NETFLIX BUSINESS CASE STUDY - by S Omprakash

Tool used : Google Collab

DATA EXPLORATION :

1) Perform initial analysis to find the number of columns and rows and data type

```
[1] !gdown 1LK6uc4SGUEft4yY350RUKLqXhV3nMqZm
```

```
Downloading...
From: https://drive.google.com/uc?id=1LK6uc4SGUEft4yY350RUKLqXhV3nMqZm
To: /content/netflix.csv
100% 3.40M/3.40M [00:00<00:00, 157MB/s]
```

```
[3] import numpy as np
    import pandas as pd

    data = pd.read_csv('netflix.csv')
```

```
data.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
```

Insights : It is observed that there are about 8807 rows and 12 columns with data type as object and release year as int64. Observing the total number of rows, we can observe that some columns have null values in it. So, columns such as director, cast, country and date_added has NULL values

Recommendation's:

Handing the Null Values are important as when we perform a columns specific manipulation it will be difficult to get correct insights. Either we can drop the NULL values using data. Dropna() function or replace the null values with some values like data.fillna()

2) Checking the NULL values in every column

```
data.isna().sum().reset_index()
```

1 to 12 of 12 entries Filter

level_0	index	0
0	show_id	0
1	type	0
2	title	0
3	director	2634
4	cast	825
5	country	831
6	date_added	10
7	release_year	0
8	rating	4
9	duration	3
10	listed_in	0
11	description	0

Show 25 per page

```
data[data.duplicated()]
```

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

Insights :

1. **Shape of data** - There are 8807 rows and 12 columns in the data
2. **Missing values** - There are few missing values in some of the columns
3. **Datatypes of all attributes** - 11 columns are containing string values, and 1 column is having integer values.

Recommendations:

We can either fill the Null values with zero or string such as ‘not specified’ or drop the rows with NULL values.

Basic Analysis

1) **How many different types of the shows are screened ?**

```
data.groupby('type')['show_id'].count()
```

type	show_id
Movie	6131
TV Show	2676

dtype: int64

Insights: The dataset includes both movies and TV shows, and it's clear that movies dominate in terms of the number of titles available. This could suggest that movies are more popular or are easier to produce than TV shows. **There are 6131 movies and 2676 TV Shows on our platform, and in terms of percentages, we have 69.6% Movies and 30.4% TV Shows.** .

Recommendation's:

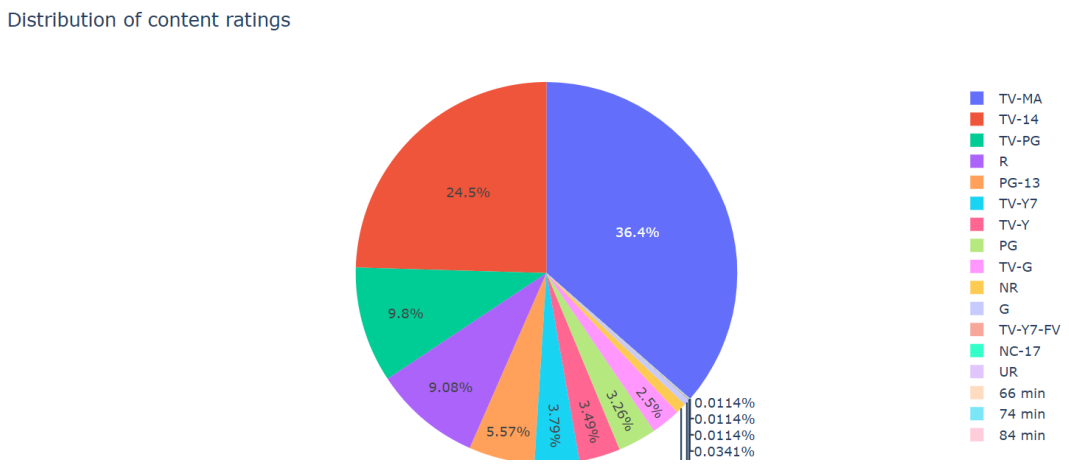
We can get Genre of the movie type so Netflix can produce the specific content and genre of the customer.

