

Business Case :

- Analyzing the data and generate insights that could help Netflix to decide which type of movies/show they produce to improve their business.

```
## importing all libraries

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

Basic Matrix

```
# analysing what is in the data
#upload a data
df = pd.read_csv('netflix.csv')
df
```

	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 l...
...

The shape of this data contains

8807 - number of rows 12 - number of columns


```
# information inside the data
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
```

The data has only one series of **integer** type and all other are **categorical** type

```
# analysing the value count of all the ratings and how many unique ratings are mentioned in the data
df['rating'].value_counts()
```




	count
rating	
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

dtype: int64

These are the **14** unique ratings categories in which category **TV-MA** has the highest number of movies and shows

```
# analyzing the unique country names and the number of shows of any particular countries having on netflix platform
df['country'].value_counts()
```



	count
country	
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

748 rows × 1 columns

dtype: int64

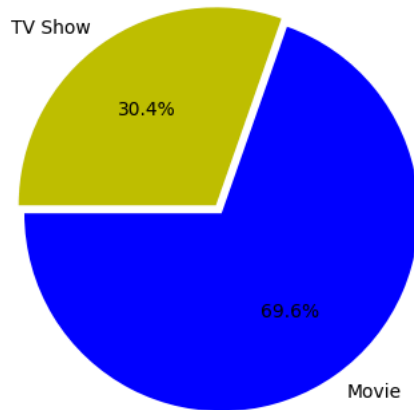
There are **748 Unique countries** on netflix which produces thier TV shows and movies on the platform and on which **United States** has the highest number of movies and TV shows on NETFLIX

Comparison of TV shows vs Movie

```
plt.title("Percentation of Netflix Titles that are either Movies or TV Shows")
g=plt.pie(df.type.value_counts(),explode=(0.025,0.025),
labels=df.type.value_counts().index, colors=['b','y'],autopct='%1.1f%%',
startangle=180)
plt.show()
```



Percentation of Netflix Titles that are either Movies or TV Shows



```
df1 = df['country'].apply(lambda x: str(x).split(',')).tolist()
df1
```



```
[[ 'United States'],
 [ 'South Africa'],
 [ 'nan'],
 [ 'nan'],
 [ 'India'],
 [ 'nan'],
 [ 'nan'],
 [ 'United States',
  'Ghana',
  'Burkina Faso',
  'United Kingdom',
  'Germany',
  'Ethiopia'],
 [ 'United Kingdom'],
 [ 'United States'],
 [ 'nan'],
 [ 'nan'],
 [ 'Germany', 'Czech Republic'],
 [ 'nan'],
 [ 'nan'],
 [ 'United States'],
 [ 'nan'],
 [ 'Mexico'],
 [ 'nan'],
 [ 'nan'],
 [ 'nan'],
 [ 'Turkey'],
 [ 'nan'],
 [ 'nan'],
 [ 'India'],
 [ 'Australia'],
 [ 'nan'],
 [ 'United States'],
 [ 'United States'],
 [ 'United States', 'India', 'France'],
 [ 'nan'],
 [ 'nan'],
 [ 'United Kingdom'],
 [ 'nan'],
 [ 'nan'],
 [ 'nan'],
 [ 'nan'],
 [ 'Finland'],
 [ 'China', 'Canada', 'United States'],
```

```
['India'],
['United States'],
['United States'],
['United States'],
['United States'],
['United States'],
['nan'],
['South Africa', 'United States', 'Japan'],
['nan'],
['United States'],
['Nigeria'],
['India'],
['Japan'],
['Japan']
```

```
df1 = pd.DataFrame(df1,index=df['title'])
df1
```

	0	1	2	3	4	5	6	7	8	9	10	11
title												
Dick Johnson Is Dead	United States	None	None	None	None	None	None	None	None	None	None	None
Blood & Water	South Africa	None	None	None	None	None	None	None	None	None	None	None
Ganglands	nan	None	None	None	None	None	None	None	None	None	None	None
Jailbirds New Orleans	nan	None	None	None	None	None	None	None	None	None	None	None
Kota Factory	India	None	None	None	None	None	None	None	None	None	None	None
...
Zodiac	United States	None	None	None	None	None	None	None	None	None	None	None
Zombie Dumb	nan	None	None	None	None	None	None	None	None	None	None	None
Zombieland	United States	None	None	None	None	None	None	None	None	None	None	None
Zoom	United States	None	None	None	None	None	None	None	None	None	None	None
Zubaan	India	None	None	None	None	None	None	None	None	None	None	None

8807 rows × 12 columns

