In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Importing Dataset

In [2]:
```
df = pd.read_csv('/kaggle/input/exploratory-data-analysis-on-netflix-data/netflix_titles_2021.csv')
```

Data head

In [3]:
```
df.head()
```

Out[3]:

	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, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	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 t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

In [4]:
```
df.shape
```

Out[4]:
```
(8807, 12)
```

Information Dataset

In [5]:
```
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
```

Duplicated Values

In [6]:
```
df.duplicated().sum()
```

Out[6]: 0

Dropping missing values

In [7]:
```
df.drop(['show_id', 'cast', 'date_added', 'description'], inplace=True, axis=1)
df.head()
```

Out[7]:

	type	title	director	country	release_year	rating	duration	listed_in
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90 min	Documentaries
1	TV Show	Blood & Water	NaN	South Africa	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	TV Show	Ganglands	Julien Leclercq	NaN	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	TV Show	Jailbirds New Orleans	NaN	NaN	2021	TV-MA	1 Season	Docuseries, Reality TV
4	TV Show	Kota Factory	NaN	India	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...

In [8]:
```
df.fillna('Unknown',inplace=True,axis=1)
```

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
type            0
title           0
director        0
country         0
release_year    0
rating          0
duration        0
listed_in       0
dtype: int64
```

Exploratory Data

In [10]:
```
df.head()
```

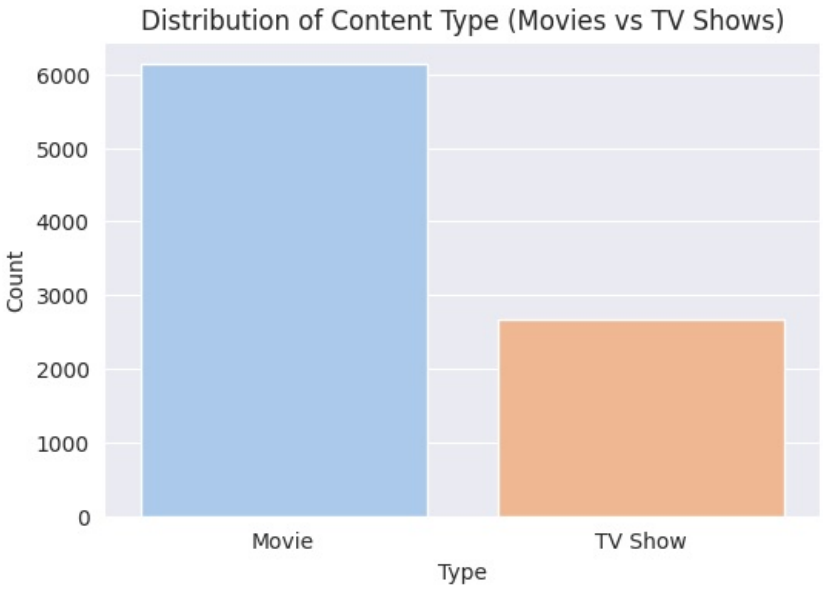
Out[10]:

	type	title	director	country	release_year	rating	duration	listed_in
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90 min	Documentaries
1	TV Show	Blood & Water	Unknown	South Africa	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	TV Show	Ganglands	Julien Leclercq	Unknown	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	TV Show	Jailbirds New Orleans	Unknown	Unknown	2021	TV-MA	1 Season	Docuseries, Reality TV
4	TV Show	Kota Factory	Unknown	India	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...

Data Visualization

In [11]:
```
sns.set_style("darkgrid")

