Business Case: Netflix - Data Exploration and Visualisation

Business Case: Netflix - Data Exploration and Visualisation

Analyse 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.

1. Defining Problem Statement and Analysing basic metrics

The business case envisages analysing of the given Netflix dataset and using different methods like visual and non-visual analysis formulate insights which will help Netflix in decision making. The business case leverages Python's robust data analytics and visualization capabilities to extract valuable insights from the data set, viewing patterns, and content performance metrics. By harnessing Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn, the case aims to gain a comprehensive understanding of user preferences, engagement trends, and market dynamics. Through data-driven analysis and visualization techniques, the case tries to optimize content recommendations to cater to diverse audience segments. This data-centric approach empowers Netflix to make informed decisions, drive subscriber retention and growth, and maintain a leading position in the ever-evolving streaming industry landscape.

The data set have the following columns:

1 Show_id : Unique ID for every Movie / Tv Show

2 Type : Identifier - A Movie or TV Show

3 Title : Title of the Movie / Tv Show

4 Director : Director of the Movie

5 Cast : Actors involved in the movie/show

6 Country : Country where the movie/show was produced

7 Date added : Date it was added on Netflix

8 Release year : Actual Release year of the movie/show

9 Rating : TV Rating of the movie/show

10 Duration : Total Duration - in minutes or number of seasons

11 Listed in : Genre

12 Description : The summary description

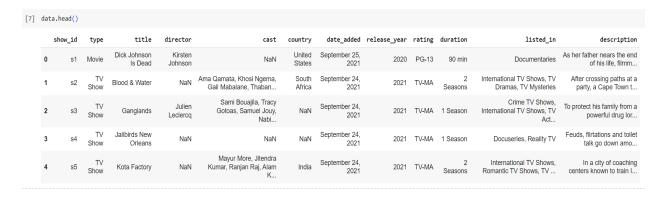
The data set is downloaded as 'netflix.csv' and saved as dataframe named 'data'.

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

!gdown https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv

Downloading...
From: https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv
To: /content/netflix.csv
100% 3.40M/3.40M [00:00<00:00, 27.8MB/s]</pre>
[3] data=pd.read_csv('netflix.csv')
```

The data sample is observed by data.head()



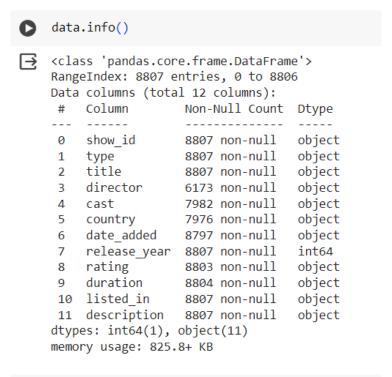
The data about the shows is divided into 12 columns and there are 8807 rows in the dataset.

2. Observations on the shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), missing value detection, statistical summary

Shape of the dataframe: data.shape showed

```
[13] data.shape
(8807, 12)
```

The basic information about dataframe. data.info()



Data type of 11 of the 12 columns are object type, only release year being int64

Detecting missing values by isna().

4307

```
data.isna().sum()
                         0
     show id
                         0
     type
     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
[18] data.isna().sum().sum()
```

The columns 'director' has most missing values of 2634 in the total 8807 rows. 'cast', 'country','date_added','rating' and 'duration' have missing values. Total number of missing values in the dataset is 4307.

The missing values were replaced with appropriate substitutes.

E.g.: data['director'].fillna('Unknown director',inplace=True)

```
[24] data['director'].fillna('Unknown Director',inplace=True) ?

data['cast'].fillna('Unknown Cast',inplace=True)
 data
```

The dataset refers to different shows from 1925 to 2021

```
[5] print(data['release_year'].max())
print(data['release_year'].min())

2021
1925
```

The shows are rated into 17 different ratings.

```
[8] data['rating'].nunique()
17
```

3. Non-Graphical Analysis: Value counts and unique attributes.

The data is subdivided into two dataframes named 'movies' and 'TVS' which stores the data of 6131 movies and 2676 TV shows respectively.

