1. Problem Statement - Analyze the data and generate insights that could help Netflix in deciding which type of shows/movies to produce and how they can grow the business in different countries

In [551…

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

Read data

In [552…

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

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

2.1 The shape of data

In [553…

```
df.shape
```

Out[553]: (8807, 12)

2.2 Data types of all the attributes, conversion of categorical attributes to 'category' (If required) , missing value detection

In [554…

```
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
```

Observation :

A. There are missing values in features such as "director" (29 %) , "cast" , "country" , "rating" , and duration

B. date_added is an 'Object' , need to convert it datetime

2.3 Missing value detection

In [555…

```
df.isnull().sum(axis=0)
```

Out[555]:

```
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
```

2.4 Statistical summary

In [556…

```
df.describe()
```

Out[556]:

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

In [557…

```
df.describe(include='object')
```

Out[557]:

	show_id	type	title	director	cast	country	date_added	rating	duration	listed_in	de
count	8807	8807	8807	6173	7982	7976	8797	8803	8804	8807	
unique	8807	2	8807	4528	7692	748	1767	17	220	514	
top	s1	Movie	Dick Johnson Is Dead	Rajiv Chilaka	David Attenborough	United States	January 1, 2020	TV-MA	1 Season	Dramas, International Movies	Pa a
freq	1	6131	1	19	19	2818	109	3207	1793	362	

Observations :

A. Director is "Rajiv Chilaka" has 19 Movies/Series on the platform , including 4528 different directors

B. "David Attenborough" has been casted in 19 movies/series

C. There are 2818 movies/series from United states

D. There are Movies/Series from 748 different countries in Netflix platform

E. Mostly 'TV-MA' rated shows are in Netflix

F. 109 movies/series being added to platform on January 1, 2020

In [558…

```
# checking duplicates records
df.loc[df.duplicated()]
```

Out[558]:

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

Observations :

No duplicate records present in the dataset, which is good.

3. Non-Graphical Analysis: Value counts and unique

attributes

In [559…
```
df["type"].value_counts()
```

Out[559]:
```
Movie      6131
TV Show    2676
Name: type, dtype: int64
```

Observation :

The dataset contains data about only Movie and TV shows

In [560…
```
df["title"].value_counts()
```

Out[560]:
```
Dick Johnson Is Dead                     1
Ip Man 2                                 1
Hannibal Buress: Comedy Camisado         1
Turbo FAST                               1
Masha's Tales                            1
                                        ..
Love for Sale 2                          1
ROAD TO ROMA                             1
Good Time                                1
Captain Underpants Epic Choice-o-Rama    1
Zubaan                                   1
Name: title, Length: 8807, dtype: int64
```

In [561…
```
df["title"].unique()
```

Out[561]:
```
array(['Dick Johnson Is Dead', 'Blood & Water', 'Ganglands', ...,
       'Zombieland', 'Zoom', 'Zubaan'], dtype=object)
```

In [562…
```
df.loc[(df["title"].isnull()) | (df["title"].isna())]
```

Out[562]:

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

Observation :

No duplicate titles are present in the dataset

No NaN /Null values for titles , which looks good

In [563…
```
df["director"].value_counts()
```

Out[563]:
```
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
```

In [564…

```
df["director"].unique()
```

Out[564]:
```
array(['Kirsten Johnson', nan, 'Julien Leclercq', ..., 'Majid Al Ansari',
       'Peter Hewitt', 'Mozez Singh'], dtype=object)
```

In [565…
```
df.loc[(df["director"].isnull()) | (df["director"].isna())][:5]
```

Out[565]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	Internatic TV Sho TV Dran TV Myste
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuse Reality
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	Internatic TV Sho Romantic Shows, T
10	s11	TV Show	Vendetta: Truth, Lies and The Mafia	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime Sho Docuse Internatic TV
14	s15	TV Show	Crime Stories: India Detectives	NaN	NaN	NaN	September 22, 2021	2021	TV-MA	1 Season	British Sho Crime Sho Docuse

Observation :

Director feature has multiple Nan values . Need to analyzed further why dierctors are null

In [566…
```
df["cast"].value_counts()
```

Out[566]:
```
David Attenborough                                                    19
Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jigna Bhardwaj, Rajesh Kava, Mousam, Swapnil    14
Samuel West                                                           10
Jeff Dunham                                                            7
David Spade, London Hughes, Fortune Feimster                           6
                                                                      ..
Michael Peña, Diego Luna, Tenoch Huerta, Joaquin Cosio, José María Yazpik, Matt Letscher, Alyssa Diaz    1
Nick Lachey, Vanessa Lachey                                            1
Takeru Sato, Kasumi Arimura, Haru, Kentaro Sakaguchi, Takayuki Yamada, Kendo Kobayashi, Ken Yasuda, Arata Furuta, Suzuki Matsuo, Koichi Yamadera, Arata Iura, Chikako Kaku, Kotaro Yoshida    1
Toyin Abraham, Sambasa Nzeribe, Chioma Chukwuka Akpotha, Chioma Omeruah, Chiwetalu Agu, Dele Odule, Femi Adebayo, Bayray McNwizu, Biodun Stephen    1
Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanana, Manish Chaudhary, Meghna Malik, Malkeet Rauni, Anita Shabdish, Chittaranjan Tripathy    1
Name: cast, Length: 7692, dtype: int64
```

In [567…
```
df["cast"].unique()
```

Out[567]:
```
array([nan,
       'Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Molaba, Dillon Windvogel, Natasha Thahane, Arno Greeff, Xolile Tshabalala, Getmore Sithole, Cindy Mahlangu, Ryle De Morny, Greteli Fincham, Sello Maake Ka-Ncube, Odwa Gwanya, Mekaila Mathys, Sandi Schultz, Duane Williams, Shamilla Miller, Patrick Mofokeng',
       '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'],
      dtype=object)
```

Observation :

Casts are comma seperated , need to process this feature before further analysis

In [568…
```
df["country"].value_counts()
```

Out[568]:
```
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
```

Observation :

Few countries are comma seperated and few are not . Hence data processing is required

In [569…
```
df.loc[(df["country"].isnull()) | (df["country"].isna())][:5]
```

Out[569]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	liste
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crim Sh Internat TV Sh TV
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docus Reali
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 Dra TV Ho TV Myst
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	Childr F M
10	s11	TV Show	Vendetta: Truth, Lies and The Mafia	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Crim Sh Docus Internat T

Observation :

Looks like for both TV Show and Movie, There are missing 'country' and 'director' . Need to reach out to business analyst to see why those are missing. Those missing data need

to be tagged appropriately and processed further.

In [570…
```
df["rating"].value_counts()
```

Out[570]:
```
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
```

In [571…
```
df["rating"].unique()
```

Out[571]:
```
array(['PG-13', 'TV-MA', 'PG', 'TV-14', 'TV-PG', 'TV-Y', 'TV-Y7', 'R',
       'TV-G', 'G', 'NC-17', '74 min', '84 min', '66 min', 'NR', nan,
       'TV-Y7-FV', 'UR'], dtype=object)
```

Observation :

A few ratings have duration value with units (i.e. 74 mins , 84 min etc.) hence pre processsing is required

In [572…
```
df.loc[df["rating"].isnull()]
```

Out[572]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration
5989	s5990	Movie	13TH: A Conversation with Oprah Winfrey & Ava ...	NaN	Oprah Winfrey, Ava DuVernay	NaN	January 26, 2017	2017	NaN	37 min
6827	s6828	TV Show	Gargantia on the Verdurous Planet	NaN	Kaito Ishikawa, Hisako Kanemoto, Ai Kayano, Ka...	Japan	December 1, 2016	2013	NaN	1 Season
7312	s7313	TV Show	Little Lunch	NaN	Flynn Curry, Olivia Deeble, Madison Lu, Oisín ...	Australia	February 1, 2018	2015	NaN	1 Season
7537	s7538	Movie	My Honor Was Loyalty	Alessandro Pepe	Leone Frisa, Paolo Vaccarino, Francesco Miglio...	Italy	March 1, 2017	2015	NaN	115 min

Observation :

A few ratings are Null . Need to analyze why

In [573…

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

Out[573]:

```
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
```

In [574…

```
df.loc[(df["duration"].isnull()) | (df["duration"].isna())]
```

Out[574]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	April 4, 2017	2017	74 min	NaN	Movies	o
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	September 16, 2016	2010	84 min	NaN	Movies	
5813	s5814	Movie	Louis C.K.: Live at the Comedy Store	Louis C.K.	Louis C.K.	United States	August 15, 2016	2015	66 min	NaN	Movies	T h

In [575…

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

Out[575]:

```
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
```

In [576…

```
df.loc[(df["listed_in"].isnull()) | (df["listed_in"].isna())]
```

Out[576]:

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

Observation :

"listed_in" needs to be pre processed as there are comma sperated values . However no null /Nan values are present .

