

[Questa foto](https://en.wikipedia.org/wiki/Billboard_Year-End_Hot_100_singles_of_1990) è concesso in licenza da [CC BY-SA](https://creativecommons.org/licenses/by-sa/3.0/)

Music Data Analysis in R

Billboard chart songs: 2010 - 2019

Nicola Pasolini | Computer Engineering

Introduction

The present analysis aims to examine in detail a dataset that reports the popularity characteristics of musical pieces published between 2010 and 2019.

The dataset used was obtained from the Kaggle platform and is accessible at the following link: [Top Spotify Songs from 2010-2019](https://www.kaggle.com/datasets/leonardopena/top-spotify-songs-from-20102019-by-year?select=top10s.csv) .

# Starting Data

The dataset is composed of several variables that provide detailed information on each song, allowing an in-depth analysis of popularity dynamics in the musical context. The key variables present in the dataset are listed below along with their meaning:

1. title: The title of the song.
2. artist: The artist responsible for the song.
3. top genre: The genre of the song.
4. year: The year the song was on the Billboard charts.
5. bpm (Beats per minute): The number of beats per minute, representing the tempo of the song.
6. nrgy (Energy): Energy levels of the song, where higher values indicate greater energy (speed, volume).
7. dnce (Danceability): The ease with which you can dance to the song, with higher values indicating greater adaptability to dancing.
8. dB (Decibel): The sound volume of the song.
9. live (Liveness): The probability that the song was recorded with a live audience.
10. val (Valence): Levels of valence, with higher values indicating a more positive tone (happy, cheerful).
11. dur (Duration): The duration of the song.
12. acous: The probability that the song is acoustic.
13. spch (Spoken): Levels of speech in the song, with higher values indicating a greater presence of spoken words.
14. pop (Popularity): The popularity of the song, with higher values indicating greater popularity.

# Objective of the Analysis

The main objective of this analysis is to understand the relationships between the musical characteristics described in the dataset and the popularity of each song.

Through the use of analytical techniques in R, I intend to explore emerging patterns and find meaningful conclusions about the evolution of musical preferences over the decade under review. The wide range of information provided by the variables will allow me to deepen my understanding of the factors that contribute to the success of a song in the contemporary music scene.

Conduct of the Analysis

# Data Preparation and Cleaning

In the initial phase of the analysis process I dedicated myself to data preparation and cleaning to ensure consistency and reliability throughout the entire analysis.

The first activity carried out corresponds to the removal of songs with zero popularity.

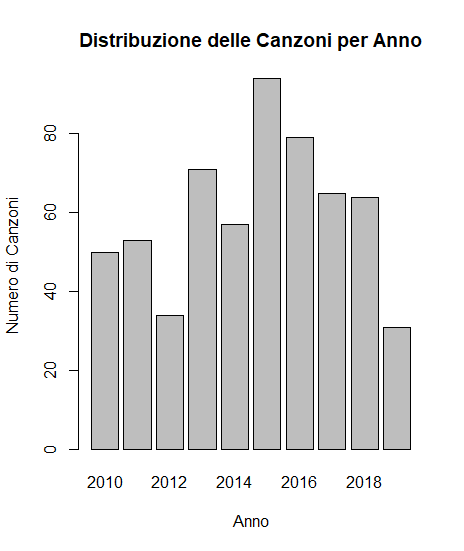
music <- subset(music, pop != 0)

I chose to exclude songs with zero popularity to ensure data integrity by avoiding negatively influencing the analysis with unrepresentative data.

Subsequently I moved on to the extraction of the Top 31 Songs by Year.

To focus on the most relevant tracks from each year, I pulled the top 31 songs based on popularity.

The number was determined considering that the number of songs available is not uniform for all years, and 31 represents the lowest common denominator.



The goal of these operations was to create a more focused subset of data, composed of the most popular tracks from each year, to analyze them in more detail. Removing songs with zero popularity helped maintain a meaningful dataset, while extracting the top 31 songs by year allowed me to focus on the most influential tracks over time.

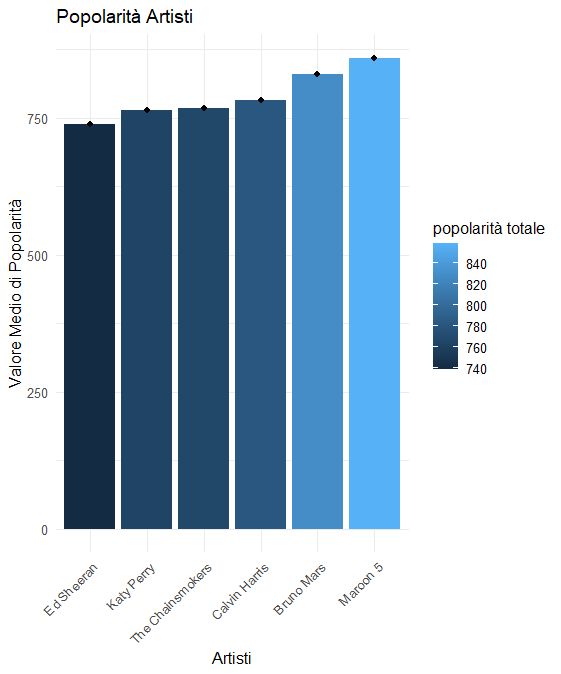
# Desciptive Analysis

# To fully understand the musical landscape of the decade, I focused the analysis on determining the **best artists and genres**, basing the evaluation on the sum of the popularity of the songs per category.

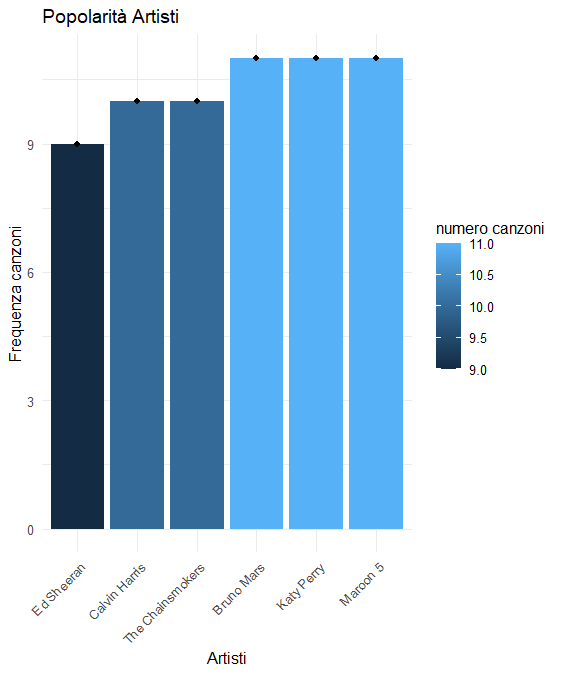
Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamente

To identify the top artists, I calculated the **sum of the popularity** of their songs in the dataset. Below, I present the results with a bar chart highlighting the top contributors to popularity over the decade:

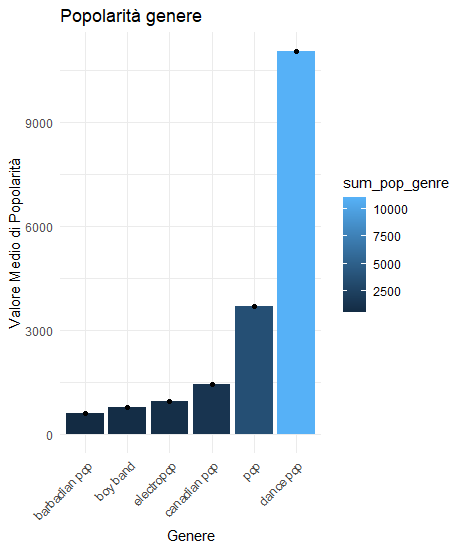


Let's now evaluate whether popularity is **consistent with the number of songs** present in the chart by the artists:



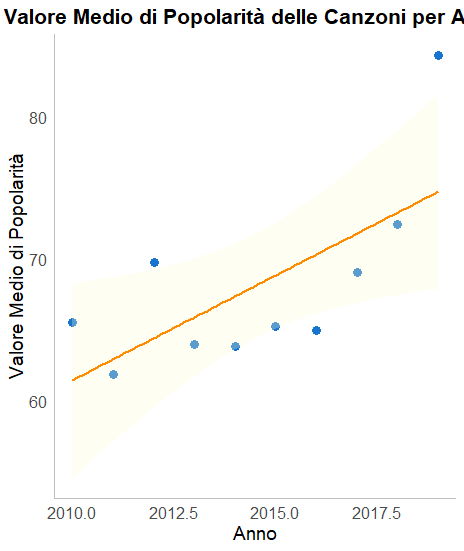
It emerges that the most listened to artist is Maroon 5, but the ranking is still very homogeneous.

