Colablink:

https://colab.research.google.com/drive/1t4cv8QG_ukpLmG_QeS-EFAaQ5ajajXMc?usp=sharing

1. Defining Problem Statement and Analysing basic metrics (10 Points)

PROBLEM STATEMENT:

To suggest which type of shows to produce and how to grow the business.

Basic Metics Analysis:

The dataset consists of 12 cloumns where

- **show_id:** gives the id of the particular show/movie
- **type**: gives the information about if it is a TV series/Movie.
- **title** : gives the title of the Series/movie
- **director**: gives the persons name who directed particular series/Movie
- **cast**: gives information about persons who worked for the series/Movie
- **country**: gives information about the countries where the series/Movie is available.
- **date_added**: gives the information about when the series/Movie is added to the netflix
- **release_year**: gives the information about the year of release of that series/Movie
- **rating**: gives the information about what is the consumer group that can watch that series/Movie(ex:kids, adults or both)
- **duration**: gives the duration of the series/Movie
- **listed_in**:gives the information about the Genre of the series/Movie
- **description**:gives the brief description about the story line about that series/Movie

Duration of data:

- The data set consists of data from (2008-01-01) to (2021-09-25) ie., 13.74 years of data.

Duplicates:

- As no.of unique records with respect to 'id' cloumn and length of data is 8807, we can conclude that there are no duplicates.

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 (10 Points)

- **Shape of data**: 8807 rows and 12 colums
- **Datatypes of all attributes:** (AT first)

```
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
```

After converting the datatype of 'date_added' column:

```
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   datetime64[ns]
 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
```

Missing DATA:

- The data consists of 8807 records and some missing values.

- For the directors column 29 % data is missing, for cast and country columns around 9% of data is missing.

Statistical summary:

	count	unique	top	freq	first	last
show_id	8807	8807	s1	1	NaT	NaT
type	8807	2	Movie	6131	NaT	NaT
title	8807	8807	Dick Johnson Is Dead	1	NaT	NaT
director	6173	4528	Rajiv Chilaka	19	NaT	NaT
cast	7982	7692	David Attenborough	19	NaT	NaT
country	7976	748	United States	2818	NaT	NaT
date_added	8797	1714	2020-01-01 00:00:00	110	2008-01-01	2021-09-25
rating	8803	17	TV-MA	3207	NaT	NaT
duration	8804	220	1 Season	1793	NaT	NaT
listed_in	8807	514	Dramas, International Movies	362	NaT	NaT
description	8807	8775	Paranormal activity at a lush, abandoned prope...	4	NaT	NaT

From the above statistics we can say:

The most Popular:

- Director is Rajiv Chilaka.
- Movie is Dick Johnson Is Dead
- Actor David Attenborough
- Movies have more popularity than TV series

3. Non-Graphical Analysis: Value counts and unique attributes (10 Points)

Movie 69.615079

TV Show 30.384921

- From the above information, it can be inferred that the data consisted almost 70% of movies and 30% of TV shows.
- United States has more no.of series/Movies
- Rajiv Chikla has directed more no.of movies/series
- Anupum kher has acted in more movies in the given dataset.
- In 2019, most no.of series/movies were added to Netflix, and in 2010 least no.of movies were added.
- Most no.of movies/series were added to Netflix on Fridays.
- Most no.of movies/series were added to Netflix in July month.

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

4.1 For continuous variable(s): Distplot, countplot, histogram for univariate analysis (10 Points)

Analysis of TV shows:

- There are more no.of shows with only 1 season.
- There are very less no.of shows with seasons greater than 4.

Analysis of Movies:

- Most of the movies have the duration in the range of 50mis to 150mins.
- Highest no.of movies have the duration around 100mins.

Analysis on Ratings:

- Almost all the ratings have more no.of movies than series.
- For TV-Y and TV-Y7 ratings , there are more no.of series than movies.

Analysis based on the release_year:

- There is rapid increase in the release of no.of movies/series from 1940 to 2018 and slight decrease towards 2020.

