

```
# Install seaborn if not already available
!pip install seaborn --quiet
```

```
# Import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Code cell output actions

```
# Set Seaborn style
sns.set(style="whitegrid")
```

```
from google.colab import files
uploaded = files.upload()
```

```
import io
# Use the actual uploaded filename from the keys of the uploaded dictionary
# The keys are typically the filenames of the uploaded files.
# In this case, the key is likely 'netflix_cleaned (1).csv' based on the traceback and global \
# We can get the first key from the dictionary, assuming only one file was uploaded.
uploaded_filename = list(uploaded.keys())[0]
```

```
df = pd.read_csv(io.BytesIO(uploaded[uploaded_filename]))
```



Choose Files netflix_cleaned.csv

- **netflix_cleaned.csv**(text/csv) - 3437779 bytes, last modified: 6/2/2025 - 100% done
Saving netflix_cleaned.csv to netflix_cleaned (2).csv

```
# Show basic structure
df.shape
```

```
# Display column info and types
df.info()
```

```
# View first 5 rows
df.head()
```



```
<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              6173 non-null   object
4   cast                 7982 non-null   object
5   country              7976 non-null   object
6   date_added            8709 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
12  added_year            8709 non-null   float64
13  added_month           8709 non-null   float64
dtypes: float64(2), int64(1), object(11)
memory usage: 963.4+ KB
```

Code cell output actions

	show_id	type	title	director	cast	country	date_added	release_year	rating
0	s1	movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-
1	s2	tv show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-14
2	s3	tv show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-14
3	s4	tv show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-14
4	s5	tv show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-14

Next steps:

[Generate code with df](#)[View recommended plots](#)[New interactive sheet](#)

```
# Null values
df.isnull().sum()
```

```
# Statistical summary
df.describe(include='all')
```



	show_id	type	title	director	cast	country	date_added	release_year	r
count	8807	8807	8807	6173	7982	7976	8709	8807.000000	
unique	8807	2	8807	4528	7692	748	1699	NaN	

Code cell output actions

top	s8807	movie	Zubaan	Rajiv Chilaka	David Attenborough	United States	2020-01-01	NaN	1
freq	1	6131	1	19	19	2818	109	NaN	
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2014.180198	
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8.819312	
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1925.000000	
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2013.000000	
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2017.000000	
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2019.000000	
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2021.000000	

```
sns.countplot(data=df, x='type', palette='Set2')
plt.title('Distribution of Content Type')
plt.show()
```

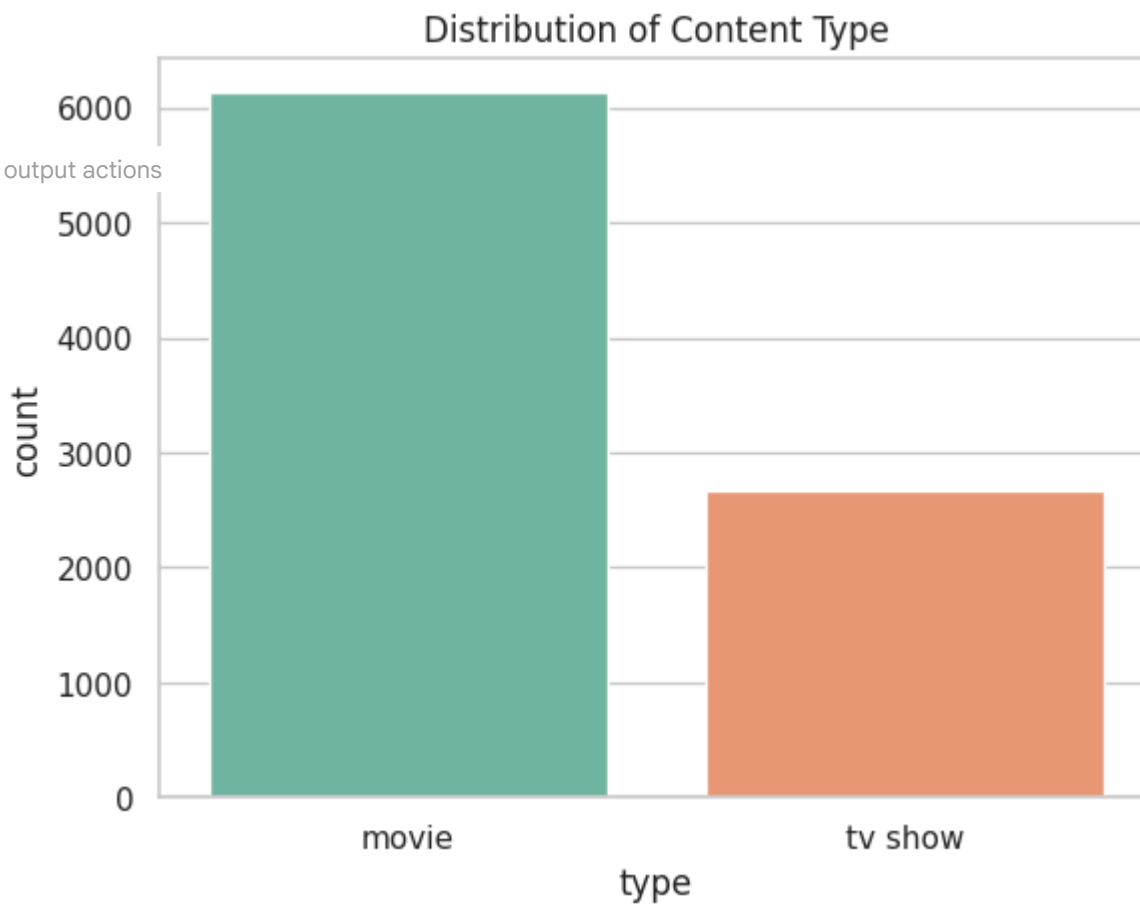


<ipython-input-6-1fc123c494c3>:1: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. As:

```
sns.countplot(data=df, x='type', palette='Set2')
```

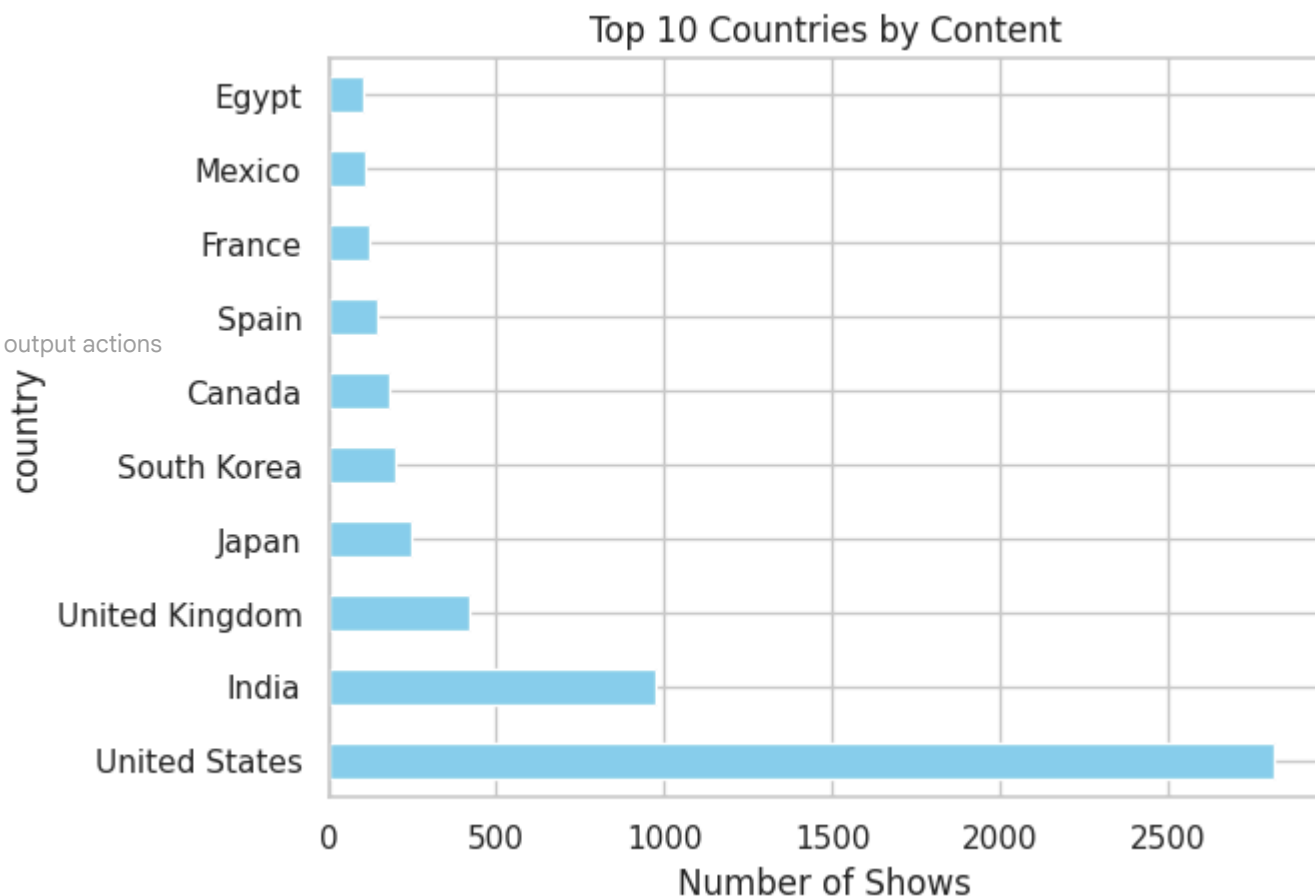
Code cell output actions



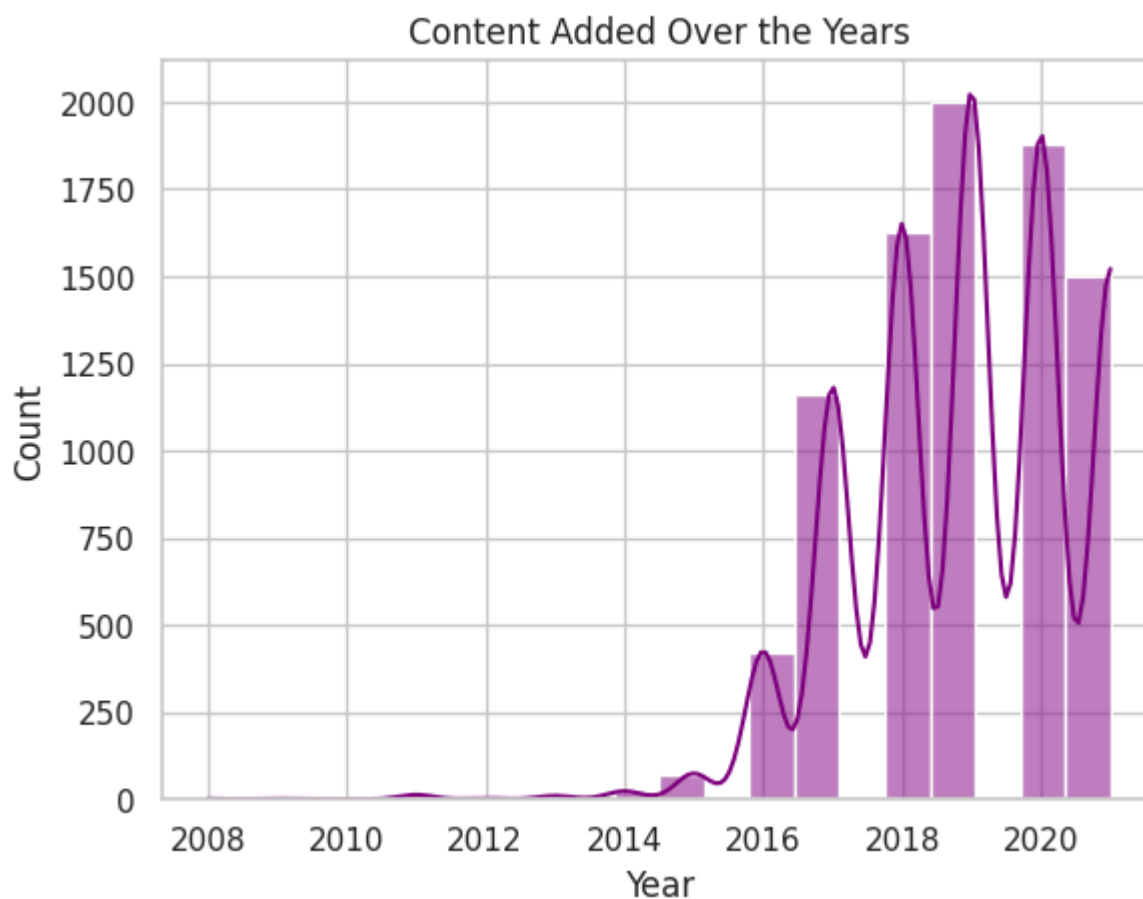
```
df['country'].value_counts().head(10).plot(kind='barh', color='skyblue')  
plt.title('Top 10 Countries by Content')  
plt.xlabel('Number of Shows')  
plt.show()
```



Code cell output actions



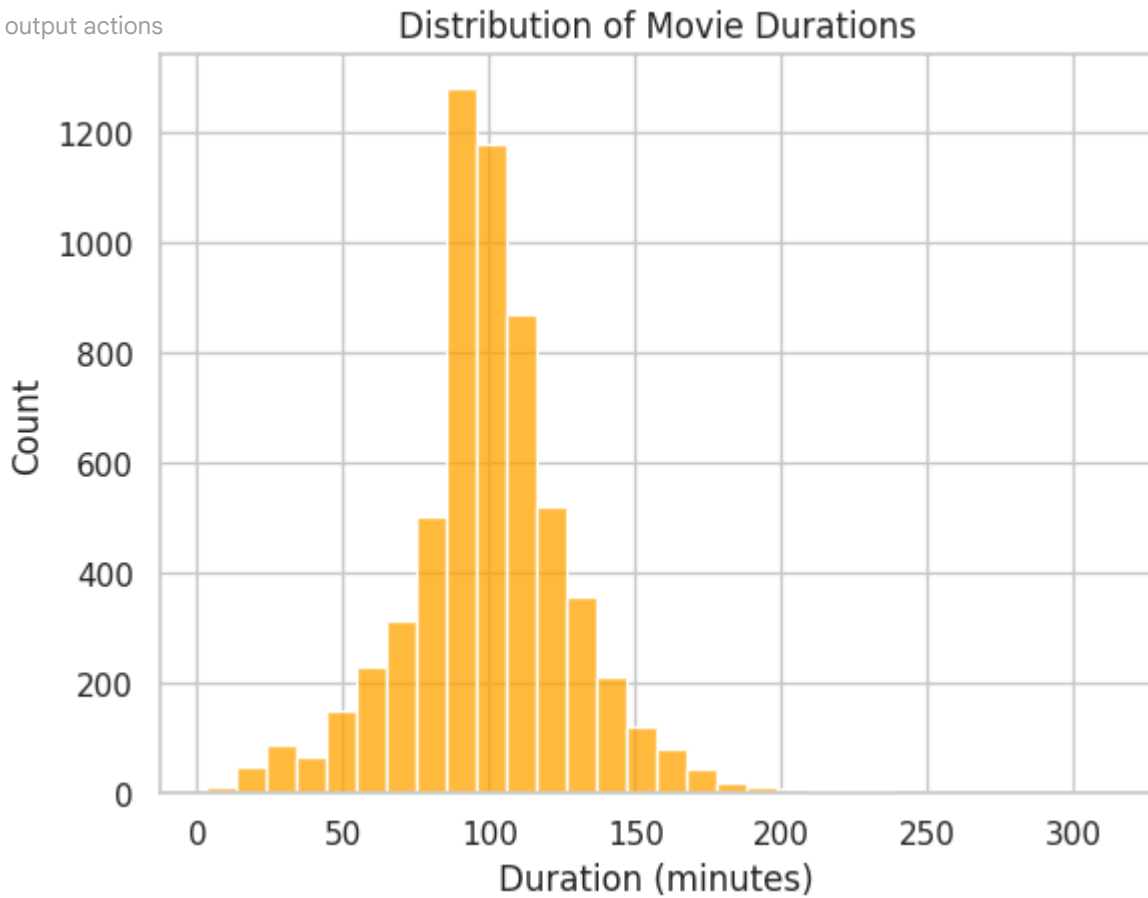
```
sns.histplot(df['added_year'], bins=20, kde=True, color='purple')  
plt.title('Content Added Over the Years')  
plt.xlabel('Year')  
plt.ylabel('Count')  
plt.show()
```



```
movies = df[df['type'] == 'movie'].copy()
movies['duration_int'] = movies['duration'].str.extract('(\d+)').astype(float)
```

```
sns.histplot(movies['duration_int'], bins=30, color='orange')
plt.title('Distribution of Movie Durations')
plt.xlabel('Duration (minutes)')
plt.show()
```

Code cell output actions



```
plt.figure(figsize=(12, 5))
sns.countplot(data=df, x='rating', hue='type', order=df['rating'].value_counts().index)
plt.title('Rating Distribution by Type')
plt.xticks(rotation=45)
plt.show()
```



```
sns.countplot(data=df, x='added_month', palette='viridis')
plt.title('Content Added by Month')
plt.xlabel('Month')
plt.ylabel('Count')
plt.show()
```

