Project Objective and Background

The purpose of this project is to answer the question, how do we predict the health of a tree in New York City? What factors contribute to the deterioration of a tree's health? Every year, NYC Parks & Rec asks volunteers to go around the city and document the living conditions of the city's trees to create a tree census.

In a city like NYC, trees provide a plethora of benefits as well as adding aesthetic value to the city's parks and streets. Trees help reduce carbon emissions, improve air quality, provide shade and lower air temperatures, reduce stormwater runoffs, and increase home values. Trees cover approximately a quarter of NYC and the city heavily invests in maintaining the well-being of these trees. However, over time, due to negligence and a variety of factors, the health of a tree can decline and eventually, is reduced to a stump (leaving a dead tree up is a dangerous as it becomes susceptible to lightning and a fire hazard).

Since the city spends millions every year on keeping the city green, it's important to have a way of monitoring the trees so that more and more trees are kept in good condition.

This project aims to create a model that can predict whether or not a tree is in need of attention due to health concerns. In order to create this model, we are using the 2015 Street Tree Census from NYC OpenData.

Data Wrangling

The dataset contains data on all 683,788 trees in NYC, living or otherwise. Each row represented a tree planted in the city and each column described an attribute of the tree. We start the data wrangling process by displaying a general overview of the dataset, looking at sample data rows, and listing all the column names. The columns are listed here along with a description and their data type.

Column Name	Description	Data Type
tree_id	Unique identification number for each tree point.	Number
tree_dbh	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	Number
curb_loc	Location of tree bed in relationship to the curb; trees are either along the curb (OnCurb) or offset from the curb (OffsetFromCurb)	String
health	Indicates the user's perception of tree health.	String
spc_common	Common name for species, e.g. "red maple"	String
stewards	Indicates the number of unique signs of stewardship observed for this tree. Not recorded for stumps or dead trees.	String
guards	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.	Text
sidewalk	Indicates whether one of the sidewalk flags immediately adjacent to the tree was damaged, cracked, or lifted. Not recorded for dead trees and stumps.	Text
problems		Text
root_stone	Indicates the presence of a root problem caused by paving stones in tree bed	Text
root_grate	Indicates the presence of a root problem caused by metal grates in tree bed	Text
root_other	Indicates the presence of other root problems	Text
trunk_wire	Indicates the presence of a trunk problem caused by wires or rope wrapped around the trunk	Text

trnk_light	Indicates the presence of a trunk problem caused by lighting installed on the tree	Text
trnk_other	Indicates the presence of other trunk problems	Text
brch_light	Indicates the presence of a branch problem caused by lights (usually string lights) or wires in the branches	Text
brch_shoe	Indicates the presence of a branch problem caused by sneakers in the branches	Text
brch_other	Indicates the presence of other branch problems	Text
borough	Name of borough in which tree point is located	Text
latitude	Latitude of point, in decimal degrees	Number
longitude	Longitude of point, in decimal degrees	Number

Keeping relevant columns

Since our purpose is to determine the health of trees, we removed any columns that did not aid us in creating our model. The columns removed are listed at the end of this report. Before removing the irrelevant columns, we checked for duplicates and found each row as unique, meaning no duplicates.

Missing values in rows

Next, we looked for any rows with missing values in any columns and found ~31,600 rows with data missing from the *health*, *spc_common*, *steward*, *guards*, *sidewalk*, and *problems* columns. Looking into these columns, we noticed 31,616 rows do not have any *health* data listed while 31,615 rows listed its trees as *Dead* or *Stump* under *status*.

Removing remaining rows with missing data

After we removed all rows where the *health* column was empty, the number of rows with missing data decreased significantly. Only *spc_common*, *guards*, *sidewalk*, and *problems* still contained a few rows with missing data. Meanwhile, we located and removed a single row that was classified as *Alive* under *status* but missing *health* data. Since our dataset is still quite large aftering removing data, we decided to drop all rows containing missing data since there is no way to guess what their columns' content.

Dropping remaining irrelevant columns

We dropped the *stump_diam* and *status* columns since all of our trees are living and stumps only apply to dead trees. The final count for our dataset at this point is 652,118 rows and 21 columns. There is still a lot of data left and our final step is to look at the distribution of tree diameters, or *tree_dbh*.

