Albert Frei
Gregorius Sukiman
Tayler Tolman

Process Book

Overview and Motivation: Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.

As we are all movie and tv show enthusiasts the topic of collecting media data and sorting it was of interest to us. While rooting through different sources of data we stumbled upon a cache of Netflix history. Netflix has recently been in the news and a hot topic among the streaming community. To this day it is probably the most popular streaming site of all time. Along with that we want to see how maturity rating has changed over time in Netflix, and if certain directors/show creators are tied to certain genres/maturity ratings. This information, while also interesting, could project the volume of maturity ratings in the future as well as inform us of past trends. Potentially pointing towards events that warped the entire landscape of the video industry!

Related Work: Anything that inspired you, such as a paper, a website, visualizations we discussed in class, etc.

Pie/donut charts get a lot of flack but I like this article about them https://www.oreilly.com/content/in-defense-of-the-pie-chart/ by Randy Olson. I think they still have their place and wanted to add one to the project.

Questions: What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?

First set of questions:

Do maturity ratings show trends over the years? Are the most popular directors sticking to one maturity rating? Are some maturity ratings more prolific than others?

Some new ones we thought of:

Can we pinpoint high/low points and tie them to real world events? Are some genres more saturated with a single maturity rating than others?

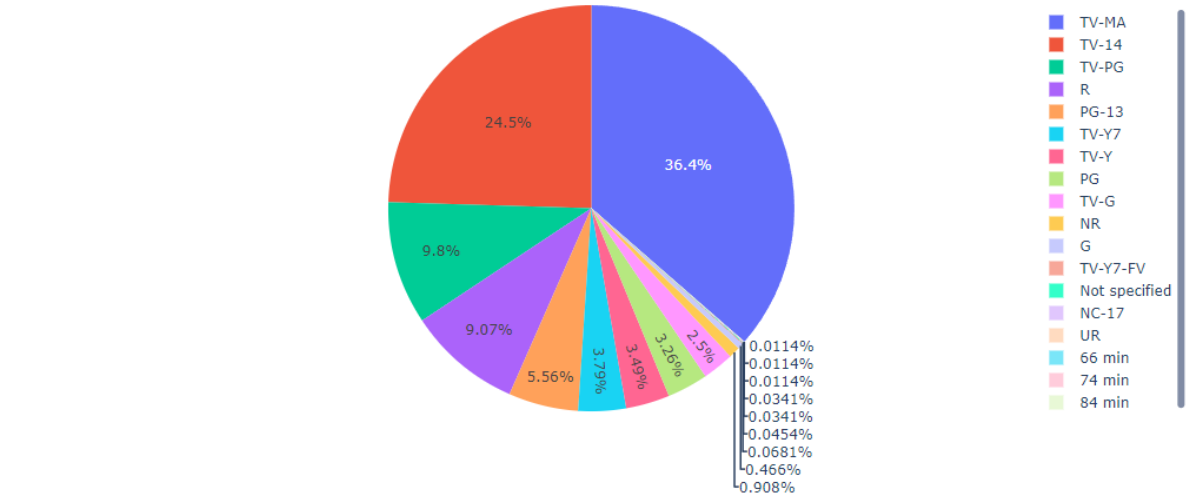
Data: Source, scraping method, cleanup, etc.

For data, we are going to collect our data from kaggle (https://www.kaggle.com/datasets/shivamb/netflix-shows). This website will give us the data in a csv file which we can download.

As far as cleanup is concerned we needed to throw at any null dates and filter to only movies for the time being (tv show ratings is a stretch goal). D3 and JS grouping/filtering methods will easily work on the csv and resulting object data. Unfortunately not all movie maturity ratings are the same in every country so we will have to ‘convert’ all movies to share the same maturity scale.

Exploratory Data Analysis: What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

There are plenty of visualizations that show movie/tv show data, even some specific to Netflix. One design in particular that caught our attention was:



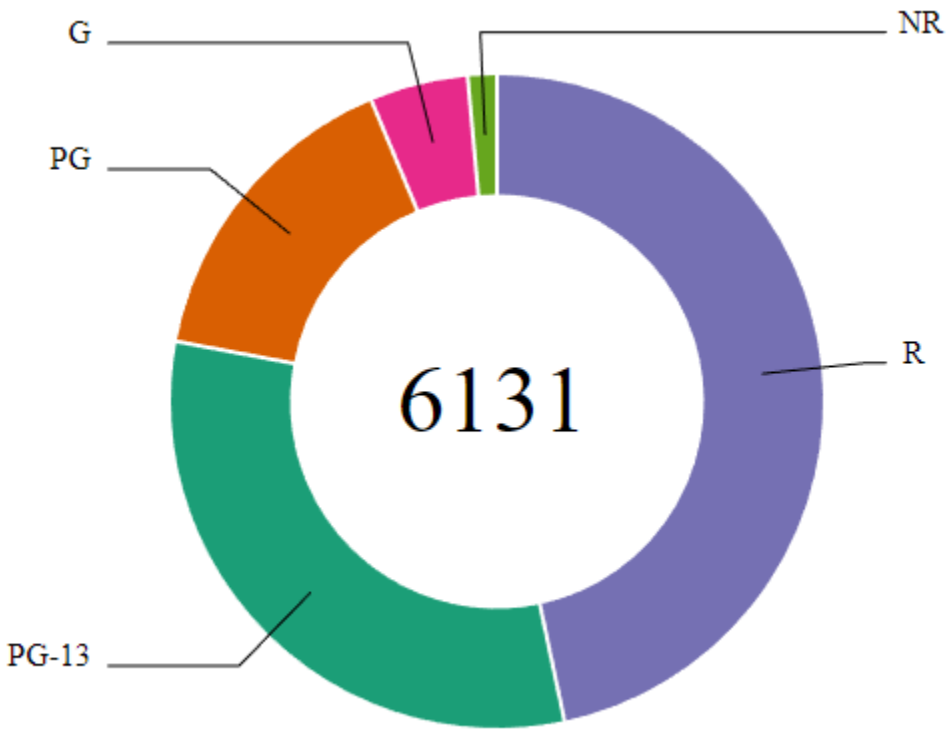
This graphic shows maturity ratings as breakdown, however we discussed how the pie chart and percentage list did not make for a good visualization. There are too many categories and the slices are rather uniform in size on the smaller spectrum. Furthermore it is rather uninteresting by itself. We tried to fix many of the design issues with the donut chart and add a story element with the passage of time.

After visualizing the data with a line-graph, we realized that the total number of movies in the data went down in 2021. This begs the question what happened? We could explore this insight, maybe 2020 was a boom year because of stay at home orders. We anticipate that we will include this observation in the storytelling aspect of the project.

After visualizing the data with the bar chart, the insight that it gave us is the list of directors that have many movies published on Netflix. This gave us a question which is: Do these directors have certain types of genres or ratings? This gave us another insight for our design which is to make some kind of filters for the maturity ratings and use it in the bar chart.

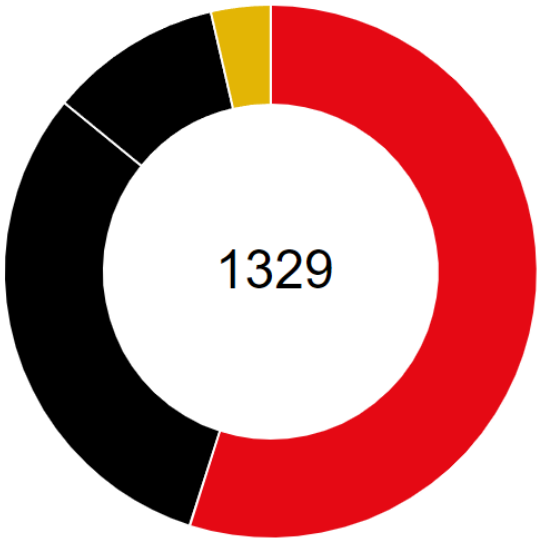
Design Evolution: What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?

First off the initial design of the donut chart:

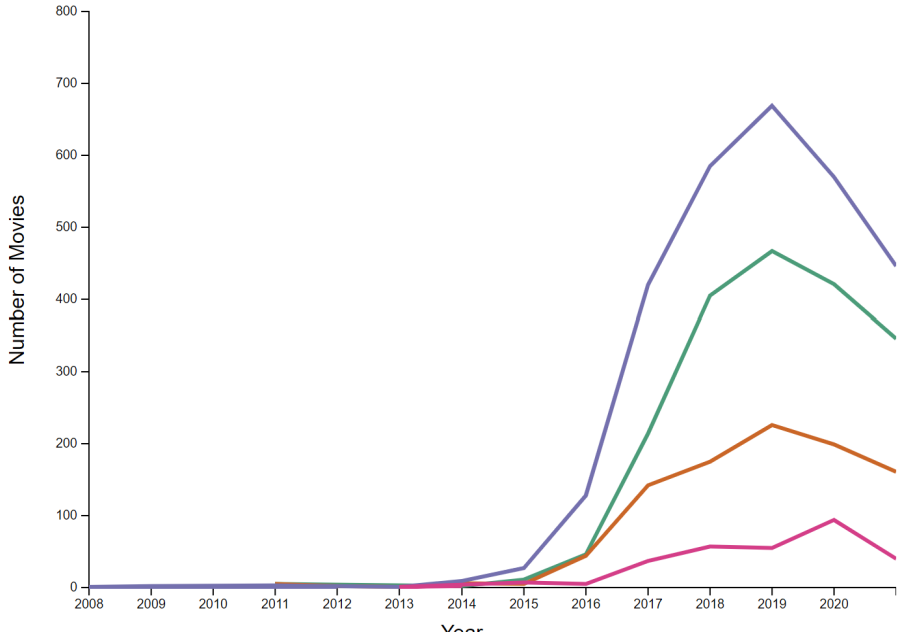


This chart shows a breakdown of the different movie maturity ratings as part of a whole for the date range selected. Pie/Donut charts obviously have their failings when it comes to slice comparison since humans have a hard time judging surface area. However, the percent of each category in our case is distinguished enough that slices are relatively easy to compare. Whether that holds true under different filtering conditions remains to be seen. Furthermore with only 5 different categories our donut chart is not over cluttered and color is used to properly distinguish categories.

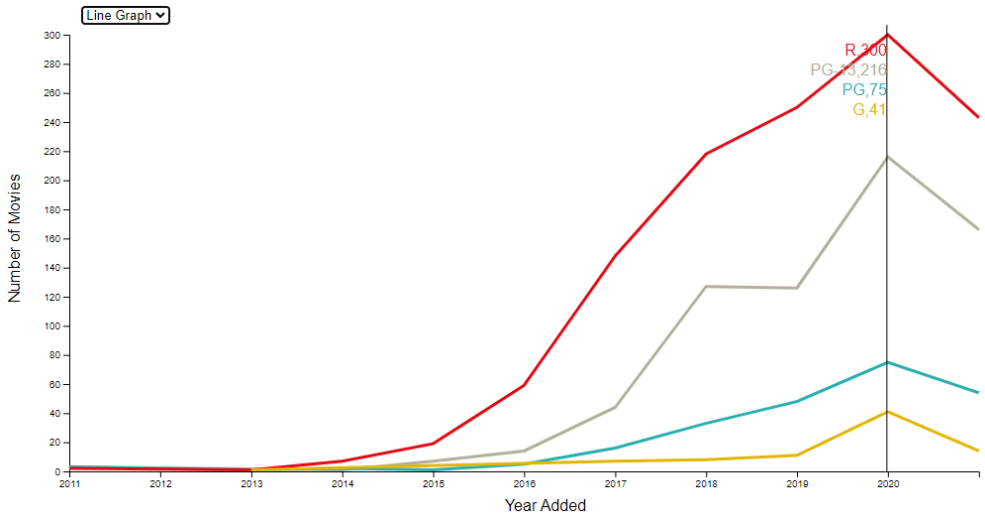
The design was later modified to remove labels, instead putting the legend on our line/scatter graph. Shown below is an example of filtering and without the labels.



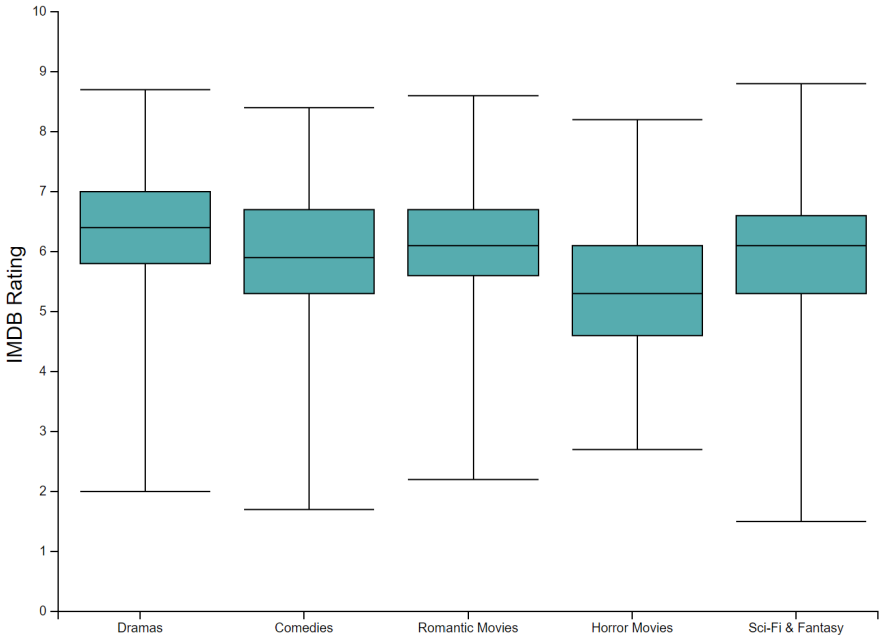
Second, the initial design of the line-graph:



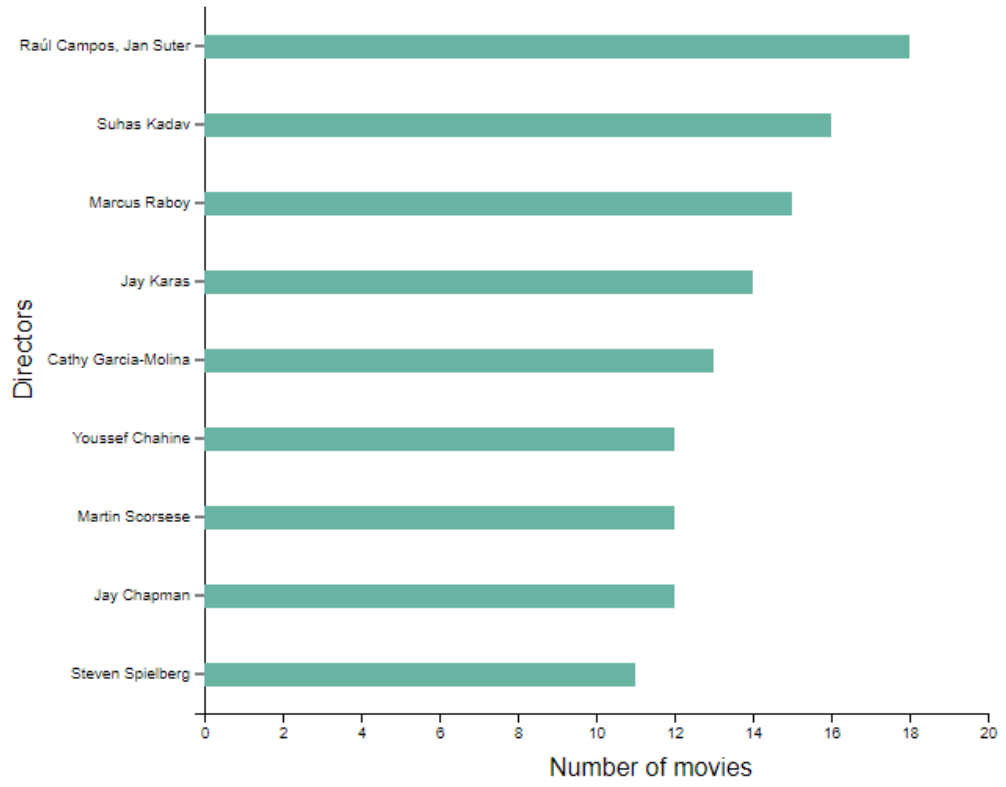
Initially we intended to filter out some of the earlier years in the line chart, but instead we ended up including all data points. Also, we added an overlay line to show us the number of movies that were added on that year to give us certain numbers of it.



We decided to add a box and whisker plot in order to explore the data by genre. The box and whisker plots IMDB rating vs genre. This way we can compare the distribution of ratings across genres.

