PROJECT PART 4:

TOPIC PROPOSAL

By

GROUP 5

PRAMOD KRISHNACHARI

JIN-HYUK SON

SANJANA MURALIDHAR

VAIJAYANTI SHRIKANT DESHMUKH

SNEHITHA TADAPANENI

Visualization of Complex Data: DS 6401

Under the guidance of,

Prof. Anya Mendenhall

(TA: Erica Zhao)

Urban Tree Distribution in NYC and their Impact on Air Pollution in 2015

1. Introduction and Background

Urban environments like New York City face growing challenges around sustainability, public health, and environmental justice [1]. Among the many factors contributing to urban livability, urban trees play a quiet but powerful role, improving air quality, reducing heat, supporting biodiversity, and enhancing psychological well-being. At the same time, air pollution remains a critical concern, especially fine particulate matter (PM2.5), which is linked to respiratory and cardiovascular diseases.

Inspired by NYC Open Data initiatives and previous projects like the NYC Street Tree Census (2015) [6] and NYC Community Air Survey (NYCCAS) [5], this project explores the intersection of **urban forestry and air quality**. Several studies and local planning discussions emphasize the importance of green infrastructure in combating urban air pollution, but few visualizations bring these datasets together at a granular, community district level.

NYC is home to over 5 million trees, and monitoring their health, diversity, and distribution offers insights into how well neighbourhoods are equipped to handle environmental stressors[2]. Air quality, on the other hand, varies widely across the city due to traffic, industrial zones, and socio-economic disparities. Exploring the spatial and relational patterns between these two urban systems may help inform better environmental policy and urban planning decisions.

The main goal of this project is not to claim causation, but rather to identify patterns that can reveal actionable insights or inform future hypotheses. The project focuses on uncovering:

- Which community districts have the highest and lowest tree coverage and biodiversity?
- What are the most frequent tree stress factors across the city?
- Are there visible associations between tree health/stewardship and air quality levels?
- Do areas with greater tree density or diversity experience better air quality?

To answer these questions, the project is structured into three phases:

1. **Tree Data Exploration** – Analyze tree density, species diversity, health, and stress indicators.
2. **Air Quality Analysis** – Examine PM2.5 levels and identify the cleanest vs. most polluted districts.
3. **Combined Insights** – Overlay both datasets to observe potential associations and generate hypotheses.

Visualizations such as maps, bar charts, heatmaps, and scatter plots will be created using Tableau, and statistical summaries like correlation coefficients are made where ever required to provide additional context.

Project Benefits

- Urban planners and environmental agencies looking to optimize green space investment.
- Public health advocates interested in the co-benefits of trees in pollution-prone neighbourhoods.
- Educators and students exploring data storytelling and spatial pattern recognition.
- Everyday New Yorkers who want to understand how their environment supports or hinders, their health.

Ultimately, this project serves as a starting point for deeper inquiry into how the urban natural environment intersects with public well-being.

2. Datasets

We are using two datasets for our analysis and study of trees and air quality in urban areas.

2.1. Dataset I: Air Quality Dataset [5]

Q. Who collected the data? Who funded the project that the data came from? Other important information?

The New York City Department of Health and Mental Hygiene (DOHMH) is the team behind gathering this dataset, as part of their ongoing efforts to keep an eye on air quality. This project is probably backed by the NYC government along with various public health and environmental initiatives aimed at tracking and enhancing the quality of air in urban areas.

Q. Why was the data created and for what purpose? Who collected the data? Who funded the project that the data came from?

The purpose of this data is straightforward: it was put together to monitor air pollution levels in various neighborhoods throughout NYC. Air pollution is a significant issue for both the environment and public health, as being exposed to these pollutants can lead to respiratory illnesses, heart problems, cancers, and even early deaths. With this dataset, policymakers, researchers, and everyday folks can get a clearer picture of pollution trends, pinpoint areas that are at higher risk, and take action to enhance air quality.

Q. What is the timeline or lineage of the data?

The dataset is updated regularly, with the most recent update on March 24, 2025. It provides historical data over multiple years, enabling time-series analysis of air quality trends and comparisons across different time periods.

Q. Define and describe the variables included in the dataset.

The dataset is packed with various attributes that detail air quality measurements across different locations and timeframes. Here are some of the key variables you'll find: -

Unique ID: This is the identifier for each record.

Indicator ID: It represents specific air quality indicators.

Name: This refers to the pollutant or air quality measure (ozone, Nitrogen dioxide, Deaths due to PM2.5, etc)

Measure: This is the actual concentration of the measured pollutant (Mean, number per km2, million miles, etc)

Measure Info: Here, you'll find extra details about the measurement (per square mile, per 100,000 adults, etc)

Geo Type Name: This indicates the geographical level, like borough or community district.

Geo Join ID: A unique identifier for geographical locations.

Geo Place Name: The name of the location where the air quality was assessed. –

Time Period: This shows the time range for the measurement.

Start_Date: The date when the measurement was recorded.

Data Value: This is the numerical value representing the air quality measurement.

Q. How large is the dataset (cases, how many and what are the variables)?

Although the specific number of records can change, the current updated version has 18025 records. This dataset includes a variety of air quality measurements taken from various locations and times, making it an invaluable tool for studying pollution trends in NYC.

Q. What locations are included in the dataset?

The dataset features air quality measurements from all over New York City, capturing data from various boroughs and neighborhoods. This makes it possible to analyze trends at different geographical levels, whether you're looking at the city as a whole, specific boroughs, or even individual neighborhoods.

Q. Describe how the datasets will help you achieve your goals/questions that posed in your introduction, including any limitations.

This dataset offers essential insights into pollution exposure in NYC, enabling researchers to spot areas with elevated pollution levels and contrast them with healthier neighborhoods. We can examine trends over time to see if air quality is getting better or worse. We can also evaluate how environmental policies, urban development, and climate changes affect air quality. Furthermore, we can explore the differences in air pollution exposure among various neighborhoods and demographic groups.

Limitations: Air pollution is shaped by a variety of factors, such as weather, traffic patterns, and industrial operations, and not all of these elements might be reflected in this dataset. This dataset offers a summary of pollution levels instead of real-time data on individual exposure. Additionally, some areas might have gaps or inconsistencies in their data, which could impact the overall analysis.

We found out that New York City is split into various geographic types, and thought Community Districts (CDs) are the perfect fit for our analysis. For the Air Quality dataset, we’ll stick with the Geo_type CD to keep things consistent. We also came across a spatial file for the CDs, which will help us create maps in Tableau. To connect this with our Tree Census data, we did a spatial join using latitude and longitude, making use of the MAKEPOINT function. In the new file, there's a column named Boro CD that corresponds to Geo_join_id in the Air Quality data. Now, we can easily merge both datasets, paving the way for more exploration and visualization.

2.2. Dataset II: NYC Tree Distribution Dataset [6]

Q. Who collected the data? Who funded the project? Other important information?

The New York City Department of Parks and Recreation (NYC Parks) led the collection of this dataset through the 2015 Street Tree Census, also known as TreesCount! This was the third decadal tree census, following previous surveys in 1995 and 2005. The project was a massive citizen science effort, with over 2,300 volunteers working alongside professional arborists to map and document street trees across NYC.

The funding for this project likely came from a mix of NYC government funds, environmental organizations, and sustainability initiatives focused on urban forestry and green infrastructure.

Q. Why was the data created and for what purpose?

The primary goal of this dataset was to create a comprehensive inventory of street trees in NYC, tracking details like species, size, health, and geographic distribution. The data serves multiple purposes, including:

- Urban Forestry Management
- Sustainability & Climate Resilience
- Research & Policy Making
- Public Engagement

Q. What is the timeline or lineage of the data?

The NYC Street Tree Census occurs every 10 years, providing historical data on urban forestry trends.