2) **Find the most common rating to the movies and TV shows .**

```
data.groupby(['type','rating'])['show_id'].count().sort_values(ascending=False).reset_index()
```

	type	rating	show_id
0	Movie	TV-MA	2062
1	Movie	TV-14	1427
2	TV Show	TV-MA	1145
3	Movie	R	797
4	TV Show	TV-14	733
5	Movie	TV-PG	540
6	Movie	PG-13	490

```
import plotly.express as px
pichart= px.pie(x,values='Count',names='rating',title='Distribution of content ratings')
pichart.show()
```



Insights : It is observed that TV-MA rating is the highest with 36.4 % percentage of the total which is the matured adult ratings followed by TV-14 and TV-PG.

3) Top 5 directors on Netflix

```
directorlist=pd.DataFrame()
```

```
[12] directorlist['Director'] = data['director'].str.split(',',expand=True).stack()
```

```
[15] group_director=directorlist.groupby("Director").size().reset_index(name='TotalCounts')
```

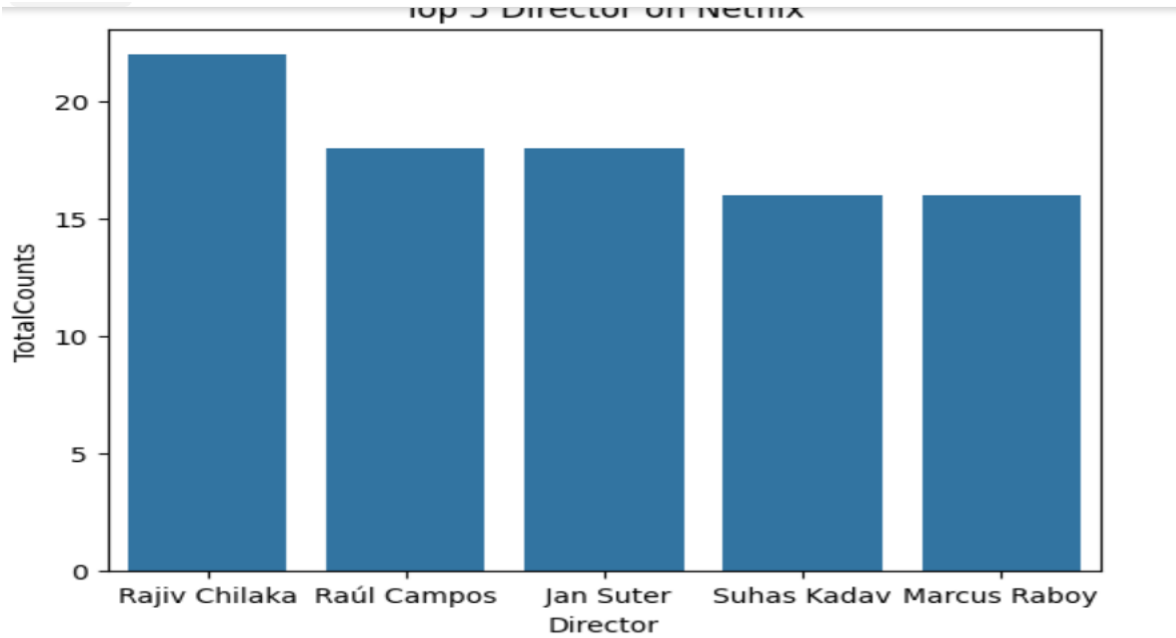
```
group_director=group_director[group_director['Director']!='not available']
group_director.sort_values(by=['TotalCounts'],ascending = False,inplace=True)
group_director
```

	Director	TotalCounts
4020	Rajiv Chilaka	22
4067	Raúl Campos	18
261	Jan Suter	18
4651	Suhas Kadav	16
3235	Marcus Raboy	16
...	...	...
5061	Yeo Siew Hua	1
5062	Yesim Ustaoglu	1
5063	Yeung Yat-Tak	1
5064	Yibrán Asuad	1
2560	Joaquín Mazón	1

```
top5=group_director.head(5)
top5
```

