ETL Project

Brice Maynard and Megan Rose

For our ETL project we decided to examine music data using a CSV file from Kaggle with the top songs on Spotify from 2010-2019 and by scraping Billboard’s Top 100 Songs, Top Streaming Songs, and Top Radio Songs from 2010 to 2019.

After scraping the webpages from Billboard’s website, we needed to modify some of the data in the CSV files to get rid of “\n\n” in front of the songs and artists. We did this modification with Pandas and re-saved the fixed data as new CSV files. After fixing the 3 Billboard files we setup tables in Postgres for each of these files and the Spotify data file that we pulled from Kaggle. We had errors importing the data from Kaggle into Postgres and after trying several different ways to fix it, we realized that we needed to resave the CSV file as UTF-8.

If we were running and examining the data, we would have cleaned up the “Genre” column on the Spotify table to make all the genres consistent across. We also would have set up primary keys in the Billboard and Spotify tables to consist of the song and year to make it unique.