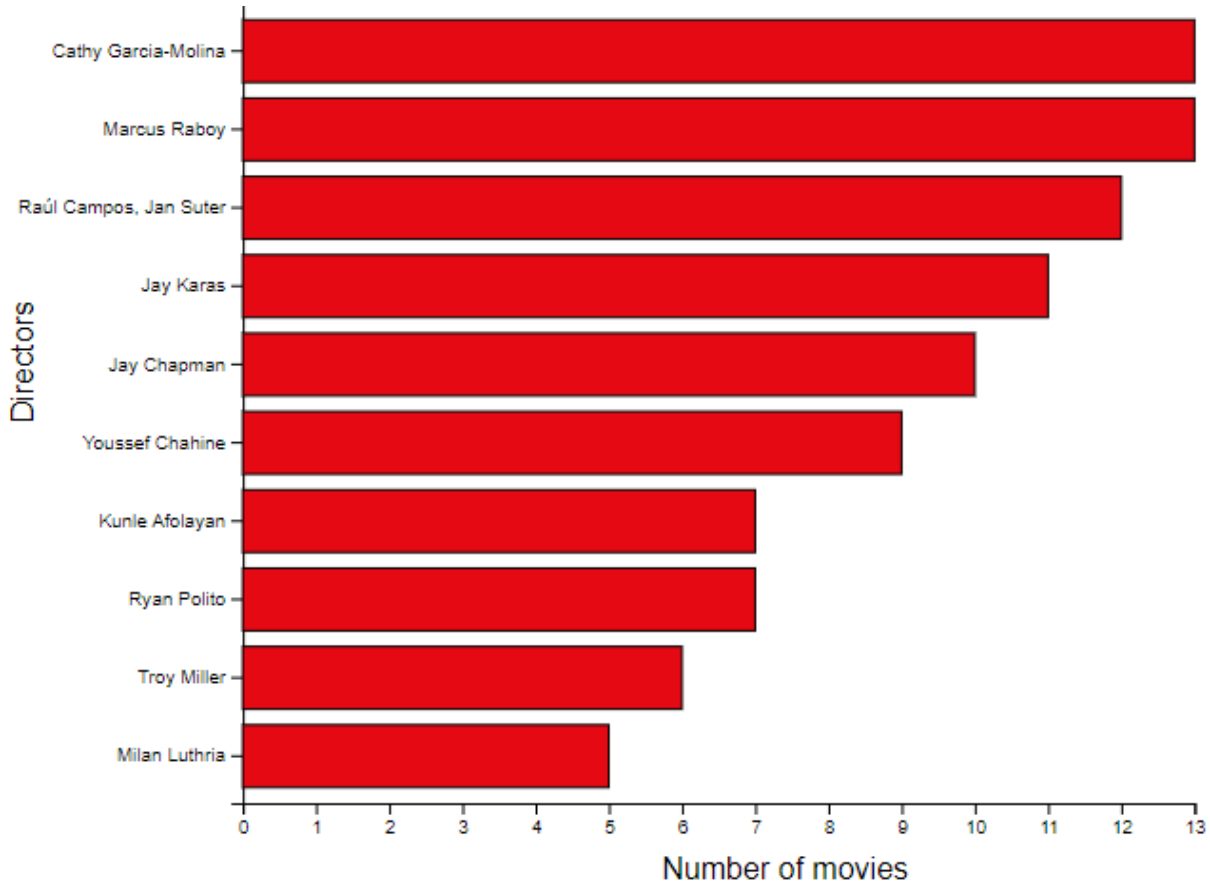


Third the initial design of the bar chart:



This design is to show the list of the directors that have the most number of movies. But, in the data, we do have some data type which represents the TV shows. Having these data we can work on something that we learn in the lecture which is having different colors for the bars, one represents the movies and one represents the TV shows.

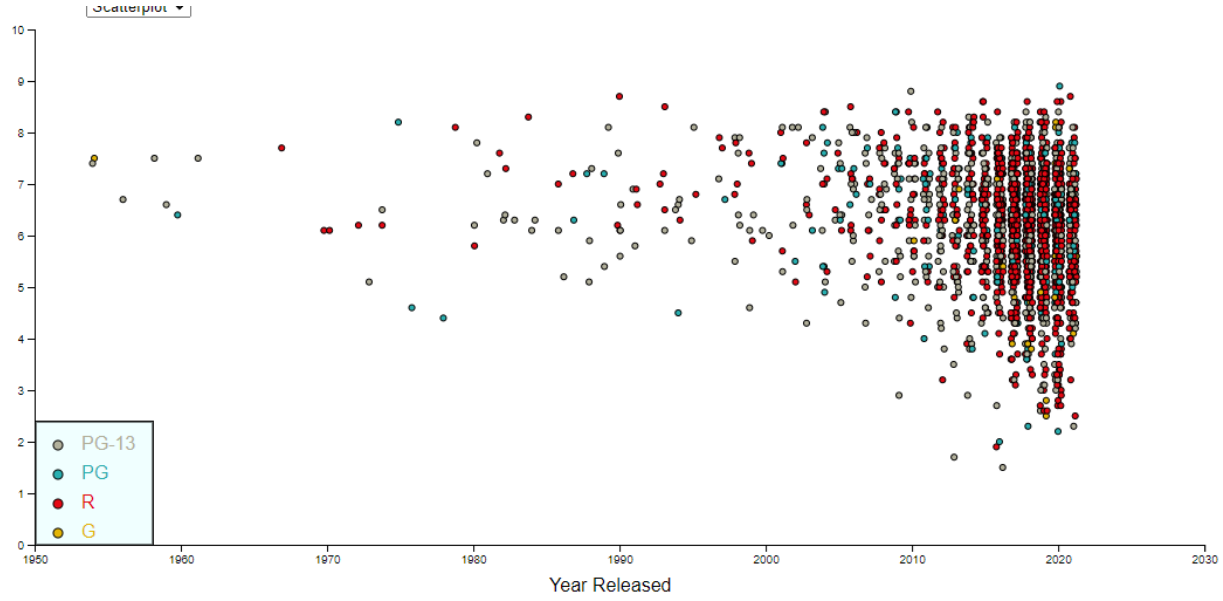
The design was later modified to focus solely on the movies only. We fixed the sizing, gave some padding, and later we gave an interaction and transition with the donut chart that changes the data every single time we filter out the donut chart. Another thing that we did is that when we hover our mouse on the bar, it will show a tooltip which shows all the list of movies that the director made and the maturity ratings. These can be seen in the two pictures below, and in the video for more detailed version:



Movies made by the Director:

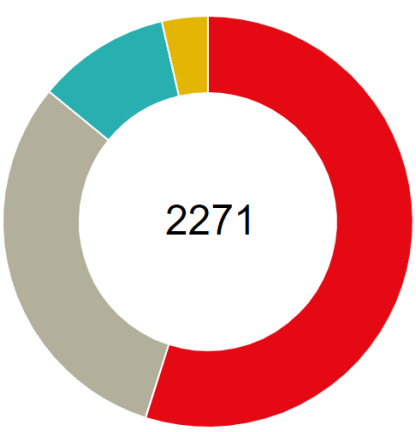
Hello, Love, Goodbye **(PG-13)**
Unexpectedly Yours **(PG-13)**
My Amnesia Girl **(PG-13)**
Three Words to Forever **(G)**
A Very Special Love **(G)**
The Hows of Us **(PG-13)**
Seven Sundays **(R)**
A Second Chance **(R)**
A Second Chance **(R)**
Four Sisters and a Wedding **(PG-13)**
It Takes a Man and a Woman **(PG)**
My Ex & Whys **(PG-13)**
She's Dating the Gangster **(PG-13)**

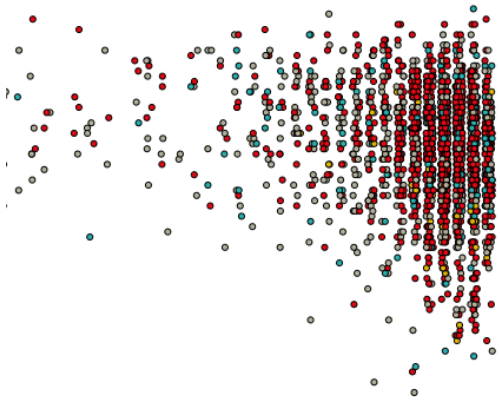
Next is the scatter plot showing movies when they came out and the rating they had at the time. These includes a tooltip (shown later) when a movie is hovered.

