Homework 1: Visualizing Street Trees in San Francisco

Kate Liang, ksl67

INFO 4310

OVERVIEW

This dataset represents all the street trees in San Francisco currently maintained by the DPW. I wanted to center the data around the types of tree species currently in SF, and I designed my visualizations with the goal of helping those with allergies navigate which areas to possibly avoid and give an overview of the different tree species over the years. My project starts off with an ordered list of the 5 most popular trees in SF and how many there are of each in the city. The first visualization below the list is a map of San Francisco, plotting the locations of all of the 5 most popular trees, with the color representing the type of tree. The second visualization is a line graph of the average diameter at breast height (DBH) of each of the 5 species over the years in comparison to the average DBH of all SF trees. The third visualization is a scatterplot of the number planted and average plot size of each of the 5 popular species every year. The goal of the three visualizations is to first give a general overview of the tree species landscape in SF by providing the reader with a list of the most popular types. Then, I wanted to show where these types were located using the map. Afterwards, I wanted to give the reader insight into how each of these species may have changed over time in their size (DBH) and land area (plot size and number planted) using a line graph and scatter plot.

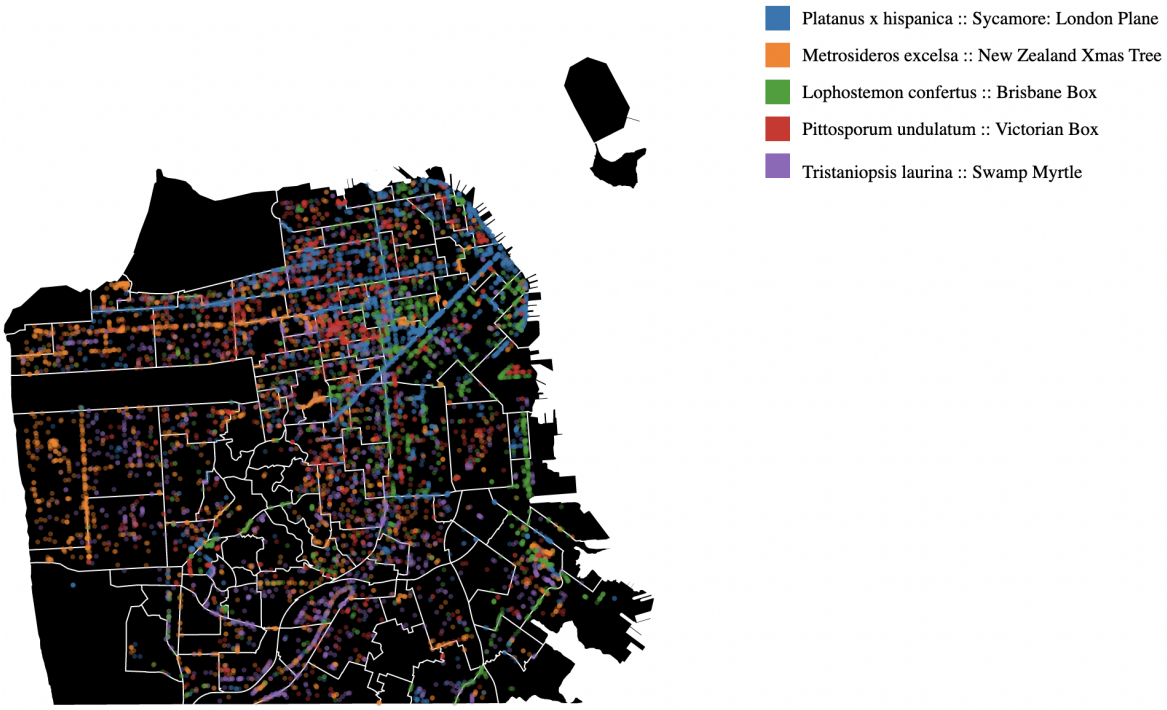
DATA PROCESSING

To process the data for the visualizations, I needed to first classify the trees by species. Then, for each species, I needed to count the total, sum up the DBH and count per year for the line graph, and sum up the plot area and number planted per year for the scatterplot. I also needed the overall sum and count of all SF trees over the years. I used a dictionary object for each of these since this data structure was best for categorizing data. After populating the objects, I calculated the average DBH for each species per year, average DBH of all SF trees each year, and average plot area of each species per year. I also sorted the total count of all the species to determine the 5 most popular trees. Additional processing was also needed due to a few challenges with the data format. For example, the date planted in the CSV was in MM/DD/YY format, so I had to add the 20 or 19 in front of the year, which may have caused some ambiguity of if a tree was planted in 1923 vs 2023. Since the CSV file was dated to 1/30/23, I assumed if a tree was planted 1/30/23, then it was planted in 2023, and all trees after as in 1923 and beyond. In addition, for the plot size, there were multiple formats. The two most common ones were “_ x _” or “Width _ ft”. I discarded any data not in these two formats and

