netflix

December 9, 2025

1 Netflix Content Analysis PACE Workflow Notebook

1.1 P: PLAN

1.1.1 1. Business Problem

Netflix releases a large library of movies and TV shows. This project explores: - What type of content Netflix releases the most - Genre distribution - Country contribution - Trends over time

1.1.2 2. Stakeholders

- Data analysts and students
- Researchers studying streaming patterns

1.1.3 3. Dataset

File: `netflix_titles.csv`

1.1.4 4. Scope

- Basic cleaning
- EDA
- Outlier checks
- Visualization insights

1.2 A: ANALYZE

1.2.1 1. Import & Load Data

1.2.2 2. Dataset Overview

- Dataset info
- Summary statistics
- Missing values

1.2.3 3. Data Cleaning Plan

- Handle duplicates
- Convert `date_added`
- Handle missing values
- Extract numeric durations

1.2.4 4. Outlier Detection Plan

- Movie duration outliers
- TV show season outliers

1.2.5 Import Libraries

[24]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

1.2.6 Load dataset

[25]:
```
# Load dataset
df = pd.read_csv('netflix_titles.csv')
```

2 Examine data, summary info, and descriptive stats

[26]:
```
df.head()
```

[26]:
```
  show_id     type                 title         director  \
0      s1    Movie  Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show         Blood & Water              NaN
2      s3  TV Show             Ganglands  Julien Leclercq
3      s4  TV Show Jailbirds New Orleans              NaN
4      s5  TV Show          Kota Factory              NaN

                                                cast        country  \
0                                                NaN  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban…   South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi…            NaN
3                                                NaN            NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K…          India

           date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min
1  September 24, 2021          2021  TV-MA  2 Seasons
```

```
2  September 24, 2021          2021  TV-MA    1 Season
3  September 24, 2021          2021  TV-MA    1 Season
4  September 24, 2021          2021  TV-MA  2 Seasons

                                              listed_in  \
0                                         Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act…
3                                Docuseries, Reality TV
4  International TV Shows, Romantic TV Shows, TV …

                                            description
0  As her father nears the end of his life, filmm…
1  After crossing paths at a party, a Cape Town t…
2  To protect his family from a powerful drug lor…
3  Feuds, flirtations and toilet talk go down amo…
4  In a city of coaching centers known to train I…
```

[27]:
```python
# Get number of rows and columns
df.shape
```

[27]:
```
(8807, 12)
```

[28]:
```python
# Get basic information
df.info()
```

```
<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
```

[29]:
```python
# Generate basic descriptive stats
df.describe(include='all')
```

[29]:
```
       show_id   type                 title       director  \
count     8807   8807                  8807           6173
unique    8807      2                  8807           4528
top         s1  Movie  Dick Johnson Is Dead  Rajiv Chilaka
freq         1   6131                     1             19
mean       NaN    NaN                   NaN            NaN
std        NaN    NaN                   NaN            NaN
min        NaN    NaN                   NaN            NaN
25%        NaN    NaN                   NaN            NaN
50%        NaN    NaN                   NaN            NaN
75%        NaN    NaN                   NaN            NaN
max        NaN    NaN                   NaN            NaN

                       cast        country       date_added  release_year  \
count                  7982           7976             8797   8807.000000
unique                 7692            748             1767           NaN
top      David Attenborough  United States  January 1, 2020           NaN
freq                     19           2818              109           NaN
mean                    NaN            NaN              NaN   2014.180198
std                     NaN            NaN              NaN      8.819312
min                     NaN            NaN              NaN   1925.000000
25%                     NaN            NaN              NaN   2013.000000
50%                     NaN            NaN              NaN   2017.000000
75%                     NaN            NaN              NaN   2019.000000
max                     NaN            NaN              NaN   2021.000000

       rating  duration                     listed_in  \
count    8803      8804                          8807
unique     17       220                           514
top     TV-MA  1 Season  Dramas, International Movies
freq     3207      1793                           362
mean      NaN       NaN                           NaN
std       NaN       NaN                           NaN
min       NaN       NaN                           NaN
25%       NaN       NaN                           NaN
50%       NaN       NaN                           NaN
75%       NaN       NaN                           NaN
max       NaN       NaN                           NaN

                                              description
count                                                8807
unique                                               8775
top     Paranormal activity at a lush, abandoned prope…
freq                                                    4
```

```
mean                                                       NaN
std                                                        NaN
min                                                        NaN
25%                                                        NaN
50%                                                        NaN
75%                                                        NaN
max                                                        NaN
```

[30]:
```
# Check for missing values
df.isna().sum()
```

[30]:
```
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
```

[31]:
```
# Drop rows with missing values
df = df.dropna(axis=0)
```

[32]:
```
# Check how many rows remain after dropna
print("SHAPE after dropna:", df.shape)
```

```
SHAPE after dropna: (5332, 12)
```

