PHS 650: Final Project

Aly, Haley, Zoe

4/26/23

Table of contents

Introduction

In the music industry, there is a consideration of what song will be popular or a ‘hit,’ as song popularity is associated with more revenue (Pham, Kyauk, and Park 2015). Thus, predicting the popularity of a song, referred to as the study of Hit Song Science, can be useful in determining which songs should receive the most investment from musicians and their labels. Random forests have been found to accurately predict which songs will be popular and determined that songs that ‘made it’ to the top charts were found to be ‘happier’ and more ‘party-like’ (Interiano et al. 2018; Middlebrook and Sheik 2019). Additionally, artist familiarity, loudness, year of release, and number of genres were also found to accurately predict the popularity of songs (Pham, Kyauk, and Park 2015).

The goal of our project is to add to study of the Hit Song Science and examine how song elements, specifically the duration and intensity, are associated with the song’s popularity. We hypothesize that shorter songs are more likely to be popular and more intense songs are more likely to be popular.

Methods

The following analysis will use data from the Tidy Tuesday Spotify Songs dataset. The dataset contains 32,883 songs in the genres of EDM, Latin, Pop, R&B, Rap, & Rock and 23 variables describing characteristics of the songs and the playlists they were found in. The data dictionary can be found here. We will be using all the songs in the Tidy Tuesday Spotify datasets. The only exclusion criteria we will apply is to remove duplicate songs, indicated by track_id. The data was accessed 4/5/2023.

The R code to import the dataset can be found below:

```
spotify_songs <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytu
```

Applying the exclusion criterion reduces the sample size from 32833 songs to 28356 songs.

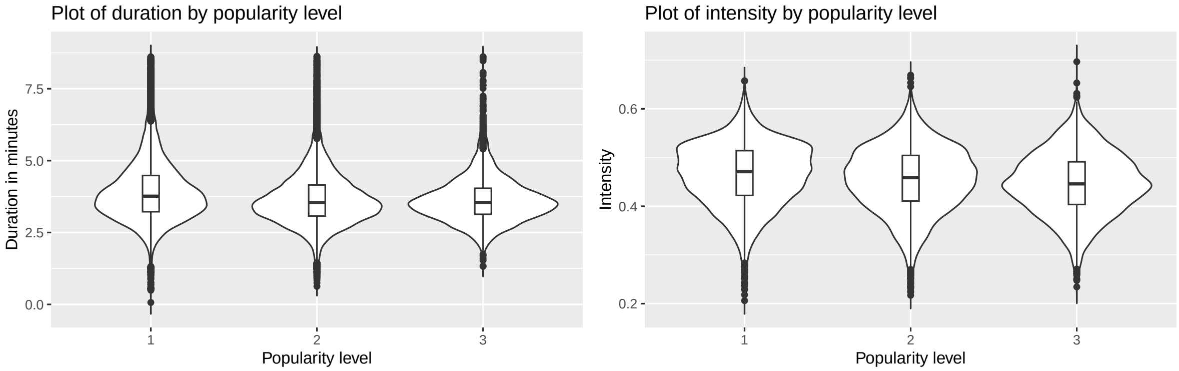
To complete the present analysis, 5 new variables were created. Table 1 describes these variables, norm.loudness, norm.tempo, itensity, duration_mins, popularity created from original loundness, tempo, and popularity variables from the Spotify dataset. The table provides the class, range, and description of each new variable.

Table 1: Spotify Data variables

Variable	Class	Range	Description
loudness	double	-46.448 to 1.275	The overall loudness of a track in decibels (dB) averaged across the entire track.
tempo	double	0-239.44	estimated track tempo in beats per minute (bpm)
norm.loudness	double	0-1	a min-max normalized spotify_songs$loudness
norm.tempo	double	0-1	a min-max normalized spotify_songs$tempo

Variable	Class	Range	Description
intensity	double	0-1	the average of energy, normalized tempo, and 1- normalized loudness, where higher scores imply higher intensity
minutes	double	0.07-8.6 minutes	The duration of song in minutes, converted from milliseconds (duration_ms)
popularity	double	1, 2, 3	Song popularity characterized into three tertiles. 1 represents low 0-33, 2 represents medium 34-66, and 3 represents high 64-100 ranges from the numeric track_popularity variable.

Results



(a) Duration

(b) Intensity

Figure 1: Duration and Intensity by Popularity Level

From Figure 1 we see that the distributions of intensity and duration are visibly skewed within popularity classes. As a result, we used a Spearman rank correlation to test each hypothesis. In Table 2 we see that both p-values are well below our pre-established cutoff of 0.05. Since we are doing multiple statistical tests, we could adjust the p-values to account for this. However, since there are only two tests and the p-values are several orders of magnitude smaller than our cutoff, this is unnecessary. Both duration and intensity were negatively correlated with popularity to a weak degree; neither coefficient was on the expected scale. Since intensity was a composite measure, exploration of how the components relate to popularity could explain why the direction of correlation was unexpected.

Table 2: Spearman Rank Correlation Results

Factor	p-value	Coefficient
Duration	$1.7345287 \times 10^{-83}$	-0.1145858
Intensity	$9.9552206 \times 10^{-66}$	-0.1014288

References

Interiano, Myra, Kamyar Kazemi, Lijia Wang, Jienian Yang, Zhaoxia Yu, and Natalia L. Komarova. 2018. "Musical Trends and Predictability of Success in Contemporary Songs in and Out of the Top Charts." *Royal Society Open Science* 5 (5): 171274. https://doi.org/10.1098/rsos.171274.

Middlebrook, Kai, and Kian Sheik. 2019. "Song Hit Prediction: Predicting Billboard Hits Using Spotify Data." arXiv. https://doi.org/10.48550/arXiv.1908.08609.

Pham, James, Edric Kyauk, and Edwin Park. 2015. "Predicting Song Popularity."