
UNIFIED MENTOR
NETFLIX
Netflix Data Analysis
Made by: Yuvraj Singh Bhadauria
UNID: UMID19052537412

ABSTRACT

In the era of digital streaming, platforms like Netflix have revolutionized how audiences' access and consume content. With thousands of titles spanning various genres, regions, and audience demographics, understanding the structure and evolution of this vast content library is essential for content strategists, data scientists, and business decision-makers. This project presents a comprehensive analysis of a cleaned dataset of Netflix titles, encompassing data from 1925 to 2021. The primary objective is to explore patterns and trends across multiple dimensions such as content type, genre distribution, regional contributions, ratings, and duration — all of which influence how Netflix curates and promotes its catalog to global users.

The project begins with a thorough data cleaning phase, addressing missing values, standardizing date formats, and handling ambiguous entries such as "Not Given" in director and country fields. This process ensures the integrity and usability of the dataset for meaningful analysis. The cleaned dataset, consisting of 8,503 entries and 10 key attributes, is then subjected to exploratory data analysis (EDA) using Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn. Each visualization is tailored to uncover a specific aspect of the dataset — for instance, identifying the predominance of movies over TV shows, the dominance of the U.S. in content production, and the prevalence of adult-targeted content via ratings like TV-MA and TV-14.

Genres are explored through tokenization of the listed_in column, revealing that Drama, Comedy, and International Movies are among the most frequent categories. Temporal analysis of the date_added field shows a sharp rise in content acquisition during the late 2010s, particularly between 2016 and 2020 — highlighting Netflix's aggressive global expansion strategy. In addition, analysis of movie durations and top directors offers a glimpse into Netflix's structural approach to runtime and recurring collaborators, respectively.

Beyond descriptive insights, the project also sets the stage for future data science applications. The structured and cleaned dataset can now be used to develop machine learning models such as content recommenders, genre classifiers, or user segmentation tools based on consumption patterns. These applications can assist streaming platforms in personalizing user experiences, optimizing content placement, and making data-driven content investment decisions.

In summary, this project demonstrates how structured exploratory analysis can convert raw metadata into actionable insights. By examining the Netflix content catalog through a data science lens, the project not only highlights historical trends but also provides a scalable foundation for advanced predictive analytics and strategic decision-making in the digital content domain.

1. Problem Statement

Netflix hosts a vast catalog of entertainment. Understanding its structure can inform content strategies. This project focuses on cleaning and analyzing Netflix data to extract meaningful patterns in content distribution, viewer targeting, and platform growth.

2. Dataset Overview

The dataset used in this project is a publicly available compilation of Netflix titles, originally sourced from Kaggle. It includes metadata for movies and TV shows available on the platform between 1925 and 2021. The dataset comprises 8,503 rows and 10 columns, with each row representing a unique piece of content on Netflix. Key attributes include title, type (Movie or TV Show), director, cast, country, date_added (to Netflix), release_year, rating (e.g., TV-MA, PG), duration, and listed_in (genres). These features provide a rich foundation for analyzing content trends, audience targeting, and geographic patterns of production and distribution.

Some fields, such as director and country, contain missing or ambiguous values (e.g., "Not Given"), which were addressed during the data cleaning phase. The listed_in column often includes multiple genres per title, making it ideal for genre frequency and diversity analysis. The date_added field was particularly useful in time-series analysis after being converted to a proper datetime format. Overall, the dataset is structured enough to allow in-depth exploration of Netflix's evolving content strategy and user engagement trends.

Furthermore, the dataset enables a detailed examination of how Netflix has expanded its global footprint. By analyzing the country attribute, it becomes possible to observe which regions contribute the most to Netflix's content library. Titles from the United States dominate the dataset, followed by countries like India, the United Kingdom, and Canada. This information provides insight into market priorities and regional production partnerships. Similarly, the rating column allows us to assess the platform's targeting of specific age groups, which in turn reflects broader content curation policies.

The time-based attributes — release_year and date_added — support both longitudinal and comparative analysis. By plotting trends over time, we can trace how Netflix's acquisition and original content strategies have evolved. For instance, a notable surge in content additions around 2016–2020 aligns with Netflix's aggressive global expansion and investment in original programming. When paired with genre and duration analysis, we also begin to understand how audience preferences may influence the volume and type of content being released.

