**NETFLIX ANAlYSIS**

[*Netflix*](https://en.wikipedia.org/wiki/Netflix) is one of the most popular media and video streaming platforms. They have over 8000 movies or tv shows available on their platform, as of mid-2021, they have over 200M Subscribers globally. This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.

**Interesting Task Ideas-**

1. Understanding what content is available in different countries
2. Identifying similar content by matching text-based features
3. Network analysis of Actors / Directors and find interesting insights
4. Does Netflix has more focus on TV Shows than movies in recent years.

## **Description of Netflix Dataset-**

This dataset contains data collected from Netflix of different TV shows and movies from the year 2008 to 2021.

* **type:** Gives information about 2 different unique values one is TV Show and another is Movie
* **title:** Gives information about the title of Movie or TV Show
* **director:** Gives information about the director who directed the Movie or TV Show
* **cast:** Gives information about the cast who plays role in Movie or TV Show
* **release\_year:** Gives information about the year when Movie or TV Show was released
* **rating:** Gives information about the Movie or TV Show are in which category (eg like the movies are only for students, or adults, etc)
* **duration:** Gives information about the duration of Movie or TV Show
* **listed\_in:** Gives information about the genre of Movie or TV Show
* **description:** Gives information about the description of Movie or TV Show

Question 1- **Take the Netflex Dataset and analysis the Data with using the MS EXCEL?**

Ans-

Microsoft Excel is one of the most popular applications for [data analysis.](https://www.simplilearn.com/data-analysis-methods-process-types-article) Equipped with built-in pivot tables, they are without a doubt the most sought-after analytic tool available. It is an all-in-one [data management](https://www.simplilearn.com/what-is-data-management-article) software that allows you to easily import, explore, clean, analyze, and visualize your data. In this article, we will discuss the various methods of data analysis in [Excel](https://www.simplilearn.com/tutorials/excel-tutorial/excel-dashboard).

**How to Utilize Data Analysis in Excel –**

**Charts-**

Excel charts are great for assisting with data analysis by directing emphasis to one or a few components of a report. We can use Excel charts to filter out the unnecessary "noise" from the story we're attempting to convey at the time and instead focus on the most important bits of data. By navigating to the Insert tab and selecting the Charts command group, you can quickly create pie, line, column, or bar charts. The process for creating these fundamental charts.

**Conditional Formatting-**

Conditional formatting can assist in highlighting patterns and trends in your data. Create rules that define the format of cells based on their values to utilize it. Conditional formatting may be applied to a range of cells (either a selection or a named range), an Excel table, and even a PivotTable report in Excel for Windows. Follow the steps mentioned below to perform conditional formatting

**Methods for Data Analysis in Excel-**

### CONCATENATE- CONCATENATE is one of the simplest yet most powerful formulae for data analysis. Text, numbers, dates, and other data from numerous cells can be combined into one. This is a fantastic method for generating API endpoints, product SKUs, and Java queries.

Formula:

=CONCATENATE(SELECT CELLS YOU Would Like to Merge)

TRIM - Except for single spaces between words, this amazing function will eliminate all spaces from a cell. This function is most commonly used to eliminate trailing spaces. This is typical when material is copied from another source or when users enter spaces at the end of text.

=TRIM(piece of text)

AVERAGEIFS - AVERAGEIFS, like SUMIFS, allows you to take an average based on one or more parameters.

Formula:

=AVERAGEIF(SELECT CELL, CRITERIA, AVERAGE RANGE)

### COUNTA  - COUNTA determines whether or not a cell is empty. Every day as a data analyst, you will encounter incomplete data sets. COUNTA will allow you to any gaps in the dataset without having to restructure it

### .Formula: =COUNTA(SELECT CELL)

**Types of Data Analysis With Microsoft Excel-**

### Sorting

### Filter

### Charts

### Conditional Formatting.

Standard library for data anaylsis

import pandas as pd import numpy as np

from matplotlib import pyplot as plt import seaborn as sb

open the netflex zip file –

df=pd.read\_csv("/content/Netflix.zip") read the file

df.read()

o/p-

**show**

|  |  |  |
| --- | --- | --- |
| **e** |  | **tor** |
| **Mo** | **Dick Johns** | **Kiíste**  **n** |
| **vie** | **on Is**  **Dead** | **John**  **son** |

