Case Study: Netflix - Data Exploration and Visualization

Submitted by: Abhishek Singh

Python Notebook Link

Business Problem:

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.

Pre-Requisites:

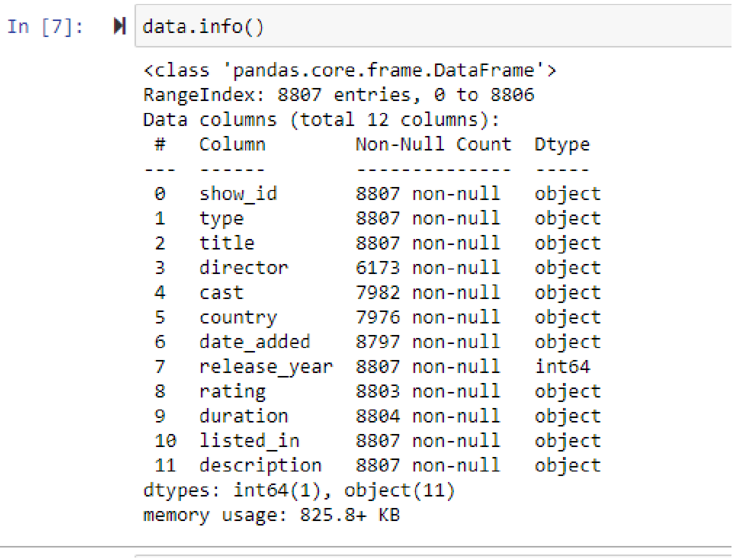
1. Required Python Libraries were imported.
2. Dataset was imported to Jupyter notebook, stored in a data frame and preliminary analysis performed on the data to understand the data size, check null values, and derive statistical observations from the data.

Initial Analysis, Observations and Data Cleansing:

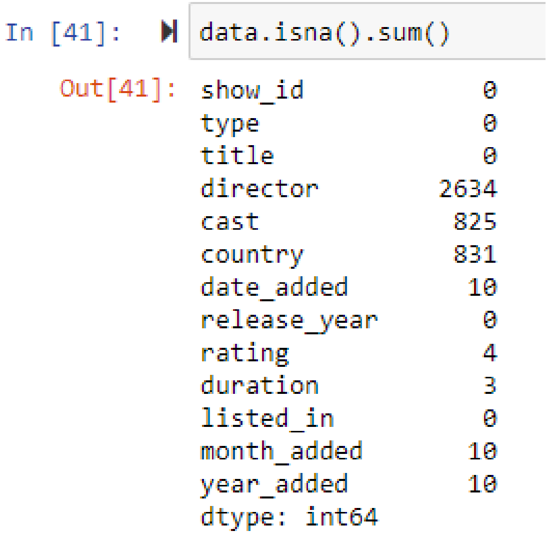
1. The dataset provided consists of a list of all the TV shows/movies available on Netflix:

Column Name	Column Description
Show_id	Unique ID for every Movie / Tv Show
Type	Identifier - A Movie or TV Show
Title	Title of the Movie / Tv Show
Director	Director of the Movie
Cast	Actors involved in the movie/show
Country	Country where the movie/show was produced
Date_added	Date it was added on Netflix
Release_year	Actual Release year of the movie/show
Rating	TV Rating of the movie/show
Duration	Total Duration - in minutes or number of seasons
Listed_in	Genre
Description	The summary description

2. Dataset dimension is 8807 x 12 (8807 Rows x 12 Columns).
3. All columns except 'release_year' are object type data. 'release_year' is a int datatype column. Columns 'director', 'cast', 'country', 'date_added', 'rating' and 'duration' have 2634, 825, 831,10, 4 and 3 Null values respectively.

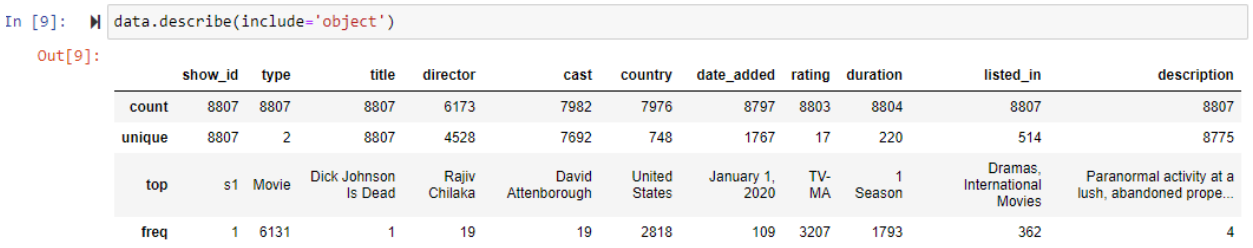


```
In [7]:  data.info()

<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
```



```
In [41]:  data.isna().sum()

Out[41]: 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
         month_added       10
         year_added        10
         dtype: int64
```

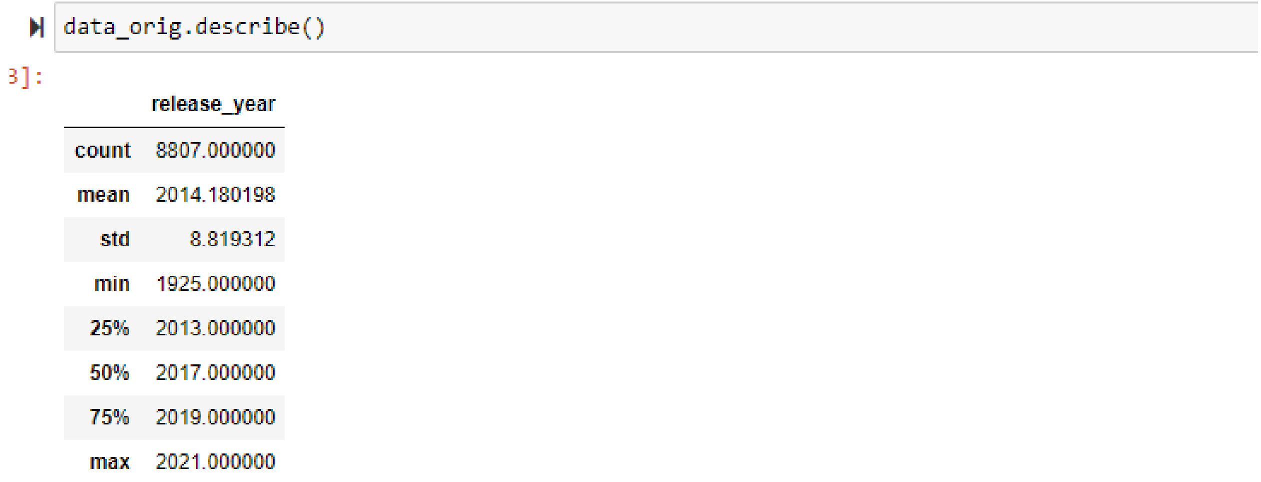
4. Column ‘show_id’ is unique identifier for each show and has non duplicate entries.
5. Value_counts for the categorical columns was checked and is attached below.



```
In [9]:  data.describe(include='object')
Out[9]:
```

	show_id	type	title	director	cast	country	date_added	rating	duration	listed_in	description
count	8807	8807	8807	6173	7982	7976	8797	8803	8804	8807	8807
unique	8807	2	8807	4528	7692	748	1767	17	220	514	8775
top	s1	Movie	Dick Johnson Is Dead	Rajiv Chilaka	David Attenborough	United States	January 1, 2020	TV-MA	1 Season	Dramas, International Movies	Paranormal activity at a lush, abandoned prope...
freq	1	6131	1	19	19	2818	109	3207	1793	362	4

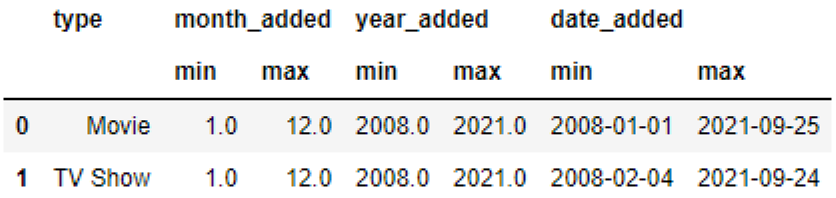
6. Statistical analysis of ‘release_year’ data was performed, and it was inferred that the oldest movie in the dataset was released in 1925 while the newest movie in 2021.



```
data_orig.describe()
3]:
```

