# Defining Problem Statement and Analysing basic metrics

Problem Statement: **to** Analyze the data and generate insights which helps deciding what type of shows/movies to produce and how to grow business in various countries

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

Number of Movies Released per Year (Last 20-30 years)
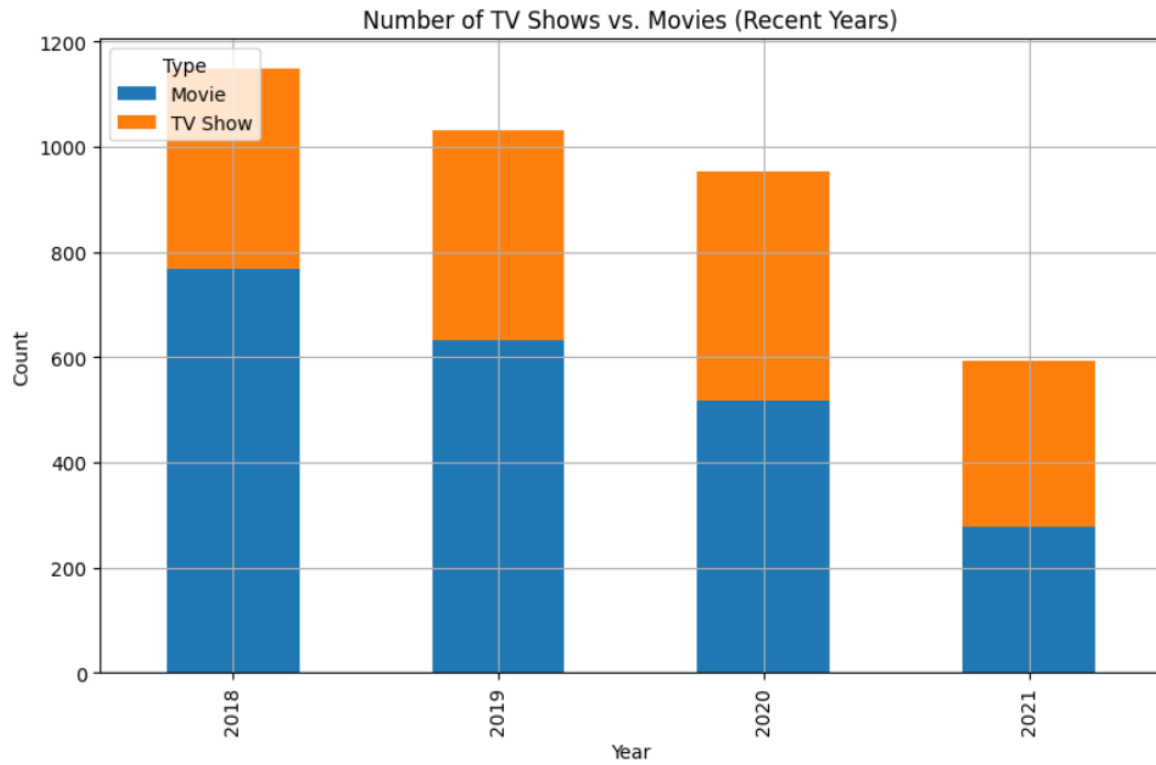
Comparison of tv shows vs. movies.
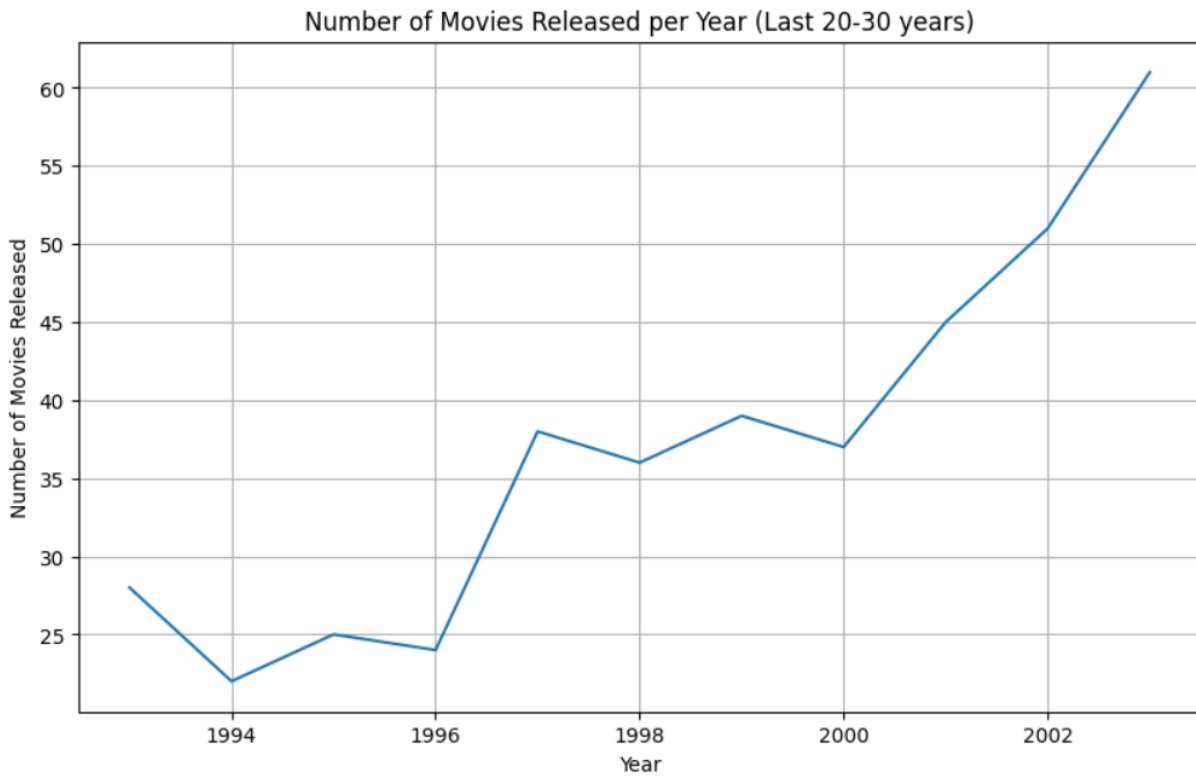
Comparison of TV Shows vs. Movies
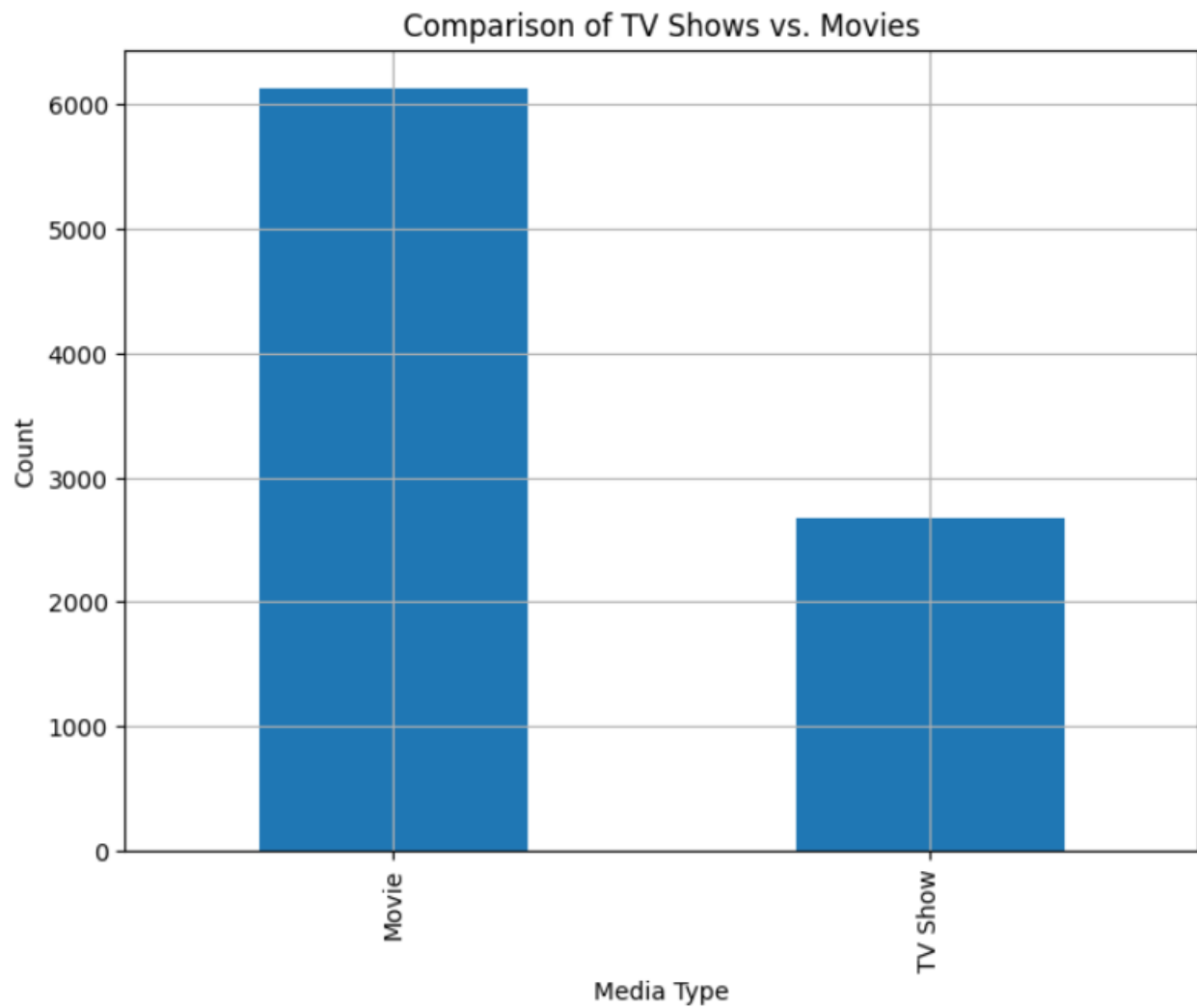
Does Netflix has more focus on TV Shows than movies in recent years