Pre-processing - Excluding feature "description" as we're doing basic analysis without NLP

In [577…

```
df.drop("description",axis = 1,inplace=True)
```

In [578…

```
# checking "description" feature was dropped or not
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 11 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
dtypes: int64(1), object(10)
memory usage: 757.0+ KB
```

In [579…

```
# Pre-processing -   "cast" , "country", "listed_in" - Needs to be melted as multiple co
#                    "rating" have duration i.e. 74 mins , 84 min etc. Some pre process

# Missing values -   "director" feature has multiple Nan values
#                    Looks like for both TV Show and Movie, country and director is nul
#                    A few ratings are Null . Need to analyze why
```

Pre-processing 1.1 - Melt comma sperated colums "cast" , "country", "listed_in"

In [580…

```
cast_list = df['cast'].apply(lambda x:str(x).split(", ")).to_list()
df_title_to_cast = pd.DataFrame(cast_list,index=df['title']) # converting casts into par
df_title_to_cast = df_title_to_cast.stack() # level wise stacking
df_title_to_cast = pd.DataFrame(df_title_to_cast)
df_title_to_cast.reset_index(inplace=True)
df_title_to_cast.drop("level_1", axis = 1, inplace=True)
df_title_to_cast.rename(columns = {0:'cast'}, inplace = True)
```

In [581…

```
df_title_to_cast.head()
```

Out[581]:

	title	cast
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

In [582…

```
df_title_to_cast.shape
```

Out[582]:

```
(64951, 2)
```

In [583…

```
country_list = df["country"].apply(lambda x: str(x).split(", ")).to_list()
df_title_to_country = pd.DataFrame(country_list,index=df["title"])
df_title_to_country = df_title_to_country.stack()
df_title_to_country = pd.DataFrame(df_title_to_country)
```

```
df_title_to_country.reset_index(inplace = True)
df_title_to_country.drop("level_1",axis = 1 , inplace= True)
df_title_to_country.rename(columns = {0:'country'}, inplace = True)
```

In [584…
```
df_title_to_country.shape
```

Out[584]:
```
(10845, 2)
```

In [585…
```
df_title_to_country.head()
```

Out[585]:

	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

In [586…
```
df["listed_in"].head()
```

Out[586]:
```
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 ...
Name: listed_in, dtype: object
```

In [587…
```
listed_in_list = df["listed_in"].apply(lambda x: str(x).split()).to_list()
df_title_to_listed = pd.DataFrame(listed_in_list,index=df["title"])
df_title_to_listed = df_title_to_listed.stack()
df_title_to_listed = pd.DataFrame(df_title_to_listed)
df_title_to_listed.reset_index(inplace= True)
df_title_to_listed.drop("level_1", axis = 1, inplace = True)
df_title_to_listed.rename(columns={0:'listed_in'},inplace = True)
```

In [588…
```
df_title_to_listed.shape
```

Out[588]:
```
(39221, 2)
```

In [589…
```
df_title_to_listed.head()
```

Out[589]:

	title	listed_in
0	Dick Johnson Is Dead	Documentaries
1	Blood & Water	International
2	Blood & Water	TV
3	Blood & Water	Shows,
4	Blood & Water	TV

In [590…

```
df.head()
```

Out[590]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	list
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documer
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	Interna TV Shov Dram Mys
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crir S Interna TV Shov
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docu Real
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	Interna TV S Roman Shows

In [591…

```
df_tiltle_cast_country = df_title_to_cast.merge(df_title_to_country,on="title")
```

In [592…

```
df_tiltle_cast_country.head()
```

Out[592]:

	title	cast	country
0	Dick Johnson Is Dead	nan	United States
1	Blood & Water	Ama Qamata	South Africa
2	Blood & Water	Khosi Ngema	South Africa
3	Blood & Water	Gail Mabalane	South Africa
4	Blood & Water	Thabang Molaba	South Africa

In [593…

```
df_tiltle_cast_country_list_in = df_tiltle_cast_country.merge(df_title_to_listed,on="tit
```

In [594…

```
df_tiltle_cast_country_list_in.head()
```

Out[594]:

	title	cast	country	listed_in
0	Dick Johnson Is Dead	nan	United States	Documentaries
1	Blood & Water	Ama Qamata	South Africa	International
2	Blood & Water	Ama Qamata	South Africa	TV
3	Blood & Water	Ama Qamata	South Africa	Shows,
4	Blood & Water	Ama Qamata	South Africa	TV

In [595…

```
df.head()
```

Out[595]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	list
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documer
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	Interna TV Shov Dram Mys
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crir S Interna TV Shov
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docu Real
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	Interna TV S Roman Shows

In [596…

```
df_selected = df[["show_id","type","title","director","date_added","release_year","ratir
```

In [597…

```
df_selected.head()
```

Out[597]:

	show_id	type	title	director	date_added	release_year	rating	duration
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	September 25, 2021	2020	PG-13	90 min
1	s2	TV Show	Blood & Water	NaN	September 24, 2021	2021	TV-MA	2 Seasons
2	s3	TV Show	Ganglands	Julien Leclercq	September 24, 2021	2021	TV-MA	1 Season
3	s4	TV Show	Jailbirds New Orleans	NaN	September 24, 2021	2021	TV-MA	1 Season
4	s5	TV Show	Kota Factory	NaN	September 24, 2021	2021	TV-MA	2 Seasons

In [598…
```
df_selected.shape
```

Out[598]:
```
(8807, 8)
```

In [599…
```
df_tiltle_cast_country_list_in.shape
```

Out[599]:
```
(384220, 4)
```

In [600…
```
df_processed = df_selected.merge(df_tiltle_cast_country_list_in,on="title")
```

In [601…
```
df_processed.head()
```

Out[601]:

	show_id	type	title	director	date_added	release_year	rating	duration	cast	country	listed_i
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	September 25, 2021	2020	PG-13	90 min	nan	United States	Documentarie
1	s2	TV Show	Blood & Water	NaN	September 24, 2021	2021	TV-MA	2 Seasons	Ama Qamata	South Africa	Internationa
2	s2	TV Show	Blood & Water	NaN	September 24, 2021	2021	TV-MA	2 Seasons	Ama Qamata	South Africa	TV
3	s2	TV Show	Blood & Water	NaN	September 24, 2021	2021	TV-MA	2 Seasons	Ama Qamata	South Africa	Shows
4	s2	TV Show	Blood & Water	NaN	September 24, 2021	2021	TV-MA	2 Seasons	Ama Qamata	South Africa	TV

In [602…
```
df_processed.shape
```

Out[602]:
```
(384220, 11)
```

Pre-processing 1.2 - Convert "date_added" feature to datetime and then feature extraction to year , month, week, day

In [603…
```
df_processed.info() # BEFORE convertion, checking data type of "date_added" feature
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 384220 entries, 0 to 384219
Data columns (total 11 columns):
 #   Column        Non-Null Count   Dtype 
---  ------        --------------   ----- 
 0   show_id       384220 non-null  object
 1   type          384220 non-null  object
 2   title         384220 non-null  object
 3   director      253193 non-null  object
 4   date_added    383798 non-null  object
 5   release_year  384220 non-null  int64 
 6   rating        384073 non-null  object
 7   duration      384217 non-null  object
 8   cast          384220 non-null  object
 9   country       384220 non-null  object
 10  listed_in     384220 non-null  object
dtypes: int64(1), object(10)
memory usage: 35.2+ MB
```

In [604…

```
df_processed["date_added"]= pd.to_datetime(df_processed["date_added"])
```

In [605…

```
df_processed.info() #  AFTER convertion, checking data type of "date_added" feature
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 384220 entries, 0 to 384219
Data columns (total 11 columns):
 #   Column        Non-Null Count   Dtype         
---  ------        --------------   -----         
 0   show_id       384220 non-null  object        
 1   type          384220 non-null  object        
 2   title         384220 non-null  object        
 3   director      253193 non-null  object        
 4   date_added    383798 non-null  datetime64[ns]
 5   release_year  384220 non-null  int64         
 6   rating        384073 non-null  object        
 7   duration      384217 non-null  object        
 8   cast          384220 non-null  object        
 9   country       384220 non-null  object        
 10  listed_in     384220 non-null  object        
dtypes: datetime64[ns](1), int64(1), object(9)
memory usage: 35.2+ MB
```

In [606…

```
df_processed.head()
```

Out[606]:

	show_id	type	title	director	date_added	release_year	rating	duration	cast	country	listed_i
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	2021-09-25	2020	PG-13	90 min	nan	United States	Documentarie
1	s2	TV Show	Blood & Water	NaN	2021-09-24	2021	TV-MA	2 Seasons	Ama Qamata	South Africa	Internationa
2	s2	TV Show	Blood & Water	NaN	2021-09-24	2021	TV-MA	2 Seasons	Ama Qamata	South Africa	TV
3	s2	TV Show	Blood & Water	NaN	2021-09-24	2021	TV-MA	2 Seasons	Ama Qamata	South Africa	Shows
4	s2	TV Show	Blood & Water	NaN	2021-09-24	2021	TV-MA	2 Seasons	Ama Qamata	South Africa	TV

In [607…
```
df_processed["date_added_year"] = df_processed["date_added"].dt.year
df_processed["date_added_month"] = df_processed["date_added"].dt.month_name()
df_processed["date_added_week"] = df_processed["date_added"].dt.weekday
df_processed["date_added_day"] = df_processed["date_added"].dt.day_name()
```

In [608…
```
df_processed.head()
```

Out[608]:

	show_id	type	title	director	date_added	release_year	rating	duration	cast	country	listed_i
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	2021-09-25	2020	PG-13	90 min	nan	United States	Documentarie
1	s2	TV Show	Blood & Water	NaN	2021-09-24	2021	TV-MA	2 Seasons	Ama Qamata	South Africa	Internationa
2	s2	TV Show	Blood & Water	NaN	2021-09-24	2021	TV-MA	2 Seasons	Ama Qamata	South Africa	TV
3	s2	TV Show	Blood & Water	NaN	2021-09-24	2021	TV-MA	2 Seasons	Ama Qamata	South Africa	Shows
4	s2	TV Show	Blood & Water	NaN	2021-09-24	2021	TV-MA	2 Seasons	Ama Qamata	South Africa	TV

In [609…
```
# df_processed.date_added_year = df_processed.date_added_year.astype(int)
```

Pre-processing 1.3 - "rating" have duration i.e. 74 mins , 84 min etc. Some pre processsing is required

In [610…
```
df_processed["rating"].value_counts()
```

Out[610]:
```
TV-MA       140910
TV-14        85908
R            42099
TV-PG        29258
PG-13        28633
PG           22095
TV-Y7        14454
TV-Y          8214
TV-G          5940
G             3204
NR            2723
NC-17          243
TV-Y7-FV       239
UR             150
74 min           1
84 min           1
66 min           1
Name: rating, dtype: int64
```

In [611…
```
df_processed["rating"].replace(to_replace=r'(^.*min$)', value='Unknown', regex=True,inpl
```

In [612…
```
df_processed["rating"].value_counts()
```

Out[612]:
```
TV-MA       140910
TV-14        85908
R            42099
TV-PG        29258
PG-13        28633
PG           22095
TV-Y7        14454
TV-Y          8214
TV-G          5940
G             3204
NR            2723
NC-17          243
TV-Y7-FV       239
UR             150
Unknown          3
Name: rating, dtype: int64
```

Pre-processing 1.4 - "duration" is string , need to translate to interger

In [613…
```
df_processed["duration"]= df_processed["duration"].str.replace(" min", "", case = True)
```

In [614…
```
df_processed[df_processed["type"] == "Movie"]["duration"].value_counts()
```

Out[614]:
```
94     6370
93     6351
97     6112
95     5849
106    5569
       ...
18        6
39        6
9         5
10        4
8         1
Name: duration, Length: 205, dtype: int64
```

In [615…
```
df_processed[df_processed["type"] == "TV Show"]["duration"].value_counts()
```

Out[615]:
```
1 Season      87664
2 Seasons     24405
3 Seasons     13072
4 Seasons      5420
5 Seasons      4649
7 Seasons      2270
6 Seasons      1555
8 Seasons       698
10 Seasons      652
9 Seasons       596
13 Seasons      445
12 Seasons      304
15 Seasons      278
11 Seasons       90
17 Seasons       75
Name: duration, dtype: int64
```

Pre-processing 1.5 - In "listed_in", Few Genres can be combined such as 'Movies' and 'Movies,' , 'Shows' and 'Shows,' etc.

In [616…
```
df_processed["listed_in"]= df_processed["listed_in"].str.replace(",", "", case = True)
```

Pre-processing 1.6 - For movies , creating new feature 'duration_catgory' from 'duration' as following categories :

Short (< 15)

Medium(16 - 30)

Regular(H) - Regular Duration for Hollywood Movies (31 to 90)

Regular(I) - Regular Duration for Bollywood Movies (91 to 150)

Long - Long (> 151)

In [617…
```
def movie_duration_categorization(x):
    x = int(x)
    if x <= 15:
        return "Short"
    elif 15 < x <= 30:
        return "Medium"
    elif 30 < x <= 90:
        return "Regular-H"
    elif 90 < x <= 150:
        return "Regular-I"
    else:
        return "Long"
```

In [618…
```
df_processed["duration"] = df_processed["duration"].fillna(0)
df_processed["duration_category"] = df_processed[df_processed["type"] == 'Movie']["durat
df_processed["duration_category"] = df_processed["duration_category"].replace(np.nan,"Un
df_processed["duration_category"].unique()
```

Out[618]:
```
array(['Regular-H', 'Unknown', 'Regular-I', 'Long', 'Medium', 'Short'],
      dtype=object)
```

5.1 Missing Value check (Treatment optional)

In [619…
```
percent_missing = df_processed.isnull().sum() * 100 / len(df_processed)
missing_value_df = pd.DataFrame({'column_name': df_processed.columns,
                                 'percent_missing': percent_missing})
missing_value_df.sort_values('percent_missing', ascending=False)
```

Out[619]:

	column_name	percent_missing
director	director	34.102077
date_added	date_added	0.109833
date_added_year	date_added_year	0.109833
date_added_month	date_added_month	0.109833
date_added_week	date_added_week	0.109833
date_added_day	date_added_day	0.109833
rating	rating	0.038259
show_id	show_id	0.000000
type	type	0.000000
title	title	0.000000
release_year	release_year	0.000000
duration	duration	0.000000
cast	cast	0.000000
country	country	0.000000
listed_in	listed_in	0.000000
duration_category	duration_category	0.000000

5.1.1 Missing Value Treatment

In [620…

```
df_processed["date_added_year"] = df_processed["date_added_year"].fillna(0)
df_processed['date_added_year'] =df_processed['date_added_year'].astype(np.int64)

df_processed["date_added_week"] = df_processed["date_added_week"].fillna(0)
df_processed['date_added_week'] = df_processed['date_added_week'].astype(np.int64)

df_processed["date_added_month"] = df_processed["date_added_month"].fillna("Unknown")
df_processed["date_added_day"] = df_processed["date_added_day"].fillna("Unknown")

df_processed["duration"] = df_processed["duration"].fillna(0)
#df_processed['duration'] = df_processed['duration'].astype(np.int64)
```

In [621…

```
df_selected = df_processed[["type","director","release_year","rating","duration","cast",
```

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

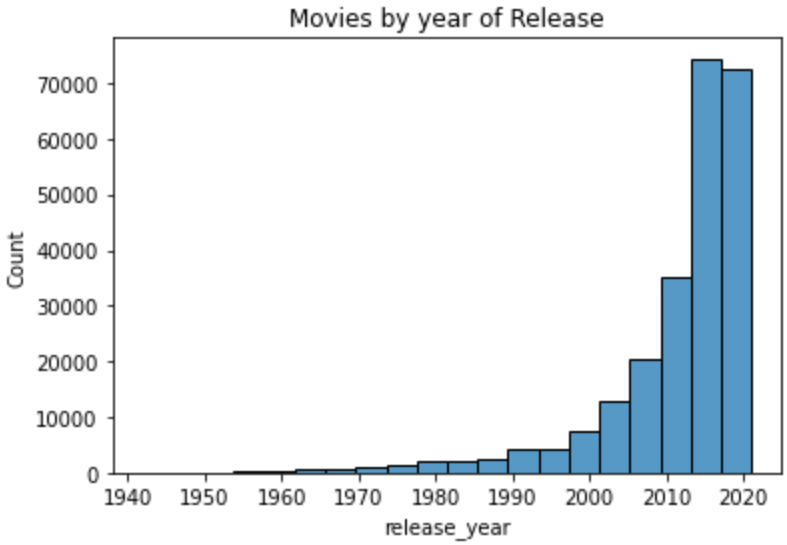
4.1 For continuous variable(s): Distplot, countplot, histogram for univariate analysis

In [622…

```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [623…
```
df_movies = df_selected.loc[df_selected['type'] == 'Movie']
df_series = df_selected.loc[df_selected['type'] == 'TV Show']
```

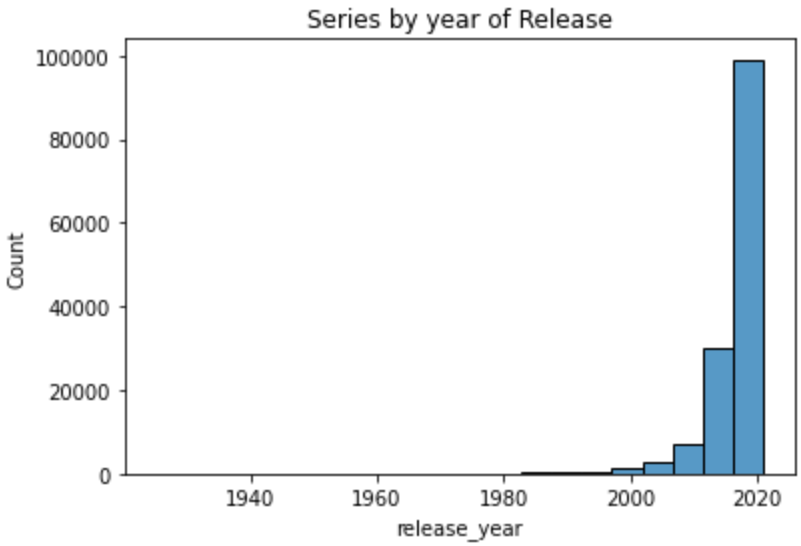
In [624…
```
sns.histplot(df_movies['release_year'],bins=20).set(title='Movies by year of Release')
plt.show()
```



In [625…
```
df_movies['release_year'].min()
```

Out[625]:
```
1942
```

In [626…
```
sns.histplot(df_series['release_year'],bins=20).set(title='Series by year of Release')
plt.show()
```

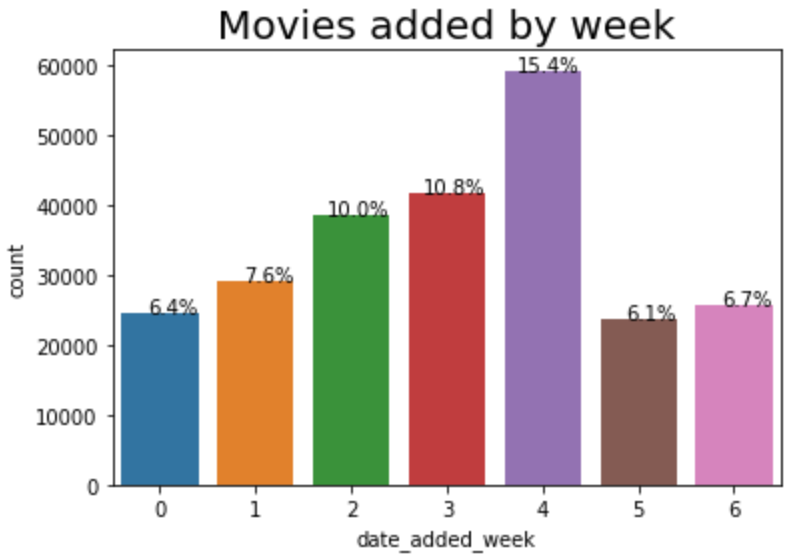


Observation :

In year 2019 - 2020 most of Movies and Series were released in Netflix

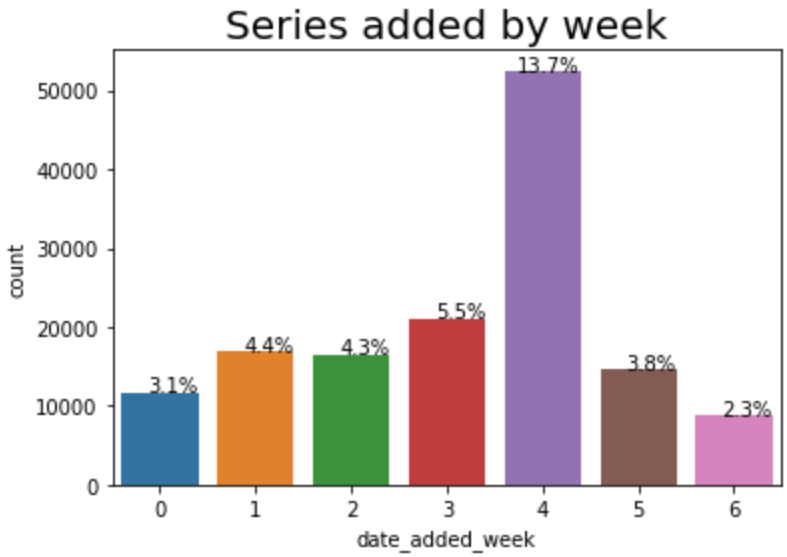
In [627…
```
total = float(len(df_selected))
ax = sns.countplot(x="date_added_week", data=df_movies)
plt.title('Movies added by week', fontsize=20)
for p in ax.patches:
    percentage = '{:.1f}%'.format(100 * p.get_height()/total)
    x = p.get_x() + p.get_width()
```

