"Movie Madness" Process Book

CS 5630
Kaden Hendrickson, u1200510@utah.edu, u1200510
Ethan Ramos, u1182994@utah.edu, u1182994

Project Repository:
https://github.com/ethramos22/dataviscourse-pr-moviemadness

Table of Contents

Background and Motivation

Growing up we loved watching movies, and have also struggled trying to find what movie to choose. There are so many options in today's world with streaming platforms as well as cable and weekly releases at the theater. We want to stay in the know of pop culture by focusing on what is popular, relevant, and rated well to help us and others decide what to watch.

Project Objectives

Benefits:
- Help users determine what show or movie to watch, by filtering data by genre, popularity, score, platforms, and more

Accomplish:
- See trends amongst popular and highly rated movies based on budget, genre, revenue, and more
- Point users to streaming platforms and places that currently offer a selected movie
- Show movies profits and losses based on budget and revenue
- Show which countries had produced a selected a movie

Answer:
- What movie is the most popular and has the highest reviews and vice versa?
- What movie is the most popular with the lowest reviews and vice versa?

Data

We'll be collecting our data from 'The Movie Database' aka TMDB (https://developers.themoviedb.org/3/getting-started). This movie database is free to use, and contains massive amounts of data related to movies, tv shows, streaming platforms, actors, geographical location, etc. We'll utilize multiple endpoints, the most relevant being the following:
- /movie/{movie_id}
- /movie/{movie_id}/images
- /movie/{movie_id}/similar
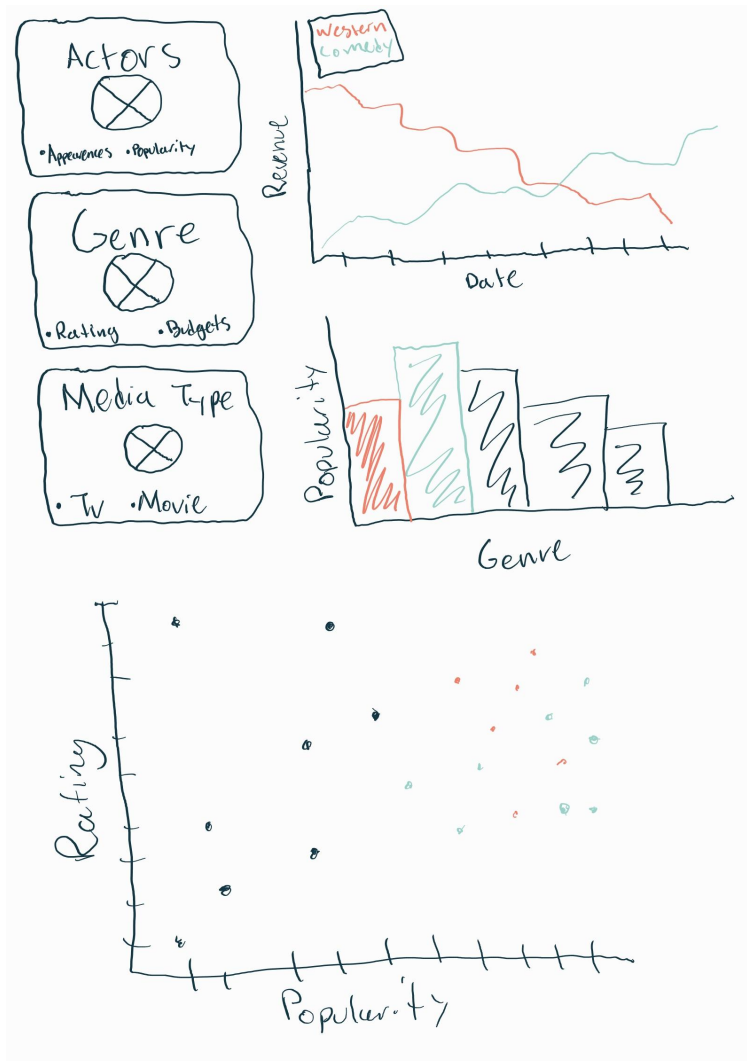- /movie/popular
- /movie/top_rated

Using TMDB will require minimum data cleanup with its clean and easy to use API. Our main derivation from the TMDB API will be for a movie Id. One of our main API calls will be a movie title. From the result we can take the "id" field to make additional queries on the details of that movie. This will lead us to the details API call for a movie. From this response we can retrieve things like genre, budget, revenue, poster url for an image of the poster, popularity rating, and runtime.

