Marzieh Rasti

Final Report
Analyzing the correlation between air quality(PM2.5 and NO2) data and public health outcomes(Asthma) in New York City

Problem Statement

The core issue addressed in this study is the relationship between air quality and respiratory problems, specifically asthma-related hospital visits, in New York City. This project aims to determine whether atmospheric pollutants significantly impact the frequency of asthma-related hospital visits among NYC's residents.

New York City, being a densely populated urban area, faces notable challenges related to air pollution. This study is crucial given the increasing concerns about environmental health, urban living, and sustainable city planning. The findings will have significant implications for policy decisions, public health initiatives, and raising individual awareness about the health impacts of air quality.

This data analysis and story-telling report is organized around the following questions of interest:

- How are NO2 and PM2.5 levels correlated with the number of hospital visits due to asthma?
- Are there specific geographic areas where the correlation between air pollution (NO2 and PM2.5) and asthma hospital visits is stronger?
- How have NO2 and PM2.5 levels changed over the years, and how does this trend correlate with the number of asthma-related hospital visits?
- Can we develop a predictive model to forecast the number of asthma-related hospital visits based on NO2 and PM2.5 levels?

Data wrangling

The data utilized in this study is sourced from the NYC Environment & Health Data Portal (NYC.gov), which provides comprehensive information on environmental pollutants and their health impacts. The dataset was downloaded from the NYC.gov website.

The raw dataset from NYC Environment & Health contains 18 columns and 2,037,616 rows. It required significant size reduction.Based on the Geography column in the CSV files that include all neighborhoods, I have selected New York, Manhattan, and Staten Island. Therefore, I dropped the other neighborhoods.

Additionally, I dropped some columns in the asthma dataset that were based on an unclear counting method, such as "Estimated annual rate per 10,000." Similarly, I dropped columns in the PM2.5 dataset that were based on an unclear measuring method, such as "10th percentile mcg/m3" and "90th percentile mcg/m3."

The dataset does not have any null values. However, the data type of some columns needed to be changed and remove some prefixes. Some datasets had different time periods. For example, air pollutant data included yearly and seasonal time periods, but the asthma data only included yearly time periods. Therefore, I decided to only keep yearly data and drop the seasonal records.

Some geographical columns did not contain any useful information and were actually repetitive, so I dropped unnecessary columns.

The final shape of my dataset is 347 rows and 8 columns.

Exploratory Data Analysis

Initially, I analyzed datasets including asthma emergency visits for adults, asthma emergency visits for children aged 4 and under, and asthma emergency visits for children aged 5-17. Finally, I chose the asthma emergency adults dataset for analysis.

In this section, I will explore the correlation between air quality and the number of hospital visits due to asthma in New York City.

