Business Case: Netflix - Data Exploration & Visualisation :

Business Problem : 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

```
#importing different libaries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

import warnings #to ignore the warnings & make our code more
representable
warnings.filterwarnings("ignore")

#Loading of dataset
df = pd.read_csv("/content/drive/MyDrive/netflix.csv")
df.head()
```

```
  show_id     type                  title         director  \
0      s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show          Blood & Water              NaN
2      s3  TV Show              Ganglands  Julien Leclercq
3      s4  TV Show  Jailbirds New Orleans              NaN
4      s5  TV Show           Kota Factory              NaN

                                                cast        country  \
0                                                NaN  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                NaN            NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India

           date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min
1  September 24, 2021          2021  TV-MA  2 Seasons
2  September 24, 2021          2021  TV-MA   1 Season
3  September 24, 2021          2021  TV-MA   1 Season
4  September 24, 2021          2021  TV-MA  2 Seasons

                                           listed_in  \
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 ...

                                         description
0  As her father nears the end of his life, filmm...
```

```
1   After crossing paths at a party, a Cape Town t...
2   To protect his family from a powerful drug lor...
3   Feuds, flirtations and toilet talk go down amo...
4   In a city of coaching centers known to train I...
```

- "Title" , "director" & "cast" columns needs to be unnested to make our analyis more accurate.
- Duration columns having data in minutes for movies and in seasons for TV shows

Attributes information:

Show_id: Unique ID for every Movie / Tv Show

Type: Identifier - A Movie or TV Show

Title: Title of the Movie / Tv Show

Director: Director of the Movie

Cast: Actors involved in the movie/show

Country: Country where the movie/show was produced

Date_added: Date it was added on Netflix

Release_year: Actual Release year of the movie/show

Rating: TV Rating of the movie/show

Duration: Total Duration - in minutes or number of seasons

Listed_in italicized text: Genre

Description: The summary description

```
df.shape #checking the count of no. of rows and columns of dataset

(8807, 12)
```

Dataset is having 8807 rows of data with 12 attributes.

```
df.info() #to check the data types of all columns and count of values
in particular column.

<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
```

- We can see that type of rating and date_added columns is "object" which should be categorical and datetime.
- More no. of missing values in cast and director columns.

Statistical summary

```
df.describe()  #to check statistical summary of numerical type data

       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
```

- 25% of the tolal data belongs to year 2019-2021
- 25% of the tolal data belongs to year 1925-2013

Insight -->Netflix should add latest Movies and TV shows to attract more customers.

```
df.describe(include = object) #to check statistical summary of
categorical type data

       show_id   type                   title       director  \
count     8807   8807                    8807           6173   
unique    8807      2                    8807           4528   
top         s1  Movie  Dick Johnson Is Dead  Rajiv Chilaka   
freq         1   6131                       1             19   

                      cast        country       date_added rating 
duration  \
count                 7982           7976             8797   8803 
8804   
```

```
unique                7692            748             1767     17
220
top     David Attenborough  United States  January 1, 2020  TV-MA  1
Season
freq                    19           2818              109   3207
1793

                           listed_in  \
count                           8807
unique                           514
top     Dramas, International Movies
freq                             362

                                              description
count                                                8807
unique                                               8775
top     Paranormal activity at a lush, abandoned prope...
freq                                                    4
```

Conclusion :-

- Show_id and Title are the unique factors.
- "Type" and "rating" column needs to be changed to categorical data.
- "United States" is having the maximun content available.

Missing value detection

```
df.isnull().sum() #checking count of null values per column.
```

```
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
```

- Lot of missing data in director, cast and country columns as compared to others.

```
for col in df:
  null_count = df[col].isnull().sum() / len(df) *100
  print(col , "-->" ,null_count)
```

```
show_id --> 0.0
type --> 0.0
title --> 0.0
director --> 29.908027705234474
cast --> 9.367548540933349
country --> 9.435676166685592
date_added --> 0.11354604292040424
release_year --> 0.0
rating --> 0.04541841716816169
duration --> 0.034063812876121265
listed_in --> 0.0
description --> 0.0
```

As we can see 30% of Director columns value are missing , we cant drop this much data. We will fill these columns with "Unknown"

```
df[["director","cast","country"]] =
df[["director","cast","country"]].fillna("Unknown") #Fillling up the
missing values

df.isnull().sum()

show_id          0
type             0
title            0
director         0
cast             0
country          0
date_added      10
release_year     0
rating           4
duration         3
listed_in        0
description      0
dtype: int64
```

I will drop these rows in which date added values are missing when I will do the analysis related to date added

```
df["rating"].value_counts() #checking unique values in rating columns.

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
```

As I can clearly see that last three values of rating should be in duration columns.

Shifting of data to the right columns

```
df.loc[(df["rating"] == "74 min") | (df["rating"] == "84 min") |
(df["rating"] == "66 min")]
df["duration"][[5541,5794,5813]] = df["rating"][[5541,5794,5813]]
df["rating"][[5541,5794,5813]] = "Nan"

df["rating"].value_counts() #checking the count of each category.
```

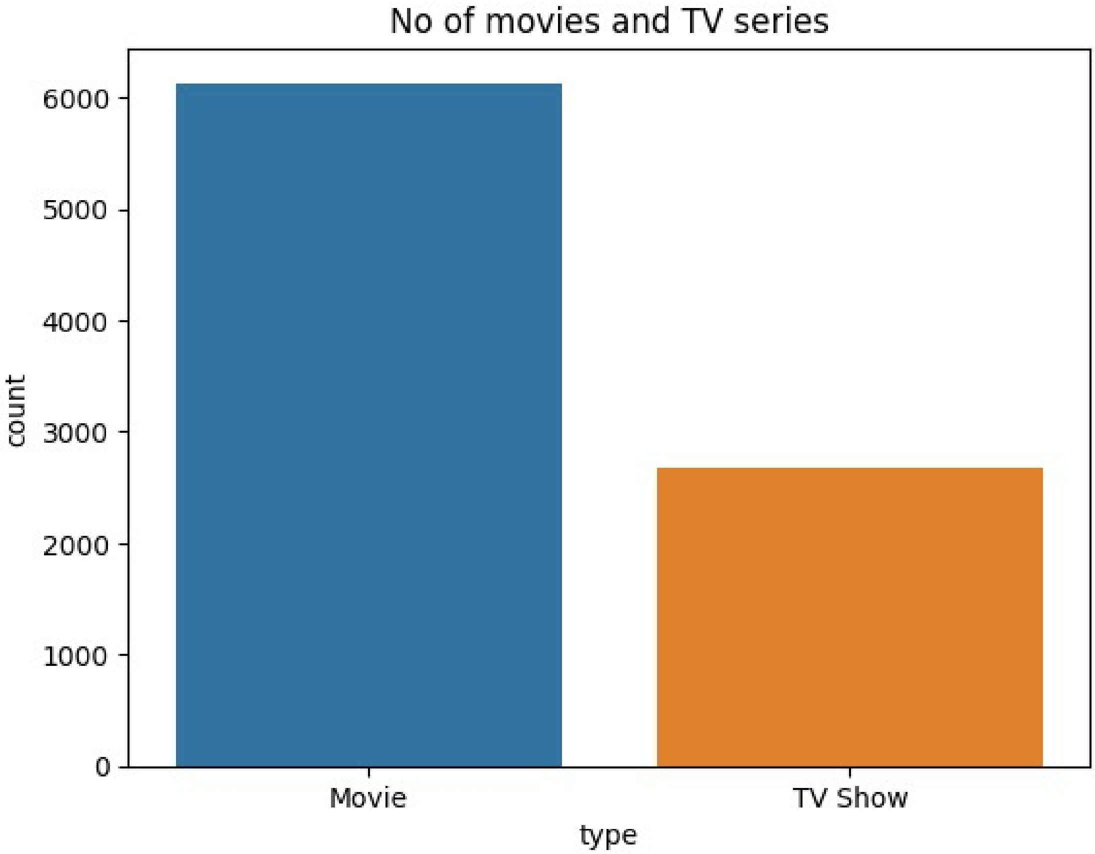
```
TV-MA        3207
TV-14        2160
TV-PG         863
R             799
PG-13         490
TV-Y7         334
TV-Y          307
PG            287
TV-G          220
NR             80
G              41
TV-Y7-FV        6
NC-17           3
Nan             3
UR              3
Name: rating, dtype: int64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

