Netflix Data Analysis

Exploring content trends, genre popularity, and distribution patterns on Netflix. This presentation provides key insights into the platform's diverse library.

Explore the full analysis in our Colab Notebook.

SS by Shubham Kumar

Dataset Overview

We began by loading the comprehensive Netflix content dataset. Initial exploration revealed its fundamental structure. This crucial first step underpins our entire analysis.

Data Loading

We ingested the raw CSV file into a pandas DataFrame. This established the foundation for our analysis.

Structural Insight

The `.info()` method provided column types and non-null counts. It helped assess data completeness and consistency.

Statistical Summary

Using `.describe()` gave us key statistics for numerical columns. This offered quick insights into data distribution.

Pandas Methods Used in the Analysis

1. **pd.read_csv()**
   Purpose: To load the Netflix dataset from a CSV file into a DataFrame.
   Usage: This is the starting point for data analysis using Pandas.

2. **df.head()**
   Purpose: Displays the first 5 rows of the DataFrame.
   Usage: Used for a quick overview of the dataset structure and sample data.

3. **df.info()**
   Purpose: Shows column names, data types, and non-null values.
   Usage: Helps assess data completeness and understand the schema.

4. **df.describe()**
   Purpose: Generates statistical summary for numerical columns.
   Usage: Useful for understanding data distribution (mean, std, min, max, quartiles).

5. **df.columns**
   Purpose: Lists all column names in the DataFrame.
   Usage: Handy to understand what features are available in the dataset.

6. **df.isnull().sum()**
   Purpose: Counts the number of missing (null) values in each column.

Dataset Overview

📌 About the Dataset

- The dataset contains **8,807 entries** and **12 columns**.
- It provides metadata about content available on **Netflix**, including both **Movies** and **TV Shows**.

📜 Key Columns

- **show_id** – Unique identifier for each title
- **type** – Whether the content is a *Movie* or *TV Show*
- **title** – Name of the content
- **director** – Director's name (can be multiple or missing)
- **cast** – Lead actors or cast members
- **country** – Country of origin
- **date_added** – Date when the content was added to Netflix
- **release_year** – Year the title was originally released
- **rating** – Age rating (e.g., TV-MA, PG-13)
- **duration** – Length in minutes (movies) or seasons (TV shows)
- **listed_in** – Genres/categories
- **description** – Short summary of the content

Data Cleaning: Missing Values

Director & Cast

Missing director and cast values were filled with "unknown".

Country

Missing country values were also replaced with "unknown".

Date Added

10 missing "date_added" entries were manually updated to datetime objects.

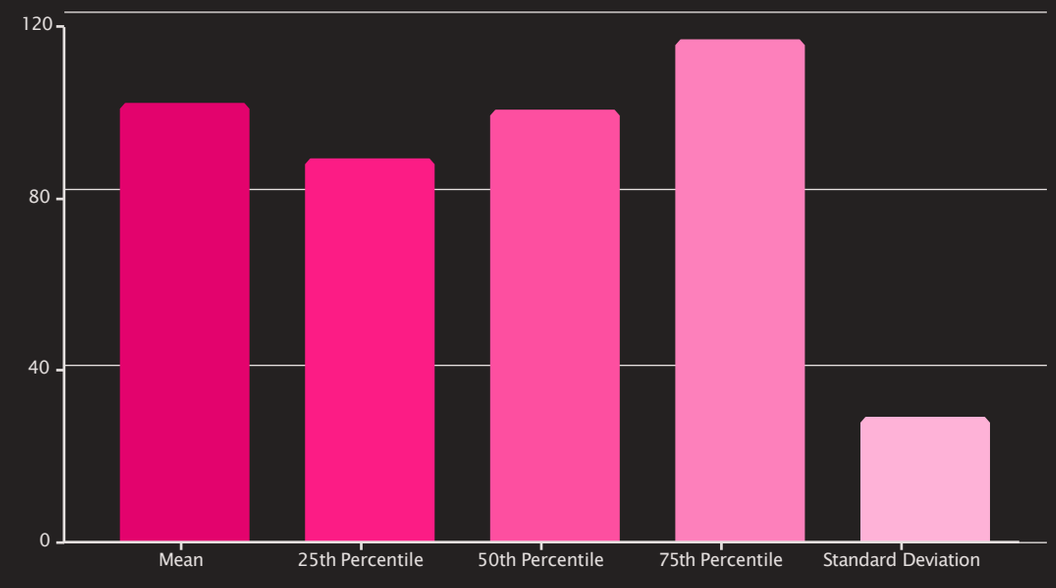
Rating & Duration

4 missing ratings and 3 missing durations were imputed based on content.

