Table of Contents

List of Figures

List of tables

List of Appendix Figures

Chapters

1.Introduction

1.1 Background and Motivation

With the advancements in digital technology, significant progress has been made in the realm of music. Today, platforms like Spotify and YouTube are where music is frequently found being listened to by us. However, more than mere auditory pleasure is provided by music; it influences how feelings are experienced and how behaviors are exhibited. For businesses such as retail stores, this becomes particularly intriguing, where customers' shopping behaviors and purchasing decisions can be influenced by the right kind of music. Thus, the pressing question that emerges is: What kind of songs should be played in stores to ensure customers are made happy and more engaged?

1.2 Problem Statement

Plenty of information is available about the types of music that people enjoy on streaming platforms like Spotify and YouTube. Knowledge of the atmosphere of the songs, whether they have a licence, and whether there is an official video is provided. These platforms can also demonstrate which songs are popular. However, utilizing this data to make decisions can be challenging. Understanding the reasons behind the popularity of various songs is necessary in order to select the best music for in-store playlists. Music that would be enjoyed by consumers, when chosen, is likely to increase their engagement and purchases. This study is an attempt to assist businesses in achieving just that.

1.3 Aims and Objectives

The main goal of this research is to determine what factors influence a song's popularity on streaming platforms and use this information to attract more customers with in-store music.

The stages that will be taken are as follows:

- To learn what aspects of a song contribute to its popularity, examine data from Spotify and YouTube.
- Get the data ready for an in-depth investigation.

- Analyse and visualise data to discover how elements like mood and song popularity are related.
- Discover the similarities and differences between popular songs, particularly as they relate to licencing and other features.
- Use insights to assist businesses in developing playlists that uplift the atmosphere in their establishments and increase clientele interaction.

1.4 Research Questions

The following questions that this study aims to address:

- What connection exists between a song's mood and popularity?
- What characteristics, particularly those related to licencing do popular songs offer?
- Can this knowledge be applied to improve in-store playlists that engage customers?

1.5 Limitations

Every research study comes with its own set of constraints, and the acknowledgment of these is essential for a comprehensive understanding of the outcomes. Some of the primary limitations of this study are outlined as follows:

Mood Labeling: In the original dataset,that gathered from Kaggle (rastelli, 2023) specific labels for genres or moods were not provided. As a result, mood labels had to be derived based on audio features present in the dataset. While songs with high danceability and energy were categorized as 'Energetic' based on this study by and those with low values as 'Relaxed', it is important to acknowledge that this classification might not always align with the listener's perception.

Temporal Relevance: Data, especially in dynamic fields like music, is recognized to be heavily time-dependent. The dataset used for this study was collected on the 7th of February,

2023. The rapidly evolving nature of music preferences is noted, and what is trending today might not necessarily be the case in subsequent months.

Lack of Real-world Testing: The means to validate the findings in actual retail environments were not available. While the research indicates certain song preferences based on platform data, a guarantee of similar customer engagement in a store setting cannot be made. Insights were derived from online music platforms, which may or may not directly translate to offline, in-store experiences.

Popularity Metrics: The study was largely grounded on metrics such as likes, comments, streams, and views from Spotify and YouTube. While these metrics provide valuable insights into online preferences, they might not offer a comprehensive representation of a song's impact on customer behavior in a physical store.

Data Imbalance: The data was found to be imbalanced in terms of song distribution across different moods. Various balancing techniques were attempted, but the results remained consistent with the imbalanced dataset. The decision was made to proceed with the original data to preserve authenticity and real-world preferences. Nonetheless, it is important to recognize that an imbalanced dataset might overshadow nuances in the less represented categories.

Understanding these limitations is emphasized as being crucial for interpreting the research outcomes. While the findings of this study offer significant insights into current music preferences on online platforms, it is understood that real-world implementation would necessitate further validation and iterative testing.

Conclusions Despite Limitations:

Despite the aforementioned limitations, this research has yielded significant insights. It underscores the potential power of music streaming platforms as tools for marketing and customer engagement. The findings of this study serve as a foundational step in understanding listener preferences, allowing businesses to harness this information for crafting ambient environments that resonate with their clientele. Such insights, though derived from a specific temporal snapshot, are invaluable for businesses seeking to leverage music as a tool for enhancing customer experience and engagement. It demonstrates the utility and potential applications of using real-time data analytics in making informed marketing decisions.

1.6 Structure of the Dissertation

This dissertation is organized as follows:

Introduction: The problem is set up and its importance is explained.

Literature Review: A review of what other researchers have discovered about popular music and its effects on consumer behavior is presented.

Methodology: The approach to analyzing the data is explained.

Data Cleaning and Preprocessing: The process of preparing the data for analysis is described.

Exploratory Data Analysis: The initial insights obtained from the data and any interesting patterns that were found are shared.

In-Depth Analysis and Results: The results of the detailed analysis of music track popularity and mood are presented.

Interpretation and Discussion: The results are discussed and compared with previous research. Consideration is also given to how the results might assist businesses in creating more effective in-store playlists.

Legal, Social, Ethical, and Professional Issues: Potential issues related to the legality, social implications, ethical concerns, and professional norms surrounding the use of music and data analytics in the industry are discussed.

Conclusion and Future Work: The research is summarized, its implications are discussed, and suggestions for possible future extensions of the work are made.

Appendices: Any additional information or data needed to understand the research is provided.

References: A list of all the resources and literature referred to in this project is presented.

2.Literature Review

1. Background Music's Impact on Live Streaming Commerce Behaviour:

Critical Perspective: The use of the SOR framework and Broadbent's filter theory provides a rigorous academic underpinning to the study, ensuring that the findings have a theoretical basis. However, the study primarily focuses on background music placement, perhaps missing other audio cues or contextual variables. While it acknowledges the significance of live streamers' explanations, more in-depth research could be beneficial on how music interacts with these explanations. Additionally, the study's focus on tempo in isolation might ignore genres or specific artists that could hold substantial influence. (S.Zhang, 2023)

Building upon this foundation, the gaps identified in the literature are sought to be addressed by the present study. Unlike the aforementioned research, which focused primarily on tempo and the placement of background music in live streaming commerce, the scope in this study is expanded to consider a broader range of audio features and their impact on consumer behavior. Specifically, genres and moods of music tracks, which were under-explored in previous studies, are examined in detail in this project. The analysis in this study, leveraging a comprehensive and up-to-date dataset collected from popular online platforms, offers a nuanced understanding of how various audio features are correlated with different moods and how these, in turn, influence song popularity. Furthermore, the importance of song licensing, which has emerged as a significant factor in the analysis, is recognized in this study and was not covered in prior research. By examining these aspects, the aim of this project is to provide a more comprehensive and actionable framework for businesses and live streamers who wish to optimize their use of background music to enhance customer engagement and sales, without their active involvement.In this manner, the present research not only builds upon existing theoretical frameworks but also extends the empirical analysis, thereby contributing to a richer, more detailed picture of the interplay between background music and consumer behavior in both online and physical retail environments, as revealed by the findings.

2. Analyzing Song Popularity using Spotify Data:

Critical Perspective: The use of multiple machine learning algorithms ensures the comprehensiveness of the study. However, the study's reliance on audio features and past artist data might not account for external factors like marketing campaigns or socio-cultural

events that influence song popularity. The omission of data concerning which specific songs were popular may limit the practical applicability of the findings.This study offers a valuable computational model for song popularity prediction. To augment this, future studies should consider integrating broader external variables to enhance the model's accuracy and applicability. (Singh, 2023)

Building upon these identified gaps, the present project endeavors to explore beyond the audio features and artist data that have been predominantly focused on in previous studies. Recognizing the potential influence of external factors, such as licensing status and mood labels, on song popularity, this project incorporates these dimensions into the analysis. Additionally, this research not only analyses the audio characteristics that are potentially correlated with song popularity, but also investigates how licensing status and mood, derived from audio features, might interact with song popularity. This approach seeks to provide a more holistic and nuanced understanding of song popularity, aiming to address the limitations of past research and offer actionable insights that are grounded in a richer, more complete dataset.

In this manner, the present project is positioned as a substantive extension of prior work, designed to fill critical research gaps and offer a more comprehensive and practical analysis of song popularity using Spotify data.

3. SpotHitPy - Song Hit Prediction using Spotify:

Critical Perspective: While the dataset is impressive in size, the study's challenges reflect the complexity of predicting musical trends. The highlighted accuracy rates, although high, suggest there's a 14% margin of error – a significant portion when dealing with industries worth billions. SpotHitPy's research underlines the potential of ML in the music industry. It's a stepping stone, urging further refinement to tackle the elusive nature of musical trends. (Dimolitsas, 2020)

In light of these identified gaps and challenges, a distinctive approach is adopted in the current project that extends beyond mere hit prediction. The project is designed to delve into various factors and their intricate relationships that may influence song popularity, rather than solely focusing on hit prediction accuracy. Integration of additional dimensions, such as song licensing, mood categorization based on audio features, and the balance of the dataset, is actively pursued in this project. By aiming to provide more contextualized and nuanced

insights into song popularity, the project is positioned to reduce the margin of error that was highlighted in previous studies like SpotHitPy.

Thus, a more robust and actionable understanding of song popularity, considering not only audio features but also licensing status and mood, is sought to be offered by this project. This approach is intended to fill a critical research gap and advance beyond the foundational work established by SpotHitPy.

4. Relationship between Spotify Audio Features and Dance Music:

Critical Perspective: The study offers nuanced insights into dance music, highlighting specific Spotify features that resonate with listeners. However, it might be leaning heavily on Spotify's pre-defined audio features, which could miss out on other potential influential factors. The cluster analysis is commendable but may need more layers of validation across diverse listener groups.This article delves deep into the auditory anatomy of dance music, setting the stage for streaming platforms to refine their algorithms. It's a call for further exploration into diverse music genres and how specific features resonate with their dedicated audiences. (Duman, 2022)

This resource used to identify audio features better in this project.In response to the call for further exploration issued by the study, this project is designed to move beyond the confines of pre-defined audio features. A multidimensional analysis approach is taken in this project, which incorporates not only Spotify's audio features but also additional factors such as song licensing and mood categorizations. By extending the investigation into how licensing status and mood, as deduced from audio features, relate to song popularity, the project seeks to fill the gaps identified in previous research. This approach aims to provide richer, more comprehensive insights and may pave the way for streaming platforms to make more informed decisions in their algorithm refinements and song recommendations.Thus, the project aspires to contribute to a deeper understanding of song popularity and listener engagement through the lens of mood categorizations, thereby responding to the call for further research issued by the aforementioned study.

5. Deep Learning vs. Machine Learning in Music Genre Classification

Critical Perspective:Innovative methodology is employed by (Fernandes, 2023)to investigate the differences between deep learning and machine learning for music genre classification. While this contributes to improving user experience on digital platforms, the study is noted

for its failure to consider consumption patterns or user behavior, focusing primarily on the technical aspects of music categorization.

In recognition of this limitation, comprehensive studies that factor in user behavior are conducted in this project. This research extends beyond technical categorizations to explore the intricate relationship between song attributes, including mood, and user engagement or popularity metrics.

6. Personalized Music Recommendations:

Critical Perspective:A unique perspective on personalized music recommendations based on single-user data on Spotify is offered by (Pipilis, 2023)However, the potential oversight of broader trends and insights from larger user communities is highlighted as a limitation.

This project aims to capitalize on this observed gap by investigating patterns in song popularity based on various attributes.

7. Content Recommendations on Spotify:

Critical Perspective:

(Kutlimuratov, 2023) is focused on data extraction, preparation, and prediction methods for content recommendations on Spotify. However, the absence of user perspective evaluation and the lack of consideration for applicability in different contexts, such as in-store music curation, are marked as significant gaps.

Our study is designed to bridge this gap by evaluating the potential utility and real-world applicability of song attributes, including licensing status and mood labels, in influencing song popularity and listener behavior.

8. Music Emotion Recognition Systems:

Critical Perspective:An innovative approach to acquiring "ground truth" for Music Emotion Recognition systems is introduced by (Gomez-Canon, 2022)However, a failure to connect these findings to practical applications or to consider their impact on marketing or music curation techniques is identified.

This research aims to contribute to this area by investigating how mood labels, as a proxy for emotional content in songs, can be used effectively in various practical applications, including marketing and music curation.

9. Sonic Seasoning: A Perspective from a BBC Podcast

Critical Perspective: An intriguing dimension suggesting that music could influence flavor perception, termed "sonic seasoning," is introduced in a BBC podcast episode. However, this perspective is noted for not having been thoroughly explored in previous studies. (minutes, 2023)

In response to this unexplored avenue, this project aims to delve deeper into the potential broader impacts of music, examining how mood labels and song attributes are related to listener behaviour and song popularity.

10.Predicting Music Popularity Using Social Media and Audio Features:

Critical Perspective: The effectiveness of incorporating both social media and audio features to anticipate the popularity of newly released tracks was explored by (Yee, 2022). Challenges like demographic disparities and time bias were identified as significant limitations, and the accuracy was found to be slightly weakened due to multiple target class predictions.

In this project, a more focused approach is employed to explore the impact of specific song attributes, such as mood and licensing status.

11. Predicting Music Preferences Using Contextual Factors and Machine Learning:

Critical Perspective: Insights into predicting preferred music types across 52 countries using prior listening habits and contextual parameters were provided by (Terroso-Saenz, 2023)

This research aims to extend beyond the prediction of music types to investigate the influence of mood and licensing status on song popularity, thereby offering additional dimensions for optimizing music industry strategies.

12. Spotify Popularity Prediction:

Critical Perspective: The relationship between a song's acoustic features on Spotify and its popularity across multiple platforms was demonstrated by (M.J.Krause, 2021)

In contrast to focusing solely on acoustic features, this study incorporates additional song attributes like mood and licensing status, aiming to generate a more comprehensive understanding of factors influencing song popularity.

13.The Effect of Background Music in Retail Stores:

Critical Perspective: How background music affects consumers' perceptions and behaviours was illustrated by (Sbai, 2019), emphasizing music's strategic use by retailers.

This research extends the inquiry into the digital domain, exploring how song attributes can potentially influence online listener behaviour and song popularity, rather than in physical retail environments.

14.The Effects of Employees' Opportunities to Influence In-store Music on Sales:

Critical Perspective: The positive effects of granting employees more control over in-store music were showcased by (Daunfeldt, 2019)

While (Daunfeldt, 2019) focused on employee control over music in a retail environment, this project aims to understand how song attributes influence the general public's listening behaviour, thereby addressing a broader audience.

15.The Impact of In-store Music on Shopper Behaviour: Journal of Business and Retail Management Research (2010)

Critical Perspective: In-store music was found to have a tangible influence on shopping behaviour in the study from the (MeghaSharma, 2023), with the music genre also playing a pivotal role in shaping consumer spending.

This research aims to extend the understanding of music's influence on behaviour to the online space, focusing on how song attributes, such as mood, correlate with song popularity and listener engagement metrics.

Overall conclusion:

This review establishes the significant and multifaceted role that music plays in influencing consumer behaviour, evident across various contexts, from online streaming platforms to in-store experiences. It highlights the nuanced impact that strategic selection, timing, and placement of music tracks can have on how consumers interact with products and engage with platforms. While individual studies provide valuable insights into specific aspects of music's influence, there is a discernible need for a more integrated and holistic approach. This comprehensive perspective is vital for painting a fuller picture of music's extensive impact across diverse scenarios. As this review has substantiated the profound influence of music in guiding consumer actions and the potential of predictive modelling, it thereby provides a

strong foundational basis for future research. These insights will be instrumental in shaping the methodologies and frameworks for further exploration into the immense potential that music holds in sculpting consumer experiences and decisions.

3.Methodology

The methodology chapter aims to provide a comprehensive overview of the systematic approach adopted for data analysis in this study. This includes the initial data exploration, data cleaning and preprocessing, deep exploratory data analysis, and the subsequent interpretation of the findings. The chapter will show the steps taken to derive insights from the dataset.

Data Description:

- Track: The song's title.
- Artist: The performing artist's name.
- Url_spotify: A hyperlink directing to the artist's Spotify page.
- Album: The name of the album to which the song belongs.
- Album_type: Specifies whether the song was released as a single or is part of an album on Spotify.
- Uri: A unique Spotify link that can be used to locate the song via the API.
- Danceability: A rating between 0.0 and 1.0 that indicates how suitable a song is for dancing, based on various musical elements such as tempo, rhythm stability, and beat strength. (Software, 2020)
- Energy: A scale from 0.0 to 1.0 that estimates the intensity and vibrancy of a track. Higher values suggest a more lively and energetic song.
- Key: The designated musical key of a track, represented using integers that correspond to standard Pitch Class notation. A value of -1 indicates that the key was undetermined.
- Loudness: A measure, in decibels (dB), representing the track's average volume level across its entire length.
- Speechiness: A score indicating the presence of spoken words in a track, with higher values suggesting more prominent speech content.
- Acousticness: A confidence score from 0.0 to 1.0 indicating the likelihood that a track is primarily composed of acoustic elements.
- Instrumentalness: An estimate of the probability that a track lacks vocal content, with values closer to 1.0 indicating a higher likelihood of being instrumental.

- Liveness: A score suggesting the probability that a track was recorded during a live performance, with higher values indicating a greater likelihood of a live recording.
- Valence: A scale from 0.0 to 1.0 indicating the emotional tone of a track, with higher values suggesting a more positive, upbeat sound. (DonSolare, 2018)
- Tempo: The calculated speed of a track, given in beats per minute (BPM). (Ben, 2021)
- Duration_ms: The total length of the track, measured in milliseconds.
- Stream: The count of how many times the song has been streamed on Spotify.
- Url_youtube: The associated YouTube video URL for the song, if available.
- Title: The title of the related YouTube video.
- Channel: The name of the YouTube channel that uploaded the video.
- Views: The total number of views for the video on YouTube.
- Likes: The total number of likes that the video has received on YouTube.
- Comments: The total number of comments posted on the video on YouTube.
- Description: The written description accompanying the video on YouTube.
- Licensed: Signifies if the video contains content that is licensed, meaning that the material was uploaded to a YouTube channel associated with a content partner and has been officially claimed by that partner. (Studiobinder, 2021)
- Official_video: A boolean value indicating whether the linked video is the official music video for the song.

3.1 Initial Data Exploration

The study commenced with a preliminary exploration of the dataset to discern its size, type of variables, and presence of missing values. The dataset contained:
20,718 rows and 28 columns.
Primary data types: float64 and object, with one column of int64.
Missing values in various columns like 'Description', 'Stream', and 'Comments'.
This initial exploration helped establish a foundation and provided context for the preprocessing phase.

