Prime Video Analysis EDA


```
In [1]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        import warnings
        warnings.filterwarnings('ignore')
```

```
In [2]: df=pd.read_csv(r"C:\Users\chitt\Downloads\amazon_prime_titles.csv\amazon_prime_t
        df
```

Out[2]:

	show_id	type	title	director	cast	country	date_added	release
0	s1	Movie	The Grand Seduction	Don McKellar	Brendan Gleeson, Taylor Kitsch, Gordon Pinsent	Canada	March 30, 2021	
1	s2	Movie	Take Care Good Night	Girish Joshi	Mahesh Manjrekar, Abhay Mahajan, Sachin Khedekar	India	March 30, 2021	
2	s3	Movie	Secrets of Deception	Josh Webber	Tom Sizemore, Lorenzo Lamas, Robert LaSardo, R…	United States	March 30, 2021	
3	s4	Movie	Pink: Staying True	Sonia Anderson	Interviews with: Pink, Adele, Beyoncé, Britney…	United States	March 30, 2021	
4	s5	Movie	Monster Maker	Giles Foster	Harry Dean Stanton, Kieran O'Brien, George Cos…	United Kingdom	March 30, 2021	
...	...	...	...	...	...	...	...	
9663	s9664	Movie	Pride Of The Bowery	Joseph H. Lewis	Leo Gorcey, Bobby Jordan	NaN	NaN	
9664	s9665	TV Show	Planet Patrol	NaN	DICK VOSBURGH, RONNIE STEVENS, LIBBY MORRIS, M…	NaN	NaN	
9665	s9666	Movie	Outpost	Steve Barker	Ray Stevenson, Julian Wadham, Richard Brake, M…	NaN	NaN	

	show_id	type	title	director	cast	country	date_added	release
9666	s9667	TV Show	Maradona: Blessed Dream	NaN	Esteban Recagno, Ezequiel Stremiz, Luciano Vit...	NaN	NaN	
9667	s9668	Movie	Harry Brown	Daniel Barber	Michael Caine, Emily Mortimer, Joseph Gilgun, ...	NaN	NaN	

9668 rows × 12 columns

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

Out[3]:

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	The Grand Seduction	Don McKellar	Brendan Gleeson, Taylor Kitsch, Gordon Pinsent	Canada	March 30, 2021	2014
1	s2	Movie	Take Care Good Night	Girish Joshi	Mahesh Manjrekar, Abhay Mahajan, Sachin Khedekar	India	March 30, 2021	2018
2	s3	Movie	Secrets of Deception	Josh Webber	Tom Sizemore, Lorenzo Lamas, Robert LaSardo, R...	United States	March 30, 2021	2017
3	s4	Movie	Pink: Staying True	Sonia Anderson	Interviews with: Pink, Adele, Beyoncé, Britney...	United States	March 30, 2021	2014
4	s5	Movie	Monster Maker	Giles Foster	Harry Dean Stanton, Kieran O'Brien, George Cos...	United Kingdom	March 30, 2021	1989

In [4]: `df.tail()`

Out[4]:

	show_id	type	title	director	cast	country	date_added	release_y
9663	s9664	Movie	Pride Of The Bowery	Joseph H. Lewis	Leo Gorcey, Bobby Jordan	NaN	NaN	1
9664	s9665	TV Show	Planet Patrol	NaN	DICK VOSBURGH, RONNIE STEVENS, LIBBY MORRIS, M...	NaN	NaN	2
9665	s9666	Movie	Outpost	Steve Barker	Ray Stevenson, Julian Wadham, Richard Brake, M...	NaN	NaN	2
9666	s9667	TV Show	Maradona: Blessed Dream	NaN	Esteban Recagno, Ezequiel Stremiz, Luciano Vit...	NaN	NaN	2
9667	s9668	Movie	Harry Brown	Daniel Barber	Michael Caine, Emily Mortimer, Joseph Gilgun, ...	NaN	NaN	2

In [5]: `df.shape`

Out[5]: (9668, 12)

