

Shade NYC

Team Members:

Carlos Anguiano - cjanguia@usc.edu, Student ID: 9385358906

Val Katritch - vjkatrit@usc.edu, Student ID: 7063534092

Project Overview

Shade cover from trees has colloquially been said to cool down urban environments. As climate change continues to impact urban environments, keeping cities cool is incredibly important for health and economic purposes (Ettinger, 2024). This is a social justice issue, as people in lower-income communities often have less access to green spaces, meaning that they will experience the “heat island effect” to a greater extent than those in higher-income regions. This project aims to analyze data from New York City Open Data to test if this theory applies to New York. In addition, we will investigate environmental injustices by analyzing socioeconomic status by zip code to determine if wealth disparity affects the tree cover.

Hypotheses:

1. Higher property values will be associated with lower HVI.
2. Higher property values will be associated with higher tree count.

Data

We acquired the data through the City of New York Open Data site, which provides many sources of data regarding New York City. After exploring the website, we discovered data on the Heat Vulnerability Index (HVI), Tree Census, and property values.

After trying some preliminary visualizations with the NYC Open Data site, we retrieved the raw data via APIs using the links below. The HVI data was retrieved from the Open Data site using the API in the CSV format, due to its nature of being a smaller data set. The property value distribution data and tree census data were much larger data sets, with 9.85 million rows counting each property and 684K samples for each tree, so we had to acquire this through a JSON format of API instead of CSV.

APIs:

1. HVI Data
 - a. <https://data.cityofnewyork.us/resource/4mhf-duep.csv>
 - b. Retrieved via API (CSV format)
 - c. Number of Data Samples: 184
2. Tree Census
 - a. <https://data.cityofnewyork.us/resource/uvpi-gqnh.json>
 - b. Retrieved via API (JSON format)
 - c. Number of Data Samples: 684,000
3. Property Value Distribution
 - a. <https://data.cityofnewyork.us/resource/yjxr-fw8i.json>
 - b. Retrieved via API (JSON format)
 - c. Number of Data Samples: Sampled 500,00 data samples from the full 9.85 million dataset

Data Cleaning, Analysis & Visualization

After we found the data on NYC Open Data and located all APIs, we used the “get” function to acquire the APIs in Jupyter notebook (we initially coded in Jupyter notebook, as this is what we learned throughout the semester, and then transitioned this to a .py file later on in the project). For each data set we cleaned up the data separately:

Data Cleanup:

For the Heat Data, we collected the full data, which only included relevant fields, including zip code (zcta20) and heat index (hvi). We read the CSV file using StringIO to convert the data into text, and then created the CSV file “heat_data_full.csv”. We then pulled that data into a pandas dataframe, sorted the data by zip code, cleaned up column titles, and wrote it into a new file with heat index by zipcode (heat_by_zip).

For the Tree Data, we pulled only the necessary columns for analysis from a JSON system, as the full data would have been unnecessary and overwhelmed the storage allotted and system memory. The columns pulled were “tree_id,” “zipcode,” and “boroname.” We pulled these fields for the entire set of 684,000 data points in the data base to get as full a representation of the trees in NYC as data allowed. The data was written into the file in 100,000 chunks to keep a consistent process resistant to failures. This was written to a CSV file “tree_data_full.csv”. To clean this data, we pulled the information into a data frame, counted unique tree instances per zip code and read this along with its respective borough name and zipcode into a cleaned CSV file titled “tree_count_by_zip.csv”.

For the socioeconomic data, the process was a little different due to the size of the data. The Property Value data had information representing 9.85 million properties, which ended up being a massive strain on run time and memory storage. So we took the API as a JSON file, and wrote a program to sample the entire database to output a representative 500,000 row sample using a reservoir and a randomizer. The data would fill the reservoir with the first 500,000 data points, 100,000 rows at a time. Each subsequent 100,000 rows of data had a chance of replacing the data in the filled the 500,000 reservoir. We used a try-except tool to safeguard against data read or processing errors, something prone to happen with a dataset of this size. If a certain 100,000 row chunk hit 5 errors, it would simply skip the chunk and move on. This loop processed the data until it had gone through all 9.85 million rows, giving us the random sample we needed. After processing every 400,000-row chunk, it would write to the output file, “property_value_sampled.csv,” to provide a backup of the reservoir if the code threw errors later down the line. Once it had gone through the entire dataset, it wrote the final reservoir to the CSV. To process the data sample, we opened it in a dataframe and computed the mean property values for each zip code to summarize the data and get a proxy for socioeconomic data in NYC. We then saved this to a new processed CSV file with average property values by zip code, thus completing the data cleanup phase of this project.

