Asthma and Pollution: Exploratory Analysis of Predictive Machine Learning Models

Ainsley Atherton, Nikki La, Jordan Ledbetter, Palaniappan Vijay Sithambaram

DAT 490 - Data Science Capstone Project

Arizona State University

---

1 Abstract

Asthma is a prevalent chronic respiratory condition that affects millions worldwide, with exacerbations often linked to environmental factors, particularly air pollution. Prior studies have established a significant correlation between air quality and increased asthma cases, especially in large urban areas. This study built upon existing research by employing machine learning techniques to deepen our understanding of how specific environmental pollutants influence asthma rates. We utilized comprehensive datasets from the Centers of Disease Control and Prevention (CDC) and the Environmental Protection Agency (EPA), to apply machine learning models, including Random Forest, and Seasonal Auto Regressive Integrated Moving Average with exogenous factors (SARIMAX). These models analyzed asthma-related health outcomes such as hospitalizations and emergency room visits, providing critical insights into the public health implications of air quality. By detecting patterns within the data, our machine learning models elucidated the connections between air quality and asthma outcomes. Our analysis confirmed significant correlations between hospitalizations and emergency room visits due to asthma and pollutants such as nitrogen dioxide (NO_2), particulate matter ($PM_{2.5}$, PM_{10}), and ozone (O_3). These findings emphasize the critical role of machine learning in identifying key environmental triggers of asthma and offer potential pathways for public health interventions and policy modifications aimed at mitigating asthma exacerbations through improved air quality standards.

2 Introduction

In recent years, the detrimental impacts of environmental degradation on human health have become an increasingly important area of scientific inquiry. Particularly, air quality has been identified as a pivotal factor affecting the prevalence and exacerbation of respiratory diseases such as asthma, leading to increased hospitalizations and emergency room visits. Studies indicate that air pollution contributes to about 334 million asthma cases worldwide, significantly exacerbating the disease's prevalence and severity. Ongoing urbanization and industrial activities continue to compromise air quality, affecting millions of people's health and well-being. This paper explores the intricate relationship between air pollution and its direct impacts on public health, particularly focusing on respiratory diseases such as asthma. Our research focuses intensely on the effects of air pollutants on individuals with asthma and individuals at risk of developing asthma, including both adults and children, who represent a particularly vulnerable segment of the population.

2.1 Current State of Air Quality and Public Health

The EPA's AirNow showcased the annual number of days when air pollution levels were classified as 'Unhealthy for Sensitive Groups' or worse, according to the Air Quality Index for a combination of ozone and $PM_{2.5}$. Figure 1 illustrates a notable decline in unhealthy air quality days across thirty-five major U.S. cities over two decades, indicating progress in air quality management.

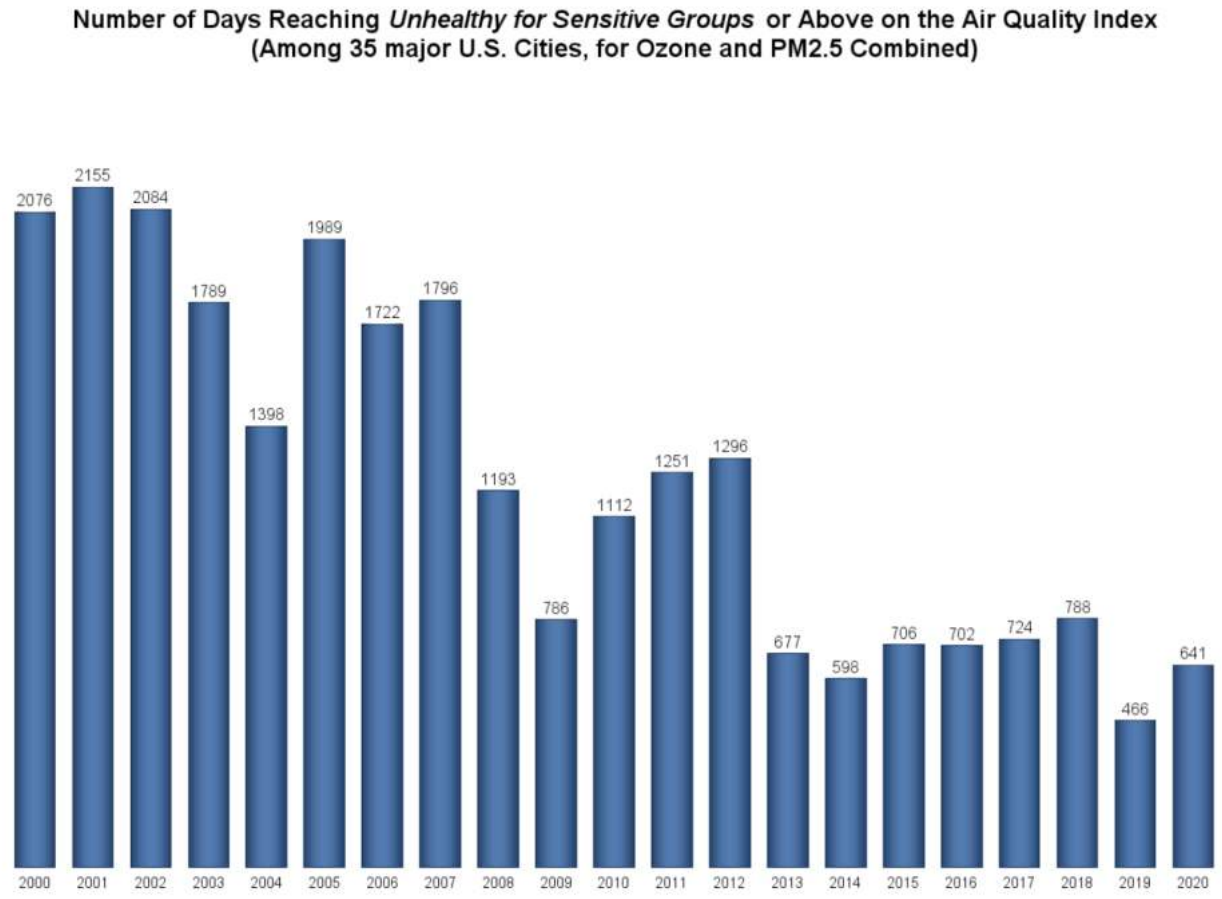

Figure 1: Trends in AQI Days for Sensitive Groups, 2000-2020 [1]

Despite decades of improvement under initiatives like the Clean Air Act, the "State of the Air" 2024 report by the American Lung Association reveals that approximately 39% of Americans, amounting to approximately 131.2 million people, live in areas with failing grades for ozone or particle pollution levels. This marks an increase of 11.7 million people compared to the previous year, emphasizing a disturbing rise in the number of individuals breathing unhealthy air on a daily basis. This deterioration in air quality is attributed to a mix of extreme weather conditions such as heatwaves, droughts, and wildfires, particularly affecting the western United States. Furthermore, the implementation of a new, more stringent EPA standard for measuring year-round fine particle pollution has revealed that more people are affected than previously recognized under older, less strict standards. The report also signifies a racial disparity in air quality impacts: while people of color constitute 41.6% of the U.S. population, they represent 56% of those living in counties with at least one failing air quality grade [19]. Pollution levels in many areas of the United States still exceed national air quality standards for common pollutants such as ozone and particulate matter, which are capable of traveling long distances and affecting air quality across state lines. These pollutants are linked

to severe health risks, including increased hospital admissions and emergency room visits for asthma attacks and other serious conditions. In response, the EPA continues to work with states to improve air quality monitoring and enforce standards that address the sources and impacts of pollution [3].

2.2 Key Air Quality Indicators

Air quality significantly impacts public health, particularly respiratory health. Our research delved into the relationship between key air quality indicators—namely particulate matter ($PM_{2.5}$, PM_{10}), nitrogen dioxide (NO_2), carbon monoxide (CO), ozone (O_3), and sulfur dioxide (SO_2)—and asthma-related health outcomes, including hospitalization rates and emergency room visits. Each of these pollutants originated from both natural and anthropogenic sources and had the potential to aggravate cardiovascular diseases, influence infant mortality rates, and provoke a range of respiratory diseases [9].

2.2.1 Particulate Matter (PM)

Particulate Matter (PM), including $PM_{2.5}$ and PM_{10}, comprises fine particles that pose serious threats to respiratory health. Originating from both natural sources, like wildfires and volcanic emissions, and anthropogenic activities, including vehicle exhaust, industrial combustion, and residential fuel burning, these particles are notorious for their ability to penetrate deep into the lungs. For asthma sufferers, the inhalation of PM can exacerbate symptoms, leading to increased frequency and severity of asthma attacks [22]. The smaller the particle size, the deeper they can infiltrate into the pulmonary system, with $PM_{2.5}$ being particularly detrimental because it bypasses the upper respiratory tract defenses and lodges directly in the mucous membrane [10]. This penetration can trigger inflammatory responses in the airways, worsening asthma conditions and potentially leading to more frequent emergency room visits and hospitalizations [22].

2.2.2 Nitrogen Dioxide (NO_2)

Nitrogen Dioxide (NO_2), primarily emitted from vehicle engines and other sources of high-temperature combustion, is a prevalent traffic-related pollutant [10, 22]. Transportation contributes up to 80% of ambient NO_2 levels, serving as a significant factor in air pollution. NO_2 is known to irritate the respiratory system, causing symptoms such as coughing, wheezing, dyspnea, and severe conditions like bronchospasm and pulmonary edema at high exposure levels. Particularly harmful to asthmatics, NO_2 exposure can exacerbate asthma symptoms and increase the frequency of hospital and ER visits due to respiratory complications. Studies show exposure to even low levels of NO_2 has been consistently found to reduce lung function and worsen asthma across various age groups, particularly children and individuals with pre-existing respiratory conditions [22].

2.2.3 Carbon Monoxide (CO)

Carbon Monoxide (CO) is produced by the incomplete combustion of fossil fuels and is prevalent in motor vehicle exhaust and industrial emissions. CO interferes with the blood's ability to carry oxygen, leading to serious health conditions such as cardiovascular diseases, neurological damage, and death at high concentrations. Chronic exposure to lower concentrations can also affect cardiovascular health and fetal development [10]. Recent studies have identified a specific association between CO exposure and moderate to severe asthma exacerbations in adults. Although similar associations have not been confirmed in children, there is evidence

showing that lower ambient levels of CO are correlated with decreased asthma-related death rates [22].

2.2.4 Sulfur Dioxide (SO_2)

Sulfur Dioxide (SO_2) is primarily emitted from the combustion of fossil fuels by power plants and other industrial facilities. As a potent respiratory irritant, SO_2 penetrates deep into the lung where it is transformed into bisulfite, causing bronchoconstriction. This process significantly worsens respiratory conditions such as bronchitis and mucus production, particularly among individuals with asthma [10]. Asthmatic individuals, especially those with allergic asthma, are highly sensitive to SO_2 and experience more severe symptoms and greater declines in lung function at lower concentrations than non-asthmatics. Research has shown a clear association between SO_2 exposure and both asthma prevalence and exacerbation of asthma symptoms in children aged 0 to 18 years, leading to increased hospital and emergency room admissions [22].

2.2.5 Ozone (O_3)

Ground-level Ozone (O_3) forms through chemical reactions with oxides of nitrogen (NO_x) and volatie organic compounds (VOC) in the presence of sunlight . It is a significant component of urban smog where its precursors—fossil fuel consumption, industrial emissions, gasoline vapors, motor vehicle exhaust, and chemical solvents—are abundant [10, 22]. With its low water solubility, which allows it to bypass the upper respiratory tract defenses and penetrate deep into the lungs, O_3 is highly reactive and poses serious threats. It interacts with antioxidants in the airway epithelial cells, leading to oxidative stress, resulting in airway inflammation, airway hyperresponsiveness, and lung failure. These effects are pronounced in asthmatic individuals, where O_3 can aggravate asthma symptoms, increase hospital and ER admissions, and exacerbate respiratory diseases. Epidemiological studies claim short-term exposure to elevated O_3 levels leads to increased emergency room visits and hospitalizations, particularly during warmer seasons. Moreover, long-term exposure to O_3 has been linked to detrimental effects on lung function development in children and progression of asthma to chronic obstructive pulmonary disease (COPD) in adults [22].

2.3 Air Quality Index

To quantify and categorize air quality in a manner that aligns with health outcomes, the Air Quality Index (AQI) is used extensively. The AQI is a standardized indicator that helps the public understand how clean or polluted the air is on a daily basis and the potential health effects concerning different levels. Below is a detailed overview of the AQI categories and their health implications, which serve as key metrics in our study. The EPA database is instrumental in our research, offering data on the number of days each county records within each AQI category.

1. **Good (0 to 50):** Air quality is satisfactory, and air pollution poses little or no risk.
2. **Moderate (51 to 100):** Air quality is acceptable; however, there may be a concern for some people who are unusually sensitive to air pollution. Sensitive individuals should consider limiting prolonged outdoor exertion.
3. **Unhealthy for Sensitive Groups (101 to 150):** Members of sensitive groups may experience health effects while the general public is less likely to be affected. Sensitive groups, including children, active adults, and people with respiratory diseases such as asthma, are advised to limit prolonged outdoor exertion.

4. **Unhealthy (151 to 200):** Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects. It is advised that children, active adults, and people with respiratory diseases such as asthma avoid outdoor exertion while everyone else should limit prolonged outdoor exertion.

5. **Very Unhealthy (201 to 300):** Everyone may experience more serious health effects. Children, active adults, and people with respiratory diseases such as asthma avoid outdoor exertion; everyone else should limit prolonged outdoor exertion.

6. **Hazardous (301 to 500):** Health warnings of emergency conditions. The entire population is likely to be affected. It is advised that everyone avoid all physical activity outdoors.

Air Quality Index (AQI)	Levels of Health Concern
0-50	Good
51-100	Modest
101-150	Unhealthy For Sensitive Groups
151-200	Risky
201-300	Harmful
301-500	Dangerous

Table 1: Air Quality Index [18]

2.4 Asthma Epidemiology

Asthma, a prevalent chronic condition characterized by chronic airway inflammation, affects approximately 334 million people globally, with incidence rates significantly higher in developed countries, sometimes reaching up to 20% [6, 16]. Symptoms such as wheezing, dyspnea, cough, and chest tightness, which are often accompanied by variable expiratory airflow limitation, define the clinical presentation of the disease. The disease impacts all age groups and does not discriminate by race or ethnicity, although these factors can influence prevalence, morbidity, and mortality rates across different regions. In the United States, asthma affects 7.6% of the population, with notable variations among ethnic groups—rates are particularly high among Black non-Hispanics and Puerto Ricans compared to Mexicans and Asians [6]. As depicted in Figure 2, data from the CDC shows trends in asthma prevalence from 2010 to 2021 among children and adults, indicating periods of fluctuation yet a generally stable pattern over recent years.

Recent epidemological research has consistently shown the adverse effects of air pollution on asthma, showing both an increase in exacerbations and new asthma diagnoses, despite overall advancements in air quality over the last several decades. NO_2, ozone, and PM, which are regulated under the US Environmental Protection Agency's Clean Air Act, have been particularly implicated. For instance, exposure to NO_2 is linked to a 14% increase in asthma exacerbations, and PM exposure contributes to approximately 8% of new asthma cases annually among children globally. Proximity to high-traffic areas, major sources of NO_2 and PM, correlates with a significant uptick in asthma symptoms, with children living within 500 meters of busy roads at a 25% higher risk of developing asthma. Implementing successful regulatory and public health interventions, such as traffic emission controls, has proven to reduce asthma morbidity by as much as 15% in affected populations, demonstrating the role of environmental exposure in asthma prevalence [15].

Globally, the prevalence of asthma varies widely, influenced by urbanization, lifestyle, and socioeconomic factors. While developed nations report higher rates, contributing factors in lower prevalence in developing regions include lifestyle, environmental exposures, and underdiagnosis. The "hygiene hypothesis" suggests that reduced microbial exposure in cleaner environments increases asthma risk by limiting early immune system training. Additionally, a significant rise in asthma prevalence is projected globally, with an expected 100 million new cases over the next decade [6].

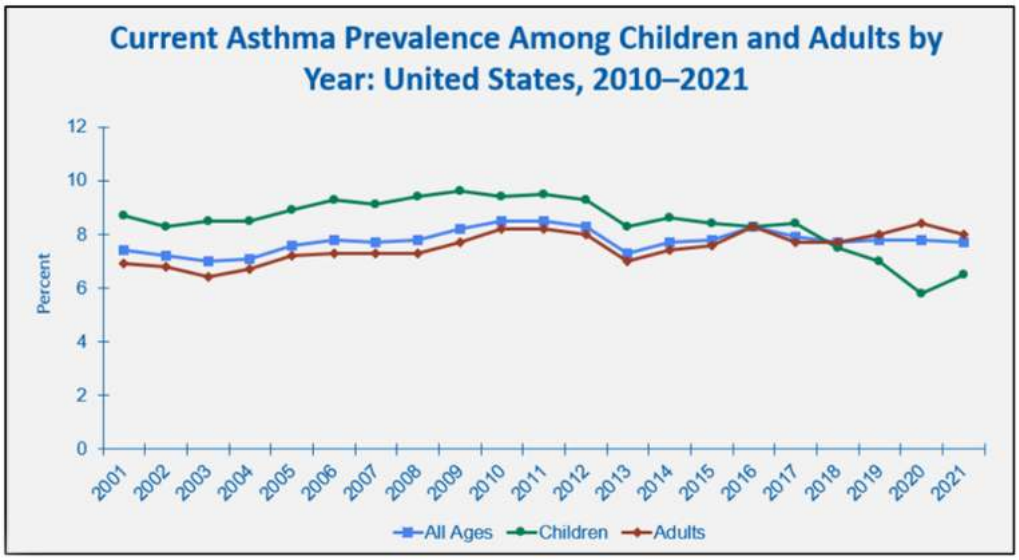

Figure 2: Trends in Asthma Prevalence by Age Groups, 2001-2021 [15]

2.4.1 Genetics, Environment, and Asthma

Asthma research continuously indicates the intricate interplay between genetic predisposition and environmental factors in the etiology of the disease. With heritability estimates for asthma susceptibility ranging from 35% to 95%, it's clear that genetics play a significant role [13]. Large-scale genomic studies have demonstrated this complexity, identifying numerous genomic regions and over 100 genes associated with allergy and asthma across various populations, emphasizing the importance of genetic makeup in the development of asthma [20, 21]. Genetic predispositions also play a crucial role in how individuals respond to these pollutants. Polymorphisms in glutathione-S-transferases (GSTs), such as the GSTM1 gene, influence the body's ability to combat oxidative stress from environmental pollutants. Research indicates that individuals with null variants of GSTM1, who lack this protective enzyme, are more susceptible to ozone and show a 12% higher rate of asthma exacerbations compared to those with active GSTM1 [15].

Allergen sensitization is another significant genetic factor where high levels of total serum immunoglobulin E, indicating allergen sensitivity, have been correlated with the incidence of asthma. The development of asthma is also intricately linked to immune responses in early life. For instance, an impairment in interferon production at an early age has been associated with an increased risk of wheeze, while other studies specify the role of polymorphisms in Toll-like receptors in shaping immune responses that may contribute to persistent asthma [20]. The environmental aspect is equally significant, with factors such as respiratory infections, airborne pollutants, and allergen sensitization playing pivotal roles in both the onset and exacerbation of asthma [13, 17]. The consistent association between postnatal exposure to environmental tobacco smoke and respiratory symptoms, including the exacerbation of asthma and its severity, exemplifies the impact of air quality on respiratory health [20]. The same goes for exposure

to animals; certain exposures in early life, such as to farm animals, have been associated with a lower risk of developing atopy and asthma, whereas domestic pets have had mixed findings [20]. Moreover, the effects of gene-environment interactions in asthma are often complex and multi-layered. Some genetic variants may modify the body's response to inhaled agents, while environmental exposures can also directly influence gene expression through epigenetic mechanisms. This suggests that lifestyle and environmental exposures are substantial players in disease susceptibility [20]. Interestingly, sex and gender have time-dependent effects on the development of asthma. Until adolescence, boys are more likely to develop asthma than girls; however, post-puberty, the trend reverses, with females exhibiting a higher incidence of asthma [20]. This shift could be influenced by hormonal changes during puberty, although the precise mechanisms remain to be fully understood.

2.5 Link Between Air Quality and Asthma

Since the 1960s, studies have established a correlation between pollution and both the incidence and severity of asthma. Modern research aligns with historical data, showing that pollutants like nitrogen dioxide (NO_2), carbon monoxide (CO), and particulate matter ($PM_{2.5}$) are linked to DNA methylation changes associated with asthma [4, 17]. Asthma, as a chronic inflammatory disease of the airways, is susceptible to the quality of inhaled air; hence, pollutants such as particulate matter (PM), nitrogen dioxide (NO_2), carbon monoxide (CO), sulfur dioxide (SO_2), and ground-level ozone (O_3) have been recognized for their roles in both triggering acute asthma exacerbations and contributing to the disease's development. The airborne particulates, especially $PM_{2.5}$ and PM_{10}, can penetrate the respiratory tract, instigating inflammation and impairing lung function, thereby elevating the risk of asthma attacks. NO_2, often associated with traffic emissions, has been correlated with increased respiratory symptoms and decreased lung function. CO, resulting primarily from incomplete combustion, impacts oxygen delivery within the body, exacerbating asthma's effects, while SO_2 can trigger bronchoconstriction and exacerbate pre-existing respiratory conditions. Ozone, a secondary pollutant formed by the reaction of sunlight with other pollutants, can cause oxidative stress and inflammation within the lungs, leading to increased emergency visits and hospitalizations for asthma patients [10, 22].

2.5.1 Air Quality Standards

Asthma prevalence, a key indicator of chronic airway inflammation, remains persistently high despite advancements in reducing air pollutant levels. The U.S. Environmental Protection Agency has established national ambient air quality standards to regulate the concentration of air pollutants, including carbon monoxide, NO_2, ozone, sulfur oxides, and PM, to levels intended to be protective of public health (see Figure 3) [23]. Notably, the acceptable limits set for NO_2 at an annual mean of 0.053ppm and ozone at an 8-hour standard of 0.08ppm, as well as PM_{10} and $PM_{2.5}$ at annual means of $50\mu g/m^3$ and $15\mu g/m^3$ respectively, mitigate the exacerbation of respiratory diseases [15]. However, emerging epidemiological evidence indicates that even levels of pollutants within these regulatory limits may not be sufficiently low to prevent adverse health outcomes. Specifically, studies have found associations between exposure to levels of NO_2 and ozone below current EPA standards and increased asthma symptoms in children, suggesting that the current standards may not fully protect vulnerable populations. For instance, Gent et al. have shown that ozone and $PM_{2.5}$ levels beneath the thresholds of 0.12ppm (1-hour) and $50\mu g/m^3$ (annual mean) are still correlated with exacerbations of asthma [9].

Carbon monoxide	Lead	NO_2	Ozone	Sulfur oxides	PM_{10}	$PM_{2.5}$ (1997 proposed standards)
9 ppm (8-h average)	1.5 μg/m^3 (quarterly average)	0.053 ppm (annual mean)	0.08 ppm (8-h standard)	0.14 ppm (24-h mean)	50 μg/m^3 (annual mean)	15 μg/m^3 (annual mean)
35 ppm (1-h average)			0.12 ppm (1-h average)	0.03 ppm (annual mean)	150 μg/m^3 (24-h mean)	65 μg/m^3 (24-h mean)

Figure 3: Primary National Ambient Air Quality Standards, United States [15]

Regional monitoring, which provides an overview of air quality based on fixed stations, might not capture the full extent of individual exposure, especially for those spending significant time outdoors, where they face a greater burden of pollutants. Personal exposure is evident in research where portable monitoring devices reveal a stronger correlation between health outcomes and pollutant levels than data from regional monitors [15].

Air pollution is proven to have an influence on asthma exacerbation and the risk of developing asthma. In children, exposure to NO_2 has been associated with respiratory symptoms like coughing and wheezing, and higher levels of this pollutant in the home environment correlate with an increased risk of respiratory disease. Asthmatic children have shown heightened vulnerability to ozone, with studies in Southern California and Atlanta linking ambient exposure to higher emergency department visits. Additionally, particulate matter, especially the finer $PM_{2.5}$, is implicated in increased asthma morbidity. A stark example of the link between air quality and health is observed in the Utah Valley, where a temporary reduction in steel mill activity, and consequently particulate emissions, coincided with a marked decrease in respiratory disease exacerbations [15]. This evidence indicates a pressing need for re-evaluation of air quality standards to better protect public health, particularly for those with preexisting conditions like asthma. The standards, while designed to safeguard, may require adjustment in light of recent findings that even lower levels of pollutants than previously thought still pose significant risks [9].

2.5.2 Controversies

The connection between air pollution and asthma, while widely acknowledged, continues to be a controversial topic of discussion. The relationship is complicated by factors such as the variability in individual sensitivty to air pollutants; some individuals may experience severe asthmatic responses to relatively low levels of pollutants, while others may not be affected at all. Some studies challenge the direct correlation, proposing that even levels below current standards can exacerbate asthma symptoms [15]. Longitudinal studies, such as the one by McConnell et al., demonstrate that children in areas with higher ozone levels are more likely to develop asthma, suggesting a significant environmental impact [11]. Similarly, indoor NO_2 levels have been linked to increased respiratory diseases in children [14]. The debate continues over which pollutants are most impactful, with studies variably highlighting NO_2, SO_2, and $PM_{2.5}$ as primary triggers. Divergent study methodologies also add to the controversy, as differences in research design, population demographics, and pollutant measurement techniques can results in inconsistent findings; this inconsistency points to a need for more targeted research to clarify the roles of specific pollutants [4, 15].

2.6 Project Plan

Previous research has demonstrated a correlation between air quality and increased asthma cases, providing a basis for the application of machine learning models. This project aimed to expand upon existing research by focusing on specific pollutants and their roles in asthma exacerbation. By leveraging extensive datasets from the CDC and EPA, we have designed a comprehensive analysis framework to explore the effects of specific air pollutants on asthma.

Our investigation was guided by several critical research questions. First, we sought to determine the most significant environmental contaminants contributing to asthma incidence, aiming to identify which pollutants were most critical in exacerbating the condition. Additionally, we analyzed how various environmental contaminants, such as particulate matter, nitrogen dioxide, sulfur dioxide, carbon monoxide, and ozone correlated differently with asthma incidence and health outcomes like hospitalizations and emergency room visits.

Furthermore, our project explored how machine learning algorithms could effectively analyze air quality and health data to understand patterns of asthma incidence. By applying these models to historical data from the EPA and CDC, we determined how asthma health outcomes were affected by air quality indicators such as pollutant exposure across different counties. We also compared the effectiveness of different machine learning models, specifically Random Forest and SARIMAX, to determine which algorithm offered the best accuracy and reliability for asthma trends. Through addressing these questions, we identified the most impactful pollutants and refined analytical techniques for better public health strategies.

3 Methods

3.1 Data Sources

Our study utilized a combination of comprehensive asthma and air quality datasets to examine the intricate relationships between environmental pollutants and asthma-related health outcomes. Specifically, we drew upon three critical sources of asthma data: yearly statistics for emergency room visits, hospitalizations, and prevalence rates of asthma. These datasets were sourced from the CDC's National Environmental Public Health Tracking Network, which provided detailed epidemiological data collected from a variety of surveys, including the Behavioral Risk Factor Surveillance System (BRFSS).

3.1.1 Search Strategy and Selection Criteria

To compile relevant literature for our investigation, we systematically explored online research databases including Springer Link, PubMed Central, BioMed Central, and ScienceDirect. Our search strategy revolved around identifying articles containing keywords associated with machine learning methodologies in conjunction with terms such as "air quality","asthma", or "air pollution". The inclusion criteria were limited to publications from 1991-2024, with a focus on the most recent ten years to ensure the relevance and timeliness of the data and studies included.

3.1.2 Asthma Datasets

The CDC National Environmental Public Health Tracking Network is a collection of data comprising of census data, the Behavioral Risk Factor Surveillance System (BRFSS), and other epidemiological surveys from which we will be extracting our datasets from. Specifically we will be querying asthma data by county and year for emergency room visits, hospitalizations, and prevalence. Shown below are three maps generated using the CDC data explorer of the

metrics listed above for the year 2018. Multiple years going back as far as 2000 were included and the datasets were exported to use for our project.

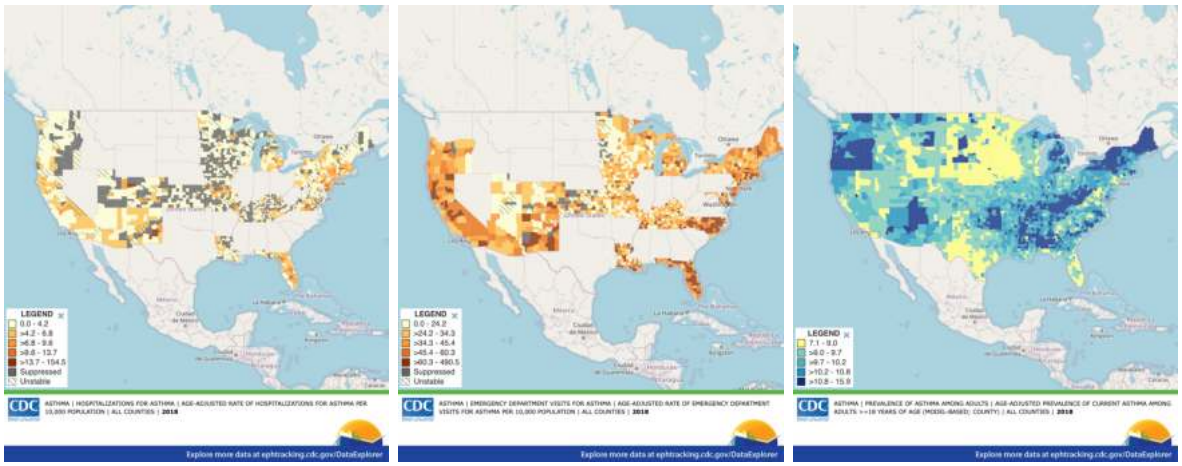

Figure 4: Distribution of Hospitalizations, ER Visits, and Asthma Prevalence for 2018 [5]

3.1.3 Air Quality Datasets

The EPA database supplies yearly air trend reports that encompass standard concentrations and emissions, speciation concentrations, spotlight concentrations, toxic concentrations, visibility data, alongside county and state FIPS codes. Our analysis will specifically target air trends spanning the years available in the asthma datasets above. The selected datasets encompass concentrations of various air pollutants, such as nitrogen dioxide (NO_2), carbon monoxide (CO), ozone (O_3), sulfur dioxide (SO_2), and particulate matter with diameters of 2.5 micrometers or less ($PM_{2.5}$), particulate matter with diameters of 10 micrometers or less (PM_{10})[2].

3.1.4 Methodological Approach

With our questions as a guide, our methodological approach involves the following steps:

1. Data Preparation: Rigorous cleaning and preprocessing of data to ensure accuracy and consistency.
2. Model Development: Application and tuning of Random Forest and SARIMAX algorithms to the prepared datasets.
3. Validation: Splitting data into training and test sets to evaluate the model's predictive capabilities on the testing data.
4. Performance Assessment: Quantifying model success with statistical metrics and interpreting results.

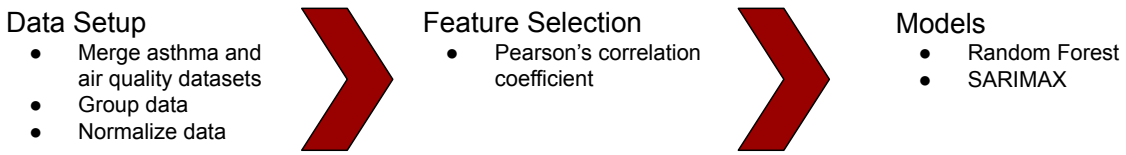

Figure 5: Data Preparation and Analysis Workflow

3.2 Data Processing

Asthma datasets spanning the years 2000 to 2021 were downloaded from the Centers for Disease Control and Prevention (CDC). The prevalence dataset was omitted due to its limited four years of data and significant number of missing values. Air quality data was obtained from the Environmental Protection Agency (EPA) data repository, consisting of yearly averages of pollutant levels by county. The two Asthma datasets were matched by FIPS code to the air quality dataset and merged. The years 2005 to 2019 were selected as they consisted of the most counties reporting data and had the most complete data. Counties for these years that had missing feature values were dropped.

The combined datasets were then grouped into three subgroups based on the slope of linear regression models fitted to each county's data:

- Group A: Counties with an increasing slope (≥ 0.1)
- Group B: Counties with a decreasing slope (≤ -0.1)
- Group C: Counties with a neutral slope (-0.1 to 0.1)

For outlier removal, the low and upper bound cutoffs were determined using the Q1 quartile, Q3 quartile, and inter quartile range shown in the equations below.

$$lowerbound = q1 - 1.5 * iqr \tag{1}$$

$$upperbound = q3 + 1.5 * iqr \tag{2}$$

Boxplots were constructed and outliers were removed on the appropriate transformed or un-transformed datasets.

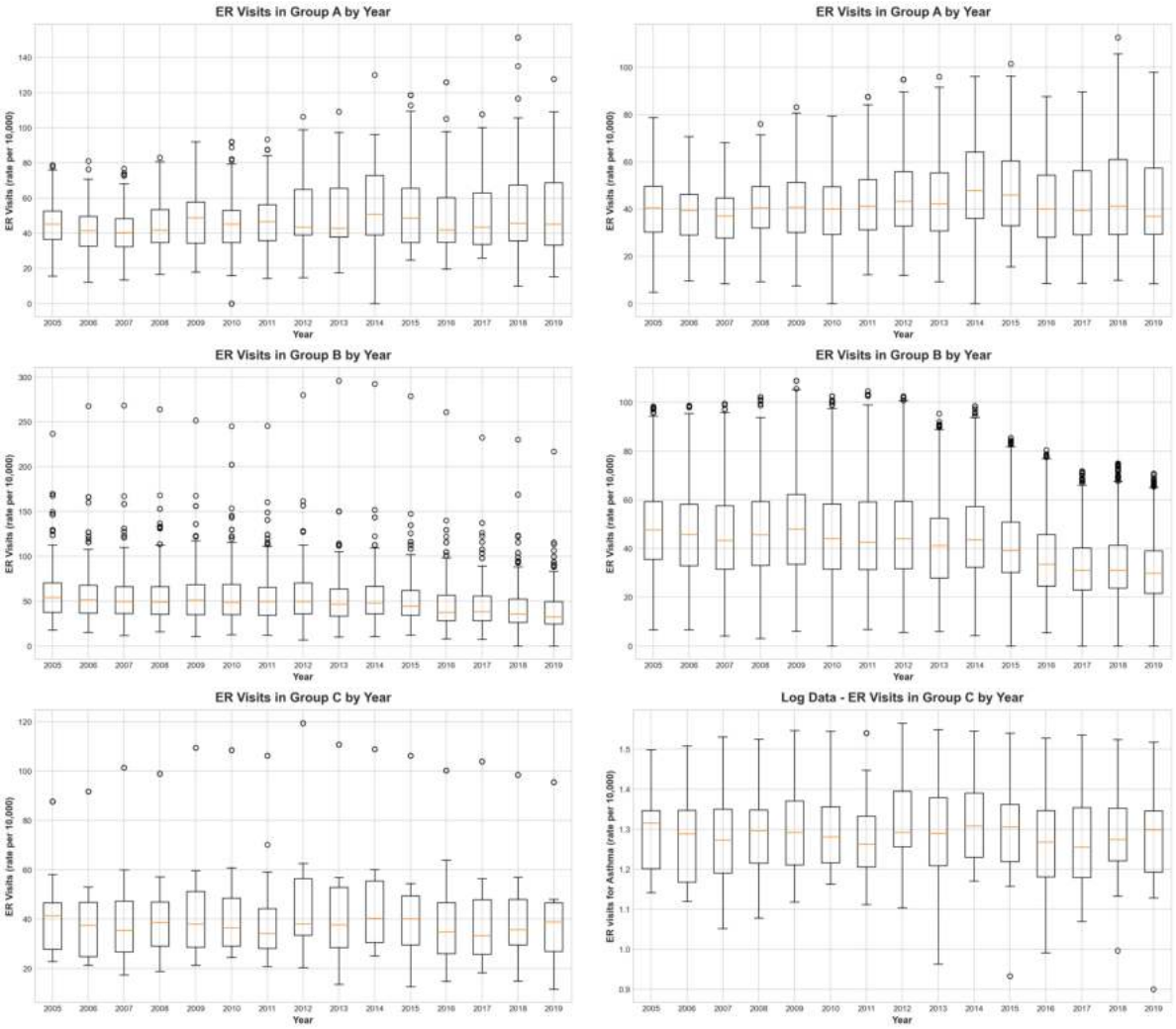

Figure 6: Emergency Room Visits, Boxplots

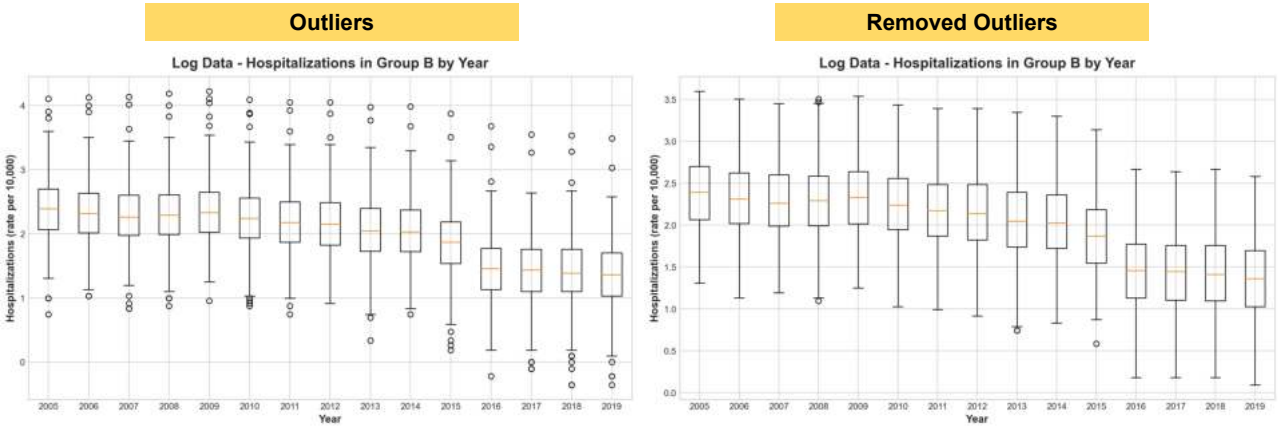

Figure 7: Hospitalization Visits, Boxplots

3.2.1 Feature Selection

In order to identify the most significant correlations in air quality measurements for asthma hospitalizations and emergency room visits, we utilized Pearson's correlation heatmap. This method provided a clear visual representation of the relationship between various environmental

pollutants and asthma outcomes. Our comprehensive list of factors include Days with AQI, Good Days, Moderate Days, Unhealthy for Sensitive Groups Days, Unhealthy Days, Very Unhealthy Days, Hazardous Days, Max AQI, 90th Percentile AQI, Median AQI, Days CO, Days NO_2, Days Ozone, Days $PM_{2.5}$, and Days PM_{10}.

Note: Our Pearson correlation heatmaps below were condensed to just a few rows to simplify visualization and highlight key findings. Additionally, the graphs below show heatmaps before outlier removal and transformations.

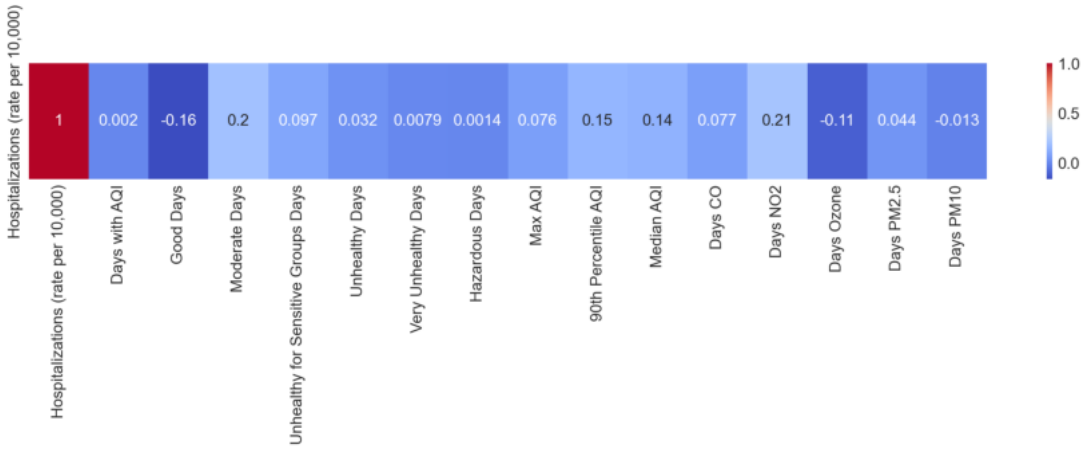

Figure 8: Pearson's Correlation Heatmap for Hospitalization Groups

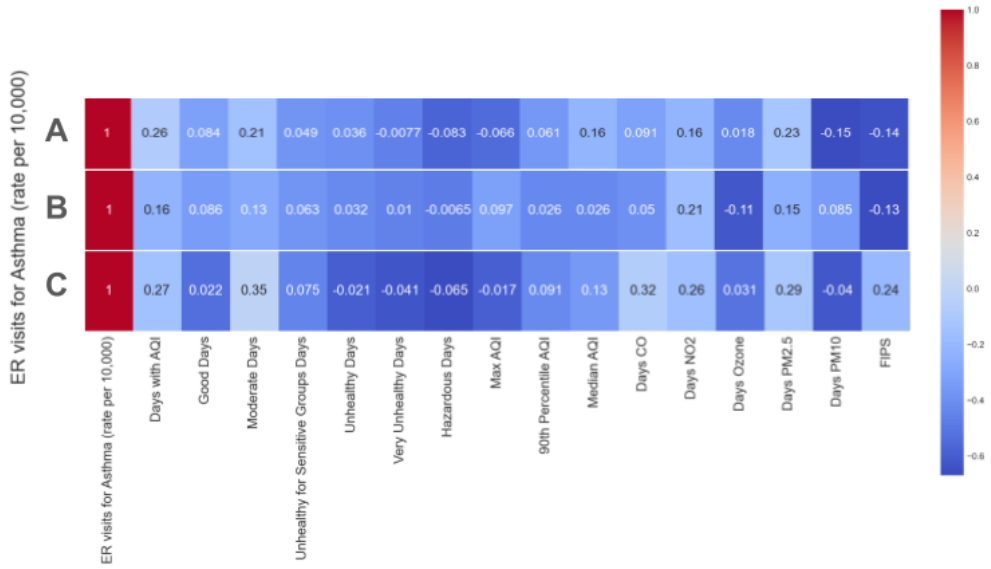

Figure 9: Pearson's Correlation Heatmap for ER Groups

The correlation analysis provided detailed insights for each of our groups:

- In Group A for ER visits, significant positive correlations were found with Days with AQI, Moderate Days, Median AQI, Days NO_2, and Days $PM_{2.5}$.
- Group B showed significant correlations for ER visits with Days with AQI, Moderate Days, Days NO_2, and Days $PM_{2.5}$.

- For Group C ER visits, positively correlated factors included Days with AQI, Moderate Days, Median AQI, Days CO, Days NO_2, and Days $PM_{2.5}$.
- In Group B for Hospitalizations, Moderate Days, 90th Percentile AQI, Median AQI, and Days NO_2 were significant.

These results suggested variable impacts of different pollutants and AQI-related factors on asthma exacerbations across the groups. While all examined features were incorporated into our predictive model, we planned on giving special attention to those with a significant correlation. To prioritize these features, we utilized Random Forests to assess the relative importance as well as guide the model training process.

3.3 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) in this study served as foundational phase in determining the suitability of modeling techniques and refining hypotheses based on observed data characteristics. Following data preparation, both univariate and multivariate analysis techniques were employed to explore the structure and relationships among the variables. For univariate analysis, histograms and box plots were generated to visualize the distributions of each variable, identifying skewness or outliers that could influence further analysis. For multivariate analysis, scatter plots and heat maps were created to probe the relationships between air quality indicators and health outcomes to facilitate a deeper understanding of these interactions [8]. The objectives of our EDA include:

- Refining initial hypotheses by identifying key trends and potential anomalies in the data.
- Guiding the selection of appropriate machine learning models, such as Random Forest and SARIMAX, tailored to the patterns and relationships identified.
- Designing the model training and validation phases to ensure our predictive models are reflective of current research studies.

The results of the EDA were expected to be instrumental in shaping the subsequent phases of the research process. We anticipated that visualizations such as bar graphs, histograms, and scatter plots would provide insights into the dynamics of asthma-related health outcomes and the impact of air quality. We used these visual tools to guide our investigative focus and assist in the development of predictive models and hypotheses.

3.4 Statistical/Machine Learning Methods

To further our understanding of the dynamics between air quality indicators and asthma health metrics, we employed predictive models, each selected for its strengths in capturing different aspects of the data's structure and underlying patterns.

3.4.1 Pearson's Correlation Coefficient

To investigate the relationship between air quality indicators and asthma-related health outcomes, this program employed Pearson's correlation coefficient to determine the most statistically significant features. Pearson's correlation coefficient is the basis of our study in determining and quantifying the degree and direction of association between the levels of specific pollutants and the incidence of hospitalizations and emergency room visits due to asthma.

Pearson's correlation coefficient was utilized with the Python Seaborn library. More specifically, Seaborn's heatmap() method was employed to create various correlation heatmaps for

the three datasets, asthma prevalence, asthma hospitalizations, and asthma ER visits. Additionally, specific county graphs were created to determine the best way to approach model construction, either on a country view, state, or county view. County specific air quality variables were associated with higher correlation values than state or country together. The .corr function in python was used to generate boxplots of correlation values by air quality feature for each county.

3.4.2 Random Forest

Random Forest, an ensemble learning method, combines multiple decision trees to improve predictive performance and controls overfitting [12]. By aggregating the predictions of numerous trees, each trained on a random subset of data, Random Forests offered insights into the nonlinear interactions between air quality indicators and health outcomes. The decision to focus on a single county stemmed from a hypothesis that localized environmental and health data might reveal patterns not easily discernible in a broader analysis. By filtering our dataset for a single county and selecting features with a high correlation, such as 'Days $PM_{2.5}$' and 'Unhealthy for Sensitive Groups Days', we hoped to observe a strong R-squared score and Mean Squared Error in our Random Forest model.

We meticulously prepared our dataset for predictive analysis by addressing missing values, normalizing, transforming, and selecting pivotal target variables. Utilizing the scikit-learn library, the data was split into a training set, used to develop the model, and a test set, reserved for evaluating its predictive performance. This split is instrumental to validate the model's predictive power and ensure its applicability in real-world scenarios. Employing the RandomForestRegressor algorithm, we ensured reproducibility by setting a predetermined number of estimators and a constant random state. With predictions made on the testing set, which encompassed county-level data, we embarked on a rigorous evaluation. We utilized metrics like R-squared score, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to quantify the model's accuracy, while also creating visualizations to juxtapose actual and predicted values. These visual aids served to reinforce the model's performance, providing an immediate, tangible understanding of its predictive accuracy.

3.4.3 SARIMAX

The SARIMAX model is a type of time series forecasting method that extends the ARIMA (AutoRegressive Integrated Moving Average) model by incorporating seasonality and external variables, offering a strong framework for understanding temporal dynamics. To build our model, we first prepared a DataFrame containing the target variable, 'Hospitalizations (rate per 10,000)', which we aimed to forecast based on historical data. Recognizing the seasonal patterns often present in healthcare data, such as asthma hospitalizations, we tailored the SARIMAX model to account for this cyclicality by setting the seasonal parameters (P, D, Q, and S) to capture annual patterns. Our choice of model parameters—p, d, q for the non-seasonal component, and P, D, Q for the seasonal component—was guided by the need to model the data's autocorrelation and non-stationarity while also accommodating seasonal trends. The order for each model was set to (1,1,1) for p, d, and q and seasonailty parameters were set to (1, 0, 0, 2). The coefficient with the highest correlation value was assigned as the exogenous variable. The SARIMAX model was then fitted to our dataset spanning from 2005 to 2019 and mdoel fit was evaluated.

4 Results

4.1 Exploratory Data Analysis

The provided bar graphs depicted two distinct trends in asthma-related health outcomes over two decades. The first graph, in Figure 10, shows a general decrease in hospitalization rates for asthma, suggesting an improvement in long-term asthma management and control. The second graph representing emergency room visits for asthma, indicates a rise in acute asthma episodes until around 2015, followed by a notable decline. The initial increase may point to growing incidences of asthma exacerbations, while the recent decrease could reflect better early intervention or changes in environmental factors impacting asthma severity. These trends, observed from the data, emphasized the evolving nature of asthma treatment and the potential impact of environmental quality on patient outcomes. Improvements in air quality typically occur gradually over an extended period, so the significant decline in asthma-related hospital visits could be partially attributed to underlying shifts in healthcare reporting across counties rather than abrupt changes in environmental conditions.

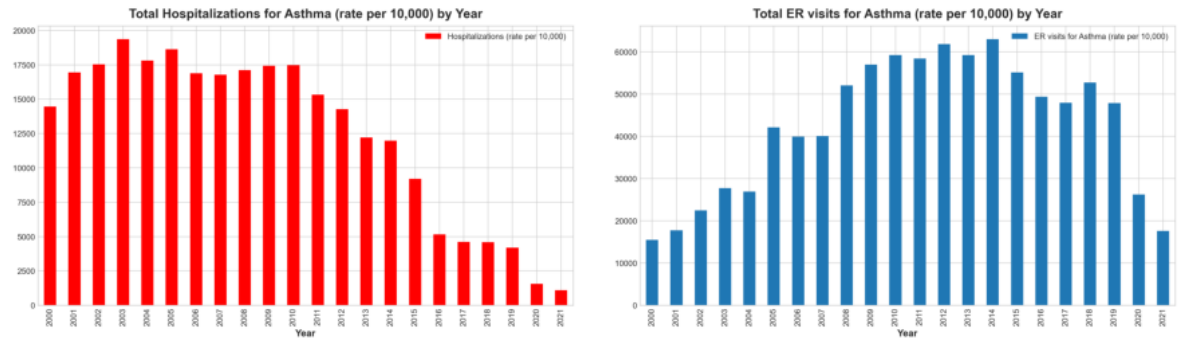

Figure 10: Total Hospitalizations and Emergency Room Visits by Year

The histograms of our three metrics, presented in Figure 11, delineates the distributions of asthma prevalence, hospitalizations, and emergency room visits, offering a comprehensive statistical perspective. The distribution of asthma prevalence follows a relatively normal distribution, clustering around a central value, suggesting a consistent asthma prevalence rate across the observed counties. In contrast, the hospitalizations graph displays a right-skewed distribution, indicating a higher frequency of counties with lower hospitalization rates and fewer with very high rates. The emergency room visits graph exhibits a similar right-skewed pattern, suggesting that while most counties experience a moderate number of visits, a smaller number experience significantly more, which could indicate regional disparities in asthma exacerbations or access to healthcare.

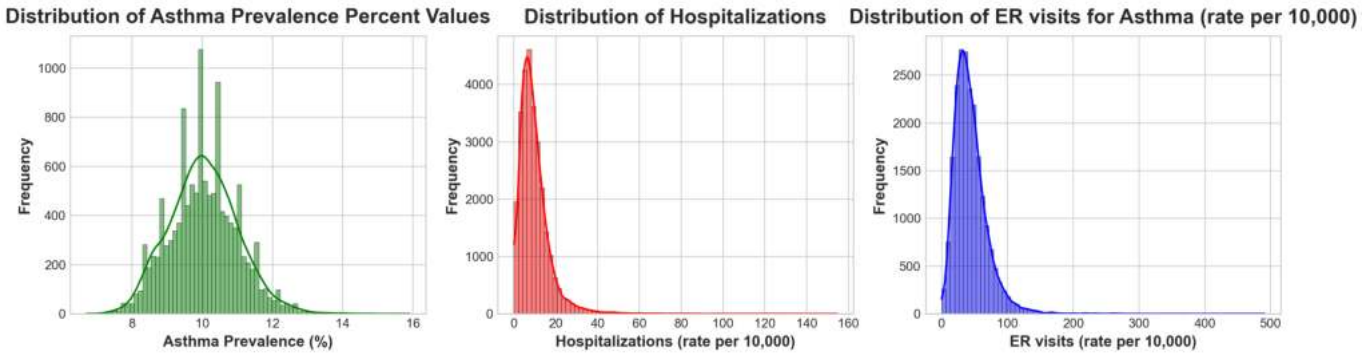

Figure 11: Distribution of Asthma Metrics

Six scatter plots exploring the relationship between days with Air Quality Index (AQI) and various levels of air quality — from good days to hazardous days — show a range of distributions (see Figure 12). These include:

- Good and Moderate Days: Both show a strong positive correlation with the total days with AQI, suggesting that as the total number of monitored days increases, the number of days categorized as good or moderate also rises. This trend might have indicated overall stable or consistent air quality conditions in most areas.
- Unhealthy for Sensitive Groups and Unhealthy Days: These plots revealed that while many areas have few to no days in these categories, some outliers experienced a high number of such days, implying regional variations in air quality that occasionally reach levels posing risks to sensitive populations.
- Very Unhealthy and Hazardous Days: These categories are relatively sparse, but where they do occur, they appear as outliers with few areas experiencing days of such poor air quality. This indicated that extremely poor air quality is not widespread but could signify localized environmental issues.

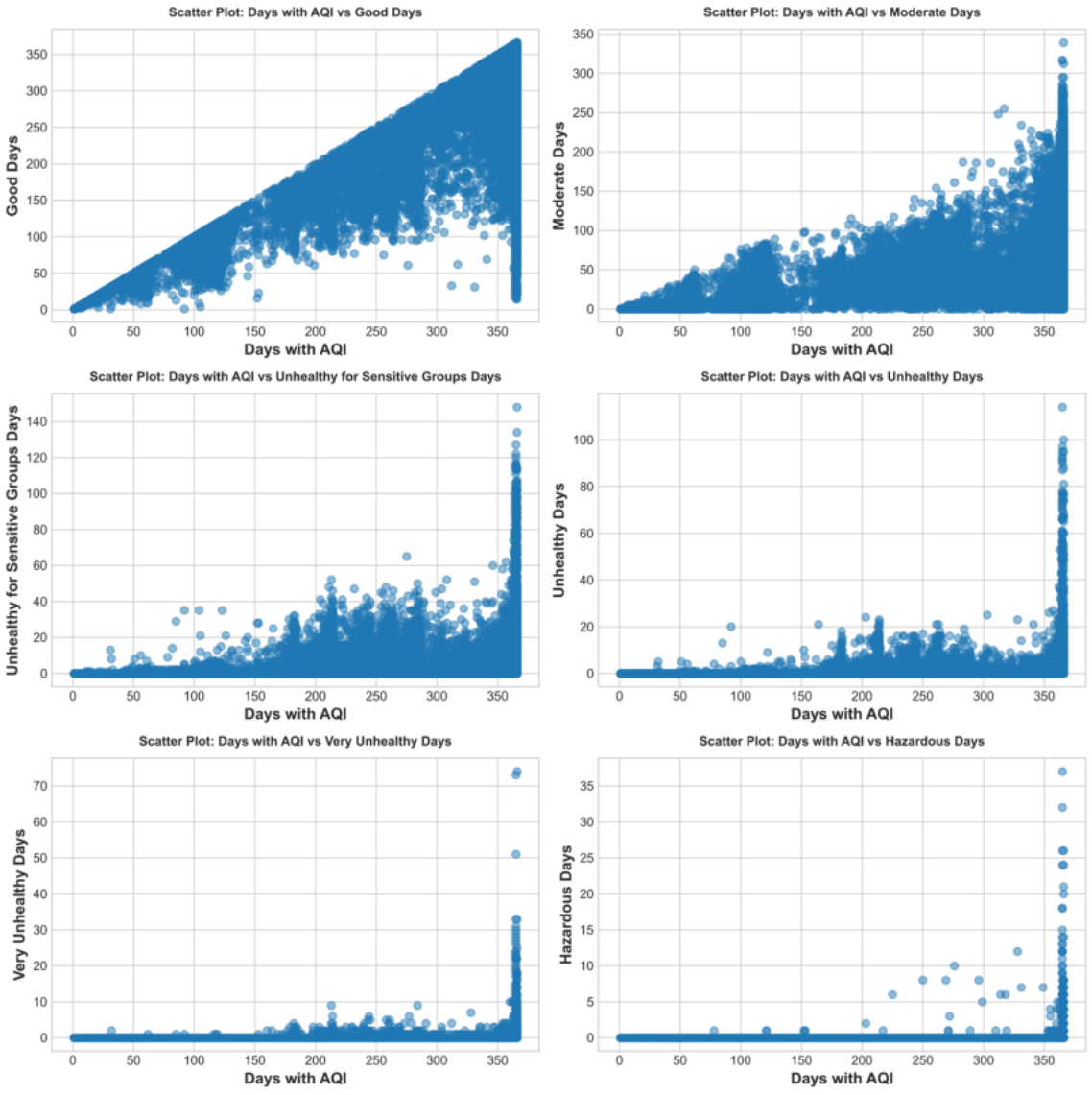

Figure 12: Days AQI vs. AQI Category for Level of Concern

The distribution of median Air Quality Index (AQI) values, reveals a bell-shaped curve that suggests a normal distribution. This pattern indicated that most counties have median AQI values clustered around the mean, reflecting a degree of uniformity in air quality across the sampled regions. Notably, the majority of the median AQI values are concentrated in the lower range of the index, suggesting that air quality is generally within moderate to good categories for these areas. The presence of fewer high-value outliers implied that significantly poor air quality is less common, yet still a concern in some regions (see Figure 13).

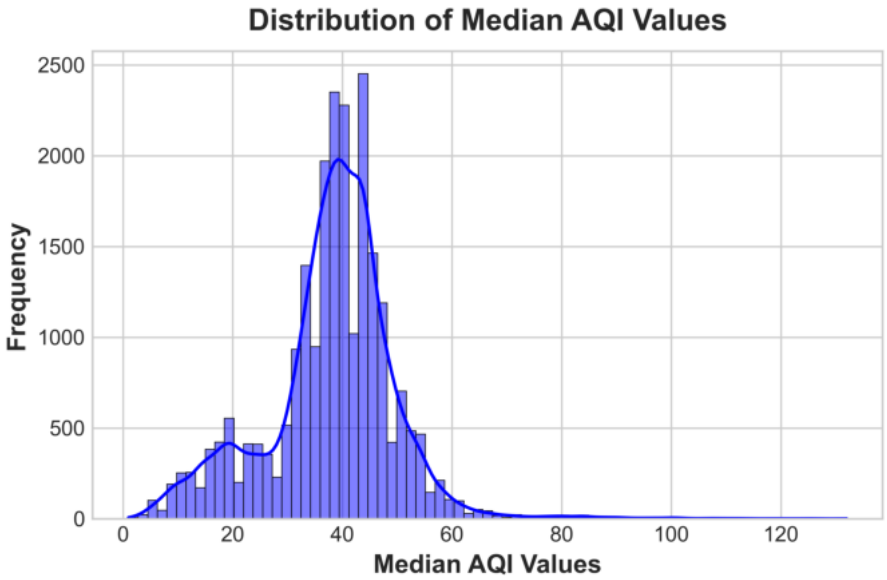

Figure 13: Distribution of Median AQI

A series of scatter plots compared median Air Quality Index (AQI) values with hospitalizations for asthma across multiple years. The plots showed a weak positive correlation in most years, as indicated by the correlation coefficients, suggesting that higher median AQI values had a slight association with increased hospitalizations for asthma. However, the dispersion of data points indicated substantial variability, implying that other factors may have also played a significant role in asthma hospitalization rates (see Figure 14).

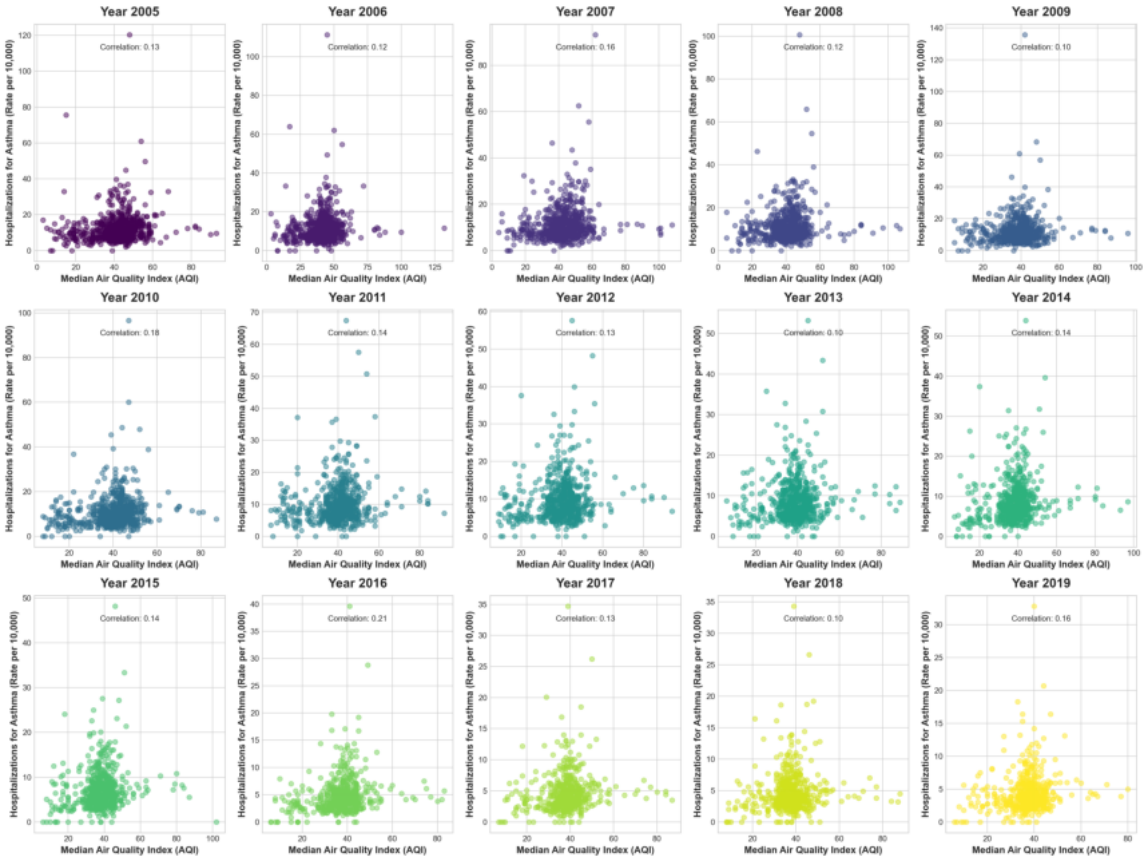

Figure 14: Scatterplots for Median AQI and Hospitalizations by Year

In Figure 15, several years' worth of scatter plots displayed the relationship between median Air Quality Index (AQI) values and ER visits for asthma. The correlation coefficients for each year are generally low, suggesting a weak relationship between the median AQI and the rate of asthma-related ER visits.

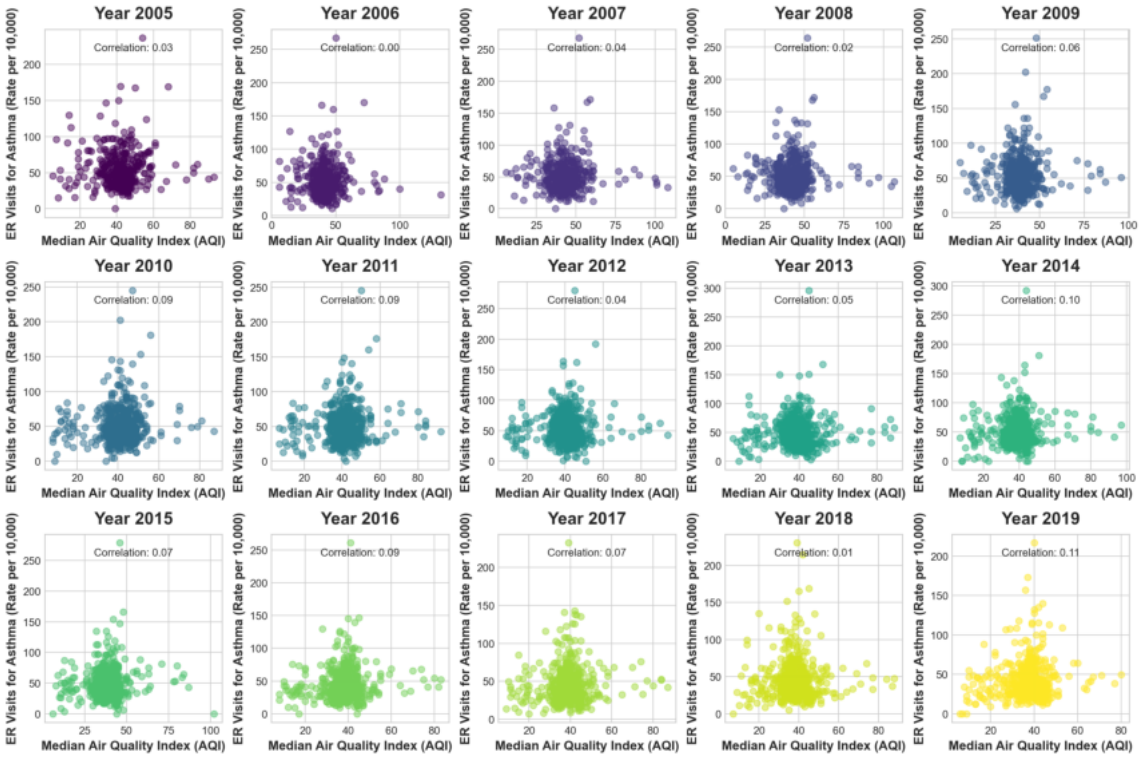

Figure 15: Scatterplots for Median AQI and ER Visits by Year

Maps from 2018 to 2021 detailed the distribution of PM2.5 levels alongside asthma prevalence rates across the United States. While higher PM2.5 levels were commonly associated with increased asthma prevalence, the maps revealed a more nuanced relationship. In some regions, elevated PM2.5 concentrations corresponded with greater asthma prevalence, which aligned with established understanding of pollution's impact on respiratory health. However, this pattern was not universally observed across all maps, indicating that PM2.5 was not the sole factor affecting asthma rates. For instance, some areas with lower PM2.5 levels still showed high asthma prevalence, pointing to the potential influence of other environmental triggers, healthcare access quality, and genetic predispositions. Over the four-year span, the fluctuations in PM2.5 did not appear to directly mirror changes in asthma prevalence, suggesting that long-term exposure or cumulative effects, rather than annual variations in pollution, might have played a more significant role in influencing asthma trends. This complexity was further compounded by inter-annual variability in both $PM_{2.5}$ and asthma prevalence, which could have been attributed to a range of factors including environmental regulations, industrial activities, and public health initiatives (see Figure 16).

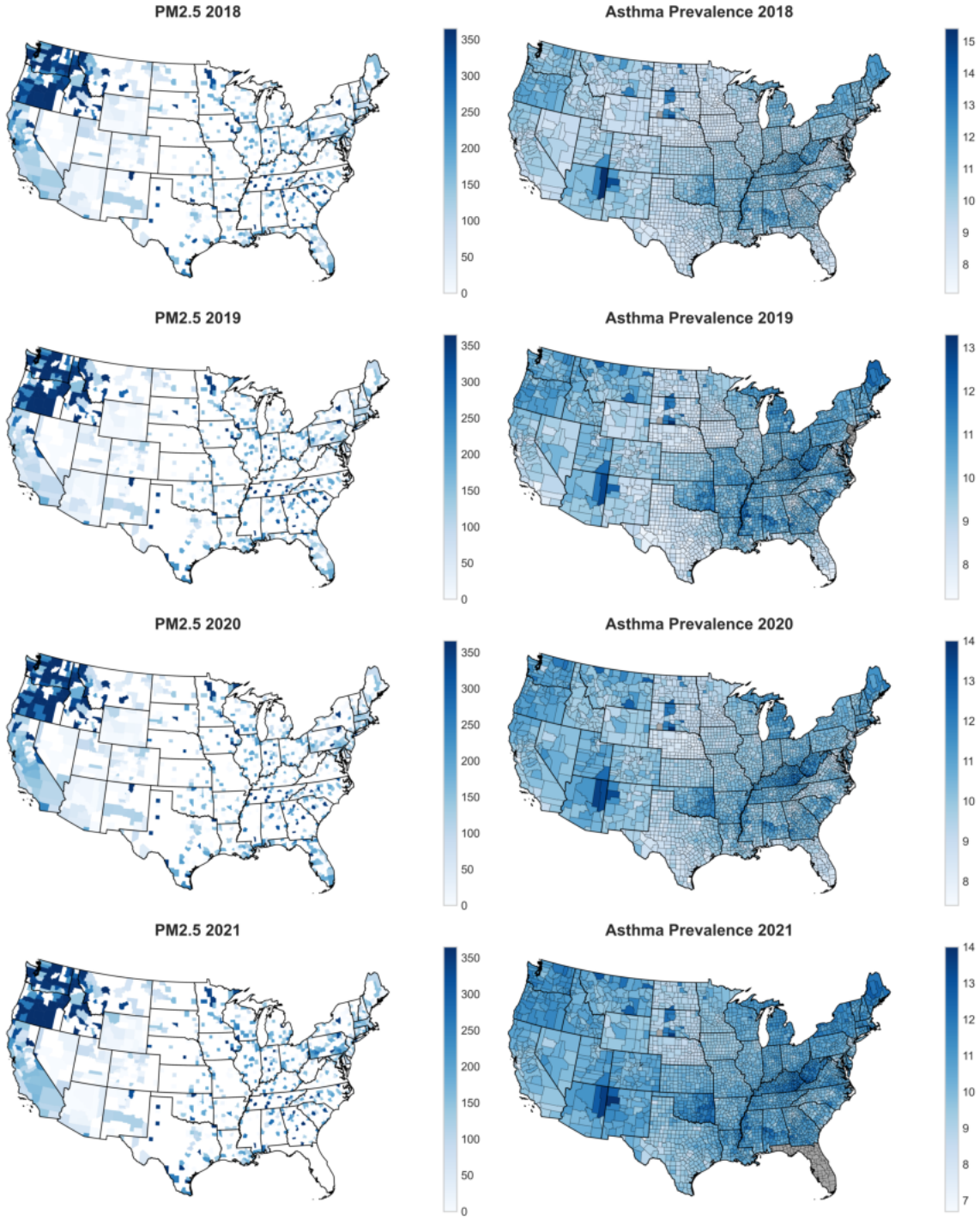

Figure 16: Daily Mean $PM_{2.5}$ Concentration vs. Asthma Prevalence

Utilizing Pearson's coefficient, our results, depicted in Figure 17, indicated that the most predictive air quality variables for asthma hospitalizations include: unhealthy for sensitive group days, median AQI, 90th percentile AQI, Days CO, and Days NO_2. The most predictive air quality variables for asthma prevalence included: unhealthy days, very unhealthy days, and max AQI. The most predictive air quality variables for ER visits for asthma include: moderate days, days CO, and days NO_2.

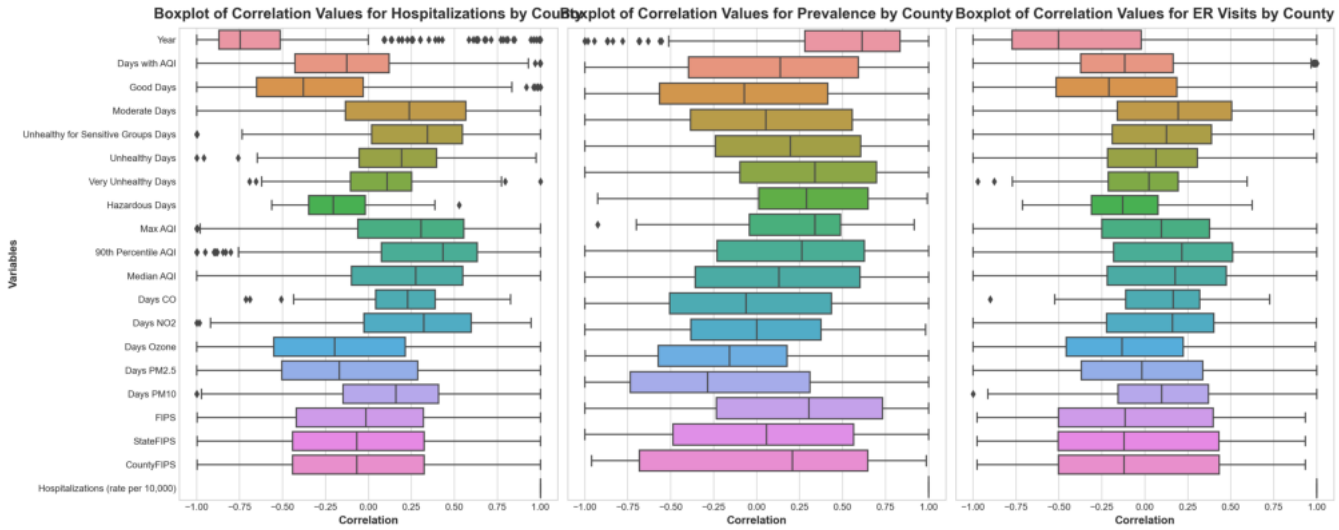

Figure 17: Correlation Boxplots for Hospitalizations, Prevalence, and ER Visits by County

It is important to note that not all counties submitted all types of air quality data, and some did not submit data for all years. Before developing our algorithm, we aim to explore how these factors. We also sought to investigate whether they could be causing anomalies, such as the negative correlations with hospitalizations and ER visits. Another notable aspect of the asthma datasets is that not all counties have submitted data for recent previous years. We theorized that both of these facts may have been contributing to some of the inconsistencies in the correlation values, specifically in the prevalence dataset as it only went back to 2018 and the hospitalization and ER visits datasets started in the year 2000.

4.2 Data Grouping

Linear regression was performed on the selected years for each county in each dataset to determine its grouping. Figure 20 below provided an example of the grouping method using both the Asthma ER visits and Hospitalizations dataset, highlighting the distinct trends observed. Emergency Room visits group A consisted of 113 counties, group B contained 582 counties, and group C had 30. Hospitalizations group A contained data for 3 counties, group B had 671 counties, and group C contained 15 counties. Due to the low number of counties in groups A and C for hospitalizations, these groups were dropped. It was decided not to include these counties in one lump dataset with the counties from group B as the trend difference would likely negatively affect model accuracy and predictability for those that fell under the group B category.

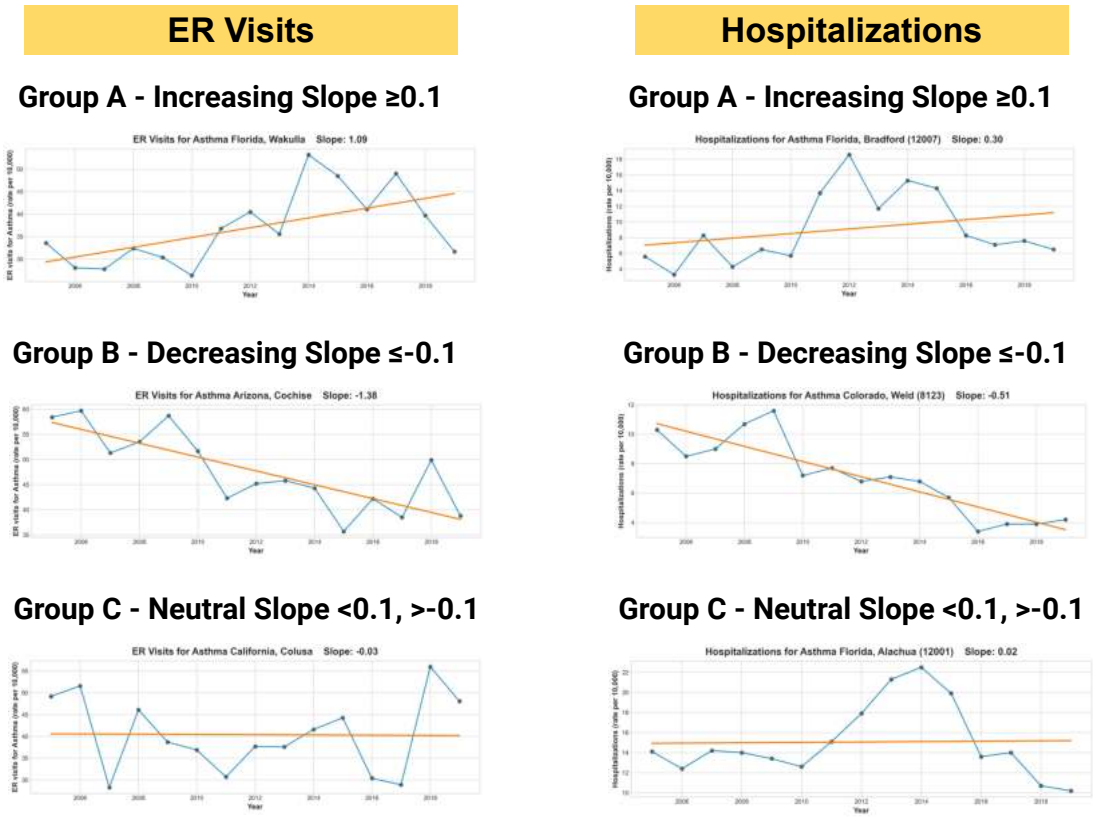

Figure 18: Categorization of Subgroups

QQ-plots generated for the ER visits groups show that groups A and B are close to normal with R^2 values of 0.994 and 0.996 respectively. For Group C, a log transformation was called for to normalize the data as the R^2 score of the data with a log transformation was higher at 0.995.

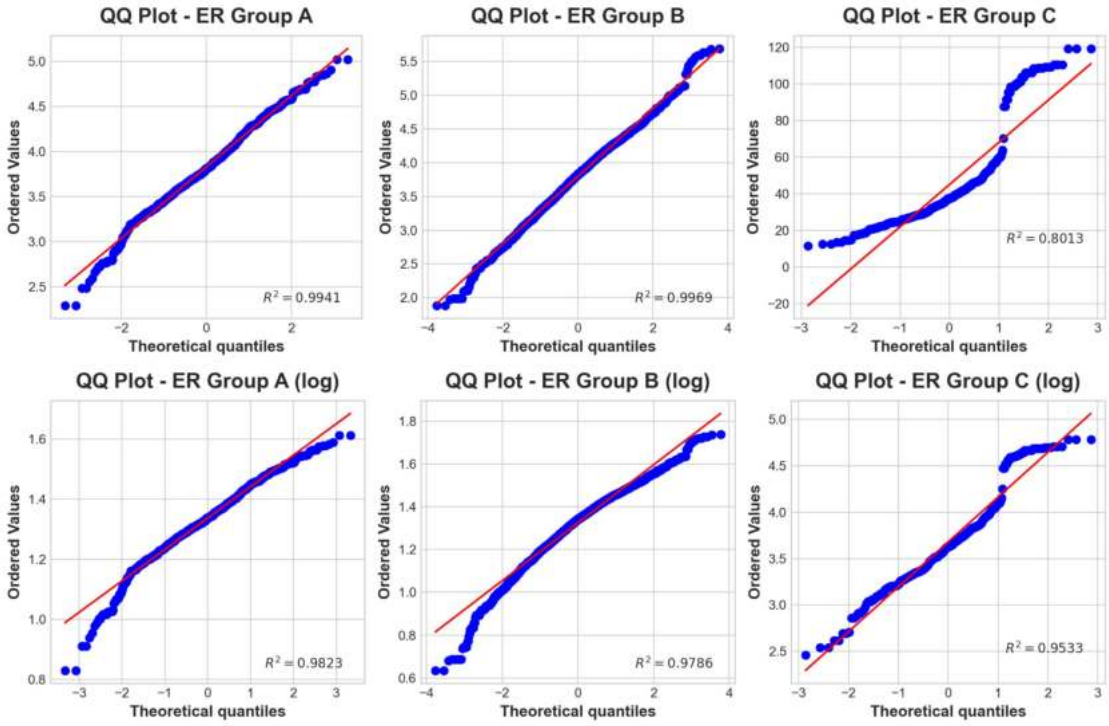

Figure 19: ER Visits QQ-plots

QQ-plots generated for the hospitalizations group B illustrated that a log transformation was necessary to normalize the data which generated a new improved R^2 score of 0.99.

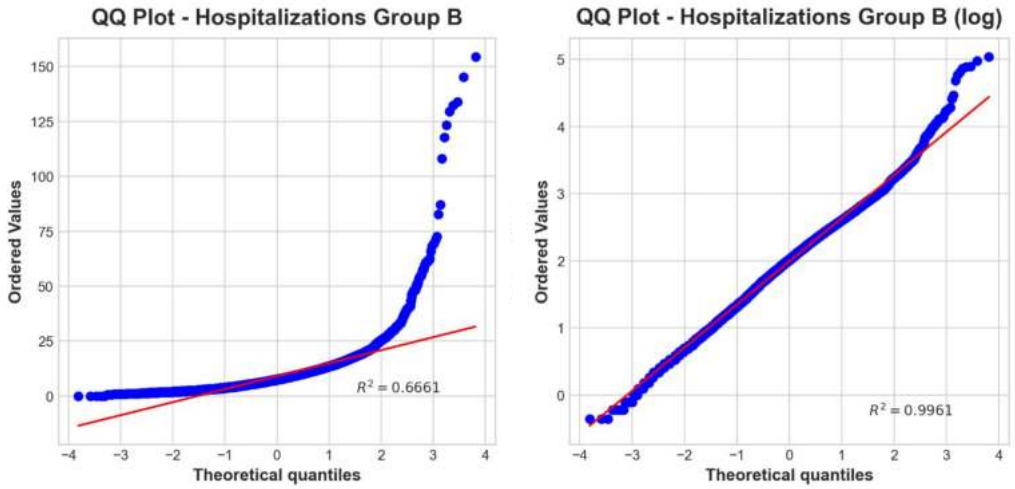

Figure 20: Hospitalizations QQ-plots

4.3 Random Forest

Table 2 summarized the performance metrics for the Random Forest models across different groups.

Random Forest Summary			
Group	R^2 Score	MSE	RMSE
ER Visits Group A	0.65	105.84	10.29
ER Visits Group B	0.68	119.77	10.94
ER Visits Group C	0.63	0.08	0.27
Hospitalizations Group B	0.25	0.26	0.51

Table 2: Performance metrics for Random Forest models

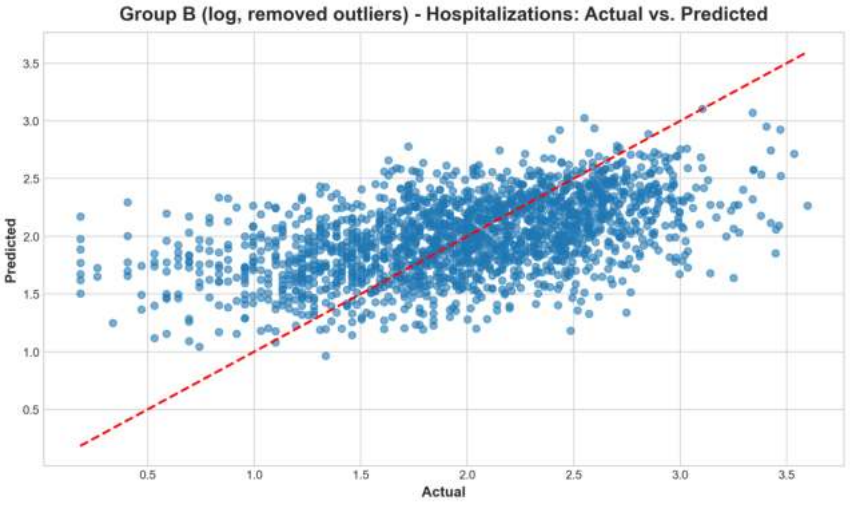

Figure 22: Random Forest for Hospitalizations Group B

The Random Forest models demonstrated varying levels of performance across the groups.

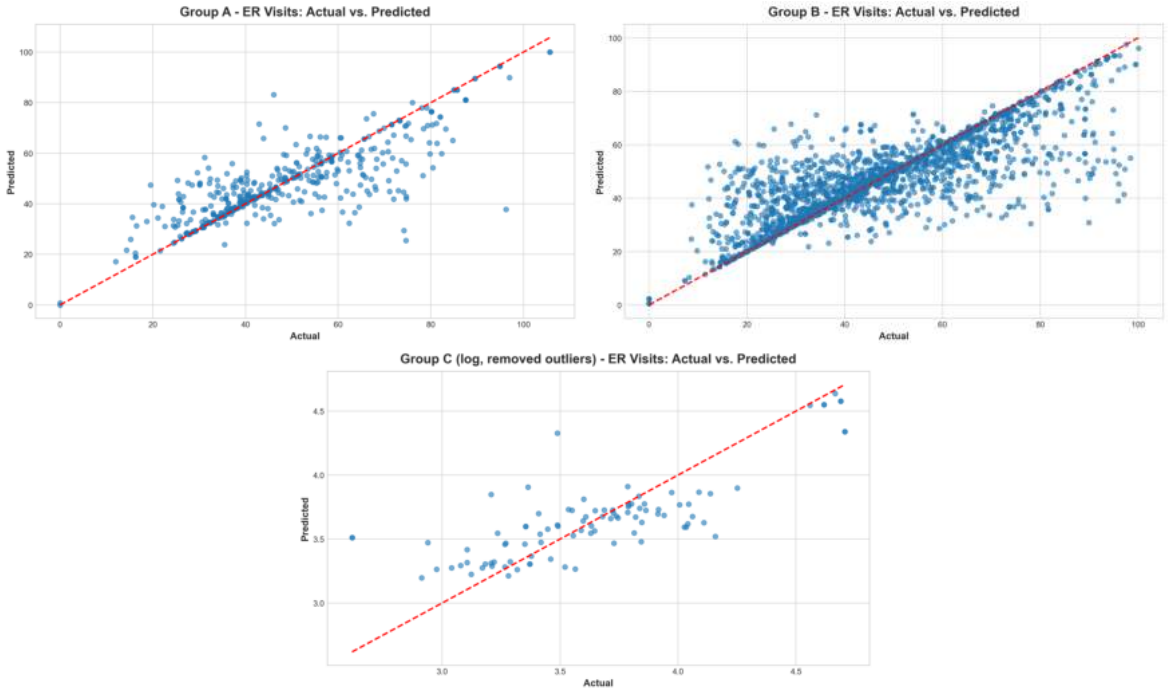

Figure 21: Random Forest Models for ER Visits Groups A, B, C

ER visits group B showed the highest performance with an R^2 score of 0.68. ER visits group A and C also performed relatively well with R^2 scores of 0.65 and 0.63, respectively. These models predicted asthma ER visits relatively accurately based on the R^2 scores and RMSE values. For Hospitalizations Group B, the performance was lower compared to groups in ER visits, with an R^2 score of 0.25 and an RMSE of 0.51.

4.4 SARIMAX

Table 3 contained performance and test values for all SARIMAX models and figures 19 and 20 contained plots of the generated models.

	ER Visits (A)	**ER Visits (B)**	**ER Visits (C)**	**Hosp (B)**
Ljung-Box (L1) (Q)	0.02	0.56	0.07	0.10
Prob(Q)	0.90	0.45	0.79	0.75
Heteroskedasticity (H)	0.22	4.01	3.06	10.64
Prob(H) (two-sided)	0.22	0.15	0.25	0.02
Jarque-Bera (JB)	0.60	1.01	0.78	3.24
Prob(JB)	0.74	0.60	0.68	0.20
Skew	0.44	-0.65	-0.02	-1.11
Kurtosis	2.51	3.11	1.84	3.81

Table 3: Performance metrics for SARIMAX Models

Notes on performance and test values: The Ljung-Box (L1) (Q) statistic measured autocorrelation in the residuals, while Prob(Q) represented the p-value associated with the Ljung-Box test, assessing the significance of autocorrelation. Heteroskedasticity (H) indicated whether the variance of the residuals changed over time, with Prob(H) (two-sided) being the p-value. Jarque-Bera (JB) and Prob(JB) were statistics that tested for normality in the residuals. Lastly, Skew measured the asymmetry of the distribution of residuals, while Kurtosis measured the "tailedness" of the distribution.

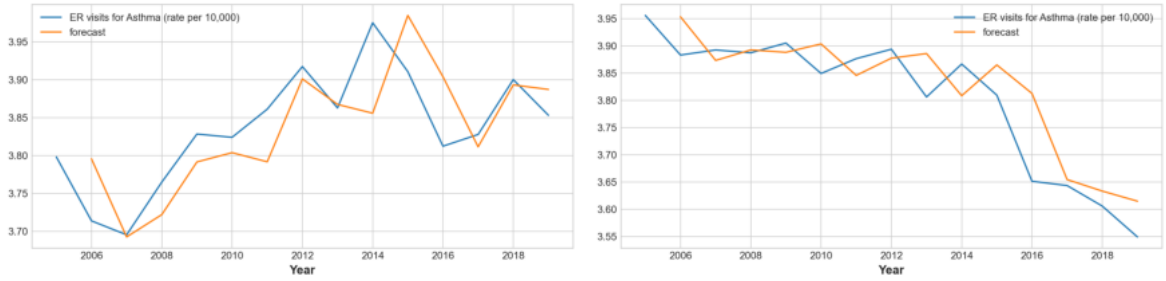

Figure 23: SARIMAX Models for ER Visits Group A and ER Visits Group B

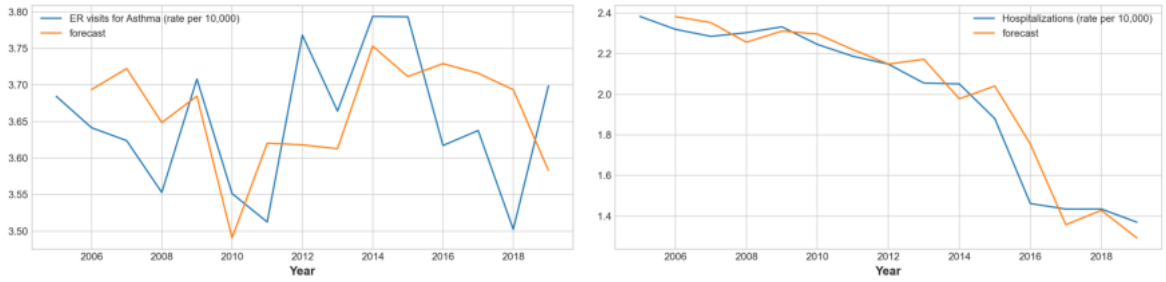

Figure 24: SARIMAX Models for ER Visits Group C and Hospitalizations Group B

The SARIMAX models were evaluated using various diagnostic statistics to assess their reliability. Notably, the Ljung-Box test indicated minimal autocorrelation in the residuals, with statistics ranging from 0.02 to 0.56 across models for ER visits and hospitalizations. Importantly, the associated p-values were above 0.05, indicating no significant autocorrelation. Furthermore, the Jarque-Bera test confirmed the normality of the residuals, with high p-values across all models, ranging from 0.60 to 3.24. However, regarding heteroskedasticity, the hospitalization group B model had considerable variability, with a statistically significant value of 10.64 in the heteroskedasticity test, while other models ranged from 0.22 to 4.01.

5 Discussion

5.1 Implications

Building upon previous studies that highlighted the significant impact of $PM_{2.5}$ and ozone on air quality and asthma [17, 4], our findings reinforce these relationships and further pinpoint them as crucial contributors to asthma exacerbations. Moreover, while our study highlighted the role of specific pollutants, such as $PM_{2.5}$ and ozone, we also recognized the broader influence of multiple air quality indices such as median AQI, max AQI, 90th percentile AQI, days with AQI, and moderate days. This broader approach to air quality assessment could offer more comprehensive insights for healthcare professionals and policymakers, enabling them to develop targeted interventions that address multiple facets of air quality rather than focusing solely on individual pollutants.

Gent et al, specifically demonstrated that $PM_{2.5}$ and ozone levels below the EPA standards of 0.12 ppm (1hr), 0.08 ppm (8hr), and $50\mu g/m^3$ (1hr), $150\mu g/m^3$ (8hr) respectively were associated with an increase of asthma symptoms [7]. This finding emphasized the limitations of current air quality standards and supported our argument for considering a wider range of varying air quality indices, not just specific pollutants.

The ethical dimension of our study centered on its potential public health implications. Identifying key pollutants and AQI related factors that have a significant impact on asthma exacerbartions enables healthcare professionals and policymakers to develop targeted interven-

tions. Such targeted efforts can not only ease the strain on the healthcare system but also potentially improve patients' quality of life.

Additionally, our findings resonate with the ongoing debate in literature. While some studies challenge the direct correlation between air pollution and asthma, arguing that even levels below standards can exacerbate symptoms [15], others like McConnell et al., demonstrate a significant environmental impact, particularly in the areas with higher ozone levels [11]. In all, this inconsistency highlights the need for targeted research to clarify the roles of specific pollutants, a point further emphasized by Buteau et al. [4].

5.2 Challenges

One challenge faced during this study was the inconsistency and incompleteness of the available data. Despite selecting years with comprehensive reported data, missing values and data gaps still remained an issue. This limitation, due to missing county values, could potentially introduce selection bias and limit the generalizability of our findings to some extent. Additionally, the absence of uniform geographic information across the counties contributed to weaker correlations and posed difficulties in drawing conclusions from our data analysis. This inconsistency was a significant factor in our decision to group the counties into three categories - increasing slopes, decreasing slopes, and neutral slopes - based on the trends in asthma-related hospitalizations and emergency room visits. This grouping strategy was employed as a method to mitigate the impact of these data gaps and to facilitate a more structured analysis of the trends observed. To enhance future research, it will be crucial to integrate more detailed and consistent data specific to our county demographics, which could provider deeper insights into the environmental factors affecting asthma outcomes. Furthermore, implementing methods to better handle or impute missing data could also improve the overall quality and reliability of our findings.

5.3 Next Steps

In our ongoing efforts to enhance the predictive capabilities of our machine learning models, we plan to concentrate on refining the algorithms and improving data utilization. This will involve several targeted strategies aimed at optimizing our current predictive models based on Random Forest and SARIMAX.

1. **Data Enrichment and Feature Engineering:** We aim to enrich our dataset by incorporating more granular data points and introducing additional variables that may influence asthma outcomes, such as weather conditions and more detailed pollutant concentration levels for each county. Feature engineering will be key in transforming raw data into more effective, predictive inputs that capture the complexities of environmental impacts on asthma.

2. **Model Tuning and Optimization:** By adjusting model parameters and experimenting with different configurations, we intend to find the most effective settings that increase the accuracy and reliability of our predictions. This will include hyperparameter tuning using methods such as grid search and cross-validation to systematically explore a range of configurations.

3. **Advanced Analytical Techniques:** We will explore the application of more complex machine learning techniques and newer algorithms that could offer better performance. Techniques such as ensemble methods and deep learning may provide new insights with improved prediction accuracy.

4. **Validation and Testing:** Enhancing our validation framework to include a cross-validation setup will help ensure that our models perform well across different subsets of data. Additionally, this will help prevent overfitting and will verify that our models can generalize well to new, unseen data.

By focusing on these areas, we hope to significantly enhance the sophistication and effectiveness of our models. This approach will help us to better understand and predict how environmental factors influence asthma, leading to more informed public health strategies and interventions.

6 Conclusions

The data analysis conducted in this study revealed intricate relationships between air quality indicators and asthma outcomes, as illustrated across various visual representations and machine learning models. While there is a discernible correlation between pollutants like $PM_{2.5}$, NO_2, and median AQI values with asthma exacerbations, the observed relationships exhibit modest strength, indicating the influence of specific air quality on respiratory health. Our findings suggested that while air quality is an influential factor, its impact on asthma must be considered alongside other socio-economic and healthcare accessibility variables. The observed trends indicate that despite the complex and varied nature of environmental influences, machine learning models like Random Forest and SARIMAX provide valuable insights into how specific pollutants can be used in generating ML models. Therefore, future interventions aimed at mitigating asthma exacerbations should adopt a multifaceted approach, focusing not only on improving air quality but also on enhancing healthcare delivery and addressing social determinants of health. This holistic approach could lead to a more significant decrease in asthma-related hospitalizations and ER visits, as we strive to understand and counteract the impacts of air pollution on public health. The groundwork laid by this research supports the potential of machine learning as an invaluable tool for informing public health strategies and developing community-targeted approaches to address the persistent issue of asthma.

7 Code Appendix

The code for this project can be found in the following GitHub repository: GitHub Repository

References

[1] *A Look Back: Ozone and PM in 2020*. URL: https://epa.maps.arcgis.com/apps/Cascade/index.html?appid=9f72fb0d74be4d398e794d1231f24ef0.

[2] Environmental Protection Agency. *AirData website File Download page*. en. Data & Tools. 2023. URL: https://aqs.epa.gov/aqsweb/airdata/download_files.html (visited on 01/20/2024).

[3] *Air Pollution: Current and Future Challenges*. Oct. 2023. URL: https://www.epa.gov/clean-air-act-overview/air-pollution-current-and-future-challenges.

[4] Stéphane Buteau et al. "Air pollution from industries and asthma onset in childhood: A population-based birth cohort study using dispersion modeling". In: *Environmental Research* 185 (June 2020), p. 109180. ISSN: 0013-9351. DOI: 10.1016/j.envres.2020.109180. URL: https://www.sciencedirect.com/science/article/pii/S0013935120300724 (visited on 01/23/2024).

[5] CDC. *Asthma Data Visualizations — CDC*. en-us. Jan. 2023. URL: https://www.cdc.gov/asthma/data-visualizations/default.htm (visited on 02/11/2024).

[6] Oladunni Enilari and Sumita Sinha. "The Global Impact of Asthma in Adult Populations". In: *Annals of Global Health* 85.1 (Jan. 2019), p. 2. ISSN: 2214-9996. DOI: 10.5334/aogh.2412. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7052341/ (visited on 02/11/2024).

[7] Janneane F. Gent et al. "Association of Low-Level Ozone and Fine Particles With Respiratory Symptoms in Children With Asthma". In: *JAMA* 290.14 (Oct. 2003), pp. 1859–1867. ISSN: 0098-7484. DOI: 10.1001/jama.290.14.1859. URL: https://doi.org/10.1001/jama.290.14.1859 (visited on 02/23/2024).

[8] IBM. *What is Exploratory Data Analysis? — IBM*. en-us. 2024. URL: https://www.ibm.com/topics/exploratory-data-analysis (visited on 02/19/2024).

[9] Kelly Livingston and Stephanie Ebbs. *EPA announces new air quality standards for particulate matter, citing health risks*. en. Feb. 2024. URL: https://abcnews.go.com/Politics/epa-announces-new-air-quality-standards-particulate-matter/story?id=107003457 (visited on 02/08/2024).

[10] Ioannis Manisalidis et al. "Environmental and Health Impacts of Air Pollution: A Review". In: 8.14 (Feb. 2020). DOI: 10.3389/fpubh.2020.00014.

[11] Rob McConnell et al. "Asthma in exercising children exposed to ozone: a cohort study". en. In: *The Lancet* 359.9304 (Feb. 2002), pp. 386–391. ISSN: 01406736. DOI: 10.1016/S0140-6736(02)07597-9. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673602075979 (visited on 02/23/2024).

[12] Manuel Méndez, Mercedes G. Merayo, and Manuel Núñez. "Machine learning algorithms to forecast air quality: a survey". en. In: *Artificial Intelligence Review* 56.9 (Sept. 2023), pp. 10031–10066. ISSN: 1573-7462. DOI: 10.1007/s10462-023-10424-4. URL: https://doi.org/10.1007/s10462-023-10424-4 (visited on 01/20/2024).

[13] James W. Mims. "Asthma: definitions and pathophysiology". en. In: *International Forum of Allergy & Rhinology* 5.S1 (Sept. 2015). ISSN: 2042-6976, 2042-6984. DOI: 10.1002/alr.21609. URL: https://onlinelibrary.wiley.com/doi/10.1002/alr.21609 (visited on 02/01/2024).

[14] Lucas M. Neas et al. "Association of Indoor Nitrogen Dioxide with Respiratory Symptoms and Pulmonary Function in Children". en. In: *American Journal of Epidemiology* 134.2 (July 1991), pp. 204–219. ISSN: 1476-6256, 0002-9262. DOI: 10.1093/oxfordjournals.aje.a116073. URL: https://academic.oup.com/aje/article/101131/Association (visited on 02/23/2024).

[15] David B. Peden. "The epidemiology and genetics of asthma risk associated with air pollution". en. In: *Journal of Allergy and Clinical Immunology* 115.2 (Feb. 2005), pp. 213–219. ISSN: 00916749. DOI: 10.1016/j.jaci.2004.12.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0091674904032191 (visited on 02/23/2024).

[16] Howard David Pettigrew et al. "The Clinical Definitions of Asthma". en. In: *Bronchial Asthma*. Ed. by M. Eric Gershwin and Timothy E. Albertson. New York, NY: Springer New York, 2012, pp. 3–18. ISBN: 9781441968357 9781441968364. DOI: 10.1007/978-1-4419-6836-4_1. URL: https://link.springer.com/10.1007/978-1-4419-6836-4_1 (visited on 02/01/2024).

[17] Mary Prunicki et al. "Exposure to NO2, CO, and PM2.5 is linked to regional DNA methylation differences in asthma". In: *Clinical Epigenetics* 10.1 (Jan. 2018), p. 2. ISSN: 1868-7083. DOI: 10.1186/s13148-017-0433-4. URL: https://doi.org/10.1186/s13148-017-0433-4 (visited on 01/30/2024).

[18] Karlapudi Saikiran et al. "Prediction of Air Quality Index Using Supervised Machine Learning Algorithms". In: (2021), pp. 1–4. DOI: 10.1109/ACCESS51619.2021.9563323. URL: https://ieeexplore.ieee.org/abstract/document/9563323/authors#authors.

[19] *State of the Air*. 2024. URL: https://www.lung.org/research/sota/key-findings.

[20] Padmaja Subbarao, Piush J. Mandhane, and Michael R. Sears. "Asthma: epidemiology, etiology and risk factors". In: 181.9 (Oct. 2009), E181–E190. DOI: https://doi.org/10.1503/cmaj.080612.

[21] Simon F. Thomsen. "Genetics of asthma: an introduction for the clinician". en. In: *European Clinical Respiratory Journal* 2.1 (Jan. 2015), p. 24643. ISSN: 2001-8525. DOI: 10.3402/ecrj.v2.24643. URL: https://www.tandfonline.com/doi/full/10.3402/ecrj.v2.24643 (visited on 02/07/2024).

[22] Angelica I. Tiotiu et al. "Impact of Air Pollution on Asthma Outcomes". In: *International Journal of Environmental Research and Public Health* 17.17 (Sept. 2020), p. 6212. ISSN: 1661-7827. DOI: 10.3390/ijerph17176212. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7503605/ (visited on 01/23/2024).

[23] OAR US EPA. *NAAQS Table*. en. Other Policies and Guidance. Apr. 2014. URL: https://www.epa.gov/criteria-air-pollutants/naaqs-table (visited on 02/01/2024).