Exploring Trends in Music Platforms: A Comparative Analysis of Key Factors for Trending Songs on Spotify and YouTube

Anna Charchyan
American University of Armenia
BS in Data Science
Email: anna_charchyan@edu.aua.am

Supervisor: Ashot Abrahamyan
American University of Armenia
Email: ashot.abrahamyan@aua.am

***Abstract*—The music industry is defined by strong branding and public image, which generate success for music-related content, artists, and labels. The music industry has exceptionally high entry barriers for new artists or labels, who often struggle with reaching popularity and success. We have an interest in gaining a thorough knowledge of why certain musical content is more popular than others. To address this inconsistency, we will try to identify the key aspects which have the most influence on the popularity of songs. This study analyses which are the most significant aspects that make a song more popular than others and explores the impact of streaming data on music industry trends and decision-making processes. Performing extensive Exploratory Data Analysis, this research aims to gain insights into the distributions and relationships between different key features. Implementation of various visualization techniques leads us to the analysis of trends and patterns based on the characteristics of the songs. Through the use of machine learning techniques, we further predicted future patterns in music consumption along with providing observations that can help us better comprehend constantly developing digital streaming platforms' trends and future strategies for commercial success.**

I. Introduction

Music has always played an unseparated role in our society as we are constantly exposed to music in our everyday lives. Music streaming platforms are now recognized as major players in the constantly developing digital media environment, each providing different options for music consumption and distribution. Spotify and YouTube platforms have a significant portion in defining musical tendencies as well as identifying which artists reach trending charts and which songs become "hits." Those platforms have various characteristics, algorithms, and audiences, which leads us to the fact that the data can vary radically from one another. Many researchers have focused on analyzing the internal structure of songs in order to better understand the fundamental factors that contribute to their popularity, employing a variety of approaches and data sets. With the development of digitization of music content, platforms, as the primary "instrument" for music listening, obtained the capacity to retrieve a broader range of information with more precision than ever before. Metrics like the total number of plays per track or the genre of the song help clarify and comprehend the industry. People are starting to apply data analytics to uncover tendencies, enabling better findings and allocation of resources. Nevertheless, our research will provide a more comprehensive understanding of the subject.

II. Data Collection & Description

Data collection was facilitated through Kaggle, the relevant datasets were downloaded, ensuring access to a diverse range of song data for comprehensive analysis. It was an important part to find the right source with accurate and usable information. After a research on key components it was decided to collect two datasets for Spotify and four datasets for Youtube. Spotify's dataset contains 7154 observations of 19 variables, and YouTube's dataset contains 225623 observations of 15 variables.An important step was to ensure that the datasets are accurate, usable and reliable, thus data cleaning was performed.

A. Data Processing

Category mapping was performed and synchronized for combining. For performing statistical analysis of music industry videos, we filtered the data by `channel_title` in YouTube's dataset. Due to the accuracy of Spotify's datasets, only a few artist names needed improvements, such as from "Beyoncé" to "Beyonce" and from "ROSALÍA" to "ROSAL". Formerly, the names of the columns in the two datasets were changed to align together in the new combined dataset. Finally, data type conversions were made.

B. Audio Features Description

Audio features from the Spotify Web API include track popularity, danceability, energy, and other parameters that describe music tracks. Track Popularity represents the popularity of the song; the higher the number, the more popular is the song. These numbers are generated by Spotify's algorithm. 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. Typically, energetic tracks feel fast, loud, and noisy. Key represents the key the track is in. Integers map to pitches using standard Pitch Class notation. Loudness represents the overall loudness of a track in decibels, which typically range between -60 and 0 dB. 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 by 0. Speechiness detects the presence of spoken words in a track. Values below 0.33 most likely represent music and other non-speech-like tracks, while values above 0.66 describe tracks that are probably made entirely of spoken words. Acousticness has 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. The closer the instrumentalness value is to 1.0, the greater the likelihood the track contains no vocal content. Liveness detects the presence of an audience in the recording. Valence has a measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration [1]. A number of important features are missing from the dataset such as gender of the artist, day released and total streaming day, which somewhat gives a small limitation on the ability for a deeper statistical analysis.

III. Methods

This section describes our approach to identifying key features of the trending songs. The development of the methods used can be divided into five main parts:

1) Data exploration and statistical analysis
2) Relationship exploration of features
3) Hypothesis testing
4) Machine learning techniques
5) Forecasting

A. Data Exploration and Statistical Analysis

The first steps in data exploration involve gathering information on the most popular artists on Spotify. This provides insights into trends and patterns that may influence the popularity of songs. With the first visualization, we gain insights into artists dominating top charts throughout the years.

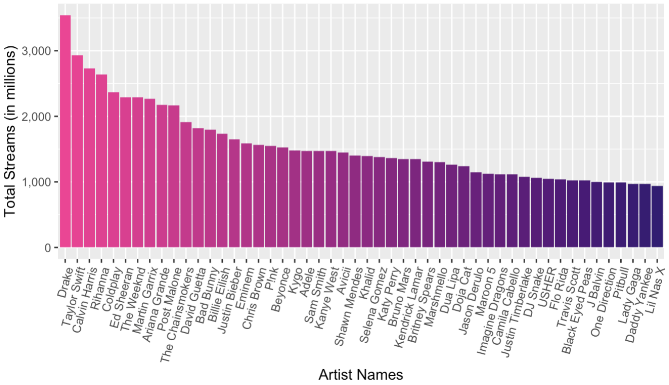

Fig. 1. Top 50 Artists by Popularity on Spotify

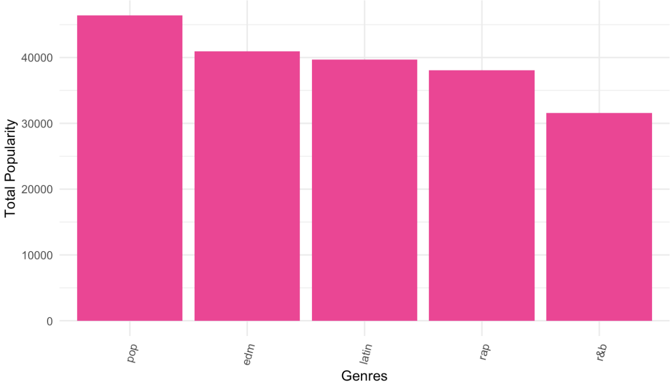

Fig. 2. Top 5 Genres by Popularity

The Fig. 1 represents the top 50 artists by popularity throughout the years with Drake, Taylor Swift, and Calvin Harris sitting in the first, second, and third spots, respectively. Each artist is ranked by their popularity on Spotify and depicted in decreasing order. By seeing artists that took first places we can see a pattern not only of reached peak popularity but also a trend in maintaining on charts for a considerable period of time. This plot gives insight into artist dominance and longevity as Spotify rankings point out the impact of digital streaming platforms on music trends, which can be a characteristic of how effectively artists use these platforms and their abilities to create and produce hits on a regular basis and adjust to evolving musical trends.Analyzing current trending musical genres could shape the production styles of new songs, by integrating components from specific genres into their music in order to attract a larger audience. Fig. 2 represents the top 5 most popular genres, which include pop in the first place, Latin taking second spot, EDM, Rap, and R&B, respectively, third, fourth, and fifth spots. This gives us an insight into trending genres of music, which can have an important influence on the trending factor for the song. Highlighting the areas with the highest listener demand can serve as a guide for record labels, producers, and musicians on what genres of music to create, support, and invest in. This gives us an insight into trending genres of music, which can have an important influence on the trending factor for the song.Determining which music styles are trending is vital for promoters and marketers who want to adapt marketing campaigns, music festivals, and concerts more efficiently. This insight helps them target their target market with greater precision, thus maximizing interaction and engagement, which leads to their commitment to song trend changes. Observation of distributions of audio features such as Tempo, valence, acousticness, energy, and danceability is crucial for understanding various applications in the music industry as well as detecting possible outliers. The density plot of danceability in Fig. 3 indicates binomial distribution, which represents two main categories of songs—less danceable and extremely danceable. This indicates a division in music preferences, with certain genres or musicians focusing on less danceable