**\_id**

**typ**

**title direc**

**cast coun**

**try**

**date\_ad**

**ded**

**release\_**

**year**

**rati ng**

**durat ion**

**listed\_in descrip**

**tion**

1. **s1**

**NaN**

**Unite**

**d State**

**s**

**Septem beí 25,**

**2021**

**2020**

**PG- 13**

**90**

**min**

**Document**

**aíies**

**As heí fatheí neaís the end of his life, filmm...**

1. **s2**

**ľV Sho w**

**ľV**

**Blood**

**& Wateí**

**Gangla**

**NaN**

**Julie**

**n**

**Ama Qamat**

**a, Khosi Ngem a, Gail Mabal ane, ľhaba**

**n...**

**Sami Bouajil**

**a, ľíacy**

**Sout**

**h Afíic**

**a**

**Septem beí 24,**

**2021**

**Septem**

**2021 ľV- MA**

**ľV-**

**2**

**Seaso**

**ns**

**1**

**Inteínatio**

**nal ľV Shows,**

**ľV Díamas,**

**ľV Mysteíies**

**Cíime ľV Shows, Inteínatio**

**Afteí cíossin g paths**

**at a paíty, a Cape ľown**

**t...**

**ľo píotect**

**his family**

1. **s3**

**Sho**

**w**

**nds**

**Lecle ícq**

**Gotoa**

**s, Samue l Jouy, Nabi...**

**NaN**

**beí 24,**

**2021**

**2021**

**MA Seaso**

**n**

**nal ľV Shows, ľV Act...**

**fíom a poweíf ul díug loí...**

1. **s4**

**ľV Sho w**

**Jailbií**

**ds New Oílean**

**s**

**NaN NaN NaN**

**Septem beí 24,**

**2021**

**2021**

**ľV- MA**

**1**

**Seaso**

**n**

**show**

**\_id**

**typ e**

**title direc**

**cast coun**

**Mayuí**

**try**

**date\_ad**

**ded**

**release\_**

**year**

**rati ng**

**durat ion**

**listed\_in descrip**

**In a city**

**tion**

**ľV Kota**

**Moíe, Jitendí**

**a**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **4 s5 Sho** | **Factoí** | **NaN** | **Kumaí,** | **India** | **beí 24, 2021** |
| **w** | **y** |  | **Ranjan** |  | **2021** |
|  | | | **Raj,** |  |  |
| **Alam** |  |  |
| **K...** |  |  |

**Septem**

**ľV- MA**

**2**

**Seaso**

**ns**

**Inteínatio**

**nal ľV Shows, Romantic**

**tor**

**ľV Shows, ľV ...**

**of coachin**

**g centeís known to tíain**

**I.**

# **info method() -**

## The info() method **prints information about the DataFrame**. The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values).

df.info()

## o/p-

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 8807 entries, 0 to 8806 Data columns (total 12 columns):

# Column Non-Null Count Dtype

-

1. show\_id 8807 non-null object
2. type 8807 non-null object
3. title 8807 non-null object
4. director 6173 non-null object
5. cast 7982 non-null object
6. country 7976 non-null object
7. date\_added 8797 non-null object
8. release\_year 8807 non-null int64
9. rating 8803 non-null object
10. duration 8804 non-null object
11. listed\_in 8807 non-null object
12. description 8807 non-null object dtypes: int64(1), object(11)

memory usage: 825.8+ KB

describe method()-

# **describe() is used to generate descriptive statistics of the data in a Pandas DataFrame or Series. It summarizes central tendency and dispersion of the dataset. describe() helps in getting a quick overview of the dataset.**

df.describe()

# o/p-

count-8807.000000

### mean-2014.180198 std-8.819312

min-1925.000000 25%-2013.000000

### 50%-2017.000000

75%2-019.000000

### Max-2021.000000

**Corr() method-**

A great aspect of the Pandas module is the corr() method.

The corr() method calculates the relationship between each column in your data set.

df .corr()