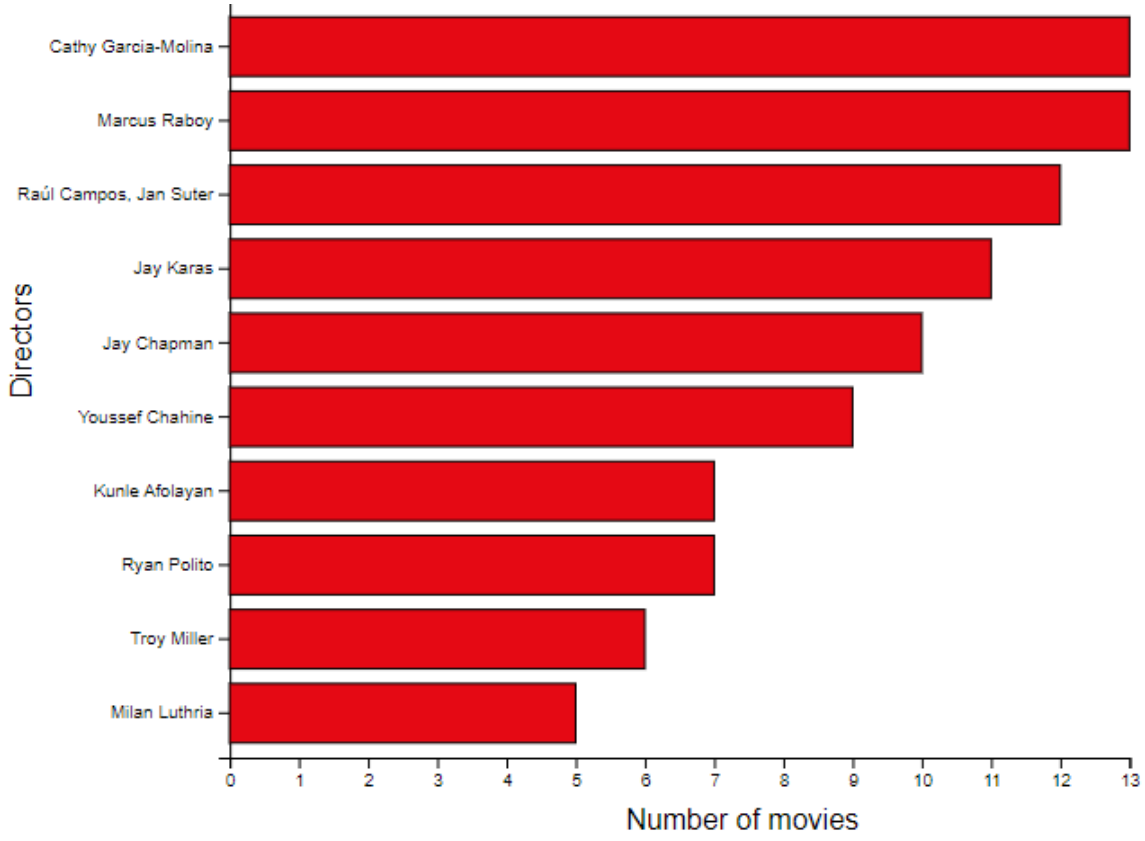


Implementation: Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

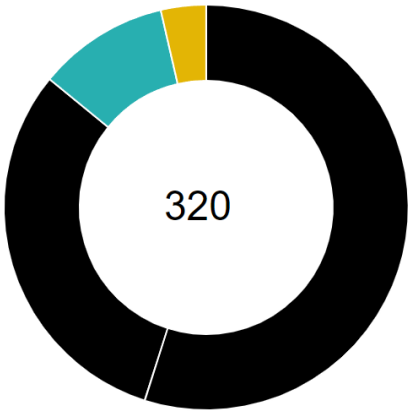
We want to change the bar chart and line chart depending on the selected maturity rating in the donut chart. Currently the pie chart has hover functionality to highlight selected elements. Hovering filters the line, bar and scatter graph to highlight the selected category and clicking should filter the graphs based on the selected category.

