NAME:MAHESH DARAKHE

BATCH:BEGINER NOVEMBER 2022

Business Problem

Analyze the data and generate insights that could help Netflix in deciding which type of shows/movies to produce and how they can grow the business in different countries

```
import numpy as np
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
from pickle import FALSE
from wordcloud import WordCloud
df=pd.read_csv('https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv')
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
```

```
df.shape
```

```
(8807, 12)
```

```
df.nunique()
```

```
show_id         8807
type               2
title           8807
director        4528
cast            7692
country          748
date_added      1767
release_year      74
rating            17
duration         220
listed_in        514
description     8775
dtype: int64
```

Insight: From the results we can se there are 8807 rows and 12 columns uniique values in each column

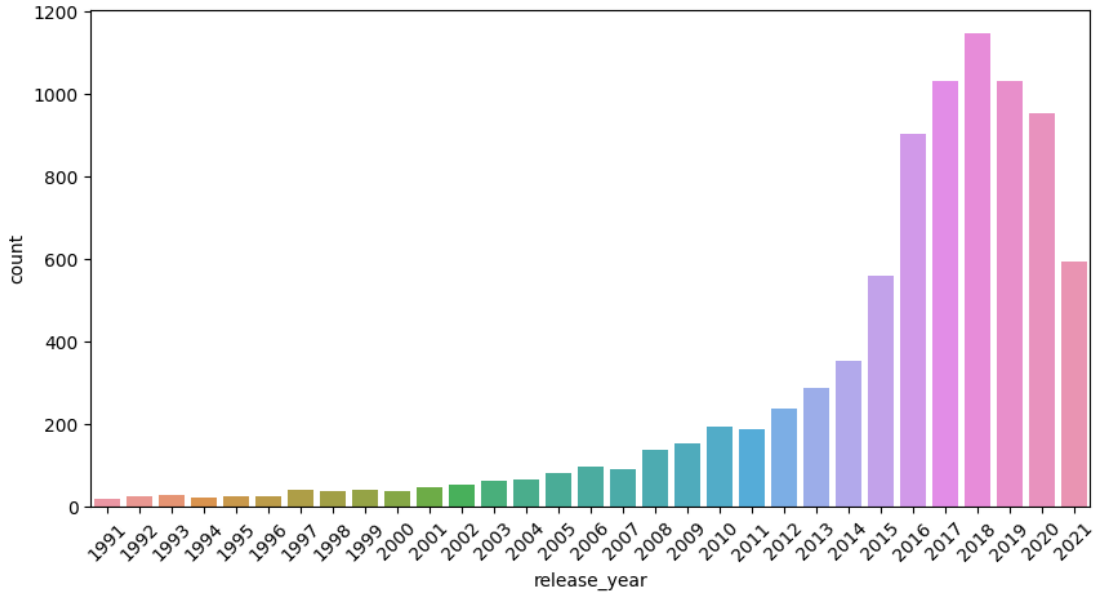
How has the number of movies released per year changed over the last 20-30 years?

```
df['release_year'].value_counts().head(30)
```

```
2018    1147
2017    1032
2019    1030
2020     953
2016     902
2021     592
2015     560
2014     352
2013     288
2012     237
2010     194
2011     185
2009     152
2008     136
2006      96
2007      88
2005      80
2004      64
2003      61
2002      51
2001      45
1999      39
1997      38
```

```
2000      37
1998      36
1993      28
1995      25
1996      24
1992      23
1994      22
Name: release_year, dtype: int64
```

```
plt.figure(figsize=(10,5))
sns.countplot(df[df['release_year'] > 1990],x = 'release_year')
plt.xticks(rotation=45)
plt.show()
```



Insight: from the above visualization we can see the release of number of movies and tv shows has chnged over the years it has peaked in year 2018and has been increasing from 2000 gradully till 2018

Comparison of tv shows vs. movies.

> Top 10 countries with their number of Movies produced

```
df['country'] = df['country'].astype(str)
df['country'] = df['country'].replace("['nan']", np.nan)
df['country'] = df['country'].str.split(', ')
df_exploded_countries= df.explode('country', ignore_index=True)
df_exploded_countries[df_exploded_countries['type']=='Movie'].groupby(by='country')['title'].count().sort_values(ascending= False).head(10)
```

```
country
United States     2751
India              962
United Kingdom     532
nan                440
Canada             319
France             303
Germany            182
Spain              171
Japan              119
China              114
Name: title, dtype: int64
```

Insight:United states has the highest no. of movies and tv shows released

> Top 10 countries with their number of Movies released

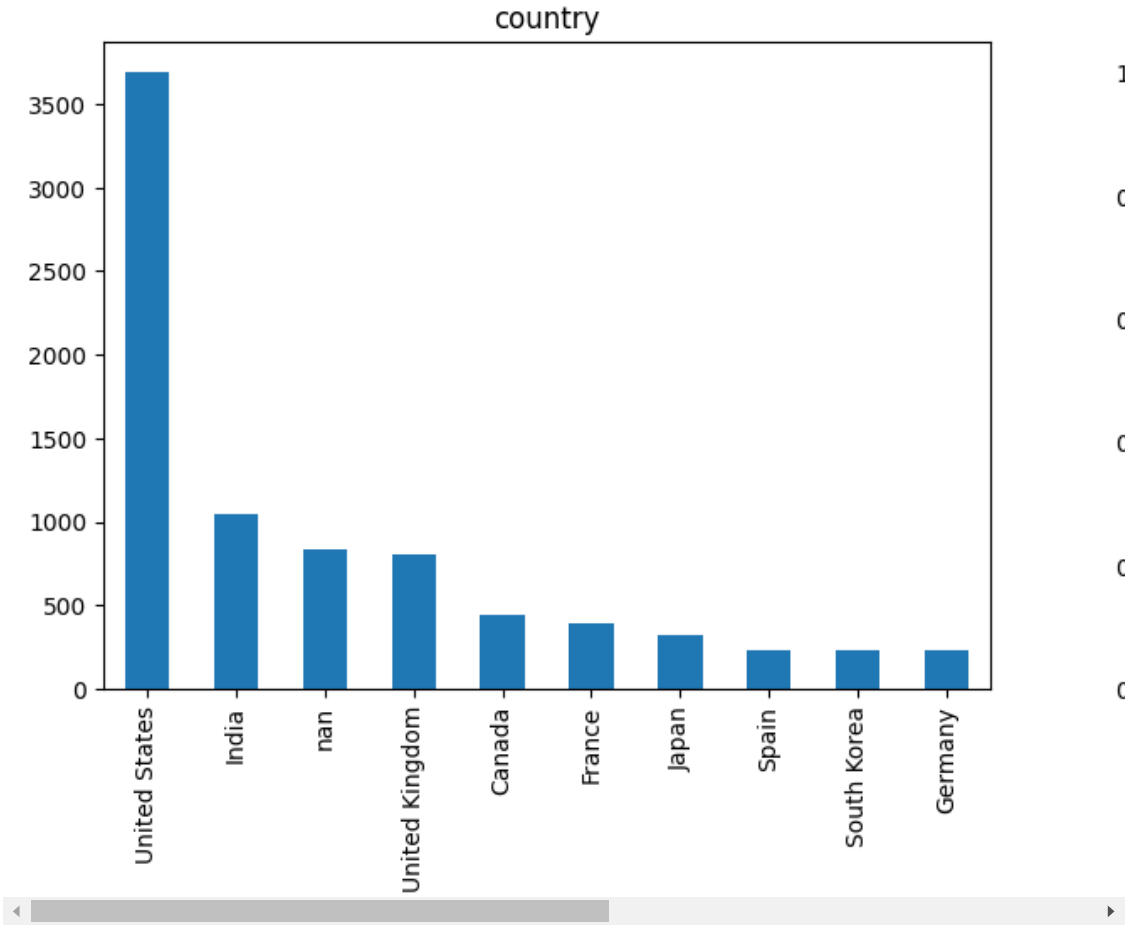
```
df_exploded_countries[df_exploded_countries['type']=='TV Show'].groupby(by='country')['title'].count().sort_values(ascending= False).head(10)
```

```
country
United States     938
nan               391
United Kingdom    272
Japan             199
South Korea       170
Canada            126
France             90
India              84
Taiwan             70
```

```
Australia          66
Name: title, dtype: int64
```

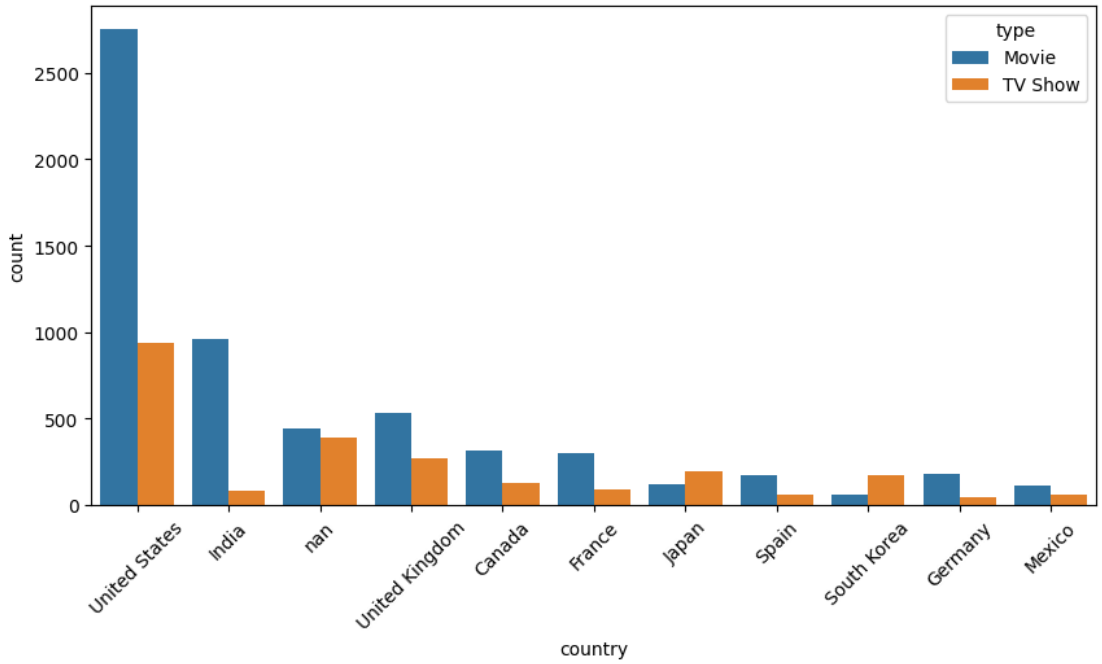
Insight:United states has the highest no. of movies andreleased

```
plt.figure(figsize=(15,5))
plt.subplot(1,2,1)
df_exploded_countries['country'].value_counts().head(10).plot(kind='bar')
plt.title('country')
plt.subplot(1,2,2)
df_exploded_countries['country'].value_counts(ascending=True).head(10).plot(kind='bar')
plt.title('country')
plt.show()
```



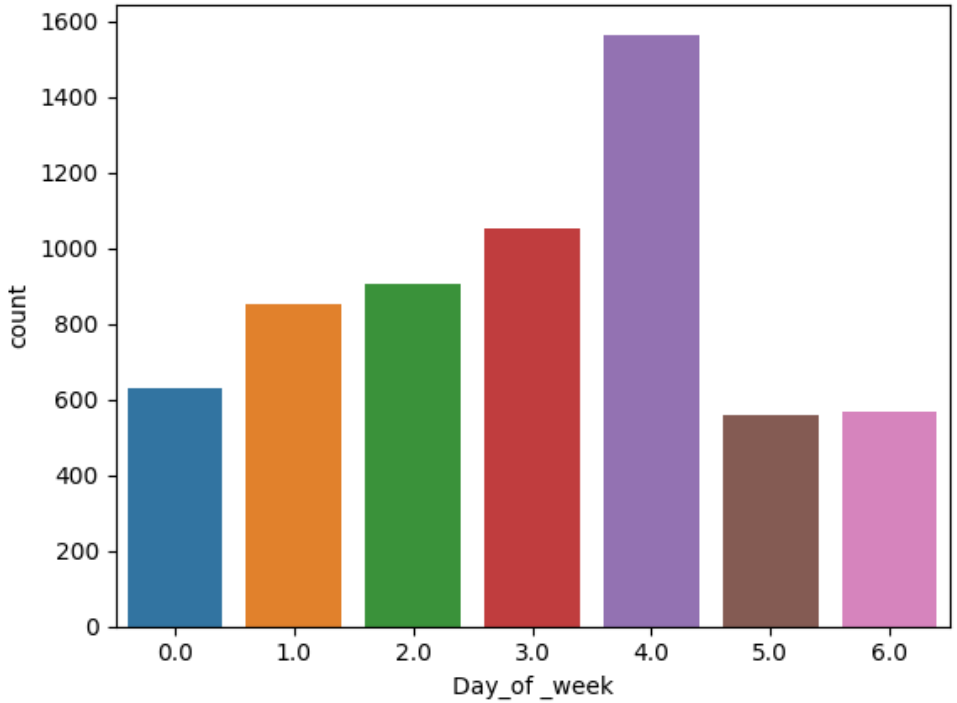
Insight:From the above representation we can conclude that the most no of movies and tv shows are produced in United States followed by india

```
plt.figure(figsize=(10,5))
sns.countplot(df_exploded_countries,x='country',hue='type',order=df_exploded_countries.country.value_counts().iloc[:11].index)
plt.xticks(rotation=45)
plt.show()
```



> What is the best Week to launch a Movie?

```
df['date_added'] = pd.to_datetime(df['date_added'],format='%B %d, %Y',errors='coerce')
df['Day_of _week'] = df['date_added'].dt.weekday
df[df['type']=='Movie'].groupby(by='Day_of _week')['title'].count()
sns.countplot(df[df['type']=='Movie'],x='Day_of _week')
plt.show()
```

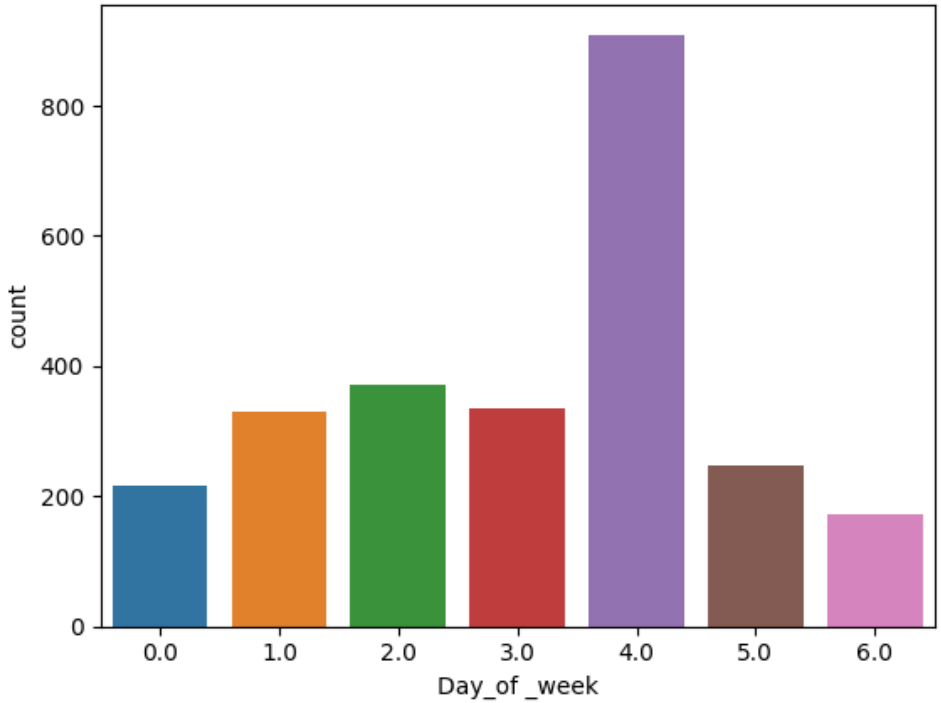


Insight:As we can see 5th dy of weekk that is friday has the most no of releases considering that it would be better if new releases are scheduled on friday

