CENTRO UNIVERSITÁRIO DE BRASÍLIA

DISCIPLINA: Introdução a R para Ciência de Dados

PROFESSOR: Dr. Wandré Nunes

ALUNO: Felipe Martins Machado Mendes
RA: 22251506

Trabalho Final
EDA do dataframe "Spotify and Youtube"

BRASÍLIA

JUNHO, 2023

In this work, the "Spotify and Youtube" database will be analyzed and explored, through the R language, using the contents learned in class during this first half of 2023.

This is a database that is available on the Kaggle platform (link: https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube) and is authored by users Salvatore Rastelli, Marco Guarisco and Marco Sallustio .

It is worth mentioning that it was updated in February 2023, and as these platforms are very active and the ranking of songs is volatile, the data presented here may not be consistent with the current scenario of platforms.

Specifications of the database

There are exactly **20718 rows** e **21 columns,** the details upon then are:

Track: song name, as seen on the Spotify platform.

Artist: name of the artist.

Url_spotify: artist's URL on Spotify.

Album: the album the song is contained in on Spotify.

Album_type: indicates if the song is relesead on Spotify as a single or contained in an album.

Uri: a spotify link used to find the song through the API.

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: is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.

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). Values typically range between -60 and 0 db.

Speechiness: detects the presence of spoken words in a track.

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..

Liveness: detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed 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, while tracks with low valence sound more negative..

Tempo: the overall estimated tempo of a track in beats per minute (BPM)..

Duration_ms: the duration of the track in milliseconds.

Stream: number of streams of the song on Spotify.

Url_youtube: url of the video linked to the song on Youtube, if it have any.

Title: title of the videoclip on youtube.

Channel: name of the channel that have published the video.

Views: number of views.

Likes: number of likes.
Comments: number of comments.
Description: description of the video on Youtube.
Licensed: Indicates whether the video represents licensed content.
official_video: boolean value that indicates if the video found is the official video of the song.

From all of those features, the most interesting ones are the ones that use AI to generate metrics for the distinc characteristics of the songs. For example, we have **Danceability**, which shows us how danceable a song is. But the question we must ask is if that AI is sharp or not?

Before we begin our analysis, it's important to say that all the graphics created during this project are at the **end of the document.**

Analysing the top 10 songs by streams and views:

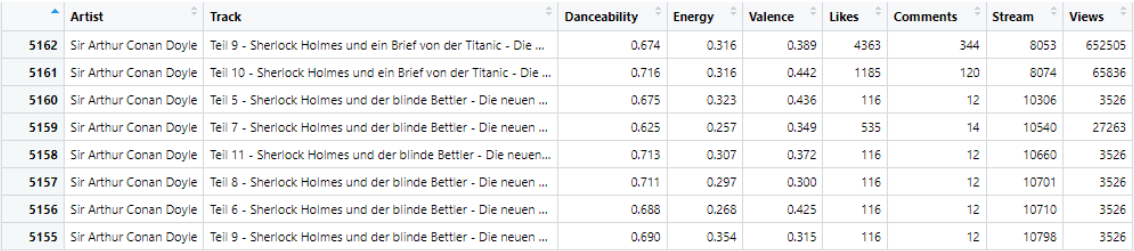
In the figures that are at the end of the document, we can see that the songs in the top of spotify are different from the ones on youtube and that the views are on a larger scale than the streams. I believe that it's because of two things: **the way spotify and youtube register a person listening to the song** and because of the **collaborations**.

Of the top 10 songs by number of streams, 4 are collaborations between two artists, and in the top 10 songs by number of views, 2 also are collaborations. In that case, I think that when you combine two influential singers and the song is good, the result will be a song which will be accessed by both of their fan bases and the stream number will skyrocket.

Youtube counts a view every time a person loads the page of the video of the song. Spotify counts a stream every time a person listens to 30 seconds of a song. In this way, the quantity of views will be a lot higher than the number of streams, just check the number of views of Despacito (8 billion - the highest one of youtube) and the number of streams (3 billion - the highest one of spotify).

