**Netflix Data Analysis Report**

27.11.2022

**----------**

**Team Members:**  
**1. Tharun Sai Bhusa - 11608033**  
**2. Rohith Kompelly - 11608398**  
**3. Rithik Reddy Kunti - 11594218**  
**4. Nikhilesh Patil - 11629719**  
**5. Kavya Yerramsetty - 11594314**

# **Overview**

Media and entertainment has become an integrated part of the lives of people, meaning that people nowadays are very enthusiastic about trying new content in terms of watching it and choosing it. With the OTT platforms, it has become very convenient for people to choose and watch any kind of content at their comfort space and time. As people are widely using OTT platforms, especially Netflix, there is a lot of data being collected by this organisation which in turn is used to improve user experience by analysing patterns and recommending the best content that matches their taste.

As a part of this assignment, I have tried to dig into some patterns of Movies & TV shows from a subset of Netflix data. Understand the trends in movies/shows based on region, duration and ratings.

**Goals**

1. Understand the data and clean the data.
2. Impute the missing values in the dataset appropriately.
3. Find interesting patterns in the dataset.

# **Data description**

The dataset has following columns :

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl.no** | **Column Name** | **Description of column.** | **Data Type** |
| 1 | show\_id | Unique ID given to identify each title. | String |
| 2 | type | Categorical column mentioning whether it is a movie or a show | String |
| 3 | title | Name of movie or show | String |
| 4 | director | Name of the Director | String |
| 5 | cast | Actors in the movie/show | String |
| 6 | country | Country where content was made | String |
| 7 | date\_added | The date when the movie/show is added on Netflix | String |
| 8 | release\_year | The year when the movie/show was released. | Integer |
| 9 | rating | Rating given by the MPAA | String |
| 10 | duration | Duration of movie/show | String |
| 11 | listed\_in | Genres under which the movie/show is listed | String |
| 12 | description | Short summary of the movie/show | String |

The dataset has information about 1500 titles.

# **Data Cleaning**

The data given has 12 columns out of which 11 columns are strings, but the underlying data type has to change understanding the data description.

|  |  |  |
| --- | --- | --- |
| **sl.no** | **Column Name** | **Changed data Type** |
| 1 | cast | List of strings |
| 2 | country | List of strings |
| 3 | year\_added, month\_added, date\_only\_added | Get year,month & date from date\_added column into new columns |
| 4 | duration | Integer (number of min/ seasons) |
| 5 | listed\_in | List of strings |

# **Imputing Missing Values**

The dataset has missing values also known as null values in a few columns. Here are the details regarding null values in the dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl.no** | **Column Name** | **Number of Missing values** | **% Missing values** | **Imputation strategy** |
| 1 | director | 480 | 32% | N/A |
| 2 | cast | 152 | 10.13% | Drop values. |
| 3 | country | 360 | 24% | Impute using director and cast columns. |
| 4 | release\_year | 30 | 2% | Impute using date\_added column. |

## director

Since the director column is not a definitive list of objects meaning the set is not limited, we choose to ignore this column and also the missing percentage is relatively very high in this column, so we can’t drop the values.

## cast

For the cast column again the column is not definitive, but the missing percentage is relatively less, so we can drop those values while analyzing data using that column.

## country

For the country column the missing percentage is high so we can’t drop the values but we can use director & cast columns to fill in the country values. The strategy is to fill in country value from an instance where the director or cast matches.

