Group 6 ETL Project

Max Izotov, Julissa Guzman, MaryLouise Pabello

November 4, 2020

Data Sources:

Our team obtained data from Kaggle to Analyze the Top Spotify Tracks of 2017 and 2018.

Data	Data Type	Data Source
Top Tracks of 2017 Spotify Data	csv	https://www.kaggle.com/nadintamer/top-tracks-of-2017
Top Tracks of 2018 Spotify Data	csv	https://www.kaggle.com/nadintamer/top-spotify-tracks-of-2018

The dataset can be used to analyze the song patterns from 2017-2018 and analyze what audio features make them popular.

Extraction:

We obtained the data from the Kaggle website as csv. These csv files are included in the repository, in the " Resources" data folder.

Transformation:

Our original data had the following columns: **id** (Spotify's unique URL for each song); **name** (a song's title); **artists**; **danceability** (a Spotify-determined numeric value for how suitable a song is for dancing); **energy** (a song's perceived intensity and activity); **key** (the song's musical pitch); **loudness** (measured in decibels); **mode** (a music term that identifies the song's musical scale as either major or minor); **speechiness** (how much spoken word is present in a song compared to instrumental); **acousticness** (a Spotify confidence value of a song's being acoustic [lacking any electronic amplification such as auto-tune]); **instrumentalness** (identifies if a song has no vocals present); **liveness** (identifies the presence of an audience in the song); **valence** (a Spotify-determined measure of a song's musical positiveness [will a person's emotions be happy when listening to a song]; **tempo** (a musical term that measures a song's speed and is measured in beats per minute); **duration_ms** (the length of a song in milliseconds); and **time_signature** (Spotify's estimate of a song's meter, a musical term that denotes how many beats are in each measure of a song).

We were only interested in keeping the columns name, artists, danceability, energy, loudness, speechiness, liveness, tempo, and duration_ms, and used Pandas to create new data frames for the cleaned data ("cleaner_2017" and "cleaner_2018"). We also renamed the columns "name" and "duration_ms" to "song_title" and "duration_min" to better reflect the information in those columns (a song's title and its duration in minutes instead of milliseconds). And then converted the remaining numerical columns (danceability, energy, loudness, speechiness, and liveness) into percentages so the value of each numerical measure was easier to understand. Our final data frames are called "cleaner_2017" and "cleaner_2018".

Load

The data was loaded into two tables in a PostgreSQL database called "ETL_project". The two tables are called "data_2017" and "data_2018".