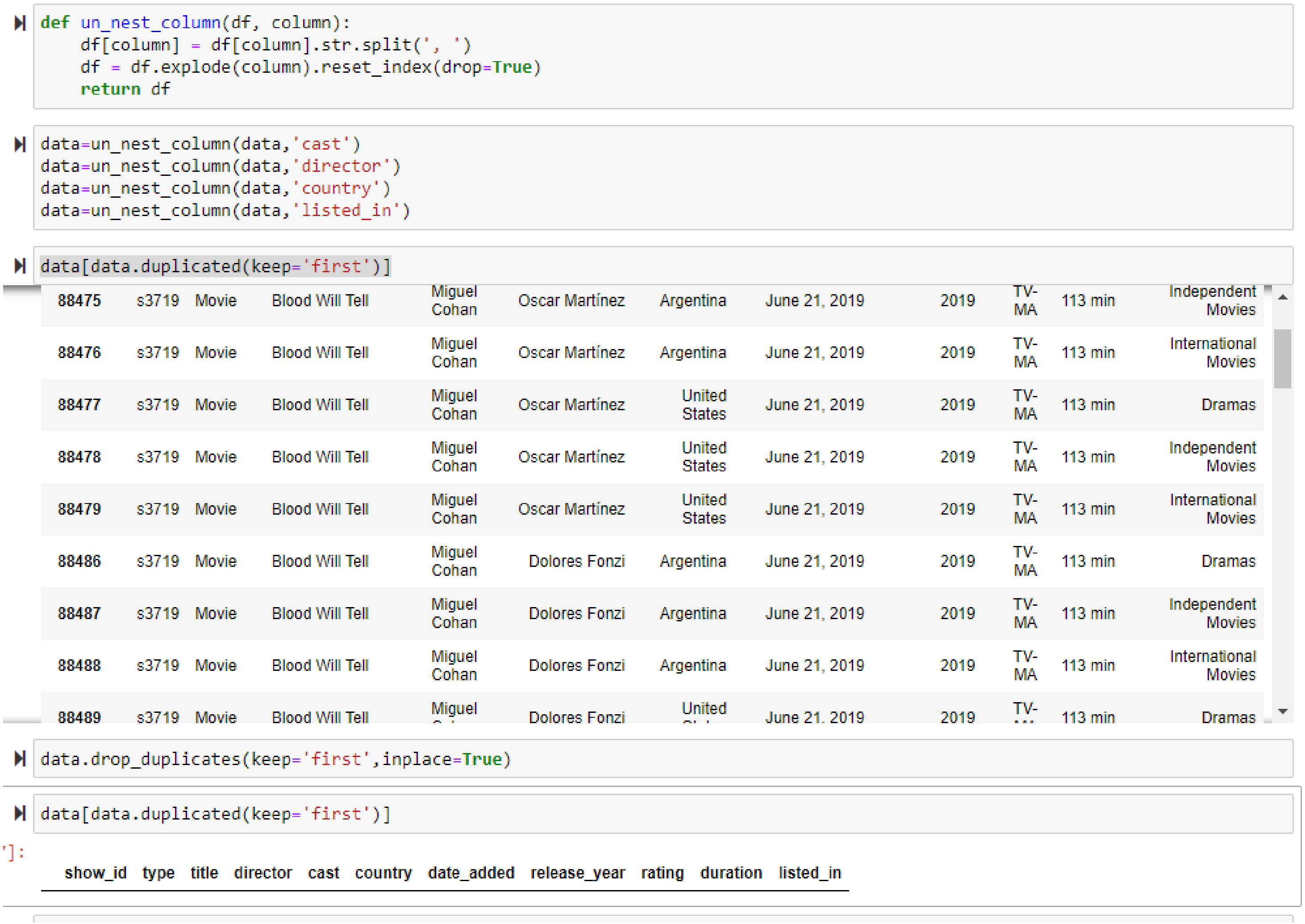
	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

7. Column ‘description’ will not be used for analysis and was dropped from the dataset.
8. Null values in ‘director’, ’cast’, ‘rating’, ‘duration and ’country’ columns were filled with string ‘unknown’ and missing ‘date_added’ values were replaced with date ‘1899-01-01’.
9. Dataset contains list of movies and TV shows added to Netflix between ‘2008-01-01’ and ‘2021-09-25’.



	type	month_added		year_added		date_added	
		min	max	min	max	min	max
0	Movie	1.0	12.0	2008.0	2021.0	2008-01-01	2021-09-25
1	TV Show	1.0	12.0	2008.0	2021.0	2008-02-04	2021-09-24

10. Columns ‘director’, ‘cast’, ‘country’, ‘listed_in’ have multiple commas separated values which are separated by creating multiple rows and assigning one value in each row and the expanded dataset is checked for any duplicate rows and deleted (if any).



```python
def un_nest_column(df, column):
    df[column] = df[column].str.split(', ')
    df = df.explode(column).reset_index(drop=True)
    return df
```

```python
data=un_nest_column(data,'cast')
data=un_nest_column(data,'director')
data=un_nest_column(data,'country')
data=un_nest_column(data,'listed_in')
```

```python
data[data.duplicated(keep='first')]
```

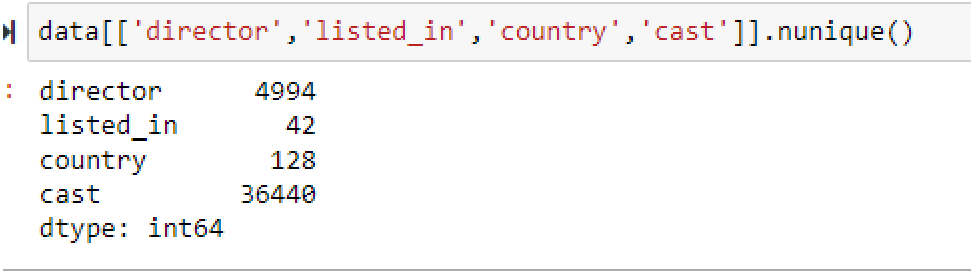
88475	s3719	Movie	Blood Will Tell	Miguel Cohan	Oscar Martínez	Argentina	June 21, 2019	2019	TV-MA	113 min	Independent Movies
88476	s3719	Movie	Blood Will Tell	Miguel Cohan	Oscar Martínez	Argentina	June 21, 2019	2019	TV-MA	113 min	International Movies
88477	s3719	Movie	Blood Will Tell	Miguel Cohan	Oscar Martínez	United States	June 21, 2019	2019	TV-MA	113 min	Dramas
88478	s3719	Movie	Blood Will Tell	Miguel Cohan	Oscar Martínez	United States	June 21, 2019	2019	TV-MA	113 min	Independent Movies
88479	s3719	Movie	Blood Will Tell	Miguel Cohan	Oscar Martínez	United States	June 21, 2019	2019	TV-MA	113 min	International Movies
88486	s3719	Movie	Blood Will Tell	Miguel Cohan	Dolores Fonzi	Argentina	June 21, 2019	2019	TV-MA	113 min	Dramas
88487	s3719	Movie	Blood Will Tell	Miguel Cohan	Dolores Fonzi	Argentina	June 21, 2019	2019	TV-MA	113 min	Independent Movies
88488	s3719	Movie	Blood Will Tell	Miguel Cohan	Dolores Fonzi	Argentina	June 21, 2019	2019	TV-MA	113 min	International Movies
88489	s3719	Movie	Blood Will Tell	Miguel	Dolores Fonzi	United	June 21, 2019	2019	TV-	113 min	Dramas

```python
data.drop_duplicates(keep='first',inplace=True)
```

```python
data[data.duplicated(keep='first')]
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in

11. Column ‘Date_added’ was converted to datetime to enable date time operations. ‘month_added’ and ‘year_added’ values were calculated from ‘date_added’ values using datetime operations.
12. Column ‘type’ categorizes the show as ‘Movie’ and ‘TV Show’. The column was converted to ‘category’ datatype for memory efficiency. Dataset was segregated into two separate datasets based on whether the entry is ‘Movie’ or ‘TV Show’.
13. Shows listed on Netflix are directed by 4993 unique directors, have 36440 unique cast members, produced across 128 unique countries in the world and across 43 unique genres.



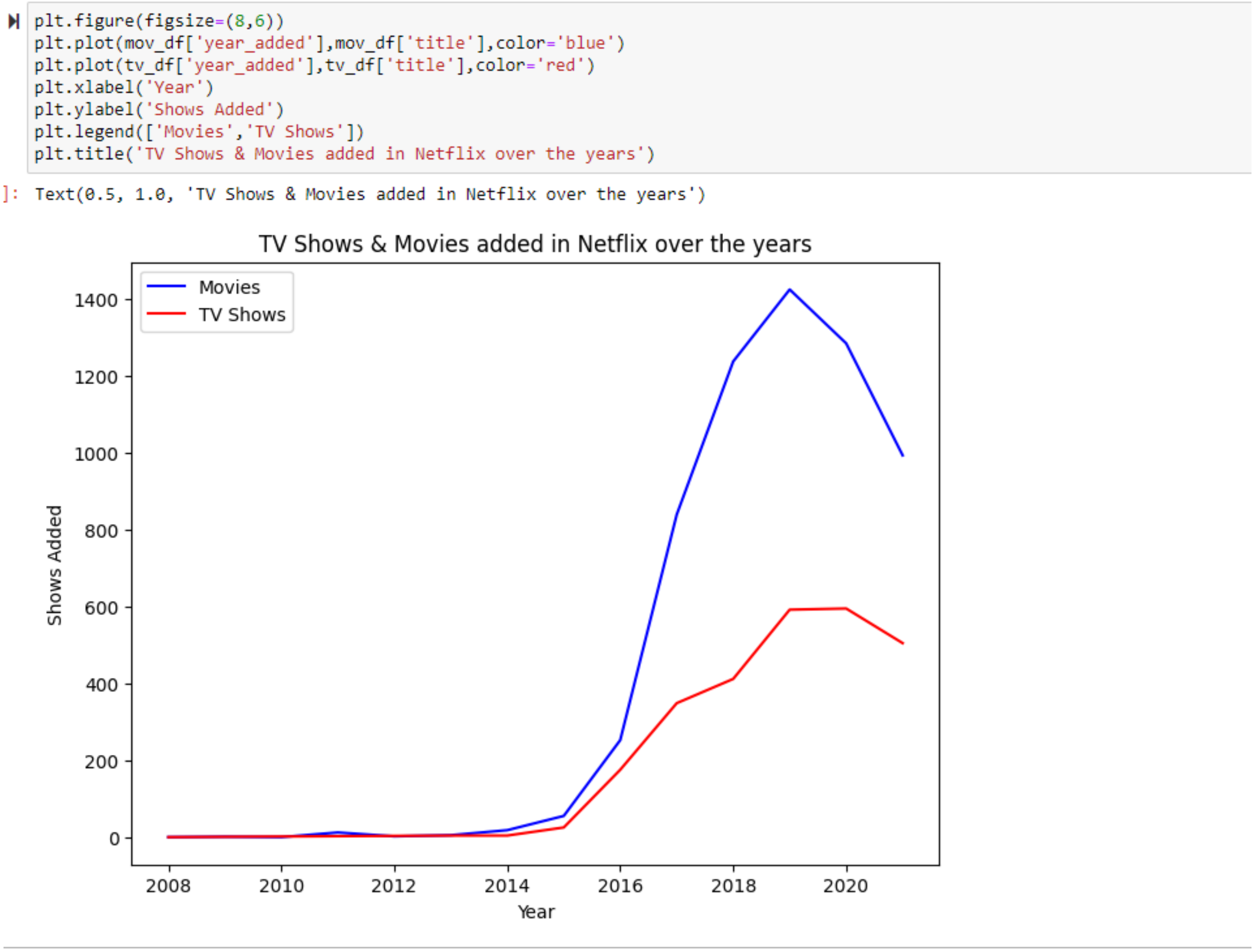
```python
data[['director','listed_in','country','cast']].nunique()
```

```
director      4994
listed_in       42
country        128
cast         36440
dtype: int64
```

Graphical Analysis, Observations and Insights:

○ Comparison of TV Shows vs. Movies

1. Univariate analysis shows that the number of movies and TV shows added to Netflix catalogues has added over the years. The number of shows beings added witnessed a sharp increase from 2014 to 2020, which could be attributed to internet access being more accessible for a larger subset of population. The slight dip in the graph for the year 2021 may be accounted to the incomplete data available for the said year.



```
plt.figure(figsize=(8,6))
plt.plot(mov_df['year_added'],mov_df['title'],color='blue')
plt.plot(tv_df['year_added'],tv_df['title'],color='red')
plt.xlabel('Year')
plt.ylabel('Shows Added')
plt.legend(['Movies','TV Shows'])
plt.title('TV Shows & Movies added in Netflix over the years')
```

```
]: Text(0.5, 1.0, 'TV Shows & Movies added in Netflix over the years')
```

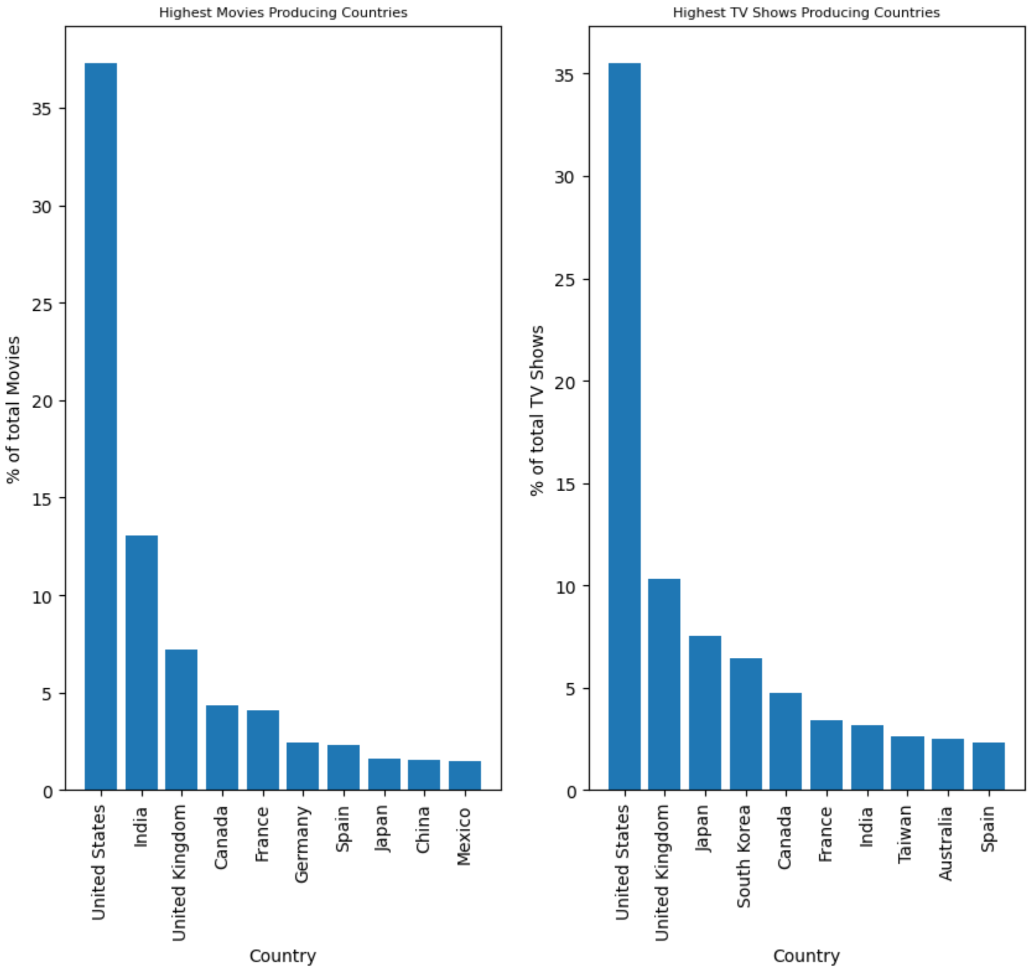
2. The available catalogue comprises of movies and TV shows primarily produced in United States (~35 % each) which caters to the English-speaking users which might be inaccessible to a large section of non-English speaking population.

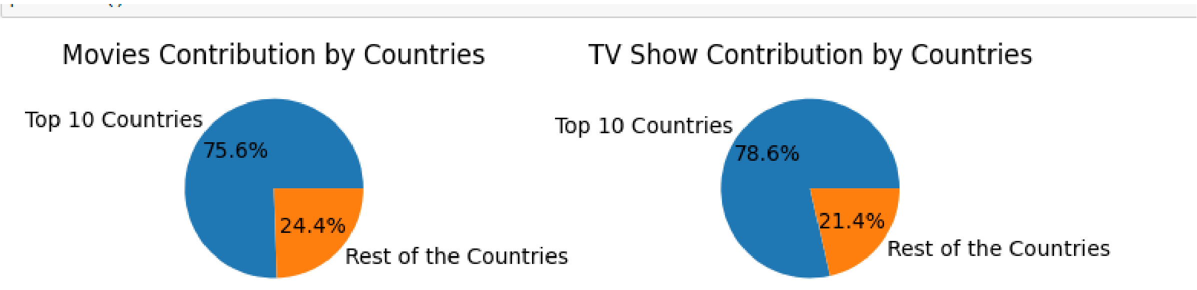
Note:

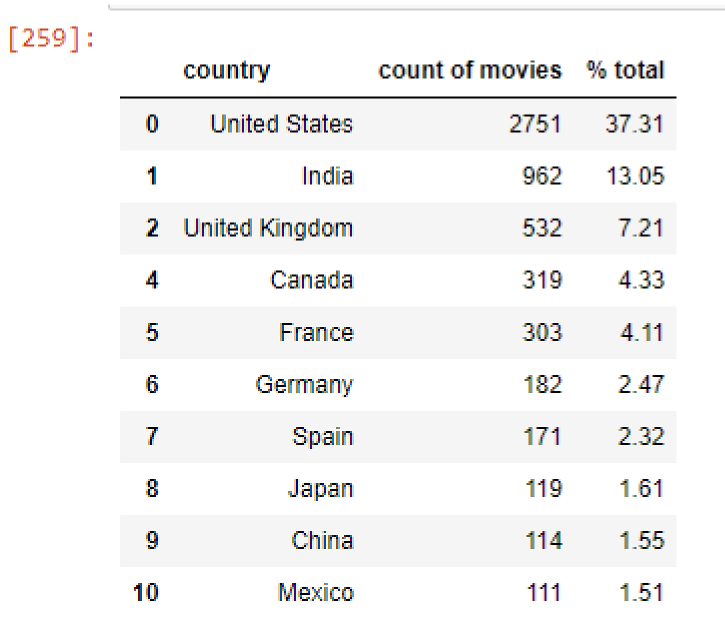
- In the visual analysis below, we have chosen only the top 10 countries producing highest percentage of movies/TV shows as the percentage contribution by the countries beyond top 10 countries is less than 2% each, hence can be ignored.

- The entries where the country information was not available were excluded from the analysis and calculation.
- The observations would require validation and a second instance of calculations if the missing country information is made available.

3. The top 10 countries contribute ~75% and ~78% of the total movies and TV shows respectively, available on the service. This further illustrates that majority of the users could be concentrated to a relatively small demographic and including regional shows from Asian and African nations could have positive impact on new users signing up for the service.

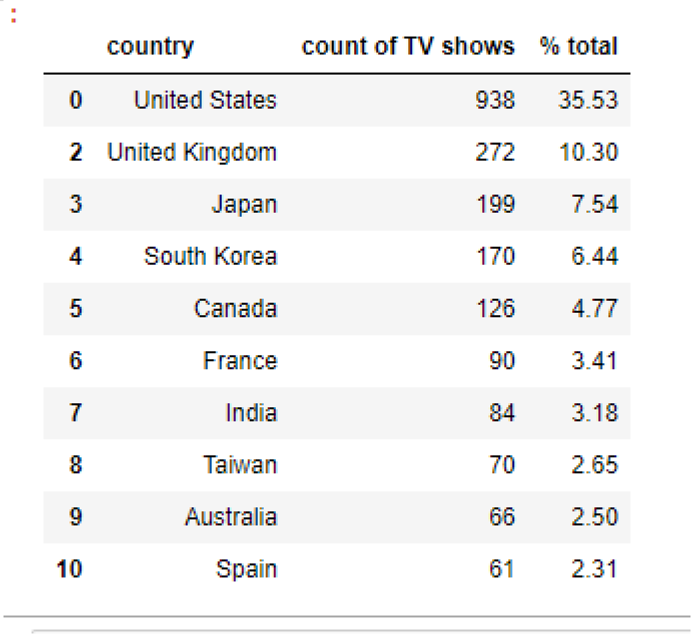






[259]:

	country	count of movies	% total
0	United States	2751	37.31
1	India	962	13.05
2	United Kingdom	532	7.21
4	Canada	319	4.33
5	France	303	4.11
6	Germany	182	2.47
7	Spain	171	2.32
8	Japan	119	1.61
9	China	114	1.55
10	Mexico	111	1.51



:

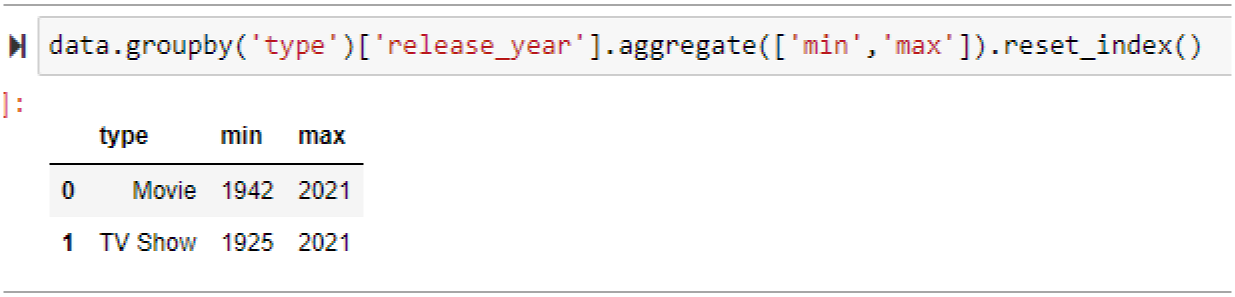
	country	count of TV shows	% total
0	United States	938	35.53
2	United Kingdom	272	10.30
3	Japan	199	7.54
4	South Korea	170	6.44
5	Canada	126	4.77
6	France	90	3.41
7	India	84	3.18
8	Taiwan	70	2.65
9	Australia	66	2.50
10	Spain	61	2.31

○ **What is the best time to launch a Movie/ TV show?**

1. Data trends show that maximum percentage of movies are added to the catalogue in the first week of the month after which the average number of added in subsequent weeks takes a slight fall. The number of movies added witness slight deviations over the weeks with regular spikes at intervals of 4 weeks, i.e., once a month. However, the quantity of movies added has a consistent negative growth during the last 6 weeks of the year relative to the rest of the year.
   Focusing attempts towards improving the new-movie collection during the last 4 to 6 weeks may incentivize users to opt into the service during the holiday season

2. Data trends show the number of new TV Shows being added is consistent throughout the year with regular spikes observed at intervals of 4 weeks.

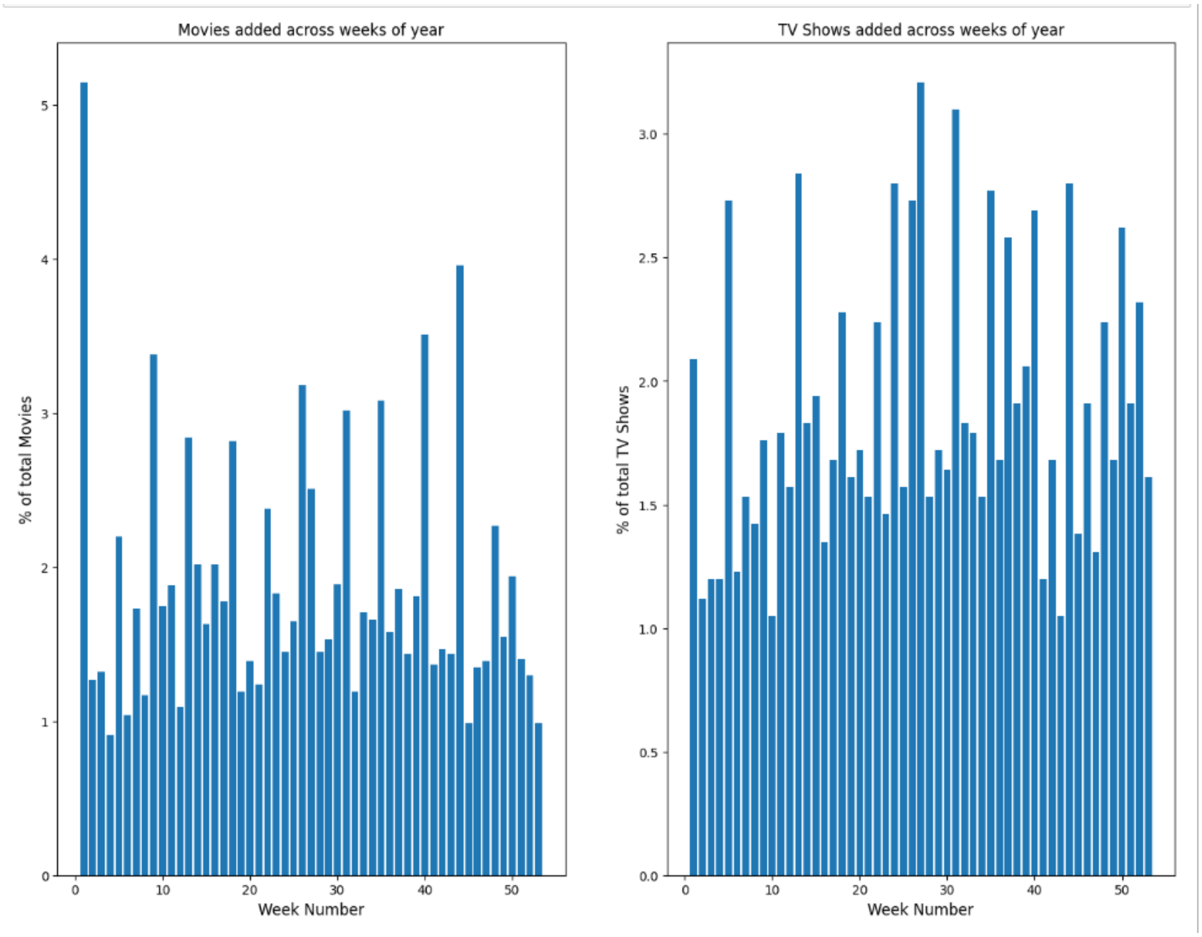
3. Furthermore, movies and TV shows available on the platform were released between the years 1942-2021 and 1925-2021.



