HW1: Static Visualization Design

Curtis Xu, kx64

Story Outline

This visualization explores how different tree species—focusing on DBH (diameter at breast height) and Planting Year—can form “tree-like” shapes based on their statistical distributions. By placing multiple species on the same graph, viewers can compare how species differ in terms of DBH or planting dates. In addition, viewing the DBH and Year graphs side by side helps reveal links between a tree’s planting year and its trunk diameter by species.

Data Processing

In this step, I use Python with NumPy and Pandas library to clean, filter, and transform the data for easier loading in D3.

1. **Data Cleaning and Filtering**
    - Removed entries that lacked valid ‘DBH’ or ‘PlantDate’ values.
    - Reformatted PlantDate into the format ’%Y-%m’ for consistency and easier parsing.
    - For the ‘qSpecies’ column, shorten the names by keeping only the common name (split at "::" and take the last part).
    - Excluded overly general entries such as "Tree(s) ::".
2. **Data Subsets Selection**
    - After Data Cleaning, we have around 10,000 rows for data visualization.
    - Focused on the top 30 common species by frequency to ensure clarity of comparison in the visualizations.
    - Data outliners, such as trees with a DBH greater than 40 or planting dates before 1970, are excluded to enhance visualization clarity.

Visual Encoding

Box Plot with Violin Plot (for DBH)

- **Marks:**
  - **Paths** (Violin Plot)
  - **Rectangles & lines** (Box Plot)
  - **Circles** (Outliers)

- **Visual Channels:**
  - **Horizontal position (x-axis):** Each species occupies a vertical band.
  - **Vertical position (y-axis):** DBH (Diameter at Breast Height).
  - **Box Plot Components:**
    - The **rectangle** vertically spans from the first quartile (Q1) to the third quartile (Q3).
    - The **median line** inside the box shows the median DBH.
    - The **whiskers** (lines extending from the box) show data within a defined range (e.g., 1.5 * IQR or min/max)
    - Circles **outside** these whiskers represent outliers.
  - **Violin Plot Components:**
    - **Shape width** indicates the frequency distribution (thicker areas indicate higher frequency).
    - **Line length** represents the full data range, from smallest to largest.
  - **Color Hue:** Used for a tree-like simulation and does not represent additional numerical meaning.

Box Plot with Scatter Plot (for Plant Year)

- **Marks:**
  - **Rectangles & lines** (Box Plot)
  - **Circles** (Outliers and Scatter Plot)

- **Visual Channels:**
  - **Horizontal position (x-axis):** Each species occupies a vertical band.
  - **Vertical position (y-axis):** Planting year (time scale).
  - **Box Plot Components:**
    - The **rectangle** vertically spans from the first quartile (Q1) to the third quartile (Q3).
    - The **median line** inside the box shows the median planting year.
    - The **whiskers** (lines extending from the box) show data within a defined range (e.g., 1.5 * IQR or min/max)
    - Circles **outside** these whiskers represent outliers.
  - **Scatter Dot Position:** Each point's vertical coordinate reflects its planting date, while the horizontal position represents its species. A small jiggle is applied to prevent overlapping circles.
  - **Color Hue:** Used for a tree-like simulation and does not represent additional numerical meaning.

Design Rationale

1. **“Tree-Like” Shapes and Opacity Management:**

   The combination of box plots, violin plots, and scatter plots creates a silhouette reminiscent of tree forms. This design choice enhances the thematic connection to trees while reinforcing the statistical distributions visually. By controlling opacity for each mark, it will create a light, visual-friendly visualization to catch the audience's attention.

2. **Dual Views for Storytelling**:

   By separating the DBH and Planting Year into two graphs, viewers can compare distributions within each metric (e.g., how certain species cluster at particular DBH ranges or planting years) and then see if there is a visual connection between the two.

3. **Tradeoff, Clarity, and Comparability**:

   Because we have too many unique species in the dataset, focusing on the top 30 species avoids overcrowding, making differences between species more apparent. Meanwhile, aligning species along the x-axis and applying identical scales for y-axis metrics makes it straightforward to compare across species.