# Musical Characteristics: An Analysis of Genre Popularity and Song Success in Music

Samantha Georgina Isaac Munoz - Student ID: 230057658
Department of Science and Technology
INM430 Principles of Data Science
City, University of London

***Abstract*—The music industry, continually evolving, now stands as one of the largest fields that many with a passion for this art wish to enter, yet they often face challenges in knowing where to start or in gaining traction for their music. This paper presents an analysis, along with three machine learning models, to determine whether there is a relationship between the popularity of a song or a musical genre and its musical characteristics. It also aims to identify how these characteristics vary across different genres. The dataset used for this study has over 30,000 songs records, featuring a total of 22 attributes, including 9 specifics musical characteristics sourced from the Spotify API. This paper seeks to provide insights that could guide aspiring musicians and industry professionals in understanding the dynamics of music popularity and the evolving trends in the music industry.**

***Keywords—Spotify, Music Genre, Popularity, Characteristics, Trends, Regression Models***

## I. Introduction

Over the years, the music industry has grown rapidly, seeing the rise of many new waves of experimentation and the freedom. This environment has offered artists the opportunity to create new musical genres, a phenomenon supported by the theory of the cultural evolution of music, which suggests a continual transformation and fusion of styles and musical forms [1].

However, it's still rare for new artists and their music to be among the most-played or popular songs. This shows a gap between new or different ideas and what most listeners like, meaning that new thigs don't always become widely popular. The characteristics of a song, such as rhythm, energy, musical notes, duration and tempo, may influence their reception of these new music.

With the rising of new streaming platforms like Spotify, Deezer or SoundCloud, it has become possible to determine what makes certain musical genres more popular amongst audiences. These platforms have transformed not just how music is consumed, but also how it is valued and promoted, creating a new ideal in the music industry that favours the availability and accessibility of a wide range of music [2]. This knowledge could be helpful to the music industry and those aspiring to enter it, providing a basic to understand and emulate the attributes of popular music.

## II. Data, Research Questions and Analysis Strategy

### A. Data

The database used for this Project originates from Kaggle, created by the user Joakim Arvidsson, where the records were obtained from the Spotify API. The dataset contains a total of 32,833 songs records, with 22 columns of features, one of which is the target column. There is a mix of numerical and categorical values indicating general information about the song, including the name, album, release date, playlist; as well as technical information such as tempo, danceability, mode, instrumentalness, among others features.

This dataset captures current trends in music that have gained popularity, making it suitable for this project. It contains direct information from one of the most popular and widely used streaming platforms today, Spotify. The only two primary limitations of this dataset are that it features predominantly mainstream music, omitting tracks that may have become popular on social media but are not part of Spotify's repertoire. Additionally, the dataset categorises songs genre based on the genre of the playlist they appear in. Therefore, this aspect will be taken into consideration, and the *'playlist_genre'* will be retained as the genre of the song, as it represents the genre under which these specific songs and the entire playlist have been classified by users.

### B. Research Questions

The main objective of this project is to identify and analyse if the musical characteristics of the songs can determine the popularity of it and how these factors vary across different musical genres. This involves delving into a comprehensive analysis of musical attributes to discern patterns and trends that influence a song's success in the music industry.

The questions that this project aims to answer are:

- Is there a correlation between the popularity of a song and its musical characteristics?
- In what ways do musical characteristics differ among music genres?
- How have musical genres evolved over time?

### C. Analysis Strategy

To achieve the goal of answering the research questions, the following analysis strategy was developed:

1. Obtain the 'spotify_songs' dataset in CSV format from Kaggle.
2. Load the dataset into a Jupyter Notebook and commence with an initial analysis to conduct data cleaning and feature engineering (correcting formats, handling missing values, removing unnecessary columns).
3. Conduct a more in-depth second analysis, focused on addressing the research questions. Explain and analyse insights from the various musical genres using informative graphs.
4. Develop two predictive models to determine the popularity of a song based on its musical characteristics (numerical data).
5. Make potential adjustments, changes, and tuning to the models to enhance its accuracy.
6. Analyse the results and draw conclusions.

## III. ANALYSIS

### A. Data Selection

During the data selection process, six columns, or features, were removed as they were irrelevant to the objective of this project. Additionally, half of these columns are IDs generated by the Spotify's API. Since the project does not involve working with the API, these IDs are not useful. Other columns where either removed or created during the cleaning and feature engineering process. At the end a total of 19 features were being retained for the analysis process and for the training of the models.