Similarly, I performed a genre analysis, calculating the sum of the popularity of songs belonging to each genre. Below, I present the results of the distribution of popularity between different musical genres:

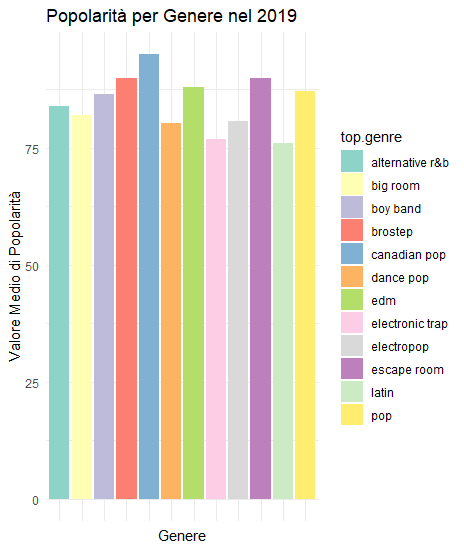


However, the situation is different for the most listened to musical genres, where as we might have expected Pop dominates the other genres, in particular in the dance variant.

Then follows the representation of the **trend in popularity** of the songs in the period considered:



You can notice a positive increase in interest in music with an interesting peak in 2019, which is why I looked into the musical genres relating to the single year mentioned:



Most listened songs in 2019 were:

Immagine che contiene testo, schermata, Policromia, diagramma

Descrizione generata automaticamente

The insight into 2019 in the context of the decade provides a more complete view of popularity dynamics and emerging musical preferences. This approach allows us to grasp the details of a particularly significant period in the musical panorama.

**Now we will return to considering the initial dataset, containing all the songs.**

# Song duration and bpm analysis

First I create macro categories in musical genres:

Immagine che contiene schermata, testo, Carattere

Descrizione generata automaticamente

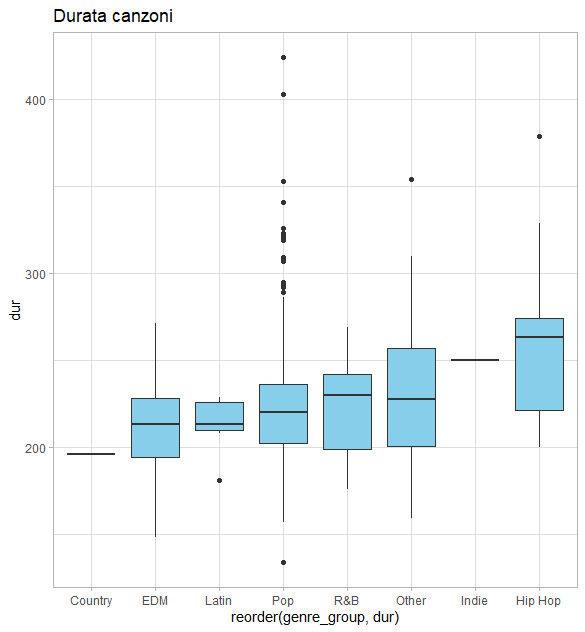
Immagine che contiene testo, Carattere, schermata, linea

Descrizione generata automaticamente

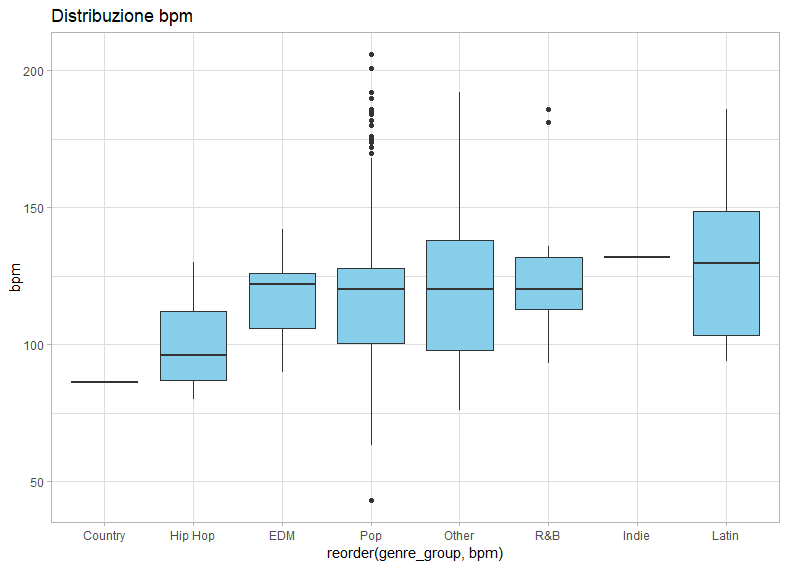
Most of the song has a duration of 3-4 mins.

There is an outlier in pop category, we have song with most/min duration.

The longest song (424s) is by Justin Timberlake “TKO” and the shortest song (134s) is by Justin Bieber “Mark My Words”.



I repeat the analysis for bpm:



Analyzing the bpm, most songs have bpm around 100-150 bpm, which some few songs from pop genre has bpm >200bpm. Those are: How Ya Doin? by Little Mix and FourFiveSeconds by Rihanna.

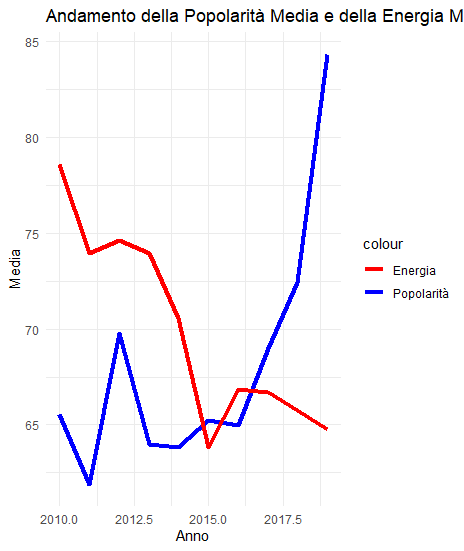
Immagine che contiene testo, schermata, Carattere

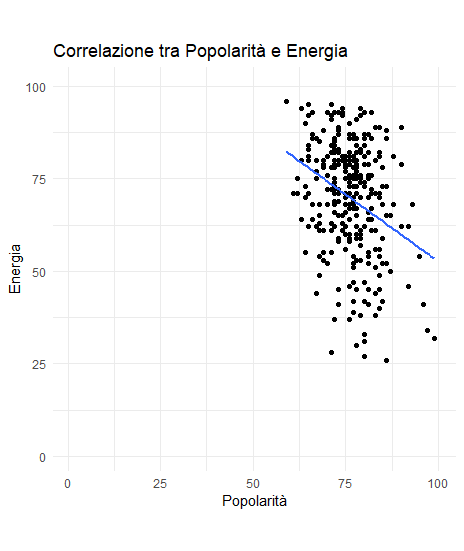
Descrizione generata automaticamente

# Correlation Analysis of Indices

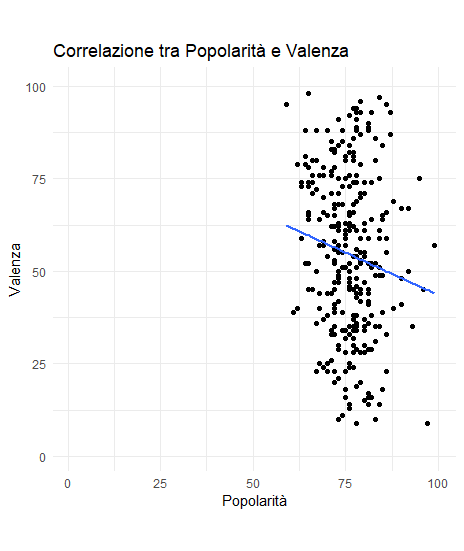
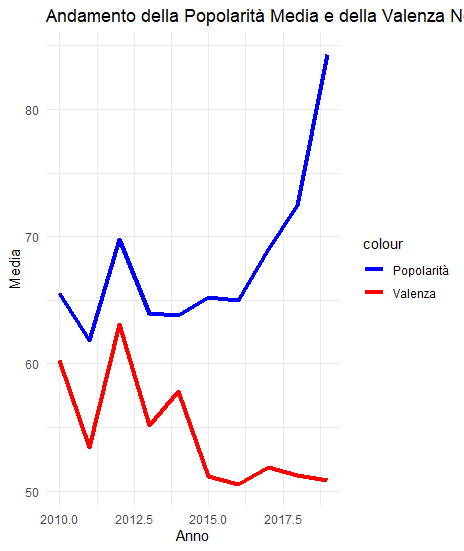
To better understand the possible connection between the musical indices and the popularity of the songs, I performed a correlation analysis between the variables.

**Correlation for the Energy Index (nrgy):** the value is negative and close to zero (-0.310190417883068), therefore it is not possible to establish a correlation and the graphs below confirm this.





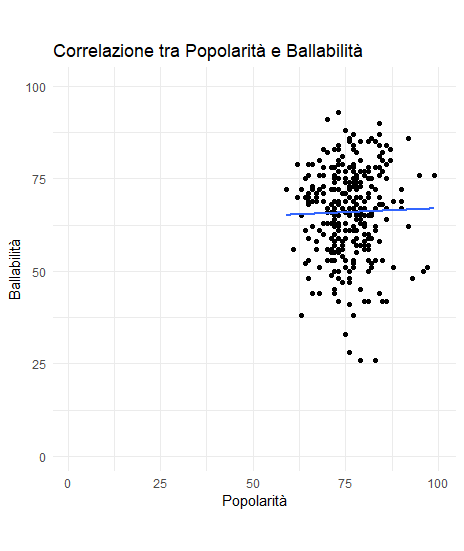
**Correlation Analysis for the Valence Index (val):** the value is negative and close to zero (-0.138666943897798), therefore it is not possible to establish a correlation and the graphs below confirm this.



**Analysis of the Correlation for the Danceability Index (dnce):** the value is positive and close to zero (0.024111726433937), it is not yet possible to establish a correlation however the situation is improving and the graphs below confirm this.

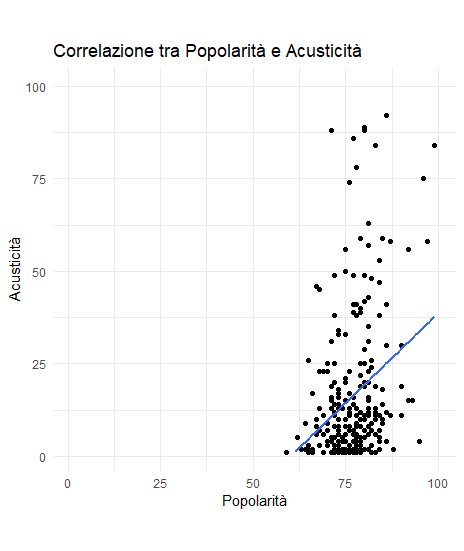
Immagine che contiene testo, diagramma, Diagramma, linea

Descrizione generata automaticamente



**Analysis of the Correlation for the Acoustic Index (acous):** the value is positive and quite close to zero (0.332733253064227), a weak positive correlation is perceived and in fact the graphs show a similarity in the trends.

Immagine che contiene testo, diagramma, Diagramma, linea

Descrizione generata automaticamente

**Correlation Analysis for the Live Index (live):** the value is negative and close to zero (-0.0687243007864658), therefore it is not possible to establish a correlation and the graphs below confirm this.

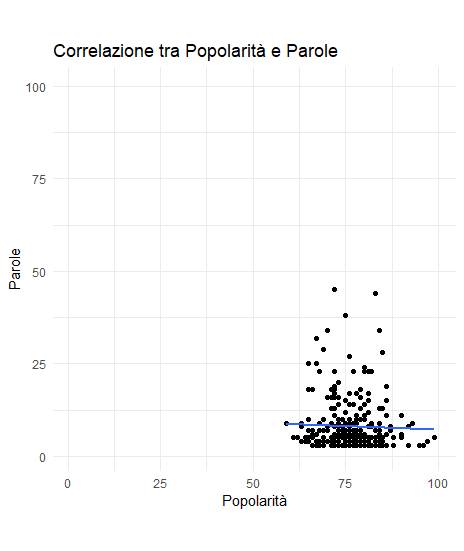
**Immagine che contiene testo, diagramma, schermata, Diagramma

Descrizione generata automaticamenteImmagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente**

**Analysis of the Correlation for the Speech Index (spch):** the value is negative and close to zero (-0.0347385480386376), therefore it is not possible to establish a correlation and the graphs below confirm this.

Immagine che contiene testo, diagramma, schermata, Diagramma

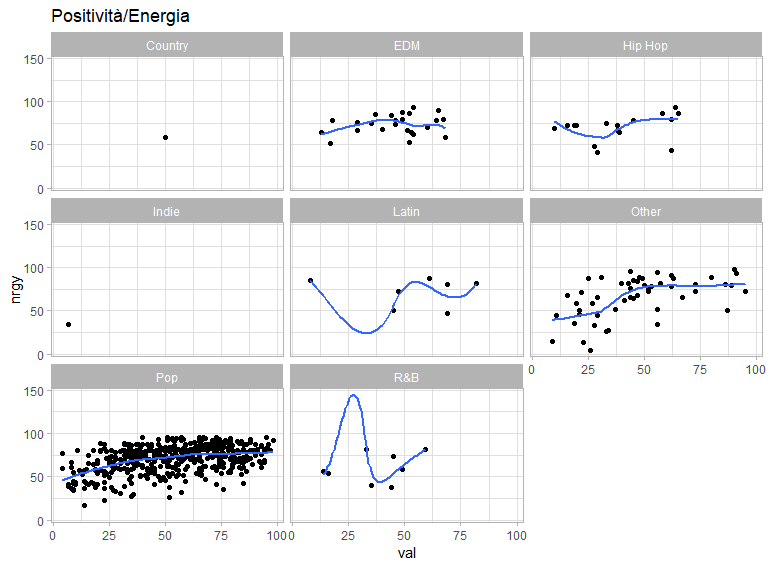
Descrizione generata automaticamente

**There are a correlation between positivity & energy (cor at 0.39), the highest the value so does the energy.**

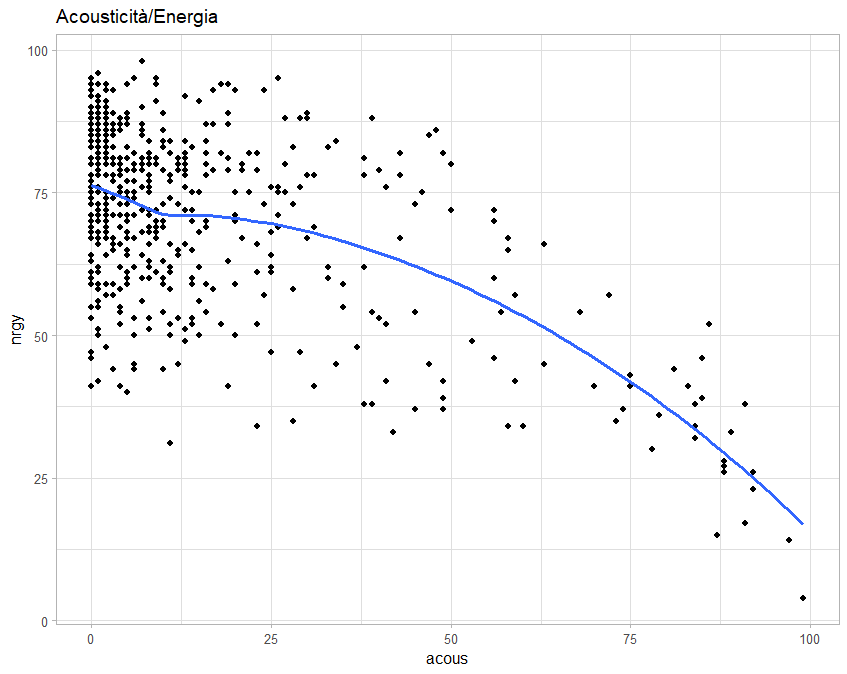
Immagine che contiene testo, diagramma, linea, schermata

Descrizione generata automaticamente

**The genre showing this are: Hip Hop and R&B.**



**The are a correlation between energy & acoustic, the lower the energy so does the acoustic (cor at -0.58)**



Conclusions

The analysis conducted on the dataset extracted from Spotify's repositories contributed to revealing multiple facets of the musical panorama of the decade, but confirms the intrinsic complexity of the music industry and its difficulty in parameterizing based on specific indicators.

Identifying artists, genres and popularity trends offers valuable viewable statistical information, however, the analysis highlights the apparent impossibility of defining a single formula for musical success.

The world of music is intrinsically diverse, influenced by multiple factors such as cultural preferences, social trends and emotional elements, which go beyond quantitative measurements.

Therefore, rather than looking for a definitive formula, the usefulness of this analysis lies in the possibility of providing statistical data and displaying trends, with the aim of offering a comprehensive picture of musical dynamics. This understanding can support informed decisions in the music industry, but it remains critical to recognize the diversity and changeability of audience preferences.

Ultimately, the analysis confirmed my initial thought that musical success is inherently complex and difficult to reduce to specific parameters.