From a modeling perspective, the dataset's structured nature and breadth of attributes make it a strong candidate for building classification models, recommendation systems, and user preference predictors.

3. Data Cleaning

Effective data analysis starts with clean, reliable data. The raw Netflix dataset, although rich in features, required several preprocessing steps to make it suitable for exploratory analysis and visualization. Below are the key data cleaning tasks carried out in this project:

3.1 Handling Missing Values

The dataset contained missing or ambiguous values in several important columns, particularly director and country.

- **Director**: About **29.44%** of the entries had missing or "Not Given" values in the director column. Instead of dropping these rows — which would lead to significant data loss — these values were replaced with the string "Unknown" to retain the record while still distinguishing it for analysis.
- **Country**: Roughly **3.27%** of the entries in the country column were missing. Given the relatively small proportion and the geographic importance of this field for content distribution analysis, these rows were removed from the dataset.

3.2 Removing Duplicates

A check for duplicate rows was conducted using the duplicated() method. Fortunately, no duplicate entries were found in the dataset, which is a positive indicator of data integrity.

3.3 Converting date_added to Datetime Format

The date_added column, which originally contained string-type date values, was converted into a proper datetime format. This transformation was essential for enabling time-series analysis, such as determining trends in how content was added to Netflix over the years.

4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a fundamental step in any data-driven project. It involves summarizing the main characteristics of a dataset, often through visualizations and descriptive statistics, to gain a better understanding of the data before applying any predictive modelling or hypothesis testing. EDA helps uncover underlying patterns, spot anomalies, test assumptions, and check for correlations between variables. It serves as the lens through which data scientists begin to understand the story behind the numbers.

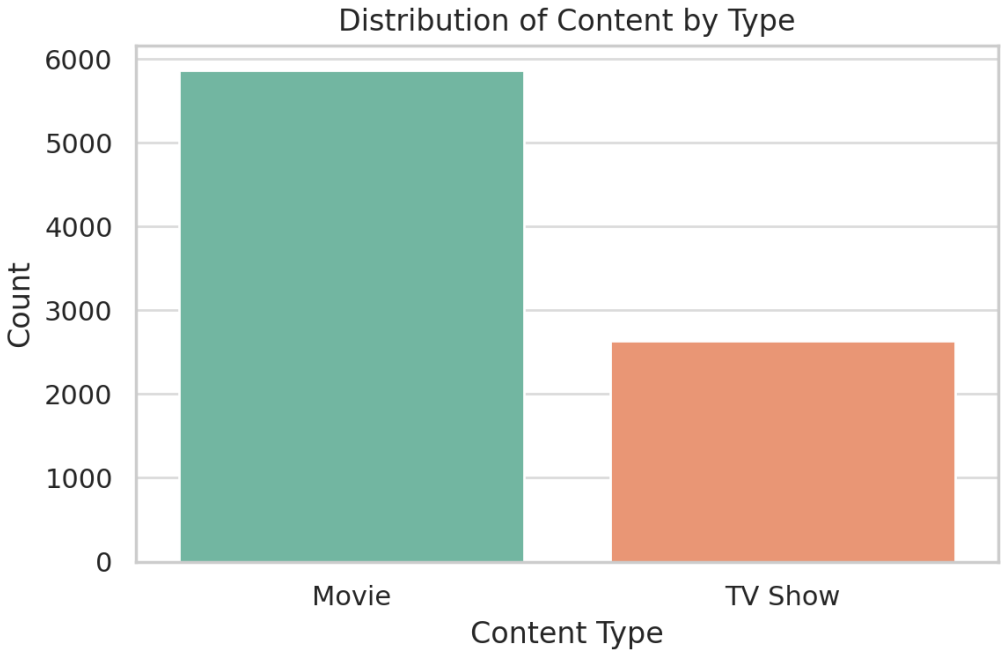
In this project, EDA was essential for analyzing the Netflix content catalog from multiple perspectives including content type, genre distribution, regional contributions, content addition timelines, viewer ratings, and movie durations. By using tools like Pandas for aggregation and Matplotlib/Seaborn for visualization, we were able to identify key trends

such as the dominance of movies over TV shows, the rise of certain genres, and the significant growth in Netflix's content additions in recent years.

Performing EDA first is critical because it guides the next steps in the analysis pipeline. It informs data preprocessing decisions, highlights what features might be important for modelling, and can even reveal errors or biases in the dataset. For instance, through EDA, we noticed the concentration of content in certain countries and genres, and a skew toward adult-rated shows insights that would otherwise remain hidden in raw data. EDA thus transforms an unstructured dataset into a meaningful foundation for storytelling, insight generation, and machine learning applications.

4.1 Content Type Distribution

Movies significantly outnumber TV Shows, indicating a preference for short-form content.

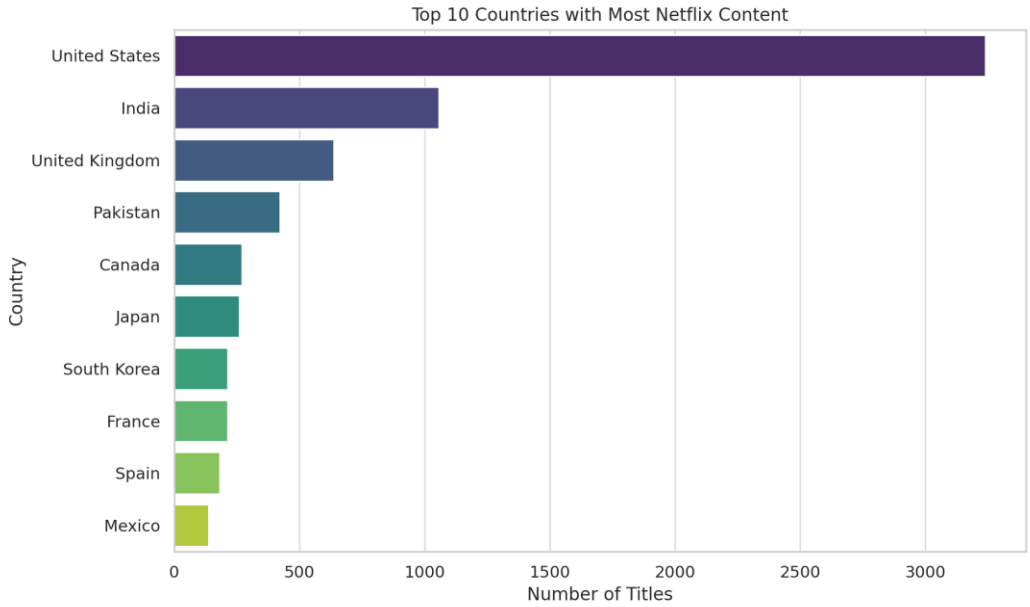


This analysis presents a comparison between the number of movies and TV shows available on Netflix. As seen in the bar chart, **movies significantly outnumber TV shows** on the platform. This clear dominance suggests that Netflix places a strategic emphasis on short-form content, possibly due to its faster production cycles, broader appeal, and suitability for casual viewing.

From a business perspective, this trend might reflect consumer behavior. Shorter content formats like movies are often easier to promote globally, require less long-term viewer commitment compared to TV shows, and can cater to a wide variety of genres and themes within a limited runtime.

4.2 Top Countries by Content Volume

The United States leads, followed by India and the UK. These regions dominate content production on Netflix.



This visualization highlights the top 10 countries contributing the most content to Netflix's global catalog. As observed, the **United States dominates** the platform, contributing far more titles than any other country over 3,000 titles. It is followed by **India**, **the United Kingdom**, and a mix of countries from Asia, North America, and Europe including Pakistan, Canada, Japan, South Korea, France, Spain, and Mexico.

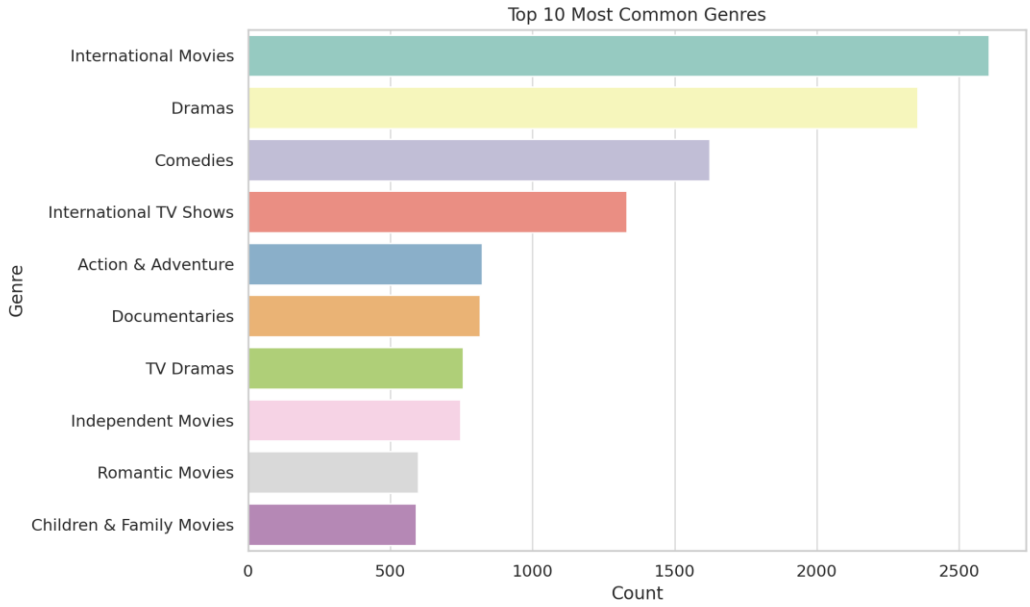
This distribution emphasizes Netflix's strong reliance on U.S.-based content, which is likely influenced by its origins as a U.S.-based company, its initial licensing agreements, and its primary audience during its early growth years. However, the presence of **India, Japan, South Korea, and Pakistan** among the top contributors shows Netflix's intentional push into diverse international markets and its investment in **regional storytelling** to appeal to local audiences.

4.3 Top Genres

Genres such as Drama, Comedy, and International Movies are most common, often with multiple tags per title.

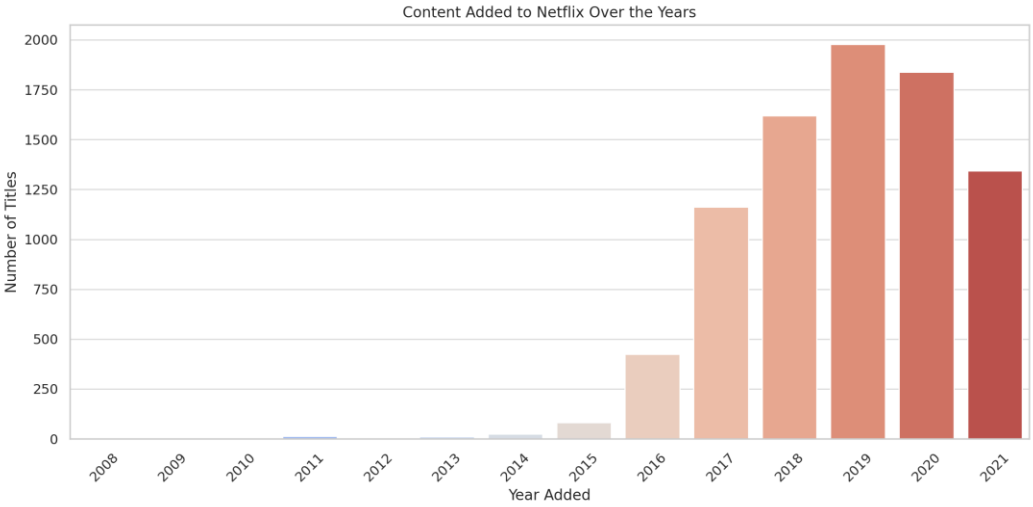
This analysis showcases the **top 10 most common genres** on Netflix by analyzing the **listed_in** column, which contains genre tags for each title. It's important to note that most titles on Netflix are tagged with **multiple genres** for example, a film can be both a "Comedy" and a "Romantic Movie." To accurately analyze this data, the genre tags were **split and exploded**, allowing for a genre-level frequency count across all titles.

