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.

-Analysis of Missing Values:

The dataset contains missing values across several columns, which require a methodical approach to address. The columns and their respective counts of null values are as follows:

~ Key: 2 missing values
~ Licensed: 491 missing values
~ Valence: 2 missing values
~ Liveness: 2 missing values
~ Speechiness: 2 missing values
~ Loudness: 2 missing values
~ Official_Video: 491 missing values
~ Tempo: 2 missing values
~ Views: 2484 missing values
~ Danceability: 2 missing values
~ Duration_MS: 2 missing values
~ Instrumentalness: 2 missing values
~ Channel: 491 missing values
~ Youtube_Info: 491 missing values
~ Comments: 593 missing values
~ Description: 911 missing values
~ Likes: 2685 missing values
~ Stream: 610 missing values
~ Acousticness: 2 missing values
~ Energy: 2 missing values

- **Handling Missing Values in Views and Likes Columns:**

The Views and Likes columns are particularly significant for analysis, necessitating careful handling of their missing values.

1. Filling Missing Values with Defaults :

When the absence of values signifies zero engagement, filling with a default value such as 0 can effectively reflect this condition. Alternatively, imputing with the median is advantageous for maintaining the dataset's overall statistical integrity, particularly in cases where outliers exist.

2. Removing Rows with Missing Values:

If a significant analysis hinges on Views and Likes data, and the missing values constitute a small proportion, removal can ensure the dataset's reliability.

3. Advanced Imputation Using Correlations:

By leveraging correlations with other features such as Stream, Duration_MS, or Description, predictive modeling can estimate missing values with reasonable accuracy.

Suggested Approach:

Perform an initial analysis to determine if the missing data exhibits any patterns (e.g., concentrated in specific categories).

Use median imputation for handling missing values in Views and Likes to account for outliers while preserving overall data structure. Rows with extensive missing values should be removed if imputation compromises reliability.

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.

-Issues Identified:

Merged columns such as Spotify_Info and Youtube_Info contain data concatenated by delimiters, which require separation into distinct components.

Steps Taken:

1. Splitting Merged Data:

Separated Spotify_Info into Spotify_Album_Link and Spotify_Track_Id, and Youtube_Info into corresponding components using delimiters.

2. Data Cleaning:

Removed extraneous prefixes, suffixes, and delimiters from split components to ensure clean, usable data.

3. Column Renaming and Refinement:

Renamed columns to accurately describe their contents, ensuring clarity and consistency.

4. Validation:

Verified that the newly split columns aligned with original data formats and eliminated any duplicates to maintain data integrity.

This step ensures the proper decomposition of merged data, enabling improved usability for downstream 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

-Issues Identified:

Inconsistencies in case sensitivity and naming conventions across column headers and data entries reduce clarity and can lead to analytical errors.

Steps Taken:

1. Standardization of Column Names:

All column names were converted to lowercase and formatted using underscores for uniformity.

2. Data Entry Formatting:

For critical columns such as Track and Artist, uniform text formatting was applied (e.g., converting all entries to uppercase or title case as appropriate).

3. Preservation of Intentional Formatting:

Columns like Description, where mixed case enhances readability, were preserved without modifications.

These measures ensure consistency across the dataset, mitigating errors caused by variations in naming conventions.

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.**

-Identified Columns:

Columns lacking relevance or containing mostly null values contribute little to analysis and should be removed.

Steps Taken:

1. Identification of Redundant or Low-Value Columns:

Columns with excessive null values, placeholders, or duplicate information were flagged for removal.

2. Rationale for Removal:

Eliminating such columns reduces noise in the dataset and enhances focus on meaningful variables.

Outcome:

The dataset was streamlined by removing redundant and irrelevant columns, improving clarity and analytical efficiency.

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.

-Issues Identified:

Certain columns, such as Danceability and Energy, were improperly stored as text instead of numeric formats.

Steps Taken:

1. Conversion to Appropriate Data Types:

Non-numeric entries were replaced with null values, and the columns were converted to numeric formats.

2. Validation:

Data type consistency across numeric columns was ensured, and anomalies were addressed through systematic cleaning.

This ensures the dataset's compatibility with statistical and machine learning models, which require correct data types.

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.**

-Issues Identified:

Some columns, such as Views, contained invalid entries (e.g., "invalid_data"), while others required scrutiny for inconsistencies.

Steps Taken:

1. Correction of Invalid Entries:

Invalid data entries were replaced with null values, and columns were converted to appropriate numeric formats.

2. Validation of Non-Numeric Columns:

Columns like Album were examined for numeric or irrelevant entries and confirmed to contain valid string data.

This ensures the integrity of both numeric and categorical data, critical for accurate analysis.

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?**

Steps Taken:

1. Identification of Duplicates:

Duplicate rows were identified based on unique identifiers and removed systematically.

2. Validation of Dependencies:

For dependent columns, duplicates were retained where necessary to preserve data relationships.

This ensures the dataset remains unique and accurate without losing critical 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.**

Steps Taken:

1. Reordering Columns:

Columns were logically reorganized to improve readability and accessibility, grouping related features together.

2. Renaming Columns:

Columns were renamed to better reflect their contents, ensuring clarity and precision.

These adjustments enhance the dataset's usability, facilitating smoother analysis and interpretation.

THANK YOU!

FOR YOUR TIME AND ATTENTION