CAValli Team: Report 2

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

Assignment 2 – Distributions

1. Data Preprocessing [in Python]

This part involved activities related to data cleaning and dealing with inconsistencies in data retrieved through an API call to introduce tree genus, family and order. Specifically:

- **searching for genus, family and order in a .json file** (returned by the API call) for each scientific name in the original dataset;
- **creating 3 dictionaries** mapping each of scientific names respectively to a list of tree genus, family and order values**;**
- **adding the 3 corresponding columns** to the original dataset**;**
- **instantiating some additional dictionaries** mapping each of tree genus, family and order values respectively to a list of scientific names associated with it.

See the Python notebook in our repository for further details.

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

We designed the requested data visualisation, the Sankey diagram.

Specifically:

- **technical choices**, they include all the decisions about the dataset we used:
  - we decided to **subdivide states in North and South** categories to simplify the visualisation;
  - in order to decrease the size of the resulting Sankey diagram and make it more readable, we decided **not to include every state and every city of our dataset and instead select only the most important ones** containing more trees (top N states/cities);
  - the same approach was adopted in relation to **tree families in the last layer** of the Sankey diagram (top N tree families).
- **stylistic choices**, they include the most aesthetic decisions:
  - we implemented a feature that allows to **click on a node of the Sankey diagram** and highlight all links getting into it;
  - each node/link shows **a tooltip** when we hover on it with the mouse, **providing information about the amount of trees** in that node/link.