```
[16] movies=data[data['type']=='Movie']
[18] TVS=data[data['type']=='TV Show']
```

Analysis on Movies

Popular Director

```
movies.explode('director').groupby('director')['show_id'].count().sort_values(ascending=False).head(7)

director
Unknown Director 188
Rajiv Chilaka 22
Raúl Campos 18
Jan Suter 18
Suhas Kadav 16
Jay Karas 15
Marcus Raboy 15
Name: show_id, dtype: int64
```

The value counts per director shows that the greatest number of movies have missing values on director field. Apart from that, the greatest number of movies are by Rajiv Chilaka.

• Popular rating of movies

Most popular movie rating is TV-MA having 2062 movies which is meant for mature audience. Not very behind is TV-14 rating with 1427 movies.

Popular location

For analysing the popular location where the movie is made from, we need to un-nest the values in the country field. i.e., there are several entries where the movie is shown to be from multiple locations. It has to be converted into rows in which there are single entry per row. For that <u>pd.explode</u> is used.

```
[41] movies.explode('country').groupby('country')['show_id'].count().sort_values(ascending=False).head(7)
     country
     United States
                        2364
     India
                         927
     Unknown Country
                         440
     United States
                         388
     United Kingdom
                         382
     Canada
                         187
      France
                         155
     Name: show_id, dtype: int64
```

Most number of movies are made in United States, with 2364 movies majorly from USA and another 388 movies partly from USA. The second most popular destination is India.

• Popular Genre

Similarly, the genre of the film, shown in the column 'listed_in' also have multiple values in single row. Here also, explode function is used before finding the value count.

The most popular genre is 'Dramas' (1600+827=2427). The genre of 'International Movie' with having almost the same count is following and just behind them are 'Comedies'.

Analysis on TV Shows

• Popular Director

```
TVS.groupby('director')['show_id'].aggregate('count').sort_values(ascending=False).head(6)

director
Unknown Director 2446
Alastair Fothergill 3
Rob Seidenglanz 2
Shin Won-ho 2
Iginio Straffi 2
Hsu Fu-chun 2
Name: show id, dtype: int64
```

The value counts per director shows that the greatest number of TV Shows have missing values on director field. It was unable to find out relation of directors and the shows from the data set.

• Popular rating of TV Shows

Most popular TV show rating is TV-MA having 1145 shows which is meant for mature audience. Following behind is TV-14 rating with 733 shows.

Popular location

For analysing the popular location where the Show is made from, we need to un-nest the values in the country field. i.e., there are several entries where the TV Show is shown to be from multiple locations. It has to be converted into rows in which there are single entry per row. For that <u>pd.explode</u> is used.

```
[47] TVS.explode('country').groupby('country')['show_id'].count().sort_values(ascending=False).head(7)
     country
     United States
                        847
     Unknown Country
     United Kingdom
                        246
     Japan
                        174
     South Korea
                        164
     United States
                         91
     Canada
                         84
     Name: show id, dtype: int64
```

Most number of TV Shows are made in United States, with 847 shows majorly from USA and another 91 shows partly from USA. The second most popular destination is UK.

• Popular Genre

Similarly, the genre of the TV show, shown in the column 'listed_in' also have multiple values in single row. Here also, explode function is used before finding the value count.

```
TVS.explode('listed_in').groupby('listed_in')['show_id'].count().sort_values(ascending=False).head(7)
listed_in
International TV Shows
                           774
 TV Dramas
                           696
International TV Shows
TV Comedies
                           461
Crime TV Shows
                           399
Kids' TV
                           388
Romantic TV Shows
                           338
Name: show_id, dtype: int64
```

The most popular genre is 'International TV show' with 1351(774+577) shows. 'TV Dramas' is following with 696 shows and just behind them are 'TV Comedies'.

• Movies and TV Shows genres popular in different months of a year.

The data is processed and classified on the basis of the genres in 'listed_in' column and month from date added.

The genres are added as a string of values, these are converted to list by using 'apply' and 'split' functions. And then using 'explode' function, the genres are separated. Still there were problem of mismatch due to trailing empty spaces which is solved by using 'strip' function. Now, using 'transform' and 'max' function, the maximum for easch genre is found and matched with the count value.

```
[67] movies['added_month']=movies['date_added_datetime'].dt.month
    movies_new=movies.explode('listed_in')
    mmg=movies_new.groupby(['added_month','listed_in'])[['show_id']].aggregate('count')
    mmg.reset_index(inplace=True)
    mmg['listed_in']=mmg['listed_in'].str.strip()
    mmg['Count'] = mmg.groupby(['listed_in'])['show_id'].transform('max')
```

	added_month	listed_in	show_id	Count
102	4	International Movies	254	254
213	7	Dramas	158	158
113	4	Comedies	127	127
272	9	Action & Adventure	93	93
115	4	Documentaries	78	78
198	7	Independent Movies	74	74
336	11	Children & Family Movies	64	64
202	7	Romantic Movies	60	60
297	10	Romantic Movies	60	60
206	7	Thrillers	51	51

```
TVS['added_month']=TVS['date_added_datetime'].dt.month
TVS_new=TVS.explode('listed_in')
TVG=TVS_new.groupby(['added_month','listed_in'])[['show_id']].aggregate('count')
TVG.reset_index(inplace=True)
TVG['listed_in']=TVG['listed_in'].str.strip()
TVG['Count'] = TVG.groupby(['listed_in'])['show_id'].transform('max')
```

```
[70] TVG[TVG['show_id']==TVG['Count']].sort_values(by='Count',ascending=False)
```

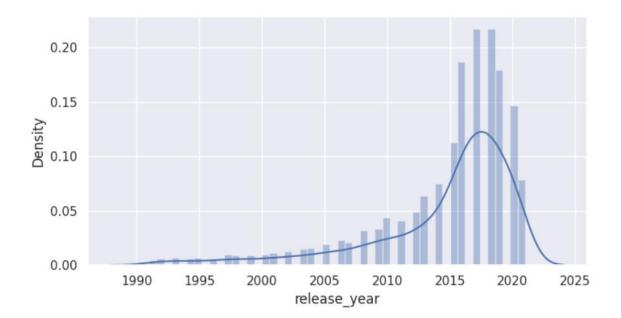
	added_month	listed_in	show_id	Count
383	12.0	International TV Shows	86	86
173	6.0	TV Dramas	72	72
372	12.0	TV Dramas	72	72
206	7.0	TV Dramas	72	72
371	12.0	TV Comedies	55	55
366	12.0	Romantic TV Shows	45	45
24	1.0	Kids' TV	44	44
182	6.0	Crime TV Shows	41	41
249	8.0	Crime TV Shows	41	41
85	3.0	British TV Shows	36	36
282	9.0	Docuseries	32	32

- 4. Visual Analysis Univariate, Bivariate after pre-processing of the data Note: Pre-processing involves unnesting of the data in columns like Actor, Director, Country
 - i. For continuous variable(s): Distplot, countplot, histogram for univariate analysis

For visual analysis based on release year, the shows after 1990 is taken since for the time period before 1990, much fewer shows are released.

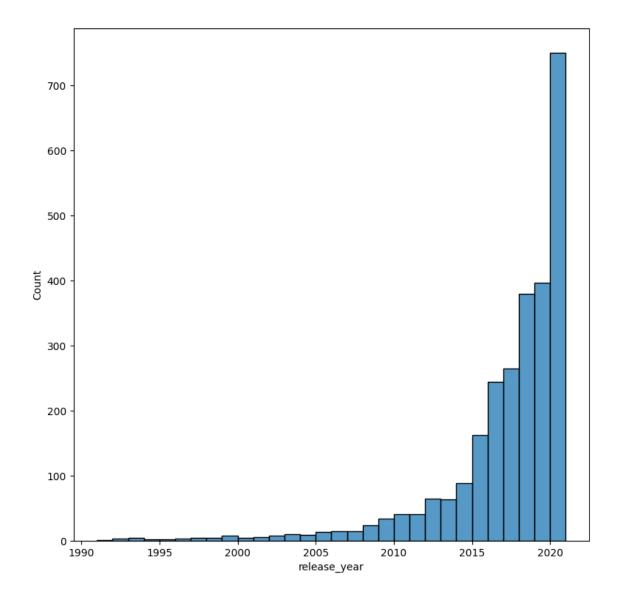
Distplot of movies released over years.

```
[ ] sns.set(rc={"figure.figsize": (8, 4)})
sns.distplot(movies[movies['release_year']>1990]['release_year'])
```



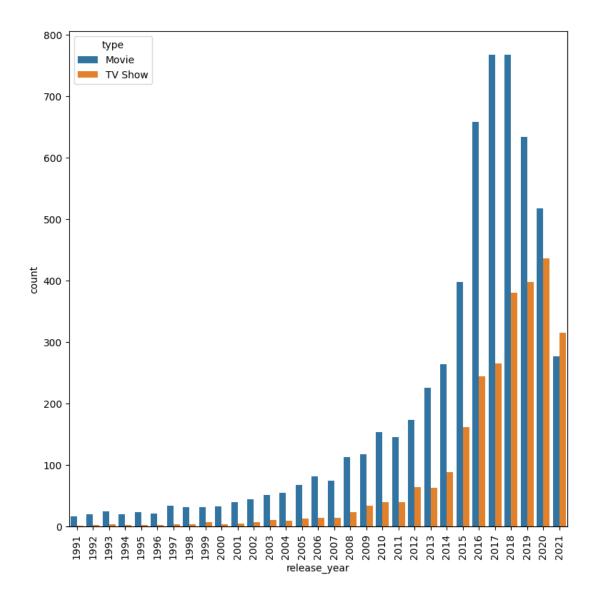
Histogram for TV shows released over the years:

```
plt.figure(figsize = (15,15))
sns.histplot(TVS[TVS['release_year']>1990]['release_year'],bins=30)
```



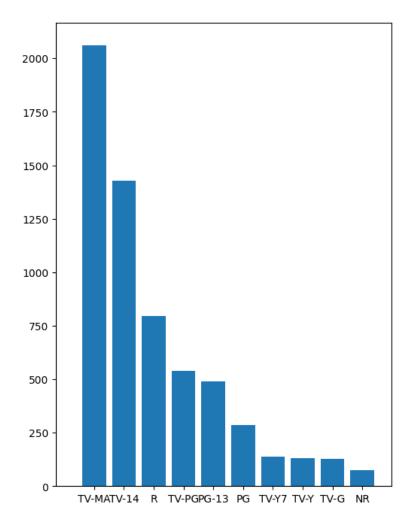
Count plot featuring movies and TV shows released over years:

```
[11] plt.figure(figsize = (9,9))
    sns.countplot(data['release_year']>1990],x='release_year',hue='type')
    plt.xticks(rotation=90)
```

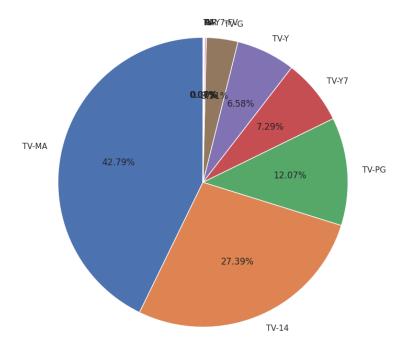


Bar chart showing the distribution of movies over different ratings

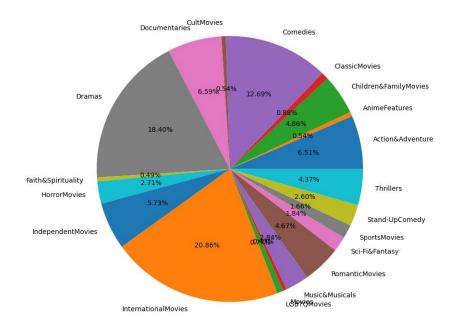
```
[33] plt.figure(figsize = (5,8))
    plt.bar(rating_count.index,rating_count)
```



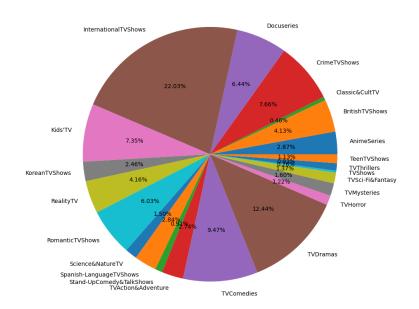
Pie chart showing distribution of TV shows over different ratings



Pie chart for Movies and TV shows for different genre



