Analyzing the 2015 NYC Tree Census

Dianna Sinicrope

8/7/2019

## Introduction

My original question was a broad question asking about the state of trees in NYC. In 2015, the NYC Parks and Recreation staff, the group TreesCount!, and volunteers conducted a tree census of all the street-side trees in NYC. They released their findings, and I was curious to explore more about the dataset. Things like the number of trees in each borough, specific species, and whether or not more trees existed in more affluent neighborhoods piqued my interest from the beginning, but I purposely didn’t dive into the data with any solid hypothesis in mind.

An exploratory data analysis was first needed to get a good grasp of what the data was and how it was setup. I found 41 variables in the dataset that address things like health, location, and size of each tree. From there I wanted to determine where trees are located, if trees are in better health in individual boroughs, and if the median income of an area correlates with either the number of trees or the health of trees in that area.

## Initial Findings

Throughout the EDA process, I discovered that there are three categories of tree status: alive, dead, and stump. Knowing these categories proved to be important since the answers to my original research questions were much more focused on alive trees, with less focus on dead trees and stumps.

## # A tibble: 3 x 2  
## status Count\_status  
## <chr> <int>  
## 1 Alive 652173  
## 2 Dead 13961  
## 3 Stump 17654

## [1] 652173

There were 652173 alive trees in NYC in 2015.

I also looked at the number of dead trees per borough, as well as the total count of dead trees in NYC.

## # A tibble: 5 x 2  
## boroname Count\_trees  
## <chr> <int>  
## 1 Queens 4440  
## 2 Brooklyn 3319  
## 3 Bronx 2530  
## 4 Staten Island 1870  
## 5 Manhattan 1802

## [1] 13961

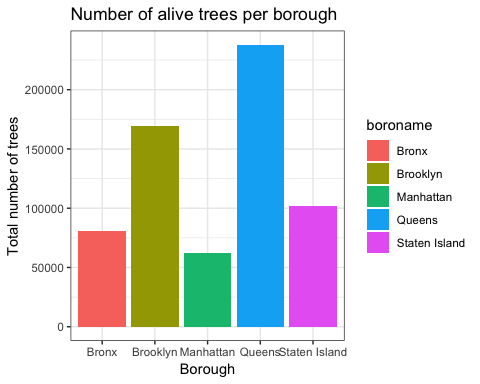
## [1] 683788

There were 13961 alive trees in NYC in 2015.

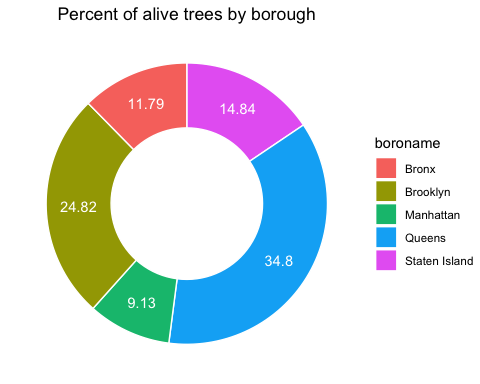
Here’s the full list of alive and dead trees as well as stumps in NYC, per borough:

## boroughs alive\_trees dead\_trees stump\_trees total  
## 1 Queens 237974 4440 8137 250551  
## 2 Brooklyn 169744 3319 4230 177293  
## 3 Staten Island 101443 1870 2005 105318  
## 4 Bronx 80585 2530 2088 85203  
## 5 Manhattan 62427 1802 1194 65423

I filtered my data for alive trees and was able to represent the alive tree per borough in two plots: a bar plot of the number of alive trees per borough and a donut chart of the percentage of alive trees in each borough. The alive trees per borough plot is particularly interesting since we can immediately see that Queens has the largest number of trees (238K) followed by Brooklyn with 170K trees. Manhattan has the smallest number of trees with 62K. Seeing the data in this way made me wonder about whether the square mileage of a borough correlates with the number of trees it has, and I’ve addressed this later in this analysis.



## # A tibble: 5 x 3  
## boroname Count\_trees percent\_of\_total\_trees  
## <chr> <int> <dbl>  
## 1 Queens 237974 34.8   
## 2 Brooklyn 169744 24.8   
## 3 Staten Island 101443 14.8   
## 4 Bronx 80585 11.8   
## 5 Manhattan 62427 9.13



## Central Tendency Statistics

I also wanted to look at central tendency statistics. On average, each borough has 130K trees. Queens and Brooklyn are the only two boroughs above this average while Manhattan, Staten Island, and the Bronx are below.

## [1] 130434.6

## Tree Species

I was also interested in seeing the most popular trees species in NYC. Out of the 133 tree species in NYC, the most popular trees species is the London planetree (with 12.7% of all trees being London planetrees), followed by the honeylocust (9.4%).

## # A tibble: 6 x 3  
## spc\_common Sum\_species percent\_of\_total\_species  
## <chr> <int> <dbl>  
## 1 London planetree 87014 12.7   
## 2 honeylocust 64263 9.4   
## 3 Callery pear 58931 8.62  
## 4 pin oak 53185 7.78  
## 5 Norway maple 34189 5   
## 6 littleleaf linden 29742 4.35

## # A tibble: 6 x 3  
## spc\_common Sum\_species percent\_of\_total\_species  
## <chr> <int> <dbl>  
## 1 black pine 37 0.01  
## 2 pitch pine 33 0   
## 3 Osage-orange 29 0   
## 4 Scots pine 25 0   
## 5 Virginia pine 10 0   
## 6 "" 5 0

Looking at each borough, I can see that in Queens (below), the most popular tree is also the London planetree, but the second most popular is the pin oak.

## # A tibble: 6 x 2  
## spc\_common Sum\_species  
## <chr> <int>  
## 1 London planetree 31111  
## 2 pin oak 22610  
## 3 honeylocust 20290  
## 4 Norway maple 19407  
## 5 Callery pear 16547  
## 6 cherry 13497

For Manhattan (below), the borough with the smallest number of trees, the most popular species was the honeylocust, followed by the Callery pear. This kind of analysis (particularly when paired with tree health statistics, which I address later in this report) could be important for botanists who are interested in figuring out tree diseases or the best environments for certain trees to thrive in.

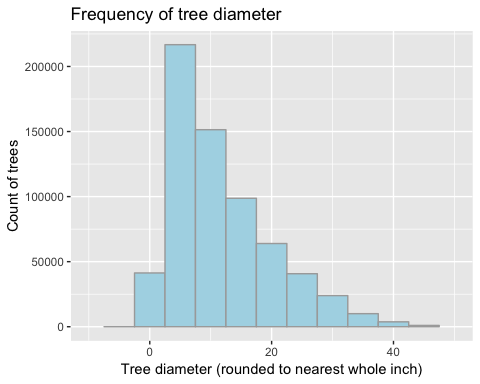
## # A tibble: 6 x 2  
## spc\_common Sum\_species  
## <chr> <int>  
## 1 honeylocust 13175  
## 2 Callery pear 7297  
## 3 ginkgo 5859  
## 4 pin oak 4584  
## 5 Sophora 4453  
## 6 London planetree 4122

## Histogram for Tree Diameter

Tree diameter is also available as a variable in the dataset. While tree diameter in itself isn’t a great indicator of exact height or age of a tree (since I’ll assume height and age are dependent on more variables than just tree diameter), I might be able to uncover some interesting questiong after analyzing it. The histogram that I’ve created shows that the majority of trees have smaller diameters. The data is skewed right with a long tail (i.e. a positive skew).

## Warning: Removed 361 rows containing non-finite values (stat\_bin).

## Warning: Removed 2 rows containing missing values (geom\_bar).



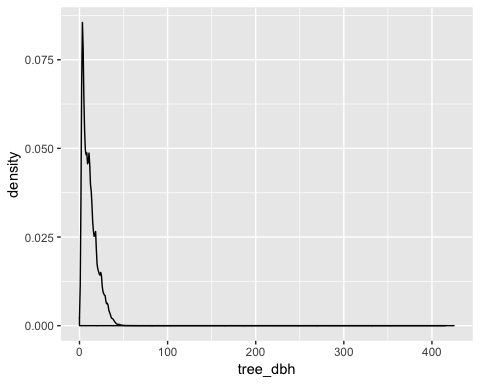
The mode, median, and mean are as follows:  
Mode: 4 in.  
Median: 10 in.  
Mean: 11.7 in.  
This increasing order of mode, median, and mean is what we would expect from a positive skew, and as such, it’s a great way to validate my visual analysis of the histogram.

