Homework 1 Report

Isabelle Breedveld, ikb22

With my data visualizations, I wanted to tell the story of the different tree species planted throughout San Francisco, and how they may differ in several dimensions, including geography, frequency, and size of the tree as measured by the standard unit of DBH, diameter at breast height. From the first bar graph describing the frequency of different tree species, it is clear that one species stands out as the clear dominant one, with a little more than 2,600 trees planted in total. Aside from this heavy outlier, there appear to be other tree species that are pretty frequently planted as well. Drawing from this graph, I ultimately wanted to investigate the top 15 most frequent species in order to uncover trends, patterns, and perform a more thorough analysis. The second bar chart depicts these top 15 species, ordered left to right from greatest to least frequency, and also color coded. To investigate the geographical and size differences, I created a map with dots referring to each tree planted, with the same color coding as in the bar graph. Moreover, the size of each dot corresponds to the diameter at breast height of the given tree, which can give a general idea of the sizes of tree species compared to others. From the map, it was clear that certain species tended to be larger than others. For example, the blue, pink, brown, and green dots appear to have much bigger dots than the other colors on the map, which correspond to the species, as referred to by their common names, Sycamore: London Plane, Red Flowering Gum, Blackwood Acacia, and Indian Laurel Fig Tree 'Green Gem' respectively. Additionally, there are some clear patterns to the planting of the trees. For example, the blue dots very clearly follow a straight line near the top of the map and also a more diagonal line cutting south. Other tree species follow a similar pattern, such as the light green dots near the top, and the light blue dots on the left side of the map. All that to say, interesting follow up questions to continue this investigation would be to look into the planting dates and if the patterns we see on the map have any correlation to year planted. Additionally, we could investigate different weather conditions in the different areas of the map, and see whether this impacts where the species are planted. For example, is there a specific reason why the blue colored dots are mostly concentrated in the north/northeast section of San Francisco?

In order to accurately generate the visualizations, I had to pre-process the data in various ways. The dataset that I started with was the one already pre-filtered by the professor to only list trees with full species names, those with latitude and longitude, valid site information, and ones with a DBH provided. I thought these filtering requirements would be in my best interests as I needed many of those fields in order to complete the visualizations. Next, to create the bar graph of species frequency, I had to find the species counts somehow. Admittedly, the easier way would have been to process the dataset even further using the python script or to find counts using SQL. However, I ended up creating a dictionary of each species and their counts. This resulted in 419 unique species in the dictionary. This number of species seemed like too many to visualize in one bar graph, so I processed this dictionary even further by filtering out those tree species with a total number of trees planted of 50 and less, which resulted in 99 tree species values, a number much more manageable for a bar chart. Once I created this chart, my next goal

was to find the top 15 most frequent species among the remaining 99 contenders. To do this, I created a list out of the filtered counts dictionary, and sorted them in order of greatest to least, and then took the first 15 values of this list. The last step was to convert this list back into a dictionary of the same form as the original, so that I could re-use the same function I had used for the bar chart previously. Lastly, when creating the map, I also added a "Position" field to the tree dataset using the projection function so that I could access the pixel locations of each tree on the map which I would use later with the data join to create the dot representations for each tree. Although I could have pre-processed the data in an earlier step (rather than in my script code), I found the dictionaries fairly easy to use.

As described above, my visualization consists of three separate graphs, two bar charts and one map of San Francisco. For the bar charts, the marks of the visualization are rectangles. In the first bar chart, "Frequency of Tree Species" (Figure 1), the visual channels are horizontal aligned position and vertically aligned length. The horizontal aligned position corresponds to the tree species, meanwhile the vertically aligned length corresponds to the number of planted trees for that specific tree species. The second bar chart, "Top 15 Most Frequent Species" shares the same visual channels as previously stated as well as another: color hue. The color hue changes according to the tree species, creating 15 distinct color hues that will be re-used in the map visualization. For the map visualization, the mark used was a dot (or circle) corresponding to each plant tree. The visual channels were aligned horizontal position and aligned vertical position, corresponding to latitude and longitude respectively. Similarly to the second bar chart, the visual channel of color hue was additionally used in order to differentiate between tree species type. The colors used for each species in the bar chart correspond to those used in the map.

All of my visualizations were selected and curated in order to facilitate the story that I outlined in the beginning of this document. First, the bar graph was an important preliminary graph to display the general frequency trends of species. As it turns out, there are several key tree species, and the rest have a pretty low frequency comparatively speaking. Thus, this graph is a great, simple way to show the reasoning to the viewer behind my design choice of zooming in on 15 of the top species in particular. Next, I wanted the map and the second bar chart to go hand in hand so that the bar chart would serve as a kind of legend for the map. The bar chart has the species (common name) labeled on the bottom and then the number of plants on the y axis as well as the specific color code for each species. I placed them side by side so that it is easy for the user to look up which color dots in the map refer to which species, without having to scroll. For the dots on the map, I decided to scale the radii from 1-10 so that the bigger circles wouldn't be too dominant and hide the other circles behind it. Moreover, I set the opacity of the circle colors to 0.8, so that they weren't fully solid in case of overlapping circles, which occurred quite frequently given the high number of trees in the dataset. I didn't set the opacity any lower because some of the colors were darker and lighter versions of the same color, like green for example, and I wanted the user to still be able to differentiate between these two. All in all, I believe that my visualizations help to facilitate the story that roughly 15 species dominate the

tree geography in San Francisco, and these species are sometimes arranged in clear lines, indicating there might have been a methodology behind planting these trees.