The Figure 1 provides a concise overview of the musical characteristics discussed in this research paper, aiding in understanding their role. These columns will be used on our models to predict if this characteristics cause an impact on a song's popularity:

| Feature | Brief description |
|---|---|
| Danceability | Indicates how suitable a song is for dancing. |
| Energy | Perceptual measure of intensity and activity. |
| Key | Musical key of a song, if the key can't be identify, it's marked as -1. |
| Loudness | Overall loudness of a song in decibels (dB). |
| Mode | Modality (major or minor) of a song. |
| Speechiness | Presence of spoken words in a track. |
| Acousticness | Likelihood of a song made with acoustic (non-electronic) instruments. |
| Instrumentalness | Indicates if the song doesn't contains vocals. |
| Liveness | Indicates if the song was performed live. |
| Valence | Musical positivity conveyed by a song. |
| Tempo | Tempo of a track in beats per minute (BPM). |

Fig. 1. Musical characteristic with the description.

### B. Data Preparation

The data preparation process was relatively simple, involving a few but significant steps in cleaning the data. A total of 23,449 records were removed, nearly a third of the entire dataset. This step was taken after an initial analysis where duplicates were searched for. No exact duplicates were found, but when checking the *'track_name'* column (which refers to the song name), several songs were discovered to be repeated in different playlists. Removing these repetitions was crucial to avoid imbalance and biased results in both the analysis and the models.

Only five rows with missing values in various columns were identified and removed. Feature engineering was then applied with the *'track_album_release'* column being converted to datetime values, and three new columns were created to separate the date into year, month, and day. Similarly, for the *'duration_ms'* column, a new column was created converting the milliseconds into minutes, and the original column was subsequently removed.

To conclude this section, a pre-processing for the data was conducted prior to the creation of the models, the data was divided into training and testing sets, allocating 70% for training and 30% for testing. Following this, normalization of the musical characteristics was performed using the *StandardScaler* form the *scikit-learn* library. This step was necessary due to the selected models and substantial variation in feature scales.

### C. Analysis and Construction of the models

During this phase, a secondary data analysis was performed which involved creating various graphs and thoroughly examining them for insight that contribute to a deeper understanding of the data and the relationship between each musical characteristic.

As part of this analysis, a correlation heatmap was created to explore potential correlations between song popularity and various musical characteristics. This visual representation allowed us to assess the strength and direction of these relationships. The goal was to identify any musical characteristics that might have a high impact on a song's popularity.

Visual representations helped in the understanding of distinctive attributes for each genre. This information was crucial for identifying patterns and trends in addressing research questions effectively.

Three supervised learning regression models were used to predict the song popularity based on various musical characteristics. Initially, a Lineal Regression model was implemented as a baseline to get a scope of a regression model performance with the data and the chosen characteristics. Then two more models, namely Random Forest Regression and Extreme Gradient Boosting (XGBoost), were developed. In addition, a random search was added to optimise the parameters for both models. The Random Forest Regression model was chosen because of its simplicity of implementation and robustness, which reduces susceptibility to errors. XGBoost selection was based on its superior performance, versatility and capacity to address various data type, enhancing the model's ability to generalize by optimizing the loss function and regularization [3].

### D. Validation of Results

To evaluate the models, four metrics were employed: $R^2$ (Coefficient of Determination), MAE (Mean Absolute Error), MSE (Mean Squared Error), and lastly RMSE (Root Mean Squared Error). These metrics were chosen for their widespread use in assessing the performance of regression models. They have become standards tools in the field of analysis due to their effectiveness and applicability.

- $R^2$ is a measure that shows the degree to which variations in a dependent variable can be predicted from independent variables [4].

$$R^2 = 1 - \frac{SSresidual}{SStotal}$$

- MAE measures the average magnitude of errors between paired observations which depict the same phenomenon [5].

$$MAE = \frac{1}{n}\sum_{i=0}^{n} |y_i - \hat{y}_i|$$

- MSE measures the closeness of data points to the fitted regression line [6].

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$$

- RMSE shows the magnitude of the errors between the models predictions and the real data points [6].

$$MSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}$$

## IV. Findings, Future Work And Conclusion

### A. Analysis Findings

The first and most significant observation made was that there is no strong correlation between musical characteristics and the popularity of a song. Figure 2 shows that most correlations are neutral or very weak, suggesting that there isn't a single dominant musical characteristic determining a song's popularity. This also could indicate that popularity is multifaceted, and other factors not considered in this project, such as marketing, social media, or cultural influence, might play a part.