- Past censuses: 1995, 2005
- Most recent census: 2015 (data collected between May 2015 – October 2016)
- Next expected census: 2025

This dataset captures a snapshot of NYC's street trees as of 2015-2016, offering a basis for long-term studies on how urban greenery evolves over time.

Key Publications Associated with the Dataset:

- Study on Tree Species and Air Quality: Research from the Lamont-Doherty Earth Observatory highlights that some tree species release volatile organic compounds (VOCs), which can contribute to ozone formation and impact air quality. This suggests that tree selection should be done carefully to maximize environmental benefits.
    - Source: Planting Some Tree Species May Worsen, Not Improve, NYC Air
- NYC Parks' Forestry Management System (ForMS): This system uses the census data to track tree conditions, support maintenance planning, and monitor urban forestry trends.
    - Source: NYC Open Data

Q. How large is the dataset?

- Total records (trees surveyed): 683,788
- Total variables: 45 (including species, health, location, and maintenance needs)
- Categorical features: 26 (such as tree species, health condition, and borough)

Key Variables in the Dataset:

- tree_id – Unique ID for each tree
- spc_common – Common name of the tree species
- tree_dbh – Diameter at breast height (in inches)
- health – Condition of the tree (Good, Fair, Poor)
- status – Tree status (Alive, Dead, Stump)
- problems – Any recorded issues (e.g., broken branches, sidewalk damage)
- latitude, longitude – Exact geographic coordinates
- borough – Borough name (Manhattan, Brooklyn, Queens, Bronx, Staten Island)
- nta_name – Neighborhood name (e.g., Whitestone, Yorkville)

Q. What locations are included in the dataset? Can it be used for analysis at different geographic levels?

This dataset focuses solely on New York City, covering all five boroughs: Manhattan, Brooklyn, Queens, Bronx, and Staten Island.

Since the dataset includes detailed geographic identifiers like neighborhoods (NTA), community districts (CD), city council districts, census tracts, and zip codes, it allows for various levels of analysis, such as:

- Street-level analysis: Examining tree health and species distribution in specific areas.
- Neighborhood comparisons: Identifying disparities in tree coverage between different communities.
- Borough-wide studies: Understanding how tree distribution varies across NYC.

However, state, national, or global-level comparisons are not possible unless combined with other tree datasets from different regions.

Q. How does this dataset help in answering our research questions? What are its limitations?

This dataset is a valuable resource for studying the relationship between urban tree distribution and environmental factors like air quality, public health, and urban heat islands. By combining it with the NYC Air Quality Dataset, we can:

- Identify areas with lower tree coverage and higher air pollution, which may require targeted environmental interventions.
- Analyze tree health trends and their correlation with various stress factors, stewards pollution levels.
- Explore how tree species distribution affects air quality, given that some trees emit volatile organic compounds (VOCs).

- Assess whether environmental policies and urban planning initiatives have led to improved greenery and air quality over time.

Limitations: The dataset only includes street trees (not those in parks or private properties), so it doesn't provide a complete picture of NYC's total tree population. Data collection occurred only in 2015-2016, meaning it does not reflect recent changes in NYC's urban forestry. Tree health assessments were subjective, based on visual inspections, which could introduce some inconsistencies. The dataset doesn't track real-time changes in tree conditions, growth, or air quality impacts.

2.3. Dataset III: NYC Community Districts [7]

Q. Who collected the data? Who funded the project that the data came from? Other important information?

This dataset is provided by the New York City Department of City Planning (NYC DCP) through their Community Profiles platform. It is part of the city's broader effort to support data-informed urban planning, policy-making, and public transparency. The data is likely funded by the NYC municipal government as part of ongoing urban development and community assessment initiatives.

Q. Why was the data created and for what purpose?

The purpose of the Community Profiles data is to summarize socio-economic, health, and environmental indicators at the Community District (CD) level. It helps planners, researchers, policymakers, and residents better understand conditions in different neighborhoods.

Q. What is the timeline or lineage of the data?

The dataset is regularly updated as part of the NYC Planning portal. While some variables are updated annually based on survey data. The latest data reflects the most recent updates to the NYC community profile dashboards, which incorporate information from 2020–2024 depending on the variable.

Q. Define and describe the variables included in the dataset:

- cd_name: The official name of the community district (e.g., *Brooklyn Community District 1*).
- cd_number: A unique numeric code representing each CD.
- cd_score: An aggregated score reflecting the overall performance.
- cd_area_score: A normalized or scaled version of the CD score, useful for comparing across varying district sizes.
- cd_area_acres: The total area of the district in acres, providing insight into geographic size and density when combined with other datasets.

Q. How large is the dataset?

The dataset contains 59 records, corresponding to the 59 Community Districts in New York City. It includes 5 core variables described above, but is often linked with broader datasets in the NYC Open Data ecosystem.

Q. What locations are included in the dataset?

All five boroughs of NYC are represented: Manhattan, Brooklyn, Queens, The Bronx, and Staten Island, broken down into their respective Community Districts (CDs).

Q. How does this dataset help achieve your goals/questions, and what are its limitations?

This dataset plays a supporting role in our analysis by:

- Providing a spatial reference and area size, which we used to normalize tree density metrics (trees per square mile).
- Potentially contributing district-level context for explaining variations in air pollution and tree health outcomes.

Limitations include its aggregated nature, it does not offer granular data on environmental features or tree attributes directly. Also, cd_score may represent a blend of indicators that need further breakdown for precise interpretation.

Merging Tree Distribution Data with Air Quality Data

To combine the Tree Census with the Air Quality Dataset, we used a spatial join based on latitude and longitude, utilizing the MAKEPOINT function. This allowed us to link tree locations to Community Districts (CDs), ensuring consistency across datasets. The Boro CD column in the tree dataset now corresponds with the Geo Join ID in the Air Quality dataset, enabling deeper analysis and visualization in Tableau.

By merging these datasets, we can map out relationships between tree density and air pollution levels, uncovering insights that can inform urban forestry policies and sustainability efforts in NYC.

3.Data Story

Phase 1: Understand tree distribution, diversity, health, and environmental stress indicators across NYC.

Key Questions:

- Which districts have the highest/lowest tree density and species diversity?
- What is the distribution of tree health across districts?
- What are the most common tree stress factors ?
- How does tree health relate to stewardship levels?

Chart 1(Map): Tree Density Across Community Districts (Trees per Square Mile)

Variables:

- Trees per sqmi (Continuous): Trees per square mile (Calculated by dividing cd_area_acre)
- Geo Place Name (Categorical): NYC Neighbourhood Name

- Longitude & Latitude (Geospatial-Continuous): Geographical Representation

Narrative: This map visualizes the distribution of tree density across NYC community districts. Darker shades represent higher tree densities, while lighter shades show sparser tree coverage. From the visualization, we can observe that districts in outer boroughs such as Queens and parts of Brooklyn have relatively higher tree densities, possibly due to residential zoning and wider sidewalks. Conversely, denser commercial areas like lower Manhattan exhibit lower tree counts, reflecting limited green space in these regions. This answers our first objective: identifying areas with the highest and lowest tree coverage.

