The Relations Between Common Air Pollutants & Asthma Emergency Calls in NYC

Z. Mokhtarzadeh, S. Prochowski

CSCI-UA 476: Processing Big Data for Analytics Final Project

The Relations Between Common Air Pollutants & Asthma Emergency Calls in NYC

Zachary Mokhtarzadeh (zem240) & Sabina Prochowski (sp5642)

Abstract: Air pollution is a significant global health concern that has been linked to several diverse health effects like asthma exacerbation. Thus, studying the association between air pollutants and asthma health outcomes in areas like New York City (NYC) where the population density and urban infrastructure contribute to high levels of air pollution, is crucial for informing public health policies and interventions. This application investigated the relationship between air pollution and asthma emergency department (ED) visits in NYC neighborhoods. The purpose of this study was to identify which air pollutants (sulfur dioxide, nitrogen dioxide, and ozone) had the strongest association with asthma health outcomes. The study used a quantitative approach of correlations, percent variance explained, and a regression model to analyze the air quality data from NYC Open Data with the asthma emergency department visits data from the Environment and Health Data Portal. From 2009 to 2014, ozone had the highest association with asthma emergency department visits, followed by nitrogen dioxide and sulfur dioxide, based on slope calculations. However, these findings were inconsistent with the correlations found in 2014 alone, suggesting that the relationship between pollutants and asthma visits is complex and varies over time. It is important to note that these findings do not control for other factors not available in the dataset such as genetic predisposition and other environmental factors, which can also be considered asthma triggers. Regardless, the findings still suggest that there is an urgent need to reduce air pollution in order to improve health outcomes, especially in specific NYC neighborhoods, to be discussed below.

Keywords: Air Pollutants, Asthma Health Outcomes, New York City, Emergency Department Visits.

1 Introduction

Ozone (O_3), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2) are identified as three of the six pollutants classified as “criteria” air pollutants (CDC, 2022). These are the leading pollutants that can have a profound toxicological impact on human health and the environment. Long and short term exposure of these pollutants has been confirmed to account for playing a major role in the incidence and progression of several respiratory and cardiovascular diseases, including asthma (Adel Ghorani-Azam et al., 2016). In our study, to analyze the relationship between these three pollutants and NYC asthma emergency department (ED) visits, we used several quantitative methods on two robust and reliable publicly accessible datasets, which ensured the quality and validity of our research. The methods utilized in our study can serve as a model for future research in NYC or other urban areas with comparable high levels of air pollution. Our findings contribute to a growing pool of evidence on the negative effects of air pollution and emphasizes the significance of the need for ongoing research in this area.

Ozone (O_3), a colorless gas that is the most ubiquitous air pollutant in the United States, is given off as not only a product of natural chemical reactions between nitrogen oxides and volatile organic compounds (VOCs), but also due to human activities (Adel Ghorani-Azam et al., 2016). Some of these human activities include but are not limited to emissions from vehicles, factories, fossil fuels, evaporation of paints, and more. EPA has confirmed that short-term ozone exposure is linked to respiratory morbidity and lung function decrements, i.e., individuals may not be able to inhale total lung capacity.[1] Moreover, long-term exposure to ozone is associated with asthma aggravation, and though the pathogenesis of asthma is not completely identified, evidence points to it being likely to be one of the causes of asthma development. A prospective study on a cohort of nonsmokers to evaluate the association between long-term ozone exposure and development of adult-onset asthma supports this as over a 15 year period, 3.2% of males and 4.3% of females reported new asthma diagnoses (William F. McDonnell et al., 1999).

Sulfur dioxide (SO_2) is another gaseous air pollutant that is produced when sulfur-containing fuel such as coal, oil or diesel is burned. According to the American Lung Association, major contributors to sulfur dioxide emissions are diesel engines, electricity generation, and other industrial processes like petroleum refining and metal processing, to name a few. Research on sulfur dioxide's contribution to asthma has had a relative decrease in attention in recent years (in comparison to pollutant, ozone which has more evidence supporting its link to causing direct air constriction and asthma exacerbation) as a PubMed article revealed a relatively small number of inhaled SO_2 particles over the past decade (Anita L. Reno et al., 2015). Hence, SO_2 is worth researching further as we have studied that the highest correlation resulted in between SO_2 and asthma ED visits in 2014 in comparison to other pollutants.

Nitrogen dioxide (NO_2) is a highly reactive gas that primarily is emitted from the burning of fuel. This implies emissions majorly come from vehicles, power plants, and off-road equipment. Breathing air that has a high NO_2 concentration can aggravate respiratory conditions like asthma, causing respiratory symptoms that contribute to higher rates of hospital admissions and emergency room visits according to EPA. Though our study is conducted on emergency department visits of adults in NYC, a comparable study in Barcelona shares that NO_2 was associated with mortality for causes of death in asthmatic patients with more than one emergency room admission for asthma, which is clearly a significant finding (J. Sunyer et al., 2002).[2]

Consequently, the potential health effects of these three air pollutants have been an important topic of research. In this paper, we analyzed the concentrations of three common air pollutants measured in mean parts per billion (ppb) with asthma emergency department visit rates in 2014 in addition to a regression analysis of the same two variables to find relationships between each neighborhood and pollutant from 2009 to 2014. The purpose of this was to better understand the extent of the impact of air pollution on human health in NYC and contribute research to the ongoing efforts to improve air quality. Before this study, our knowledge was quite limited other than what we already knew from past physical science courses and general environmental concerns regarding air pollution.

This project aims to provide valuable insights into the relationship of air pollution and asthma emergency visits in an effort to highlight the importance of continued air quality monitoring and management. This study had the potential to advance our knowledge in comprehending the complex relationship between the two variables. By analyzing the data of three major air

[1] This study by EPA explains how ozone reacts in the respiratory tract and how response may vary among individuals, but common symptoms include cough, throat irritation, pain / burning / discomfort during breathing, and chest tightness (EPA, 2023).

[2] A study on patients in Barcelona who have passed away in 1985-1995 who had visited the emergency department of one of four hospitals in the city for asthma related reasons, had concluded nitrogen dioxide and ozone has exacerbated severe asthma and potentially contributed to death among those asthmatic patients.

pollutants across multiple neighborhoods in NYC, we identified notable trends in pollution levels. Additionally, our use of regression analysis provides a more in-depth understanding of the relationship between pollutants and asthma emergency call rates over time. We intended our findings to have the potential to advance our knowledge in identifying which pollutants have a greater impact on human health. Then, we highlighted the neighborhoods which are affected most (in terms of their regression relationship with the number of asthma ED visits) by each pollutant. With further research and additional case studies, we can learn more about why we are observing certain neighborhoods with higher levels of air pollutants that are contributing to more asthma department visits.

The main contributions of this paper as summarized as follows:

- We provide an analysis of the concentration of three common air pollutants (O_3, SO_2, and NO_2) in multiple NYC neighborhoods and their correlation (along with percent variance explained) with asthma ED visits.
- We highlight the top 3 neighborhoods most affected by each pollutant based on our regression analysis.
- We discuss the differences in our findings of the contributing air pollutants found in our correlation analytic and regression analytic.

To drive a deeper investigation into how air pollution is affecting asthma emergency visit rates, we were motivated to build data analytics from two datasets, one containing data about air quality (records of different air pollutants measured in NYC neighborhoods across several years) in certain neighborhoods in NYC and the other containing asthma emergency department visits (records of emergency department visits in NYC neighborhoods along with standardized measures of those visits based on age groups and population also across the same years as the other dataset) among adults in NYC neighborhoods. We stored those two datasets in the Hadoop Distributed File System (HDFS), and after, performed multiple cleaning jobs on each of our two datasets. We removed any rows with NULL values in our datasets as well as columns of our datasets that were not directly relevant to our study. We also ensured to remove commas and other obstructions that would get in the way of conducting methods on numerical data. Additionally, Zachary reformatted the date columns for the air quality dataset to enable us to use this shared column between the datasets in joining the two datasets (in addition to joining on with another variable which was the neighborhood geoID, to be further discussed in upcoming sections). While our datasets contain data from 2005 to 2018, we decided to use data for the years 2009 to 2014 as these were the only years that provided the most relevant data for the air pollutants we were studying; further explanation for our reasoning behind this can be found in the analytic stages section below.

Upon the completion of cleaning and formatting the data, we profiled each of the datasets to make sure our datasets were correctly cleaned so that we could maintain a quality standard of goodness and ensure they were ready for a join. Zachary conducted all of his cleaning jobs with MapReduce for the air quality dataset while Sabina also conducted a clean in Hive in addition to MapReduce for the asthma emergency department visits dataset. Following that, once we both ingested our data into Hive, we joined our two datasets on the year column and geoID, which is a unique NYC neighborhood identifier. After a successful joining of the datasets, we profiled our data once again in Hive to make sure our data maintained the same quality. There, we developed and performed our analytics of correlation, percent variance, and regression using HiveQL. After assessing our correlation between data values and slopes, we checked for correctness and assessed our findings against research and related work. To better understand our data pipeline, here is our Data Flow Diagram displaying the steps we took from start to finish in performing our analytics.

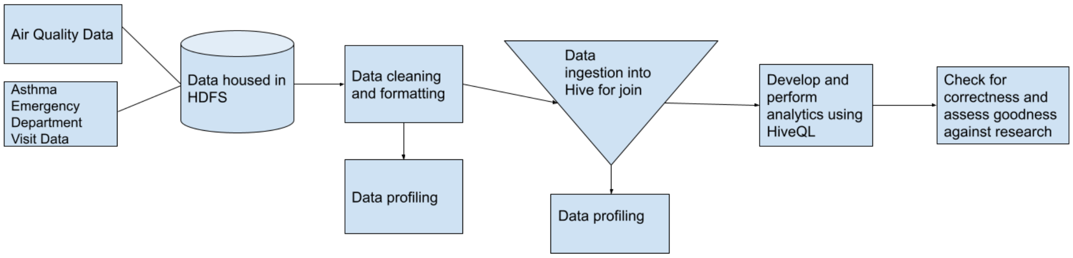

Air Quality Data
Asthma Emergency Department Visit Data
Data housed in HDFS
Data cleaning and formatting
Data profiling
Data ingestion into Hive for join
Data profiling
Develop and perform analytics using HiveQL
Check for correctness and assess goodness against research

Figure 1: Data Flow Diagram

2 Motivation

With the increasing urbanization and industrialization of society, poor air quality has been an increasingly pressing global concern that demands greater attention. There is an urgency to address how air pollution affects human health, especially in highly populated urban areas like NYC, where air quality is compromised from traffic congestion, industrial emissions, and other human activities detrimental to the environment. Thus, we chose to investigate the air quality where we live, in New York City, to determine if there are any outstanding associations between air pollutants and asthma emergency department visits in NYC neighborhoods. This topic is worth investigating as air pollution is a major environmental and public health issue, with room for further research on what areas it contributes most to and why this is happening as well. This research can provide valuable evidence and recommendations on what areas are in need of more attention. That way, policymakers and public health professionals can prioritize allocating interventions to the most affected populations. This is relevant as pollution is unequally distributed across different locations across the United States. In fact, neighborhoods that were identified as the most and least polluted in 1981 continue to be the same neighborhoods to be the most and least polluted more than 30 years later in 2016 (Johnathan Colmer et al., 2020).[3]

Despite past measures that had been implemented in order to improve air quality, socioeconomically disadvantaged areas are still more likely to have had more exposure to air pollution at any given time. According to a more recent study conducted by Nyc.gov, though there is a clear pattern in the United States that the worst air quality tends to be located in the poorest neighborhoods, the most polluted neighborhoods in NYC, Midtown and Lower

[3] This study observed how fine particulate air pollution is closely correlated with disparities between distinct demographic and socioeconomic groups over the span of 36 years, taking account over 65,000 U.S. census tracts. Although there has been a substantial decrease in absolute disparities, relative disparities between the different areas remain, indicating a need for continued efforts to address this issue.

Manhattan, are actually one of the wealthiest neighborhoods in the city.[4] However, this does not undermine the fact that in NYC and anywhere else in the United States, it still holds true that adults living in high-poverty neighborhoods experience more air pollution related health issues. Thus, this is worth studying as it is crucial to attend to parts of the city where health outcomes are the poorest. This is because socioeconomically disadvantaged neighborhoods have access to fewer resources like healthcare which can imply that they may not be able to protect themselves from air pollution's adverse effects (Nyc.gov, 2022). Also, these neighborhoods often have higher baseline rates of health conditions, including those induced by air pollution, hence; populations in these neighborhoods are more likely to have existing health problems that are either worsened by air pollution or contribute to asthma exacerbations and / or development (Nyc.gov, 2022).

The motivation behind our research is thus to observe the potential harmful effects of air pollution on human health, particularly with regard to asthma. The detrimental health effects of exposure to air pollutants have been well-documented, including asthma and other respiratory problems. The high incidence of asthma emergency department visits in New York City, coupled with the city's long-standing problems with air quality, made this topic a pressing issue for us to investigate. Our team was motivated to study the relationship between air pollution and asthma emergency department visits in NYC to better understand how air pollution can directly affect human health. This is precisely why we chose to investigate this topic as we believed it would be beneficial to use data analytics to help understand real-world problems. Additionally, we wanted to explore the potential of using data analytics to inform potential decisions that are in process of being taken that relate to reducing air pollution and improving public health. By understanding this essential relationship, we hope to provide insights that can ultimately help inform decision-makers in this space. Ultimately, our goal is to contribute to the ongoing efforts to understand how air pollution can change public health outcomes and to use data analytics to help make the effects clear.

Sabina's interest in this topic stems from her personal experience of being diagnosed with adult-onset asthma earlier this year. Since she was born and raised in NYC, she was particularly interested in looking into the contributions of air pollutants in asthma health outcomes over time. She was also curious to find out more about how her neighborhood was affected by each air pollutant as her neighborhood is known to have poor air quality. Through her research, Sabina aimed to raise awareness about the detrimental effects of air pollution on public health and contribute to ongoing efforts to improve air quality in her hometown.

Zachary's interest in this topic comes from an experience of having a large amount of his family members having asthma, including his mother. Since moving to NYC, Zachary has become more and more curious about how air quality can affect someone as the air in NYC is much different than his home city's. Through his research, Zachary wants to use data to help people understand how air pollution can directly affect someone's health.

[4] This study uses the same data we use in our study from the New York City Community Air Survey, but in the years of 2015-2017 with a focus on an air pollutant, PM2.5, we did not use (since it used a different measure of air pollutant concentration) to analyze the neighborhood-by-neighborhood differences in air quality and highlights neighborhoods with the worst air quality (Nyc.gov, 2022).

3 Related Work

3.1 Air Pollution in Neighborhoods with Lower SES

It is commonly seen that specific demographics in the United States, such as Blacks, Asians, Hispanics, Latinos, and low-income populations, are exposed to elevated levels of dangerous fine particulate air pollution (PM2.5) compared to other groups.[5] Moreover, previous research highlights the disparities in exposure to air pollution among different population / income groups and further indicates that racial and ethnic minorities as well as lower-income communities are at a greater risk of premature death due to PM2.5 exposure (Nicole Rura et al., 2022). Thus, this information made us look into if the neighborhoods we found to have the top contributing air pollutants would be neighborhoods consisting of low income or communities of color. However, after further research, there are studies such as this one by Anjum Hajat et al. in 2013 that suggest that the association between air pollution and socioeconomic status (SES) can differ depending on the metropolitan area being studied. In New York City, researchers found a positive association between pollutant concentrations and SES. This means that neighborhoods with higher SES have higher levels of pollution, which is contrary to what is often shown in research that suggests low-income neighborhoods are disproportionately affected by pollution in the U.S. This research highlights the importance of considering SES as a confounding variable in studies of air pollution and health, and suggests that efforts to reduce pollution should not only target low-income neighborhoods, but also address the sources of pollution in higher-income areas. Interestingly enough, our results show that the top neighborhoods with the highest associations are in fact a mixture of neighborhoods of varying SES.

3.2 Promising Outcomes in Health Impact Assessments

There are promising outcomes in implementing health impact assessments to substantially improve disparities in public health benefit. Nyc.gov used data modeling tools to find that a health impact assessment referred to as the “80 x 50” policy[6] that estimates the effects of reducing 80% of greenhouse emissions by 2040 would reduce the fine particulate matter pollutant throughout NYC regardless of neighborhood poverty level. Nyc.gov further explains that although reductions in PM2.5 would be similar across neighborhoods, high poverty neighborhoods would benefit most from improvements in air quality. Hence, local efforts to improve air pollution like reducing vehicle traffic or incentivizing residents to replace current efforts that generally contribute to pollutant emissions like converting to cleaner heating oil should be prioritized in neighborhoods with the highest rates of poverty (Nyc.gov, 2022). Hopefully, our research can inspire other researchers and professionals in the field to contribute to a further study to detect if this similar process would also translate to the three pollutants in our study.

3.3 Higher Ozone Concentrations According to Time of Year

Our results from our regression analysis showed that O_3 had the highest slope calculations, followed by NO_2, and then SO_2. The higher the slope indicates a more significant impact of a

[5] This study assessed 17 years’ worth of demographic data for disparities in air pollution exposure among racial/ethnic and income groups across the U.S. population (Abdulrahman Jbaily et al., 2022).

[6] This policy is an effort to confront climate change in NYC and other regions around the world. It essentially encourages transitioning to cleaner energy sources and decreasing the use of fossil fuels in transportation, infrastructure, and more.

pollutant. The slopes were found between each pollutant's mean parts per billion (ppb) value and asthma department visit for each neighborhood in 2009 to 2014. The findings from our slope calculations share consistencies with prior research. An article published by Pediatric Research states that "it was impossible to disentangle the true impact of summer O_3 concentration on summer urgent asthma visits" (Lovinsky-Desir, 2018). This article shared similar findings to us, noting that O_3, NO_2, and SO_2, greatly contribute to asthma visits, but they emphasize that O_3 is much more strongly present in asthma attacks during summer concentrations. Hence, it is not much of a surprise that air pollution affects asthma visit rates, but it is important to note when and what pollutants can contribute to asthma attacks / hospitalizations.

3.4 Further Research Supporting Ozone as a Asthma Contributor in NYC

Although our research uses data on adults 18 years or older, the following research conducted on children is still relevant as it investigates the impact of pollution on respiratory health in New York. This study done by Shao Lin et al. in 2008 investigated the impact of chronic exposure to high ozone levels on childhood asthma admissions in New York State. They accomplished this by following a birth cohort in New York State from 1995 to 1999 and identified births and asthma admissions through the New York State Integrated Child Health Information System and joined that data with ambient ozone data from the New York State Department of Environmental Conversation. Their results showed that asthma admissions were strongly associated with increased ozone levels for all chronic exposure indicators. They found stronger associations amongst younger children, low socio demographic groups, and New York City residents. Their research shows that exposure to ambient ozone can increase the risk of asthma admissions among children and lower SES groups.

3.5 Similar Study Conducted in Another Urban Area with Ozone

Another relevant study on asthma and emergency department visits, not in NYC, but in another urban area, Seattle was conducted by Therese F. Mar et al. in 2009. Their objective was to determine whether ozone exposure in Seattle is associated with increased use of hospital emergency departments. They used hospital data on daily asthma cases for all ages that were obtained between 1998 and 2002. They also obtained ozone and fine particulate matter (PM2.5) data from local air agencies. Similar to our analysis, regression models were used to understand the association between asthma visits to emergency departments and air pollutants. Through their research it has been made clear that breathing in ozone can make asthma worse, especially for children in their area. They found that high levels of ozone were linked to more trips to the emergency department for both kids and adults, which aligns with our research quite closely.

3.6 Similar Study on Our 3 Air Pollutants' Effects on Asthma ED Visits

This following study aligns best with our work as it studies the same three air pollutants according to asthma ED visits. In 2021, Xue-yan Zheng et al. searched through electronic databases to retrieve studies that investigated the risk of asthma-related emergency room visits and hospital admissions associated with short-term exposure to O_3, NO_2, or SO_2. They conducted a sensitivity analysis by excluding the studies with a high risk of bias. Through their research they had found out that short-term exposure to daily O_3, NO_2, and SO_2 was associated with an increased risk of asthma exacerbation. The location of where the study was conducted is unfortunately not specified in the article. However, it is still relevant to our project as it is evident that all three pollutants have an impact on respiratory health, resulting in higher numbers of hospital admissions and ED visits for asthma related reasons.

4 Datasets

4.1 Air Quality Dataset

As mentioned previously, we used two datasets to accomplish our three analytics. One was an air quality dataset provided by NYC Open data that contained records of common pollutant exposures in New York City neighborhoods from 2005 to 2018. The size of the data was approximately 16.1K records. Below you will find a table representing the dataset.

Field Name	Data Type	Brief Description
Unique ID	int	Identifies the record
Indicator ID	string	Identifier of the type of measured value depended on the time and area
Name	string	Name of the indicator
Measure	string	How it is measured
Measure Info	string	Information on the measurement
Geo Type Name	string	Geography type for neighborhoods
Geo Join ID	string	Identifier of the neighborhood geographic area
Geo Place Name	string	Name of neighborhood
Time Period	string	description of the time
Start_Date	string	Date_value for the start of the time period. Always a date value, and we will be using these dates over the Time Period ones
Data Value	decimal	The actual data value for the indicator
Message	string	notes to apply to the data

Table 1: Air Quality Dataset Representation

After cleaning the Air Quality dataset, we were left with the following fields: Unique ID, Indicator ID, Name, Measure, Measure Info, Geo Type Name, Geo Join ID, Geo Place Name, Start_Date (reformatted), and Data Value.

4.2 Asthma Emergency Department Visits Dataset

The second dataset was the asthma emergency department visits dataset provided by the Environment and Health Data Portal that contains records of asthma emergency department visits among adults in NYC neighborhoods from 2005 to 2018. It is a relatively small dataset, storing 580 records. Below you find a representation of the records of the dataset.

Field Name	Data Type	Brief Description
Time	int	The year the asthma emergency department visit took place
GeoType	string	Geography type
GeoID	int	Unique identifier acts as a geography ID for each entry
GeoRank	int	The geography ranking based on location in NYC (from 0 to 4)
Geography	string	Neighborhood
Age-adjusted rate per 10,000	decimal	Number of asthma-related emergency department visits among NYC adults, divided by the population in 5 year age group, quotients are multiplied by the proportion of the 2000 US population in each age group.
Estimated annual rate per 10,000	decimal	Number of asthma-related emergency department visits among NYC adults, divided by the population of adults; expressed as cases per 10,000 residents
Number	string (to be converted to int)	Number of asthma-related emergency department visits among NYC adults (18 years and older)

Table 2: Asthma Emergency Department Visits Dataset Representation

After cleaning this dataset, we were left with all fields besides GeoType.

5 Analytics Stages

5.1 Data Cleaning & Profiling

In order to maintain a frictionless and quality analytic, we had to clean our data and profile it in order to understand and organize our data. In the air quality dataset, we removed extraneous columns and pollutants we were not interested in pursuing. We removed the Time Period and message columns as we were only concerned with the year of our analytic rather than the seasonal differences. We also reformatted the Start_Date column in order to make it easier to read and work with our other data. As for profiling, we found distinct values of the Geo Type Name, Measure, Measure Info, and Start_Date columns. We also filtered through the data to rename pollutants to “SO_2”, “NO_2”, and “O_2” based on the distinct values we noted since names were not consistent, for example, with some having just “SO_2” or “sulfur dioxide” or other trailing spaces. This would ensure it would be easy to capture all of the data for each pollutant accurately and efficiently. It was important to understand the distinct values of those columns due to seeing the different ways data was collected. We had also performed some other statistical analysis that did not end up being meaningful to our research so we discluded it in the paper.

In both datasets, we removed rows with NULL values in our datasets as well as columns of our datasets that were not directly relevant to our study. We also ensured to remove commas and other obstructions that would get in the way of conducting methods on numerical data. We also replaced one of the geoIDs that was mistakenly given to two different neighborhoods (Bronx and New York City), which carried through in both datasets. To handle this, we replaced the geoID of New York City with 0 and kept the geoID of Bronx as 1 to prevent issues in our join and analysis. All of our cleaning was done primarily in MapReduce and Hive.

The cleaning specific to the asthma emergency department visits dataset removed rows with NULL values, removed the GeoRank column as it would not contribute to our analysis, and created another column from the standardized rate across age groups, which is the age-adjusted rate and as a new binary column, where the value is 1 if the age-adjusted rate per 10,000 is greater than the mean of that column Sabina found earlier for her profiling, and 0 otherwise. This could be a meaningful analysis because it could help identify areas that may require more attention and resources to address asthma-related health issues. Ultimately, we did not use this column in our final analysis as we decided to focus on two other variables with more numerical results rather than binary data.

As for profiling, Sabina found distinct counts of several column-like distinct "time" values to see how many years are included in this dataset and distinct GeoIDs because that was intended to be used for joining our datasets together. She then ran distinct "geography" as they are supposed to align with the GeoID. Sabina noticed that there were 47 distinct Geo IDs, but 48 different neighborhoods, so there was an inconsistency. To analyze this further, she then selected distinct "GeoId, Geography" and found that a GeoId of 1 was repeated, one for New York City as a whole and another for Bronx, so she decided to clean this as she mentioned above. She then found the means and medians of the numerical data columns, but found it did not make sense to find the mode for her dataset as the other columns are themselves, calculations or counts, so they are unique values that should not or would not provide any meaningful information to find the modes of. These averages and medians are not included here as they do not provide meaningful data either as they are very general overviews of the dataset and thus, do not contribute to the direction of this research topic.

5.2 Data Ingestion

After our data ingestion into Hive, since we had reformatted the years in the air quality dataset and then replaced one of the geoID that was mistakenly given to two different neighborhoods (in both datasets), we were able to easily join the two datasets on the year as well as the GeoID.

After joining the dataset, we also conducted further profiling for our analysis. We noticed that the dataset held a great amount of different air pollutants that also had varying measures of pollutant concentrations. Thus, we narrowed down the data after researching which were the most common air pollutants and decided to choose three of the top pollutants that also shared the same type of measurements of air pollutant concentration records to maintain consistency. Ultimately, we decided to only study the air pollutants that had data values measured in the same metrics, as mean parts per billion as these were most plentiful. In our profiling, we conducted a select distinct to analyze this with distinct pollutants, measure, and measure_info columns. The table below was part of our profiling process as it provides the counts of pollutant records per year. As one can see, since 2005, 2008, 2016, and 2018 all held at least one account of 0 records for one of the pollutants. Rows of those years were discarded in our final analysis as we wanted to ensure there were records for each pollutant in every year in order to have a more reliable regression analysis that would have a more fair distribution amongst the pollutants per each as year as to not leave any room for skewed results due to missing data.

Year	**NO_2 Count**	**O_3 Count**	**SO_2 Count**
2005	0	240	0
2008	232	0	116
2009	348	356	116
2010	348	116	116
2011	348	116	116
2012	348	356	116
2013	348	116	164
2014	232	116	116
2016	232	116	0
2018	348	116	0

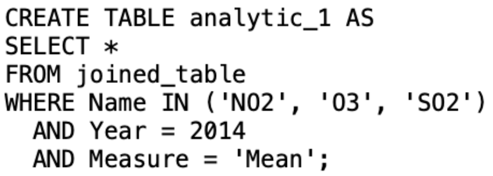
Table 3: Counts of Air Pollutant Records Per Year in Joined Table

Our final Hive table for our first two analytics that found the correlations and percent variance in the year 2014 had a total of 464 records. Our final Hive table prior to the creation of the regression table of years 2009 to 2014 had a total of 3,364 records.

5.3 Analytics & Discussion

Based on what we found in our profiling after the join table, we first started out with finding the correlation between each pollutant's mean ppb concentration with the standardized number of asthma-related ED visits based on population and age groups. Specifically, this column was originally calculated in the original dataset in the Environment and Health Data Portal as the "Number of asthma-related emergency department (ED) visits among NYC adults (18 years and older), divided by the population in 5 year age groups, using NYC DOHMH intercensal estimates; quotients are multiplied by the proportion of the 2000 US population in each age group. Age-adjusted rate is the sum of the weighted age-specific rates; expressed as cases per 10,000 residents." We decided to use this column over the other standardized column based only on age or raw number of visits as we believed it was more representative of the overall population's health status.

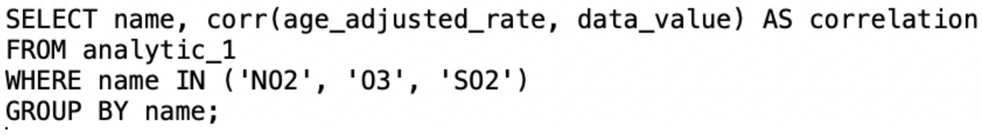
In this context, correlation refers to the strength of the relationship between the level of pollutants' air concentrations and asthma department visits whereas percent variance explained offers a measure of how much variability in the asthma ED visits can be accounted for by the variability in the mean concentrations of each pollutant. It is important to find both metrics because even if two variables are highly correlated, they can still explain only a small amount of the variability in each other, which would imply their relationship may not be meaningful or useful. Hence, with both calculations, we can gain a more comprehensive understanding of the relationship between the pollutants' concentrations and asthma ED visits.



```
CREATE TABLE analytic_1 AS
SELECT *
FROM joined_table
WHERE Name IN ('NO2', 'O3', 'SO2')
  AND Year = 2014
  AND Measure = 'Mean';
```

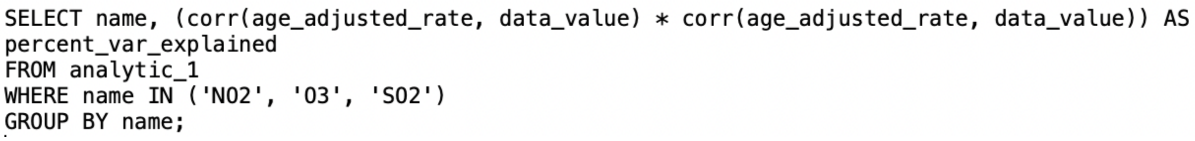
Figure 1: Hive Table Creation for Correlation & Percent Variance Explained Computation

The code necessary for this analytic was completed by first creating a Hive table that only consisted of the three air pollutants we were interested in, only keeping the pollutants that were measured as the mean parts per billion in 2014 (Figure 1). Then, we used select statements to compute correlations and percent variance explained grouped by the name of the pollutant (we previously renamed for this reason so we have clean analytics). Please refer to those select statements in Figures 2 and 3.



```
SELECT name, corr(age_adjusted_rate, data_value) AS correlation
FROM analytic_1
WHERE name IN ('NO2', 'O3', 'SO2')
GROUP BY name;
```

Figure 2: Select Statement for Correlation



```
SELECT name, (corr(age_adjusted_rate, data_value) * corr(age_adjusted_rate, data_value)) AS
percent_var_explained
FROM analytic_1
WHERE name IN ('NO2', 'O3', 'SO2')
GROUP BY name;
```

Figure 3: Select Statement for Percent Variance Explained

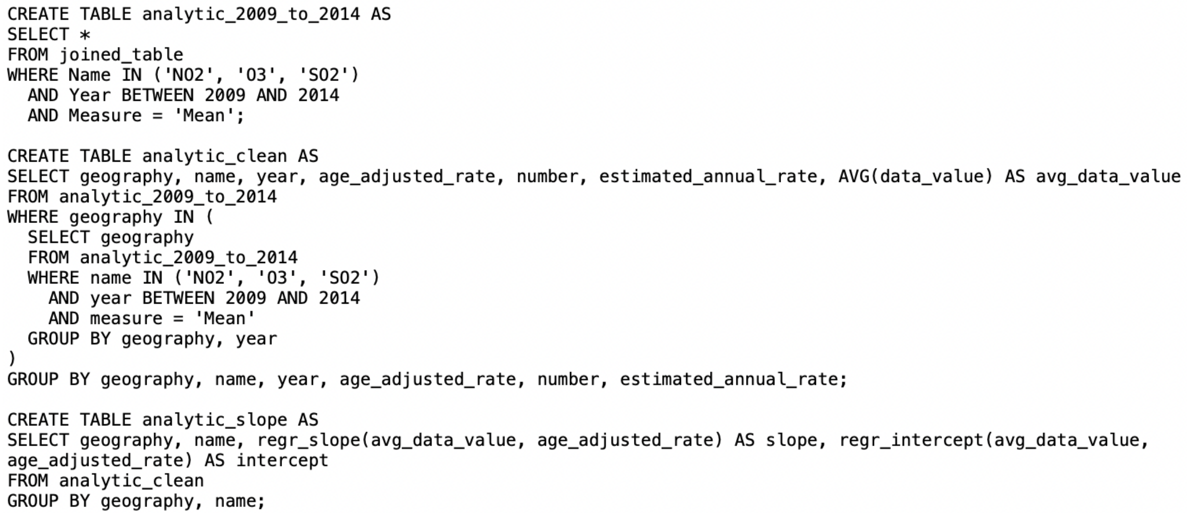
Table 4 provides the results. SO_2 had the strongest relationship with asthma-related ED visits among the three pollutants. The results show a weak positive correlation for NO_2 (0.137), a weak negative correlation for O_3 (-0.217), and a moderate positive correlation for SO_2 (0.295). As for percent variability, NO_2 explains only 1.87% of the variance in asthma-related ED visits, while

O_3 explains 4.70% and SO_2 explains 8.69%. Though O_3 had a negative correlation, it had a higher percent variance than NO_2, indicating that it is a bigger contributor to variation in asthma visits.

Pollutant	Correlation	Percent Variance Explained
NO_2	0.137	1.87%
O_3	-0.217	4.70%
SO_2	0.295	8.69%

Table 4: Correlations and Percent Variance Explained of Each Air Pollutant in 2014

The negative correlation between O_3 and asthma-related ED visits suggests that as O_3 levels increase, asthma-related ED visits are to decrease. This may seem counterintuitive, as O_3 is a known respiratory irritant and exposure to high levels can cause respiratory distress, chest pain, and coughing. We did not expect to see this result. However, it is important to note that correlation does not necessarily imply causation, and other factors may be at play. One important factor is an individual's unique susceptibility to air pollution. Some may be more sensitive to air pollution and may experience more severe health effects even at lower levels of exposure. Additionally, genetic predisposition can increase one's risk of experiencing negative health effects from air pollution. Lastly, this is also a consensus of all neighborhoods, which may skew results due to outliers or neighborhoods that have much air pollutant exposure. Also, social determinants of health as mentioned previously can be another valid explanation since low-income neighborhoods and communities with people of color may experience disproportionately higher levels of air pollutants and may have less access to healthcare and resources to mitigate the impacts of air pollution. Also please note, as commonly known, correlation does not imply causation!



```
CREATE TABLE analytic_2009_to_2014 AS
SELECT *
FROM joined_table
WHERE Name IN ('NO2', 'O3', 'SO2')
  AND Year BETWEEN 2009 AND 2014
  AND Measure = 'Mean';

CREATE TABLE analytic_clean AS
SELECT geography, name, year, age_adjusted_rate, number, estimated_annual_rate, AVG(data_value) AS avg_data_value
FROM analytic_2009_to_2014
WHERE geography IN (
  SELECT geography
  FROM analytic_2009_to_2014
  WHERE name IN ('NO2', 'O3', 'SO2')
    AND year BETWEEN 2009 AND 2014
    AND measure = 'Mean'
  GROUP BY geography, year
)
GROUP BY geography, name, year, age_adjusted_rate, number, estimated_annual_rate;

CREATE TABLE analytic_slope AS
SELECT geography, name, regr_slope(avg_data_value, age_adjusted_rate) AS slope, regr_intercept(avg_data_value,
age_adjusted_rate) AS intercept
FROM analytic_clean
GROUP BY geography, name;
```

Figure 4: Create Table Steps for Regression Analysis

After that, we created another Hive table using the same specifications as before for the table for the previous analytics, except that we expanded it to include years 2009 to 2014. Then, based off of that table, we created another table to select columns such as geography, name, year, and several rates and averages from the previous table. We decided to find the average of the column since there were multiple records for certain pollutants due to measurements taken at different times of the year. This approach would produce a more consolidated dataset so that our results would not be skewed because of the varying amounts of records per pollutant per year in each neighborhood. Thus, the impacts of measurements taken at different times of year can be better accounted for. Lastly, we create a table to store our regression analytic for easy viewing of the results. It groups the data by geography and pollutant name to calculate the regression slope and intercept for each neighborhood and pollutant. We kept the necessary columns for analysis like the neighborhood name, pollutant name, year, the standardized ED visit rate, and data value of mean parts per billion. Since there are 47 different neighborhoods with 3 records of each pollutant, it would take several pages to list out a total of 141 rows. Hence, we provide a visualization of the top 3 neighborhoods with the highest slopes for each pollutant instead down below since that is what we are primarily looking for in our analysis.

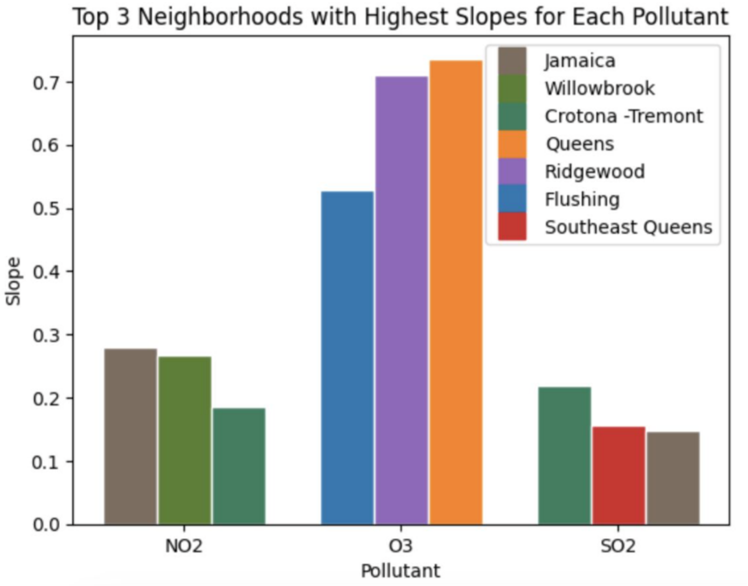


Figure 5: Regression Analysis

For NO_2, the neighborhoods of Jamaica and Willowbrook had the highest regression slopes at 0.279 and 0.267 respectively, suggesting a stronger association between NO_2 concentration and asthma ED visits in these neighborhoods. The neighborhood of Crotona-Tremont also had a positive coefficient, but with a lower value of 0.185.

For SO_2, the neighborhood of Crotona-Tremont had the highest regression at 0.218, followed by Southeast Queens with a coefficient of 0.156, and Jamaica with a coefficient of 0.147.

For O_3, the neighborhoods of Queens, Ridgewood, and Flushing had the highest regression coefficients at 0.736, 0.711, and 0.529 respectively, indicating a stronger association between O_3 concentration and asthma ED visits in these neighborhoods.

First, it's worth noting that the slopes for ozone (O_3) are significantly higher than those for nitrogen dioxide (NO_2) and sulfur dioxide (SO_2) in all three neighborhoods. This indicates that a unit increase in O_3 concentration is associated with a larger increase in asthma ED visits compared to the same unit increase in NO_2 or SO_2 mean concentration.

Regarding the discrepancies between our correlation analysis and the regression results, we speculate that the association between pollutant concentrations and asthma ED visits is in fact not linear and a more complex model is needed to accurately capture the relationship. Changes in pollutant concentrations over time may affect the strength and direction of their association with asthma ED visits. It is also worth noting that correlation analysis only measures the strength of the linear relationship between two variables, whereas regression analysis can capture more complex relationships. This is why we may trust our regressions more as it captures a picture over time rather than a correlation in a singular year and because it aligns with what we have found in related work.

Two neighborhoods, Crotona-Tremont and Jamaica, are shared between NO_2 and SO_2 with relatively high slopes in both pollutants. This may indicate that these neighborhoods are particularly impacted by multiple pollutants and may be especially vulnerable to air pollution’s health effects. Further research could explore the reasons why these neighborhoods are affected by both pollutants and potential interventions that could hopefully mitigate their exposure.

After performing the analytics, we conducted research to analyze these specific neighborhoods. Based on the information provided from the NYC Department of City Planning, Jamaica and Crotona-Tremont are generally considered to be lower-income neighborhoods. Southeast Queens is a mix of middle and lower-income communities, while Willowbrook, Ridgewood, and Flushing are generally considered to be middle to upper income neighborhoods. It's important to note that income levels can vary widely within neighborhoods. Thus, our analysis presented that there are mixed results of air pollutants most affecting both low income neighborhoods and middle to upper income neighborhoods. There could be several reasons for the mixed results we are seeing in the regression analysis. The effects of air pollution on health outcomes can be complex and multifactorial, and may not always be directly correlated with income. Our dataset does not capture all of the relevant variables that could explain the relationships you are observing. Regardless, it is important to prioritize those individuals and communities with little to no access to resources to protect themselves against air pollutants’ effects.

6 Conclusion

As air quality ranks in the top five health concerns of NYC residents (Sarah Johnson et al., 2020), this is an important line of research. In our study, it was found that ozone had the highest association with asthma visits between the three pollutants based on regression, followed by nitrogen dioxide, and sulfur dioxide. Interestingly, this was contrary to the order of correlations observed among the pollutants. The findings highlight the urgent need to reduce air pollution levels in NYC, particularly in specific neighborhoods where air pollution is higher in order to help improve health outcomes for vulnerable populations. This study demonstrates the importance of studying the relationship between air pollution and general health outcomes in urban areas like NYC. By placing our study within the context of past research on the health effects of air pollution, it has been shown that the findings from this study contribute to a growing body of knowledge in this area. This analysis emphasized the need for a continued effort to reduce air pollution levels and improve overall health outcomes, particularly for vulnerable populations such as children, people of color, and low-income individuals.

The significance in the findings lie in their potential to inform public health policies and interventions aimed at reducing air pollution levels and improving asthma and overall health outcomes in NYC. If the air pollutants with strongest association with asthma visits can be identified, that valuable information could be used to target interventions and policies in specific neighborhoods and reduce health disparities. This is particularly important given the rising levels of air pollution combined with asthma prevalence in urban areas around the world. This analysis suggests new ways of thinking about the relationship between air pollution and asthma health outcomes in urban areas. As stated before, by understanding where air pollutants are affecting areas more strongly, there could be policies based on the locations and specific pollutants in order to target those areas and lessen the disparity based on location and socioeconomic status. Additionally, there should be further research in order to understand the complex relationship between air pollutants and asthma health outcomes to help identify other factors that may contribute to asthma health outcomes in urban areas. By adopting a comprehensive and data-driven approach to air pollution management, there could be massive health outcome improvements and it could reduce the burden of asthma on vulnerable populations in urban areas.

Future work we would consider would be to analyze correlations for each pollutant for each year so that we can see trends or patterns instead of just looking at the most recent year. If the correlations for a particular pollutant and asthma visits were consistently high over several years, this may imply a more stable relationship between the two variables. On the other hand, if the correlations vary significantly from year to year, this could indicate a more complex relationship that may be influenced by external factors that change over time. In some ways, we were also limited by what was publicly available regarding asthma and ED visits, because we would have liked to provide more up-to-date insights into the relationship between air pollution and ED visits, especially considering that trends have likely changed in recent years like with COVID, when human activities may have led to less pollutant emissions.

Acknowledgements

We would like to thank the hosts of our datasets, Environment and Health Data Portal and NYC Open Data for hosting publically available datasets. We would also like to thank the HPC team at NYU for helping us share our tables in Hive and being able to run our analysis on Dataproc. Lastly, thank you to Professor Malavet for guiding us through our troubles in this project and for an enriching semester in Big Data at NYU.

7 Works Cited

American Lung Association. “Sulfur Dioxide.” *Www.lung.org*, 12 Feb. 2020, www.lung.org/clean-air/outdoors/what-makes-air-unhealthy/sulfur-dioxide.

Balali-Mood, Mahdi, et al. “Effects of Air Pollution on Human Health and Practical Measures for Prevention in Iran.” *Journal of Research in Medical Sciences*, vol. 21, no. 1, 1 Sept. 2016, p. 65, www.ncbi.nlm.nih.gov/pmc/articles/PMC5122104/, https://doi.org/10.4103/1735-1995.189646.

Boston, 677 Huntington Avenue, and Ma 02115 +1495-1000. “Racial, Ethnic Minorities and Low-Income Groups in U.S. Exposed to Higher Levels of Air Pollution.” *News*, 12 Jan. 2022, www.hsph.harvard.edu/news/press-releases/racial-ethnic-minorities-low-income-groups-u-s-air-pollution/.

CDC. “Air Quality - Air Pollutants.” *CDC*, 21 Nov. 2022, www.cdc.gov/air/pollutants.htm.

Colmer, Jonathan, et al. “Disparities in PM2.5 Air Pollution in the United States.” *Science*, vol. 369, no. 6503, 30 July 2020, pp. 575–578, https://doi.org/10.1126/science.aaz9353.

Hajat, Anjum, et al. “Air Pollution and Individual and Neighborhood Socioeconomic Status: Evidence from the Multi-Ethnic Study of Atherosclerosis (MESA).” *Environmental Health Perspectives*, vol. 121, no. 11-12, Nov. 2013, pp. 1325–1333, www.ncbi.nlm.nih.gov/pmc/articles/PMC3855503/, https://doi.org/10.1289/ehp.1206337. Accessed 8 Dec. 2019.

“Health Impact Assessments.” *Environment & Health Data Portal*, 15 Dec. 2022, a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-stories/hia/.

Jbaily, Abdulrahman, et al. “Air Pollution Exposure Disparities across US Population and Income Groups.” *Nature*, vol. 601, no. 7892, 12 Jan. 2022, pp. 228–233, https://doi.org/10.1038/s41586-021-04190-y.

Johnson, Sarah, et al. “Assessing Air Quality and Public Health Benefits of New York City’s Climate Action Plans.” *Environmental Science & Technology*, vol. 54, no. 16, 14 July 2020, pp. 9804–9813, https://doi.org/10.1021/acs.est.0c00694.

Koenig, Jane Q. “Air Pollution and Asthma.” *Journal of Allergy and Clinical Immunology*, vol. 104, no. 4, 1 Oct. 1999, pp. 717–722, reader.elsevier.com/reader/sd/pii/S0091674999702800?token=7274DCE807E8BC423DC83A2BEDE05FFB02DDD22F16138C7FEAA7500ACC76FBFBB03985F27115A41EF603B394D954C751, https://doi.org/10.1016/S0091-6749(99)70280-0. Accessed 14 Mar. 2021.

Lin, Shao, et al. “Chronic Exposure to Ambient Ozone and Asthma Hospital Admissions among Children.” *Environmental Health Perspectives*, vol. 116, no. 12, Dec. 2008, pp. 1725–1730, https://doi.org/10.1289/ehp.11184. Accessed 9 Mar. 2020.

Mar, Therese F., and Jane Q. Koenig. “Relationship between Visits to Emergency Departments for Asthma and Ozone Exposure in Greater Seattle, Washington.” *Annals of Allergy, Asthma & Immunology*, vol. 103, no. 6, Dec. 2009, pp. 474–479, https://doi.org/10.1016/s1081-1206(10)60263-3. Accessed 5 June 2020.

“NYC Community District Profiles.” *Communityprofiles.planning.nyc.gov*, communityprofiles.planning.nyc.gov/bronx/6.

Reno, Anita L., et al. “Mechanisms of Heightened Airway Sensitivity and Responses to Inhaled SO2 in Asthmatics.” *Environmental Health Insights*, vol. 9s1, Jan. 2015, p. EHI.S15671, https://doi.org/10.4137/ehi.s15671.

Sunyer, J. “Effect of Nitrogen Dioxide and Ozone on the Risk of Dying in Patients with Severe Asthma.” *Thorax*, vol. 57, no. 8, 1 Aug. 2002, pp. 687–693, https://doi.org/10.1136/thorax.57.8.687.

US EPA. “Basic Information about NO2 | US EPA.” *US EPA*, 6 July 2016, www.epa.gov/no2-pollution/basic-information-about-no2#What%20is%20NO2.

US EPA,OAR. “Health Effects of Ozone in the General Population | US EPA.” *US EPA*, 21 Mar. 2016, www.epa.gov/ozone-pollution-and-your-patients-health/health-effects-ozone-general-population.

Zheng, Xue-yan, et al. “Short-Term Exposure to Ozone, Nitrogen Dioxide, and Sulphur Dioxide and Emergency Department Visits and Hospital Admissions due to Asthma: A Systematic Review and Meta-Analysis.” *Environment International*, vol. 150, May 2021, p. 106435, https://doi.org/10.1016/j.envint.2021.106435.