```
data.groupby('type')['release_year'].aggregate(['min','max']).reset_index()
```

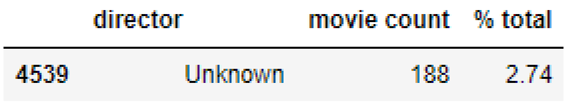
]:

	type	min	max
0	Movie	1942	2021
1	TV Show	1925	2021

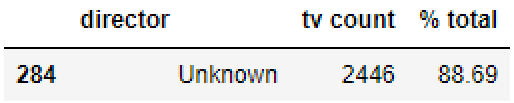


- **Analysis of actors/directors of different types of shows/movies**

1. The director information is not available for ~3% of movies and ~88% of TV Shows and consequently these movies & TV shows were not included for the analysis.

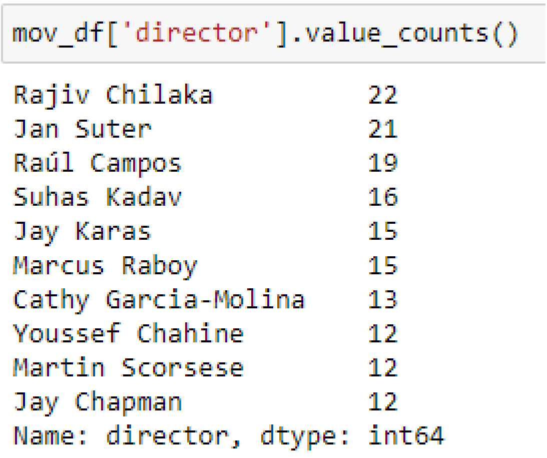


	director	movie count	% total
4539	Unknown	188	2.74

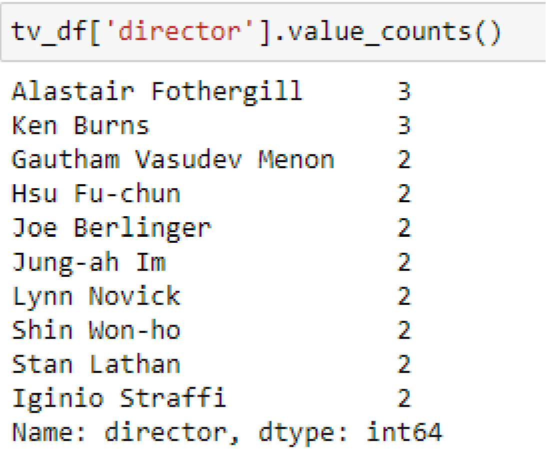


	director	tv count	% total
284	Unknown	2446	88.69

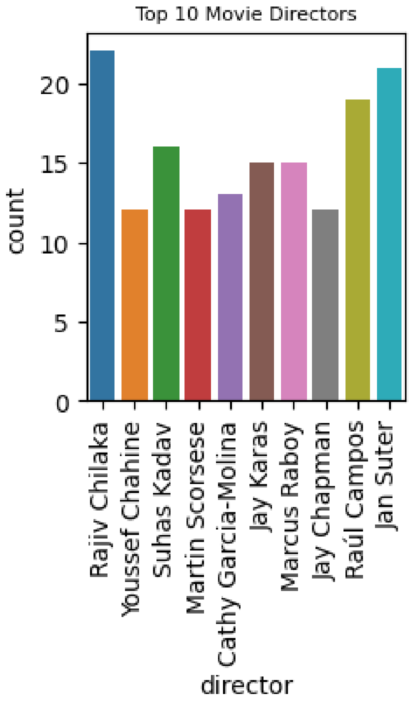
2. Data trends show that directors have worked on at most 3 unique TV shows available in the catalogue. This could suggest a lack of availability of consistent shows on the platform. This might impact recurrent subscription of users interested in the works of a particular director.
3. Movie catalogue includes multiple movies created by a single director which is a sign of a healthy movie catalogue.

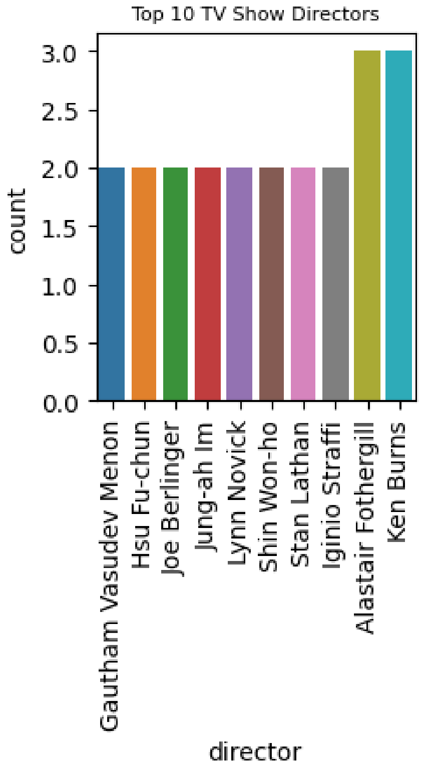


```
mov_df['director'].value_counts()

Rajiv Chilaka           22
Jan Suter               21
Raúl Campos             19
Suhas Kadav             16
Jay Karas               15
Marcus Raboy            15
Cathy Garcia-Molina     13
Youssef Chahine         12
Martin Scorsese         12
Jay Chapman             12
Name: director, dtype: int64
```

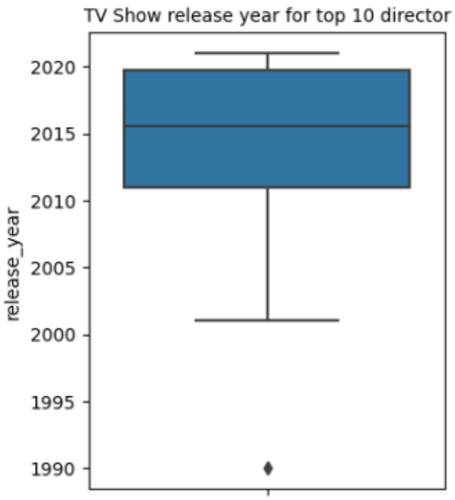


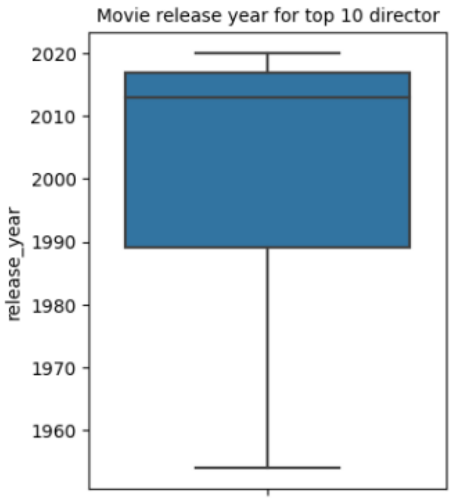
```
tv_df['director'].value_counts()