plt.figure(figsize=(6, 4))
sns.countplot(x=df['type'], palette="pastel")
plt.title("Distribution of Content Type (Movies vs TV Shows)")
plt.xlabel("Type")
plt.ylabel("Count")
plt.show()
```



Add a new column with the new category names

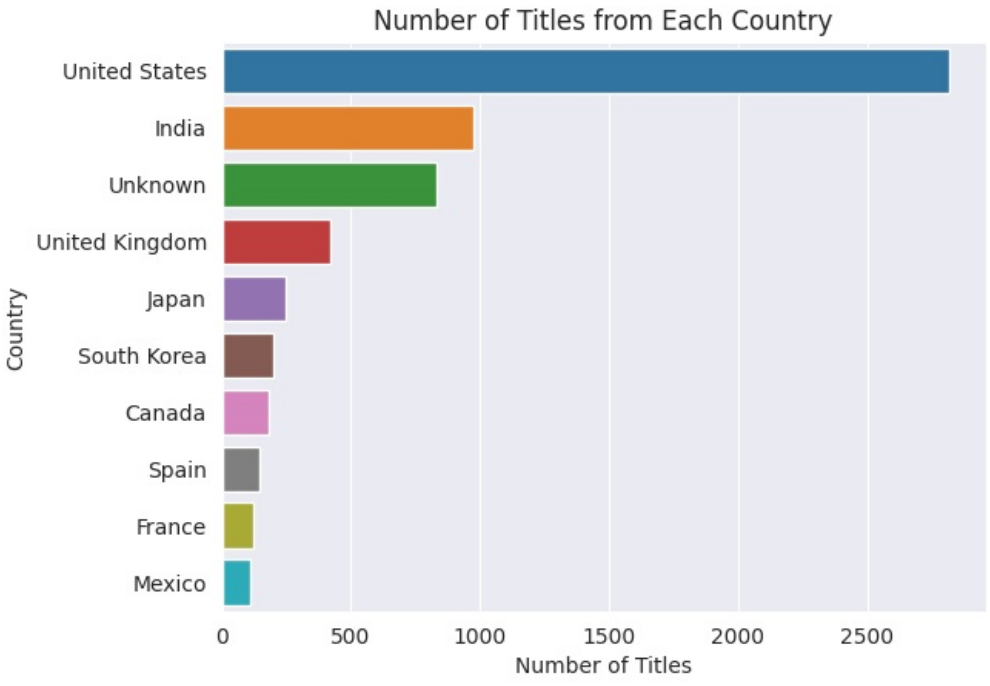
In [12]:
```
n_largest_titles = df['country'].value_counts().nlargest(10).index.tolist()
df['country_filtered'] = df['country'].apply(lambda x: x if x in n_largest_titles else "Other")
df.head()
```

Out[12]:

	type	title	director	country	release_year	rating	duration	listed_in	country_filtered
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90 min	Documentaries	United States
1	TV Show	Blood & Water	Unknown	South Africa	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	Other
2	TV Show	Ganglands	Julien Leclercq	Unknown	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	Unknown
3	TV Show	Jailbirds New Orleans	Unknown	Unknown	2021	TV-MA	1 Season	Docuseries, Reality TV	Unknown
4	TV Show	Kota Factory	Unknown	India	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	India

Plot the chart

In [13]:
```
ax = sns.countplot(y='country_filtered', data=df.dropna(subset='country'), order=n_largest_titles)
ax.set_xlabel('Number of Titles')
ax.set_ylabel('Country')
ax.set_title('Number of Titles from Each Country');
```



How many movies did each director direct?

Just like above, we'll only plot the top directors.

In [14]:
```
director_value_counts = df['director'].value_counts()
director_value_counts[director_value_counts >= 10]

