Spotify & YouTube Music Dataset Cleaning

Introduction

The dataset comprises information from both Spotify and YouTube. This document outlines the data cleaning process conducted using Power BI, focusing on handling missing values, fixing irregularities, correcting data types, and ensuring data consistency.

Steps for Data Cleaning

1. Retrieving and Organizing Data

- Imported the CSV file into Power BI.
- Entered the Power Query editor to begin the data transformation process.
- Columns from Spotify and YouTube were mixed, so we rearranged them by using the **Move** function, grouping Spotify-related columns on the left and YouTube columns on the right.
- Unnecessary columns were removed to improve clarity and focus on relevant data.

2. Renaming Columns

- Renamed columns to follow consistent formatting.
    - Example: `Youtube_Info` → `youtube_info`
- Renaming was done by double-clicking the column headers and typing the new names.

3. Handling Duplicate Rows

- Checked for duplicate entries using **Column Profiling**.
    - Initial column profiling based on 1000 rows showed no duplicates, but checking the entire dataset revealed duplicates.
- The **Index Column** was used to identify and remove duplicates.
- **Remove Duplicates** was applied, reducing the dataset from 20,718 rows to 19,682.

4. Handling Missing Values

- Missing values were present in several columns like `Views`, `Likes`, `Description`, `Streams`, and `Comments`.

- We decided not to fill missing textual data (e.g., `Description`, `Comments`) as in real-world scenarios, many videos may lack this information. No significant bias was introduced by leaving them empty for this analysis.

5. Fixing Merged Columns

- Columns like `spotify_info` and `youtube_info` contained multiple pieces of information merged into one, separated by delimiters.
    - For `spotify_info`, data was separated by the pipe symbol (`|`). We used **Split Column by Delimiter** and specified the pipe symbol to split it into `Track_ID` and `Track_Name`.
    - For `youtube_info`, we used **Extract Text by Number of Characters** as YouTube links had consistent character lengths (44 characters). This split the YouTube link from the song title.

6. Correcting Data Types

- **Views**: Initially stored as text, needed to be numeric. Attempted conversion failed due to an "invalid" entry.
    - Replaced the "invalid" value with `null` using **Replace Values**, then successfully converted the column to a numeric data type.
- **Danceability** and **Energy**: These columns were supposed to be numeric but contained invalid entries like "NaN".
    - Changed data type back to text to detect "NaN", replaced them with `null`, and then re-converted to numeric.

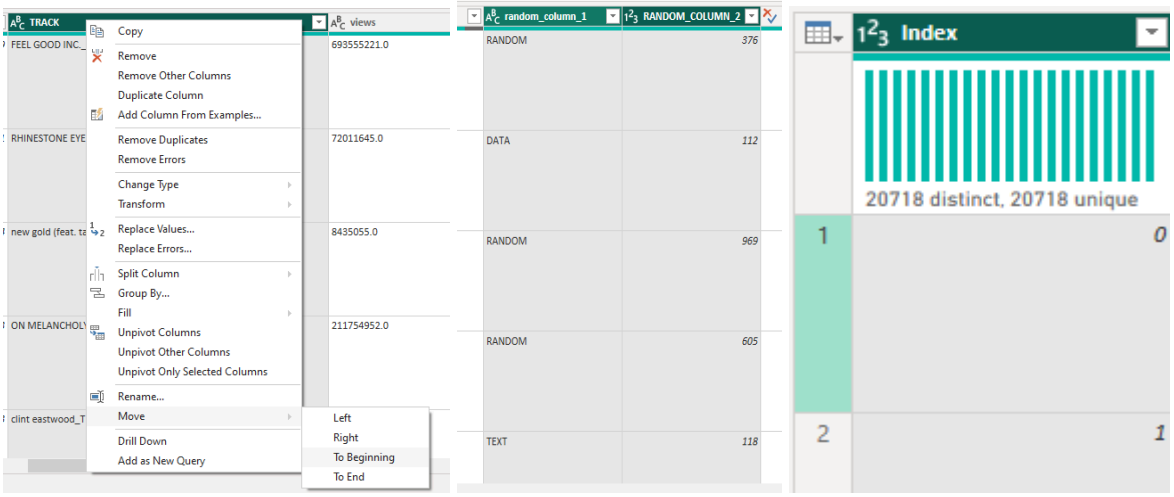
7. Handling Text Irregularities

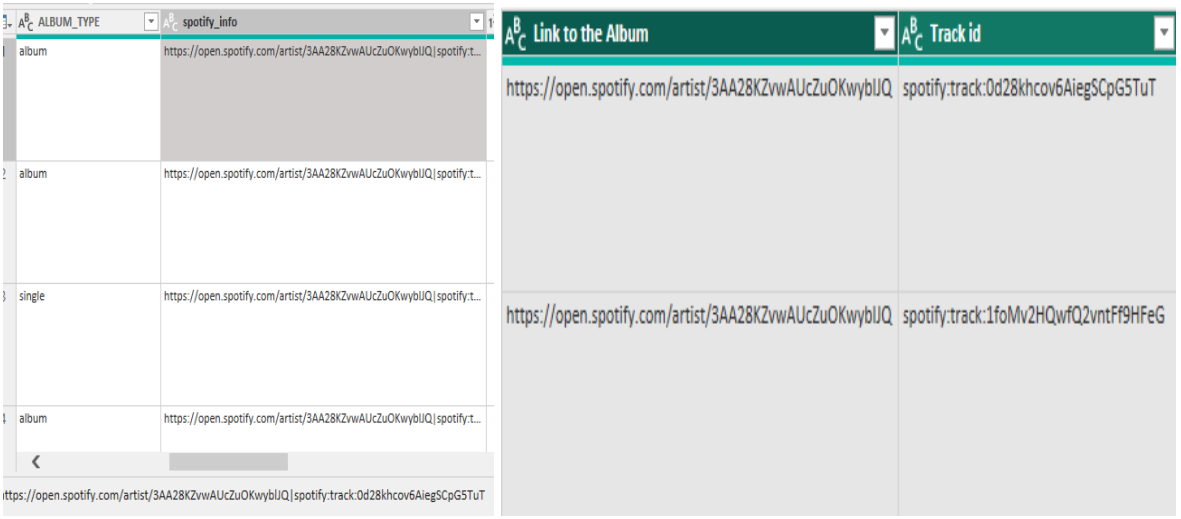
- In the `Track` and `Artist` columns, every entry had "_track" or "_artist" appended.
    - Used **Extract Text After Delimiter** to remove the suffix "_track" and "_artist" from each entry.

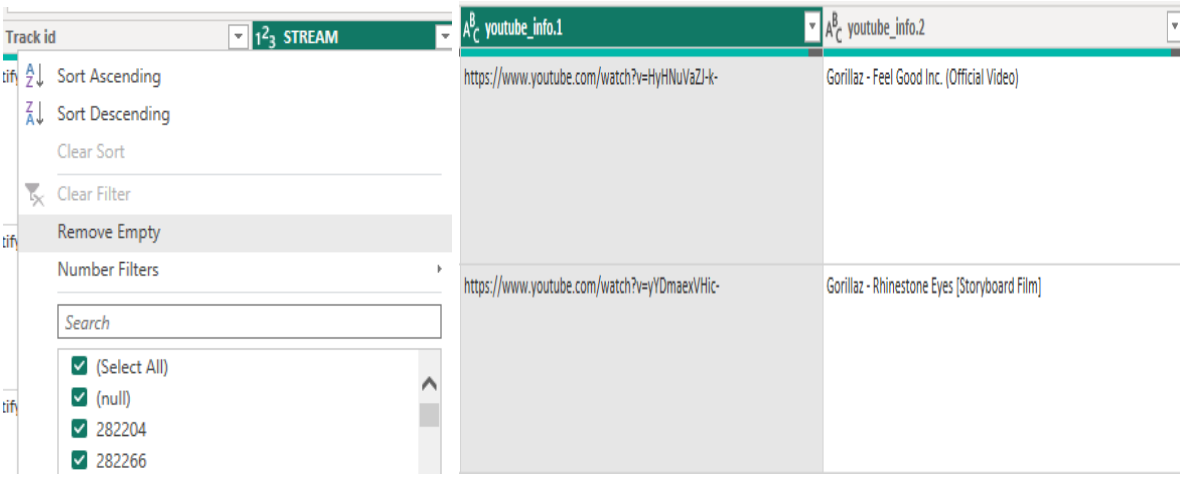
Conclusion

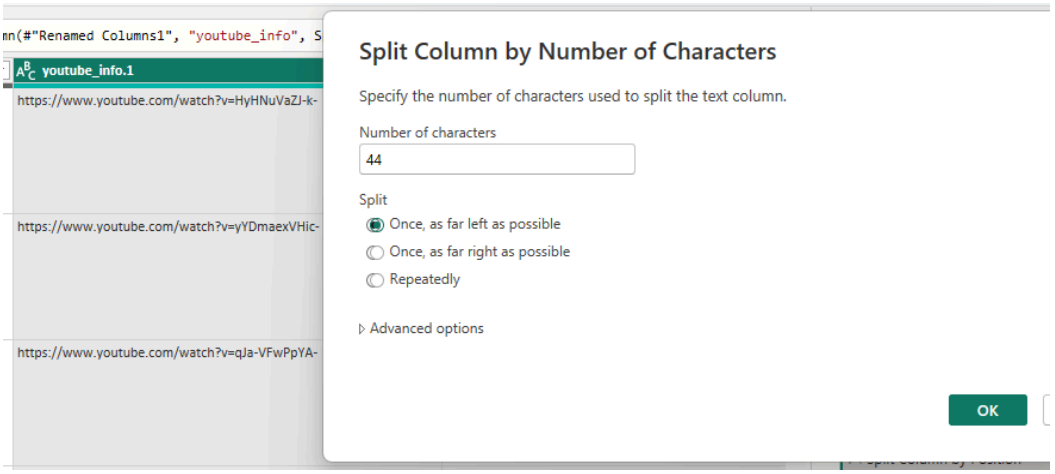
By following these steps, the dataset is now clean, ready for analysis, and consistent. The process tackled issues like duplicates, merged columns, missing values, and data type errors while maintaining the integrity of the original dataset.