Alastair Fothergill       3
Ken Burns                 3
Gautham Vasudev Menon     2
Hsu Fu-chun               2
Joe Berlinger             2
Jung-ah Im                2
Lynn Novick               2
Shin Won-ho               2
Stan Lathan               2
Iginio Straffi            2
Name: director, dtype: int64
```



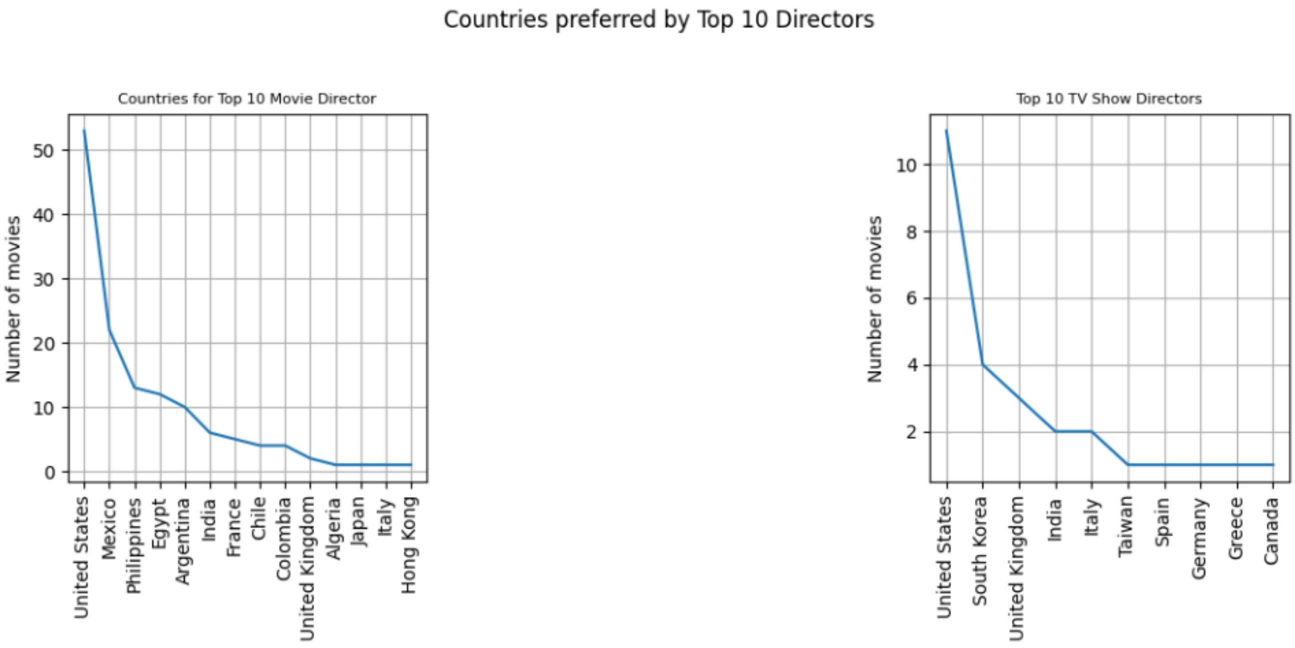


4. Further analysis shows that the movies/TV Shows available for the top 10 directors each were primarily released between 2010 to 2020, i.e., these are relatively new movies since the older TV Shows (released before 2010) and older movies (released before 1990) are far and few in numbers. This concentrated window of movie/TV Show release years could discourage users who prefer older classics from turning to recurring users, since the available catalogue for the most recurring directors might not entice them.





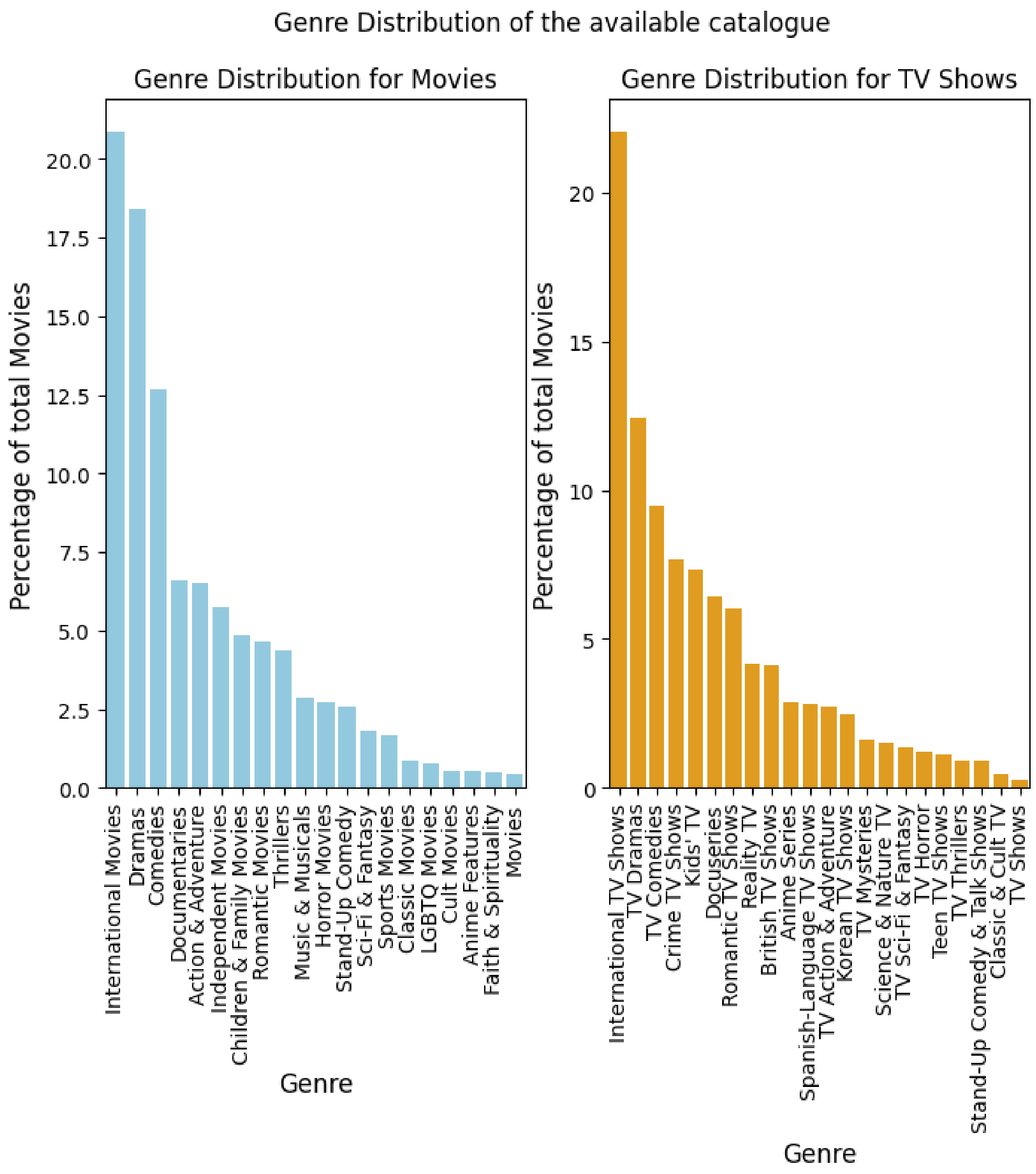
5. Data trends between top 10 directors and the countries where the associated movies/TV Shows are produced suggest reinforces that the country with highest contribution to movie catalogue have the largest number of unique movies by the directors as well. However, there are also some countries which have overall very small contribution to the catalogue, but the top 10 directors prefer to produce movies in those countries, such as Philippines, Egypt, Argentina.

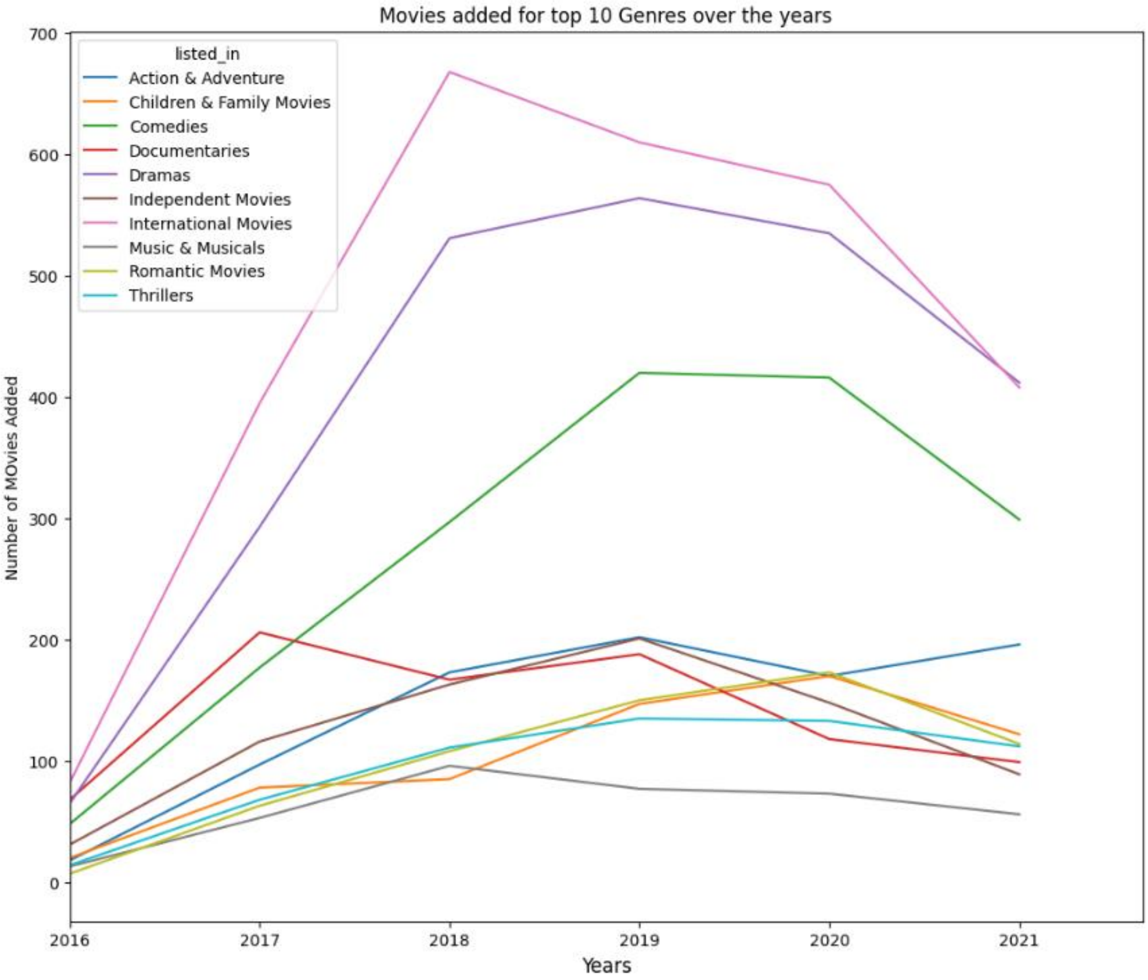


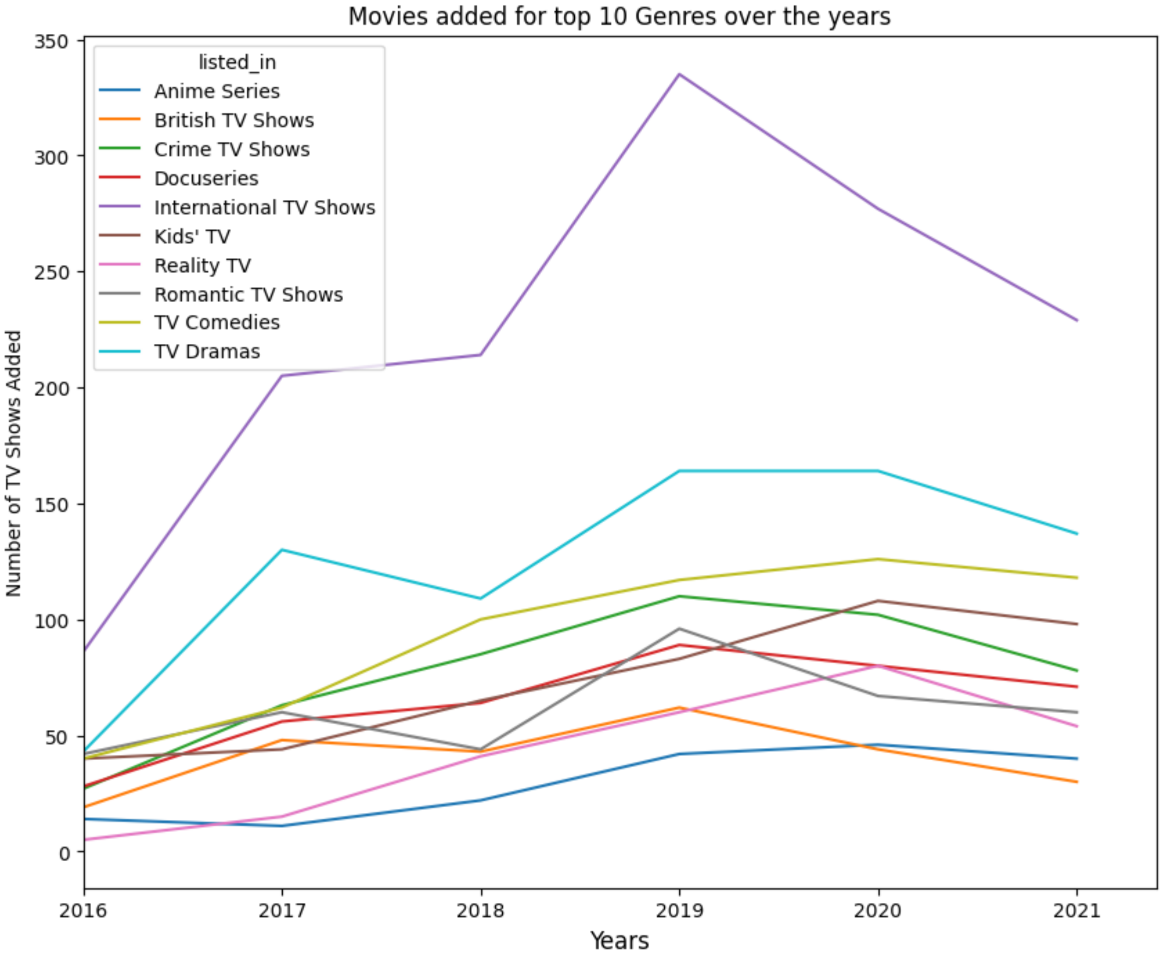
- **Which genre movies are more popular or produced more**

1. Graphical Analysis of the available movies and TV Shows show that Drama, Comedy, Documentaries and Romantic Fiction are the most widely available genres across all the available movies and TV shows, whereas, Comedy, Horror, Classic and Cult Movies and Talk Shows are among the least available options on the platform. This further elucidates that users who prefer older classical movies/shows should not be incentivized to make recurring purchase.
2. Further analysis of the top 10 movie genres shows that from 2016 onwards, the movie catalogue witnesses a linear increase however three of the 10 genres (International Movies, Drama & Comedies) witnessed majority of the influx. The other genres are undersaturated and the catalogue could be improved to attract new users to the platform.
