

✓ Netflix Data Exploration Business Case

- Name - Sanjesh Chourasia
- Submission date - 28/09/2024

1. Project Overview

Netflix, a global leader in video streaming, offers over 8,000 movies and TV shows to more than 200 million subscribers worldwide. This project aims to analyze data from 8,807 Netflix titles to uncover insights that can guide decisions about content production, release timing, and business growth strategies. Through this analysis, we will answer key questions about popular content types, release patterns, and country-specific production trends.

Dataset

[Netflix Dataset](#)

The dataset consists of movies and TV shows available on Netflix. Each entry includes details such as:

- Show ID: Unique ID of the show
- Type: Movie or TV show
- Title: Title of the show
- Director: Director of the show
- Cast: Actors involved
- Country: Country of production
- Date Added: When it was added to Netflix
- Release Year: Year of release
- Rating: TV rating (e.g., PG, TV-MA)
- Duration: Number of seasons or runtime in minutes
- Listed In: Genre
- Description: A brief summary of the show

This dataset provides a broad range of information to analyze, from content type and release dates to cast and country-specific trends.

✓ 2. Importing the libraries and the dataset

In this project, I am using **NumPy** for working with arrays, **Pandas** for handling and analyzing data with DataFrames, and both **Matplotlib** and **Seaborn** for creating visualizations and charts."

```
1 # Importing libraries.
2 import numpy as np
3 import pandas as pd
4 import matplotlib.pyplot as plt
5 import seaborn as sns
6
7 # reading csv file from google drive.
8 from google.colab import drive
9 drive.mount('/content/drive')
```

↳ Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

```
1 # Loading data.
2 df = pd.read_csv('/content/drive/My Drive/python data analyst project scaler/netflix.csv')
3 df.head(2)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documenta
					Ama Qamata,						Internati

Next steps:

Generate code with df

View recommended plots

New interactive sheet

```
1 df.shape
```

```
(8807, 12)
```

Dataset contain 8807 rows and 12 columns.

```
1 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
```

```
1 df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
release_year	8807.0	2014.180198	8.819312	1925.0	2013.0	2017.0	2019.0	2021.0

```
1 # list of columns.
2 df.columns
```

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

```
1 # columns with datatype int.
2 numerical_column_name = df.select_dtypes(include=['number']).columns
3 print('Numerical Columns:', numerical_column_name)
```

```
Numerical Columns: Index(['release_year'], dtype='object')
```

```
1 # column with datatype object.
2 object_column_name = df.select_dtypes(include=['object']).columns
3 print('Object Columns:', object_column_name)
```

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

```
1 #unique value
2 df.nunique()
```




	0
show_id	8807
type	2
title	8807
director	4528
cast	7692
country	748
date_added	1767
release_year	74
rating	17
duration	220
listed_in	514
description	8775

dtype: int64

3. Data Cleaning

```
1 # Creating copy fo df.
2 df_c1 = df.copy()
```

```
1 # null values (before cleaning data)
2 null = df_c1.isnull().sum()
3
4 # percentage null value
5 null_ = round((null/len(df_c1))*100,2)
6
7 # concating null and null_ columns.
8 con = pd.concat([null,null_],axis=1,keys=['null_values','per_null'])
9 con
10
```

	null_values	per_null	
show_id	0	0.00	  
type	0	0.00	
title	0	0.00	
director	2634	29.91	
cast	825	9.37	
country	831	9.44	
date_added	10	0.11	
release_year	0	0.00	
rating	4	0.05	
duration	3	0.03	
listed_in	0	0.00	
description	0	0.00	

Next steps:

[Generate code with con](#)[View recommended plots](#)[New interactive sheet](#)

In our dataset, the director, cast, and country columns contain a majority of the null values.

```
1 # Replacing null values with 'Unknown'
2 df_c1['director'].fillna('Unknown_director', inplace=True)
3 df_c1['cast'].fillna('Unknown_cast',inplace=True)
4 df_c1['country'].fillna('Unknown_country', inplace=True)

1 # Replacing null values with 'Unknown' and changing datatype of date column.
2 df_c1['date_added'].fillna('Unknown_date_added', inplace=True)
3 df_c1['date_added'] = pd.to_datetime(df_c1['date_added'], errors='coerce')

1 # Replacing null values in rating column with mode value.
2 df_c1['rating'].fillna(df['rating'].mode(), inplace=True)

1 # splitting and changing datatype of duration.
2 df_c1[['duration', 'duration_type']] = df_c1['duration'].str.split(' ', expand=True)
3 df_c1['duration'].fillna(0, inplace=True)
4 df_c1['duration_type'].fillna('Unknown_duration_type', inplace=True)

1 df_c1.head(2)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown_cast	United States	2021-09-25	2020	PG-13	90
1	s2	TV Show	Blood & Water	Unknown_director	Ama Qamata, Khosi Ngema, Gail Mabalane,	South Africa	2021-09-24	2021	TV-MA	2

Next steps: [Generate code with df_c1](#) [View recommended plots](#) [New interactive sheet](#)

4. Analysing data

1. Analysing each categorical variable both using graphical and nongraphical analysis.

1.1 Number of sessions in each Tv shows

