EDA ANALYSIS ON NETFLIX_TITLES DATASET

1.Introduction:

The dataset used in this analysis is the **Netflix Movies and TV Shows** dataset (by Shivam Bansal, Kaggle). It contains detailed information about content available on Netflix up to 2021, covering both movies and television series. Key features include: *title, content type, director, cast, country, release year, date added*, *rating, duration, listed genres*, and *description.*
As of mid-2021, Netflix had over **8,000 titles** and more than **200 million subscribers worldwide**. This tabular dataset is valuable for analyzing Netflix's content portfolio—a mix of global and regional titles across time, genres, and formats

The aim of this EDA report is to:

- Provide a **comprehensive overview** of the dataset.
- Perform data cleaning and feature preparation to ensure reliability.
- Explore key patterns such as content growth over the years, genre distribution, country-wise production trends, and duration analysis.
- Highlight insights that inform and justify building a content-based recommender system.

By diving into structured data (e.g., types, release years, ratings) and unstructured metadata (e.g., genre lists, textual descriptions), this EDA lays the foundation for a recommendation engine that explains *why* certain titles are similar—making your final project both technically robust and business-relevant.

2.Dataset Overview:

The Netflix dataset, sourced from Kaggle, provides a comprehensive list of movies and TV shows available on the Netflix platform as of 2021. It is a structured CSV file that contains detailed metadata for each title, allowing for various types of analysis, including content distribution, trends over time, and recommendation systems.

2.1.Dataset Structure

The dataset consists of **one file**: netflix_titles.csv.

Column Name	Description
show_id	Unique ID assigned to each show/movie
type	Indicates whether the entry is a Movie or a TV Show
title	Title of the movie or TV show
director	Name(s) of the director(s)
cast	Lead actors/actresses in the title
country	Country of origin
date_added	Date when the title was added to Netflix
release_year	The year the title was released
rating	Age classification (e.g., TV-MA, PG, R)
duration	Duration of the movie (in minutes) or number of seasons for TV shows
listed_in	Genre(s) or categories assigned to the title
description	Short summary or description of the title

2.2.Dataset Dimensions:

- **Rows:** 8,807
- **Columns:** 12

```python
import pandas as pd

# Load the dataset
df = pd.read_csv('netflix_titles.csv')

# View top rows
df.head()

# Check data types and null values
df.info()
df.isnull().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

```
memory usage: 825.8+ KB
```

	0
show_id	0
type	0
title	0
director	2634
cast	825
country	831
date_added	10
release_year	0
rating	4
duration	3
listed_in	0
description	0

dtype: int64

2.3.Key Features for Analysis

- **Temporal Attributes:** date_added and release_year help us study trends over time.
- **Categorical Attributes:** type, rating, country, and listed_in are essential for classification and content filtering.
- **Text Attributes:** title and description can be used for content-based recommendation systems.
- **Relational Attributes:** cast and director allow us to examine content based on key contributors.

3. Data Cleaning and Preprocessing:

Before performing any analysis, it is essential to clean and preprocess the dataset to ensure accuracy and consistency in the results. The following steps were taken:

3.1. Handling Missing Values

Several columns in the dataset contained missing values:

- director: Many entries were NaN, especially for TV Shows. These were filled with "Not Available".
- cast: Missing values were filled with "Not Available".
- country: Some entries lacked a country, which were filled with "Unknown".
- date_added: Missing values were imputed using "Not Available" or retained for time-based filtering if needed.
- rating: A small number of entries were missing; these were filled with "Not Rated" to maintain consistency.

3.2. Standardizing Column Formats

- date_added was converted to datetime format to allow time-series analysis.
- duration was split into duration_int and duration_type (e.g., minutes or seasons) for better granularity.
- Text columns such as title, director, and cast were converted to lowercase for uniformity.

3.3. Removing Duplicates

- Duplicate entries (based on the same title and type) were dropped to prevent biased analysis.

3.4 Extracting Useful Features

To improve analysis and build a recommendation system:

- release_year was converted to numeric.
- listed_in was split into multiple genres using a delimiter (,).
- New columns like year_added and month_added were extracted from date_added.
- A genre_primary column was created using the first listed genre for high-level analysis.

3.5 Encoding for Modeling

- For recommendation modeling, we used TF-IDF encoding on the description column.
- genre, cast, and director were considered for similarity-based recommendations using cosine similarity.