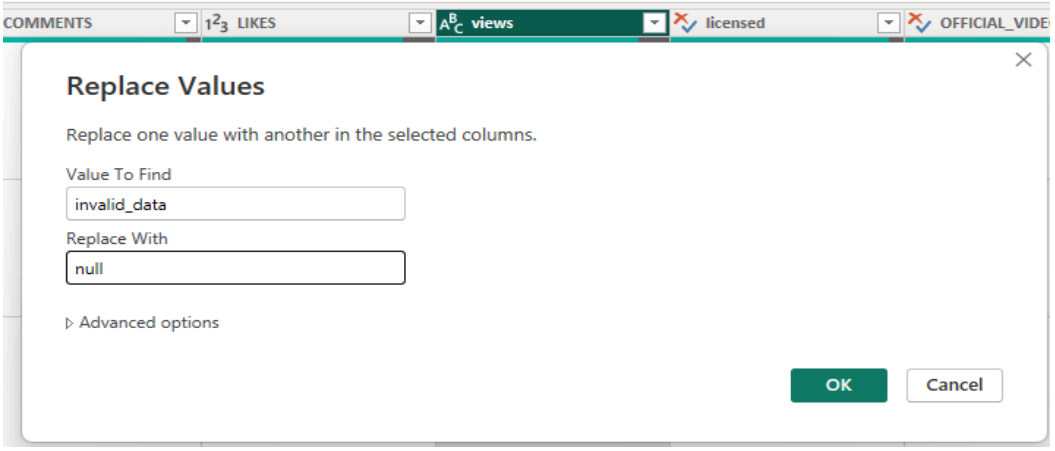
Screenshots

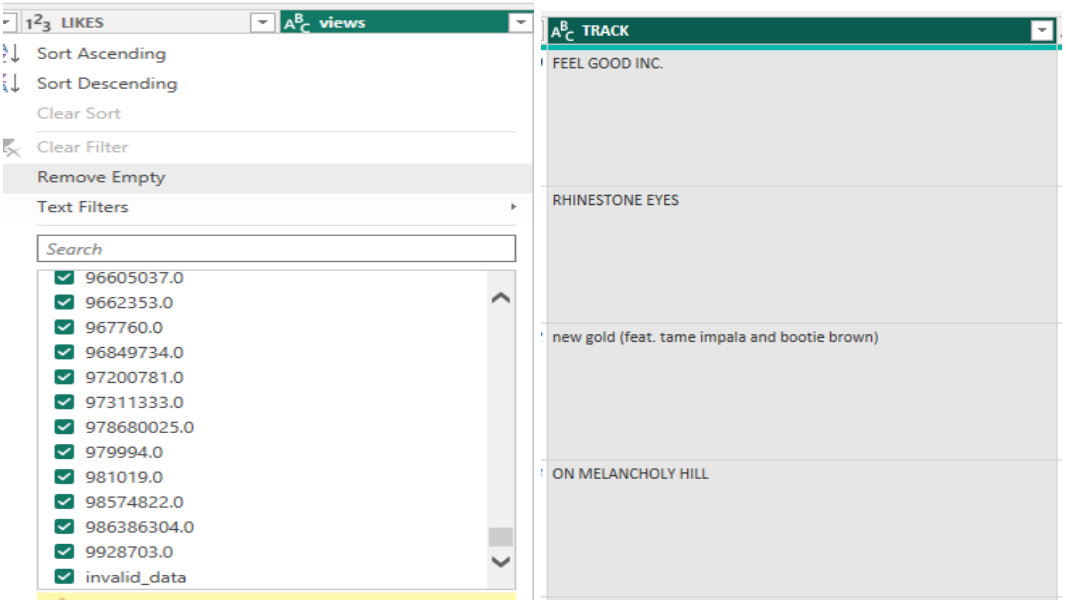







mn(#"Renamed Columns1", "youtube_info", S
youtube_info.1
https://www.youtube.com/watch?v=HyHNuVaZJ-k-
https://www.youtube.com/watch?v=yYDmaexVHic-
https://www.youtube.com/watch?v=qJa-VFwPpYA-
Split Column by Number of Characters
Specify the number of characters used to split the text column.
Number of characters
44
Split
Once, as far left as possible
Once, as far right as possible
Repeatedly
Advanced options
OK


COMMENTS
LIKES
views
licensed
OFFICIAL_VIDE
Replace Values
Replace one value with another in the selected columns.
Value To Find
invalid_data
Replace With
null
Advanced options
OK
Cancel


LIKES
views
Sort Ascending
Sort Descending
Clear Sort
Clear Filter
Remove Empty
Text Filters
Search
96605037.0
9662353.0
967760.0
96849734.0
97200781.0
97311333.0
978680025.0
979994.0
981019.0
98574822.0
986386304.0
9928703.0
invalid_data
TRACK
FEEL GOOD INC.
RHINESTONE EYES
new gold (feat. tame impala and bootie brown)
ON MELANCHOLY HILL