Pairwise relationships in the dataset

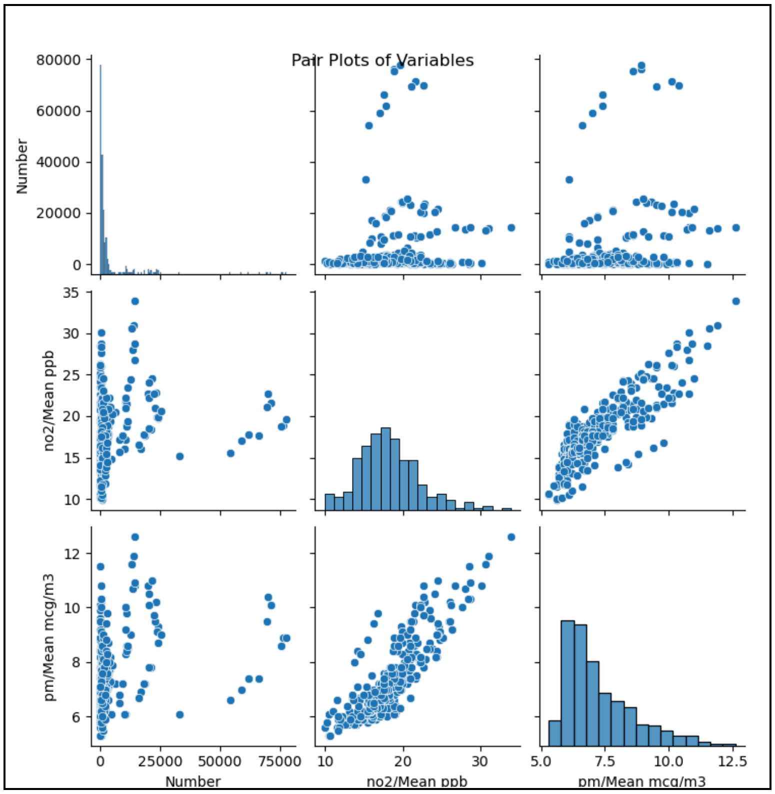

Figure 1: Pairwise relationships in the dataset

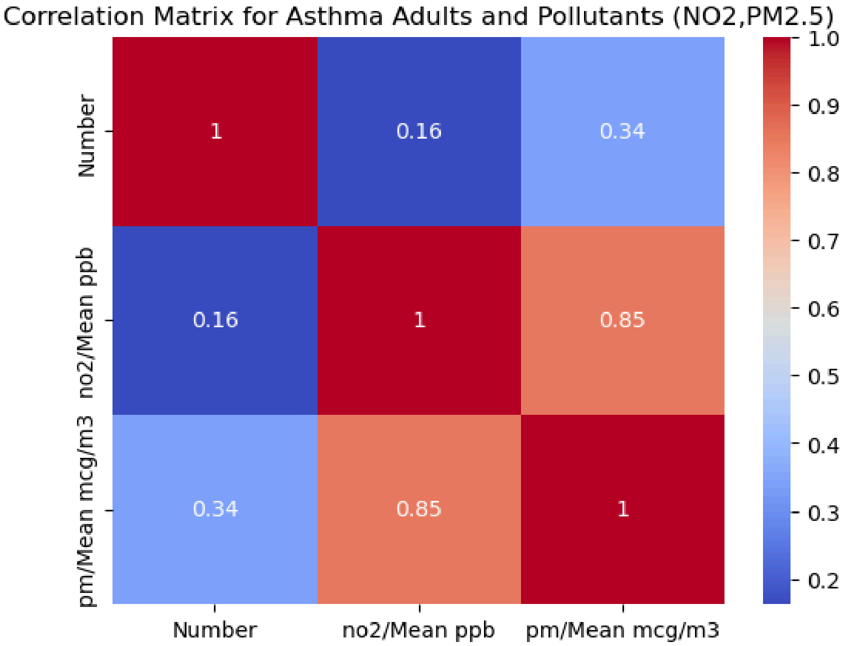

Figure 2: Correlation between different features in the dataset

Based on the above plots(figure1&2), we can infer about the relationships between the variables:

Number vs. NO2/Mean (ppb):

There is a positive correlation between the number of hospital visits and NO2 levels. As the concentration of NO2 increases, the number of hospital visits also tends to increase. However, the spread suggests some variability in the relationship.

Number vs. PM/Mean (mcg/m3):

There is a positive correlation between the number of hospital visits and PM levels. Similar to NO2, higher concentrations of particulate matter are associated with more hospital visits.

NO2/Mean (ppb) vs. PM/Mean (mcg/m3):

- There is a strong positive correlation between NO2 and PM levels. This indicates that higher concentrations of NO2 are often associated with higher concentrations of particulate matter, suggesting that these pollutants may share common sources or conditions that lead to their increase.

Geographic Analysis of Pollutants Levels

In this section, I wanted to analyze the distribution of the average levels of pollutants across the different boroughs of New York City, highlighting geographic variations in air pollutant concentration.