```
1 # filtering tv show data.
2 tv = df_c1[df_c1['type'] == 'TV Show']
3
4 # grouping and calculating count of duration.
5 tv_g = tv.groupby(['duration'])['title'].count().sort_values(ascending=False).reset_index()
6 tv_g.head()
```

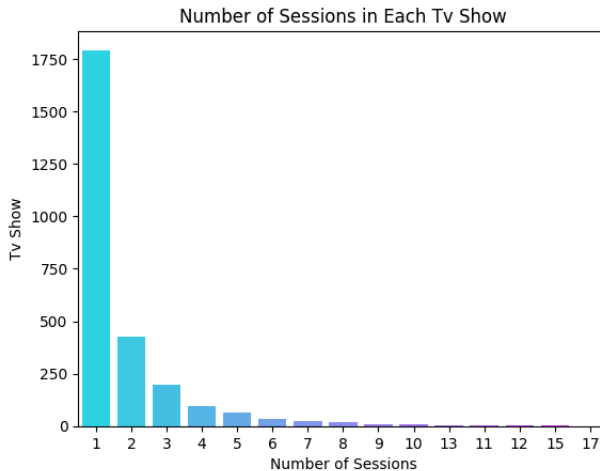
	duration	title
0	1	1793
1	2	425
2	3	199
3	4	95
4	5	65

Next steps: [Generate code with tv_g](#) [View recommended plots](#) [New interactive sheet](#)

```

1 sns.barplot(x='duration', y='title', data=tv_g, palette='cool', hue='duration', legend=False)
2 plt.xlabel('Number of Sessions')
3 plt.ylabel('Tv Show')
4 plt.title('Number of Sessions in Each Tv Show')
5 plt.show()

```



Most TV Shows has only one session.

1.2 Rating.

```

1 rating = df_c1.groupby(['rating'])['title'].count().sort_values(ascending=False).reset_index().head(10)
2 rating

```



	rating	title	
0	TV-MA	3207	
1	TV-14	2160	
2	TV-PG	863	
3	R	799	
4	PG-13	490	
5	TV-Y7	334	
6	TV-Y	307	
7	PG	287	
8	TV-G	220	
9	NR	80	

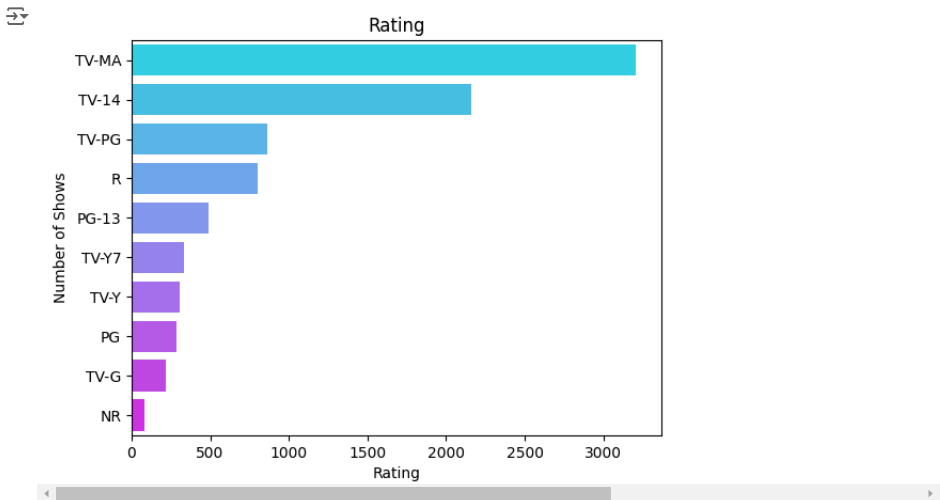
Next steps:

[Generate code with rating](#)
[View recommended plots](#)
[New interactive sheet](#)

```

1 sns.barplot(y='rating', x='title', data=rating, palette='cool', hue='rating', legend=False)
2 plt.xlabel('Rating')
3 plt.ylabel('Number of Shows')
4 plt.title('Rating')
5 plt.show()

```



The majority of shows on Netflix are produced for an adult audience, with a TV-MA rating.

1.3. Length of movies.

```
1 movies = df_c1[df_c1['type'] == 'Movie']
```

```
1 # converting datatype into int
2 movies['duration'] = movies['duration'].astype(int)
```

<ipython-input-57-2821d4f4f440>: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

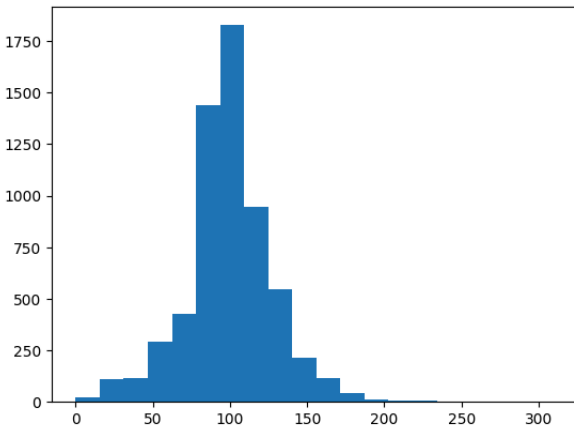
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
movies['duration'] = movies['duration'].astype(int)