3.2 Data Preprocessing

3.2.1 Column Filtering

To make the dataset more manageable and directly relevant to the analysis objectives, several columns ('Unnamed: 0', 'Url_spotify', 'Url_youtube', 'Uri', 'Description') were dropped.

3.2.2 Handling Missing Data

Post column-filtering, the dataset still contained missing values. The strategy adopted was to drop rows with missing values in critical columns. This ensured that the dataset was complete for the deep exploratory phase. The cleaned dataset comprised 19,549 rows and 23 columns.

3.2.3 Feature Engineering

A 'popularity' feature was engineered by aggregating 'Views', 'Stream', 'Likes', and 'Comments'. This composite metric provided a holistic measure of a song's popularity.

Additionally, the 'Duration_ms' column was transformed to minutes for clearer interpretation.

3.3 Deep Exploratory Data Analysis

3.3.1 Feature Scaling and Clustering

For a focused analysis, audio feature columns were chosen and scaled. The clustering algorithm, K-means, was employed to group songs into categories or 'moods'. Using the elbow and silhouette methods, three clusters were found optimal: 'Energetic', 'Relaxed', and 'Neutral'.

3.3.2 Statistical Analysis

Given the distribution imbalance across mood categories, Welch's ANOVA—a variant of ANOVA that's robust against heteroscedasticity—was used. This was followed by the Games-Howell posthoc test to pinpoint specific group differences in popularity.

3.3.3 Visualization Tools

To better interpret the clusters and their characteristics, several visualization methods were employed:

- Violin, box, and bar plots showcased mood popularity.
- Parallel coordinate plots and scatter plot matrices illustrated the mood categories.
- KDE plots highlighted popularity distributions.

3.4 Additional Observations and Techniques

Beyond the primary EDA and clustering, the analysis also incorporated:

- Noting of class imbalances and their implications on analysis conclusions.
- Exploration of correlations, especially around the engineered 'popularity' metric.
- Diverse metrics like song 'type', artist prominence, and mood-key relationships.

3.5 Challenges and Considerations

During the course of this study, several challenges were encountered. Most notably, the dataset lacked mood or genre labels, necessitating the derivation of moods from audio features, a method that might not always reflect a listener's actual sentiment. The data, sourced on 7th February 2023, could become less relevant over time due to the rapidly evolving nature of music trends. Furthermore, there was an absence of practical validation in real-world retail settings; thus, the insights gathered from online platforms might not seamlessly apply to in-store environments. Heavy reliance on popularity metrics, such as those from Spotify and YouTube, posed another limitation, potentially missing out on capturing a song's tangible impact on in-store consumers. Lastly, the dataset's imbalanced song distribution across different moods posed a potential bias, even though efforts were made to mitigate it.

In conclusion, the methodology adopted for this study is iterative and comprehensive, beginning with a broad overview of the dataset and delving deeper into specific areas of interest. Through a combination of data preprocessing techniques, clustering, statistical tests, and diverse visualizations, the methodology facilitated the extraction of nuanced insights about song popularity and moods.

4. Exploratory Data Analysis (Initial Phase)

The Exploratory Data Analysis (EDA) chapter offers a fundamental examination of the dataset before diving into the preprocessing phase. This initial exploration focuses on understanding the dataset's intrinsic features, recognizing anomalies, and outlining potential challenges.

4.1 Dataset Overview with display()

The display(df_without_cleaning) function was executed to get a visual overview of the dataset. This glance provided:

- A snapshot of the first few rows, highlighting columns and general data formats.
- Visual cues about potential missing values or discrepancies in some columns.

4.2 Descriptive Statistics with describe()

The df_without_cleaning.describe() function provided summary statistics for the dataset:

- Central Tendency: Mean values for each column gave a central reference point.
- Dispersion: Standard deviation values indicated the spread of data around the mean.
- Position: Minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values showcased data distribution.

4.3 Data Information with info()

Using the df_without_cleaning.info() function, we gained insight into:

- Total number of entries and columns in the dataset.
- The datatype of each column, helping identify columns with mixed datatypes or those that might need conversion.
- The number of non-null values in each column, offering a preliminary hint about columns with missing data.

4.4 Missing Values Analysis

The df_without_cleaning.isna().sum().sort_values(ascending=False) function was executed to identify and rank columns based on missing data:

- Columns with the most missing values were listed at the top, giving a clear picture of which columns might require imputation or other handling strategies.

- The extent of missing values in some columns can influence decisions on whether to keep, drop, or impute the data.

4.5 Duplicate Rows Inspection

To ensure data integrity and avoid redundancies, duplicate rows were checked using:

```
print(f'Is there any duplicate rows? {df_without_cleaning.duplicated().any()}')
print(f'There are {len(df_without_cleaning)-len(df_without_cleaning.drop_duplicates()):,} duplicate rows')
```

The analysis confirmed whether the dataset contained any duplicate entries and, if present, the number of such entries.

4.6 Data Type Analysis

Data type distribution was explored using print(df_without_cleaning.dtypes.value_counts()):

This allowed for a breakdown of how many columns belong to each data type (e.g., float64, int64, object).

Such an overview is essential for planning data conversions or handling specific data types during preprocessing.

4.7 Columns Overview

The df_without_cleaning.columns command provided a list of all columns present in the dataset. This inventory helped:

- Understand the kind of information encapsulated within the dataset.
- Plan for potential column operations like drops, merges, or renames during the preprocessing phase.

To conclude,the initial EDA phase, as outlined in this chapter, sets the stage for data preprocessing. By harnessing a range of Python commands and functions, a detailed reconnaissance of the dataset was achieved. Such thorough preliminary analysis ensures that the following data preprocessing steps are well-informed, methodical, and aligned with the dataset's characteristics.

5.Data Cleaning and Preprocessing

This chapter delves into the fundamental steps of data cleaning and preprocessing carried out on our music dataset. Ensuring data quality through these procedures is paramount as it underpins the robustness of our subsequent exploratory and statistical analyses.

5.1 Initial Dataset Examination

An extensive dataset, sourced from Kaggle and comprising 20,718 songs (rows) each associated with a set of 28 attributes (columns), served as the starting point of our journey. This dataset contained a mix of numerical and categorical data. Upon initial examination, it was observed that certain columns, including the 'album_type' column which categorizes songs as part of an 'album', 'single', or 'compilation', contained missing values that had the potential to skew the analysis.

5.2 Data Cleaning: Discarding Irrelevance, Handling Missing Data, and Renaming

The data cleaning process was initiated with the removal of irrelevant columns such as 'Unnamed: 0', 'Url_spotify', 'Url_youtube', 'Uri', and 'Description'. These columns were determined to lack analytical value for the study, resulting in a more streamlined set of 23 columns.

Subsequently, missing data was addressed by excluding rows with missing values in key columns, which yielded a cleaned dataset consisting of 19,549 songs.

Additionally, the 'album_type' column was renamed to 'Type' to enhance clarity and interpretability. This newly named 'Type' column indicates whether a song is part of an 'album', 'single', or 'compilation'. Upon analysis, 'album' was identified as the most common type, indicating that the majority of the songs in the dataset are part of an album.

5.3 Feature Engineering: Adding Value to the Dataset

Feature engineering was identified as a crucial phase of preprocessing. A new column, 'popularity', was derived by aggregating 'Views', 'Stream', 'Likes', and 'Comments', providing a comprehensive gauge of song popularity.

Additionally, the 'Duration_ms' column was transformed from milliseconds to 'minutes', thereby enhancing the readability of the data.

5.4 Standardization: Ensuring Equitable Feature Representation

The audio feature columns were selected for standardization using a standard scaler. This step was taken to ensure that all features, irrespective of their original range, carried equal weight in the modeling phase, preventing biases towards features with larger ranges.

5.5 Data Imbalance: Retaining Realism

An imbalance in the distribution of songs across the mood categories was observed. Instead of downsampling or oversampling, the original, imbalanced dataset was retained, as it was believed to more accurately represent real-world scenarios.

5.6 Wrapping Up: The Road to Quality Data

The completion of these steps ensured that the dataset was ready for further examination and testing. Through data cleaning and preprocessing, irrelevant and missing data were eliminated, valuable features were engineered, ranges were standardized, class imbalance was addressed, and interpretability was enhanced. These meticulous steps established the foundation for the subsequent exploratory data analysis, clustering, and statistical testing, ultimately culminating in the final conclusions and recommendations

6. In-Depth Analysis and Results

6.1 Exploring Correlations

In the in-depth analysis, an exploration of potential correlations among song features and the engineered 'popularity' metric was initiated. However, this initial step did not reveal any strong correlations between 'popularity' and the different audio features, suggesting a complex interplay of multiple factors contributing to song popularity.

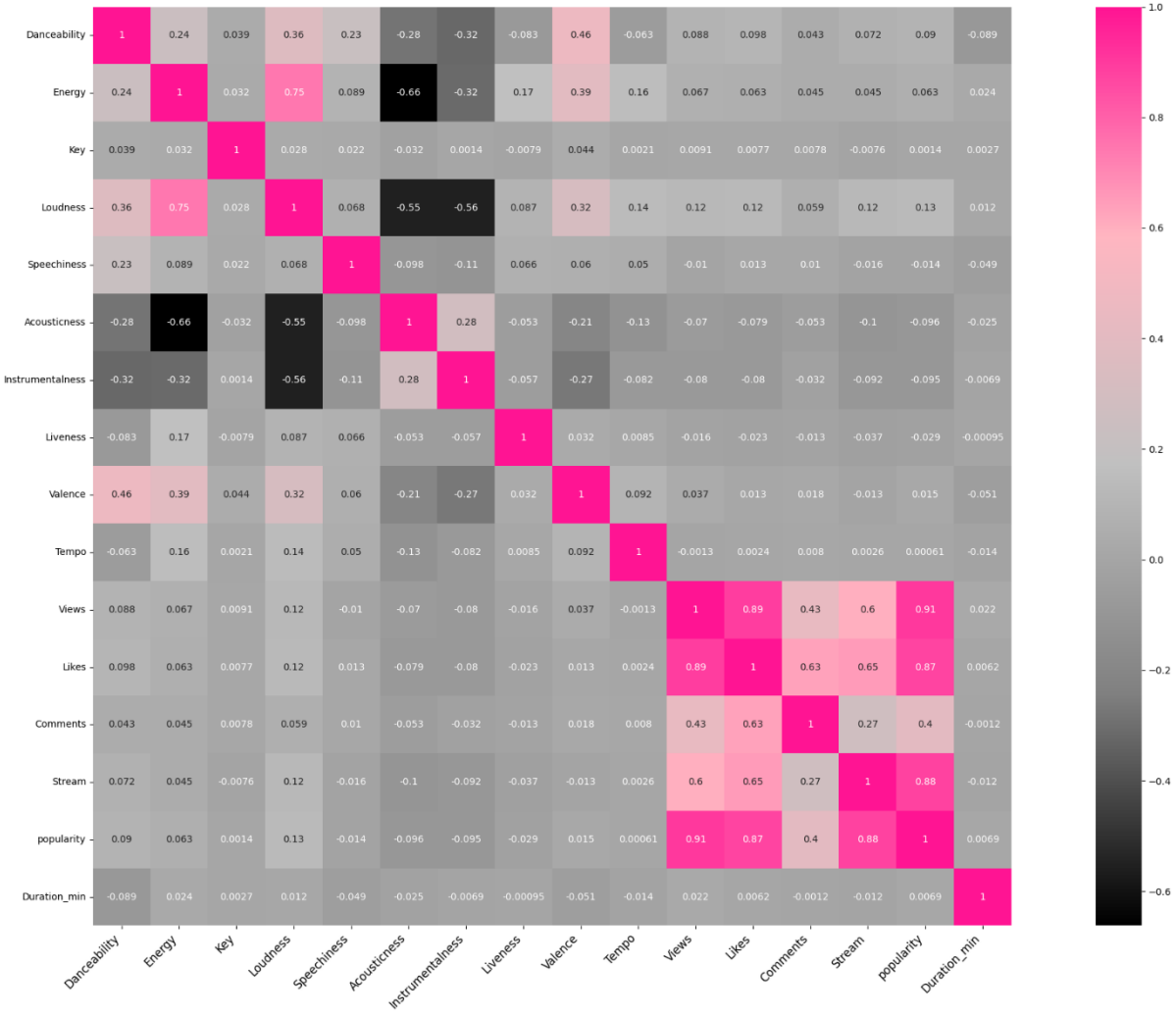


Figure 1 Correlation Heatmap of Cleaned Dataset Features

As is observed from the correlation heatmap, no strong correlation between popularity and audio features is evident. However, the correlation between popularity and popularity features is presented below:

```
correlations = df_cleaned[[ 'popularity','Views', 'Likes', 'Stream', 'Comments']].corr()
# Display the correlation coefficients
```

```
print(correlations['popularity'])
```

output: popularity   1.000000

Views   0.908384

Likes   0.872657

Stream   0.881058

Comments   0.397336

Name: popularity, dtype: float64

```
fig, axes = plt.subplots(2, 4, figsize=(20, 12))
for i, (metric, color) in enumerate(zip(metrics, colors)):
    # Scatter plot
    sns.scatterplot(data=df_cleaned, x='popularity', y=metric, ax=axes[0, i], color=color)
    axes[0, i].set_title(f'Scatter plot: Popularity vs {metric}')

  # Regression plot
  sns.regplot(data=df_cleaned, x='popularity', y=metric, ax=axes[1, i], color=color)
  axes[1, i].set_title(f'Regression plot: Popularity vs {metric}')
```

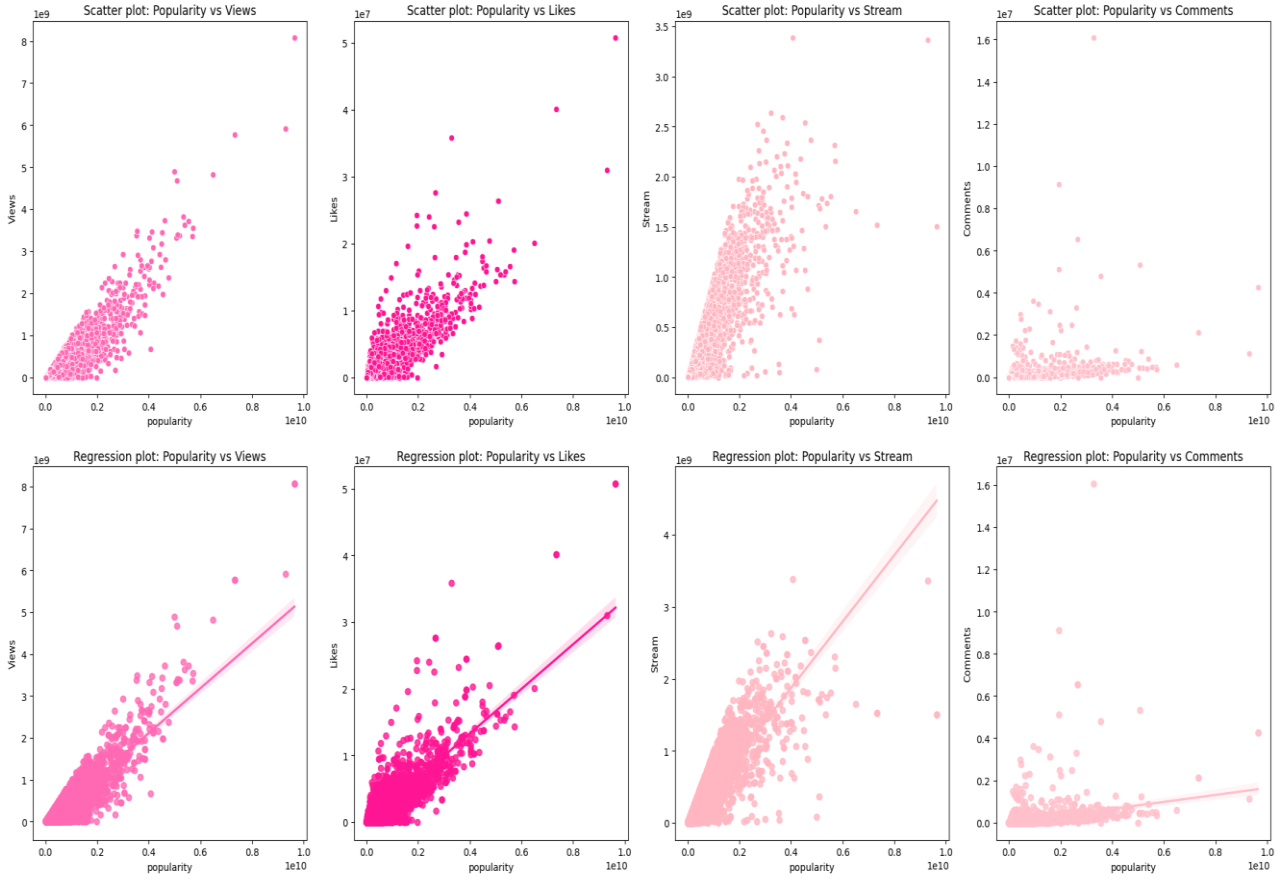


Figure 2 Comparative Analysis of Popularity against Key Metrics: Scatter and Regression Plots

the slope for the Stream metric appears to be steeper than the rest, despite its correlation coefficient being lower than Views.

Understanding this discrepancy is crucial for anyone aiming to leverage these metrics in enhancing song popularity. The slope difference can be attributed to various factors:

1. **Data Scale Variation**: If the Stream metric typically involves smaller values compared to Views or Likes, a slight increase in popularity might reflect as a more noticeable change in streams.
2. **Data Distribution**: The distribution of data points, including the presence of outliers, impacts the slope of the regression line. It's possible that a few highly popular songs with exceptionally high stream counts are influencing the overall trend.
3. **Measurement Units**: Each metric's scale or unit of measurement might differ. If Views and Likes are counted in larger units (like thousands or millions) compared to Streams, it naturally affects the regression slope's steepness.

   In essence, while the correlation values provide a snapshot of the linear relationship's strength between variables, the actual visual representation in regression plots offers a more nuanced picture, influenced by the underlying data's characteristics. For artists or stakeholders aiming

to boost song popularity, it's essential to consider both these aspects. Prioritizing strategies that focus on increasing streams might yield more immediate or pronounced results, even if the long-term goal is to elevate views or likes.

