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

In [2]:
```
df = pd.read_csv(r'/Users/siyonabansal/Downloads/netflix.csv')
df.head()
```

Out[2]:

	show_id	type	title	director	cast	country	date_added	release_year	rating
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA

In [3]:
```
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
```

In [4]:
```
df.describe() #only for numeric columns
```

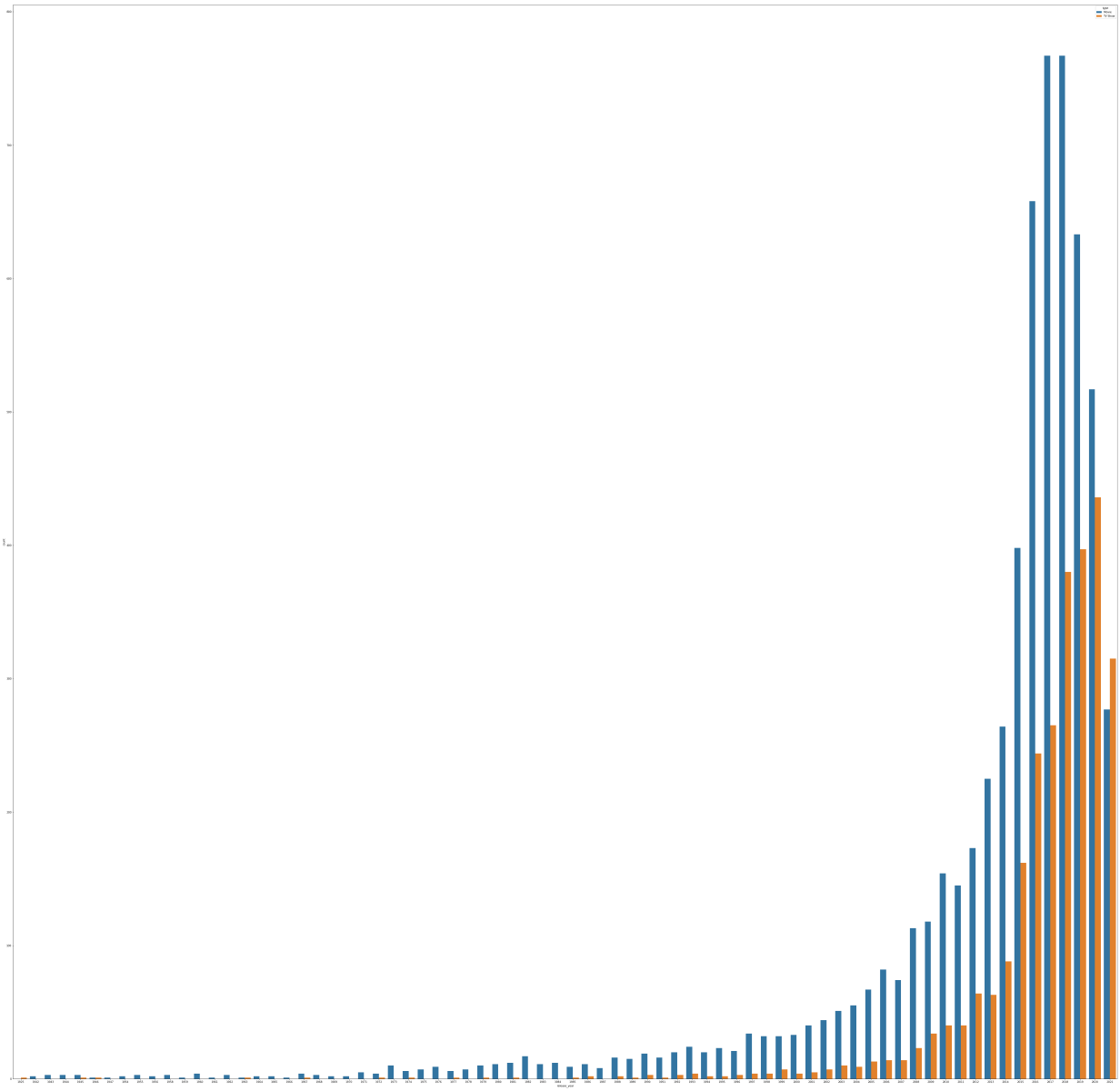
Out[4]:

	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

In [5]:
```
g1 = df.groupby(["type"]).size().reset_index(name='Number')
g1

plt.figure(figsize=(75,75))
sns.countplot('release_year', data=df, hue='type')
```

```
/Users/siyonabansal/opt/anaconda3/lib/python3.9/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[5]:
```
<AxesSubplot:xlabel='release_year', ylabel='count'>
```



we can see that only since 2021 the shift has gone towards tv shows.

In [6]:
```
fd = df[df['release_year']>2020]
fd.head()
```

Out[6]:

	show_id	type	title	director	cast	country	date_added	release_year	rating
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA

	show_id	type	title	director	cast	country	date_added	release_year	rating
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	NaN	September 24, 2021	2021	TV-MA

In [7]:
```
#Does Netflix has more focus on TV Shows than movies in recent years.

fd['type'].value_counts()
```

Out[7]:
```
TV Show    315
Movie      277
Name: type, dtype: int64
```

In [8]:
```
#therefore, we can conclude that ott platforms are shifting towards tv shows
```

In [9]:
```
#Understanding what content is available in different countries

udf = df.copy()
udf.drop(['show_id','description', 'title', 'director','cast','date_added','d
#removig enteries with no countries
udf.drop(udf[udf['country'].isnull()].index, axis=0, inplace=True)
udf
```

Out[9]:

	type	country	release_year	rating	listed_in
0	Movie	United States	2020	PG-13	Documentaries
1	TV Show	South Africa	2021	TV-MA	International TV Shows, TV Dramas, TV Mysteries
