Analyzing the Correlation Between Air Quality (PM2.5 and NO2) Data and Public Health Outcomes (Asthma) in New York City

By: Marzieh Rasti

Data Science Capstone Project

Springboard

The Problem

Relationship between air quality and respiratory problems, specifically asthma-related hospital visits, in New York City.

Research Questions:

- **How are NO2 and PM2.5 levels correlated with the number of hospital visits due to asthma?**
- **How do pollutants levels vary across different GeoTypes in New York City?**
- **Can we develop a predictive model to forecast the number of asthma-related hospital visits based on NO2 and PM2.5 levels?**

Stakeholders

Data

TimePeriod	GeoType	GeoID	GeoRank	Geography	Estimated annual rate per 10,000	Number
2020	CD	101	6	Financial District (CD1)	16.6*	7*
2020	CD	102	6	Greenwich Village and Soho (CD2)	3.2*	1*
2020	CD	103	6	Lower East Side and Chinatown (CD3)	73.0	47
2020	CD	104	6	Clinton and Chelsea (CD4)		
2020	CD	105	6	Midtown (CD5)		
2020	CD	106	6	Stuyvesant Town and Turtle Bay (C		
2020	CD	107	6	Upper West Side (CD7)		
2020	CD	108	6	Upper East Side (CD8)		
2020	CD	109	6	Morningside Heights and Hamilton		
2020	CD	110	6	Central Harlem (CD10)		
2020	CD	111	6	East Harlem (CD11)		
2020	CD	112	6	Washington Heights and Inwood (C		
2020	CD	201	6	Mott Haven and Melrose (CD1)		
2020	CD	202	6	Hunts Point and Longwood (CD2)		
2020	CD	203	6	Morrisania and Crotona (CD3)		
2020	CD	204	6	Highbridge and Concourse (CD4)		

TimePeriod	GeoType	GeoID	GeoRank	Geography	10th percentile mcg/m3	90th percentile mcg/m3	Mean mcg/m3
Annual Average 2022	CD	101	6	Financial District (CD1)	6.7	7.8	7.2
Annual Average 2022	CD	102	6	Greenwich Village and Soho (CD2)	7.6	9.3	8.4
Annual Average 2022	CD	103	6	Lower East Side and Chinatown (CD3)	6.1	8.6	7.2
Annual Average 2022	CD	104	6	Clinton and Chelsea (CD4)	6.7	9.0	7.8
Annual Average 2022	CD	105	6	Midtown (CD5)	8.2	9.7	9.1
Annual Average 2022	CD	106	6	Stuyvesant Town and Turtle Bay (CD6)	6.7	8.5	7.5
Annual Average 2022	CD	107	6	Upper West Side (CD7)	5.9	6.3	6.1
Annual Average 2022	CD	108	6	Upper East Side (CD8)	6.1	6.8	6.4
Annual Average 2022	CD	109	6	Morningside Heights and Hamilton Heights (CD9)	6.1	6.4	6.2
Annual Average 2022	CD	110	6	Central Harlem (CD10)	6.1	6.4	6.2
Annual Average 2022	CD	111	6	East Harlem (CD11)	6.0	6.3	6.2
Annual Average 2022	CD	112	6	Washington Heights and Inwood (CD12)	6.1	6.6	6.3
Annual Average 2022	CD	201	6	Mott Haven and Melrose (CD1)	5.8	6.5	6.1
Annual Average 2022	CD	202	6	Hunts Point and Longwood (CD2)	6.3	6.8	6.5
Annual Average 2022	CD	203	6	Morrisania and Crotona (CD3)	6.2	6.4	6.3
Annual Average 2022	CD	204	6	Highbridge and Concourse (CD4)	5.8	6.3	6.0

Data Source: https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/air-quality/?id=2023#display=summary

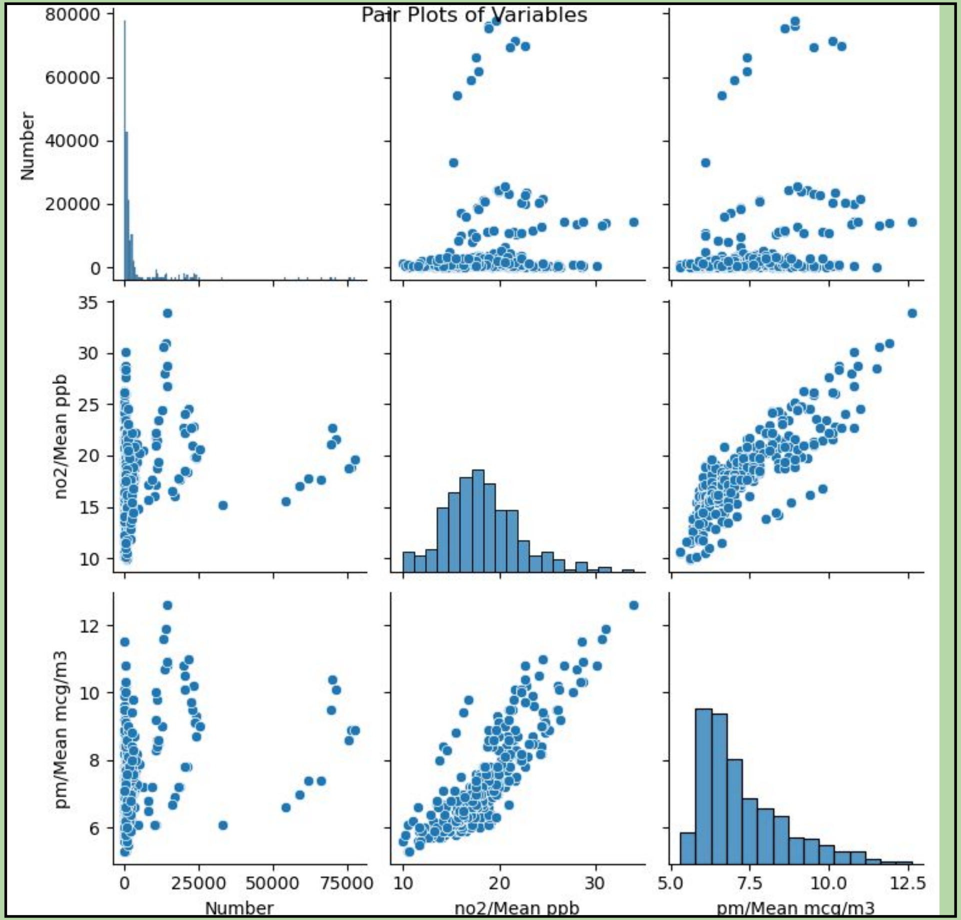
Data Collection Areas(New York City Boroughs)

Data Wrangling

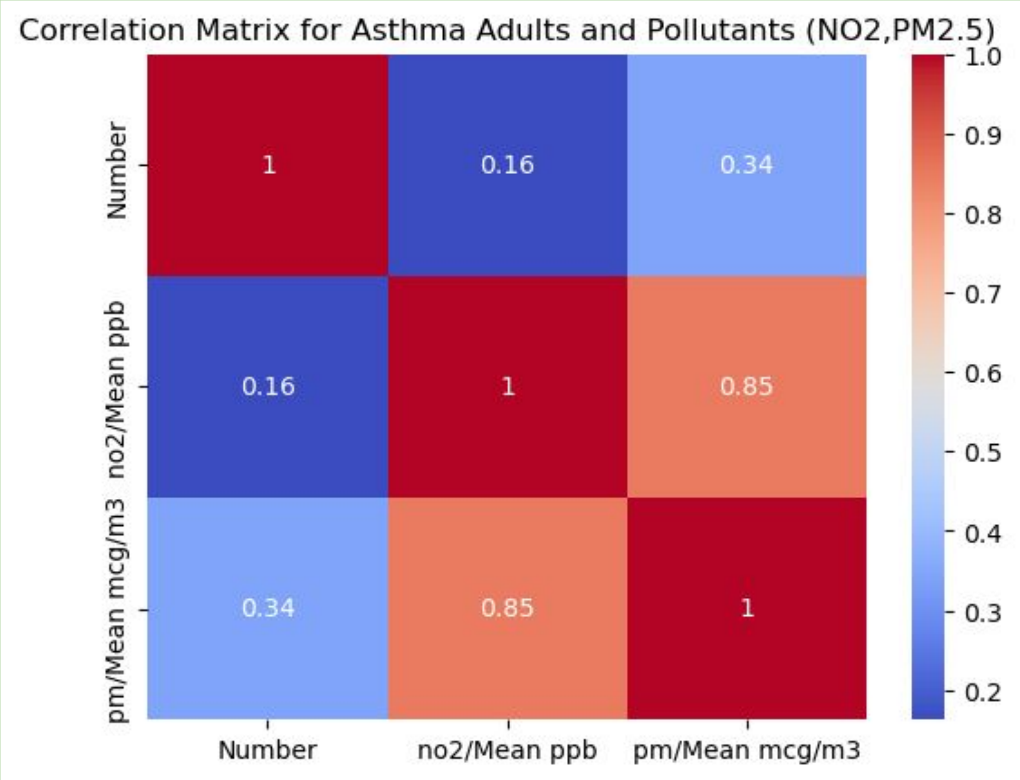
Original dataset had 2,037,616 rows and 18 columns

- Drop columns that were based on an unclear measuring method
- To ensure consistency, the yearly records chosen and discarded the seasonal data
- The dataset does not have any null values
- Drop some geographical columns did not contain any useful information
- Target Value: Number of Adult Asthma

Exploratory Data Analysis

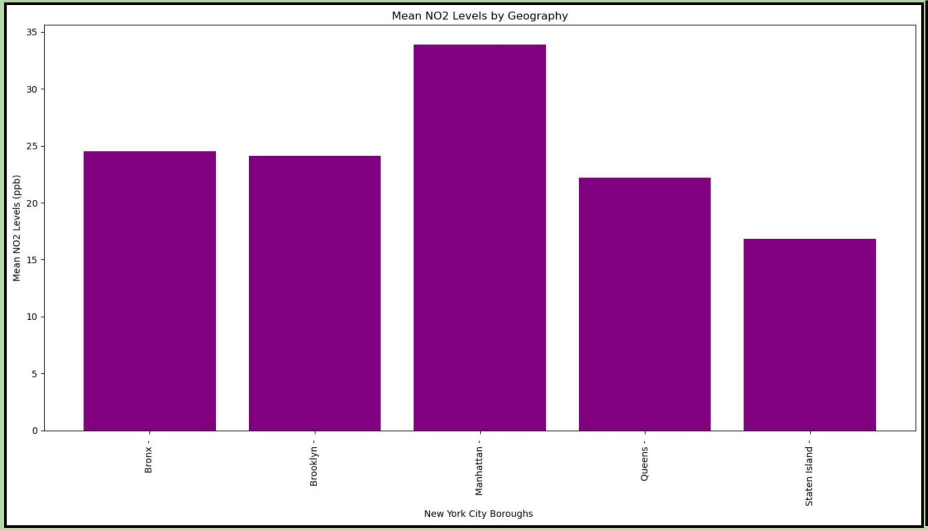


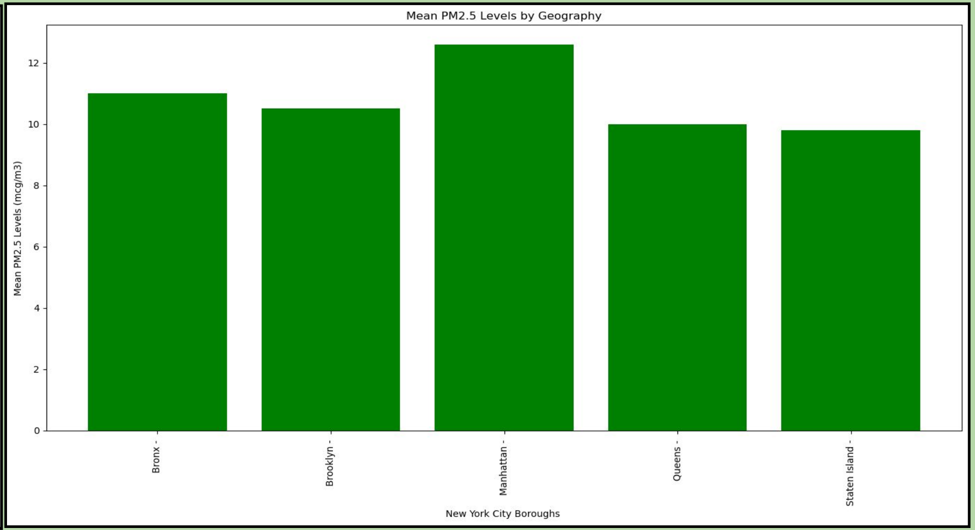
Pairwise relationships in the dataset



Correlation between different features in the dataset

Geographical Distribution of Pollutants Levels&Asthma





Geographical Distribution of Mean NO2 Levels in NYC Boroughs

Geographical Distribution of Mean PM 2.5 Levels in NYC Boroughs

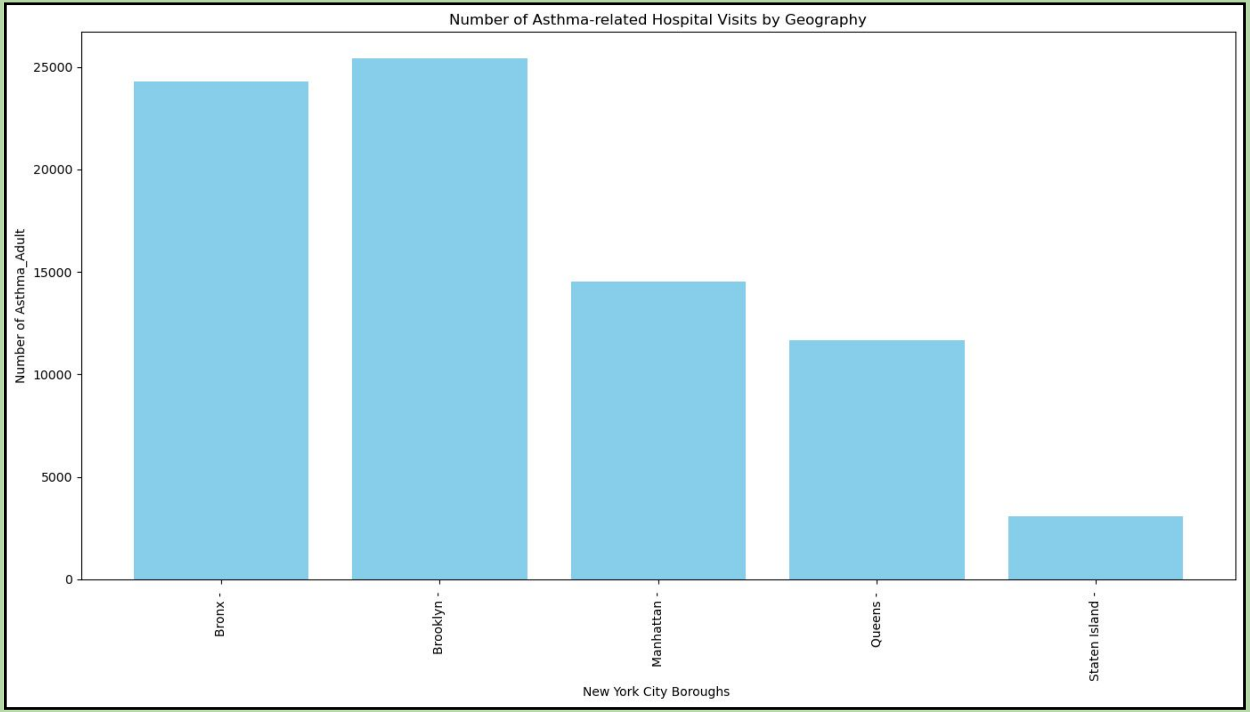


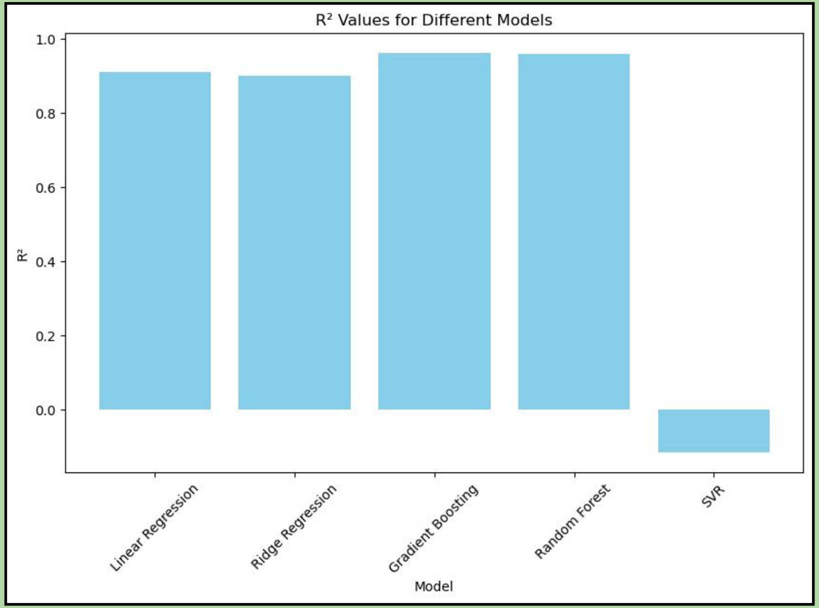
Figure 5: Geographical Distribution of the Number of Asthma-related Hospital Visits in NYC Boroughs

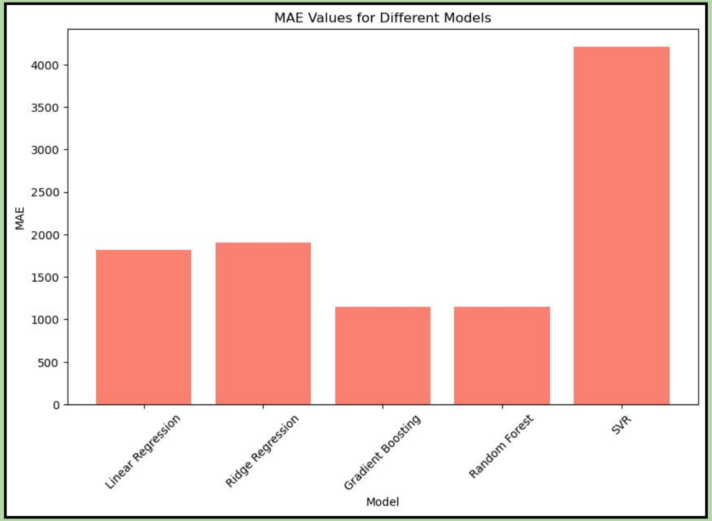
Machine Learning Modeling

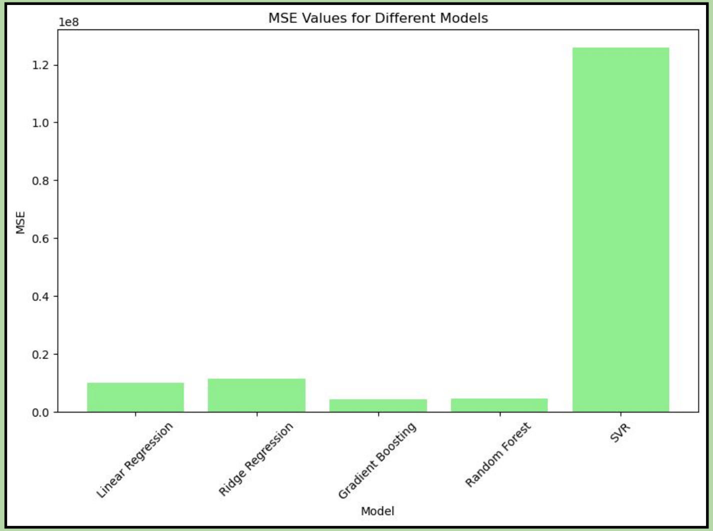
Type: Supervised Learning

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

Comparison Models





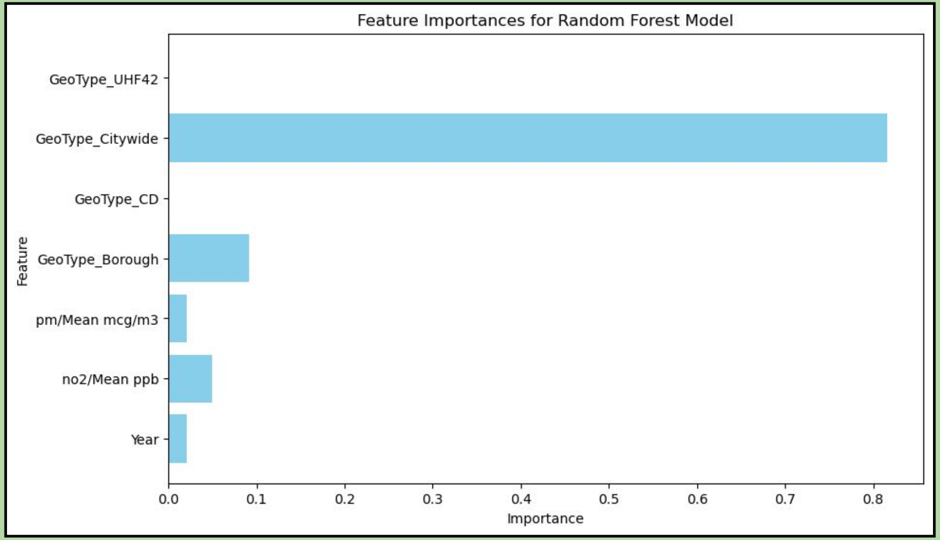


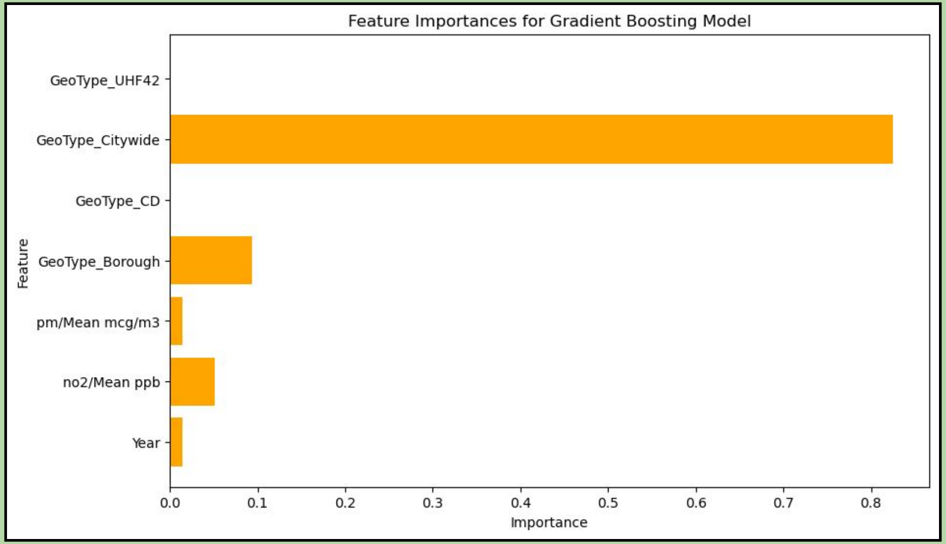
Comparison and Model Selection

Model	R2	MAE	MSE
Linear Regression	0.910860	1820.996477	1.005953e+07
Ridge Regression	0.899823	1906.950252	1.130502e+07
Gradient Boosting	0.963038	1142.768278	4.171197e+06
Random Forest	0.960578	1145.995271	4.448820e+06
SVR	-0.115649	4208.299780	1.259013e+08

SVR model is the **worst**, **Random Forest** and **Gradient Boosting** are the **best** models

Feature importances for Selected Models



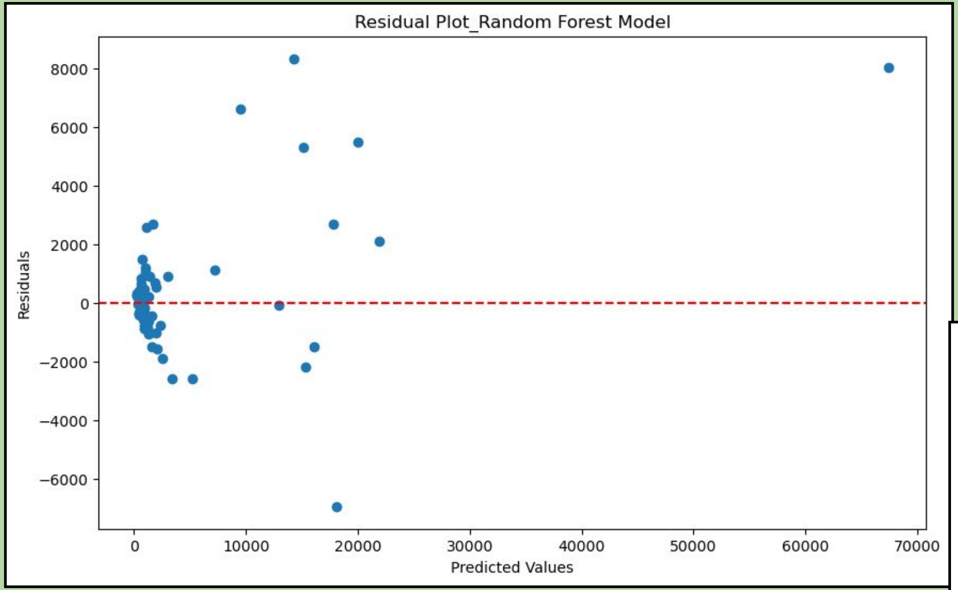


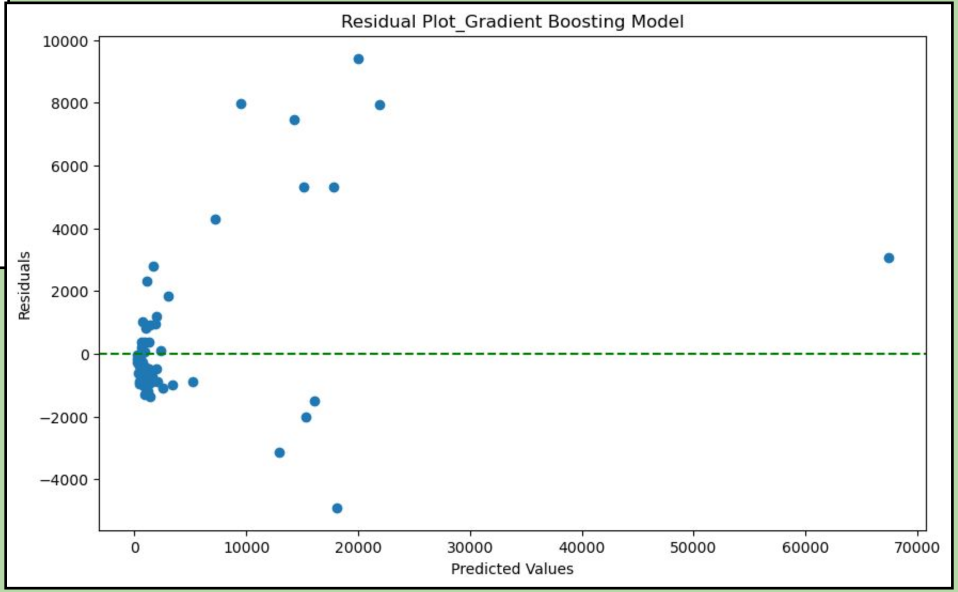
Applying Grid search CV for hyperparameter Tuning

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

Differences between the actual and predicted values





Conclusion

In this modeling project, multiple machine learning algorithms were applied to predict asthma-related hospital visits in New York City based on various air quality and geographical features. The models evaluated include Linear Regression, Ridge Regression, Gradient Boosting, Random Forest, and Support Vector Regression (SVR).

The performance metrics (R2, MAE, and MSE) showed that both Gradient Boosting and Random Forest performed exceptionally well, with R2 values of 0.962973 and 0.960578, respectively. These models also demonstrated lower MAE and MSE values compared to the other models, indicating better prediction accuracy and precision.

- The feature importance plots revealed that geographical features, particularly GeoType_Citywide, had the most significant impact on the predictions, followed by air quality metrics such as pm/Mean mcg/m3 and no2/Mean ppb.
- Hyperparameter tuning using Grid Search CV was performed to optimize the models. For Random Forest, the best parameters included max_depth: 10, max_features: 'sqrt', and n_estimators: 200. For Gradient Boosting, the optimal parameters included learning_rate: 0.1, max_depth: 5, and n_estimators: 100.
- Residual plots indicated that while the models performed well, there are still some outliers.

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.

Thank You!