The purpose of the visualizations is to find out the factor that DBH is the most sensitive to. The DBH of a tree is the diameter at the tree's breast height, so we are interested in finding the factor that is the most important to a tree's size. The data provides several factors and attributes we can look into, such as the physical location, the legal status (Permitted or DPW maintained), the caretaker, and the plot size. In order to find the most sensitive factor among those, we need to visualize the relationship between each of these attributes with the DBH separately. The most sensitive factor will demonstrate a clear trend or distinction for different DBH in the visualization.

I modified the provided processing code for a more thorough data filtering. For the numeric attributes like latitude, longitude and DBH, I checked to make sure that the data values are valid numbers using a helper function is_float(). For plot size, there isn't a standard string format that all the values follow. But most of the values take on the form of "length x width" or "width xft". I filtered out all the data points that do not follow one of these two formats. The filtered data points are further modified by adding the position attribute, which projects the latitude and longitude into pixel positions. I updated the qCareTaker and qLegalStatus attributes to group all the small group sizes as "other". This prevents us from having too many colors on maps and pie charts, which can be a burden for users to process mentally. I added a PlotArea attribute that calculates the numeric plot area from the string value PlotSize. For PlotSize values of format "width xft", I assumed that the plot is a square of x ft by x ft. Lastly, I updated the DBH attribute by casting the string values into integer values.

The first set of visualizations is a map of SF with trees represented by the colored circles on the map and a pie chart representing the legal status distribution. The visual channels are the horizontal and vertical position, color, and radius of the circles and the areas in the pie chart. The position represents the physical location of the trees, so users can easily tell where the large trees are located. The color represents the legal status of the trees. This representation creates a connection between the map and the pie chart as their color keys are identical. So when users see these two visualizations side by side, they can easily make the connection between legal status and DBH of trees. The radius represents the DBH of the trees, so larger trees are represented by larger circles. This representation is intuitive for human understanding because we use area as a channel to represent physical sizes of trees. The large circles also catch more attention and users can immediately focus on the parts of the map with larger circles. Using pie chart area is an easy way of illustrating distribution, especially when a few groups take up the majority of data. The second set of visualizations is identical to the first set except that the colors represent the caretaker (or the owner) of the trees. Therefore, the rationale behind the visual channel selection is identical to the first set of visualizations. I intentionally chose to use the same visual channels because the goal of these visualizations is to find out what is the most sensitive factor to DBH. By keeping other aspects of the visualizations constant and only varying the variable of interest (which is the color channel), we can have a more clear understanding of which factor causes DBH to vary more. The last visualization is a scatter plot of plot area vs DBH. This visualization takes on a different form because plot area is on a continuous numeric scale rather than discrete categorical data, so it cannot be easily represented by color and pie charts. The visual channels for this graph are vertical and horizontal positions and colors of the circles. The position represents the relationship between plot area and DBH, and a direct relationship should result in a linear shaped graph. The color represents caretakers. This is because I have seen the correlation between caretakers and DBH from the second set of visualizations, and thus would like to further verify this relationship in the last visualization.

These visualizations convey an interesting information about tree DBH. From observing the physical locations of trees on maps, we can see that the large trees are located near each other and are distributed on only a few roads of SF. From the first set of visualizations, we can see that there isn't a clear correlation between the legal status and the DBH of trees. The majority of trees are DPW maintained, and there are both large and small trees of this category. However, the second set of visualizations is different. We see that most of the trees are private and only a small number of trees are DPW. So if there is no correlation between caretakers and DBH, we should expect to see more of the large trees are private than are DPW owned. As we can see from the map, this is not the case. Most of the large trees (represented by the larger circles) are DPW owned (represented by blue). Therefore, we can clearly see the correlation between caretakers and DBH of trees. From the last visualization, we did not see a direct relationship between the plot area and the DBH (which should be represented by a linear graph). However, from the distribution shape of the data, we can tell that the majority of trees have plot areas between 9 ft^2 and 100 ft^2, and almost all large trees fall within this range too. And we can see that most large trees (circles with higher vertical positions) are DPW (colored blue), which further verified our observation from the second set of visualizations.