In [6]: `df.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9668 entries, 0 to 9667
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       9668 non-null   object
 1   type          9668 non-null   object
 2   title         9668 non-null   object
 3   director      7585 non-null   object
 4   cast          8435 non-null   object
 5   country       672 non-null    object
 6   date_added    155 non-null    object
 7   release_year  9668 non-null   int64
 8   rating        9331 non-null   object
 9   duration      9668 non-null   object
 10  listed_in     9668 non-null   object
 11  description   9668 non-null   object
dtypes: int64(1), object(11)
memory usage: 906.5+ KB
```

In [7]: `df.isnull().sum()`

Out[7]:
```
show_id            0
type               0
title              0
director        2083
cast            1233
country         8996
date_added      9513
release_year       0
rating           337
duration           0
listed_in          0
description        0
dtype: int64
```

In [8]: `df.describe()`

Out[8]:

	release_year
count	9668.000000
mean	2008.341849
std	18.922482
min	1920.000000
25%	2007.000000
50%	2016.000000
75%	2019.000000
max	2021.000000

In [9]: `df.isnull().sum()/len(df)*100`

```
Out[9]:   show_id          0.000000
          type             0.000000
          title            0.000000
          director        21.545304
          cast            12.753413
          country         93.049235
          date_added      98.396773
          release_year     0.000000
          rating           3.485726
          duration         0.000000
          listed_in        0.000000
          description      0.000000
          dtype: float64
```

In [10]:
```python
df.duplicated().sum()
```

Out[10]: 0

In [11]:
```python
df.drop('description', axis=1, inplace = True)
```

In [12]:
```python
df.head()
```

Out[12]:

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	The Grand Seduction	Don McKellar	Brendan Gleeson, Taylor Kitsch, Gordon Pinsent	Canada	March 30, 2021	2014
1	s2	Movie	Take Care Good Night	Girish Joshi	Mahesh Manjrekar, Abhay Mahajan, Sachin Khedekar	India	March 30, 2021	2018
2	s3	Movie	Secrets of Deception	Josh Webber	Tom Sizemore, Lorenzo Lamas, Robert LaSardo, R...	United States	March 30, 2021	2017
3	s4	Movie	Pink: Staying True	Sonia Anderson	Interviews with: Pink, Adele, Beyoncé, Britney...	United States	March 30, 2021	2014
4	s5	Movie	Monster Maker	Giles Foster	Harry Dean Stanton, Kieran O'Brien, George Cos...	United Kingdom	March 30, 2021	1989

In [13]:
```python
df['director'].fillna('unavailable',inplace = True)
df['cast'].fillna('unavailable',inplace = True)
df['country'] = df['country'].fillna(df['country'].mode()[0])
df['date_added'] = df['date_added'].ffill()
df['rating'] = df['rating'].fillna(df['rating'].mode()[0])
```

In [14]:
```python
df.isnull().sum()
```

Out[14]:
```
show_id         0
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
dtype: int64
```

```
In [15]: df['date_added'] = pd.to_datetime(df['date_added'] , format='%B %d, %Y')
```

```
In [16]: df.head(5)
```

Out[16]:

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	The Grand Seduction	Don McKellar	Brendan Gleeson, Taylor Kitsch, Gordon Pinsent	Canada	2021-03-30	2014
1	s2	Movie	Take Care Good Night	Girish Joshi	Mahesh Manjrekar, Abhay Mahajan, Sachin Khedekar	India	2021-03-30	2018
2	s3	Movie	Secrets of Deception	Josh Webber	Tom Sizemore, Lorenzo Lamas, Robert LaSardo, R...	United States	2021-03-30	2017
3	s4	Movie	Pink: Staying True	Sonia Anderson	Interviews with: Pink, Adele, Beyoncé, Britney...	United States	2021-03-30	2014
4	s5	Movie	Monster Maker	Giles Foster	Harry Dean Stanton, Kieran O'Brien, George Cos...	United Kingdom	2021-03-30	1989

```
In [17]: df['rating'].unique()
```

```
Out[17]: array(['13+', 'ALL', '18+', 'R', 'TV-Y', 'TV-Y7', 'NR', '16+', 'TV-PG',
                '7+', 'TV-14', 'TV-NR', 'TV-G', 'PG-13', 'TV-MA', 'G', 'PG',
                'NC-17', 'UNRATED', '16', 'AGES_16_', 'AGES_18_', 'ALL_AGES',
                'NOT_RATE'], dtype=object)
```