4	TV Show	India	2021	TV-MA	International TV Shows, Romantic TV Shows, TV ...
7	Movie	United States, Ghana, Burkina Faso, United Kin...	1993	TV-MA	Dramas, Independent Movies, International Movies
8	TV Show	United Kingdom	2021	TV-14	British TV Shows, Reality TV
...	...	...	...	...	...
8801	Movie	United Arab Emirates, Jordan	2015	TV-MA	Dramas, International Movies, Thrillers
8802	Movie	United States	2007	R	Cult Movies, Dramas, Thrillers
8804	Movie	United States	2009	R	Comedies, Horror Movies

	type	country	release_year	rating	listed_in
8805	Movie	United States	2006	PG	Children & Family Movies, Comedies
8806	Movie	India	2015	TV-14	Dramas, International Movies, Music & Musicals

7976 rows × 5 columns

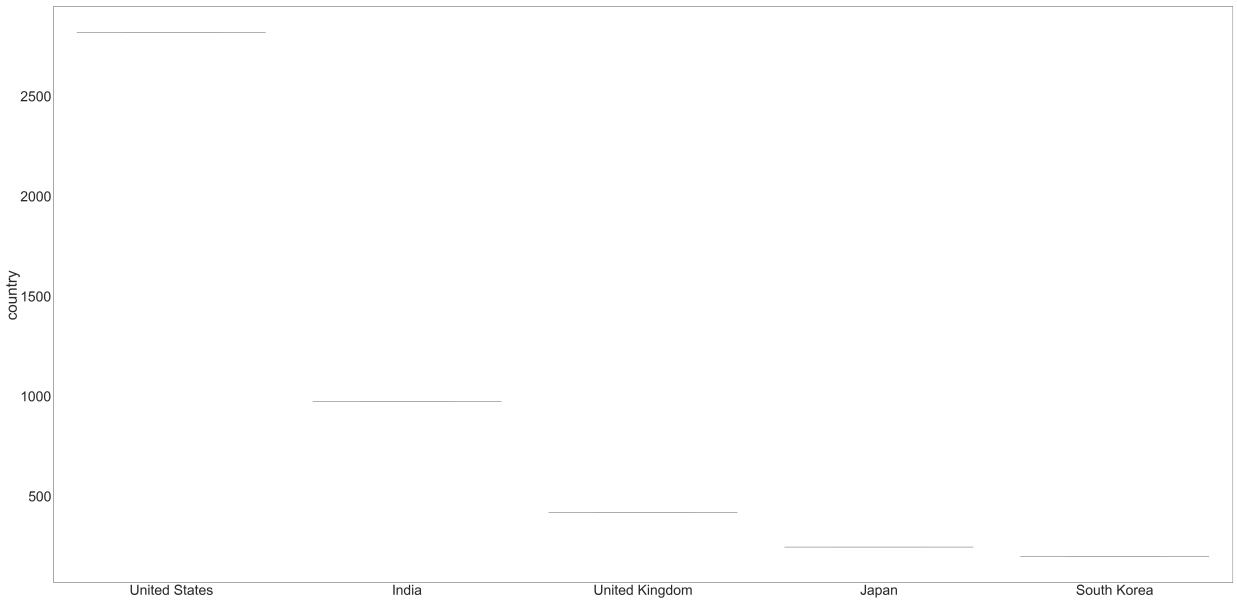
In [10]:
```
#most content is made in US
r= udf['country'].value_counts().head()
rdf = pd.DataFrame(r)
rdf
```

Out[10]:

	country
United States	2818
India	972
United Kingdom	419
Japan	245
South Korea	199

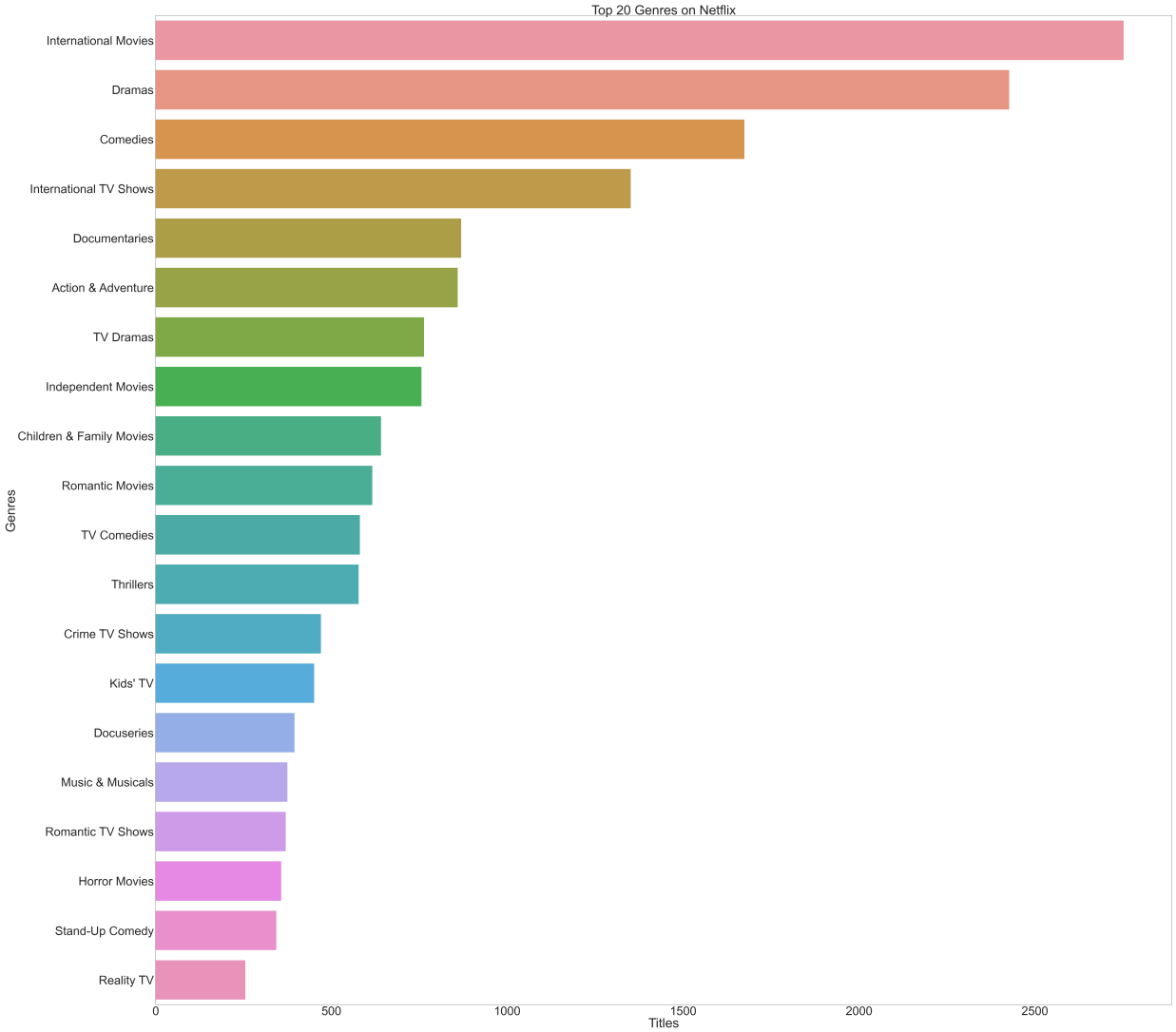
In [11]:
```
plt.figure(figsize=(100,50))

sns.set(style="white", font_scale=6)

sns.boxplot(x=rdf.index,y=rdf.country, data = rdf, palette="Reds")
```

Out[11]:
```
<AxesSubplot:ylabel='country'>
```



In [12]:
```
#Identifying similar content by matching text-based features
#top genres on netflix
genres = df.set_index('title').listed_in.str.split(', ', expand=True).stack()
plt.figure(figsize=(100,100))
g = sns.countplot(y = genres, order=genres.value_counts().index[:20])
plt.title('Top 20 Genres on Netflix')
plt.xlabel('Titles')
```

