**ETL Project Report**

Aim: To create tables to analyze the Grammy awards data to show correlation between the billboards top 100 and the Grammy awards.

Data Extraction:

Kaggle was used as the main data source. The billboard data was from billboard charts and the Grammy awards data was from the Grammy awards file. Both files were in CSV format.

Data Cleanup: (Transformation)

Clean up included

* Renamed the song column to Song/Nominee and dropped an unnamed column from the billboard data frame.
* Renamed the column artist and nominee to Artist and Song/Nominee respectively, dropped published\_at, title, updated\_at, workers and img columns from the grammy data frame.
* Merged the billboard and the grammy data frames using the merge function in pandas.
* Dropped the duplicates from the merged data frame.
* Used group by function to calculate the sum of the weeks a particular song or an artist was on the billboard chart and created separate data frames for popular artists, popular songs and awards.
* Renamed columns on each of these data frames so that they are more applicable for creating a relational database tables in SQL.
* In pgadmin, created tables popular artists, popular songs and awards weeks to hold the data.

Loading:

We used a relational database postgres sql to load our tables. We created a connection to the pgadmin using an engine from the pandas to populate the values into the 3 tables. Our tables include 1. Popular artists, 2. Popular songs and 3. Awards.

Application of these tables for future analysis:

1. These tables can be used to show if songs, artists, and albums that win Grammy awards actually spend any time in the billboards top 100.
2. The Grammy awards are voted on by the industry, musicians, musical engineers, musical artists, music producers, record labels etc.
3. Billboard top 100 is made up of songs that are funneled by consumer driven applications like record play, streaming, music downloads etc.