

In [15]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [16]:

```
df=pd.read_csv('https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/df')
```

Out[16]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	Seasons
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, ...	NaN	September 24, 2021	2021	TV-MA	Season

In [17]:

df.head(5)

Out[17]:

	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 [18]:

df.shape

Out[18]:

(8807, 12)

Analysis : In this dataset there are 8807 rows and 12 columns and this dataset segregated later as movie show and tv shows using head function we can get starting first five rows replace NaN values with unknown and fill with missing values.

In [19]:

Handling missing values

In [20]:

```
df['director']=df['director'].fillna('Unknown director')
```

In [21]:

```
df['country']=df['country'].fillna('Unknown country')
```

In [22]:

```
df['cast']=df['cast'].fillna('Unknown actor')
```

In [23]:

```
df['rating']=df['rating'].fillna('Unknown rating')
```

In [24]:

```
df['date_added']=pd.to_datetime(df['date_added'])
```

In [25]:

```
df
```

Out[25]:

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown actor	United States	2021-09-25	2020
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown country	2021-09-24	2021
3	s4	TV Show	Jailbirds New Orleans	Unknown director	Unknown actor	Unknown country	2021-09-24	2021
4	s5	TV Show	Kota Factory	Unknown director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021
...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2019-11-20	2007
8803	s8804	TV Show	Zombie Dumb	Unknown director	Unknown actor	Unknown country	2019-07-01	2018
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	2019-11-01	2009
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	2020-01-11	2006

	show_id	type	title	director	cast	country	date_added	release_year
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanu	India	2019-03-02	2015

1. Find the counts of each categorical variable both using graphical and non- graphical analysis.

8807 rows × 12 columns

In [26]:

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

Out[26]:

```
Movie      6131
TV Show    2676
Name: type, dtype: int64
```

In [27]:

```
df['title'].value_counts()
```

Out[27]:

```
Gringo: The Dangerous Life of John McAfee    1
Magic Snowflake                             1
Thithi                                        1
A Flying Jatt                               1
Thackeray                                    1
..
Either Me Or My Auntie                      1
Chris Brown: Welcome to My Life             1
El club de los buenos infieles              1
Oprah Presents When They See Us Now         1
14 Cameras                                  1
Name: title, Length: 8807, dtype: int64
```

In [28]:

```
df['director'].value_counts()
```

Out[28]:

```
Unknown director      2634
Rajiv Chilaka         19
Raúl Campos, Jan Suter 18
Suhas Kadav           16
Marcus Raboy          16
...
John Whitesell        1
Todor Chapkanov       1
Álvaro Delgado-Aparicio L. 1
D Ho                  1
Bilal Lashari         1
Name: director, Length: 4529, dtype: int64
```

In [29]:

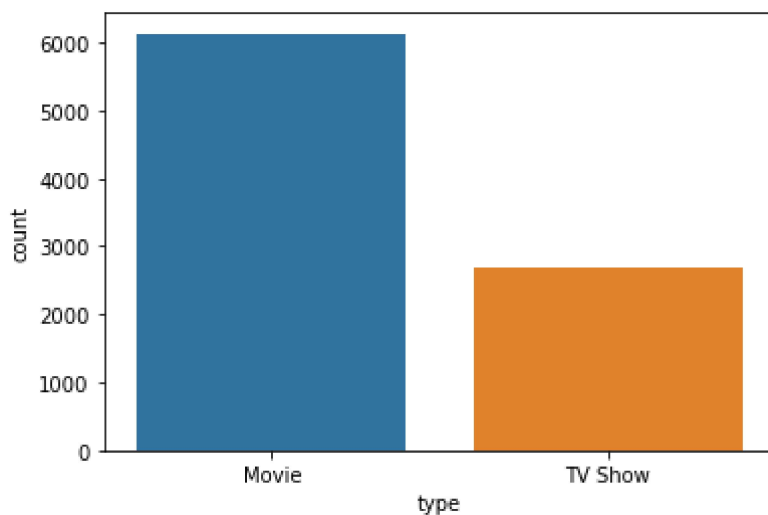
```
df['listed_in'].value_counts()
```

Out[29]:

```
Dramas, International Movies      362
Documentaries                    359
Stand-Up Comedy                  334
Comedies, Dramas, International Movies  274
Dramas, Independent Movies, International Movies  252
...
Horror Movies, International Movies, Sci-Fi & Fantasy  1
Comedies, LGBTQ Movies, Music & Musicals  1
Action & Adventure, Anime Features  1
British TV Shows, TV Comedies, TV Dramas  1
Dramas, Faith & Spirituality, Sports Movies  1
Name: listed_in, Length: 514, dtype: int64
```