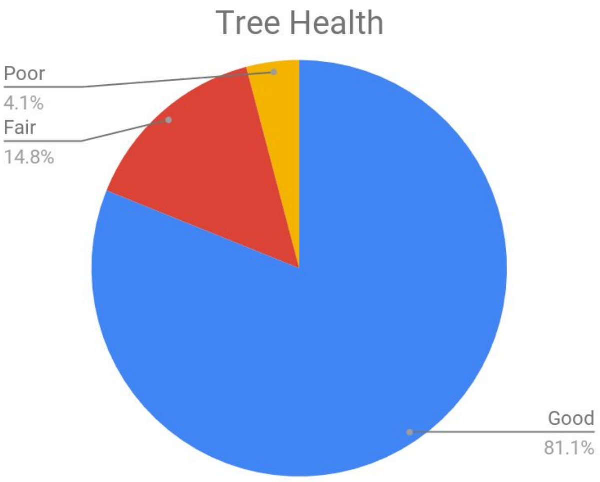
Addressing distribution of tree diameters

The diameters are measured in centimeters and their distribution overall is very skewed to the right, which means that the majority of trees have a smaller circumference and there are outliers in the data. Looking at the distribution of diameters, the range is from 0 to 425 inches. Right away, we can remove any trees with 0 as their diameter and 425 is too extreme of a value. Just 67 trees have diameters over 100 inches, and we removed them from our dataset since their impact is minimal overall. Our remaining trees all have diameters less than 100 inches. Even then, upon closer inspection, at the 75th percentile, the diameter of a tree is 16 inches. But since there are some trees whose diameters get quite large as they grow older, we decided to keep the maximum at 100 in order to include those trees.

After cleaning our dataset, we saved the remaining data in a new file for visualization and storytelling.

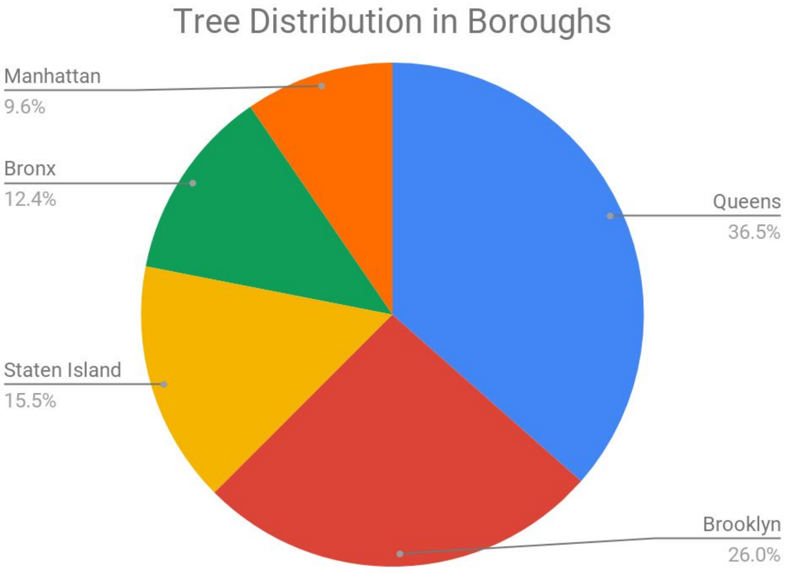
Data Storytelling and Visualization

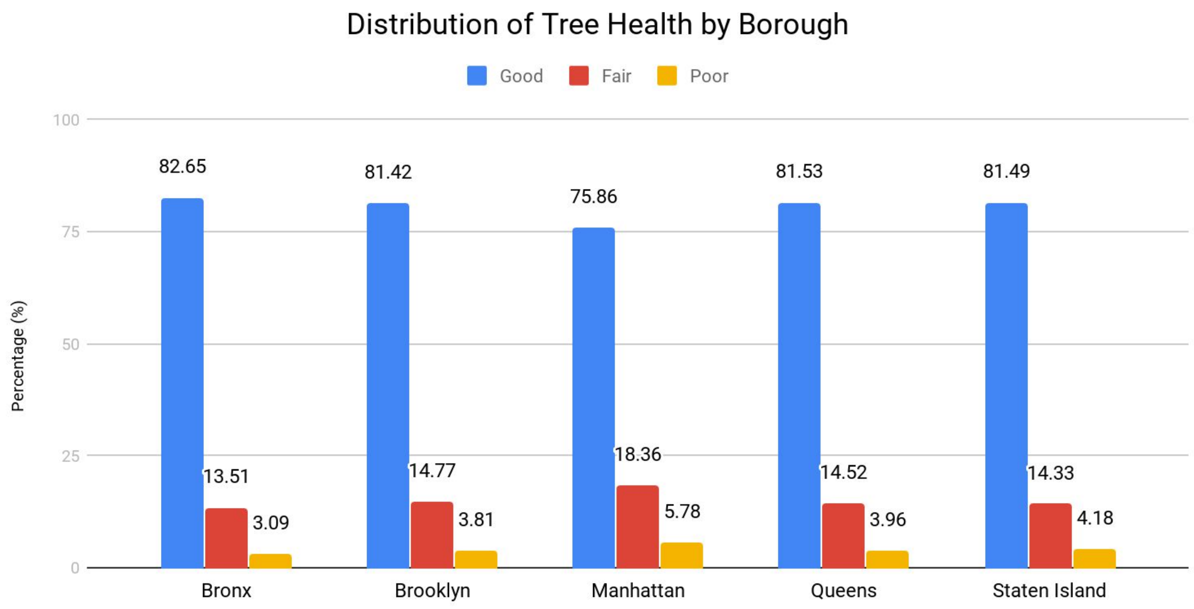
Our goal here is to identify any trends, patterns, and anomalies in the data using visualizations. We start by verifying that the data is clean by looking at a general overview and some sample rows. The column that we are interested in building our model around is the health column, and the majority of trees are in good health.



Boroughs

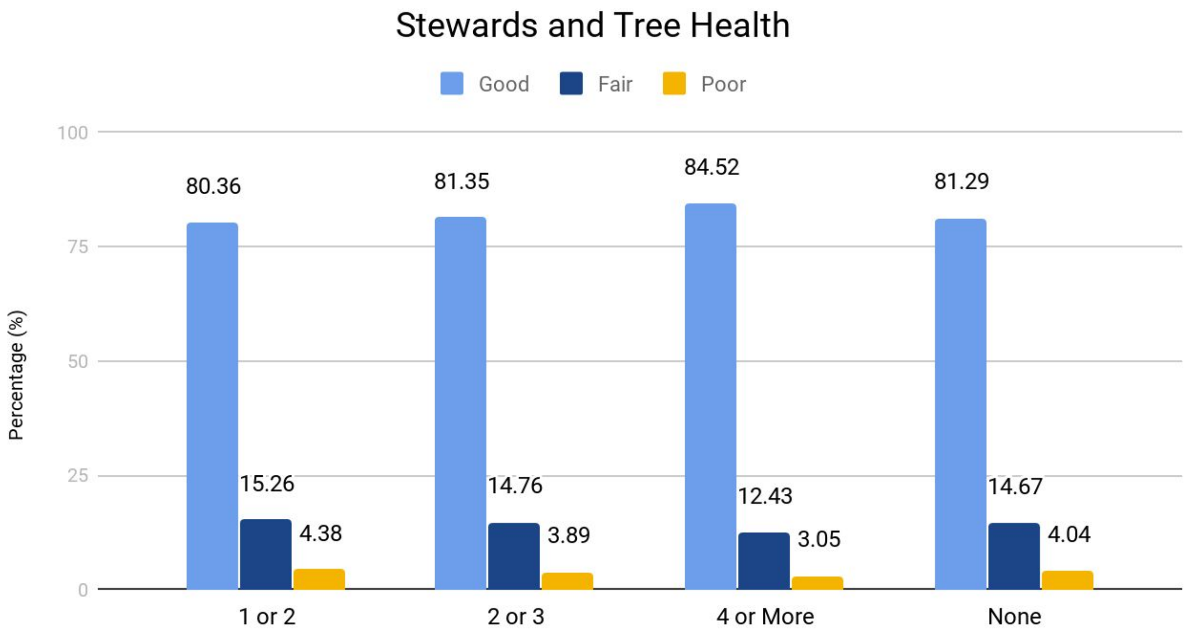
We looked at which borough the trees are in and their health statuses there. From the pie chart, the majority of trees are in Queens, followed by Brooklyn, Staten Island, the Bronx, and Manhattan.

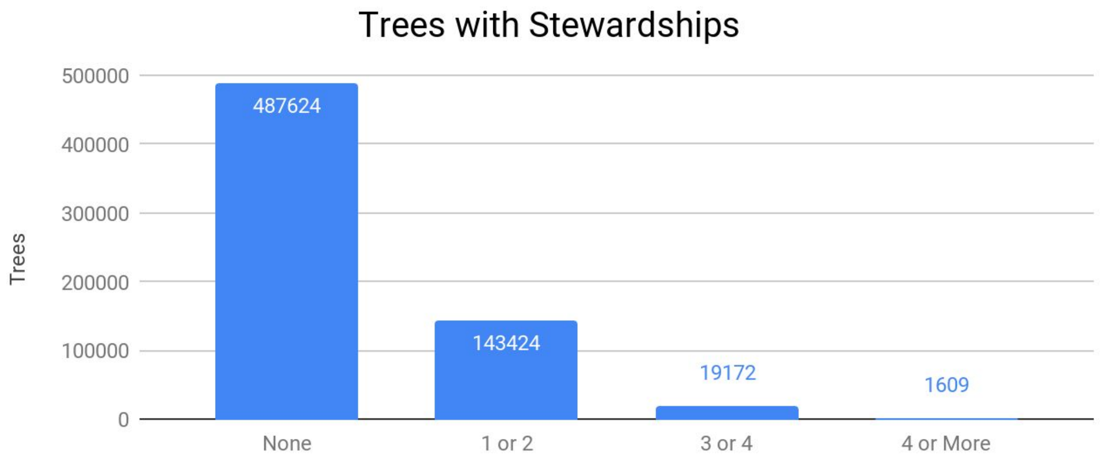




The consensus shows that the majority of trees in each borough are in good health, with the Bronx having the highest percentage of good, healthy trees. The borough with the lowest percent of good trees and highest percent of poor and fair trees is Manhattan.

Stewardships

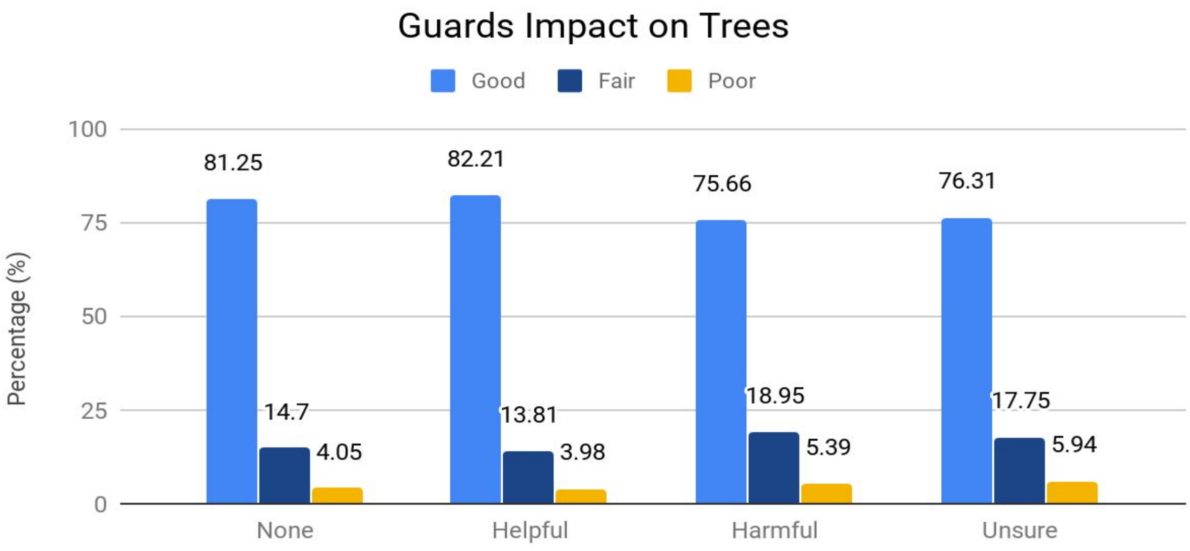




A steward is an individual assigned to care for the tree and monitor for any signs of distress. The majority of trees do not have a steward, yet the highest number of poor trees belong to trees with 1 or 2 stewards. It is worth noting that trees with 4 or more stewards have the highest percent of good trees and the lowest percentages of fair and poor trees. Only 1,609 trees have 4 or more stewards and 19,172 trees have 3 or 4 stewards, which means that the vast majority of trees are not assigned to any designated individual for monitoring.

Guards

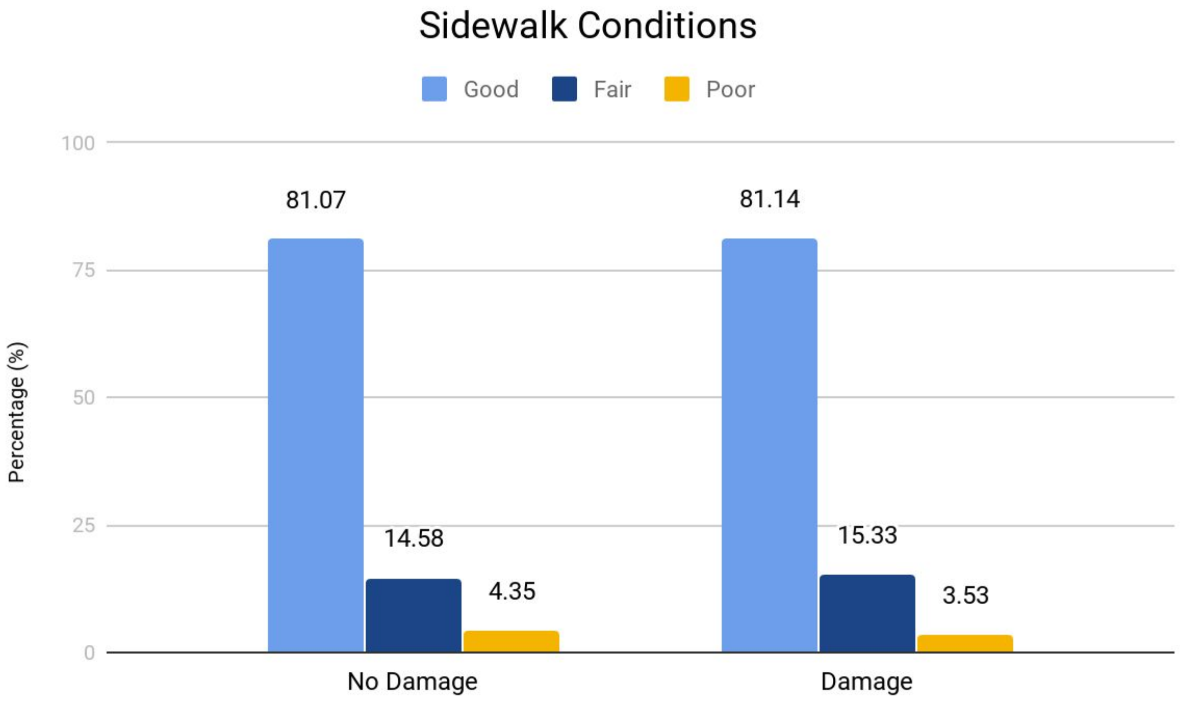
Tree guards are fences built around a tree and its soil, establishing a perimeter. They protect the tree and its contents from urban disruptors such as animal waste and reduce soil compaction.



Sidewalks

Trees are often planted adjacent to sidewalks and the condition of the sidewalk is recorded as either ‘Damaged’, or ‘No Damage’. The majority of sidewalks are not damaged in the city, a sign of good upkeeping.

Sidewalk Condition	Number of Trees	Percentage
No Damage	464,699	71.29%
Damaged	187,130	28.71%

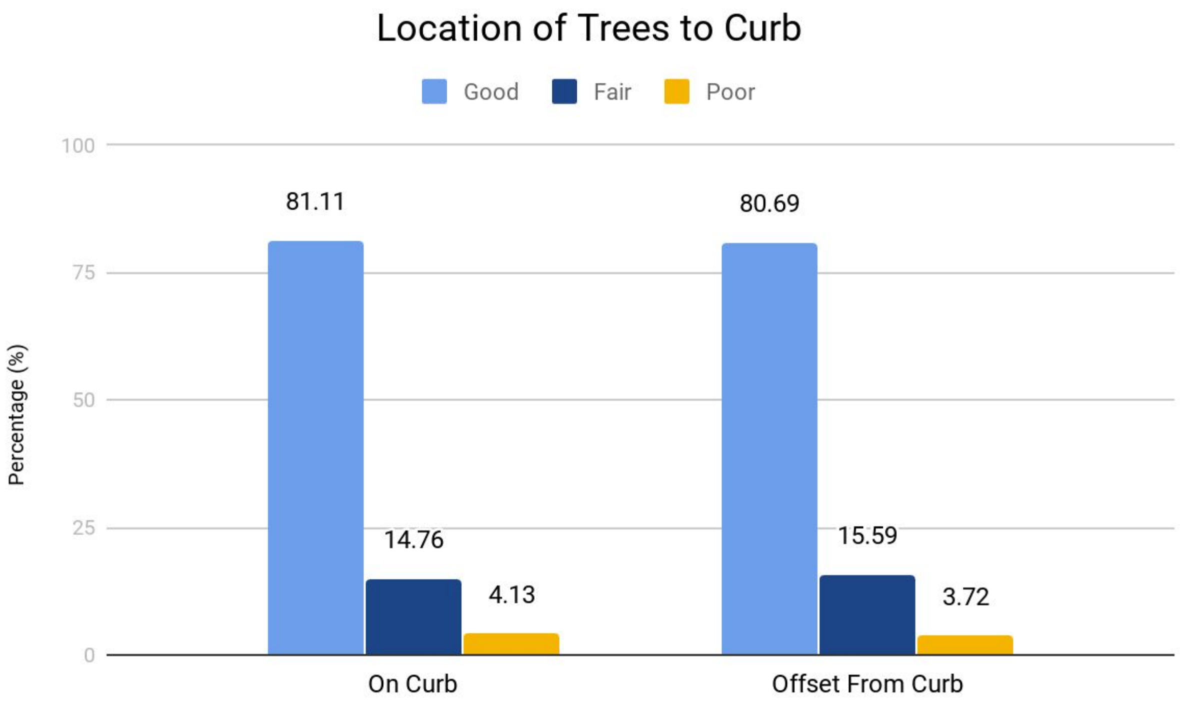


Sidewalk condition does not appear to have a large impact on tree health since the percentages of good, fair, and poor trees are roughly identical. Damaged sidewalks have slightly less poor trees, >1% difference, and slightly more good trees, with the same difference.

Curb Location

Tree beds are either along the curb (on curb) or offset from the curb. The vast majority of trees are on curb while only a small number of trees are offset.

Location	Number of Trees	Percentage
On Curb	625,973	96.03%
Offset From Curb	25,856	3.97%

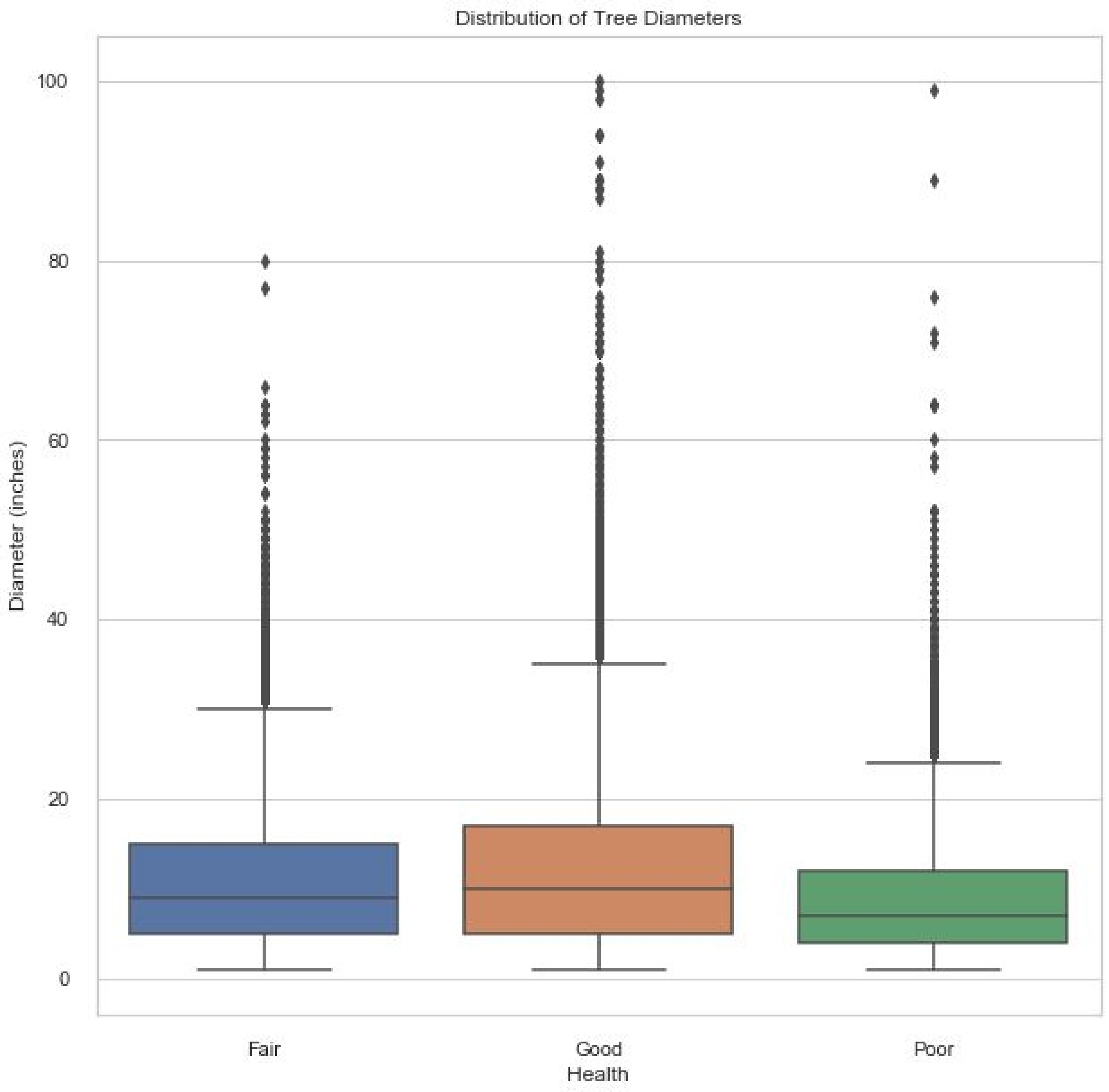


The tree's location to the curb and its health is again very similar, despite the huge difference in the number of trees being on curb and offset from the curb.

Tree Diameter Distribution

Good trees have larger diameters when comparing their averages and medians. The difference between good and fair tree diameters, average and median, is about one inch and the difference between fair and poor trees is almost two inches. There is evidence here that healthier trees have wider diameters than their counterparts.

