Project Report

Fall 2022 – MITA Capstone Project

OTT Content Analysis

Submitted By

Pradnya Tipare
RUID: 203004532

Presented To:

Prof. Sergei Schreider

Table of Contents

Introduction

The way we consume videos has undergone massive changes. Now we have multiple OTT platforms such as Netflix, Amazon Prime Video, Hulu, and Disney+ to stream TV shows and movies online. In 2019, the OTT market was valued at 85.16 billion USD, and it is expected to reach 194.20 billion USD by 2025. Under COVID-19, many countries introduced social distancing measures that forced theaters to limit the number of audiences or even shut down and that encouraged people to stay at home, accelerating the increase in OTT platform subscriptions. Therefore, we thought it was the right time to analyze different OTT platforms and provide useful information for people not able to decide which platform fits them best.

With overabundance of information and multiple criteria to compare various OTT platforms, it has become increasingly difficult for users to find the best fit for their tastes. Through this study, we investigated different OTT platform data sets to provide users with insights into each platform to determine which services to subscribe to. This study presents an analysis of four major OTT platforms — Netflix, Amazon Prime, Disney+ and Hulu. Amongst multiple factors affecting online streaming subscriptions, we mainly analyzed the content on the platforms like the genres, TV Rating and cast using Jupyter Notebook to explore and clean the dataset and Power BI to visualize the insights.

The distinguishable factor between the Netflix and Amazon Prime Video was the age group. Netflix had adult films while Amazon Prime Video had Teens films the most. Disney+ was, undoubtedly, best suitable for animations and Hulu also had adult movies the most. Each OTT platform had its own distinct characteristics. Also, Netflix had most of the International Content making it a most sought-after platform for people from all over world.

Motivation

In recent years, the advent of various OTT platforms has introduced a novel issue: the difficulty in choosing which OTT platform to subscribe to. Netflix, Amazon Prime, and Disney+ are some of the many OTT services that are well-known to the public, but the number of services is growing as localized OTT platforms like Watcha (South Korea) and Voot (India), Hotstar (India) are joining the line.

As these platforms are coming up with new ways to stand out among competitors by presenting original content, it is evident that more customers are being lost in deciding which platform would be suitable for their use. Moreover, most of the available recommendation systems are focused on suggesting the content but not the platforms that hold and provide those contents. To ease the choice dilemma, our study aims to present a guideline for choosing the appropriate OTT platform that fits one's personal preferences.

Dataset

Dataset used for the analysis consists of different TV shows and movies from the year 2008 to 2021 for the 4 OTT Platforms:

1. **Amazon Prime**
   The Amazon Prime Video Content dataset consists of both TV shows and movies that are available on Amazon Prime as of 2021, with 9668 rows including the header and 12 columns. The columns include show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, and description.

2. **Netflix**
   The Netflix Content dataset consists of both TV shows and movies that are available on Netflix as of 2021, with 8807 rows including the header and 12 columns. The columns include show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, and description.

3. **Disney+**
   The Disney+ Content Dataset consists of both TV shows and movies that are available on Disney+ as of 2021 from its launch in 2019, with 1450 rows including the header and same 12 columns as of Netflix and Amazon Prime.

4. **Hulu**
   The Hulu Content Dataset consists of both TV shows and movies that are available on Hulu as of 2021, with 3073 rows including the header and same 12 columns as of Netflix and Amazon Prime.

Each Platform had different tables. So merging all tables together to form a single table was required. Added a new column OTT to describe which OTT platform data belongs to.

Data Description:

- Show_id : unique ID for every Movie/ TV Show
- Type: Identifier - A Movie or TV Show
- Title: Name of the Movie/ TV show
- Director : Director of the Movie
- Cast : Actors involved in the Movie/ TV Show
- Country : Country where the movie/ Show was produced
- Date_added: Date it was added on the platform
- 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: Description of the movie or show

Data Exploration

1. Importing Libraries

```
In [3]:
import numpy as np
import pandas as pd
import matplotlib as plot
```

2. Merging and loading the dataset

```
In [17]:
df = pd.concat(
    map(pd.read_csv, ['amazon_prime_titles.csv', 'disney_plus_titles.csv','hulu_titles.csv','netflix_titles.csv']), ignore_index=True)
df
```

Out[17]:

	OTT	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	Amazon Prime	s1	Movie	The Grand Seduction	Don McKellar	Brendan Gleeson, Taylor Kitsch, Gordon Pinsent	Canada	30-Mar-21	2014	NaN	113 min	Comedy, Drama	A small fishing village must procure a local d...
1	Amazon Prime	s2	Movie	Take Care Good Night	Girish Joshi	Mahesh Manjrekar, Abhay Mahajan, Sachin Khedekar	India	30-Mar-21	2018	13+	110 min	Drama, International	A Metro Family decides to fight a Cyber Crimin...
2	Amazon Prime	s3	Movie	Secrets of Deception	Josh Webber	Tom Sizemore, Lorenzo Lamas, Robert LaSardo, R...	United States	30-Mar-21	2017	NaN	74 min	Action, Drama, Suspense	After a man discovers his wife is cheating on ...
22996	Netflix	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	11-Jan-20	2006	PG	88 min	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...
22997	Netflix	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	2-Mar-19	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty...

22998 rows × 13 columns

After merging all the OTT platforms data, we can see there are total 22998 entries and 13 columns to work on.

