NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Personalized Music Recommendation System

Data Analysis Report

MSDSP 498 – Capstone Project

Group 2

Abhigna Mallepally, Anishka Agarwal, Edwin Daniels, Sachin Sharma

January 26, 2025

Introduction

In the age of digital media, personalized recommendation systems have become an integral part of the user experience across various platforms. These systems aim to provide tailored content suggestions, thereby enhancing user engagement and satisfaction. Music streaming services like Spotify have revolutionized the way users discover and consume music by leveraging sophisticated recommendation algorithms. This project aims to develop a personalized music recommendation system by integrating multiple machine learning (ML) and deep learning (DL) models, using user interaction data and Spotify's popular songs metadata. By employing a combination of collaborative filtering, content-based filtering, matrix factorization, and neural network techniques, the system strives to deliver highly accurate and relevant music recommendations and address key challenges such as cold-start recommendations, playlist generation, and engagement-driven insights while ensuring scalability and user satisfaction.

Overview of Dataset

Two datasets will be used to support the development of our recommendation system:

1. **User Data:** We collected the last 50 songs listened to by users using the Spotify REST API. This data provides insight into individual music preferences, which helps in generating personalized recommendations.
2. **Training Data:** We compiled a dataset of 1 million songs from Spotify available on Kaggle (https://www.kaggle.com/datasets/amitanshjoshi/spotify-1million-tracks/data). This dataset contains 19 features between 2000 and 2023 and has data of a total of 61,445 unique artists and 82 genres. This data is used to get our songs recommendation using different models, ensuring that the recommendations are diverse and cover a wide range of music genres and artists.

Table 1: Column level information of the datasets

Column Name	Description	User Data	Training Data
User_name	Name of the user	✓	
Artist_name	Name of the artist	✓	✓
Track_name	Track or record name	✓	✓
Track_id	Unique id from Spotify for a track	✓	✓
Popularity	Track popularity (0 to 100)	✓	✓
Year	Year released (2000 to 2023)	✓	✓
Played_at	Time when the user last played the track	✓	
Genre	Genre type of the song	✓	✓
Danceability	Track suitability for dancing (0.0 to 1.0)	✓	✓
Energy	The perceptual measure of intensity and activity (0.0 to 1.0)	✓	✓
Key	The key, the track is in (-1 to -11)	✓	✓
Loudness	Overall loudness of track in decibels (-60 to 0 dB)	✓	✓
Mode	Modality of the track (Major '1'/ Minor '0')	✓	✓
Speechiness	Presence of spoken words in the track	✓	✓
Acousticness	Confidence measure from 0 to 1 of whether the track is acoustic	✓	✓
Instrumentalness	Whether tracks contain vocals (0.0 to 1.0)	✓	✓
Liveness	Presence of audience in the recording (0.0 to 1.0)	✓	✓
Valence	Musical positiveness (0.0 to 1.0)	✓	✓
Tempo	Tempo of the track in beats per minute (BPM)	✓	✓
Time_signature	Estimated time signature (3 to 7)	✓	✓
Duration_ms	Duration of track in milliseconds	✓	✓

Data Collection

Before applying EDA techniques, we extracted user Spotify data. To do this, Spotify SDK tools were employed to extract and analyze the user's most recent played tracks (fifty tracks).

- Libraries such as ‘spotipy’ ‘pandas’ ‘datetime’ ‘counter’ were used to collect user data.
- Input authentication details such as CLIENT_ID, CLIENT_SECRET, REDIRECT_URI of the user was used for authentication with the Spotify API (Spotify Developer).
- The user's most recently played tracks (up to 50) were fetched. Helper functions were setup to extract the required audio features for the tracks.
- The extracted data was converted into a dataframe and saved as a csv file for further EDA analysis.

User Data EDA

The purpose of this exploratory data analysis is to gain insight into the user's music preferences, trends in audio features, and listening habits using data from their Spotify playlist. The analysis explored various dimensions such as song genres, popularity, audio features, and temporal listening patterns. Various visualization methods were employed to make data insights more comprehensible, such as bar plots, histograms, scatter plots, line plots, and pie charts. Each visualization aimed to highlight specific aspects of the dataset, from genre distribution to trends over time.

1. **Missing Values**: The initial step addressed missing values in the genre column by filling them with "Unknown." This ensures that no data is lost when performing genre-based analyses, such as visualizations or aggregations. Without this step, null values could have introduced errors or biases, especially in grouping operations or filtering. After this, a check confirmed that all columns had zero missing values, making the dataset ready for further exploration.

2. **Basic Descriptive Statistics:** Basic descriptive statistics were calculated for all numerical columns, offering a snapshot of the dataset's distribution. The popularity column had an average value of 54.66, with a range from 25 to 88, showing a mix of moderately and highly popular songs. The year column indicated that user listened more to songs which were released between 2012 and 2024, with a median year of 2023, suggesting a preference for recent tracks. Audio features such as danceability (mean: 0.59) and energy (mean: 0.26) provided insights into the overall characteristics of the music, which leaned toward less energetic but moderately danceable tracks.

Table 2: Basic Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
popularity	50.00	54.66	10.70	25.00	49.00	52.00	57.75	88.00
year	50.00	2022.32	2.56	2012.00	2022.00	2023.00	2024.00	2024.00
danceability	50.00	0.59	0.13	0.35	0.48	0.58	0.67	0.90
energy	50.00	0.26	0.25	0.02	0.08	0.14	0.43	0.93
key	50.00	5.24	3.37	0.00	2.25	5.00	8.00	11.00
loudness	50.00	-15.66	6.71	-25.79	-20.62	-18.13	-8.37	-2.52
mode	50.00	0.78	0.42	0.00	1.00	1.00	1.00	1.00
speechiness	50.00	0.08	0.05	0.02	0.05	0.07	0.08	0.34
acousticness	50.00	0.81	0.29	0.04	0.75	0.98	0.99	1.00
instrumentalness	50.00	0.60	0.44	0.00	0.00	0.90	0.94	0.96
liveness	50.00	0.13	0.09	0.08	0.11	0.11	0.12	0.71
valence	50.00	0.37	0.18	0.06	0.24	0.33	0.47	0.86
tempo	50.00	109.54	31.00	65.64	83.85	104.54	128.36	202.50
duration_ms	50.00	181720.74	52558.24	122727.00	140904.25	166910.50	196121.25	351588.00
time_signature	50.00	3.92	0.34	3.00	4.00	4.00	4.00	5.00

3. **Genre Distribution**: A bar plot was generated to visualize the distribution of genres in the dataset. The graph highlighted the most common genres and their relative frequency. This analysis revealed which genres were dominant, potentially indicating the user's preferences. For instance, genres like chill guitar and filmi appeared more frequently, indicating the user’s dominant music styles, while less frequent genres like desi pop and desi hip hop indicate niche interests.

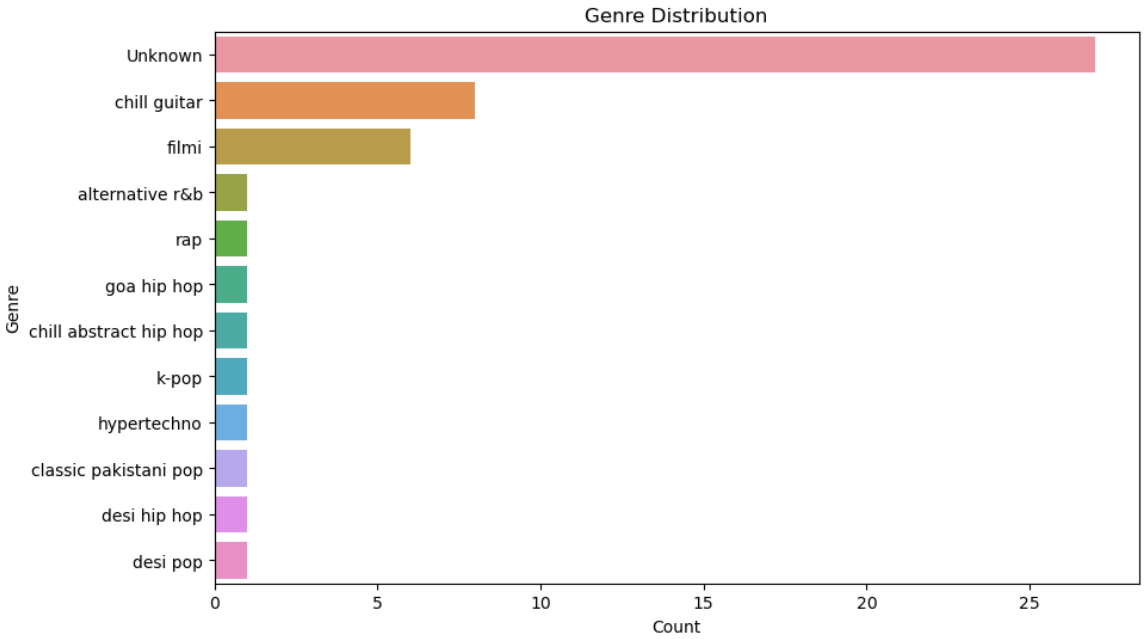


Figure 1: Genre Distribution

4. **Popularity Analysis**: The distribution of song popularity was analyzed using a histogram. The graph showed a skew toward moderately popular songs, with a gradual decrease in frequency for highly popular tracks. This trend suggests that the dataset primarily consists of songs with average popularity scores, with a smaller proportion of standout hits. The histogram also helped identify any unusual clustering or outliers in popularity.

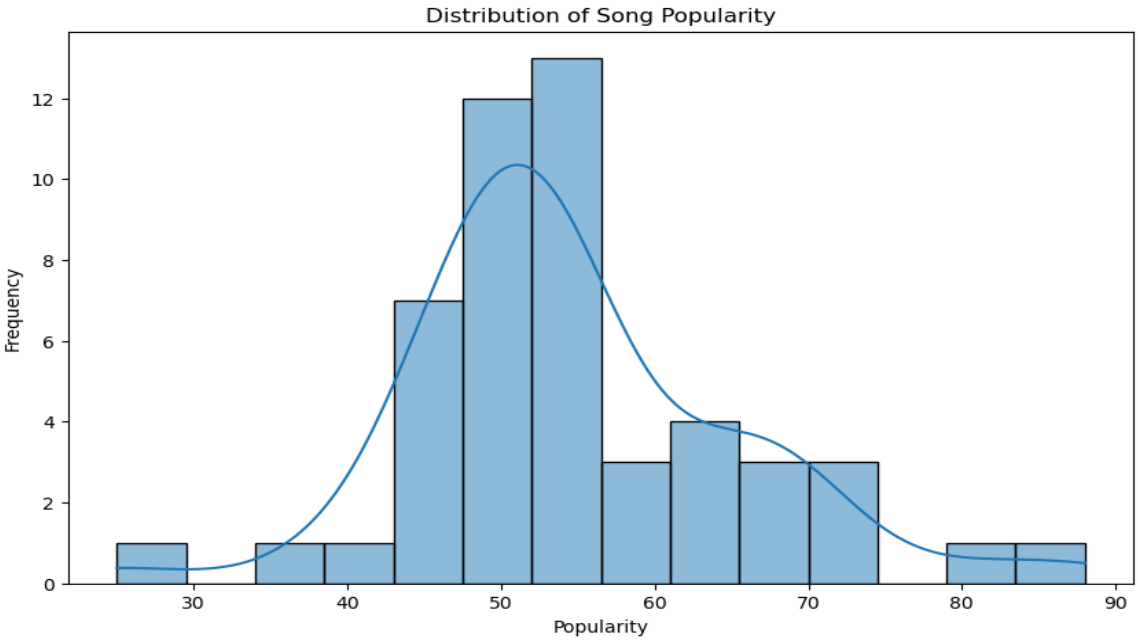


Figure 2: Popularity Analysis

5. **Audio Features**: Histograms for audio features like danceability, energy, loudness, and tempo revealed their distributions. For instance, danceability had a fairly uniform spread, while energy was skewed toward lower values, indicating a preference for less intense tracks. The loudness histogram showed most values clustering in the quieter range, and tempo exhibited a peak around 110 BPM, reflecting typical preferences for mid-tempo songs. These distributions provided a comprehensive understanding of the user's musical preferences.

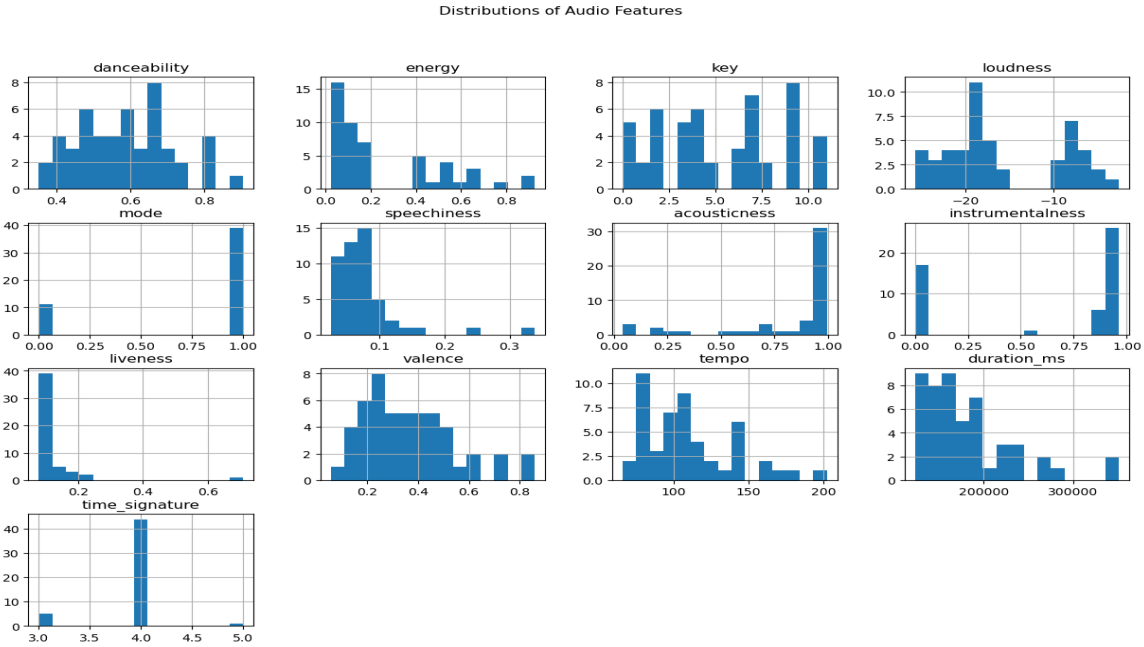


Figure 3: Audio Features Analysis

6. **Year Analysis**: The release years of songs were visualized using a histogram. Most tracks were concentrated between 2020 and 2024, with a noticeable peak around 2023. This trend indicates a preference for contemporary music. The distribution's tail extending to 2012 suggests that the dataset also includes some older tracks, albeit in smaller proportions, which may reflect occasional interest in older music.

