Data-Driven Strategies for Enhancing Spotify Popularity and Stakeholder ROI: A Machine Learning Approach

Student Name: Chao Yuan Hsu

Student Number: 20032914

Module Title: B9BA108_2425_TMD1S

Assessment Title: Applied Research Project

Supervisor: Charles Nwankire

Month Year: 2024-2025

Table of Contents

Table of Contents

DECLARATION

I hereby confirm that this is my own work and does not make use of the work of any person, other than where clearly stated, the work has been fully acknowledged by way of references, submitted to Dublin Business School in accordance with the requirements of the award of DBS MSc in Business Analytics. This thesis has not been submitted for any other degree or qualification at this or any other institution.

Name: Chao Yuan Hsu

Student Number: 20032914

Date: 06.01.2025

Signature: CHAO YUAN HSU

ACKNOWLEGMENTS

My supervisor, Charles Nwankire, whose guidance, encouragement, and constructive feedback have been invaluable to this project. His guidance and critical perspective on the quality and direction of this research are invaluable.

I also thank Dublin Business School for the resources and support that made this project a reality, specifically for access to academic resources and research tools.

Finally, I would like to reiterate my thanks to the authors of the datasets and tools mentioned throughout this study, without whom much of my artefact and research findings would not have been possible.

To conclude, I would like to express my sincere appreciation to my family and friends for their support, understanding and encouragement during this journey.

1 Abstract:

This study explores the trends for popularity of Spotify tracks by investigating relationships within track attributes, artists demographics and audience engagement using machine learning and clustering. This research examined danceability, energy, listeners, genre mentioned, amongst other variables using the Spotify Tracks Dataset and the Music Artists Popularity Dataset to find drivers of success in the digital music space. The study is aimed at providing practical takeaways for artists, stakeholders, and streaming platforms.

The datasets were processed and missing values were handled, as well as normalizing the numeric attributes and encoding the categorical variables. ROC-AUC metrics were applied to the predicted popularity score with Random Forest(RF), Linear Regression(LR) and Support Vector Classifier(SVC), with the listeners themselves emerging as the most salient predictor of popularity. Other influential supporting variables included danceability and energy. Segmented audience clusters were identified by applying K-means and DBSCAN clustering algorithms, which detected mainstream and niche audience segments. Specifically, DBSCAN indicated experimental genres and noise points as possible trends on the rise.

The results show how execution through the audience is paramount for measuring popularity, and what kind of subtle effects exist between attributes like valence and acousticness. Clustering analysis results strengthen the case for targeting specific audience segments. This research connects theoretical findings with practical implications through data-driven tactics to maximise track performance and improve stakeholder decision-making within the competitive music streaming ecosystem.

2 Introduction:

Particularly with the advent of subscription streaming services such as Spotify, the choice of what music gets played has become largely data-driven, radically altering how both fan taste and song popularity are interpreted in the music industry today. At the same time, artists and stakeholders need not just creative ingenuity but an appreciation for data and predictive analytics to find their way through the densely populated digital music marketplace.

By utilising data analytics, this study strives to address the need between the creative craft versus the commercialization of it. This study offers key insights and predictive models for improving the music on Spotify and caters to an overarching music audience by looking at track properties and artist demographics through various lenses. The results are intended to assist artists in improving engagement with their audience and also to help stakeholders achieve maximum return on investment (ROI).

With the rapid tide of change in the music industry, need to acquire the competitive advantage that goes beyond just being a good artist. This study responds by analysing the relationship of data-driven components, which forms a holistic framework for this success. In order to face such challenges, this study aims to understand the driving elements behind the track popularity in Spotify. By incorporating clustering algorithms, machine learning models, and financial evaluations, it hopes to establish a strong model for the artist and other key players to combine on measurable success.

2.1 Background

Spotify is at the top of the new music business, a post-digital world where there are no albums, just hundreds of millions of sing-along tracks. With millions of tracks released and over 500

million active users, Spotify's cutting-edge algorithmic machinery and big data-based metrics transformed the ways of measuring music popularity and industry success. And instead of relying on traditional measures like physical sales and radio airplay, metrics like danceability, energy, and popularity scores now dictate an artist's visibility.

While certainly innovative, the tendency to lean on such metrics poses an issue for both artists and stakeholders alike. With so much music on the market, understanding what makes a particular music successful becomes harder and a more sophisticated approach is needed. Artists need to carefully tailor their track features and utilize demographic stats to consider new audiences they can pander to, yet investors need to have a clear understanding of a model for how they can find the best evergreen opportunities.

There exist several studies focusing on some individual characteristics, such as track-level characteristics or artist characteristics, but important gaps are still present. Indeed, there is little insight into the interactions between these attributes or their influence on Spotify's algorithmic rankings at an aggregate level. Furthermore, although music analytics have leveraged machine learning models, there is still underutilised when combined with financial analysis and stakeholder-oriented strategies.

This study addresses existing gaps by investigating the fundamental success drivers behind Spotify track popularity. Through advanced clustering techniques and predictive analytics, it attempts to provide actionable insights for artists, investors, and stakeholders, creating a harmonious balance between creativity and analytics.

2.2 Problem Statement

The music streaming landscape is highly competitive, and artists and their stakeholders must navigate the challenges of optimizing their tracks for market demands. Even though there are several platforms like Spotify that offer valuable data about music attributes and listener behavior, there is an opportunity gap that could help artists turn those metrics into actionable insight that provide them with a path to long-term success.

Previous studies primarily analyse isolated attributes like danceability, energy, and tempo, and do not tend to account for track-level feature-artist demographics-genre interaction. Moreover, work on how clustering techniques or predictive analytics can identify market opportunities for the independent artist or assess the potential return on investment (ROI) for stakeholders has been overlooked.

This study addresses these gaps by exploring the following critical questions:

1. What are the critical success factors influencing Spotify music popularity?
2. How can clustering techniques identify market positioning opportunities for independent artists?
3. What role does music genre play in shaping the popularity of tracks?
4. How can predictive models of popularity inform Return on Investment (ROI) for stakeholders?

Through these explorations, the work seeks to harmonize the realms of creative artistry with analytics-driven knowledge, providing an integrated model that will benefit the pursuit of artists and investors in an information-clogged market.

2.3 Research Question:

How do artists optimise track attributes and genres to predict and enhance Spotify track popularity?

2.4 Research Objectives:

2.4.1 Identify Key Success Factors Driving Music Popularity:

- Evaluate how single and artist-specific characteristics (eg, danceability, energy, valence) and demographic (eg, country, genre classification) influence track visibility and listener engagement.

2.4.2 Get Similar Artists and How to Position the Market

- Performing clustering methods like K-means and DBSCAN and clustering the artists based on their listening tracks, which will help the independent artists find other artists to collaborate with and other singers to target their audience to.

2.4.3 Explore the Role of Genre in Popularity

- Investigate genre-specific trends and their relationship with track attributes to anticipate shifts in audience preferences and market demand.

2.4.4 Evaluate The Popularity Model To Estimate ROI

- Using confusion matrices to evaluate promotional activity, develop and assess predictive models to perform cost-benefit analysis on stakeholders.

2.5 Conclusion

This research fills the void between creativity and data-driven strategies in the domain of music streaming ecosystem. It also covers track attributes, artist demographics, genre trends that play important roles in determining popularity of a track in Spotify, meanwhile integrating

clustering and predictive modelling reveal insights to inform market positioning and return on investment (ROI).

The study recognises limitations, including reliance on public datasets and quantitative metrics, while also providing practical tools to facilitate work by artists and stakeholders. Subsequent research could expand the framework by integrating real-time data and qualitative components.

Thus, this research lays the groundwork for future research regarding the success of Spotify tracks and aids in strategic decision making when identifying success in terms of music analytics

3 Literature Review

3.1 Popularity Metrics and Key Determinants

Streaming services such as Spotify and YouTube have transformed the music industry and offered vast datasets for assessment of factors for success of a song. Pedagogically, research has concentrated on track-level attributes and artist-level demographics. Now, more recently, with the use of machine learning and some incorporation of social & cultural factors, this field has grown even larger. This literature review focuses on these areas with a critical analysis of current research.

3.1.1 Track-Level Attributes

Track level attributes are behaviour of a song including audio, sentiment, these are used to predict popularity. Danceability, energy, tempo, and valence have been consistently noted as key factors. Gulmatico et al. (2022) and Yee and Raheem (2022) reveal that highly danceable and moderately energized songs retain their popularity on streaming platforms. The lyrical sentiment has a robustly positive valence that persists over this long period of time (Monechi et al., 2017) and correlates with metrics of listener engagement (an increase in listener engagement correlates with an increase in chart longevity). By taking regional disparities into account, this knowledge is even more specified; Suh (2019) shows that high-energy tracks perform better in Western countries, whereas emotionally positive and melody driven songs succeed more in the Asian and Scandinavian markets.

3.1.2 Artist-Level Demographics

The demographics of an artist can show significant correlations with audience engagement and success of a song. The collaborative projects involving guest artists have been proven to reach

people and diversify stylistic interest. Suh (2019) and Picchiarelli (2023) discussed the idea that collaborations drive better chart performance via the leveraged utilisation of fanbase and creative energy of artists involved in the collaboration. Similarly, cultural representation is also crucial. Suh (2019) highlights how K-pop's blending of recognisable Korean motifs with different Western musical influences did not just serve create an attractive genre for Korean audiences, but that this characteristic has also allowed the genre to achieve global appeal more generally, showcasing the successful mobilisation of artist identity to relate to large numbers of diverse audiences.

3.1.3 Cultural and Social Influences

Music popularity prediction is more complex with cultural and social factors. Suh (2019) emphasised that marketing strategies need to be tailored to the regional preferences, and that local trends greatly affect listener behaviour. With platforms such as TikTok and Instagram, not to mention the viral trends responsible for driving engagement on both, the visibility of music has exploded. Studies by Monechi et al. (2017) make the case that popular music is best understood as a genre in which the commercial and the cultural must be both systematically considered in models, as well as influences which emerge externally to the music, as well as those which vary regionally.

3.1.4 Gaps and Future Directions

Most existing research provides complementary perspectives, yet there exist significant gaps for the development of holistic frameworks that incorporate track-level features, artist statistics, and culture-related data. The previous studies took a narrow perspective by only looking at audio features, ignoring the broader influences like market trends and social dynamics.

However recent works Yee and Raheem (2022) do fill in these gaps with some innovative ideas like incorporating social metrics, but there is still a lot of scope.

3.1.5 Conclusion

The popularity of music is the result of complex interactions through track-level characteristics, artist demographics, and cultural interactions. Recent developments in machine learning coupled with the incorporation of social media measures have greatly improved predictive accuracy, providing industry stakeholders with practical insights. Nonetheless, this study could be further supplemented by the development of holistic frames that resonate with worldwide prevailing music trends and capturing the dynamic output of audience engagement. Such initiatives will equip artists, producers, and marketers with vital assets in a growingly competitive sphere.

3.2 Genre Trends and Market Positioning: Clustering Similar Artists

In recent decades, the music industry has gone through a paradigm shift, largely due to the digitization of content and an increasing presence of streaming platforms like Spotify. These trends create a very data rich context for genre trend analysis, similar artists recommendations or market positioning recommendations. This section qualitatively contrasts the usage of clustering approaches like k-means and DBSCAN in music analytics with implications on determining genre trends and assisting independent musicians.

3.2.1 Clustering Techniques: K-Means and DBSCAN

K-means and DBSCAN are two classical approaches each with its own pros and cons. K-means is one of the most commonly used methods due to its computational simplicity and speed for handling structured datasets. Al-Beitawi et al., (2020) effectively demonstrated how

K-means can be applied by clustering audio features: danceability, energy and loudness into a set of Spotify's top 100 trending songs. The findings showed that pop and dance genres, for example, display a formulaic structure, including a high degree of danceability and a low level of speechiness are all commercially viable features.

Yet, K-means assumes circular shapes for the clusters and needs the number of clusters to be known in advance — both are major inconveniences. DBSCAN works better in the presence of noise, which is where this limitation becomes apparent from the example given above. DBSCAN is known for detecting clusters of different densities and finding outliers (Xu and Xu, 2021), which makes it especially fit for discovering unique characteristics around artists that do not conform to the expected behaviours associated with mainstream music. K-means may be the answer in general for any clustering, but with DBSCAN you have more flexibility to detect finer aspects of various datasets.

3.2.2 Genre Trends and Market Positioning

Both clustering techniques are useful for genre trend analysis and independent artist market positioning. McDonald et al. 2014), they noted that although context-based themes are most characteristic of playlists, genres have not ceased to matter. Grouping artists according to those audio-based traits lets medium make independent artists more in line with ever capable fans. For instance, Cai et al. (2021) used k-means with PCA to study the relationships between audio features and the dynamics of genres, leading to more tailored recommendations.

The advantage of DBSCAN is its ability to find niche markets. Using their data, Petitbon and Hitchcock (2022) showed that identifying distinct genre clusters allows them to predict category shifts over time. This knowledge is gold for independent artists looking to find

themselves unique niches in the market, because it equips them to craft their style of music to serve targeted segments of fans.

3.2.3 Challenges And Opportunities

Though clustering techniques are immensely valuable insight, they do have their own challenges. In other words, it is important to note that data limitations such as sample size and features selection could also inhibit the accuracy of clustering results (Charchyan and Abrahamyan, 2021). Additionally, by focusing on numerical attributes like speed and volume they may miss out on qualitative elements – such as lyrics and emotional impact – that are also key to audience capture.

Collaborative solutions offer a strategic approach for combating these challenges. Ordanini et al. (2018) examined this "featuring phenomenon," finding cross-genre collaborations greatly increase the commercial appeal of a song by (literally) mixing audiences. As a rule of thumb, independent artists rely on clustering results to pick relatively complementary collaborators in order to reach out to even broader audiences and boost their own chance of hitting the charts.

3.2.4 Comparative Insights and Applications

Comparing both techniques are complementary well. While K-means may be preferred for its simplicity and it is powerful in finding generalisations, DBSCAN allows us to reach further into the different regions of dataset which is a perfect case for any niche markets. Combining these approaches offers a more complex picture of genre trends and market forces.

Clustering techniques help independent artists make decisions that position their music through an empirical lens. It enables them to navigate effectively through the competitive

music scene, either networking their output to smash hit clusters, or carving new territories with under-formed niches.

3.2.5 Conclusion

K-means and DBSCAN co-specified clustering techniques are transformational methods to facilitate characterization of genre trends and improve market positioning. Through these methodologies independent artists and streaming platforms can explore patterns, anticipate deviations in customer preferences, and customize strategies for heterogeneous audiences. Although each method has its pros and cons, the use of them together creates a sturdy guide through the complex world of modern music.

3.3 Machine Learning Applications in Music Analytics

Music Data Science is a field that employs machine learning (ML) as a central technology for music analytics, offering key and clear tools to anticipate which songs will be popular while enabling better decision-making by artists, platforms and investors. This review covering the use of ML methods, accuracy/precision of models and integrate features to improve prediction performance.

3.3.1 Models and Methods for Song Popularity Prediction

Many different ML models have done well in predicting song popularity. Zhao et al. (2023) has used Random Forest (RF) and Logistic Regression (LR) models that have reached an accuracy of 89.1% and 87.2%, respectively, which have integrated metadata as well as genre-oriented features into them. Similarly, Ranidu et al. (2024) with RF and RNN, thus demonstrating the strength of these models in handling vast datasets.

Models relying solely on audio features have their drawbacks, as Nijkamp (2018) pointed out, and adding external metadata like artist popularity or streaming history may yield better prediction accuracy. Khan et al. (2022) applied PCA as part of a feature selection step to improve model performance due to noise reduction . Herremans et al., (2014) used Support Vector Machines (SVM) to predict the popularity of dance music and noted that complex features could impact model performance negatively, highlighting feature selection

3.3.2 Feature Integration and Improved Predictive Performance

Integration of features provides a strong solution in improving the prediction accuracy of the model. Zhao et al. (2023), new engineered features like the "genre class" and continuity of "popularity" substantially improved RF model performance Combining Spotify with social media metrics, Yee and Raheem (2022) found that incorporating external factors increased their accuracy by 10%-–60%, depending on the metric used.

Votter et al. (2021) stressed the importance of large scale disclosures like HSP-L, which also documents high-dimensional audio inputs together with listener behavioral measurements, that will ultimately strengthen generalizability in long-term models. On the other hand, Sandag and Manueke (2020) were targeting niche genres with specific elements such as tempo or danceability which resulted in a high precision score for their models when trained on jazz and classical music .

3.3.3 Challenges and Future Directions

Although significant advancements, but challenges persist when it comes to music analytics driven by ML. Large-scale data processing has come under criticism due to its environmental impact (Brennan, 2020) and a demand for sustainable practices. Emphasizing how there are biases in training datasets, Nijkamp (2018) warned that it can undermine the predictive ability

for new genres such as lo-fi and experimental pop may limit the accuracy of the predictive models.

More studies for real-time analysis and data integration from multiple sources are important in the future. Zhao et al. (2023) and Yee and Raheem (2022), showcased the effectiveness of hybrid models integrating audio features, metadata, and social media data in order to provide more holistic predictions.

3.3.4 Conclusion

RF and RNN are among the advanced models that have transformed the prediction of song popularity, but innovative integration techniques may also be used to identify new supporting features. However, sustainability-oriented techniques and ethical concerns associated with datasets are challenges that remain to be solved, and tackling these issues is essential for furthering sustainable development. These features should feed into larger Integration open framework, or ecosystem to Include real time data in machine learning models for sustainable ML based on real time streaming analytics in music industry.