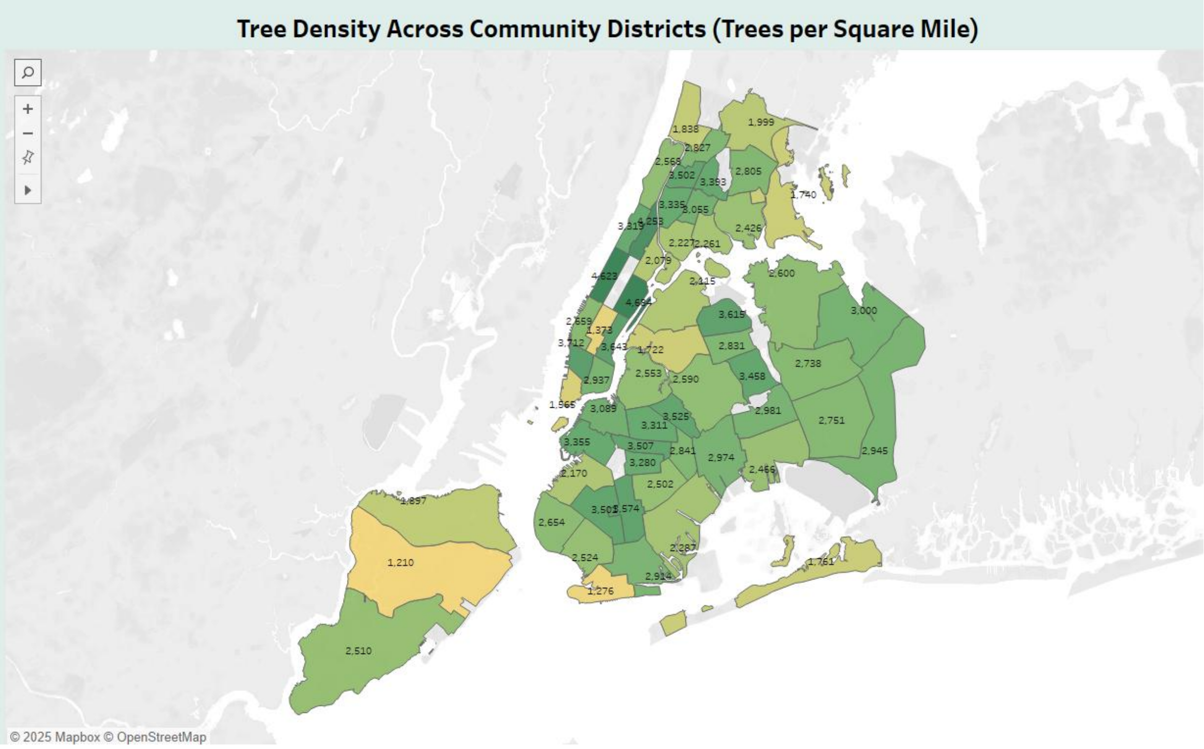


Figure 1: Tree Density Across Community Districts(Trees per Square Mile)

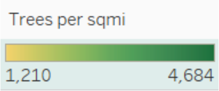


Chart 2(Map): Species Diversity Across Community Districts

Variables:

- Spc Common (Categorical): Specie Common Name.
- Geo Place Name (Categorical): NYC Neighbourhood Name.
- Longitude & Latitude (Geospatial-Continuous): Geographical representation

Narrative: This map displays species diversity by community district, calculated using a biodiversity index. A high index suggests a wider variety of tree species. The map helps us understand ecological resilience districts with greater diversity are better protected from species-specific diseases or pests. The chart reveals notable diversity in areas like Staten

Island and portions of the Bronx, while some central districts rely heavily on a few dominant species. This aligns with our goal to explore species diversity patterns.

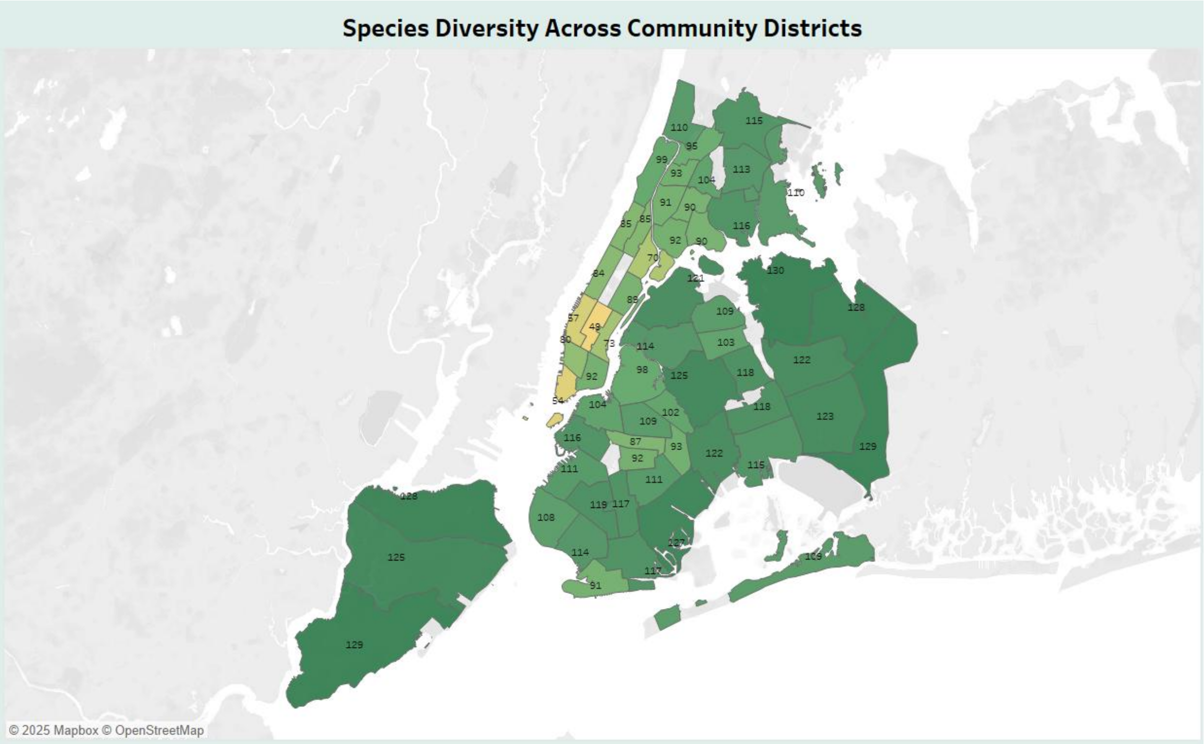


Figure 2: Species Diversity Across Community Districts

Number of Species

48 130

Chart 3(Map): Tree Density and Species Diversity Across NYC

Variables:

- Spc Common (Categorical): Specie Common Name.
- Geo Place Name (Categorical): NYC Neighbourhood Name.
- Longitude & Latitude (Geospatial-Continuous): Geographical representation.

Narrative: This map visualizes tree density and species diversity across New York City's community districts, using a colour gradient to represent tree density (trees per square mile) and red circles to indicate species diversity, the larger the circle, the greater the number of species. From the visualization, it's clear that districts in Queens and parts of Brooklyn exhibit both high tree density and greater biodiversity, suggesting robust urban forestry efforts and ecological resilience. In contrast, central and southern areas of Manhattan show lower tree density and species diversity, likely due to space limitations and commercial development. Staten Island, while showing moderate tree density, features relatively high species diversity in several areas, possibly due to more residential zoning and natural spaces. This chart supports our objective of identifying community districts with the highest and lowest tree coverage and biodiversity, offering insight into the distribution and ecological quality of NYC's urban forest.

