

Netflix Content Analytics Project

```
# These are the basic libraries we need for analysis
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats

# Display settings (so that large tables show all columns)
pd.set_option("display.max_columns", None)
```

2. Load & Explore Data

```
# Load dataset
data = pd.read_csv("/content/netflix_titles.csv")

# Quick shape & preview
print("Rows, Columns:", data.shape)
```

Rows, Columns: (8807, 12)

data.head(5)

1 to 5 of 5 entries

Filter

?

index	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmmaker Kirsten Johnson stages his death in inventive and comical ways to help them both face the inevitable.
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Molaba, Dillon Windvogel, Natasha Thahane, Arno Greeff, Xolile Tshabalala, Getmore Sithole, Cindy Mahlangu, Ryle De Morny, Greteli Fincham, Sello Maake Ka-Ncube, Odwa Gwanya, Mekaila Mathys, Sandi Schultz, Duane Williams, Shamilla	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town teen sets out to prove whether a private-school swimming star is her sister who was abducted at birth.

Next steps: [Generate code with data](#) [View recommended plots](#) [New interactive sheet](#)


3. Data Cleaning

```
# Filling missing values
data['director'] = data['director'].fillna("Unknown Director")
data['cast'] = data['cast'].fillna("Not Available")
data['country'] = data['country'].fillna("Unknown Country")
data['rating'] = data['rating'].fillna("Not Rated")
data['duration'] = data['duration'].fillna("Unknown")
```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 14 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 8807 non-null object

```
4 cast 8807 non-null object
5 country 8807 non-null object
6 date_added 8709 non-null datetime64[ns]
7 release_year 8807 non-null int64
8 rating 8807 non-null object
9 duration 8807 non-null object
10 listed_in 8807 non-null object
11 description 8807 non-null object
12 duration_int 8804 non-null float64
13 duration_type 8807 non-null object
dtypes: datetime64[ns](1), float64(1), int64(1), object(11)
memory usage: 963.4+ KB
```


```
# Convert date_added → datetime
data['date_added'] = pd.to_datetime(data['date_added'], errors="coerce")
data
```



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Not Available	United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmmaker...
1	s2	TV Show	Blood & Water	Unknown Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thabane...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown Country	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lord...
3	s4	TV Show	Jailbirds New Orleans	Unknown Director	Not Available	Unknown Country	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down among...
4	s5	TV Show	Kota Factory	Unknown Director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city where coaching centers are known to train I...
...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, ...	United States	2019-11-20	2007	R	158 min	Cult Movies, Dramas, ...	A political cartoonist, a crim...

Next steps: [Generate code with data](#) [View recommended plots](#) [New interactive sheet](#)


```
# Drop duplicates
data = data.drop_duplicates()
data
```



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Not Available	United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmmaker Kirsten Johnson turns the camera on herself and her family to explore the meaning of life, death, and legacy.
1	s2	TV Show	Blood & Water	Unknown Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thabane...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town teen sets out to prove whether a local mercenary can deliver the goods.
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabil...	Unknown Country	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lord, a young man must become a gangster.
3	s4	TV Show	Jailbirds New Orleans	Unknown Director	Not Available	Unknown Country	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down among the inmates of the Louisiana State Prison.
4	s5	TV Show	Kota Factory	Unknown Director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city where coaching centers are known to train IIT aspirants, a group of friends start a coaching center for IIT aspirants.
...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Pattinson, ...	United States	2019-11-20	2007	R	158 min	Cult Movies, Dramas, ...	A political cartoonist, a criminal profiler, and a police detective hunt for a serial killer in San Francisco.

Next steps: [Generate code with data](#) [View recommended plots](#) [New interactive sheet](#)

```
# Strip spaces in text columns
for col in ['title','director','cast','country','rating','listed_in']:
    data[col] = data[col].astype(str).str.strip()
data
```




	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Not Available	United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmmaker Kirsten Johnson
1	s2	TV Show	Blood & Water	Unknown Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thabane	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town teen
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi	Unknown Country	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lord
3	s4	TV Show	Jailbirds New Orleans	Unknown Director	Not Available	Unknown Country	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down among the inmates
4	s5	TV Show	Kota Factory	Unknown Director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city where coaching centers are known to train IIT aspirants
...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal,	United States	2019-11-20	2007	R	158 min	Cult Movies, Dramas,	A political cartoonist, a criminal

Next steps: [Generate code with data](#) [View recommended plots](#) [New interactive sheet](#)

```
# Extract numbers and units from duration
data['duration_int'] = data['duration'].str.extract(r'(\d+)')
data['duration_int'] = pd.to_numeric(data['duration_int'], errors="coerce")
data['duration_type'] = data['duration'].str.extract(r'([a-zA-Z]+)')

data
```



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Not Available	United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmmaker Kirsten Johnson...
1	s2	TV Show	Blood & Water	Unknown Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thabane...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabil...	Unknown Country	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lord...
3	s4	TV Show	Jailbirds New Orleans	Unknown Director	Not Available	Unknown Country	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down among the inmates of the Louisiana State Prison...
4	s5	TV Show	Kota Factory	Unknown Director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city where coaching centers are known to train IIT aspirants...
...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Pattinson, ...	United States	2019-11-20	2007	R	158 min	Cult Movies, Dramas, ...	A political cartoonist, a criminal, and a police officer...

