HW1 Submission

After spending some time looking at the dataset in excel and the console, I came up with a couple of different ideas for what I wanted to show. I really wanted to maximize the number of visual channels I was using to display a lot of different attributes about the dataset. Initially, I considered plotting the dimension of the tree's plot compared to the diameter of the tree, but after reviewing the data in the PlotSize column, I realized this would take a lot of cleaning. I also thought about the different ways I could use color as a visual channel to represent the data: breaking it up by diameter, species, even plant date.

Eventually, I settled on doing a map visualization of the data using longitude, latitude, diameter, and species to show something interesting about the data. I wanted to see how the diameters and densities of different trees varied and if their species played into that at all. I used the filtered data for some initial explorations in Python to figure out how many species I was working with, how many counts there were of each tree species, and which ones were the most common. This later informed my decisions to break the visualization into a few components.

I began by plotting all the trees as circles on the map of San Francisco and built a scale to scale the circles based on the diameter of the tree. The initial range of the scale was 0-9999, so I assumed the 9999 was a bad point and removed it from the data. Still, the visualization ended up showing me all circles of about the same size and one massive circle. After exploring the data more in Excel, I found that one tree's diameter was 200, while the average was around 10, so this was making the data act strange. I also ended up removing a couple trees at 125, so my new scale was ranging from 0 to 106. This allowed the differences in diameters to genuinely show up in my visualization.

I decided to use scaleOrdinal to represent the different tree species. Because there are 408, I knew I would not be able to give each one a new color, so this scheme was the best one to show such an expansive group of colors. After scaling my circles by diameter and coloring for species, I realized the map was a bit cluttered and difficult to glean information from as a static visualization. There, I decided to break it down further by the top 5 most common tree species in the dataset. I figured out which color each of those 5 were in the initial plot and used those colors to plot them in their individual maps for consistency. I ended up needing to adjust the range of the scales for a few of the species' diameter scales so the circles did not show up overwhelmingly large. These modified scales allow the users to observe the difference in diameters between trees.

I ended up breaking the visualizations down because they allowed me to examine something more interesting about the data than the clutter that all the data together shows. I kept the initial plot in there because I felt it was important to show the biodiversity and make up while providing a launching point for my data. Because I broke up the five most common species, the audience is able to perceive different patterns in where these trees are planted, the density with which they are planted, and about how large they are comparatively. This breakdown demonstrated that these trees are not scattered randomly, but that they all have individual journeys in San Francisco.