# there are eleven directors with more than 10 titles so we use those
n_largest_directors = director_value_counts.nlargest(11).index
n_largest_directors
```

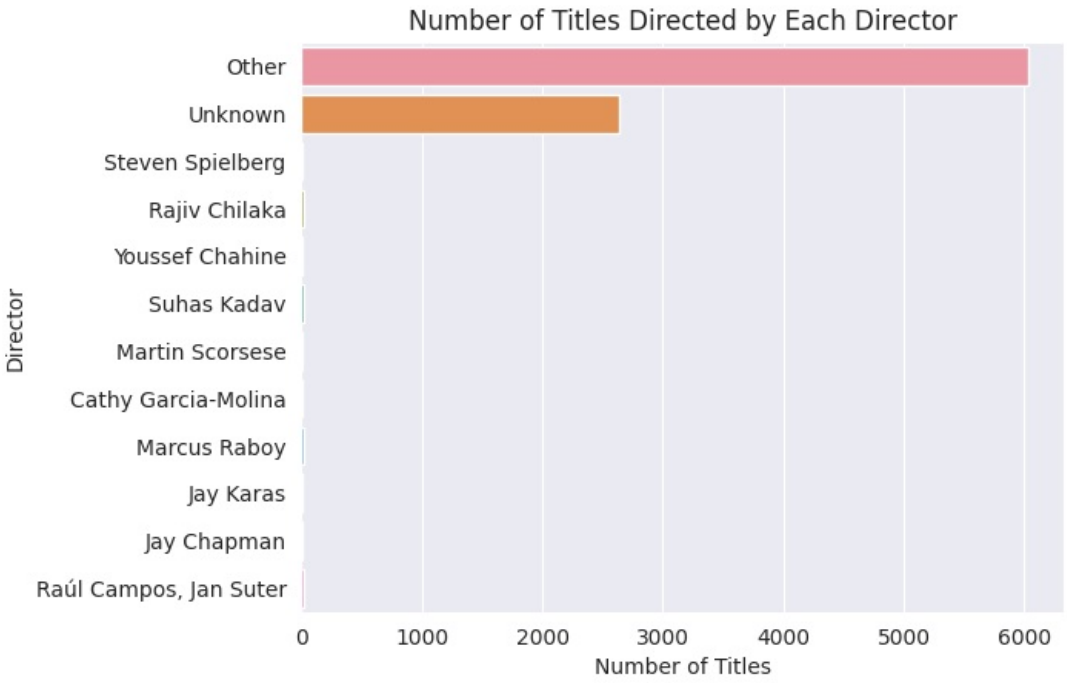
Out[14]:
```
Index(['Unknown', 'Rajiv Chilaka', 'Raúl Campos, Jan Suter', 'Suhas Kadav',
       'Marcus Raboy', 'Jay Karas', 'Cathy Garcia-Molina', 'Jay Chapman',
       'Youssef Chahine', 'Martin Scorsese', 'Steven Spielberg'],
      dtype='object', name='director')
```

In [15]:
```
df['director_filtered'] = df['director'].apply(lambda x: x if x in n_largest_directors else "Other")
df.head()
```

Out[15]:

	type	title	director	country	release_year	rating	duration	listed_in	country_filtered	director_filtered
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90 min	Documentaries	United States	Other
1	TV Show	Blood & Water	Unknown	South Africa	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	Other	Unknown
2	TV Show	Ganglands	Julien Leclercq	Unknown	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	Unknown	Other
3	TV Show	Jailbirds New Orleans	Unknown	Unknown	2021	TV-MA	1 Season	Docuseries, Reality TV	Unknown	Unknown
4	TV Show	Kota Factory	Unknown	India	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	India	Unknown

In [16]:
```
ax = sns.countplot(y='director_filtered', data=df.dropna(subset='director'))
ax.set_xlabel('Number of Titles')
ax.set_ylabel('Director')
ax.set_title('Number of Titles Directed by Each Director');
```



In [17]:
```
df.head()
```

Out[17]:

	type	title	director	country	release_year	rating	duration	listed_in	country_filtered	director_filtered
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90 min	Documentaries	United States	Other
1	TV Show	Blood & Water	Unknown	South Africa	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	Other	Unknown
2	TV Show	Ganglands	Julien Leclercq	Unknown	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	Unknown	Other
3	TV Show	Jailbirds New Orleans	Unknown	Unknown	2021	TV-MA	1 Season	Docuseries, Reality TV	Unknown	Unknown
4	TV Show	Kota Factory	Unknown	India	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	India	Unknown

