In [3]:

```
import os
import pandas as pd
print(os.getcwd())
```

```
/Users/ayeshasiddiqha
```

In [5]:

```
os.chdir("/Users/ayeshasiddiqha/Downloads")
```

In [8]:

```
netflix_data = pd.read_csv("netflix.csv")
print(netflix_data.head())
print("--------------------------------------------")
print(netflix_data.columns.to_list())
```

```
  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...  
--------------------------------------------
['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
'release_year', 'rating', 'duration', 'listed_in', 'description']
```

In [9]:

```
print(netflix_data.shape)
print(netflix_data.info())
```

```
(8807, 12)
<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
None
```

In [10]:

```
# Check for missing values
print(netflix_data.isnull().sum())

# Summary of the dataset
print(netflix_data.describe(include='all'))
```

```
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
       show_id   type   title       director                cast  \
count     8807   8807    8807           6173                7982
unique    8807      2    8807           4528                7692
top      s8807  Movie  Zubaan  Rajiv Chilaka  David Attenborough
freq         1   6131       1             19                  19
mean       NaN    NaN     NaN            NaN                 NaN
std        NaN    NaN     NaN            NaN                 NaN
min        NaN    NaN     NaN            NaN                 NaN
25%        NaN    NaN     NaN            NaN                 NaN
50%        NaN    NaN     NaN            NaN                 NaN
75%        NaN    NaN     NaN            NaN                 NaN
max        NaN    NaN     NaN            NaN                 NaN

              country       date_added  release_year rating  duration  \
count            7976             8797   8807.000000   8803      8804
unique            748             1767           NaN     17       220
top     United States  January 1, 2020           NaN  TV-MA  1 Season
freq             2818              109           NaN   3207      1793
mean              NaN              NaN   2014.180198    NaN       NaN
std               NaN              NaN      8.819312    NaN       NaN
min               NaN              NaN   1925.000000    NaN       NaN
25%               NaN              NaN   2013.000000    NaN       NaN
50%               NaN              NaN   2017.000000    NaN       NaN
75%               NaN              NaN   2019.000000    NaN       NaN
max               NaN              NaN   2021.000000    NaN       NaN

                           listed_in  \
count                           8807
unique                           514
top     Dramas, International Movies
freq                             362
mean                             NaN
std                              NaN
min                              NaN
25%                              NaN
50%                              NaN
75%                              NaN
max                              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
```

In [12]:

```
# Convert columns to 'category'
categorical_cols = ['type', 'rating', 'listed_in']
for col in categorical_cols:
    netflix_data[col] = netflix_data[col].astype('category')
```

In [14]:

```
netflix_data.head(3)
```

Out[14]:

	show_id	type	title	director	cast	country	date_added	release_y
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2

In [15]:

```
print(netflix_data.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   category
 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   category
 9   duration      8804 non-null   object  
 10  listed_in     8807 non-null   category
 11  description   8807 non-null   object  
dtypes: category(3), int64(1), object(8)
memory usage: 674.8+ KB
None
```

In [17]:

```
# Movies vs. TV Shows
print(netflix_data['type'].value_counts())
```