## [1] 4

## [1] 10

## [1] 11.70949

I’ve also created a density plot and plotted a density plot over the histogram to see the distribution shape of the data and to make sure that I’ve used an appropriate number of bins for my histogram. The density plot also shows the same right skew I mention above.

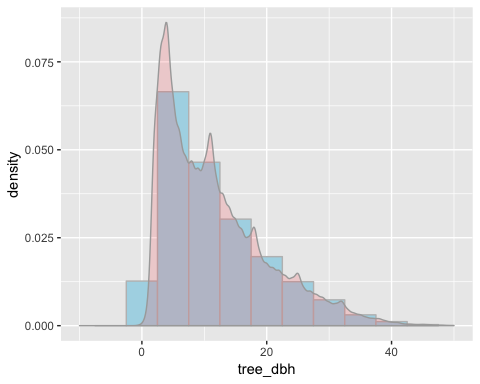


Overlapping the density plot with the histogram is a good way to make sure that I got a good estimate for the number of bins in the histogram.

## Warning: Removed 361 rows containing non-finite values (stat\_bin).

## Warning: Removed 361 rows containing non-finite values (stat\_density).

## Warning: Removed 2 rows containing missing values (geom\_bar).



## Health of trees

In general, trees in NYC are between fair and good health, leaning more toward good health. One a 0 to 2 scale (with 0 being poor health and 2 being good health), trees in NYC are around the 1.76 mark. Overall though, trees in the Bronx are the healthiest (1.79 on the health scale) and trees in Manhattan are the least healthy (1.70 on the health scale). It would be interesting to pair this data perhaps with pollution data or soil health to see if there’s a correlation or suspected cause.

## health\_avg  
## 1 1.769785

## # A tibble: 5 x 2  
## boroname health\_avg  
## <chr> <dbl>  
## 1 Bronx 1.79  
## 2 Brooklyn 1.78  
## 3 Queens 1.78  
## 4 Staten Island 1.77  
## 5 Manhattan 1.70

I found it particularly interesting when I started looking at the health of specific species of trees (below). The pitch pine and the arborvitae are in nearly perfect health at the top of the list ranking 2.0 and 1.9 respectively on the health scale. But the bottom of the list for the least healthy trees are what surprised me the most. The Norway maple, which I had ranked as the 5th most popular tree in NYC, is also one of the least healthy trees in NYC. With 34K trees citywide, the Norway maple ranked a sad 1.5 on the health scale, meaning that most trees were either ranked as poor or fair health. This low number led me to hypothesize that there might be something going on with Norway maples. It could be a range of issues from disease to age (it could have been more popular to plant them years ago, and now they are all coming to the end of their natural, expected lifespans) or people could be cutting down a large area of them to make room for buildings. The next step would be to narrow down on the boroughs that have the most Norway maples since the cause of their below-average health could be based on location.

## # A tibble: 6 x 3  
## spc\_common health\_avg count\_trees  
## <chr> <dbl> <int>  
## 1 pitch pine 2 33  
## 2 arborvitae 1.91 328  
## 3 Osage-orange 1.90 29  
## 4 red pine 1.89 106  
## 5 false cypress 1.88 108  
## 6 Atlas cedar 1.87 87

## # A tibble: 6 x 3  
## spc\_common health\_avg count\_trees  
## <chr> <dbl> <int>  
## 1 eastern hemlock 1.57 88  
## 2 Amur cork tree 1.53 183  
## 3 katsura tree 1.52 911  
## 4 Norway maple 1.51 34189  
## 5 pond cypress 1.49 181  
## 6 black pine 1.41 37

When I segment for Norway maples in specific boroughs, I see that these trees in Manhattan, Queens, and the Bronx are the least healthy types of Norway maples. I also paired the data with average diameter, thinking that perhaps bigger (and generally, older) trees were in poorer health, but the average diameter of the Norway maples doesn’t seem to affect the average health within each borough.

