Assignment 4 - Data Science

Team 8 - Fred Buchanan, Vedant Gannu, Shepard Gordon, Alexander He

December 17, 2020

1. **Choose an investigation and identify pre-existing source of data that can address a particular data science goal (7%)**
   a) Choose, and state, the goal and reasons why the datasets were chosen and how they were found and managed, Min 3-4 sentences.

The goal of our project was to study and analyze data related to one of the United Nation's 17 sustainable goals [1]. Our team chose data from NYC's Open Data website to better understand how effective NYC is at meeting UN goal 11 ("Sustainable Cities and Communities''). The datasets chosen were 1) the 2015 Street Tree Census [2] and 2) the NYC Property Valuation and Assessment Dataset [3], both of which were previously collected and catalogued for NYC. These datasets were selected to analyze correlations that may exist between tree count density and property value. Our group decided to search for any correlation between tree density and property values in zip codes. A positive correlation would encourage developers to plant more trees to increase their property values, leading to more trees and thus an overall more sustainable and healthy city. The tree data was downloaded as a CSV file, and the property valuation data was downloaded as a Microsoft Access Database file which was converted to txt file format and then saved as a CSV file.

b) Document and discuss the data formats and any metadata standards/ conventions in use, and the method(s) of discovery and access and how they helped or hindered the process, Min 3-4 sentences.

For the 2015 Street Trees Census dataset, the site where it was downloaded from (NYC Open Data) provided a data dictionary file defining the column fields. This well documented metadata helped us understand all values in the dataset and can also be used by future users seeking to understand our derived values.

For the property valuation data, NYC Open Data provided a Property Valuation and Assessment Data Dictionary in Excel format. This excel sheet provided short descriptions about all Property Valuation and Assessment Data columns along with agency information, number of downloads, and the date of last recorded update to the data.

In terms of data formats, our team saved the excel data files as CSV files due to the portability and versatility of the CSV format. We left the original dataset metadata files in their respective format since they already contained the needed information for provenance, but additional data sets that we derived were documented in Dublin Core metadata standard for easy

readability.

2. Data Analysis (10%)

a) Develop and state two particular questions/hypotheses related to the goal of the investigation and that can be answered using the datasets under consideration. Design an analysis study (preliminary, full and post) to answer these questions and document the analysis design, Min 3-4 sentences (3%)

The team's first hypothesis is that there is a positive linear correlation between the density of trees in an area zip code and the mean property value in the region marked by that zip code in New York City. The associated null hypothesis is that there is no correlation. Our team expected a correlation based on the team's initial understanding that urban trees are often considered a quality of life improvement associated with wealthy neighborhoods. As a result, the team expected wealthy citizens (who live in high property value buildings) to have more trees around their properties.

The team's second question is whether or not another property of trees better correlates with mean land value. The team explored this question using a random forest classification algorithm. The associated null hypothesis to this question is that no aggregate properties of trees explain mean property value.

b) Provide a description of the choices of tools/methods used or a description of any code or scripts written, and describe how your results were stored and managed, Min 3-4 sentence. Submit your code to course GitHub repository for evaluation (3%)

The team chose to primarily use the Python programming language for this project due to everyone in the group being proficient with the language. The project performed rigorous numeric and statistical analysis through Python libraries including NumPy, SciPy and Pandas. Our Python scripts were stored in a GitHub repository, along with data artifacts, metadata, associated visualization, and numbered to achieve easy reproducibility. For each dataset that was generated for analysis, python scripts and CSV files were created to

c) Perform the analysis in a form that can be validated and describe the steps and results your group took to ensure this validation, Min 3-4 sentences (4%)

Mean property value was calculated from the full value of all properties in each zip code, except for zip codes with less than 50 buildings. Smaller zip codes were excluded from analysis for being outliers. A simple count of trees was used for tree density. A simple count of trees by species by zip code was also performed. These data were merged by zip code before analysis.

A simple linear regression was used to calculate the correlation between the number of trees and the mean property value. We found an R^2 of 0.081, an extremely low value suggesting that the number of trees has no predictive power over the mean property value of a zip code. This caused us to reject our initial hypothesis. These data suggested that higher valued neighborhoods do not have more trees than lower valued neighborhoods.

We then wanted to find if other data had better predictive power over mean property value. To determine this, a random forest classifier was trained to predict mean property value from the number of trees of each species in a given zip code. The Gini importance values from this classifier were examined. The number of Callery Pear species trees in a zip code was found to have the most predictive power over mean property value. The Purple Leaf Plum species was then found to be a distant second followed closely by the Norway Maple and Northern Red Oak species.

3. Presentation/Visualization (8%)

a) Prepare presentation / visualization of both the data (and any metadata, information) and the results of the analysis and describe them, Min 2-3 sentences. (3%)

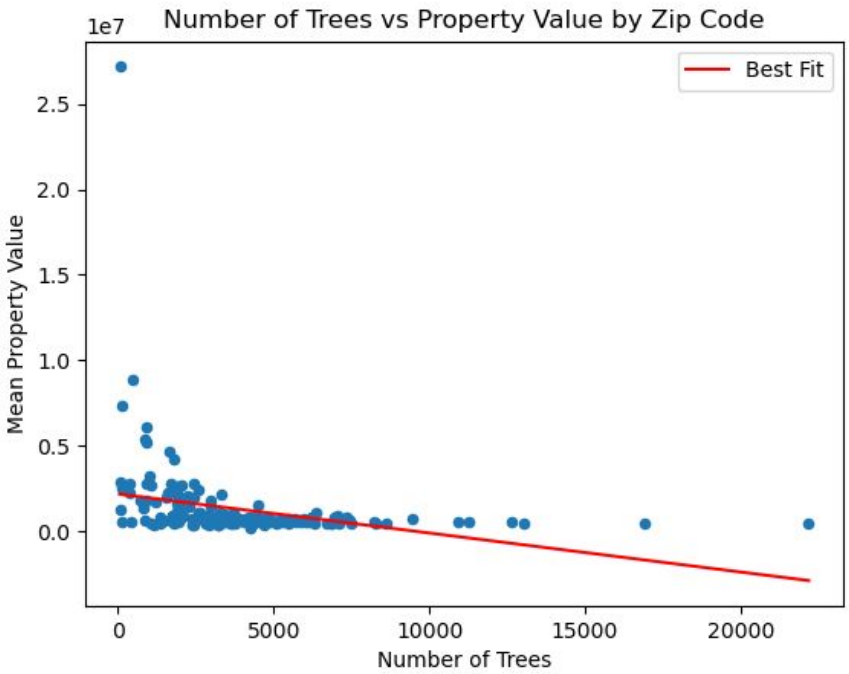

Figure 1: Number of Trees by Zip Code vs Mean Property Value. Although we initially hoped for some correlation here, the best fit line is mostly flat, indicating the number of trees is not a good predictor of mean property value.

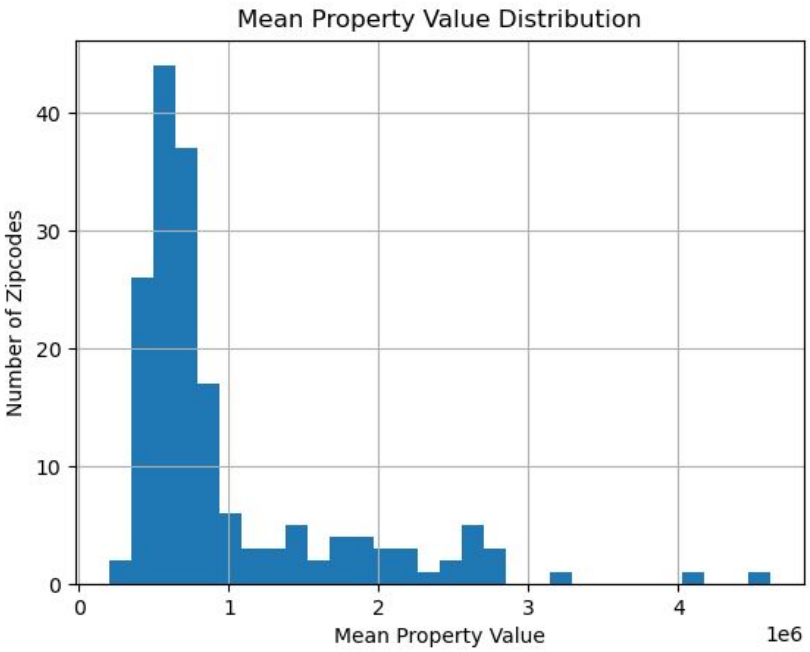

Figure 2: Distribution of mean property value by zip code.