In [18]:
```
df_movies = df.loc[df['type'] == 'Movie'].copy().reset_index(drop=True)
df_movies.head()
```

Out[18]:

	type	title	director	country	release_year	rating	duration	listed_in	country_filtered	director_filtered
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90 min	Documentaries	United States	Other
1	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Unknown	2021	PG	91 min	Children & Family Movies	Unknown	Other
2	Movie	Sankofa	Haile Gerima	United States, Ghana, Burkina Faso, United Kin...	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies	Other	Other
3	Movie	The Starling	Theodore Melfi	United States	2021	PG-13	104 min	Comedies, Dramas	United States	Other
4	Movie	Je Suis Karl	Christian Schwochow	Germany, Czech Republic	2021	TV-MA	127 min	Dramas, International Movies	Other	Other

Checking if all the rows have the same unit (min)

In [19]:
```
df_movies.duration.str.contains('min').sum() == df_movies.dropna(subset='duration').shape[0]
```

Out[19]:
```
False
```

Converting the durations to numbers

In [20]:
```
df_movies['duration'] = pd.to_numeric(df_movies['duration'].apply(lambda x: str(x).split()[0]), errors='coerce'
df_movies.head()
```

Out[20]:

	type	title	director	country	release_year	rating	duration	listed_in	country_filtered	director_filtered
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90.0	Documentaries	United States	Other
1	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Unknown	2021	PG	91.0	Children & Family Movies	Unknown	Other
2	Movie	Sankofa	Haile Gerima	United States, Ghana, Burkina Faso, United Kin...	1993	TV-MA	125.0	Dramas, Independent Movies, International Movies	Other	Other
3	Movie	The Starling	Theodore Melfi	United States	2021	PG-13	104.0	Comedies, Dramas	United States	Other
4	Movie	Je Suis Karl	Christian Schwochow	Germany, Czech Republic	2021	TV-MA	127.0	Dramas, International Movies	Other	Other

Create a new boolean column

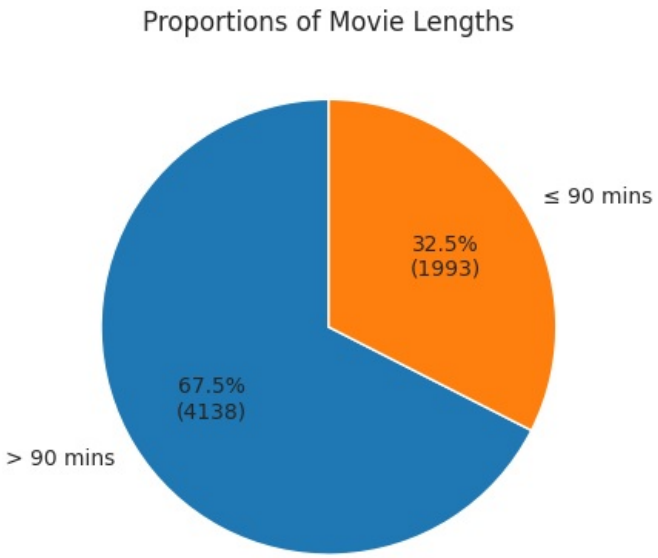
In [21]:
```
df_movies['longer_than_90'] = df_movies['duration'] > 90
```

