Netflix Movies & TV Shows – Exploratory Data Analysis (EDA)

Internship: Data Science Internship

Submitted by: Syeda Faizah

Week 4 Task Report – Netflix EDA

Tools & Technologies Used:
Python, Pandas, NumPy, Matplotlib, Seaborn

1. Introduction

Netflix offers a wide variety of movies and TV shows, making it one of the largest streaming platforms in the world.
The goal of this project is to analyze Netflix's content library using **Exploratory Data Analysis (EDA)** to extract meaningful insights about content type distribution, release trends, country-wise production, movie durations, and rating patterns.

This analysis helps in understanding **content trends over time**, **geographical production patterns**, and **user-oriented content insights**, which can be valuable for business and data-driven decision-making.

2. Dataset Description

The dataset contains Netflix Movies and TV Shows and includes the following columns:

Column	Description
show_id	Unique identifier for each title
type	Content type – Movie or TV Show
title	Name of the title
director	Director of the title
cast	Main actors/actresses
country	Country of production
date_added	Date when added to Netflix
release_year	Original release year
rating	Audience rating (e.g., PG, TV-MA)
duration	Duration in minutes (for movies) or seasons (for TV Shows)
listed_in	Genre/category of the content
description	Short synopsis

3. Data Preprocessing

To ensure accurate analysis, the following steps were taken:

- **Missing Values:**
  - director**,** cast**,** country → filled with 'Unknown'
  - rating → filled with mode value
  - date_added → filled with mode after conversion
- **Whitespace Removal:** Leading/trailing spaces removed from string columns to avoid parsing issues.
- **Date Conversion:**
  - 'date_added' converted to 'datetime'

    - Extracted ‘year_added’ and ‘month_added’ for trend analysis
- **Duration Processing:**
    - Extracted numeric value from ‘duration’ column
    - Classified each entry as 'Minute' (Movie) or 'Season' (TV Show)

4. Feature Engineering

- year_added → year when the content was added to Netflix
- month_added → month of addition
- duration_num → numeric duration for movies or number of seasons for TV Shows
- duration_type → distinguishes Movies vs TV Shows

These features were essential for **time-based trend analysis**, **duration distribution**, and **content type comparisons**.

5. Visualizations and Insights

5.1 Distribution of Content on Netflix

Insight:

- Movies dominate Netflix’s library compared to TV Shows.

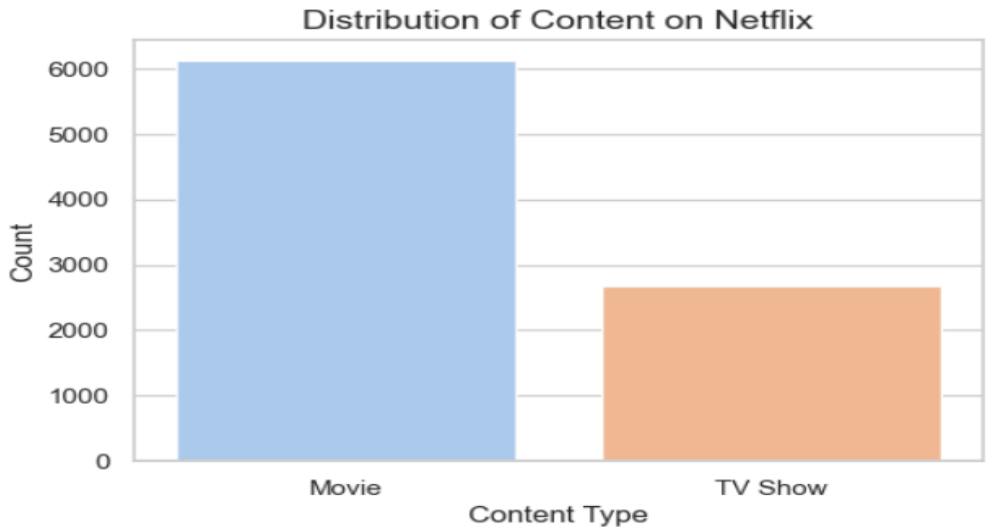

Fig 5.1: Distribution of Movies vs TV Shows

5.2 Content Added Over the Years

Insight:

- Rapid growth in Netflix content is observed after 2016, indicating platform expansion and investment in original content.

