Analysis of the Popularity of Spotify Songs from 2010-2019 with Multilevel Model

Wang Jingyao

2022-12-05

Abstract

Spotify is a music player that is used all over the world. It contains a very large number of songs, but there are always a few songs that get the most play. In this report, we selected the dataset "Spotify Top 100 Songs of 2010-2019" and analyze the factors affecting the popularity of songs by fitting a multilevel model, using the release time of songs as a group.

Introduction

People have music everywhere in their lives, music playback app on cell phones, stereo at home, or background sound in restaurants. Sometimes doctors also use music therapy to treat patients. Music can soothe people's emotions and also make people more passionate. From early CDs to later MP3s, and now smartphones, the way people listen to songs is becoming more and more convenient. Different people have different preferences for music. For example, some people like the blues, some like country music, and some like songs in certain languages. Spotify is a worldwide music playback app which contains tens of thousands of songs. You can listen and download your favorite music on this app. This app makes it very convenient to hear songs from all over the world. Although there is a very large number of songs, there are always a few songs at the top of the playlist. Spotify has a module on the home page called "The best of 2022", which showcases the most popular songs of the year. But what factors influence the annual changes of this list? This report will use the multilevel model to analyze the factors that influence the popularity of music each year.

Method

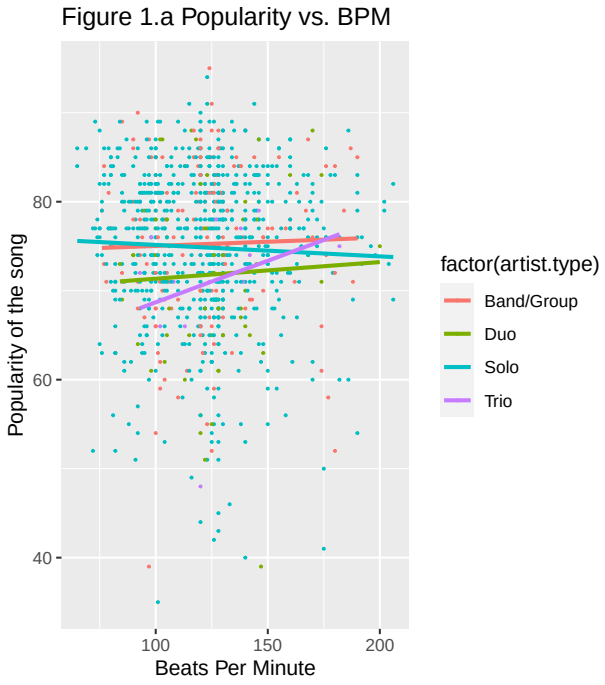
Data Cleaning

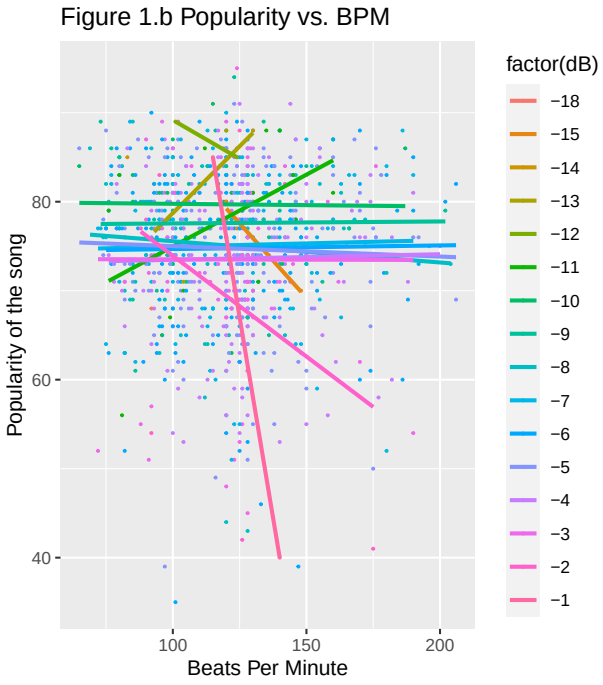
The data that we choose is named "Spotify Top 100 Songs of 2010-2019", which is download from Kaggle (https://www.kaggle.com/datasets/muhmores/spotify-top-100-songs-of-20152019?select=Spotify+2010+-+2019+Top+100.csv) and it was collected on March 20th, 2022. The orginal data has 1003 observations and 17 variables. We now cleaning the data by omit NA firstly, which have 1000 observations left. By checking the uniqueness of the year that the song has been released, it contains year that is to early or with no information. We use `filter()` to skim the data and now we have 999 observations left. The last step is to find out the duplication of the songs' title. By using `duplicated()` for the `title` column, our final dataset has 944 observations and 17 variables. To fit the model better, we calculated the time difference between the year 2022, which is the year that the data was collected, and the time of song release, and add it in the dataset of the column with name`time.released`. The table below demonstrates the explaination of each variable name in this dataset:

