École Polytechnique fédérale de Lausanne

COM-480 – Data visualization – Spring 2024

Project : First Milestone

Hanwen Zhang
Heling Shi
Jingren Tang

1 Dataset

The dataset "Top Hits Spotify from 2000-2019" encompasses comprehensive information for about 2000 songs, offering a rich resource for analysing the music industry. The dataset is clean with no missing values. Each entry in the dataset contains the artist's name, song title, and other key features relevant to its popularity, musical attributes, and lyrical content. These characteristics include the track's duration, explicit content, release year, and popularity rating. Musical features such as danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, genre, and the overall estimated tempo of a track in beats per minute (BPM). The rich information enable the data scientists to explore and understand the trend, preference, and evolution over the first 2 decades of the 21st century. The dataset, popular on Kaggle and consisting of approximately 2000 rows, contains no NaN values. Therefore, it minimizes the need for extensive cleaning.

2 Problematic

This project delves into the evolution of popular music characteristics over time, highlighting variations in song popularity and the corresponding key musical traits such as speechiness and danceability. Time-series analysis will be used to illustrate these changes, dynamic distribution graphs be plotted to showcase the variety in musical attributes across different periods. Additionally, we will provide a straightforward, intuitive visualization of how these characteristics have varied over the years, without deep analysis. This approach aims to offer a clear and immediate understanding of music trends and their evolution.

— **Motivation :** Identify changes in preferences for song styles and lyrics across different years, enabling songwriters and listeners to deepen their understanding.

— **Target Audience :** Songwriters and listeners interested in the trends of music creation.

3 Exploratory Data Analysis

3.1 Dataset Overview

- There are 2000 songs in our dataset, each with 18 attributes. No missing values in the dataset.
- The dataset spans songs from 1998 to 2020. However, due to the limited entries for 1998 (1 song) and 2020 (3 songs), we will not take 1998 and 2020 into consideration in the future analysis.
- There is a total of 835 different artists counted.

Here are two bar charts shown. Figure 1 illustrates the distribution of song counts per year from 2000 to 2019, with noticeably sparse data for the years 1998 and 2020. Figure 2 illustrates the top 15 artists ranked by the number of songs they have produced in the dataset. Rihanna has the highest count(25 songs), followed by Drake(23 songs) and Eminem(21 songs).

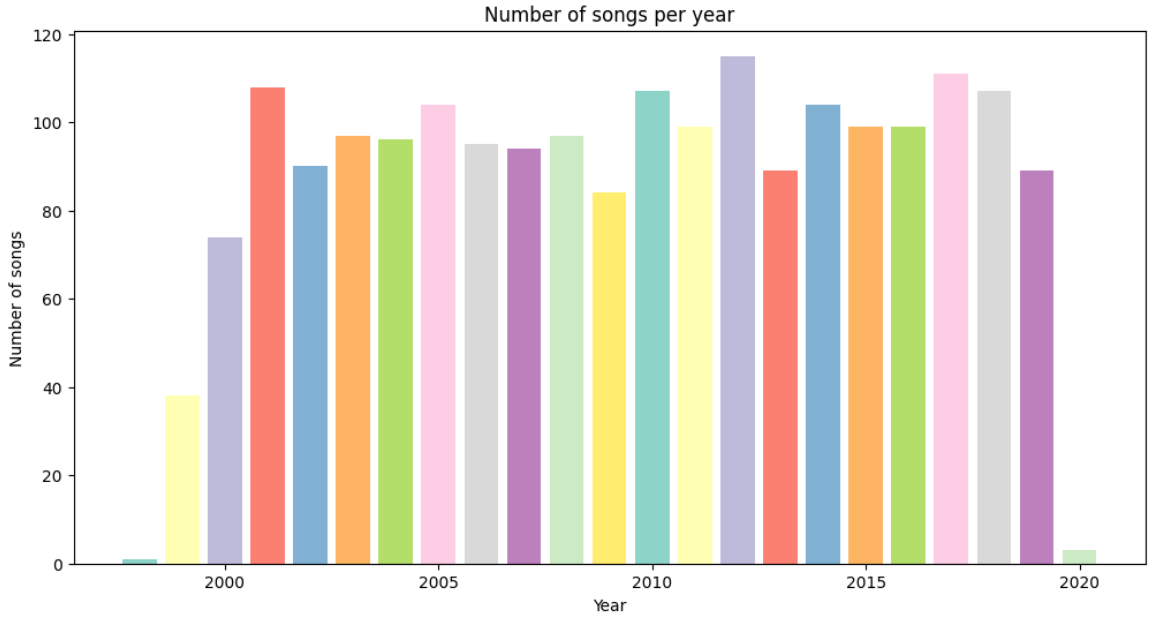

FIGURE 1

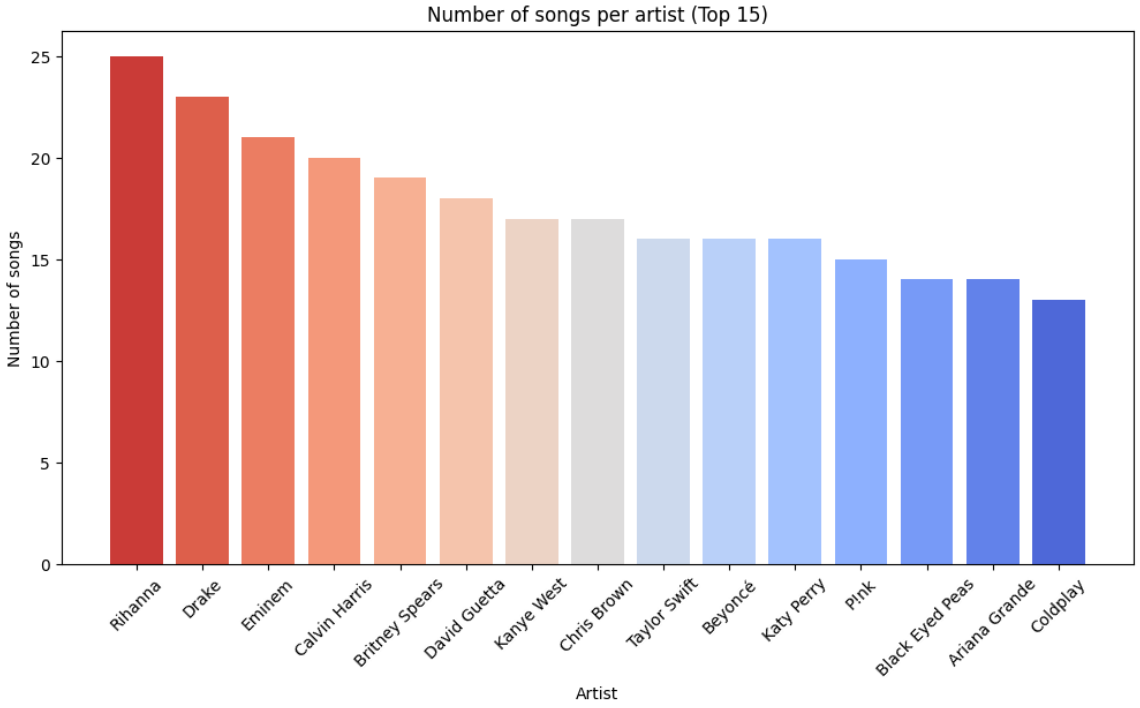

FIGURE 2

3.2 Features Distribution

The dataset's features can be categorized into two types : categorical and numerical. For categorical features, use pie charts to represent the proportional distribution of each category. For numerical features, use histograms to depict the distribution.

3.2.1 Feature Engineering

- The introduction of the original dataset mentions that although speechiness and instrumentality are numerical variables, they can be categorized into types by given thresholds, we convert these two variables into categorical variables to show the distribution.
- For the attribute "Genre", a song in the dataset can include multiple genres, we split and count the total number of occurrences of each genre.

3.2.2 Categorical Feature

- The majority of songs in the dataset are characterized as non-explicit, primarily musical rather than speech-centric, and predominantly vocal.
- The songs in different keys and modes are relatively evenly distributed.
- Pop is the most predominant genre, followed by hip-hop and R&B.

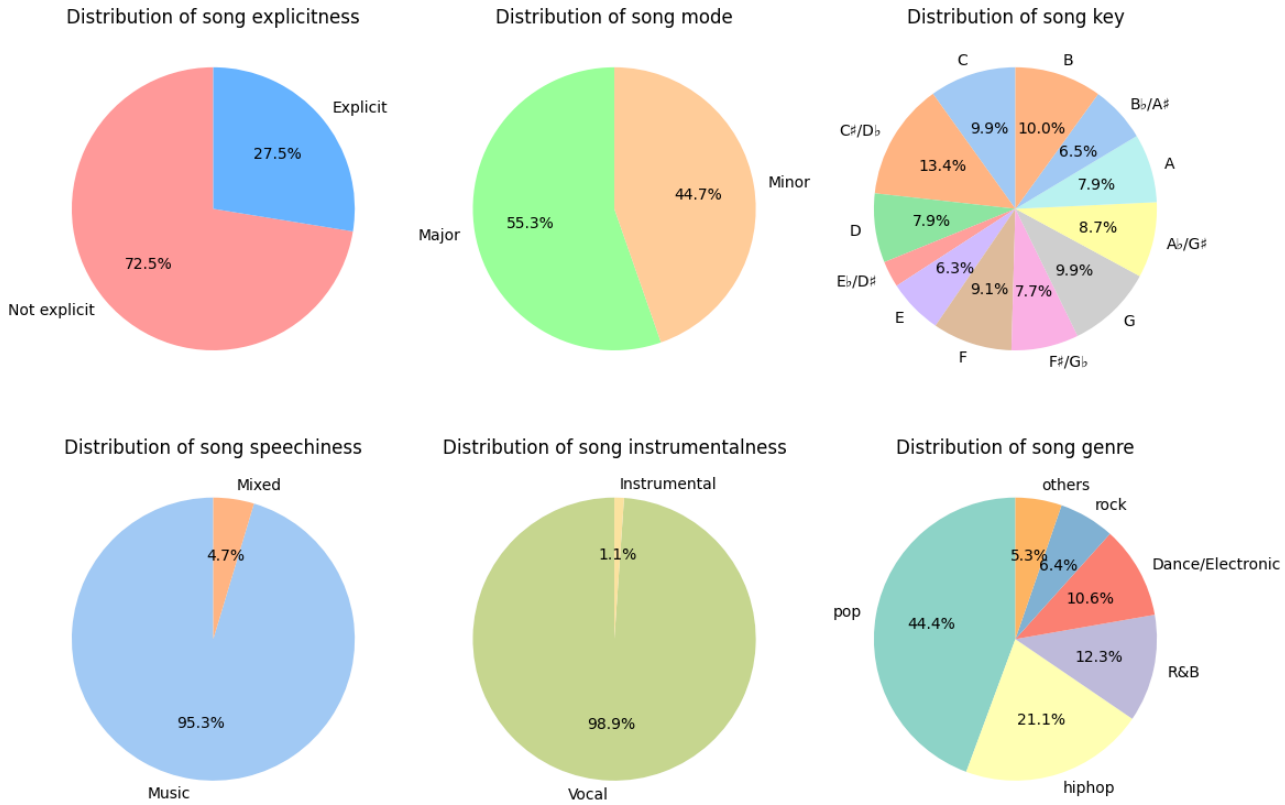

FIGURE 3

3.2.3 Numerical Feature

The histograms display the distribution of musical attributes including acousticness, popularity, tempo, energy, loudness, danceability, liveness, valence, and song duration.

- The majority of the songs in the dataset have lower acousticness and higher popularity.
- The valence of the songs are diverse.
- The songs were recorded mainly in the studio.(predominant low liveness)

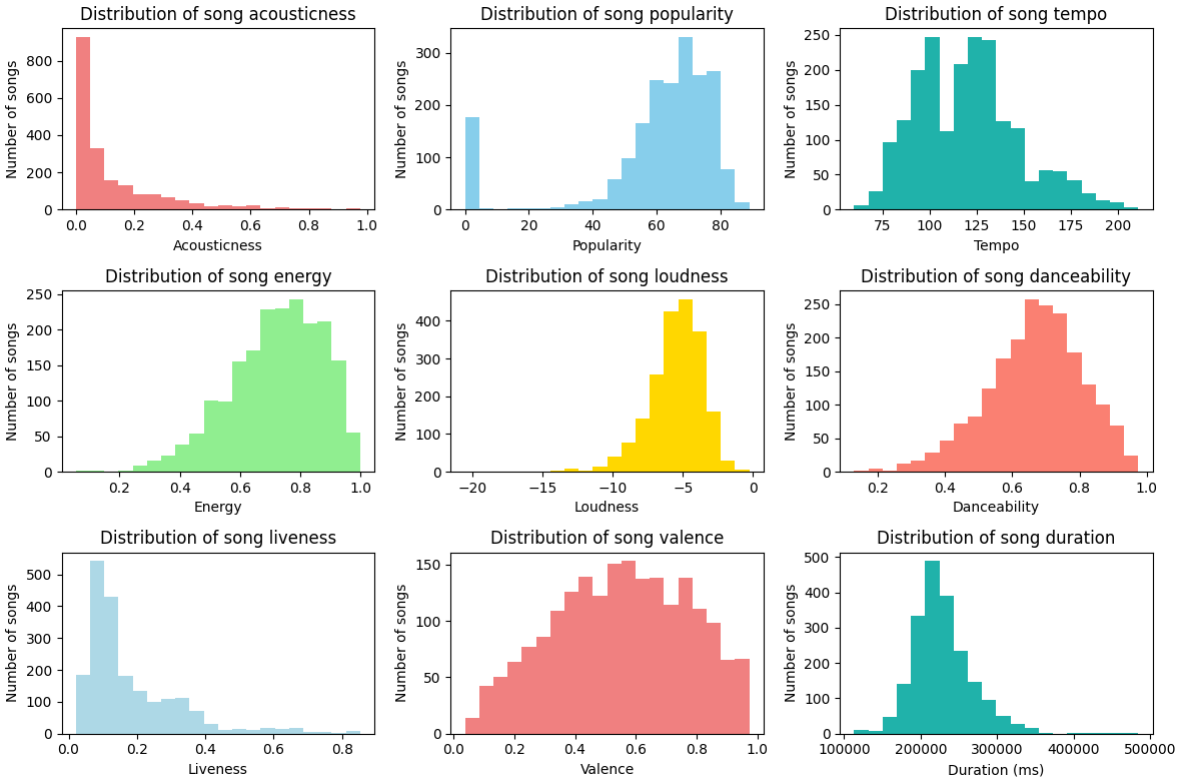

FIGURE 4

4 Related work

4.1 Previous work with this data

As a high quality dataset on Kaggle, many users posted their processes and results of data analysis and visualisation using python.

A Kaggle user, Varun Sai Kanuri[1] created a heatmap to visualize the pairwise correlation of columns of the dataset and identified those with high relevance(loudness and energy, energy and dancability, speechiness and popularity). He created a treemap of singers playlist, in which each rectangle represents an artist, and they are grouped and labeled by genre. The size of each rectangle corresponds to the popularity of the artist.

Ahmed Hafez[2] showed the average popularity of the songs of the 10 artists with the most songs included, as well as the distribution of chords in these popular music.

4.2 Originality

Although some attempts at visualisation have been made, they tend to be exploratory, without a clear theme, and do not convey information very efficiently and intuitively.

Building on previous efforts, our work not only presents a clearly defined theme but also can convey more meaningful information through enriched interactive methods. Additionally, by integrating other datasets (such as corresponding music videos on YouTube and song lyrics), we are further able to explore the characteristics of these songs.

4.3 Source of inspiration

A treemap can be utilized to illustrate the frequency of each song in the dataset, each rectangle representing a song [3]. The rectangle size reflects the song's occurrence rate : larger rectangles stands for more frequent appearances. Colors differentiate genres or artists. Also, using a box plot visualization can showcases the distribution of various musical attributes across the songs. Each box represents a different attribute, such as Energy, Danceability, Speechiness, Valence, Loudness, and Tempo.

Références

[1] Kanuri, V. S. (2023). Spotify Data Visualization. Kaggle.
https ://www.kaggle.com/code/varunsaikanuri/spotify-data-visualization/notebook

[2] Hafez, A. (2023). Spotify Data Analysis and Visualization. Kaggle.
https ://www.kaggle.com/code/ahmedtronic/spotify-data-analysis-and-visualization

[3] Bailey, B. (2024, March 24). Taylor Swift Lyrics TreeMap. Tableau Public. Retrieved from https://public.tableau.com/app/profile/blair.bailey/viz/TaylorSwiftLyricsTreeMap/Dashboard.