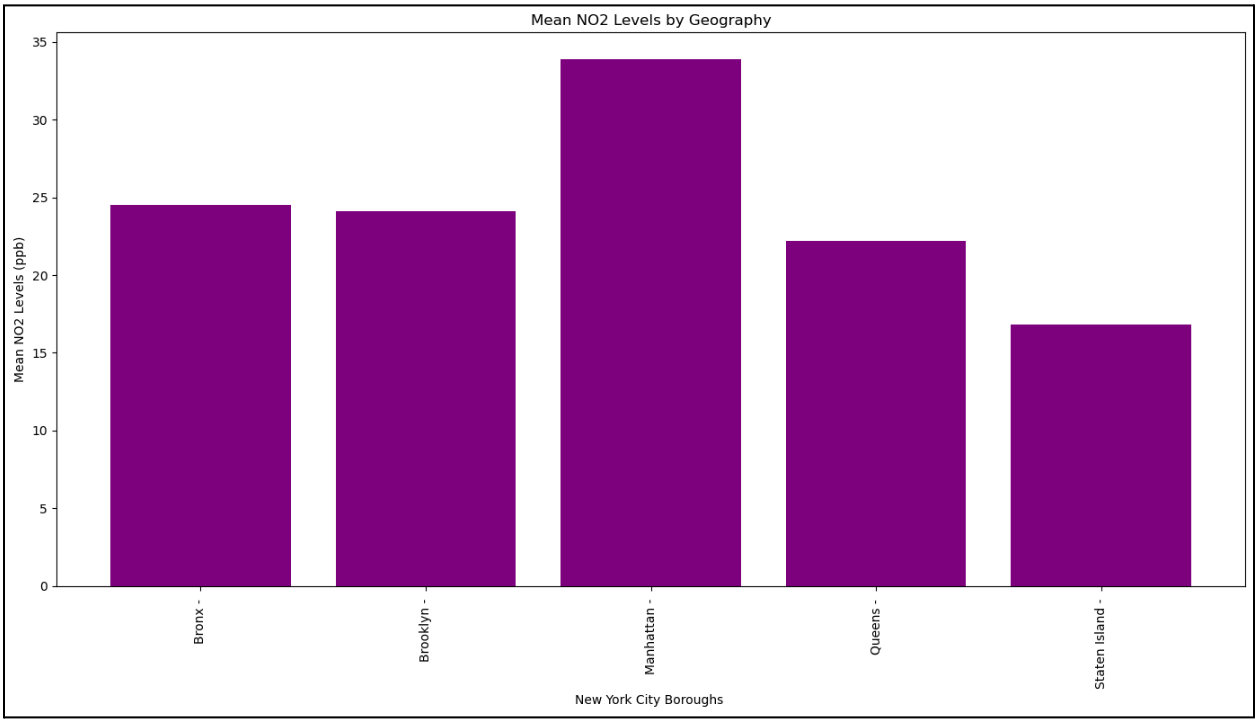

Figure 3:Geographic Distribution of Mean NO2 Levels in NYC Boroughs

- The high NO2 levels in Manhattan can be attributed to dense traffic, high population density, and significant commercial activities.
- Bronx and Brooklyn also exhibit elevated NO2 levels, possibly due to similar urban density and traffic patterns.
- Lower NO2 levels in Queens and Staten Island suggest these areas might have less traffic congestion and more open spaces, leading to better air quality

