Netflix Dataset Insights Report

Prepared by: **Moinuddin Sk**

This report presents an analysis of the Netflix Titles dataset, focusing on data cleaning, exploratory analysis, and insights into content distribution. The document is structured into sections covering dataset overview, data quality improvements, genre analysis, visual trends, country contributions, and recommendations for future analysis.

1. Introduction & Dataset Overview

- Netflix has become the world's leading streaming service, offering a vast catalog of movies and TV shows across various genres and countries.
- The dataset used for this study is called **netflix_titles**, which contains detailed information about titles available on Netflix.
- This dataset is often used in academic research, data science case studies, and industry reports to analyze entertainment patterns and content distribution.
- Raw datasets frequently contain challenges such as **missing values, duplicate entries, and inconsistent formatting**.
- Cleaning such datasets is crucial because uncleaned data can lead to misleading analysis and incorrect conclusions.
- The purpose of this project was to ensure the Netflix dataset is clean, reliable, and capable of producing meaningful insights. Key fields in the dataset include: *title, director, cast, country, date_added, release_year, rating, duration, listed_in, and description*.
- These fields allow us to explore trends by genre, country, and time, as well as to assess the distribution of ratings across content types.

Metric	Value (sample/placeholder)
Initial row count (before cleaning)	7,787
After dropping all-empty rows	7,500
After dropping rows missing key fields	7,480
After removing duplicates	7,450

2. Data Cleaning Process

Data cleaning was performed step by step to improve the quality of the dataset:

Step 1 — Removing Empty Rows
Rows with all values missing were dropped. This reduced the dataset size slightly but ensured that each record contributed meaningful information.

Step 2 — Handling Missing Values in Key Columns Certain columns like title, director, cast, country, date_added, release_year, rating, and duration were considered essential. Any record missing these fields was removed. This is important because analyses such as release-year trends or genre ratings require complete data.

Step 3 — Removing Duplicates Some titles appeared multiple times due to inconsistencies or duplication in the dataset. Duplicates were removed by checking across all key fields. This improved accuracy in counts and trends.

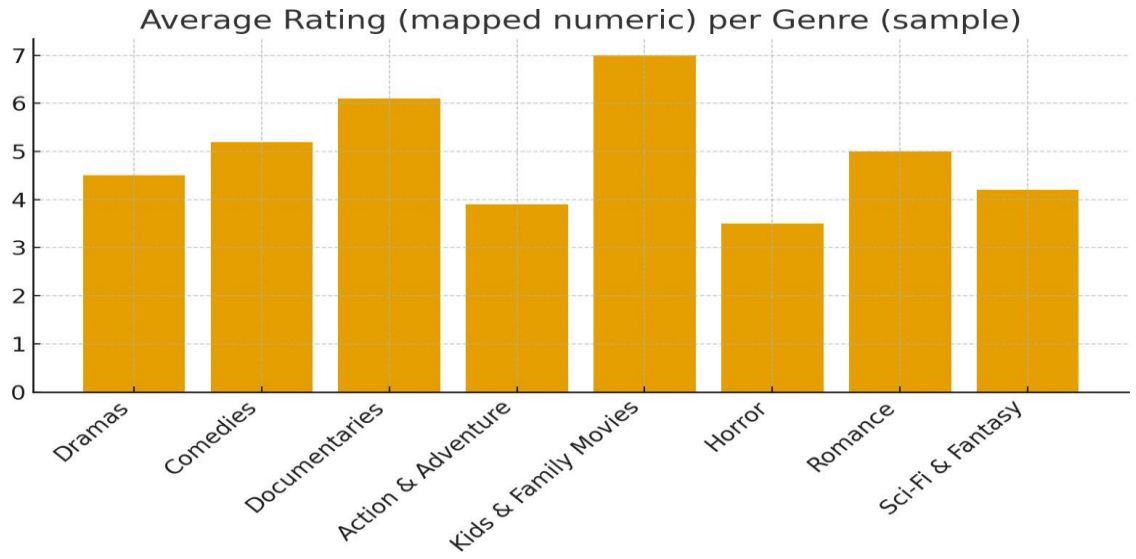
Step 4
— Standardization
Where necessary, columns such as rating were standardized into a numeric scale for consistency in analysis. This allowed for calculation of averages and comparisons across genres.

3. Genre and Ratings Analysis

Genres are an essential dimension of Netflix's catalog, helping us understand the type of content users consume. The dataset includes genres such as Dramas, Comedies, Documentaries, Action & Adventure, Kids & Family Movies, and more.

