P2 Final Report

Team Members

Eric Freitag (eef49)
Dana Gong (dg588)
Jiadi Huang (jh2649)
Mei-Jen Lee (ml2298)

Data Description

Our team used multiple open source datasets:

Dataset 1: 2015 Tree Census from NYC Open Data

- This dataset was very large, containing 684k rows and 45 columns and originally included variables from tree health, type, who sighted it, etc.
- We cleaned it down to just over 87,000 rows and 8 columns to make it fit our submission size and be more manageable.
    - To clean the data, we subsetted to only trees that were alive, their health was considered good, and their diameter to be greater than 20 inches (as per the generic definition of a large tree)
    - Additionally, we dropped most of the descriptive columns to only give us tree_id, block_id, nta (neighborhood tabulation area), borough, and latitude and longitude.
    - To calculate the tree number per community district, we accumulated each row which shares the same community district code and then averaged it.

Dataset 2: NYC Air Quality from NYC Open Data

- This dataset contains 16,122 rows and 12 columns on different measures of air quality. For our purposes, we decided to use the measure for particulate matter (PM2.5). We also only include rows that included air quality measurements by community district to fit our other datasets.
- To allow this dataset to be more easily used with our geojson file, we added the air quality data as a feature in our geojson file using an online tool.

Dataset 3: NYC poverty data from NYC open data

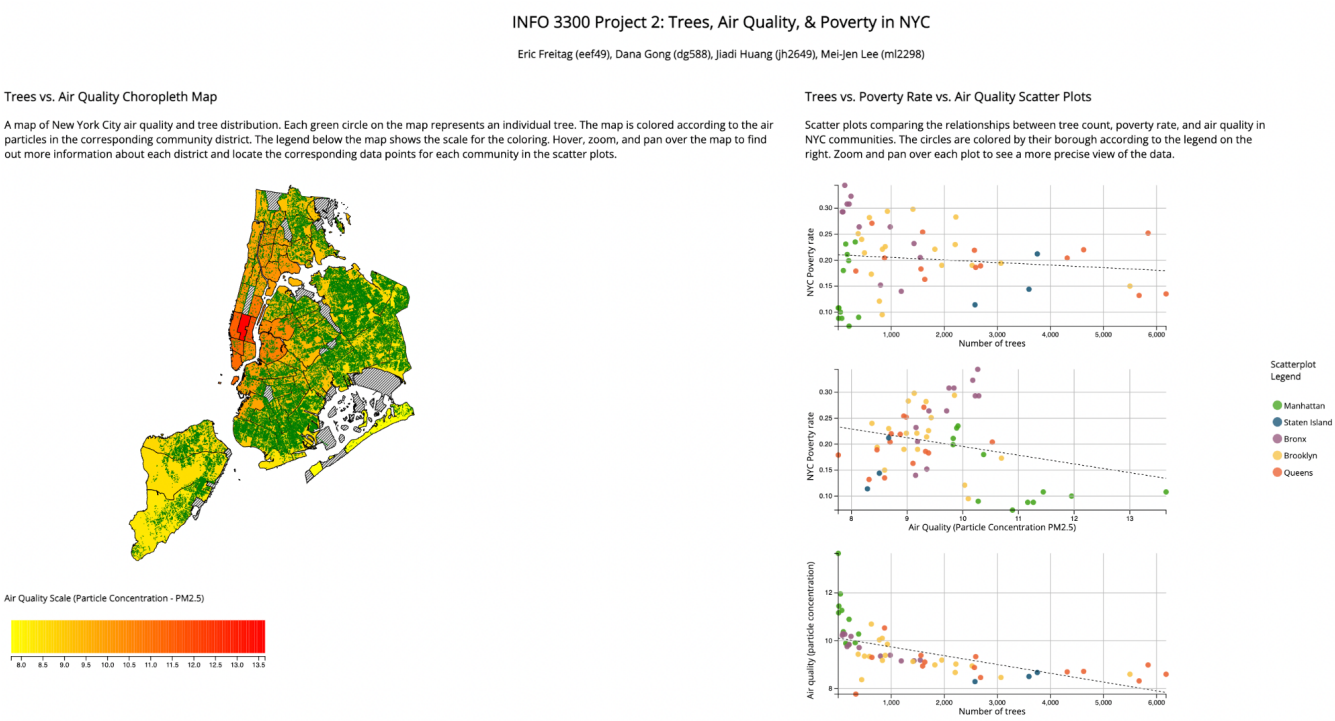
- This dataset contains 385 rows and 52 columns on the poverty rate in each neighborhood area. We averaged the poverty rate in each community district (community districts are areas governed by separate community boards created in 1975) and found the mapping between the community district in this dataset and the other datasets.
    - Each data point from this dataset is a neighborhood area instead of a community district, so we accumulated each row which shared the same community district. After processing the data, we found the average poverty rate based on the community district.
    - The community district in each dataset has different forms. In the poverty dataset, the community district is formed by a borough code followed by a number. We

mapped this to a three digit number. For example, 102 means Manhattan community district 02.

- Some of the rows contain more than 1 community district (for example, MN Community district 1 & 2). In these cases, we split the string and processed them separately.

Additional Files:
File 1: Boundaries of NYC Community Districts from NYC Open Data

Visual Design Rationale



Visually, we wanted to create somewhat of a dashboard with different views of data.
For our choropleth map, we chose to use a sequential color scale with red filling the areas with lowest air quality to yellow filling the areas with the best air quality. This seemed the most intuitive to us as redder would indicate more hazardous, and we indicated this in our air color scale. We got confirmation about using a divergent color scale in office hours. In this plot, our marks were circles to indicate trees, and community district shapes to indicate different areas. In areas where districts were not residential and did not include air quality data, we used a diagonal hatch pattern to distinguish them from our map (these areas include major community districts like Central Park, JFK, and LaGuardia Airport).