```
df1 = df1.stack()
df1
```

	0
title	
Dick Johnson Is Dead	0 United States
Blood & Water	0 South Africa
Ganglands	0 nan
Jailbirds New Orleans	0 nan
Kota Factory	0 India
...
Zodiac	0 United States
Zombie Dumb	0 nan
Zombieland	0 United States
Zoom	0 United States
Zubaan	0 India

10845 rows × 1 columns

dtype: object

Tells the show/movie with thier country names

```
df1 = pd.DataFrame(df1)
df1
```



0

title

Dick Johnson Is Dead	0	United States
Blood & Water	0	South Africa
Ganglands	0	nan
Jailbirds New Orleans	0	nan
Kota Factory	0	India
...
Zodiac	0	United States
Zombie Dumb	0	nan
Zombieland	0	United States
Zoom	0	United States
Zubaan	0	India

10845 rows × 1 columns

```
df2 = pd.merge(df1, df, how='inner', on='title')
df2.shape
```



(10845, 13)

df2.head(5)



	title	0	show_id	type	director	cast	country	date_added	release_year	rating	duration	listed_in	descript
0	Dick Johnson Is Dead	United States	s1	Movie	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her fa near: end c life, film
1	Blood & Water	South Africa	s2	TV Show	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	A cross paths party, a C Towr
2	Ganglands	nan	s3	TV	Julien	Sami Bouajila, Tracy Gotsis	NaN	September	2021	TV-MA	1 Season	Crime TV Shows, International	To protect family fro

Merged the new country column to the previous dataframe

```
# #dropping one of the country column(old one)
# df2.drop(columns='country', inplace = True)
# df2.rename({0 : 'country'}, inplace= True, axis= 1)
df2.rename({0 : 'country1'}, inplace= True, axis=1)
```

```
# df2.head(3)
df2.head(2)
```



	title	country1	show_id	type	director	cast	country	date_added	release_year	rating	duration	listed_in	descrip
0	Dick Johnson Is Dead	United States	s1	Movie	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her fa near: end c life, film
						Ama Qamata,						International	

```
df3= df2['listed_in'].apply(lambda x :str(x).split(',')).tolist()
df3
```



```
['Documentaries'],
['International TV Shows', 'TV Dramas', 'TV Mysteries'],
['Crime TV Shows', 'International TV Shows', 'TV Action & Adventure'],
```

```

['Docuseries', ' Reality TV'],
['International TV Shows', ' Romantic TV Shows', ' TV Comedies'],
['TV Dramas', ' TV Horror', ' TV Mysteries'],
['Children & Family Movies'],
['Dramas', ' Independent Movies', ' International Movies'],
['Dramas', ' Independent Movies', ' International Movies'],
['Dramas', ' Independent Movies', ' International Movies'],
['Dramas', ' Independent Movies', ' International Movies'],
['Dramas', ' Independent Movies', ' International Movies'],
['Dramas', ' Independent Movies', ' International Movies'],
['British TV Shows', ' Reality TV'],
['Comedies', ' Dramas'],
['Crime TV Shows', ' Docuseries', ' International TV Shows'],
['Crime TV Shows', ' International TV Shows', ' TV Action & Adventure'],
['Dramas', ' International Movies'],
['Dramas', ' International Movies'],
['Children & Family Movies', ' Comedies'],
['British TV Shows', ' Crime TV Shows', ' Docuseries'],
['TV Comedies', ' TV Dramas'],
['Documentaries', ' International Movies'],
['Crime TV Shows', ' Spanish-Language TV Shows', ' TV Dramas'],
['Thrillers'],
['International TV Shows',
 ' Spanish-Language TV Shows',
 ' TV Action & Adventure'],
['Crime TV Shows', ' Docuseries', ' International TV Shows'],
['International TV Shows', ' TV Action & Adventure', ' TV Dramas'],
['Comedies', ' International Movies'],
['Children & Family Movies'],
['Comedies', ' International Movies', ' Romantic Movies'],
['Docuseries', ' International TV Shows', ' Reality TV'],
['Comedies', ' International Movies', ' Music & Musicals'],
['Comedies'],
['Horror Movies', ' Sci-Fi & Fantasy'],
['Thrillers'],
['Thrillers'],
['Thrillers'],
['Dramas', ' Independent Movies', ' International Movies'],
['TV Comedies'],
['British TV Shows', ' International TV Shows', ' TV Comedies'],
['International TV Shows', ' TV Dramas', ' TV Thrillers'],
["Kids' TV"],
['Dramas', ' International Movies', ' Thrillers'],
['Action & Adventure', ' Dramas', ' International Movies'],
["Kids' TV", ' TV Comedies'],
['Action & Adventure', ' Dramas'],
['Action & Adventure', ' Dramas'],
['Action & Adventure', ' Dramas'],
["Kids' TV"],
["Kids' TV", ' TV Sci-Fi & Fantasy'],
['Action & Adventure', ' Classic Movies', ' Dramas'],
['Dramas', ' Horror Movies', ' Thrillers'],
['Action & Adventure', ' Horror Movies', ' Thrillers'],
['Action & Adventure', ' Horror Movies', ' Thrillers'],
['Documentaries']

```

splitting all the categories by which we get to know which show or movie comes under which documentary

