Analysis of Air Quality and Demographic Characteristics in Boston: A Data Science Exploration

MBTA Transit Air Quality Team B

Sinforiano (Sammy) Terada, Class of 2024, syterada@bu.edu

Chao-Jen (Jackson) Chiu, M.S Class of 2026, cjchiu@bu.edu

Jonathan Suarez, Class of 2024, jwsuarez@bu.edu

Naveen Vaidyamath, Class of 2024, naveen02@bu.edu

Zixuan (Cathy) Wang, Class of 2024, zxwang12@bu.edu

Table of Contents

Introduction

Project Focus and Overall Goal

The primary objective of our project is to explore the impact of Boston's transportation systems, including buses and roads, on air quality. We aim to understand how air quality varies across the city over time, particularly in relation to transit access. Crucially, we're examining the interplay of this air quality data with demographic aspects such as race and ethnicity, income levels, housing and population density, and social vulnerability. Our goal is to unravel the complex ties between transportation, air quality, and public health in the context of Boston.

Importance of the Project

This project holds substantial importance due to its potential to enhance public health, mitigate environmental damage, foster equity, and improve urban planning practices in Boston. With public transportation often linked to health challenges, particularly in marginalized communities, our analysis seeks to inform strategies that can reduce air pollution and its negative impacts, thereby benefiting these communities.

Our Approach

Initially, our approach was to gather and analyze a diverse array of data sets:

1. **Geospatial Data:** Geographic locations of roads, public transportation stops, and residents.

2. **Air Quality Sensor Data:** Measurements of air pollutants such as PM2.5, PM10, nitrogen dioxide, etc.
3. **Demographic Data:** Information on race/ethnicity, area median income, housing density, population density, and social vulnerability.
4. **Health Data:** Health outcomes like asthma and lung cancer rates.
5. **Transportation Data:** Public transportation details including transit times and traffic data.
6. **Social Vulnerability Index Data:** Geographic data to assess social vulnerability in different areas.

Midway through the project, we encountered significant limitations and inconsistencies with the initial data sets. This challenge necessitated a shift to new data sources and a revised approach to our base question. Our focus shifted to an extension proposal that sought to delve deeper and uncover more nuanced findings, observations, conclusions, and potential solutions.

We pivoted to using the Pollution Proximity Index (PPI) data set, provided by the Metropolitan Area Planning Council (MAPC). Despite its lack of specific location identifiers, we leveraged a dataset from MAPC that maps each town in the MAPC to its "Community Type" classification. These classifications of the areas had PPI data for each 250 square meter grid of all of the MAPC Massachusetts. This dataset allowed us to investigate correlations between income, population density, housing density, and "Community Type" classifications, alongside public transit accessibility and types within these classifications.

One limitation of this revised approach is the inability to analyze annual changes and health outcomes due to the PPI data being limited to 2020. Additionally, the provided CDC data

lacked the necessary location specificity for our project. Our revised question for the extension project became: “What is the relationship between the community types (municipalities mapping to community types) in Boston and the PPI data with regards to race distributions, median income, and MBTA infrastructure?”

Through this project, we strive to contribute to a deeper understanding of how Boston's transportation infrastructure impacts air quality and, in turn, affects various demographic groups. Despite the challenges posed by data limitations, our team remains committed to providing meaningful insights that could influence future urban planning and public health initiatives in Boston.

Base Question Analysis

Introduction

In the realm of urban planning and public health, understanding the relationship between air quality and demographic factors is crucial. This report reflects on our initial attempt to analyze how areas with poor air quality in Boston compare to those with better air quality, focusing on demographics such as race/ethnicity, area median income, housing density, population density, and social vulnerability. This early stage analysis was a part of our learning process in a data science class, conducted before obtaining more refined data.

Data Cleaning and Collection Process

Our project involved a multifaceted approach to data cleaning and collection, tailored to each data source's unique characteristics. The base project provided links for collecting data on air quality, transit, demographics, health, and income.

1. **Pollution Proximity Index (PPI) Data:** The PPI data was obtained through a straightforward CSV download. Our task primarily involved filtering out null values to ensure the cleanliness and usability of this dataset.

2. **Social Vulnerability Index:** Similar to the PPI data, the social vulnerability index was a dataset provided by the City of Boston. This dataset was downloaded and then cleaned for further use, ensuring its relevance and accuracy for our analysis.

3. **Air Quality Sensor Data:** The collection of air quality sensor data posed significant challenges. We utilized the AirNow API to gather yearly data for 30 zip codes in Boston. However, the API's limitations only allowed one day's worth of data per zip code every 8 seconds. Consequently, collecting a year's worth of data for all zip codes took an extensive amount of time – over 24 hours in total. To expedite this process, we employed multiple Python scripts with different API keys, reducing the data collection time to approximately 5 hours. Despite this effort, we encountered issues with the uniformity of AQI and OZONEAQI data across different days and zip codes, which is an ongoing concern we are addressing with other teams, TPMs, and our professor.

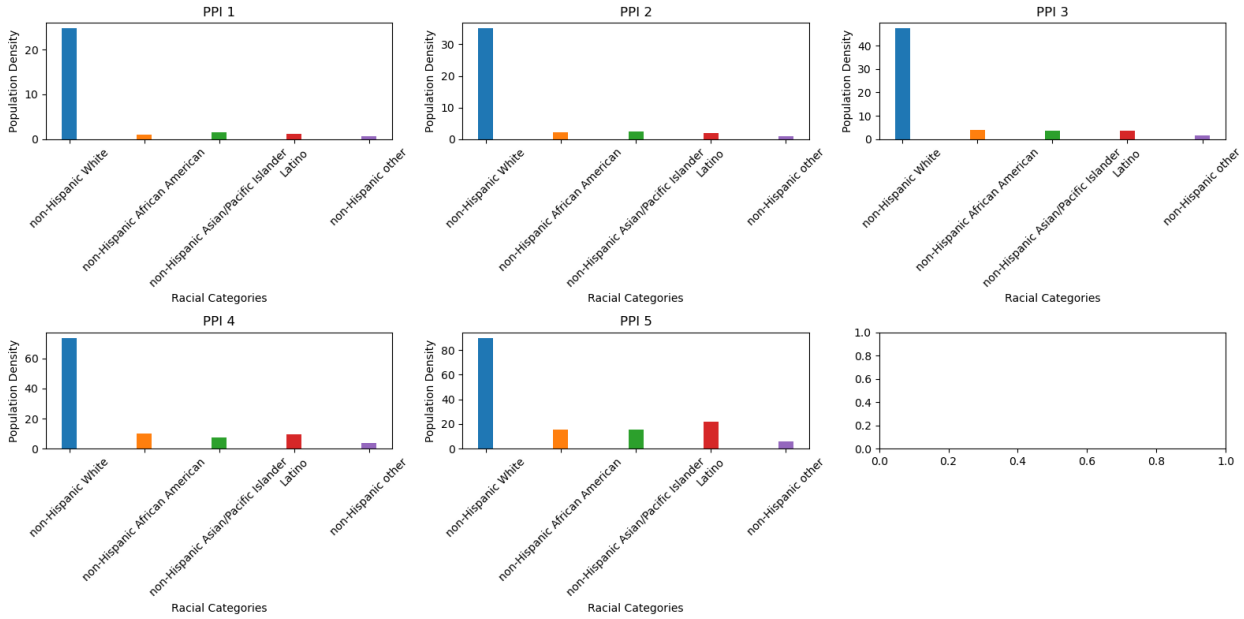
4. **Census API for Population Data:** We utilized the Census API for analyzing population data, calculating housing and population density for Boston zip codes. The results were visualized using a shapefile, creating a map of Boston's counties delineated by zip codes. The map was color-coded, ranging from tan to dark red, to indicate varying levels of population and housing density, with darker shades representing higher values.

5. **Social Vulnerability Data Integration:** For the social vulnerability dataset, we mapped the names of cities to their corresponding zip codes. This allowed us to extract relevant data from those zip codes that were also included in the air quality dataset. The analysis of social vulnerability was then conducted based on this extracted data, providing insights into how different areas of Boston varied in terms of social risks and needs.

This data cleaning and collection process was integral to our initial analysis, providing the foundation upon which our findings and insights were built. Despite the challenges encountered, particularly with the air quality sensor data, this phase of the project was crucial in shaping our understanding of the complex relationship between air quality and demographic characteristics in Boston.

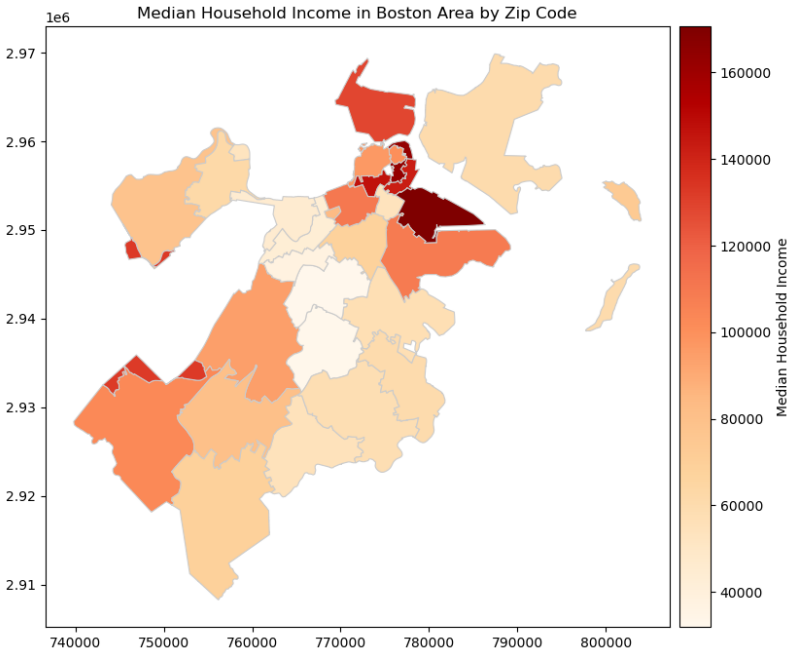
Findings

Race/Ethnicity and Air Quality



Our preliminary analysis, based on the PPI data, suggested a noticeable correlation between air quality and racial demographics in Boston. We observed an increase in the proportion of minority populations as air quality decreased. The data indicated a significant rise in the population of all racial categories, except non-Hispanic Whites, particularly between PPI levels 3 to 5.

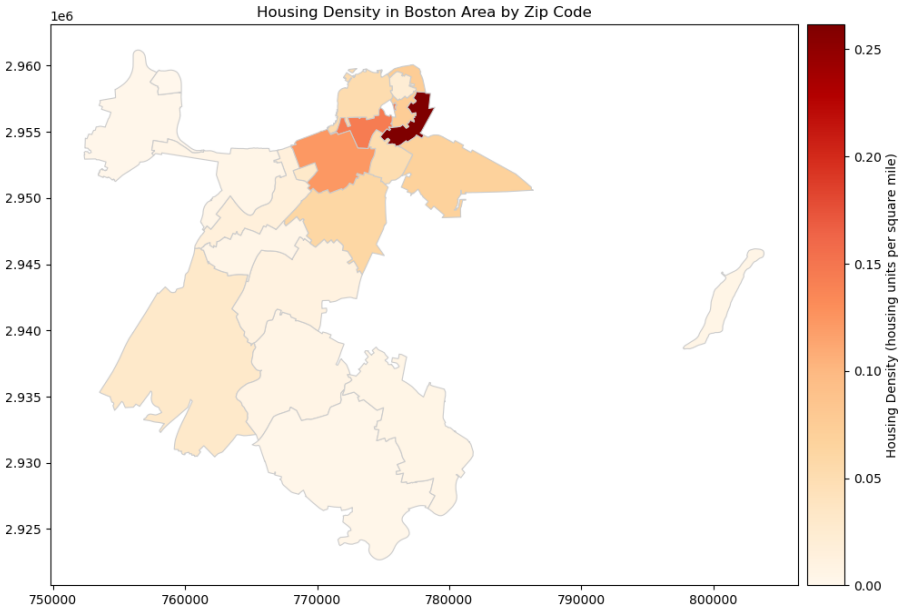
Area Median Income/Income and Air Quality



In this early analysis phase, we explored the median income across various zip codes in Boston to gauge economic disparities in relation to air quality. The K-means clustering graph

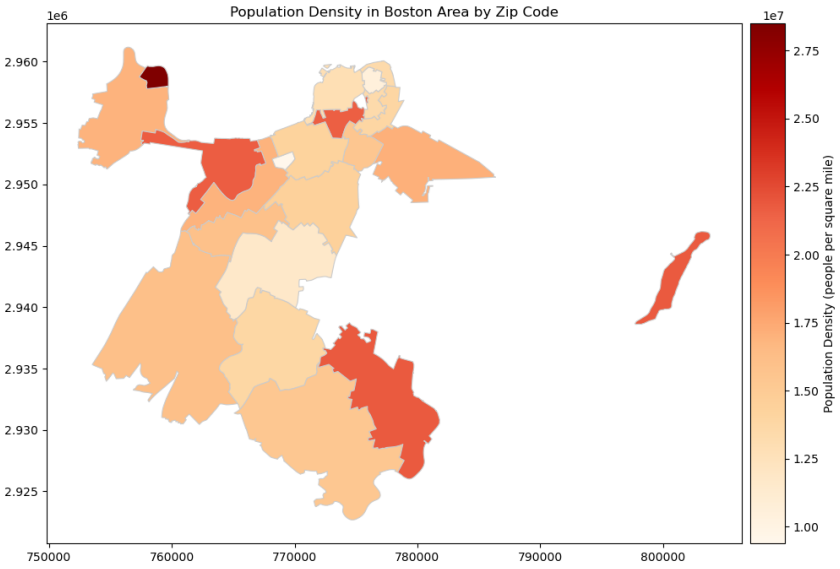
from this period showed that the air quality index did not significantly vary by zip code, mostly remaining within the 20 to 60 range. We considered the possibility that the COVID-19 pandemic's impact on transit activity in 2020 might have influenced these findings.

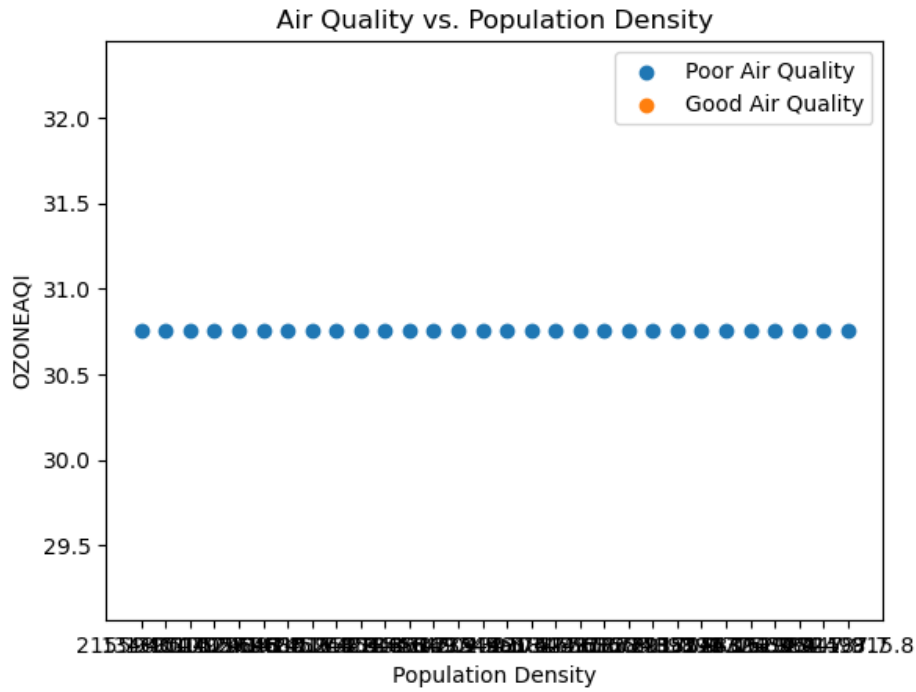
Housing Density and Air Quality



Our initial shaded map analysis of Boston's housing density revealed a potential correlation between housing density and income levels. The northeastern region, characterized by high housing density, seemed to align with wealthier areas, suggesting a potential link between higher-density areas and higher-income neighborhoods. This early finding was intriguing for its implications for urban planning and housing policy in relation to air quality and transit infrastructure.

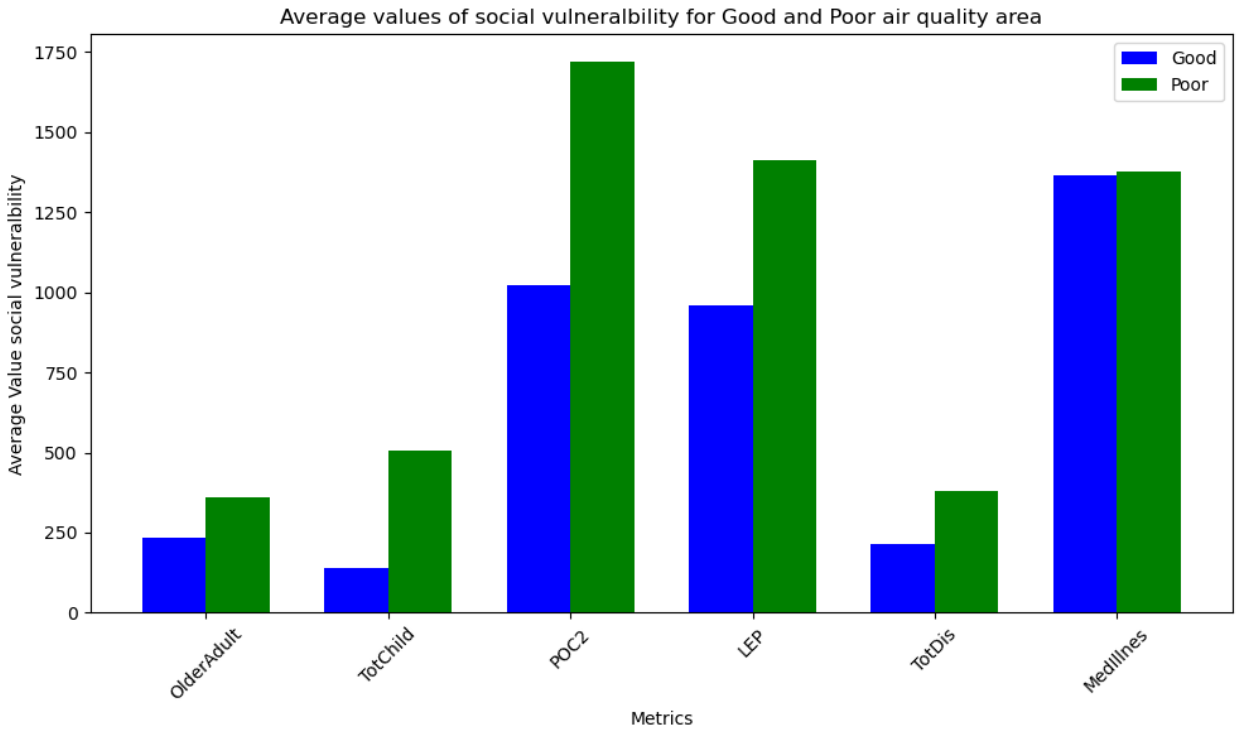
Population Density and Air Quality





Using zip codes to approximate population density, our initial findings showed that areas with higher population density did not exhibit a clear correlation with air quality indicators like OZONEAQI/PM2.5AQI. This was an interesting observation, challenging the conventional assumptions about urban density and air quality.

Social Vulnerability and Air Quality



In this phase, we compared areas with differing air quality based on social vulnerability metrics. Our preliminary data showed that metrics such as the proportion of older adults, children, people of color, and individuals with limited English proficiency were higher in areas with poor air quality. This indicated a potential strong link between poor air quality and increased social vulnerability.

Issues and Limitations

Our initial study faced several significant limitations, which are crucial to acknowledge:

1. **Data Quality and Completeness:** The quality and completeness of our data were potential limitations, particularly the omission of the influence of specific and special events. This raised concerns about the representativeness of our findings and the potential for overlooking critical variables that could affect air quality.

2. **Ambiguity of Correlation and Causation:** Another major challenge was distinguishing between correlation and causation in our analysis. This ambiguity made it difficult to draw definitive conclusions about the relationships we observed between air quality and demographic factors.

3. **External Factors:** Factors such as industrial pollution, meteorological conditions, and regional transportation patterns were not fully accounted for in our analysis. These external factors could significantly influence air quality and thereby impact our findings.

4. **Generalizability Issues:** There was also a concern about the lack of generalizability of our findings, given the specific transportation and population dynamics in Boston. This limitation was important to consider when extrapolating our results to other contexts.

5. **Sensor Data Inconsistencies:** We noticed inconsistencies in the air pollutant sensor data. Some sensors located outside Boston city limits and those near modes of transportation other than buses raised questions about the scope of our analysis. Additionally, not all sensors had complete data sets, and there was uncertainty about the necessary categories of pollutants to include and their units of measurement.

6. **Public Transit Data and Organized Dataset Needs**: The public transit information, mainly MBTA stops in Boston, was available in map form. However, the lack of an organized dataset posed challenges, particularly concerning the potential inaccuracies of human counts and the need for more structured data for analysis

.

7. **Uniformity of AQI Data**: A notable concern was the uniformity of AQI data across all zip codes for the year 2020. While our API fetching code seemed to function correctly, the identical AQI readings across zip codes were perplexing and raised questions about the accuracy and reliability of this data.

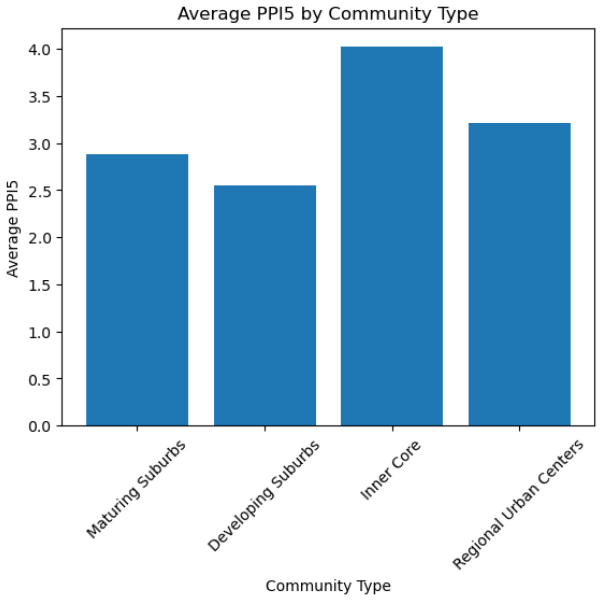
Our initial analysis offered valuable insights into the complex relationship between air quality and demographic characteristics in Boston. It highlighted the potential disproportionate impact of poor air quality on minority populations, lower-income groups, and socially vulnerable individuals. While these findings were preliminary and subject to the limitations of our early dataset, they underscored the importance of continued analysis and the need for more accurate data for informed policy-making and urban development.

Extension Project

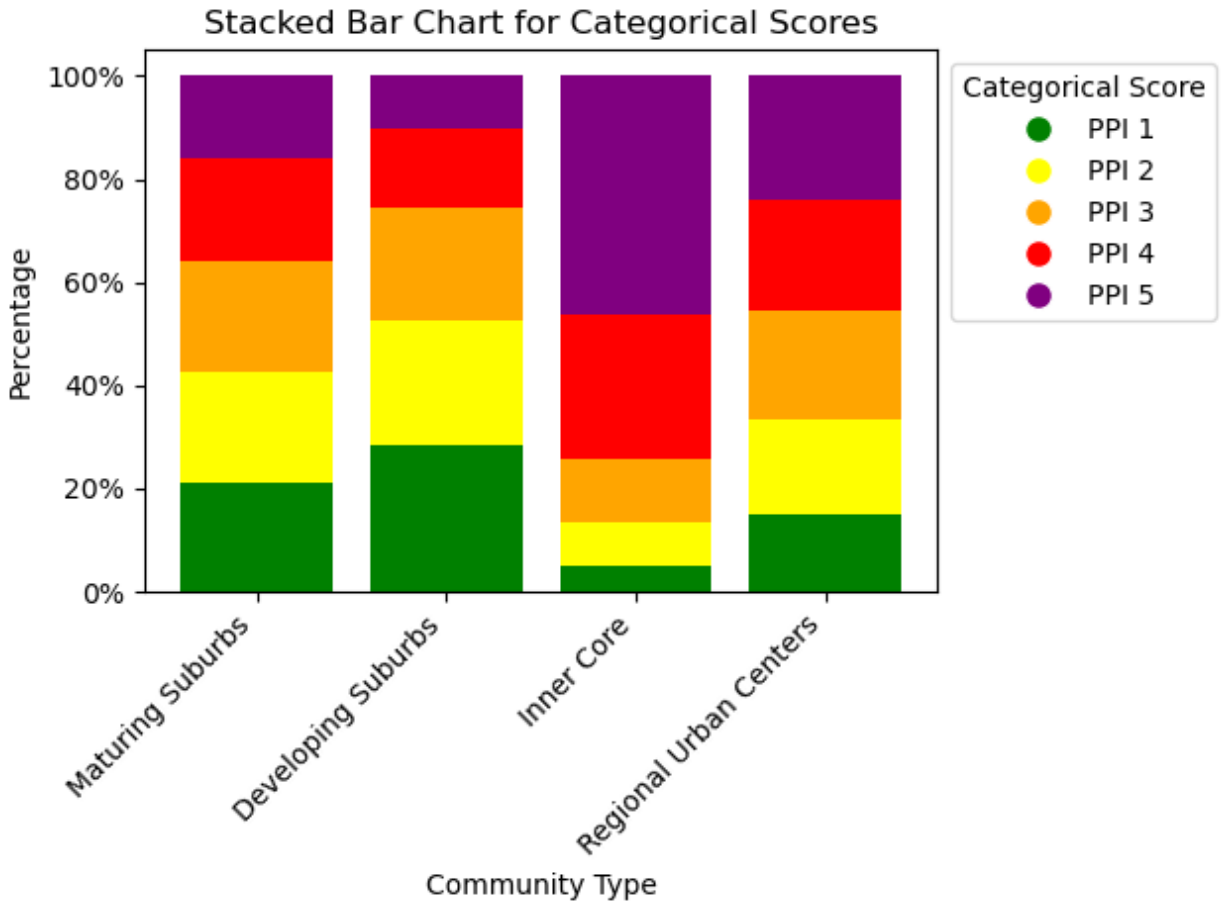
Extension Project Question: "What is the relationship between the community types (municipalities mapping to community types) in Boston and the PPI data with regards to race distributions, median income, and MBTA infrastructure?

The extension project was not an adaptation but an evolution of our analytical scope, specifically tailored to harness the Pollution Proximity Index (PPI) data provided by the Metropolitan Area Planning Council (MAPC). This new direction was embarked upon to deepen our understanding of the environmental and demographic dynamics within Boston. Our endeavor was to establish a more nuanced correlation between community types, as defined by their unique characteristics, and the air quality as reflected by the PPI data. The essence of this project was to unravel the underlying patterns that connect race distributions, income levels, and the availability and type of MBTA infrastructure across various Boston communities. By constructing a crosswalk dataset to effectively link zip codes with their respective community types, we aimed to overcome the limitations of the PPI dataset's lack of specific locational details. This approach was designed to illuminate the differential impacts of air quality on diverse communities, thereby offering a more detailed and contextualized understanding of how public infrastructure and socio-economic factors intertwine to shape the living conditions in Boston.

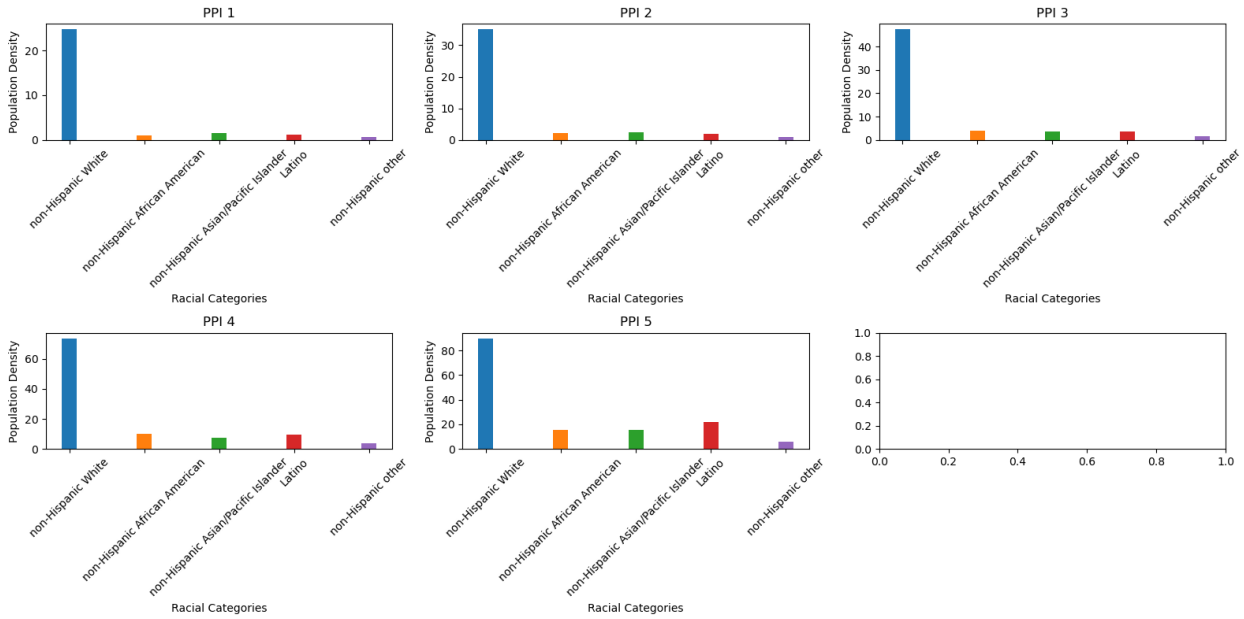
Pollution Proximity Index Analysis By Community Types



Based on the MAPC PPI Dataset, we produced the graph above, which demonstrates the relationship between the average PPI per community type. We can see that Inner Core and Regional Urban Centers (RUC) have the highest PPI compared to the suburb community types, with Inner Core being just below 4.0 and RUC below 3.5. This is due to the fact that these communities are closer to cities and with more streets, highways, and public transportation means more carbon emissions produced by vehicles. Whereas we see both Maturing and Developing suburbs have PPI of below 3.0 and 2.5 respectively. This is due to less exposure to these environments and generally having more green land than transportation infrastructure in the municipalities.

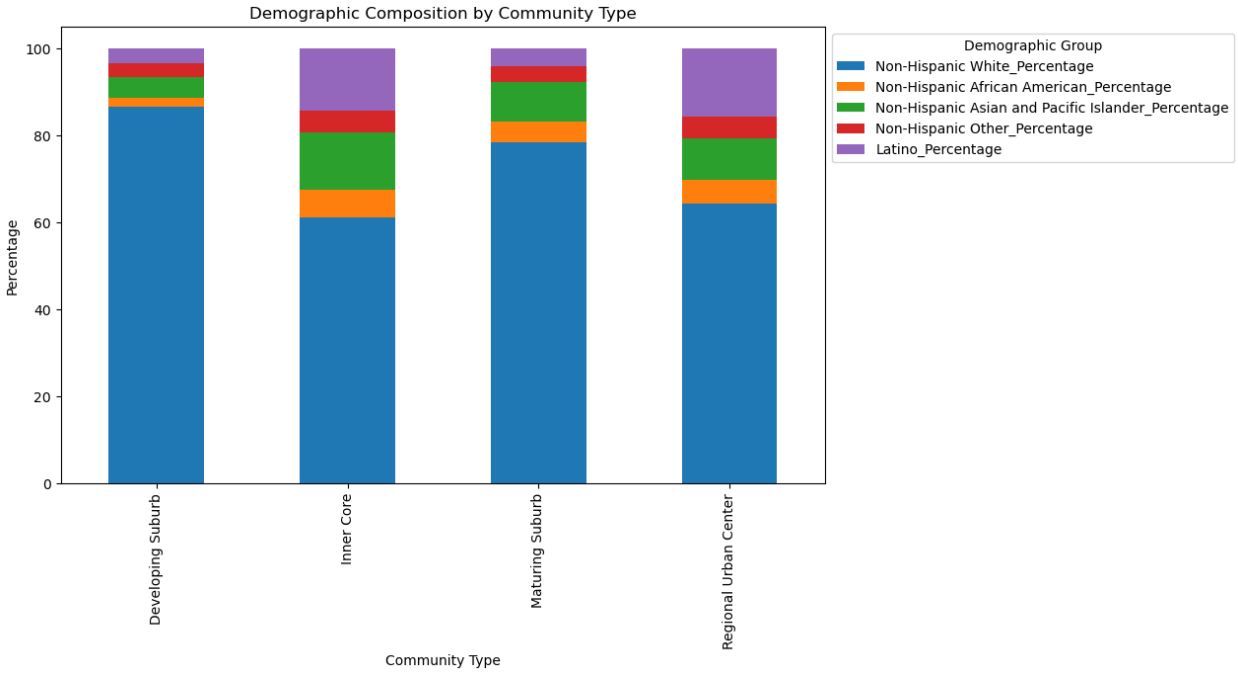


This graph also uses the MAPC PPI Dataset, and demonstrates the different categories as percentages of the total PPI in the community types. We see that Inner core has a large distribution of level 5 and 4 PPI in these communities, whereas the rest are pretty equal distributions, with more percentage of level 1 PPI for Maturing and Developing Suburbs

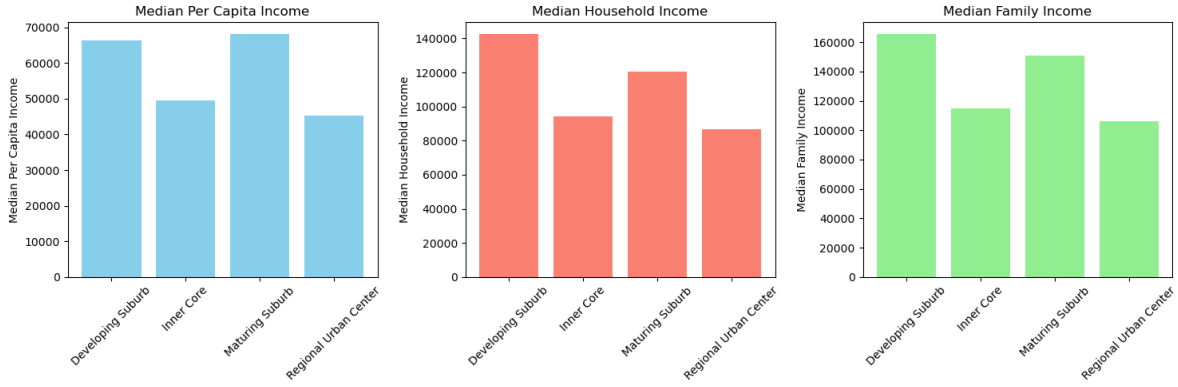


This graph also uses the MAPC PPI Dataset, and was used for the base question analysis. However, taking into account all the new analyses in this section, we provide more insights on this graph. As mentioned previously, it shows the different population densities of racial categories when the PPI level increases. The 5 different racial categories here are Non-Hispanic White, Non-Hispanic African American, Non-Hispanic Asian/Pacific Islander, Latino, and Non-hispanic other. The population density of the Non-Hispanic White seems to stay the same across all levels of PPI, due to Massachusetts being a predominantly white state. However, as the PPI level increases, we can see a clear increase in the population density of minorities (Non-Hispanic African American, Non-Hispanic Asian/Pacific Islander, Latino, and Non-hispanic other). As elaborated on in later sections, minority groups along with lower income individuals tend to be grouped into these Inner Core and Regional Urban Center areas, and are exposed to worse levels of PPI.

Demographics Analysis by Community Type



As we can see, a much higher proportion of minorities live in Inner Core and Regional Urban Center towns. This data set comes from the MAPC website that gives the racial demographics for each municipality by population number. We were able to match these municipalities with their Community Type to get exact percentages of racial breakdown per Community Type. We can see that the two suburban Community Types have an overwhelming percentage of Non-Hispanic Whites (over 80% each), while there is a much higher percentage of minority groups in Inner Core and Regional Urban Centers. One limitation from this dataset is that it gives the population per town with a margin of error. We ignored the margin of error and grouped the town by their Community Type and summed the columns before making the percentages. One limitation here could be that the total population of different Community Types could be very different which may cause the data to look a little bit skewed.

