Final Written Report

AUTHOR
Tiffany Kuo, Salih Awouda, Caroline Klein

PUBLISHED
December 7, 2022

NYC Tree Census

Introduction & data

Low-income neighborhoods often lack the same resources available in more privileged areas. One of the disadvantages some of these neighborhoods face is a lack of tree coverage, which deprives them of many psychological and physical benefits. Other studies have found that low-income and minority neighborhoods are disproportionately subject to this lack of trees. For example, one study found that neighborhoods with 90% or more of their residents living in poverty (Greenberg, Sept. 16, 2020). Trees absorb carbon dioxide and offset the effects of climate change, leaving those with less tree coverage victims of greater heat and pollution. Low-income neighborhoods face a similar issue when they have a larger proportion of unhealthy trees compared to wealthier neighborhoods.

With this research, we hope to reveal potential instances of this inequity in New York City in order to encourage the production of solutions if needed. We will be investigating the question: How are the health of trees affected by the income of the area they are in? We predict that the higher the income of the zipcode region, the healthier the trees will be because the area will have more resources for tree upkeep and protection. The null hypothesis would be that there is no relationship between the income of a neighborhood and the health of its trees.

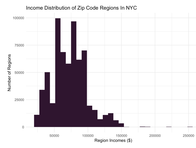
The primary data we will use is the 2015 NYC Tree Census from NYC Open Data. This census was conducted by volunteers and staff from NYC Parks & Recreation in 2015. These contributors documented every tree in New York City as a row, as well as information about the characteristics of the tree. The key variables we will be analyzing are species, apparent health (categorized as poor, fair, or good), zip code, and presence of stewardship or guards surrounding the tree.

We will also use a data set of the median incomes of different types of households in different zip codes of New York City. This data was collected in 2019 by the U.S. Census Bureau via a community survey, and was found on the website CCC New York. Each row represents a zip code and the value is the median income for a type of household (all households, families, families with children, and families without children) in dollars. We will be using the columns representing zip code and median income for all households.

Methodology

In order to investigate this question, we used the variable in the tree census representing the health of the tree. Trees were categorized as having poor, fair, and good health, which we will quantify as having a score of 1, 2 or 3 when processing the data. The following is a distribution visualization for tree health:

The key variable we will be examining the effect of is the income of the zip code region the tree is in. We will be organizing income into 4 income brackets with an equal number of trees in each category when we process the data. The distribution of these incomes is represented in the following graph:



The relationship between tree health and income can be seen through this visualization:

This plot does not seem to definitively support our hypothesis because the health of the trees appear to have a similar distribution across all zip code income brackets. It appears that for the first 3 income brackets, the number of healthy (Good) trees increases and number of unhealthy (Poor) trees decreases as the region income increases. However, the income range with the most trees of "Good" health and the least trees of "Poor" health is the second wealthiest, not the wealthiest group of zip code regions.

However, to test these apparent results, we will utilize summary statistics to calculate the difference in mean tree health score of each income bracket. We chose this method in order to compare the mean of the lowest income bracket to the mean of the higher income brackets. In this way, we can test our hypothesis that higher income corresponds to better tree health. With m_1 representing the mean of the lowest income bracket and m_k representing the means of each of the higher income brackets, the null hypothesis is $H_0 : m_k - m_1 = 0$ and the alternate hypothesis is $H_A : m_k - m_1 > 0$.

Income Bracket	Mean Tree Health
$0-$54278	2.753199
$54279-$70793	2.782665
$70794-$84713	2.791617
$84714-$250001	2.748717

Income Bracket	Difference in Mean Tree Health
First and Second Income Bracket	0.0294661
First and Third Income Bracket3	0.0384117
First and Fourth Income Bracket	-0.0044817

We also created two linear regression models to see how the health of a tree is affected by income, the species of a tree (whether it was a London planetree, the most common species, or not), the number of signs of stewardship for the tree, the status of the sidewalk, and the kinds of guards present near the tree. This statistic method helps with our investigation by controlling for these variables. For the first regression, we used income as a categorical variable using the four different brackets. In the second regression, we used income as a continuous variable and accounted for the possibility of it being a curvilinear relationship by squaring the income.

Term	Estimate	STD Error	Statistic	P-Value
(Intercept)	2.7962017	0.0021419	1305.4942932	0.0000000
income_bracket$54279-$70793	0.0281222	0.0017841	15.7628059	0.0000000
income_bracket$70794-$84713	0.0376613	0.0017793	21.1662543	0.0000000
income_bracket$84714-$250001	-0.0004655	0.0018155	-0.2563845	0.7976541
type	-0.0500514	0.0018762	-26.6763008	0.0000000
stewardsTor2	-0.0041611	0.0017075	-2.4369567	0.0148117
stewardsTor4	0.0170553	0.0042111	4.0500421	0.0000512
stewardsTor4orMore	0.0529587	0.0128428	4.1236140	0.0000373
sidewalkDamage	0.0055038	0.0014072	3.9112046	0.0000918
guardsHarmful	-0.0622818	0.0038606	-16.1327824	0.0000000
guardsUnsure	-0.0666873	0.0059230	-11.2472184	0.0000000
guardsHelpful	0.0167832	0.0027315	6.1442102	0.0000000

