Song Data Analysis

ShaNiece Twitty

2025-02-10

Introduction

This dataset, sourced from Spotify and obtained via Kaggle, contains 2,000 tracks released between 1998 and 2020. I chose this sample due to my love for music and curiosity about potential trends. Through exploratory data analysis, I aimed to identify visible patterns and changes in music over time that could be effectively visualized. The dataset includes specific tracks and features variables related to music, some of which I hadn't realized could be measured. However, I am not completely confident in the set as you will see in the rest of the report, there appears to be some biases in the data that may affect the analysis.

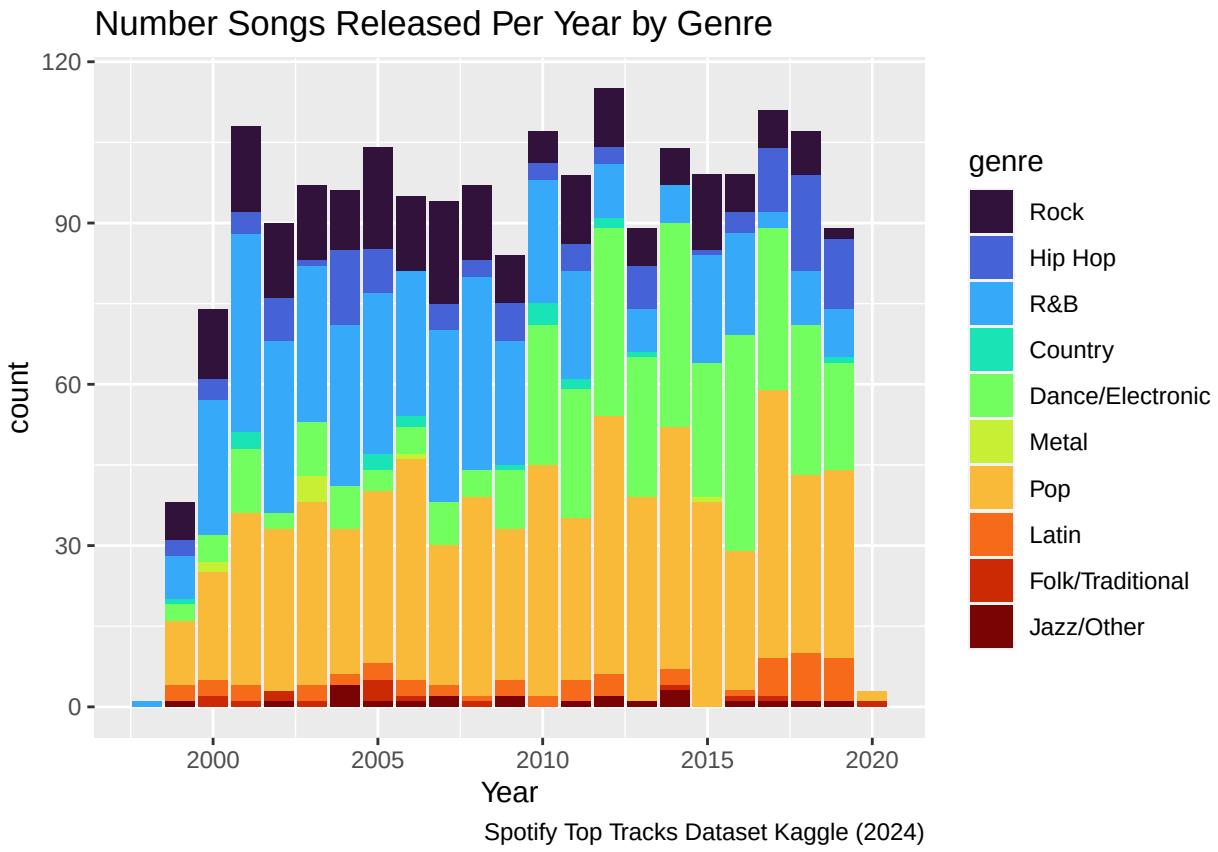
Data Cleaning

To further my exploratory analysis of the sample the data set underwent cleaning for easier usage. Initially, each genre was represented in its own individual binary column. To alter this, the genre columns were pivoted and a new categorical column was created. However, some songs were assigned to multiple genres and the new column lead to entry duplication. To resolve this the distinct function applied to keep only one of the tracks. Additionally, a genre priority system was implemented using the combine function to aid in assigning one genre to each song based on a predefined hierarchy. The final data set, "Songs2.0", includes these edits and has been structured to facilitate analysis.

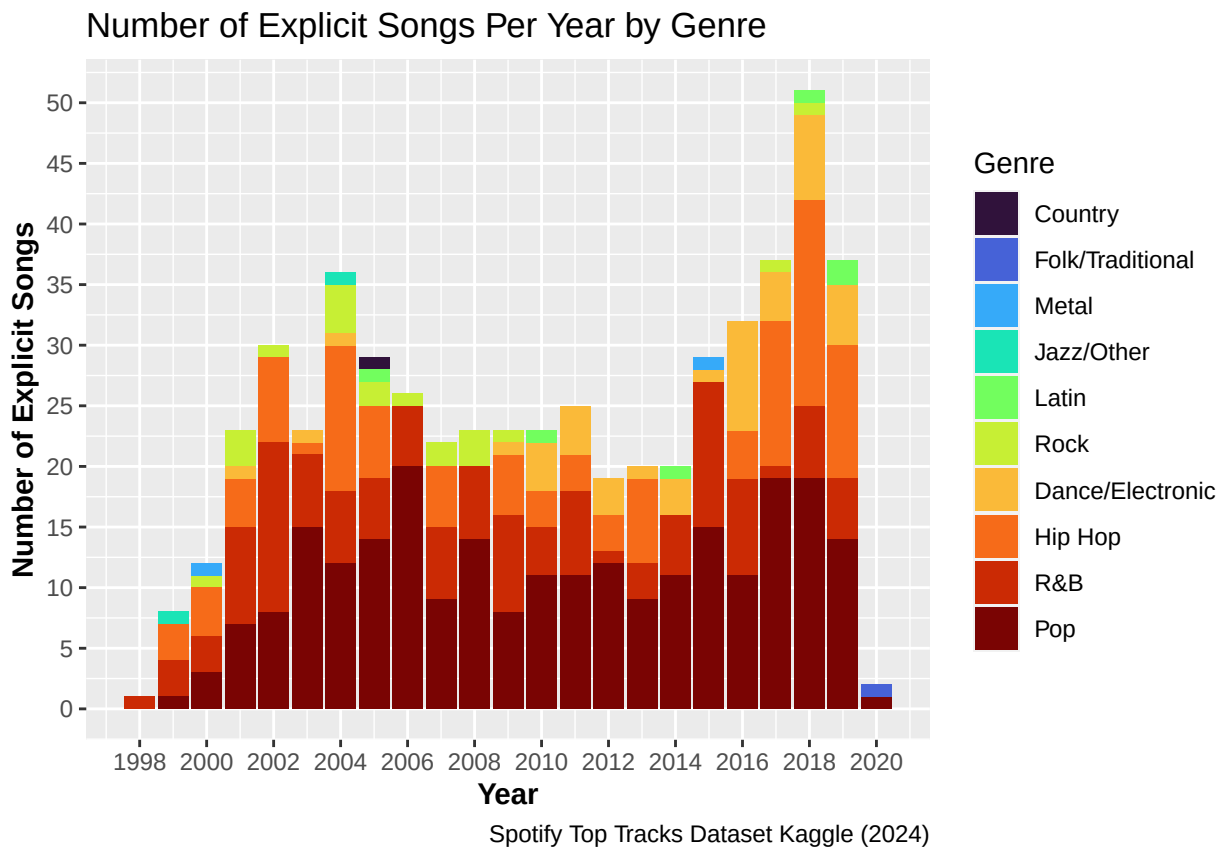
Table 1: Description of Variables

Variable	Description
entry_number	The entry number of the track
duration	Duration of the track in milliseconds
explicit	Indicates whether the lyrics or content of the song are considered explicit (Boolean expression represented in binary)
year	Release year of the track
popularity	A measure of the track's popularity (higher values indicate more popular 0-100)
danceability	Describes how suitable the track is for dancing (0.0 least danceable, 1.0 most danceable)
energy	Represents the perceived intensity and activity of the track (0.0 to 1.0)
key	The key the track is in (integers map to standard pitch class notation)
loudness	The overall loudness of the track in decibels (dB)
mode	Indicates the modality of the track (0 for minor, 1 for major)

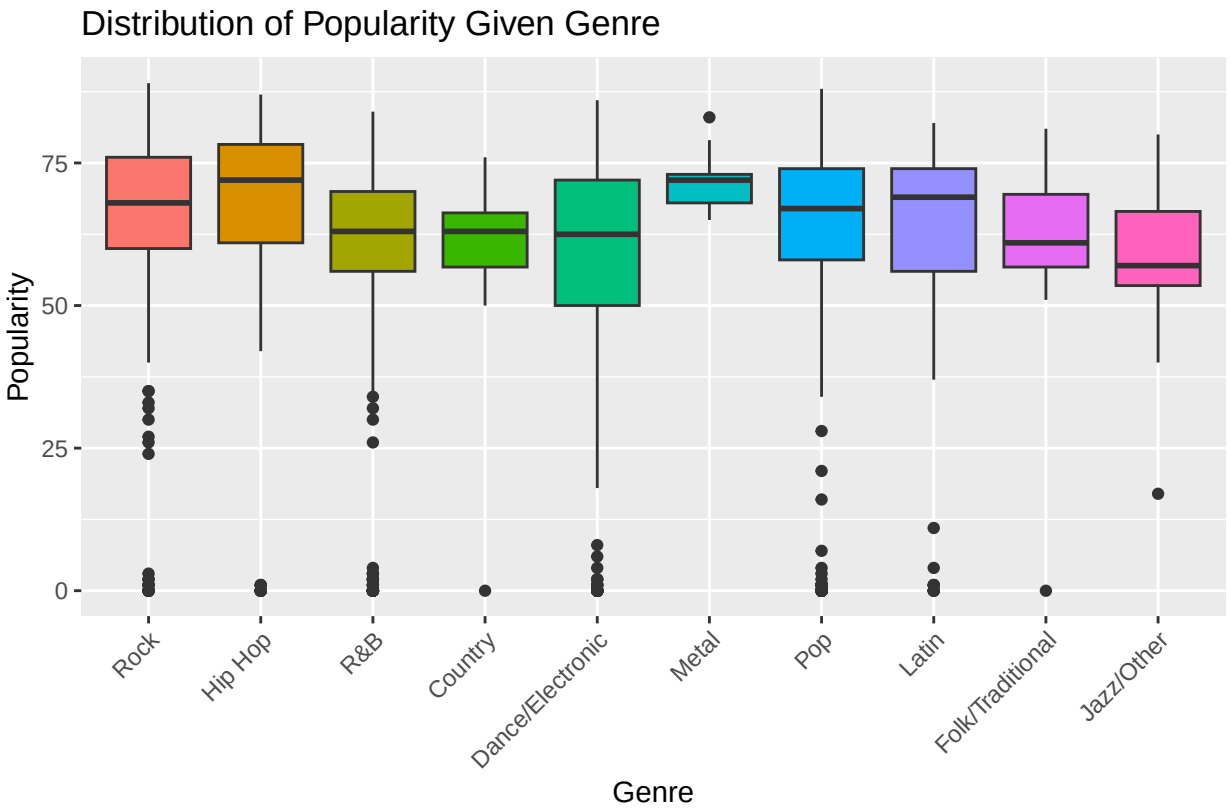
Variable	Description
speechiness	Detects the presence of spoken words in the track (0.0 to 1.0)
acousticness	A confidence measure of whether the track is acoustic (0.0 to 1.0)
instrumentalness	Scale of the amount of vocals on the track (0.0 to 1.0)
liveness	Detects the presence of an audience in the recording (0.0 to 1.0)
valence	Describes the musical positiveness conveyed by the track (0.0 to 1.0). Tracks with high valence sound more positive (happy, cheerful, euphoric), while tracks with low valence sound more negative (sad, depressed, angry)
tempo	The overall estimated tempo of the track in beats per minute (BPM)
genre	The category of artistic composition, based on for style and subject matter



This stacked bar chart visualizes the annual count of songs released, with colors representing the contribution of each genre to the total, relative to the data. The chart highlights the dominance of the Pop genre throughout the years, with R&B making a notable contribution and Dance/Electronic gaining prominence from 2010 onward.



This stacked bar chart visualizes the annual count of explicit songs released, with colors representing each genre's contribution to the total, relative to the data. The chart highlights Pop as the dominant genre, followed by Hip-Hop and R&B.



This box plot displays the distribution of popularity ratings for tracks, grouped by genre. Each box has a relatively small interquartile range, though all genres exhibit outliers. Metal has the smallest interquartile range and the highest average popularity, however has limited contribution to the total number of released songs, as seen in previous graphs. In contrast, Dance/Electronic shows the widest range in popularity ratings.

Conclusion

Given more time for this analysis, I am unsure if statistical tests are worth exploring. However, one key area of interest would be calculating the proportions to determine the relative probability of a track belonging to a specific genre, given that it is explicit. Additionally, I would analyze the annual distribution of genre popularity through visualizations.

One notable limitation is that the data source does not specify how popularity is calculated. It is possible that genres with fewer releases, such as Metal, maintain more consistent popularity due to a possible more highly targeted audience. This trend is evident in the box plot, but it would be valuable to examine how it holds up with a larger data set that offers equal genre representation.

The data set appears to be heavily skewed toward Pop, R&B, and Dance/Electronic, suggesting possible over representation. However, this could also reflect the actual distribution of released songs—for example, more Pop songs may be released annually than Jazz. Investigating broader music industry trends could provide further insight into this imbalance.

STwittyDataLab/SongsRepository