SCALER NETFLIX - DATA EXPLORATION AND VISUALISATION
BUSINESS CASE STUDY

AUTHOR: DHANANJAY YADAV | DSML FEB 2023

CASE_STUDY COLAB
LINK: COLAB LINK

EVALUATION CRITERIA AND POINTS:

---

1. Defining Problem Statement and Analyzing basic metrics 10 Points

2. Observations on the shape of data, data types of all the attributes, conversion of categoricalattributes to 'category' (If required), missing value detection, statistical summary 10 Points

3. Non-Graphical Analysis: Value counts and unique attributes 10 Points

4. Visual Analysis - Univariate, Bivariate after pre-processing of the data

(Note: Pre-processing involves unnesting of the data in columns like Actor, Director, Country)

- **4.1 For continuous variable(s): Distplot, countplot, histogram for univariate analysis 10 Points**
- **4.2 For categorical variable(s): Boxplot 10 Points**
- **4.3 For correlation: Heatmaps, Pairplots 10 Points**

5. Missing Value & Outlier check (Treatment optional) 10 Points

6.0 Insights based on Non-Graphical and Visual Analysis 10 Points

- **6.1 Comments on the range of attributes**
- **6.2 Comments on the distribution of the variables and relationship between them**
- **6.3 Comments for each univariate and bivariate plot**

7. Business Insights 10 Points Should include patterns observed in the data along with what you caninfer from it

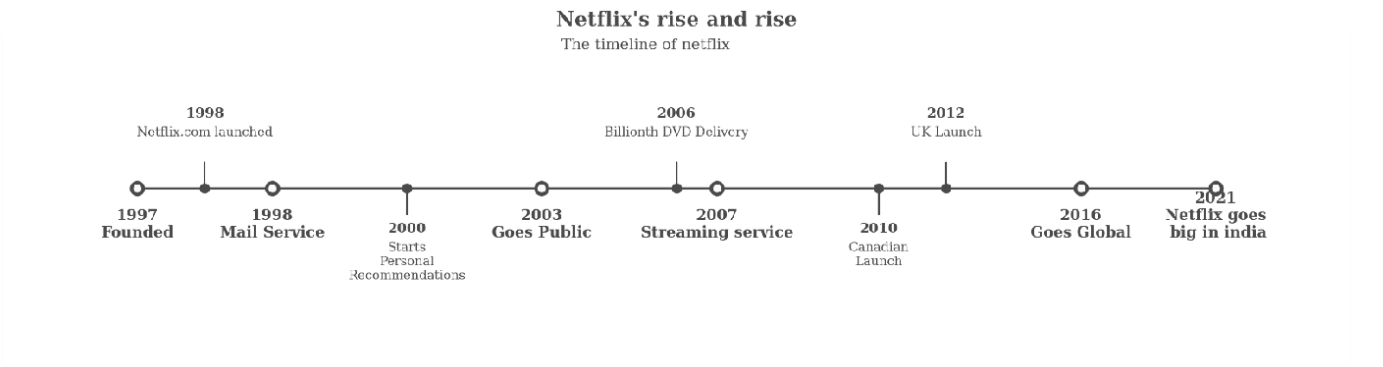
8 Recommendations 10 Points - Actionable items for business. No technical jargon. Nocomplications. Simple action items that everyone can understand.

ABOUT NETFLIX

For years, Netflix has been known for its innovation and early-entry into different segments of the entertainment and media industry. Netflix began as the first DVD-by-mail service, but soon became the first in other categories as well including a DVD subscription plan and then a streaming service. Over 50 percent of adults living in the United States currently have a Netflix subscription (Steveliesman). Netflix has been a front runner in the streaming industry for years, but their success has attracted new competitors to join the streaming industry. Disney+ is a new entrant that has taken the streaming industry by storm. Although Netflix is still very popular and successful, without strategic planning and the addition of new features and content, Netflix could lose its hold of the streaming industry.

Netflix was formed in California in 1997 by Reed Hastings and Marc Randolph. Hastings and Randolph came up with the idea of a starting a DVD rental service via mail. Netflix.com was launched in 1998 as the first DVD rental site. A year later, Netflix debuted its subscription service that offered its members unlimited DVD rentals on a monthly basis. In 2002, Netflix made its IPO at a selling price of one dollar per share under the NASDAQ ticker, NFLX. A milestone of one million accounts was reachedby 2003, and the number of subscribers doubled over the next few years. In 2007, Netflix introduced its streaming service.

Their introduction of the streaming service became a turning point for Netflix. They became a sensation in the entertainment and media industry. Netflix thrived through its mail system and then grew exponentially with its streaming service. Netflix was innovative; they offered a totally new approach to watch movies and television series. They eventually became so popular that traditional DVD rental leaders, like Blockbuster, simply could not compete and were forced to file for bankruptcy. A few years later, Netflix expanded into the international market by offering their service in Canada. In 2013, Netflix expanded operations and began producing original content, like ‘House of Cards’ and ‘Orange Is the New Black.’ That same year, Netflix received 31 primetime Emmy nominations. They were the first ever internet TV network to be nominated for a Primetime Emmy (McFadden). Netflix continued to expand into many new territories over the next few years, while offering content in 62 different languages. Netflix is 1 currently available in 190 different countries (Where is Netflix available?). It also has 193 million subscribers as of July 2020 (Moody). Netflix has grown into a multibillion-dollar company



Business Problem:

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

1. Defining Problem Statement and Analysing basic metrics:

Problem Statement:

Netflix is a multi-national streaming company which produces movies and tv web series all around the year and all around the globe

- This project aims to build a movie and tv shows recommendation mechanism and data analysis within Netflix.
- Analyzing the Netflix dataset and comparing India with different countries and take conclusion through visualand descriptive analysis

Basic metrics:

Importing Libraries:

```
# importing libraries

import pandas as pd
import numpy as np
import plotly


import matplotlib as mpl
import matplotlib.pyplot as plt


import seaborn as sns


plt.rcParams['figure.dpi'] = 200
```

Loading the Dataset:

```
df=pd.read_csv("/netflix.csv")
```

First 10 values:

```
df.head(10)
```

	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...
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	NaN	September 24, 2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	The arrival of a charismatic young priest brin...
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	NaN	September 24, 2021	2021	PG	91 min	Children & Family Movies	Equestria's divided. But a bright-eyed hero be...
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies	On a photo shoot in Ghana, an American model s...
8	s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Ho...	United Kingdom	September 24, 2021	2021	TV-14	9 Seasons	British TV Shows, Reality TV	A talented batch of amateur bakers face off in...
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	September 24, 2021	2021	PG-13	104 min	Comedies, Dramas	A woman adjusting to life after a loss contend...

Objects summary:

```
df.describe(include=[np.object])
```

```
<ipython-input-28-554b7518cb2b>:1: DeprecationWarning: `np.object` is a deprecated alias for the builtin `object`. To silence this warning, use `object
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  df.describe(include=[np.object])
```

	show_id	type	title	director	cast	country	date_added	rating	duration	listed_in	description
count	8807	8807	8807	6173	7982	7976	8797	8803	8804	8807	8807
unique	8807	2	8804	4528	7692	748	1767	17	220	514	8775
top	s1	Movie	15-Aug	Rajiv Chilaka	David Attenborough	United States	January 1, 2020	TV-MA	1 Season	Dramas, International Movies	Paranormal activity at a lush, abandoned prope...
freq	1	6131	2	19	19	2818	109	3207	1793	362	4

Number summary:

```
df.describe(include=[np.number])
```

	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

2. Observations on the shape of data, data types of all the attributes, conversion of categoricalattributes to 'category' (If required), missing value detection, statistical summary

Shape of the dataset:

```
df.shape
```

```
(8807, 12)
```

Print the name of columns:

```
df.columns
```

```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

length of data:

```
len(df)
```

```
8807
```

checking datatypes:

```
df.dtypes
```

```
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
dtype: object
```

```
df.info
```

```
<bound method DataFrame.info of      show_id     type                  title         director  \
0         s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1         s2  TV Show           Blood & Water              NaN
2         s3  TV Show               Ganglands  Julien Leclercq
3         s4  TV Show  Jailbirds New Orleans              NaN
4         s5  TV Show            Kota Factory              NaN
...      ...      ...                     ...              ...
8802   s8803    Movie                 Zodiac    David Fincher
8803   s8804  TV Show             Zombie Dumb              NaN
8804   s8805    Movie              Zombieland  Ruben Fleischer
8805   s8806    Movie                    Zoom     Peter Hewitt
8806   s8807    Movie                  Zubaan      Mozez Singh

                                                  cast        country  \
0                                                  NaN  United States
1     Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2     Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                  NaN            NaN
4     Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India
...                                                ...            ...
8802  Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...  United States
8803                                               NaN            NaN
8804  Jesse Eisenberg, Woody Harrelson, Emma Stone, ...  United States
8805  Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...  United States
8806  Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...          India
```

```
0                                    Documentaries
1       International TV Shows, TV Dramas, TV Mysteries
2     Crime TV Shows, International TV Shows, TV Act...
3                             Docuseries, Reality TV
4     International TV Shows, Romantic TV Shows, TV ...
...                                                 ...
8802                 Cult Movies, Dramas, Thrillers
8803          Kids' TV, Korean TV Shows, TV Comedies
8804                        Comedies, Horror Movies
8805              Children & Family Movies, Comedies
8806    Dramas, International Movies, Music & Musicals

                                            description
0     As her father nears the end of his life, filmm...
1     After crossing paths at a party, a Cape Town t...
2     To protect his family from a powerful drug lor...
3     Feuds, flirtations and toilet talk go down amo...
4     In a city of coaching centers known to train I...
...                                                 ...
8802  A political cartoonist, a crime reporter and a...
8803  While living alone in a spooky town, a young g...
8804  Looking to survive in a world taken over by zo...
8805  Dragged from civilian life, a former superhero...
8806  A scrappy but poor boy worms his way into a ty...
```