## 

## 

## released\_year

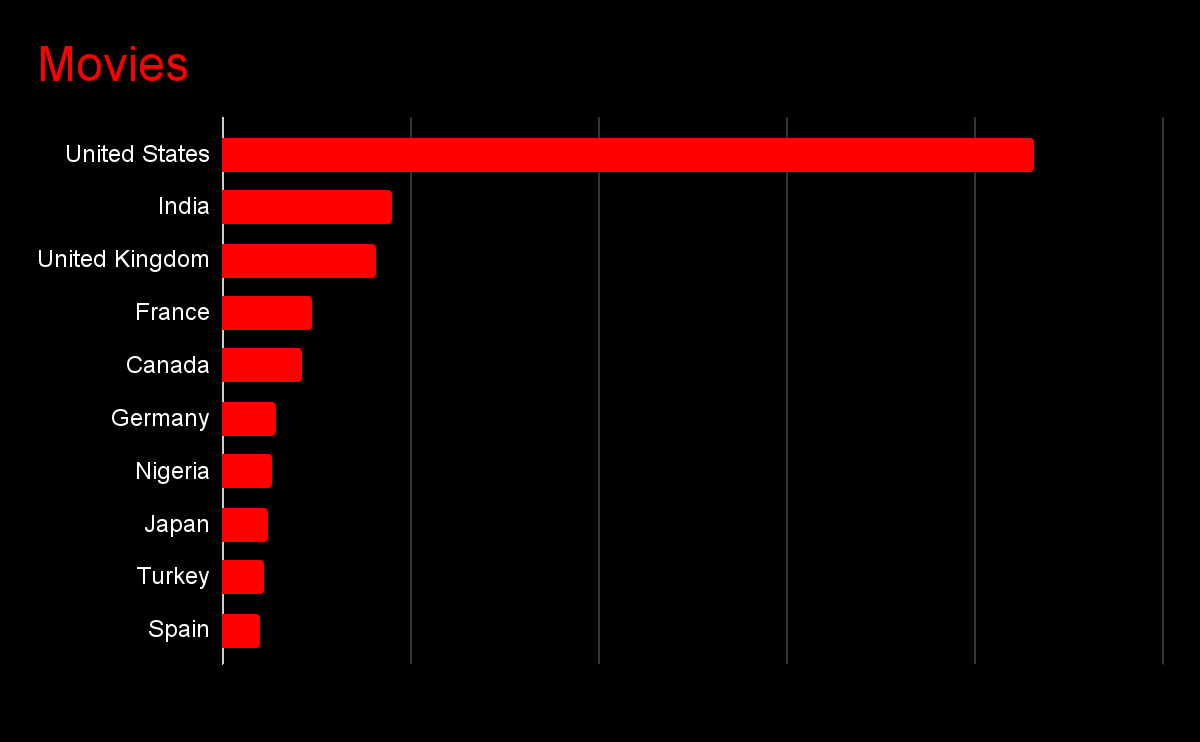
For release year the missing percentage is very less(2%), we can use mode imputing strategy. That is imputing the missing values with the most repeated value in the column.

