etl_group_project
Week 13 group project

This project was completed by Clay, Mac and Don

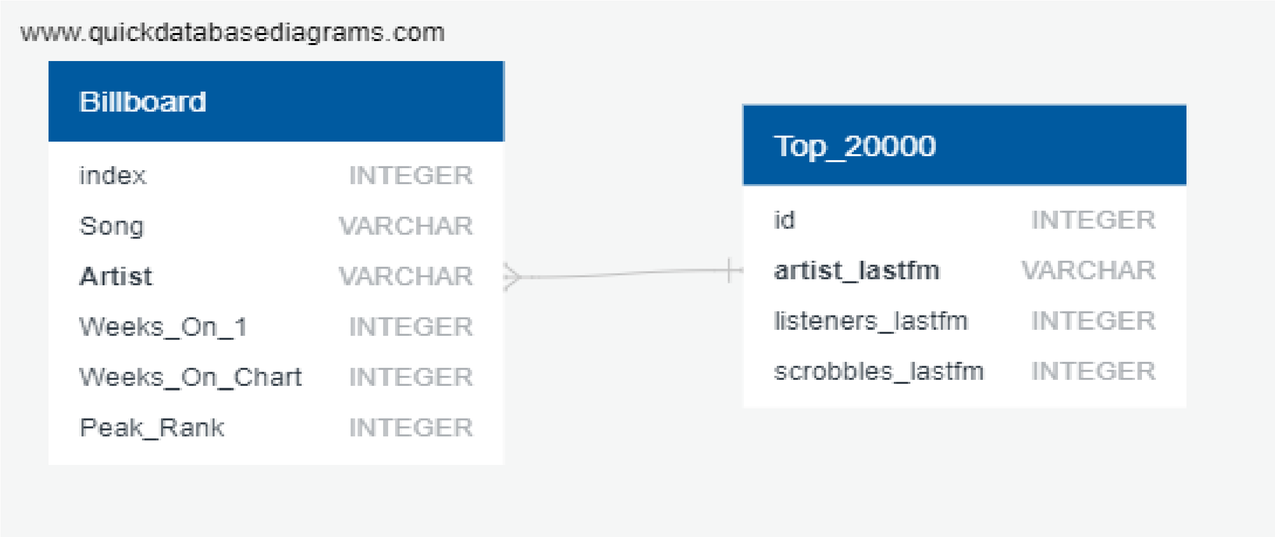
The project delivers a database with data about music artist and the number of listeners and scrobbles they currently have as well as data on the

We used 2 different datasources, both from Kaggle. The Billboard Top 100 data came from https://www.kaggle.com/saberianz/billboard-charts and the "listener and scobble" data came from https://www.kaggle.com/pieca111/music-artists-popularity.

The download of the music artists database incuded data on all artists that have at least one listener or scrobble (streaming event) and contained well over 1 million artists. Since this file was too large to load into our github repository we created code in the combined_code jupyter notebook and limited the data to the Top 20000 artists sorted by # of listeners. You can view the original data set here: https://drive.google.com/open?id=1ouK0pTZxhyrcRDj_yxiKvQ2ZPnA5uwcZ.

As part of the initial data transformation we noticed that some artists were included in the file more than once based upon the artist have more than one country in the original country_mb column. We excluded the duplicates and eliminated non-useful columns by creating a pandas dataframe and then converting to a .csv file that was then was uploaded/pushed as top_20000.csv into our github repository.

The database schema was then defined using quickdatabasediagrams.com with that code exported to postgresSql "QuickDBD-Music_artist.sql". The schema shows in the "QuickDBD-Music_artist.png" image file.



www.quickdatabasediagrams.com

Billboard	
index	INTEGER
Song	VARCHAR
Artist	VARCHAR
Weeks_On_1	INTEGER
Weeks_On_Chart	INTEGER
Peak_Rank	INTEGER

Top_20000	
id	INTEGER
artist_lastfm	VARCHAR
listeners_lastfm	INTEGER
scrobbles_lastfm	INTEGER

The tables were created in postgreSQL and then loaded.

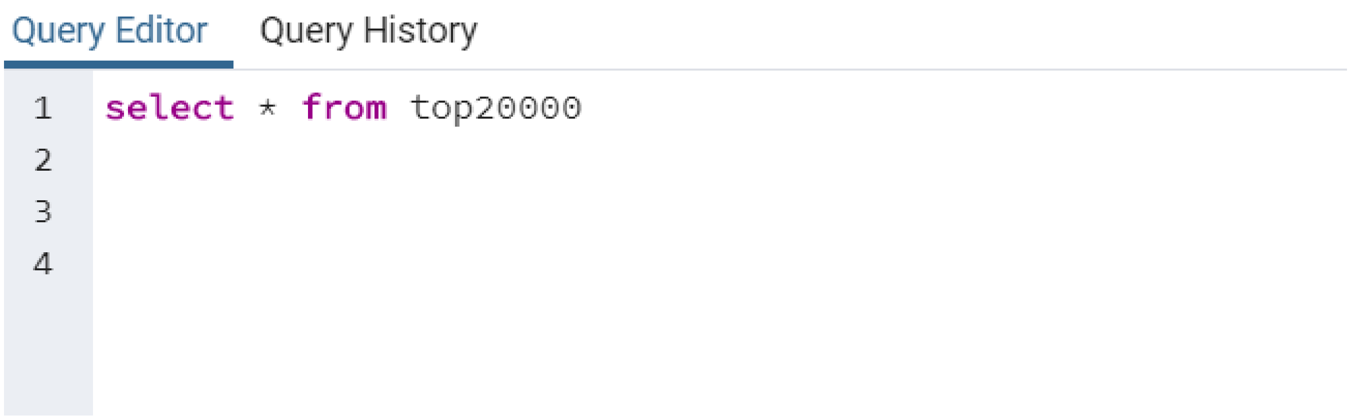
Extract artist.csv **Transform - Top_20000.c**	**Load into Table** **top20000**

DataSource: https://www.kaggle.com/pieca111/music-artists-popularity.		Pandas dataframe > .csv file	
mbid		id	id
artist_mb	removed - same as artist_lastfm		artist_lastfm
artist_lastfm		artist_lastfm	listners_lastfm
country_mb	removed		scrobbles_lastfm
country_lastfm	removed		
tags_mb	removed		
tags_lastfm	removed		
listeners_lastfm		listners_lastfm	
scrobbles_lastfm		scrobbles_lastfm	
ambiguous_artist	removed		
# of records = 1,466,083		records = 20,000	records = 20,000

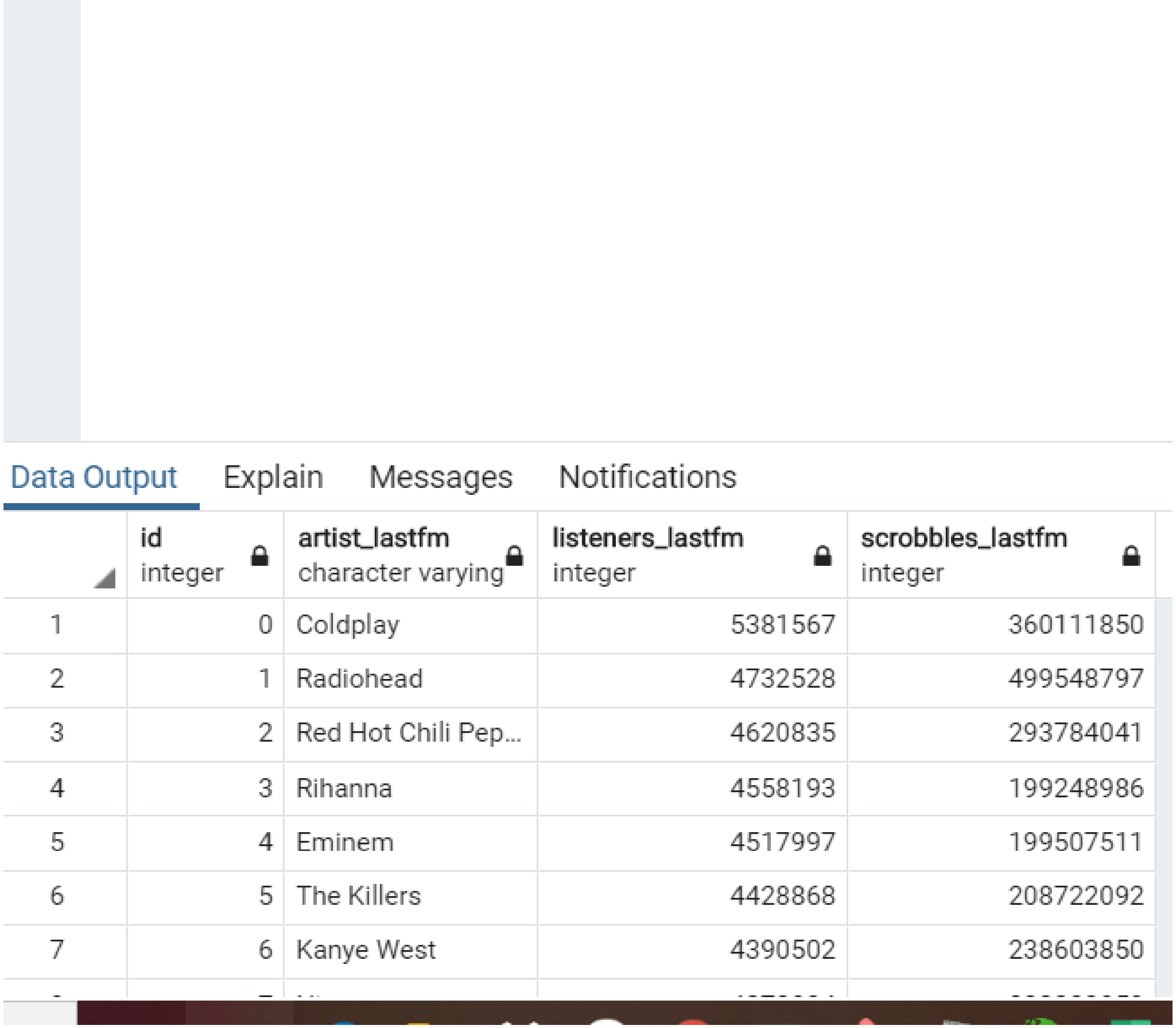
Extract: Billboard.csv			**Load into Table top20000**
https://www.kaggle.com/saberianz/billboard-charts			
id			id
Song			Song
Artist			Artist
Weeks On #1			Weeks On #1
Weeks On Chart			Weeks On Chart
Peak Rank			Peak Rank
# of records = 1951			# of records = 1951

Below is the resulting screenshots of the database tables

top20000

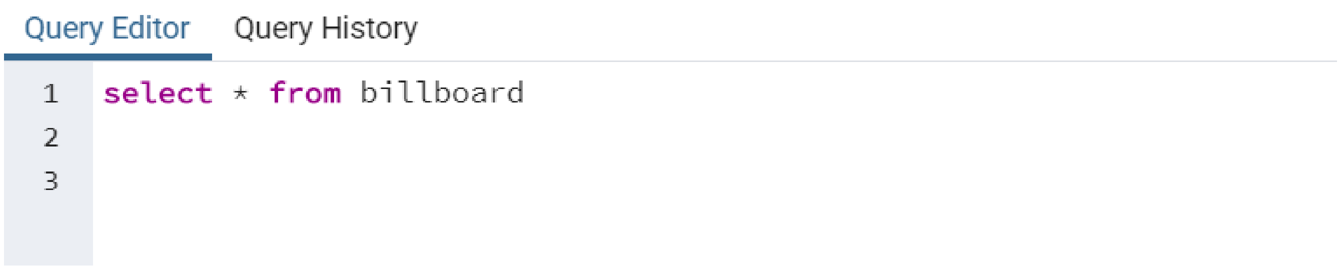


```
select * from top20000
```

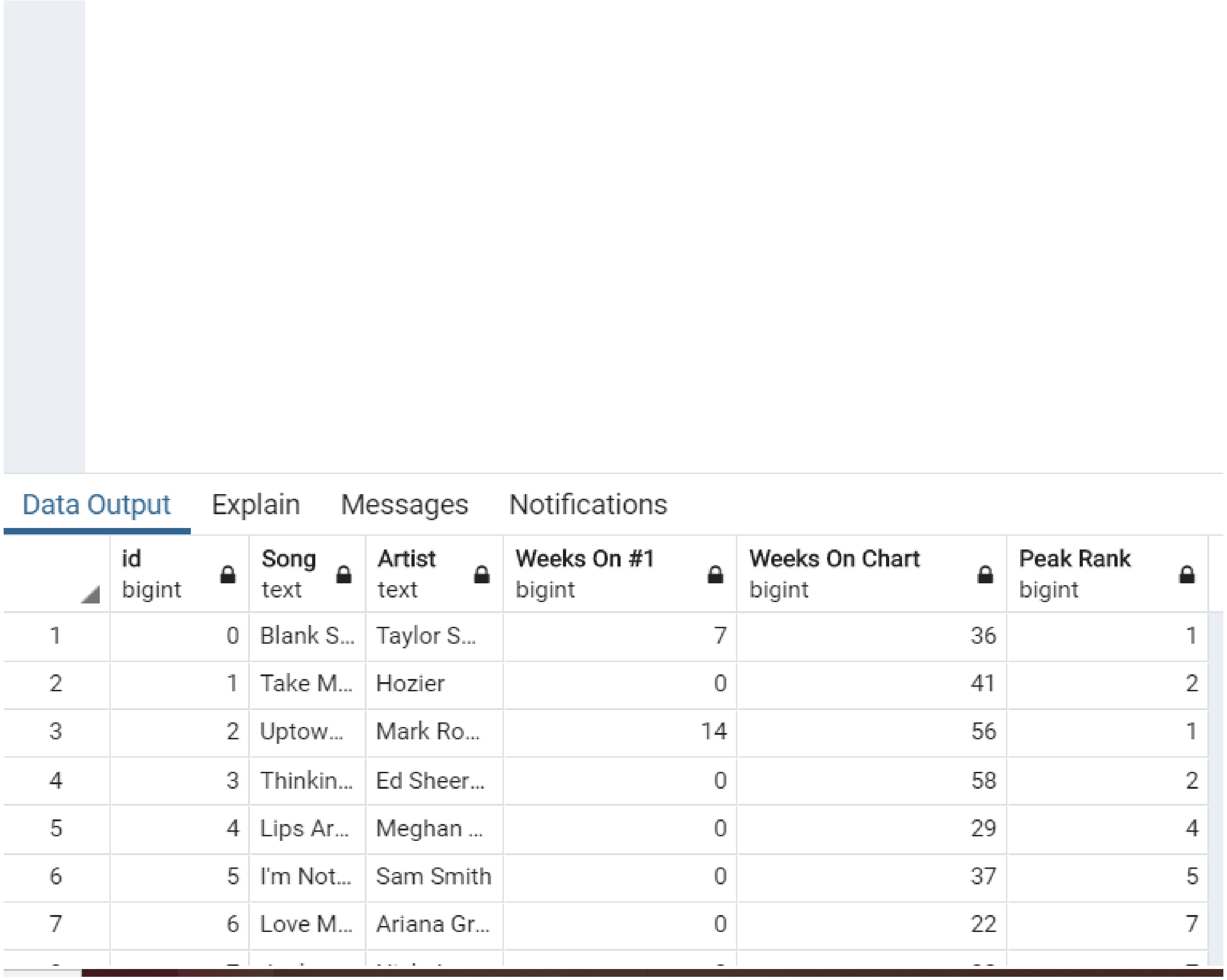


	id integer	artist_lastfm character varying	listeners_lastfm integer	scrobbles_lastfm integer
1	0	Coldplay	5381567	360111850
2	1	Radiohead	4732528	499548797
3	2	Red Hot Chili Pep...	4620835	293784041
4	3	Rihanna	4558193	199248986
5	4	Eminem	4517997	199507511
6	5	The Killers	4428868	208722092
7	6	Kanye West	4390502	238603850

billboard



```sql
select * from billboard
```



	id bigint	Song text	Artist text	Weeks On #1 bigint	Weeks On Chart bigint	Peak Rank bigint
1	0	Blank S...	Taylor S...	7	36	1
2	1	Take M...	Hozier	0	41	2
3	2	Uptow...	Mark Ro...	14	56	1
4	3	Thinkin...	Ed Sheer...	0	58	2
5	4	Lips Ar...	Meghan ...	0	29	4
6	5	I'm Not...	Sam Smith	0	37	5
7	6	Love M...	Ariana Gr...	0	22	7