```
In [18]: df['rating'] = df['rating'].replace({'AGES_16_':'16+' , 'AGES_18_':'18+' , 'NOT_
```

```
In [19]: df['type'].value_counts()
```

```
Out[19]: type
         Movie      7814
         TV Show    1854
         Name: count, dtype: int64
```

In [20]:
```python
df.rename(columns={'listed_in':'categories'} , inplace = True)
```

In [44]:
```python
df['duration_min'] = df['duration'].apply(lambda x : int(x.split()[0]) if 'min'
```

In [46]:
```python
df['duration_seasons'] = df['duration'].apply(lambda x : int(x.split()[0]) if 'S
```

In [48]:
```python
df.describe()
```

Out[48]:

|	date_added	release_year	duration_min	duration_seasons
count	9668	9668.000000	9668.000000	9668.000000
mean	2021-06-04 18:47:48.680182016	2008.341849	73.801303	0.191663
min	2021-03-30 00:00:00	1920.000000	0.000000	0.000000
25%	2021-03-30 00:00:00	2007.000000	41.000000	0.000000
50%	2021-05-22 00:00:00	2016.000000	86.000000	0.000000
75%	2021-08-06 00:00:00	2019.000000	101.000000	0.000000
max	2021-10-10 00:00:00	2021.000000	601.000000	29.000000
std	NaN	18.922482	51.060792	1.014932

In [50]:
```python
df.describe(include='object')
```

Out[50]:

	show_id	type	title	director	cast	country	rating	duration
count	9668	9668	9668	9668	9668	9668	9668	9668
unique	9668	2	9668	5774	7928	86	21	219
top	s1	Movie	The Grand Seduction	unavailable	unavailable	United States	13+	1 Season
freq	1	7814	1	2083	1233	9249	2454	1342

In [56]:
```python
sns.countplot(data=df, x='type',hue='type')
plt.show()
