Boston Transit and Air Quality Deliverable 2 Report

Team C

Background and Data Collection

Boston's dense transportation network raises concerns about its impact on air quality. As car emissions dominate Massachusetts's environmental challenges, it's crucial to understand the impact that transportation has on Boston's air quality. Furthermore, the socio-economic disparities in experiencing this air quality remain a concern. This project aims to investigate the link between transportation infrastructure and the air quality, and its varied effects across neighborhoods.

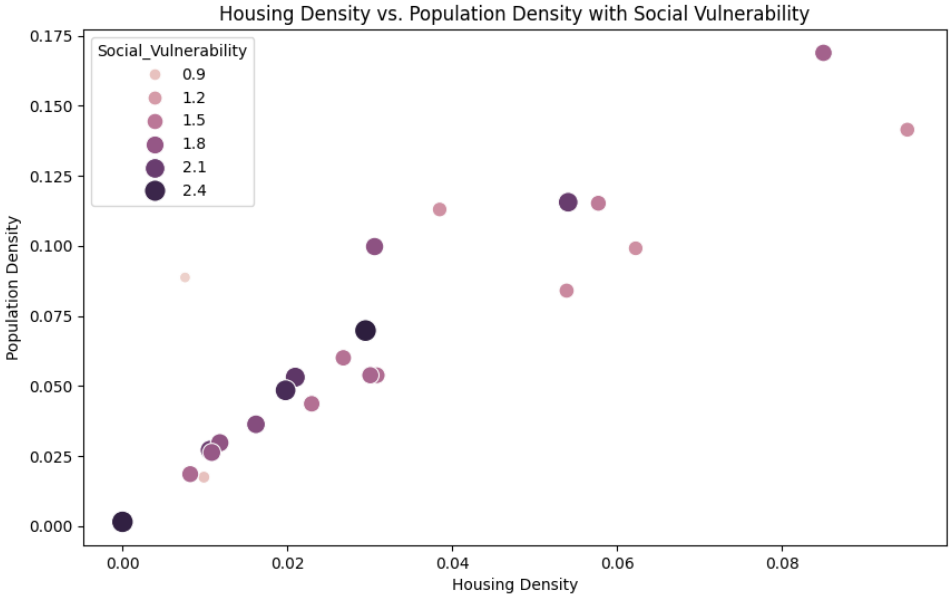
Our work thus far has centered around four sources of information:

1. Air Quality Index (AQI): We used a bash script to fetch data from the AirNow API, collecting hourly air quality readings from 2020 to 2023. This data was consolidated into a single CSV file for preprocessing. To clean the data, we removed rows and columns containing only NaN values (e.g., PM10_value column). We then focused on relevant columns (AQI values for PM2.5, NO2, and OZONE) and created a new dataframe incorporating these along with neighborhood names. To address null values, we replaced them with the mean of the corresponding feature. Finally, we grouped the data by neighborhood and calculated the average AQI over the four-year span.
    a. PurpleAir: While we primarily used the AirNow API for gathering air quality data, we also attempted to use PurpleAir's API to get additional data and sensors. However, the PurpleAir API had a very large issue in that all the air quality collection sites were privately owned, meaning that most of them were inaccessible for public use. Of the ones that were publicly available, the periods of time they were active for were very sporadic and unpredictable, leading us to eventually conclude that PurpleAir's air quality data would be too difficult to

work with to draw any meaningful conclusions out of due to the inconsistency of timing for the data collection sites.

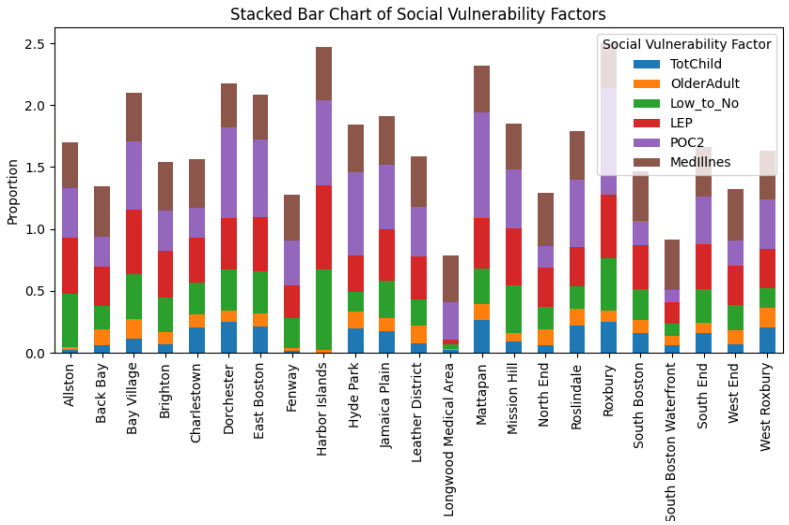
2. Pollution Proximity Index (PPI): We analyzed and visualized pollution proximity data for Boston using a dataset on road proximity. After loading this data from a CSV file, we conducted a basic inspection, outputting the top 5 rows and summary statistics to an output.txt file.
3. Social Vulnerability Index: We downloaded data from Climate Ready Boston, focusing on geographic and social vulnerability information for each location. This information was used to calculate the percentage of the population in each socially vulnerable category and the housing and population densities. We aggregated this data to create a social vulnerability index for each neighborhood, applying equal weights to all categories.
4. Incomes: We gathered median income data for all 23 Boston neighborhoods from BostonPlans.org and Census.gov for 2021. Our initial plan was to extend this data collection to 2020-2023 for a comprehensive analysis alongside the AQI data. However, due to issues with the AirNow API and changes in the project scope, this extension was not realized.
5. Census Data: We collected census data for all zip codes in Boston all the way back to 2011. This includes gender, age, race(ACS), and citizenship status. We initially gathered these data in hopes to draw conclusions between areas with better air quality and poor quality based on these characteristics. However, due to issues with the AirNow API and AQI data collection, we were unable to utilize these data either.