The bar chart reveals that **International Movies** top the list, followed by **Dramas** and **Comedies**. This indicates that Netflix prioritizes **globally diverse content** and **emotionally engaging storytelling**, both of which appeal to a wide range of audiences. The high volume of international content also reflects Netflix's strategy to penetrate and retain users in non-English speaking markets by offering culturally relevant entertainment.



4.4 Content Addition Over the Years

Content additions peaked during 2019–2020, reflecting Netflix's aggressive global expansion.



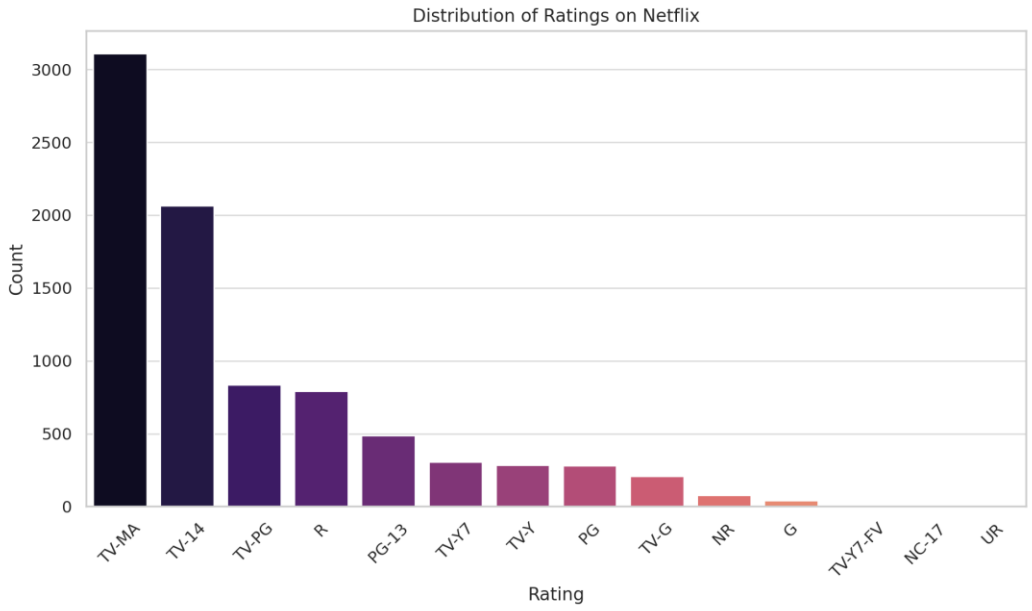
Between 2015 and 2020, Netflix transformed from a primarily U.S.-based streaming service into a dominant global entertainment platform. The sharp spike in content additions during these years aligns with the company's expansion into more than 190 countries and the

launch of Netflix Originals like *Stranger Things*, *Narcos*, and *Sacred Games*. These investments weren't just about volume they were strategic moves to offer more localized and diverse storytelling to international audiences.

This analysis highlights **Netflix's growth trajectory** and its efforts to remain competitive by frequently refreshing its catalog. The slight drop in content additions after 2020 could be attributed to **pandemic-related production delays**, shifts in content strategy, or a greater emphasis on quality over quantity.

4.5 Ratings Distribution

Most content is aimed at mature audiences with TV-MA and TV-14 ratings.



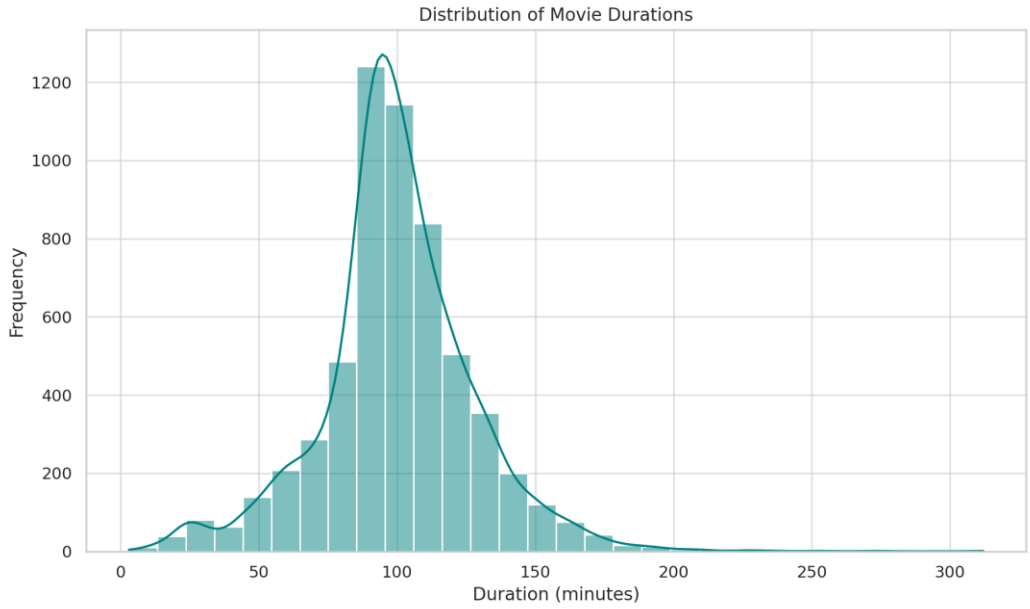
This visualization explores the distribution of content ratings on Netflix, revealing how titles are categorized based on their suitability for various age groups. The most frequent ratings are TV-MA and TV-14, which together represent a substantial portion of the platform's offerings. These two ratings indicate that Netflix's catalog is heavily geared toward mature audiences, with a strong emphasis on adult and young adult content.

- TV-MA (Mature Audience) is applied to content that may contain graphic violence, strong language, or sexual content—intended strictly for adult viewers.
- TV-14 is suitable for viewers over 14 and may include moderately strong language or violence.

The high concentration of these ratings suggests a strategic content focus: Netflix seems to prioritize edgier, high-engagement content that appeals to teens, millennials, and adult demographics the most active streaming audience segments.

4.6 Movie Duration Analysis

Most movies range between 80–120 minutes, aligning with standard feature lengths.



This histogram visualizes the distribution of movie durations on Netflix, offering insight into how long the typical movie on the platform runs. The data shows a clear **concentration of titles between 80 and 120 minutes**, which aligns with the standard runtime for most feature films globally. The distribution follows a **right-skewed pattern**, where the frequency of movies sharply declines after the 2-hour mark.

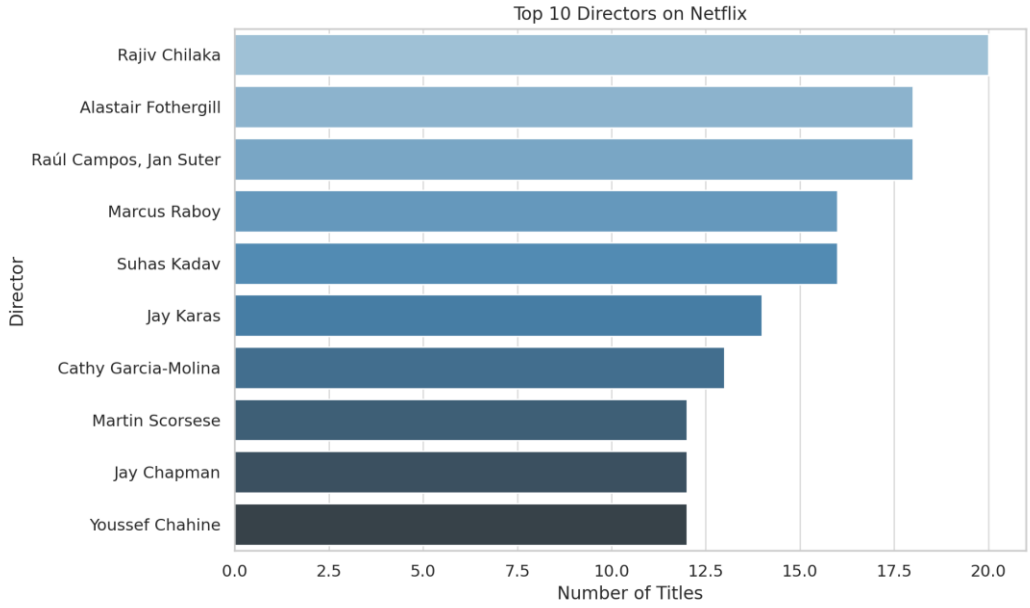
The peak of the histogram occurs around the **90-minute** mark, indicating that a significant number of Netflix movies are designed to be concise, digestible, and accessible for a wide range of viewers. This duration is also optimal for mobile or evening viewing, where time constraints influence user preferences.