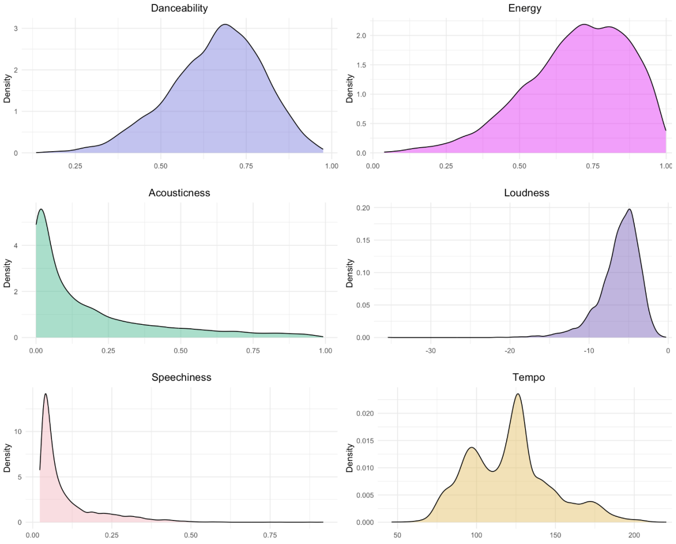

Fig. 3. Density Plots

songs by concentrating on more relaxing listening settings, while others may concentrate on more danceable songs which are suited for venues such as clubs and parties. The density plot of energy indicates beta distribution which depicts that the majority of songs focus on certain energy degrees, with fewer songs at the lowest and highest peaks. Thus, we can conclude that there is an extensive preference for balanced levels of energy in songs.The density plot of Acousticness and Speechiness represents the lognormal distribution, which shows that, while the majority of tracks have a lower score of both characteristics, there is a large tail of tracks that have significantly higher parameters. This can represent characteristics of certain genres, such as rap, which has high speechiness, or folk, which has high acousticness. The density plot of loudness represents the Pareto distribution, which suggests that a mere group of songs is significantly louder compared to the general preferences of the audience. This tendency may be associated with genres that emphasize significant impact and presence via volume. This can be noticed particularly in electronic or rock genres. The Density plot of Tempo represents bimodal distribution and proposes two declared Tempo ranges for popular songs. The pattern may reflect a difference among genre preferences—for instance, one peak might indicate fairly slow-tempo, relaxed songs, while another might indicate dancing music. The determination of mode and the corresponding common Tempo may affect the sentimental and emotive plangency of the melody (see Figure 17 in the Appendix). Both histograms of minor and major modes display a wide range of tempos, with a major peak at 120 beats per minute. However, the minor mode has lower peaks at both slower and faster tempos. This implies that the speed of songs composed in minor mode changes, presumably expressing a variety of styles and feelings associated with minor tonality. Meanwhile, the major mode has fewer songs at much lower or higher tempos than the minor mode. This may imply a tendency for this speed in modern dance and pop genres, which usually use major tonality for their lively and cheerful features. The insignificant contrast in prevalence between minor and major modes provides insight that a song's keynote might not be a critical aspect for determining its popularity(See Figure 18 in the Appendix). Setting up this data, the set of plots visualizing changes in valence, danceability, energy, and tempo of the tracks over a decade provides an in-depth representation of the way musical characteristics have changed. The rise in danceability, illustrated in Fig. 4, starting from 2016 demonstrates an increasing public demand for songs that are captivating and appropriate for vibrant social settings. Likewise, the tempo plot displays notable changes, including a late increasing trend, which supports the idea of a trend toward speedy songs. In contrast, the valence plot, which reflects the positiveness of the songs, shows a significant increase following a period of decrease, indicating a shift towards songs that reflect a lighter and happier mood. However, notwithstanding the increase in the positiveness of the songs, the energy plot has decreased in the last few years, indicating that even though songs get more lively and more joyful, they are additionally embracing a less thrilling, calmer energy pattern. These trends are valuable for multiple stakeholders in the music industry. For artists and producers, understanding these trends can guide the creative process to align with current listener preferences, potentially increasing the marketability of their music. Understanding these patterns can help artists and producers adjust their production methods to current audience preferences, thereby enhancing the potential success of their music. Record companies, streaming providers, and playlist curators can utilize this data to design releases in a more effective way and modify their genres, song attributes, and promotional strategies to better reflect their listeners' shifting interests, which leads to song success. The charts depicting statistical analysis of the YouTube dataset show two

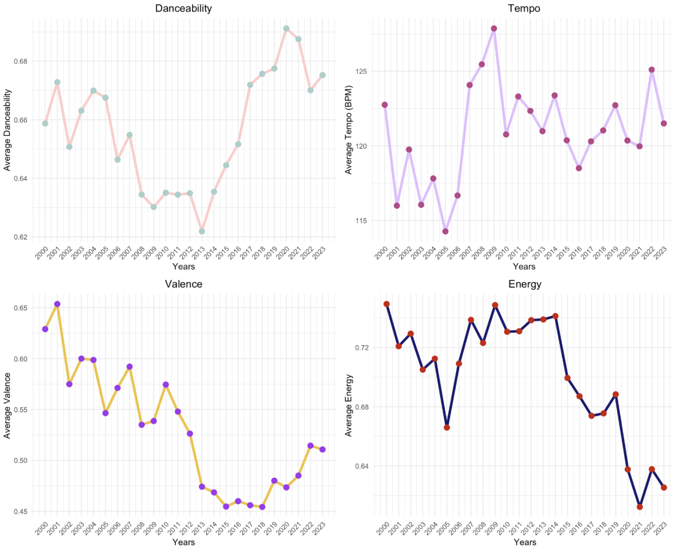

Fig. 4. Change in Song Characteristics Over the Years

independent but linked elements of channel depiction. The first chart illustrates the top 50 YouTube channels based on total views, with Marvel Entertainment and MrBeast placing in first spots. This demonstrates the dominant position of particular creators of content who are able to reach a large audience through parameters such as the popularity of their brand, reputation, content diversity, and constant interaction with viewers. We have identified a total of 367 VEVO artists in our dataset, excluding radio channels. The artists' main pages are intricately linked to their VEVO counterparts, ensuring a seamless navigation experience. Furthermore, all VEVO videos are set up to redirect users to the artists' main pages. This visualization focuses exclusively on videos that are branded under VEVO, highlighting their unique distribution and reach. Figure 6 looks at the success of prominent VEVO channels by comparing the number of videos each channel has created to the total number of days the videos have been popular. ”VEVO” channels have a huge number of clips and significant trending days, suggesting an effective strategy of frequent content launches that keep viewers engaged over an extended period. This implies a clear link between active content creation and continuous exposure, as well as popularity. By combining these analyses, we can conclude that reputable YouTube channels, even if it is widespread amusement such as Marvel and MrBeast or music related such as VEVO channels, exhibit characteristics such as proliferative content production and effective collaboration that have popularity among large audiences.

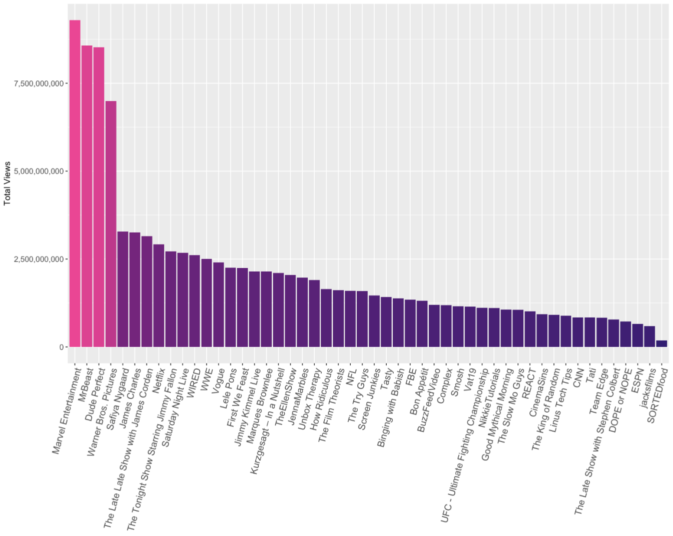

Fig. 5. Top 50 YouTube Channels by Total Views

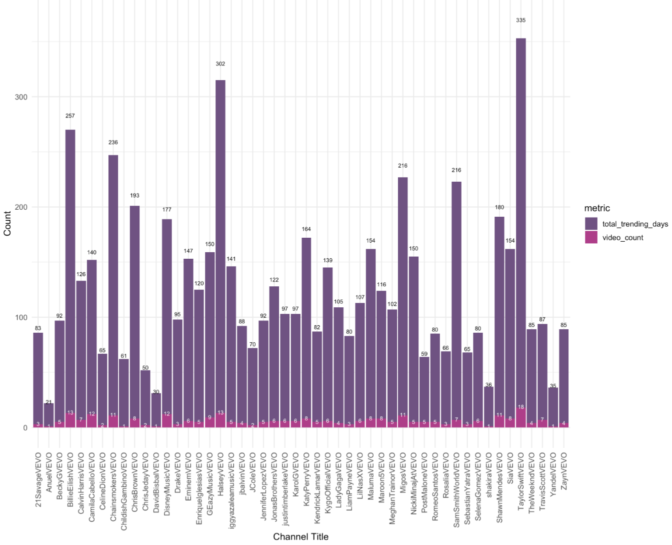

Fig. 6. Top 30 VEVO Channels: Video Count vs. Trending Days

B. *Relationship exploration of features*

In the sections that came before it, we constructed an initial expiration about distinct characteristic distributions and how they impact the areas of demand. Developing on this base, we now focus on the relationship and interactions between these traits. This part, named 'Relationship Exploration of Features,' is focused on discovering the chain of correlations that are present between distinct features in the data we have collected. By methodically examining these interactions, we hope to unearth more profound conclusions. We will use a number of data analysis tools, ranging from scatter plots and correlation matrices to more advanced regression models, each designed to demonstrate best the type and intensity of the correlations between variables of interest. Throughout this comprehensive study, we desire to bring readers to an exhaustive overview of the dynamics, therefore improving the totality of the findings. The correlation graphs are displayed in Fig. 7 analyze the relationship between positiveness and energy across a variety of music genres. Plot includes EDM, Latin, Pop, R&B, Rap, and Rock genres. Charts demonstrate specific trends in how various genres mix certain musical characteristics. For example, genres such as Latin and Pop have a large clustering in the upper right region, indicating a significant

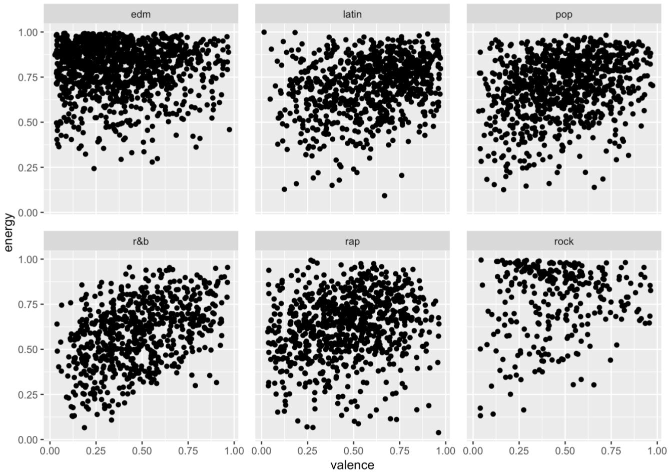

Fig. 7. Correlation between Positivity and Energy

relationship between positivity and increased energy, which is consistent with their energetic and dynamic nature. This trend demonstrates a preference across these categories for music that is both vibrant and strongly uplifting. In contrast, the Rock and Rap genres show broader scatter plots, indicating a less consistent relationship between energy and positivity. Rap music, as shown above, has a wide distribution over the energy axis, independent of valence, reflecting a wide range of expressive emotions, from high-intensity festival melodies to more somber or aggressive pieces. Rock music naturally displays high levels of energy, irrespective of valence, exhibiting the style's emphasis on tension and strength. Ensuing the examination of the correlation across valence and energy among genres, we evaluated the relationships among distinct audio characteristics and song popularity in order to obtain a deeper study of what attracts audiences. The correlation coefficients estimated among different musical characteristics and song popularity produce beguiling outcomes. Insignificantly, danceability has practically no association (0.068) with song popularity, despite valence having a somewhat positive relationship (0.108) in contrast to other factors. Nevertheless, loudness (0.058), liveness (-0.040), song duration (-0.059), acousticness (0.070), speechiness (-0.009), instrumentalness (-0.242), Tempo (-0.031) and energy (-0.132), have varying associations, ranging from negative to practically no relationship[1]. This emphasizes the diverse aspect of music consumption, which demonstrates how a complex combination between components define a track's success. These results indicate that the pattern of popularity can not be attributable to a particular one factor. The composite values for correlation demonstrate that a single musical component does not persistently regulate a track's popularity on the platform. Therefore, what we have discovered reveals that song popularity is determined by a number of criteria rather than one particular variable. This understanding is crucial for comprehending the larger dynamics of music consumption, emphasizing the significance of taking a comprehensive strategy when evaluating the success of the song.

Further exploration of how distinct musical characteristics impact one another reveals that the relationships between these factors have a major impact on the listening experience, revealing intricate interactions within diverse musical environments.The adverse correlation between acousticness and energy (-0.57)[2] suggests that acoustic songs tend to be more subdued and calmer than amplified or electronic music, which is consistent with standards within styles that emphasize organic sounds. Inversely, the significant positive relationship among energy and loudness (0.71)[2] indicates that as songs get more dynamic, they also tend to be noisy, a trait typical in categories such as rock and techno that try to create a dramatic effect. Furthermore, the balanced positive relationship within valence and danceability indicates an interplay within a positivity and the song's rhythm, which is compatible due to its frequent utilization in joyful and festive environments. Furthermore, the slightly adverse relationship that exists between energy and instrumentalness indicates that intensely instrumental songs may not necessarily be high-energetic, particularly in more relaxed or classical music situations ((see Figure 19 in the Appendix).

