Problem Statement:

Air quality is a critical concern for the wellbeing of Boston residents. Poor air quality could be detrimental to public health, causing respiratory issues such as Asthma. To address the issue of air quality around Boston, we are looking at how public transportation can play a crucial role in reducing emissions and increasing air quality. We aim to look at Boston divided into neighborhoods and see the factors behind the variability in air quality to help make guided decisions on where to focus efforts and policies to enhance air quality.

Data Processing Steps:

For each of the datasets below, we cleaned the data by replacing null values by the mean, median, or getting rid of them depending on the context. Additionally, we added relevant data to our final combined summary by neighborhood csv file and displayed data on a folium map for easy visualization (neighborhoodmap.html)

Proximity to Roads (PPI Index)

- Download CSV file to analyze the data
- Understood data definitions, looked at relation between community types and ppi

Air Quality Sensor Data (Google Maps)

- Pulled Sensor Data from Google Maps API by zip code. Has many more sensors to access data from compared to AirNow but is limited to 30 days history.
- Merged the available files and linearized the data such that each day's data for a given zip code is one a single line.

MBTA Transit Data

- Pulled data about MBTA Stops across Boston into a geojson file.
- Mapped the data to understand the spread of stations across the city.

Census Data:

- Pulled 2017 - 2022 (excluding 2020) ACS census data from U.S. Census Bureau for every census tract and Boston neighborhood in Suffolk County.

Social Vulnerability

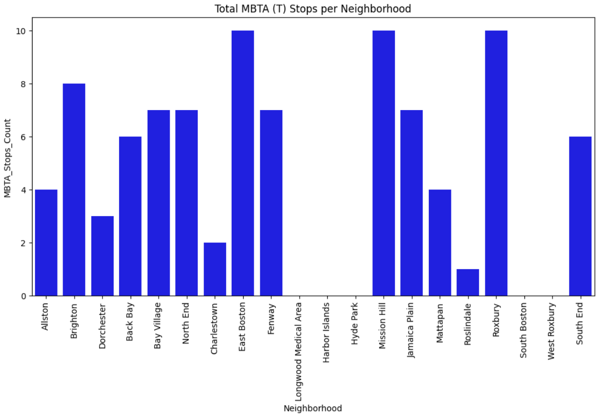
- Acquired housing and population density from Analyze Boston.
- Mapped Social Vulnerability data onto a folium map to see how it varies across the city and make use of the geojson file.

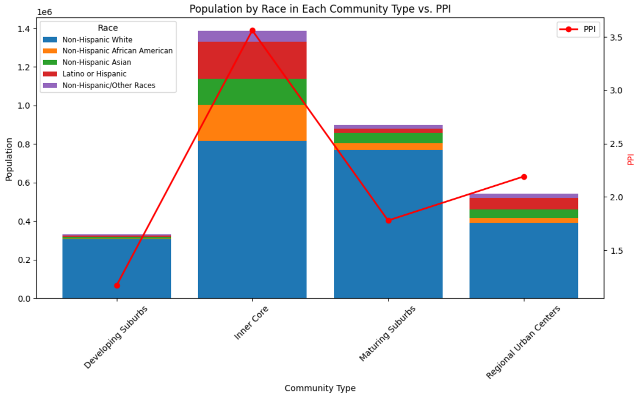
CDC Places Health Data:

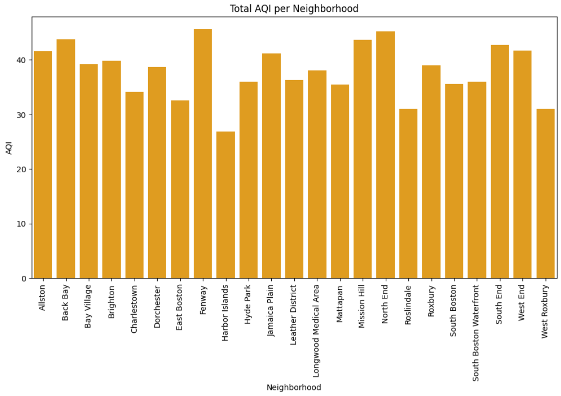
- Found health data divided by zip code.
- Filtered data by neighborhood to understand prevalence of health conditions vs AQI.

