WS #5 - Joining

Monday, September 15, 2025

DS 002R - Jo Hardin

Name: _____

Names of people you worked with: _____

Share one or two adjectives that reflect your state of being at this moment.

Task: Consider a database which holds all the music in Spotify. Brainstorm possible separate data tables (rectangular data frames) which might exist in the database.

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

Solution:

1. [**songs**] (each row is a different song) with columns: song name, length, artist name, album name
2. [**albums**] (each row is a different album) with columns: album name, number of songs, artist name, playing time, producer, genre
3. [**artists**] (each row is a different artist) with columns: artist name, age, 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?