Data-Driven Insights: Analyzing Spotify Music Data for Strategic Decision-Making

---

Executive Summary

This report presents an in-depth analytical study of Spotify music data with a focus on supporting decision-making through data-driven insights. The project was undertaken as part of the "Decision Support System" course and aims to demonstrate how music data can be leveraged using statistical and machine learning techniques to inform strategic decisions in music production, curation, and user engagement.

The dataset, sourced from Kaggle, includes two primary components: **track-level metadata** and **artist-level information**. After comprehensive preprocessing and integration of these datasets, the final combined dataset enabled multi-dimensional analysis across audio features, artist popularity, and release trends.

Through **exploratory data analysis (EDA)** and **data visualization**, several key patterns were identified. Popularity was found to have moderate correlations with features such as danceability, energy, and loudness, whereas instrumental or acoustic tracks were generally less popular. Trends over time showed that modern music is becoming more energetic and danceable, aligning with evolving listener preferences.

The report also implements **machine learning models**, specifically **KMeans clustering** to categorize tracks into musical styles and **K-Nearest Neighbors (KNN)** to create a functional recommendation engine based on user-defined preferences. These models simulate real-world decision support systems that can assist streaming platforms, producers, and marketers in making evidence-based choices.

In conclusion, this report highlights the effectiveness of combining data science and machine learning techniques in building intelligent systems that support strategic decisions in the music industry.

TABLE OF CONTENTS

1. INTRODUCTION

In the era of digital transformation, data-driven approaches have become vital for strategic decision-making in various domains, including the music industry. With the widespread use of streaming platforms like Spotify, vast amounts of music-related data are generated every day. This data provides a valuable opportunity to analyze listener behavior, musical trends, and content characteristics.

The objective of this project is to demonstrate how Spotify data can be harnessed to support intelligent decision-making processes using **data analysis**, **visualization**, and **machine learning models**. The project leverages tools and techniques covered in the **Decision Support System (DSS)** course to explore patterns in music characteristics, artist influence, and song popularity.

To facilitate this analysis, a structured and cleaned Spotify dataset was sourced from Kaggle. The dataset includes metadata on tracks and artists, covering over a million records. After performing data preprocessing and integration, the project undertakes a comprehensive analysis and builds a basic recommendation system capable of suggesting songs based on user preferences. The insights gained from this work can guide professionals in content strategy, playlist curation, and personalized music recommendations.

The following subsections detail the data source, cleaning methodology, and the visual tools used to uncover underlying patterns.

1.1 Data Source

The data used for this study was sourced from **Kaggle**, an open data platform. The dataset provides detailed and structured information about Spotify's music catalog, and is divided into two primary files:

- **tracks.csv**: Contains metadata and audio features for individual tracks. Key attributes include:
    - id, name, popularity, release_date, duration_ms, explicit

- Audio features such as danceability, energy, tempo, loudness, acousticness, etc.
- Track-level artist IDs and artist names

- **artists.csv**: Provides metadata about artists on Spotify, including:
  - id, name, popularity, followers, and associated genres

These two datasets were merged using the **artist ID** fields (id_artists in the track dataset and id in the artist dataset), enabling a combined perspective on both tracks and their respective creators. This integration was essential to gain holistic insights into both track-level and artist-level trends.

1.2 Data Preprocessing and Cleaning

Before conducting any meaningful analysis, it was essential to ensure the dataset was clean, consistent, and properly structured. Data preprocessing helps improve the quality of insights and ensures the robustness of any models or visualizations built on top of the data. This stage was carried out using Python's **Pandas** and **NumPy** libraries, which are widely used for data manipulation and cleaning tasks.

The following preprocessing steps were performed:

Removal of Duplicate Records

Duplicate rows were identified and removed from both tracks.csv and artists.csv. This step was critical to avoid data redundancy and to maintain the integrity of statistical results.

Handling Missing Values

Missing or null values were systematically addressed:

- Rows containing **missing values in essential columns** such as popularity, release_date, followers, or audio features were removed.
- For less critical columns or sparse fields, missing values were either **imputed** with defaults (e.g., zero or mean values) or **excluded** from model training.

- **Dropping Irrelevant Columns**

Certain columns that were not useful for the analysis or modeling—such as internal IDs that did not provide analytical value—were dropped to reduce dimensionality and improve focus.

- **Data Type Formatting**

Some columns required data type corrections:

- The release_date column was converted from string to datetime format.
- Numerical fields were explicitly cast to proper numeric types to ensure compatibility with machine learning algorithms.

- **Dataset Merging**

The final step of preprocessing involved merging the two datasets:

- The id_artists column from the tracks.csv file was matched with the id column from the artists.csv file.
- The merged dataset created a unified structure, enabling analysis of track features in conjunction with artist popularity and followers.

After completing these steps, the final dataset was significantly cleaner and more structured, with fewer nulls, no duplicates, and properly formatted features. This provided a reliable foundation for further exploratory analysis and machine learning applications.

1.3 Data Visualization

Data visualization played a pivotal role in uncovering patterns, identifying relationships, and communicating insights derived from the Spotify dataset. Visual analytics enables stakeholders to make faster and more informed decisions by providing a clear view of complex data relationships.

To support the analysis, the following Python libraries were utilized:

- **Matplotlib**: For basic plotting and customization.
- **Seaborn**: For advanced statistical visualizations with built-in aesthetics.
- **Pandas**: For generating inline visual summaries of grouped or aggregated data.

