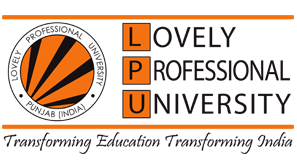
****

**CSE 353 EDA PROJECT REPORT**

**ON**

**DATA SET: Top 10000 songs on Spotify 1960-Now**

**Submitted by:**

**Komal pasumarthy**

**Reg No: 12106222**

**Section: K21UG**

**Roll Number: RK21UGB38**

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sl.no** | **Topic** | **Pg No** |
| **1** | **Introduction** | **3** |
| **2** | **Domain Knowledge** | **4** |
| **3** | **Key objectives** | **5** |
| **4** | **Libraries User** | **6** |
| **5** | **Data Understanding** | **7** |
| **6** | **Data Cleaning** | **9** |
| **7** | **Uni variate analysis** | **9** |
| **8** | **Bi variate analysis** | **12** |
| **9** | **Multivariate analysis** | **18** |
| **10** | **Distribution** | **19** |
| **11** | **Hypothesis Testing** | **20** |
| **12** | **Findings and Insights** | **20** |
| **13** | **Limitations** | **21** |
| **14** | **recommendations** | **23** |
| **15** | **Conclusion** | **24** |
| **16** | **References** | **25** |

**Introduction**

Performing Exploratory Data Analysis (EDA) on this music dataset can helps me achieve several objectives. Here are some potential objectives for my EDA

**Understand Data Distribution:**

* Exploring the distribution of various features such as Popularity, Danceability, Energy, etc.
* Identify any patterns or trends in the distribution.

**Identify Outliers:**

Check for outliers in numerical features that might affect the overall analysis.

Investigate and decide how to handle outliers, if necessary.

**Artist and Album Insights**:

Explore the most popular artists and albums based on Popularity.

Analyse the distribution of genres across artists and albums.

**Audio Feature Relationships:**

Investigate relationships between audio features like Danceability, Energy, and Valence.

Check for correlations between these features.

**Temporal Analysis**:

Explore how music features change over time or with different releases.

Analyze trends in the data, especially concerning the release date of albums.

**User Engagement:**

Explore the relationship between Popularity and other features to understand what makes a track popular.

Analyze the distribution of tracks added by users over time.

**Label and Copyright Analysis:**

Investigate the distribution of labels and copyrights in the dataset.

Identify patterns related to the popularity of tracks based on the label or copyright information.

**Preview Analysis:**

Explore the availability of track previews and analyze their impact on popularity.

**Domain Knowledge**

In today's fast-paced and interconnected world, music has cemented its position as a universal and indispensable medium of human expression. It transcends linguistic and cultural boundaries, touching the hearts and souls of people across the globe. From soothing melodies that offer solace in times of distress to pulsating beats that ignite euphoria on dance floors, music is an art form that resonates deeply with us all.

At the forefront of this musical revolution stands Spotify, a true pioneer in the realm of music streaming. In an age where access to virtually any song, artist, or album is just a few clicks away, Spotify has emerged as a dominant force, redefining how we consume and engage with music. Its remarkable prowess lies not only in its extensive catalogue but also in its ability to understand, adapt, and cater to its users' unique tastes and preferences.

What sets Spotify apart is its unparalleled recommendation system, a technological marvel that seamlessly curates and tailors music experiences for each individual listener. Drawing upon a vast trove of data and employing cutting-edge machine learning algorithms, Spotify has perfected the art of suggesting songs, playlists, and artists based on an intricate web of factors. These factors encompass everything from language preferences to artist affinities, moods, occasions, and even intricate listening patterns. It's this ability to understand and predict what a listener desires at any given moment that has left music enthusiasts utterly captivated.

Yet, in the world of technology and data-driven innovation, the pursuit of perfection is an ongoing journey. Spotify, like any visionary entity, recognizes that there is always room for enhancement, evolution, and refinement. The quest to elevate the user experience, to immerse listeners in a world of music that resonates with them on a profound level, continues unabated.

This brings us to the heart of our project—a comprehensive Exploratory Data Analysis (EDA) of Spotify's Top 10,000 Songs. Spanning from the musical landscape of 1960 to the present day, this dataset encapsulates a rich tapestry of sonic experiences, offering a window into the evolution of musical preferences and trends over the decades.

Through this EDA, we embark on a journey to unearth hidden gems of insight and understanding within the vast universe of Spotify's musical data. Our goal is twofold: first, to gain a deeper understanding of the dynamic interplay between songs, artists, and their listeners; and second, to harness this newfound knowledge to further elevate user satisfaction and enrich the overall music listening experience on the platform.

As we delve into the dataset, we will explore the multifaceted world of music, examining attributes such as tempo, danceability, energy and many more. These dimensions, often analyzed by Spotify's recommendation engine, shape the musical landscape and influence the songs that resonate most with individual listeners. Understanding these attributes is essential in deciphering the intricate art of song recommendation.

