IDS 702 Team Black Project

Emma Wang, Pragya Raghuvanshi, Lorna Aine, Eric Rios Soderman

2022-10-21

EXPLORING POPULARITY AND EXPLICITY OF 2010s MUSIC THROUGH MUSICALITY.

I. Data Overview

The data set used in this research is a subset of a larger spotify dataset[link] that contained xxx number of tracks. It contains x number of observations/ tracks and x number of variables for songs between 2010 and 2019 thus the 2010s decade. Our focus in this research aims to infer which features of music can measure the popularity of songs using songs fro the 2010s decade and predict where a song is explicit or not based on these features. Our Research Questions are as under:

-What are the musical attributes that gauged the popularity of songs in the 2010s? *Dependent Variable (continuous) = popularity Independent Variables = acousticness, danceability, energy, instrumentation, tempo, loudness, and speechiness*

-To what extent can the musicality of a song predict whether a song will be explicit or non-explicit?

Dependent Variable (categorical) = Explicitness Independent Variables = danceability, energy, speechiness, and tempo In RQ2, we aim to "predict" whether the song is explicit or not by the four features of music: dancebility, energy, speechiness, tempo.

Project Proposal : Provide the chief characteristics of your data, including sample size, number of variables, and source. Include your research questions here.

Description of Dataset The original dataset contains 586672 rows, that is song tracks and 23 variable columns. Potential challenges faced with the dataset were: 1. Inconsistent data types for dates column which required cleaning of dates to read them better. 2. Some cleaning of artist names required missing values to be dropped, cleaning of brackets from the names to make them consistent for future use. 3. Conversion of duration variable from ms to minutes to facilitate easy interpretation. 4. Speechiness variable levels above 0.66 categorise as speech tracks like podcasts and poetries. Hence, they have been dropped. 5. Trash tempo values recorded as 0 had to be removed.

II.Primary relationship of interest

Table 1: Dimensions

586,672	23

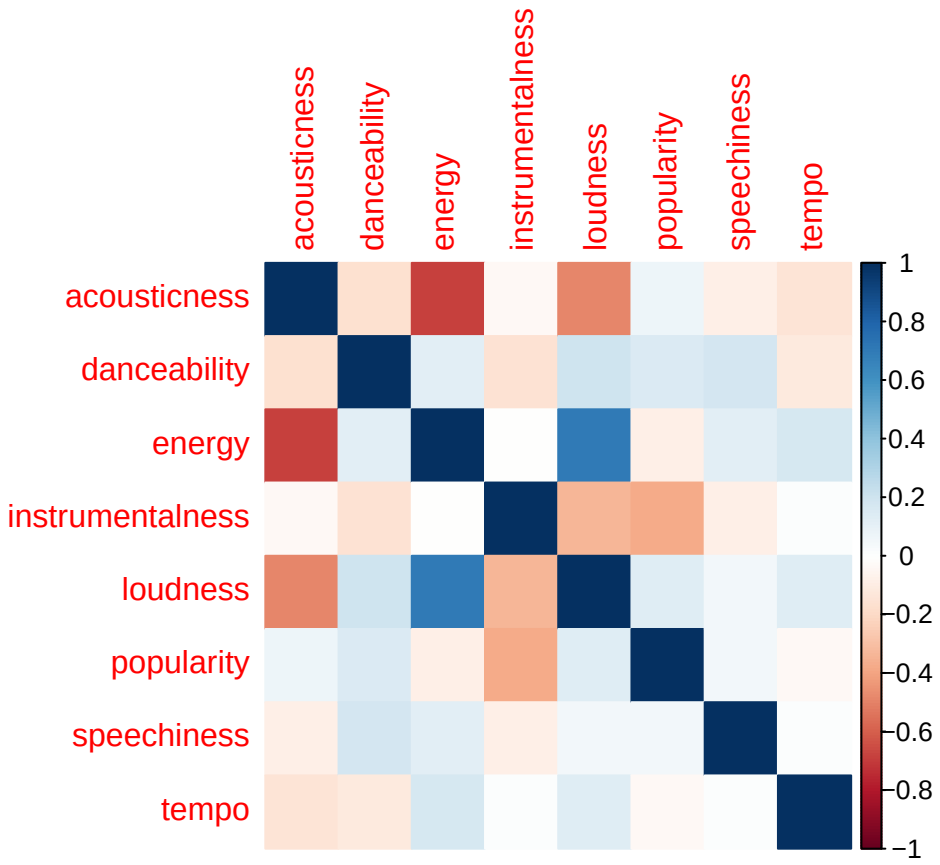
From the table below, we can observe the mean, median, SD, minimum and maximum values of variables separately for explicit and non explicit categories. While mean values of acousticness, instrumentalness, tempo are higher for non explicit songs, values for energy, danceability, loudness, speechiness are higher for

