- Outlines the story you want to tell with the data

The goal of this visualization is to convey two pieces of information to readers: (1) to show how municipal trees are distributed around the San Francisco area; (2) to show the year every municipal tree was planted. To deliver the first information, I used the topoJson to draw the area map and used dots to represent the positions of every single tree on the map. To deliver the second information, I used different colors to distinguish trees planted in different year periods.

By reading the visualization, readers can tell most municipal trees were planted before 1955. Some trees were planted from 2000 to 2015 and there are very few newly-planted trees in the SF area. Also, we can see trees planted from 2000 to 2015 gathered in a few spots (the upper left and the middle left parts of the map), which potentially suggests these areas are comparatively new communities.

- Explains any processing you did of the data

After settling on visualizing plant dates of municipal trees, instead of diving right into the visualization, I took a close look at the 'PlantDate' column of the original dataset. According to the DATASET_INFO, if the tree was planted before 1955, it was marked as NA. Therefore, I first filled in NA with `'1/1/54 0:00'` following the format of 'PlantDate'.

Second, I utilized the str.split() function from Python to extract the year. One tricky thing about the 'PlantDate' is that it only gives the last two digits of the year, which would cause trouble if I wish to compare the age of trees later. Therefore, I added the string '20' to years that are less than 23 and '19' to others. If the year is '08', it will also convert to '2008' as I utilized the number of digits to handle such edge cases.

Third, I wish to assign a label to these municipal trees based on the year they were planted. I chose the gap of 15 years. In the end, all trees were assigned to one of the following labels: "Post 1955", "From 1955 to 1970","From 1970 to 1985","From 1985 to 2000", "From 2000 to 2015", and "Most Recent Years".

Finally, I added a new column to the original dataset named "PlantYear". The Python script I used to process data is in the folder labeled as "clean.py"; the cleaned dataset is named as "Street_Tree_List-2022-01-30_CLEANED.csv".

- Identifies the visual encodings you use in your image to link data to visual channels

Position: The position of each dot on the map was based on the longitude and latitude of the tree.

Color Hue: One of six different color hues was assigned to each dot based on its plant years.

- Provides a rigorous rationale for your design choices and explain how they help to facilitate the communication of the story you want to tell.

Basic Layout: After deciding that I want to show readers how municipal trees spread over the area of San Francisco, I deemed a distribution map the most effective visualization tool in this case.

Color: I want to utilize different color hues to demonstrate another dimension of this dataset, which is the plant year of municipal trees. When choosing colors, I wish the color could match the context of the dataset. In this case, I use dots to symbolize every single tree. Therefore, I applied the color of green and lime to show trees that were planted in earlier years. I applied colors of orange, yellow and brown to symbolize the color of the soil. Finally, I chose light blue as the color of the map to build a strong contrast between dots and the background.

Visual Channels: I used dots to represent each data point. Due to the large set of data points from the dataset, I reduced the radius of each dot to 1 to best avoid overlapping.