Information Visualization Project

Students: Catalina Chirita, Andra Acsintoae

1. Dataset Description

Link to Data Set: https://www.kaggle.com/datasets/shivamb/netflix-shows.
Dataset Description: The dataset is composed of Movies and TV Shows on Netflix, one of the most popular media and video streaming platform. The Netflix dataset is composed on 8807 Movies and TV Shows listed on Netflix, that have been released since 1940s to the beginning of 2021. Each sample has different properties such as the type, the genre, the country on which it was produced, the date of the release, the rating or its description. Below we can take a look at the fields and samples presented on the dataset:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...
1	s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly...	United States	November 16, 2017	2009	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi...	In a postapocalyptic world, rag-doll robots hi...
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	2008	PG-13	123 min	Dramas	A brilliant group of students become card-coun...

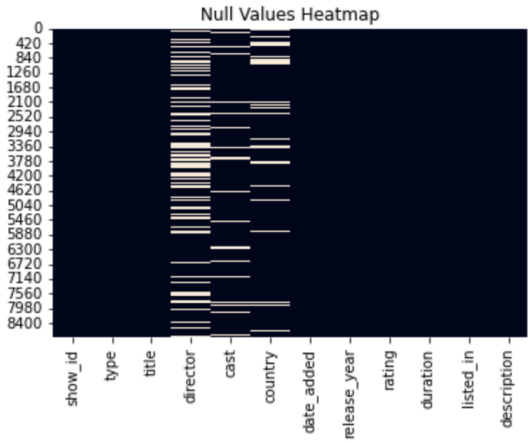
2. Dataset Preparation for Data Analysis and Visualization

In the cleaning part of the dataset, we removed the nan and null values. First, we take a look at the non-null values corresponding to each field out of the total of 8807 recordings (6131 recordings of type 'Movie' and 2676 and type 'TV Show'). Since we can see there are some null entries for some of the fields, we also plot a Null Values Heatmap in order to understand the impact those might have on the overall dataset.

```
pd_train.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
```

There are 8807 entries and 12 columns. There are a few columns that contain null values ('director', 'cast', 'country', 'date_added', 'rating')



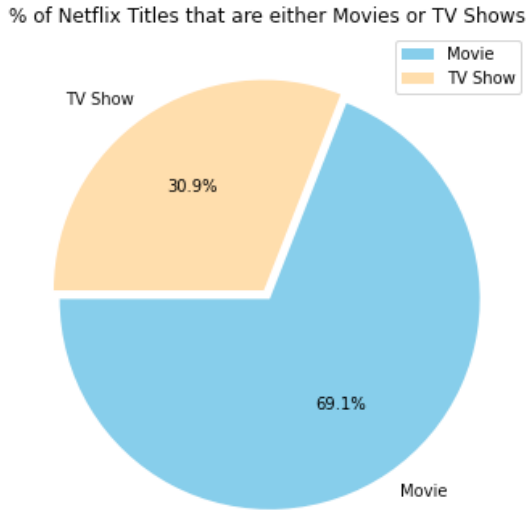
Above in the heatmap and table we can see that there are quite a few null values in the dataset.

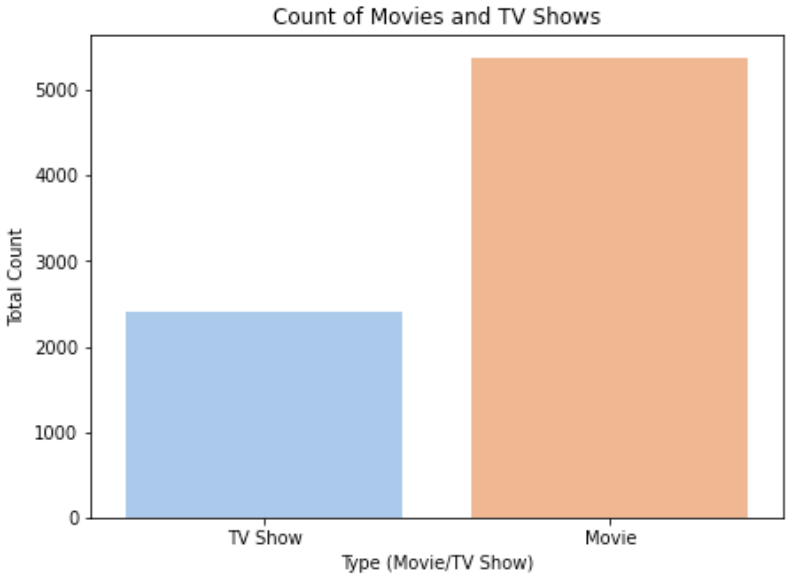
Additional to removing the null values, in the processing part we also create two new data frames derived from the initial one split by type, thus resulting an overview of TV Shows and another one of Movies. Besides under the overall content on Netflix, we will also compare during our analysis the properties of Movies versus TV Shows.

3. Exploratory Data Analysis

3.1. High Level Understanding of the Dataset

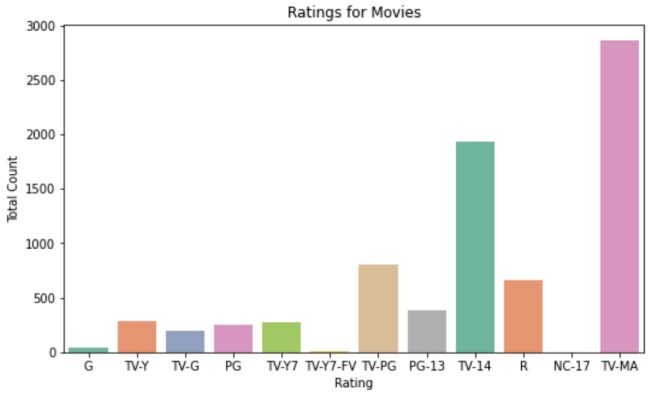
For the high level understanding of the dataset, we wanted to understand what are the discrepancies between Movies and TV Shows in terms of content present on Netflix. For this step we have used both pie plot and box plot and, even though the assumption was that most people rely on Netflix for its variety of TV Shows, we can see in the plots below that Movies outperform TV Shows in terms of number.