3. Analysis of TV Shows for the top 10 genres of shows indicate a similar picture, wherein a huge volume of shows added under the International TV Shows and TV Dramas, while the rest of the genres witnessed an influx of less than 100 shows being added at their peak. The

TV show catalogue could be looked upon and improved as it comprises of a large user base of recurring customers owing to the seasonal nature of most TV shows.

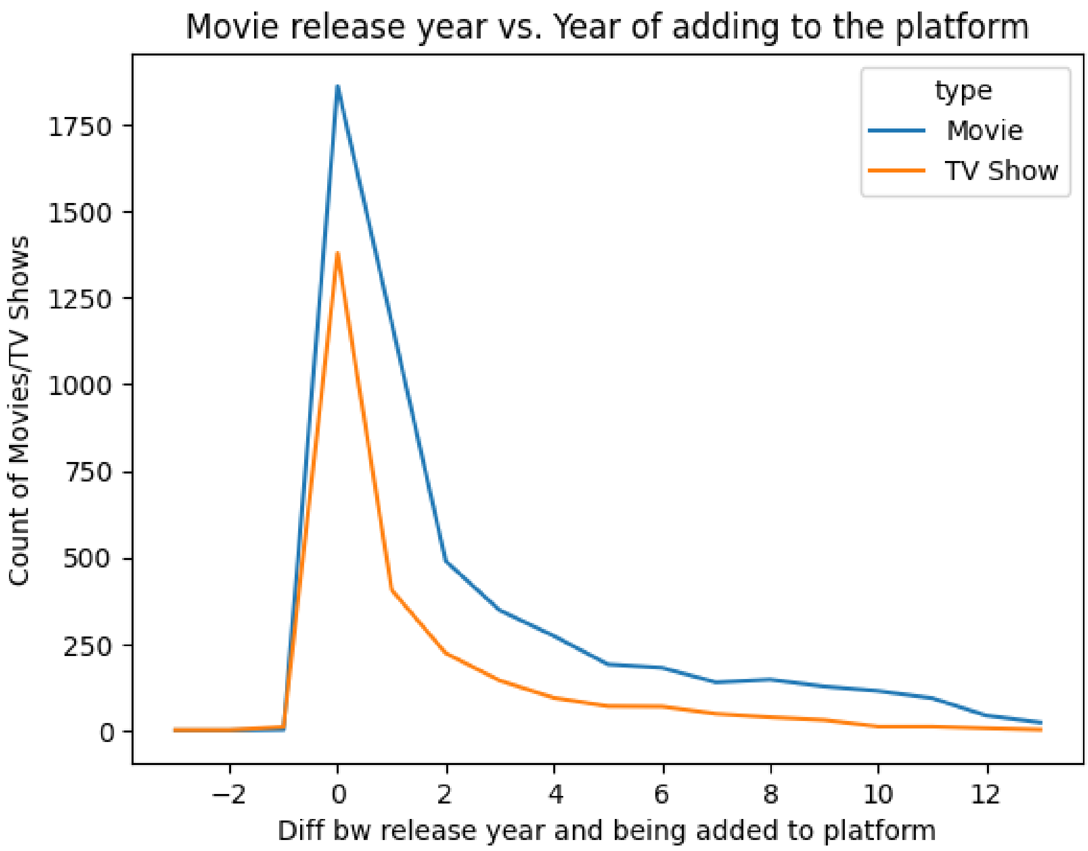



Movies added for top 10 Genres over the years
listed_in
Action & Adventure
Children & Family Movies
Comedies
Documentaries
Dramas
Independent Movies
International Movies
Music & Musicals
Romantic Movies
Thrillers
Number of MOvies Added
700
600
500
400
300
200
100
0
2016
2017
2018
2019
2020
2021
Years

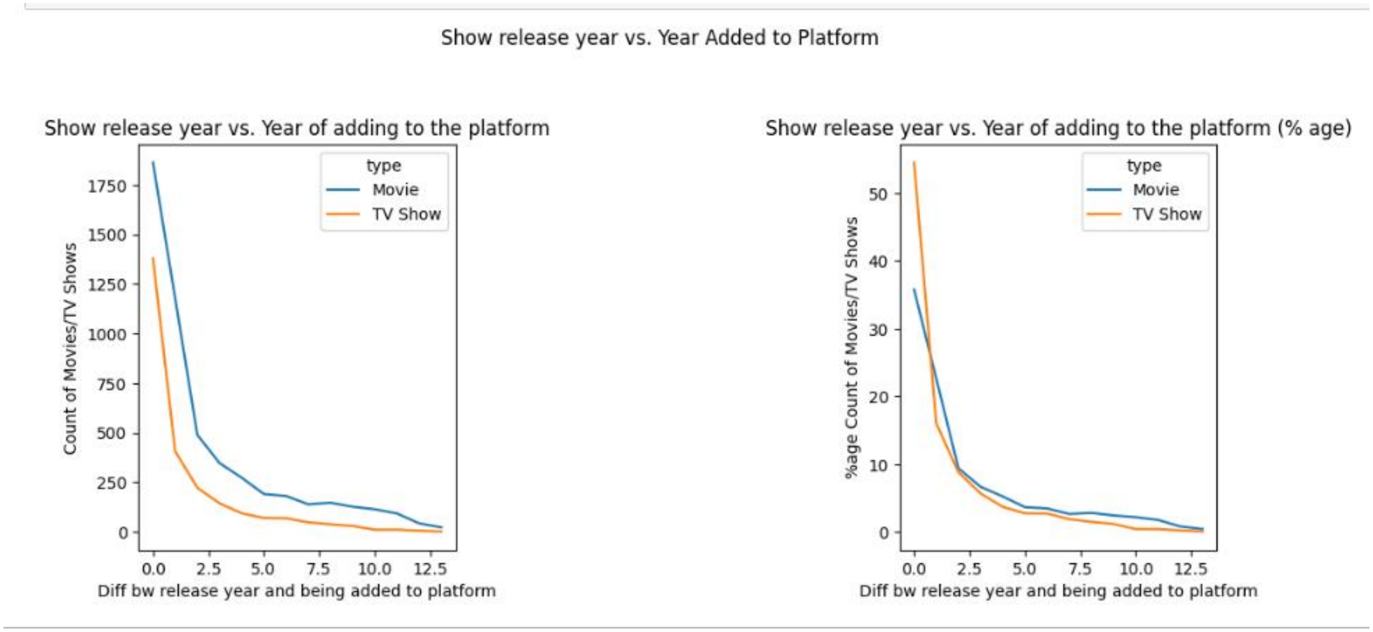

Movies added for top 10 Genres over the years
listed_in
Anime Series
British TV Shows
Crime TV Shows
Docuseries
International TV Shows
Kids' TV
Reality TV
Romantic TV Shows
TV Comedies
TV Dramas
Number of TV Shows Added
350
300
250
200
150
100
50
0
2016
2017
2018
2019
2020
2021
Years

- **Find After how many days the movie will be added to Netflix after the release of the movie**

1. Statistical analysis shows that shows have been getting added to the platform since 2008-01-01. We shall consider the movies and TV shows that have been released since 2008 to analyze the average difference between a movie being released and it being subsequently added to the platform. The entries with date_added data missing (189 in total) are not being considered for the analysis



2. Graphical analysis yields that there are discrepancies in either the 'date_added' or 'release_year' information of some tables, as the difference of the 2 values is going in the negative range. These entries are discarded for further analysis
3. Analyzing the data shows that majority (~38 % of movies and ~50% of TV Shows) are added to the catalogue withing the first year of show being released. However, the remaining shows get added more than 5 years since they were released. The ideal difference between a show being released and it being added should be lesser than 3 years, as movie/TV show enthusiasts would prefer their respective show be available as early as possible after the show release