```

df3 = pd.DataFrame(df3,index= df2['title'])
df3

```



	0	1	2
title			
Dick Johnson Is Dead	Documentaries	None	None
Blood & Water	International TV Shows	TV Dramas	TV Mysteries
Ganglands	Crime TV Shows	International TV Shows	TV Action & Adventure
Jailbirds New Orleans	Docuseries	Reality TV	None
Kota Factory	International TV Shows	Romantic TV Shows	TV Comedies
...
Zodiac	Cult Movies	Dramas	Thrillers
Zombie Dumb	Kids' TV	Korean TV Shows	TV Comedies
Zombieland	Comedies	Horror Movies	None
Zoom	Children & Family Movies	Comedies	None
Zubaan	Dramas	International Movies	Music & Musicals

10845 rows × 3 columns

These are show or movies comes under various categories example - **Blood & Water** comes under **International TV Shows ,TV dramas and TV Mysteries**

```
df3 = df3.stack()
df3 = pd.DataFrame(df3)
df3
```



0

title		
Dick Johnson Is Dead	0	Documentaries
Blood & Water	0	International TV Shows
	1	TV Dramas
	2	TV Mysteries
Ganglands	0	Crime TV Shows
...
Zoom	0	Children & Family Movies
	1	Comedies
Zubaan	0	Dramas
	1	International Movies
	2	Music & Musicals

23754 rows × 1 columns

```
#merging data
```

```
df4 = pd.merge(df3, df2, how = 'inner', on = 'title')
df4.drop(columns = 'listed_in', inplace = True)
df4.rename({0:'listed_in'}, axis = 1, inplace = True)
# we can dropping show id and description as of now
df4.drop(columns = ['show_id', 'description'], inplace = True)
df4.shape
```



(37790, 11)