```
#Conversion of categorical attributes to 'category' and 'datetime'
df["date_added"] = pd.to_datetime(df["date_added"])
df =df.astype({"type" : "category", "rating" : "category"})
```

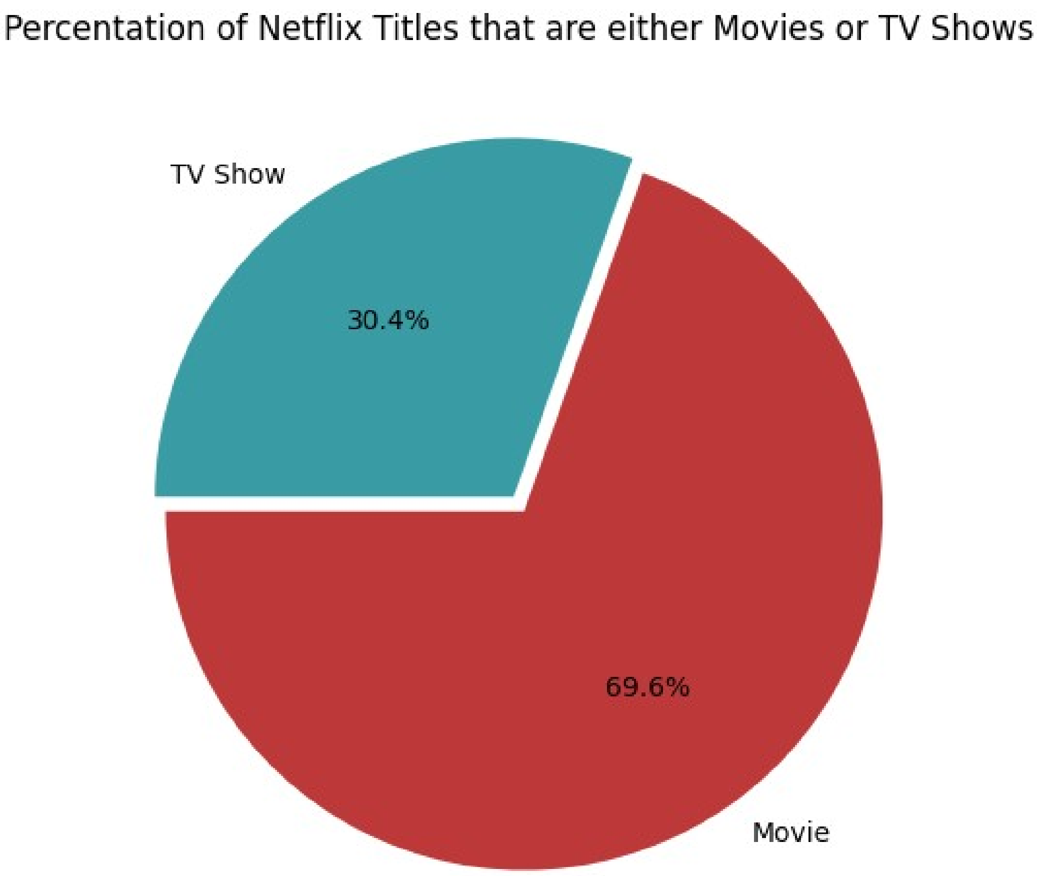
Univariate Analysis

```
df_datetime = df.copy()
df_datetime['Year'] = df.date_added.dt.year  #adding new columns to
the dataframe --> year , month , weekday
df_datetime['month'] = df.date_added.dt.month
df_datetime['day'] = df.date_added.dt.day_name()
```

```
sns.countplot(x = "type" , data = df_datetime) #countplot to count the
no of movies and tv shows available.
plt.title("No of movies and TV series")
plt.show()
```

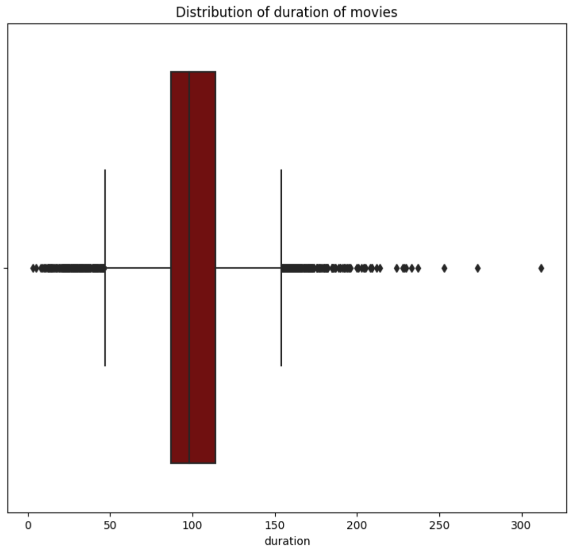


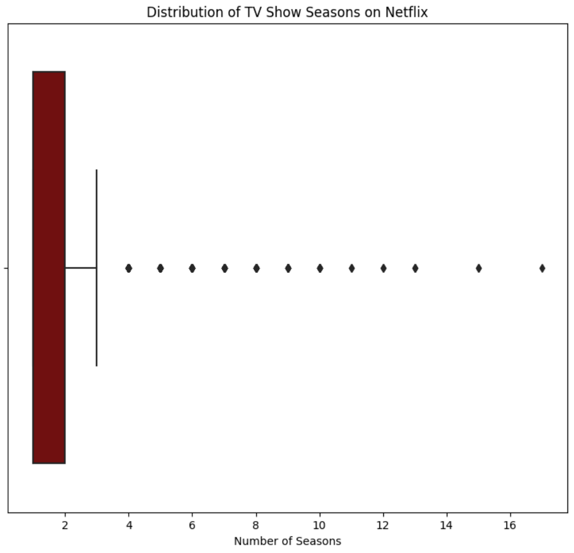
```
plt.figure(figsize=(12,6))
plt.title("Percentation of Netflix Titles that are either Movies or TV
Shows")
g = plt.pie(df_datetime.type.value_counts(),explode=(0.025,0.025),
labels=df_datetime.type.value_counts().index,
colors=['#bd3939','#399ba3'],autopct='%1.1f%%', startangle=180)
plt.show()
```



Immense difference between the count of no of movies and TV show.

```
plt.figure(figsize=(20,8))
duration_df = df.loc[df["duration"].str.contains("min")== True]
["duration"].apply(lambda x: x.split()[0]).astype(int)  # splting the
movies duration as its type is string , extracting the numeri value
and converting it into int type
plt.subplot(1,2,1) #subplots to make the data look easy for
comparison.
sns.boxplot(x=duration_df , color = "maroon")
plt.title("Distribution of duration of movies")
tv_show_df = df[df["duration"].str.contains("Season", na=False)]
seasons_df = tv_show_df["duration"].apply(lambda x: int(x.split()[0]))
plt.subplot(1,2,2)
sns.boxplot(x=seasons_df, color="maroon")
plt.xlabel("Number of Seasons")
plt.title("Distribution of TV Show Seasons on Netflix")
plt.show()
```

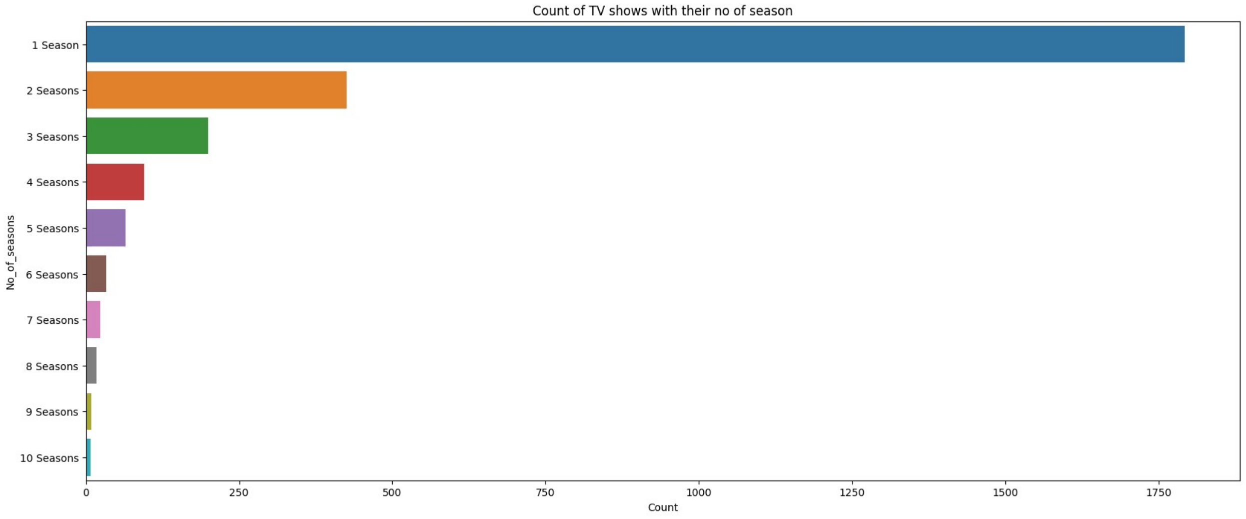




Conclusion -

- Average duration of movies are around 100 min
- TV shows mostly are having 1 or 2 seasons.
- There are lot of outliers present in movies as compare to TV shows

