NETVIZ

Ten years ago Netflix was only present in the United States. Nowadays, the world-wide streaming platform is available in more than 190 countries. Its catalogue is rapidly growing not only to catch up with recent movies and TV shows releases, but also to adapt their content to the local market by producing original content. We expect that the evolution of their library reflects this strategy.

Our Story

When looking for available datasets about Netflix and other streaming services, we didn't really know the direction in which we were going or what we wanted to show. Each available dataset was either incomplete or outdated. This is when we decided to build our own dataset, using an unofficial Netflix catalogue, Flixwatch. After making sure it was allowed, we scrapped the content corresponding to all movies and TV shows available worldwide, along with all possible attributes. With data in hand, the next steps were exploring it and getting acquainted with it. After a few brainstorming sessions, we decided on our main theme, the evolution of Netflix's release strategy, and built our storyline along two main directions. First, the geographical and cultural expansion of the streaming platform. Then, the evolution of the content itself following that expansion. Additionally, we thought it would be interesting to complement that with an exploratory tool, to satisfy more curious readers. With that in mind, we started drafting a few sketches of our visualizations and of our website in general. We quickly agreed on a dark, sleek and minimalistic design, centered around the red color, hinting at the design of the streaming platform.

Then came the implementation of the final product. We first designed a basic skeleton, using smooth scrolling pages, each containing one part of the visualization. This way, each part of the visualization can be integrated with our storyline, and separated from the rest, to avoid having a cluttered and dense website. This scrolling navigation choice made it easy to tell a story, mimicking the flipping of pages in a children's book. Throughout the whole implementation process, we made sure (or tried to at least) stay consistent in terms of design choices. This was to ensure a consistent and pleasing visual identity of the website. Generally speaking, the implementation went smoothly, and stayed true to our original sketches.

Once all the visualizations were set, we linked them using a well thought, light and user friendly scenario. The story immerses the reader in the website, by mentioning a familiar feeling of endlessly scrolling through Netflix to find a show. This assimilation should captivate the audience and encourage it to continue browsing the website.

Challenges

We faced some challenges from the very beginning of this project. It took us some time to settle on the subject of our data visualization. Once that was settled, we spent more time finding a dataset, then decided to scrape our own data. In the initial versions of the data, we noticed missing movies and TV shows. After a few tries and towards the end of the semester, we eventually converged to a complete dataset. This hindered our development speed, but thankfully, our visualization could be engineered with “wrong” data.

One important type of challenge we faced was related to the size of the data. Netflix’s catalogue is very large. Very large. It contains 25,852 items from which 20,211 are movies and 5,641 are TV shows, all representing 95 different languages. With such an amount of data, it was difficult to come up with insightful visualizations. We often had to rethink ways to select subsets of data that made sense for the visualizations we had in mind. We also had to simplify some parts, and tone down our initial expectations, because dealing with so much data was not computationally efficient nor appropriate for reactive and interactive visualizations.

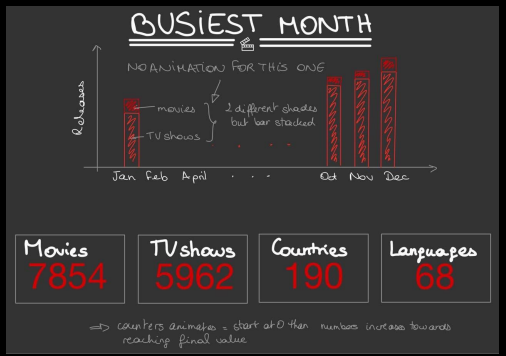
Finally, we encountered obvious technical difficulties regarding the website’s implementation. For example, starting animations on the page and not on the loading of the entire website, the positioning and sizing of plots in the window, or the integration of a navigation bar on our visualizations. The list is long, but thankfully, we have been able to fix most issues. One of the major issues we haven’t been able to fix is the multi-browser compatibility of our website, and its behaviour when fitting the window on smaller screens. Specifically, we tested our code on Google Chrome and Microsoft Edge, and noticed some issues when using Firefox.

