Spotify

Music

BEHIND THE MUSIC: CLEANING AND STRUCTURING STREAMING DATA

PRESENTED BY ADNAN HAFEEZ

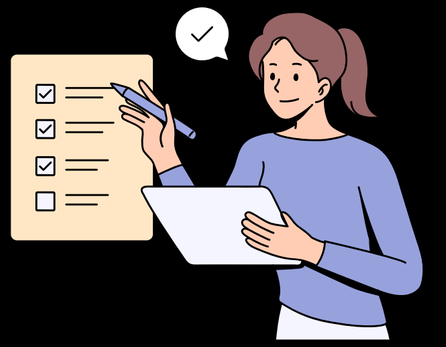
ABOUT PROJECT

This project was created with the goal of transforming messy, inconsistent music streaming data into a clean and structured format, ready for meaningful analysis and business intelligence. Focusing exclusively on data from Spotify and YouTube Music, we used Power BI's Power Query Editor to address common data quality issues such as missing values, inconsistent formats, invalid entries, and structural anomalies.

Our approach was driven by a deep understanding of the importance of data integrity in the music streaming industry—where millions of records influence decisions about user behavior, artist performance, and content strategy. By cleaning and standardizing the dataset, we've built a strong foundation for accurate reporting, visualization, and data-driven insights.

This project reflects our commitment to data accuracy, usability, and storytelling through clean data. Whether you're a data analyst, music industry researcher, or BI professional, our work offers a practical example of preparing raw streaming data for analytical success.

DATA CLEANING



IDENTIFY AND HANDLE MISSING VALUES

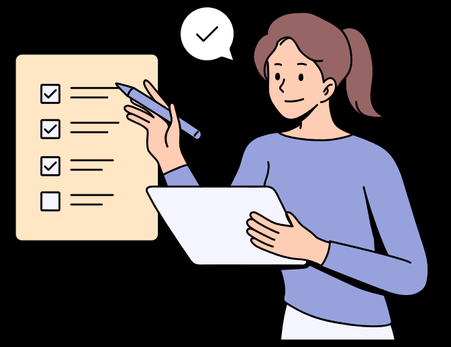
Column Name	Total Missing Values (Based on Column Profiling)
Likes	2685
Views	2484
Description	911
Stream	610
Comments	593
Licensed	491
Official Video	491
Channel	491
Youtube_Info	491
Key	2
Valence	2
Liveness	2
Speechiness	2
Loudness	2
Tempo	2
Danceability	2
Duration_Ms	2
Instrumentalness	2
Acousticness	2

- DIRTY DATA— LIKE MISSING VALUES, TYPOS, OR INCONSISTENT FORMATS—CAN LEAD TO INCORRECT INSIGHTS. CLEANING ENSURES THE DATASET TRULY REFLECTS REAL-WORLD BEHAVIORS

- STAKEHOLDERS RELY ON DATA FOR DECISIONS. CLEAN DATA ENSURES THEY BASE THOSE DECISIONS ON FACTS, NOT FLAWS.

Music

DATA CLEANING



HANDLING MISSING VALUES IN LIKES AND VIEWS

COLUMN - VIEWS

ABOUT

THE NUMBER OF TIMES A VIDEO OR TRACK HAS BEEN PLAYED—CRITICAL FOR MEASURING POPULARITY AND AUDIENCE REACH.

HANDLING APPROACH

SOLUTION

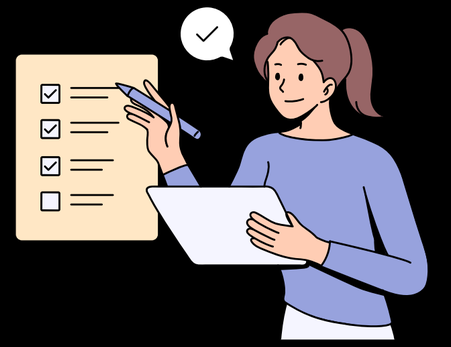
- MISSING VALUES WERE FILLED WITH 0, ASSUMING THE CONTENT HASN'T RECEIVED ANY VIEWS YET.

