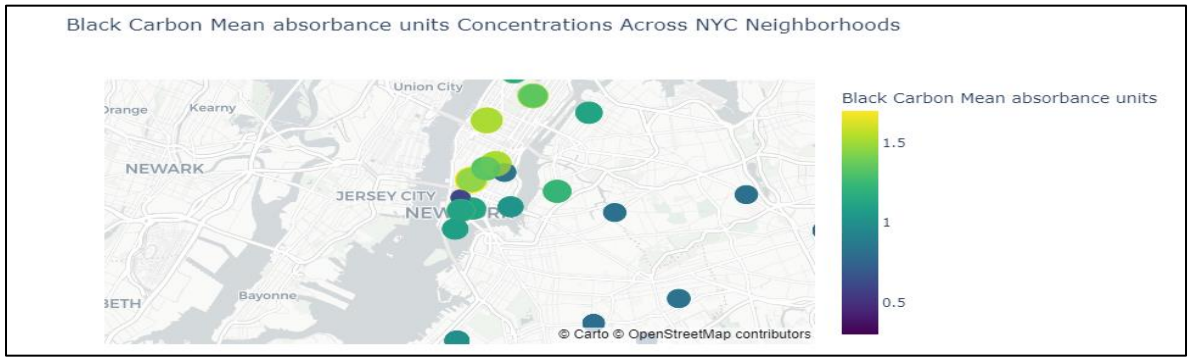
- Seasonal Analysis of each Pollutants

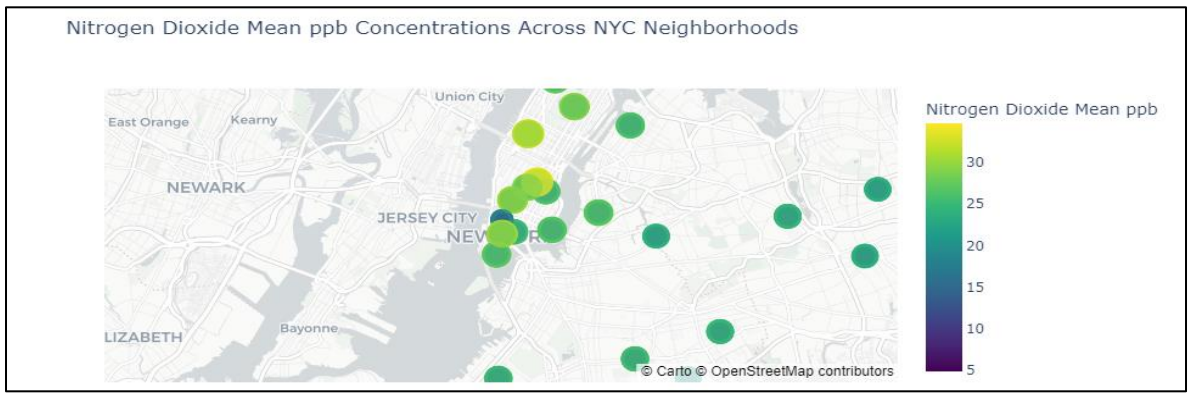


- Geographical Analysis of each pollutant

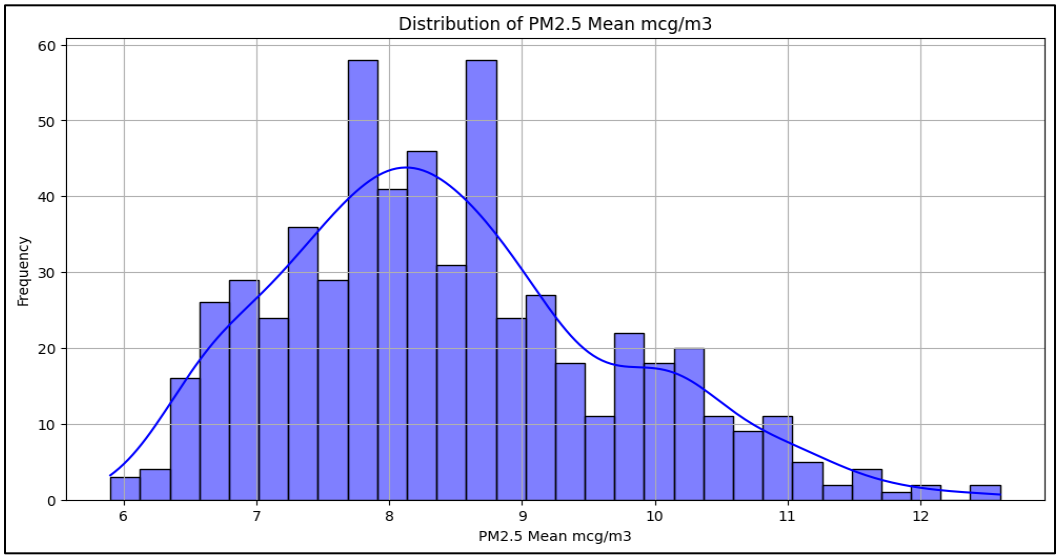


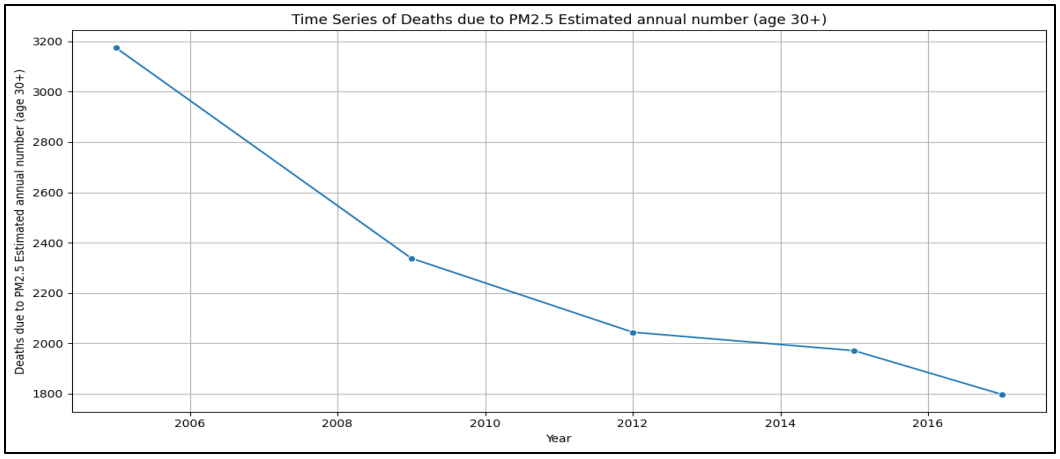






- Frequency Distribution and Time Series of PM2.5





3) Predictive Modeling

1. The Random Forest Regressor (RFR) was implemented to assess its performance in predicting PM2.5 levels. This ensemble learning method is robust against overfitting and captures non-linear relationships effectively.
2. Gradient Boosting Regressor (XGBoost) was chosen for its ability to handle complex, non-linear relationships in the data. The predictive models address the research question of forecasting future pollutant levels with high accuracy, aiding proactive air quality management.

Model Development:

- Splitting the dataset into training (80%) and testing (20%) subsets.
- Hyperparameter tuning using grid search to optimize model performance.

Evaluation Metrics:

- R-squared to measure goodness-of-fit.
- Mean Absolute Error (MAE) to assess prediction accuracy.

4) Health Impact Analysis: Correlating pollutant trends with public health data to quantify the impact of air quality on health outcomes. This step provides actionable insights into how pollution affects public health, informing targeted interventions for at-risk populations.

- How our Methodology Answers Research Questions:
    - The combination of EDA and correlation analysis identifies key pollutants and their temporal/spatial patterns, addressing questions about major contributors to poor air quality.
    - Predictive modeling provides accurate forecasts of pollutant levels, enabling proactive measures against future risks.
    - Health impact analysis bridges the gap between environmental data and public health outcomes, offering a comprehensive understanding of pollution's effects on human well-being.
    - This structured methodology ensures a robust framework for analyzing air quality dynamics while addressing both environmental and public health concerns effectively.

RESULTS

The dataset comprises two primary components:

- Air Quality Data: This includes pollutant concentrations such as PM2.5 (fine particulate matter), NO2 (nitrogen dioxide), nitric oxide, and black carbon. The data also contains geographic identifiers and timestamps to allow for temporal and spatial analysis.
- Health Data: This includes health-related statistics such as deaths and hospitalizations attributed to PM2.5 exposure, segmented by age groups and regions.