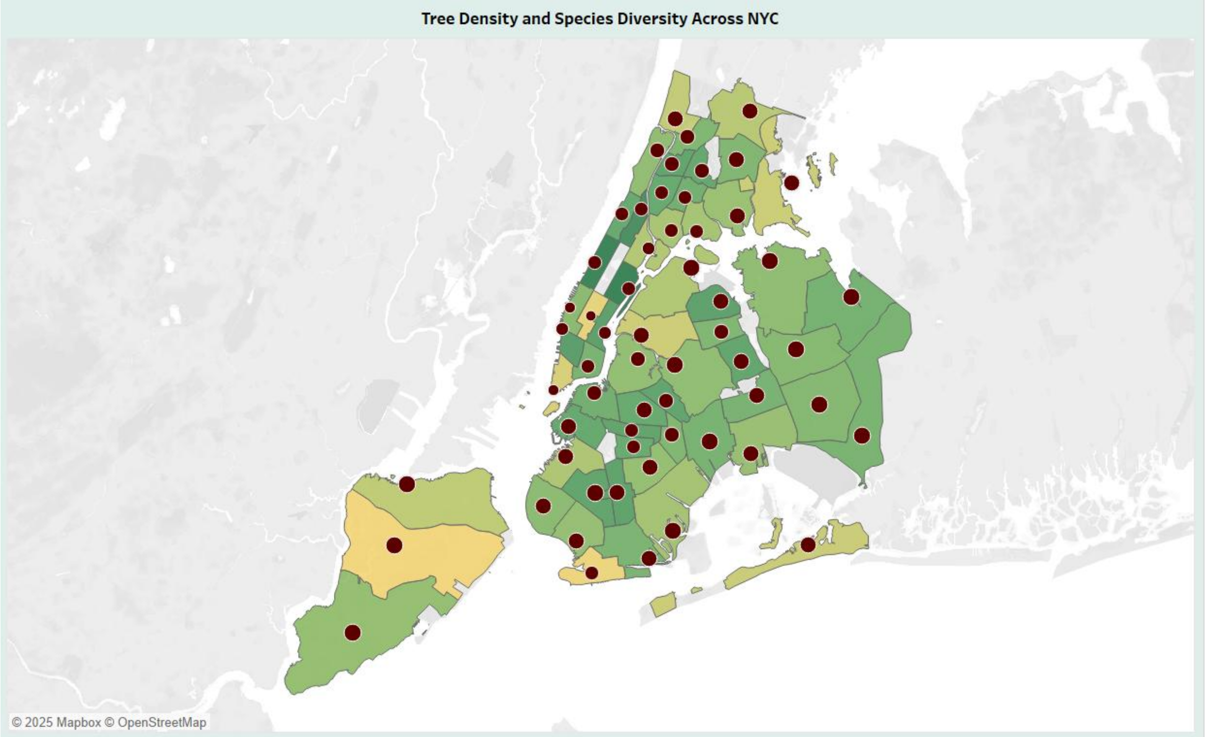

Tree Density and Species Diversity Across NYC
© 2025 Mapbox © OpenStreetMap

Figure 3: Tree Density and Species Diversity Across NYC

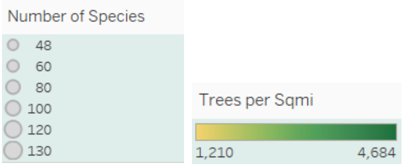


Chart 4(Bar Chart): Top 10 Most Common Tree Species

<u>Variables:</u>

- Tree Id (Continuous): Unique Discrete value.
- Spc Common (Categorical): Specie Common Name

<u>Narrative:</u> This bar chart displays the top 10 most common tree species across New York City, ranked by the total number of trees recorded in the 2015 Street Tree Census. The London planetree stands out as the most prevalent species, with over 87,000 individual trees, followed by honey locust, Callery pear, and pin oak, each with substantial counts exceeding 50,000. Other frequently found species include Norway maple, little leaf linden, and cherry, while ginkgo and Sophora round out the list with lower, yet still significant, numbers. This distribution indicates that while NYC's urban forest includes a range of species, it is somewhat dominated by a few, which could pose risks related to biodiversity loss and vulnerability to species-specific pests or diseases. This chart supports our analysis of species composition and helps highlight the importance of encouraging greater biodiversity in future tree planting strategies to improve ecological resilience across the city.