```
[17]
    TVS_genre=TVS['listed_in'].explode().reset_index()
    TVS_genre['listed_in']=TVS_genre['listed_in'].str.replace(' ','')
```

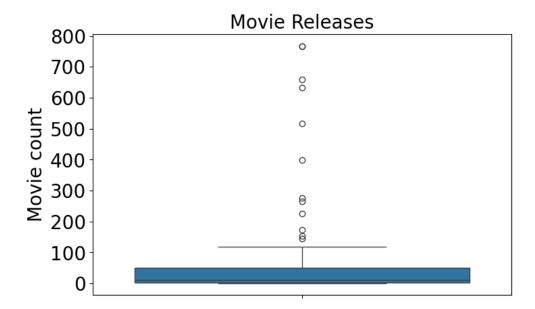


ii. For categorical variable(s): Boxplot

The data of movies pertaining to different years has been plotted as a box plot.

```
[27] mbox=movies.groupby('release_year')['show_id'].count().reset_index()
```

```
[38] plt.figure(figsize=(8,5))
    sns.boxplot(y = mbox["show_id"])
    plt.yticks(fontsize=20)
    plt.ylabel('Movie count', fontsize=20)
    plt.title('Movie Releases', fontsize=20)
```

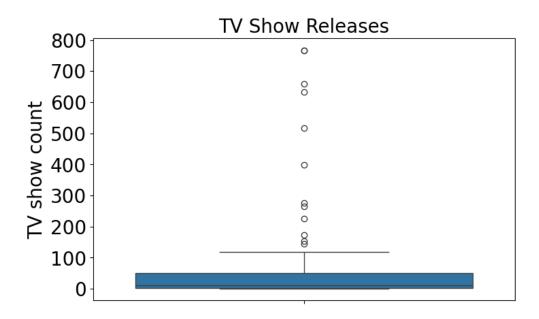


Here, the movies in the data set have been from 1930s to 2021. In this large range, median of number of movies per year comes to a low value as compared to the total number. This is because of the large time range. All years except near to the last decade have a low frequency of movies per year. Most of the valuable data is found to be outliers in the box plot.

Same trend was observed in the case of TV shows also.

```
[36] TVbox=TVS.groupby('release_year')['show_id'].count().reset_index()

[39] plt.figure(figsize=(8,5))
    sns.boxplot(y = mbox["show_id"])
    plt.yticks(fontsize=20)
    plt.ylabel('TV show count', fontsize=20)
    plt.title('TV Show Releases', fontsize=20)
```

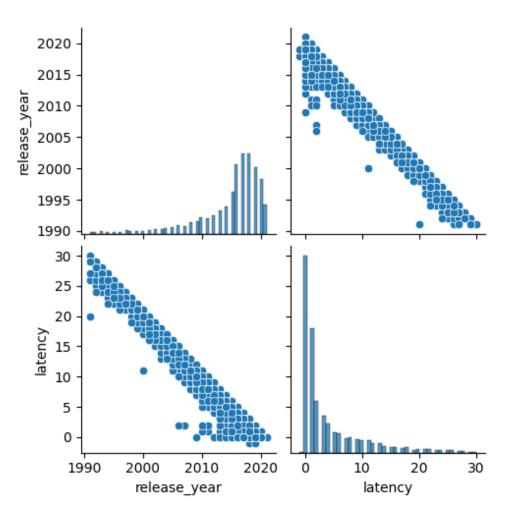


iii. For correlation: Heatmaps, Pairplots

The latency in adding a show to Netflix is calculated from the release year and year from date added.

Pair plot between year released and latency

sns.pairplot(movies[movies['release_year']>1990][['release_year','latency']])

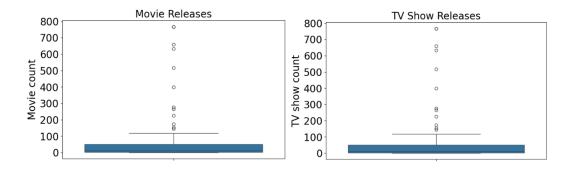


5. Missing Value & Outlier check

Missing values have been taken care of by using 'fillna()' function.