```
    y = p.get_height()
    ax.annotate(percentage, (x, y),ha='right')
plt.show()
```



In [628…

```
total = float(len(df_selected))
ax = sns.countplot(x="date_added_week", data=df_series)
plt.title('Series added by week', fontsize=20)
for p in ax.patches:
    percentage = '{:.1f}%'.format(100 * p.get_height()/total)
    x = p.get_x() + p.get_width()
    y = p.get_height()
    ax.annotate(percentage, (x, y),ha='right')
plt.show()
```



Observation :

Both movies and Series are more often added on Friday(i.e. week day # 4)

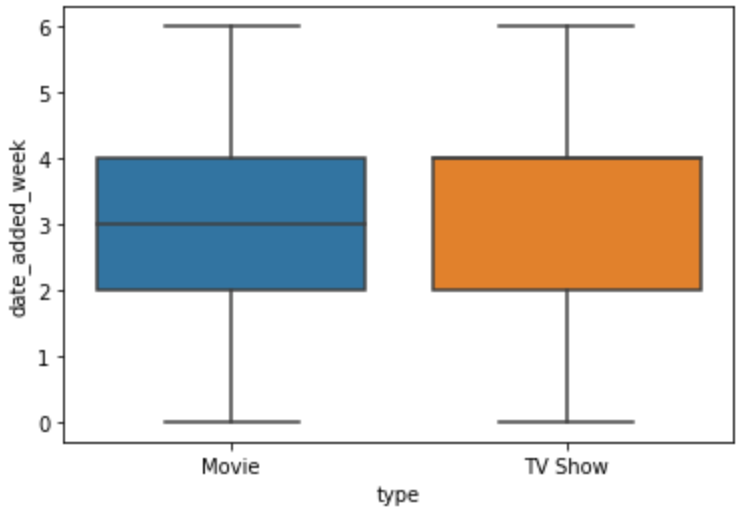
4.2 For categorical variable(s): Boxplot

In [629…

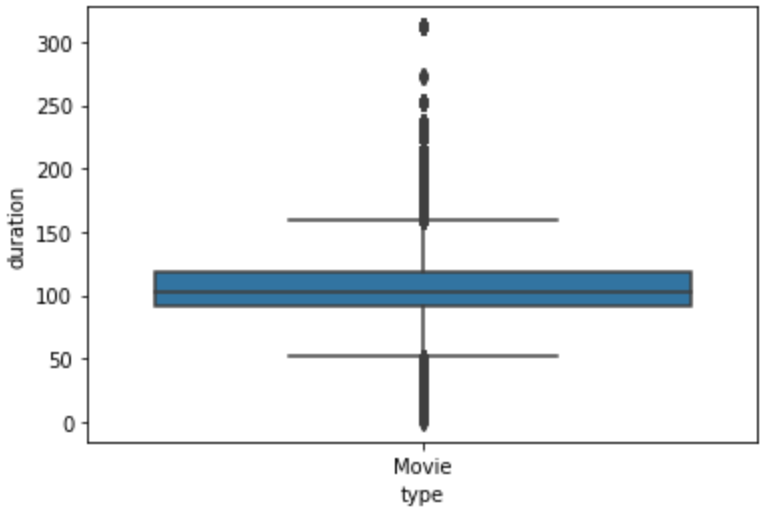
```
df_movies['duration'] = df_movies['duration'].astype(np.int64)
```

```
C:\Users\hp\AppData\Local\Temp/ipykernel_19412/1770795435.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  df_movies['duration'] = df_movies['duration'].astype(np.int64)
```

In [630…
```
sns.boxplot(x="type",y='date_added_week',data=df_processed)
plt.show()
```



In [631…
```
sns.boxplot(x="type",y='duration',data=df_movies)
plt.show()
```

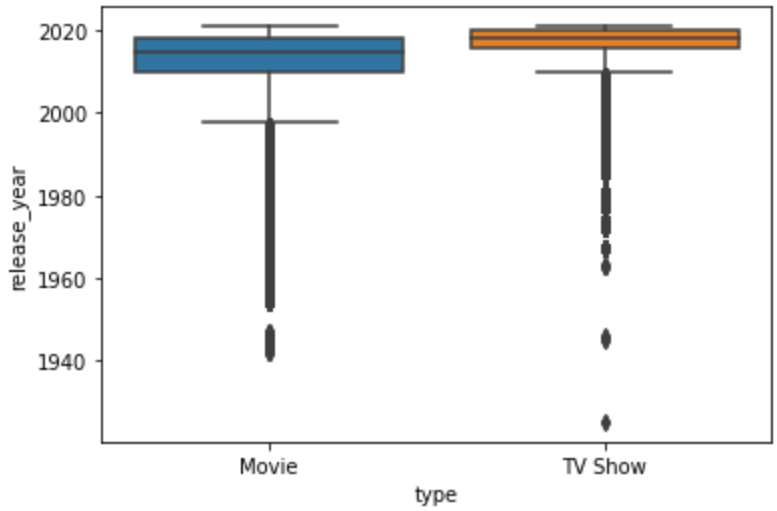


Observation :

Wide range of shows/movies - from very short to very long duration, even across different ratings

In [632…
```
sns.boxplot(x="type", y="release_year", data=df_selected)
```

Out[632]:
```
<AxesSubplot:xlabel='type', ylabel='release_year'>
```

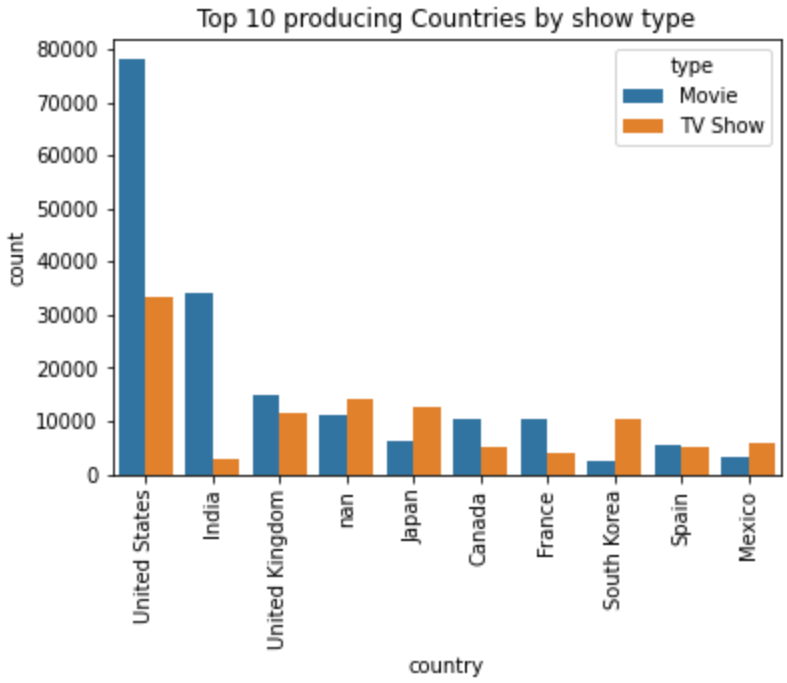


Observation :

Both shows/movies data are skewed , with many outliers on each type

In [633…

```
sp = sns.countplot(x="country",data=df_selected,hue="type",order=df_selected.country.val
sp.set(title='Top 10 producing Countries by show type')
plt.xticks(rotation=90)
plt.show()
```



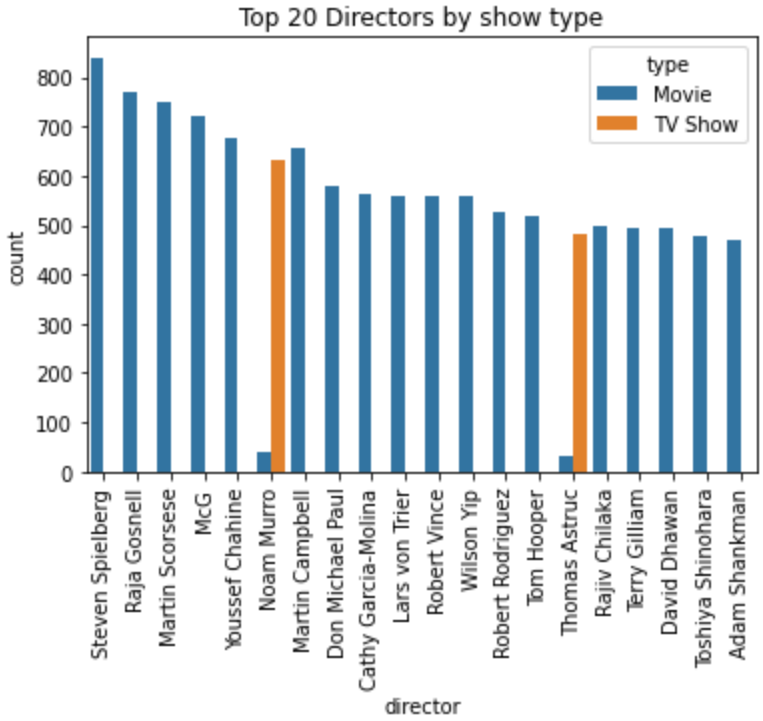
Observation :

A. United states and India produce more movies than TV Shows

B. Japan , South Korea and Mexio produce more TV Shows than movies

In [634…

```
sns.countplot(x="director",data=df_selected,hue="type",order=df_selected.director.value_
plt.xticks(rotation=90)
plt.show()
```



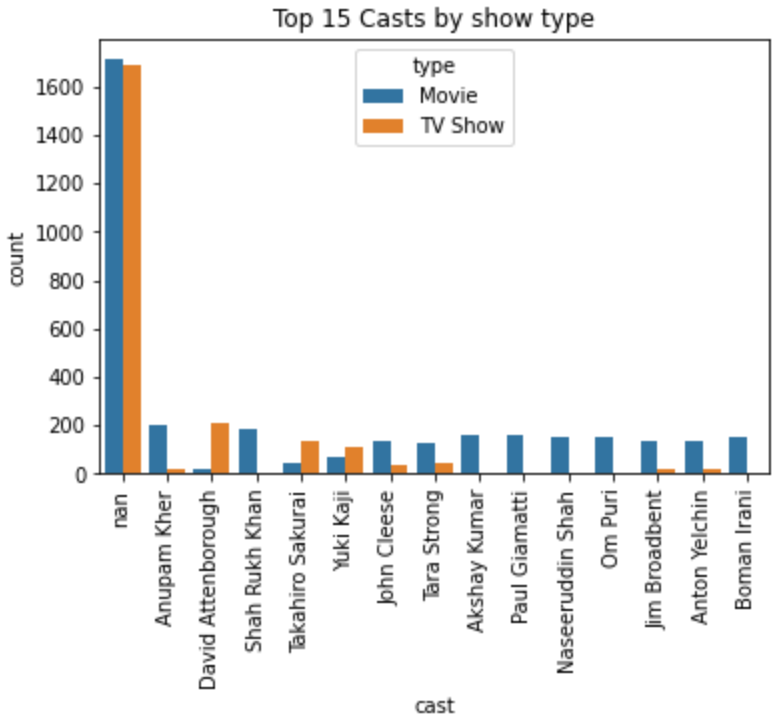
Observation :

A. Noam Murro and Thomas Astruc more often produces TV Shows

B. Steven Spielberg , Raja Gosnell have more movies in Neflix platform

In [635…

```
sns.countplot(x="cast",data=df_selected,hue="type",order=df_selected.cast.value_counts()
plt.xticks(rotation=90)
plt.show()
```



Observation :

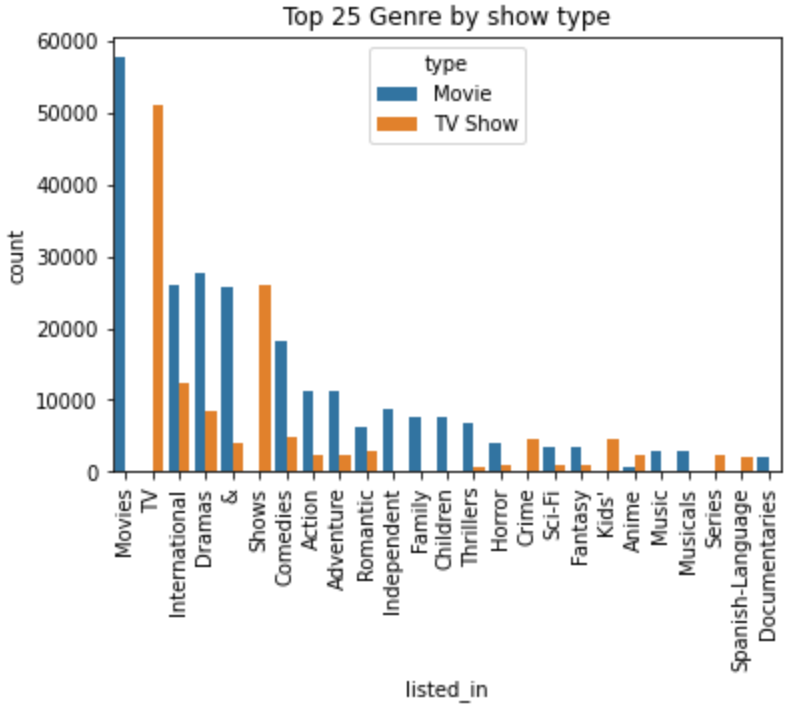
A. Netflix is showing more movies of Anupam Kher , Shah Rukh Khan , Akshyay Kumar , Boman Irani etc.

B. Many Movies/shows have no cast information tagged, which can be improved for deep dive analysis

C. David Attenborough, Takahiro Sakurai , Yuki Kaji have more TV shows than movies in Netflix platform

In [636…

```
sns.countplot(x="listed_in",data=df_selected,hue="type",order=df_selected.listed_in.valu
plt.xticks(rotation=90)
plt.show()
```



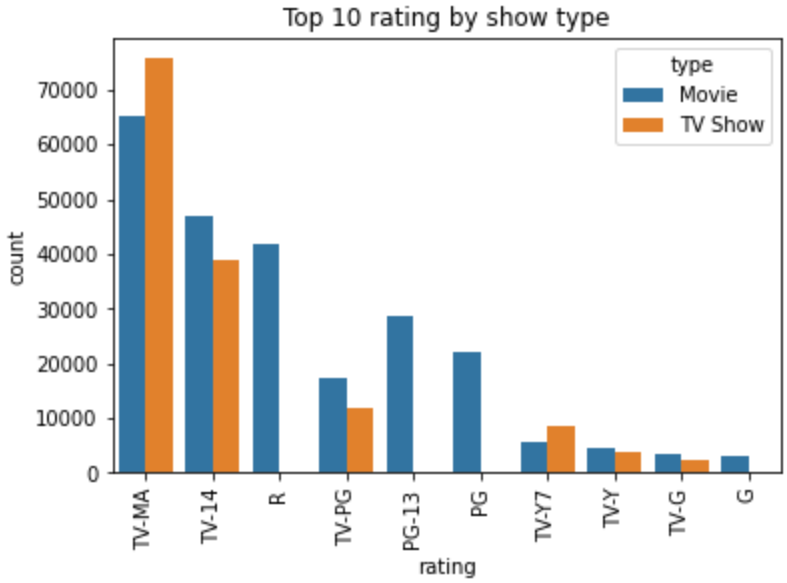
Observation :

A. Less Comedy,Thriller, and no Children "shows" in Netflix platform, which can be explored

B. Too many Genres to choose from

In [637…

```
sns.countplot(x="rating",data=df_selected,hue="type",order=df_selected.rating.value_cou
plt.xticks(rotation=90)
plt.show()
```



In [638…

```
df_movies = df_processed.loc[(df_processed["type"] =="Movie")]
df_movies['duration'] =df_movies['duration'].astype(np.int64)
```

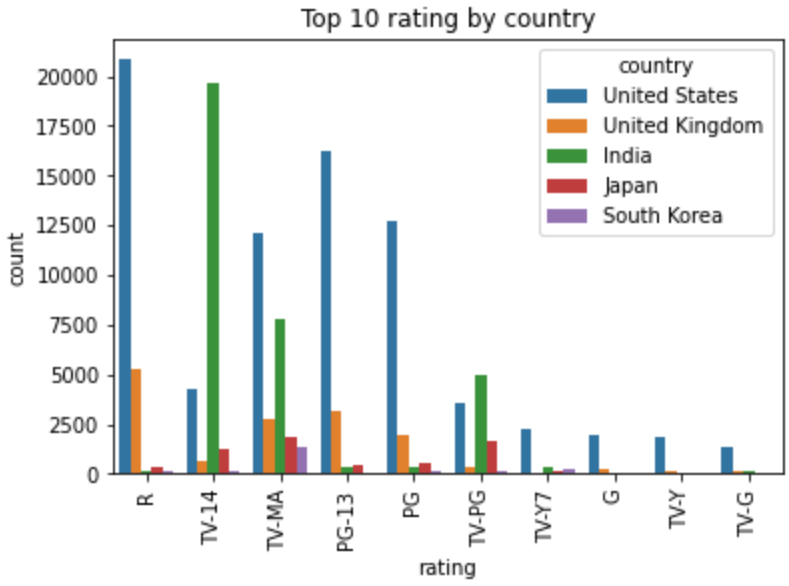
```
C:\Users\hp\AppData\Local\Temp/ipykernel_19412/3073083368.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  df_movies['duration'] =df_movies['duration'].astype(np.int64)
```

In [639…

```
df_movies_intresting_countries = df_selected.loc[(df_selected["type"] =="Movie") & ((df_
```

In [640…

```
ax = sns.countplot(x="rating",data=df_movies_intresting_countries,hue="country",order=df
ax.set(title='Top 10 rating by country')
sns.move_legend(ax, "upper right")
plt.xticks(rotation=90)
plt.show()
```



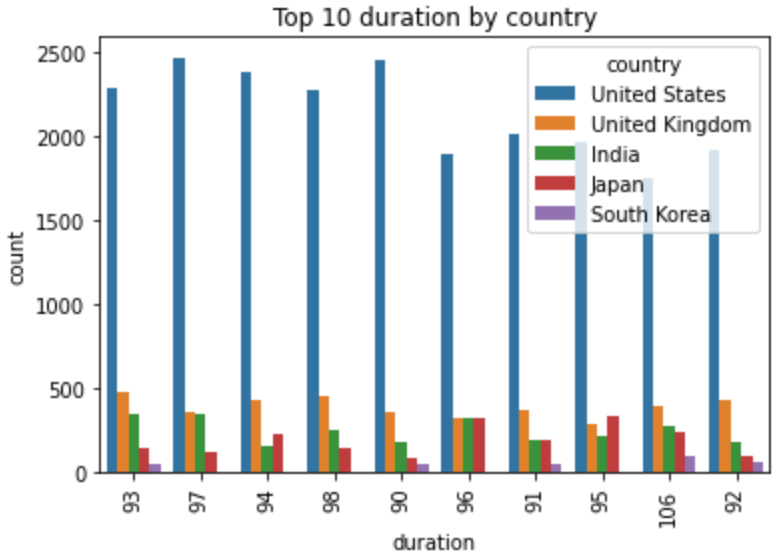
Observation :

A. In India - Other rated movies (which are frequently published in US and other countries) can be listed to see viewers response.

B. Likewise category movies in other countries can be listed in other countries , with audio or subtitle

In [641…