```
# Clean missing values
df['country'] = df['country'].fillna('Unknown')
df['cast'] = df['cast'].fillna('Not Specified')

# Strip and convert dates
df['date_added'] = df['date_added'].str.strip()
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
```

```
df['date_added'].isna().sum()
```

```
np.int64(10)
```

4.Exploratory Data Analysis (EDA) :

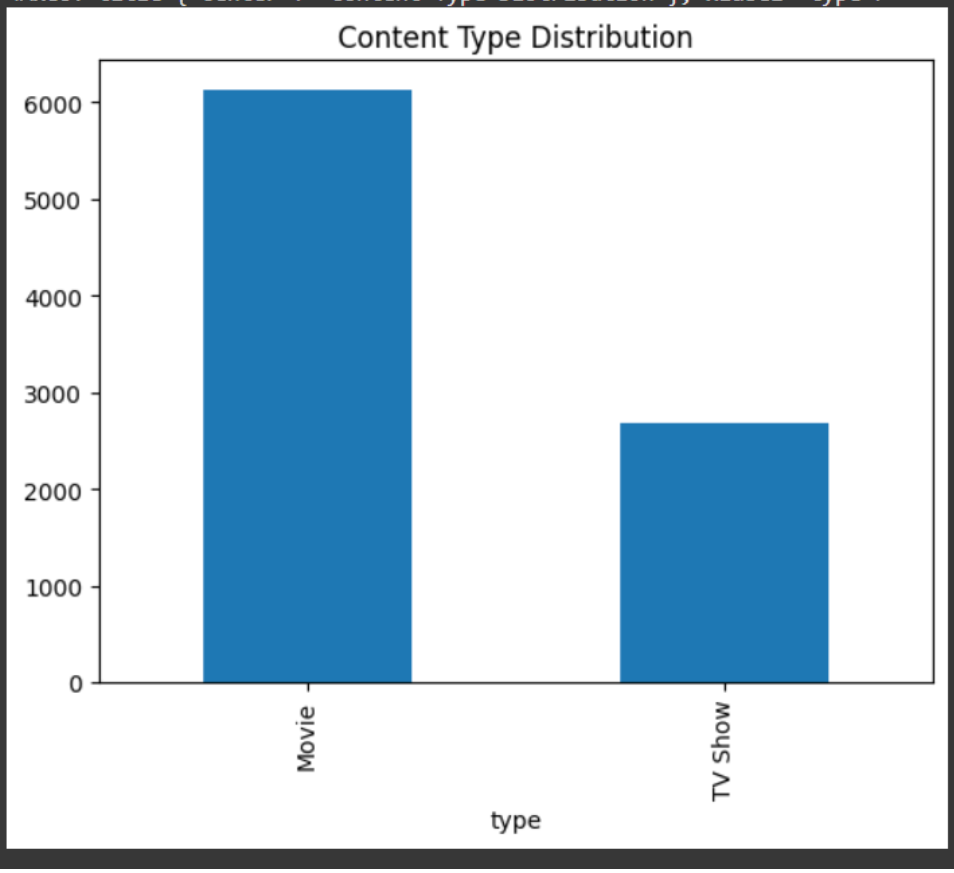
In this section, we explore the Netflix dataset to uncover trends, patterns, and insights related to the content available on the platform. The analysis helps us understand what types of content are available, how they are distributed over time and geography, and who the key contributors (directors, actors) are.

4.1.Distribution of Content Types

Netflix offers both **Movies** and **TV Shows**. Understanding the proportion between the two helps us know what dominates the platform.