In a comparable manner in the industry of video content on YouTube, the analysis demonstrates significant positive associations among the number of dislikes and the number of comments (corellation equals to 0.93), along with a proportionally significant relationship within the total number of dislikes and comments (0.96), demonstrating that clips that elicit greater engagement produce a greater number of dislikes and likes[3]. This indicates that participation, likable or hateful, has a major impact on a video's prominence (see Figure 20 in the Appendix). Furthermore, the overall number of views has substantial positive relationships with comments (0.87), dislikes (0.88), and likes (0.89), indicating that increased participation in any form corresponds with larger viewing statistics[3]. Trending days also have significant relationships with views (0.64) and likes (0.72), implying that clips that stay in trending positions for a longer period of time gather a greater number of views and likes[4]. Likewise, a significant relationship between total trending days and video count (0.91)[4] suggests that channels with a greater number of videos receiving their content trend over a greater duration demonstrate the overall effect of possessing a greater resource library, which increases the likelihood of video trending. The visual representation of relationships with trending days can be obsereved in Fig.8. The findings from different industries highlight how intricate relationships between parameters have significant effects on audience involvement as well as material popularity among platforms. Establishing on earlier comprehensive findings that point out intricate relations among multiple features that have a significant impact on listeners' involvement along with song popularity throughout Spotify, the collection of scatterplots demonstrates particular inter-

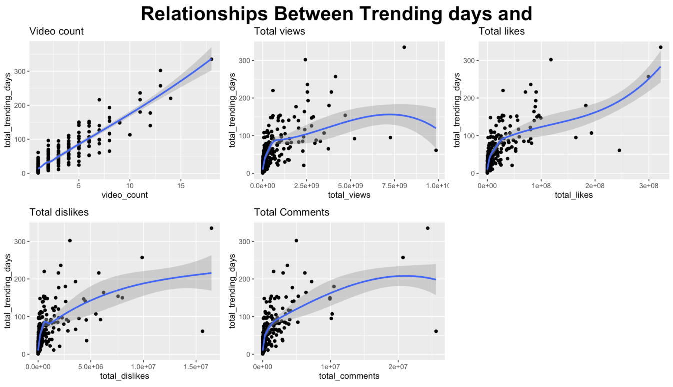

Fig. 8. YouTube caharcteristic relstionships

[1] R Markdown chunk: 'correlation-result', file: 'Spotify_Analysis.Rmd'.

[2] R Markdown chunk: 'heatmap-matrix', file: 'Spotify_Analysis.Rmd'.

[3] R Markdown chunk: 'heatmap-matrix', file: 'Youtube_Analysis.Rmd'.

[4] R Markdown chunk: 'relationships-features', file: 'Youtube_Analysis.Rmd'.

actions among musical features according to two tonalities, mode 1 (major) and mode 0 (minor). The first plot of Fig. 9 illustrates a significant positive correlation within loudness and energy, which is ordinary since louder audio recordings are more energetic. The points are tightly clustered and develop together, displaying that when tracks become louder, they typically suggest more energy. Another plot depicts a slightly positive relationship within danceability and energy, demonstrating that songs that have a greater likelihood of being danceable likewise tend to be more energetic, however with significant deviation, indicating the presence of energetic songs which are not always danceable and vice versa. These scatterplots, which represent the interactions within minor and major tonal systems, contribute to our comprehension of the manner in which various musical characteristics interact to form the listening engagement. This reinforces prior studies and demonstrates audio features interact with each other to determine music's overall effect and popularity.

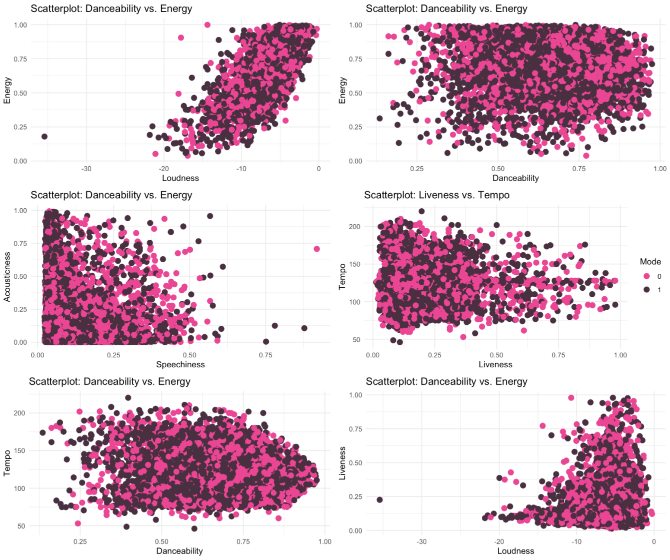

Fig. 9. Scatterplots: Spotify

C. Hypothesis testing

In our study, we intend to test hypotheses in order to evaluate the impact of different characteristics on the rate at which tracks become trendy. This analysis is guided by four key hypotheses: the Genre Impact Hypothesis, the Song Features Hypothesis, the Video Count Hypothesis, and the Views, Likes, Dislikes, and Comments Hypothesis. The Genre impact Hypothesis proposes that a song's genre has no significant influence on its Spotify trending speed (Null Hypothesis, H0), whereas the Alternative Hypothesis proposes that certain types of genres trend quicker (H1). In a similar vein, the Song Characteristics Hypothesis analyzes whether a track's essential musical components, such as loudness, duration, and Tempo, etc, impact its rise to the top trending charts (H1), as opposed to the null hypothesis (H0), which states that these characteristics have no impact on trending rate. The Video Count impact Hypothesis proposes that Video count does not significantly affect the total trending days(H0), whereas the Video count significantly influences the total trending days, with an increase in video count leading to an increase in trending days (H1). In a similar form, the last hypothesis states that the total views, likes, dislikes, and comments do not significantly influence the total trending days (H0). Vs. the Alternative Hypothesis, which declares that these metrics significantly influence the total trending days. To evaluate these hypotheses, we are going to apply linear regression, particularly multivariate regression. Linear regression is a method of statistical analysis for modeling the interaction between a dependent variable and one or more independent variables by modeling a linear equation of the data that is observed. Multivariate regression deepens this notion by containing various independent variables. The multivariate linear regression formula is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