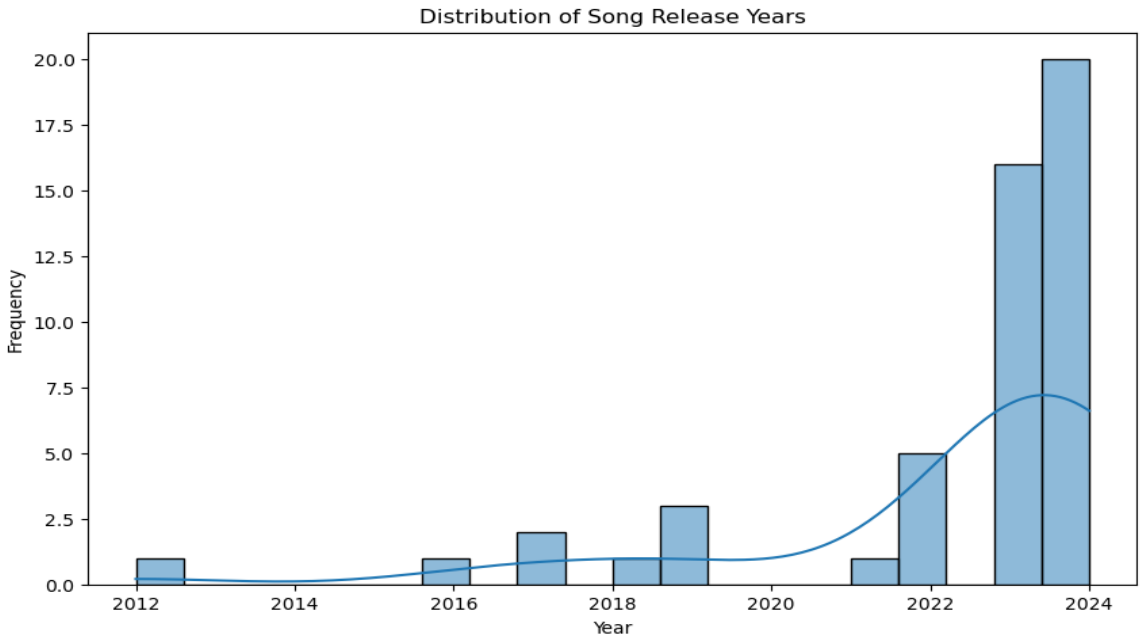


Figure 4: Year Analysis

7. **Top Artists**: A bar plot showcased the most frequently listened-to artists in the dataset. This visualization highlighted the user's musical preferences by identifying which artists appeared most often. For instance, Adeben and Arijit Singh dominate the graph indicating a strong preference for their work, while a more balanced distribution across multiple artists indicates varied tastes.

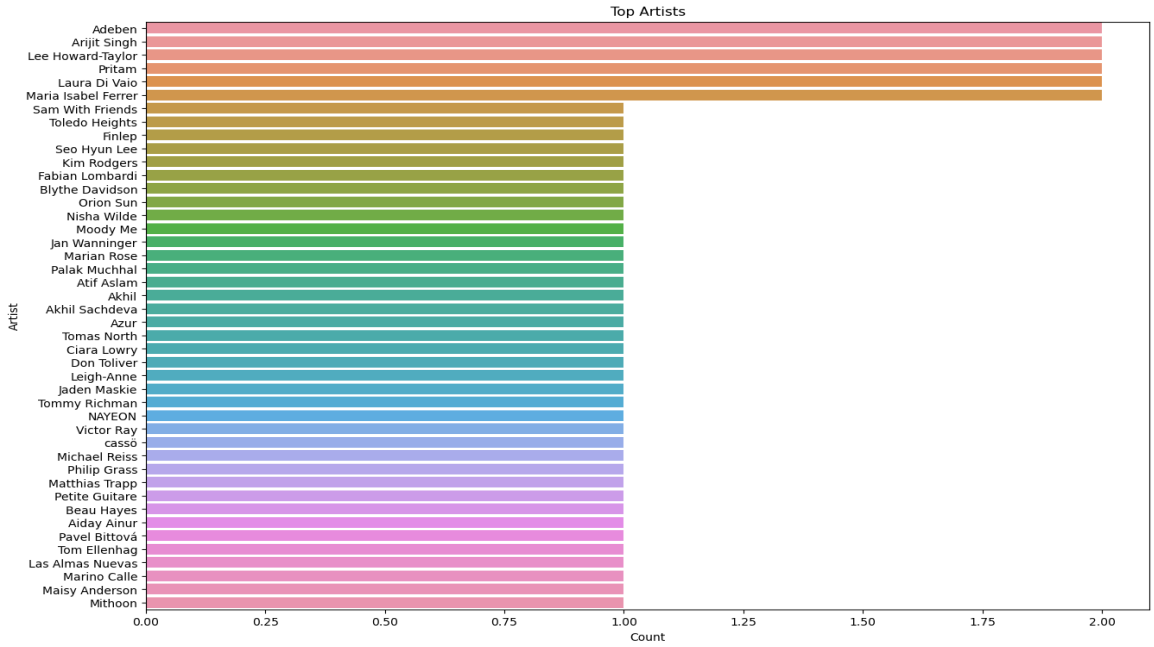


Figure 5: Top Artists

8. **Heatmap of Correlations**: A correlation heatmap was created to explore relationships between audio features. Strong positive correlations were observed between features like danceability and valence (positivity), indicating that happier songs tend to be more danceable. On the other hand, negative correlations, such as between loudness and acousticness, suggested that quieter tracks are often more acoustic. These insights help understand how audio characteristics interact and contribute to song classification.

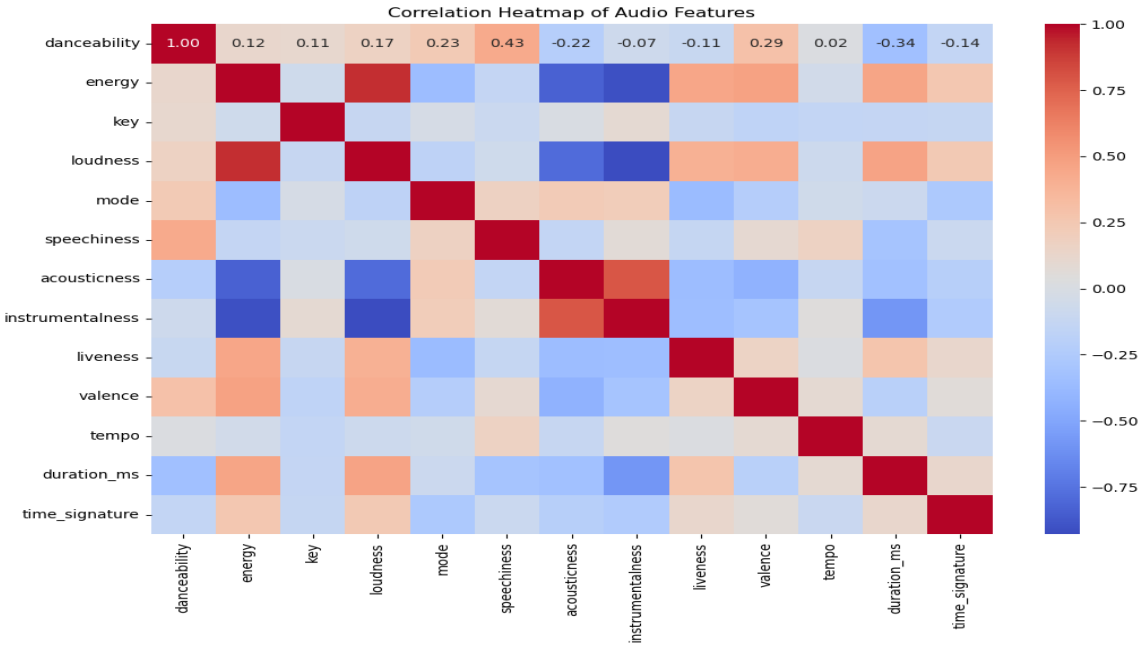


Figure 6: Heatmap of Correlations

9. **Time-Series Analysis**: The played_at column was used to examine listening trends over time, specifically by analyzing the hour of the day. A bar plot revealed distinct peaks, such as higher listening activity during mornings and lower activity at night. This behavior suggests a preference for listening during transit hours, likely while traveling to work, with reduced activity during sleeping hours.

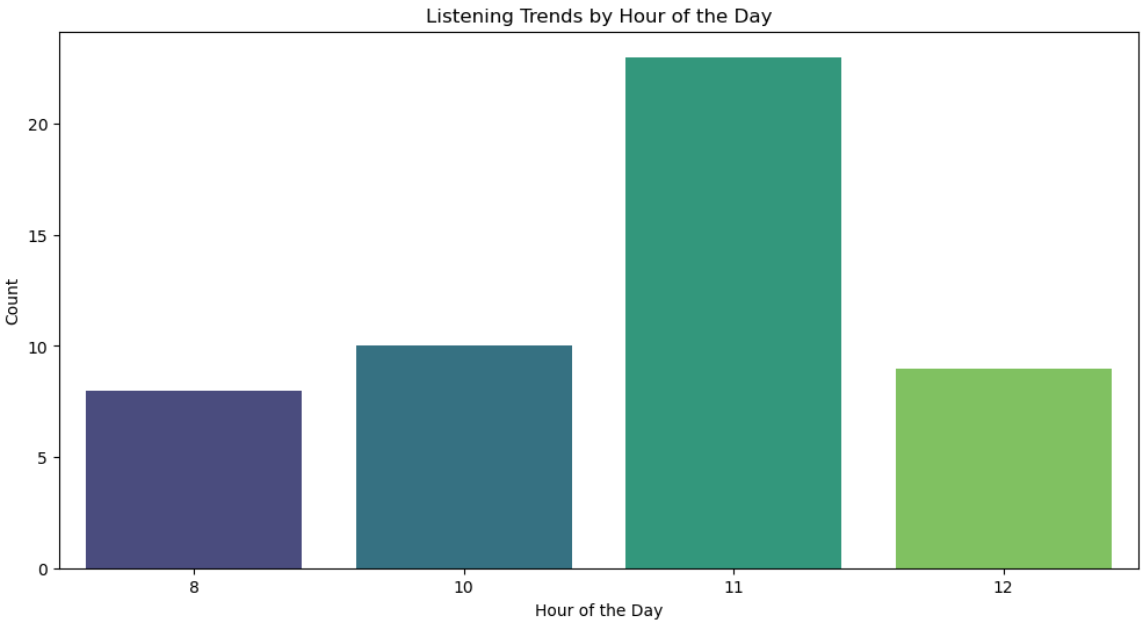


Figure 7: Time-Series Analysis

10. **Cluster Analysis**: K-means clustering was performed on audio features to group songs into similar clusters. A scatter plot of danceability versus energy showed how tracks were grouped, with three clusters identified. For example, one cluster might represent calm, acoustic tracks, while another could include high-energy, danceable songs. This analysis helps segment the dataset and provides recommendations or personalized playlists based on song characteristics.

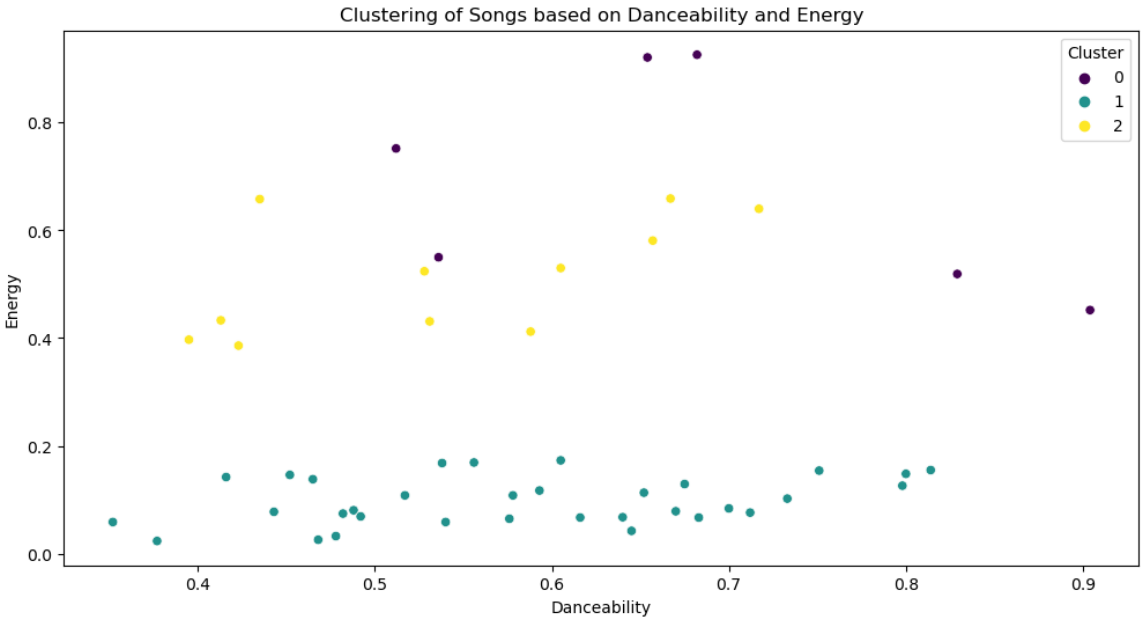


Figure 8: Cluster Analysis

11. **User Listening Behavior**: Listening behavior was further analyzed by categorizing listening times into morning, afternoon, evening, and night. A bar plot showed that the morning was the most active period, followed by the afternoon, reflecting typical travel-time listening habits. Evening and night listening were less frequent, possibly indicating lower engagement after work hours. This analysis provides valuable insights into the user's daily listening patterns.