In [30]:

```
sns.countplot(data=df,x='type')
plt.show()
```



Based on the above chart it is seen that majority of the audience watch Movie than TV show that is 6000 audience watch Movie and somewhere around 2800 watch Tv show calculating the percentage around 60% watch Movie and 28% watch TV show

2. Comparison of tv shows vs. movies

1. Find the number of movies produced in each country and pick the top 10 countries.

In [31]:

```
movie_df=df.loc[df['type']=='Movie']  
movie_df
```


Out[31]:

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown actor	United States	2021-09-25	2021
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	Unknown country	2021-09-24	2021
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	2021-09-24	1990
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	2021-09-24	2020
12	s13	Movie	Je Suis Karl	Christian Schwochow	Luna Wedler, Jannis Niewöhner, Milan Peschel, ...	Germany, Czech Republic	2021-09-23	2020
...
8801	s8802	Movie	Zinzana	Majid Al Ansari	Ali Suliman, Saleh Bakri, Yasa, Ali Al-Jabri, ...	United Arab Emirates, Jordan	2016-03-09	2016
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2019-11-20	2006
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	2019-11-01	2009
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	2020-01-11	2020
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	2019-03-02	2019

6131 rows × 12 columns

In [32]:

```
data=movie_df.groupby('country')['title'].count().sort_values(ascending=False)
data
```

Out[32]:

```
country
United States    2058
India            893
Unknown country  440
United Kingdom   206
Canada          122
...
India, Mexico    1
India, Malaysia  1
India, Japan      1
India, Iran       1
Zimbabwe         1
Name: title, Length: 652, dtype: int64
```

Top 10 countries

In [33]:

```
data[:10]
```

Out[33]:

```
country
United States    2058
India            893
Unknown country  440
United Kingdom   206
Canada          122
Spain            97
Egypt            92
Nigeria         86
Indonesia        77
Japan            76
Name: title, dtype: int64
```

2. Find the number of Tv-Shows produced in each country and pick the top 10 countries.

In [34]:

```
tv_df=df.loc[df['type']=='TV Show']  
tv_df
```

Out[34]:

	show_id	type	title	director	cast	country	date_added	release_year
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown country	2021-09-24	2021
3	s4	TV Show	Jailbirds New Orleans	Unknown director	Unknown actor	Unknown country	2021-09-24	2021
4	s5	TV Show	Kota Factory	Unknown director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	Unknown country	2021-09-24	2021
...
8795	s8796	TV Show	Yu-Gi-Oh! Arc-V	Unknown director	Mike Liscio, Emily Bauer, Billy Bob Thompson, ...	Japan, Canada	2018-05-01	2015
8796	s8797	TV Show	Yunus Emre	Unknown director	Gökhan Atalay, Payidar Tüfekçioğlu, Baran Akbu...	Turkey	2017-01-17	2016
8797	s8798	TV Show	Zak Storm	Unknown director	Michael Johnston, Jessica Gee-George, Christin...	United States, France, South Korea, Indonesia	2018-09-13	2016
8800	s8801	TV Show	Zindagi Gulzar Hai	Unknown director	Sanam Saeed, Fawad Khan, Ayesha Omer, Mehreen ...	Pakistan	2016-12-15	2012

	show_id	type	title	director	cast	country	date_added	release_year
8803	s8804	TV Show	Zombie Dumb	Unknown director	Unknown actor	Unknown country	2019-07-01	2018

In [35]:

```
data1=tv_df.groupby('country')['title'].count().sort_values(ascending=False)
data1
```

