SPRINGBOARD DATA SCIENCE CAREER TRACK

CAPSTONE PROJECT #1 - DATA WRANGLING

SILVIA MAIONE

Initial cleaning

In the first part of this project the data is analyzed to identify and deal with any missing values. Also, since many features are of object type, they are converted to numerical type, so that they can be used when applying machine learning algorithms.

The Pandas module is used to convert the dataset, which is in a spreadsheet format, and do the initial analyses. Methods such as “info()”, “describe()”, etc. are used to have an idea of columns composition and preliminary statistic information, to identify unique names and their frequency. Some columns are removed from the dataset already at this step because they are considered not relevant for the purpose of this project. The documentation that accompanies the dataset gives more information about such columns and it is used to support this choice. This also allows a partial reduction of the dataframe size. For example, the columns with geocoded information are removed, together with the “state” column. The latter has the same value for each entry; the columns with geographical information are dropped because the purpose isn’t to find a specific correlation between trees’ health and exact position. Only longitude and latitude are kept.

Most of the entries are of text type. They are checked against any typos or misspellings. After any necessary correction, they’re ready to be converted. Such conversion to numerical type is done using label encoding and/or one hot encoding.

A check for duplicates is also done, and if any are found, they are removed.

This initial cleaning step leads to a reshaped version of the dataframe, that can be used for exploratory data analyses.

Missing data

The preliminary analysis shows that there are missing values. Any rows that have a missing value in the tree health column are removed, since the tree health itself is the target variable to predict. After this reduction, a few columns have less than 5 missing values. The corresponding rows are removed, since such number is very low with respect to the number of entries. One last column instead has a about 50 missing values. This is the "problems" column, which has over 200 unique values. Some of them might be grouped, since they refer to similar issues, but the most frequent one is undefined, i.e. "None". Although the number of missing entries is low with respect to the total, instead of being dropped, they are filled with the most common ("None"). Also, the information on trees' problem is present in other features, so at the end this one might not be used at all.

Some features, for example "steward" and "guard", can have a "None" value too, which doesn't indicate an actual missing value, but a specific attribute. For this reason, they are not replaced nor dropped.

Outliers

The only numerical feature, apart from longitude/latitude and the identifying code of the borough, is the tree diameter's value. By looking at the results of the "describe()" method call, it seems there might be some outliers. A box plot shows several values being much higher than the last quartile, which is 16. Values above 100 seem unlikely, so the entries above about 90 should be considered as errors or typos, and thus be dropped. They represent a very small percentage of the data, even after the reduction done so far, so dropping them can be considered acceptable.