In this model, y is the dependent variable we aim to predict. The intercept, denoted as β_0, anchors the regression line, while the coefficients $\beta_1, \beta_2, \ldots, \beta_n$ represent the effect of each independent variable $x_1, x_2, \ldots, x_n$ on y. The term ϵ indicates the residual error, accounting for differences between the predicted and observed values. Employing the above-mentioned regression model, we can assess the degree of significance and nature of the interactions within song genres or musical characteristics and the rate at which tracks get popular on Spotify. By analyzing the regression model's coefficients, we can evaluate the impact of each factor and whether or not the results support the alternative or null hypotheses. The first set of results from the Genre Impact Hypothesis demonstrated that musical genres do have a significant influence on the rate at which songs become trending, with genres such as Rock, Latin, Rap, R&B, Pop, and Latin having statistically significant positive parameters, with values such as 8.60 for Pop and 7.69 for Latin[5]. The evidence presented here supports rejecting the notion of a null hypothesis (H0) in favor of the Alternative Hypothesis (H1), suggesting that specific musical genres trend more quickly on Spotify than others. At the exact same time, the Song Characteristics Hypothesis results demonstrate that qualities like danceability, energy, instrumentalness, and duration_ms have significant values, demonstrating a decisive effect on the rate of tracks achieving trending charts. Danceability is positively related to popularity, with a one-unit increase resulting in a 12.46 rise in popularity[5]. Energy, on the contrary, is negatively correlated, implying that increased energy leads to lesser popularity. Predictors such as key, mode, speechiness, acousticness, liveness, valence, and Tempo are statistically insignificant at the 0.05 level[5], implying that their effect on the trending rate is doubtful. Loudness and instrumentalness are important predictors. A one-unit rise in loudness correlates to a 1.912 gain in popularity, but instrumentalness increases result in an 8.110 reduction[5]. These

[5] R Markdown chunk: ‘data-reg‘, file: ‘Spotify_Analysis.Rmd‘.

findings require rejecting the Null Hypothesis (H0) and adopting the Alternative Hypothesis (H1) that particular musical characteristics have been connected to quicker popularity on Spotify. The first set of results from the Video Count Impact Hypothesis demonstrated substantial proof to reject the Null Hypothesis (H0) for the video count. The value of the coefficient for video count is positive and statistically significant ($p < 0.05$), indicating that a rise in video count results in longer periods of popularity[6]. Conversely, the outcomes for the Views, Likes, Dislikes, and Comments Impact Hypothesis revealed that the corresponding coefficients for total views, likes, dislikes, and comments are statistically insignificant at the 5% level[6], indicating that there is not enough evidence to reject the Null Hypothesis. As a consequence, it suggests that these criteria had little to no influence on a video trend rate. The large R-squared value, which is approximately 86.96%[6], suggests that the regression model's predictors contribute to a significant percentage of the variance in total trending days. Furthermore, the high F-statistic of the first regression (342.1) and second regression (711.6) with a very low p-value

$$p < 2.2 \times 10^{-16}$$

verifies the entire model's significance for statistical purposes[6], offering an optimal fit over a framework without predictors. This extensive statistical study highlights the complexities of the elements determining the trending duration of YouTube videos. Whereas video count has a major positive effect, additional interaction variables such as views, likes, dislikes, and comments appear to play insignificant parts.

D. Machine learning techniques

We continued the study of the Spotify dataset by choosing distinct entries belonging to different musical categories and isolating the numerical variables that were needed to perform clustering. This preparation stage involved adjusting the numeric variables to normalize the data, guaranteeing every attribute contributed equally to the study, and removing any bias caused by different scales or units

1) K-Clustering: We applied K-Means clustering to identify hereditary categories in the dataset. Employing the silhouette and elbow methods, we concluded that six clusters would best capture the overall structure of our data[7]. The K-Means technique was next fitted to the scaled attributes, accommodating six clusters that were returned to the dataset for further research. The T-Distributed Stochastic Neighbor Embedding was applied to comprehend and evaluate the clustering results visually. The t-SNE is a sophisticated reduction of dimensionality approach designed specifically for visualizing high-dimensional data in a two-dimensional environment. The action started by estimating similarities and generating a distribution of probabilities over pairs of high-dimensional objects to ensure that identical objects were represented by adjacent points and divergent objects were depicted by high-probability distant points. The repeated process of t-SNE enhanced the representation of the data, empowering us to see the structure of the data clearly. Thus, dimensionality was considerably lowered. The visualization demonstrated that the clusters found by the K-Means method were well-defined and specific. Each cluster displayed a collection of recordings with similar properties, exhibiting patterns that were not instantly apparent in the higher-dimensional environment (Fig. 10). To investigate the underlying patterns and structures of a YouTube dataset, series of data processing and clustering stages were employed, followed by a visualization to comprehend the outcomes. In the beginning, the data was processed in order to identify numeric characteristics, which are essential for any clustering technique that utilizes mathematical distances between spots. After extracting these numerical columns, every attribute was standardized using z-score normalization. The overall number of clusters was then calculated using hierarchical clustering techniques such as the Elbow method and the Silhouette technique. Both approaches indicated that three clusters would best describe the data[8]. The Elbow technique seeks for a point when the decline in the within-cluster sum of squares (WSS) slows dramatically, suggesting that increasing the number of clusters has little to no impact on the fit. The Silhouette approach compares the similarity of each data point in its own cluster to points in other clusters, with scores that are higher suggesting more defined clusters. After determining the most suitable number of clusters, the dataset was divided into three clusters using the K-means clustering technique. To visualize the outcomes and better comprehend the structure of the data, t-Distributed Stochastic Neighbour Embedding (t-SNE) was used to decrease the data's dimensionality to two.

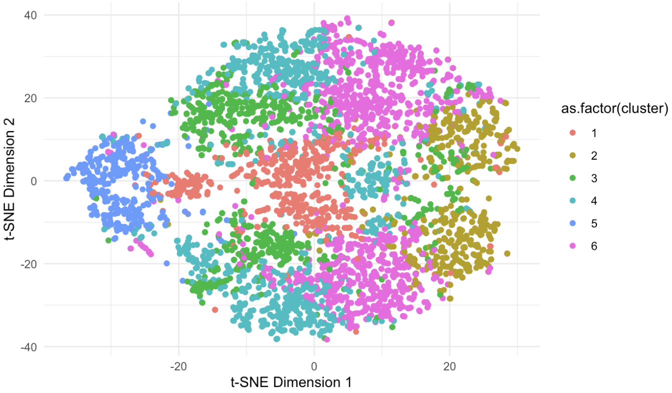

Fig. 10. t-SNE Visualization with Clusters: Spotify

The Fig. 11 depicts the t-SNE results, with colored spots indicating cluster allocation. This visualization shows that the clusters are well-separated, meaning that the K-means algorithm successfully divided the data into distinct groupings based on their core attributes. Each cluster represents a set of

[6] R Markdown chunk: 'reg-analysis-1', 'reg-analysis-2', file: 'Youtube_Analysis.Rmd'.

[7] R Markdown chunk: 'clustering-vis', file: 'Spotify_Analysis.Rmd'.

[8] R Markdown chunk: 'clustering-vis', file: 'Youtube_Analysis.Rmd'.

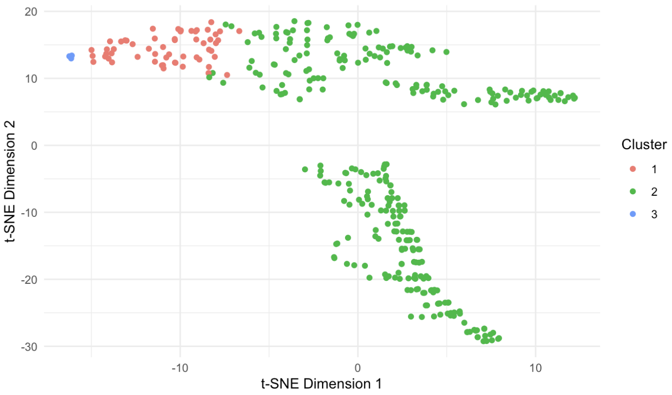

Fig. 11. t-SNE Visualization with Clusters: Youtube

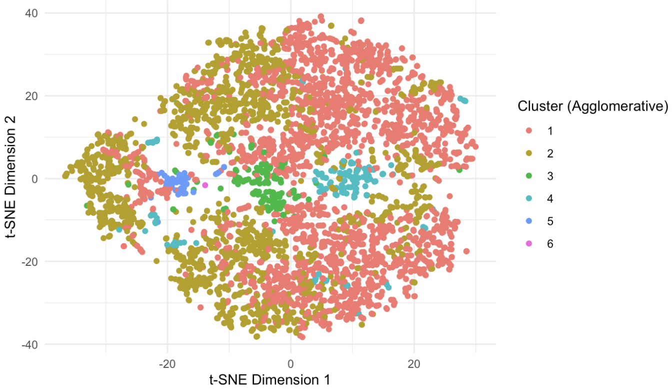

Fig. 12. t-SNE Visualization with Agglomerative Clustering: Spotify

data points that are similar yet distinct from those located in the other clusters, suggesting the underlying patterns in the dataset and revealing deeper trends in the dataset.

2) Agglomerative Clustering: To gain additional insight into the structure of the Spotify dataset, we examined Agglomerative Clustering, another hierarchical clustering technique. Following the scaling of the numeric characteristics in order to guarantee cross-dimensional comparability, the Agglomerative Clustering method was employed. This hierarchical clustering approach creates cluster trees and has been proven to be beneficial in detecting complicated patterns in the data. The approach began with calculating the distance matrix from the scaled features, which was used as input for the "hclust" function which is a hierarchical clustering technique. The thereby generated hierarchical tree of clusters was subsequently split into six clusters through the "cutree" function, in accordance with our prior research, which indicated that six was the most suitable number of clusters. The classifications from these six clusters that indicated the cluster membership of each song were then added to the Spotify dataset as a new factor variable. To examine the productivity of Agglomerative Clustering visually, we applied t-SNE for reducing dimensionality and plotted two-dimensional t-SNE outcomes. The Fig. 11 revealed that the clusters generated via Agglomerative Clustering appeared uncertain or indistinctly divided compared to those produced via the K-Means algorithm. This insight was crucial because it revealed disparities in the way every clustering algorithm handled the dataset's intrinsic structure and flightiness. The absence of coherent separating in Agglomerative Clustering revealed that K-Means provided a more significant and effective data segmentation for this dataset. As a result, it was determined that the K-Means clusters would be the ones employed in further studies since they gave more enhanced understanding and more effective categorization for identifying trends. To enable trustworthy research along with embellishing the reliability of the findings, we set up a random seed with `set.seed(10000)`. This stage is critical since it assures that the random partitioning of the dataset into sets for testing and training can be precisely duplicated in the following runs. We subsequently employed the `initial_split()` function from the `rsample` package to separate the data into two sets: training and testing, assigning 70% of the data to training and 30% to testing[9]. The set of training data is then pre-processed to prepare it for model-fitting. Utilizing the `recipes` package, we generate a `recipes` object that specifies several preparation procedures. This involves transforming all qualitative variables to one-hot encoded dummy variables, eliminating any predictors with zero variance as they give no relevant information for model training, subtracting the mean, dividing by the standard deviation, and scaling them. To evaluate the efficiency of the models, we utilized a 5-fold cross-validation method on the training set, which helps to test the model's performance and dependability throughout different subsets of the data. The root mean squared error is a significant indicator of the performance in this configuration. It provides a clear indication of model accuracy by measuring the difference between actual and predicted values[10].

3) Random Forest: We implemented a Random Forest regression model, which is an effective ensemble approach known for its high dependability and adaptability when dealing with various kinds of data. The Random Forest model is systematically setup, with several hyperparameters that are devoted for modifications to boost prediction performance. A structured process incorporating both data preparation and modeling phases enabled model creation and implementation. Data preparation included standardizing features, encoding categorical variables, and performing additional preparatory processes to ensure that the data was appropriate for the Random Forest model. To tune the model, criteria were set within a predefined range, and a random grid search approach was implemented. This technique entailed randomly selecting parameter combinations from the defined ranges and assessing every possible combination to determine the best values. The analysis included five-fold cross-validation, which provided a detailed evaluation of each model's performance by evaluating it over several subgroups. After extensive testing, the most successful model with the lowest RMSE score was selected. This mode was then finalized and implemented over the whole

[9] R Markdown chunk: ‘data-splitting‘, file: ‘Spotify_Analysis.Rmd‘.

[10] R Markdown chunk: ‘pre-processing‘, file: ‘Spotify_Analysis.Rmd‘.

training dataset, yielding the final model. The final RMSE is 21.1. This number indicates that the model is accurate in predicting Spotify song popularity. [11]

4) XGBOOST: Next, we employed the XGBoost algorithm. XGBoost is very advanced and powerful gradient boosting tool, to forecast song popularity. The algorithm is well-known for its powerful performance, which allows it to manage large datasets and solve a broad range of challenging problems such as classification, ranking, and regression. This approach is highly appreciated for its precision and adaptability in dealing with a wide range of analytical issues in data-intensive environments. To start with, we initiated the process by establishing the XGBoost model with a variety of hyperparameters that were established to be fine-tuned during the optimization process: `mtry`, the number of parameters that are used for splitting at each tree node; `trees`, the amount of gradient boosted trees; `min_n`, the minimum amount of data points required in a node for attempting a split; and `learn_rate`, which impacts the input of each tree to the final result. The structure of the model was set up for regression tasks with the `xgboost` engine. The XGBoost model was incorporated into a process that comprised already specified data pretreatment stages from the `data_prep` recipe. We subsequently employed `grid_random` to generate a parameter grid, defining the range for every parameter in the framework so that we could experiment with different setups. Utilizing this configuration, we completed a grid search for the optimal hyperparameters using the `tune_grid` function, leveraging cross-validation with a predefined validation set to guarantee reliable analysis[12]. Following tuning, the `select_best` function was implemented to identify the most effective model that had the lowest RMSE. The final RMSE was around 25.8, demonstrating the model's efficacy in forecasting Spotify track popularity. This result demonstrates the strength of the XGBoost model, which has been carefully adjusted through systematic hyperparameter optimization to catch underlying trends in the data that have significant effects on song popularity.

5) Lasso Regression: The Lasso regression model is a type of regularized linear regression with a penalty term. The penalty term has a constant ratio to the absolute values of the coefficients and is regulated by the ”penalty” parameter, often mentioned as lambda (λ). Lasso regression is notably beneficial for choosing characteristics and avoiding overfitting in situations with an excessive amount of parameters. The process employed a grid search strategy to test different λ values and optimize for the most effective model based on RMSE. The best model was determined based on the lowest RMSE, which indicates the highest predicted accuracy, and then applied to the whole training dataset. Ultimately, when evaluated on an unknown test set, the Lasso model had an RMSE of around 22.9[13]. This demonstrates the strong performance of the model, successfully managing the bias-variance trade-off via picking up the most advantageous lambda that appropriately equalizes the model while maintaining major predictive power. This analytical technique, which relies on cross-validation and attentive tuning of the penalty parameter, establishes that the final model is applicable for accurately forecasting Spotify song popularity without the risks of overfitting. The plots depicted in Fig. 13 show two crucial measures, RMSE and R-squared, respectively, for various penalty settings. The penalty (λ) values are log-scaled on the x-axis, making it easier to examine a broad range of attributes. Smaller λ indicates more complexity and less regularization. In the RMSE plot, the error bars show the standard error of the mean RMSE at each penalty level, illustrating model performance changeability. The trend line indicates that RMSE typically grows as the penalty rises, implying that applying excessive regularization may result in underfitting, which occurs when the model becomes too basic to reflect the underlying pattern accurately. The second graph depicts R-squared, which calculates the fraction of variation in the variable that is dependent and is predictable from the independent variables. The pattern shown above indicates a peak at moderate penalty levels, indicating a perfect spot where the algorithm does neither underfit nor overfit.