```
df['type'].value_counts().plot(kind='bar', title='Content Type Distribution')
```

```
<Axes: title={'center': 'Content Type Distribution'}, xlabel='type'>
```

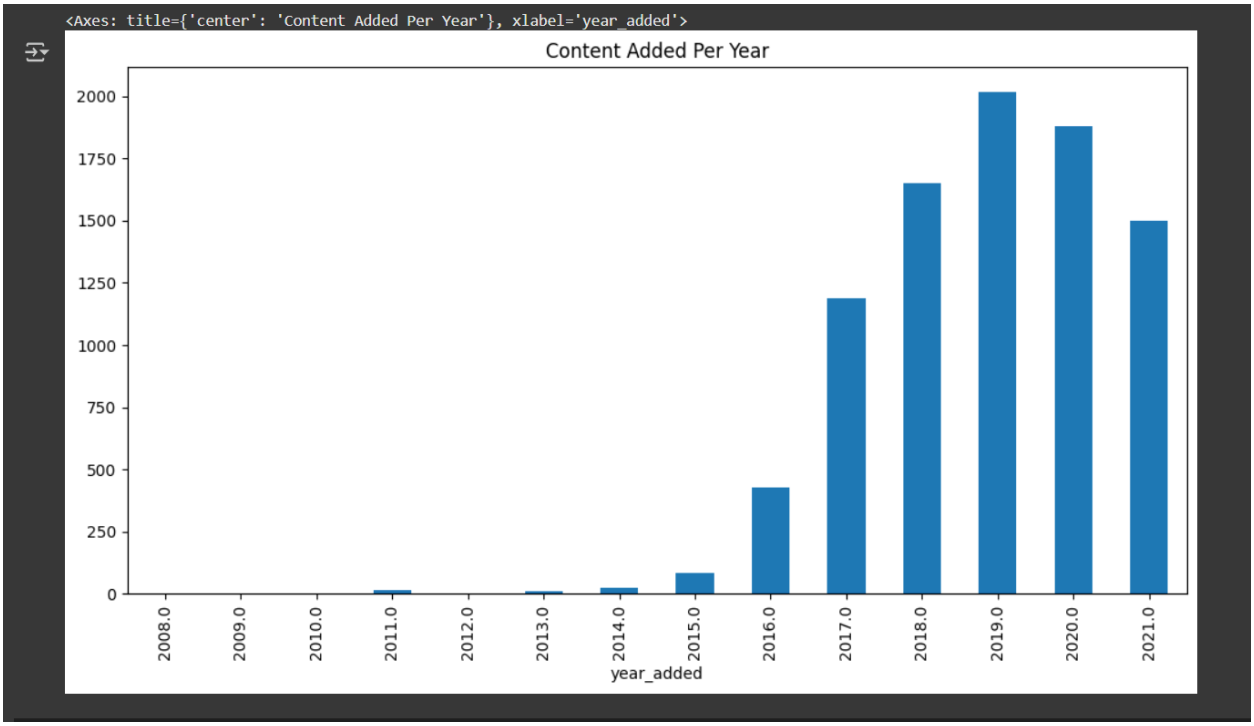


Insight: Movies dominate the platform, with roughly two-thirds of the total content, while TV Shows make up the remaining one-third.

4.2. Content Added Over the Years

Analyzing the date_added column reveals how Netflix's content library has grown over time.

```
[ ]  df['year_added'] = df['date_added'].dt.year
     df['year_added'].value_counts().sort_index().plot(kind='bar', figsize=(12,6), title='Content Added Per Year')
```



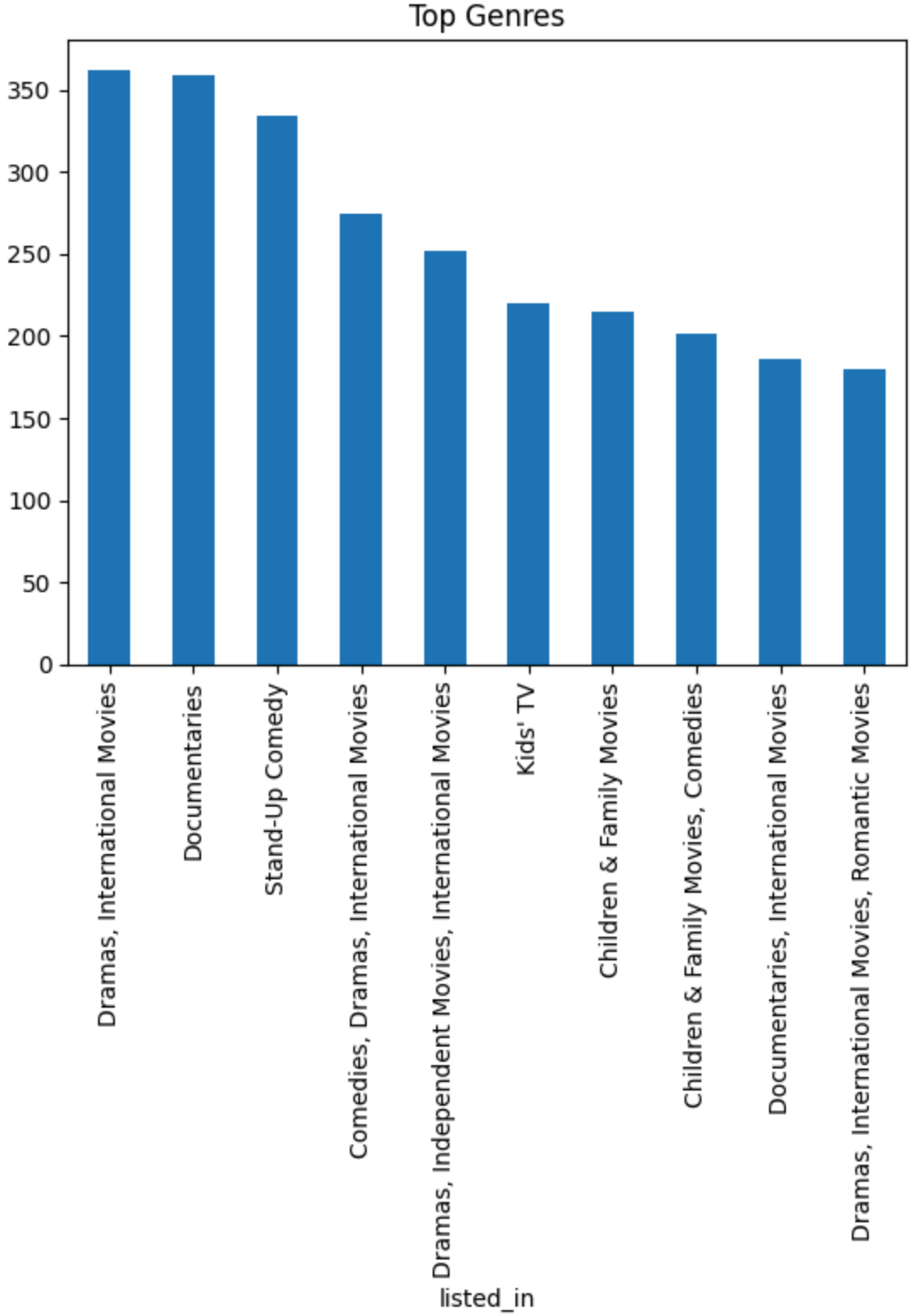
Insight: There was a noticeable surge in content additions after 2015, with the peak around 2019–2020. This reflects Netflix's aggressive expansion and content acquisition during that time.

4.3 Top Genres by Count

The listed_in column contains genre tags. By exploding the multi-genre strings, we can identify the most common genres.

```
df['listed_in'].value_counts().head(10).plot(kind='bar', title='Top Genres')
```

