Netflix Data: Cleaning, Analysis and Visualisation

By Manuel Cabrera

Netflix is a streaming service that offers a diverse catalogue of movies and TV shows; this data (obtained from Kaggle) consists of content that Netflix has added to its service from 2008 to 2021.

This project aims to clean the data using Python (Jupyter Notebook), analyse and perform visualisation on the dataset provided.

1.0 Discussing the dataset

This dataset consists of the following variables:

- **show_id**: Netflix ID of the media.
- **Type**: Movie or TV Show.
- **title**: Title of the media.
- **director**: Director of the media.
- **country**: Country in which the movie was made.
- **date_added**: Date in which the media was added.
- **release_year**: Year in which the media was released.
- **rating**: Age rating of the media.
- **duration**: Duration of the media.
- **listen_in**: Classification given by Netflix.

2.0 Libraries and gathering of the data

For this project the only libraries that were used were: *pandas, numpy, seaborn* and *matplotlib*.

Since the data was given on a comma separated value (csv) format, this was easily brought as a dataframe using *pandas*

3.0 Assessing

There were 4 quality and tidiness issues with the dataset:

- Variable ‘date_added’ had the wrong data type.
- Variable ‘duration’ had the wrong data type.
- Variable ‘listed_in’ contained several variables per observation.
- There are two types of observations, TV shows and movies.

4.0 Cleaning

This is how the data was cleaned:

- **Variable 'date_added' had the wrong data type** – using pandas we changed 'date_added' data type to 'datetime' data type.
- **Variable 'listed_in' contained several variables per observation** – each of the variables was extracted and added into a separate variable.
- **There are two types of observations, TV shows and movies** – the dataset was split into two groups, observations on movies and observations on TV shows.
- **Variable 'duration' had the wrong data type** – after the dataset was split into TV shows and movies, the 'duration' variable was extracted as an integer.

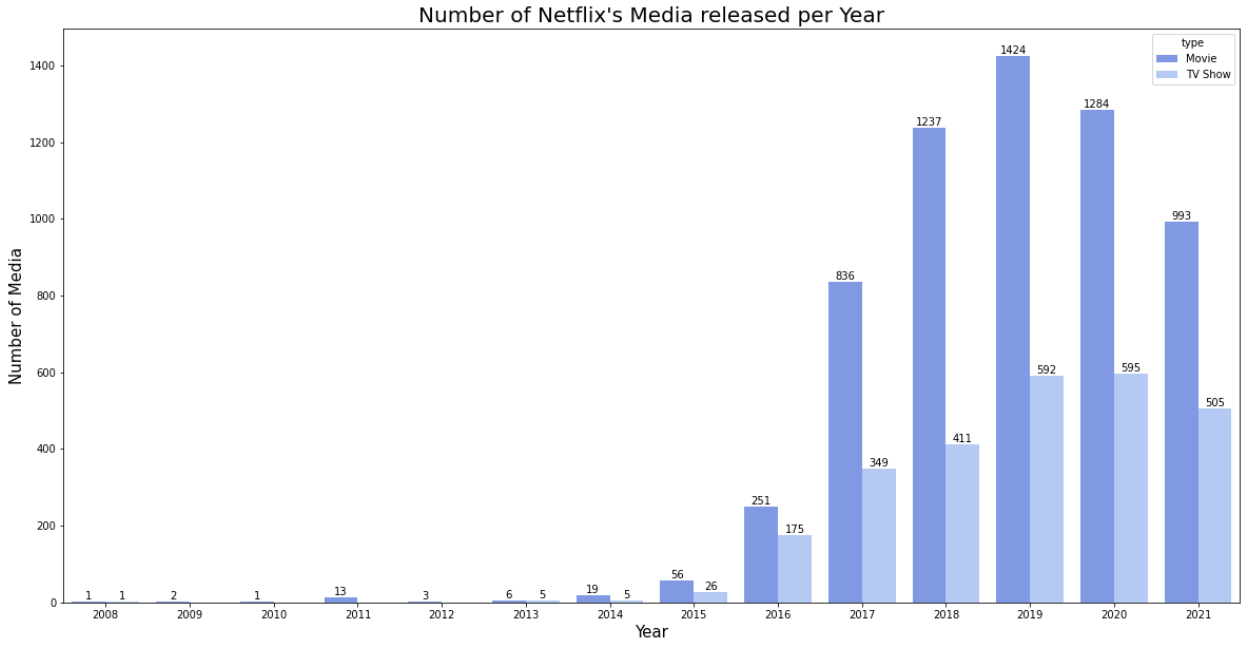
5.0 Analysing and Visualising the data

Prior to generating the visualisations, we set the following questions:

5.1 What type of media has Netflix produced the most?

For both TV shows and Movies there has been a steady increase since the start of 2008; the only big drop happening in 2021 possibly due to the economic impact of COVID.