```
[8807 rows x 12 columns]>
```

Make a copy of the dataset:(just observation on dropping null values)

```
df1 = df.copy()
```

```
[19] df1.shape
```

```
(8807, 12)
```

Drop NULL values:

```
df1=df1.dropna()
df1.shape
```

```
(5332, 12)
```

Print first 10 values:

```
df1.head(10)
```

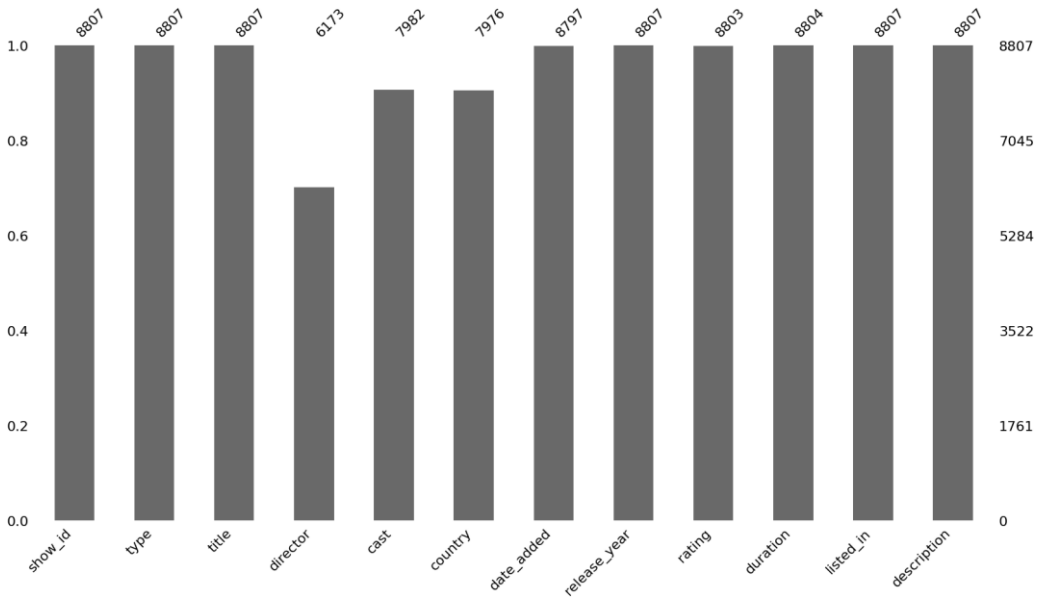
	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies	On a photo shoot in Ghana, an American model s...
8	s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Ho...	United Kingdom	September 24, 2021	2021	TV-14	9 Seasons	British TV Shows, Reality TV	A talented batch of amateur bakers face off in...
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	September 24, 2021	2021	PG-13	104 min	Comedies, Dramas	A woman adjusting to life after a loss contend...
12	s13	Movie	Je Suis Karl	Christian Schwochow	Luna Wedler, Jannis Niewöhner, Milan Peschel, ...	Germany, Czech Republic	September 23, 2021	2021	TV-MA	127 min	Dramas, International Movies	After most of her family is murdered in a terr...
24	s25	Movie	Jeans	S. Shankar	Prashanth, Aishwarya Rai	India	September	1998	TV-14	166 min	Comedies, International	When the father of the man she loves

MISSING VALUES:

```
df.isnull().sum()/len(df)*100
```

```
show_id         0.000000
type            0.000000
title           0.000000
director       29.908028
cast            9.367549
country         9.435676
date_added      0.113546
release_year    0.000000
rating          0.045418
duration        0.034064
listed_in       0.000000
description     0.000000
dtype: float64
```

```
import missingno as msno
msno.bar(df, figsize=(20,10))
plt.show()
```



On the left side of the plot, the y-axis scale ranges from 0.0 to 1.0, where 1.0 represents 100% data completeness. If the bar isless than this, it indicates that we have missing values within that column.

On the right side of the plot, the scale is measured in index values. With the top right representing the maximum number ofrows within the data frame.

Along the top of the plot, there are a series of numbers that represent the total count of the non-null values within thatcolumn.

"In the above plot director, cast and country have more missing value compared to other like rating, date added, duration"

The statsistical summary:(total 11 objects and one number)Objects summery:

```
df.describe(include=[np.object])
```

```
<ipython-input-28-554b7518cb2b>:1: DeprecationWarning: `np.object` is a deprecated alias for the builtin `object`. To silence this warning, use `object
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  df.describe(include=[np.object])
```

	show_id	type	title	director	cast	country	date_added	rating	duration	listed_in	description
count	8807	8807	8807	6173	7982	7976	8797	8803	8804	8807	8807
unique	8807	2	8804	4528	7692	748	1767	17	220	514	8775
top	s1	Movie	15-Aug	Rajiv Chilaka	David Attenborough	United States	January 1, 2020	TV-MA	1 Season	Dramas, International Movies	Paranormal activity at a lush, abandoned prope...
freq	1	6131	2	19	19	2818	109	3207	1793	362	4

Number summary:

```
df.describe(include=[np.number])
```

	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

conversion of categorical attributes to 'category':

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 15 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   datetime64[ns]
 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        
 12  day_added     8797 non-null   float64       
 13  year_added    8797 non-null   float64       
 14  month_added   8797 non-null   float64       
dtypes: datetime64[ns](1), float64(3), int64(1), object(10)
memory usage: 1.0+ MB
```

3. Non-Graphical Analysis: Value counts and unique attributes:

Check unique values:

```
df.nunique()
```

```
show_id         8807
type               2
title           8804
director        4528
cast            7692
country          748
date_added      1767
release_year      74
rating            17
duration         220
listed_in        514
description     8775
dtype: int64
```

Check for Duplicate values:

```
df.duplicated().sum()
```

```
0
```

```
for i in df.columns:
 print("The Unique values and nuniques of ",str(i),"are",df[i].unique(),df[i].nunique())
 print("...................")
```

```
The Unique values and nuniques of  show_id are ['s1' 's2' 's3' ... 's8805' 's8806' 's8807'] 8807
..................
The Unique values and nuniques of  type are ['Movie' 'TV Show'] 2
..................
The Unique values and nuniques of  title are ['Dick Johnson Is Dead' 'Blood & Water' 'Ganglands' ... 'Zombieland'
 'Zoom' 'Zubaan'] 8804
..................
The Unique values and nuniques of  director are ['Kirsten Johnson' nan 'Julien Leclercq' ... 'Majid Al Ansari'
 'Peter Hewitt' 'Mozez Singh'] 4528
..................
The Unique values and nuniques of  cast are [nan
 'Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Molaba, Dillon Windvogel, Natasha Thahane, Arno Greeff, Xolile Tshabalala, Getmore Sithole, Cindy N
 'Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabiha Akkari, Sofia Lesaffre, Salim Kechiouche, Noureddine Farihi, Geert Van Rampelberg, Bakary Diombera
 ...
 'Jesse Eisenberg, Woody Harrelson, Emma Stone, Abigail Breslin, Amber Heard, Bill Murray, Derek Graf'
 'Tim Allen, Courteney Cox, Chevy Chase, Kate Mara, Ryan Newman, Michael Cassidy, Spencer Breslin, Rip Torn, Kevin Zegers'
 'Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanana, Manish Chaudhary, Meghna Malik, Malkeet Rauni, Anita Shabdish, Chittaranjan Tripathy'] 7692
```

```
df.columns
```

```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description',
       'day_added', 'year_added', 'month_added'],
      dtype='object')
```

Value counts:

Value counts of "type":

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

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

Value counts of "show_id":

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

```
s1       1
s5875    1
s5869    1
s5870    1
s5871    1
        ..
s2931    1
s2930    1
s2929    1
s2928    1
s8807    1
Name: show_id, Length: 8807, dtype: int64
```

Value counts of "title":

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

```
15-Aug                                  2
Feb-09                                  2
22-Jul                                  2
The Ridiculous 6                        1
Mike Epps: Don't Take It Personal       1
                                       ..
Good Time                               1
Captain Underpants Epic Choice-o-Rama   1
We Bare Bears                           1
To All the Boys: P.S. I Still Love You  1
Zubaan                                  1
Name: title, Length: 8804, dtype: int64
```

Value counts of "director":

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

```
Rajiv Chilaka                     19
Raúl Campos, Jan Suter            18
Marcus Raboy                      16
Suhas Kadav                       16
Jay Karas                         14
                                  ..
Raymie Muzquiz, Stu Livingston     1
Joe Menendez                       1
Eric Bross                         1
Will Eisenberg                     1
Mozez Singh                        1
Name: director, Length: 4528, dtype: int64
```

Value counts of "country":

```
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
```

Value counts of "date_added":

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

```
January 1, 2020      109
November 1, 2019      89
March 1, 2018         75
December 31, 2019     74
October 1, 2018       71
                     ...
December 4, 2016       1
November 21, 2016      1
November 19, 2016      1
November 17, 2016      1
January 11, 2020       1
Name: date_added, Length: 1767, dtype: int64
```

Value counts of "release_year":

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

```
2018    1147
2017    1032
2019    1030
2020     953
2016     902
        ...
1959       1
1925       1
1961       1
1947       1
1966       1
Name: release_year, Length: 74, dtype: int64
```

Value counts of "rating":

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

```
TV-MA       3207
TV-14       2160
TV-PG        863
R            799
PG-13        490
TV-Y7        334
TV-Y         307
PG           287
TV-G         220
NR            80
G             41
TV-Y7-FV       6
NC-17          3
UR             3
74 min         1
84 min         1
66 min         1
Name: rating, dtype: int64
```

Value counts of "duration":

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

```
1 Season     1793
2 Seasons     425
3 Seasons     199
90 min        152
94 min        146
             ...
16 min          1
186 min         1
193 min         1
189 min         1
191 min         1
Name: duration, Length: 220, dtype: int64
```

Value counts of "listed_in":