```
1 movies.describe()
```

	date_added	release_year	duration	
count	6131	6131.000000	6131.000000	
mean	2019-05-07 03:32:47.639863040	2013.121514	99.528462	
min	2008-01-01 00:00:00	1942.000000	0.000000	
25%	2018-04-01 00:00:00	2012.000000	87.000000	
50%	2019-06-19 00:00:00	2016.000000	98.000000	
75%	2020-07-23 12:00:00	2018.000000	114.000000	
max	2021-09-25 00:00:00	2021.000000	312.000000	
std	NaN	9.678169	28.369284	

```
1 plt.hist(movies['duration'], bins=20)
```

```
array([2.200e+01, 1.130e+02, 1.170e+02, 2.930e+02, 4.280e+02, 1.439e+03,
       1.827e+03, 9.450e+02, 5.440e+02, 2.120e+02, 1.170e+02, 4.300e+01,
       1.300e+01, 9.000e+00, 5.000e+00, 1.000e+00, 1.000e+00, 1.000e+00,
       0.000e+00, 1.000e+00]),
array([ 0. , 15.6, 31.2, 46.8, 62.4, 78. , 93.6, 109.2, 124.8,
       140.4, 156. , 171.6, 187.2, 202.8, 218.4, 234. , 249.6, 265.2,
       280.8, 296.4, 312. ]),
<BarContainer object of 20 artists>)
```



Average duration of most movies on Netflix varies from 90 min to 120 min.

2. Comparison of TV Shows and Movies.

```
1 type1 = df_c1['type'].value_counts()
2 type1.reset_index()
```

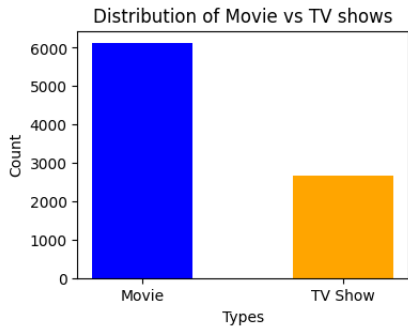
	type	count
0	Movie	6131
1	TV Show	2676

There are total 6131 movies and 2676 TV Shows in Netflix.

```
1 # calculating number of movies and tv shows
2 type1 = df_c1['type'].value_counts()
3 print('Distribution of Movies and Tv Shows',type1)
4
5 # plotting graph
6 plt.figure(figsize=(4,3))
7 plt.bar(type1.index, type1.values , color= ['blue','orange'], width = 0.5, align = 'center')
8 plt.xlabel('Types')
9 plt.ylabel('Count')
10 plt.title('Distribution of Movie vs TV shows')
11 plt.show()
```

↕ Distribution of Movies and Tv Shows type

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

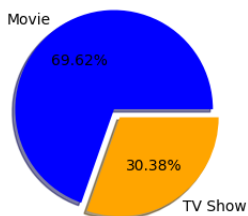


Data is distributed among Movies and Tv shows. There are total 6131 Movies and 2676 TV Shows.

```
1 plt.figure(figsize=(4,3))
2 plt.pie(data=type, x=type1.values, labels=type1.index, colors= ['blue','orange'], autopct='%0.2f%%',explode = [
3 plt.title('Distribution of Movie vs TV shows')
4 plt.show()
5
```



Distribution of Movie vs TV shows



Distribution of Movies is 69.62% and TV Shows is 30.38%.

2.1. Finding the number of movies produced in each country and pick the top 10 countries.

```
1 # selecting movies data only.
2 movies = df_c1[(df_c1['type'] == 'Movie') & (df_c1['country'] != 'Unknown_country')]
3 movies.head(2)
```



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	li:
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown_cast	United States	2021-09-25	2020	PG-13	90	Docum
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United	2021-09-24	1993	TV-MA	125	Inde Inter

Next steps:

[Generate code with movies](#)[View recommended plots](#)[New interactive sheet](#)

```
1 # unnesting country columns
2 movies['country1'] = movies['country'].str.split(', ')
3 movies['country1']
4
5 # Explode the genres into individual rows
6 movies= movies.explode('country1')
7
```

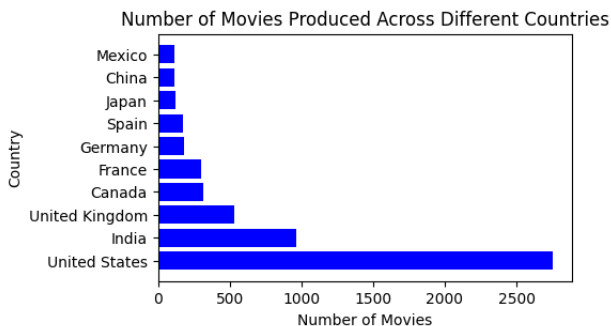
```
1 # group by each country and counting unique numbers of movies.
2 c = movies.groupby(['country1'])['release_year'].count().sort_values(ascending=False).head(10)
3 c.reset_index()
```



	country1	release_year
0	United States	2751
1	India	962
2	United Kingdom	532
3	Canada	319
4	France	303
5	Germany	182
6	Spain	171
7	Japan	119
8	China	114
9	Mexico	111