Analysing the metrics of the AI

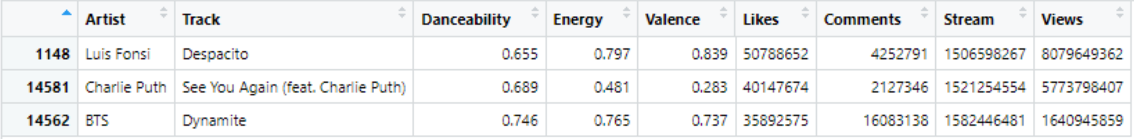
Well, the feelings that a song brings to a person are very subjective. Asking an AI to give a score about how much energy a music brings, or how danceable it's, is an interesting thing to do, but can an AI be up to it?
In some cases, yes, but in another, the AI just messes things up pretty badly. Take for example, there is a register of an audiobook in the data. It's just a guy reading a story of Sherlock Holmes in german. The AI got there and gave it 0.713 of danceability, which is quite a high score.



	Artist	Track	Danceability	Energy	Valence	Likes	Comments	Stream	Views
5162	Sir Arthur Conan Doyle	Teil 9 - Sherlock Holmes und ein Brief von der Titanic - Die ...	0.674	0.316	0.389	4363	344	8053	652505
5161	Sir Arthur Conan Doyle	Teil 10 - Sherlock Holmes und ein Brief von der Titanic - Die ...	0.716	0.316	0.442	1185	120	8074	65836
5160	Sir Arthur Conan Doyle	Teil 5 - Sherlock Holmes und der blinde Bettler - Die neuen ...	0.675	0.323	0.436	116	12	10306	3526
5159	Sir Arthur Conan Doyle	Teil 7 - Sherlock Holmes und der blinde Bettler - Die neuen ...	0.625	0.257	0.349	535	14	10540	27263
5158	Sir Arthur Conan Doyle	Teil 11 - Sherlock Holmes und der blinde Bettler - Die neuen...	0.713	0.307	0.372	116	12	10660	3526
5157	Sir Arthur Conan Doyle	Teil 8 - Sherlock Holmes und der blinde Bettler - Die neuen ...	0.711	0.297	0.300	116	12	10701	3526
5156	Sir Arthur Conan Doyle	Teil 6 - Sherlock Holmes und der blinde Bettler - Die neuen ...	0.688	0.268	0.425	116	12	10710	3526
5155	Sir Arthur Conan Doyle	Teil 9 - Sherlock Holmes und der blinde Bettler - Die neuen ...	0.690	0.354	0.315	116	12	10798	3526

But this AI mistake seems to be only in the Danceability feature. I saw other metrics like Valence in other songs that I totally agree with.

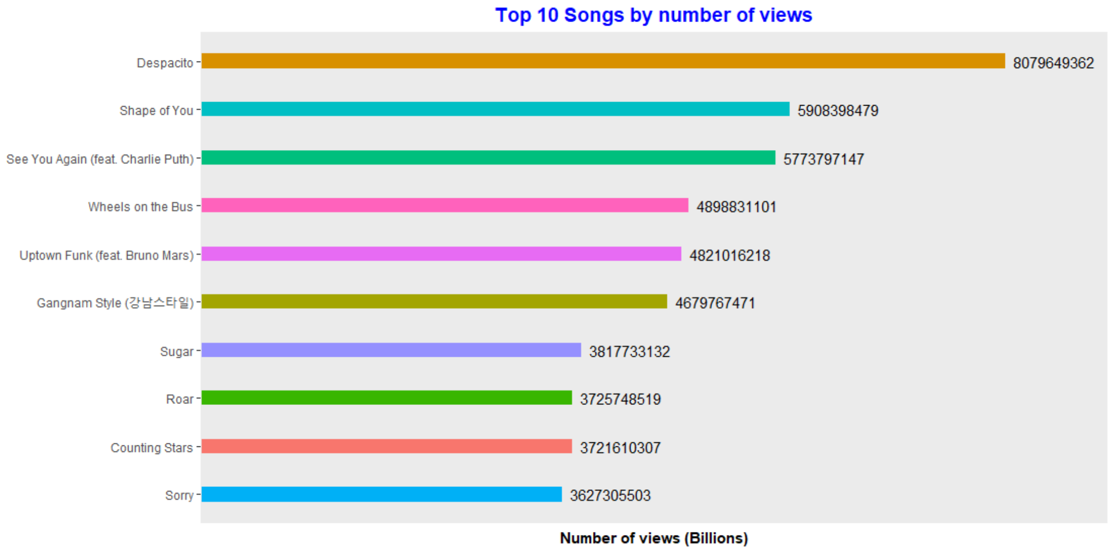
And in some of the graphs that I made, we can see a clear and logic relation between some of the variables, like the energy and loudness one, so I'll give a point for the AI for that.

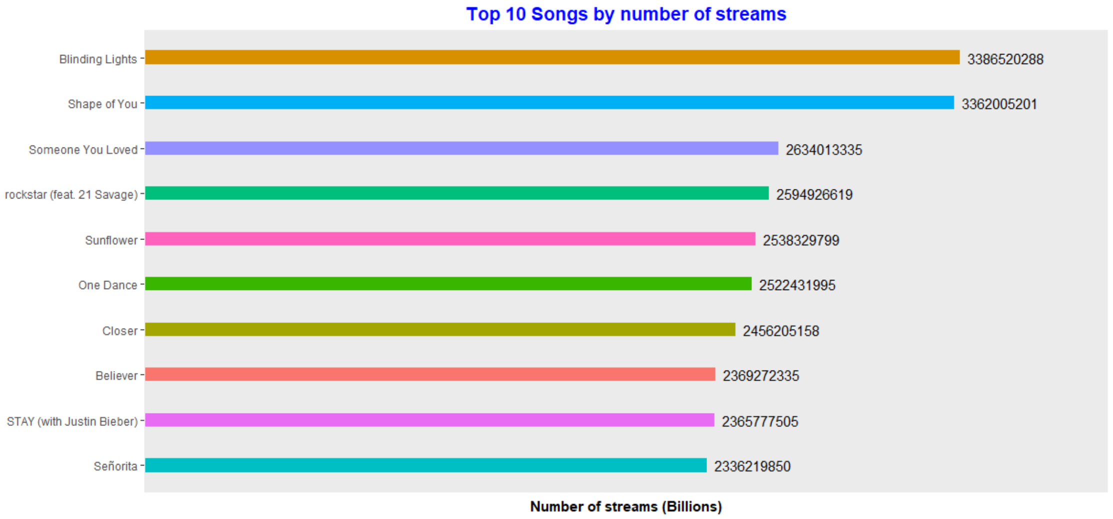


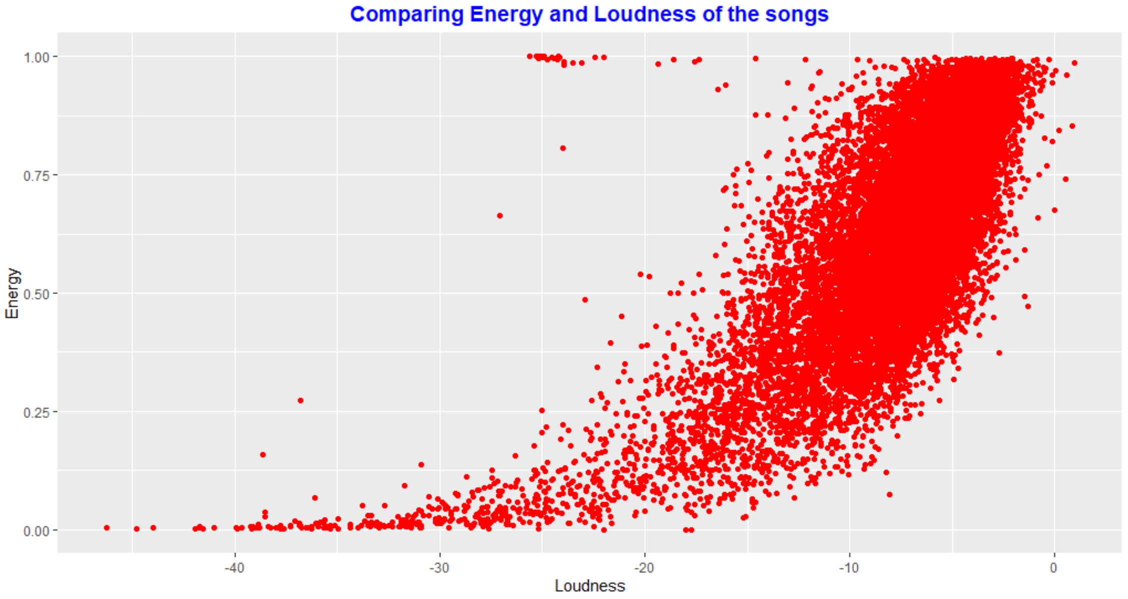
	Artist	Track	Danceability	Energy	Valence	Likes	Comments	Stream	Views
1148	Luis Fonsi	Despacito	0.655	0.797	0.839	50788652	4252791	1506598267	8079649362
14581	Charlie Puth	See You Again (feat. Charlie Puth)	0.689	0.481	0.283	40147674	2127346	1521254554	5773798407
14562	BTS	Dynamite	0.746	0.765	0.737	35892575	16083138	1582446481	1640945859

