Deliverable #2 Report

Problem Statement

Examine the influence transportation infrastructure has on the air quality and climate of Boston and surrounding 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
    a. **g250mm_id**: Refers to a 250 x 250 mm area of population in Boston
    b. **nhwi_10**: Refers to the not-hispanic white population based on Census 2010 data
    c. **nhaa_10:** Refers to not-hispanic african-american population based on Census 2010 data
    d. **nhapi_10:** Refers to not-hispanic population asian-pacific islander population based on Census 2010 data
    e. **lat_10:** Refers to Latino population based on Census 2010 data
    f. **nhother_10:** Refers to other racial population based on Census 2010 data
    g. **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**
    a. 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.
        i. Made the column names more readable
        ii. Merged with the AQI data so that everyone can work collaboratively.

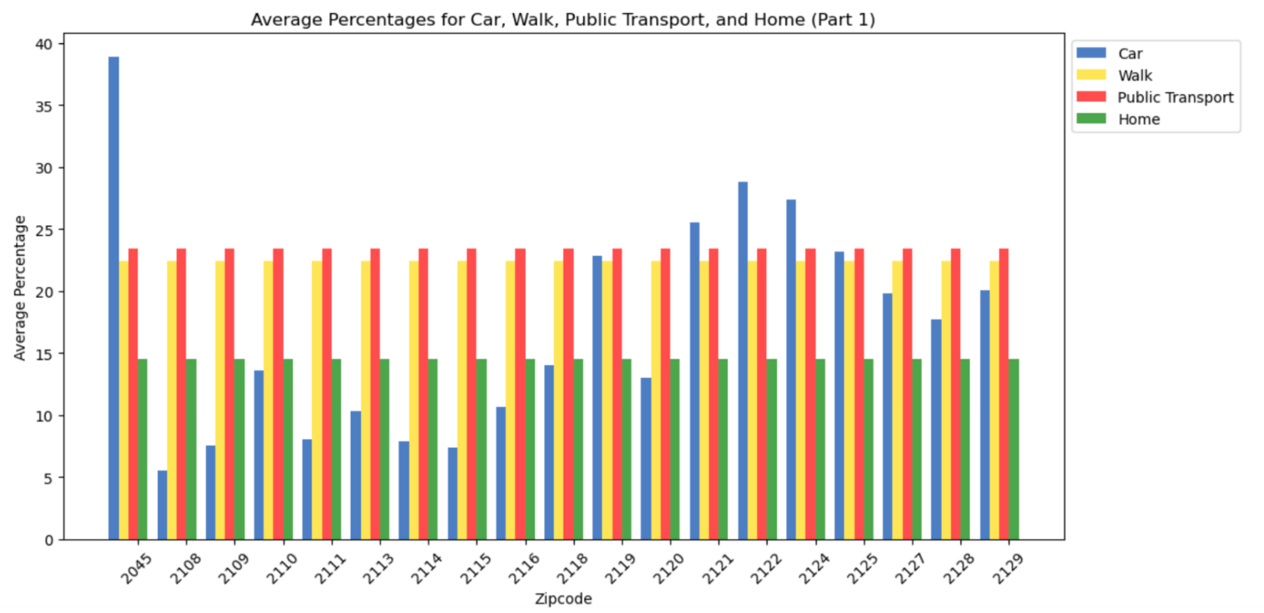
2. **AQI Data**

a. 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.
   i. First we converted the .db file to .csv.
   ii. Got rid of all the NaN values.
   iii. Removed unnecessary features.
   iv. Made a new dataset that contains the average of PM and OZONE values for each zip code.

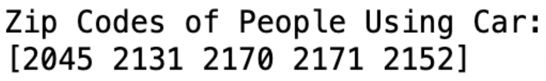
Exploratory Data Analysis - Base Questions:

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?***

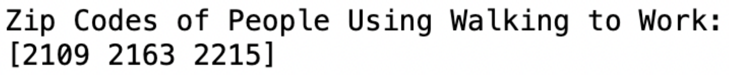
In order to observe the yearly change in air quality for Boston residents, we would need to explore proximity. However, the current PPI data does not include this information. Hence, we explored transportation mediums used in the 45 different zipcodes and explored the impact these had on PM2.5 levels. Based on the grouped bar graphs below, we see that the Car seems to be the most popular medium of transport, 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'.



