

Name: \_\_\_\_\_

Consider the database which holds all the music in Apple Music. Brainstorm possible separate data tables (rectangular dataframes) which might exist in the database.

- Come up with at least 3 different tables (the three tables should all have different observational units (i.e., row types)).
- For each table, describe the observational unit (that is, what is a row?).
- For each table, provide at least four variables (columns), some of which could be used to join the data tables.
- Indicate which variable(s) would be used to link the data tables.
- Rank order the tables from most rows to least rows.

Solution:

**songs** (each row is a different song) with columns: song name, length, **artist name**, **album name**

**albums** (each row is a different album) with columns: **album name**, number of songs, **artist name**, playing time, producer, genre

**artists** (each row is a different artist) with columns: **artist name**, age, gender, number of top Billboard hits, number of albums

**songs > albums > artists**

A few notes:

- For a table where a row is an artist, there shouldn't be a column called "album". Why not?
- Great to use column names that differentiate the variables. For example, instead of "title" use "album title" and "song title."
- Although both the songs and albums data tables include "artist name", we couldn't (or wouldn't want to) join on that variable. Why not?