DATA CLEANING AND TRANSFORMATION

Project – Spotify + YouTube Music

Dataset - (“https://drive.google.com/file/d/1qanyuwEzkwEJ73vDJHk4ZlWE0JUG7udb/view”)

Introduction

This report provides a detailed explanation of the full data cleaning and transformation process applied to the Spotify–YouTube dataset. Each step includes the affected columns, the issue identified, the approach used to correct it, and the reasoning behind each transformation. The goal is to produce a clean, structured, and analysis-ready dataset suitable for Power BI dashboards and insights.

Project Summary

The objective of this project was to convert an unclean, inconsistent dataset into a structured, analysis-ready format suitable for reporting and dashboard creation in Power BI.

Key operations included:

- Handling missing values
- Splitting merged columns
- Correcting case-sensitivity issues
- Removing irrelevant fields
- Standardizing data types
- Fixing invalid entries
- Eliminating duplicates
- Reordering and renaming columns for clarity

Through these steps, the dataset was fully optimized for professional reporting and advanced analytics.

DATA CLEANING AND TRANSFORMATION

1. Handling Missing Values

Columns Involved:

- Views, Likes, and other columns containing Null values

Issue Identified:

- Several rows contained missing or blank entries, especially in engagement fields such as Views and Likes.

Approach

- Used Power Query → Transform → Replace Values to convert missing values in Views and Likes to 0.
- Checked all other columns using Column Quality and cleaned or filled missing values where required.

Why This Approach:

- 0 is a meaningful value for both Views and Likes (indicating no engagement).
- Prevents loss of rows due to Null values.
- Ensures smooth calculations, aggregations, and visualizations in Power BI.

---

2. Fixing Merged or Combined Columns

Columns Involved:

- Spotify_Info, Youtube_Info

Issue Identified:

- Both columns contained multiple pieces of data merged together using delimiters or fixed patterns.

Approach:

- Applied Split Column → By Delimiter for Spotify_Info (using "|").
- For Youtube_Info, used Split Column → By Delimiter or By Number of Characters, depending on pattern.
- Cleaned the resulting columns with Trim, Clean, and removed unwanted prefixes/suffixes.

Why This Approach:

- Splitting restores original attributes such as IDs, links, and metadata.
- Improves clarity and allows proper modeling in Power BI.
- Ensures each field contains only one meaningful value.

---

3. Standardizing Text Fields and Column Names

Columns Involved:

- All column names, Artist, Track, Album

Issue Identified:

- Column names used inconsistent casing and spacing. Text values in Artist and Track were not formatted uniformly.

Approach:

- Converted all column names to lowercase_with_underscores using Transform → Format.
- Applied Capitalize Each Word for Artist and Track columns.
- Used Trim and Clean to remove extra spaces and non-printable characters.

Why This Approach:

- Consistency in naming prevents formula and relationship errors.
- Properly formatted names look professional and improve readability.
- Standard text formatting ensures accurate grouping and filtering.

---

4. Removing Irrelevant or Random Columns

Columns Involved:

- Irrelevant fields such as unnamed columns or random metadata fields

Issue Identified:

- Dataset contained columns with random values, placeholders, or auto-generated index numbers.

Approach:

- Used Remove Columns to delete fields with no analytical value.
- Identified the “unnamed: 0” column. Instead of deleting it, recognized it could serve as Track_ID, so it was renamed and retained.

Why This Approach:

- Removes noise and reduces dataset complexity.
- Prevents unnecessary load on Power BI.
- Preserves useful identifiers for primary key or row-level uniqueness.

---

5. Fixing Inconsistent Data Types

Columns Involved:

- Danceability, Energy, Tempo, Duration_ms, Views, Likes

Issue Identified:

- These numeric columns were incorrectly stored as text due to symbols, formatting issues, or inconsistent data entry.

Approach:

- Removed text characters using Replace Values (e.g., %, letters, “k”).
- Changed data type using Transform → Data Type → Decimal Number / Whole Number.
- Handled conversion errors with Replace Errors → Null and cleaned them afterward.

Why This Approach:

- Ensures accurate numeric calculations.
- Prevents Power BI visual errors.
- Makes measures like SUM, AVERAGE, and MAX work correctly.

---

6. Correcting Invalid Data Entries

Columns Involved:

- Views, Album

Issue Identified:

- The Views column contained entries such as "invalid_data" and “nan”. The Album column included numeric or irrelevant text entries.

Approach:

- Replaced invalid entries in Views with Null → then converted to numeric.
- Cleaned the Album column using Remove Errors, Trim, and manual corrections to keep only valid album names.

Why This Approach:

- Prevents incorrect metrics from affecting analysis.
- Ensures that Views and Album columns contain only meaningful and valid values.
- Supports reliable filtering and visualizations.

---

7. Removing Duplicate Records

Columns Involved:

- Track_ID, Track, Artist, Album

Issue Identified:

- Some rows appeared multiple times, which would inflate engagement numbers.

Approach:

- Used Remove Duplicates in Power Query.
- Used Track_ID or a combination of Track + Artist + Album as the unique key.

Why This Approach:

- Duplicate-free data ensures accurate insights.
- Avoids inflated View/Like counts.
- Produces a dependable dataset for modeling.

---

8. Reordering and Renaming Columns

Columns Involved:

- Track_ID, Track, Artist, Album, Views, Likes, Danceability, Energy, Tempo, Duration_ms, other metadata fields

Issue Identified:

- The default order of columns did not follow a meaningful or logical structure.

Approach:

- Reordered columns into logical groups:
- Track Details → Engagement Metrics → Audio Features → Additional Metadata
- Renamed unclear or technical column names to professional, readable names.

Why This Approach:

- Makes the dataset more intuitive and easier to work with.
- Supports better readability and cleaner Power BI dashboards.
- Helps users quickly understand the dataset structure.

Conclusion

The dataset has been thoroughly cleaned and transformed using Power BI's Power Query Editor. Each major issue—missing values, merged columns, inconsistent text, wrong data types, invalid entries, duplicates, and unclear structure—was systematically corrected.

The final dataset is:

✓ Clean

✓ Structured

✓ Consistent

✓ Professional

✓ Ready for Power BI dashboards and analysis

Thank You

Submitted By : Manish Kumar