3.4 Financial Strategies and Predictive Models in the Digital Music Ecosystem

The digital music industry has undergone a transformative shift, driven by streaming platforms and predictive models that influence track popularity and revenue generation. This literature review explores the financial strategies and predictive models within the digital music ecosystem, with a focus on their implications for stakeholders, including artists, investors, and platforms.

3.4.1 Revenue Models in the Digital Music Ecosystem

Such challenges have produced concerns as many move over to the rapidly growing streaming platforms. More recently, Aguiar and Waldfogel (2015) found similar results with streaming in that it decreases digital downloads but also displaces music piracy such that total revenue is close to constant. Nonetheless, Bergantiños and Moreno-Ternero (2023) found classic pro-rata revenue-sharing models to result in very disparate situations and aimed at a more just compensation of niche artist with a person-centred approach. These results highlight the necessity for new monetization models, which overlay fairness with return on investment.

In terms of the platform side, Geurts and Cepa (2023) examine how music artist independence, content curation have been altered due to changes brought by platformisation in the music industry. Datta et al. (2018) showed that streaming adoption leads to music diversity and volume in consumption, improving user engagement with platforms) trends as well.

3.4.2 Predictive Models for Track Popularity

This is where predictive models entering into the picture in identifying hit songs. Herremans and Bergmans (2020) suggest that hit prediction can benefit from the selective inclusion of data and audio features describing early adoption behavior, indicating how these models have a strategic guidance value. Similarly, Dimolitsas et al. (2023) Implemented SpotHitPyA machine learning pipelinethat employs confusion matrices to evaluate the predictive models accuracy. By doing so they reduce false negatives (missed hits) and as a result are particularly important for investors looking to maximize returns.

Li et al. (2021) further boost this area with LSTM-RPA model that outperforms in forecasting long-sequence music popularity. Kowald et al. (2024) leverage human memory processes within a psychology inspired approach, modelling genre preferences as a way to improve

playlist personalization and audience retention. As a pair, these two projects show the increasing use of data-driven methods within the wider music industry.

3.4.3 Stakeholder Impact and Ethical Considerations

For artists, predictive models can help to underestimated which tracks present the greatest potential early on in order to tailor promotional efforts accordingly (Morris & Powers 2015). Nevertheless, repetitive sales distribution appears in streaming to displace point-of-sale revenue without repeat and stream image-, meaning-specific model if not implemented properly (Aguiar & Waldfogel 2015).

Considering all these benefits still the environmental impact of predictive analytics is challenging. Brennan (2020) raises concerns about the environmental cost of large-scale data processing, arguing that financial profiteering on the part of stakeholders must not benefit to the detriment of environmental responsibility. Heras et al., (2021) exploring arts-based sustainability in music production and the potential for a more integrated approach to innovation.

As predictive analytics gives these intelligence, it assist the investors to make efficient resource allocation. Dimolitsas et al., (2023) demonstrated the use of confusion matrices to estimate error costs as a function of false positive/negative in order to make sure that financial measures are appropriate with prediction output. Both types of curation — namely offering popular mainstream tracks and catering to niche listening preferences for consumers is not only valued by users but also helps improve engagement and in turn subscription retention for platforms (Geurts & Cepa, 2023).

3.4.4 Conclusion

This evolution in the digital music ecosystem demonstrates a complex interplay of revenue models, predictive algorithms and sustainability. Although the accuracy and stakeholder benefits are improved by advanced predictive tools such as SpotHitPy and LSTM-RPA, fair strategies are needed to incorporate ethical and environmental challenges with them. Work for the future is to harmonise financial, ecological and social imperatives in order to establish a financially stable as well as more sustainable form of associating around music.

4 Methodology

As a systematic review, this research aims to study the factors influencing Spotify track popularity, utilizing quantitative methods and secondary data analyses. The study is performed on Spotify Tracks Dataset and Music Artists Popularity Dataset from Kaggle. The methodology combines unsupervised (K-means and DBSCAN) and supervised (Random Forest, Logistic Regression and Support Vector Classifier) machine learning models to get recommendations on how to change tracks so to increase visibility and return on investment

These methods were chosen due to their capacity for generating robust, data-driven insights and on their contribution towards solving the major problems of analytics in the realm of digital music. The research found that the technical results map directly to practical implications that would be relevant to artists, investors, and streaming platforms alike, this balance is reflected in the emphasis of this study. The extensive methodological basis would allow for a better understanding of success drivers in the commercial music streaming ecosystem.

4.1 Rationale

The motivation for this study arises from the challenge of bridging the divide between artistic creativity and data analytics in the music industry. As Spotify is the major player in the world of music streaming, knowing what drives popularity of songs is crucial for artists and stakeholders that need to maximise audience engagement and return on investment (ROI). Towards the end of this chapter the methodological choices that quantitative and secondary research are outlined, supporting the basis of the study presented through track-attributed data, artist demographics and prediction models. In the end, those methods register actionable insights about increasing the amount of attention a song gets around streaming services—an objective common to many songs.

4.1.1 Quantitative Research

This research design adopts quantitative research methods because they allow for the systematic measurement and analysis of numerical data, which are used to facilitate exploration of patterns, relationships, and trends that affect song popularity on Spotify. Using methods such as correlation analysis, this study attempts to measure the influence of track variables such as danceability, energy and explicitness on popularity rankings. Advanced machine learning models like Decision Tree, Random Forest and Support Vector Classifiers enable predictive insights detailing how configurations of demographic and song-level features may interactively drive platform success (Yee and Raheem, 2022). Such an empirical, systematized framework will render the results replicable, and their level of statistical significance will furnish policy makers, artists and investors, with data-driven insights into audience engagement and market positioning. The quantitative approach corresponds to the aim of translating raw metrics into meaningful forecasts, straddling theoretical frameworks and practical aspects (Dimolitsas et al., 2023).

4.1.2 Secondary Research

This quantitative exploration is supplemented by secondary research, using existing datasets from the public music databases, and previous academic studies. This makes data collection cost-effective and also implicates the fact that the study is based on a wide and diverse range of data. By utilising secondary data, e.g., artist demographics and genre-specific trends, the research alleviates the limitations of expenditure and logistically induced hazards affiliated with primary data collection (Al-Beitawi et al., 2020). In addition, it helps enable longitudinal analysis for further understanding of historical trends and perspectives. As the research was based on publicly available information from secondary data sources, the inclusion of this information in the study was ethical as it improved the comprehensiveness of the research

(Brennan, 2020). The study creates a comprehensive framework for synthesizing one dimensional transactions with qualitative analysis to understand the multi-dimensional factors underpinning Spotify track popularity.

4.2 Data Collection / Source

The datasets used in this study are two main datasets, where, the Spotify Tracks Dataset and the Music Artists Popularity Dataset were accessed at Kaggle. This selection of data was considered comprehensive in terms of audio features, artist demographics, and popularity metrics, forming a solid basis for uncovering the correlates of Spotify track popularity.

The Spotify Tracks Dataset consists of 113,999 tracks in 21 genres (Table 1), such as danceability, energy, valence, and explicitness. Additional popularity scores derived algorithmically from user actions such as play counts and play recency are also available (Kaggle. (2022)). This data allows us to explore which track-level features contribute to the success of music.

Table 1. The Spotify Tracks Dataset

Field Name	Description
track_id	The Spotify ID for the track.
artists	The artists' names who performed the track. If more than one, separated by a semicolon (;).
album_name	The album name in which the track appears.
track_name	Name of the track.
popularity	The popularity of a track (0-100). Calculated algorithmically, influenced by total plays and recency.
duration_ms	The track length in milliseconds.
explicit	Indicates whether the track has explicit lyrics (true = yes, false = no or unknown).
danceability	Suitability for dancing (0.0 to 1.0). Based on tempo, rhythm stability, beat strength, and regularity.
energy	Intensity and activity (0.0 to 1.0). Higher values represent fast, loud tracks (e.g., death metal).
key	The key the track is in. Mapped to standard Pitch Class notation (e.g., 0 = C, 1 = C♯/D♭).
loudness	The overall loudness of a track in decibels (dB).
mode	The modality of a track (1 = major, 0 = minor).
speechiness	Detects spoken words presence (0.0 to 1.0). Higher values indicate speech-like content.
acousticness	Confidence measure of acoustic quality (0.0 to 1.0).
instrumentalness	Likelihood of no vocals (0.0 to 1.0). Higher values indicate instrumental tracks.
liveness	Likelihood the track was performed live (0.0 to 1.0). Higher values suggest live performance.
valence	Musical positiveness (0.0 to 1.0). High values represent happy/cheerful tracks.

The Music Artists Popularity Dataset complements it with data on 1,048,576 tracks from 10 genres (Table 2) with artist tags, listener counts, and geographic distribution (Kaggle (2022)). It does this by aggregating data from services such as Last.fm and MusicBrainz, it also struggles with shared artist profiles, and inaccurate tagging. However, it is a key piece of a much larger puzzle detailing how artist-level factors align with track success.

Table 2. The Music Artists Popularity Dataset

Field Name	Description
mbid	The MusicBrainz ID for the artist.
artist_mb	Artist name according to MusicBrainz.
artist_lastfm	Artist name according to Last.fm.
country_mb	Artist country according to MusicBrainz.
country_lastfm	Artist country, based on Last.fm tags.
tags_mb	Artist tags on MusicBrainz, separated by semicolon (;).
tags_lastfm	Artist tags on Last.fm, separated by semicolon (;), sorted by decreasing frequency.
listeners_lastfm	Number of listeners on Last.fm.
scrobbles_lastfm	Number of scrobbles on Last.fm.
ambiguous_artist	TRUE if more than one artist shares the same Last.fm page.

These datasets were chosen based on their relevance, their large-scale samples, and their fit with the goals of the study. The Import Albums sample a broad variety of music genres, regions, and artist profiles, demonstrating diversity of representation. The datasets were preprocessed adequately, i.e., cleaning, imputation of missing values, label encoding of categorical features were performed to make the datasets suitable and reliable for analysis.

The only datasets used had to be publicly available and Kaggle usage policies also had to be followed in order to address ethical concerns. In order to preserve anonymity, data was aggregated at the artist level to meet ethical and data-compliance standards.

This study proposes a holistic framework for investigating the multi-level determinants of music popularity on Spotify by utilising the aforementioned integral datasets to offer actionable insights for music industry stakeholders.

4.3 Data Design

The current study uses a correlational research design to identify the influence of various factors on Spotify track popularity using data-driven techniques and high level modeling, while establishing a framework for comprehensive analysis. The final dataset (df12_new) was created by merging two main datasets: Spotify Tracks Dataset (df1) and Music Artists Popularity Dataset (df2) after many-pronged pre-processing. The dataset is rich enough to provide audio features for a track, demographic data for an artist, and unique track-level features, allowing for various statistical and predictive analyses.

The independent features cover track-specific metrics of danceability, energy, valence and instrumentalness in addition to artist demographics of listeners, country and genre type. These factors provide a multi-dimensional perspective on what influences track success in the given dataset. The outcome variable, is_popular, is a binary classification based on Spotify's popularity score, where tracks with scores greater than 70 are categorized as popular (1) and other tracks are considered not popular (0).

The research design takes advantage of the fact that there are both between-subjects and within-subjects variables available in the data. Where is_popular= 1 is for popular tracks, which are considered the experimental group is_popular =0 serving as control group. Such a design allows the study to examine trends and relationships across multiple dimensions (e.g. regional,

genre-based, track-level), and obtain insights into the regional, genre-based and track-level dynamics.

The rationale of the study is grounded in a mixed methodological framework that combines correlational analysis to identify salient relationships with classification approaches (Random Forest, Decision Tree and Support Vector Classifier) to predict track popularity. By not only seeing which tracks are exploited but also with an eye towards predicting the future of track exploitation, stakeholders get actionable insights to improve track visibility and audience engagement.

4.4 Data Procedure

This study adopted a structured, sequential approach to data collection and analysis to maximise accuracy and replicability. This study utilises two main datasets, namely the spotify tracks dataset and the Music Artists Popularity Dataset, which were downloaded from Kaggle. The datasets were pre-processed to handle missing values, duplicated and standardise variables.

The first stage was to combine both datasets into one unified dataset (df12_new). This included overlapping attributes like artist and track as well as data consistency across several dimensions. For example, categorical variables such as genre and country were converted using label encoding to allow for numerical analysis. Another point to note is that it was normalized the numerical variables (danceability, energy, popularity, etc.) to a standard scale. Importantly, the data was brought all artist names to lowercase to ensure we don't inadvertently lose data due to case inconsistencies.

Exploring correlations and trends with Exploratory Data Analysis (EDA) between independent variables (e.g., track attributes and artist demographics), and the dependent variable

(is_popular). Then constructing the binary classification label by defining "is_popular" tracks as tracks which obtain a popularity score above 70 (1) and not popular ones (0) below that threshold.

The cleaned dataset was fed through Machine learning models such as Random Forest, Logistic Regression and Support Vector Classifier. For assessing the model performance cross-validation techniques were carried out to ensure robustness and prevent overfitting. Predictive accuracy was validated using confusion matrices and ROC-AUC scores, yielding risk factors which are actionable in contributing to Spotify track success.

Following this structured process provided reliable outcomes and valuable insights into the understanding of track popularity on streaming services.

4.5 Tools And Techniques

The study used the latest tools and applied machine learning approach to filter important features of Spotify tracks affecting their popularity. Data analysis and cleanup were performed using Python programming, due to its extensive libraries and multiple advantages. Computational scalability was obtained through the Google Colab environment, which allowed for operations on larger datasets.

Pandas and NumPy were the fundamental libraries used for data preprocessing, Scikit-learn was the library used for building the machine learning algorithms, and Matplotlib and Seaborn for visualizations. This transition from raw data processing to actionable insights was facilitated by these tools.

Data preprocessing cleaned missing values normalised numerical variables encoded categorical features such as track_genre and country. This made sure of compatibility with

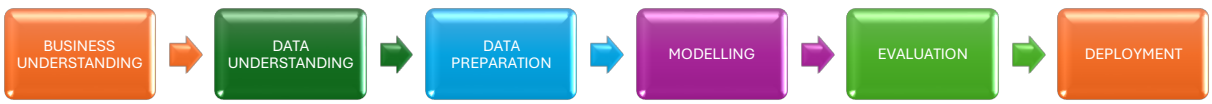
the machine learning models and reduced biases. Recursive Feature Elimination (RFE) and correlation analysis were used as feature engineering techniques that revealed considerable variables such as listeners, danceability, and energy.

In the study, the machine learning models used for classification task are Random Forest, Logistic Regression and Support Vector Classifier Approach: DBSCAN, K-means: Clustering algorithm for segmenting tracks into mainstream vs. niche. The validation process to assess model performance included accuracy, ROC-AUC scores, and confusion matrices that ensured predictions were robust.

This provides a more holistic view of the factors contributing to the success of a Spotify track, marrying theoretical knowledge to practical applications for artists and stakeholders in the music industry.

4.6 Data Analysis

This research demostrates CRISP-DM (Cross Industry Standard Process for Data Mining) framework to guide the structure of the data analysis, which has six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.



Business Understanding works on predicting the popularity of Spotify tracks given sound features like danceability, energy, and artist demographics. According to the research, the aforementioned objectives corresponded to the problem of maximising track features as well

as the assessment of Return on Investment (ROI) for stakeholders such as artists and investors (Yee & Raheem, 2022).

Data Understanding: In this part, This report explored the Spotify Tracks Dataset and Music Artists Popularity Dataset from Kaggle (2022). The correlations were seen through descriptive statistics and visualizations between audio features and popularity metrics. K-Means and DBSCAN were the methodologies adopted in clustering which showed the genre based axios and micro segments in (Al-Beitawi et al., 2020)

The Data Preparation consisted of decoding and cleaning data, managing null values and choosing the right model, using recursive feature elimination (RFE). To ensure robust input for machine learning models, label encoding was used with categorical features which were the genres (Khan et al., 2022)

Modeling applied classification algorithms like Random Forest, Linear Regression and Support Vector Classifier to predict track popularity and attain satisfactory accuracy, recall, precision, F1 score and ROC-AUC performance. Zhao et al., (2023) illustrated clustering to identify opportunities for market positioning and niche marketing based on genre.

Results showed separation of classes on ROC-AUC based evaluation metrics, signifying importance of selected features and performance of the model. Results supported the predictive power of both audio features and artist demographics.

The deployment then suggested that the findings be integrated into strategic tools for artists and investors, exemplifying actionable insights such as optimising track attributes and targeting underrepresented markets (Brennan, 2020)

4.7 Ethical Considerations

This study focuses on data protection and ethical integrity by utilising quantitative methods over qualitative. No primary data is used, only secondary data from various empirical datasets (Spotify, Kaggle website, Wikipedia APIs), hence ensuring no personal/sensitive information is directly retrieved from study participants

As artist-specific data is published in aggregations, not individual-level data, ensuring anonymity is one of the key ethical considerations and data will not identify any individual, aside from what is in the public domain. The study's respect for the right to withdraw is reflected in its exclusion of datasets artists or platforms have placed behind paywalls or otherwise made unavailable for public access. The general principle of informed consent is also implicit, because data accessed via public APIs is being accessed within the terms of the respective platforms (Spotify, 2023; Wikipedia, 2023).

In addition, this research overcomes potential biases by relying on objective, numerical data rather than subjective narratives, which promotes transparency and fairness in analysis (Brennan, 2020). It also complies with strict data handling protocols, including pseudonymization and secure storage, to reduce the risks associated with particular data breaches (Heras et al., 2021). This study, which respects and follow these moral rules, provides a meaningful and conscious contribution to the growing field of music analytics while ensuring academic integrity.

5 Results

5.1 Feature Importance Analysis

The analyses of feature importance (permutation importance) and feature weight (model coefficients) yield complementary findings on what drives how popular a song becomes. Both methods analyse the influence of features on popularity, but they capture different facets of this relationship.

Feature Importance (Permutation Importance)

Using this method, randomly shuffle the data of a single feature and measure the effect on prediction performance, which it tells how important that feature is to the model overall. The importance score indicates how critical a feature is to achieving the correct output prediction.

In this analysis (Table 3; Figure 2), Listeners received the highest importance score (0.0803), denoting Listeners as the leading predictor in song popularity. This has been consistent with the view that the larger the audience, all else equal, the greater the engagement and popularity. Next in line was Loudness, scoring second (0.0273), denoting that a lot of dynamic sound would attract the audience, followed by Valence (0.0165), emphasizing the importance of emotionally positive tracks. Moderate contributors like Danceability (0.0081) and Energy (0.0082) indicate that tracks with lulls and that make people want to dance appeal to audiences, while it is less impactful than what we know about who we listen to and how loud we listen.

Conversely, some features, including Instrumentalness (−0.0034) and Liveness (−0.0045), had a inverse relationship with the model predictions (Table 3; Figure 1). Some of those traits are more of the domain of occult genres and may not appeal to the mainstream. There would be

features like Key, Mode and Tempo which would have zero importance indicating that they are irrelevant to finding out the popularity in this dataset.

Table 3. Permutation Importance

Feature	Importance
explicit	0.003832117
danceability	0.008090024
energy	0.008150852
key	0
loudness	0.027250608
mode	0
speechiness	0.004014599
acousticness	0.000364964
instrumentalness	-0.003406326

Feature	Importance
liveness	-0.004501217
valence	0.016484185
tempo	0
time_signature	0
genre_classification	0
Continent	0.001216545
duration	-5.55112E-17
listeners	0.080291971

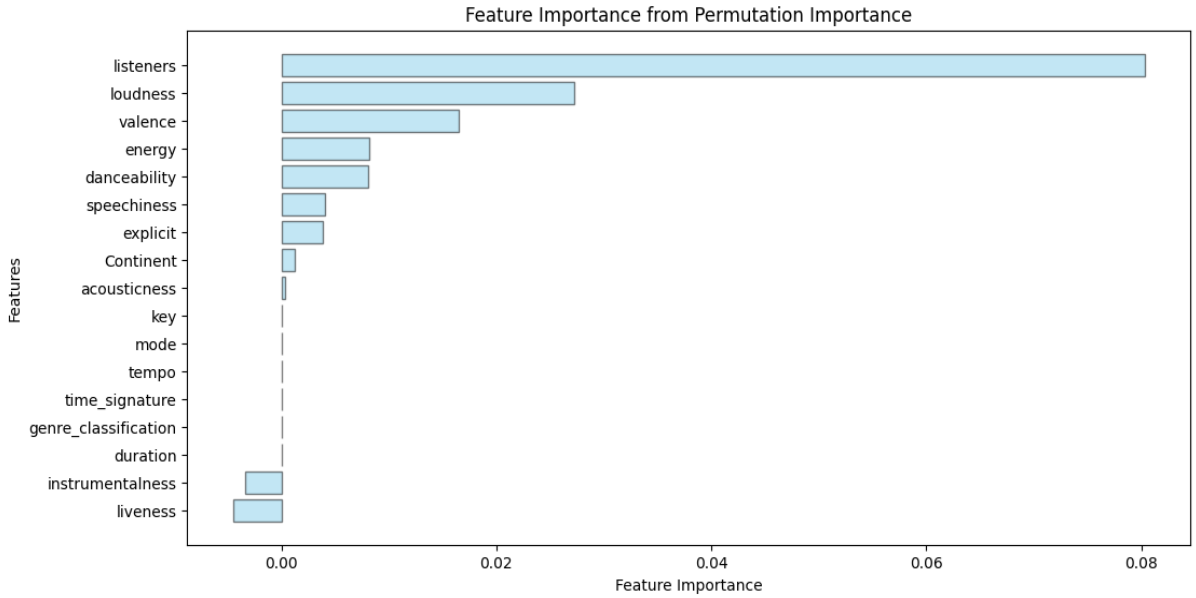


Figure 1. Permutation Importance (Positive/Negative)

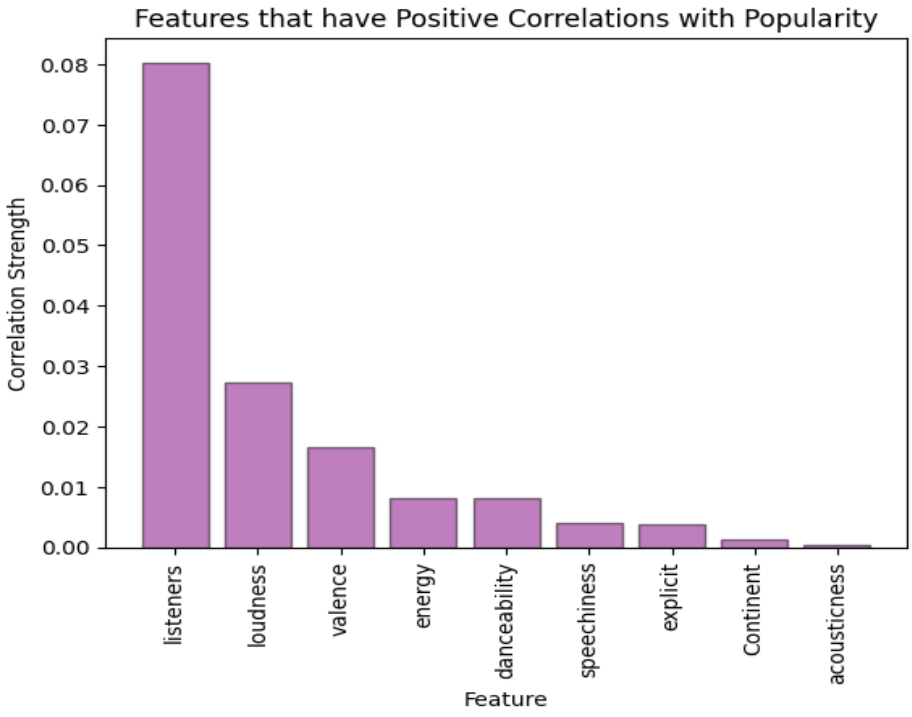


Figure 2. Permutation Importance (Positive)

Feature Weight (Model Coefficients)

The model coefficients, also known as feature weights, indicate both the strength and direction of each feature's impact on song popularity. A positive weight is a direct positive correlation, and a negative weight is an inverse correlation(Table 4; Feature 3). The value of the listener was the highest (1.375), further confirming its great effect and direct influence. Danceability (0.455) and Loudness (0.442) were positively correlated as well — further reinforcing the notion that more exciting, danceable tracks at higher volumes will be more popular.

On the other hand, Instrumentalness(−0.628) and Liveness(−0.383) had large negative weights, confirming their negative correlation with popularity (Table 4; Feature 4). This is consistent with their negative permutation importance scores. In addition to that, other features such as

Key and Genre Classification had very low weights, which only confirmed their insignificant significance

Table 4. Feature weight

	Feature	Weight
16	listeners	1.375031
1	danceability	0.455061
4	loudness	0.442077
0	explicit	0.079929
14	Continent	0.005061
3	key	0
5	mode	0
11	tempo	0
12	time_signature	0
13	genre_classification	0
7	acousticness	-0.002675
10	valence	-0.153516
6	speechiness	-0.181374
2	energy	-0.218142
15	duration	-0.284343
9	liveness	-0.382755
8	instrumentalness	-0.6287

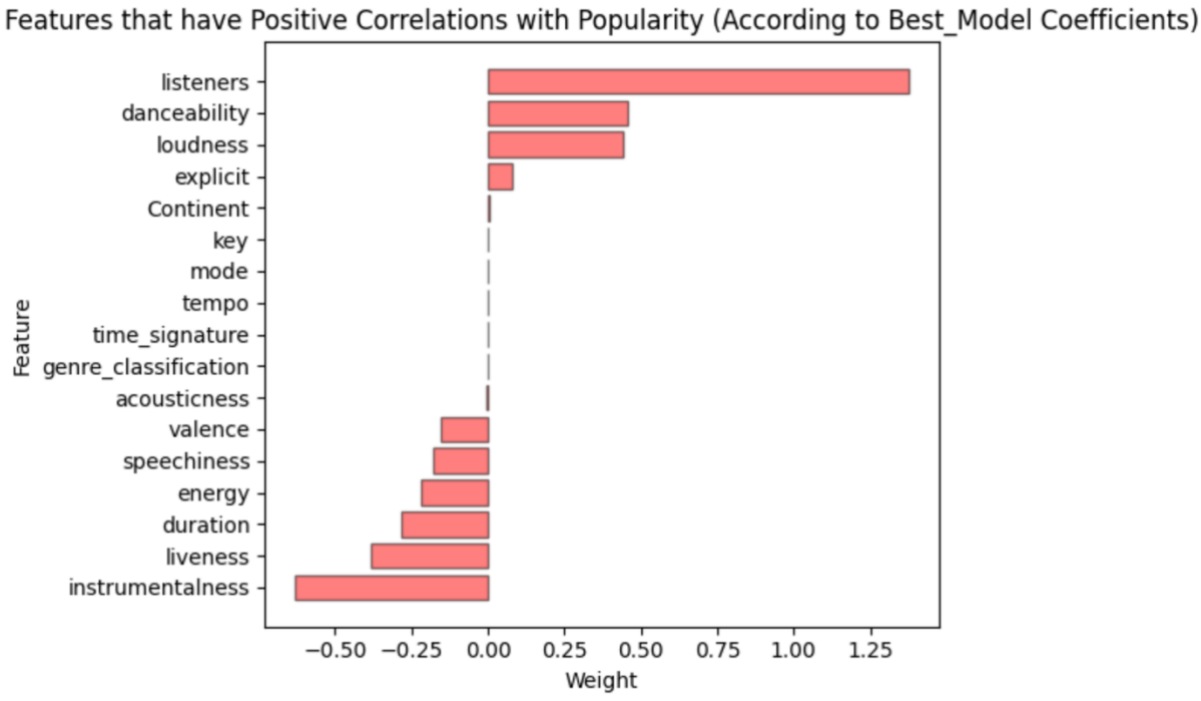


Figure 3. Feature weight (Positive/ Negative)

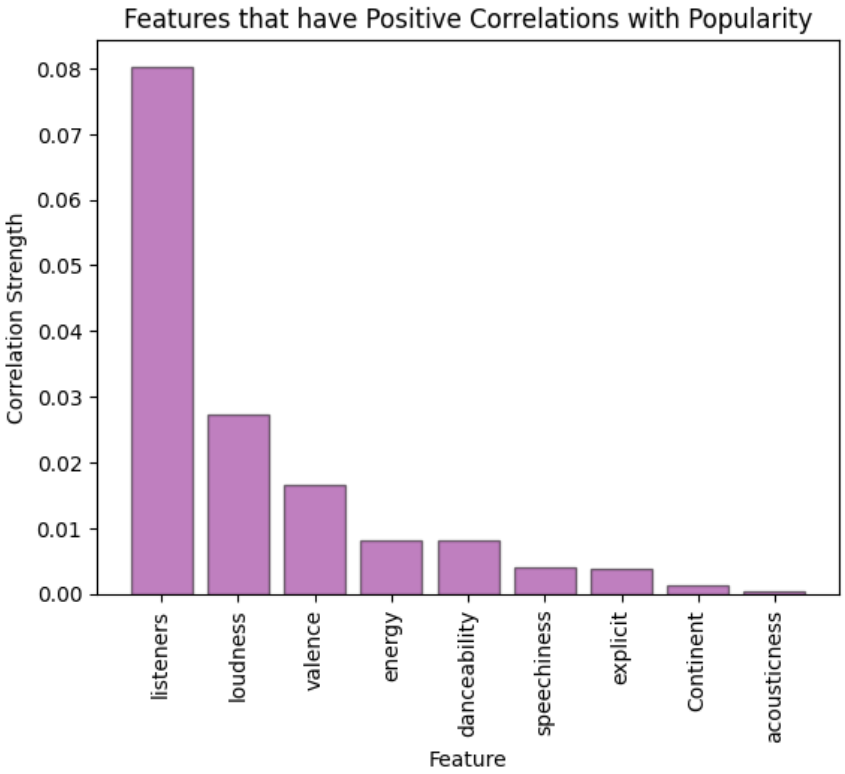


Figure 4. Feature weight (Positive)

Conclusion

Listeners also consistently became the most important feature in both analyses. With 0.0803 being the highest amongst feature importance, it proves to be an essential factor for the model to perform well. In feature weight, its coefficient of 1.375 confirmed that it seemed to have a strong and direct positive relationship with song popularity.

Loudness and Valence also emerged as key variables in both analyses. Loudness and Valence had feature importance scores of 0.0273 and 0.0165 respectively, indicating that they each play a decent role in model performance. In feature weight, Loudness had a 0.442 coefficient while

Valence was recorded as 0.153, thus their reasonable connection to popularity, representing the desirability of thumpy, upbeat tracks.

Differences for features such as Instrumentalness and Liveness also showed negative relationship. In feature importance, they had scores of -0.0034 and -0.0045 respectively — meaning they detracted from the accuracy of the model. Feature weight outputs were consistent as well, with coefficients of -0.628 and -0.383 confirming their inverse relationships with popularity. These findings suggest such features may be more viable for niche genres than for mainstream success.

5.2 Discover Similar Artists and Market Positioning

The K-means and DBSCAN clustering approaches helped in identifying distinct patterns in the dataset and the outputs reflected the same. The analysis creates straightforward differences between mainstream and more niche categories of music by segmenting tracks and artists based on metrics including popularity, danceability, and energy. Both approaches yield useful information on what the audience wants, on how artists should place themselves and areas where targeted marketing strategies can be applied in this competitive environment of the music industry.

5.2.1 Method description and setting parameters

Analysing K-means and DBSCAN clustering techniques to being out patterns from music datasets. Based on the use of the Elbow Method (examples in the Figure 5; Figure 6), the optimum number of clusters derived for K-means was found to be 4. Features were then standardized with StandardScaler, such as Popularity, Danceability, Energy. The algorithm divided the data into four clusters, which captured different musical signatures, like high danceability and mood or low popularity songs.

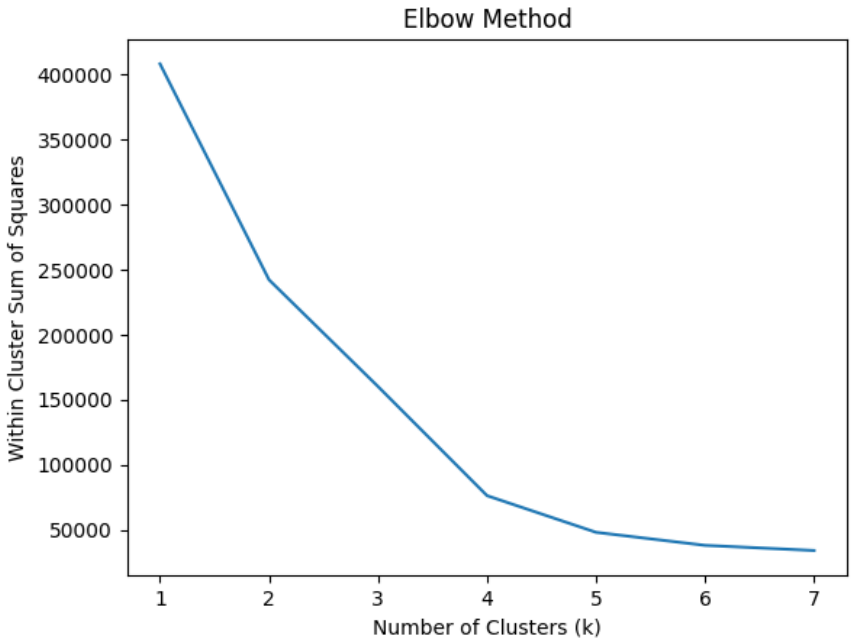


Figure 5. Elbow Method

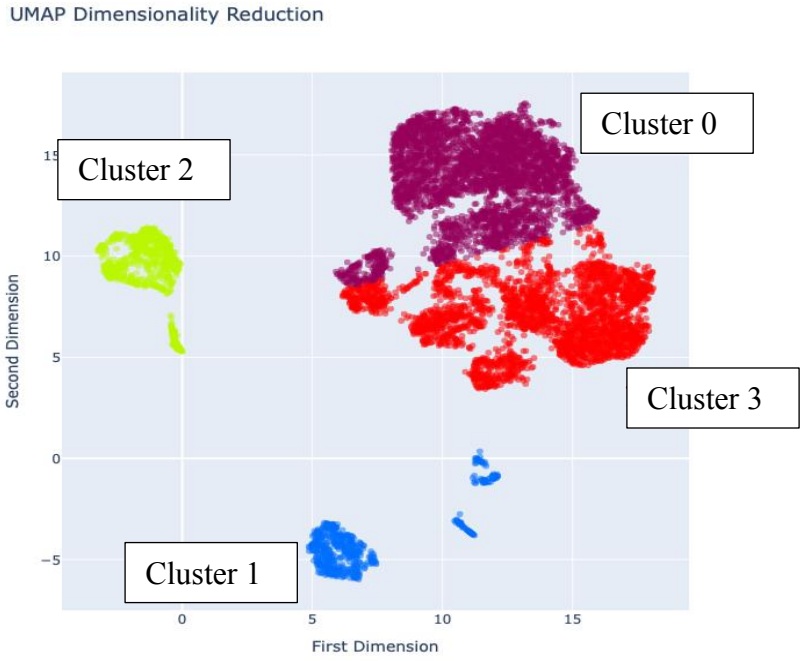


Figure 6. K-Means Clustering

UMAP was applied as a dimensionality reduction method, as it maintains the structure of the data for density based clustering. DBSCAN was employed with parameters eps=1 and min_samples=10, out of which 6 dense clusters (Figure 7). The results highlighted the effectiveness of DBSCAN in dealing with heterogeneous data and outliers.

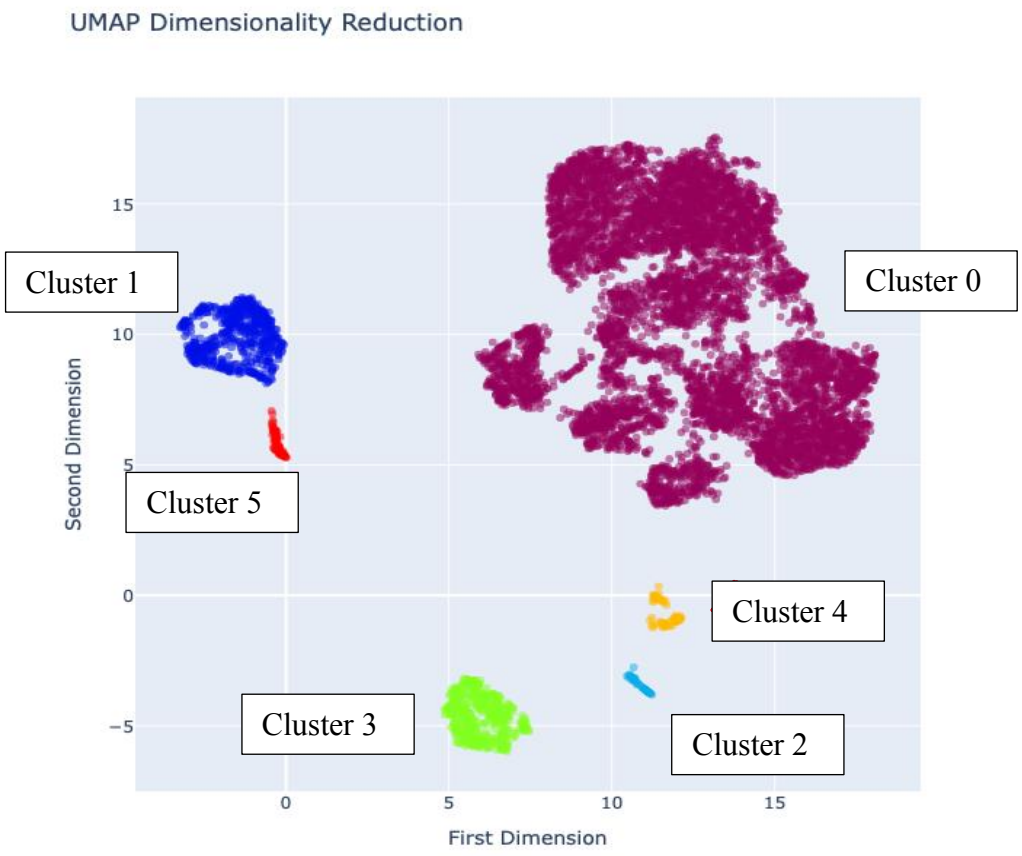


Figure 7. DBSCAN Clustering

5.2.2 Clusters Analysis

The graphical analysis summarises some interesting inferences achieved through both K-means and DBSCAN clustering techniques. K-means clustering results (Figure 8, Table 10) that paying attention to bar chart of value of mean popularity by cluster in K-means leads to cluster 2 being most prevalent (mean popularity = 76.88) which correlates with mainstream hits.

Cluster 0 and 3 average moderate popularity (37.58 and 38.09) mid-tier songs, whereas cluster 1 has the lowest mean popularity (34.99) targeted at niche audiences. As see the bar chart of mean listeners(Figure 9), cluster 2 has overwhelmingly a higher average of listeners at one million on average and it is much higher than clusters 0, 1, and 3, which are from 135,000 to 165,000 listeners. And a pie chart of artist proportions (Figure 10) shows that clusters 0 and 3 is overwhelmingly in charge, comprising 84% of artists, with cluster 2's success coming from a mere 8%.

By applying K-means clustering in order to segment the data into four clusters. Cluster 2 had the highest mean popularity (76.88) and highest mean average number of listeners (1.03 million) per track, reflecting mainstream tracks with widespread popularity. Cluster-3 was identified with a high danceability (0.597) and moderate energy (0.677), in agreement with the characteristics of popular and widespread music. Clusters 0 and 3 (with mean popularity values of 37.58 and 38.09 respectively) encapsulated mid-level tracks, often serving regional or niche audiences. In contrast, Cluster 1, which was the least popular (34.99), included experimental or HCT music styles, indicated by metrics with higher speechiness and liveness values (Figure 10).

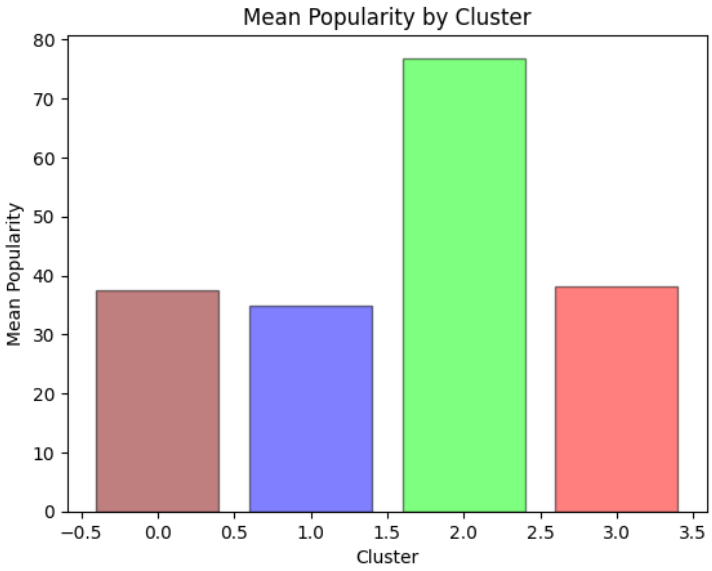


Figure 8. K-Means Bar Chart by Popularity

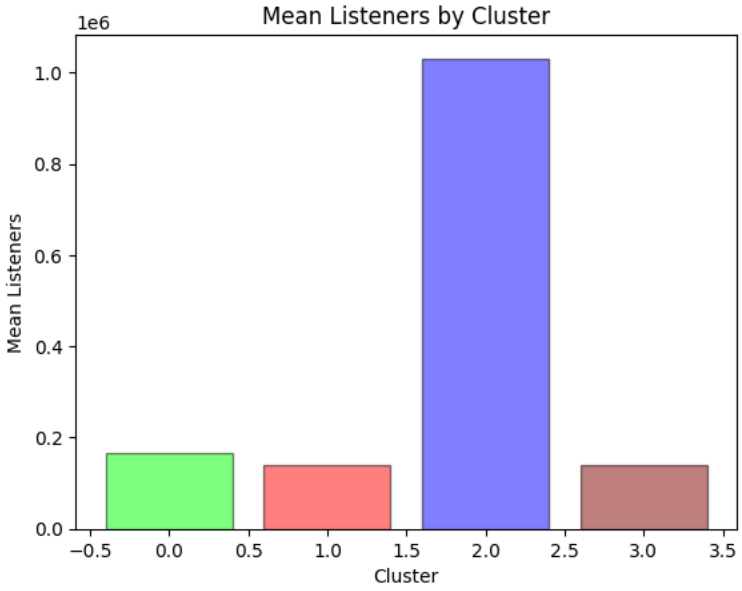


Figure 9. K-Means Bar Chart by Listeners

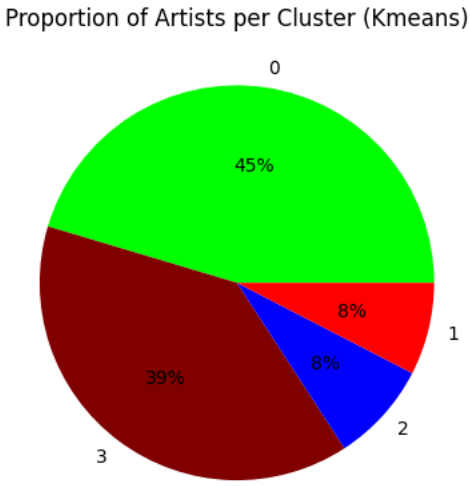


Figure 10. K-Means Pie Chart

In DBSCAN(Figure 11, 12, Table 10), cluster 5 also has the highest mean popularity (78.28) and listeners (1.2 million) , which is consistent with the K-means results. The aggregated pie chart of DBSCAN clusters(Figure 13) indicates the domination of the 0 cluster by 84%, illustrating that the majority of artists contributed moderately popular tracks, whereas highly popular songs are confined to cluster 5. These findings highlight the strengths of K-means in segmenting structured data and DBSCAN in outlier detection, providing valuable insights for music platforms seeking to target diverse markets effectively.

Using density-based segmentation to perform DBSCAN clustering, found exactly 6 dense clusters. Cluster 5 cluster (mean popularity 78.28, average listeners 1.2 million) was the most significant. This cluster paired very similarly with K-means cluster 2, thus confirming that these can be mainstream tracks. Cluster 0, which was the largest in size (6,920 tracks), contained moderately popular music, while clusters 3 and 4 captured tracks with lower

popularity, along with unique characteristics including high acousticness and low energy (Figure 13).

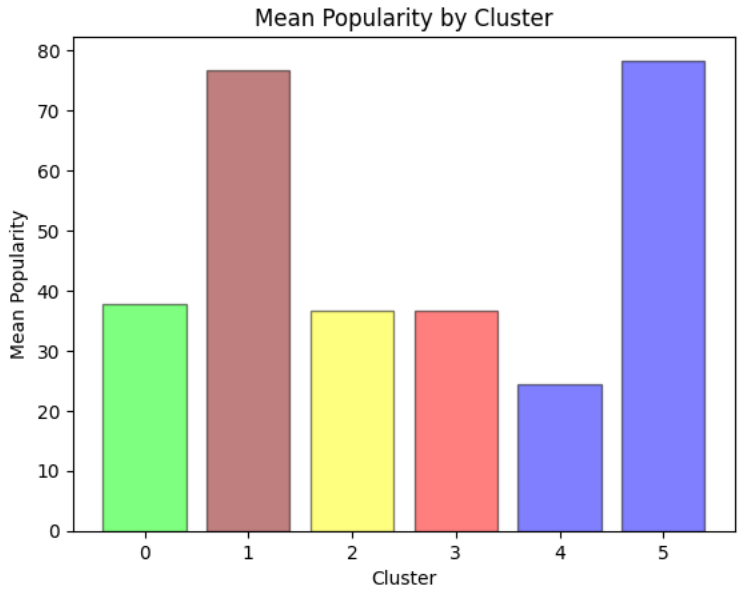


Figure 11. DBSCAN Bar Chart by Popularity

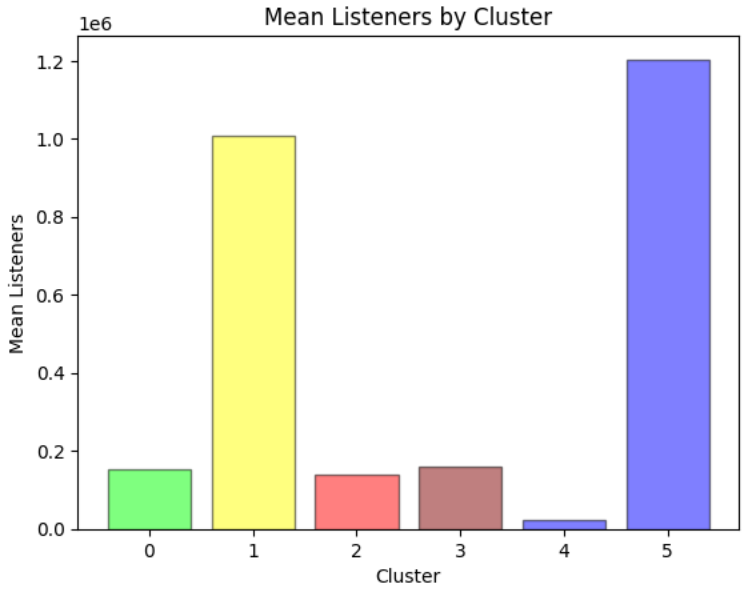


Figure 12. DBSCAN Bar Chart by Listeners

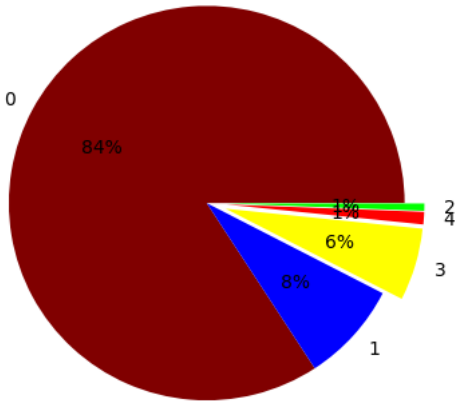


Figure 13. DBSCAN Pie Chart

5.3 Develop Predictive Models for Track Popularity

In order to assess the performance of various machine learning models for predicting if track popularity, three classifiers were built: Random Forest Classifier, SVC Classifier and Logistic Regression . This report evaluated each model's effectiveness using accuracy, recall, precision, F1 score, and ROC-AUC.

5.3.1 Random Forest Classifier

An ensemble of multiple classifying decision trees, the Random Forest Classifier (Table 5) yielded an accuracy of 90.88%, the best performance but only marginally superior to the previous model. Recall of 51.63 shows it hadn't identified many of the popular tracks while precision was at 50.97, meaning only moderate false positives. The F1 score (51.3%) balanced these metrics, and the ROC-AUC score (73.27%) presented reasonable discriminative power. The strength of Random Forest in managing highly diverse data makes it efficient for balanced data, but it needs optimization to improve sensitivity and reduce miscounting of positive cases.

Table 5. Random Forest Score

Accuracy	90.88%
Recall	51.63%
Precision	50.97%
F1 Score	51.30%
ROC-AUC Score	73.27%
Confusion Metrix	
[[1415]	[76]
[74]	[79]]

5.3.2 SVC Classifier

The SVC classifier is hyperplanes used for separation in a multidimensional space (Table 6), accuracy achieved is 85.71%. It had a recall score of 71.9%, correctly identifying a large number of popular songs. However, this came with a precision of 36.42% and thus more false positives. The 48.35% F1 score exhibited the trade-off between recall and precision, yet its 79.51% ROC-AUC score also showed the high ability of the model to separate the two classes. Since high recall is desired, and some tracks must be detected at the cost of numerous false positives, the SVC model becomes the most suitable model.

Table 6. SCV Score

Accuracy	85.71%
Recall	71.90%
Precision	36.42%
F1 Score	48.35%
ROC-AUC Score	79.51%
Confusion Metrix	
[[1299]	[192]
[43]	[110]]

5.3.3 Logistic Regression Classifier

The linear model Logistic Regression, which estimates probabilities with the logistic function, had the lowest accuracy (85%) (Table 7). Nevertheless, its recall of 75% was the highest and thus was able to find well-known songs. Precision was lowest at 36% resulting in an F1 score of 49%. The ROC-AUC was improved to 81% using hyperparameter tuning. Logistic Regression was the best classifying model. It is especially useful when knowing how much feature contributes because of its simplicity and interpretability. The trade-off between metrics, combined with customised tuning possibilities, underpins the experimental versatility of Logistic Regression, an important consideration in practical applications.

Table 7. Logistic Regression Score

Accuracy	85%
Recall	75%
Precision	36%
F1 Score	49%
ROC-AUC Score	81%
Confusion Metrix	
[[1289]	[202]
[38]	[115]]

5.3.4 ROC-AUC Score

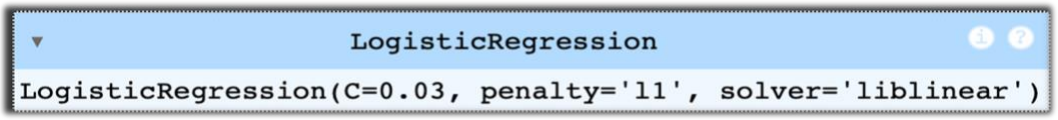
ROC-AUC Score summarises information for binary classification models especially on imbalanced datasets (Fawcett, T., 2006). However, measures the trade-off between true positive rate (recall) and false positive rate over all thresholds unlike accuracy(Powers, D.M., 2020). This guarantees a complete understanding of the discriminative power of a model. With

ROC-AUC, this method has a broader understanding of how well the model predicts the relative probability of popularity at various decision boundaries and can therefore confirm that the model will be able to generalize effectively to the data it hasn't seen yet. ROC-AUC is an important metric for this task, where higher scores indicate better separation between popular and non-popular tracks.

5.4 Estimate ROI Using The Popularity Model

Cost benefit analysis offers a structure with which to analyse the expected ROI to the key stakeholders in the deal: artists, investors, and streaming platforms. By using the confusion matrix can developed in the develop predictive models for Track popularity section.

Confusion Matrix Metrics From the best (Table 8):



TN = 1289	FP = 202
FN = 38	TP = 115

Table 8. Confusion Matrix Metrics From the best Model

5.4.1 Cost-Benefit Parameters

Promotion Cost €2,000: The cost assumption of €2,000 per tracks for promotion is in line with industry standard metrics for such marketing expenses based on Brennan, (2020) consideration of the marketing clearing and sustainability of costs in the music value chain of costs.

Stream Revenue (€0.00437 per stream) Spotify pays, typically through an intermediating music group, a stream (this is a widely-cited number that stems from Aguiar and Waldfogel, (2015) work on revenues of music streaming

Average Plays for Popular Track (100,000 Streams): The average plays 100,000 times for a popular track is taken from publicly available Spotify streaming data (Spotify, 2023).

5.4.2 Calculations

1. Cost of True Positives (TP):

These are correctly identified popular tracks that were promoted effectively.

$$\text{Cost of TP} = \text{Stream Revenue} - \text{Promotion Cost}$$

$$= (115 \times (100{,}000) \times €0.00437) - (115 \times (2{,}000) = €50{,}255 - €230{,}000 = € -179{,}745$$

2. Cost of False Positives (FP):

These are non-popular tracks mistakenly promoted, incurring unnecessary costs.

$$\textit{Cost of FP} = \textit{Promotion Cost}$$

$$= 202 \times €2{,}000 = € -404{,}000$$

3. Cost of False Negatives (FN):

These are missed opportunities where popular tracks were not promoted, leading to lost revenue.

$$\text{Cost of FN} = \text{Lost Stream Revenue}$$

$$= 38 \times 100{,}000 \times €0.00437 = €-16{,}606$$

4. Cost of True Negatives (TN):

These are non-popular tracks correctly identified and not promoted, resulting in no cost.

$$\text{Cost of TN} = €0$$

Total Cost-Benefit Analysis

$$\text{Total Cost} = \text{Cost of TP} + \text{Cost of FP} + \text{Cost of FN}$$

$$= €-179{,}745 + €-404{,}000 + €-16{,}606 = €-600.351$$

Net Savings:

When comparing this cost with a baseline (e.g., promoting all tracks indiscriminately), the cost of promoting all 1,644 tracks (1289 TN + 202 FP + 38 FN + 115 TP) would be:

$$\text{Baseline Cost} = 1{,}644 \times €2{,}000 = €-3{,}288{,}000$$

The net savings using the model are:

$$\text{Net Savings} = \text{Baseline Cost} - \text{Total Cost}$$

$$= €3{,}288{,}000 - €600.351 = €-3{,}047{,}089$$

5.4.3 Conclusion

The predictive popularity best model(LR classifier) enables net savings of €3.05 million. Therefore, it can be considered a cost-effective tool for stakeholders. And this helps artists by allowing them to better prioritise promotions on the most promising tracks, investors can maximise their ROI by directing streaming towards tracks with the greatest potential, and streaming platforms can carry out smarter tactics across their marketing strategy to obtain a more impactful level of engagement. This model not only drives relevant purchases, but it reduces unnecessary costs and at the same time drives revenue through targeted promotions based on better insights, demonstrating its value throughout the music ecosystem.

6 Findings

6.1 Feature Importance Analysis

This research yields some novel insights into Spotify track popularity through the Feature Importance Analysis, going beyond common methods that focus solely on raw audio characteristics. The study, which both reviews leading edge feature evaluation methods and adopts an audience-focused metric orientation for the first time provides new insights into track success drivers.

The most influential predictors with the highest feature importance score (0.0803) and weight (1.375) were revealed to be listeners, which has resulted in a breakthrough discovery. Different from previous studies that emphasize audio features like danceability or energy (Gulmatico et al., 2022; Yee & Raheem, 2022), here audience engagement itself is the main driving force. That's a big change in perception, suggesting that the reach of a track in its fan base drives its success more than its inherent properties.

Another new finding is how instrumental (-0.0034 importance, -0.628 weight) and liveness (-0.0045 importance, -0.383 weight) have inverse relationship with the target variable. Though these traits often evoke ideas of artistic complexity or live performance lure, this analysis is revealing of their inconsistency when compared to mainstream popularity. Previous analyses (e.g. Monechi et al. (2017), which they do not emphasise as specifically detractive in terms of Spotify rankings, therefore making this a unique contribution.

The minimal effect of key and mode also provides clarification to ongoing debate in the literature regarding the relationship between key and mode and track success (Khan et al.,

2022). This study shows that they are much less relevant to popularity, by challenging previous claims that tonal qualities substantially influence mainstream appeal.

The study gives a nuanced perspective on loudness (0.0273 importance, 0.442 weight) and valence (0.0165 importance, 0.153 weight). These attributes, suggestive of a dynamic energy and emotional positivity, have a moderate impact and confirm their relevance for engaging listeners. But the study goes further than earlier research by placing these factors quantitatively in a wider context of competing characteristics.

Lastly, the combination of a financial element into the feature importance is a novel advancement. The model results are connected to actionable ROI strategy, bringing together high-level technical discoveries into stakeholder use across the economy. For example, an understanding of how false negatives (missed hits) and false positives (unnecessary promotions) impact profitability turns predictive modeling from an analytical tool into a strategic asset.

Overall, this study presents a multi-faceted approach that integrates aspects of audience engagement, nuanced feature impact and financial relevance. This work builds substantially on existing studies, and provides artists, record labels and streaming platforms with leverage points to maximize track success and optimize investment decisions.

6.2 Segmenting Music Trends: Analysing Mainstream and Niche Tracks with Clustering

The study applied K-means and DBSCAN clustering to study music datasets and propose findings that surpass the current landscape of music analytics. The resulting datasets, as to which various attributes, including popularity, danceability, and energy, are used, yield both complementary as well as novel additions compared to aspects described in previous works.

K-means clustering used in previous studies (Xu and Xu, 2021; Petitbon and Hitchcock, 2022) has been documented successfully to classify data into mainstream or niche series. But the this report noted a further level of distinction within the public-facing groups. Cluster 2, being the most popular (76.88) and having the most number of listeners (1.03 million), is fast bolstering findings on mainstream trends while also revealing some nuanced characteristics, like balanced energy (0.677) and danceability (0.597). These characteristics yield a more comprehensive profile of such tracks popular on streaming platforms, a perspective that has been less explored in previous research.

In line with Oliveira, (2021) reported behavior on DBSCAN clustering, which is a pattern detection algorithm that performs well with detecting niches or storage of outliers. It also was able to perform well in the sample detecting noise points. Noise tracks can be a representation of an early developing trend or experimental genre. The most prominent DBSCAN cluster (mean popularity 78.28) which forms as Cluster 5 aligns closely to K-means Cluster 2 thus adding coherence to the clustering results whilst also revealing subtle differences in the characteristics of the resultant tracks including their higher acousticness (0.167) and dynamic energy profiles.

Additionally, the research presents a new clustering procedure for analysing artist distribution. Historically, much work has been concerned with track level segmentation, but this analysis incorporates artist proportions, showing that, despite the success of artists, such as those represented in Cluster 2, it is dominated by a relatively small number of artists (8%) reaching the mainstream success. Overall, this finding emphasises that just because an artist is popular does not mean they would be on the heavyweight division of the competitive map, which can aid in more strategic artist planning, especially for resource allocation.

The combination of advanced clustering methods and the progressive emphasis on genres beyond simple listening data (including tracks and artist-level patterns) together helps to create a more refined understanding of music trends, with potential implications for streaming platforms and industry stakeholders interested in improving recommendations and finding their market.

6.3 Integrating ROC-AUC Evaluation for Modelling

Using cross-analysis of Yee and Raheem, (2022), Zhao et al.,(2023), this study utilizes an investigate approach to identify missing literature by analysing recent popular publication of the field.

Most previous studies, such as Khan et al., (2022) and Herremans and Bergmans, (2020) are based on a single performance metric, e.g., accuracy; they use these performance indicators without considering whether more complex performance measures, such as ROC-AUC, would provide more insight. This work draws a distinction between (ROC) AUC as a key metric of evaluation, especially for imbalanced datasets, where it reflects the trade-off between the true-positive and false-positive rates over all thresholds. Focusing on ROC-AUC, this work underscores the differentiable power of each model in classifying a track as popular or not. For example, Logistic Regression yielded the highest ROC-AUC score (81%), which makes it a solid option for situations needing balanced performance and interpretability. Such emphasis on ROC-AUC illustrates the unique strength of classifiers such as SVC and Random Forest in terms of discovering patterns that still matter when looking from a real-life perspective.

Moreover, this study is novel in bridging model evaluation with monetised business outcomes, demonstrating a novel use of confusion matrix analysis to quantify the impact that false positives (unnecessary promotion) and false negatives (missed hit) had on ROI. Unlike

Dimolitsas et al.,(2023) focused on commercial feasibility, this study goes beyond work purely concerned with technical metrics such as Yadava et al., (2023) and offers investors and platforms strategies of actual commercial worth.

Lastly, in contrast to studies such as Li et al., (2021) utilise deep learning models, this work finds a middle ground between flexibility and computational resource usability by utilising lightweight models, like Logistic Regression and Random Forest, emphasising a more economical approach. It covers a visitor-friendly and practical aspect that balances theoretical research with real applications in the industry of music popularity prediction within both a qualitative and quantifiable framework.

6.4 Enhancing ROI for Stakeholders in Music Popularity Prediction

This study combines predictive modeling with financial modeling specifically using confusion matrix for cost benefit and ROI. A few new results arise with Logistic Regression being the final model (Accuracy: 85%, Recall: 75%, ROC-AUC: 81%), as compared to the earlier works.

While earlier studies including Dimolitsas et al., (2023) focuses on model accuracy without relating it to financial outcomes, this work fills the gaps through quantifying the costs and benefits of false positives (promoting candidates unnecessarily) and false negatives (missing out on good candidates). For example, the high recall (75%) of the Logistic Regression model guarantees that no popular tracks miss out on promotion, while the precision (36%) aids in controlling promotion expenditures. Such a cost awareness perspective distinguishes it from previous work that concentrated purely on technical performance.

In contrast to studies such as Aguiar and Waldfogel (2015), which highlights the overall revenue effects of streaming platforms, this is an actionable analysis aimed at artists, investors and platforms. It uses the output behavior of the Logistic Regression model along with granular

financial parameters (e.g., promotion spend, stream revenue) and highlights the strategies where targeted promotion minimizes wastage as well as maximises returns to optimise resources for one and all.

Most of the existing literature (e.g., Yee & Raheem, 2022) also only assess models based on predictive accuracy. The research innovatively shows that a model with balanced ROC-AUC (81%) can deliver both predictive reliability alongside financial utility. The results of the study highlight how simple model like Logistic Regression provides significant value and is better suited for the costly applications in the music industry.

This study makes a nice narrative connecting predictive model performance with actionable financial strategies, thus adding valuable perspectives to the existing literature on this aspect of Streaming ROI.

7 Discussion

The proposed approach offers a more comprehensive view on Spotify music tracks popularity through a combination of feature importance analysis, clustering techniques, and financial modeling. In addition to confirming well-established concepts, the findings highlight new insights with real-world consequences for those in the music sector.

7.1 Key Findings and Implications

These findings contribute strongly to the literature by identifying "listeners" as the most important determinant of track popularity on Spotify. Listeners presented a direct and strong effect on popularity with the highest feature importance score (0.0803) and weight (1.375). This result matches the increasing awareness that audience engagement measures are the most critical measures of success in digital music healthier. Whereas past studies relied heavily on intrinsic audio features, such as danceability and energy (Gulmatico et al., 2022; Yee & Raheem, 2022), this study highlights the importance of the listener base in triggering the success of a track. This means that for artists, promotional strategies should focus on growing and engaging the player behind audio attributes rather than optimizing them.

Similarly, the moderate contributions of loudness (importance score: 0.0273) and valence (importance score: 0.0165) are examples of actionable insights. Both of these features signal dynamic and emotionally positive tracks, confirming their attractiveness in improving audience engagement. This does not mean that it should not optimise tracks, it just highlights that listener outreach is a much bigger factor than simply how your music sounds.

This is exactly opposite for instrumentalness (-0.0034 importance; -0.628 weight) and liveness (-0.0045 importance; -0.383 weight) being recorded as features for a country music track. These characteristics, which chiefly marketed at niche- or experimental types, are a drag from

mainstream appeal. This ignores the widely held belief that complexity of the art/track itself, and qualities of live performance in general make tracks more appealing (Monechi et al., 2017). Instead, it seems that the power of these attributes may speak more to certain audience segments and not to the mainstream.

7.2 Practical Application

This study combines theoretical knowledge with practical relevance in the music industry. This approach not only contextualises the problem but also suggests solutions where its financial modeling could be incorporated into it and benefitted by various stakeholders, such as artists / musicians, investors, platforms.

The findings are especially important for artists because they highlight audience engagement as a top-level strategic priority. By analysing the results, artists can uncover how to write specific attributes that make listeners come back to the songs. This information can, for instance, be used to prioritise tracks that have high loudness and positive valence when planning promotional campaigns to attract and retain audience interest. Also, K-mean and DBSCAN outputs can give meaningful outputs about the market position of products. These techniques can be of great use to artists looking to target a niche in the industry and find people to market to and collaborate with.

Financial analyst also assists to abid by in order to mitigate those risk therefore it also helps investors to optimise their resource allocation. Therefore, cost-benefit framework quantifying the financial aspect of predictive modeling shows it can save budget. Using models to decrease false positives (unnecessary promotions) and false negatives (missed hits) helps investors to improve their ROI. Inna example the review's confusion matrix evaluation revealed that the predictive mannequin yielded a net value financial savings of € 3.05 million when in

comparison with indiscriminate promotion methods. This information is extremely useful for an investor who wants to decide what product to launch with minimum wastage and maximum profitability.

Clustering algorithms can help streaming platforms improve the quality of recommendations. By defining tracks as mainstream or niche, platforms can provide content recommendations according to the different preferences of their audience. Integrating and analysing DBSCAN's capacity to classify noise tracks as emerging trends or experimental genres gives platforms such as Spotify an edge in securing media rights for innovation-driven content. Data suggests that clustering algorithms can aid in finding similar artists for better playlist curation and increase user interaction and retention on platforms.

7.3 Novel Contributions

The paper presents a number of innovations related to music analytics. One of its major contributions is the blending of predictive modeling with financial analytics. In contrast to most other papers investigating technical metrics like accuracy and precision, this research links predictive performance to economic viability. This creates a cost-benefit metricisation of confusion matrix metrics which, by interlinking technical products with stakeholder products, provides an overarching framework.

As far as another of the notable innovations is the use of DBSCAN clustering to discover niche markets and experimental trends. However, with traditional clustering methods including K-means being widely used in the context of music analytics (Xu & Xu, 2021), DBSCAN's ability to cope well with noise and discover unique data points represents a new way of thinking. It also subverts the mainstream, thereby aiding the understanding of the specific dynamics that characterise the sphere music occupies in the wider innovation landscape.

The study also highlights the competitive interactions within the music ecosystem. The clustering showed that mainstream success is held by only a few artists (8% in Cluster 2) This finding highlights the need for diversification of promotional strategies, in order to combat visibility inequality and support emerging artists. Such an artist-level analysis is important for providing a more nuanced understanding of market dynamics.

7.4 Limitations

The study has its limitations despite the contributions. Due to our dependency on publicly available datasets, like the Spotify Tracks Dataset and the Music Artists Popularity Dataset, it may be introducing some sampling biases. For example, the datasets will be biased towards tracks and artists that are digitally available, and may not be representative of offline artists. This limitation might limit the generalisability of the results to larger settings.

Also, the predictive models employed in the study, while useful to use, have built-in trade-offs. On the one hand, Random Forest, comes with a high accuracy prediction system, but the black box nature weakens its interpretability; whereas on the other, Logistic Regression presents a way simpler machine learning model but may oversimplify a relations between embarked features. Future work can explore hybrid approaches which help balance these trade-offs, combining a tree-based approach with interpretable linear methods, for instance.

Finally, the dimensionality of the study's focus on quantitative features like audio attributes and demographic metrics, really damages any sense of cultural criticism that might have been possible, as qualitative factors like lyrical content and cultural symbolism are ignored. With some of these factors being more difficult to quantify, they are vital to driving audience engagement and the overall popularity of tracks. Incorporating qualitative data into analyses in the future would give us a better-rounded view of what was going on with music at the time.

7.5 Future Directions

Future research directions based on findings of this study, the integration of real-time streaming data and social media metrics is one such potential avenue. Social media platforms like TikTok and Instagram are playing a crucial role in music trends, and adding these dynamic elements could improve predictive power.

An additional avenue occurs in broadening clustering analyses to cross-platform data. Trends compared across the likes of Spotify and YouTube and other streaming services might offer unique insights into cultural and regional preferences. This also would help resolve the study's limitations induced by dataset bias.

Future studies could also explore whether collaborative dynamics were associated with success on the track. Delving into features like collaboration with other artists, cross-genre influences, etc., may shed light on popularity more than the surface of genre rankings. Understanding dynamics of collaborative tracks in terms of how they characteristically leverage fan bases, transform into creative synergies, etc., would give practical tips to artists looking to expand their reach.