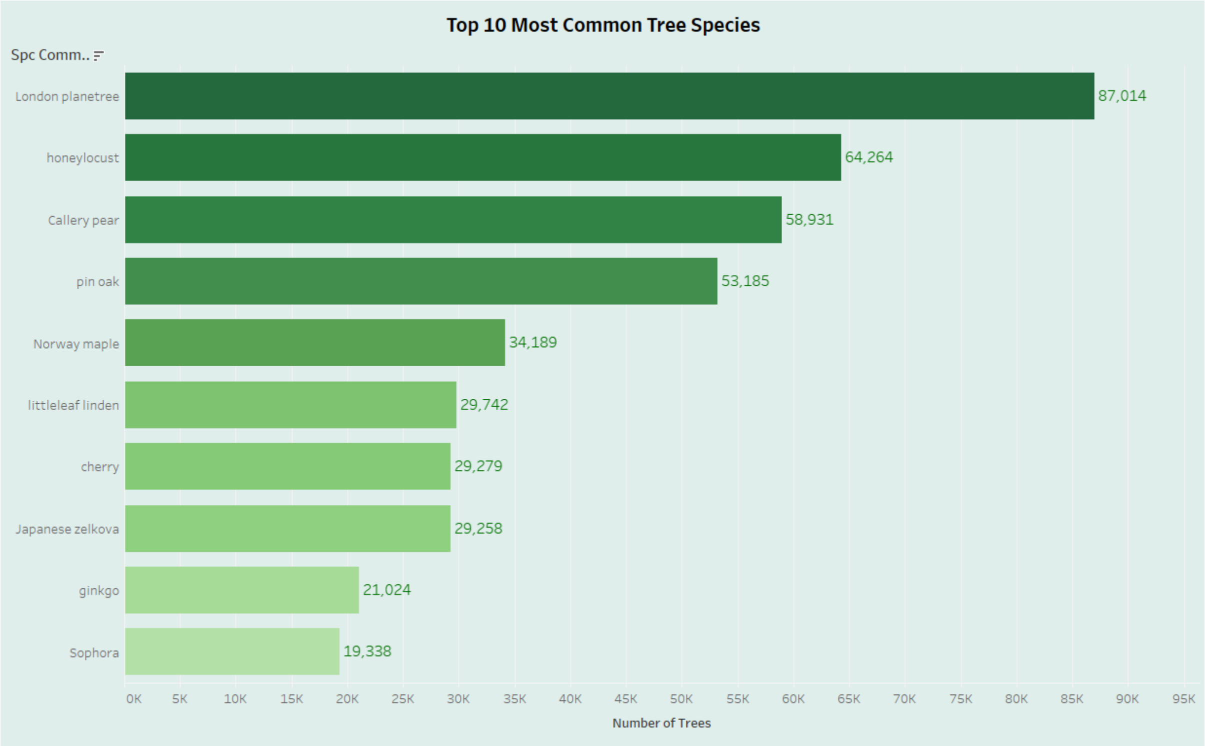


Figure 4: Top 10 Most Common Tree Species

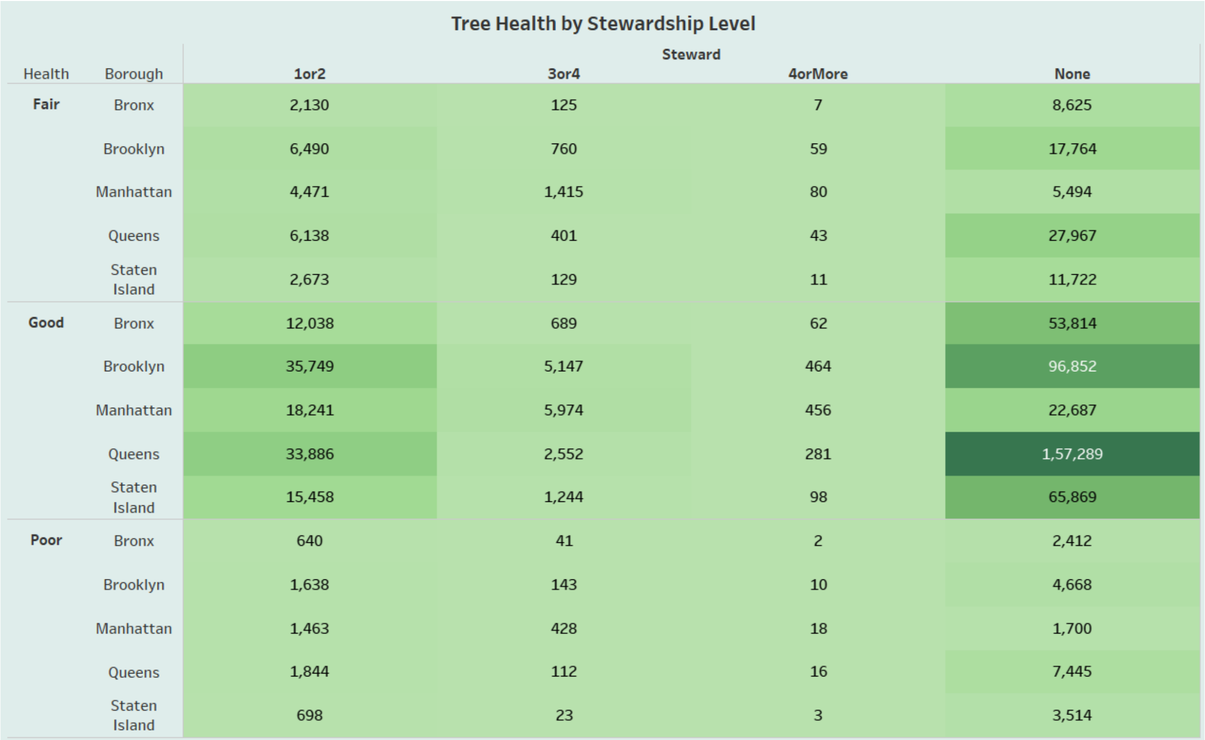
Chart 5(Table): Tree Health by Stewardship Level

Variables:

- Steward (Categorical): Organization, group, or individual responsible for maintaining trees (e.g., ½, ¾, none)
- Health (Categorical): Tree Health Status (e.g., Good, Fair, Poor).
- Tree Id (Continuous): Unique discrete value.

Narrative: This table visualizes the relationship between tree health conditions (categorized as *Good*, *Fair*, or *Poor*) and stewardship levels (None, 1or2, 3or4, 4orMore) across NYC's five boroughs. Stewardship refers to the level of care provided by individuals or groups, and the data suggests a positive association between higher stewardship and better tree health. Across all boroughs, trees with *no stewards* have the highest counts in each health category, which is expected given the larger sample size. However, when comparing proportions, trees categorized as *Good* are more frequent in stewarded groups, especially in boroughs like Brooklyn and Queens, where active stewardship (especially in the 1or2 and 3or4 ranges) is correlated with healthier tree populations. For example, in Brooklyn, over 35,000 trees with 1or2 stewards are in *Good* condition, while over 5,000 are stewarded at the 3or4 level. Conversely, *Poor* health categories show significantly lower numbers among stewarded trees, indicating that higher stewardship involvement may contribute to improved tree health. This chart supports our research objective of exploring how community engagement influences urban tree vitality and highlights the importance of promoting local stewardship programs to maintain a healthy urban canopy.

While trees with no stewardship outnumber all other groups due to the large baseline population, it's noteworthy that trees with 1 or 2 stewards consistently show higher counts across all health categories (*Good*, *Fair*, and *Poor*) compared to those with 3 or 4 or 4 or more stewards. This pattern is not necessarily a reflection of effectiveness, but rather a matter of distribution, most trees that do receive care are typically maintained by individuals or small groups, while fewer trees are under the care of highly engaged or organized stewardship efforts (like community organizations or environmental groups). Thus, the lower counts in the '4 or more' category stem from the rarity of such intensive stewardship, not its impact. In fact, when viewed proportionally, trees in the '4 or more' steward group tend to show a higher percentage in *Good* condition, suggesting that while fewer in number, deep engagement may be associated with better outcomes. This reinforces the idea that both the presence and intensity of community care play a role in urban tree health.



Tree Health by Stewardship Level

		Steward			
Health	Borough	1or2	3or4	4orMore	None
Fair	Bronx	2,130	125	7	8,625
	Brooklyn	6,490	760	59	17,764
	Manhattan	4,471	1,415	80	5,494
	Queens	6,138	401	43	27,967
	Staten Island	2,673	129	11	11,722
Good	Bronx	12,038	689	62	53,814
	Brooklyn	35,749	5,147	464	96,852
	Manhattan	18,241	5,974	456	22,687
	Queens	33,886	2,552	281	1,57,289
	Staten Island	15,458	1,244	98	65,869
Poor	Bronx	640	41	2	2,412
	Brooklyn	1,638	143	10	4,668
	Manhattan	1,463	428	18	1,700
	Queens	1,844	112	16	7,445
	Staten Island	698	23	3	3,514

Figure 5: Tree Health by Stewardship Level

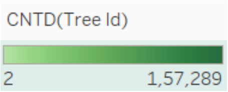


Chart 6(Bar Chart): Percent of Trees Affected by Each Stress Factor

Variables:

- Tree Id (Numeric): Unique discrete value.
- Stress factors (Categorical): These represent different environmental or factors effecting trees.

Narrative: This bar chart highlights the percentage of trees affected by various stress factors across New York City, helping to fulfill our project's goal of identifying the most common environmental and physical challenges faced by urban trees. The most significant stressor is stones around tree beds, impacting 21.47% of all recorded trees, potentially obstructing root expansion and water absorption. The next most common issue is branches tangled in lights (9.56%), followed by generalized trunk and root-related damages labeled as TrunkOther (4.99%) and RootOther (4.65%). Other issues like wires and ropes, metal grates, and rare cases such as sneakers tied to trees appear with lower frequencies but still signify urban interference with natural growth.

This chart is crucial in guiding both city agencies and community stewards toward targeted interventions. Understanding that a large proportion of stress comes from physical infrastructure (stones, lights, grates) informs where to prioritize design changes or maintenance efforts. Overall, it directly supports our objective of exploring what stress factors are most prevalent in NYC's street trees and how urban conditions might be contributing to declining health in certain areas.

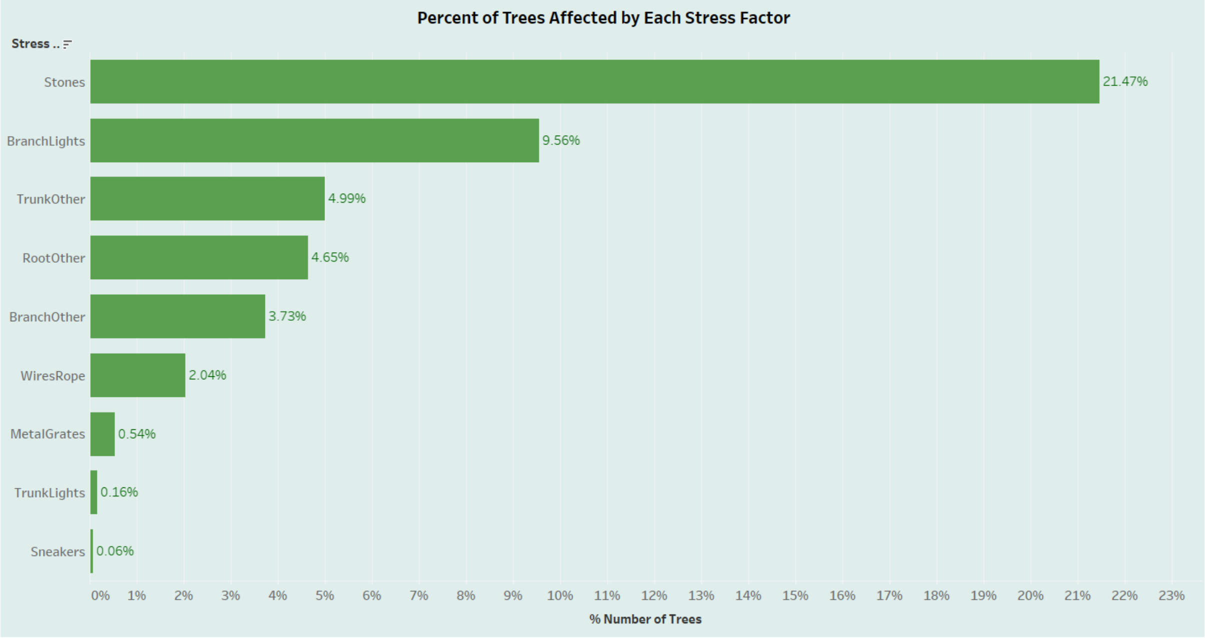


Figure 6: Percent of Trees Affected by Each Stress Factor

Phase 2: Analyze spatial and temporal variations in pollution levels across NYC.