explicit songs. We can also infer that popularity of explicit songs is minutely higher than non explicit songs. High values of SD for tempo and popularity indicate that the data points are spread out in relation to the mean value. Low values of SD for danceability, energy, speechiness indicate that the data points are clustered around the mean. Nearly equal value of median and mean for danceability, energy , tempo indicate that the data points are more or less evenly distributed.

	Non-Explicit	Explicit	total
	(N=92440)	(N=12327)	(N=104767)
acousticness			
Mean (SD)	0.295 (0.300)	0.227 (0.224)	0.287 (0.293)
Median [Min, Max]	0.181 [0, 0.996]	0.151 [0.00000115, 0.993]	0.176 [0, 0.996]
danceability			
Mean (SD)	0.599 (0.154)	0.687 (0.147)	0.609 (0.156)
Median [Min, Max]	0.609 [0.0532, 0.988]	0.708 [0.0620, 0.986]	0.620 [0.0532, 0.988]
energy			
Mean (SD)	0.657 (0.224)	0.680 (0.162)	0.660 (0.217)
Median [Min, Max]	0.692 [0.0000203, 1.00]	0.685 [0.0000202, 1.00]	0.691 [0.0000202, 1.00]
instrumentalness			
Mean (SD)	0.0975 (0.254)	0.0212 (0.110)	0.0885 (0.243)
Median [Min, Max]	0.00000368 [0, 1.00]	0 [0, 0.989]	0.00000251 [0, 1.00]
tempo			
Mean (SD)	123 (28.2)	120 (29.4)	122 (28.3)
Median [Min, Max]	124 [31.3, 230]	120 [46.7, 220]	124 [31.3, 230]
loudness			
Mean (SD)	-7.32 (3.90)	-6.71 (2.56)	-7.25 (3.77)
Median [Min, Max]	-6.52 [-54.8, 1.93]	-6.40 [-29.0, 1.26]	-6.50 [-54.8, 1.93]
speechiness			
Mean (SD)	0.0752 (0.0754)	0.188 (0.130)	0.0885 (0.0912)
Median [Min, Max]	0.0461 [0.0220, 0.658]	0.159 [0.0229, 0.658]	0.0497 [0.0220, 0.658]
popularity			
Mean (SD)	38.1 (19.5)	47.7 (17.4)	39.2 (19.5)
Median [Min, Max]	42.0 [0, 92.0]	50.0 [0, 94.0]	42.0 [0, 94.0]

```
library(psych)

RQ1_relation <- c("popularity", "acousticness", "danceability", "energy", "instrumentalness", "tempo",
df1 = subset[RQ1_relation]
#sum(is.na(df1))
#pairs.panels(df1)
cordf1 = cor(df1)
corrplot(cordf1, method = 'color', order = 'alphabet')
```




```
RQ2_relation <- c("explicit_fac", "danceability", "energy", "speechiness","tempo")
df2 = subset[RQ2_relation]
#sum(is.na(df2))
#pairs.panels(df2)
```

Definitions of variables of interest

-**Popularity** is calculated by an algorithm that is based on how many times a track has been played and how recent those plays were. This is the response variable of interest for research question 1 (Spotify, 2022).