I think more research could be fascinating. A quick Google search shows websites like, “Norway Maple – New York Invasive Species Information” and, “The Norway Maple: New York’s Ultimate Weed | Tarrytown, NY Patch” so it’s especially interesting that the fifth most popular tree in NYC has such a bad reputation and is in such poor health.

## # A tibble: 5 x 4  
## boroname health\_avg count\_trees diameter\_avg  
## <chr> <dbl> <int> <dbl>  
## 1 Staten Island 1.70 4127 11.9  
## 2 Bronx 1.62 3376 13.5  
## 3 Brooklyn 1.48 6989 14.6  
## 4 Queens 1.47 19407 15.0  
## 5 Manhattan 1.40 290 10.2

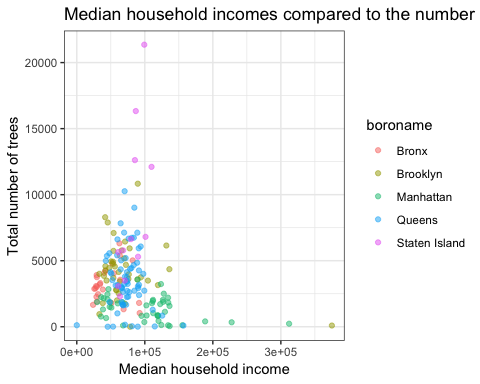
## Correlation with income

I also joined the tree census data with median incomes per zip code. I was expecting more than what I ended up seeing here, though. I created a scatterplot and a correlation matrix, but there was barely any correlation between the number of trees in a zip code and that zip code’s median household income.

## # A tibble: 6 x 4  
## # Groups: boroname, zip\_city [3]  
## boroname zip\_city income2 count\_trees  
## <chr> <chr> <int> <int>  
## 1 Brooklyn Brooklyn 375000 92  
## 2 Manhattan New York 312118 230  
## 3 Manhattan New York 227647 338  
## 4 Manhattan New York 188682 405  
## 5 Queens Long Island City 156769 90  
## 6 Manhattan New York 155312 117

At first glance, I don’t see a correlation with income and count of trees, but I’ve also created a scatter plot to check visually.

## Warning: Removed 3 rows containing missing values (geom\_point).



I don’t see a correlation here between the variables, but I’ll create a correlation matrix just to make sure.

## income2 count\_trees  
## income2 1.00 -0.16  
## count\_trees -0.16 1.00

This report also shows that the two variables barely share any correlation.

## Correlation with square mileage

Since the income of a zip code didn’t have a strong correlation with the number of trees, I decided to look at whether a link existed between a borough’s square mileage and its number of trees.

There are more trees per square mile in Manhattan than any other borough (2,838 compared to the second ranked Brooklyn with 2,424 trees per square mile), so I decided that maybe my zip code income data was too granular and not ideal since it doesn’t have square mileage associated with it. Instead of using median household income by zip code, I found median household income by borough on Wikipedia and compared that with trees per square mile for each borough. It turns out that there is a mild correlation between these two variables.

## # A tibble: 5 x 4  
## # Groups: boroname [5]  
## boroname squaremiles count\_trees trees\_per\_sqmile  
## <chr> <int> <int> <dbl>  
## 1 Manhattan 22 62427 2838.  
## 2 Brooklyn 70 169744 2425.  
## 3 Queens 108 237974 2203.  
## 4 Bronx 42 80585 1919.  
## 5 Staten Island 58 101443 1749.

## Classes 'grouped\_df', 'tbl\_df', 'tbl' and 'data.frame': 5 obs. of 5 variables:  
## $ boroname : chr "Bronx" "Brooklyn" "Manhattan" "Queens" ...  
## $ squaremiles : int 42 70 22 108 58  
## $ count\_trees : int 80585 169744 62427 237974 101443  
## $ trees\_per\_sqmile: num 1919 2425 2838 2203 1749  
## $ hhmedian : int 36593 52782 79781 62008 76244  
## - attr(\*, "groups")=Classes 'tbl\_df', 'tbl' and 'data.frame': 5 obs. of 2 variables:  
## ..$ boroname: chr "Bronx" "Brooklyn" "Manhattan" "Queens" ...  
## ..$ .rows :List of 5  
## .. ..$ : int 1  
## .. ..$ : int 2  
## .. ..$ : int 3  
## .. ..$ : int 4  
## .. ..$ : int 5  
## ..- attr(\*, ".drop")= logi TRUE

## # A tibble: 5 x 5  
## # Groups: boroname [5]  
## boroname squaremiles count\_trees trees\_per\_sqmile hhmedian  
## <chr> <int> <int> <dbl> <int>  
## 1 Manhattan 22 62427 2838. 79781  
## 2 Staten Island 58 101443 1749. 76244  
## 3 Queens 108 237974 2203. 62008  
## 4 Brooklyn 70 169744 2425. 52782  
## 5 Bronx 42 80585 1919. 36593

Here’s the correlation matrix showing that the variables are mildly correlated.

## trees\_per\_sqmile hhmedian  
## trees\_per\_sqmile 1.00 0.33  
## hhmedian 0.33 1.00

## Implications

My initial EDA of this dataset is a good start to uncovering the potential of the data within. My data could be used by environmentalists, botanists, and dendrologists as either a sample of trees on the East Coast or as a study of trees in NYC. They could use the information to determine the average diameter of trees and how certain types of trees grow in certain conditions. Even having the location of each stump could prove useful for a person interested in the history of trees. Stumps provide the ability to count the number of rings/ years a tree was alive, and we could combine the years a tree’s age with its diameter in order to get a good estimate of how old any particular tree is. I doubt that it would be exact, but predicting their ages based on other variables could be done.

This tree census could also have implications for homeowners and the government (or anyone else interested in planting trees in NYC). Species of trees could be chosen for the best health, and I would caution people against planting the Norway maple. Or, if I owned a tree service company, I’d absolutely offer my services to any area with a high percentage of trees in poor health.

## Limitations

A few limitations of my analysis should be addressed. First, my datasets are from different years. The tree census is from 2015, the income by zip code is for the first part of 2019, the Wikipedia information for square mileage is from 2018, and the U.S Census Bureau’s income by borough dataset was an average of 2013-2017, in 2017 dollars. Ideally, I would have all datasets from the same year, but I don’t think my initial findings in this report would be considerably different if I had had better access to updated data. Second, my own lack of knowledge for trees from a scientific standpoint should be pointed out. For a more in-depth project, it would be helpful to work with a botanist or a dendrologist who might be able to better identify key datapoints about the trees dataset.

## References

2015 Street Tree Census - Tree Data: NYC Open Data. (2015). Retrieved from <https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh>

2015 Street Tree Census- Tree Data Dictionary. (2015). Retrieved from <https://data.cityofnewyork.us/api/views/ncgg-4chj/files/f963ffec-5cbc-45c4-b2bd-bab2d3c53f1b?download=true&filename=TreesCount2015> Street Tree Census - Tree Data Dictionary.pdf

Bedell, D. (2011, November 15). The Norway Maple: New York’s Ultimate Weed. Retrieved August 7, 2019, from <https://patch.com/new-york/tarrytown/the-norway-maple-new-york-s-ultimate-weed>

Boroughs of New York City. (2019, August 04). Retrieved August 7, 2019, from <https://en.wikipedia.org/wiki/Boroughs_of_New_York_City>

Median HH Income by zip for NJ and NYC (2019)[XLSX]. (2019, August 2). Claritas, LLC.

New York Invasive Species (IS) Information. (2019, July 2). Retrieved from <http://nyis.info/invasive_species/norway-maple/>

U.S. Census Bureau QuickFacts: New York County (Manhattan Borough), New York; Bronx County (Bronx Borough), New York; Queens County (Queens Borough), New York; Kings County (Brooklyn Borough), New York; Richmond County (Staten Island Borough), New York; New York city, New York. (n.d.). Retrieved August 7, 2019, from <https://www.census.gov/quickfacts/fact/table/newyorkcountymanhattanboroughnewyork,bronxcountybronxboroughnewyork,queenscountyqueensboroughnewyork,kingscountybrooklynboroughnewyork,richmondcountystatenislandboroughnewyork,newyorkcitynewyork/HSG010218>