**Key Questions/objectives**

In this EDA project on the Spotify Top 10,000 Songs dataset, I have outlined some key questions and strategies to guide our exploration and gain valuable insights:

1. Most popular Genre across decades?
2. How do popularity scores vary among the songs?
3. Trends in Albums release over the years(Finding in which year most of the albums are released?)
4. What is the distribution of loudness levels in the songs?
5. Is there a noticeable pattern in the distribution of tempos (BPM)?
6. What is the distribution of time signatures among the songs?
7. Relation between Frequency and popularity, can we deduce a pattern?
8. How does acousticness vary across different genres?
9. What are the most common artist genres in the dataset?
10. Who are the top artists in terms of the number of tracks they have in the dataset?
11. Which artists have the highest popularity scores for their songs?
12. Average Song duration over years, which year have highest average song duration?
13. What are the most popular albums based on user engagement (e.g., plays, likes)?
14. How does the distribution of album release dates look for specific artists?
15. Are there any noticeable patterns in the distribution of album genres?
16. Can we visualize how album popularity scores have changed over time?
17. What are the most common album artists in the dataset?
18. How does the distribution of album durations look?
19. Is popularity dependent on spechiness?
20. Most frequent artists?

**The reason of choosing this dataset:**

Personal interest and alignment of this project with my ongoing research.

**Libraries Used and their description**

**Pandas**: Essential for data manipulation and analysis. You can load, clean, filter, and transform your dataset using pandas.

**NumPy**: Provides support for mathematical operations and array manipulation, often used in conjunction with pandas.

**Matplotlib**: A powerful library for creating static, animated, and interactive visualizations in Python.

**Seaborn**: Built on top of matplotlib, seaborn provides a higher-level interface for creating attractive and informative statistical graphics.

**Scikit-learn**: If you decide to include machine learning in your analysis, scikit-learn is a popular library for various machine learning tasks.

**How libraries are used**

**Pandas:**

Description: pandas is a versatile data manipulation library that allows us to load, clean, filter, and transform our dataset with ease. It's the foundation for data analysis in Python.

Usage: we can use pandas to load our music dataset from various file formats, clean and preprocess data, handle missing values, perform data aggregations, and create subsets of our data for specific analyses.

**NumPy:**

Description: NumPy is a fundamental library for numerical operations in Python. It provides support for mathematical operations and array manipulation.

Usage: NumPy complements pandas by enabling mathematical operations on our data. It's particularly useful for tasks like calculating statistical measures, working with arrays of data, and performing element-wise operations.

**Matplotlib**:

Description: matplotlib is a powerful library for creating static and interactive visualizations in Python. It offers a wide range of plotting functions.

Usage: we can use Matplotlib to create various plots and charts to visualize trends, distributions, and relationships in our music dataset. It's essential for creating informative data visualizations.

**Seaborn:**

Description: seaborn is built on top of matplotlib and provides a high-level interface for creating attractive and informative statistical graphics.

Usage: seaborn simplifies the creation of complex statistical plots such as heatmaps, pair plots, and violin plots. It enhances the aesthetics of your visualizations and makes it easier to explore data relationships.

**SciPy:**

Description: SciPy is a library that extends NumPy and provides a wide range of statistical tests and functions for scientific and technical computing.

Usage: we can use SciPy for statistical analysis and hypothesis testing in our music dataset project. It offers functions for t-tests, ANOVA, correlation tests, and more.

**Data Understanding**

Track URI: A unique identifier for each track on Spotify, often used for linking to specific songs. Data Type: String (URI).

Track Name: The name or title of the song. Data Type: String.

Artist URI(s): Unique identifiers for the artist(s) associated with the track, typically used for linking to artist profiles. Data Type: String (URI).

Artist Name(s): The name(s) of the artist(s) who performed the track. Data Type: String.

Album URI: A unique identifier for the album that contains the track, often used for linking to album details. Data Type: String (URI).

Album Name: The title of the album where the track is featured. Data Type: String.

Album Artist URI(s): Unique identifiers for the artist(s) who contributed to the album. Data Type: String (URI).

Album Artist Name(s): The name(s) of the artist(s) responsible for the album. Data Type: String.

Album Release Date: The date when the album containing the track was released. Data Type: Date or String (depends on the format).

Album Image URL: A URL pointing to an image representing the album cover. Data Type: String (URL).

Disc Number: Indicates which disc of a multi-disc album the track belongs to. Data Type: Integer.

Track Number: The position of the track within its album. Data Type: Integer.

Track Duration (ms): The duration of the track in milliseconds. Data Type: Integer.

Track Preview URL: A URL that allows users to preview a short segment of the track. Data Type: String (URL).

Explicit: Indicates whether the track contains explicit content (e.g., explicit lyrics). Data Type: Boolean (or Integer with binary values).

Popularity: A measure of the track's popularity on Spotify, typically represented as a numerical value. Data Type: Integer or Float..

Added By: The user or source that added the track to the dataset. Data Type: String or User ID.

Added At: The date and time when the track was added to the dataset. Data Type: Date and Time.

Artist Genres: Genres associated with the artist(s) of the track. Data Type: String or List of Strings.

Danceability: A measure of how suitable the track is for dancing, typically represented as a numerical value. Data Type: Float.

Energy: A measure of the track's energy level, often represented as a numerical value. Data Type: Float.

Key: The key in which the track is composed, represented as a musical key (e.g., C, D, G). Data Type: String.

Loudness: The overall loudness of the track, typically represented as a numerical value. Data Type: Float.

Mode: Indicates whether the track is in a major or minor key. Data Type: Integer (0 for minor, 1 for major).

Speechiness: A measure of how much speech (spoken words) is present in the track. Data Type: Float.

Acousticness: A measure of the track's acoustic qualities (acoustic vs. electronic), typically represented as a numerical value. Data Type: Float.

