

The background of the image is a dark, atmospheric photograph of a live music concert. Red stage lights create sharp beams and rays against a black sky. A band is visible on stage, with one member in the foreground playing a guitar. The overall mood is energetic and focused on the music.

SPOTIFY & YOUTUBE MUSIC

CASE STUDY



ABOUT THE PROJECT



- Collected and analyzed a dataset containing music statistics from Spotify and YouTube Music.
- Focused on cleaning and preparing the dataset to ensure accuracy, consistency, and usability.
- Applied various data cleaning techniques, including:
 - Handling missing values
 - Splitting merged columns
 - Standardizing formats
 - Correcting data types
 - Removing irrelevant and duplicate entries
- The goal was to transform raw, messy data into a clean, structured dataset ready for further analysis and insights.

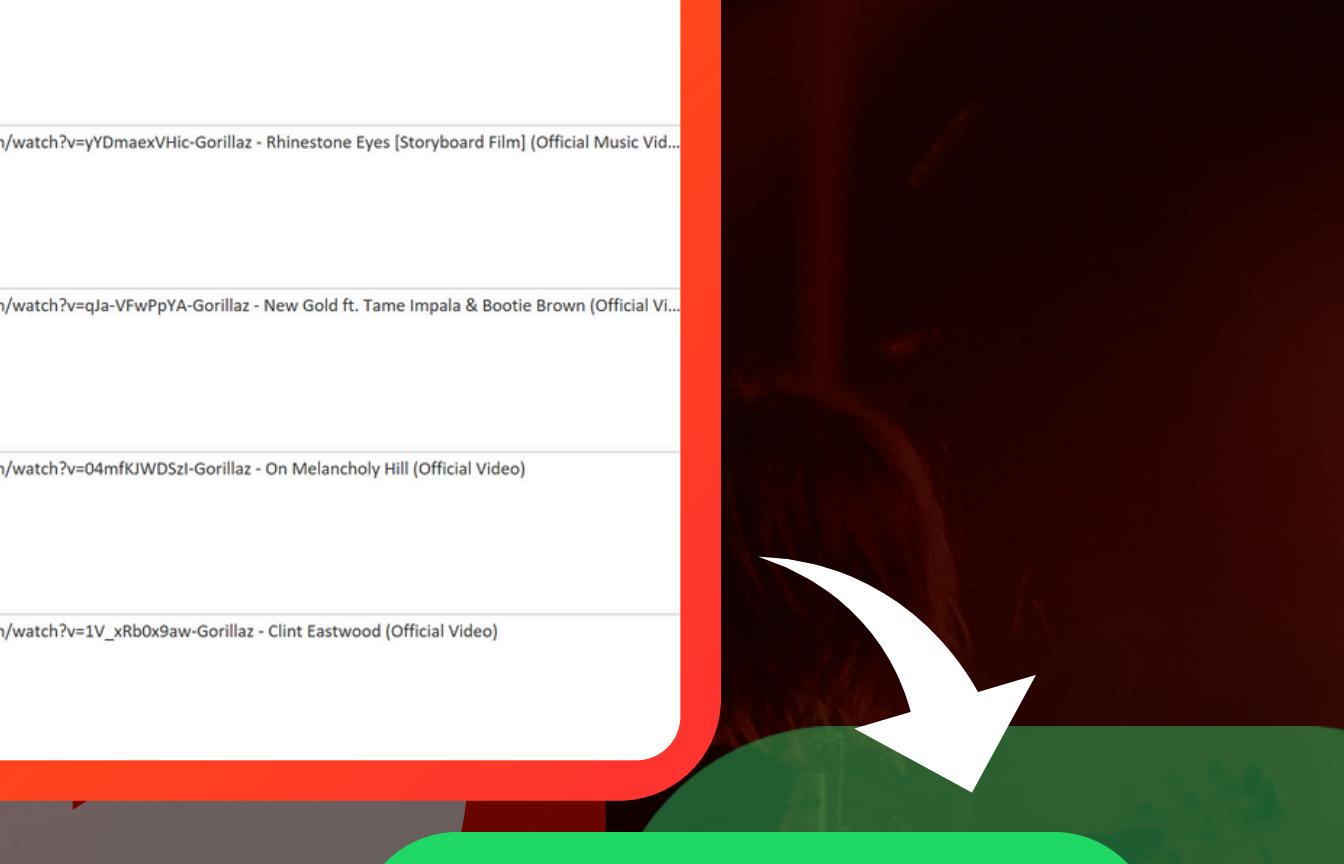




1. HANDLED MISSING VALUES

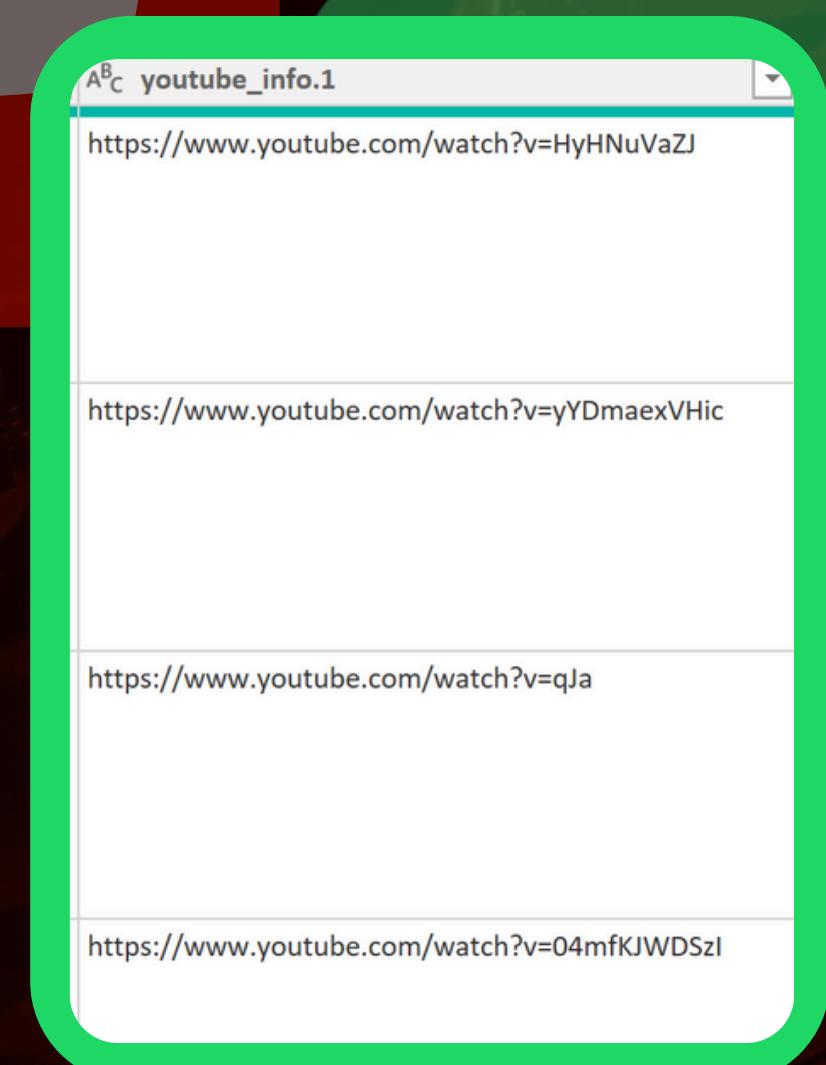
- Identified missing values in columns such as Views and Likes.
- Removed rows with few missing values and imputed with mean/median when missing data was significant.
- Ensured analysis accuracy and avoided bias from empty entries.





2. FIXED MERGED COLUMNS

- Split Spotify_Info and Youtube_Info into original components (e.g., Danceability, Energy, Tempo and Views, Likes, Comments).
- Removed unnecessary delimiters and extra characters.
- Organized the data to be clean and structured.





3. STANDARDIZED CASE AND NAMING

- Converted all column names to a consistent format (lowercase_with_underscores).
- Corrected artist and track names to have consistent casing (e.g., “ED SHEERAN” → “Ed Sheeran”).
- Improved readability and consistency of the dataset.

A ^B C artist	A ^B C trak	1.2 valence
ARTIST_GORILLAZ	FEEL GOOD INC._TRACK	0.772
ARTIST_GORILLAZ	RHINESTONE EYES_TRACK	0.852
ARTIST_GORILLAZ	NEW GOLD (FEAT. TAME IMPALA AND BOOTIE BROWN)_TRACK	0.551
ARTIST_GORILLAZ	CLINT EASTWOOD_TRACK	0.525
ARTIST_GORILLAZ	NEW GOLD (FEAT. TAME IMPALA AND BOOTIE BROWN) - DOM DOLLA ...	0.358



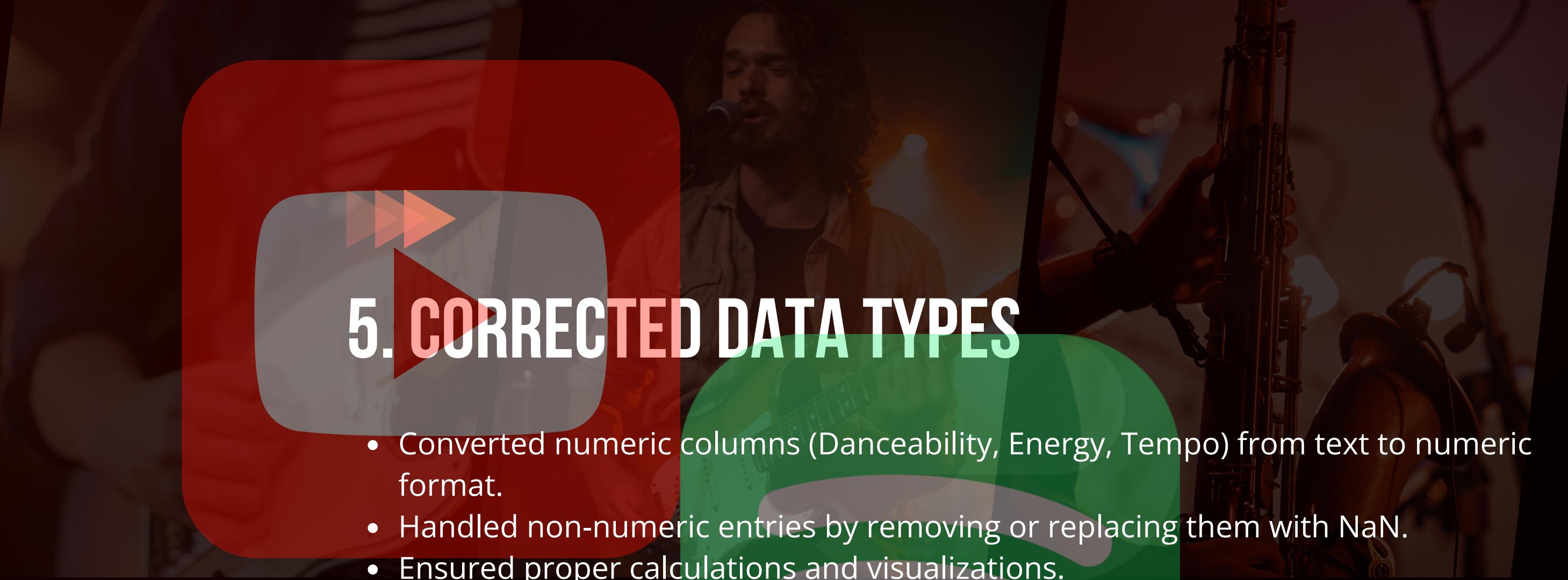
A	B	C	random_column_1	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	random_column_2	random_column_3
RANDOM																										376		
DATA																										112		
RANDOM																										969		
RANDOM																										605		
TEXT																										118		
DATA																										794		



4. REMOVED IRRELEVANT COLUMNS

- Deleted random or irrelevant columns such as Unnamed:0 and id_generated.
- Cleaned unusual values inside important columns.
- Kept only meaningful columns for analysis.





5. CORRECTED DATA TYPES

- Converted numeric columns (Danceability, Energy, Tempo) from text to numeric format.
- Handled non-numeric entries by removing or replacing them with NaN.
- Ensured proper calculations and visualizations.



6. FIXED INVALID ENTRIES

- Replaced invalid Views values like invalid_data with NaN, then imputed using statistical measures.
- Checked the Album column to ensure no numeric or irrelevant entries existed.
- Corrected all values to be valid and logical.



7. REMOVED DUPLICATES

- Used `drop_duplicates()` to detect and remove duplicate rows.
- Ensured the dataset contained only unique and accurate records.



8. REORDERED AND RENAMED COLUMNS

- Reorganized columns logically:
- Metadata: artist_name, track_name, album
- Spotify stats: danceability, energy, tempo...
- YouTube stats: views, likes, comments...
- Renamed unclear columns to descriptive names (e.g., yt_views instead of vws).
- Enhanced clarity and usability of the dataset.





HUDA HOSAWEE