Next steps: [Generate code with data](#) [View recommended plots](#) [New interactive sheet](#)


```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 14 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               8807 non-null   object
4   cast                   8807 non-null   object
5   country                8807 non-null   object
6   date_added             8709 non-null   datetime64[ns]
7   release_year           8807 non-null   int64
8   rating                 8807 non-null   object
9   duration               8807 non-null   object
10  listed_in              8807 non-null   object
11  description             8807 non-null   object
12  duration_int           8804 non-null   float64
13  duration_type          8807 non-null   object
dtypes: datetime64[ns](1), float64(1), int64(1), object(11)
memory usage: 963.4+ KB
```

4. Exploratory Data Analysis (EDA)

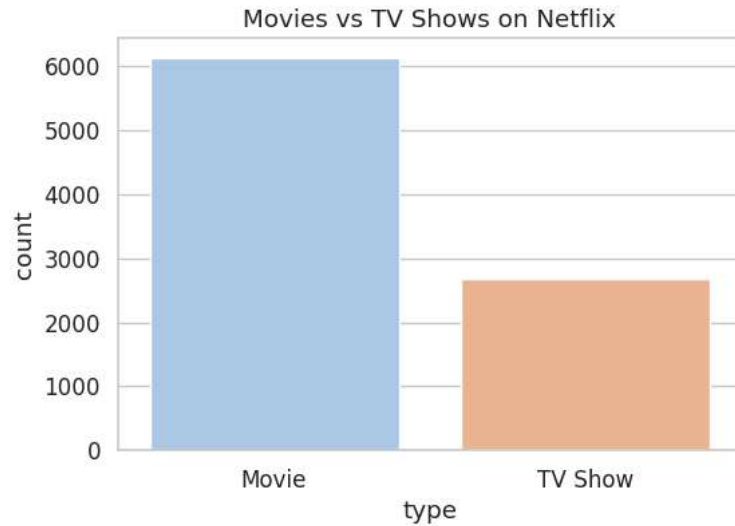
```
# Movies vs TV Shows count

plt.figure(figsize=(6,4))
sns.countplot(x='type', data=data, palette="pastel")
plt.title("Movies vs TV Shows on Netflix")
plt.show()
```

 /tmp/ipython-input-147980436.py:4: FutureWarning:

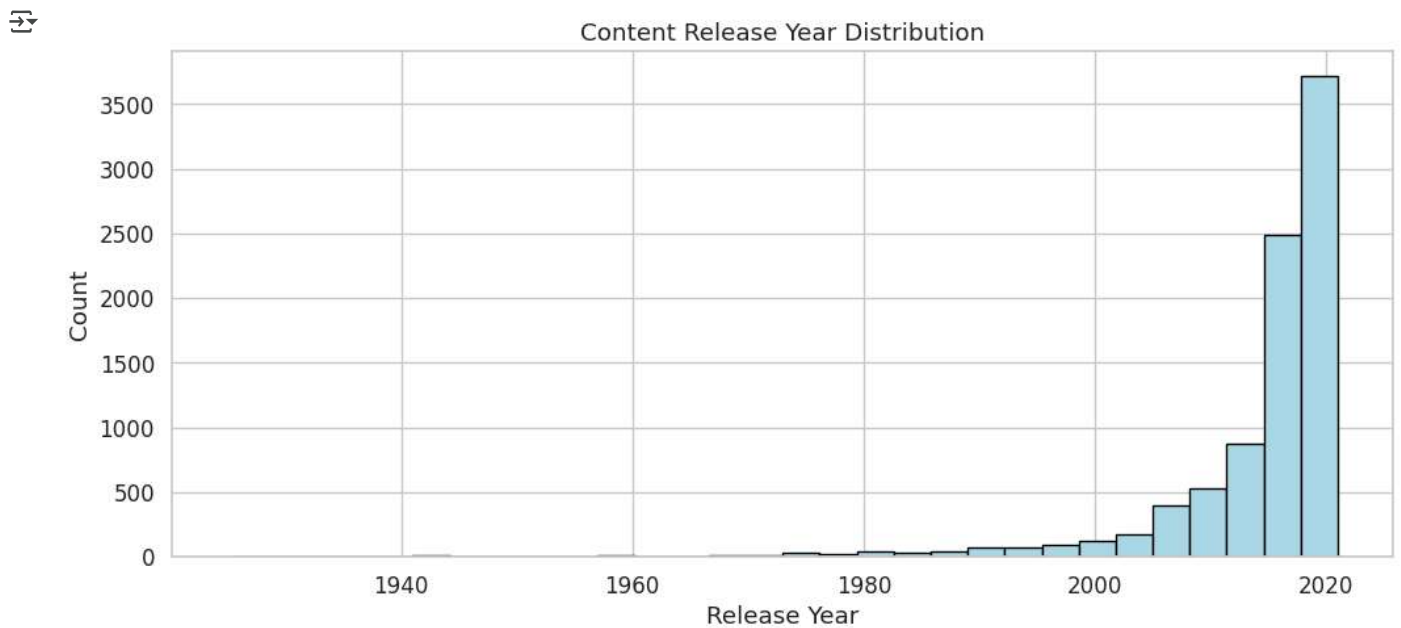
Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `l

```
sns.countplot(x='type', data=data, palette="pastel")
```