Defining Features

- Air Quality Dataset:
    - TimePeriod: Seasonal or annual timeframes.
    - Pollutants: Metrics like PM2.5 (mcg/m^3), NO2 (ppb), black carbon absorbance.

    - Geographic Information: Region identifiers (e.g., UHF42 codes) and rankings.
- Health Dataset:
    - Health Outcomes: Metrics such as asthma emergency visits and cardiovascular hospitalizations.
    - Rates: Annual rates per 100,000 for various health outcomes.

Summary Descriptions

- The air quality dataset spans multiple years, capturing seasonal variations in pollutant levels.
- PM2.5 and NO2 are critical pollutants with notable seasonal spikes.
- Health data reveals correlations between high pollutant levels and increased rates of asthma.

Preparedness for Hypotheses and Models

- The cleaned datasets allowed for effective exploratory data analysis (EDA) to identify trends, correlations, and seasonal patterns in pollutant levels.
- Feature engineering prepared the data for machine learning models, enabling accurate predictions of future air quality levels.
- Health data integration supported hypothesis testing regarding the relationship between pollution levels and health outcomes.

Key Insights from Data Preparation

The preprocessing steps ensured a high-quality dataset suitable for answering research questions about air quality dynamics, forecasting pollutant levels, and understanding their health impacts. The structured approach allowed seamless integration into predictive modeling workflows while maintaining analytical rigor.

Model Results for Air Quality Data

1) Gradient Boosting Regressor: The Gradient Boosting Regressor (XGBoost) was selected as the primary predictive model due to its ability to handle non-linear relationships and achieve high accuracy. The model was trained on the scaled dataset, and its performance was evaluated using metrics such as Mean Squared Error (MSE) and R-squared.

Initial Results:

- Mean Squared Error (MSE): 0.17

- R-squared: 0.91

  These results indicate that the model explains 91% of the variance in PM2.5 levels, demonstrating strong predictive performance.

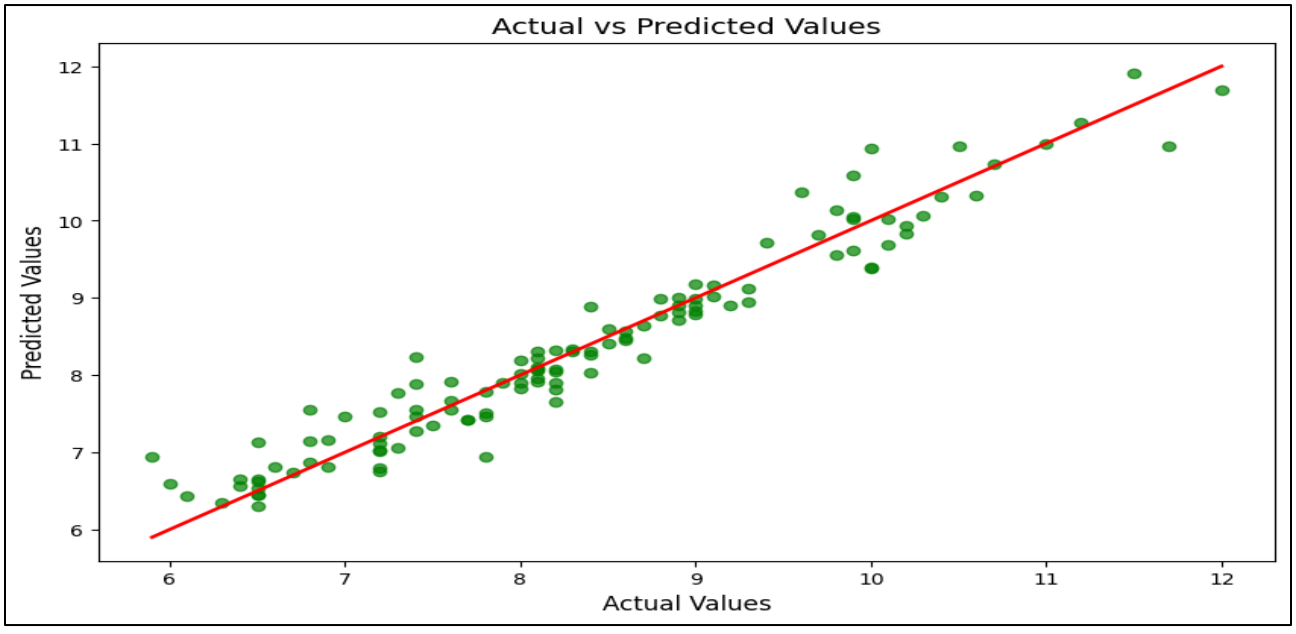
2) Hyperparameter Tuning

To further improve model accuracy, hyperparameter tuning was conducted using GridSearchCV. The best parameters were determined as follows:

- Learning Rate: 0.2
- Max Depth: 3
- Number of Estimators: 300
- Post-Tuning Results:
- MSE: 0.11
- R-squared: 0.94

The tuned model explained 94% of the variance in PM2.5 levels, reflecting a significant improvement in predictive accuracy.

```
Best Parameters: {'learning_rate': 0.2, 'max_depth': 3, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 300}
Best Cross-Validation Score (MSE): 0.11053275738633213
Test Set Mean Squared Error (MSE): 0.10791860858236887
Test Set R-squared (R2) Score: 0.9402737971001806
```



Feature Importance:

- Black Carbon absorbance emerged as the most significant predictor, contributing approximately 74% to the model's predictive power.
- Other key features included "Year" (11.6%) and "Nitrogen Dioxide Mean ppb" (7.3%), highlighting the importance of temporal factors and NO2 levels in predicting PM2.5.

Model Results for Health Data

1) Gradient Boosting Regressor: The Gradient Boosting Regressor (XGBoost) was selected as the primary predictive model due to its ability to handle non-linear relationships and achieve high accuracy, and since it performed the best among all the models we used. The model was trained on the scaled dataset, and its performance was evaluated using metrics such as Mean Squared Error (MSE) and R-squared.

Initial Results:

- Mean Squared Error (MSE): 63.14
- R-squared (R2) Score: 0.93

These results indicate that the model explains 93% of the variance in deaths due to PM2.5 (age 30+) demonstrating strong predictive performance.

2) Hyperparameter Tuning

To further improve model accuracy, hyperparameter tuning was conducted using GridSearchCV. The best parameters were determined as follows:

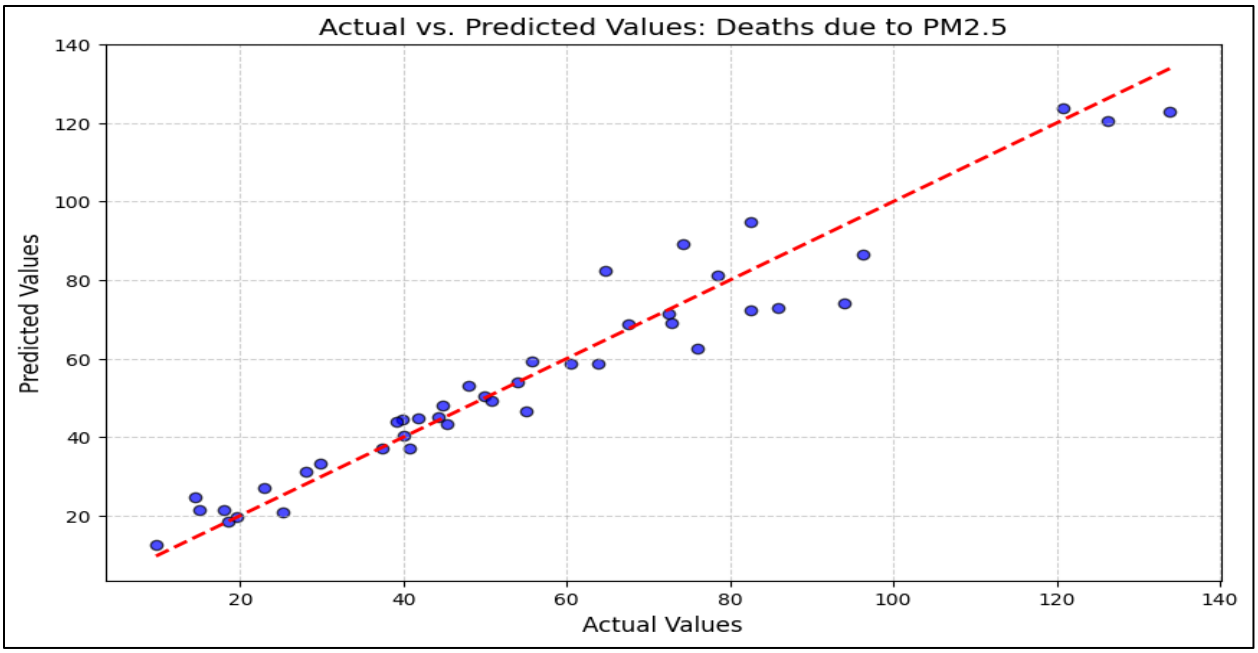
- Learning Rate: 0.2
- Max Depth: 3
- min_samples_leaf: 1
- min_samples_split: 3
- Number of Estimators: 300