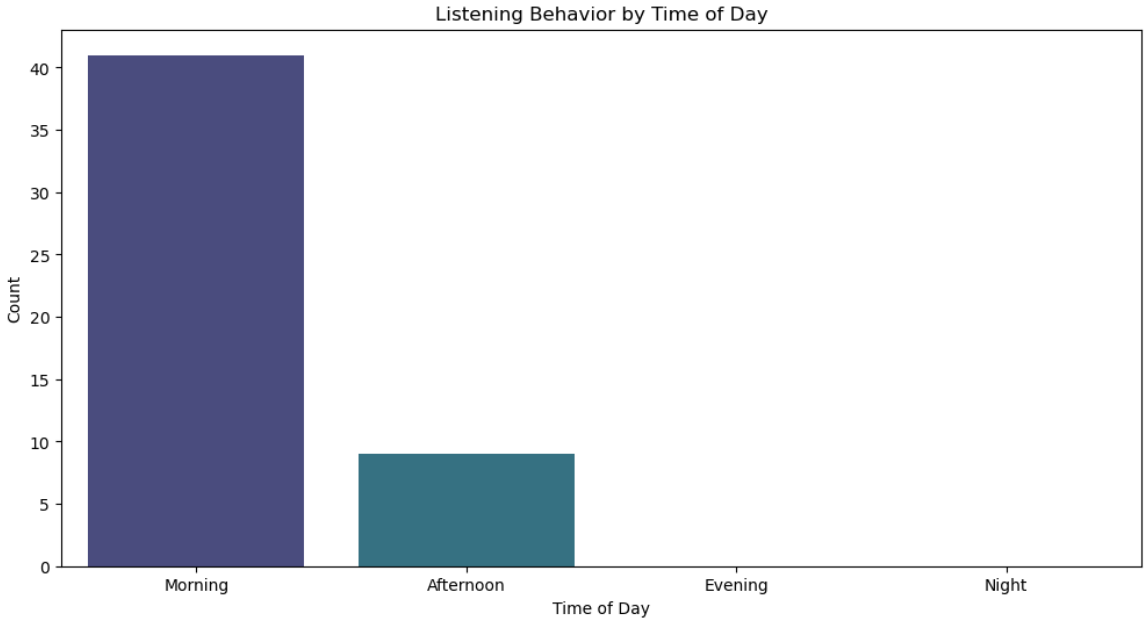


Figure 9: User Listening Behavior

The exploratory data analysis of the user dataframe provided valuable insights into the dataset's structure, user preferences, and listening habits. Key takeaways include a preference for contemporary music (with most tracks from recent years), a skew toward moderately popular songs, and a strong inclination for certain genres and artists. Analysis of audio features revealed patterns such as a preference for mid-tempo, less energetic tracks, with specific characteristics like higher acousticness and danceability. Time-series analysis and clustering added further depth, uncovering peak listening hours (evening and night) and grouping songs based on their audio characteristics. These findings not only enhance our understanding of the user’s musical preferences but also lay the groundwork for personalized recommendations, playlist creation, or predictive modeling. Overall, this detailed exploration helped in understanding user preferences and evolving trends, which informed music recommendations and potential industry strategies.

Spotify 1 Million Track Dataset EDA

This exploratory data analysis examines a dataset containing over 1 million music tracks, exploring various attributes such as artist names, track titles, genres, and audio features (e.g., danceability, energy, loudness). By analyzing these attributes, the goal is to uncover meaningful insights about musical characteristics, artist trends, genre distributions, temporal changes in popularity and audio features and evolving listener preferences. The analysis also investigates relationships between audio features and popularity, identifies dominant genres and artists, and examines the temporal evolution of musical characteristics. Techniques like clustering and feature correlation analysis provide further depth, segmenting tracks into distinct musical styles and highlighting how features interact. This EDA provides a holistic view of the dataset, offering actionable insights that can inform tasks like building music recommendation systems, curating personalized playlists, and predicting listener preferences.

1. **Initial Cleaning and Filtering**: Initial cleaning identified missing values in artist_name (15) and track_name (1), which were removed to ensure a complete dataset. The track_id column was verified for uniqueness, confirming no duplicate entries. Additionally, all columns were checked for consistent data types, ensuring reliable analysis in subsequent steps. The dataset was sorted by the year column, revealing a range of years between 2000 and 2023. Filtering was performed to refine the data: rows with loudness > 0 were removed, focusing on valid audio features. Similarly, tracks with time_signature < 3 were excluded to maintain quality in musical complexity. These actions ensured the dataset was ready for deeper exploration.

2. **Basic Descriptive Statistics:** Basic descriptive statistics were calculated for all numerical columns, offering a snapshot of the dataset's distribution. The average popularity score was 18.38, indicating a mix of lesser-known and moderately popular tracks. Danceability

averaged 0.54, and energy was slightly higher at 0.64, suggesting that the dataset contained moderately upbeat tracks. These statistics offered a foundation for understanding the dataset's musical landscape.

Table 3: Basic Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
popularity	1158550.00	18.38	15.89	0.00	5.00	15.00	29.00	100.00
year	1158550.00	2011.95	6.80	2000.00	2006.00	2012.00	2018.00	2023.00
danceability	1158550.00	0.54	0.18	0.00	0.41	0.55	0.68	0.99
energy	1158550.00	0.64	0.27	0.00	0.45	0.69	0.87	1.00
key	1158550.00	5.29	3.56	0.00	2.00	5.00	8.00	11.00
loudness	1158550.00	-8.99	5.68	-58.10	-10.84	-7.46	-5.28	0.00
mode	1158550.00	0.64	0.48	0.00	0.00	1.00	1.00	1.00
speechiness	1158550.00	0.09	0.13	0.00	0.04	0.05	0.09	0.97
acousticness	1158550.00	0.32	0.36	0.00	0.01	0.15	0.64	1.00
instrumentalness	1158550.00	0.25	0.37	0.00	0.00	0.00	0.61	1.00
liveness	1158550.00	0.22	0.20	0.00	0.10	0.13	0.29	1.00
valence	1158550.00	0.46	0.27	0.00	0.23	0.44	0.67	1.00
tempo	1158550.00	121.37	29.77	0.00	98.79	121.92	139.90	249.99
duration_ms	1158550.00	249598.55	149462.19	2073.00	181120.00	225760.00	286933.00	6000495.00
time_signature	1158550.00	3.89	0.47	0.00	4.00	4.00	4.00	5.00

3. **Tracks Per Year**: The number of tracks released annually was analyzed, revealing a peak in 2018 with 55,774 tracks. After 2020, track releases declined significantly, with only 38,078 tracks in 2023. This trend likely reflects shifts in the music industry or global circumstances like the COVID-19 pandemic. A line chart visualized these changes, highlighting high-activity periods.

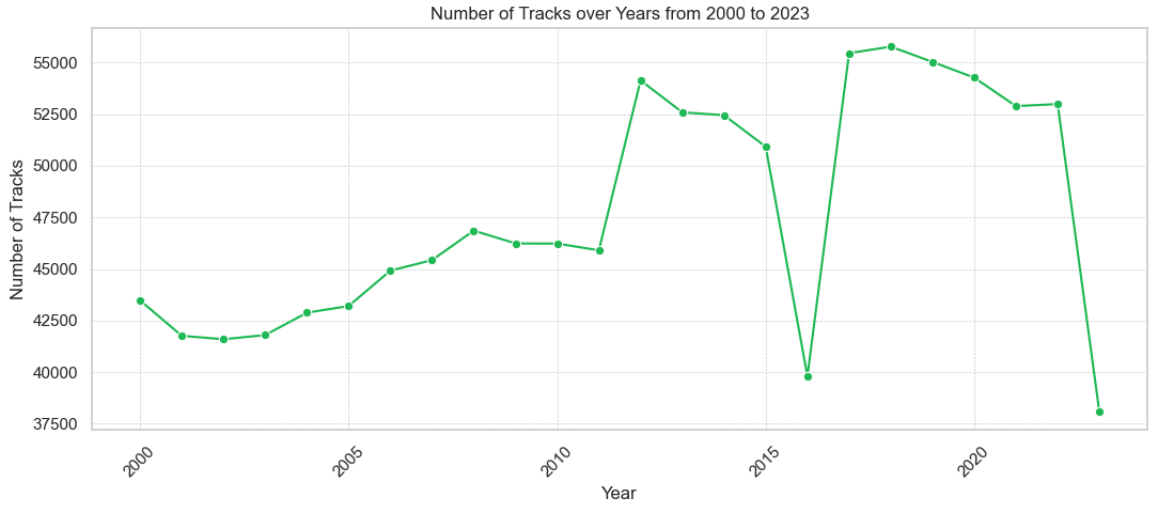


Figure 10: Number of tracks over years

4. **Artists Per Year**: The dataset showed a peak in unique artist contributions in 2020, with 14,190 artists releasing tracks. This period aligns with the rise of independent music production and streaming platforms, empowering more artists to release music. A line chart visualized these trends, indicating a decline in artist contributions post-2020.

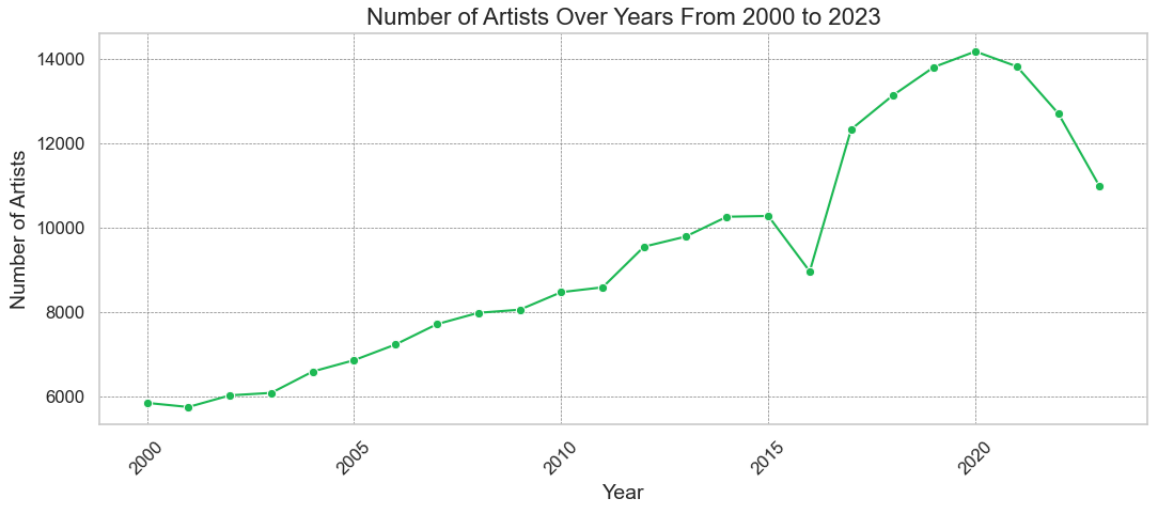


Figure 11: Number of artists released tracks over years

5. **Tracks by Genre**: The dataset included 82 unique genres, with black metal being the most represented, comprising 1.88% of all tracks. In contrast, songwriter was the least represented genre. A pie chart visualized the proportion of the top 10 genres, highlighting diversity in musical styles.

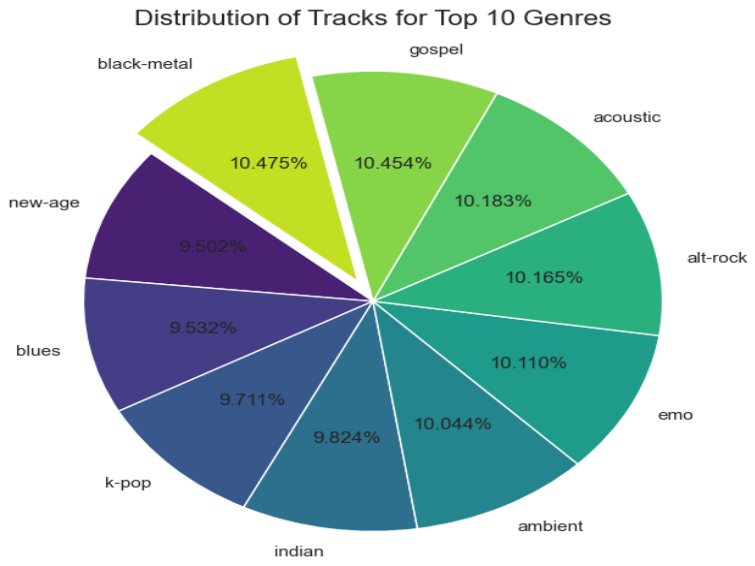


Figure 12: Number of Tracks in each Genre

6. **Popularity Analysis**: Pop had the highest average popularity score (55.67), followed by hip-hop (46.32) and rock (46.22). Among artists, Harry Styles, Rauw Alejandro, and Billie Eilish ranked highest, with consistent success across their tracks. A bar plot showcased the top 10 artists by average popularity, emphasizing their influence.

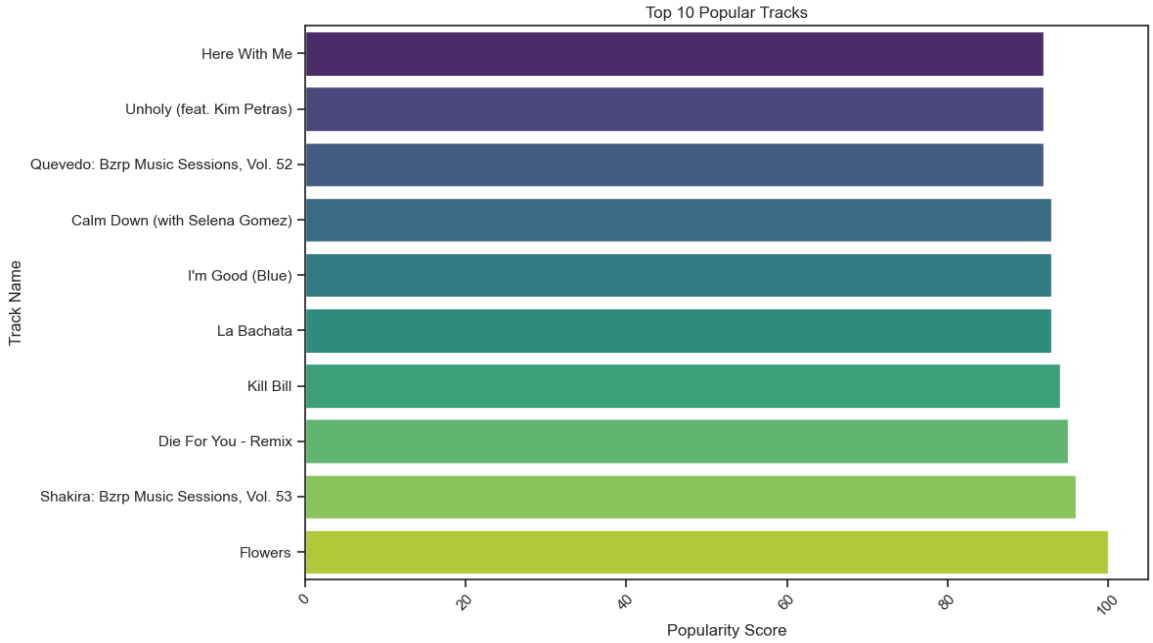


Figure 13: Most popular tracks

7. **Song Duration Trends**: Track duration trends over time revealed a gradual decline, reflecting a preference for shorter, more digestible tracks in the streaming era. Tracks released after 2020 were generally shorter than those from earlier years. A line chart visualized these changes, highlighting the industry's adaptation to listener preferences.

