Process Book

Benjamin Mastripolito, Alper Sahistan

Repository: https://github.com/benpm/moviz

Live web app: https://benpm.github.io/moviz

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.*

We like stories that individual movies tell, but we also think movies collectively tell us stories. Visualizing the movie data, such as budget, rating, release date, oscar, and genre, can show us interesting information about individual movies, trends, industry direction, and much more. We also want a practical tool that we can use to look up movies to decide what to watch that day.

We think that rather than letting user search through uninteresting visualizations, we could tell several interesting stories we pre-defined while giving room to explore meaningfully. Therefore we decided to lead user through curated three different views which can be switched between.

Related Work

> *Anything that inspired you, such as a paper, a web site, visualizations we discussed in class, etc.*

We were inspired by many different film visualizations that we've seen, such as Movie Colors, as well as visualization techniques we were introduced to in different forms in class, such as heatmaps and stacked area charts.

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?*

There are multiple things we are trying to convey through our visualization:

- What film genres compose oscar nominations and winners over its history?
- How can we display film data for both film discovery and for learning interesting trends from film history?
- How has the market share of various film studios changed over history?
- What benefit is gained to a film·s popularity, commercial success, and rating by having a larger budget?
- How have different measures of success, measured together, changed for films over the decades?
- What relationship do measures of financial success and measures of popularity share among films over the years?

These questions evolved as we discovered new aspects of our data, as well as decided to add more functionality. For example, as we realized that the combined metric of profit might be a more direct indicator of success than revenue, we began investigating the relationship between profit and other measures of success, such as ratings. This revealed new outliers in the data.

Data

> *Source, scraping method, cleanup, etc.*

There are three data sets that we are using as our raw data sources:

- https://www.imdb.com/interfaces/
- https://www.kaggle.com/datasets/danielgrijalvas/movies
- https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset

All data processing is done in /data_processing using pandas. These are the primary operations that are performed on the data:

- Remove all rows from the main movie dataset without corresponding rotten tomatoes scores
- Remove all rows without one or more of the following fields: `"released", "budget", "gross", "score", "year", "genre", "company"`
- Remove all duplicates of name and year
- Determine the oscar status of each film in the remaining table and assign that to a new field
- Count oscar nominations for each film and add it to a new field
- Compute profit from budget and revenue
- Combine similar production company names using fuzzy string matching
- Add inflation-adjusted versions of monetary fields

Later, when we decided to utilize semantic zooming in our scatterplot, we also wrote a script (simulate.py) for taking the scatterplot dots and performing basics physics simulation and hierarchical combination of near points.

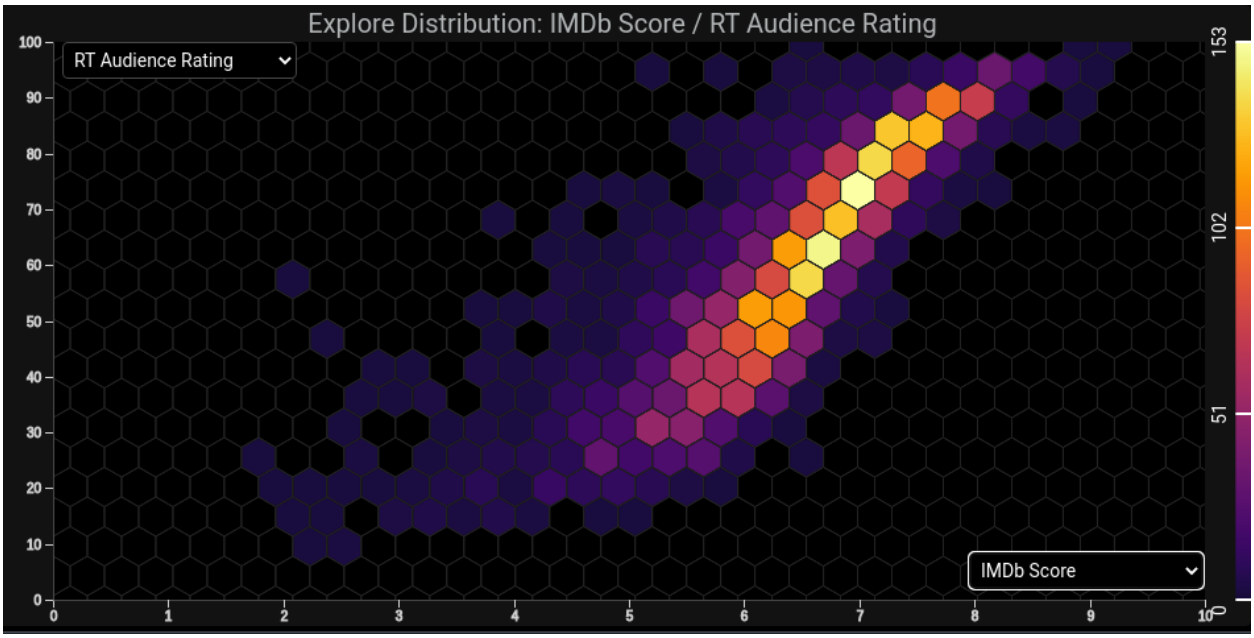
Exploratory Data Analysis

> *What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?*

The scatterplot view was the first used to visualize the data. We realized that the distribution of scores between IMDb, rotten tomatoes tomatometer and rotten tomatoes audience are very different. Tomatometer is extremely uniform.

We also noticed irregularities in the budget data. As shown below there is some aggressive rounding of film budget:

The next quirk of the data we noticed was a nearly linear correlation between the three score metrics, which was most visible within the heatmap view:



Another thing we noticed was that monetary metrics such as budget, revenue, and profit were very fit for logarithmic scaling. We used this information to inform our choice of axis scaling.

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?*

Our design changed significantly over the course of the project. We'll start with our initial sketches.

Sketches

Visualization for Data Science - Vis ideas

Data

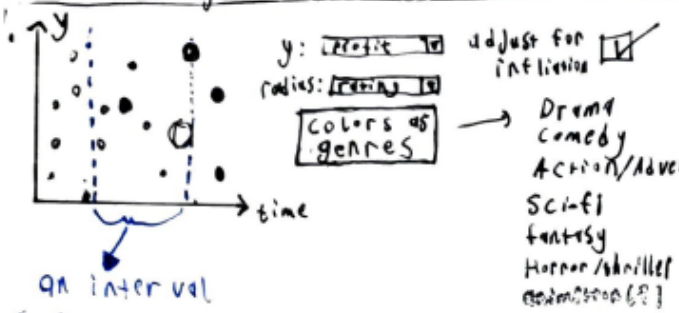
Plain

Name, date, rating, genre, cost, earnings, studio(?)

Derived

Profit, inflation adjusted: cost, earnings, profit, rating/cost, rating/profit, rating/earnings

---



Drama
Comedy
Action/Adventure
Sci-fi
Fantasy
Horror/thriller
animation(?)

Selector can be binded to another vis → or many

Posibble problems:

- Cluttering
- Color channel may not be over-used. (shapes/icon maybe)
- Solution to cluttering: Zooming?, time range?
- Too plain?
- Search bar? filters?

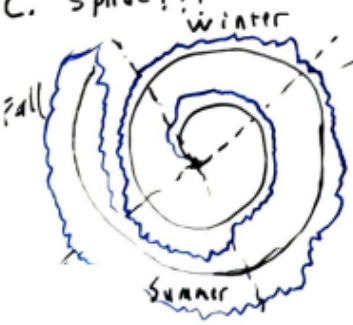
a.