6.2 Cluster Analysis: Grouping Similar Songs

Moving forward, K-means clustering was applied to the audio feature columns with the goal of segmenting the songs into groups that shared similar auditory characteristics. The elbow and silhouette methods were utilized to determine the optimal number of clusters, which pointed to a two-cluster solution. However, to capture a greater variety of musical styles and ensure more nuanced interpretations, a three-cluster solution was opted for. Each cluster, or 'mood', was labeled as 'Energetic', 'Relaxed', or 'Neutral', based on the average audio feature characteristics of the songs in that group.

```
# feature selection for clustering
features = ['Danceability', 'Energy', 'Key', 'Loudness', 'Speechiness', 'Acousticness',
'Instrumentalness', 'Liveness', 'Valence', 'Tempo']
X = df_cleaned[features]
# Scale the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
# Store the SSE (Sum of Squared Errors) and silhouette scores
sse = []
silhouette_scores = []
# Iterate from 2 to 10 to find the optimal number of clusters
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X_scaled)
    sse.append(kmeans.inertia_)
    silhouette_scores.append(silhouette_score(X_scaled, kmeans.labels_))
# Plot the elbow method and silhouette score to find the optimal number of clusters
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(15,5))
ax1.plot(range(2, 11), sse, marker='o', color='pink', markerfacecolor='hotpink')
ax1.set_xlabel('Number of clusters')
ax1.set_ylabel('SSE')
```

```
ax1.set_title('Elbow Method')
ax2.plot(range(2, 11), silhouette_scores, marker='o', color='pink', markerfacecolor='hotpink')
ax2.set_xlabel('Number of clusters')
ax2.set_ylabel('Silhouette Score')
ax2.set_title('Silhouette Score Method')
plt.show()
```

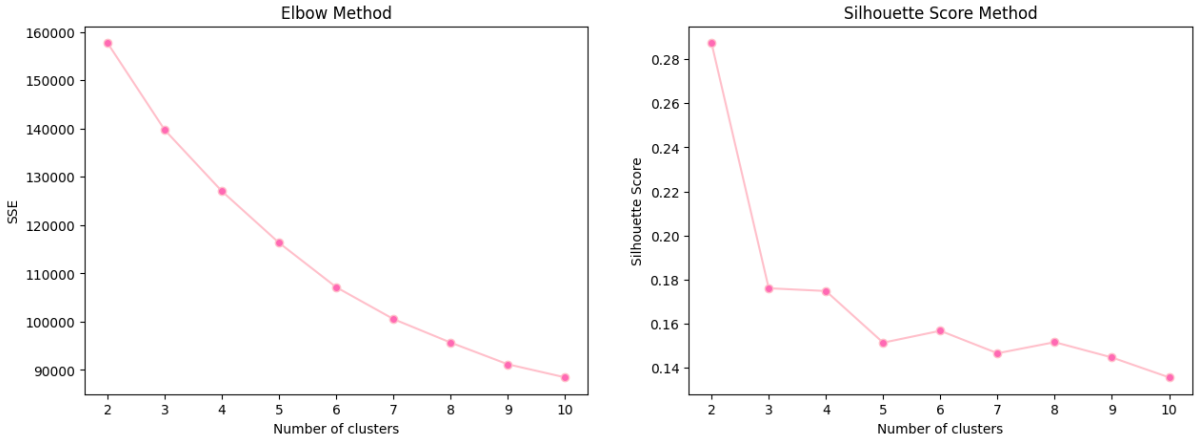


Figure 3 Evaluation of Optimal Cluster Count: Elbow and Silhouette Methods

```
# Apply K-Means clustering
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans.fit(X_scaled)

# Added the cluster labels to the dataframe
df_cleaned['mood'] = kmeans.labels_

# Calculation of mean values for each feature per cluster
cluster_means = df_cleaned.groupby('mood')[features].mean()
```

The purpose of this calculation is to understand the average characteristics of each mood category based on the selected features. A summary is provided of how each mood differs, on average, across various features. This kind of analysis is especially useful when data has been clustered into categories (like moods) and there is a desire to get a sense of the "typical" characteristics of each cluster.

```
colors = {0: 'grey', 1: 'deeppink', 2: 'lightpink'}
n_features = len(features)
n_cols = 5  # Set the number of columns
n_rows = (n_features + n_cols - 1) // n_cols  # Calculate the number of rows
fig, axes = plt.subplots(nrows=n_rows, ncols=n_cols, figsize=(20, 10))
# Flatten the axes to make it easier to iterate
axes_flat = axes.flatten()
# For each feature, create a subplot
for i, feature in enumerate(features):
    ax = axes_flat[i]

    for mood in cluster_means.index:
        ax.bar(str(mood), cluster_means.loc[mood, feature], color=colors[mood])

    ax.set_title(f'Average {feature} in each Mood')
    ax.set_xlabel('Mood')
    ax.set_ylabel(f'Average {feature}')

for i in range(len(features), len(axes_flat)):
    fig.delaxes(axes_flat[i])

plt.tight_layout()
plt.savefig('Moods.png')  # Saves the whole figure
plt.show()
```

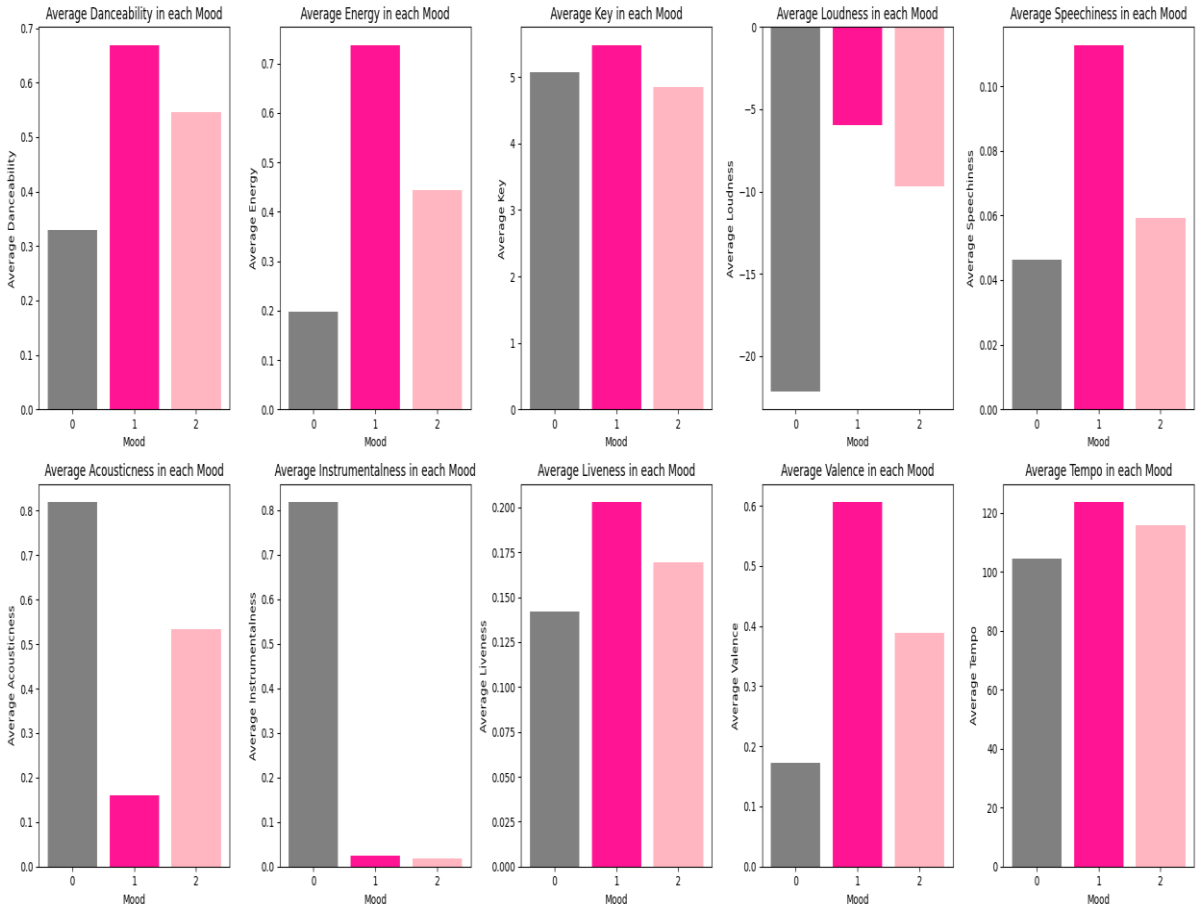


Figure 4 "Feature-wise Average Comparison Across Moods

In an endeavor to understand the nuances of mood categorizations, an exploration was conducted to examine how various audio features correlate with the mood labels: 'Relaxed', 'Neutral', and 'Energetic'. To depict this, a parallel coordinates plot was leveraged, which visually emphasizes the relationship and distinguishes patterns across multiple dimensions simultaneously.

1. **Danceability:** 'Energetic' songs exhibit a higher danceability quotient, as one might expect, while 'Relaxed' tracks tend to be less danceable.
2. **Energy:** Not surprisingly, 'Energetic' songs have a surge in energy, contrasting sharply with the low-energy profile of 'Relaxed' tracks.
3. **Loudness:** Both 'Energetic' and 'Neutral' songs have higher loudness levels, with 'Relaxed' songs trailing behind, indicating a softer, more subdued auditory experience.
4. **Speechiness:** There isn’t a distinct pattern observable for speechiness, suggesting that spoken words or rap elements don't have a strong differentiation across moods.
5. **Acousticness:** 'Relaxed' songs seem to have a higher acoustic element, implying a preference for natural and organic sounds in this category.

6. **Instrumentalness:** While all moods have tracks with varying degrees of instrumental elements, 'Relaxed' songs stand out with a prominent peak, hinting at the prevalence of tracks without vocals.
7. **Liveness:** No striking differentiation across moods, indicating that the presence of live audience sounds isn’t a strong mood differentiator.
8. **Valence:** A measure of musical positiveness, where 'Energetic' tracks unsurprisingly lead the pack with high valence. 'Relaxed' songs, on the other hand, span a wide spectrum, suggesting diverse emotional content.
9. **Tempo:** The pace of 'Energetic' songs is notably higher, aligning with the inherent nature of such tracks, while 'Relaxed' tracks tend to have a slower tempo, reinforcing their calm ambiance.

Based on these visualizations, the labels were chosen. For instance, mood number 1, which is high in average danceability and energy, was named as 'Energetic' mood songs.

```
mood_labels = {0: 'Relaxed', 1: 'Energetic', 2: 'Neutral'}
# Applied the mapping to create a new column with mood labels
df_cleaned['mood_label'] = df_cleaned['mood'].map(mood_labels)
# Create the parallel coordinates plot
plt.figure(figsize=(12, 6))
parallel_coordinates(mood_data_norm, 'mood_label', color=[colors[i] for i in
mood_data_norm['mood_label'].unique()])
# Set plot title and labels
plt.title('Relationship Between Mood Labels and Mood-Based Features')
plt.xlabel('Mood-Based Features')
plt.ylabel('Normalized Values')
# Show the plot
plt.show()
```

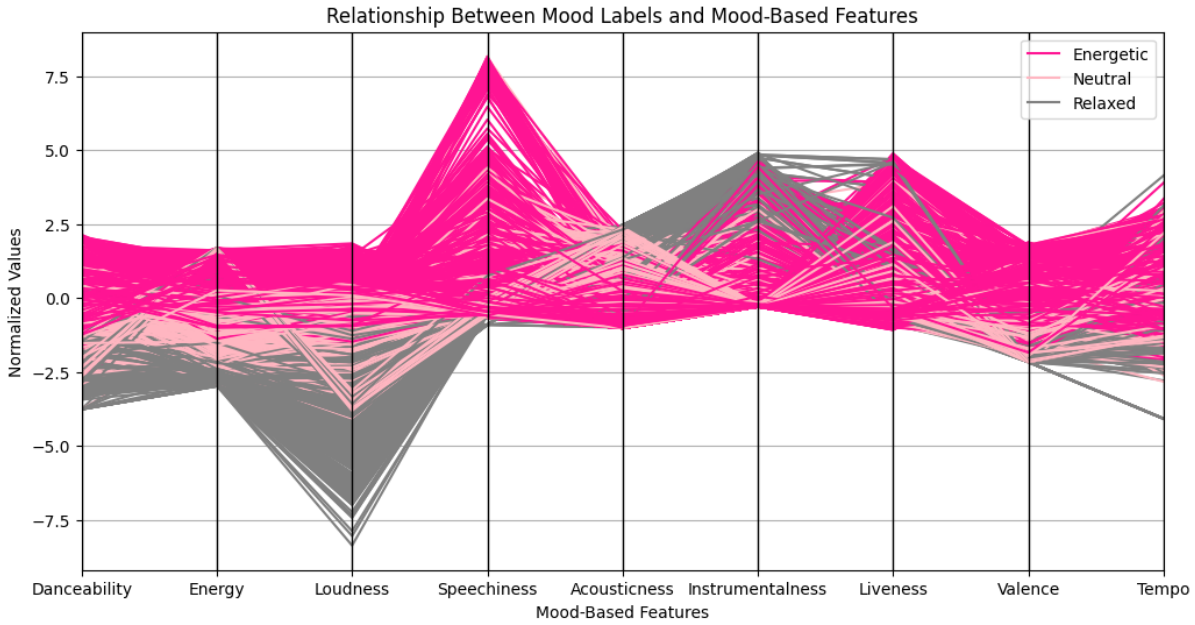


Figure 5 Parallel Coordinates Visualization of Normalized Feature Distribution Across Mood Labels

But again, a strong correlation between the mood column and audio features could not be observed.

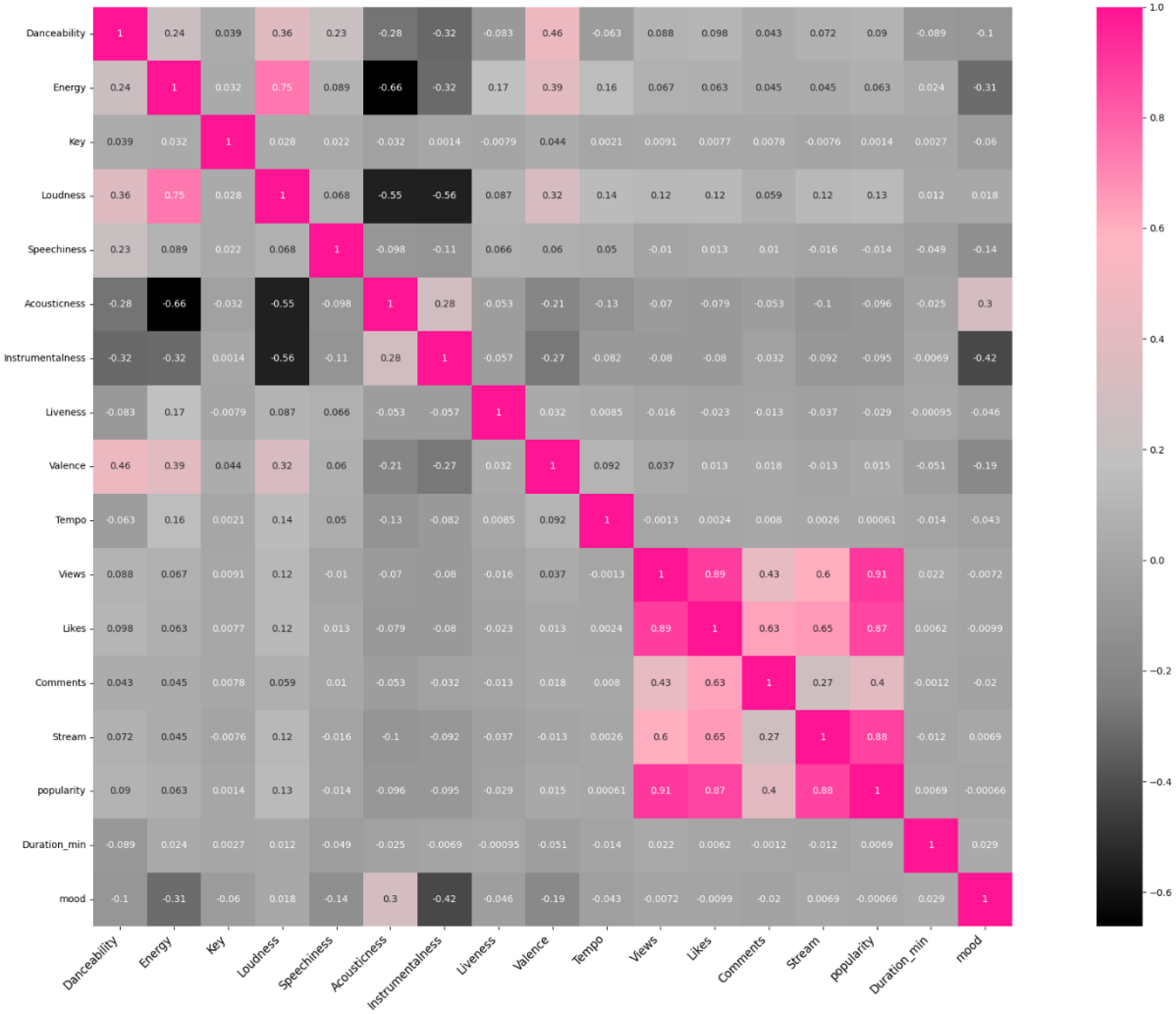


Figure 6 Correlation Heatmap of Dataset Features with mood feature

In the next two steps, it is mentioned why, despite the lack of strong correlation between mood, popularity, and audio features, there are findings that contradict this lack of correlation.

6.3 Mood Popularity: Deciphering Trends

Our visualizations brought to light a fascinating trend: the dominance of 'Energetic' songs. Not only were they more numerous, but they also outscored 'Neutral' and 'Relaxed' songs in terms of popularity.

```
import matplotlib.pyplot as plt
import seaborn as sns

# Create a 3x1 grid of subplots
fig, axes = plt.subplots(nrows=3, ncols=1, figsize=(10, 15))

# Plot the boxplot on the first subplot
sns.boxplot(x="mood_label", y="popularity", data=df_cleaned, palette=["deeppink",
"lightpink", "grey"], ax=axes[0])
axes[0].set_title("Boxplot of Popularity by Mood")

# Plot the violin plot on the second subplot
sns.violinplot(x="mood_label", y="popularity", data=df_cleaned, palette=["deeppink",
"lightpink", "grey"], ax=axes[1])
axes[1].set_title("Violin Plot of Popularity by Mood")

# Plot the bar plot on the third subplot
sns.barplot(x="mood_label", y="popularity", data=df_cleaned, palette=["deeppink",
"lightpink", "grey"], ax=axes[2])
axes[2].set_title("Barplot of Popularity by Mood")

# Adjust the layout
plt.tight_layout()
plt.show()
```

From the visualization below, it is concluded that Energetic songs tend to be more popular compared to other labels. However, as was observed earlier, there isn't a strong correlation between them. Therefore, it was decided to check if there is an imbalance in the data or not. This is because if a larger number of Energetic songs are observed compared to other moods, this might be the reason for noticing this difference between moods.