Before 2017, the number of TV Shows and Movies brought to the streaming service was on par. However, after 2017 the company started introducing more movies into the service more than doubling TV Shows in the amount of content.



5.2 Which country produced the most of Netflix's media? And what are the most popular genres for countries that produced media?

TV Shows

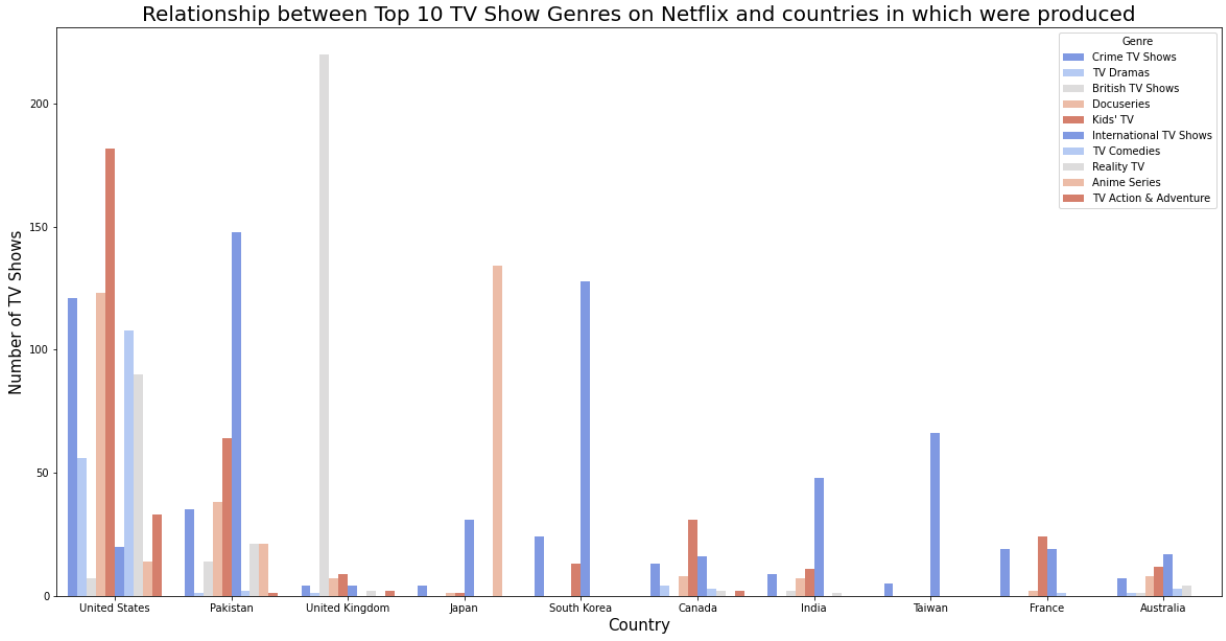
The top 5 countries that produced most of the TV show content were the US, Pakistan, the UK, Japan, and South Korea.

The most popular TV genres overall seem to be Kid's TV and International TV Shows. Considering that Netflix is an American company, it makes sense that shows produced outside of the US are considered 'International TV Shows.

US has a big diversity of shows produced but most of them were TV Action & Adventure, followed by Docuseries and Crime TV Shows.

For Pakistan and South Korea, both the most produced genres were Kid's TV and International TV Shows.

However, the UK and Japan have the biggest production of Reality TV shows and Anime series respectively.