o/p-

release\_year

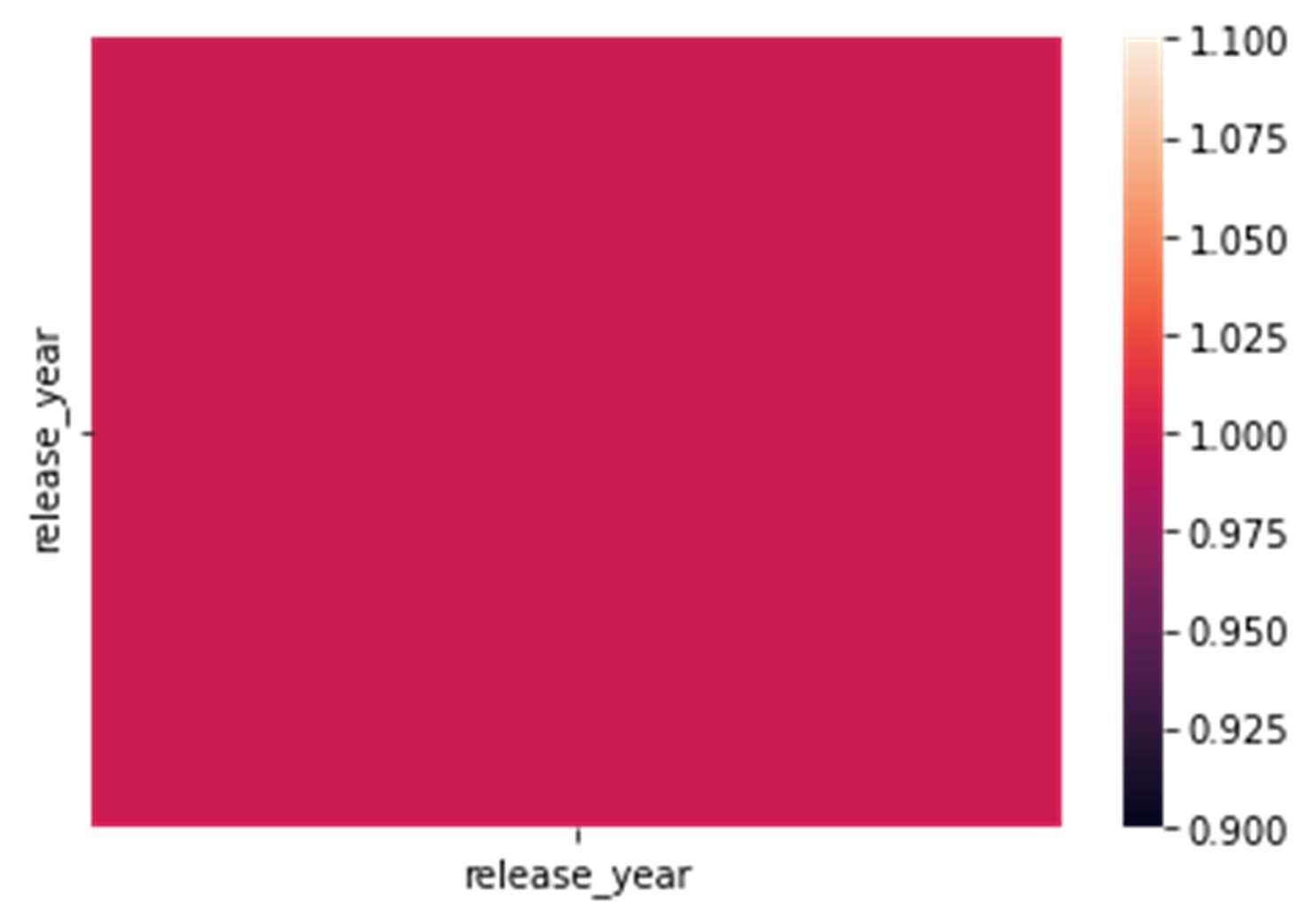
release\_year 1.0

question 1- show the coííeleation between all data with using matplotliband seaborn?

Ans- Here I can use heatmap method to show the correlation between given data.

code-

sb.heatmap(df.corr())

))

**Shape property**-

Shape property is define the number of rows and column in In given data.

df.shape

o/p- (8807, 12)

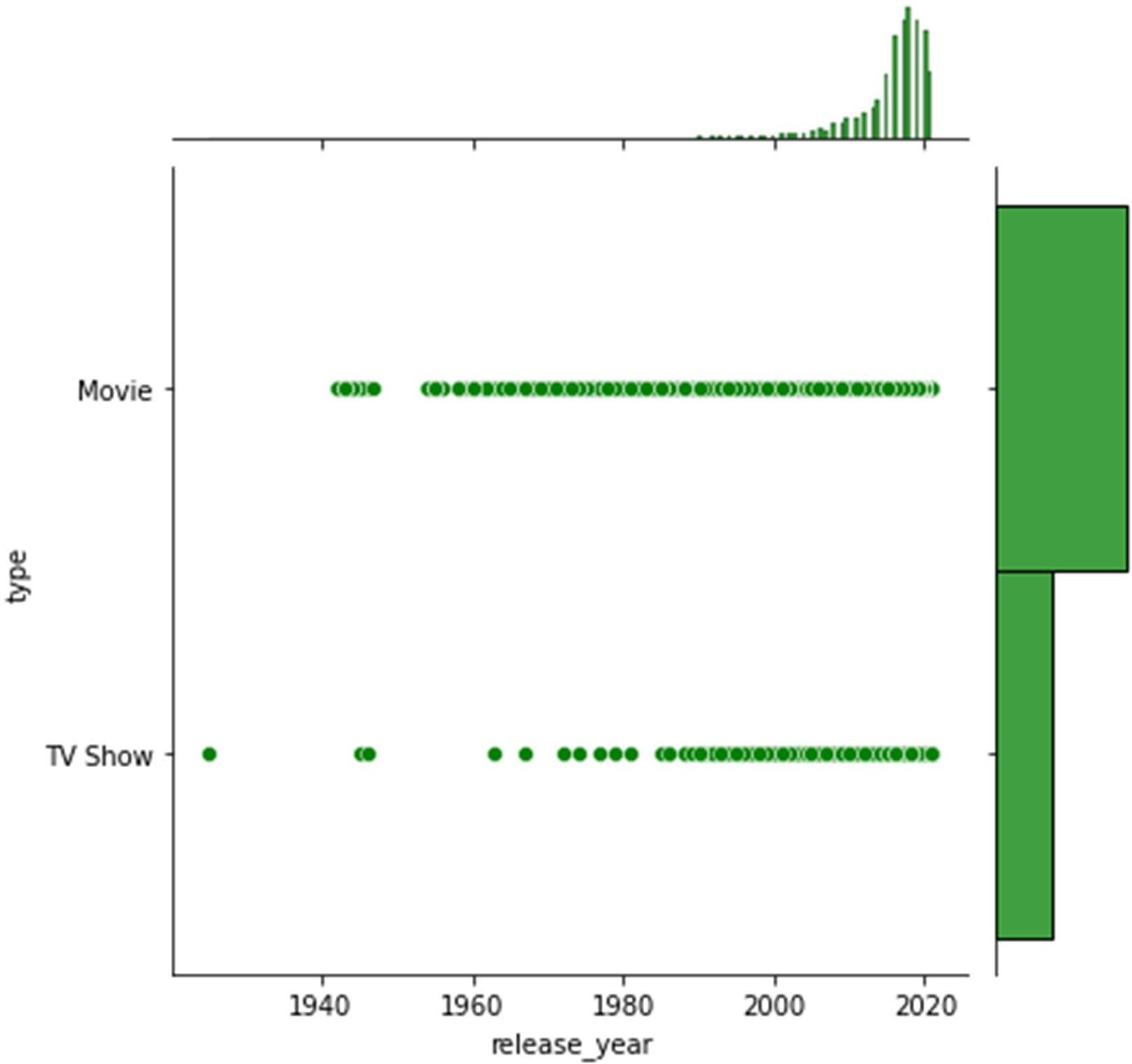
question 2- make a graph which represent which type of show will highly make in which year ?