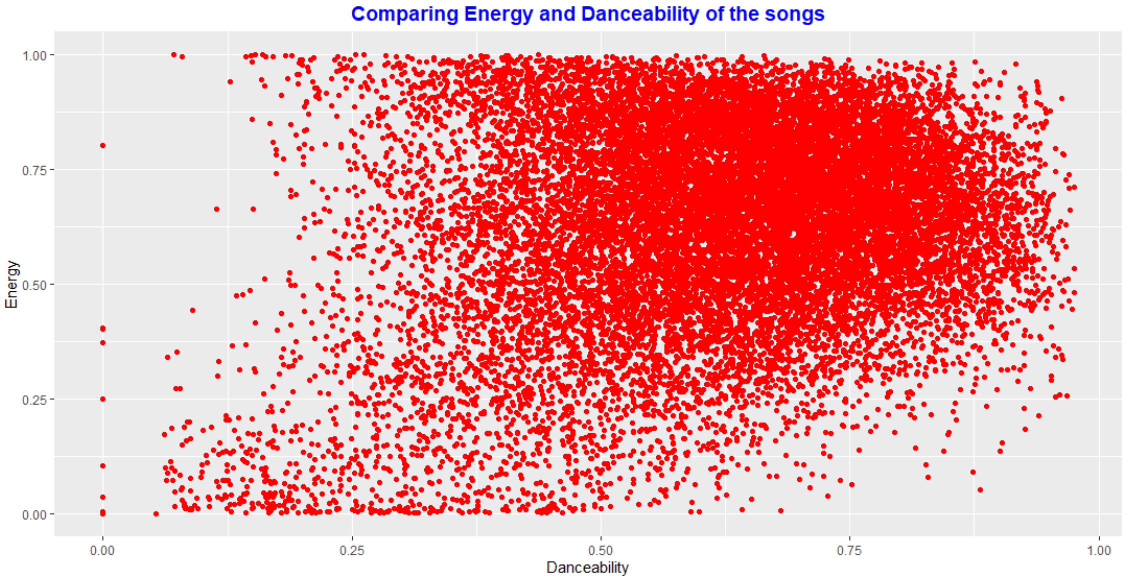
For example, the music See You Again, it's a sad song about a sad event. It has a low Valence and a low Energy, but quite a high score for Danceability.

Graphics:


Top 10 Songs by number of views
Despacito
8079649362
Shape of You
5908398479
See You Again (feat. Charlie Puth)
5773797147
Wheels on the Bus
4898831101
Uptown Funk (feat. Bruno Mars)
4821016218
Gangnam Style (강남스타일)
4679767471
Sugar
3817733132
Roar
3725748519
Counting Stars
3721610307
Sorry
3627305503
Number of views (Billions)


Top 10 Songs by number of streams
Blinding Lights
3386520288
Shape of You
3362005201
Someone You Loved
2634013335
rockstar (feat. 21 Savage)
2594926619
Sunflower
2538329799
One Dance
2522431995
Closer
2456205158
Believer
2369272335
STAY (with Justin Bieber)
2365777505
Señorita
2336219850
Number of streams (Billions)


Comparing Energy and Loudness of the songs
Energy
1.00
0.75
0.50
0.25
0.00
-40
-30
-20
-10
0
Loudness


Comparing Energy and Danceability of the songs
Energy
1.00
0.75
0.50
0.25
0.00
Danceability
0.00
0.25
0.50
0.75
1.00