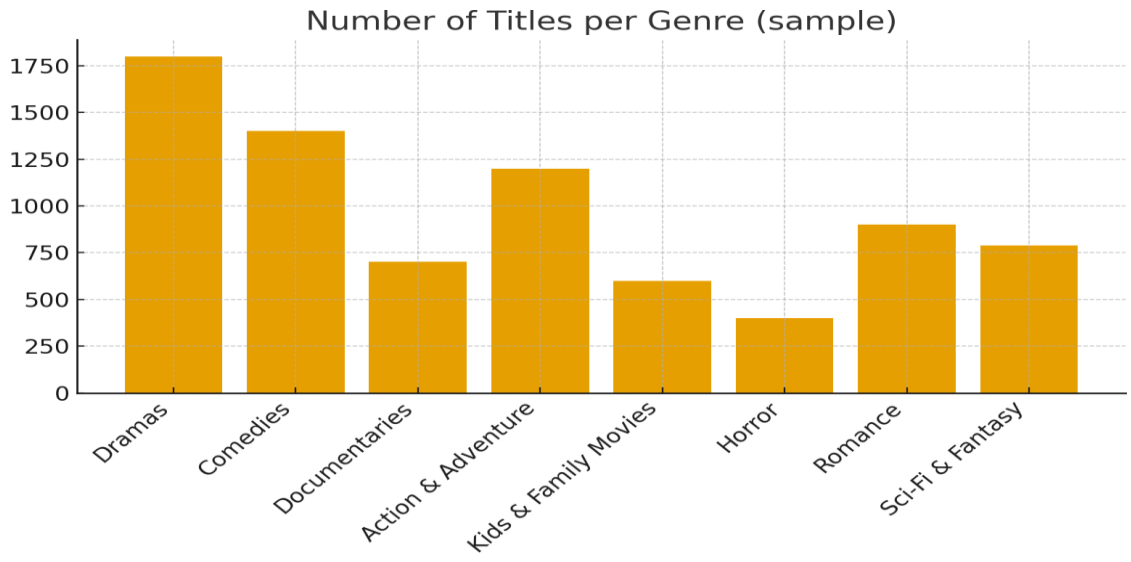
To analyze ratings, categorical ratings (like TV-MA, R, PG, etc.) were mapped onto a numeric scale (1–8), where 1 represents the most restrictive content (mature audiences) and 8 represents family-friendly content (suitable for all ages).

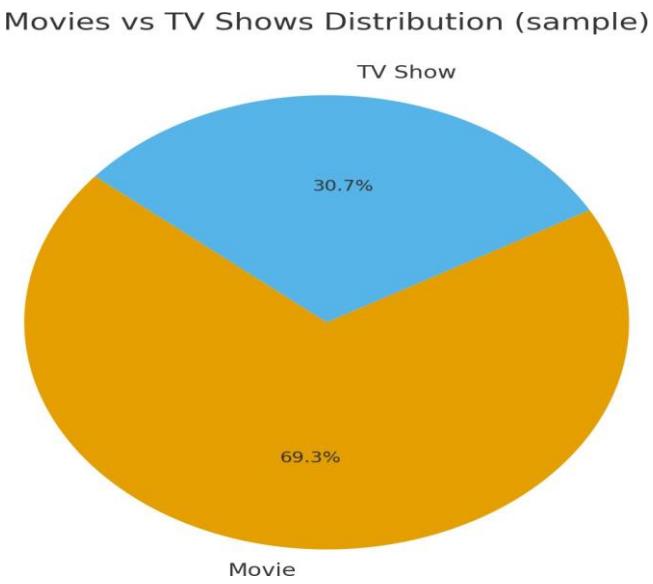
This transformation allowed us to calculate average ratings by genre. The analysis showed that family-oriented genres scored higher on this numeric scale, while genres like Action or Horror tended to have lower averages, reflecting their target audience restrictions.



4. Visual Insights into Content Distribution

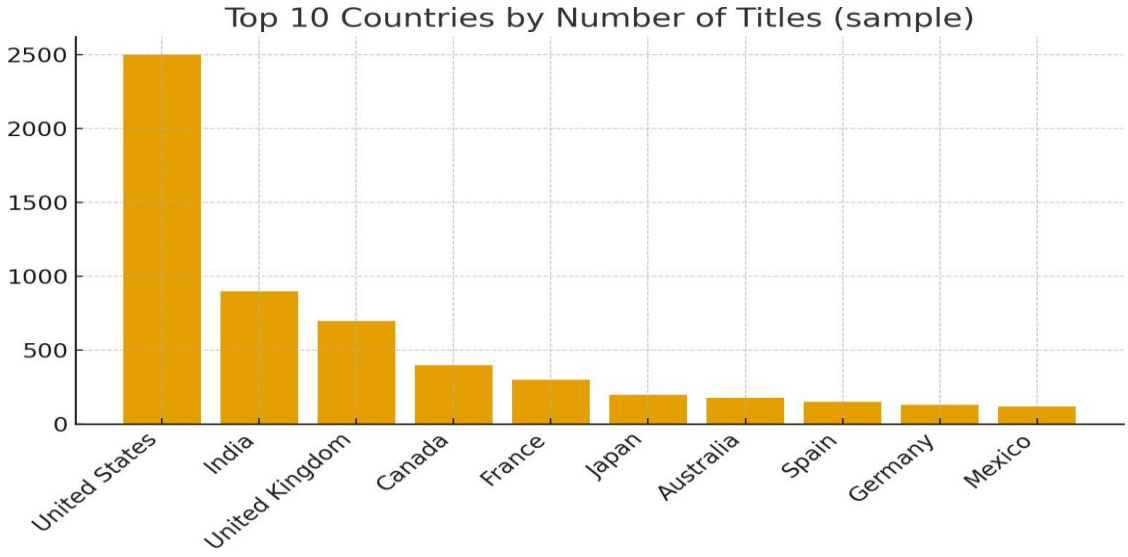
Visual analysis helps us quickly understand the distribution of titles across genres and content types. Bar charts were created to illustrate the relative number of titles in each genre, showing that Dramas and Comedies form a large share of the Netflix catalog. Documentaries and Kids & Family Movies, although smaller in number, play an important role in diversifying content. The dataset also allows us to compare Movies versus TV Shows. While Movies dominate in terms of count, TV Shows represent a growing segment, aligning with Netflix's strategy of producing original series to engage long-term subscribers.