Movies longer than 150 minutes are relatively rare, which may be a reflection of both user attention span and platform strategy. Longer formats tend to perform better in theatrical settings or are broken into episodic content on streaming platforms. From a platform design and recommendation perspective, understanding this duration trend is important. It helps tailor content suggestions to the time users are likely to commit. For example, a user browsing in the evening may be more inclined to select a 90-minute film over a 2.5-hour one.

Business-wise, this distribution confirms that Netflix prioritizes **efficient storytelling** — enough time to build a compelling narrative without the fatigue associated with longer runtimes. For content producers, this insight can help guide script length, pacing, and editing decisions tailored for streaming audiences.

6.7 Top Directors by Title Count

Directors like Rajiv Chilaka and Alastair Fothergill appear most frequently, known for children's and documentary content.



This visualization ranks the top 10 directors on Netflix based on the number of titles attributed to them. By filtering out records where the director was listed as "Unknown", we focused on named contributors to uncover the most frequently featured filmmakers on the platform. At the top of the list are **Rajiv Chilaka** and **Alastair Fothergill**, both of whom have a distinctive presence on Netflix.

- **Rajiv Chilaka** is known for his work in children's programming, particularly the *Chhota Bheem* series, which contributes to his high title count.
- **Alastair Fothergill** is a leading producer of nature documentaries, responsible for widely acclaimed series like *Our Planet*, which reflect Netflix's investment in high-quality educational and environmental content.

The presence of other directors such as **Raúl Campos & Jan Suter** (known for Latin American stand-up specials), **Marcus Raboy**, and **Jay Chapman** suggests that Netflix has also heavily invested in **comedy and live performance content**. The list also includes globally recognized names like **Martin Scorsese**, reflecting the platform's ability to attract critically acclaimed directors for original films and documentaries.

This analysis provides insight into **Netflix's content strategy from a production perspective**. Rather than being dominated solely by mainstream film directors, the list shows a deliberate balance of creators across genres—children's content, documentaries, comedy specials, and award-winning cinema. It demonstrates Netflix's commitment to **niche audiences as well as prestige storytelling**.

7. Conclusion & Future Scope

This analysis provides a clear view of Netflix's content strategy and growth trajectory over the years. By exploring a wide array of variables from content type and genre to geographic distribution, content duration, and maturity ratings we have uncovered patterns that reflect both user demand and strategic decision-making by the platform. The dominance of movies over TV shows, the strong representation of content from the United States and India, the popularity of genres like Drama and International Movies, and the focus on mature audience ratings together outline Netflix's evolving positioning as a global content leader.

One of the key takeaways is the evident shift in Netflix's approach to content acquisition and production, especially during the 2016–2020 period. The sharp rise in content additions during these years indicates a phase of rapid scaling and market expansion. Additionally, the diversity of directors, runtime trends, and genre tags demonstrate Netflix's effort to balance quantity with diversity offering content that caters to both mass and niche audiences.

These insights do not just serve as descriptive observations; they offer a solid foundation for advanced machine learning applications. For instance:

- A **content recommendation system** can be built using user watch history, genre preferences, and content duration.
- **Classification models** can be trained to predict content type (movie or TV show), genre, or even regional success based on metadata features.
- **Clustering algorithms** could be applied to group similar types of content or user profiles for targeted marketing.
- **Forecasting models** can help estimate future content demand by region or genre, guiding production and licensing decisions.

In future iterations, integrating **user engagement data** (such as ratings, watch time, or likes) would allow for deeper sentiment analysis and popularity scoring. Additionally, comparing this dataset with data from competitors like Amazon Prime or Disney+ could uncover unique positioning opportunities or content gaps.

In conclusion, this project bridges the gap between raw metadata and meaningful business intelligence. It not only reflects Netflix's historical content patterns but also sets the stage for data-driven decision-making, product optimization, and strategic innovation in the competitive streaming industry.