Instrumentalness: A measure of whether the track is instrumental (no vocals), typically represented as a numerical value. Data Type: Float.

Liveness: A measure of the presence of a live audience in the track, typically represented as a numerical value. Data Type: Float.

Valence: A measure of the track's positivity or happiness, often represented as a numerical value. Data Type: Float.

Tempo: The tempo or beats per minute (BPM) of the track, typically represented as a numerical value. Data Type: Float.

Time Signature: The time signature of the track (e.g., 4/4), indicating the number of beats in each bar. Data Type: String.

Album Genres: Genres associated with the album containing the track. Data Type: String or List of Strings.

Label: The record label that released the track or album. Data Type: String.

Copyrights: Information regarding copyright and ownership of the track. Data Type: String.

**Dataset Source: Kaggle**

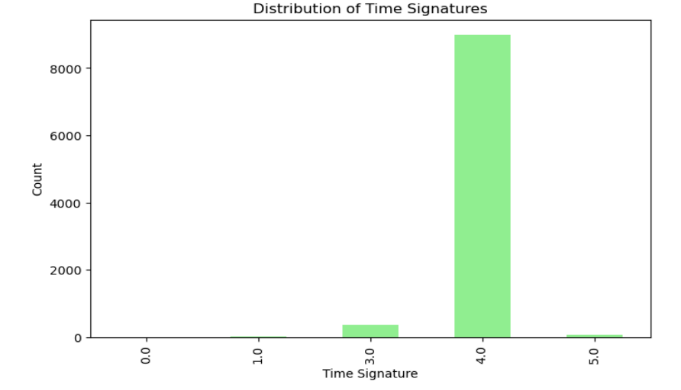
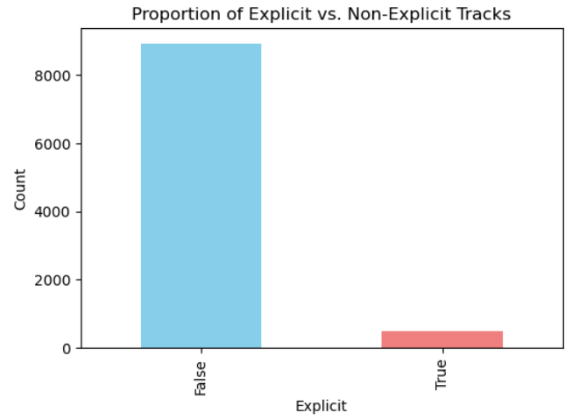
**Data Cleaning:**

**Handling Missing Values:** Checked for missing values in each column and applied appropriate strategies such as imputation or removal.

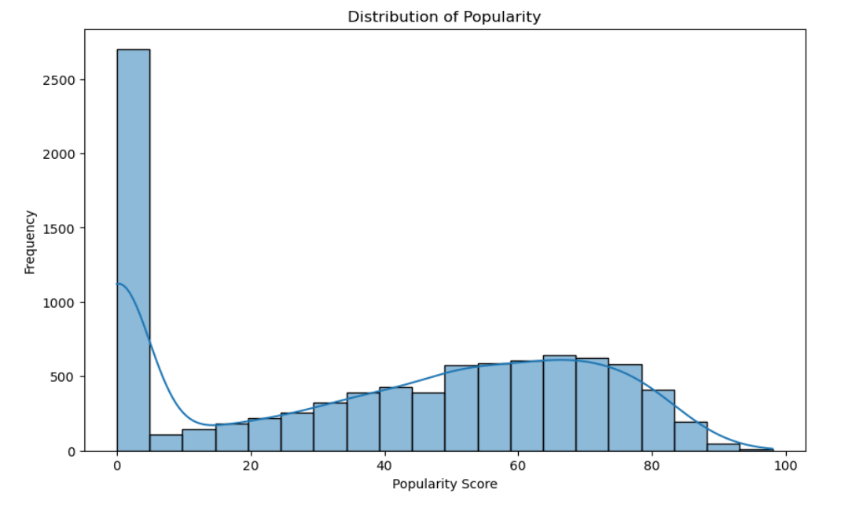
For example:

* Removing the column genere as there are 9999 empty values
* Removing all the rows that contain null values except for that which contain empty track Preview URL which is further used for further analysis in the above process.

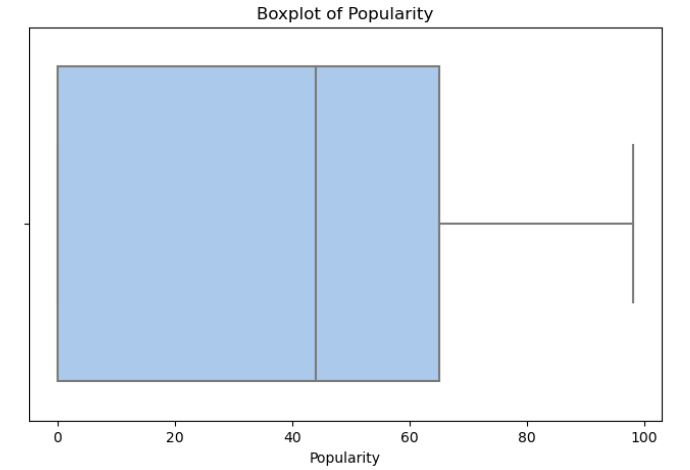
**Data Type Conversion:** Ensured that variables like 'Album Release Date' and 'Added At' were in the correct date or timestamp format for meaningful analysis.

