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

Primary Question/s for our analysis:

- How can you predict the popularity of songs?
- Is there a single variable that accurately predicts popularity? Is there a most influential variable?

Regression / Supervised Learning

Our group plans to perform supervised learning, specifically multiple linear regression to analyze our dataset. Since the data is already labeled and the target variable is already known, our model will map the relationships between features to make predictions. In our case, we plan to predict the popularity of a song, which is represented by popularity points. Popularity points were initially calculated based on a song's rank and time on the Billboard Hot 100, and essentially represents the "success" of a song on the Billboard (the higher the better!). Specifically, we plan to use various song characteristics from our dataset, for example, valence, tempo, or binary variables to predict popularity points. The binary variables will be feature engineered and one hot encoded to indicate the presence or absence of certain words of interest within a song's title, such as "love", allowing us to examine their potential influence on a song's popularity.

Models and Algorithms in Analysis

Our dataset consists of many features that describe a song, so we may decide to use principal component analysis (PCA) to reduce the dimensionality of our data. We understand that many features of a song may be related, for example, songs with higher valence may also have a higher tempo. Thus, to reduce many correlated and redundant variables, we may use PCA to transform our model to contain principal components that capture the most variance. The downside of using PCA would be the loss of interpretability of the result. Without PCA, we could look at our regression coefficients and more confidently understand the contribution of each of our chosen variables. But with PCA, we would have to interpret our principal components, which are a combination of multiple features, making it more difficult to understand the effects of individual variables.

What does "success" mean?

In one direction, we may use a train test split when classifying data and look at the Sum of Squared Errors (SSE) to evaluate our model's performance. Lower values would indicate that our predictions are closer to actual values. In another direction, we may look at the R-squared value to understand how much of the variability in the target variable, popularity, is explained by our chosen features. A higher R-squared indicates

better fit. We may also apply our model onto new songs and see if their real life Billboard score is analogous to our predicted popularity points.

Weaknesses and Concerns

A few weaknesses we identified regarding this project relate to the concept of song "popularity" as well as other weaknesses. Below is a list of concerns:

- Popularity might depend on other variables that we cannot account for, such as seasonal popularity. We can see this through Mariah Carey's "All I Want For Christmas Is You".
- Objective rank may neglect information that would let us know how popular it is, it may ignore downloads. Plays also are not a good measure of popularity and cannot actually teach us about people's trends or how they like them.
- A number for popularity is difficult to deal with, older songs that have been low on the list for more years will generally be higher than new songs that are incredibly popular, which is an interesting bias towards older songs. We may have to take into account the age of the song, although if a song has been on the billboard for many years it probably is very well-known.
- Many of the variables are subjective, making it interesting to think of who is encoding and deciding these metrics, like danceability or valence.
- Potential learning point: a wide scope can easily sneak up on you and make predictions extremely difficult

When having too large a scope, it becomes possible to overlook details in defining a question and performing analysis. Throughout the project, we hope to maintain clarity in our questions and analysis process, to make meaningful predictions.

Results

Similar to measuring success, we will present our results by evaluating the accuracy of our model's predictions. We may use the R-squared value to understand how much variance of the target variable is explained by our features. Additionally, we may create visualizations to compare and showcase the predicted versus true values. We may also utilize graph coefficients from the multilinear regression equation to understand each feature's contribution to defining the target variable.