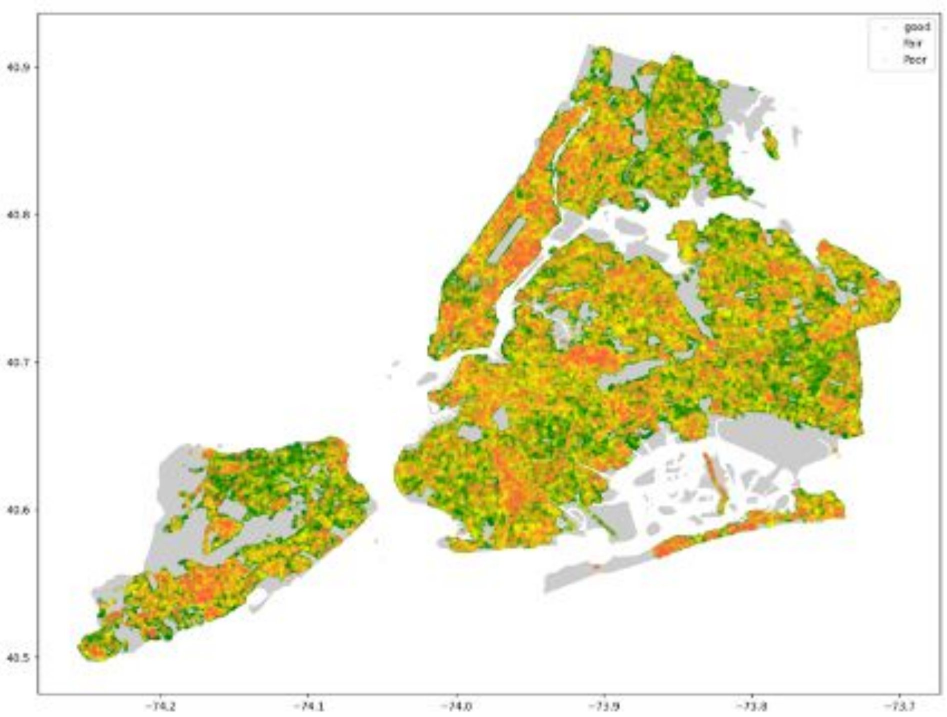
NYC Tree Census: Exploratory Data Analysis and Visualization

Kiyan Mohebbizadeh (km3826), Anne Wei (yw3939), Wenxi Zhang (wz2615), Zhening Zhang (zz3040), Tal Zussman (tz2294)

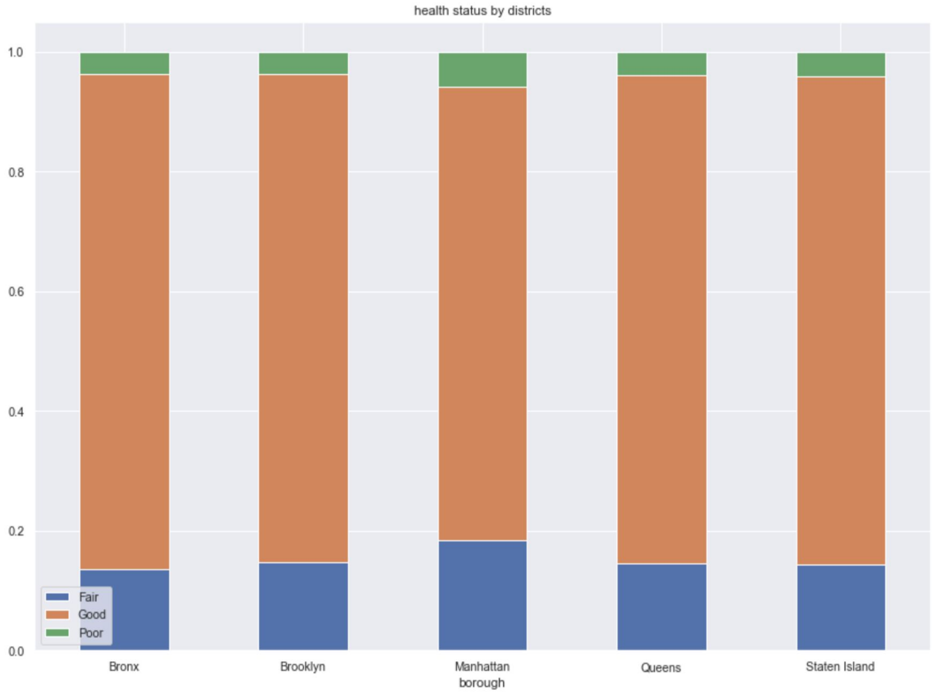
Initial Data Exploration: Map of Tree Health

From this map we can see that tree health is somewhat correlated to the location of the tree. The trend seems to show that more urban areas as well as coastal areas have overall worse tree health. Looking at Manhattan, Brooklyn, and the coastal areas, we can see a concentration of yellow and orange. In contrast the more suburban and less crowded areas of NYC (Queens, Staten Island and Harlem) there is a higher concentration of green signifying better tree health.