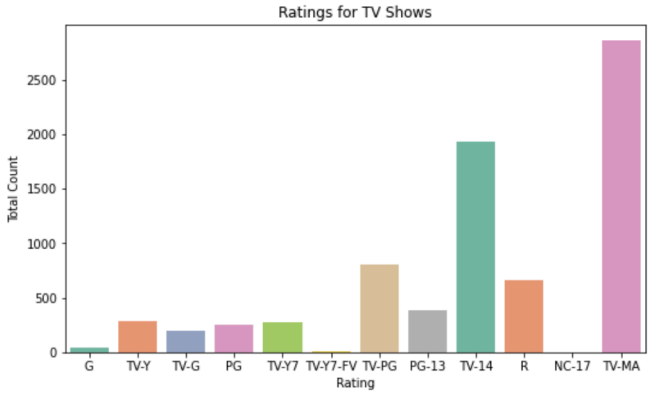


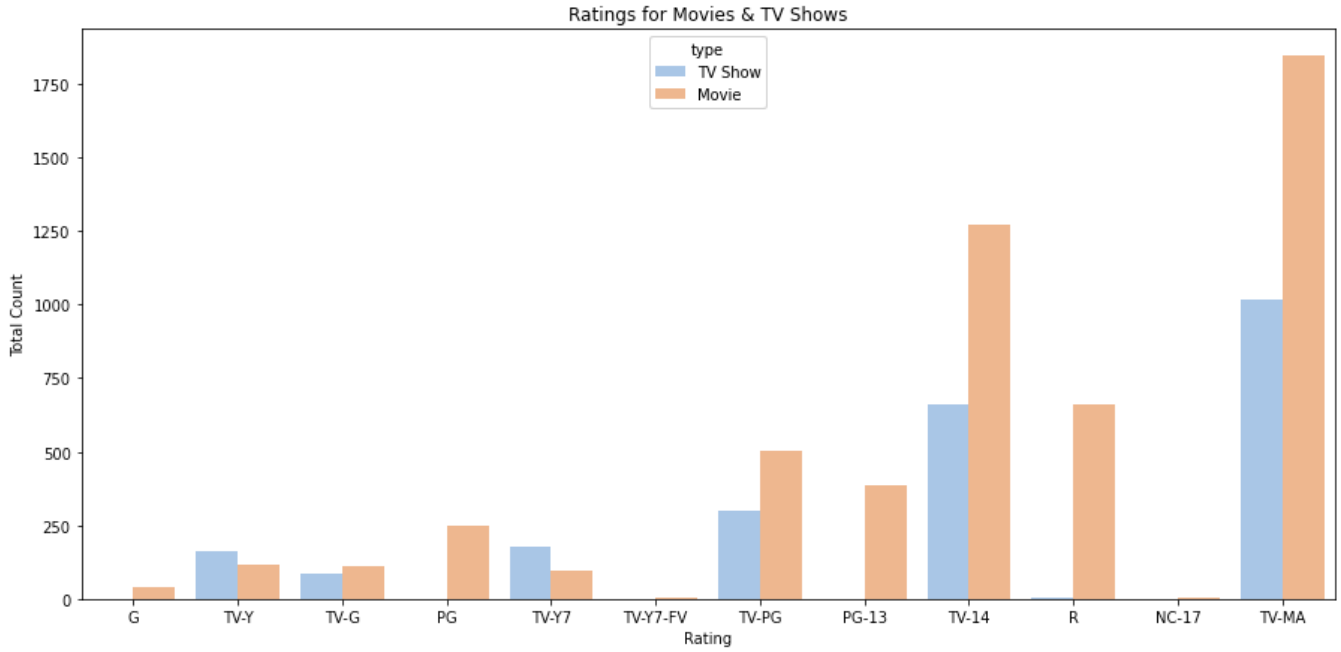


3.2. Low Level Understanding of the Dataset
Netflix Ratings

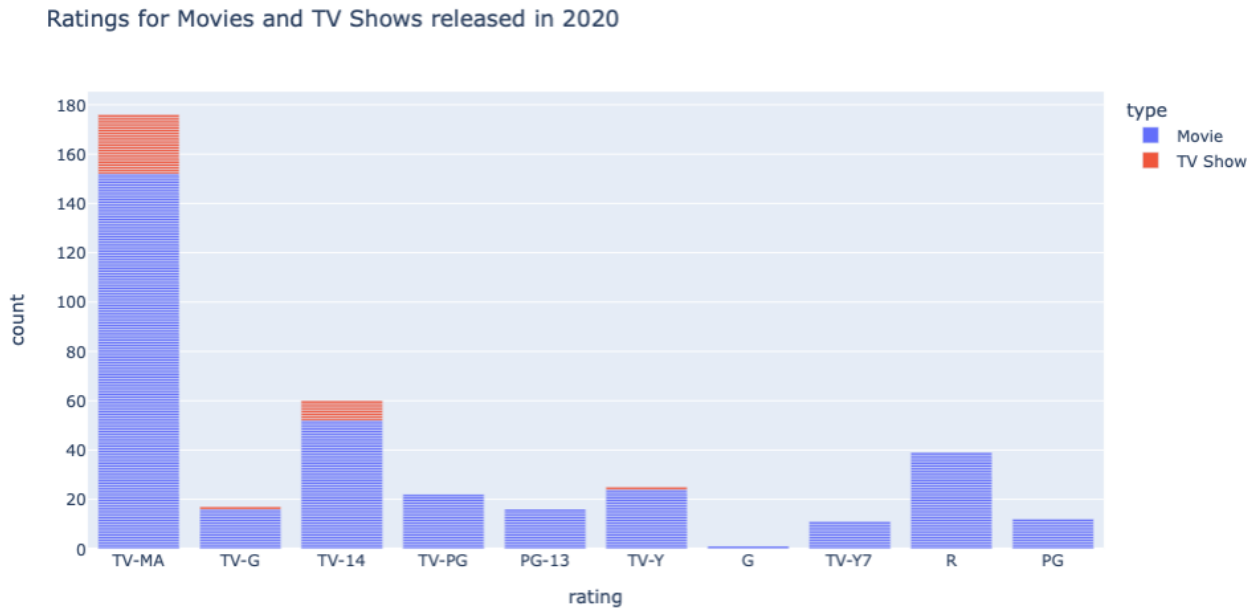
On this part we have analysed the ratings on both Movies and TV Shows, where the ratings were based on the film rating system and fell under one of the following categories: 'G', 'TV-Y', 'TV-G', 'PG', 'TV-Y7', 'TV-Y7-FV', 'TV-PG', 'PG-13', 'TV-14', 'R', 'NC-17', 'TV-MA'. Each of the categories is representative to the audience the content targets and the ranking will be based on the age of the respective members from the youngest to the oldest.





