Case Study :-

Spotify & YouTube Music

05/09/2025

Presented by

Sharad Mittal

Questions for Data Cleaning:

1. Identify and Handle Missing Values:
2. Fix Irregularities in Merged Columns:
3. Correct Case Sensitivity and Naming Conventions:
4. Remove or Handle Irrelevant Columns:
5. Handle Inconsistent Data Types:
6. Address and Fix Invalid Data Entries:
7. Check for and Remove Duplicate Rows:
8. Reorder and Rename Columns for Clarity:

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.

Columns with Missing Values:

Column Name	Missing Values	Column Name	Missing Value
LIKES:	2685	licensed, OFFICIAL_VIDEO, CHANNEL, youtube_info	491
STREAM	610	key, valence, liveness	2
VIEWS	2484	speechiness, loudness, tempo, danceability	2
Description	911	DURATION_MS, ENERGY	2
COMMENTS	593	acousticness, instrumentalness	2

Steps:-

1. Identify Missing Values :- Inspected all columns and found missing values.
2. Handle Missing Values in Views :- Replaced missing values with 0 to reflect no engagement.
3. Approach:- using 0 in the palace of null values means there're no views .
4. Handle Missing Values in Likes :- Replaced missing values with 0 to reflect no engagement.
5. Approach:- using 0 in the palace of null values means there're no views .
6. Cross Check:- Rechecked the dataset to confirm no missing values remained.
7. Approach:- Ensures data consistency and completeness for analysis.

2. Fix Irregularities in Merged Columns: ·

1. 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? ·
2. After splitting, remove any unnecessary delimiters or prefixes/suffixes that do not belong.

Key Steps:-
1: Split the Spotify_Info and Youtube_Info Columns :-
Used the Text to Columns feature to split the data.
Components Identified: Spotify_Info: Spotify Link , Spotify ID. Youtube_Info: Youtube Link , Youtube video Title.
Approach : Splitting ensures individual components can be analyzed separately.

Clean the Split Data :-
use delimiters (|) pipe symbol for sportify_info and Hyphen for youTube_info to split the data accurately.

Verify Changes:-
Reviewed the new columns to ensure data integrity and no residual delimiters
Confirmed data is clean and ready for analysis.

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.

Key Steps:

1. Renamed columns to lowercase with underscores for uniformity and tool compatibility.
2. Standardized Artist and Track columns to title case for consistency and to avoid duplicates caused by case differences.
3. Verified and removed duplicates after case standardization.

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.

Key Steps:- Inspecting and Identifying Columns:

1. Irrelevant Columns Removed:
2. Columns Retained:
3. Handling Random Data
4. Checked the views and likes columns(for eg.)
5. Confirmed no irrelevant columns remain.
6. Ensured all data entries are consistent and usable.

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.

Key Steps:-

1. Identified columns (Danceability, Energy) with incorrect data types.
2. Converted columns to numeric format.
3. Verified the conversion.
4. Ensures no invalid data remains, and the data type is consistent.

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.

Key Steps:-

1. Identified invalid entries in the Views column.
2. Issue: Non-numeric entries like "invalid_data."
3. Action: Replaced invalid entries with the median of the column.
4. Justification: The median is less affected by outliers and provides a reliable replacement.
5. Checked the Album column for inconsistencies.
6. Issue: Numeric entries and irrelevant data.
7. Action: Replaced invalid entries with "Unknown" and standardized the column to title case.
8. Justification: Ensures consistency and prevents irrelevant data from affecting analysis.
9. Verified changes.
10. Rechecked both columns to confirm all values are valid and consistent.

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

Key Steps:-

1. Identified duplicate rows.
2. Action: Inspected the dataset for exact duplicates across all columns.
3. Approach: Prevents redundant data from skewing analysis.
4. Removed duplicate rows.
5. Action: Used the "Remove Duplicates" checking key columns (e.g., Artist, Track, Views).
6. Approach: Ensures concise data without losing relevant information.
7. Validated data uniqueness.
8. Action: Re-inspected the dataset to confirm no duplicates remain.
9. Approach: Ensures each row represents unique and accurate information.

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.

Key Steps:-

1. Reordered columns for readability.
2. Action: Rearranged columns to group metadata, metrics, and categorical data logically.
3. Approach: Improves usability for analysis by organizing related columns together.
4. Renamed ambiguous column names.
5. Action: Updated names to be descriptive and consistent (e.g., Views → total_views).
6. Approach: Enhances clarity and aligns with standard naming conventions.
7. Verified the changes.
8. Action: Reviewed the dataset to confirm the order and names were correct.
9. Approach: Ensures the dataset is ready for analysis with no confusion about column purposes.

Thank you

We Respect your valuable time .

SHARAD MITTAL