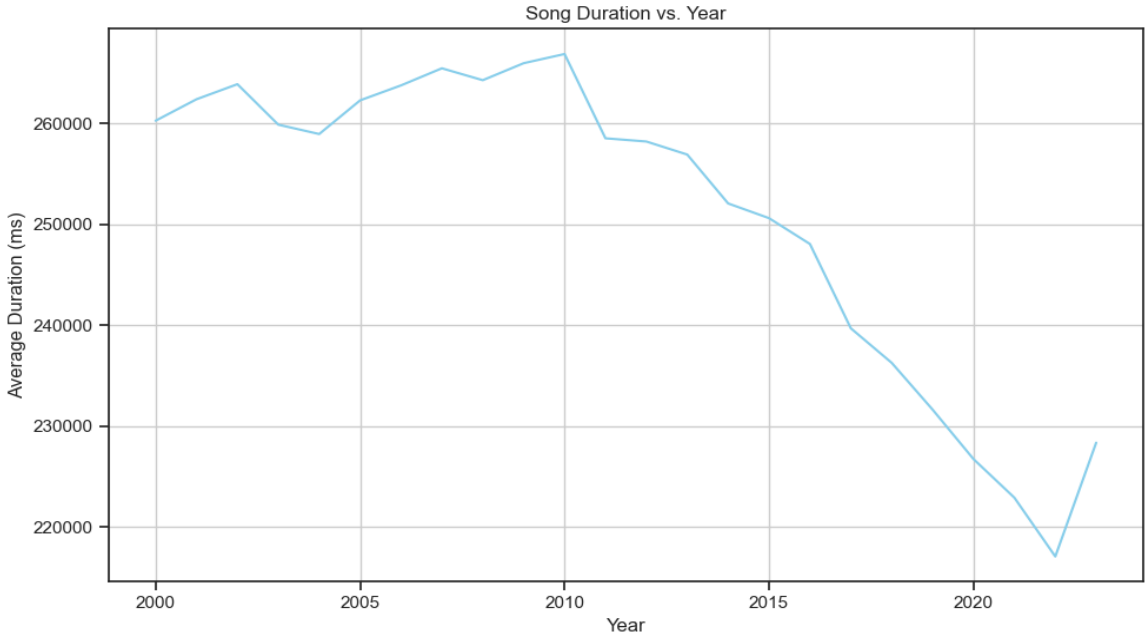


Figure 14: Song Duration vs. Year

8. **Top Artists by Popularity**: Artists were ranked based on average popularity scores. Harry Styles topped the list with an average score of 75.68, followed by Rauw Alejandro (71.92) and Billie Eilish (68.27). A bar plot highlighted the leading artists, offering insights into their consistent appeal.

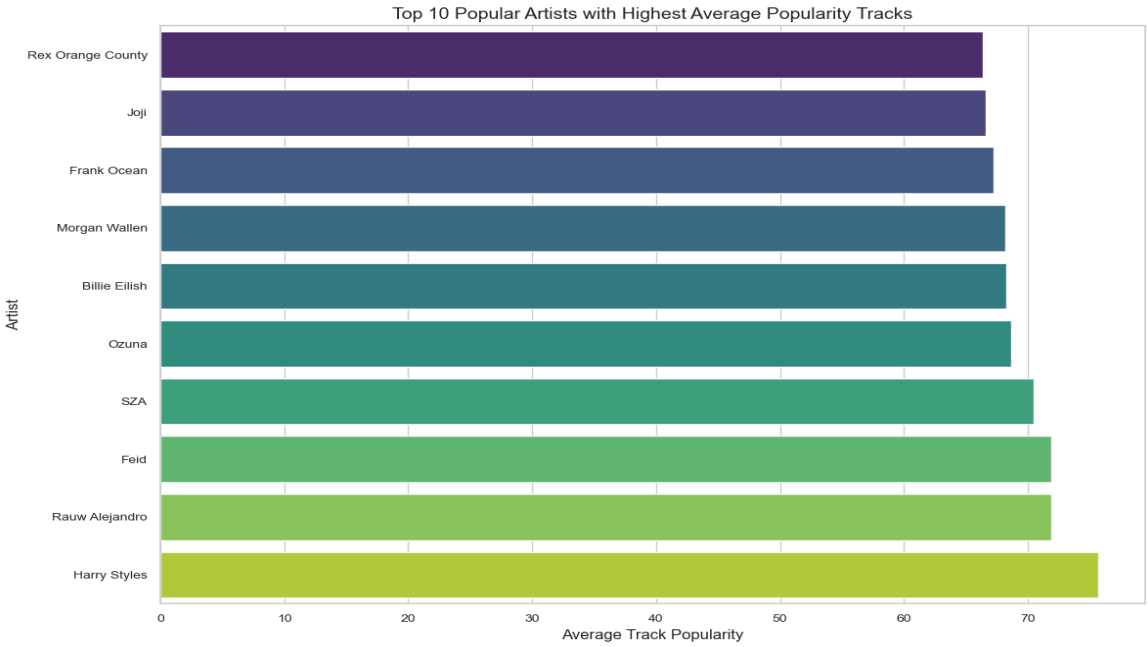


Figure 15: Top 10 artists: With highest Average popularity tracks

9. **Popularity vs. Audio Features**: Relationships between popularity and features like danceability, energy, and valence were explored. Popular tracks tended to have higher danceability and energy scores, indicating listener preferences for lively, engaging tracks. Pairplots visualized these relationships, providing actionable insights for music producers.

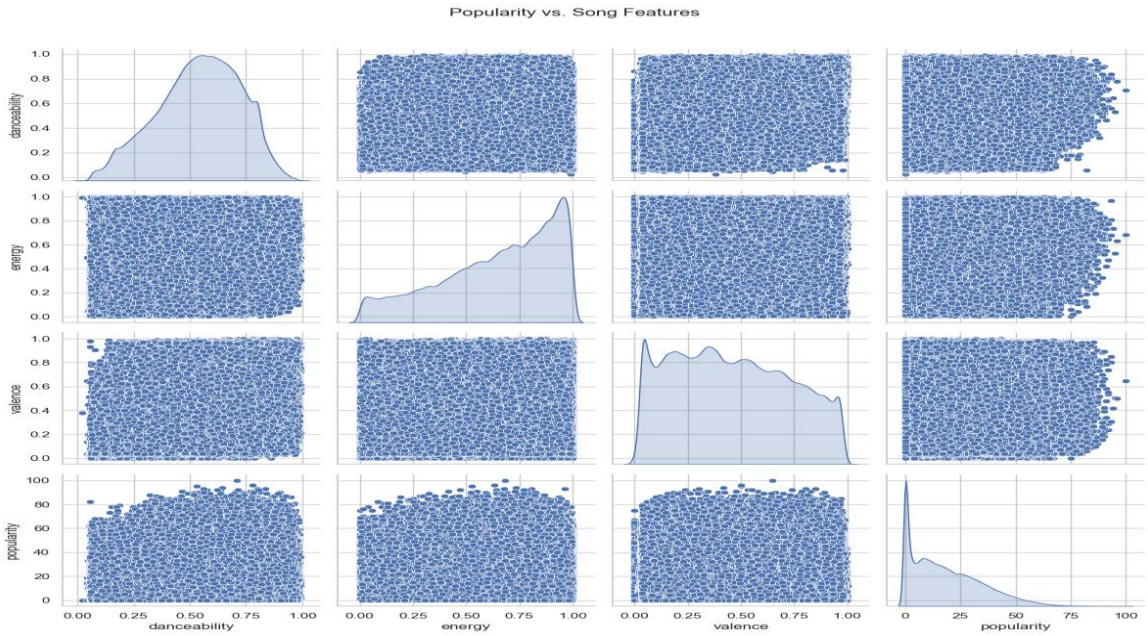


Figure 16: Popularity vs. Song Features

10. **Top Genres by Popularity**: Genres were ranked by their average popularity scores, with pop (55.67) leading, followed by rock (46.22) and hip-hop (46.32). This analysis identified the genres that resonate most with listeners, helping understand broader trends in music preferences.

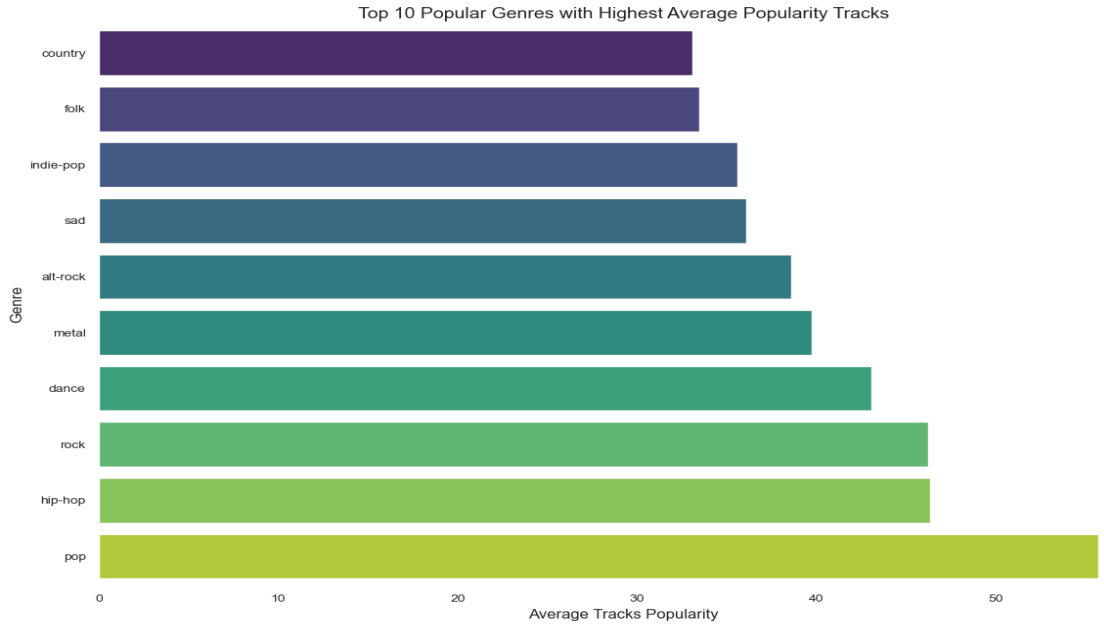


Figure 17: Top 10 Genres: With highest Average popularity tracks

11. **Loudness Trends**: A steady decline in loudness was observed over the years, reflecting the industry's shift away from the "loudness war." This aligns with the trend of prioritizing dynamic range and audio quality. A line chart visualized the decline in average loudness scores over time.

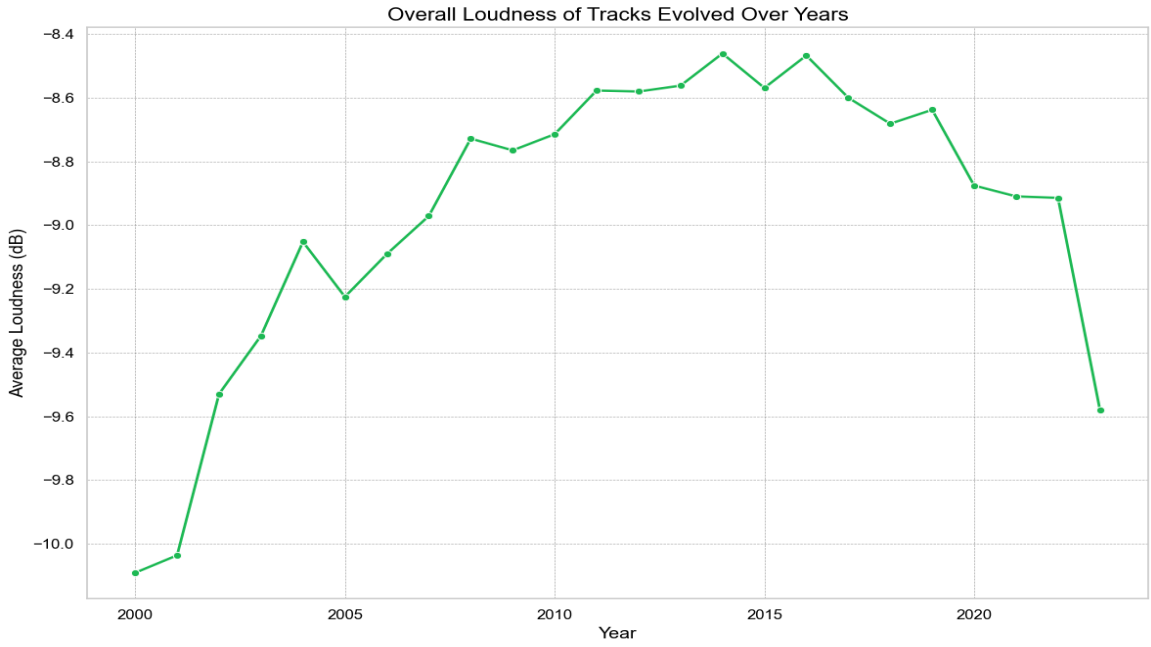


Figure 18: Loudness of tracks evolved over years

12. **Genre Trends Over Time**: The evolution of genres was visualized, showing significant growth in hip-hop and pop over the past two decades. Conversely, older genres like gospel and blues saw declines in representation. A stacked area chart illustrated these dynamic shifts, emphasizing the changing nature of listener preferences.

