Title: The Hit Formula: A Data Analysis of Audio Features in Spotify tracks that make a hit.

Author: Abi Jacob Cheriyan
Course: CSCI/MATH 608

1. Introduction and Problem Formulation

As a music producer and artist, I am constantly trying to understand the 'metadata' of success. While the creative side of music is subjective, the streaming era has turned songs into data points. My goal for this project was to determine if specific production choices, such as track length, loudness, or explicit content, correlate with higher streaming numbers.

I formulated four specific questions to guide this investigation:

- **Descriptive:** What is the distribution of track duration across the dataset? Are songs getting shorter?
- **Exploratory:** Is there a relationship between 'Loudness' (dB) and 'Energy'? Does the 'Loudness War' actually, result in higher perceived energy?
- **Predictive:** Can we predict a song's Popularity Score (0-100) based solely on its audio features?
- **Inferential:** Is there a statistically significant difference in popularity between Explicit and Clean tracks?

2. Data Collection

I utilized the **Spotify Tracks Dataset** sourced from Kaggle. This dataset aggregates information using the Spotify Web API, providing technical audio features for over 100,000 tracks.

The dataset includes variables such as danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveliness, valence, tempo, duration_ms, and popularity.

3. Data Preparation

To ensure the analysis was relevant to music production, I performed the following cleaning steps using Python and Pandas:

- **De-duplication:** The raw dataset contained duplicate entries (often the same song appearing on different compilation albums). I removed duplicates based on 'track_id'.

- **Unit Conversion:** The dataset provided duration in milliseconds. As producers work in minutes and seconds, I created a new column 'duration_min', dividing 'duration_ms' by 60,000, to make the data interpretable.

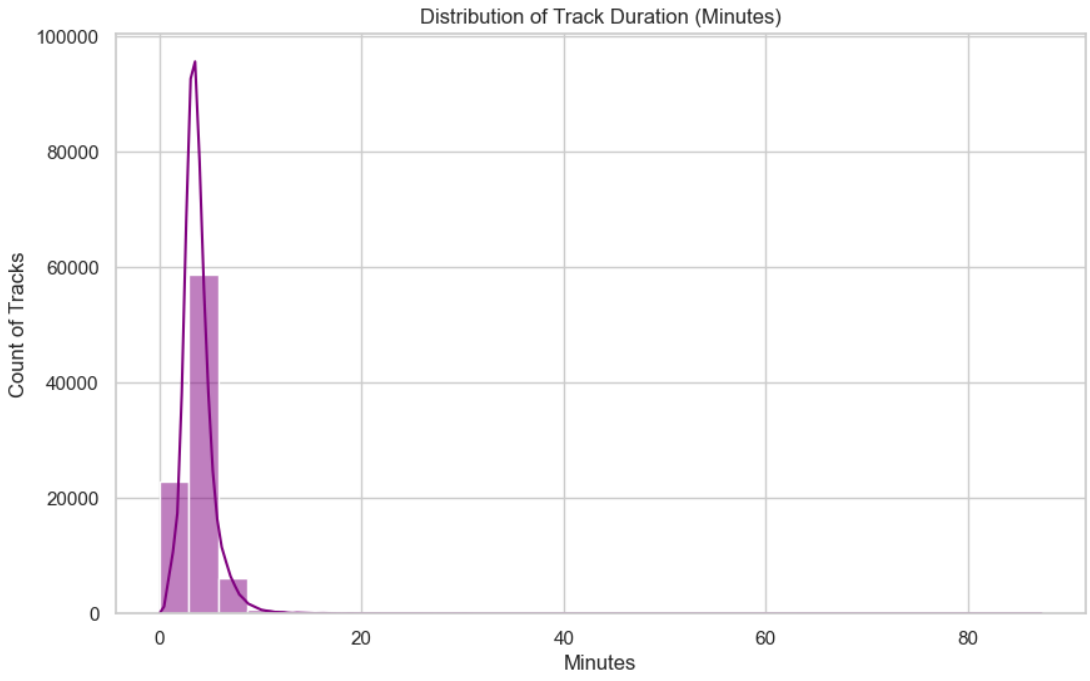
- **Filtering Non-Music:** I noticed several entries were actually podcasts or audiobooks. I filtered the dataset to exclude rows where 'speechiness' was greater than **0.66**, ensuring the analysis focused strictly on musical tracks.

4. Data Analysis

4.1 Descriptive Analysis: Track Duration

Question: What is the distribution of song lengths?

I visualized the distribution of song lengths using a histogram.