JUSTIFICATION

- PRESERVES ALL ROWS IN THE DATASET, AVOIDING DATA LOSS.
- A "0" IS A CLEAR, INTERPRETABLE VALUE MEANING "NO VIEWS."
- HELPS MAINTAIN ACCURACY IN STATISTICAL SUMMARIES (SUM, AVERAGES).
- REFLECTS LOW OR NO ENGAGEMENT RATHER THAN UNDEFINED DATA.

Music

DATA CLEANING



HANDLING MISSING VALUES IN LIKES AND VIEWS

COLUMN - LIKES

ABOUT

INDICATES USER APPROVAL OR ENJOYMENT OF CONTENT—AN IMPORTANT ENGAGEMENT METRIC.

HANDLING APPROACH

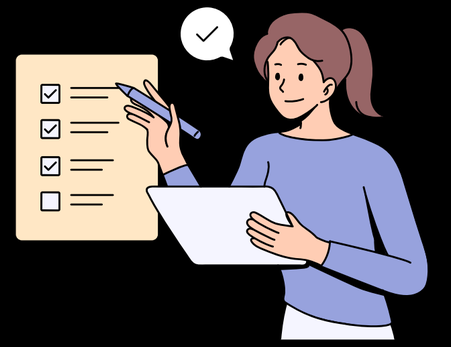
SOLUTION

- MISSING VALUES WERE ALSO FILLED WITH 0, IMPLYING THAT NO LIKES HAVE BEEN RECORDED YET.

JUSTIFICATION

- ENSURES DATA COMPLETENESS WITHOUT REMOVING CONTEXT FROM OTHER COLUMNS.
- 0 CLEARLY COMMUNICATES A LACK OF INTERACTION.
- SUPPORTS MEANINGFUL AND UNDISTORTED ANALYTICAL SUMMARIES.

DATA CLEANING



FIX IRREGULARITIES IN MERGED COLUMNS

COLUMN - SPOTIFY_INFO

ABOUT

SPOTIFY_INFO IS A LEADING DIGITAL MUSIC STREAMING SERVICE THAT PROVIDES USERS WITH ACCESS TO A VAST LIBRARY OF SONGS, PODCASTS, AND OTHER AUDIO CONTENT.

JUSTIFICATION

- CHOOSING DELIMITERS FOR SPLITTING MERGED DATA IN THE SPOTIFY_INFO COLUMN (OR ANY OTHER COLUMN CONTAINING COMBINED DATA) IS A KEY PART OF THE DATA CLEANING AND PREPROCESSING PROCESS.

HANDLING APPROACH

SOLUTION

- SPLITTING THE COLUMNS BY DELIMITERS: CHOOSING THE PIPE (|) DELIMITER FOR SPLITTING MERGED DATA IS A STRATEGIC DECISION BASED ON SEVERAL PRACTICAL AND TECHNICAL FACTORS.

Music

DATA CLEANING

FIX IRREGULARITIES IN MERGED COLUMNS

COLUMN – YOUTUBE_INFO

ABOUT

YOUTUBE IS ONE OF THE WORLD'S LARGEST AND MOST INFLUENTIAL VIDEO-SHARING PLATFORMS, ALLOWING USERS TO UPLOAD, VIEW, AND SHARE VIDEOS SPANNING A WIDE VARIETY OF GENRES, TOPICS, AND INTERESTS.

JUSTIFICATION

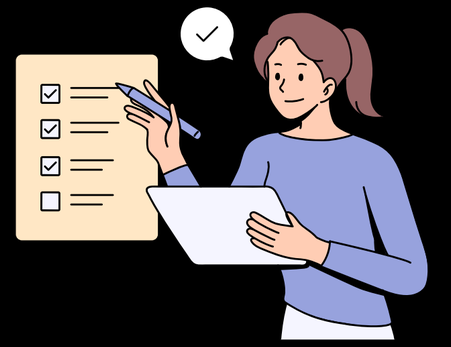
- USING A FIXED NUMBER OF CHARACTERS TO SPLIT THE DATA CAN BE HELPFUL WHEN THE DATA FOLLOWS A CONSISTENT PATTERN OR FORMAT WHERE COMPONENTS HAVE PREDICTABLE LENGTHS.