```
1 # plotting bar graph.
2 plt.figure(figsize=(5,3))
3 plt.barh(c.index,c.values, color=['b'])
4 plt.xlabel('Number of Movies')
5 plt.ylabel('Country')
6 plt.title('Number of Movies Produced Across Different Countries')
7 plt.show()
```



The United States produces the highest number of movies (2,751), followed by India with 962 and the United Kingdom with 532.

2.2. Finding the number of Tv-Shows produced in each country and pick the top 10

```
1 tv = df_c1[(df_c1['type'] == 'TV Show') & (df_c1['country'] != 'Unknown_country')]
2 tv.head(2)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	lis
1	s2	TV Show	Blood & Water	Unknown_director	Ama Qamata, Khosi Ngema, Gail Mababane, Thaban... Mayur	South Africa	2021-09-24	2021	TV-MA	2	Inter TV I M

Next steps: [Generate code with tv](#) [View recommended plots](#) [New interactive sheet](#)

```
1 tv['country1'] = tv['country'].str.split(', ')
2 tv['country1']
3 # Explode the genres into individual rows
4 tv= tv.explode('country1')
5
```

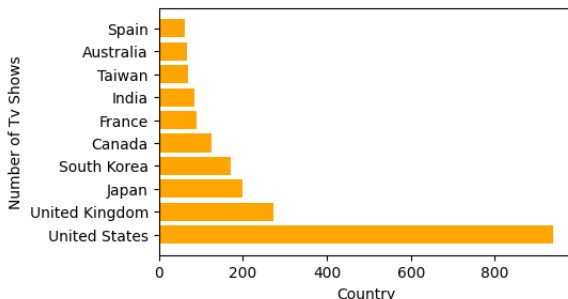
<ipython-input-68-4662fed466>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
tv['country1'] = tv['country'].str.split(', ')

```
1 b = tv.groupby(['country1'])['release_year'].count().sort_values(ascending=False).head(10)
2 print(b.reset_index())
3
4 plt.figure(figsize = (5,3))
5 plt.barh(b.index,b.values,color='orange')
6 plt.xlabel('Country')
7 plt.ylabel('Number of Tv Shows')
8 plt.title('Tv Shows Produced Across Different Countries')
9 plt.show()
```

```
country1  release_year
0  United States      938
1  United Kingdom     272
2      Japan         199
3  South Korea        170
4      Canada        126
5      France         90
6      India          84
7      Taiwan         70
8  Australia          66
9      Spain          61
```

Tv Shows Produced Across Different Countries



The United States produces the highest number of TV shows (938), followed by the United Kingdom with 272 shows and Japan with 199 shows.

3. What is the best time to launch a TV show?

```
1 #creating copy of df_c1 dataset.
2 date = df_c1.copy()
3 date.head(2)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown_cast	United States	2021-09-25	2020	PG-13	90
1	s2	TV Show	Blood & Water	Unknown_director	Ama Qamata, Khosi Ngema, Gail Mabalane,	South Africa	2021-09-24	2021	TV-MA	2

Next steps: [Generate code with date](#) [View recommended plots](#) [New interactive sheet](#)

```
1 # chacking data type of date_added column.
2 date.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      8709 non-null   datetime64[ns]
7   release_year    8807 non-null   int64
8   rating          8803 non-null   object
9   duration        8807 non-null   object
10  listed_in       8807 non-null   object
11  description      8807 non-null   object
12  duration_type    8807 non-null   object
dtypes: datetime64[ns](1), int64(1), object(11)
memory usage: 894.6+ KB
```

```
1 date['date_added'].isnull().sum()
```

```
98
```

```
1 # dropping null values from date_added column
2 date.dropna(subset=['date_added'], inplace=True)
```

```
1 date['date_added'].isnull().sum()
```

```
0
```

```
1 # creating day_added, month_added, year_added, and week_added column.
2 date['day_added'] = date['date_added'].dt.day
3 date['month_added'] = date['date_added'].dt.month
4 date['year_added'] = date['date_added'].dt.year
5 date['week_added'] = date['date_added'].dt.isocalendar().week
```

```

1 # Create a week-of-month function
2 def week_of_month(date):
3     first_day = date.replace(day=1)
4     return (date.day + first_day.weekday()) // 7 + 1
5
6 # Apply the function to the date_added column
7 date['week_of_month'] = date['date_added'].apply(week_of_month)
8
9

```

3.1. Finding which is the best week and month to release the Tv-show.