Exploratory Data Analysis and Visualizations

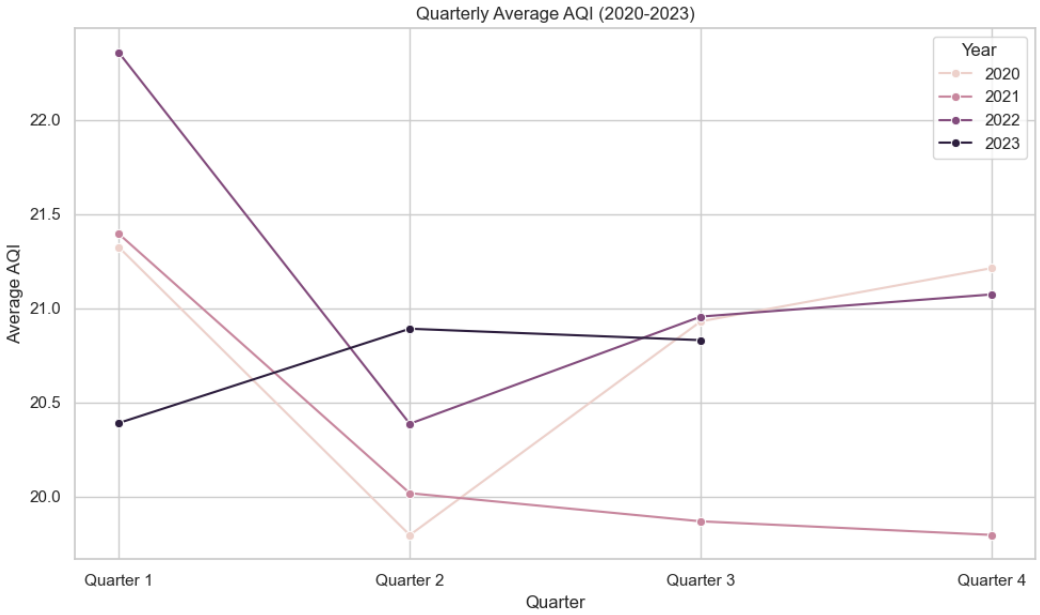


- The scatter plot is mapping housing density against population density for neighborhoods in Boston, with the size of each data point reflecting the Social Vulnerability Index (SVI). A clear positive correlation is depicted between housing and population density, indicating that as the number of housing units per area increases, the population density tends to rise correspondingly. The SVI, represented by the varying sizes of the plot points, seems to initially increase with density, suggesting that denser areas may experience higher social vulnerability. However, the analysis indicates a turning point around (0.03, 0.075) on the graph where the SVI begins to decrease as the density continues to increase. This could suggest that beyond a certain density threshold, neighborhoods may see a decline in social vulnerability, potentially due to better access to resources, infrastructure, and community resilience measures. The presence of an outlier near the origin suggests that there are neighborhoods with low density but disproportionately high social vulnerability, an outlier that would need further investigation. Understanding why this is the case could reveal insights into how social vulnerability is affected by factors other than density. The graph suggests a complex

relationship between housing density, population density, and social vulnerability, with implications for urban development and social policy.

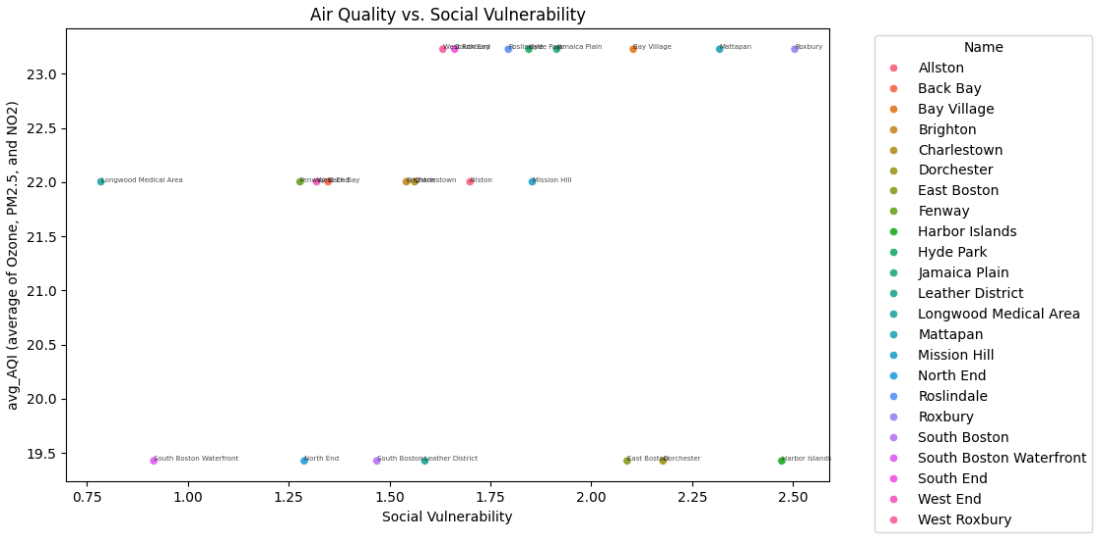


- The graph above displays every neighborhood in Boston along with the proportion of the population in each of the six social vulnerability factors: young children, older adults, those with low to no income, those with limited English proficiency, people of color, and those with medical illnesses.
- In every neighborhood, POC2 (people of color) and MedIllness (people with medical illness) take about the same proportion of the index. In comparison to the total population overall, people of color statistically have lower income, which means a high proportion of POC2 is correlated with a high proportion of Low_to_No incomes.
- Individuals with lower income may be restricted to which neighborhoods they can choose. This means that people may be unable to afford housing in a more excellent neighborhood or air purifiers. Additionally, if the neighborhood has lousy air quality, people with medical issues will be more likely to suffer.