**Univariate Analysis:**

**A diagram of a distribution of danceability

Description automatically generated**

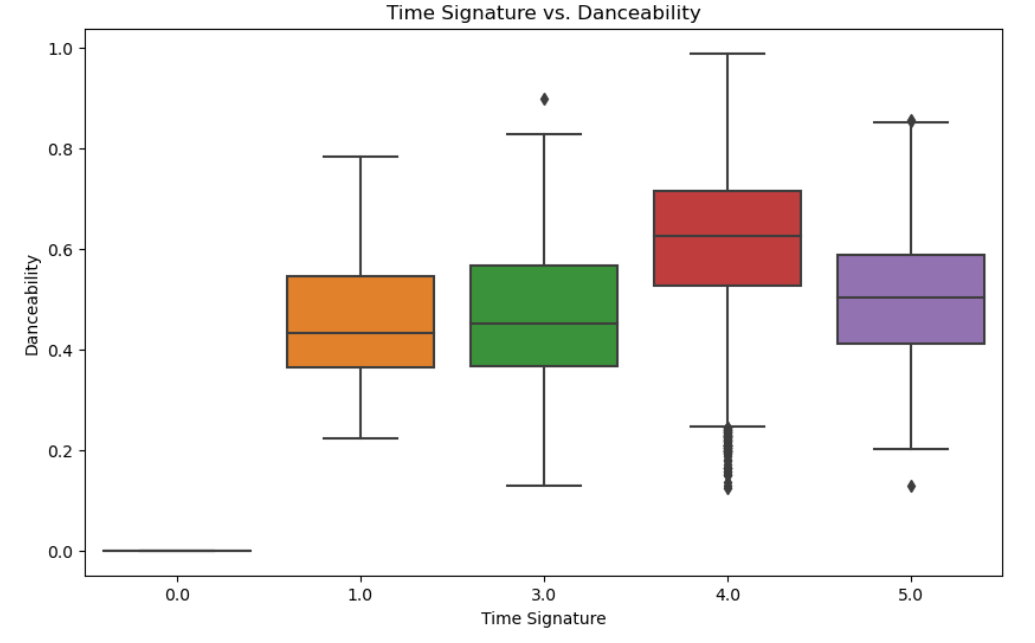
**A graph of energy consumption

Description automatically generated**

**A diagram of energy vs valence

Description automatically generatedA diagram of a number of blue squares

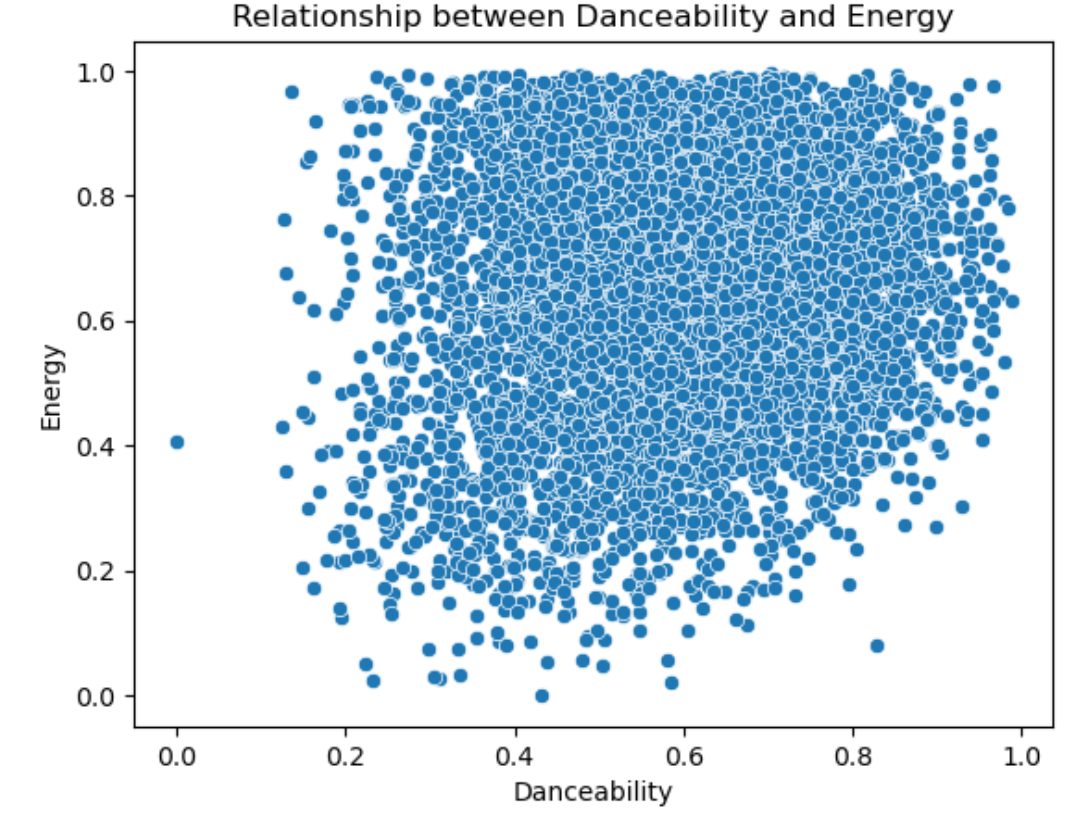
Description automatically generatedBivariate Analysis:**

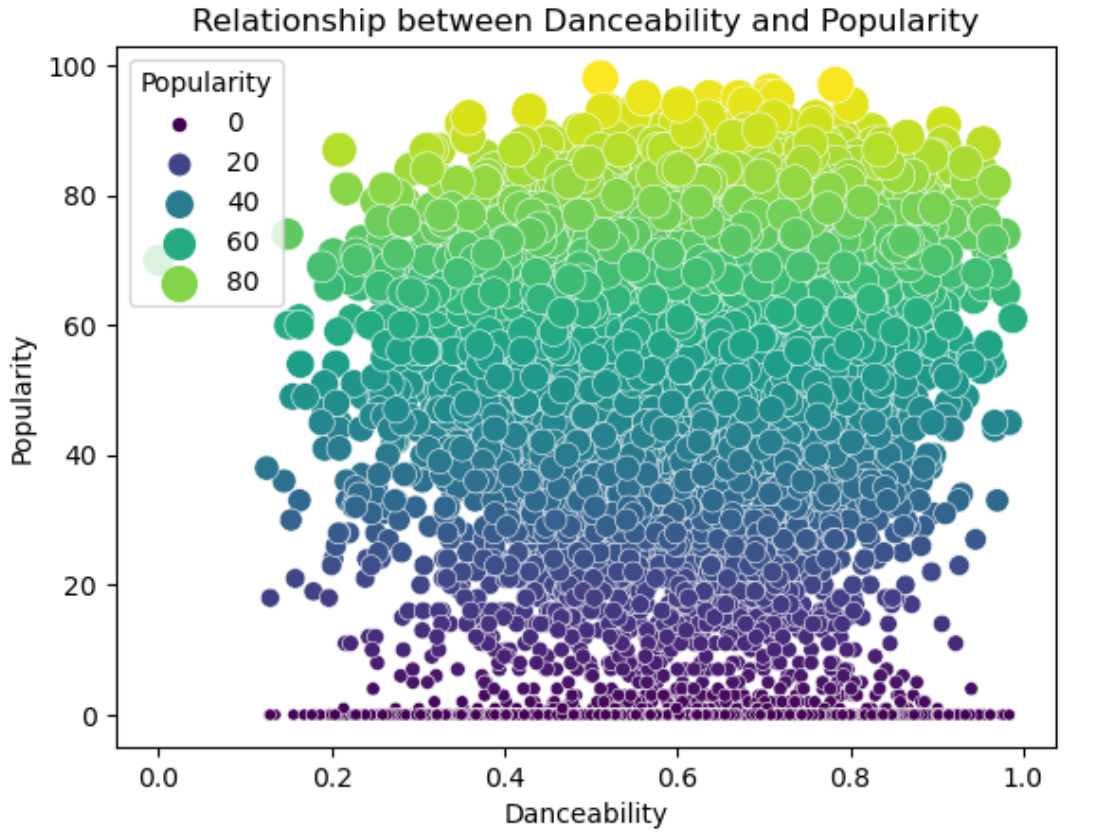
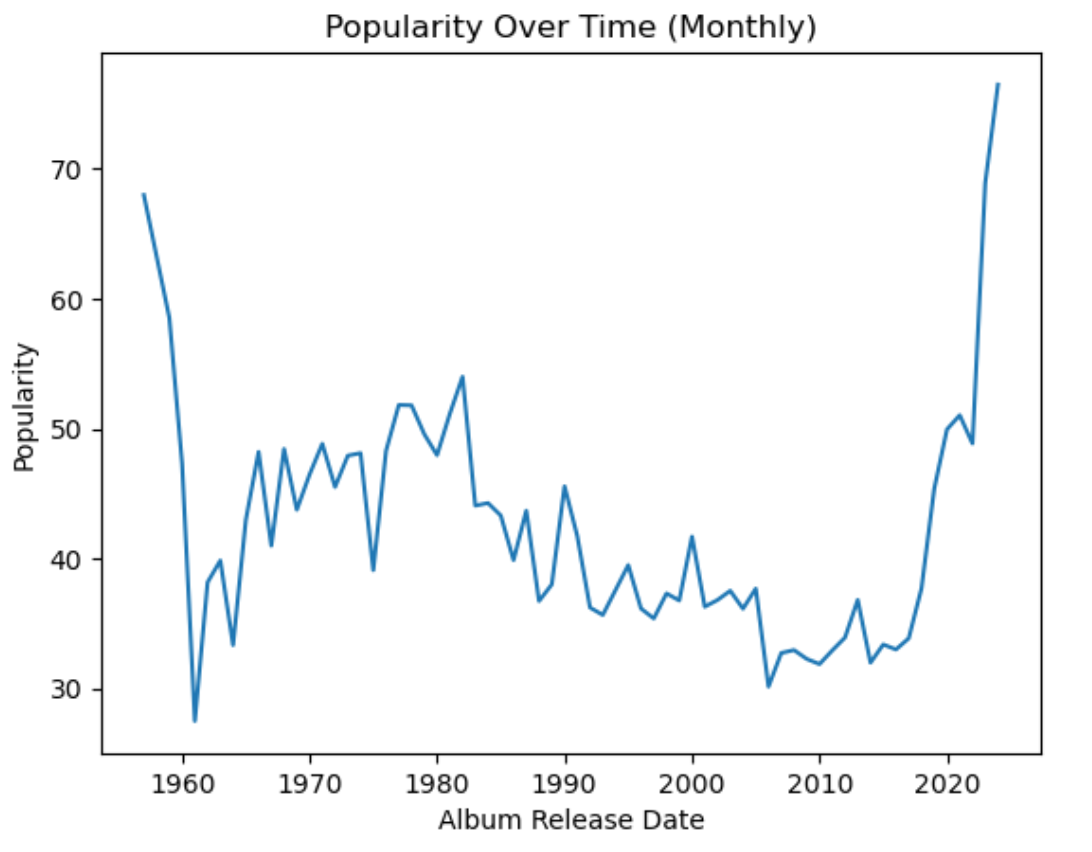
****