3. Finding if dataset contains null values

```
In [18]: df.drop('description',axis = 1,inplace = True)
         df.isna().sum()
```

```
Out[18]: OTT                0
         show_id            0
         type               0
         title              0
         director        8259
         cast            5321
         country        11499
         date_added      9554
         release_year       0
         rating           864
         duration         482
         listed_in          0
         dtype: int64
```

There are a total of 35,979 null values across the entire dataset with 8259 missing points under “director” 5321 under “cast”, 11499 under “country”, 9554 under “date_added”, 864 under “rating” and 482 under “duration”. We will have to handle all null data points before we can dive into EDA.

Data Cleaning

Data Cleaning means the process of identifying incorrect, incomplete, inaccurate, irrelevant, or missing pieces of data and then modifying, replacing, or deleting them as needed. Data Cleansing is considered as the basic element of Data Science.

1. Dropping description column as it of no use is the analysis

2. Handling null values

```
In [19]: df = df.fillna('NA')
         df.isna().sum()
```

```
Out[19]: OTT             0
         show_id         0
         type            0
         title           0
         director        0
         cast            0
         country         0
         date_added      0
         release_year    0
         rating          0
         duration        0
         listed_in       0
         dtype: int64
```

For null values, the easiest way to get rid of them would be to delete the rows with the missing data. However, this wouldn't be beneficial to our EDA since there is loss of information. Since 'director', 'cast', and 'country' contain most null values, I choose to treat each missing value as unavailable i.e., NA. After replacing, we can see that there are no more null values in the dataset.

3. Changing datatype

   In the duration column, there appears to be a discrepancy between movies and shows. Movies are based on the duration of the movie and shows are based on the number of seasons. To make EDA easier, I converted the values in these columns into integers for both the movies and shows datasets.

4. Replacing values in rating

```
In [20]: df.rating.unique()
```

```
Out[20]: array(['NA', '13+', 'ALL', '18+', 'R', 'TV-Y', 'TV-Y7', 'NR', '16+',
       'TV-PG', '7+', 'TV-14', 'TV-NR', 'TV-G', 'PG-13', 'TV-MA', 'G',
       'PG', 'NC-17', 'UNRATED', '16', 'AGES_16_', 'AGES_18_', 'ALL_AGES',
       'NOT_RATE', 'TV-Y7-FV', 'NOT RATED', '2 Seasons', '93 min',
       '4 Seasons', '136 min', '91 min', '85 min', '98 min', '89 min',
       '94 min', '86 min', '3 Seasons', '121 min', '88 min', '101 min',
       '1 Season', '83 min', '100 min', '95 min', '92 min', '96 min',
       '109 min', '99 min', '75 min', '87 min', '67 min', '104 min',
       '107 min', '84 min', '103 min', '105 min', '119 min', '114 min',
       '82 min', '90 min', '130 min', '110 min', '80 min', '6 Seasons',
       '97 min', '111 min', '81 min', '49 min', '45 min', '41 min',
       '73 min', '40 min', '36 min', '39 min', '34 min', '47 min',
       '65 min', '37 min', '78 min', '102 min', '129 min', '115 min',
       '112 min', '61 min', '106 min', '76 min', '77 min', '79 min',
       '157 min', '28 min', '64 min', '7 min', '5 min', '6 min',
       '127 min', '142 min', '108 min', '57 min', '118 min', '116 min',
       '12 Seasons', '71 min', '74 min', '66 min', 'UR'], dtype=object)
```

The rating column has some values like 80min which clearly does not belong there. So, removing them.
Also renaming values in rating for better understanding.

```
In [21]: # df['rating'] = df['rating'].astype('string')
df['rating'] = df['rating'].apply(lambda x: np.NAN if (('min' in x) or ('Seasons' in x) or ('Season' in x)) else x)
df.fillna('NA',inplace = True)
df['rating'] = df['rating'].replace({'PG-13':'Teens-Age above 13','13+':'Teens-Age above 13', 'TV-MA':'Adults','18+':'Adults', 'PG':'Kids-with parenta
df.rating.unique()
```

```
Out[21]: array(['Not Rated', 'Teens-Age above 13', 'General Audience', 'Adults',
       'Kids', 'Kids-Age above 7', 'Teens-Age above 16',
       'Kids-with parental guidance', 'Teens-Age above 14'], dtype=object)
```

5. Exploding values from column Genre and Country

```
In [22]: df.country.unique()
```

```
Out[22]: array(['Canada', 'India', 'United States', 'United Kingdom', 'France',
       'Spain', 'NA', 'Italy', 'United Kingdom, France',
       'United States, Italy', 'United States, India',
       'United Kingdom, United States',
       'United States, United Kingdom, Germany', 'Japan',
       'China, United States, United Kingdom',
       'Denmark, United Kingdom, Czech Republic, Netherlands',
       'United States, Ireland', 'United States, United Kingdom, Canada',
       'Croatia, Slovenia, Serbia, Montenegro', 'Japan, Canada',
       'United States, France, South Korea, Indonesia',
       'United Arab Emirates, Jordan'], dtype=object)
```

In the Genre and Country column there are some entries which contains multiple values in the single cell. This disobeys 1NF form of normalization. So, splitting them in different rows to use it in the visualizations later.