Movie Runtime Insights

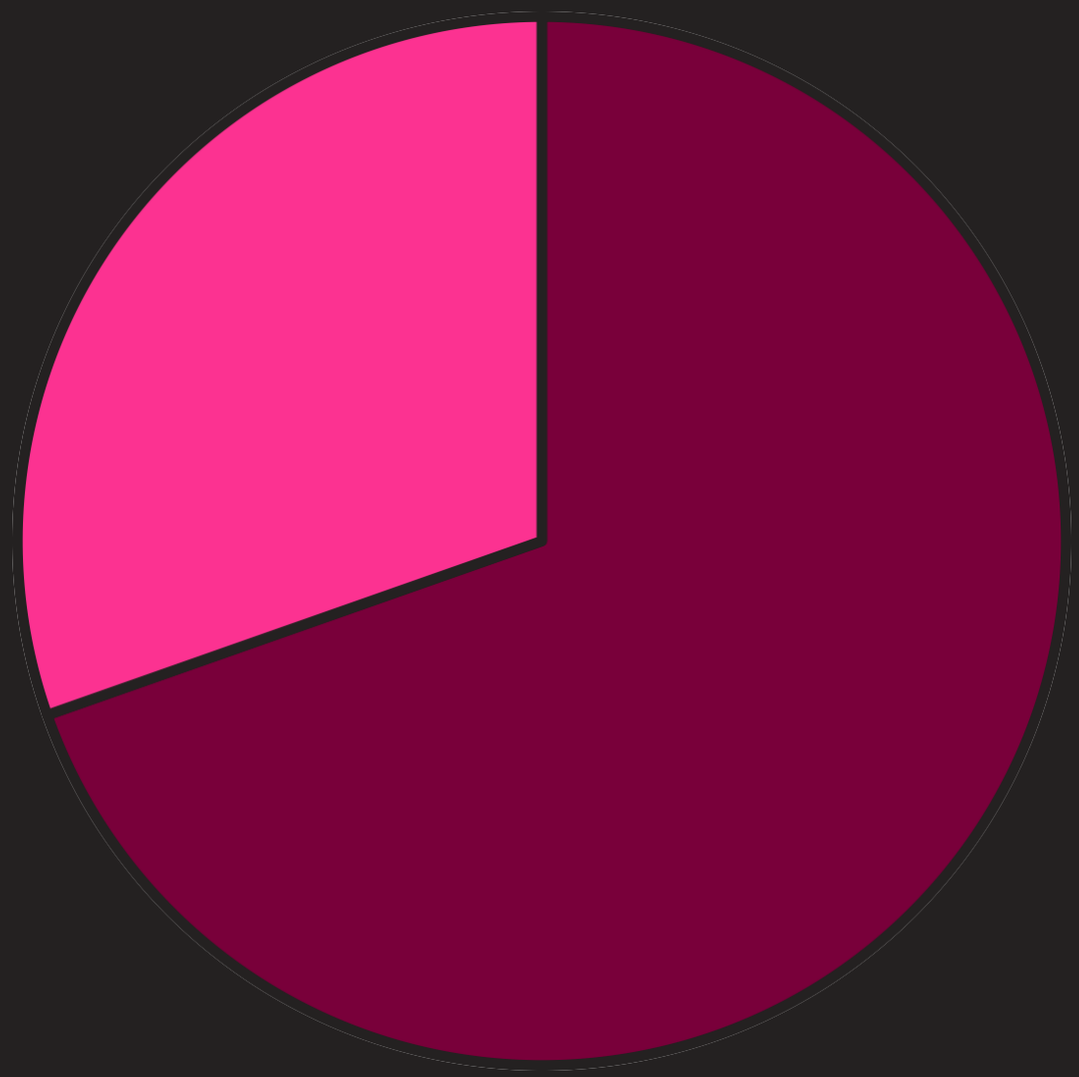
The average movie runtime is approximately 99.56 minutes. Most movies fall between 75 and 137 minutes.

