Joël Lingg, Ferdinand Ytteborg, and Fredrik Hægermark

# Milestone 1 - Video streaming platforms

## Dataset

We found multiple datasets on Kaggle that consisted of the main video streaming providers; Netflix[1], Amazon Prime[2], Disney+[3], and Hulu[4]. The preprocessing part is fairly simple as the quality of the datasets are quite good. However, there are a few rows that consist of empty cells. In other words, lack of information. So we have to decide whether to keep them or simply delete these rows. Luckily the amount of rows is quite low, so simply deleting these would not be that big of a deal, as we have a lot more data in the dataset. If time permits, we plan to add information from the IMDb database to our datasets in order to get information about ratings and genres of the different movies. In that way our data set would be even more complete and it will allow us to do even more data visualizations.

## Problematic

With our work we want to support users in their decision on which streaming service they will find the most attractive content. First of all, a comparison of the overall facts and figures (i.e. number of content) of the streaming services should be provided. To lead on, we'd like to compare other statistics such as ratings, amount of seasons and how fast new content is created. In that way the user will have a brief overall look at the different actors in the market, and will see what differentiates them.

In the next part we would like to use some of the data from the IMDb dataset regarding ratings as well as genres. In this way we could visualize which streaming provider that has the highest rated content, as well as if there are particular genres that they are better in compared to the other ones. For instance we could compare Netflix's Action movies with Amazon Prime's, and also their Romcom movies, and see whether one provider has better ratings in one particular genre. In that way the user can decide which streaming provider to choose based on what types of genres they like.

In the final part, we'd like to connect streaming providers with things like actors or directors. For example, we'd like to have a look at which actors are purely found on one streaming service and which ones are all over it.

During all the things mentioned above, we'd like to give the user the possibility to tailor the site or parts of the site to their preference. This can include

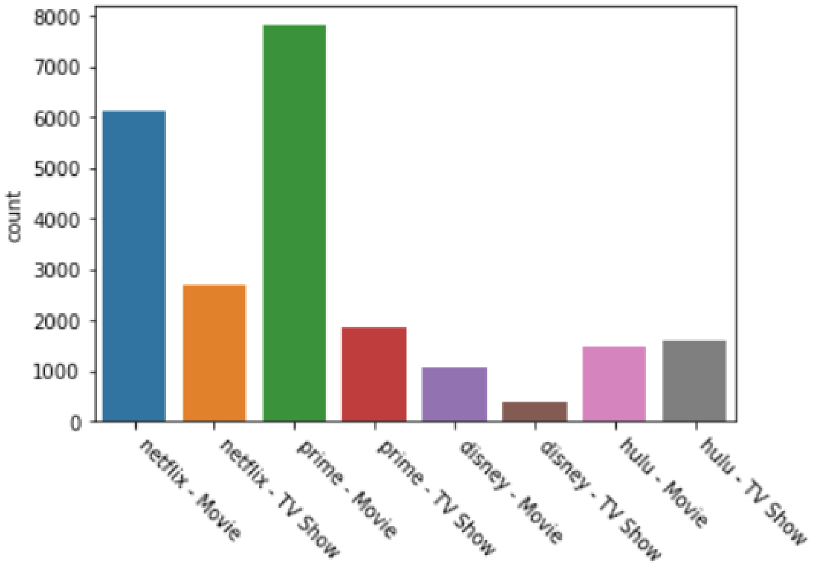
- Time limits: Movies should be older than 1990
- Genre: Only include the genres Mystery, Science Fiction and Thriller
- Actors: Only include my most beloved actors
- Director: I don't like this Director

[1] https://www.kaggle.com/datasets/shivamb/netflix-shows
[2] https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows
[3] https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows
[4] https://www.kaggle.com/datasets/shivamb/hulu-movies-and-tv-shows

Doing this should allow the user to get some insights about the streaming services. If time permits, we'd like to take also into account the cost of streaming services (e.g. in Switzerland & Norway[5]) to decide which one has the best value for money.