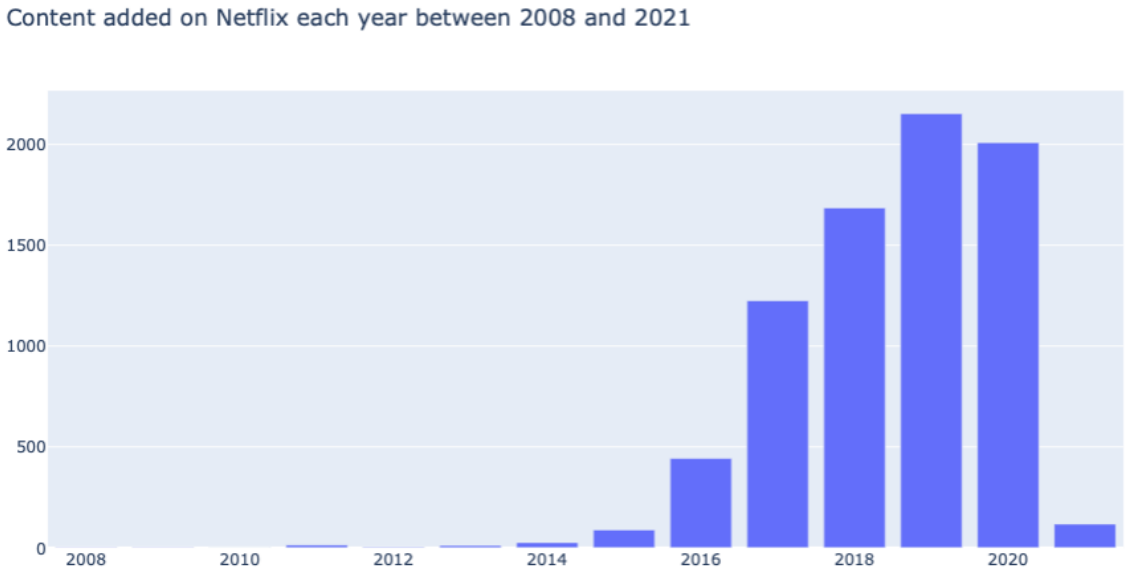


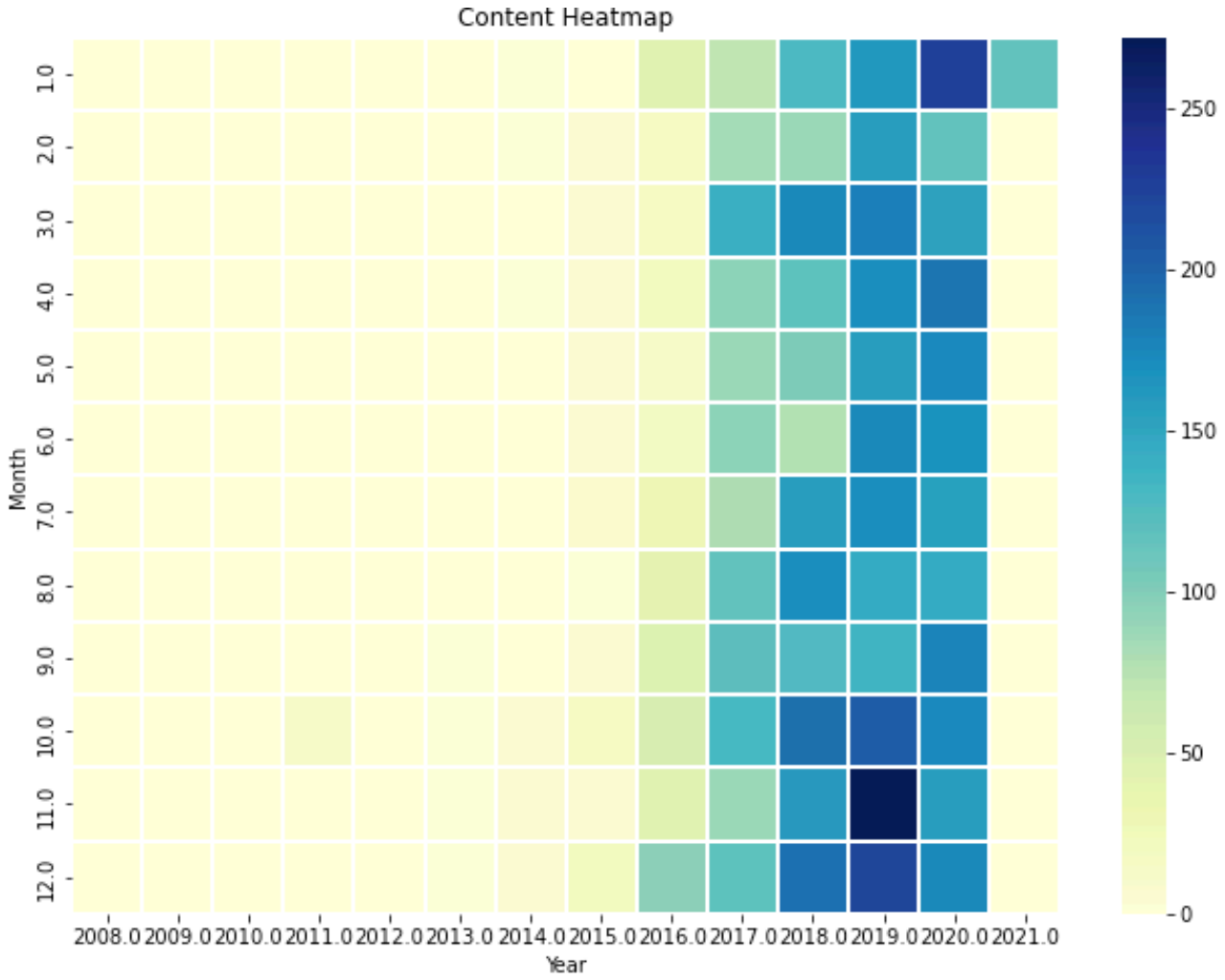
The histograms above represent the number of movies and TV Show rated under each category through the history. As we can see, there is close to no incline towards creating content for the General Audience (G), as in general it is not recommended to target a broad audience. On this note, the preference lies under creating content for the adult audience or for Mature Accompanied (MA) when children are below a certain age. The preference seems to be proportioned amongst both TV Shows and Movies, and it was kept, as we can see on the interactive histogram below, thorough out time.