```

1 # Filtering data based on tv show
2 date_tv = date[date['type'] == 'TV Show']
3 date_tv.head(2)

```



	show_id	type	title	director	cast	country	date_added	release_year	rating	dur
1	s2	TV Show	Blood & Water	Unknown_director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown_country	2021-09-24	2021	TV-MA	

Next steps: [Generate code with date_tv](#) [View recommended plots](#) [New interactive sheet](#)

```

1 # chacking rows and columns.
2 date_tv.shape

```





(2578, 18)

```

1 count_week = date_tv.groupby(['year_added', 'month_added', 'week_added', 'week_of_month'])['show_id'].count().sort_index()
2 count_week.head()

```



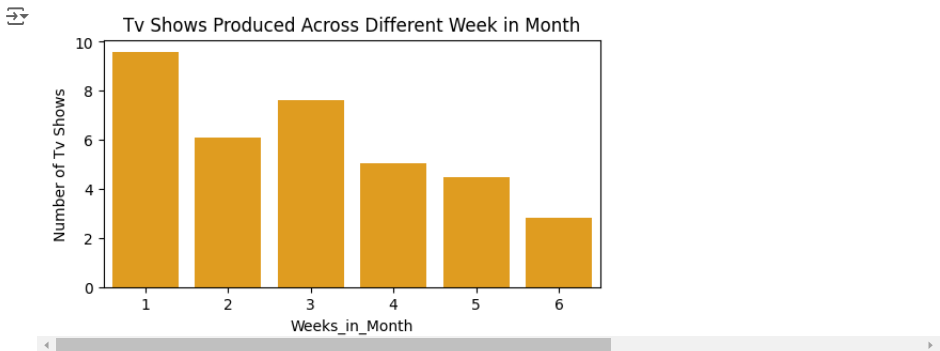
	year_added	month_added	week_added	week_of_month	show_id	
0	2021	7	27	2	42	
1	2021	6	24	3	41	
2	2019	10	40	1	27	
3	2021	4	15	3	27	
4	2017	8	31	1	24	

Next steps: [Generate code with count_week](#) [View recommended plots](#) [New interactive sheet](#)

```

1 plt.figure(figsize=(6,3))
2 sns.barplot(x='week_of_month', y='show_id', data=count_week, color = 'orange',errorbar=None)
3 #plt.bar(x = count_week['week_added'], height = count_week['show_id'],color='red') # Changed y to height
4 plt.xlabel('Weeks_in_Month')
5 plt.ylabel('Number of Tv Shows')
6 plt.title('Tv Shows Produced Across Different Week in Month')
7 plt.show()

```

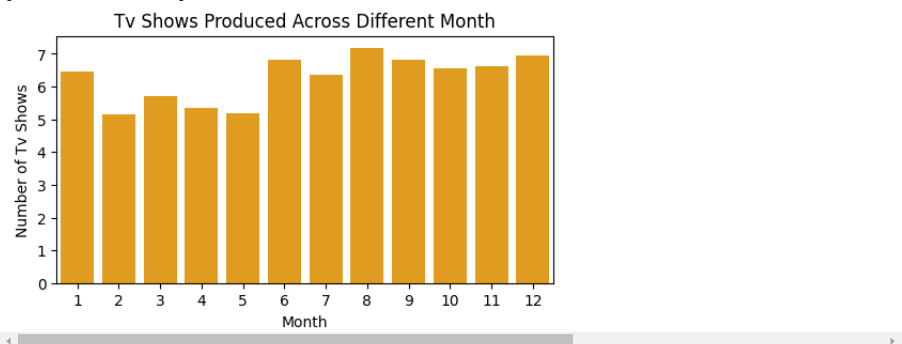


The highest number of tv shows are released in the first week of the month, making it the best time to release tv shows.

```
1 # Convert the index to columns
2 print(count_week)
3 plt.figure(figsize=(6,3))
4 sns.barplot(x='month_added', y='show_id', data=count_week, color = 'orange',errorbar=None)
5 #plt.bar(x = count_week['week_added'], height = count_week['show_id'],color='red') # Changed y to height
6 plt.xlabel('Month')
7 plt.ylabel('Number of Tv Shows')
8 plt.title('Tv Shows Produced Across Different Month')
9 plt.show()
```

	year_added	month_added	week_added	week_of_month	show_id
0	2021	7	27	2	42
1	2021	6	24	3	41
2	2019	10	40	1	27
3	2021	4	15	3	27
4	2017	8	31	1	24
...
408	2020	3	11	4	1
409	2018	4	15	4	1
410	2018	4	18	6	1
411	2020	2	8	5	1
412	2008	2	6	2	1

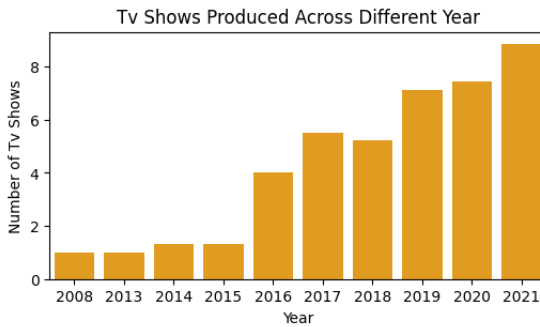
[413 rows x 5 columns]



The highest number of tv show relised in October.