Another main query for our project will be "Get Popular" to retrieve an array of the most popular movies. This query can return an array of the movies most popular by view count, and is updated on a daily basis. Similarly we will use "Get Top Rated" and "Get Upcoming" to populate our movie list with relevant data.

For our data processing we can use the returned json objects to bind to d3 elements and javascript objects. Packaging the movie into its own object for later use on multiple visualizations.

Visualization Design (Kaden & Ethan)

- Interactive feature (in the shape of a card) on the left (choose genre, actor, industry, media type, etc.) on the right we get interactive visualizations related to the feature you chose (profit, budget, rating, popular actors)

Movies

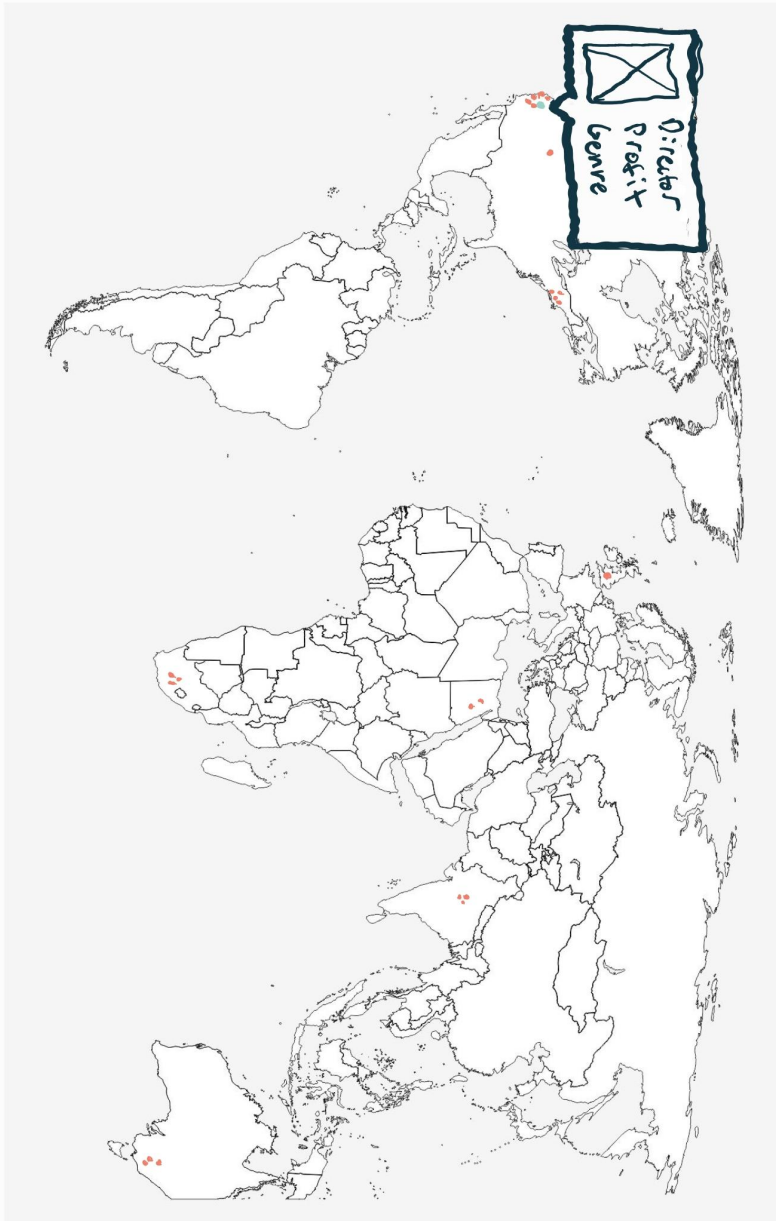|	
☒	Title . Profit . Genre
☒	Title . Profit . Genre
☒	Title . Profit . Genre
☒	Title . Profit . Genre
☒	Title . Profit . Genre
☒	Title . Profit . Genre