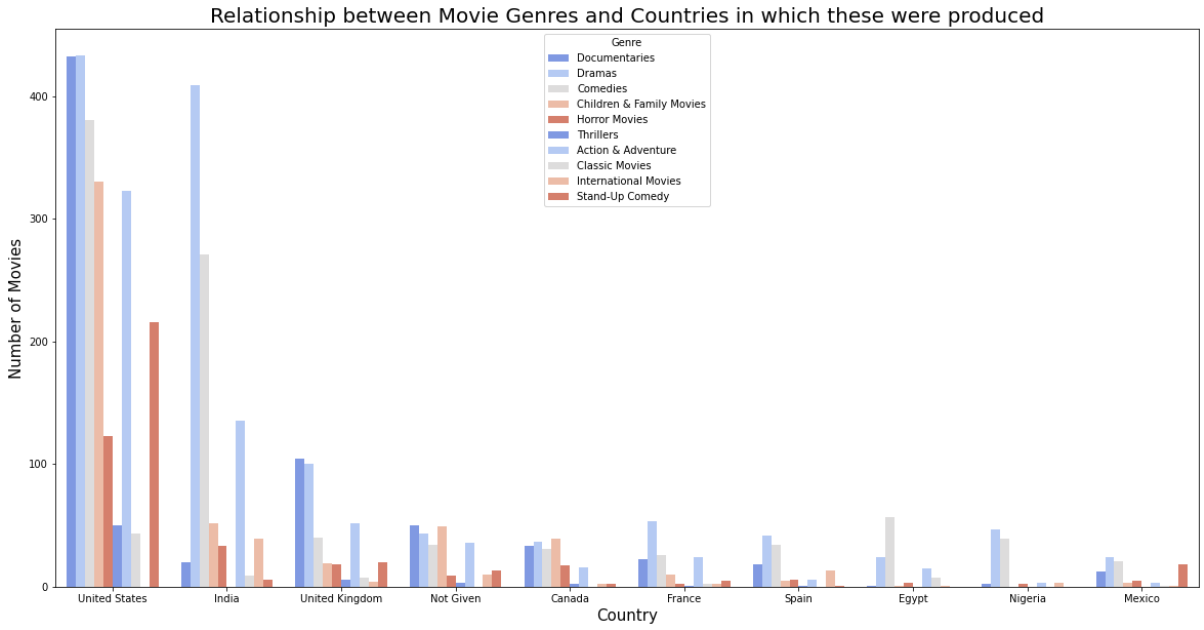
Movies

The top 5 countries that produced most of Movies were the US, India, the UK, not given, and Canada. The most popular movie genre overall seems to be drama and comedy.

Most of the media produced in the US were documentaries, dramas, and comedies. Whilst for India, it was mainly dramas followed by comedies.

The US and the UK have one of the biggest ratios in movies produced against documentaries.

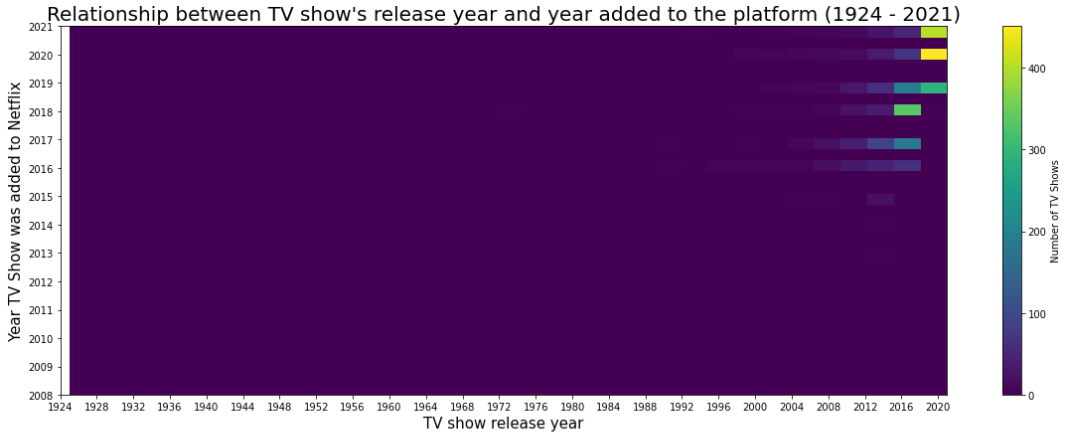
Furthermore, it seems that Stand-Up Comedy is also a big genre in the US; when compared to other countries.

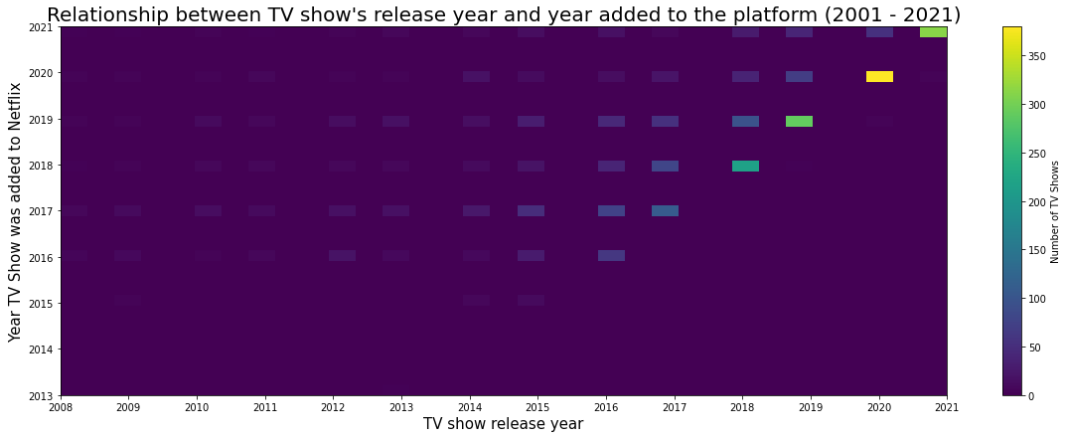


5.3 What is the relationship between the year a media was made and when added to the Netflix platform?

TV Shows

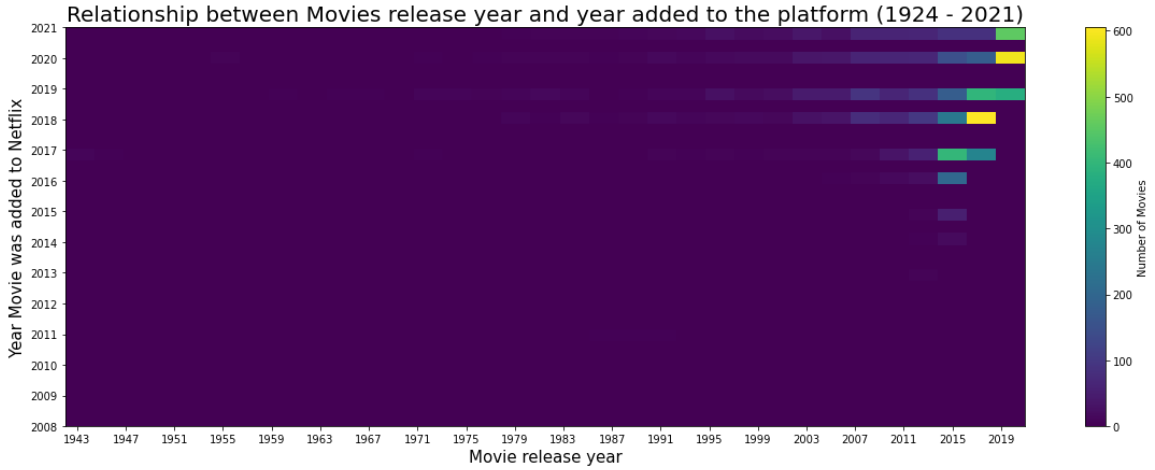
Initially, Netflix would not bring recently produced TV shows into the service. It was up until 2016, in the service started producing its own TV shows and bringing recently filmed TV shows into its streaming service.

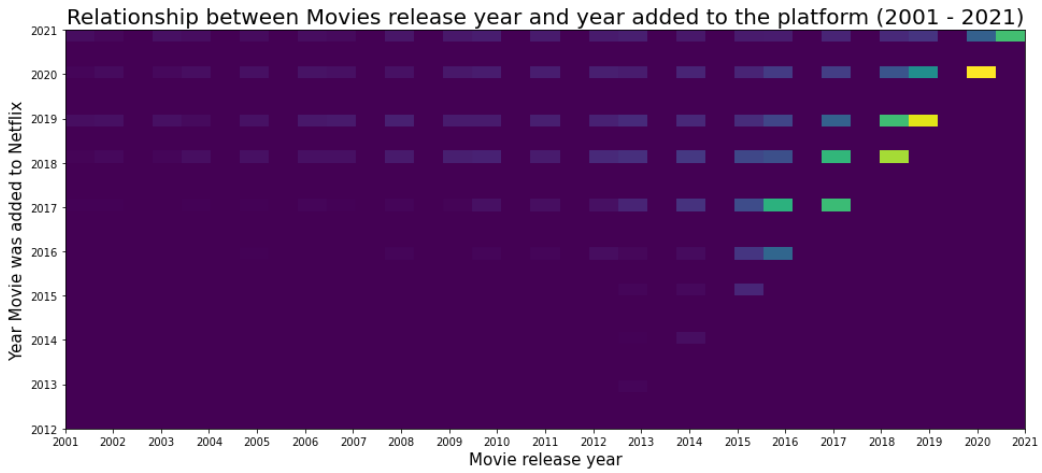




Movies

Starting from 2015, there seems to be a clear increase in the number of movies released in the same year as they were added to the platform. There is also a slight decrease in 2021 but this would have been likely due to COVID - since not many movies were made in 2020 and released in 2021.



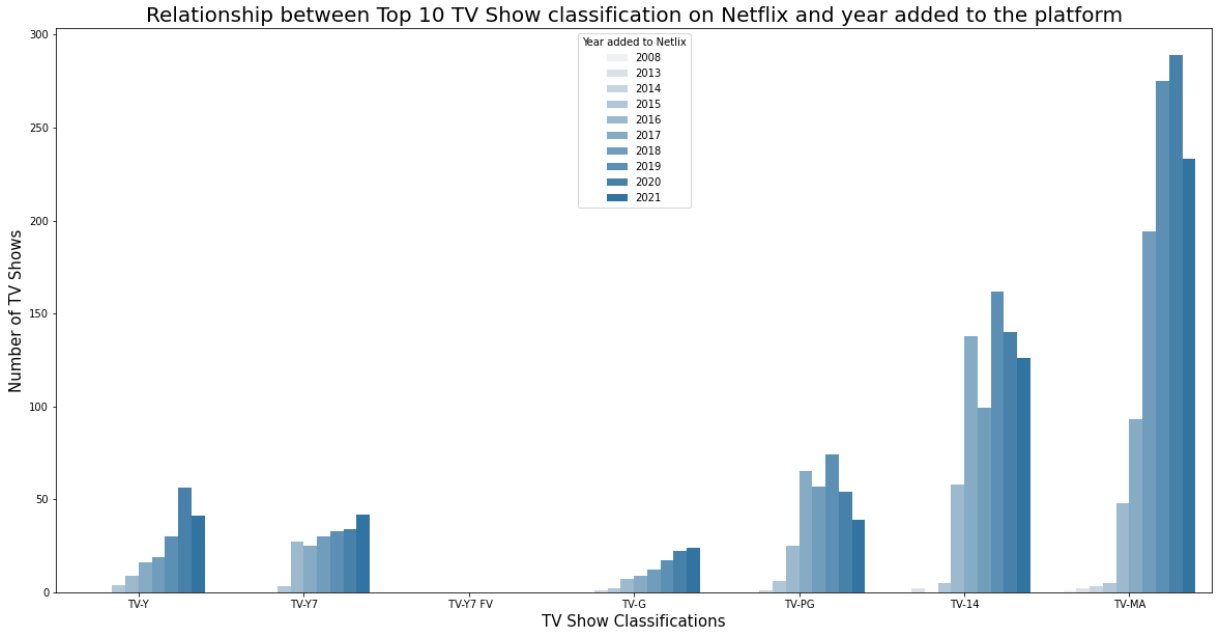


5.4 Has Netflix's media classification changed over time?

TV Shows

Most of Netflix's TV shows are categorised as TV-MA (primarily) and TV-14 (secondarily). TV-MA had been steadily increasing up until 2021; whilst in the other hand TV-14 has had a sporadic growth with a decrease starting from 2020 - This possible due to TV Shows pushing their classification to a more mature audience (TV-MA).

The only shows that had a steady increase with no drops whatsoever were the ones classified as TV-G and TV-Y7.

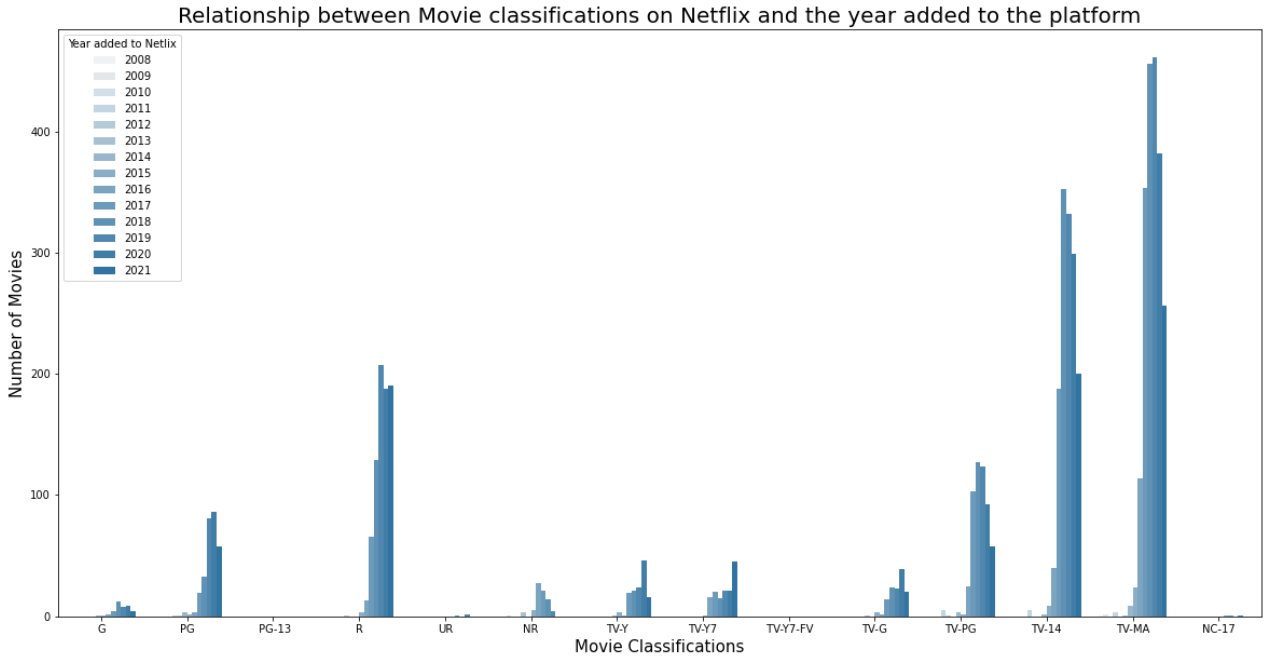


Movies

Movies had both movie classification and TV classification meaning that some of the movies that are part of Netflix were movies made for straight to TV.

There seems to be that most of the movies available on Netflix would be for mature-aged people (18+, R and TV-MA).

It seems that Netflix did not want to bring teen movies released in theatres (PG-13) to its platform but instead bring R-rated movies. This strategy is different to the straight to TV movies, since the second most popular classification would be TV-14.



5.5 What are the most popular genres for Netflix media?

TV Shows

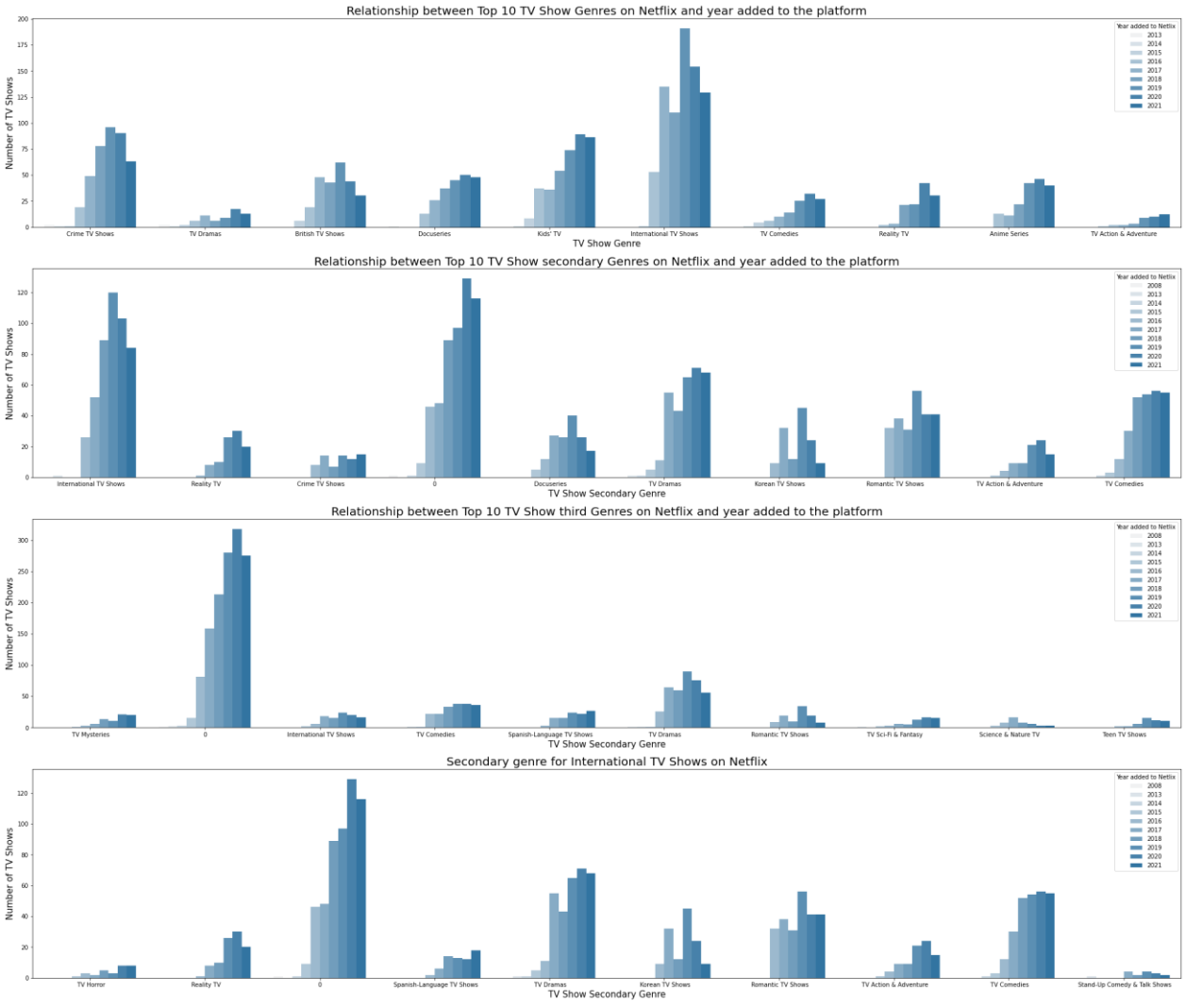
For this analysis an assumption had to be made, the first category that appears on a TV show would be classified as the 'main' category, followed by the secondary and the third.

Most of the TV shows produced were International TV Shows followed by Crime shows and Kids TV.

For secondary genres, most of TV shows do not have a secondary genre. But the rest were mainly classified as International TV Shows, TV dramas and TV comedies.

For third genres, Most of TV shows do not have a third genre. But the rest were mainly classified as TV Dramas, TV comedies and Romantic TV shows.

Since International TV Shows is a broad genre, we also performed an analysis on the secondary genres for International TV Shows. The result was that most of them do not have a secondary genre. But the rest were mainly classified as TV Dramas, Romantic TV shows and TV comedies.



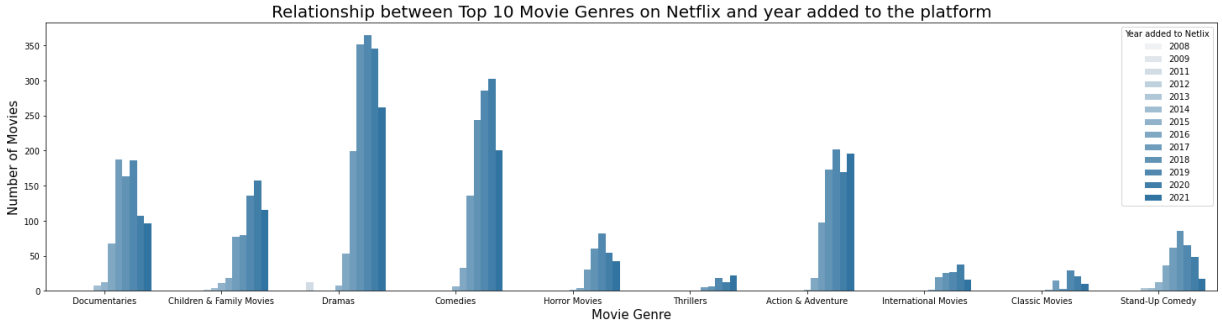
Movies

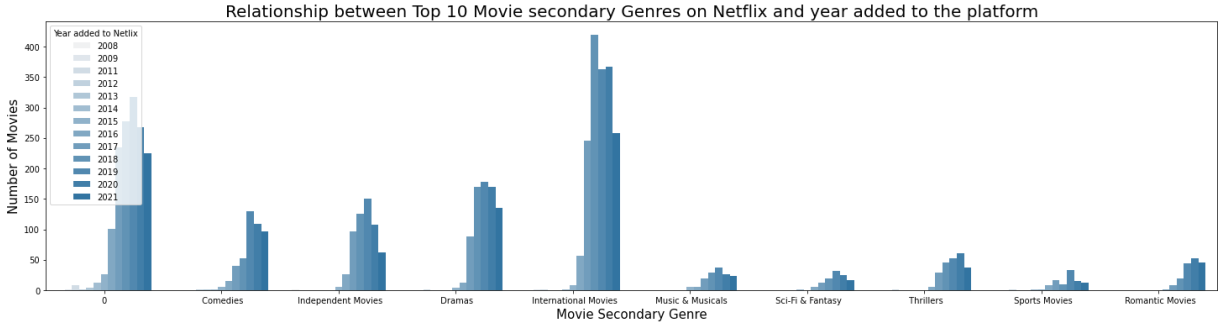
For this analysis an assumption had to be made, the first category/genre that appears on a Movie would be classified as the 'main' category, followed by the secondary and the third.

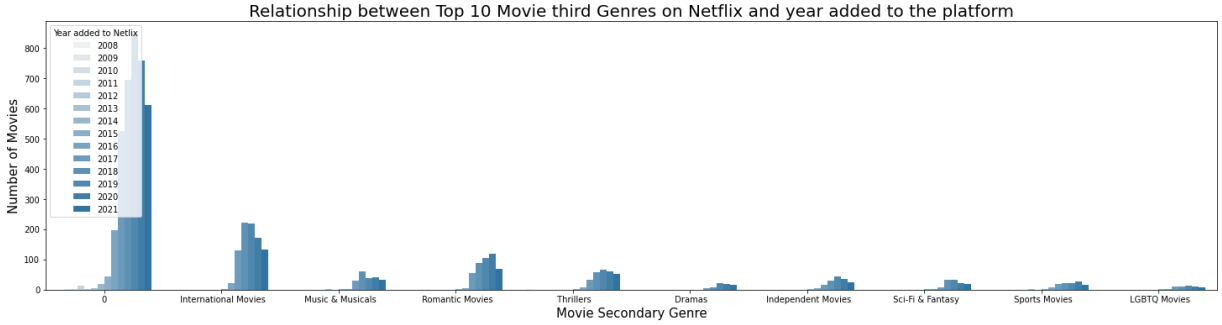
Most of the movies produced were Dramas, Comedies and Action & Adventure.

