Project Report

Netflix Content Strategy Analysis

REG # FA23-BCS-126

NAME: MANAHILL KHITAB

SECTION 5C

1. Project Overview & The Dataset

The goal of this project was to analyze the Netflix library to understand how the company has changed its content strategy over time. The dataset contains roughly **8,800 rows**, representing movies and TV shows available on the platform as of 2021. It includes details like the title, director, cast, country of origin, release year, and the date it was added to Netflix.

2. Data Problems & Challenges

When I first loaded the data, it was not ready for analysis. There were several issues:

- **Missing Values:** Many rows were missing the director, cast, and country.
- **Formatting Issues:** The date_added column was a string (text) instead of a date format, making it impossible to track growth over time.
- **Multi-value Columns:** The country and listed_in (genres) columns often contained multiple items (e.g., "United States, United Kingdom"). Counting these as they were would have led to inaccurate results.

3. Data Cleaning Process

To fix these problems, I performed the following steps:

- **Imputation:** I filled in the missing director, cast, and country values with labels like "Unknown" or "No Director" so the rows wouldn't be deleted.
- **Standardization:** I converted the date_added column into a proper Python **Datetime** format and extracted the **Year Added**.
- **Exploding Data:** I split the multi-country and multi-genre strings into individual items. This ensured that if a movie was made by three countries, each country got credit in my analysis.

4. Features & Variables

I worked with a total of **12 original columns** and created **2 new features** to deepen the analysis:

1. **Content Age:** Calculated by subtracting the release_year from the year_added.
2. **Is Mature:** A True/False column to identify adult-rated content (TV-MA, R, NC-17).

5. Analysis & Visualization Methodology

I used the **Pandas** library for data manipulation and **Seaborn/Matplotlib** for creating visuals. The analysis was divided into three main parts:

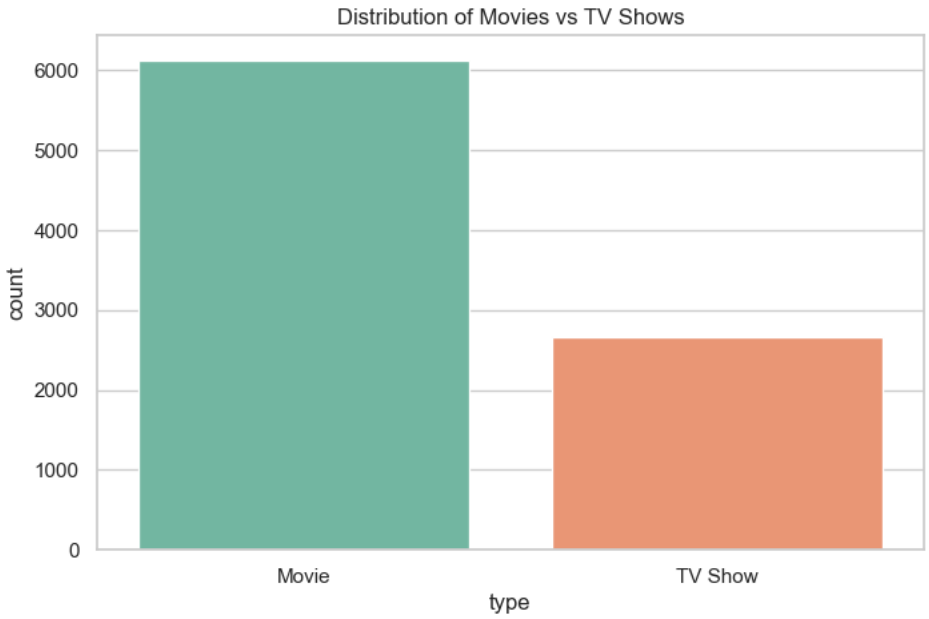
- **Composition:** Looking at the split between Movies and TV Shows.
- **Geography:** Mapping which countries produce the most content.

- **Trends:** Tracking growth, genre diversity, and maturity shifts over time.

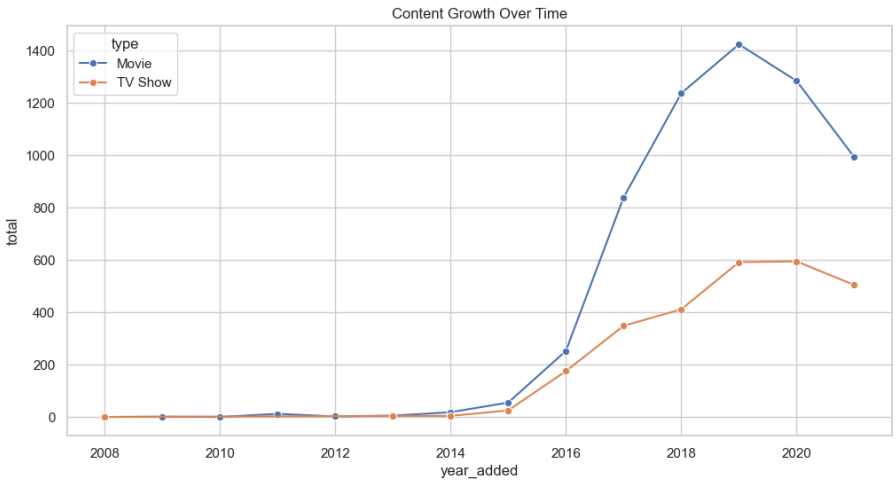
6. Key Charts & Results

I created several plots to visualize my findings:

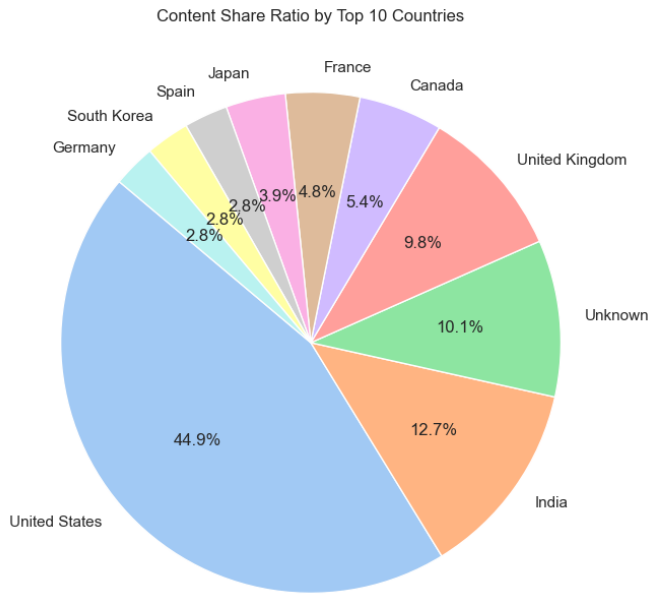
- **Bar Charts:** Showed that Movies dominate the volume, but the United States and India are the clear production leaders.



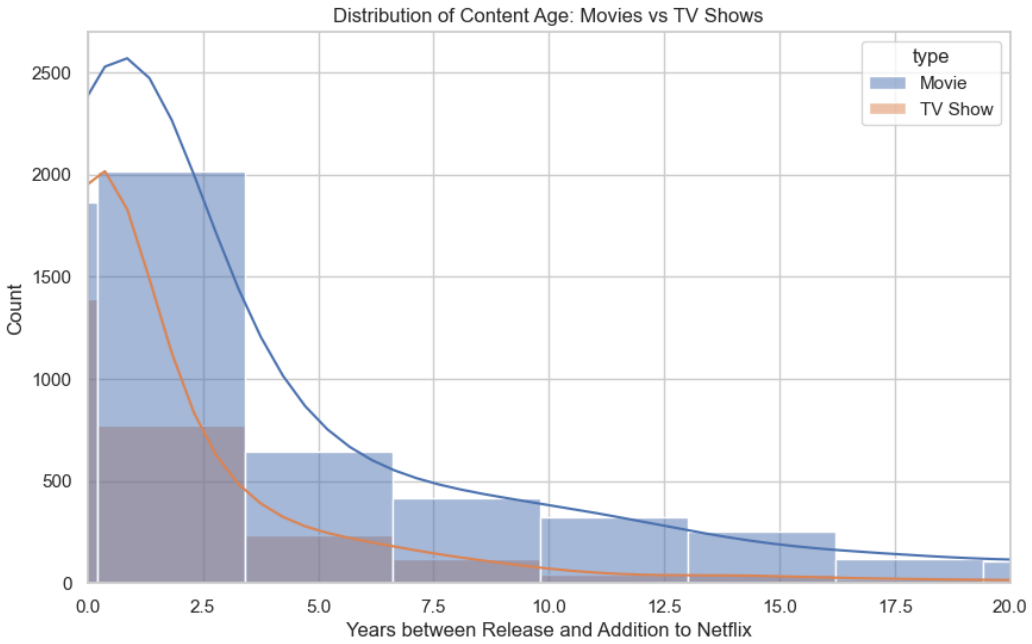
- **Line Graphs:** Revealed a massive spike in content additions between 2016 and 2019, followed by a pandemic-related dip.



- **Pie Charts:** Showed the "Share of Content," highlighting that Netflix still relies heavily on the US for nearly 40% of its library.



- **Histograms:** Used for **Content Age**, proving that Netflix prefers adding "fresh" content (less than 2 years old) to keep the library feeling current.



7. Final Strategic Conclusion

The story of this data shows a platform in transition. Netflix has moved from being a repository for old licensed movies to a global production powerhouse.

Main Conclusion: Netflix is successfully pivoting toward **Mature, International TV Series**. This strategy helps them retain adult subscribers and lowers their risk if US production slows down. However, their heavy reliance on US content remains a strategic bottleneck that they must continue to address by investing in international markets like South Korea and India.