Problemset

The data directory contains:

- the top 200 charts of 28 March 2023 for a selection of countries (`charts_xx.xlsx`)
- basic track metadata (`tracks.csv`)
- track lyrics (`lyrics.csv`)
- artist metadata (`artist.json`)

Exercise 1

- Study all datasets provided.
- What are questions that could be analyzed using these datasets?
- Specifically, what are questions that could be analyzed by suitably combining and consolidating all datasets into one?
- How should such a combined data set look like: What defines a single row? What are the relevant key columns that can be used to join the different data sets?

Exercise 2

Consolidate all datasets into one. Identify "problems" in the data in terms of data quality and think about possible solutions. In particular, consider aspects such as:

- Merging
- Reshaping/pivoting
- Missing values
- Type conversion

Exercise 3

Answer questions such as:

- What is the total number of streams of all chart songs?
- What is the average danceability per country?
- How many songs of the German charts contain the word "love"?
- Identify rows or columns with (many) missing values. What are possible ways of handling these cases?
- What are the top genres of the German charts? (Difficult! You need to make several assumptions on the way)