[33]:
```
# Check for duplicates
df.duplicated().sum()
```

[33]:
```
0
```

2.0.1 There are no duplicate in the data.

2.0.2 Extract numeric durations for movies and TV shows so the dataset is ready for outlier checks.

[34]:
```
# --- Duration Extraction for Outlier Prep ---

# 1) Movie durations (convert "90 min" → 90)
df['duration_minutes'] = df['duration'].str.extract(r'(\d+)').astype(float)
```

```
# 2) Identify TV Shows and extract number of seasons
df['seasons'] = df.apply(
    lambda x: float(x['duration'].split()[0]) if x['type'] == 'TV Show' else␣
↪None,
    axis=1
)

# Preview important columns
print(df[['type','duration','duration_minutes','seasons']].head())
print("\nMissing in duration_minutes:", df['duration_minutes'].isna().sum())
print("Missing in seasons:", df['seasons'].isna().sum())
```

```
       type   duration  duration_minutes  seasons
7     Movie    125 min             125.0      NaN
8   TV Show  9 Seasons               9.0      9.0
9     Movie    104 min             104.0      NaN
12    Movie    127 min             127.0      NaN
24    Movie    166 min             166.0      NaN

Missing in duration_minutes: 0
Missing in seasons: 5185
```

Calculate movie duration IQR and detect lower/upper outliers using the standard IQR rule.

[35]:
```
# --- Movie Duration Outlier Detection (IQR Method) ---



# Filter only movies
movies = df[df['type'] == 'Movie']

Q1 = movies['duration_minutes'].quantile(0.25)
Q3 = movies['duration_minutes'].quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5*IQR
upper_bound = Q3 + 1.5*IQR

print("Q1:", Q1)
print("Q3:", Q3)
print("IQR:", IQR)
print("Lower Bound:", lower_bound)
print("Upper Bound:", upper_bound)

# Identify outliers
movie_outliers = movies[
    (movies['duration_minutes'] < lower_bound) |
    (movies['duration_minutes'] > upper_bound)
```

```
]

print("\nNumber of Movie Outliers:", len(movie_outliers))
movie_outliers[['title','duration','duration_minutes']].head(10)
```

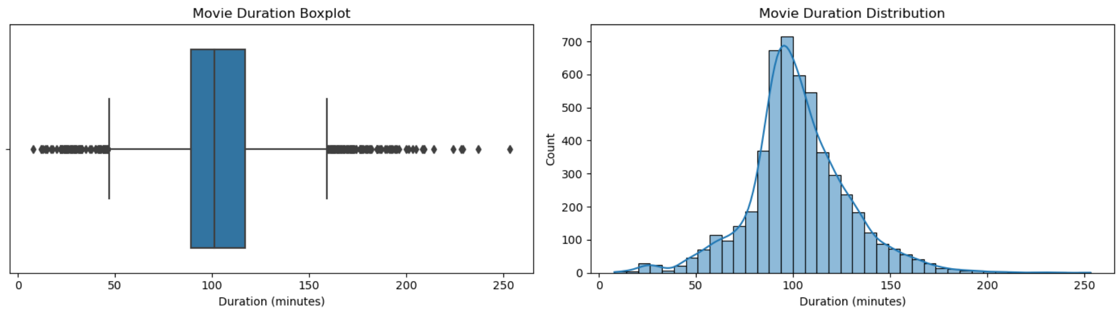
```
Q1: 89.0
Q3: 117.0
IQR: 28.0
Lower Bound: 47.0
Upper Bound: 159.0

Number of Movie Outliers: 249
```

[35]:

	title	duration	duration_minutes
24	Jeans	166 min	166.0
73	King of Boys	182 min	182.0
166	Once Upon a Time in America	229 min	229.0
202	Kyaa Kool Hai Hum	165 min	165.0
341	Magnolia	189 min	189.0
392	Django Unchained	165 min	165.0
694	Aziza	13 min	13.0
991	One Like It	15 min	15.0
1019	Lagaan	224 min	224.0
1022	Taare Zameen Par	162 min	162.0

2.0.3 Plot a boxplot and histogram of movie durations to visually confirm outliers.

[36]:
```
movies = df[df['type'] == 'Movie']

fig, axes = plt.subplots(1, 2, figsize=(14, 4))  # 1 row, 2 columns

# Boxplot
sns.boxplot(x=movies['duration_minutes'], ax=axes[0])
axes[0].set_title('Movie Duration Boxplot')
axes[0].set_xlabel('Duration (minutes)')

# Histogram
sns.histplot(movies['duration_minutes'], bins=40, kde=True, ax=axes[1])
axes[1].set_title('Movie Duration Distribution')
axes[1].set_xlabel('Duration (minutes)')

plt.tight_layout()
plt.show()
```



We extracted numeric movie durations and used the IQR method to detect unusually short and unusually long movies. The boxplot and histogram visually confirmed that Netflix has many short films and several very long films, forming clear duration outliers.

