"DATA CLEANING AND TRANSFORMATION"

(Using Power BI)

INTRODUCTION

Spotify and YouTube Music are two of the world's leading music streaming platforms, each playing a key role in shaping the modern digital music experience. Spotify focuses on audio streaming, providing users with millions of tracks, personalized playlists, and recommendations based on listening habits.
YouTube Music offers a wider range of content — from official songs to user-created videos — allowing users to explore music visually and audibly.

Both platforms attract millions of listeners globally, each with unique tastes and preferences.
This project aims to combine and analyze data from Spotify and YouTube Music to identify trends, understand audience behavior, and prepare the dataset for deeper analysis and visualization using Power BI.

PROJECT SUMMARY

The **Spotify and YouTube Music Data Cleaning Project** focuses on preparing a unified dataset for accurate analysis.
The main goal was to clean, standardize, and organize the data to ensure reliability and consistency.
Key tasks included handling missing and inconsistent values, separating merged columns, correcting case sensitivity, standardizing data types, and removing irrelevant or duplicate entries.

After completing the cleaning process, the dataset was optimized for visualization in Power BI.
This clean and structured dataset allows for detailed insights into music trends, artist performance, and audience engagement across both streaming platforms.

Tools Used

- Microsoft Power Query
- Power BI (for future visualization)

1. IDENTIFY AND HANDLE MISSING VALUES

- Examine the dataset for any missing values. Which columns contains null values.
- How should missing values in the Views and like columns be handles? Should they be filled with a default values, removed, or handled in another way? Justify your approach.

Solution:

Handling "Views" Column

- **Meaning:** Represents the number of times a song or video has been played on the platform.
- **Action Taken**: All missing or blank values were replaced with 0 using the *Replace Values* option in Power Query.
- **Reason for Choosing 0:**
  → It ensures that rows with missing data are not removed.
  → A value of 0 logically represents "no views recorded yet."
  → Helps maintain uniformity and avoids errors during calculations or visualizations in Power BI.

Handling "Likes" Column

- **Meaning:** Indicates how many users have liked a particular song or video.
- **Action Taken:** Missing values were replaced with 0 using Power Query transformation.
- **Reason for Choosing 0:**
  → Represents that the video might not have received any likes yet.
  → Keeps the dataset complete without deleting any rows.
  → Ensures better accuracy during aggregation and comparison in Power BI reports.

Justification:

Replacing missing values with 0 helps maintain **data completeness** and **consistency** without losing any records.
It ensures smooth functioning of calculations, accurate visualizations, and reliable analysis in Power BI.

---

2. FIX IRREGULARITIES IN MERGED COLUMNS

- The Spotify_Info and Youtube_Info columns contain merged data separated by delimiters. Split these columns back into their original components. What are the original components, and how can you ensure that the split data is clean and accurate?
- After splitting, remove any unnecessary delimiters or prefixes/suffixes that do not belong

Solution:

Handling "Spotify_Info" Column

- **Issue Identified:**
  The Spotify_Info column contained multiple pieces of information (such as track ID, artist ID, and link) merged together and separated by a delimiter.
- **Action Taken:**
  Used Split Column → By Delimiter in Power Query, selecting the "|" (pipe) symbol as the separator.
  This separated the data into clear fields — for example:
  - Spotify_Link
  - Artist_ID
  - Track_ID

- **Data Cleaning Step:**
  After splitting, extra spaces and unnecessary characters were trimmed using Transform → Format → Trim/Clean to ensure each field was accurate and properly formatted.

Handling "Youtube_Info" Column

- **Issue Identified**:
  The Youtube_Info column also had merged text data — such as video ID, video type, and metadata — either separated by symbols or fixed-length structures.
- **Action Taken:**
  Used Split Column → By Number of Characters when data parts were fixed-length, or By Delimiter if a consistent separator was found.
  Renamed resulting fields for clarity, such as:
  - Youtube_VideoID
  - Youtube_Type
  - Youtube_Metadata
- **Data Cleaning Step:**
  Removed redundant prefixes or suffixes that did not add analytical value.

Justification:

Splitting these merged columns restored the original data structure, making it easier to analyze and visualize in Power BI.
It enhanced data readability, ensured accuracy, and allowed smoother relationships between tables during the modeling phase.

3.CORRECT CASE SENSITIVITY AND NAMING

CONVENTIONS:

• The column names have inconsistent case sensitivity (some are uppercase, others lowercase). Standardize all column names to follow a consistent format (e.g., all lowercase with underscores).

• Fix any data entries where case sensitivity might affect consistency (e.g., artist names or track titles). Ensure that the Artist and Track columns are formatted consistently.

Solution:

Standardizing Column Names:

Renamed all column headers in **Power Query Editor** to maintain a consistent format.
Converted all names to **lowercase** and replaced **spaces, special characters (such as ":" or ".")** with **underscores** using the *Transform → Format → Lowercase* and *Replace Values* options.
This ensures uniform naming conventions for easy reference in Power BI formulas and visuals.

Formatting Text Columns ("Artist" and "Track"):

Applied **Transform → Format → Capitalize Each Word** on the *Artist* and *Track* columns.
This standardized all entries (e.g., "gorillaz" → "Gorillaz") to enhance readability and maintain consistency in visualizations.

Checking for Anomalies:

Scanned the dataset for duplicate or incorrectly formatted artist and track names.
Verified that the changes were applied correctly using *Data Preview* before loading the cleaned data back into Power BI.

Justification:

Consistent column naming and properly formatted text values improve data clarity and professionalism.
They also make it easier to perform further transformations, apply DAX formulas, and build clear and accurate Power BI reports.

4. REMOVE OR HANDLE IRRELEVANT COLUMNS:

- Identify and remove any irrelevant or randomly generated columns that do not provide useful information for analysis. Which columns should be removed, and why?
- If any random data exists in relevant columns, clean or remove those entries.

Solution:

- **Identifying Irrelevant Columns:**

Used the Power Query Editor to inspect all columns in the dataset.
Identified unnecessary columns such as *random_column_1* and *random_column_2*, which contained random or null values with no analytical significance.

- **Approach:**

Applied the Remove Columns option from the *Home* tab in Power Query to delete these irrelevant fields.
This helps simplify the dataset and avoid unnecessary processing in Power BI.

- **Handling Semi-Relevant Columns:**

Detected the column *“unnamed:0”* containing unique sequential values.
Rather than deleting it, recognized that it could serve as a unique identifier (Track ID) for each record.
Thus, retained it for potential use in identifying or merging datasets.

- **Cleaning Random Data in Relevant Fields:**

Scanned other columns (e.g., Artist, Album, Track) for any random or meaningless text.
Used the Replace Values or Remove Rows → Remove Errors/Blanks options to clean irregular entries.

• Justification:
Eliminating irrelevant columns and cleaning random data improves dataset quality and performance.
It ensures that only meaningful, accurate, and analysis-ready data is loaded into Power BI for visualization and reporting.

5. HANDLE INCONSISTENT DATA TYPES:

- Some columns that should be numeric (e.g., Danceability, Energy) are stored as text. Convert these columns back to numeric format. What steps would you take to identify and fix any issues that arise during this conversion?
- Ensure that all numeric columns are in the correct format and handle any non-numeric values or anomalies.

Solution:

• Identifying Affected Columns:
In Power Query, checked the *Data Type* icons beside each column to identify fields such as *Danceability* and *Energy* that were incorrectly stored as text.

• Cleaning Non-Numeric Characters:
Used the Replace Values option to remove unwanted symbols like %, k, or text strings that prevented numeric conversion.
Also, filtered out blank or non-numeric entries before transformation.

• Converting Data Type:
Selected the affected columns → Transform Tab → Data Type → Decimal Number to ensure proper numeric format.
When conversion errors appeared, used Replace Errors → Replace with 0 or Null to maintain consistency.

• Verification:
Sorted each column to confirm numeric sorting worked correctly and previewed values to ensure proper decimal formatting.

• Justification:
Maintaining numeric data types ensures accurate aggregation and smooth DAX calculations (e.g., SUM, AVERAGE).
This step enhances analytical reliability and prevents issues in Power BI dashboards or visualizations.

6. ADDRESS AND FIX INVALID DATA ENTRIES

- Check the *Views* column for any entries labeled as "invalid_data" or other incorrect values. Replace or clean these entries and justify your method.
- Verify that all values in the *Album* column are correctly labeled and do not contain numeric or irrelevant data.

Solution:

1) Invalid Entries in "Views" Column

- **Approach:**
  Used Power Query → Transform → Replace Values to replace "invalid_data" and "nan" with null.
  Then changed the column type to Whole Number using Transform → Data Type → Whole Number.
  If any fractional data existed (e.g., 1.5k), converted the column to Decimal Number.
- **Why:**
  Ensures that *Views* only contains valid numerical data for accurate calculations and aggregation.
  Prevents errors during DAX formula execution and report generation in Power BI.

2 Cleaning the “Album” Column

- **Approach:**
  Checked for invalid or numeric entries in the *Album* column.
  Used Remove Rows → Remove Errors/Blanks and replaced irrelevant entries with null.
  Applied Transform → Format → Trim/Clean to remove extra spaces.
- **Why:**
  Ensures all album names are properly formatted text values, improving clarity and consistency across the dataset.

Justification:

By cleaning invalid data, replacing random text with nulls, and setting proper data types, the dataset becomes structured, consistent, and reliable.
This process minimizes analytical errors and ensures accurate insights in Power BI reports.

7. CHECK FOR AND REMOVE DUPLICATE ROWS:

• Identify and remove any duplicate rows in the dataset. How can you ensure that the remaining data is unique and accurate?

Solution:

Approach:

- In Power Query, use **Home → Remove Rows → Remove Duplicates**.
- Based on earlier analysis, the **“unnamed: 0”** column serves as a **unique identifier** for each row.
- Apply the duplicate removal step using this column to ensure each record appears only once.

Why:

- The **“unnamed: 0”** column contains sequential numeric values (e.g., 1, 2, 3, …), which represent unique row IDs.
- Removing duplicates based on this column ensures all remaining rows are **unique and accurate**, preventing data redundancy.

Justification:

By removing duplicate records, the dataset becomes **cleaner, more reliable, and easier to analyze**.
Unique data improves accuracy in aggregations, summaries, and all subsequent Power BI reports.

8. REORDER AND RENAME COLUMNS FOR CLARITY:

• Reorder the columns in a logical sequence to improve the dataset's readability and usability. What order makes the most sense for this dataset?

• Rename columns where necessary to ensure that their names clearly reflect the data they contain.

Solution:

Reordering Columns

Approach:

- Used Power Query → Manage Columns → Choose Columns → Reorder Columns
- Arranged columns in a logical sequence for better readability, such as: Track_ID → Track → Artist → Album_Name → Views → Likes → Danceability → Energy → Tempo

Why:

- Grouping related fields together (Track, Artist, Album) improves understanding of song details.
- Numeric metrics (Views, Likes, Energy) are placed together for easier analytical comparison.
- Logical order enhances navigation and readability while preparing dashboards in Power BI.

Renaming Columns

Approach:

- Used Power Query → Transform → Rename Columns to rename unclear column headers.
- **Example:**
  - "youtube_info.2" → "YouTube_Song_Title"
  - "unnamed: 0" → "Track_ID"
  - "track.name" → "Track"

Why:

- Descriptive and properly formatted names make the dataset self-explanatory.
- Clear column titles prevent confusion during DAX formula creation or visual analysis.
- Aligns with naming conventions (no spaces, consistent capitalization, underscores).

Justification:

Reordering and renaming columns enhance data clarity, usability, and presentation quality.
These steps ensure that the dataset follows professional standards and is easy to interpret for analysis, reporting, and dashboard creation in Power BI.

Conclusion

The complete data cleaning and transformation process in Power BI (Power Query) successfully converted a raw and inconsistent Spotify–YouTube Music dataset into a clean, structured, and analysis-ready format.
Through a series of systematic steps — including removing duplicates, fixing data types, correcting invalid entries, renaming columns, and ensuring consistency — the dataset now maintains both accuracy and professional quality.

Key improvements achieved:

- All columns follow proper naming conventions and data types.
- Invalid, missing, or irrelevant values were removed or standardized.
- Numeric, text, and categorical fields were formatted for accurate insights.
- Duplicates and random noise were eliminated to ensure data integrity.
- Columns were reordered and renamed for improved readability and reporting efficiency.

These transformations ensure that the dataset is reliable, consistent, and ready for visualization in Power BI dashboards.
Overall, the process enhances data quality and ensures that any analysis or insight derived from it is accurate, meaningful, and trustworthy.

THANKYOU!

Submitted By: ***GAJENDRA SINGH PANWAR***