Netflix Dataset Analysis Report

1. Dataset Overview

The Netflix dataset contains information about movies and TV shows, including attributes such as title, director, cast, country, release year, rating, duration, and genres. Most fields are categorical, some are numeric, and one is date-time (date added).

2. Missing Values

Director and Cast are frequently missing. Country is missing for some records, especially older titles. Date Added is missing in certain entries, which may affect trend analysis.

3. Content Distribution

Type: Most titles are movies (≈70%), while TV shows account for the rest. Release Year: A majority of titles were released after 2010, with rapid growth after 2015. Ratings: The most frequent ratings are TV-MA, TV-14, PG, and R. Movies are often PG or R, while TV shows are mostly TV-MA or TV-14. Duration: Movies are measured in minutes (mostly between 80–120). TV Shows are measured in seasons (usually 1–3).

4. Country Insights

The United States contributes the most titles, followed by India, the UK, and Canada. Regional differences are visible: Indian titles lean toward drama and romance, while US titles focus more on documentaries and comedies.

5. Genre Trends

The most common genres are drama, comedy, action, documentary, and international films. Many titles span multiple genres (e.g., drama + thriller).

6. Growth Patterns

The number of titles added each year increased significantly after 2015. Kids’ shows and reality TV are common among TV series, while action and drama dominate movies.

7. Feature Engineering

To prepare the dataset for modeling, several transformations were performed: - Label Encoding: Converted categorical values such as ratings into numeric codes. - Frequency Encoding: Replaced countries with their frequency counts. - One-Hot Encoding for Genres: Split multi-genre values into separate binary columns. - Date Features: Extracted year and month from date added. - Duration Normalization: Converted durations into numeric fields (minutes for movies, seasons for TV shows). - Cast/Director Features: Created binary flags for frequent actors or directors (optional).

8. Normalization

Normalization was applied to transform categorical and textual data into a structured, numeric-friendly format: 1. Categorical Features: Ratings and countries were label encoded (e.g., PG → 3, R → 4). Countries were also frequency encoded (e.g., US = 3000, India = 2000). 2. Genres: One-hot encoding converted multi-genre entries into binary columns (e.g., Drama, Action → Drama = 1, Action = 1). 3. Date Features: Extracted year_added and month_added for temporal analysis. 4. Duration: Converted text-based durations into numeric values (Movies: '90 min' → 90; TV Shows: '2 Seasons' → 2). 5. Cast/Director Features (optional): Popular actors/directors were encoded into binary fields (e.g., has_Spielberg = 1).

9. Summary

The analysis shows that Netflix is dominated by movies, though TV shows have grown in recent years. The US and India are the biggest contributors, with ratings centered around TV-MA, TV-14, and PG. Drama and comedy are the most frequent genres. Feature engineering and normalization prepared the dataset by encoding categorical values, splitting genres, normalizing durations, and extracting time features. This transformed dataset is now well-suited for predictive modeling, clustering, or recommendation systems.