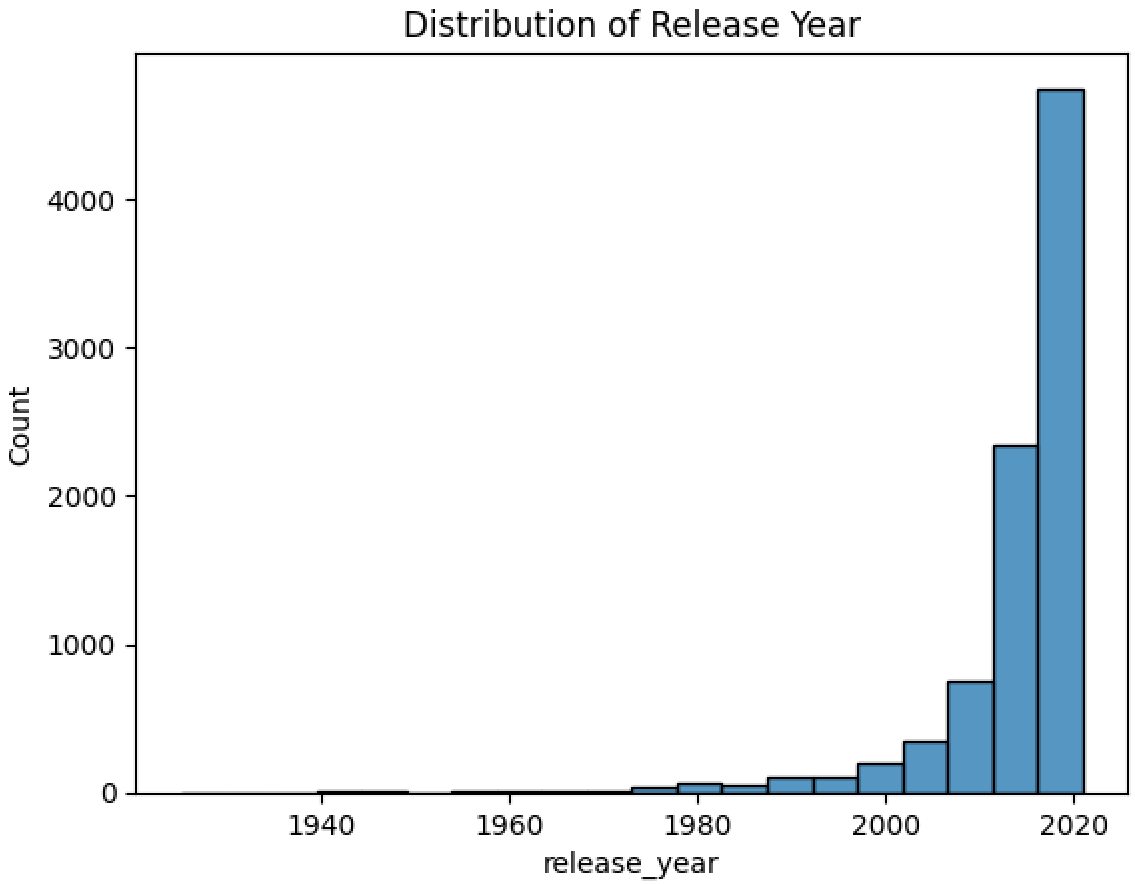
```
# Top genres
print(netflix_data['listed_in'].value_counts().head(10))

# Top countries
print(netflix_data['country'].value_counts().head(10))
```

```
type
Movie      6131
TV Show    2676
Name: count, dtype: int64
listed_in
Dramas, International Movies                        362
Documentaries                                       359
Stand-Up Comedy                                     334
Comedies, Dramas, International Movies              274
Dramas, Independent Movies, International Movies    252
Kids' TV                                            220
Children & Family Movies                            215
Children & Family Movies, Comedies                  201
Documentaries, International Movies                 186
Dramas, International Movies, Romantic Movies       180
Name: count, dtype: int64
country
United States     2818
India              972
United Kingdom     419
Japan              245
South Korea        199
Canada             181
Spain              145
France             124
Mexico             110
Egypt              106
Name: count, dtype: int64
```

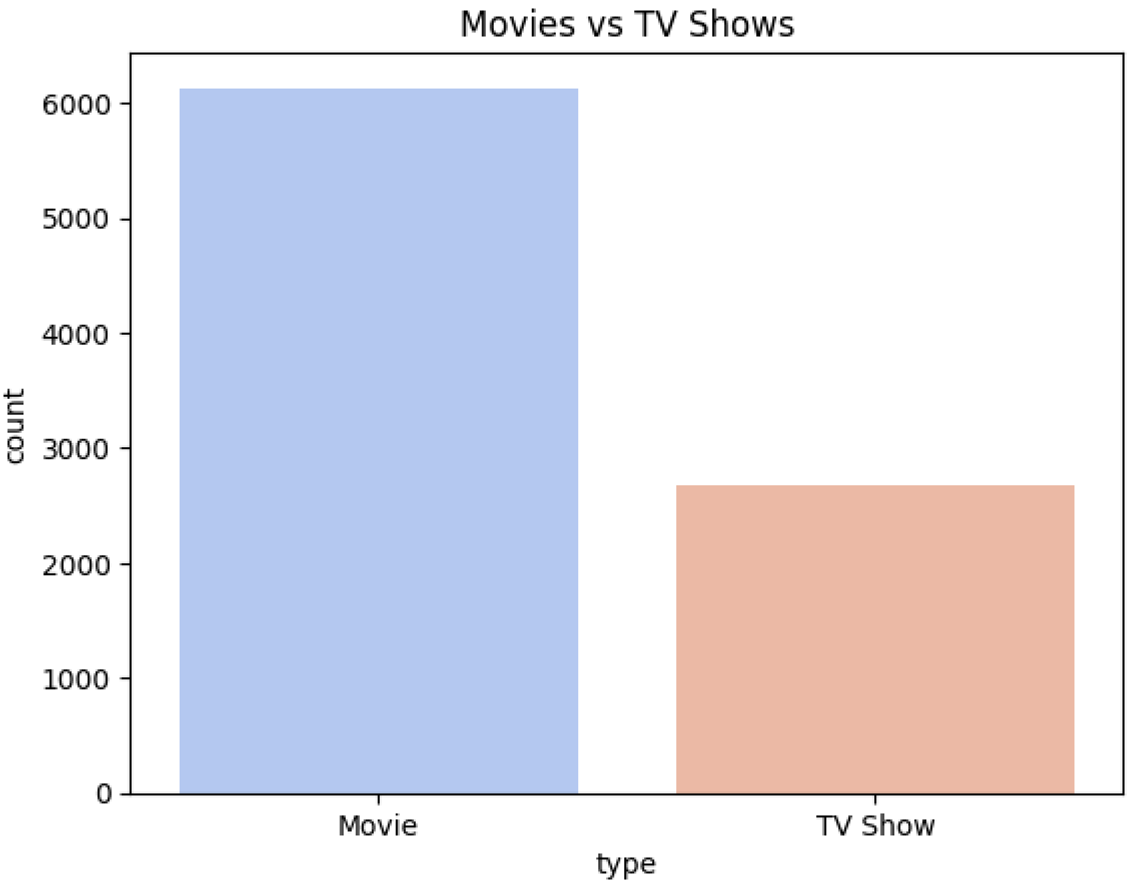
In [19]:

```
import matplotlib.pyplot as plt
import seaborn as sns

# Distribution of Release Year
sns.histplot(netflix_data['release_year'], kde=False, bins=20)
plt.title('Distribution of Release Year')
plt.show()

# Countplot for Type
sns.countplot(data=netflix_data, x='type', palette='coolwarm')
plt.title('Movies vs TV Shows')
plt.show()
```

```
Matplotlib is building the font cache; this may take a moment.
```



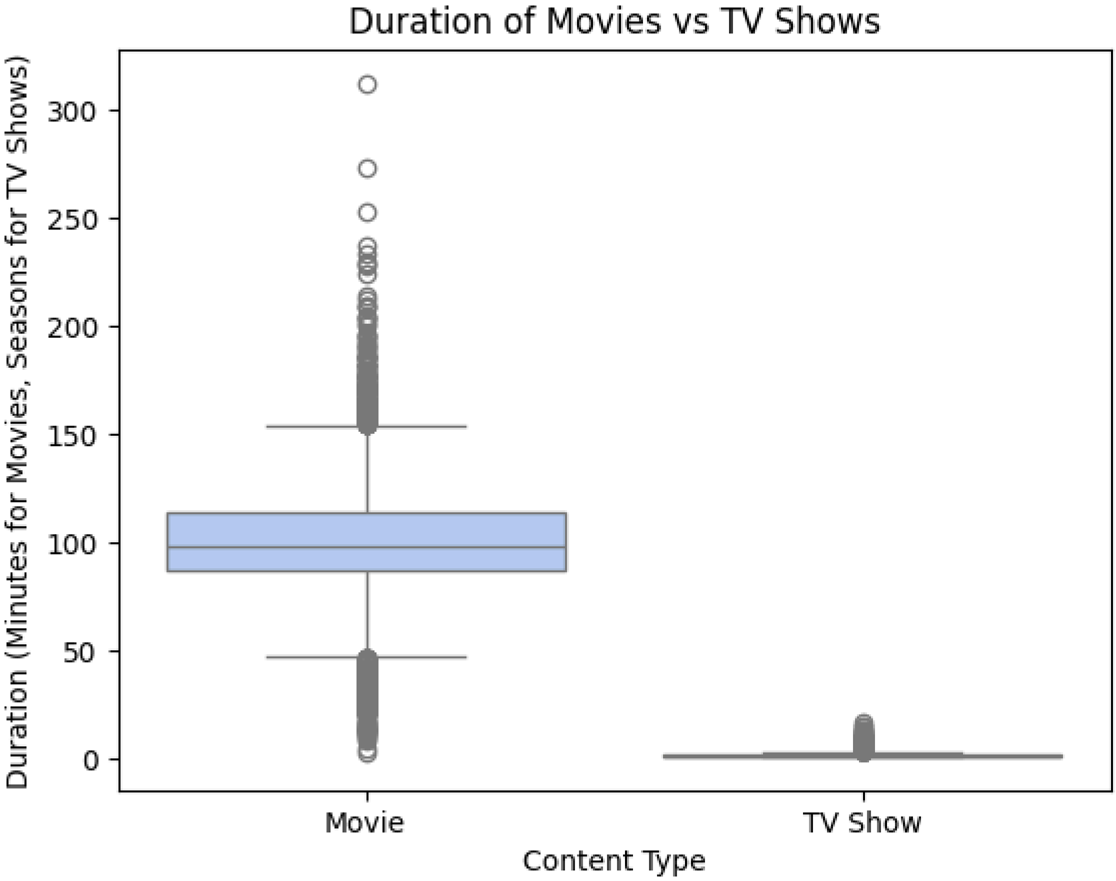
```
/var/folders/c8/n9hz87597yz68gbmzzks3v_00000gn/T/ipykernel_39921/203462252
6.py:10: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be remove
d in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for
the same effect.

  sns.countplot(data=netflix_data, x='type', palette='coolwarm')
```



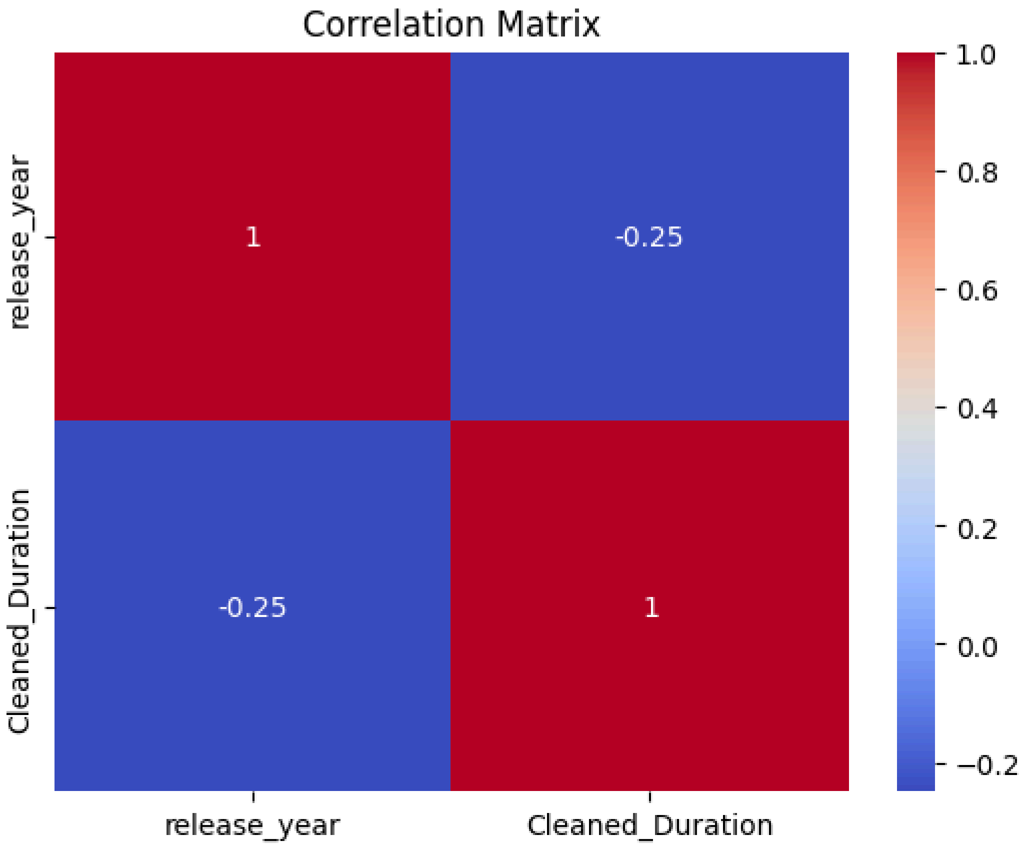
In [21]:

```
import numpy as np

# Function to clean and preprocess the 'Duration' column
def clean_duration(row):
    if isinstance(row, str):
        if 'min' in row:  # For Movies
            return int(row.replace(' min', ''))
        elif 'Season' in row:  # For TV Shows
            return int(row.split(' ')[0])  # Extract the number of season
    return np.nan  # Handle any unexpected format

# Apply cleaning function to create a new column
netflix_data['Cleaned_Duration'] = netflix_data['duration'].apply(clean_d

# Drop rows with missing or invalid durations
netflix_data = netflix_data.dropna(subset=['Cleaned_Duration'])

# Boxplot of Cleaned Duration by Type
sns.boxplot(data=netflix_data, x='type', y='Cleaned_Duration', palette='c
plt.title('Duration of Movies vs TV Shows')
plt.ylabel('Duration (Minutes for Movies, Seasons for TV Shows)')
plt.xlabel('Content Type')
plt.show()
```

```
/var/folders/c8/n9hz87597yz68gbmzzks3v_00000gn/T/ipykernel_39921/144192200
7.py:19: FutureWarning: 

Passing `palette` without assigning `hue` is deprecated and will be remove
d in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for
the same effect.

  sns.boxplot(data=netflix_data, x='type', y='Cleaned_Duration', palette
='coolwarm')
```



