Homework 1: Visualizing San Francisco Trees

INFO 4310
Stephanie Zhang

Overview

The main story I was trying to convey with the data was the distribution and spread of different species around San Francisco. However, the original data had over 404 unique species (sometimes with just one tree per species) and I knew it would create too many overlaps and overwhelm a user from just the first sight of the map. I decided to filter the data to the most common four species. The radius of the circles that each represent a tree from these species is proportional to the diameter of the tree. This is an important story to tell because while we know San Francisco is agriculturally diverse, details of the specific locations of different species of trees can help us with understanding city planning or can be used to identify landmarks and significant trees. Many of the trees in this database are the only tree of that species. This can raise questions regarding invasive species that were brought from around the world. The most common trees picked are more used for city planning and I was expecting them to appear in parks or other green spaces. I also wanted to look at the age of these trees as San Francisco was ranked first by per capita income and sixth by aggregate income as of 2021 and is undergoing ongoing and advanced gentrification. A bird-eye's view of the Bay shows only 3 main parks which didn't show up in the previous visualization. This made me curious as to who the caretakers of these trees were. Many landmark trees are also located on private property. I was tackling the average age of trees based on their caretakers. This was to help see the differences between how well-maintained the trees are.

Data Processing

In my preprocessing steps for the first visualization, I first separated the common name from the qSpecies column using the Split() function and kept the column after the ":: " characters. Then I value_counted() the number of trees to get the names of the four most common species and created a new dataframe with tree records belonging only to those four species. In D3, I cast the DBH value to a Number and put that through my radius scale. I was still curious about the age of the trees in the database. However, because of all the null values and the uneven spread of null value PlantDate across the different species, I grouped the trees by their qCaretaker.

To get the age of the trees, I first converted the PlantDate column to a string and then to a datetime column. Then I calculated the number of day differences between the day I filtered the data and the datetime in the PlantDate column. This generated the difference between the two dates in days. Then, I divided the number of days by 365 into a new column called "age". This tree collects the age of the tree in years. To represent numerical data along a categorical axis, I decided to implement a bar plot and find the average age. I wanted to group by a different categorical column to show more features of the dataset and decided on “qCaretaker” because it had 0 null values and because it had an interesting selection of caretakers. Unsurprisingly, the landmark tree average age was highest.

Visual Encodings and Design Rationale

For the first visualization, the four colors I picked to represent the different tree species were picked to contrast with the black background of the San Francisco map. I chose to plot each tree as a circle because not only did I want the mark to be easily interpreted as a pin for location but also because I wanted each radius to reflect how big the tree was. Traditionally, we look for the number of rings to identify the age of the tree or we measure the diameter of the tree to see how big it is. I wanted to use circles because it encodes the cross-section of a tree which it is easier to measure and count the age of the tree. The opacity of each circle was set to 0.4 to help accommodate many different trees in the same area. Originally, I was thinking of making the radius of each circle dependent on the age of the tree. However, I quickly realized that the PlantDate was null for a lot of the trees with some disproportionate spread across the different tree species. This encouraged me to look toward other columns of data that were collected, which preferably had no null values. The diameter at breast height was even better than the age as a radius because they intuitively represent each other. Seeing as how the diameter is just double the radius of the tree, I customized each circle on the map to have a radius proportional to the diameter measured. This lets us see not only the positions and specific species of trees belonging to the most prevalent four species but also how big each tree is. The overlap of circles does not mean the trees are right next to each other and the exaggeration of the radius was to ensure that the points would still appear on the black background. From the first visualization, we can see that Sycamore: London Plane trees are specially planted together along the bordering neighborhoods of northern San Francisco, and rough sections of these trees show two different diameters. This suggests that these groups of trees were probably planted together in two separate efforts.

For the second visualization, I wanted to use a different color to represent age compared to the four I picked for the first visualization to avoid the confusion that they represent the same column. I kept the bar graph the same size as the map to make sure the data was displayed to an appropriate size. Because the names of the caretakers take up a lot of horizontal space, I angled it at 45 degrees to increase the font size. According to the visual channel order of human interpretation, length along an aligned vertical axis is most discernible to the human eye so that’s why I used a bar graph. While the bars are not sorted in order, it’s easy to see that the average age of Permitted Site trees is greater than that of DPW Maintained trees.