SPOTIFY DATASET ANALYSIS

****Author**: Mohan Krishna**

1. Introduction

This project provides a detailed analysis of the popular Spotify tracks dataset, with the aim of exploring relationships between various musical features and predicting the popularity of songs using different machine learning models.

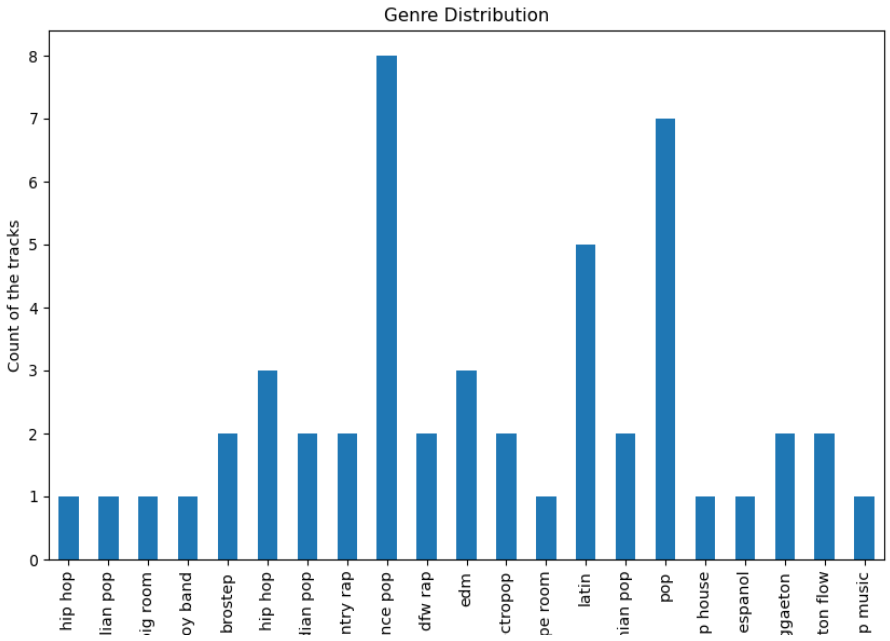
2. Data Preprocessing

Before diving into the analysis, we cleaned the dataset by removing unnecessary columns and renaming variables for clarity.

3. Exploratory Data Analysis

3.1 Genre Distribution

Chart: Bar plot representing the distribution of tracks across different genres.