```
df_TV_season = df.loc[df["duration"].str.contains("Season")== True ,
"duration" ].value_counts().reset_index()[:10]  #filtering out top 10
values of TV shows using string.
df_TV_season.rename(columns = {"index" : "No_of_seasons" ,
"duration" : "Count"}, inplace = True) #renaming the columns
plt.figure(figsize=(20,8))
sns.barplot(y = "No_of_seasons" , x = "Count" , data = df_TV_season)
plt.title("Count of TV shows with their no of season")
plt.show()
```



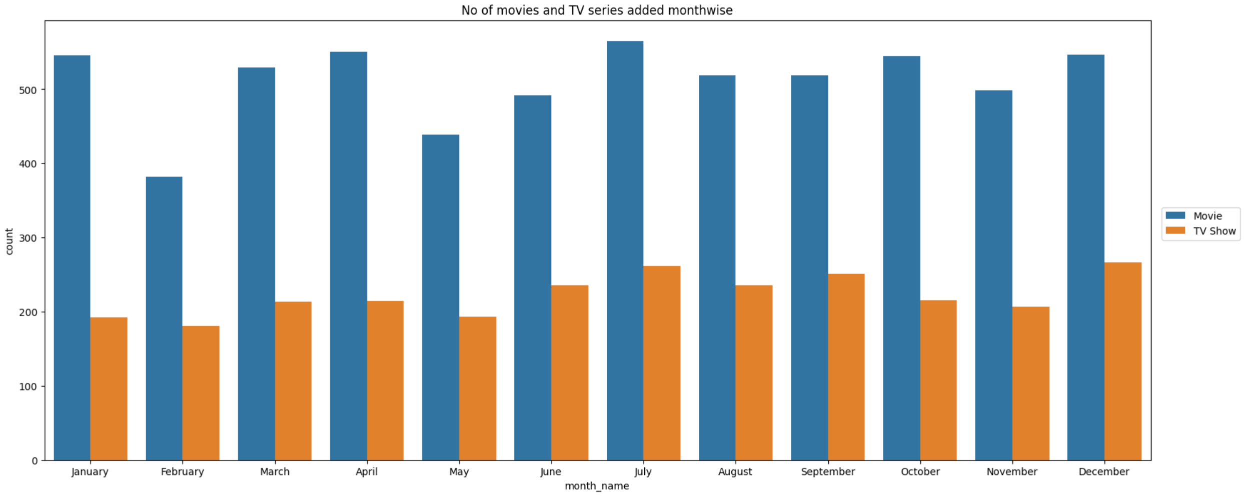
Mostly TV shows have only one season.

Bivariate Analysis

```
df_datetime = pd.DataFrame(df)
df_datetime['Year'] = df.date_added.dt.year
df_datetime['month'] = df.date_added.dt.month
df_datetime['day'] = df.date_added.dt.day_name()
df_datetime_month = df_datetime.sort_values(by ="month")
df_datetime_month['month_name'] = df.date_added.dt.month_name()
```

Analysis of number of content added on Netflix over the period

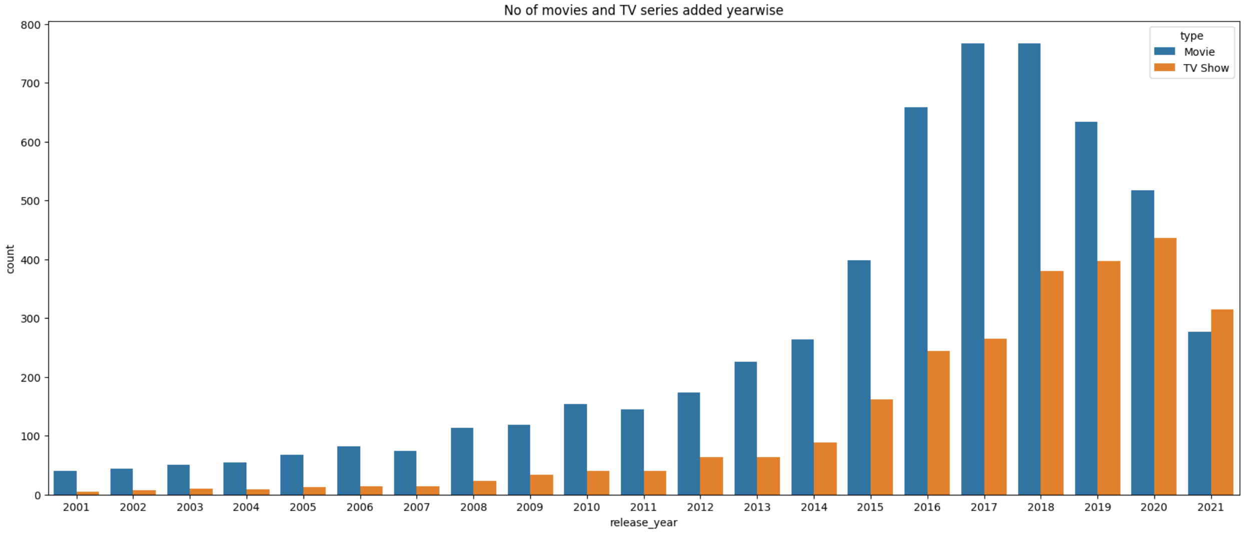
```
plt.figure(figsize=(20,8)) #defining fig size fot the graph image
sns.countplot(x = "month_name" , data = df_datetime_month , hue =
"type")
plt.title("No of movies and TV series added monthwise") #title name of
the plot
plt.legend(loc=(1.01,0.5))
plt.show()
```



Conclusion :-

- July and December are the months when most content was added becasue no of TV shows durind these two months are maximum among all.
- No of movies added per month is greater then no of TV shows added per month.

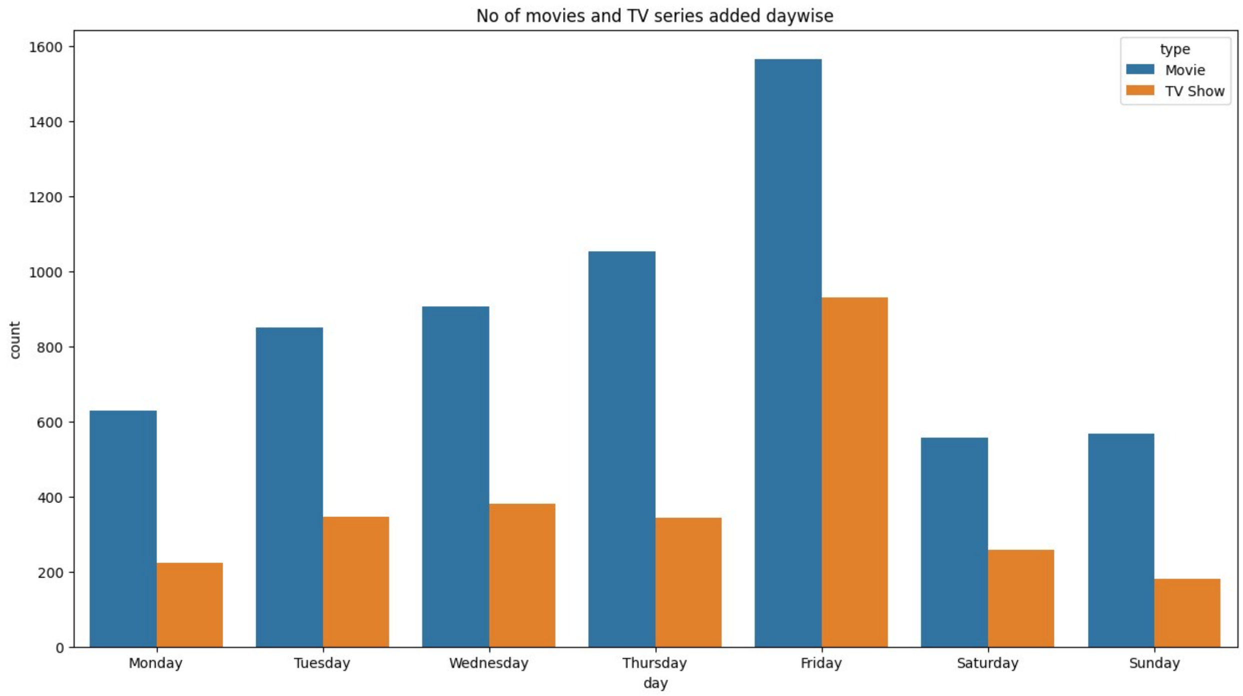
```
plt.figure(figsize=(20,8))
df_year = df.loc[df['release_year']>2000] #used masked to get out data
for movies and TV shows released after 2000
sns.countplot(x='release_year', data = df_year, hue='type')
plt.title("No of movies and TV series added yearwise")
plt.show()
```



Conclusion :-

- In 2020 , maximum no. of TV shows are added followed by 2019 & 2018.
- More no of movies added on Netflix after "2015"
- We can see in 2021 count of movies add drop significanty ,maybe due to COVID pandemic.

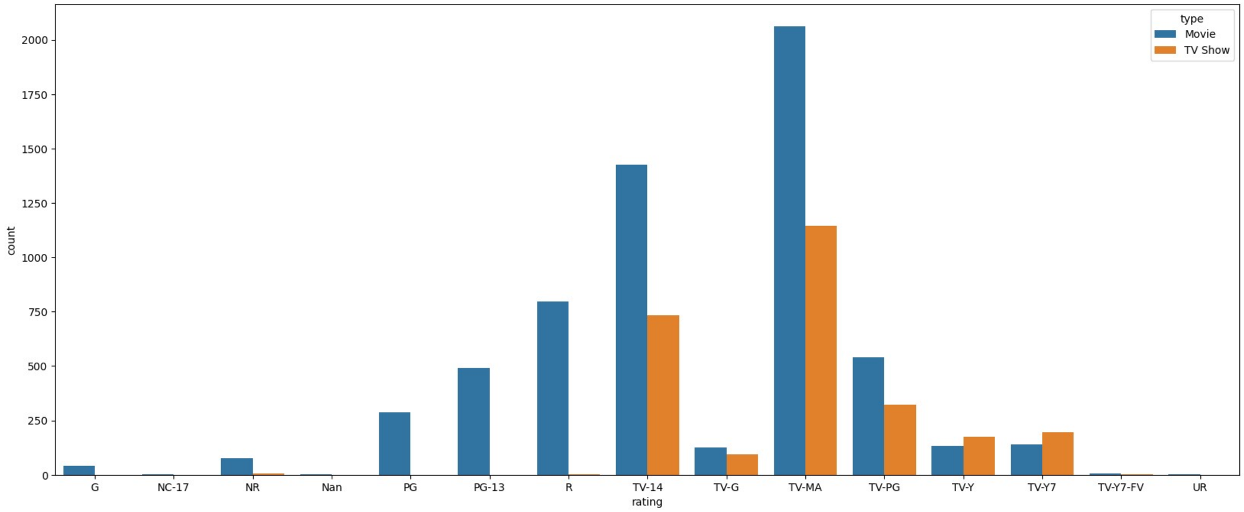
```
plt.figure(figsize=(15,8))
sns.countplot(x = "day" , data = df_datetime , hue = "type" ,
order=["Monday" , "Tuesday" , "Wednesday", "Thursday", "Friday",
"Saturday" ,"Sunday"])
plt.title("No of movies and TV series added daywise")
plt.show()
```



Conclusion :- Most of the content added on netflix on "Friday" followed by Thursday as weekend appraches after these days.

```
print('PG-13 -----> Parental Guidance with Adult Themes[Parental
Guidance]',
'TV-MA -----> Mature Audience[Only for Adults]',
'PG -----> Parental Guidance without Adult Themes[Parental Guidance]',
'TV-14 -----> Contents with Parents strongly cautioned.',
'TV-PG -----> Parental guide suggested[Parental Guidance]',
'TV-Y -----> Children suited content[General Audience & Kids]',
'TV-Y7 -----> Children of age 7 and older[General Audience & Kids]',
'R -----> Strictly for Adults[Only for Adults]',
'TV-G -----> Suitable for all audiences[General Audience & Kids]',
'G -----> General Audience films[General Audience & Kids]',
'NC-17 -----> No one seventeen and under admitted[Only for Adults]',
'NR -----> Not rated movies[Not Rated]',
'TV-Y7-FV -----> Children of age 7 and older with fantasy
violence[General Audience & Kids]',
'UR -----> recut version of rated movie[Not Rated]', sep = '\n')

df_rating = df[df["rating"].isnull()== False]
df_rating.reset_index(inplace = True)
plt.figure(figsize=(20,8))
sns.countplot(x ="rating" , data = df_rating , hue = "type")
plt.show()

PG-13 -----> Parental Guidance with Adult Themes[Parental Guidance]
TV-MA -----> Mature Audience[Only for Adults]
PG -----> Parental Guidance without Adult Themes[Parental Guidance]
```

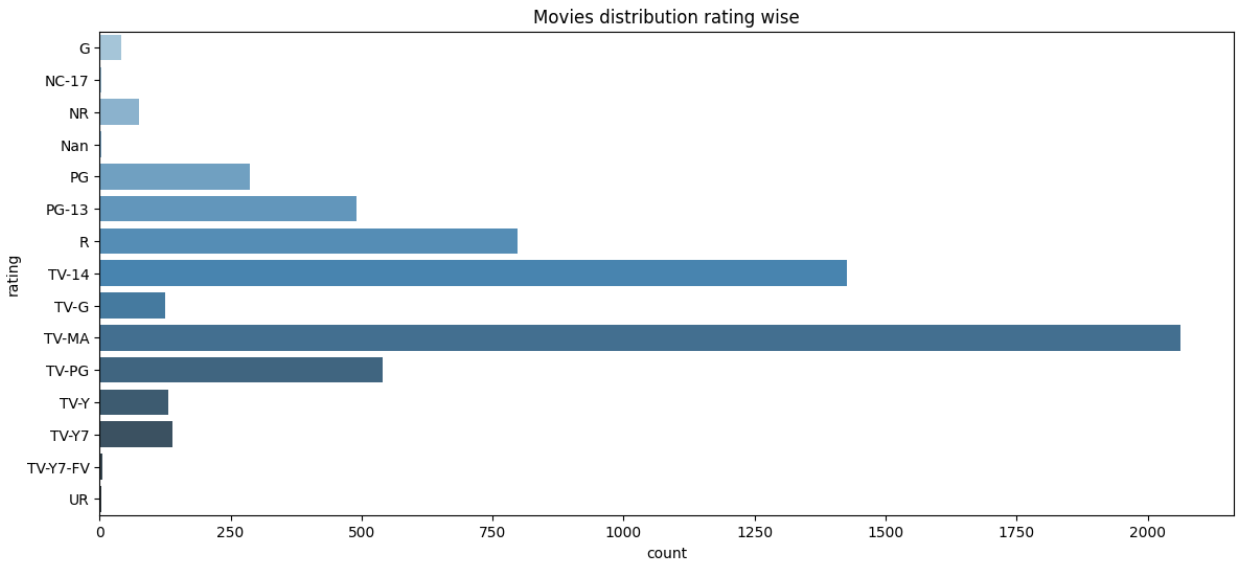
```
TV-14 -----> Contents with Parents strongly cautioned.
TV-PG -----> Parental guide suggested[Parental Guidance]
TV-Y -----> Children suited content[General Audience & Kids]
TV-Y7 -----> Children of age 7 and older[General Audience & Kids]
R -----> Strictly for Adults[Only for Adults]
TV-G -----> Suitable for all audiences[General Audience & Kids]
G -----> General Audience films[General Audience & Kids]
NC-17 -----> No one seventeen and under admitted[Only for Adults]
NR -----> Not rated movies[Not Rated]
TV-Y7-FV -----> Children of age 7 and older with fantasy
violence[General Audience & Kids]
UR -----> recut version of rated movie[Not Rated]
```



Conclusion :-

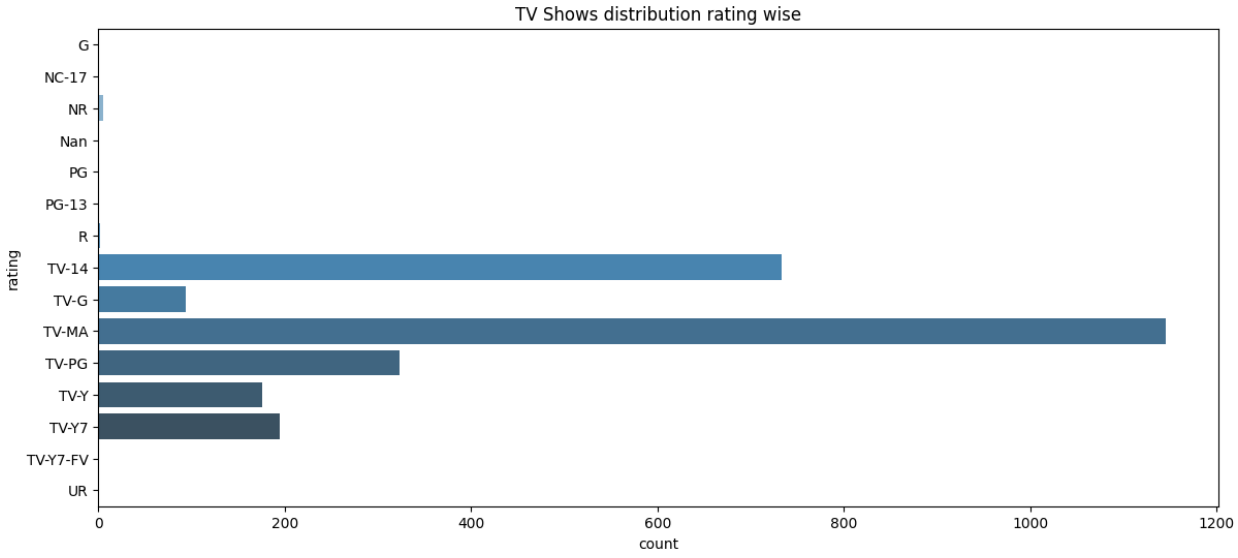
- Mostly TV shows and movies are belongs to TV-MA & TV-14 rating.
- Mostly content available on netflix is for adults and teenagers.

