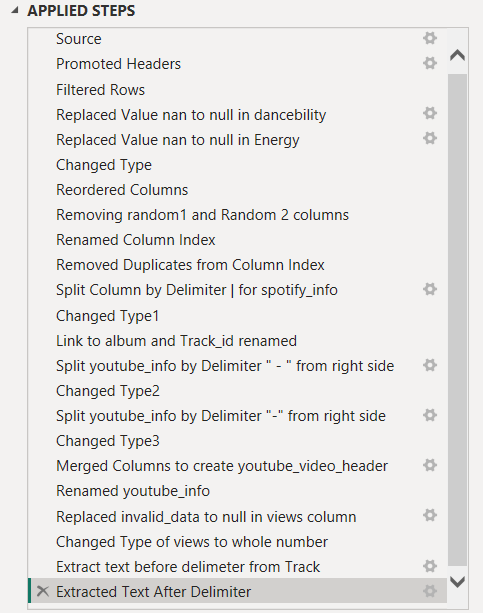
Spotify & YouTube Music

(Dataset cleaning and transformation Approach)

-by Shivani Nagar

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

* These are the steps that were applied while cleaning the data.



**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.

**Approach :**

**NOTE :** For null values – It is better to drop all the values if we are using the data for machine learning purpose.

Else,

* Basically, there are 2 rows for columns(Duration\_ms,key,valenceliveliness,speechiness,loudness,dancebility,instrumentalness,acousticness,energy,tempo) ----> But, the number of streams is given, and the number is quiet impressive and therefore we can not simply delete the rows, because for future purpose, we may want to find out average of streams by artist.
* For youtube\_info column : There are many blank values, but for the same row, spotify link is present. Must be the song is present on spotify but not on youtube. So, in this case we may split the tables for youtube and spotify information if we want visualisation separately for both platforms.
* For comments,views,likes :
* Replace invalid\_data to null in views.  
  We can find out views for the values when comments or likes are given. It can be calculated by taking out the percentage of likes/comments compared to views for an artist and then that values can be filled respectively.  
  And, For the rows where likes,comments and views all are null, we can keep them null, as it won’t mess up with the average(aggregate function) calculation.

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.

**Approach :**

* For column spotify\_info ---> Split by delimeter |
* For column youtube\_info ----> First, Split by delimeter “ – “ from right side. Then, again split first column youtube\_info by delimeter “-“ from right.
  + Then, merge both splitted columns by delimeter “-“ to get the youtube header.
* Change views type to whole number.

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.

**Approach :**

* Capitalise each word for column Track and artist.
* Rename the column headers to have a common format.

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.

**Approach :**

* Removed columns random\_1 and random\_2
* Replaced irrelevant data in columns like : Extracted text after delimeter : from track\_id column.  
  Also, removed duplicates from column index.

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.

**Approach :**

* Replaced nan value to null in dancebility and Energy column and then converted their type to whole number.
* Also, replace invalid\_data with null in views column.

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.

**Approach :**

* Use the replace values option from Transform tab and replace invalid\_data to null. It is done because views need to be in whole numbers format and invalid\_data entery would hamper the conversion and would raise errors for the same.
* In the Album column : 123456 is an entry that might be due to unavailability of data, we can replace such values with null or data not retrieved properly.

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?

**Approach :**

* Rename column unnamed to Index.
* Move the column in front and remove duplicates and null values if any,from this column.
* Now, this column will act as Primary key for the whole table and it is the property of primary key that it contains unique and not null values.

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.

**Approach :**

* Reordering the columns in such a way, to bring all the column related to spotify data in the beginning and all the youtube related data at the back.
* Example columns related to spotify : (Track, Album,Album\_type,Link\_to\_album,Track\_id,Duration\_ms,stream,key,valence,liveness,speechiness,loudness,dancebility…..)
* Example columns related to youtube : (Channel,Artist,Youtube\_info,Youtube\_video\_header,description,comments,likes,views,licensed,official\_video)

THANK YOU.