For our scatterplots, our channels are vertical and horizontal x and y alignment, as well as different colors to indicate the different boroughs. Our marks are the circles indicating the different community districts. We used linear scale because it shows the data distribution clearly.

We wanted to not only relate back to the map through interactive elements which will be discussed in the next section, but visually group different boroughs together. In our scatter plots, we plotted a regression line to visualize the trend displayed from our data more clearly.

Interactive Design Rationale

We want to allow the user to explore individual communities on the choropleth map. Thus we included the zoom and pan interaction for the map. Users can zoom and pan to explore more and see the distribution of trees more clearly on the map. We also decided to include a hover interaction for the choropleth map. By hovering on a community, the user can see the exact information for which borough the community is located and also the air particle concentrations for the community. The corresponding points for the communities on the three scatterplots are also shown by thickening the stroke width of the data point when hovering over the community, so users can see the corresponding air quality, the poverty rate, and the number of trees in the data distribution at the same time clearly. We are trying to show the correlation between the three variables: NYC poverty rate, number of trees and air quality (Particles Concentrations PM2.5). On the scatterplot, the user can also zoom and pan to see their x and y coordinates more clearly on the graph.

The Story

The origin of our idea was a body of research indicating lower concentrations of trees and green areas in lower income areas. American Forests, Washington, D.C.-based conservation nonprofit, released a nationwide analysis last month showing that low-income neighborhoods and communities of color have significantly less tree canopy and lack critical infrastructure that, with a number of many community health indicators, is a major factor in air pollution and quality. Although our visualizations are not a great indicator of the current research, it helps support the growing body of knowledge surrounding the environment and urbanism by visualizing air quality, tree density and poverty in the largest city in America. There are some major limitations to our insights that include: reducing our tree census data to just about ⅛ of the original dataset, and the fact that major areas that were not residential were not covered such as Central Park and airports.

Our choropleth map visualization shows the disparity in the distribution of trees and its corresponding air quality (PM2.5/particulate matter) across the community districts in NYC. It is shown that areas in Manhattan have a far lower concentration of trees and worse air quality, despite the presence of Central Park (which is left out of the data). Further from Manhattan, the air quality becomes better, especially as you get to the outskirts of the city. Ideally, we would like to compare this to asthma or other air quality related health issues, as research shows that poorer air quality can worsen asthma symptoms.

In addition to our choropleth map, we created scatter plots to isolate the relationships between the three variables that we were examining: tree count, poverty rate, and air quality. We also showed different colors for the five boroughs in NYC. The most significant result that we saw from these plots was that there is a clear correlation between the number of trees in a community and the air quality in that community. This makes a lot of sense because research shows that the presence of trees in a community actively helps reduce the amount of contamination in the air. Furthermore, areas that have less trees are often more densely populated, like the center of Manhattan, so there is likely more pollution being produced in these areas of the city. The other two scatter plots that we created illustrated results that were quite surprising. Rather than lower income areas having less trees and lower air quality like our background research suggested, the New York City data that we analyzed demonstrated the opposite results. As the poverty rate increases, the number of trees in a community grows slightly and the air quality grows significantly. These results were most likely due to the fact that lower income communities are often less developed and have less commercial infrastructure, which leads to less pollution being generated in these areas. Furthermore, New York City prioritized the planting of trees under Mayor Bloomberg's MillionTreesNYC initiative, so NYC may be an exception to the general trend of lower income neighborhoods having less trees.

Team Contributions

Dana Gong (dg588):

- Cleaned tree census and air quality datasets
- Created choropleth map of air quality in NYC and plotted tree points
- Calculated and plotted regression lines for scatter plots
- Edited visual layout and color scales

Mei-Jen Lee (ml2298):

- Draw two scatter plots and their paired interactivity, poverty vs trees & poverty vs air quality.
- Show the corresponding data points in the scatter plot when hovering over an area.
- Map Community district data among air, trees, and poverty dataset to pair the interactivity

Jiadi Huang (jh2649)

- Worked on Pan and Zoom for all the graphs
- Worked on interactivity for hovering on community district and retrieving the corresponding data
- Created a legend for the color scale

Eric Freitag (eef49)

- Created the scatter plot comparing air quality and number of trees.
- Designed the pan and zoom interactions for the scatter plots
- Worked on the mouseover features for the choropleth map.

- Commented the code base and organized code from multiple contributors to make it readable for outside users.

(Rough) Breakdown of Work:

Data Cleaning:

- Overall, since we were working with very, very large and messy datasets, a lot of time was spent on cleaning and making sure our data was usable. Additionally, since we used multiple datasets, we had to map them out through cleaning and in our js file to make the columns match each other.
- 9 hours total for all 3 datasets

Choropleth Map

- 6 hours

Scatterplots

- 6 hours

Scales/Legend/Regression Line

- 4 hours

Interactivity

- 6 hours

Debugging

- This, along with data cleaning, took the most time as we were formatting things differently and also parts of code built off one another and we had to trace back to find why things were not showing as we expected
- 8 hours

Visual Design and Editing Layout

- To reconcile differences in layout/color/design patterns, we took some time at the end of our project to make sure everything looked cohesive.
- 4 hours