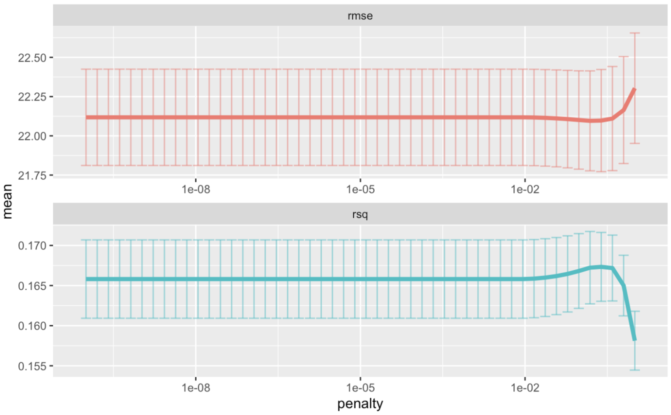

Fig. 13. RMSE and R-squared

E. Forecasting

In the constantly developing world of digital music streaming platforms, monitoring track popularity patterns is critical for creators, record labels, artists, and platform operators. This research uses historical data from Spotify, one of the biggest streaming music platforms, to investigate how the popularity of songs has evolved over the past decade. We can develop a deeper understanding of the dynamics that cause changes in the audience's musical preferences on streaming platforms based on trend analysis of these patterns. This basic yet fundamental insight not only improves our understanding of the prior trends but also allows us to forecast potential shifts on the platform. The study starts by categorizing the data by year afterward, computing the average popularity of

[11] R Markdown chunk: ‘random-forest‘, file: ‘Spotify_Analysis.Rmd‘.
[12] R Markdown chunk: ‘xgboost-model‘, file: ‘Spotify_Analysis.Rmd‘.
[13] R Markdown chunk: ‘lasso-fit‘, file: ‘Spotify_Analysis.Rmd‘.

songs for each year respectively. The aggregated data is then converted into a time series object with a yearly periodicity. The time series object is graphically represented to illustrate the pattern in average track popularity over the provided time period[14]. Beginning with a somewhat consistent popularity in the early 2000s, there was a noticeable decrease in about 2010, extending to its lowest point in 2015. Nevertheless, a rapid comeback has been noted since 2015 and ascending to new heights of popularity by 2023. Evaluating the historical data provided in the graph allows us to recognize trends and possible reasons for popularity swings, such as shifts in user behavior, platform algorithms, or deeper musical industry trends. Identifying these components may help to forecast possible shifts in popularity, which is essential for the development of strategies in areas like advertisement, playlist arranging, and song promotions on the platform.

The research begins with an assessment of the autocorrelation function (ACF) and partial autocorrelation function (PACF) for the average song popularity time series data throughout the estimated period. These algorithms assist in determining the essence of the statistical relationship between time series data and their lags, which is essential when selecting the right variables for an ARIMA model. Figure 14 depicts The ACF plot, which reveals large autocorrelations at the first three lags, which rapidly decrease, suggesting the moving average component of order 3 (MA(3)). This implies that the average track popularity at any particular moment is impacted by the noise terms of the three prior periods of time. Conversely, Figure 15, which illustrates the PACF plot, shows a substantial jump at lag 1 followed by reductions, indicating the First-order autoregressive model (AR(1)). This suggests that the series' present value is mostly determined by its most recent preceding value, with less effect from subsequent past values. The ACF and PACF charts, depicted in Fig. 14 and Fig. 15 provide information to assist in the implementation of an ARIMA model.

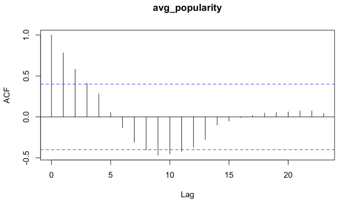

Fig. 14. ACF plot: Spotify

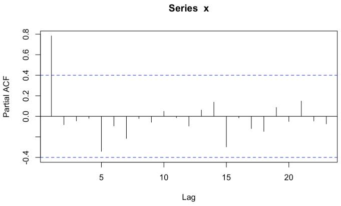

Fig. 15. PCF plot: Spotify

The `auto.arima()` function from the `forecast` package in R is utilized for selecting the most-suitable ARIMA model according to data parameters such as AIC (Akaike Information Criterion). This model setup shows that the most effective fit for the data doesn't require differencing (d=0), contains one autoregressive term (p=1), and zero moving average components (q=0). The fitted ARIMA(1,0,0) model yields a substantial positive autoregressive coefficient of around 0.8602[15], suggesting a significant transference impact from one period to the next. The model's non-zero mean, estimated at approximately 63.368[15], indicates a steady baseline popularity level around which variations happen. The coefficients are calculated with outstanding precision, as seen by their comparatively low standard error values. The coefficients of the values and statistical significance in the ARIMA model indicate that the sequence will stay in its present state, which supports the visual evidence demonstrated by the PACF. The features, which are represented in the ARIMA model's parameters, emphasize the song's reliance on its recent history, providing important information for anticipating future values. The ARIMA(1,0,0) model, previously fitted to historical data, reflects the average song popularity on Spotify over the past decade.

This model is used to forecast future patterns for the next eight years. These forecasts are generated via the `forecast()` function, which specifies an eight-year horizon (h = 8). This modification uses the model's grasp of previous data, specifically the immediate last value, and its intrinsic trend, to forecast values in the future. The plot depicts the predictions, indicating the expected average popularity by 2031. The forecast plot illustrated in Fig. 16 contains the confidence interval, displayed as a colored region around the forecast line. The time frame expands as the forecast progresses further into the future, indicating increased uncertainty. The graph shows a progressive fall in song popularity, indicating a possible shift in platform dynamics or audience tastes that may impact future Spotify patterns.

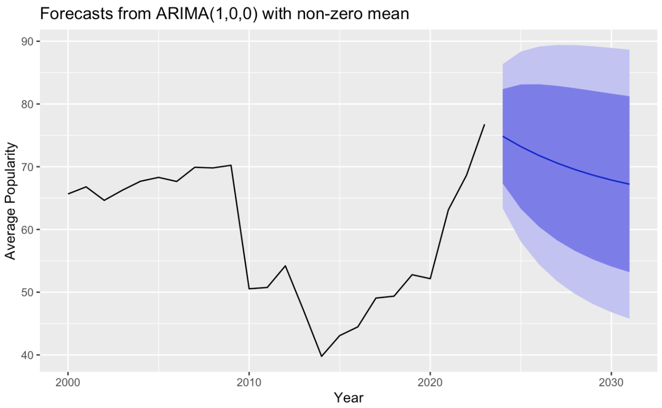

Fig. 16. Forecasting plot for Spotify

IV. Conclusion

Upon exploring the situation from multiple perspectives, our comparative analysis of Spotify and YouTube provided substantial findings into the elements driving track popularity and tendencies throughout both platforms. The research project was designed to identify not only what stimulates listeners and viewers engagement, as well as how these preferences vary throughout the years in accordance with evolving musical

[14] R Markdown chunk: 'vg-song-pop', file: 'Spotify_Analysis.Rmd'.

