21-05-2025

Project Report

Netflix TV Shows vs Movies - Exploratory Data Analysis

Submitted By:

Muhammad Areeb (21F-BSCS-20)
Muhammad Hamza (21F-BSCS-29)
Muhammad Bilawal (21F-BSCS-32)
Mohammad Ahmed Siddiqui (21F-BSCS-49)

Submitted To:
Engr. Imran Alvi

Course:
Introduction to Data Science

Dawood University of Engineering and Technology

Table of Contents

Abstract:

This project presents an Exploratory Data Analysis (EDA) of the Netflix Titles dataset to uncover patterns and trends in the platform's global content library. The analysis process included data loading, cleaning, and transformation tasks such as handling missing values, converting date fields, and creating new derived features for better interpretability. Univariate and bivariate visualizations were used to examine distributions and relationships across various features including content type, release periods, countries of origin, genres, durations, and rating classifications.

The study leverages Python libraries such as Pandas, Matplotlib, and Seaborn to conduct detailed visual and statistical analysis. A structured approach was followed to identify inconsistencies in the dataset, generate time-based insights, and segment categorical variables into meaningful components. Custom visualizations were used to evaluate content trends over time, explore genre frequencies, analyze regional production patterns, and assess audience rating categories. The project emphasizes the importance of visual storytelling in making large datasets interpretable and actionable, serving as a basis for further business intelligence or content recommendation strategies in the streaming media domain.

Project aims and objectives:

The aims and objectives of the project can be categorized into two main levels:

Overall Aim:

The overall aim of the project is to perform a comprehensive exploratory data analysis on the Netflix Titles dataset in order to identify patterns, trends, and relationships within the data that can help understand Netflix's content distribution, audience targeting strategies, and temporal production dynamics.

This aim defines the high-level goal of using data-driven insights to interpret the structure and evolution of Netflix's content offerings.

Project Objectives:

The objectives break down the aim into specific, measurable, achievable, relevant, and time-bound (SMART) goals that need to be achieved to accomplish the overall aim. Here's a breakdown of the project's objectives:

- **Understand dataset structure and quality:** This objective focuses on exploring the dataset schema, identifying missing or inconsistent values, and preparing the data for meaningful analysis through cleaning and preprocessing.

- **Perform data cleaning and feature engineering:** This objective involves dropping irrelevant columns, filling or handling missing values, converting date fields, and creating new attributes (e.g., year added, duration type) to enhance the dataset's analytical value.

- **Analyze content types and formats:** This objective examines the distribution and characteristics of Movies versus TV Shows, providing insights into content format priorities on the platform.

- **Explore geographic and production trends:** This objective investigates the countries contributing most to the content library, and how content production has varied over time.

- **Investigate genre and duration distributions:** This objective focuses on identifying dominant genres and common durations for both Movies and TV Shows to understand viewing preferences.

- **Evaluate content rating classifications:** This objective analyzes the frequency and distribution of content ratings (e.g., TV-MA, TV-14) to infer target audience demographics.

- **Visualize findings using informative plots:** This objective emphasizes the use of data visualization tools to present insights clearly, aiding interpretation and communication of patterns.

These objectives act as the specific steps toward achieving the overall aim of producing a well-structured, insight-rich exploratory analysis of the Netflix Titles dataset.

Dataset Description:

The dataset used in this project was obtained from a CSV file named netflix_titles.csv, which contains information about Movies and TV Shows available on Netflix. This dataset is publicly accessible and was originally published on the Kaggle platform under the title **"Netflix Movies and TV Shows"**.

Features/Columns Overview:

The dataset includes 12 columns, each representing different attributes of the shows listed on Netflix:

- **show_id**: A unique identifier assigned to each show.
- **type**: Indicates whether the content is a Movie or a TV Show.
- **title**: The name of the Movie or TV Show.
- **director**: Name(s) of the director(s) involved in the production.
- **cast**: List of main actors/actresses featured in the show.
- **country**: The country or countries where the show was produced.
- **date_added**: The date the show was added to Netflix.
- **release_year**: The year the show or movie was originally released.
- **rating**: The maturity rating (e.g., TV-MA, PG, R) assigned to the content.
- **duration**: Duration of the Movie in minutes or number of seasons for TV Shows.
- **listed_in**: Categories or genres associated with the show (e.g., Dramas, Comedies).
- **description**: A brief synopsis or description of the content.

