**Netflix - Data Exploration and Visualization**

Start by exploring a few questions: What type of content is available in different countries?

1. How has the number of movies released per year changed over the last 20-30 years?

**A graph of a number of data

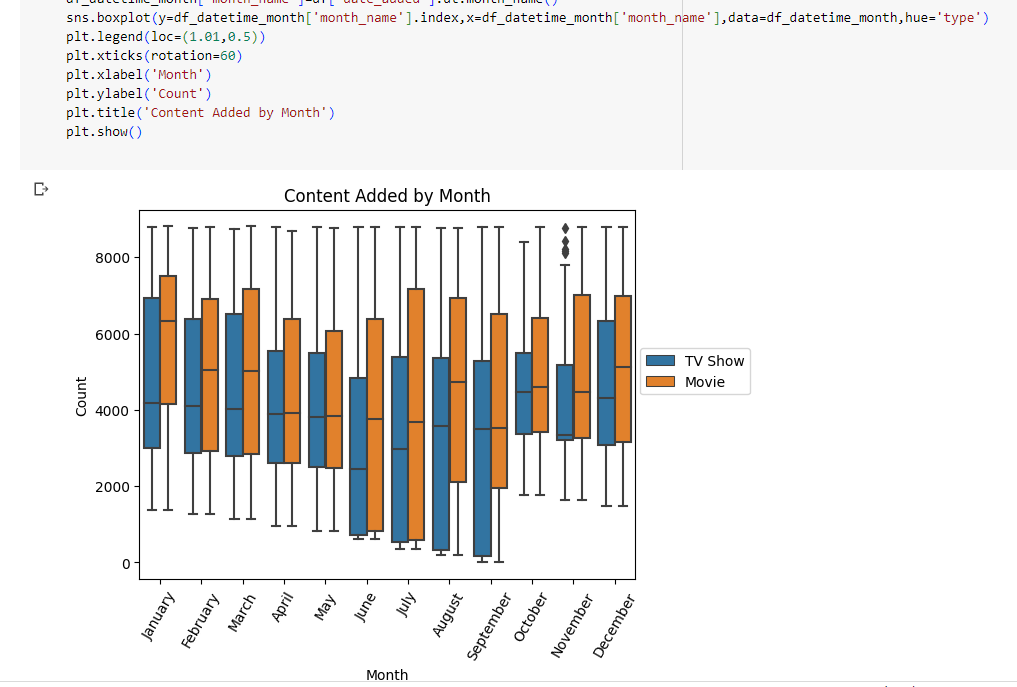
Description automatically generated with medium confidence**

2.Comparison of tv shows vs. movies.

A screen shot of a graph

Description automatically generated

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



5.Does Netflix has more focus on TV Shows than movies in recent years

A graph with blue and orange bars

Description automatically generated

Conclusion: There is deep decline in TV shows production due to pandemic .

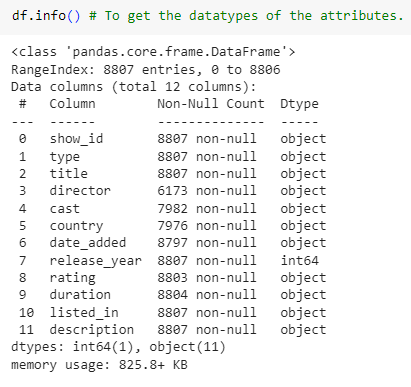
6.Understanding what content is available in different countries

A screen shot of a computer

Description automatically generated

1. Defining Problem Statement and Analyzing basic metrics

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.



A close up of a number

Description automatically generated

Shape of a dataframe is 8807

A close-up of a computer screen

Description automatically generated

df.colomns—Gives the columns names.

df.info()🡪provides how many null values and datatypes.

Column Duration has 2 values 1. Minutes 2. Seasons

Few of the Direcot,actor and country data is missing.

Date\_added column is defined as categorical object which should be datetime object.

Few of the values wrongly entered for the column’s duration and rating.

1. **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**

**A close-up of a red line

Description automatically generated**

Netflix dataset has 8807 rows and 12 attributes.

**Datatypes**

A screenshot of a computer

Description automatically generated

**Conclusion**

We have observed 2 datatypes ‘object’ and ‘int’ in the dataset.

