Report and Code Submission

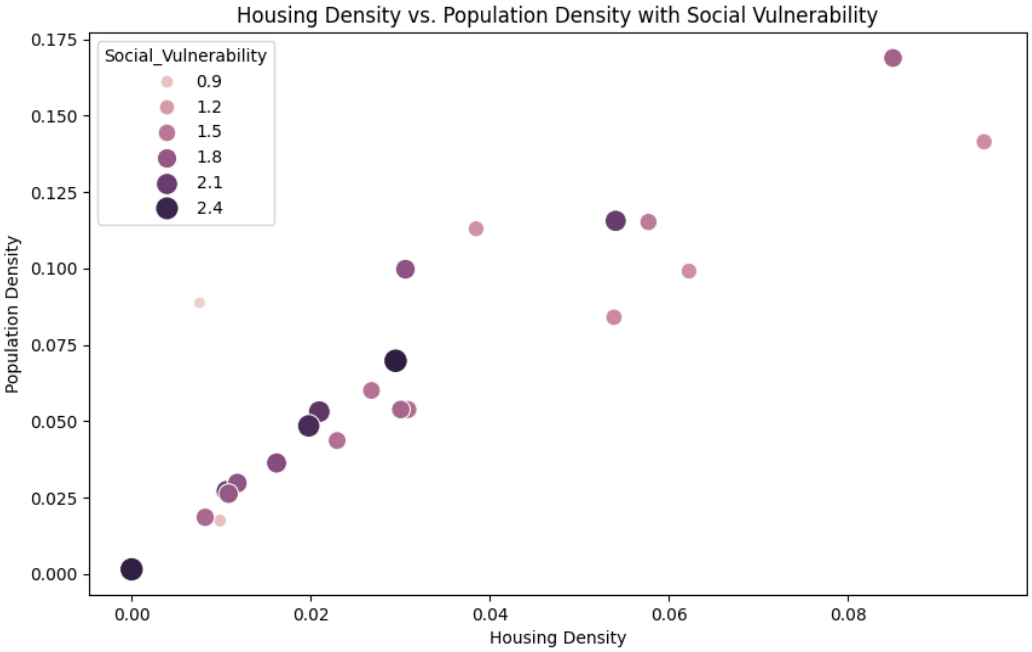
For each deliverable, you must submit both a report and the associated code. Your report should include the following sections for Deliverable 1:

1. A brief introduction to your problem statement.
2. Details of the data collection or cleaning steps you've undertaken.
3. Exploratory Data Analysis (EDA).
4. If your analysis has led to answers for any of the questions or if you've formulated hypotheses, especially for at least questions.
5. Individual contributions of each team member. We recommend that each team member writes 3-4 lines about their contributions, which can then be compiled into the report.