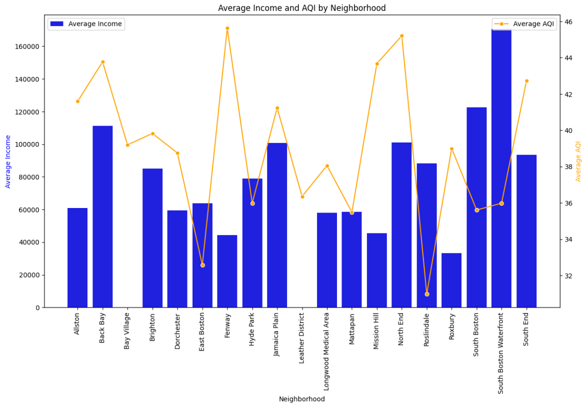
Exploratory Data Analysis:
https://github.com/BU-Spark/ds-boston-transit-air-quality/tree/team-f-dev (Working Branch) - All newly refactored data since del. 1 is in the Google API folder and Health Data folder.

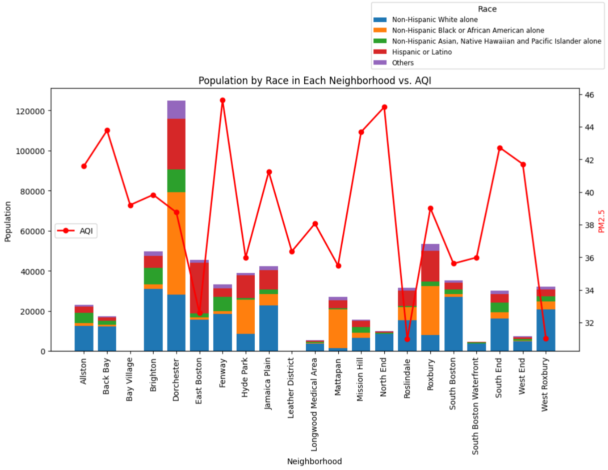
Our exploration of the above data sets yielded a few observations of note. This included the fact that some areas did not have MBTA stops. We came to this graph after having divided the data by tracts so there may be minor variations in locations by neighborhood. These areas in general did not have more than one MBTA stop. Additionally, we were having difficulty mapping the PPI dataset but were able to graph it to give a general overview of the insights. We also took a look at factors such as Race, Income, and AQI per neighborhood. (More Explained Later)











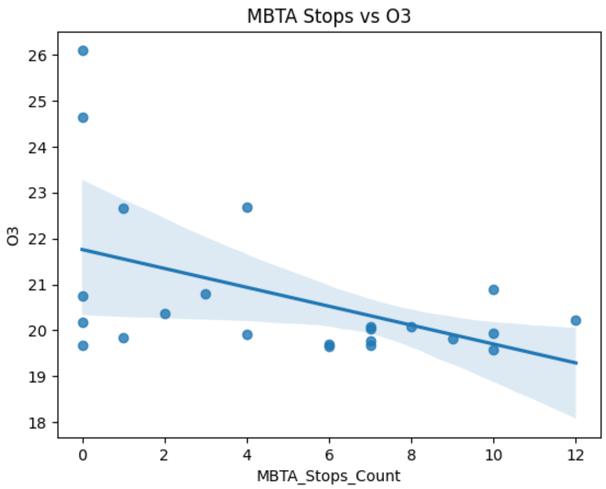
Base Questions:

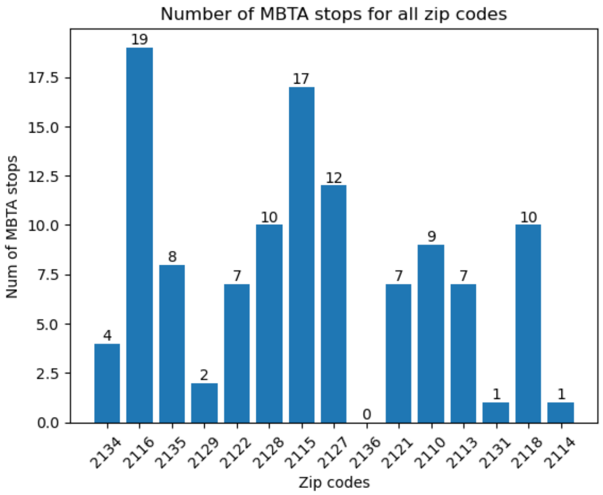
Q1: What is the yearly change in air quality for Boston residents based on their proximity to different types of transportation infrastructure, specifically, proximity to public transportation options or proximity to roads?

In deliverable 1, we performed an analysis between air quality and one popular public transportation in Boston which is the MBTA. We analyzed the AQI_OZONE and AQI_PM2.5 (data from AirNow API) versus the number of stations in that area. However, after we revised it, we found the raw data as below, but some areas are having the same zip code (e.g, Back Bay and Bay Village are having the same zip code 2116.

In deliverable 2, we fixed this issue and grouped the data by zip code. Then, we got the data as below. In addition, we gathered data from the Google Air Quality API about more detailed information about the air quality in the past 30 days. To prevent having multiple points for zip codes that don't have any stops, we truncated the data according to the zipcodes in the above graph. Then, we plot the Ozone levels with respect to the number of stops. It seems to have a slight negative relationship between Ozone levels and Stops. When we calculated the r-squared value, it yields -0.531.

However, we also found that the AQI has a slight positive correlation with the number of MBTA stops with a r-square value of 0.493. It seems that building more MBTA might lead to bad air quality, but there are other factors that should be considered. There may be other factors such as traffic density because more transportation services often occur in downtown areas with large traffics.





PPI Data:

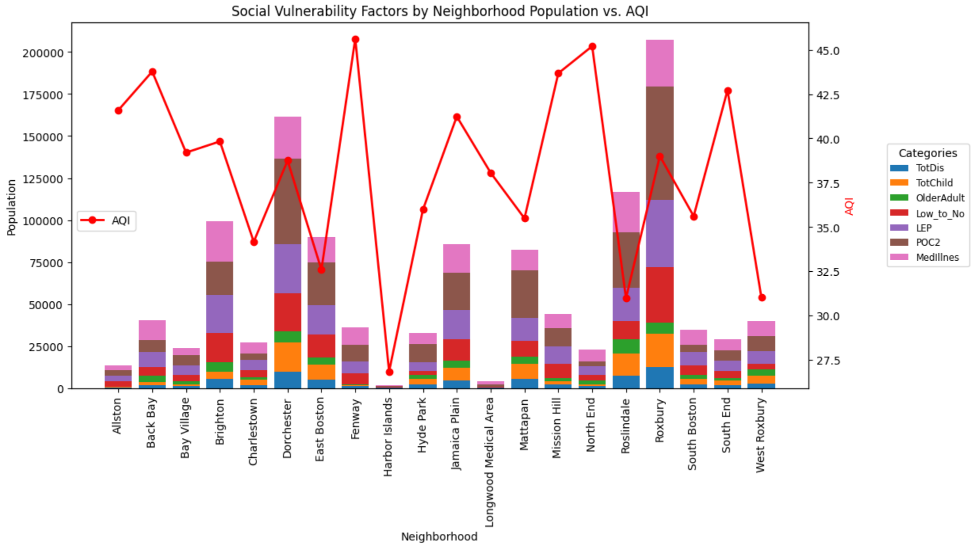
- **Developing Suburbs:** Outskirts of the city that are of high growth. Lowest Population, Lowest PPI.
- **Inner Core:** Central Part of the City. Highest Population, Highest PPI.
- **Maturing Suburbs:** Outskirts of the city that have undergone significant development and have matured over time. Second highest Population, Second Lowest PPI.
- **Regional Urban Centers:** Regions within a city that serve as an economic, cultural, or administrative hub. Second Lowest Population, Second Highest PPI.

In general, areas with higher population have higher PPI, however, it could be assumed that Regional Urban Centers have more PPI due to movement or people commuting to these areas for work.

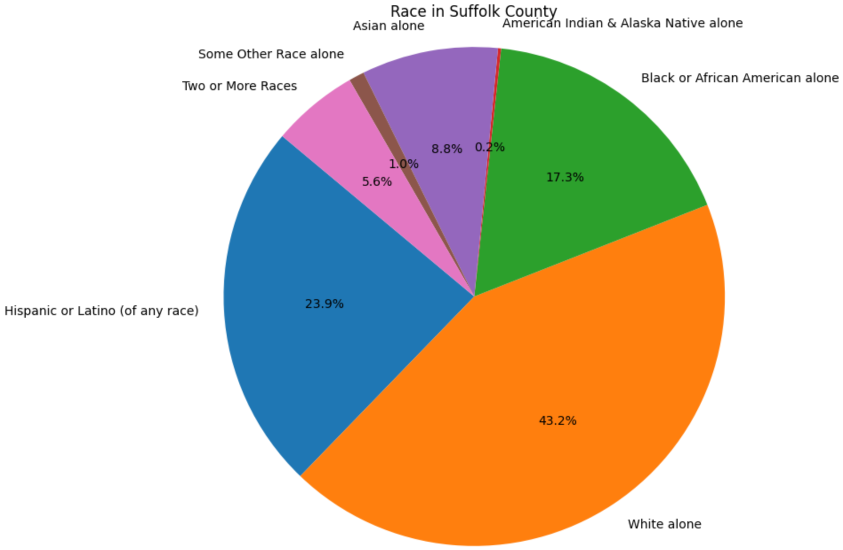
Q2: How do areas in Boston with poor air quality compare to areas with better air quality based on different demographic characteristics?

In Deliverable 1, we conducted an initial assessment of air quality in various neighborhoods within Boston, focusing on Mattapan, Roslindale, West Roxbury/Roxbury, and Hyde Park. At that time, the data indicated that these neighborhoods had a red Air Quality Index, suggesting poor air quality. However, we now have access to updated information from the Google Air Quality API, which shows that all neighborhoods in Boston currently have good air quality. However, it's important to note that this new data only covers the past 30 days, and it may not provide a comprehensive view of air quality trends over a longer time frame Therefore, it's crucial to consider the potential impact of a more extensive dataset when analyzing the relationship between air quality and various demographic characteristics. The information we found for housing and population density was not helpful to answer this question.

Based on the Social Vulnerability Factors by Neighborhood Population vs AQI below, it seems like people of color are a common vulnerability factor in each neighborhood.



Based on the 2022 American Community Survey 1-Year Estimates from the Census, the median household income for Suffolk County is $85,358. Based on the 2022 American Community Survey 1-Year Estimates from the Census, this the race demographic in Suffolk County:

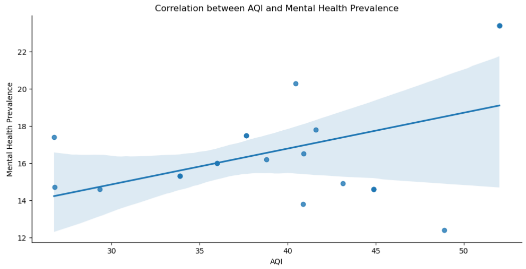


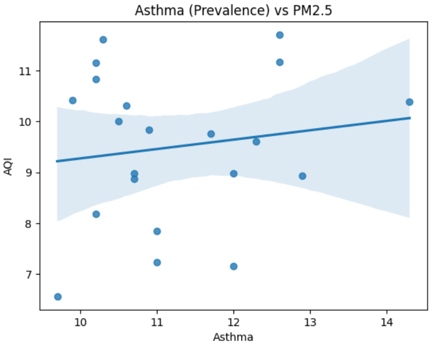
Q3:What is the relationship between health data and what are the trends in yearly change in air quality for Boston residents by neighborhood, zip code.