```
df4.head(2)
```

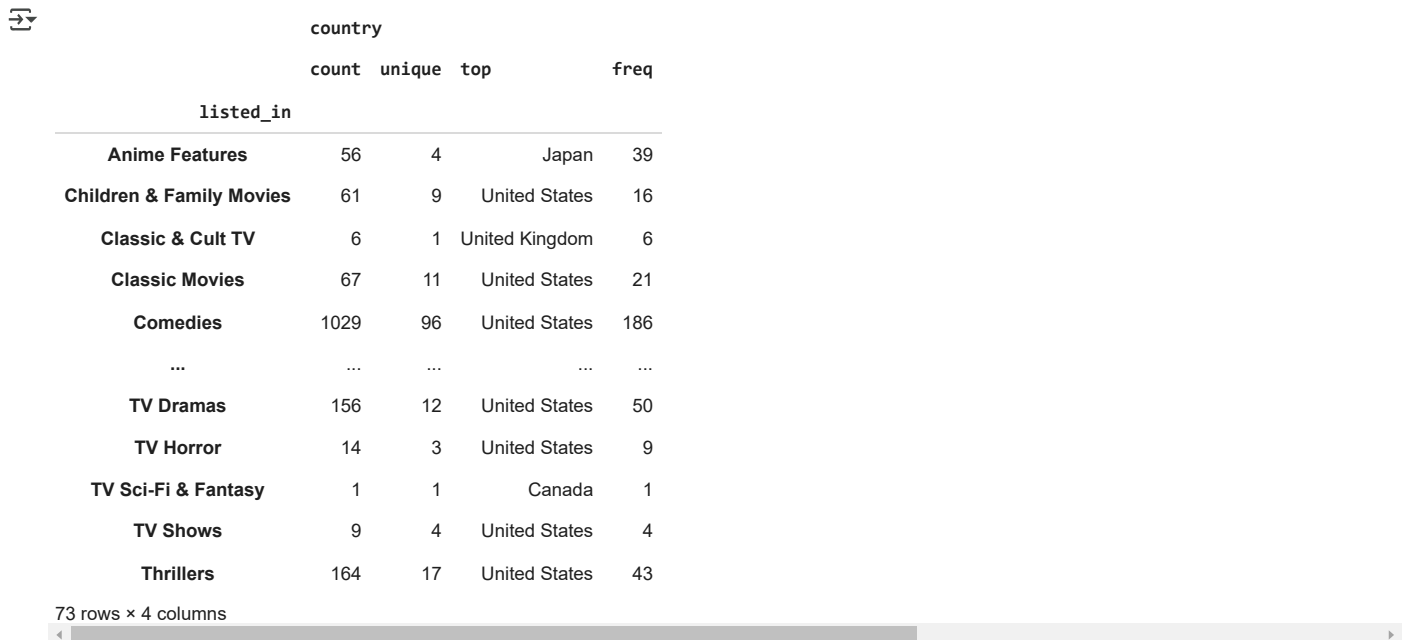


	title	listed_in	country1	type	director	cast	country	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Documentaries	United States	Movie	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min

Ama Qamata,

What type of Content is available in different countries

```
findings = df4.groupby(by = ['listed_in'])[['country']]
findings.describe()
```



	country			
	count	unique	top	freq
listed_in				
Anime Features	56	4	Japan	39
Children & Family Movies	61	9	United States	16
Classic & Cult TV	6	1	United Kingdom	6
Classic Movies	67	11	United States	21
Comedies	1029	96	United States	186
...
TV Dramas	156	12	United States	50
TV Horror	14	3	United States	9
TV Sci-Fi & Fantasy	1	1	Canada	1
TV Shows	9	4	United States	4
Thrillers	164	17	United States	43

73 rows × 4 columns

These are the insights of our findings example- anime features has 56 count and more in Japan

```
findings.nunique()
```



	country
listed_in	
Anime Features	4
Children & Family Movies	9
Classic & Cult TV	1
Classic Movies	11
Comedies	96
...	...
TV Dramas	12
TV Horror	3
TV Sci-Fi & Fantasy	1
TV Shows	4
Thrillers	17

73 rows × 1 columns

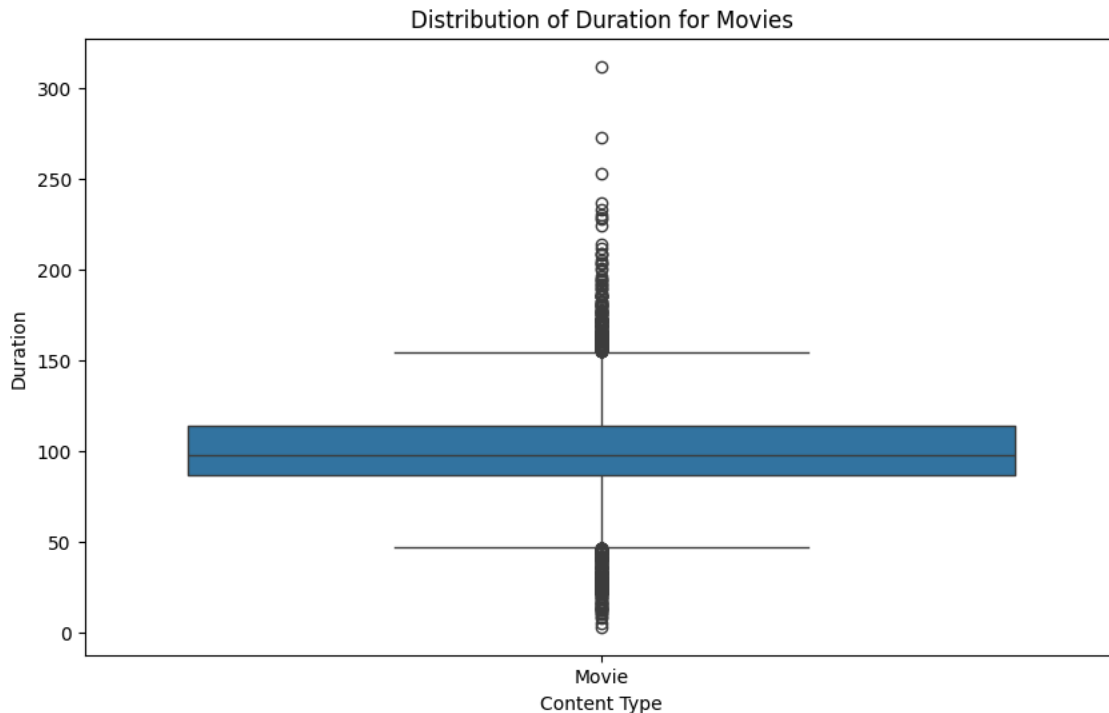
.Missing Value & Outlier check