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

```
Dramas, International Movies                            362
Documentaries                                           359
Stand-Up Comedy                                         334
Comedies, Dramas, International Movies                  274
Dramas, Independent Movies, International Movies        252
                                                       ...
Kids' TV, TV Action & Adventure, TV Dramas                1
TV Comedies, TV Dramas, TV Horror                         1
Children & Family Movies, Comedies, LGBTQ Movies          1
Kids' TV, Spanish-Language TV Shows, Teen TV Shows        1
Cult Movies, Dramas, Thrillers                            1
Name: listed_in, Length: 514, dtype: int64
```

Value counts of "description":

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

```
Paranormal activity at a lush, abandoned property alarms a group eager to redevelop the site, but the eerie events may not be as unearthly as they think.    4
Challenged to compose 100 songs before he can marry the girl he loves, a tortured but passionate singer-songwriter embarks on a poignant musical journey.    3
A surly septuagenarian gets another chance at her 20s after having her photo snapped at a studio that magically takes 50 years off her life.
3
Multiple women report their husbands as missing but when it appears they are looking for the same man, a police officer traces their cryptic connection.     3
Secrets bubble to the surface after a sensual encounter and an unforeseen crime entangle two friends and a woman caught between them.
2


..
Sent away to evade an arranged marriage, a 14-year-old begins a harrowing journey of sex work and poverty in the slums of Accra.
1
When his partner in crime goes missing, a small-time crook's life is transformed as he dedicates himself to raising the daughter his friend left behind.     1
During 1962's Cuban missile crisis, a troubled math genius finds himself drafted to play in a U.S.-Soviet chess match – and a deadly game of espionage.      1
A teen's discovery of a vintage Polaroid camera develops into a darker tale when she finds that whoever takes their photo with it dies soon afterward.
1
A scrappy but poor boy worms his way into a tycoon's dysfunctional family, while facing his fear of music and the truth about his past.
1
Name: description, Length: 8775, dtype: int64
```

4. Visual Analysis - Univariate, Bivariate after pre-processing of the data

Unnesting data

unnesting of type:

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

	index	type
0	Movie	6131
1	TV Show	2676

unnesting of show id:

```
df['show_id'].value_counts().reset_index() # value count of show id
```

	index	show_id
0	s1	1
1	s5875	1
2	s5869	1
3	s5870	1
4	s5871	1
...	...	...
8802	s2931	1
8803	s2930	1
8804	s2929	1
8805	s2928	1
8806	s8807	1

8807 rows × 2 columns

unnesting of title:

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

	index	title
0	15-Aug	2
1	Feb-09	2
2	22-Jul	2
3	The Ridiculous 6	1
4	Mike Epps: Don't Take It Personal	1
...	...	...
8799	Good Time	1
8800	Captain Underpants Epic Choice-o-Rama	1
8801	We Bare Bears	1
8802	To All the Boys: P.S. I Still Love You	1
8803	Zubaan	1

8804 rows × 2 columns

unnesting of director:

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

	index	director
0	Rajiv Chilaka	19
1	Raúl Campos, Jan Suter	18
2	Marcus Raboy	16
3	Suhas Kadav	16
4	Jay Karas	14
...	...	...
4523	Raymie Muzquiz, Stu Livingston	1
4524	Joe Menendez	1
4525	Eric Bross	1
4526	Will Eisenberg	1
4527	Mozez Singh	1

4528 rows × 2 columns

unnesting of country:

```
df['country'].value_counts().reset_index()
```

	index	country
0	United States	2818
1	India	972
2	United Kingdom	419
3	Japan	245
4	South Korea	199
...	...	...
743	Romania, Bulgaria, Hungary	1
744	Uruguay, Guatemala	1
745	France, Senegal, Belgium	1
746	Mexico, United States, Spain, Colombia	1
747	United Arab Emirates, Jordan	1

748 rows × 2 columns

unnesting of date_added:

```
df['date_added'].value_counts().reset_index()
```

	index	date_added
0	2020-01-01	110
1	2019-11-01	91
2	2018-03-01	75
3	2019-12-31	74
4	2018-10-01	71
...	...	...
1709	2017-02-21	1
1710	2017-02-07	1
1711	2017-01-29	1
1712	2017-01-25	1
1713	2020-01-11	1

1714 rows × 2 columns

unnesting of year column:

```
df['release_year'].value_counts().reset_index()
```

	index	release_year
0	2018	1147
1	2017	1032
2	2019	1030
3	2020	953
4	2016	902
...	...	...
69	1959	1
70	1925	1
71	1961	1
72	1947	1
73	1966	1

74 rows × 2 columns

unnesting of rating column:

```
df['rating'].value_counts().reset_index()
```

	index	rating
0	TV-MA	3207
1	TV-14	2160
2	TV-PG	863
3	R	799
4	PG-13	490
5	TV-Y7	334
6	TV-Y	307
7	PG	287
8	TV-G	220
9	NR	80
10	G	41
11	TV-Y7-FV	6
12	NC-17	3

13	UR	3
14	74 min	1
15	84 min	1
16	66 min	1

unnesting of duration:

```
df['duration'].value_counts().reset_index()
```

	index	duration
0	1 Season	1793
1	2 Seasons	425
2	3 Seasons	199
3	90 min	152
4	94 min	146
...	...	...
215	16 min	1
216	186 min	1
217	193 min	1
218	189 min	1
219	191 min	1

220 rows × 2 columns

unnesting of listed_in column:

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

	index	listed_in
0	Dramas, International Movies	362
1	Documentaries	359
2	Stand-Up Comedy	334
3	Comedies, Dramas, International Movies	274
4	Dramas, Independent Movies, International Movies	252
...	...	...
509	Kids' TV, TV Action & Adventure, TV Dramas	1
510	TV Comedies, TV Dramas, TV Horror	1
511	Children & Family Movies, Comedies, LGBTQ Movies	1
512	Kids' TV, Spanish-Language TV Shows, Teen TV S...	1
513	Cult Movies, Dramas, Thrillers	1

514 rows × 2 columns

unnesting of description:

```
df['description'].value_counts().reset_index()
```

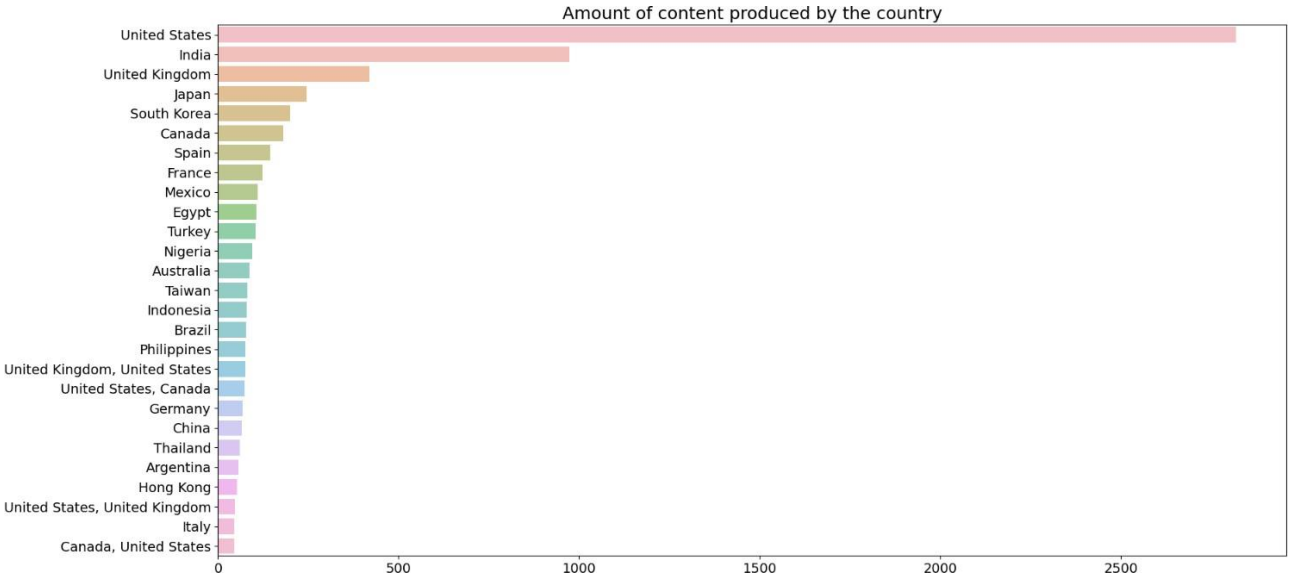
	index	description
0	Paranormal activity at a lush, abandoned prope...	4
1	Challenged to compose 100 songs before he can ...	3
2	A surly septuagenarian gets another chance at ...	3
3	Multiple women report their husbands as missin...	3
4	Secrets bubble to the surface after a sensual ...	2
...	...	...
8770	Sent away to evade an arranged marriage, a 14-...	1
8771	When his partner in crime goes missing, a smal...	1
8772	During 1962's Cuban missile crisis, a troubled...	1
8773	A teen's discovery of a vintage Polaroid camer...	1
8774	A scrappy but poor boy worms his way into a ty...	1

8775 rows × 2 columns

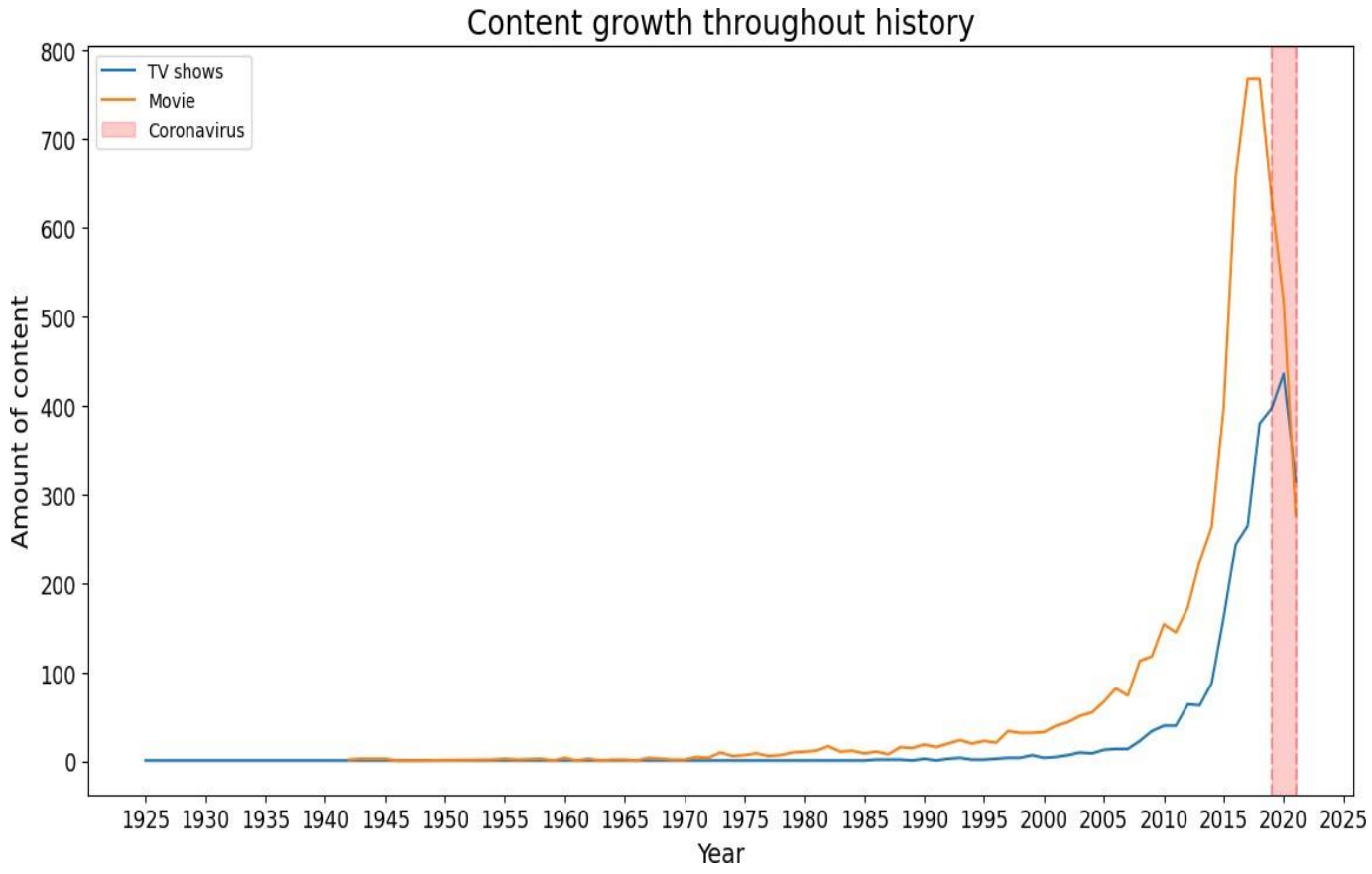
4.1 For continuous variable(s): Distplot, countplot, histogram for univariate analysis:barplotting the number of content per each country:

```
countries = df['country'].value_counts()[df['country'].value_counts(normalize=True)> 0.005]
list_countries = list(countries.index)
```

```
plt.figure(figsize=(20,10))
plt.title('Amount of content produced by the country', fontsize=18)
plt.tick_params(labelsize=14)
sns.barplot(y=countries.index, x=countries.values, alpha=0.6)
plt.show()
```

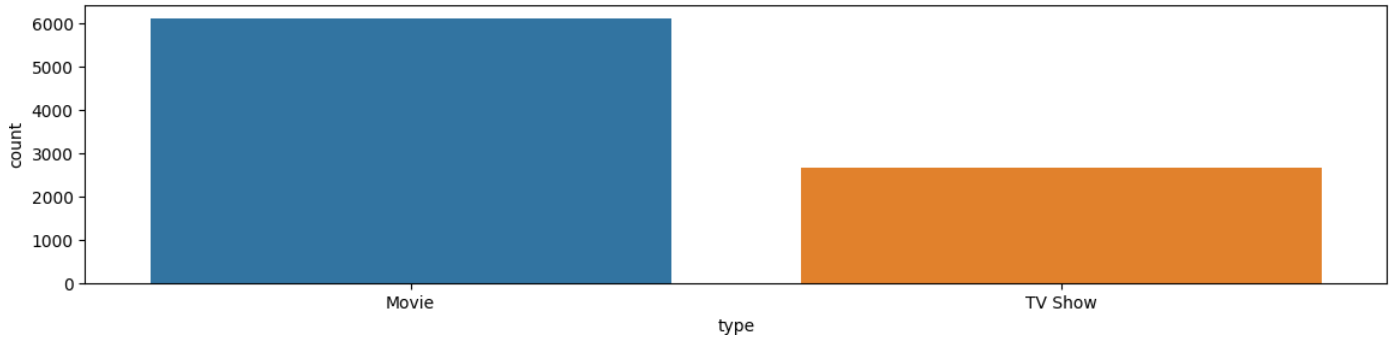


```
[51] TVshows = df[df['type'] == 'TV Show']
     Movie = df[df['type'] == 'Movie']
     TVshows_progress = TVshows['release_year'].value_counts().sort_index()
     Movie_progress = Movie['release_year'].value_counts().sort_index()
     plt.figure(figsize=(14, 7))
     plt.plot(TVshows_progress.index, TVshows_progress.values, label='TV shows')
     plt.plot(Movie_progress.index, Movie_progress.values, label='Movie')
     plt.axvline(2019, alpha=0.3, linestyle='--', color='r')
     plt.axvline(2021, alpha=0.3, linestyle='--', color='r')
     plt.axvspan(2019, 2021, alpha=0.2, color='r', label='Coronavirus')
     plt.xticks(list(range(1925, 2026, 5)), fontsize=12)
     plt.title('Content growth throughout history', fontsize=18)
     plt.xlabel('Year', fontsize=14)
     plt.ylabel('Amount of content', fontsize=14)
     plt.yticks(fontsize=12)
     plt.legend()
     plt.show()
```



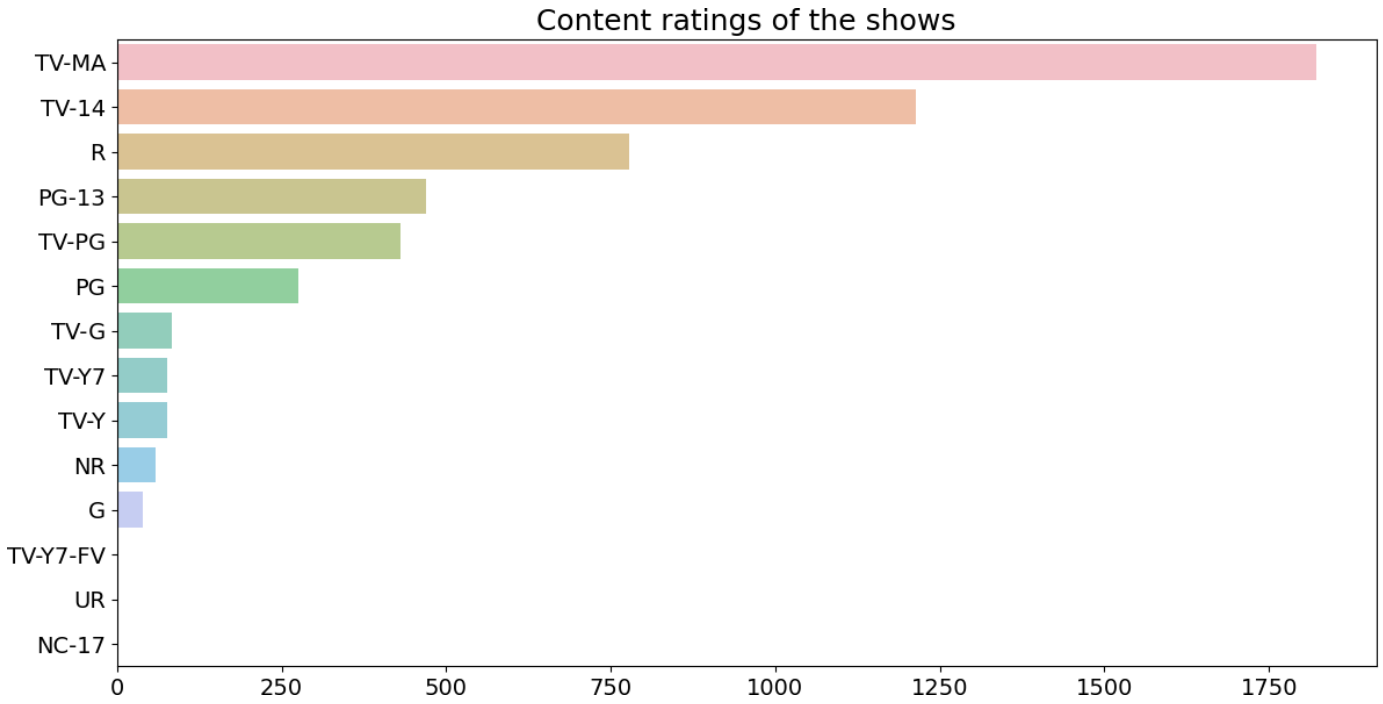
Countplot:

```
plt.figure(figsize=(14, 3))
sns.countplot(x='type',data = df)
```

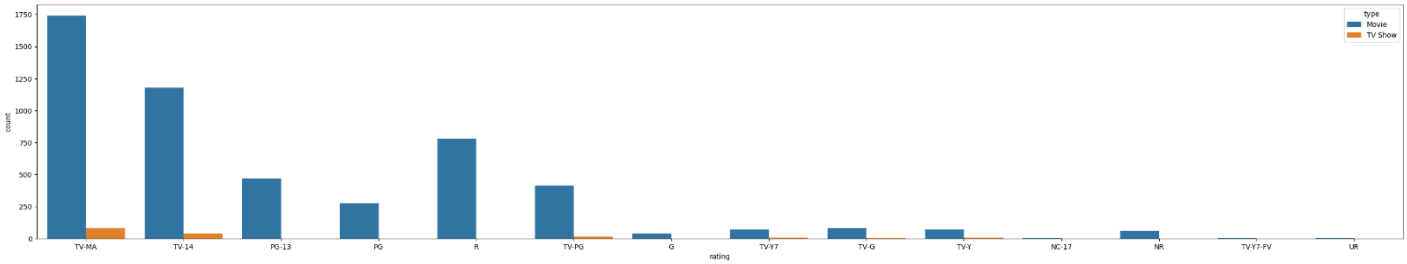


Barplot:

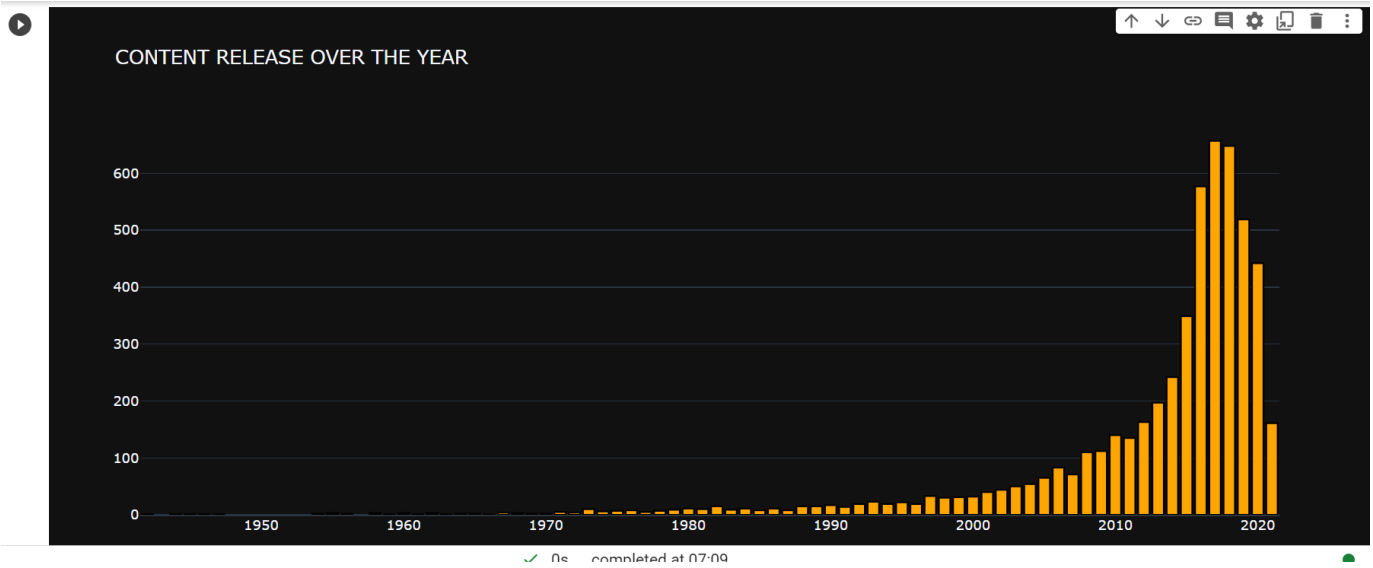
```
df.dropna(inplace=True)
rating = df['rating'].value_counts()
plt.figure(figsize=(14,7))
plt.title('Content ratings of the shows', fontsize=18)
plt.tick_params(labelsize=14)
sns.barplot(y=rating.index, x=rating.values, alpha=0.6)
```



```
plt.figure(figsize = (35,6))
sns.countplot(x='rating',data = df,hue='type')
```



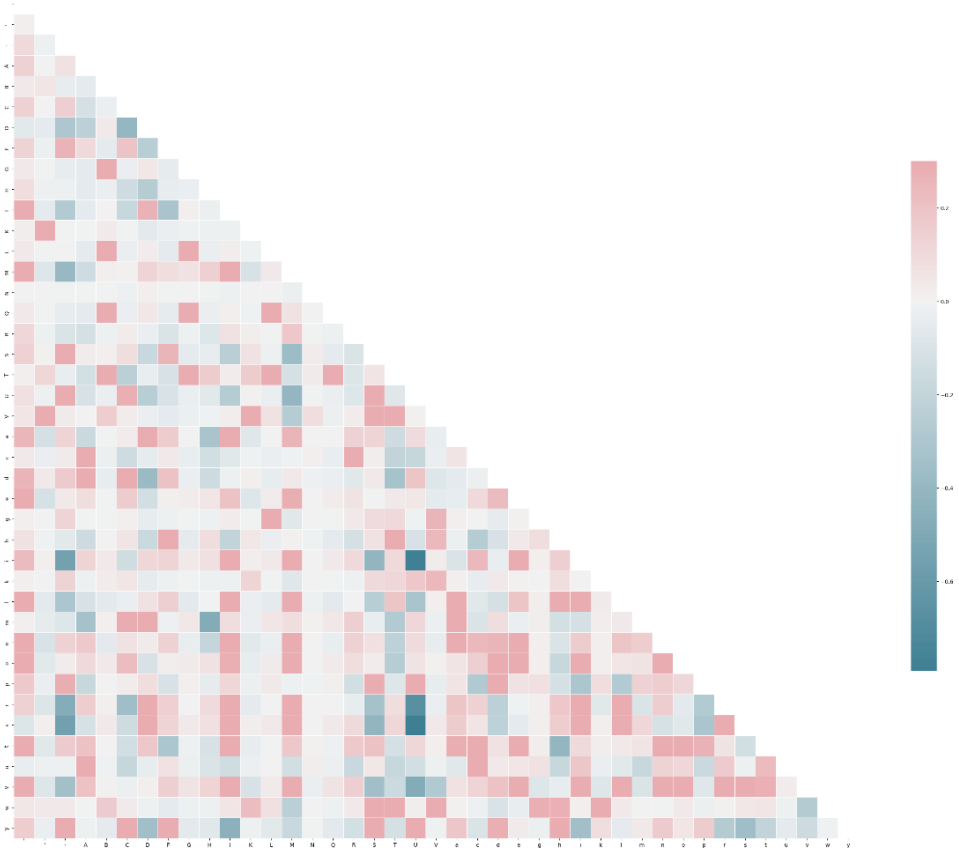

```
temp_df1 = df['release_year'].value_counts().reset_index()
import plotly.graph_objects as go
trace1 = go.Bar(
 x = temp_df1['index'],
 y = temp_df1['release_year'],
 marker = dict(color = 'rgb(255,165,0)',
 line=dict(color='rgb(0,0,0)',width=1.5)))
layout = go.Layout(template= "plotly_dark",title = 'CONTENT RELEASE OVER THE YEAR')
fig = go.Figure(data = [trace1], layout = layout)
fig.show()
```



0s completed at 07:09

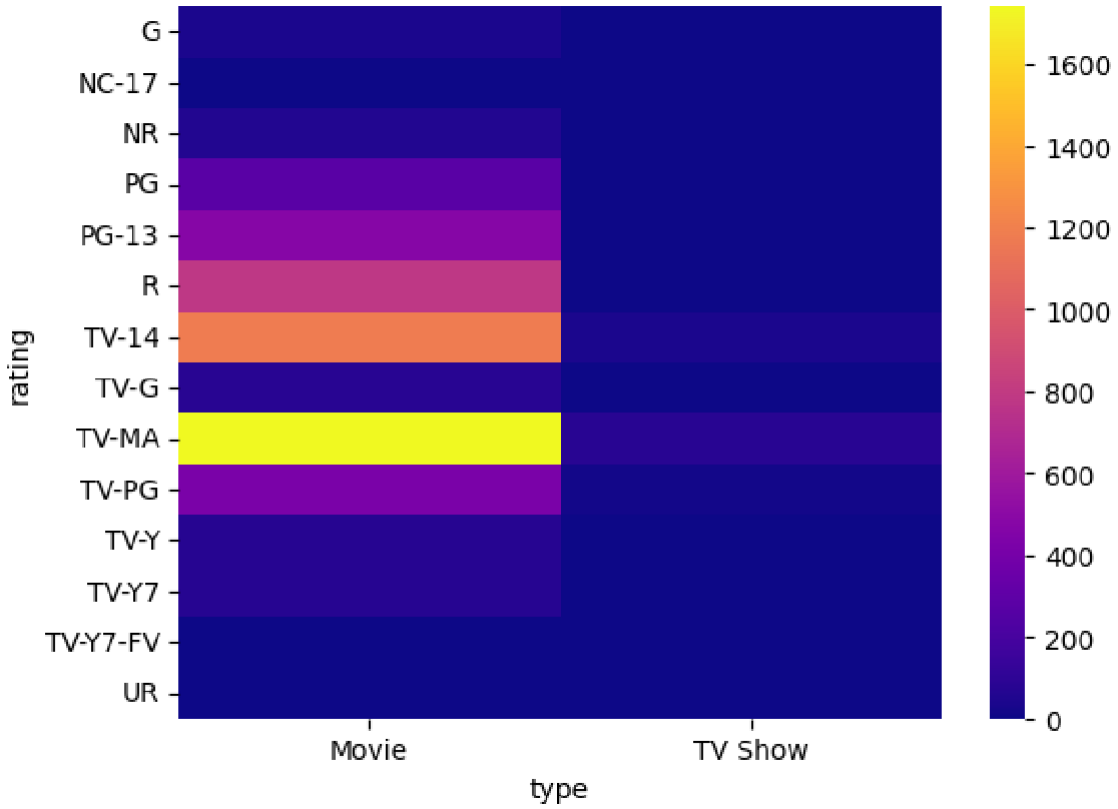
```
# bold('**HEATMAP(Correlation)**')

from sklearn.preprocessing import MultiLabelBinarizer # Similar to One-Hot Encoding
data= df['listed_in'].astype(str).apply(lambda s : s.replace('&',' ').replace(',', ' '))
test = data
mlb = MultiLabelBinarizer()
res = pd.DataFrame(mlb.fit_transform(test), columns=mlb.classes_)
corr = res.corr()
mask = np.zeros_like(corr, dtype=np.bool)
mask[np.triu_indices_from(mask)] = True
f, ax = plt.subplots(figsize=(35, 34))
cmap = sns.diverging_palette(220, 10, as_cmap=True)
sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.3, center=0,square=True, linewidths=.5,
cbar_kws={"shrink": .5})
plt.show()
```

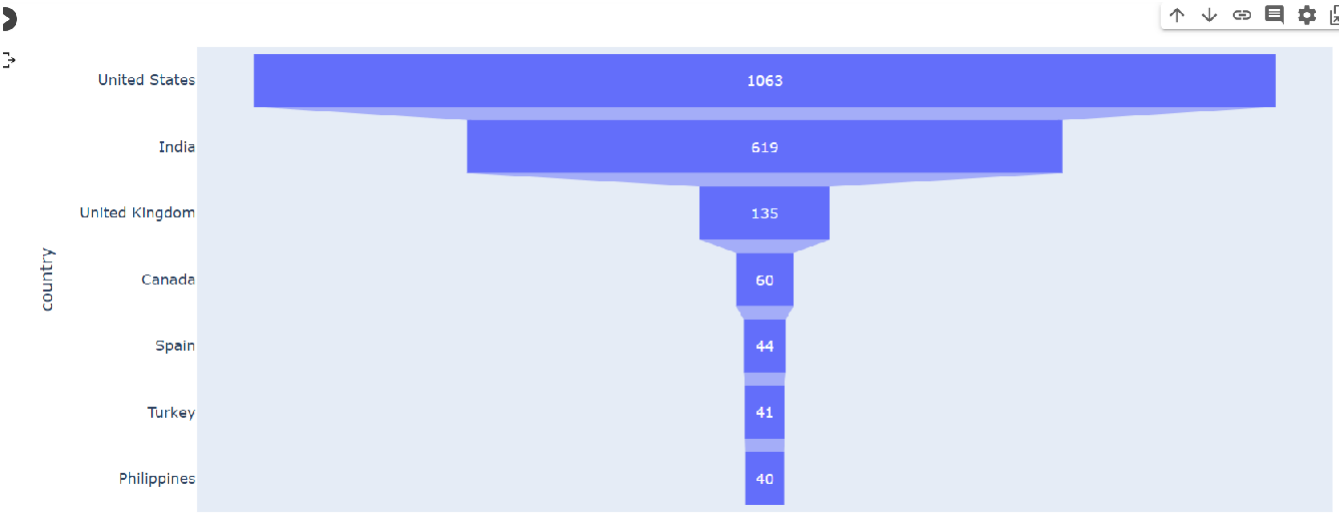


"The above graph is shows that international movies with target audience children are low"

```
colormap = plt.cm.plasma
sns.heatmap(pd.crosstab(df["rating"], df["type"]), cmap = colormap)
```



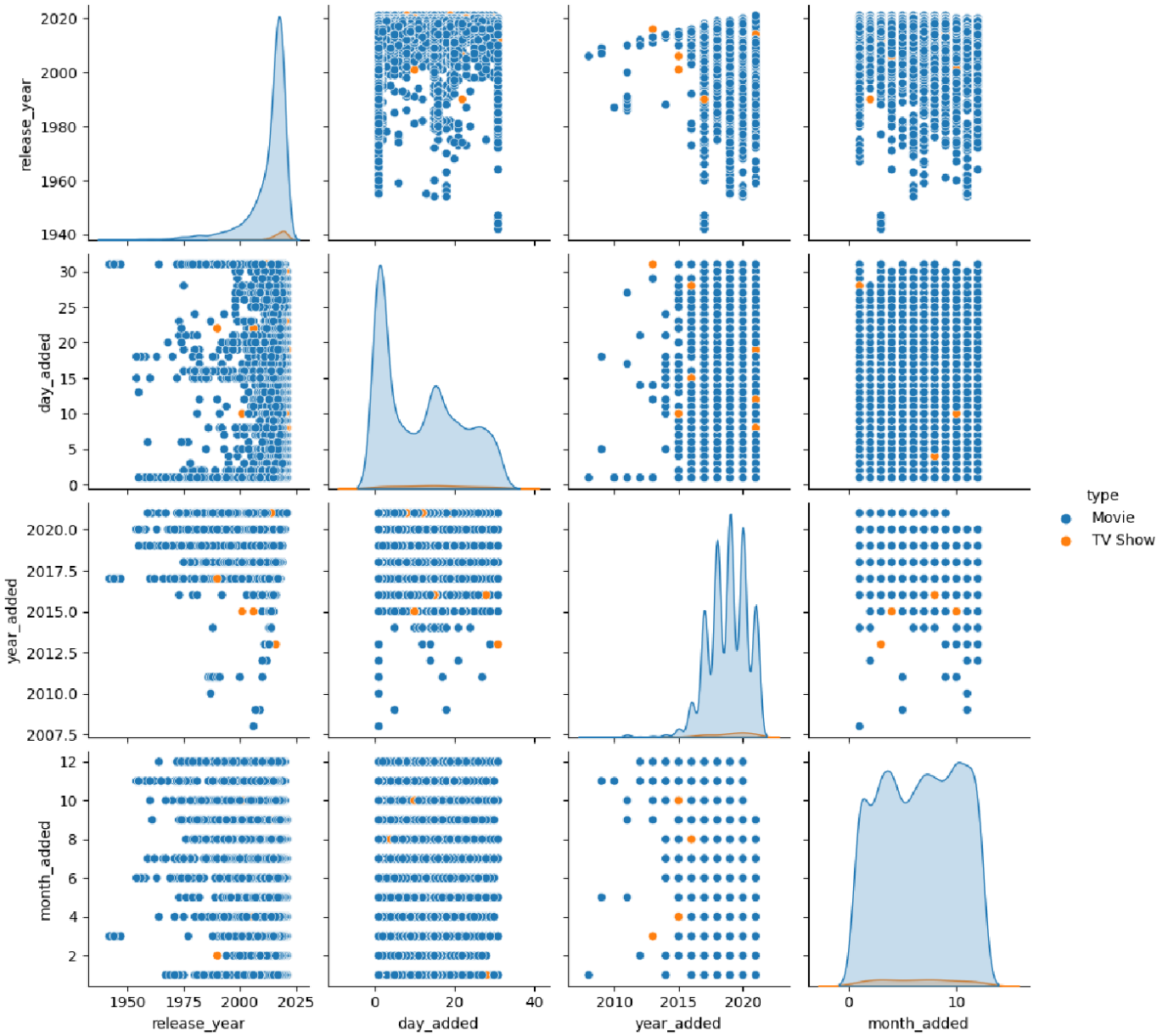

```
# plt.figure(figsize = (35,6))
df = pd.read_csv('/netflix.csv')
import plotly.express as px
data = dict(number=[1063,619,135,60,44,41,40],
 country=["United States", "India", "United Kingdom", "Canada", "Spain",'Turkey','Philippines'])
fig = px.funnel(data, x='number', y='country')
fig.show()
```



“the above graph shows most content produced by usa followed by india”

pair plot of type and released year:

```
plt.figure(figsize = (35,6))
sns.pairplot(mf,hue='type')
```

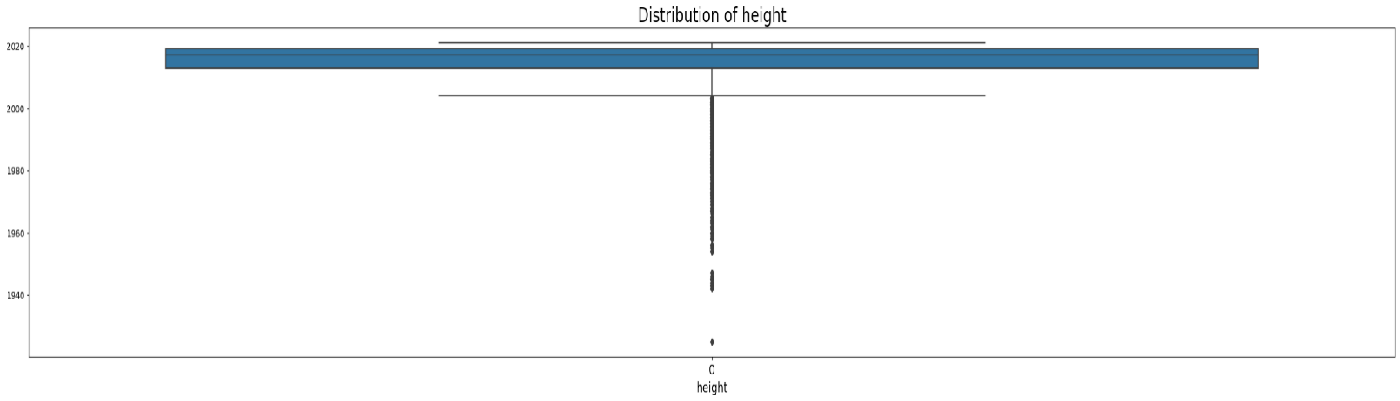


5. Missing Value & Outlier check (Treatment optional):

```
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
```

```
plt.figure(figsize = (35,6))
# box plot of the release year
ax = sns.boxplot(df['release_year'])
# notation indicating an outlier
ax.annotate('Outlier', xy=(190,0), xytext=(186,-0.05), fontsize=14,
 arrowprops=dict(arrowstyle='->', ec='grey', lw=2), bbox = dict(boxstyle="round"))
# xtick, label, and title
plt.xticks(fontsize=14)
plt.xlabel('height', fontsize=14)
plt.title('Distribution of height', fontsize=20)
```