Recomendation:its best to add movies on friday

> What is the best Week to launch a TV Show?

```
df[df['type']=='TV Show'].groupby(by='Day_of _week')['title'].count()
sns.countplot(df[df['type']=='TV Show'],x='Day_of _week')
plt.show()
```



Insight:As we can see 5th dy of weekk that is friday has the most no of releases considering that it would be better if new releases are scheduled on friday

Recomendation:its best to add TV Shows on friday

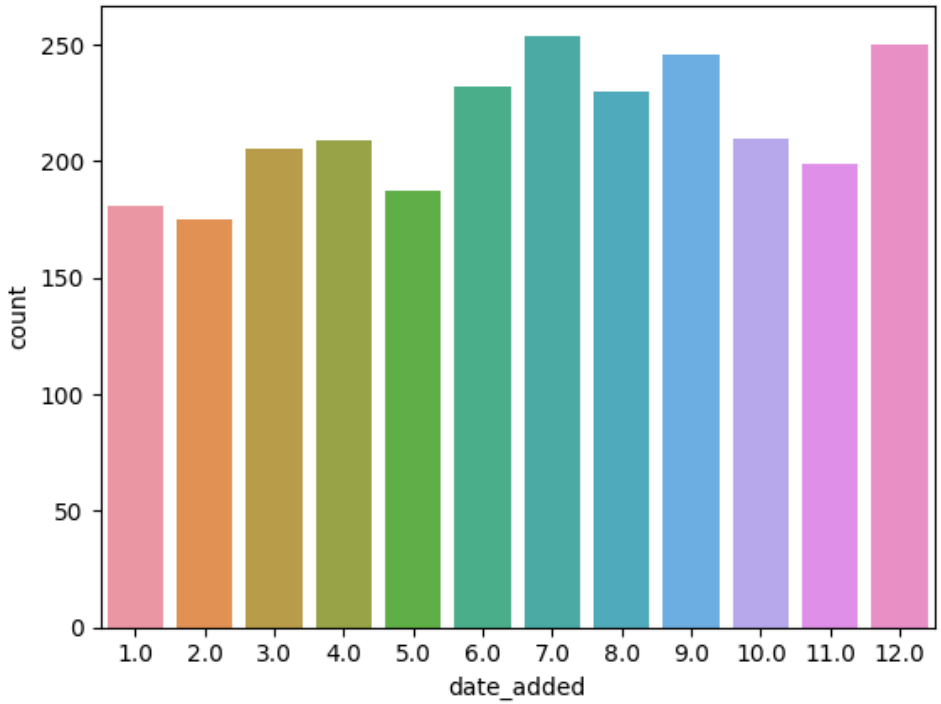
What is the best Month to launch a TV Show?

```
df[df['type']=='TV Show'].groupby(by=df['date_added'].dt.month)['title'].count()
```

```
date_added
1.0     181
2.0     175
3.0     205
4.0     209
5.0     187
6.0     232
7.0     254
8.0     230
```

```
9.0     246
10.0    210
11.0    199
12.0    250
Name: title, dtype: int64
```

```
sns.countplot(df[df['type']=='TV Show'],x=df[df['type']=='TV Show']['date_added'].dt.month)
plt.show()
```



Insight:best month to launch a tv show would be 7th month that is july

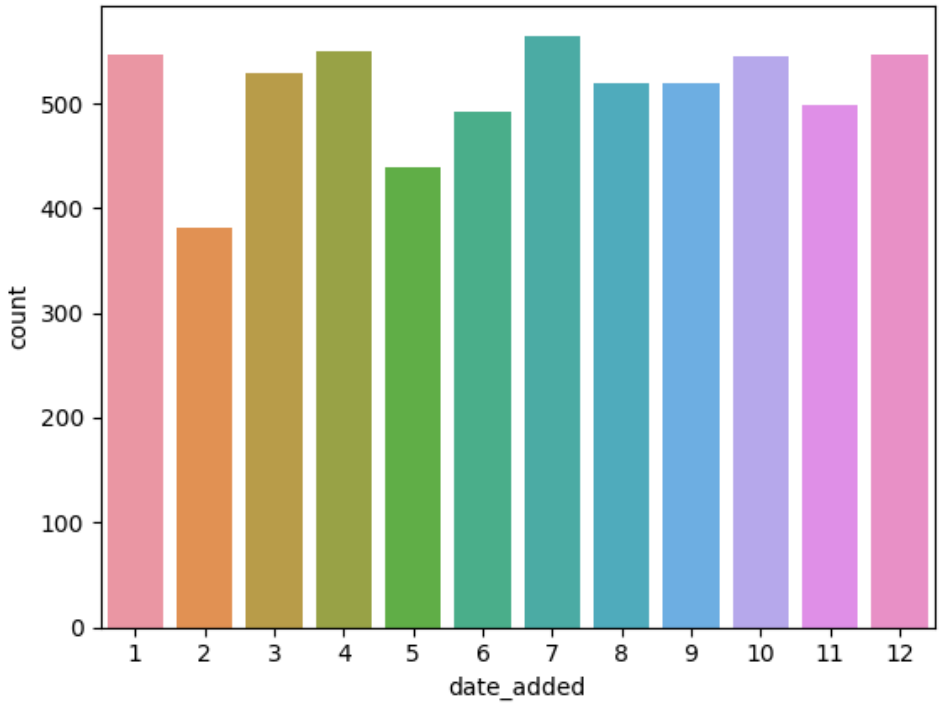
Recomendation:its best to add TV Shows in July

What is the best Month to launch a Movie?

```
df[df['type']=='Movie'].groupby(by=df['date_added'].dt.month)['title'].count()
```

```
date_added
1.0     546
2.0     382
3.0     529
4.0     550
5.0     439
6.0     492
7.0     565
8.0     519
9.0     519
10.0    545
11.0    498
12.0    547
Name: title, dtype: int64
```

```
sns.countplot(df[df['type']=='Movie'],x=df[df['type']=='Movie']['date_added'].dt.month)
plt.show()
```



Insight:best month to launch a Movie would be 7th month that is July

Recomendation:its best to add a Movie in July

> Analysis of actors/directors of different types of shows/movies

```
df['cast'] = df['cast'].astype(str)
df['cast'] = df['cast'].replace("['nan']", np.nan)
df['cast'] = df['cast'].str.split(', ')
df_exploded_cast= df.explode('cast', ignore_index=True)
df_exploded_cast.groupby(by='cast')['title'].count().sort_values(ascending= False).head(10)
```

```
cast
nan                825
Anupam Kher         43
Shah Rukh Khan      35
Julie Tejwani       33
Naseeruddin Shah    32
Takahiro Sakurai    32
Rupa Bhimani        31
Om Puri             30
Akshay Kumar        30
Yuki Kaji           29
Name: title, dtype: int64
```

Insight:Anupam kher was casted in most number of the Movies and tv shows combined

```
from pickle import FALSE
df.groupby(by='director')['title'].count().sort_values(ascending= False).head(10)
```

```
director
Rajiv Chilaka             19
Raúl Campos, Jan Suter    18
Suhas Kadav               16
Marcus Raboy              16
Jay Karas                 14
Cathy Garcia-Molina       13
Jay Chapman               12
Youssef Chahine           12
Martin Scorsese           12
Steven Spielberg          11
Name: title, dtype: int64
```

Insight:Rajiv Chilaka directed most number of the Movies and tv shows combined

Which genre movies are more popular or produced more

```
all_genres = ' '.join(df['listed_in'].dropna())
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(all_genres)
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```

Insight:TV Shows,International Movies,Dramas internatonal,action adventure are the most popular and most produced geners

Recomendation:adding new movies and tv shows from above geners is recomended.

Find After how many days the movie will be added to Netflix after the release of the movie

