Spotify: Predicting Song Popularity Using Audio Features

1. Introduction

a) **Problem statement**:

Spotify is a constantly growing music streaming service founded in 2006. While it carries a wealth of music catalog, it continues to be inundated with new tracks every day. Our goal is to predict the popularity of songs to refine the quality of the recommendation system.

b) **Background**:

According to Spotify CEO Daniel Ek, there were “nearly 40,000” new tracks being uploaded daily. In February 2021, the CEO said that this number has increased to over 60,000. With such an overwhelming supply of new music, music streaming services would need to retain listeners' engagement by being able to recommend potentially popular music as soon as possible, sometimes even before the song actually becomes popular; one would need an efficient way to predict song popularity.

Furthermore, having a model that can accurately predict a song’s popularity will also help artists, music producers, and audio engineers to craft their track to better target their potential consumers.

c) **Goal**:

This project aims to develop a system to predict popularity supported by data-driven analysis. Spotify provides fourteen audio features, including danceability, valence, energy, tempo, loudness, speechiness, instrumentalness, liveness, acousticness, duration, year, time signature, key, and mode. These audio features will be used to explore and understand patterns in popular songs to predict popularities. Based on our predictions, correlations between

acousticness and loudness, danceability and loudness, danceability and popularity, danceability and tempo, duration and energy, duration and instrumentalness, duration and speechiness, and popularity and speechiness are likely to increase in popular songs as years pass.

2. Datasets

a) US 1921- 2020 dataset:

The audio feature dataset for top charting songs from 1921 to 2020 has been sourced from Kaggle. The data was read in as a DataFrame called **df_US20**.

df_US20 contained 20 features. They could be categorized into four categories:

1) Mood: danceability, valence, energy, tempo
2) Properties: loudness, speechiness, instrumentalness
3) Context: liveness, acousticness
4) Others: popularity, song name, artist name, song id, artist id, explicitness, duration, release date, key, mode (musical mode), and time signature.

b) US 2021 dataset:

The dataset for top charting songs from 2021 has also been sourced from Kaggle. The data was read in as a DataFrame called **df_US21**.

df_US21 contained 13 features, which included: Unnamed: 0, position, track name, streams, chart start date, chart end date, album name, release date, artist, features, song duration(ms), explicit, url.

c) Audio features of US top charting songs in 2021 via Spotify Web API:

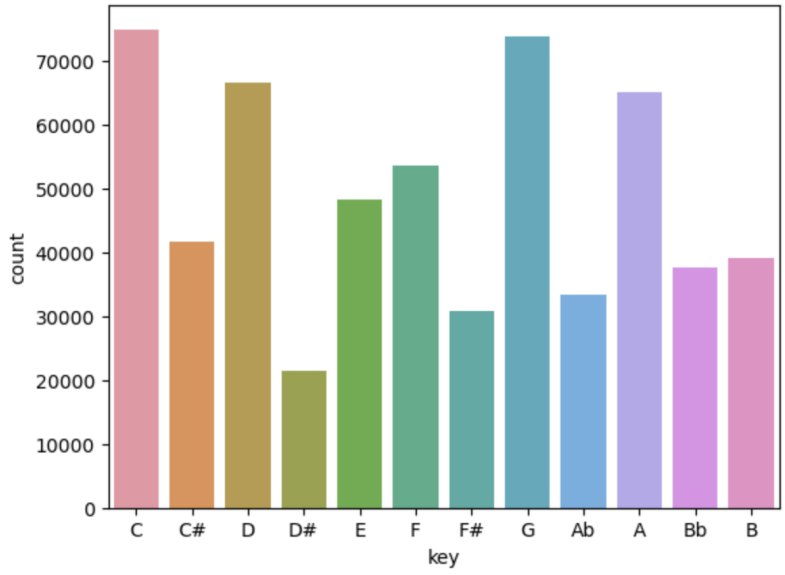
df_US21 data was initially sourced to gather audio feature data for top charting songs in 2021. However, **df_US21** did not have the audio features we needed for this project. We sourced song titles and their artist names from **df_US21** to acquire their audio features using Spotify Web API. The acquired data was saved as a DataFrame called **US21**.