Post-Tuning Results:

- MSE: 53.83
- R-squared: 0.94

The tuned model explained 94% of the variance in deaths due to PM2.5 (age 30+), reflecting a significant improvement in predictive accuracy.

```
Best Parameters: {'learning_rate': 0.2, 'max_depth': 3, 'min_samples_leaf': 1, 'min_samples_split': 3, 'n_estimators': 300}
Best Cross-Validation Score (MSE): 68.18
Test Set Mean Squared Error (MSE): 53.83
Test Set R-squared (R2) Score: 0.94
```



Feature Importance:

- Cardiovascular hospitalizations due to PM2.5 (age 40+) emerged as the most significant predictor, contributing approximately 65.8% to the model's predictive power.
- Followed by respiratory hospitalizations (20+) at 20.4%. Other hospitalization and asthma emergency visit data show moderate importance, while temporal features like StartYear and EndYear have minimal impact, and GeoRank contributes nothing.

Interpretation of Outputs

The GBR model demonstrated high accuracy in predicting PM2.5 levels, with Black Carbon absorbance identified as the most critical factor influencing air quality. The inclusion of temporal variables such as "Year" and pollutant-specific metrics like NO2 further improved the model's robustness. The GBR Model Prediction for Health Data Protection – the main key insights are Cardiovascular and respiratory hospitalizations due to PM2.5 dominant predictive power, while temporal features and GeoRank have negligible impact.

Addressing Research Questions

- The models successfully answered the research questions by identifying key pollutants (e.g., PM2.5, NO2) and their seasonal trends.
- Forecasting future air quality levels was achieved with high accuracy (R-squared =0.94), enabling proactive interventions.
- Correlation analysis linked pollutants to health outcomes, providing actionable insights into public health impacts.

Limitations and Future Work

- While the models performed well, additional data such as real-time weather conditions or traffic patterns could further enhance predictive capabilities.
- Incorporating advanced deep learning models like LSTMs could improve long-term forecasting accuracy.
- Expanding health impact analysis by integrating more granular health data would provide deeper insights into pollution-related risks.

DISCUSSION

Data-Driven Public Health Strategies: By integrating pollutant trends with health data, this study demonstrates the potential for data-driven approaches in addressing environmental health challenges. Future work could include real-time monitoring systems to identify high-pollution events and proactively mitigate their impacts (Van Donkelaar et al., 2010).

Progress in Air Quality Improvement: While overall pollutant levels have declined in NYC over the past decade, seasonal peaks during winter and hotspots in densely populated areas persist. This calls for year-round monitoring and seasonal emission control policies (Dominici et al., 2006).

Impact of PM2.5 on Vulnerable Populations: Long-term PM2.5 exposure is linked to increased hospitalization rates for asthma and other respiratory diseases, particularly among children and the elderly in underserved neighborhoods. These findings align with studies linking PM2.5 to disproportionate health burdens in urban communities (Brunekreef & Holgate, 2002).

Limitations

Data Availability:

- Health data was only available for **PM2.5**, limiting a comprehensive evaluation of other pollutants' health impacts.
- Gaps in real-time air quality data reduce the ability to assess immediate pollutant fluctuations.

Geographic Scope

- While the analysis focused on NYC, data granularity for some neighborhoods was insufficient, which could lead to the underrepresentation of certain areas.