The market share of studios between years "1963 - 1989"

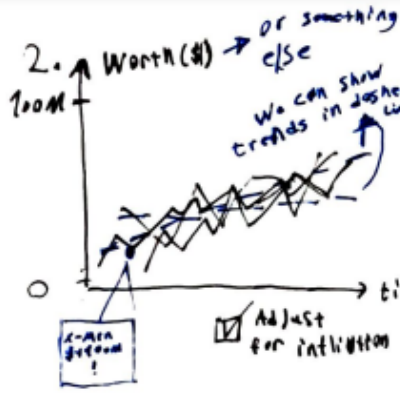
Average movie rating per 5 years for selected region

→ Maybe this can be a seperate thing without interaction

c. Spiral!!!



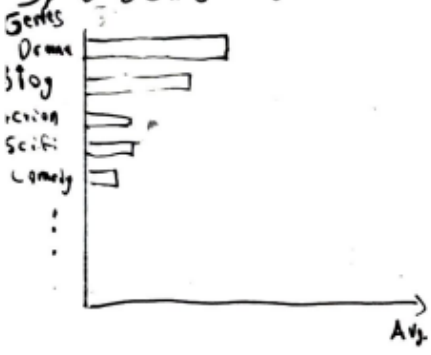
→ Earnings, rating or Profit there will be cycle
Summer is the block buster Season

2. 

Studios Profits over years
Each studio different color
Each data point is interactive
mouse sliders can open up a dialog displaying movies for that point in specific along side their profit.

a) A time range selector can show movie ratings

3) Oscars Vs Genre → I think this is a veeeery good idea
Maybe as a side vis!!!



→ Can be improved
Open to ideas!!!

1. Movie Discovery View

a) Scatter Plot – Main

y = ratings

X = time

area = ?

Color = 4 colors → nominated
won
nominated best pic.
won best pic.

b) Stacked Line Chart – Sub

y = Oscar nominations, Oscar winners

X = time (selected from main)

mark = Lines (Stacked) → Genres

Color = ?

2. Budget & Studio View

a) Scatter Plot – Main

y = Profit, cost, earnings ——→ Line (Avg.) → moving

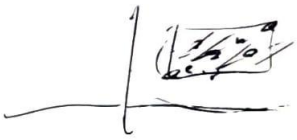
X = time

Color = green/red (redundant)

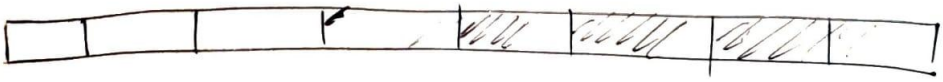
b) Pie chart – Sub

angle = Avg. Profit for selected period of time (Market Share)

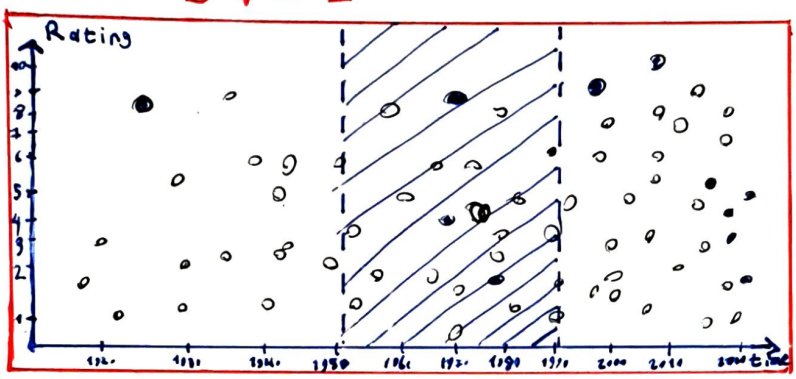
Colors, text = Studios



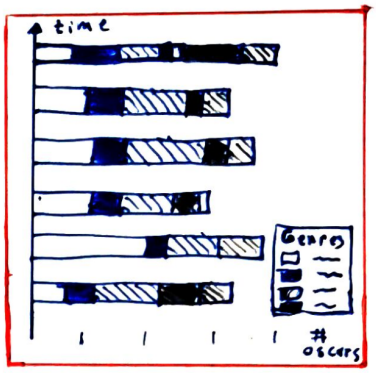
3.



SVG I



SVG II



X = release date

y = Rating: IMDB, or Rotten Tomatoes...

Color = Oscar Nominee, Oscar winner, Best Pic. Nominee, Best Pic. winner

Interaction #1: Time Range selection

Interaction #2: Hover on point to get further info about a movie

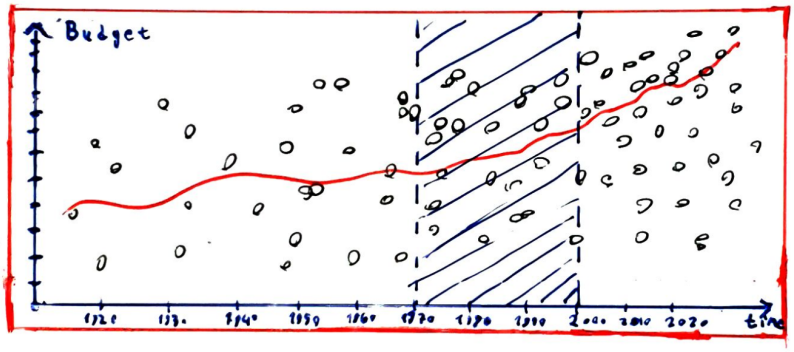
for selected range →

y = Years in range

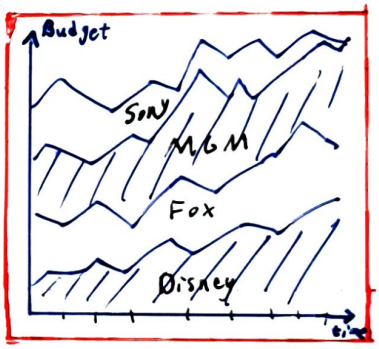
Length = # Oscars or # Nominations

Colors = Genres

SVG I



SVG II



X = release date

y = Budget, cost or Profit

Color = green to red to emphasize earnings (redundant)

Interaction #1: Time range selection

Interaction #2: Hover on point to get further info about a movie

Interaction #3: Toggle avg. trend line

Interaction #4: Adjust for inflation

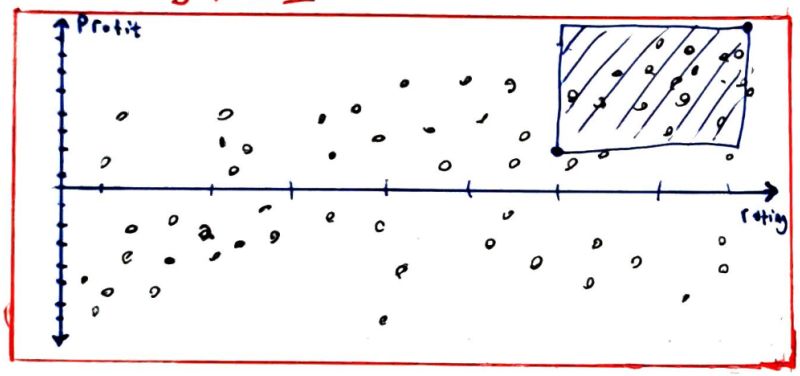
for selected range →

y = Studio's: Market Share, spending or profit

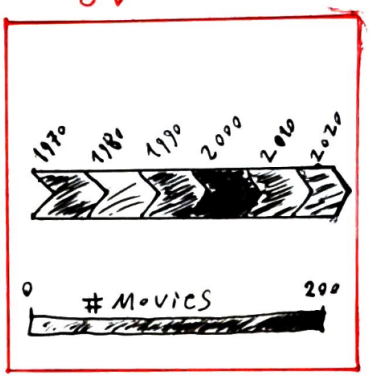
X = time selected in range

Colors/Areas: Studios

SVG I



SVG II



X = rating: IMDB or Rotten Tomatoes

y = Profit, Budget or cost

Color = Red/green profit (redundant)