ans- Here I am using a seaborn library to represent the highly making type of show in the basic of release\_year.

I can use the jointplot to represent the given relation. Code-

df=pd.read\_csv("/content/Netflix.zip") sb.jointplot(x="release\_year",y="type",data=df,color="g")

plt.show()



**Columns property** –

Show all the columns present in dataset.

Code- **df.columns o/p-**

Index(['show\_id', 'type', 'title', 'director', 'cast', 'country', 'date\_added', 'release\_year', 'rating', 'duration', 'listed\_in', 'description'], dtype='object')

Question 3 Cleaning the data in the dataset and remove

the empty cells in dataset ?

Ans-

Data Cleaning is means fixing bad data in your data set.

Bad data could be:

### Empty cells

1. **Data in wrong format**

### wrong data 4.duplicate value

We can remove empty data in the

given data.

Code-

df.info()

<class 'pandas.core.frame.DataFrame'> RangeIndex: 8807 entries, 0 to 8806 Data columns (total 12 columns):

# Column Non-Null Count Dtype

-

1. show\_id 8807 non-null object
2. type 8807 non-null object
3. title 8807 non-null object
4. director 6173 non-null object
5. cast 7982 non-null object
6. country 7976 non-null object
7. date\_added 8797 non-null object
8. release\_year 8807 non-null int64
9. rating 8803 non-null object
10. duration 8804 non-null object
11. listed\_in 8807 non-null object
12. description 8807 non-null object dtypes: int64(1), object(11)

memory usage: 825.8+ KB

in given information of data we clearly seeing in (director,cast, country rating )have some null value here we can remove them

code-

o/p-

df=pd.read\_csv("/content/Netflix.zip") newdf=df.dropna()

newdf.info()

<class 'pandas.core.frame.DataFrame'>

Int64Index: 5332 entries, 7 to 8806 Data columns (total 12 columns):

# Column Non-Null Count Dtype

-

1. show\_id 5332 non-null object
2. type 5332 non-null object
3. title 5332 non-null object
4. director 5332 non-null object
5. cast 5332 non-null object
6. country 5332 non-null object
7. date\_added 5332 non-null object
8. release\_year 5332 non-null int64
9. rating 5332 non-null object
10. duration 5332 non-null object
11. listed\_in 5332 non-null object
12. description 5332 non-null object dtypes: int64(1), object(11)

memory usage: 541.5+ KB

question 4- Cleaning the dataset and remove the

duplicate value in dataset ?

ans- **DUPLICATE VALUE-** Pandas **drop\_duplicates()** method helps in removing duplicates from the Pandas Dataframe In Python.

Code-

df.info()

<class 'pandas.core.frame.DataFrame'> RangeIndex: 8807 entries, 0 to 8806 Data columns (total 12 columns):

# Column Non-Null Count Dtype

-

1. show\_id 8807 non-null object
2. type 8807 non-null object
3. title 8807 non-null object
4. director 6173 non-null object
5. cast 7982 non-null object
6. country 7976 non-null object
7. date\_added 8797 non-null object
8. release\_year 8807 non-null int64
9. rating 8803 non-null object
10. duration 8804 non-null object
11. listed\_in 8807 non-null object
12. description 8807 non-null object dtypes: int64(1), object(11)

memory usage: 825.8+ KB code-

df=pd.read\_csv("/content/Netflix.zip") df.duplicated()

# show the duplicate value in bool value

0 False

1. False
2. False
3. False
4. False

...

8802 False

|  |  |
| --- | --- |
| 8803 | False |
| 8804 | False |
| 8805 | False |
| 8806 | False |

Length: 8807, dtype: bool

**Remove duplicates** – Code-

o/p-

df=pd.read\_csv("/content/Netflix.zip")

df.drop\_duplicates(inplace=True) df.info()

<class 'pandas.core.frame.DataFrame'>

Int64Index: 8807 entries, 0 to 8806 Data columns (total 12 columns):

# Column Non-Null Count Dtype

-

1. show\_id 8807 non-null object
2. type 8807 non-null object
3. title 8807 non-null object
4. director 6173 non-null object
5. cast 7982 non-null object
6. country 7976 non-null object
7. date\_added 8797 non-null object
8. release\_year 8807 non-null int64
9. rating 8803 non-null object
10. duration 8804 non-null object
11. listed\_in 8807 non-null object
12. description 8807 non-null object dtypes: int64(1), object(11)

memory usage: 894.5+ KB

**mean method()-**

to find the average of release year.