Out[35]:

country	
United States	760
Unknown country	391
United Kingdom	213
Japan	169
South Korea	158
...	
Italy, Canada, France	1
Italy, Germany	1
Italy, South Africa, West Germany, Australia, United States	1
Italy, United Kingdom	1
Uruguay, Germany	1
Name: title, Length: 197, dtype: int64	

Top 10 countries

In [36]:

```
data1[:10]
```

Out[36]:

country	
United States	760
Unknown country	391
United Kingdom	213
Japan	169
South Korea	158
India	79
Taiwan	68
Canada	59
France	49
Spain	48
Name: title, dtype: int64	

3. What is the best time to launch a TV show?

Type *Markdown* and LaTeX: α^2

2. Find which is the best month to release the Tv-show or the movie. Do the analysis separately for Tv-shows and Movies

In [39]:

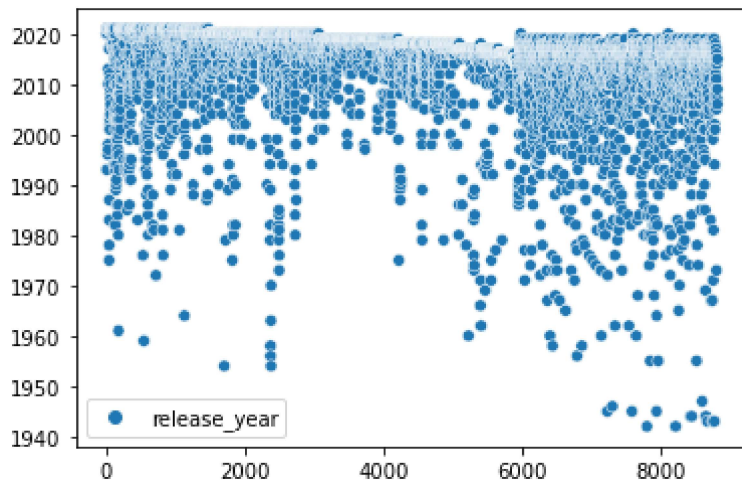
```
df
df_new=df.copy()
df_new=df_new.loc[df_new['type']=='Movie']
df['months']=df_new['date_added'].dt.month_name()
df_new.groupby('months')['title'].count().sort_values(ascending=False).head(1)
```

Out[39]:

```
months
July      565
Name: title, dtype: int64
```

In [40]:

```
sns.scatterplot(data=df_new)
plt.show()
```



4. Analysis of actors/directors of different types of shows/movies.

1. Identify the top 10 directors who have appeared in most movies or TV shows.

In [44]:

```
director_movie=movie_df.groupby('cast')['title'].count().sort_values(ascending=False)
director_movie
```

Out[44]:

```
cast
Unknown actor
475
Vatsal Dubey, Julie Tejjwani, Rupa Bhimani, Jigna Bhardwaj, Rajesh Kava, Mo
usam, Swapnil
13
Samuel West
10
Jeff Dunham
7
Kevin Hart
6

...
Hailee Steinfeld, Woody Harrelson, Kyra Sedgwick, Haley Lu Richardson, Bla
ke Jenner, Hayden Szeto, Alexander Calvert, Eric Keenleyside
1
Hailee Steinfeld, Jessica Alba, Samuel L. Jackson, Sophie Turner, Dove Cam
eron, Toby Sebastian, Gabriel Basso, Thomas Mann, Rachael Harris, Dan Fogl
er, Steve-O, Rob Huebel, Finesse Mitchell, Jaime King      1
Haifa Wehbe, Bassem Samra, Mohamed Lotfy, Ahmed Fathy, Nagua Fouad, Ahlam
Al-Gritly, Karim El Abnoudi, Salah Abdullah, Hakim, Usha
1
Hafsia Herzi, Ash Stymest, Karole Rocher, Paul Hamy, Ira Max, Lindsay Kara
moh, Myriam Djeljeli
1
Şöpe DİRİSÜ, Wunmi Mosaku, Matt Smith, Malaika Wakoli-Abigaba
1
Name: title, Length: 5446, dtype: int64
```

In [77]:

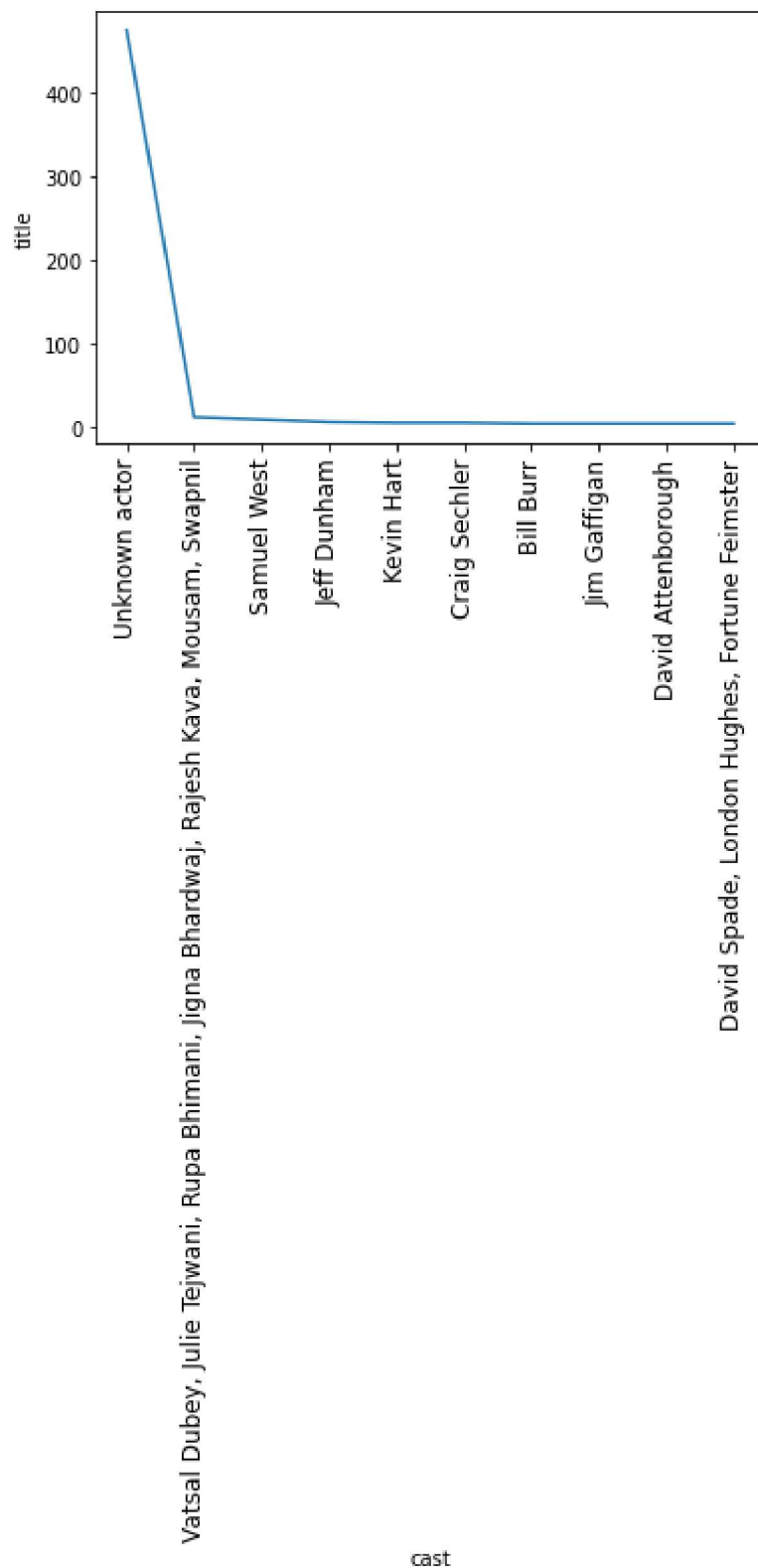
```
top_10=director_movie[:10]  
top_10
```

