Data visualization: Personal report

Joris LIMONIER

January 12, 2022

Contents

1 Sankey diagram by Joris LIMONIER

1.1 Introduction

1.1.1 Get the project

My submission for the project of data visualization consists of a Sankey diagram. It is available here [1] and should be viewed in the browser (tested on Chrome), on a computer. The code for the project is posted on GitHub.

1.1.2 Background on Sankey diagrams

Sankey diagrams are defined as follows: "Sankey diagrams are a type of flow diagram in which the width of the arrows is proportional to the flow rate." [1] In our case, we have four columns that are linked by flows, representing the number of albums or the number of songs.

1.1.3 Code structure

The diagram was created with the following structure:

- Python [2] for data pre-processing (cleaning, formatting, writing).
- CSV for the data relative to each passage from one column to another (4 columns, therefore 3 CSV files).
- JSON to group the CSV data and put them in a format that can be fed to the D3.js library.
- JavaScript (Vanilla) to perform last-minute grouping and filter-relative updates
- D3.js [3] for the interactive data visualization.

Let us first detail the data preprocessing part.

1.2 Data pre-processing

1.2.1 Raw data

We grab the following raw data from the CSV Wasabi datasets that were downloaded.

- Album field
    - _id (album id)

[1] If the link is dead, go to https://jorislimonier.github.io/, navigate to "Projects", then look for "Collaborative Data Visualization"

  - id_artist
  - genre

- Artist field
  - _id (artist id)
  - type (*e.g.* “Person”, “Orchestra”, “Group”, ...)
  - gender (male, female, unknown)

- Song field
  - id_album

1.2.2 Clean the data

The number of NaN values varies significantly form column to column. The Id columns don’t have any but some columns have many, making them sometimes tedious to work with. Overall however, the dataset is pretty clean. Barely a few Id’s are misformatted and some genres have unexpected characters.

Let us call “column transfer” (*CT*) the passage from one column to another. This is what we feed into D3.js and this is the part with the most impact on the diagram.
We clean the data in *sankey.py*, which contains one class per *CT* :

- *TypeGender* : goes from the artist type to the gender of the artist.
- *GenderAlbums* : goes from the gender of the artist to the number of albums produced.
- *AlbumsSongs* : goes from the number of albums produced to the average number of songs per album.

The same file also contains the following class:

- *Sankey* : contains several functions to go from the individual CT classes to writing the data in the final format.

Each of the *CT* classes writes a CSV file. Usually, Sankey diagrams have three columns:

- source : the origin node for this flow
- target : the target node for this flow
- value : the flow quantity (*i.e.* the number of elements that go from a given source to a given target)

The *TypeGender CT* does indeed have these three columns. The *GenderAlbums* and *AlbumsSongs CT* 's on the other hand, are composed of a fourth column:

- genre : the genre of the album

This genre column allows to perform the filter operation.

1.2.3 Transfer the data

The *Sankey* class has a $write\_final\_data$ method that takes the three CSV's and writes the format in a JSON file (sankey-genre.json), in the format demanded by D3.js. The format is as follows:

- nodes:
  - index
  - name
- links:
  - source
  - target
  - genre
  - value

where the set of the node indices is simply a bijection between the unique node names and $\mathbb{N}_{\geq 0}$.

1.2.4 Set up filtering

Filtering is performed in *sankey-filter.js*. We allow for a filtering that only takes place for the two last *CT* 's (*i.e.* *GenderAlbums* and *AlbumsSongs*), because the *TypeGender CT* concerns artists, not albums (and obviously, artists don't have a genre).
We use JavaScript for the filtering part because we need to update the graph when genres are added/removed from the filter list. On each change of the selection dropbox (using an *eventListener*), we go through sankey-genre.json and check for each element in the "links" list if it is in the list of genres selected by the user. If so, we group it with all other elements going from the same source node to the same target node and sum their values.

1.2.5 Group genres

The grouping of genres is performed in *sankey.py*. It contains a dictionary-like object that is of the following structure:

- keys: the new group name.
- values: a list of strings. If the old genre contains one of these strings, it will be assigned to the new group which is named after the *key*.

1.3 Data Visualization

1.3.1 General intention

This visualization tries to stay close to Schneiderman's mantra: "Overview, Zoom and Filter on demand" [4]. It appeared however that the zoom feature didn't really find any meaningful usecase for this Sankey diagram, which is why it has been omitted.
Table 1 details the actions available for each of the general task categories.

	User task	Details
1.	Overview	The artists gender per artist type The number of albums per artist gender The number of songs-per-album per number of albums
2.	Selecting	Drag and drop nodes to rearrange them Hover over the links to flow value, source and target
3.	Filter	Only show data for one or more album genres

Table 1: User tasks for the Sankey Diagram.

1.3.2 Overview

Figure 1 shows an overview of what the default view looks like. The boxes on top show column names. They could probably be inferred from the node labels, but it is meant to facilitate the User Experience.

1.3.3 Select

If the user wishes to rearrange the nodes from each column, they can drag and drop them. This allows to isolate a single segment of the data for a temporary visualization, *e.g.* if someone wants to see how many female made 3 vs 4 albums with 2 songs/album. The other informations can be dragged to the bottom to disregard them. Figure 2 shows such an operation. Note that this is more of a "quick-and-dirty" way of visualizing that more

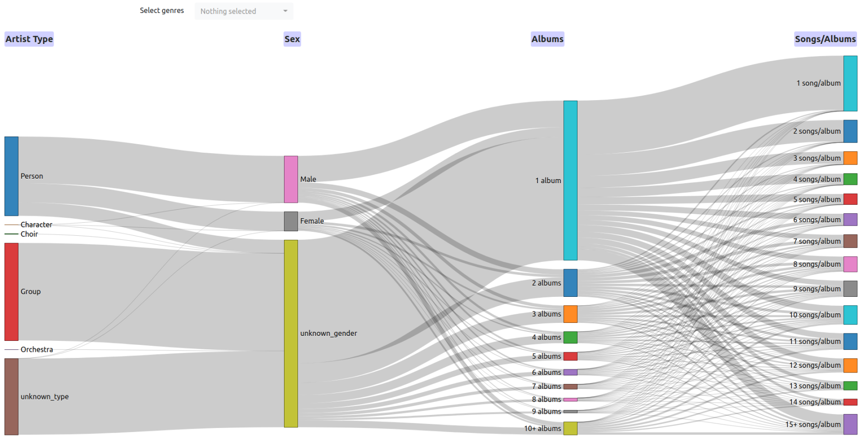


Figure 1: Overview of the default Sankey diagram.

females made 3 albums than 4 albums, rather than a real use case for most of users.

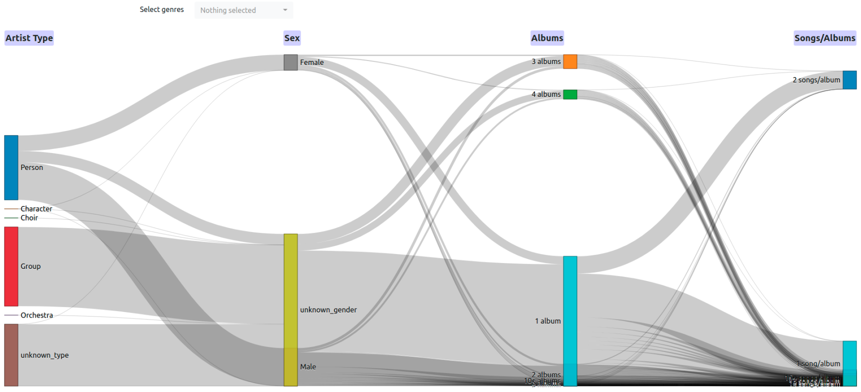
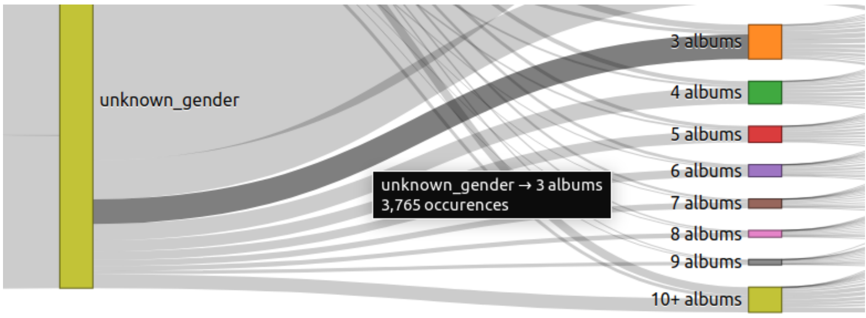


Figure 2: Display of the “quick-and-dirty” isolation

Moreover, if the user wants to have more information on one specific link, *e.g.* knowing how many artists of unknown gender produced 3 albums, this can be done by hovering over the link. Then a tooltip appears (see Figure 3), showing complementary information in the following format:

$$\text{source} \rightarrow \text{target}$$
$$n \text{ occurences}$$

where n is the link value from the given source to the given target.

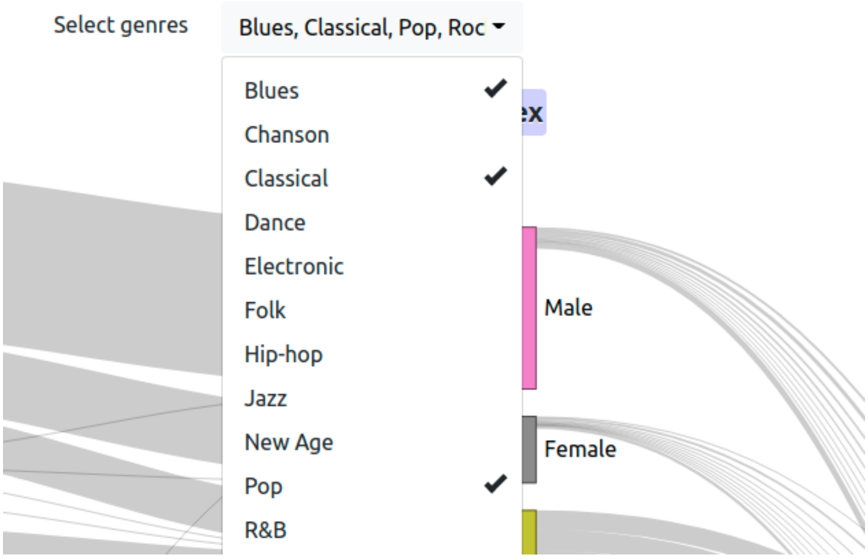


unknown_gender → 3 albums
3,765 occurences

Figure 3: Display of the tooltip on hover

1.3.4 Filter

Subsequently, the user can choose to show the data only for a few genres. The selection dropdown is displayed in Figure 4.



Select genres: Blues, Classical, Pop, Roc

- Blues ✓
- Chanson
- Classical ✓
- Dance
- Electronic
- Folk
- Hip-hop
- Jazz
- New Age
- Pop ✓
- R&B

Figure 4: Display of the filter options where “Blues”, “Classical” and “Pop” are selected

When the user clicks to add/remove a genre from the list, the diagram is recomputed and redisplayed on the page. The width of the links update accordingly, as shown in Figure 5. As mentioned previously, only albums have a genre (artists don’t), which is why the left-most *CT* remains untouched after filtering.

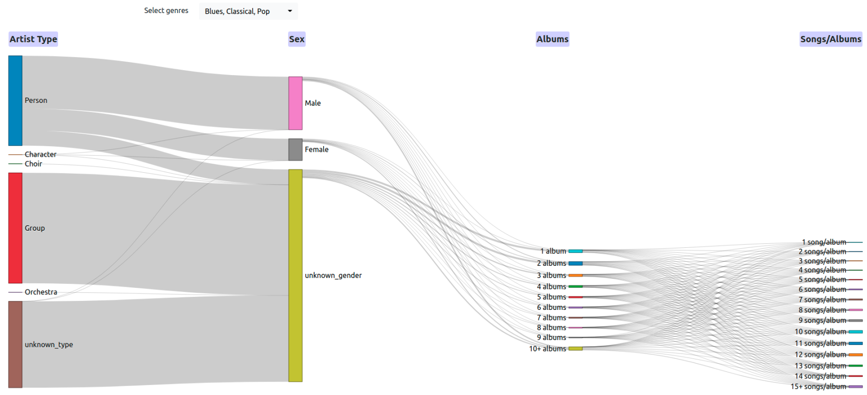


Figure 5: Display of the diagram where “Blues”, “Classical” and “Pop” are selected

1.4 User test

1.4.1 Perform tasks

The user test was performed on 3 testers who agreed that their information is being gathered. They were aged 22-24, with a Bachelor’s degree in a scientific discipline.
The users were asked the following questions:

1. Task 1
   (a) How many men made 5 albums.
   (b) On a scale from 1 to 5 (1 means very easy, 5 means very difficult), how hard was that task?
2. Task 2
   (a) How many rap solo artists are women.
   (b) On a scale from 1 to 5 (1 means very easy, 5 means very difficult), how hard was that task?
3. Task 3
   (a) Can you tell how many solo artists made 3 albums? Why/Why not?
   (b) On a scale from 1 to 5 (1 means very easy, 5 means very difficult), how hard was that task?

Task 1 and 2 were performed successfully by the participants, who found it rather easy to complete (one of them mentions that some time was needed to get familiar with the diagram, since he was not familiar with them).

The third task is a trap since there is no *CT* that allows to give this information. 2 participants identified that this was impossible, the third one said that the were unable to perform the task (therefore they didn't understand that the task was undoable).

1.4.2 Debriefing

The users reported that overall, they enjoyed interacting this visualization (3/3) and that they would recommend it to a friend (3/3). Their least favorite part of the data was that the "unkown_genre" genre is overrepresented. The way to improve this issue is either to reduce the number of missing values in the data set (but this is a data collection matter, which is not my main concern), or to better group the genres. This can be achieved by two ways: grouping more genres under the ones already present, and adding more grouped genres.

One user mentioned that they would have liked to see smooth animations when selecting other genres. This is probably doable through better knowledge of D3.js, but since this was my very first experience with JavaScript at all and since the time was constrained, the final version of my Sankey diagram doesn't contain transitions.

References

[1] Sankey diagram. Page Version ID: 1062562883. URL: https://en.wikipedia.org/w/index.php?title=Sankey_diagram&oldid=1062562883.

[2] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[3] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D³: data-driven documents. 17(12).

[4] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343. ISSN: 1049-2615. doi:10.1109/VL.1996.545307.

2 Homework for 20/10/21

A paragraph describing the users.

Users could be people insterested in the music industry, who want to compare men and women artists for sociological interpretation.

The list of visual tasks supported by users and the visualization goals.

User task	**Details**
Overview	Flow between columns
Zoom	TBD
Filter	Filter by genres

The list of (raw) attributes you will need from the WASABI dataset you are going to use.

The following attributes will be needed:

- Album field
  - _id (album id)
  - id_artist
- Artist field
  - _id (artist id)
  - type (*e.g.* “Person”, “Orchestra”, “Group”, “Choir”, “Other” or “”)
  - gender (male, female, unknown)
  - members (check this one, it may be the name of the members of a band)
- Song field
  - id_album
  - genre

The informal description of the processing of the raw data in order to make it to fit in the visualization technique. This might include calculated variables you must add in the process.

Clustering Artist - *type* variable to make *is_band* boolean variable

The name of visualization technique and the name of the member of the group who is going to implement it. Associate the visualization technique with the visual goal.

Sankey diagram with the following columns:

- Single artist or Band
- Male, Female, Unknown
- Number of Albums
- Number of Songs

A visual mapping of variables available in your data set (after data processing) and the visual variable available in the visualization technique you have chosen.

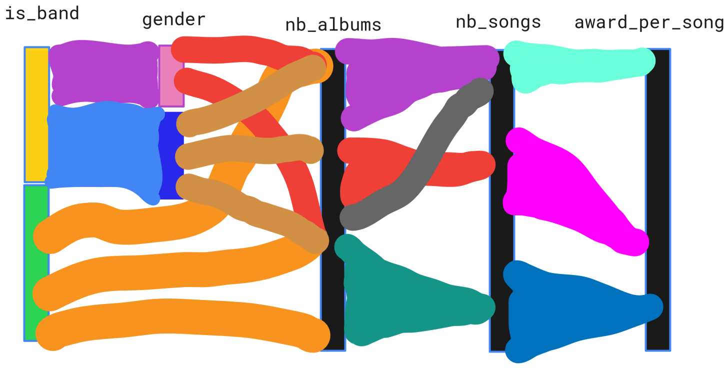


Figure 6: Representation of the Sankey diragram (JL)

3 UX Protocol

TODO:

1. Write text that will be read (to remove bias in the way I say things)
2. Test app on some people (minimum 3)

3.1 Presentation & training

1. Ask written consent for recording
2. Tell people why they are here and what we will do

3. Give the following information:

   Age

   Sex

   Highest level of education reached

4. Present Sankey diagram and explain what a Sankey diagram is.

5. Ask if anything is unclear?

3.2 User test

1. Task 1
   - (a) How many men made 5 albums.
   - (b) On a scale from 1 to 5 (1 means very easy, 5 means very difficult), how hard was that task?

2. Task 2
   - (a) How many rap solo artists are women.
   - (b) On a scale from 1 to 5 (1 means very easy, 5 means very difficult), how hard was that task?

3. Task 3
   - (a) Can you tell how many solo artists made 3 albums? Why/Why not?
   - (b) On a scale from 1 to 5 (1 means very easy, 5 means very difficult), how hard was that task?

3.3 Debriefing

1. What are your three favorite feature?

2. What is your three least favorite feature?

3. Would you recommend this application to a friend?

4. What would you do differently?