```
<Axes: title={'center': 'Top Genres'}, xlabel='listed_in'>
```


Top Genres
350
300
250
200
150
100
50
0
Dramas, International Movies
Documentaries
Stand-Up Comedy
Comedies, Dramas, International Movies
Dramas, Independent Movies, International Movies
Kids' TV
Children & Family Movies
Children & Family Movies, Comedies
Documentaries, International Movies
Dramas, International Movies, Romantic Movies
listed_in

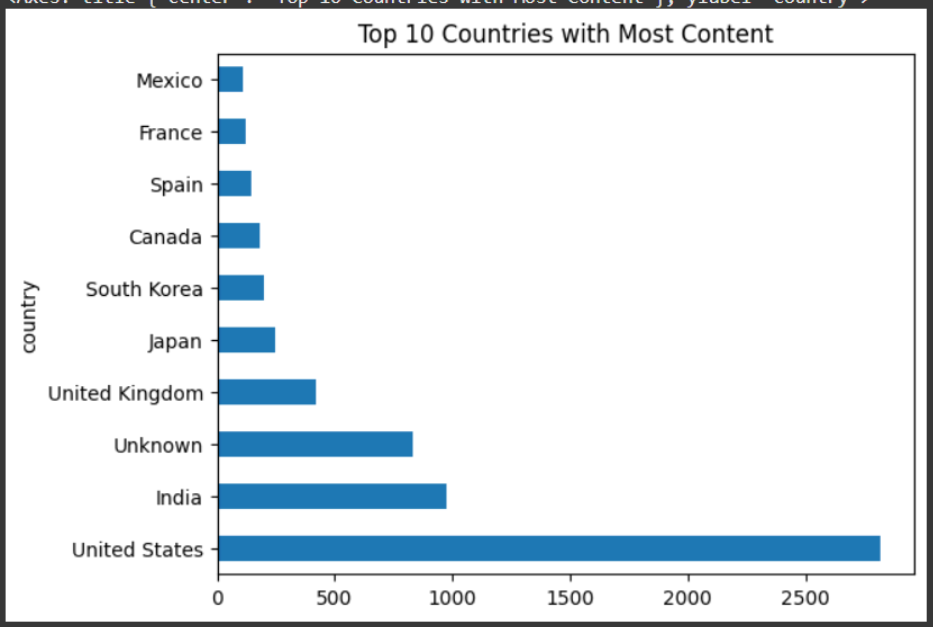
Insight: The most frequent genres include Documentaries, Dramas, Comedies, and International content, showing Netflix's global strategy and genre diversification.

4.4 Top Countries Producing Content

Using the country column, we can find which countries are the most frequent content producers.

```
top_countries = df['country'].value_counts().head(10)
top_countries.plot(kind='barh', title='Top 10 Countries with Most Content')
```

```
<Axes: title={'center': 'Top 10 Countries with Most Content'}, ylabel='country'>
```



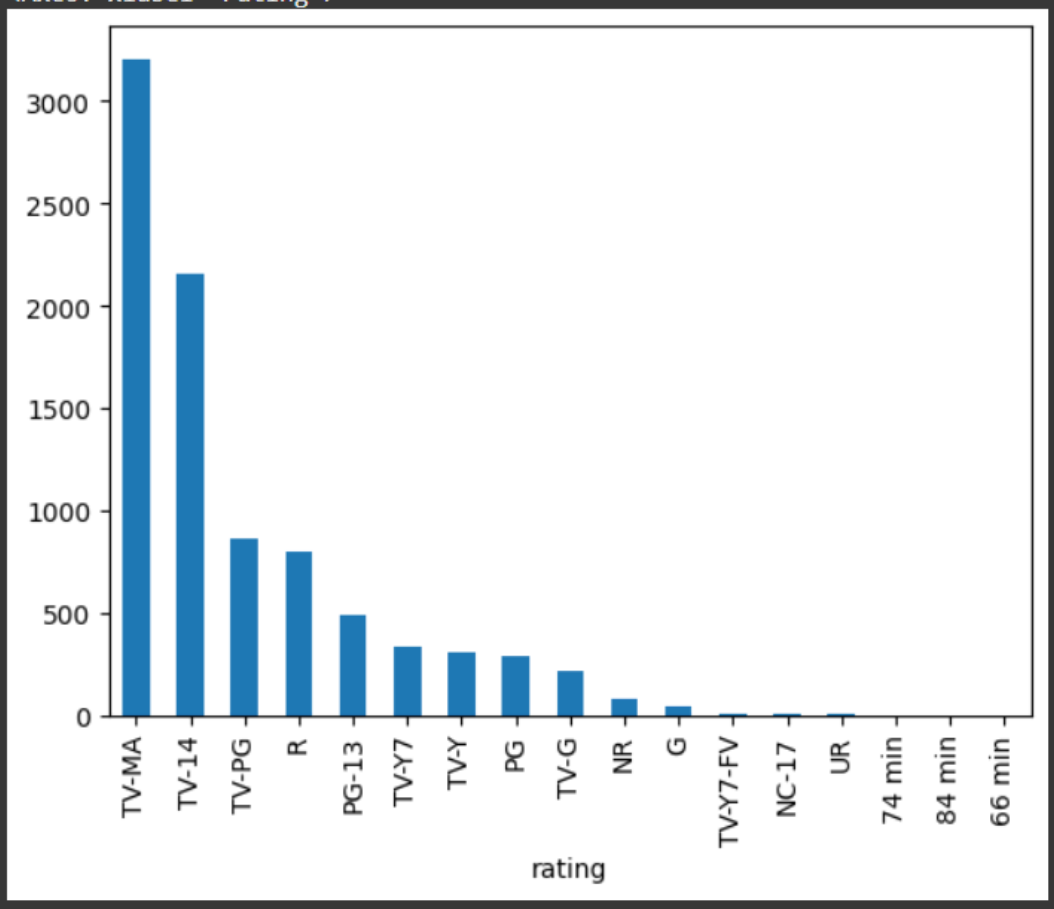
Insight: The United States dominates content production, followed by India, the UK, and Canada. This reflects both Netflix's home market and its expansion into Bollywood and international markets.

4.5. Ratings Distribution

Different types of content cater to different age groups. Exploring the rating column shows how Netflix classifies its content.