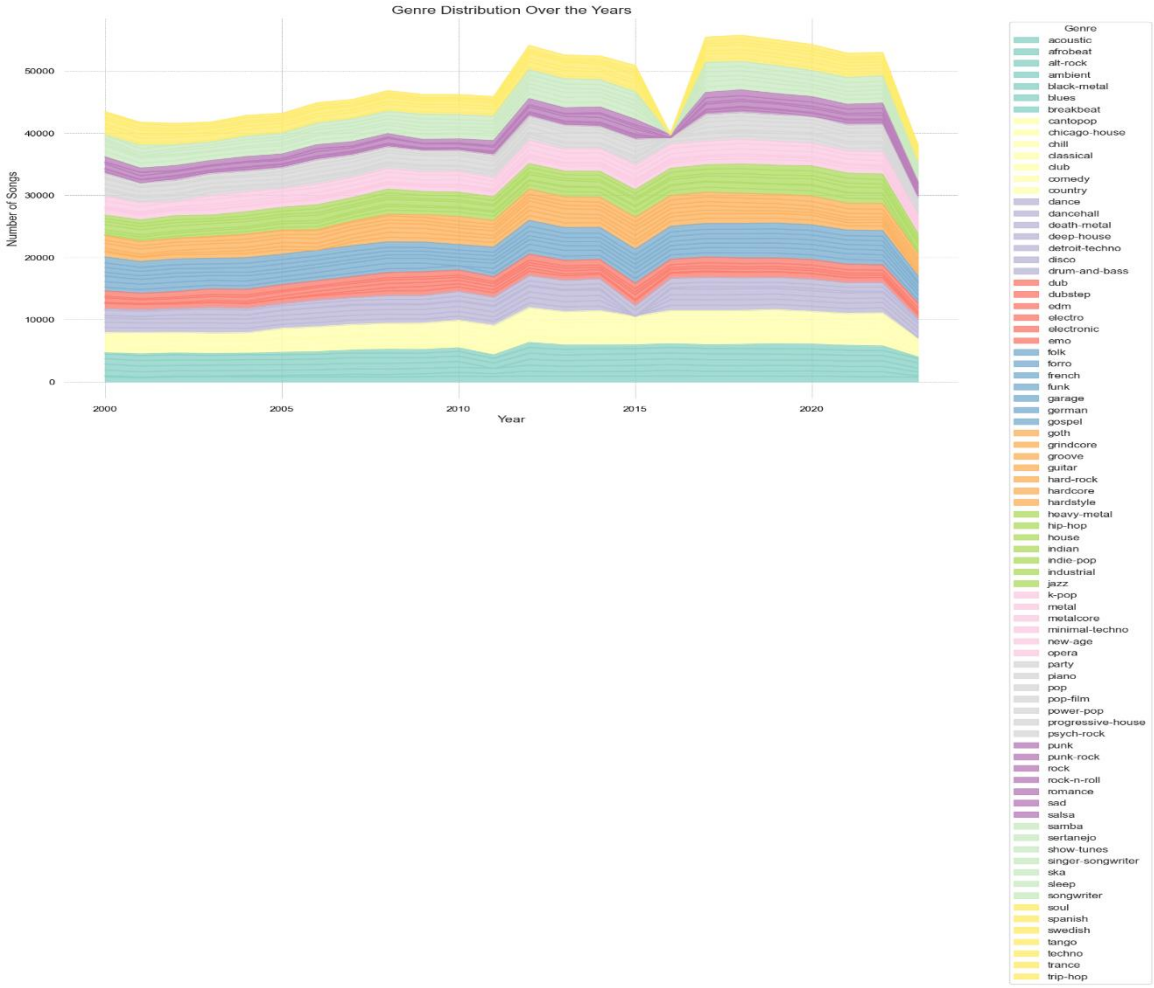


Figure 19: Genre Distribution over the years

13. **Danceability Trends**: Danceability scores showed a slight upward trend, indicating increasing listener preference for rhythmically engaging tracks. This growth aligns with the rising popularity of dance and electronic music genres. A line chart captured these temporal changes.

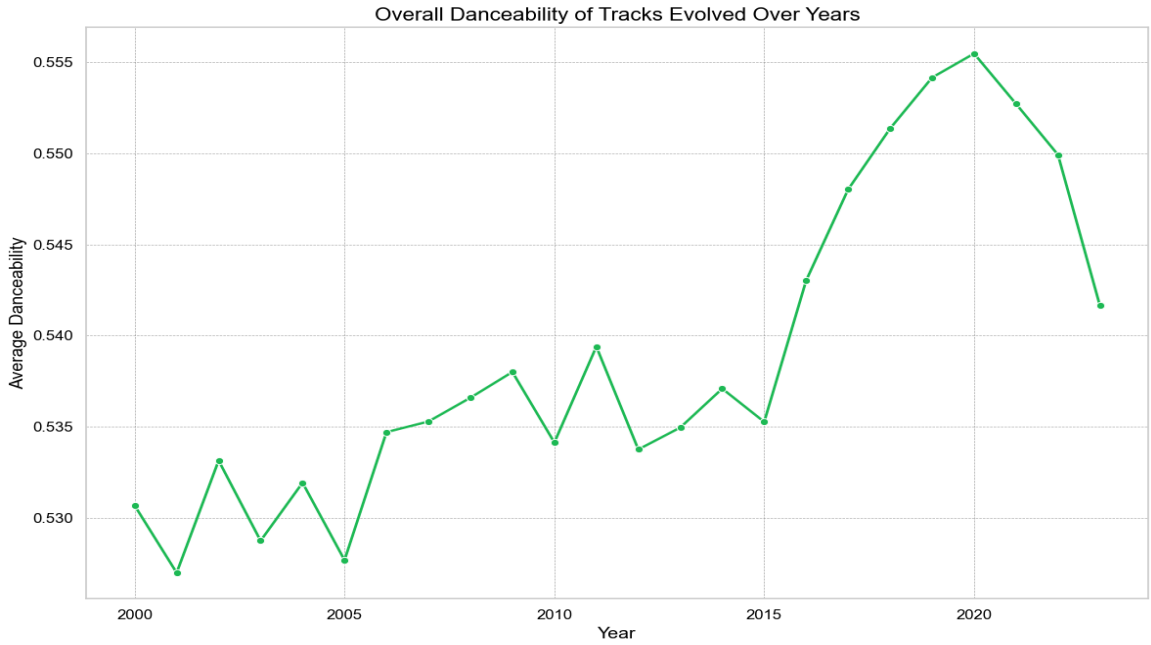


Figure 20: Danceability of tracks evolved over years

14. **Track Duration vs. Popularity**: Tracks of moderate length were more popular, while excessively long or short tracks tended to have lower popularity scores. A scatter plot visualized the correlation between track duration and popularity, providing insights into optimal track lengths for engagement.

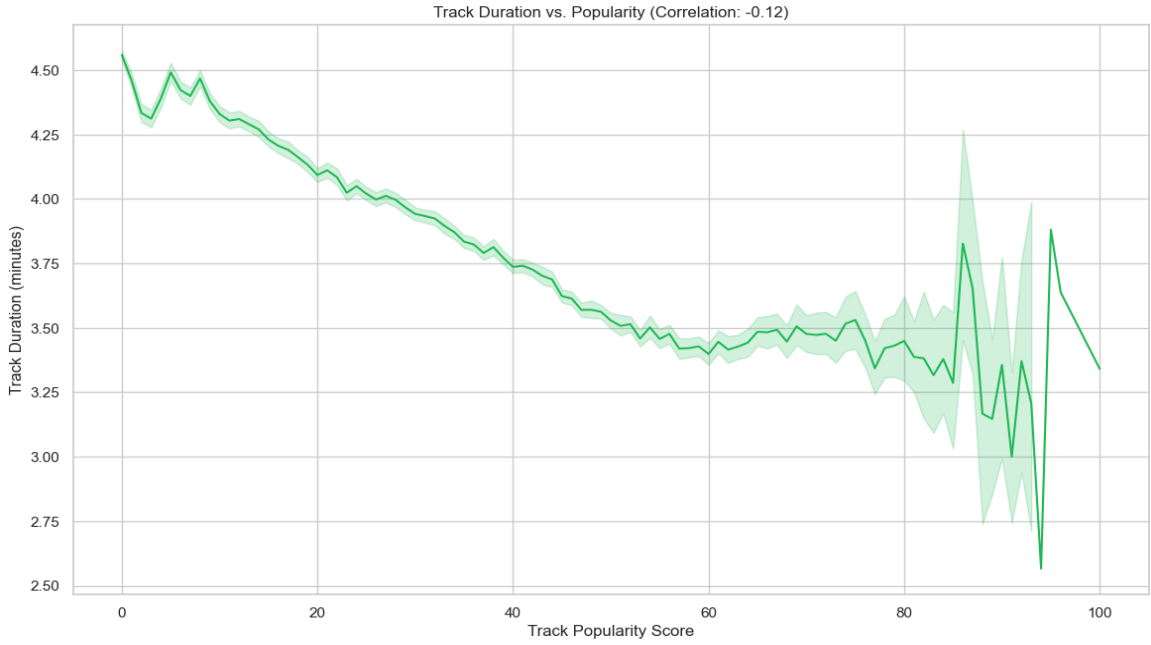


Figure 21: Track duration vs. Popularity

15. **Top Tracks by Genre**: The most popular tracks in each genre were identified. For example, in hip-hop, AP Dhillon's "Excuses" and "Brown Munde" stood out. This analysis offered a genre-specific view of listener preferences and standout performers.

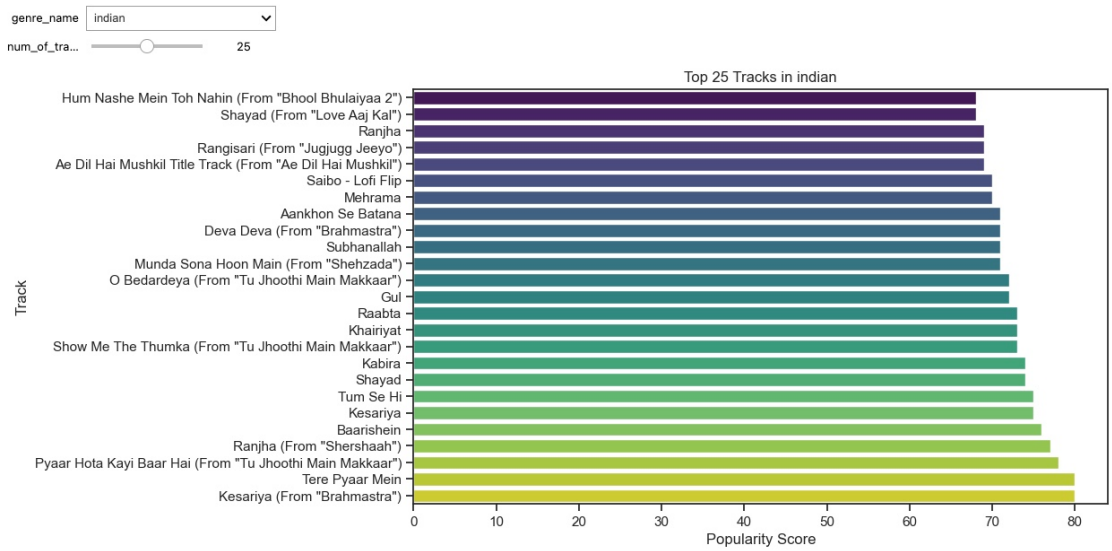


Figure 22: Top tracks in each genre

16. **Top Tracks for Top 10 Artists**: The analysis identified the most popular tracks from the top 10 artists by average popularity scores. Artists such as Harry Styles, Rauw Alejandro, and Billie Eilish were consistently at the top. For example, Harry Styles' "As It Was" and Billie Eilish's "Bad Guy" ranked among the most popular tracks. A bar plot visualized the top tracks for each artist, showcasing their contributions to the dataset. This analysis provided insight into which songs contributed most significantly to each artist’s popularity.

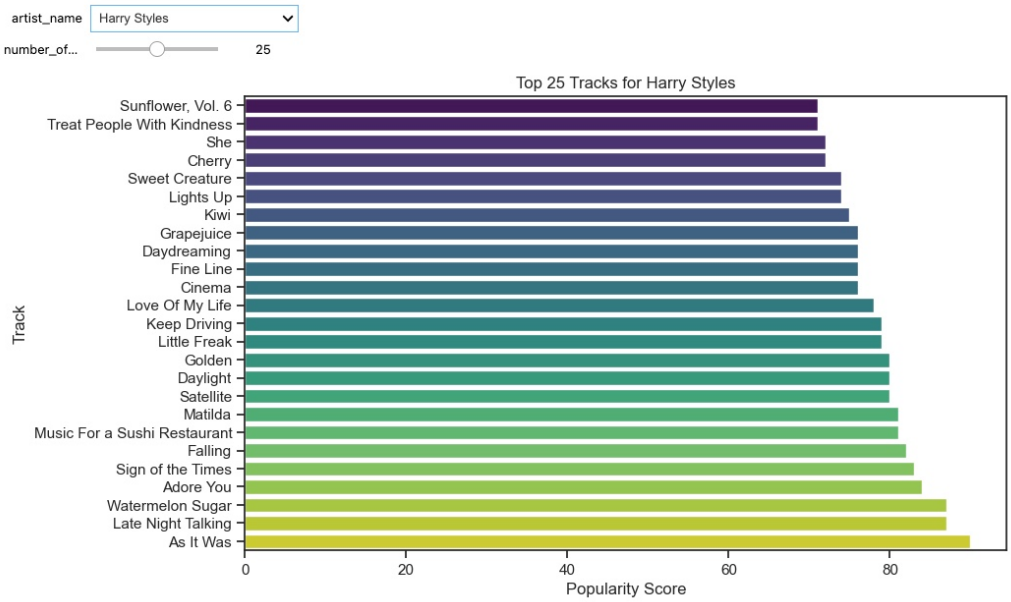


Figure 23: Top tracks for top 10 artists

17. **Cluster Analysis**: K-means clustering grouped tracks into four clusters based on features like danceability, energy, and valence. These clusters represented distinct musical styles, such as high-energy dance tracks or mellow acoustic compositions. Scatter plots visualized these clusters, providing a clear segmentation of musical characteristics.