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 have some level of correlation 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). In order to measure its impact on AQI levels, we attempted to merge our AQI data with transport data to get a deeper understanding of any correlations or trends. We only considered values where the percent 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 were the most prominent. Thus, we can hypothesize that people used a car lived 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.

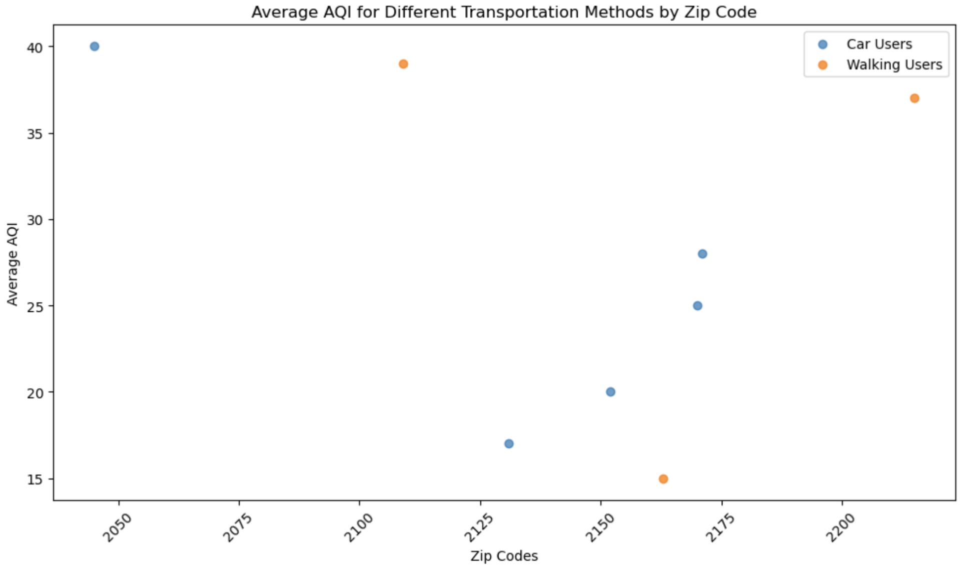


```
Zip Codes of People Using Car:
[2045 2131 2170 2171 2152]
```



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

Once we got these zipcodes, the next step was to explore how these translated to AQI changes, if any. From the graph below, there is no correlation between transport and air quality.

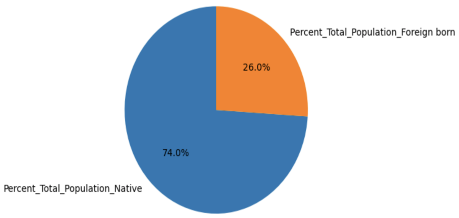


However, this inference might change since we have found new AQI data(derved from 16 sensors) that can inform our analysis better.

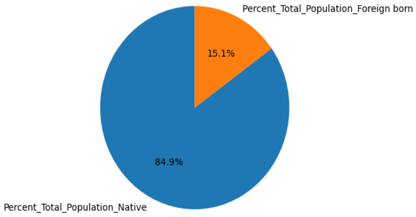
2) ***How do areas with poor air quality compare to areas with better air quality based on different demographic characteristics, specifically:***
   a) ***Race/ethnicity (ACS)?***
      i) I first started the analysis by comparing columns for Natives and Foreigners. There are five different unique ozone and PM2.5 levels. For both metrics we looked at the rows with highest and lowest

ozone and PM2.5 levels. As expected there were multiple rows that have max or min metric levels. So we took the average of the values. In districts with the highest ozone level, while natives constitute 74.0% of the population, foreigners constitute 26.0% of the population. Whereas in districts with the lowest ozone level, while Natives constitute 84.9% of the population, foreigners constitute 15.1% of the population. So we can conclude that 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.