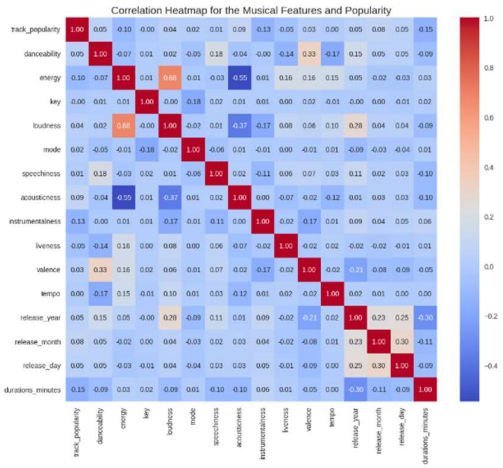


Fig. 2. Correlation Heatmap for the Musical Features and Popularity.

Pop, Rap, and Latin genres have the highest average popularity among genres, indicating a greater trend towards these styles. On the other hand, EDM shows the lowest average popularity. However, examining both bar graphs (Figure 3) can refine this perspective by revealing that within the EDM genre, the '*pop edm*' subgenera is the most popular, aligning with the high average popularity of the Pop genre.

An important point to note is that innovation and the fusion of genres, along with the rising of new subgenres each year, may be playing a key role in what currently resonates with listeners. This dynamic of music and their constant evolution could be a significant factor in shaping the trends of music popularity.

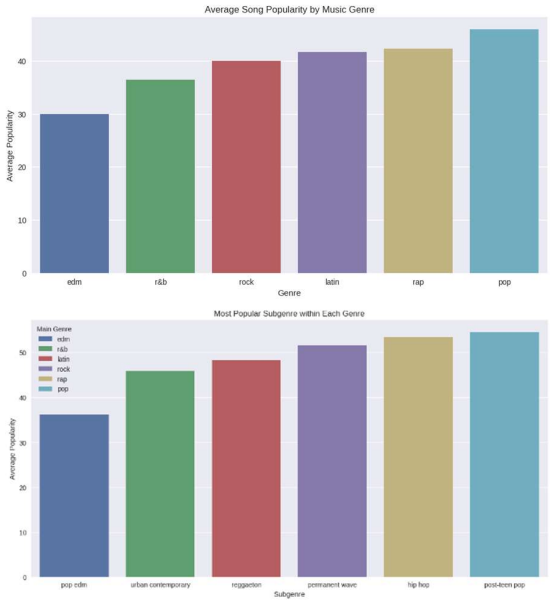


Fig. 3. Most Popular Genre and Subgenre within Each Genre

Pop and EDM exhibit higher danceability among all genres, while Rock and EDM lead in terms of energy and loudness, aligning with their often intense and energetic styles. Rap is distinguished by its prominence of spoken lyrics, reflecting the significance of the words in the genre. Latin and R&B show greater acoustic elements, suggesting a strong presence of organic elements in their music. EDM is characterized by high instrumentality, indicative of a focus on electronic production. Liveliness is low across all genres, implying few live recordings. Pop and Latin stand out in positivity, associated with lively rhythms. The tempo in EDM genre varies widely, demonstrating its versatility and experimental tendencies.

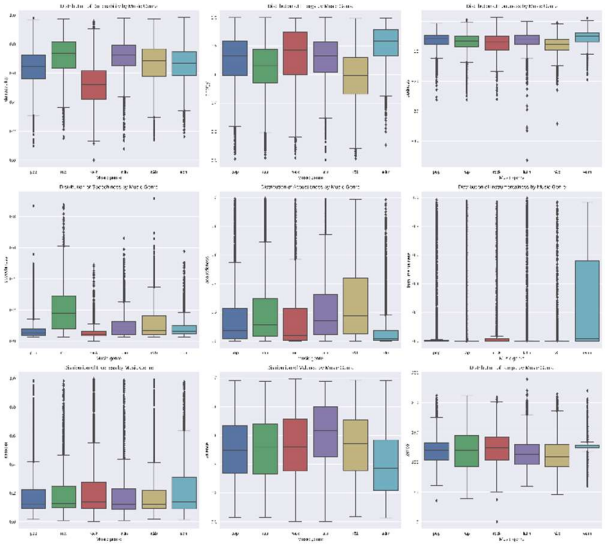

Fig. 4. Distribution of the Musical Characteristics by Music Genre.

The genres R&B and Rap have gained popularity in recent times, Rock has maintained a constant presence with a slight decline, Latin genre shows a recent surge in the last few decades, Pop remains stable demonstrating adaptability to trends, and EDM, after peaking in the 90's, is regaining popularity. These patterns underscore the fluidity of popular music and its interaction with cultural and technological changes over time.