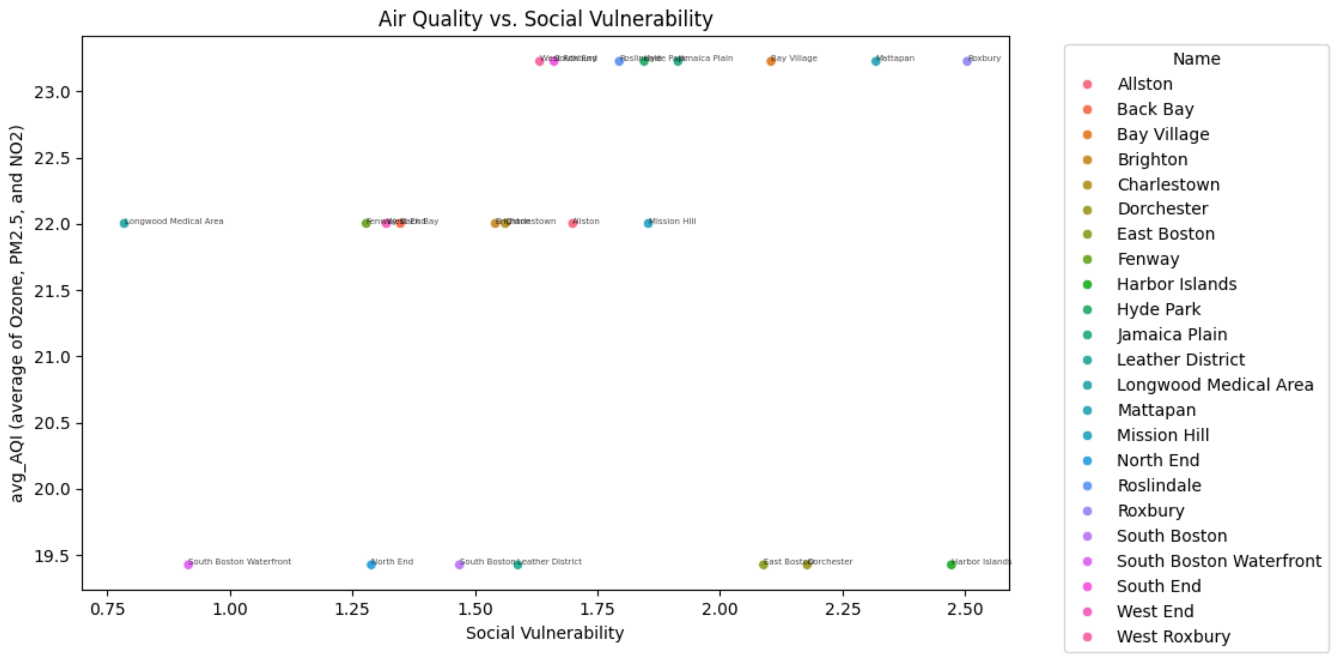
1. Boston's dense transportation network raises concerns about its impact on air quality. As car emissions dominate Massachusetts's environmental challenges, it's crucial to understand their relationship with urban air pollution. Furthermore, the socio-economic disparities in experiencing this air quality remain a concern. This project aims to investigate the link between transportation infrastructure and the air quality, and its varied effects across neighborhoods.
2. Our work thus far has centered around three main sources of information:
    a. Air Quality Index (AQI): Using a bash script, we fetched the data from the AirNow API to collect air quality readings for every hour of every day for all of 2021. We collected all the readings into a single file to begin preprocessing. To preprocess the collected data, we first remove rows and columns that only contain `NaN` values (e.g. PM10_value column). Then, we focus on the columns that we are interested in (columns with AQI values for PM2.5, NO_2 and OZONE) by creating a new dataframe that has these columns plus the name of different neighborhoods. There are lines with null values, so we filled the null value with the mean of the corresponding feature. After doing all the steps mentioned above, we grouped the rows with the same neighborhood and applied a mean function to show the average AQI of that specific neighborhood for 2021.
    b. Pollution Proximity Index (PPI): For the proximity to roads csv file we wrote code to analyze and visualize pollution proximity data for Boston. After loading the dataset from a CSV file, it provides a basic inspection by printing the top 5 rows and summary statistics to an output.txt file. It generates a histogram of the PPI and a bar chart depicting the average PPI by commute type, saving both as image files. It also utilizes the Folium library to create two maps: a basic map of Boston and a heatmap based on the PPI, latitude, and another column and then saved as HTML files.
    c. Social Vulnerability Index: For this file, we began by downloading data from Climate Ready Boston to get general geographic information and social vulnerability information for each location. Using this, we converted the given

data to get the percentage of the population for each socially vulnerable category, as well as the housing and population densities. This data was then aggregated to get an overall social vulnerability index for every neighborhood (using equal weights for all categories, though this can easily be changed).