Did we stick to the original plan?

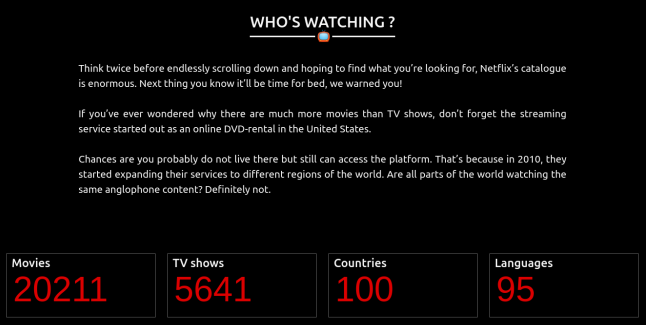
Generally, we closely followed our initial structure and sketches. The main layout of the website didn’t change throughout the project. However, we deviated from some of our original sketches, to deliver a clean and attractive final product.

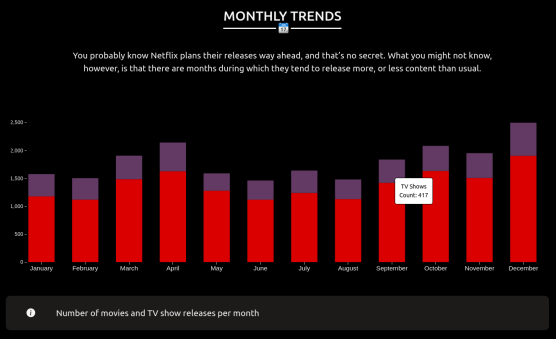
Introduction page

Before diving in the subject, we give a little introduction to our storyline and our dataset, showing in what month Netflix releases the most content and some basic statistics about our data. Initially, we had



planned to have a single page for that part, but the result was too dense. We therefore split that page into two: one showing global counts about the dataset, and introducing the storyline, and the other showcasing the monthly trends of releases.

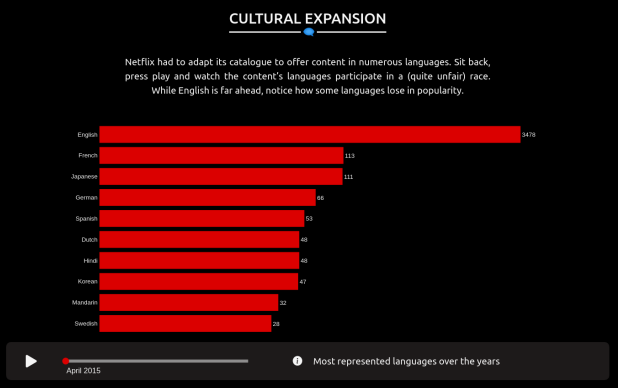


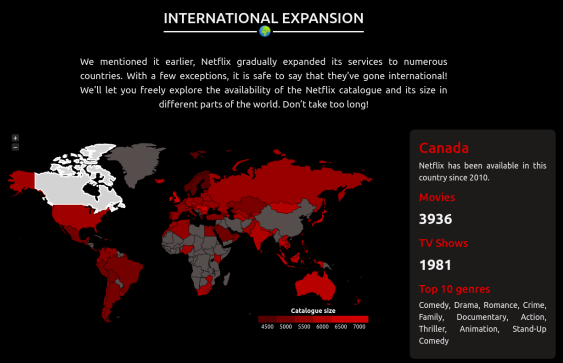


Cultural and Geographical expansion

This first part shows how Netflix expanded geographically and culturally.

