Netflix Content Strategy Analyzer – Data Cleaning & Insights

1. Data Cleaning Steps

Data cleaning is the first and most critical step in building a reliable Netflix content strategy analyzer. Cleaning ensures that the dataset is consistent, accurate, and ready for analysis. Below are the detailed 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

Once the data is cleaned, we calculate key performance metrics that will help Netflix optimize its content strategy. These metrics provide actionable insights into the type of content available, country distribution, and audience preferences.

- **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

Machine learning models work best with numeric data. Categorical columns (like country, type, and rating) must be encoded. We use One-Hot Encoding to convert text categories into numeric dummy variables.

- **3.1 Identify Categorical Columns:** Use df.select_dtypes(include=['object']) to list all categorical features.
- **3.2 Apply One-Hot Encoding:** Use OneHotEncoder(sparse_output=False, handle_unknown='ignore') to transform each category into binary columns.
- **3.3 Merge Encoded Data:** Drop original categorical columns and concatenate the encoded DataFrame to get the final numeric dataset.

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 key metrics provide valuable business insights such as which countries and genres to invest in, what content types are in demand, and how Netflix’s content library is evolving over time. This serves as the foundation for building a robust content strategy analyzer.