Here the count of songs for each mood is observed:

```
df_cleaned['mood_label'].value_counts()
```

Energetic 13435 Neutral 5302 Relaxed 812 Name: mood_label, dtype: int64

As we see there is a significant imbalance in dataset.

In the preliminary exploration through heatmaps and bar charts, some interesting patterns and relationships were observed, leading to a suspicion of a potential imbalance in the data. This suspicion was confirmed by the subsequent count. Recognizing such an imbalance is identified as an essential part of the exploratory data analysis, as it informs the subsequent steps that will be taken in data preprocessing and model training.

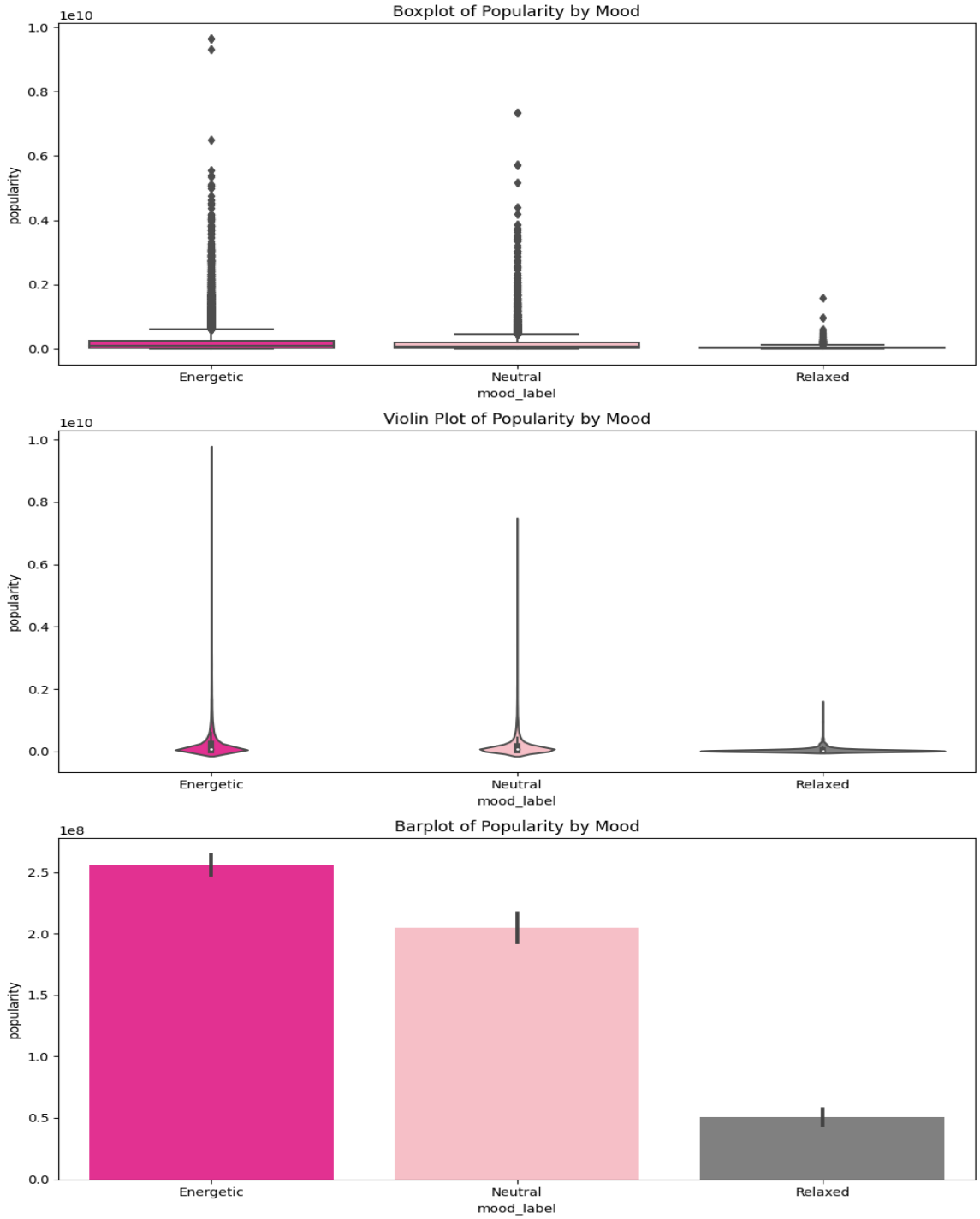


Figure 7 Comparative Distribution of Song Popularity Across Mood Labels: Boxplot, Violin Plot, and Bar Plot

6.4 Addressing Data Imbalance: Mood Distribution and its Impact on Analysis

The density of the label was first checked:

```
plt.figure(figsize=(12, 6))
for label in df_cleaned['mood_label'].unique():
    sns.kdeplot(df_cleaned[df_cleaned['mood_label'] == label]['popularity'], label=label, color=colors[label])

plt.legend()
plt.show()
```

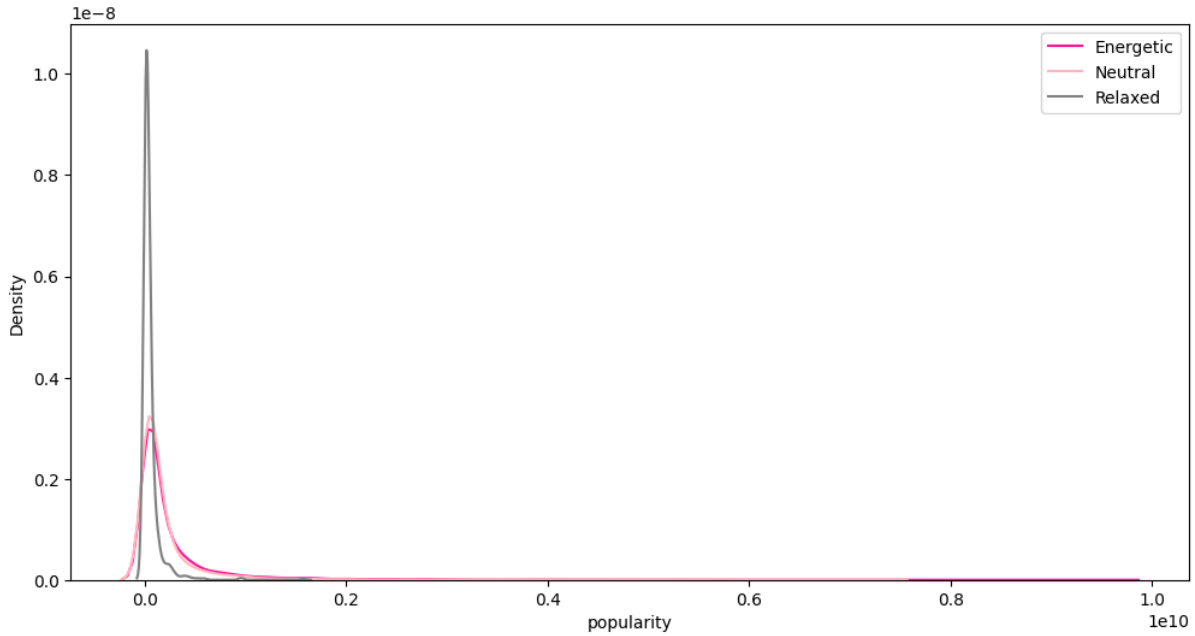


Figure 8 KDE Analysis of Song Popularity Across Different Moods

In our analysis, the Kernel Density Estimation (KDE) plot illustrated distinct patterns among different moods. 'Energetic' and 'Neutral' songs displayed a broad range of popularity ratings, indicating a possible overrepresentation in the dataset. On the other hand, the 'Relaxed' category had a tighter, more peaked distribution, denoting a consistent yet lower popularity.

To address the potential imbalance, two central data balancing techniques were employed:

downsampling and oversampling, to ascertain if the patterns held across different balancing approaches.

Downsampling:

- **Advantages**: It retains the integrity of original data points, ensuring the dataset's genuine characteristics are preserved. It's also computationally more straightforward.
- **Limitations**: There's an inherent reduction of data, which can be significant for underrepresented categories like 'Relaxed'.

Oversampling:

- **Advantages**: It ensures all original data is included.
- **Limitations**: By duplicating existing entries, certain patterns might get unduly emphasized, leading to potential misinterpretations during the analysis phase. Furthermore, with oversampling, the dataset is populated with numerous replicated data points, which, in essence, can be considered "artificial."

Surprisingly, the primary insights, namely the dominance of the 'Energetic' label, remained consistent across both methodologies, reinforcing our initial observations.

Given these findings, the decision was made to continue with the original imbalanced dataset due to the following reasons:

1. Authenticity of Data: The natural distribution might reflect genuine market dynamics, and artificially balancing it could obscure such insights.
2. Preservation of Information: Although downsampling reduces the size of the dataset, its consistency across both methods confirmed that the original dataset offers valuable insights.
3. Real-world Preference: The intrinsic inclination towards 'Energetic' songs is valuable and could be masked upon artificially balancing the data.

While balanced data is typically crucial for unbiased model training, understanding the nuances of the imbalance is deemed to be equally essential. The decisions were deeply rooted in preserving the originality of the data and extracting credible insights.

Given our project's focus on data visualization and analysis, the choice between downsampling (losing data) and oversampling (potentially introducing artificial patterns) became pivotal. The KDE plot insights, coupled with the intent to ensure genuine data representation, made downsampling the preferable choice. The inherent risk of introducing artificial data through oversampling wasn't ideal for our visualization objectives, where each data point's authenticity is paramount.

As for **SMOTE**, while it's an often-utilized method for addressing imbalances, especially in modelling, I encountered problems running it due to the presence of categorical data in our dataset. Its methodology of generating synthetic samples doesn't gel well with non-numerical data, presenting hurdles in its application. Beyond this operational challenge, introducing

synthetic data through SMOTE could have led to misleading patterns in a visualization-centric endeavour.

While ensuring balanced data is typically vital for model training to prevent bias, it is equally important to understand the nature and implications of the imbalance. In this instance, our explorative steps and subsequent decisions were rooted in maintaining the integrity and authenticity of the data while gleaning reliable insights.

```
# Find the number of samples in the smallest group
n_samples = min(len(df_energetic), len(df_relaxed), len(df_neutral))
# Downsample all groups to match the smallest group
df_energetic_downsampled = resample(df_energetic, replace=False, n_samples=n_samples, random_state=123)
df_relaxed_downsampled = resample(df_relaxed, replace=False, n_samples=n_samples, random_state=123)
df_neutral_downsampled = resample(df_neutral, replace=False, n_samples=n_samples, random_state=123)
# Combine all downsampled groups back into a single dataframe
df_balanced = pd.concat([df_energetic_downsampled, df_relaxed_downsampled, df_neutral_downsampled])
# Display new class counts
print(df_balanced.mood_label.value_counts())
```

output: Energetic 812

Relaxed 812

Neutral 812

Name: mood_label, dtype: int64

```
# Violin plot of popularity against mood labels
plt.figure(figsize=(10, 6))
sns.violinplot(x='mood_label', y='popularity', data=df_balanced)
plt.title('Violin Plot of Popularity against Mood Labels')
plt.show()
```

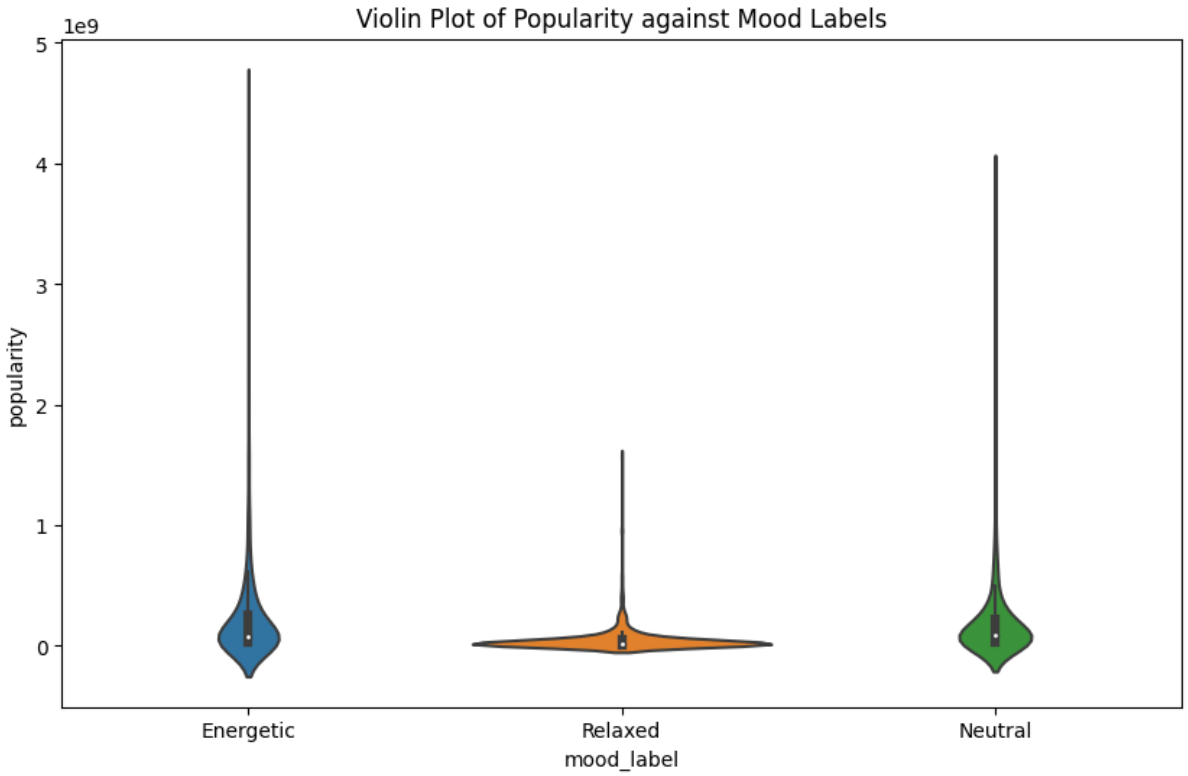


Figure 9 - Violin Plot of Popularity vs. Mood Labels on Downsampled Balanced Data

```
# Bar plot of average popularity against mood labels
plt.figure(figsize=(10, 6))
sns.barplot(x='mood_label', y='popularity', data=df_balanced, ci=None)
plt.title('Bar Plot of Average Popularity against Mood Labels')
plt.show()
```

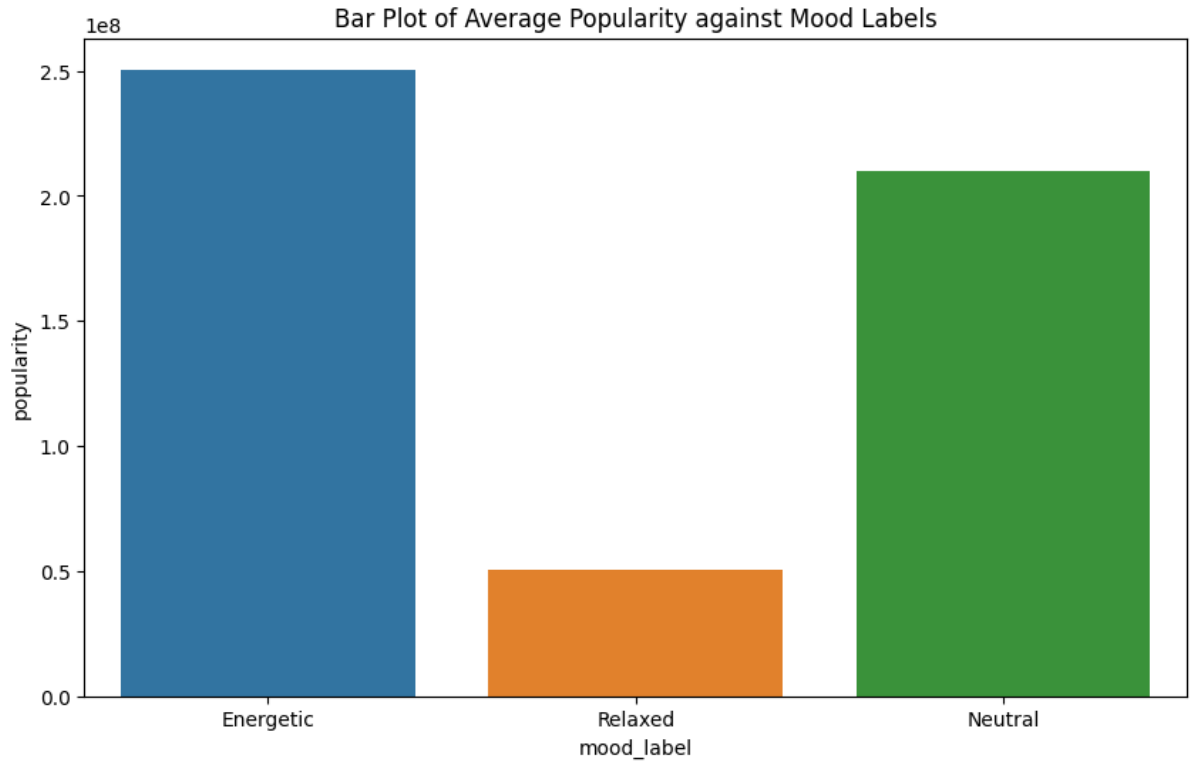


Figure 10 - Bar Plot of Average Popularity vs. Mood Labels on Downsampled Balanced Data

```
from imblearn.over_sampling import RandomOverSampler
# Separate features and target variable
X = df_cleaned.drop('mood_label', axis=1)
y = df_cleaned['mood_label']
ros = RandomOverSampler(random_state=123)
X_ros, y_ros = ros.fit_resample(X, y)
# Combining the features and target variable into a single dataframe after oversampling
df_balanced_ros = pd.concat([pd.DataFrame(X_ros), pd.DataFrame(y_ros)], axis=1)
# Display new class counts
print(df_balanced_ros.mood_label.value_counts())
```

output: Energetic 13435

Neutral 13435

Relaxed 13435

Name: mood_label, dtype: int64

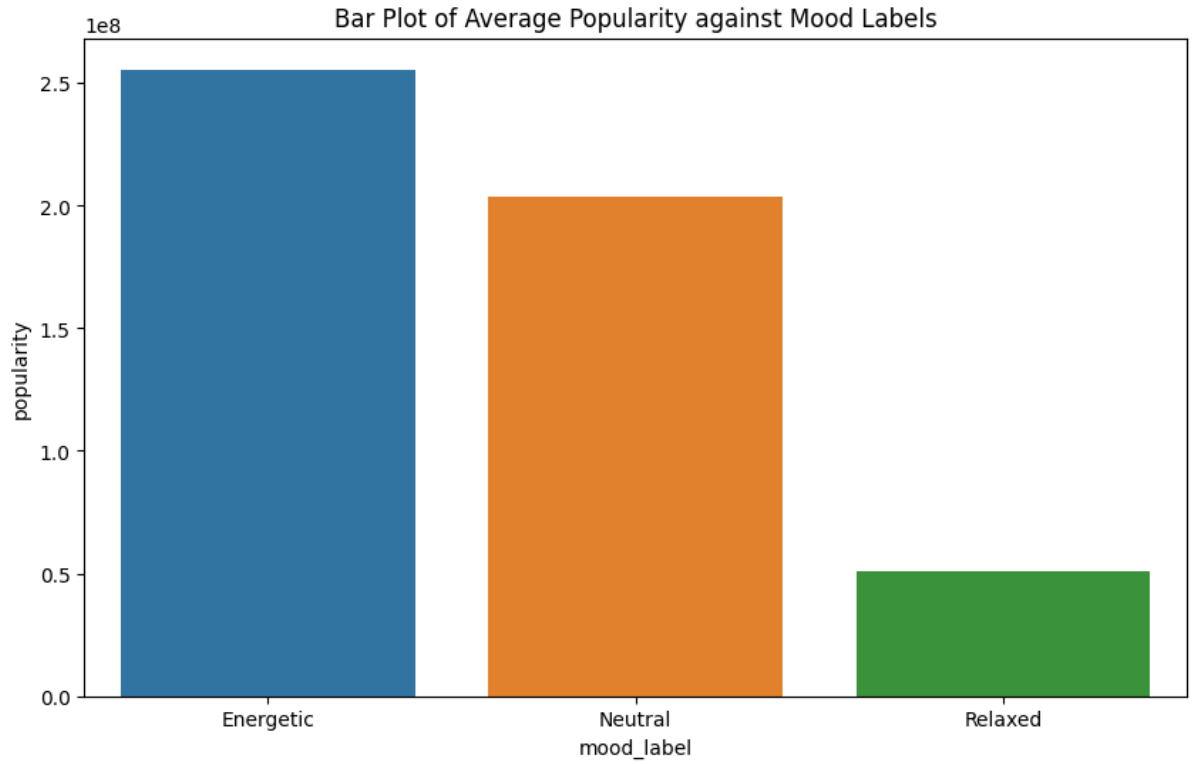


Figure 11 - Class Distribution after Random Oversampling Barplot

It was observed that even in balanced data, the 'Energetic' label is more popular. Therefore, the analysis was continued with the imbalanced data, as it is considered more real and accurate.