```
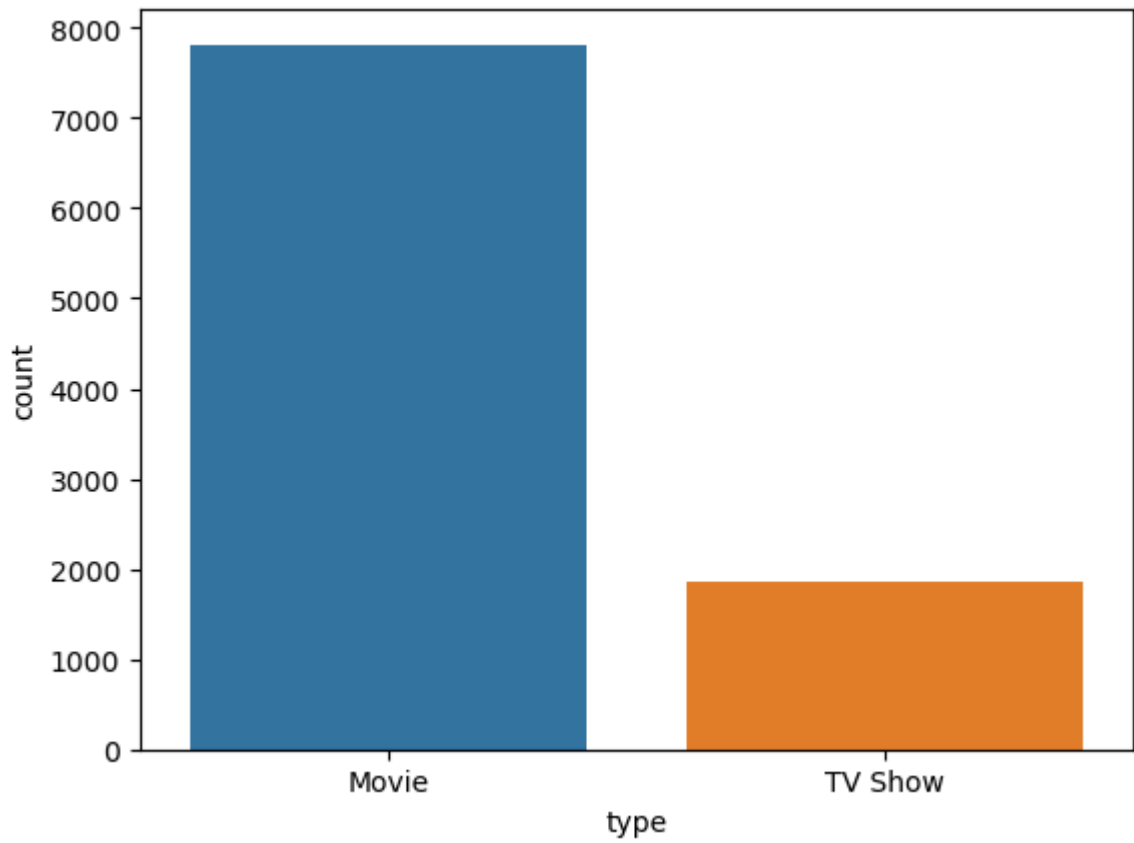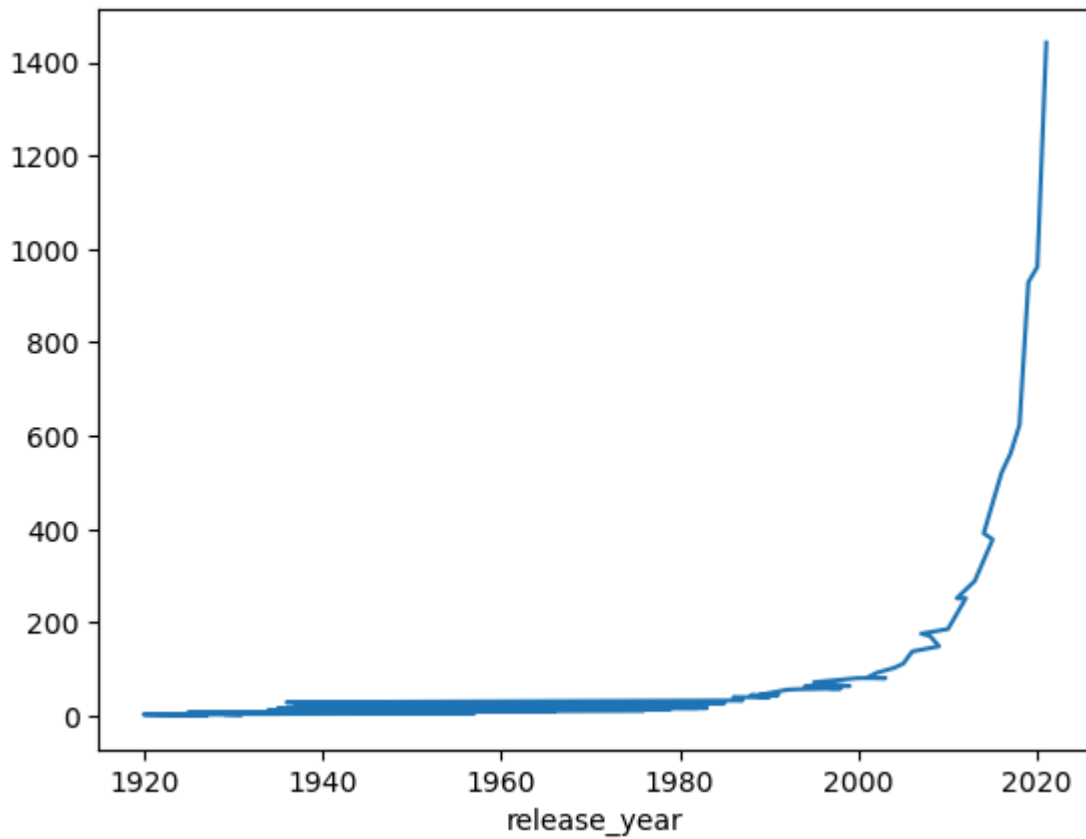
```
In [62]:  df['categories'].value_counts()[:10].plot.bar()
          plt.xticks(rotation=45)
          plt.show()
```
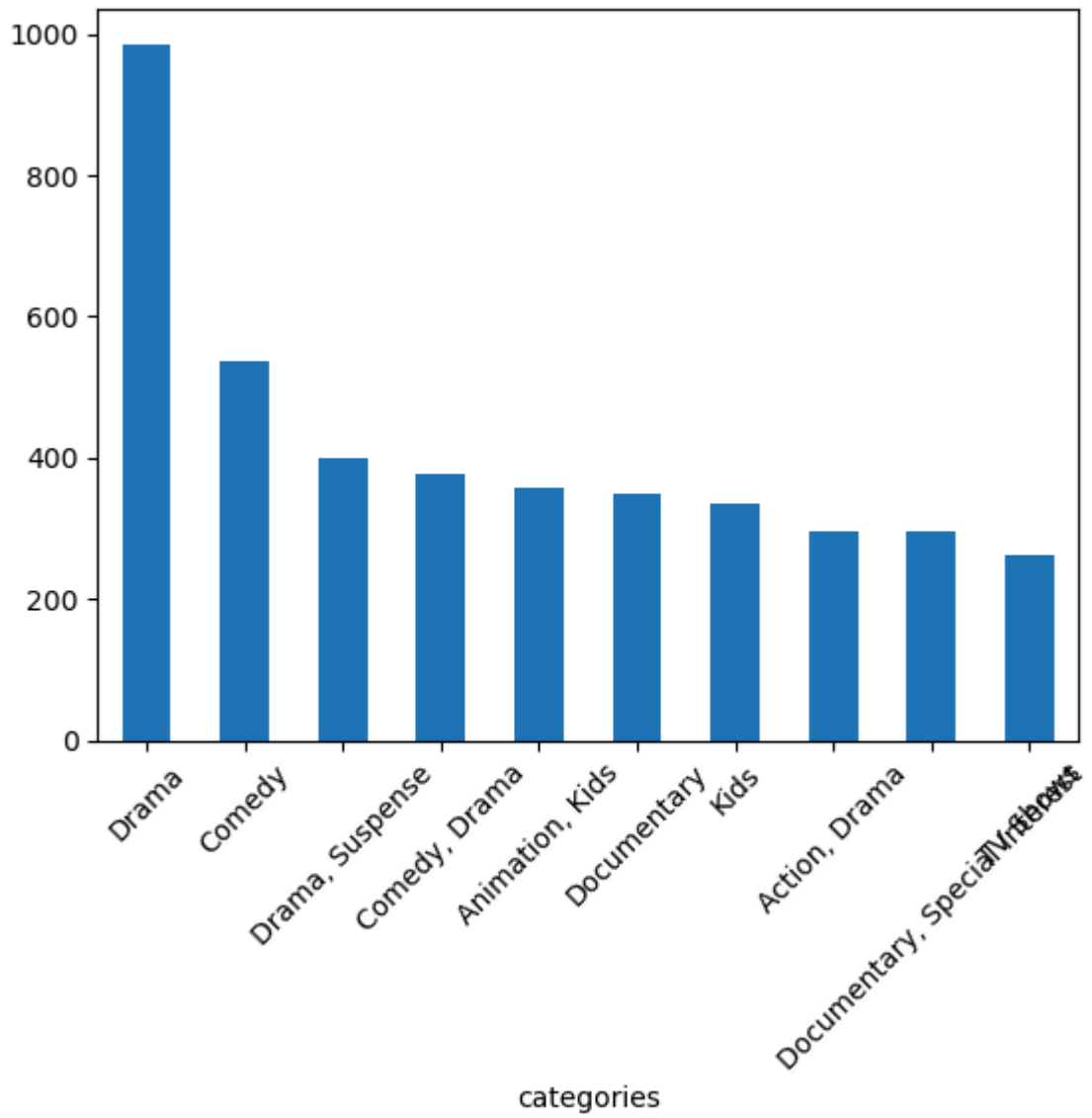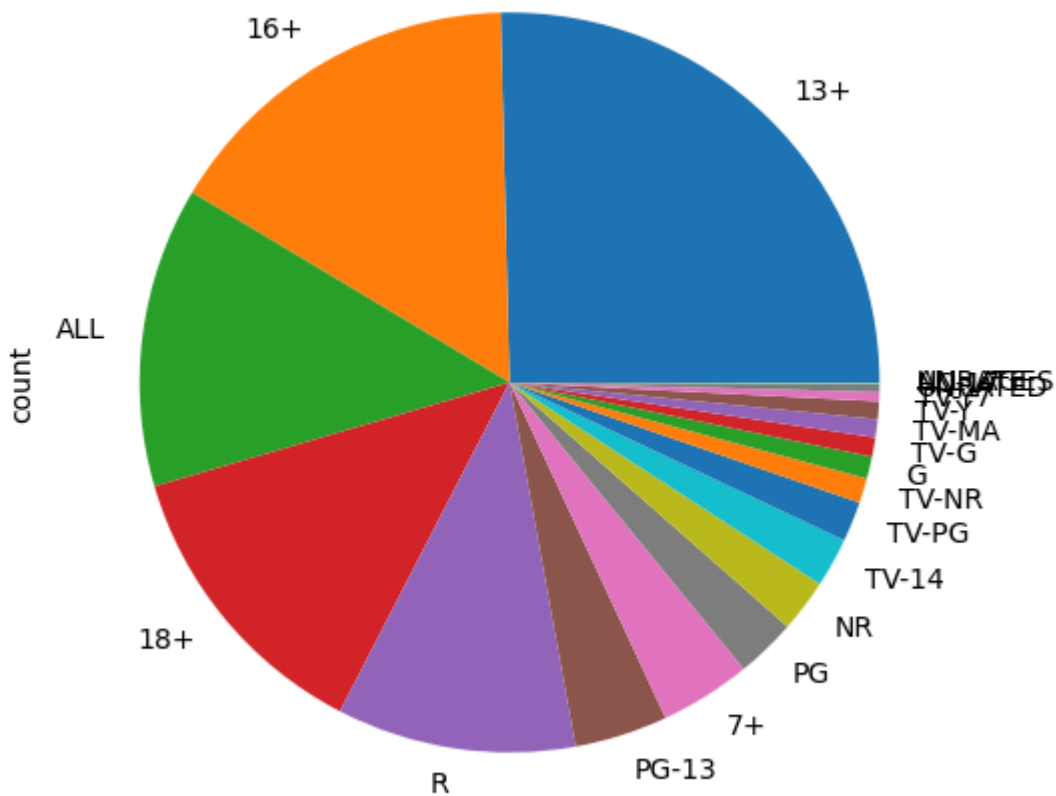
```
In [64]: df['release_year'].value_counts().plot.line()
         plt.title('Content Relesed Over the Years')
         plt.show()
```
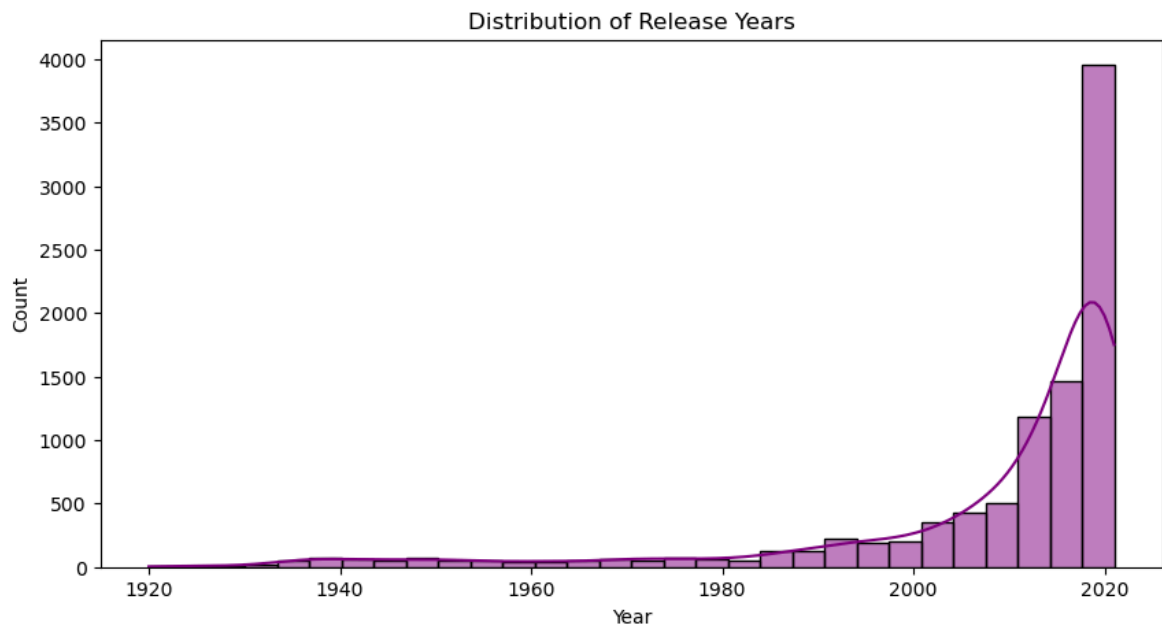
Content Relesed Over the Years