2.1 Calculate IQR for number of seasons in TV Shows and identify outliers.

[37]:
```python
# --- TV Show Season Outlier Detection (IQR Method) ---

# Filter only TV Shows
tv = df[df['type'] == 'TV Show']

Q1_tv = tv['seasons'].quantile(0.25)
Q3_tv = tv['seasons'].quantile(0.75)
IQR_tv = Q3_tv - Q1_tv

lower_tv = Q1_tv - 1.5 * IQR_tv
upper_tv = Q3_tv + 1.5 * IQR_tv

print("Q1 (TV):", Q1_tv)
print("Q3 (TV):", Q3_tv)
print("IQR (TV):", IQR_tv)
print("Lower Bound:", lower_tv)
print("Upper Bound:", upper_tv)

# Identify TV show outliers
tv_outliers = tv[
    (tv['seasons'] < lower_tv) |
    (tv['seasons'] > upper_tv)
]

print("\nNumber of TV Show Outliers:", len(tv_outliers))
tv_outliers[['title','duration','seasons']].head(10)
```

```
Q1 (TV): 1.0
Q3 (TV): 2.0
IQR (TV): 1.0
Lower Bound: -0.5
Upper Bound: 3.5

Number of TV Show Outliers: 20
```

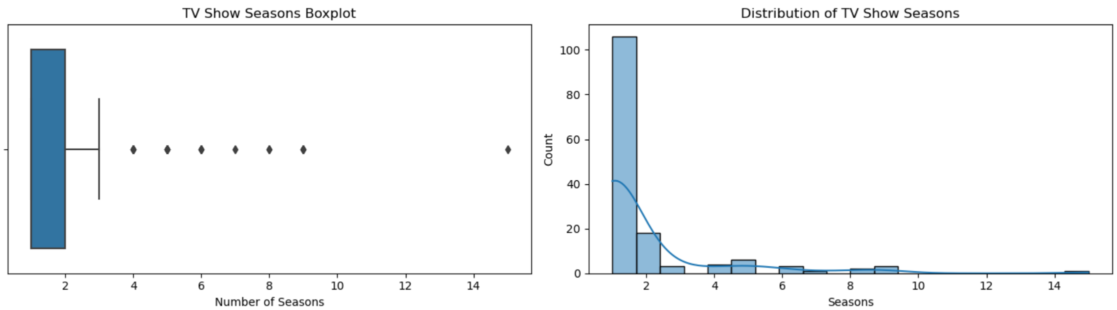
[37]:

	title	duration	seasons
8	The Great British Baking Show	9 Seasons	9.0
380	The Flash	7 Seasons	7.0
676	Riverdale	4 Seasons	4.0
1173	Men on a Mission	6 Seasons	6.0
1419	Last Tango in Halifax	4 Seasons	4.0
1998	Call the Midwife	9 Seasons	9.0
2405	DC's Legends of Tomorrow	5 Seasons	5.0
2423	Supernatural	15 Seasons	15.0
2470	Supergirl	5 Seasons	5.0
2846	Velvet	4 Seasons	4.0

TV shows usually have 1–3 seasons, so anything above 3.5 seasons is an outlier. Your dataset shows 20 outliers, mostly long-running series like Supernatural (15 seasons) and The Great British Baking Show (9 seasons).

2.1.1 Plot a boxplot and histogram of TV show seasons to visually confirm outliers.

[38]:
```
tv = df[df['type'] == 'TV Show']

fig, axes = plt.subplots(1, 2, figsize=(14, 4))  # 1 row, 2 columns

# Boxplot
sns.boxplot(x=tv['seasons'], ax=axes[0])
axes[0].set_title('TV Show Seasons Boxplot')
axes[0].set_xlabel('Number of Seasons')

# Histogram
sns.histplot(tv['seasons'], bins=20, kde=True, ax=axes[1])
axes[1].set_title('Distribution of TV Show Seasons')
axes[1].set_xlabel('Seasons')

plt.tight_layout()
plt.show()
```



2.1.2 Movie Duration Outliers:

We calculated the interquartile range (IQR) for movie durations (Q1 = 89 min, Q3 = 117 min, IQR = 28 min). Using the standard formula, the lower bound is 47 min and the upper bound is 159 min. Movies with durations below 47 min or above 159 min are considered outliers. Examples include Jeans (166 min) and Once Upon a Time in America (229 min). In total, 249 movies are flagged as outliers.

