CS 506 Final Project: Air Quality & Transport

Team E: Doruk Savasan(doruks@bu.edu, 2025), Can Erozer(caner@bu.edu, 2024),
Maxwell Higa(mhiga@bu.edu, 2024), Medha Dhir(mdhir@bu.edu, 2024)

Table of Contents

Introduction

Improving air quality is a critical concern in Boston, particularly for marginalized communities that bear the brunt of its adverse effects, resulting in elevated rates of asthma and lung diseases. One potential solution to address poor air quality involves transitioning public transportation to alternative energy sources. This shift holds promise in significantly reducing harmful emissions and stabilizing air quality. However, implementing this transition is a complex undertaking due to the indispensable role transportation plays in the lives of Boston residents.

Understanding the intricate links between transportation infrastructure, air quality, and residents' health is paramount. Such insights are essential for guiding necessary actions to enact changes in Boston effectively. Moreover, recognizing the profound impact of air quality on the health of Bostonians emphasizes the need to develop policies for a cleaner and more environmentally sustainable city. Recognizing and analyzing the correlations between air quality and overall health outcomes are pivotal in formulating effective strategies.

Our project encompasses two primary objectives: firstly, to comprehend the correlations and impact of transportation infrastructure on overall air quality in Boston. Secondly, upon establishing a baseline correlation, we aimed to assess health outcomes, such as air quality and the percentage of residents suffering from asthma and lung diseases by zip code.

Project Goal

Investigate the relationship between transportation infrastructure and its impact on Boston's air quality and climate and surrounding Boston neighborhoods.

Data Collection

To conduct the analysis of transportation infrastructure and resulting air quality, a plethora of data is needed, some of which were provided to us by the City of Boston:

1. **Proximity to roads** (PPI Roads): Contains data about the spatial patterns of residents living in close proximity to roads with the highest levels of vehicle air pollution emissions across the MAPC region
    i. **g250mm_id**: Refers to a 250 x 250 mm area of population in Boston
    ii. **nhwi_10**: Refers to the not-hispanic white population based on Census 2010 data
    iii. **nhaa_10:** Refers to not-hispanic african-american population based on Census 2010 data

iv. **nhapi_10:** Refers to not-hispanic population asian-pacific islander population based on Census 2010 data
v. **lat_10:** Refers to Latino population based on Census 2010 data
vi. **nhother_10:** Refers to other racial population based on Census 2010 data
vii. **ppi5:** Refers to air pollution emissions(**0: lowest, 5: highest**)

2. **Air Quality Data** (AQI Dashboard): Since the dashboard gets real-time data, use the AirnowAPI that feeds into the dashboard:AirNow API

3. **Census(Transport, Income, Household Size) Data** (Census Bureau): In order to measure the AQI impact on population, it is necessary to get census data. The dataset we collected (DP02/03/04/05) provided valuable insights into the modes of transportation, income, and households for 45 zipcodes in Boston for the year 2021. The image below shows the specific data recorded by the dataset.

Data Cleaning

1. **Census Data**
   i. In order to have more organized data, we ran a Python script that dropped empty values and pre-processed data that was relevant for 2021.
   ii. Made the column names more readable
   iii. Merged with the AQI data so that everyone can work collaboratively.
2. **AQI Data**
   i. In order to have more organized data, we ran a Python script that converted the aqi.db files into csv for easy access via pandas.
   ii. First we converted the .db file to .csv.
   iii. Got rid of all the NaN values.
   iv. Removed unnecessary features.
   v. Made a new dataset that contains the average of PM and OZONE values for each zip code.

Base Questions & Analysis

1) ***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?***

To observe the yearly change in air quality for Boston residents, we would need to explore proximity. However, the current PPI data does not include this information. Hence, we explored transportation mediums used in the 45 different zip codes and explored the impact these had on PM2.5 levels. Based on the grouped bar graphs below, the Car seems to be the most popular transport medium, followed by public transportation. What can be further studied is the impact this can have on air quality and how Boston combats this since the air quality for 2021 is relatively 'Good'.

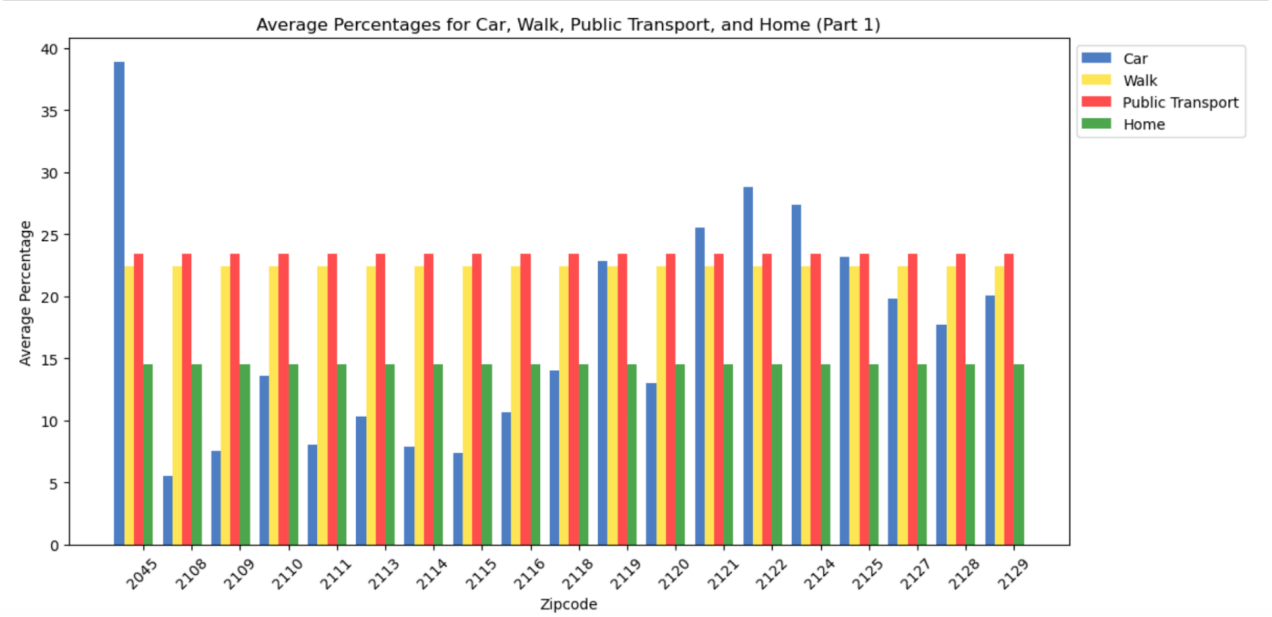


Figure #1*: The bar chart above depicts the largest modes of transportation(Car, Walk, Public Transportation, or WFH) across 45 zip codes in the Boston area. This helped us understand the demographics of Bostonians and their proximity to roads, since 'Car' emerged as the largest form of transport.*

Although the overall air quality in Boston is relatively good, we hypothesize that as we look deeper into each zip code, the variance in air quality between zip codes will correlate to the relative distance to high-emission areas (roads, highways, and public transportation lines). Furthermore, PM2.5 for 2021 stayed relatively the same across the year since the overall air quality for 2021 is 'Good'(29-33 AQI). To measure its impact on AQI levels, we attempted to merge our AQI data with transport data to better

understand any correlations or trends. We only considered values where the percentage of residents in a zipcode using a specific type of transport was greater than 50%. Using this to filter the data, Car, and Walk appeared to be the preferred methods of transport. The image below shows the zipcodes and average AQI values where either Car/Walk was the most prominent. Thus, we can hypothesize that people who use a car live in areas that might be further away from public transport and find it efficient to have a Car. On the other hand, the people who walked to work were those who lived close by and didn't need public transport.

```
Zip Codes of People Using Car:
[2045 2131 2170 2171 2152]
```

```
Zip Codes of People Using Walking to Work:
[2109 2163 2215]
```

***Figure #2**: The code snippets above depict the most common zipcodes of Boston residents using Car or Walking to work. This statistic enables us to explore the zipcodes and figure our a trend or correlation between the area and relative road proximity. We suspect residents with a Car are likelier to live farther away from larger roads.*

Once we got these zip codes, the next step was to explore how these translated to AQI changes, if any. The graph below shows no correlation between transport and air quality.

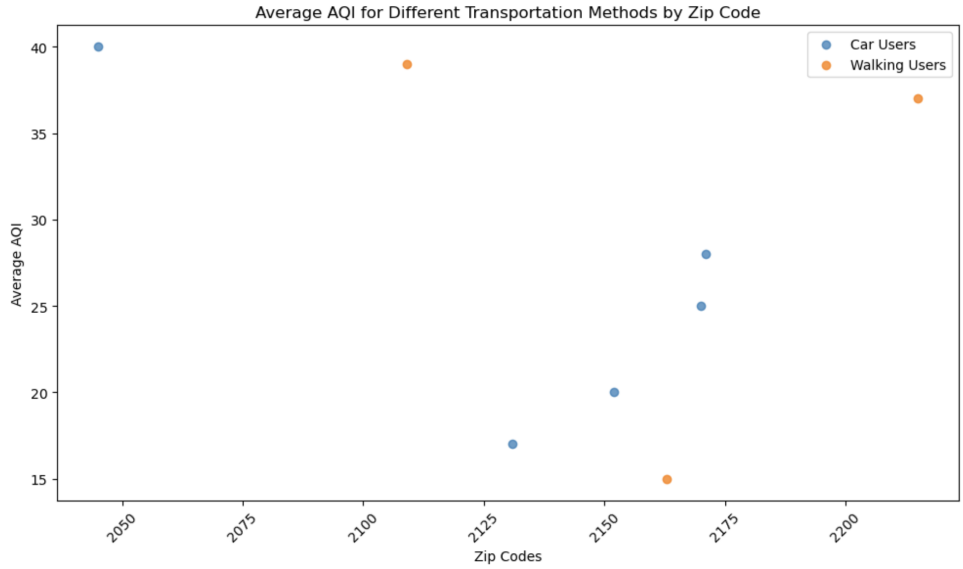


***Figure #3**: The scatter plot above aims to explore a correlation between two modes of transport: car users and walking users against the average AQI levels for 12 zipcodes, to strengthen our analysis. Based on the graph, it is clear no correlation exists, meaning our AQI data might need to be updated.*

However, this inference might change since we have found new AQI data(derved from 16 sensors) that can inform our analysis better.

2) ***How do areas with poor air quality compare to areas with better air quality based on different demographic characteristics, specifically:***
   a) ***Race/ethnicity (ACS)?***
      i) I first started the analysis by comparing columns for Natives and Foreigners. There are five different unique ozone and PM2.5 levels. We looked at the rows with the highest and lowest ozone and PM2.5 levels for both metrics. As expected multiple rows had max or min metric levels. So we took the average of the values. In districts with the highest ozone level, natives constitute 74.0% of the population, foreigners constitute 26.0%. Whereas in districts with the lowest ozone level, Natives constitute 84.9% of the population, foreigners constitute 15.1%. So as the percentage of foreigners increases the air quality tends to get poorer. However, when observing the PM2.5 metric, we reached an opposite conclusion.

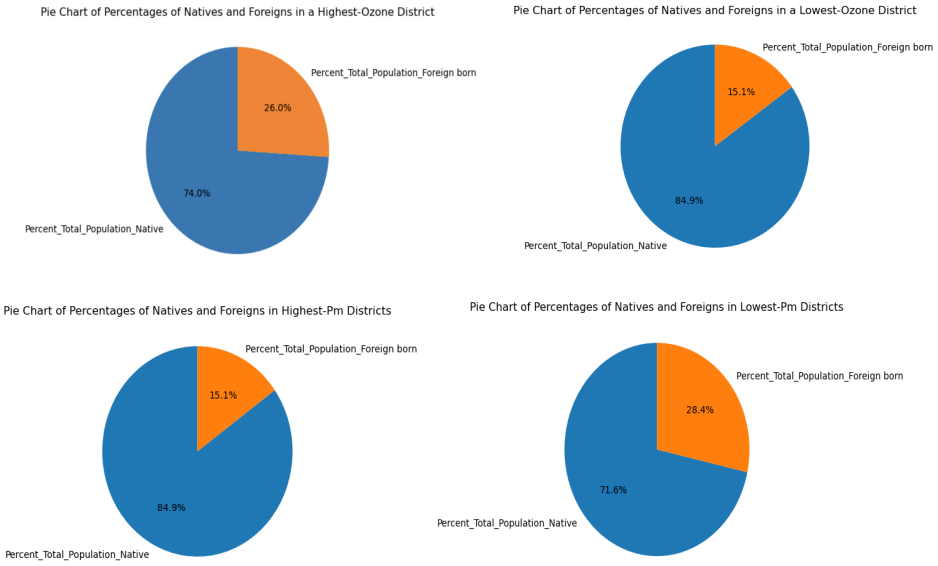


Figure #4: *The pie charts above depicts the splits between the ozone districts and the population distributions.*

      ii) We then proceeded our analysis by comparing districts by looking at columns for Hispancs and Non-Hispanics. Because there were so many columns that belong to this category we decided to

combine columns for Hispancs and Non-Hispanics and compared two columns. As can be seen from the pie charts, as the percentage of Hispanic population increases the air quality gets poorer by looking at the ozone metric. But again, we made a different conclusion when we looked at the PM2.5 metric.

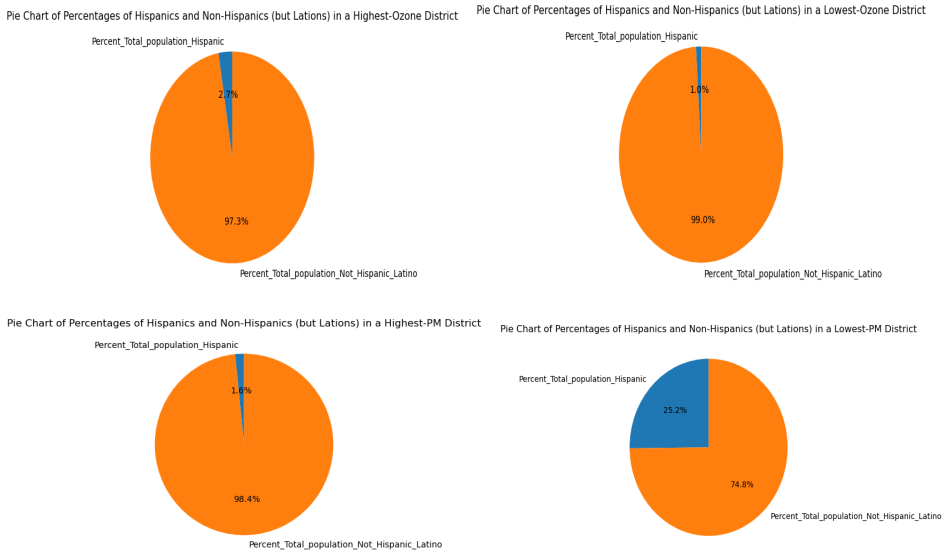


Figure #5: *After exploring the correlations between PM districts for Foreigners and Natives, we explored the same trend for Hispanic and non-Hispanic (but Latino) populations.*

Area median income/ income

iii) We decided to take a different approach because we couldn't reach a consistent conclusion when working on columns for race/ ethnicity. The unique ozone levels are 31.076, 31.084, 31.533, 31.939, 35.124. Acknowledging that the first four ozone levels are close, we thought it would be reasonable to combine and take the average of rows with these ozone levels. But still, there was an inconsistency when looking at the ozone and PM2.5 metrics. Here are the results: by looking at the ozone metric, even though there is no direct correlation between air quality and income levels, we can say that in places with good air quality, the percentage of people who make 200k is the highest; and in places with poor air quality, the percentage of people who make 50k-74.9k is the highest. However, by looking at the PM2.5 metric, now people with the highest income dominate the highest polluted areas.

Therefore, in order to make a concise conclusion, we need to further investigate what are the biggest factors of ozone and PM2.5 metrics which exceeds the scope of our project.

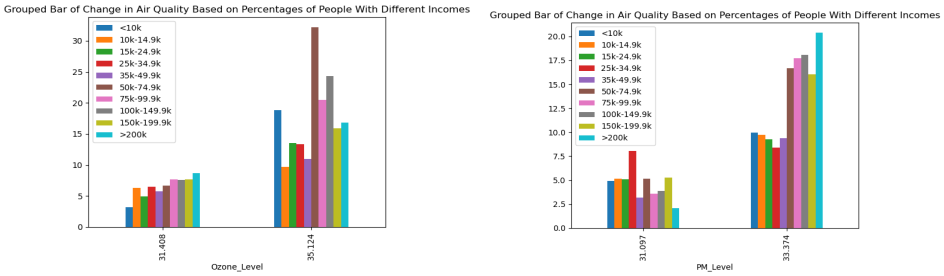


Figure #6: *Here we looked at how people with different incomes are being affected by air quality levels.*

b) *Housing & Population Density*

i) Overall, there is not much of a correlation between PM2.5 AQI and housing density or population density, except the outliers. This means that most densely populated and densely housed areas do have relatively worse mean air quality. In both graphs, the zip code 02118 (South End to South Boston) has the worst air quality and highest density, while the zip code 02128 (East Boston) has the best air quality and lowest density

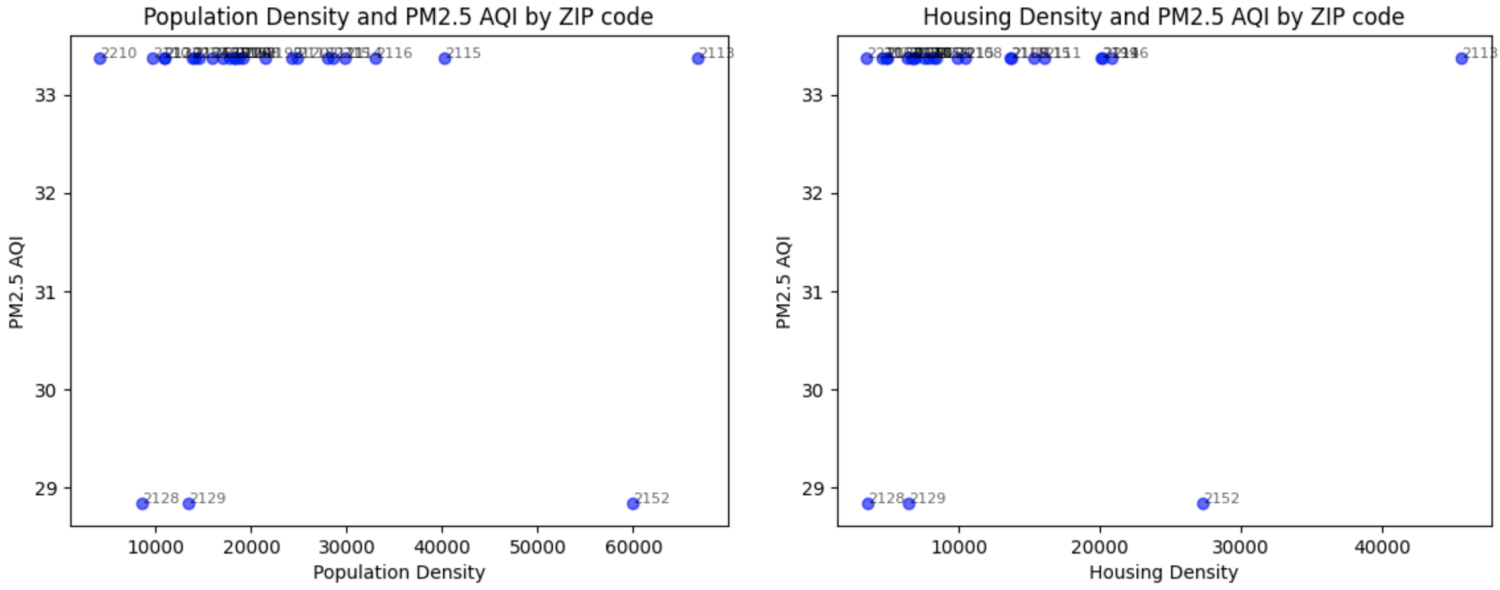


Figure #7: *The above plots show the relationship between air quality levels and population and housing densities. While there is not a very significant correlation between housing/population density and PM2.5 AQI, the extreme*

points in the data set do follow a trend: the least dense zip codes (02128, 02129) have a better air quality relative to the most dense zip codes (02118).

c) *Social vulnerability*

i) Given that Boston is demographically quite diverse with 53% people of color, comparing the social vulnerability factor between different races have provided deep insights into the disproportionate effects of air quality and proximity to transportation between white people and people of color. The data indicates that over 45% of Black and Asian residents and over 50% of Latinx residents reside in areas with the worst air quality (PPI of 5), versus less than 30% of white residents. We hope to explore other factors of social vulnerability as we move into the extension project, specifically health data that would elaborate on inequities between different demographics in Boston. By conducting research into respiratory diseases during the extension project, we were able to further conclude that people of color in Boston are disproportionately affected by air quality. Black and Hispanic residents are particular suffered from a much higher asthma prevalence rates, which can be attributed to living in areas with the worst PPI.

3) *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.*

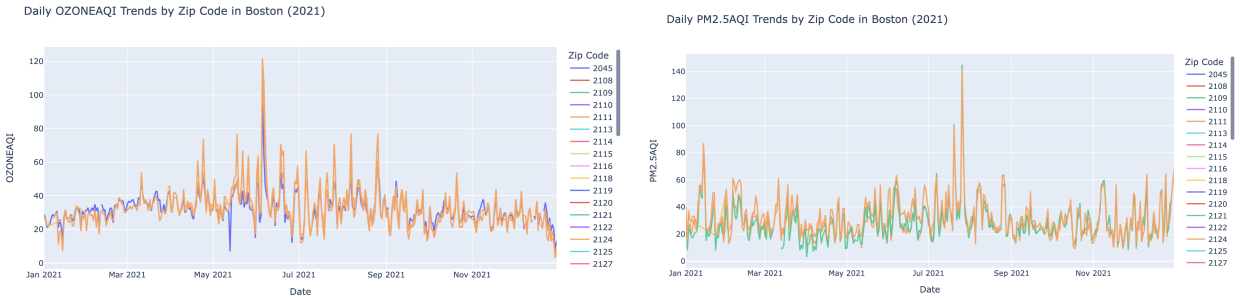


Figure #8: *These plots help us analyze the yearly changes in PM2.5 and OZONE levels for each of the zip codes used in our project. We were able to capture seasonal and yearly trends.*

The analysis primarily focuses on two pollutants - Ozone (OZONEAQI) and particulate matter of size 2.5 microns or less (PM2.5AQI). The data was grouped by zip code and date to calculate mean values. Interactive plots above were created to show daily trends in OZONEAQI and PM2.5AQI across different zip codes in Boston. These plots reveal variations in air quality over time and differences between neighborhoods.

	zip_code	ozone_mean	ozone_max	ozone_min	Max_Ozone_Date	Min_Ozone_Date
0	2045	35.1236	122	4	2021-06-05	2021-12-30
1	2108	31.0838	112	3	2021-06-05	2021-12-30
2	2109	31.0756	112	3	2021-06-05	2021-12-30
3	2110	31.0838	112	3	2021-06-05	2021-12-30
4	2111	31.0838	112	3	2021-06-05	2021-12-30
5	2113	31.0838	112	3	2021-06-05	2021-12-30

	zip_code	pm25_mean	pm25_max	pm25_min	Max_PM2.5_Date	Min_PM2.5_Date
0	2045	29.0168	133	8	2021-07-26	2021-11-27
1	2108	33.3699	141	11	2021-07-26	2021-10-17
2	2109	33.3736	141	11	2021-07-26	2021-10-17
3	2110	33.3699	141	11	2021-07-26	2021-10-17
4	2111	33.3699	141	11	2021-07-26	2021-10-17
5	2113	33.3699	141	11	2021-07-26	2021-10-17

Figure #9: *We used the above table to find a trend in daily air quality levels.*

For each zip code, average, maximum, and minimum OZONEAQI levels were calculated along with the dates of these extreme values. Similar statistical measures were calculated for PM2.5AQI, providing a comprehensive view of the air quality in terms of these two pollutants.

The highest and lowest AQI levels for both pollutants were further analyzed to determine their corresponding seasons. This gives insight into seasonal variations in air quality. Bar charts were plotted to visualize the distribution of the highest and lowest AQI levels across different seasons. This helps in understanding which seasons experience poorer air quality.

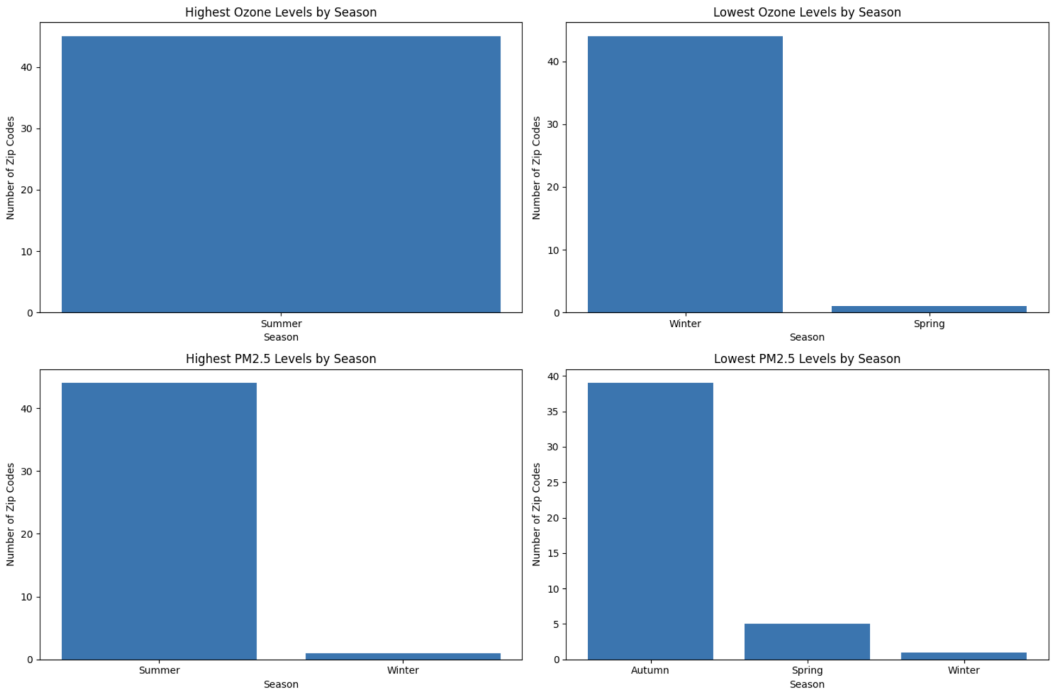


Figure #10: *These bar graphs illustrate the PM2.5 AQI and Ozone levels by season, respectively. The majority of zip codes had the highest PM2.5 AQI and Ozone levels in the summer, and lowest in the autumn, with a few outliers in the spring and winter.*

The average OZONE AQI and PM2.5 AQI were calculated for each zip code. Combined AQI Score: A combined AQI score, derived as an average of Ozone and PM2.5 AQI, was computed to provide a single metric representing overall air quality. Best and Worst Air Quality: Zip codes with the best (lowest) and worst (highest) combined AQI scores were identified, highlighting neighborhoods with relatively better or poorer air quality.

```
(    zip_code  Ozone_Avg   PM25_Avg  Combined_AQI
 42      2474  31.083799  28.839335     29.961567
 18      2129  31.083799  28.839335     29.961567
 17      2128  31.083799  28.839335     29.961567
 30      2145  31.083799  28.839335     29.961567
 31      2152  31.083799  28.839335     29.961567,
     zip_code  Ozone_Avg   PM25_Avg  Combined_AQI
 29      2144  31.532738  33.160819     32.346778
 23      2138  31.083799  33.369863     32.226831
 32      2163  31.083799  33.369863     32.226831
 1       2108  31.083799  33.369863     32.226831
 24      2139  31.083799  33.369863     32.226831)
```

Figure #11 *: The table displays the top 5 zip codes with the best and the worst air quality levels.*

After examining these fundamental questions, an issue arose: the AQI (Air Quality Index) data wasn't entirely accurate as it relied on readings from only 3-5 sensors

around Boston and neighboring areas, potentially leading to a lack of precision in determining air quality. Therefore, acquiring more reliable and comprehensive data was crucial before proceeding with our extension. In our extension phase, our primary objective was to solidify the earlier analysis by utilizing updated and improved air quality data.

Extension Analysis

Introduction

To conduct a more extensive analysis of air quality and its implications for Boston residents, our focus was on investigating the health outcomes of individuals in the area. Prior studies have highlighted a significant correlation between lung diseases and substandard air quality. To ensure more precise findings, we replaced the AQI data obtained from the AirQuality API with air quality data sourced from Aqicn. This alternative dataset draws information from 10 sensors distributed across the Boston region, thereby implying greater accuracy and reliability in our analysis. Furthermore, we merged this new aqi AQI data with health outcomes data from CDC in order to successfully analyze the correlations that exist between

Problem Statement: What is the relationship between these **yearly changes in air quality for Boston residents and health outcomes** (e.g., asthma rates, lung cancer rates)

Data Collection

1. **Aqicn- Air Quality Data <LINK>**
    a. Mean estimations of PM2.5 values for 12 zipcodes
    b. 5 number summary statistics for 2023.

2. **CDC - Health Outcomes Data <LINK>**
    a. Provides details about the biggest health issues of Boston residents by zipcode.

3. **Census Health Data <LINK>**
    a. Provided health outcomes by zipcode
    b. Data was divided by Health Status, Conditions

Main Question

What is the relationship between these yearly changes in air quality for Boston residents and health outcomes (e.g. asthma rates, lung cancer rates, other health outcomes)?

Step 1: **Exploratory Data Analysis(EDA)**
We conducted basic EDA for health data to explore the health outcomes for Boston residents. Before we can explore the impact of air quality on health outcomes of residents, we need to understand the trends displayed in the health data.

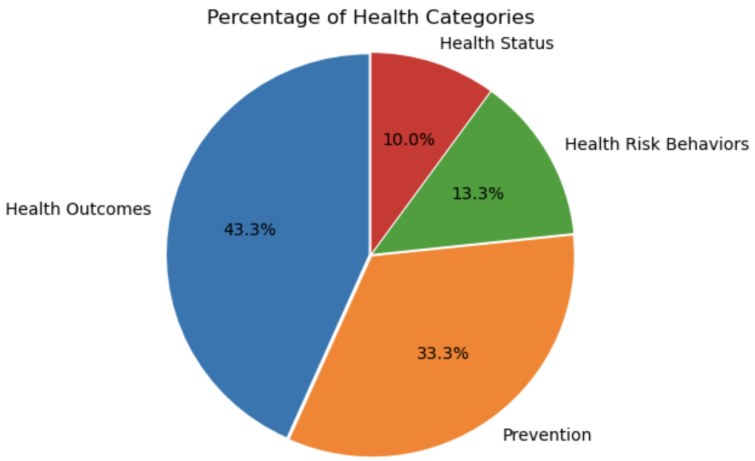


***Figure # 12**: The pie chart above summarizes the various percentages of different health categories for Boston residents. More than 50% of the people have preventive health conditions, indicating that preventive category should be explored further to look for a correlation.*

Based on the pie chart above, we can easily conclude that ~77% of Boston residents have health outcomes or prevention based health conditions which need to be addressed. However, in order to understand the necessary steps to combat this, we need to delve deeper into the health 'conditions' of each of these categories. The mind-maps below attempt to describe the conditions found in each of the 4 categories in order to inform our analysis further.

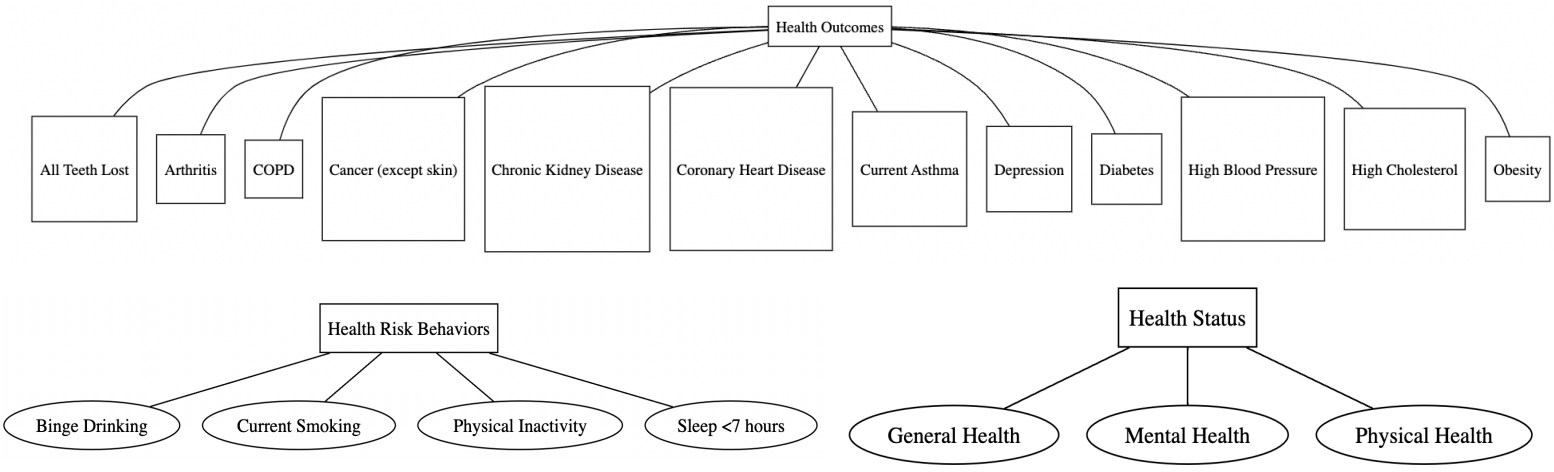


***Figure #13**: The charts above use the census health data to examine what are the largest health outcomes of Boston residents based on AQI levels. The most common types of diseases found among residents were Cancer, Kidney Diseases, Diabetes, Depression, High Blood Pressure*

All of the 'Conditions' had the same count(38), but what brought about the **higher** percentages in health outcomes and prevention are the number of conditions present are higher than health status and health risk behaviors(visible in the graphs above) These are just preliminary insights. In order to fully understand the role this plays in air quality, we need to have a location-based analysis and measure a regression model to understand health outcomes and air quality.

Step 2**: Preliminary Insights**

Using the new AQI data from 10 additional sensors around Boston neighborhoods, we conducted a two-fold analysis: exploring data by zip codes and data for aggregated Boston to discover correlations and trends. However, the data created an issue since we had a 5-number statistic summary (minimum, Q1, median, Q3, maximum), and needed a method to calculate the mean. Using an estimation function, we generated:

Source:https://www.math.hkbu.edu.hk/~tongt/papers/SMMR2018.pdf

$$\bar{X}(w_1, w_2) \approx \left(\frac{2.2}{2.2 + n^{0.75}}\right)\frac{a+b}{2} + \left(0.7 - \frac{0.72}{n^{0.55}}\right)\frac{q_1 + q_3}{2} + \left(0.3 + \frac{0.72}{n^{0.55}} - \frac{2.2}{2.2 + n^{0.75}}\right)m$$

Using this formula, we estimated the mean for each provided zip code as well as Boston as a city and were able to visulaize the yearly changes in AQI:

Zipcode	PM2.5_2022	PM2.5_2023	PM2.5_Change
2111	6.73753	9.4645	2.72697
2113	7.9084	9.96361	2.05521
2118	8.30109	10.5777	2.27656
2124	7.21404	9.10869	1.89465
2127	7.75032	10.5325	2.78215
2128	7.08768	9.64689	2.55921
2130	7.12675	8.7106	1.58385
2135	6.94934	8.87072	1.92138
2139	5.53174	9.11105	3.57931

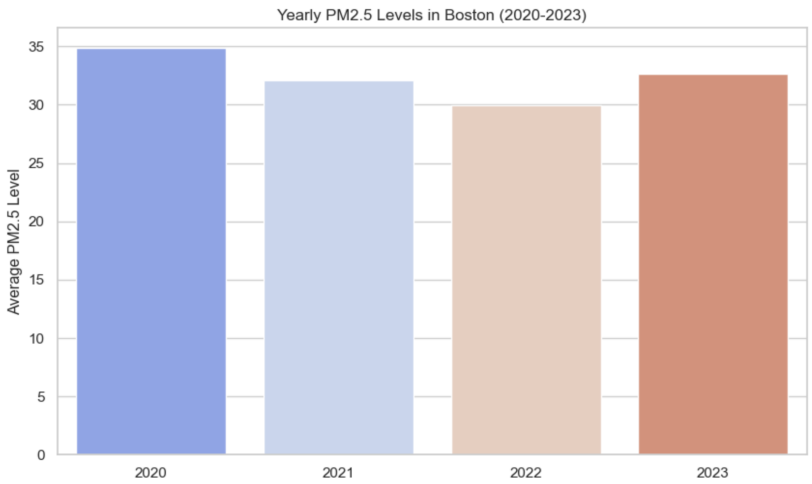


Figure #14*: The above table and bar chart visualizes the PM2.5 changes from year 2022 to 2023 for each of the 9 zip codes we analyzed and from 2020 to 2023 for Boston.*

Step 3: Main Insights: Diseases that are Strongly Correlated with Poor Air Quality:

Depression: The relationship between air quality and depression is complex and can be influenced by various factors. Exposure to pollutants in bad air can lead to inflammation and oxidative stress, which are known to affect mental health. Moreover, living in areas with poor air quality can limit opportunities for outdoor activities, which is important for mental well-being.

Binge Drinking: The direct link between air quality and binge drinking is less clear. However, it's possible that the stress and mental health challenges associated with living in areas with poor air quality could indirectly contribute to behaviors like binge drinking as a coping mechanism.

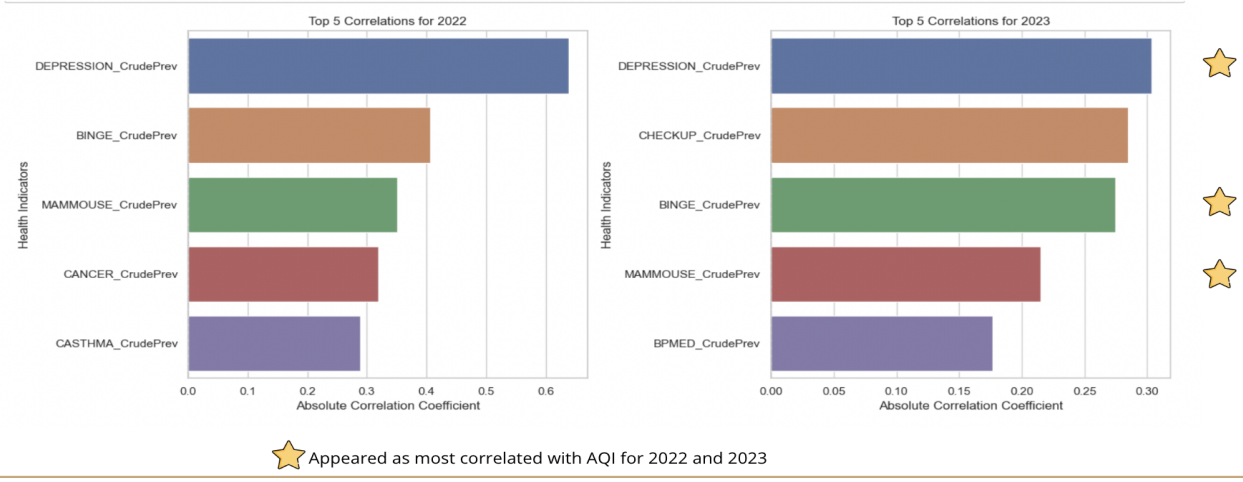
Cancer: Certain air pollutants, such as benzene, formaldehyde, and particulate matter, are known carcinogens. Long-term exposure to these pollutants can increase the risk of developing various types of cancer, including lung cancer.

Asthma: Poor air quality is a well-known trigger for asthma. Pollutants like ozone, nitrogen dioxide, and particulate matter can irritate the airways, leading to asthma attacks and exacerbating existing asthma symptoms.

High Blood Pressure: Studies have shown a correlation between air pollution and an increased risk of high blood pressure (hypertension). Pollutants can cause inflammation and oxidative stress, which can lead to changes in blood vessel function, contributing to hypertension.

Diabetes: Research indicates that air pollution can increase the risk of developing diabetes. This may be due to inflammation caused by air pollutants, which can affect insulin resistance, a key factor in the development of type 2 diabetes.

Kidney Diseases: Recent studies suggest a potential link between air pollution and kidney disease. The kidneys filter the blood to remove waste and excess fluids, and pollutants in the blood can increase the burden on these organs, potentially leading to kidney damage over time.



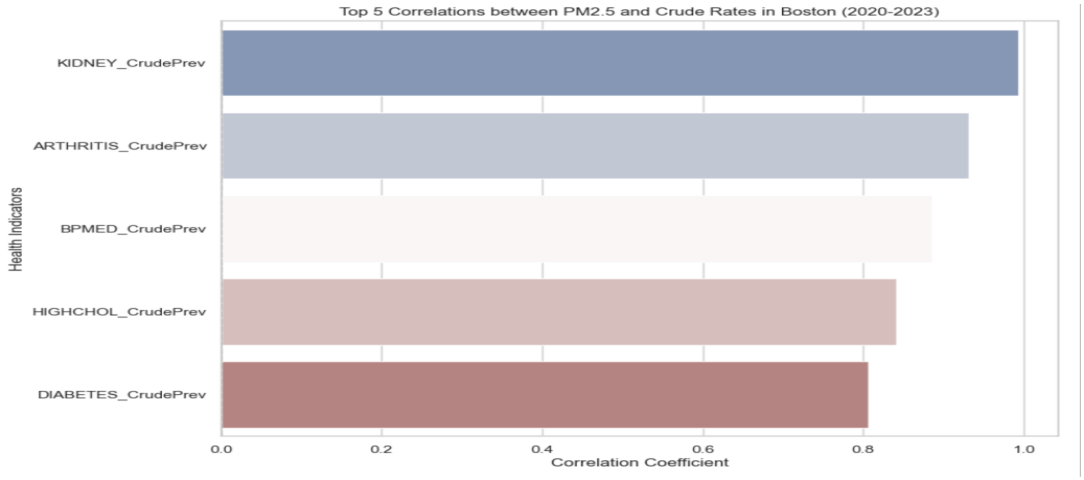


Figure #15*: The charts above use the census health data to examine what are the largest health outcomes of Boston residents based on AQI levels. The most common types of diseases found among residents were Cancer, Kidney Diseases, Diabetes, Depression, High Blood Pressure*

Step 4: Summary

In summary, bad air quality not only affects respiratory health but can also have far-reaching consequences on physical and mental health. The mechanisms vary from direct physiological effects, like inflammation and oxidative stress, to more indirect effects, such as behavioral changes and increased stress levels. The severity of these impacts often depends on factors like the duration and intensity of exposure, individual health status, and socio-environmental conditions. Bad air quality can have a significant impact on various aspects of human health, either directly through physiological changes or indirectly through behavioral or societal factors.

Conclusion

This extended analysis endeavors to delve deeper into air pollution's intricate and diverse impacts. By amalgamating the realms of environmental science, public health, and social considerations, the study aims to unearth intricate patterns that could serve as fundamental pillars for well-informed decision-making and the community's overall welfare. By unraveling the complex interplay between these interdisciplinary factors and the air quality, specifically in the Boston area, we aspire to equip the city with invaluable insights. These insights can pave the way for the development of equitable policies that cater to the diverse demographics within the city, ensuring a fair and balanced approach to environmental and public health challenges faced by different communities. Ultimately, this comprehensive understanding seeks to foster a healthier environment and promote all Boston residents' well-being.

Individual Contributions

Medha

- Explored trends for base question 1: yearly trends in air quality and proximity to transport.
- Looked at transportation data to understand the impact, if any, on overall air quality.
- Worked on the health outcomes and CDC data analysis as part of the extension to explore the health and transportation trend further.
- Worked with other team members to pre-process data and distribute work accordingly.

Dk (Doruk Savasan)

- Fetched the air quality data using the AirNow API.
- Collected all the census data sets. (DP02/03/04/05)
- Collected Health dataset for the extension projects.
- Preprocessed and merged all the necessary datasets.
- Did exploratory analysis in the beginning to better understand the project goal and tools.
- Explored trends for base question 3, the yearly changes in air quality.
- Created and maintained a project repository for collaborative working.
- Started looking at correlations between air quality and health outcomes.
- Discovered 9 new air quality sensors in Boston which I will incorporate into the project soon.
- Explored how air quality relates to diseases on the cdc datasets.
- Graphed how the prevalence of each disease was affected by air quality levels for each zip code and Boston.
- Analyzed yearly air quality changes for Boston from 2020 to 2023 and visualized how each disease was affected by the changes.
- Worked on the report, explained how each disease might relate to poor air quality.

Max

- Explored trends for base question 2, the relationships between air quality population density, housing density, and social vulnerability
- Explored ways to transform geographical data to be in terms of zip codes in boston rather than census tracts or neighborhood names
- Explored some data about crime rates in 2021 to find relationships between crime rates and air quality as a possible part of the extension project
- Ran regressions to find some sort of correlation between crime rates by zip code and their associated air quality, but no significant results
- Analyzed health data by race to find relationships to each racial demographic's social vulnerability

Can

- Explored trends for base question 2, the relationships between air quality and race/ethnicity and area median income.
- Experimented the ways of combining related columns for clearer visualization of findings.
- Explored ways to make our finding more reliable due to the lack of sensors.
- Started working on gathering new AQI data for the extension part of our project.