Key Questions:

- What are the spatial patterns in PM2.5 or NO2 levels?
- Which districts consistently show higher or lower pollution?

Chart 7(Map): Air Pollution Levels by Community District

Variables:

- Name (Categorical): Name of the gas present (eg, NO2, Fine particle)
- Data Value (Continuous): NO2/ PM2.5 Pollution Concentration (µg/m3).
- Geo Place Name (Categorical): NYC neighbourhood name.
- Latitude & Longitude (Geospatial-Continuous): Exact geographic coordinates of tree.

Narrative: This map shows average PM2.5 levels across NYC's community districts. Higher concentrations are depicted in red, while cleaner areas are in pale orange. We observe consistently higher pollution levels near industrial zones and major traffic corridors, such as parts of the Bronx and areas adjacent to highways. Cleaner districts are generally located near parks or waterfronts. This helps answer our second-phase questions around spatial patterns of air pollution and identifies chronically affected areas.

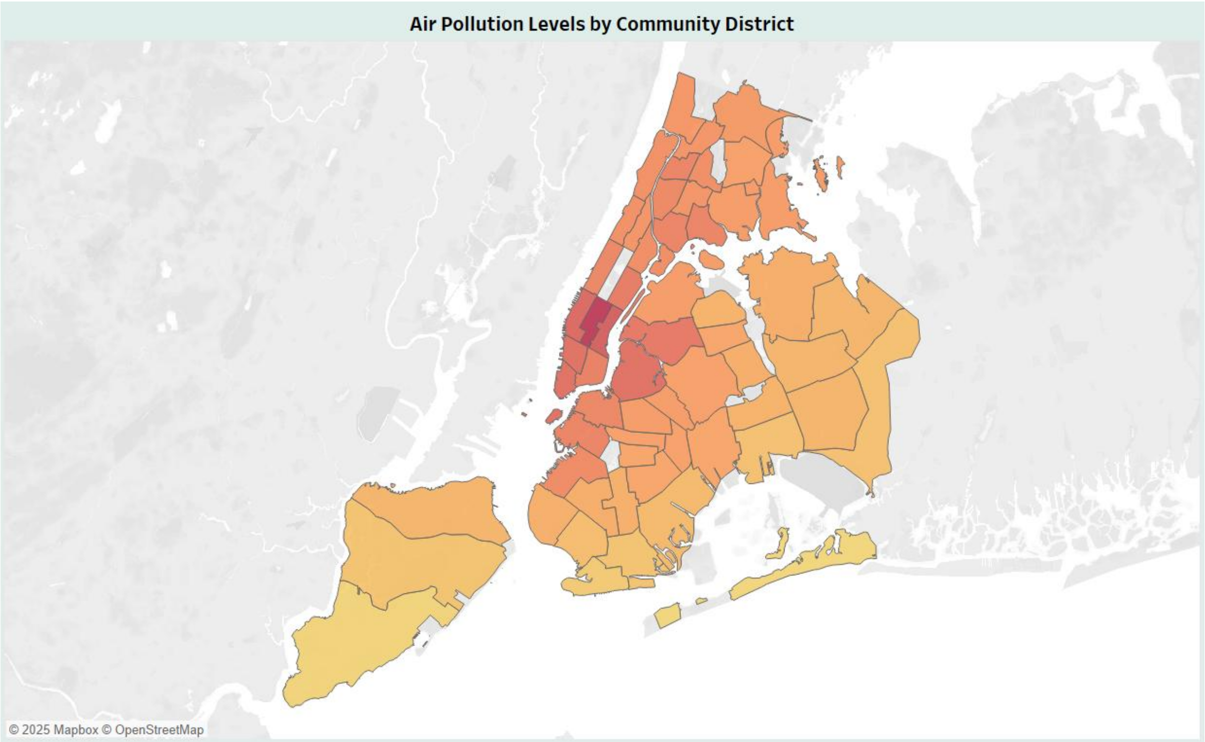
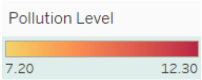

Air Pollution Levels by Community District
© 2025 Mapbox © OpenStreetMap

Figure 7: Air Pollution Levels by Community District



Phase 3: Overlay tree data with air quality metrics to reveal potential associations.

Key Questions:

- Do districts with more diverse or dense tree populations show better air quality?
- Does greater tree species diversity correlate with better air quality in NYC's community districts?

Chart 8(Scatter Plot): Tree Density vs Air Pollution by Community District

Variables:

- Name (Categorical): Name of the gas present (eg, NO2, Fine particle)
- Geo Place Name (Categorical): NYC neighbourhood name
- Trees per sqmi (Continuous): Number of trees per square mile in districts
- Data Value (Continuous): NO2/ PM2.5 Pollution Concentration (µg/m3).

Narrative: This scatter plot examines the relationship between tree density and average PM2.5 levels in each community district. A general downward trend suggests a negative correlation, districts with higher tree density often report lower pollution. While correlation does not imply causation, the pattern implies potential benefits of urban greenery in mitigating air pollution. This finding aligns with our hypothesis that tree density may improve air quality.

