Step-1

Defining problem statement and analyzing basic metrics

Netflix is one of most popular streaming platforms which have more than thousands of movies and TV shows in it.

Problem Statement: -

Our aim is to analyze the data and give valuable insights to company to produce what type of content and how the grow business in different countries.

1 a) **To analyze the data we need to read the data set file**

```
[ ]  # uploading the data from google drive ans giving access to the file.
     from google.colab import drive
     drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
[ ]  # reading or loading the file as dataframe.
     df=pd.read_csv('/content/drive/MyDrive/Netflix-case study/netflix.csv')
```

b) **We need to import libraries like pandas , numpy , matplotlib, seaborn to cleaning and visualize the data**

```
[ ]  # importing the required libraries

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

c) **Analyzing the data using df.head()**—shows top 5 rows of data

```
[ ] # just analyasing the data(Given columns)
    #step-1
    df.head()
```

	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...

So we have columns called show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, description.

df.sample(5) -randomly pick data from dataframe

```
[ ] #data preview sample data pciking randomly
    df.sample(5)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
8778	s8779	Movie	Yes or No 2	Saratswadee Wongsomphet	Supanart Jittaleela, Sushar Manaying, Nisa Boo...	Thailand	November 8, 2018	2012	TV-PG	112 min	International Movies, LGBTQ Movies, Romantic M...	No longer university girls, Kim and Pie face n...
5675	s5676	TV Show	Miss Panda & Mr. Hedgehog	NaN	Dong-hae Lee, Seung-ah Yoon, Jin-hyuk Choi, So...	South Korea	December 15, 2016	2012	TV-PG	1 Season	International TV Shows, Korean TV Shows, Roman...	When a gifted patissier with a gloomy past mee...
202	s203	Movie	Kyaa Kool Hai Hum	Sangeeth Sivan	Tusshar Kapoor, Riteish Deshmukh, Isha Koppika...	India	August 27, 2021	2005	TV-MA	165 min	Comedies, International Movies, Music & Musicals	Longtime friends Rahul and Karan head to Mumba...
757	s758	Movie	Breaking Boundaries: The Science Of Our Planet	Jonathan Clay	David Attenborough, Johan Rockström	United States	June 4, 2021	2021	TV-PG	74 min	Documentaries	David Attenborough and scientist Johan Rockstr...
6514	s6515	TV Show	Cold Case Files Classic	NaN	Bill Kurtis	United States	September 15, 2020	1999	TV-MA	1 Season	Crime TV Shows, Docuseries	Through forensic science and criminal psycholo...

Step-2

Observation on shape, datatype, missing value detection, statistical summary

Shape- **df.shape**--we have 8807 rows and 12 columns

df.info()- To se the data types of each column

df.isna() or df.isnull()- To see whether do we have any missing values in table

df.isna().sum()- To see how many records were missing in each column

statistical summary – df.describe().transpose()—To see the mean, min, max, and to have them in transposed manner.

```
[ ] # finding null values
    df.isna()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	False	False	False	False	True	False	False	False	False	False	False	False
1	False	False	False	True	False	False	False	False	False	False	False	False
2	False	False	False	False	False	True	False	False	False	False	False	False
3	False	False	False	True	True	True	False	False	False	False	False	False
4	False	False	False	True	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...
8802	False	False	False	False	False	False	False	False	False	False	False	False
8803	False	False	False	True	True	True	False	False	False	False	False	False
8804	False	False	False	False	False	False	False	False	False	False	False	False
8805	False	False	False	False	False	False	False	False	False	False	False	False
8806	False	False	False	False	False	False	False	False	False	False	False	False

8807 rows × 12 columns

```
[ ] # count of null values in each column
    df.isna().sum()
```

	0
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

```
[ ] # in total we have 8807 rows and 12 columns
    #step 2 of analysing the data
    df.shape
```

```
(8807, 12)
```

```
[ ] # datatypes of each column.
    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.describe().transpose()# staistical data
```

	count	mean	min	25%	50%	75%	max	std
date_added	8797	2019-05-17 05:59:08.436967168	2008-01-01 00:00:00	2018-04-06 00:00:00	2019-07-02 00:00:00	2020-08-19 00:00:00	2021-09-25 00:00:00	NaN
release_year	8807.0	2014.180198	1925.0	2013.0	2017.0	2019.0	2021.0	8.819312

Checking for duplicate values and datatypes of each column

```
[7] df.dtypes
```

	0
show_id	object
type	object
title	object
director	object
cast	object
country	object
date_added	object
release_year	int64
rating	object
duration	object
listed_in	object
description	object

```
[9] df.duplicated().sum()
```

```
0
```

Step-3

Handling missing values and replacing them.

If we observe we have wrong format of date, we need to change it into right format (yyyy-mm-dd)

Adjust Data Types and Fill in Missing Values

. Verify data types make sense.All except release_year are objects/strings as expected.

. Established information from previous test.

The following are missing data

. Duration. . Rating. . Date_added. . Cast. . Country. . Director.