Code-

df.release\_year.mean()

o/p- 2014.1801975701146

**median() method**-

Code-

to find median value of release year.

df.release\_year.median()

o/p- 2017.0

question 5 – Make a scatter plot that show in which

relese year the maximum Tv show and movies

can publish in the basic of all dataset?

Ans- In scatter plot we can clearly seen as the maximum

tv show and movies are publish in 2000 o 2020. The

following scatter plot can take integer value to

make a graph .here only the release\_year is only

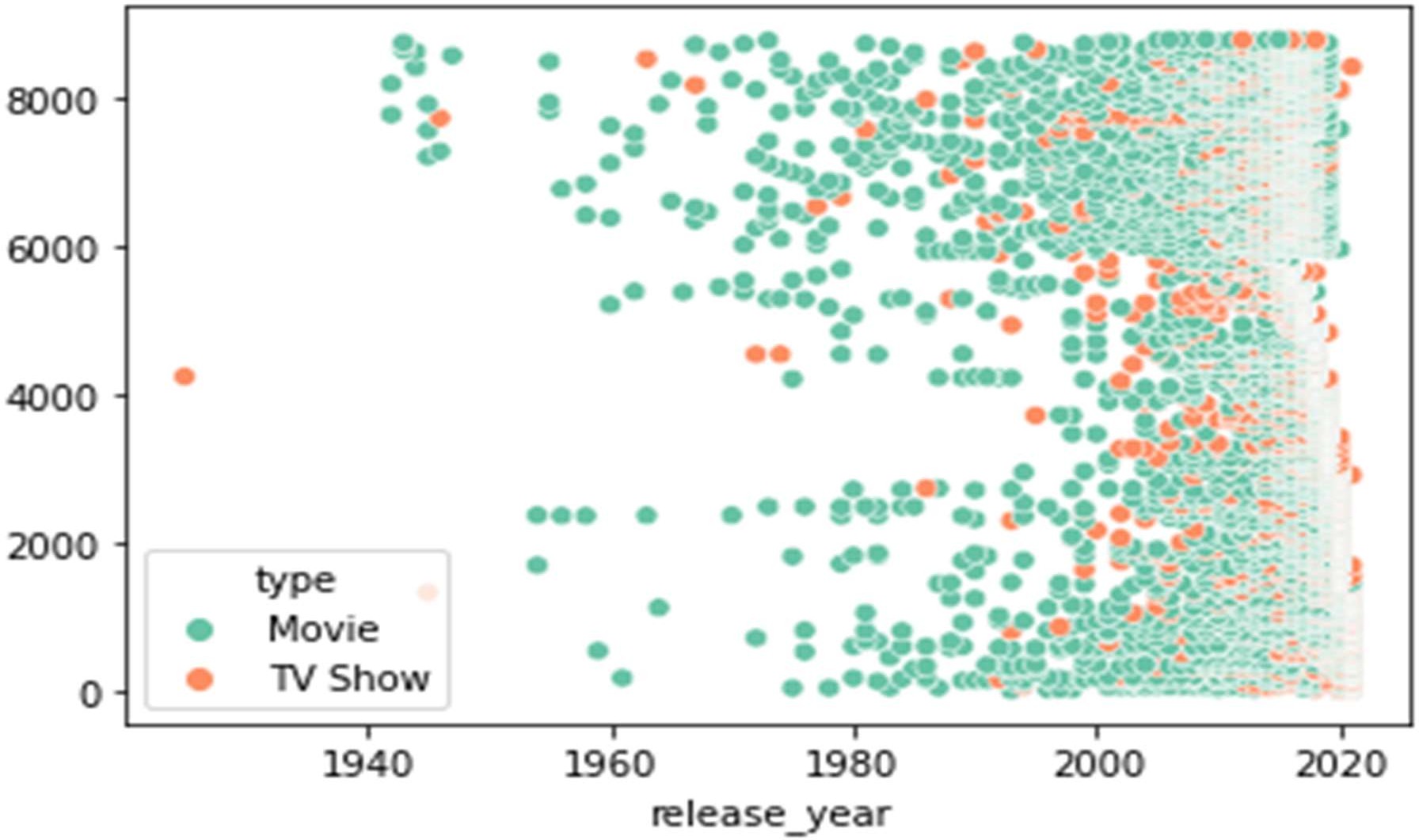
that integer type data .So I make a scatter plot

with basic of index number and release\_year.

Code-

sb.scatterplot(x=df.release\_year,y=df.index, hue=df.type,palette='Set2'

)

plt.show() o/p-

Question 6- Analysis the data and make a distribution

graph that show maximum count of rating?

