Regana Alicka & Lauren McKinzie : Spotify Data ETL

Almost 250 million users use Spotify every month. Part of Spotify’s pull is their method of recommending songs based on a user’s listening history that accounts for song features such as valence, danceability, energy, and speechiness. Since Spotify is so hugely popular, it is useful to look at what song features are associated with the most commonly streamed songs on the service. Spotify’s API and the spotipy python wrapper make it possible to access and organize Spotify data that can then be imported into postgres for analysis.

A few difficulties occurred as we tried to clean the data. We first attempted to upload the CSV file in postgres, however the song titles included apostrophes and quotes continually causing errors in the upload process. We moved over to upload the data set using pandas, which ultimately was a success, but involved a unique looping code to replace the apostrophes with commas. See here: df["Track Name"] = df["Track Name"].map(lambda x: str(x).replace("","").replace(""","").replace(",","")) - This code ultimately allowed us to successfully upload our code into postgres. Another obstacle we encountered was using the unique syntax of the spotipy library and Spotify API to extract the correct information we wanted. In order to make the necessary API calls, we had to isolate the Spotify IDs that each track is indexed by within Spotify’s API by splicing the ID number off the end of the Spotify URL associated with each song.

We choose postgres because our data was structured for SQL storage. Our data was already organized in tables.