# Project 2 Proposal

## Overview

Music plays an important role in the lives of people all around the world, evident with Spotify's 345 million monthly active users and 155 million subscribers. People's music preferences vary, and they listen to music for a variety of reasons. Among the three of us, our music choices are diverse and can be measured in various ways to show our similarities and differences in how we like our music. Do the music tracks and artists we have in common have similar audio features, and are these audio features popular among highly reviewed albums?

## Names of team members

- Anand Patel
- Andrés de la Rosa
- Lina Yang

## Name of team's GitHub repository

https://github.com/UC-Berkeley-I-School/Project2_DeLaRosa_Patel_Yang

## Primary Datasets

We will have two primary Spotify datasets. We will be accessing our own Spotify libraries through the Spotify API. The data set will include tracks, artists, albums, and their corresponding audio features, such as tempo, loudness, danceability, acousticness, etc. The audio features and descriptions can be found on Spotify's developer documentation here or summary in section AudioFeaturesObject: Audio Features on Spotify Track.

We already pulled the first 100 tracks from each of our libraries; extracted the song's audio features, artist info, and track information; and saved them to CSVs. The variables in these datasets can be found below. These data sets will be labeled by team member name and combined.

We will also be accessing the Spotify API to get albums and their audio features that are found in the Pitchfork (P4K) CSV mentioned below. We will be focusing on albums from the Pitchfork data set from December 6th 2016 to December 6th, 2017 only. This is the first 1260 lines of the P4K data set.

## Supplemental Dataset

A CSV sourced from Kaggle with Pitchfork reviews from January 5th, 1999 to December 6th, 2017. Source here.

## Initial Plots, Figures, Tables

- With our own music libraries, we plan on to:
- Briefly describe our data. Showing its shape, how many tracks, amount of different artists each member has on its playlist, average time per song.
- Grouped bar plots comparing each team member's musical preferences by corresponding audio feature. For example, on average, Anand listens to music with more acoustics than Lina.
- Scatter plot comparing our average team musical preference with what is popular on Spotify.
- Histograms and radar charts show the distribution of audio features in each of our members' music tracks. We will assign numerical ranges to explore whether our music preferences vary

and whether we like more "boring" music or more "lively" music based on an equation that involves music features, such as energy, danceability, tempo, and loudness.

- With the pitchfork reviews, we plan to:
- Heatmap of correlations between P4K score, best music label, audio features from Spotify. Already shared in Heat Map of Combined Spotify/P4K 2017 Dataset.
- Scatter plot of albums by score vs. popularity. This partitions the data set into 4 parts based on relation to mean score and popularity.
- For each of 4 partitions, plots showing genre breakdown and audio feature characteristics vs. the score.
- Group data by genre, sort by score. Compare the top albums for each genre by score against their audience-centric audio features like valence, danceability, or popularity.

**Exploration of Variables and Expected Insights**

In the Our Personal Music Libraries Analysis, some of the features we will explore are: acousticness, danceability, energy, instrumentals, liveness, loudness, speechiness, tempo, and valence. The values of each of these features range from 0.0 to 1.0. A higher score for each feature indicates a higher likeliness for the track to have those feature characteristics. How we score boringness will be based on loudness, tempo, energy, and danceability. In regards to variety, we will use a sample standard deviation of the audio features. We will compare our music libraries with each other to see the similarities and differences.

In the Pitchfork Analysis, we will explore the album scores, album genre, popularity, valence, danceability, and other audio features. We want insight into what are the characteristics of albums that are both popular and critically acclaimed. What kinds of albums might be critically acclaimed but not very popular. What features do these albums have? What albums are overrated: popular but not critically acclaimed. What kind of features do these albums have? What albums are unpopular and poorly received by critics, and what features do they have?

We expect that the 2017 P4K reviews yield a high amount of unpopular, highly received albums since they primarily review indie music. We might also expect the dataset to have more rock albums reviewed, since Pitchfork started by reviewing indie rock albums.

Anand: Add the Spotify pitchfork: 4 subdata sets to genre to characteristics analysis

**Report Plan and Organization**

Our report will be in two parts. Part 1 will be an analysis on our own personal music libraries. Part 2 will be an analysis of the combined data set for pitchfork 2017 reviews and album audio characteristics from Spotify. After garnering insights for both parts, we will summarize our findings in a white paper.

**Appendix:**

P4K & Spotify Dataset Approval:

Original open source data set: Pitchfork review data

- https://www.kaggle.com/ermoore/pitchfork-reviews-through-12617
- ~ 1260 lines because focusing on 2017 reviews only.
- Combining this data set with Spotify album data on the 2017 reviewed albums
    - Variables:

.

| artist | best | date | genre | review | score |
|---|---|---|---|---|---|

Spotify data set for pitchfork

- For each album in the 2017 pitchfork reviews, we wrote code to use the Spotify Web API to search that album by name & artist.
- The tracks for the album will be pulled
    - Each track will have audio features pulled
- =For each album, the audio features of the tracks will be averaged.

Spotify/Pitchfork combined data set

- Already made, cleaned, and saved to csv.
- Rows: for each 2017 pitchfork album that was found on Spotify. Reduces the dataset from ~1250 albums to ~1060, since not every album is on Spotify.
- Columns/variables: everything from the pitchfork review + Spotify variables. Below are the spotify variables:

| album_uri | album_name | album_artist | acousticness | danceability | energy | instrumentalness | liveness | loudness | speechiness | tempo | valence | popularity | duration_ms | is_explicit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

- Interesting questions:
- We can see if the score of the album has a relationship with the album popularity.
- We can look at the relationship between album score and other audio features
- Then we can investigate which genre's of albums have higher score
- By grouping by genre, we could investigate audio characteristics of the best reviewed albums in each.
- Partition the dataset into 4 parts by popularity and P4K score: most popular & best score (the best albums), least popular & best score (the obscure hits), most popular & least scores (the overrated), and least popular & least scores (the ignorables).
    - For each category, we can look at the genre breakdown
    - See if the albums had certain audio feature traits
    - See what their P4K reviews said.

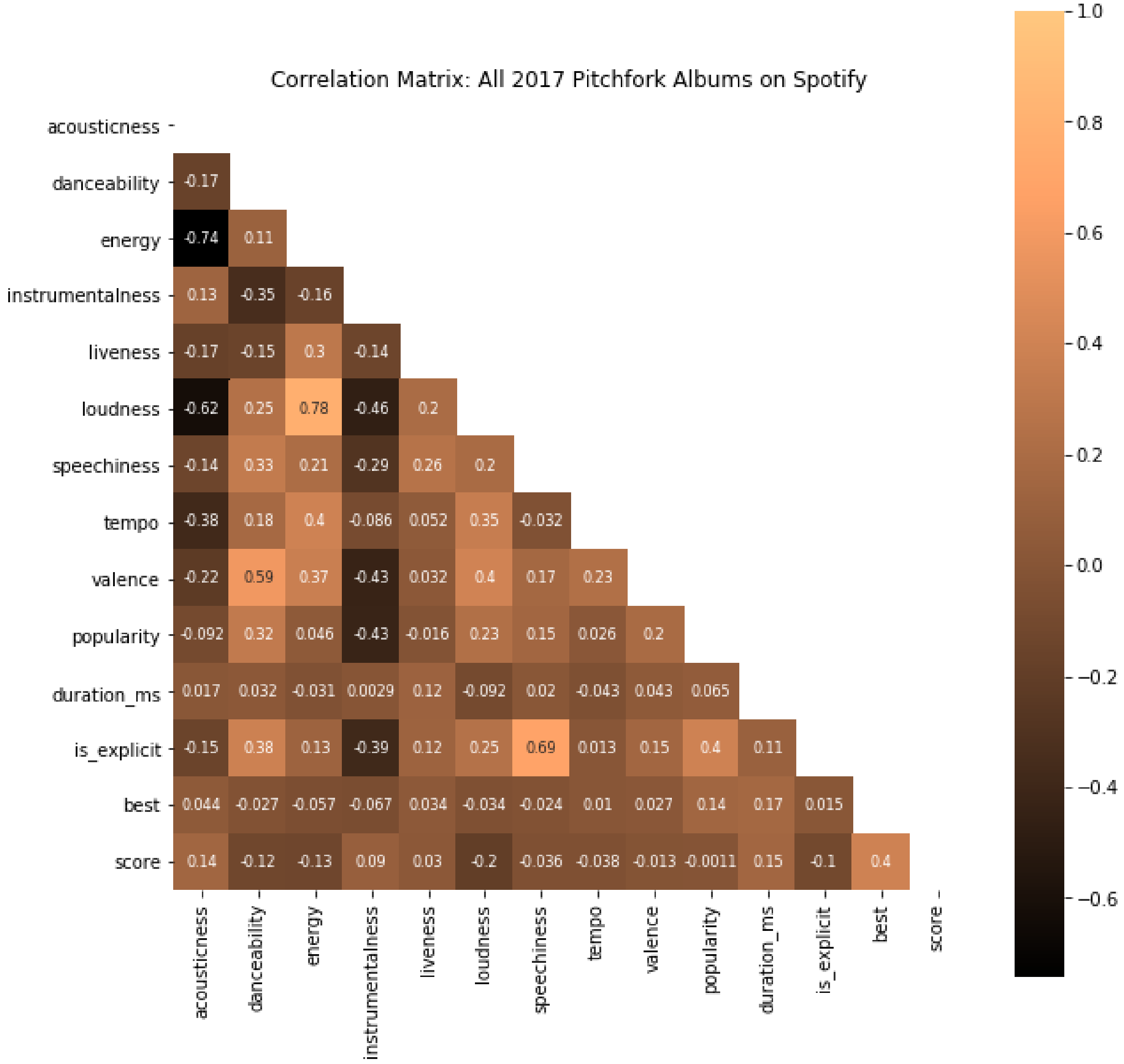
All of the data is available and cleaned now for the CSV. Plotting and visualization is ready to happen. We already generated a plot of a heatmap of the correlation matrix between Spotify and P4K variables in the 2017 spotify & P4K combined data set (See Heat Map of Combined Spotify/P4K 2017 Dataset). Only additional thing to clean might be converting the review date to a datetime value.

AudioFeaturesObject: Audio Features on Spotify Track

| KEY | TYPE |
|---|---|
| acousticness<br>A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. | Float |
| analysis_url<br>An HTTP URL to access the full audio analysis of this track. An access token is required to access this data. | String |
| danceability<br>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. | Float |
| duration_ms<br>The duration of the track in milliseconds. | Integer |
| energy<br>Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy. | Float |
| id<br>The Spotify ID for the track. | String |
| instrumentalness<br>Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. | Float |
| key<br>The key the track is in. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. | Integer |

| | |
|---|---|
| liveness<br>Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. | Float |
| loudness<br>The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db. | Float |
| mode<br>Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. | Integer |
| speechiness<br>Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. | Float |
| tempo<br>The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. | Float |
| time_signature<br>An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). | Integer |
| track_href<br>A link to the Web API endpoint providing full details of the track. | String |
| type<br>The object type: “audio_features” | String |
| uri<br>The Spotify URI for the track. | String |
| valence<br>A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). | Float |

Heat Map of Combined Spotify/P4K 2017 Dataset


Correlation Matrix: All 2017 Pitchfork Albums on Spotify
acousticness
danceability -0.17
energy -0.74 0.11
instrumentalness 0.13 -0.35 -0.16
liveness -0.17 -0.15 0.3 -0.14
loudness -0.62 0.25 0.78 -0.46 0.2
speechiness -0.14 0.33 0.21 -0.29 0.26 0.2
tempo -0.38 0.18 0.4 -0.086 0.052 0.35 -0.032
valence -0.22 0.59 0.37 -0.43 0.032 0.4 0.17 0.23
popularity -0.092 0.32 0.046 -0.43 -0.016 0.23 0.15 0.026 0.2
duration_ms 0.017 0.032 -0.031 0.0029 0.12 -0.092 0.02 -0.043 0.043 0.065
is_explicit -0.15 0.38 0.13 -0.39 0.12 0.25 0.69 0.013 0.15 0.4 0.11
best 0.044 -0.027 -0.057 -0.067 0.034 -0.034 -0.024 0.01 0.027 0.14 0.17 0.015
score 0.14 -0.12 -0.13 0.09 0.03 -0.2 -0.036 -0.038 -0.013 -0.0011 0.15 -0.1 0.4
acousticness
danceability
energy
instrumentalness
liveness
loudness
speechiness
tempo
valence
popularity
duration_ms
is_explicit
best
score