Importing Libraries

In [1]:
```python
# Importing Necessary Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Reading CSV File

In [2]:
```python
# Reading CSV File
df = pd.read_csv(r"C:\Users\Ajith\Desktop\Datasets\netflix_titles.csv")
df.head(5)
```

Out[2]:

	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...

Info of the DataFrame

In [3]:
```python
# Info of the DataFrame
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
```

Null values in the DataFrame

In [4]:
```python
# Null values in the DataFrame
df.isna().sum()
```

Out[4]:
```
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
```

Columns in the DataFrame

In [5]:
```python
# Columns in the DataFrame
df.columns
```

Out[5]:
```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

Shape of the DataFrame

In [6]:
```python
# Shape of the DataFrame
df.shape
```

Out[6]:
```
(8807, 12)
```

Dropping Unnecessary Column

In [7]:
```python
# Dropping show_id and description column
df.drop(["show_id","description"], axis=1, inplace=True)
```

Null values in Country Column

In [8]:
```python
# Most frequent value in the column
mode_country = df["country"].mode()[0]
mode_country
```

Out[8]:
```
'United States'
```

In [9]:
```python
# Verifying the absence of Null values
df["country"] = df["country"].fillna(mode_country)
```

Null values in date_added Column

In [10]:
```python
mode_date = df["date_added"].mode()[0]
mode_date
```

Out[10]:
```
'January 1, 2020'
```

In [11]:
```python
# Verifying the absence of Null values
df["date_added"] = df["date_added"].fillna(mode_date)
```

Null values in Rating Column

In [12]:
```python
# Most frequent value in the column
mode_rating = df["rating"].mode()[0]
mode_rating
```

Out[12]:
```
'TV-MA'
```

In [13]:
```python
# Filling in the Null values with most frequent value
df["rating"] = df["rating"].fillna(mode_rating)
```

Null values in Duration Column

In [14]:
```python
# Most frequent value in the Column
mode_duration = df["duration"].mode()[0]
mode_duration
```

Out[14]:
```
'1 Season'
```

In [15]:
```python
# Filling in the Null values with most frequent value
df["duration"] = df["duration"].fillna(mode_duration)
```

Null values in Director Column

In [16]:
```python
# We don't have much data to fill Null values in this column
# So, We are filling in the Null values with "Not Mentioned"
df["director"] = df["director"].fillna("Not Mentioned")
```

Null values in Cast Column

In [17]:
```python
# We don't have much data to fill Null values in this column
# So, We are filling in the Null values with "Not Mentioned"
df["cast"] = df["cast"].fillna("Not Mentioned")
```

Null values in the DataFrame

In [18]:
```python
# Final Null values in the DataFrame
df.isna().sum()
```

Out[18]:
```
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
```

Removing unwanted values from Rating Column

In [19]:
```python
# Unique values in the rating column
# We can see "74min, 84min, 66min" in the rating column which makes no sense
# So, We have to remove them
df["rating"].unique()
```

Out[19]:
```
array(['PG-13', 'TV-MA', 'PG', 'TV-14', 'TV-PG', 'TV-Y', 'TV-Y7', 'R',
       'TV-G', 'G', 'NC-17', '74 min', '84 min', '66 min', 'NR',
       'TV-Y7-FV', 'UR'], dtype=object)
```

In [20]:
```python
# Modifying DataFrame with rows which don't include "min" in rating column
df = df[~df["rating"].str.contains("min")]
```

In [21]:
```python
# Unique values in the rating column after the transformation
df["rating"].unique()
```

Out[21]:
```
array(['PG-13', 'TV-MA', 'PG', 'TV-14', 'TV-PG', 'TV-Y', 'TV-Y7', 'R',
       'TV-G', 'G', 'NC-17', 'NR', 'TV-Y7-FV', 'UR'], dtype=object)
```

Converting date_added Column to DateTime Format

In [22]:
```python
# Removing (,) commas appearing in the date_added column
# With (,) It will be hard for Pandas to convert the Date to DateTime Format
df["date_added"] = df["date_added"].str.replace(",","")
```

In [23]:
```python
# Changing the date_added column to DateTime Format
df["date_added"] = pd.to_datetime(df["date_added"])
```

Extracting Year, Month and Date from date_added Column

In [24]:
```python
# Extracting Year from date_added column
df["year"] = df["date_added"].dt.year
```

In [25]:
```python
# Extracting Month Names from date_added column
df["month"] = df["date_added"].dt.month_name()
```

In [26]:
```python
# Extracting Date from date_added column
df["date"] = df["date_added"].dt.day
```

Dropping Unnecessary Column

In [27]:
```python
# Dropping date_added, listed_in and cast column
df.drop(["date_added","listed_in","cast"], axis=1, inplace=True)
```

Final DataFrame

In [28]:
```python
# Final DataFrame
df.head(5)
```

Out[28]:

	type	title	director	country	release_year	rating	duration	year	month	date
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90 min	2021	September	25
1	TV Show	Blood & Water	Not Mentioned	South Africa	2021	TV-MA	2 Seasons	2021	September	24
2	TV Show	Ganglands	Julien Leclercq	United States	2021	TV-MA	1 Season	2021	September	24
3	TV Show	Jailbirds New Orleans	Not Mentioned	United States	2021	TV-MA	1 Season	2021	September	24
4	TV Show	Kota Factory	Not Mentioned	India	2021	TV-MA	2 Seasons	2021	September	24

In [30]:
```python
df
```

Out[30]:

	type	title	director	country	release_year	rating	duration	year	month	date
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90 min	2021	September	25
1	TV Show	Blood & Water	Not Mentioned	South Africa	2021	TV-MA	2 Seasons	2021	September	24
2	TV Show	Ganglands	Julien Leclercq	United States	2021	TV-MA	1 Season	2021	September	24
3	TV Show	Jailbirds New Orleans	Not Mentioned	United States	2021	TV-MA	1 Season	2021	September	24
4	TV Show	Kota Factory	Not Mentioned	India	2021	TV-MA	2 Seasons	2021	September	24
...	...	...	...	...	...	...	...	...	...	...
8802	Movie	Zodiac	David Fincher	United States	2007	R	158 min	2019	November	20
8803	TV Show	Zombie Dumb	Not Mentioned	United States	2018	TV-Y7	2 Seasons	2019	July	1
8804	Movie	Zombieland	Ruben Fleischer	United States	2009	R	88 min	2019	November	1
8805	Movie	Zoom	Peter Hewitt	United States	2006	PG	88 min	2020	January	11
8806	Movie	Zubaan	Mozez Singh	India	2015	TV-14	111 min	2019	March	2

8804 rows × 10 columns

All the Movies uploaded on Netflix

In [31]:
```python
# DataFrame with type is equal to "Movies"
movies_df = df[df["type"]=="Movie"]
movies_df.head(5)
```

Out[31]:

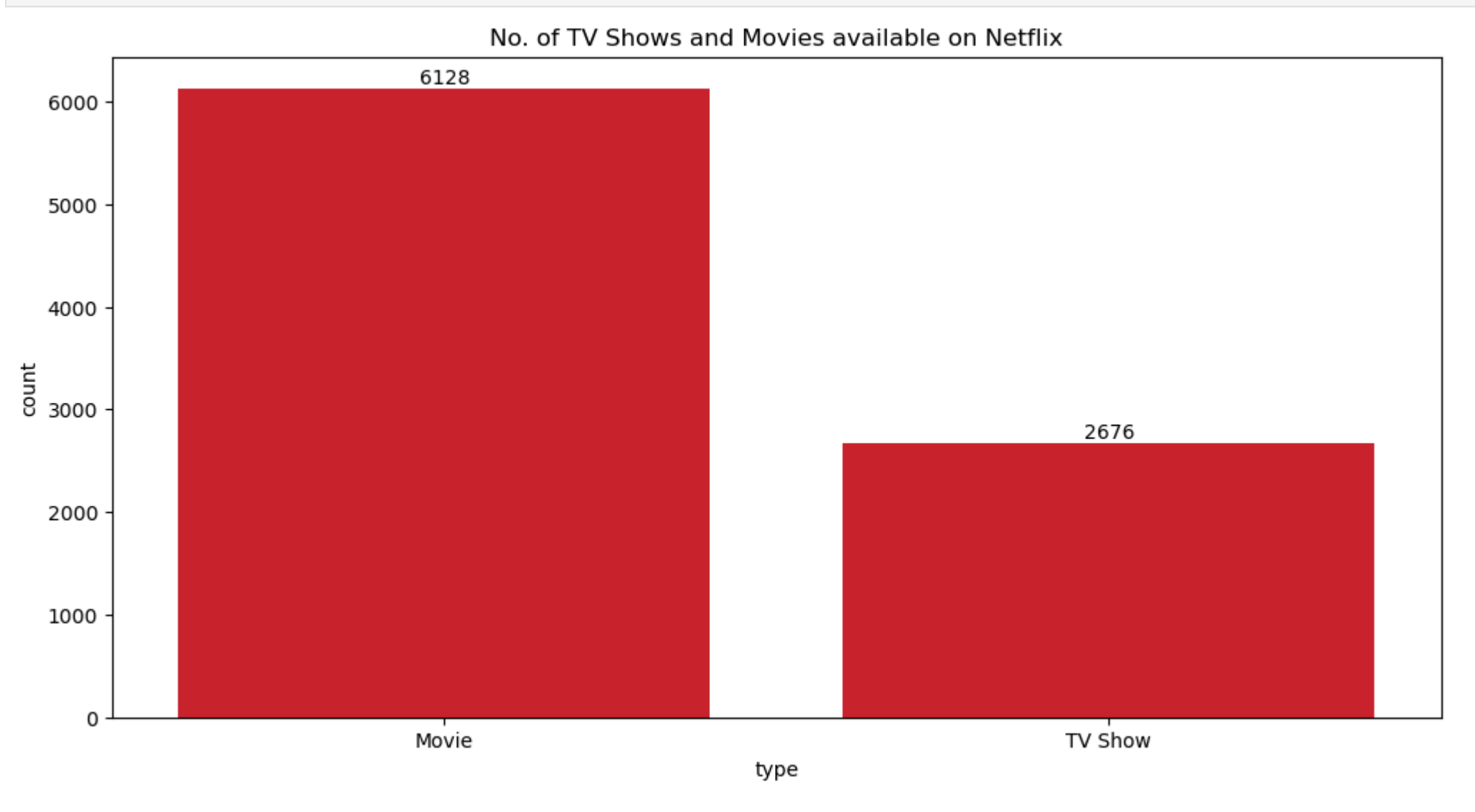
	type	title	director	country	release_year	rating	duration	year	month	date
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90 min	2021	September	25
6	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	United States	2021	PG	91 min	2021	September	24
7	Movie	Sankofa	Haile Gerima	United States, Ghana, Burkina Faso, United Kin...	1993	TV-MA	125 min	2021	September	24
9	Movie	The Starling	Theodore Melfi	United States	2021	PG-13	104 min	2021	September	24
12	Movie	Je Suis Karl	Christian Schwochow	Germany, Czech Republic	2021	TV-MA	127 min	2021	September	23

No. of TV Shows and Movies available on Netflix

In [32]:
```python
# Count of unique values in the type column
df["type"].value_counts()
```

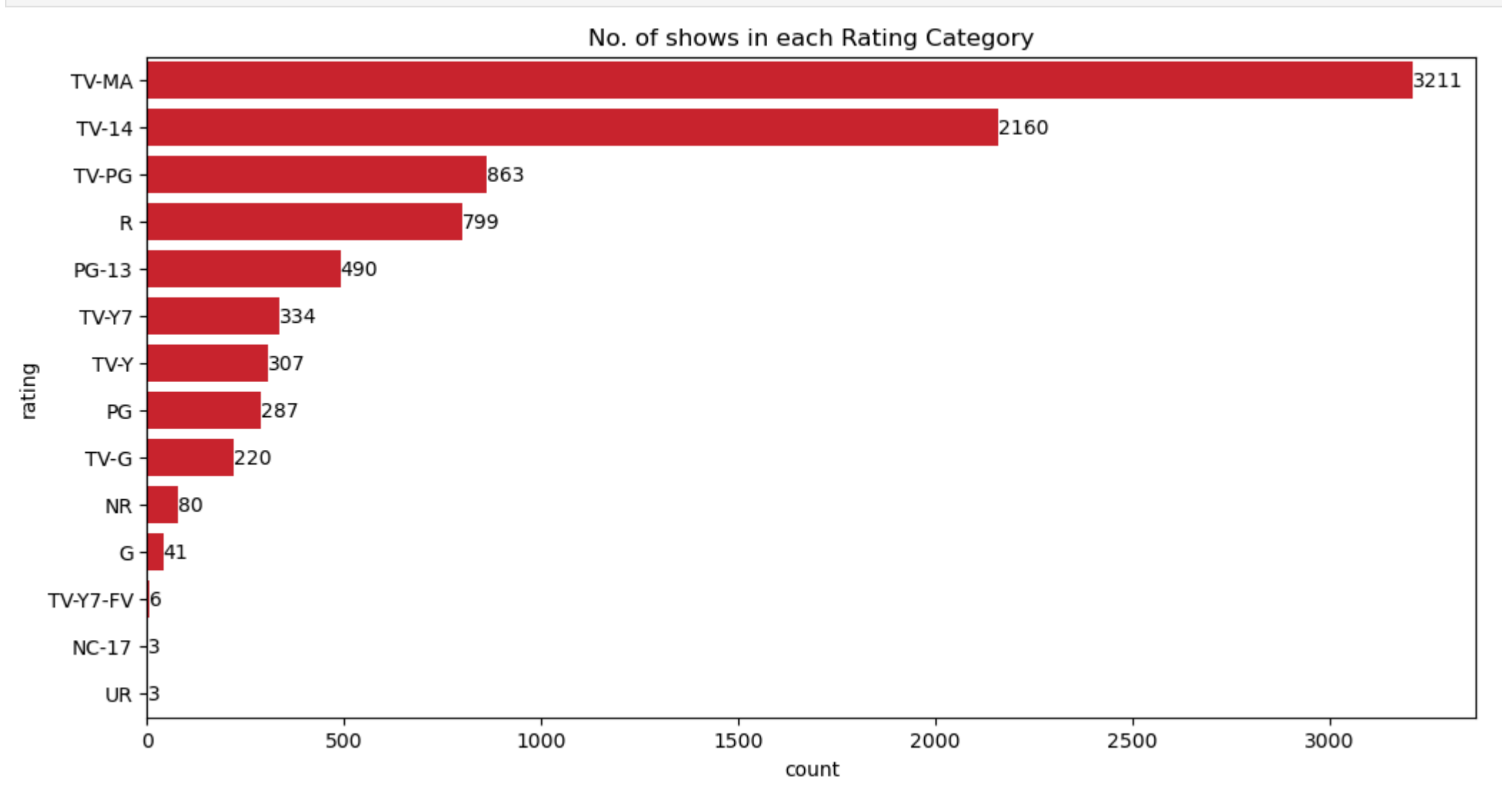
Out[32]:
```
Movie      6128
TV Show    2676
Name: type, dtype: int64
```

In [33]:
```python
fig, ax = plt.subplots(figsize=(12, 6))
ax = sns.countplot(data=df, x="type", color="#E50914", order=df["type"].value_counts().index)
ax.set_title("No. of TV Shows and Movies available on Netflix")
ax.bar_label(ax.containers[0], fontsize=10)
plt.show()
```



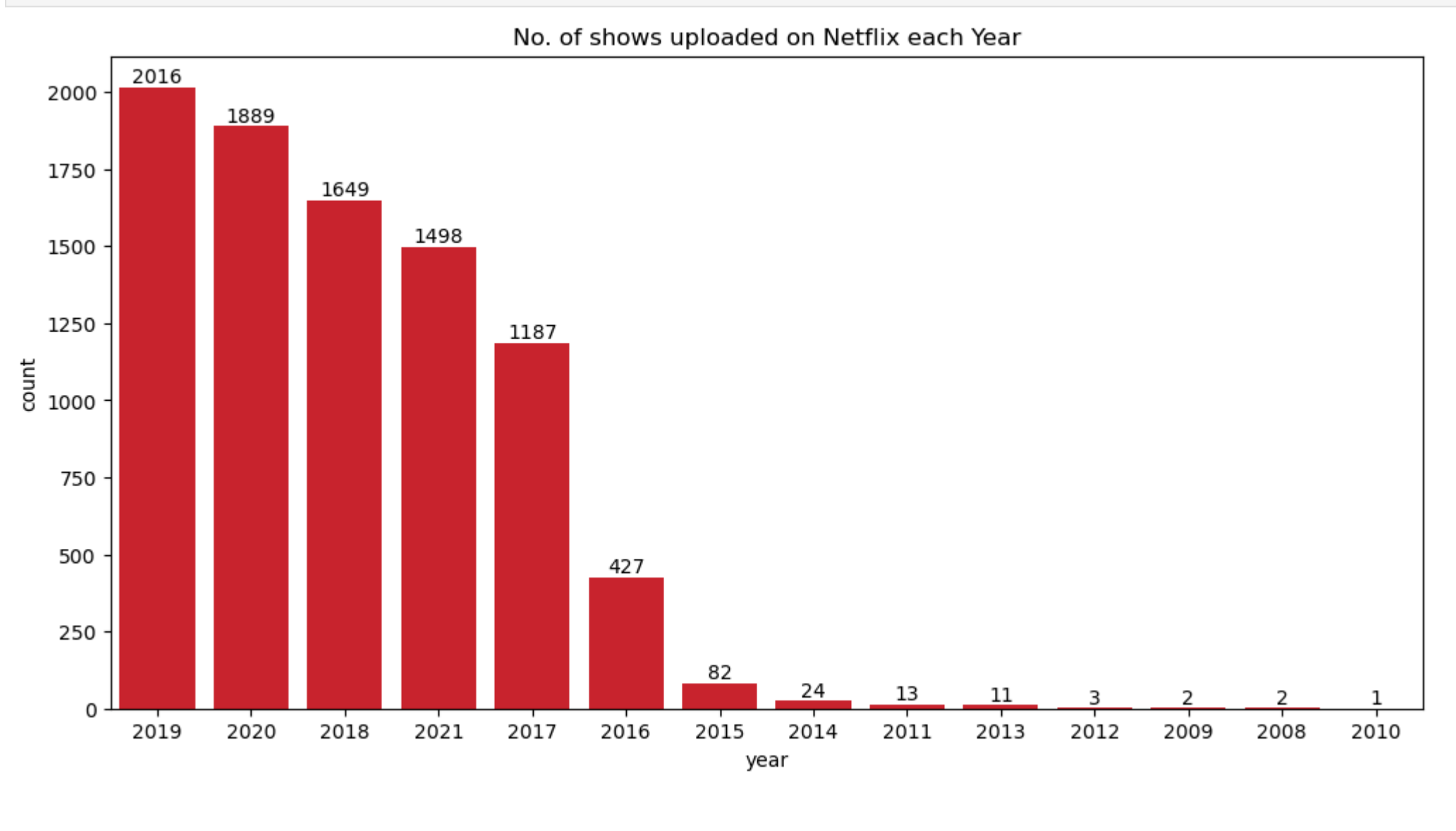
No. of shows in each Rating Category

In [34]:
```python
fig, ax = plt.subplots(figsize=(12, 6))
ax = sns.countplot(data=df, y="rating", order=df["rating"].value_counts().index, color="#E50914")
ax.set_title("No. of shows in each Rating Category")
ax.bar_label(ax.containers[0], fontsize=10)
plt.show()
```



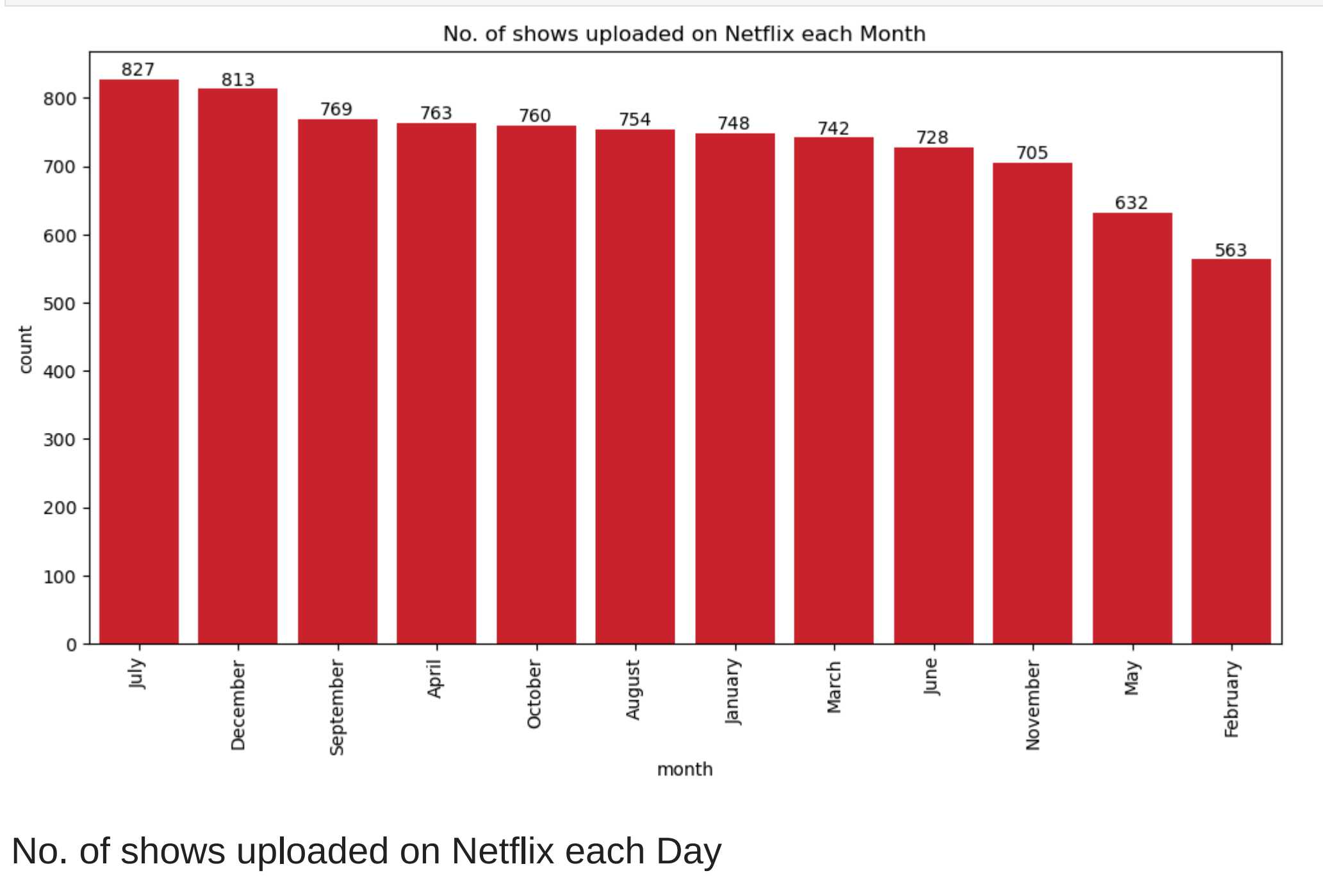
No. of shows uploaded on Netflix each Year

In [35]:
```python
fig, ax = plt.subplots(figsize=(12, 6))
ax = sns.countplot(data=df, x="year", order=df["year"].value_counts().index, color="#E50914")
ax.set_title("No. of shows uploaded on Netflix each Year")
ax.bar_label(ax.containers[0], fontsize=10)
plt.show()
```



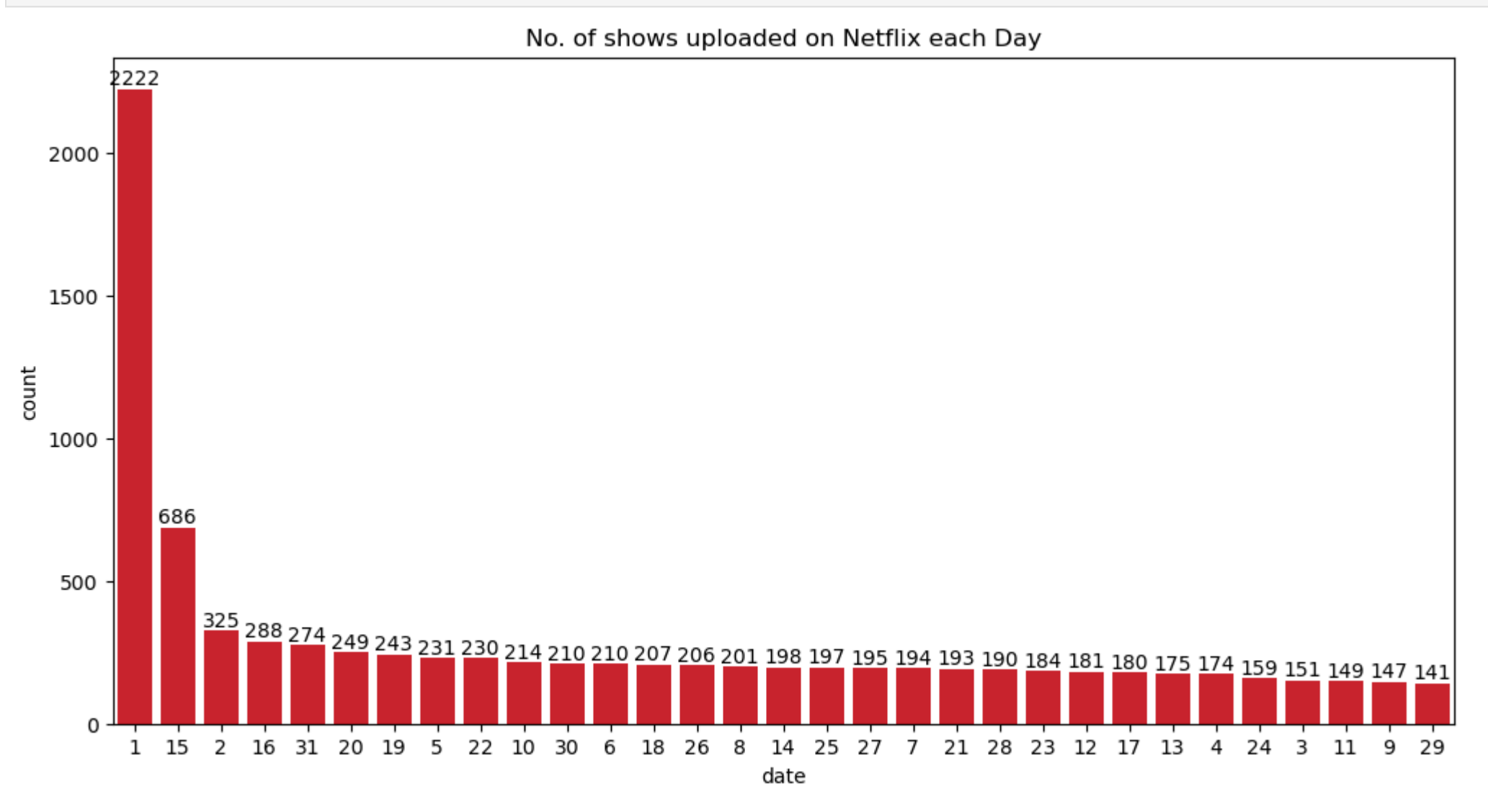
No. of shows uploaded on Netflix each Month

In [36]:
```python
fig, ax = plt.subplots(figsize=(12, 6))
ax = sns.countplot(data=df, x="month", order=df["month"].value_counts().index, color="#E50914")
ax.set_title("No. of shows uploaded on Netflix each Month")
ax.tick_params(axis='x', rotation=90)
ax.bar_label(ax.containers[0], fontsize=10)
plt.show()
```



