Project 2 Proposal
Ashton Cho, Deric Liang, Wes Morberg
https://github.com/UC-Berkeley-I-School/Project2_Cho_Liang_Morberg

Overview of dataset

The dataset we propose to use is a unique list of the most streamed songs on Spotify in 2023, up to July 14. The dataset is named 'spotify-2023.csv' in our Github repository (note: the data contains some encoding issues - the file 'proposal_data_exploration.ipynb' contains code to address these issues without compromising the data). The dataset contains several categories of information about the songs:

- Basic information: song name, artist, release date,
- Streaming information: streams, number of playlists the song is present in (across multiple platforms),
- Music information: BPM, key, mode, and metrics of song characteristics (measuring instrumentation percentage, speech percentage, danceability percentage, etc.).

Overview of Final Report plan

Using this data, we plan to answer the following guiding question: What are some of the common characteristics of the songs that have been popular this year? This guiding question informs the sub-questions listed in the "Overview of variables for exploration" section below. We may combine this dataset with the following information, but would appreciate guidance on whether the scope is large enough without the supplemental data:

- Genre data to inform questions on how genre relates to song popularity and variables which may influence popularity
- Billboard data to observe correlations with billboard chart placement
- Data from tunebat.com to gather more information about a song, such as duration, album format, loudness, explicitness, or label.

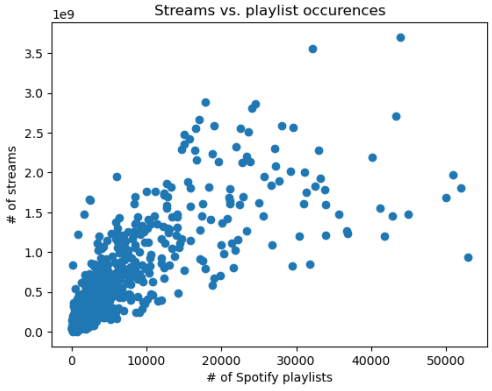
Overview of variables for exploration

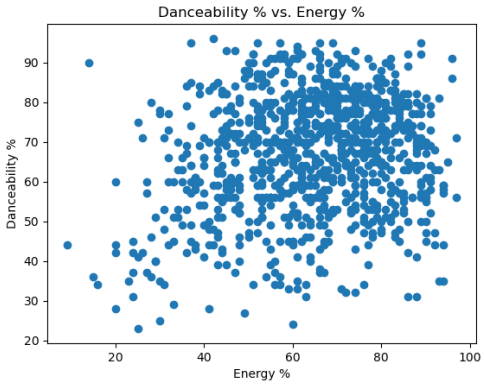
Below, we outline some of the variables in the data for which we plan to explore. We describe the information in each variable, and sub-bullets outlining the questions associated with each variable.

- *artist_name*: Name of the artists on the song.
    - What is the count of artists with multiple hits?
    - Are there any differences in the number of Male/Female/transgender artists (if we can find data on this or have time to manually enter the data)?
    - Can we join genre data onto a song based on the artist?
- *streams:* The number of streams the song has on Spotify.

    - Using *artist_count* (the number of artists on a song), do we find that more popular songs tend to be more collaborative?
    - Do we observe a correlation with *bpm* (beats per minute of a song)?
    - Combining *key* (music scale label) and *mode* (music tone - major or minor) into a *key_final* variable, do we observe any *key_final* categories with a higher stream count, or is the distribution uniform?
    - Do we need to think about converting this to streams per month to account for varying release timings?
- Combining *key* (music scale label) and *mode* (music tone - major or minor) into a *key_final* variable, are there any *key_final* categories that are more or less common, or is the distribution uniform?
- *released_year*: The year the song was released.
    - Do we observe outliers? We would expect most songs to have a 2022 or 2023 release date.
- *released_month* and *released_day*: The month and day a song is released.
    - Do we see non-uniform distributions? If so, what are some possible explanations?
- *danceability_%* and *energy_%*: Measures the danceability and energy level of a song.
    - Are these variables substitutable in terms of information provided?
- *instrumentalness_%* and *speechiness_%*: Measures the instrumental and speech level of a song.
    - Would it be correct to hypothesize that more popular music has a higher *speechiness_%* than *instrumentalness_%* due to the popularity of rap music?

Initial plots, figures, and tables





	artist_count	released_year	released_month	released_day	bpm
count	953.000000	953.000000	953.000000	953.000000	953.000000
mean	1.556139	2018.238195	6.033578	13.930745	122.540399
std	0.893044	11.116218	3.566435	9.201949	28.057802
min	1.000000	1930.000000	1.000000	1.000000	65.000000
25%	1.000000	2020.000000	3.000000	6.000000	100.000000
50%	1.000000	2022.000000	6.000000	13.000000	121.000000
75%	2.000000	2022.000000	9.000000	22.000000	140.000000
max	8.000000	2023.000000	12.000000	31.000000	206.000000

key	count
C#	120
G	96
NaN	95
G#	91
F	89
B	81
D	81
A	75
F#	73
E	62
A#	57
D#	33

mode	count
Major	550
Minor	403