Initial Shape and Structure of the Dataset:

When first loaded, the dataset contains 8,807 rows and 12 columns. The data includes both textual and date/time fields, with categorical variables like type, rating, and country. Upon inspection, several columns such as director, cast, and country contain missing values, which were handled during the data cleaning process. The date_added column also required conversion from string to datetime format for effective time-based analysis.

Libraries Implemented:

1. **Pandas** – For data loading, cleaning, manipulation, and analysis.
2. **NumPy** – For numerical operations and handling missing data.
3. **Matplotlib** – For creating basic visualizations such as bar and line plots.
4. **Seaborn** – For generating advanced statistical and aesthetically pleasing plots.
5. **Collections (Counter)** – For counting and analyzing occurrences of categorical data.

Data Preprocessing:

- Loaded the dataset using Pandas from the CSV file netflix_titles.csv.
- Handled missing values in columns like director, cast, country, and date_added by filling them with placeholders or leaving them as-is.
- Removed duplicate records to ensure data uniqueness.
- Converted date_added column to datetime format and extracted year_added and month_added.
- Split the duration column into two new columns: duration_int (numeric part) and duration_type (minutes or seasons).
- Standardized text fields like title, cast, and listed_in by converting them to lowercase.

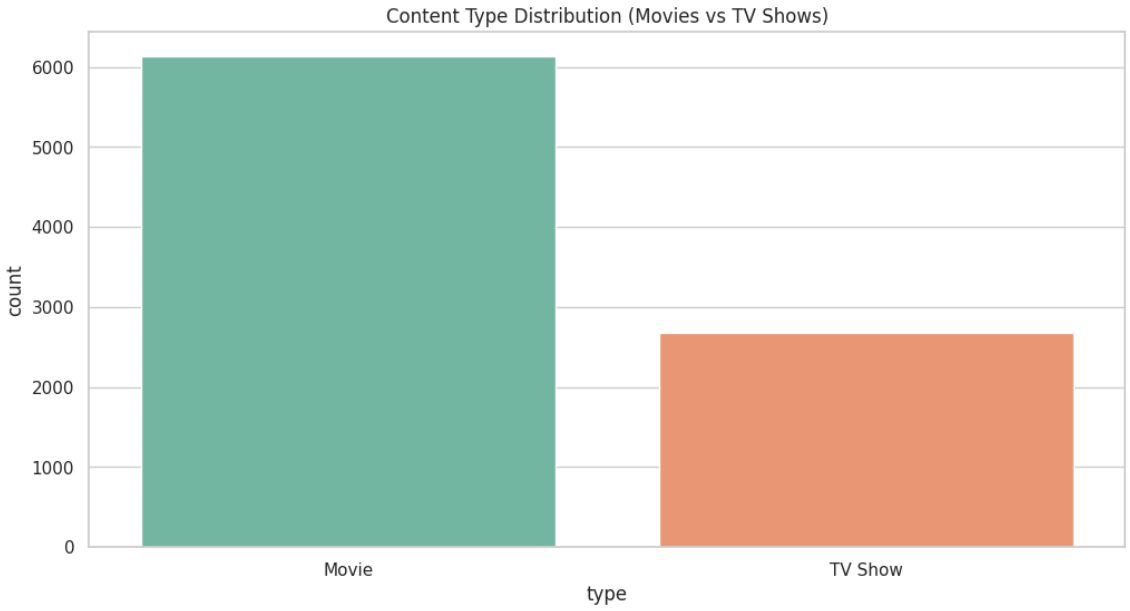
Exploratory Data Analysis (EDA):

EDA involves visually and statistically examining the dataset to uncover patterns, trends, and insights. It helps in understanding the structure of the data and supports better decision-making.