The standard deviation for movie runtime is 28.28 minutes, indicating some variability.



Content Type Distribution

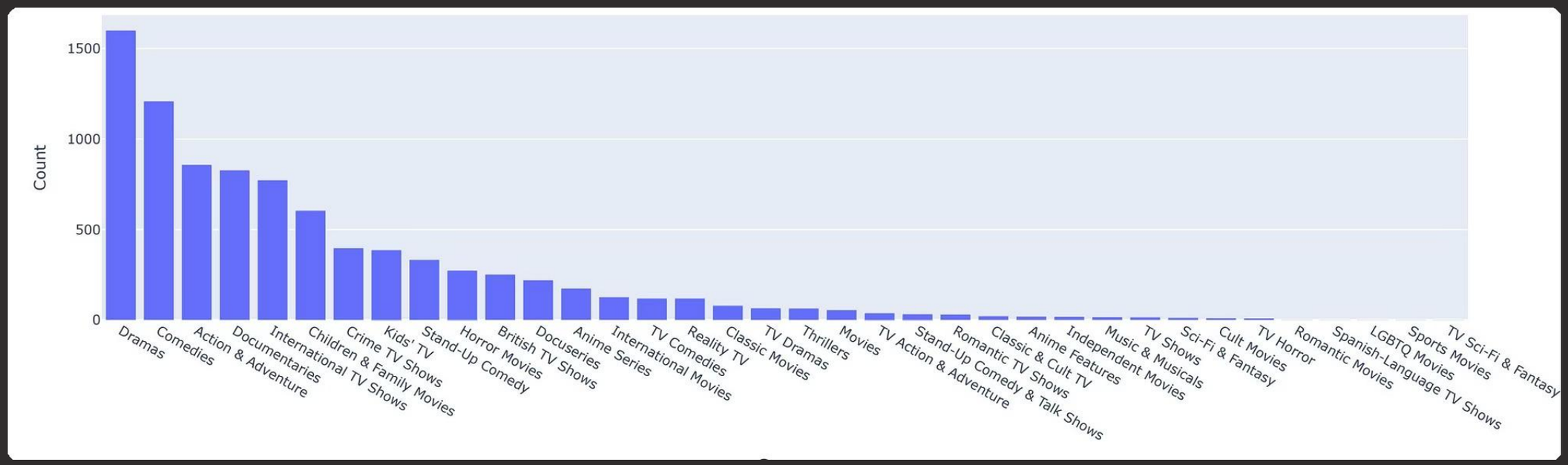
Netflix primarily features Movies, with 6131 titles. TV Shows account for 2676 titles.