Recent year it have equal focus on tv shows and movies which shows that tv shows focus is increasing yearly

Number of TV Shows vs. Movies (Recent Years)

Understanding what content is available in different countries

	country	listed_in
0	, France, Algeria	Dramas, Independent Movies, International Movies
1	, South Korea	International TV Shows, TV Dramas
2	Argentina	Crime TV Shows, International TV Shows, Spanis...
3	Argentina, Brazil, France, Poland, Germany, De...	Dramas, International Movies, Thrillers
4	Argentina, Chile	Dramas, Independent Movies, International Movi...
...	...	...
743	Venezuela	Documentaries, International Movies
744	Venezuela, Colombia	Documentaries, International Movies
745	Vietnam	Dramas, International Movies, Romantic Movies,...
746	West Germany	Documentaries, International Movies
747	Zimbabwe	Comedies, International Movies, Romantic Movies

748 rows × 2 columns

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 Data:   (8807, 12)
```
data types of all the attributes:

```
show_id         object
type            object
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
duration        object
listed_in       object
description     object
dtype: object
```

conversion of categorical attributes to 'category'

```
df['country'] = df['country'].astype('category')
df['country']
```

```
0         United States
1          South Africa
2                   NaN
3                   NaN
4                 India
          ...
8802      United States
8803                NaN
8804      United States
8805      United States
8806              India
Name: country, Length: 8807, dtype: category
Categories (748, object): [', France, Algeria', ', South Korea', 'Argentina', 'Argentina, Brazil, France, Poland, Germany,
D..., ..., 'Venezuela, Colombia', 'Vietnam', 'West Germany', 'Zimbabwe']
```

missing value detection:

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

statistical summary :

```
df.describe()
```

	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

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

```
df['director'].value_counts()
```

```
director
Rajiv Chilaka                        19
RaÃºl Campos, Jan Suter              18
Marcus Raboy                         16
Suhas Kadav                          16
Jay Karas                            14
                                     ..
Raymie Muzquiz, Stu Livingston        1
Joe Menendez                          1
Eric Bross                            1
Will Eisenberg                        1
Mozez Singh                           1
Name: count, Length: 4528, dtype: int64
```

```
df['director'].unique()
```

```
array(['Kirsten Johnson', nan, 'Julien Leclercq', ..., 'Majid Al Ansari',
       'Peter Hewitt', 'Mozez Singh'], dtype=object)
```

```
df['type'].value_counts()
```

```
type
Movie      6131
TV Show    2676
Name: count, dtype: int64
```

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

Distribution of Content by Country