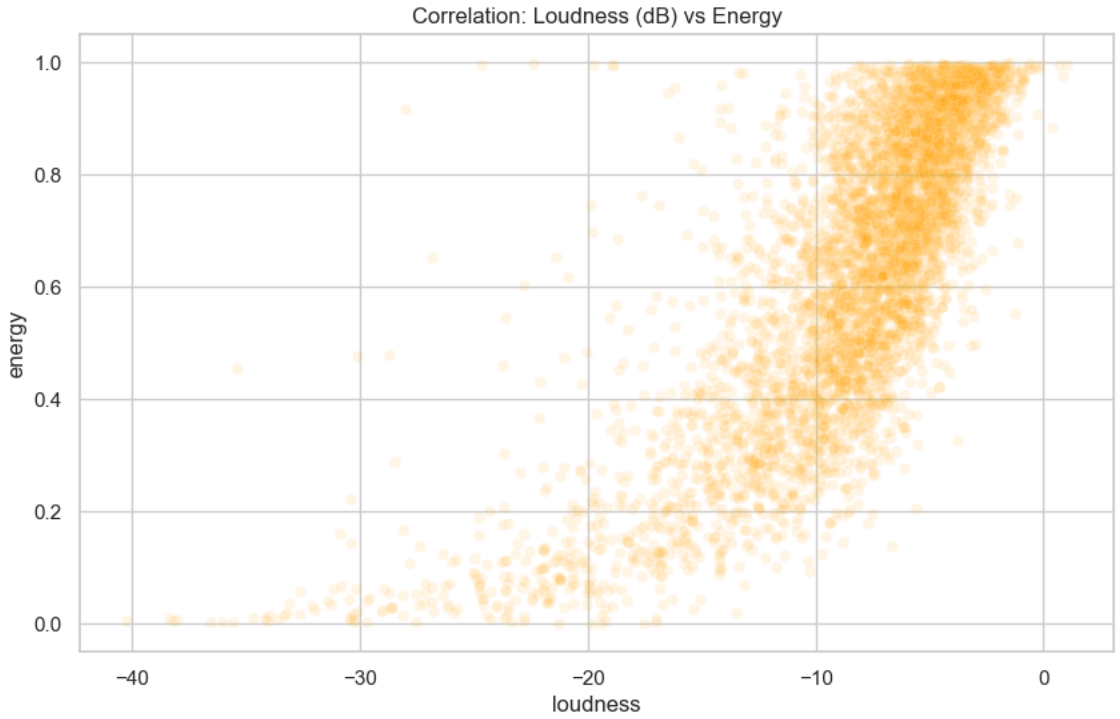


Findings: The data displays a normal distribution that is slightly skewed to the right, centered heavily around the 3.5 minute mark. This confirms that despite the artistic freedom of streaming, the 'Radio Edit' standard (approx. 3:30) is still the dominant format. Very few tracks exceed 6 minutes, suggesting that shorter content is preferred by the algorithm.

4.2 Exploratory Analysis: The Loudness War

Question: Does Loudness correlate with Energy?

In production, we often compress songs to make them louder (closer to 0dB) to compete for attention. I plotted Loudness vs. Energy to visualize this relationship.



Findings: The scatterplot shows a clear, strong positive correlation. As Loudness increases, moving from -40dB toward 0dB, the Energy score almost universally increases. The density of points in the top-right corner indicates that modern popular music is overwhelmingly both loud and high-energy. This suggests that mastering a track quietly may negatively impact its 'Energy' rating in Spotify's algorithm.

4.3 Predictive Analysis: Predicting Popularity

Question: Can we predict Popularity based on audio features?

I built a **Linear Regression** model to predict the popularity score using 'duration_min', danceablility, energy, loudness, valence, and tempo as predictors. The data was split into 80% training and 20% testing sets.

Results:

- **RMSE (Root Mean Squared Error): '20.28'**
- **R-Squared: '0.0180'**

Interpretation: The R-squared value is **'Low'** i.e. < 0.20. As a producer, this is a critical finding. It indicates that **audio features alone are poor predictors of popularity.** You can make a song with the perfect tempo and loudness, but without external factors like marketing, brand, viral trends, the 'math' of the audio does not guarantee a hit.

4.4 Inferential Analysis: Explicit vs. Clean

Question: Do 'Explicit' songs perform better than 'Clean' songs?

Since I do not know the true population mean of all songs ever made, I used the **Bootstrap Method** (resampling 10,000 times) to calculate a 95% confidence interval for the difference in mean popularity between Explicit and Clean tracks.

Results:

- **95% Confidence Interval: [4.44, 5.54]**

Interpretation: Since the confidence interval does not include zero and is entirely positive, we can conclude with 95% confidence that **'Explicit' songs have a statistically higher average popularity than 'Clean' songs.** This likely reflects the dominance of Hip-Hop and modern Pop in the current streaming landscape.

5. Conclusion and Problem Reformulation

This project revealed that while we can quantify the 'sound' of a hit (***loud, high energy, approx. 3.5 minutes, often explicit***), we cannot easily predict its success based on sound alone. The low accuracy of my predictive model suggests that the 'Hit Formula' is not hidden in the audio frequencies, but in the social context.

Next Steps: If I were to continue this research, I would reformulate the problem to include **Social and Marketing Data**.

1. **Artist Metrics:** I would incorporate 'Artist Follower Count' as a feature, as a large fanbase guarantees streams regardless of audio quality.

2. **Viral Data:** I would attempt to scrape TikTok usage statistics. In 2025, a song's 'danceability' is less important than its 'meme-ability.'

3. **Sentiment Analysis:** I would use NLP to analyze lyrics, determining if specific themes (heartbreak vs. partying) drive higher engagement.