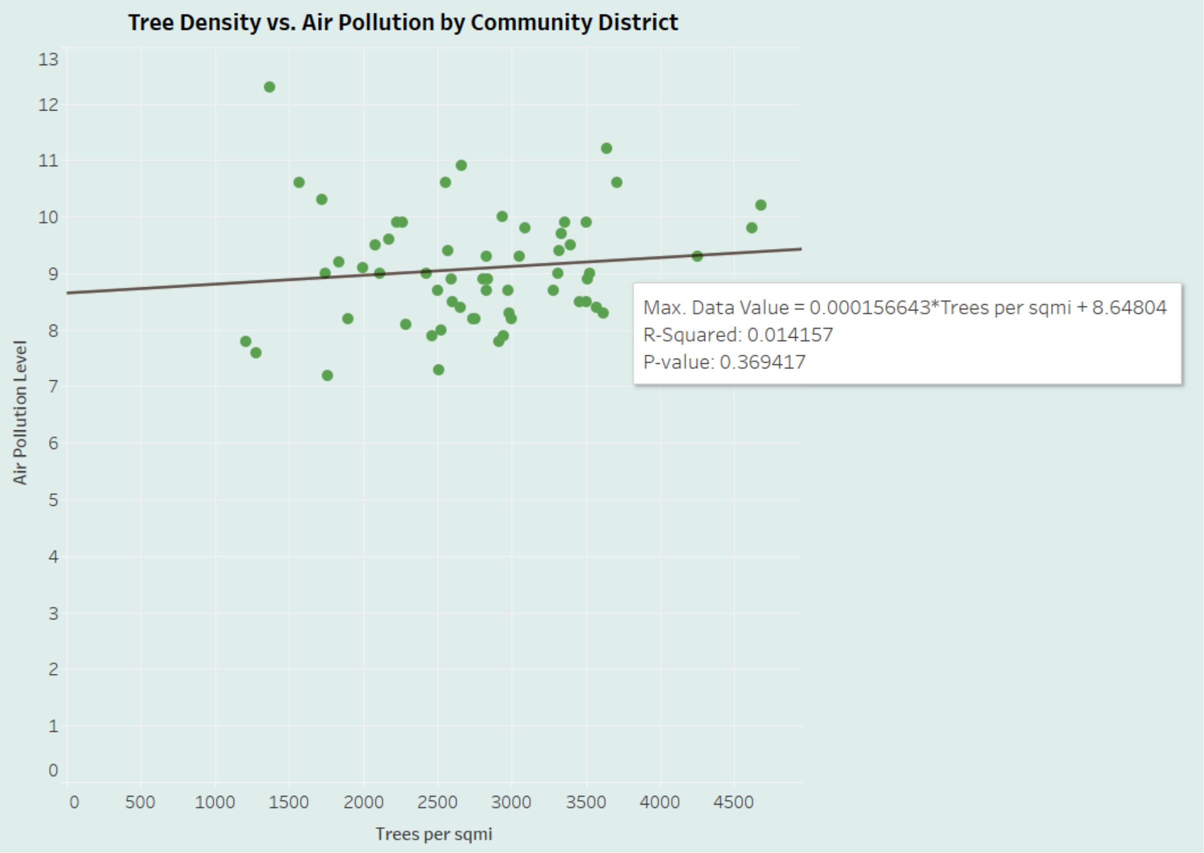


Figure 8: Tree Density vs Air Pollution by Community District

Chart 9(Scatter Plot): Species Diversity vs Air Pollution by Community District

Variables:

- Spc Common (Categorical):
- Data Value (Continuous): NO2/ PM2.5 Pollution Concentration (μg/m^3).
- Name (Categorical): Name of the gas present (eg, NO2, Fine particle).
- Geo Place Name (Categorical): NYC neighbourhood name.

Narrative: This scatter plot visualizes each community district with the x-axis representing the number of tree species (diversity) and the y-axis showing air pollution levels (likely PM2.5 concentration). The size of each point represents tree density (trees per square mile). Visually, there's a slight downward trend, districts with higher species diversity generally experience lower air pollution levels. While not every point follows this pattern, a majority of districts with over 110 species seem to fall into the lower pollution range (~7–9), while those with fewer species cluster higher on the pollution scale. This initial observation suggests a potential inverse relationship between biodiversity and pollution levels, laying the groundwork for hypothesis generation.

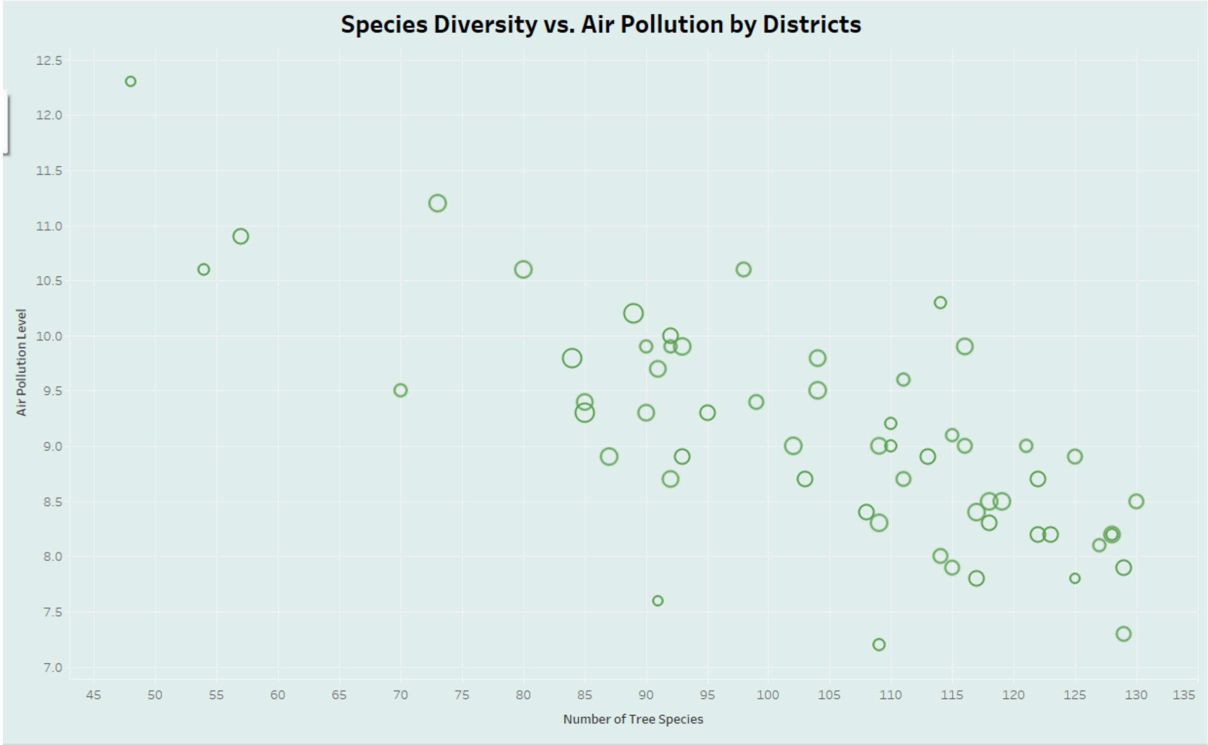


Figure 9: Species Diversity vs Air Pollution by Districts

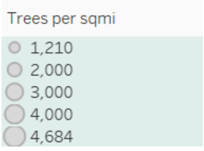
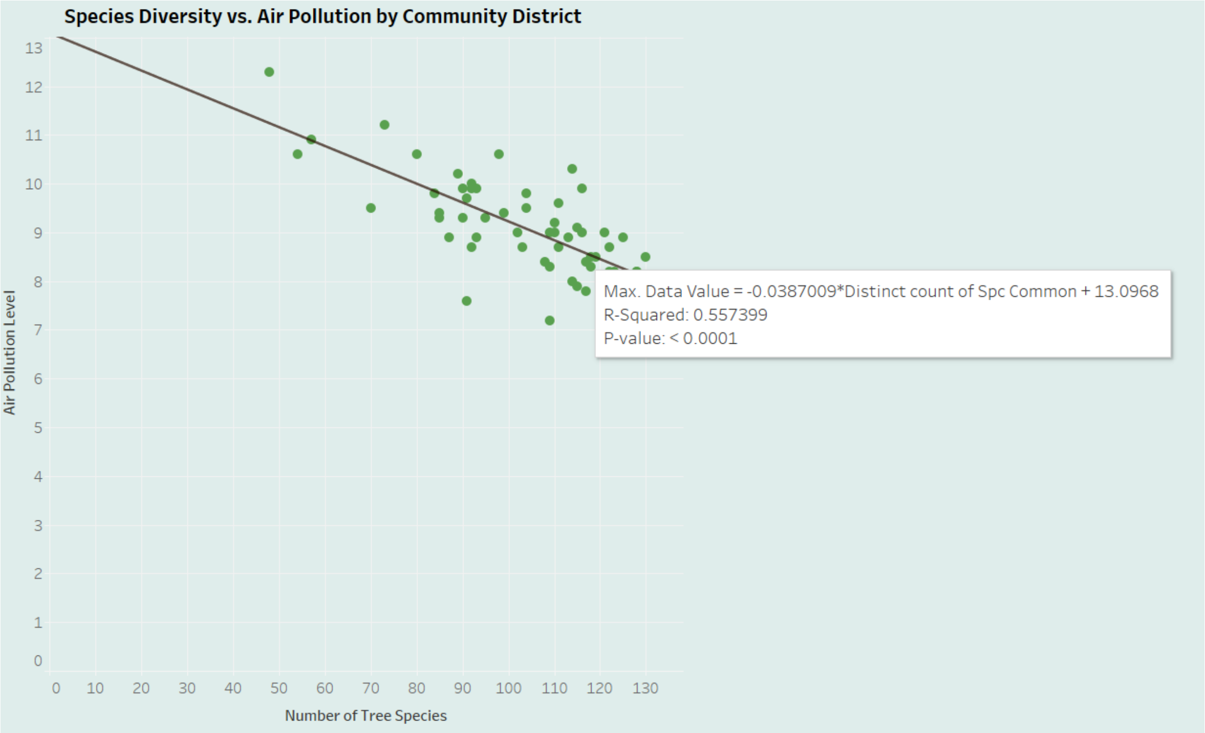


