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:

Proximity to Roads (PPI Index)
- Download CSV file to analyze the data
- Understood the key to the document and pulled key insights from it.

Air Quality Sensor Data
- Pulled Sensor Data from AirNow Api by zip code
- 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.
- Combined census data with air quality data.

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.

Exploratory Data Analysis: (attached with this file on gradescope and on github)

https://github.com/BU-Spark/ds-boston-transit-air-quality/tree/team-f-dev (Working Branch)

https://github.com/BU-Spark/ds-boston-transit-air-quality/tree/team-f (Main Branch)

- **DataByNeighborhood.csv:** Merged data for Social Vulnerability, MBTA Stop data, and Air Quality by neighborhood.
- **/Merged Data/mergedmap.html**: Combined HTML Map on folium to visualize the Air Quality, Social Vulnerability, and MBTA Stop data.
- Folders have an **.ipynb** with some exploratory data analysis for each data set.
- **Graphs** added within this report

Metric Definitions:

- **PPI:** A score that takes into account the five factors: the volume, speed, and fleet mix of traffic on a given roadway; distance of residents to that roadway; and proximity to other high-emitting roadways nearby.
- **Particulate Matter:** Tiny solid or liquid particles suspended in the air
    - **PM2.5**: Particulate Matter 2.5 micrometers or smaller
    - **PM10**: Particulate Matter 10 micrometers or smaller
- **OZONE:** Ozone is a molecule composed of three oxygen atoms.

For each of the metrics above, the lower the score, the better.

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?

Based on the air quality data from 2023 for various Zip Codes in Boston and their respective number of MBTA stops, we found a moderate negative correlation between ozone levels (AQI_OZONE) and the number of public transportation options (MBTA_Stops_Count). This suggests that areas with a greater number of public transportation options may have better ozone air quality. However, no significant relationship was found between PM2.5 levels and the number of MBTA stops, as indicated by a negligible correlation.

The spatial analysis has produced a correlation matrix between the average Air Quality Index (AQI) for ozone and PM2.5 and the number of MBTA stops within each ZipCode. Here are the key points from the correlation matrix:

There is a moderate negative correlation between the average AQI_OZONE and the MBTA_Stops_Count $r = -0.578$. This suggests that areas with more MBTA stops tend to have lower average ozone levels, which might imply that increased availability of public transportation could be associated with better ozone air quality.

The correlation between average AQI_PM2.5 and MBTA_Stops_Count is negligible $r=-0.006$, indicating no clear relationship between PM2.5 levels and the number of public transportation options.

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.

How do areas with poor air quality compare to areas with better air quality based on different demographic characteristics, specifically:

- **Race/ethnicity (ACS)?**
- **Area median income/ income**
- **Housing density**
- **Population density**
- **Social vulnerability**

- Zip Codes with Red AQI (Unhealthy for Sensitive Groups):
    - 2126 - Mattapan
    - 2131 - Rosindale
    - 2132 - West Rosindale/Roxbury
    - 2136 - Hyde park

1. **Mattapan**
    a. Race/ethnicity (ACS)
        i. Has a majority Non-Hispanic Black or African American alone population
    b. Area median income/ income
        i. 58k
    c. Housing density & Population density - the information found wasn't helpful
    d. Social vulnerability
        i. The most vulnerable groups is People of Color
2. **Rosindale**
    a. Race/ethnicity (ACS)
        i. Has a majority Non-Hispanic White Alone population
    b. Area median income/ income
        i. 88k
    c. Housing density & Population density - the information found wasn't helpful
    d. Social vulnerability
        i. People of Color
3. **West Rosindale/Roxbury**
    a. Race/ethnicity (ACS)
        i. Has a majority Non-Hispanic White Alone population
    b. Area median income/ income
        i. 106k
    c. Housing density & Population density - the information found wasn't helpful
    d. Social vulnerability
        i. Medical Illness & POC2
4. **Hyde Park**
    a. Race/ethnicity (ACS)
        i. Has a majority Non-Hispanic Black or African American alone population

b. Area median income/ income
    i. 78k
c. Housing density & Population density - the information found wasn't helpful
d. Social vulnerability
    i. People of Color

—----------

In comparing areas with poor air quality to those with better air quality, we can examine several demographic characteristics for different zip codes, namely race/ethnicity, area median income/income, housing density, population density, and social vulnerability. Here's a breakdown of how these factors differ in the specified areas:

****Mattapan (Zip Code 2126):****
- ***Race/Ethnicity (ACS):*** Mattapan has a majority population of Non-Hispanic Black or African American individuals, indicating a predominantly Black community in this area.
- ***Area Median Income/Income:*** The median income in Mattapan is approximately $58,000, which suggests a moderate income level for its residents.
- **Housing Density & Population Density:** Unfortunately, specific information regarding housing and population density was not available for Mattapan.
- ***Social Vulnerability:*** Mattapan is identified as having one of the most vulnerable populations, with People of Color facing higher vulnerability to environmental and social challenges in this area.

****Rosindale (Zip Code 2131):****
- ***Race/Ethnicity (ACS):*** Rosindale has a majority population of Non-Hispanic White Alone individuals, indicating a predominantly White community.
- ***Area Median Income/Income:*** The median income in Roslindale is higher, standing at approximately $88,000, which reflects a relatively higher income level in this area compared to Mattapan.
- ***Housing Density & Population Density:*** Unfortunately, specific information regarding housing and population density was not available for Rosindale.
- ***Social Vulnerability:*** The impact of social vulnerability, particularly in terms of People of Color, is an important aspect to consider in Roslindale as well.

