CS 109A Final Project: Milestone 2

Tejal Patwardhan, Akshitha Ramachandran, Grace Zhang

1. What is your group #?
   - Final Project - Group Number #49
2. Have you met/communicated with your fellow teammates?
   - Yes, we have met and have been communicating with each other.
3. Have you met/communicated with your assigned TF? If not, please provide a reason.
   - Yes, we have emailed our TF (Rashmi) and have scheduled a meeting for Wednesday at noon to review our progress.
4. Has your team formulated a well-defined question to address in your project, based on the project description and references? If so, please write down the question. If your team hasnt done this yet, thats okay!
   - **Description**: We are best friends and have a shared playlist together consisting of some songs from the million songs database. We are looking to create a similar playlist and will do so in this project. We will create our dataset by introducing a new binary variable into a training set from the million songs database that indicates whether the song should be in our playlist, classifying songs in our training set as worthy or not-worthy. We will then train various models to predict songs that should be in the playlist, and later compare the models to select the best one. To train our models, we will explore features including artist, album, year, emotion, and lyrics. Finally, we will run our model on an excluded subset from the million songs playlist to create a new playlist we can jam with together!
   - **Response variable**: The response variable will be a binary variable indicating whether the song is in the playlist (or not).
   - **Dataset**: We will work with the Million Playlist Dataset and extract our own subset of songs from said set for training, testing, and validation purposes.
   - **Models**: We will consider the following models:
     - Logistic Regression, with:
       - Stepwise Variable Selection
       - Ridge Variable Selection
       - Lasso Variable Selection
       - Interaction Terms
       - Higher-degree Polynomial terms
     - Neural networks

- And more if we learn them in time!

5. Briefly describe your teams plans for work to be completed by Nov 28 (milestone three). Please assign specific tasks to team members and deadlines for when these tasks are to be completed.

   - **Week of October 22**: Write out plans and collect data. We all did this already, and will be using the data from here: `https://labrosa.ee.columbia.edu/millionsong/sites/default/files/AdditionalFiles/unique_terms.txt`
   - **Week of October 29**: Data cleaning and filtration, determining the subset of data that will be most useful for our analysis (all together).
   - **Week of November 5**: Data exploration. Conduct EDA and data visualizations to determine whether there are any important features or connections between the data.
   - **Week of November 12**: Build logistic regression models and ensure that the data used is appropriate for these analyses.
   - **Week of November 19**: Evaluate models and determine which one best fits our project goals. Write up milestone 3 and submit to Canvas.

   Right now, we are at an early enough stage in the process that we are planning to do this work together, sharing a laptop and collaborating in the ideation process in person. For later steps, such as writeup and figure creation, we will assign more specific tasks to specific people.