x-axis ▼   y-axis ▼

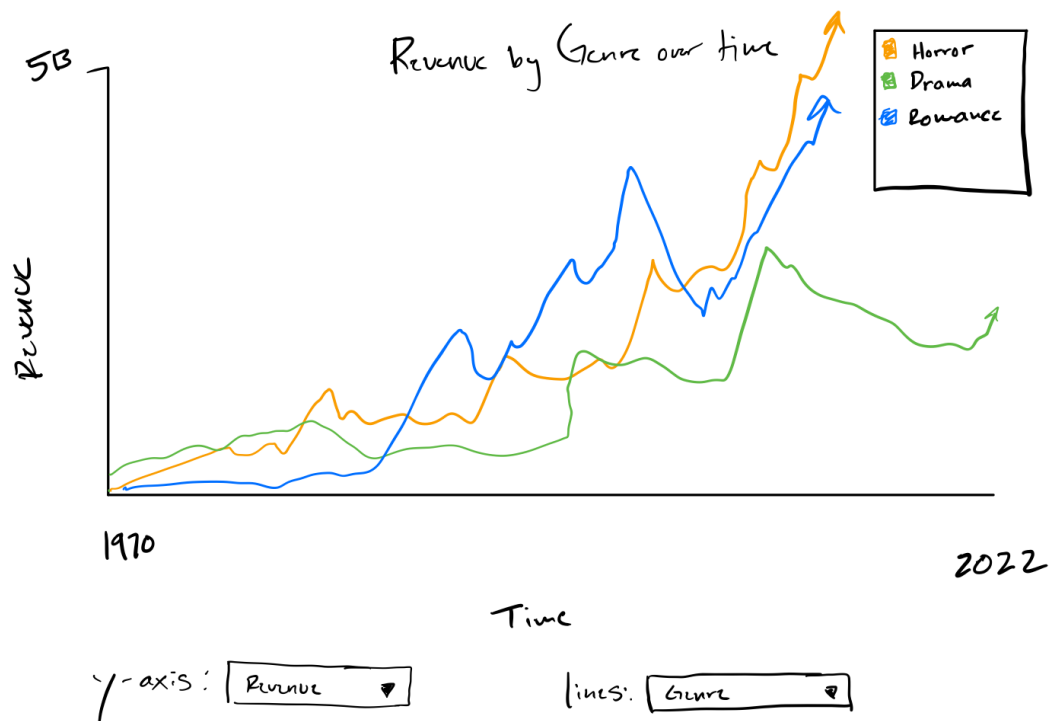Rating

Comedy
Drama
western

Popularity
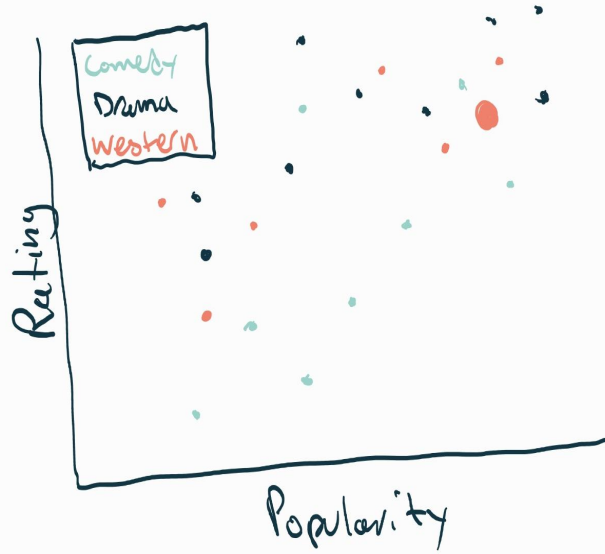
- Map with pins on production location of popular movies. When a location gets clicked, a view appears with more detailed information about the movie/movies that were produced in that country.


Movies by production location

Director
Profit
Genre

- Line chart that maps quantitative movie data over time for different genres (ie. change in budget/revenue/production time/ etc. over time for drama movies). This visualization could allow changing in the y axis or in the data show by manipulating a dropdown menu

Revenue by Genre over time

Legend:
- Horror
- Drama
- Romance

y-axis: [Revenue ▼]   lines: [Genre ▼]