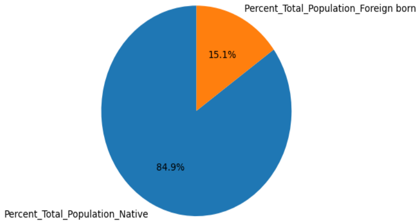
Pie Chart of Percentages of Natives and Foreigns in a Highest-Ozone District



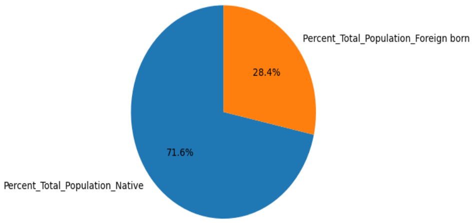
Pie Chart of Percentages of Natives and Foreigns in a Lowest-Ozone District



Pie Chart of Percentages of Natives and Foreigns in Highest-Pm Districts

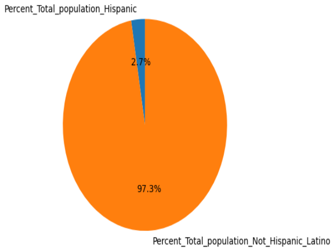


Pie Chart of Percentages of Natives and Foreigns in Lowest-Pm Districts

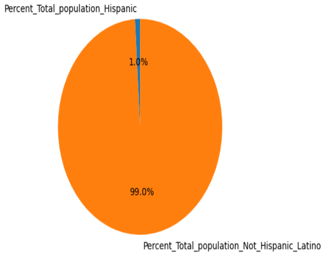


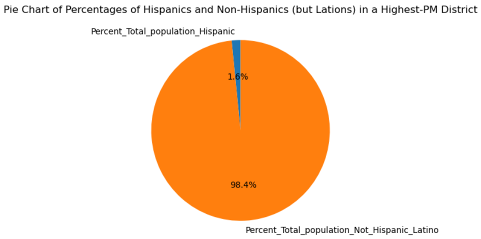
ii) We then proceeded our analysis by comparing districts by looking at columns for Hispancs and Non-Hispanics. Because there were so many columns that belong to this category we decided to combine columns for Hispancs and Non-Hispanics and compared two columns. As can be seen from the pie charts, as 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.

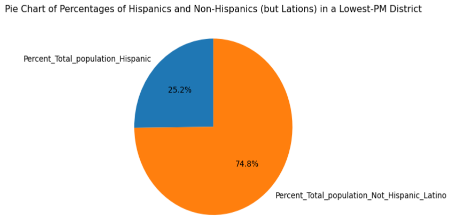
Pie Chart of Percentages of Hispanics and Non-Hispanics (but Lations) in a Highest-Ozone District



Pie Chart of Percentages of Hispanics and Non-Hispanics (but Lations) in a Lowest-Ozone District



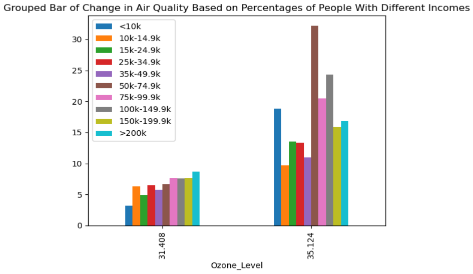

Pie Chart of Percentages of Hispanics and Non-Hispanics (but Lations) in a Highest-PM District

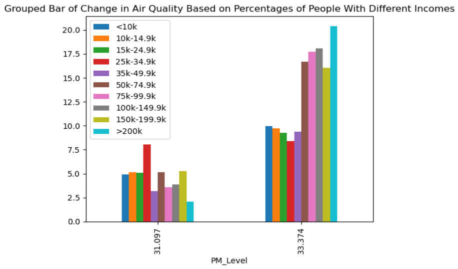

Pie Chart of Percentages of Hispanics and Non-Hispanics (but Lations) in a Lowest-PM District

b) *Area median income/ income*

i) Because we couldn't reach a consistent conclusion when working on columns for race/ ethnicity, we decided to take a different approach. The unique ozone levels are 31.076, 31.084, 31.533, 31.939, 35.124. Acknowledging that the first four ozone levels are close to each other, we thought it would be reasonable to combine and take the average of rows that have these ozone levels. But still, there was an inconsistency when looking at the ozone and PM2.5 metric. Here are the results: by looking at the ozone metric, even though there is no a 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.

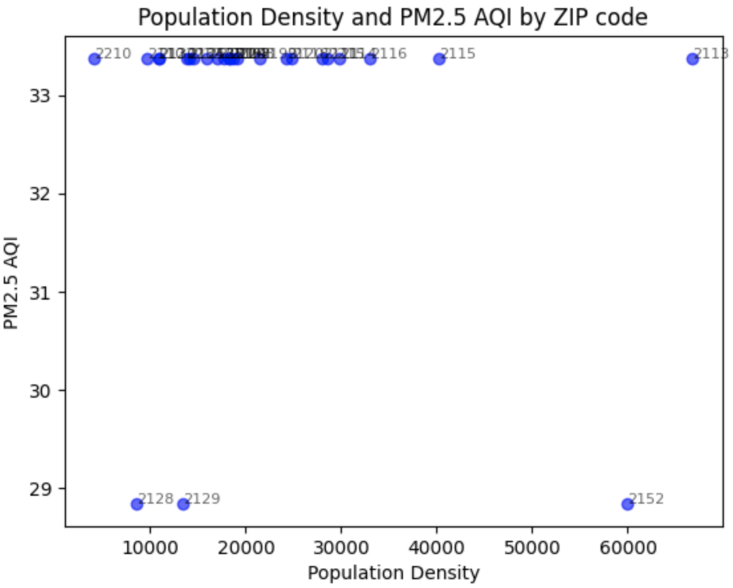
Therefore, in order to make a concise conclusion, we need to further investigate what are the biggest factors of ozone and PM2.5 metrics which exceeds the scope of our project.

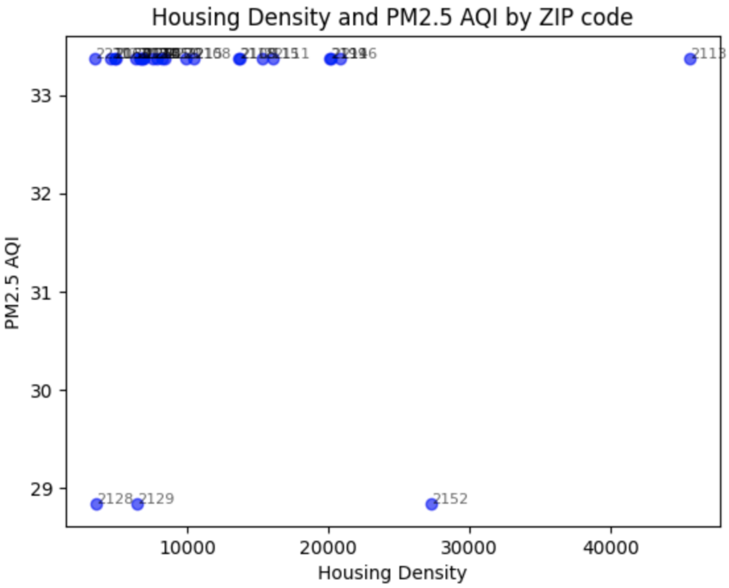

Grouped Bar of Change in Air Quality Based on Percentages of People With Different Incomes


Grouped Bar of Change in Air Quality Based on Percentages of People With Different Incomes

c) *Housing & Population Density*

i) Overall, there is not much of a correlation between PM2.5 AQI and housing density or population density, with the exception of the outliers. This means that most densely populated and densely housed areas do have relatively worse of a 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