Top Genres on Netflix

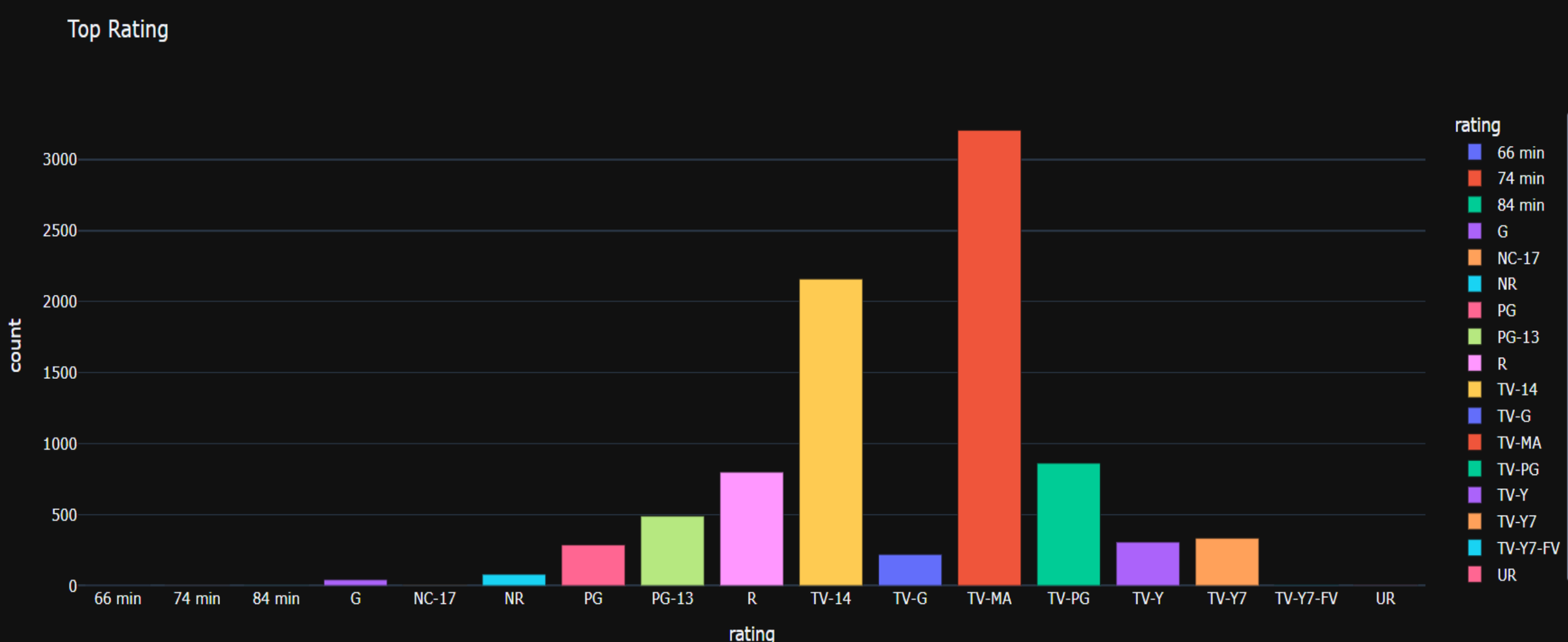
Dramas lead with 1600 titles, followed by Comedies (1210) and Action & Adventure (859). Documentaries and International TV Shows are also highly represented.

- **Dramas: 1600**
- **Comedies: 1210**
- **Action & Adventure: 859**
- **Documentaries: 829**
- **International TV Shows: 774**



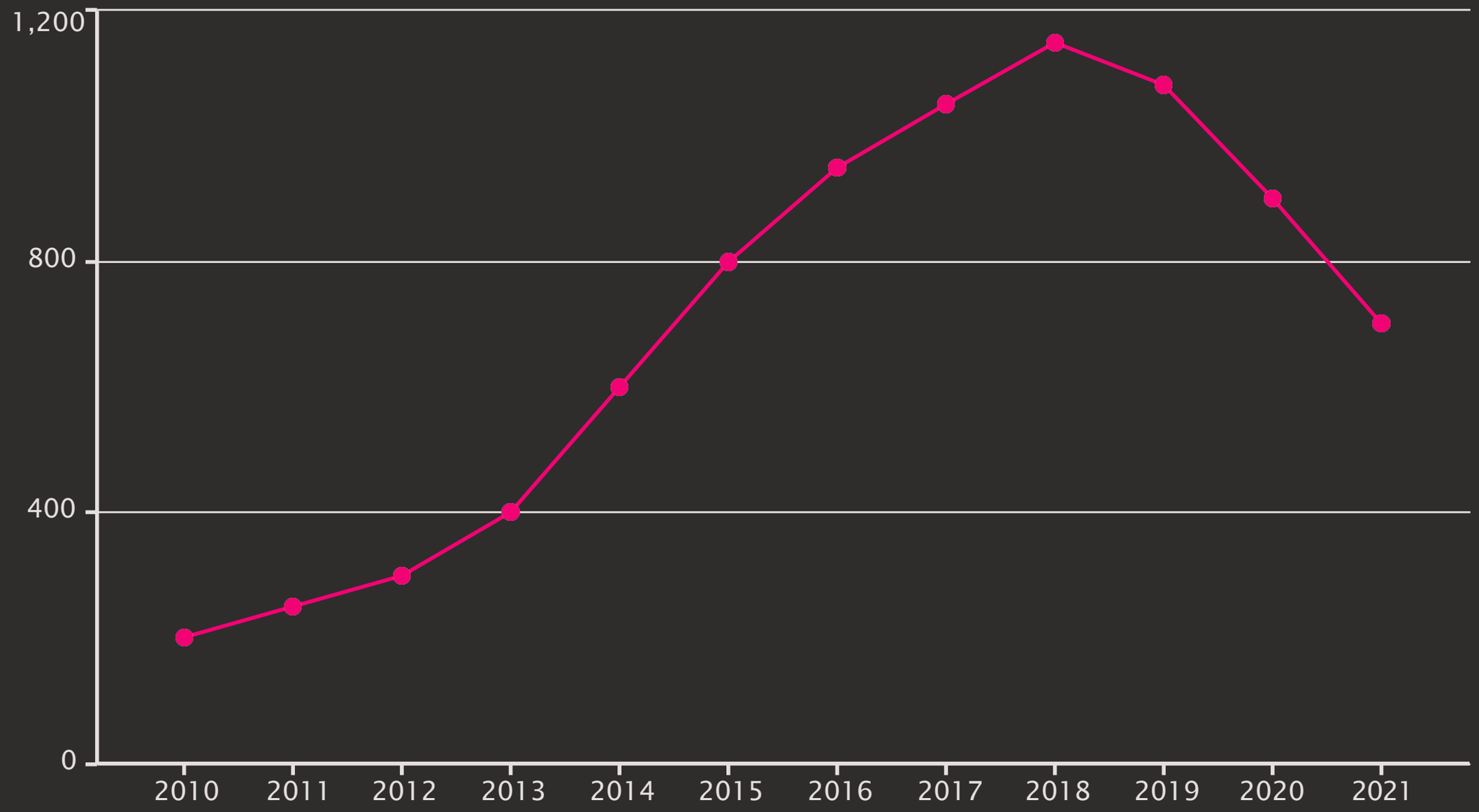
📊 Top Content Ratings on Netflix

- **TV-MA** is the most common rating, with over **3,000 titles**, indicating a large volume of mature content.
- **TV-14** follows next, highlighting Netflix's focus on teenage and adult audiences.
- **R** and **PG-13** ratings are also significant, showing a mix of restricted and parental-guidance content.
- Family-friendly ratings like **TV-Y**, **TV-Y7**, and **TV-G** are present but much fewer in number.
- Very few titles are rated **G**, **NR**, or **NC-17**, suggesting limited content for very young or adult-only audiences.
- A few outlier labels like **“84 min”, “66 min”** are possibly data errors or misclassified entries.



Content Release Year Trends

The volume of content added to Netflix has steadily increased over the years. As the chart below illustrates, 2018 saw the most releases with 1147 titles.