3. Data Cleaning and Data Wrangling

1) In **df_US20** and **US21**, column 'mode' has been renamed to 'm_mode' (standing for musical mode) to avoid confusion with statistical mode.
2) **df_US20**:
    a) There were null values in the "name" feature. This was dropped as song_id serves the same purpose.
    b) Column 'release_date' was renamed to 'year'; we also only kept year from 'release_date' data, because we only wanted to visualize trends throughout years, rather than specific dates.
3) **df_US21**:
    a) There was no audio feature data we were after. As such, only the track names and artist names were extracted to use with Spotify Web API.
4) **US21**:
    a) Created a column and imputed 2021, as all the songs in this datasets are from 2021.

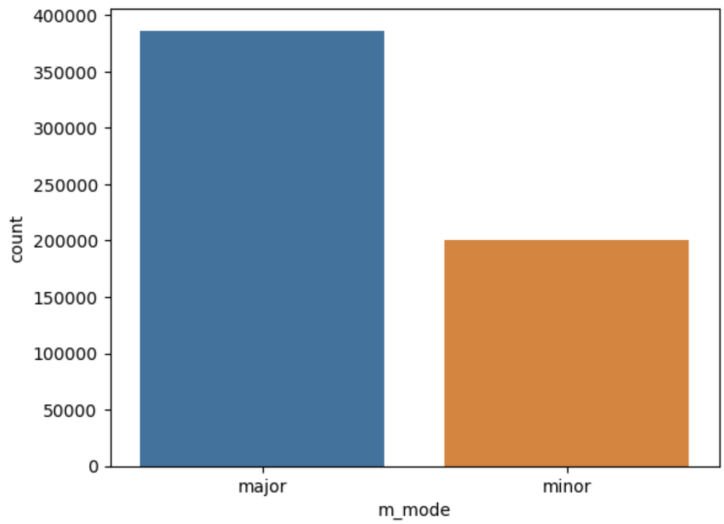
4. Exploratory Data Analysis and Initial Findings

1) **Keys**:



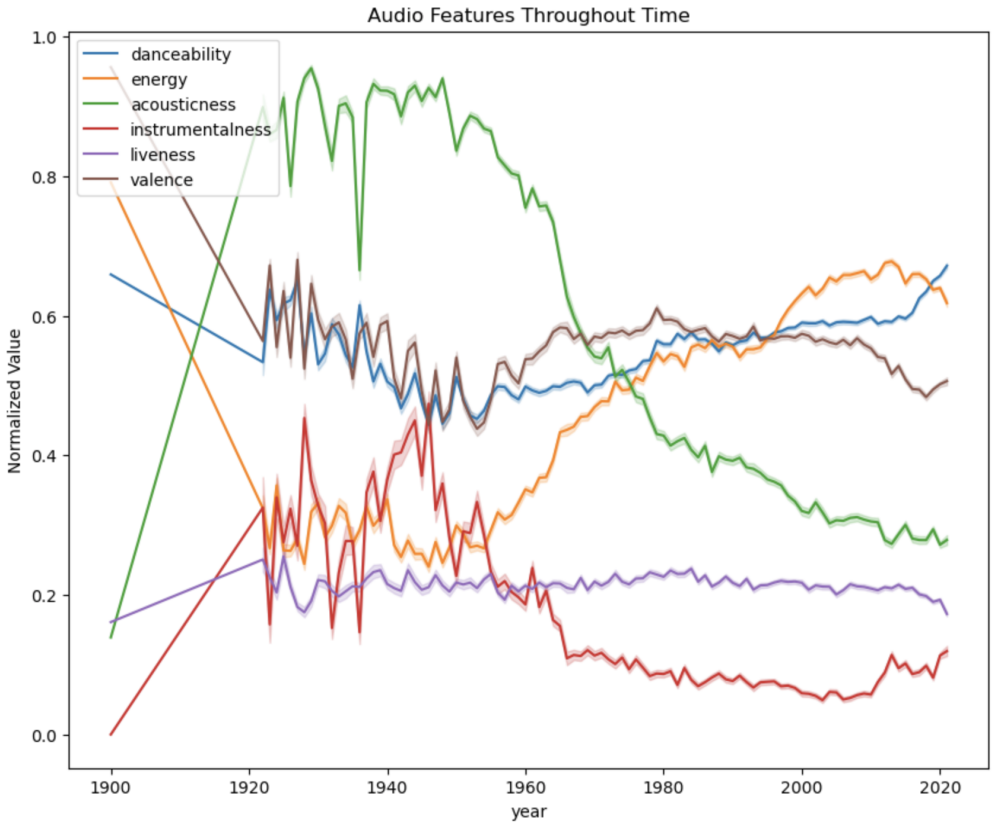
The most common keys in popular songs were C and G. This makes sense, because these two keys are keys that most people find easier to sing along to.