HANDLING APPROACH

SOLUTION

- SPLITTING THE COLUMNS BY NUMBER OF CHARACTERS: FIXED-LENGTH DATA IF THE FIRST PART OF THE MERGED DATA (E.G., A URL) ALWAYS HAS THE SAME NUMBER OF CHARACTERS, SPLITTING BY A FIXED CHARACTER LIMIT CAN RELIABLY SEPARATE COMPONENTS.

DATA CLEANING



CORRECT CASE SENSITIVITY AND NAMING CONVENTIONS

CLEANING APPROACH

- CONVERTED ALL COLUMN NAMES TO LOWERCASE.
- REPLACED SPACES, COLONS, AND SPECIAL CHARACTERS WITH UNDERSCORES FOR READABILITY AND CONSISTENCY.

🎵 ARTIST & TRACK FORMATTING

- STANDARDIZED ARTIST AND TRACK COLUMNS USING TITLE CASE (E.G., GORILLAZ → GORILLAZ) VIA:
- TRANSFORM → FORMAT → CAPITALIZE EACH WORD

Music

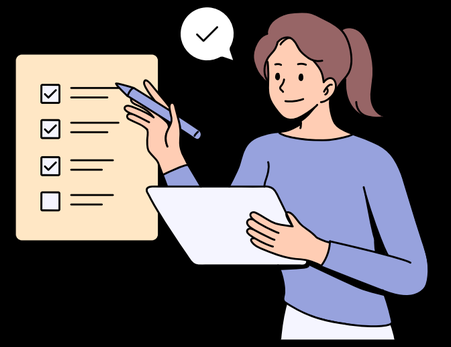
DATA CLEANING

CORRECT CASE SENSITIVITY AND NAMING CONVENTIONS

WHY STANDARDIZE CASE & NAMING CONVENTIONS ?

- PROMOTES CLEAN, CONSISTENT, AND ANALYSIS-READY DATA.
- ENSURES COMPATIBILITY WITH POWER BI, SQL, AND OTHER TOOLS.
- ENHANCES CLARITY IN VISUAL REPORTS AND DASHBOARDS.

DATA CLEANING



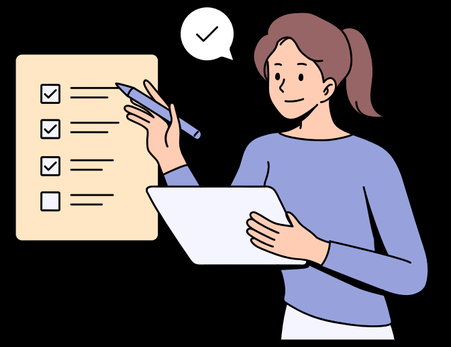
REMOVE OR HANDLE IRRELEVANT COLUMNS

THE DATASET INCLUDES SEVERAL COLUMNS THAT ARE RANDOMLY GENERATED, OR IRRELEVANT TO THE PROJECT'S ANALYTICAL GOALS.

CRITERIA FOR REMOVAL

- COLUMNS THAT WERE DERIVED INTO MORE USEFUL FEATURES (E.G., SPLITTING YOUTUBE_INFO).
- FIELDS CONTAINING RANDOM OR NON-INFORMATIVE DATA.
- COLUMNS WITH OVER 50% MISSING DATA AND NO STRONG ANALYTICAL VALUE.

DATA CLEANING



HANDLE INCONSISTENT DATA TYPES

SOME COLUMNS EXPECTED TO HOLD NUMERIC VALUES (E.G., DANCEABILITY, LOUDNESS) ARE INCORRECTLY STORED AS TEXT, WHICH AFFECTS ANALYSIS, SORTING, AND CALCULATIONS IN POWER BI.

APPROACH

- IDENTIFIED NUMERIC COLUMNS STORED AS TEXT (E.G., DANCEABILITY, ENERGY, TEMPO).
- REPLACED INVALID ENTRIES (E.G., "N/A", BLANKS) WITH NULL OR 0 AS APPROPRIATE.
- CONVERTED TO CORRECT TYPES USING POWER QUERY → TRANSFORM → DATA TYPE.

Music

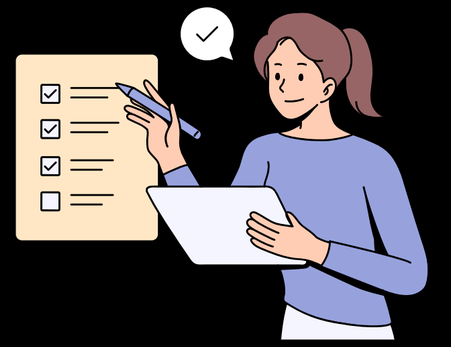
DATA CLEANING

HANDLE INCONSISTENT DATA TYPES

WHY DEALING WITH INCONSISTENT DATA TYPES IS IMPORTANT?

- ENSURES ACCURATE CALCULATIONS AND VISUALIZATIONS.
- ENABLES USE OF NUMERIC FUNCTIONS LIKE AVERAGES, SUMS, AND COMPARISONS.
- PREVENTS ERRORS IN CHARTS, KPIS, AND AGGREGATIONS IN POWER BI.

DATA CLEANING



ADDRESS AND FIX INVALID DATA ENTRIES

VIEWS AND ALBUM COLUMN

- REPLACED "INVALID_DATA" AND OTHER NON-NUMERIC VALUES WITH 0.

- REMOVED ENTRIES THAT WERE PURELY NUMERIC OR IRRELEVANT (E.G., "12345").

- ENSURED ALL ENTRIES ARE TEXTUAL AND MEANINGFUL ALBUM TITLES AND USED DATA FILTERS AND POWER QUERY TEXT FUNCTIONS TO DETECT AND CLEAN ANOMALIES.

Music

DATA CLEANING

ADDRESS AND FIX INVALID DATA ENTRIES

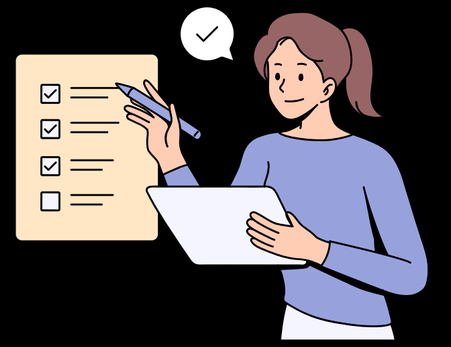
BENEFITS

- IMPROVES DATA ACCURACY: ENSURES METRICS LIKE VIEWS ARE CORRECTLY REPRESENTED AND NOT SKEWED BY TEXT ERRORS.

- ENABLES RELIABLE ANALYSIS: CLEAN ALBUM AND VIEWS DATA SUPPORTS GROUPING, FILTERING, AND AGGREGATIONS IN POWER BI.

- SUPPORTS SMOOTH DATA MODELING: CONSISTENT DATA TYPES REDUCE TRANSFORMATION COMPLEXITY AND ERRORS DURING MODEL CREATION.

DATA CLEANING



CHECK FOR AND REMOVE DUPLICATE ROWS

DUPLICATE ROWS CAN:

- SKEW ANALYSIS (E.G., INFLATED VIEWS/LIKES)
- MISREPRESENT UNIQUE CONTENT
- AFFECT DASHBOARD ACCURACY

APPROACH

- USED POWER QUERY EDITOR → REMOVE ROWS → REMOVE DUPLICATES
- CHECKED ACROSS KEY IDENTIFIERS SUCH AS: TRACK, ARTIST, ALBUM, VIEWS, LIKES, DURATION_MS
- VERIFIED WITH ROW COUNT BEFORE AND AFTER CLEANUP

Music

DATA CLEANING

CHECK FOR AND REMOVE DUPLICATE ROWS

BENEFITS

- IMPROVES DATA INTEGRITY: ENSURES EACH RECORD REPRESENTS UNIQUE CONTENT
- PREVENTS METRIC INFLATION: AVOIDS DOUBLE-COUNTING VIEWS, LIKES, OR STREAMS
- ENHANCES ANALYSIS ACCURACY: PRODUCES MORE RELIABLE AGGREGATIONS AND INSIGHTS
- OPTIMIZES PERFORMANCE: REDUCES UNNECESSARY DATA LOAD FOR FASTER VISUALS IN POWER BI

DATA CLEANING

REORDER AND RENAME COLUMNS FOR CLARITY

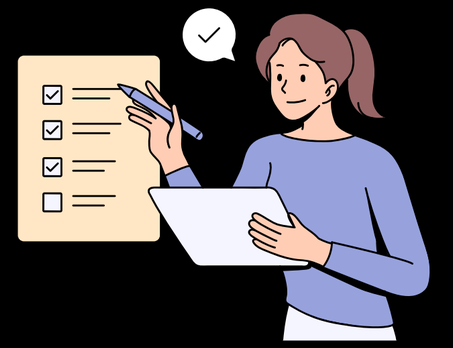
REORDERING COLUMNS

REARRANGED COLUMNS INTO A LOGICAL SEQUENCE FOR IMPROVED READABILITY AND ANALYSIS FLOW:

1. TRACK, ARTIST, ALBUM
2. DURATION_MS, VIEWS, LIKES, COMMENTS
3. DANCEABILITY, ENERGY, ACOUSTICNESS, INSTRUMENTALNESS, LIVENESS, VALENCE, SPEECHINESS, LOUDNESS, TEMPO
4. ADDITIONAL FLAGS LIKE OFFICIAL_VIDEO, LICENSED, DESCRIPTION (IF RETAINED)

Music

DATA CLEANING



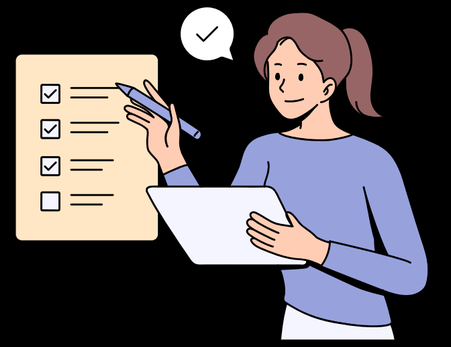
REORDER AND RENAME COLUMNS FOR CLARITY

RENAMING COLUMNS

- STANDARDIZED ALL COLUMN NAMES TO LOWERCASE WITH UNDERSCORES FOR CONSISTENCY (E.G., TRACK NAME → TRACK, DURATION (MS) → DURATION_MS).

- REMOVED SPECIAL CHARACTERS AND WHITESPACE TO ALIGN WITH NAMING BEST PRACTICES USED IN DATA MODELING AND ANALYTICS TOOLS.

DATA CLEANING



ADDRESS AND FIX INVALID DATA ENTRIES

BENEFITS

- IMPROVES READABILITY: EASY TO NAVIGATE AND INTERPRET THE DATASET.
- ENHANCES USABILITY: LOGICAL ORDER HELPS USERS FIND KEY FIELDS QUICKLY.
- CONSISTENCY: CLEAN, UNIFORM NAMES ENSURE BETTER COMPATIBILITY ACROSS TOOLS (E.G., POWER BI, SQL).
- FACILITATES ANALYSIS: STREAMLINED STRUCTURE SUPPORTS MORE EFFICIENT DASHBOARDING AND STORYTELLING.

THANK YOU !

Ultimately, this cleaned dataset offers a solid foundation for exploring user engagement trends across Spotify and YouTube Music. By applying these cleaning practices, we enable deeper, data-driven storytelling and empower more confident decision-making in future analytics projects.

Next Slide