```
1 plt.figure(figsize=(6,3))
2 sns.barplot(x='year_added', y='show_id', data=count_week, color = 'orange',errorbar=None)
3 #plt.bar(x = count_week['week_added'], height = count_week['show_id'],color='red') # Changed y to height
4 plt.xlabel('Year')
5 plt.ylabel('Number of Tv Shows')
```

```
6 plt.title('Tv Shows Produced Across Different Year')
7 plt.show()
```



Highest number of tv shows relised in year 2021.

3.2. Finding which is the best week and month to release the Movies

```
1 date_movie = date[date['type'] == 'Movie']
2 date_movie.head(2)
```



	show_id	type	title	director	cast	country	date_added	release_year	rating	durati
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown_cast	United States	2021-09-25	2020	PG-13	
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	Unknown_country	2021-09-24	2021	PG	

Next steps:

[Generate code with date_movie](#)
[View recommended plots](#)
[New interactive sheet](#)

```
1 date_movie.shape
```



```
(6131, 18)
```

```
1 count_movies = date_movie.groupby(['year_added', 'month_added', 'week_added', 'week_of_month'])['show_id'].count()
2 count_movies.head()
```



	year_added	month_added	week_added	week_of_month	show_id
0	2020	1	1	1	100
1	2018	10	40	1	89
2	2019	11	44	1	87
3	2018	3	9	1	75
4	2019	12	1	6	67

Next steps:

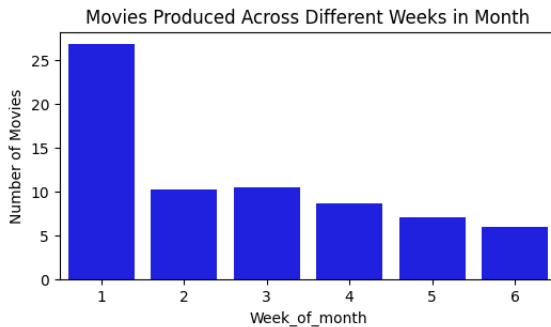
[Generate code with count_movies](#)
[View recommended plots](#)
[New interactive sheet](#)

```
1 plt.figure(figsize=(6,3))
2 sns.barplot(x='week_of_month', y='show_id', data=count_movies, color = 'blue',errorbar=None)
```

```

3 plt.xlabel('Week_of_month')
4 plt.ylabel('Number of Movies')
5 plt.title('Movies Produced Across Different Weeks in Month')
6 plt.show()

```

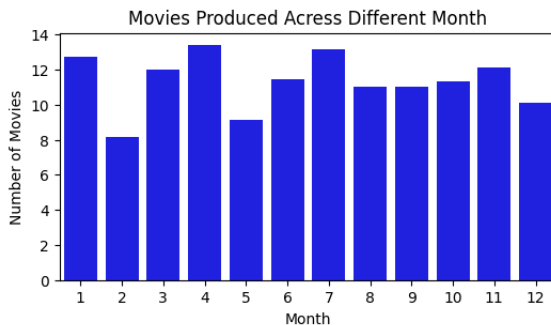


The highest number of movies are released in the first week of the month, making it the best time to release a movie.

```

1 plt.figure(figsize = (6,3))
2 sns.barplot(x='month_added',y='show_id',data=count_movies,color='blue',errorbar=None)
3 plt.xlabel('Month')
4 plt.ylabel('Number of Movies')
5 plt.title('Movies Produced Across Different Month')
6 plt.show()

```

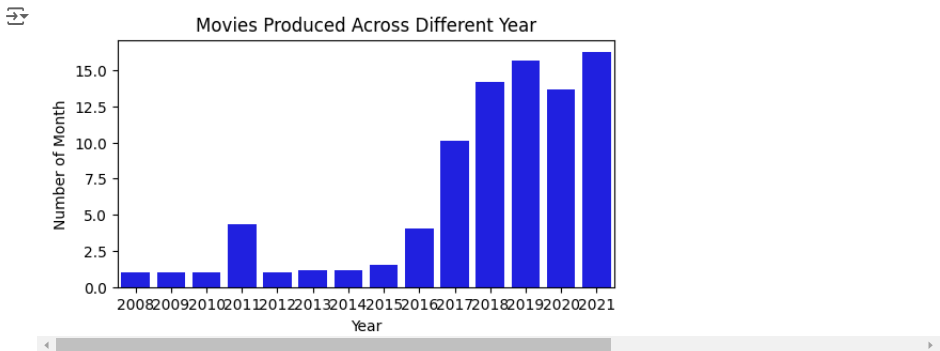


The highest number of movies are released in January, April, and July.

```

1 plt.figure(figsize=(6,3))
2 sns.barplot(x='year_added',y='show_id',data=count_movies,color='blue',errorbar=None)
3 plt.xlabel('Year')
4 plt.ylabel('Number of Month')
5 plt.title('Movies Produced Across Different Year')
6 plt.show()

```



Highest number of movie rekised in 2021 followed by 2019.

4. Analysis of actors/directors of different types of shows/movies.

4.1. Identify the top 10 actors who have appeared in most movies or TV shows.

