Examining the Relationship Between Tree Quality and Socioeconomic Status in New York City

Charlie Cai lc4181@nyu.edu

Stephen Zhang cz1906@nyu.edu

CSCI-UA 476 Project

Abstract

Trees are not just aesthetically pleasing, but they also perform a critical function in maintaining life on our planet. Access to green spaces was essential because it provide critical habitats for people and help preserve the ecosystem. Recently, researchers had explored various factors that influenced the quality of trees, and household income was one such factor. This paper examined the relationship between income and the presence of trees in New York City. We hypothesized that higher income communities enjoyed greater access to green and healthy urban environments, while lower income communities faced a lack of such privileges. We conducted a correlational analysis to investigate the relationship between income and the quality of trees, including factors such as trunk size. Additionally, we utilized machine learning techniques to further explore this relationship. Our study failed to reject the null hypothesis, and our results indicated insufficient evidence to support a correlation between income and the presence of trees in New York City. Additionally, our findings showed that lower income communities exhibited a higher abundance of trees compared to their higher income counterparts, which suggested that the urbanization efforts in NYC may have prioritized tax income generation over fostering healthier neighborhoods. By exploring the relationship between income and tree distribution, our study shed light on the urbanization strategies employed in NYC. Understanding these patterns could inform policymakers and urban planners in their efforts to promote equitable access to green spaces and enhance the overall well-being of all communities.

Keywords: Correlation, Linear Regression, Decision Tree, Households Income, Tree Quality, MapReduce, PySpark

Introduction

Trees are crucial for humanity. Green spaces provide clean oxygen to the environment, which is vital not only for mammals such as humans but also for supporting a diverse range of animals and insects. Moreover, green spaces aid in maintaining and improving the ecosystem. Rainwater is absorbed by trees and other plants and then released back into the air as transpiration, creating the necessary environment for all life to thrive. Additionally, green spaces can positively impact people's emotions by providing spaces for outdoor activities and enhancing their happiness. Finally, green spaces can generate economic benefits by creating job opportunities. Given all those benefits, access to green spaces should be available to all citizens, and people’s awareness of the importance of trees needs to be raised.

There were extensive effort studying factors that contribute to better quality of green spaces. In a thought-provoking article published by The New York Times titled "Since When Have Trees Existed Only for Rich Americans?", the issue of tree accessibility and its correlation

with socioeconomic status is raised. In the article, the researchers examined the impacts of household income on tree quality and numbers in different neighborhoods, and they found out that individuals with higher incomes tend to have greater access to green spaces compared to those residing in lower-income communities. Inspired by this discussion, we seek to investigate whether a similar relationship exists in the context of New York City.

We hypothesize that for areas in NYC, such as Wall Street, Soho, and the Midtown financial district, with higher household income, there should be a greater abundance and quality of trees, indicating by the greater number of health trees and the truck size. On the other hand, for lower socio-economic areas, the trees could be less healthier and small in truck size. This hypothesis is based on the assumption that areas with higher family income will generate a higher tax income, which in turns boast more extensive tree plantingand coverage compared to lower income communities. To test our hypothesis, we collected data from NYC OpenData, which contains information about the quantity and quality of trees across different neighborhoods arranged by postal codes, or so-called Zip Codes. In addition, we collected data on area's household income across different neighborhoods in NYC from the U.S. Census Bureau. By preprocessing two datasets, we were able to group two datasets together based on the Zip codes, and then we investigated the relationship between tree quality and family income, which their correlation was calculated and machine learning models—Linear Regression and Decision Tree Regression—were implemented to further investigate and verify this relationship. Our findings suggest that there is no correlation between tree quality and family income, and the results fail to reject the null hypothesis.

Overall, this study can be used to inform policymakers on decisions around urban planning and resource allocation, as the results will promote more sustainable and equitable access to green spaces across different neighborhoods in NYC. Meanwhile, we hope that the study could potentially draw conclusions about the mental health impact and personal well-being of the lack of access to green spaces in the urban environment.

Related Work

Green Space Impact In "The Environmental Injustice of Green Gentrification: The Case of Brooklyn's Prospect Park", the paper highlights the issue of green gentrification in Prospect Park, Brooklyn. It explains how the park's renovation led to an increase in property values in the surrounding neighborhoods, ultimately leading to the displacement of low-income and minority residents. The paper also raises concerns about the equitable distribution of green space and the potential negative impacts of green gentrification.

Green Space Benefits In "Air Quality Effects of Urban Trees and Parks", the paper explores the positive impact of trees and parks on air quality in urban areas. The study shows that trees and green spaces can significantly reduce air pollution, particularly in areas with high levels of vehicular traffic. The article also highlights the need for policymakers to prioritize the planting and preservation of urban trees and parks to improve air quality and promote public health.

Green Space and Income In "Since When Have Trees Existed Only for Rich Americans?", the article argues that trees and parks have historically been associated with affluent neighborhoods, leading to a lack of greenery in low-income areas. The article also emphasizes the importance of equitable access to green spaces for all residents, regardless of their socioeconomic status. The author suggests that policies should be implemented to increase tree planting and preservation in underserved communities to improve public health and promote environmental justice.

Median Household Income and Relaivte Tree Density In the study, the researchers conclude that "results are statistically insignificant and do not directly point to issues of lack of investment in poor communities... MN04 and MN05 are the midtown business districts, little room for trees to be a priority... Residential high-income districts appear to have higher tree-densities."

Mapping Urban Trees and Income in Manhattan The article discusses a research project that used satellite imagery and socioeconomic data to analyze the distribution of trees and income levels in Manhattan. It claims, "the analysis showed that the NYC Park's department has achieved a fairly equitable distribution of trees across its city."

Datasets

To explore these questions, our research relies on three primary datasets as inputs for our analysis.

Firstly, we utilize the comprehensive tree count data collected by the NYC Department of Parks and Recreation in 2015. This dataset (**Table 1**) provides information on tree species, diameter, and the perceived health of each tree. With a staggering 684,000 trees counted, this dataset offers valuable insights into the city's urban forest, in total of 45 columns with information such as tree truck's diameter, status of tree, etc. However, it's important to note that the census includes only trees located in public spaces, excluding those in private areas. (Detailed Schema is at Appendix in **Table 1**)

Our second dataset (**Table 2**) originates from the U.S. Census Bureau, American Community Survey 1-Year Estimates, Table S1903 (2005-2019, 2021), which provides median household income data categorized by community districts. There are a total of 4161 records. (Detailed schams is at Appendix in **Table 2**)

Our third dataset is primarily used for data processing purpose. As the income data is categorized by community districts, while the tree data is categorized by zip codes, a third dataset means to establish a correspondence between income and tree data. We collected zip codes corresponding to each community district from reliable online sources. By linking the zip codes to the median household income data, we created a comprehensive dataset that associates income with specific zip codes.

Analytic Stages

Data Collection

To gather the necessary data for our research, we obtained two primary datasets: the 2015 Street Tree Census - Tree Data and Median Income dataset. During the initial investigation, we observed that both datasets were well-formatted and contained minimal null values, which facilitated a smoother analysis process. These findings assured us of the initial data's quality and reliability, thus enhancing the validity of our research outcomes.

To ensure compatibility between the income data, which is categorized by community districts, and the tree data, which is categorized by zip codes, we conducted a meticulous process of matching the two datasets by collecting the zip codes corresponding to each community district from the internet For instance, the district of Astoria (Q1) encompasses zip codes 11101, 11102, 11103, 11105, and 11106.

Data Preparation

All data preparation stages were carried out utilizing the NYU High Performance Computing's (HPC) DataProc platform, employing Java MapReduce (MR) for efficient processing. Employing the power and capabilities of the NYU High Performance Computing's DataProc platform, alongside Java MapReduce, allowed us to effectively preprocess the data, ensuring its suitability for subsequent analysis, and handle the large volumes of data involved in our research.