Release year histogram

```
plt.figure(figsize=(12,5))
data['release_year'].hist(bins=30, color='lightblue', edgecolor='black')
plt.xlabel("Release Year")
plt.ylabel("Count")
plt.title("Content Release Year Distribution")
plt.show()
```



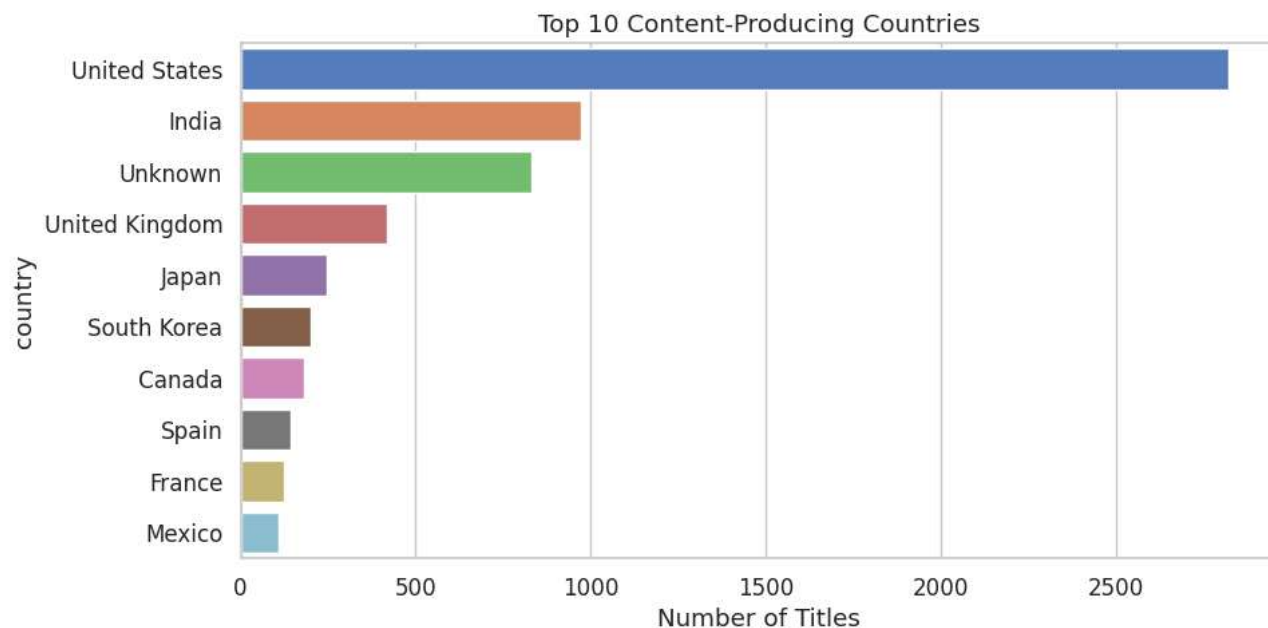
Top 10 countries

```
top_countries = data['country'].value_counts().head(10)
plt.figure(figsize=(10,5))
sns.barplot(x=top_countries.values, y=top_countries.index, palette="muted")
plt.title("Top 10 Content-Producing Countries")
plt.xlabel("Number of Titles")
plt.show()
```

 /tmp/ipython-input-1356345901.py:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `l

```
sns.barplot(x=top_countries.values, y=top_countries.index, palette="muted")
```




