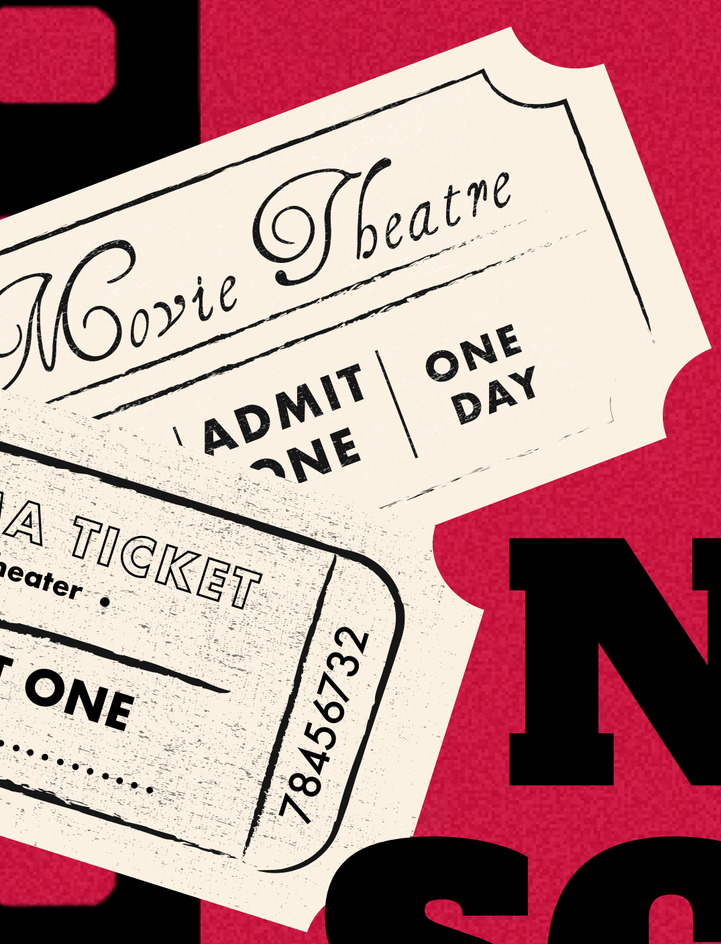
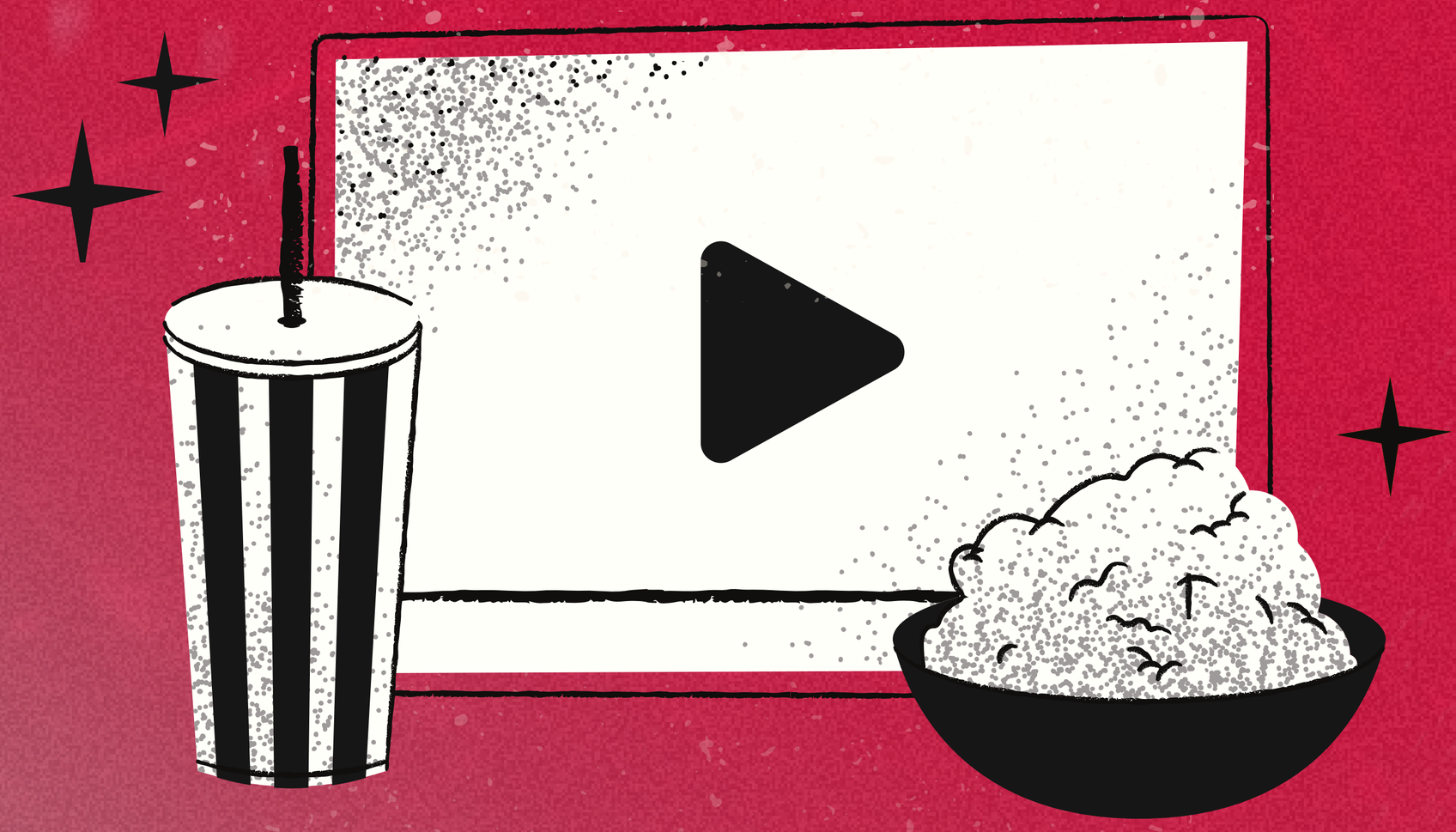