To prepare the Median Income dataset, we focused on extracting the rows that corresponded to "All Households" and the year "2015." This selection ensured that our analysis encompassed various household types, including Families, Families with Children, and Families without Children, which are all represented within the "All Households" category. Furthermore,

by narrowing down the data to the specific year of 2015, we aligned the income dataset with the corresponding tree dataset.

Similarly, for the tree dataset, our MapReduce process targeted rows that pertained to living trees. By excluding the data related to dead trees, which lack health values, we maintained the integrity and relevance of our analysis. Both MapReduce processes generated new files, which were subsequently utilized in the data profiling stage.

Table 3 and **Table 4** shows the resulting schamas of data preparation stage. We extracts DBH, health, and zipcode from tree dataset. From median income, we got location and data. The tables show field name, data type, and description of the field.

Data Profiling

The data profiling stage was conducted on the NYU High Performance Computing's DataProc platform using Java MapReduce (MR). This stage aimed to gain a deeper understanding of the datasets and extract key metrics for further analysis.

In one MapReduce process, we focused on profiling the DBH using the cleaned tree dataset from the previous stage. By grouping the data by zip code, we summed all the DBH values for each corresponding zip code and calculated the average DBH. This profiling process provided valuable insights into the distribution and average size of trees within each zip code. In another MapReduce process, we aimed to profile the tree health by counting the number of trees categorized as good, fair, and poor in each corresponding zip code.

As the tree dataset already included zip codes as primary keys, we encountered the challenge of associating zip code information with the Median Income dataset, which lacked this information. To overcome this, we introduced another dataset containing city names and zip code pairs. Mapper one read the city name and income pair, while Mapper two read the city name and zip code pair. Through a reducer process, we joined the income and zip code data using the city name as the key. The resulting output provided a dataset with zip code and income pairs.

Data Ingestion

In our final dataframe, the key identifier is the zip code, with each row containing columns such as median household income, tree diameter, counts of trees perceived as good, fair, and poor. It will be used to analyze the relationship between income and tree characteristics, we employed both linear regression and decision tree models. Each zip code serves as a data point within the models for various comparisons, including income vs. tree diameter, income vs. counts of good trees, fair trees, and poor trees.

Data Analysis & Result Modeling

The data analysis and result modeling stage involved the utilization of PySpark. We employed the spark-submit functionality on the NYU High Performance Computing's DataProc platform to execute the analysis at scale. The datasets obtained from the data profiling stage, including income, tree health, and DBH, were imported into PySpark as data frames.

Linear Regression Analysis

As shown in **Figure 2** at the Appendix, the code snippet performs linear regression analysis on a dataset to predict the diameter at breast height (TreeDBH) of a tree based on the median income of the area where the tree is located using PyShark. It started by selecting the "MedianIncome" column as the input feature and the "TreeDBH" column as the target variable. It then created a VectorAssembler object to assemble the features into a vector column named "features". The data is split into training and testing sets using an 80/20 ratio and a seed value of 42 to prevent overfiting. A Linear Regression model was then created with the "features" column as the input and "TreeDBH" as the label. The model was trained on the training data and used to predict on the testing data. Finally, the model's performance was evaluated using the mean squared error (MSE) and R-squared (R2) metrics, using the RegressionEvaluator class, as shown in **Table 5** at Appendix.

The similar procedures were performed for variables other than the“TreeDBH". From the results, for the variable "Income & TreeDBH," the R-Squared value of -0.12 suggests that the model explains very little of the variance in the dependent variable (income). Regarding the variable "Income & Good Quality Tree," the R-Squared value of -0.01 suggests that the model has no explanatory power for income when considering good quality trees. For the variable "Income & Fair Quality Tree," the R-Squared value of 0.01 indicates a marginal improvement in the model's explanatory power for income variation related to fair quality trees, although it remains quite low. Lastly, for the variable "Income & Poor Quality Tree," the R-Squared value of 0.01 suggests that the model explains a small portion of the income variance associated with poor quality trees.

In summary, the linear regression analysis shows that the models’ performance are generally poor for predicting income based on tree-related variables. The mean squared errors are relatively high for most cases, indicating significant prediction errors. Additionally, the R-Squared values are very low or negative, suggesting that the models have limited explanatory power for the income variance observed in the dataset. However, from those models’ slopes, we observed a weak negative correlation, meaning that every increase in median income will result in lower/less tree-related attributes, but results are statistically insignificant, suggesting that there we do not have insufficient evidence to support a correlation between income and tree quality. More details will be provided in the next section.

Decision Tree Regression Analysis

As shown in **Figure 3** at the Appendix, the code snippet creates a decision tree regression model using the DecisionTreeRegressor class. In order to maximize the model's performance, a hyperparameter tuning steps was implemented, which the model was then configured with a maximum depth of 15, a minimum number of instances per node of 10, a seed value of 42. Same as before, the data was splitted to have a ratio of 0.8 for training data and 0.2 for testing data. Afterwards, the input feature column was set to "features" while the label column was set to "TreeDBH". The model was trained on the training data using the fit() method. Next, the model was used to predict on the test data using the transform() method. The model's performance was evaluated using the mean squared error (MSE) metric, using the RegressionEvaluator class. The MSE and R-squared (R2) values are computed, as shown in **Table 6** at Appendix.

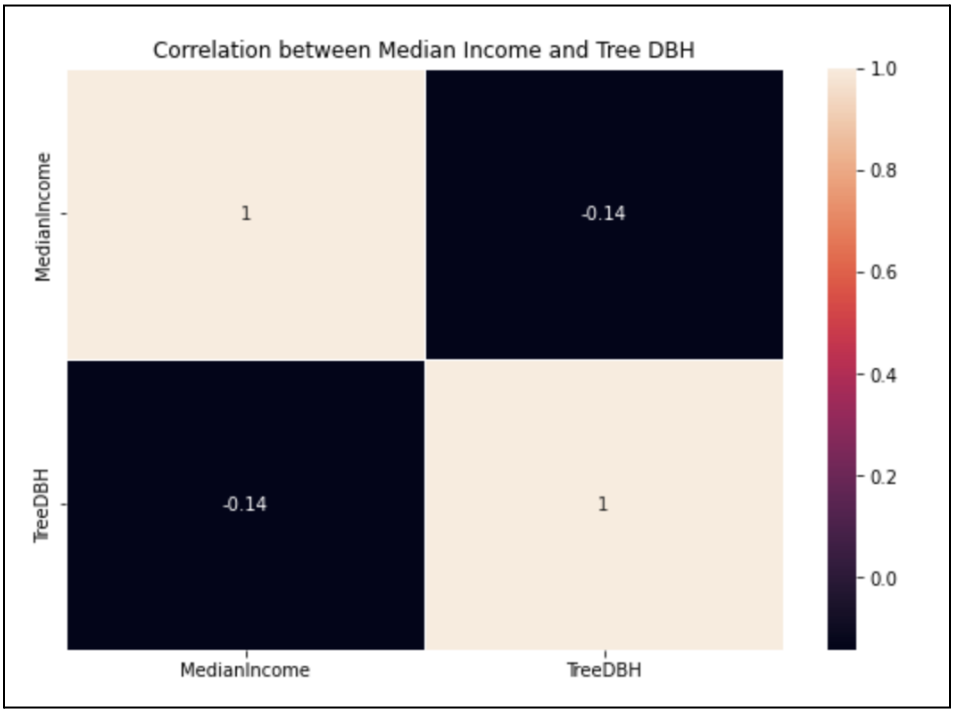
The similar procedures were performed for variables other than the"TreeDBH". From the results, for the variable "Income & TreeDBH," the R-Squared value of 0.24 indicates that the model explains approximately 24% of the variance in the dependent variable (income) when considering DBH. Regarding the variable "Income & Good Quality Tree," the mean squared suggests significant prediction errors in the model's performance. The R-Squared value of 0.20 indicates that the model explains approximately 20% of the income variance related to good quality trees. For the variable "Income & Fair Quality Tree,"the R-Squared value of 0.02 has a low explanatory power of the relationship. Lastly, for the variable "Income & Poor Quality Tree," the R-Squared value of 0.00 suggests that the model does not explain any meaningful variance in income when considering poor quality trees.

