Lab 2 Data Proposal

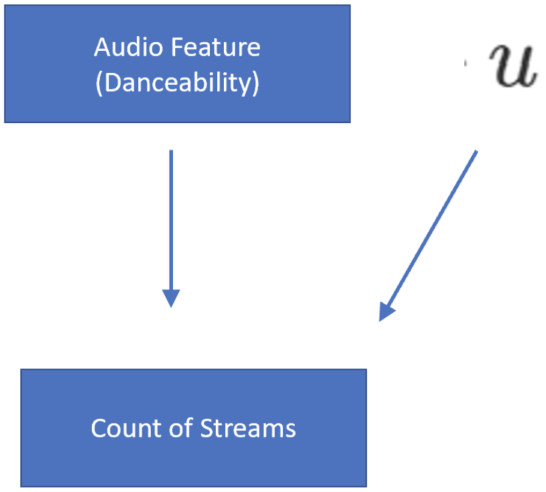
Saniya Lakka, Megan Martin, Andrew Sandico

Introduction:

As part of the data scientist team for Acme, Inc, we are supporting the product Spotify to maximize the number of streams by identifying the key audio feature that causes the amount of streams. Since Spotify monetizes by usage (how many times a song is played and how long a user stays on the app), our work is motivated to identify the Spotify feature that causes the most streams.

Approach:

To operationalize our work, we will be leveraging (or creating) a data set from Spotify that identifies audio features for songs. Our conceptual model that explains the causation of more streams is that the specific audio feature of danceability would have the largest R2 to predict number of streams.



Short Model : Streams = B1 danceability + u

Danceability as defined by spotify: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

Control variables: genre, accousticness, energy, instrumentalness, liveness, loudness, mode, speechiness, tempo, and valence.

Spotify provides a description of these music features here: https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features

Categorical variables to include: artist, genre

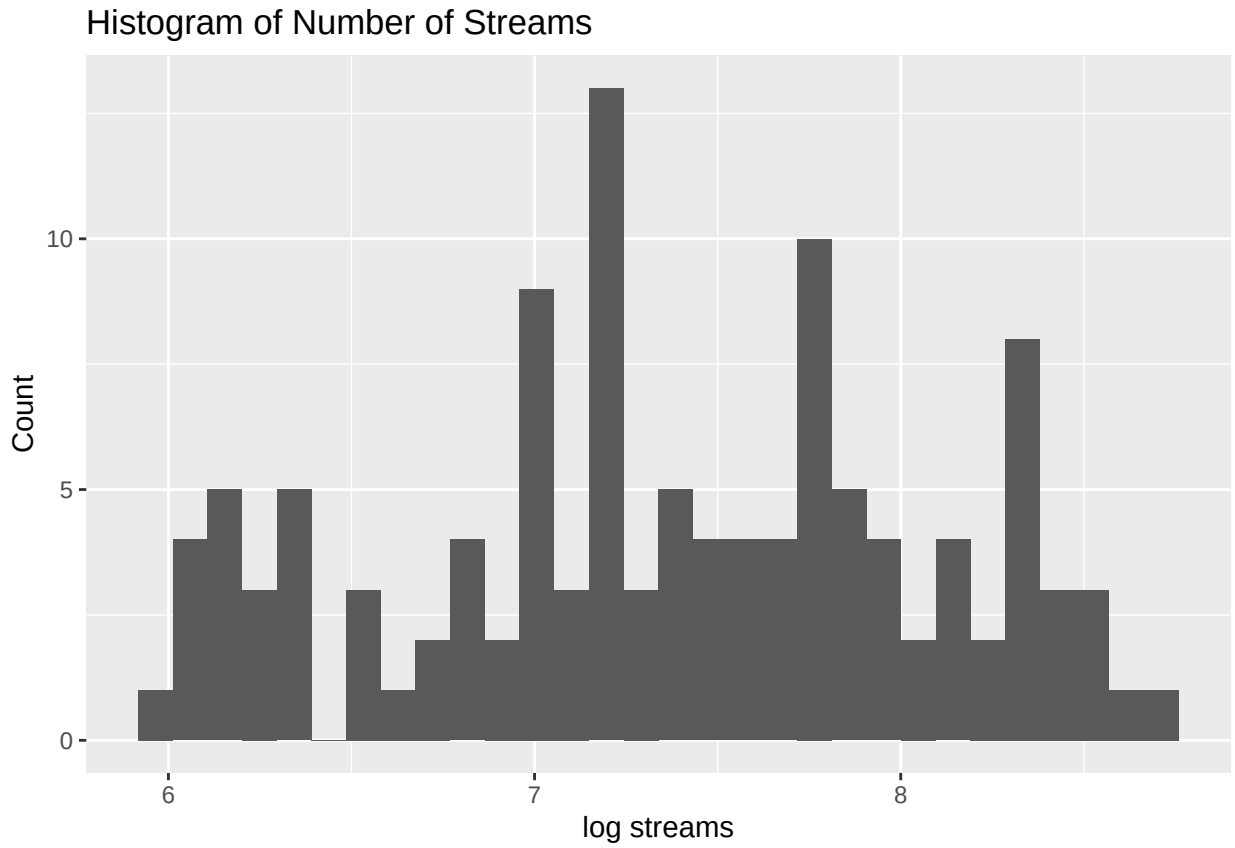
Initial Data Exploration:

We combined two datasets for this proof of concept. The first dataset contains attributes of songs from 2017 and was obtained from here. The second dataset contained streaming counts for the top 200 songs by

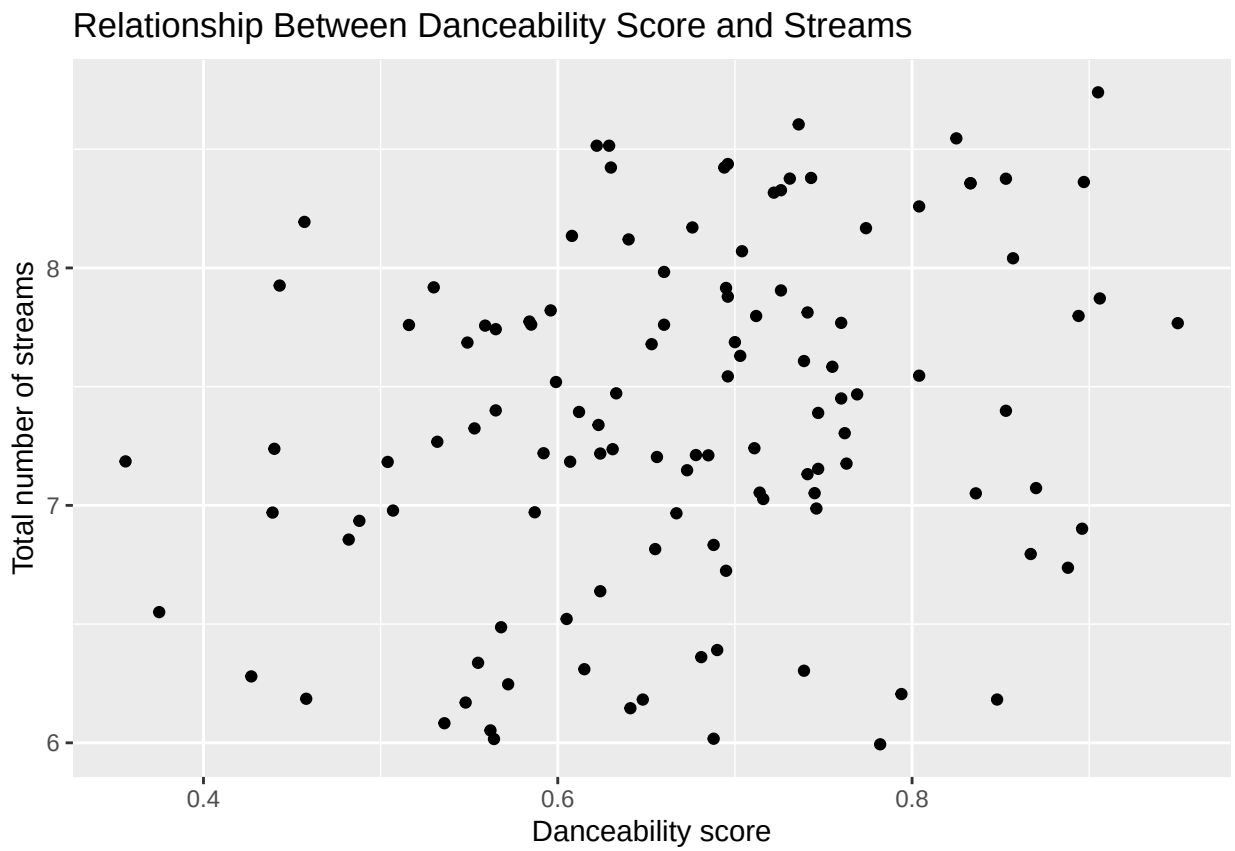
week from 2010-2017 and was obtained from here. Because the song streaming dataset contained weekly streaming numbers, the dataset was collapsed by song to provide a total streaming count by song. Finally, the streaming count dataset was joined with the song attribute dataset. The final dataset contains 118 songs from 2017 with streaming totals and music attributes for which this analysis can be conducted. The column names are listed below:

```
##  [1] "acousticness"     "danceability"     "duration_ms"      "energy"
##  [5] "instrumentalness" "key"              "liveness"         "loudness"
##  [9] "mode"             "speechiness"      "tempo"            "time_signature"
## [13] "valence"          "target"           "song_title"       "artist"
## [17] "Streams"          "id"               "date"             "sum_streams"
```

Plotting the log transformation of the total streams provides the following distribution:



Finally, plotting our proposed primary feature (danceability) by total streams provides the following distribution:



After approval of this dataset, we will attempt to pull attribute data for songs directly from the spotify using the API.