2) **Musical modes:**



More popular songs are in major mode.

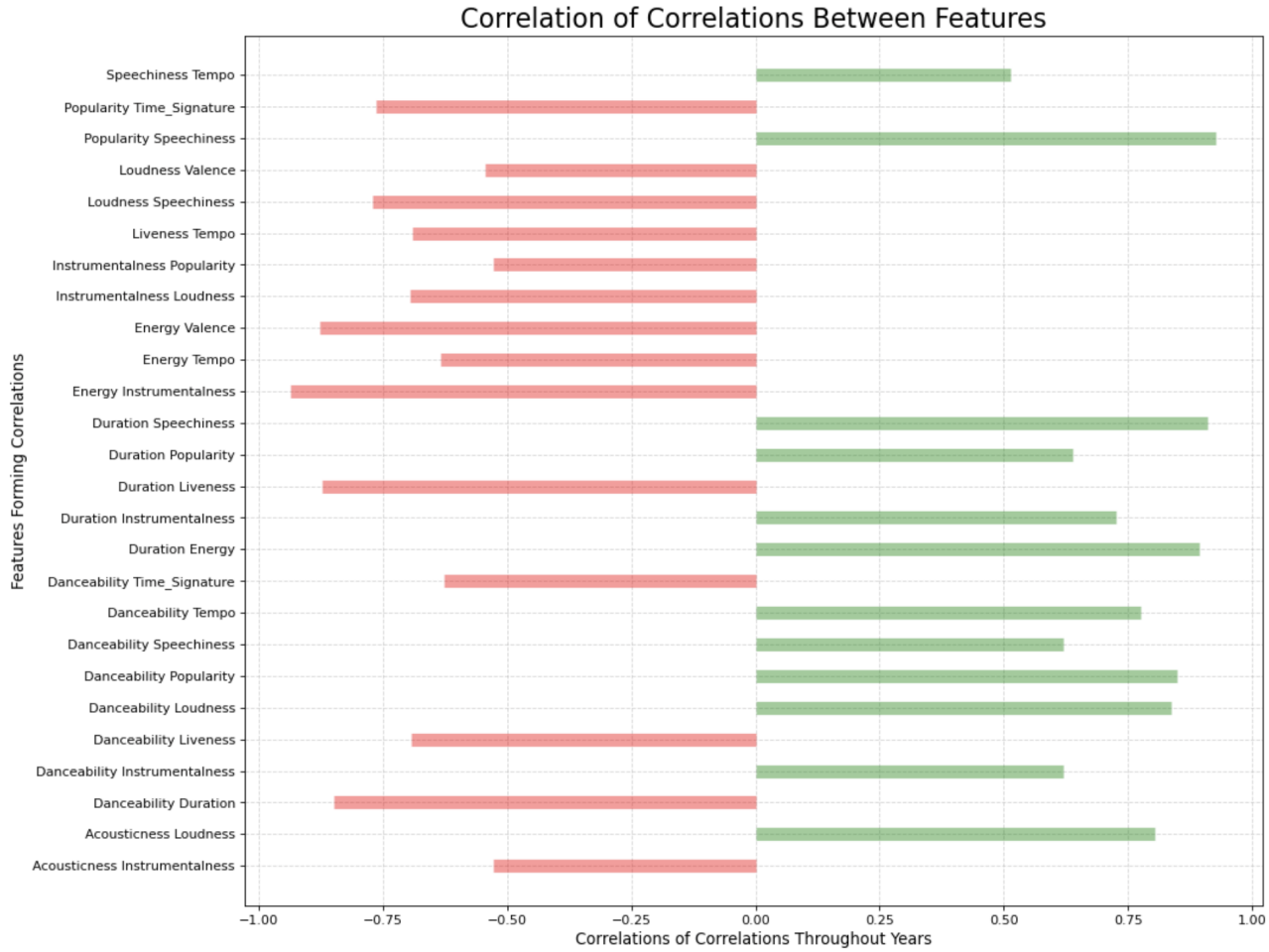
3) **Individual Features Throughout Time:**



