Project Deliverable #1 - Project Proposal

Background and context to the problem statement

Music can be an important part of people's daily life. For our project, we would like to investigate what makes a song popular and use machine learning to predict the popularity of a new song. There are many factors that could contribute. This includes intrinsic factors to the music itself, such as tempo, genre, key, mood of the lyrics, as well as the artist who released the music and the year in which the music came out.

Identification, description and source of the data set(s) plan to use

We plan to use a dataset from kaggel.com. This dataset is collected from Spotify's api and contains 11,450 songs' data, from 1986 through 2023. This dataset would be a great resource for our project as it contains a vast amount of song data from various genres and dates. The dataset mainly contains 37 features, which include numeric features such as loudness, danceability, and categorical features such as album_type, artist_genres, and text features such as track_name. As it is collected from Spotify's API, the data is reliable and in relative high quality.

Proposed ML techniques to solve the problem

We would like to explore various regression models to predict a song's popularity level. We will try out models like linear regression, ridge regression, Lasso regression, decision tree regression, random forest regression, and principal component analysis. It would also be interesting to use feature selection, to better understand which variables impact popularity.