Spotify & YouTube Music

Dataset Link -
https://drive.google.com/file/d/1qanyuwEzkwEJ73vDJHk4ZlWE0JUG7udb/view

Questions for Data Cleaning:

1. Identify and Handle Missing Values:

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

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.

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.

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.

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.

6. Address and Fix Invalid Data Entries:

- Check the Views column for any entries labeled as "invalid_data" or any other incorrect values. Replace these entries and justify your method.
- Ensure that all values in the Album column are correctly labeled and that there are no numeric entries or irrelevant data.

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?

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.

Instructions for Submission:

1. For each issue identified, fix the irregularities to ensure the data is clean and consistent.
2. Document each step taken during the cleaning process and explain why the chosen methods were applied.