NANDO N' BASS- Spotify Tracks Dataset Analysis

João Tavares Fernando 07/2024

This brief written report serves as an accompanying analysis for the insights gathered and presented on the PowerPoint presentation, delving further on the sources, methodology and findings.

NANDO N' BASS is a music production studio that wants to better understand how to work inside the studio during the recording sessions, but also to know if they should be working with artists in specific genres. They acquired a dataset consisting of various different tracks and their respective characteristics/variables, extracted from Spotify using the Spotify Web API on October 2022. A link to the dataset, as well as a detailed description of the different variables, can be found in the annex section, at the end of the written report.

The data file was loaded into MySQL Workbench, where most of the exploratory analysis was conducted. PowerBI was used for visual analysis and presentations, as well as creating some of the data tables shared. The correlation matrix was generated using Pandas on a JupyterNotebook. All the relevant files and scripts will be shared with both the presentation and the written analysis.

The cleaning process was fairly straightforward, with no Null values present in the dataset, and only one Blank value. Upon further inspection, the Blank value referred to an entry with no artist, album or track name. After searching Spotify using the track ID, the track was shown as "unavailable", as such, this entry was removed.

After cleaning, we were left with 113999 entries, containing 89740 unique tracks over 114 different genders.

For easier analytical work, popularity was divided into ten different intervals, each spaced ten popularity points apart. As an example, Popularity Rank 10 encompasses all songs with popularity greater than or equal to 90.

Looking at the distribution of tracks across popularity ranks, we see that less than 1% of unique tracks has popularity greater than 80, and as such, we should be cautious when extrapolating our findings to more tracks.

The most and least popular musical genres are also in line with what one would presume, though some comments are needed. "Pop-film" is mostly comprised of Bollywood songs, which are relevant only to a very specific, yet numerous, type of listener. Genres like "British" also pose a potential problem with our dataset. Each song can be allocated to multiple genres, an area which is in itself highly subjective. As an example, "The Beatles" tracks can be categorized as "Rock, Psych Rock or British", while tracks by "Iron Maiden", which is also a band from England, are categorized as "Metal or Hard Rock".

The multiple characteristics/variables related to a track where divided into numeric and non-numeric data, based on whether they represented numeric quantitative values or categorical data, and where evaluated using different approaches.

A song is classified as Explicit if its lyrics contain sensitive or adult topics. While a superficial view shows that being explicit has a slight positive impact on popularity, with 20% of explicit tracks having Popularity Rank 7 or higher, compared to 12% of non explicit tracks, we also need to be aware that less than 9% of all tracks are considered explicit. We can't really take this information at face value, seeing as certain genres are more prone to being explicit than others as well as the impact that being an instrumental track has on this variable. Besides these analytical conclusions there are also external factors that influence the popularity of an explicit song, like it being played on the radio or being age restricted.

Looking at the other non numeric variables key, mode and time signature, we need to be aware of the technical depth and meaning these have in the context of music theory and song construction, and as such, provide more aerial insight, leaving these details to the musical experts at the studio. Nonetheless, we find that there is a very big over representation of major scales, in particular those in the key of Sol, Re and Do; as well as "normal" time signatures (those in 4/4). These findings reflect common musical knowledge. Times outside of 4/4 tend to have an arrhythmic feeling to them, and the popular keys, modes and time signatures tend to have a more familiar sound to them. A slight change can make a song sound unnatural or exotic. Listeners can be very sensitive to different sounding works and tend to be wary of exploratory sounding music, or music that is not like those they are accustomed to. Though these are not immediately reflected on the popularity of the tracks and these variables should be taken into consideration with other characteristics of tracks themselves, as well as other external factors.

Looking at the numeric variables and grouping by popularity, we can see that certain combinations stand out. Popular songs tend to be slightly shorter, as well as more danceable and energetic. This makes it so they are easier to remember and listen on repeat, as well as provide them with more listening opportunities outside of active music consumption, like in sport events, or background music in stores and malls. Popular songs also tend to be louder, less acoustic and not instrumental, making them all around "catchier" and memorable. Lastly, though there is little variation across the ranks, higher popularity songs tend to have average valence (positivity).

When trying the same exercise across musical genres instead of popularity rank, we find that the same outcomes are not reached. When grouping tracks through genres you get a more defined and strict set of characteristics, as for example, most metal songs will not sound like reggae music. Taking for example, emo music and latin music. The two contrasting genres have variable values close to each other, but end up having a difference of almost 40 popularity points. As it stands, the dataset is not prepared to provide workable insights regarding the numeric variable's impact on the music genres themselves.

A correlation matrix was created to better understand how the different variables are related with each other and how they impact popularity. Regarding popularity, we can see that no variable has a very high influence, positive or negative. With this in mind, we can still divide the variables in those which have a positive impact (Explicit, Danceability, Loudness, Tempo) and negative impact (Speechiness, Acousticness, Instrumentalness, Valence). Besides the focus on how popularity itself is affected, we can also see that some variables have high correlations between themselves. There is a positive correlation between explicit and speechiness, which makes sense, as songs without speech or lyrics, are unable to be explicit. Danceabilty and Energy are both highly positively correlated with loudness and valence, however, Danceability and Energy are not highly correlated with each other. This makes sense, seeing as being energetic does not make a song danceable. For example, metal music is highly energetic but not exactly

danceable. Acousticness has a high negative impact on both energy and loudness. Live songs also have a positive correlation with energy and speechiness, due to the noise and interaction with the live audience.

While all variables are deserving of an in depth analysis by themselves, Speechiness stands out in the sense that it has a relevant negative correlation with popularity but the results are very homogeneous across both popularity ranks as well as genres, with values rarely going over 0.15. In fact, over 96% of tracks have speechiness under 0.33, meaning they classify as "music" as opposed to spoken word or something like a poem. Looking at a scatter plot of music gender over popularity and speechiness, we can see that the genres with higher speechiness are "Comedy, Kids and Children". Seeing as this variable is this skewed, it would likely benefit from some mathematical normalization.

Having gone over our data set, we can recommend the following:

- No need to shy away from explicit songs. But no need to push for them.
- Regarding musical structure, there is little incentive to deviate from the norm, but success can also be found in niche communities.
- Energy has a positive impact but is not equal to danceability (metal VS house).
- Higher danceability, and loudness should lead to a higher popularity.
- Higher acousticness and instrumental should lead to a lower popularity.

While this analysis can be seen as simplistic over such a rich and complex topic, with external factors not being taken into account, as well as the fact that the non numeric variables are deeply related to complex music theory (which in itself is heavily influenced by geographical and historical context), we were still able to provide actionable and direct insights for the music producers. Moving forward, further action can be taken in the following areas:

- Expand the dataset:
    - None of the variables has a very high correlation value with popularity, we can expand the amount of tracks we have, or even our timeframe.
    - We may be missing relevant variables, like release year, which could add to this analysis.
- Delve deeper into the different variables:
    - Valence having a negative correlation is curious. After looking at the genres, "Party" and "Kids" show very high valence values, but very low popularity, skewing this results. In this case, Natural Language Processing and Sentiment Analysis could be good approaches.
- Advanced predictive models:
    - By doing a more thorough statistical approach, like looking at the distributions of the different variables or dimensionality reduction, we could get some deeper insights with the data own.
- Redefining musical genres:
    - At a superficial level, the current data set does not allow for a deep analysis focused on musical genres. By clustering genres we could shine a new light on this topic.

Annex

Source:

Spotify Tracks Dataset

https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset/data

Description of variables:

track_id: The Spotify ID for the track

artists: The artists' names who performed the track. If there is more than one artist, they are separated by a ;

album_name: The album name in which the track appears

track_name: Name of the track

popularity: The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Duplicate tracks (e.g. the same track from a single and an album) are rated independently. Artist and album popularity is derived mathematically from track popularity.

duration_ms: The track length in milliseconds

explicit: Whether or not the track has explicit lyrics (true = yes it does; false = no it does not OR unknown)

danceability: Danceability describes how suitable a track is for dancing 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

energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale

key: The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1

loudness: The overall loudness of a track in decibels (dB)

mode: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0

speechiness: 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

acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic

instrumentalness: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content

liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live

valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry)

tempo: 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 and derives directly from the average beat duration

time_signature: An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of 3/4, to 7/4.

track_genre: The genre in which the track belongs