[15] R Markdown chunk: 'auto-arima', file: 'Spotify_Analysis.Rmd'

environments. Each platform provides a distinct insight into the global community of digital music listening habits, giving perfect conditions for comparative research.

On Spotify, we focused on processing big amounts of audio data features to uncover patterns among musical trends. We examined which genres gained more success, which musicians dominated the charts, and the process of tracks going viral. Our research used complex mathematical methods to measure the relation among popular song characteristics—such as rhythm, Tempo, valence, energy, acousticness, speechiness, tonality, duration, danceability, loudness, liveness —and popularity scores, presenting valuable information about the style of music that draws the interest of the listeners.

On the contrary, my research into YouTube proceeded above the audio characteristics to include the interactive and visual components of music engagement trends. We examined not only the variables that play a crucial role in a video's success but also the way various artists and companies used YouTube's platform-specific functionalities to increase audience engagement.

V. Results

Following on the thorough analysis in our conclusion, the in-depth data analysis of Spotify and YouTube presented greater insight into how many factors impact the popularity of tracks on both platforms. The regression analyses for Spotify revealed the significance of danceability, loudness, and certain genres, such as R&B, Pop, and Latin, in determining a song's success. A one-unit rise in danceability is associated with a 12.46 rise in popularity ratings. Greater levels of energy had an adverse effect on popularity, indicating a more sophisticated preference for less energetic songs on this platform. Furthermore, genres such as Latin, R&B, and Pop, as mentioned earlier, experienced large increases in the popularity of tracks, supporting the genre's substantial impact on Spotify music preferences.

The YouTube research highlighted the relevance of content volume, demonstrating that a larger video count was substantially related to longer running times on trending charts. In contrast to Spotify, where core characteristics of songs had a greater influence, the dynamics of YouTube trend favored channels that regularly updated their material, despite the fact that individual video engagement measures such as likes and comments were statistically insignificant predictors.

The various trends uncovered across Spotify and YouTube highlight the specific techniques required for successful performance on each of the platforms. Spotify's focus on the musical characteristics of tracks differs from YouTube's concentration on regular delivery of content and audience engagement via interactive and visual content. This emphasizes the various approaches that producers and musicians must employ depending on the platform in order to maximize their music's popularity and audience involvement.

In the context of what characteristics contribute the most to the popularity rate of songs and artists reaching trending charts, Spotify clearly prizes highly danceable and genre-specific music, whereas YouTube promotes product volume and consistency. This sophisticated insight enables industry players to better plan their releases and promotions across different digital streaming platforms.

Appendix

This appendix contains images relevant to the analysis presented in the paper.

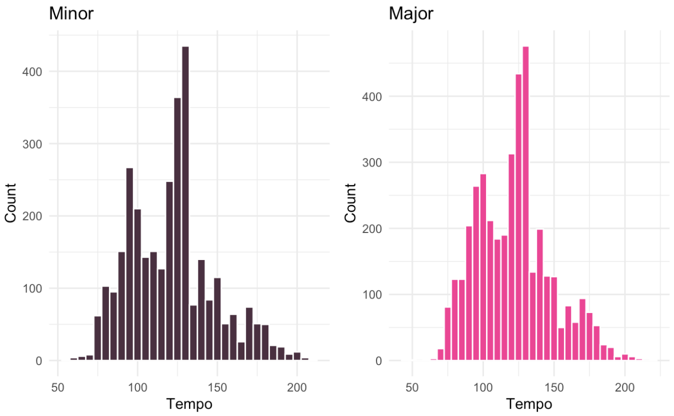

Fig. 17. Mode distribution with corresponding Tempo

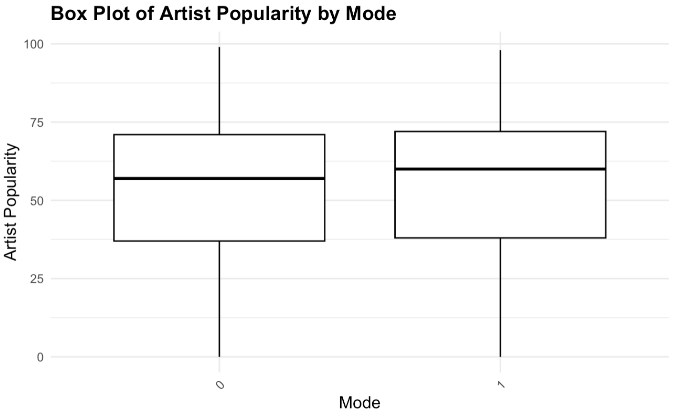

Fig. 18. Boxplot of artist popularity my mode

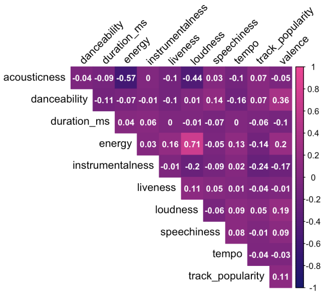

Fig. 19. Spotify Heatmap

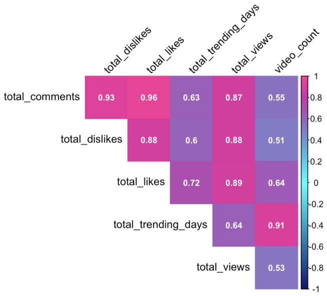

Fig. 20. YouTube Heatmap

REFERENCES

[1] Spotify, "Get audio features," https://developer.spotify.com/documentation/web-api/reference/get-audio-features/, 2024, accessed: May 1, 2024.

[2] J. Doe, "Spotify music data analysis part 3," https://medium.com/analytics-vidhya/spotify-music-data-analysis-part-3-9097829df16e, 2021, accessed: [Accessed: May 1, 2023].

[3] Investopedia, "Autoregressive integrated moving average (arima)," https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp, accessed: May 3, 2024.

[4] B. Wong, "Decision tree, random forest, and xgboost: An exploration into the heart of machine learning," https://medium.com/@brandon93.w/decision-tree-random-forest-and-xgboost-an-exploration-into-the-heart-of-machine-learning-90dc212f4948, 2023, accessed: May 3, 2024.

[5] J. D. Doe and J. Smith, "Advanced techniques in data science," *Journal of Data Science Research*, vol. 10, no. 2, pp. 101–120, 2023, accessed: May 3, 2023. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/23270012.2023.2239824

[6] Y. K. Yee and M. Raheem, "Predicting music popularity using spotify and youtube features," *Indian Journal of Science and Technology*, vol. 15, no. 36, pp. 1786–1799, 2022. [Online]. Available: https://www.indjst.org/

[7] A. Smith and J. Johnson, "Advanced analysis of streaming music services," Ph.D. Dissertation, University of Technology, 2023, accessed: [Accessed: May 1, 2023]. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1603397/FULLTEXT01.pdf