	Director	TotalCounts
4020	Rajiv Chilaka	22
4067	Raúl Campos	18
261	Jan Suter	18
4651	Suhas Kadav	16
3235	Marcus Raboy	16

```
import seaborn as sns
import matplotlib.pyplot as pt
sns.barplot(x='Director',y='TotalCounts',data=top5)
pt.title('Top 5 Director on Netflix')
pt.show()
```



4) **TOP Actor on the Netflix**

```
cast = data[['cast','title','show_id']]
cast['cast'] = cast['cast'].str.split(',')
cast_explode = cast.explode('cast').reset_index(drop=True)
cast_explode.groupby(['cast'])['show_id'].count().sort_values(ascending=False)
```

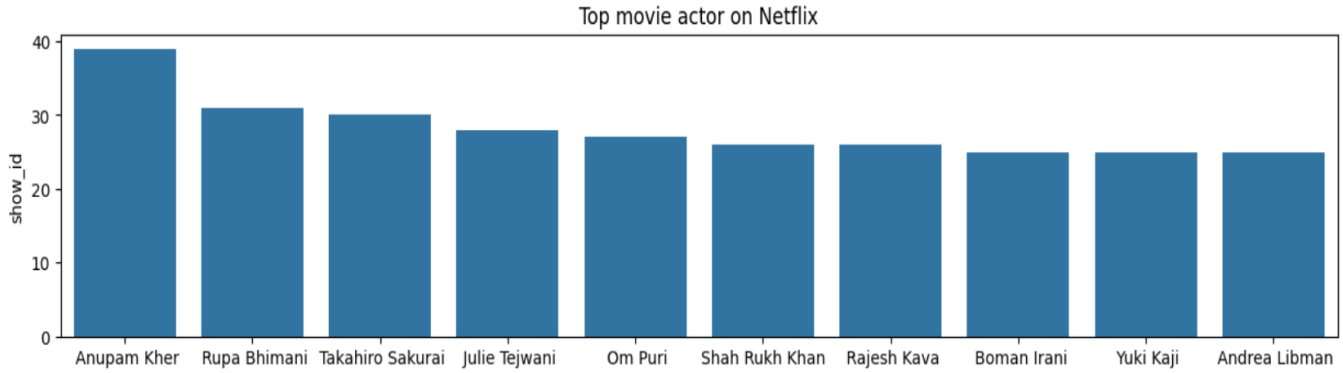
cast	show_id
Anupam Kher	39
Rupa Bhimani	31
Takahiro Sakurai	30
Julie Tejwani	28
Om Puri	27
Shah Rukh Khan	26
Rajesh Kava	26
Boman Irani	25
Yuki Kaji	25
Andrea Libman	25

dtype: int64

```
[57] top5actor=top5actor.to_frame()
```

```
pt.figure(figsize=(15,3))
sns.barplot(x='cast',y='show_id',data=top5actor)
pt.title('Top movie actor on Netflix')
```

```
Text(0.5, 1.0, 'Top movie actor on Netflix')
```



Insights:

From the graph, we can find that Anupam Kher has acted in most of the content and Rupa Bhimani with 31. Maybe they are the most popular actors or they suitable to the character of the movies so director is more willing to cast them.