```
netflix_movies_df = df[df.type.str.contains("Movie")]
netflix_movies_df['duration'] = netflix_movies_df['duration'].str.extract('(\d+)',
expand=False).astype(float)
# Creating a boxplot for movie duration
plt.figure(figsize=(10, 6))
sns.boxplot(data=netflix_movies_df, x='type', y='duration')
plt.xlabel('Content Type')
plt.ylabel('Duration')
plt.title('Distribution of Duration for Movies')
plt.show()
```



```
<ipython-input-190-ae75f7f6abb0>: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
`netflix_movies_df['duration'] = netflix_movies_df['duration'].str.extract('(\d+)',`



How has the number of movies released per year changed over the last 20-30 years?

df4

	title	listed_in	country1	type	director	cast	country	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Documentaries	United States	Movie	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min
1	Blood & Water	International TV Shows	South Africa	TV Show	NaN	Ama Qamata, Khosi Ngema, Gail Mabalan... Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons
2	Blood & Water	TV Dramas	South Africa	TV Show	NaN	Ama Qamata, Khosi Ngema, Gail Mabalan... Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons
3	Blood & Water	TV Mysteries	South Africa	TV Show	NaN	Ama Qamata, Khosi Ngema, Gail Mabalan... Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons
4	Ganglands	Crime TV Shows	nan	TV Show	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season
...
37785	Zoom	Children & Family Movies	United States	Movie	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma... Tim Allen,	United States	January 11, 2020	2006	PG	88 min

```
df4['date_added'] = pd.to_datetime(df4['date_added'], format='%B %d, %Y', errors='coerce')
df4.head()
```

	title	listed_in	country1	type	director	cast	country	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Documentaries	United States	Movie	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min
1	Blood & Water	International TV Shows	South Africa	TV Show	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons
2	Blood & Water	TV Dramas	South Africa	TV Show	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons

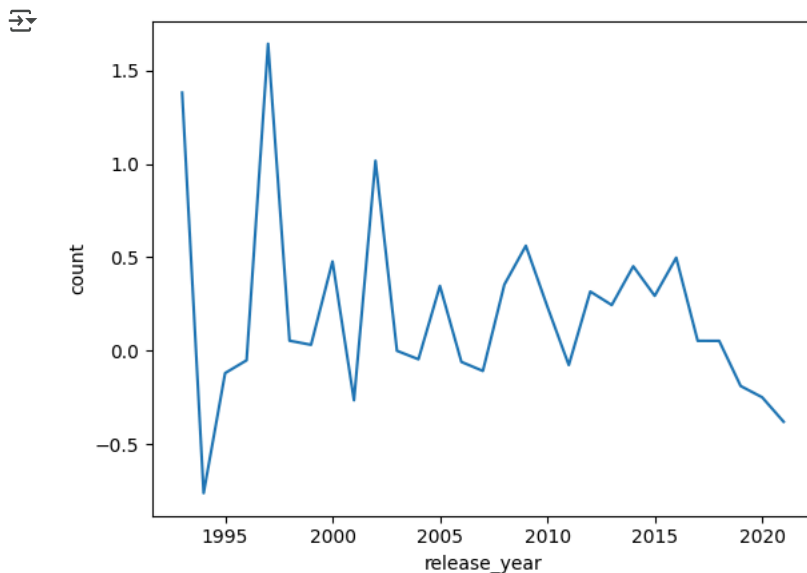
```
df4.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37790 entries, 0 to 37789
Data columns (total 11 columns):
#   Column          Non-Null Count  Dtype
---  -
0   title            37790 non-null  object
1   listed_in        37790 non-null  object
2   country1         37790 non-null  object
3   type             37790 non-null  object
4   director         29265 non-null  object
5   cast             34571 non-null  object
6   country          36068 non-null  object
7   date_added       37506 non-null  datetime64[ns]
8   release_year     37790 non-null  int64
9   rating           37784 non-null  object
10  duration         37787 non-null  object
dtypes: datetime64[ns](1), int64(1), object(9)
memory usage: 3.2+ MB
```

```
release_year_data = df4['release_year'].value_counts().sort_index(ascending = True)
percentage_change_yearly = release_year_data.tail(30).pct_change()
percentage_change_yearly = pd.DataFrame(percentage_change_yearly).reset_index()
percentage_change_yearly.head()
```

```
release_year  count
0            1992    NaN
1            1993  1.381356
2            1994 -0.761566
3            1995 -0.119403
4            1996 -0.050847
```

```
sns.lineplot(data= percentage_change_yearly, x = 'release_year', y = 'count')
plt.show()
```

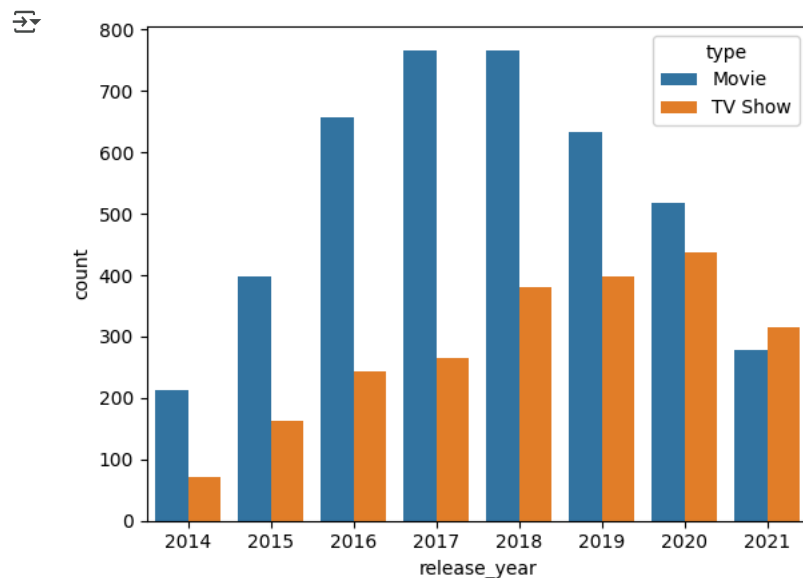


Comparison of TV shows vs Movies.

```
df4.head()
```

	title	listed_in	country1	type	director	cast	country	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Documentaries	United States	Movie	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min
1	Blood & Water	International TV Shows	South Africa	TV Show	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons
2	Blood & Water	TV Dramas	South Africa	TV Show	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons
						Ama Qamata.					

```
df5 = df.sort_values(by = 'release_year', ascending = False)
sns.countplot(data = df5.head(6500), x = 'release_year', hue = 'type')
plt.show()
```