```
df['complete_release_year'] = pd.to_datetime(df['release_year'].astype(str) + '-01-01')
df['difference']=df['date_added']-df['complete_release_year']
print(df['difference'].mode())
```

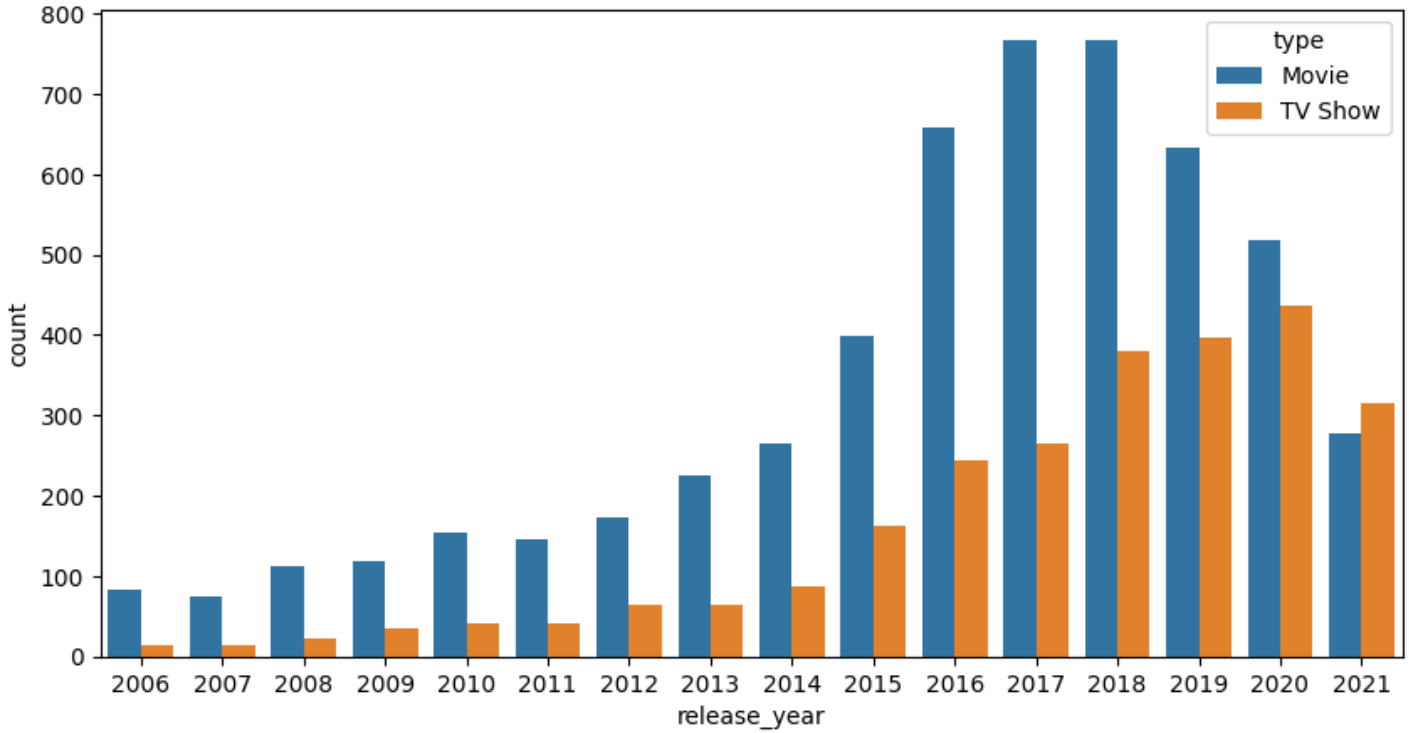
```
0   334 days
Name: difference, dtype: timedelta64[ns]
```

Insight:The median number of days after the release to be added on netflix is 334 so it takes 334 days on a median scale for a released movies to be added to netflix

Does Netflix has more focus on TV Shows than movies in recent years

```
plt.figure(figsize=(10,5))
sns.countplot(df[df['release_year']>2005],x='release_year',hue='type')
```

```
<Axes: xlabel='release_year', ylabel='count'>
```



Insight:The focus on Movies gradually increased after 2011 it peaked in 2017-2018 and the focus on TV shows has peaked in 2020 overall the focus on movies was more till 2018 and for tv shows it was in 2018 after that its gradually decreasinfg but the focus on movies has been drastically decreasing after 2018

✓ 1s completed at 2:11 AM