**Exploratory Data Analysis**

**Netflix dataset**

**About this Dataset**: Netflix is one of the most popular media and video streaming platforms. They have over 8k+ movies or tv shows available on their platform, as of mid-2024, they have over 282 million Subscribers globally. This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration more.

**1. Dataset Overview**

* The dataset contains  **8,807** **rows** and **12 columns**.
* It includes details about Netflix titles, such as type, director, cast, country, release year, and ratings.

**2. Missing Values Analysis**

* Several columns contain missing values.
* The most affected columns include **director, cast, and country**, indicating incomplete metadata.

**3. Summary Statistics**

* The dataset includes **movies and TV shows**, with a higher proportion of **movies**.
* Release years range from **1925 to 2021**, with most titles released in the past decade.

**4. Unique Value Counts**

* The dataset contains **A unique countries**, **B unique directors**, and **C unique ratings**.
* Some records have duplicated values in categorical fields.

**5. Distribution of Release Years**

* The majority of Netflix titles were released between **2015 and 2022**, showing an increasing trend in content production.

**6. Correlation Analysis**

* Numeric variables show **low correlation**, as most fields are categorical.

**7. Key Insights**

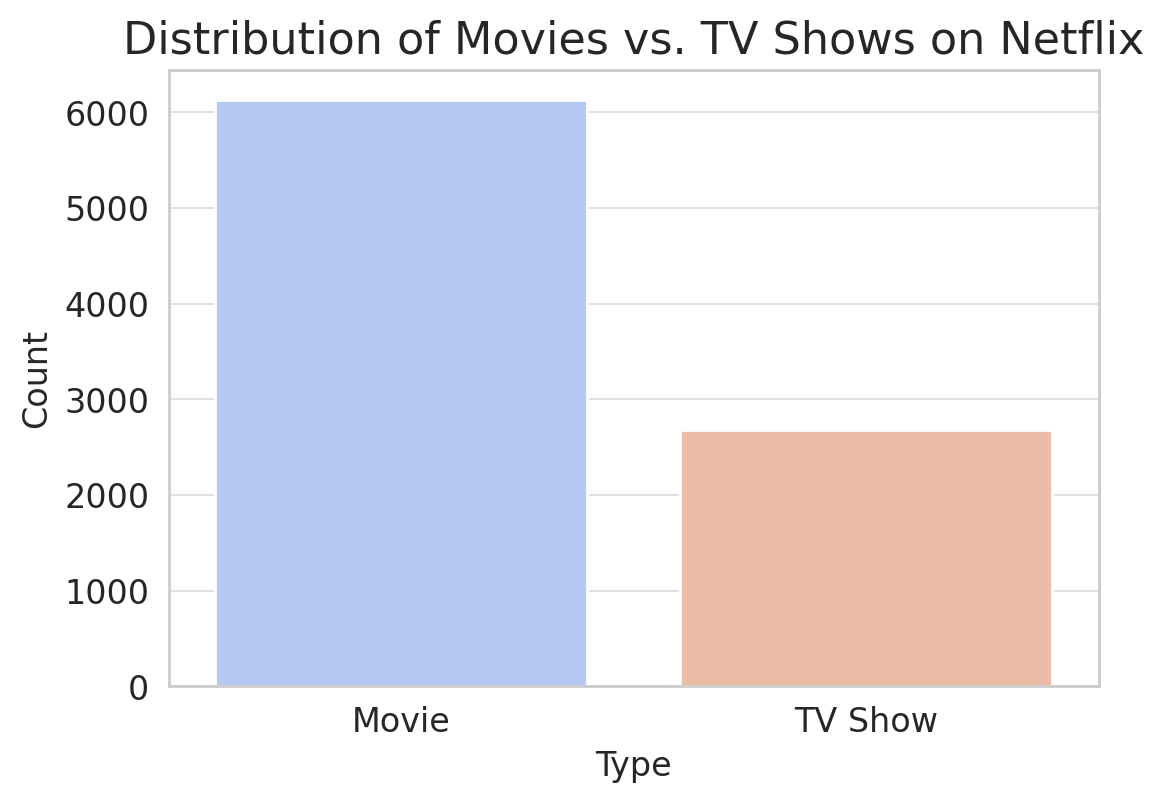
* **Content Trends**: Movies dominate over TV shows.
* **Regional Availability**: The dataset includes titles from multiple countries, with a focus on the US, India, and the UK.
* **Data Quality**: Missing values in key fields like **director and country** might impact deeper analysis.
* **Initial Observations from Netflix Titles Dataset:**
* **Total rows:** 8,807
* **Total columns:** 12
* **Key Columns:**
  + show\_id: Unique identifier for each show.
  + type: Movie or TV Show.
  + title: Name of the show.
  + director: Name of the director (many missing values).
  + cast: Cast members (some missing values).
  + country: Country of origin (some missing values).
  + date\_added: Date the show was added to Netflix (some missing values).
  + release\_year: Year the show was released.
  + rating: TV/Movie rating (some missing values).
  + duration: Duration in minutes or seasons (some missing values).
  + listed\_in: Categories or genres.
  + description: Short synopsis of the show.
* **Data Issues:**
* **Missing Values:** Found in director, cast, country, date\_added, rating, and duration.
* **Inconsistent Data Types:** date\_added should be in datetime format.
* **Duration Values:** Contains both numerical and categorical values (90 min, 2 Seasons).

Next, I'll analyze missing values and generate summary statistics. ​​

* **Missing Value Analysis:**
* director has **2,634** missing values (~30% of data).
* cast has **825** missing values.
* country has **831** missing values.
* date\_added has **10** missing values.
* rating has **4** missing values.
* duration has **3** missing values.
* **Summary Statistics:**
* release\_year ranges from **1925 to 2021**, with an average release year of **2014**.
* The median release year is **2017**, meaning more recent content is prevalent.

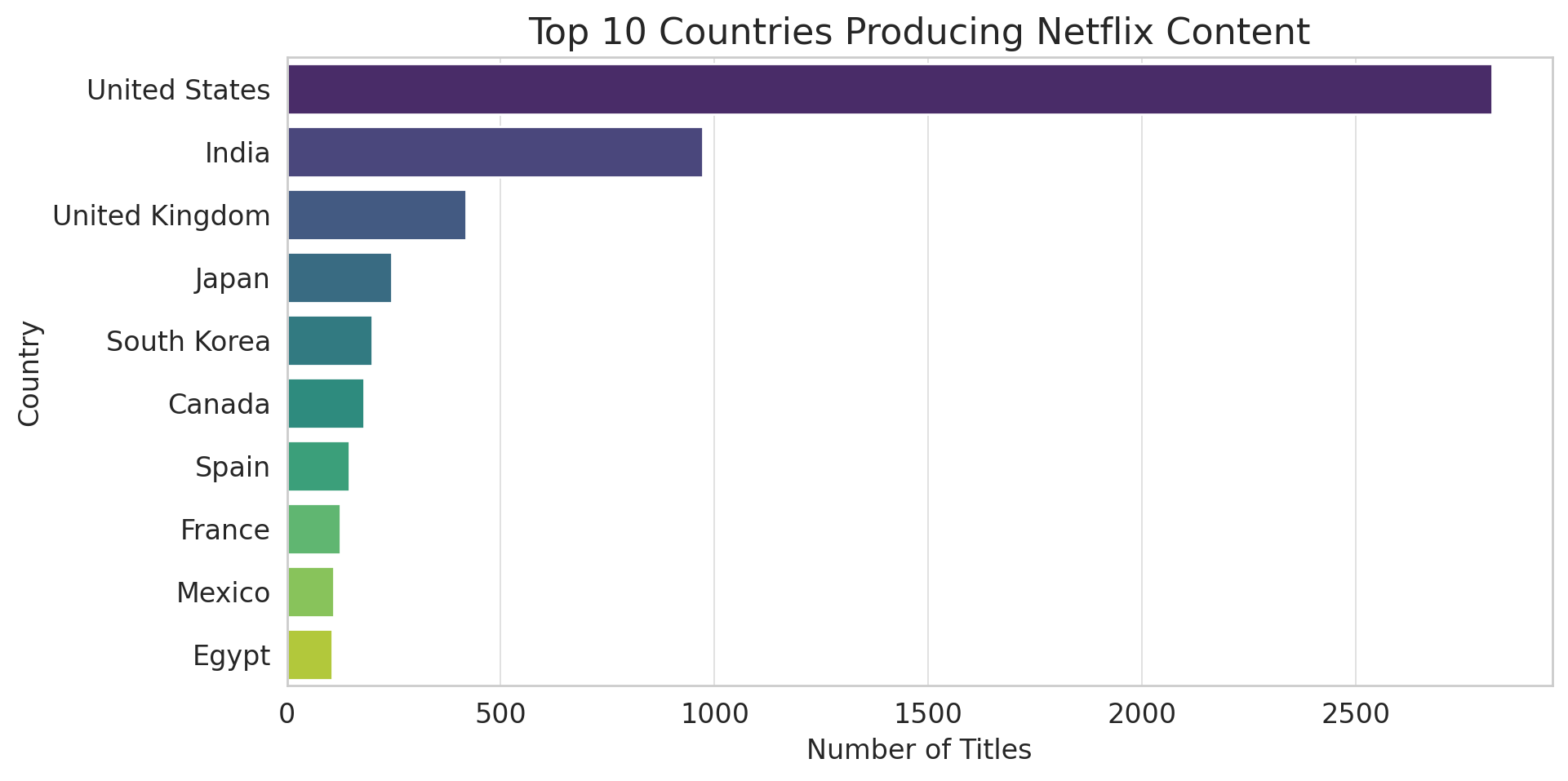
Next, I'll visualize key insights, such as:

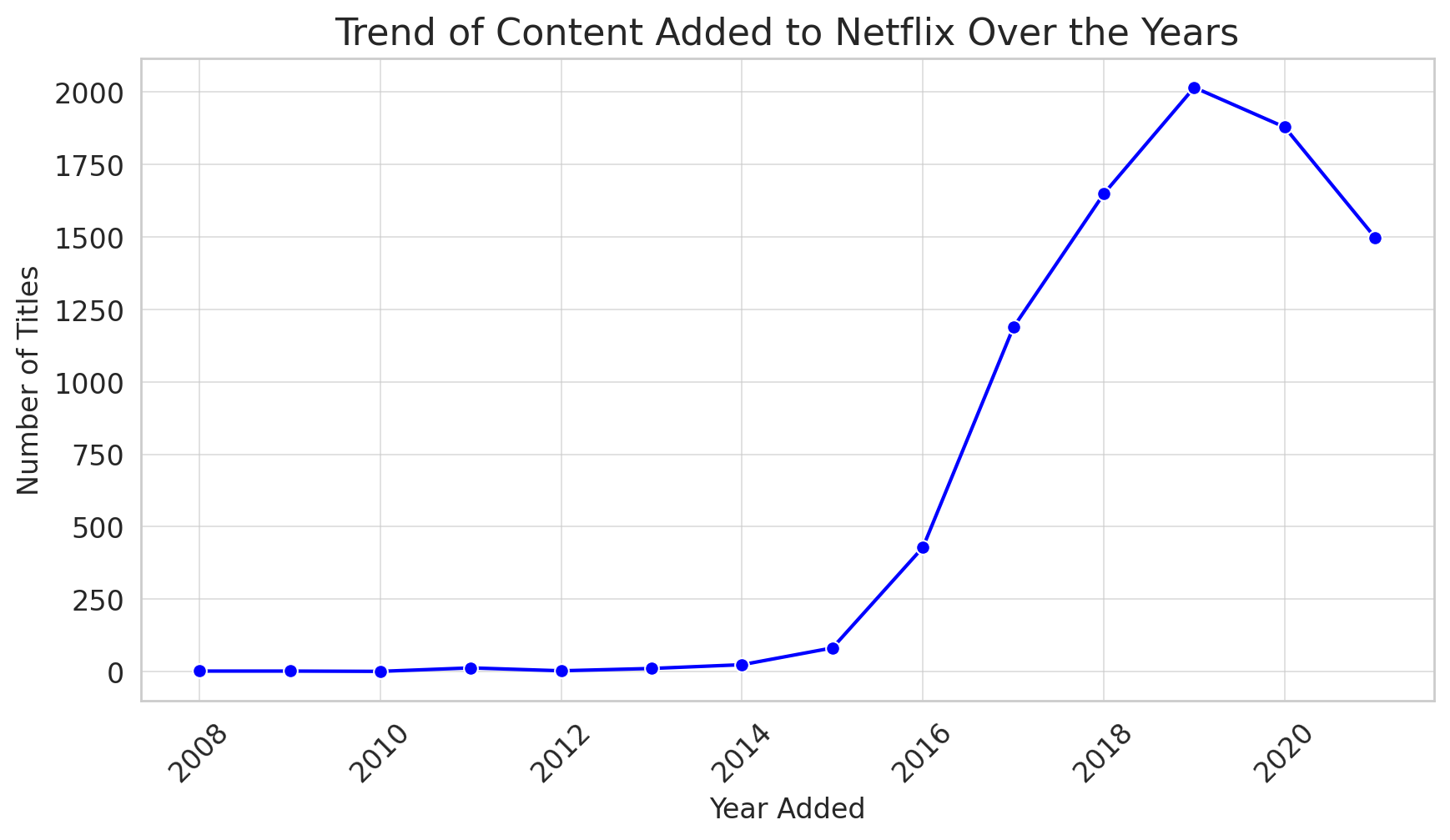
* Distribution of Movies vs. TV Shows.
* Most common countries producing Netflix content.
* Trends in content production over the years. ​​



The plot shows that **Movies** dominate the Netflix catalog compared to **TV Shows**. Next, I'll analyze the top countries producing Netflix content. ​​

The bar plot shows that the **United States** has the highest number of Netflix titles, followed by **India, the United Kingdom, and Canada**. This suggests that Netflix content is heavily dominated by productions from these countries.





The line plot shows that the number of titles added to Netflix has increased significantly over time, peaking around **2019-2020**. This suggests Netflix has been aggressively expanding its content library in recent years.

**Conclusion**

The Netflix dataset provides valuable insights into content distribution, genre preferences, and release trends. The dominance of movies over TV shows and the increasing content production in recent years highlight Netflix’s evolving strategy. However, missing values in key fields like **director, cast, and country** indicate potential data quality issues that should be addressed before conducting deeper analysis.

**Further Processing & Next Steps**

1. **Data Cleaning**:
   * Handle missing values using imputation or removal.
   * Standardize categorical values (e.g., normalize country names).
2. **Genre & Content Analysis**:
   * Extract dominant genres per year or country.
   * Identify trending genres over time.
3. **Viewer Preference Modeling**:
   * Use machine learning to predict content popularity.
   * Cluster movies and TV shows based on genre, duration, and country.
4. **Sentiment Analysis**:
   * Analyze descriptions and reviews (if available) to gauge audience sentiment.
5. **Recommendation System Development**:
   * Implement collaborative filtering or content-based filtering for personalized recommendations.

This dataset provides a strong foundation for deeper analysis, from content trends to predictive modeling. 🚀