For that purpose, we first displayed the representation of languages in Netflix's content across time, movies and TV shows combined. Our initial idea was to animate our bars in a "race-like" fashion. We successfully implemented this, but with a slight change: we didn't display the flag of the country corresponding to the language. We thought it would be more meaningful to display the count corresponding to the language at the current time. We also implemented a navigation bar at the bottom that helps the user seek through the animation.

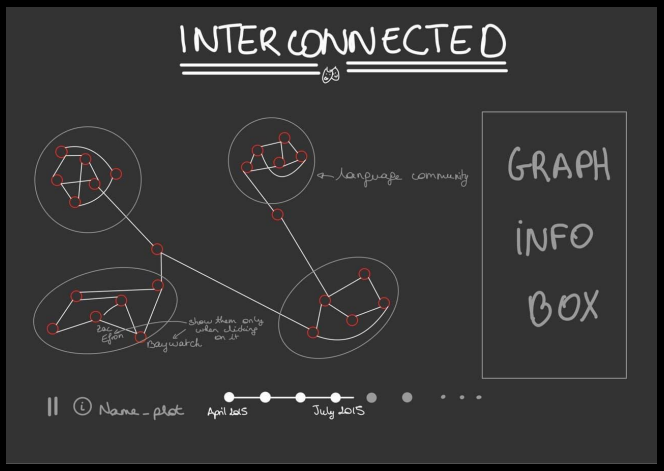




In the second visualization we used a clickable map that displays insights about a particular country, in order to show how Netflix expanded geographically and culturally. The map is color coded with respect to the content available in a particular country. Countries colored in gray were either missing from our

dataset, or do not benefit from Netflix's services. In the side box, we included, when possible, and after selecting a country, the year in which Netflix was available in that country, the number of movies and TV shows available in it and the top 10 most represented genres.

The last visualization in the "geographical and cultural expansion" axis is a network showcasing the connection between movies, depending on their cast.
We underestimated the complexity of the implementation of this visualization, as we had initially planned to display all movies. We didn't take into account the size of our data, and therefore had to make a selection on the nodes to display. We selected the 10 languages with the most movies, and selected the hundred best movies, sorted with respect to their metacritic and IMDb ratings. With this filtering, we managed to construct a reasonably-sized graph, where two movies are linked if they share at least an actor. We colored the nodes by the audio of the associated movie, and were able to visualize different *communities.* We put aside the time dimension of the visualization as it was too complex.

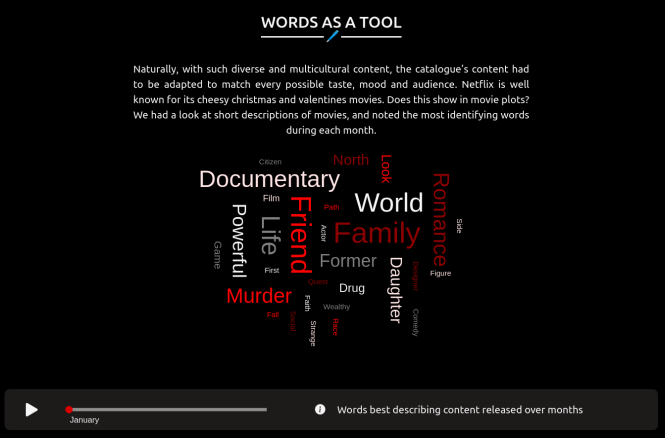


Evolution of the content

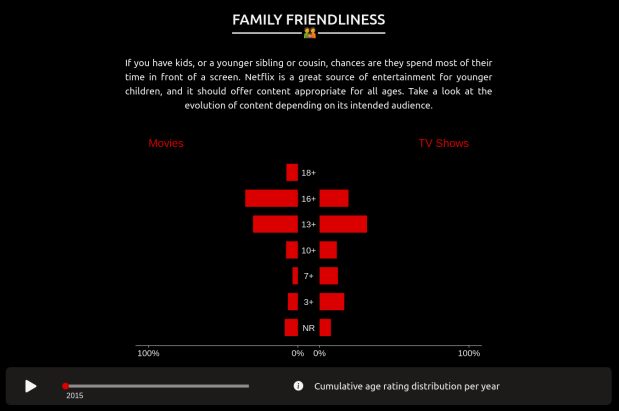
The second part is to show how Netflix adapted their content following their international expansion.

