Spotify + YouTube Music

Data Cleaning & Preparation Report (Power BI)

DATA ANALYSIS PROJECT

Project Overview

Dataset Context: A combined dataset containing music track metrics from two major platforms, Spotify and YouTube Music, including engagement metrics (views, likes, comments, streams) and metadata (licenses, album info).

Objective: To transform raw, unstructured data into a high-integrity, clean dataset suitable for accurate analytical reporting and visualization in Power BI, ensuring all metrics are reliable and contextually accurate.

Data Cleaning & Preparation

Question 1: Identify and Handle Missing Values

🔍 **ISSUE IDENTIFIED**

The dataset contained missing (null) values across multiple columns, including `license`, `views`, `likes`, `comments`, and `streams`. These nulls were not uniform in nature and required context-based handling rather than a single generic approach.

⚙️ **ACTIONS TAKEN**

1. Handling Null Values in `license` Column

Rows containing null values in the license column were removed.

Reasoning: A detailed inspection showed that rows with a null license value were also missing most core analytical attributes such as artist name, track, key, and engagement metrics. In many cases, more than 90% of the row data was absent. Filling or imputing such rows would have resulted in fabricated or misleading data. Therefore, row removal was the most reliable and data-integrity-preserving choice. This step alone resolved approximately 60–70% of the missing data problem.

2. Handling Null Values in Views, Likes, Comments, and Streams

For views, likes, comments, and streams, rows were not removed, as these records contained complete contextual information (artist, track, channel). A separate artist-level aggregation table was created to calculate the average engagement metrics per artist. This table was merged back into the main dataset, and conditional logic was applied:

- If a metric value was null → replace it with the artist-level average
- If the value was already present → retain the original value

📋 **JUSTIFICATION**

Music engagement metrics are artist-dependent, not randomly distributed. Using a global dataset average would distort performance for both high-profile and low-profile artists. Artist-level averaging preserves natural engagement patterns, avoids bias, and ensures analytical realism.

✅ **OUTCOME**

- No meaningful data was lost
- Engagement metrics were completed without distortion
- Dataset became analysis-ready while preserving real-world behavior

Question 2: Fix Irregularities in Merged Columns

🔍 **ISSUE IDENTIFIED**

Spotify and YouTube columns contained merged values separated by delimiters, limiting analytical usability.

⚙️ **ACTIONS TAKEN**

- Spotify columns were split using the `|` delimiter
- YouTube columns were split using the `-` delimiter

Each platform was handled separately due to different metadata structures. After splitting, unnecessary fragments and prefixes were removed, retaining only analysis-relevant fields.

📋 **JUSTIFICATION**

Merged columns violate the principle of atomic data, where each column should represent a single attribute. Proper splitting ensures accurate filtering, clean grouping, and reliable downstream analysis.

✅ **OUTCOME**

Unstructured text fields were converted into clean, structured columns, improving analytical flexibility.

Question 3: Correct Case Sensitivity and Naming Conventions

🔍 **ISSUE IDENTIFIED**

Several column names were in uppercase. Additionally, `artist` and `track` columns contained redundant keywords embedded within values.

⚙️ **ACTIONS TAKEN**

1. Column Name Standardization

All column names were manually converted to lowercase for consistency and model readability.

2. Cleaning Artist Column

Pattern identified: `artist_artistname`. Using delimiter-based extraction, text before `_` was removed, retaining only the artist name.

3. Cleaning Track Column

Pattern identified: `trackname_track`. Text after `_` was removed to retain the clean track name.

📋 **JUSTIFICATION**

Removing redundant system-generated keywords ensures that text fields represent pure business entities, improving grouping, sorting, and visualization clarity.

✅ **OUTCOME**

Naming consistency was achieved, and text fields became clean, human-readable, and analysis-ready.

Question 4: Remove or Handle Irrelevant Columns

🔍 **ISSUE IDENTIFIED**

Several columns did not provide analytical value and increased model complexity.

⚙️ **ACTIONS TAKEN**

The following columns were removed:

- `random_1`
- `random_2`
- `unnamed`
- YouTube URLs / info
- Spotify URLs / track IDs

📋 **JUSTIFICATION**

Random and unnamed columns contained no business logic. URLs and IDs are non-aggregatable and do not contribute to performance analysis. All relevant insights were already captured via engagement metrics and artist attributes. Retaining such fields would clutter the model without adding insight.

✅ **OUTCOME**

A leaner, more focused dataset aligned with the project's analytical goals.

Question 5: Handle Inconsistent Data Types

🔍 **ISSUE IDENTIFIED**

Some numeric columns required explicit validation to ensure correct data types, including `danceability`.

⚙️ **ACTIONS TAKEN**

All columns were reviewed and converted to their appropriate data types:

- Numeric metrics → numeric
- Categorical attributes → text
- Boolean flags → true/false

Danceability inconsistencies were already resolved during Question 1 by removing rows lacking sufficient supporting data.

📋 **JUSTIFICATION**

Explicit data type enforcement prevents silent aggregation errors and ensures calculation accuracy in Power BI.

✅ **OUTCOME**

The dataset achieved full data-type integrity with no unresolved inconsistencies.

Question 6: Address and Fix Invalid Data Entries

🔍 **ISSUE IDENTIFIED**

Views Column: Invalid non-numeric entries surfaced during data-type conversion.

Album Column: Some album values appeared numeric (e.g., 123456) but were stored as text and showed no Power BI errors.

⚙️ **ACTIONS TAKEN**

Views Column: These values were converted to null and handled using the same artist-level averaging logic defined in Question 1. No separate imputation logic was introduced.

Album Column: No values were removed or modified.

📋 **JUSTIFICATION**

For the Album column, the field is descriptive metadata. Numeric-looking text does not imply invalid data. Removing such rows would cause unjustified data loss.

✅ **OUTCOME**

Valid but imperfect metadata was preserved, maintaining dataset completeness.

Question 7: Check for and Remove Duplicate Rows

🔍 **ISSUE IDENTIFIED**

Certain track names appeared more than once.

⚙️ **ACTIONS TAKEN**

Duplicate checks showed that repeated track names belonged to different artists or channels, representing distinct songs rather than duplicated records. Track name alone was not treated as a unique identifier. No duplicate rows were removed.

📋 **JUSTIFICATION**

Removing records based only on matching track names would result in data loss and analytical bias.

✅ **OUTCOME**

Dataset was confirmed to be free of true duplicates.

Question 8: Reorder and Rename Columns for Clarity

⚙️ **ACTIONS TAKEN**

Columns were reordered into a logical analytical sequence:

1. Identification & descriptive attributes
2. Audio features
3. Engagement metrics

Column names were reviewed and renamed only where clarity improved (e.g., `view` → `views`, `like` → `likes`).

📋 **JUSTIFICATION**

Logical column ordering improves readability, modeling efficiency, and long-term maintainability without altering data values.

✅ **OUTCOME**

The dataset is clean, consistent, analysis-ready, and suitable for accurate Power BI visualization and reporting.

Final Data Quality Summary

Dataset Readiness	Analytical Reliability	Suitability
100% Prepared for Modeling	High (Context-Aware Imputation)	Optimized for Power BI Visualization

Final Analyst Note

All transformations were performed with a focus on data integrity, contextual relevance, and analytical reliability rather than cosmetic cleaning. Decisions regarding null handling and duplicates were driven by the business context of music streaming data to ensure that the final output accurately reflects artist performance without introducing statistical bias.

Prepared By

Chandra Prakash Choudhary

LinkedIn Profile GitHub Profile