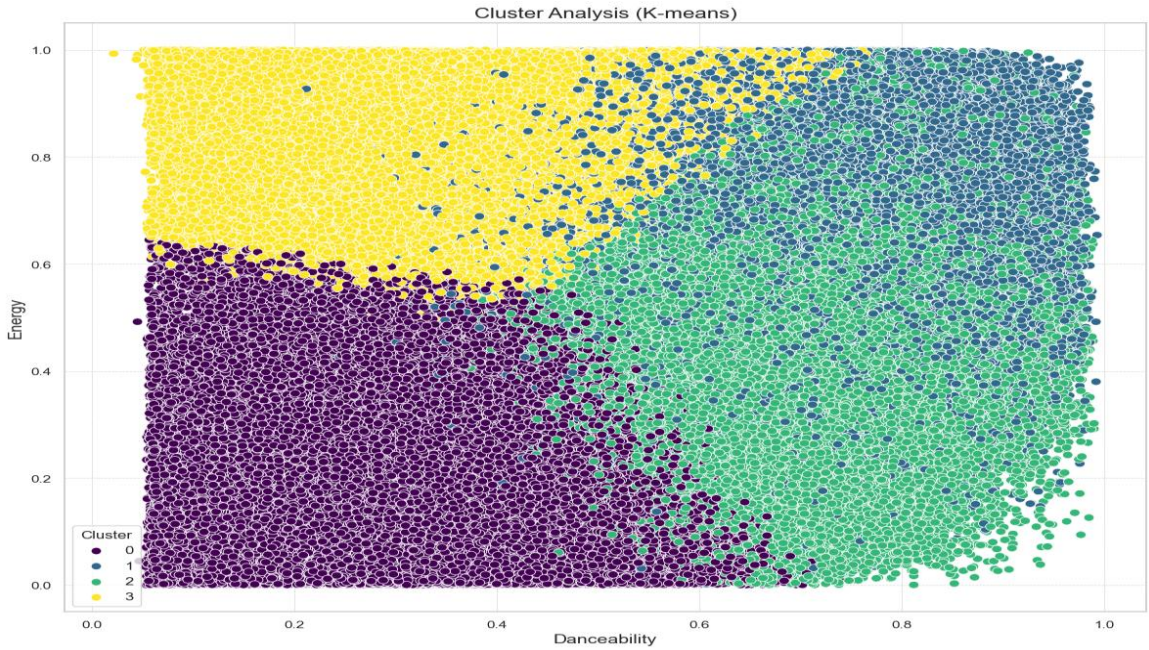


Figure 24: Cluster Analysis (K-means)

18. **Artist Popularity**: Artist popularity trends were examined over time. For specific artists like AP Dhillon, the analysis showcased their popularity trajectory year by year. A line chart visualized the artist's performance, indicating growth periods and potential declines. For instance, AP Dhillon's discography showed consistent popularity in recent years, with hits like "Brown Munde" dominating charts. Additionally, genre analysis revealed that certain artists heavily contributed to specific genres (e.g., AP Dhillon to hip-hop), highlighting their influence within these categories.

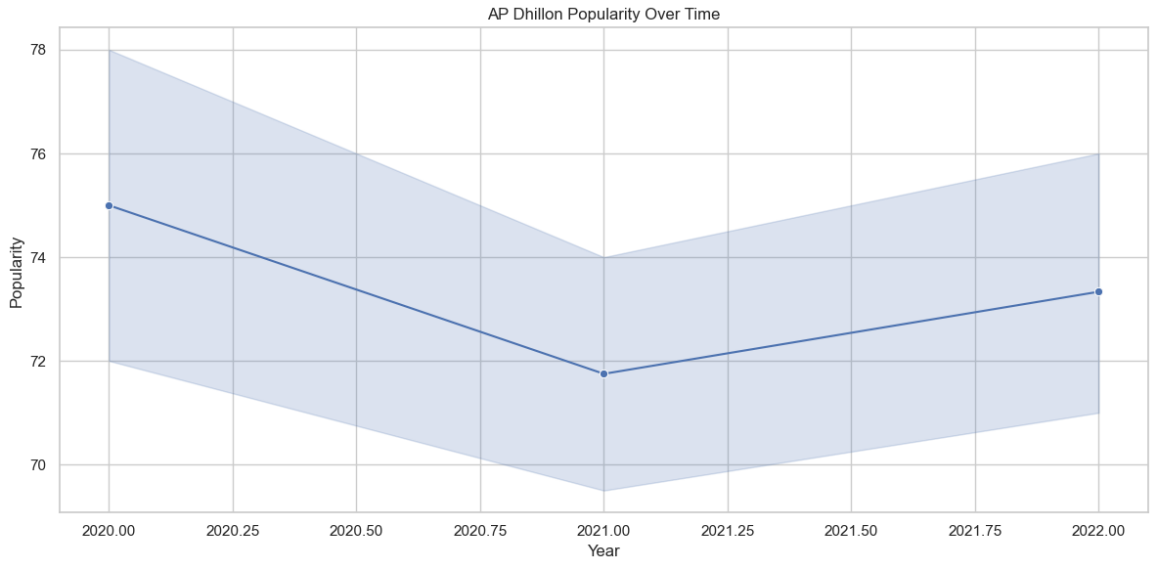


Figure 25: Artist popularity

19. **Speechiness vs. Instrumentalness**: The relationship between speechiness and instrumentalness was explored to analyze the trade-off between lyrical content and instrumental focus. Tracks with high speechiness values leaned toward lyrical or spoken content, such as rap or podcasts. In contrast, tracks with high instrumentalness were typically acoustic or classical compositions. A scatter plot visualized this relationship, showing distinct groupings for these two styles. This analysis revealed the dataset's diversity, accommodating both vocal-heavy and purely instrumental tracks.

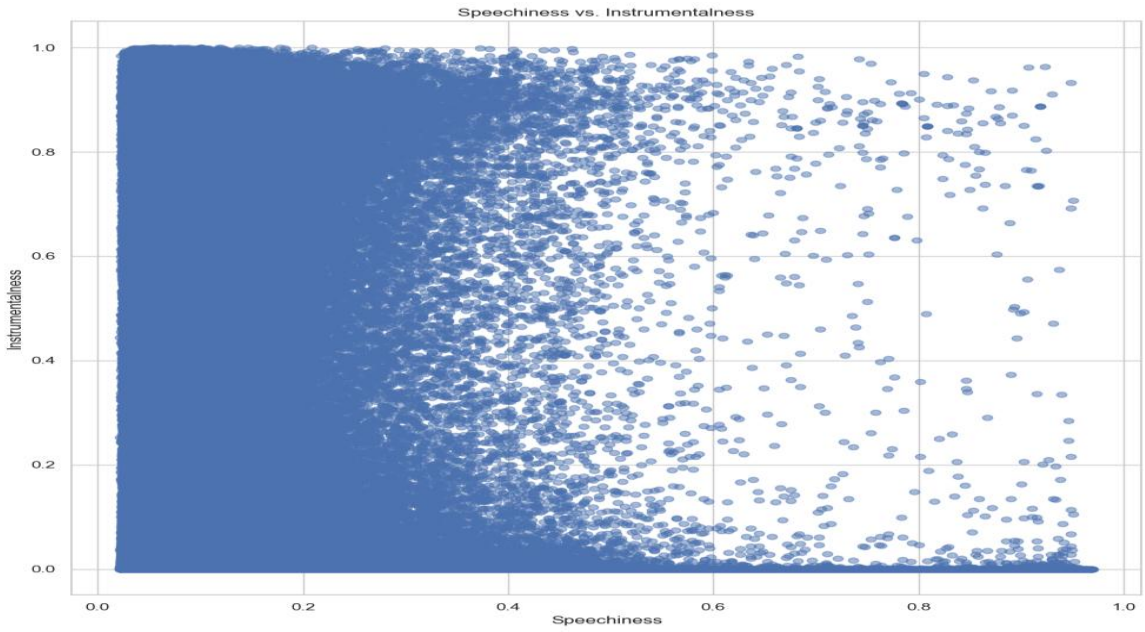


Figure 26: Speechiness vs. Instrumentalness

20. **Energy vs. Loudness**: The top 25 genres in the dataset were analyzed for their global popularity. Genres like pop, hip-hop, and rock emerged as the most dominant in terms of both track count and average popularity. Scatter plots compared audio features like energy and loudness across these genres, showing how their characteristics differed. For instance, hip-hop tracks had higher energy and loudness, while classical and acoustic genres showed higher instrumentalness. A multi-faceted visualization highlighted the diverse appeal of these genres worldwide.

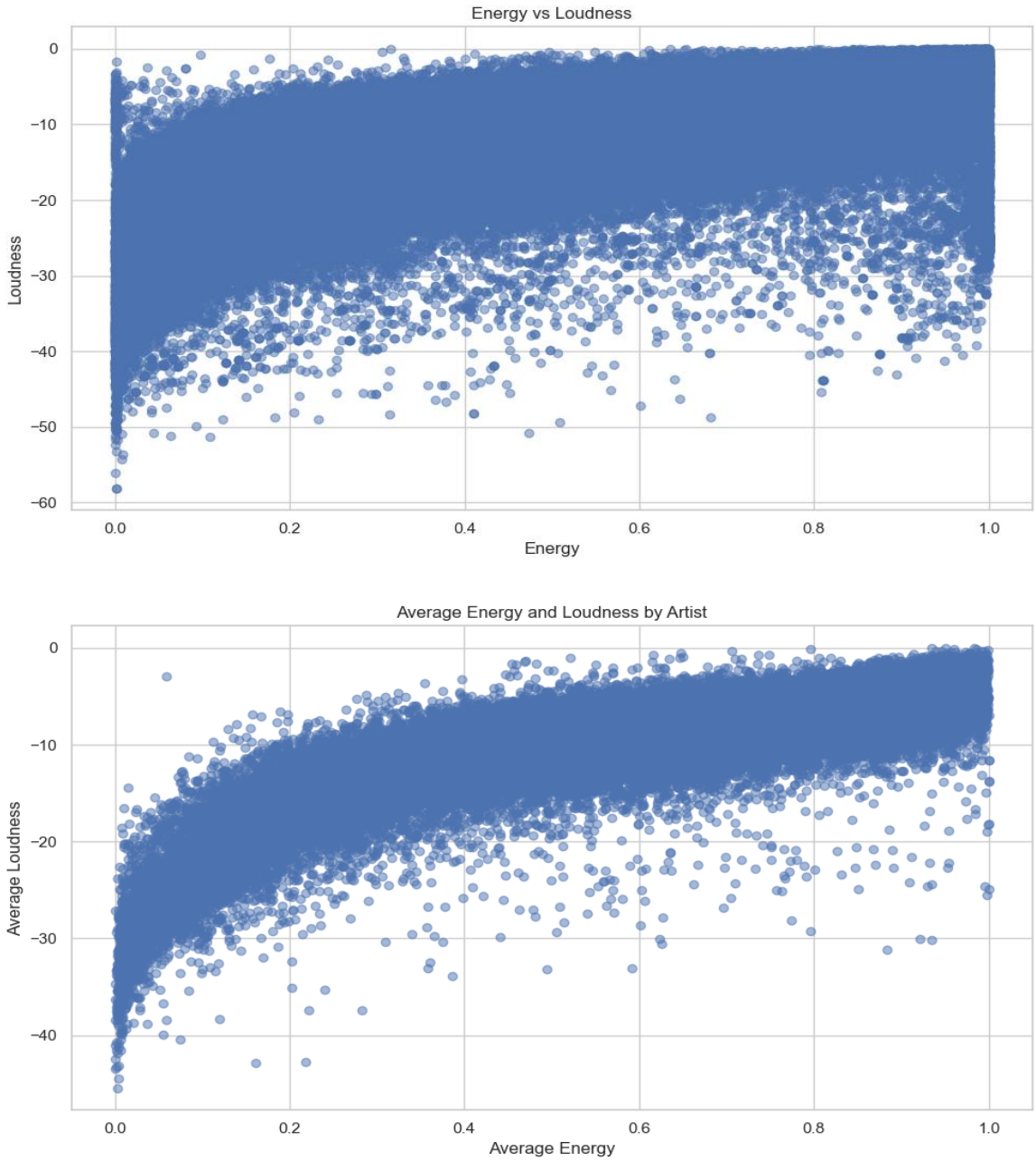
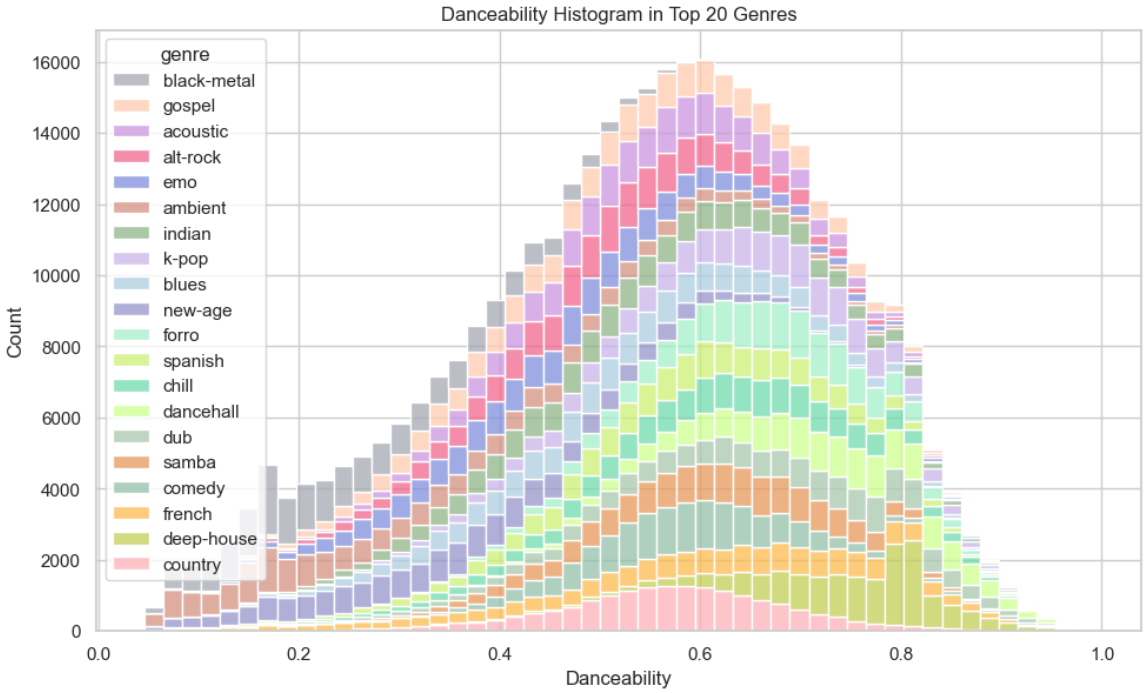
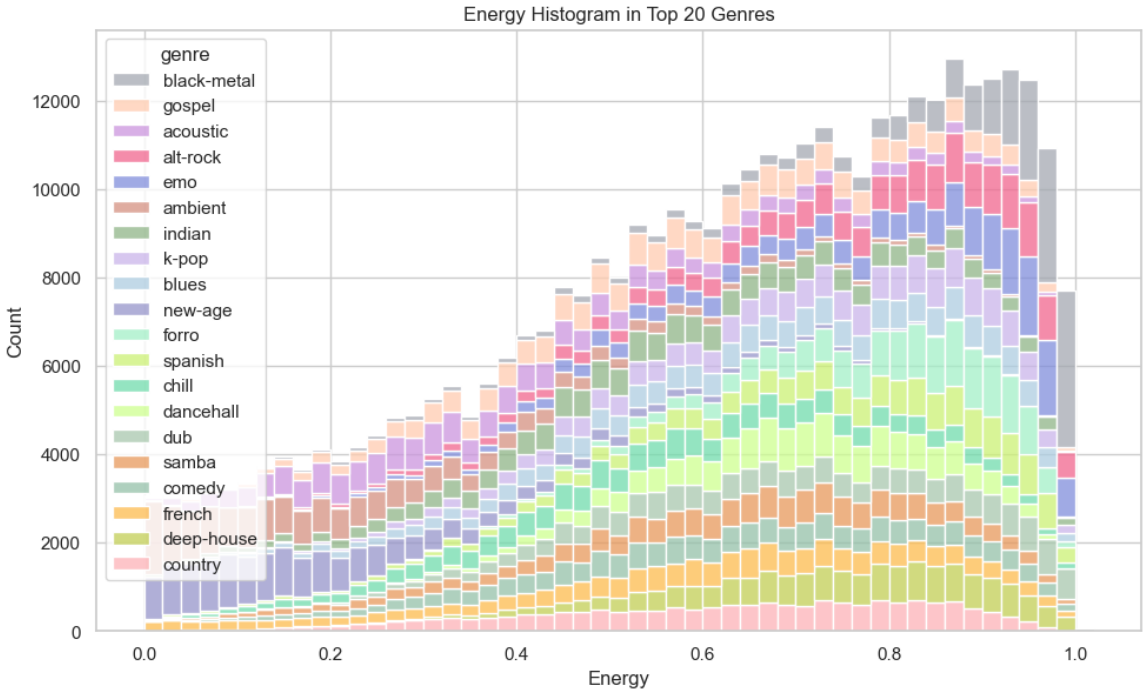
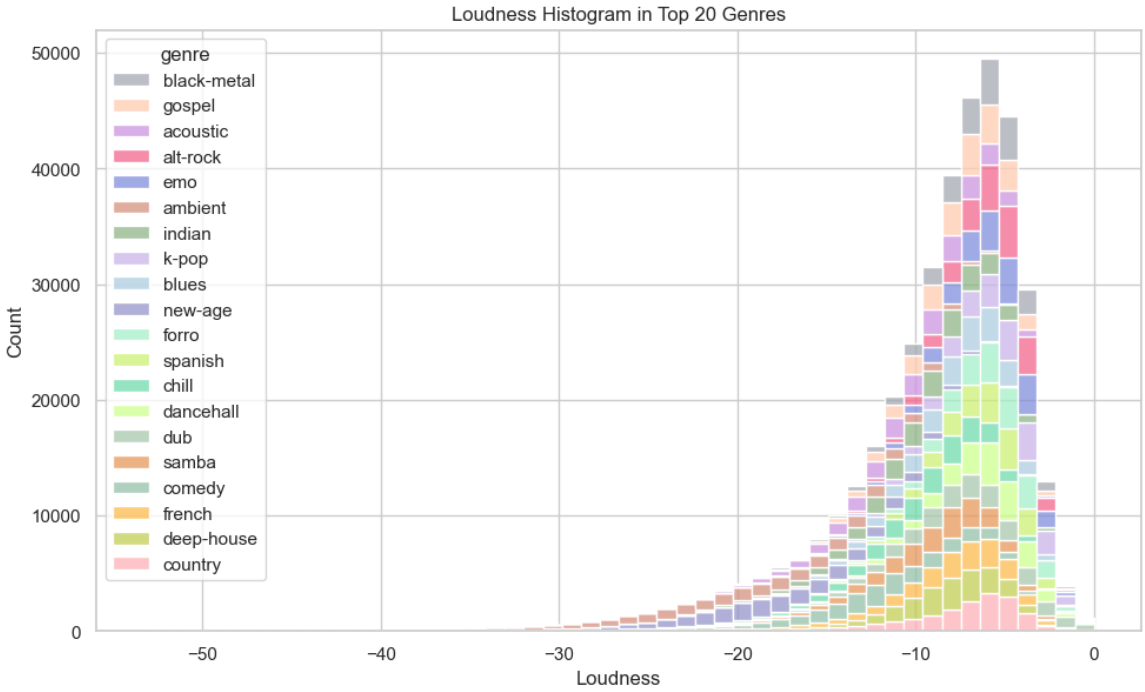