- **Final Design:**
  - Will contain a list of movies that can be filtered by popularity, top rated, upcoming, etc. The list will contain visualizations about the movie info (rating, genre, language, adult content, etc.). A panel in the middle of the screen will display the movie, and all the details we can find about it. The panel on the right of the screen will display several visualizations, and will highlight the currently selected movie in each of the visualizations - to show it compares to the other movies in the list. Potentially, we can create another visualization on the bottom of the screen that will show geographical or timeline data. Maybe even actors being connected to one another.
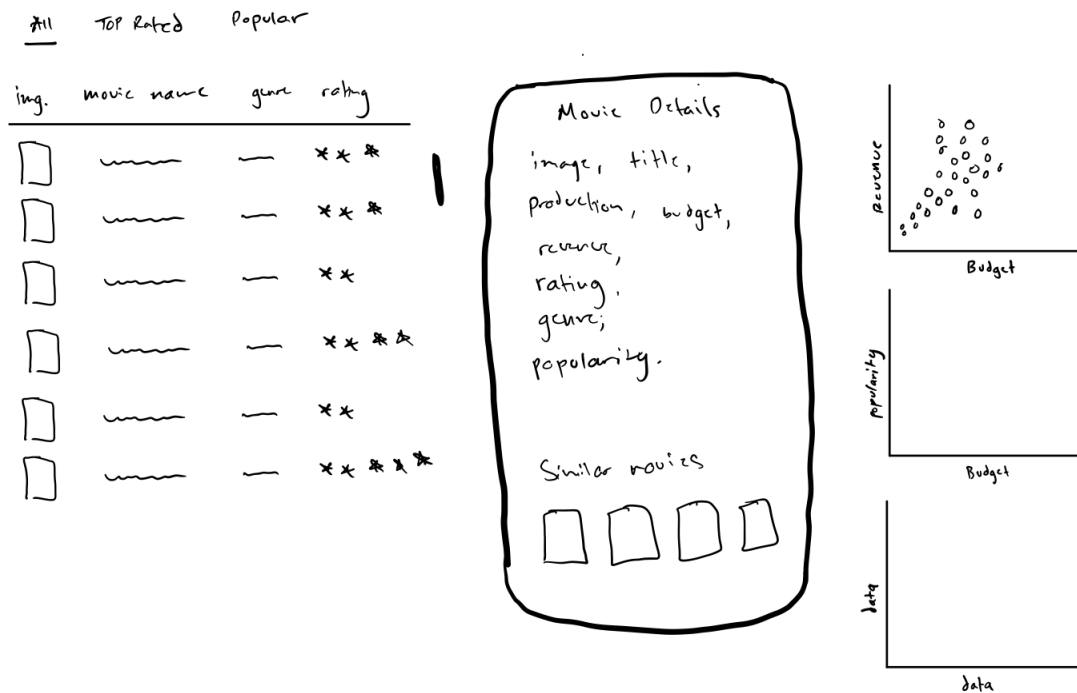
Actor Connectedness
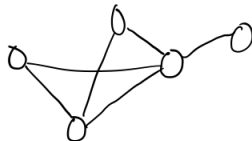
Exploratory Data Analysis

Our initial exploration of the data solely involved parsing through the json to compare attributes and see what kinds of visualizations we could create. Once we had a general idea of the data we were going to use, we used scatterplots to plot different attributes against each other. After testing out plotting multiple combinations of attributes, we were able to find which combinations were most insightful to the user (Budgets vs. Revenue, Budget vs. Rating, and Revenue vs. Rating)