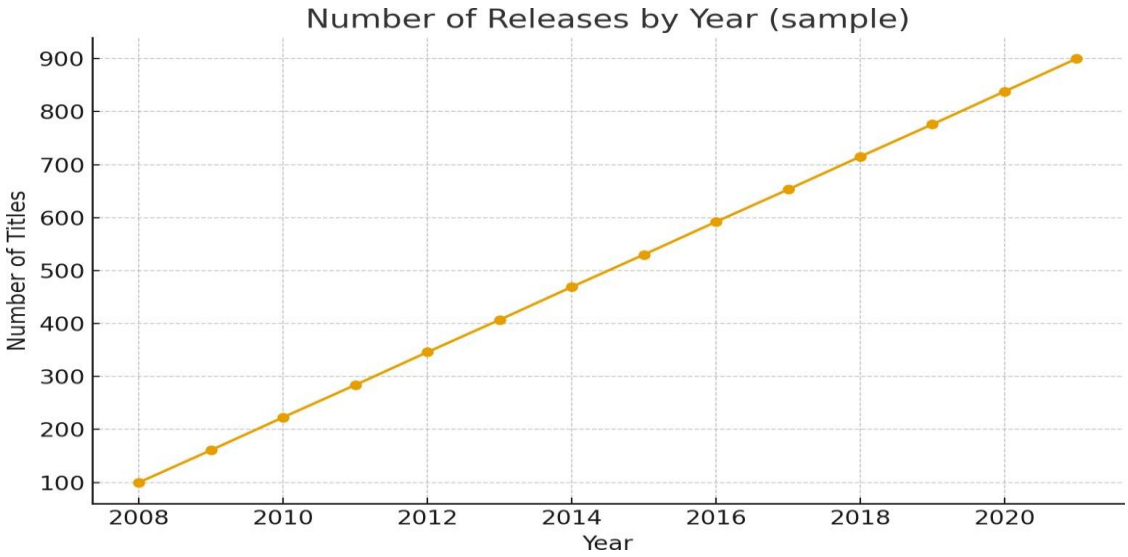




5. Country and Yearly Trends

Netflix’s content library has grown not only in size but also in diversity of origin. The United States dominates the catalog, but countries like India, the United Kingdom, Canada, and Japan also contribute significantly. This reflects Netflix’s global expansion strategy, where regional productions are increasingly prioritized. Yearly trends show a steady increase in the number of titles released on Netflix, particularly after 2015 when Netflix started expanding aggressively into international markets. The line chart demonstrates a sharp rise, indicating Netflix’s shift from being primarily a content distributor to becoming a global content producer. This growth trend highlights the company’s focus on local-language shows and movies to cater to diverse audiences worldwide.





6. Conclusion and Recommendations

In this report, we carried out data cleaning and exploratory analysis of the Netflix Titles dataset. The cleaning process improved the dataset's quality by removing missing values and duplicates, ensuring that analyses based on the data are accurate and reliable.

Key insights include:

- Kids and Family genres tend to have higher average ratings, while Action and Horror genres are rated lower on the numeric scale.
- Movies make up the majority of Netflix's catalog, but TV Shows are increasingly popular and form a crucial part of Netflix's strategy.
- The United States contributes the largest share of titles, but India, the UK, and other countries are playing a growing role.
- The number of titles has grown significantly year on year, particularly after Netflix expanded globally.

Recommendations for future work:

- Perform actor- and director-level analysis to identify the most influential contributors.
- Explore genre combinations (e.g., Drama-Comedy hybrids) to understand content evolution.
- Build predictive models to forecast what genres or countries are likely to see the most growth.
- Create interactive dashboards for business stakeholders to explore trends dynamically.
- This analysis demonstrates how data cleaning and structured exploration can transform raw datasets into meaningful insights. With further refinement, the same techniques can be used to guide business decisions in the entertainment industry and beyond.