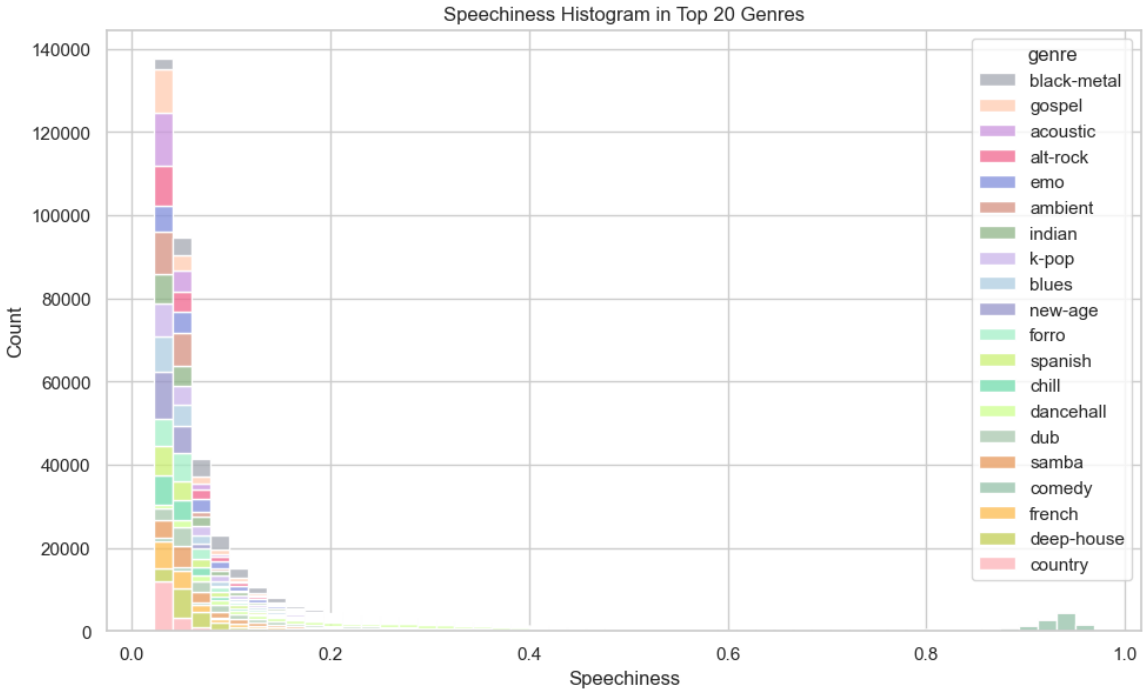
Figure 27: Energy vs. Loudness

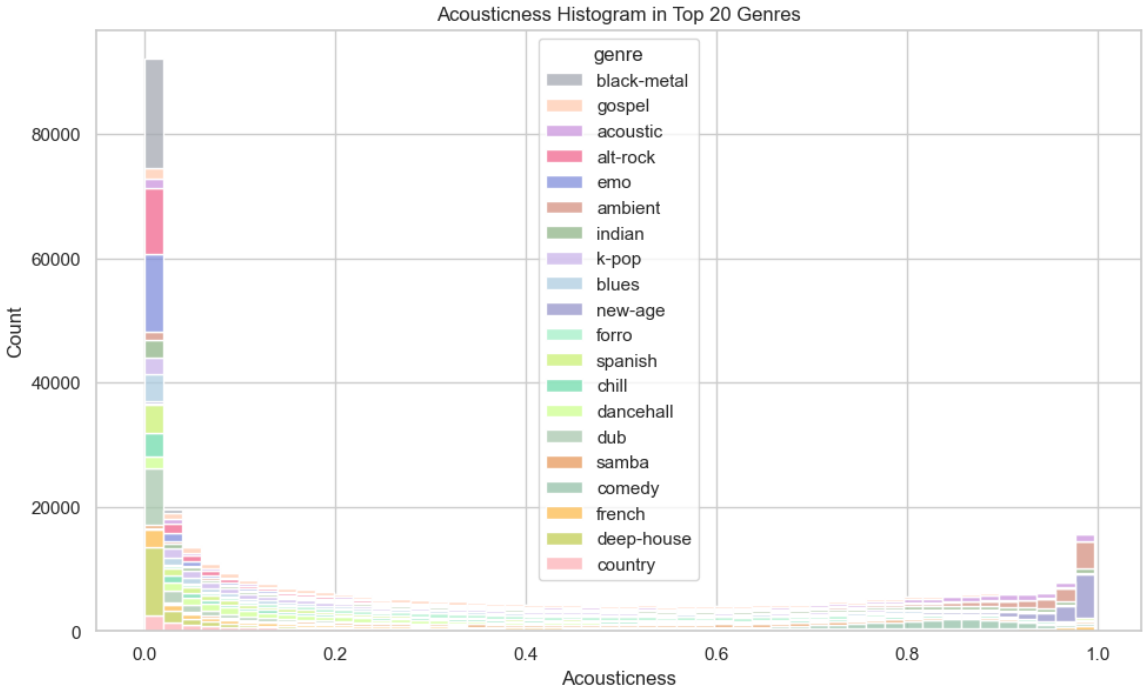
21. **Audio Feature Histograms**: Histograms for features like loudness, energy, and valence revealed their distributions. For instance, energy scores were skewed toward higher values, indicating a preference for lively tracks. These visualizations offered a comprehensive understanding of musical characteristics within the dataset.

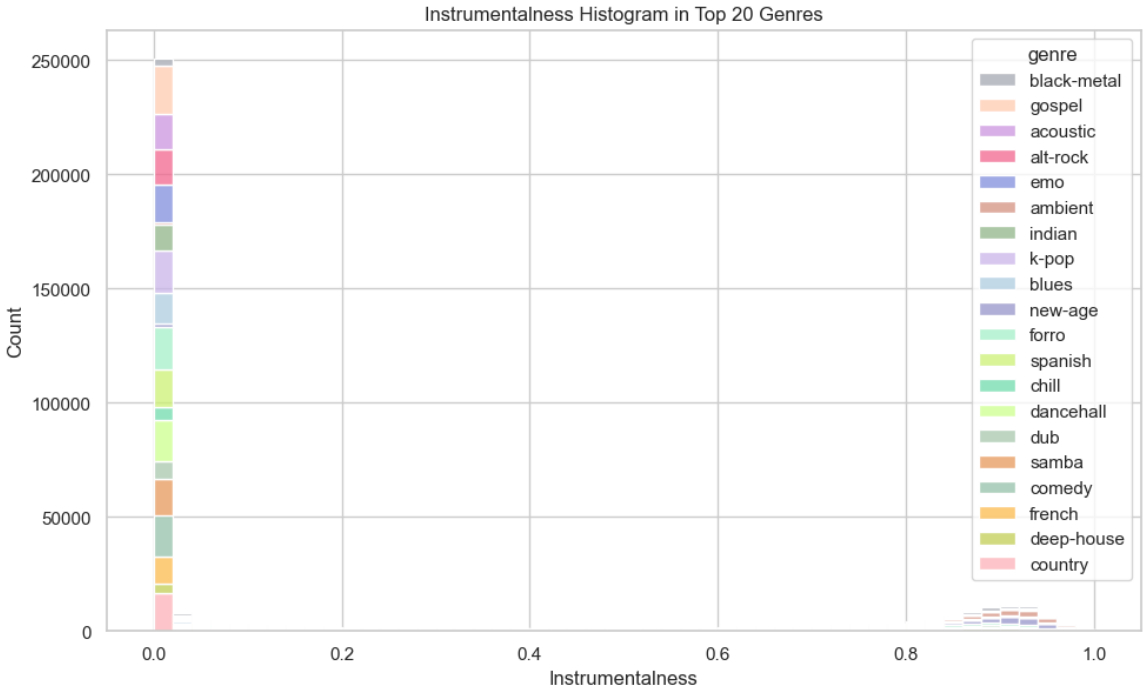

Danceability Histogram in Top 20 Genres
genre
black-metal
gospel
acoustic
alt-rock
emo
ambient
indian
k-pop
blues
new-age
forro
spanish
chill
dancehall
dub
samba
comedy
french
deep-house
country
Count
Danceability
0
2000
4000
6000
8000
10000
12000
14000
16000
0.0
0.2
0.4
0.6
0.8
1.0

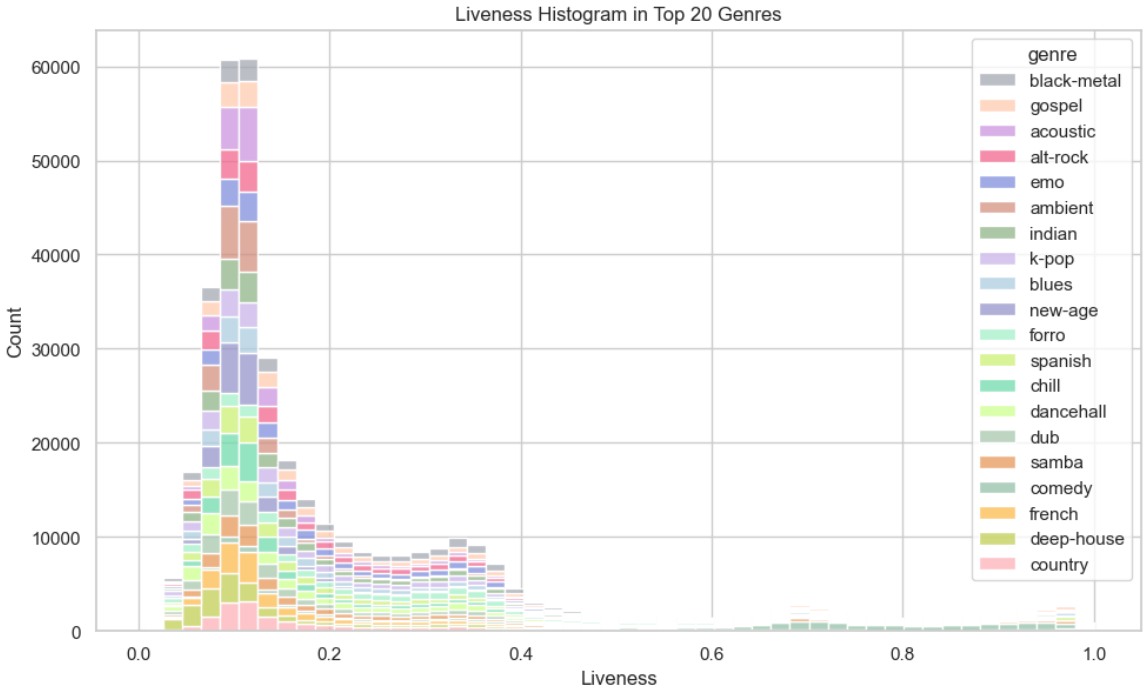

Energy Histogram in Top 20 Genres
genre
black-metal
gospel
acoustic
alt-rock
emo
ambient
indian
k-pop
blues
new-age
forro
spanish
chill
dancehall
dub
samba
comedy
french
deep-house
country
Count
Energy
0
2000
4000
6000
8000
10000
12000
0.0
0.2
0.4
0.6
0.8
1.0