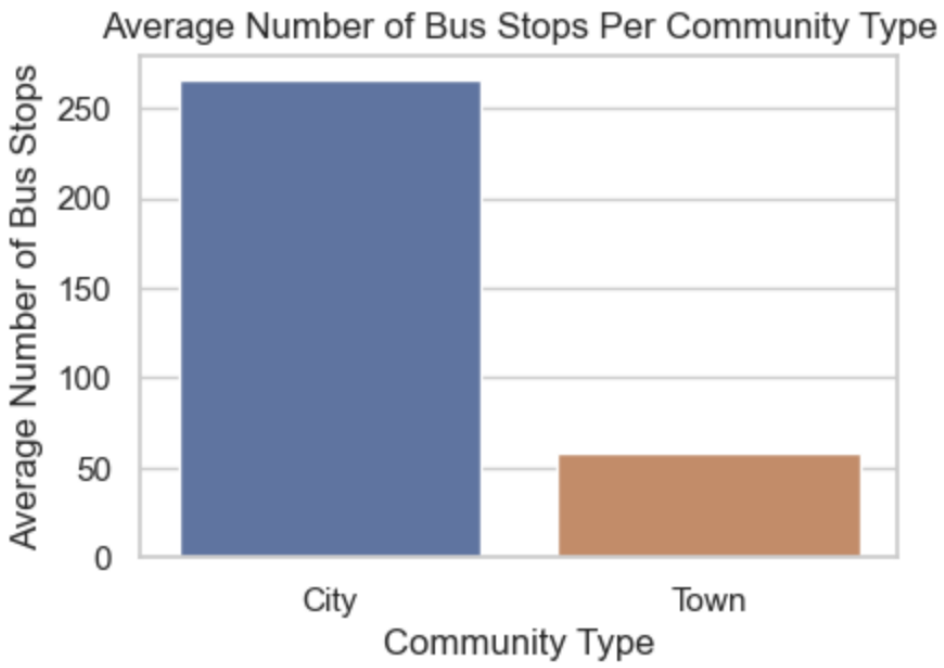


As seen here, the community types with the higher proportion of minority groups, Inner Core and Regional Urban Centers, have significantly lower income levels than the two suburban categories. The data for these graphs was scraped from a public source, and was joined with the same data set as the demographics. One potential shortcoming of the data that impacts the appearance of the graph is that the maximum of the scale was $250,000. If there was a certain town that had a medium income higher than that, it was listed at $250,000+, which had to be cleaned to simply 250,000. This means it's possible that the disparities between the suburban categories could be higher, or it's possible that the disparities are not as bad as it seems.

MBTA Infrastructure Analysis by Community Types

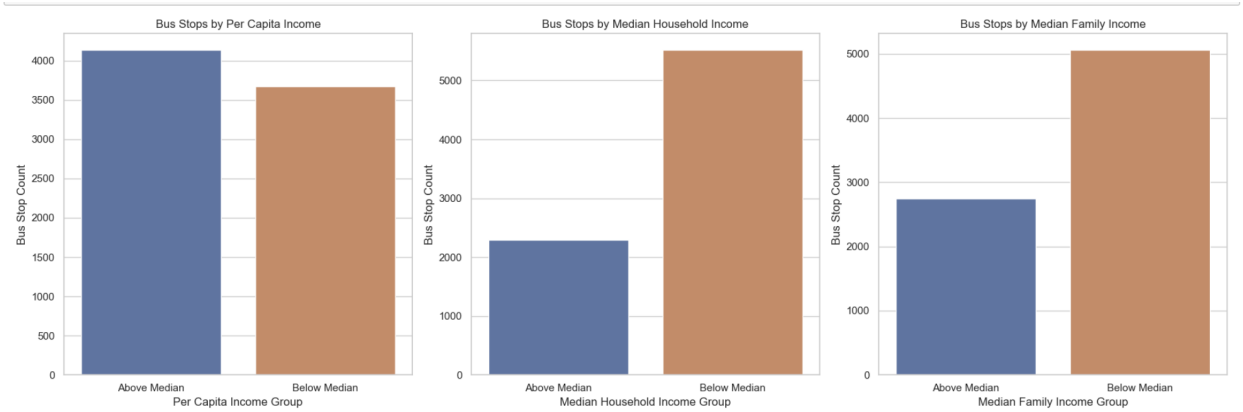
Process: To generate the graph of average number of bus stops per community type, I categorized each bus stop's city/town to the city group or town group by comparing the bus stops dataset and cities and towns dataset. Then I counted the average number of bus stops for each group and generated the graph.

Finding: We can see that the average number of bus stops in cities is a lot more than in towns. This could have some relationship with the air quality levels.