NETFLIX SCRAPERS

[Patrick Louis Aldover
Jérémy Chaverot
Jason Mina]



EPFL - COM480

INTRODUCTION

Nowadays, during our free time, we all indulge in various forms of entertainment to relax and escape from our daily routines. Among these, movies and TV shows hold a significant place. Based on our shared interests and this observation, we decided to focus our data analysis on films and TV shows to highlight the different factors influencing their success. Specifically, our focus is on actors and actresses, as these celebrities play a crucial role in marketing, production, quality, and ultimately the viewer's opinion.

How to come up with a novel and original idea?

That’s the question we asked ourselves at the very beginning. It’s commonsense that cinema is a domain that has been seen and seen again. That’s why to go off the beaten tracks, our interest has turned to a brand-new way of consuming entertainment. Since we often use Netflix to watch our favorite movies and shows, we decided to focus on movies and shows from this platform, which constitutes an original approach to the subject. Indeed Netflix comes in as a new important place for the movie industry. Nowadays, movies don’t necessarily even premiere in theaters anymore; their global release can happen solely on Netflix, the American subscription video on-demand streaming service. Take, for example, Don’t Look Up, a $75 million-budget film released in 2021 starring Leonardo DiCaprio, Jennifer Lawrence and Thimothée Chalamet. Moreover, Netflix stands out as the pionerring platform that provides TV shows with a global audience, regardless of their size or budget.

In a nutshell, what is to be seen ?

First, we guide you through the statistics of your favorite actors on Netflix, tracing their careers and examining the evolution of their success over time using IMDb and TMDb scores of the movies they have starred in. Next, we'll delve behind the scenes by exploring the influence of co-stars with a network graph, giving you a clearer picture of all the interactions and connections. Using this data, we'll then take a broader view to study the geographic influence on viewer opinions with a dynamic world map showing the birthplaces of actors. Finally, our last visualization will take you to the heart of IMDb scores, breaking down the scores each year into various ranges to reveal the main tags that emerge.

Is the visualization made for you ?

We mainly target three groups of people. First and foremost, movie and show enthusiasts can use our visualizations to discover (possibly new) actors that had a positive impact on the movies and shows they starred in. Conversely, these enthusiasts can take a look at our visualized pre-processed data to be aware of actors that negatively influenced previous movies and shows. Our second target group involves movie and TV critics. Similar to movie and show enthusiasts, they can utilize our visualizations to assist them to write reports and articles about an actor’s (un)favorable movie/TV history. Finally, we developed our visualization website also for the statistical aspect: Our illustration can assist researchers to determine key factors of a movie’s or show’s success.

The dataset

The primary dataset used for our data visualizations is the Netflix TV Shows and Movies dataset found on Kaggle. This data gathered by Victor Soeiro contains information about Netflix movies and shows and their actors. The information contained are recent and up to July 2022.

LINE CHART

Description

As an introductory visualization, we visualized the impact of a single actor/director via a line chart. The x-axis denotes the release year of the movie/show, while the y-axis indicates the IMDB/TMDB scores. On the website, the green point indicates the score of a single movie/show and the red point indicates the mean score of all movies released in a specific year. When hovering the mouse over a green point, a tooltip listing additional details about the movie/show (Title, Rating, Runtime, Description) appears. To highlight the impact of a single actor/director, we drew a horizontal line in the line chart, describing the mean IMDB/TMDB score across all movies/shows.

Design

The implemented line chart slightly deviates from our provided sketch. First, we used slightly different colors to match the color scheme of our website. Second, we added radio buttons above the line chart to switch between IMDB and TMDB scores. And third, we added grid lines to simplify the readings of the points.

