- # Project Proposal

1. Introduction

The taste of music has changed throughout history. Now that access to new music is easier than ever, music platforms are finding specific trends and traits to outperform competitors and attract more users using machine learning. In this project, we want to dive into the field and figure out for ourselves how music platforms use the data they collected to give their users better experience in music recommendations, how they categorize music into micro-genres and much more. One motivational example could be that of Spotify mobile App, how they have segmented their archive into thousands of micro-genres.

2. Research Questions:

a. Which genre should a song belong to based on specific characteristics?(from dataset -2)
b. Predict the year of the song based on different characteristics like album cover etc. (from dataset -1)
c. Given a song, what are some other recommendations that would suit the same user's taste? (either dataset)
d. Can we predict the popularity of a song based on given features? (from dataset -1)
e. Which genre got famous/changed according to year and why? (from dataset -1)
f. Which feature affects the popularity of a song the most?(from dataset -1)
g. Who are the popular artists? (from dataset -1)
h. Predicting if a new artist is going to be popular based on its audio features(from dataset -1)

3. Methodology

Data Cleaning & Exploration:

- Make sure each column is in the correct data type.
- No missing values in the training/testing data
    - Either removing rows/columns, or filling missing values with column median/mean and such
- Dealing with outliers

Feature Engineering :

- Correlations in columns
- Adjusting columns according to different scenarios
- Data visualizations using R libraries

Data Wrangling and Modeling :

- Making use of all the different models mentioned in the Model Selection section below

4. **Metrics:**

   As we will be training our model to predict different characteristics, we are planning to use k-fold cross-validation and then test our dataset to measure the performance of our model. We will use classification and clustering to solve for our research questions. For classification problems we plan to use F1 score and possibly confusion matrix to represent the performance of our model and as for clustering, random index and silhouette score are some pretty good metrics we will consider.

• Project Outline

1. **Literature Review**
   Some of the related works include:
   a. Genre Classification of Spotify Songs using Lyrics, Audio Previews, and Album Artwork (https://cs229.stanford.edu/proj2017/final-reports/5242682.pdf): this paper analyzes the data from Spotify to classify the genre of a song based on the songs lyrics, audio previews, and album artwork. The model they built achieved 91.75% accuracy on the test dataset. The possible genres were Metal, Christian, Country and Rap.

2. **Data Source and References**

   Dataset 1:
   Core Information: This dataset is from Kaggle (

   Core information: The dataset is from Kaggle (https://www.kaggle.com/datasets/lehaknarnauli/spotify-datasets?resource=download&select=tracks.csv). This dataset has audio features of over 500,000 songs. This is an open source and can be downloaded as a CSV file from Kaggle.

   Metadata: This dataset has two subsets of data, artists & tracks. Here are the details for each dataset:

   Tracks: It consists of 20 columns and has 586,672 rows. Each column defines a specific feature of the song. Some of the relevant columns for our purpose are as follows:

Field Name	Type	Brief Description
ID	String	A unique identifier for the song
name	String	The name of the song
popularity	Numeric	Defines the popularity of the song. The value is between 0 to 100
duration_ms	Numeric	Defines the duration of the song in milliseconds
artists	String	The name of the artist
id_artists	String	A unique identifier for the artist
danceability	Numeric	Defines the danceability of the song. The value is between 0 to 1
energy	Numeric	Defines the energy of the song. The value is between 0 to 1
loudness	Numeric	Defines the loudness of the song. The value is between -60 to 6
speechiness	Numeric	Defines the speechiness of the song. The value is between 0 to 1
acousticness	Numeric	Defines the energy of the song. The value is between 0 to 1
liveness	Numeric	Defines the liveness of the song. The value is between 0 to 1
tempo	Numeric	Defines the tempo of the song. The value is between 0 to 250

Artists: It consists of 5 columns and has 1,104,349 rows. Each column defines a specific feature of the song. Some of the relevant columns for our purpose are as follows:

Field Name	Type	Brief Description
ID	String	A unique identifier for the artist. This column can be joined with the id_artist from the Tracks dataset
followers	Numeric	The number of followers the artist has
name	String	The name of the artist

According to Kaggle, this document has some missing or mismatched data in the 'Track' CSV file, specifically the 'name' column. This information can be found on this link. As we work on the dataset, we will verify this information as well.

Dataset 2:

Core information: The dataset is from Kaggle (https://www.kaggle.com/datasets/mrmorj/dataset-of-songs-in-spotify?resource=download). This dataset again has audio features of different songs. This is an open source and can be downloaded as a CSV file from Kaggle.

Metadata: This dataset has two subsets of data, genres_v2 & playlists. Here are the details for each dataset:

genres_v2: It consists of 22 columns and has 42,306 rows. Each column defines a specific feature of the song. Some of the relevant columns for our purpose are as follows:

Field Name	Type	Brief Description
ID	String	A unique identifier for the song
genre	String	Defines the genre of the song
song_name	String	Defines the name of the song
danceability	Numeric	Defines the danceability of the song. The value is between 0 to 1
energy	Numeric	Defines the energy of the song. The value is between 0 to 1
loudness	Numeric	Defines the loudness of the song. The value is between -35 to 4
speechiness	Numeric	Defines the speechiness of the song. The value is between 0 to 1
acousticness	Numeric	Defines the acoustic-ness of the song. The value is between 0 to 1
liveness	Numeric	Defines the liveness of the song. The value is between 0 to 1
tempo	Numeric	Defines the tempo of the song. The value is between 55 to 220

playlists: It consists of 2 columns and only has 40 rows. The columns and its description are as follows:

Field Name	Type	Brief Description
playlist	String	A unique identifier for a playlist
genre	String	Defines the genre of the song

According to Kaggle, this document has some missing or mismatched data in the 'genres_v2' dataset, specifically the 'song_name' column. This information can be found on this link. As we work on the dataset, we will verify this information as well.

3. Data Processing

Both the datasets are CSV file and the size of both files are approximately 350 MB combined. This dataset consists of string and numeric values, so importing them in R Studio would be easy. Some of the features in the dataset will not be relevant to our analyses and will be removed from the dataset. Before carrying out the analysis we will:

a. Make sure that we do not have any duplicate rows.
b. Remove any outliers.
c. For any missing value, we will not consider that feature for that row in the analysis.

If the performance of the best model at the end still isn't up to the standard we will always go back to the data itself and do some more explorations and feature engineering like selecting only the relevant columns in the model.

4. Data stylized facts

We included a clustering problem in our research questions with unsupervised learning. We can show how the features interact with each other on a 2-d graph via dimensionality reduction methods such as principal component analysis.

5. Model Selection

For problems regarding classification, we will use models including but not limited to:

- Decision trees
- Naive Bayes
- K-nearest neighbor
- Support vector machines

For problems regarding clustering, we will use models including but not limited to:

- KMeans
- DBScan
- Agglomerative
- Gaussian Mixture

The majority of this project happens in model selection. We will be doing lots of hyperparameter tunings for each possible model and there will be visualizations for picking the best parameters and overall model selection. We will choose models based not only on the resulting metrics but also the exact scenario based on each question.

6. Software packages and Dataset

Software: We are planning to use R-Studio to answer the questions mentioned in the project proposal section.

Libraries/packages: We are planning to use the following libraries (but not limited to): ctree, caret, ggplot2, plotly, corrplot, tidyverse, dplyr, kernLab, e1071, DataExplorer, Caret