```
ax = sns.countplot(x="duration",data=df_movies_intresting_countries,hue="country",order=
ax.set(title='Top 10 duration by country')
sns.move_legend(ax, "upper right")
plt.xticks(rotation=90)
plt.show()
```

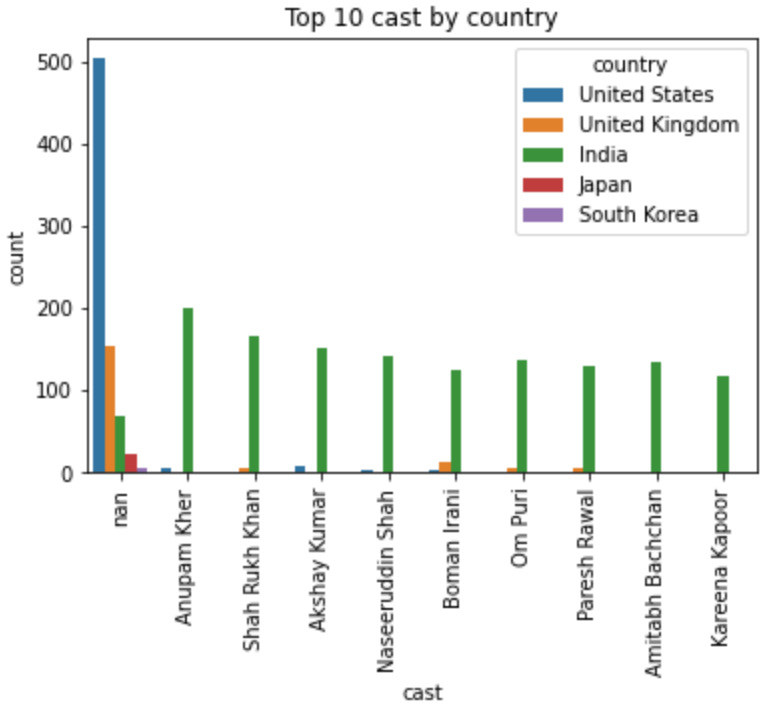


Observation :

A. Duration ranging from 90 to 106 minutes are generally being listed across top countries

In [642…

```
ax = sns.countplot(x="cast",data=df_movies_intresting_countries,hue="country",order=df_r
ax.set(title='Top 10 cast by country')
sns.move_legend(ax, "upper right")
plt.xticks(rotation=90)
plt.show()
```

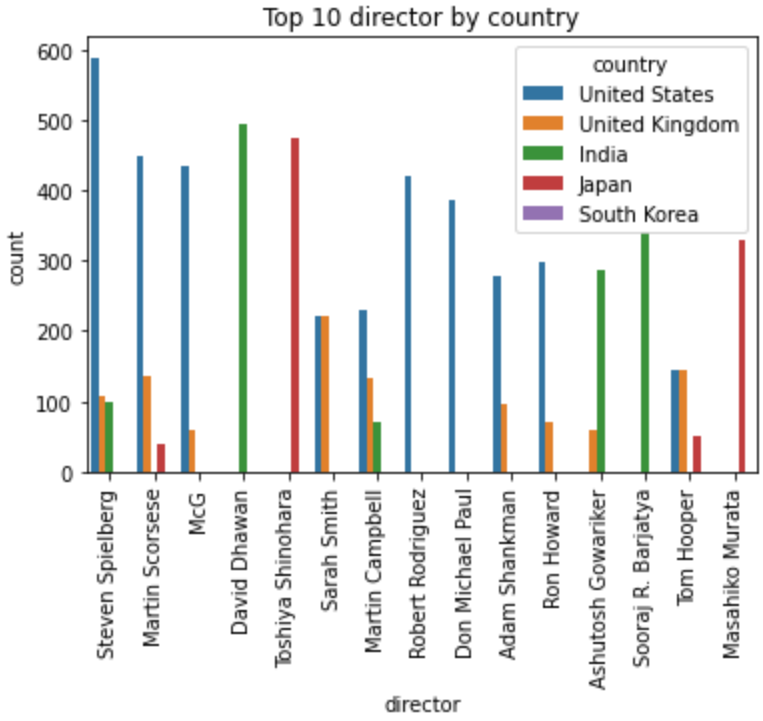


Observation :

A. Except India - There is not many cast centric listing in other countries

In [643...

```
ax = sns.countplot(x="director",data=df_movies_intresting_countries,hue="country",order=
ax.set(title='Top 10 director by country')
sns.move_legend(ax, "upper right")
plt.xticks(rotation=90)
plt.show()
```



Observation :

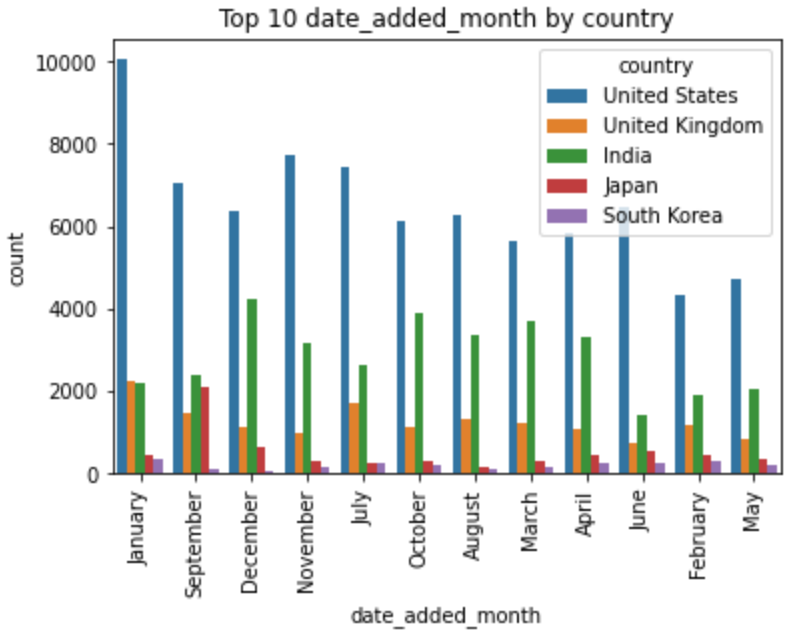
A. India - David Dhawan , Asutosh Gowariker and Sooraj R. Barjatya are mostly listed directors

B. Japan - Toshiya Shinohara , Masahiko Murata are mostly listed directors

C. Sahra Smith has been listed in both US and UK

In [644…

```
ax = sns.countplot(x="date_added_month",data=df_movies_intresting_countries,hue="country
ax.set(title='Top 10 date_added_month by country')
sns.move_legend(ax, "upper right")
plt.xticks(rotation=90)
plt.show()
```



Observation :

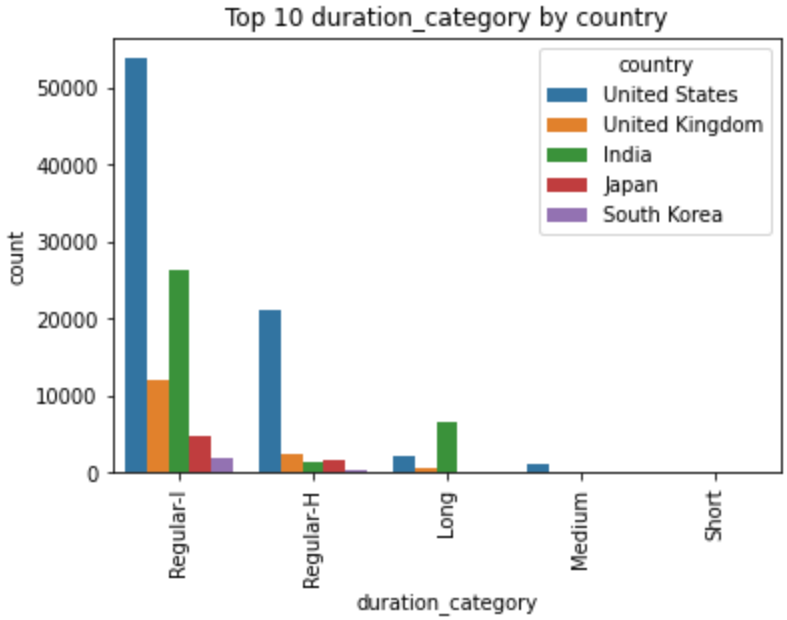
A. US - Listing is relatively high in Jan , Sep, Nov , July and June

B. Japan - September has highest listing

C. India - Most listing in Dec, Oct, Nov, Aug , March , April etc.

In [645…

```
ax = sns.countplot(x="duration_category",data=df_movies_intresting_countries,hue="countr
ax.set(title='Top 10 duration_category by country')
sns.move_legend(ax, "upper right")
plt.xticks(rotation=90)
plt.show()
```



Observation :

A. Any specific pattern is not observed wrt. duration_category

B. Customization based on country can be explored

In [646…

```
#categorical_features = df_movies.select_dtypes(exclude=['int64','float64']).columns
#categorical_features[1:] # Excluding identity column show_id
df_movies_intresting_countries["duration_category"]
```

Out[646]:

```
0          Regular-H
427        Regular-I
428        Regular-I
429        Regular-I
430        Regular-I
             ...
384215     Regular-I
384216     Regular-I
384217     Regular-I
384218     Regular-I
384219     Regular-I
Name: duration_category, Length: 135773, dtype: object
```

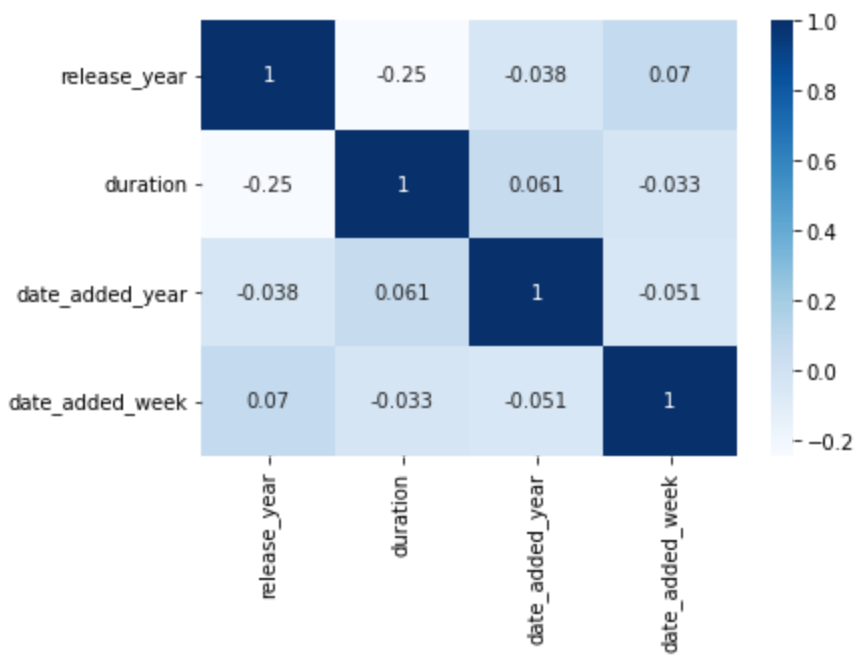
4.3 For correlation: Heatmaps, Pairplots

In [647…

```
sns.heatmap(df_movies.corr(), cmap="Blues", annot=True)
```

Out[647]:

```
<AxesSubplot:>
```



Observation :

A. Little correlation bteween added week and year of release

B. likewise , corr between date added year and duration

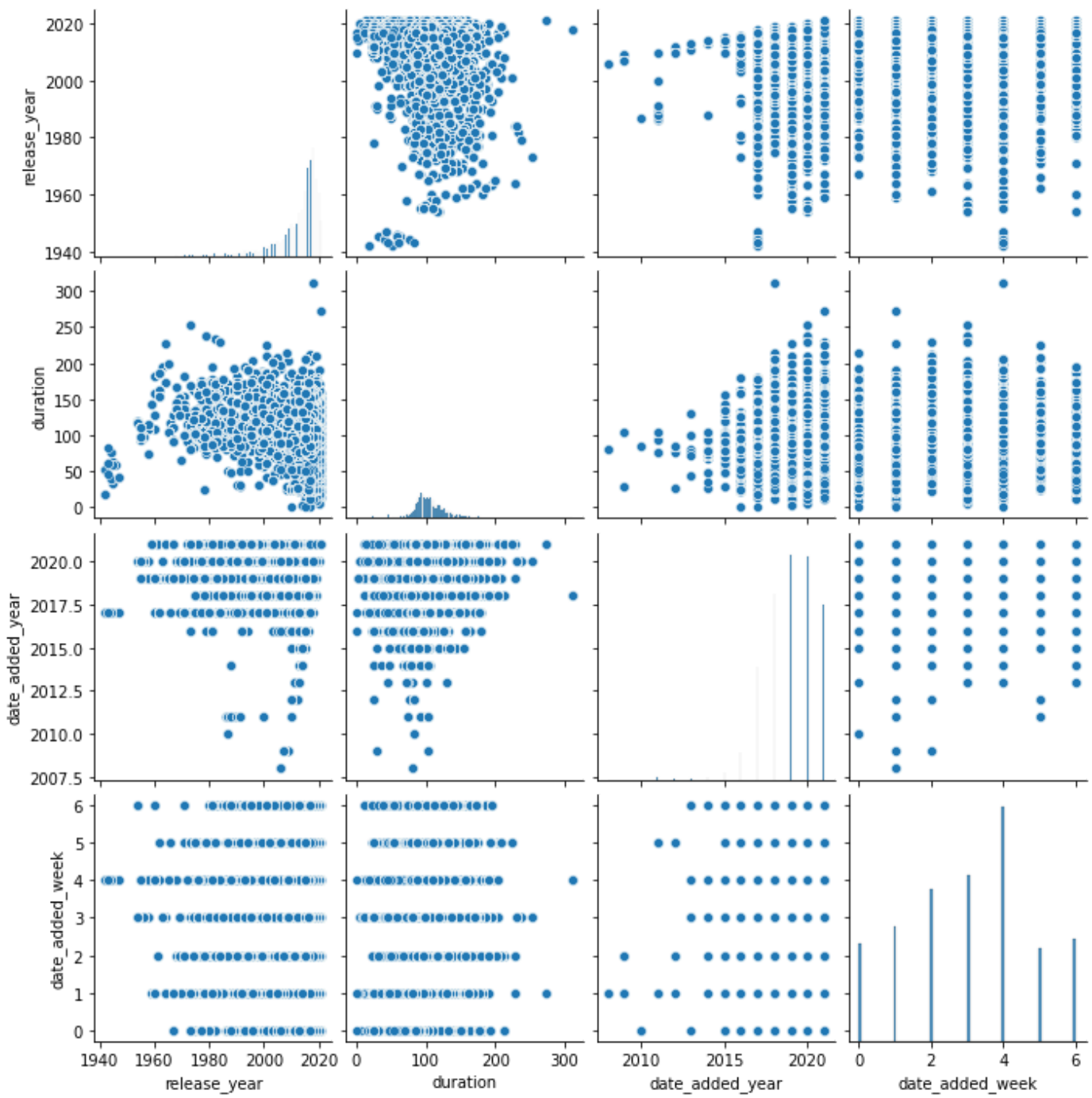
C. Negative corr between duration and year of release .

In [648…

```
sns.pairplot(data=df_movies)
plt.plot()
```

Out[648]:

```
[]
```

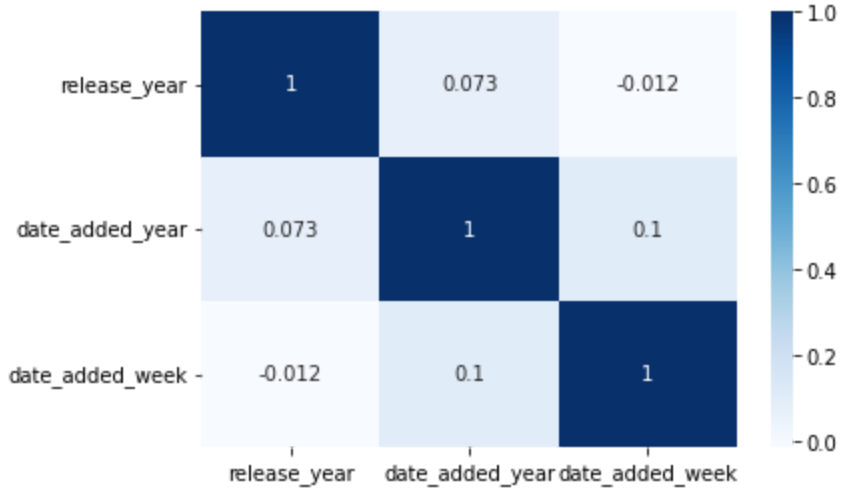


In [649…

```
sns.heatmap(df_series.corr(), cmap="Blues", annot=True)
```

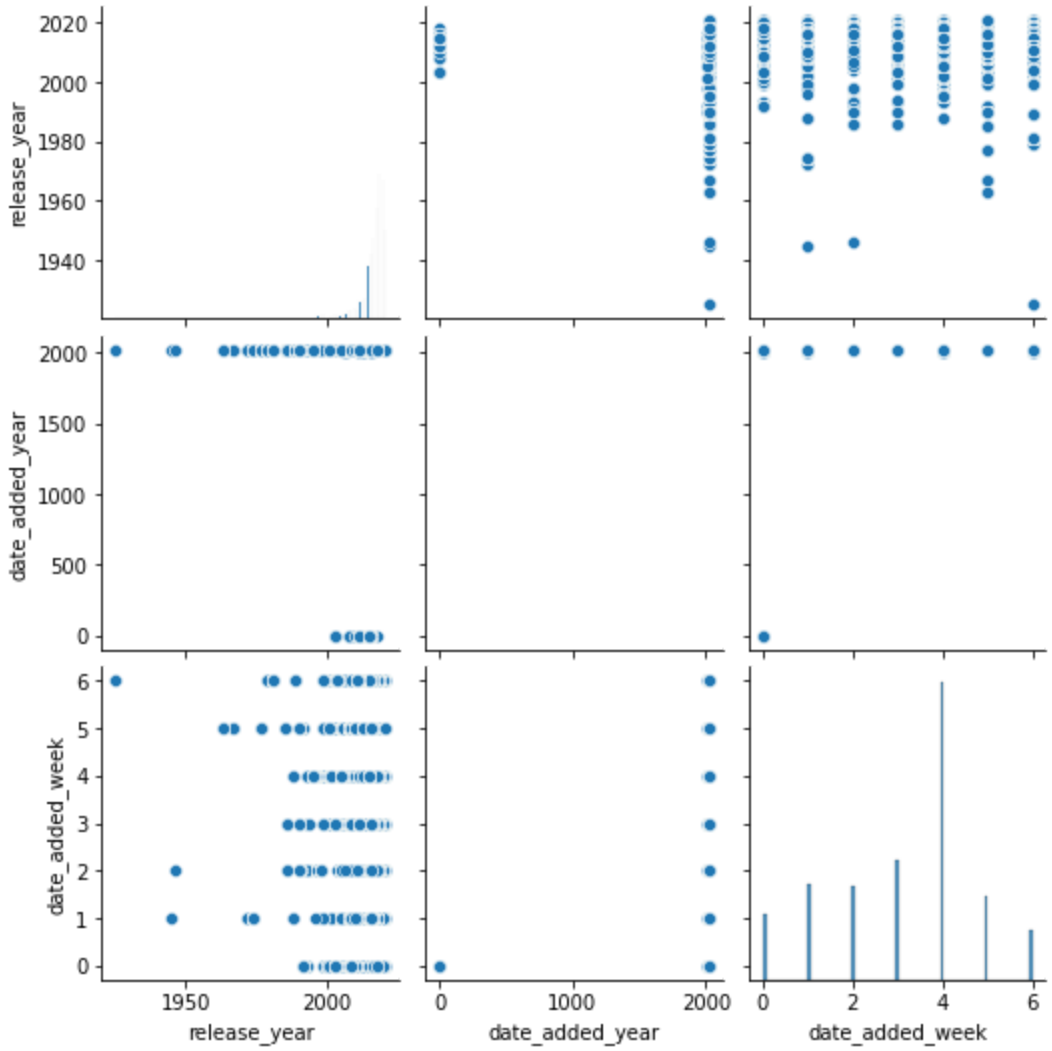
Out[649]:

<AxesSubplot:>



In [650…
```
sns.pairplot(data=df_series)
plt.plot()
```

Out[650]:
```
[]
```



5.2 Outlier check

In [651…
```
# Z score
from scipy import stats
```

In [652…
```
z_release_year = np.abs(stats.zscore(df_selected['release_year']))
df_selected.loc[np.where(z_release_year > 3)]
```

Out[652]:

	type	director	release_year	rating	duration	cast	country	listed_in	date_added_year	date_a
2095	Movie	Steven Spielberg	1975	PG	124	Roy Scheider	United States	Action	2021	
2096	Movie	Steven Spielberg	1975	PG	124	Roy Scheider	United States	&	2021	
2097	Movie	Steven Spielberg	1975	PG	124	Roy Scheider	United States	Adventure	2021	
2098	Movie	Steven Spielberg	1975	PG	124	Roy Scheider	United States	Classic	2021	
2099	Movie	Steven Spielberg	1975	PG	124	Roy Scheider	United States	Movies	2021	
...	...	...	...	...	...	...	...	...	...	
383524	Movie	Mu Chu	1973	NR	81	Nan Chiang	Hong Kong	Action	2016	
383525	Movie	Mu Chu	1973	NR	81	Nan Chiang	Hong Kong	&	2016	
383526	Movie	Mu Chu	1973	NR	81	Nan Chiang	Hong Kong	Adventure	2016	
383527	Movie	Mu Chu	1973	NR	81	Nan Chiang	Hong Kong	International	2016	
383528	Movie	Mu Chu	1973	NR	81	Nan Chiang	Hong Kong	Movies	2016	

10167 rows × 13 columns

Observation :

There is outliers for feature "release_year" as there few old movies starting from 1975 onwards

6. Insights based on Non-Graphical and Visual Analysis

- **Overview** - The dataset contains data about only Movie and TV shows
- **Identity features** - 'show_id', 'title' are identity features , won't add much value for data analysis. Hence can be ignored
- **Data Quality** - feature **'director'** contains **many empty** values (34 % of entire dataset)
  - Both TV Show and Movie, there are missing 'country' and 'director'
  - **'ratings'** has **non-homoginious data**
  - 'duration' value for movies in string (i.e. 74 mins , 84 min etc.)
  - TV shows in terms of seasons number , hence pre processsing is must.
- **Preprocessing** - **'casts', 'country' , 'listed_in'** have **comma seperated values** , which needs to be pre-processed before further analysis.
- **Visual Analysis**

  - In year **2019 - 2020 most of Movies and Series were released** in Netflix
  - Both movies and Series are **more often added on Friday**(i.e. week day # 4)

- **Wide range** of shows/movies - from **very short to very long duration**, even across different ratings
- Both shows/movies **data are skewed , with many outliers** on each type
- **United states and India** release **more movies** than TV Shows in Neflix platform
- **Japan , South Korea and Mexico** release **more TV Shows** than movies in Neflix platform
- **Noam Murro and Thomas Astruc more often** releases **TV Shows** in Neflix platform
- **Steven Spielberg , Raja Gosnell** have **more movies** in Neflix platform
- Netflix is showing **more movies of Anupam Kher , Shah Rukh Khan , Akshyay Kumar , Boman Irani** etc.
- **Many** Movies/shows have **no cast information tagged**, which can be improved for deep dive analysis
- **David Attenborough, Takahiro Sakurai , Yuki Kaji** have **more TV shows** than movies in Netflix platform
- **Less Comedy,Thriller, and no Children(i.e. "TV shows")** in Netflix platform, which can be explored
- **Too many Genres to choose from**
- **In India** - **Other rated movies** (which are frequently published in US and other countries) have **fewer release**.
- Likewise category movies in other countries can be listed in other countries , with audio or subtitle
- Duration ranging from 90 to 106 minutes are generally being listed across top countries
- **Except India** - There is **not many cast centric listing** in other countries
- **India** - **David Dhawan , Asutosh Gowariker and Sooraj R. Barjatya** are **mostly listed directors**
- Japan - Toshiya Shinohara , Masahiko Murata are mostly listed directors
- Sahra Smith has been listed in both US and UK
- **US** - Listing is relatively high in **Jan**(New year) , Sep, Nov , **July** (Independence day) and June
- **Japan - September has highest listing**
- **India** - **Most listing in Dec, Oct, Nov, Aug**(second half of the year) , March , April etc.
- No specific pattern is not observed wrt. duration_category

In [653…