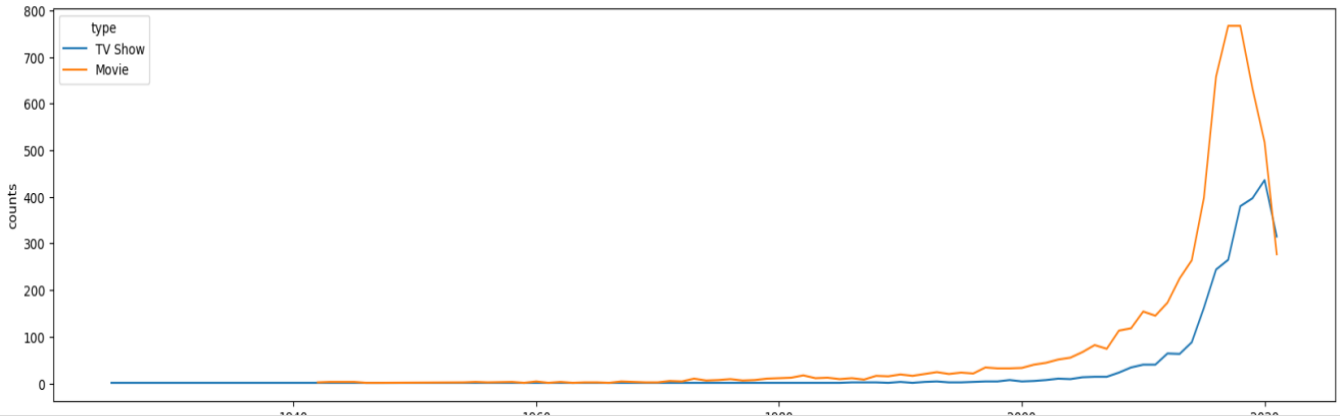
Recommendations: We have identified actors and director who have consistently been part of successful content, which we have given in insights section. We should collaborate with these professionals more frequently as they have a track record of attracting viewers.

5) Trend of the content produced on Netflix over years

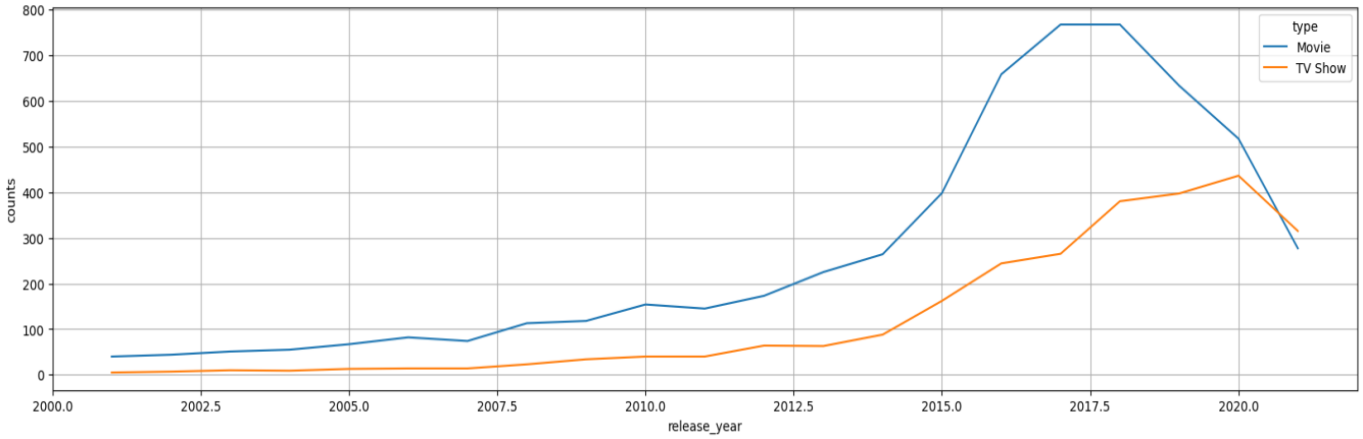
```
yearvstype=data[['type','release_year']]
yearvstype=yearvstype.groupby(['release_year','type']).size().reset_index(name='counts')
```

```
[44] import seaborn as sns
     pt.figure(figsize=(20,5))
     sns.lineplot(x='release_year',y='counts',data=yearvstype,hue='type')
```

```
<Axes: xlabel='release_year', ylabel='counts'>
```




```
import seaborn as sns
pt.figure(figsize=(20,5))
sns.lineplot(x='release_year',y='counts',data=yearvstype,hue='type')
pt.grid()
```



Insights :

If we consider from 1925, Netflix started with movies content type. In the mid of 1935 they started also looking into Tv series. In specific if we look from 2000 to 2020 both has better performed good. But the movies content has done better than the TV shows until 2020. In specific year 2020 TV shows is doing best than the movies.