5. Statistical Analysis

We apply basic statistical concepts here.

```
# Descriptive stats for duration
print("Duration statistics:")
print(data['duration_int'].describe())

# Chi-Square test → relation between type and rating
crosstab = pd.crosstab(data['type'], data['rating'])
chi2, p, dof, expected = stats.chi2_contingency(crosstab)
print(f"\nChi-Square test: chi2={chi2:.2f}, p={p:.4f}")

# T-test → compare avg movie duration vs TV shows seasons
movies = data[data['type']=="Movie"]['duration_int'].dropna()
shows = data[data['type']=="TV Show"]['duration_int'].dropna()
t_stat, p_val = stats.ttest_ind(movies, shows, equal_var=False, nan_policy='omit')
print(f"\nT-test: t={t_stat:.2f}, p={p_val:.4f}")
```

 Duration statistics:

count	8804.000000
mean	69.846888
std	50.814828
min	1.000000
25%	2.000000
50%	88.000000
75%	106.000000
max	312.000000
Name: duration_int, dtype: float64	

Chi-Square test: chi2=1048.25, p=0.0000

T-test: t=269.69, p=0.0000

6. Feature Engineering

We create extra useful columns.

```
# Year & month when content was added
data['year_added'] = data['date_added'].dt.year
data['month_added'] = data['date_added'].dt.month

# Binary column → is it a movie?
data['is_movie'] = (data['type']=="Movie").astype(int)
```

```
data[['title', 'type', 'year_added', 'month_added', 'is_movie']].head()
```

1 to 5 of 5 entries Filter ?

index	title	type	year_added	month_added	is_movie
0	Dick Johnson Is Dead	Movie	2021.0	9.0	1
1	Blood & Water	TV Show	2021.0	9.0	0
2	Ganglands	TV Show	2021.0	9.0	0
3	Jailbirds New Orleans	TV Show	2021.0	9.0	0
4	Kota Factory	TV Show	2021.0	9.0	0

Show 25 per page



Like what you see? Visit the [data table notebook](#) to learn more about interactive tables.

✓ NETFLIX DATA ANALYTICS PROJECT

```
# 1. IMPORT LIBRARIES
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

sns.set(style="whitegrid", palette="muted", font_scale=1.1)
plt.rcParams['figure.figsize'] = (12,6)

# -----
# 2. EXTRACT (LOAD DATA)
# -----
# Replace with your CSV file path
data = pd.read_csv('/content/netflix_titles.csv')

print("Data loaded successfully!")
print(f"Shape: {data.shape}")
print(data.head())

# -----
# 3. TRANSFORM (CLEANING & FEATURE ENGINEERING)
# -----

# 3a. Handle missing values
numeric_cols = data.select_dtypes(include=np.number).columns
data[numeric_cols] = data[numeric_cols].fillna(data[numeric_cols].median())

categorical_cols = data.select_dtypes(include='object').columns
data[categorical_cols] = data[categorical_cols].fillna('Unknown')

# 3b. Convert 'duration' to numeric
if 'duration' in data.columns:
    def extract_duration(x):
        try:
            return int(''.join(filter(str.isdigit, str(x))))
        except:
            return 0
    data['duration_int'] = data['duration'].apply(extract_duration)

# 3c. Create 'is_movie' column
if 'type' in data.columns:
    data['is_movie'] = data['type'].apply(lambda x: 1 if str(x).lower() == 'movie' else 0)

# 3d. Process genres
if 'listed_in' in data.columns:
    data['genre_list'] = data['listed_in'].str.split(', ')
    genre_df = data.explode('genre_list')

# 3e. Convert 'date_added' to datetime
if 'date_added' in data.columns:
    data['date_added'] = pd.to_datetime(data['date_added'], errors='coerce')
    data['month_added'] = data['date_added'].dt.month
    data['year_added'] = data['date_added'].dt.year
```

Data loaded successfully!
 Shape: (8807, 12)

	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...

```

```

# -----
# 4. EXPLORATORY & ADVANCED ANALYTICS
# -----

# 4a. Movies vs TV Shows
print("\nMovies vs TV Shows Count:")
print(data['is_movie'].value_counts())
sns.countplot(x='is_movie', data=data)
plt.title('Count of Movies vs TV Shows')
plt.xlabel('Is Movie (1=Movie, 0=TV Show)')
plt.show()

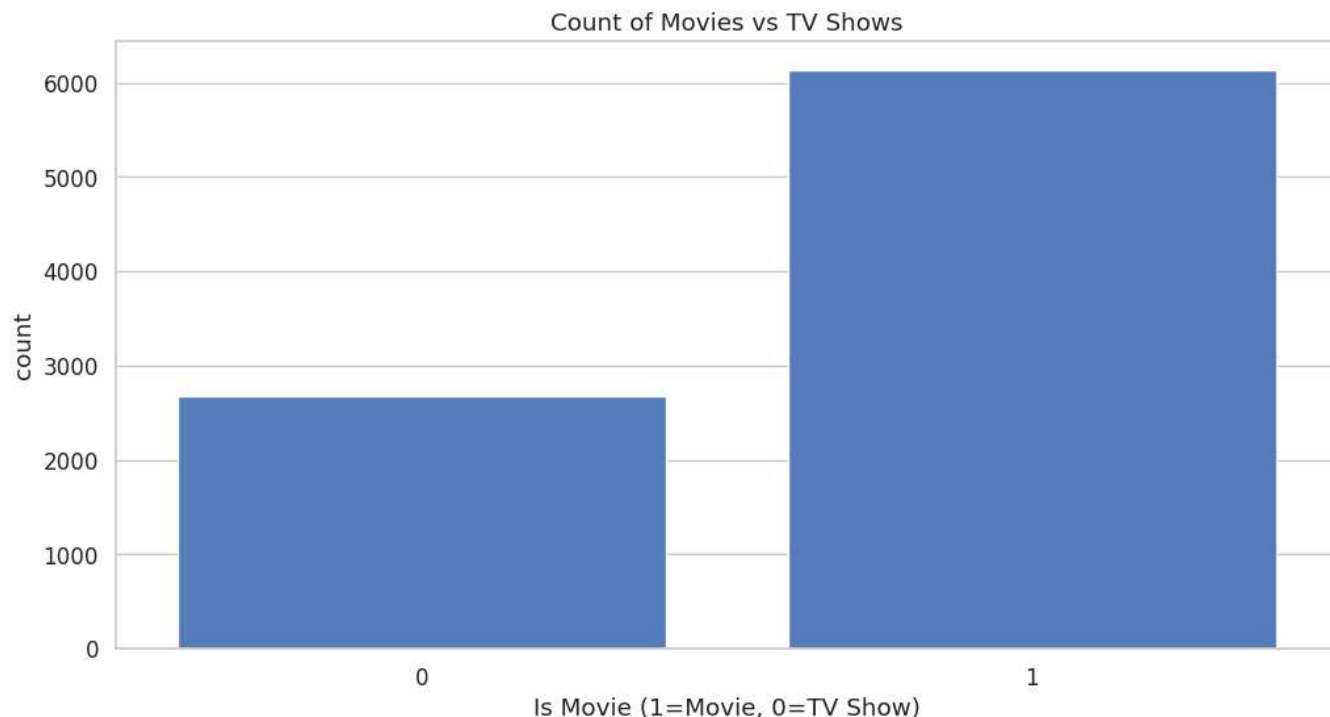
```



```

Movies vs TV Shows Count:
is_movie
1      6131
0      2676
Name: count, dtype: int64

```

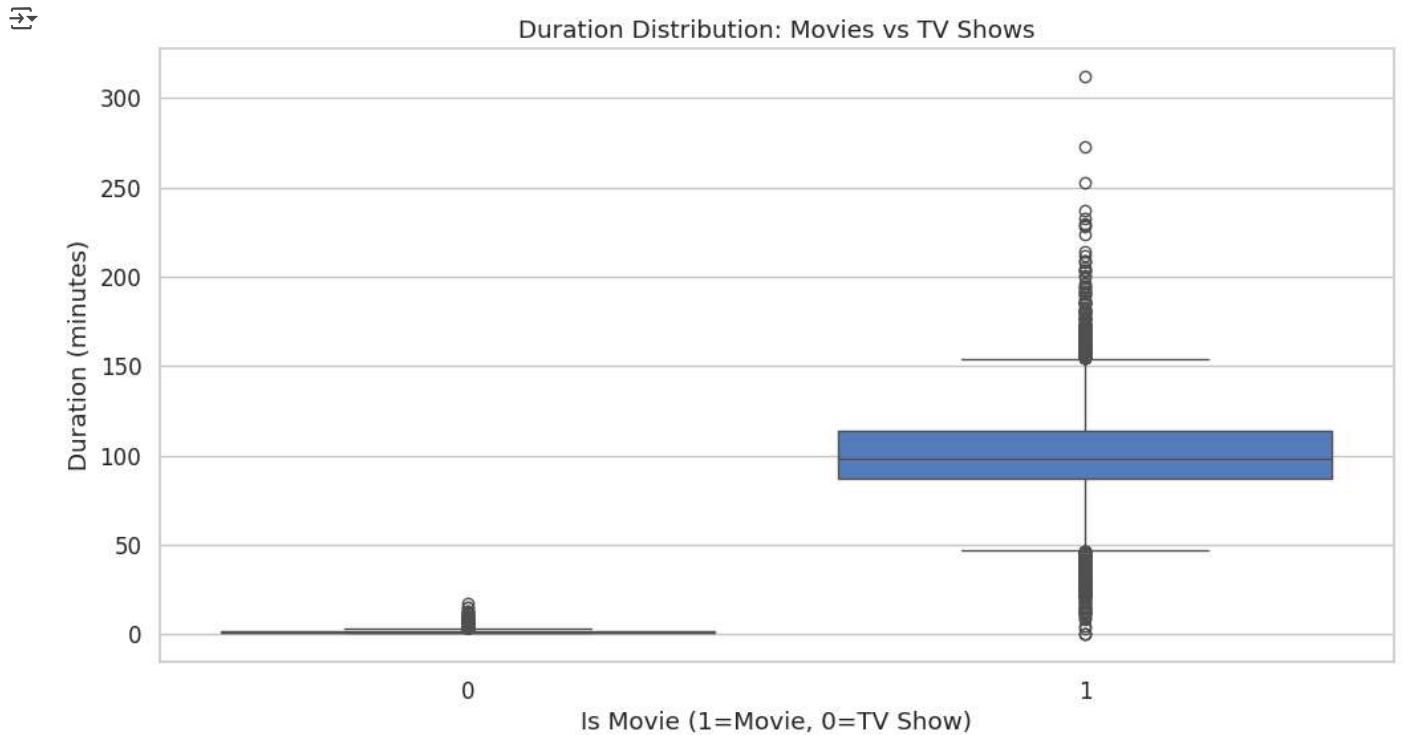


```

# 4b. Duration Analysis
sns.boxplot(x='is_movie', y='duration_int', data=data)
plt.title('Duration Distribution: Movies vs TV Shows')
plt.xlabel('Is Movie (1=Movie, 0=TV Show)')
plt.ylabel('Duration (minutes)')

```

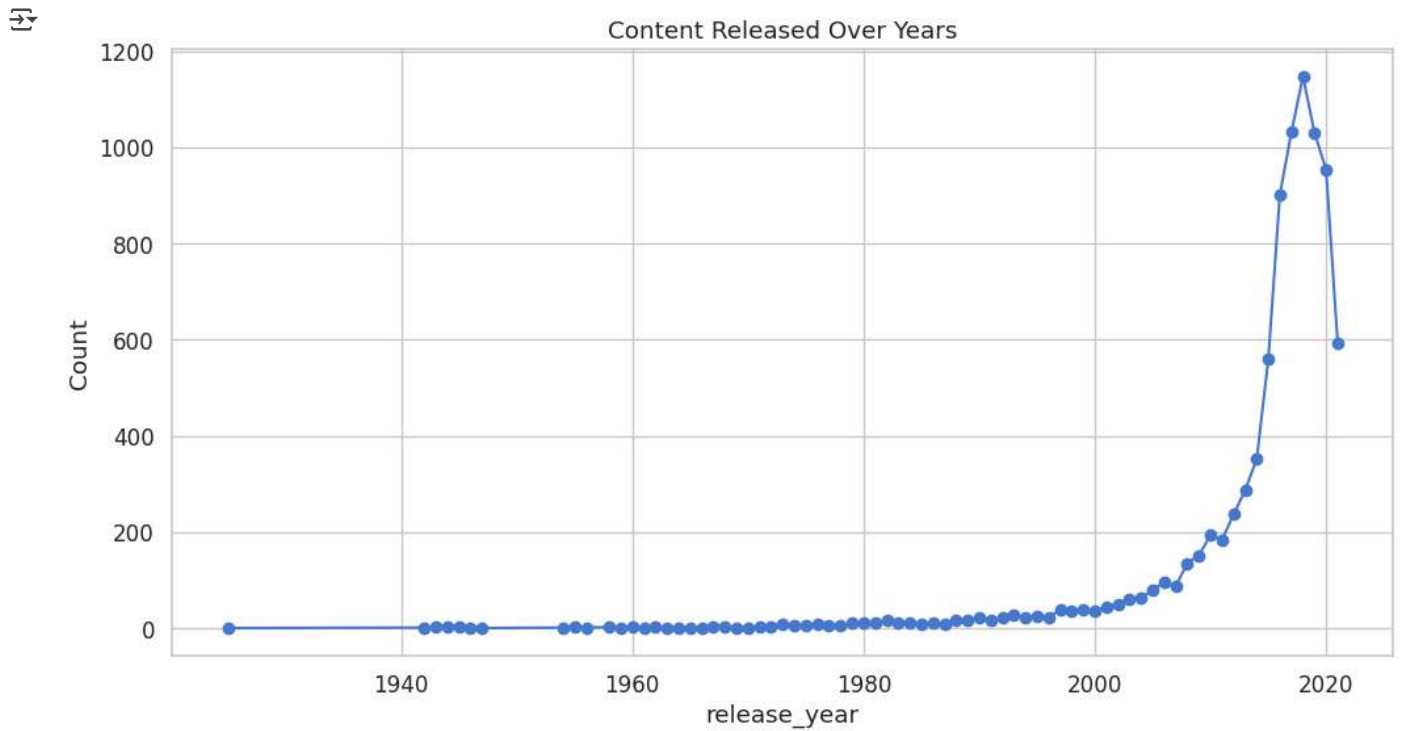
```
plt.show()
```



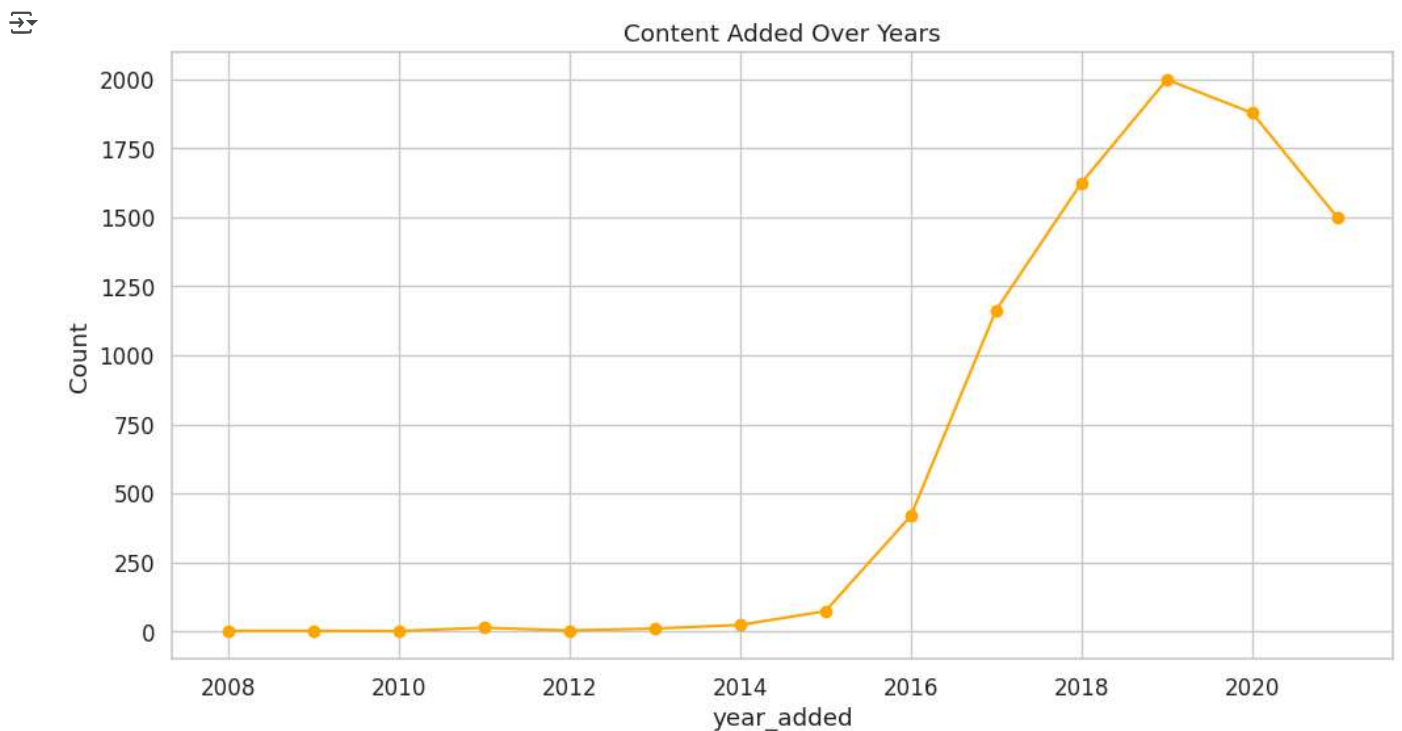
```
avg_duration = data.groupby('is_movie')['duration_int'].mean()
print("\nAverage Duration by Type:\n", avg_duration)
```