Interaction #1: Select an area - - - - - - - - - -

Interaction #2: Toggle adjust for inflation

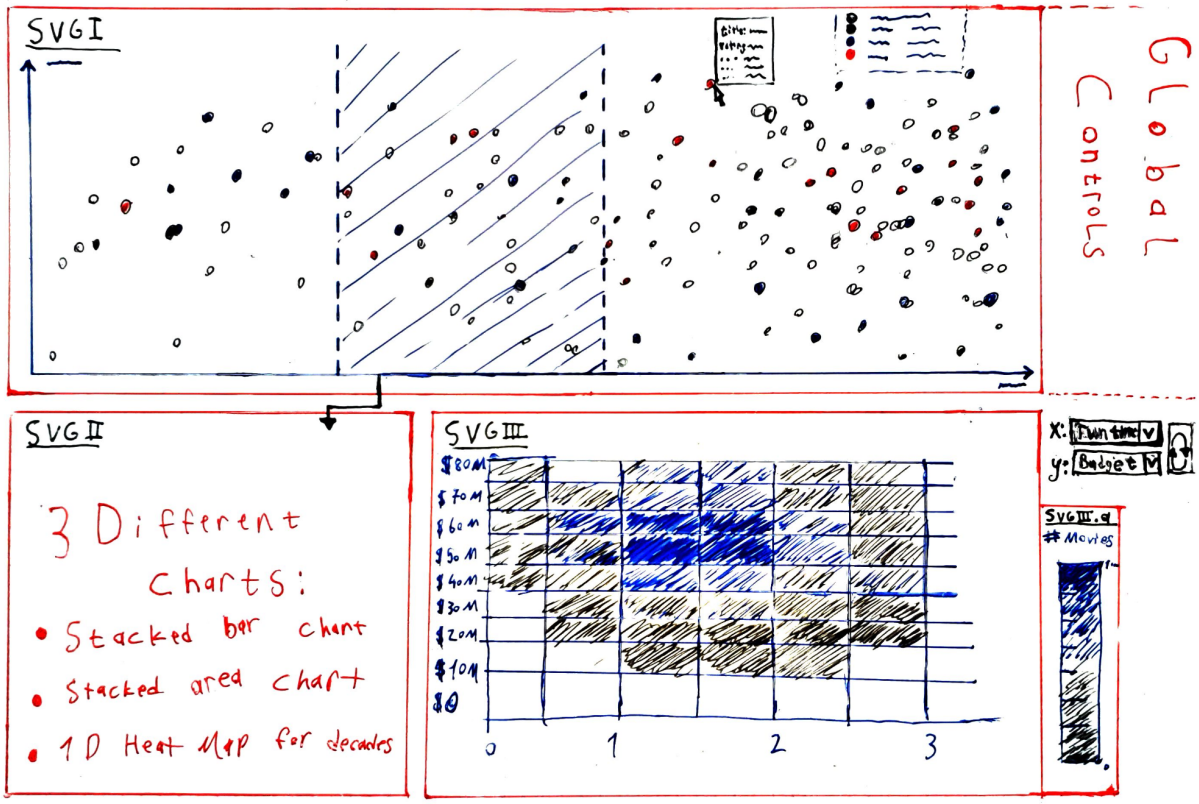
Interaction #3: Hover to reveal movie info

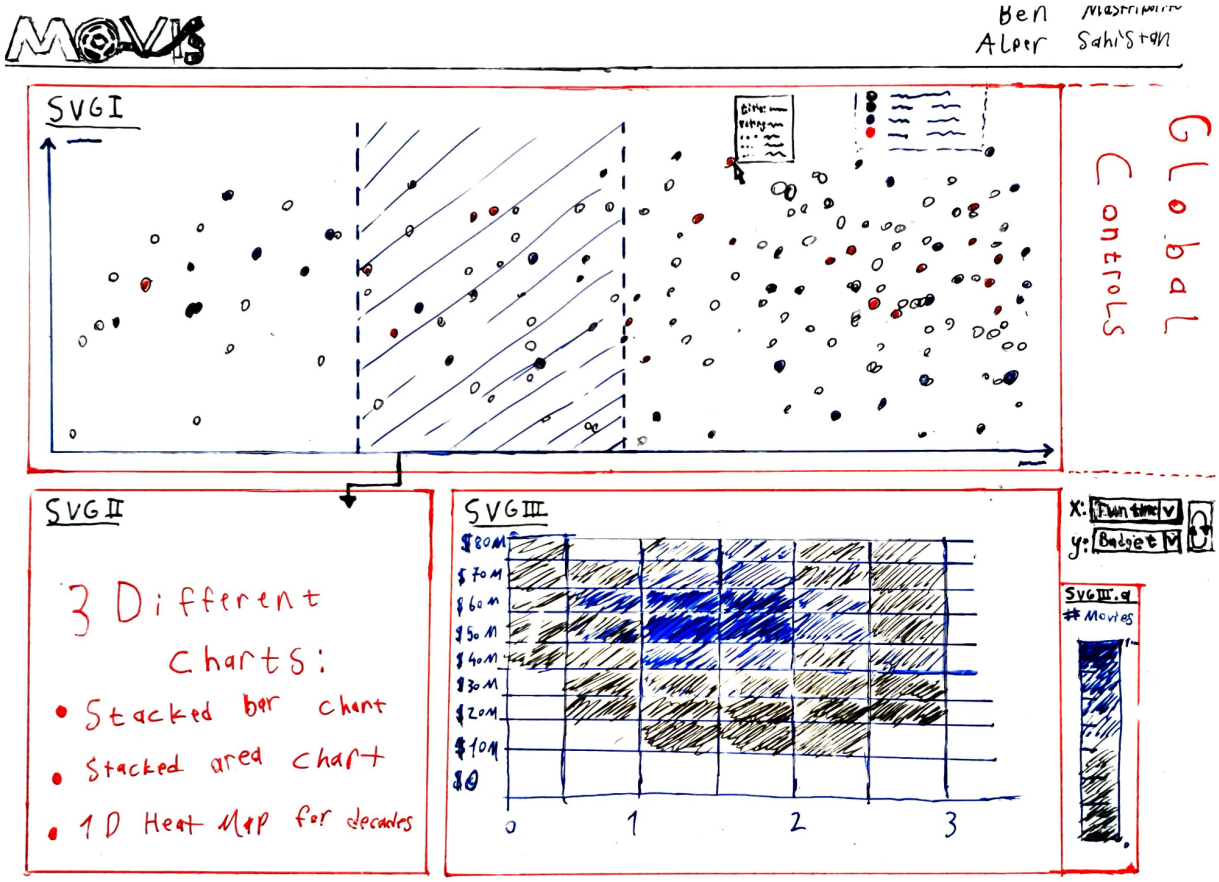
→ X = decade range for selected movies' release date