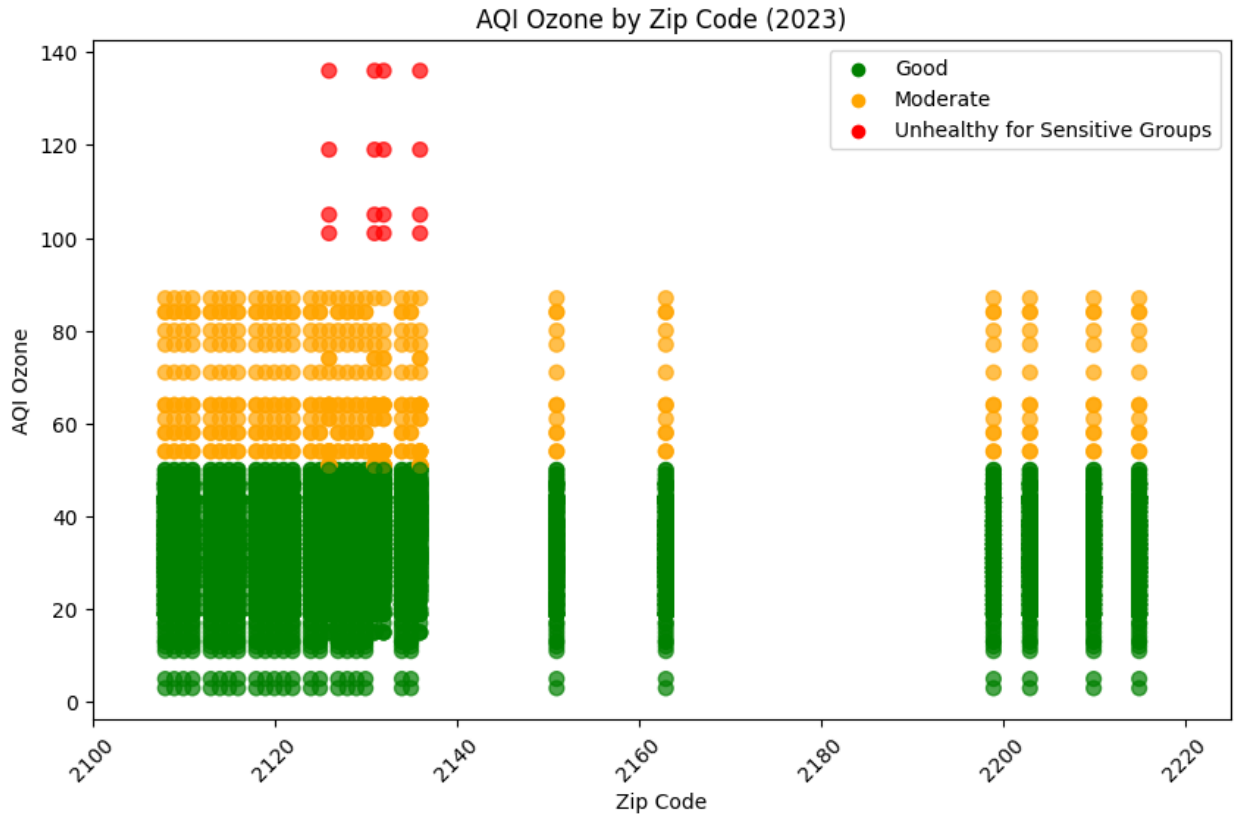
****West Roslindale/Roxbury (Zip Code 2132):****
- ***Race/Ethnicity (ACS):*** West Rosindale/Roxbury has a majority population of Non-Hispanic White Alone individuals, which is similar to Rosindale.
- ***Area Median Income/Income:*** West Rosindale/Roxbury exhibits a higher median income, around $106,000, signifying a relatively affluent area.
- ***Housing Density & Population Density:*** Specific information regarding housing and population density was not available for this area as well.
- ***Social Vulnerability:*** The social vulnerability in this area appears to be influenced by medical illness and People of Color (POC2).

****Hyde Park (Zip Code 2136):****
- ***Race/Ethnicity (ACS):*** Hyde Park, like Mattapan, has a majority population of Non-Hispanic Black or African American individuals, indicating a predominantly Black community.
- ***Area Median Income/Income:*** The median income in Hyde Park is approximately $78,000, reflecting a moderate income level for residents.
- ***Housing Density & Population Density:*** Unfortunately, specific information regarding housing and population density was not available for Hyde Park.
- ***Social Vulnerability:*** Similar to other areas, social vulnerability is influenced by the presence of People of Color.

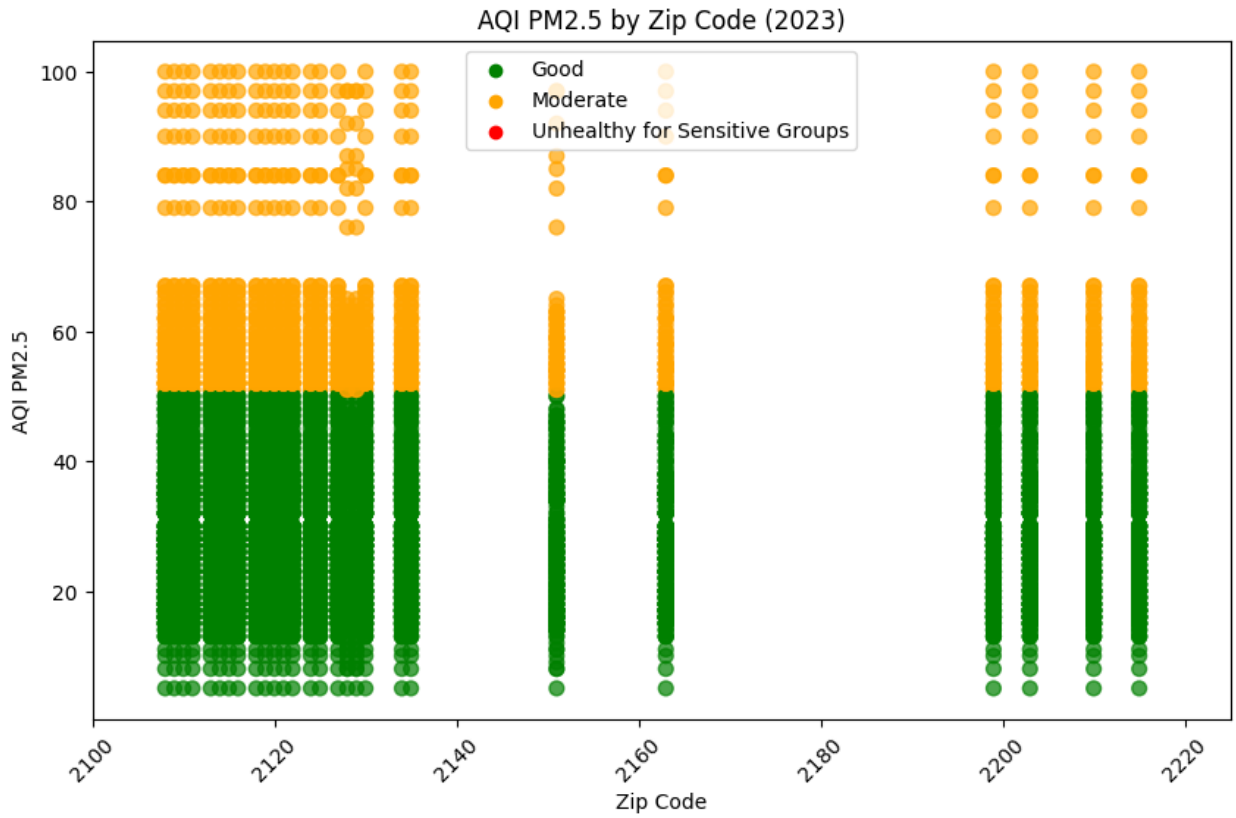
In summary, different zip codes with varying air quality levels demonstrate distinctive demographic characteristics, including racial composition and income levels. However, detailed data on housing and population density was not available. Social vulnerability, particularly among People of Color, appears to be a common concern in these areas, emphasizing the need for targeted interventions and policies to address disparities and improve environmental and social well-being.