Data Insights:

- 10 countries across the world produce 75% of all the content available on the platform. The extremely concentrated demographic impacts new user experience negatively for the users not familiar with the language, pop-culture and entertainment industry of these countries
- New movies/TV Shows are added to the platform regularly every week throughout the year, however there is decline in the number of movies added during the last 4 weeks of the year, thus impacting the application usage during festive holiday and alienating the family user base to a small extent.
- A major volume (~88%) of director information is not available for TV shows in the catalogue which impacts the data insights generated.
- Majority of TV Shows available were produced and released in the 2010s while the movies are primarily from the period of 1990 to 2015. There is scope for introduction of older classical movies/shows for the audience that might be seeking such content.
- Some directors seem to prefer a few countries that have very small contribution to the overall contribution indicating there in untapped potential in the media industry from such countries.
- The content available on the platform is concentrated within a few top genres. Users looking for genres such as Anime, Cult and Classics and Talk Shows have limited options to choose from.

Recommendations:

- Introduction of local and regional movies/TV Shows from countries with least contribution now, such as Mexico, Chile, Egypt, would encourage users from these specific demographics to opt in to the the service.
- Adding new family/holiday genre movies around the Christmas weeks to attract large group gatherings and user base (e.g. families) to sign up for the service.
- The availability of director information for the TV shows missing the director value would help provide better insights in the data
- Adding movies/TV Shows produced in the 1940s and 1950s would incentivize enthusiasts of older classical movies to become recurring users.
- Adding more contents in the genres that are currently undersaturated such as Anime, Talk Shows, Cult and Classic movies would attract a wider user base to perform recurring transactions on the platform.