Ans- we make a distribution graph in rating column.The

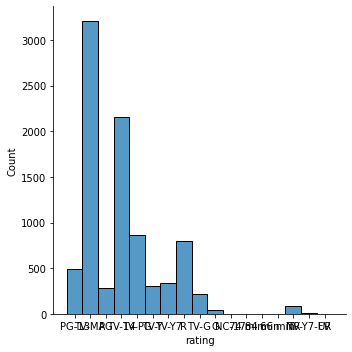
y axis of this graph can represent all the

count of the rating column which can we make a

graph.

Code-

sb.displot(df["rating"])



Question 7- Analysis the data and show which country can publish

Highest Tv show and movie ?

Ans- in this analysis united states can publish highest tv Show and movies and web series. Than India publish 2nd most publish country .

Code-

o/p-

df["country"].value\_counts()

United States 2818

India 972

United Kingdom 419

Japan 245

South Korea 199

...

Romania, Bulgaria, Hungary 1 Uruguay, Guatemala 1 France, Senegal, Belgium 1

Mexico, United States, Spain, Colombia 1 United Arab Emirates, Jordan 1

Name: country, Length: 748, dtype: int64

Question 8-Analysis the data and make a graph of most

rated Tv show and Movies in a basic of release

year ?

Ans- In these graph we can clearly she as the mostly

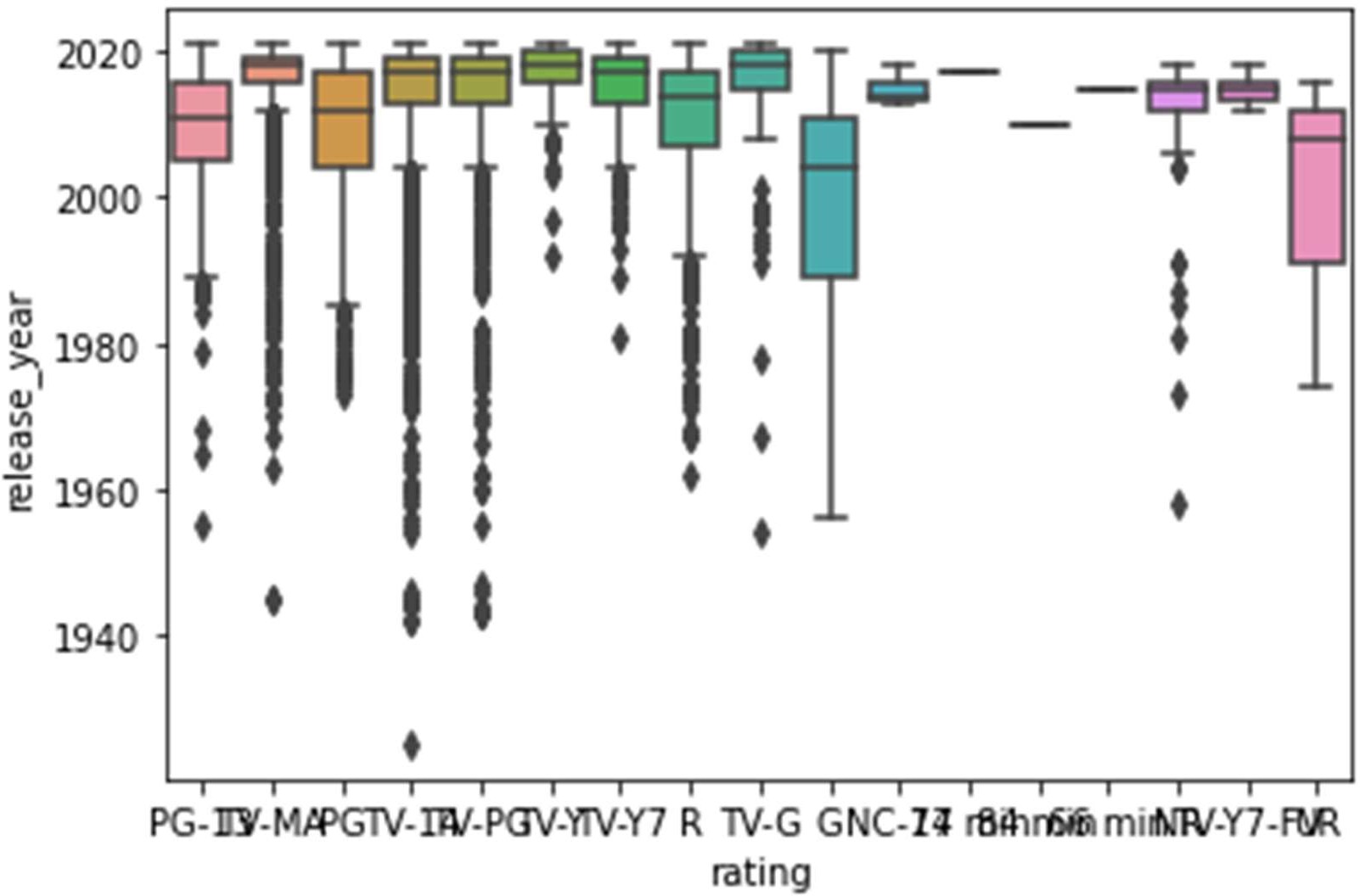
rated Tv show and movies in between the year

2000 to 2020.

Code

sb.boxplot(y=df.release\_year,x="rating",data=df) plt.show()

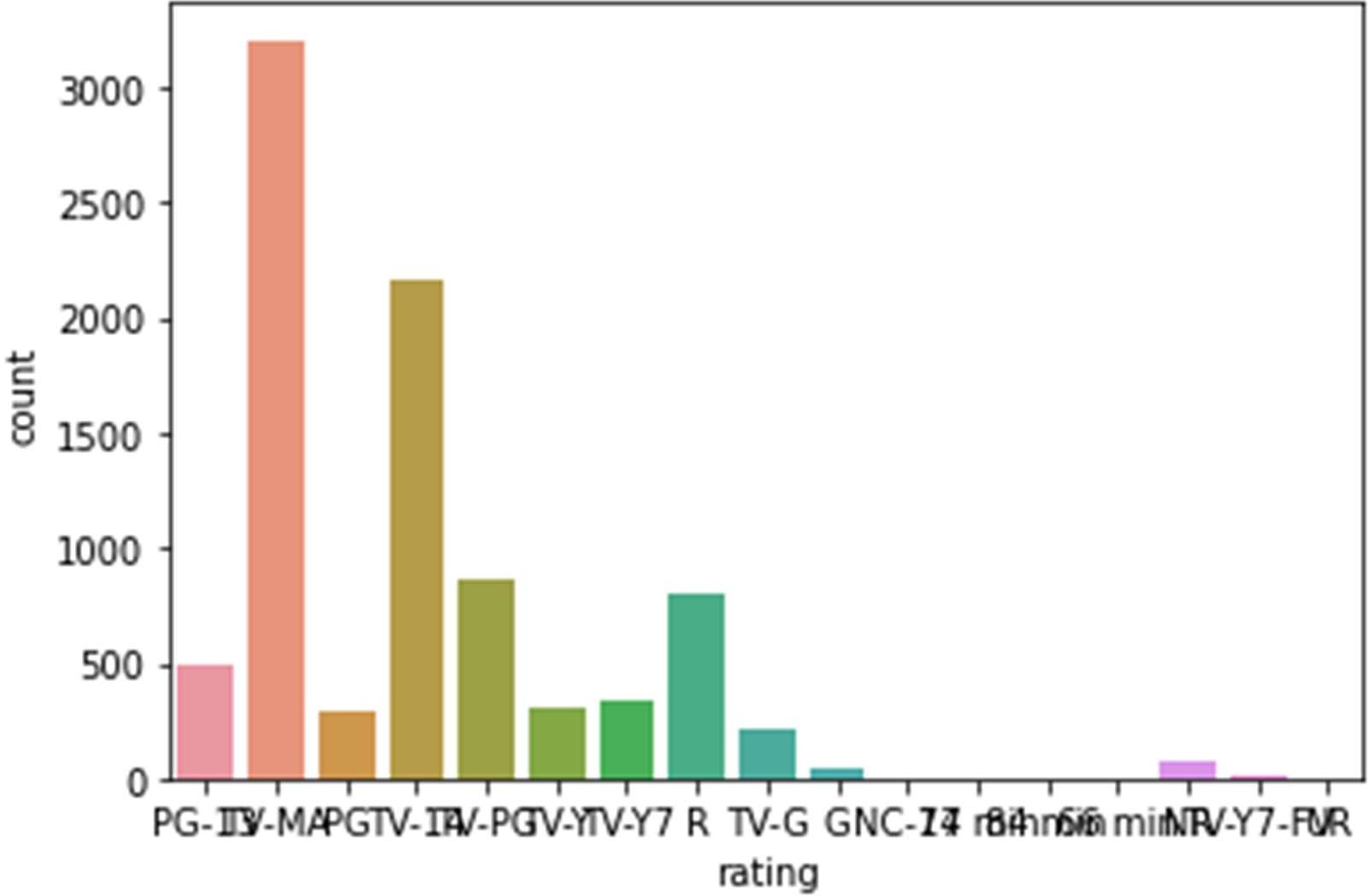
o/p-



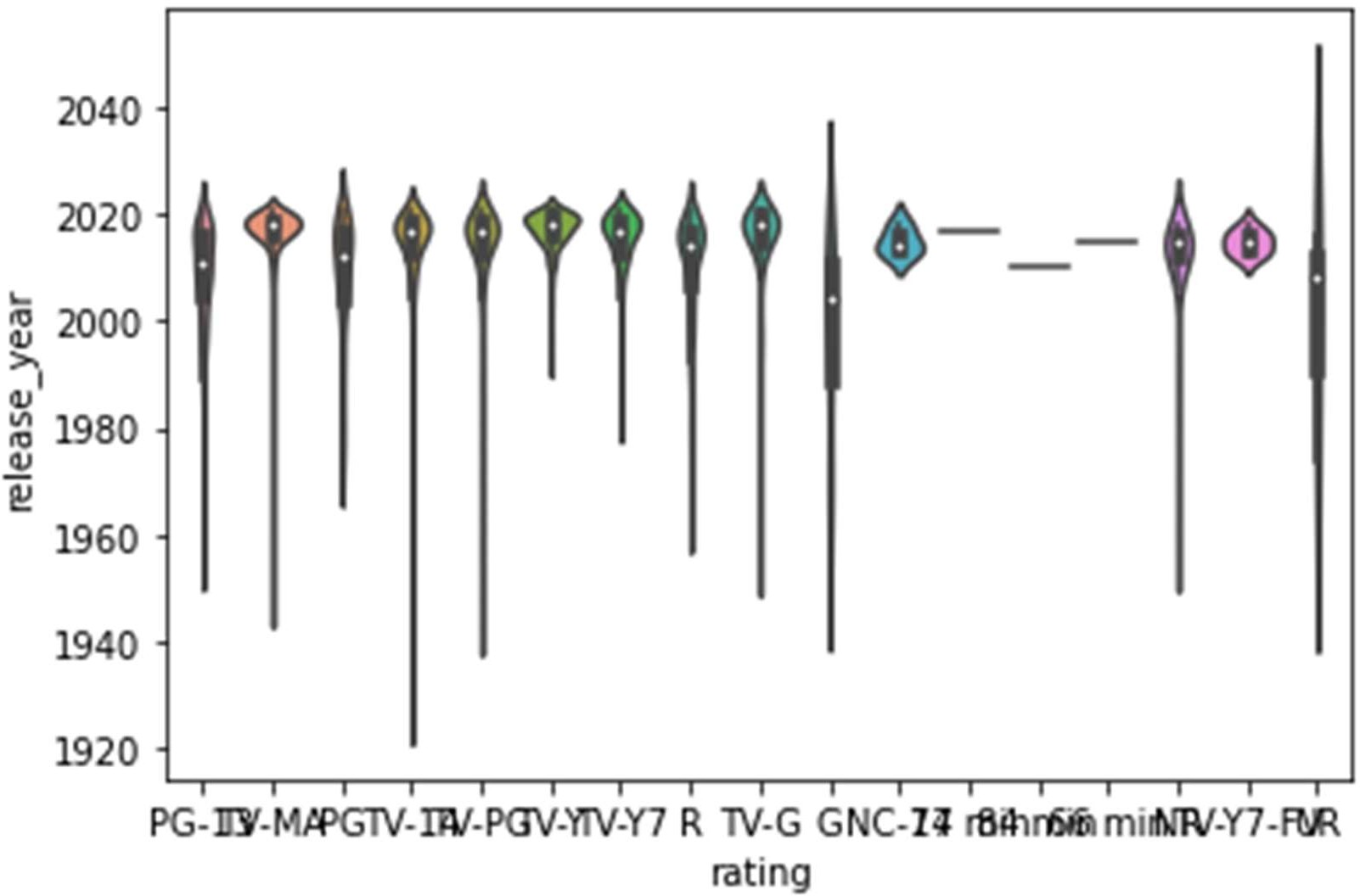
Countplot-

o/p-

sb.countplot(x="rating",data=df) plt.show()



violin plot-

o/p-

sb.violinplot(y="release\_year",x="rating",data=df) plt.show()

Question 9-Analysis the data and show which the type of

data take most unique value in a dataset.

Ans- actors has the most unique value followed by title and directors.

Code-

df.T.apply(lambda x:x.nunique(),axis=1)

o/p-

show\_id 8807

type 2

title 8807

director 4528

Actors 7692

country 748

date\_added 1767

release\_year 74

rating 17

duration 220

listed\_in 514

description 8775 dtype: int64.

Question 10- show only the Titles of all Tv show and movies that were released in india only ?

Ans – In this question be can use the Filtering to find

TV show & movies which can make and publish by

India.

Code-

df[(df['type']=='TV Show') & (df['country']=='India')]['title']

**#that code can say show the tv show**

**of india.**

o/p-

4 Kota Factory

39 Chhota Bheem

50 Dharmakshetra