Out[77]:

```
cast  
Unknown actor  
475  
Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jigna Bhardwaj, Rajesh Kava, Mo  
usam, Swapnil      13  
Samuel West  
10  
Jeff Dunham  
7  
Kevin Hart  
6  
Craig Sechler  
6  
Bill Burr  
5  
Jim Gaffigan  
5  
David Attenborough  
5  
David Spade, London Hughes, Fortune Feimster  
5  
Name: title, dtype: int64
```


In [78]:

```
sns.lineplot(data=top_10)
plt.xticks(rotation=90,fontsize=12)
plt.show()
```



2. Identify the top 10 directors who have appeared in most movies or TV shows.

In [63]:

```
director_tv=movie_df.groupby('director')['title'].count().sort_values(ascending=False)
director_tv
```

Out[63]:

```
director
Unknown director      188
Rajiv Chilaka         19
Raúl Campos, Jan Suter 18
Suhas Kadav           16
Marcus Raboy          15
...
Jose Javier Reyes      1
Jos Eduardo Giordano, Sergio Goyri Jr. 1
Josef Fares            1
Joseph Cross           1
Kief Davidson, Pedro Kos 1
Name: title, Length: 4355, dtype: int64
```

In [65]:

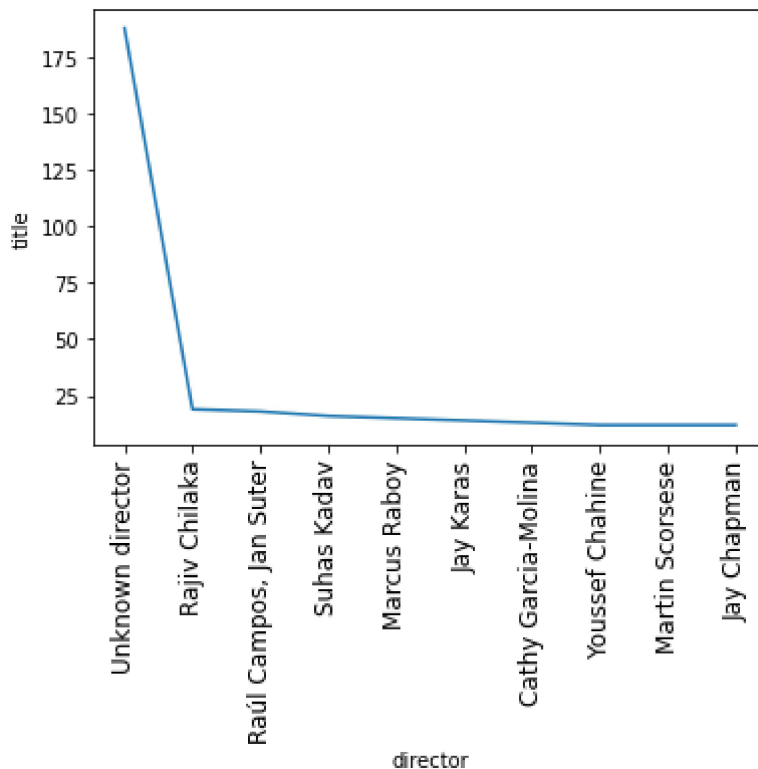
```
top_10=director_tv[:10]
top_10
```

Out[65]:

```
director
Unknown director      188
Rajiv Chilaka         19
Raúl Campos, Jan Suter 18
Suhas Kadav           16
Marcus Raboy          15
Jay Karas             14
Cathy Garcia-Molina   13
Youssef Chahine       12
Martin Scorsese        12
Jay Chapman           12
Name: title, dtype: int64
```

In [74]:

```
sns.lineplot(data=top_10)
plt.xticks(rotation=90, fontsize=12)
plt.show()
```



Type *Markdown* and LaTeX: α^2

5. Which genre movies are more popular or produced more

In [46]:

```
gen=df[['type','listed_in']]
Genre=gen.set_index('type').listed_in.str.split(', ',expand=True).stack()
d2=Genre.reset_index()
d2.rename({0:'Genre'},axis=1,inplace=True)
```

In [47]:

```
d2[d2['type']=='Movie'].Genre.value_counts().head(10)
```

Out[47]:

International Movies	2624
Dramas	1600
Comedies	1210
Action & Adventure	859
Documentaries	829
Dramas	827
Independent Movies	736
Romantic Movies	613
Children & Family Movies	605
Thrillers	512

Name: Genre, dtype: int64

In [48]:

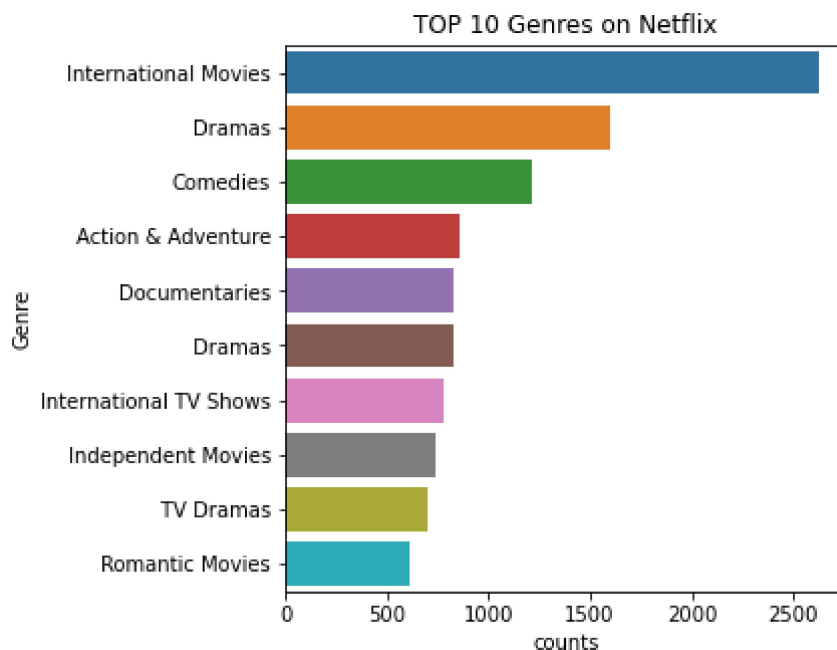
```
d2[d2['type']=='TV Show'].Genre.value_counts().head(10)
```

Out[48]:

```
International TV Shows    774
TV Dramas                696
International TV Shows    577
TV Comedies              461
Crime TV Shows           399
Kids' TV                 388
Romantic TV Shows        338
British TV Shows         253
Docuseries               221
Anime Series             176
Name: Genre, dtype: int64
```

In [50]:

```
Genre=df.set_index('title').listed_in.str.split(',',expand=True).stack()
plt.figure(figsize=(5,5))
g=sns.countplot(y=Genre,order=Genre.value_counts().index[:10])
plt.title('TOP 10 Genres on Netflix')
plt.xlabel('counts')
plt.ylabel('Genre')
plt.show()
```



6. Find After how many days the movie will be added to Netflix after the release of the movie (you can consider the recent past data)

In [61]:

```
def release(x):
    return x['date_added']-x['release_year'].mean()
```

In [55]:

```
show_df = df.loc[(df['type']=='TV Show')]
show_df.head(2)
```

Out[55]:

	show_id	type	title	director	cast	country	date_added	release_year	rating
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV M/
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown country	2021-09-24	2021	TV M/

Insights:

The most content type on Netflix is movies, The popular streaming platform started gaining traction after 2014. Since then, the amount of content added has been increasing significantly, The country by the amount of the produces content is the United States, The most popular director on Netflix , with the most titles, is Jan Suter. International Movies is a genre that is mostly in Netflix, The largest count of Netflix content is made with a “TV-14” rating, The most popular actor on Netflix TV Shows based on the number of titles is Takahiro Sakurai, The most popular actor on Netflix movie, based on the number of titles, is Anupam Kher.

Recommendations:

Most of the recommender systems study users by using their history. Recommender systems have two primary approaches. They are collaborative filtering or content-filtering. Collaborative filtering relies on the concept that people who liked something in the past would also like the same experience in the future. Contentbased filtering methods are useful in places where information is known about the item but not about the user. It functions as a classification task-specific to the user. It models a classifier to model the likes and dislikes of the user concerning the characteristics of an item.