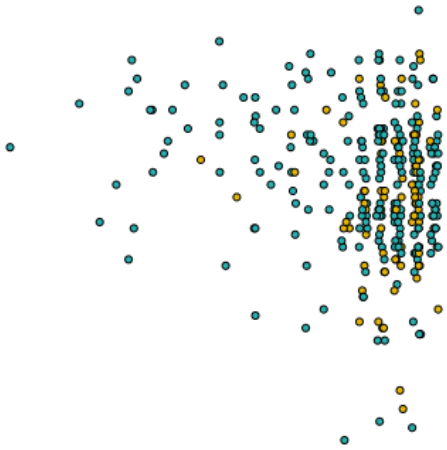

2271

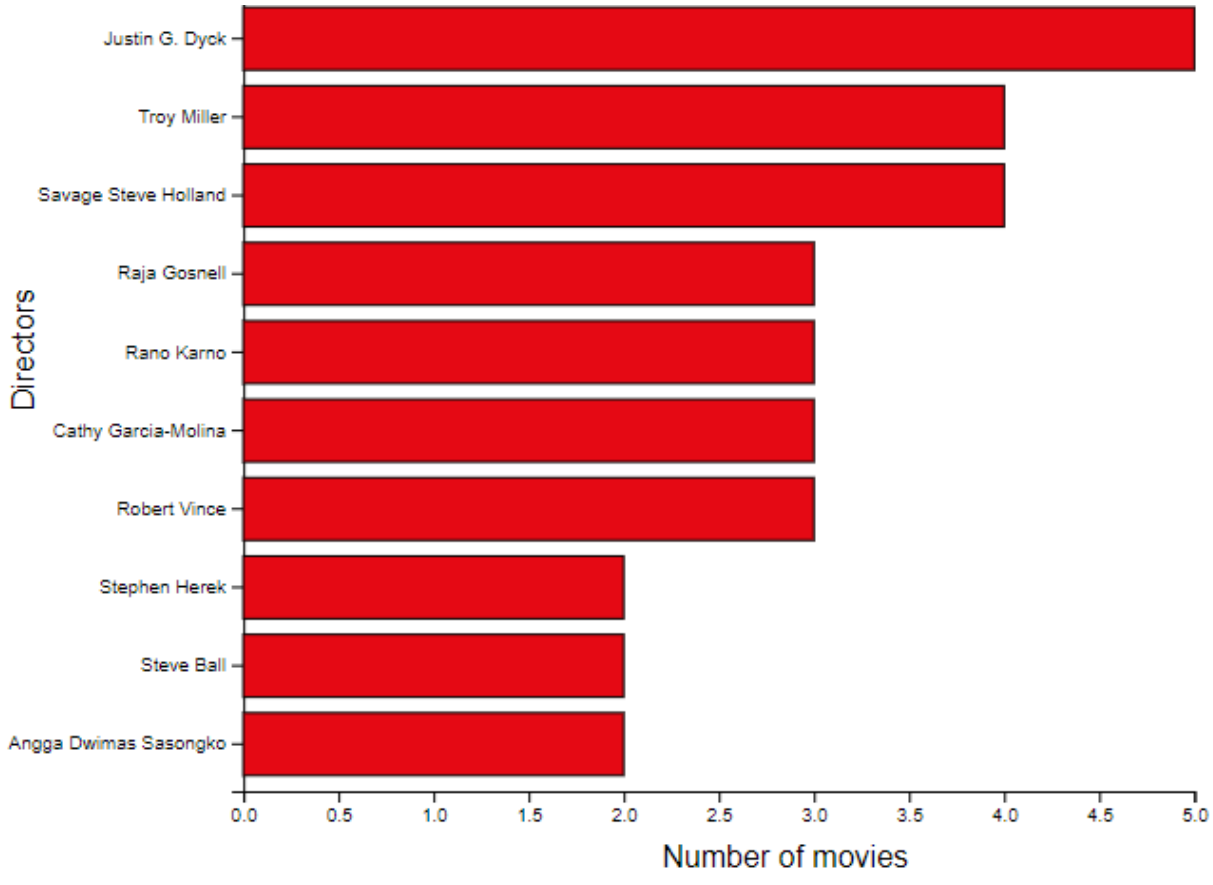



Cathy Garcia-Molina
Marcus Raboy
Raúl Campos, Jan Suter
Jay Karas
Jay Chapman
Youssef Chahine
Kunle Afolayan
Ryan Polito
Troy Miller
Milan Luthria
Directors
0
1
2
3
4
5
6
7
8
9
10
11
12
13
Number of movies

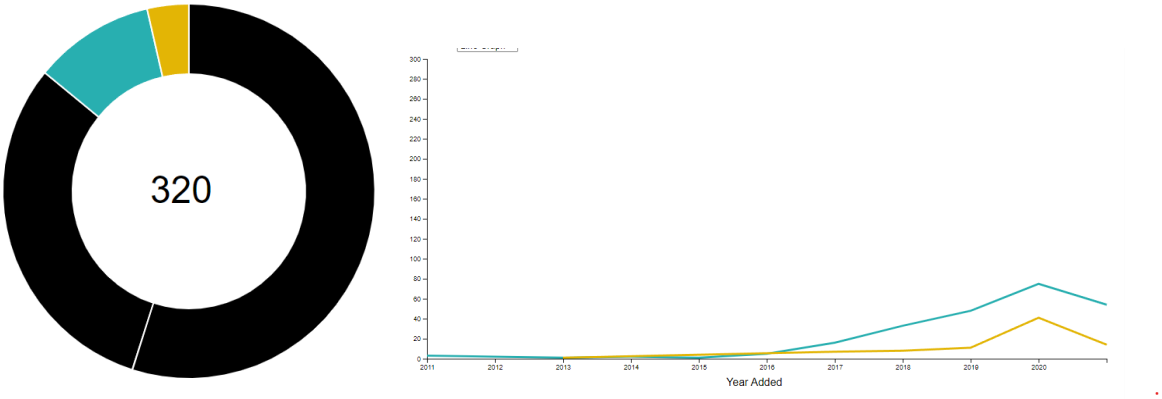
vs


320

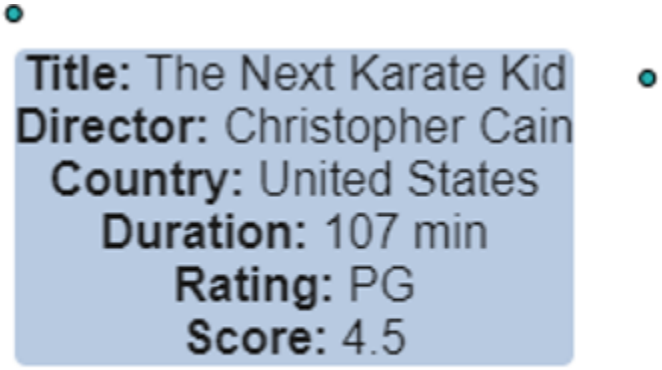



Justin G. Dyck
Troy Miller
Savage Steve Holland
Raja Gosnell
Rano Karno
Cathy Garcia-Molina
Robert Vince
Stephen Herek
Steve Ball
Angga Dwimas Sasongko
Directors
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Number of movies

Of course the line graph will follow suite



For the scatterplot specifically additional information can be found through a tooltip



The color scheme that we used in the final submission was chosen with color blindness in mind. The initial color chosen was the Netflix red. This was put into a complementary color scheme tool and we tested generated colors with the Google Chrome developer tool that simulates different forms of color blindness. While red is typically a problem color for color blind individuals, we made sure that no two colors are going to be indistinguishable for the chromatically impaired.

Evaluation: What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

So far that rated R shows are the most produced on Netflix, and, if we extrapolate, the world. A large portion of movies on Netflix are not in english. The directors with the most movies on Netflix are a mix of directors you probably haven't heard of and big name hollywood directors.

Netflix also tends to include modern movies, with our data showing a huge spike of acquisition of movies sometime after 1990. Additionally all old movies seem to have higher ratings, indicating that bad movies were probably not acquired or forgotten about completely. Furthermore the movie maturity rating seemed to have no correlation with the average user score of a movie. A G rated movie is just as likely to be bad as R rated.

However we can say that horror movies on netflix tend to have lower median ratings than other common genres and Sci-fi movies surprisingly have high user ratings, but also one of the largest ranges between max and min.