No. of shows uploaded on Netflix each Day

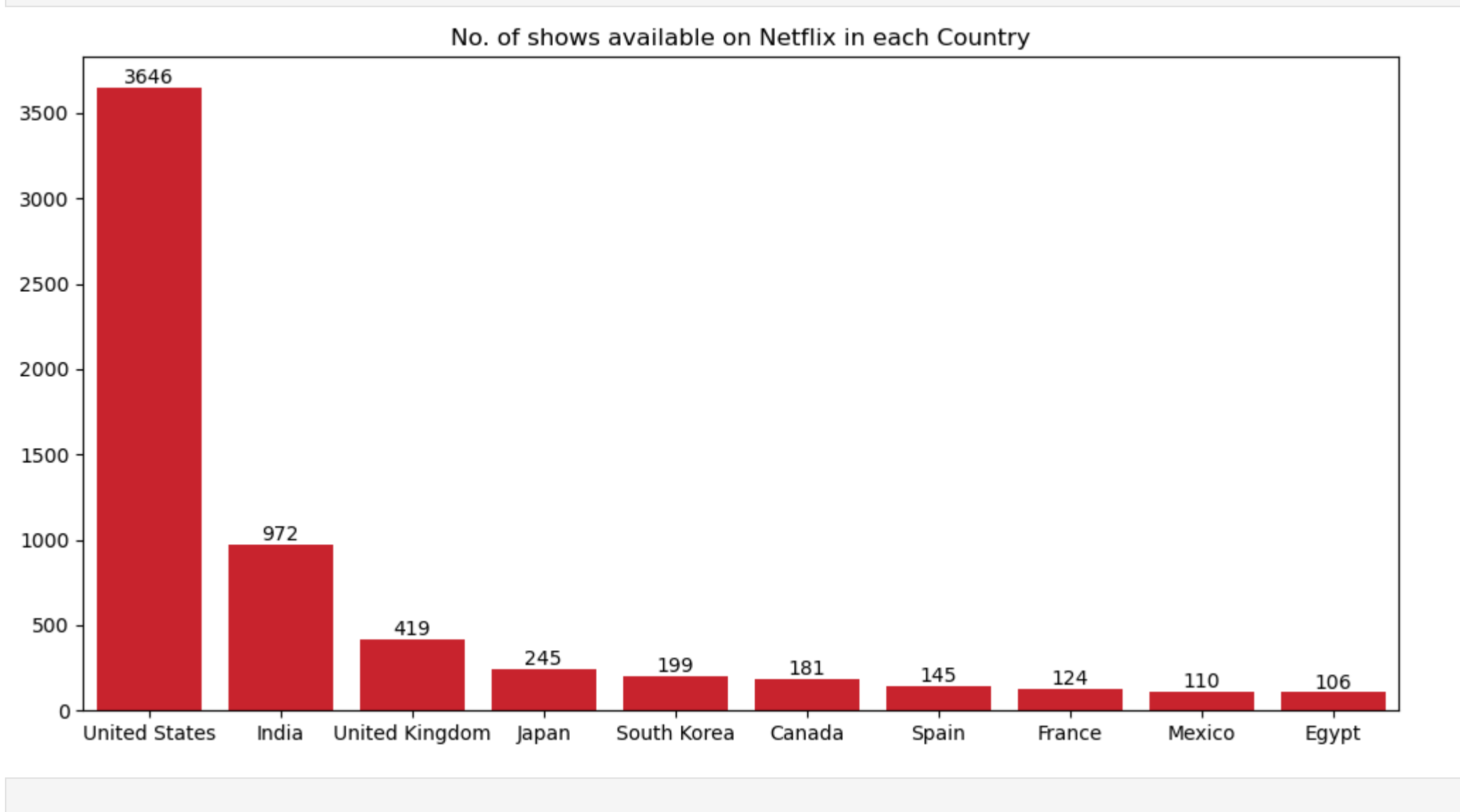
In [37]:
```python
fig, ax = plt.subplots(figsize=(12, 6))
ax = sns.countplot(data=df, x="date", order=df["date"].value_counts().index, color="#E50914")
ax.set_title("No. of shows uploaded on Netflix each Day")
ax.bar_label(ax.containers[0], fontsize=10)
plt.show()
```



No. of shows available on Netflix in each Country

In [38]:
```python
country_counts = df["country"].value_counts().head(10)
```

In [39]:
```python
fig, ax = plt.subplots(figsize=(12, 6))
ax = sns.barplot(x=country_counts.index, y=country_counts.values, color="#E50914")
ax.set_title("No. of shows available on Netflix in each Country")
ax.bar_label(ax.containers[0], fontsize=10)
plt.show()
```



In [ ]: