Case Study

# NETFLIX DATA ANALYSIS

Akanksha Trivedi

Scaler Academy

# ABSTRACT

Netflix is a provider of entertainment services. The company offers Tv shows and movies such as original series, documentaries and feature films through an internet subscription on TV, computer and mobile services. It began its operations in 1997, founded by two tech entrepreneur Reed Hastings and Marc Randolph. The Company's head office is in Los Gatos, California. Over the period with growth of internet users and the decline of DVD sales and rental services, it changed its business model to video on demand. From 2012 onwards, it started producing its original TV-series and movies.

Netflix collects huge amounts of data from a vast variety of subscriber based. It collects data such as the location of a user, content watched by the user, user interests, the data searched by the user, and the time at which user watched. In this paper, we analyze the data and generate insights that could help Netflix deciding which type of shows/movies to produce and how they can grow the business in different countries.

# Data Science Analysis

## Import Libraries

Importing the libraries we need-

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

## Loading the dataset

Using Pandas Library ,we will load the csv file
df = pd.read_csv("netflix.csv")

# first five data

```
df.head()
```

| | 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 I... |

## Length of data

```
len(df)
```

```
8807
```

## Let's check the NaN values in some columns:-

```
df.isna().sum()
```

```
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
```

## Observation from above steps

The dataset contains over 8807 titles, 12 descriptions and we can see that there are nan values in some columns as follows-

1. director with 2634 nan values
2. country with 831 nan values
3. cast with 825 nan values
4. date_added with 10 nan values
5. rating with 4 nan values
6. duration with 3 nan values

## To get all attributes

```
df.columns
```

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

## The shape of data

```
df.ndim
```

```
2
```

## Data types of all the attributes

```
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
```

## Data cleaning operations

Identifying the missing values

```
print("Columns with missing values:-")
print(df.isnull().any())
```

```
Columns with missing values:-
show_id         False
type            False
title           False
director         True
cast             True
country          True
date_added       True
release_year    False
rating           True
duration         True
listed_in       False
description     False
dtype: bool
```

## No data or Null data Count

```
df.isnull().sum().sum()
```

```
4307
```

There are 4307 null values across the entire dataset.

## Imputation method

```
[28] df.director.fillna("no director", inplace = True)
     df.cast.fillna("no cast", inplace = True)
     df.country.fillna("no country", inplace = True)
     df.dropna(subset = ["date_added","rating","duration"],inplace = True)
```

## Missing values

```
df.isnull().any()
```

```
show_id         False
type            False
title           False
director        False
cast            False
country         False
date_added      False
release_year    False
rating          False
duration        False
listed_in       False
description     False
dtype: bool
```

Finally, there are no more missing values in the datagram.

## Top directors on Netflix-Movies

```
[23] dft = df.groupby(['director','type']).size().unstack(fill_value = 0)
     dft.sort_values(['Movie','TV Show'],ascending = False)
```

| type / director | Movie | TV Show |
|---|---|---|
| no director | 187 | 2434 |
| Rajiv Chilaka | 19 | 0 |
| Raúl Campos, Jan Suter | 18 | 0 |
| Suhas Kadav | 16 | 0 |
| Marcus Raboy | 15 | 1 |
| ... | ... | ... |
| Vijay Roche | 0 | 1 |
| Vijay S. Bhanushali | 0 | 1 |
| Vikramaditya Motwane, Anurag Kashyap | 0 | 1 |
| Wouter Bouvijn | 0 | 1 |
| Yasuhiro Irie | 0 | 1 |

## Top directors on Netflix-TV Shows

```
[25] dft.sort_values('TV Show',ascending = False)
```

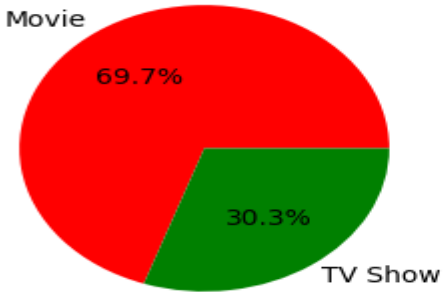
| type / director | Movie | TV Show |
|---|---|---|
| no director | 187 | 2434 |
| Alastair Fothergill | 0 | 3 |
| Ken Burns | 0 | 2 |
| Rob Seidenglanz | 0 | 2 |
| Stan Lathan | 2 | 2 |
| ... | ... | ... |
| Harry Chaskin | 1 | 0 |
| Harry Elfont, Deborah Kaplan | 1 | 0 |
| Harshavardhan Kulkarni | 1 | 0 |
| Haruka Fujita | 1 | 0 |
| Şenol Sönmez | 2 | 0 |

4527 rows × 2 columns

A quick grouping of director vs type of content we get that director **Alastair Fothergill** has done second most movies and Tv shows for Netflix platform over the years.

## Let's compare the total number of movies and TV shows

```
plt.figure(figsize= (5,3))
plt.pie(df.type.value_counts(),labels = df.type.value_counts().index,
colors  = ['red','green'],autopct ='%2.1f%%')
plt.show()
```



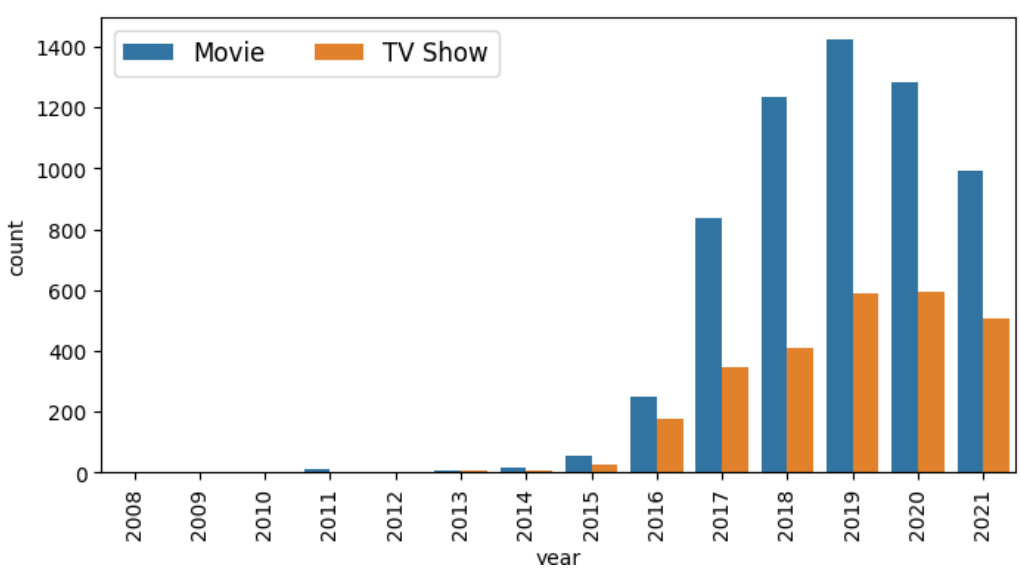

There are far more movie titles (69.7%) that TV shows titles(30.3%) in terms of title.

## Datatype changes of date_added

```
[42] df['date_added'] = pd.to_datetime(df['date_added'])
     df['date_added_year'] = pd.to_datetime(df['date_added']).dt.year
```

## Compare with respect to years

```
plt.figure(figsize=(8, 4))
plt.xticks(rotation=90)
counts=df.loc[:,['type','date_added','title']].drop_duplicates().reset_index(drop=True)
counts['year']=counts['date_added'].dt.year
sns.countplot(data=counts[counts['date_added'].dt.year>=2001], x='year', hue='type')
plt.legend(bbox_to_anchor=(0, 1), loc='upper left', fontsize='large',ncol=2)
plt.show()
```



We can see that Netflix dominated market after 2015 and over the years , Movies dominated over TV shows and most number of Movies and TV Shows released in the year 2019. So we can know that Netflix has increasingly focused on movies rather than TV shows in recent years.

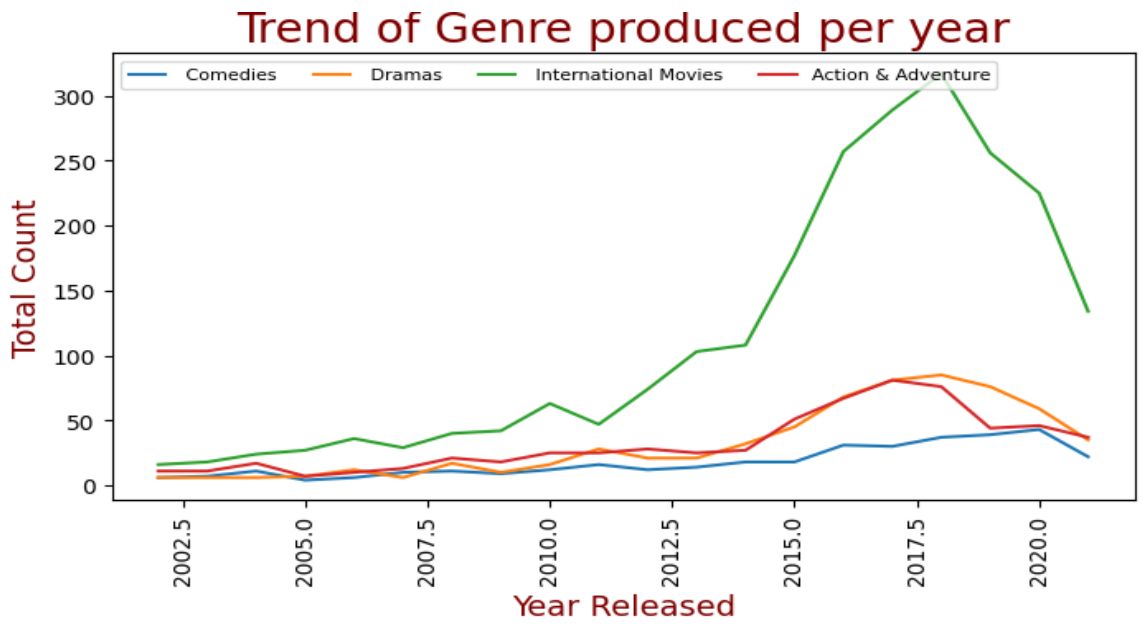
## Top 4 genre on Netflix

Below are the queries used to get top 4 genres

```
[59] df['listed_in'] = df['listed_in'].str.split(',')
     df = df.explode('listed_in')
```

```
df5 = df.groupby(['release_year', 'listed_in']).size().reset_index(name='Totalcount')
top4_listed_in = df5['listed_in'].value_counts().index[:4]
top4_data = df5.loc[(df5["listed_in"].isin(top4_listed_in))]
plt.figure(figsize=(8, 4))
plt.xticks(rotation=90)
sns.lineplot(data=top4_data[top4_data['release_year']>2001],x='release_year',
y='Totalcount',hue='listed_in')
plt.title('Trend of Genre produced per year',color='maroon',fontsize=20)
plt.xlabel('Year Released',color='maroon',fontsize=14)
plt.ylabel('Total Count',color='maroon',fontsize=14)
plt.legend(bbox_to_anchor=(0, 1), loc='upper left', fontsize= 8,ncol=4)
```

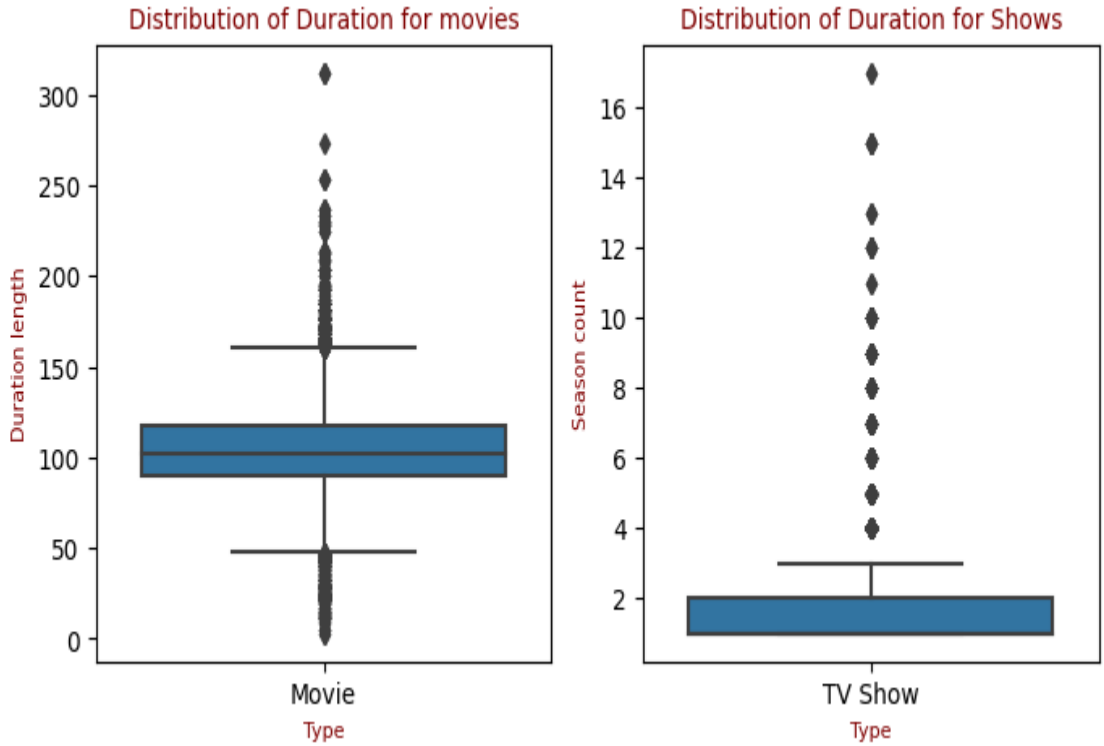
Graph plotted for above is as below.



From the above graph, international movies take the first place followed by dramas and comedy.

## Duration distribution for Movies and TV shows

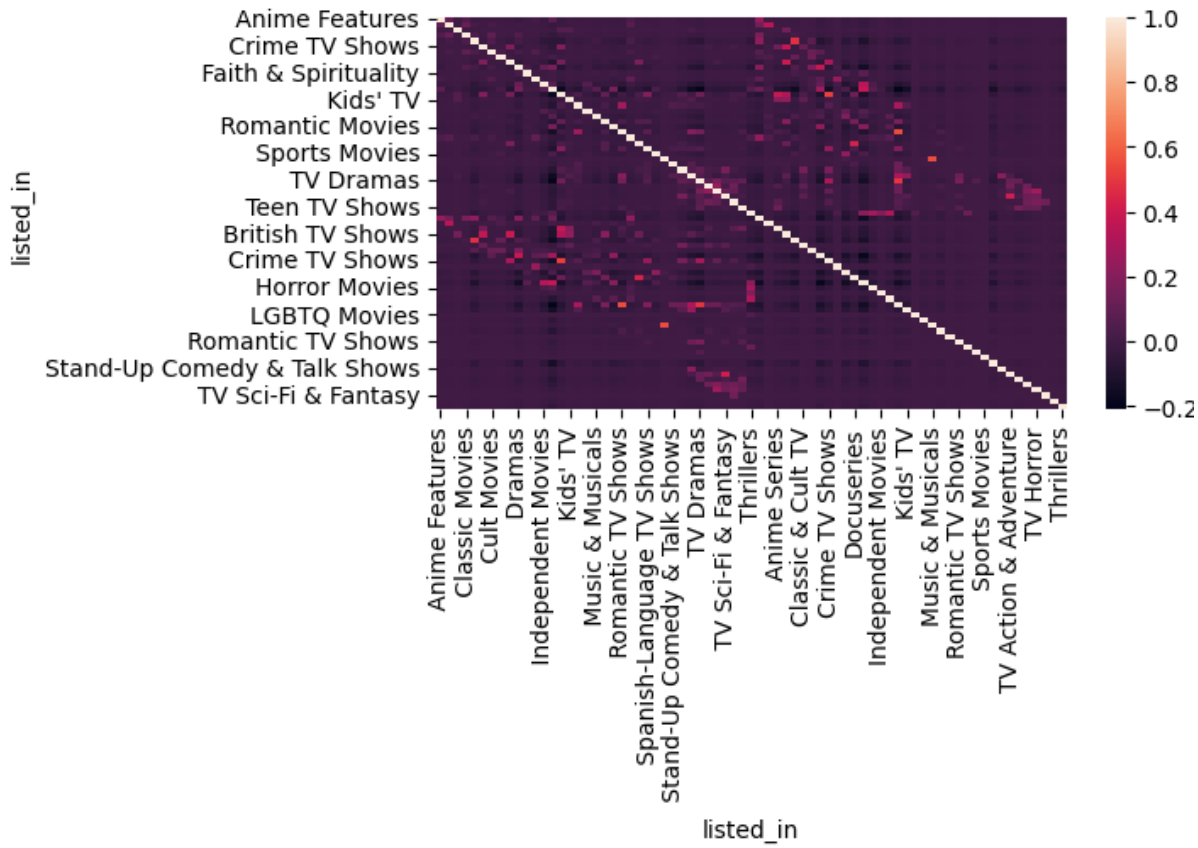
```
[90] plt.figure(figsize=(8,4))
     plt.subplot(1,2,1)
     sns.boxplot(x='type',y='duration_new',data=df[df['type'] == 'Movie'])
     plt.xlabel('Type',color='maroon',fontsize=8)
     plt.title('Distribution of Duration for movies',color='maroon',fontsize=10)
     plt.ylabel('Duration length',color='maroon',fontsize=8)
     plt.subplot(1,2,2)
     sns.boxplot(x='type',y='duration_new',data=df[df['type'] == 'TV Show'])
     plt.title('Distribution of Duration for Shows',color='maroon',fontsize=10)
     plt.xlabel('Type',color='maroon',fontsize=8)
     plt.ylabel('Season count',color='maroon',fontsize=8)
```




