**Top Songs of the 2010’s**

Fox Aufenger

Anne Ramer

Bridget Reardon

Extract

* Extracted Kaggle data set Top Spotify songs from 2010-2019 By Year
  + <https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year>
* Read the CSV in Jupyter Notebook from Billboard and Insider websites
  + <https://www.billboard.com/charts/decade-end/hot-100>
  + <https://www.insider.com/best-songs-every-year-2017-8>

Transform

* We experienced trouble with the UTF-8 coding from Kaggle. We tried using the both encoding='mac\_roman' and encoding='cp1252' to read the CSV into Pandas (pd.read\_csv in pandas). We found encoding=' ISO-8859-1'.
* Top 100 billboard- difficulty with marketing ads and how to extract text from div (find errors) and , bs4.element.ResultSet errors when trying to use pd.read\_json. We ran each item as its own for loop, then merged into one data frame.
* Sorted dates by year
* From the Insider article, all text was delivered as one string. We had trouble separating the string as some titles were missing quotes. We used the split function.
* Construct tables inside PgAdmin
* Dropped the years we did not want before merging the tables
* Convert strings (years) to integers in columns

Load

* Created a relational database in SQL
* Using SQL, performed a union to merge the data into one dataframe.
* Using Pandas, (pd.concat) we merged the data into one dataframe.
* Pandas returned all of the songs, while SQL dropped one of the songs.

Complications with process

* We wanted to demonstrate various data types combined and had trouble finding a corresponding csv file. Thus we are not able to list from top to bottom (#1, #2) but rather general top songs of the year from various sites
* One of the songs was dropped from our dataset while trying to perform union from the CSV to the JSON file

Additional Ideas

* Sorted combined dataframe to show all duplicates to compare years and sources
* With more time, we would like to determine the cause of the dropped song
* Find if any titles or artists are shared between sites / frequency
* Find a way to combine billboard list to df in one for loop