```
Average Duration by Type:
is_movie
0      1.764948
1     99.528462
Name: duration_int, dtype: float64
```

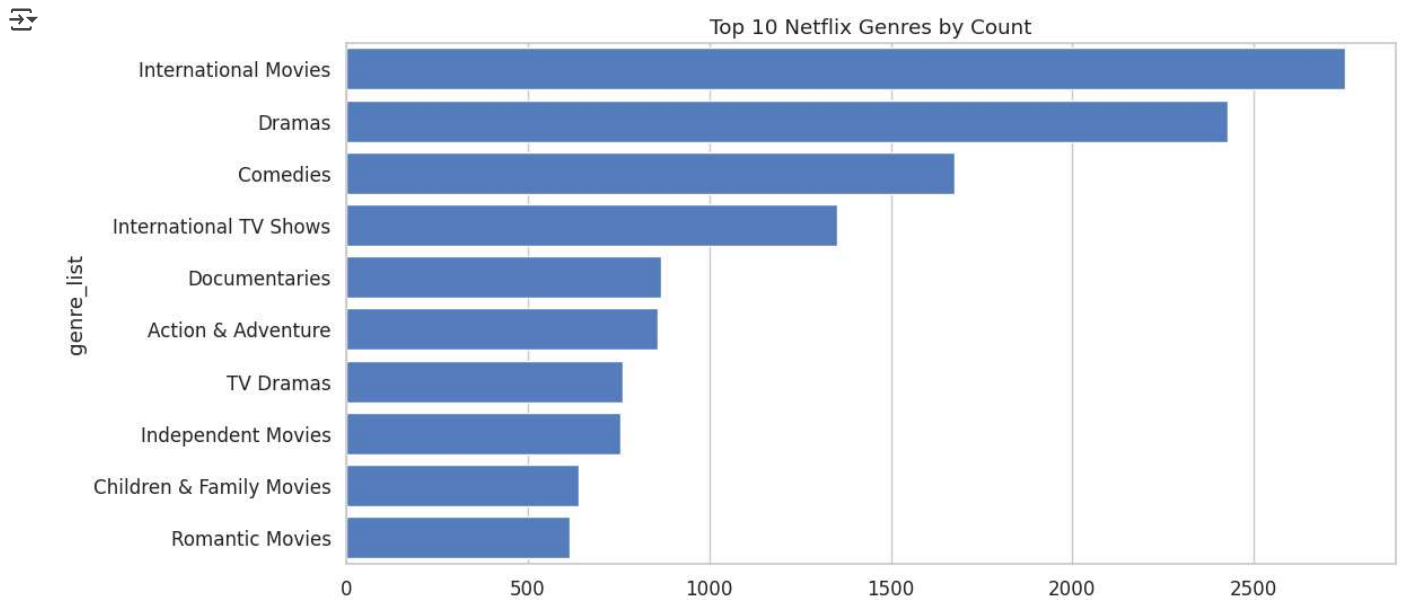
```
# 4c. Release Year Trend
if 'release_year' in data.columns:
    yearly_counts = data.groupby('release_year').size()
    yearly_counts.plot(kind='line', marker='o', title='Content Released Over Years')
    plt.ylabel('Count')
    plt.show()
```




```
# 4d. Date Added Trend
if 'year_added' in data.columns:
    added_counts = data.groupby('year_added').size()
    added_counts.plot(kind='line', marker='o', color='orange', title='Content Added Over Years')
    plt.ylabel('Count')
    plt.show()
```



```
# 4e. Genre Analysis
# Top genres by count
top_genres_count = genre_df['genre_list'].value_counts().head(10)
sns.barplot(x=top_genres_count.values, y=top_genres_count.index)
plt.title('Top 10 Netflix Genres by Count')
plt.show()
```



```
# Top genres by average duration
top_genres_duration = genre_df.groupby('genre_list')['duration_int'].mean().sort_values(ascending=False).head(10)
print("\nTop 10 Genres by Average Duration:\n", top_genres_duration)
```



```
Top 10 Genres by Average Duration:
genre_list
Classic Movies      118.646552
Action & Adventure  113.515716
Dramas              113.051092
Romantic Movies     110.573052
International Movies 110.349927
Thrillers           107.166378
Music & Musicals    106.125333
Sci-Fi & Fantasy    106.016461
Faith & Spirituality 105.584615
Cult Movies         104.521127
Name: duration_int, dtype: float64
```

```
# 4f. Country Analysis
if 'country' in data.columns:
    top_countries = data['country'].value_counts().head(10)
    sns.barplot(x=top_countries.values, y=top_countries.index)
    plt.title('Top 10 Countries by Content Count')
    plt.show()
```



Top 10 Countries by Content Count

```
# 4g. Correlation Analysis (numeric features)
numeric_cols = data.select_dtypes(include=np.number).columns
corr_matrix = data[numeric_cols].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Correlation Matrix of Numeric Features")
plt.show()
```