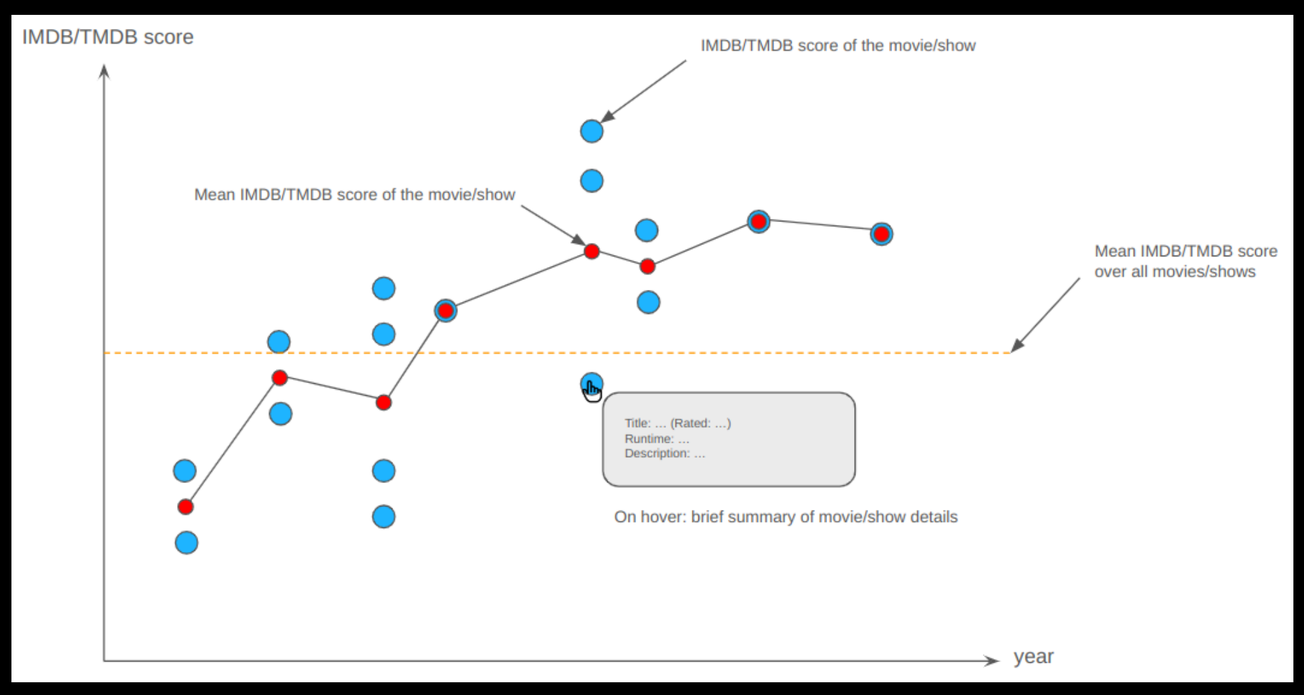


Figure 1: The line chart sketch.

