ETL Project

Brice Maynard and Megan Rose

**E:** For this project we decided to gather and examine music data over the last decade. Kaggle had a CSV file with Spotify Top Songs 2010 – 2019, but more data was necessary. We discovered the Billboard website had information available, however there was no API or file to easily download. In order to get the information needed, we had to use Beautiful Soup to scrape several webpages to gather the data. There were three main webpages that we scraped: Top Streaming Songs, Top Radio Songs, and Top Songs. Each of the main webpages had to be looped through several times in order to scrape the data. The base URL changed when a different year was selected so this had to change with every loop and break after the year 2019 was scraped. After the data was scraped and in a dataframe in Pandas, we exported the dataframes to CSV files.

**T:** Once the data was collected, we started to transform the data so it could be imported into Postgres as easily as possible. If we opened the Billboard CSV files in Excel, the Artist and Title did not show up. However, if the file was opened in Notebook, the Artist and Title did show up. We later discovered, after printing in Pandas, that the data had “\n” and “\n\n” at the beginning and end of the fields. The data was modified in Pandas to drop these unnecessary characters and exported the data to CSV files yet again. Additionally, the dataframe generated column of 0-X was dropped in Pandas before exporting the CSV files.

**L:** Following the cleaning of the Billboard datasets, tables were setup in Postgres for each of the CSV files. Four tables were created: one for the Kaggle Spotify data and 3 for the Billboard files. Importing the Spotify CSV file that was downloaded from Kaggle proved to be troublesome. Every time the file was imported it produced the error “invalid byte sequence for encoding UTF8.” It was determined that there was an accent on the e in Beyonce’s name that was causing the error. After trying several failed modifications in Postgres, the CSV file was resaved as UTF-8. This simple fix allowed the file to be imported into Postgres with no further issues. Since the files had headers and were already in “table” format, Postgres was used as this was the easiest way to get the data into a querying tool.

**Notes:** If a full analysis of the data was to be completed, the “Genre” column on the Spotify table would have been cleaned using Pandas. There were several different “pop” genres in the data which would make analysis difficult. For instance, there was pop, dance pop, Australian pop, etc. In Pandas, the data would have been modified to categorize like genres to one main category. Primary keys would have also been established in the four tables that were created. The primary keys would have been a combination of the song and year. A primary key on song was initially done when the tables were created, but it was discovered at least one song was a top hit in 2010 and 2011. Making the primary key a combination of the song and year would have eliminated any record importing errors.