```
[19] # updating date_added column
     df['date_added'] = pd.to_datetime(df['date_added'].str.strip(), format='%B %d, %Y')
```

```
[20] df.head()
```

	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	2021-09-25	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	2021-09-24	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	2021-09-24	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	2021-09-24	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	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

Now we need to replace NaN with unavailable in every categorical column.

```
[21] # Handling Missing Values
     # Replacing NaN with unavialable to clear missing values (categorical)
     # Now we need to check numerical missing values

     df.fillna({'rating':'Unavailable','cast':'Unavailable','country':'Unavailable','director':'Unavailable'},inplace=True)
```

Now we need to replace the missing values in numerical columns with mean, mode or most recent value.

Date_added column has some missing values . So, here I have replaced them with most recent date

```
[24] # We need replace empty date_added rows with either mode or most recenrt date .That can be found using max()

     recetn_date=df['date_added'].max()
```

```
[25] df.fillna({'date_added':recetn_date})
```

Step-4

Non-Graphical Analysis and unnesting the data

```
[22] #3. Non-Graphical Analysis: Value counts and unique attributes

     unique_values=df.nunique()
     unique_values.name='count_of_values'

     unique_values
```

	count_of_values
show_id	8807
type	2
title	8807
director	4529
cast	7693
country	749
date_added	1714
release_year	74
rating	18
duration	220
listed_in	514
description	8775

dtype: int64

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

array(['Movie', 'TV Show'], dtype=object)
```

```
[23] no_of_movies_in_each_type=df['type'].value_counts()
     no_of_movies_in_each_type
```

type	count
Movie	6131
TV Show	2676

dtype: int64

a) Here we are finding the unique count of values for each column, and unique types of data present.

b) No. of movies directed by each director individually

```
[24] # checking out the no of movies directed by each director
     movies_by_each_director =df_director=df['director'].str.split(',').explode().value_counts().head(10)
     movies_by_each_director.reset_index()
     movies_by_each_director.name='count_of_movies'
     movies_by_each_director
```

director	count_of_movies
Unavailable	2634
Rajiv Chilaka	22
Jan Suter	18
Raúl Campos	18
Marcus Raboy	16
Suhas Kadav	16
Jay Karas	15
Cathy Garcia-Molina	13
Martin Scorsese	12
Jay Chapman	12

C) No. of movies acted by actor/cast

```
movies_by_each_cast=df_cast=df['cast'].str.split(',').explode().value_counts().head(10)
movies_by_each_cast.reset_index()
movies_by_each_cast.name='count_of_movies'
movies_by_each_cast
```

cast	count_of_movies
Unavailable	825
Anupam Kher	39
Rupa Bhimani	31
Takahiro Sakurai	30
Julie Tejwani	28
Om Puri	27
Shah Rukh Khan	26
Rajesh Kava	26
Andrea Libman	25
Paresh Rawal	25

d) Country and there total no of movies or TV Shows produced.

```
[28] count_in_each_country=df_country=df['country'].str.split(',').explode().value_counts().head(10)
     count_in_each_country.reset_index()
     count_in_each_country.name='count_of_movies'
     count_in_each_country
```

country	count_of_movies
United States	3211
India	1008
Unavailable	831
United Kingdom	628
United States	479
Canada	271
Japan	259
France	212
South Korea	211
France	181

dtype: int64

If We observe we have more number of NaN which is unavailable data in few columns which is almost 30% of data is missing or not given.

So clearing them by replacing them with mean, mode, median or unavailable values.

We need to change the date format into particular format which system accepts.

Insights on non-graphical analysis : -

- **a)** We have more no of missing values
- **b)** Most happening director – Rajiv Chilaka
- **c)** Most happening actor- Anupam Kher
- **d)** Most no of content were released in year 2018
- **e)** Country which released most of content-USA followed by India
- **f)** Most of movies got added into Netflix in 7th month of every year

Recommendations based on Non-Graphical Analysis: -

a) India as a country need to concentrate most on Producing TV Shows.

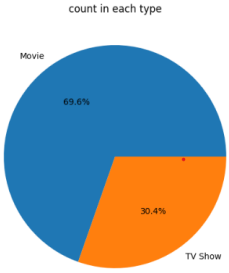
- **b)** Most of tv shows and movies have genre of international movies followed by drama. So we need to try keep on adding new movies of such genre and also need to concentrate on least watched genre and need to analyze it and try to add them back.
- **c)** Highest genre- TV-MA preceded by TV-14
- **d)** Best time to release the movies into Netflix were 7th month (July) followed by 12th month(December)

Step-5

Visual Analysis

- **a) Pie Chart**

```
plt.figure(figsize=(8,6))
plt.pie(data=df,labels=df['type'].value_counts().index,x=df['type'].value_counts(),autopct='%1.1f%%')
plt.title('count in each type')
plt.show()
```



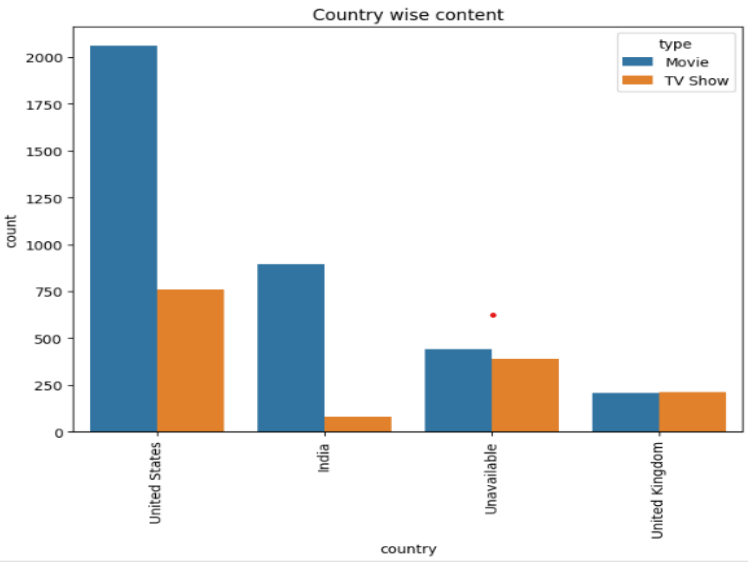
If we observe the above pie chart most of content available in Netflix are Movie TypeLooking into this chart we can say that audience were of movies type, and Netflix need to add more no TV shows which attract the audience with good genre and ratings

b) Countplot using hue

United states produce more no of movies and TV shows followed by india in movies and UK in Tv Shows.

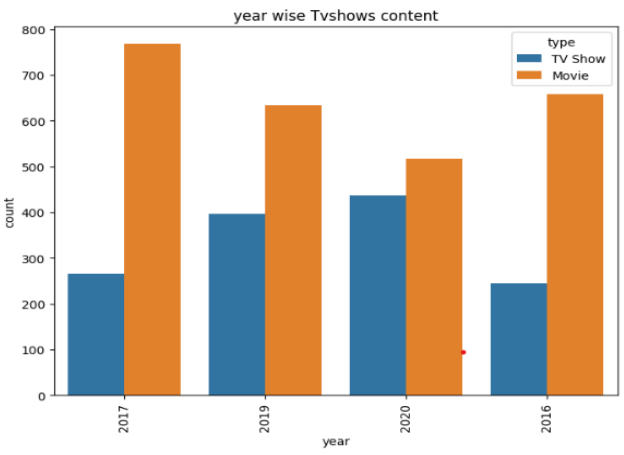
```
# including both tv shows and movies

plt.figure(figsize=(8,6))
sns.countplot(data=df,x='country', order=df['country'].str.split(',').explode().value_counts().index[0:4],hue='type')
plt.title('Country wise content')
plt.ylabel('count')
plt.xticks(rotation=90)
plt.xlabel('country')
plt.show()
```



c) Year wise analysis

```
plt.figure(figsize=(8,6))
sns.countplot(data=df,x='release_year',order=df['release_year'].value_counts(ascending=False).index[1:5],hue='type')
plt.title('year wise Tvshows content')
plt.ylabel('count')
plt.xticks(rotation=90)
plt.xlabel('year')
plt.show()
```

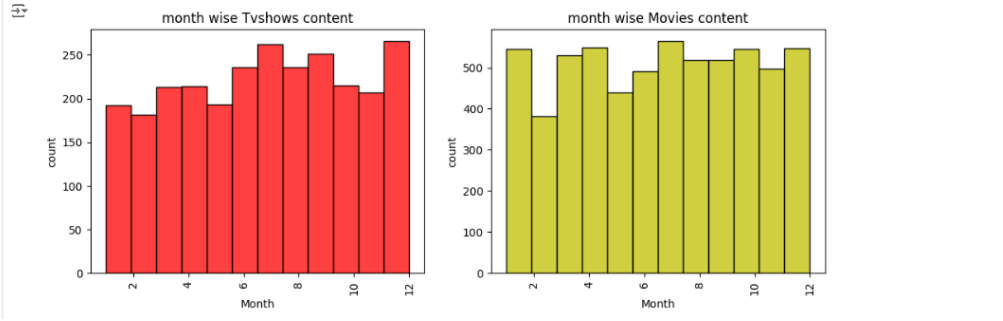


a) Most of releases happened between 2012 and 2021 among which most of them were released in 2018(Combined both TV shows and Movies)
b) 2020 has highest tv shows & 2017 has highest movies in Netflix
c) seems to be Netflix is concentrating more on Movies than TVshows in 2016 and 2017
d) seems to be Netflix is concentrating more on TV shows than moviesin 2019 and 2020

d) Month wise histogram analysis

```
df2=df[df['type']=='TV Show']
plt.figure(figsize=(12,4))
plt.subplot(1,2,1)
sns.histplot(data=df2, x='Month',bins=12,color='red')
plt.title('month wise Tvshows content')
plt.ylabel('count')
plt.xticks(rotation=90)
plt.xlabel('Month')
df3=df[df['type']=='Movie']

plt.subplot(1,2,2)
sns.histplot(data=df3, x='Month',bins=12,color='y')
plt.title('month wise Movies content')
plt.ylabel('count')
plt.xticks(rotation=90)
plt.xlabel('Month')
plt.show()
```

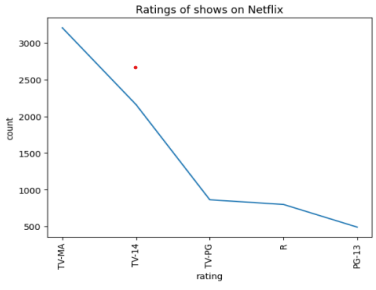


7th month has most number of movies released and 12th month has most number of TV Shows released.
Least in February

e) Line plot of ratings

```
sns.lineplot(data=df['rating'].value_counts()[0:5])
plt.title('Ratings of shows on Netflix')
plt.xlabel('rating')
plt.xticks(rotation=90)
```

```
([0, 1, 2, 3, 4],
 [Text(0, 0, 'TV-MA'),
  Text(1, 0, 'TV-14'),
  Text(2, 0, 'TV-PG'),
  Text(3, 0, 'R'),
  Text(4, 0, 'PG-13')])
```



TV-MA rated movies count was more in Netflix.

f) Dist plot

Huge number of movies/TV shows got released in year 2018

```
sns.distplot(df['release_year'],color='gold')
plt.show()
```

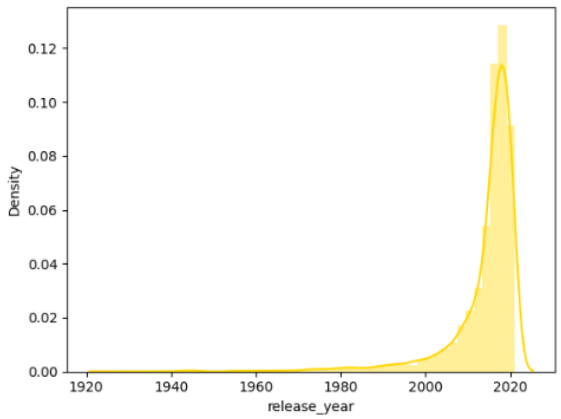
```
<ipython-input-43-d906f2f5f77b>:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df['release_year'],color='gold')
```



g) Most no of movies directed by each director. (Box Plot)

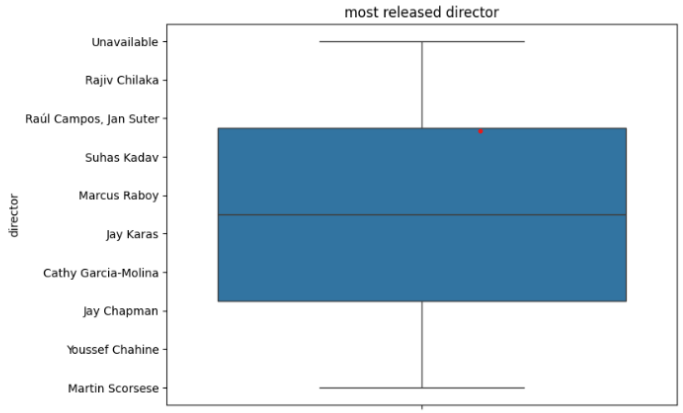
We have so many unavailable data which in case the top director might get changed if we have right information of data.

Most no of movies directed- Rajiv Chilaka

```
plt.figure(figsize=(8,6))
sns.boxplot(y=df['director'].value_counts().index[:10])
plt.title('most released director')
plt.ylabel('director')

#
```

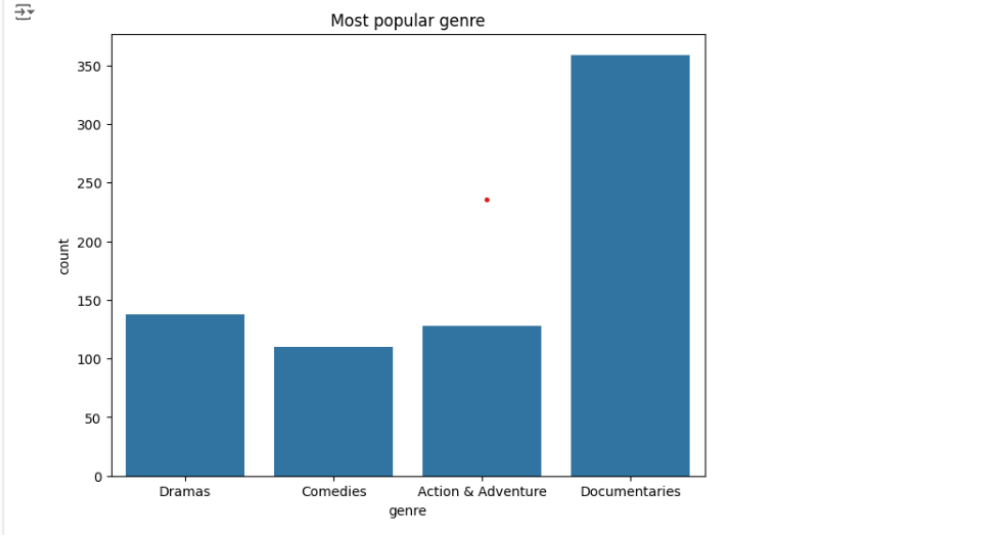
```
Text(0, 0.5, 'director')
```



h) Countplot on genre

Top genre were international movies followed by Dramas

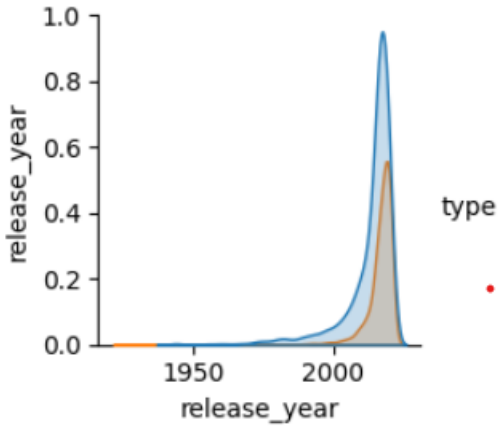
```
plt.figure(figsize=(8,6))
sns.countplot(data=df,x='listed_in',order=df['listed_in'].str.split(',').explode().value_counts().index[1:5])
plt.title('Most popular genre')
plt.xlabel('genre')
plt.ylabel('count')
plt.xticks(rotation=0)
plt.show()
```



i) Pair plot-The only possible pair plot

```
sns.pairplot(data=df,hue='type')
```

```
<seaborn.axisgrid.PairGrid at 0x7fb5c4c3faf0>
```



Step-6

Business Insights

. Netflix most valued director was Rajiv Chilaka
. Most valued actor was Anupam Kher
. Most of them have TV-MA genre content.
. Best time to add movies into Netflix were 7th month and 12th month.
. Most content from USA, INDIA, UK
. Most of releases happened between 2012 and 2021 among which most of them were released in 2018 (Combined both TV shows and Movies)
. 2020 has highest tv shows & 2017 has highest movies in Netflix
. seems to be Netflix is concentrating more on Movies than TVshows in 2016 and 2017
. seems to be Netflix is concentrating more on TV shows than movies in 2019 and 2020

Step-7

Recommendations

a) India as a country need to concentrate most on Producing TV Shows.
b) Most of tv shows and movies have genre of international movies followed by drama. So we need to try keep on adding new movies of such genre and also need to concentrate on least watched genre and need to analyze it and try to add them back.
c) We can add most of the content Having genre- TV-MA preceded by TV-14
d) Best time to release the movies into Netflix were 7th month (July) followed by 12th month (December)