```
1 actor = date.copy()
2 actor = actor[actor['cast'] != 'Unknown_cast']
```

```
1 actor['actor1'] = actor['cast'].str.split(',')
2 actor = actor.explode('actor1')
3 actor.head(2)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	list
1	s2	TV Show	Blood & Water	Unknown_director	Ama Qamata, Khosi Ngema, Gail Mabalan... Thaban...	South Africa	2021-09-24	2021	TV-MA	2	Intern TV s TV D My
1	s2	TV Show	Blood & Water	Unknown_director	Ama Qamata, Khosi Ngema, Gail Mabalan... Thaban...	South Africa	2021-09-24	2021	TV-MA	2	Intern TV s TV D My

```
1 actor_count = actor.groupby(['actor1'])['title'].count().sort_values(ascending=False).reset_index().head(10)
2 actor_count
```




	actor1	title
0	Anupam Kher	39
1	Rupa Bhimani	31
2	Takahiro Sakurai	29
3	Julie Teiwani	28
4	Om Puri	27
5	Rajesh Kava	26
6	Shah Rukh Khan	26
7	Paresh Rawal	25
8	Yuki Kaji	25
9	Boman Irani	25



Next steps:

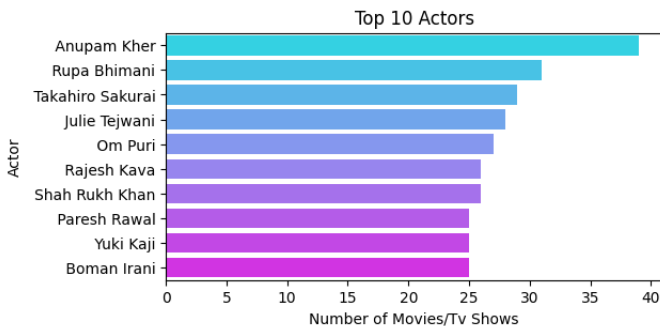
[Generate code with actor_count](#)



[View recommended plots](#)

[New interactive sheet](#)

```
1 plt.figure(figsize=(6,3))
2 sns.barplot(x='title', y='actor1', data=actor_count, palette='cool', hue='actor1', legend=False)
3 plt.xlabel('Number of Movies/Tv Shows')
4 plt.ylabel('Actor')
5 plt.title('Top 10 Actors')
6 plt.show()
```



Anupam Kher is top actor who apper in 39 shows followed by Rupa Bhimani (31 show) and Takahiro Sakurai (29 show).

4.2. Identify the top 10 directors who have appeared in most movies or TV shows.

```
1 director = date.copy()
2 director = director[director['director'] != 'Unknown_director']
```

```
1 director['director1'] = director['director'].str.split(',')
2 director = director.explode('director1')
3 director.head(2)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	durati
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown_cast	United States	2021-09-25	2020	PG-13	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabl...	Unknown_country	2021-09-24	2021	TV-MA	

Next steps:
 [Generate code with director](#)
[View recommended plots](#)
[New interactive sheet](#)

```

1 director_count = director.groupby(['director1'])['title'].count().sort_values(ascending=False).reset_index().ho
2 director_count

```

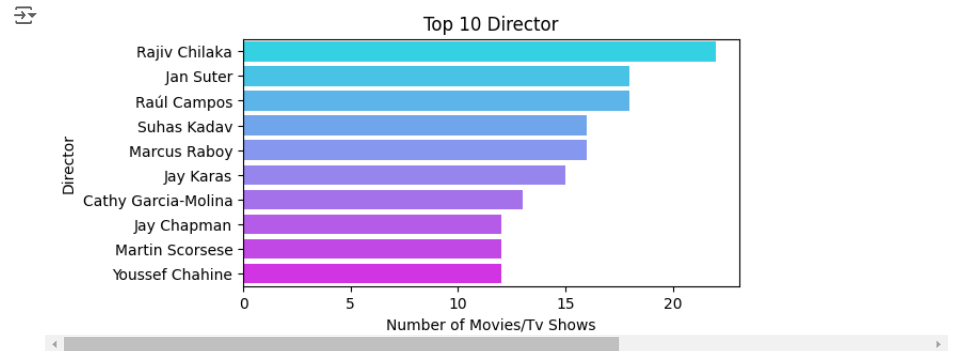
	director1	title	
0	Rajiv Chilaka	22	
1	Jan Suter	18	
2	Raúl Campos	18	
3	Suhas Kadav	16	
4	Marcus Raboy	16	
5	Jay Karas	15	
6	Cathy Garcia-Molina	13	
7	Jay Chapman	12	
8	Martin Scorsese	12	
9	Youssef Chahine	12	

Next steps:
 [Generate code with director_count](#)
[View recommended plots](#)
[New interactive sheet](#)

```

1 plt.figure(figsize=(6,3))
2 sns.barplot(x='title', y='director1', data=director_count, palette='cool', hue='director1', legend=False)
3 plt.xlabel('Number of Movies/Tv Shows')
4 plt.ylabel('Director')
5 plt.title('Top 10 Director')
6 plt.show()

```



Rajiv Chilaka is top director followed by Jan Suter and Raúl Campos.

5. Which genre movies/tv shows are more popular or produced more

```
1 # creating copy of data as genre
2 genre = date.copy()
3 genre['genre1'] = genre['listed_in'].str.split(',')
4 genre = genre.explode('genre1')
5 genre.head(2)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown_cast	United States	2021-09-25	2020	PG-13	90
1	s2	TV Show	Blood & Water	Unknown_director	Ama Qamata, Khosi Ngema, Gail Mabalanane, Thabane...	South Africa	2021-09-24	2021	TV-MA	2