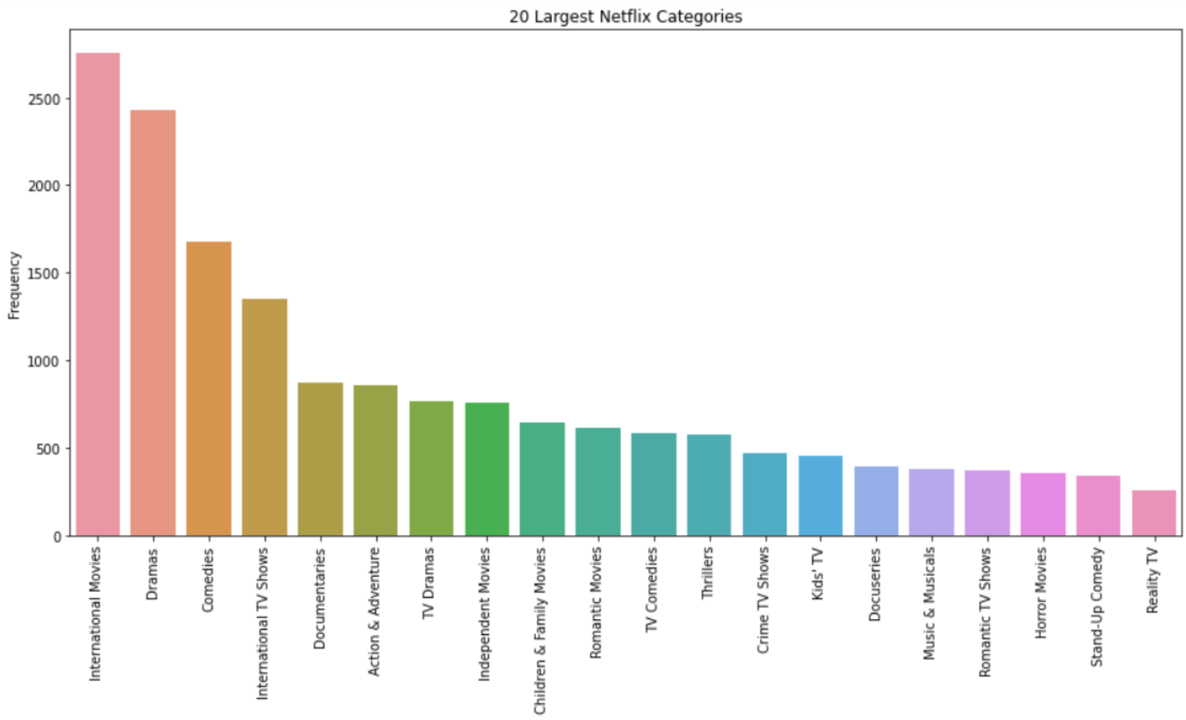
# Exploratory Data Analysis

One plot that's really interesting to look at is the one that shows the number of series and movies on the different platforms. The amount of series and movies might play a huge role in regard to deciding on which streaming provider to choose, as some people might favor series more than movies and vice versa. From the graph below we can see that Netflix and Prime dominate the market in terms of movies, and there is a huge difference in terms of the distribution of movies and series for the two platforms.

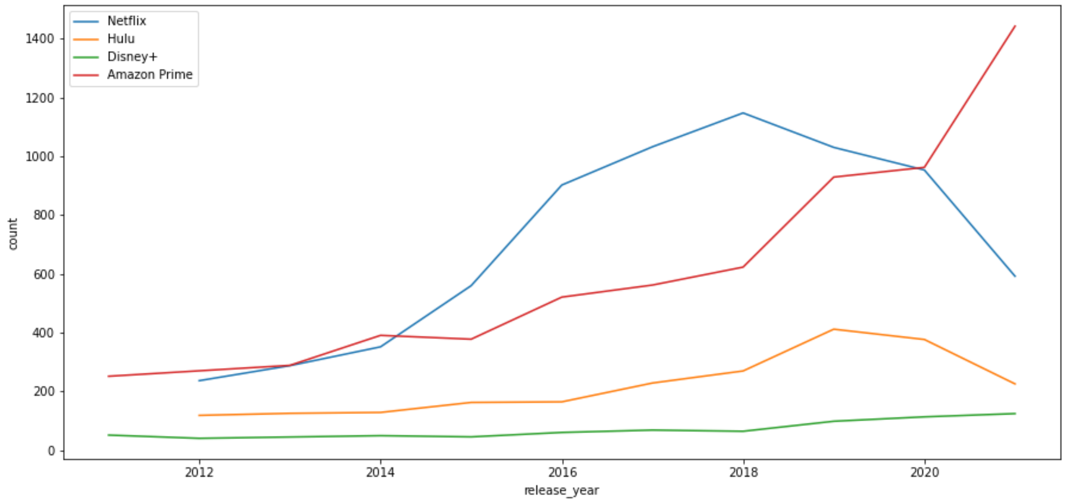


It would be interesting to look at what kind of different genres and categories the different streaming providers contain. Below you would find an example of the different categories Netflix consists of. We can already get an idea of what kind of categories that is dominated by this streaming provider. Netflix has a lot of international movies. It would be interesting to compare this plot to those of the other ones.

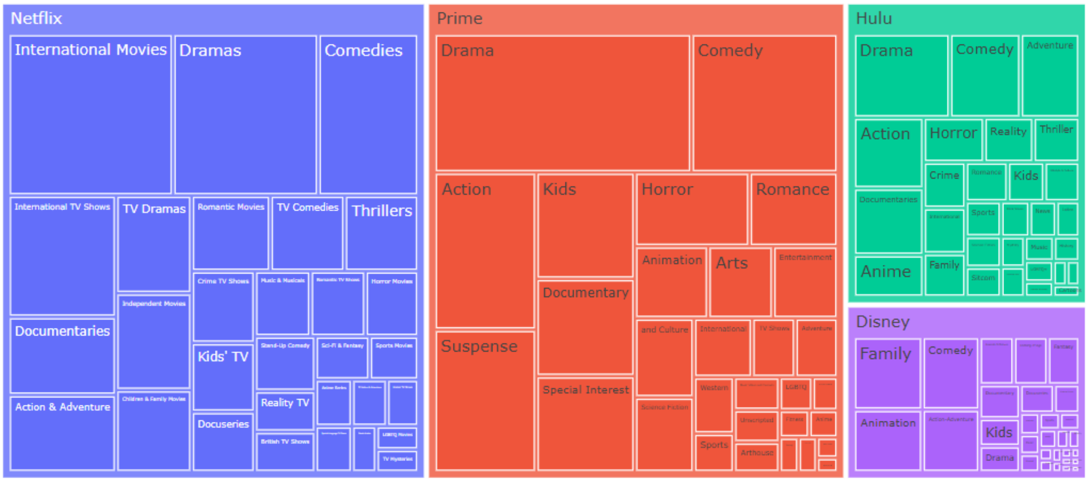
[5] Those are the home countries of the authors of this document.



Another interesting plot to take a look at is the one that shows how many movies and series each provider has released over the previous years. This gives the user some idea of what to expect in terms of new releases. We can see that Netflix previously was way above the other ones, but has declined the previous two years. Prime however has ramped up their acquisitions, and has now surpassed Netflix.



The tree plot is also quite interesting as it gives a great overview over both the biggest provider in general as well as in terms of the different categories. We can see that Netflix and Prime are quite similar in the amount of content, and way in front of the other two. We also see what the biggest categories are for each provider and how they compare to the others.



# Related work

Kaggle provides a feature to share and search code linked to a dataset. We leverage this feature to find previous work done with our datasets. With Netflix being in business for a long time, it is no surprise that there is a lot of work done with the dataset. Most of them are indeed concerned with visualization. However, when moving on to the other datasets, much less projects leverage them. The Hulu set has only 2 code projects available. It needs to be mentioned that there has been work done on the Disney+ dataset to compare it with Netflix. However, our goal is to compare all four of them to find the one streaming service to rule them all. There has been done some[6] really inspiring visualizations with one dataset. However, we'll try to generate visualizations which compare the services directly instead of putting the same visualizations for all providers side by side.

The fifth dataset[7] has been leveraged to do some side to side comparison[8]. However, as mentioned, we'd like to focus on doing single visualizations comparing the services instead of putting facts & figures side by side. Additionally, the visualizations we've found are fairly static. We'd like to make our website more adaptive allowing the user to get info about his favorite actors or about his most beloved years.

Of course, there are websites trying to answer the question which streaming service will be the best. However, as far as we can tell, most of them seem to focus on doing a textual comparison without visualization. Our goal is to do the opposite. Compare them visually and interactively with text to explain the visualizations if needed.

[6] https://www.kaggle.com/code/subinium/storytelling-with-data-netflix-ver
[7] https://www.kaggle.com/datasets/ruchi798/movies-on-netflix-prime-video-hulu-and-disney
[8] https://www.kaggle.com/code/ruchi798/movies-and-tv-shows-eda

We found two sites with some visualizations[9], but they focus on the questions “Best value for money per country?” and “What do current users think?”. Both of those topics are not the main focus of our work, but may be an idea to investigate if time permits.

[9] https://www.comparitech.com/tv-streaming/netflix-vs-hulu-vs-amazon/
https://observer.com/2021/09/amazon-apple-netflix-disney-hbo-max-streaming-wars/