```
df_selected.describe(include=object)
```

Out[653]:

	type	director	rating	duration	cast	country	listed_in	date_added_month	date_added_day	dur
count	384220	253193	384073	384220	384220	384220	384220	384220	384220	
unique	2	4528	15	221	36440	128	43	13	8	
top	Movie	Steven Spielberg	TV-MA	1 Season	nan	United States	Movies	July	Friday	
freq	242047	840	140910	87664	3398	111473	57642	37498	111675	

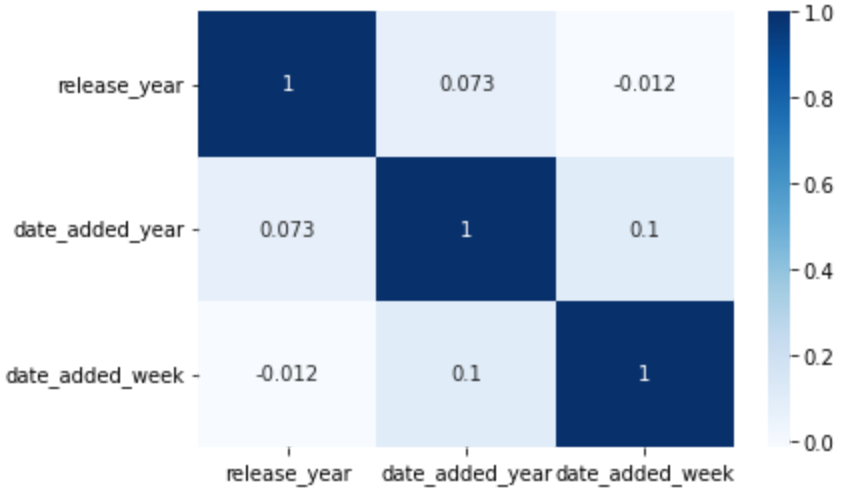
6.1 Comments on the range of attributes

- **Range of attributes**
  - **Release year** ranges from 1942 to 2021 , but most of movies are released in 2019 (75 %) and 2020
    - 50 % released in 2017
    - 25 % released in 2012
  - **Date added year** most of movies are added in 2020 (75 %) and latest by 2021
  - **Top stats** ranges from 0 to 6 most of movies are added in Friday(75 %) and latest by 2021
    - type : Movie (i.e. less TV shows)
    - director : Steven Spielberg
    - rating : TV-MA
    - season : 1 Season
    - country: United states
    - Month added - **July (may be due to US independence day i.e. holiday season)**
    - duration category - Regular I(ranges from 91 minutes to 150 minutes)
  - **Casting information** is missing for majority listing
  - **Range** release and date added are not uniform , very much skewed around 2019 and 2020

6.2 Comments on the distribution of the variables and relationship between them

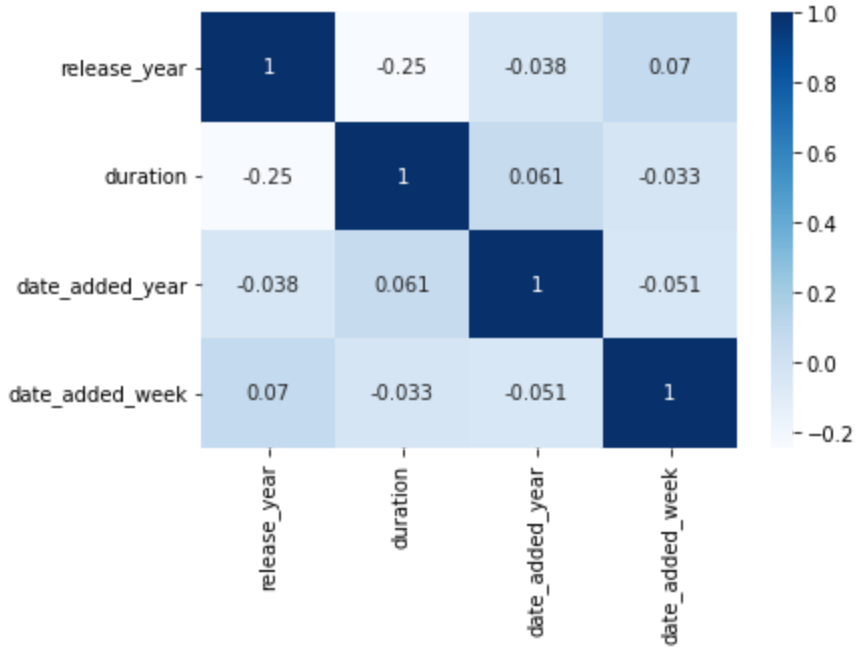
In [654…
```
sns.heatmap(df_series.corr(), cmap="Blues", annot=True)
```

Out[654]:
```
<AxesSubplot:>
```



In [655…
```
sns.heatmap(df_movies.corr(), cmap="Blues", annot=True)
```

Out[655]:
```
<AxesSubplot:>
```



- **Distribution of variables**
  - **Overall Data spread not uniform** , very much **skewed around 2019 and 2020**
    - **type** : more data point of movies than TV shows
    - **country** : more data points in United states and India , less in other countries
    - **rating** : less data points across categories and countries
- **Relationship between variables**
  - **Movies**
    - **release_year and date-added-year** - There is minor negative correlation (i.e.-0.038)
    - **release_year and date-added-week** - There is minor positive correlation (i.e. 0.07)
    - **release_year and duration** - There is decent negative correlation (i.e. -0.25)
  - **Series**
    - **release_year and date-added-year** - There is minor positive correlation (i.e. 0.073)
    - **release_year and date-added-week** - There is minor negative correlation (i.e. -0.012)
- **Categorical variables** - relationship has been mentioned in section 6 above

6.3 Comments for each univariate and bivariate plot

- **Categorical variables** - Comments for each univariate and bivariate plot have been mentioned in section 6 above

7. Business Insights - Should include patterns observed in the data along with what you can infer from it

- **Business Insights**
  - **Country wise focus**: More focus in US and India considering the relative volume of listing in those countries

- **Categories** : Movies catogory listing is primary and "TV Shows" are catching up based on pattern observed in other countries like Japan , Korea etc.
- **Holiday list**: More fresh listing towards holidays - such as weekends or holiday seasons in July(i.e. specific to US) , December etc.
- **Region specific customization for listing** . e.g . more actor centric listing in India etc.
- **Range**: Wide range of movies to target all category of viewers e.g. different ratings , different Genres, different durations etc.

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

- **Recommendations**
  - **Cutomized Genres**: There are too many options for consumers , if selected Genres or recommendation based on review pattern in the region would help
  - **Duration** : Average duration is very high for most of the listing . Short duration movies /series is an opportunity if some viewers don't have much time to spend
  - **New categories beyond Movies / TV shows** : More specificalized Genres with **short duration** can be experimented
    - **Business conferences**
    - **Technology innovations**
    - **Audio content** : like podcast or audio streaming platform can be one option that consumers can explore
  - **More customization**: More customization based on country , viewers pattern etc. would help get competitive advantage