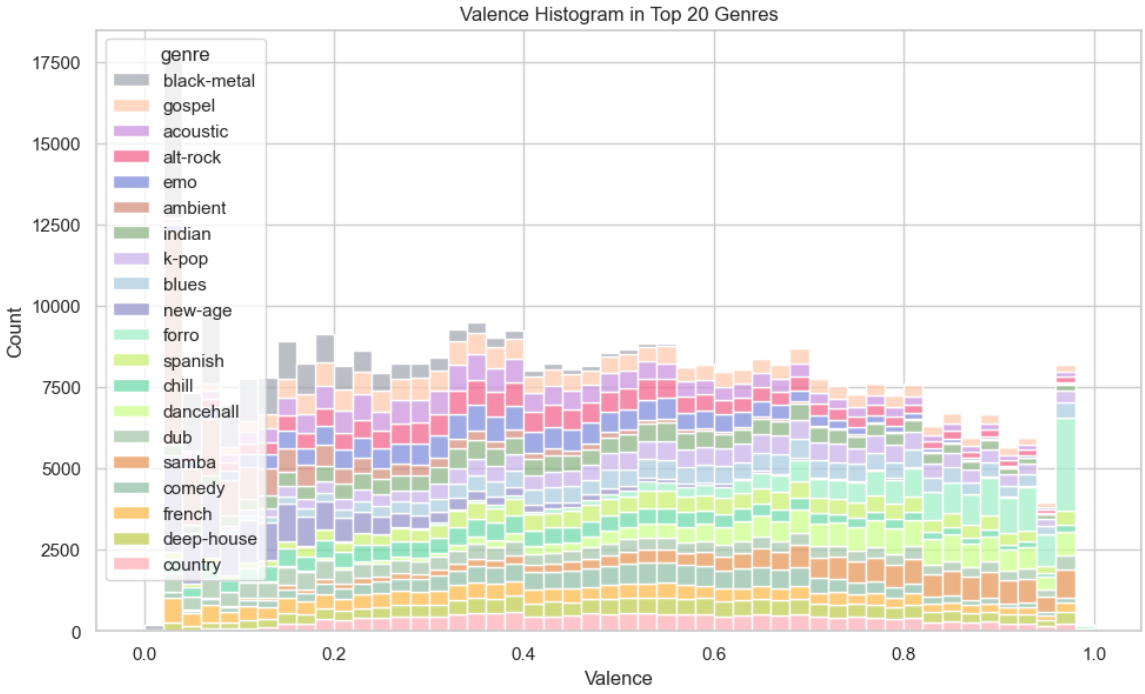

Loudness Histogram in Top 20 Genres
genre
black-metal
gospel
acoustic
alt-rock
emo
ambient
indian
k-pop
blues
new-age
forro
spanish
chill
dancehall
dub
samba
comedy
french
deep-house
country
Count
50000
40000
30000
20000
10000
0
−50
−40
−30
−20
−10
0
Loudness


Speechiness Histogram in Top 20 Genres
genre
black-metal
gospel
acoustic
alt-rock
emo
ambient
indian
k-pop
blues
new-age
forro
spanish
chill
dancehall
dub
samba
comedy
french
deep-house
country
Count
140000
120000
100000
80000
60000
40000
20000
0
0.0
0.2
0.4
0.6
0.8
1.0
Speechiness


Acousticness Histogram in Top 20 Genres
genre
black-metal
gospel
acoustic
alt-rock
emo
ambient
indian
k-pop
blues
new-age
forro
spanish
chill
dancehall
dub
samba
comedy
french
deep-house
country
Count
0
20000
40000
60000
80000
0.0
0.2
0.4
0.6
0.8
1.0
Acousticness


Instrumentalness Histogram in Top 20 Genres
genre
black-metal
gospel
acoustic
alt-rock
emo
ambient
indian
k-pop
blues
new-age
forro
spanish
chill
dancehall
dub
samba
comedy
french
deep-house
country
Count
0
50000
100000
150000
200000
250000
0.0
0.2
0.4
0.6
0.8
1.0
Instrumentalness


Liveness Histogram in Top 20 Genres
Count
Liveness
0
10000
20000
30000
40000
50000
60000
0.0
0.2
0.4
0.6
0.8
1.0
genre
black-metal
gospel
acoustic
alt-rock
emo
ambient
indian
k-pop
blues
new-age
forro
spanish
chill
dancehall
dub
samba
comedy
french
deep-house
country


Valence Histogram in Top 20 Genres
Count
Valence
0
2500
5000
7500
10000
12500
15000
17500
0.0
0.2
0.4
0.6
0.8
1.0
genre
black-metal
gospel
acoustic
alt-rock
emo
ambient
indian
k-pop
blues
new-age
forro
spanish
chill
dancehall
dub
samba
comedy
french
deep-house
country

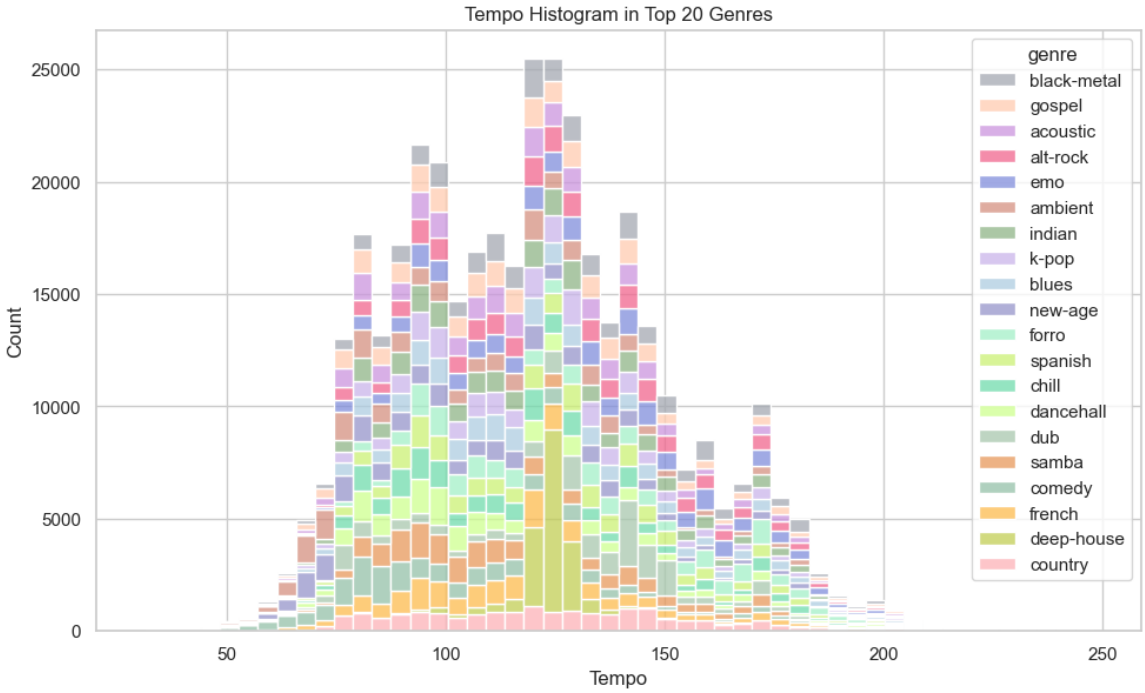


Figure 28: Audio Features Histograms

The exploratory data analysis of the 1 million tracks dataset revealed valuable insights into the structure and trends within the music industry. The dataset spans over two decades (2000–2023), capturing diverse musical styles, evolving listener preferences, and the dynamic nature of audio characteristics. Pop and hip-hop emerged as dominant genres, with consistent popularity over the years. Leading artists such as Harry Styles, Rauw Alejandro, and Billie Eilish consistently performed well, contributing significantly to their respective genres. Popular tracks tended to have higher danceability and energy, reflecting listener preferences for lively, rhythmic compositions. Temporal trends indicated a gradual decline in loudness and a slight rise in danceability, suggesting evolving production techniques and audience expectations. Track releases peaked in 2018, with artist activity reaching its zenith in 2020. However, a decline in releases and artist contributions was observed post-2020, potentially reflecting shifts in industry dynamics. K-means clustering highlighted distinct musical styles, while correlation analyses revealed meaningful relationships between audio

features, such as the positive link between danceability and energy. This analysis offers a foundation for advanced tasks like recommendation systems, playlist generation, and predictive modeling. By leveraging these insights, stakeholders can better understand listener behavior, identify emerging trends, and create personalized music experiences. The findings emphasize the richness of the dataset and its potential to drive innovation in music analytics.

Data Gaps

Identifying data gaps is a crucial part of analyzing any dataset, as it helps highlight potential limitations and areas for improvement in the data collection and processing pipeline. In this context, the gaps in the user-collected and Million Song Dataset can impact the effectiveness of downstream tasks, such as recommendation systems, trend analysis, or user profiling. The following gaps outline areas where the data is either incomplete or limited, along with the implications for insights and system performance:

1. **Missing User Interaction Data**
    - **Gap**: The user-collected dataset does not contain user interaction types (e.g., likes, skips, repeats), which are critical for understanding user preferences in a recommendation system.
    - **Impact**: Lack of interaction granularity could reduce the accuracy of collaborative filtering and limit personalized recommendations.
2. **Incomplete User Profiles**
    - **Gap**: User data lacks demographic information such as age, location, or language preferences, which could provide additional insights for tailoring recommendations.

- **Impact**: Without this data, recommendations might not account for regional or cultural preferences, reducing user satisfaction.

3. **Limited Genre Diversity**
   - **Gap**: The genre column in both datasets may not provide detailed subgenres or cross-genre classifications (e.g., fusion genres).
   - **Impact**: The recommendation engine might oversimplify recommendations, leading to repetitive or generic suggestions.
4. **Temporal Information**
   - **Gap**: While the played_at column exists in the user-collected data, it lacks contextual temporal details such as listening trends across different time blocks (morning, evening, weekends).
   - **Impact**: The system might fail to capture and adapt to time-based listening patterns.
5. **Popularity Bias**
   - **Gap**: Both datasets rely heavily on the popularity column, which may disproportionately favor highly popular tracks and ignore niche or less-known songs.
   - **Impact**: Recommendations might be skewed toward mainstream music, potentially alienating users seeking diversity.
6. **Data Size Imbalance**
   - **Gap**: The user-collected dataset has far fewer records compared to the Million Song Dataset, which can lead to underrepresentation of user-specific preferences.
   - **Impact**: Cold-start problems may persist due to insufficient user-specific data.
7. **Potential Missing Values**

- **Gap**: Columns like artist_name, track_name, and track_id in the Million Song Dataset contain missing values (Non-Null Count indicates some records are incomplete).
- **Impact**: Missing track or artist information might result in errors or incomplete recommendations.

Associated Risks

While working with the user-collected and Million Song Dataset, several risks emerge that could impact the quality and reliability of downstream processes like recommendation systems. Each risk, if left unaddressed, can hinder the system's ability to deliver accurate, diverse, and scalable results, ultimately affecting user satisfaction and engagement. Identifying these risks and implementing effective mitigation strategies is critical for developing robust and user-centered solutions: These risks stem from inherent challenges such as:

1. **Data Sparsity Risk**
    - **Risk**: Sparse user-specific data could lead to poor collaborative filtering performance, especially for new users.
    - **Mitigation**: Implement hybrid filtering to combine collaborative and content-based approaches.
2. **Cold-Start Problem**
    - **Risk**: Insufficient user history in the user-collected dataset could result in generic recommendations.
    - **Mitigation**: Use genre-based recommendations or sentiment filters to cater to new users.

3. **Bias in Recommendations**
    - **Risk**: Over-reliance on popularity metrics may exclude diverse or lesser-known songs, resulting in biased suggestions.
    - **Mitigation**: Introduce diversity constraints in the recommendation algorithm.
4. **Scalability Challenges**
    - **Risk**: The imbalance between the user-collected data size and the Million Song Dataset could impact the scalability of the system.
    - **Mitigation**: Utilize sampling or feature engineering to balance data processing loads.
5. **Incomplete or Incorrect Data Risk**
    - **Risk**: Missing or incomplete records in the Million Song Dataset might lead to incorrect recommendations.
    - **Mitigation**: Implement data validation and imputation techniques during preprocessing.

Impact of Data Gaps on Final Project

Data gaps can significantly affect the success and usability of the final project, particularly in the domain of music recommendation systems. These gaps introduce challenges that impact key functionalities and user experience, limiting the project's effectiveness and potential to stand out in a competitive landscape. Below is a breakdown of the specific impacts of these gaps:

1. **Reduced Recommendation Accuracy**: Incomplete user data and sparse interactions limit the system’s ability to provide highly accurate and personalized recommendations, which is a core feature of the project.

2. **Cold-Start Challenges**: The system may struggle to engage new users effectively due to insufficient mechanisms for addressing the cold-start problem.
3. **User Dissatisfaction**: Repetitive or generic recommendations resulting from biases and missing contextual data might lead to a poor user experience.
4. **Loss of Competitive Edge**: The inability to handle niche or diverse user preferences may impact the project's ability to differentiate itself from existing music recommendation systems.
5. **Scalability and Operational Issues**: The data imbalance could result in scalability issues, increasing system latency during recommendation generation for large datasets.

Next Steps to Address Data Gaps

To overcome the identified data gaps and mitigate their associated risks, a clear and actionable plan is essential. The following steps outline strategies to enrich the dataset, improve algorithm performance, and ensure scalability:

1. **Data Augmentation**: Enrich user-collected data by gathering more interaction types (likes, skips, playlists) and demographics.
2. **Feature Engineering**: Create derived features like "time-based listening trends" and "genre diversity score" to enhance recommendation personalization.
3. **Handling Missing Data**: Apply imputation techniques (mean/mode for numerical features, "unknown" for categorical features) to address missing values in the Million Song Dataset.
4. **Algorithm Adjustments**: Incorporate diversity constraints and hybrid recommendation approaches to mitigate biases and cold-start issues.

5. **Balance Data Size**: Use stratified sampling or synthetic data generation to balance the dataset and ensure scalable operations.