We used Air Quality Sensor Data and an integrated health dataset for our analysis. The air quality data includes measurements such as the AQI, as well as parameters such as ozone and

PM2.5, recorded by Boston neighborhood or postcode. The exploratory data analysis corrected previous data inconsistencies by grouping data by zip code and incorporating detailed air quality information from the Google Air Quality API. The health data cover a variety of health indicators, including the prevalence of medical conditions, as well as demographic and socio-economic indicators for different neighborhoods in Boston. Exploratory analyses were conducted to generate summary statistics and visualizations to understand the distribution and trends of air quality and health indicators. We conducted a series of correlation analyses to explore the relationship between AQI and a variety of health indicators, including mental health prevalence, chronic diseases, and lifestyle factors.

After calculating the correlation values between AQI values and various health indicators, we obtained the following results:





The analysis showed a statistically significant moderate positive correlation between AQI/PM2.5 and the prevalence of mental health problems (MHLTH_CrudePrev and Asthma). This suggests that areas with poorer air quality (higher AQI) may have a higher prevalence of mental health problems. Other health indicators, including those related to chronic physical health conditions such as COPD, coronary heart disease and stroke, as well as lifestyle factors such as smoking and obesity, are negatively correlated with the AQI.This correlation suggests that neighborhoods with higher AQI levels may have a higher prevalence of these health conditions.
The analysis showed a significant moderate positive correlation between AQI and the prevalence of mental health problems,areas with poorer air quality (higher AQI and Pm2.5) may have a higher prevalence of mental health and Asthma problems.

Extension Proposal:

In pursuit of a comprehensive understanding of factors influencing air quality in Boston, it became apparent that the lack of variation in air quality was proving to be a problem. The AirNow API only had three active sensors and even with 5 years of historic data, showed very little variation in air quality between areas. Therefore, for deliverable 2, we had focused on refactoring our code to make use of the Google Maps Air Quality API. This specifics regarding the dataset has been explained in previous sections.

It also became apparent that some key elements with regards to public transport and infrastructure. We were initially provided with MBTA Transit Data, but this only covered the MBTA Subway (T) System, it did not cover buses - a crucial aspect of public transport. Busses help reduce cars on the road by packing more people per vehicle and in turn help reduce emissions. We plan to explore how the presence of bus stops and routes affect the air quality in a given neighborhood.

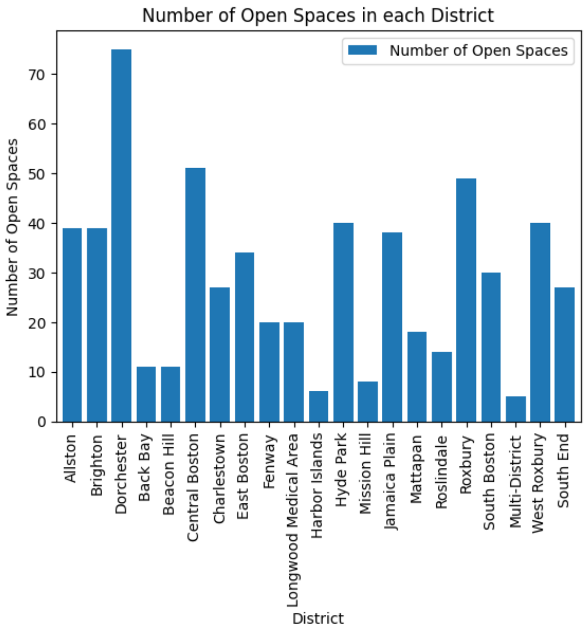
Additionally, people tend to make a conscious decision to use bikes instead of cars to help reduce their carbon footprint. Boston has a robust shared bicycle system. Therefore, we thought it would be a helpful contributing factor to see not only air quality but also health factors in each neighborhood.

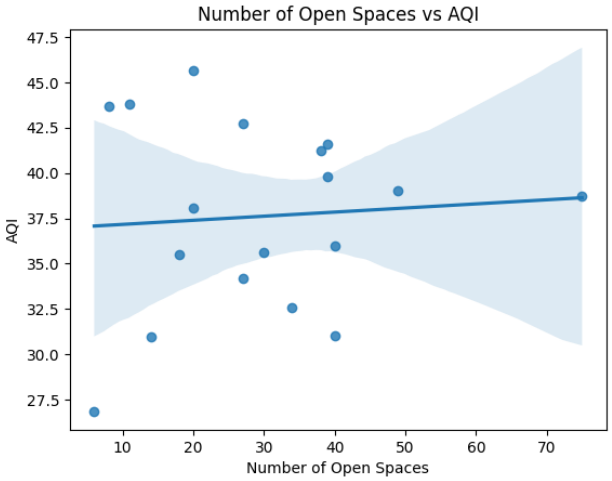
Finally, as proposed by the professor in discussion during office hours regarding the lack of substantial data and correlation for Air Quality, we decided to make use of proxies to help assess air quality across Boston. One proxy is the open green spaces, encompassing parks, playgrounds, and conservation areas, across Boston's neighborhoods. These spaces play a pivotal role in enhancing air quality in urban environments because they act as natural filters, capturing particulate matter and reducing airborne pollutants. They also serve as carbon sinks by absorbing carbon dioxide through photosynthesis. By analyzing the distribution, accessibility, size, and quality of green spaces using data from Analyze Boston, we aim to evaluate their potential impact on air quality improvement within different Boston neighborhoods. This analysis will help identify correlations between green space presence and air quality, guiding urban planning and policy decisions to prioritize the preservation and expansion of these areas for the benefit of public health and environmental sustainability.

We are also examining the zoning classifications within each neighborhood, categorizing them into commercial, residential, and industrial zones. These zoning categories are essential factors that influence air quality. Commercial areas often generate increased traffic and emissions, residential zones may be susceptible to localized pollution sources, and industrial districts can be sources of various pollutants. By analyzing zoning data, we intend to identify areas with distinct zoning characteristics and evaluate their potential contributions to variations in air quality. This additional layer of analysis will help us gain a more understanding of the relationship between land use and air quality dynamics across Boston's neighborhoods.

**Note*: Although we initially planned to utilize zoning as a proxy for assessing air quality, we encountered a limitation – the zoning data available to us was not recent and dated back to the early 2000s. This gap in the data restricted our ability to provide a current assessment of zoning's impact on air quality. We plan to dig further to find up-to-date relevant information to help factor zoning data and its influence on air quality in Boston's neighborhoods.

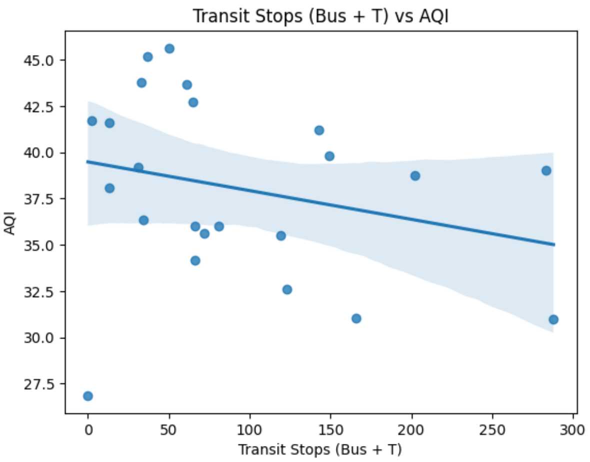
**Note:* The Analyze Boston data set for open spaces is divided by districts rather than neighborhoods. Since there are over 500 open spaces, it is hard to pinpoint what open space belongs to which neighborhood. However, for the purpose of consistency of our current graphs, we split the districts into neighborhoods. For example, Allston-Brighton is a district, but we split the information to represent the Allston neighborhood and the Brighton neighborhood. We did this for the Allston-Brighton, Back Bay-Beacon Hill, and the Fenway-Longwood districts.