In [22]:
```
# Correlation cleaned duration Heatmap
numerical_cols = ['release_year', 'Cleaned_Duration']
correlation_matrix = netflix_data[numerical_cols].corr()

sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

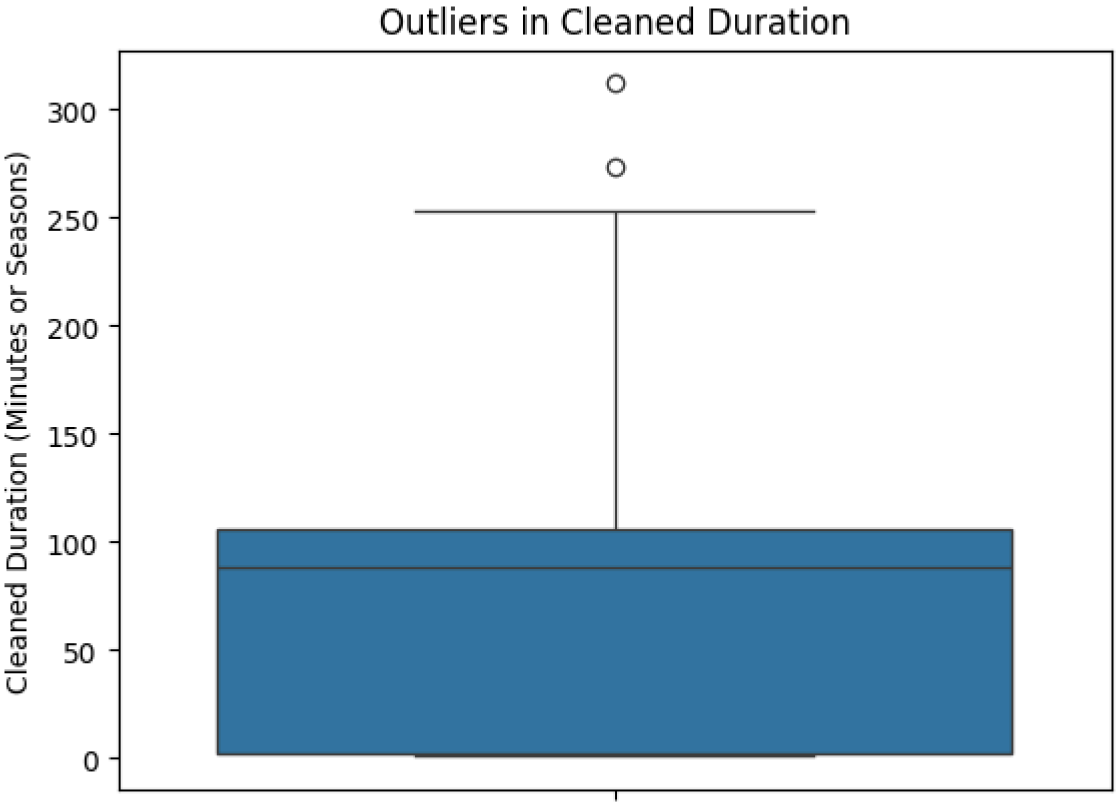


In [26]:

```
# Missing Values
missing_cols = netflix_data.columns[netflix_data.isnull().any()]
print(netflix_data[missing_cols].isnull().sum())

# Fill missing Country with 'Unknown'
netflix_data['country'] = netflix_data['country'].fillna('Unknown')

# Check for Outliers in Duration
sns.boxplot(data=netflix_data, y='Cleaned_Duration')
plt.title('Outliers in Cleaned Duration')
plt.ylabel('Cleaned Duration (Minutes or Seasons)')
plt.show()
```

```
director      2634
cast           825
date_added      10
rating           4
dtype: int64
```

```
/var/folders/c8/n9hz87597yz68gbmzzks3v_00000gn/T/ipykernel_39921/248797419
8.py:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  netflix_data['country'] = netflix_data['country'].fillna('Unknown')
```


Outliers in Cleaned Duration
Cleaned Duration (Minutes or Seasons)
0
50
100
150
200
250
300

```
In [ ]:
```