Trends before 1922 were ignored, because there is only one observation before 1922. Songs were high in acousticness around 1920s-1950s. There were downward trend in acousticness from then on. Valence and danceability values showed similar patterns in 1920s-1960s. They then showed slight divergence until the 1990s, but both showed upward trends. Valence and danceability then showed another divergence; danceability increased while valence decreased. Energy remained consistent throughout the 1920s-1950s, staying around 0.3. It then experienced a constant upward trend until the 2020s. This explosive increase in energy can be explained by the rise of electronic music, such as EDM, hip hop, and all of their subgenres. Liveness stays consistently

low, around 0.2 throughout 1920s-2020s. This makes sense, because more polished, "radio-ready" studio recordings are likely to be popular. Instrumentalness hovers around 0.2-0.4 between 1920-1950, then decreases until the 1970s, then stays around 0.1 until 2020. This makes sense, because more popular tracks are likely to be full songs with vocalists, rather than empty instrumentals.

4) Correlation-Time Correlation:



There were twelve combinations of features with positive correlation over 0.5, eight with strong positive correlation over 0.7. This included 'Acousticness Loudness', 'Danceability Loudness','Danceability Popularity', 'Danceability Tempo', 'Duration Energy', 'Duration

Instrumentalness', 'Duration Speechiness', and 'Popularity Speechiness'. Correlations between these combinations are more likely to increase as years pass.

5. Future Work

1) **Potential Application**:
    a) This predictive model for popularity can be used on its own to recommend new songs that you have little information about their popularity.
    b) This model can also serve as a useful addition to the recommendation system by allowing us to more efficiently recommend new songs with very little information about their popularity.
    c) Finally, this model can help artists and audio technicians to learn and engineer their tracks to have audio features in popular songs.
2) **Future Work**:
    a) **Exploring the relationship between streams of songs and artists' standing in different social media-** Aside from "how good" a song is, a song's popularity can be impacted by its artist's status. Some people may initially listen to a song only because it is released by their favorite artists and grow to like it, because they listened to it so often.
    b) **Focusing on individual genres of songs and developing models that predict popularity using audio features within each genre-** one may be able to better predict popularity after training the model with genre specific data. Current model is trained on songs from a number of different genres. This can hinder model performance, as different genres each have characteristic audio features.

References:

1) https://www.billboard.com/pro/how-much-music-added-spotify-streaming-services-daily/

2) https://developer.spotify.com/discover/