Spotify: Music Recommendation System

1. Introduction

a) **Problem statement**:

Spotify has been revolutionizing and dominating the global music streaming industry since its birth in 2006. However, with the recent increase in the number of competitors, Spotify has seen a slow decline in its market share. Our goal is to refine the quality of Spotify's music recommendation system.

b) **Background**:

SinceAccording to MIDiA, Spotify has seen a slow decrease in its dominance in the music streaming industry, with its current share at 31%, down from 34% in 2019. Such shift in market dominance is further supported by the growth in Q2 2021. While Spotify only grew 20%, Amazon Music and YouTube Music grew 25% and 50%, respectively.

This points to a need for improvements to better retain old listeners and attract new customers. While there are numerous ways to accomplish this, including better interfaces, having an effective recommendation system will play a critical role.

c) **Goal**:

This project aims to develop a data-driven recommendation system to better cater to users' tastes. supported by data-driven analysis. We will be developing three different recommendation systems: content-based filtering, collaborative filtering, and model. 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 songs to

recommend similar music. We also compare users' listening histories and provide recommendations based on user similarities. Lastly, hybrid system combine content-based filtering model and collaborative filtering model. We first use a content-based filtering model to select songs that are similar to the user's recent listening history. We then use a collaborative filtering model to narrow down from the original recommendation based on other similar users' listening history.

References:

1) https://www.forbes.com/sites/eamonnforde/2022/01/19/spotify-comfortably-remains-the-biggest-streaming-service-despite-its-market-share-being-eaten-into/?sh=7fa39fed3474
2) https://techcrunch.com/2022/01/20/spotify-subscription-numbers-up-youtube-music-tidal/

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_cont**.

df_cont 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) User ID Playlist Dataset:

The dataset for users listening/ playlist pattern was also sourced from Kaggle. The data was read in as a DataFrame called **df_collab**.

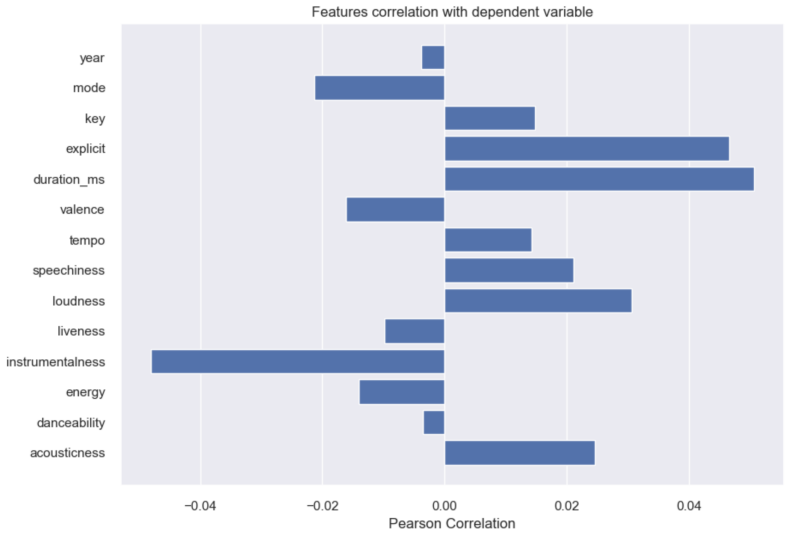
df_collab contained 4 features, which included: 'user_id', 'artistname', 'trackname', and 'playlistname'.

3. Data Cleaning and Data Wrangling

1) In **df_collab**:
    a) Quotes and white spaces were removed from column names and data.
    b) Data without playlist names were removed
    c) Data without track names and artists names were removed
    d) Columns were renamed to match those of **df_cont**:
        i) 'artistname' was renamed to 'artist'.
        ii) 'trackname' was renamed to 'song'.

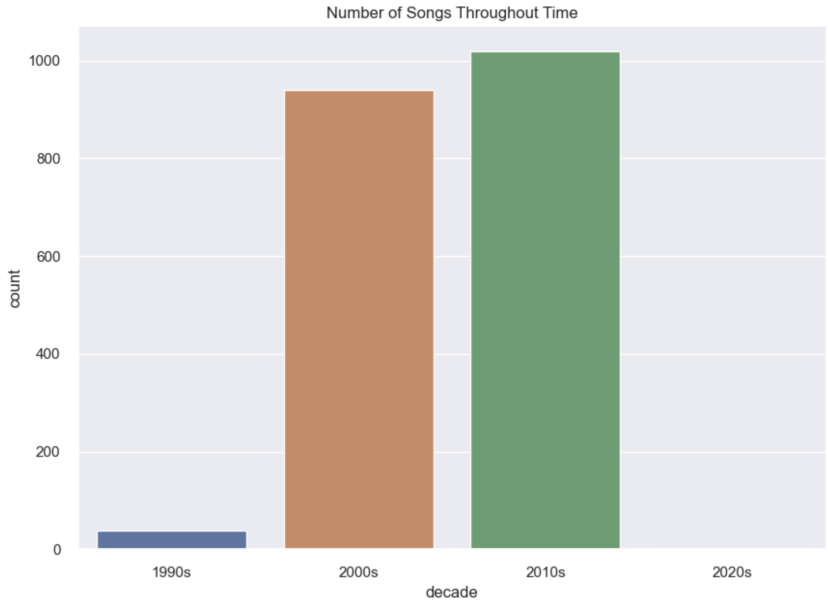
4. Exploratory Data Analysis and Initial Findings

1) **Feature Correlations with Popularity:**



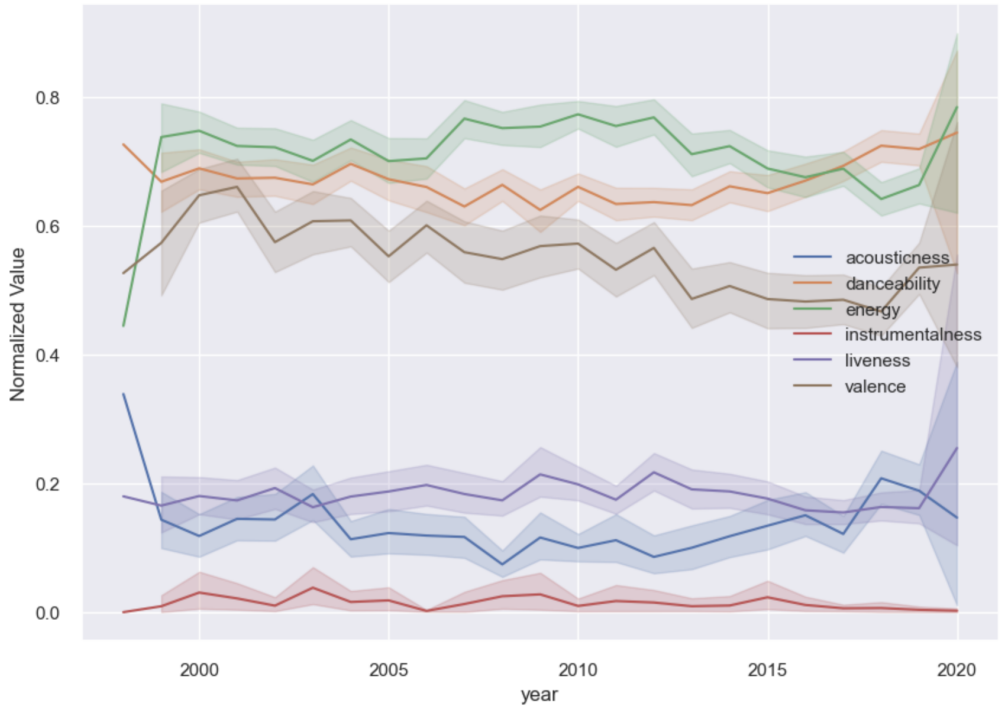
There isn't any feature with significant correlation with popularity.

2) **Number of songs**:



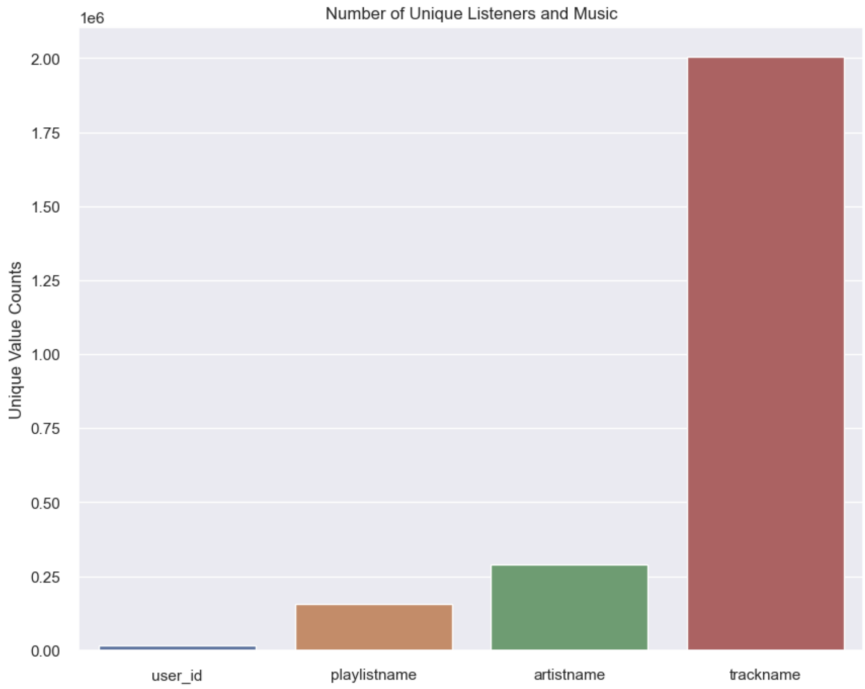
The number of top-charting songs increased from the 1990s to the 2000s. This is followed by a slight increase in the number of songs in the 2010s. There were very few songs in the 2020s.

3) Individual Features Throughout Time:



Acousticness started around 0.3-0.4, followed by a linear drop to 0.1-0.2. The acousticness stayed in that range throughout, with slight increase during 2015- 2020 region. Valence experienced overall decline from 0.6-0.7 range to 0.4-0.6 range. Danceability stayed constant 2000- 2015 within 0.6 - 0.8 range. Energy remained consistent throughout the 1920s-1950s, around 0.7. While it experienced a decline between 2010-2017, it recovered to upward of 0.7. Instrumentalness stayed consistent in the 0.0- 0.1 range.

4) **Comparison of number of unique listeners, playlists, artists, and tracks:**



There are more unique tracks and unique artist names than unique listeners and unique playlists. This makes intuitive sense, given that listeners tend to add more than one song to each of their playlists. There are more unique playlists than unique listeners. This may be because some listeners may have more than one playlist. Finally, there are more tracks than artists. This may happen if listeners saved more than one track from a given artist.

5. Future Work

1) **Potential Application**:

   a) This recommendation system can be used on its own to recommend songs to users we have little information about.

b) This model can be used to effectively and tastefully recommend songs which we have little information about.

2) **Future Work**:

a) **Supplement with a model that predicts song popularity to alleviate cold start problems**- it is important to be able to recommend songs that users would enjoy. One factor that plays a critical role is song's popularity. Being able to predict songs' popularities given audio features and making initial recommendations based on popularity prediction may offer potential solution to a cold start problem.