Tree Health	Average Diameter (inches)	Median Diameter (inches)	Number of Trees
Good	11.97	10	528,582
Fair	10.97	9	96,451
Poor	9.02	7	26,796



Tree Species

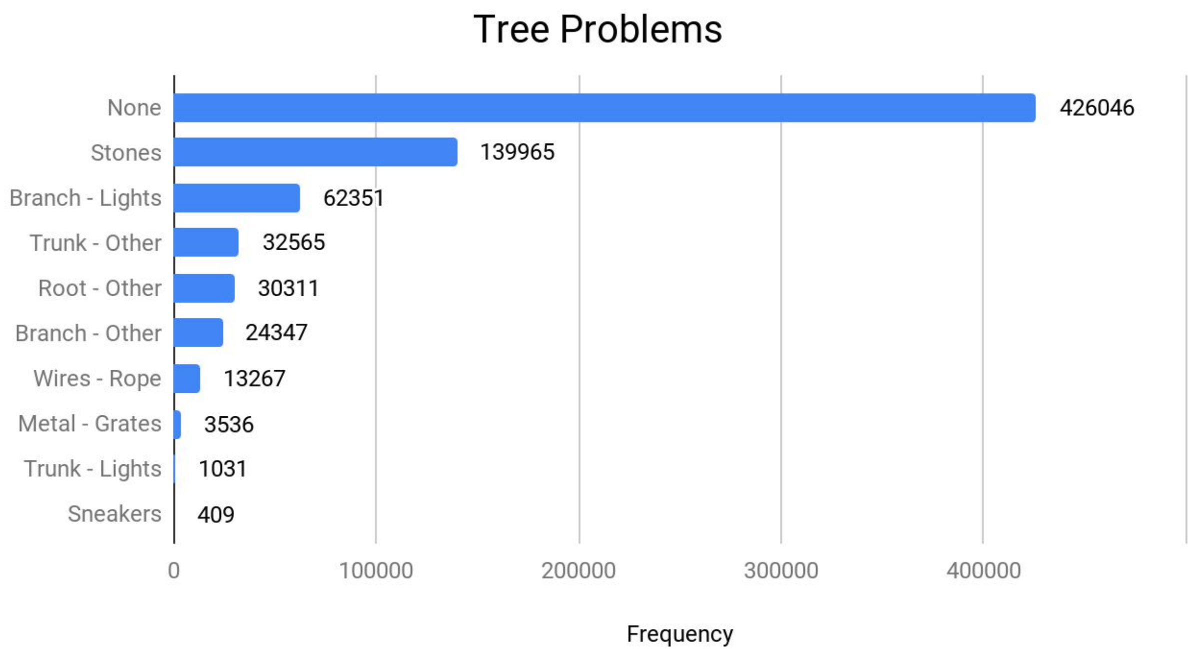
There are 132 unique species of trees grown in the city. London planetree is the most popular tree in NYC, followed by honeylocust.

Tree Species	Count
London planetree	86,997
Honeylocust	64,246
Callery pear	58,898

Pin oak	53,167
Norway maple	34,179
Littleleaf linden	29,733
Japanese zelkova	29,251
Cherry	29,248
Gingko	21,012
Sophora	19,332