```
plt.figure(figsize=(14,6))
movies_ratingwise = df.loc[df["type"] == "Movie" , ["type" ,
"rating"]]
sns.countplot( y="rating" , data =movies_ratingwise,
palette="Blues_d" )
plt.title("Movies distribution rating wise")
plt.show()
```



Conclusion : Mostly movies are belongs to TV-MA & TV-14 rating.

```
plt.figure(figsize=(14,6))
movies_ratingwise = df.loc[df["type"] == "TV Show" , ["type" ,
"rating"]]
sns.countplot( y="rating" , data =movies_ratingwise,
palette="Blues_d" )
plt.title("TV Shows distribution rating wise")
plt.show()
```



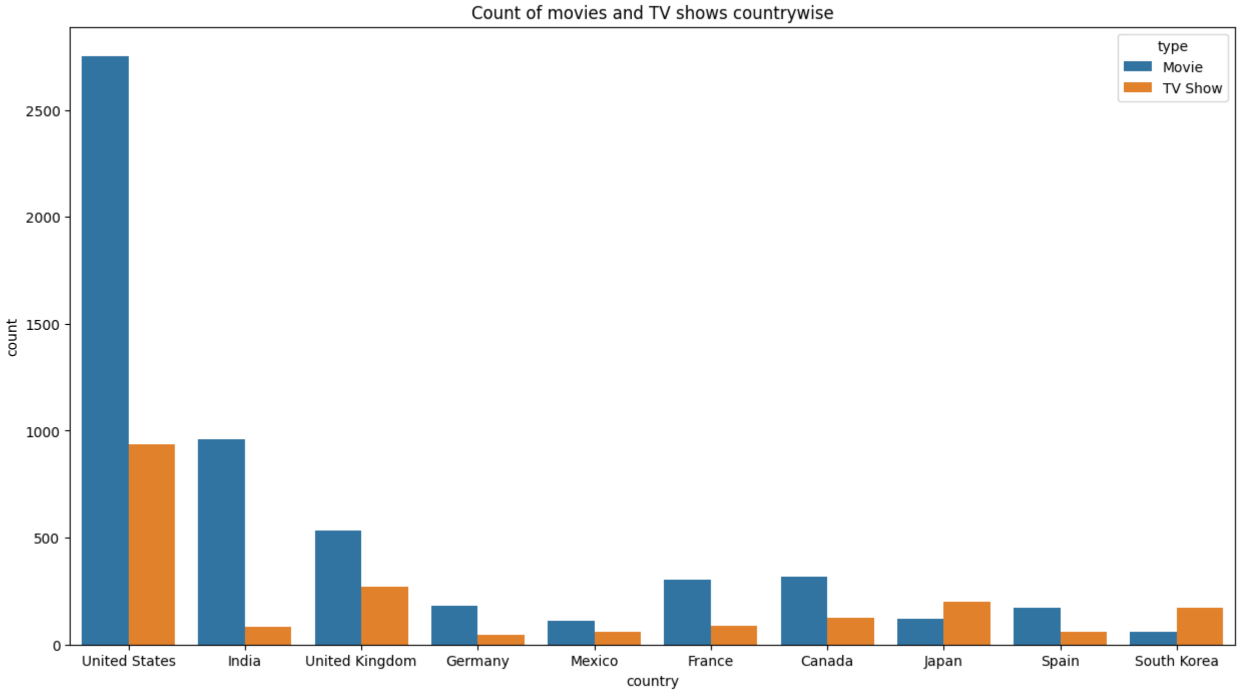
Conclusion :- Mostly TV Shows are belongs to TV-MA & TV-14 rating.

```
director = df["director"].apply(lambda x : str(x).split(",
")).tolist()  #exploding the nested data in directors column.
df_director = pd.DataFrame(director, index = df["title"])
```

```
df_director= df_director.stack()
df_director = df_director.reset_index()
df_director.drop(columns ="level_1" , inplace = True) #droping the
columns
df_director.columns = ["title" , "director"] #renaming the columns
df_fav_director = df.merge(df_director , on = "title" ) #merging of
the dataframes
df_fav_director.head(4)
```

```
  show_id     type                     title       director_x  \
0      s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show           Blood & Water          Unknown
2      s3  TV Show               Ganglands  Julien Leclercq
3      s4  TV Show  Jailbirds New Orleans          Unknown

                                                 cast        country  \
0                                             Unknown  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...        Unknown
3                                             Unknown        Unknown

  date_added  release_year rating   duration  \
0 2021-09-25          2020  PG-13     90 min
1 2021-09-24          2021  TV-MA  2 Seasons
2 2021-09-24          2021  TV-MA   1 Season
3 2021-09-24          2021  TV-MA   1 Season

                                           listed_in  \
0                                      Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act...
3                             Docuseries, Reality TV

                                         description    Year  month
day  \
0  As her father nears the end of his life, filmm...  2021.0    9.0
Saturday
1  After crossing paths at a party, a Cape Town t...  2021.0    9.0
Friday
2  To protect his family from a powerful drug lor...  2021.0    9.0
Friday
3  Feuds, flirtations and toilet talk go down amo...  2021.0    9.0
Friday

        director_y
0  Kirsten Johnson
1          Unknown
2  Julien Leclercq
3          Unknown
```

```python
#exploding country column
country = df["country"].apply(lambda x: str(x).split(", ")).tolist()
#exploding the country column
df_country = pd.DataFrame(country, index = df["title"])
df_country = df_country.stack()
df_country = df_country.reset_index()
df_country.drop(columns = "level_1" , inplace = True)
df_country.columns = ["title" , "country"]

Country_wise_trend = df.merge(df_country , on = "title") #making new dataframe by merfing df_country and original dataframe.
Country_wise_trend.drop(columns = "country_x" , inplace = True)
Country_wise_trend.rename(columns = {"country_y" : "country"}, inplace = True)
Country_wise_trend = Country_wise_trend.loc[Country_wise_trend["country"] != "Unknown"]
top10_country = Country_wise_trend["country"].value_counts().head(10).reset_index()
top10_country.rename(columns = {"index" :"country" , "country" : "count"}, inplace = True)
Country_wise_trend = Country_wise_trend.merge(top10_country, how = "inner" , on = "country")
plt.figure(figsize = (15,8))
sns.countplot(x ="country" , data =Country_wise_trend , hue = "type" )
plt.title("Count of movies and TV shows countrywise")
plt.show()
```



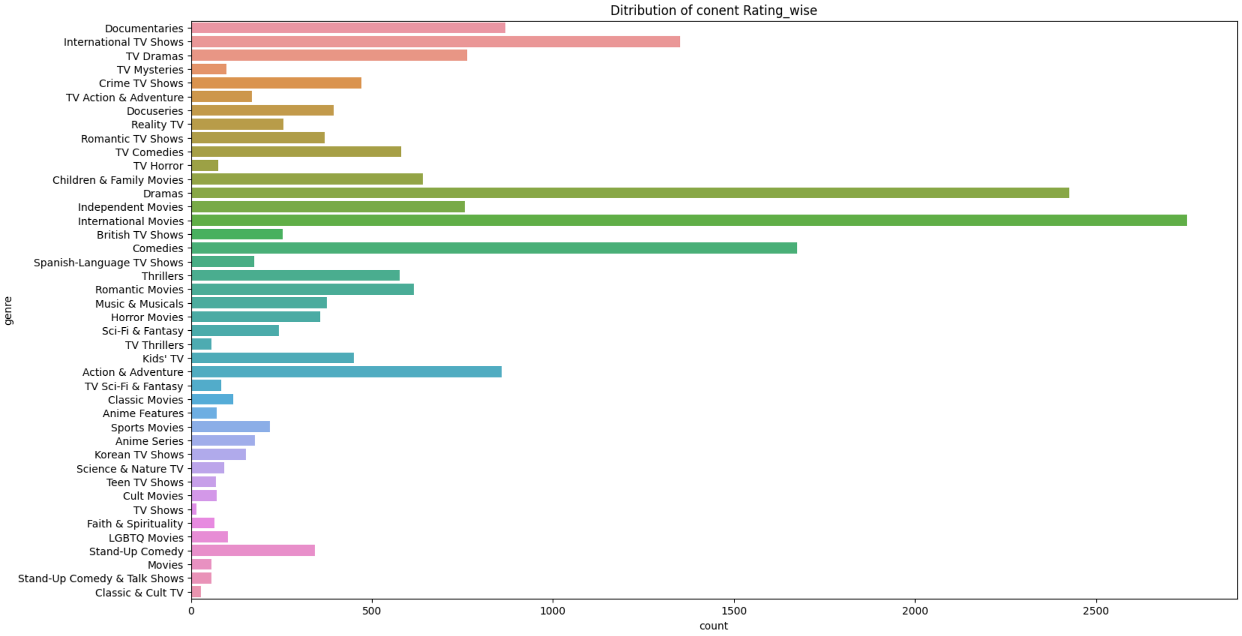
Conclusion :-

- Netflix should target to add more TV shows in Unites states and India as compare to movies.
- Netflix should target to add more movies in Japan and South Korea.

```
#exploding listed_in column
listed_in = df["listed_in"].apply(lambda x: str(x).split(",
")).tolist()
df_genre = pd.DataFrame(listed_in, index = df["title"])
df_genre = df_genre.stack()
df_genre = df_genre.reset_index()
df_genre.drop(columns = "level_1" , inplace = True)
df_genre.columns = ["title" , "genre"]
df_genre.head()
```

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

```
plt.figure(figsize = (18,10))
sns.countplot(y = "genre" , data =df_genre )
plt.title("Ditribution of conent Rating_wise")
plt.show()
```



Most appearing category in netflix movies and TV shows are:-

- International Movies
- Dramas
- Comedies

- International TV show

Non-Graphical Analysis

```
director_countrywise= df_fav_director.merge(df_country , on = "title")
director_countrywise= director_countrywise.drop(columns =
["director_x" , "country_x" ])
director_countrywise.rename(columns = {"director_y": "director" ,
"country_y" : "country"}, inplace = True)
director_countrywise =
director_countrywise.loc[director_countrywise["director"] !=
"Unknown"]
director_countrywise.reset_index(inplace= True)
director_countrywise.head()

   index show_id     type                               title  \
0      0      s1    Movie                Dick Johnson Is Dead
1      2      s3  TV Show                           Ganglands
2      5      s6  TV Show                       Midnight Mass
3      6      s7    Movie  My Little Pony: A New Generation
4      7      s7    Movie  My Little Pony: A New Generation

                                                cast date_added
release_year  \
0                                            Unknown 2021-09-25
2020
1  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi... 2021-09-24
2021
2  Kate Siegel, Zach Gilford, Hamish Linklater, H... 2021-09-24
2021
3  Vanessa Hudgens, Kimiko Glenn, James Marsden, ... 2021-09-24
2021
4  Vanessa Hudgens, Kimiko Glenn, James Marsden, ... 2021-09-24
2021

  rating  duration                                          listed_in
\
0  PG-13    90 min                                      Documentaries

1  TV-MA  1 Season  Crime TV Shows, International TV Shows, TV Act...

2  TV-MA  1 Season                 TV Dramas, TV Horror, TV Mysteries

3     PG    91 min                           Children & Family Movies

4     PG    91 min                           Children & Family Movies


                                         description    Year  month
day  \
0  As her father nears the end of his life, filmm...  2021.0    9.0
```

```
Saturday
1  To protect his family from a powerful drug lor...  2021.0    9.0
Friday
2  The arrival of a charismatic young priest brin...  2021.0    9.0
Friday
3  Equestria's divided. But a bright-eyed hero be...  2021.0    9.0
Friday
4  Equestria's divided. But a bright-eyed hero be...  2021.0    9.0
Friday

          director        country
0  Kirsten Johnson  United States
1  Julien Leclercq        Unknown
2    Mike Flanagan        Unknown
3    Robert Cullen        Unknown
4   José Luis Ucha        Unknown
```

```
country = director_countrywise['country'].value_counts()
[:6].index.tolist()
print(' Top 2 Directors of Top 5 Countries')
print('\n')
for val in country:
  if val != 'Unknown':
    print(f'**{val}**')

print(director_countrywise.loc[director_countrywise['country']==val,
'director'].value_counts()[:2])
    print('\n')
```

```
 Top 2 Directors of Top 5 Countries


**United States**
Jay Karas       15
Marcus Raboy    15
Name: director, dtype: int64


**India**
Anurag Kashyap    9
David Dhawan      9
Name: director, dtype: int64


**United Kingdom**
Alastair Fothergill    4
Edward Cotterill       4
Name: director, dtype: int64


**Canada**
```

```
Justin G. Dyck       8
Mike Clattenburg     5
Name: director, dtype: int64


**France**
Thierry Donard      5
Youssef Chahine     4
Name: director, dtype: int64
```

Conclusion :

- Anurag Kashyap and David Dhawan are the most famous directors for Inida.
- Jay Karas and Marcus Raboy are the most famous directors in United States.

```
director_countrywise["director"].value_counts().head(3)

Rajiv Chilaka    22
Jan Suter        21
Raúl Campos      19
Name: director, dtype: int64
```

Conclusion : "Rajiv Chilaka" is the most famous director among all followed by Jan Suter

```
#exploding cast column
cast = df["cast"].apply(lambda x : str(x).split(", ")).tolist()
df_cast = pd.DataFrame(cast,  index = df["title"])
df_cast = df_cast.stack()
df_cast = df_cast.reset_index()
df_cast.drop(columns = "level_1" , inplace = True)
df_cast.columns = ["title" , "cast"]
df_fav_cast = df.merge(df_cast , on = "title" )

cast_countrywise= df_fav_cast.merge(df_country , on = "title")
cast_countrywise= cast_countrywise.drop(columns = ["cast_x" ,
"country_x"])
cast_countrywise = cast_countrywise.rename(columns = {"cast_y" :
"cast" , "country_y" : "country"})
cast_countrywise = cast_countrywise.loc[cast_countrywise["cast"] !=
"Unknown"].reset_index() #making new dataframe by dropping all rows
whose cast is unknown and then resetting the index..00
cast_countrywise.head()

   index show_id     type          title director date_added
release_year  \
0      1      s2  TV Show  Blood & Water  Unknown 2021-09-24
2021
1      2      s2  TV Show  Blood & Water  Unknown 2021-09-24
```

```
2021
2      3       s2  TV Show  Blood & Water  Unknown 2021-09-24
2021
3      4       s2  TV Show  Blood & Water  Unknown 2021-09-24
2021
4      5       s2  TV Show  Blood & Water  Unknown 2021-09-24
2021

  rating   duration
listed_in  \
0  TV-MA  2 Seasons  International TV Shows, TV Dramas, TV Mysteries

1  TV-MA  2 Seasons  International TV Shows, TV Dramas, TV Mysteries

2  TV-MA  2 Seasons  International TV Shows, TV Dramas, TV Mysteries

3  TV-MA  2 Seasons  International TV Shows, TV Dramas, TV Mysteries

4  TV-MA  2 Seasons  International TV Shows, TV Dramas, TV Mysteries


                                         description    Year  month
day  \
0  After crossing paths at a party, a Cape Town t...  2021.0    9.0
Friday
1  After crossing paths at a party, a Cape Town t...  2021.0    9.0
Friday
2  After crossing paths at a party, a Cape Town t...  2021.0    9.0
Friday
3  After crossing paths at a party, a Cape Town t...  2021.0    9.0
Friday
4  After crossing paths at a party, a Cape Town t...  2021.0    9.0
Friday

               cast      country
0        Ama Qamata  South Africa
1       Khosi Ngema  South Africa
2     Gail Mabalane  South Africa
3    Thabang Molaba  South Africa
4  Dillon Windvogel  South Africa

country_actor = cast_countrywise['country'].value_counts()
[:6].index.tolist()
print(' Top 2 Actors of Top 5 Countries')
print('\n')
for val in country:
  if val != 'Unknown':
    print(f'--{val}--')
    print(cast_countrywise.loc[cast_countrywise['country']==val,
'cast'].value_counts()[:2])
    print('\n')
```

```
 Top 2 Actors of Top 5 Countries


--United States--
Tara Strong          22
Samuel L. Jackson    22
Name: cast, dtype: int64


--India--
Anupam Kher       40
Shah Rukh Khan    34
Name: cast, dtype: int64


--United Kingdom--
David Attenborough    17
John Cleese           16
Name: cast, dtype: int64


--Canada--
John Paul Tremblay    14
Robb Wells            14
Name: cast, dtype: int64


--France--
Wille Lindberg    5
Benoît Magimel    5
Name: cast, dtype: int64
```

Conclusion :-

- These are the top two cast of these countires.
- Netflix has added more content for India in which cast are- Anupam Kher or Shah Rukh Khan.

```
cast_countrywise["cast"].value_counts().head(5) #value_counts of the
cast columns to get the most famous actors
```

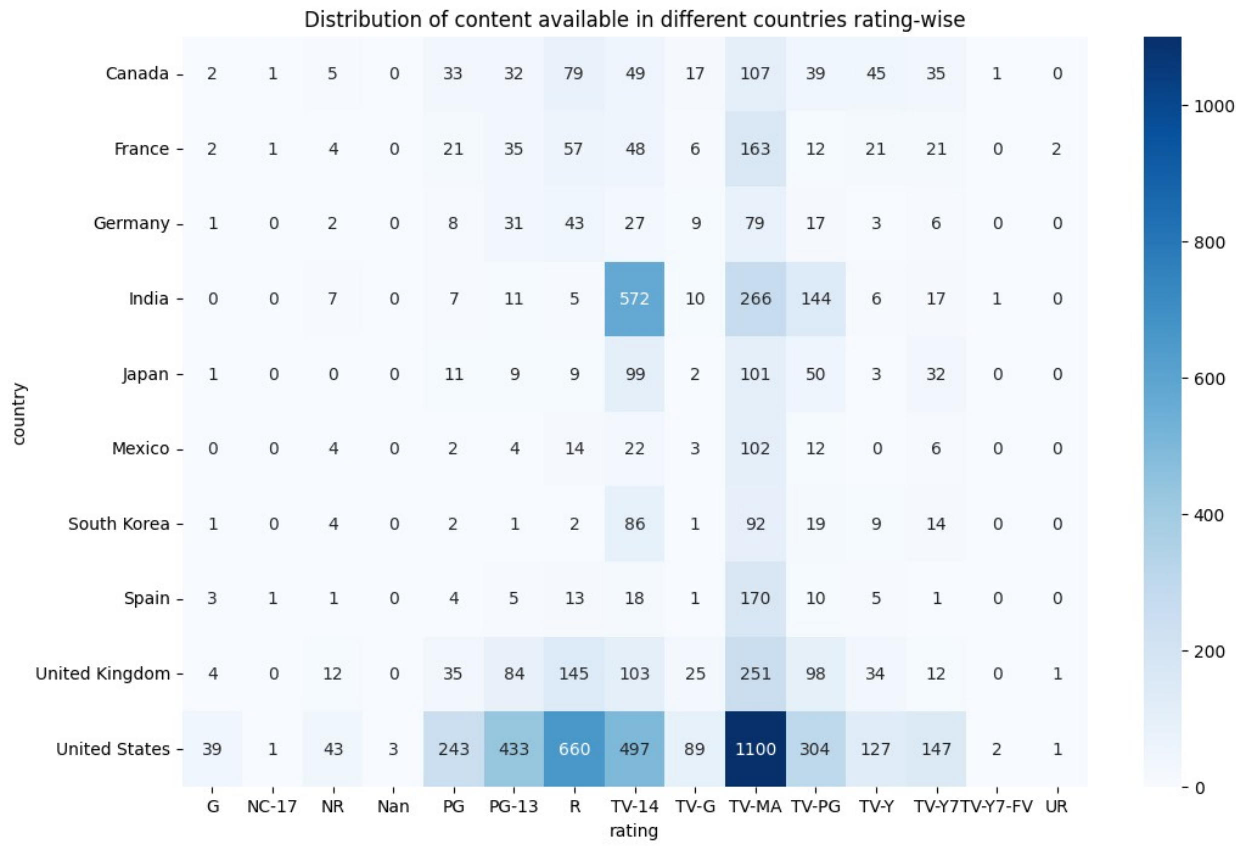
```
Anupam Kher           46
David Attenborough    45
Vincent Tong          42
John Cleese           40
```

```
Tara Strong           39
Name: cast, dtype: int64
```

These are the top five actors and most famous actor.

Heatmap

```
df_trend_country = df.merge(df_country , on = "title")
df_trend_country.drop(columns = "country_x" , inplace = True)
df_trend_country.rename(columns = {"country_y":"country"}, inplace =
True)

temp = df_trend_country['country'].value_counts()[:11].reset_index()
temp.rename(columns = {'index':'country', 'country':'count'},
inplace=True)
country_list = temp['country'].tolist()
df_top10country =
df_trend_country.loc[df_trend_country['country'].isin(country_list)]
df_top10country = df_top10country.loc[df_top10country["country"]!
="Unknown"] #dropping of rows whose value is unknown.

# Group the data by "country" and "rating" and count the occurrences
of each rating in each country
heat_rating = df_top10country.groupby(["country",
"rating"]).size().reset_index(name='count')

# Create a pivot table to get "rating" as columns and "count" as
values for each "country"
heat_rating = heat_rating.pivot_table(index="country",
columns="rating", values="count", fill_value=0)

# Create the heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(heat_rating, annot=True, cmap="Blues", fmt="d")
plt.title("Distribution of content available in different countries
rating-wise")
plt.show()
```



Conclusion :-

- Top 10 countries are having most content that belongs to TV-MA (Adults Category)
- India and United States are having large content in TV-14 category.
- United Kingdom and United States are having large content in R category.

```
genre_country_df= df_trend_country.merge(df_genre , on= "title")
genre_country_df.head(5)

  show_id     type                    title         director  \
0      s1    Movie  Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show          Blood & Water          Unknown
2      s2  TV Show          Blood & Water          Unknown
3      s2  TV Show          Blood & Water          Unknown
4      s3  TV Show              Ganglands  Julien Leclercq