Color = # movies in given decade

MOVIS

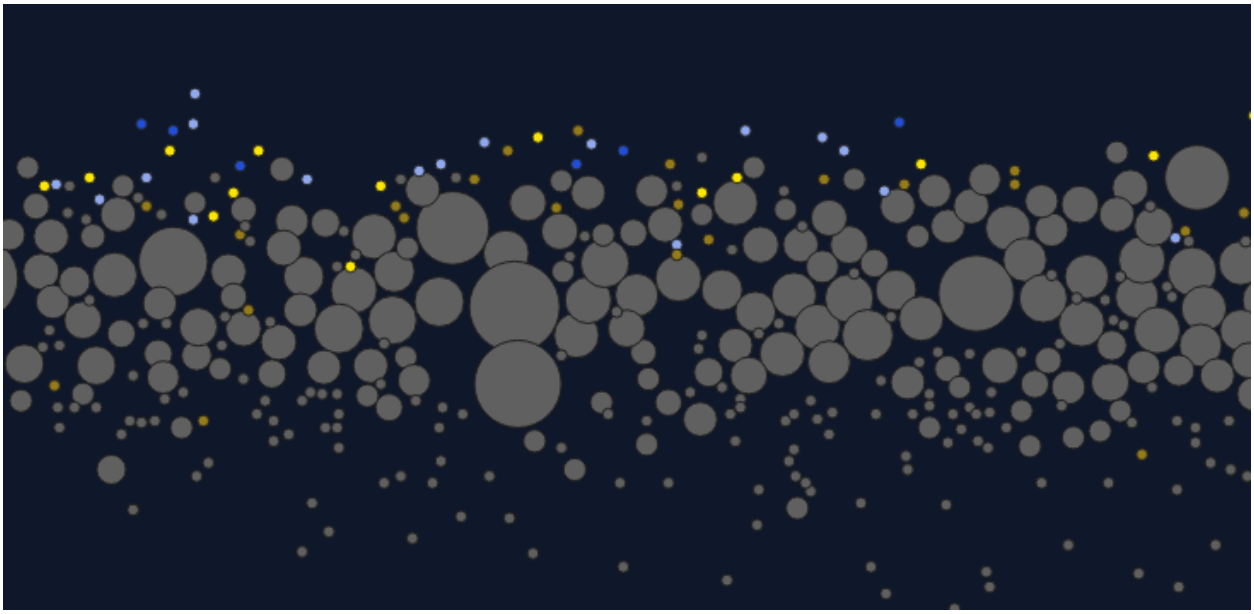
Ben Alper
Sahil Shah





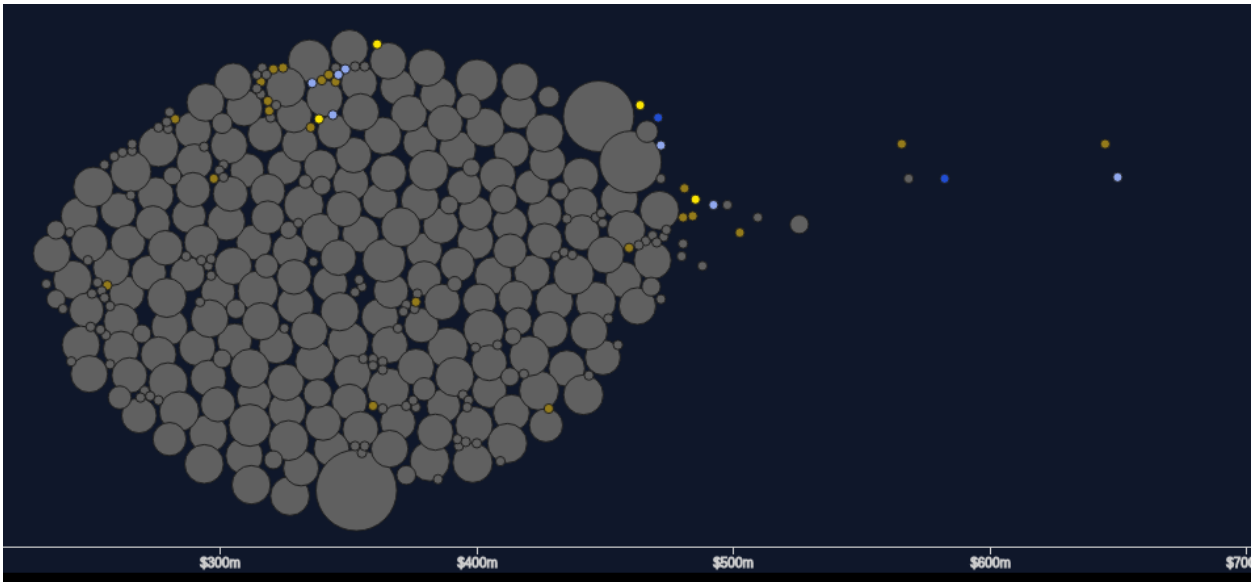
Scatterplot

The largest divergence from our original designs was the scatterplot. In the images above, we show the scatterplot as a typical dots-on-a-plane plot. After our discussion with our assigned TA, we decided that the high degree of overlapping was undesirable, so we decided to attempt a **semantic zooming** approach. We achieved this by coalescing dense regions of circles into larger circles when the user is more zoomed out:

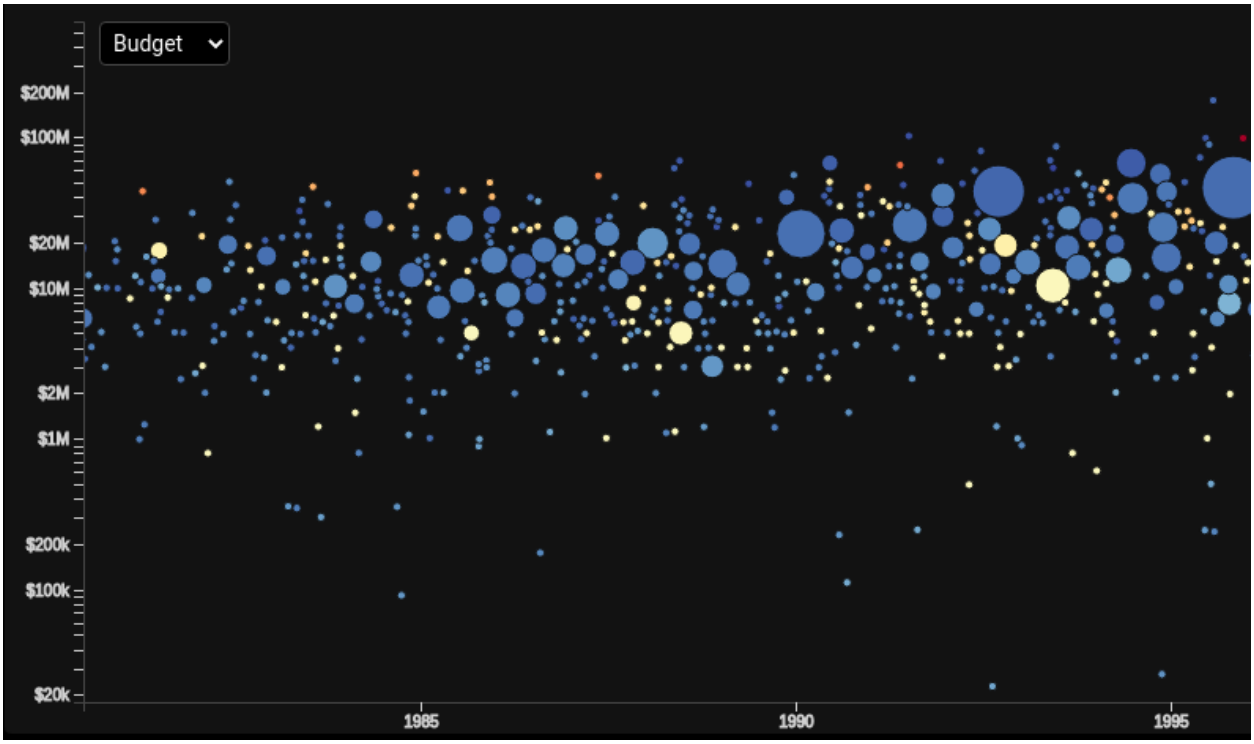


We decided to pre-compute the entirety of the grouped scatterplot out of fear of performance issues. We achieved this by writing a Python script that loaded all the data and performed one physics simulation per zoom level per combination of axes, then saved everything out to a CSV.

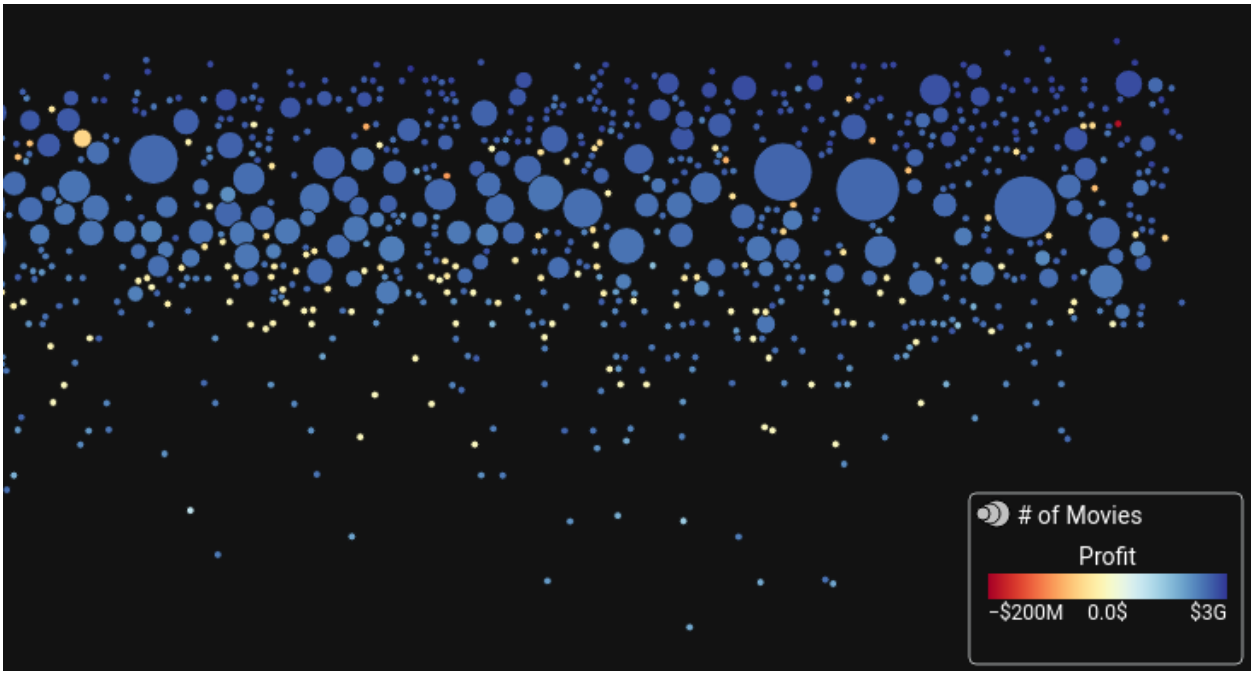
The next design iteration came after we noticed the effect the grouping and simulation had on budget and revenue plots:

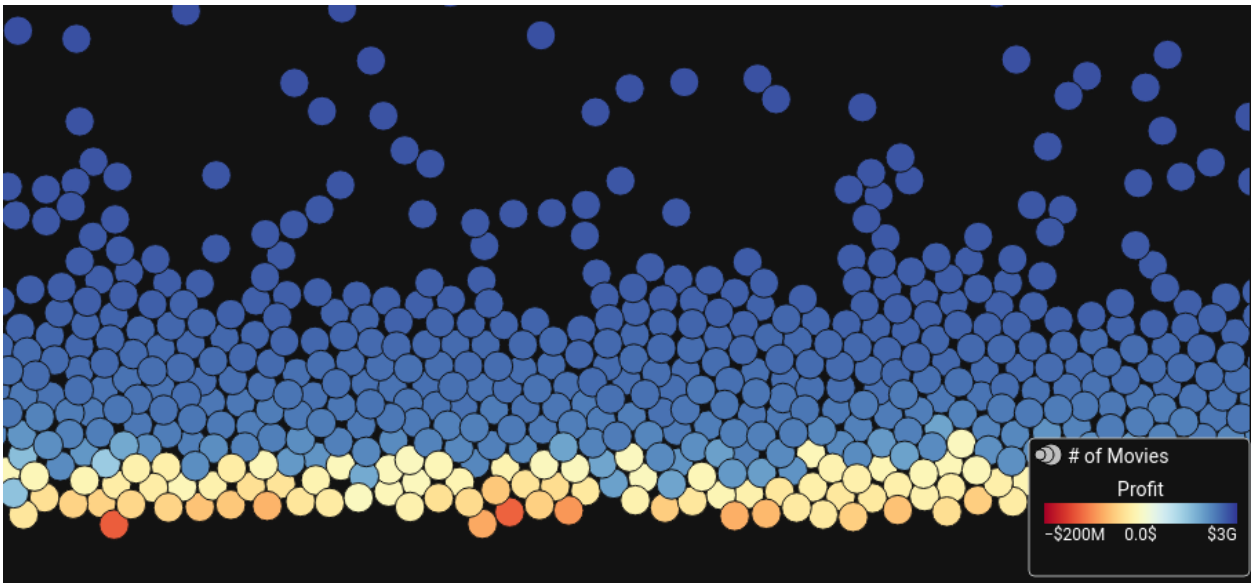


We saw this as a very clear indicator that we should be using log-scale for these particular axes:



Next, we decided to color the dots in the Movie Economy and Cost v Quality views by their profit. We went through several iterations of colormapping schemes before landing on something we both liked:





This process was a balancing act between a best-case visual encoding of profit while still maintaining a pleasing look.