```
data['cast'].fillna('Unknown Cast',inplace=True)
data['director'].fillna('Unknown Director',inplace=True)
data['country'].fillna('Unknown Country',inplace=True)
data['rating'].fillna('NA',inplace=True)
data['duration'].fillna('NA',inplace=True)
```

In the box plot drawn using count of movies over the years, outliers have been spotted.



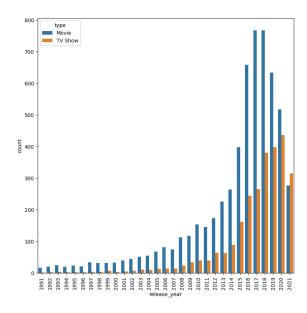
6. Insights based on Non-Graphical and Visual Analysis

i. Comments on the range of attributes

- The data contains 8807 entries corresponding to different Movies and TV shows.
- There are 12 different attributes corresponding to each show.
- Among them 6131 records correspond to movies and 2676 records are about TV shows.
- The data contains details about shows released from 1925 to 2021.
- The dataset had 4307 missing values. Predominantly in the 'director' column.
- The most popular location is observed to be USA, followed by India.
- Most popular genre in movies is 'Dramas' followed by 'International Movies', 'Comedy Movies' and 'Action & Adventure' movies.
- Most popular genre in TV shows is 'International TV shows' followed by 'TV Drama's and 'TV comedies.'
- In the case of both movies and TV shows, the greatest number of shows are rated 'TV-MA'.

ii. Comments on the distribution of the variables and relationship between them

• The distribution of shows released over the years has been observed.



The distribution shows an exponential increase in the number of shows released over the last two decades, especially after 2010. When analysed for movies and TV shows separately, the number of movies shows a peak around mid-2010s and the TV shows appear to be increasing except 2021, since the data regarding 2021 is not complete, it can be inferred that number of TV shows is still going to a peak.

• The date of addition of movies to Netflix and the number of movies added in different months are analysed.

	added_month	listed_in	show_id	Count
102	4	International Movies	254	254
213	7	Dramas	158	158
113	4	Comedies	127	127
272	9	Action & Adventure	93	93
115	4	Documentaries	78	78
198	7	Independent Movies	74	74
336	11	Children & Family Movies	64	64
202	7	Romantic Movies	60	60
297	10	Romantic Movies	60	60
206	7	Thrillers	51	51

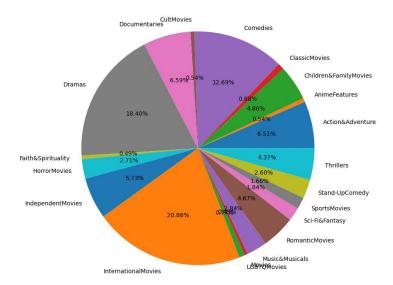
The peak number of different genre of movies added to Netflix is in the table above. The months with more peak values are July and April. But Children & Family movies are found to be added most in November. It denotes the relation between release of movies in Netflix and holidays.

Similarly, the addition of TV shows is analysed across different months.

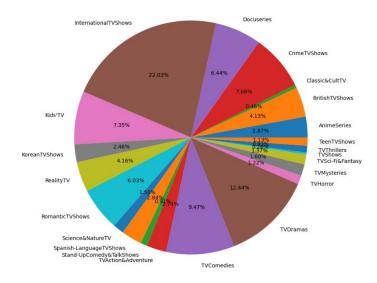
	added_month	listed_in	show_id	Count
383	12.0	International TV Shows	86	86
173	6.0	TV Dramas	72	72
372	12.0	TV Dramas	72	72
206	7.0	TV Dramas	72	72
371	12.0	TV Comedies	55	55
366	12.0	Romantic TV Shows	45	45
24	1.0	Kids' TV	44	44
182	6.0	Crime TV Shows	41	41
249	8.0	Crime TV Shows	41	41
85	3.0	British TV Shows	36	36
282	9.0	Docuseries	32	32

Here the release of shows in Netflix show a prominent trend over different genres releasing maximum number of shows in December.

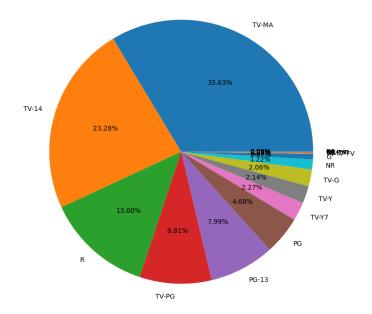
 The distribution of movies over genres shows that more than 50% of all shows are belonging to International Movies, Dramas or Comedies. Other major genres are Action & Adventure, Documentaries, Children& Family Movies and Independent movies.