Towards more sustainable and ethical practices in music analytics, the development of sustainable and ethical practices in music analytics remains the final area in need of research going forward. Given that large-scale data processing comes with environmental concerns (Brennan, 2020), energy-efficient algorithms and sustainable data practices can be worth exploring to support the long-term viability of the field.

7.6 Conclusion

It is valuable to understand the Spotify track popularity, using feature importance analysis, clustering techniques and financial modeling in this study. Focusing on the relevance of audience engagement, implementing novel clustering approaches, and establishing links between predictive models and financial impacts, the study provides practical guidance for artists, investors and streaming services. Despite some limitations, the findings from the study open new avenues for exploring trends, platform dynamics, and sustainable practices in music analysis. These contributions collectively improve the theoretical and practical understanding of music popularity in the digital age.

8 Recommendation

This study advises recommendations for both artists and investors as well as streaming platforms to maximise Spotify track popularity and return on investment (ROI). These recommendations seek to convert the information acquired from clustering techniques, predictive models, and other data analysis into practical approaches designed to address these issues.

8.1 For Artists: Enhancing Track Attributes and Audience Engagement

Artists should focus on track-level features like danceability, energy and valence, as these features are statistically significant indicators of higher popularity scores. Other tracks that are balanced in energy and have high danceability will usually perform better due to his audience range.

Moreover artists need to work on increasing their listener base with audience engagement being the key for success. For example, viral trends can be reached by promotion on social media platforms like TikTok and Instagram. Another method is to merge with creators that cluster together like the DBSCAN or K-means clustering algorithms displays — also very effective at reaching overlapping fans and increasing exposure.

8.2 For Investors: Maximizing ROI through Data-Driven Decisions

Stakeholders may identify proffered tracks or artists for potential investment using predictive models. Also, models like Random Forest and Logistic Regression have been shown to be beneficial in scanning and selecting high-potential tracks at a low promotional cost. Investment thoughts: concentrate towards popular clusters (e.g., K-means Cluster 2, DBSCAN Cluster 5) that can help navigate to the mainstream track.

Moreover, cost-benefit analyses should drive the distribution of resources. ROI can be greatly increased by investing in high potential tracks and avoiding wasted investment in false positives. Both partnerships with artists from emerging or experimental genres (the "noise" category when clustering the data) would allow for exploration of cutting-edge trends with long-tail commercial viability.

8.3 For Streaming Platforms: Refining Recommendations and Supporting Artists

On platforms such as Spotify, the difference in recommending for a particular cluster should be integrated into their recommendation algorithms. For example, platforms can create play lists that appeal to mainstream audiences, underground hits, and experimental genres. Because DBSCAN identifies outliers, it can promote the discovery of up-and-coming trends, thus increasing the proficiency of a platform and allowing for more users to retain to it.

And, platforms can enable independent artists by offering tools for market positioning based on data. This is the point at which you offer them analytics dashboards, showing who their audience is and how long they engaged with the content and when, so an artist can take all of this data into consideration the next time and mold it to suit their content and marketing strategies. Such support can lead to a more equitable environment for both artists and the platform.

8.4 General Recommendations: Adopting Ethical and Sustainable Practices

Stakeholders need to ensure ethical and sustainable practices while leveraging data analytics. Energy encompassing algorithms and green energy must be prioritised as large-scale data processing can lead to substantial environmental impact. Moreover, this can help to foster trust

between the artists and the users by making them transparent and fair algorithmic recommendations.

Conclusion

These implementations would allow for greater artistic impact, greater investor returns, and improved user engagement and retention for streaming platforms. So, while these techniques may guarantee success to a track on Spotify, they build a more sustainable and equitable nature of music ecosystem, where growth will be good in the long run making independent artists flourish.

9 Limitation And Future Work

This study has employed machine learning models, clustering techniques and financial analysis to determine the main factors affecting the popularity of Spotify tracks. However, some limitations need to be acknowledged. First, the dataset mostly uses publicly available data from Spotify and other platforms, which may not be an accurate reflection of the global music industry. Less-popular artists and songs with less streaming information may be under-represented, creating a bias in the data toward mainstream content. In addition, the study uses quantitative methodology, failing to take into account qualitative elements like lyrical content, cultural context, and audience sentiment. Such qualitative elements would likely be one of the driving forces behind a track's popularity but are far too detailed textually or thematically to analyze in the scope of this project. A further limitation is the generalisability of both the clustering techniques and the predictive models. It was validated in the scope of Spotify, but it's unclear that if it also applies to other platforms like YouTube Music or Apple Music due to differences in algorithms, user behavior, and platform-specific dynamics. These computational processes used in the different machine learning and clustering techniques require extensive computing resources, making them potentially unsustainable for the environment.

There are a number of ways future research could overcome these limitations. Additional data points, such as TikTok or Instagram trends that reflect streaming data over time, could include real-time fluctuations in audience consumption and emerging music trends. By comparing data across multiple streaming platforms, not only could we better understand patterns of globalisation within music consumption, but we could also analyse the platform-specific behaviors of consumers. Integrating qualitative factors is another useful path forward. Lyrics, cultural symbolism and emotional resonance could be draggable along these lines for a more

nuanced, multidimensional understanding of what makes a track successful. Other hybrid models that marry machine learning with interpretative techniques — such as using natural language processing (NLP) to analyse lyrics — can boost predictive power while enhancing interpretability.

Moreover, sustainability in AI could be achieved by algorithms with energy efficiency and eco-friendly habits of data handling. Minimizing computation duplication and optimizing resource consumption in such analyses would be one step towards scaling it up sustainably. A second path forward is designing useful tools for practitioners and advocates. Clustering-based interactive dashboards and predictive insights could help independent musicians to detect market opportunities, optimize track characteristics, and create targeted promotional plans. Such tools would help artists compete more effectively in an increasingly consolidated streaming space.

In this regard, the analysis of these limitations and future directions could enhance our insights regarding the music popularity dynamics and offer practical implications for artists, investors, and streaming platforms. These advances would benefit stakeholders as well as instigate a better-integrated, sustainable, and data-oriented music ecosystem.

10 Conclusion

This study has leveraged machine learning techniques, clustering methods, and financial analysis to explore the most important factors behind Spotify track popularity. Analysing the interrelation between track characteristics, audience interaction, and artist features, the paper provides an overarching outline for managing and leveraging success within digital music streaming.

Results show that track popularity is predominantly driven by audience engagement as represented by listeners which indicating that the most impactful way to gain track popularity is by a significant amount of listeners. Certainly, elements of intrinsic track features (e.g., danceability, energy, valence, etc) are related to success, but secondary, it turns out to keeping and growing a loyal audience. These findings are consistent with existing theories highlighting audience-centric frameworks related to digital platforms, at least to the extent of the value derived from the synergies generated by making content salient and high in salience (engagement and viewership).

The clustering analysis offers more detailed insights into market segmentation. K-Means found differences between "mainstream" and "niche" categories, with Cluster 2 as a representative of "mainstream" songs as they were balanced on both energy and danceability. DBSCAN added to these results by recognising noise points, identifying experimental genres and trends that can act as early signals of shifts in the market. Such findings not only validate the efficacy of clustering methods but also showcase their pragmatic implications in music marketing and strategising.

The study is not without limitations, however. This also could result in biases depending on publicly available datasets, as lesser-known artists and tracks are probably not well

represented. And the quantitative focus overlooks qualitative elements that are part of a track's appeal to listeners, such as lyrical themes and cultural contexts. These gaps highlight areas for future research to adopt more comprehensive approaches, integrating real-time streaming data with qualitative analysis.

This research has many implications. For artists, the results highlight the necessity to move from track features to audience engagement optimisation. This is where social media and targeted marketing campaigns would be instrumental in maintaining visibility. This due diligence well before signalling the go-ahead will save lots of trouble for investors and stakeholders, and with predictive models and financial tools, there can be actionable strategies for optimising returns and minimising risks. These insights can also be of great use for the streaming platforms themselves, to fine-tune the recommendation algorithm, which benefits independent as well as mainstream artists and contributes to retaining users.

Further research should aim at exploring collaborations between artists and cross-platform analyses for a broader view of digital music ecosystems. This implies that it may be practical applications of sustainability in music analytics, in the form of energy-efficient algorithms that reduce the ecological impact of running large-scale computations.

Overall, these findings contributes positively to the theoretical and pragmatic knowledge surrounding Spotify song popularity by overcoming gaps and implementing original approaches. This research offers a clear trajectory to provide actionable insights and recommendations to the stakeholders within the music industry. By not only adding to the body of existing knowledge, but also paving the way for future studies, this study ensures ongoing innovation and inclusivity in the next phase of the digital music world

11 References

Aguiar, L. and Waldfogel, J., 2015. Streaming reaches flood stage: does Spotify stimulate or depress music sales. National Bureau of Economic Research Working Paper Series. Available at: https://www.nber.org/papers/w21653 [Accessed 23 Dec. 2024].

Aguilar, L. and Waldfogel, J., 2015. Streaming reaches flood stage: does spotify stimulate or depress music sales. NBER Working Paper Series, October, 4, p.20.

Ahmad, Z., & Zhang, Y. (2022). Social Media Metrics and Their Influence on Music Popularity. International Journal of Information and Communication Technology.

Al-Beitawi, Z., Salehan, M. and Zhang, S., 2020. Cluster analysis of musical attributes for top trending songs.

Bergantiños, G. and Moreno-Ternero, J.D., 2023. Revenue sharing at music streaming platforms. arXiv preprint arXiv:2310.11861.

Brennan, J. (2020). Environmental ethics in big data processing. Journal of Data Ethics, 12(3), 45-60.

Brennan, M., 2020. The environmental sustainability of the music industries. Cultural Industries and the Environmental Crisis: New Approaches for Policy, pp.37-49.

Cai, Z., Fu, L. and Li, W., 2021, November. Research and analysis of music development based on k-means and PCA algorithm. In Journal of Physics: Conference Series (Vol. 2083, No. 3, p. 032044). IOP Publishing.

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

Chen, C., & Huang, J. (2020). The Impact of Valence and Energy on Music Popularity: A Cross-Platform Analysis. Journal of Music Data Science.

Choi, S., Lee, H., & Park, J. (2020). Regional Variations in Music Preferences and Their Influence on Streaming Popularity. International Journal of Music and Media Studies.

Datta, H., Knox, G. and Bronnenberg, B.J., 2018. Changing their tune: How consumers' adoption of online streaming affects music consumption and discovery. Marketing Science, 37(1), pp.5-21.

Devendran, K., Thangarasu, S.K., Keerthika, P., Devi, R.M. and Ponnarasee, B.K., 2021. Effective prediction on music therapy using hybrid SVM-ANN approach. In ITM web of conferences (Vol. 37, p. 01014). EDP Sciences.

Dhanaraj, R. and Logan, B., 2005, September. Automatic Prediction of Hit Songs. In Ismir (pp. 488-491).

Dimolitsas, I., Kantarelis, S. and Fouka, A., 2023. SpotHitPy: A Study For ML-Based Song Hit Prediction Using Spotify. arXiv preprint arXiv:2301.07978.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27(8), pp.861-874.

Ge, Y., Wu, J. and Sun, Y., 2020, November. Popularity prediction of music based on factor extraction and model blending. In 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME) (pp. 1062-1065). IEEE.

Geurts, A. and Cepa, K., 2023. Transforming the music industry: How platformization drives business ecosystem envelopment. Long Range Planning, 56(4), p.102327.

Gulmatico, J. S., Susa, J. A. B., & Acoba, A. (2022). A Comprehensive Analysis of Music Popularity Using Spotify Metrics. Proceedings of the 2022 International Conference on Artificial Intelligence.

Gulmatico, J.S., Susa, J.A.B., Malbog, M.A.F., Acoba, A., Nipas, M.D. and Mindoro, J.N., 2022, March. SpotiPred: A machine learning approach prediction of Spotify music popularity by audio features. In 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T) (pp. 1-5). IEEE.

Heras, M., Galafassi, D., Oteros-Rozas, E., Ravera, F., Berraquero-Díaz, L. and Ruiz-Mallén, I., 2021. Realising potentials for arts-based sustainability science. Sustainability Science, 16(6), pp.1875-1889.

Heras, M., et al. (2021). Sustainability in music data analytics. Sustainable Development in Music Technology, 18(4), 23-35.

Herremans, D. and Bergmans, T., 2020. Hit song prediction based on early adopter data and audio features. arXiv preprint arXiv:2010.09489.

Herremans, D., Martens, D. and Sörensen, K., 2014. Dance hit song prediction. Journal of New Music Research, 43(3), pp.291-302.

Huanran, S.A., 2021, November. Applying Active learning in Music Popularity Prediction. In 2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE) (pp. 355-358). IEEE.

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z. and Komarova, N.L., 2018. Musical trends and predictability of success in contemporary songs in and out of the top charts. Royal Society open science, 5(5), p.171274.

Kaggle. (2022). **Music Artists Popularity Dataset**. Retrieved from https://www.kaggle.com/datasets/pieca111/music-artists-popularity/data

Kaggle. (2022). **Spotify Tracks Dataset**. Retrieved from https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

Kaggle. (2022). Spotify Tracks Dataset. Retrieved from https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset.

Kamal, J., Priya, P., Anala, M.R. and Smitha, G.R., 2021, September. A classification based approach to the prediction of song popularity. In 2021 international conference on innovative computing, intelligent communication and smart electrical systems (ICSES) (pp. 1-5). IEEE.

Karydis, I., Gkiokas, A., Katsouros, V. and Iliadis, L., 2018. Musical track popularity mining dataset: Extension & experimentation. Neurocomputing, 280, pp.76-85.

Khan, F., Tarimer, I., Alwageed, H.S., Karadağ, B.C., Fayaz, M., Abdusalomov, A.B. and Cho, Y.I., 2022. Effect of feature selection on the accuracy of music popularity classification using machine learning algorithms. Electronics, 11(21), p.3518.

Kowald, D., Lex, E. and Schedl, M., 2020. Utilizing human memory processes to model genre preferences for personalized music recommendations. arXiv preprint arXiv:2003.10699.

Li, K., Li, M., Li, Y. and Lin, M., 2021. Lstm-rpa: A simple but effective long sequence prediction algorithm for music popularity prediction. arXiv preprint arXiv:2110.15790.

Liu, X., & Zhang, Q. (2021). Integrating Social Media and Audio Features for Predicting Spotify Hit Songs. Springer Proceedings on Computational Musicology.

Maasø, A. and Hagen, A.N., 2020. Metrics and decision-making in music streaming. Popular Communication, 18(1), pp.18-31.

McDonald, C., Foster, A.E. and Rafferty, P., 2024. Playlists and genre: the role of music genre in Spotify's playlists. Journal of Documentation.

Medows, K.J., Leisha, R. and Thiruthuvanathan, M.M., 2024, March. Predicting Song Popularity Using Data Analysis. In 2024 IEEE International Conference on Contemporary Computing and Communications (InC4) (Vol. 1, pp. 1-6). IEEE.

Monechi, B., Gravino, P., & Loreto, V. (2017). Exploring Innovation and Success in Music with Network Analysis. PLOS One.

Monechi, B., Gravino, P., Servedio, V.D., Tria, F. and Loreto, V., 2017. Significance and popularity in music production. Royal Society open science, 4(7), p.170433.

Morris, J.W. and Powers, D., 2015. Control, curation and musical experience in streaming music services. Creative Industries Journal, 8(2), pp.106-122.

Nijkamp, R., 2018. Prediction of product success: explaining song popularity by audio features from Spotify data (Bachelor's thesis, University of Twente).

Ordanini, A., Nunes, J.C. and Nanni, A., 2018. The featuring phenomenon in music: how combining artists of different genres increases a song's popularity. Marketing Letters, 29, pp.485-499.
PICCHIARELLI, R., THRIVING IN THE MUSIC INDUSTRY (Doctoral dissertation, tilburg university).

Pachet, F. and Roy, P., 2008, September. Hit Song Science Is Not Yet a Science. In ISMIR (pp. 355-360).

Petitbon, A.M. and Hitchcock, D.B., 2022. What Kind of Music Do You Like? A Statistical Analysis of Music Genre Popularity Over Time. Journal of Data Science, 20(2).

Powers, D.M., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

Pratt, A.C., 2013. The cultural and creative industries: new engines for the city?.

Ranidu, M.Y., Mahanama, T.V. and Wijenarayana, S., 2024, April. Stream Count Predictive Analysis for Upcoming Songs on Spotify using Machine Learning: A Systematic Literature Review. In 2024 International Research Conference on Smart Computing and Systems Engineering (SCSE) (Vol. 7, pp. 1-5). IEEE.

Reisz, N., Servedio, V.D. and Thurner, S., 2022. To what extent homophily and influencer networks explain song popularity. arXiv preprint arXiv:2211.15164.

Sandag, G.A. and Manueke, A.M., 2020, October. Predictive models for popularity of solo and group singers in Spotify using decision tree. In 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS) (pp. 1-5). IEEE.

Sebastian, N. and Mayer, F., 2024. Beyond Beats: A Recipe to Song Popularity? A machine learning approach. arXiv preprint arXiv:2403.12079.

Song, Y., & Zhao, X. (2019). Machine Learning Models for Music Popularity Forecasting: A Case Study on Chinese Streaming Platforms. Journal of Music Technology and Applications.

Spotify, 2023. Spotify API Terms of Use. Available at: https://developer.spotify.com/ [Accessed 23 Dec. 2024].

Spotify. (2023). Spotify API terms of use. Retrieved from https://developer.spotify.com/

Suh, B.J., 2019. International music preferences: An analysis of the determinants of song popularity on Spotify for the US, Norway, Taiwan, Ecuador, and Costa Rica.

Vötter, M., Mayerl, M., Specht, G. and Zangerle, E., 2021, November. Novel datasets for evaluating song popularity prediction tasks. In 2021 IEEE International Symposium on Multimedia (ISM) (pp. 166-173). IEEE.

Wikipedia. (2023). Wikipedia API documentation. Retrieved from https://www.mediawiki.org/wiki/API:Main_page

Wu, L., Ma, K., & Zhang, H. (2022). Exploring the Role of Collaborative Networks in Music Popularity Using Machine Learning. ACM Transactions on Multimedia Computing.

Xu, Y. and Xu, S., 2021. A clustering analysis method for massive music data. Modern electronic technology, 5(1), p.8.

Yee, Y.K. and Raheem, M., 2022. Predicting music popularity using spotify and youtube features. Indian Journal of Science and Technology, 15(36), pp.1786-1799.

Yee, Y.K., & Raheem, M. (2022). Predicting music popularity using Spotify and YouTube features. Indian Journal of Science and Technology, 15(36), 1786–1799.

Zhao, M., Harvey, M., Cameron, D., Hopfgartner, F. and Gillet, V.J., 2023, March. An analysis of classification approaches for hit song prediction using engineered metadata features with lyrics and audio features. In International Conference on Information (pp. 303-311). Cham: Springer Nature Switzerland.

Zhao, M., Harvey, M., Cameron, D., et al. (2023). An analysis of classification approaches for hit song prediction using engineered metadata features with lyrics and audio features. Springer Nature Switzerland.

de Oliveira, G.P., 2021. Analyses of musical success based on time, genre and collaboration.

12 Appendices

Table 9. KMeans Descriptive Statistics

popularity:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	40.350669	22.146336	0.0	24.0	36.0	58.0	96.0
1	2588.0	62.444745	11.294475	0.0	55.0	61.0	70.0	98.0
2	2707.0	29.719616	13.242524	0.0	22.0	31.0	40.0	58.0
3	2252.0	29.464476	13.206963	0.0	21.0	30.0	40.0	68.0

duration_ms:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	208414.307578	74173.051864	31533.0	167151.00	200600.0	243288.00	609226.0
1	2588.0	225176.942427	66883.541701	45760.0	184179.50	217133.0	253611.75	643127.0
2	2707.0	232128.531215	80934.620627	32986.0	183926.50	220305.0	266303.50	859345.0
3	2252.0	251434.238455	152253.251252	30680.0	184487.75	230381.0	289319.75	4563897.0

explicit:

kmeans cluster	count	unique	top	freq
0	673	2	True	622
1	2588	2	False	2587
2	2707	1	False	2707
3	2252	2	False	2250

danceability:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	0.569588	0.195100	0.0000	0.435	0.576	0.724	0.966
1	2588.0	0.568262	0.154802	0.0748	0.467	0.573	0.679	0.970
2	2707.0	0.549920	0.162075	0.0588	0.441	0.560	0.671	0.934
3	2252.0	0.529079	0.183114	0.0513	0.401	0.546	0.667	0.982

energy:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	0.745986	0.216942	0.00066	0.61000	0.7940	0.938	1.000
1	2588.0	0.633546	0.245057	0.00209	0.45700	0.6725	0.842	0.998
2	2707.0	0.715333	0.225090	0.02990	0.55800	0.7680	0.909	1.000
3	2252.0	0.653172	0.257597	0.00181	0.47075	0.6980	0.880	0.999

key:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	5.315007	3.480648	0.0	2.0	5.0	8.0	11.0
1	2588.0	5.214838	3.546508	0.0	2.0	5.0	8.0	11.0
2	2707.0	4.825268	3.508955	0.0	2.0	5.0	8.0	11.0
3	2252.0	5.748668	3.458310	0.0	3.0	6.0	9.0	11.0

loudness:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	-7.026371	4.088721	-35.120	-8.55300	-6.1430	-4.39800	-1.115
1	2588.0	-8.164412	4.544006	-37.859	-10.04000	-7.2095	-5.04900	1.023
2	2707.0	-7.354786	3.449500	-23.405	-9.24750	-6.7750	-4.94550	1.028
3	2252.0	-8.946608	5.118648	-43.957	-11.11675	-7.7725	-5.46775	4.532

mode:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	0.589896	0.492218	0.0	0.0	1.0	1.0	1.0
1	2588.0	0.633308	0.481995	0.0	0.0	1.0	1.0	1.0
2	2707.0	0.990395	0.097550	0.0	1.0	1.0	1.0	1.0
3	2252.0	0.193162	0.394866	0.0	0.0	0.0	0.0	1.0

speechiness:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	0.204520	0.241029	0.0000	0.056800	0.1080	0.247000	0.961
1	2588.0	0.061526	0.055939	0.0232	0.033500	0.0422	0.063525	0.532
2	2707.0	0.084000	0.110471	0.0230	0.035350	0.0485	0.082500	0.947
3	2252.0	0.072095	0.067136	0.0228	0.036175	0.0485	0.076100	0.762

acousticness:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	0.227468	0.301449	0.000001	0.001860	0.0635	0.406	0.994
1	2588.0	0.288905	0.319462	0.000001	0.015875	0.1440	0.534	0.996
2	2707.0	0.243215	0.299035	0.000000	0.003980	0.0821	0.453	0.994
3	2252.0	0.293087	0.339741	0.000000	0.004640	0.1125	0.564	0.996

instrumentalness:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	0.103724	0.253039	0.0	0.000000	0.000007	0.00572	0.973
1	2588.0	0.134698	0.283568	0.0	0.000000	0.000102	0.03075	0.985
2	2707.0	0.125735	0.269389	0.0	0.000000	0.000051	0.02750	0.983
3	2252.0	0.318796	0.392516	0.0	0.000006	0.009450	0.80100	0.988

liveness:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	0.265609	0.231343	0.0116	0.1020	0.1640	0.35400	0.984
1	2588.0	0.180446	0.146559	0.0222	0.0938	0.1210	0.22425	0.980
2	2707.0	0.226655	0.195891	0.0145	0.0985	0.1470	0.30600	0.988
3	2252.0	0.194796	0.160365	0.0176	0.0943	0.1275	0.25700	0.974

valence:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	0.453450	0.252098	0.00000	0.25000	0.4390	0.64200	0.980
1	2588.0	0.488430	0.251716	0.02780	0.28200	0.4780	0.69025	0.992
2	2707.0	0.531666	0.263863	0.02630	0.32200	0.5410	0.75150	0.983
3	2252.0	0.462352	0.279562	0.00785	0.21775	0.4435	0.69700	0.984

tempo:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	120.479520	31.691840	0.000	94.97800	119.0020	142.07900	217.282
1	2588.0	121.965876	29.050788	48.718	99.97800	120.8665	139.84350	210.534
2	2707.0	125.944362	29.946231	42.186	102.65000	124.8060	144.82650	215.149
3	2252.0	123.968973	30.065872	54.669	100.22475	123.0040	142.08875	215.513

time_signature:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	3.720654	0.870723	0.0	4.0	4.0	4.0	5.0
1	2588.0	3.910355	0.311612	3.0	4.0	4.0	4.0	5.0
2	2707.0	4.004064	0.161930	3.0	4.0	4.0	4.0	5.0
3	2252.0	3.878330	0.367893	3.0	4.0	4.0	4.0	5.0

listeners:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	266366.814264	589389.286882	0.0	7521.00	35120.0	200571.00	4732528.0
1	2588.0	448029.212519	641700.943214	0.0	43960.25	181025.5	577418.25	5381567.0
2	2707.0	114614.691910	255744.463295	0.0	8106.50	34263.0	106324.50	3820581.0
3	2252.0	87895.124334	153449.883909	0.0	8389.00	31704.0	101625.00	1689609.0

Continent:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	1.809807	1.037722	0.0	1.0	1.0	3.0	4.0
1	2588.0	1.916151	0.982019	0.0	1.0	2.0	3.0	4.0
2	2707.0	1.933875	0.992053	0.0	1.0	2.0	3.0	4.0
3	2252.0	2.003996	0.985223	0.0	1.0	2.0	3.0	4.0

is_popular:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	0.335810	0.472624	0.0	0.0	0.0	1.0	1.0
1	2588.0	0.970247	0.169937	0.0	1.0	1.0	1.0	1.0
2	2707.0	0.002216	0.047036	0.0	0.0	0.0	0.0	1.0
3	2252.0	0.006217	0.078618	0.0	0.0	0.0	0.0	1.0

genre_numeric:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	673.0	6.353640	3.668932	0.0	3.0	8.0	10.0	11.0
1	2588.0	7.110896	3.292538	0.0	4.0	8.0	10.0	11.0
2	2707.0	6.393424	3.477517	0.0	2.0	8.0	9.0	11.0
3	2252.0	6.218028	3.463714	0.0	3.0	8.0	10.0	11.0

Table 10. DBSCAN Descriptive Statistics

popularity:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	53.800000	2.774887	51.0	52.0	53.0	55.0	58.0
0	7517.0	40.909804	20.057391	0.0	26.0	40.0	56.0	98.0
1	49.0	36.693878	20.614724	0.0	19.0	39.0	54.0	73.0
2	565.0	42.346903	22.684000	0.0	26.0	38.0	61.0	96.0
3	84.0	24.476190	3.906979	19.0	22.0	24.0	27.0	39.0

duration_ms:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	158575.400000	45644.833906	101635.0	125384.0	164213.0	186885.0	214760.0
0	7517.0	235686.568312	104458.162029	30680.0	184240.0	222240.0	267306.0	4563897.0
1	49.0	217187.877551	115259.684623	31533.0	131480.0	200125.0	258106.0	590666.0
2	565.0	209340.814159	67119.812452	47674.0	171746.0	202000.0	241767.0	609226.0
3	84.0	193500.214286	96256.235850	32986.0	123546.0	181795.5	261451.0	490400.0

explicit:

dbscan cluster	count	unique	top	freq
-1	5	1	False	5
0	7517	2	False	7514
1	49	2	False	45
2	565	1	True	565
3	84	2	True	53

danceability:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	0.711400	0.097784	0.6000	0.61800	0.7340	0.80100	0.804
0	7517.0	0.549890	0.167183	0.0513	0.44000	0.5610	0.67300	0.982
1	49.0	0.385878	0.188637	0.0000	0.22500	0.4020	0.48400	0.757
2	565.0	0.587526	0.195296	0.1270	0.44500	0.6000	0.74500	0.966
3	84.0	0.549762	0.089446	0.2520	0.49925	0.5575	0.62025	0.719

energy:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	0.374800	0.070205	0.29500	0.335	0.358	0.410	0.476
0	7517.0	0.668860	0.244739	0.00181	0.496	0.715	0.880	1.000
1	49.0	0.440440	0.351760	0.00066	0.116	0.433	0.727	0.998
2	565.0	0.772814	0.181132	0.22500	0.640	0.807	0.944	1.000
3	84.0	0.727357	0.191679	0.33000	0.585	0.762	0.896	0.994

key:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	7.000000	5.522681	0.0	2.0	11.0	11.00	11.0
0	7517.0	5.234668	3.524227	0.0	2.0	5.0	8.00	11.0
1	49.0	5.653061	3.326353	0.0	3.0	6.0	8.00	11.0
2	565.0	5.362832	3.554694	0.0	2.0	5.0	9.00	11.0
3	84.0	4.642857	3.068074	0.0	2.0	4.0	6.25	11.0

loudness:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	-11.646400	2.540045	-14.452	-13.480	-12.2680	-9.434	-8.598
0	7517.0	-8.088926	4.417759	-43.957	-10.024	-7.1650	-5.116	4.532
1	49.0	-13.472714	7.566973	-35.120	-19.221	-11.1760	-7.432	-3.373
2	565.0	-6.092869	2.528608	-13.964	-7.861	-5.7560	-4.202	-1.115
3	84.0	-11.310631	4.961812	-22.709	-14.532	-11.0385	-6.750	-2.660

mode:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	1.000000	0.000000	1.0	1.00	1.0	1.0	1.0
0	7517.0	0.629506	0.482969	0.0	0.00	1.0	1.0	1.0
1	49.0	0.632653	0.487078	0.0	0.00	1.0	1.0	1.0
2	565.0	0.568142	0.495774	0.0	0.00	1.0	1.0	1.0
3	84.0	0.750000	0.435613	0.0	0.75	1.0	1.0	1.0

speechiness:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	0.415180	0.382319	0.0472	0.0947	0.3000	0.74500	0.889
0	7517.0	0.069747	0.066513	0.0228	0.0346	0.0459	0.07350	0.762
1	49.0	0.064516	0.063855	0.0000	0.0356	0.0428	0.05720	0.353
2	565.0	0.144714	0.115290	0.0242	0.0588	0.1050	0.20900	0.724
3	84.0	0.896798	0.078224	0.5240	0.8855	0.9220	0.93725	0.961

acousticness:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	0.609000	0.172035	0.335000	0.57200	0.6360	0.7300	0.772
0	7517.0	0.271785	0.318374	0.000000	0.00718	0.1070	0.5090	0.996
1	49.0	0.587707	0.402993	0.000001	0.04590	0.7540	0.9480	0.994
2	565.0	0.138121	0.201282	0.000001	0.00102	0.0297	0.2070	0.897
3	84.0	0.789226	0.136429	0.134000	0.74550	0.8160	0.8705	0.985

instrumentalness:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	0.476405	0.438964	0.0	0.000026	0.737000	0.75400	0.891000
0	7517.0	0.186844	0.327371	0.0	0.000000	0.000248	0.19700	0.988000
1	49.0	0.375763	0.415782	0.0	0.000017	0.065200	0.87100	0.973000
2	565.0	0.090961	0.232076	0.0	0.000000	0.000012	0.00564	0.965000
3	84.0	0.000027	0.000129	0.0	0.000000	0.000000	0.00000	0.000869

liveness:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	0.147440	0.118887	0.0741	0.0971	0.102	0.1050	0.359
0	7517.0	0.199626	0.168108	0.0145	0.0953	0.129	0.2650	0.988
1	49.0	0.214131	0.177490	0.0116	0.1000	0.114	0.3090	0.713
2	565.0	0.217501	0.169394	0.0197	0.0982	0.148	0.3120	0.983
3	84.0	0.753361	0.191867	0.0803	0.6580	0.777	0.9205	0.984

valence:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	0.374160	0.345829	0.03880	0.0800	0.276	0.688	0.788
0	7517.0	0.496482	0.266177	0.00785	0.2730	0.494	0.718	0.992
1	49.0	0.331912	0.289976	0.00000	0.1000	0.229	0.594	0.980
2	565.0	0.468773	0.249666	0.02150	0.2780	0.460	0.656	0.974
3	84.0	0.412133	0.211888	0.03510	0.2545	0.391	0.562	0.873

tempo:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	133.355200	51.419886	77.998	95.0160	122.033	171.73900	199.990
0	7517.0	124.061565	29.684780	42.186	100.9320	123.016	141.98600	215.513
1	49.0	111.880388	39.046228	0.000	86.1160	109.605	139.43800	217.282
2	565.0	123.511573	30.064708	64.086	96.9860	120.071	144.19800	213.778
3	84.0	99.240262	29.689923	55.346	78.6615	91.346	115.58075	180.710

time_signature:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	3.800000	0.447214	3.0	4.0	4.0	4.0	4.0
0	7517.0	3.936677	0.289946	3.0	4.0	4.0	4.0	5.0
1	49.0	0.959184	0.199915	0.0	1.0	1.0	1.0	1.0
2	565.0	3.980531	0.217928	3.0	4.0	4.0	4.0	5.0
3	84.0	3.452381	0.936602	1.0	3.0	3.5	4.0	5.0

listeners:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	47718.400000	43291.244182	873.0	3951.0	57805.0	83373.0	92590.0
0	7517.0	221739.962352	446447.910537	0.0	12878.0	56036.0	205244.0	5381567.0
1	49.0	140042.265306	195502.363150	333.0	14348.0	51050.0	205405.0	918197.0
2	565.0	302732.844248	633766.963315	0.0	8467.0	39038.0	258246.0	4732528.0
3	84.0	23836.214286	28271.265317	74.0	2513.5	10997.0	37635.5	114167.0