column names	explanation
title	Song's Title
artist	Song's artist
top.genre	Genre of song
year.released	Year the song was released
added	Day song was added to Spotify's Top Hits playlist
bpm	Beats Per Minute - The tempo of the song
nrgy	Energy - How energetic the song is
dnce	Danceability - How easy it is to dance to the song
dB	Decibel - How loud the song is
live	How likely the song is a live recording
val	How positive the mood of the song is
dur	Duration of the song
acous	How acoustic the song is
spch	The more the song is focused on spoken word
pop	Popularity of the song (not a ranking)
top.year	Year the song was a top hit
artist.type	Tells if artist is solo, duo, trio, or a band

Exploratory Data Analysis

Before we fitting the model, we would like to select the variables by checking the relationship between each different factors and the popularity of the song. We choose `artist type` and `dB` as our random factors. The plots below demonstrates the relationship on this two group levels.





The two plots above demonstrates the relationship between the popularity of a song and the BPM on the artist type level and decibel level separately. From Figure 1.a, we can notice that the intercept does not change much which we assume that BPM is not a essential factor that influence the popularity of the song. In Figure 1.b, most of the intercepts stay the same but some intercepts are changing along with the decibel changes.

Figure 2.a Popularity vs. Energy

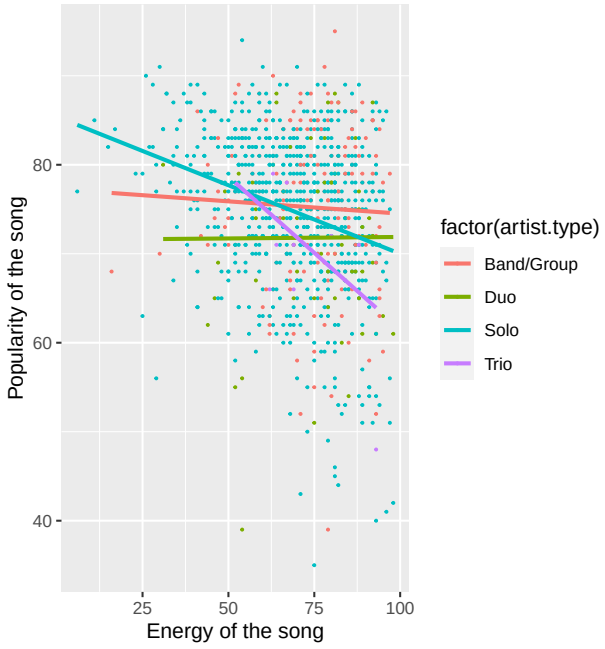


Figure 2.b Popularity vs. Energy

Figure 3.a Popularity vs. Live recording

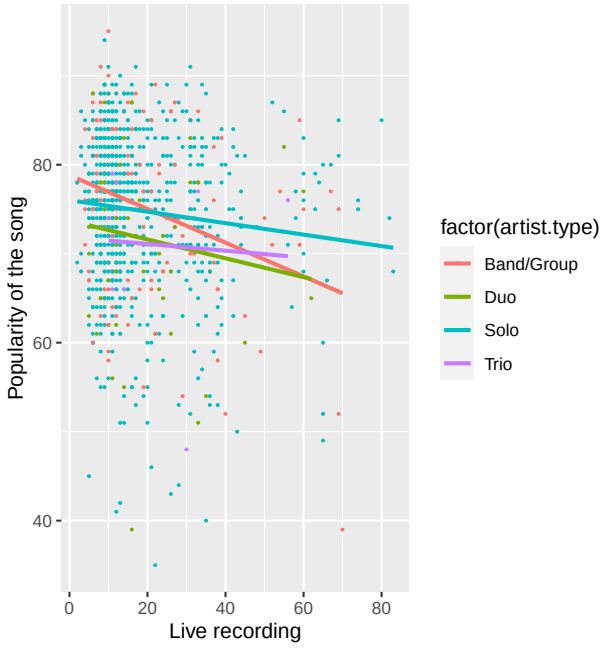


Figure 3.b Popularity vs. Live recording

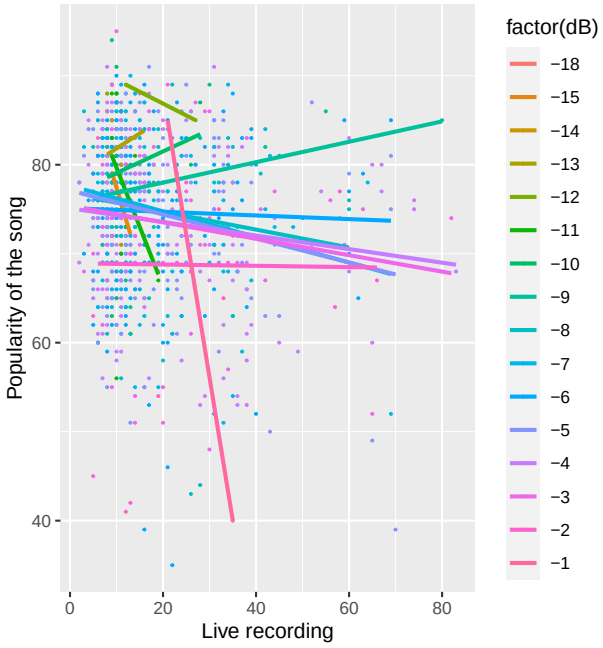


Figure 2 shows how the popularity of a song varies with the energy of a song and Figure 3 indicates the relationship between popularity and how likely a song is live recording.The overall trend is decreasing with this two factors. Based on the level of dB, the overall trend is decreasing but there are a few intercept is climbing. The intercept vaires with different dB.

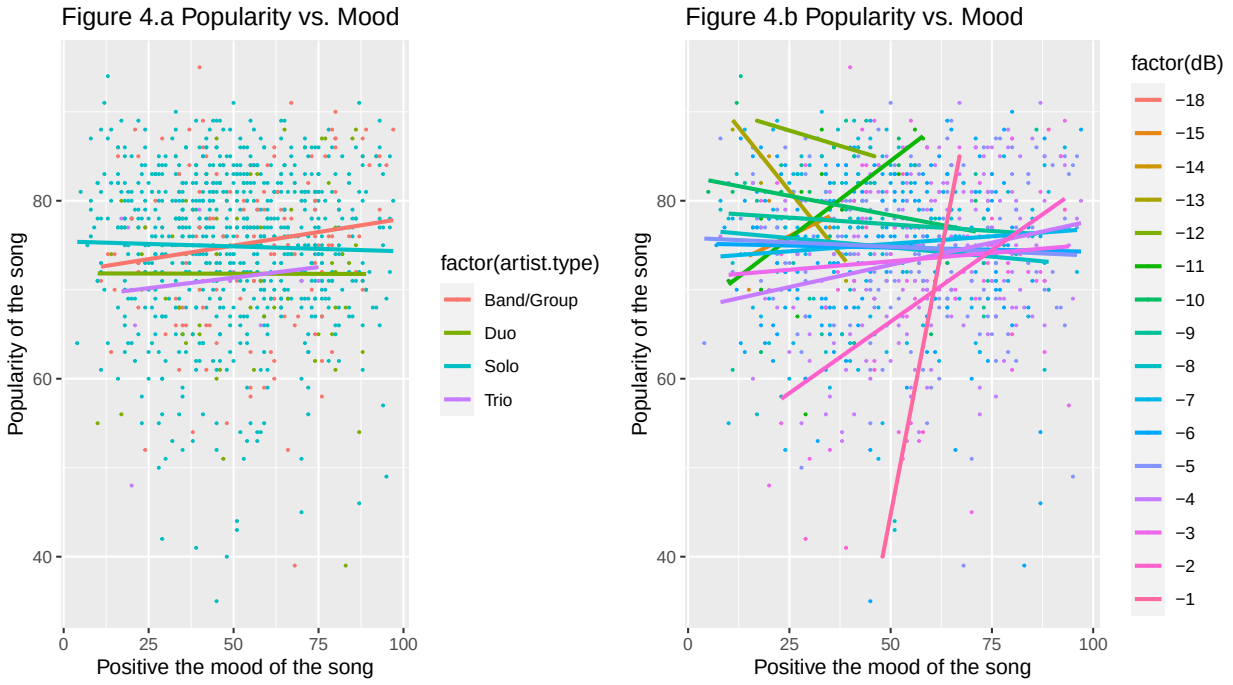
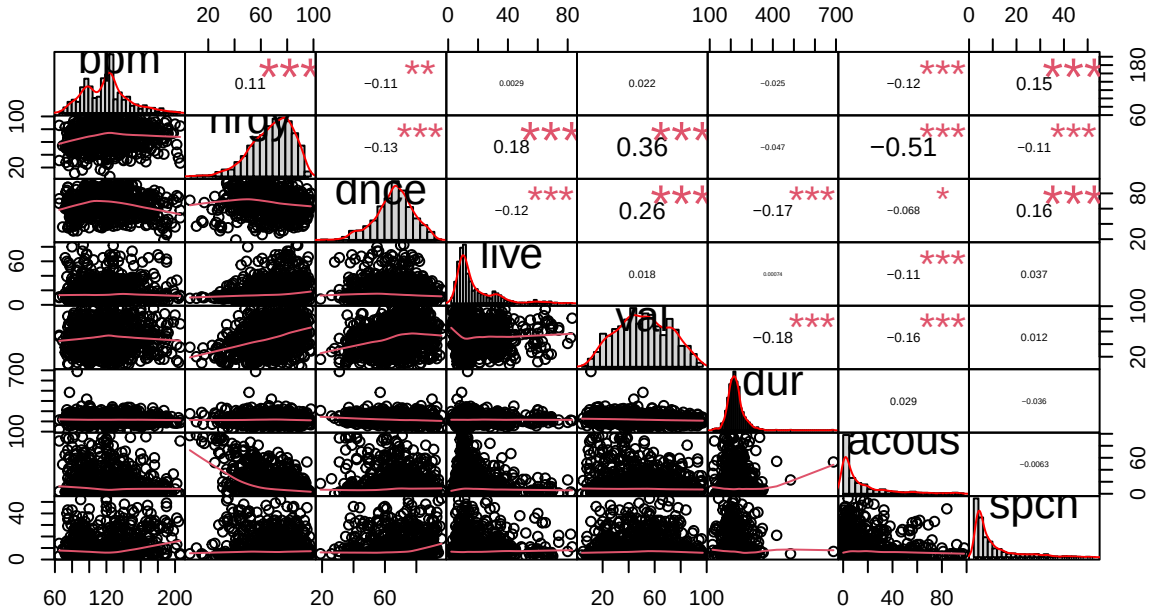


Figure 4 shows how the mood of a song has an impact on the popularity of the song. A higher value indicates a more positive mood, in both figure 4.a and 4.b, it has a upward trend based on `artist type` and `dB` levels.

