Spotify & Youtube Data Cleaning Project

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.

Objective: Address missing or null values to maintain data integrity.

Which columns have missing values?

– Columns with missing values include Licensed, Official Video, Views, YouTube Info, Comments, Description, Likes, and Stream.

What is the approach to handle missing values in quantitative metrics?

– For columns like Views, Likes, and Stream, rows with missing values are removed to ensure accurate performance analysis.

How should missing values in textual data be managed?

– For columns such as Description and Comments, missing or blank entries are retained as they do not significantly affect the analysis.

Summary: Missing values in quantitative columns were handled by removing affected rows, while textual data was retained as is.

---

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.

Objective: Separate merged columns into their original components for accuracy.

What are the components of the Spotify_Info and YouTube_Info columns?

– Spotify_Info contains Spotify Link and Spotify Track ID. YouTube_Info contains YouTube Link and YouTube Video Title.

What method should be used to separate these components?

– Use delimiters (|)pipe symbol for Spotify_Info and character length for YouTube_Info to split the data accurately.

How should the separated components be validated?

– Verify that links are functional and IDs are accurate for Spotify, and ensure video titles are correctly extracted for YouTube.

Summary: Separated Spotify_Info into Spotify Link and Track ID, and YouTube_Info into YouTube Link and Video Title, ensuring accuracy and cleanliness.

---

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.

Objective: Standardize column names and data entries for consistent formatting.

How should column names be standardized?

– Column names should be capitalized in title case (e.g., "Spotify Info" instead of "spotify_info").

What approach should be taken for data entry formatting?

– Text entries such as artist names and track titles should be capitalized consistently.

What are the benefits of standardizing formatting?

– Enhances clarity, consistency, and readability of the dataset, making it easier to work with.

Summary: Column names and text entries have been standardized to ensure consistency and improve dataset usability.

---

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.

Objective: Eliminate columns that do not contribute to the analysis.

Which columns are considered irrelevant?

– Random Column 1 and Random Column 2

What actions were taken regarding irrelevant columns?

– Removed the irrelevant columns to streamline the dataset.

How was random data handled in relevant columns?

– Verified and cleaned any random data in relevant columns to maintain data quality.

Summary: Removed irrelevant columns and ensured no random data remained in the relevant columns.

---

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.

Objective: Convert columns with inconsistent data types to their correct numeric format.

Which columns are affected by incorrect data types?

– Views, Danceability, Energy

How should the columns be converted to the correct numeric format?

– Use Power BI's "Change Data Type" feature to convert columns to decimal format.

What issues might arise during conversion?

– Potential non-numeric values or anomalies; use Power BI tools to filter and correct these issues.

How can we ensure correct formatting after conversion?

– Verify with a sample of rows to ensure formatting accuracy and validate no text values remain.

Summary: Converted Views, Danceability, and Energy columns to decimal format, correcting any non-numeric values.

---

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.

Objective: Handle invalid entries in the Views column and ensure correct labeling in the Album column.

What issues are present in the Views column?

– Contains invalid entries like "invalid_data."

How should invalid data in the Views column be managed?

– Replace "invalid_data" with null and then convert the column to numeric format.

What is the current state of the Album column, and how should it be corrected?

– Verify no numeric or irrelevant entries exist and clean the data to ensure proper labeling.

Summary: Managed invalid data in Views by replacing with null and ensured correct labeling in the Album column.

---

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?

Objective: Identify and remove duplicate rows to maintain data uniqueness and accuracy.

How were duplicate rows identified?

– Initial sample check showed no duplicates; full dataset review revealed duplicates.

What steps were taken to remove duplicate rows?

– Used the Index Column to identify and remove duplicates.

How was data accuracy ensured after removing duplicates?

– Final verification ensured that data remained unique and no critical data was removed.

---

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.

Objective: Improve dataset readability and usability by reorganizing and renaming columns.

How should the columns be reordered for better clarity?

– Columns should be grouped logically, with Spotify-related columns listed first and YouTube-related columns listed afterward.

What is the approach for renaming columns?

– Column names should be standardized to capitalize each word for consistency. For instance, "spotify_info" should be renamed to "Spotify Info."

What are the key columns to be retained and organized?

– Key columns for Spotify include Track, Album, Album Type, Key, and others. For YouTube, key columns include Channel, Views, Likes, Comments, and others.

Summary: Columns have been reordered to group related data and renamed for consistency, enhancing clarity and usability.


Thank
you