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” that appeared at the beginning and end of each artist and song title. We did this modification with Pandas and re-saved the fixed data as new CSV files. The “\n\n” prevented the data from showing up when the CSV files were opened, and the data did not show up in SQL when imported into tables. We also dropped the “record” column when saving new CSV files as this column was unnecessary.

Following the fixing of the 3 Billboard files, we set up 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. Once we re-saved the CSV file, we were able to import the data into the Spotify table. We decided to load the data as tables in SQL since the CSV files had headers and it was easier to read in the tables format.

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 throughout the years. 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. That way when examining data, it would have been easier to compare top songs on each of the Billboard lists to see how they compared to other top lists.