Design Evolution

Distribution Chart Addition
**[entry by Ethan, 11/7]**
Based on feedback from Haihan Lin, an overall distribution chart will be necessary for our project. Not only will this add more visualization, but also show overview data of all the selected movies. Firstly we can show how many movies have a specific genre. The current dataset based on most popular movies, currently playing movies, and top rated movies. Data was collected the week of halloween and we can see that horror is the top genre.
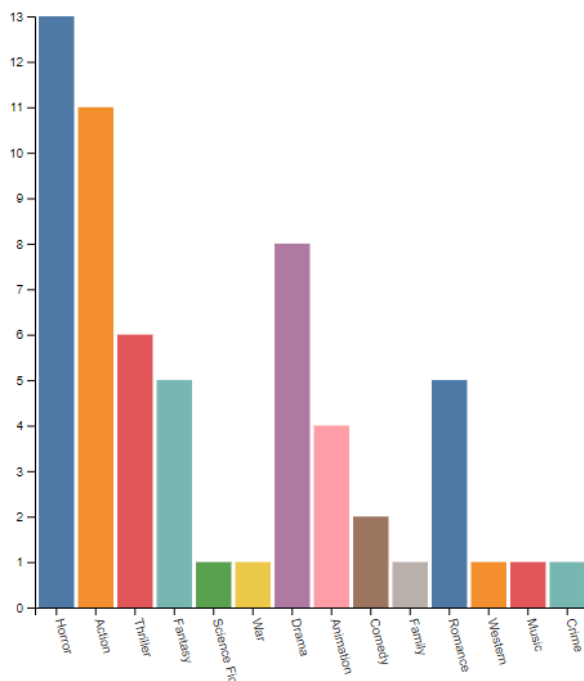
Movie Genre Color Scheme
[entry by Ethan, 11/9/22]
Using the d3 set color scheme we would get duplicate colors for different genres, due to the limitations of using a preset scheme as tableau10 would only have colors for 10 values. We knew our data needed an ordinal color scale with a value for each genre. Color coding genre's will help with our other visualizations on our web page, for example to color code the dots on our dot plots.
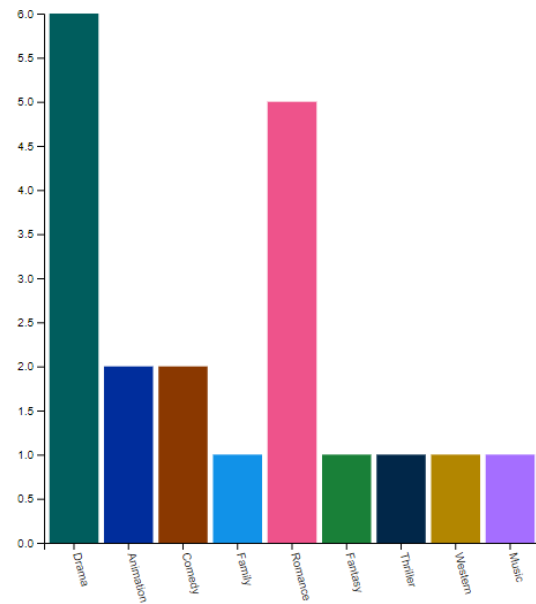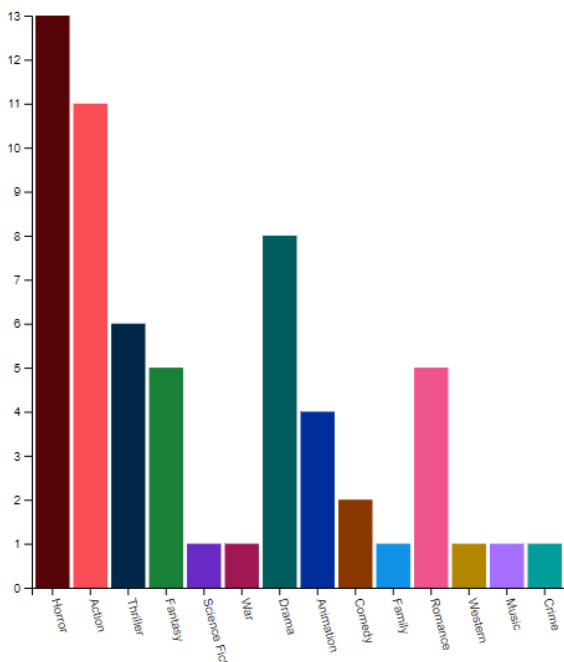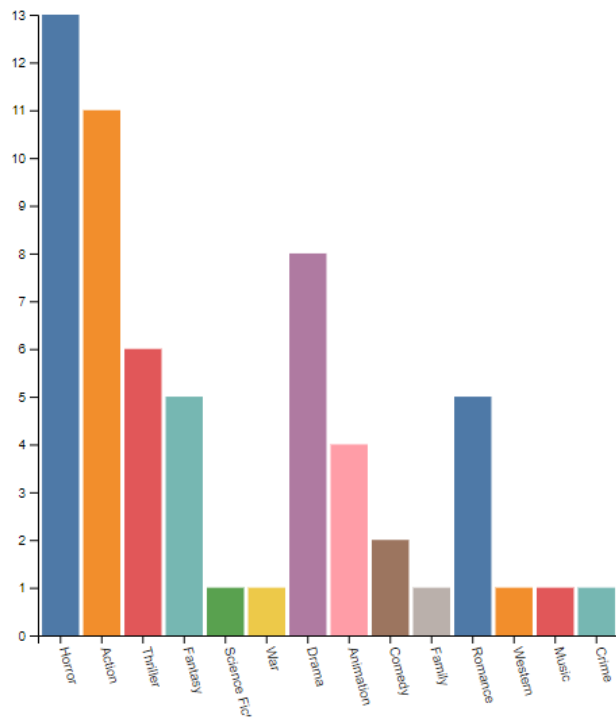
Using color palettes, we could get an x amount of color values for an ordinal scale. From our data we had 13 genres. We created our range of these 13 color values to map 1-to-1 to the keys of the data grouped by genre. When the data would change from selection and interaction, genres will

always have the same color. As we can see above, changing the list of movies from popular to top rated, the Drama genre becomes the top genre.

Table item selection
[Kaden, 11/9/22]

Clicking on an item in the tables displays more information about the selected movie in the detail card to the right of the screen. This allows us to show more information than what the table can hold, including the movie poster, movie description, and eventually a list of 'similar movies'.