The column “date\_added” Dtype at present is object and this should needs to be categorized as Datetime.

Observed more missing values in director, cast and country when compared to other attributes.

**Statistical Summary**

A screenshot of a computer

Description automatically generated

**Conclusion:**

25% of total data belongs to the year 1925-2012.

25% of total data belongs to the year 2019-2021

Insight:- Netflix should add more latest movies to attract more no of customers.

A screenshot of a computer

Description automatically generated

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 maximum content available.

**Missing Value Detection**

A screenshot of a computer

Description automatically generated

**Conclusion:**

Lots of missing data found in director, cast and country columns as compared to others.

Percentage contribution of missing values:

A screenshot of a computer code

Description automatically generated

**Conclusion:**

Director share is 29% of null values followed by country and cast and we cannot drop this data which may cause inconsistency. So we need to do fill the NULL values for director, cast and country with No data.

**Handling Missing Values**

Replace blank countries with the most common country.

Replace director name and cast with no data .

A screenshot of a computer program

Description automatically generated

Drop invalid and duplicate data:

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

Now the dataset is cleaned with missing values ,invalid and duplicates.

**Category Change:**

Date\_added dtype should not be a object(string) . We need to convert object dtype to datetime dtype.

A group of colorful text boxes

Description automatically generated with medium confidence

**Additional Data Cleaning**

A screenshot of a computer

Description automatically generated

We observed duration values as NaN for the director “Louis C K “ and the values in ratings should in duration. This may be due to input error .Now we need to copy the rating values to duration and make a rating column as No data for the director Louis C K.

Check to make sure there is no other content with the same director to avoid accidental overwriting.

A screenshot of a computer

Description automatically generated

Since there is no extra data , we can start overwriting the duration .

A screenshot of a computer

Description automatically generated

Now we have overwritten the ratings and duration and requirement is fulfilled for director Louis C.K.

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

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

**Conclusion :**

* Anurag Kashyap and David Dhawan are the most famous directors for India.
* Jay Karas and Marcus Raboyare, the most famous directors in United States.

**Top 2 directors**

A screenshot of a computer code

Description automatically generated

**Conclusion:**

Rajiv Chilaka is the most famous director among all followed by Jan Suter.

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

**Conclusion :-**

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

A screen shot of a computer

Description automatically generated

**Conclusion:**

Anupam Kher is the topmost actor among all followed by David Attenborough.

What type of shows that have been watched on Netflix

Column ‘type’ Analysis :

A white rectangular box with green text

Description automatically generated with medium confidence

Conclusion:

Two types of shows being produced on NetScaler i.e., Movies and TV show. Movies content has higher number when compared to TV Show.

Country Column Analysis:

A screenshot of a computer

Description automatically generated

Conclusion:

We have gathered the top 10 countries where Netflix provided content(both movies & TV Shows). United States tops for Netflix content followed by India and United Kingdom.

Check the type of content based on Country:

Netflix Movie content produced to Countries:

A screenshot of a computer

Description automatically generated

Conclusion:

The highest Netflix movie content country is United States followed by India.

Netflix TV shows content produced to Countries:

A screenshot of a computer

Description automatically generated

The highest Netflix TV shows content country is United States followed by United Kingdom.

Content Ratings:

A screenshot of a computer

Description automatically generated

Conclusion:

TV-MA certified content has been most produced by Netflix followed by TV-14

Release year:

A screenshot of a computer

Description automatically generated

Conclusion:

Above data reveals that year 2018 has highest number of Netflix content releases done followed by Year 2017 and 2019.

Popular Genres Analysis:

A screen shot of a computer

Description automatically generated

Conclusion:

The most viewed Netflix content is Dramas ,International Movies followed by Documentaries and Stand-Up comedy.

**Summary:**

Netflix has more number of movies than TV Shows

Most number of movies produced by United States followed by India which is the 2nd most number of movies in Netflix.

Most of the content Movies & TV Shows are for Mature Audience.

2018 is the year where Netflix released a lot of content when compared to previous Years.

International Movies and movies are the most liked and most popular Genres on Netflix.

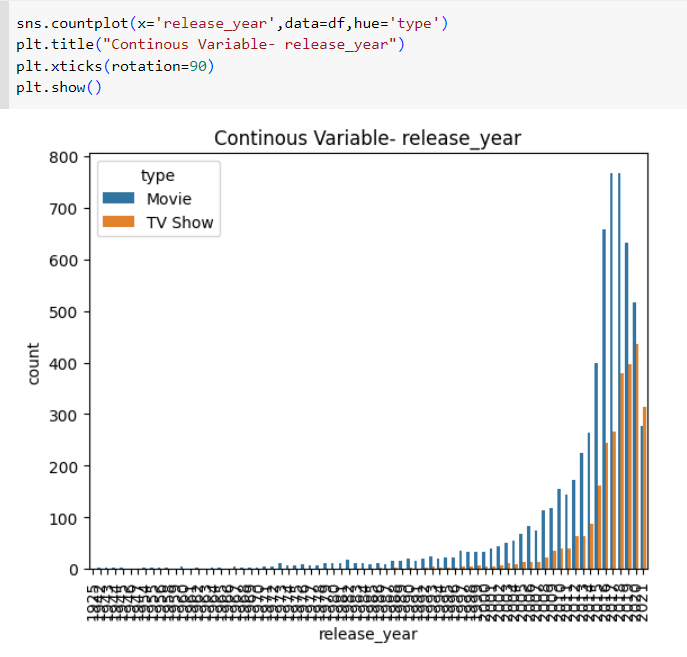
**4. Visual Analysis - Univariate, Bivariate after pre-processing of the data**

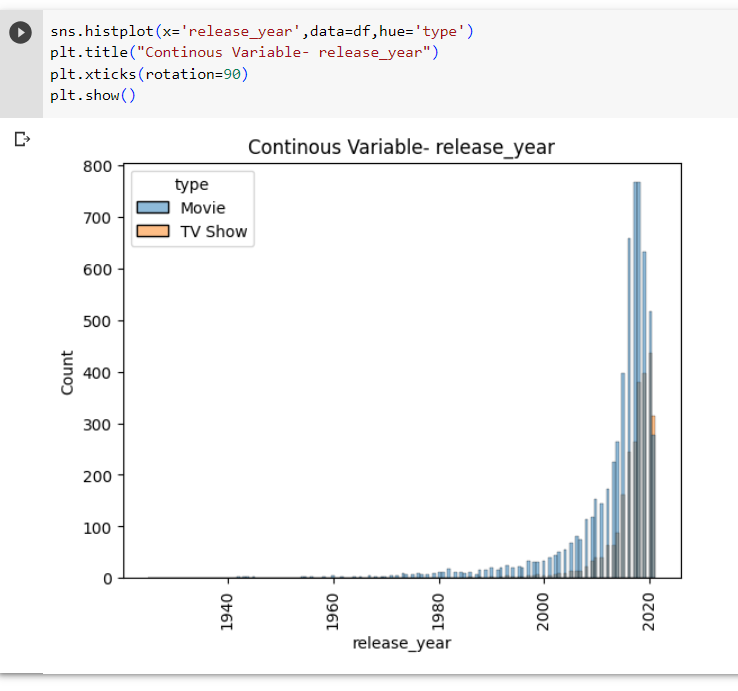
**Univariate:**

For continuous variable(s):

A graph of a number of movies

Description automatically generated with medium confidence





Conclusion:

From the above visualization, we can observe that there is a significant increase of Netflix content production from year 2018-2021.

Among movies and TV Shows, Movies has a seen a significant growth.

4.2 For categorical variable(s):

Duration:

A screenshot of a computer

Description automatically generated

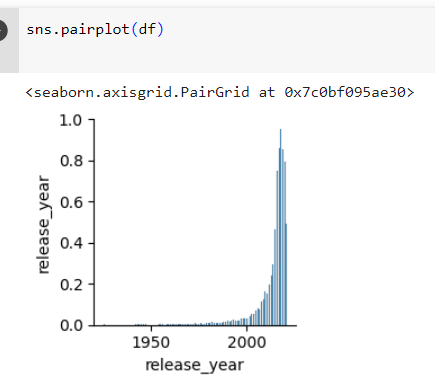
**Conclusion:**

Average duration of movies is 100 mins.

TV shows mostly are having 1 or 2 Seasons.

There are lot of outliers present in movies as compared to TV shows.