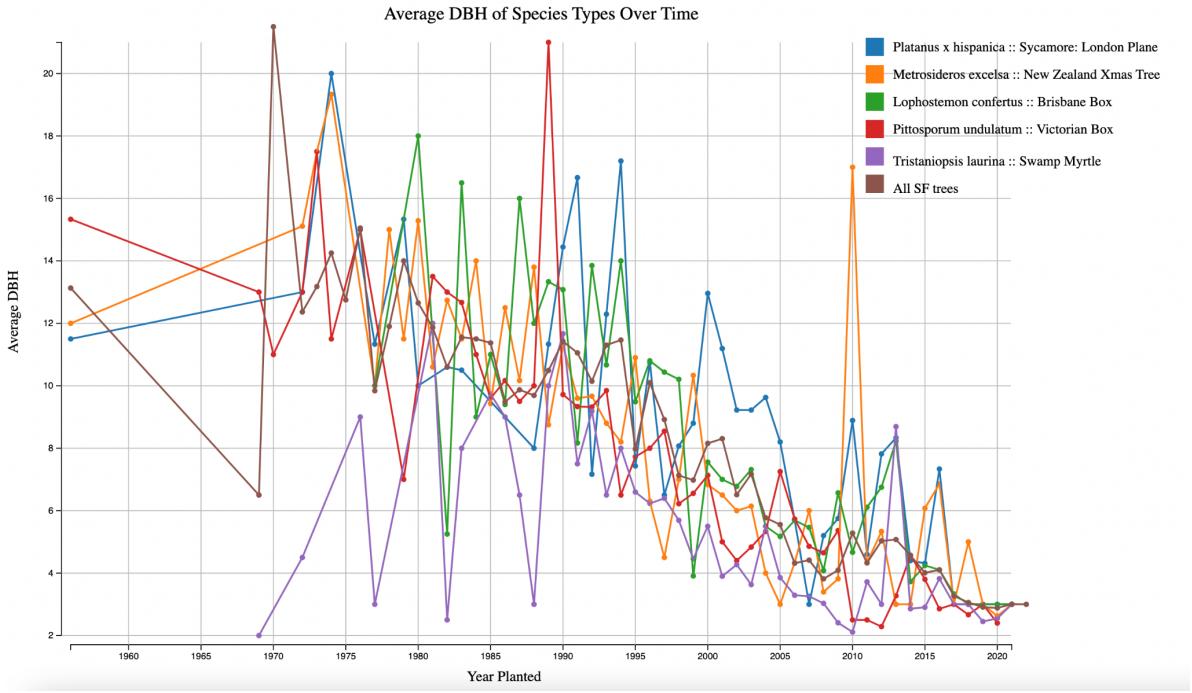
made the assumption that all plot areas were squares and in feet when summing up the total plot area.