In [23]:
```
# df.explode(df.assign(country=df.country.str.split(',')),'country')
# df['country']=df.country.apply(lambda x:x[0::].split(','))
df = df.join(df.country.str.split(', ',expand = True).stack().reset_index(drop=True,level=1).rename('Country')).drop('country',axis=1)
df = df.join(df.listed_in.str.split(', ',expand = True).stack().reset_index(drop=True,level=1).rename('Genre')).drop('listed_in',axis=1)
df
```

Out[23]:

	OTT	show_id	type	title	director	cast	date_added	release_year	rating	duration	Country	Genre
0	Amazon Prime	s1	Movie	The Grand Seduction	Don McKellar	Brendan Gleeson, Taylor Kitsch, Gordon Pinsent	30-Mar-21	2014	Not Rated	113 min	Canada	Comedy
0	Amazon Prime	s1	Movie	The Grand Seduction	Don McKellar	Brendan Gleeson, Taylor Kitsch, Gordon Pinsent	30-Mar-21	2014	Not Rated	113 min	Canada	Drama

6. Replacing values in Genre

Genre Column contains values like Drama, Dramas so making them one.

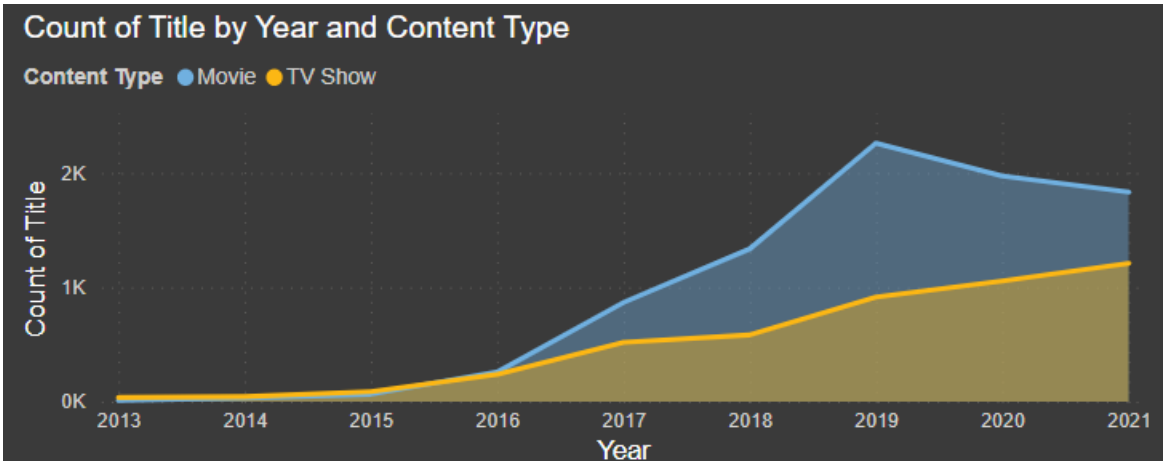
In [50]:
```
df['Genre'] = df['Genre'].replace(dict.fromkeys(['Action','Action-Adventure','TV Action & Adventure','Adventure'],'Action & Adventure'))
df['Genre'] = df['Genre'].replace({'Dramas':'Drama','Anime Features':'Anime','Anime Series':'Anime','Animation':'Anime','Comedies':'Comedy','Crime TV
```

In [51]:
```
a = df['Genre'].unique()
a.sort()
a
```

Out[51]:
```
array(['Action & Adventure', 'Adult Animation', 'Animals & Nature',
       'Anime', 'Anthology', 'Arthouse', 'Arts', 'Biographical',
       'Black Stories', 'British TV Shows', 'Buddy', 'Cartoons',
       'Children & Family Movies', 'Classic & Cult TV', 'Classics',
       'Comedy', 'Coming of Age', 'Concert Film', 'Cooking & Food',
       'Crime', 'Cult Movies', 'Dance', 'Disaster', 'Documentary',
       'Drama', 'Entertainment', 'Faith & Spirituality', 'Family',
       'Fantasy', 'Fitness', 'Game Show / Competition', 'Game Shows',
       'Health & Wellness', 'Historical', 'Horror', 'Independent Movies',
       'International Content', 'Kids', "Kids' TV", 'Korean TV Shows',
       'LGBTQ+', 'Late Night', 'Latino', 'Lifestyle',
       'Lifestyle & Culture', 'Medical', 'Military and War', 'Movies',
       'Music', 'Music Videos and Concerts', 'Mystery', 'News', 'Parody',
       'Police/Cop', 'Reality TV', 'Romance', 'Sci-Fi & Fantasy',
       'Science & Nature TV', 'Science & Technology', 'Science Fiction',
       'Series', 'Sitcom', 'Sketch Comedy', 'Soap Opera / Melodrama',
       'Spanish-Language TV Shows', 'Special Interest', 'Sports',
       'Spy/Espionage', 'Stand-Up Comedy', 'Stand-Up Comedy & Talk Shows',
       'Superhero', 'Survival', 'Suspense', 'TV Shows', 'Talk Show',
       'Teen', 'Thriller', 'Travel', 'Unscripted', 'Variety', 'Western',
       'Young Adult Audience', 'and Culture'], dtype=object)
```

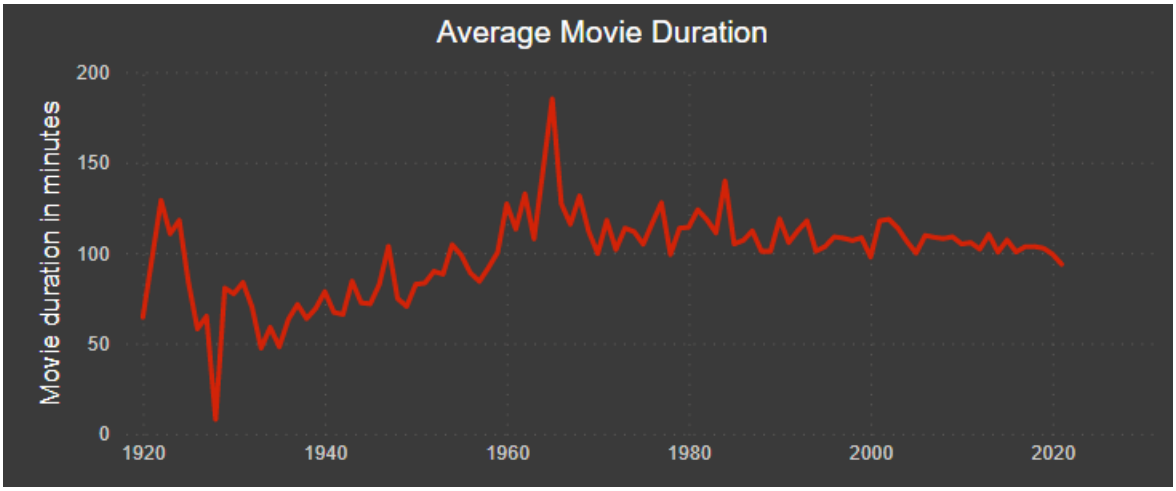
Exploratory Data Analysis

1. Amount of Content (Content Type) as a Function of Time



Based on the timeline above, we can conclude that the streaming platform started gaining traction after 2014. Since then, the amount of content added has been increasing significantly. The growth in the number of movies on Platforms is much higher than that on TV shows. Movies reached a peak of 2263 movies in 2019 and then again decreased. But on the either side the number of TV Shows are continuously increasing with the year.

2. Average Movie Duration

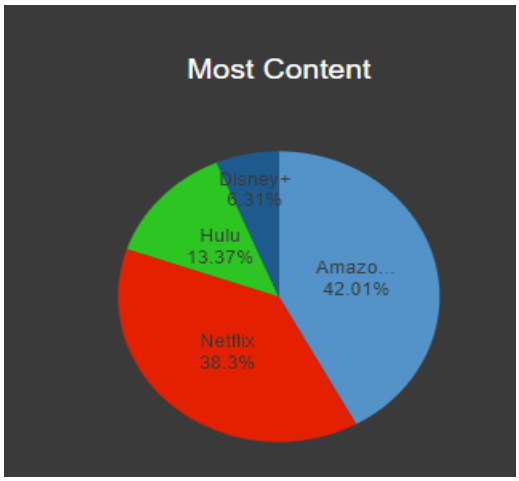


- This graph shows a trend of the average movie duration throughout the years from 1920 to 2021.
- Experiments were made during the early days of cinema for its length to understand its audience.

- Once the audience were more comfortable with the idea of cinemas as a form of entertainment, the length kept on increasing.
- Until 1965, when the optimum length (around 120 minutes) was reached, the productions decided to keep it stable for a long time.
- But the trend in that stable time is a slight dip owing to the reduced attention span of the population as we go on.

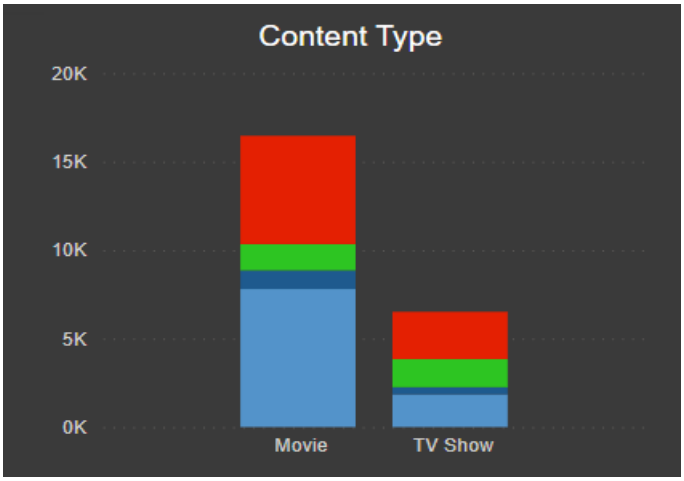
Comparison

1. Amount of Content on Platforms



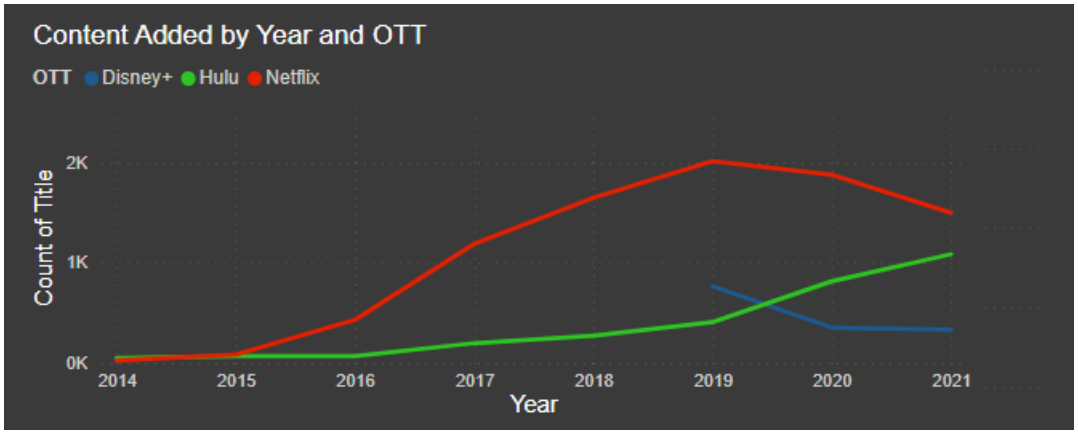
Amazon Prime has the highest content amongst all the platforms followed by Netflix and lowest on Disney+ as it was recently launched in 2019.