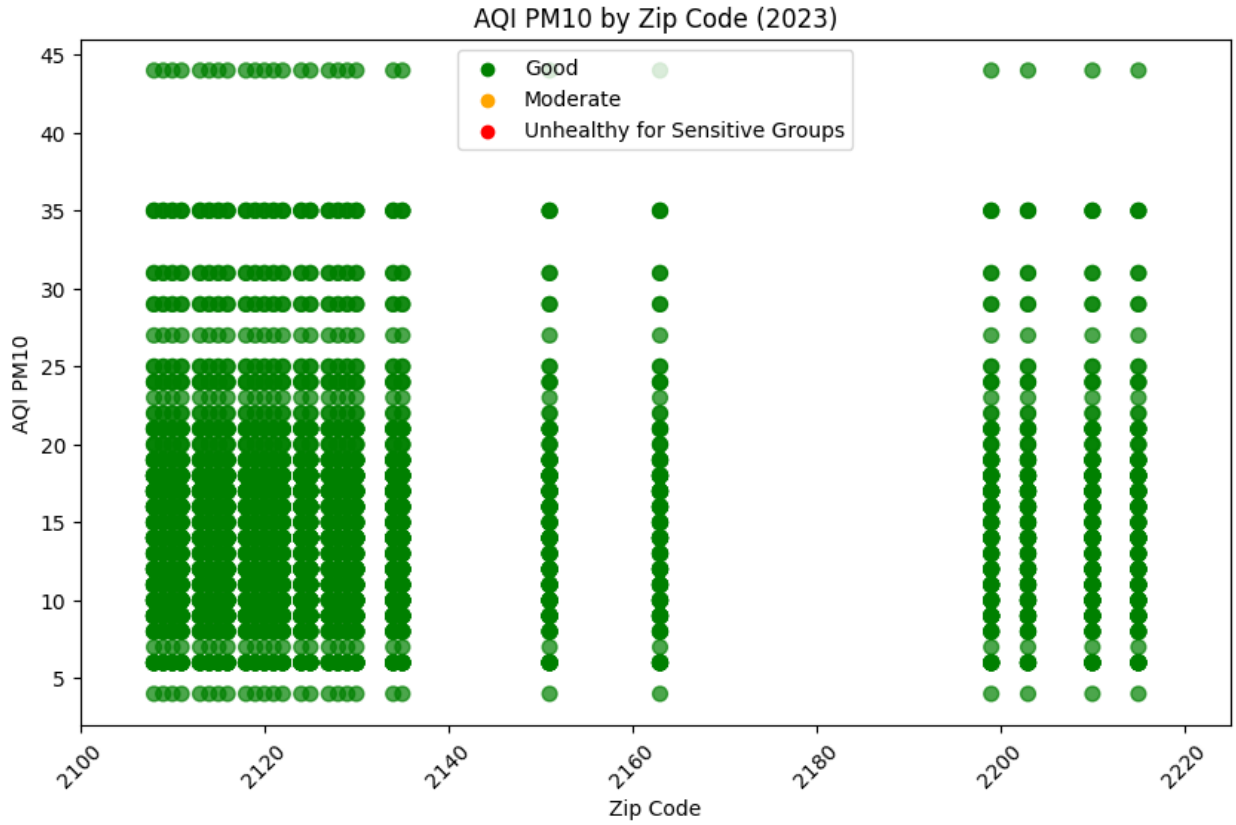
Graphs Continue on Next Page

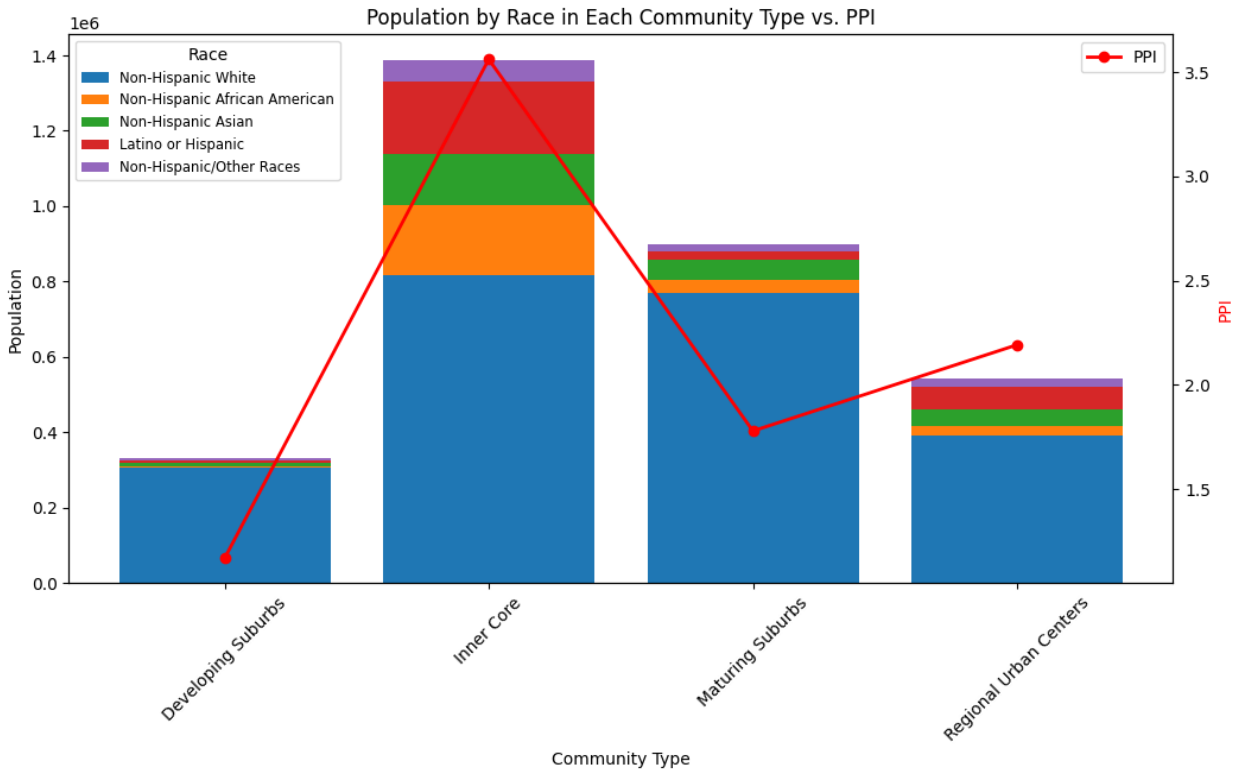
Graphs Generated:

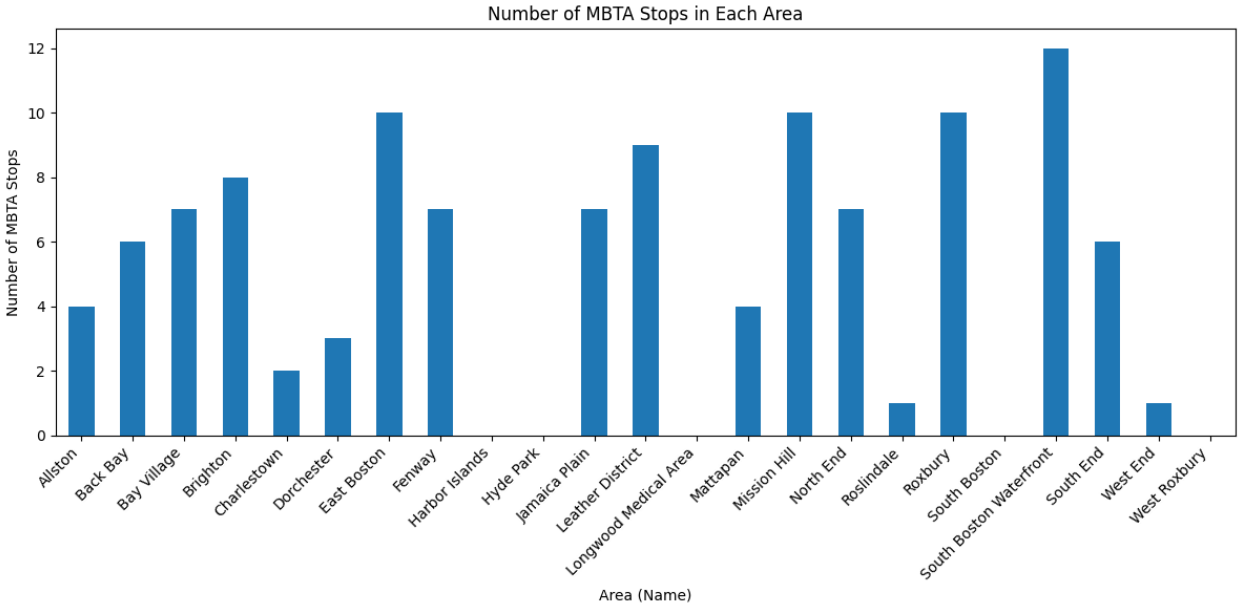


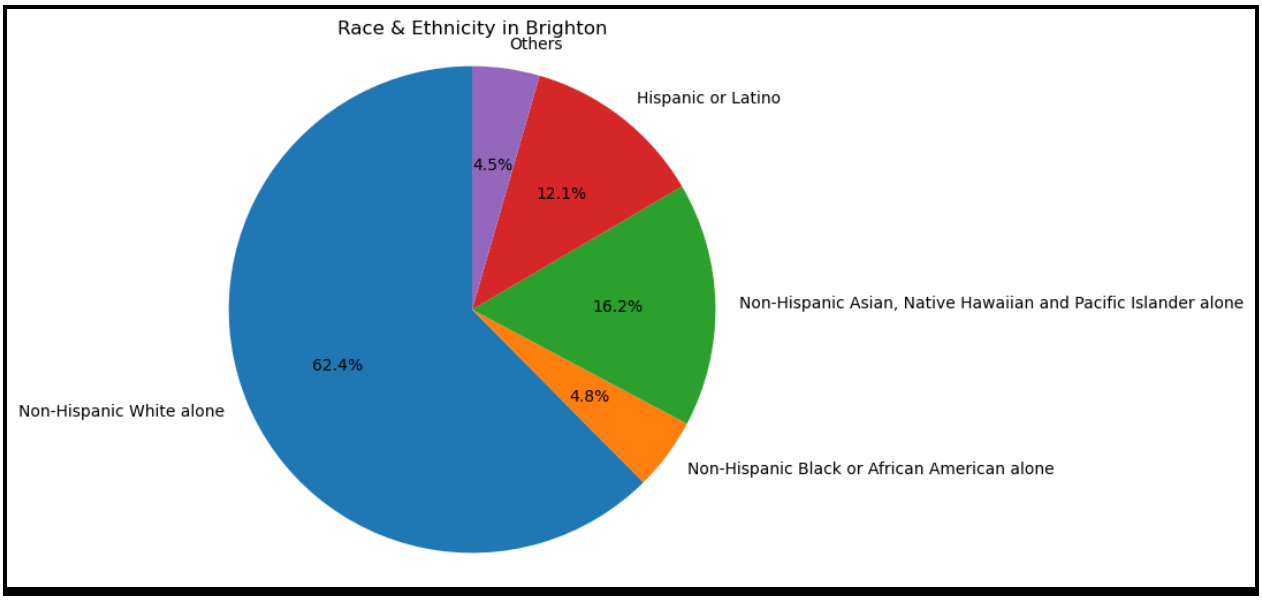
Context for AQIGraphs:
The graph shows the entire year so one point is for each day in the zip code. The dataset shows the max occurrence so on average there was good quality on average and there were few days with unhealthy quality.