4.3 For correlation: Heatmaps, Pairplots



A screenshot of a computer

Description automatically generated

Bivariate

Analysis of number of contents added on to Netflix over the period

A screenshot of a computer screen

Description automatically generated

**Conclusion :-**\*

* July, August and September are the months when most content was added because no of TV shows during these 3 months are maximum among all.
* June, July and August are the months where most content added.
* No of movies added per month is greater than no of TV shows added per month.

A screenshot of a graph

Description automatically generated

* Conclusion:
* More no of movies were added during 2019 followed by 2018 and 2017.
* More no of TV shows added in 2018.
* We can see drop of count of movies from 2019 onwards due to pandemic.

A screenshot of a graph

Description automatically generated

**Conclusion:**

Most of the content is added on Netflix on "Friday" followed by Thursday as weekend approaches after these days.

A screenshot of a graph

Description automatically generated

Conclusion:

Mostly TV shows and movies are belongs to TV-MA & TV-14 rating.

Mostly content available on Netflix is for adults and teenagers.

A computer code with many colorful text

Description automatically generated with medium confidence A graph of a number of people

Description automatically generated

Conclusion:

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

A computer code with black text

Description automatically generated A colorful bar chart with numbers

Description automatically generated with medium confidence

Conclusion:

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

* International Movies
* Dramas
* Comedies
* International TV show

5. Missing Value & Outlier check (Treatment optional)

Missing Value Detection

A screenshot of a computer

Description automatically generated

Conclusion:

Lots of missing data found in director, cast and country columns as compared to others.

Percentage contribution of missing values:

A screenshot of a computer code

Description automatically generated

Conclusion:

Director share is 29% of null values followed by country and cast and we cannot drop this data which may cause inconsistency. So we need to do fill the NULL values for director, cast and country with No data.

Handling Missing Values

Replace blank countries with the most common country.

Replace director name and cast with no data .

A screenshot of a computer program

Description automatically generated

Drop invalid and duplicate data:

A screenshot of a computer code

Description automatically generated

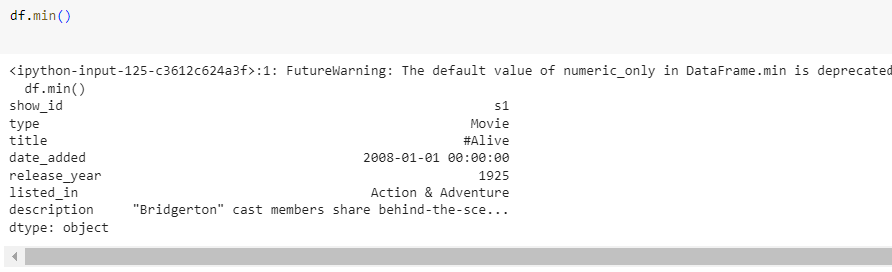
A screenshot of a computer

Description automatically generated

Now the dataset is cleaned with missing values ,invalid and duplicates.

6. Insights based on Non-Graphical and Visual Analysis

6.1 Comments on the range of attributes



A screenshot of a computer

Description automatically generated

Min() and Max() function selects the min and max value respectively from each attribute which is of any dtype.

6.2 Comments on the distribution of the variables and relationship between them

Every attribute is correlated and analyzing will be easy to find t.

For example, to find high rating in a year for movie or TV shows , we can establish relationship using replots.

A screen shot of a graph

Description automatically generated

6.3 Comments for each univariate and bivariate plot

* 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 available 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 movies are 100 min duration.
* Netflix should add more movies for United States and India falling in category of international 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 should add TV Show on Friday than any other weekday.
* As per 2021 data, count of TV shows 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.

7. Business Insights

* 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 available 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 compare to other countries.
* 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 shows as compare to movies.

8. Recommendations

*Movies* :-

* Preferred movies duration is between 90-100 minutes.
* Netflix should add more movies for United States and India falling in category of international 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 should add TV Show on Friday than any other weekday.

*TV Show*:-

* Preferred movies duration is 1-2 seasons.
* Netflix should focus on countries like Japan, South Korea and France in TV shows , as they prefer TV shows over movies.
* Netflix should add TV Show on Friday than other weekday.
* As per 2021 data, count of TV shows 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.