We started with a world cloud displaying the most frequent words used in the description of the movies and TV shows released over the twelve months, all years combined. The implementation went smoothly, but the data preparation wasn't a piece of cake.

The initial idea was to perform keyphrase extraction, but the results were not satisfactory. Instead, we processed the item descriptions and kept words that are only particular to a specific month, but not frequent in the full data. That way, we were able to find words characterizing a single month. The results are quite satisfactory: we observe love vocabulary in February, and Christmas-related words in December.

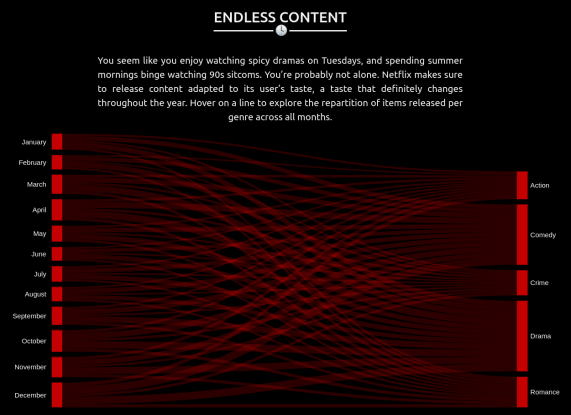


We decided to display the words in two rotations, and to set the size of the word according to its importance. As mentioned before, the navigation bar helps start the animation and allows you to go back and forward in time.



The next visualization shows the proportion of content that Netflix made available over the years, depending on its age rating. We didn't encounter any issues with the implementation of this visualization and stayed true to our original sketch.

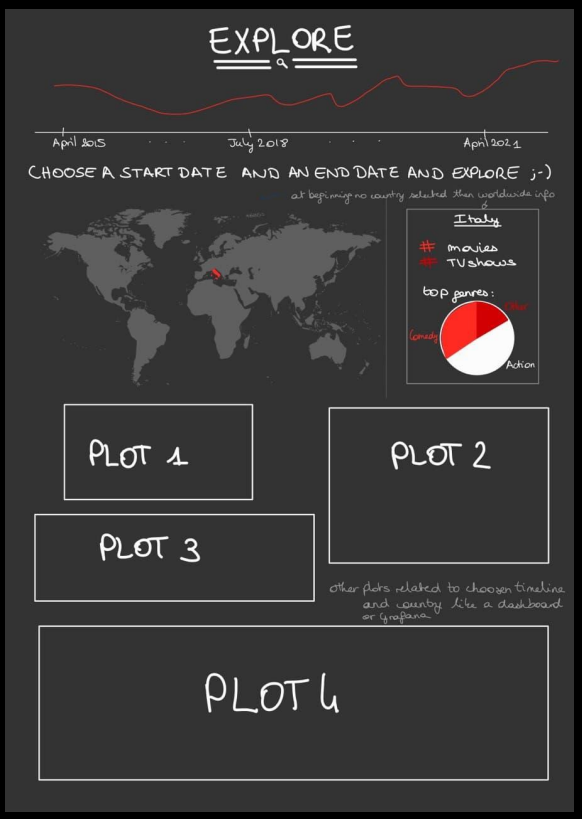
Finally, for the last visualization of this part, we display the proportion of content released in each month to the top 5 most represented genres using a Sankey diagram. We made some little modifications compared to what we had initially planned. Displaying the proportions regarding all the genres resulted in a very cluttered and unreadable plot. Unfortunately, we could not implement the selection of a certain flow.

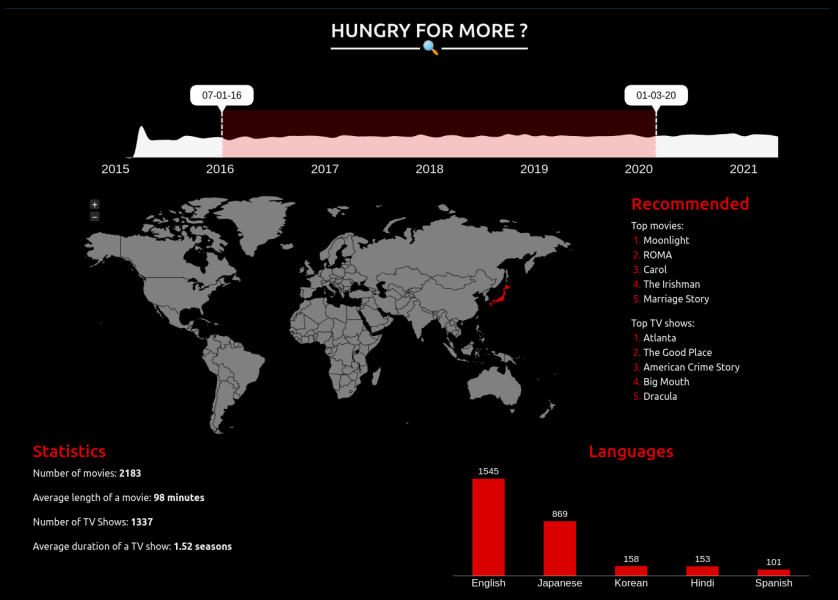


Exploratory tool

Last but not least, to give the user even more insight on the data, we wanted to include an exploratory tool, in which the user could select an interval in time and a country, and would get in return statistical insights about the selection.

The major issue we faced was to implement a quick way of filtering data according to the user query, but it was eventually sorted out. Because of the density of the data, we decided to only display basic statistics: the top 3 best rated movies and TV shows as well as the number of available movies and TV shows in the selected period of time and for the selected country.





Contributions

Overall, we all worked together but at different levels depending on our skills.

Milestone 1

For the first milestone, we all participated in the search of the dataset. Miloš scrapped the data and we all contributed equally to the EDA. Finally, Eloïse and Karim clean the EDA prepared submission of the milestone.

Milestone 2

For the second milestone, Karim and Miloš worked together on setting up the website. On the other hand, we all produced sketches and ideas for the visualizations. Eloïse drew the final version of them. At this point we started splitting the work between team members according to visualizations. We each implemented static versions of what we envisioned. Eloïse worked on the stacked bar plot and the racing bars. Karim worked on the network and the map. Miloš worked on the mirrored horizontal bar plot and the word cloud. Finally, Karim redacted the milestone ,then Eloïse and Miloš helped with the relecture.

Milestone 3

For the third and final milestone, we all continued working on our previously static visualizations.

- Eloïse prepared the data and implemented the visualizations of the stacked bar plot, the animated counters, the racing bar plot and the sankey diagram.
- Karim prepared the data and implemented the visualization of the map and the network. He also prepared the data for the word cloud.
- Miloš prepared the data and implemented the mirrored horizontal bar plot and the explanatory tool, and thanks to Karim's data he implemented the visualization of the word clouds. He had to re-scrap the data during this milestone and also implemented the navigation bar that the other team members reused all over the website.

In addition to the technical implementation, Eloïse wrote the process book and Karim wrote the user storyline that is visible on the website story. Miloš helped the other team members when they had technical questions about their code.

General Assessment

Throughout this project, we managed to equally and efficiently divide all the work among team members. However, that doesn't mean we lacked communication. We organized online meetings at least once or twice a week to update each other on our progress, and even managed to meet in-person at EPFL! We regularly consulted each other and all agreed on directions to take regarding our different visualizations.