6.5 Verification and Statistical Analysis

To statistically verify the visual observations made, a Welch's ANOVA test was executed, followed by a Games-Howell posthoc test. The reason for choosing these techniques was to address the imbalances found in the dataset. The results corroborated the initial finding: 'Energetic' songs enjoyed significantly more popularity than both 'Neutral' and 'Relaxed' songs, and 'Neutral' songs outperformed 'Relaxed' songs.

To evaluate the differences in popularity between the different mood labels, statistical tests were employed. Specifically, Welch's ANOVA and the Games-Howell posthoc test were utilized.

Welch's ANOVA:

Analysis of Variance (ANOVA) tests the hypothesis that the means of two or more groups are equal. Welch's ANOVA is a variation of the standard one-way ANOVA, specifically designed to handle situations where the assumption of equal variances is not met (heteroscedasticity).

In this case, Welch's ANOVA has been used because it's more robust to unequal variances and unequal sample sizes in different groups.

results:

Source	ddof1	ddof2	F	p-unc	np2
mood_label	2	5612.686595	750.038491	2.080912e-289	0.008796

Table 1- Welch's ANOVA results

With an extremely small p-value (2.080912e-289), the null hypothesis that all group means are equal can be rejected. This indicates that significant differences in the popularity scores among the three mood categories exist.

Games-Howell Posthoc Test:

After determining that significant differences between the groups exist, it became necessary to identify which specific groups differ. The Games-Howell posthoc test was applied, which, like Welch's ANOVA, does not assume equal variances and equal sample sizes. This makes it particularly appropriate for this data.

A	B	Mean(A)	Mean(B)	diff	se	T	df	pval	hedges
Energetic	Neutral	2.555789e+08	2.047009e+08	5.087806e+07	7.275393e+06	6.993169	11189.126412	9.808154e-12	0.106348
Energetic	Relaxed	2.555789e+08	5.056323e+07	2.050157e+08	5.528495e+06	37.083454	4484.137491	3.051004e-12	0.424357
Neutral	Relaxed	2.047009e+08	5.056323e+07	1.541376e+08	6.837032e+06	22.544525	5349.625269	2.253198e-12	0.385034

Table 2 - Games-Howell Posthoc

All pair-wise comparisons show significant differences in popularity scores between each pair of mood categories. For example, the popularity of 'Energetic' mood songs is significantly higher than that of 'Neutral' and 'Relaxed' mood songs.

In contrast, other tests like Tukey's HSD or standard ANOVA would not have been appropriate for this case because these tests assume homogeneity of variances and equal sample sizes. As the data did not meet these assumptions, The more robust Welch's ANOVA and Games-Howell tests were opted for

This analysis confirms initial assumption that there is a significant relationship between the mood of the song and its popularity.

6.6 Mean Popularity Analysis by Mood Label

```
# Group the dataframe by 'mood_label' and calculate the mean popularity
mean_popularity = df_cleaned.groupby('mood_label')['popularity'].mean()
print(mean_popularity)
```

At this stage of the research, a comprehensive analysis of the relationship between the mood of songs and their popularity was performed. The mean popularity of songs for each mood label was calculated to offer a general understanding of the data distribution. The results are as follows:

```
mood_label
Energetic    2.555789e+08
Neutral      2.047009e+08
Relaxed      5.056323e+07
Name: popularity, dtype: float64
```

These figures provide an interesting insight: on average, songs labeled as 'Energetic' have the highest popularity, followed by 'Neutral', with 'Relaxed' songs bringing up the rear. It's crucial to note, however, that these are average values and can be influenced by outliers or extreme values.

This analysis serves as a starting point for understanding the underlying trends and relationships between song mood and popularity. It suggests that there is a significant difference in the popularity of songs based on their mood labels.

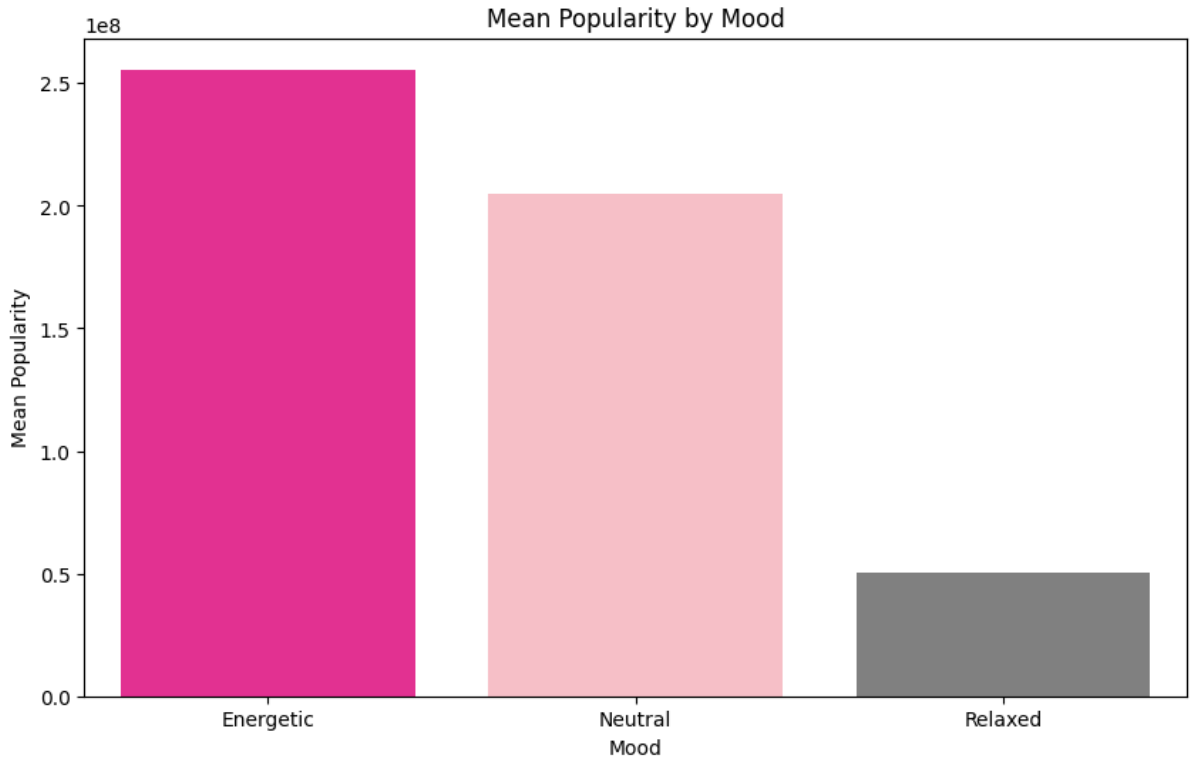


Figure 12 - Bar Chart of Mean Popularity Across Different Moods

This was also done for the median:

```
# Group the dataframe by 'mood_label' and calculate the mean popularity
median_popularity = df_cleaned.groupby('mood_label')['popularity'].median()
print(median_popularity)
```

```
mood_label
Energetic    87946118.0
Neutral      76546181.5
Relaxed      21817028.0
Name: popularity, dtype: float64
```

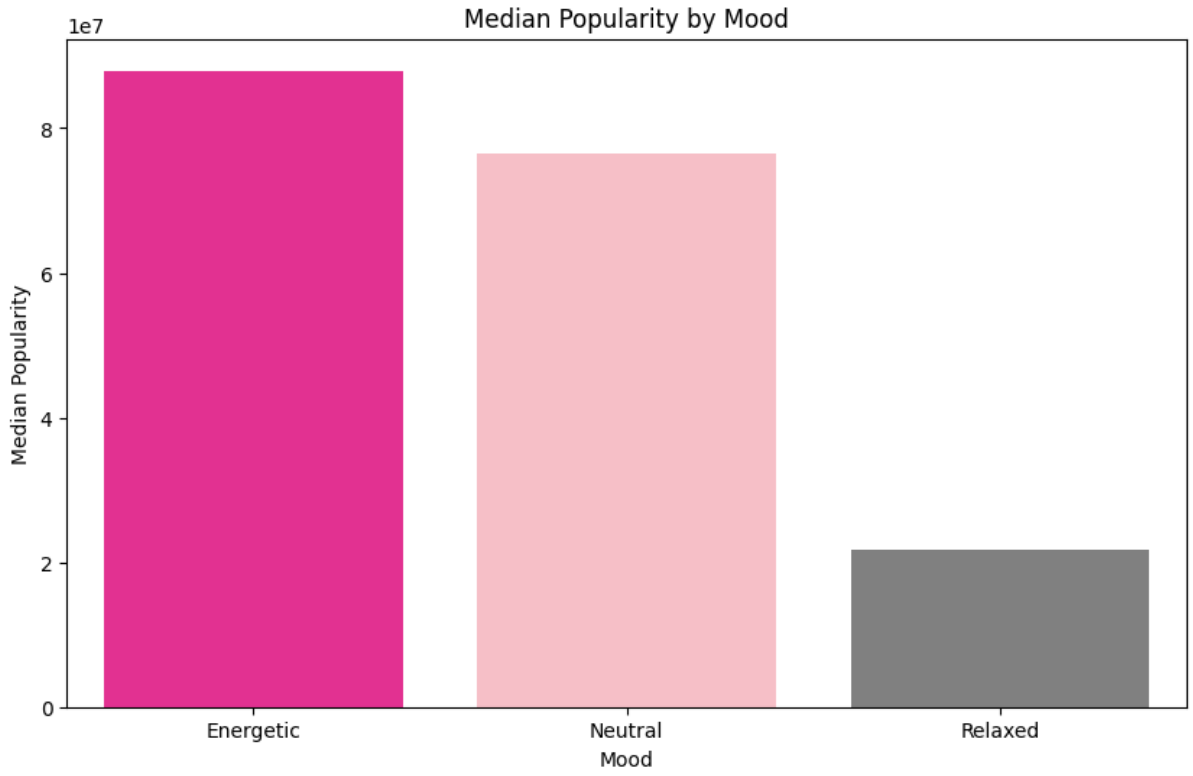


Figure 13 - Bar Chart of Median Popularity Across Different Moods

6.7 Analysis of Popularity Metrics by Mood Category

For a comprehensive understanding of how different moods impact the overall popularity of songs, various metrics that collectively define popularity were analyzed. These metrics include the number of 'Likes', 'Views', 'Streams', and 'Comments' a song has received. The bar plots represent the mean values of each of these metrics, grouped by the mood of the song, namely 'Energetic', 'Neutral', and 'Relaxed'. This approach facilitates a direct comparison

across moods for each popularity metric.

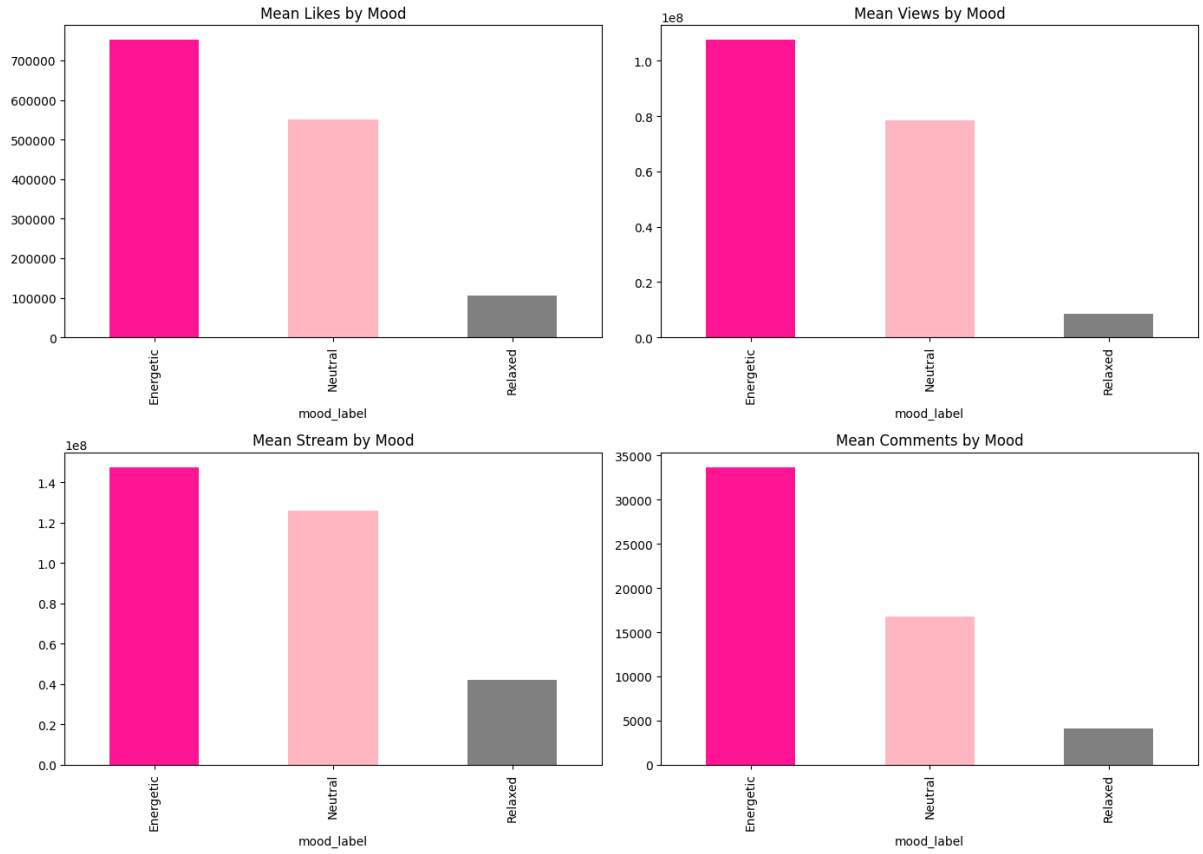


Figure 14 - Comparison of Average Likes, Views, Stream, and Comments Across Different Moods

this visual analysis strengthens the hypothesis that the mood of a song, particularly when it's energetic, has a strong correlation with its popularity. Such insights can be pivotal for artists, music producers, and platforms aiming to curate or produce content that resonates most with their target audience.

6.8 Popularity and Licensing Insights

Song licensing refers to the granting of rights by the copyright holder to use a particular track in specific ways. In the context of the dataset, licensed songs have been granted such rights, whereas unlicensed songs have not. An analysis was conducted to determine the influence of song licensing on its popularity, and the data was visualized using boxplot and barplot, comparing both categories. From the visual representation, the distribution indicates that licensed songs tend to have a distinct pattern in popularity compared to their unlicensed counterparts. This suggests that licensing might have an association with how listeners perceive the popularity of a song, although further investigation might be required to

determine causality.