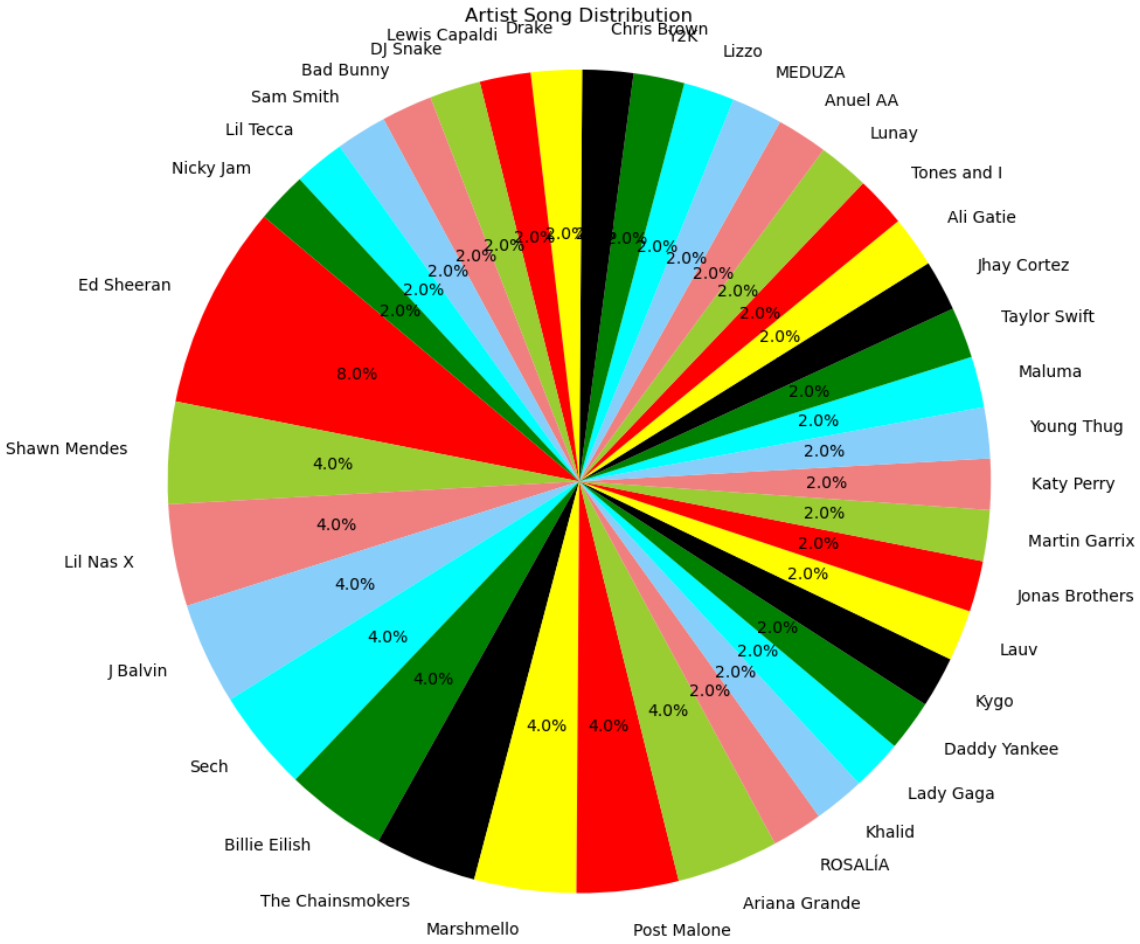


Explanation: This bar chart shows the most common genres in the dataset. Genres such as dance pop, pop, and Latin are the most frequent.

Conclusion: Dance pop and pop are the most prevalent genres.

3.2 Artist Song Distribution

Chart: Pie chart showing the distribution of songs by different artists.

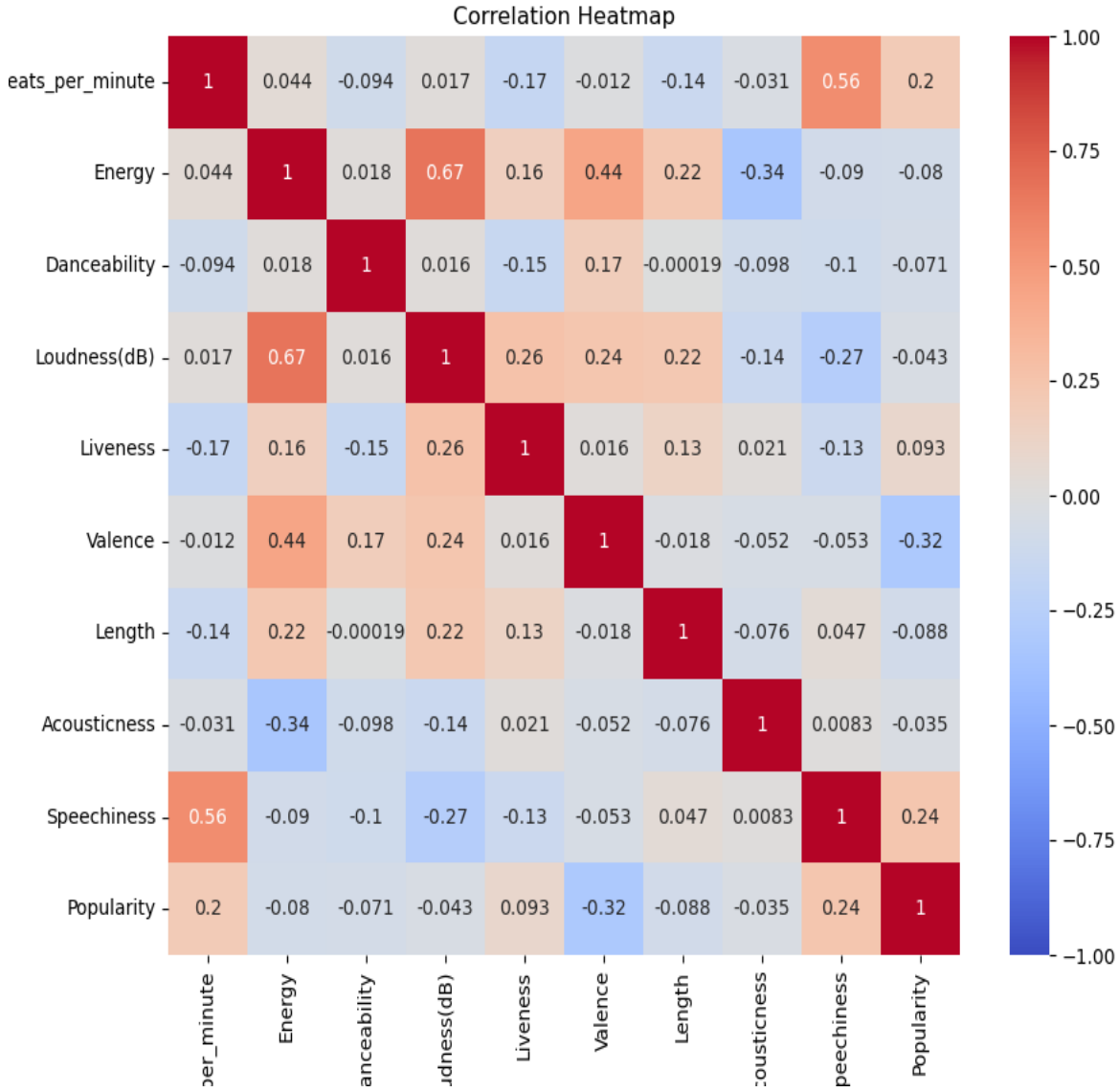


Explanation: This pie chart highlights which artists have the highest number of songs in the dataset.

Conclusion: Artists like Ed Sheeran and Post Malone have a larger presence in the dataset.

3.3 Correlation Heatmap

Chart: Heatmap illustrating the correlation between numerical features.



Explanation: The heatmap helps visualize relationships between variables such as Energy, Danceability, and Popularity.

Conclusion: There is a notable positive correlation between Energy and Loudness.

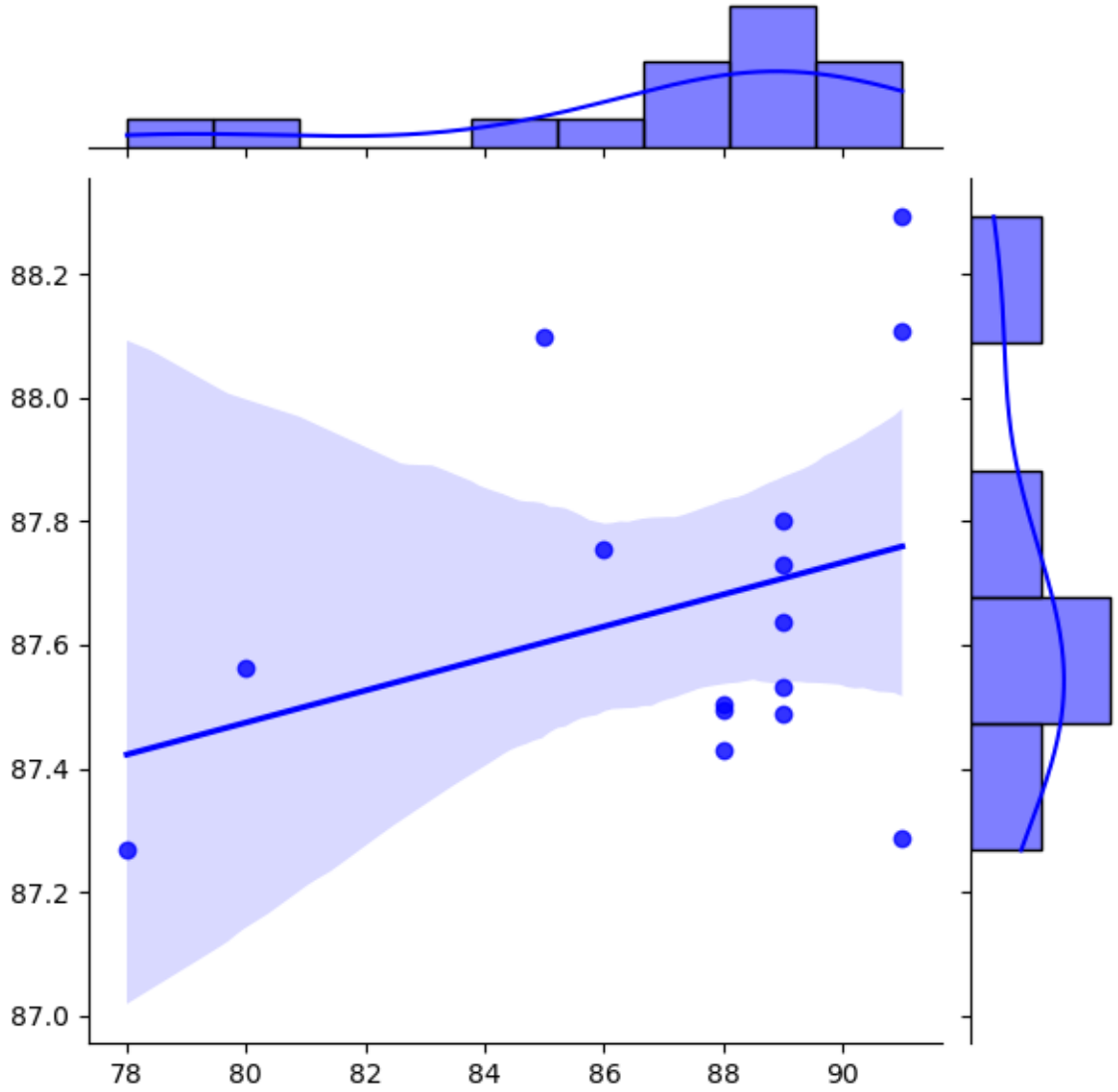
4. Predictive Modeling

4.1 Linear Regression

Model: A linear regression model was used to predict the popularity of tracks based on features like Energy, Danceability, and Loudness.
Conclusion: The model showed moderate prediction accuracy with a mean squared error of 13.02.

4.2 Naive Bayes

Model: A Gaussian Naive Bayes classifier was implemented to classify tracks into popularity categories.



Conclusion: The Naive Bayes model had lower accuracy compared to linear regression, suggesting that linear models perform better on this dataset.

5. Conclusion

This analysis provided insights into how different musical features relate to each other and how they impact the popularity of songs. Linear regression showed better performance in predicting popularity.