The visualizations focused on key dimensions such as track popularity, feature correlations, temporal trends, and artist influence.

2. ANALYSIS AND INSIGHTS

2.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was carried out to reveal patterns, trends, and relationships within the Spotify dataset. This phase was pivotal in interpreting the data both musically and strategically, setting the stage for data-driven recommendations and potential machine learning applications.

The analysis focused on three key dimensions:

- **Track-level** audio characteristics and popularity
- **Artist-level** influence through followers and engagement
- **Temporal** evolution of music trends.

2.2 Popularity Distribution and Listener Behaviour

The histogram titled *Track Popularity Distribution* illustrates how Spotify tracks are distributed across the platform's popularity scale, which ranges from 0 to 100. The distribution reveals several important patterns in user engagement and listening behaviour.

A significant number of tracks have a popularity score of 0, indicating that a substantial portion of the songs on Spotify receive minimal or no user engagement. Following this, the number of tracks gradually increases, peaking within the popularity range of 20 to 40.

Beyond this peak, the count steadily declines, with only a small fraction of tracks achieving a popularity score above 80.

This pattern reflects a classic *long-tail distribution*, which is typical of digital content platforms. In this case, a small number of highly popular tracks account for a large portion of total user streams, while the majority of available tracks remain relatively underexposed or niche.

From a listener behaviour standpoint, this suggests that user engagement is heavily concentrated around a limited set of hit songs. Spotify's recommendation systems, curated playlists, and algorithmic exposure likely reinforce this disparity by promoting tracks that already perform well. As a result, many songs, despite being available on the platform, struggle to gain visibility or consistent streams.

This insight highlights the competitive nature of digital music streaming and underscores the importance of strategic positioning and exposure for artists seeking higher engagement on platforms like Spotify.

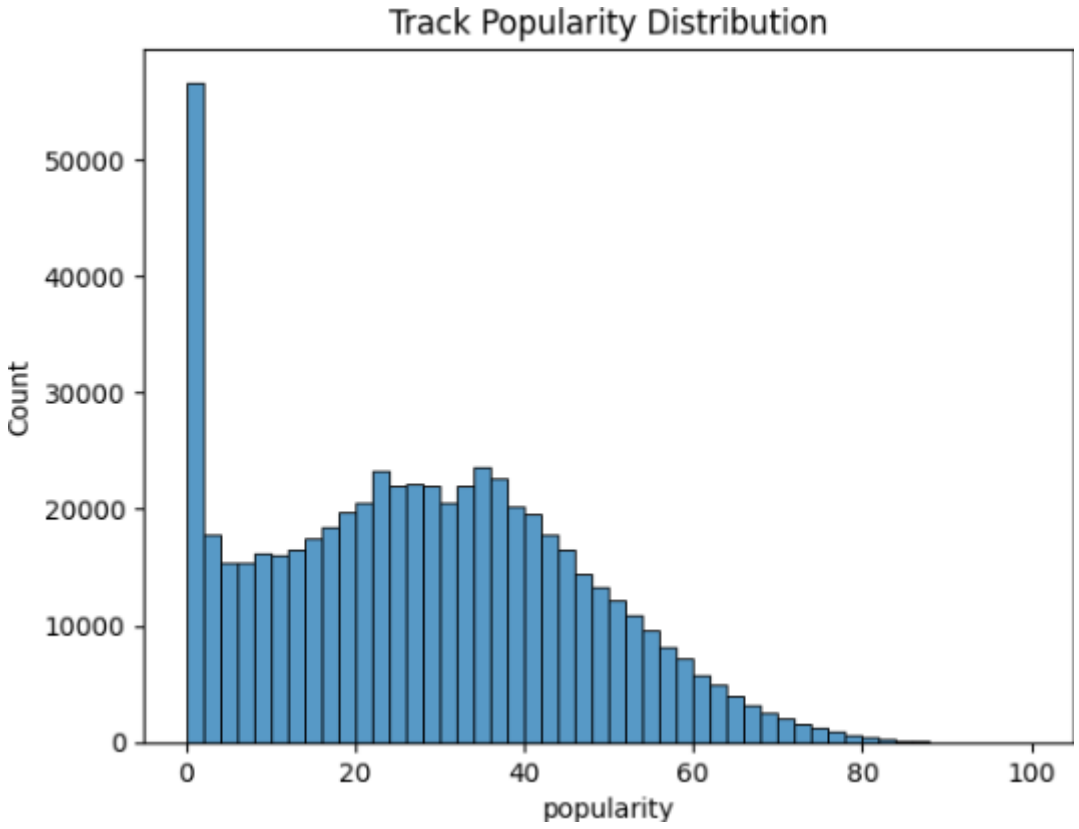

Fig 1: Track Popularity Distribution

2.3 Correlation Analysis of Audio Features

To explore the relationships among audio features and their influence on track popularity, a detailed correlation matrix was generated using Spotify's numerical song attributes. This matrix provides valuable insights into how individual features relate to one another and to overall track performance.

2.3.1 General Observations

The matrix reveals several notable relationships between audio characteristics:

- **Danceability** shows a moderate positive correlation with **valence** (0.53) and **energy** (0.24), suggesting that upbeat and lively songs tend to be more danceable.
- **Energy** and **acousticness** are strongly negatively correlated (−0.72), indicating that high-energy tracks are less likely to be acoustic.
- **Instrumentalness** and **acousticness** exhibit a positive relationship, while both show negative correlations with features like **danceability** and **energy**, reinforcing the contrast between electronic/popular tracks and acoustic/instrumental ones.

2.3.2 Influence on Popularity

From the matrix, it is evident that **no single audio feature has a strong direct correlation with popularity**. The most notable positive relationship is with the feature **energy**, which has a correlation coefficient of **+0.3** with popularity. This suggests that high-energy tracks tend to be slightly more popular, likely due to their appeal in mainstream and upbeat listening contexts.

Other features, such as **danceability** (0.19), **loudness** (implied through energy), and **valence** (0.14), show only weak correlations with popularity. Meanwhile, features like **acousticness**, **instrumentalness**, and **speechiness** have either negligible or slightly negative correlations, indicating limited influence on a track's commercial success.

2.3.3 Key Insight

While certain musical attributes like energy and danceability align loosely with popularity, the overall weak correlations indicate that **popularity is influenced by a combination of factors beyond audio features alone**—including marketing, artist recognition, playlist inclusion, and algorithmic recommendations. This underscores the complex and multifactorial nature of success on streaming platforms like Spotify.

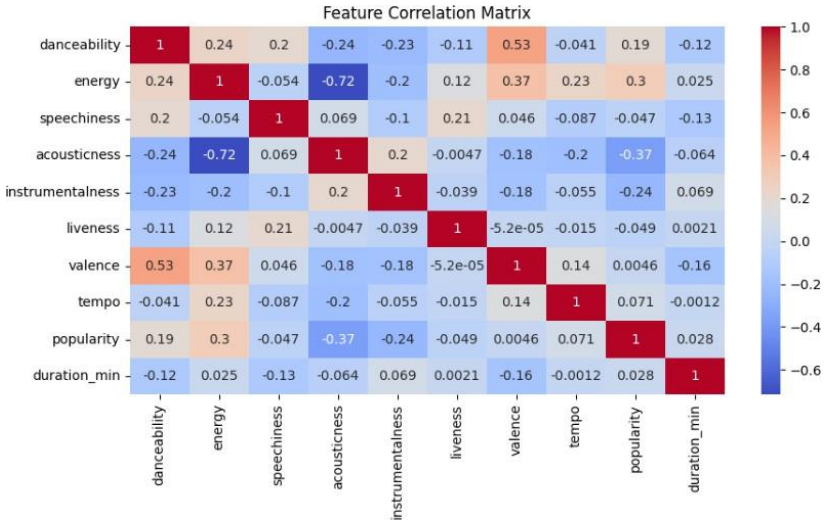

Fig 2: Feature Correlation Matrix

2.4 Impact of Explicit Content on Popularity

A boxplot analysis comparing explicit (1) and non-explicit (0) tracks reveals that **explicit tracks tend to have a higher median popularity** and a wider interquartile range, indicating more variability and a stronger presence among high-popularity songs.

2.4.1 Key Observations

- Explicit tracks show greater clustering at higher popularity levels.
- The upper quartile and whiskers for explicit songs are higher than non-explicit ones.
- Outliers are present in both categories, though explicit songs dominate the upper range.

2.4.2 Interpretation

This trend may reflect the popularity of explicit content in mainstream genres like hip-hop and pop, especially among younger and Western listeners. However, the overlapping distributions suggest that **explicitness alone is not a strong predictor of popularity**—factors like rhythm, artist influence, and marketing play significant roles.

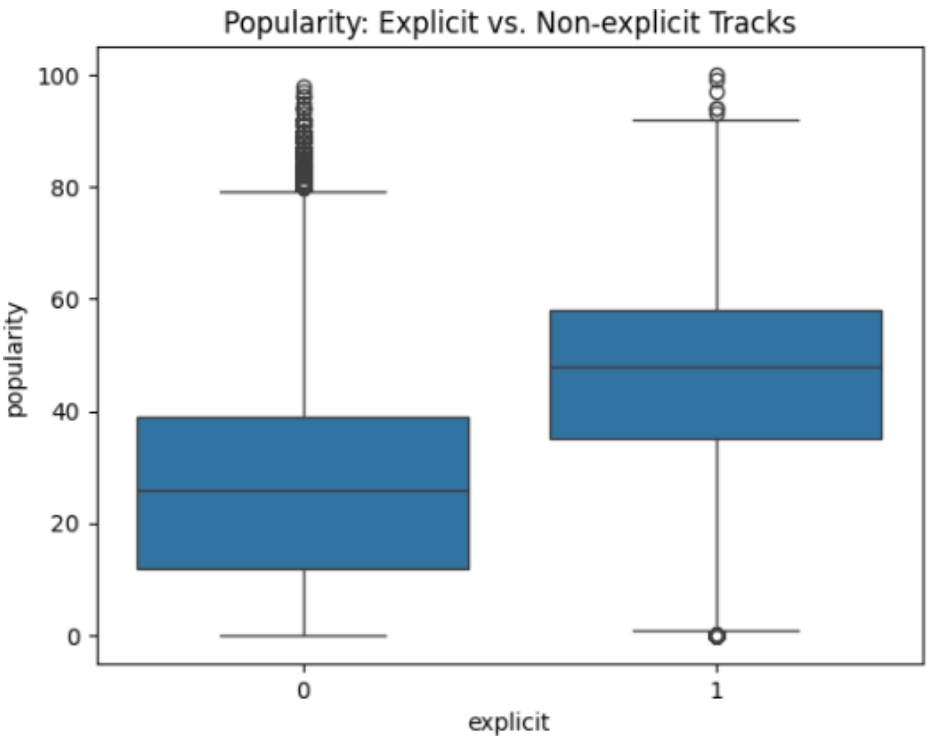

Fig 3: Popularity Explicit vs Non-explicit Tracks

2.5 Temporal Trends in Music Features

An examination of musical characteristics across more than a century reveals significant shifts in the evolution of music, shaped by both cultural changes and technological advancements. This section discusses the temporal trends observed in three key features: popularity, danceability, and energy. These trends reflect not only artistic developments but also changing modes of music consumption and distribution.

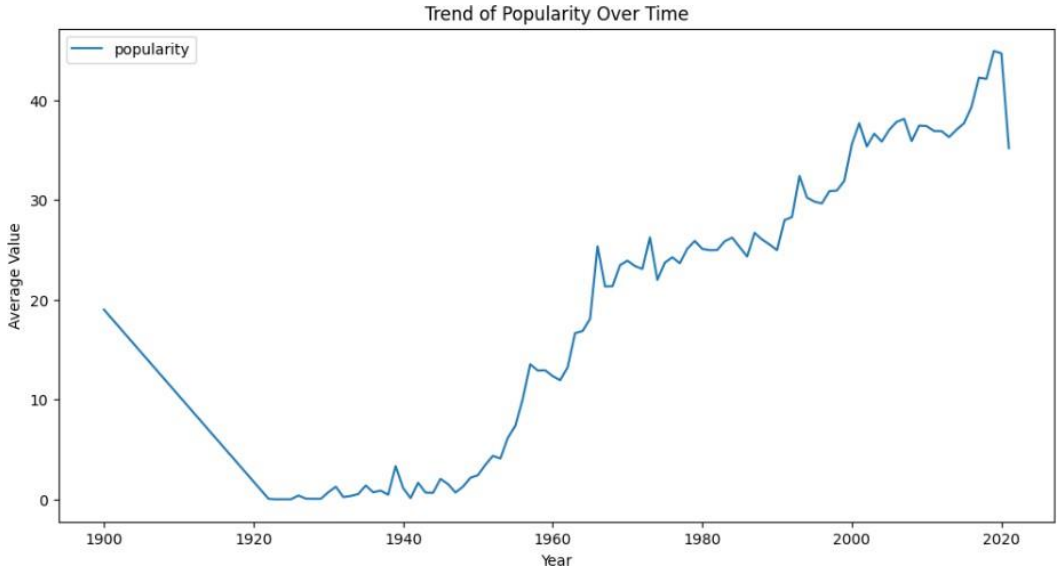

Fig 4: Trend Popularity Over Time

The trend in **popularity** over time illustrates a complex trajectory. From 1900 to around 1950, popularity showed a gradual decline, followed by a period of stagnation. This early trend likely reflects both the limitations of available data and the relatively nascent state of the music recording industry. Beginning in the 1950s, a slow upward movement in popularity is observed, coinciding with the expansion of commercial music through radio broadcasting, vinyl records, and the emergence of globally influential genres such as rock and soul. A more pronounced and sustained increase appears after the year 2000, aligning with the rise of digital music platforms and the transformation of how audiences access and engage with music. The peak in popularity near 2020 corresponds to the widespread use of algorithmic recommendations, streaming playlists, and the global reach of social media—all of which have amplified the visibility and accessibility of popular tracks.

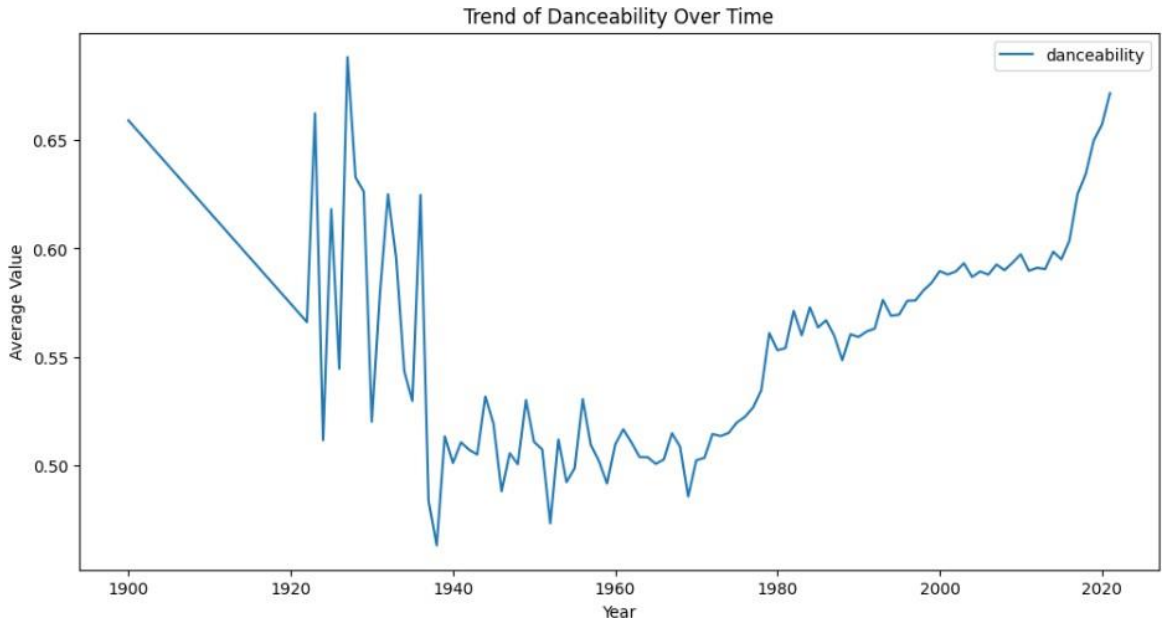

Fig 5: Track of Danceability Over Time

In contrast, the evolution of **danceability** reveals a different yet complementary pattern. While the early to mid-20th century shows inconsistent and erratic changes in danceability, a more discernible trend emerges from the 1980s onward. During this period, the average danceability of tracks begins to rise, corresponding with the growing influence of genres such as disco, funk, hip-hop, and early electronic dance music, all of which emphasize rhythm and movement. This upward trajectory continues steadily into the 21st century. By 2020, danceability reaches its highest values, reflecting a musical landscape increasingly oriented toward rhythmically engaging content. This change aligns with the growing dominance of pop and EDM in mainstream charts, as well as the popularity of short-form video platforms where dance-driven content thrives.

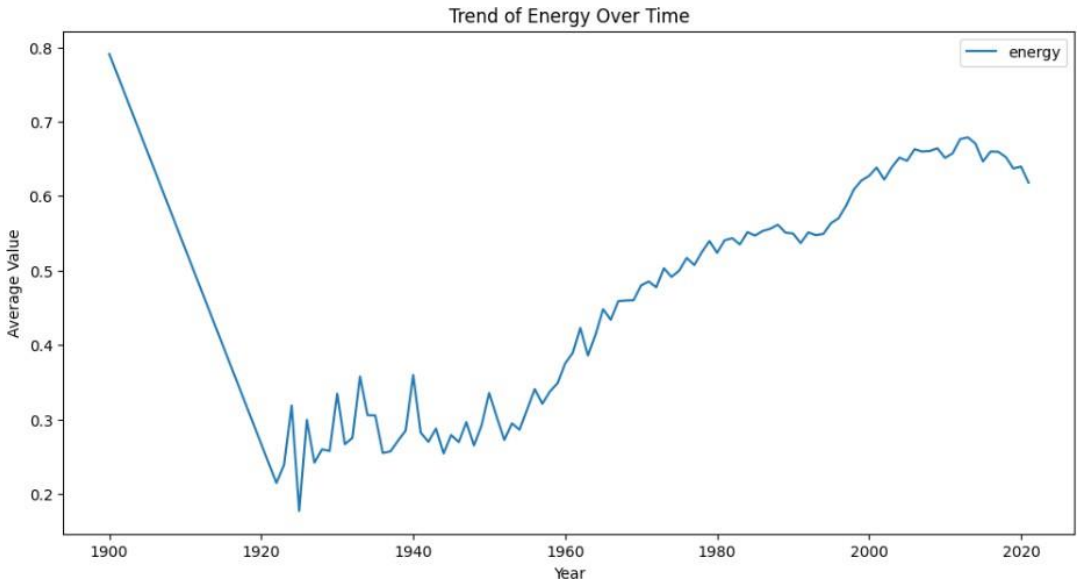

Fig 6: Trend of Energy Over Time

The third feature, **energy**, measures the perceived intensity and dynamic quality of a song. Initially, energy values were relatively low in the early decades of the 20th century, likely due to the acoustic instrumentation and recording limitations of the time. A noticeable shift begins in the 1960s, with energy levels rising steadily through the remainder of the century. This trend mirrors the advent and proliferation of high-intensity genres such as rock, punk, rap, and techno, all of which introduced more aggressive instrumentation and higher tempos. In the digital era, particularly from the early 2000s onwards, energy levels remain consistently high. This reflects both stylistic preferences and modern production techniques that emphasize volume, pace, and sonic impact, designed to maintain listener attention in a competitive streaming environment.

Collectively, the analysis of these three features shows a clear transformation in the character of popular music. The simultaneous rise in popularity, danceability, and energy over recent decades suggests that contemporary music increasingly caters to listeners seeking high engagement and emotional stimulation. These developments have been strongly influenced by the digitalization of music distribution, the role of data-driven personalization in music platforms, and the cultural emphasis on visually and sonically captivating content.

In summary, the temporal progression of musical features reflects a shift toward more energetic, rhythmically driven, and widely appealing compositions. This evolution is tightly coupled with the broader technological context in which music is produced, shared, and consumed, marking a significant departure from the trends observed in earlier periods of recorded music history.

2.6 Growth in Music Production Over Time

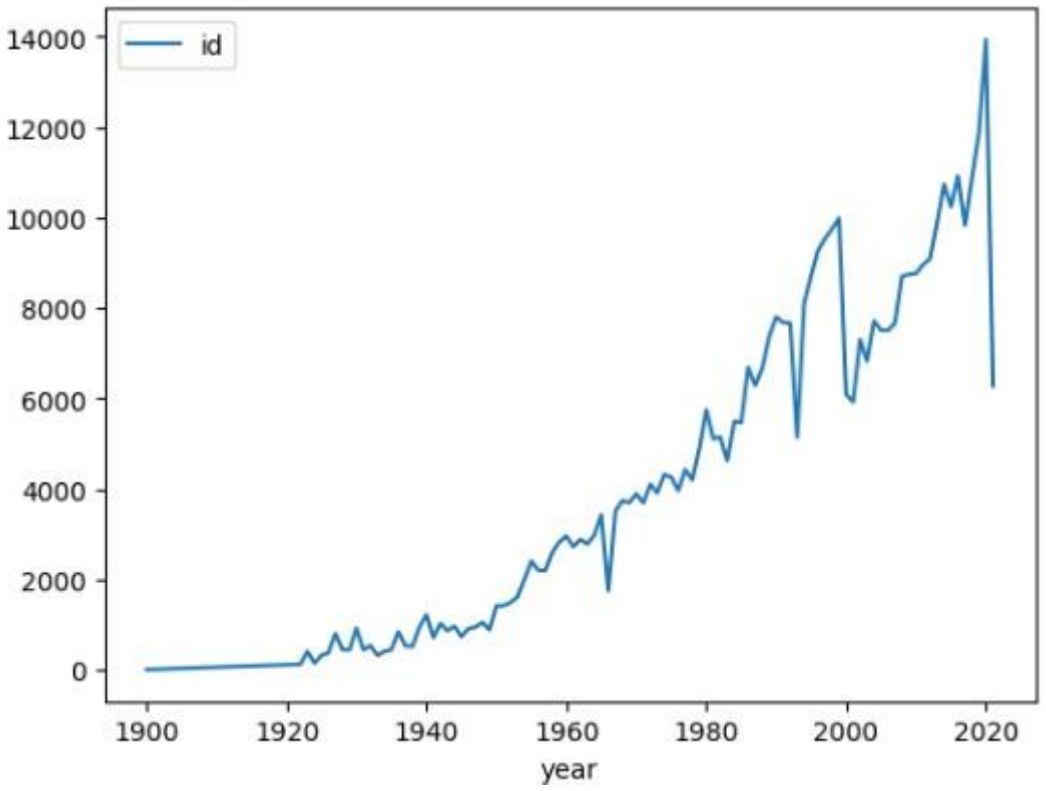

Fig 7: Track of Music Over Time

The plot illustrating the number of tracks produced each year from 1900 to 2020 reveals a clear and sustained upward trend, indicating the growing scale of music production over the past century. In the early 1900s, the volume of music releases was relatively low, reflecting both the limited recording infrastructure and the constraints of physical media distribution. However, starting in the mid-20th century, the number of new tracks began to rise steadily, with noticeable acceleration from the 1960s onward.

This growth becomes particularly pronounced in the early 2000s, culminating in an exponential increase in music production over the past two decades. The sharp rise in the number of songs released after 2015 aligns closely with the widespread adoption of digital distribution platforms, which significantly lowered the barriers to entry for artists. The emergence of streaming services, online music stores, and social media channels has enabled independent musicians to publish content directly to global audiences, bypassing traditional gatekeepers such as record labels and radio stations.

Notably, the peak observed around 2020 underscores the digital era's role in democratizing music creation and distribution. While minor fluctuations appear due to data limitations or reporting inconsistencies, the overall trajectory clearly illustrates a dramatic expansion in the volume of global music output.

This trend reflects broader shifts in the music industry—from analog to digital, from centralized control to decentralized access—and emphasizes the importance of data-driven platforms in shaping the modern musical ecosystem.

2.7 Artist Popularity and Followers

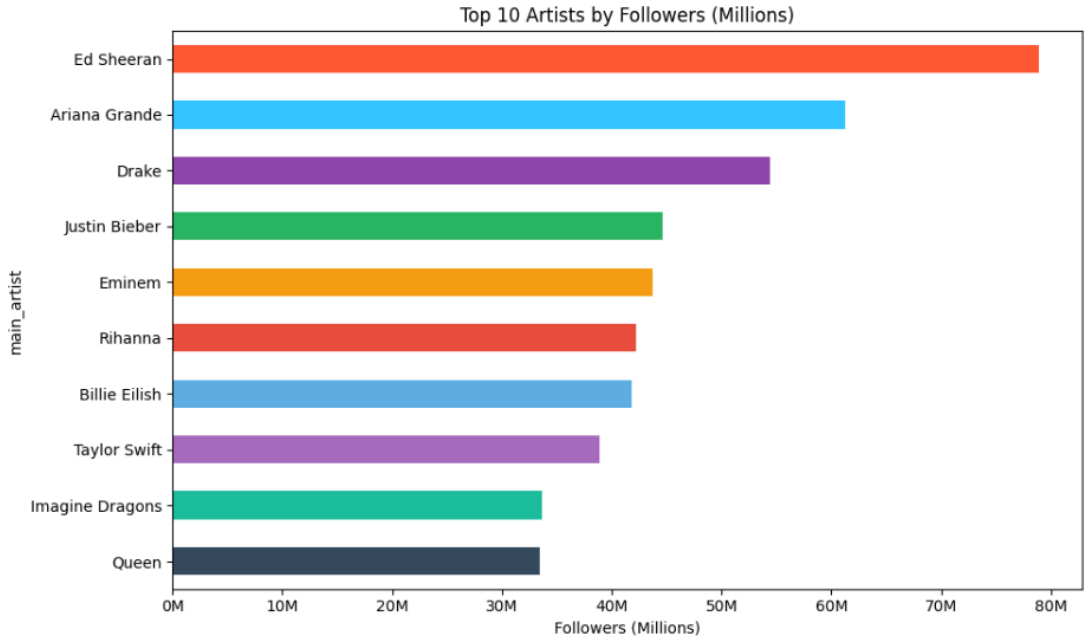

Fig 8: Track Popularity Distribution

The analysis of artist follower counts highlights the significant role of audience reach in shaping musical success. The bar chart presented above illustrates the top ten most-followed artists, with figures measured in millions. Ed Sheeran leads the list by a substantial margin, followed by other globally recognized artists such as Ariana Grande, Drake, Justin Bieber, and Eminem.

A clear pattern emerges from this distribution: artists with large follower bases tend to consistently produce tracks with higher average popularity scores. This relationship underscores the interplay between personal brand strength and commercial performance. Artists with substantial followings benefit from enhanced visibility on streaming platforms,

where recommendation algorithms often amplify content from well-known figures. Furthermore, their releases are more likely to receive immediate attention, leading to accelerated streaming counts, social media engagement, and playlist placements.

In addition to algorithmic support, long-established fan loyalty contributes significantly to sustained engagement with new releases. These artists often possess strong cross-platform presence and media influence, further reinforcing their reach. As a result, the follower count serves as a reliable proxy for measuring an artist's influence within the digital music ecosystem.

This analysis highlights how modern music consumption is increasingly influenced by artist visibility, brand identity, and platform-driven discovery—factors that now play a central role in determining track success and longevity in the industry.

3. RECOMMENDATIONS

Based on the comprehensive analysis of Spotify's track and artist data, several actionable recommendations are proposed to support decision-making in music production, marketing strategies, and personalized content delivery. These recommendations integrate both analytical insights and machine learning techniques to enhance the effectiveness of decision support systems in the music domain.

3.1 Data-Driven Content Strategy

The feature-based analysis of musical trends over time indicates that popularity is closely associated with high energy and danceability levels. Music producers and artists are encouraged to emphasize these audio characteristics in their compositions to align with current listener preferences. Songs that are rhythmically engaging and energetic tend to perform better in terms of popularity and listener retention.

Additionally, the role of explicit content warrants strategic consideration. While tracks labeled as explicit exhibit marginally higher average popularity, this impact is likely influenced by genre norms and audience demographics. Tailoring content to suit specific target segments—considering factors such as age, geography, and platform—can improve relevance and reach.

Timing also plays a significant role in maximizing visibility. With the number of new tracks released annually increasing sharply—particularly in the digital age—artists should aim to release new material during peak engagement periods or in coordination with cultural and seasonal events to stand out amid high competition.

3.2 Machine Learning-Based Recommendations

To complement manual insights, a **K-Nearest Neighbors (KNN)** recommendation model was developed to suggest tracks based on user preferences or reference songs. This system serves as a proof-of-concept for personalized music recommendation engines.

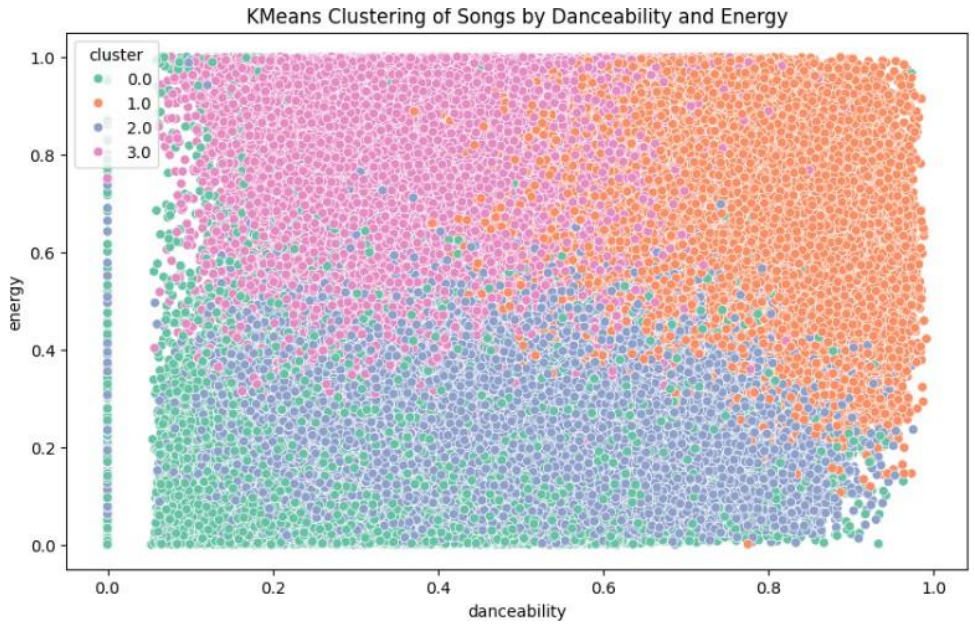

Fig 9: KMeans Clustering of Songs by Danceability and Energy

The model uses the following features for training: **year of release**, **danceability**, **energy**, and **artist follower count**. These attributes were selected for their strong correlation with popularity and their ability to capture the temporal, acoustic, and social dimensions of music.

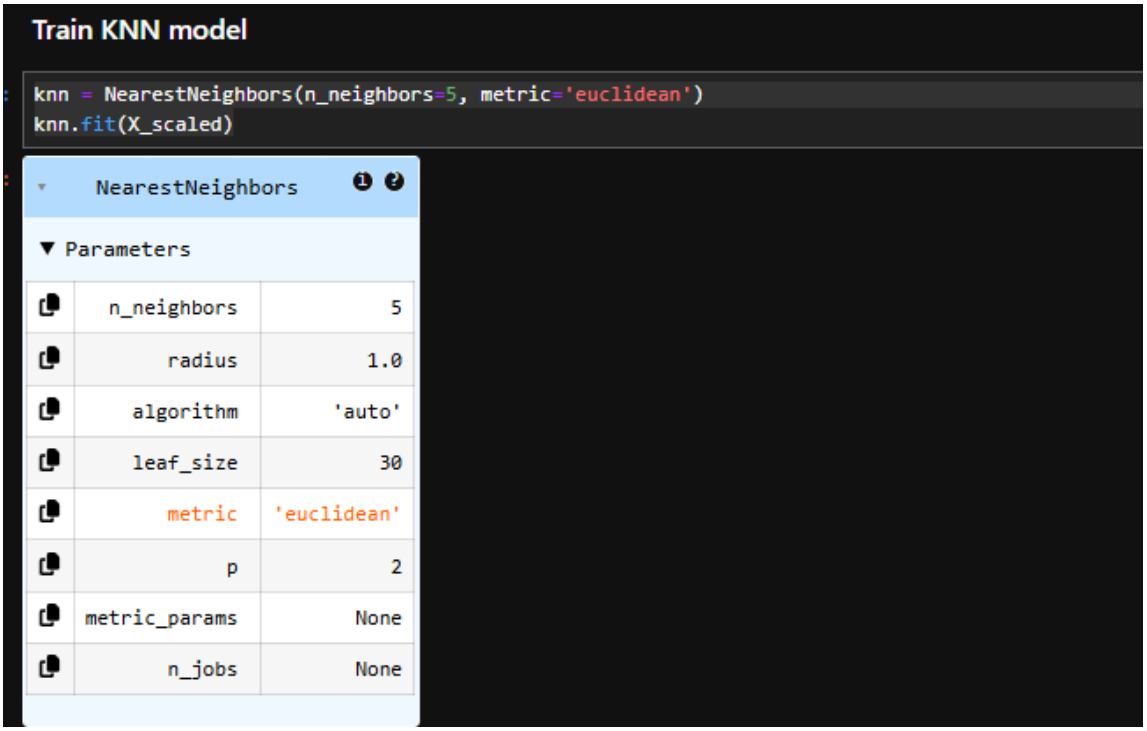

Fig 10: KNN Train Model

After scaling the input features, the KNN model identifies the five most similar tracks to a given input profile. Users provide values for release year, danceability, energy, and artist followers, after which the system returns a curated list of similar songs. Each recommendation includes the song title, artist, popularity score, follower count, and core feature values. This approach allows for real-time, context-aware suggestions based on measurable musical characteristics.

Additionally, a **KMeans clustering model** was implemented to segment songs into groups representing different musical moods or production styles. These clusters can inform playlist generation, content categorization, and listener targeting by identifying cohesive musical themes across tracks.

The integration of such models into a Decision Support System (DSS) offers a dynamic, scalable method to enhance user experience and support strategic decisions in music curation and promotion.

3.3 Decision Support System (DSS) Applications

The insights and algorithms developed can be applied in various decision-support contexts within the music industry:

- **Playlist Curation**: Streaming platforms and curators can leverage clustering and similarity algorithms to build thematic playlists—such as upbeat workout mixes or mellow evening sets—aligned with listener moods and habits.
- **Artist Development**: Music labels and talent managers can use historical feature-popularity patterns to guide artists in refining their sound, aligning with trends that resonate most with audiences.
- **Marketing and Promotion**: By analyzing the relationship between follower count and track popularity, promotional resources can be allocated toward emerging artists with rising influence to maximize ROI.
- **User Experience Optimization**: Personalized recommendations powered by machine learning can improve engagement and retention by delivering content that matches users' musical taste profiles.

4. CONCLUSION

This project demonstrated the power of leveraging Spotify's extensive music dataset to support informed decision-making in the music industry. Through systematic data preprocessing, detailed exploratory analysis, and effective visualization, key insights were uncovered regarding track popularity, artist influence, and evolving musical trends.

The study highlighted that energetic, danceable tracks tend to perform better on the platform, while artist follower counts strongly correlate with track success. Explicit content shows nuanced effects, emphasizing the importance of audience targeting. Temporal analysis revealed a significant growth in music releases, underscoring the need for strategic content timing.

The development of machine learning models, including clustering and recommendation algorithms, showcased practical applications of data-driven approaches within a Decision Support System framework. These models enable personalized music suggestions and strategic playlist curation, enhancing both user experience and industry efficiency.

Overall, this work underscores the value of integrating data analytics with music domain knowledge to foster smarter content creation, marketing, and user engagement strategies in the digital streaming ecosystem.

5. SOURCES

1. https://developer.spotify.com/documentation/web-api/
2. https://spotifycharts.com/
3. https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db
4. https://www.kaggle.com/datasets/leonardopena/top-spotify-songs-from-20102019-by-year
5. https://pandas.pydata.org/docs/
6. https://matplotlib.org/stable/index.html
7. https://seaborn.pydata.org/
8. https://scikit-learn.org/stable/

6. APPENDICES

Appendix A: Merged Dataset Columns

Column Name	Description
id	Track ID
name	Track Name
popularity	Track Popularity (0–100)
duration_ms	Duration in milliseconds
explicit	Boolean flag for explicit content
artists	List of contributing artists
id_artists	Artist ID(s)
release_date	Track release date
danceability	Musical rhythm and groove (0–1)
energy	Intensity of the track (0–1)
loudness	Volume level in decibels
speechiness	Presence of spoken words
acousticness	Probability of acoustic sound
instrumentalness	Likelihood of being instrumental
liveness	Audience presence factor
valence	Musical positivity (0–1)
tempo	Beats per minute
duration_min	Track length in minutes
year	Extracted release year
main_artist	Primary artist
followers	Artist follower count