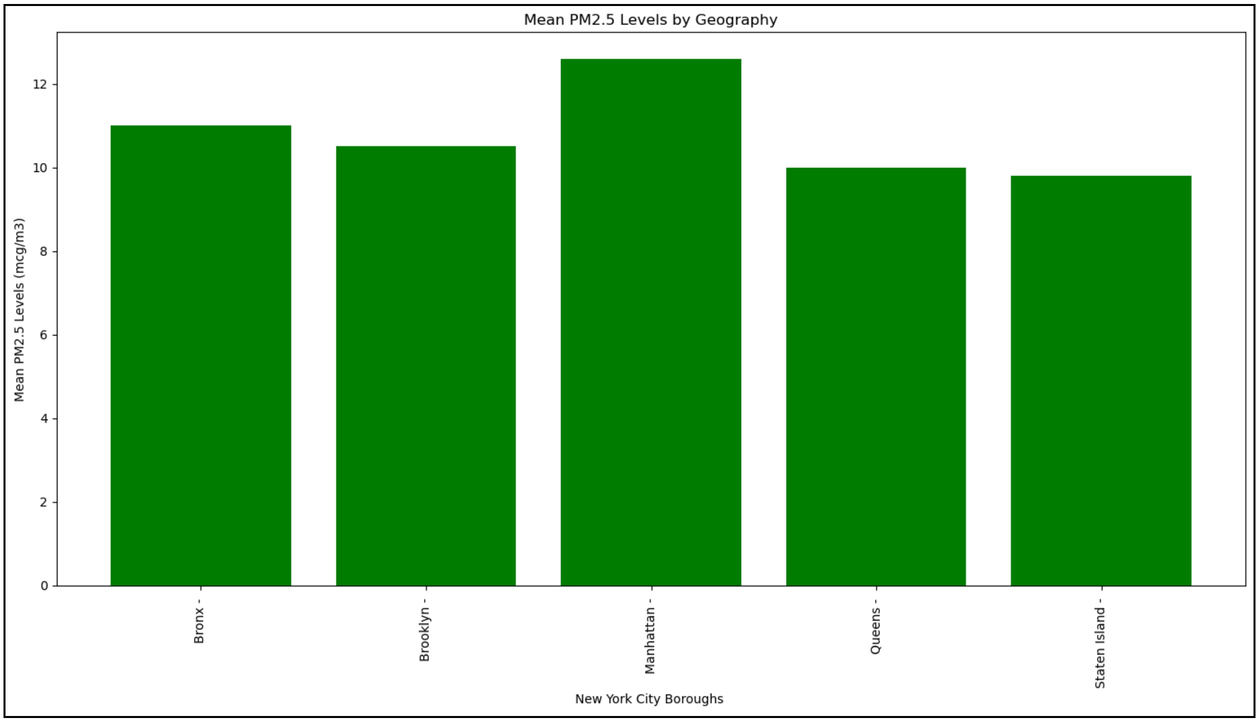

Figure 4:Geographic Distribution of Mean PM 2.5 Levels in NYC Boroughs

- The elevated PM2.5 levels in Manhattan could be due to high traffic emissions, construction activities, and other urban sources.
- Similar PM2.5 levels in Bronx, Brooklyn, and Queens might result from comparable urban activities and traffic patterns.
- The lower PM2.5 levels in Staten Island may reflect its lower population density, lesser industrial activities, and greater green spaces.

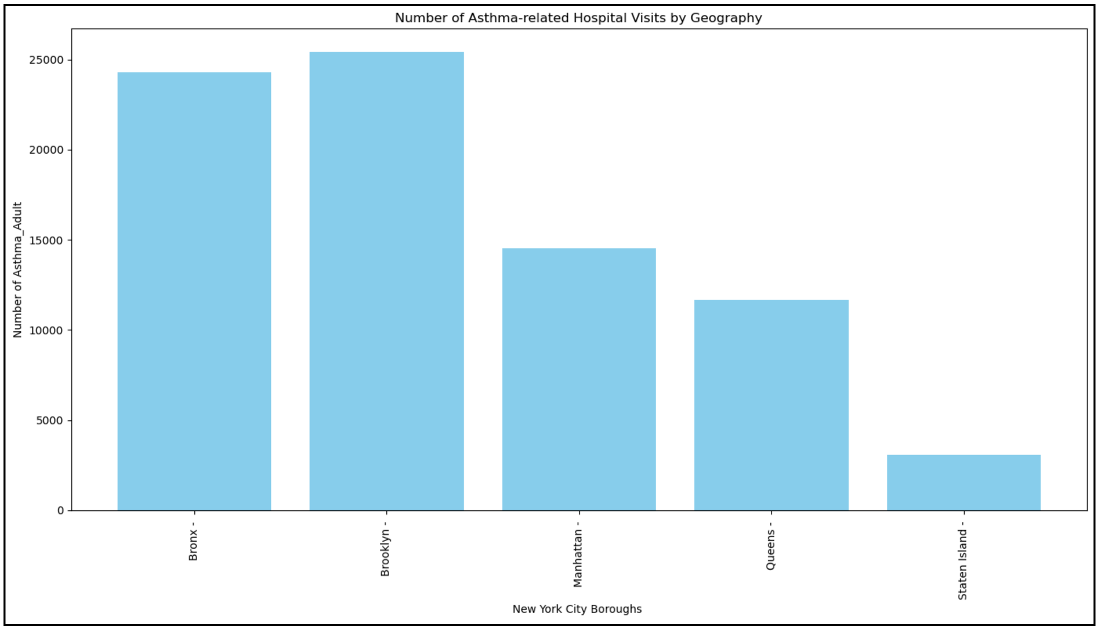

Figure 5: Geographic Distribution of the Number of Adults with Asthma Visits in NYC Boroughs

- The high number of asthma-related hospital visits in Bronx and Brooklyn suggests a significant impact of air pollution on respiratory health in these areas.
- The moderate number of visits in Manhattan could indicate better healthcare access or different demographic factors influencing hospital visit rates.
- Lower hospital visit numbers in Queens and Staten Island might be due to their better air quality and possibly lower population density.

Modeling

This is a regression problem, in supervised learning. Here we have used the following regression models:

1. Multiple Linear Regression
2. Ridge Regression (Regularized Linear Regression)
3. Gradient Boosting Machines
4. Random Forest
5. Support Vector Regression Model(SVR)

