ETL Project Proposal

Quality versus Quantity:

Grammy Winners vs. Billboard Top 100 for 2009-2019

Group Members: Julia Headlee, Jaime Jimenez, Alicia Perez

The Billboard data we are extracting will be coming from a database downloaded from Kaggle <https://www.kaggle.com/danield2255/data-on-songs-from-billboard-19992019>. The Grammy data will be scraped from their website <https://www.grammy.com/>. Alternatively, if the Grammy website does not provide full nomination information for previous years, we will use Wikipedia to pull previous year nomination data. We are looking to pull most of the categories for the Grammys and we will convert that into a table with the Grammy year, song title, artist, award category, win/loss, and any other category we find necessary.

We will transform the data using Pandas and taking only the data from 2009 through 2019. There are several columns in the Billboard data that we don’t find necessary and will be dropped. For example, there are writing credits, features, and lyrics columns. We currently do find those useful for the hypothesized analysis and they will be dropped. There are also many duplicate entries as it is a weekly list, so we likely be doing a group by to make the data more concise. That will depend on other circumstances we discuss.

We will load the data into Postgres. The tables made in Pandas/Jupyter Notebook will be put into Postgres, and we will load the data from there. We will have to make a composite key with the song title and the artist as we don’t want to overwrite or mess with any data that has the same song title from different artists.