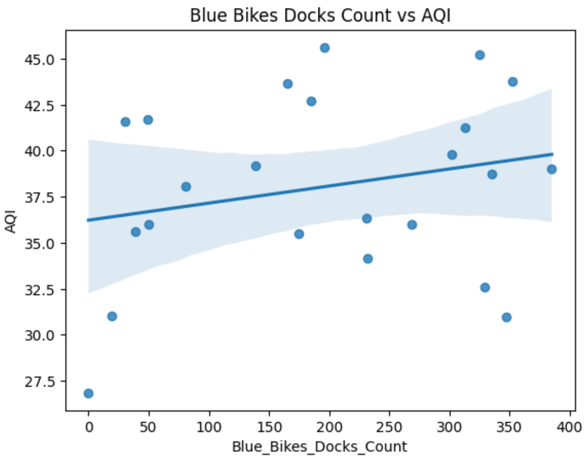




Multi-district applies to open spaces that go through multiple neighborhoods like the Charles River Reservation.

At first glance, open space data showed a very weak positive correlation between the number of open spaces and AQI, however, open spaces could also mean undeveloped land, and therefore we believe that looking into finding a relevant green space data set could provide us with more helpful information.





Our initial analysis of MBTA T stops vs AQI concluded that neighborhoods with more T stops had worse air quality. This was not expected. The use of bus stop data helped clarify our findings about AQI vs Transit Stops. There are many more bus stops spread out across the city. As a result, we saw an improvement in air quality when there were more stops present.

On the other hand, as the number of blue bike docks increased, the air quality worsened. The relationship here had a weak positive correlation, meaning that it is not a direct causal relationship. There would be many other factors contributing to the AQI.

In conclusion, as we try to find more relevant and accessible data to more accurately come to meaningful conclusions, we think that the expansion of our scope to use Bus Stop data was a step forward in understanding the intricate relationship between Boston's public transit and its Air Quality. While we did not find a meaningful relationship between Blue Bike docks, we believe that further research with regards to health and green space data could help us come to better conclusions. We plan on supporting our open space data research with specific information with regards to green spaces available, and we plan on looking for reliable and up to date information on Boston's zoning by neighborhood to create a proxy to air quality to help support the initial lack of data from the AirNow Air Quality API.

Individual Contribution

Anulika Nnadi:
For the presentation, I made the project motivation, data collections, and q2 insights slides. Using some of the graphs that Hemanshu created, I also made the visualization & insights for the extension slide. I helped with research for the extension project and found the dataset for open spaces in Boston from Analyze Boston and created the corresponding graph after manually processing/splitting the data. I worked on the data processing steps section and the extension project section. I tailored Q2 to match our new data and created the graph for the race demographics in Boston.

Ziliang Wang: In Deliverable 2, I was responsible for collecting Google air quality data to help the group answer the main question, while I used the data to answer the main question 3, as well as creating a portion of the PowerPoint for the presentation, where I was responsible for the limitations/challenges as well as the questions 3 sections.

Xinzhu Liang:
For this deliverable, I and Ziliang collected the data from the google air quality api from Oct 19 - Nov 17. Also, I updated question 1 to include analysis about the google air quality data.
For the presentation, I am responsible for the data collection page, and the question 1 insight page.

Hemanshu Bhojwani:
I worked on refactoring all the code from deliverable 1 to work with the new Google Air Quality data. I processed the health data set and the extension data sets on top of the air quality, ppi, social vulnerability, and transit data for deliverable 1. I created graphs relating to the extension project, as well as health, AQI vs Stops, social vulnerability, and income graphs. In the report, I worked on the problem statement, data processing, EDA, PPI and extension sections and I created a summary neighborhood map in folium. In the presentation, I worked on the conclusion and extension slides.