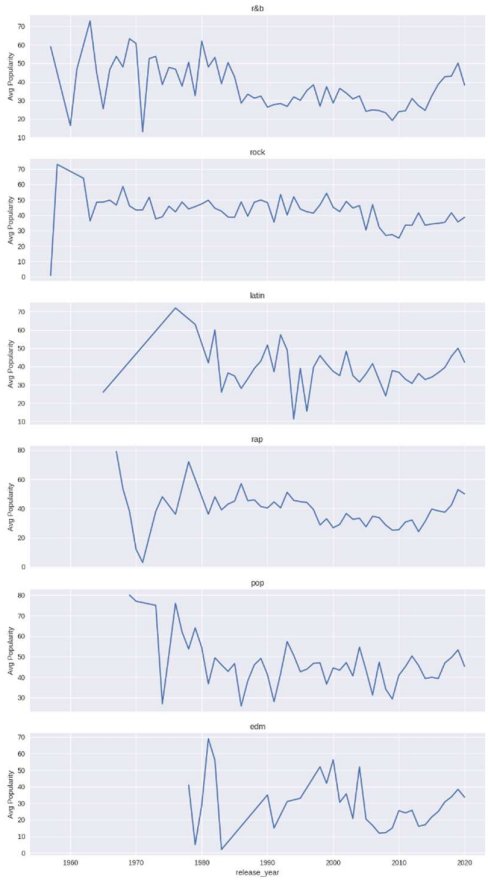


Figure 4. – Popularity of Genres Across the Years

## B. Models Findings

Figures 5 and 6 displays the performance of the three regression models, where none of them achieved good results in predicting popularity based on the nine musical characteristics used.

| Models | R2 | MAE | MSE | RMSE |
|---|---|---|---|---|
| **Lineal Regression** | 0.068 | 18.832 | 511.119 | 22.608 |
| **Random Forest** | 0.618 | 11.832 | 209.707 | 14.481 |
| **XGBoost** | 0.196 | 17.367 | 440.755 | 20.994 |

Fig. 5. Scores for the Training Data.

| Models | R2 | MAE | MSE | RMSE |
|---|---|---|---|---|
| **Lineal Regression** | 0.054 | 18.891 | 516.793 | 22.733 |
| **Random Forest** | 0.079 | 18.552 | 502.97 | 22.427 |
| **XGBoost** | 0.088 | 18.439 | 498.298 | 22.323 |

Fig. 6. Scores for the Test Data.

## C. Conclusion

The findings from the analysis and the outcomes of the models reveal that there is no direct relationship between the popularity of a song and its musical characteristics. This suggests that popularity depends on factors beyond these features, such as music production, marketing, social media, among others. Additionally, music has evolved and grown significantly over the years, always presenting a wide variety of new genres influenced by the context and era in which they emerge, which in turn affects their popularity. These results are a good start point for further analysis on what makes a song popular, as this study was limited to only musical characteristics and did not extend to social aspects or deeper industry factors.

## D. Future Work

It's necessary to acquire more datasets that include recent music and more than just technical metrics. Additionally, focusing on data beyond technical values and seeking information from sources other than solely Spotify will provide a more comprehensive understanding. Furthermore, conducting a deeper analysis for each genre to observe their evolution through time and identify the most prevalent characteristics in the most popular songs of each genre would be potentially revealing.

# V. Word Counts Per Section

| Section | Word Count |
|---|---|
| Abstract | 142 |
| Introduction | 214 |
| Research questions | 97 |
| Data | 203 |
| Analysis | 832 |
| Findings | 587 |

# References


[1] P. E. Savage, 'Cultural evolution of music', Palgrave Commun, vol. 5, no. 1, Art. no. 1, Feb. 2019, doi: 10.1057/s41599-019-0221-1.

[2] P. D. Townsend, *The evolution of music through culture and science*, First edition. Oxford, United Kingdom: Oxford University Press, 2020.

[3] H. Tian, H. Cai, J. Wen, S. Li, and Y. Li, 'A Music Recommendation System Based on logistic regression and eXtreme Gradient Boosting', in *2019 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2019, pp. 1–6. doi:10.1109/IJCNN.2019.8852094

[4] 'R-Squared', Corporate Finance Institute. Accessed: Dec. 15, 2023. [Online]. Available: https://corporatefinanceinstitute.com/resources/data-science/r-squared/

[5] C. Willmott and K. Matsuura, 'Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance', *Climate Research*, vol. 30, p. 79, Dec. 2005, doi: 10.3354/cr030079

[6] M. Khan and S. Noor, 'Performance Analysis of Regression-Machine Learning Algorithms for Predication of Runoff Time', *Agrotechnology*, vol. 08, no. 01, 2019, doi: 10.35248/2168-9881.19.8.187.