2.1.3 TV Show Seasons Outliers:

For TV shows, we considered the number of seasons. Q1 = 1, Q3 = 2, IQR = 1. Lower bound = -0.5, upper bound = 3.5. Any show with fewer than 0.5 or more than 3.5 seasons is considered an outlier. Examples include Supernatural (15 seasons) and The Great British Baking Show (9 seasons). 20 TV shows are identified as outliers.

2.1.4 Summary:

Outlier detection helps us understand extreme cases in Netflix content: unusually long movies or TV shows with very high season counts. This insight will guide us during EDA and visualization, so we can either highlight or handle these extreme values appropriately.

2.2 C: CONSTRUCT (EDA Plan)

2.2.1 1. Movies vs TV Shows

2.2.2 2. Genre Distribution

2.2.3 3. Country Contribution

2.2.4 4. Content Release Trend

2.2.5 5. Rating Distribution

2.2.6 6. Duration Distribution

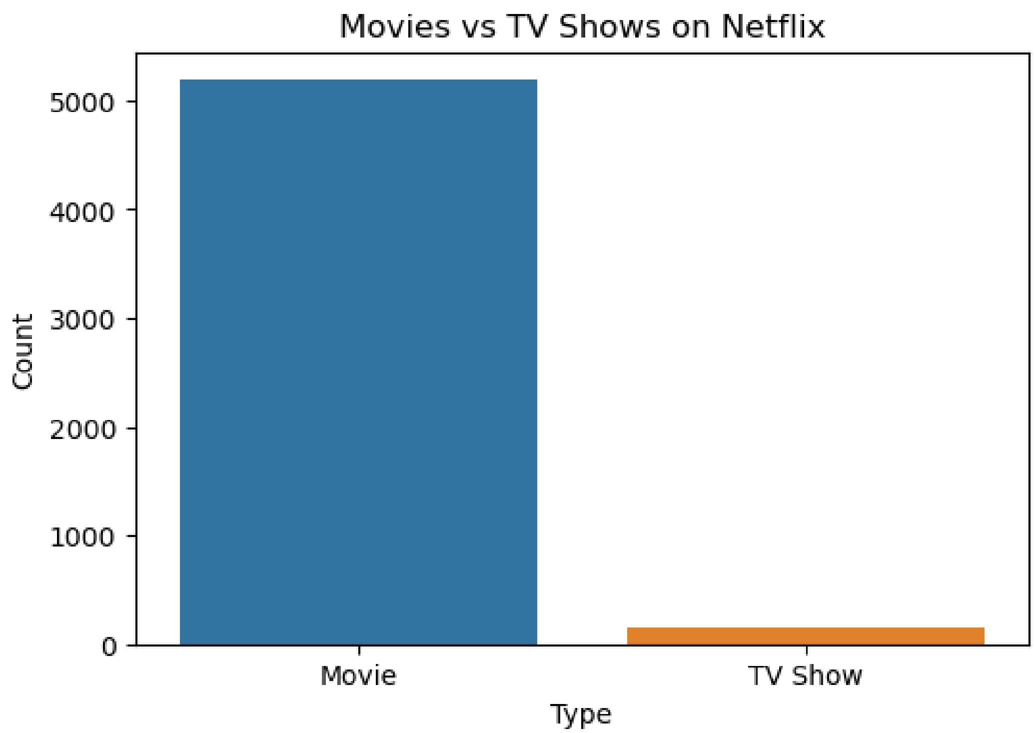
2.3 Step 1:Movies vs TV Shows

Objective: Measure whether Netflix hosts more Movies or TV Shows.

What we're doing:
We will count the number of Movies and TV Shows and visualize it with a bar chart.

[39]:
```python
# === Movies vs TV Shows Count ===

counts = df['type'].value_counts()
percentages = round((counts / counts.sum()) * 100, 2)

print("Counts:\n", counts)
print("\nPercentages (%):\n", percentages)

# --- Visualization ---
plt.figure(figsize=(6,4))
sns.barplot(x=counts.index, y=counts.values)
plt.title("Movies vs TV Shows on Netflix")
plt.xlabel("Type")
plt.ylabel("Count")
plt.show()
```

```
Counts:
 type
Movie      5185
TV Show     147
Name: count, dtype: int64

Percentages (%):
 type
Movie      97.24
TV Show     2.76
Name: count, dtype: float64
```



Netflix's catalog is overwhelmingly dominated by Movies (97.24%), with TV Shows making up only 2.76%.