```
plt.ylabel('Genres')
plt.show()
```



In [13]:

```
genres_country = df.set_index('country').listed_in.str.split(', ', expand=True
gc= pd.DataFrame(genres_country)
gc.reset_index()
gc.rename(columns={0 :'genre'}, inplace=True)
gc
```

Out[13]:

country	genre
United States	Documentaries
South Africa	International TV Shows
South Africa	TV Dramas
South Africa	TV Mysteries
NaN	Crime TV Shows
...	...
United States	Children & Family Movies
United States	Comedies
India	Dramas
India	International Movies

country	genre
India	Music & Musicals

19323 rows × 1 columns

In [22]:
```
gc.groupby(['country'])['genre'].count()
```

Out[22]:
```
country
, France, Algeria                                      3
, South Korea                                          2
Argentina                                            139
Argentina, Brazil, France, Poland, Germany, Denmark    3
Argentina, Chile                                       5
                                                    ...
Venezuela                                              2
Venezuela, Colombia                                    2
Vietnam                                               20
West Germany                                           2
Zimbabwe                                               3
Name: genre, Length: 748, dtype: int64
```

In [31]:
```
#Network analysis of Actors / Directors and find interesting insights
df['director'].fillna('no director')
df['cast'].fillna('no info on cast')

df.head()
```

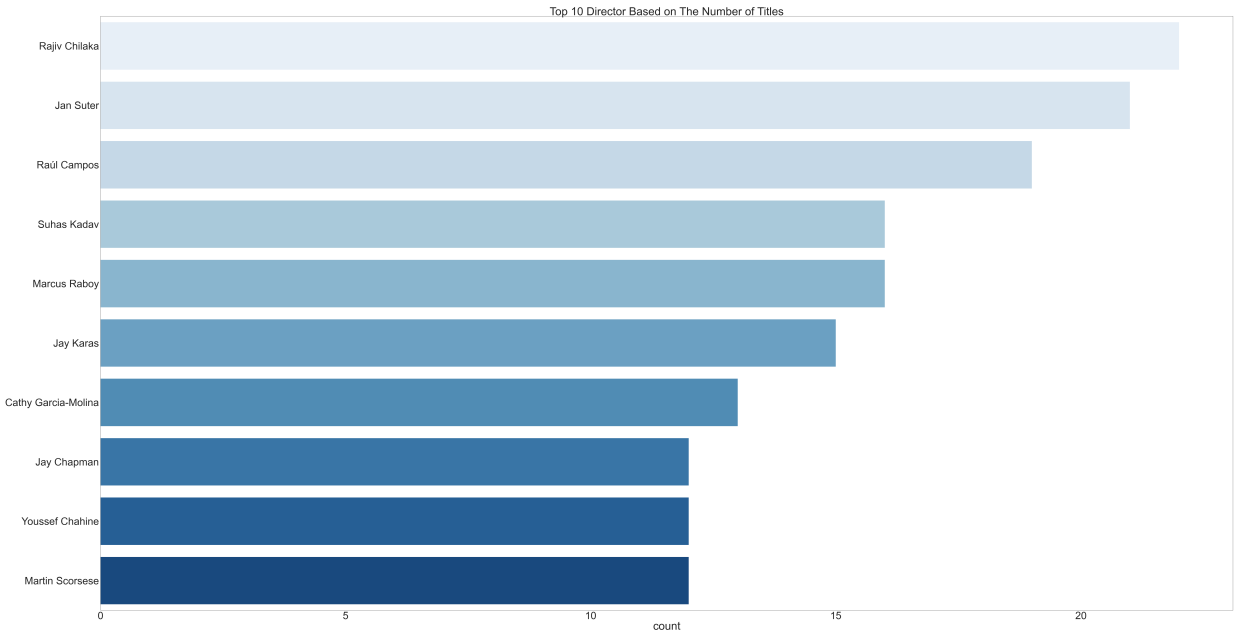
Out[31]:

	show_id	type	title	director	cast	country	date_added	release_year	rating
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan	India	September 24, 2021	2021	TV-MA

show_id	type	title	director	cast	country	date_added	release_year	rating
				Raj, Alam K...				

In [35]:

```
directors = df[df.director != 'No Director'].set_index('title').director.str.
plt.figure(figsize=(130,70))
plt.title('Top 10 Director Based on The Number of Titles')
sns.countplot(y = directors, order=directors.value_counts().index[:10], palet
plt.show()
```



In [ ]: