Deliverable 2 Report

Sinforiano Terada, Naveen Vaidyamath, Jonathan Suarez, Chao-Jen Chiu, Zixuan Wang

City of Boston: Transit and Air Quality B

Table of Contents

Problem Statement, Data Cleaning and Collection steps

The goal of this project is to analyze yearly health data and proximity to diverse transportation infrastructure such as public transport and roads to investigate the relationship between these data and the disparate impacts on the residents of Boston.

For data cleaning and collection, we had a different process for each source. The base project had given some links to collect data for air quality, transit, and demographics, health and income data. The given Pollution Proximity Index (PPI) data was a simple CSV download where all we had to do was filter out some null values. The social vulnerability index was similar to the PPI data where it was a set provided by the City of Boston that was simply downloaded and cleaned before use. For air quality sensor data, we used the AirNow API to collect the yearly data for 30 zip codes. We have found many constraints with collecting this data, such as the API only allowing us to request one day's worth of data for one zip code every 8 seconds. This means that collecting the data took 8 seconds x 365 days x 30 zip codes = 87600 seconds, which totals over 24 hours. Luckily, we found a way to fetch the data in just 5 hours by running multiple python scripts to collect a zip code with different API keys. However, after collecting that data we found that all of the AQI (air quality index) and OZONEAQI (O-zone AQI) to be same, correlating to each day of each zip code, this is an ongoing issue that we are currently addressing with other teams in our project and the TPMs and professor.

Using the cleaned data, we conducted an analysis of population data using the Census API, calculating both housing and population density for Boston zip codes. We visualized the results by creating a shapefile which allowed for a more intuitive presentation. The final outcome is a map of Boston's counties delineated by zip codes and color-coded from tan to dark red to indicate the varying levels of population and housing density. The darker shades represent higher values.

For the social vulnerability dataset, we transferred the names of cities to their belonging zip codes and extracted those zip code’s data that are also included in the air quality data which we downloaded. The further analysis about social vulnerability was built based on the extracted data.

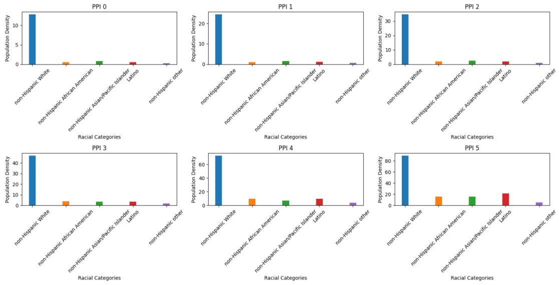
Exploratory Data Analysis

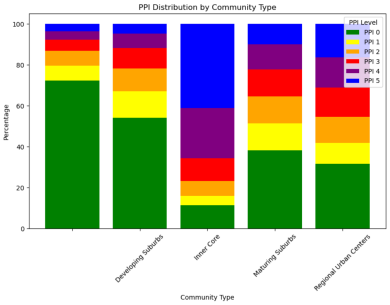
Our exploratory data analysis included looking at how air quality changes among racial demographics, air quality based on income and zip code, housing density and air quality based on zip code, and how air quality might relate to social vulnerability.

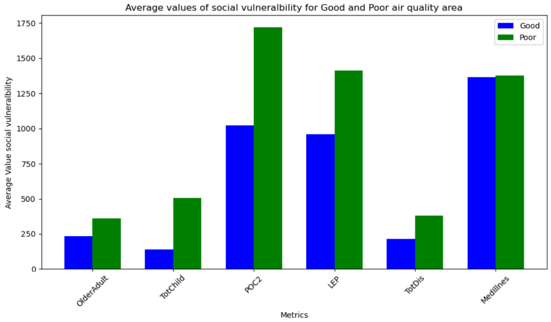
The PPI data ended up having the most impressive findings. The first graph shown below shows how the population density of different racial groups changes for each PPI level (0-5 with 5 being the worst air quality). The population density of minority groups was considerably higher in PPI levels 4 and 5 compared to that 0-3. The second graph shows the PPI levels based on where the grids were located. Each location in the data was assigned a community type that indicates the location's proximity to the city. As seen in the graph, the inner core communities have by far the worst air quality.

The third plot demonstrates social vulnerability. We counted and ranked the data according to the Category Name from each zip code (including Good and Moderate). We extracted half of the zipcode data with the most Good as the good group and the other half as the poor group, and plotted a bar graph based on the zip code data they belonged to in terms of social vulnerability. These social vulnerability indexes include Older Adult, Children, People of Color, Limited English Proficiency, Low to No Income, People with Disabilities, and Medical Illness. We can observe that these indicators for the good group are usually smaller than those for the poor group. This suggests that Social Vulnerabilities play an important role in air quality.

As mentioned in the previous section, the other data explorations did not show anything interesting since we had issues with the air quality sensor data, which we will elaborate on in the next section.


PPI 0
PPI 1
PPI 2
PPI 3
PPI 4
PPI 5
Population Density
Racial Categories
non-Hispanic White
non-Hispanic African American
non-Hispanic Asian/Pacific Islander
Latino
non-Hispanic other
0
5
10
20
30
40
60
80


PPI Distribution by Community Type
Percentage
0
20
40
60
80
100
PPI Level
PPI 0
PPI 1
PPI 2
PPI 3
PPI 4
PPI 5
Developing Suburbs
Inner Core
Maturing Suburbs
Regional Urban Centers
Community Type


Average values of social vulneralbility for Good and Poor air quality area
Average Value social vulneralbility
0
250
500
750
1000
1250
1500
1750
Good
Poor
OlderAdult
TotChild
POC2
LEP
TotDis
Medilines
Metrics

Finding Underlying Patterns and Base Questions

Base Question 1: 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.

With regards to race and ethnicity, we can see from the PPI data that as air quality decreases, the proportion of minorities living in these areas increases. While the population density of nonHispanic Whites remains the dominant racial group throughout all of the PPI levels, the data shows a drastic increase in the number of all other racial categories particularly between levels 3 and 4, and the numbers increase even further between levels 4 and 5.

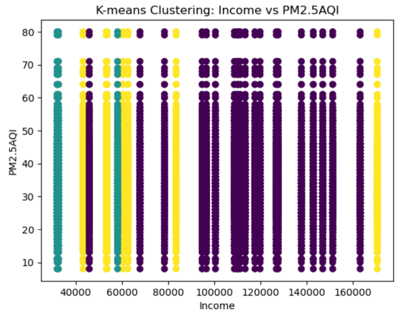
We were able to demonstrate the median income by zip code in Boston to get a better understanding of which areas are considered wealthier and other areas poorer. As we can see with the K-means clustering graph, It doesn't seem that the air quality index differs as much by zip code, as it is mostly clustered in the 20 to 60 range. This could be due to the fact that 2020 is when the COVID pandemic happened, and transit activity was at its lowest it has ever been in the history of the United States. However, there was still typical transit activity in the months January to March, so there is opportunity to perform more in-depth analysis on the true Air Quality for those months.

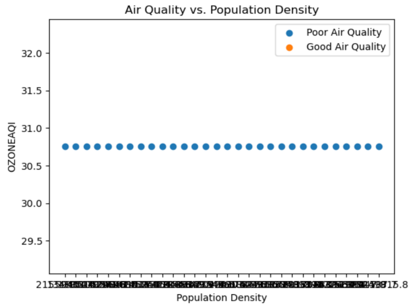
As demonstrated through our shaded map analysis, we've uncovered interesting trends in Boston's housing density. The northeastern region of Boston stands out with the highest housing density, indicating a greater concentration of houses per square mile. In contrast, as we move south and west, the housing density decreases, suggesting less compact living conditions in those areas. What's particularly noteworthy is the correlation between the darker-shaded regions on the housing density map and some of the more expensive areas on the median household income graph. This correlation suggests that areas with higher housing density often coincide with higher-income neighborhoods in Boston The implications of this correlation are significant, indicating a potential connection between housing density and income levels in Boston. It raises the possibility that higher housing density areas might be more appealing to individuals with higher incomes, potentially due to factors such as proximity to urban amenities and employment opportunities. This finding holds considerable weight for urban planning and housing policy decisions, especially concerning air quality and transit. To provide a seamless transition to our conclusion, it's crucial to recognize that understanding the relationship between housing density and income can significantly inform decisions related to affordable, healthy urban development, as well as investments in infrastructure. Such decisions can collectively contribute to creating more equitable living conditions for all residents of Boston, ultimately fostering a vibrant and inclusive urban landscape."

We can use zip codes to represent areas with higher and lower population densities. Next, we analyze the data in conjunction with the AQI data, creating a scatter plot to determine the correlation between the two. Based on the current data, correlation between Population Density and OZONEAQI/PM2.5AQI : 0. Based on our shaded map analysis, we have uncovered trends in Boston's population density. The northwestern and southeastern regions of Boston stand out with the highest population density. In contrast, other areas show less significant differences in population density. In general, we can expect areas with high population density to be lively urban areas with more developed infrastructure, potentially resulting in poorer air quality.

Regarding the social vulnerability comparison between good air quality area and poor air quality area, we compared them on the following metrics: older adults(OlderAdult), children(TotChild), people of color(POC2), limited English proficiency(LEP), lower income(Low_to_No), people of disabilities(TotDis), and medical illness(MedIllness). We analyzed the zip codes based on the overlap of the zipcodes in the social vulnerability dataset that are assigned by the place names and the zip codes present in our air quality dataset, which is a total of 10. First, we counted the zip codes according to their CategoryName (which includes whether the daily air quality is Good or Moderate), and based on the ratio, we classified the 5 zip codes with the most Good as good air quality areas, and the 5 zip codes with the least Good as poor air quality areas. Based on this, we calculated an average for each metrics for the good and poor groups and displayed the result in a bar chart. We can see that in each of the metrics comparisons, the data for the POOR area is much higher than the GOOD area. This shows that air quality has an impact on social vulnerability metrics. There is a strong correlation between the presence of poor air quality and people such as children, elderly, and low income adults who live in those areas.

Now for the aforementioned problems. We collected the air quality data based on Boston area zip codes. The collection was distributed across the different groups because it took a whole group's computers running for hours to collect a single year's worth of data. After collecting this data, we all noticed that there are no meaningful differences across the years and zip codes in air quality. Our plots for the air quality data are shown below.





The base questions “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?” and “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?” would be answered as “there is no difference” because all of the data across all zip codes and years are the same. Since there is no yearly change in air quality, our answer to the first question is “There is no change” and the answer for the second question is “There is no relationship because there are no trends in yearly changes”.

Extension Proposal

We had to modify our extension proposal a little bit since we had so much trouble with the air quality sensor data. Although we had viable air quality data with the PPI data set, it did not have any sort of identifying information as to where each row is located. The only sort of information given was the data point's "Community Type". This data set was provided by the Metropolitan Area Planning Counsel (MAPC), which is a collection of over 100 Boston-area towns in eastern Massachusetts. We were able to dig around on their site and found a data set that maps each town to its "Community Type" classification. Our next step will be to create a crosswalk data set where we find the zip codes for all of the towns and map them to their "Community Type" classification. This way, we can try to map our income by zip code, population density by zip code, and housing density by zip code findings to the "Community Type" classifications to see if there are any conclusions we can draw at the "Community Type" level. Furthermore, we can investigate how accessible public transit is in each "Community Type" as well as what type of public transit is available in each classification.

If we are able to create our crosswalk table, we would be able to stick to the spirit of the original extension project. One short-coming will come in analyzing annual changes because the PPI data is only for 2020. As a byproduct of this shortcoming, will aso not be able to analyze health outcomes from this data because of the constraint on when this data was collected. Furthermore, the CDC data given to us does not specify location to a degree necessary for this project.

Our question to answer for our extension project is: "What is the relationship between the community types (zip codes mapping to community types) in Boston and the PPI data with regards to race distributions, median income, and MBTA infrastructure?

Visualization and Insights for Extension Proposal

Since most of our energy has been spent coming up with a suitable modification, we have not yet started analyzing the data yet. In lieu of presenting findings, we will now present our new data sets to further clarify our extension proposal.

Link 1 below shows all of the towns in the MAPC and their Community Type on page 4. The towns right of the red boundary are the ones in the MAPC and the towns that we will be focusing on for the rest of the project. These towns are all listed in link 2. This dataset that we will use as the basis for our crosswalk table. We will be focusing on the columns “Municipality Name”, “Community Type (2008 Classification)”, “Regional Planning Authority Acronym”, and “MBTA Community Type”. We will focus on the towns with the “Regional Planning Authority Acronym” as MAPC, as well as use their “MBTA Community Type” to aid in our analysis.

Link 3 is where we will likely find the best help in mapping the zip codes. We have used this website previously to help look at data and gain some insight before creating our own analysis. We will use this website again for their free zip code database to help our mappings.

Finally, we will use the same census data, social vulnerability data, and MBTA transit data that we used in Deliverable 1 to continue our analysis of these topics using the Community Type geographic descriptor as given to us by the MAPC in the PPI data.

[1]https://www.mapc.org/wp-content/uploads/2017/09/Massachusetts-Community-Types-Summary-July_2008.pdf

[2] https://datacommon.mapc.org/browser/datasets/444

[3]https://simplemaps.com/city/boston/zips/education-college-or-above

Individual Contributions

Sinforiano Terada:

I wrote almost all of this report along with all of the extension proposal in addition to the PPI data analysis from deliverable 1 and helping collect the AQI data (also part of deliverable 1).

Zixuan Wang

I helped to collect part of the AQI data and MBTA dataset. Also, I add up the data preparation part with what I have done for analysis and my analysis process for social vulnerability part.

Jonathan Suarez

I helped write the report as well as provided some direction on where to go with the project by discussing in the slack channel with different teams. I primarily helped by collecting almost all of the air quality data and spending hours coming up with the right strategies to reduce the fetching time by 20 hours.

Naveen Vaidyamath

I was responsible for acquiring the census data and using it to generate information on housing and population in Boston. My final result was a shapefile map of Boston, which was shaded to represent variations in population and housing density xacross the city. We utilized this map to draw conclusions about the diverse communities in Boston and to explore correlations and impacts related to other community factors, such as air quality, race, public health, and more. Also produced the presentation.

Chao-Jen Chiu

I was responsible for collecting AQI data, including attempting to use other APIs such as the Purple Air API to obtain different AQI datasets, although the results were not satisfactory. Additionally, I was also responsible for issues related to population density.