“There is a lot of historical content which is not preferred by modern audience”

Check for NULL Values:

```
df.isnull().sum()
```

```
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
```

+ Code + Text

replacing null values:

```
df['cast'].fillna(df['cast'].mode(), inplace = True)
```

Treatment for null values:

```
[22] #unnesting the directors column, i.e- creating separate lines for each director in a movie
     constraint1=df['director'].apply(lambda x: str(x).split(', ')).tolist()
     df_new1=pd.DataFrame(constraint1,index=df['title'])
     df_new1=df_new1.stack()
     df_new1=pd.DataFrame(df_new1.reset_index())
     df_new1.rename(columns={0:'Directors'},inplace=True)
     df_new1.drop(['level_1'],axis=1,inplace=True)
     df_new1.head()
```

	title	Directors
0	Dick Johnson Is Dead	Kirsten Johnson
1	Blood & Water	nan
2	Ganglands	Julien Leclercq
3	Jailbirds New Orleans	nan
4	Kota Factory	nan

```
[23] #unnesting the cast column, i.e- creating separate lines for each cast member in a movie
     constraint2=df['cast'].apply(lambda x: str(x).split(', ')).tolist()
     df_new2=pd.DataFrame(constraint2,index=df['title'])
     df_new2=df_new2.stack()
     df_new2=pd.DataFrame(df_new2.reset_index())
     df_new2.rename(columns={0:'Actors'},inplace=True)
     df_new2.drop(['level_1'],axis=1,inplace=True)
     df_new2.head()
```

	title	Actors
0	Dick Johnson Is Dead	nan
1	Blood & Water	Ama Qamata
2	Blood & Water	Khosi Ngema
3	Blood & Water	Gail Mabalane
4	Blood & Water	Thabang Molaba

```
[24] #unnesting the listed_in column, i.e- creating separate lines for each genre in a movie
     constraint3=df['listed_in'].apply(lambda x: str(x).split(', ')).tolist()
     df_new3=pd.DataFrame(constraint3,index=df['title'])
     df_new3=df_new3.stack()
     df_new3=pd.DataFrame(df_new3.reset_index())
     df_new3.rename(columns={0:'Genre'},inplace=True)
     df_new3.drop(['level_1'],axis=1,inplace=True)
     df_new3.head()
```

	title	Genre
0	Dick Johnson Is Dead	Documentaries
1	Blood & Water	International TV Shows
2	Blood & Water	TV Dramas
3	Blood & Water	TV Mysteries
4	Ganglands	Crime TV Shows

```
#unnesting the country column, i.e- creating separate lines for each country in a movie
constraint4=df['country'].apply(lambda x: str(x).split(', ')).tolist()
df_new4=pd.DataFrame(constraint4,index=df['title'])
df_new4=df_new4.stack()
df_new4=pd.DataFrame(df_new4.reset_index())
df_new4.rename(columns={0:'country'},inplace=True)
df_new4.drop(['level_1'],axis=1,inplace=True)
df_new4.head()
```

	title	country
0	Dick Johnson Is Dead	United States
1	Blood & Water	South Africa
2	Ganglands	nan
3	Jailbirds New Orleans	nan
4	Kota Factory	India

```
[26] #merging the unnested director data with unnested actors data
     df_new5=df_new2.merge(df_new1,on=['title'],how='inner')
     #merging the above merged data with unnested genre data
     df_new6=df_new5.merge(df_new3,on=['title'],how='inner')
     #merging the above merged data with unnested country data
     df_new=df_new6.merge(df_new4,on=['title'],how='inner')


     #replacing nan values of director and actor by Unknown Actor and Director
     df_new['Actors'].replace(['nan'],['Unknown Actor'],inplace=True)
     df_new['Directors'].replace(['nan'],['Unknown Director'],inplace=True)
     df_new['country'].replace(['nan'],[np.nan],inplace=True)
     df_new.head()
```

	title	Actors	Directors	Genre	country
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa

```
[27] #merging our unnested data with the original data
     df_final=df_new.merge(df[['show_id', 'type', 'title', 'date_added',
           'release_year', 'rating', 'duration']],on=['title'],how='left')
     df_final.head()
```

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	90 min
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons

```
[28] #now checking nulls
     df_final.isnull().sum()
```

```
title           0
Actors          0
Directors       0
```

```
df_final.loc[df_final['duration'].isnull(),'duration']=df_final.loc[df_final['duration'].isnull(),'duration'].fillna(df_final['rating'])

df_final.loc[df_final['rating'].str.contains('min', na=False),'rating']='NR'

df_final.isnull().sum()
```

```
title              0
Actors             0
Directors          0
Genre              0
country        12497
show_id            0
type               0
date_added       158
release_year       0
rating            67
duration           0
dtype: int64
```

```
[30] #Ratings can't be in min, so it has been made NR(i.e- Non Rated)
     df_final.loc[df_final['rating'].str.contains('min', na=False),'rating']='NR'
     df_final['rating'].fillna('NR',inplace=True)
     pd.set_option('display.max_rows',None)
```

```
#just an attempt to observe nulls in date_added column
df_final[df_final['date_added'].isnull()].head()
```

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration
139473	A Young Doctor's Notebook and Other Stories	Daniel Radcliffe	Unknown Director	British TV Shows	United Kingdom	s6067	TV Show	NaN	2013	TV-MA	2 Seasons
139474	A Young Doctor's Notebook and Other Stories	Daniel Radcliffe	Unknown Director	TV Comedies	United Kingdom	s6067	TV Show	NaN	2013	TV-MA	2 Seasons
139475	A Young Doctor's Notebook and Other Stories	Daniel Radcliffe	Unknown Director	TV Dramas	United Kingdom	s6067	TV Show	NaN	2013	TV-MA	2 Seasons
139476	A Young Doctor's Notebook and Other Stories	Jon Hamm	Unknown Director	British TV Shows	United Kingdom	s6067	TV Show	NaN	2013	TV-MA	2 Seasons
139477	A Young Doctor's Notebook and Other Stories	Jon Hamm	Unknown Director	TV Comedies	United Kingdom	s6067	TV Show	NaN	2013	TV-MA	2 Seasons

in duration column, it was observed that the nulls had values which were written in corresponding ratings column, i.e-you can't expect ratings to be in min. So the duration column nulls are replaced by corresponding values in ratings column

```
[32] #date added column is imputed on the basis of release year,i.e- suppose there's a null for date_added
     #when release year was 2013.So below piece of code just checks the mode of date added for release year=2013
     # and imputes in place of nulls the corresponding mode

     for i in df_final[df_final['date_added'].isnull()]['release_year'].unique():
       imp=df_final[df_final['release_year']==i]['date_added'].mode().values[0]
       df_final.loc[df_final['release_year']==i,'date_added']=df_final.loc[df_final['release_year']==i,'date_added'].fillna(imp)
```

```
[33] #country column is imputed on the basis of director,i.e- suppose there's a null for country
     #when we have a director whose other movies have a country given.So below piece of code just checks the mode of
     #country for the director
     # and imputes in place of nulls the corresponding mode

     for i in df_final[df_final['country'].isnull()]['Directors'].unique():
       if i in df_final[~df_final['country'].isnull()]['Directors'].unique():
         imp=df_final[df_final['Directors']==i]['country'].mode().values[0]
         df_final.loc[df_final['Directors']==i,'country']=df_final.loc[df_final['Directors']==i,'country'].fillna(imp)
```

So we imputed the country column on the basis of directors whose other movie titles had countries given. But there might bedirectors who have only one occurence in our data. In that scenario, I have used Actors as a basis. i.e- for this Actor majorly acts in movies of which country? Imputation has been done on this basis. For remaining rows, country has been filled as Unknown Country

```
for i in df_final[df_final['country'].isnull()]['Actors'].unique():
  if i in df_final[~df_final['country'].isnull()]['Actors'].unique():
    imp=df_final[df_final['Actors']==i]['country'].mode().values[0]
    df_final.loc[df_final['Actors']==i,'country']=df_final.loc[df_final['Actors']==i,'country'].fillna(imp)
#If there are still nulls, I just replace it by Unknown Country
df_final['country'].fillna('Unknown Country',inplace=True)
df_final.isnull().sum()
```

```
title           0
Actors          0
Directors       0
Genre           0
country         0
show_id         0
type            0
date_added      0
release_year    0
rating          0
duration        0
dtype: int64
```

```
[35] df_final.head()
```

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	90 min
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons

```
[36] df_final['duration'].value_counts()
```

```
112 min        2594
85 min         2486
89 min         2420
124 min        2310
86 min         2213
4 Seasons      2134
116 min        2122
118 min        2119
119 min        2075
87 min         2063
109 min        2020
113 min        1990
120 min        1845
117 min        1770
121 min        1728
5 Seasons      1698
111 min        1667
144 min        1563
114 min        1529
127 min        1505
115 min        1444
123 min        1398
125 min        1299
122 min        1298
84 min         1268
128 min        1241
130 min        1216
126 min        1205
81 min         1203
83 min         1192
133 min        1169
137 min        1122
82 min         1100
136 min        1092
132 min        1047
131 min         913
135 min         851
7 Seasons       843
129 min         837
75 min          794
148 min         671
140 min         658
6 Seasons       633
79 min          629
139 min         617
143 min         608
80 min          586
134 min         572
145 min         549
149 min         540
138 min         540
74 min          517
78 min          506
141 min         495
72 min          470
142 min         464
46 min          451
77 min          447
150 min         442
172 min         432
158 min         424
73 min          408
76 min          408
151 min         395
147 min         379
163 min         371
154 min         356
146 min         342
162 min         333
54 min          323
153 min         300
71 min          297
70 min          289
8 Seasons       286
157 min         284
155 min         275
68 min          263
9 Seasons       257
24 min          252
161 min         230
166 min         228
```

```
10 Seasons      220
156 min         214
58 min          197
176 min         192
152 min         186
168 min         178
165 min         177
171 min         174
160 min         169
185 min         166
22 min          162
69 min          160
44 min          149
181 min         144
173 min         144
63 min          141
180 min         133
159 min         132
13 Seasons      132
26 min          128
170 min         120
177 min         117
23 min          116
60 min          114
28 min          113
64 min          113
12 Seasons      111
```

```
#removing mins from data
df_final['duration']=df_final['duration'].str.replace(" min","")
df_final.head()
```