-**Explicitness** is whether a song contains inappropriate words such as curse words and sexually explicit words that are unacceptable to play in some public settings. 1 is the value identifying a song as explicit, while 0 implies that a song is non-explicit. This is the dependent variable for the second research question (Spotify, 2022).

-**Acousticness* is a confidence measure from 0.0 to 1.0 of how much of the track is composed with acoustic instruments. 1.0 represents high confidence the track is acoustic (Spotify, 2022).

-**Danceability** is a rating of a track's suitability for dancing. This metric is based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable (Spotify, 2022).

-**Energy* is a perceptual measure of intensity and activity. Energetic tracks typically feel fast, loud, and noisy (Spotify, 2022).

-**Instrumentation** pertains to whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumentals in this context, while Rap or spoken word tracks are considered "vocal". As the instrumentalness value approaches 1.0, there is a greater likelihood that the track contains no vocal content. In

addition, values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0 (Spotify, 2022).

-**Tempo** refers to the overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece, which derives directly from the average beat duration (Spotify, 2022).

-**Loudness** measures the overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological association of physical strength (amplitude). Lastly, the values typically range between -60 and 0 db (Spotify, 2022).

-**Speechiness** detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks (Spotify, 2022).

Analysis of variables

To better understand the data set, the popularity variable was binned into 5 groups ranging from least popular to more popular across group 1 to group 5.

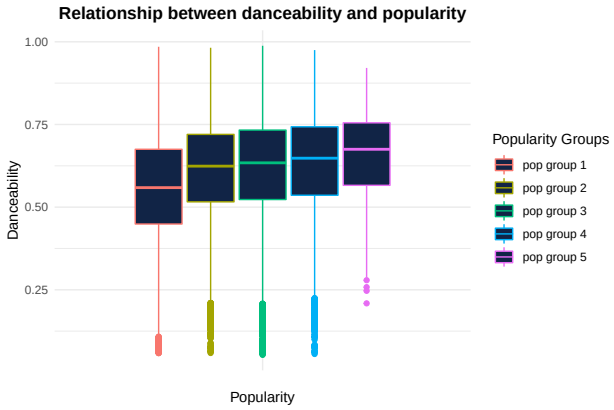


Fig 1

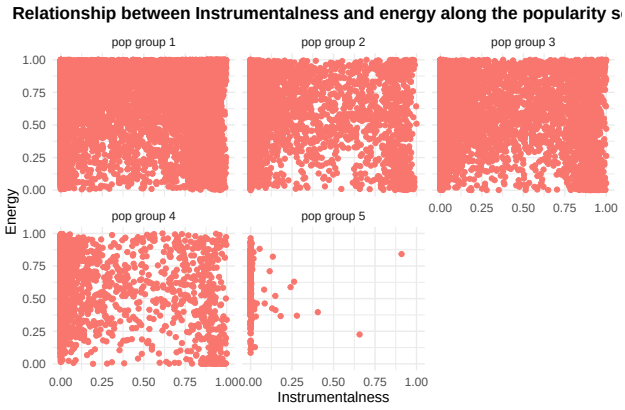


Fig 2

Fig 1: In this chart we see that the average danceability increases a long the popularity scale of songs. Fig 2: In this chart we see that as songs get more popular the energy remains evenly distributed but the Instrumentalness reduces with an exception of few outliers although the relationship becomes insignificant

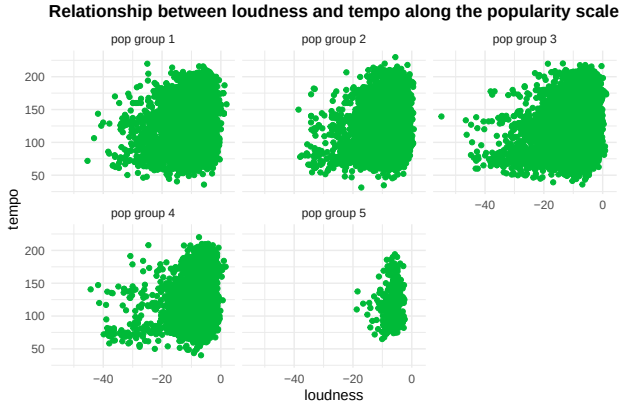


Fig 3

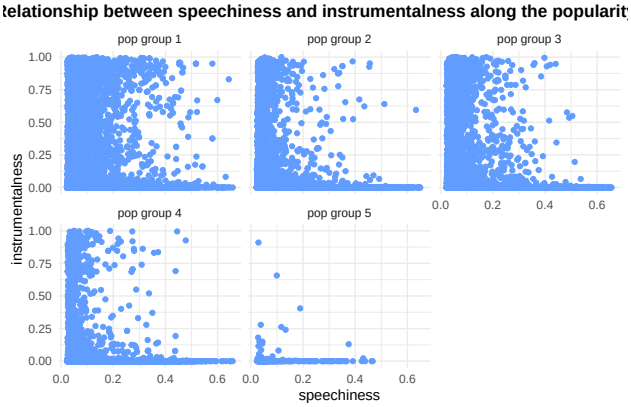


Fig 4

Fig 3: The tempo and loudness recipe for a popular song lies between 50-200 bpm for tempo units and -20 to 0 for loudness decibles. Fig 4: As songs gets popular the spechiness remains evenly distributed but the instrumentalness disappear From the above EDA, we see that variables like instrumentalness will ultimately become less signficant.

EDA Q2: Fig 5 : explicit content in music has risen over the years Fig 6: the energy in explicit songs is centered around the mean(calc) while non explicit songs are skewed to the higher energy(calc)

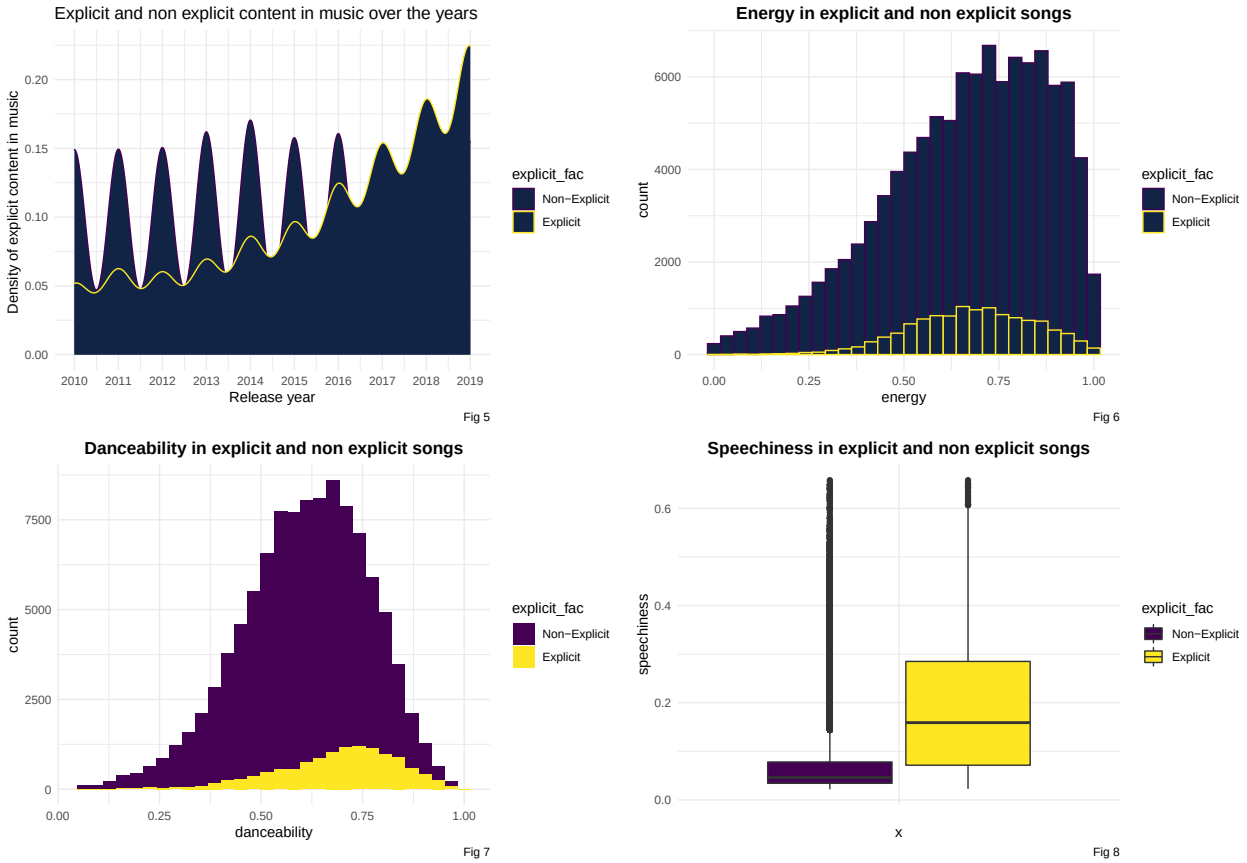


Fig 7: the danceability songs is centered around the mean(calc) Fig 8: The average speechiness in explicit songs is way less in explicit songs. In conclusion, the chosen variables would give us a great classification of explicit and non explicit songs moving forward.

Relationship to Research Questions

First and foremost, we choose the variables for our research questions based on prior, domain knowledge of musical terminology. For the first research question, which concerns the features that popularized songs during the 2010s, a series of aforementioned predictors were chosen. What will follow is the justification for this "a priori" selection for both research questions.

The reasoning for choosing these specific predictors (acousticness, danceability, energy, instrumentation, tempo, loudness, and speechiness) to predict song popularity is the weight of their importance. Popular songs, whether they are an emotional ballad or a dance track, all have certain features to keep the listeners engaged and interested to repeat listening to these tracks (Leviatan, 2017). The tempo, energy and loudness indicate the pacing and sonic impact and pleasantness of the track. The speechiness, danceability and instrumentation (which also includes acoustic choices or "acousticness") dictate melody choices, chord progressions, instrument choices, wordings, vocal lines and more types sonic layers. However, this last point is very nuanced because it pertains to the genre choice of the producers. There are very popular songs with high instrumentation, no words and low danceability, such as songs from classical music. On the other hand, Pop and Rock songs vary their levels of instrumentation and acousticness and speechiness to deliver

the best possible songs. Lastly, if the song is aimed towards a festive audience, such as a club song, then prioritizing danceability governs the levels of instrumentation and speechiness and lack of acousticness, and this prioritization varies by genre (Androids, 2017). In conclusion, the interplay of these factors influence the popularity of songs by helping them become easily memorable and enticing.

As for the second research question, the explicitness of tracks is strongly swayed by other factors. A very logical approach to predicting explicitness was first looking at the high levels of speechiness in songs. For example, rap songs rank high in this metric because the verses are composed of a spoken word format over a series of 8 or 16 bars, and each bar is a rap line (Edwords), while singing doesn't have to adhere to the "1 bar = 1 line" rule; thus, speechiness became the metric of most importance. In addition, songs in this genre tend to include explicit content, often sexual, in the lyrics (Tayag, 2017). Second to this metric, the other predictors of danceability, energy and tempo were considered as helpful in predicing explicitness. The energy and danceability of the song collude with speechiness to infer if a track could have explicit language. For example, a song with low energy and low danceability may or may not be less likely to have explicit language than a song with high energy and danceability, holding the speechiness level constant, and this is a relationship we wish to investigate as well. As for tempo, music genres that are known to include explicit language follow specific tempos. For instance, Trap songs usually have a tempo of 140 bpm (Burchell, 2019).

In the research, we have two main Research Questions: RQ1. What are the musical attributes that gauged the popularity of songs in the 2010s? *Dependent Variable (continuous) = popularity Independent Variables = acousticness, danceability, energy, instrumentation, tempo, loudness, and speechiness* In RQ1, we aim to "infer" what features of music could better measure the popularity.

RQ2. To what extent can the musicality of a song predict whether a song will be explicit or non-explicit?

Dependent Variable (categorical) = Explicitness Independent Variables = danceability, energy, speechiness, and tempo In RQ2, we aim to "predict" whether the song is explicit or not by the four features of music: dancebility, energy, speechiness, tempo.

III.Other characteristics

Briefly describe other variables in the data. If there are many, do not list them all. Rather, describe the types of variables that are present (e.g., "demographic information")

A few variables such as key and time signature are part of this dataset, although they were not chosen as the predictors. Most of the remaining variables in this dataset pertain to the artist name, the song title, the modalities and key of a song, the duration and the release dates. Nonetheless, some variables were still very rich in information in terms of prediction. To illustrate, one could predict the scale of a song based on the popularity in addition to other predictors.

Some were not chosen because they were differently coded, such as time signature, which has an extremely limited set of plausible values (lacks signatures like 6/8). In contrast, some offered little or irrelevant information. Liveness is one such case. It parametrizes a song's performance as a live or studio quality recording, and, given that the songs that play on the radio tend to be studio songs, we opted to not use this variable as a predictor.

IV. Potential Challenges

Describe aspects of the data that may present challenges in the modeling stage. For example, might certain categorical variables need to be collapsed? Is there a lot of missingness? Could the size of the dataset present model selection challenges?

Eric's findings from cleaning : Yes, there is a bit of missingness. Some dates have years, but not months and days recorded. This meant that we had to manually fix them or give them arbitrary values. Some artist names are missing, which is notable but currently not obstructive. We also do not know if some of the tempos of the higher songs were incorrectly recorded.

Modifications based on the research questions:

1. Our research question pertains to the songs of the 2010s decade. Hence, post subsetting the original dataset to our research interest, the dataset contains 104767 rows and 23 columns.
2. Bad tempo data in the subset for around 148 tracks have been dropped. The column names depicting are variables of interest and the rows depicting the number of observations.
3. For our first research question we have popularity as dependent variable which is continuous variable. Our second research questions predicts Explicitness(dependent variable) on basis of predictor variables. Explicit here is a categorical variable and hence, we will factor it into non explicit, and explicit.

V. Bibliography (Citations)

Androids (2017, October 13). An Idiot's Guide to EDM Genres. Retrieved October 20, 2022, from https://www.complex.com/music/an-idiots-guide-to-edm-genres/

Burchell, C. (2019, May 27). 10 Tips for Making Your First Trap Beat. Inverse. Retrieved October 20, 2022, from https://flypaper.soundfly.com/produce/10-tips-for-making-your-first-trap-beat/#

Edwords, E. (n.d.). Rap Song Structure Is TOO Important To Ignore. Retrieved October 20, 2022, from https://rhymemakers.com/rap-song-structure/

Leviatan, Y. (2017, July 27). Making Music: The 6 Stages of Music Production. Waves. Retrieved October 20, 2022, from https://www.waves.com/six-stages-of-music-production

Spotify (2022). Spotify Web API Reference | Spotify for Developers. https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features

Tayag, Y. (2017, May 17). Expert on Male Psychology Explains How Pop Got Sexually Explicit. Retrieved October 20, 2022, from https://www.inverse.com/article/31842-pop-music-sexually-explicit-lyrics-rap-hip-hop