2. Content by Content Type



All the platforms focus more on Movies than TV Shows. Amazon Prime has highest number of movies with 7801 movies and Netflix has highest number of TV Shows with 2674 Shows.

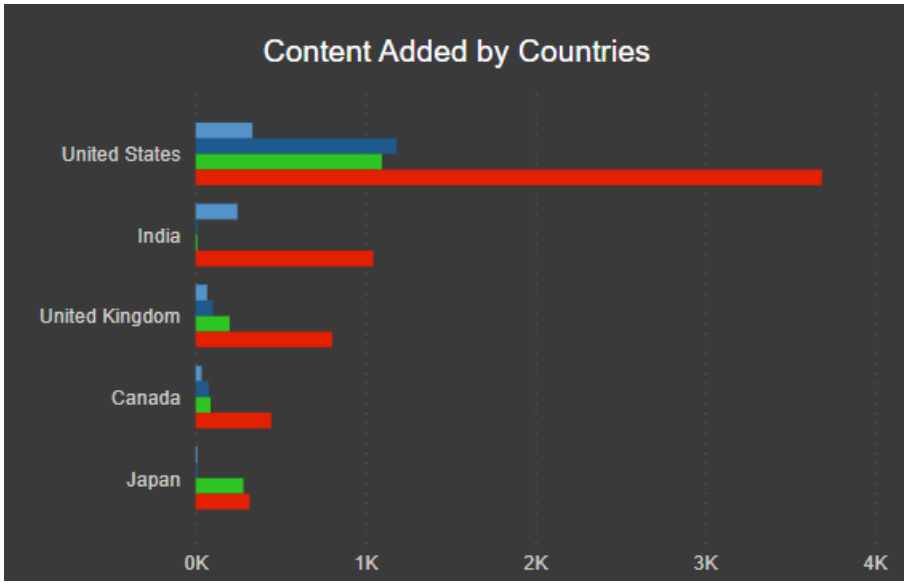
3. Content Added by Year on Platforms

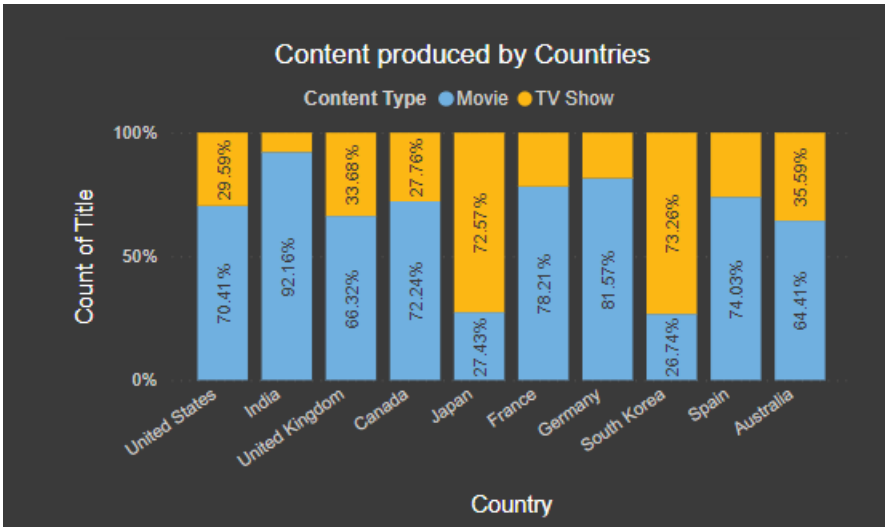


In 2007, Netflix introduced streaming media and video on demand. We see a slow in the beginning but then it picked up in 2014-2015 and there is a rapid increase till 2018.

By 2018, the content on Netflix was 13 times of 2007 year's content. But it has declined since 2019 since the beginning of covid. The other factor could be - In 2019, Disney plus was also launched. Films and television series produced by The Walt Disney Studios and Walt Disney Television, such as Marvel movies moved to Disney plus.

4. Countries by the Amount of the Produced Content

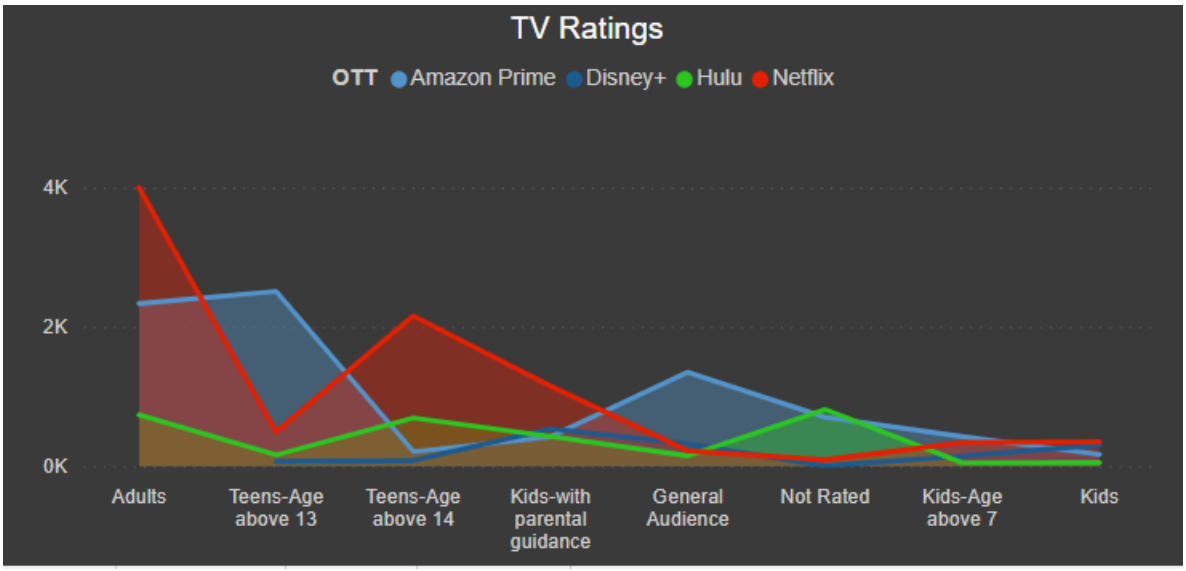




As expected,

- Unites States tops the chart followed by India, United Kingdom, and Canada.
- Interestingly, the content available in India is heavily skewed towards movies, confirming the intuition about big influence of bollywood in-house movie production.
- South Korea has the highest percentage of TV shows

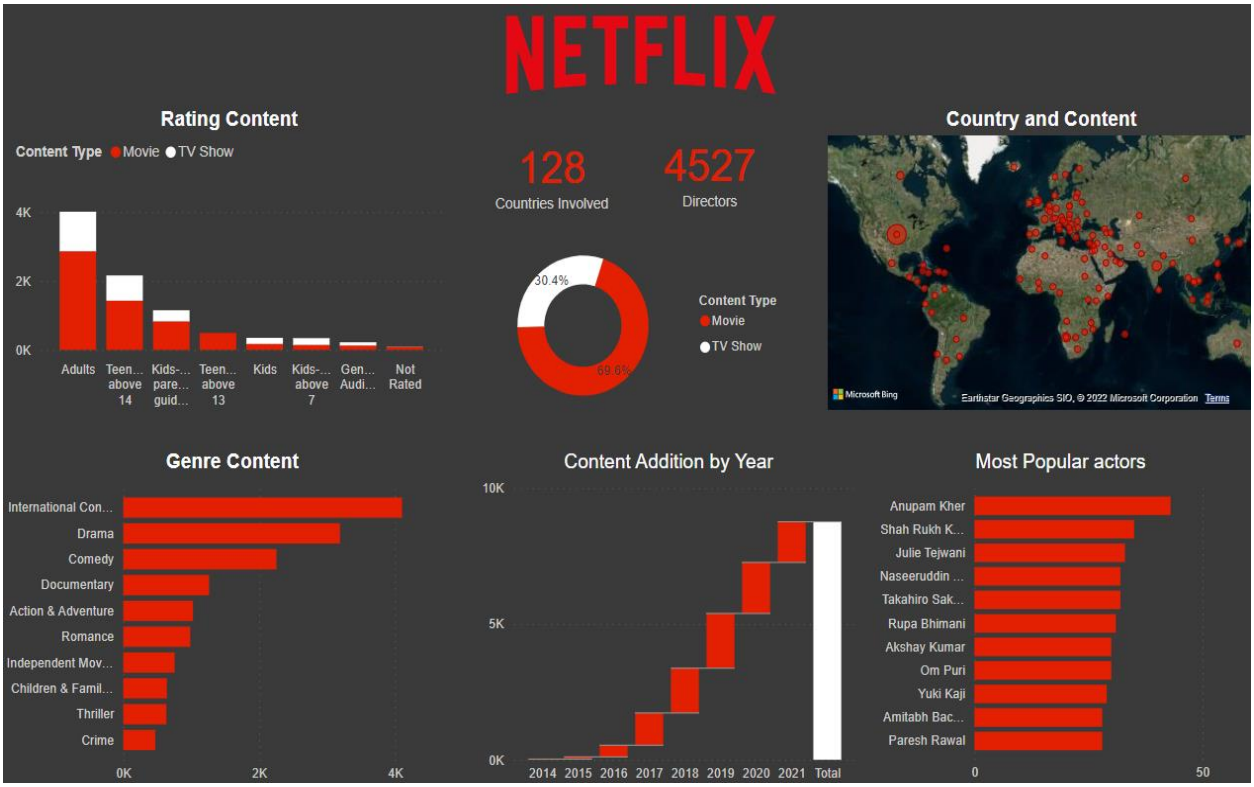
5. Amount of Content by Rating



It seems the most content on Platforms caters to adults and then teens. Netflix had adult films while Amazon Prime Video had Teens films the most. Disney+ was, undoubtedly, best suitable for animations and Hulu also had adult movies the most

Platform Specific Visualizations

1. Netflix



Netflix is one of the most popular streaming platforms.