To combine the data, we opened a dataframe for each of our 3 sets of data by defining its file path. In each dataframe, we ensured the zip codes were strings and standardized them to ensure they were only 5 digits, as some of the zip codes were reported differently. We then merged the dataframes by zip code, wrote to “merged_data_by_zip.csv”, allowing us to compare the three factors and analyze the data.

Data Analysis:

To analyze the data, we pulled the data from the merged CSV and used the describe function to get a statistical summary, and calculated a Pearson correlation test using pandas. The results printed as such:

	Heat Index	Tree Count	Average Property Value
count	178.000000	178.000000	1.780000e+02
mean	3.050562	3827.696629	1.748939e+06
std	1.411304	2895.891272	4.118348e+06
min	1.000000	30.000000	2.269399e+05
25%	2.000000	1960.250000	6.072860e+05
50%	3.000000	3315.500000	8.140393e+05
75%	4.000000	4886.750000	1.449682e+06
max	5.000000	22186.000000	3.952853e+07

Correlations:

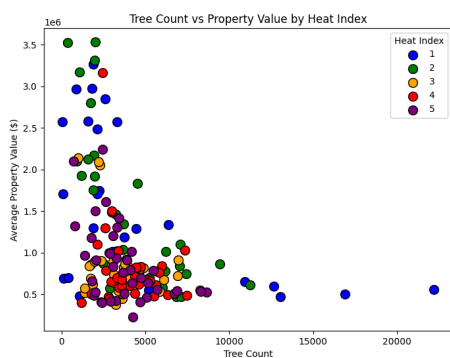
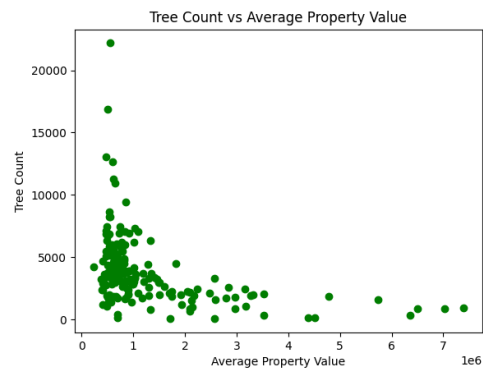
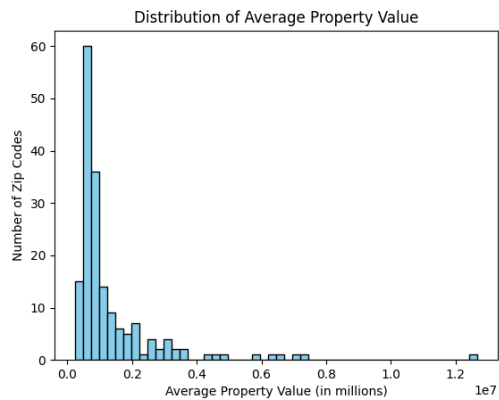
	Heat Index	Tree Count	Average Property Value
Heat Index	1.000000	-0.020000	-0.226298
Tree Count	-0.020000	1.000000	-0.265619
Average Property Value	-0.226298	-0.265619	1.000000

Visualizations:

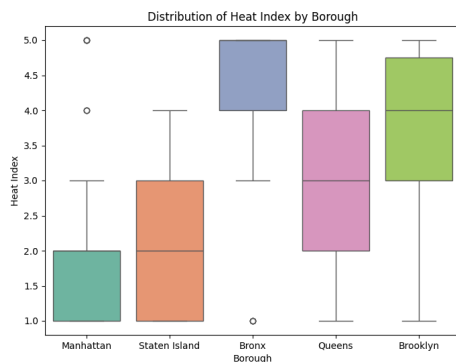
After calculating the statistical analysis, we created 5 visualizations using matplotlib. The data, as is, did not provide very clear visualizations. To show visually clear trends in our figures, we had to limit the data to various levels by figure. Our histogram filtered the data to keep only points within 3 standard deviations to simplify the data for visualization purposes. The scatter plot had data within 2 standard deviations, and the multivariable scatterplot had data within 0.5 standard deviations. Though this may not be the most statistically sound, it provided the most clear visuals, which we ultimately believed to be the priority.