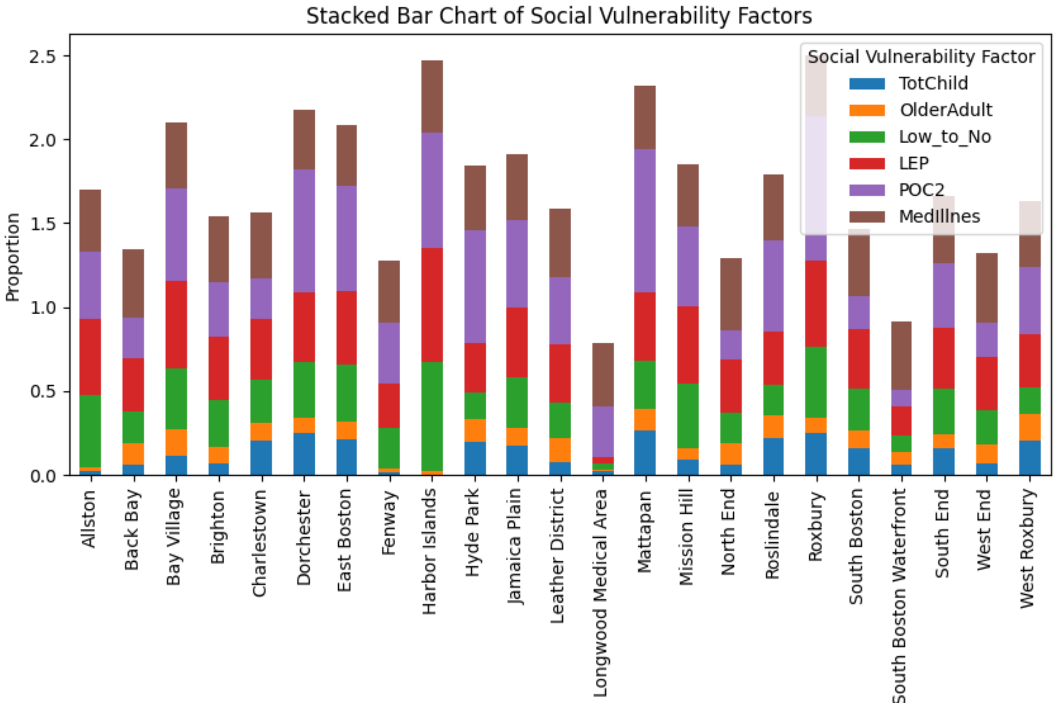
3.



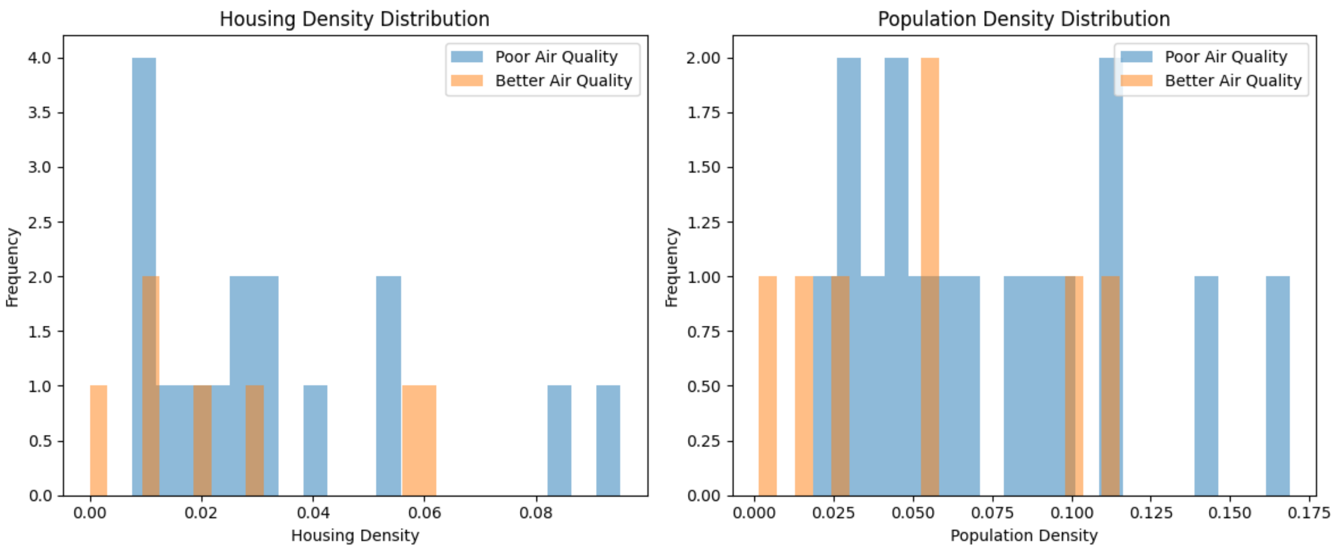
- Graph maps the housing density (x-axis) against the population density (y-axis) of each Boston neighborhood, and plots points with varying size based on each neighborhood's Social Vulnerability Index.
- We see that housing density and population density are very strongly correlated, while in general–and ignoring the possible outlier close to (0,0)–neighborhoods tend to increase in social vulnerability as density increases up until around (0.03, 0.075), then the social vulnerability of neighborhoods begins to decrease as density increases.