For the secondary genres, most of the TV shows are classified as International Movies. But the rest were mainly classified as None, dramas and independent movies.

For the third genre, most TV shows do not have a third genre. But the rest were mainly classified as international movies.



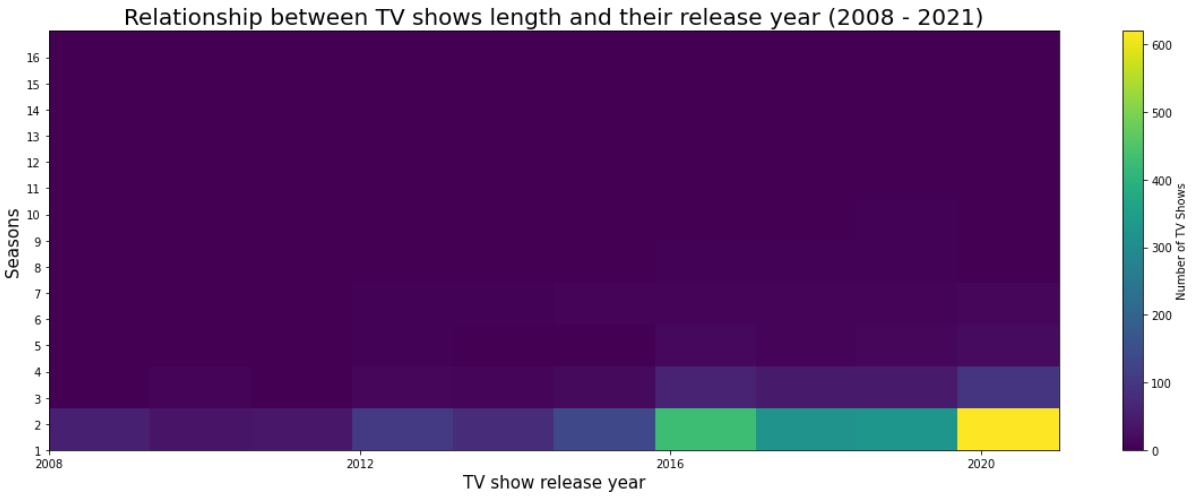


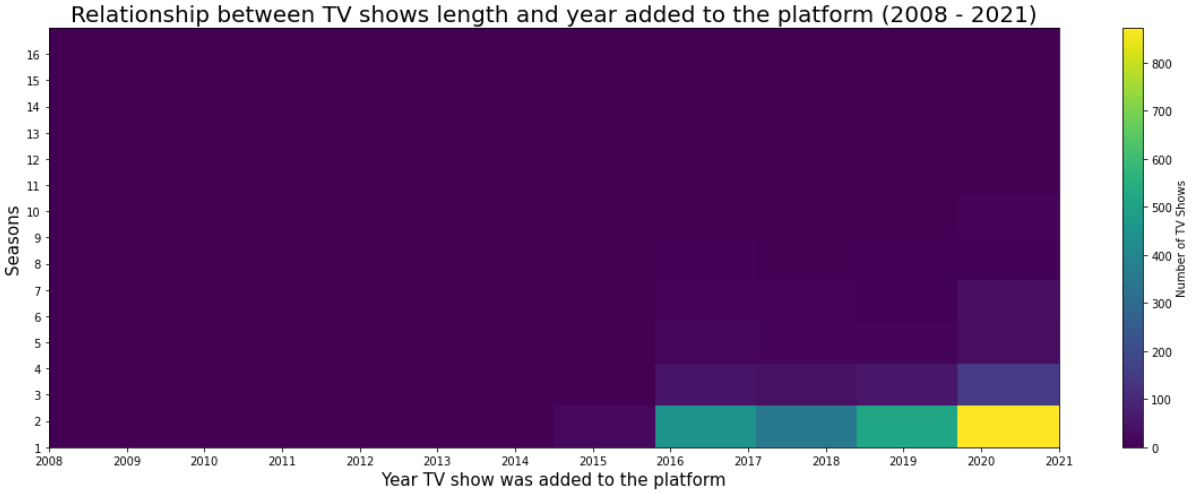


5.6 Has the length of TV seasons or Movie's length changed over time?

TV Shows

For both release year and year added to the platform, it seems to be that most of the TV shows just last for a single season.





Movies

For movies release year and length, most of the movie's length range between 70 and 120min, however overtime the length of the movies has decreased.

For movies length and year it was added to the platform, it follows a similar trend in which most of the movies last between 70 and 120min. However, it also seems like most of the movies seem to be gradually increasing in length.