1. It has 128 countries which contributes to the content available on Netflix, which is highest among all the platforms.
2. There are about 4,000++ movies and almost 2,000 TV shows, with movies being the majority. There are far more movie titles (69.6%) that TV shows titles (30.4%) in terms of title.
3. We explored the amount of content Netflix has added throughout the previous years. Based on the timeline above, we can conclude that the popular streaming platform started gaining traction after 2014. Since then, the amount of content added has been increasing significantly. About 1800+ new content was added in 2019 and 2020.
4. The largest count of Netflix content is made for adults followed by teens and movies for kids with parental guidance.
5. From the Genre Content graph, we can see that lot of International Content is available on Netflix making it all in one platform for people from all countries. Content for Drama

and Comedy is also available on Netflix in large amount. Documentaries are the new form of content which people like to watch a lot.

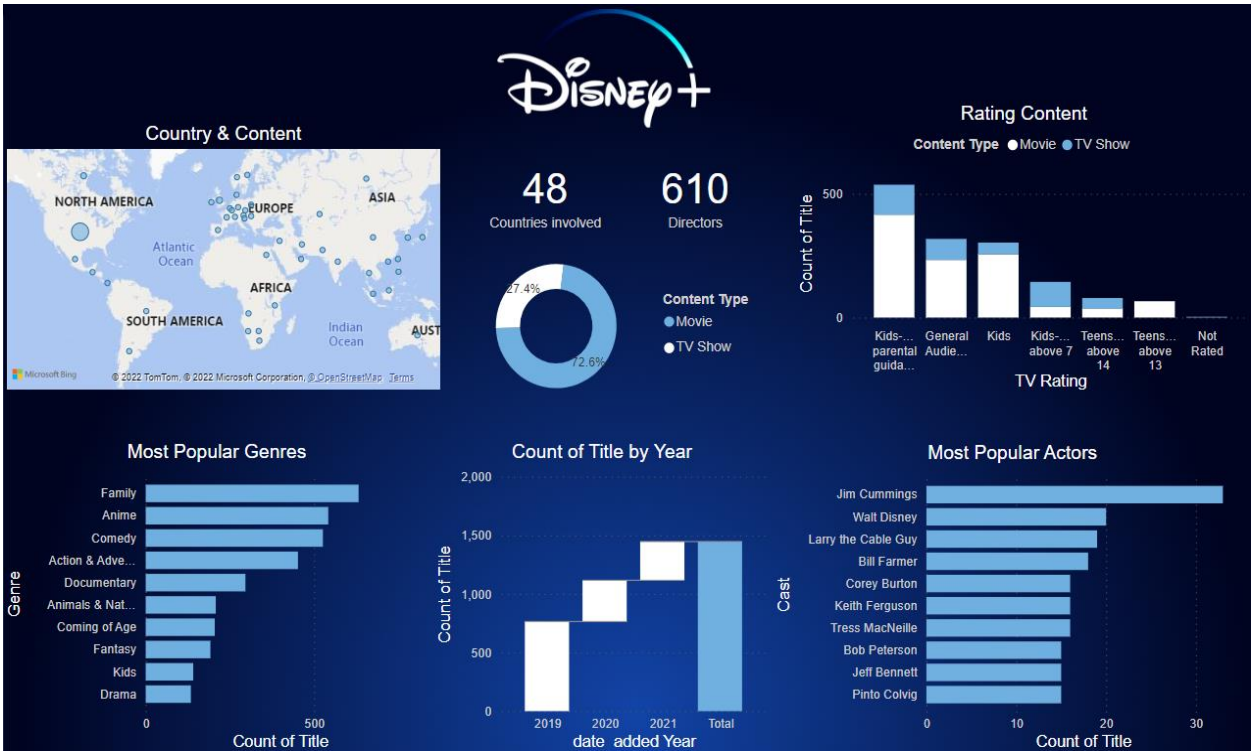
6. As International Content is most content available on Netflix, we can see from the top 10 most popular actors 9 are Indian actors, concluding Bollywood produces more multi-starrer movies.
7. United States contributes the most to the content on Netflix, followed by India and UK

2. Amazon Prime

1. 46 countries contribute to the content of Amazon Prime with United States contributing the most followed by India.
2. Amazon Prime contains almost 80% movies of total content and TV Shows just 20%.
3. Amazon Prime has a balanced distribution in terms of age rating. It has a roughly similar number of films categorized as Adults, General audience, and kids (parental guidance recommended).
4. From Most popular genres we can see drama made up the biggest proportion with comedy and action following.
5. Amazon Prime contains content which very released recently. The amount of content is increasing significantly which each release year as the content produced each year is also increasing.