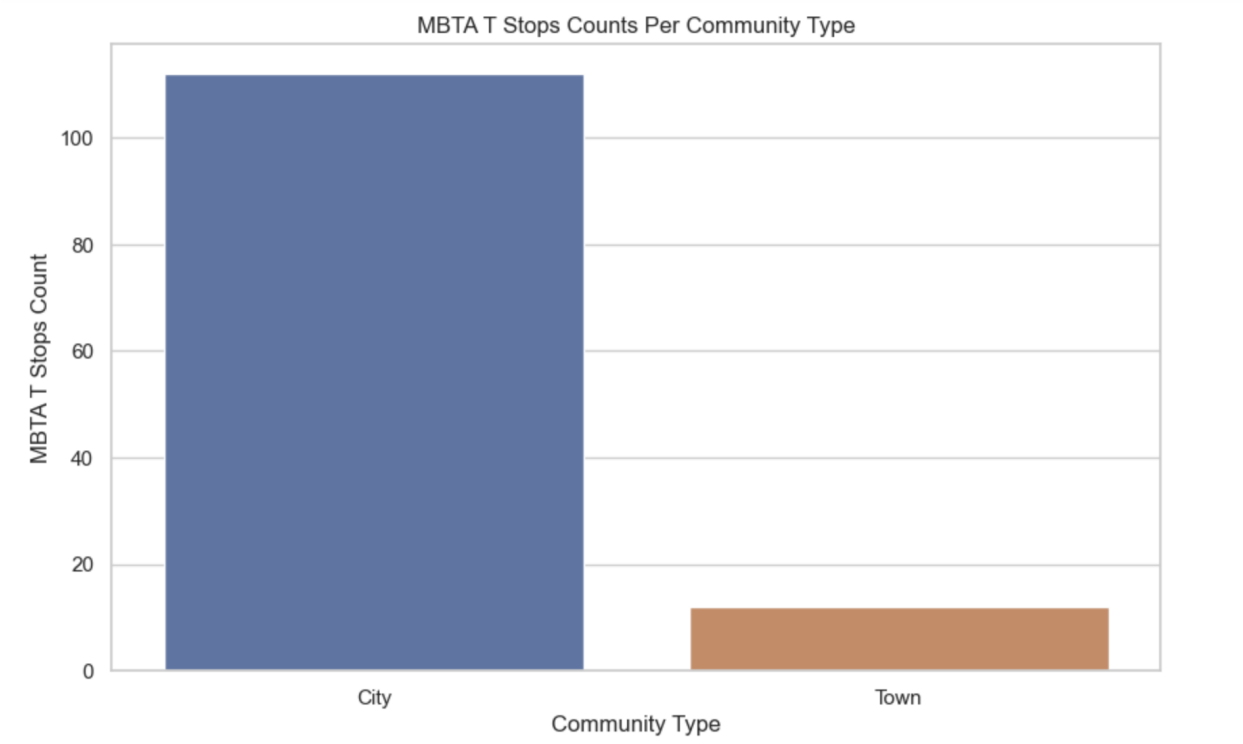
Process: There are three aspects of income data we examined, which are per capita income, median household income, and median family income. In each category, I found its median value, and calculated the counts of bus stops in municipalities with above median income and below median income.

Finding: We can discover that in median household and family income dataset, high income households or families municipalities tend to have less bus stops. However, in the per capita income analysis, we found out the above median per capita income municipalities tend to have more bus stops than below median municipalities have.



Process: To produce this graph, I used the MBTA train stop dataset that we created and calculated the counts of T stops in municipalities and put them into the city and town categories.

Finding: Similar to the result we got in bus stop analysis, this graph illustrates that city has more subway stations than towns.



Conclusion

Our exploration into the intersection of air quality, transportation, and demographic characteristics in Boston culminates in a series of insightful discoveries and significant implications. The data we've analyzed paints a telling picture of how urban infrastructure, notably transportation systems, correlates with variations in air quality across different Boston communities. This study not only highlights the environmental challenges faced by urban areas but also brings to light the social disparities embedded within these challenges.

Through our initial analysis and subsequent extension project, we uncovered a notable relationship between areas with poor air quality and higher concentrations of minority populations. This finding is crucial in understanding the broader implications of environmental justice and public health in urban settings. It signals a clear need for targeted policy interventions, especially in regions where vulnerable communities are disproportionately affected by poor air quality.

The extension project provided deeper insights, particularly in understanding how different community types within Boston experience varying levels of air quality. The nuanced correlation we established between community types, as delineated by unique characteristics, and the PPI data contributes to a more comprehensive understanding of urban environmental dynamics. It lays the groundwork for more informed and equitable urban planning strategies that prioritize the well-being of all residents.

However, our research journey was met with certain limitations, notably in data accuracy and the scope of analysis. These limitations underscore the importance of continuous improvement in data collection methodologies and analytical techniques in urban studies. They also highlight the potential for further research, particularly in refining the understanding of how

socio-economic factors and public infrastructure intersect to shape the air quality and, by extension, the quality of life in urban communities.

Our study, while limited in scope, provides a critical stepping stone towards a more holistic understanding of the complex interplay between urban environments, public health, and societal dynamics. The insights gained from this research are not only academically enriching but also carry significant implications for urban policy and planning, paving the way for creating healthier, more equitable cities.

Individual Contributions

Naveen Vaidyamath

Naveen delved into race-related data from Boston, examining demographic information vital for understanding how different racial and ethnic groups are affected by biking and air quality. Further, Naveen explored the means of travel to work data to gain insights into daily transportation methods. His efforts in coding led to the generation of plotted maps using shapefiles and the development of a script for accessing the 2022 census API. Naveen's work on population data, housing information, and zip code details contributed significantly to datasets related to housing and population density. He also initiated work on focusing on the interrelation of population, housing, and air quality. He took a role in crafting much of the deliverable presentations and the outline and base project components of the final paper.

Chao-Jen (Jackson) Chiu

Jackson focused on studying the Technical Memorandum for MAPC Research Brief and summarizing its content, particularly concerning racial disparities and air pollution. He utilized the air quality API to create the required dataset and played a key role in Deliverable 1, focusing on population density and its correlation with AQI data. Jackson collaborated with Naveen for analytics work and attended Professor Galletti's office hours with Jonathan to discuss and clarify the extension project proposal. He also took on the introduction, and contribution section of the paper.

Zixuan (Cathy) Wang

As the team leader in the initial phase, Cathy explored the Public Transit of Boston data and analyzed how MBTA stops might influence air quality and pollution in Boston. Her leadership was instrumental in dividing the team into groups focused on different project questions. Cathy was responsible for the social vulnerability part, generating a bar chart to show social vulnerability metrics in areas with varying air quality. She also drafted responses for Checkpoint A and strategized on utilizing data for other project questions.

Jonathan Suarez

Jon dedicated his time to analyzing the social vulnerability index dataset, identifying connections between different population metrics. He created Python scripts to discover correlations between vulnerability factors and area characteristics. Jon also contributed significantly to fetching yearly air quality data for 30 zip codes (which took 20 hours) and analyzed relationships between air quality and median income per zip code using k-means clustering. He also helped construct the extension question, communicated and gathered feedback from Professor Galletti and other project teams, and help build the outline of this paper.

Sinforiano (Sammy) Terada

Sammy focused on understanding the background information on air pollutants data. He assisted the team in fetching air quality data and analyzed the PPI data for Deliverable 1, particularly addressing the race/ethnicity aspect of the project. Sammy's involvement was crucial in organizing the next steps for developing the extension project and conducting preliminary analyses. He also scraped the income data and made many of the visuals for the final report.