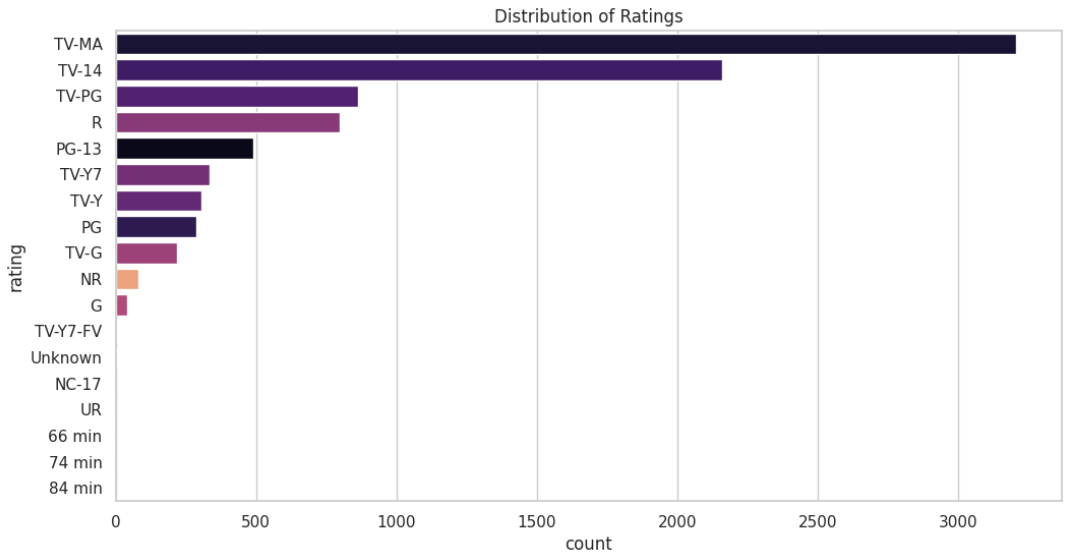
Univariate Analysis:

Focuses on analyzing individual columns to understand their distributions and dominant values.

