Netflix Content Strategy Analyzer – Data Cleaning & Insights

1. Data Cleaning Steps

- 1.1 Load the Dataset: Import the dataset using pandas.read_csv(). If the file is hosted online, use its direct URL.
- 1.2 Remove Duplicates: Duplicate records can lead to incorrect insights. Use drop_duplicates() to ensure each record is unique.
- 1.3 Handle Missing Values: Missing data can distort analysis. Use SimpleImputer to replace missing numeric values with the mean and categorical values with the most frequent value.
- 1.4 Verify Data Types: Convert columns to correct data types (numeric, object, datetime) so operations work as expected.
- 1.5 Remove Irrelevant Columns: Drop columns not required for analysis (e.g., unnecessary IDs, URLs).
- 1.6 Standardize Text Data: Convert all text data to lowercase and strip whitespace for uniformity.

2. Key Metrics and Insights

- 2.1 Total Titles Available: The overall count of titles (movies + TV shows). This gives the size of Netflix's library.
- 2.2 Content Type Distribution: Measure the proportion of movies vs TV shows to understand platform focus.
- 2.3 Top Countries: Identify the top 5 countries contributing most to Netflix's library.
- 2.4 Popular Genres: Analyze the listed_in column to find which genres dominate the library.
- 2.5 Recent Release Trends: Observe the number of titles released year by year to study Netflix's content growth.
- 2.6 Ratings Distribution: Count how many titles fall under each rating (PG, R, TV-MA, etc.) to understand audience targeting.

3. Normalization for Categorical Data

- 3.1 Identify Categorical Columns: Use df.select_dtypes(include=['object']) to list all categorical features.
- 3.2 Apply Encoding Methods: Different techniques are available to transform categorical values into numeric form.
- 3.3 Merge Encoded Data: Drop original categorical columns and concatenate the encoded DataFrame to get the final numeric dataset.

3.4 Types of Categorical Encodings

- Label Encoding: Assigns each unique category an integer value. Suitable for ordinal data but may mislead models when used on nominal data since it imposes an order.
- One-Hot Encoding: Creates binary columns (0/1) for each category. Prevents ordinal relationships but can increase dataset dimensionality significantly when categories are many.
- Ordinal Encoding: Maps categories to integers based on a defined order. Works well for ordered features like ratings (e.g., Low < Medium < High).
- Frequency Encoding: Replaces categories with their frequency count in the dataset. Useful for high-cardinality categorical features.

4. Benefits of Cleaning & Normalization

- Ensures accurate and trustworthy insights.
- Prepares the dataset for machine learning and analytics.
- Avoids errors caused by null values or inconsistent categories.
- Improves model accuracy and performance.

5. Conclusion

By following a structured data cleaning and normalization process, we create a reliable Netflix dataset ready for analytics and machine learning. The encoding techniques allow categorical data to be efficiently transformed into numeric form, enabling models to interpret them correctly. This serves as the foundation for building a robust content strategy analyzer.