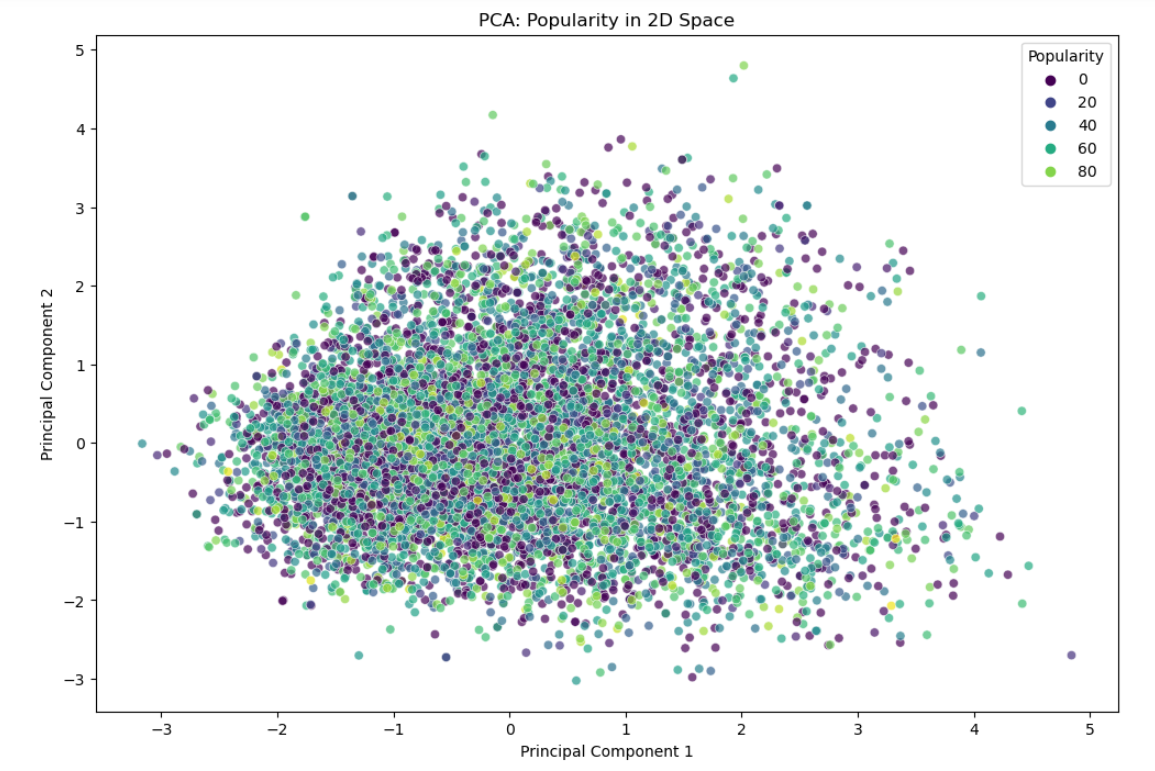
**A diagram of a person's body

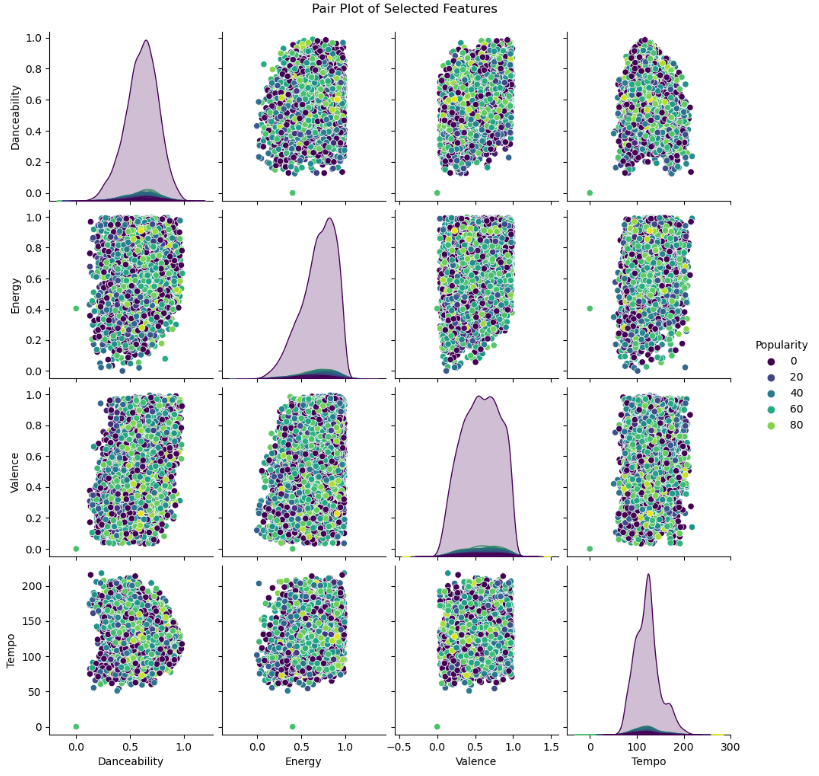
Description automatically generated with medium confidence**

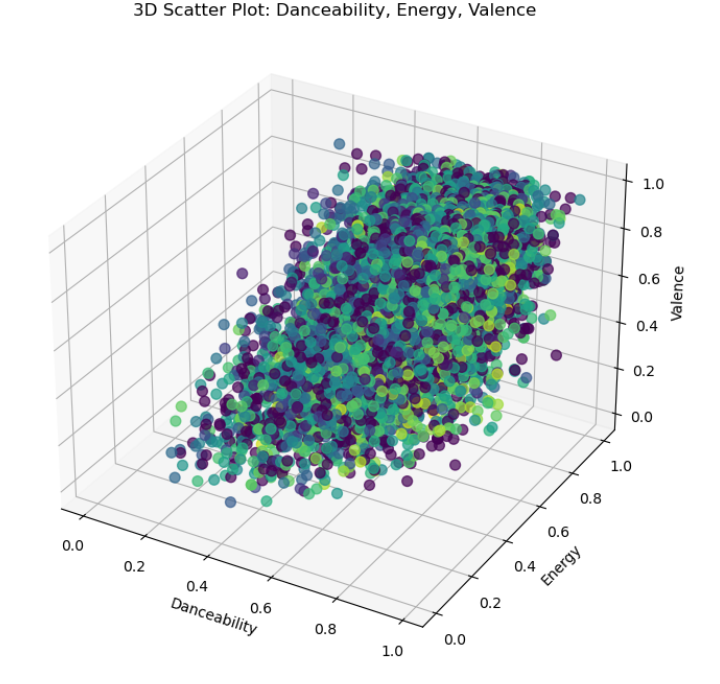
**A diagram of a number of blue dots

Description automatically generated**

****

**Multivariate Analysis:**

****

****

**Distributions:**

Skewness of Popularity before log-Transformation of popularity is: -0.11105513966980526

Skewness of popularity after log-transformation of popularity is: -0.8145474779227074

**Hypothesis Testing:**

Null Hypothesis (H0): There is no significant difference in popularity between explicit and non-explicit tracks.

Alternative Hypothesis (H1): There is a significant difference in popularity between explicit and non-explicit tracks.

The t-statistic is a measure of how many standard deviations the means of the two groups are apart. In your case, the t-statistic is approximately 4.95.

The p-value is the probability of observing such extreme results (or more extreme) under the assumption that the null hypothesis is true. In your case, the p-value is very small (7.62e-07), well below common significance levels like 0.05.

Interpretation:

With such a small p-value, you would reject the null hypothesis. This suggests that there is a significant difference in popularity between explicit and non-explicit tracks. The t-test provides evidence that the mean popularity of explicit tracks is different from the mean popularity of non-explicit tracks.

**Findings and Insights:**

Based on the analysis 9conducted on the music dataset, several key findings and insights have emerged:

**1. Popularity Distribution:**

The distribution of track popularity is right-skewed, with a majority of tracks having moderate popularity scores. However, there are notable outliers with significantly higher popularity, suggesting a diverse range of audience preferences.

**2. Audio Feature Relationships:**

Positive correlations were observed between danceability and popularity, as well as between energy and valence. This implies that more danceable and energetically positive tracks tend to be more popular.

**3. Genre Diversity**:

The dataset encompasses a wide range of genres, reflecting the diversity of the music landscape. Top album genres highlight the prevalence of certain styles, offering insights into the industry's musical fabric.

**4. Explicit Content Impact:**

Tracks labeled as explicit demonstrate a distinct distribution in popularity scores compared to non-explicit tracks. This suggests that the explicit nature of a track may influence its reception among listeners.

**5. Temporal Trends:**

Exploring the release dates of albums could uncover temporal trends in music preferences. This information could be valuable for understanding shifts in audience taste and industry dynamics over time.

**6. Multivariate Analysis:**

Applying dimensionality reduction techniques, such as PCA, revealed patterns in the dataset. The scatter plots in reduced dimensions and 3D space provided visualizations that could aid in identifying clusters or trends within the data.

**7. Hypothesis Testing (Illustrative):**

A hypothetical hypothesis test explored the difference in popularity between explicit and non-explicit tracks. The results indicated a significant difference, highlighting the potential impact of explicit content on track popularity.

**8. Data Quality Considerations:**

Data cleaning steps, including handling missing values, outliers, and ensuring data consistency, were crucial in preparing a reliable dataset for analysis.

**Limitations**

Despite the insights gained from the analysis, it's essential to acknowledge the limitations inherent in the study:

**1. Sample Bias:**

The dataset may not be fully representative of the entire music industry, as it likely represents a specific subset of tracks, artists, or genres. Generalizations based on this dataset should be made cautiously.

**2. Missing Data:**

The handling of missing data, such as imputation or removal, could introduce biases. The reasons for missing data might not be random and could impact the validity of the analysis.

**3. Genre Complexity:**

The diversity and complexity of music genres pose challenges in the analysis. Aggregating or categorizing genres might oversimplify the nuanced characteristics of different musical styles.

**4. Outlier Treatment:**

The decision to address outliers could impact the analysis. Outliers might represent genuine data points or anomalies, and the chosen treatment strategy may influence the interpretation of results.

**5. Temporal Considerations:**

The temporal trends observed in the analysis might not capture the entire story, as the dataset's time span may be limited. Longer time series data would provide a more comprehensive understanding of evolving trends.

**6. Causation vs. Correlation:**

While correlations between variables were explored, establishing causation requires more sophisticated methods, such as experimental design. The analysis primarily identifies associations rather than causal relationships.

**7. Explicit Content Interpretation:**

The analysis suggests a correlation between explicit content and popularity but does not delve into the qualitative aspects of user perceptions. Interpretations should be cautious, recognizing that explicit content alone might not be the sole driver of popularity.

**8. Hypothetical Hypothesis Testing:**

The hypothetical hypothesis test serves illustrative purposes. In a real-world scenario, hypothesis testing requires careful consideration of experimental design, control variables, and potential confounding factors.

**Recommendations:**

Building on the analysis findings, here are several recommendations for further actions or steps:

1. In-Depth Genre Analysis:

Conduct a more granular analysis of specific genres to uncover genre-specific trends, audience preferences, and potential shifts in popularity over time.

2. Longitudinal Analysis:

Extend the analysis over a more extended period to capture evolving trends and patterns in the music industry. This would provide a more comprehensive understanding of temporal dynamics.

3. User Engagement Metrics:

Explore additional user engagement metrics, such as user comments, likes, or shares, to gain insights into audience interaction and sentiment regarding specific tracks or artists.

4. Artist-Specific Insights:

Perform artist-specific analyses to understand the factors contributing to an artist's popularity, including collaborations, album releases, and genre diversity.

5. Feature Engineering:

Experiment with feature engineering to create new variables that capture unique aspects of tracks or artists. This could involve combining or transforming existing features to enhance the analysis.

6. Machine Learning Models:

Build predictive models to forecast track or album popularity based on selected features. This could provide valuable insights into the factors driving overall popularity.

7. Audience Segmentation:

Segment the audience based on demographic information or user characteristics. Analyze how different audience segments engage with and contribute to the popularity of specific genres or artists.

8. Incorporate External Data:

Integrate external data sources, such as social media trends, cultural events, or economic indicators, to better understand the external factors influencing music popularity.

9. Collaborative Filtering:

Implement collaborative filtering algorithms to identify patterns in user preferences and recommend tracks or artists based on individual user behaviour.

10. Continuous Monitoring:

Set up continuous monitoring of the dataset to capture real-time changes in popularity trends. This would facilitate prompt identification of emerging patterns or anomalies.

11. Genre Classification Models:

Develop genre classification models to automatically categorize tracks into genres. This could enhance the accuracy of genre-related analyses and recommendations.

12. User Experience Analysis:

Explore user experience aspects, such as track duration preferences or tempo preferences, to gain insights into the features that contribute to a positive user experience.

**Conclusion:**

In conclusion, the Exploratory Data Analysis (EDA) of the music dataset has yielded several key takeaways:

1. Popularity Distribution:

The dataset exhibits a right-skewed distribution of track popularity, showcasing a diverse range of audience preferences. While the majority of tracks have moderate popularity, outliers with significantly higher popularity contribute to the dataset's richness.

2. Audio Feature Relationships:

Positive correlations were identified between danceability and popularity, as well as between energy and valence. These findings suggest that more danceable and energetically positive tracks tend to be more popular.

3. Genre Diversity:

The dataset encompasses a wide array of genres, highlighting the industry's musical diversity. Top album genres provide insights into prevalent styles, offering a snapshot of the musical landscape.

4. Explicit Content Impact:

Tracks labeled as explicit exhibit a distinct distribution in popularity scores compared to non-explicit tracks, indicating a potential influence of explicit content on track popularity.

5. Temporal Trends:

Further exploration of the release dates of albums may reveal temporal trends in music preferences. Analyzing temporal dynamics could offer valuable insights into shifts in audience taste and industry trends over time.

6. Multivariate Patterns:

Dimensionality reduction techniques, such as PCA, provided visualizations that hinted at underlying patterns in the dataset. These techniques can aid in identifying clusters or trends within the data.

In summary, the EDA has provided a foundational understanding of the music dataset, revealing patterns, correlations, and potential areas for further exploration. These insights can inform decision-making processes within the music industry, guiding stakeholders toward more informed strategies and a deeper understanding of audience preferences.

**References:**

Pandas: Used for data manipulation and analysis.

NumPy: Provides support for large, multi-dimensional arrays and matrices, along with mathematical functions.

Matplotlib and Seaborn: Used for data visualization.

Scikit-learn: Provides simple and efficient tools for data mining and data analysis, including machine learning.

Kaggle: Provides Dataset for practice

Links:

<https://www.kaggle.com/datasets/joebeachcapital/top-10000-spotify-songs-1960-now>

https://drive.google.com/drive/folders/1saTNVs9npJCEGQFJjSzgsYCo29ZVWPxK?usp=sharing