From the above graph , we could see that average movie duration is around 110 minutes and average TV Show duration is 1 to 2 seasons. we can see that most movies fall within a reasonable duration range, with few outliers exceedingly approximately 2.5 hours. This suggests that most movies on Netflix are designed to fit within a standard viewing time. For TV shows, the box plot reveals that most shows have one to four seasons, with very few outliers having longer durations. This aligns with the earlier trends, indicating that Netflix focuses on shorter series formats.

## Genre correlations

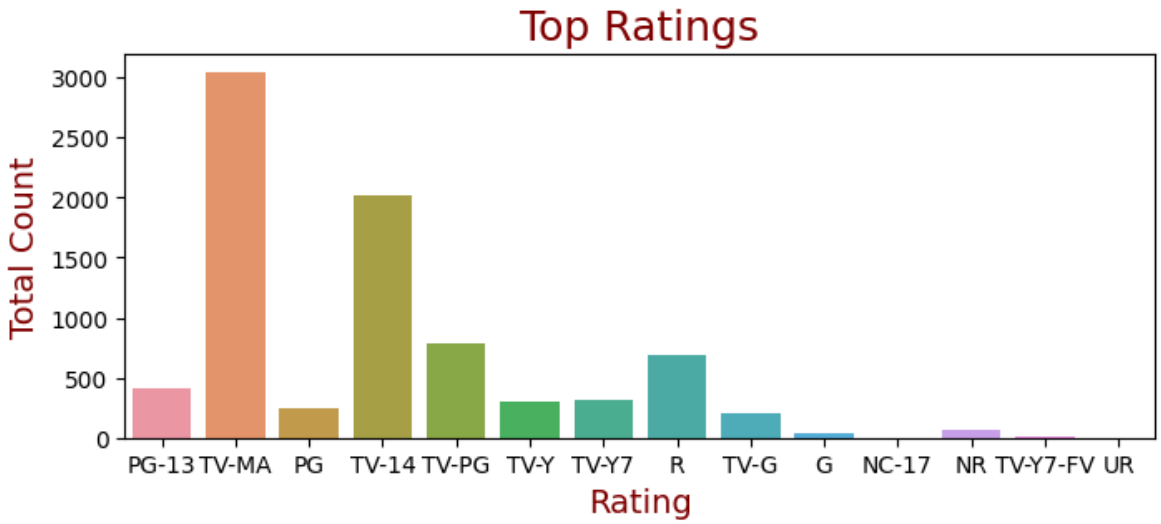
```
dfa = df.groupby(['show_id','listed_in']).size().unstack(fill_value = 0)
plt.figure(figsize = (6,3))
sns.heatmap(dfa.corr())
plt.show()
```



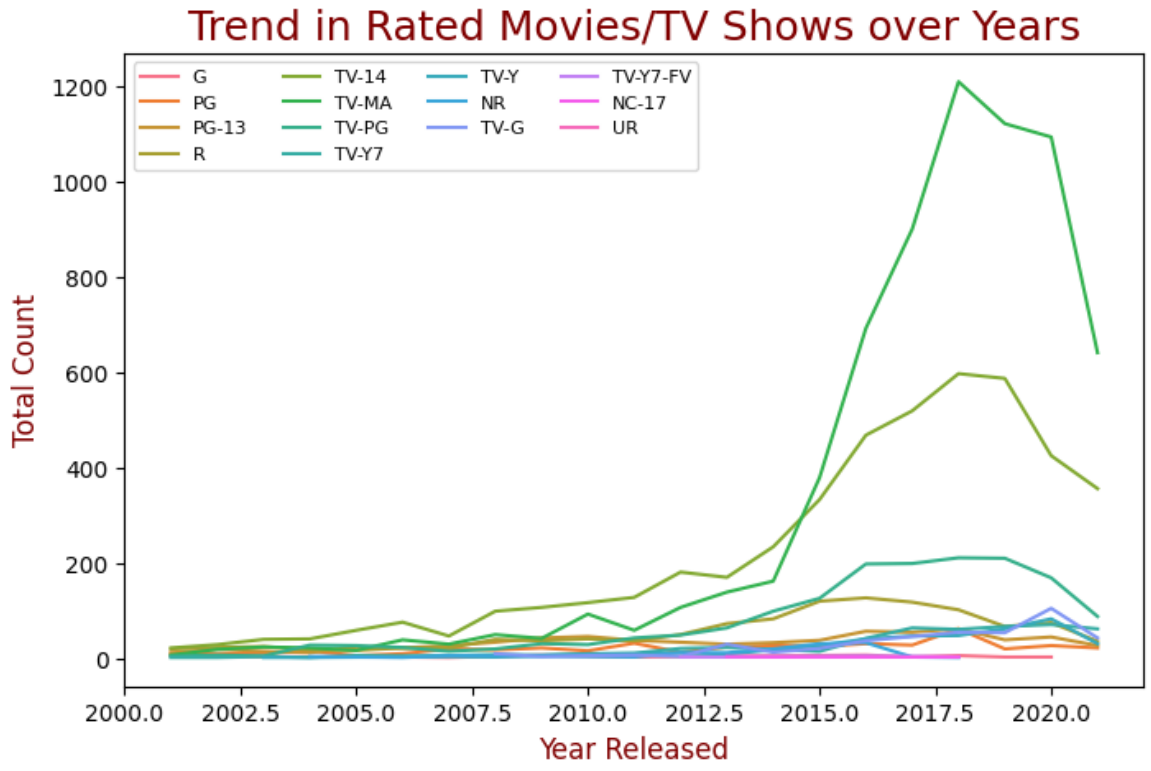

By analyzing the heatmap, we can identify strong positive correlations between specific genres. These correlations provide insights into viewer preferences and content interconnections.

## Movies/TV Shows rating comparison over the years

```
dfb=df.loc[:,['title','rating']].drop_duplicates()
dfb=df.groupby(['release_year','rating']).size().reset_index(name='TotalCount')
plt.figure(figsize=(8,3))
sns.countplot(x='rating',data=df6)
plt.title('Top Ratings',color='maroon',fontsize=18)
plt.xlabel('Rating',color='maroon',fontsize=14)
plt.ylabel('Total Count',color='maroon',fontsize=14)
plt.show()
```




```
plt.figure(figsize=(8,6))
sns.lineplot(x='release_year',y='TotalCount',hue='rating',data=dfb[(dfb['release_year']>2000)])
plt.title('Trend in Rated Movies/TV Shows over Years',color='maroon',fontsize=18)
plt.xlabel('Year Released',color='maroon',fontsize=12)
plt.ylabel('Total Count',color='maroon',fontsize=12)
plt.legend(bbox_to_anchor=(0, 1), loc='upper left', fontsize= 8,ncol=4)
plt.show()
```