In summary, the decision regression analysis shows some improvement compared to the linear regression analysis for predicting income based on tree-related variables. The mean squared errors are generally lower, indicating reduced prediction errors. Additionally, the R-Squared values are higher, suggesting that the model explains a larger proportion of the income variance observed in the dataset. However, the overall explanatory power of the model remains relatively low, indicating that other factors beyond tree-related variables might influence income more significantly. However, from those models' tree diagram, the graph suggest that lower income communities exhibited a higher abundance of trees compared to higher income communities, but results are also statistically insignificant, suggesting that there we do not have insufficient evidence to support this relationship. More details will be provided in the next section.

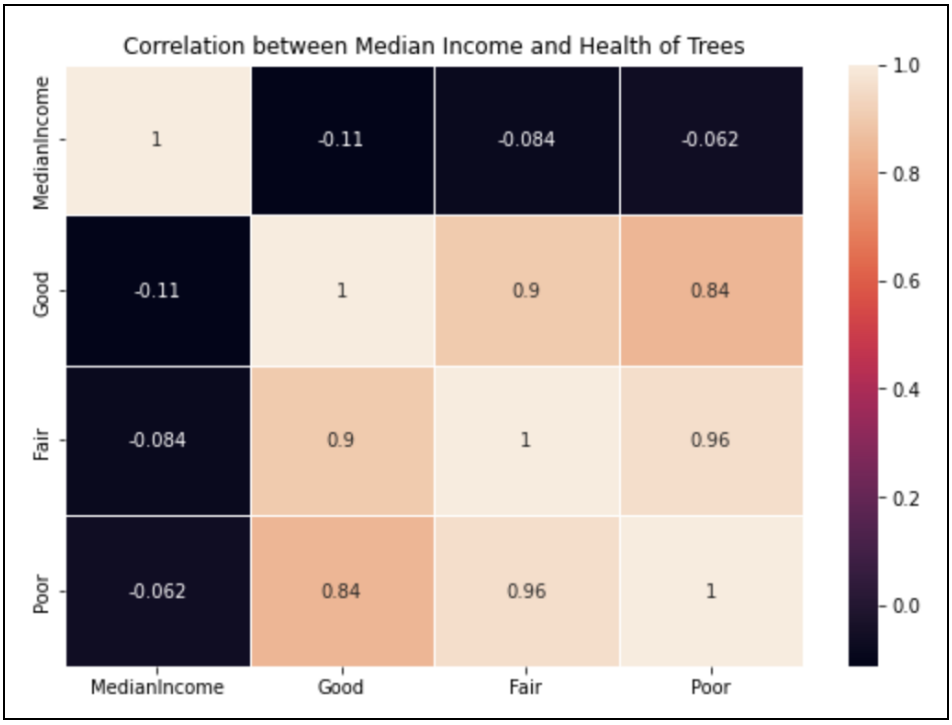
Graphs and Result

Correlation Analysis

There is a weak negative correlation, as shown in this heatmap. When income is 1, DBH is -0.14, meaning that tree truck's diameter increases when neighborhood's median income decreases.

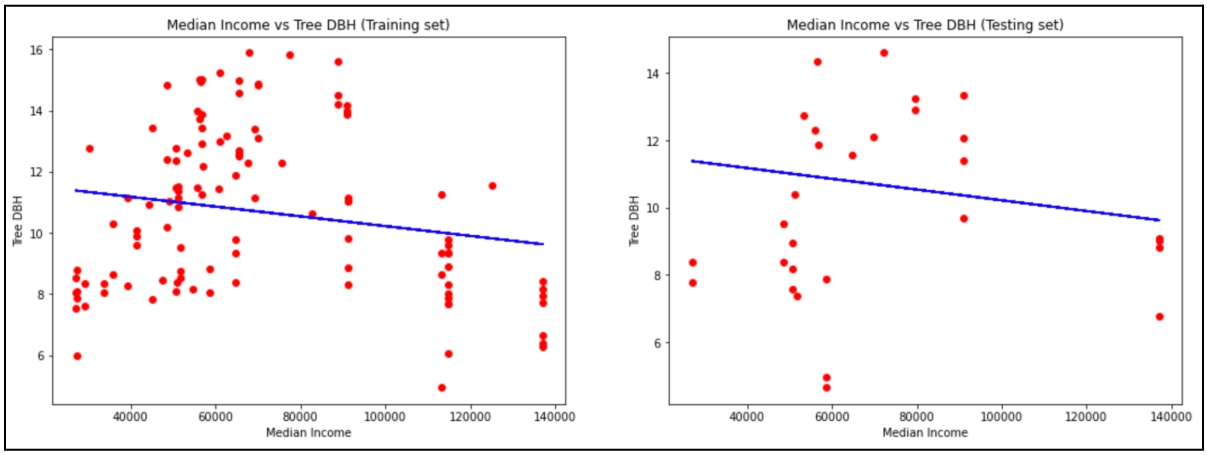


There are also weak negative correlations between number of healthy trees, fair trees, poor trees and median income. When median income is 1, number of healthy tree is -0.11, number of fair tree is -0.084, and number of poor tree is -0.062. This suggests that when income increases, good, fair, and poor tree counts decrease.