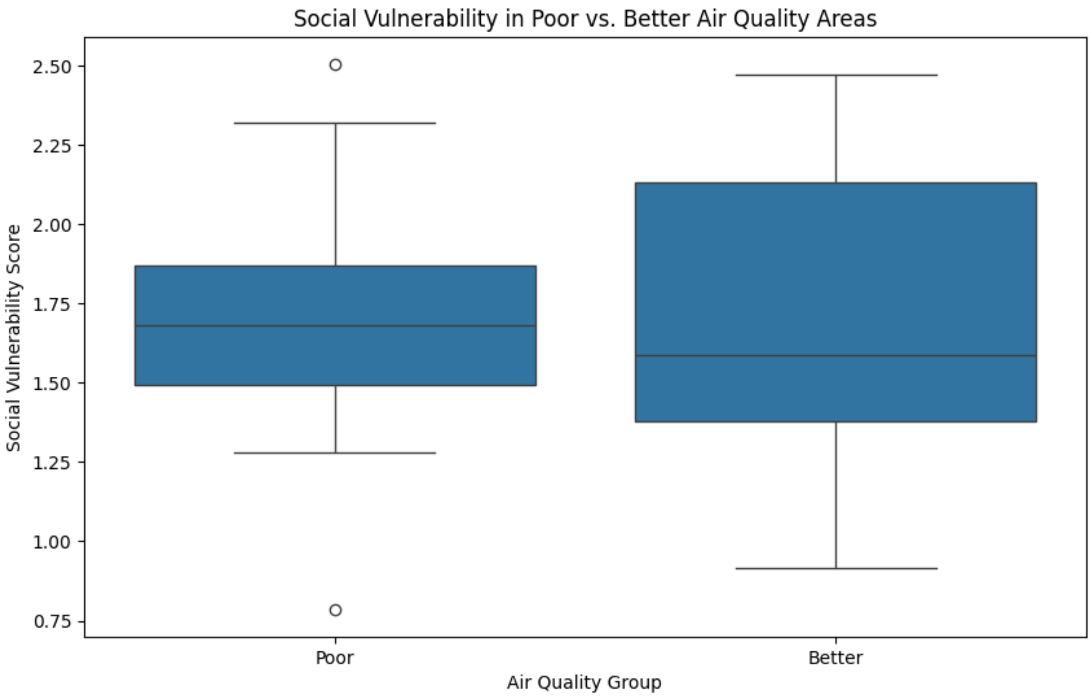
- For each neighborhood, we plot the social vulnerability against the average of Ozone, PM2.5 and NO2 AQIs. We notice the clear three horizontal lines for average AQI since we only have three active monitoring sites.
- Based on the graph, we are not able to clearly spot a correlation between social vulnerability and air quality. Those with higher air quality tend to have a larger variance and those with lower air quality tend to be more vulnerable. But beyond that we are not able to draw clear conclusions.
- The graph may have generalizability issues due to the lack of monitoring sites we currently have. Resulting in us grouping multiple neighborhoods into one monitoring site by distance.



- The graph above takes every neighborhood in Boston and shows the proportion of the five social vulnerability factors. In the 23 neighborhoods, there is not a strong correlation with age.
- In every neighborhood, POC2 (people of color) and MedIllness (people with medical illness) take about the same proportion of the index. In comparison to the total population overall, people of color statistically have lower income, which means that a higher percentage of their population is near or around the poverty line,
- Individuals with lower income may be restricted to which neighborhoods they can choose. This means that people may be unable to afford housing in a more excellent neighborhood or air purifiers. Additionally, if the neighborhood has lousy air quality, people with medical issues will be more likely to suffer.