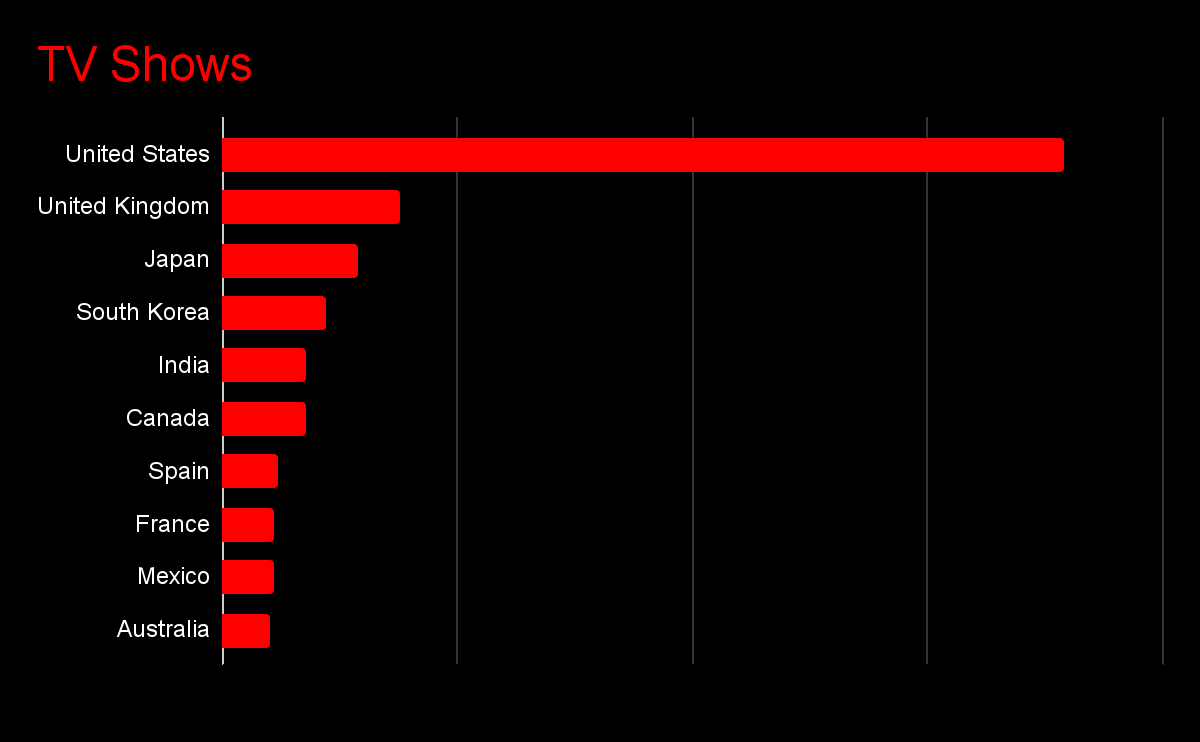
# **Finding Interesting patterns in the data**

The dataset has 1500 unique titles, filmed across 73 countries with over 9641 actors in 42 genres, where 985 of them are movies & 515 of them are TV shows released over the past 53 years..

## What are the top countries that produce content

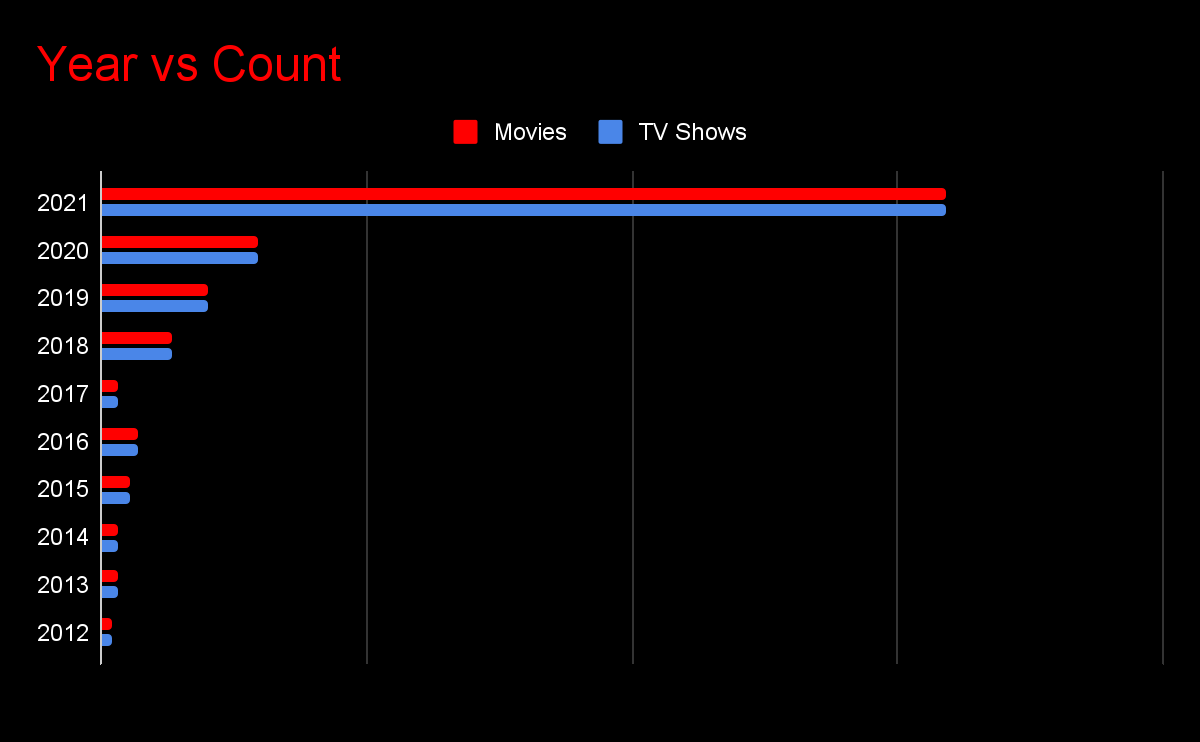
The United States tops in producing content in both Movies and TV Series. But Second position for movies is taken by India & for TV shows by the United Kingdom. We can see in the following plots the top ten countries producing movies and TV shows. Few countries produce more content through movies compared to TV shows and few countries vice-versa.

For example France,Germany,Nigeria and Turkey focus on movie content whereas South Korea,Mexico,Australia focus more on TV shows. 



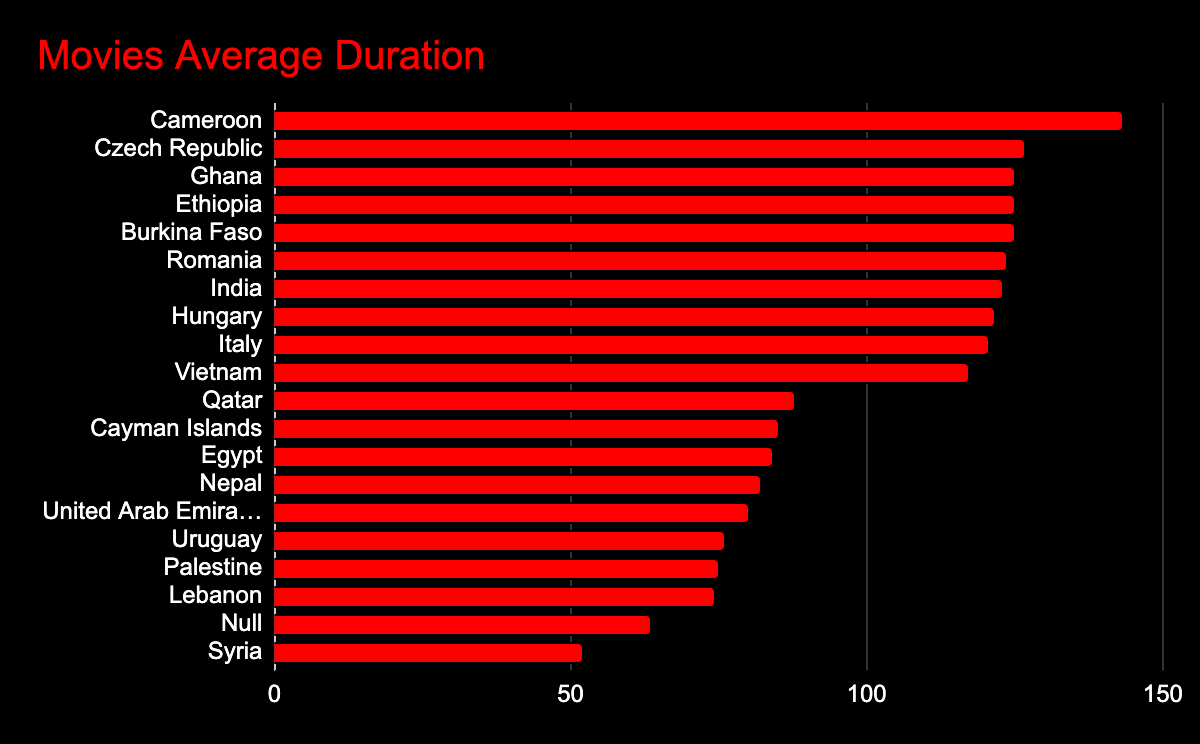
## How did content progress with time

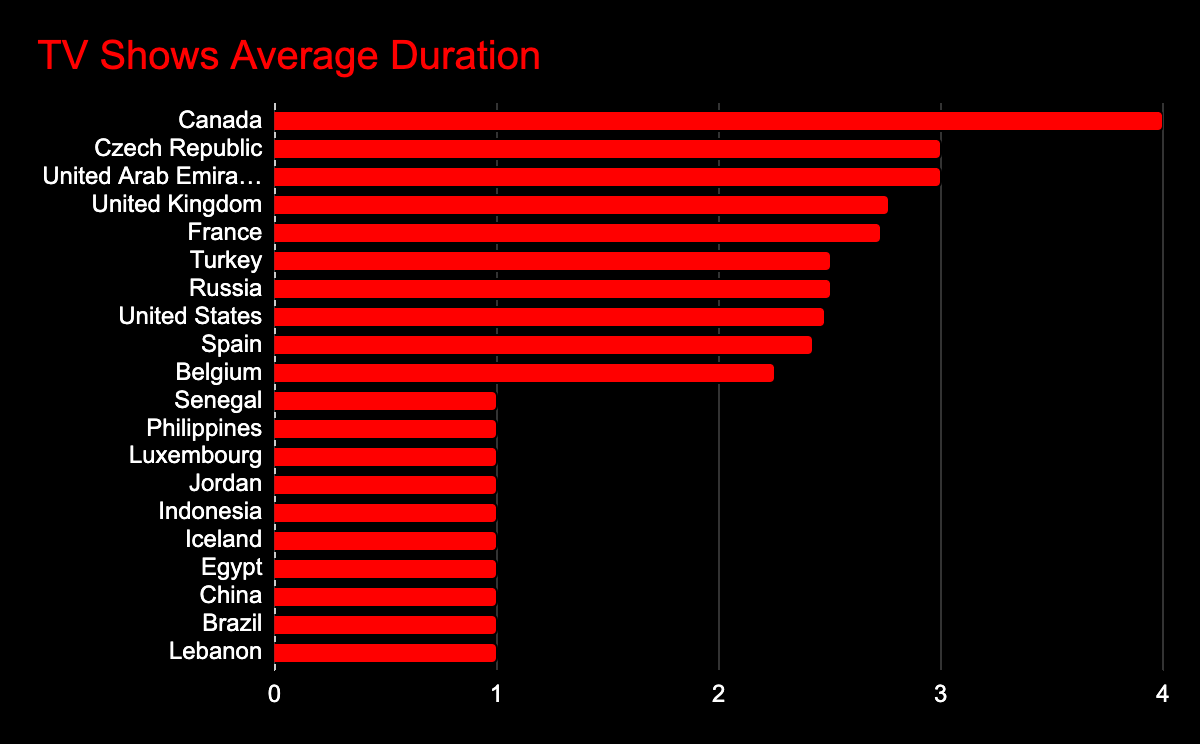
The trend shows that the content produced in both movies and TV shows has been increasing exponentially in the past 20 years. In 2021 the number of TV shows released crossed the number of movies, showing in future TV shows could go a long way. Even though movies have existed for a very long time, TV shows started in the 2000's have surpassed movies.



## What is average watch time across different countries

The average watch time varies across countries , from the following graphs we can see top 10 & bottom 10 movies and TV shows according to average duration. This trend could be explained as few countries have good profits from the entertainment industry and are thus encouraged and few countries have restrictions on the content they produce, so it’s tough to increase the duration in those countries.

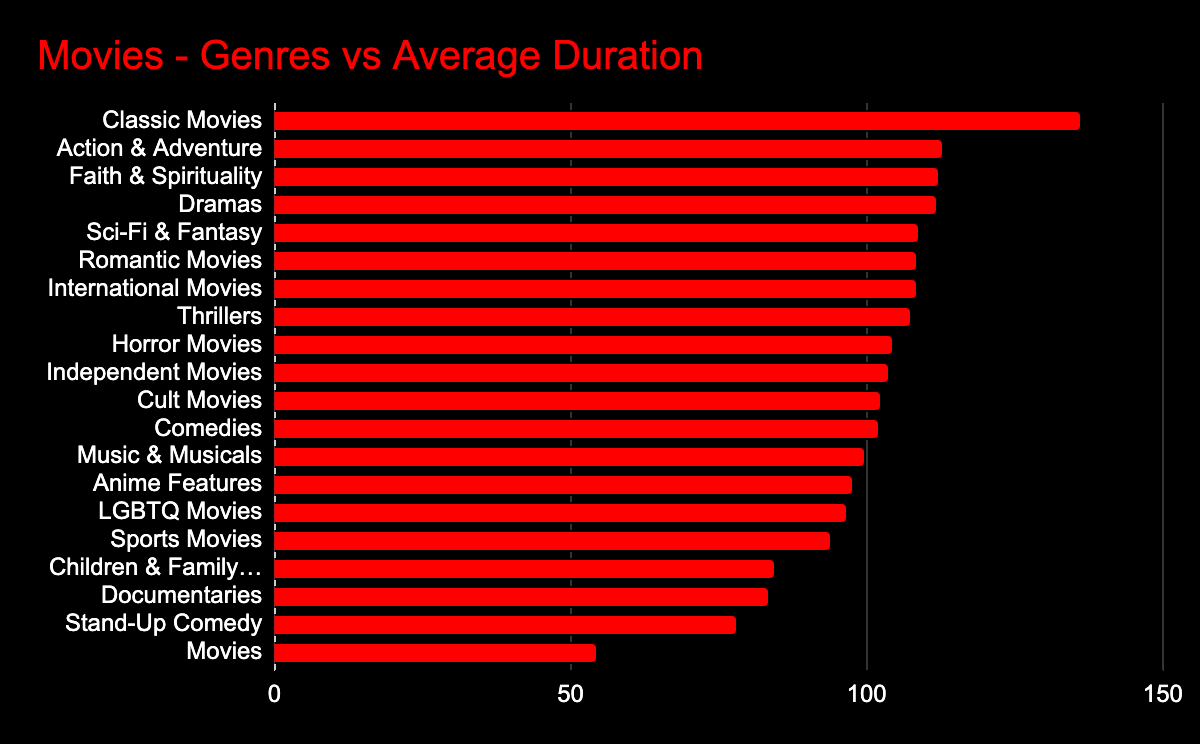


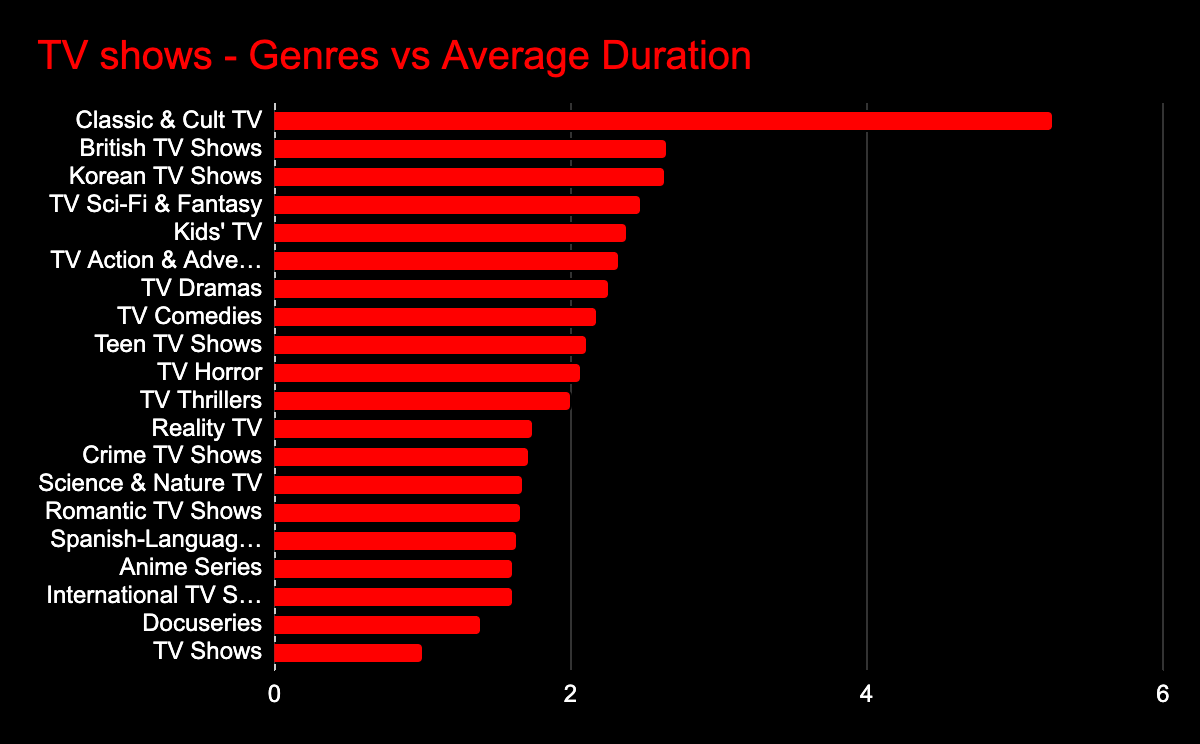


## What is average watch time across different genres

In movies Classic genre, Action & Adventure have more average duration compared to comedy,Animes & documentaries.

In TV shows Classic & cult, British & Korean TV shows have more duration compared to Thrillers & Reality TV shows.





## How did different genres dominate different times

Comparatively Thrillers, Children & Family Movies have been produced in 2000’s compared to 1940’s when Action & Adventure, Classic Movies & Dramas have dominated. This helps us understand how the audience's tastes have shifted in the past 60 years.

Similarly in TV shows Kids TV dominated last century whereas in 2000’s Anime Series, TV Dramas & Teen shows are more.

## How did average watch time change as time progressed

There is no particular trend in the duration over the last 10 years, it is almost around 90 min ± 10 min, compared to 1940-1980 where the duration is almost around 110 min ± 10 min. So we can say movie durations have been reduced in the last 60 years.

For TV shows the average length is 2 seasons in the last 10 years, but in the past there are far less TV shows with utmost 1 season. TV shows have started late after the movies started. So this makes sense.

## Is there a relation between genre and rating

R rated movies are mostly from Horror,Action & Adventure, Classic & Cult genres. Sci-Fi & fantasy and Sports movies are mostly PG-13 rated in both movies & TV shows.

## Is there a relation between country and rating

There are 10 types of ratings for movies and 6 types of ratings for TV shows. Since most of the content is from the US, a maximum number of shows or movies with different ratings have been produced by the US, except for TV-14(India), TV-PG(Japan) and TV-G(Indonasia) in movies.

## Is there a relation between country and genre

From 43 genres in the data, 20 are for movies and 23 for TV shows. The content for large number of these genres is dominantly produced in US, except for few genres like :

1. International Movies - India
2. Anime Features - Japan
3. International TV shows - Japan
4. Anime Series - Japan
5. Stand up Comedy talk Shows - South Korea
6. Classic & Cult TV - Argentina

There are also some regional genres, specific to countries like

1. British TV shows - United Kingdom
2. Spanish Language TV shows - Spain
3. Korean TV shows - South Korea

## How many movies/TV shows have been listed in more than one genre

There are 1156 (397 TV Shows & 759 Movies )listed as more than one genre, that is around 77% of content belongs to multiple genres.