The first linear model can be written out as follows:

$$health = 2.7962017 + 0.0281222 \times income2 + 0.0376613 \times income3 + -4.654692 \times 10^{-4} \times income4 + -0.0500514 \times xpecies + -0.0041611 \times stewardsTor2 + 0.0170553 \times stewardsTor4 + 0.0529587 \times stewardsplus + 0.0055038 \times sidewalkdamage + -0.0622818 \times guardsharmful + -0.0666873 \times guardsunsure + 0.0167832 \times guardshelpful$$

The created linear model above predicts the health score (1 for poor, 2 for fair, 3 for good) for a tree based on a multitude of factors. The intercept is 2.7962017, which means that if everything is kept at baseline (species is London planetree, median income of the neighborhood is $0-54278, no stewardship for the tree, undamaged sidewalk, and no guards), the health score should be 2.7962017. There are also different coefficients for different income levels: if the income is $54279-70793, it is 0.0281222, if $70794-84713, it is 0.0376613, if $84713 or greater, it is -4.654692×10^{-4}. The type coefficient, -0.0500514, is split into two categories: London planetree (0) or other species (1), which means it would be in use when the tree is not a London planetree. The steward coefficients are different for the different number of stewardship shown on each tree, so we can see that for the more stewardship shown on each tree, the healthier it tends to be, as the coefficient for 4 or more stewards, 0.0529587, is higher than that of 2 and 1/2. Furthermore, guards seem to beone of the most important in determining tree health. Having harmful guards harms the health of the tree, but helpful guards increases the health by 0.0167832.

In our next regression model, we will be examining the effect of the raw income (in thousands of dollars) and the squared income to account for the potential curvilinear relationship.

Term	Estimate	STD Error	Statistic	P-Value
(Intercept)	2.7725362	0.0041135	674.010840	0.0000000
raw_rescaled	0.0012367	0.0000910	13.585607	0.0000000
type	-0.0523742	0.0018727	-27.967644	0.0000000
stewardsTor2	-0.0031204	0.0017094	-1.825458	0.0679324
stewardsTor4	0.0190330	0.0042236	4.506311	0.0000066
stewardsTor4orMore	0.0531505	0.0128482	4.136793	0.0000352
sidewalkDamage	0.0061696	0.0014072	4.384347	0.0000116
guardsHarmful	-0.0627000	0.0038794	-16.162381	0.0000000
guardsUnsure	-0.0654883	0.0059254	-11.052093	0.0000000
guardsHelpful	0.0167019	0.0027494	6.074766	0.0000000
Income_rescale	-0.0000081	0.0000005	-14.928835	0.0000000

The linear model can be written out as follows:

$$health = 2.7725362 + 0.0012367 \times income + -0.0523742 \times xpecies + -0.0031204 \times stewardsTor2 + 0.019033 \times stewardsTor4 + 0.0531505 \times stewardsplus + 0.0061696 \times sidewalkdamage + -0.0627 \times guardsharmful + -0.0654883 \times guardsunsure + 0.0167019 \times guardshelpful + -8.0723589 \times 10^{-6} \times incomesq$$

The second model is similar to the first except for two variables. First, the income brackets have been replaced with a continuous variable of raw income, `raw_rescaled`, which has been divided by 1000 to make the model simpler. This variable has the coefficient of 0.0012367. The second difference is the inclusion of the rescaled income variable squared to account for a curvilinear relationship. The variable is represented by `income_rescale`, and its coefficient is $-8.0723589 \times 10^{-6}$.

Results

From our results, we can conclude that there is a small but not negligible impact of income on the health of trees in NYC. From the third visualization showing the relationship between income bracket and tree health, we can see a small increase of Good health trees in the first tree brackets, but with a fall in the last one. Thus, we visually see that the income variable does have an impact on health (outcome variable).

Although visually the distribution of Poor, Fair, and Good health seem similar for each category, our summary statistics provide further evidence that there is a relationship between income and tree health. We calculated the difference in the means between the lowest income group and each of the higher income groups. We found that the second and third highest had an average tree health score of 0.0294661 and 0.0384117 higher than the lowest income group, but the highest income bracket had a mean tree health score of -0.0044817 lower. By simulating the null hypothesis and testing for independence, we found that there is an incredibly low probability that these differences are due to chance because the p-values were equal to 0, 0, and 0. While this demonstrates a statistically significant relationship between income and tree health, the difference in means between the highest and lowest categories calls this positive correlation into question because the low income areas have better average tree health than the wealthy areas.

With our regressions, we can understand the numerical impact of income on health level. With the first regression, `income_health`, we can see that when the income increases to the second bracket, the health score is 0.0281222 higher than that of the first bracket. When it increases to the third bracket, the score is 0.0376613 higher. But when the income bracket increases to the fourth bracket, the score is -4.654692×10^{-4} lower. After accounting for other variables like species, guards and stewardship, the regression suggests that there may be a positive linear relationship between income and health, even though the coefficient became negative at the fourth bracket. This made us wonder if there was a curvilinear relationship instead.

With our second regression, `income_rescale_health`, the income variable is continuous (we rescaled it to be divided by 1000 to make it simpler). We also included the square of the income to account for the possibility of it being a curvilinear relationship. In this model the coefficient for the income is 0.0012367 which shows that for every increase in $1000, the health increases by that value. The trends remain similar for the other variables with slight changes in values. For the square of the income which accounts for the curvilinear relationship, the variable has a coefficient of $-8.0723589 \times 10^{-6}$. The resulting model shows that for every increase in $1000 for the income there is an effect of $0.0012367 + -8.0723589 \times 10^{-6} \times income + -8.0723589 \times 10^{-6}$. Since $-8.0723589 \times 10^{-6}$ is very close to zero it would suggest that the model is mostly linear.

After comparing the R^2 value for the two regressions, we can see that the R^2 for the first regression, 0.0033846, is greater than that of the second one, 0.0025605. Unfortunately, since the R^2 value is so small, it means that only 0.32845533% of the variance in the data is predicted by our first model. Interestingly, most of the p-values we received from the regression are close to 0. Since our p-values are mostly very close to 0, especially those of the first three income brackets, this means that we should reject our null hypothesis that the income variable does not affect the health outcome.

However, outliers do exist: in our first regression, the variable for the highest income bracket and the variable for guards with unsure helpfulness are the only ones with p-values higher than 0.05. The p-value for the highest income bracket is 0.7976541, which is extremely high, and the coefficient for this bracket is -4.654692×10^{-4}, which breaks the trend of the coefficients increasing as the income bracket increases. This suggests that there may be several outlier neighborhoods in the income bracket that have very unhealthy trees. With the unreliable coefficient for this income bracket, it suggests there is a chance that the trend of health becoming better as income increases still stands.

Overall, we can not fully confirm our hypothesis because although there is an upwards trend for the most part, the highest income zip code regions consistently demonstrate that they have less healthy trees.

Discussion

From our analysis, we learned that our research question is somewhat supported by the difference in means and p-values calculated in our statistical arguments. There seems to be an increase, albeit slight, in the mean tree health for the first to the third income brackets in the regression which would suggest that the hypothesis where tree health gets better as the neighborhood's income increases is supported. The low p-values, the differences in mean, and the visualization of the relationship between the exploratory & outcome variables support this hypothesis as well. However, we come into problems when delving deeper into the values for each statistical model and what they mean.

A major flaw in our model is the inability of our regression to predict accurate outcomes. Our R^2 for our `income_health` regression is 0.0032846, which means only 0.32845533% of the variance in the data is explained by our first model. This is highly inaccurate and suggests that a linear regression is not the best choice for the modeling.

Both residual models from the two regressions have the same trends, and we will discuss them in general. The residuals models show a major flaw in our investigation: the dataset itself. Because the observations for health (the outcome variable) were separated into only three categories (Poor, Fair, and Good), the residual plot is also separated into three separate lines with similar negative slopes. The points in the residual plot fall into parallel slopes because as the fitted value increases, it becomes more different from the actual value in similar slopes for the trees with scores of 1 & 2 and closer for the trees of 3 points. For example, for the Good health trees, the higher the score gets (2.7 to 2.9), the closer the residual moves to 0, as the fitted value gets closer to the actual score 3. We can see a large flaw in our model in that the lowest fitted value goes to only 2.6, but the trees with Poor health have an actual score of 1 and trees with Fair health have an actual score of 2. This means that for all the trees whose actual values are 1, the model's predictions are all 1.7 to 1.9 points away. However, we could argue that this makes sense because we found the average health of trees in one area, and many more trees are observed having Good health compared to Fair and Poor, which raises the average score.

This data set itself has many flaws which make this analysis difficult. In order to investigate this question more effectively in the future, we should try to find data with a more detailed scale to measure tree health so that the outcome variable has more than 3 possible values. Even with the results we have, it is hard to say that "one income group tends to have trees with scores 0.0294661 higher" means anything when in reality these trees are only categorized as Poor, Fair, or Good in the original data, not with detailed numeric scores. Additionally, this data is questionable because apparent health of a tree is subjective and vague, especially considering much of this data is collected by volunteers rather than professionals, and there is a great possibility for human error in how this tree data is collected.

A major way we could improve on our models in a future investigation is to use an ordinal logistics regression. Even better, we could use a classification approach model which uses multiple predictors such as income, guards, stewardship, to predict the health of the tree. A classification model reads some input and generates an output that classifies the input into some category, which would be much more helpful because our outcome variable has only three categories. A linear regression is not a good way to predict the health of trees because there are so many factors involved. If we were to do the project over again, we would use these alternative methods. We would also try to have a clearer idea of what we were looking for and how we were planning to calculate it from the start.

In conclusion, although we found some statistically significant results, they are not practically significant. There are too many issues with the data set and the statistical processing to claim a positive relationship between the variables with any level of certainty. Once the mentioned improvements to our analysis are put in place, it would be interesting to rerun these tests. We could also analyze different major cities and see if the results are consistent.