Tree Problems

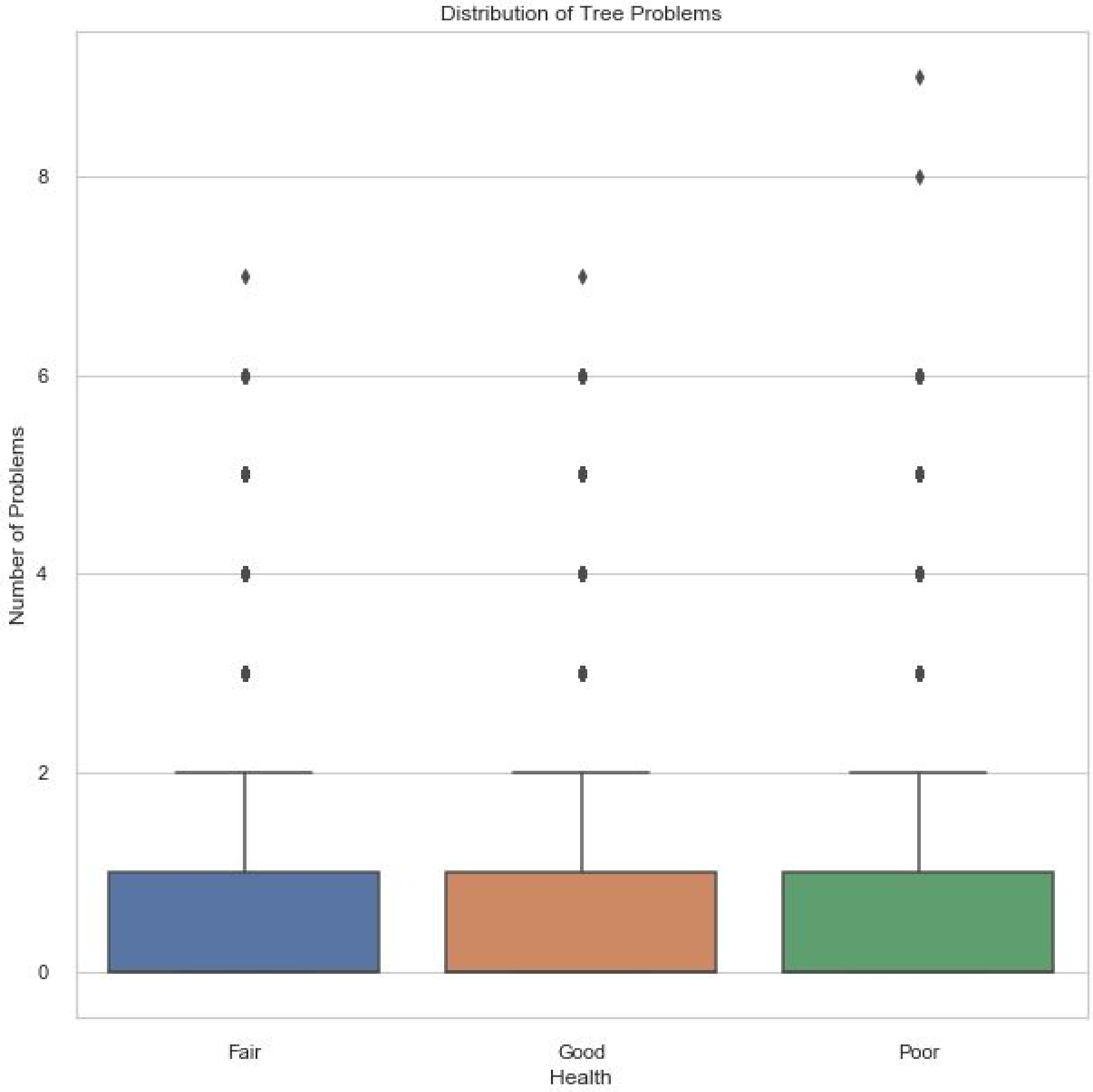
As trees get older, they may become damaged by natural elements, development, other trees, and human contact. It’s important to look into these problems as they very likely can impact the health of a tree. The majority of trees do not have any problems since they are mostly in good health, but for the trees that do, they can have up to nine defined problems. The graph below shows how times a problem / none has been recorded. Some trees may have multiple afflictions, like Sneaker and Branch - Lights, and each one is recorded in their category.



Impact of Problems on Tree Health

After identifying the potential problems that trees can have, we now look into whether or not the health of a tree is related to the number of problems that tree has.

Health	Count	Mean	Min	25%	50%	75%	Max
Good	528,582	0.43	0	0	0	1	7
Fair	96,451	0.64	0	0	0	1	7
Poor	26,796	0.68	0	0	0	1	9



Since the majority of trees do not have any problems, we removed them and counted the remaining trees.

Health	Count	Mean	Min	25%	50%	75%	Max
Good	171,955	1.32	1	1	1	2	7
Fair	42,527	1.46	1	1	1	2	7
Poor	11,301	1.62	1	1	1	2	9

It looks like tree health is linked to the number of problems the tree has. Looking at the mean, the number increases as the health of the tree decreases. As a result, we created a new column called 'num_problems' that tells us how many problems the tree has and inserted it into the dataset.

This table shows the removed columns and their description.

Column Name	Description	Data Type
block_id	Identifier linking each tree to the block in the blockface table/shapefile that it is mapped on.	Number
created_at	The date tree points were collected in the census software.	Date & Time
stump_diam	Diameter of stump measured through the center, rounded to the nearest inch.	Number
status	Indicates whether the tree is alive, standing dead, or a stump.	Text
spc_latin	Scientific name for species, e.g. "Acer rubrum"	Text
user_type	This field describes the category of user who collected this tree point's data.	Text
address	Nearest estimated address to tree	Text
postcode	Five-digit zip code in which tree is located	Number
zip_city	City as derived from zip code. This is often (but not always) the same as borough.	Text
community board	Community board in which tree point is located	Number
borocode	Code for borough in which tree point is located: 1 (Manhattan), 2 (Bronx), 3 (Brooklyn), 4 (Queens), 5 (Staten	Number

	Island)	
cncldist	Council district in which tree point is located	Number
st_assem	State Assembly District in which tree point is located	Number
st_senate	State Senate District in which tree point is located	Number
nta	This is the NTA Code corresponding to the neighborhood tabulation area from the 2010 US Census that the tree point falls into.	Text
nta_name	This is the NTA name corresponding to the neighborhood tabulation area from the 2010 US Census that the tree point falls into.	Text
boro_ct	This is the boro_ct identifier for the census tract that the tree point falls into.	Number
state	All features given value 'New York'	Text
x_sp	X coordinate, in state plane. Units are feet.	Number
y_sp	Y coordinate, in state plane. Units are feet.	Number
council district		Number
census tract		Number
bin		Number
bbl		Number