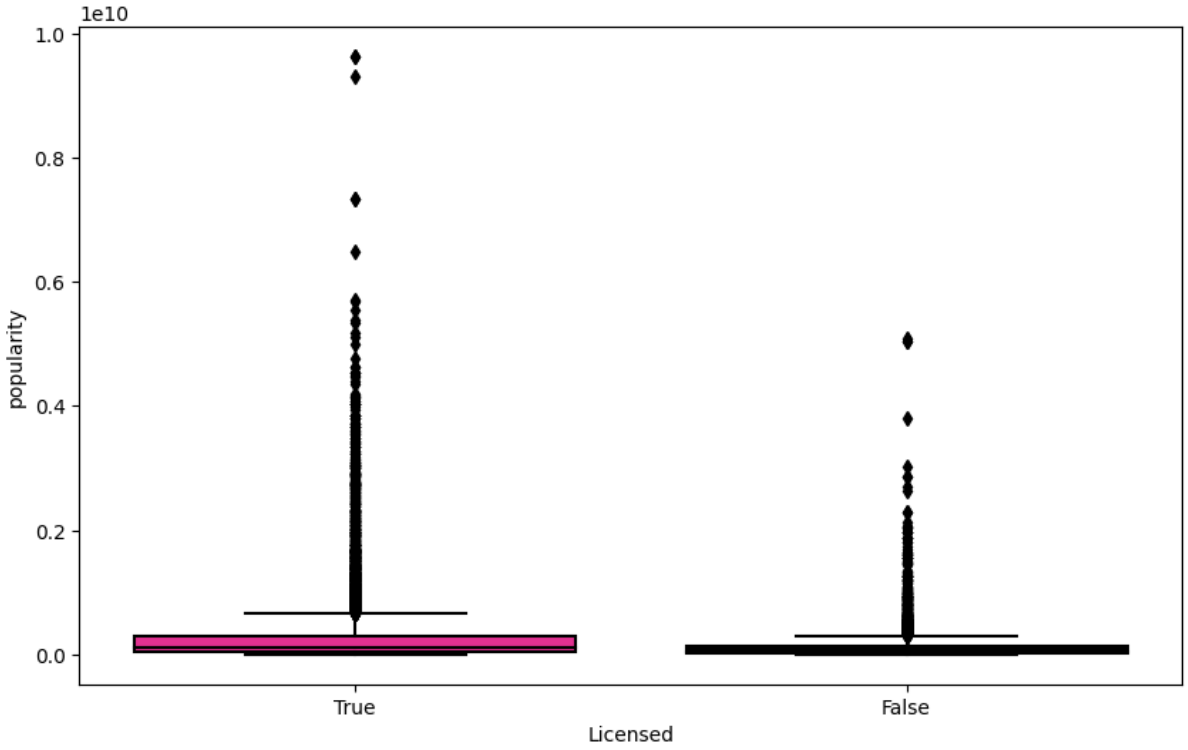


Figure 15 - Popularity Distribution Based on Licensing Status Boxplot

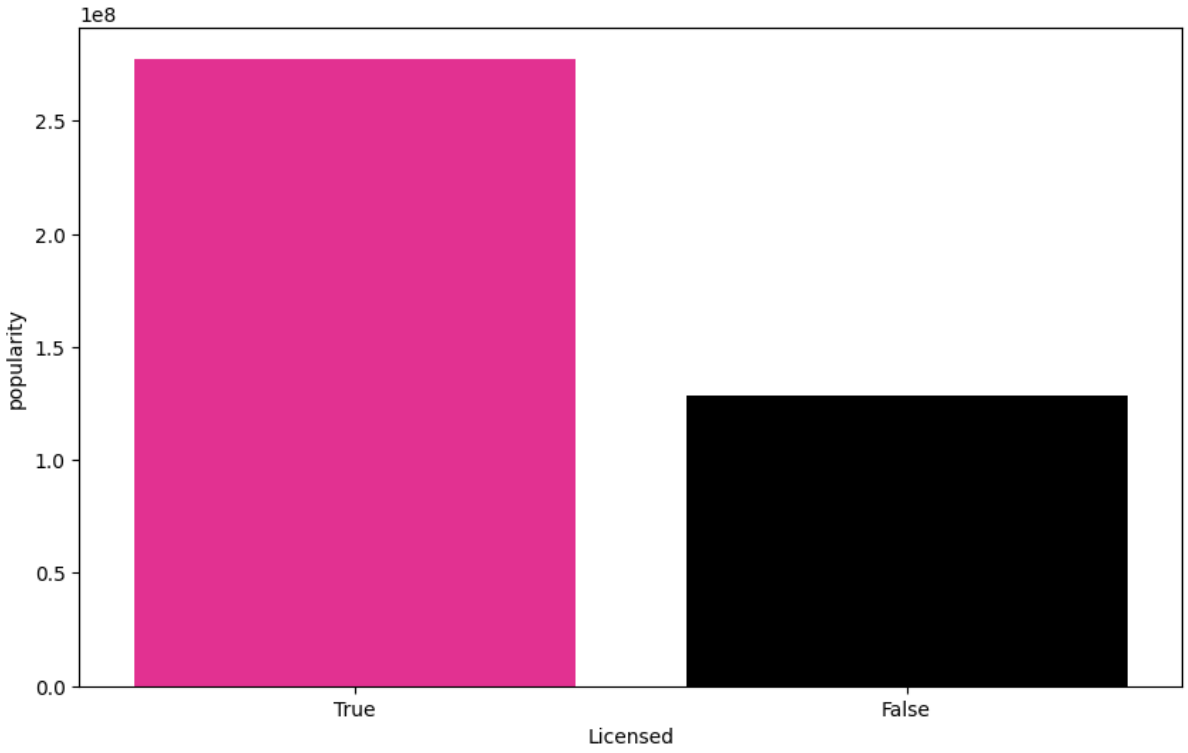


Figure 16 - Average Popularity Comparison Based on Licensing Status Barplot

Diving deeper into the realm of licensed songs, an effort was made to understand how the mood distribution might differ for songs with granted rights. A horizontal bar plot was used to depict the findings, which displayed the count of songs corresponding to each mood label within the licensed category. The graph accentuates that 'Energetic' songs dominate the licensed category, followed by 'Neutral' and 'Relaxed' tracks. This could hint at a potential preference or trend within the music industry for licensing more upbeat or 'Energetic' tracks, perhaps due to their perceived commercial appeal. This specific inclination within licensed songs reinforces the importance of considering licensing when analyzing song attributes and their potential influence on popularity.

```
licensed_songs = df_cleaned[df_cleaned['Licensed'] == True]
mood_counts = licensed_songs.groupby('mood_label').size().reset_index(name='Count')
# Define a color dictionary
colors = {'Energetic': 'deeppink', 'Relaxed': 'grey', 'Neutral': 'lightpink'}
plt.figure(figsize=(10, 6))
# Use the color dictionary in the barplot
sns.barplot(x='Count', y='mood_label', data=mood_counts, palette=colors)
plt.xlabel('Number of Licensed Songs')
plt.ylabel('Mood Label')
plt.title('Number of Licensed Songs by Mood Label')
plt.show()
```

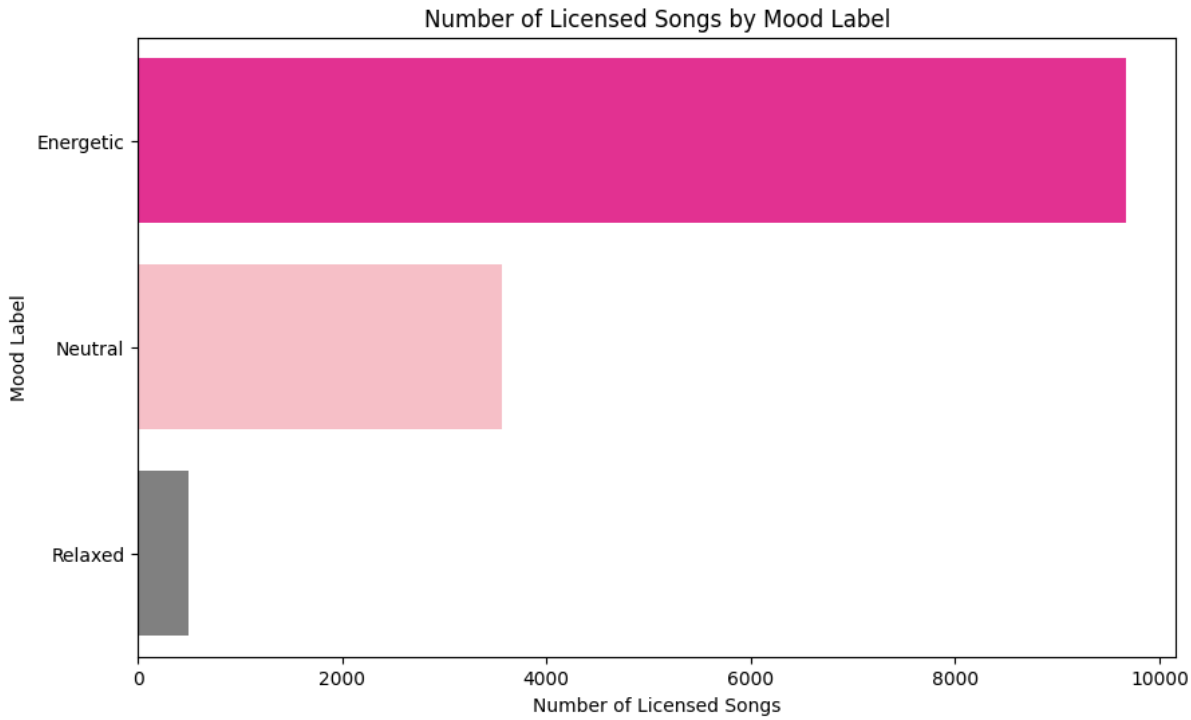


Figure 17 - Licensed Song Distribution Across Mood Labels

This relationship between licensing and popularity was also analyzed for balanced data, and similar results were obtained once again.

```
#Group by 'Licensed' column and calculate mean popularity
licensed_popularity = df_balanced.groupby('Licensed')['popularity'].mean().reset_index()
# Create bar chart
plt.figure(figsize=(10, 6))
sns.barplot(x='Licensed', y='popularity', data=licensed_popularity)
plt.title('Average Popularity by License Status')
plt.ylabel('Average Popularity')
plt.xlabel('License Status')
plt.show()
```

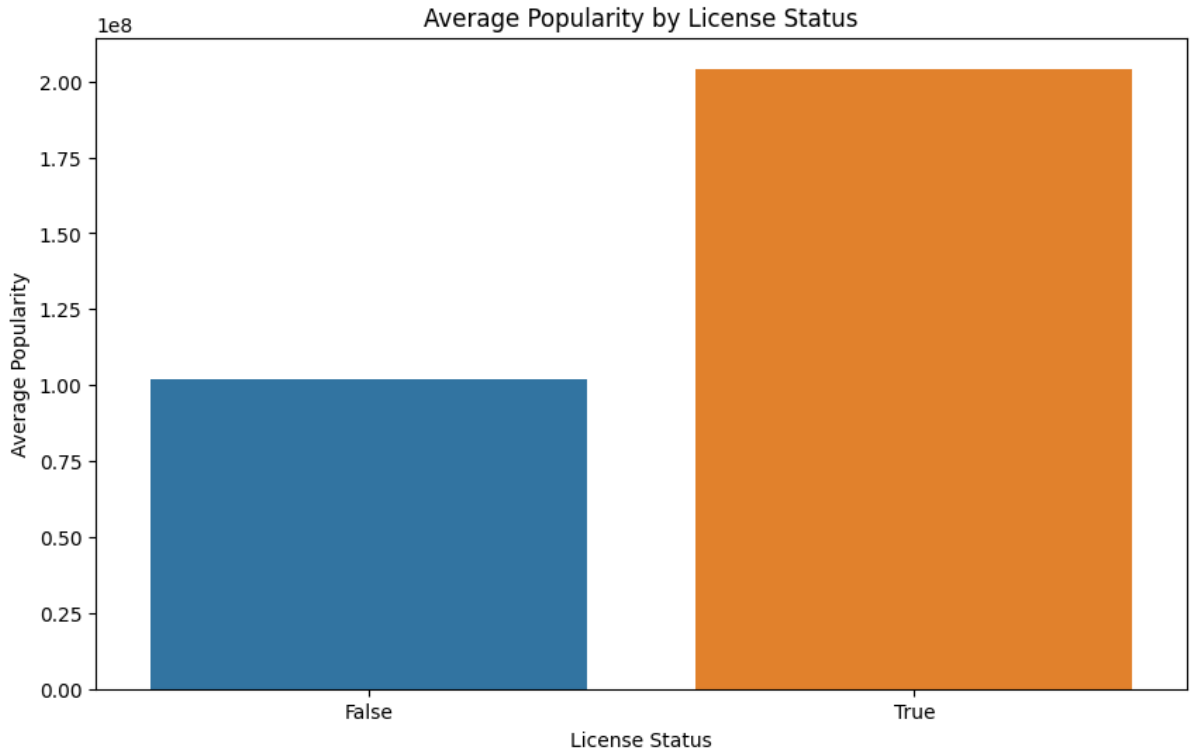


Figure 18 - Average Popularity Comparison Between Licensed and Unlicensed Songs in Balanced Data

As can be seen, regardless of whether the data is balanced or imbalanced, possessing a license appears to be an important factor in increasing a song's popularity.

6.9 Popularity Distribution Across Song Types

From the visual representation, it's evident that among various song types, 'Album' garners the highest mean popularity. This highlights a significant trend: listeners exhibit a preference for songs that are part of a complete album. Such songs, often intricately woven with a central theme or narrative, could resonate more with the audience due to their holistic musical experience.

For in-store environments, where ambiance and prolonged engagement are pivotal, playing tracks from popular albums can be a strategic choice. Given their demonstrated popularity, these tracks could potentially keep the shoppers engaged longer, promoting a more extended stay and, by extension, increased purchasing potential. Furthermore, selecting from top albums can ensure a consistent mood and thematic resonance, enhancing the overall in-store ambiance and experience for customers.

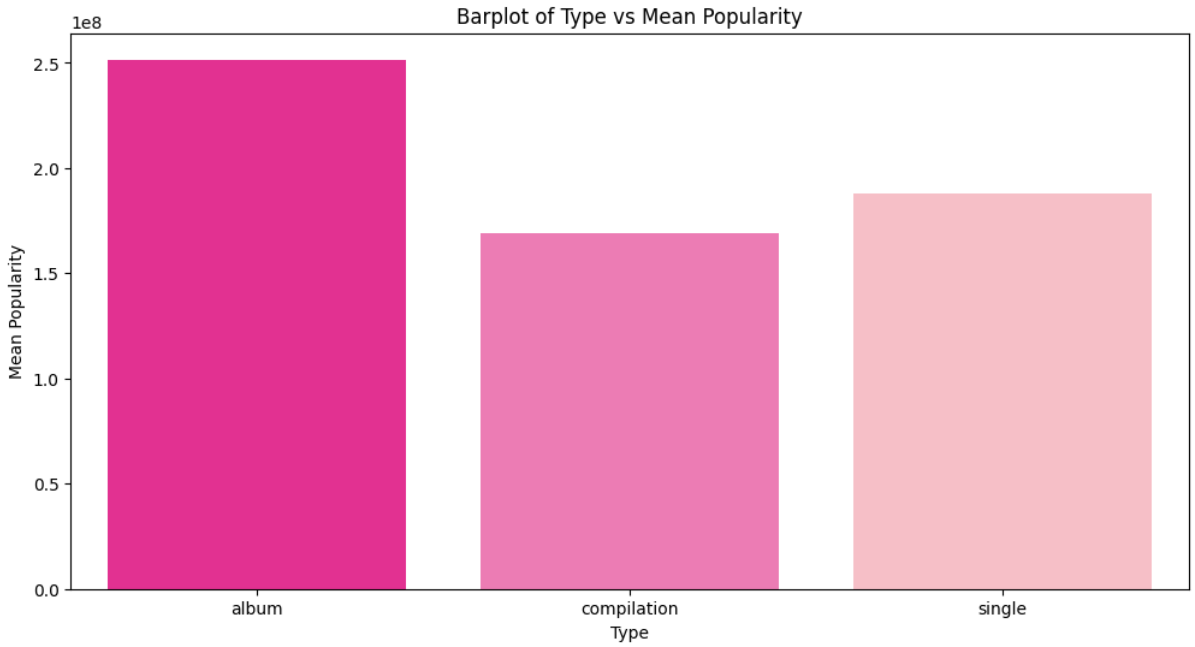


Figure 19 - Mean Popularity Distribution Across Different Types

Again, when visualized for the balanced data, the results remained consistent: the possession of a license appears to be a significant factor in increasing a song's popularity

```
mean_popularity_balanced = df_balanced.groupby('Type')['popularity'].mean()
```

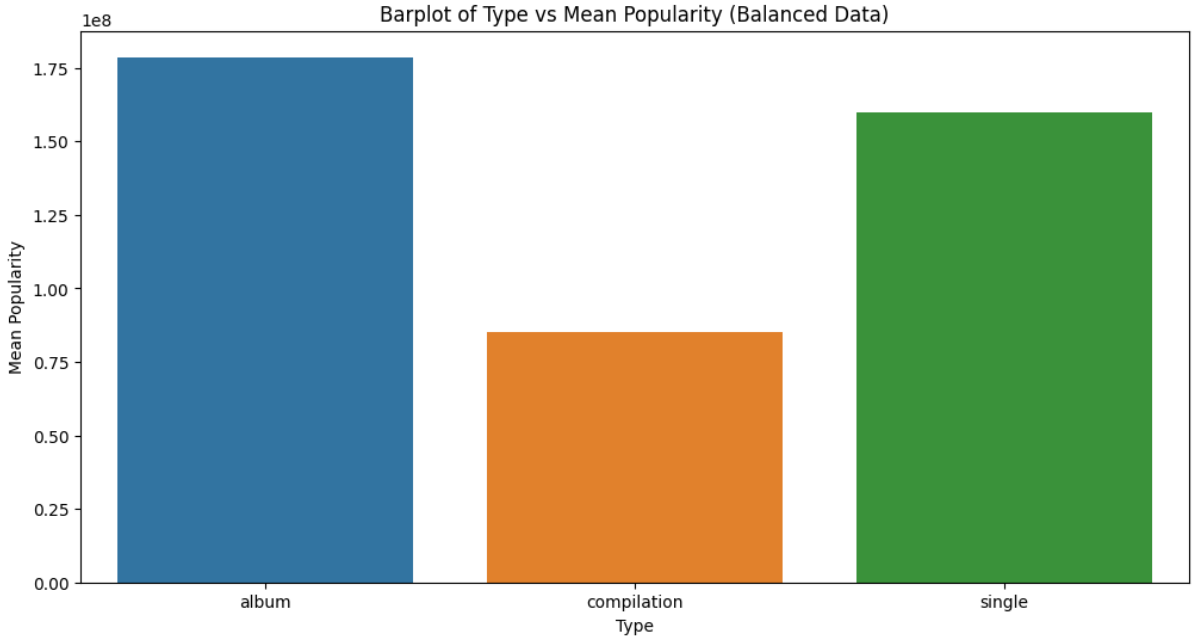


Figure 20 - Mean Popularity Distribution Across Different Types (Using Balanced Data)

Influence of Album Type on Song Popularity Across Mood Labels

The bar plot showcases the distribution of song types (such as albums, singles, and compilations) across different mood categories. The colors denote different moods, with 'Energetic' represented in 'deeppink', 'Relaxed' in 'grey', and 'Neutral' in 'lightpink'. From the visualization, it's evident that songs categorized under the 'Energetic' mood have a significant count of album-type songs. Given the prior understanding that 'Energetic' songs tend to be more popular among listeners, and considering the dominance of the 'Album' type in this category, it's compelling to infer that album songs are a preferred choice among listeners, especially those resonating with more lively, energetic vibes.

```
# Create a color list based on the order of mood labels in the DataFrame
color_list = [color_dict[mood] for mood in mood_album_counts['mood_label'].unique()]
# Create a bar plot
plt.figure(figsize=(12, 6))
sns.barplot(x='mood_label', y='Count', hue='Type', data=mood_album_counts,
palette=color_list)
plt.title('Number of Each Type in Each Mood Category')
plt.show()
```

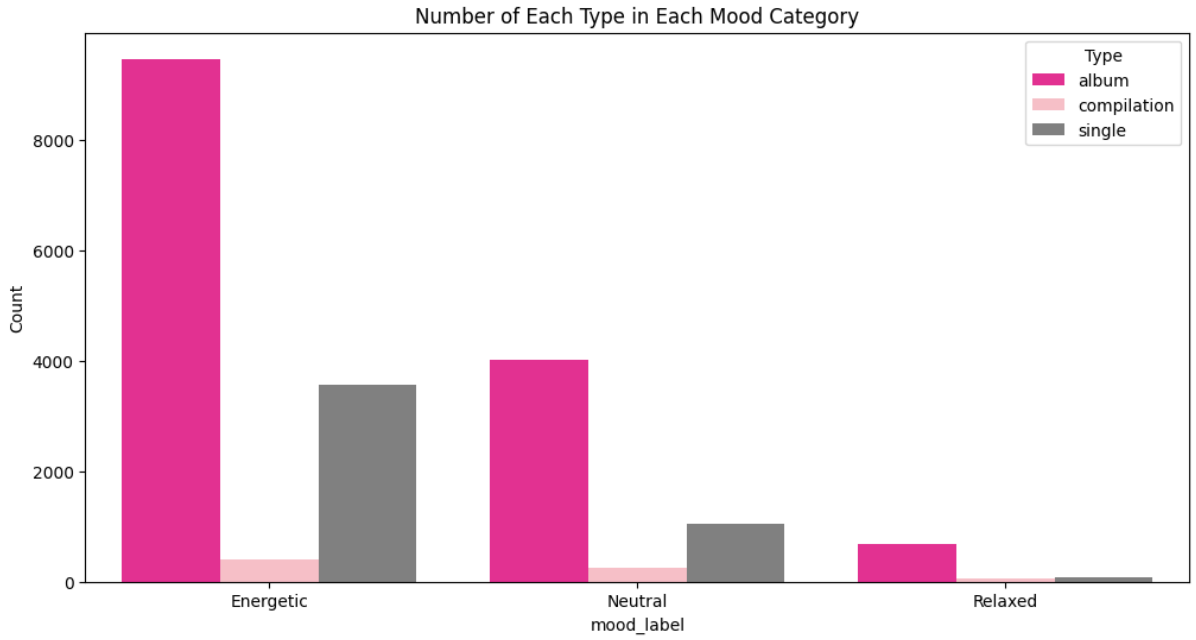


Figure 21 - Type Distribution Across Different Moods

```
mood_album_pivot = pd.pivot_table(mood_album_counts, values='Count',
index='mood_label', columns='Type', fill_value=0)
```

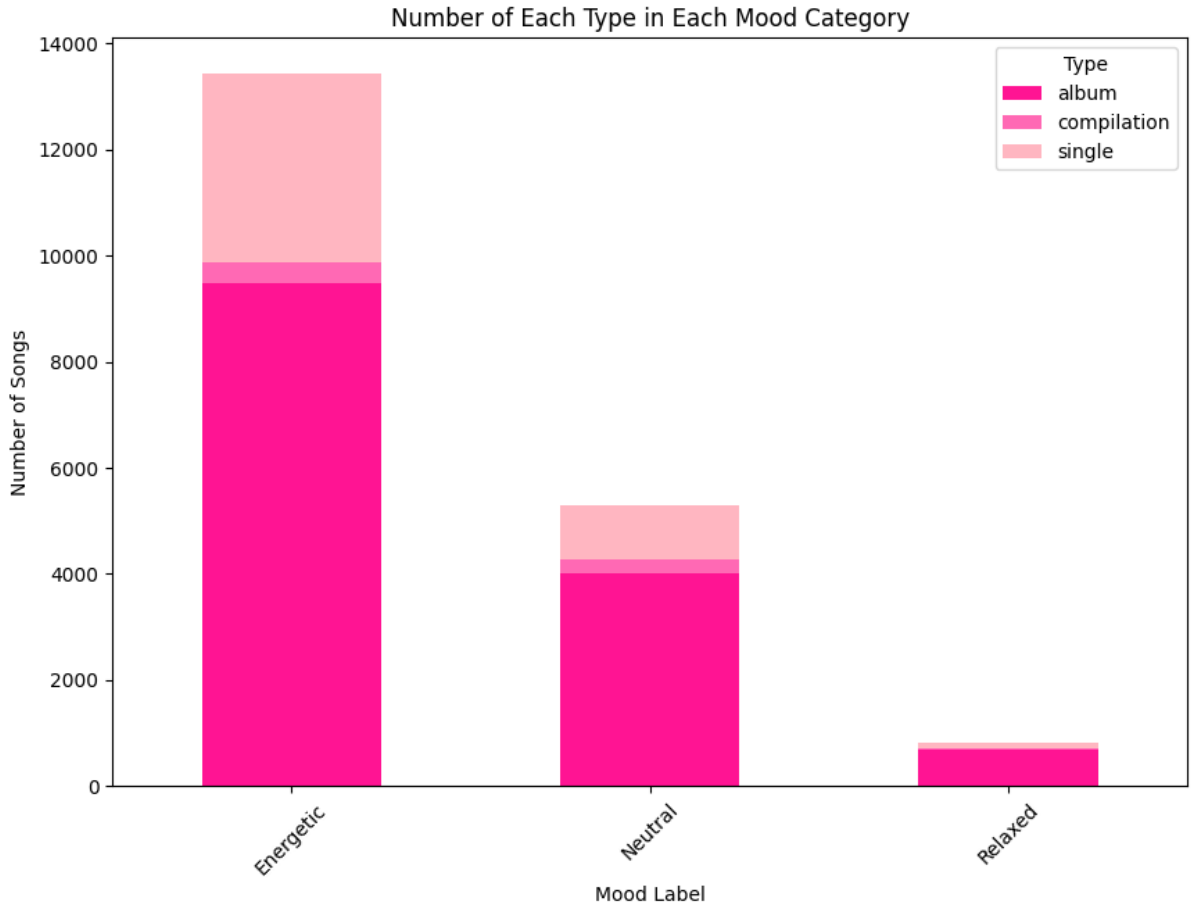


Figure 22 - MoodCategory vs SongTypes StackedBarplot

In essence, this chapter provides rich insights into what drives a song's popularity on music streaming platforms, particularly highlighting the substantial influence of song mood on listener preferences. This knowledge has far-reaching implications, especially in the context of curating in-store music playlists to enrich customer experiences and interactions.

7.Interpretation and Discussion

7.1 Listener Preferences and the Predominance of Energetic Tracks

At the crux of our findings, 'Energetic' tracks have emerged as the clear frontrunners in terms of popularity. With a significant lead over 'Neutral' and 'Relaxed' songs, it becomes evident that there's an innate predilection among listeners for more upbeat, vivacious music. This inclination can be attributed to several reasons:

- Energetic music could serve as a source of motivation, especially in a world that is increasingly fast-paced.
- Such tracks might resonate more with the current zeitgeist, reflecting societal values of passion, determination, and enthusiasm.
- From a physiological perspective, upbeat songs often trigger the release of dopamine, a neurotransmitter associated with pleasure and reward.

7.2 Licensed Songs and Their Indomitable Appeal

The analysis showcases a telling correlation between song licensing and popularity. Licensed songs, which have been granted rights by copyright holders, consistently chart higher in terms of listener engagement. This could be because:

- Licensed tracks might be of superior quality or produced by renowned artists, drawing more listeners.
- They might also receive more promotion, becoming more accessible and familiar to the audience.
- There's a possibility that listeners equate licensing with authenticity, leading to increased trust and engagement.

7.3 Crafting the Perfect In-Store Playlist

An offshoot of our primary research is the tangible application in the realm of in-store music curation. Businesses can strategically leverage these insights to:

- Prioritize 'Energetic' tracks, capitalizing on their undeniable allure to foster a vibrant shopping environment.
- Incorporate licensed music, benefiting from their proven popularity metrics.

- Ensure a mix of moods, with a significant lean towards the upbeat, crafting a holistic yet engaging ambiance.
  The reverberations of such a meticulously curated playlist could manifest in prolonged customer engagement, an enhanced shopping experience, and potentially, augmented sales.

7.4 The Distinctive Nature of This Research

In the vast expanse of music research, several elements set this study apart:

- **Use of Existing Data:** By harnessing available data, the research is not only feasible but also extendable for future endeavours.
- **Objective Analysis:** The focus remains unwaveringly on clear popularity metrics, offering an unfiltered view of prevailing music trends.
- **Sidestepping Feedback Flaws:** Traditional feedback mechanisms are fraught with biases and inaccuracies. This study's reliance on empirical data circumvents these issues.
- **Direct Practical Implementation:** The findings aren't confined to theoretical realms. Their real-world application, especially in shaping in-store music experiences, is invaluable for businesses.
- **Unwavering Faith in Data:** At the heart of this research lies an unwavering belief in data's power to elucidate music preferences, eliminating the need for subjective tools like surveys.
  In essence, this research not only illuminates the prevailing musical inclinations but also paves the way for tangible applications in business environments.

Certainly, discussing the decision to not use balanced data and the reasons for doing so would be vital for the reader to understand the methodological choices and any potential biases or limitations of the research.
Let's include a section in your discussion chapter that sheds light on this:

7.5 Data Authenticity and the Choice Against Balancing Techniques

One of the pivotal methodological decisions in this research was to maintain the integrity of the original data distribution without resorting to data balancing techniques. Here's why:

1. **Representation of Real-world Scenarios**: Balancing techniques, though beneficial in some contexts, can often distort the actual distribution of data in real-world scenarios. By preserving the authentic distribution, our research offers insights that are directly applicable and reflective of the real-world music landscape.

2. **Avoiding Overgeneralization**: Techniques like over-sampling or under-sampling, commonly employed to achieve balanced data, can sometimes lead to overgeneralization or loss of vital information, respectively. Avoiding such techniques ensures that our results are not influenced by artificially inflated or reduced data points.
3. **Emphasis on Natural Patterns**: By refraining from modifying the dataset to achieve balance, we underscore our commitment to identifying natural patterns and trends in the data. This is crucial for drawing genuine insights and understanding the true relationship between song moods and their popularity.
4. **Avoidance of Potential Biases**: Balancing techniques can inadvertently introduce biases. For instance, over-sampling might amplify noise or outliers, while under-sampling could overlook them. By working with the original dataset, we ensure that our results aren't skewed by such potential biases.
5. **Acknowledgment of Imbalances**: Recognizing and analyzing data as it is – even if imbalanced – allows us to shed light on the very imbalances that might be of interest. For instance, the dominance of 'Energetic' tracks in our dataset wasn't masked or adjusted, but rather highlighted and deeply explored.

Music, in its myriad forms, has always been reflective of societal tendencies. By delving deep into the relationship between song moods and their popularity, this research unfurls the predominant musical preferences of contemporary society. Beyond mere numbers, it offers invaluable insights for businesses, especially those keen on harnessing the magic of music to elevate consumer experiences. Through a data-centric lens, this study spotlights the undeniable influence of 'Energetic' tracks, the allure of licensed songs, and the immense potential of these findings in crafting in-store musical landscapes.

Drawing from and building upon the existing literature, our research provides fresh insights into the intricate dynamics of song popularity. We've contextualized our findings within prior research, offering both continuity and novel contributions to the field. Future endeavors might integrate more nuanced factors, such as sociological and psychological elements, to deepen the understanding of music consumption patterns.

By weaving in the research of (Sbai, 2019)about the influence of fast music on arousal and impulse buying, the chapter now establishes a robust connection between prior studies and current findings, especially in the context of energetic music's popularity.

8.Legal, Social, Ethical and Professional issues

1. Legal Issues: The realm of music data is known to be laced with legal intricacies, especially when topics like copyright and data privacy are brought up.

- Copyright Concerns: Care was taken in the curation of our dataset to ensure no breach of copyright laws. Specifically, a focus was placed on song metrics rather than the actual songs, thus avoiding direct copyright conflicts.
- Data Privacy: Priority was given to privacy. In alignment with data privacy principles, personally identifiable information was avoided in our dataset. As a result, strict adherence to regulations, including the General Data Protection Regulation (GDPR), was ensured. (University, 2022)

2. Social Issues :Music is recognized not just as entertainment but as a significant influence on society, culture, trends, and personal behaviors. (Peralta, 2021)

- Music Industry Trends: Insights were provided by our findings, especially the rise of 'Energetic' tracks, which could potentially steer future music production directions.
- Audience Impact: Despite the reliance on data in our analysis, the importance of understanding the variability and subjectivity of human preferences was recognized. An emphasis was placed on the need for music to cater to a broad spectrum of tastes, beyond just popular trends.

3. Ethical Issues :When exploring music data, an ethical stance was deemed essential throughout the research process.

- Dataset Biases: An awareness of potential biases, like the possible underrepresentation of specific music moods or genres, was maintained. A conscious decision was made not to artificially balance our dataset, preserving its authenticity while also recognizing inherent biases.
- Misuse Prevention: Concerns were raised regarding the potential for the industry to become overly focused on 'Energetic' songs based on our findings. A proactive emphasis was placed on the value of varied musical landscapes, warning against a disproportionate focus on any single mood or genre. (PLoSOne, 2020)

4. Professional Issues: A commitment to professional integrity was evident in all research efforts.

- Adherence to Standards: The approach, grounded in industry benchmarks, followed best practices for data analysis.
- Transparency and Accuracy: A commitment was made throughout the research to maintain transparency. Each decision, potential shortcoming, and rationale was clearly presented, ensuring that the thought process could be easily understood by readers. (globalresearchcouncil, -)

To conclude In this chapter, a spotlight was shone on the multiple facets of considerations – legal, social, ethical, and professional – intertwined with music data research. Through the acknowledgment and addressal of these complexities, the research was shown to be founded not just on data but also on integrity and foresight.

9.Conclusion and Future Work

Music in the Digital Age: In today's technology-driven world, the way music is listened to has been transformed. Big platforms such as Spotify and YouTube have been instrumental in reshaping the manner in which songs are engaged with. In this evolving context, an investigation was conducted to understand the factors determining a song's popularity and its implications for commercial entities.

9.1 Key Discoveries:

The Power of Mood:

Aim: Explore the relationship between song moods and their popularity rankings.

Insight: 'Energetic' tracks have a unique allure, consistently dominating both in volume and ranking, signifying listeners' preference for invigorating melodies.

Attributes of Chart Dominance:

Aim: Identify attributes consistent among top-ranked songs, emphasizing licensing and song types.

Insight: Licensed songs and those part of an album narrative consistently outperformed others, emphasizing their elevated visibility and inherent appeal.

In-store Playlist Potency:

Aim: Translate insights into optimizing in-store playlists.

Insight: Understanding the predilection for 'Energetic' tracks and licensed songs allows businesses to craft in-store auditory experiences that deeply resonate with their customers, promising heightened engagement and a superior shopping environment.

9.2 Future Work:Continuing The Musical Exploration

Understanding Listener Choices:

Knowing what songs people like is a start, but we should also explore why they like them to better understand their feelings and thoughts.

Tech's Influence on Music:

As technology changes, it's important to see how it affects the way people listen to music. This will help predict future trends in music listening habits.

Cultural Resonance:

Music is interwoven with cultural nuances. Investigating how cultural and regional variations shape song preferences can provide a panoramic view of global listening patterns.

Temporal Evolution:

Mapping the evolving preferences across eras can highlight the influence of societal, technological, and cultural shifts on music tastes.

From Theory to Reality:

Validating findings in tangible settings, like integrating curated playlists in commercial venues, will measure the actual impact on consumer behaviors. Furthermore, discerning online listening habits from in-store preferences is crucial, as the ambiance and context may lead to divergent choices.

Holistic Comprehension:

Future efforts should strive to present a rounded understanding of the listener-music relationship, accounting for varied factors like mood, cultural background, age, and distinct musical tastes.

To recap:

A detailed look was taken into music data, showing what people like to listen to nowadays. From what was found out, there are clear suggestions that businesses and those who love music can follow. Even so, there's a lot more about music that hasn't been found out yet. As music changes with technology, the goal is to keep understanding these changes. This helps make sure everyone can keep enjoying music.

References

Ben, 2021. *musictheoryacademy.com.* [Online]
Available at: https://www.musictheoryacademy.com/how-to-read-sheet-music/tempo/
[Accessed - - 2021].

Daunfeldt, S.-O., 2019. *The Effect of Employees'Opportunities to Influence in-store Music on Sales,* Stockholm: s.n.,

Dimolitsas, I., 2020. *SpotHitPy:A Study for ML-based song hit predicton using spotify.,* Athens: National Technical Univrsity of Athens.

DonSolare, 2018. *community.spotify.com/.* [Online]
Available at: https://community.spotify.com/t5/Spotify-for-Developers/Valence-as-a-measure-of-happiness/td-p/4385221
[Accessed 11 february 2018].

Duman, D., 2022. Music we move to:spotify audio features and reasons for listening. *PLoS ONE,* 17(9), p. 18.

Fernandes, S., 2023. *Comparison of deep learning and machine learning in music genre categorization,* -: national college of Ireland.

globalresearchcouncil, -. *globalresearchcouncil.org/.* [Online]
Available at:
https://globalresearchcouncil.org/fileadmin/documents/GRC_Publications/grc_statement_principles_research_integrity_FINAL.pdf
[Accessed - - -].

Gomez-Canon, J., 2022. TROMPA-MER:an open dataset for personalized music emotion recognition. *Ambient Intelligence and Huminized computing,* -(-), p. 22.

Kutlimuratov, A., 2023. *Music recommender system.* Tashkant, Tashkant university of information technologies.

M.J.Krause, A., 2021. *Elucidation of the relationship between a song's Spotify Descriptive Metrics and its popularity on Various Platforms.* -, -.

MeghaSharma, 2023. The Effect Of Retail Store Environment And Personality Traits On Impulsive Buying Behaviour. *Dogo Rangsang Research Journal,* 14(2), p. 13.

minutes, B. 6., 2023. *Sonic seasoning.* [Sound Recording] (BBC).

Peralta, L., 2021. *www.savethemusic.org.* [Online]
Available at: https://www.savethemusic.org/blog/how-does-music-affect-society/
[Accessed 3 November 2021].

Pipilis, G., 2023. *creating &evaluating a music recommender system without access to multiple user data,* Athens: -.

PLoSOne, 2020. *www.ncbi.nlm.nih.gov.* [Online]
Available at:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7451523/#:~:text=The%20thematic%20analysis%20generated%206,private%20versus%20public%20conceptualizations%20of
[Accessed 27 August 2020].

rastelli, S., 2023. *kaggle.com.* [Online]
Available at: https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube
[Accessed 7 february 2023].

S.Zhang, D., 2023. The impact of background music and live streamers'explanations on consumer behaviour in live streaming commerce:an SOR perspective. *Journal of Retailing and Consumer Services,* 75(-), p. 12.

Sbai, I. H., 2019. *An Exploration of the effect of background Music in Retail stores on Moroccan Consumers'perception and behaviors.* -, Springer International Publishing.

Singh, S., 2023. An exploratory study on analyzing the popularity of songs using spotify data. *Innovative Ideas and Invention in computer science,* 8(7), p. 13.

Software, 2020. *software.com.* [Online]
Available at: https://www.software.com/src/explore-the-data-behind-your-most-productive-music-for-coding
[Accessed 19 May 2020].

Studiobinder, 2021. *studiobinder.com.* [Online]
Available at: https://www.studiobinder.com/blog/what-is-music-licensing/
[Accessed 18 April 2021].

Terroso-Saenz, 2023. Predicting music preferences using contextual factors and machine learning. *Entertainment Computing,* 44(-), p. 16.

University, T., 2022. *www.futurelearn.com.* [Online]
Available at: https://www.futurelearn.com/info/courses/key-topics-in-digital-transformation/0/steps/261259
[Accessed - - 2022].

Yee, J. R., 2022. predicting music popularity using spotify and youtube features. *Indian journal of science and technology,* 15(36), p. 14.

Bibliography

Arul. (2022). Music Classification Using Machine Learning. Journal of Audio Analysis.

Atzil. (2023). Effects of Sociality and Affective Valence on Brain Activation: A Meta-Analysis. Journal of Neurological Research.

Avinash Navlani. (2021). Python for Data Analysis. Python Programming Press.

Baig. (2021). Machine Learning Techniques and Their Applications. Machine Learning Journal.

Banachewicz. (2022). Advanced Python Programming. Tech Publications.

Bergin. (2018). Python Programming for Beginners. Tech Education Press.

Campbell. (2020). Data Analysis with Python. Wiley Press.

Chen, C.-h. (2007). Handbook of Data Visualization. Berlin: Springer Science & Business Media.

Dong, G. (2018). Feature Engineering for Machine Learning and Data Analytics. California: CRC Press.

evidencen. (2023). How to engineer a new feature in Python using pandas. Retrieved from https://evidencen.com/how-to-engineer-a-new-feature-in-python-using-pandas/

Fernando Terroso-Saenz. (2023). Prediction of Song Success Using Machine Learning. Music Data Analysis Journal.

FreeCodeCamp. (2021). Python Machine Learning Tutorial. Retrieved from https://www.youtube.com/watch?v=GPVsHOlRBBI&t=1s

Gomes, I. P. I. S. I. A. M. A.-Y.-O. M. (., 2021). Keeping the Beat on: A Case Study of Spotify. Trends and Applications in Information Systems and Technologies, 1366(10), pp. 5-33.

gov.uk. (2020). Review into bias in algorithmic decision-making. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957259/Review_into_bias_in_algorithmic_decision-making.pdf

gov.uk. (2021). Ethics transparency and accountability framework for automated decision-making. Retrieved from https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making/ethics-transparency-and-accountability-framework-for-automated-decision-making

Healy, K. (2018). Data Visualization: A Practical Introduction. Princeton University Press.

Hu. (2010). Function of Lyrics in Music Mood Classification. Music Psychology Journal.

Hwang. (2019). Using Python and R for Enhanced Marketing Strategy. Marketing Science Review.

Jandaghian. (2023). Data Analysis and Visualization with Python. Python Programming Press.

Jacqueline Kazil. (2016). Data Analysis with Python. Wiley Press.

jain, P. (2022). Getting started with feature engineering. Retrieved from https://www.analyticsvidhya.com/blog/2020/10/getting-started-with-feature-engineering/

Khannade. (2023). Emotion Recognition in Music. Music Data Analysis Journal.

L. Vardo. (2023). "Predicting Song Success: Understanding Track Features and Predicting Popularity Using Spotify Data. 2023 22nd International Symposium INFOTEH, 10(11), p. 6.

Lawton, G. Data Preprocessing. Retrieved from https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing

Maroely. (2023). The Effect of Customized Music in Advertising. Marketing and Music Journal.

openclassrooms. (2021). Train a supervised machine learning model. Retrieved from https://openclassrooms.com/en/courses/6389626-train-a-supervised-machine-learning-model/6398776-create-new-features-from-existing-features

Patel. Music Marketing Web Resource.

Prince, C. w. (2022). Python Data Analysis Tutorial. Retrieved from https://www.youtube.com/watch?v=bynsxAbjImQ

ProjectPro. (2022). 8 Feature engineering techniques for machine learning. Retrieved from https://www.projectpro.io/article/8-feature-engineering-techniques-for-machine-learning/423

Provost. (2013). Data Mining and Data-Analytic Thinking. Data Science for Business.

salvatorerastelli. (2023). Spotify and YouTube Dataset. Retrieved from https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube

Schwabish. (2021). Data Visualization with Python. Python Programming Press.

Simplilearn. (2021). Data Analysis with Python Tutorial. Retrieved from https://www.youtube.com/watch?v=7kPqESo1vRw&t=2s

SMIECH, D. (2023). Spotify YouTube EDA and OLS Regression. Retrieved from https://www.kaggle.com/code/danielsmiech/spotify-youtube-eda-and-ols-regression

tableau. (2023). Business Intelligence. Retrieved from https://www.tableau.com/en-gb/learn/articles/business-intelligence

Ukdataservice. Copyright in data. Retrieved from https://ukdataservice.ac.uk/learning-hub/research-data-management/rights-in-data/copyright/

Walker. (2020). Advanced Python Programming. Python Programming Press.

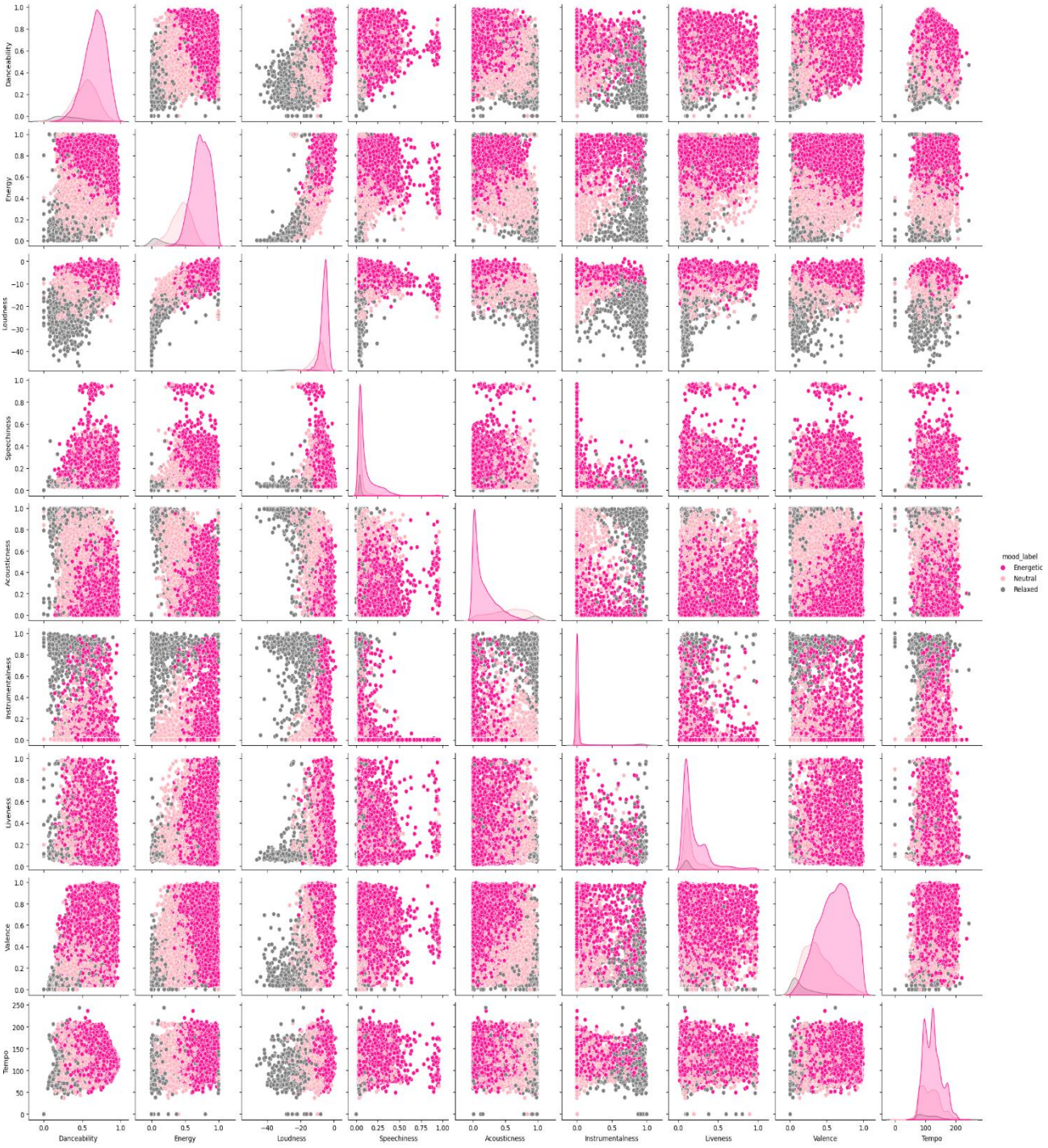
Zave, P. (2003). An experiment in feature engineering. In: S. b. archive, ed. Programming Methodology. New York: Springer, pp. 7-17.

Appendices

A.additional plots and insights:

Mood-Based Scatter Plot Matrix of Song Attributes:

This matrix offers a visual representation of how each song attribute pairs with every other, colored distinctively based on the song's mood. By examining this matrix, one can discern patterns and relationships between the features, potentially identifying how certain attributes might influence or correlate with a song's mood.

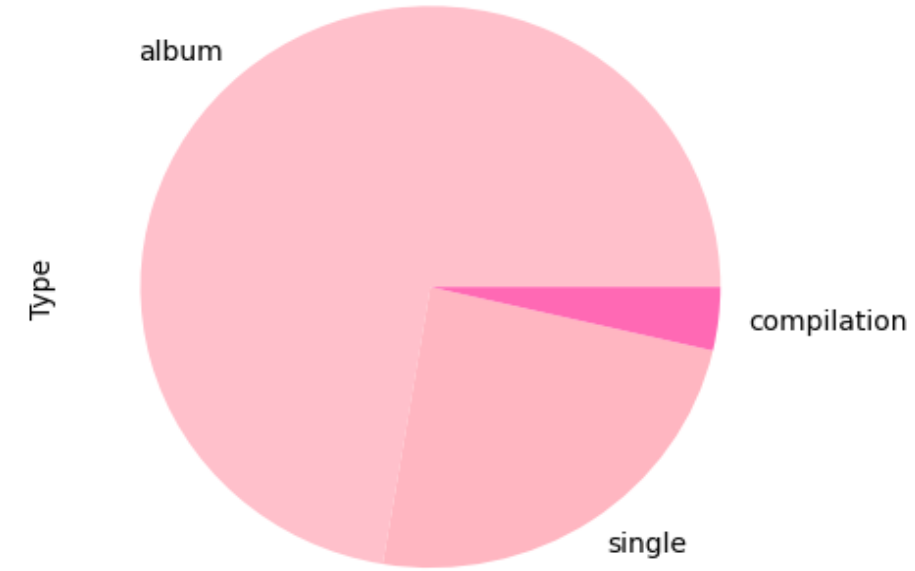


Appendix Figure 1 - Mood-based Feature Distributions and Relationships

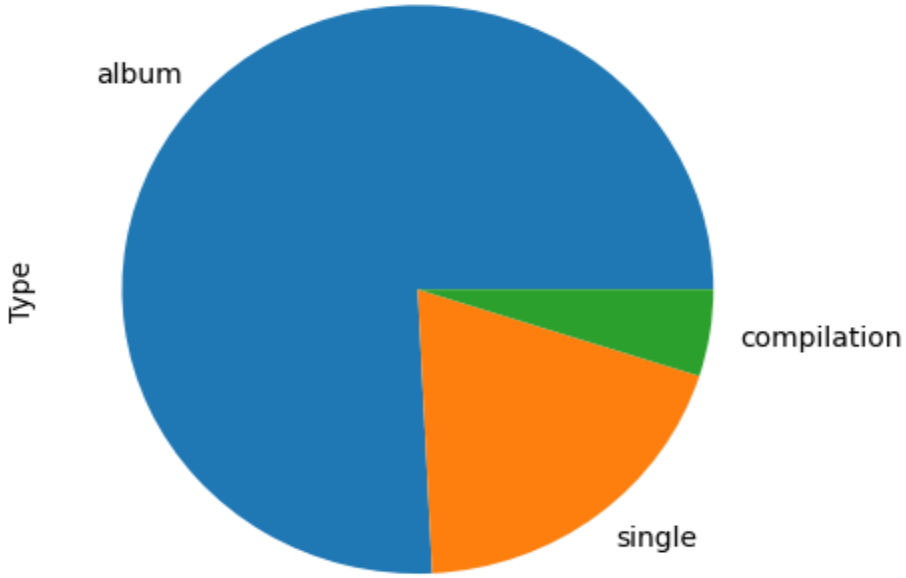
Distribution of Song Types

The pie chart visualizes the distribution of different song types within the dataset. Each segment of the pie represents a specific song type, such as "album", "single", or "compilation". The varying shades of pink denote different types, with the size of each segment indicating the proportion of that type in relation to the whole dataset. Interestingly,

when analyzing both the balanced and unbalanced data, the distribution patterns remained consistent, suggesting that the proportions of song types are robust against sampling variations. This chart provides an intuitive snapshot of the prevalence of each song type, allowing readers to quickly discern which types are more common than others.



Appendix Figure 2 - Proportional Distribution of Song Types



Appendix Figure 3 - Proportional Distribution of Song Types Balanced Data

Analyzing Top Danceability Songs for Cluster Validation

The songs with the highest "Danceability" scores were extracted from the dataset, a metric typically used to denote the suitability of a song for dancing. The top 4 songs, based on their danceability scores, were selected to inspect and verify the accuracy of the clustering process in grouping songs perceived as highly danceable. The dataset **df_cleaned** was sorted in descending order based on the 'Danceability' feature, and the top 4 rows were selected. Only the 'Artist', 'Track', and 'Danceability' columns of the resulting data (**top_danceable_songs**) were displayed. By inspecting these songs, an intuitive check on the clustering algorithm was conducted. If songs with high danceability were found to be closely grouped within a specific cluster, it would indicate that the clustering was functioning as anticipated. Further, the presence of well-known upbeat artists and tracks in this list would hint at the clustering's capability to group songs in a way that matched the conventional understanding of what makes a track 'danceable'.

```
# Display the top danceable songs along with the artists
print(top_danceable_songs[['Artist', 'Track', 'Danceability']])
```

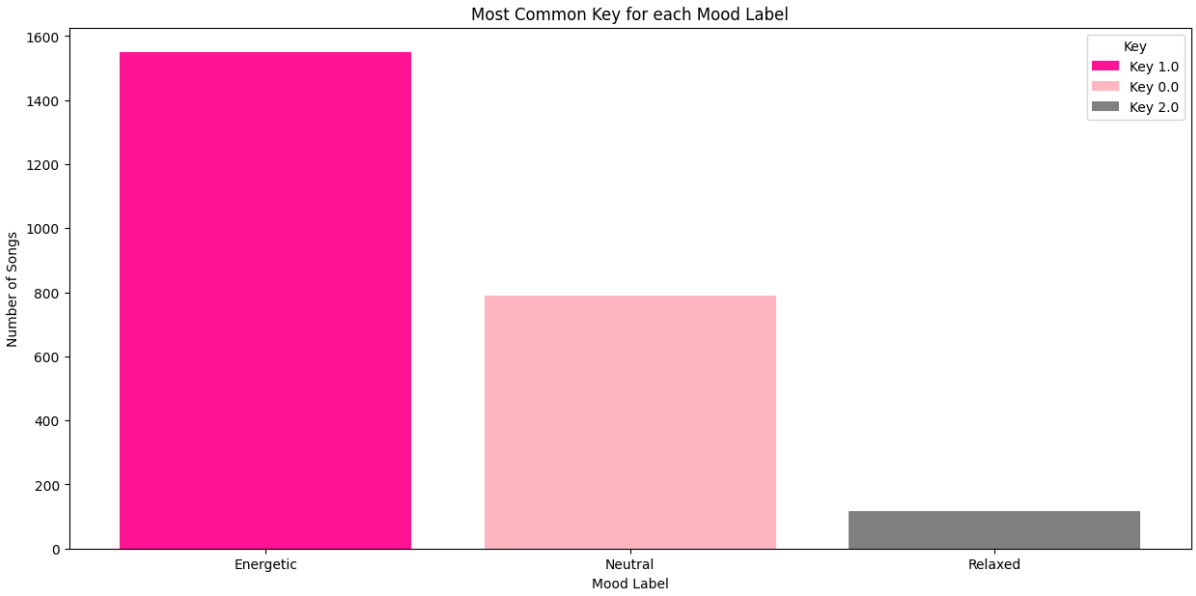
output:

Artist	Track	Danceability
BIA	CAN'T TOUCH THIS	0.975
Dave	Funky Friday	0.975
Timberland	Give It To Me	0.975
Xavier Wulf	Psycho Pass	0.973

Key Signatures and Song Mood

The visualization depicts the most prevalent musical "key" associated with songs for each mood category. In music, a "key" determines the group of pitches, or scale, that forms the basis of a music composition in classical, Western art, and Western pop music. It plays an integral role in setting the tone and mood of a piece. The "key" is typically represented by a number or a combination of numbers and letters in music datasets. From the chart, we observe that the predominant key for songs labeled as "Energetic" is 1.0. This insight suggests that songs in this specific key tend to have characteristics that are perceived as energetic, indicating a potential correlation between the musical key and the mood it evokes in listeners.

```
# Add labels and title
plt.xlabel('Mood Label')
plt.ylabel('Number of Songs')
plt.title('Most Common Key for each Mood Label')
plt.legend(title='Key')
# Display the plot
plt.show()
```



Appendix Figure 4 - Predominant Musical Key by Mood Category

B. google colab file link:

https://colab.research.google.com/drive/1iweA6kNWpg9PR-Mp2buFKmG3DL5ukYgE?usp=sharing