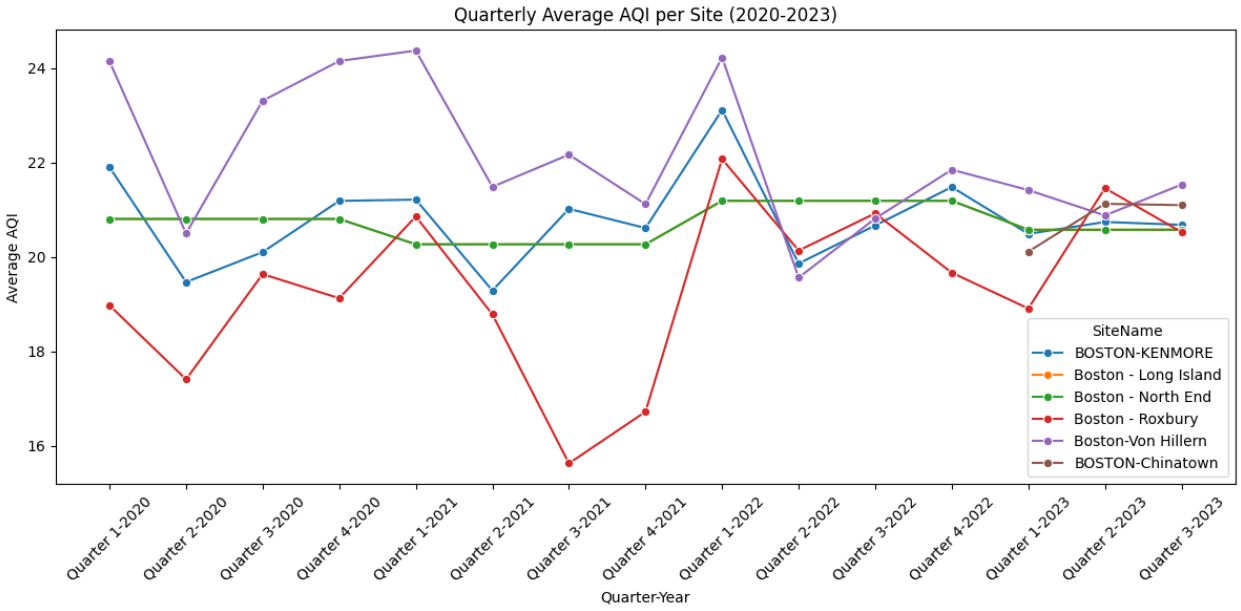


- The data suggest that the air quality is generally good, with little variations between sites, which led to the decision to utilize an average of all collected data points to represent the quarterly average AQI for each year. 2021 was characterized by a notably better air quality compared to the other three years. This improvement could be largely connected to the impact of the COVID-19 pandemic, particularly during periods when quarantine measures were most strict. The decrease in transportation due to more people working from home may be a reason values were lower. Air quality in the first quarter and last quarter of each year is worse when compared to the other quarters. This can be attributed to the colder temperatures, which lead to an increased use of private vehicles over public transportation, and a rise in electricity consumption as heating needs increased. A significant decrease in the AQI value was recorded in the second quarter of 2020, the start of COVID-19 pandemic.
- In 2020, the AQI standard deviation (SD) is the lowest among the four years at approximately 3.79, suggesting that air quality were relatively stable around the mean for that year. This stability, however, was disrupted in 2021, where the SD increased to approximately 4.45, indicating a wider variance in air quality. This could be attributed to the gradual lifting of pandemic-related restrictions, leading to increased traffic and industrial activity. 2022 showed a significant reduction in the SD to 2.92, which implies a return to more consistent air quality levels. This might reflect the adaptation to a new normal with stable economic activities and consistent environmental regulations. For 2023, the SD is approximately 3.72, which is close to the value in 2020. This suggests that while there is some variation in AQI values, the air quality has returned to a level of consistency similar to the pre-pandemic period.
- The varying SD indicate that while the overall air quality remained relatively good, the consistency of the air quality experienced notable changes. The year 2021 stands out as

having the greatest variability, which could be linked to transitional environmental and economic activities as the world was adjusting to the pandemic's impact. Subsequent years show a trend towards stabilization, with 2022 displaying the most uniform air quality levels within this dataset.