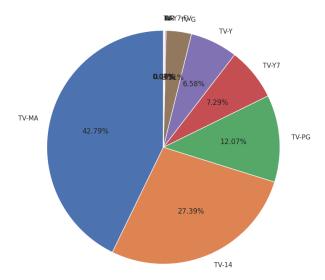
• Genre-wise analysis of TV shows denote that more than 60% of TV shows are International TV shows, TV Dramas or RV Comedies. Docuseries, Kids TV and Crime TV shows are other popular genres.



• The movies are distributed over different rating on the basis of the audience allowed to watch them. The pie diagram of movies across ratings shows that almost 70% of all the movies are TV-MA, TV-14 or R rating.



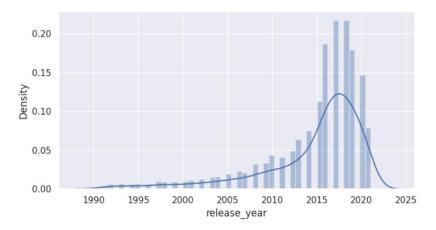
• TV shows across different ratings shows the following distribution



Here also about 70% of all shows comes under only two ratings – TV-MA and TV-14.

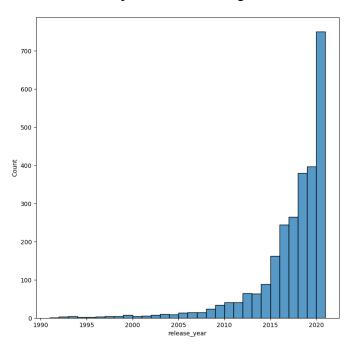
iii. Comments for each univariate and bivariate plot

• The dist-plot of movie numbers across years show a varying trend which peaks around mid-2010s and then starting to decrease



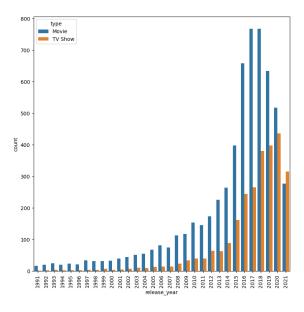
This plot is drawn avoiding the pre 90s movies which were observed to be outliers in the basis of frequency of movies per year.

• The release of TV shows is depicted in the histogram.



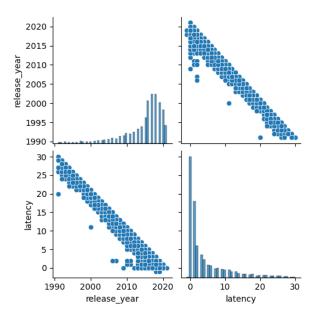
The growing trend of TV shows is exponential and is observed to be continuing still.

 The trends of number of releases of movies and TV shows across the years can be observed in the count plot below.



It clearly shows the difference in frequency of movies and TV shows and as well as the growing trend of both. TV shows were less compared to number of movies in yearly 90s. But, number of TV shows show a rapid growth through the 2010s and by 2021, the number of TV shows is more than the number of Movies.

• The latency of shows being added to Netflix is found by the difference in year added and year released. The year released and latency corelation is shown in the following pair plot.



The plot shows that the movies which are released after 2010 has started to show zero latency i.e., those are added to Netflix on the same year they got released.

7. Business Insights

- There is a high rate of increase in number of shows added in Netflix during the years following 2010. During these years there were high advancements in mobile technology and smart phones became cheaper and common. Common people have opted to use smart phones and in turn opted to be viewers of the platform in hand, rather than other media like movie theatre or home theatre. Also the emerging technologies of Smart TVs also must have helped this growth. This is a promising trend for a platform like Netflix.
- The increase in addition of shows to Netflix shows different trends for movies and TV shows. The number of movies were much larger as compared to the number of TV shows in early 2010s. This might be because back then the smart phone technology wasn't common and the TV shows which were popular were those few ones which got popular through TV itself. Later on, due to the emergence of platforms like Netflix, there has been an increase in production of TV shows to grab the viewers via these new platforms. By 2021, number of TV shows released surpassed the number of movies.
- Among movies, in analysis across different genres, 'International Movies' are largest in number. 'Dramas' and 'Comedies' have also great number of movies and these three themselves constitute the 50% of all movies. 'Anime features' and 'Faith and Spirituality' are the least preferred genre of movies among all. In the medium range, 'Action and Adventure' films and 'Documentaries' are equally preferred. 'Romantic movies', 'Childrens and family' movies and 'Independent movies' combinedly constitute 18-20% of the total number of movies and thus seems significant.
- Similar trend is observed in case of TV shows as well. 'International TV shows' are the most prominent in TV shows. 'TV Comedies' follows with a significant fan base. 'Crime TV shows' and 'Docuseries' and 'Romantic TV shows' constitute about 30% of all the TV shows which is an interestingly different trend as compared to movies. 'Classic cult TV' is least preferred TV show genre.
- The months with more movies added to Netflix are July and April. But Children & Family movies are found to be added most in November. It denotes the relation between release of movies in Netflix and holidays. In case of TV shows, the release of shows in Netflix show a prominent trend over different genres releasing maximum number of shows in December. These trends indicate that there is a strong corelation between addition of shows to Netflix and the holiday-vacation months. The viewers will be available and will be looking for content to watch during these seasons. This can be effectively utilised to increase visibility of different genres of shows.

8. Recommendations.

- The growth pattern of movies and TV shows are showing a promising development over the past years. A platform like Netflix could really make use of the trend and try to expand their viewer base. Both movies and TV shows are finding demand in the phase of transition of viewers from big screen to smart phones and smart TVs.
- Another interesting trend is the growth of TV shows surpassing the growth of movies released in a year. This trend is in development for the last few years under analysis. Netflix can shift its priority towards producing more TV shows which is up to the taste of the viewers to make use of this high demand.
- In case of movies, large scale movies with internationally well-known directors and cast are observed to be most welcomed by the viewers. 'International Movies' are a top priority for future productions. But strong emphasis should be given to other genres like 'Dramas' and 'Comedies' which attracts large number of film enthusiasts to the platform.
- International TV shows is itself the highest priority in case of TV shows. 'TV Dramas' are attracting ample audience and should be taken with greater importance. Other genres like 'Crime TV shows' and 'Docuseries' have the potency to attract a good number of viewers too.
- Time of addition of shows to the platform shall be chosen wisely considering the viewers availability and general spirit. We can make use of the vacation and holiday times to push more contents since then the viewers will be available and expecting entertainment. Family movies find it optimum to be added before holidays where thrillers comedies and other genres find the vacations to be the right time to release in the platform.
- Similarly, most of the TV shows are received with much enthusiasm when released in the platform near to the holiday month of December.