# Spotify Data Analysis: Investigating Musical Features and Popularity Trends Using Machine Learning

Zahra Momeni
Sapienza University
Visual Analytics final project
Email: momeni.2110005@studenti.uniroma1.it

***Abstract*—The rapid expansion of music streaming platforms has led to vast amounts of audio data, providing an opportunity to analyze and understand listener preferences. This study explores Spotify's extensive track dataset to identify key musical attributes influencing song popularity. Using a dataset of over 586,672 songs, we analyze features such as danceability, energy, valence, and acousticness to uncover correlations with track success. Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and K-Means clustering are applied to extract patterns, which are visualized using scatter plots, heatmaps, and interactive time-series visualizations. The findings provide insights for music producers, streaming platforms, and analysts, helping to optimize song attributes and enhance playlist curation.**

***Index Terms*—Music analytics, machine learning, clustering, dimensionality reduction, Spotify, recommender systems.**

## I. Introduction

Music streaming platforms generate vast amounts of data on song characteristics, including tempo, danceability, energy, and popularity. Identifying meaningful patterns within high-dimensional music data remains challenging. This study applies Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and K-Means clustering to investigate the correlation between audio features and song popularity. Our research provides insights into which musical attributes contribute most to commercial success, using Spotify data for analysis.

## II. Related Work

The application of machine learning techniques for music analysis has been widely explored in prior research, particularly in areas such as music trend analysis, genre classification, and recommender systems. This section reviews three relevant studies— [2], [3], and [4]—and compares them to the present study, Spotify Data Analysis.

The study [2] applies a BN-content analysis model and normalization algorithms to examine the evolution of music trends. It finds that education and technology significantly influence music development, with higher education institutions playing a major role in shaping musical styles. Additionally, it discusses how digital platforms impact music distribution and audience engagement. While this study and Spotify Data Analysis both leverage data-driven approaches, their focus differs. The former emphasizes cultural and educational impacts on music trends, whereas Spotify Data Analysis investigates feature-based correlations with song popularity, emphasizing audio characteristics rather than external influences.

The second study, [3], explores clustering techniques such as K-Means and Hierarchical Clustering for genre classification. The research highlights the effectiveness of dimensionality reduction (PCA) in improving clustering performance and demonstrates that feature selection, particularly Mel-Frequency Cepstral Coefficients (MFCCs), enhances classification accuracy. Similar to Spotify Data Analysis, this study employs clustering and dimensionality reduction to analyze high-dimensional music datasets. However, while [3] primarily focuses on genre classification, Spotify Data Analysis extends clustering applications to song popularity analysis, exploring how different audio attributes contribute to listener engagement.

The third study, [4], investigates K-Means clustering and cosine similarity for personalized music recommendations. It focuses on identifying feature-based similarities among songs to enhance recommendation accuracy and discusses cold start and data sparsity challenges in recommender systems. This study aligns with Spotify Data Analysis in its application of clustering algorithms and similarity measures, but their objectives differ. While [4] seeks to enhance personalized recommendations, Spotify Data Analysis aims to understand the relationship between musical attributes and song popularity, making it more exploratory in nature rather than focused on user personalization.

In summary, prior studies have demonstrated the effectiveness of clustering, dimensionality reduction, and feature-based analysis in music data research. However, while existing works primarily focus on music evolution, genre classification, and recommender systems, Spotify Data Analysis uniquely investigates the correlation between musical features and popularity using PCA, t-SNE, and K-Means clustering. By integrating these techniques, this study contributes to a deeper understanding of how musical attributes influence a track's success, offering insights relevant to music producers, streaming platforms, and industry analysts.

## III. Data Preparation and Processing

The dataset utilized in this study is sourced from Spotify's track metadata, comprising 586,672 records with 20 features that encapsulate song attributes, popularity metrics, and audio characteristics. The dataset includes three primary categories of features: track metadata (e.g., song ID, name, artist, and release date), musical features (e.g., danceability, energy, acousticness), and popularity metrics (e.g., explicit content flag, popularity score). Given the high-dimensional nature of the data, preprocessing and normalization steps were performed to ensure consistency and accuracy in the subsequent analysis.

### A. Data Cleaning and Normalization

To improve data quality, several preprocessing steps were applied. Missing values in track names were replaced with *Unknown*, and duplicate entries were removed as necessary. Artist names were cleaned by eliminating unnecessary brackets and formatting inconsistencies. The duration column was converted from milliseconds to seconds for better interpretability, and release dates were reformatted into a standard datetime structure.

For numerical consistency, all numeric attributes were normalized using MinMaxScaler, transforming them into a 0-1 range to ensure uniformity across different scales. This preprocessing step optimizes the dataset for clustering, visualization, and correlation analysis. After processing, the cleaned dataset was stored for further analysis.

## IV. Visualization and Interaction

The interactive visualization dashboard is designed to allow users to explore patterns within the Spotify dataset using dimensionality reduction and clustering techniques. The primary users of this tool include music producers, streaming platform analysts, and data scientists, who can leverage the insights to understand track popularity, musical feature relationships, and song recommendations.

### A. Main Components of the Visualization

1) **Dimensionality Reduction and Scatter Plot**
   - Users can choose between PCA (Principal Component Analysis) and t-SNE (t-Distributed Stochastic Neighbor Embedding) to reduce high-dimensional data into a lower-dimensional space for visualization.
   - The scatter plot dynamically updates based on user selections, showing how songs cluster based on their features while color-coding data points according to popularity.
   - PCA provides an interpretable projection, where the explained variance is displayed to show how much information is retained. t-SNE focuses on preserving local similarities and allows users to adjust the perplexity parameter for fine-tuning the clustering results.
2) **Interactive Song Analysis**
   - Users can select points on the scatter plot to analyze song characteristics and correlations.
   - When multiple songs are selected, the system calculates correlations between popularity and key features, identifying which attributes contribute most to a track's success.
   - A breakdown of the most common artists and song details within the selected cluster is displayed, helping users discover trends among similar songs.
3) **Song Recommendation System**
   - Users can search for a song and receive recommendations based on cosine similarity and nearest neighbors analysis.
   - The system finds songs with similar audio features and presents a ranked list of recommendations.

- This feature is particularly useful for playlist curation and content-based music recommendations.

### B. *User Interaction*

- **Parameter Selection:** Users can modify the number of PCA components or adjust the perplexity for t-SNE to explore different visual representations of the data.
- **Dynamic Filtering:** Songs can be selected directly from the scatter plot, allowing for a detailed feature analysis of custom selections.
- **Search & Recommendation:** Users can input a song title to receive automated recommendations based on feature similarity.

This visualization tool enhances the interpretability of music data analysis by providing an interactive, data-driven exploration of track popularity, feature correlations, and similarity-based recommendations. It facilitates decision-making for industry professionals and supports a deeper understanding of song attributes influencing success in streaming platforms.

## V. Analytics and Insights

This section presents the analytical findings derived from correlation analysis, dimensionality reduction, clustering, and trend exploration of the Spotify dataset. By leveraging machine learning techniques such as Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and K-Means clustering, we extract key insights into the relationships between track attributes and their popularity.

### A. *Feature Correlations with Popularity*

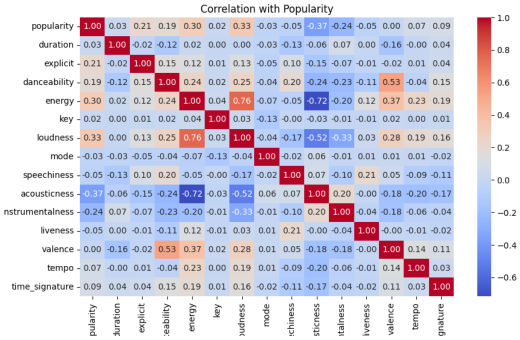

Fig. 1. Correlation Heatmap of Song Features with Popularity.

A correlation heatmap was constructed to identify the relationships between various song features and popularity. The analysis revealed the following key patterns:

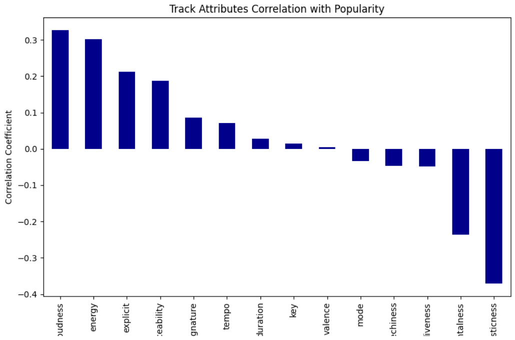

Fig. 2. Bar Plot of Feature Correlations with Popularity.

#### 1) *Positive Correlations with Popularity:*

- **Loudness (0.33):** Louder tracks tend to be more popular, indicating that energetic and high-intensity songs resonate more with listeners.
- **Energy (0.30):** High-energy songs exhibit a strong correlation with popularity, suggesting that lively and dynamic tracks gain wider appeal.
- **Danceability (0.19):** Danceable songs have a moderate positive impact on popularity, reinforcing the significance of rhythm in audience engagement.
- **Explicit Content (0.21):** Explicit tracks show a slight positive correlation with popularity, possibly reflecting audience preferences for contemporary music trends.

#### 2) *Negative Correlations with Popularity:*

- **Acousticness (-0.37):** Songs with high acoustic content tend to be less popular, implying a preference for electronically produced or high-energy tracks.
- **Instrumentalness (-0.24):** Instrumental-heavy songs are generally less favored, suggesting a strong audience preference for vocal-based tracks.
- **Valence (-0.16):** Songs with a higher happiness score (valence) exhibit a slight negative correlation with popularity, indicating that neutral or moody tones may be more engaging.

#### 3) *Weak or No Correlation:*

- **Musical Key (0.02) and Mode (-0.03):** Neither the key nor mode (major/minor) significantly influences a song's popularity.
- **Tempo (0.07):** There is no meaningful correlation between a song's speed and its popularity.

These findings highlight the importance of energy, loudness, and danceability in predicting a song's success, while attributes such as acousticness and instrumentalness tend to have a negative impact.

### *B. Trends in Music Releases and Popularity Over Time*

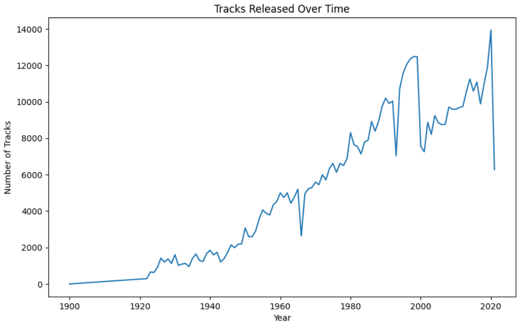

Fig. 3. Tracks Released Over Time.

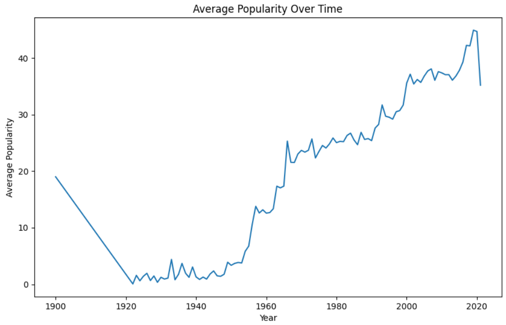

Fig. 4. Average Popularity Over Time.

#### *1) Historical Growth of Music Releases:*

- **Before 1960:** The number of song releases grew at a slow and steady rate, largely due to technological constraints and limited distribution channels.
- **Post-1960 Acceleration:** A significant increase in song releases coincides with the rise of modern music genres, improved recording technology, and global music distribution.
- **Exponential Growth Post-2000:** The advent of digital music production and online streaming platforms (e.g., Spotify, YouTube, SoundCloud) led to a dramatic increase in releases.
- **Recent Volatility (2000–2020):** The number of tracks released fluctuates, likely influenced by industry shifts such as streaming dominance and external factors (e.g., COVID-19 pandemic).

#### *2) Average Popularity Over Time:*

- **Early Years (Pre-1940s):** The dataset shows a sharp decline in popularity, potentially due to limited metadata availability or streaming biases.
- **Gradual Growth (1940s–1960s):** As music became more widely accessible via radio and records, song popularity increased.
- **Rapid Popularity Surge (1970s–2000s):** This period marks the ”golden era” of modern music, driven by mass media (radio, TV, cassette tapes, CDs).
- **Peak Popularity (2000s–2010s):** The digital streaming revolution led to unprecedented accessibility, pushing song popularity to new heights.
- **Recent Decline (Post-2020):** A slight dip in recent years may reflect data sparsity (newer songs lacking sufficient streaming metrics) rather than an actual trend.

These trends underscore the impact of technological advancements and digital transformation in shaping the global music industry.

### *C. Dimensionality Reduction and Visualization*

*1) PCA (Principal Component Analysis):* PCA was applied to reduce the dataset’s dimensionality while retaining the most significant variance. The PCA scatter plot revealed the following:

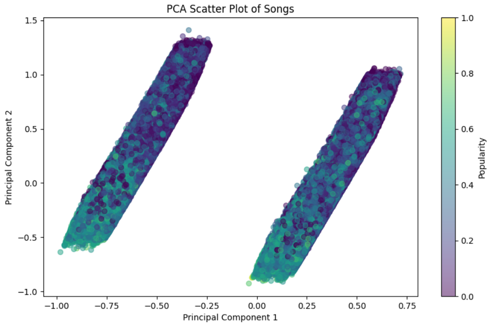

Fig. 5. PCA Scatter Plot of Songs.

- **Two Principal Components Capture Key Variance:** The first two principal components (PC1, PC2) account for the largest variance, preserving the dataset’s essential structure.
- **Bimodal Distribution:** The data exhibits two primary clusters, suggesting natural groupings based on features like loudness, energy, and danceability.
- **Popularity Gradient:** High-popularity tracks are dispersed across both clusters, indicating that popularity is not strictly bound to a single feature set.
- **Feature Contributions:** The PCA results suggest that loudness, energy, and danceability contribute

the most to variance, aligning with correlation findings.

*2) t-SNE (t-Distributed Stochastic Neighbor Embedding):* t-SNE was utilized for non-linear dimensionality reduction, emphasizing local relationships. The t-SNE scatter plot highlights:

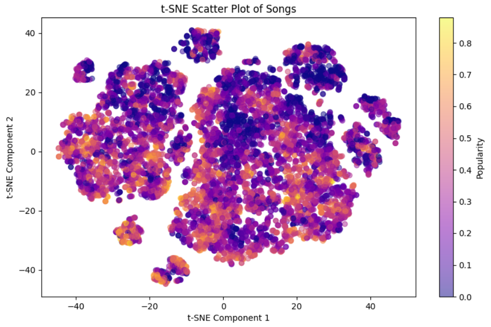

Fig. 6. t-SNE Scatter Plot of Songs.

- **Clear Clustering Patterns:** The dataset forms distinct clusters, reinforcing that certain song attributes naturally group together.
- **Local Similarity Preservation:** Unlike PCA, t-SNE maintains close relationships between similar data points, making it better suited for discovering hidden song groupings.
- **Popularity Distribution:** High-popularity tracks are spread across multiple clusters, suggesting that different types of songs can achieve success in different ways.

**Key Differences Between PCA and t-SNE:**

- PCA captures global variance, making it useful for understanding how features contribute to overall trends.
- t-SNE focuses on local clusters, making it ideal for discovering song groupings based on similar musical characteristics.

## *D. Clustering Analysis (K-Means)*

*1) Cluster Interpretation (PCA-Based K-Means):* K-Means clustering on PCA-reduced data reveals four distinct clusters:

- Cluster 0: Low-energy, high-acousticness tracks, likely representing older, softer, or instrumental songs.
- Cluster 1: High-energy, danceable tracks, aligning with modern, popular, and upbeat songs.

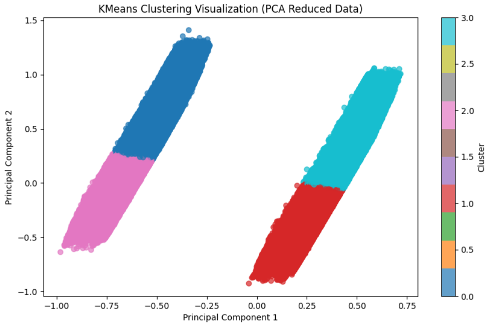

Fig. 7. K-Means Clustering Visualization (PCA Reduced Data).

- Cluster 2: The most popular cluster, characterized by high loudness, energy, and danceability, representing mainstream hits.
- Cluster 3: Older, low-popularity, acoustic-heavy songs with a tendency toward instrumental music.

*2) Cluster Interpretation (t-SNE-Based K-Means):* Applying K-Means to t-SNE-reduced data provided additional insights into localized song clusters:

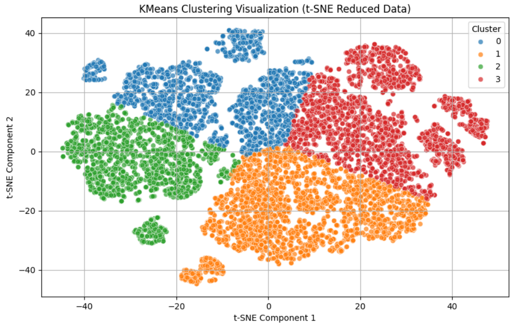

Fig. 8. K-Means Clustering Visualization (t-SNE Reduced Data).

- Cluster 0: Low-popularity, acoustic, instrumental-heavy tracks.
- Cluster 1: High-popularity, energetic dance tracks, often in major keys.
- Cluster 2: Highly popular, synthetic music with electronic production.
- Cluster 3: Low-popularity, acoustic-focused songs with high instrumentalness.

**Key Takeaways from Clustering:**

- Popular songs exhibit high energy, loudness, and danceability.
- Acoustic and instrumental-heavy tracks tend to be less favored in mainstream audiences.

- Songs naturally group into clusters that align with their musical characteristics.

## VI. Conclusion and Future Improvements

### A. Conclusion

This study presents a comprehensive Spotify data analysis utilizing machine learning techniques to explore song popularity, feature correlations, and clustering patterns. By applying dimensionality reduction (PCA, t-SNE) and K-Means clustering, we identified the key factors influencing a track's success and uncovered meaningful song groupings based on their musical attributes.

The correlation analysis revealed that loudness, energy, and danceability are the strongest positive predictors of song popularity, while acousticness and instrumentalness negatively impact mainstream success. The historical analysis of music releases and popularity trends demonstrated the significant influence of technological advancements, particularly the digital streaming revolution post-2000.

Dimensionality reduction techniques provided valuable visual insights into the dataset's structure. PCA captured global variance, emphasizing how different audio features contribute to popularity, while t-SNE highlighted local clusters, revealing distinct song groupings. The clustering analysis further reinforced these findings, categorizing songs into four major groups based on their acoustic and energetic properties. Modern, high-energy tracks emerged as the most popular, whereas acoustic-heavy, instrumental tracks were generally less favored.

These insights are particularly relevant for music producers, streaming platform analysts, and industry professionals, offering guidance on track optimization, audience engagement, and playlist curation. The findings also contribute to the broader field of music analytics and recommendation systems, demonstrating the effectiveness of machine learning techniques in understanding musical trends and listener preferences.

### B. Future Improvements

While this study provides valuable insights, several areas for improvement and further research remain:

- **Expansion of Feature Analysis:** Incorporating additional audio features such as melodic patterns, lyrical content, and genre classification could enhance the understanding of song popularity. Sentiment analysis on lyrics could provide deeper insights into the emotional tone of popular tracks.
- **Refining Clustering Models:** Exploring alternative clustering techniques such as DBSCAN or hierarchical clustering could provide more granular segmentation of songs. Implementing hybrid models combining genre classification and feature-based clustering could refine recommendations for playlist generation.
- **Temporal Analysis of Music Trends:** A decade-wise breakdown of music feature evolution could provide insights into how listener preferences have changed over time. Investigating how seasonality and external events (e.g., viral trends, award seasons) influence song popularity would be beneficial.
- **Integration with Real-World Music Recommendation Systems:** Implementing a content-based recommendation system using similarity measures could personalize music discovery for users. Combining collaborative filtering techniques with feature-based clustering could enhance playlist curation in streaming platforms.
- **Deep Learning Applications:** Incorporating neural networks (e.g., autoencoders for feature extraction or deep learning for genre classification) could improve the accuracy of predictions. Leveraging natural language processing (NLP) for lyric-based music classification could further enrich the dataset.

## VII. Final Thoughts

This study highlights the power of data-driven approaches in music analytics, demonstrating how machine learning techniques can uncover hidden patterns and optimize music recommendations. By refining the models and expanding feature analysis, future research can further enhance music discovery, personalization, and predictive modeling in the evolving digital landscape.

As the music industry continues to evolve with AI-driven innovations and streaming dominance, leveraging advanced analytics will be crucial for understanding audience behavior and shaping the future of music consumption.

## References

[1] Spotify Datasets for Music Analysis: https://www.kaggle.com

[2] The Development Trend of Music Art Based on Content Analysis Method: https://www.researchgate.net/publication/371282936_The_development_trend_of_music_art_based_on_content_analysis_method

[3] Analysis of Music Genre Clustering Algorithms: https://dc.uwm.edu/cgi/viewcontent.cgi?article=3844&context=etd

[4] Content-Based Filtering Technique Using Clustering Method for Music Recommender Systems: https://www.researchgate.net/publication/387823545_Content-Based_Filtering_Technique_using_Clustering_Method_for_Music_Recommender_Systems

[5] PCA (Principal Component Analysis): https://scikit-learn.org/stable/modules/decomposition.html#pca

[6] t-SNE (t-distributed Stochastic Neighbor Embedding): https://scikit-learn.org/stable/modules/manifold.html#t-sne

[7] K-Means Clustering Algorithm: https://scikit-learn.org/stable/modules/clustering.html#k-means

[8] Pandas Library: https://pandas.pydata.org

[9] Plotly for Visualization: https://plotly.com/python

[10] Matplotlib Library: https://matplotlib.org