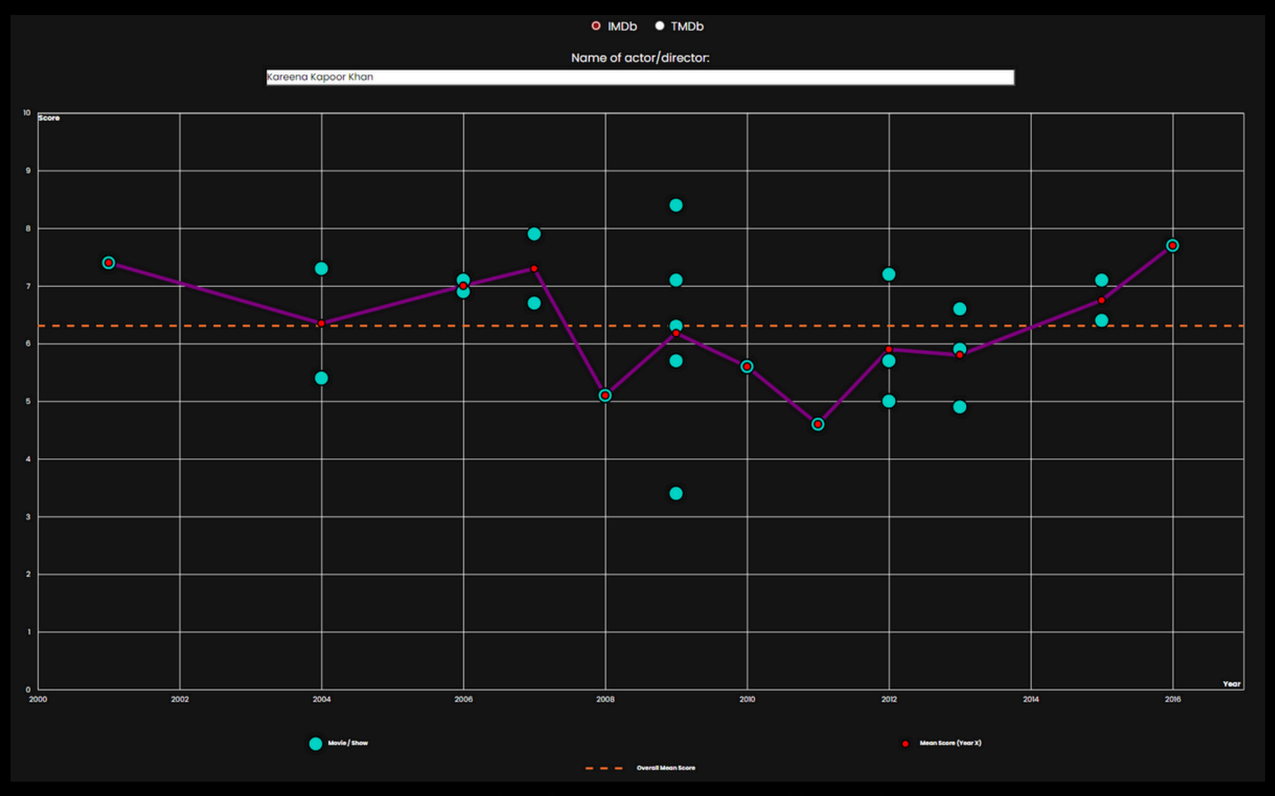


Figure 2: The implemented line chart.

Challenges

The main challenge of this visualization was creating a smooth transition of the points and lines when switching between IMDB and TMDB scores. We achieved this by utilizing the transition and duration functionalities of D3.js.

ACTOR/DIRECTOR NETWORK

Description

To illustrate which co-actors/directors worked especially well with actors/directors, we visualized an actor/director network. Each node describes an actor/director or a movie/show. The visualization is interactive, allowing a user to zoom in and out of the visualization and move individual nodes around. The links between the nodes are colored based on the IMDB/TMDB score of the movie/show.

Design

The implemented network slightly deviates from our provided sketch by using different colors. Apart from that, we implemented the network as intended.

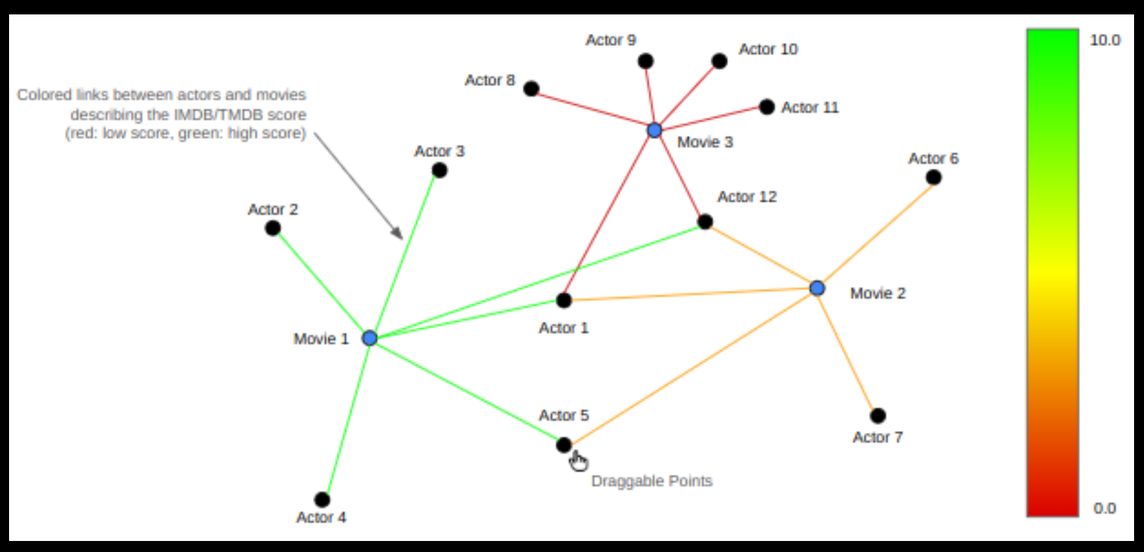


Figure 3: The actor/director network sketch.

