1. Project Members

Michael Kinkley - mkinkley
Miri Hyman - mirhyman
Erik Hoberg - ehoberg

2. Dataset name

miri_playlist

3. Description of the Dataset

The dataset contains a little over 13,000 tracks with audio features for each track. We want to explore how the various track features contribute to the popularity of a track (the output being the number of people who listened to the track). We believe that this problem is not too hard because the data was found online in an easy to use format without much manipulation required. It does not contain an unreasonable number of features or rows and has a clearly defined output label for each track. This data is not too easy because we included over 13,000 rows and no feature seems to jump out at overly correlating with the output label. In order to ensure that this dataset is of high quality, we tried to throw out subjective features from the original dataset or features that were incomplete for some rows. We chose features that we thought might have been related to the popularity.

4. Format

Each input instance contains 8 features. The 8 features, from left to right, are [acousticness, danceability, energy, instrumentalness, liveness, speechiness, tempo, valence].
A full description of these can be found here: https://developer.spotify.com/web-api/get-audio-features/

5. Training and Test Groups

We chose to divide 20% of our data into a training set and use the remaining 80% as training data. We performed this divide by choosing the top 20% of the trackids and assumed that this gave us a relatively random split, since a sorting on trackids would not make much sense for any of the other features. We chose a high proportion of the data as test data in order to ensure that a model over this data would have enough information to develop an effective algorithm. We did not proportion any data into developement data because we decided that if this dataset were to be chosen we could just redistribute some of the test data into developement data before training begins.

6. Steps taken to ensure dataset is ethical

We believe our dataset is ethical because trying to predict how popular a song will be is not targetting any specific group of people or leading to conclusions that might harm anyone. Our dataset is also freely available to the public, so there are no ethical concerns about our data collection or use of this data.