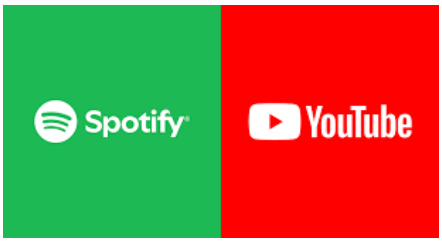
Power BI Project – YouTube + Spotify Dataset Evaluation through Power Query



Project Overview

This task focused on ensuring that the dataset containing information from Spotify and YouTube is clean and ready for further analysis. Our primary goals were to address issues such as missing values, inconsistent formatting, irrelevant data, and invalid entries, thus improving the dataset's usability. Below is a detailed, step-by-step breakdown of the cleaning and transformation process.

1. Identify and Handle Missing Values:

Question:

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 Data: - Stream (3%) - Channel (2%) - YouTube Info (2%) - Description (4%) - Views (11%) - Likes (12%) - Comments (3%) - Licensed (2%) - Official Video (2%)

Solution :

Official_Video, Views contain null values, there is nan i.e. not a number in Danceability and Energy column so we are replacing “nan” with "null". In this type of dataset we are not sure if the null values are zero or it is the data that is unavailable so instead of replacing null values with zero we are keeping the null values as it is because we don't know if the YouTube views likes or comments are hidden or if the data is just not available.

2. Fix Irregularities in Merged Columns:

Question :

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.

Solution :

Spotify_Info column contains the spotify album link as well as the spotify track id, they can be separated by splitting columns using the delimiter “|” at every occurrence. This way we'll have the Spotify_Album_Link and the Spotify_Track_Id column respectively.

Youtube_Info column contains the Youtube video album link as well the track title, they can be separated by splitting columns using the number delimiter 44 as youtube link in this dataset has 44 characters, so after splitting columns the youtube_track_link and youtube_track_title will be two different columns. Regarding suffixes in the columns, we can remove it by extracting text using a delimiter to extract text before suffix "-".

3. Correct Case Sensitivity and Naming Conventions:

Question :

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.

Solution :

For all column names to follow consistent format here we will transform the case to capitalizing each word for the column headers as well as the content in the columns for our personal convenience and to ensure consistent formatting in our dataset.
Hence the column names after consistent formatting would be : Track, Album, Album_Type, Spotify_Album_Link, Spotify_Track_Id, Duration_Ms, Stream, Key, Valence, Liveliness, Speechiness, etc.

4. Remove or Handle Irrelevant Columns:

Question :

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.

Solution :

There were two columns Random Column 1 and Random Column 2 which were

identified as randomly generated columns that are irrelevant in this dataset and won't be of any valid use for analysis. Deleted these two columns for our convenience.

5. Handle Inconsistent Data Types:

Question :

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.

Solution :

Danceability and energy are measured in the decimal format so to convert the text columns into decimal format will transform the format of the entire column and there were no errors while dealing with this sort of conversion.

6. Address and Fix Invalid Data Entries:

Questions :

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.

Solutions:

Album Column Validation: A thorough validation process ensured that all entries in the Album column were correct and free of numeric or irrelevant data.

7. Check for and Remove Duplicate rows.

Questions :

Identify and remove any duplicate rows in the dataset. How can you ensure that the remaining data is unique and accurate?

Solutions:

Using the "Keep Duplicates" feature: Before removing duplicates, use the “Keep Duplicates” feature in Power Query. This will allow you to see all rows that have duplicates and review which ones are flagged. This is helpful for checking which records are being considered as duplicates.
To identify duplicate roles from the data set we will select all the columns from the entire data set and remove the duplicate through the option provided to remove duplicates. After removing the duplicates, the remaining data should be unique and accurate.

8. Reorder and Rename Columns for Clarity:

Questions :

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.

Solution :

Here’s a logical sequence for the parameters based on categorization and flow of the dataset :

1. Track

2. Artist
3. Album
4. Album Type
5. Spotify Info
6. YouTube Info
7. Description
8. Duration MS
9. Stream
10. Views
11. Likes
12. Comments
13. Licensed
14. Channel
15. Key
16. Tempo
17. Loudness
18. Energy
19. Danceability
20. Acousticness
21. Instrumentalness
22. Liveliness
23. Speechiness
24. Balance

This sequence starts with the general information about the track, artist, and album, followed by platform-specific information (Spotify and YouTube), and then moves to detailed metrics and features related to audio analysis of the popular apps.

Conclusion:

After implementing these cleaning and transformation steps, the dataset is now free from inconsistencies, irrelevant information, and invalid entries. The data is standardized

and ready for analysis. We have successfully minimized the impact of missing or erroneous data, maintaining its integrity for meaningful insights.

Contact me:

LinkedIn:

https://www.linkedin.com/in/shlok-vashishth-307b9a15b/