- For each neighborhood, we plot the social vulnerability against the average of Ozone, PM2.5 and NO2 AQIs. We notice the clear three horizontal lines for average AQI since we only have three active monitoring sites.
- Based on the graph, we are not able to clearly spot a correlation between social vulnerability and air quality. Those with higher air quality tend to have a larger variance and those with lower air quality tend to be more vulnerable. But beyond that we are not able to draw clear conclusions.
- The graph may have generalizability issues due to the lack of monitoring sites we currently have, resulting in us grouping multiple neighborhoods into one monitoring site by distance.



- The above graph depicts the average AQI for all available sites for each quarter from 2020 to 2023. However, there are some instances of missing records from the dataset. In particular, the orange line (Long Island) is missing because the site is inactive for all four years, and the site at Chinatown only began recording data at the start of 2023.
- Overall, the AQI values collected at North End are pretty steady in the past four years while the values collected at Von Hillern and Roxbury fluctuate a lot compared to the other sites.
- However, note that while the variations in the graph may look large, the entire range of AQI only ranges from ~16 at best to ~24 at worst, which overall exhibits very small variance in air quality over these four years. From this, we can conclude that while there are some small differences in air quality from year to year, the overall air quality in Boston has not changed significantly over recent years.

Extension Proposal

Project Focus on Sensor Proximity: Shifting our project's focus to areas near the sensors entails concentrating our efforts on analyzing and addressing air quality concerns in immediate proximity to the monitoring devices. This approach would involve a more localized and granular examination of air quality, allowing us to uncover nuances that may not be evident when looking at broader geographic areas.

By honing in on sensor proximity, we can:

- Identify and target specific areas with air quality issues, potentially leading to more efficient interventions.
- Provide hyper-localized air quality information, which could be valuable for communities, businesses, and policymakers.
- Explore trends and patterns in air quality variations at a finer scale, offering a deeper understanding of environmental factors affecting specific regions.

This modification could result in a more targeted and actionable project, as we delve into the intricacies of air quality within close reach of the sensors.

Exploring Alternative Proxies for Air Quality: Investigating alternative proxies for air quality involves considering various factors beyond traditional air quality metrics. These proxies can provide additional context and insights into the overall environmental health of specific areas. Some examples include:

- Proximity to Roads: Analyzing how close an area is to major roadways can shed light on the influence of vehicular emissions on air quality. This information can be critical for urban planning and traffic management.
- Public Transport: Examining the accessibility and usage of public transportation systems in an area can help us understand the potential impact of reduced private car usage on air quality and sustainability.
- Different Industries: Assessing the presence and types of industries in an area can provide insights into sources of pollution and environmental impact.

Understanding the relationship between industrial activities and air quality is essential for addressing pollution at its source.

Exploring these alternative proxies for air quality offers a more comprehensive view of the factors contributing to environmental conditions. By doing so, we can develop a richer understanding of the complex interplay between various variables and air quality, potentially leading to more effective strategies and solutions. However these are just ideas brought up during our discussion with Professor Galletti and while these alternative proxies for air quality are promising, it's important to note that their feasibility relies on the availability of relevant datasets or information.

Extension Proposal Insights and Research

As part of our preliminary research into the extension proposal, we attempted to look into other data sources to see if they would be appropriate for the revised project, and if they would have any improved data compared to the base sources.