Next steps:

[Generate code with genre](#)

[View recommended plots](#)

[New interactive sheet](#)

```
1 genre_count = genre.groupby(['genre1'])['title'].count().sort_values(ascending=False).reset_index().head(10)
2 genre_count
```

	genre1	title
0	International Movies	2624
1	Dramas	1600
2	Comedies	1210
3	Action & Adventure	859
4	Documentaries	829
5	Dramas	827
6	International TV Shows	761
7	Independent Movies	736
8	TV Dramas	679
9	Romantic Movies	613

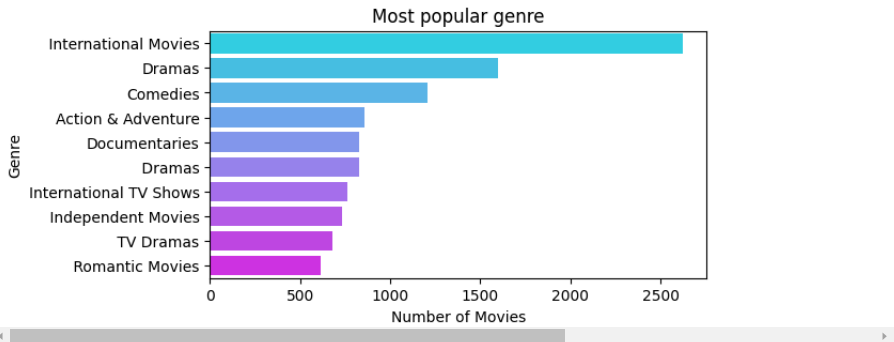
Next steps:

[Generate code with genre_count](#)

[View recommended plots](#)

[New interactive sheet](#)

```
1 plt.figure(figsize=(6,3))
2 sns.barplot(x='title',y='genre1', data=genre_count, palette='cool', hue='genre1', legend=False)
3 plt.xlabel('Number of Movies')
4 plt.ylabel('Genre')
5 plt.title('Most popular genre')
6 plt.show()
```



```
1 from wordcloud import WordCloud
2
3 plt.subplots(figsize=(6,`))
4 wordcloud=WordCloud(
5     background_color="white",
6     width=1920,
7     height=1080).generate(" ".join(genre['genre1']))
8 plt.imshow(wordcloud)
9 plt.axis("off")
10 plt.savefig("country.png")
11 plt.show()
```



1. Number of Seasons in TV Shows:

- ## 2. TV Show Ratings:

- ### 3. Movie Duration:

- #### 4. Comparison of TV Shows and Movies:

- ### 5. Top 10 Countries Producing Movies:

- The United States leads with 2,751 movies, followed by India with 962 and the United Kingdom with 532.

6. Top 10 Countries Producing TV Shows:

- The United States produces the most TV shows (938), followed by the United Kingdom (272) and Japan (199).

7. Best Time to Release TV Shows and Movies:

◦ TV Show Analysis:

- **Best Week:** Most TV shows are released in the first week of the month, making it the best time to launch new shows.
- **Best Month:** October sees the highest number of TV show releases.
- **Best Year:** 2021 had the highest number of TV show releases, showing an increasing trend in production over the years.

◦ Movie Analysis:

- **Best Week:** The first week of the month is also the best time to release movies.
- **Best Months:** January, April, and July are the months with the most movie releases.
- **Best Year:** 2021 had the most movie releases, followed by 2019, indicating consistent growth in movie production over time.

8. Top 10 Actors Appearing in Movies or TV Shows:

- Anupam Kher appears in the most shows (39), followed by Rupa Bhimani (31) and Takahiro Sakurai (29).

9. Top 10 Directors with the Most Movies or TV Shows:

- Rajiv Chilaka is the top director, followed by Jan Suter and Raúl Campos.

10. Most Popular Genres:

- The most popular genre on Netflix is International Movies, followed by Dramas and Comedies.

7. Recommendations for Netflix to Grow Business in Different Countries:

1. Expand Production in Growing Markets:

- Since India and the United Kingdom are among the top producers of both movies and TV shows, Netflix should continue to invest in local content and expand partnerships with filmmakers in these regions.

2. Focus on Popular Genres:

- International Movies, Dramas, and Comedies are the most popular genres. Producing more content in these genres will likely attract a broader audience.

3. Target Adult Audience:

- With the majority of TV shows rated for adult audiences, Netflix should focus on producing more adult-themed content, such as dramas and thrillers, while also considering expanding family-friendly content to reach more viewers.

4. Optimize Release Times:

- Since the first week of the month and certain months (January, April, July, and October) show the highest number of releases, Netflix can strategically plan new releases during these periods to maximize viewership.

5. Promote Top Actors and Directors:

- Netflix should consider highlighting shows and movies featuring top actors like Anupam Kher and directors like Rajiv Chilaka in global promotions, especially in regions where these talents are well known.

6. Increase Production in Emerging Markets:

- While the United States dominates content production, there is potential for growth in emerging markets like