```
In [66]:  plt.figure(figsize = (8,6))
          df['rating'].value_counts().plot.pie()
          plt.show()
```
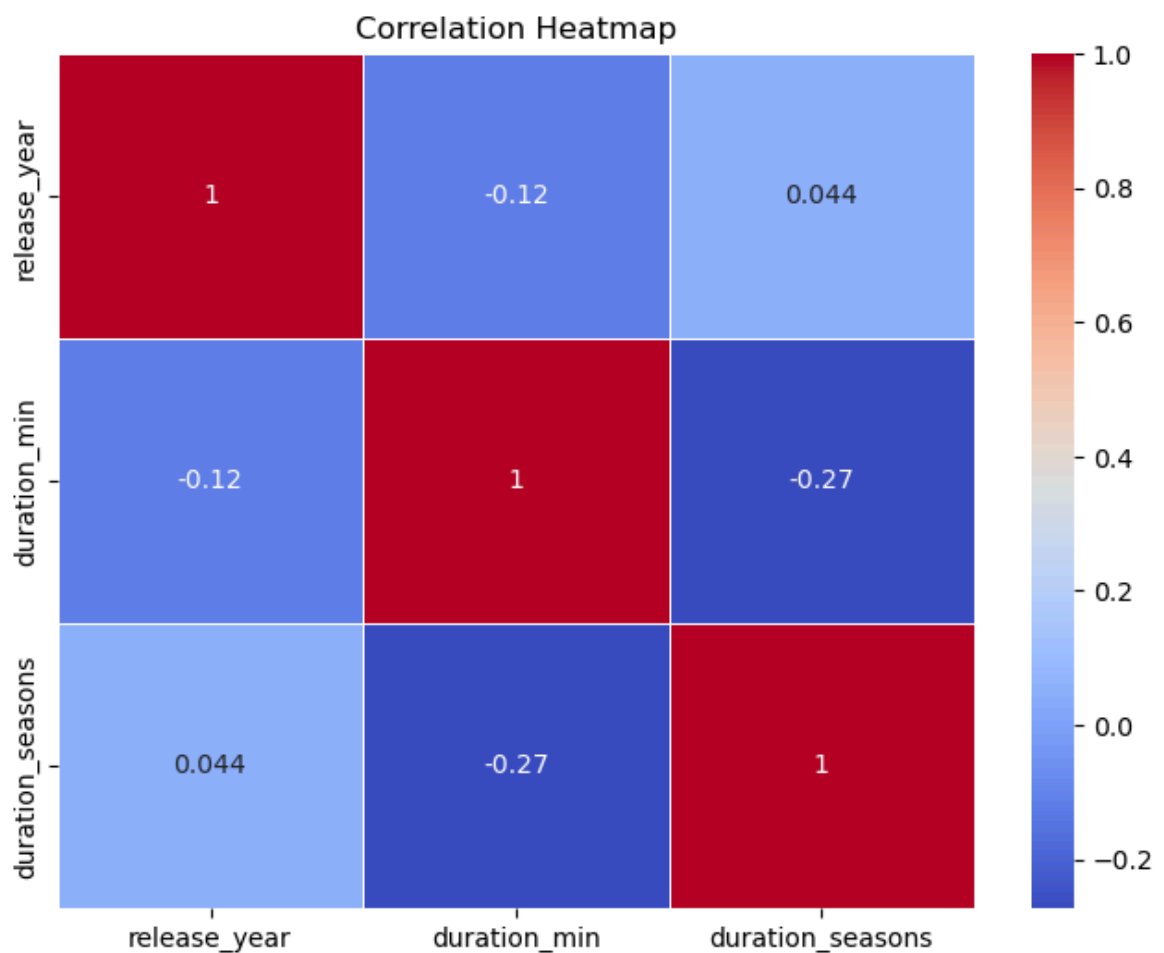
In [74]:
```python
# Movie release years distribution
plt.figure(figsize=(10,5))
sns.histplot(df['release_year'], bins=30, kde=True, color='purple')
plt.title('Distribution of Release Years')
plt.xlabel('Year')
plt.ylabel('Count')
plt.show()
```


Distribution of Release Years

In [84]:
```python
# Top 10 content-producing countries
plt.figure(figsize=(10,5))
df['country'].value_counts().nlargest(10).plot(kind='bar', color='orange')
plt.title('Top 10 Content Producing Countries')
plt.xlabel('Country')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```

Top 10 Content Producing Countries
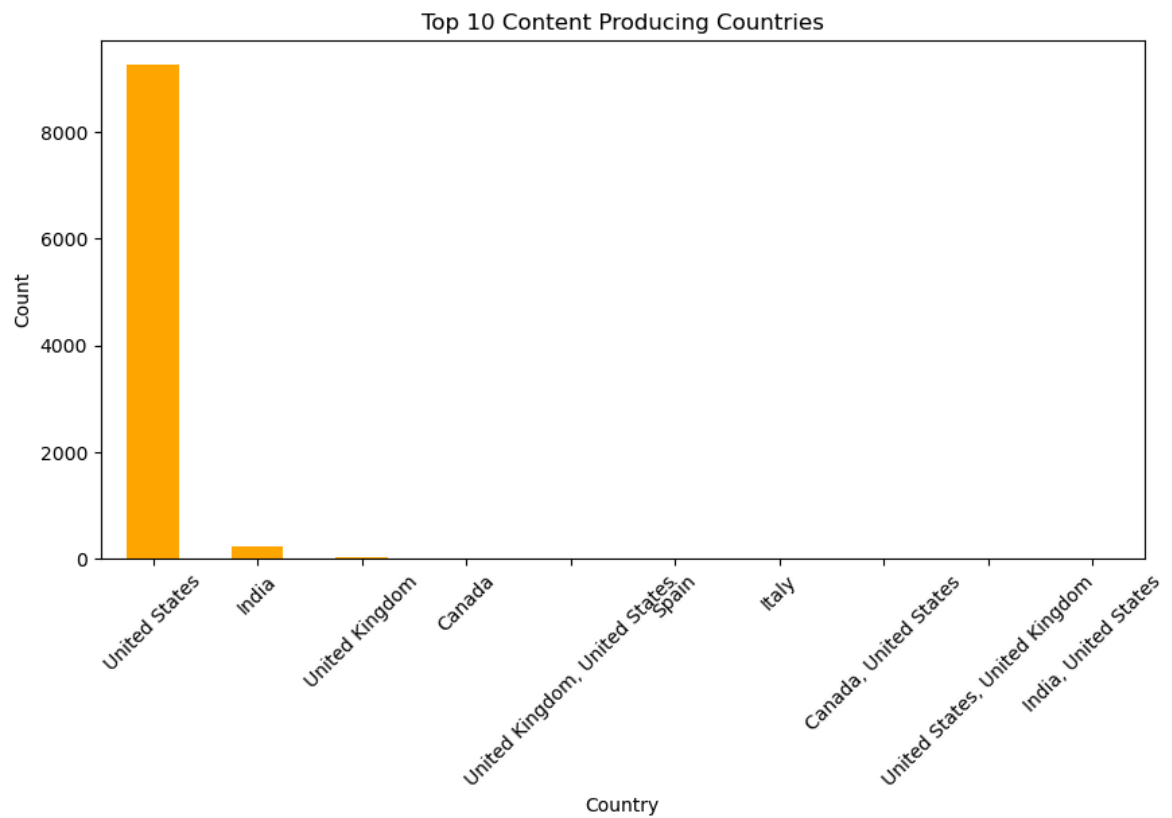

```
In [86]:   # Correlation heatmap
           plt.figure(figsize=(8,6))
           sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm', linewidths=
           plt.title('Correlation Heatmap')
           plt.show()
```

Correlation Heatmap

Completed