Correlation Check

We now want to check the correlation between each pair of variables. The figure below is a correlation plot. The closer the number is to -1 or 1, the greater the correlation between the two variables is.



Fitting Model

We take the popularity as the dependent variable of the model and select the independent variables from our dataset. From the correlation plot, the factor energy has a higher correlation with acoustics which is -0.51, so we only select energy as one of the variables of the model. The other factor that we choose is how likely the song is a live recording. Whether a song is recorded live or not will influence the effect of the song, such as the background sound, the volume of vocal, etc., which will cause people to be willing to listen to it or not. We also choose how positive the mood of the song is to be our variables. Music makes people's life more colorful, and music with different emotions will bring people different feelings. Time has always been a factor influencing popularity, so we also include `time.released` as a variable in our model, which is the time since the song has been released until now. These are all the fixed effects that we choose for our multilevel model, we also include `artist.type` and `dB` as our random factors.

```
model <- lmer(pop ~  nrgy + live + val + time.released + (1|artist.type) +
                (1|dB), data = spotify2)
```

VIF Results

We are now checking the variance inflation factor which helps us to detect multicollinearity in our analysis. This can help us to make sure that the variables that we selected does not have strong correlation. All the results are close to 1 which means that there is no strong correlation between each varaibles.

nrgy	live	val	time.released
1.267289	1.044383	1.151505	1.096528

Fixed Effects

The table below is the result of the fixed effects of our model. All the P-values are significant which are smaller than 0.05.

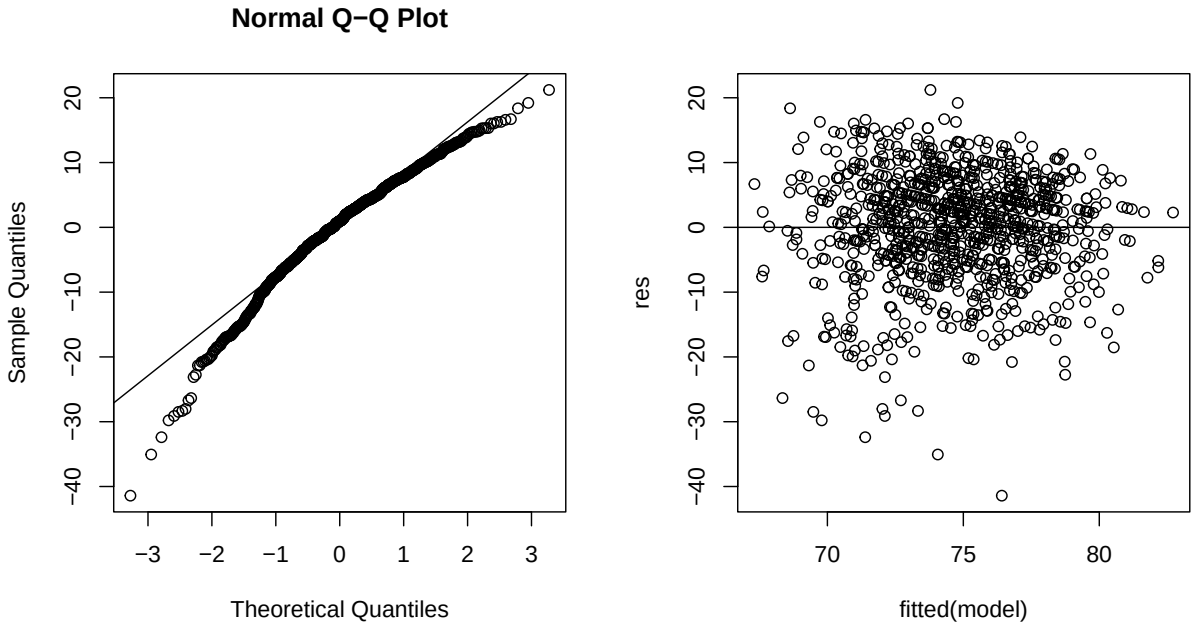
	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	84.33698	1.67145	14.89866	50.457	< 2e-16 ***
nrgy	-0.11671	0.01965	938.53282	-5.940	4.00e-09 ***
live	-0.05576	0.02075	936.12460	-2.688	0.00732 **
val	0.04102	0.01384	937.27356	2.963	0.00312 **
time.released	-0.43563	0.09643	937.72803	-4.518	7.06e-06 ***

Random Effects

Groups	Name	Variance	Std.Dev.
dB	(Intercept)	0.000	0.000
artist.type	(Intercept)	3.127	1.768
Residual		71.695	8.467

Model Checking

For model checking, we create Normal Q-Q Plot and the Residuals vs. Fitted Plot. The data is normally distributed since the points falls on the 45 degree reference line. For the Residuals vs. Fitted Plot, the points are randomly distributed, which indicates that the model is good and valid.



Result

Model with fixex Effects

$$pop = 84.33698 - 0.11671 \cdot nrgy - 0.05576 \cdot live + 0.04102 \cdot val - 0.43563 \cdot time.released$$

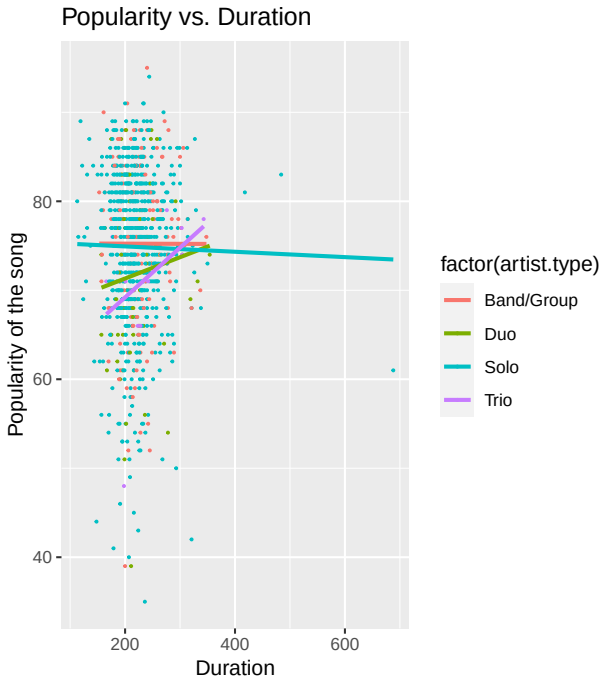
According to the multilevel model, we can get an idea about the factors that influence the popularity of a song. The `intercept` of this model is 84.33698 indicates that the popularity of a song is expected to be 84.33698 while the other variables are zero. The coefficient of `nrgy` is -0.11671, which means that the popularity of songs and energy are negatively related. Similarly, the coefficient of `live` is -0.05576, which demonstrates that the more likely a song is live recorded, the less popular it will be. For the variable `val`, the coefficient is 0.04102, which means that a more positive mood song would increase 0.04102 of popularity. The time of a song's release also shows a negative correlation with its popularity. When the release time of a song increases by one year, the popularity decreases by 0.43563.

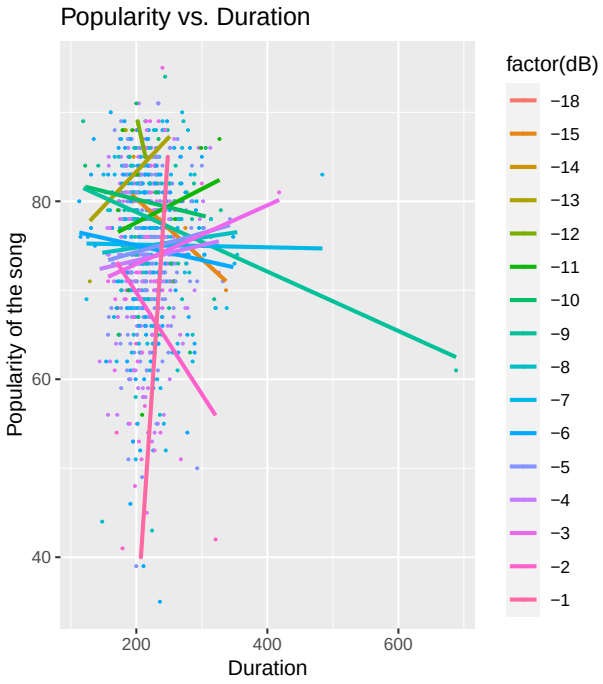
Discussion

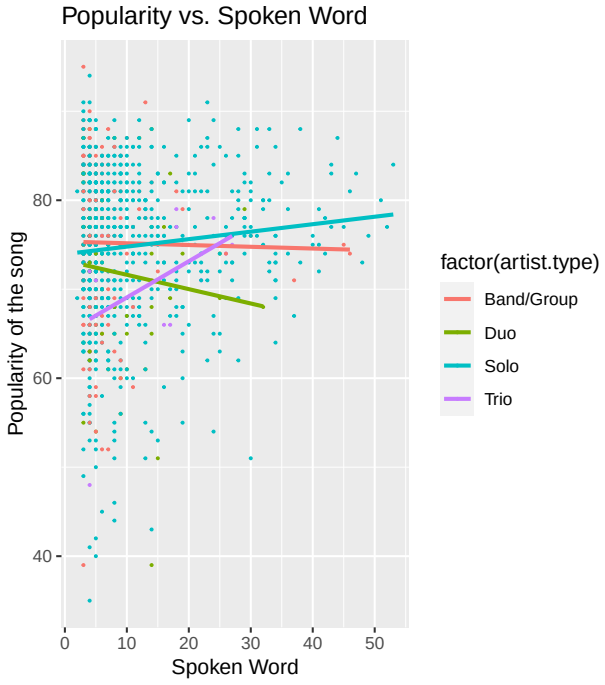
Based on our result, we know that a song with a more positive mood is more likely to have a higher popularity. Since popularity is a very fast-changing thing, we can find that the earlier the song is released, the less popular it is. And we can notice that people prefer to listen to less energetic music. When a song is less likely a live recording, its popularity rises.

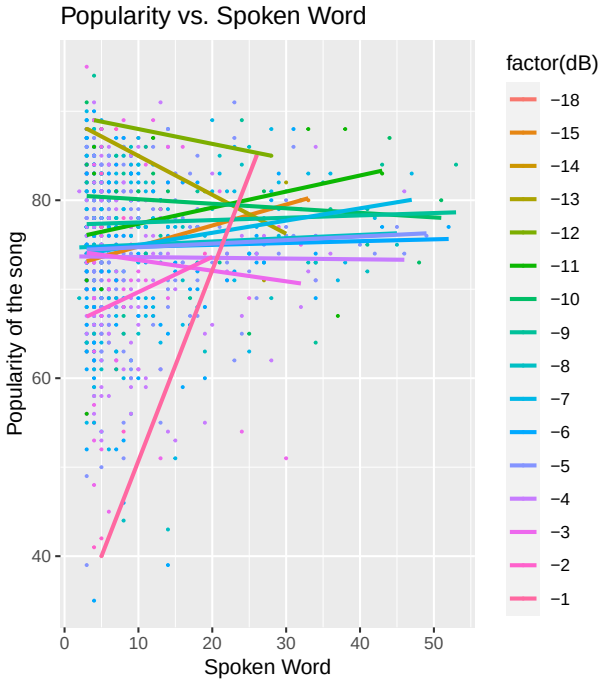
We choose 4 fixed effects and 2 random effects in this model. Although the model performs well in the model checking, there are still some limits of our analysis. Due to the rise of short videos in recent years, some songs topped the list probably because of its ability to serve as a very suitable BGM. in our analysis, we did not take into account the impact of the development of the Internet on the popularity of the songs.

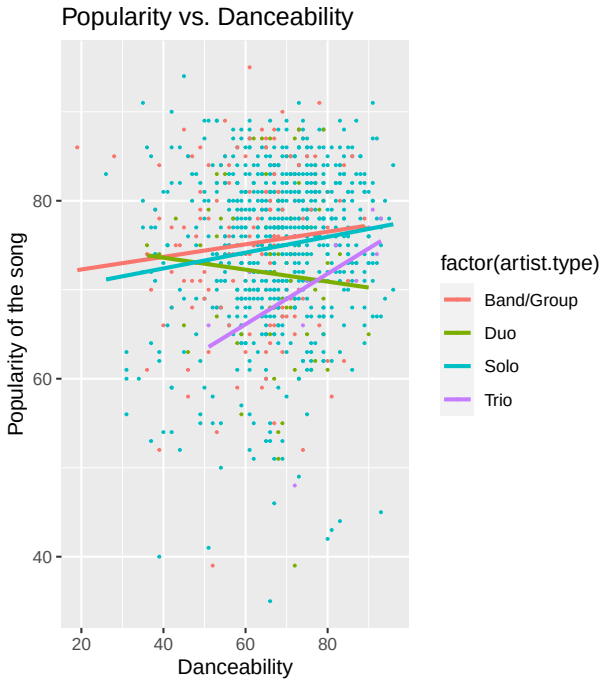
Appendix

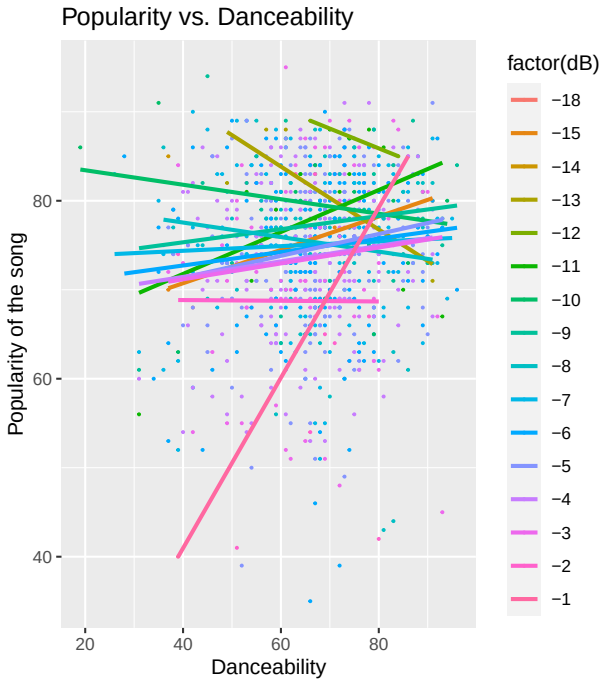

Popularity vs. Duration
Popularity of the song
Duration
80
60
40
200
400
600
factor(artist.type)
Band/Group
Duo
Solo
Trio


Popularity vs. Duration
Popularity of the song
Duration
80
60
40
200
400
600
factor(dB)
−18
−15
−14
−13
−12
−11
−10
−9
−8
−7
−6
−5
−4
−3
−2
−1


Popularity vs. Spoken Word
Popularity of the song
Spoken Word
80
60
40
0
10
20
30
40
50
factor(artist.type)
Band/Group
Duo
Solo
Trio


Popularity vs. Spoken Word
Popularity of the song
Spoken Word
80
60
40
0
10
20
30
40
50
factor(dB)
−18
−15
−14
−13
−12
−11
−10
−9
−8
−7
−6
−5
−4
−3
−2
−1


Popularity vs. Danceability
Popularity of the song
Danceability
80
60
40
20
40
60
80
factor(artist.type)
Band/Group
Duo
Solo
Trio


Popularity vs. Danceability
Popularity of the song
Danceability
80
60
40
20
40
60
80
factor(dB)
−18
−15
−14
−13
−12
−11
−10
−9
−8
−7
−6
−5
−4
−3
−2
−1