VISUAL ENCODINGS

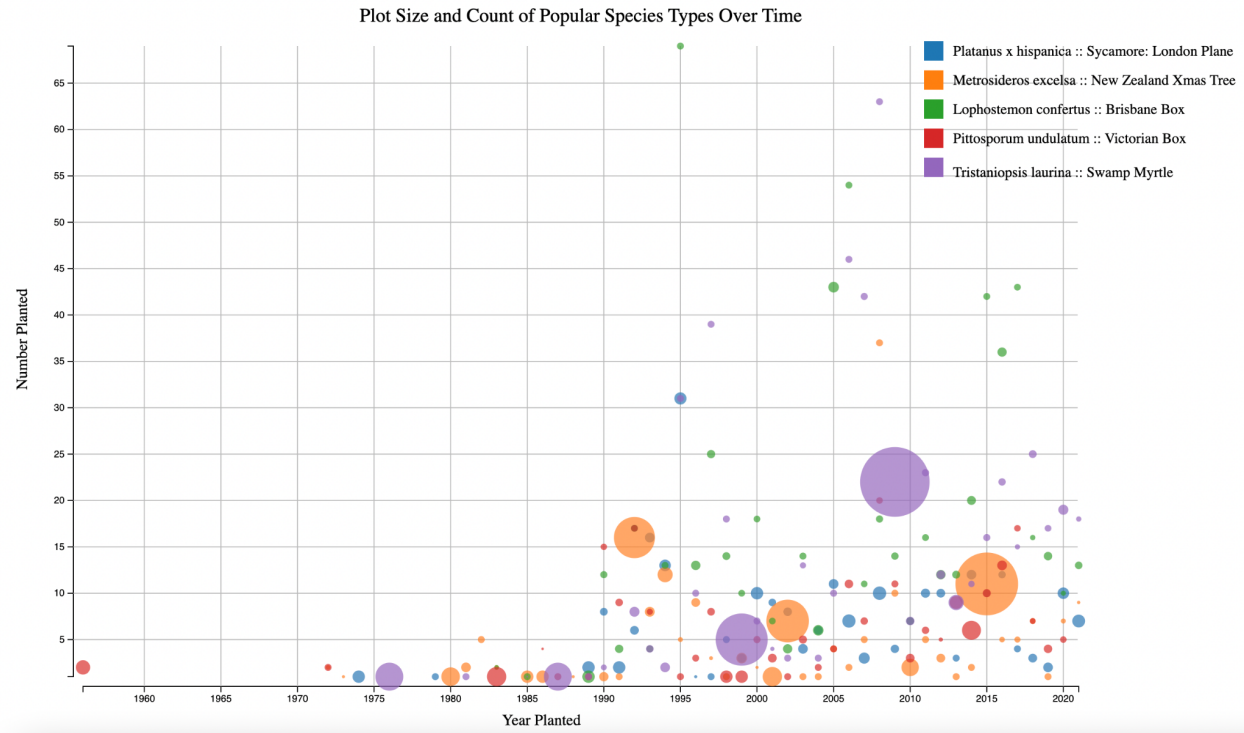
Below are images of my visualizations:



In this visualization, the visual encodings I use to link data to visual channels are the color hue to represent the species type, and the unaligned vertical and horizontal position on the map, which indicate where a specific tree (dot) lies in the city.



This line graph uses aligned horizontal position on the X axis to show which year a species type was planted and aligned vertical position on the Y axis to show the average DBH of a species type. This graph also uses color hue to represent the species type.



Similar to the previous graph, this scatterplot uses aligned horizontal position on the X axis to show which date a species type was planted and color hue to represent a species type. The scatter plot also uses aligned vertical position to show the number planted of a species type that year and area to show the average plot size.

DESIGN CHOICES

Initially, I wanted to group the trees by the district they were in using their latitude and longitude. However, this was technically challenging because I wasn't sure what the latitude and longitude ranges of each of the 117 neighborhoods were. Instead, I decided to focus on how each species changed over time in terms of count and size. However, after processing the data, I noticed that there were over 200 species types. Creating a scale for the line graph and understanding the graph from the user's viewpoint with 200+ different colors would be a huge nightmare. Thus, I decided to focus on the most popular trees in SF and thought that 5 was a good number. Taking the top 5 trees is useful since it gives information on trees that are planted frequently (in comparison to a tree that is just planted 1 or 2 times), and the number is small enough for the user to keep track of when reading the visualizations.

When deciding which visualizations to create, I wanted to include a map, which could help a user understand where a tree type is planted across SF and if one species is densely populated in one area. Using a standard radius across all dots would allow the reader to focus on the location of the dot. The line graph and scatter plot shows the average DBH and plot area over time, helping the user understand how a popular species grew. I wanted to do a line graph for the average DBH over time since a line would be easiest for the reader to understand trends. In contrast, I wanted to do a scatter plot for the average plot size and number planted because there was less of a trend in the data, and plot size and number planted are correlated since they both relate to the amount of land area a tree species can. A larger radius representing a larger plot size is more intuitive. This scatterplot goes in hand with the map because the user can then picture the amount of land and where it is. One challenge I ran into with the scatter plot was the size of the radius. Correlating the radius with the plot area made the graph unreadable since the circles would overlap each other, so I had to find a number to scale down to make the dots small enough to make the plot readable and large enough to make the colors distinguishable and entities noticeable.

The colors across the visualizations remained the same to maintain persistence, and the colors representing each species contrasted from one another to make identifying and analyzing a particular species easier. Putting these visualizations vertically tied the story together. When the user first loads the page, they understand the current state of SF trees - which species are popular and where they are located. As they scroll, they can learn how each of these changed over time in terms of size and land area.