6. Amazon Prime contains most content starring Maggie Binkley, followed by content starring Gene Autry and Champion.

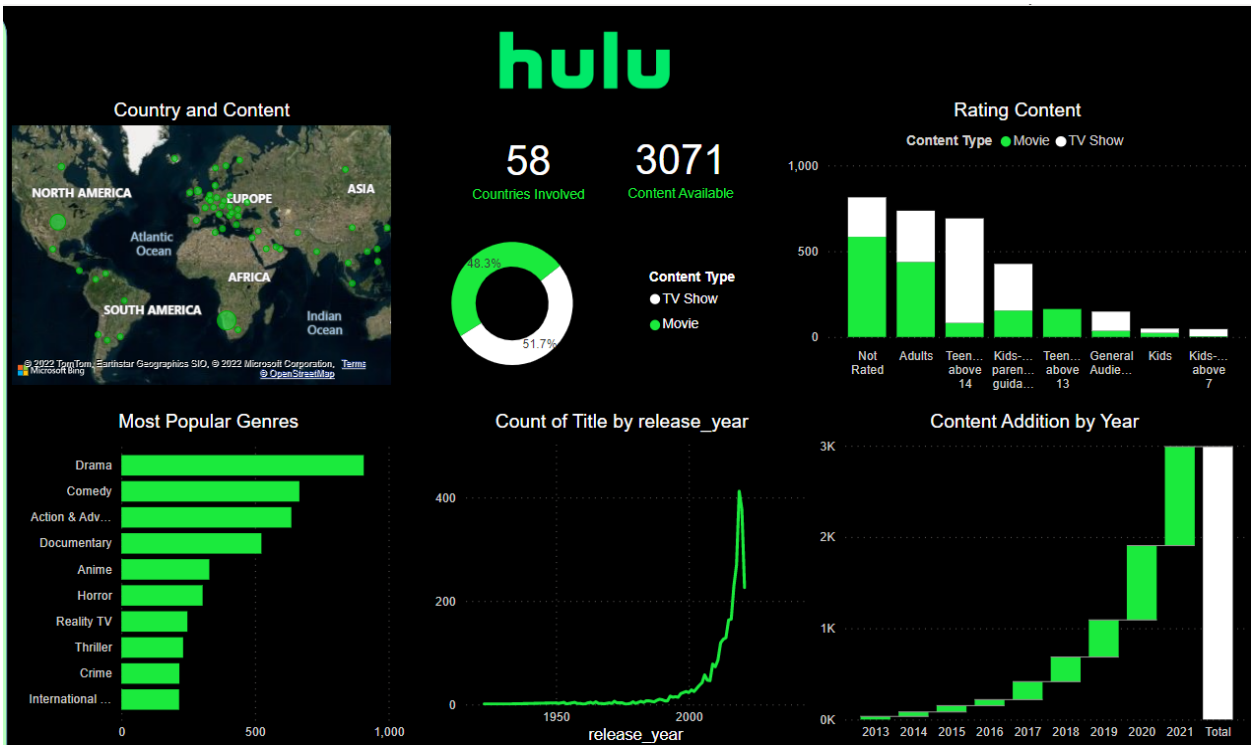
3. Disney+



Disney+ was launched in 2019, so it has very less content.

1. 48 Countries contribute to the content available on Disney+ with United States Most contributing, followed by UK and Canada.
2. About 3/4th of content on Disney+ is Movies, while TV shows are just 1/4th.
3. As Disney+ was launched in 2019, it added a lot of content in that year and kept adding content in the following years.
4. Disney+ is famous for its animation and kid's content, which we can clearly see from the Rating content graph. It mostly has content for kids and teenagers. We can see general audience content is available but adult movies are not at all there.
5. Being kids OTT platform, the genre of most of the content is Family, Anime, and Comedy.

4. Hulu



1. Movies and TV Shows contribute almost 50-50
2. 58 countries contribute for content on Hulu with United States contributing the most, followed by Japan and United Kingdom.
3. Hulu contain lot of Adult and Kids content but very less general audience. It has lot of TV shows for teens above age 14.
4. We can clearly see amount of content added on Hulu is increasing significantly every year.
5. The amount of content added on Hulu is according to release year. It contains most content which were released recently but it also includes content which very released in early years of cinema production.
6. Hulu contains drama content the most followed by comedy and action content. It also includes documentaries and international content.

Conclusion

Disney+ showed the most characteristic distribution. Yet, it had much smaller number of contents compared to the other two, which could be a strong push factor. It is still important to look at the shape of distribution because it signifies what to expect in the future. Disney+ clearly seems to focus on family, comedy, adventure, and animation films. It also has some action, drama, documentary, and fantasy movies but almost no movies in other genres. People who enjoy various genres may want to avoid Disney+.

1. Amazon Prime and Netflix contains almost same amount of content. Movies are still dominant over TV Shows but there is slight drop in movies content from 2019 but TV show contents are increasing consistently.

2. Generally, Netflix and Amazon Prime had a similar distribution in terms of genre; drama made up the biggest proportion with comedy and action following. However, Netflix still had more content across almost all genres with Amazon prime notably outnumbering Netflix only in action and romance films.

3. We notice that Netflix has an overwhelming count of films that are rated for adult audiences, and unsuitable for children under 17. It has very few films rated as general audience. It has around 2000+ films that are suitable for teenagers of age 14 and above. Amazon Prime has a balanced distribution in terms of age rating. It has a roughly similar number of films categorized as Adults, General audience, and kids (parental guidance recommended). Hulu also contain lot of Adult and Kids movies but very less general audience. Finally, although not noticeable because it has a small number of films overall, Disney+ has no content labeled Adults and a small number of films rated TV-14. All other films are rated either kids with parental guidance, which could be watched by children with or without parental guidance.

4. Netflix stands out when it comes to the number of countries contributing to the content, it has more than double the countries contributing than other platforms. United States contribute most for content on all platforms followed by India contributing most on Netflix and Amazon Prime and almost nothing on Hulu and Disney+.

References

1. Netflix Dataset: https://www.kaggle.com/datasets/shivamb/netflix-shows
2. Amazon Prime Video: https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows
3. Disney+ Dataset: https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows
4. Hulu Dataset: https://www.kaggle.com/datasets/shivamb/hulu-movies-and-tv-shows