ALY 6110
Data Management and Big Data

Module 5
Spotify and YouTube Data Analysis

Prof. Sohrob Milani

Submitted by
Khushi Doshi
Krutika Patel
Yash Tailor

2024-12-08

Introduction

Problem Statement

With the rapid growth of digital music streaming platforms like Spotify and YouTube, the music industry is shifting toward data-driven decision-making. Understanding the drivers of song popularity across platforms is crucial for artists, record labels, and marketers. By identifying patterns and trends in audio features and engagement metrics, stakeholders can optimize their strategies to enhance audience engagement, reach, and revenue generation.

This analysis addresses the question: **What factors influence song popularity across Spotify and YouTube, and how can these insights guide stakeholders to achieve higher engagement?**

Dataset Selection

The "Spotify and YouTube Dataset" from Kaggle was chosen for its comprehensive collection of over 20,000 songs, including:

- **Engagement Metrics:** Views, streams, likes, comments—indicators of user interaction.
- **Audio Features:** Danceability, energy, acousticness, tempo, and more—key elements influencing user preferences.
- **Song Metadata:** Artist, album type, release year—providing contextual information.

The dataset is ideal for this project as it enables a holistic exploration of song popularity by combining quantitative measures of audience engagement with qualitative audio characteristics.

Real-World Application

This analysis offers direct benefits for key stakeholders:

- **Artists and Record Labels**:
    - Optimize music production by identifying popular audio features like energy and danceability.
    - Target niche audiences, such as children, to achieve cross-platform success (e.g., Cocomelon's dominance in kid-friendly content).
- **Streaming Platforms**:
    - Improve user retention by curating playlists that reflect listener moods and preferences.
    - Increase replayability by focusing on high-energy tracks during workouts or parties.
- **Marketers**:
    - Design seasonal campaigns to leverage peak audience engagement times.

- Use viral marketing on social media platforms to amplify track reach and popularity.

Methodology

Steps Undertaken

1. **Data Cleaning**:
   - Missing values in audio features (e.g., danceability, energy) were imputed with the median to preserve statistical integrity.
   - Missing engagement metrics (views, likes, comments, streams) were imputed with zero and flagged using indicator columns.

```
> # Check for missing values
> sapply(spotify_youtube_dataset, function(x) sum(is.na(x)))
                x           artist      url_spotify            track            album 
                0                0                0                0                0 
       album_type              uri     danceability           energy              key 
                0                0                2                2                2 
         loudness      speechiness     acousticness instrumentalness         liveness 
                2                2                2                2                2 
          valence            tempo      duration_ms      url_youtube            title 
                2                2                2                0                0 
          channel            views            likes         comments      description 
                0              470              541              569                0 
         licensed   official_video           stream 
                0                0              576 
```

Figure 1: Missing Values by Variable

2. **Feature Engineering**:
   - **Engagement Score**: Summed views, likes, comments, and streams to create a composite engagement metric.
   - **Popularity Classification**: Created a binary target variable is_popular based on the median engagement score.
3. **Exploratory Data Analysis (EDA)**:
   - Generated visualizations to explore relationships among features and uncover patterns in song popularity.
4. **Modeling**:
   - Trained Logistic Regression, SVM (linear, radial, polynomial kernels), and Random Forest models.
   - Used Principal Component Analysis (PCA) for dimensionality reduction.

5. **Evaluation Metrics**:
    - Compared models based on accuracy, precision, recall, and F1 score.

Visualizations and Insights

Missing Values by Variable

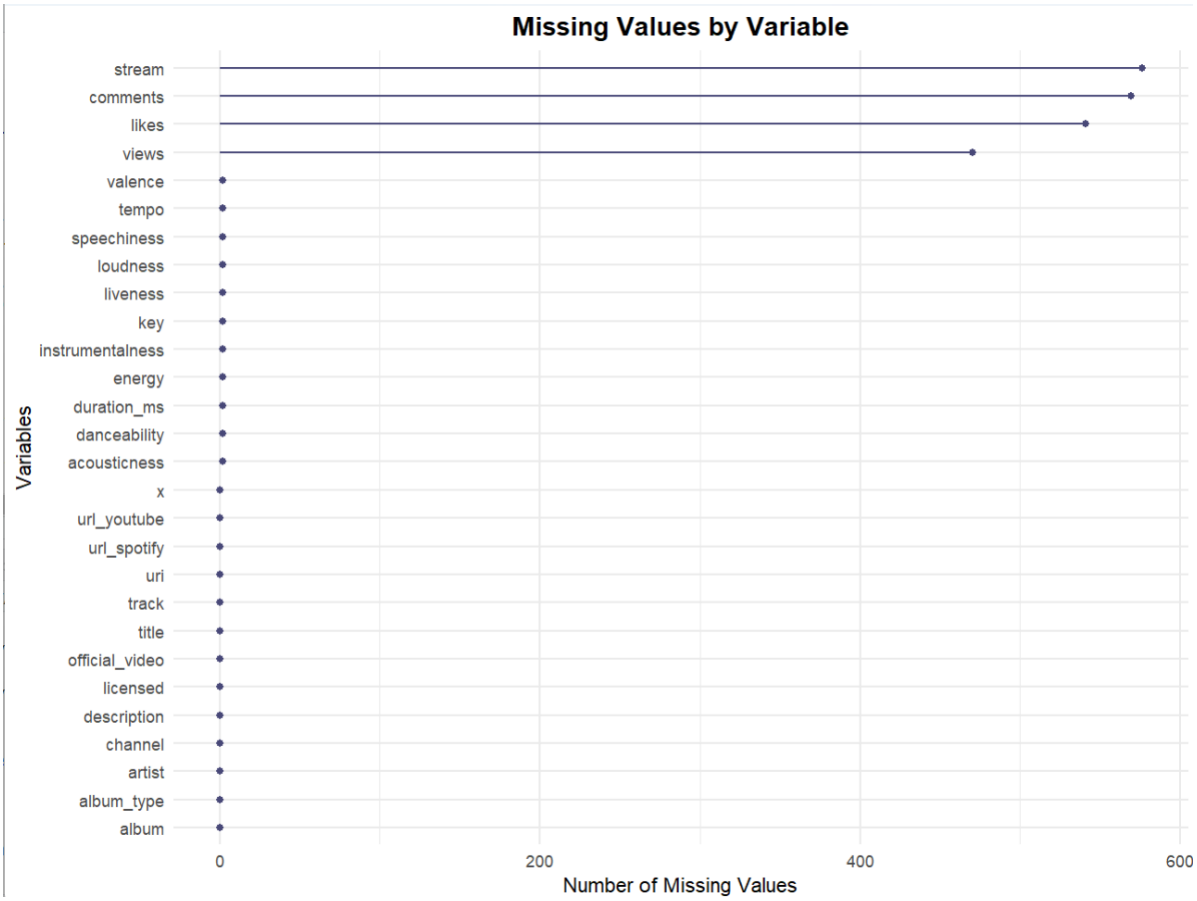

Figure 2: Missing Values by Variable Highlights the variables with missing data, particularly engagement metrics like views, likes, and streams, which were imputed.

Insights:

- Engagement metrics (views, likes, comments, streams) had significant missing values, requiring imputation.
- This step ensured data completeness for robust analysis.

Correlation Matrix

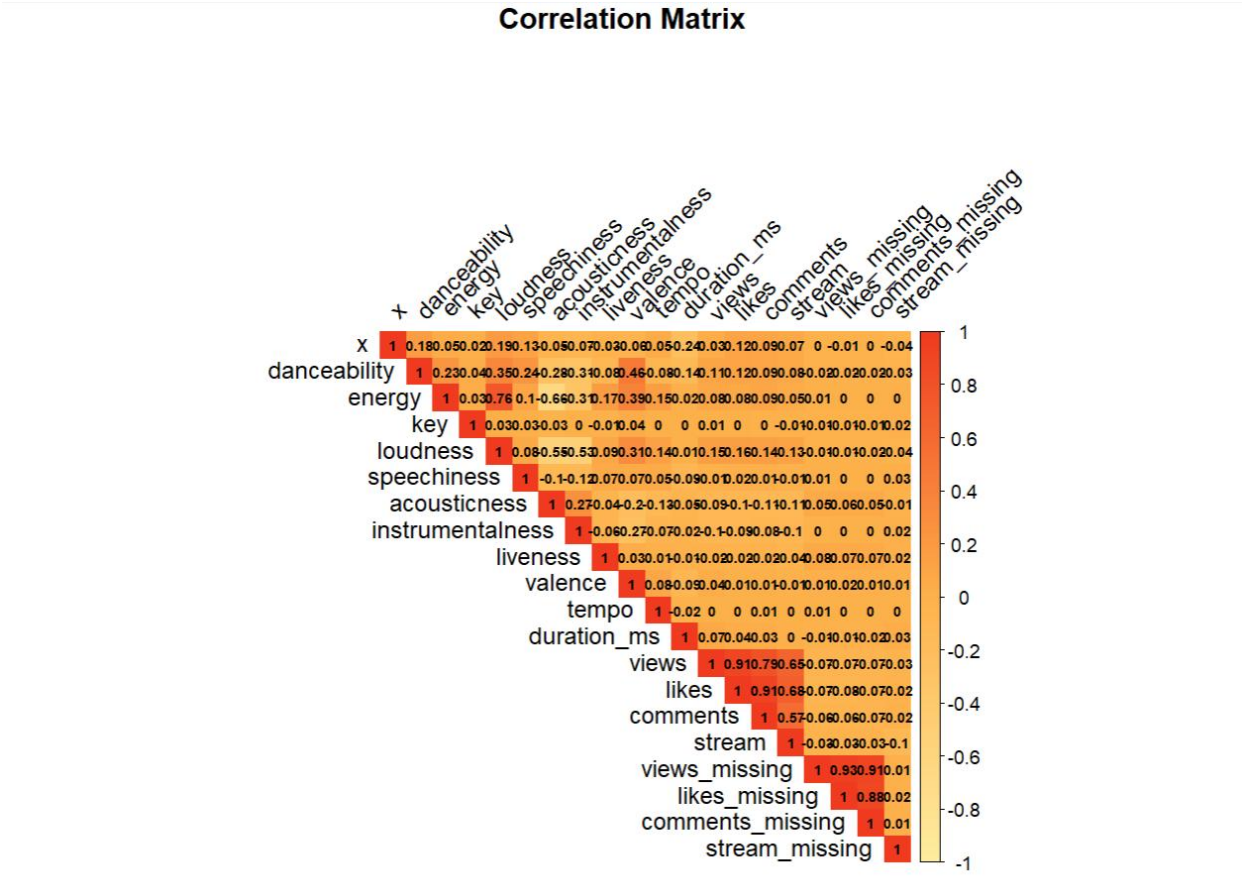

Figure 3: Correlation Matrix Displays the pairwise correlation between numerical features in the dataset. Strong correlations are observed among engagement metrics such as views, likes, and streams.

Insights:

- Strong positive correlations were observed among views, likes, and streams, justifying their combined use in the engagement_score.
- Weak correlations between audio features and engagement metrics suggest other factors influence popularity.

Density Plot of Danceability

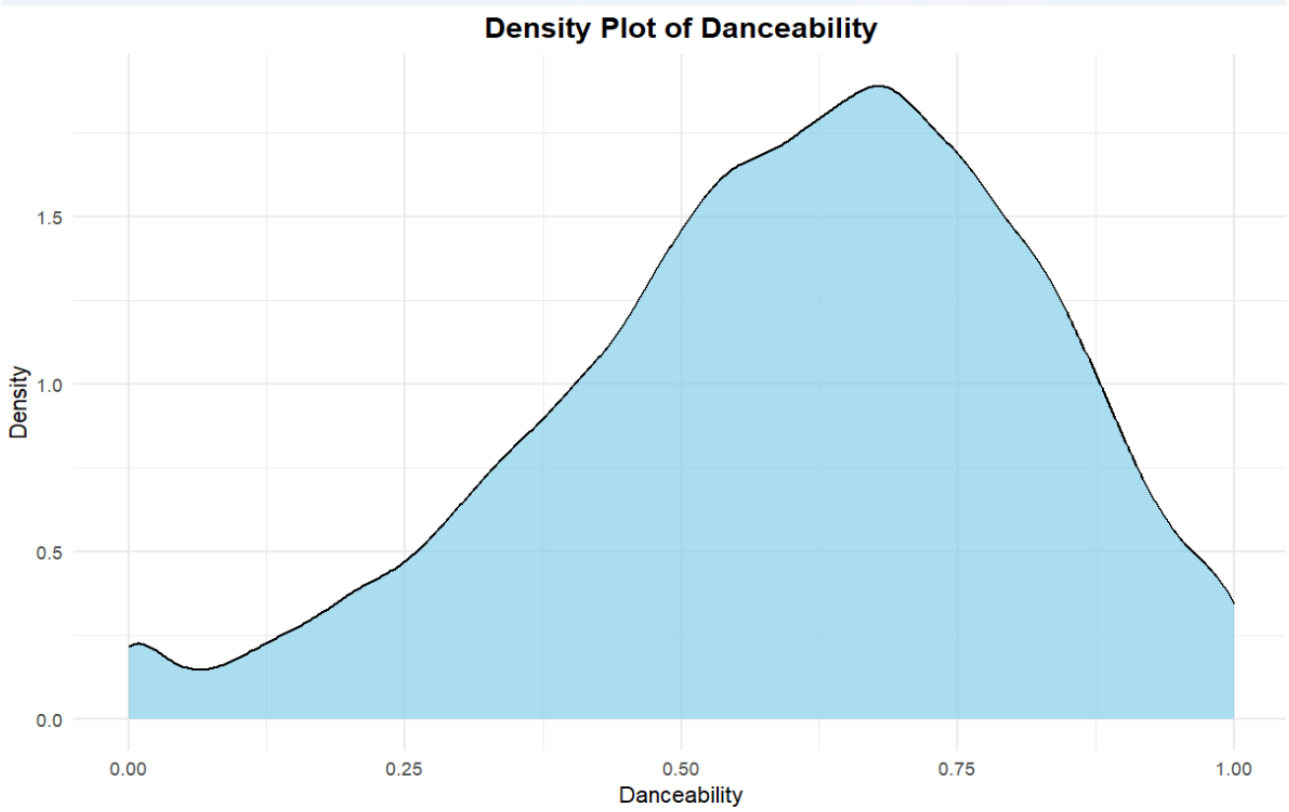

Figure 4: Density Plot of Danceability Shows the distribution of danceability scores, indicating that most songs have moderate-to-high danceability, preferred by audiences.

Insights:

- Most songs have moderate to high danceability (0.5–0.8), reflecting audience preferences for rhythmic tracks.

Density Plot of Energy

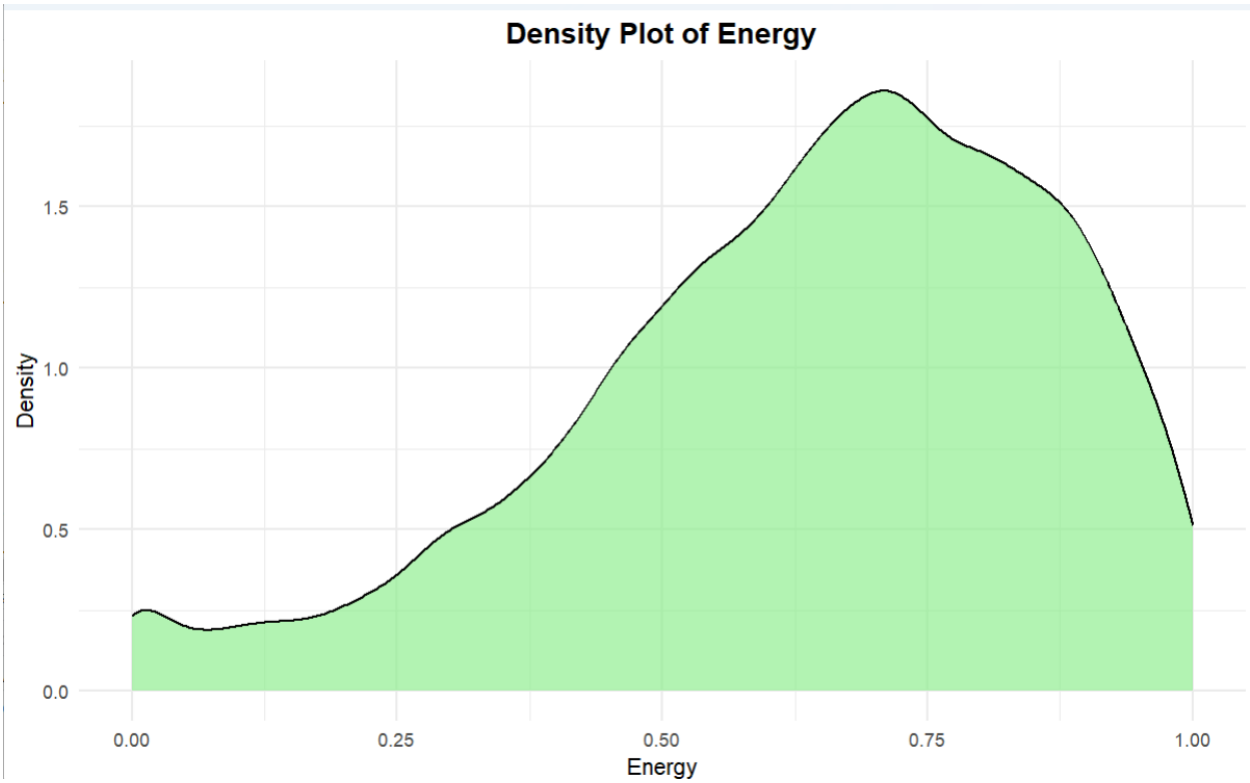

Figure 5: Density Plot of Energy Visualizes the energy levels of songs, with the majority having high energy, reflecting listener interest in upbeat tracks.

Insights:

- High-energy tracks dominate, aligning with the popularity of upbeat genres like pop and dance music.

Energy vs. Danceability (Colored by Views)

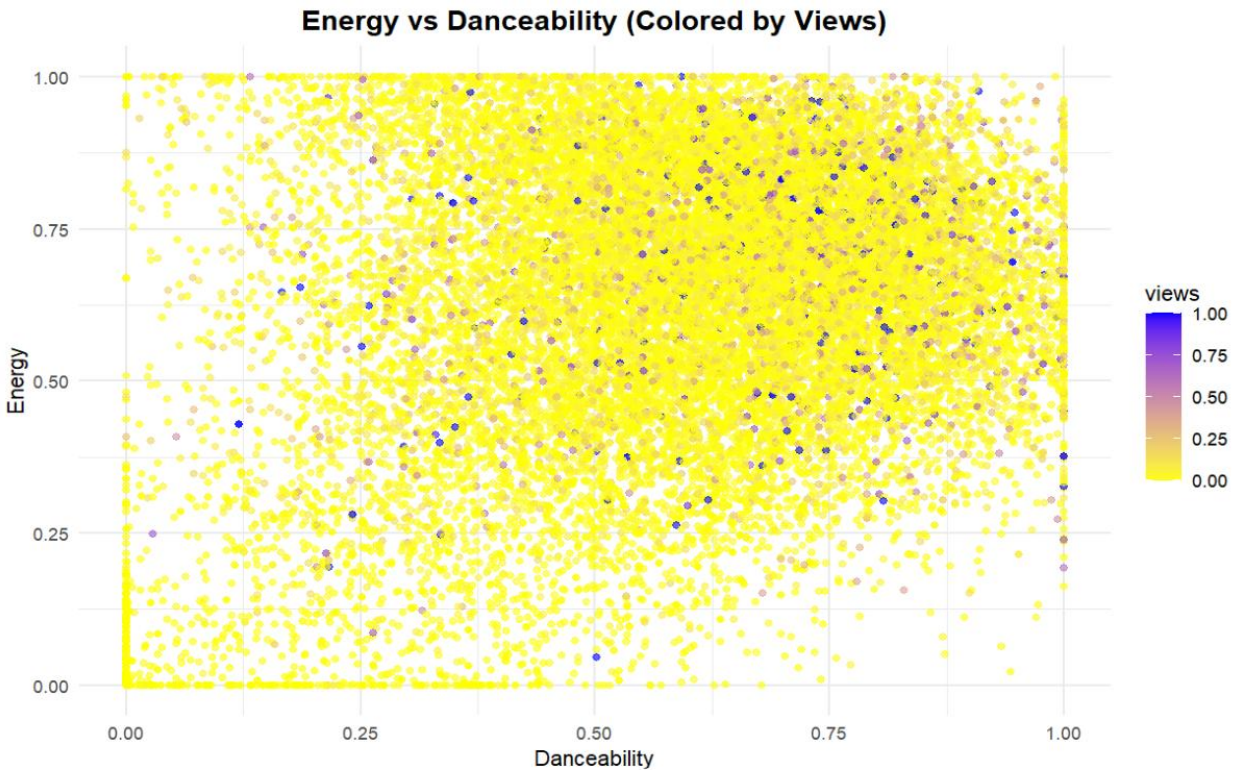

Figure 6: Energy vs Danceability (Coloured by Views) Explores the relationship between energy and danceability, with the colour intensity representing views. Popular songs balance high energy and danceability.

Insights:

- Songs with balanced energy and danceability tend to have higher views, indicating their combined importance for engagement.

Scatter Plot of Views vs. Likes

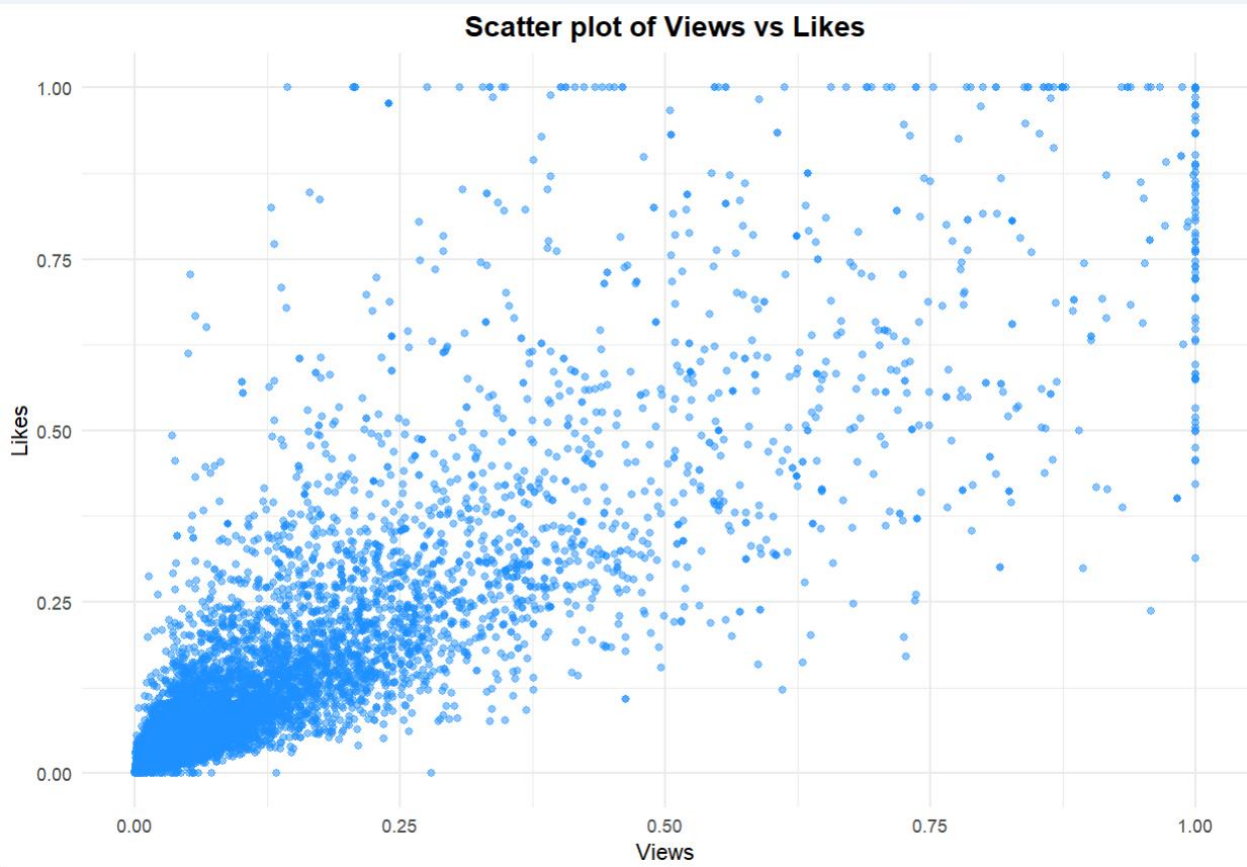

Figure 7: Scatter Plot of Views vs Likes Displays a positive correlation between views and likes, indicating that higher visibility leads to increased audience appreciation.

Insights:

- A strong positive relationship between views and likes confirms that visibility directly translates to audience appreciation.

Top 10 Artists by Total Views

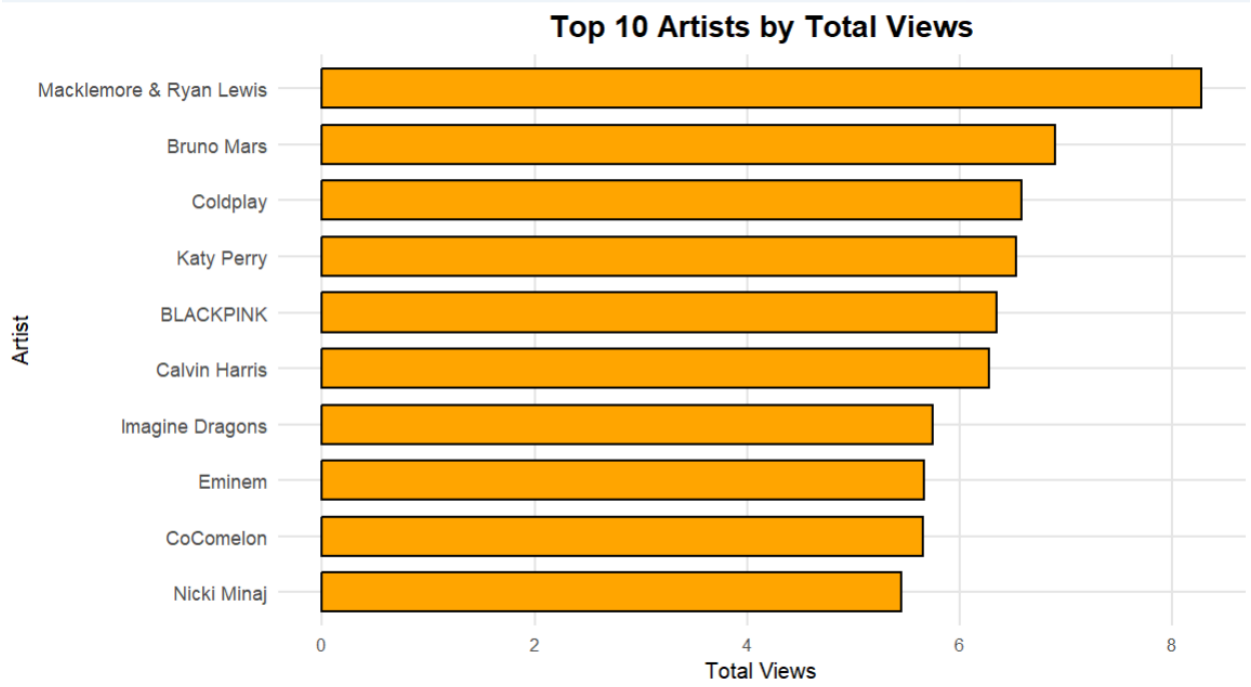

Figure 8: Top 10 Artists by Total Views Bar chart ranking the top 10 artists based on their total views. Macklemore & Ryan Lewis, Bruno Mars, and Coldplay lead the list.

Insights:

- Macklemore & Ryan Lewis, Bruno Mars, and Coldplay dominate views, showcasing the appeal of diverse music styles.

Frequency of Album Types

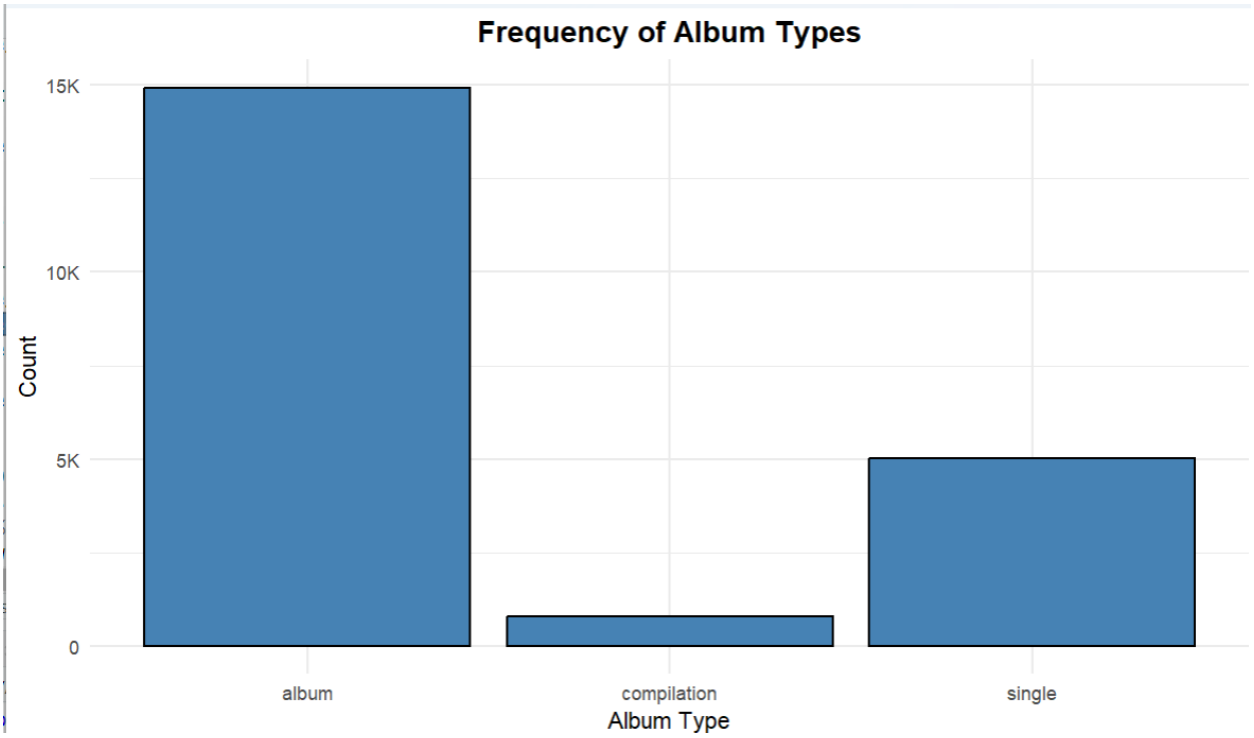

Figure 9: Frequency of Album Types A bar chart showing the count of album types. Albums dominate the dataset, followed by singles and compilations.

Insights

- This bar chart displays the frequency of different album types. Full albums are the most common, followed by singles and compilations. This distribution reflects how artists prefer releasing complete albums over standalone tracks.

Scatter Plot: Views vs Likes (Colored by Streams)

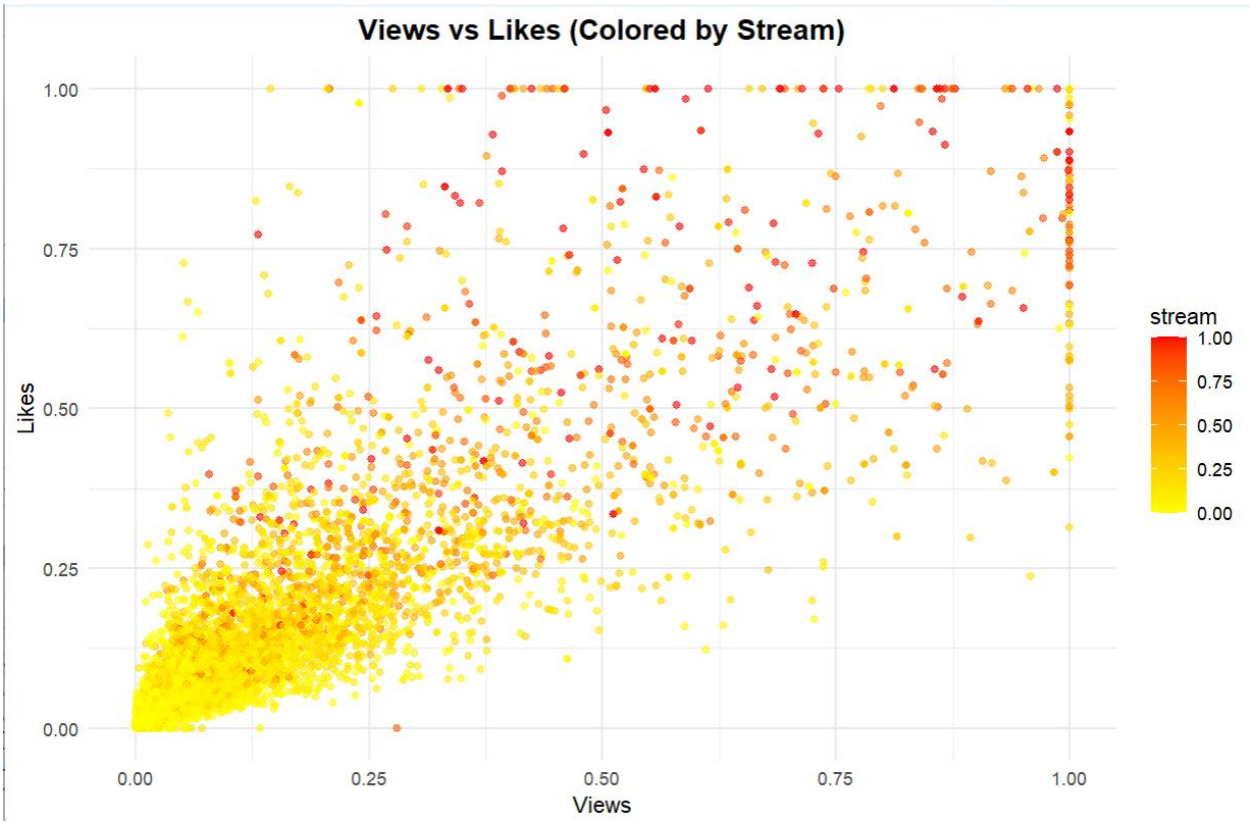

Figure 10: Views vs Likes (Coloured by Stream) Examines the relationship between views and likes, with stream counts represented by colour intensity. Higher streams correlate with higher views and likes.

Insights:

- This scatter plot investigates the relationship between views and likes, with color intensity representing streams. Higher streams are associated with higher views and likes, suggesting a strong connection between these metrics.

Key Insights from the Dashboard

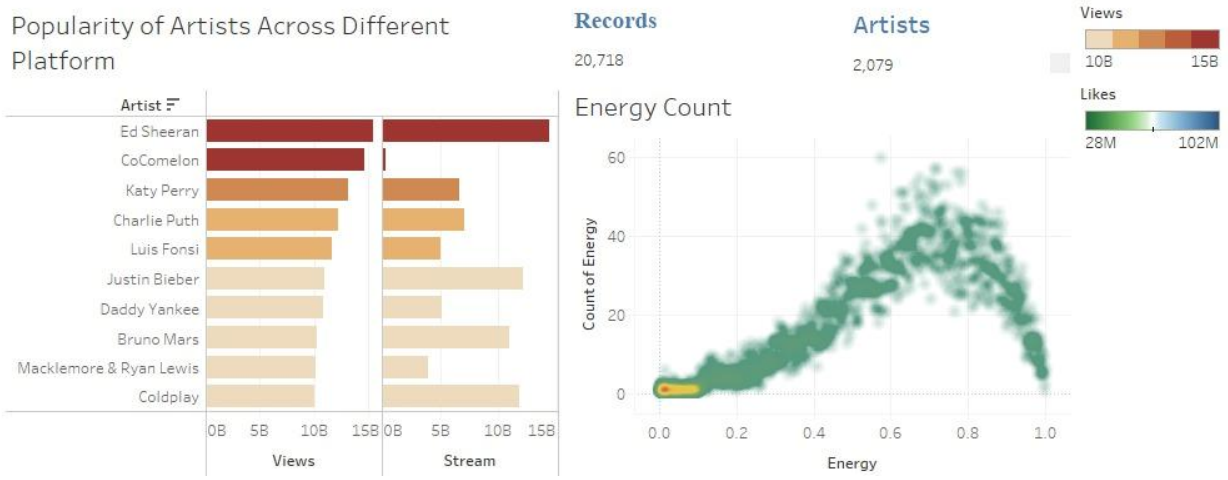

Figure 11: Top Artists by Platform Popularity & Energy and Engagement Patterns

Top Artists by Platform Popularity

A bar chart showcasing the **Top 10 Artists** by total YouTube views and Spotify streams.

Key Observations:

- Ed Sheeran dominates across both platforms with over 15 billion views/streams.
- Other top artists include Cocomelon, Katy Perry, and Charlie Puth.

High-performing artists have a strong presence across both platforms, reflecting their global appeal and consistent audience engagement.

Energy and Engagement Patterns

Energy is a measure of the intensity and activity level of a track. It quantifies how dynamic or powerful a song feels based on its sonic attributes.

How It's Calculated:

Energy is derived from several audio features, including:

1. **Dynamics**: Variation in loudness throughout the track.
2. **Tempo**: Speed of the beat (measured in beats per minute, BPM).
3. **Loudness**: Overall volume level of the track.

4. **Frequency Range**: Presence of high-energy frequencies like percussion.

Tracks with higher energy values are typically loud, fast, and have dense instrumentation, while low-energy tracks tend to be soft, slow, and calm.

Why It's Important:

1. **Listener Preferences**:
    - High-energy songs are often chosen for activities like workouts or parties, driving higher engagement.
    - Low-energy songs appeal to listeners during relaxing or contemplative moments.
2. **Cross-Platform Success**:
    - Energy directly impacts a song's replayability and emotional resonance, which are critical for popularity on platforms like Spotify and YouTube.
3. **Playlist Curation**:
    - Streaming platforms use energy metrics to create mood-based playlists (e.g., "High-Energy Workout" or "Relaxing Chill").
4. **Genre Classification**:
    - High-energy tracks are common in genres like EDM, rock, and hip-hop.
    - Low-energy tracks dominate in classical and acoustic genres.

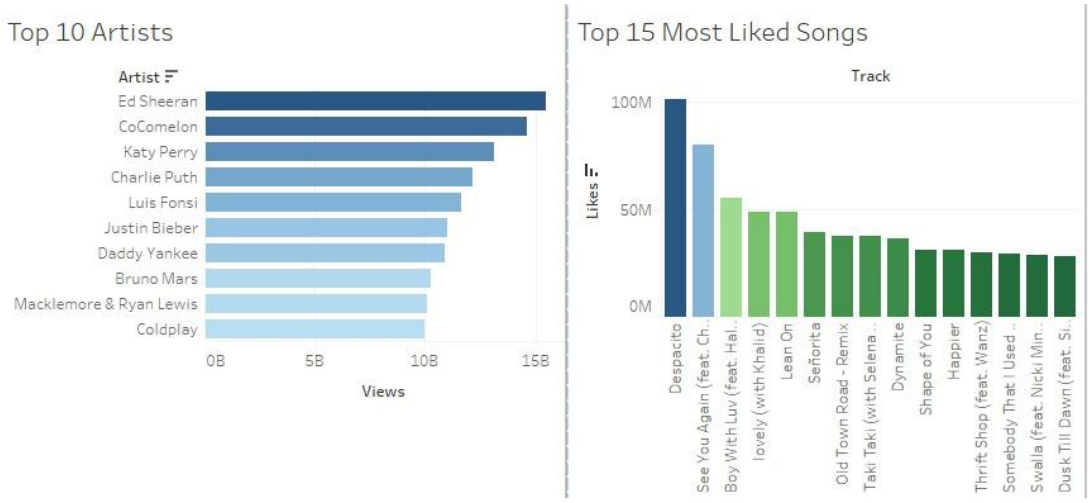

Figure 12: Top Artists by Views & Most Liked Songs on YouTube

Top Artists by Views

A bar chart showcasing the **Top 10 Artists** ranked by total YouTube views.

Key Observations:

- **Ed Sheeran** is the leading artist with over 15 billion views, showcasing his global appeal and consistent fan engagement.
- **Cocomelon**, a children's music content creator, ranks second, reflecting the immense popularity of kid-focused content.
- Other artists like **Katy Perry**, **Charlie Puth**, and **Justin Bieber** also feature prominently, driven by their diverse musical styles and wide-reaching fanbases.

High-performing artists across YouTube are those who consistently release engaging content that resonates with global audiences. Niche content creators, like Cocomelon, demonstrate the power of targeted audience engagement in driving massive views.

Most Liked Songs on YouTube

A bar chart visualizing the **Top 15 Liked Songs** on YouTube.

Key Observations:

- **"Despacito"** leads with over 100 million likes, reflecting its status as a global phenomenon with wide cultural appeal.
- **"See You Again"** follows closely, with its emotional theme and association with the "Fast & Furious" franchise contributing to its success.
- Songs like **"Shape of You"**, **"Old Town Road Remix"**, and **"Boy with Luv"** attract significant likes, showcasing their ability to capture audience attention through catchy beats and unique styles.

Songs with emotional resonance, upbeat tones, and cross-cultural appeal tend to attract higher user engagement on YouTube. International artists like BTS demonstrate the increasing global influence of non-English music in driving massive engagement.

Understanding Key Metrics

1. **Energy**:
    - **Definition**: Measures the intensity and activity level of a track. High-energy tracks are loud, fast, and rhythmically intense.
    - **Importance**:
        - High-energy tracks dominate in genres like pop, rock, and EDM, driving higher engagement.

        - Platforms use energy metrics to curate playlists such as "High-Energy Workout" or "Relaxing Chill."

2. **Danceability**:
    - **Definition**: Indicates how suitable a track is for dancing, based on rhythm stability and beat strength.
    - **Importance**:
        - Tracks with moderate-to-high danceability appeal to larger audiences and drive higher replayability.
        - Critical for party and club playlists.
3. **Acousticness**:
    - **Definition**: Reflects the likelihood of a track being acoustic.
    - **Importance**:
        - High-acousticness tracks resonate well during relaxing or introspective moments.
        - Dominates genres like classical and acoustic.
4. **Tempo**:
    - **Definition**: The speed of a track, measured in beats per minute (BPM).
    - **Importance**:
        - Faster tempos align with energetic activities like workouts.
        - Slower tempos are ideal for calming and reflective moods.

Machine Learning Models and Comparison

To predict song popularity and identify the factors influencing it, multiple machine learning models were implemented. Below is a detailed explanation of each model, its approach, and how it performed in this analysis.

Logistic Regression with Lasso Regularization

Overview:

- **Logistic Regression** is a statistical model used for binary classification. In this case, it predicts whether a song is popular (is_popular = 1) or not (is_popular = 0).
- **Lasso Regularization (L1)** penalizes the magnitude of coefficients, effectively performing feature selection by shrinking less important features to zero.

Steps Taken:

1. Features were standardized to ensure comparability.
2. The best lambda (regularization strength) was determined using cross-validation.
3. The final model was trained using the optimal lambda.

Performance:

```
> cat("Final Logistic Regression Model Performance:\n")
Final Logistic Regression Model Performance:
> cat("Accuracy:", final_metrics[1], "\n")
Accuracy: 0.9946886
> cat("Precision:", final_metrics[2], "\n")
Precision: 0.9942113
> cat("Recall:", final_metrics[3], "\n")
Recall: 0.9951714
> cat("F1 Score:", final_metrics[4], "\n")
F1 Score: 0.9946911
```

```
> # ----------------------- Logistic Regression Performance Metrics ----------------------
> test_pred_class <- as.factor(ifelse(test_pred_prob > 0.5, 1, 0))
> logistic_metrics <- calc_metrics(y_test, test_pred_class)
> cat("Logistic Regression Metrics:\n", logistic_metrics, "\n")
Logistic Regression Metrics:
 0.9946886 0.9942113 0.9951714 0.9946911
```

Insights:

- Logistic Regression emerged as the best-performing model, achieving the highest accuracy and balanced metrics.
- Its ability to select relevant features (via Lasso) enhanced its predictive power.

Support Vector Machines (SVM)

Overview:

- **SVM** is a supervised learning algorithm that finds the hyperplane that best separates data points into classes. Different kernels were used to capture linear and non-linear relationships:
    - **Linear Kernel**: Assumes linear relationships between features.
    - **Radial Kernel**: Captures non-linear patterns using a radial basis function.
    - **Polynomial Kernel**: Models polynomial relationships between features.

Steps Taken:

1. Data was standardized to meet SVM's requirements.

2. Hyperparameters (C, gamma, and degree for polynomial kernel) were optimized using cross-validation.
3. Models were trained using the optimal hyperparameters.

Performance:

```
> cat("Tuned Linear Kernel SVM Accuracy:", linear_accuracy, "\n")
Tuned Linear Kernel SVM Accuracy: 0.9939643

> cat("Tuned Radial Kernel SVM Accuracy:", radial_accuracy, "\n")
Tuned Radial Kernel SVM Accuracy: 0.9707871

> cat("Tuned Polynomial Kernel SVM Accuracy:", poly_accuracy, "\n")
Tuned Polynomial Kernel SVM Accuracy: 0.9780299
```

```
> cat("Linear Kernel SVM Metrics:\n", linear_metrics, "\n")
Linear Kernel SVM Metrics:
 0.9939643 0.9927746 0.9951714 0.9939715
> # Radial SVM Metrics
> y_pred_radial <- predict(classifier_radial, newdata = test_data[-ncol(test_data)])
> radial_metrics <- calc_metrics(test_data$is_popular, y_pred_radial)
> cat("Radial Kernel SVM Metrics:\n", radial_metrics, "\n")
Radial Kernel SVM Metrics:
 0.9707871 0.9568885 0.9859971 0.9712247
> # Polynomial SVM Metrics
> y_pred_poly <- predict(classifier_poly, newdata = test_data[-ncol(test_data)])
> poly_metrics <- calc_metrics(test_data$is_popular, y_pred_poly)
> cat("Polynomial Kernel SVM Metrics:\n", poly_metrics, "\n")
Polynomial Kernel SVM Metrics:
 0.9780299 0.9700855 0.98648 0.978214
```

Insights:

- **Linear Kernel SVM** closely matched Logistic Regression in performance, making it a viable alternative.
- The **Radial Kernel** captured non-linear relationships but underperformed compared to the linear kernel due to overfitting on high-dimensional data.
- **Polynomial Kernel** provided slightly better performance than radial but fell short of logistic regression.

Random Forest

Overview:

- **Random Forest** is an ensemble learning method that builds multiple decision trees and aggregates their predictions for improved accuracy and robustness.
- It excels at handling non-linear relationships and interactions between features.

Steps Taken:

1. Cross-validation was used to tune the mtry hyperparameter (number of features considered at each split).
2. A forest of 50 trees was constructed for prediction.

Performance:

```
> # ------------------------ Random Forest Performance Metrics --------------
> rf_metrics <- calc_metrics(test_data$is_popular, rf_predictions)
> cat("Random Forest Metrics:\n", rf_metrics, "\n")
Random Forest Metrics:
 0.9669242 0.9493494 0.98648 0.9675586
```

Insights:

- Random Forest performed well in recall, indicating its ability to correctly identify popular songs.
- However, its overall accuracy and precision were slightly lower compared to Logistic Regression and SVM with a linear kernel.

Model Comparison

```
> print("Model Comparison Summary:")
[1] "Model Comparison Summary:"
> print(model_comparison)
                Model  Accuracy Precision    Recall  F1_Score
1        SVM - Linear 0.9939643 0.9927746 0.9951714 0.9939715
2        SVM - Radial 0.9707871 0.9568885 0.9859971 0.9712247
3    SVM - Polynomial 0.9780299 0.9700855 0.9864800 0.9782140
4 Logistic Regression 0.9946886 0.9942113 0.9951714 0.9946911
5       Random Forest 0.9669242 0.9493494 0.9864800 0.9675586
```

The table below summarizes the performance of all models:

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9947	0.9942	0.9952	0.9947
SVM - Linear	0.9939	0.9928	0.9952	0.9940
SVM - Radial	0.9708	0.9569	0.9860	0.9712
SVM - Polynomial	0.9780	0.9701	0.9865	0.9782
Random Forest	0.9669	0.9493	0.9865	0.9676

Key Observations:

1. **Logistic Regression with Lasso**:
    - The best model with the highest accuracy, precision, recall, and F1 score.
    - Simple, interpretable, and computationally efficient.
2. **SVM with Linear Kernel**:
    - Performed nearly as well as logistic regression.
    - A good alternative if a robust, non-probabilistic classifier is needed.
3. **Random Forest**:
    - Performed well in terms of recall, making it suitable for scenarios where minimizing false negatives is critical.
4. **SVM with Radial and Polynomial Kernels**:
    - While capturing non-linear relationships, they underperformed relative to linear methods due to overfitting and the dataset's characteristics.

Model Limitations

1. **Logistic Regression**:
    - May oversimplify complex relationships between features.
    - Performs best on datasets with linear separability, limiting its scope for highly non-linear data.
2. **SVM**:
    - Computationally expensive for large datasets, especially with non-linear kernels.
    - Radial and polynomial kernels tend to overfit when applied to high-dimensional datasets.
3. **Random Forest**:
    - Less interpretable compared to simpler models like Logistic Regression.
    - Performance can degrade when applied to datasets with highly correlated features.

Recommendations

1. **For Artists and Record Labels**:
    - **Optimize Audio Features**: Create tracks with high energy, moderate-to-high danceability, and emotional themes to maximize listener engagement.

- **Leverage Seasonal Trends**: Release songs during peak listening periods (e.g., summer for upbeat music or holiday seasons for emotional tracks).
- **Experiment with Niche Content**: As evidenced by Cocomelon, explore targeted audiences such as children or cultural niches to expand reach.

2. **For Streaming Platforms**:
   - **Refine Playlist Algorithms**: Incorporate energy, engagement metrics, and genre preferences into playlist curation to enhance user retention.
   - **Dynamic Recommendations**: Offer personalized recommendations based on time of day (e.g., high-energy tracks for mornings and relaxing tracks for evenings).
   - **Engagement Monitoring**: Use dashboards to track the performance of recommended tracks and refine strategies in real time.
3. **For Marketers**:
   - **Use Data-Driven Campaigns**: Promote tracks with high engagement metrics (likes, views, streams) to achieve maximum ROI on marketing efforts.
   - **Collaborate with Artists**: Sponsor high-performing artists and tracks for cross-promotion opportunities.
   - **Target Social Media**: Use snippets of high-energy tracks in viral marketing campaigns on platforms like TikTok and Instagram.
4. **For Researchers and Developers**:
   - **Incorporate Demographics**: Future datasets should include listener demographics to analyze trends by region, age, and gender.
   - **Apply Advanced Models**: Explore ensemble models like Gradient Boosting or XGBoost for more nuanced predictions of song popularity.

Future Work

1. **Incorporate Listener Demographics**:
   - Include features like age, gender, and location to understand audience-specific preferences.
   - Analyze regional variations in song popularity.
2. **Explore Temporal Trends**:
   - Use time-series analysis to capture how song popularity evolves over time.
   - Identify seasonal trends or event-driven spikes in engagement.

3. **Evaluate Additional Features**:
    - Integrate external data, such as artist collaborations, music video attributes, and social media mentions.
4. **Implement Real-Time Analytics**:
    - Build real-time dashboards to provide actionable insights for artists and marketers.

Conclusion

This analysis reveals the power of data-driven insights in the music industry, highlighting the critical role of engagement metrics and audio features in shaping song popularity. Logistic Regression with Lasso emerged as the most effective predictive model, offering actionable guidance for stakeholders.

By implementing the recommended strategies, artists and platforms can better align their offerings with audience preferences, resulting in increased engagement and profitability. Future research focusing on demographic data, temporal trends, and advanced machine learning models will further refine these insights, unlocking new opportunities for innovation in music analytics.

References

- Kaggle. (n.d.). *Spotify and YouTube Dataset*. Retrieved from https://www.kaggle.com/
- Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), 293–302**.**
- Xu, C., Choi, K., Lee, J., & Yoo, C. D. (2020). Investigation of audio-based music popularity prediction using deep learning. *International Society for Music Information Retrieval Conference (ISMIR)*.
- YouTube Creators Blog. (2021). *Understanding audience engagement: A guide for creators*. Retrieved from https://blog.youtube/