What is the best time to launch a TV show

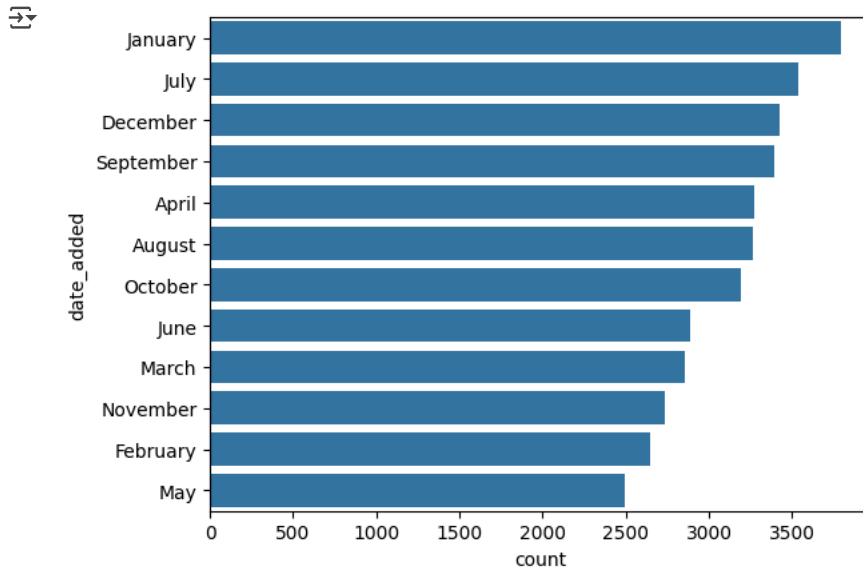
```
df4.head()
```

	title	listed_in	country1	type	director	cast	country	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Documentaries	United States	Movie	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min
1	Blood & Water	International TV Shows	South Africa	TV Show	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons
2	Blood & Water	TV Dramas	South Africa	TV Show	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons
						Ama Qamata.					

```
df6 = df4.sort_values(by = 'release_year', ascending = False)
month = df6['date_added'].dt.month_name()
month = month.value_counts()
month = month.reset_index()
month.head()
```

	date_added	count
0	January	3791
1	July	3535
2	December	3425
3	September	3397
4	April	3269

```
sns.barplot(data = month, y = 'date_added', x = 'count',)
plt.show()
```



In month of **January** the highest number of show would be release as there is more count

Summary of all the insights we cover through out the Project :

1- Basic matrix

- Shape - 8807 rows * 12 columns
- info - there are 10 columns all of OBJECT type and 1 column is of integer type
- there are 14 unique rating categories in which **TV-MA** is the category which has the most movies /shows
- there are 748 unique countries of which produced show/movies are on netflix.

2- Comparison of TV shows vs Movie

- TV- shows - 30.4%
- Movie- 69.6%

3-Non- Graphical Analysis-

- Ratings value counts - 14 unique categories
- countries - 748 unique countries
- listed_in categories - 73

4-Exploratory Analysis and Visualization

- Pie plot: Netflix Content By Type Analysis entire Netflix dataset consisting of both movies and shows. compare the total number of movies and shows in this dataset by which - [30.4% - TV Shows , 69.6%- Movies]
- Line chart - compare the TV show and movies changing during the time
- Countplot - compare the released categories over time for example there are 750+ movies released and 250+ tv shows were released

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

- Boxplot - Duration Distribution for Movies and TV Shows Analysing the duration distribution for movies and TV shows allows us to understand the typical length of content available on Netflix.

6- Insights based on Non-Graphical and Visual Analysis Ratings value counts - 14 unique categories

- countries - 748 unique countries
- listed_in categories - 73

7- Business Insights : With the help of this article, we have been able to learn about

1. Quantity: Our analysis revealed that Netflix had added more movies than TV shows, aligning with the expectation that movies dominate their content library.
2. Content Addition: July emerged as the month when Netflix adds the most content, closely followed by December, indicating a strategic approach to content release.
3. Genre Correlation: Strong positive associations were observed between various genres, such as TV dramas and international TV shows, romantic and international TV shows, and independent movies and dramas. These correlations provide insights into viewer preferences and content interconnections.
4. Movie Lengths: The analysis of movie durations indicated a peak around the 1960s, followed by a stabilization around 100 minutes, highlighting a trend in movie lengths over time.
5. TV Show Episodes: Most TV shows on Netflix have one season, suggesting a preference for shorter series among viewers.

8-RECOMMENDATIONS

- Netflix has to focus on TV Shows also because there are people who will like to see tv shows rather than movies
- By approaching the top director we can plan some more movies/tv shows in order to increase the popularity
- Not only reaching top director we can also see the director with less no of movies and having high rating as there may be some financial issues or anything so inorder to get good content netflix can reach to them and netflix can produce the movie and give the director a chance.
- We have seen most no of international movies genre so need to give priority to other genres like hooro,comedy..etc
- In TV Shows we may focus on thriller genre which will be helpfull for having more no of seasons