INFO 5311 HW1 Writeup

Visualization of Caretakers of Municipal Trees and Diameter of Trees

mmc324

The Story

After exploring the data, I was interested in the different caretakers of the trees. I want to show the user where different caretakers' trees are located in San Francisco and how large their trees are, which is represented by diameter at breast height (DBH). Maybe there is a large cluster of a certain group of caretaker's trees, or an area with specifically large trees. If a caretaker has larger trees, does that mean they are necessarily better or healthier trees? Upon exploring the data, I noticed that two of the caretakers were especially prominent: Private and DPW. This meant that they would overtake most of the trees on the map, which is quite prominent in the visualization. It appears that DPW has larger trees that are typically placed in straight lines in San Francisco. Private appears to have more randomly scattered trees all over San Francisco with a few clusters of larger trees. Many of the other caretakers had less than 100 trees. In a map highlighting the less common caretakers, it is notable to see how the caretaker Rec/Park appears to have large trees and caretaker Port's trees are clustered by the edge of San Francisco. I also decided to create separate maps for each of Private and DPW to help the user differentiate between the two entities.

Then, I became interested in the species of the tree, since a caretaker may specifically own many of the same species which would influence how large the tree is. In the balloon plot, I want to present the relationship between the top 5 most popular caretakers, and the top 10 most popular species, and how large the average tree's DBH is. For example, are some caretakers' trees of the same species larger than others? I also expected DPW and Private to have larger trees, since they seem to have most of the trees. However, there were some unexpected insights where the caretaker Rec/Park has seemingly larger southern magnolia trees than the other caretakers. Additionally, it seems like the species sycamore: london plane trees are larger than others, regardless of caretaker. DPW appears to have the largest average Indian Laurel Fig Tree 'Green Gem'. DPW consistently has larger average diameters than Private which is represented on the map.

It is really interesting to see how caretakers tend to have clusters of trees, or straight lines of trees throughout San Francisco. We can also see trees owned by certain caretakers that are larger than others. Additionally, there is not necessarily a relationship between species and diameter of the tree, which could be influenced by caretaker but also other factors that could be further investigated.

Data Processing

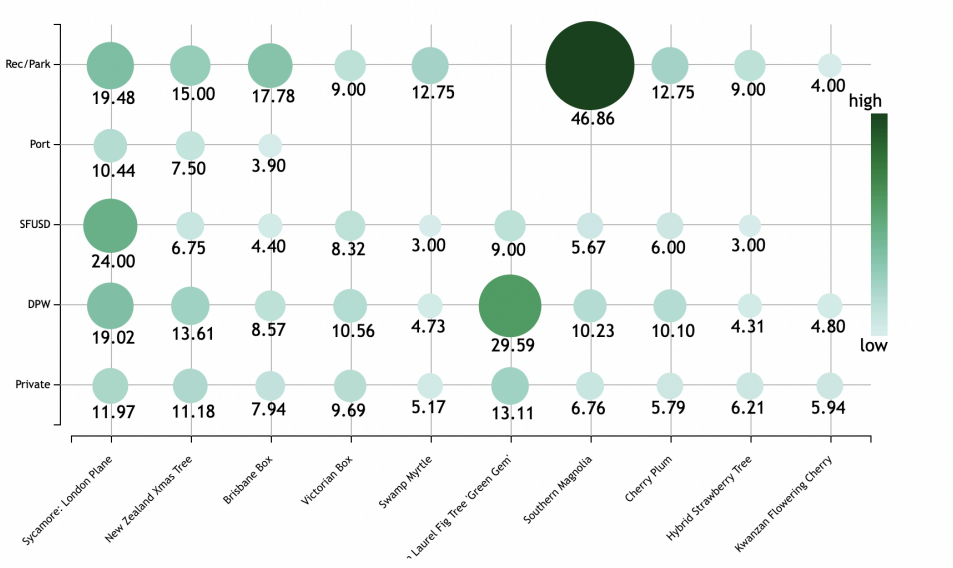
First I wanted to get a list of all the unique caretakers and how many trees they owned using loops, arrays, and dictionaries. I also converted the diameter at breast height, latitude, and longitude attributes into numbers, as they were previously inconsistent. I required different datasets for each of the four maps. I filtered out Private and DPW in the dataset of the less common caretakers. Then I have two datasets of just Private caretaker and DPW caretaker

data. For the balloon plot, I created a dictionary of each species and how many trees of each existed. I sorted the dictionary by the number of trees and selected the top 10 species. Since there were fewer caretakers, I was able to see which caretakers had the most trees, so I created an array of the top 5 caretakers. I filtered my dataset for only the top 10 species and top 5 caretakers. I altered each species name in the dataset to only reference their common name. I then created a dictionary, which groups each species-caretaker pair and calculates the average tree diameter.

Visual Encodings and Design Choices

For the maps, the **marks** are represented by circles on the map. I represent the caretaker through the visual channel of color hue. I originally used SchemeSet2, but realized I needed 22 colors. I was inspired by SchemeCategory20, since it is an already categorical color scale and makes it easier to differentiate between different caretakers. I also included a legend for the two maps with multiple caretakers, so the user can see which caretaker is referenced by a point. The **visual channels** for location of the trees are represented by an unaligned

vertical and horizontal position. The diameter at breast height is represented through the radius of the data point. I decided to include a radius difference on this plot because I think it is easy to see large differences in size on the chart so the user can see where larger trees may be located. I use a maximum radius of 7 because my experimentation with larger radii led to very difficult to view maps. However, I understand that there are many overlapping circles, which makes it difficult to see some of the points on the original map. I tried to supplement this issue with the 3 other maps, which would help the user see more detailed information about different caretakers. On the map with less popular caretakers, it is significantly easier to see the locations and sizes of their trees, as everything is so spread out. I decided to include a map for each of DPW and Private because they did tend to clash with each other on the original map. I believe that these maps make it significantly easier to see their data points. Even though the original map has many overlapping points, I still think it is interesting to see all of the data combined onto a map. It is interesting to see how the DPW trees intersect the Private trees, but are also larger. I still believe the original map can bring out useful insights, despite the messiness.



For the balloon plot, **marks** are represented by circles on the chart. The **visual channels** for species are represented by aligned horizontal position, while caretaker is represented by aligned vertical position. The diameter at breast height is represented by the radius of the circles and color hue, where lighter color is lower diameter and darker color is larger diameter. I thought the size of the trees would best be represented by the radius of the circle because it is intuitive that a larger point represents a larger diameter. I think color hue helps to supplement this decision as a second method of seeing the change in diameter, if the radius differences are visually too miniscule to see. I also added text labels with the average radius size to give the numerical value, which helps to clarify the difference between similar circle sizes. I chose the blue-green sequential scale because I felt that it matched the color

scheme of the original map, which is mostly green and orange, resulting in a more cohesive visualization. I also included a color legend to help the user see that higher values are related to darker colors. I chose a balloon plot because I knew I wanted to represent the relationship between a numerical value which would be grouped by two categorical variables. I think the balloon plot is very good at delivering the insights of these variables.