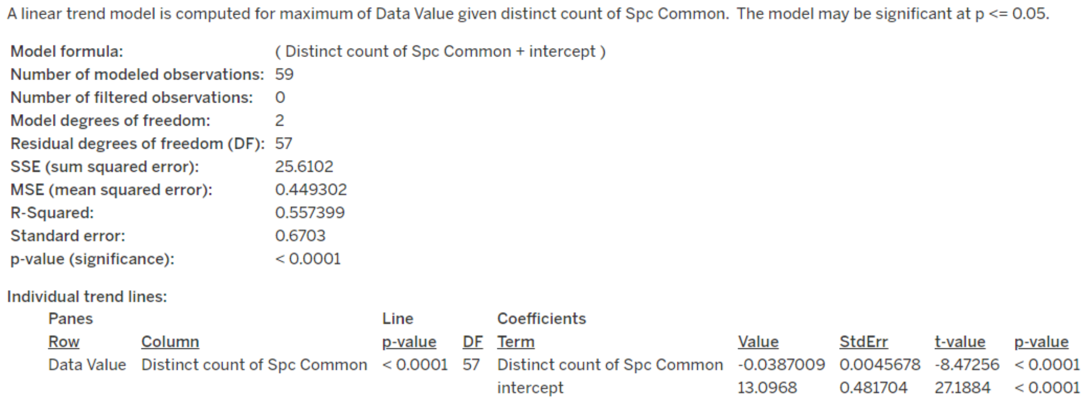
Chart 10: Statistical Significance shown for the above plot.

<u>Narrative:</u> To statistically validate the trend seen in the first plot, the second chart incorporates a regression line. The equation, R-squared value (0.557), and p-value (< 0.0001) provide quantitative support. The negative slope (-0.0387) confirms that air pollution levels decrease as the number of tree species increases. An R^2 of 0.557 indicates a moderate to strong linear relationship, explaining over 55% of the variation in pollution levels based on species diversity alone. The extremely low p-value further confirms that this relationship is statistically significant and not due to random chance.



Max. Data Value = -0.0387009*Distinct count of Spc Common + 13.0968
R-Squared: 0.557399
P-value: < 0.0001

Figure 10: Species Diversity vs Air Pollution by Community District



```
A linear trend model is computed for maximum of Data Value given distinct count of Spc Common.  The model may be significant at p <= 0.05.

Model formula:                          ( Distinct count of Spc Common + intercept )
Number of modeled observations:   59
Number of filtered observations:  0
Model degrees of freedom:               2
Residual degrees of freedom (DF): 57
SSE (sum squared error):                25.6102
MSE (mean squared error):               0.449302
R-Squared:                              0.557399
Standard error:                         0.6703
p-value (significance):                 < 0.0001

Individual trend lines:
      Panes                                        Line              Coefficients
      Row          Column                          p-value    DF   Term                            Value         StdErr       t-value     p-value
      Data Value   Distinct count of Spc Common    < 0.0001   57   Distinct count of Spc Common    -0.0387009    0.0045678    -8.47256    < 0.0001
                                                                   intercept                       13.0968       0.481704     27.1884     < 0.0001
```

Figure 11: Summary of Observations

4.Summary and Conclusion

Our project set out to explore the intersection of urban forestry and air quality in New York City by analysing tree distribution, biodiversity, health, and environmental stress indicators alongside district-level air pollution data. Rather than seeking causality, our aim was to uncover spatial and relational patterns that could inform environmental planning and policy.

Through our analyses and visualizations, we identified notable disparities in tree density and species diversity across NYC's community districts. Outer boroughs such as Queens and Brooklyn showed both high tree density and biodiversity, while central districts in Manhattan lagged, likely due to limited green infrastructure. The top 10 most common tree species were heavily dominated by a few types, such as the London planetree and honey locust, indicating a potential vulnerability to species-specific threats.

Stress factor analysis revealed that over 21% of trees face physical barriers like stones, while issues such as lights and trunk damage further complicate tree health. Notably, our stewardship analysis demonstrated that trees cared for by even 1 or 2 stewards were more frequently in 'Good' condition, underscoring the importance of community involvement in urban greenery maintenance.

The overlay of tree and air quality data brought forward compelling insights: districts with greater tree species diversity and higher tree density tended to exhibit lower levels of air pollution (PM2.5). Our regression analysis confirmed a statistically significant negative correlation between biodiversity and air pollution, with an R^2 value of 0.557 and a p-value < 0.0001. These findings support the notion that diverse and well-maintained urban forests can play a role in mitigating air pollution, while also fostering public health and environmental equity.

In sum, our project highlights how data visualization and spatial analysis can bring clarity to complex urban systems, emphasizing the co-benefits of urban forestry not only for aesthetics and ecology but also for clean air and resilient communities. These insights can guide city planners, environmental agencies, and local communities in making informed, data-driven decisions to create greener and healthier urban environments.

5. References

[1] EnviroTech (September, 2024), Growing Problem: NYC study suggests trees aren't always the answer. https://envirotecmagazine.com/2024/09/03/growing-problem-new-york-study-suggests-trees-arent-always-the-answer/

[2] Krajick, K. (2024, August 5). Planting some tree species may worsen, not improve, NYC … Lamont-Doherty Earth Observatory. https://lamont.columbia.edu/news/planting-some-tree-species-may-worsen-not-improve nyc-air-says-new-study

[3] Yuan Lai, Constantine E. Kontokosta (March, 2019), The impact of urban street tree species on air quality and respiratory illness: A spatial analysis of large-scale, high resolution urban data https://www.sciencedirect.com/science/article/abs/pii/S135382921830621X

[4] Vittoria Traverso (May, 2020), The best trees to reduce air pollution https://www.bbc.com/future/article/20200504-which-trees-reduce-air-pollution-best

[5] Environment & Health Data Portal (2015), Air Quality https://a816-dohbesp.nyc.gov/IndicatorPublic/

[6] NYC Open Data, Tree Distribution NYC 2015 https://data.cityofnewyork.us/Environment/Air-Quality/c3uy-2p5r/about_data

[7] NYC Planning, Community District Profiles. https://communityprofiles.planning.nyc.gov/

6. Author Contributions

Conceptualization & Research Design: Snehitha Tadapaneni, Pramod Krishnachari, Jin Hyuk Son, Sanjana Muralidhar, Vaijayanti Shrikant Deshmukh

Data Collection & Processing: Jin-Hyuk Son, Snehitha Tadapaneni, Sanjana Muralidhar

Data Analysis & Visualization: Jin-Hyuk Son, Snehitha Tadapaneni, Sanjana Muralidhar, Vaijayanti Shrikant Deshmukh, Pramod Krishnachari

Literature Review: Vaijayanti Shrikant Deshmukh, Pramod Krishnachari

Writing – Original Draft: Snehitha Tadapaneni, Pramod Krishnachari

Writing – Review & Editing: Vaijayanti Shrikant Deshmukh, Jin Hyuk Son, Sanjana Muralidhar