                                                cast date_added
release_year  \
0                                            Unknown 2021-09-25
2020
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban... 2021-09-24
2021
2  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban... 2021-09-24
2021
3  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban... 2021-09-24
```

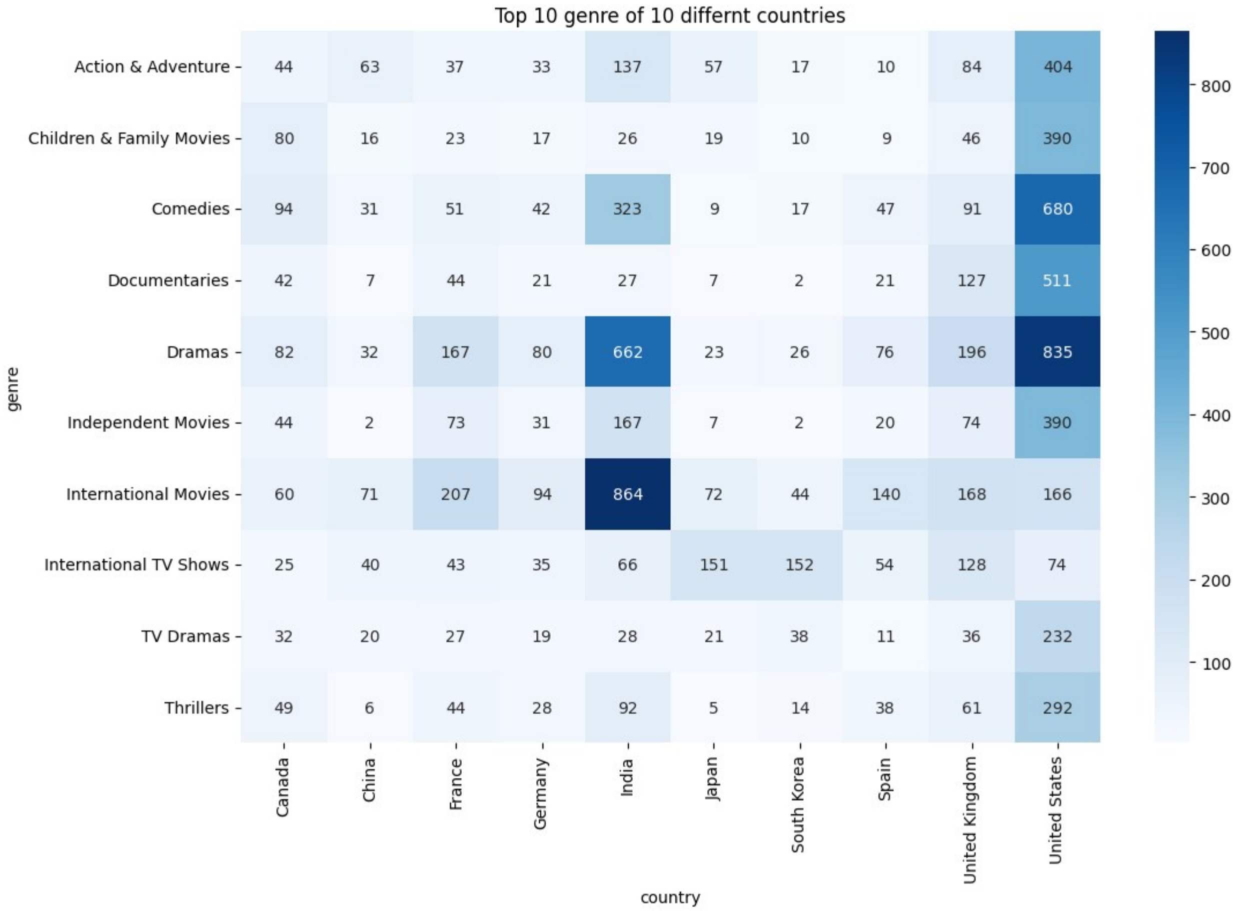
```
2021
4  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi... 2021-09-24
2021

  rating   duration                                          listed_in
\
0  PG-13     90 min                                      Documentaries

1  TV-MA  2 Seasons    International TV Shows, TV Dramas, TV Mysteries

2  TV-MA  2 Seasons    International TV Shows, TV Dramas, TV Mysteries

3  TV-MA  2 Seasons    International TV Shows, TV Dramas, TV Mysteries

4  TV-MA   1 Season  Crime TV Shows, International TV Shows, TV Act...


                                         description    Year  month
day  \
0  As her father nears the end of his life, filmm...  2021.0    9.0
Saturday
1  After crossing paths at a party, a Cape Town t...  2021.0    9.0
Friday
2  After crossing paths at a party, a Cape Town t...  2021.0    9.0
Friday
3  After crossing paths at a party, a Cape Town t...  2021.0    9.0
Friday
4  To protect his family from a powerful drug lor...  2021.0    9.0
Friday

         country                   genre
0  United States           Documentaries
1   South Africa  International TV Shows
2   South Africa               TV Dramas
3   South Africa            TV Mysteries
4        Unknown          Crime TV Shows
```

```
temp_genre = genre_country_df['genre'].value_counts()
[:10].reset_index()
temp_genre.rename(columns = {'index':'genre', 'genre':'count'},
inplace=True)
genre_list = temp_genre['genre'].tolist()
df_top10_genre =
genre_country_df.loc[genre_country_df['genre'].isin(genre_list)]
df_top10_genre.head()
```

```
  show_id     type                 title         director  \
0      s1    Movie  Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show         Blood & Water          Unknown
2      s2  TV Show         Blood & Water          Unknown
```

```
5      s3  TV Show           Ganglands  Julien Leclercq
9      s5  TV Show        Kota Factory          Unknown

                                                cast date_added
release_year  \
0                                            Unknown 2021-09-25
2020
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban... 2021-09-24
2021
2  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban... 2021-09-24
2021
5  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi... 2021-09-24
2021
9  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K... 2021-09-24
2021

  rating   duration                                          listed_in
\
0  PG-13     90 min                                      Documentaries

1  TV-MA  2 Seasons    International TV Shows, TV Dramas, TV Mysteries

2  TV-MA  2 Seasons    International TV Shows, TV Dramas, TV Mysteries

5  TV-MA   1 Season  Crime TV Shows, International TV Shows, TV Act...

9  TV-MA  2 Seasons  International TV Shows, Romantic TV Shows, TV ...


                                         description    Year  month
day  \
0  As her father nears the end of his life, filmm...  2021.0    9.0
Saturday
1  After crossing paths at a party, a Cape Town t...  2021.0    9.0
Friday
2  After crossing paths at a party, a Cape Town t...  2021.0    9.0
Friday
5  To protect his family from a powerful drug lor...  2021.0    9.0
Friday
9  In a city of coaching centers known to train I...  2021.0    9.0
Friday

         country                   genre
0  United States           Documentaries
1   South Africa  International TV Shows
2   South Africa               TV Dramas
5        Unknown  International TV Shows
9          India  International TV Shows
```

```
df_top10_genre = df_top10_genre.loc[df_top10_genre["country"] !=
"Unknown"]
df_top10_genre["country"].value_counts()[:10]

temp_c = df_top10_genre["country"].value_counts()[:10].reset_index()
temp_c.rename(columns = {'index':'country', 'country':'count'},
inplace=True)
country_list = temp_c["country"].tolist()
df_top10_genre_countrywise =
df_top10_genre.loc[df_top10_genre['country'].isin(country_list)]
df_top10_genre_countrywise.head()

heat_genre= pd.DataFrame(df_top10_genre_countrywise.groupby("genre")
["country"].value_counts())
heat_genre.rename(columns = {"country" : "count"}, inplace = True)
heat_genre.reset_index(inplace = True)
heat_genre_final = heat_genre.pivot("genre" , "country" , "count")
plt.figure(figsize = (12,8))
sns.heatmap(heat_genre_final , annot = True,  cmap="Blues", fmt = "d")
plt.title("Top 10 genre of 10 differnt countries")
plt.show()
```



Conclusion :-

- For India, netflix should add more content of genre International movies, Comedies and Dramas.
- For United States , Netflix should add more content of genre Dramas and Comedy.
- For Canada, Netflix should add more content of genre Comedies, Dramas and Children & family movies.

Summary :-

- Netflix added more movies as compare to TV shows
- Content for United States on netflix is maximum as compare to other countries.
- Netflix content is mostly availabe for adults only
- Most popular genres in recent years are International movies, Dramas, Comedies, International TV Shows and Action & Adventure.
- In 2021 , there is significant amount of drop in content added due to COVID pandemic.

*Most of viewers of Netflix is from United States followed by India & United Kingdom

Movies:-

- In United States , India and United kingdom movies are more popular as comapre to other countires.
- Almost same no. of movies are added on netflix every month.
- Mostly movies are of "100 min" duration.
- Top people casted in Movies are from India.
- "Rajiv Chilakaa" is the most famous director among all.

TV Shows :-

- TV Shows mostly are having season 1 and season 2 respectively.
- For Japan and South Korea, netflix should focus more on TV showes as compare to movies

Recommendations :

Movies :-

- Preferd movies duration is between 90-100 minutes.
- Netflix should add more movies for United States and India falling in category of Internation movies and comedies.
- Netflix should add more movies for United States and India having rating of TV-MA & TV-14.
- Top three countries where movies added are United States, India & United Kingdom.
- Netflix shoud add movies on Friday than any other weekday.

TV Show:-

- Preferd TV Show duration is 1-2 seeasons.
- Netflix should focus on countries like Japan, South Korea and France in TV shows , as they prefer TV shows over movies.

- Netflix shoud add TV Show on Friday than other weekday.
- As per 2021 data, count of TV showes are more than movies , this means people wants more web-series as they have for leisure time may be due to work from home scenario.