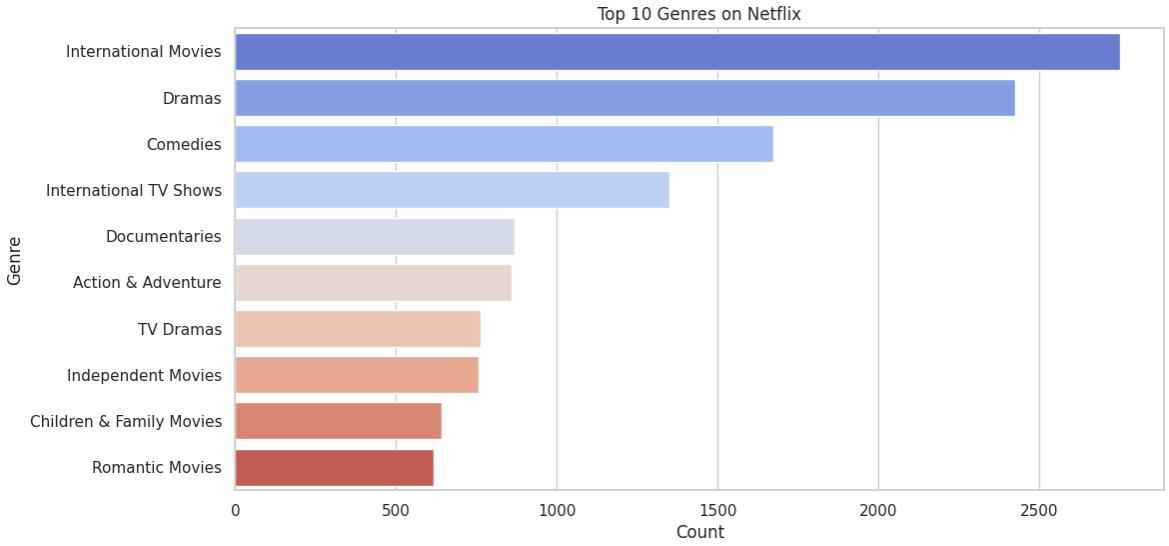
1. **Content type distribution**



2. **Rating distribution**



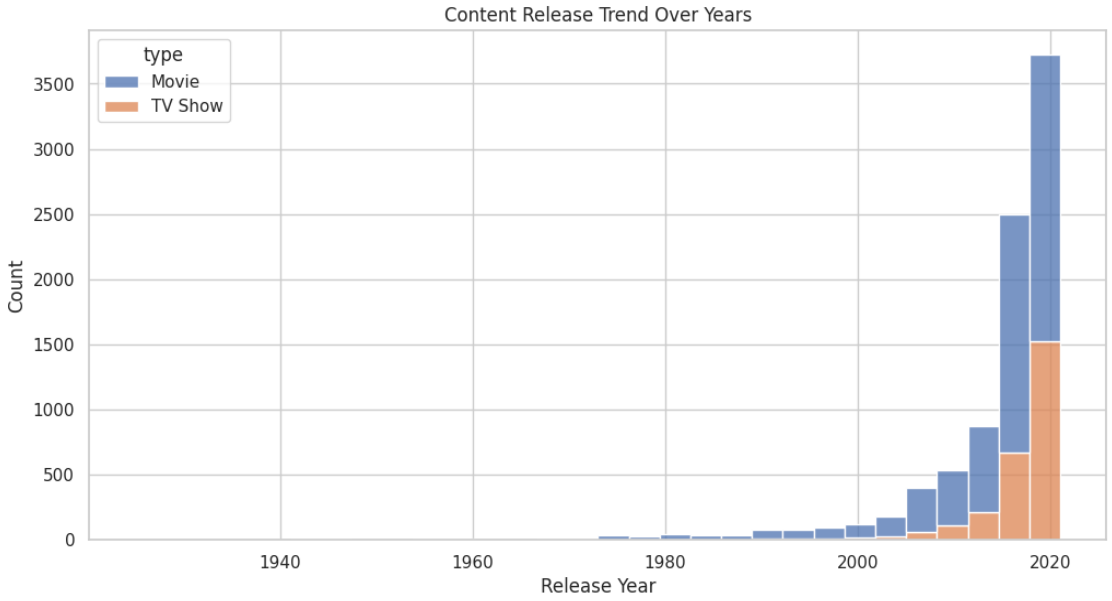
3. Genre Counts



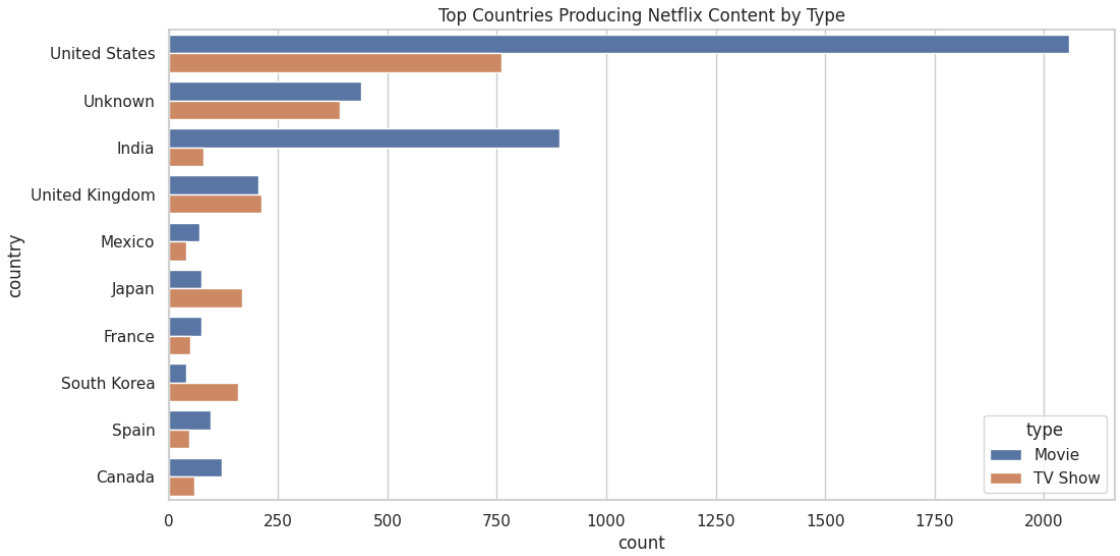
Bivariate Analysis:

Examines the relationships between two or more variables to identify correlations and trends.

