About the Data

1. What is in your data?

Here is the data we will be using for this project.
You can better understand the variables within this data here.

The data we will be dealing with in this project is Spotify data. This data was found in Kaggle, but was originally pulled from the Spotify API. Each row in the data is one song that was released between 2004 and 2019. In this data, the variable we will be trying to predict is popularity (track_popularity). We will use the following predictors to help predict popularity: album release date, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and duration. Below is a data dictionary to help us familiarize ourselves with the data.

Variable	**Variable Name in Data**	**Description**
Popularity	track_popularity	Calculated by an algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are.
Album Release Date	track_album_release_date	Provides the date an album was first released.
Danceability	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	energy	A value between 0 and 1 that represents the perceived intensity and activity level of a song, essentially indicating how energetic or upbeat a track feels, with higher values signifying more energetic songs that are typically fast, loud, and noisy.
Key	key	The key that the song is in
Loudness	loudness	Provides the overall loudness of a track in decibels (dB)
Mode	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
Speechiness	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.
Acousticness	acousticness	A confidence measure of how likely a track is to be acoustic, ranging from 0.0 to 1.0. A score of 1.0 indicates a high confidence that the track is acoustic.

Instrumentalness	instrumentalness	An audio feature that predicts if a song contains no vocals. The instrumentalness value is a number between 0.0 and 1.0, with values closer to 1.0 indicating a greater likelihood of no vocals.
Liveness	liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
Valence	valence	A numerical value between 0 and 1 that indicates how positive or "happy" a song is perceived to be, with higher values signifying a more positive mood and lower values indicating a more negative mood
Tempo	tempo	The estimated overall tempo of a track in beats per minute (BPM)
Duration	duration_ms	A number that represents the length of the section in seconds

2. How will these data be useful for studying the phenomenon you're interested in?

This data provides several variables that will allow us to determine and predict the trends of music across a span of 15 years. This will allow us to analyze and observe how musical preferences can evolve over time such as through shifts in genre or production styles. Using a machine learning model will allow us to forecast popularity trends in the future based on the variables above. We expect to build a model that will be able to analyze the numerical inputs of each variable and predict the popularity of the hypothetical song.

This type of analysis could be useful to music producers, artists, or anyone who is putting money and time into song production. It can let them know of the factors that may cause their song to be more popular and have more success, which can inform their artistic decisions.

3. What are the challenges you've resolved or expect to face in using them?

During our data cleaning process, we did run into a few challenges that we were able to work through. Firstly, some of our categorical variables were labels with numbers. We wanted to make sure that we didn’t misinterpret these variables to be quantitative variables later on in our study, so we decided to relabel these variables with the class they actually represent, instead of just meaningless numbers. Additionally, certain variables were not useful for analysis, so we decided to remove them from the data. Lastly, we spent a good amount of time cleaning the date released variable. First of all, the year variable had years ranging from 1960 all the way to 2019. Most songs were made past 2000, so we decided to limit this analysis to just songs in the 2000s. Additionally, we split the date variable into just a year and a month variable, because we thought these two variables would be more helpful for analysis.

Since some of these variables are a bit difficult to understand (as they are based on models/probabilities themselves), I think we may run into issues with interpretation of our model and predictor values. Even our response variable is based on an algorithm. This may cause issues when attempting to understand what our results mean. Also, some of these variables seem to be a bit subjective.

Additionally, because some of our variables are likely related (such as instrumentalness and speechiness), I predict that we may have some issues with multicollinearity in our data. Multicollinearity occurs when variables in a model are giving redundant information in predicting the response variable. To address this issue, we should probably assess the between-predictor correlations, so see how correlated our predictor variables are with each other.