Jaya Priya J

Netflix Insights

Data Cleaning Steps (using Pandas)

During preprocessing, the following data cleaning steps were applied to the Netflix dataset:

1. **Handling Missing Values and Unknowns**
   - Replaced all "Unknown" values with proper NaN.
   - Checked for missing values in each column.
2. **Duplicate Removal**
   - Dropped duplicate rows to ensure unique records.
3. **Date Conversion**
   - Converted the date_added column to **datetime** format for time-based analysis.
4. **Text Standardization**
   - Removed leading/trailing whitespaces from title.
   - Converted listed_in (genres) to **lowercase** for consistency.
   - Normalized rating column (stripped spaces and converted to uppercase).
5. **Feature Engineering**
   - Extracted numeric values and units from the duration column (e.g., "90 min" → 90 + "min", "2 Seasons" → 2 + "seasons").
   - Created new columns: duration_value and duration_unit.

Metrics Used in Your Code

1. **Content Type Distribution**;

   → Count of **Movies vs TV Shows**.
2. **Yearly Release Trend:**

   → Number of titles released each year.
3. **Top Genres (Categories):**

   → Top 10 most frequent genres/categories.
4. **Country Distribution**:

   → Top 10 countries contributing content.

5. **Duration Analysis**
    - For Movies → runtime in minutes (duration_value).
    - For TV Shows → number of seasons.