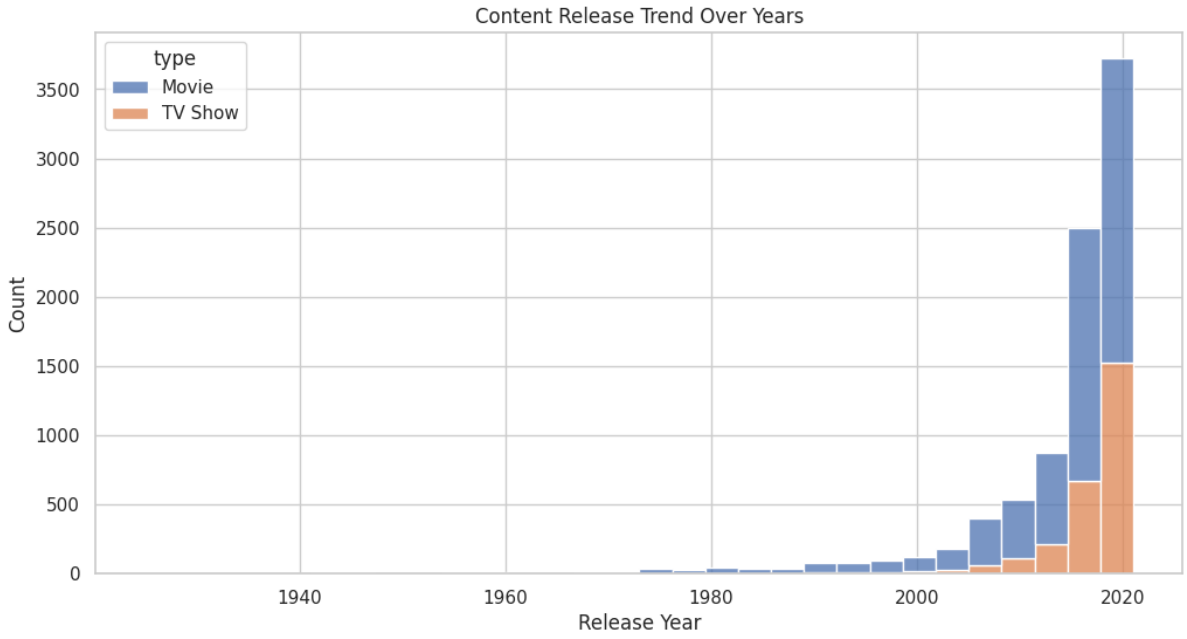
1. Year vs number of releases



2. Country vs content type

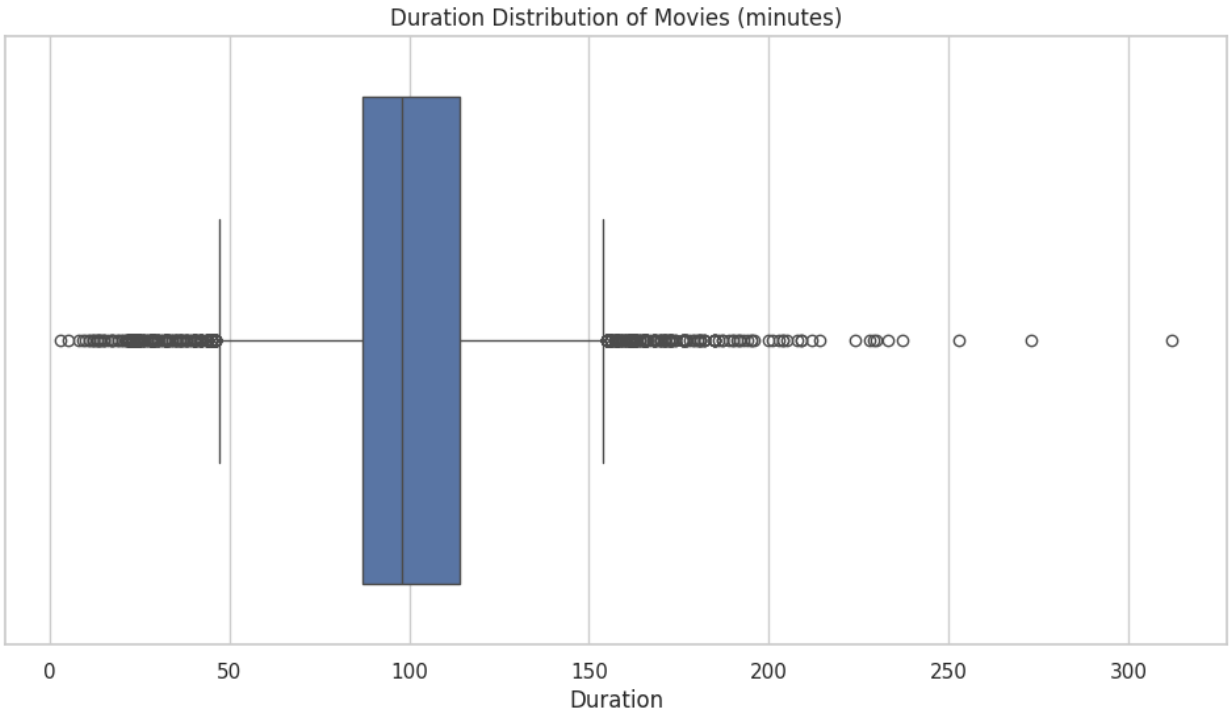


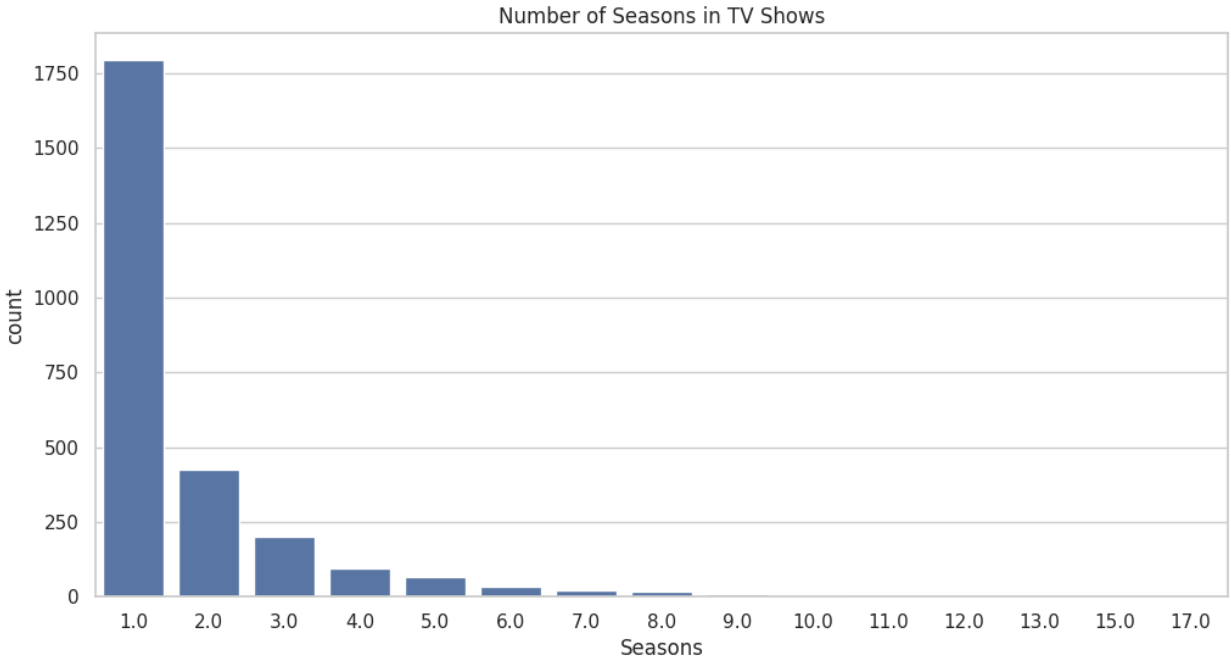
3. Year vs Number of Releases



Duration Analysis:

Explores the typical lengths of Movies and the number of seasons in TV Shows.





Summary of Insights

Based on the exploratory data analysis, the following key insights were derived:

1. Netflix has more Movies than TV Shows.
2. The US, India, and the UK contribute the most content.
3. Dramas, International Shows, and Comedies are dominant genres.
4. Most Movies are 90-120 minutes. TV Shows have 1-2 seasons.
5. Peak content production occurred between 2018-2020.
6. Ratings like TV-MA and TV-14 are most common, suggesting a large young adult audience.
7. Content additions to Netflix peaked in 2019.

Conclusion:

This exploratory data analysis project provided valuable insights into the structure and trends of Netflix's content library. By examining the dataset, we discovered that Movies dominate over TV Shows, with the majority of content originating from the United States, India, and the United Kingdom. Genres like Drama, Comedy, and International Shows were found to be the most prevalent, catering largely to a young adult audience, as suggested by the common TV-MA and TV-14 ratings.

The analysis also highlighted that Netflix experienced its highest surge in content additions between 2018 and 2020. Most movies typically have a runtime of 90–120 minutes, while TV Shows usually span one to two seasons. These findings not only help understand content distribution on Netflix but also reflect broader trends in global entertainment consumption.

Overall, the project demonstrated the power of EDA in extracting actionable insights from raw data and sets a strong foundation for further studies, such as user behavior analysis or content recommendation systems.