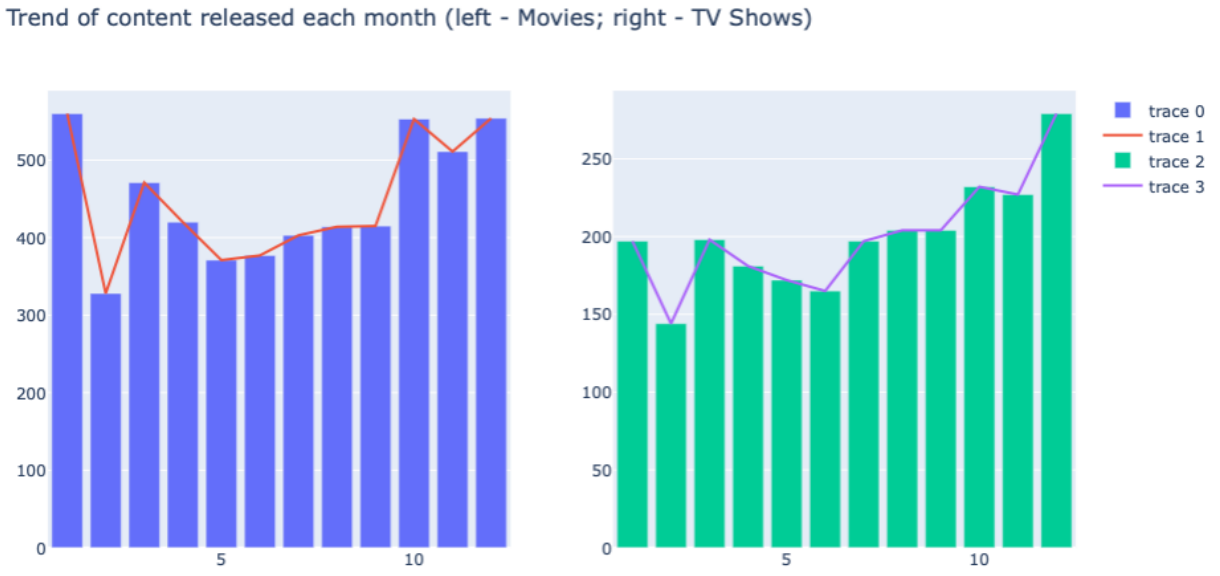
Overview of content added on Netflix throughout time

Now we will take a look at the content Netflix has added on the streaming platform throughout the years. From the interactive histogram plotted and the above heatmap, it is clear that once the stream platform became popular in 2016, more and more content has been added, with an increase up until 2020, visible as well in the heatmap, with most content added at the end of 2019 and beginning of 2020 . In 2021 the data stops at January since that is the last month available in the dataset.



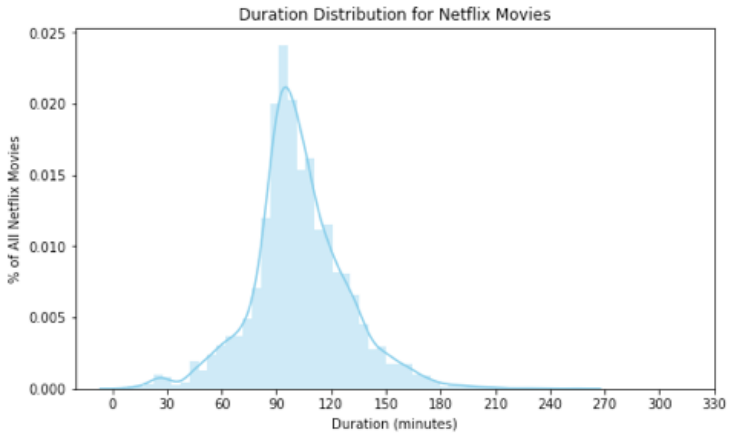


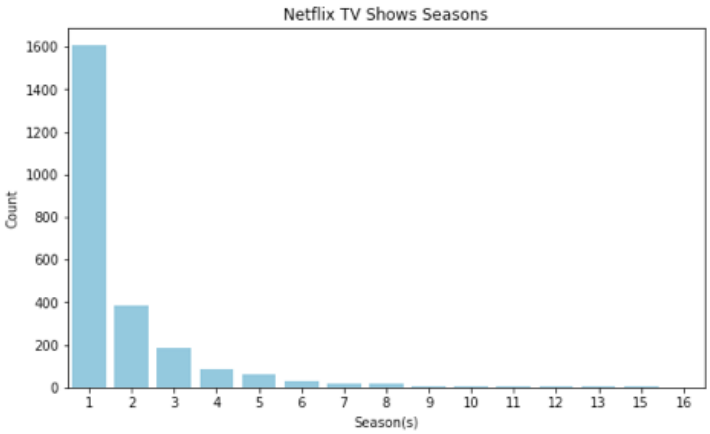
Another interesting aspect towards analysing content added by Netflix through out time is the trend of adding content across months. In the below graphs we can see that the highest count of both movies and TV Shows has been released on December and January, when producers expect people to be in-doors and spending holidays with their families, where one of the well-known activities is watching either Christmas movies or any other type of movies and TV Shows. Moreover, we can see that the lowest number of movies and TV shows released was around summer time, when people are usually spending times outdoors and not so much in front of the TV.



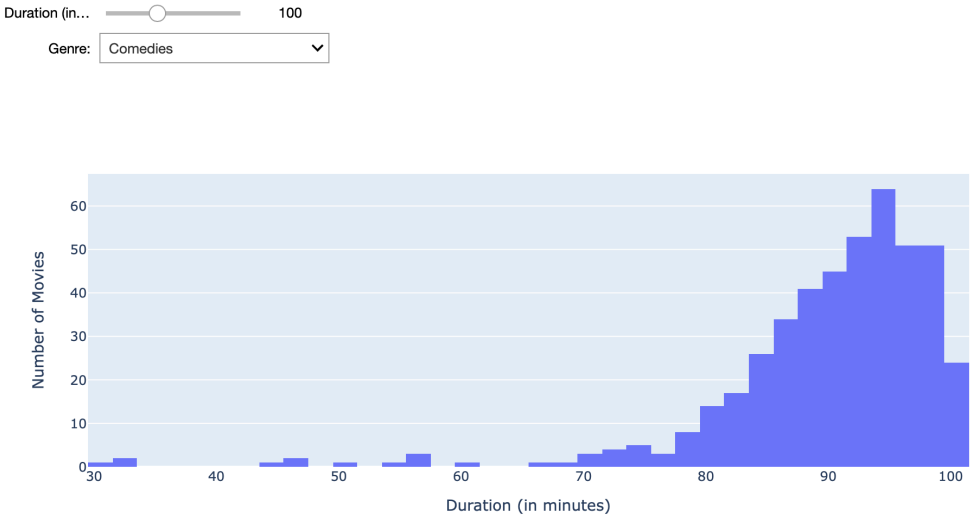
Netflix Movies and TV Show duration

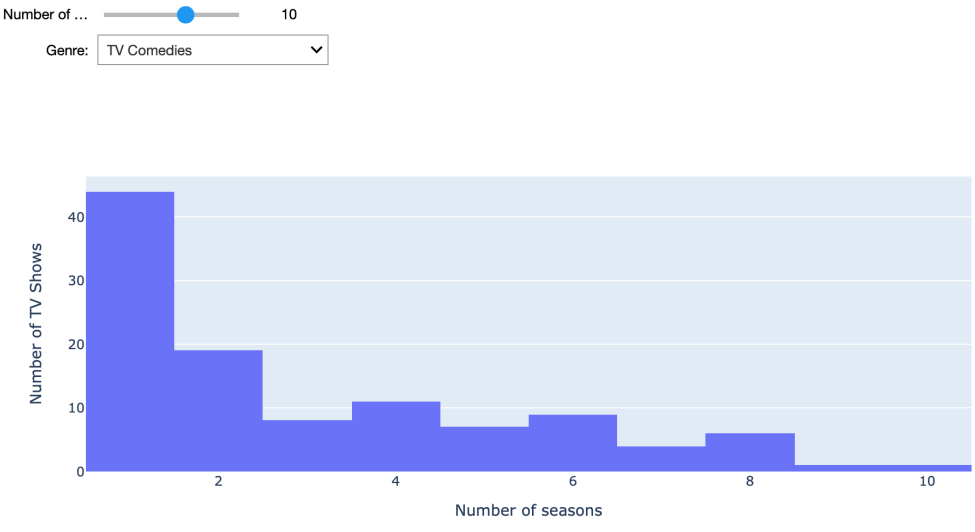
For the below representation we had to divide the data set into two other sets: for movies and TV shows, we can see in the section for movies a normal distribution, the average viewing time being 90 minutes. On the other hand, on the right side, for TV shows, the difference is that they are measured in seasons, which we can see that the graph tends to be inclined towards season 1, which means that most TV shows have mostly one season.



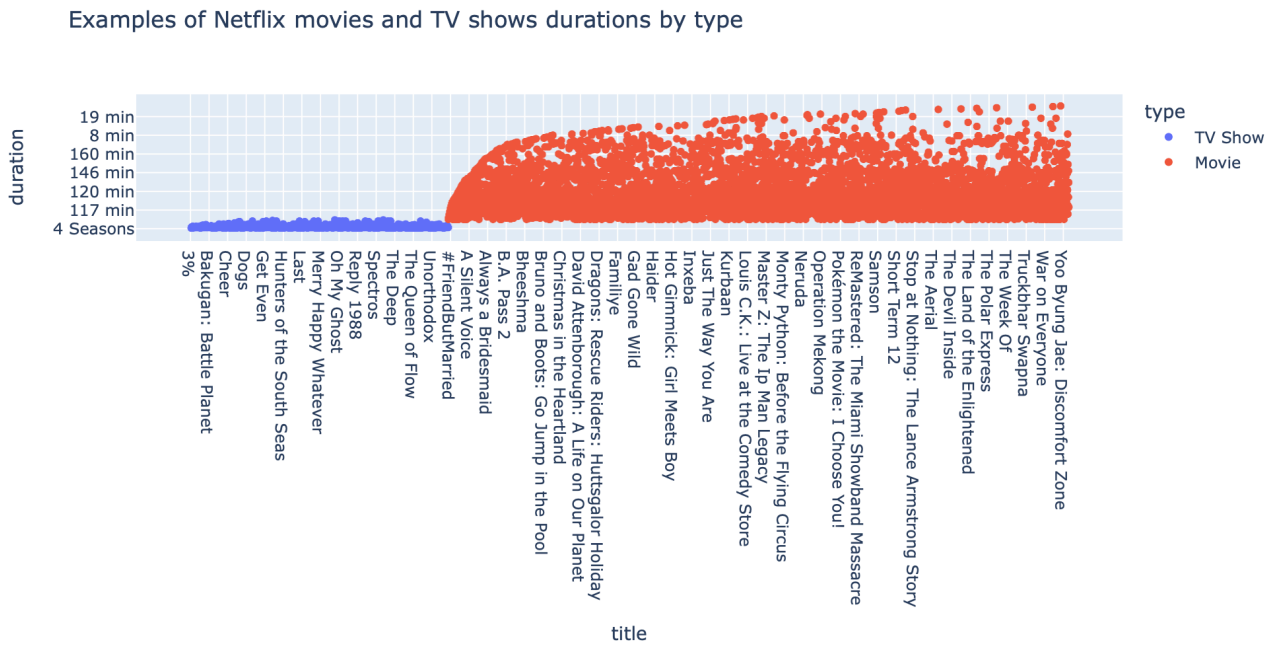


Moreover, on the Jupyter Notebook we can also play with an interactive histogram displayed below for both TV Shows and Movies, in order to filter TV Shows and Movies by genre and check the number of seasons in a TV Show or the duration of movies by genre type.



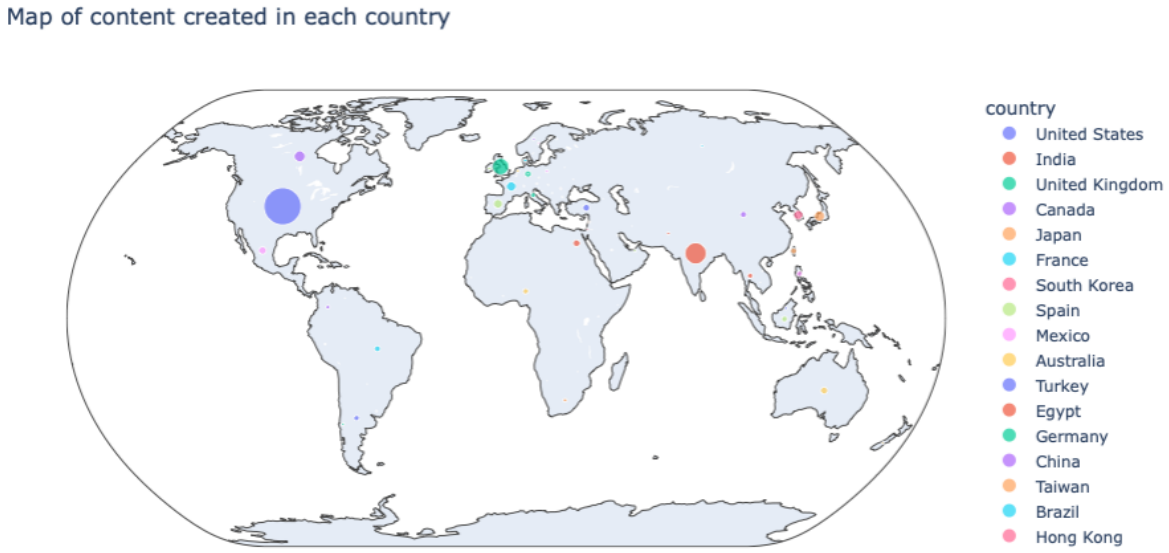


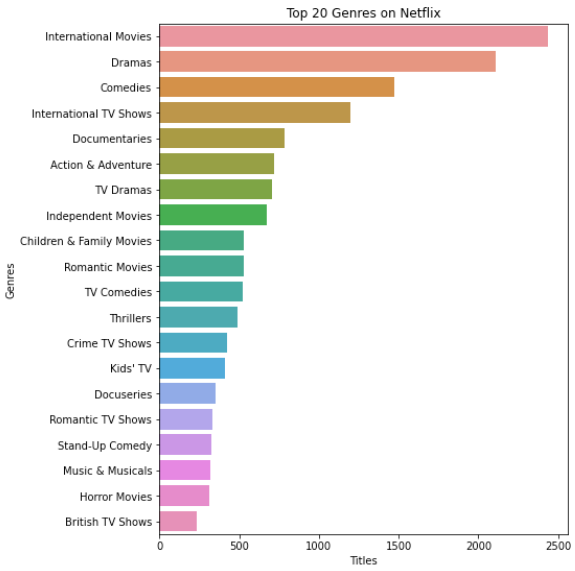
Finally, on the last scatter plot we have an overview of duration of Movies and TV Shows by type, that can also be experienced interactively on Jupyter Notebook.

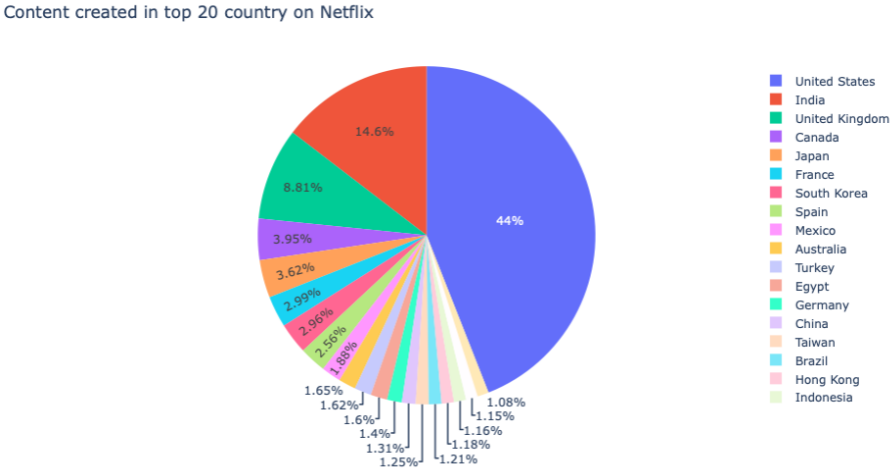


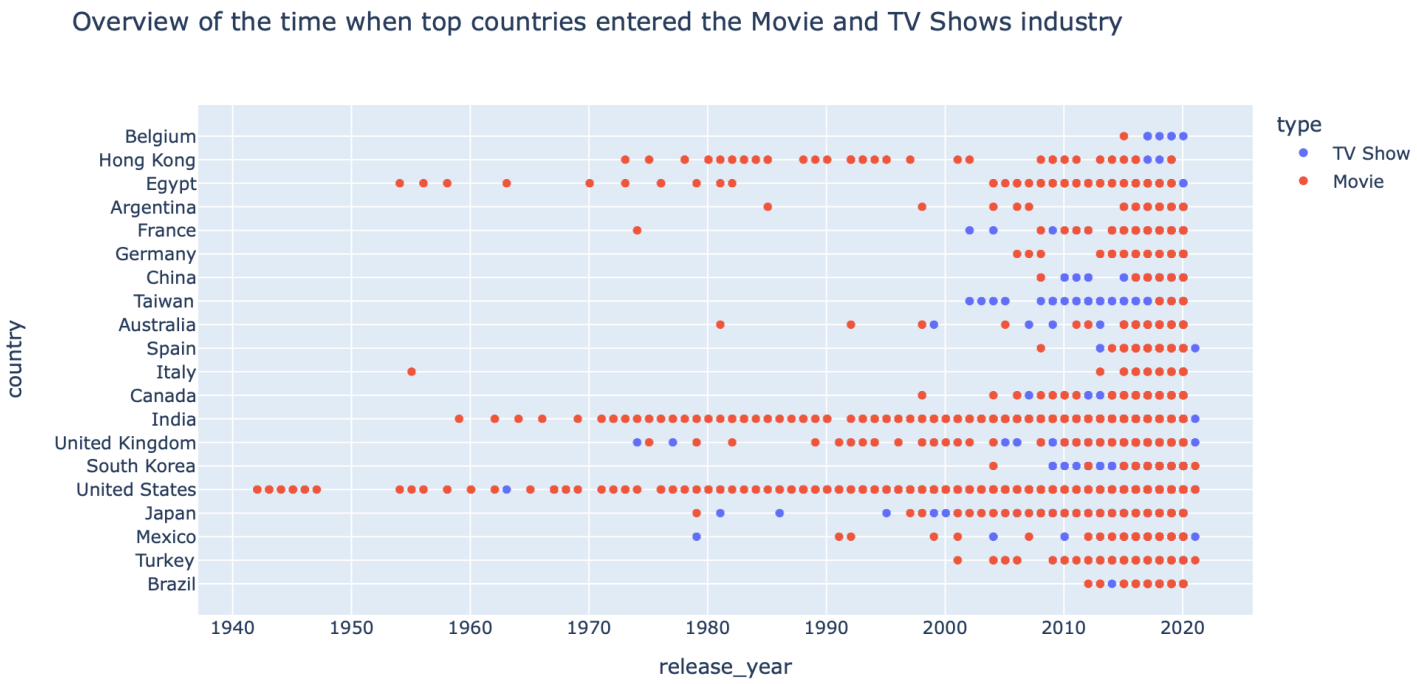
Countries roles in the Movies and TV Shows Industry

In the below interactive map and plots we can see that United States and India are in the top when it comes to movies creation and to the time countries began movie creation. Another interesting fact is that, even though United Kingdom comes three in terms of amount of content created, there are countries such as Italy and Egypt that started creating movies way before, in 1950, whereas United Kingdom only entered the industry on 1970. Moreover, there is no surprise that United States was the first country creating TV Shows, given their well-known sitcoms.



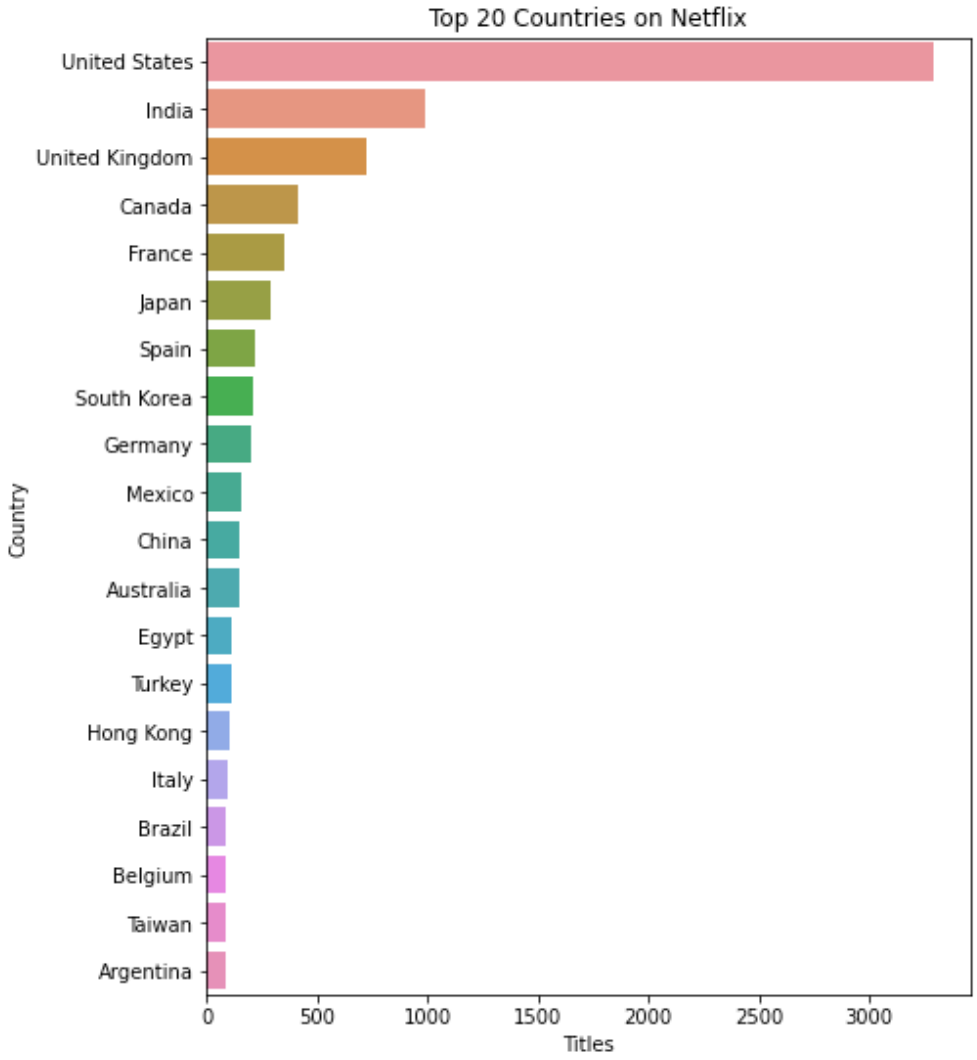

Top 20 Genres on Netflix
International Movies
Dramas
Comedies
International TV Shows
Documentaries
Action & Adventure
TV Dramas
Independent Movies
Children & Family Movies
Romantic Movies
TV Comedies
Thrillers
Crime TV Shows
Kids' TV
Docuseries
Romantic TV Shows
Stand-Up Comedy
Music & Musicals
Horror Movies
British TV Shows
Genres
0
500
1000
1500
2000
2500
Titles

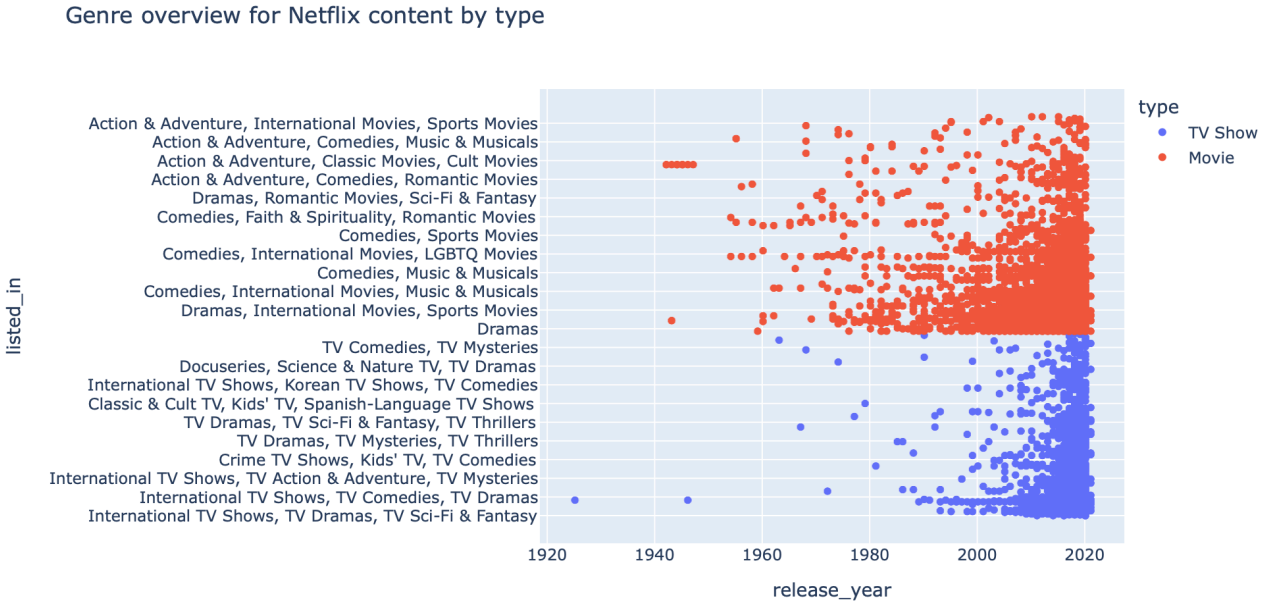

Content created in top 20 country on Netflix
14.6%
8.81%
3.95%
3.62%
2.99%
2.96%
2.56%
1.88%
1.65%
1.62%
1.6%
1.4%
1.31%
1.25%
1.21%
1.18%
1.16%
1.15%
1.08%
44%
United States
India
United Kingdom
Canada
Japan
France
South Korea
Spain
Mexico
Australia
Turkey
Egypt
Germany
China
Taiwan
Brazil
Hong Kong
Indonesia


Overview of the time when top countries entered the Movie and TV Shows industry
country
Belgium
Hong Kong
Egypt
Argentina
France
Germany
China
Taiwan
Australia
Spain
Italy
Canada
India
United Kingdom
South Korea
United States
Japan
Mexico
Turkey
Brazil
1940
1950
1960
1970
1980
1990
2000
2010
2020
release_year
type
TV Show
Movie

Overview of Movies and TV Shows Genres

In terms of film genres, it seems that International films come first, followed by the very preferred by the audience, Dramas and Comedies. Even if we notice that the United States will get the highest score in the number of content we can see that Netflix has chosen to release a lot of international movies. The reason would be that a good part of Netflix users are outside the US





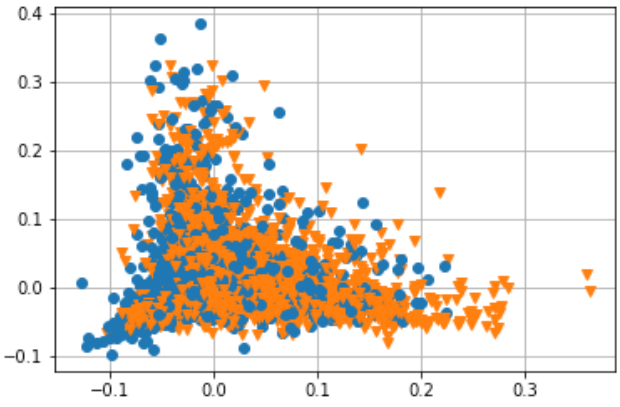
4. Models

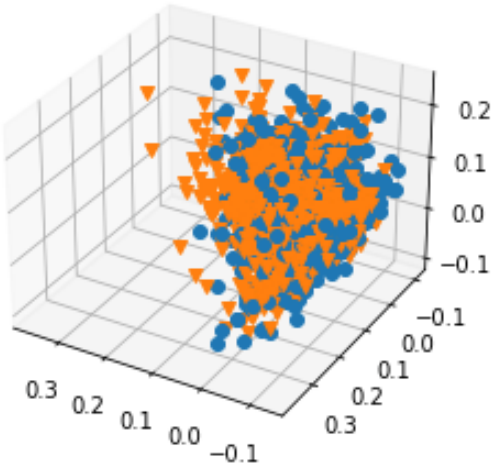
4.1. Prediction of type (Movie or TV Shows) and genre based on description

Further on, we are curious to understand whether there is a correlation between the description of the Netflix Data and the type, i.e. Movie and TV Shows, or the genre. Thus, we will proceed with two experiments:
Experiment 1: The purpose of the experiment will be to predict Type (Movie or TV Show) based on the Description added to each sample.
Experiment 2:
- The purpose of the experiment will be to predict Genre based on the Description added to each sample.
- For labelling purposes, where the Movies/TV Shows is listed under multiple Genres, we will select the main Genre and not take into consideration the adiacent ones.

For both experiments we will pre-process the Description text data and transform it to a TF-IDF matrix in order to improve our accuracy on prediction.

Moreover, we have displayed a 2D and 3D Data Distribution Plot for Experiment 1, as seen below. We managed to plot the test data by reducing its dimensionality via PCA algorithm and plotted its distribution based on the corresponding labels, 'type' for the first experiment and 'genre' for the second.





The data has been split in 80% training data and 20% training data, where for the first Experiment we have checked that the test data is somewhat balanced (given that, in the data analysis chapter we have seen that there are distinctively more movies on the dataset than TV Shows): Number of test data is 1558 out of which 468 TV Shows and 1090 Movies.

After splitting the data, we have trained and evaluated the algorithms seen in the table below, with the corresponding results, where the evaluation metric has been chosen to be accuracy:

Experiment	Model	Regularization Parameter	Accuracy
Predict type	Random Forest Classifier		73.812
Predict type	SVM	C=1	75.160
Predict type	Logistic Regression	C=1	74.775
Predict type	Logistic Regression	C=10	**75.545**

Predict genre	Random Forest Classifier		37.869
Predict genre	SVM	C=1	32.220
Predict genre	Logistic Regression	C=1	40.500
Predict genre	Logistic Regression	C=10	**41.912**

Based on training time and results Logistic Regression is the best model in terms of trade-off between computing time and accuracy. Moreover, SVM takes longer to train the model and, as we can see above, the results are similar to the other models. Overall, Logistic Regression had the best performance on both experiments, where C = 10.

4.2. Prediction of type based on description through Deep Learning Approach

For the second modeling task, we applied similar transformations as in the previous one to remove the unused English words, we removed the punctuation, change the words in the lower style and remove the numbers to get as simple a description as possible for each movie or TV show. In the end we got a specific list of words in the 'description' feature.

In the modeling part we have used a pre-trained Bert model in English, and trained it for 5 epochs, at a learning rate of 1e-6. We have chosen this model because it is one of the most powerful classification models in natural language processing. The Bert model is a bidirectional model that manages to learn information from a sequence from right to left and from left to right. It is a pre-trained model with approximately 3300M unlabelled data (words from Wikipedia and BooksCorpus).

In the following representations you can see the accuracy of the model after each era and the accuracy of the validation set.

```
Epochs: 1 | Train Loss:  0.175 | Train Accuracy:  0.511 | Val Loss:  0.172 | Val Accuracy:
0.559
100%|██████████| 1071/1071 [4:05:35<00:00, 13.76s/it]
Epochs: 2 | Train Loss:  0.164 | Train Accuracy:  0.610 | Val Loss:  0.162 | Val Accuracy:
0.634
100%|██████████| 1071/1071 [4:05:12<00:00, 13.74s/it]
Epochs: 3 | Train Loss:  0.143 | Train Accuracy:  0.725 | Val Loss:  0.148 | Val Accuracy:
0.677
100%|██████████| 1071/1071 [4:05:05<00:00, 13.73s/it]
Epochs: 4 | Train Loss:  0.119 | Train Accuracy:  0.798 | Val Loss:  0.151 | Val Accuracy:
0.673
100%|██████████| 1071/1071 [14:28:12<00:00, 48.64s/it]
Epochs: 5 | Train Loss:  0.096 | Train Accuracy:  0.860 | Val Loss:  0.147 | Val Accuracy:
0.712
```

```
Test Accuracy:  0.715
```