Movie collection selection
[entry by Kaden, 11/9/22]

Users are able to toggle between different collections of movies, including 'All', 'Now Playing', 'Popular', and 'Top Rated'. Toggling between these selections also changes the distribution chart and the smaller dot-plots at the bottom of the screen, to display data relevant to the selected movie collection.

Movie Name	Genre	Rating	Language	Revenue
				\$0 \$2,000,000,000
The Godfather	Drama, Crime	87%	English	\$245,066,411
The Shawshank Redemption	Drama, Crime	87%	English	\$28,341,469
The Legend of Hei	Animation, Fantasy, Action	87%	Chinese	\$48,045,728
The Godfather Part II	Drama, Crime	86%	English	\$102,600,000
Schindler's List	Drama, History, War	86%	English	\$321,365,567
Dilwale Dulhania Le Jayenge	Comedy, Drama, Romance	86%	Hindi	\$100,000,000
Impossible Things	Family, Drama	86%	Spanish	\$0

Displaying revenue in the overview table
[entry by Kaden, 11/9/22]

Initially, we thought that revenue would be a good metric to list in the table, to give viewers an 'at-a-glance' understanding of how much money the movie generated. We quickly found out however that the range of values was \$0 to \$3 billion, which is a hard range to visualize inside an 80 pixel svg. We decided it would make more sense to display the movie runtime - as it is a more digestible metric, and is possibly a more valuable value for the user to get 'at-a-glance.'

Movie Name	Genre	Rating	Language	Revenue
				\$0 \$2,000,000,000
Terrifier 2	Horror, Thriller	71%	English	\$5,325,078

Limiting Bar Chart Color
[entry by Ethan, 11/23/22]

Previously we thought encoding each genre with its own set color would be a good encoding for the bars and other visualizations like the dot plot to show genres across all the visualizations. This became too distracting from the data and a very weak encoding that didn't mean enough to justify

the color. We have reverted to a single color of light gray with a light red to change on hover to show which bar you are hovering on. This will work well with our tooltip.
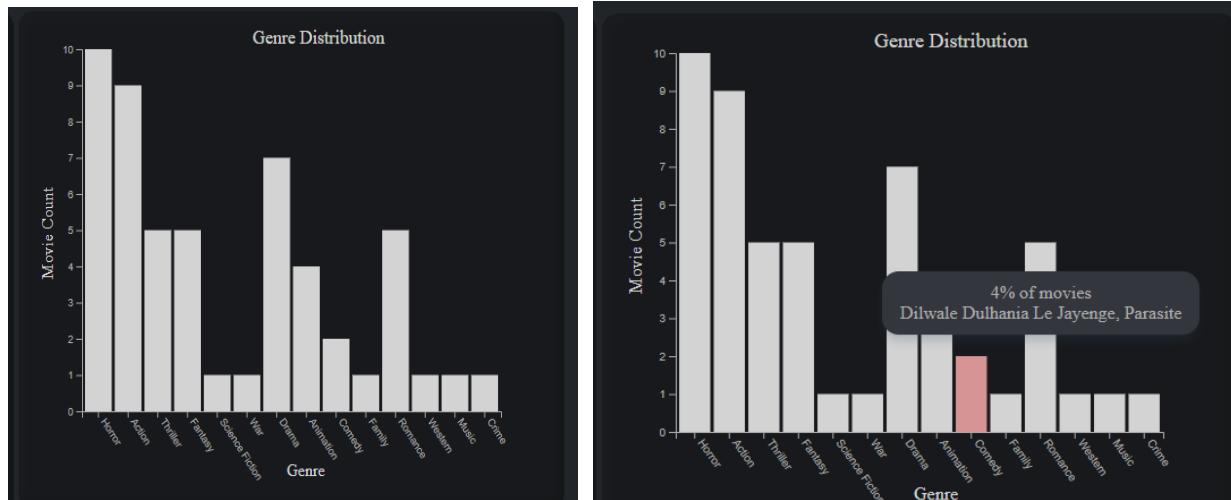
Movie Poster Card
[entry by Ethan, 11/25/22]
In our final design sketch we have the image of the movie poster represented in the table, but moving it over to the middle section with the movie section made more sense, as we can make the image bigger and draw more attention to a movie selected from the table.
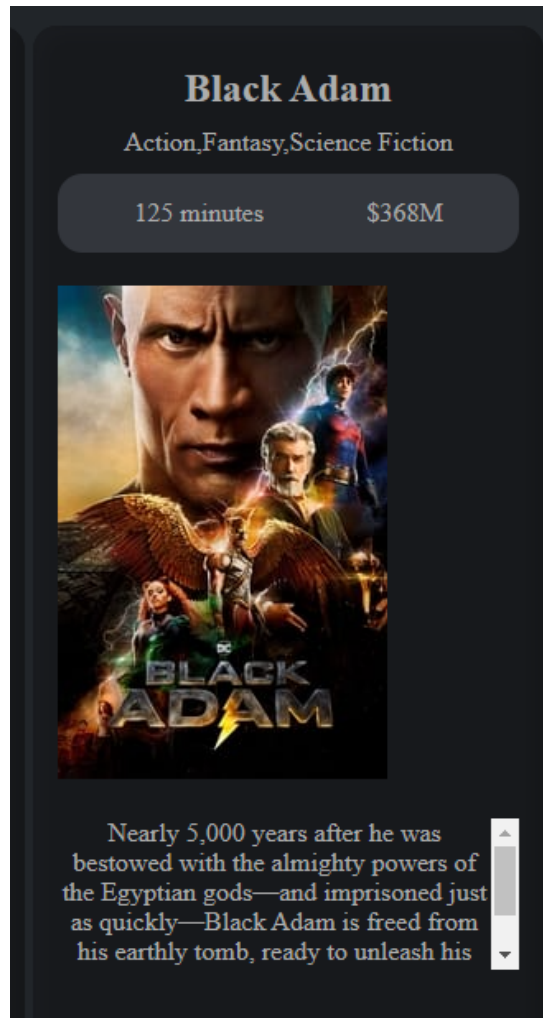
Removing Related Movies
[entry by Ethan, 11/28/22]