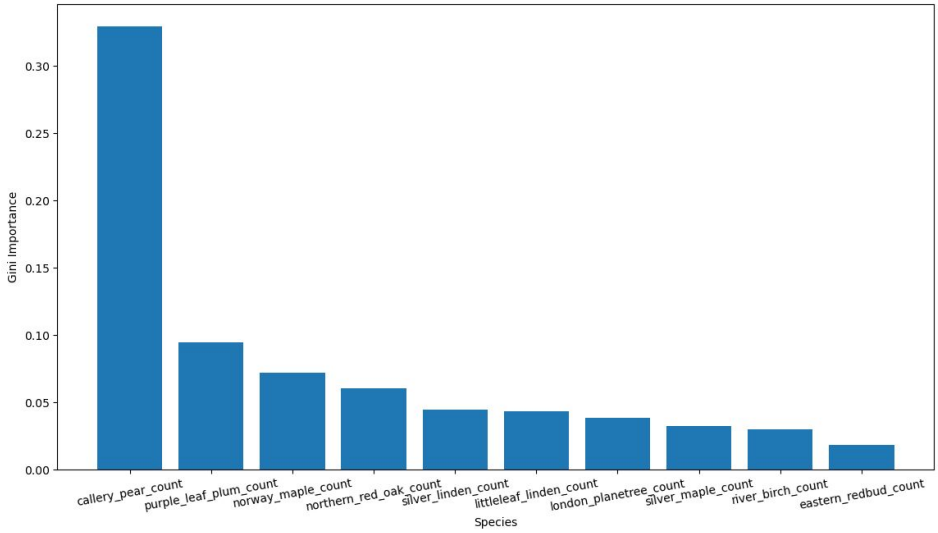

Figure 3: Gini Importance of the count of each species by zip code. Determined using a Random Forest Regressor trained on the counts and predicting the mean property value. In short, high importance indicates greater predictive power.

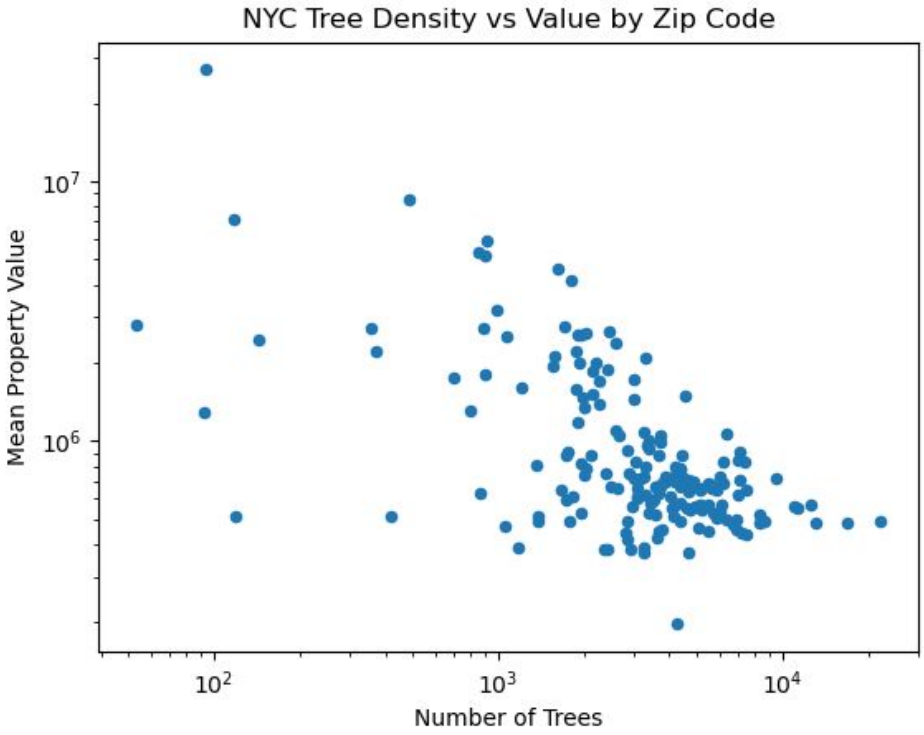

Figure 4: Number of Trees vs Mean Property Value on a log log scale. No correlation is noticed.

b) Document the management of the presentation / visualization products and any associated metadata, etc. Min 2-3 sentences (2%)

The presentation and poster for the project were created through team collaboration on Google Docs. Once both documents were completed, they were saved to the team's Google Doc folder. This location will enable both documents to best explain the aims of the project as well as contextualize and document the project's progression.

c) Describe how your presentation/visualization supports the goal of the data science investigation and highlight any value that was gained, Min 3-4 sentences (3%)

Using a scatter plot, we are able to visually observe the lack of correlation between the two axes. It also allowed us to recognize outliers for removal. Furthermore, the bar chart of importance provides a clear visual ranking of species impact on mean property value.

4. Describe your overall data management plan for the results for questions 1,2, and 3

using the 9 categories from assignment 2, Min 1-2 sentences for each category (5%)

Creation of logical collections: Our original datasets come from the NYC Open Data website accessible as downloadable Excel and MS Access files in table column format. The Tree Census data was downloaded and saved to CSV format, and the Property Valuation Assessment data was initially downloaded as an MS Access File then exported into a txt format and later changed into a CSV. Two CSV files were generated from these initial files, one for comparing tree densities per zip code and the other for comparing mean property valuations per zip code. Once these two new CSV files were created, they were inner joined on the Zipcode field.

Physical data handling: The original physical data is maintained on Google Drive. The Tree Census data is stored as a CSV file while Property Valuation and Assessment Data is kept as a text file. The metadata or information files associated with the original data sets is kept in maintained on Google Drive as well. Any derived data sets from the original data are stored as CSV files in the team's Github in the "data" directory. All stored files are publicly available for anyone to look at or download.

Interoperability support: Data files were saved as CSV files, while Python Pandas DataFrame were utilized for data creation and manipulation. This combination enabled the files to be easily transferable for anyone who wants to explore the data. One issue our team encountered is the original data files (Tree Census and Property Valuation) could not be stored on Github since the data file sizes exceed the Github repo limit. To remedy this issue, the team opted to store the original data files on Google Drive.

Security support: All source files and data related files are kept on Google Drive and Github, both of which by default provide basic forms of security.

Data ownership: Since the original data comes from NYC OpenData, a public source for NYC related data collections, the data we used is also available to anyone via URL or through the Google Drive/Github repository. The team did not change or update any data, but instead produced derivations of the data and created new data sets from the original data.

Metadata collection: Associated metadata files from the original source data files are available in Google Drive as well as the Github repo under the "metadata" folder. These public and accessible files, metadata and analysis metadata can be converted to standardized metadata files either in Dublin core or a custom metadata format.

Persistence: The data is primarily maintained and housed in Google Drive since it provides enough storage capacity for our large data files. In terms of management, the original data is subject to change since NYC OpenData updates their datasets at least once every year. If any future team continues the project for future study, that team should pull updates from NYC OpenData to ensure the original datasets are up to date.

Knowledge and information discovery: The primary goal of our analysis is to see if there's any type of correlation between average property value and the number of trees for zipcode in NYC, or if other metrics could better predict average property value. In addition we attempted to find a correlation between species type and average property value, and even used a random forest classifier to see if species can predict property value . While no positive correlation was found between average property value and number of trees, the project is still valuable since it can help policy makers narrow down their search for positive incentives to encourage influential

stakeholder in NYC to make the city more sustainable. The group also did find minor correlation between tree species type and property value.

Data distribution and publication: At the moment the derived data will be available on Github in the “data” folder of the repo. There are no mechanisms to “advertise” the study and analysis until a decision to work on a publication is organized. For now all materials are publicly available provided public access links to the resources.

5. Create a poster (poster templates are available on LMS). Please submit your poster on LMS using the same naming scheme mentioned above. Mandatory peer-evaluation form must be submitted within 12 hours of the final project submission on LMS in order to receive the presentation and participation grade. (10%)

Citations Used

1. THE 17 GOALS | Sustainable Development. (n.d.). Retrieved December 16, 2020, from https://sdgs.un.org/goals
2. Department of Parks and Recreation (DPR). (2016, June 3). 2015 Street Tree Census - Tree Data. Retrieved December 11, 2020, from https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/pi5s-9p35
3. Department of Finance (DOF). (2020, May 26). Property Valuation and Assessment Data. Retrieved December 11, 2020, from https://data.cityofnewyork.us/City-Government/Property-Valuation-and-Assessment-Data/yjxr-fw8i