- For all neighborhoods in Boston, the AQI values are below 50, which is considered good air quality. In order to better compare the air quality of all neighborhoods, we set a threshold of 20 to distinguish between neighborhoods with good and better air quality. The above graphs describe the relationship between housing density/population density and air quality.
- All neighborhoods with better air quality have relatively low housing and population density, while for neighborhoods with relatively worse air quality does not have a clear trend, rather the data points can be found in almost all housing and population density.



- The box plot shows the social vulnerability score for poor air quality groups and better air quality groups.

  Poor Air Quality Group: the line inside the box indicates the median is around 1.65, which means that half of the areas with poor air quality have social vulnerability score 1.65 or lower. The data range mainly from 1.5 to 1.9. There is an outlier near the social vulnerability score 0.75, which indicates that there is a poor air quality area has relatively low social vulnerability score

  Better Air Quality Group: the line inside the box indicates the median is around 1.55, which means that half of the areas with better air quality have social vulnerability score 1.55 or lower. The median is closer to the bottom of the box which indicates it is a right-skewed distribution and the data is more dense at the lower side of the median.

  Comparing these two air quality groups, the medians for both groups are relatively close but the better air quality group has slightly lower social vulnerability score. However, there isn't a strong contrast between air quality groups and social vulnerability scores.

4. As of writing this (10/29) so far our analysis has not led to a concrete answer to any of the questions, however we have formulated some hypotheses for question 2: **How do areas with poor air quality compare to areas with better air quality based on different demographic characteristics?**
   a. Hypothesis: Higher population density strongly correlates with higher housing density, and appears to be loosely correlated with a higher social vulnerability index. We predict that higher population/housing density will be correlated with worse air quality in general due to increased density causing worse traffic and congestion for all transportation systems.
   b. Hypothesis: There appears to be little to no correlation between social vulnerability and air quality of neighborhoods. Neighborhoods with better air quality appear to have a greater variance in terms of social vulnerability, but beyond that it is hard to draw a strong conclusion between these two factors.

5. Individual contributions: (3-4 sentences)
   a. Matias: For deliverable 1, I helped fetch the Air Quality index data using the bash script written by our team. I also collected the data from the Proximity to Roads paper and wrote code to analyze it by writing code to find the median, average, and standard deviation. I also wrote code in order to visualize the data in the form of graphs. Lastly, I wrote part of the report and now I am currently working on fetching all the data for every single day from the Air Quality Index in order to compare how the data changes yearly.

b. Yuchen: For deliverable 1, I cleaned and processed data related to social vulnerability and created graphs for analysis. I also wrote a bash script to fetch daily air quality data in 2021 and aggregate data related neighborhoods (the ones that appear in social vulnerability dataset) that we are interested in to a single csv file for easier access. Then, I combined the air quality data with the social vulnerability data to draw connections between the two.
c. Eric: For Deliverable 1, I worked on fetching and processing the Air Quality Index data for our analysis. Using the aforementioned bash script, I collected yearly data from the AirNow website, initially for every day of 2021 but eventually for every month over the entire range of years from 2014 to present day. I then created a Jupyter notebook to convert all the data into a consistent format (as the .csv files changed formats partway through the data), allowing it to be processed into a graph for further analysis.
d. Peter: For deliverable 1, I helped fetch the Air Quality Index data from AirNow. Then, I worked on fetching the MBTA transit data from ArcGIS. However, this has proven to be unsuccessful due to lack of permission to acquire the data. Next, I worked on getting some census data that we might be interested in. I am currently working on a way to get the MBTA transit data and doing further analysis on the census data.
e. Steve: For Deliverable 1, I fetched the Air Quality Index data using a bash script. I conducted exploratory data analysis, particularly focusing on the relationship between the Air Quality Group and the social vulnerability score. Given the limitations of our current data, we were unable to fully address the questions, but I also contributed to formulating the hypothesis for the questions.