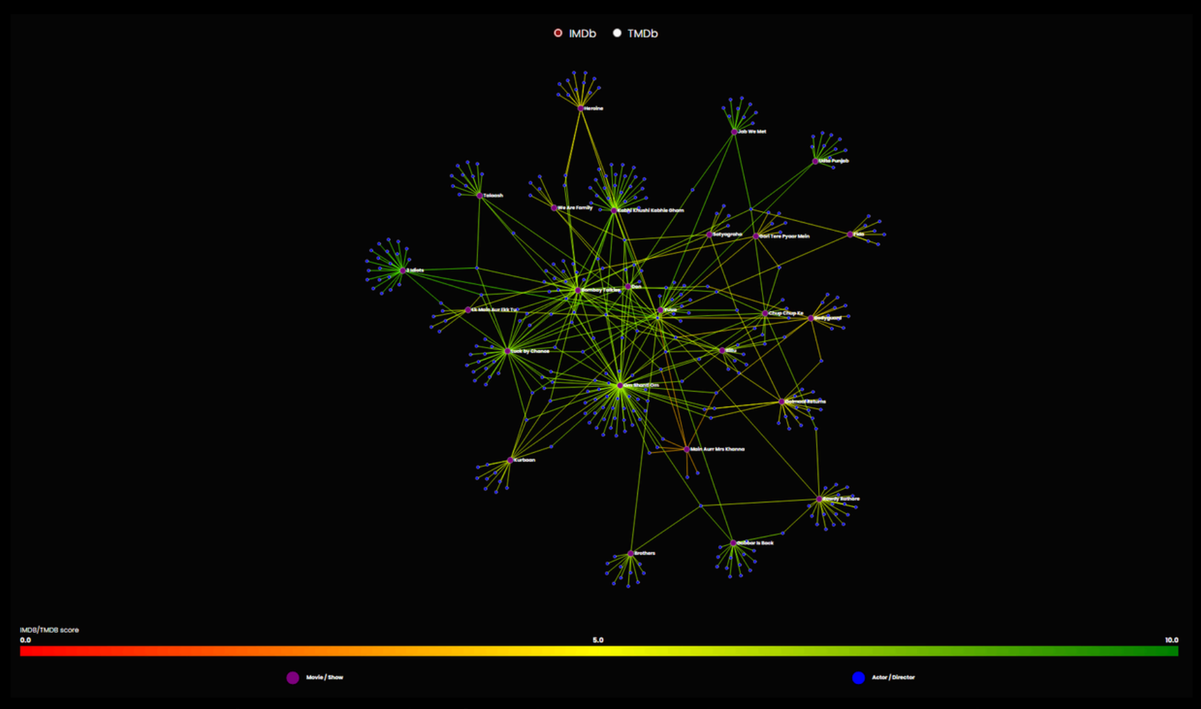


Figure 4: The implemented actor network.

Challenges

We encountered two challenges while implementing the actor/director network. First, we want to make the visualization interactive. We achieved this by utilizing the force graph of D3.js. We created two node sets to distinguish between movies/shows and actors/directors. Once again, we introduced a radio button to switch between IMDB and TMDB scores.

The second challenge we faced included overlapping labels and links. We initially hid the actor/director labels to make the network graph user-friendly. Once the mouse hovers over a node, the label appears. We also increased the distance between the nodes to avoid cluttered networks. However, we realized the overlapping links were caused by actors/directors often cooperating with the same co-actors/co-directors. In conclusion, avoiding link overlaps with the given datasets is nearly infeasible.

WORLD MAP

Description

This visualization aims to present an interactive world map highlighting statistics about the geographic distribution of actors' birthplaces in the Netflix dataset. Starting with a global overview, the map uses darker colors to indicate higher actor counts in each country. Our dynamic map features tooltips that display information when a user hovers over a country, and zooms in for a detailed view of birthplace distribution upon clicking. This functionality helps identify emerging patterns or trends, and to identify locations that serve as hotspots or incubators for actors.

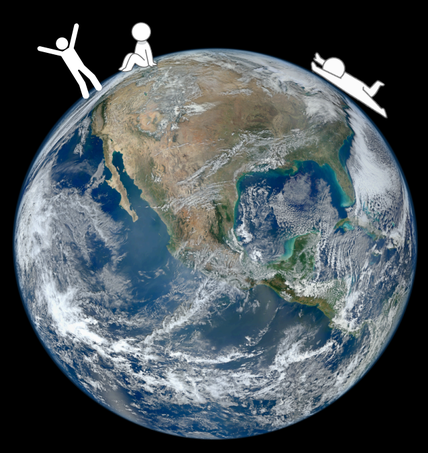
Design

To bring our ideas to life, we found a Choropleth Map in the D3.js graph gallery that perfectly suited our needs. We then added a hover effect to display statistics for the country of interest when the mouse is over it. To ensure the visualizations were interconnected, we made it so that specifying an actor in the Line Chart visualization highlights the actor's birthplace on the map.

Additionally, clicking on any country zooms in by changing the projection center and scale. This reveals the aforementioned bubble map. To enhance interactivity and readability, we also included a slider to filter birthplaces and set a minimum number of actors or directors.

The colors were slightly adjusted to match the global Netflix theme on the website. Aside from this and the legend format, the final visualization closely aligns with our initial sketch.