4.2 For categorical variable(s): Boxplot (10 Points)

- Most no.of movies belongs to international movie Genre(2624) which are followed by dramas,comidies, action and adventure.
- Most no.of TV Shows belong to TV shows Genre which are followed by TV horror, Classic and Cult Category

4.3 For correlation: Heatmaps, Pairplots (10 Points)

- From the pairplot , it is observed that more no.of movies/shows are released on 1st of the month.
- Most of the movies/series are added to Netflix in the months of July, December,April, January and on Fridays

5. Missing Value & Outlier check (Treatment optional) (10 Points)

Missing DATA:

- The data consists of 8807 records and some missing values.
- For the directors column 29 % data is missing, for cast and country columns around 9% of data is missing.

Outlier Check:

- In the box plot between Geners and Year , It was inferred that most of the medians of the boxplots lies round the year 2019.
- Most of the outliers lie below the year 2014.

6. Insights based on Non-Graphical and Visual Analysis (10 Points)

6.1 Comments on the range of attributes:

Show_id :

- It is unique and consists of 8807 values and gives the total no.of movies and TV shows put together.

Type:

- It gives the information about wether particular record is a movie or TV Series.

- Highest no.of shows have only one season.
- There are almost very less no.of shows with more than 4 seasons.
- Many shows are TV shows and horror geners.
- There are more no.of movies than tv shows.

Title:

- It gives the information about the name of the title which again is a unique value.

Director:

- This column has most % of missing values which is almost 30%
- Rajiv Chilaka has directed more no.of Movies/Series.

Cast:

- This column gives the information about who are the artists that worked for particular movie/series.
- Anupam kher has worked in many no.of movies/shows

Country:

- It gives the information about in which countries a particular Movie/series is available.
- USA, India, UK have more movies/series available followed by others.

Date_added:

- It will give information about what is the range of data set which is from 2008-01-01 to 2021-09-25
- The trends on which movies/series are released can be known.
- The data set shows that most of the movies/series were release in the months of july, December, January and april months, 1^{st} of the month, and on Fridays.

Release_year:

- In which year a movie/series was released can be inferred.
- From this we can estimate the correlation of no.of movies/series releasing every year are increasing or decreasing ie., we can get the positive/negative correlation between the time and the movie/series count.
- In this data set the no.of movies/series increased exponentially till 2019 and there after showed a slight decreasing trend.

- From this we can study what were the factors that contributed for the growth and that contributed for the decline in that particular period and take necessary step for being inline.

Rating:

It gives the information about the type of consumers that a particular movie/series belong to.

- Most of the movies/series belonged to TV-MA.

Duration:

- For most of the movies the duration was between 50 mins to 150 mins.
- Highest no.of movies were of duration around 100mis.
- Most of the series were of only one season, and very less series have more than 4 seasons.

Listed_in:

- Gives the information about a particular Genre.
- Most of the movies belonged to international movies Genre, followed by drama, action and others.
- Most of the series belonged to TV shows followed by horror and others.

Description:

- It gave a brief story line of the particular movie/Genre.

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

- The no.of movies/shows showed a positive correlation with year till 2019 and showed a negative correlation after that till 2020 and showed a high distribution towards the last decade ie., 2010 to2020.
- The month variable has shown a repeating trend in the no.of addition of movie/series.
- Most of the movies/series were added on Fridays and showed almost even distribution for other das

6.3 Comments for each univariate and bivariate plot.

- There are more no.of movies compared to series.

- Most of the movies have duration around 100 mins, ie 75mins to 125 mins, with 100mins being the highest.
- Most of the series have only one season
- There are almost few no.of series having morethan 4 seasons.

7. Business Insights (10 Points) - Should include patterns observed in the data along with what you can infer from it

- Movies can be more profitable than shows, since there was an increasing trend in the no.of movies added to Netflix with idle duration around 100mins.
- Series with only one season can be profitable as there is decreasing trend in the no.of seasons.
- New content can grab the attention of consumers rather than extending the series.
- Factors for decline in the no.of movies/series should be traced.
- In the same way , factors for increase in the no.of movies/series should be studied.

8. Recommendations (10 Points) - Actionable items for business. No technical jargon. No complications. Simple action items that everyone can understand.

- Budget estimates should be considered while buying or producing a movie/series.
- A 2 sec survey can be run on youtube and such platforms ,while buying a particular movie.
- Team which consists of all sorts of consumer can be built to analyse the consumer pulse before buying a movie.
- New talent should be encouraged by providing scholarships and internships as this can develop a win-win strategy for film-makers and producers and consumers by optimizing the cost.
- Promotions and campaigns can be run on all the online platforms by providing discounts to the users.