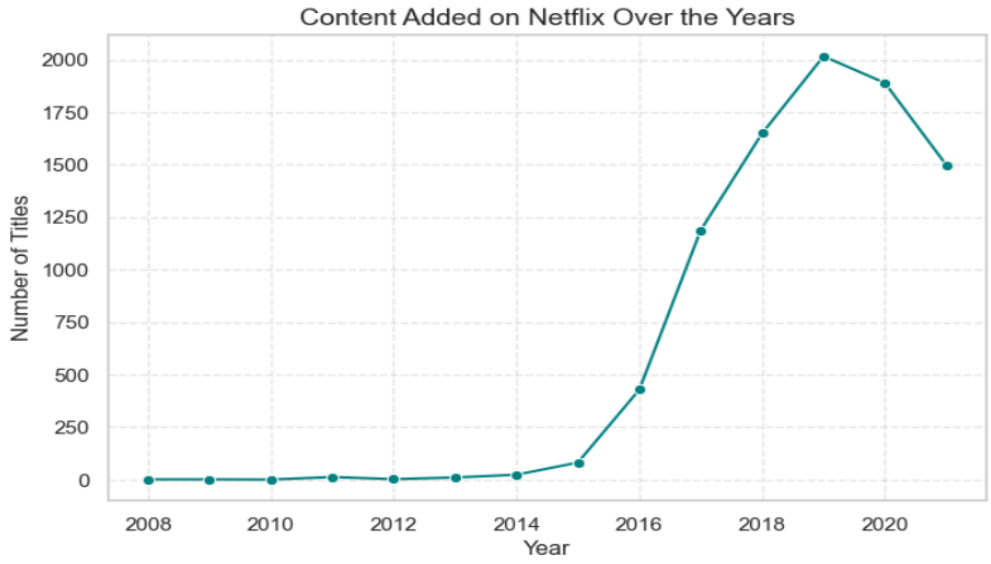

Fig 5.2: Number of titles added over the years

5.3 Top 10 Countries Producing Netflix Content

Insight:

- The USA dominates content production, followed by India and the UK.
- Suggests Netflix's focus on English-language content and expanding global library.

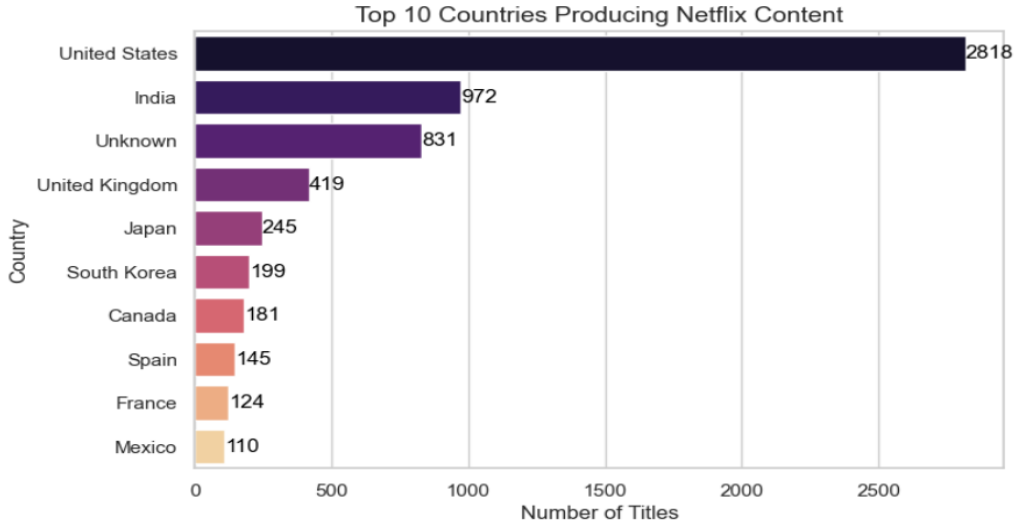

Fig 5.3: Top 10 countries producing Netflix content

5.4 Movie Duration Distribution

Insight:

- Most movies are between 80–120 minutes, reflecting standard feature-length content.
- Very few movies exceed 200 minutes, indicating selective long-format content.

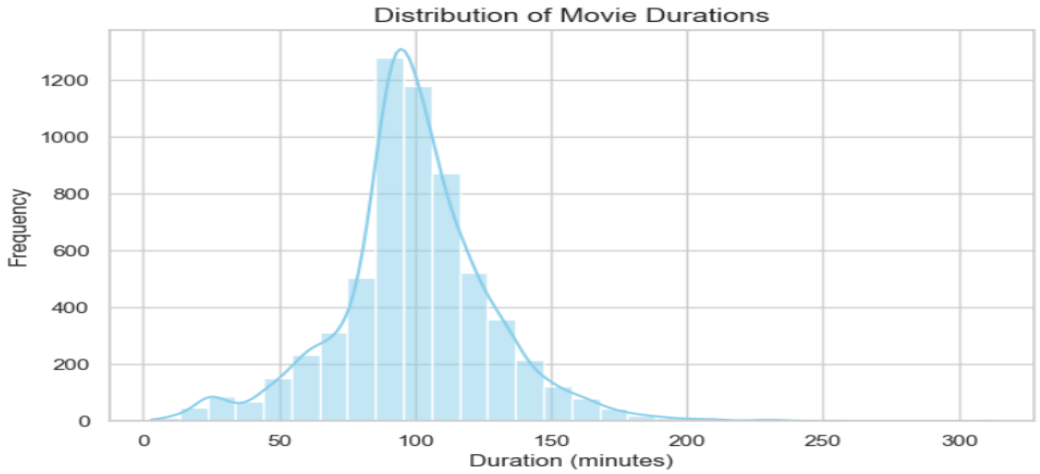

Fig 5.4: Distribution of movie durations

5.5 Ratings Distribution by Content Type

Insight:

- TV Shows dominate mature content ratings (TV-MA), while Movies have a more balanced rating distribution.
- This may guide Netflix in content rating strategies and parental controls.

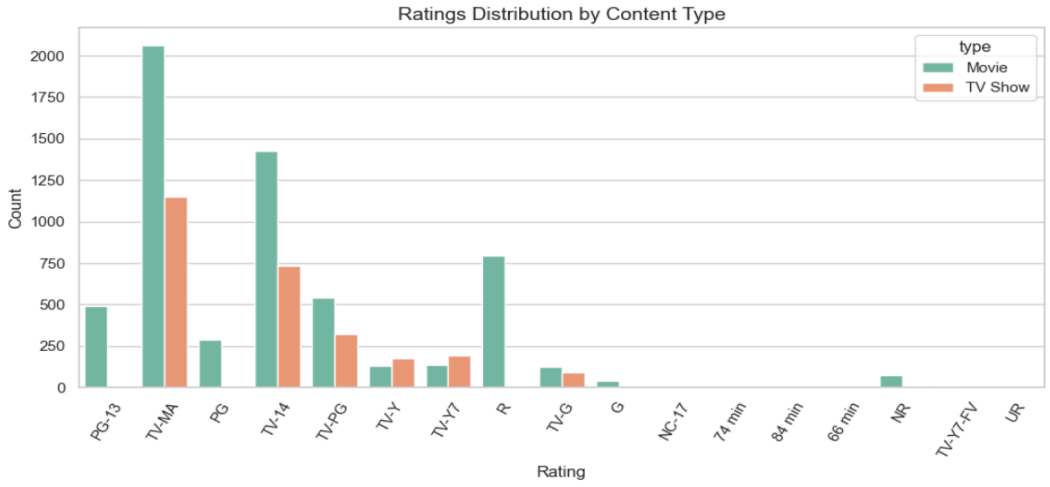

Fig 5.5: Ratings distribution by content type

5.6 Monthly Content Additions

Insight:

- Content additions peak mid-year (June–August), suggesting seasonal release patterns for new titles.
- Could help in planning promotional campaigns.

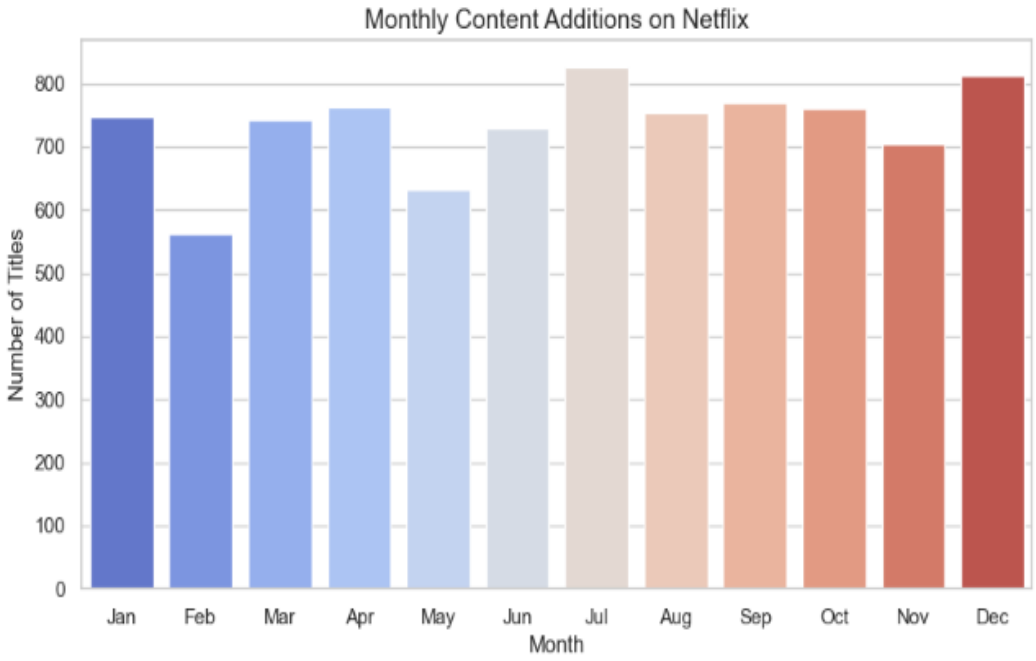

Fig 5.6: Monthly trend of content additions

6. Summary of Key Insights

- **Movies dominate** Netflix content library.
- Rapid content growth observed **after 2016**.
- **USA is the leading content producer**, followed by India and the UK.
- Most movies have **standard durations of 80–120 minutes**.
- **TV Shows dominate mature ratings**, while movies are more diverse.
- Content additions **peak mid-year**, indicating seasonal strategies.

7. Conclusion

This EDA project provides a **comprehensive overview** of Netflix's content library:

- Key patterns in **type, duration, ratings, country of production, and time trends** were analyzed.
- Insights can inform **content acquisition, production planning, and release strategies**.

- This project demonstrates practical application of **Python, Pandas, and visualization libraries** in a real-world streaming platform dataset.

8. Future Scope

- Analyze **genre-wise trends** to identify popular categories.
- Study **cast/director influence** on content popularity.
- Build **predictive models** for content success based on historical data.
- Incorporate **user ratings and reviews** for deeper insights.