1. Our team began by examining the Google Map Air Quality API, as recommended by our Project Manager through Slack. Initially, we were encouraged by its promise of ongoing calculations of air quality metrics within areas measuring 500 x 500 meters, offering precise data granularity. However, further investigation raised concerns about the extent of historical data accessibility, specifically whether we can retrieve data beyond a 720-hour (30-day) window, either per query or in total. Additionally, a notable drawback of this API is the cost of $0.005 per request.
2. CDC Datasets:
    a. We also attempted to use data from the Center for Disease Control as another source for Boston's Air Quality. However, the data given by this source is far too general to be of use for this project.. The CDC appears to separate air quality by county over all of Massachusetts, meaning that we would be able to find an aggregate air quality record for all of Boston, but not for any specific neighborhood or region within Boston.
    b. The CDC also has the Local Data for Better Health datasets for every year. These datasets are comprehensive data that provide ZIP Code Tabulation Area level estimates for various health-related metrics across the United States. While this source also initially appeared promising, none of the listed measures were related to measuring air quality, or even to any health conditions/statistics related to poor

air quality. As such, this source would not give any insight to any questions related to Boston's air quality or transportation.

3. Other sources:
    a. We also have tried looking for other credible sources online. One of such is aqicn.org. They offer particle pollution data from current all the way back to 2014. This initially looked promising but we soon realized that they use data from AirNow. We have proven before that AirNow data is worthless to us, which means aqicn.org is also of no use.

Individual Contributions

a. Matias: For Deliverable 2, I efficiently retrieved four years' worth of data from the AirNow API using an updated bash script, investing over 40 hours in this task. I also developed code to analyze and visualize this data. Eric and I collaborated on crafting most of checkpoint A. I led discussions on data limitations and challenges, addressing core and extension questions. Currently, I'm working closely with team A to explore and analyze new datasets, seeking ways to adapt our project for the semester's remainder. I'm actively communicating updates and outcomes via Slack to TAs and other teams working on Transit and Air Quality. Furthermore, I contributed significantly to the Deliverable 2 report, summarizing our progress in addressing project questions, adapting to scope changes, and outlining future structural needs.
b. Yuchen: For Deliverable 2, I created the script to fetch the hourly data, and mainly worked on exploring and analyzing the newly retrieved hourly data for the year 2020 to 2023. I preprocessed the dataset to create average AQI values for different years and different sites, which are the main columns used to create plots for analyzing. Besides that, I also explored the new data sources (e.g. PurpleAir, Google air quality and CDC health data). I found that Google source does seem to be a good fit for our purpose but according to the documentation, it may not include the data that is older than 30 days.
c. Eric: For Deliverable 2, I worked with Matias to create and refine the slides used in the Early Insights presentation. I assisted with analyzing more of AirNow's Air Quality data and creating the graphs used in the previous deliverables. I also participated in the discussion regarding the Transit + Air Quality extension project and the limitations of the current datasets in answering the base/extension questions, and worked with Matias in communicating the revised project details and answering questions for the other Transit + Air Quality teams on Slack. In addition to performing further analysis on the data we previously gathered, I am

also performing research into other potential sources of data we could utilize for the revised project.

d. Peter: For Deliverable 2, I worked on fetching census data from census.gov. However these data were not used in the final deliverable due to issues. Then I looked at google maps api in hopes to find some workable data. I also looked for other credible sources that we can fetch AQI data from including aqicn.org, mapc.org and other cites.
e. Steve: For Deliverable 2, I worked on data and graph analysis first. I prepared and added slides for the initial insights presentation. Additionally, I engaged in identifying and addressing limitations within our current datasets. Although I explored the CDC's Local Data for Better Health datasets to establish a correlation between air quality and health conditions, I found the relationship was not strongly supported by the data. I also participated in discussions about the extension project. Lastly, I collaborated with the rest of the team to complete and finalize the Deliverable 2 report.