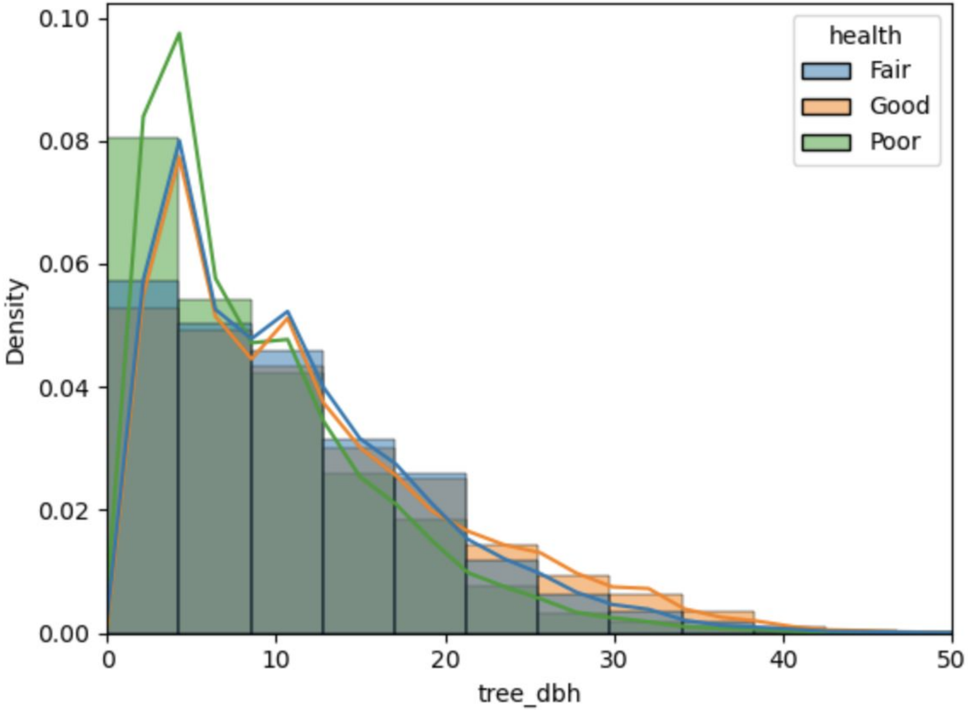
Initial Data Exploration: Health by Borough

In this stacked barchart we can see that regardless of the borough, the proportion of poor, fair, and good trees is pretty similar. The only noticeable difference is in Manhattan with lowest proportion of good health trees and the higher proportions of fair and poor health trees. Still, the difference in proportional tree health is relatively constant throughout the boroughs.



Initial Data Exploration: Health, Density and Diameter

DBH is diameter at breast height (4.5 feet off the ground). In this chart we can see that trees with smaller diameters tend to be the ones with poor health, whereas the trees with the larger diameter are more likely to have better health. We can see this especially clearly with the trend lines mapped. When the diameter is small, poor health trees are by far the most frequent. As the diameter gets larger we see the good health line surpass and rise above the fair and poor health lines around a diameter or 20

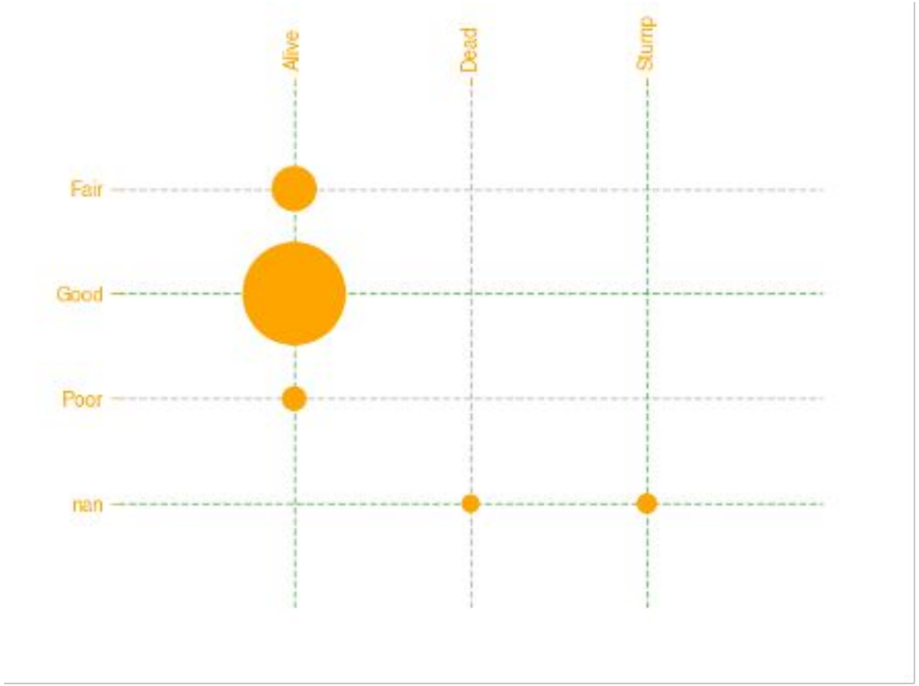


Cleaning: Missing Values

This diagram shows the missing (NaN) values by each category. We can see that all of the missing values for health are a result of a tree being a stump or dead.

That means that for health purposes we can consider all the NaN values as the tree being dead.

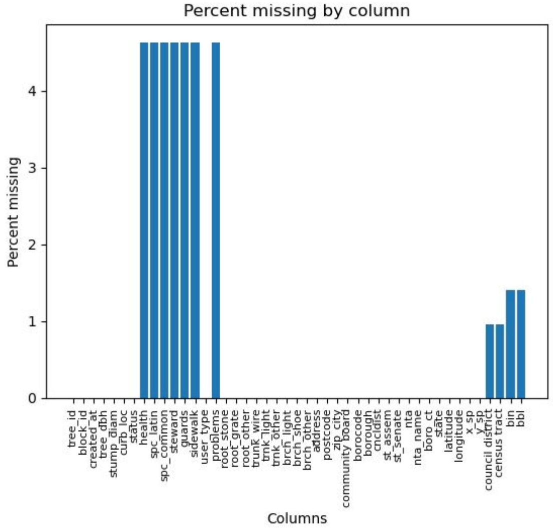
Looking at the data for dead trees, we noticed that the data is sparse. For this reason we will be removing dead trees and stumps from consideration and focus our model on predicting the live tree health.

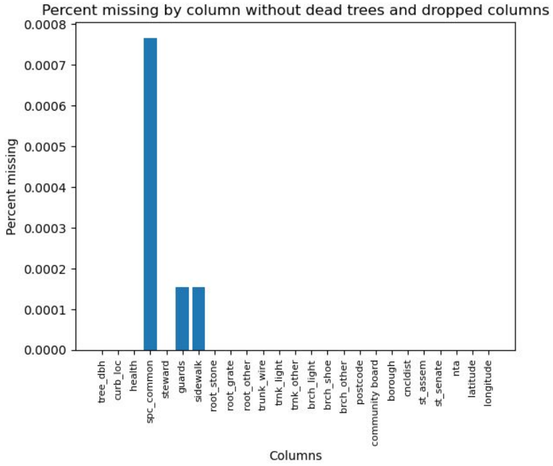


Cleaning: Dropped columns

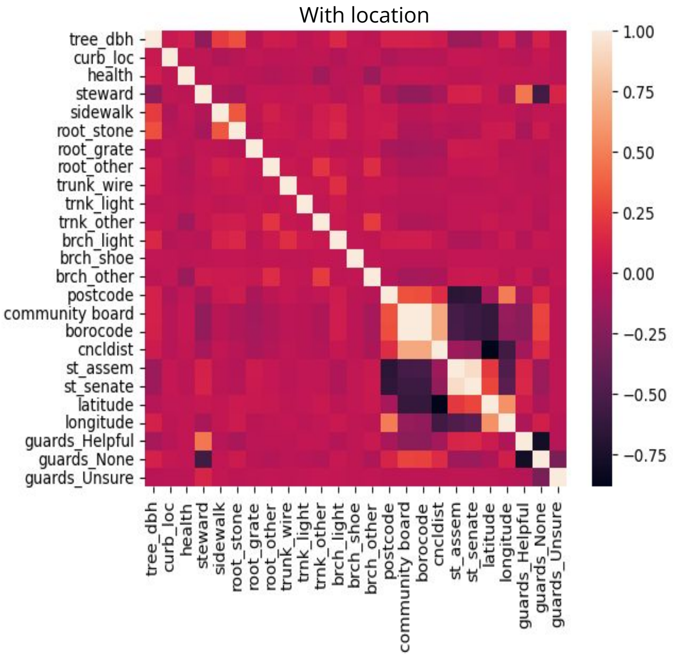
Many of the columns in the dataset are directly related to other columns in the dataset, and as such can be removed (for example, ‘borough’ and ‘borocode’).

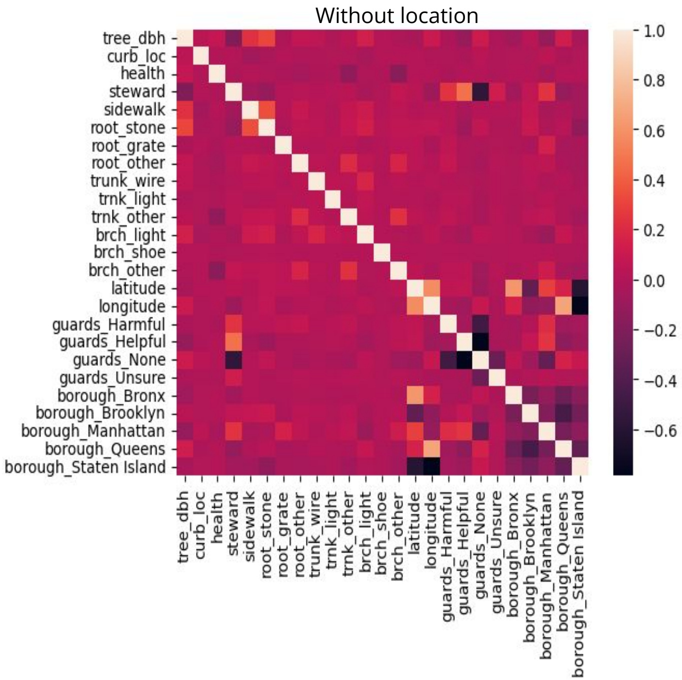
After removing such columns and dead trees, there were only 6 rows with missing values, which is a very small fraction of the total data, meaning that these rows can safely be removed.





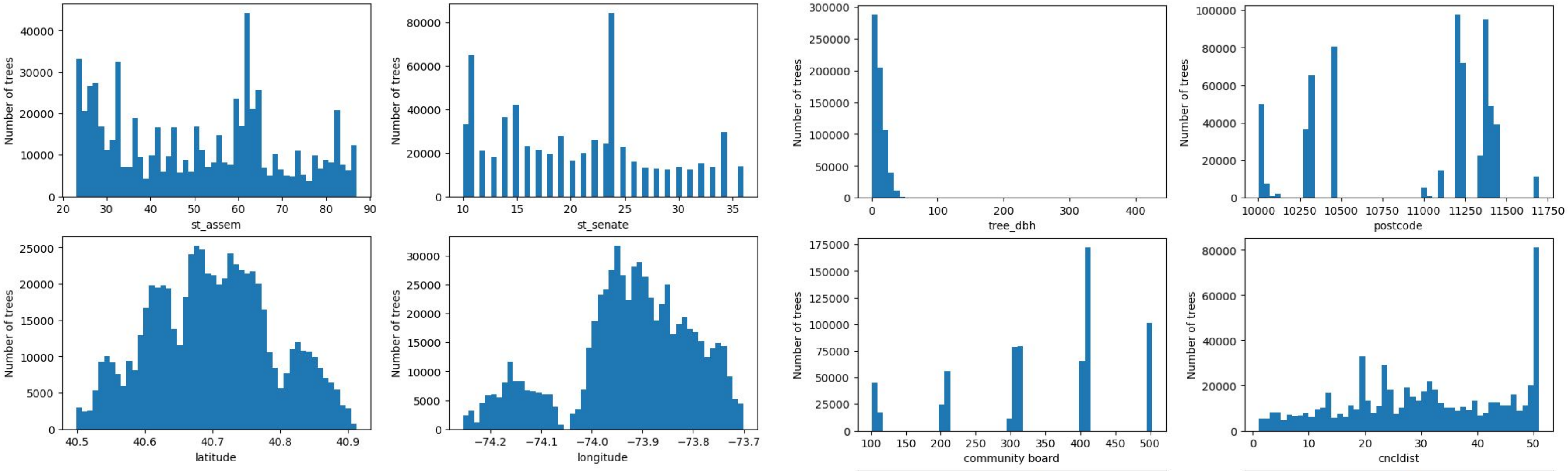
Cleaning: Correlation



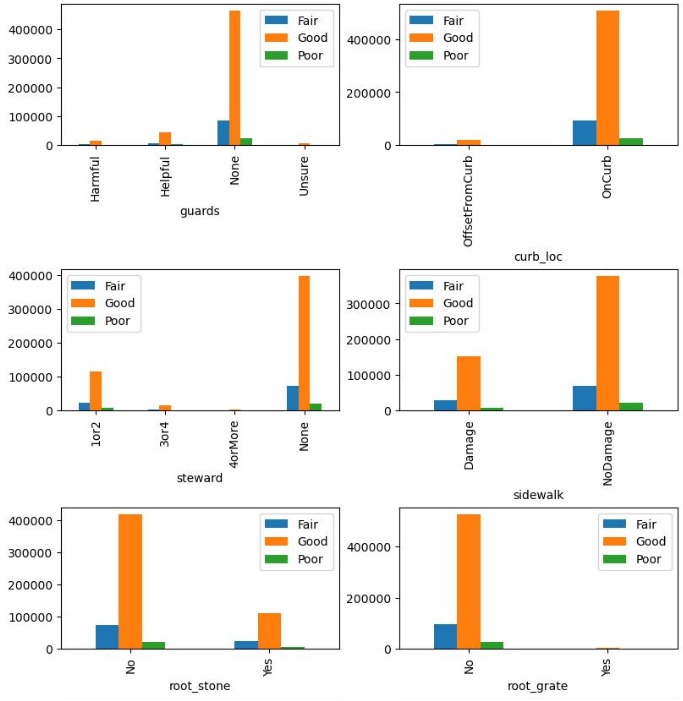


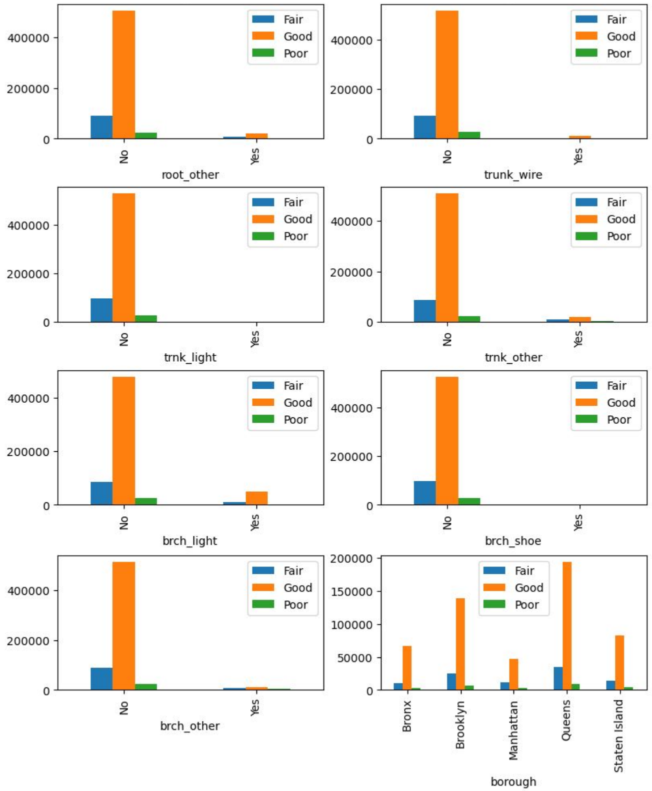
There were several different columns corresponding to location, which were highly correlated with each other. These can be safely removed to reduce the computational cost of applying models.

Insights from Data Exploration: Distribution for Numerical Variables



Insights from Data Exploration: Distribution for Categorical Variables





Machine Learning Techniques Proposed

Since this is a classification problem, we plan to evaluate the following techniques:

- Logistic regression with ordinal responses
- Decision Tree
- Random Forest
- K-nearest-neighbor
- Support Vector Machine
- Naive Bayes Classification

Additionally, since the dataset is imbalanced, we will undersample the data.