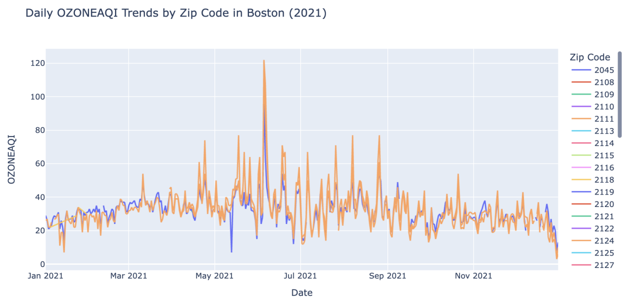


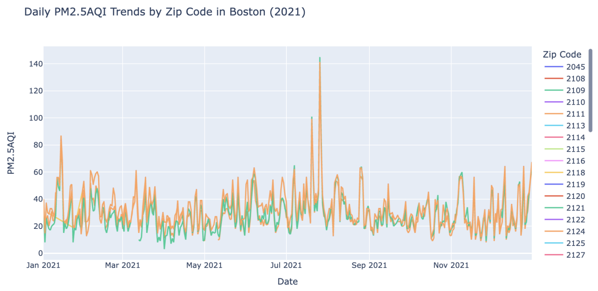


d) ***Social vulnerability***

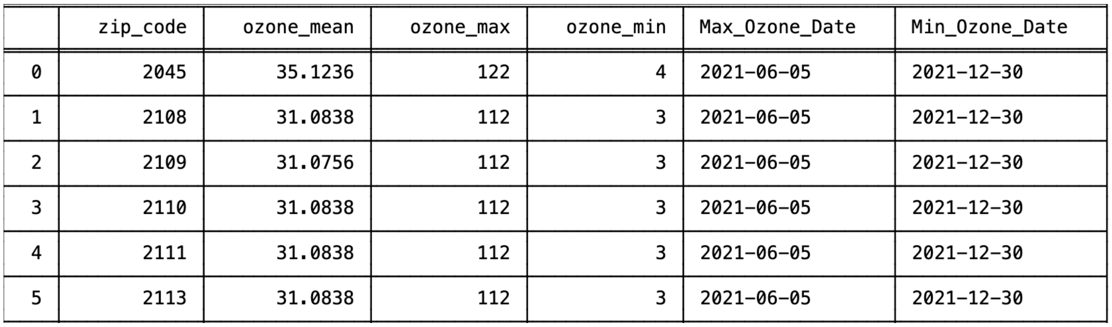
i) Given that Boston is demographically quite diverse with 53% people of color, comparing the social vulnerability factor between different races have provided deep insights into the disproportionate effects of air quality and proximity to transportation between white people and people of color. The 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.

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.***

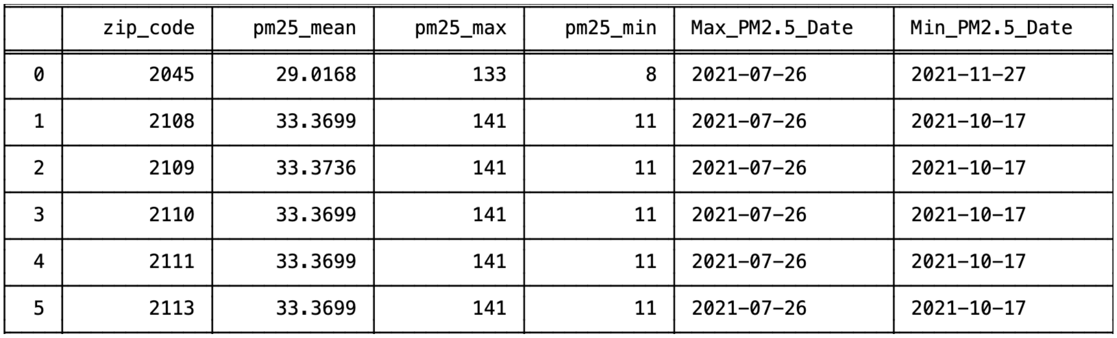




The analysis primarily focuses on two pollutants - Ozone (OZONEAQI) and particulate matter of size 2.5 microns or less (PM2.5AQI). The data was grouped by zip code and date to calculate mean values. Interactive plots above were created to show daily trends in OZONEAQI and PM2.5AQI across different zip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



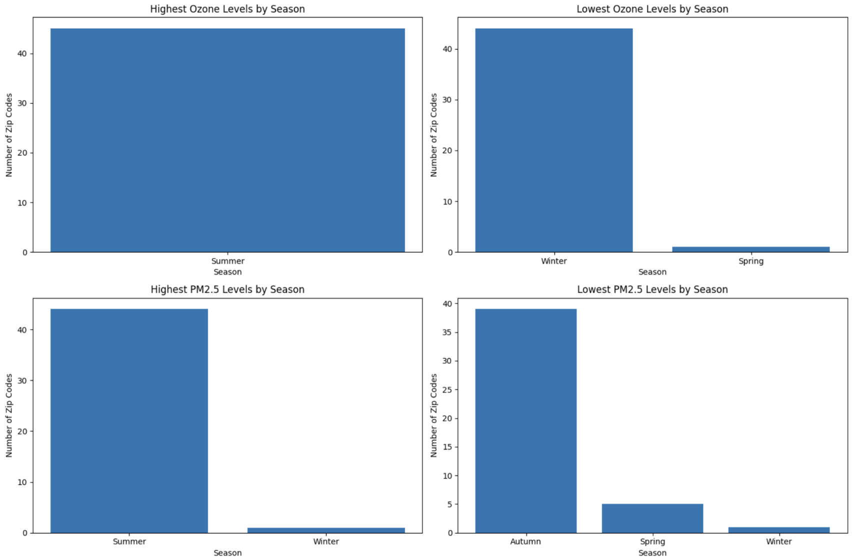
	zip_code	ozone_mean	ozone_max	ozone_min	Max_Ozone_Date	Min_Ozone_Date
0	2045	35.1236	122	4	2021-06-05	2021-12-30
1	2108	31.0838	112	3	2021-06-05	2021-12-30
2	2109	31.0756	112	3	2021-06-05	2021-12-30
3	2110	31.0838	112	3	2021-06-05	2021-12-30
4	2111	31.0838	112	3	2021-06-05	2021-12-30
5	2113	31.0838	112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



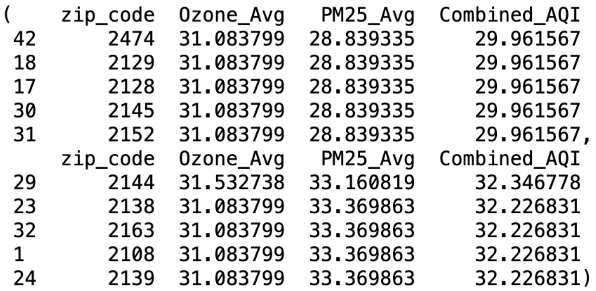
	zip_code	pm25_mean	pm25_max	pm25_min	Max_PM2.5_Date	Min_PM2.5_Date
0	2045	29.0168	133	8	2021-07-26	2021-11-27
1	2108	33.3699	141	11	2021-07-26	2021-10-17
2	2109	33.3736	141	11	2021-07-26	2021-10-17
3	2110	33.3699	141	11	2021-07-26	2021-10-17
4	2111	33.3699	141	11	2021-07-26	2021-10-17
5	2113	33.3699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

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.



The average OZONEAQI and PM2.5AQI 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)
```

After exploring these base questions, we feel that our analysis can be solidified further by exploring data relating to the impact air quality has on health outcomes of residents. We found health data based on Boston zipcodes with provides insights into the health outcomes of Boston residents. As part of the extension, we aim to exploring the relationship between AQI and health outcomes to discover trends.

Extension Project

We propose an extension of the project to delve deeper into the societal implications of air pollution. This study will focus on understanding the relationship between air quality changes (specifically PM 2.5) and health outcomes, as well as crime rates(if there is enough time) in Boston.

Objectives

1. **Health Impact Assessment:** Investigate the correlation between yearly changes in air quality and various health outcomes, such as asthma, lung cancer, and other pollution-related health conditions.
    a. **LocationID:** 45 Boston zipcodes
    b. **Category:** Health Categories(Preventive, Outcome, Risk Behavior)
    c. **Small_Question_Text:** Description of the specific health outcomes of residents

2. **Air Quality and Crime Analysis**: Explore the potential relationship between air quality fluctuations and crime rates.

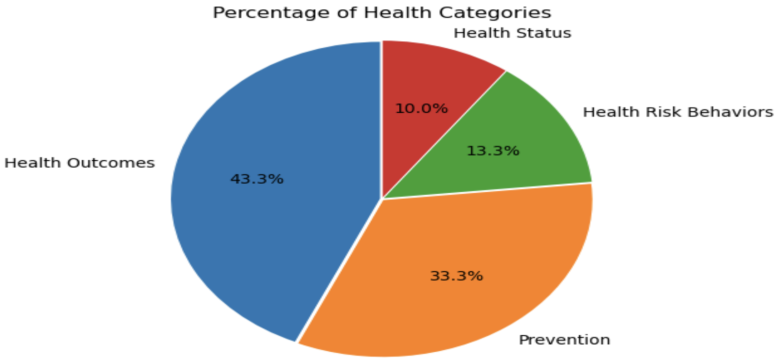
Methodology

Data Collection

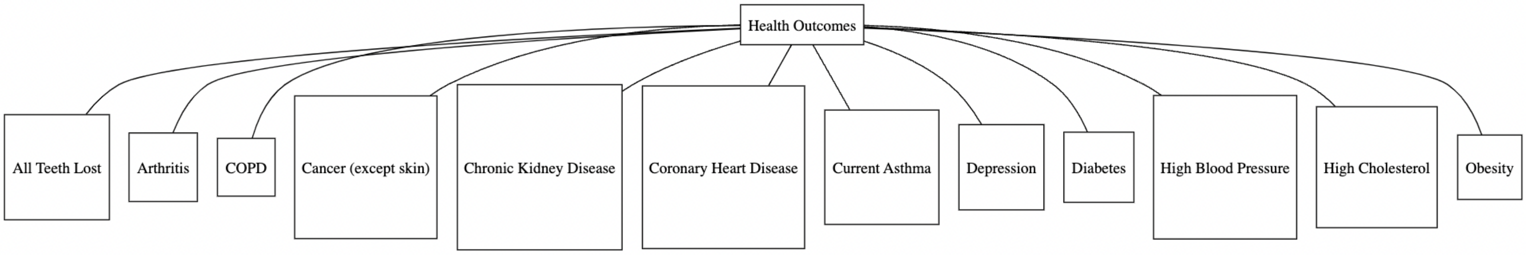
Health Data: We will gather data on respiratory and cardiovascular diseases, lung cancer incidents, and other relevant health issues from local health departments and national databases.
Crime Data: Crime statistics will be sourced from Boston's law enforcement agencies or publicly available crime databases.

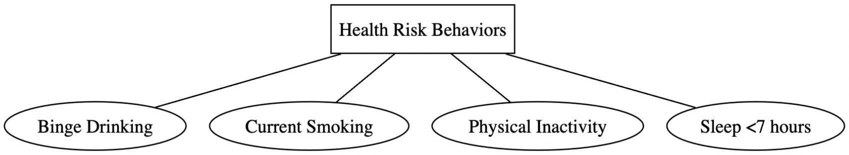
Exploratory Data Analysis

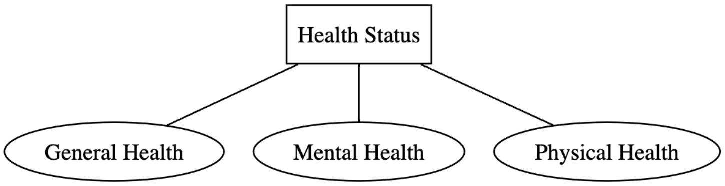
So far, we have conducted some EDA for health data to explore the health outcomes for Boston residents. Before we can explore the impact of air quality on health outcomes of residents, we need to understand the trends dispalyed in the health data.



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.







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.

Next Steps:

Base Questions

1. In order to give a more in- depth answer to the first base question, the following steps can be taken: Basically, the first base question asks the yearly change between air quality based on proximity to different types of transportation infrastructure. However, the given PPI dataset doesn’t include useful information regarding the proximity. And further search for this information from different datasets seems not to yield any useful results too. So one way to deal with this is to estimate the proximity by the modes of transport from the Transport dataset. If this is done for each zip code, we would plot yearly change versus proximity. Eventually, we can create a linear regression of a polynomial regression model to answer this question.
2. We found more air quality sensors in Boston which will be used to estimate the corresponding zipcode’s air quality measures. This way we will have more reliable data which will make our findings more accurate.

Extension Project

1. **Correlation Analysis:** Use statistical methods to analyze the correlation between air quality data and health outcomes. Techniques like Pearson correlation and regression analysis will be pivotal.

2. **Spatial Analysis:** Employ GIS (Geographic Information Systems) to visually represent the spatial distribution of health outcomes and crime rates in relation to air pollution levels.
3. **Temporal Analysis:** Analyze how these relationships (health outcomes and crime rates) vary over different times of the year and compare these patterns with air quality trends.
4. **Predictive Modeling:** If a significant relationship is identified, develop predictive models to forecast health outcomes or crime rates based on air quality data.

Impact

1. **Public Health and Policy Making:** Insights from this study can guide policymakers in implementing targeted interventions to mitigate the health impacts of air pollution.
2. **Community Awareness and Education:** The findings can be used for community education, emphasizing the importance of air quality and its direct implications on health and social well-being.
3. **Academic and Scientific Contribution:** This extension will contribute to the academic field by providing a comprehensive analysis of the environmental, health, and social dimensions of air quality.

Conclusion

This extended analysis aims to deepen our understanding of the multifaceted impacts of air pollution. By integrating environmental science with public health and social considerations, the study seeks to uncover patterns that could be pivotal for informed decision-making and community welfare. Through exposing relationships between these interdisciplinary factors and the air quality around Boston, we hope that the city will be better equipped to ensure equitable policies for all demographics.

Indiviual 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.
- Also focussed on exploring health outcomes data based on zip codes,which is part of the extension project.

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 15 new air quality sensors in Boston which I will incorporate into the project soon.

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

Can

- Explored trends for base question 2, the relationships between air quality and race/ethnicity and area median income.
- 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.