As seen in the photo above, we aren't including a list of related movies on the movie card as shown in the design sketch. We will move this feature to a stretch goal that will be nice to have, but unsure if we can add it to our data and visualization in time.
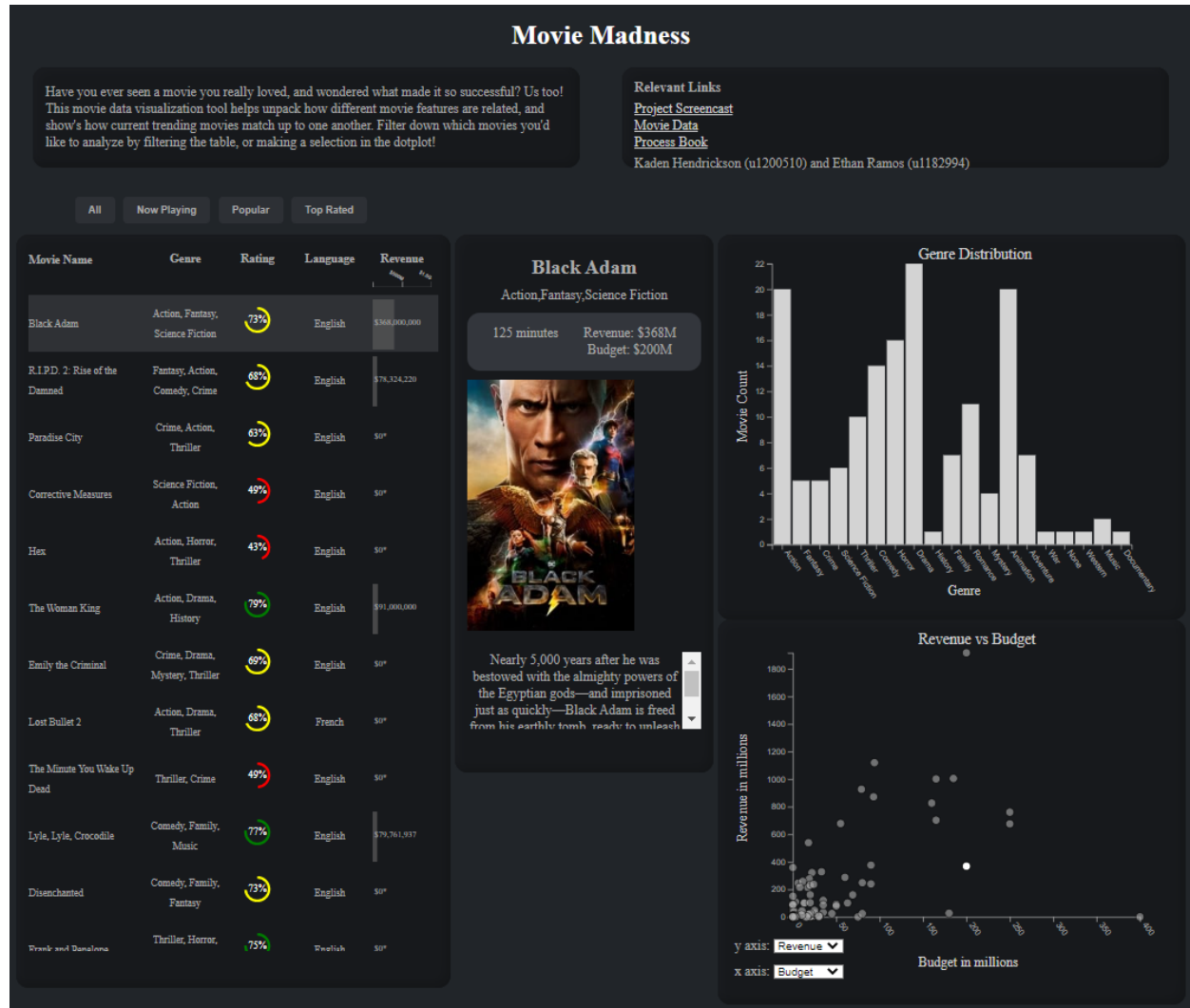

