Project Summary

Overview

Your summary should begin with an overview of your topic, including a description of the problem. Here are Here are a few questions that you should consider addressing

- ✓ What is the topic of your project?
- ✓ Why is it important to tackle this problem in your project?
- ✓ What is your dataset about? What are some key variables of interest?

The topic of our project is the prediction of a song's chart ranking. We want to address the difficulties independent musicians, producers, writers, and other artists in the music community face during a song's creative process by providing an analysis and metrics dashboard of current and historically successful tracks. Such a dashboard allows users to gain valuable insights without the need to be familiar with Spotify's API, its underlying database, and statistical and machine learning methods. Our dataset includes two main subsets: Chart Performance and Track Features. The Chart Performance dataset includes entries for each song appearing on the given chart with the following attributes:

- **Chart type**
- **World region**
- **Chart date**
- **Chart position**
- **Trend**
- **Track title**
- **Track artist**
- **Track streams**
- **Icon URL**
- **Spotify URL**

Using each track's unique Spotify URL, we then query the Spotify API and accumulate audio feature data. This is stored in the Track Features database, which includes tuples representing tracks and the following attributes:

- **Track ID**
- **duration_ms**
  - The duration of the track in milliseconds.
- **Key**
  - The estimated overall key of the track.
- **Mode**

    - Indicares the modality (major or minor) of a track, the type of scale from which its melodic content is derived.
- **Time_signature**
    - 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).
- **Acousticness**
    - A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- **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.
- **Energy**
    - A measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity
- **Instrumentalness**
    - Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context.
- **Liveness**
    - Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
- **Loudness**
    - 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.
- **Speechiness**
    - Detects presence of spoken words in a track.
- **Valence**
    - Measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.
- **Tempo**
    - 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.

Questions/Hypotheses

- ✓ What questions have you formulated regarding your dataset? These questions may tie back to the problem statement (just present as questions this time).
- ✓ What hypotheses are you going to test?
    - Have you formed any hypotheses regarding potential causal relationships among some of the variables in your dataset? If yes, state your hypotheses here.

The questions we have formulated regarding our dataset surround how audio features related to a tracks chart success. We will answer:

- What are the typical values of audio features for tracks within a given music genre?
- Which features are most important to a track's success within a music genre?
- Which features are most important to a track's overall success in competition with all genres?

We have yet to form specific hypotheses regarding potential causal relationships among the variables in our dataset. Before we do so, we much conduct further, exploratory data analysis to explore how our observed datasets are distributed. This will allow us to form hypotheses about the relationship between each audio feature and the track's overall, maximum achieved chart ranking. We will then test the statistical significance of these relationships using the appropriate hypothesis testing methods.

Data Analysis Plan

- ✓ What are some analyses you are planning to conduct? This is tentative and can change later.
  - Are you planning to predict an outcome based on some of your variables? If yes, explain here.
  - Are you planning to use machine learning techniques? If yes, briefly describe your plans here.

We are planning to use machine learning techniques to learn the optimal weights for each feature in a track's feature vector. Using these learned optimal weights, we can create a polynomial to predict the success of a song by two metrics: total streams and high achieved chart ranking. To predict total streams, we will use perform regression analysis using the sklearn library to output some continuous value prediction. To predict maximum chart position, we will perform classification analysis, again using the sklearn library, to predict a discrete chart label, ranging from the highest chart position (1), to the lowest chart position (200), and finally a label of 'Uncharted'.