Variable Integration:

- Socio-economic, industrial, and traffic pattern data were not incorporated, potentially leaving out important contributing factors.

Open Questions and Concerns

- How do specific interventions (e.g., traffic restrictions or industrial regulations) impact pollutant levels across different regions?
- What are the long-term health effects of sustained exposure to pollutants like PM2.5 and NO2?
- Can real-time data from sensors improve forecasting accuracy and enable more timely interventions?

Future Research Directions

Industrial and traffic pattern data were not incorporated, potentially leaving out important contributing factors. By addressing these limitations and exploring new research directions, future studies can provide deeper insights into air quality dynamics and their implications for public health and urban planning.

CONCLUSION

This project has provided a comprehensive analysis of air quality patterns in New York City, focusing on pollutant distribution, seasonal trends, and health impacts. By leveraging datasets from data.gov and nyc.gov, the study identified PM2.5 and NO2 as key pollutants influencing air quality and demonstrated significant seasonal variations. The analysis also revealed strong correlations between pollutant levels and health outcomes, such as increased asthma rates during periods of high PM2.5 concentrations. These

findings underscore the importance of targeted interventions during high-risk periods to mitigate adverse health effects.

This study highlights significant temporal and spatial disparities in air pollutant concentrations (PM2.5, NO2, SO2, BC) across New York City from 2009 to 2022. The correlation between PM2.5 exposure and adverse health outcomes reinforces the need for targeted interventions to reduce pollutant levels and protect public health. Vulnerable communities, particularly in areas like the upper east of NYC, experience disproportionate exposure to harmful pollutants.

The study did not fully explore the impact of external factors like traffic or industrial emissions on pollutant levels. Long-term health impacts of sustained exposure to pollutants remain unaddressed due to data limitations. Future research could integrate real-time traffic and industrial data to refine predictive models. Incorporating more granular health data would enhance the understanding of pollution-related health risks. Advanced modeling techniques, such as deep learning approaches, could further improve forecasting accuracy. This study provides actionable insights for policymakers and public health officials, emphasizing the need for proactive air quality management to safeguard public health.

REFERENCES

1) En Xin Neo, Khairunnisa Hasikin, Mohd Noriznan Mokhtar, Khin Wee Lai, Azizan, M. M., Sarah Abdul Razak, & Hizaddin, H. F. (2022). Towards Integrated Air Pollution Monitoring and Health Impact Assessment Using Federated Learning: A Systematic Review. Frontiers in Public Health, 10. https://doi.org/10.3389/fpubh.2022.851553
2) Sokhi, R., Moussiopoulos, N., Baklanov, A., Bartzis, J., Coll, I., Finardi, S., Friedrich, R., Geels, C., Grönholm, T., Halenka, T., Ketzel, M., Maragkidou, A., Matthias, V., Moldanova, J., Ntziachristos, L., Schäfer, K., Suppan, P., Tsegas, G., Carmichael, G., & Franco, V. (2022). atmospheric chemist, awarded the Nobel Prize in Chemistry 1995; Mario Molina (1943-2020), atmospheric chemist, awarded the No-bel Prize in. Atmos. Chem. Phys, 22(7), 4615–4703. https://doi.org/10.5194/acp-22-4615-2022
3) The Blueprint for a Sustainable Urban Air Quality Strategy. (2023). Clarity.io. https://www.clarity.io/blog/the-blueprint-for-a-sustainable-urban-air-quality-strategy

4) Kelly, F. J., & Fussell, J. C. (2015). Air Pollution and Public hHealth Emerging Hazards and Improved Understanding of Risk. Environmental Geochemistry and Health, 37(4), 631–649. https://doi.org/10.1007/s10653-015-9720-1

5) Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air Pollution Forecasts: An Overview. International Journal of Environmental Research and Public Health, 15(4). https://doi.org/10.3390/ijerph15040780

6) Ruzmyn Vilcassim, & Thurston, G. D. (2023). Gaps and future directions in research on health effects of air pollution. 104668–104668. https://doi.org/10.1016/j.ebiom.2023.104668

7) NYC Environmental Health. (n.d.). Nyccas.cityofnewyork.us. https://nyccas.cityofnewyork.us/nyccas2022/report/3