Three Plots to One
[entry by Kaden, 11/25/22]

In the original design we included three separate plots to the right of our movie detail card. This made our page very crammed and hard to read data points on the three small charts. We have changed to one dot plot where a user can change the y and x axis to see different distributions. We can also highlight the current selected/shown movie on the plot easier with interaction.

Revenue vs Budget

y axis: Revenue
x axis: Budget

Budget in millions

Revenue in millions
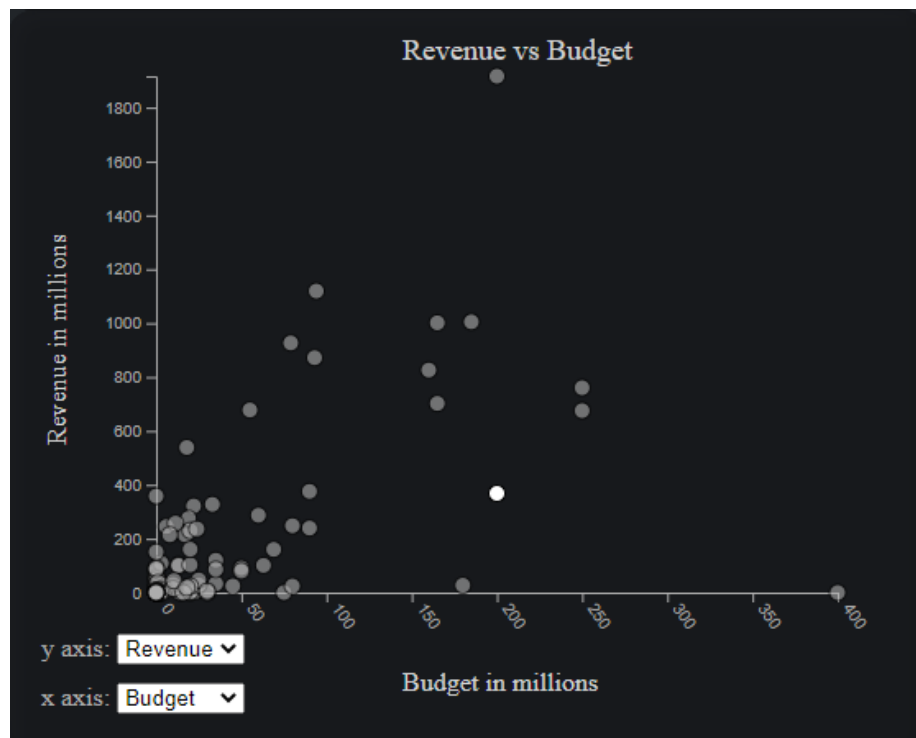
Implementation

Overview

Here is our whole site. We stayed very close to our original design while adding new visualizations and focusing on interaction. We can see our project layed out into four main areas; movie list table, movie detail card, genre distribution card, and dotplot.

Movie Table

| Movie Name	Genre	Rating	Language	Revenue
				$500M $1.0G
Black Adam	Action, Fantasy, Science Fiction	73%	English	$368,000,000
R.I.P.D. 2: Rise of the Damned	Fantasy, Action, Comedy, