Air Quality & Transit Team D Deliverable 2

Problem Statement

This report investigates the variations in air quality across Boston's different zip codes and its implications on public health, examining data spanning from 2017 to 2022. We aim to understand how Air Quality Index (AQI) levels, particularly Ozone and PM2.5, correlate with demographic factors, disease prevalence, and transit patterns in the city. The analysis seeks to identify trends and potential health risks associated with varying AQI levels, using AQI data, census information, and health statistics from the CDC. The findings are intended to inform policy decisions and interventions aimed at improving air quality and mitigating its adverse effects on the residents of Boston.

Data Used in the Analysis

AQI data was obtained through the AirNow API (https://docs.airnowapi.org/). We collected the zip-code specific AQI data for the years 2017, 2018, 2019, 2021 and 2022. Since the original base questions required us to analyze how different parts of Boston are affected by AQI, we initially downloaded the data specific to zip codes in Boston. Furthermore, we collected certain census data from https://data.census.gov/, specifically the DP02, DP03, DP04, DP05 tables. Finally disease data was obtained from CDC's 'Local Data for Better Health 2021' dataset.

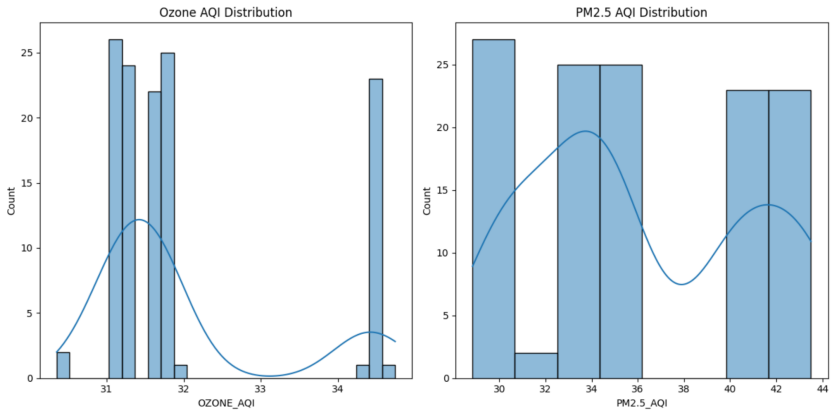
Challenges

The analysis of air quality in Boston using zip code-specific census data and AQI metrics faces several challenges. Firstly, zip-code specific census data is only accessible as a 5-year estimate for 2021, lacking annual granularity to examine how demographic changes might correlate with AQI variations over time. Additionally, the process of data collection is time-intensive, adding complexity to the analysis. The availability and distribution of AQI monitoring sites result in non-granular data, limiting the depth of spatial analysis. Further, certain zip codes with excessive null values were excluded, potentially skewing the dataset. Finally, filtering the dataset for Boston-specific data further restricted the scope of the analysis, as some relevant data became unavailable due to null values, narrowing the range of possible insights.

Exploratory Data Analysis

First, to understand the seriousness of the lack of granularity in our data, we analyzed our AQI data in this dataset. The distributions of both Ozone and PM2.5 AQI (Figure 1) values show a fairly normal distribution, with Ozone AQI having a more concentrated distribution around the mean compared to PM2.5 AQI.

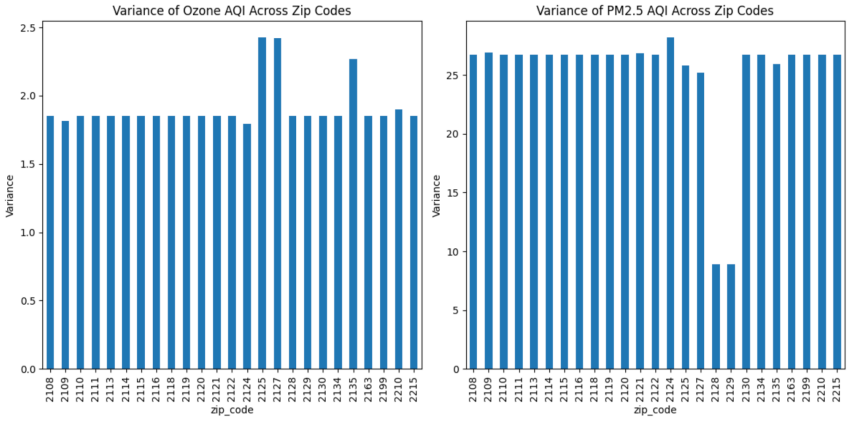
Figure 1



Looking at the standard deviation of both metrics, we see that the Ozone AQI has a standard deviation of 1.24 and the PM2.5 AQI has a standard deviation of 4.70. This suggests that our Ozone AQI metric has values that are not too far apart from each other which means there is a severe lack of granularity in our data.

There is some variance in both Ozone and PM2.5 AQI across different zip codes, as seen in the variance bar charts (Figure 2). However, the variance is relatively low, especially for Ozone AQI, suggesting that the AQI values might be quite similar across different zip codes.

Figure 2



To further understand the extent of this issue, we looked into the number of unique Ozone and PM2.5 AQI values in our dataset. Our dataset contains a total of 125 records, each representing AQI data for a specific zip code and year. The number of unique AQI values in the dataset is relatively low with 13 unique values for Ozone AQI and 18 for PM2.5 AQ. Considering the dataset covers multiple zip codes over several years, having only 13 unique Ozone AQI values and 18 unique PM2.5 AQI values indicates a high degree of data repetition. This suggests that many zip codes have the same recorded AQI values, which is unusually repetitive for a variable

expected to have a wide range due to local environmental differences. Furthermore, the highest occurrence of a single value in our Ozone AQI records is 24, while in our PM2.5 AQI records, the maximum frequency for a single value is 22. Again, these numbers suggest that our AQI data lack a serious amount of granularity and could potentially cause our further analysis to be incorrect or insufficient.

Lastly, the Ozone and PM2.5 AQI has a correlation of -0.504781. The negative correlation between Ozone and PM2.5 AQI is moderate, indicating that these two metrics tend to move in opposite directions but not strongly so. The negative correlation between PM2.5 and Ozone AQI is influenced by various factors. Increased PM2.5 levels result from factors like vehicle emissions, industrial activity, and adverse weather conditions. Conversely, Ozone AQI levels can decrease due to vehicle emissions in traffic-heavy areas. Socioeconomic disparities and public transit usage also impact these pollutants differently. We will try to analyze these correlations further in our report.

AQI Relationships

We will start our exploration with a linear regression analysis. By applying linear regression models to these variables, we aim to discern potential trends in air pollution over time. The analysis could shed light on whether air quality is improving, deteriorating, or remaining stable as years progress.

The regression analysis results (Figure 2) shows a clear trend: PM2.5 AQI is decreasing over the years, while Ozone AQI is increasing. However, it is possible to observe that the relationships are heavily influenced by the outliers that are pulling the regression line upwards / downwards. In order to get a better understanding of air quality trends over time, we will remove these outliers and run regressions again on the new data.

Figure 2

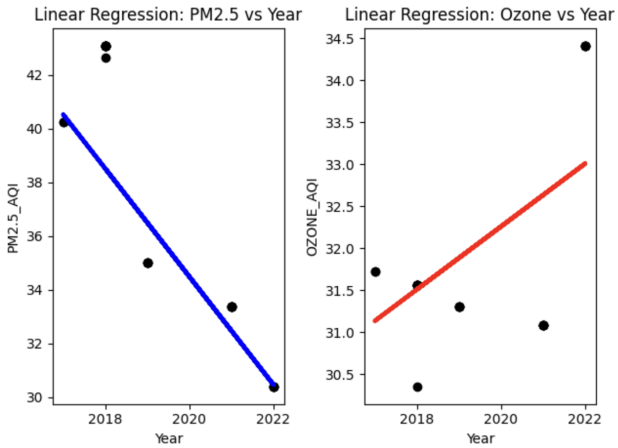
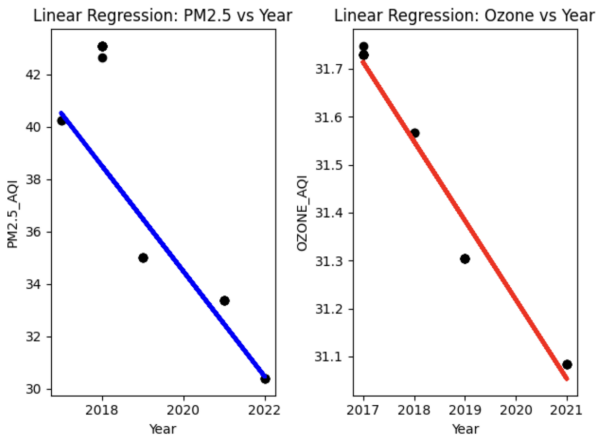


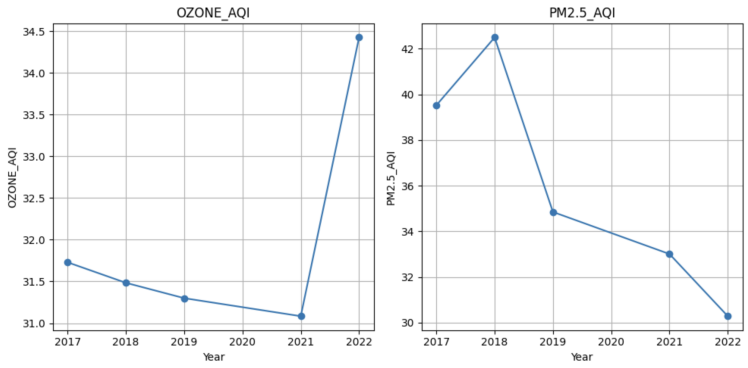
Figure 3



The revised regression analysis (Figure 3), after the removal of outliers, shows a consistent decrease in both PM2.5 AQI and Ozone AQI over time. This downward trend for both pollutants could suggest that overall air quality in Boston is improving. This improvement could be due to a variety of factors, such as successful environmental policies, a transition to cleaner energy sources, advancements in emissions-reducing technologies, or even changes in public behavior such as reduced usage of personal vehicles or increased use of public transportation. However, it should also be noted that the outliers could also be representing a trend of more unhealthy air quality days and the removal of these data points could be obscuring occasional but significant spikes in pollution levels. These spikes could result from transient events like wildfires, industrial incidents, or variations in weather patterns that significantly impact air quality. As we continue to monitor and analyze these trends, it becomes increasingly evident that the interplay between regulatory initiatives and technological advancements is playing a pivotal role in shaping a healthier and more environmentally conscious future.

The fluctuations in AQI (Figure 4), particularly the increase in Ozone AQI in 2022 and the general decrease in PM2.5 AQI from 2018, could be influenced by factors such as changes in local policies, economic activities, or environmental conditions.

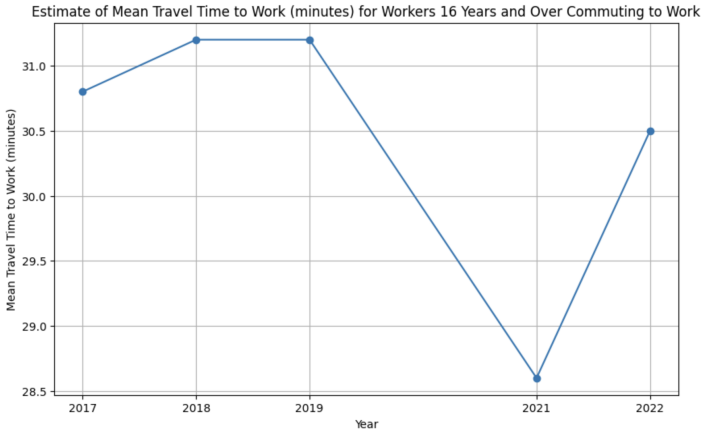
Figure 4



According to the U.S. Environmental Protection Agency's (EPA), there has been a significant decrease in PM2.5 levels since 1990. Specifically, the annual concentration of PM2.5 has decreased by 37% from 2000. This can explain the downward PM2.5 AQI trend, yet we will further investigate any additional causes for this. Moreover, EPA's New England regional office confirmed in October 2022 that "Based on preliminary data collected between March and September 2022, there were 24 days when ozone monitors in New England recorded ozone concentrations above levels considered healthy." (EPA). Most of the pollution responsible for ozone formation comes from large combustion sources, vehicles, and various everyday activities such as using household products. In order to investigate if there is any explanation for these AQI changes, we created another dataset. This dataset contains AQI metrics as columns and years as rows, and the data is Boston specific as opposed to zip code areas in Boston. With this dataset, we will not be able to find a relationship among different parts of Boston and Air Quality, however we will have the chance to see if any trend in Boston in general is affecting the AQI changes.

The only transportation census data that is available for the entirety of Boston is the mean travel time to work. Variations in this commute time are indicative of traffic flow changes, potentially influencing air quality. For example, extended commutes often coincide with increased traffic congestion, augmenting vehicular emissions, and consequently affecting Ozone AQI. As depicted in Figure 5, there's a notable escalation in mean travel time to work in the latest year. This uptick could correlate with a rise in Ozone AQI levels, possibly due to higher emissions from prolonged vehicle operation as more time is spent in transit, leading to greater ozone formation from vehicle exhaust.

Figure 5



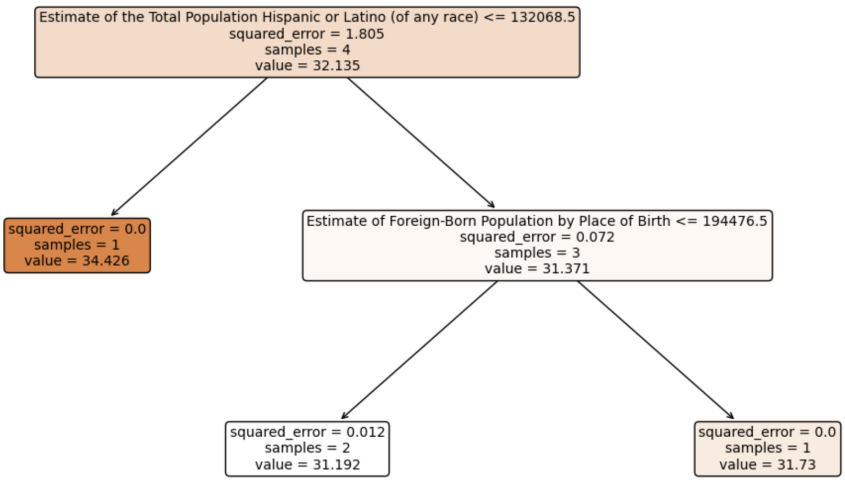
Furthermore, we analyzed the relationship between the adult population in Boston and how it relates to air quality changes over time. Our analysis showed a correlation of -0.574009 between the percentage of the total population 18 years and over and Ozone AQI, as well as a correlation of 0.768753 between the percentage of the population 18 years and over and PM2.5 AQI. Here is what it could mean.

A negative correlation between Ozone AQI with the adult population percentage may hint at adult populations favoring areas with better air quality or more stringent air quality regulations. Adults often settle in areas that offer a higher quality of life, which may include environmental considerations like cleaner air. This demographic might also be more politically active, advocating for policies that lead to improved air quality, thus contributing to lower Ozone AQI levels. The positive correlation with PM2.5 AQI and the adult population percentage might be attributed to urbanization. Areas with higher concentrations of adults are likely to be urban centers with dense traffic and industrial activities that contribute to PM2.5 emissions. Adults might commute more for work and contribute to higher vehicular emissions, leading to increased PM2.5 levels. Additionally, urban areas with high adult populations can have more construction and development projects that contribute to particulate matter in the air.

In order to get a better understanding of what other census data could have a relationship with AQI, we made a decision tree analysis. This is going to help us understand what breaking points in the census data affect air quality. The decision tree analysis (Figure 6) decisively indicates a

correlation between Boston's air quality and specific demographic groups. The consistent emergence of the Hispanic or Latino population as a primary node underscores a stark reality: these communities are likely situated in areas where air quality is poorer. This isn't a matter of chance; it's a reflection of systemic patterns where minority groups often reside in neighborhoods that bear the brunt of environmental neglect, situated near high-emission zones like factories and congested highways. Furthermore, the tree's splits based on foreign-born populations suggest these individuals are disproportionately affected by air quality issues, likely due to socio-economic factors that place them in high-risk environments for pollution exposure.

Figure 6



Extension Proposal

In an effort to enhance the granularity and accuracy of our study on the impact of air quality on public health in Boston, our team plans to integrate an additional, more detailed AQI data source from AQICN (https://aqicn.org/city/boston/). This source provides extensive coverage with numerous sensors across various zip codes in Boston, offering a deeper insight into the local air quality variations. Furthermore, our team has conducted some data analysis on CDC & AQI data. Our preliminary analysis, utilizing CDC and existing AQI data, has already revealed promising correlations between specific diseases and AQI levels. By enriching our dataset with this more granular AQI data, combined with detailed census information and CDC health statistics, we aim to construct a comprehensive and nuanced understanding of how air quality specifically impacts health outcomes in different parts of Boston. This multidimensional approach is going to help us uncover new insights and strengthen the evidence base for targeted public health interventions in the city.

Visualizations for Extension Proposal

Our team extracted zip code specific disease data from a dataset provided by CDC which we later merged with zip-code specific AQI data that we had previously collected. In order to analyze air quality's effects on the population, we selected certain diseases that could be linked to air quality and analyzed these relationships.

One disease that is directly related to air quality is asthma. In order to understand the relationship between asthma prevalence and air quality we created the scatter plots seen below (Figure 8 & 9).

Figure 8

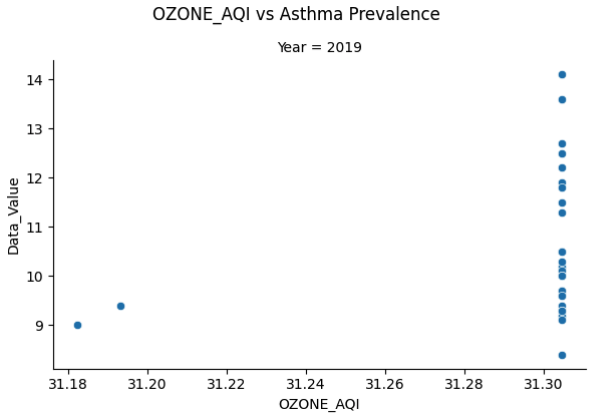


Figure 9

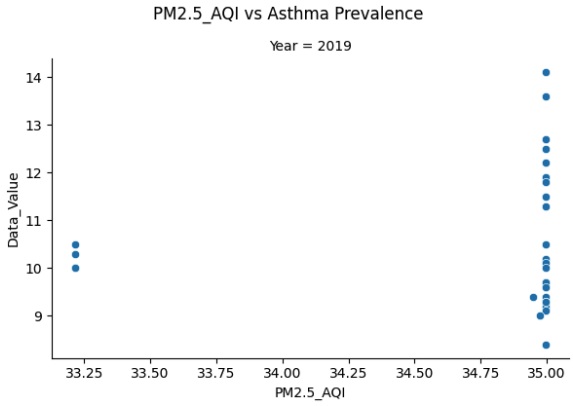


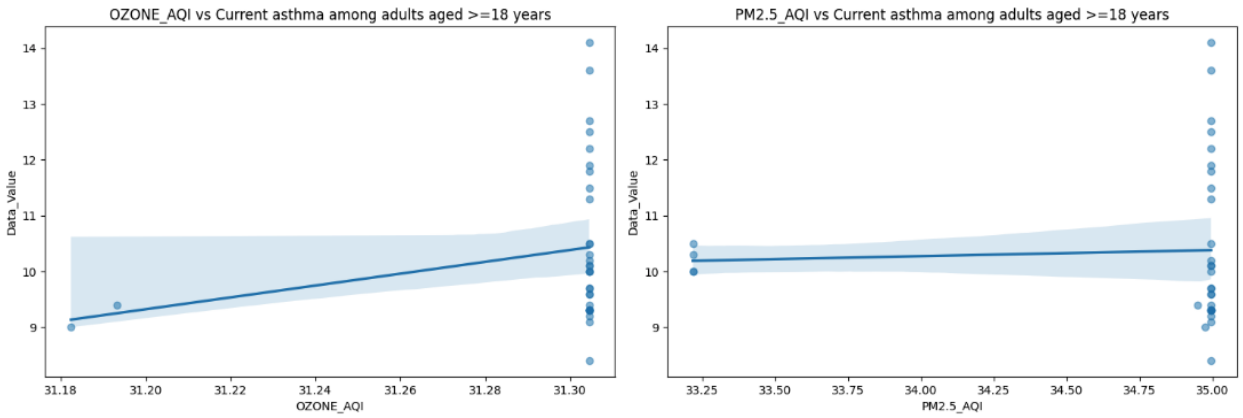
Figure 8 displays a discernible upward trend, indicating a positive correlation between OZONE_AQI and asthma prevalence rates for the year 2019. Specifically, as the OZONE_AQI values incrementally rise, there is a notable increase in the prevalence of asthma cases. This pattern suggests a potential association where higher concentrations of ozone in the air may correspond to an increase in asthma incidence among the population. The nature of this correlation hints that elevated ozone levels, often a result of various pollutants reacting under sunlight, could exacerbate respiratory conditions, thereby contributing to the frequency or severity of asthma symptoms. It's also plausible that certain areas with higher ozone levels, perhaps due to traffic congestion or industrial activities, may concurrently experience elevated rates of asthma, suggesting a geographical and environmental dimension to public health concerns.

The visualization comparing PM2.5 Air Quality Index (PM2.5_AQI) to asthma prevalence in 2019 (Figure 9) suggests a positive association. As PM2.5_AQI increases, there is a visible clustering of higher asthma prevalence rates. This trend implies that areas with poor air quality, as indicated by higher levels of particulate matter 2.5, could be linked to an increase in asthma cases. PM2.5 particles are known to penetrate deep into the respiratory tract, potentially exacerbating asthma and other respiratory diseases. The plot reinforces the idea that particulate pollution is a significant health risk, especially for those with pre-existing respiratory conditions.

It's important to consider that fine particulate matter can come from various sources, including vehicle emissions, industrial processes, and natural events like wildfires, all contributing to the overall air quality and possibly correlating with the distribution of asthma cases observed in the data.

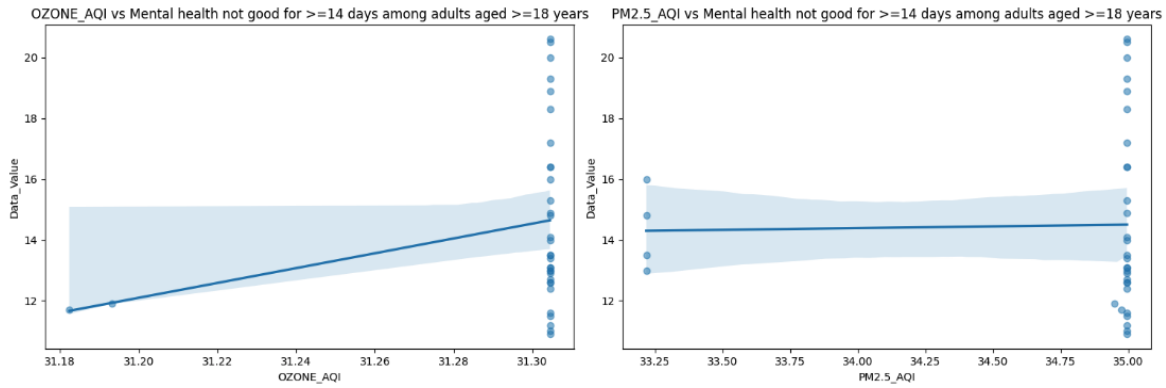
In order to investigate any further relationships between diseases and air quality, our team decided to make a regression analysis of how other diseases' prevalence changes with air quality. It is important to note that in these regression graphs, the shaded area represents the confidence interval, providing a range where the true regression line may lie with a certain level of confidence. It should also be noted that because of the lack of granularity in this data certain data points are seen as outliers. With the current Air quality data at hand, a better analysis was not possible however, as mentioned above our team will enrich the data to prevent any outliers skewing the relationships in the next steps of this project.

Figure 10



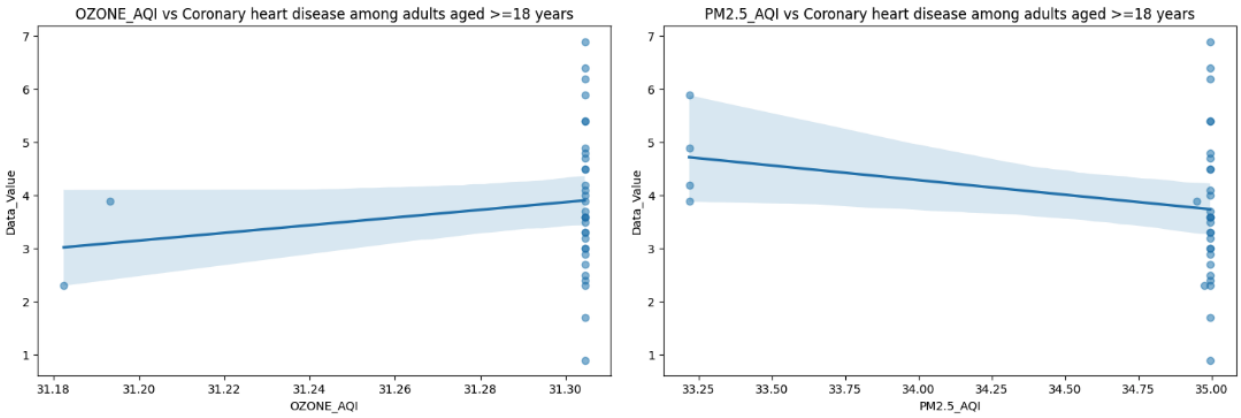
In figure 10, the left plot shows a positive linear trend between OZONE_AQI and asthma prevalence, indicated by the upward slope of the regression line. This suggests that as the ozone level increases, the rate of asthma among adults also tends to increase. The right plot presents a less steep positive linear trend between PM2.5_AQI and asthma prevalence, which could indicate a weaker correlation as compared to ozone. The confidence interval is broader here, reflecting more uncertainty around the estimated regression line.

Figure 11



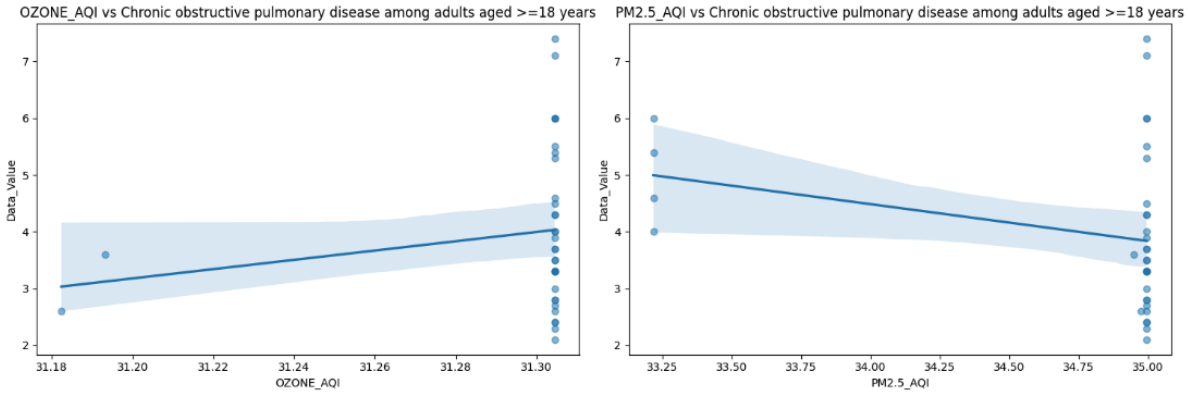
Emerging evidence indicates that air pollution can have a negative impact on mental health, contributing to conditions like depression and anxiety. In Figure 11, the plot on the left indicates a positive correlation between OZONE_AQI and reports of poor mental health days. In the case of ozone, a known irritant to the respiratory system, the implication of a connection to mental health may be less direct but could be understood through the lens of general well-being. If higher ozone levels aggravate respiratory conditions, the consequent discomfort and health concerns could lead to stress or exacerbate existing mental health issues. Additionally, high ozone days often trigger advisories that limit outdoor activities; such restrictions could negatively affect mental health by reducing opportunities for exercise and social interactions that are beneficial for mental well-being. The plot on the right exhibits a less pronounced positive trend, with the regression line having a flatter slope compared to the OZONE_AQI plot. For PM2.5, the relationship with mental health outcomes can be considered in terms of both physical and psychological pathways. Fine particulate matter can cross the blood-brain barrier, potentially causing inflammation and oxidative stress that have been linked to neurodegenerative diseases and mood disorders. Moreover, living in areas with high PM2.5 levels might also be associated with socioeconomic factors that contribute to stress and limited access to mental health resources.

Figure 12



In figure 12, the left plot shows a modest upward trend, suggesting a potential association where higher OZONE_AQI levels could correlate with an increased prevalence of coronary heart disease. Ozone is a powerful oxidant, and exposure to higher levels could lead to vascular inflammation and oxidative stress, which are known risk factors for coronary heart disease. The right plot indicates a more pronounced negative trend, which might seem counterintuitive since PM2.5 exposure is typically associated with adverse cardiovascular outcomes. However, this could suggest a non-linear relationship or indicate the influence of other confounding factors not accounted for in this simple regression model. It is also possible that at higher PM2.5 levels, other protective factors or interventions are in place that mitigate the risk, such as improved healthcare access or public health policies that are activated when pollution reaches critical levels.

Figure 13



In figure 13, the left plot exhibits a positive trend, suggesting a relationship where higher levels of ozone are associated with an increased prevalence of COPD. Ozone's role as a respiratory irritant could exacerbate existing COPD conditions or contribute to their development by damaging the lung tissue, which aligns with existing research on the health impacts of ozone exposure. The right plot, in contrast, shows a negative trend between PM2.5_AQI and COPD prevalence. Typically, we expect higher PM2.5 levels to be detrimental to lung health due to the particles' ability to penetrate deep into the lung passageways and interfere with respiratory function. The observed negative correlation might point to complexities in the data that are not captured by a simple linear model, such as variable exposure times, the presence of other pollutants, or the effectiveness of public health interventions at higher pollution levels. Another possible explanation that was previously mentioned is that the lack of granularity in the data prevents the true relationship between Pulmonary diseases and AQI to be observed.

In conclusion, our comprehensive investigation into the impacts of air quality on various health outcomes in Boston has revealed meaningful correlations that warrant further exploration. Our visual analyses, particularly the regression plots for asthma, mental health, coronary heart disease, and COPD, have highlighted potential relationships between increased pollutant levels and heightened disease prevalence. Especially notable is the positive trend observed between ozone levels and both asthma and COPD prevalence, which aligns with current scientific understanding of ozone as a respiratory irritant. The complex patterns observed with PM2.5 underscore the intricate interplay between environmental pollutants and health outcomes, and the negative trends noted for heart disease and COPD with PM2.5 warrant a closer look to fully grasp the underlying dynamics. As we prepare to delve deeper into this research, we aim to refine our methods and data quality to mitigate the influence of outliers and enhance the accuracy of our findings.

Individual contributions

Mithat Kus: Led the collection and analysis of Air Quality Index (AQI) data from the AirNow API, focusing on the years 2021, and 2022. Wrote the python script to fetch the aqi data. To account for the long runtime of this code (around 36 hours for each year) the code was specifically designed to continually save the progress to a database and continue from wherever it left off in the case of an interruption. Conducted the exploratory data analysis, including the creation of distribution and variance bar charts for Ozone and PM2.5 AQI. Performed the initial linear regression analysis to identify trends in air pollution over time, and was responsible for the analysis revision after outlier removal. Coordinated with Maria in interpreting the regression analysis results and drawing conclusions about air quality trends in Boston.

Maria Eusse Henao: Collected AQI data for the year 2019. Responsible for acquiring and analyzing census data from data.census.gov, focusing on demographic factors like DP02, DP03, DP04, DP05 tables. Responsible for writing the code for the initial regression and decision tree analysis. Led the investigation into the relationship between air quality and various health outcomes, including asthma, mental health, heart disease, and COPD, using CDC data. Collaborated with Mithat in developing and interpreting the regression plots and scatter plots for health outcomes. Took a significant role in writing and structuring the final report, ensuring clarity and coherence in presenting the findings.

Sanath Bhimsen: Collected AQI data for the year 2018. Focused on collecting and analyzing transportation and transit data as it relates to air quality, including mean travel times and other relevant metrics. Assisted Maria in analyzing the adult population demographics and its correlation with AQI. Contributed to the visualization efforts, particularly in the creation of graphs and charts related to transportation data and demographic analysis. Participated in preliminary discussions and planning of the extension proposal, specifically in identifying potential data sources.

Chengkai Yang: Collected AQI data for the year 2017. Assisted in the collection of zip-code specific disease data from the CDC and played a role in the initial stages of data merging and cleaning. Supported Sanath in analyzing transit data and helped in the decision tree analysis focusing on demographic groups and their relationship with air quality (Figure 6). Involved in the logistical aspects of the project, including organizing meetings, managing documentation, and facilitating communication between team members. Contributed to the review and editing of the final report, ensuring accuracy and completeness of the information presented.