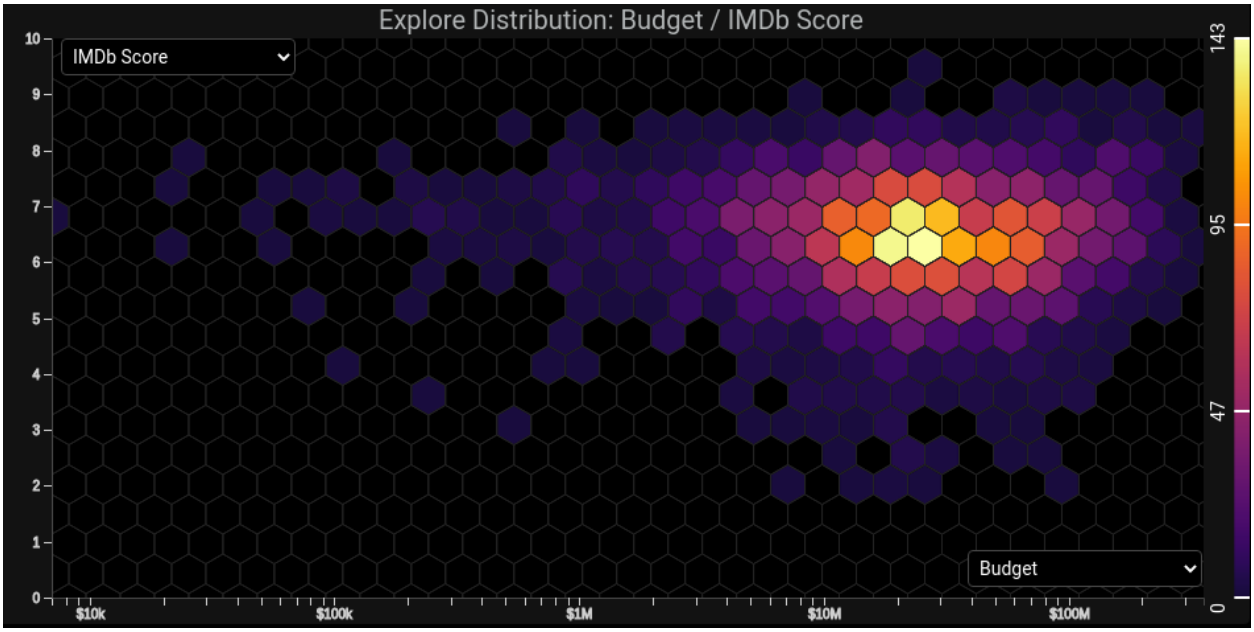
Dashboard

The dashboard went through many iterations. Our initial design included controls in a panel on the right side of the screen, but we eventually decided there weren't enough elements in the panel to warrant that usage of space, so we moved the controls to the navbar:

This included many iterations on the appearance of the view mode selector.

Heatmap

The heatmap began as grid of squares, but we eventually decided it would be more interesting as a grid of hexagons.



Implementation

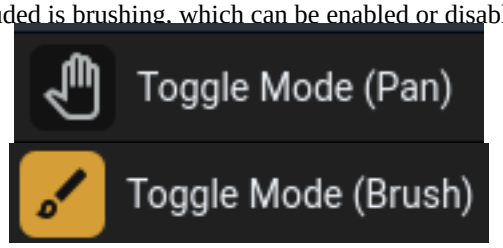
Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

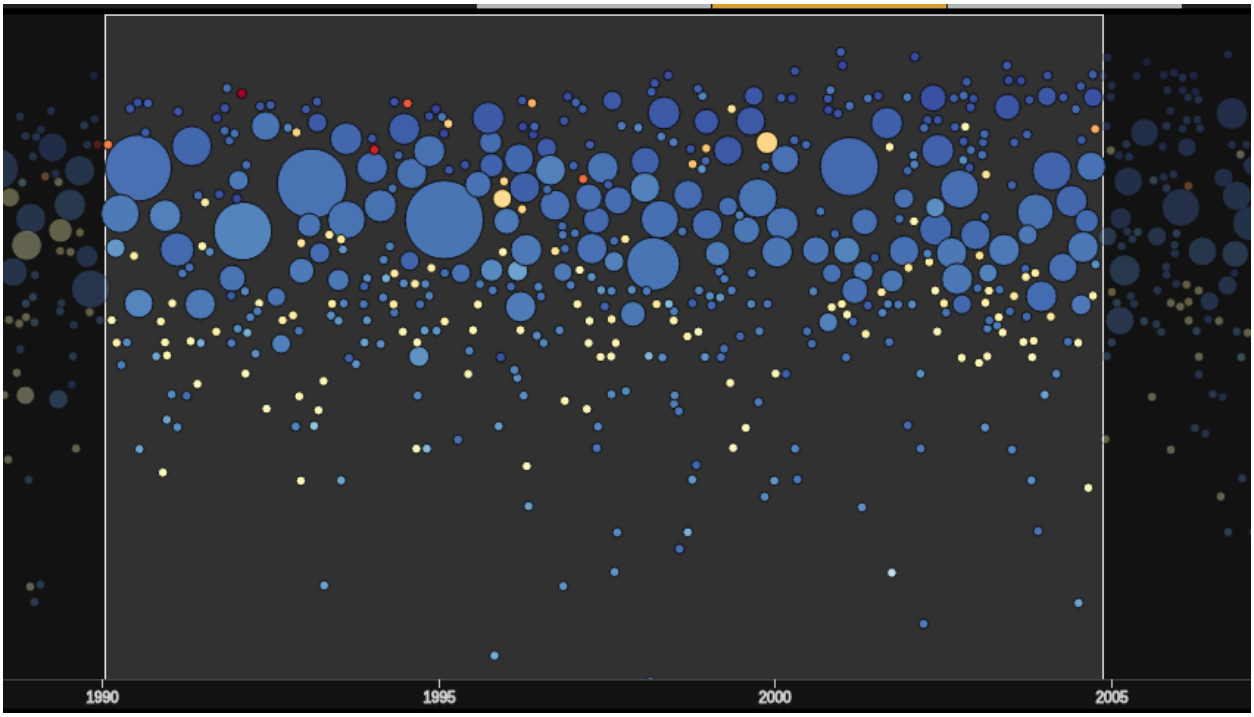
Scatterplot

The intent of the scatterplot is to visualize the distribution of films on various axes such as budget, profit, various ratings, release date, etc. The scatterplot is navigable through the use of zooming and panning. Users can also hover over data points to get detailed information on individual films, which makes it a good tool for discovering films, especially outliers.

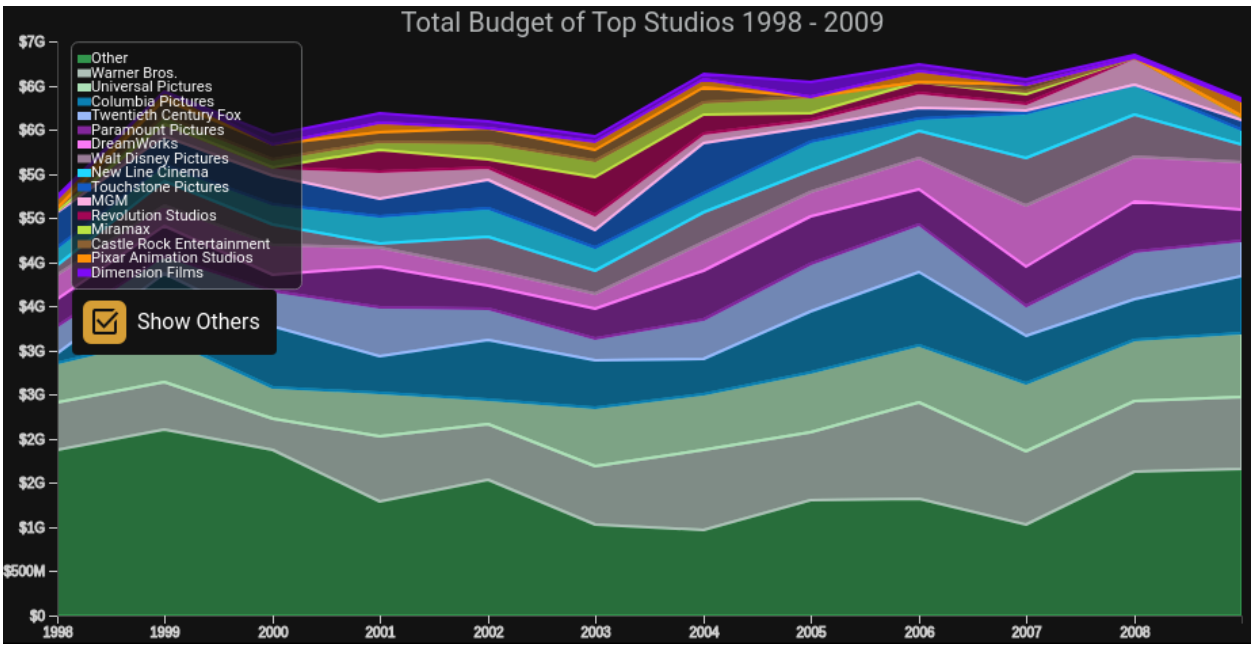
The second iteration of the scatterplot involved collapsing dense regions of data points (circles) into larger circles as part of a semantic-zooming scheme.

Another interaction which we included is brushing, which can be enabled or disabled through a button in the toolbar:





This allows the user to select movies in the scatterplot. In the first two view modes, this interaction provides a way to constrain X axis of the secondary plot, which is reflected in the X axis and the plot title:

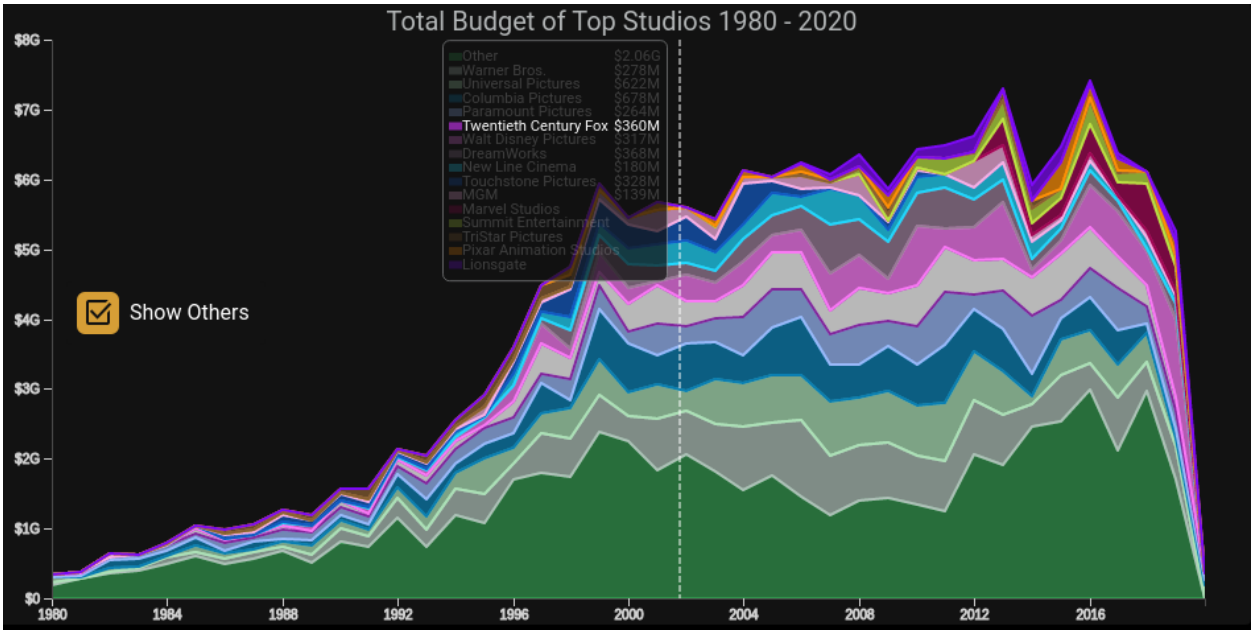


Heatmap

The heatmap is intended to give a more density-focused view on similar axes to the scatterplot. This allows the user to get a more high-level description of the distribution of films on several different metrics.

Production Company Market Share

The Movie Economy view mode's secondary plot shows the market share by various metrics for production companies throughout the years in our data set:

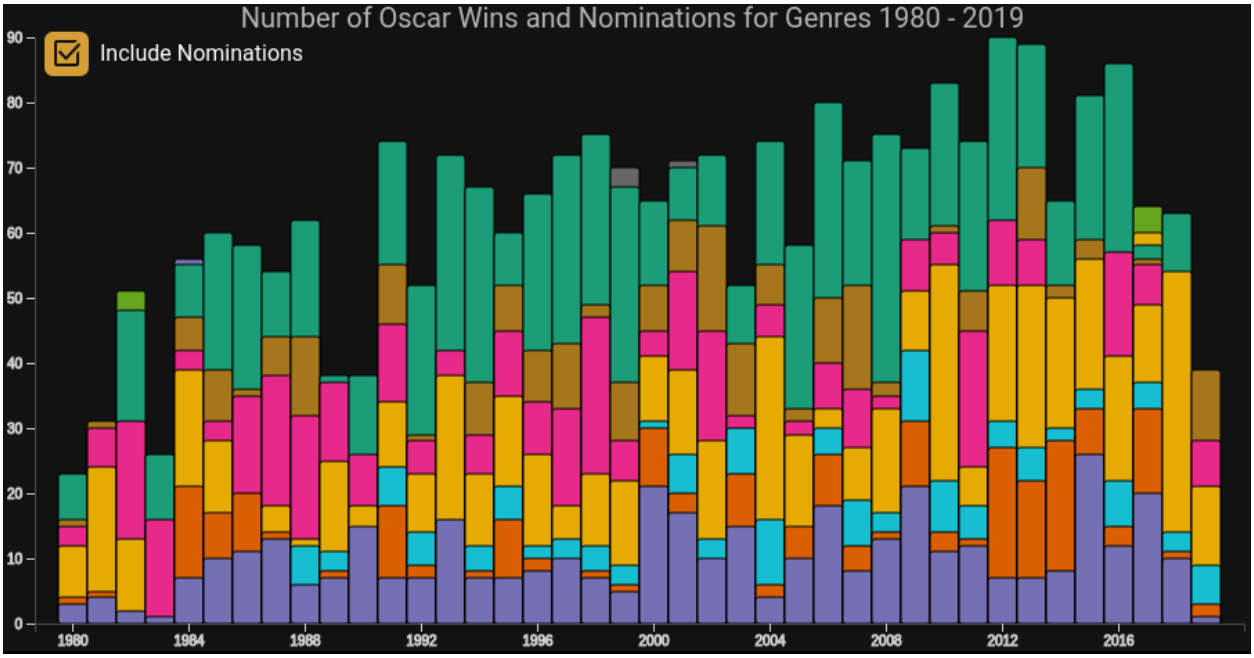


This plot can be interacted with in several ways:

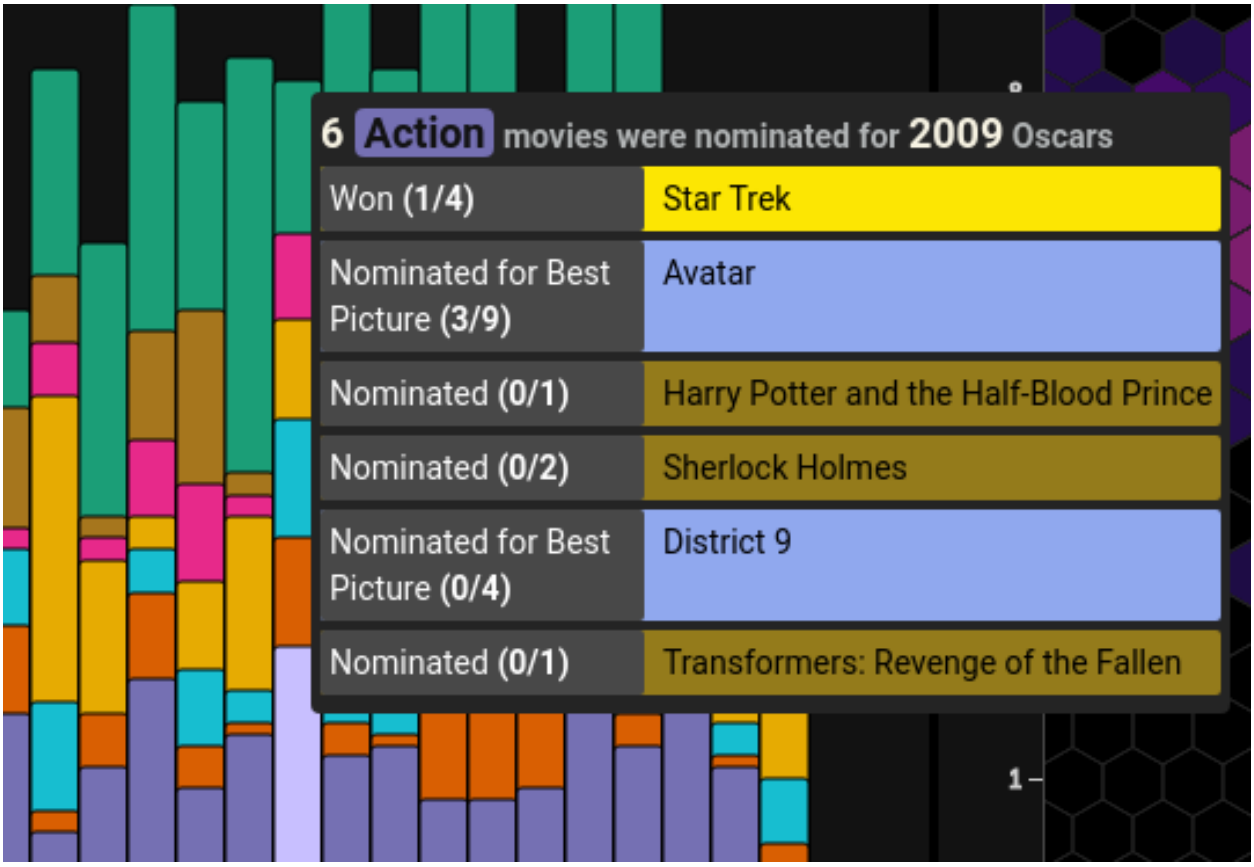
- Clicking the "Show Others" toggle button will toggle the appearance of the combined studios which are lower than the top 15
- Moving the mouse through the visualization will display the companies at that position in order of market share
- Hovering over a particular production company in the chart will highlight it in the tooltip
- Brushing over the scatterplot will change the range of the X axis (year)

Oscars Genre Distribution

The distribution of Oscar nominations and wins by genre is the first secondary plot we created:



In this plot, the inclusion of nominations without wins can be enabled or disabled by the user. There is also a tooltip shown when hovering over the bars:



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?*

We were able to answer our initial questions and more through the implementation of many different highly interactive views on the large dataset.

Our visualization provides a wide variety of tools for learning about the film industry and discovering interesting outliers on many different axes. However, further improvements could certainly be made in several areas:

- Our dataset was relatively small: only about 4000 films. Using a larger dataset would allow us to draw more confident conclusions, but would certainly cause performance issues with our current implementation. If we chose to use a larger one, we would have to optimize significant parts of our code.
- There are several visual issues that could be improved. Mainly, circles in the scatterplot often overlap due to the way in which we generate and display the data. To improve this we would implement dynamic physics to the data to accommodate different screen sizes.
- Tooltips for individuals films would benefit greatly from additional information such as posters and descriptions. This would give the tool a much more viable film discovery usage.
- Plots could have a more shared interactivity. For example, hovering on the heatmap could highlight films contained within a cell on the scatterplot.

Scatterplot

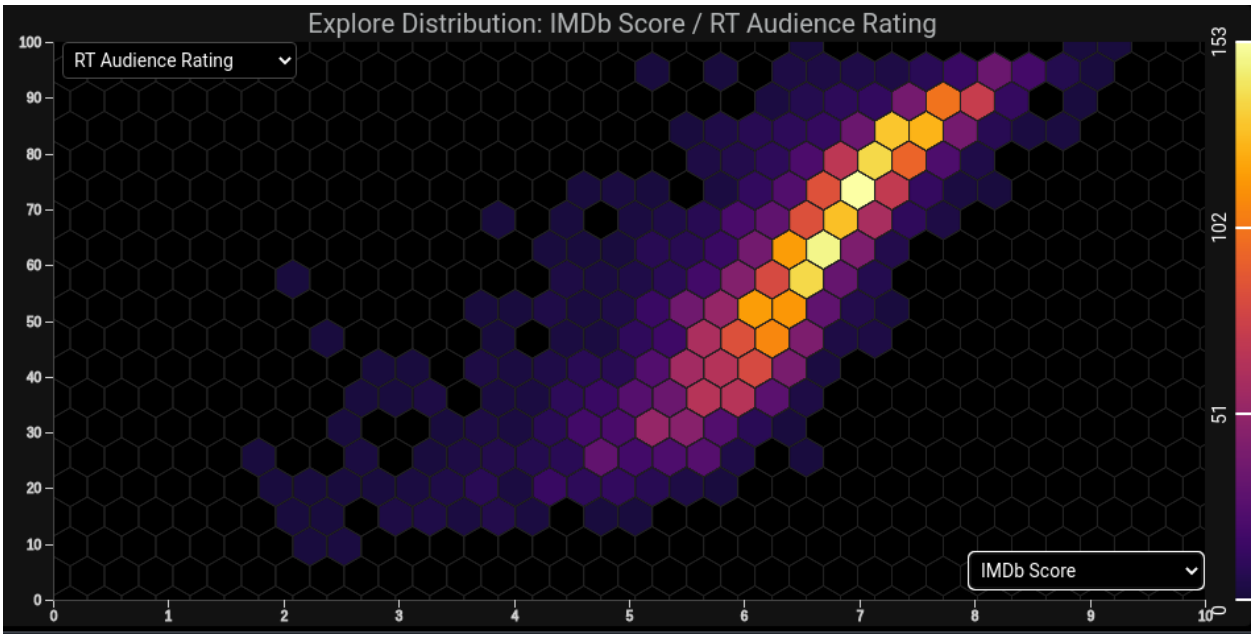
We learned a huge amount about our data through the scatterplot view, including but not limited to:

- Our data set has only approximate figures for budget and profit, rounding off many digits of accuracy
- The Rotten Tomatoes "Tomatometer" metric is significantly more well-distributed than the other ratings metrics, IMDb score and Rotten Tomatoes audience score.
- There are some very poorly rated films nominated for the Oscars.

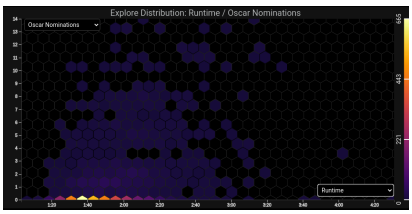
Heatmap

Through the heatmap we were able to learn the relationships between various metrics, such as:

- All rating scales have a more-or-less linear relationship with each other:



- Number of oscar nominations have a strong relationship with runtime:



Oscar Genres

The Oscar genres stacked bar chart was very informative with respect to how the popularity of various genres within the Oscars has changed over the years. For example, no fantasy film has earned an Oscar since 1999 (Sleepy Hollow).