Recommendations:

While movies dominate the platform, continue to invest in high-quality movies to maintain and expand the movie library. We should consider user preferences and genres that have been successful in the past.

6)Top 10 Content Consuming Countries on Netflix

```
import pandas as pd

df=pd.read_csv('netflix.csv')
```

```
constraint = df["country"].apply(lambda x: str(x).split(", ")).tolist()
df_country = pd.DataFrame(constraint, index = df["title"])
df_country = df_country.stack()
df_country = pd.DataFrame(df_country)
df_country.reset_index(inplace = True)
df_country = df_country[["title", 0]]
df_country.columns = ["title", "country"]
mostdirector=pd.DataFrame()
```

```
[109] mostdirector =df_country.groupby('country').size().reset_index(name='count').sort_values(by='count',ascending=False)
```

```
[118] mostdirector.drop(['level_0','index'],axis=1,inplace=True)
```

	country	count
0	United States	3689
1	India	1046
2	nan	831
3	United Kingdom	804
4	Canada	445
...	...	...
123	Mongolia	1
124	Somalia	1
125	Ethiopia	1
126	Botswana	1
127	Poland,	1

128 rows × 2 columns

Insights : We can see that "United States" is the top consumer of our content where total 3689 number of content titles have been watched. So we have high number of audiences from USA and India. There is a huge Entertainment market in these countries

Recommendations:

We increase the type of the content to the target audiences and enhance the personalised recommendations of new content to them.

7) Top 10 Release Years on Netflix.

```
result_release_year = df.groupby(["release_year"])["title"].nunique()
result_release_year = result_release_year.sort_values(ascending = False)
result_release_year.head(10)
```

release_year	title
2018	1147
2017	1032
2019	1030
2020	953
2016	902
2021	592
2015	560

Insights: we can notice that in the year 2018 highest number of contents are produced.

8) Top 10 Genres on the Netflix

```
result_listed_in = df_listed_in.groupby(["listed_in"])["title"].nunique()
result_listed_in = result_listed_in.sort_values(ascending = False)
result_listed_in.head(10)
```

listed_in	title
International Movies	2752
Dramas	2427
Comedies	1674
International TV Shows	1351
Documentaries	869
Action & Adventure	859
TV Dramas	763
Independent Movies	756

Insights: International Movies stands first and most people also like the genre Dramas**.** If any director wishes to do a content, they can check the top 10 Genre so to get a decent revenue.

9) Comparison of TV Shows vs. Movies

Top 10 countries and the number of Movies produced in each country.

```
movies_data = df[df['type'] == 'Movie']
movie_counts = movies_data['country'].value_counts().reset_index()
movie_counts.columns = ['Country', 'Num_of_Movies']
```

```
top_10_movies_countries = movie_counts.head(10)
print(top_10_movies_countries)
```

```
          Country  Num_of_Movies
0   United States           2058
1           India            893
2  United Kingdom            206
3          Canada            122
4           Spain             97
5           Egypt             92
6         Nigeria             86
7       Indonesia             77
8          Turkey             76
9           Japan             76
```

Top 10 countries and the number of TV Shows produced in each country

```
tv_show_data = df[df['type'] == 'TV Show']
tv_show_counts = tv_show_data['country'].value_counts().reset_index()
tv_show_counts.columns = ['Country', 'Num_of_TV_Shows']
top_10_tv_show_countries = tv_show_counts.head(10)
print(top_10_tv_show_countries)
```

```
          Country  Num_of_TV_Shows
0   United States              760
1  United Kingdom              213
2           Japan              169
3     South Korea              158
4           India               79
5          Taiwan               68
6          Canada               59
7          France               49
8       Australia               48
9           Spain               48
```

Insights: We can observe that USA stands top in both the type of content, where in Movies type India stands second but coming to TV series UK stands second**.** We must consider the data as it is specific to the country so we can encourage the director to choose either Tv series or Movies with respect to the country.