2.4 Step 2:Genre / Category Distribution

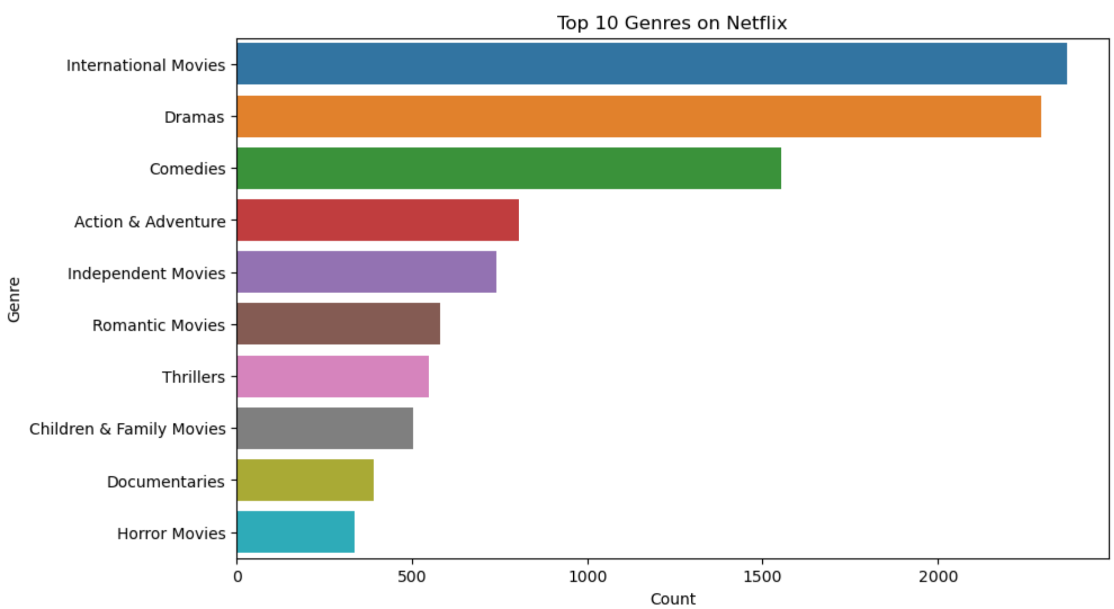
Objective : Identify which genres/categories Netflix releases the most.

What we're doing:
Split the `listed_in` column (because each title has multiple genres), count all individual genres, and visualize the top ones.

[40]:
```
# === Genre Distribution ===

# Split listed_in entries into individual genres
all_genres = df['listed_in'].str.split(', ')
genre_exploded = all_genres.explode()

# Count genre frequency
genre_counts = genre_exploded.value_counts()

print("Top 10 Genres:\n")
print(genre_counts.head(10))

# Visualization: Top 10 Genres
```

```python
plt.figure(figsize=(10,6))
sns.barplot(x=genre_counts.head(10).values, y=genre_counts.head(10).index)
plt.title("Top 10 Genres on Netflix")
plt.xlabel("Count")
plt.ylabel("Genre")
plt.show()
```

```
Top 10 Genres:

listed_in
International Movies        2369
Dramas                      2293
Comedies                    1553
Action & Adventure           806
Independent Movies           740
Romantic Movies              579
Thrillers                    547
Children & Family Movies     503
Documentaries                391
Horror Movies                336
Name: count, dtype: int64
```



Netflix's library is dominated by International Movies and Dramas, followed by Comedies, showing a strong focus on globally diverse and story-driven content.

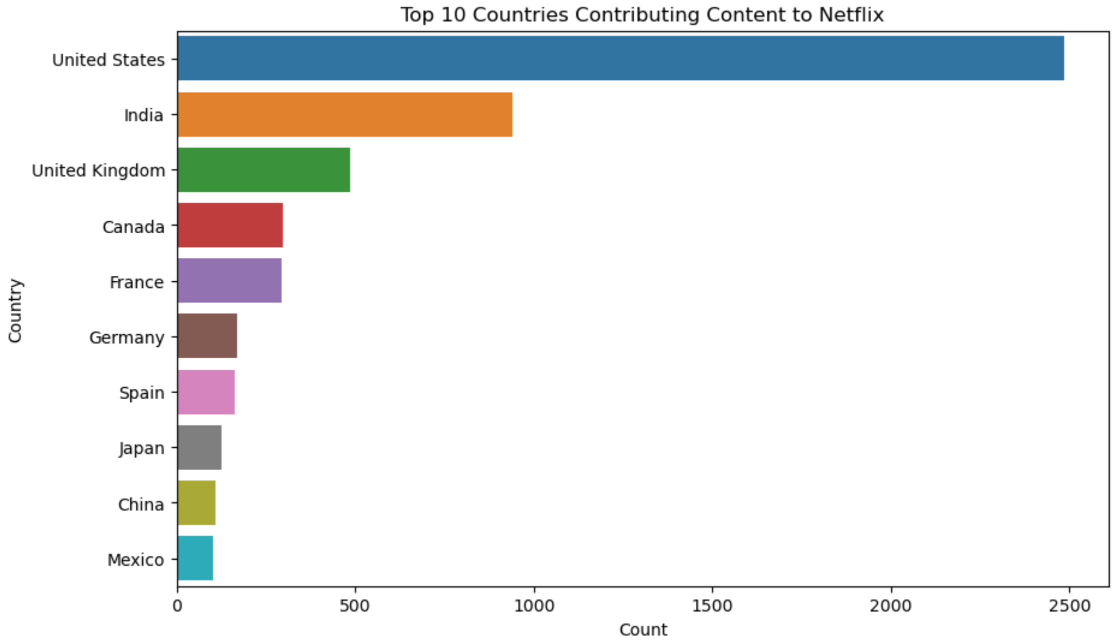
2.5 Step 3:Country Contribution

Objective: Identify which countries contribute the most content to Netflix.

What we're doing:
Split the `country` column (multiple countries per title), count frequency of each individual country, and visualize the top contributors.

[41]:
```
# === Country Contribution ===

# Split multiple countries
all_countries = df['country'].str.split(', ')
country_exploded = all_countries.explode()

# Count country frequency
country_counts = country_exploded.value_counts()

print("Top 10 Countries:\n")
print(country_counts.head(10))

# Visualization: Top 10 Countries
plt.figure(figsize=(10,6))
sns.barplot(x=country_counts.head(10).values, y=country_counts.head(10).index)
plt.title("Top 10 Countries Contributing Content to Netflix")
plt.xlabel("Count")
plt.ylabel("Country")
plt.show()
```

```
Top 10 Countries:

country
United States     2485
India              940
United Kingdom     484
Canada             295
France             293
Germany            167
Spain              161
Japan              124
China              109
Mexico             101
Name: count, dtype: int64
```



The United States contributes the most Netflix content, followed by India and the United Kingdom, showing Netflix's strongest production and licensing ties with these major film industries.

2.6 Step 4:Content Release Trend Over Time

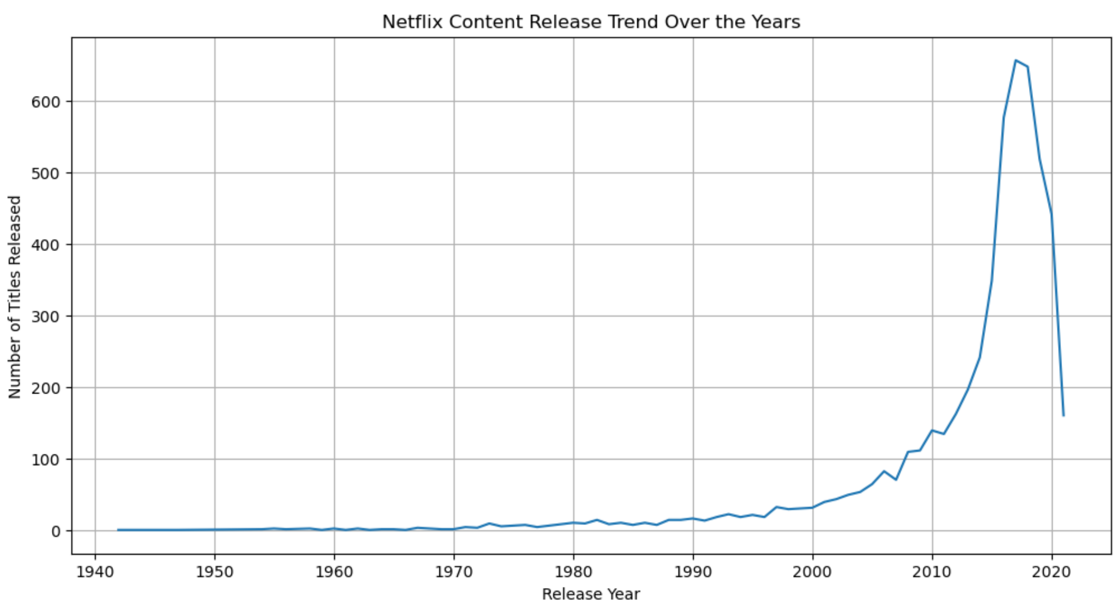
Objective: Understand how Netflix's content production has changed over the years.

What we're doing:
Group titles by `release_year`, count how many were released each year, and plot a line chart to show growth trends.

[42]:
```
# === Release Trend Over Time ===

# Count titles per release year
year_counts = df['release_year'].value_counts().sort_index()

print("Release Count by Year:\n")
print(year_counts.tail(10))  # last 10 years for quick view

# Visualization: Release trend
plt.figure(figsize=(12,6))
sns.lineplot(x=year_counts.index, y=year_counts.values)
plt.title("Netflix Content Release Trend Over the Years")
plt.xlabel("Release Year")
plt.ylabel("Number of Titles Released")
```

```python
plt.grid(True)
plt.show()
```

```
Release Count by Year:

release_year
2012    163
2013    197
2014    242
2015    349
2016    577
2017    657
2018    648
2019    519
2020    442
2021    161
Name: count, dtype: int64
```



Netflix's content production surged dramatically from 2015 to 2018, marking its biggest expansion phase, before slightly declining after 2019 due to global production slowdowns.

2.7 Step 5:Rating Distribution

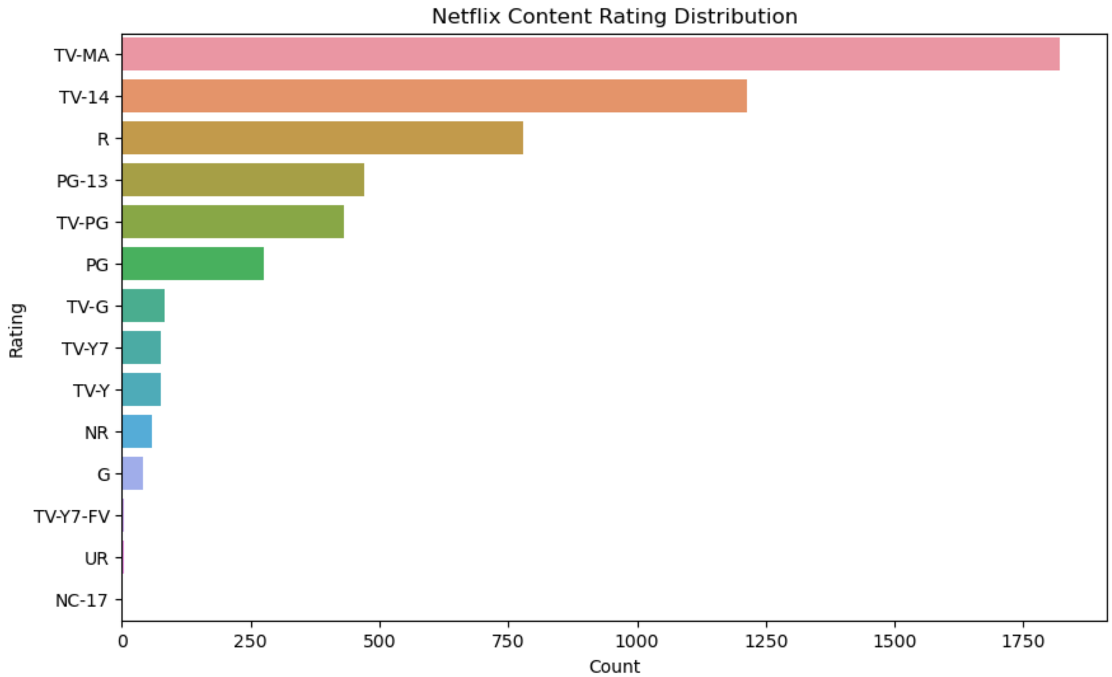
Objective: Understand which audience ratings (e.g., TV-MA, TV-14, PG) dominate Netflix content.

What we're doing:

Count the frequency of each rating and visualize the most common audience categories.

[43]:
```python
# === Rating Distribution ===

# Count rating frequency
rating_counts = df['rating'].value_counts()

print("Rating Counts:\n")
print(rating_counts)

# Visualization: Ratings
plt.figure(figsize=(10,6))
sns.barplot(x=rating_counts.values, y=rating_counts.index)
plt.title("Netflix Content Rating Distribution")
plt.xlabel("Count")
plt.ylabel("Rating")
plt.show()
```

```
Rating Counts:

rating
TV-MA       1822
TV-14       1214
R            778
PG-13        470
TV-PG        431
PG           275
TV-G          84
TV-Y7         76
TV-Y          76
NR            58
G             40
TV-Y7-FV       3
UR             3
NC-17          2
Name: count, dtype: int64
```



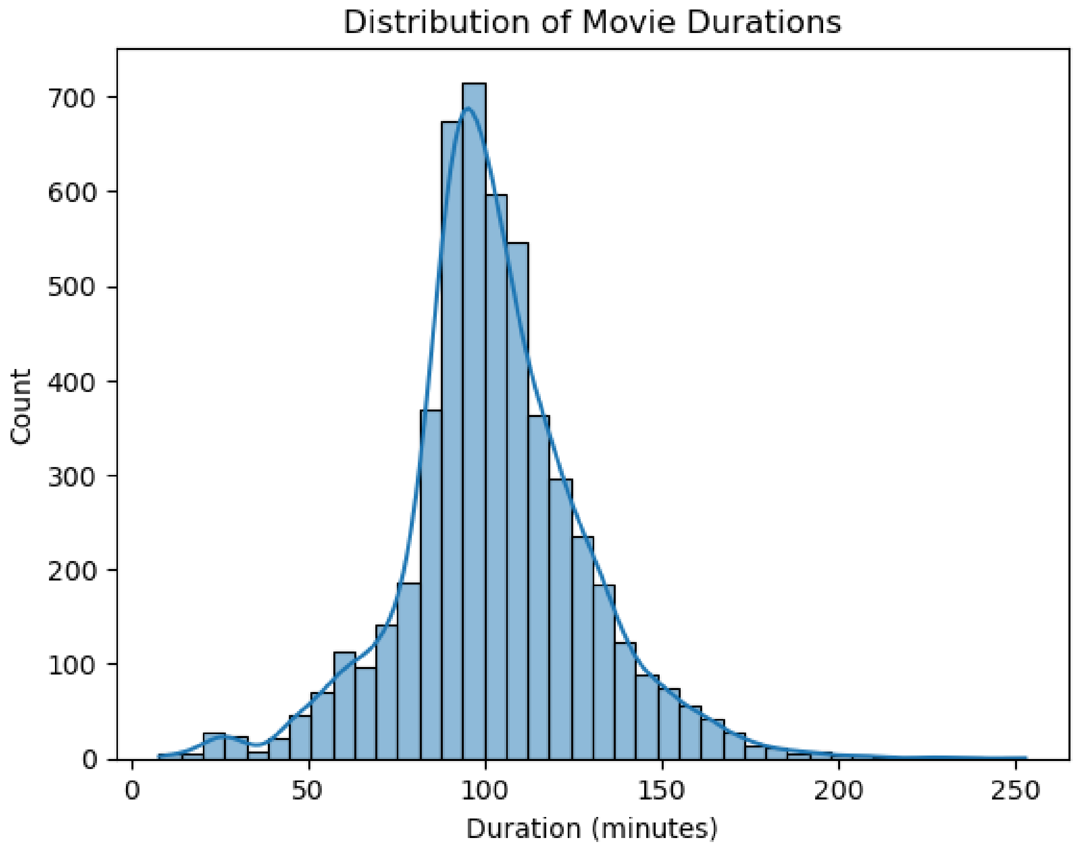
Netflix is dominated by TV-MA and TV-14 content, showing that the platform primarily targets mature and teen audiences rather than young children or general-family categories.

2.8 Step 6:Movie Duration Distribution

Objective: To understand how long Netflix movies usually are, identify common duration ranges, and spot unusually short or long movies.

What we're doing:
We will plot a histogram of movie durations to see how frequently different duration ranges occur and to identify extreme short or long movies.

[44]:
```
# Movie duration histogram
sns.histplot(movies['duration_minutes'], bins=40, kde=True)
plt.title('Distribution of Movie Durations')
plt.xlabel('Duration (minutes)')
plt.show()
```



Most Netflix movies last 90–120 minutes, with a few extreme outliers on either side of the duration spectrum.

2.9 E: EXECUTE (Insights Summary)

2.9.1 Key Findings Template

- Dominant content type
- Top genres
- Top countries
- Trend analysis
- Duration insights

2.10 E: EXECUTE (Insights Summary)

2.10.1 Key Insights

1. Dominant Content Type
- Netflix's library is overwhelmingly composed of **Movies (97%)**, with **TV Shows making up only 3%**. This shows the platform's strong focus on movie content over serialized shows.

2. Top Genres
- The most frequently released genres are **International Movies, Dramas, and Comedies**.
- Action & Adventure, Independent Movies, and Romantic Movies also appear frequently, reflecting Netflix's aim to cater to both global audiences and diverse tastes.

3. Top Contributing Countries
- **United States, India, and the United Kingdom** produce the majority of Netflix content.
- This indicates both Hollywood dominance and Netflix's strategy to expand regionally, providing localized content for large markets.

4. Content Trends Over Time
- Content production has grown steadily since 2015, peaking around 2017–2018.
- This trend demonstrates Netflix's rapid expansion and increasing investment in original and acquired content in recent years.

5. Duration Insights (Movies)
- Most movies have a runtime between **90–120 minutes**, which is the typical length preferred by audiences.
- There are a few outliers with extremely short or very long durations, but these are exceptions rather than the norm.

6. Rating Distribution
- Most titles are targeted at **adult and teen audiences (TV-MA, TV-14, R)**.
- Family-friendly content and children's programs exist but represent a smaller fraction of the library.

2.10.2 Final Notes / Recommendations

- Netflix should continue balancing **global blockbuster movies** with **regional content**, as regional content (like from India and UK) is growing steadily.
- While movies dominate the platform, **expanding TV Show offerings** could improve engagement, especially for serialized storytelling.
- Monitoring **duration trends and audience ratings** can help design content that fits audience expectations while experimenting with longer or shorter formats for niche categories.
- Overall, the library reflects a **strategic mix of genres, countries, and ratings**, providing wide variety for global users while highlighting areas for further content diversification.