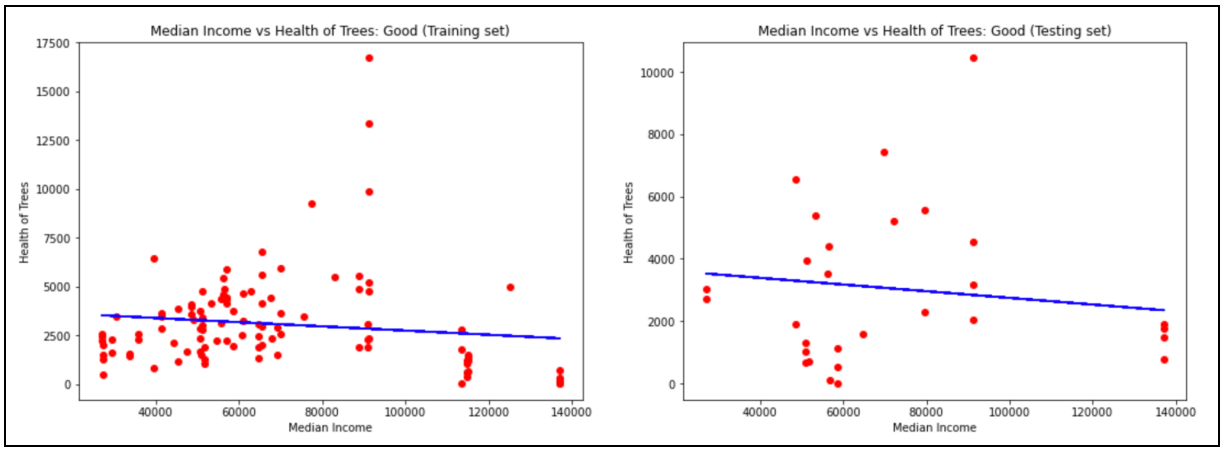


Linear Regression Analysis

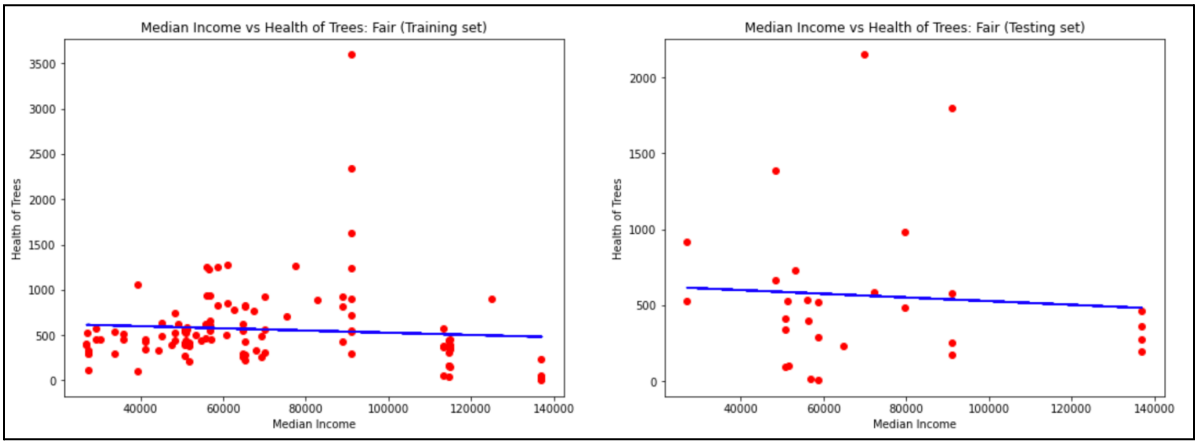
Based on the scatter plots and slopes, tree truck's diameter has a weak negative linear relationship (-1.60e-05) with median income. However, the results is statistically insignificant.



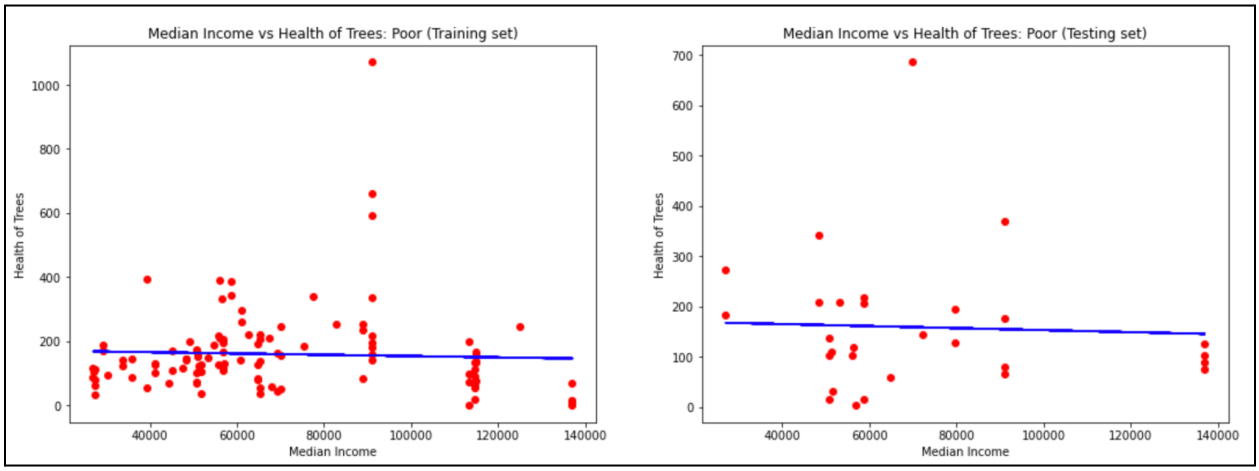
Based on the scatter plots and slopes, number of healthy tree has a weak negative linear relationship (-0.01) with median income. However, the results is statistically insignificant.



Based on the scatter plots and slopes, number of fair tree has a weak negative linear relationship (-1.2e-03) with median income. However, the results is statistically insignificant.

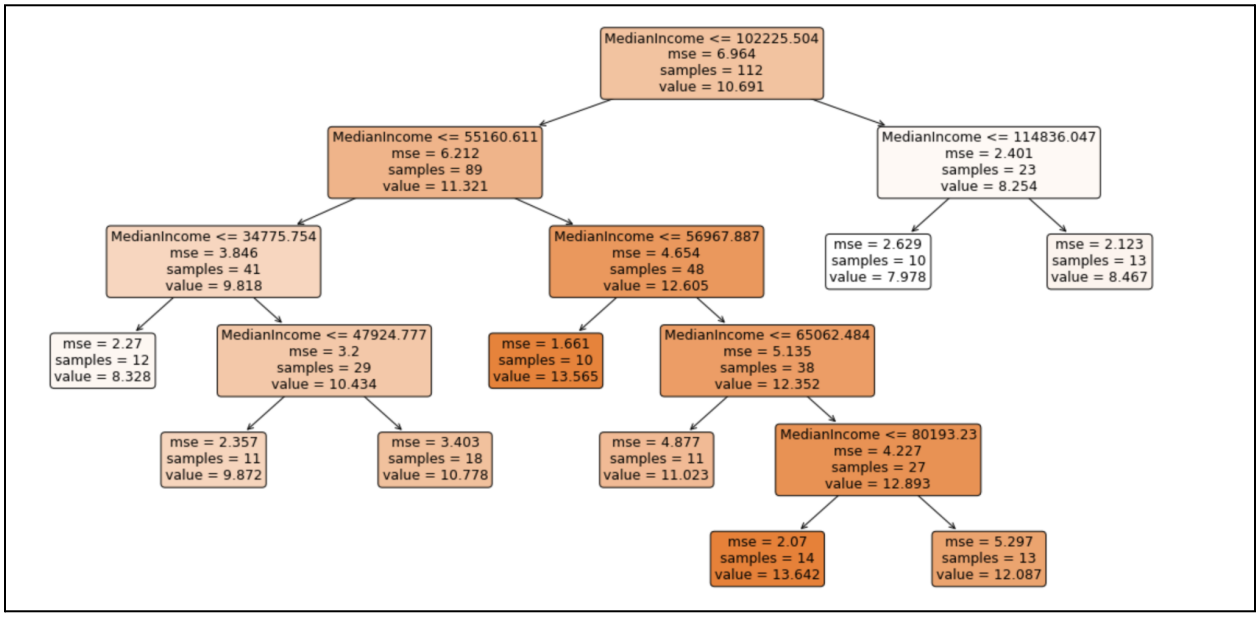


Based on the scatter plots and slopes, number of poor tree has a weak negative linear relationship (-2.01e-04) with median income. However, the results is statistically insignificant.

