Museum of Contemporary Art Chicago

DS3500: Advanced Programming with Data
Prof. Rachlin

Introduction

In this assignment we survey the artists represented at the Museum of Contemporary Art (MCA) Chicago. The attached dataset is a JSON file containing over 10,000 artists whose works are displayed at the museum. Your task is to generate Sankey visualizations using the plotly python library that show connections between artist nationality, the *decade* when they were born, and their gender. Which countries, decades and genders are most represented? You may reuse, modify, or extend any code presented in class. For example, you should feel free to modify the parameters to the **make_sankey** function we develop together in class.

Specific Tasks

1. Convert the data into a Panda's dataframe containing three columns: nationality, gender, and the *decade* the artist was born. For example, if they were born between 1940 and 1949, you need only store the decade as 1940. Make sure the decade is stored as a float or int, not a string!

2. Aggregate the data, counting the number of artists grouped by both nationality and decade.

3. Filter any rows where the decade is 0 (presumably unknown) or where there is missing data.

4. Filter out rows whose artist count is below some threshold. You'll want to experiment with this value to produce a visually appealing visualization. I suggest trying a starting threshold around 20.

5. Generate a Sankey diagram with nationality on the left (sources) and decade of birth on the right (targets)

6. Repeat steps 2-5, but this time count the number of artists grouped by nationality and gender. Your Sankey diagram will show nationalities on the left and gender on the right.

7. Repeat steps 2-5, grouping by gender and decade. Your Sankey diagram will display gender on the left and decade of birth on the right.

8. Extend the Sankey functionality we developed in class to take an arbitrary *list of columns* instead of just two specific column parameters. Now you can create multi-layered Sankey diagrams! Demonstrate your code by including a Sankey diagram with all three layers: nationality, gender, and decade of birth (in any order). Note: It is ok if some of the "nodes" of a given category do not line up. This is an artifact of the plotly Sankey layout algorithm. But your output should be *readable*, not a cluttered jungle of nodes and edges that are impossible to distinguish. (This is probably the most challenging part of the assignment.)

9. Write a ½ to 1-page interpretation of your results. What insights do you glean from your visualizations? Reflect on what your visualizations tell us about diversity, inclusion, and bias in the art world.

Considerations

A. When implementing steps 5-7, consider ways to make your code more reusable so that you aren't scripting the same tasks repeatedly. (Some of these steps should probably be consolidated into their own function!)

B. In general, as we transition from intermediate to advanced-level programming, we want to be thinking not only about functions and modularity, but the construction of packages and libraries. You should isolate visualization functions that have the potential for future reuse in separate python files (viz.py? sankey.py?) and then *import* these libraries into your main assignment code.

C. Your code will be graded on correctness, efficiency, readability, documentation, and modularity, reusability, and your text analysis. Use of ChatGPT or other online resources to help you code is acceptable so long as you site your sources. Use of ChatGPT to compose your essay is not allowed. Avoid ChatGPT whenever you are expressing yourself in writing. *Use your own voice!*

What to submit

Submit your python code as python (.py) files, not a Jupyter Notebook! Also include your three bi-level sankey diagrams, one multi-layered Sankey diagram, and your written analysis.