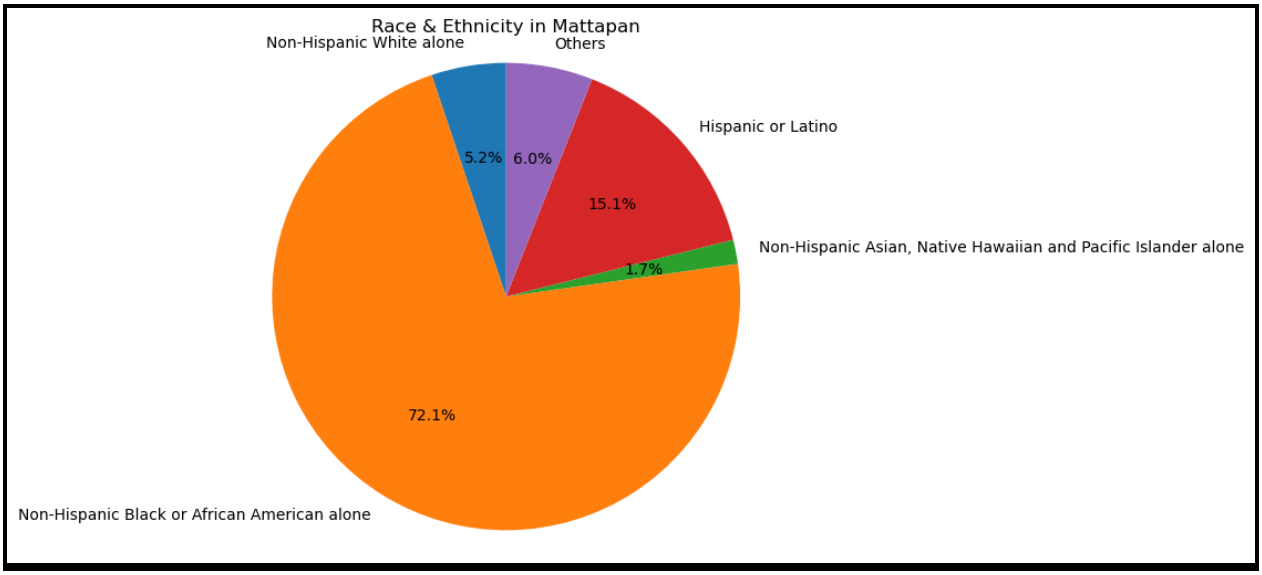

AQI PM2.5 by Zip Code (2023)
Good
Moderate
Unhealthy for Sensitive Groups
AQI PM2.5
Zip Code
100
80
60
40
20
2100
2120
2140
2160
2180
2200
2220

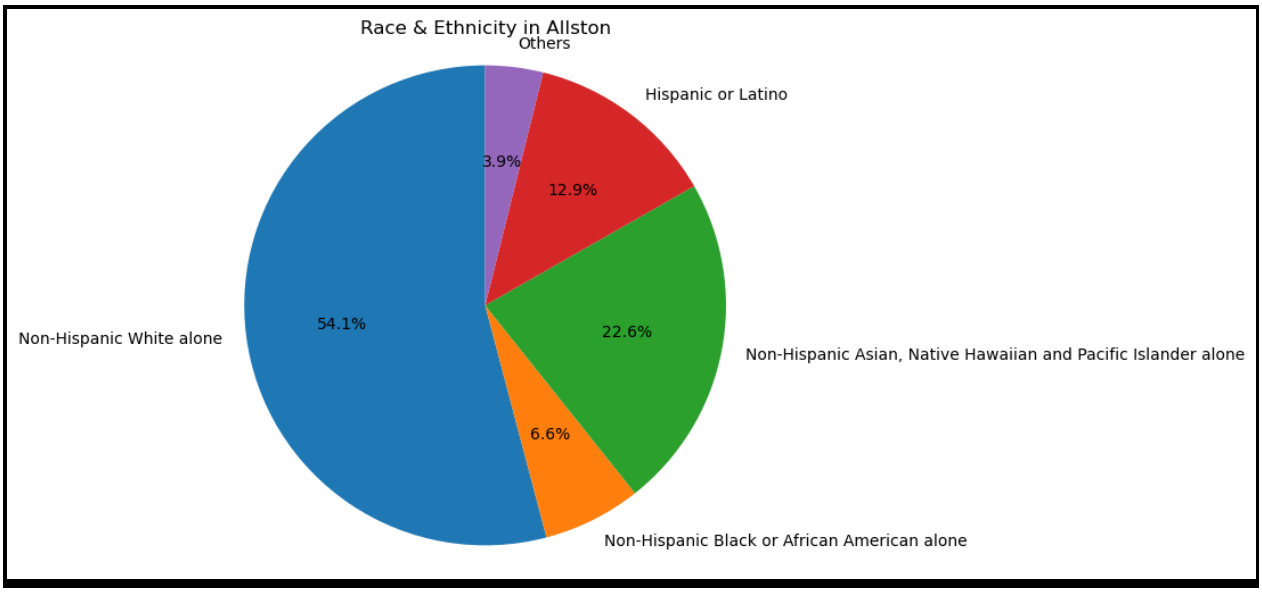

AQI PM10 by Zip Code (2023)
Good
Moderate
Unhealthy for Sensitive Groups
AQI PM10
Zip Code
45
40
35
30
25
20
15
10
5
2100
2120
2140
2160
2180
2200
2220

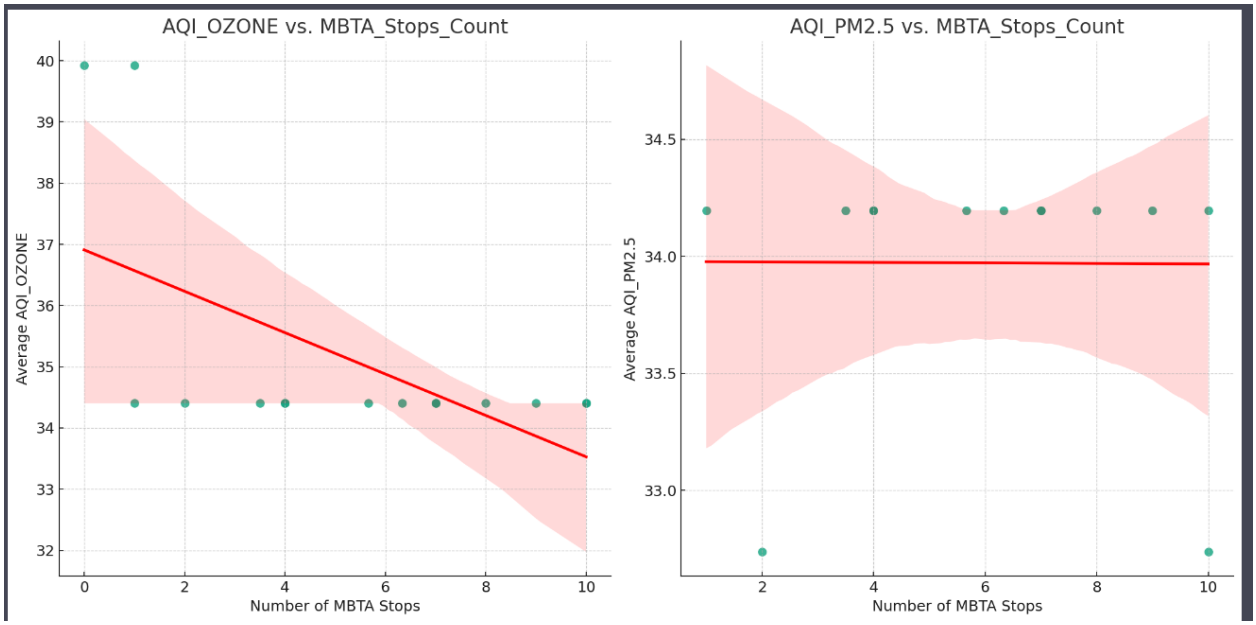

Population by Race in Each Community Type vs. PPI
1e6
Race
Non-Hispanic White
Non-Hispanic African American
Non-Hispanic Asian
Latino or Hispanic
Non-Hispanic/Other Races
PPI
Population
PPI
Developing Suburbs
Inner Core
Maturing Suburbs
Regional Urban Centers
Community Type

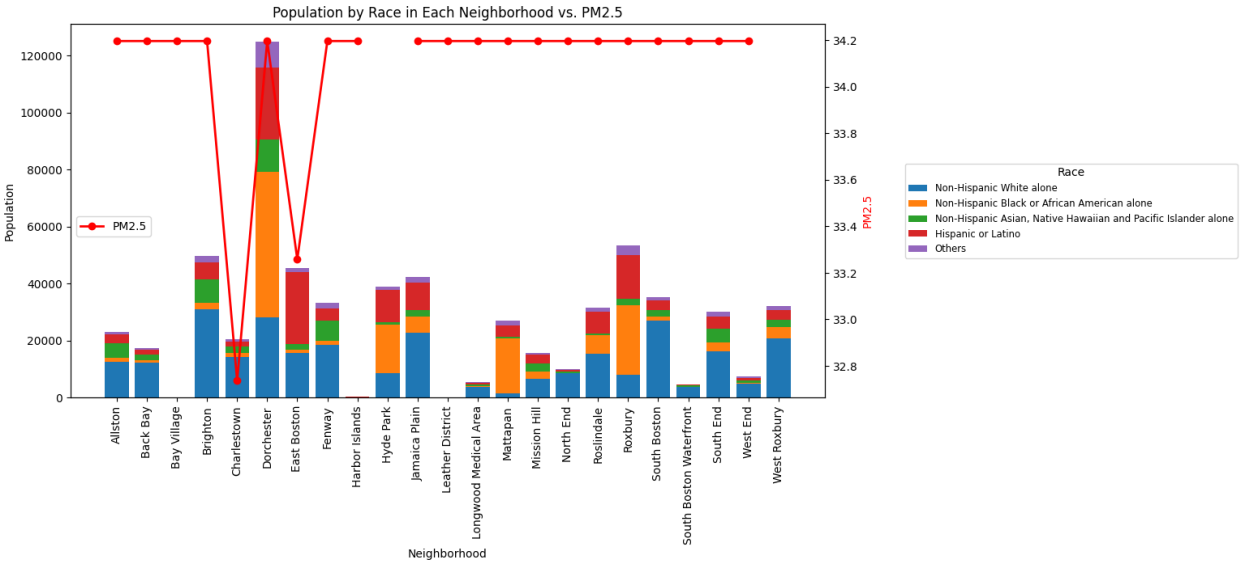

Number of MBTA Stops in Each Area
Number of MBTA Stops
Allston
Back Bay
Bay Village
Brighton
Charlestown
Dorchester
East Boston
Fenway
Harbor Islands
Hyde Park
Jamaica Plain
Leather District
Longwood Medical Area
Mattapan
Mission Hill
North End
Roslindale
Roxbury
South Boston
South Boston Waterfront
South End
West End
West Roxbury
Area (Name)

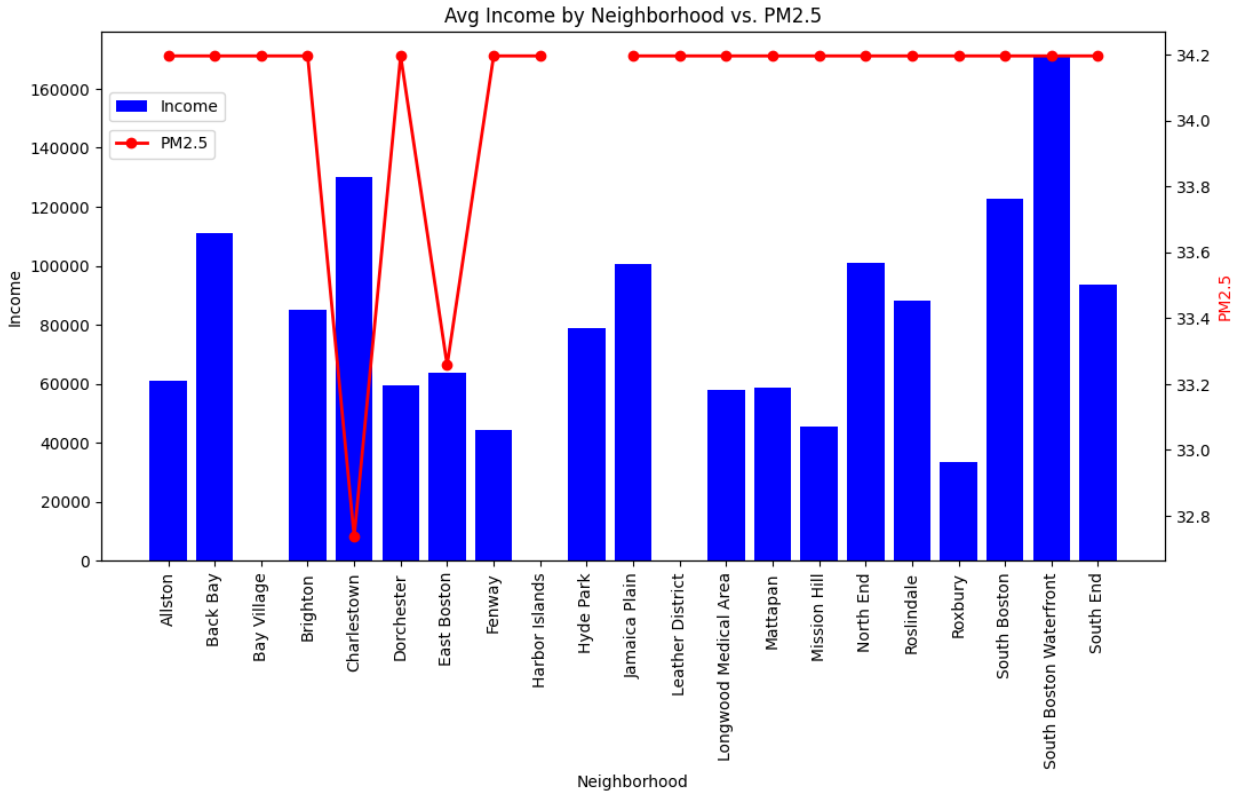

Race & Ethnicity in Brighton
Others
Hispanic or Latino
4.5%
12.1%
16.2%
Non-Hispanic Asian, Native Hawaiian and Pacific Islander alone
62.4%
4.8%
Non-Hispanic White alone
Non-Hispanic Black or African American alone

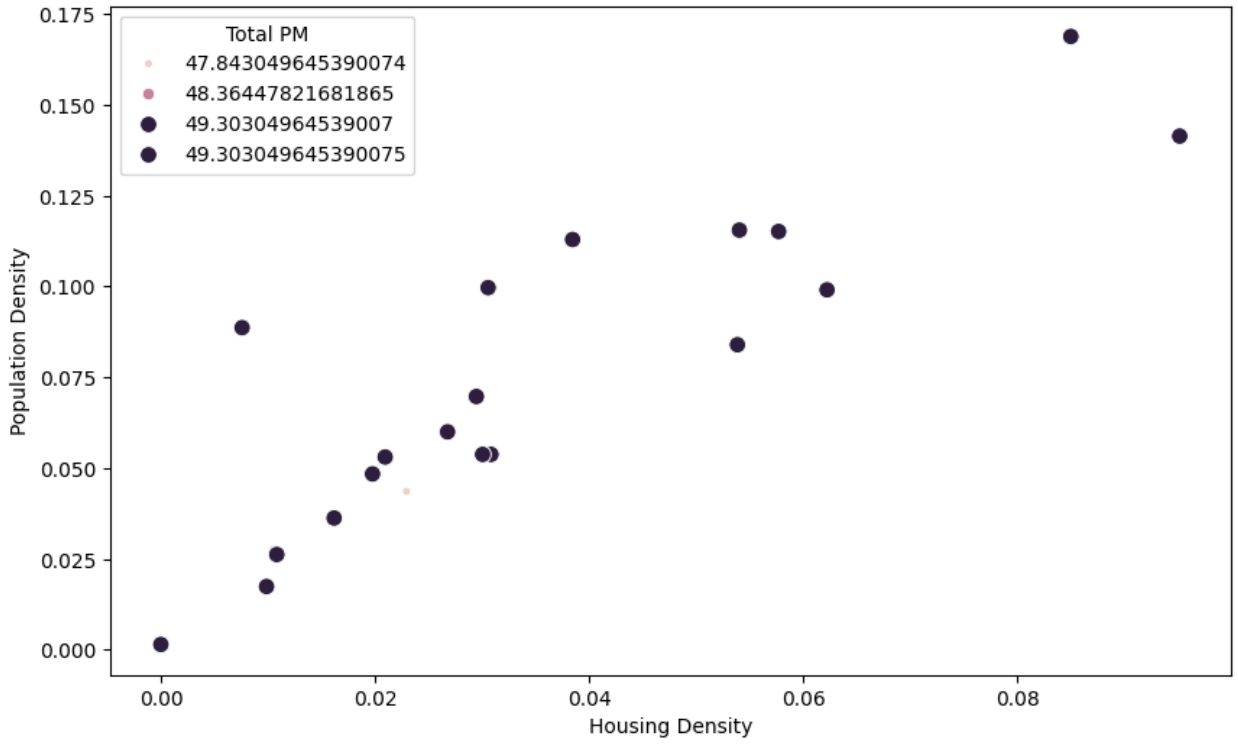

Race & Ethnicity in Mattapan
Non-Hispanic White alone
Others
Hispanic or Latino
5.2%
6.0%
15.1%
Non-Hispanic Asian, Native Hawaiian and Pacific Islander alone
1.7%
72.1%
Non-Hispanic Black or African American alone

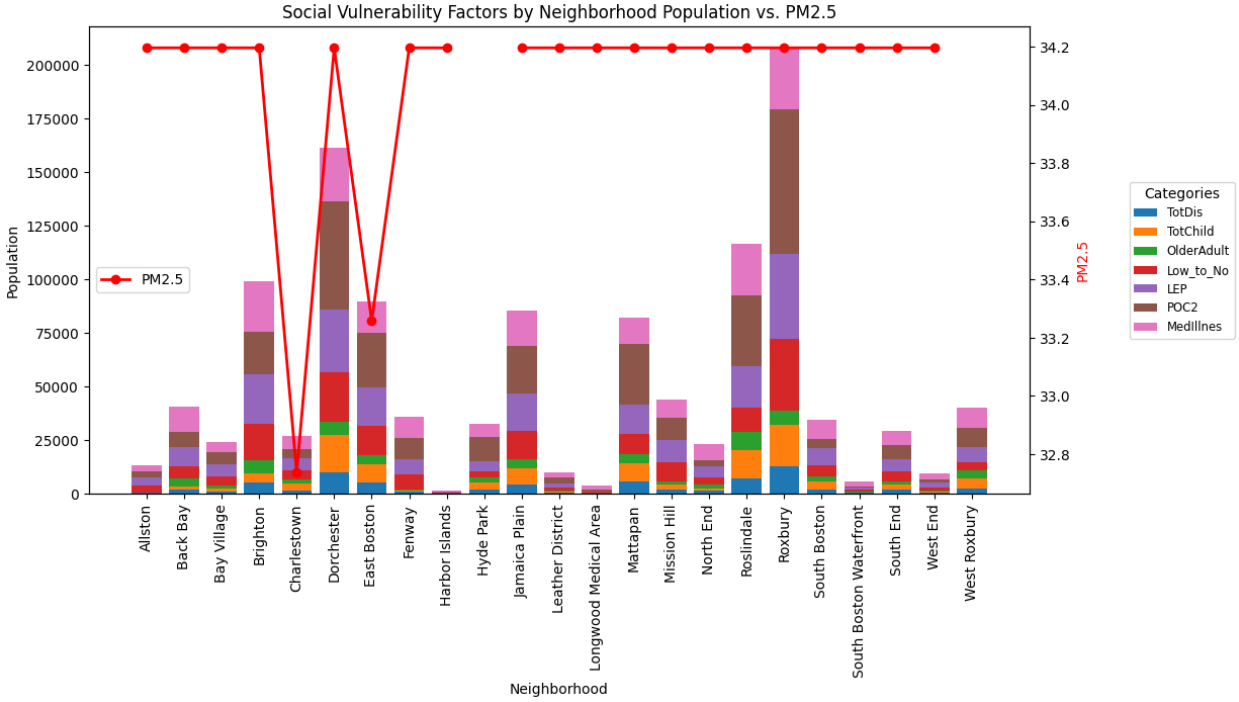

Race & Ethnicity in Allston
Others
Hispanic or Latino
3.9%
12.9%
54.1%
22.6%
6.6%
Non-Hispanic White alone
Non-Hispanic Asian, Native Hawaiian and Pacific Islander alone
Non-Hispanic Black or African American alone


AQI_OZONE vs. MBTA_Stops_Count
Average AQI_OZONE
Number of MBTA Stops
AQI_PM2.5 vs. MBTA_Stops_Count
Average AQI_PM2.5
Number of MBTA Stops


Population by Race in Each Neighborhood vs. PM2.5
Population
PM2.5
Neighborhood
Race
Non-Hispanic White alone
Non-Hispanic Black or African American alone
Non-Hispanic Asian, Native Hawaiian and Pacific Islander alone
Hispanic or Latino
Others
Allston
Back Bay
Bay Village
Brighton
Charlestown
Dorchester
East Boston
Fenway
Harbor Islands
Hyde Park
Jamaica Plain
Leather District
Longwood Medical Area
Mattapan
Mission Hill
North End
Roslindale
Roxbury
South Boston
South Boston Waterfront
South End
West End
West Roxbury


Avg Income by Neighborhood vs. PM2.5
Income
PM2.5
Neighborhood
Allston
Back Bay
Bay Village
Brighton
Charlestown
Dorchester
East Boston
Fenway
Harbor Islands
Hyde Park
Jamaica Plain
Leather District
Longwood Medical Area
Mattapan
Mission Hill
North End
Roslindale
Roxbury
South Boston
South Boston Waterfront
South End


Total PM
47.843049645390074
48.36447821681865
49.30304964539007
49.303049645390075
Population Density
Housing Density
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.00
0.02
0.04
0.06
0.08


Social Vulnerability Factors by Neighborhood Population vs. PM2.5
Population
PM2.5
Neighborhood
Categories
TotDis
TotChild
OlderAdult
Low_to_No
LEP
POC2
MedIlnes
Allston
Back Bay
Bay Village
Brighton
Charlestown
Dorchester
East Boston
Fenway
Harbor Islands
Hyde Park
Jamaica Plain
Leather District
Longwood Medical Area
Mattapan
Mission Hill
North End
Roslindale
Roxbury
South Boston
South Boston Waterfront
South End
West End
West Roxbury

Individual Contribution:

Anulika Nnadi:
In this report, my primary role was to gather crucial data from the Census for every census tract in Suffolk County pertaining to race/ethnicity (ACS), area median income/income, and poverty. I encountered some challenges in filtering the questions and sought assistance from Michelle and went CS506 during office hours to overcome these obstacles. Using the Census and ACS data, I organized the data I found by constructing new data sets pertaining to income, race & ethnicity, and poverty rates for all neighborhoods in Suffolk County so it would be easier for my teammates to create graphs using that information. Additionally, I contributed to the report by creating graphs that illustrated the various demographic aspects of each neighborhood in Suffolk County. By analyzing these graphs and the data my team collected, I addressed question 2 to ensure the report's accuracy and comprehensiveness.

Ziliang Wang:
In this report, my responsibilities included collecting a shapefile of the 250-meter gridded population data with associated PPI scores, and using AirNow api to collect and merge 2023 year Boston area air quality data (PM2.5,PM 10, Ozone).In this step, the first challenge I faced was using api to get the data,after group discussion, we find out the efficient way to collect the data.After collect the data. After our group collected all the data, I started to answer question 1. I conducted an analysis on the air quality data for Boston in 2023, focusing on how proximity to transportation infrastructure like public transit stops and roads affects air quality.

Xinzhu Liang
In this report, I am responsible for grabbing the data from the AirNow API. I created a program to extract all air quality data for the Boston area from the API for 2023. After I gained the data, I found it hard to plot the data because I made the query based on zip codes, and the coordinates it returned are the same. So I also included a “zip-to-coordinates” file online to make conversions to coordinates. After that, I could plot the air quality values across the map.

Hemanshu Bhojwani
I worked on processing the Transit Data, Social Vulnerability and Proximity to roads data sets. I also worked on merging the Transit Data, Social Vulnerability, and Air Quality datasets into a folium map. I created graphs related to all these datasets, including census data which can be seen in our github. I worked on processing the Air Quality data, merging them with datasets to create a summarized csv file by neighborhood. Manipulated data to find nearest air quality sensor and number of stops from each tract and then divided data by zip code/neighborhood. In the report I worked on the problem statement, data collection, metric definitions, PPI analysis in q1 and assisted with q2, and created the AQI graphs, Housing Density vs Population Density graph, MBTA stops by area, and PPI Graph.