66 Raja Rasoi Aur Anya Kahaniyan

69 Stories by Rabindranath Tagore

...

8173 Thackeray

8235 The Calling

8321 The Golden Years with Javed Akhtar

8349 The House That Made Me

8775 Yeh Meri Family

Name: title, Length: 79, dtype: object

**#that code can say show the movie of india.**

**Code-**

df[(df['type']=='Movie') & (df['country']=='India')]['title']

o/p-

24 Jeans

105 Angamaly Diaries

114 Anjaam

116 Dhanak

118 Gurgaon

...

8772 Yamla Pagla Deewana 2

8773 Yanda Kartavya Aahe

8798 Zed Plus

8799 Zenda

8806 Zubaan

Name: title, Length: 893, dtype: object

**Question 11- show top 10 directors ,who gave the highest number of Tv show & movies to netflix?**

**Ans-** in this question be can use the filtering to find the top 10 director who gave highest tv show and movie on Netflix.

Code-

df['director'].value\_counts().head()

o/p-

Rajiv Chilaka 19

Raúl Campos, Jan Suter 18

Marcus Raboy 16

Suhas Kadav 16

Jay Karas 14

Cathy Garcia-Molina 13

Martin Scorsese 12

Youssef Chahine 12

Jay Chapman 12

Steven Spielberg 11

Name: director, dtype: int64

***Thank You for Watching..*🤗**