Milestone 2
Group 52
Aileen Villalpando, Colin Harvey, and Clara Cousins
TF: Matteo Zhang

Scope of Work and Preliminary EDA Agenda

1. What is your group #? Group 52
2. Have you met/communicated with your fellow teammates? Yes, we met on Sunday, October 13th from 3:30-5pm to brainstorm about the question we plan to address in this project. We have also worked together in a shared google document to plan the details of our project.
3. Have you met/communicated with your assigned TF? If not, please provide a reason. Yes, we met with Matteo Zhang on Monday, October 21st from 1:00-1:25pm.
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 hasn't done this yet, that's okay! We will seek to answer the following question: "Can the subject matter/ main topic of the song (derived from lyrics) be predicted from the sound qualities of the song (beats per minute, danceability, etc.) and can a classifier that accomplishes this prediction task be used to make playlists containing songs with related subject matter?"
5. Briefly describe your team's plans for work to be completed by Nov 20th (tentative milestone three). Please assign specific tasks to team members and deadlines for when these tasks are to be completed. For example,
   - Chris and Pavlos will complete data collection and cleaning by Nov 14.
   - Kevin and Eleni will review references and develop a preliminary EDA strategy by Nov 14.
   - Eleni, Chris, Kevin, and Pavlos will complete individual EDA assignments by Nov 18.
   - Chris will write the description of data for milestone 3 by Nov 19.
   - Pavlos will summarize the individual EDA findings for milestone 3 by Nov 19.
   - Kevin will write the final version of milestone 3 and submit to Canvas on Nov 20.

   Overall project plan: We intend to use the Spotify API to get tracks for about 10-20 different playlists with each playlist having songs that are unified by a different theme (such as love, dance, driving, etc.). We will listen to all songs while reading the lyrics for them and manually assign each song a subject matter label (from a list of about 10-20 labels, with each label defined by about 20 indicator words - i.e., words that are related to the label). Possible subject matter labels (and related indicator words) include: songs about loving someone (indicators: "love", "mine", "care", "baby", "heart", etc.), songs about hating someone (indicators:

“hate”, “anger”, “unfair”, “cheated”, etc.), songs about going out to a party (indicators: “party”, “dance”, “down”, “club”, “floor”, etc.), songs about religion (indicators: “God”, “Jesus”, “temple”, “Heaven”, “spirit”, etc.), songs about drinking (indicators: “alcohol”, “glass”, “cup”, “drink”, “drunk”, etc.), songs about road trips (indicators: “drive”, “car”, “road”, “wheel”, “seat”, etc.). We will then use Beautiful Soup to extract the lyrics from each song from the lyrics database at https://lyrics.fandom.com/wiki/LyricWiki and determine how many times each indicator word appears in the lyrics. We will then set the “true” label for each song’s subject matter to the label corresponding to the indicator word that appears the most in the lyrics. We will then use the Spotify API to extract sound qualities (including beats per minute, danceability, duration, etc.) for each song. We will build a classifier (such as multiclass logistic regression, perhaps with features from PCA of all the available sound quality variables) that aims to predict subject matter label from sound qualities. We will train this model using cross-validation. To use the model for automatic playlist generation, we will prompt the user to specify the subject matter label that best matches their preference at the time and provide a handful of songs that they like (a short playlist). We will extract the sound qualities for each song in the user’s base playlist using the Spotify API and, after extracting the sound qualities for others songs in Spotify (songs not used to train, validate, or test our model as well as not in the user’s short playlist), apply a clustering approach determine ~10 songs from this fresh database of new songs that have similar sound qualities to each song in the user’s short playlist. We will then use our classifier to predict the subject matter label for each of these 10 songs (10 songs per 1 song from the user’s playlist). Finally, we will recommend as a playlist for the user all of the songs that have predicted labels that are consistent with the user’s preference specified previously. Although the task of generating the playlist is based on a classifier that was trained and tested previously (and so is known to be robust), we can validate the playlists generated by extracting the lyrics for each song in the recommended playlist, determining the “true” subject matter label, and comparing our predicted labels with the “true” ones. We can also validate the playlists by seeing if the genres of the recommended songs are similar and related (qualitatively) to the assigned subject matter label.

Timeline and tasks for each group member:

- Colin creates 20 subject matter labels and thinks of 10 indicator words associated with each label by 10/28.
- Colin compiles a list of songs from 10-20 Spotify playlists that have songs that are likely to be thematically related to our subject matter labels (goal is to have the same number of songs for each subject matter label) by 10/30..
- Aileen determines which sound quality features we will extract for each song and uses the Spotify API to extract sound quality features for any one song (goal here is to get the process/pipeline set up for finding sound quality features for each song) by 10/30.

- Clara uses Beautiful Soup to extract lyrics for any one song and find the “true” label for that song (goal here is to get the process/pipeline set up for finding subject matter label for each song) by 10/30.
- Clara will get the lyrical subject matter predictions for each song that we intend to use for training our classifier by 11/6.
- Aileen will collect the sound qualities data for each song that we intend to use for training our classifier by 11/6.
- Colin will clean the data for both lyrical content and sound qualities (remove outliers, normalize features as appropriate, etc.) and make a baseline classifier model to predict subject matter label from sound qualities and report accuracy by 11/12.
- Colin will plot the distribution of songs in each subject matter label and show how songs are distributed across sound quality variables by 11/14.
- Aileen will write the description of data for milestone 3 by 11/18.
- Colin will write the methods of EDA for milestone 3 by 11/18.
- Clara will write the revised project question for milestone 3 by 11/18.
- Full group will check over and make revisions to the milestone 3 write up in preparation for submission by 11/20. Colin will submit to Canvas.

7. **Project A only:** include the physical location of the sensors' broader area (Cambridge, Somerville, etc) We are assigned Project D.