In the design phase, we abandoned including a toggle button to switch between color representations indicating the number of people born in each country and representing the average IMDb/TMDb score. This decision was made because using colors to represent average scores wouldn't have been very informative, given the minimal differences in average scores across countries. Instead we put this infoformation in the tooltip.



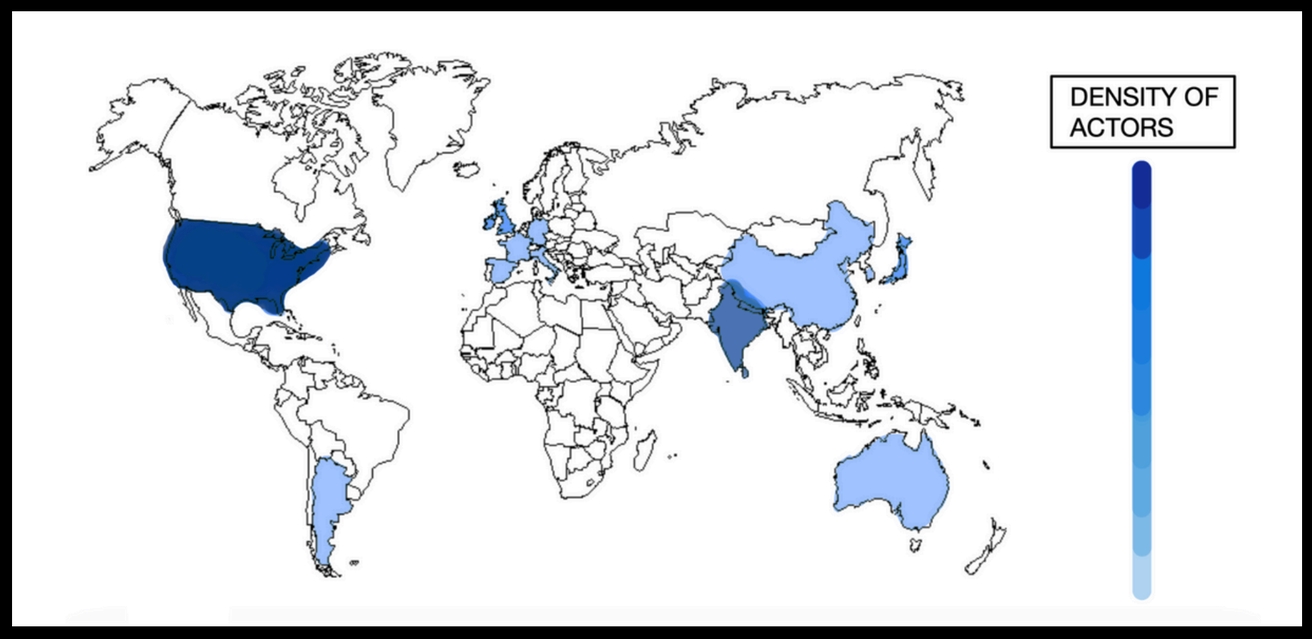


Figure 5: The world map sketch.

click

Figure 6: The bubble map sketch.

Challenges

The main challenge for this visualization was the dataset. Initially, our dataset lacked information on the actors' birthplaces. To address this, we used the TMBD API to retrieve the birthplace strings for each actor. Next, we processed these strings into the correct format and used the NINJA API to obtain the corresponding latitude and longitude coordinates. For the places that were still not found — due to reasons such as the country no longer existing, misspellings, or non-English names — we made a third request using the GPT Wrapper API to process and reformat the data, then resubmitted it to the NINJA API. Using this method, we successfully retrieved the latitude and longitude for the birthplaces of 43,440 out of 54,314 unique actors and directors!

Another significant challenge was implementing the zoom feature when clicking on a country. This required having the latitude and longitude of each country's center point, as well as a scale to adjust the zoom factor, which we initially lacked. To address this, we found two datasets: one providing the latitude and longitude of each country's center, and another containing extensive data for each country, including its area in square meters. Using the latter, we inferred the scale factor by assuming the land was uniformly distributed in a circular shape around the center point.

BAR/BUBBLE ACTOR CHARTS

Description

To visualize the density of actors in each rating range during different years, the number of actors involved in movies produced during each year is plotted, and for each year, the IMDB the number of actors involved in movies falling in a certain IMDB range is shown in the stacked bar chart. The chart is interactive using Plotly.js, and one can zoom and inspect different parts of the years and IMDB ratings. For each rating range during a year, the bar chart can be clicked on which takes one to the next chart which is the bubble chart.

Design

As mentioned in the actor/director network, the implemented network slightly deviates from our provided sketch by using different colors.

