SPRINGBOARD DATA SCIENCE CAREER TRACK

CAPSTONE PROJECT #1 – EXPLORATORY DATA ANALYSES AND VISUALIZATION

SILVIA MAIONE

After the initial part, consisting in data wrangling and clean up, the data set is ready to be analyzed. The main goal of this project is to create a model to predict the health of a tree, but also to answer some questions, such as, where most of the trees are located, what influences their health, what the most common species are.

For these purposes, using the cleaned data frame, the health of a tree is plotted vs other features, such as trees' count and location, evidence of stewardship and who conducted the survey. The initial hypothesis is that the position with respect to the curb and the evidence of stewardship might affect the health of the tree. For example, if the tree has been attended to, its health might benefit. The plot seems to suggest otherwise, since most of the trees either don't have signs of stewardship or very few. Also, the report on the health of the tree isn't affected by who did the survey. What is emerging from these initial plots is also that the data set is skewed towards the "good" health, so it will need to be balanced for the purpose of creating the model.

A variable that is investigated is the tree's diameter. The idea is that it can be an indication of the tree's health, i.e. that there is a positive correlation. A bar plot seems to confirm that, so this is a variable that is going to be used in the model. Though, further analyses, for example a bubble plot showing the health vs the diameter and the location (borough), indicates that trees having the same diameter but located in different areas are classified differently. This suggest that the location should be considered as a variable in the model as well. Also, the model itself might need to be applied to the various boroughs separately.

The tree's branch, root and trunk related problems are considered variables that can influence the health. Their possible correlations are analytically investigated. The hypothesis to confirm is that they indeed have some effect.

Another initial hypothesis is that most of the trees are in Manhattan. This comes from the presence of Central Park. It turns out that Manhattan is last in terms of trees' count, Queens being first. A quick check on Wikipedia confirms that Queen's land area is much bigger than Manhattan's, thus suggesting that this might be the actual reason for the difference. Also, plotting the first 20 species in terms of their number, shows that there are common species in both neighborhoods.

In summary, initial plots confirm that location and diameter should be considered in the model, since they seem to have a strong correlation to the tree's health. Also, other variables will be selected based on statistical methods to find correlations.

The main hypothesis is that trees with larger diameter and located in bigger areas are healthier. The influence of other factors will also be investigated, and a model will be built to verify such hypothesis.