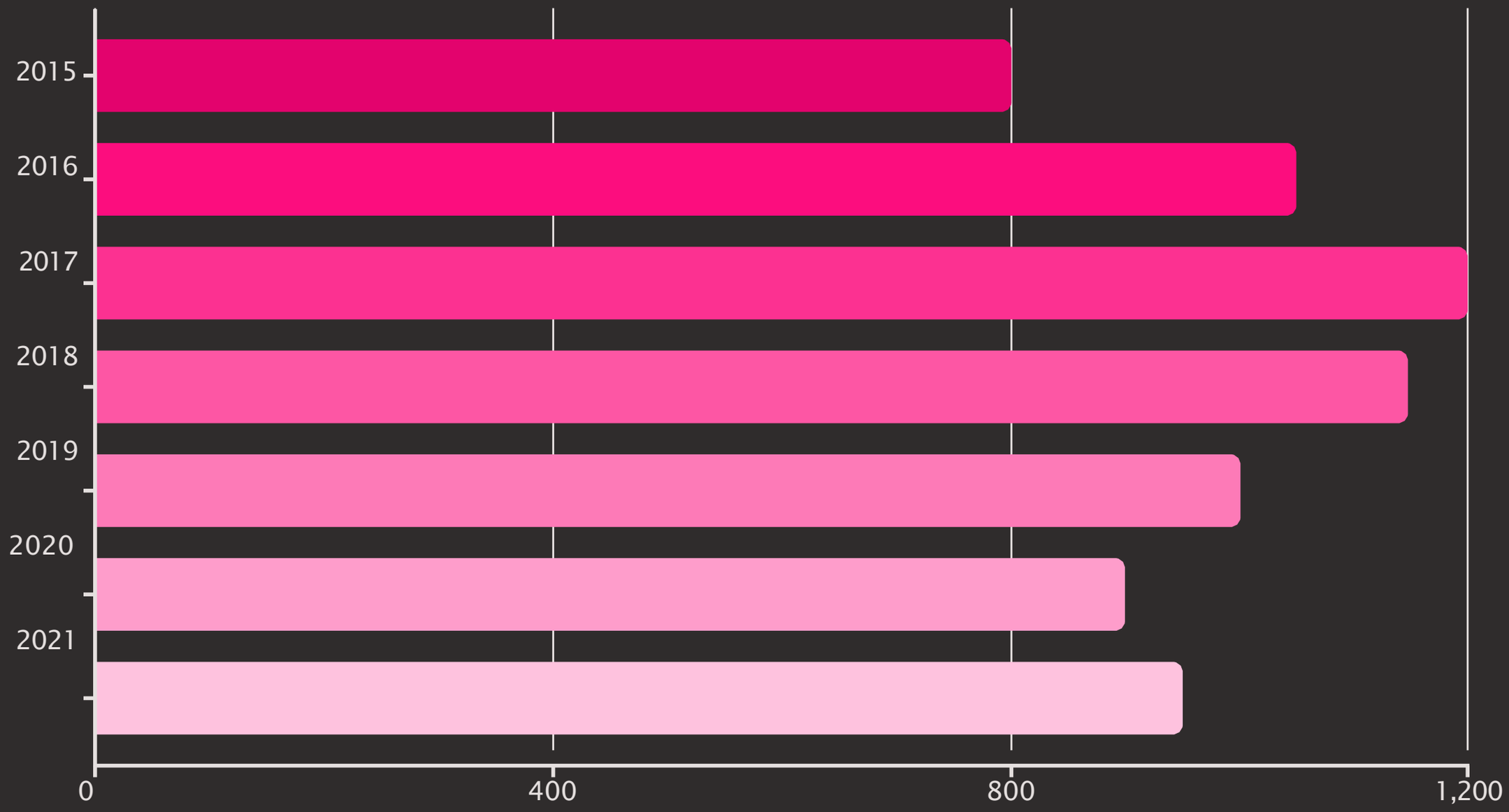


Global Content Distribution

The United States is a significant contributor.
India is emerging as a major content creator.

Content Added Over Time

The number of titles added to Netflix has shown a consistent upward trend. This demonstrates Netflix's commitment to expanding its content library.



Key Takeaways

- **Diverse Content**

  Netflix offers a wide range of international movies and TV shows.

- **Genre Popularity**

  Dramas, comedies, and documentaries are the most common genres.

- **Growth Trend**

  Content volume has steadily increased, especially for movies.

- **Content Ratings**

  A balance exists between mature and family-friendly options.

📌 Conclusion

1.Diverse Library: Netflix hosts a vast range of content, with a strong representation of both movies and TV shows, catering to a global audience.

2.Dominant Genres: Dramas, comedies, and documentaries are the most prevalent genres, suggesting a focus on storytelling and informative content.

3.Global Presence: While the U.S. remains a major contributor, countries like India are emerging as significant content producers, showing Netflix's global expansion.

4.Consistent Growth: The number of titles added has increased year-over-year, especially around 2018, reflecting Netflix's investment in original and diverse content.

5.Content Duration: Most movies are within a viewer-friendly runtime of 75–137 minutes, indicating a focus on engaging but concise storytelling.

6.Data Challenges: The presence of missing values required careful cleaning and imputation, showing the importance of preprocessing in any real-world dataset.