CAValli Team: Report 1

- Alessandro Longo – 5697430
- Vittorio Bartolomeo Secondin – 4798279
- Christian Dagnino – 4663694

Assignment 1 – Comparing Categories

1. Data Preprocessing [in Python]

This part involved the analysis of the overall dataset size and implied different choices to be taken in relation to possible inconsistencies in the given data and/or to the presence of null values for some of the columns. Specifically:

- **removing columns having an amount of null values ≥ 88%** of the total number of observations;
- **exploring variables**:
  - *state* → **filled null values with values inferred** from the corresponding entry in the *greater_metro* column;
  - *location_type* → null values **converted into *no_info* values** (one of the classes specified in this field);
  - *scientific_name* and *common_name* → taking into account all possible combinations of these two attributes being null at the same time or separately, we **inferred all missing common names** or **mapped distinct variants to one single common name,** we manually **corrected possible misspellings** and **removed meaningless values**, in the end we **removed all remaining observations with null values in both columns**;

- *tree_species* → a new column added to aggregate trees in **families, each one denoted by a unique name** (ex. oak, malus, pyrus, …);

- height_M → we plotted the distribution to **spot outliers, removed zero values and negative values**;
- diameter_breast_height_CM → we plotted the distribution to spot outliers, removed zero values and negative values.

2. Towards data visualisation [in Python]:

- **sketching data visualisations and exporting datasets** as .csv files:

  - bar chart with all cities;

  - specific bar chart w.r.t. each city;

  - stacked bar charts and small multiples w.r.t. cities in California, since that's the state having more trees than the rest of the states and more cities, in order to get a wider overview;

  - heatmap based on a dedicated dataset.

See the Python notebook in our repository for further details.

---

3. Website setting [in JS, HTML, CSS]

We designed the requested data visualisations. Specifically:

- **technical choices**, they include all the decisions about datasets we used for each visualisation:

- clearly stated in **subtitles of plots**;
- almost all decisions were **arbitrary and related to data preprocessing** (ex. selecting California as the reference state, setting the amount of the most widespread trees to be shown to 5 and not any other quantity, …).

- **stylistic choices**, they include the most aesthetic decisions about data visualisations:

  - choices that are not so pressing, such as the colours we picked up and applied to all stacked bar charts;

  - more relevant choices concerning:

    1. count and **average height** presented in the bar chart by means of a tooltip even if we suppose that the values may not be reliable;

    2. since the *others* tree type is so prominent and could hinder the incisive visualisation, we opted to **make it an optional field**;

    3. in the heatmap we chose the **greyscale to represent the density of tree types** because it proved to be more effective in highlighting the variations in comparison with other colour scales (ex. green scale).