Netflix Content Strategy Analyzer

Data Cleaning, Normalization, Encoding, EDA, and Feature Engineering Steps

1. Data Cleaning

Data cleaning ensures the dataset is accurate and consistent before analysis. Key steps include:

• **Handling Missing Values:** Replace missing categorical data with mode or “Unknown”; drop rows with too many missing values.
• **Removing Duplicates:** Identify and remove duplicate rows based on title, type, and release year.
• **Correcting Data Types:** Convert columns such as ‘release_year’ to integers and ‘date_added’ to datetime format.
• **Trimming Whitespaces:** Remove unnecessary spaces and standardize text fields like director, cast, and country.
• **Handling Inconsistencies:** Standardize genres and countries (e.g., “United States” vs “USA”).

2. Normalization

Normalization ensures uniform scaling and consistency in the dataset.

• **Text Normalization:** Convert text to lowercase for uniformity.
• **Duration Standardization:** Separate duration into numeric values and units (e.g., “90 min”, “3 Seasons”).
• **Country Grouping:** Normalize country names and group rare ones into “Other”.
• **Rating Normalization:** Standardize maturity ratings (e.g., “TV-MA” vs “TV-Ma”).

3. Encoding

Encoding converts categorical variables into numeric formats.

• **Label Encoding:** Convert binary variables such as “type” (Movie=0, TV Show=1).
• **One-Hot Encoding:** Use for multi-category columns like genres, ratings, and countries.
• **Frequency Encoding:** Replace categories with their frequency counts (useful for directors, genres).
• **Ordinal Encoding:** Apply to ordered data like ratings (G < PG < PG-13 < R < TV-MA).

4. Exploratory Data Analysis (EDA)

EDA identifies major insights and trends in Netflix’s content strategy.

• **Content Growth Over Time:** Plot number of titles added per year.
• **Genre Analysis:** Identify the most common and trending genres.
• **Ratings Distribution:** Visualize different maturity ratings.
• **Type Analysis:** Compare number of movies vs TV shows.
• **Country Analysis:** Examine which countries contribute most content.
• **Correlations:** Study relationships between release year, duration, and genre trends.

5. Feature Engineering

Feature Engineering transforms raw data into valuable features for analysis.

- **Content Length Category:** Categorize duration as Short, Medium, Long.
- **Release Decade:** Derive new feature grouping release years by decade.
- **Original vs Licensed:** Detect Netflix Originals from title or description keywords.
- **Primary Genre:** Extract first-listed genre for simplification.
- **Country Region:** Group countries into major continents or regions.
- **Temporal Features:** Extract month and year from “date_added”.
- **Popularity Indicators:** Use director or actor frequency as a proxy for popularity.

Conclusion

By following these preprocessing and analytical steps, the Netflix dataset becomes structured and ready for detailed visualization, trend discovery, and machine learning analysis. These processes are essential to understanding Netflix’s global content strategy and market behavior.