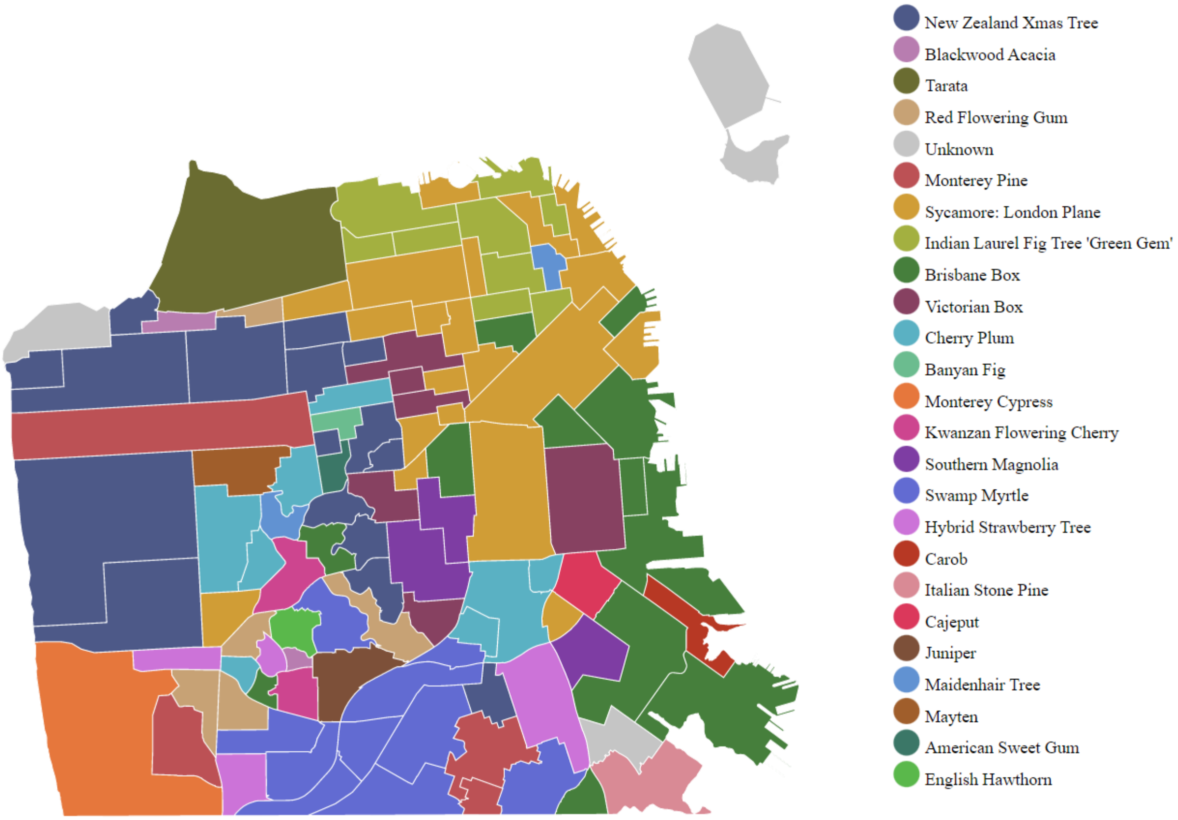
INFO 4310 HW1

Yvette Hung (yh387)



After some exploration of the data, I found that there were many species of trees planted throughout the city and was interested to see if there were any patterns as to which areas tended to plant what trees. For each neighborhood, its color represents the most common type of tree in that area. It turns out that the distribution is diverse: as many as 24 different species represent the 117 neighborhoods of the city. There seem to be clusters of communities that tend to focus on one particular tree. For instance, swamp myrtle is popular in the southern part of San Francisco. There may be many factors influencing these distributions, including regional differences in geography or community.

Processing on the trees dataset was done by using Jupyter notebook to create and export a smaller csv file that only left four fields: tree ID, species, latitude, and longitude. In addition, all empty values in the species column were replaced with “Unknown”. This became handy when calculating the list of most common species and prevented any exceptions. The unknown value can also be misleading: an area colored gray might represent a neighborhood with no trees, or one in which most of the trees are of unknown species. However, I found it an acceptable tradeoff to have a way to deal with missing values in the data.

Since the most common species of a neighborhood is an aggregated value, I decided to color entire areas rather than coloring individual trees. Displaying all the trees would either result in using too many colors (there are 419 different species in total) or information would have to be explicitly omitted (if only trees on the list of most common species were colored) to convey the same information as coloring areas.

Choosing the colors was difficult as 25 were needed, though initially I hoped that only a few species would be common across the city. I considered not labeling the colors because of the large number, but decided that the information should at least be available for users, especially since the visualization is not interactive. I used an online tool (https://medialab.github.io/iwanthue/) to select colors. The advantage is that the colors are distinct enough to get the general message across (that certain areas prefer certain trees). Meanwhile, a disadvantage is that colors and species were matched somewhat randomly, and I had to manipulate the scale to match “Unknown” with gray, the conventional color for “no data”.

In general, there are some drawbacks of this visualization such as the density of trees not being shown, but as a static map provides limited space to present data, I decided to focus on showing neighboring areas with the same majority tree species instead.