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	90
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons

```
df_final['duration_copy']=df_final['duration'].copy()
df_final1=df_final.copy()
```

```
df_final1.loc[df_final1['duration_copy'].str.contains('Season'),'duration_copy']=0
df_final1['duration_copy']=df_final1['duration_copy'].astype('int')
df_final1.head()
```

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration	duration_copy	Modified_Added_date	n
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	90	90	2021-09-25	
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons	0	2021-09-24	
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons	0	2021-09-24	
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons	0	2021-09-24	
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons	0	2021-09-24	

```
[ ]  df_final1['duration_copy'].describe()

count        204571
unique          220
top        1 Season
freq          35635
Name: duration_copy, dtype: object
```

6. Insights based on Non-Graphical and Visual Analysis:

Comments on the range of attributes, Comments on the distribution of the variables and relationship betweenthem, Comments for each univariate and bivariate plot:

- International Movies, Dramas and Comedies are the most popular.
- We have 70:30 ratio of Movies and TV Shows in our data
- countries, such as Cambodia and Cambodia, or United States and United States, are shown as differentcountries. They should have been same
- US, India, UK, Canada and France are leading countries in Content Creation on Netflix
- Most of the highly rated content on Netflix is intended for Mature Audiences, R Rated, content not intended foraudience under 14 and those which require Parental Guidance
- The duration of Most Watched content in our whole data is 80-100 mins. These must be movies and shows having only 1 Season.
- Anupam Kher, SRK, Julie Tejwani, Naseeruddin Shah and Takahiro Sakurai occupy the top stop in Most Watchedcontent.
- Net content release which are later uploaded to Netflix has increased since 1980 till 2020 though later reduced
- certainly, due to COVID-19
- International TV Shows, Dramas and Comedy Genres are popular across TV Shows in Netflix
- International TV Shows, Dramas and Comedy Genres are popular across TV Shows in Netflix
- United States is leading across both TV Shows and Movies, UK also provides great content across TV Shows andMovies. Surprisingly India is much more prevalent in Movies as compared TV Shows.
- Moreover, the number of Movies created in India outweigh the sum of TV Shows and Movies across UK since India wasrated as second in net sum of whole content across Netflix.
- So, it seems possible to conclude that the popular ratings across Netflix includes Mature Audiences and thoseappropriate for over 14/over 17 ages.
- Moreover, there are no TV Shows having a rating of R
- Across TV Shows, shows having only 1 Season are common as soon as the season length increases, the number ofshows decrease and this definitely sounds as expected
- Across movies 80-100,100-120 and 120-150 is the ranges of minutes for which most movies lie. So quite possibly 80-150 mins is the sweet spot we would be wanting for movies.
- Code Text
- Takahiro Sakurai, Yuki Kaji and other South Korean/Japanese actors are the most popular actors across TV Shows
- Our Bollywood actors such as Anupam Kher, SRK, Naseeruddin Shah are very much popular across movies on Netflix
- Ken Burns, Alastair Fothergill, Stan Lathan, Joe Barlinger are popular directors across

TV Shows on Netflix

- Rajiv Chilka, Jan Suter, Raul Campos, Suhas Kadav are popular directors across movies
- Till 2019, overall content across Netflix was increasing but due to Covid in 2020, though TV Shows didn't take a hitthen Movies did take a hit. Well later in 2021, content across both was reduced significantly
- TV Shows are added in Netflix by a tremendous amount in mid weeks/months of the year, i.e.- July
- Movies are added in Netflix by a tremendous amount in first week/last month of current year and first month of nextyear
- Dramas, Comedy, Kids 'TV Shows, International TV Shows and Docuseries, Genres are popular in TV Series in USA
- Dramas, Comedy, Documentaries, Family Movies and Action Genres in Movies are popular in USA
- So, it seems possible to conclude that the popular ratings across Netflix includes Mature Audiences and thoseappropriate for over 14/over 17 ages in both Movies and TV Shows in USA
- Across movies 80-100,100-120 is the ranges of minutes for which most movies lie. So quite possibly 80-120 mins is thesweet spot we would be wanting for movies in USA
- Vincent Tong, Grey Griffin and Kevin Richardson are the most popular actors across TV Shows in USA
- In USA, number of shows remained the same in 2021 as they were in 2020 while number of movies declined:
- TV Shows are added in Netflix by a tremendous amount in July and September in USA
- Movies are added in Netflix in USA by a tremendous amount in first week/last month of current year and first monthof next year
- In USA, though both Movies and Shows have reduced in 2021, the amount of decrease in number of TV Shows is smallas compared to Movies

- **The Most Popular Actor Director Combination in Movies Across USA are:-**
    - 'Smith Foreman and Stanley Moore'.‘Marlon Wayans and Michael Tiddes’.'Adam Sandler and Steve Brill'. 'Maisie Benson and Stanley Moore'. 'Ashleigh Ball and Ishi Rudell'. 'Tara Strong and Ishi Rudell'. 'Rebecca Shoichet and Ishi Rudell'.'Kerry Gudjohnsen and Alex Woo'. 'Kerry Gudjohnsen and Stanley Moore'.'Paul Killam and Alex Woo'. 'Paul Killam and Stanley Moore'.'Andrea Libman and Ishi Rudell'.'Kevin Hart and Leslie Small'. 'Maisie Benson and Alex Woo'. 'Alexa PenaVega and Robert Rodriguez'.'Tabitha St. Germain and Ishi Rudell'
- **The Second Most Popular Actor Director Combination in Movies Across USA are:-**
    'Rory Markham and Mike Gunther', 'Erin Mathews and Steve Ball', 'Danny Trejo and Robert Rodriguez','Jeff Dunham and Michael Simon'
- **Popular Actors in TV Shows in India are: -**
    'Rajesh Kava', 'Nishka Raheja', 'Prakash Raj', 'Sabina Malik', 'Anjali', 'Aranya Kaur', 'Sonal Kaushal', 'Chandan Anand','Danish Husain'
- **Popular actors across Movies in India:-**
    'Anupam Kher', 'Shah Rukh Khan', 'Naseeruddin Shah', 'Akshay Kumar', 'Om Puri',

'Paresh Rawal',
'Julie Tejwani',
'Amitabh
Bachchan',
'Boman Irani',
'Rupa Bhimani',
'Kareena
Kapoor',
'Ajay Devgn',
'Rajesh Kava',
'Kay Kay
Menon'

- **Popular Directors Across Movies in India:-**
  'Gautham Vasudev
  Menon','Abhishek
  Chaubey', 'Sudha
  Kongara', 'Rathindran
  R Prasad', 'Sankalp
  Reddy',
  'Sarjun',
  'Soumendra
  Padhi','Srijit
  Mukherji',
  'Tharun Bhascker Dhaassyam'
- **Popular directors across movies in India:-**
  'Rajiv Chilaka',
  'Suhas Kadav',
  'David Dhawan',
  'Umesh Mehra',
  'Anurag
  Kashyap', 'Ram
  Gopal Varma',
  'Dibakar
  Banerjee', 'Zoya
  Akhtar', 'Tilak
  Shetty',
  'Rajkumar
  Santoshi',
  'Priyadarshan',
  'Sooraj R.
  Barjatya',
  'Ashutosh
  Gowariker','Milan
  Luthria'
- In India, TV Shows were increasingly being added till 2020, though the addition of shows reduced in 2021.
- In India, Movies were increasingly added till 2018 but it has been a huge downhill since then. Now that'spreposterous, since and something has to be recommended to the Netflix Team with regards to that.

- **The Most Popular Actor Director Combination in Movies Across India are:-**
  'Rajesh Kava and Rajiv Chilaka', 'Julie Tejwani and Rajiv Chilaka', 'Rupa Bhimani and Rajiv Chilaka', 'Jigna Bhardwaj and Rajiv Chilaka','Vatsal Dubey and Rajiv Chilaka', 'Mousam and Rajiv Chilaka', 'Swapnil and Rajiv Chilaka', 'Saurav Chakraborty and Suhas Kadav','Smita Malhotra and Tilak Shetty', 'Anupam Kher and David Dhawan', 'Salman Khan and Sooraj R. Barjatya',

7. Business Insights:

- The analysis shows us that there is high number of movies produced per year than tv s
- corona virus has the impacted the content quantity
- the USA and INDIA are the top 2 countries content wise
- the content targeted in India is teens while the content being targeted at USA is adult
- lack of child content produced in India
- India and south Korean have similar taste and USA and UK audience have similar taste 7. lack of diverse content forIndian audience

8. Recommendations

- The most popular Genres across the countries and in both TV Shows and Movies are Drama, Comedy andInternational TV Shows/Movies, so content aligning to that is recommended.
- Add TV Shows in July/August and Movies in last week of the year/first month of the next year.
- For USA audience 80-120 mins is the recommended length for movies and Kids TV Shows are also popular alongwith the genres in first point, hence recommended.
- For UK audience, recommended length for movies is same as that of USA (80-120 mins)
- The target audience in USA and India is recommended to be 14+ and above ratings while for UK, itsrecommended to be completely Mature/R content.
- Add movies for Indian Audience, it has been declining since 2018.
- Anime Genre for Japan and Romantic Genre in TV Shows for South Korean audiences is recommended.
- While creating content, take into consideration the popular actors/directors for that country. Also take intoaccount the director-actor combination which is highly recommended.