```
/usr/local/lib/python3.11/dist-packages/pandas/core/computation/expressions.py:73: RuntimeWarning: invalid value
encountered in greater
  return op(a, b)
```

Creating a function so that the chart shows the respective

In [22]:
```
def movie_autopct(pct):
    total = sum(df_movies['longer_than_90'].value_counts())
    count = int(round(pct * total / 100.0))
    return f'{pct:.1f}%\n({count})'

# Plotting the chart
df_movies['longer_than_90'].value_counts().plot.pie(
    labels=['> 90 mins', '≤ 90 mins'],
    autopct=movie_autopct,
    startangle=90,
    ylabel=''
)
plt.title('Proportions of Movie Lengths')
plt.show()
```



In [23]:
```
df.head()
```

Out[23]:

	type	title	director	country	release_year	rating	duration	listed_in	country_filtered	director_filtered
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90 min	Documentaries	United States	Other
1	TV Show	Blood & Water	Unknown	South Africa	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	Other	Unknown
2	TV Show	Ganglands	Julien Leclercq	Unknown	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	Unknown	Other
3	TV Show	Jailbirds New Orleans	Unknown	Unknown	2021	TV-MA	1 Season	Docuseries, Reality TV	Unknown	Unknown
4	TV Show	Kota Factory	Unknown	India	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	India	Unknown

What is the shortest movie?

In [24]:
```
df_movies.sort_values(by='duration').head(10)
```

Out[24]:

	type	title	director	country	release_year	rating	duration	listed_in	country_filtered	director_filtered
2402	Movie	Silent	Limbert Fabian, Brandon Oldenburg	United States	2014	TV-Y	3.0	Children & Family Movies, Sci-Fi & Fantasy	United States	Other
1777	Movie	Sol Levante	Akira Saitoh	Japan	2020	TV-14	5.0	Action & Adventure, Anime Features, Internatio...	Japan	Other
972	Movie	Cops and Robbers	Arnon Manor, Timothy Ware-Hill	United States	2020	PG-13	8.0	Dramas	United States	Other
1017	Movie	Canvas	Frank E. Abney III	United States	2020	G	9.0	Children & Family Movies, Dramas	United States	Other
2276	Movie	American Factory: A Conversation with the Obamas	Unknown	United States	2019	TV-G	10.0	Documentaries	United States	Unknown
1873	Movie	Calico Critters: Everyone's Big Dream Flying i...	Momoko Kamiya	Unknown	2019	TV-Y	11.0	Children & Family Movies	Unknown	Other
4229	Movie	Calico Critters: A Town of Dreams	Momoko Kamiya	Unknown	2017	TV-Y	11.0	Children & Family Movies	Unknown	Other
2400	Movie	Cosmos Laundromat: First Cycle	Mathieu Auvray	Netherlands	2015	TV-MA	12.0	Dramas, International Movies, Sci-Fi & Fantasy	Other	Other
3008	Movie	Zion	Floyd Russ	United States	2018	TV-PG	12.0	Documentaries, Sports Movies	United States	Other
439	Movie	Besieged Bread	Soudade Kaadan	Unknown	2015	TV-14	12.0	Dramas, International Movies	Unknown	Other

In [25]:
```
df.head()
```

Out[25]:

	type	title	director	country	release_year	rating	duration	listed_in	country_filtered	director_filtered
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90 min	Documentaries	United States	Other
1	TV Show	Blood & Water	Unknown	South Africa	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	Other	Unknown
2	TV Show	Ganglands	Julien Leclercq	Unknown	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	Unknown	Other
3	TV Show	Jailbirds New Orleans	Unknown	Unknown	2021	TV-MA	1 Season	Docuseries, Reality TV	Unknown	Unknown
4	TV Show	Kota Factory	Unknown	India	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	India	Unknown

In [26]:
```
df_movies.head()
```

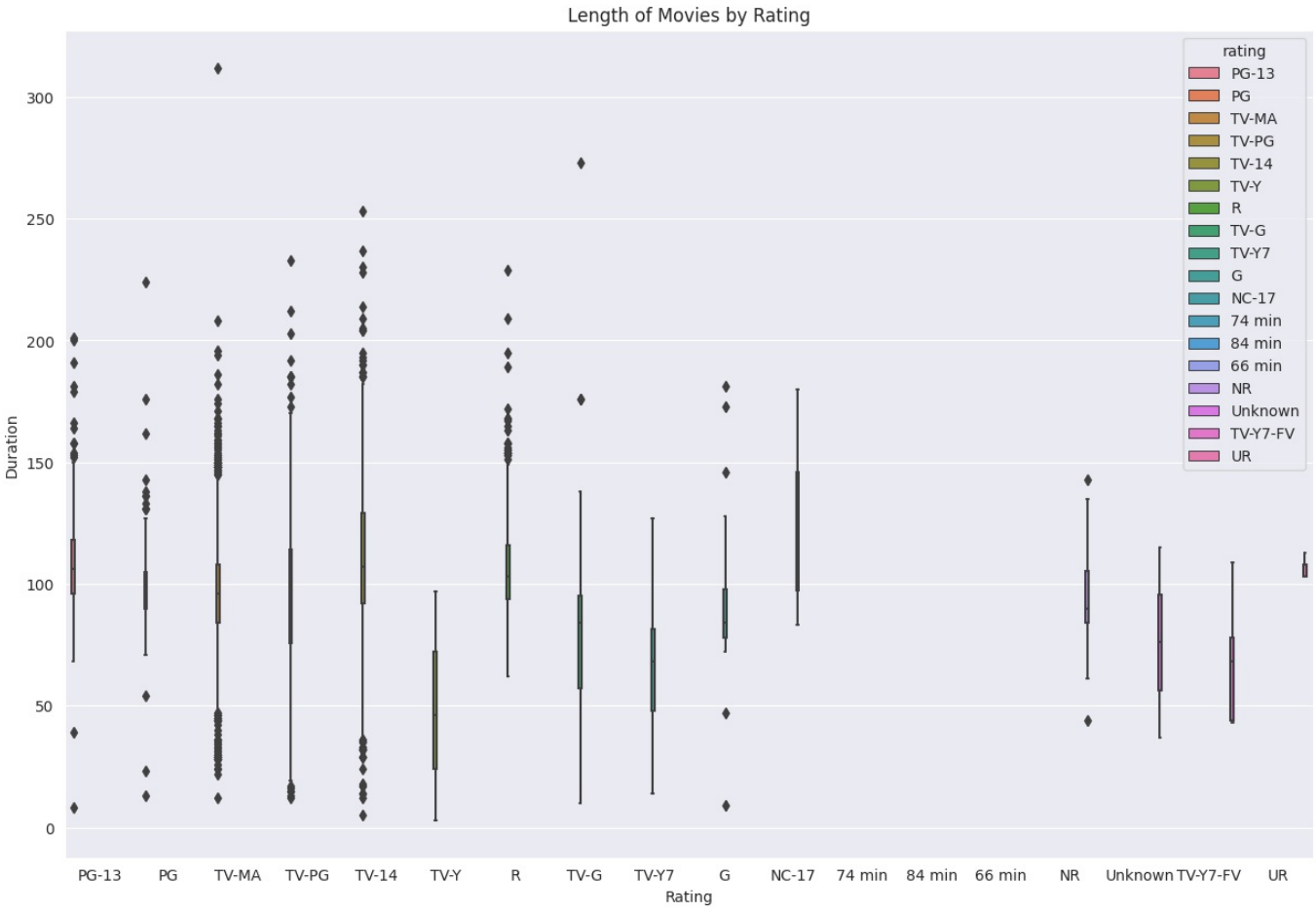
Out[26]:

	type	title	director	country	release_year	rating	duration	listed_in	country_filtered	director_filtered	longer_
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90.0	Documentaries	United States	Other	
1	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Unknown	2021	PG	91.0	Children & Family Movies	Unknown	Other	
2	Movie	Sankofa	Haile Gerima	United States, Ghana, Burkina Faso, United Kin...	1993	TV-MA	125.0	Dramas, Independent Movies, International Movies	Other	Other	
3	Movie	The Starling	Theodore Melfi	United States	2021	PG-13	104.0	Comedies, Dramas	United States	Other	
4	Movie	Je Suis Karl	Christian Schwochow	Germany, Czech Republic	2021	TV-MA	127.0	Dramas, International Movies	Other	Other	

Length of Movies

A boxplot seems appropriate to me in this scenario. It can show how the ratings relate to the length of the movies, while giving us valuable insights regarding the outliers.

In [27]:
```
plt.figure(figsize=(15, 10))
sns.boxplot(x='rating', y='duration', data=df_movies, hue='rating', palette='husl')
plt.xlabel('Rating')
plt.ylabel('Duration')
plt.title('Length of Movies by Rating');
```



get the count and rating into a dataframe

In [28]:
```
rating_count_ser = df.groupby('rating').duration.count()
rating_count_df = rating_count_ser.to_frame().reset_index()
rating_count_df.columns=['rating', 'duration_count']
rating_count_df
```

Out[28]:

	rating	duration_count
0	66 min	1
1	74 min	1
2	84 min	1
3	G	41
4	NC-17	3
5	NR	80
6	PG	287
7	PG-13	490
8	R	799
9	TV-14	2160
10	TV-G	220
11	TV-MA	3207
12	TV-PG	863
13	TV-Y	307
14	TV-Y7	334
15	TV-Y7-FV	6
16	UR	3
17	Unknown	4

In [29]:

```
plt.figure(figsize=(15, 10))
sns.barplot(data=rating_count_df, x='rating', y='duration_count')
plt.xlabel('Rating')
plt.ylabel('Count')
plt.title('Number of Titles by Rating');
```