Project by: *Aarav Shroff*

Dataset: Netflix Movies and TV Shows
Author: Shivam Bansal
Dataset link: https://www.kaggle.com/datasets/shivamb/netflix-shows

Research Question: How has Netflix's content library evolved over time in terms of content type, genres, and geographic distribution?

Why this question: Not predictive , Answerable with visuals , Broad enough for storytelling, Specific enough to stay focused

Problem Definition: With the rise of streaming services, the way people consume movies and television shows has changed dramatically, and Netflix has become one of the largest media distributors in the world. As the organization has grown and expanded its reach to include more original productions, the nature of its catalog has likely changed in important ways. Understanding these changes can provide valuable insight into the larger trends in global media production and distribution.

The goal of this project is to explore the evolution of the Netflix catalog in terms of the types of media that the organization offers, the genres of media that the organization offers, and the geographic distribution of media production. By understanding the trends and patterns in these areas, it is possible to gain a deeper understanding of the evolution of the Netflix catalog and the way that the organization has changed over time.

This topic is important to media analysts, professionals, and anyone looking to understand the way that streaming services impact global media trends. Through the process of exploratory data analysis, it is possible to understand the trends and patterns that have emerged in the Netflix catalog and gain valuable insight into the way that the organization has evolved over time.

Data Description: The dataset used for this project is the Netflix Movies and TV Shows dataset found on Kaggle, which was compiled by Shivam Bansal. Each row in this dataset corresponds to a single title that is available for streaming on Netflix, including both movies and television shows.

The key attributes found in this dataset include the title name, the type of content (movie or TV show), the country where the content was produced, the date it was added to Netflix, the release year for the title, the rating for the title, the length of the title, and the genre classification for the title. These attributes allow for analysis of the types of changes that have occurred to the platform as a whole.

There are thousands of rows in this dataset, making it a broad overview of what types of content are available to users on Netflix. However, there are instances where there are missing values

for the data in this dataset, including the country and cast columns. This creates a degree of uncertainty when analyzing the trends found in this dataset. Furthermore, the genre classifications are based upon what the platform defines for each title and may not accurately reflect the overall content found within each title.

Cleaning & Preparation: Before performing analysis on the data, a number of preprocessing steps had to be taken to ensure that the data was in a suitable form to be analyzed. First off, the date_added column had to be changed to a datetime format to allow for analysis based on this feature. This led to a new feature being created, showing the year a title was added to Netflix.

The issue of missing values in the data had to be handled carefully rather than simply ignoring it. When dealing with categorical data such as country and genre, it was assumed that a missing value could be classified as “Unknown.”

Further, there were a number of titles in the data that had been produced in multiple countries. To simplify this feature, a single country from a list of countries provided in a title’s production details was taken as the main country. This oversimplification might have been a drawback, but it was a necessary evil to ensure easy visualization.

Data Understanding & Visualization: To answer the research question, I created visualizations focused on time trends, content type differences, geographic distribution, and genre composition:

- Titles added over time (line chart) to examine overall growth and changes in catalog expansion.
- Movies vs TV shows over time (multi-line chart) to compare how content type changed as Netflix grew.
- Top countries by number of titles (bar chart) to understand geographic concentration.
- Top country trends over time (multi-line chart for top 5 countries) to see whether geographic contributions changed.
- Top genres overall (bar chart using exploded genres) to identify dominant content categories.
- Top genres by content type (grouped bar chart) to compare genre patterns between movies and TV shows.

Each visualization was chosen to directly connect to part of the research question rather than being decorative.

Storytelling & Interpretation: The time-based charts track the growth of the library over the years, including the periods of rapid growth. We can also analyze the growth of Movies compared to TV Shows over the years. This will give an idea of whether the growth is more towards the addition of Movies or TV Shows. An increase in the growth of TV Shows over the years might indicate a focus on TV shows to retain customers or to create more original TV content.

The geographic distribution charts track whether the growth is concentrated in a few countries or whether it is becoming more diverse over the years. A high concentration of growth in a few countries might indicate the growth is more towards leveraging licensing opportunities and production capabilities in those countries. An increase in the representation of more countries over the years might indicate a broader reach of the platform.

The genre charts track the growth of the library across various genres. This will give an idea of the genres dominating the library. It will also give an idea of whether there is a significant difference between the genres of Movies and TV Shows. For example, a strong presence of “International Movies” or “TV Dramas” might indicate a focus on those genres. Note that these genres are based on the platform's definition and might not reflect an objective definition of the genre.

Limitations,Ethics & Reflection:
Limitations and assumptions

- This dataset reflects Netflix’s catalog, not the entire film/TV industry, so conclusions only apply to what appears on Netflix.

- Missing values (especially country and sometimes date_added) reduce confidence in geographic and time-based conclusions.

- Using primary_country simplifies multi-country titles and may underrepresent co-productions.

- Genre labels are not standardized and may change based on how Netflix categorizes titles.

Bias and ethical considerations

- Netflix’s catalog is shaped by licensing, business strategy, and market power, so overrepresentation of certain countries may reflect systemic inequalities in global media distribution.
- Missing metadata may not be random; smaller studios or non-mainstream titles may be less documented, which can bias apparent patterns.

Reflection / future work

One surprise to explore further is how quickly the catalog appears to grow during certain years and whether that correlates with particular types of content or with certain regions. If I had more time, I would analyze differences between release_year and year_added to see how long it takes older titles to appear on Netflix, and I would bring in external context to better explain observed shifts.

References & AI Use Transparency:

Dataset - Shivam Bansal. Netflix Movies and TV Shows. Kaggle dataset: https://www.kaggle.com/datasets/shivamb/netflix-shows

Other resources - pandas documentation (datetime conversion, string split/explode), matplotlib documentation (basic plotting)

AI transparency - I used ChatGPT (GPT-5.2) to help plan the analysis structure, draft portfolio wording, and sanity-check cleaning/visualization steps. All code was run and verified by me, and interpretations were written based on the outputs from my notebook.