**Project2**

Meredith Gray & Morgan Blanchard

Project 2 revolves around ETL: extracting, transforming, and loading data. For this project, we wanted to find the most popular songs in the United States then compile data so we can dive into their musical characteristics. To complete this task, we found two different datasets. One of them includes Spoitfy’s top songs and the second dataset includes musical traits about each song including danceability, key, tempo, duration, valence, acousticness, and energy to name a few. This information can be used to answer questions such as: do the most popular songs have a certain tempo range, is the danceability high or low for the top 50 songs, or are popular songs more likely to have a shorter duration?

Our first dataset is Spotify top songs from 2017-2021 and can be found here: <https://www.kaggle.com/dhruvildave/spotify-charts>

Our second dataset contains the musical data as well as links to the songs from soundcloud and can be found here: <https://www.kaggle.com/josshhh/spotify-songs-with-soundcloud-links>

These datafiles are too large to be uploaded onto Github. Therefore, if you need to run the code it will be necessary to download the files from the links above. The first data file will be stored in the Data folder and named 'charts.csv'. The second data file will also be stored in the Data folder and will be named 'Soundcloud\_data.csv'. The second data file does not download as a csv so it will need to be converted before the code will run.

**Extract and Transform**

The big picture of this portion of the project includes downloading the datasets from Kaggle, uploading them into a jupyter notebook, cleaning each dataset, merging the datasets based on one criterion, then cleaning the final dataset to display the information we want.

During the extraction, we downloaded these two datasets and read them into our jupyter notebook using pandas. Once we could look at the data in pandas we were ready to transform the data into readable, useable merged datasets.

The first dataset (charts.csv) originally includes nearly 20 million rows. To clean the Spotify data we filtered the data to only include top songs from the past two years (2019-2021), only showed songs that were charted in the US, got rid of any rows with missing data, and narrowed to the top 50 songs charted each day. After performing those manipulations, there were still 2.5 million rows of song data. Many of these remaining rows were duplicates as the top songs were pulled each day between 2019-2021. To limit the repeated songs, we decided to only keep the songs that were charted on the first day of each month. This left us with 1700 rows of data and we felt satisfied to move on to cleaning the second dataset.

The second dataset included songs from Spotify, a link to the song on Soundcloud, as well as a variety of musical data and song data particular to Spotify, such as number of streams. This dataset included 17000 rows with 19 columns of data after we removed rows with missing data and columns with extraneous data (can be found on charts.csv) such as track artist. In order to merge these data frames, we had to rename the song title column to match because that is the criterion we wanted to merge on.

After the merge we dropped any rows with incomplete data, repeating song titles, and remade the index.

**Load**

To load the data onto MongoDB, we created a connection, then inserted our csv data into a dictionary that can be stored as a database. The code printed a line of code if the load went successful and a different line of code if something didn’t work. Once the code ran, we verified everything was loaded into MongoDB successfully.

**Conclusion**

Performing the ETL process is a valuable tool while learning to navigate data analytics. This can make data more consolidated for viewing and storing data. We have successfully extracted data about the most popular songs, transformed that data into a readable and meaningful dataset, then loaded it into another database for storage.