Netflix_Case_Study

Importing the libraries

In [70]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Loading the dataset

In [71]:
```
!gdown https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/
```

```
Downloading...
From: https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/or
iginal/netflix.csv
To: /content/netflix.csv
100% 3.40M/3.40M [00:01<00:00, 1.95MB/s]
```

In [72]:
```
# Importing the dataset
df = pd.read_csv('netflix.csv')
df
```

Out[72]:

	show_id	type	title	director	cast	country	date_added	rel
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	
...	...	...	...	...	...	...	...	
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	

	show_id	type	title	director	cast	country	date_added	rel
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	

8807 rows × 12 columns

In [73]:
```
df.shape
```

Out[73]:
```
(8807, 12)
```

In [74]:
```
# Checking for missing values in the dataset
df.isnull().sum()
```

Out[74]:

	0
show_id	0
type	0
title	0
director	2634
cast	825
country	831
date_added	10
release_year	0
rating	4
duration	3
listed_in	0
description	0

dtype: int64

we have missing values for the following columns(director,cast,country,date_addded,rating and duration)

Handling Missing Values in Specific Columns

Filling NaN value with the column Cast, Director, Country, Rating and Durationto the initital dataframe

In [75]:
```
country_mode = df['country'].mode()[0]
country_mode
```

Out[75]:
```
'United States'
```

In [76]:
```
movie_rating = df.loc[df['type'] == 'Movie', 'rating'].mode()[0]
movie_rating
```

Out[76]:
```
'TV-MA'
```

In [77]:
```
df["new"]=df.loc[df["type"]=="movie","rating"]
```

In [78]:
```
tv_rating = df.loc[df['type'] == 'TV Show', 'rating'].mode()[0]
tv_rating
```

Out[78]:
```
'TV-MA'
```

In [79]:
```
movie_duration_mode = df.loc[df['type'] == 'Movie', 'duration'].mode()[0]
movie_duration_mode
```

Out[79]:
```
'90 min'
```

In [80]:
```
tv_duration_mode = df.loc[df['type'] == 'TV Show', 'duration'].mode()[0]
tv_duration_mode
```

Out[80]:
```
'1 Season'
```

- For the ‘director’ and ‘cast’ columns, we replace missing values with 'unknown director’ and "unknown cast"to maintain data integrity and avoid any bias in the analysis.
- In the ‘country’ column, we fill in missing values with the mode (most frequently occurring value) to ensure consistency and minimize data loss.
- For the ‘rating’ column, we fill in missing values based on the ‘type’ of the show. We assign the mode of ‘rating’ for movies and TV shows separately.
- For the ‘duration’ column, we fill in missing values based on the ‘type’ of the show. We assign the mode of ‘duration’ for movies and TV shows

separately.

In [81]:
```
df['director'].fillna('Unknown Director',inplace = True)
df['cast'].fillna('Unknown cast',inplace = True)
df['rating'] = df.apply(lambda x: movie_rating if x['type'] == 'Movie' and pd.
                        else tv_rating if x['type'] == 'TV Show' and pd.isna(x
                        else x['rating'], axis=1)
df['duration'] = df.apply(lambda x: movie_duration_mode if x['type'] == 'Movie
                          and pd.isna(x['duration'])
                          else tv_duration_mode if x['type'] == 'TV Show'
                          and pd.isna(x['duration'])
                          else x['duration'], axis=1)


df['country'].fillna( country_mode, inplace=True)



df
```

```
/tmp/ipython-input-1812104441.py:1: FutureWarning: A value is trying to be set
on a copy of a DataFrame or Series through chained assignment using an inplace
method.
The behavior will change in pandas 3.0. This inplace method will never work bec
ause the intermediate object on which we are setting values always behaves as a
copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.me
thod({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, t
o perform the operation inplace on the original object.


  df['director'].fillna('Unknown Director',inplace = True)
/tmp/ipython-input-1812104441.py:2: FutureWarning: A value is trying to be set
on a copy of a DataFrame or Series through chained assignment using an inplace
method.
The behavior will change in pandas 3.0. This inplace method will never work bec
ause the intermediate object on which we are setting values always behaves as a
copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.me
thod({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, t
o perform the operation inplace on the original object.


  df['cast'].fillna('Unknown cast',inplace = True)
/tmp/ipython-input-1812104441.py:13: FutureWarning: A value is trying to be set
on a copy of a DataFrame or Series through chained assignment using an inplace
method.
The behavior will change in pandas 3.0. This inplace method will never work bec
ause the intermediate object on which we are setting values always behaves as a
copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.me
thod({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, t
o perform the operation inplace on the original object.


  df['country'].fillna( country_mode, inplace=True)
```

Out[81]:

	show_id	type	title	director	cast	country	date_added	rel
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown cast	United States	September 25, 2021	
1	s2	TV Show	Blood & Water	Unknown Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	United States	September 24, 2021	
3	s4	TV Show	Jailbirds New Orleans	Unknown Director	Unknown cast	United States	September 24, 2021	
4	s5	TV Show	Kota Factory	Unknown Director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	
...	...	...	...	...	...	...	...	
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	
8803	s8804	TV Show	Zombie Dumb	Unknown Director	Unknown cast	United States	July 1, 2019	
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	

	show_id	type	title	director	cast	country	date_added	rel
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	

8807 rows × 13 columns

In [82]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 13 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      8807 non-null   object
 4   cast          8807 non-null   object
 5   country       8807 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8807 non-null   object
 9   duration      8807 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
 12  new           0 non-null      object
dtypes: int64(1), object(12)
memory usage: 894.6+ KB
```

Un-nesting the dataset

unnesting column of director with titles

In [83]:
```
split_director = df['director'].str.split(',', expand=True)
result_df = pd.concat([df['title'], split_director], axis=1)
dfdirector=pd.melt(result_df, id_vars=["title"], value_name="director")
dfdirector.drop(['variable'], axis=1, inplace=True)
unique_director=dfdirector.drop_duplicates()
```

```
dfunique_dr= unique_director.dropna()
dfunique_dr
```

Out[83]:

	title	director
0	Dick Johnson Is Dead	Kirsten Johnson
1	Blood & Water	Unknown Director
2	Ganglands	Julien Leclercq
3	Jailbirds New Orleans	Unknown Director
4	Kota Factory	Unknown Director
...	...	...
95585	Movie 43	Rusty Cundieff
102764	Walt Disney Animation Studios Short Films Coll...	Mike Gabriel
103787	HALO Legends	Hiroshi Yamazaki
104392	Movie 43	James Gunn
111571	Walt Disney Animation Studios Short Films Coll...	Mark Henn

9612 rows × 2 columns

Unnesting column of cast with **titles**

In [84]:

```
split_cast = df['cast'].str.split(',', expand=True)
result_df = pd.concat([df['title'], split_cast], axis=1)
dcast=pd.melt(result_df, id_vars=["title"], value_name="cast")
dcast.drop(['variable'], axis=1, inplace=True)
unique_cast=dcast.drop_duplicates()
dcastunique= unique_cast.dropna()
dcastunique
```

Out[84]:

	title	cast
0	Dick Johnson Is Dead	Unknown cast
1	Blood & Water	Ama Qamata
2	Ganglands	Sami Bouajila
3	Jailbirds New Orleans	Unknown cast
4	Kota Factory	Mayur More
...	...	...
417703	Black Mirror	Jon Hamm
424590	Social Distance	Ayize Ma'at
426510	Black Mirror	Oona Chaplin
433397	Social Distance	Lovie Simone
435317	Black Mirror	Rafe Spall

64949 rows × 2 columns

Unnesting column of country with **titles**

In [85]:

```
split_country = df['country'].str.split(',', expand=True)
result_df = pd.concat([df['title'], split_country], axis=1)
dfcountry=pd.melt(result_df, id_vars=["title"], value_name="country")
dfcountry.drop(['variable'], axis=1, inplace=True)
unique_country=dfcountry.drop_duplicates()
dfcountryunique= unique_country.dropna()
dfcountryunique
```

Out[85]:

	title	country
0	Dick Johnson Is Dead	United States
1	Blood & Water	South Africa
2	Ganglands	United States
3	Jailbirds New Orleans	United States
4	Kota Factory	India
...	...	...
78859	The Look of Silence	Germany
85496	Barbecue	Sweden
87666	The Look of Silence	Netherlands
94303	Barbecue	United States
103110	Barbecue	Uruguay

10850 rows × 2 columns

In [86]:

```
split_listed_in = df['listed_in'].str.split(',', expand=True)
result_df = pd.concat([df['title'], split_listed_in], axis=1)
dflist=pd.melt(result_df, id_vars=["title"], value_name="listed_in")
dflist.drop(['variable'], axis=1, inplace=True)
dfunique_list=dflist.drop_duplicates()
dfuniquelist= dfunique_list.dropna()
dfuniquelist
```

Out[86]:

	title	listed_in
0	Dick Johnson Is Dead	Documentaries
1	Blood & Water	International TV Shows
2	Ganglands	Crime TV Shows
3	Jailbirds New Orleans	Docuseries
4	Kota Factory	International TV Shows
...	...	...
26414	Zindagi Gulzar Hai	TV Dramas
26415	Zinzana	Thrillers
26416	Zodiac	Thrillers
26417	Zombie Dumb	TV Comedies
26420	Zubaan	Music & Musicals

19323 rows × 2 columns

Merging and create the final table

In [87]:
```
d=pd.merge(dfunique_dr, dcastunique,on="title", how="outer")
d1=pd.merge(d, dfcountryunique, on="title", how="outer" )
d2=pd.merge(d1, dfuniquelist, on="title", how="outer"  )
d2
```

Out[87]:

	title	director	cast	country	listed_in
0	#Alive	Cho Il	Yoo Ah-in	South Korea	Horror Movies
1	#Alive	Cho Il	Yoo Ah-in	South Korea	International Movies
2	#Alive	Cho Il	Yoo Ah-in	South Korea	Thrillers
3	#Alive	Cho Il	Park Shin-hye	South Korea	Horror Movies
4	#Alive	Cho Il	Park Shin-hye	South Korea	International Movies
...	...	...	...	...	...
202053	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Yang Jeong-hwa	United States	Children & Family Movies
202054	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Jeon Tae-yeol	United States	Children & Family Movies
202055	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Shin Yong-woo	United States	Children & Family Movies
202056	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Lee So-young	United States	Children & Family Movies
202057	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	So-yeon	United States	Children & Family Movies

202058 rows × 5 columns

In [88]:
```
d3=df[["title", "show_id", "type", "date_added", "release_year", "rating", "du
d3
```

Out[88]:

	title	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	s1	Movie	September 25, 2021	2020	PG-13	90 min
1	Blood & Water	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
2	Ganglands	s3	TV Show	September 24, 2021	2021	TV-MA	1 Season
3	Jailbirds New Orleans	s4	TV Show	September 24, 2021	2021	TV-MA	1 Season
4	Kota Factory	s5	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
...	...	...	...	...	...	...	...
8802	Zodiac	s8803	Movie	November 20, 2019	2007	R	158 min
8803	Zombie Dumb	s8804	TV Show	July 1, 2019	2018	TV-Y7	2 Seasons
8804	Zombieland	s8805	Movie	November 1, 2019	2009	R	88 min
8805	Zoom	s8806	Movie	January 11, 2020	2006	PG	88 min
8806	Zubaan	s8807	Movie	March 2, 2019	2015	TV-14	111 min

8807 rows × 7 columns

In [89]:

```
df_final=pd.merge(d2, d3, on="title", how="inner")
df_final
```

Out[89]:

	title	director	cast	country	listed_in	show_id	type	date_adde
0	#Alive	Cho Il	Yoo Ah-in	South Korea	Horror Movies	s2037	Movie	Septemb 8, 202
1	#Alive	Cho Il	Yoo Ah-in	South Korea	International Movies	s2037	Movie	Septemb 8, 202
2	#Alive	Cho Il	Yoo Ah-in	South Korea	Thrillers	s2037	Movie	Septemb 8, 202
3	#Alive	Cho Il	Park Shin-hye	South Korea	Horror Movies	s2037	Movie	Septemb 8, 202
4	#Alive	Cho Il	Park Shin-hye	South Korea	International Movies	s2037	Movie	Septemb 8, 202
...	...	...	...	...	...	...	...	
202053	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Yang Jeong-hwa	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202054	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Jeon Tae-yeol	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202055	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Shin Yong-woo	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202056	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Lee So-young	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202057	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	So-yeon	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201

202058 rows × 11 columns

In [90]:
```
df_final
```

Out[90]:

	title	director	cast	country	listed_in	show_id	type	date_adde
0	#Alive	Cho Il	Yoo Ah-in	South Korea	Horror Movies	s2037	Movie	Septemb 8, 202
1	#Alive	Cho Il	Yoo Ah-in	South Korea	International Movies	s2037	Movie	Septemb 8, 202
2	#Alive	Cho Il	Yoo Ah-in	South Korea	Thrillers	s2037	Movie	Septemb 8, 202
3	#Alive	Cho Il	Park Shin-hye	South Korea	Horror Movies	s2037	Movie	Septemb 8, 202
4	#Alive	Cho Il	Park Shin-hye	South Korea	International Movies	s2037	Movie	Septemb 8, 202
...	...	...	...	...	...	...	...	
202053	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Yang Jeong-hwa	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202054	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Jeon Tae-yeol	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202055	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Shin Yong-woo	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202056	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Lee So-young	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202057	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	So-yeon	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201

202058 rows × 11 columns

In [91]:
```
df_final.dropna(inplace=True)
```

In [92]:
```
df_final.duplicated().sum()
```

Out[92]:
```
np.int64(0)
```

In [93]:
```
df_final
```

Out[93]:

	title	director	cast	country	listed_in	show_id	type	date_adde
0	#Alive	Cho Il	Yoo Ah-in	South Korea	Horror Movies	s2037	Movie	Septemb 8, 202
1	#Alive	Cho Il	Yoo Ah-in	South Korea	International Movies	s2037	Movie	Septemb 8, 202
2	#Alive	Cho Il	Yoo Ah-in	South Korea	Thrillers	s2037	Movie	Septemb 8, 202
3	#Alive	Cho Il	Park Shin-hye	South Korea	Horror Movies	s2037	Movie	Septemb 8, 202
4	#Alive	Cho Il	Park Shin-hye	South Korea	International Movies	s2037	Movie	Septemb 8, 202
...	...	...	...	...	...	...	...	
202053	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Yang Jeong-hwa	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202054	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Jeon Tae-yeol	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202055	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Shin Yong-woo	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202056	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Lee So-young	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202057	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	So-yeon	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201

201900 rows × 11 columns

In [94]:
```
df_final["duration"]=df_final["duration"].str.split(' ').str[0]
df_final['duration'] = df_final['duration'].astype(int)
```

In [95]:
```
df_final
```

Out[95]:

	title	director	cast	country	listed_in	show_id	type	date_adde
0	#Alive	Cho Il	Yoo Ah-in	South Korea	Horror Movies	s2037	Movie	Septemb 8, 202
1	#Alive	Cho Il	Yoo Ah-in	South Korea	International Movies	s2037	Movie	Septemb 8, 202
2	#Alive	Cho Il	Yoo Ah-in	South Korea	Thrillers	s2037	Movie	Septemb 8, 202
3	#Alive	Cho Il	Park Shin-hye	South Korea	Horror Movies	s2037	Movie	Septemb 8, 202
4	#Alive	Cho Il	Park Shin-hye	South Korea	International Movies	s2037	Movie	Septemb 8, 202
...	...	...	...	...	...	...	...	
202053	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Yang Jeong-hwa	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202054	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Jeon Tae-yeol	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202055	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Shin Yong-woo	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202056	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Lee So-young	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201
202057	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	So-yeon	United States	Children & Family Movies	s7109	Movie	Septemb 1, 201

201900 rows × 11 columns

We convert the ‘date_added’ column to datetime format using pd.to_datetime() to enable further analysis based on date-related attributes.

In [96]:
```
df_final["date_added"] = pd.to_datetime(df_final['date_added'], format='%B %d,
df_final["date_added"] = pd.to_datetime(df_final['date_added'])
df_final
```

Out[96]:

	title	director	cast	country	listed_in	show_id	type	date_adde
0	#Alive	Cho Il	Yoo Ah-in	South Korea	Horror Movies	s2037	Movie	2020-09-(
1	#Alive	Cho Il	Yoo Ah-in	South Korea	International Movies	s2037	Movie	2020-09-(
2	#Alive	Cho Il	Yoo Ah-in	South Korea	Thrillers	s2037	Movie	2020-09-(
3	#Alive	Cho Il	Park Shin-hye	South Korea	Horror Movies	s2037	Movie	2020-09-(
4	#Alive	Cho Il	Park Shin-hye	South Korea	International Movies	s2037	Movie	2020-09-(
...	...	...	...	...	...	...	...	
202053	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Yang Jeong-hwa	United States	Children & Family Movies	s7109	Movie	2018-09-(
202054	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Jeon Tae-yeol	United States	Children & Family Movies	s7109	Movie	2018-09-(
202055	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Shin Yong-woo	United States	Children & Family Movies	s7109	Movie	2018-09-(
202056	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	Lee So-young	United States	Children & Family Movies	s7109	Movie	2018-09-(
202057	최강전사 미니특공대 : 영웅의 탄생	Young Jun Lee	So-yeon	United States	Children & Family Movies	s7109	Movie	2018-09-(

201900 rows × 11 columns

Q1. Defining Problem Statement and Analysing basic metrics

Our primary objective is to identify the most promising types of shows to produce, thereby maximizing our growth potential in the entertainment industry. To achieve this goal, we need to comprehensively analyze the data and derive actionable insights that will enable us to make informed decisions on content creation and business expansion.

Analysing Basic Metrices

In [97]:
```
df_final.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 201900 entries, 0 to 202057
Data columns (total 11 columns):
 #   Column        Non-Null Count   Dtype         
---  ------        --------------   -----         
 0   title         201900 non-null  object        
 1   director      201900 non-null  object        
 2   cast          201900 non-null  object        
 3   country       201900 non-null  object        
 4   listed_in     201900 non-null  object        
 5   show_id       201900 non-null  object        
 6   type          201900 non-null  object        
 7   date_added    200312 non-null  datetime64[ns]
 8   release_year  201900 non-null  int64         
 9   rating        201900 non-null  object        
 10  duration      201900 non-null  int64         
dtypes: datetime64[ns](1), int64(2), object(8)
memory usage: 18.5+ MB
```

In [98]:
```
unique_counts = df_final.nunique()
print(unique_counts)
```

```
title            8797
director         5121
cast            39261
country           197
listed_in          73
show_id          8797
type                2
date_added       1699
release_year       74
rating             17
duration          210
dtype: int64
```

Q2. 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

In [99]:
```
df_final.shape
```

Out[99]:
```
(201900, 11)
```

In [100…
```
df_final.dtypes
```

Out[100…

	0
title	object
director	object
cast	object
country	object
listed_in	object
show_id	object
type	object
date_added	datetime64[ns]
release_year	int64
rating	object
duration	int64

dtype: object

In [101…

```
# Checking for missing values in the dataset
df_final.isnull().sum()
```

Out[101…

	0
title	0
director	0
cast	0
country	0
listed_in	0
show_id	0
type	0
date_added	1588
release_year	0
rating	0
duration	0

dtype: int64

- There is no need of changing any column to categorical column.

- Missing value already detected and fixed my filling it with unknown data and remaining filled with mode values We are going to ignore the missing value of "date_added" since very few data are missing

In [105…
```
df_final.drop_duplicates(inplace=True)
```

In [106…
```
#Statistical summary
df_final.describe().round(2)
```

Out[106…

	date_added	release_year	duration
count	200312	201900.00	201900.00
mean	2019-06-24 18:04:32.167418880	2013.45	77.74
min	2008-01-01 00:00:00	1925.00	1.00
25%	2018-07-01 00:00:00	2012.00	4.00
50%	2019-09-13 00:00:00	2016.00	95.00
75%	2020-09-15 00:00:00	2019.00	112.00
max	2021-09-25 00:00:00	2021.00	312.00
std	NaN	9.02	51.46

In [107…
```
df_final.describe(include="datetime")
```

Out[107…

	date_added
count	200312
mean	2019-06-24 18:04:32.167418880
min	2008-01-01 00:00:00
25%	2018-07-01 00:00:00
50%	2019-09-13 00:00:00
75%	2020-09-15 00:00:00
max	2021-09-25 00:00:00

submitted by Kanimozhi