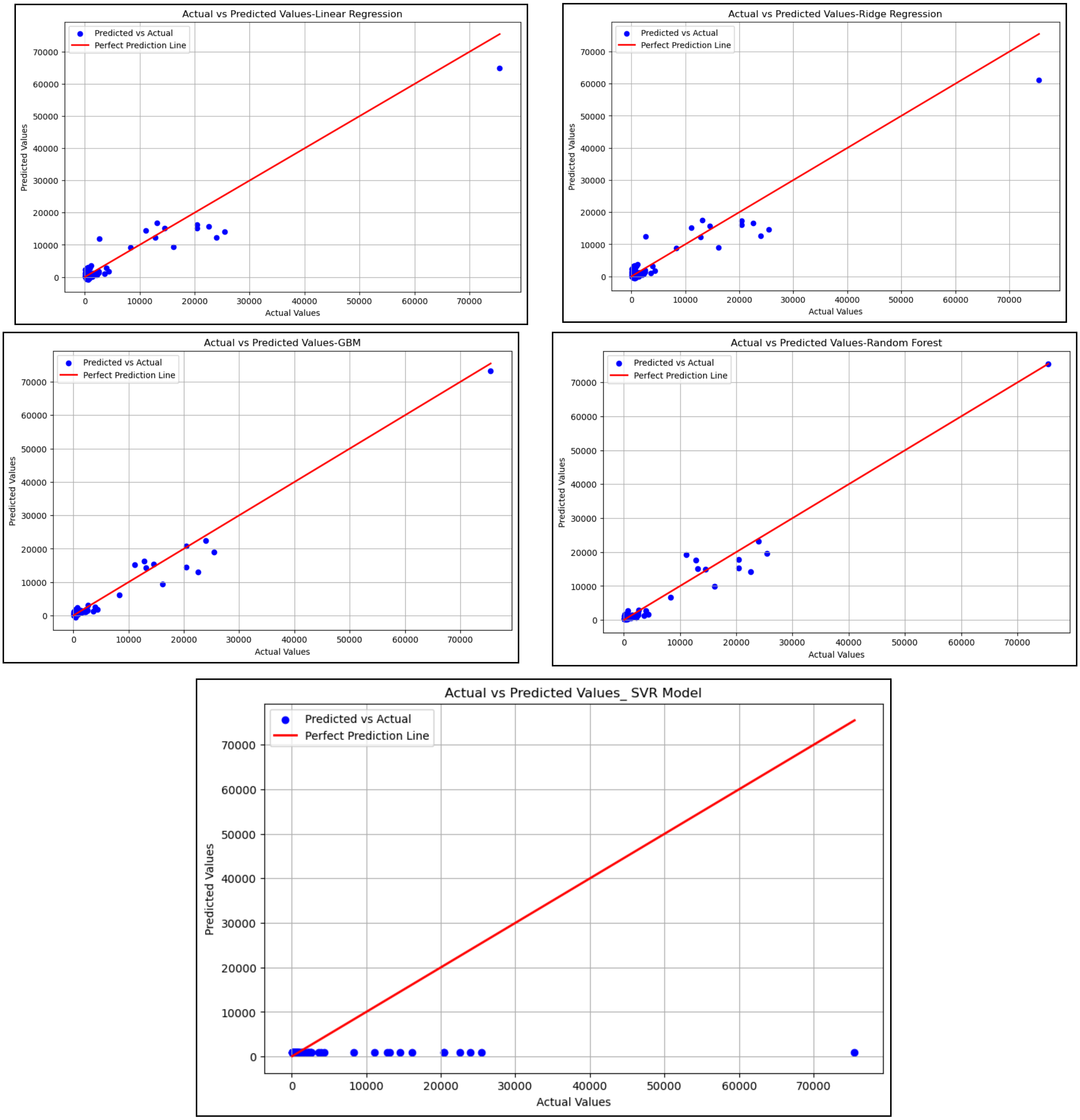

Figure 6: Actual vs Predicted Values for Adult Asthma Visits using different Models

Based on the above figure (Figure 6), all models except SVR show strong predictive performance, with predicted values closely aligning with actual values. The SVR model shows more dispersion and less accuracy, especially for higher actual visit values.

Comparison and Model Selection

I applied different ML models above and evaluated their performances using cross-validation for both the training and test data.

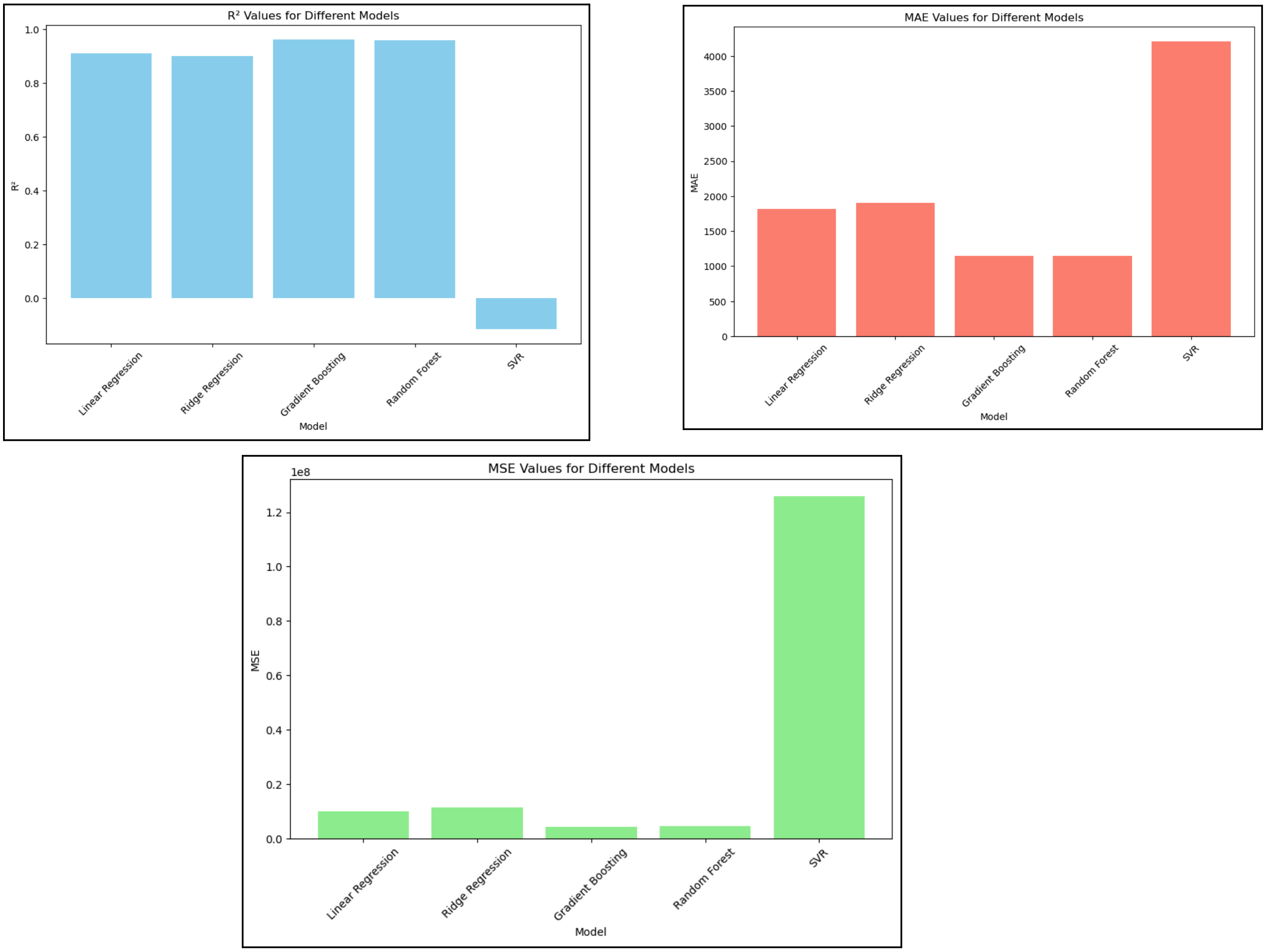

Figure 7: Model Performance Evaluation Metrics for Predicting Adult Asthma Visits

Based on the above metrics results(figure 7), **Random Forest** and **Gradian Boosting** are the best models among the ones compared. They have the highest R². However, the high values of MSE and MAE suggest that the actual predictions may still be quite far from the actual values. To make the process more accurate I want to use feature selection in the next step.

Feature importances for Selected Models

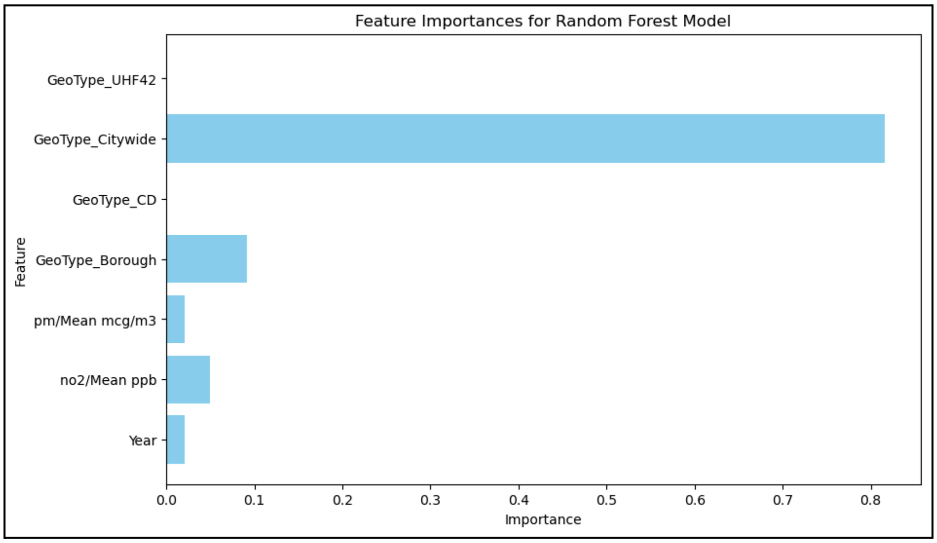

Figure 8:Feature importances for Random Forest Model

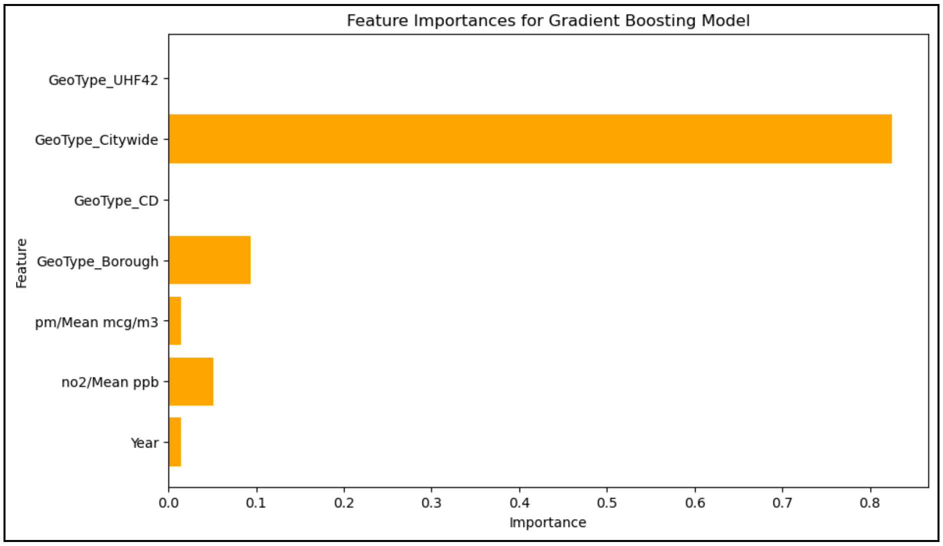

Figure 9:Feature importances for Gradient Boosting Model

Based on the above plot, the Random Forest model and the Gradient Boosting model exhibit the same feature importance.Therefore features'GeoType_UHF42' and 'GeoType_CD' are not too important for the Model, so we have to drop these features and use most relevant features to predict our target variable.So, we can use reduce dataset taht we created for the random forest model and retrain the model with reduced features.

Applying Grid search CV for hyperparameter Tuning

I applied hyperparameter tuning for both models selected. Results of the performance metrics for two machine learning models (table 1)shows both models have high Best Score and Test Score values, indicating strong performance in predicting asthma-related hospital visits. The Random Forest model slightly outperforms the Gradient Boosting model, as indicated by the marginally higher Best Score and Test Score values. The small difference between Best Score and Test Score for both models suggests that they generalize well and are not overfitting to the training data.

Table 1: Results of the performance metrics for two machine learning models

Model	Best Score	Test Score	R2
Random Forest	0.945868	0.954561	0.954561
Gradient Boosting	0.950065	0.944047	0.944047

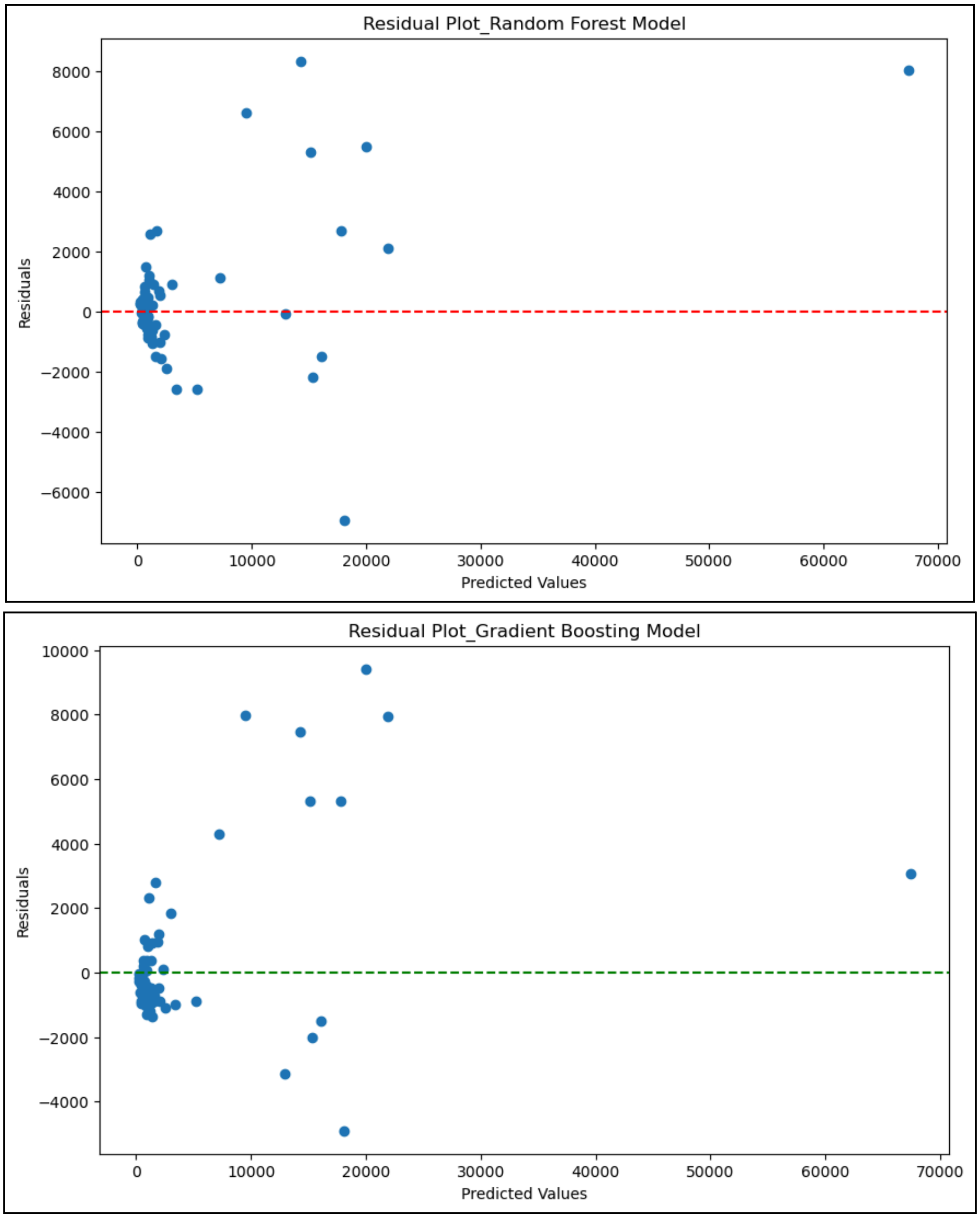

Figure 10:Residual Plot for Machine Learning Models

The residuals (the differences between the actual and predicted values) are centered around zero, which is a good sign. This indicates that the model is generally unbiased. However, there is a spread of residuals, with some large positive and negative residuals. This could suggest that the model might be underestimating or overestimating some points.

The residuals seem to be fairly evenly distributed, with no clear pattern or trend. This is desirable, as it indicates that the model's errors are random rather than systematic. There are a few points with very high positive and negative residuals, which can be considered outliers. These are data points where the model's predictions are significantly different from the actual values. These outliers can impact the model's performance metrics and may warrant further investigation to understand their nature (e.g., data entry errors, special cases, etc.).

Future Overseeing

- Further explore and engineer features that could improve model performance, such as incorporating additional air quality metrics or socioeconomic factors.
- Collect more data, especially from air pollutants, different time periods or additional geographical areas, to improve model generalization.
- Explore advanced algorithms such as XGBoost, LightGBM, or neural networks to potentially capture complex relationships within the data.
- Investigating the temporal dynamics of air quality and asthma exacerbations by incorporating time-series analysis could reveal seasonal or temporal trends that are not captured by static models. This could help in understanding how different times of the year or specific weather conditions affect asthma incidence.
- Conducting more detailed geospatial analysis using advanced GIS tools could help in identifying specific areas within the city that are more prone to poor air quality and higher asthma rates. This could inform targeted interventions and policy decisions.