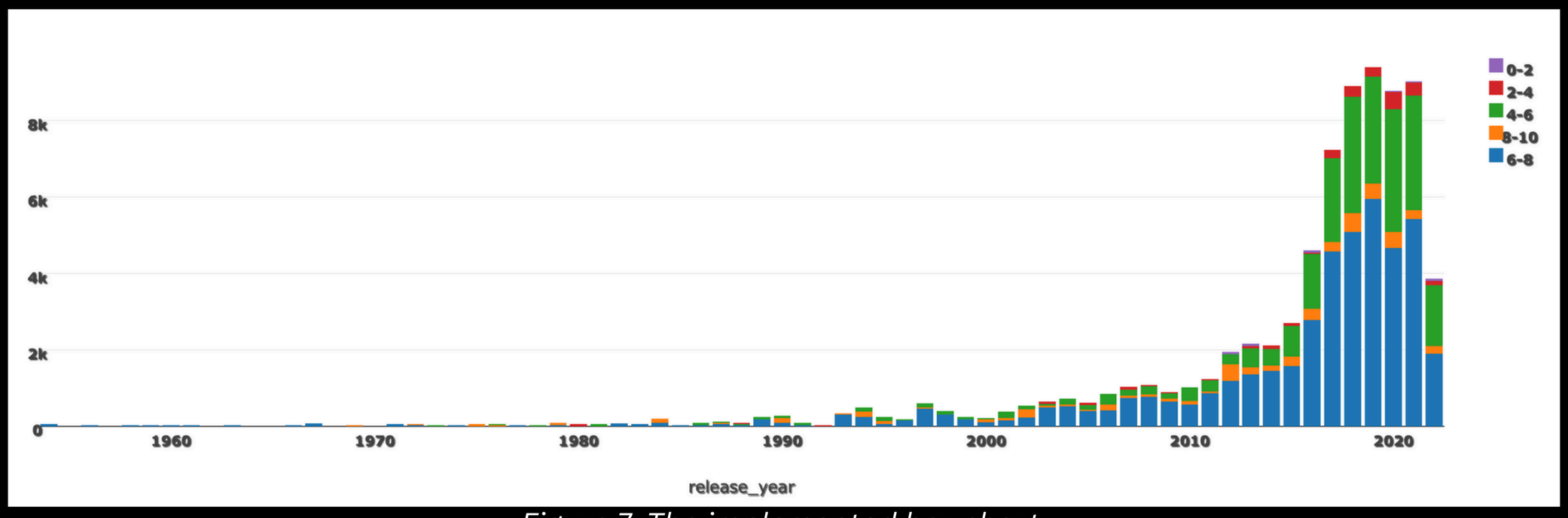
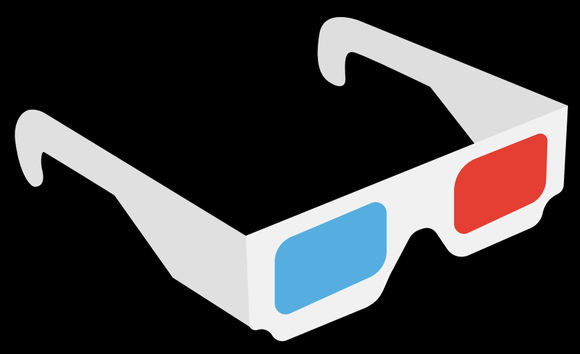


Figure 7: The implemented bar chart.

Challenges

The main challenge of this visualization was to preprocess the data and transform it into a form that can be used for the different IMDB ranges for each year. To make the bar chart interactive so that the viewer would be able top navigate through it we used plotly.js. Another big challenge was to find a way to transition from the barchart to the bubble chart when clicked in a certain range in a certain year.



BAR/BUBBLE ACTOR CHARTS

Description

The main purpose of the bubble chart was to visualize the density of actors in each Genre, possibly showing how certain genres are more affected by actors than others. In this interactive bubble chart, the different genres of movies falling into the clicked-on IMDB rating range during a year is shown, with the diameter of each bubble being proportional to the number of actors falling in this genre. Once clicked, the idea is to show the top 5-10 actors in this genre for this year.

Figure 8: The bubble chart sketch.

Design

The initial idea was to show only some of the top genres in certain IMDB Range, however, with a bubble chart one can see all different genres with their different importances .

Challenges

The biggest challenge was the transition between the bar chart and the bubble chart, and the transition back to the bar chart. Another challenge was how to dynamically allocate the different sizes based on the number of actors in a certain genre.

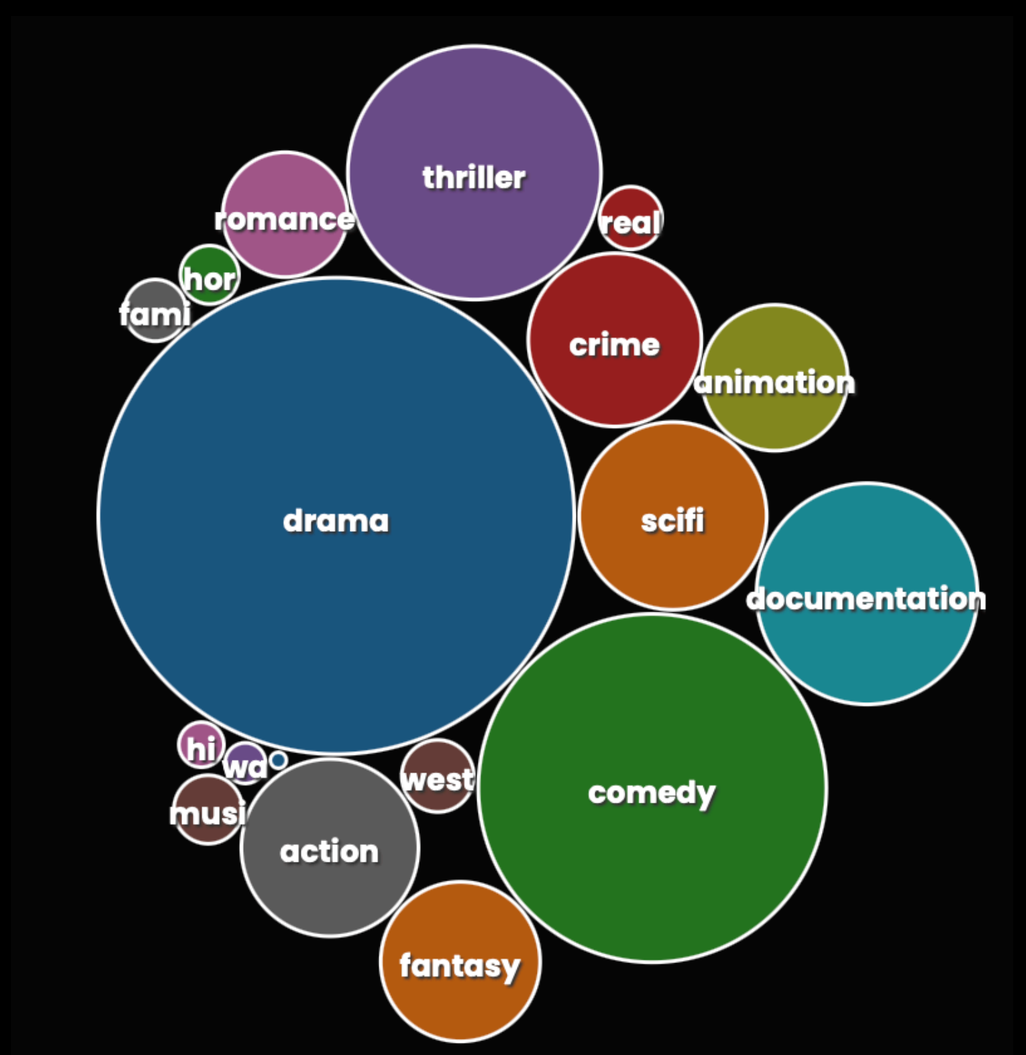
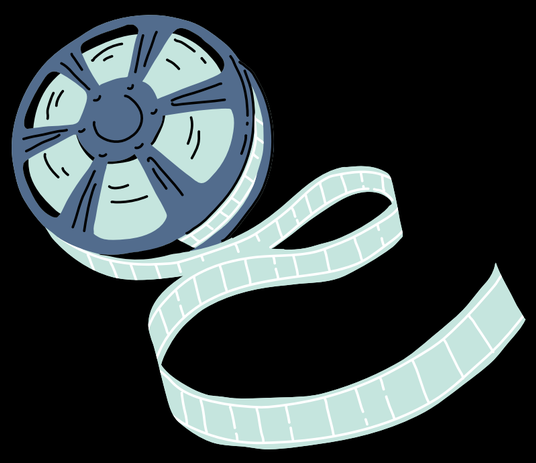


Figure 9: The implemented bubble chart.



TECHNOLOGIES

We used Python for the exploratory data analysis and for preprocessing.

Using HTML, CSS, and Javascript allowed us to structure and design our website.

D3.js enabled us to create interactive and aesthetically pleasing visualizations.

Plotly was mainly used to visualize the interactive bar chart.

PEER ASSESSMENT

All three members had to pre-process the dataset for their visualizations. The process book and reports were a joint effort. Individual tasks were assigned as follows:

Patrick Louis Aldover worked on the line chart and the network graph. He also made a minor contribution to the world map and the bar chart visualizations. Both he and Jérémy worked on the overall design of the website. He also wrote the text for the screencast.

Jérémy Chaverot worked on the world map visualization. He was also mainly responsible for the initial design of the website. Moreover he took care of recording the screencast.

Jason Mina worked on the bar chart and the bubble chart.

GitHub repository Website link Screencast link