Continent:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	1.800000	0.836660	1.0	1.0	2.0	2.0	3.0
0	7517.0	1.950246	0.987104	0.0	1.0	2.0	3.0	4.0
1	49.0	1.979592	1.050591	1.0	1.0	1.0	3.0	4.0
2	565.0	1.858407	1.052414	0.0	1.0	1.0	3.0	4.0
3	84.0	1.297619	0.740885	1.0	1.0	1.0	1.0	4.0

is_popular:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	1.000000	0.000000	1.0	1.0	1.0	1.0	1.0
0	7517.0	0.336038	0.472384	0.0	0.0	0.0	1.0	1.0
1	49.0	0.367347	0.487078	0.0	0.0	0.0	1.0	1.0
2	565.0	0.368142	0.482727	0.0	0.0	0.0	1.0	1.0
3	84.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0

genre_numeric:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
-1	5.0	5.800000	4.549725	1.0	2.0	5.0	10.0	11.0
0	7517.0	6.609020	3.417154	0.0	3.0	8.0	10.0	11.0
1	49.0	5.428571	3.662877	0.0	3.0	5.0	10.0	11.0
2	565.0	7.093805	3.176272	0.0	4.0	8.0	10.0	11.0
3	84.0	0.071429	0.373383	0.0	0.0	0.0	0.0	2.0

popularity:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	37.581882	17.862475	0.0	25.0	37.0	52.0	70.0
1	622.0	34.987138	18.086832	0.0	23.0	32.0	50.0	73.0
2	678.0	76.877581	5.093924	71.0	73.0	76.0	80.0	98.0
3	3189.0	38.088743	17.574611	0.0	25.0	39.0	53.0	70.0

duration:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	3.805892	1.311434	0.511333	3.012742	3.625267	4.376108	14.322417
1	622.0	3.460789	1.346925	0.525550	2.675975	3.312658	4.103246	10.979333
2	678.0	3.762637	0.874626	1.142033	3.225213	3.644217	4.208896	8.679767
3	3189.0	4.094361	2.218123	0.621550	3.117800	3.811933	4.620000	76.064950

explicit:

kmeans cluster	count	unique	top	freq
0	3731	1	False	3731
1	622	2	True	546
2	678	2	False	600
3	3189	2	False	3188

danceability:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	0.556122	0.161789	0.0566	0.448	0.5680	0.67400	0.934
1	622.0	0.555712	0.192128	0.0000	0.432	0.5565	0.69625	0.966
2	678.0	0.597732	0.148268	0.1500	0.502	0.5960	0.70750	0.951
3	3189.0	0.535689	0.175724	0.0513	0.421	0.5450	0.66500	0.982

energy:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	0.664182	0.244935	0.00815	0.48700	0.7030	0.87600	1.000
1	622.0	0.748615	0.221073	0.00066	0.60925	0.8005	0.94300	1.000
2	678.0	0.677645	0.193847	0.04000	0.53525	0.7110	0.83775	0.986
3	3189.0	0.672897	0.252576	0.00181	0.49700	0.7290	0.89400	0.999

key:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	4.917448	3.494560	0.0	2.0	5.0	8.0	11.0
1	622.0	5.308682	3.501261	0.0	2.0	5.0	8.0	11.0
2	678.0	5.253687	3.485858	0.0	2.0	5.0	8.0	11.0
3	3189.0	5.603638	3.532028	0.0	2.0	6.0	9.0	11.0

loudness:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	-8.173219	4.231870	-32.190	-10.13600	-7.2840	-5.22400	1.028
1	622.0	-7.373432	4.379063	-35.120	-8.88350	-6.3025	-4.48600	-1.115
2	678.0	-7.118260	3.210185	-22.390	-8.91375	-6.4880	-4.87975	-1.399
3	3189.0	-8.155748	4.786352	-43.957	-10.15400	-7.1520	-4.97400	4.532

mode:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	0.993835	0.078283	0.0	1.0	1.0	1.0	1.0
1	622.0	0.593248	0.491623	0.0	0.0	1.0	1.0	1.0
2	678.0	0.634218	0.482004	0.0	0.0	1.0	1.0	1.0
3	3189.0	0.202258	0.401746	0.0	0.0	0.0	0.0	1.0

speechiness:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	0.070070	0.068449	0.0230	0.034150	0.0449	0.07295	0.727
1	622.0	0.242156	0.280692	0.0000	0.060625	0.1155	0.29300	0.961
2	678.0	0.067167	0.067653	0.0239	0.034100	0.0439	0.06845	0.724
3	3189.0	0.071818	0.069179	0.0228	0.035700	0.0480	0.07650	0.889

acousticness:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	0.286336	0.319584	0.000000	0.009075	0.1330	0.54600	0.996
1	622.0	0.257425	0.324278	0.000001	0.001573	0.0801	0.47925	0.994
2	678.0	0.217455	0.251062	0.000005	0.012950	0.1045	0.34200	0.991
3	3189.0	0.264444	0.326653	0.000000	0.004130	0.0784	0.49600	0.996

instrumentalness:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	0.170229	0.314147	0.0	0.000000	0.000126	0.122500	0.988
1	622.0	0.111965	0.263421	0.0	0.000000	0.000005	0.007112	0.973
2	678.0	0.040200	0.155112	0.0	0.000000	0.000011	0.000967	0.982
3	3189.0	0.233515	0.354974	0.0	0.000002	0.001300	0.511000	0.987

liveness:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	0.209331	0.182164	0.0145	0.09665	0.132	0.27600	0.988
1	622.0	0.293003	0.252479	0.0116	0.10400	0.182	0.37675	0.984
2	678.0	0.166392	0.124846	0.0222	0.08740	0.119	0.20300	0.839
3	3189.0	0.195017	0.157086	0.0176	0.09540	0.129	0.26400	0.982

valence:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	0.518837	0.268069	0.00785	0.29550	0.5200	0.74500	0.992
1	622.0	0.438528	0.250341	0.00000	0.23925	0.4245	0.61475	0.980
2	678.0	0.546752	0.239838	0.03640	0.36175	0.5435	0.74525	0.979
3	3189.0	0.461092	0.264347	0.02350	0.23500	0.4430	0.67700	0.984

tempo:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	123.655553	29.620073	42.186	100.27550	123.136	140.25750	215.149
1	622.0	120.119854	32.370746	0.000	94.87625	117.925	142.04925	217.282
2	678.0	121.222416	28.234104	48.718	100.00525	118.718	139.87650	206.247
3	3189.0	124.985170	29.993485	54.669	101.61200	123.905	144.20300	215.513

time_signature:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	4.013401	0.140212	3.0	4.0	4.0	4.0	5.0
1	622.0	3.667203	0.906330	0.0	4.0	4.0	4.0	5.0
2	678.0	3.957227	0.229827	3.0	4.0	4.0	4.0	5.0
3	3189.0	3.844152	0.386221	3.0	4.0	4.0	4.0	5.0

listeners:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	1.651453e+05	303738.595135	0.0	12712.5	52966.0	170276.00	3820581.0
1	622.0	1.386853e+05	291743.283890	0.0	6033.5	27855.5	114928.25	2485167.0
2	678.0	1.031348e+06	979932.780135	0.0	199710.0	759783.5	1606147.25	5381567.0
3	3189.0	1.396341e+05	253430.795176	0.0	9930.0	42766.0	144016.00	2495391.0

Continent:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	0.592603	0.816183	0.0	0.0	0.0	1.0	4.0
1	622.0	0.636656	0.976276	0.0	0.0	0.0	1.0	4.0
2	678.0	0.587021	0.940132	0.0	0.0	0.0	1.0	4.0
3	3189.0	1.099091	1.190812	0.0	0.0	1.0	1.0	4.0

is_popular:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
1	622.0	0.003215	0.056659	0.0	0.0	0.0	0.0	1.0
2	678.0	1.000000	0.000000	1.0	1.0	1.0	1.0	1.0
3	3189.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0

genre_classification:

kmeans cluster	count	mean	std	min	25%	50%	75%	max
0	3731.0	6.667113	3.863785	0.0	3.0	8.0	10.0	11.0
1	622.0	6.081994	4.238999	0.0	1.0	8.0	10.0	11.0
2	678.0	7.721239	3.194418	0.0	6.0	8.0	11.0	11.0
3	3189.0	6.582314	3.813716	0.0	3.0	8.0	10.0	11.0

Table 10. DBSCAN Clustering Description

popularity:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	37.815462	17.730920	0.0	25.0	38.0	52.0	70.0
1	600.0	76.695000	4.934155	71.0	73.0	76.0	80.0	98.0
2	49.0	36.693878	20.614724	0.0	19.0	39.0	54.0	73.0
3	489.0	36.621677	18.705823	0.0	24.0	35.0	52.0	70.0
4	84.0	24.476190	3.906979	19.0	22.0	24.0	27.0	39.0
5	78.0	78.282051	6.040765	71.0	74.0	76.0	81.0	96.0

duration:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	3.938829	1.792992	0.511333	3.057021	3.706667	4.482783	76.064950
1	600.0	3.774957	0.881580	1.142033	3.228750	3.653133	4.212217	8.679767
2	49.0	3.619798	1.920995	0.525550	2.191333	3.335417	4.301767	9.844433
3	489.0	3.485359	1.222797	0.794567	2.821550	3.343333	4.035317	10.979333
4	84.0	3.225004	1.604271	0.549767	2.059100	3.029925	4.357517	8.173333
5	78.0	3.667866	0.818301	2.184400	3.203883	3.595825	4.030550	5.908667

explicit:

dbscan cluster	count	unique	top	freq
0	6920	2	False	6919
1	600	1	False	600
2	49	2	False	45
3	489	1	True	489
4	84	2	True	53
5	78	1	True	78

danceability:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	0.546706	0.168649	0.0513	0.43600	0.5590	0.67100	0.982
1	600.0	0.588152	0.142914	0.1740	0.49950	0.5895	0.68825	0.951
2	49.0	0.385878	0.188637	0.0000	0.22500	0.4020	0.48400	0.757
3	489.0	0.573753	0.197192	0.1270	0.43200	0.5770	0.73100	0.966
4	84.0	0.549762	0.089446	0.2520	0.49925	0.5575	0.62025	0.719
5	78.0	0.671423	0.167871	0.1500	0.57400	0.7000	0.80025	0.950

energy:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	0.668198	0.248505	0.00181	0.4920	0.7160	0.88400	1.000
1	600.0	0.674713	0.196993	0.04000	0.5280	0.7095	0.83800	0.986
2	49.0	0.440440	0.351760	0.00066	0.1160	0.4330	0.72700	0.998
3	489.0	0.783147	0.181562	0.24300	0.6470	0.8260	0.95000	1.000
4	84.0	0.727357	0.191679	0.33000	0.5850	0.7620	0.89600	0.994
5	78.0	0.700205	0.167077	0.22500	0.5845	0.7325	0.83375	0.950

key:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	5.233671	3.528243	0.0	2.00	5.0	8.00	11.0
1	600.0	5.260000	3.504216	0.0	2.00	5.0	8.00	11.0
2	49.0	5.653061	3.326353	0.0	3.00	6.0	8.00	11.0
3	489.0	5.388548	3.579935	0.0	2.00	5.0	9.00	11.0
4	84.0	4.642857	3.068074	0.0	2.00	4.0	6.25	11.0
5	78.0	5.205128	3.362741	0.0	2.25	5.5	8.00	11.0

loudness:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	-8.165167	4.495577	-43.957	-10.14725	-7.2280	-5.12800	4.532
1	600.0	-7.227792	3.276513	-22.390	-9.08150	-6.5435	-4.94275	-1.399
2	49.0	-13.472714	7.566973	-35.120	-19.22100	-11.1760	-7.43200	-3.373
3	489.0	-6.085928	2.553111	-13.958	-7.86100	-5.7150	-4.21300	-1.115
4	84.0	-11.310631	4.961812	-22.709	-14.53200	-11.0385	-6.75000	-2.660
5	78.0	-6.275705	2.505893	-13.964	-7.97625	-5.9900	-4.11075	-2.052

mode:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	0.629046	0.483095	0.0	0.00	1.0	1.0	1.0
1	600.0	0.638333	0.480884	0.0	0.00	1.0	1.0	1.0
2	49.0	0.632653	0.487078	0.0	0.00	1.0	1.0	1.0
3	489.0	0.562372	0.496602	0.0	0.00	1.0	1.0	1.0
4	84.0	0.750000	0.435613	0.0	0.75	1.0	1.0	1.0
5	78.0	0.602564	0.492535	0.0	0.00	1.0	1.0	1.0

speechiness:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	0.070875	0.068787	0.0228	0.034800	0.04610	0.074400	0.889
1	600.0	0.059612	0.052063	0.0239	0.033375	0.04205	0.062325	0.463
2	49.0	0.064516	0.063855	0.0000	0.035600	0.04280	0.057200	0.353
3	489.0	0.147502	0.113600	0.0242	0.061300	0.10700	0.211000	0.610
4	84.0	0.896798	0.078224	0.5240	0.885500	0.92200	0.937250	0.961
5	78.0	0.125286	0.123649	0.0255	0.042700	0.07130	0.140750	0.724

acousticness:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	0.276248	0.323022	0.000000	0.006457	0.1060	0.52525	0.996
1	600.0	0.223998	0.254943	0.000005	0.015775	0.1185	0.35650	0.991
2	49.0	0.587707	0.402993	0.000001	0.045900	0.7540	0.94800	0.994
3	489.0	0.132977	0.198802	0.000001	0.000662	0.0247	0.19200	0.897
4	84.0	0.789226	0.136429	0.134000	0.745500	0.8160	0.87050	0.985
5	78.0	0.167124	0.213715	0.000022	0.008398	0.0558	0.30675	0.768

instrumentalness:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	0.199393	0.335047	0.0	0.000000	0.000353	0.284000	0.988000
1	600.0	0.042374	0.160968	0.0	0.000000	0.000011	0.001172	0.982000
2	49.0	0.375763	0.415782	0.0	0.000017	0.065200	0.871000	0.973000
3	489.0	0.104760	0.249203	0.0	0.000000	0.000014	0.013500	0.965000
4	84.0	0.000027	0.000129	0.0	0.000000	0.000000	0.000000	0.000869
5	78.0	0.023477	0.098252	0.0	0.000000	0.000005	0.000289	0.695000

liveness:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	0.202735	0.171201	0.0145	0.095900	0.130	0.2710	0.988
1	600.0	0.163698	0.121294	0.0222	0.086475	0.119	0.1985	0.839
2	49.0	0.214131	0.177490	0.0116	0.100000	0.114	0.3090	0.713
3	489.0	0.221827	0.171984	0.0197	0.100000	0.154	0.3200	0.983
4	84.0	0.753361	0.191867	0.0803	0.658000	0.777	0.9205	0.984
5	78.0	0.187115	0.148656	0.0327	0.093800	0.128	0.2475	0.776

valence:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	0.492226	0.267892	0.00785	0.26600	0.4870	0.71500	0.992
1	600.0	0.545554	0.240722	0.03640	0.37000	0.5415	0.74375	0.979
2	49.0	0.331912	0.289976	0.00000	0.10000	0.2290	0.59400	0.980
3	489.0	0.453745	0.249701	0.02150	0.25300	0.4360	0.63100	0.974
4	84.0	0.412133	0.211888	0.03510	0.25450	0.3910	0.56200	0.873
5	78.0	0.555967	0.234240	0.06840	0.35625	0.5900	0.74900	0.966

tempo:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	124.268291	29.797952	42.186	100.98875	123.5645	142.03625	215.513
1	600.0	121.726918	28.501144	48.718	100.24675	119.9235	139.90225	206.247
2	49.0	111.880388	39.046228	0.000	86.11600	109.6050	139.43800	217.282
3	489.0	124.532164	30.531905	64.086	98.01500	122.0200	146.93600	213.778
4	84.0	99.240262	29.689923	55.346	78.66150	91.3460	115.58075	180.710
5	78.0	117.341628	25.931031	71.912	94.71675	115.0500	135.76700	184.981

time_signature:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	3.935405	0.294018	3.0	4.0	4.0	4.0	5.0
1	600.0	3.951667	0.236835	3.0	4.0	4.0	4.0	5.0
2	49.0	0.959184	0.199915	0.0	1.0	1.0	1.0	1.0
3	489.0	3.975460	0.229511	3.0	4.0	4.0	4.0	5.0
4	84.0	3.452381	0.936602	1.0	3.0	3.5	4.0	5.0
5	78.0	4.000000	0.161165	3.0	4.0	4.0	4.0	5.0

listeners:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	1.533888e+05	2.819405e+05	0.0	11554.25	47126.0	157318.25	3820581.0
1	600.0	1.008941e+06	9.530238e+05	0.0	207104.00	745786.5	1554844.75	5381567.0
2	49.0	1.400423e+05	1.955024e+05	333.0	14348.00	51050.0	205405.00	918197.0
3	489.0	1.582780e+05	3.189993e+05	0.0	7105.00	30223.0	142649.00	2485167.0
4	84.0	2.383621e+04	2.827127e+04	74.0	2513.50	10997.0	37635.50	114167.0
5	78.0	1.203707e+06	1.159034e+06	0.0	134585.00	932275.0	2029731.50	4732528.0

Continent:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	0.826012	1.037421	0.0	0.0	1.0	1.0	4.0
1	600.0	0.631667	0.975149	0.0	0.0	0.0	1.0	4.0
2	49.0	0.693878	0.983279	0.0	0.0	0.0	1.0	4.0
3	489.0	0.709611	1.012994	0.0	0.0	0.0	1.0	4.0
4	84.0	0.178571	0.541322	0.0	0.0	0.0	0.0	4.0
5	78.0	0.243590	0.488462	0.0	0.0	0.0	0.0	2.0

is_popular:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
1	600.0	1.000000	0.000000	1.0	1.0	1.0	1.0	1.0
2	49.0	0.040816	0.199915	0.0	0.0	0.0	0.0	1.0
3	489.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
4	84.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
5	78.0	1.000000	0.000000	1.0	1.0	1.0	1.0	1.0

genre_classification:

dbscan cluster	count	mean	std	min	25%	50%	75%	max
0	6920.0	6.628035	3.840748	0.0	3.0	8.0	10.0	11.0
1	600.0	7.676667	3.192561	0.0	6.0	8.0	11.0	11.0
2	49.0	5.285714	3.446012	0.0	3.0	5.0	9.0	11.0
3	489.0	7.200409	3.752693	0.0	3.0	9.0	11.0	11.0
4	84.0	0.035714	0.186691	0.0	0.0	0.0	0.0	1.0
5	78.0	8.064103	3.208510	0.0	7.0	9.0	11.0	11.0