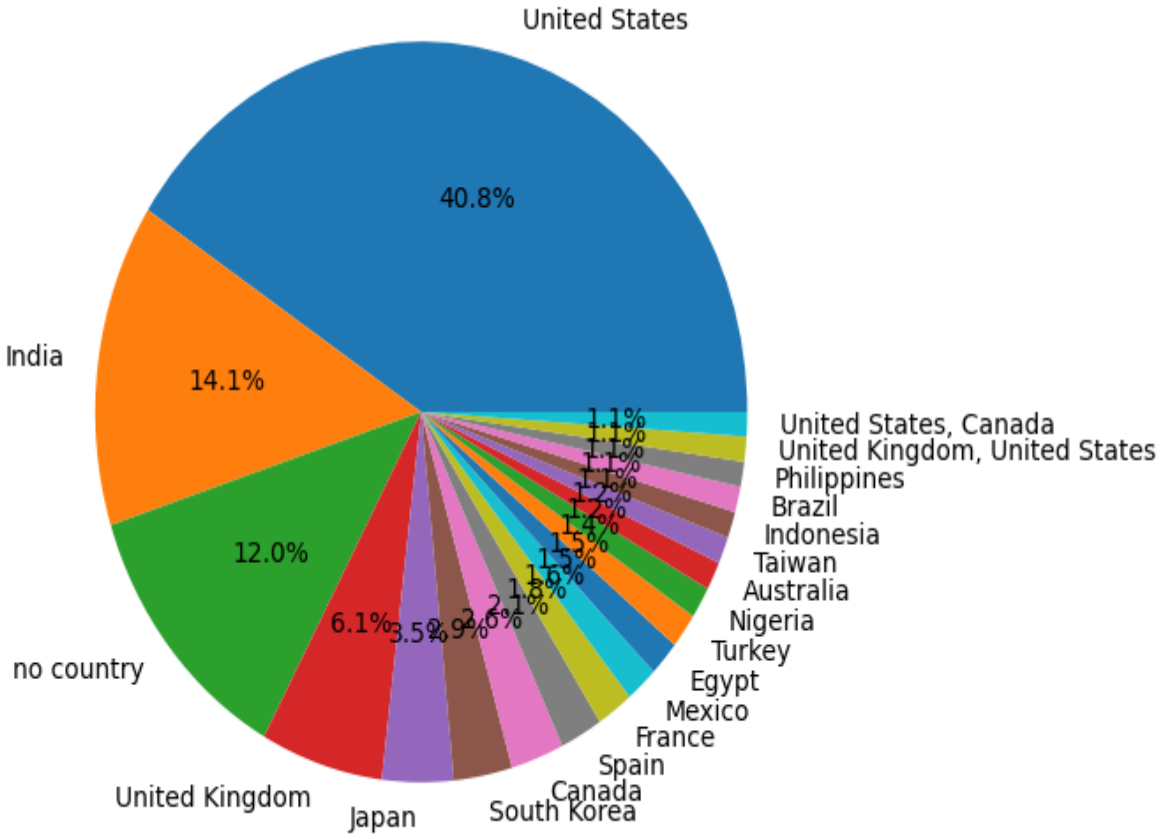



From above graph, TV-MA rated contents are more produced by Netflix over the recent years. We can see that a sharp rise in TV-MA , TV-14 rated content after 2013, indicating Netflix producing more content for Adult audiences and substantial rise in TV-PG rated content after 2017 indicating Netflix wants trying to capture other section of audiences along with adults.

## Content distribution in country

```
[26] dfc=df['country'].value_counts().reset_index().rename(columns={'index':'Country','country':'Total Count'})
```

```
plt.figure(figsize=(10,6))
plt.pie('Total Count',labels='Country',autopct='%2.1f%%',data=dfc.head(20))
plt.show
```




From the above graph, we could see most number of movies being produced in United States (40.8%) and next most dominant market being India(14.1%). Netflix has less content in rest of the countries.

## Recommendations

As we know that Netflix has established brand name and successful market strategies. In this paper we have discussed regarding business model of Netflix. There could be certain changes which can be applied for significant growth of the platform-

1. Netflix should produce more content for wider audiences including audience in the range 12 to 17.
2. Netflix should focus on TV Shows also because there are people who would like to see tv shows rather than movies.
3. Netflix needs to give priority to other genres like horror, comedy etc.
4. With most seasons being 2 in average, it can be averaged to 5 and more in years with more quality content.
5. It should also focus on increasing average duration of movies to 2 hours with good content.
6. By approaching the top director, we can plan some more movies/tv shows in order to increase the popularity.
7. Mainly the release in platform should focus on the festival holidays, year end and weekends which is to be mainly focused.
8. Netflix should start focusing other countries producing multilingual and regional content.
9. Contents in the country which has very less movies released should be increased and attract people of that country by making their native TV Shows.
10. In 2019, Most of the movies released in platform so we need to go on increasing this value to attract people by showing that getting subscription is useful as Netflix is releasing more movies per year.
11. 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 in order to get good content Netflix can reach to them and Netflix can produce the movie and give the director a chance.
12. In TV Shows we may focus on thriller genre which will be helpful for having more no of seasons.