First, we plotted a histogram with the filtered data of the Average Property Values in New York to show the wide range of property values, and how most properties are below \$0.5 million (Figure 1a). We then created a scatter plot of the tree count vs. average property values, which was first filtered to create a clearer scatter plot (Figure 1b). Then we created a multivariable scatter plot to correlate tree count by average property value (in millions), color-coded by heat index (Figure 1c). Following this, we created a heatmap of average values by borough using Seaborn, showing the heat index, tree count, and property values per borough (Figure 1d). Finally, we made a box plot demonstrating the distribution of heat index by borough (Figure 1e).

Figure 1 (a-e).



Borough	Heat Index	Tree Count	Average Property Value
Bronx	4	3408	1097442
Brooklyn	4	4663	858742
Manhattan	2	1495	3803756
Queens	3	4155	1361208
Staten Island	2	8776	500749



Discussion:

As shown from the statistical findings and the visualizations, the phenomenon initially described and the hypotheses investigated are not strongly demonstrated in this data, and partly even disproven. There were weak connections relating heat index and average property value, and tree count with property value, with correlation values of -0.226 and -0.266 , respectively. The correlation between heat index and average property value indicates as property value increases, heat index decreases slightly, which does support our first hypothesis. The tree count and property value correlation indicates that higher property values are associated with slightly fewer trees, which goes against our second hypothesis.

The Heat Vulnerability Index describes the risk of death from heat, which is calculated by looking at a neighborhood's surface temperature, green space, home air conditioning, and income (NYC.gov). This is a complicated calculation, making it difficult to pinpoint a single cause of increased vulnerability. Figure

It does show that the heat vulnerability indexes of each borough are variable, with Manhattan being the lowest (least vulnerable), and the Bronx and Brooklyn being the highest (most vulnerable). This makes sense, considering their economic differences and access to air conditioning. However, our findings seem to indicate that green space (tree coverage) is not as big a factor as it is described in the heat island effect. From our statistical analysis, the heat index and tree count show an almost zero correlation of -0.020. This could indicate that this phenomenon does not apply to New York City in the same way it does to other cities like Los Angeles (Connolly, 2023). For Los Angeles, this is most likely because of the way the city was built. Los Angeles is a much younger city than New York, and was built with more access to technology because of this. New York and Los Angeles are also in different climates, and the native climate of Los Angeles has many fewer trees, and more shrubs and flowers. This means that having access to trees in Los Angeles means access to resources to grow these trees that do not naturally occur, which is expensive to maintain. Meanwhile, New York naturally had forests, making it much less of an accessibility issue.

Changes from Original Proposal

Almost everything stayed the same from our proposal, but we added some specific hypotheses to narrow down our research. We also added some more visualizations than we had initially stated based on what was necessary. We were unable to include a map, due to restrictions on time and skill level.

Future Work

If we had more time with this project and more resources, we would have liked to create a map showing the distribution of HVI Index, property values, and tree coverage to see if there were any specific regions where this phenomenon could be focused, rather than doing a general statistical summary. This would give us more information, and would create a better visualization for the reader. However, we did not learn how to create maps in this course, so this would have been above our skill level to complete.

Although our results were statistically insignificant, it does point out that tree coverage does not directly correlate to HVI or property value. It would be useful for future studies to be completed to discover where, if anywhere, this phenomenon occurs, and to understand why exactly tree coverage can be a social justice issue.

Citations

Connolly, Rachel, et al. "The Association of Green Space, Tree Canopy and Parks with Life Expectancy in Neighborhoods of Los Angeles." *Environment International*, vol. 173, no. 107785, 13 Mar. 2023, p. 107785, www.sciencedirect.com/science/article/pii/S0160412023000582?via%3Dihub, <https://doi.org/10.1016/j.envint.2023.107785>.

Ettinger, A.K., Bratman, G.N., Carey, M. et al. Street trees provide an opportunity to mitigate urban heat and reduce risk of high heat exposure. *Sci Rep* 14, 3266 (2024).
<https://doi.org/10.1038/s41598-024-51921-y>

"Interactive Heat Vulnerability Index." Environment & Health Data Portal, a816-dohbsp.nyc.gov/IndicatorPublic/data-features/hvi/.