Group Participants: Maya Uwaydat, Teagan Britten, Gracie Williams, Aileen Kent, Nyla Upal, Jessica Ni, Ruth Melese

1. What is in your data?

We have two datasets, both from Kaggle: one has Billboard Hot 100 data, and the other has data pertaining to features of each track, taken directly from the Spotify API. The Spotify dataset includes data such as the song performer, the song title, and the Spotify track ID, in addition to characteristics such as tempo, track duration, danceability, valence (i.e. happiness), energy, key, and instrumentalness, among other variables. The Billboard dataset contains information regarding the time(s) a song scored on the Billboard Hot 100 and the rank of the song on the Billboard.

Additionally, we created our own data by calculating the "success" of each song on the Billboard, which we appended to the Spotify dataset (eventually renamed to the "song attributes" dataset). Success is measured by its rank on the billboard and the amount of time that it stays on the Billboard. The lower the rank, the less popularity points the song earns, and vice versa. Each song is assigned a different number of points for the weeks it remains on the billboard and the most successful songs will have the highest number of points. The total point values will be added to the Spotify data frame that we have (this formula is explained in more detail in the answer to question 2).

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

Our aim is to see what compositional elements predict a song's success—given the song includes the word "love" in the title (though the exact word we use may be revised).

We will measure a song's success by looking at its popularity through a) its rank on the billboard, and b) the length of time it maintains its rank. A song is awarded a certain number of "popularity points" in a week according to its rank on the Billboard Hot 100. A song with a rank of 100 will be awarded 1 point for that week; a song with rank 1 will be awarded 100 points for that week; a song with rank 7 will be awarded 94 points for that week; and so on and so forth. The number of points a song earns has a negative linear correlation with its rank. Each song will have their popularity points totaled.

Our Spotify data provides us with *plenty* of compositional elements to look at in a song to predict its success (i.e. predict high "popularity point scores"). Because we have so many compositional elements to account for when predicting success, we are considering using principal component analysis (PCA) to reduce the dimensionality of our attributes data. Or we could focus on just a *few* variables, if PCA winds up being too technical.

It would probably be most prudent to look at how certain attributes of songs predict their success *over time*; trends in the music industry shift, and a song about love from the 1960s would likely sound nothing like a song about love from the 2020s. We could (and probably should) also look at the predictors of song success overall and compare them to predictors of love songs' success.

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

One challenge we navigated was using two different data sets; it was difficult to know just how we wanted to analyze song attribute data and Billboard data together, but we resolved the ordeal by creating the popularity points formula and assigning each song a quantity of points. We are also currently having some issues retrieving songs that have only been on the Billboard top 100 for only one week; they are not showing up when we attempt to peek at parts of our Billboard dataset with our code.

As far as analytics is concerned, there are also likely confounding variables that may affect our prediction of a song's popularity. Additionally, Spotify's API will also sometimes label songs' attributes in ways that seem illogical to humans, which may make analytics more difficult: genres are sometimes tagged incorrectly according to widely accepted genre definitions, or tracks with high calculated "valence" (happiness) score may sound sad to their audiences.

Refer to codebook in the repository to see more general EDA on our data and some data visualizations