```
df['rating'].value_counts().plot(kind='bar')
```

```
<Axes: xlabel='rating'>
```



Insight: TV-MA and TV-14 ratings are the most common, indicating a strong focus on mature audiences.

4.6. Most Frequent Directors and Actors

By analyzing the director and cast columns (after cleaning), we can identify who collaborates most often with Netflix.

```
df['director'].value_counts().head(10)
df['cast'].str.split(', ').explode().value_counts().head(10)
```

cast	count
Not Specified	825
Anupam Kher	43
Shah Rukh Khan	35
Julie Tejwani	33
Naseeruddin Shah	32
Takahiro Sakurai	32
Rupa Bhimani	31
Om Puri	30
Akshay Kumar	30
Yuki Kaji	29

dtype: int64

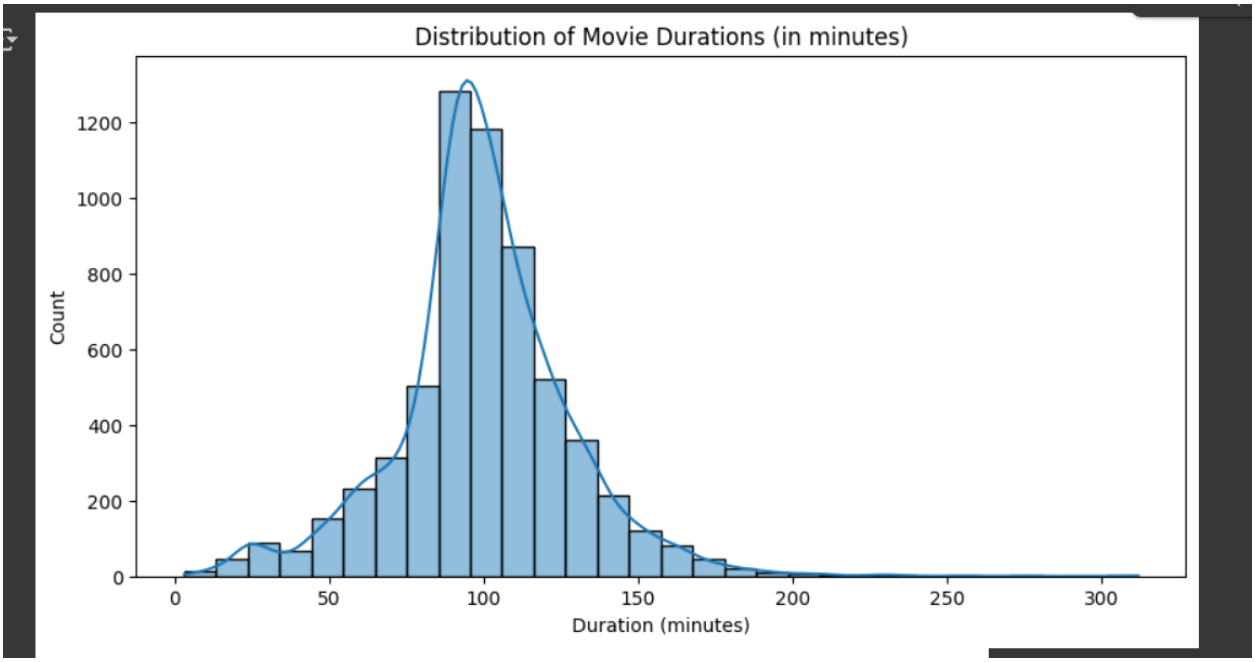
Insight: Specific directors and actors appear frequently, which may indicate long-term collaborations or popular names in Netflix Originals.

4.7 Duration analysis

Understand the distribution of content duration—split into Movies (minutes) and TV Show

Insights:

- Most movies are between 80–120 minutes.



- Many TV shows are limited series (1–2 seasons), with fewer long-running series.

5.Recommendations:

Based on the exploratory data analysis performed on the Netflix Shows dataset, here are some key recommendations:

1. Content Strategy

- **Focus on TV Shows:** Since TV Shows and Movies are almost equally present, but TV Shows tend to have higher viewer engagement due to episodic structure, Netflix can prioritize high-quality series production.
- **Optimize Duration:** Most movies have a duration between **80–120 minutes**, which appears to be the sweet spot. New content should generally aim for this range to match viewer preferences.

1. Content Strategy

- **Focus on TV Shows:** Since TV Shows and Movies are almost equally present, but TV Shows tend to have higher viewer engagement due to episodic structure, Netflix can prioritize high-quality series production.

- **Optimize Duration:** Most movies have a duration between **80–120 minutes**, which appears to be the sweet spot. New content should generally aim for this range to match viewer preferences.

3. Language Diversification

- While English dominates, there's an increasing demand for content in **non-English languages** (like Spanish, Hindi, Korean).
- Netflix can consider **localizing** more content and investing in **regional originals**.

4. Seasonal Release Planning

- Peak release months observed in the dataset are **July and December**.
- Recommendation: Align marketing efforts and big releases with these periods to maximize visibility and subscriber retention.

5. Genre Targeting

- The most common genres include **Dramas, Comedies, Documentaries, and Action**.
- Netflix can segment user bases by genre preferences and recommend personalized content, especially focusing on trending genres.

6. Tagline Improvements

- The dataset shows inconsistent or missing values in the **description/tags** fields.
- Recommendation: Standardize and enrich taglines and metadata for all content to improve **searchability** and **recommendation engine accuracy**.

7. Content Age Classification

- A wide range of maturity ratings were observed.
- Recommendation: Enhance **content filtering options** and **parental controls** for a better family-friendly experience.

6.Conclusion:

This exploratory data analysis of the Netflix dataset provided valuable insights into the content library of one of the world's leading streaming platforms. Through detailed examination and visualization of attributes like content type, genres, countries, release years, durations, and ratings, we uncovered key trends such as:

- A significant focus on TV Shows in recent years.
- The dominance of the United States in content production.
- Most content being targeted toward adult and teen audiences.
- Movies generally having a duration around 90 minutes, while TV Shows are usually described by seasons.

Our findings also highlighted missing or inconsistent values in fields like country, cast, and duration, which were addressed during preprocessing. While this analysis helps in understanding Netflix's content strategy and user offerings, it also opens avenues for more complex modeling — such as recommendation systems, sentiment analysis, or predictive modeling of viewer trends.