Netflix Dataset Insights and Data Cleaning Report

# 1. Introduction

This document presents insights derived from the Netflix dataset along with the detailed data cleaning steps performed using Pandas in a Databricks environment. The goal of this analysis is to prepare a cleaned dataset for reliable insights and further modeling.

# 2. Data Cleaning Steps

## 2.1 Remove Duplicates

- Removed duplicate rows to ensure unique records.  
- Specifically dropped duplicates based on 'title' and 'release\_year'.

## 2.2 Handle Missing Values

- Replaced missing values in 'director', 'cast', 'country', and 'rating' with 'Unknown'.

## 2.3 Converted Dates

- Converted 'date\_added' to datetime format, coercing errors to NaT.  
- Created a binary column 'date\_missing' to indicate missing dates.

## 2.4 Exploded Multi-Value Columns

- Split listed\_in (genres) into multiple rows for better analysis of each genre separately.

## 2.5 Handled Outliers in Duration

- Cleaned *duration* by removing text like "min" and "Season(s)".

## 2.6 Column Transformation and Normalization

- Removed extra spaces in type.

- Mapped ratings into groups: *Kids, Family, Teens, Adults, Unknown*.

- Standardized country names (e.g., USA → United States).

- Used one-hot encoding for type.

- Grouped rare countries (appearing < 20 times) into "Other".

# 3. Insights from the Dataset

- Movies dominate the Netflix library compared to TV Shows.  
- United States and India are the top content-producing countries.  
- Drama, International Movies, and Comedies are the most common genres.  
- Significant proportion of content is rated for Teens and Adults.  
- The number of releases has increased significantly in recent years, peaking after 2015.

# 4. Cleaned Dataset Output

- The cleaned dataset was saved at: /Volumes/workspace/default/netflix/cleaned\_netflix.csv  
- This dataset is now ready for exploratory data analysis (EDA) and predictive modeling.