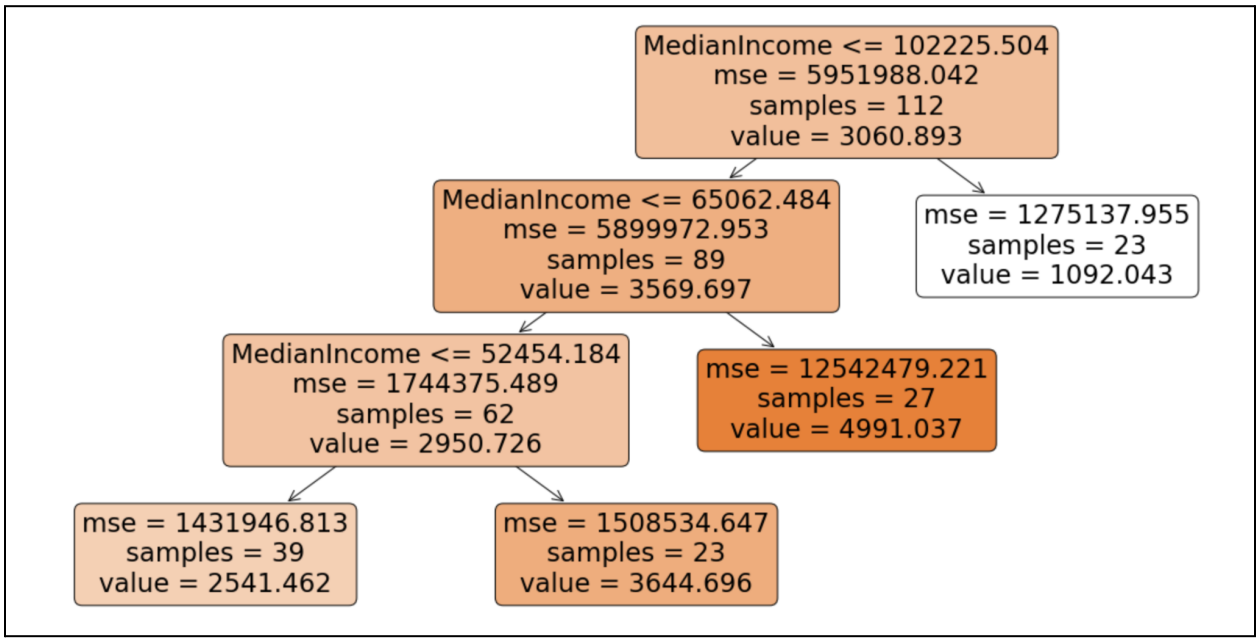


Decision Tree Regression Analysis

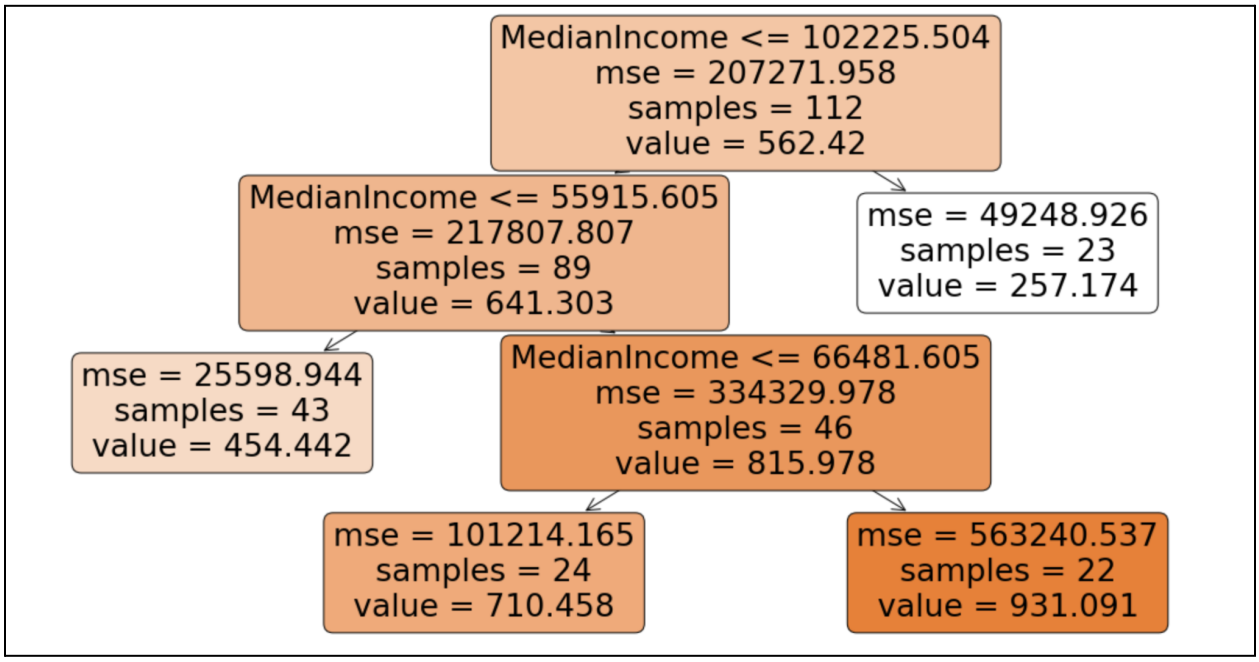
Based on the tree plots, the subtrees suggest that the tree truck's diameter decreases (indicated by value) as medican income increases (indicated by Median income). However, the result is statistically insignificant.



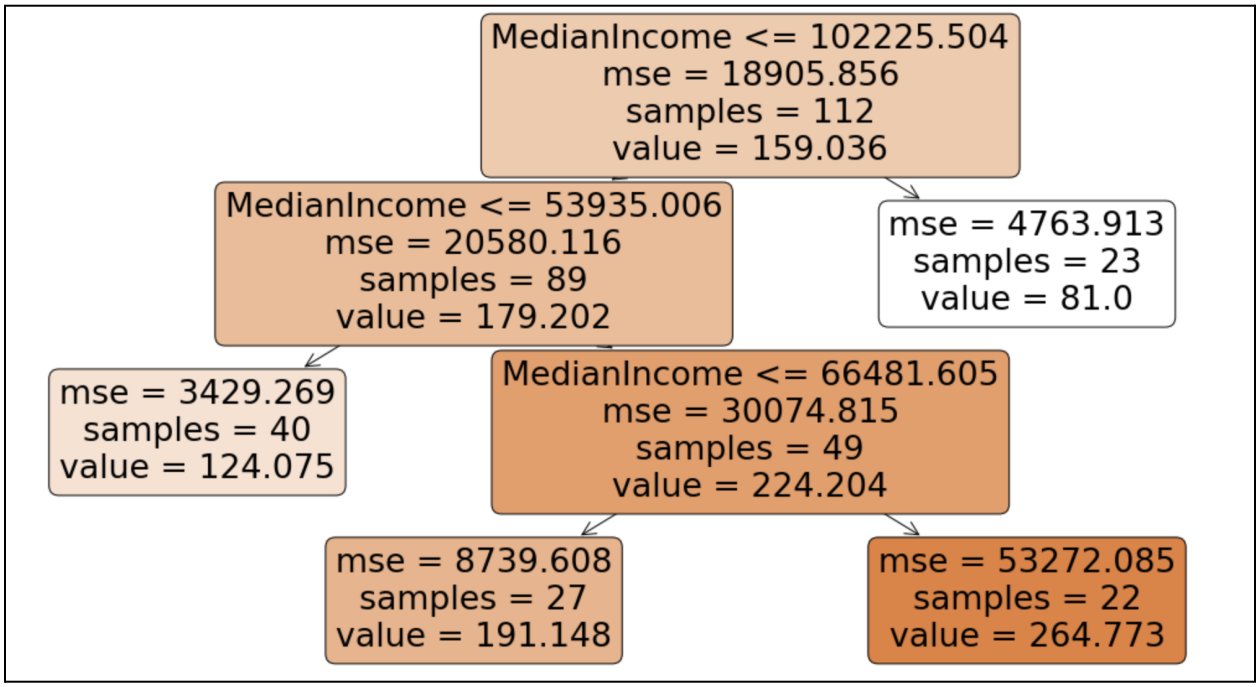
Based on the tree plots, the subtrees suggest that the number of healthy tree decreases (indicated by value) as medican income increases (indicated by Median income). However, the result is statistically insignificant.



Based on the tree plots, the subtrees suggest that the number of fair tree decreases (indicated by value) as medican income increases (indicated by Median income). However, the result is statistically insignificant.



Based on the tree plots, the subtrees suggest that the number of poor tree decreases (indicated by value) as medican income increases (indicated by Median income). However, the result is statistically insignificant.



Conclusion & Further work

Throughout this paper, we have investigated the relationship between income and trees in New York City, aiming to answer the fundamental question: Is there a relationship between household income and tree quality. Our findings indicates that there is a potentially negative correlation between them, as the results obtained from our analysis indicate a weak negative linear relationship between household income and both DBH and tree health. The correlation observed in these pairs of variables points towards a negative association, but the results are statistically insignificant. Meanwhile, the overall performance of the models used in this study was found to be poor, suggesting that income alone may not be the sole determinant of tree quality in the urban environment and the collected data might not sufficiently support the conclusion of the hypothesis.

During the process of further verification and explanation of our analysis, we conducted additional research on the data. As shown in **Figure 4** at Appendix, tree count distribution indicates Tottenville has the most trees of all the cities. Tottenville, as part of Staten Island in NYC, has more open lands to plant trees, but lower income compare to Tribeca and Lower East Side. Indeed, this helps to verify and rationalize our previous findings. We conclude that the lower income communities tends to have more open lands to grow trees. However, the higher income communities are more crowded and have less lands for tree planning and cultivation. As a results, we observe a potentially weak negative correlation between those variables. In addition, **Figure 5 and 6** at Appendix displays the number of trees per neigboirhood. The blue dots represent less than 10,000 trees, and red dots represent more than 10,000 trees. Lower income communities like Queens, Brooklyn, and Staten Island have more trees than Higher income communities like Downtown Manhattan and Upper East and West Side, which further supports our conclusion from the analysis.

To further investigate the relationship between household income and tree quality, more complex models such as neural networks could be utilized as our current models fail to capture the linear relationship between these variables. In a nutshell, the findings reveal that there is an unequal access to green spaces for all residents, regardless of their socioeconomic status. Therefore, policymakers and urban planners should prioritize building sustainable and accessible urban environments that prioritize equitable access to green spaces instead of solely considering socioeconomic factors. Creating sustainable, equitable, and healthy green spaces should be the ultimate goal for the New York City.

Acknowledgement

We are grateful to the support of NYU HPC's Dataproc Team, NYC OpenData, U.S. Census Bureau, American Community Survey, and Scikit-learn (Machine Learning in Python), especially the help and support from Professor Ann Malavet for her lectures, suggestions, and inspirations.

References

- Margolin, J., Bratslavsky, A., Bivas, J., Zia, I., & Monteith, M. (2020, December 4). Median household income vs tree density in NYC. STEAM Festival 2020. Retrieved May 2, 2023, from https://eportfolios.macaulay.cuny.edu/steamfest2020/2020/12/04/median-household-income-vs-tree-density-in-nyc/
- Dempsey, C. (2017, June 5). Mapping urban trees and income in Manhattan. Geography Realm. Retrieved May 2, 2023, from https://www.geographyrealm.com/mapping-urban-trees-and-income-in-manhattan/cccnewyork. (n.d.). Median Incomes. Keeping Track Online THE STATUS OF NEW YORK CITY CHILDREN. Retrieved May 6, 2023, from https://data.cccnewyork.org/data/table/66/median-incomes
- Department of Parks and Recreation (DPR). (2017, October 4). 2015 street tree census - tree data: NYC open data. 2015 Street Tree Census - Tree Data | NYC Open Data. Retrieved May 6, 2023, from https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh
- Leahy, I., & Serkez, Y. (2021, June 30). Since when have trees existed only for Rich Americans? The New York Times. Retrieved May 6, 2023, from https://www.nytimes.com/interactive/2021/06/30/opinion/environmental-inequity-trees-critical-infrastructure.html
- Gould, Kenneth & Lewis, Tammy. (2012). The environmental injustice of green gentrification: the case of Brooklyn's prospect park. The World in Brooklyn: Gentrification, Immigration, and Ethnic Politics in a Global City. 113-146.
- Nowak, D. A. V. I. D. J., & Heisler, G. M. (2010). *Air Quality Effects of Urban Trees and Parks*. National Recreation and Park Association. Retrieved May 8, 2023, from https://www.nrpa.org/globalassets/research/nowak-heisler-research-paper.pdf

Appendix

Table 1: 2015 Street Tree Census - Tree Data

Field Name	Data Type	Description
tree_id	Integer	Unique identification number for each tree point.
block_id	Integer	Identifier linking each tree to the block in the blockface table/shapefile that it is mapped on.
created_at	Date	The date tree points were collected in the census software.
tree_dbh	Integer	Diameter of the tree, measured at approximately 54" / 137cm above the ground. Data was collected for both living and dead trees; for stumps, use stump_diam
stump_diam	Integer	Diameter of stump measured through the center, rounded to the nearest inch.
curb_loc	Domain value: OnCurb, OffsetFromCurrb	Location of tree bed in relationship to the curb; trees are either along the curb (OnCurb) or offset from the curb (OffsetFromCurb)
status	Domain value: Alive, Dead, Stump	Indicates whether the tree is alive, standing dead, or a stump.
health	Domain value: Good, Fair, Poor	Indicates the user's perception of tree health.
spc_latin	Text	Scientific name for species, e.g. "Acer rubrum"
spc_common	Text	Common name for species, e.g. "red maple"
steward	Domain value: 1or2, 3or4, 4orMore, None	Indicates the number of unique signs of stewardship observed for this tree. Not recorded for stumps or dead trees.

Field Name	Data Type	Description
guards	Domain value: Harmful, Helpful, None, Unsure	Indicates whether a guard is present, and if the user felt it was a helpful or harmful guard. Not recorded for dead trees and stumps.
sidewalk	Domain value: Damage, NoDamage	Indicates whether one of the sidewalk flags immediately adjacent to the tree was damaged, cracked, or lifted. Not recorded for dead trees and stumps.
user_type	Domain value: Volunteer, TreesCount Staff NYC Parks Staff	This field describes the category of user who collected this tree point's data.
root_stone	Domain value: Yes, No	Indicates the presence of a root problem caused by paving stones in tree bed
root_grate	Domain value: Yes, No	Indicates the presence of a root problem caused by metal grates in tree bed
root_other	Domain value: Yes, No	Indicates the presence of other root problems
trunk_wire	Domain value: Yes, No	Indicates the presence of a trunk problem caused by wires or rope wrapped around the trunk
trnk_light	Domain value: Yes, No	Indicates the presence of a trunk problem caused by lighting installed on the tree
trnk_other	Domain value: Yes, No	Indicates the presence of other trunk problems
brch_light	Domain value: Yes, No	Indicates the presence of a branch problem caused by lights (usually string lights) or wires in the branches
brch_shoe	Domain value: Yes, No	Indicates the presence of a branch problem caused by sneakers in the branches
brch_other	Domain value: Yes, No	Indicates the presence of other branch problems
address	Text	Nearest estimated address to tree

Field Name	Data Type	Description
zipcode	Integer	Five-digit zipcode in which tree is located
zip_city	Text	City as derived from zipcode. This is often (but not always) the same as borough.
cb_num	Integer	Community board in which tree point is located
borocode	Domain value: 1(Manhattan) ,2(Bronx), 3(Brooklyn), 4(Queens), 5(Staten Island)	Code for borough in which tree point is located
boroname	Domain value: Manhattan, Bronx, Brooklyn, Queens, Staten Island	Name of borough in which tree point is located
cncldist	Integer	Council district in which tree point is located
st_assem	Integer	State Assembly District in which tree point is located
st_senate	Integer	State Senate District in which tree point is located
nta	Text	This is the NTA Code corresponding to the neighborhood tabulation area from the 2010 US Census that the tree point falls into.
nta_name	Text	This is the NTA name corresponding to the neighborhood tabulation area from the 2010 US Census that the tree point falls into
boro_ct	Text	This is the boro_ct identifyer for the census tract that the tree point falls into.
state	Text	All features given value 'New York'
latitude	Double	Latitude of point, in decimal degrees
longitude	Double	Longitude of point, in decimal degrees

Field Name	Data Type	Description
x_sp	Double	X coordinate, in state plane. Units are feet.
y_sp	Double	Y coordinate, in state plane. Units are feet

Table 2: Median Income

Location	**Text**	**City name**
Household Type	Domain value: All Households, Families, Families with Children, Families without Children	The type of income
TimeFrame	Integer	The year of the data
DataFormat	Text	All features given value 'Dollar'
Data	Integer	Income in dollars

Table 3: 2015 Street Tree Census - Tree Data

Field Name	Data Type	Description
tree_dbh	Integer	Diameter of the tree, measured at approximately 54" / 137cm above the ground. Data was collected for both living and dead trees; for stumps, use stump_diam
health	Domain value: Good, Fair, Poor	Indicates the user's perception of tree health.
zipcode	Integer	Five-digit zipcode in which tree is located

Table 4: Median Income

Field Name	Data Type	Description
Location	Text	City name
Data	Integer	Income in dollars

Table 5: Linear Regression Analysis Performance Evaluation

Linear Regression Analysis	Mean Squared Error	R-Squared
Income & TreeDBH	7.68	-0.12
Income & Good Quality Tree	5932745.15	-0.01
Income & Fair Quality Tree	236931.98	0.01
Income & Poor Quality Tree	17613.32	0.01

Table 6: DecisionTreeRegressor Analysis Performance Evaluation

DecisionTreeRegressor Analysis	Mean Squared Error	R-Squared
Income & TreeDBH	5.22	0.24
Income & Good Quality Tree	4700625.81	0.20
Income & Fair Quality Tree	234905.51	0.02
Income & Poor Quality Tree	17815.88	0.00

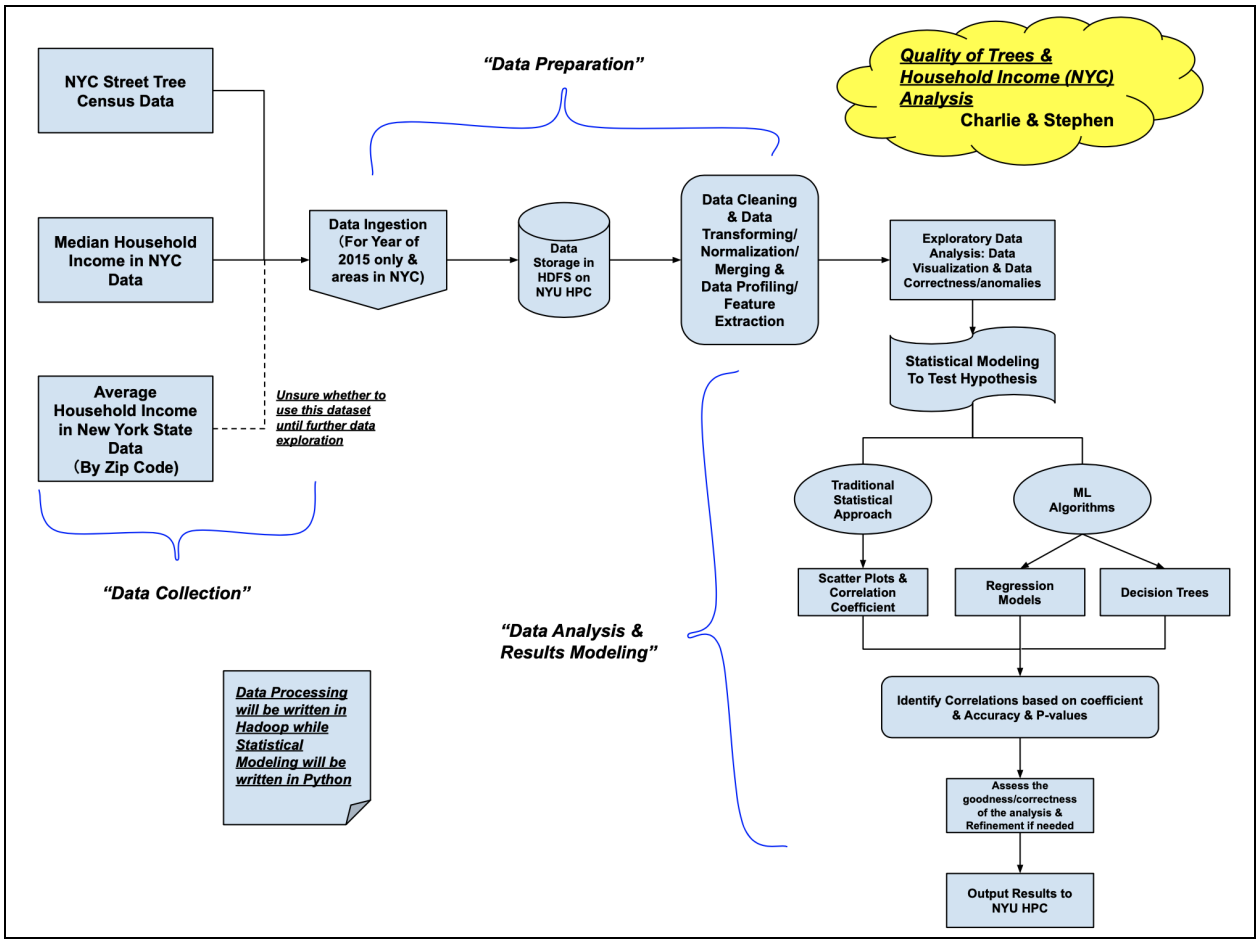


Figure 1: Design Diagram

```
# Select columns for input features and target variable
assembler = VectorAssembler(inputCols=["MedianIncome"], outputCol="features")
data = assembler.transform(final).select("TreeDBH", "features")
# Split data into training and testing sets
train_data_dbh, test_data_dbh = data.randomSplit([0.8, 0.2], seed=42)
# Create a Linear Regression model
lr = LinearRegression(featuresCol="features", labelCol="TreeDBH")
# Train the model
model = lr.fit(train_data_dbh)
# Make predictions on the testing data
predictions = model.transform(test_data_dbh)
# Evaluate the model's performance
evaluator = RegressionEvaluator(labelCol='TreeDBH', predictionCol='prediction', metricName='mse')
mse = evaluator.evaluate(predictions, {evaluator.metricName: 'mse'})
r2 = evaluator.evaluate(predictions, {evaluator.metricName: 'r2'})
```

Figure 2: Linear Regression Code Snippet

```
# Create a decision tree model
dt = DecisionTreeRegressor(maxDepth=15, minInstancesPerNode=10, seed=42, featuresCol="features", labelCol="TreeDBH")
# Train the model
model = dt.fit(train_data_dbh)
# Make predictions on the test data
predictions = model.transform(test_data_dbh)
# Evaluate the model using Mean Squared Error
evaluator = RegressionEvaluator(labelCol='TreeDBH', predictionCol='prediction', metricName='mse')
mse = evaluator.evaluate(predictions, {evaluator.metricName: 'mse'})
r2 = evaluator.evaluate(predictions, {evaluator.metricName: 'r2'})
```

Figure 3: Decision Tree Regressor Code Snippet

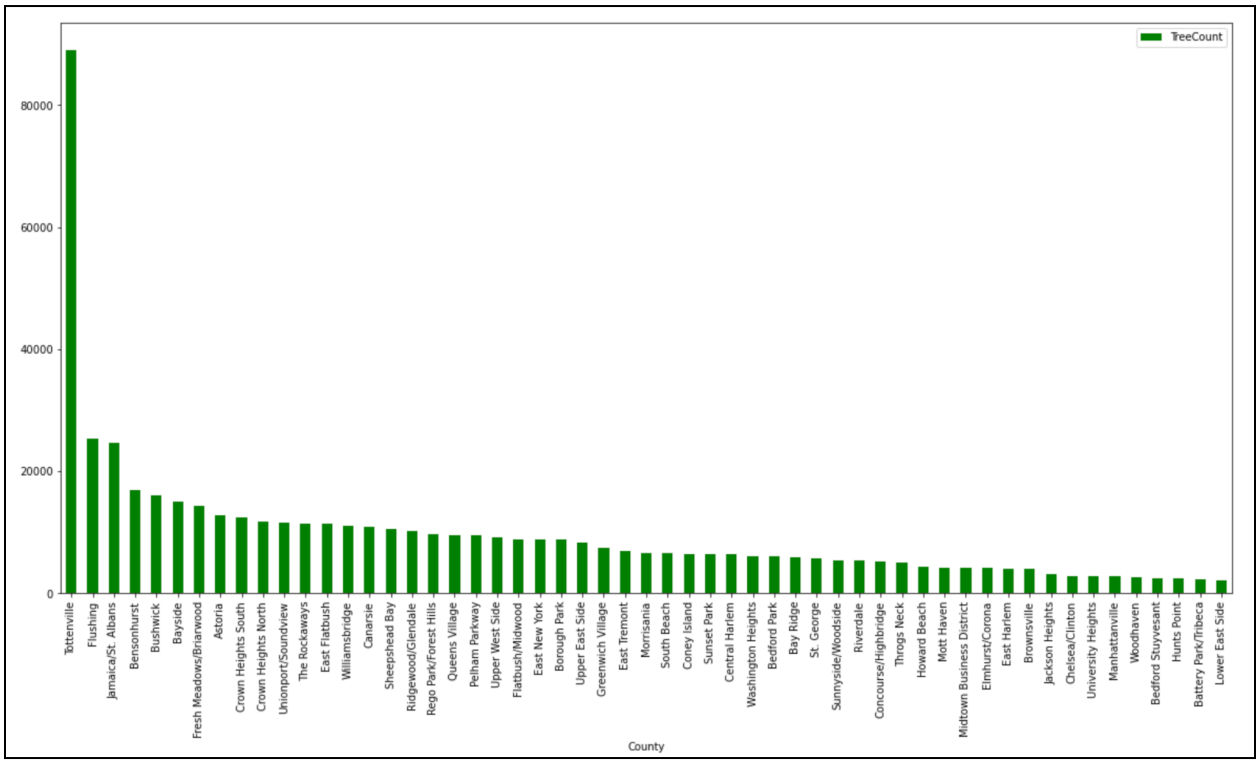


Figure 4: Tree Count Distribution by City Name

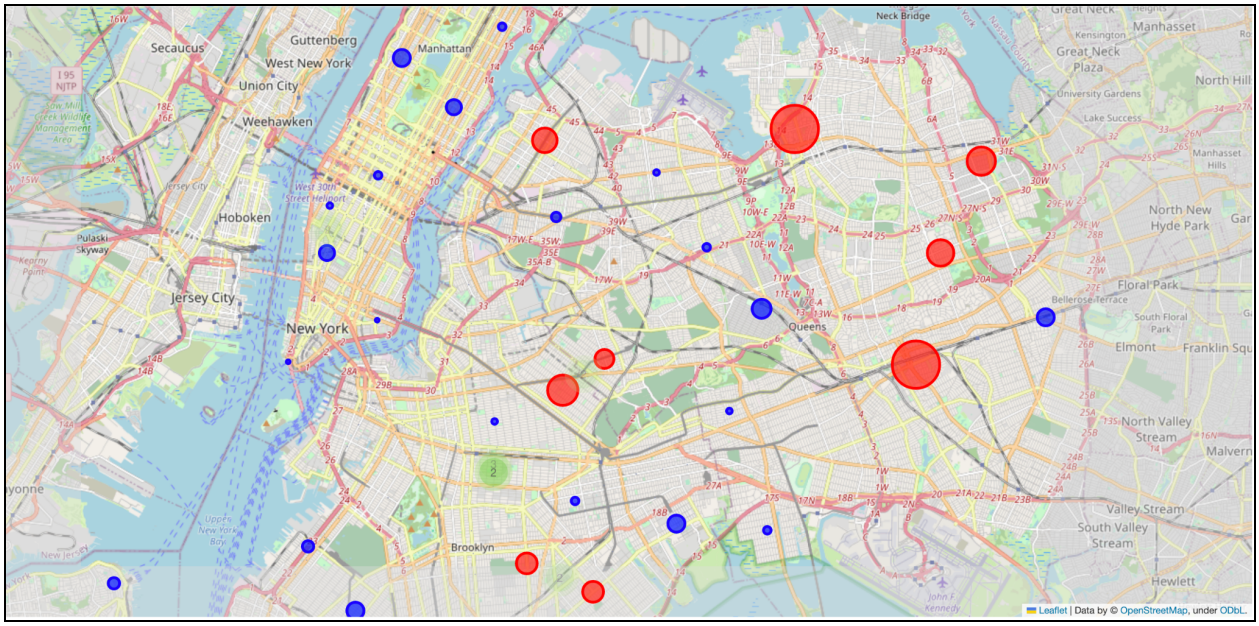


Figure 5: Tree Count Distribution on the Map

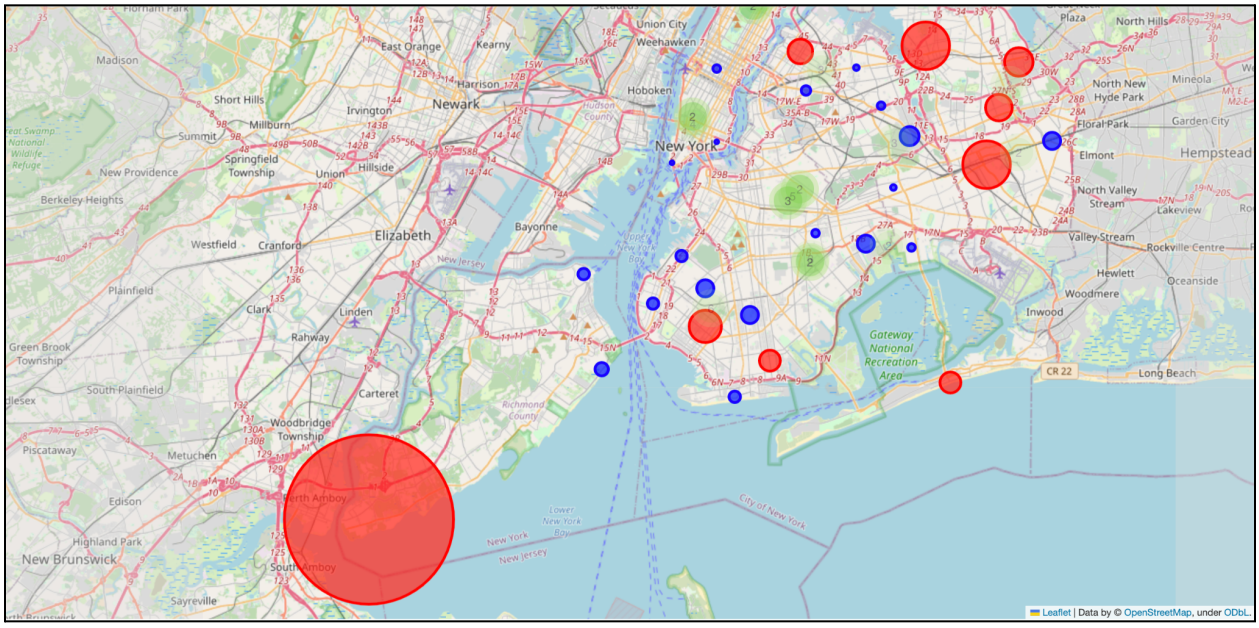


Figure 6: Tree Count Distribution on the Map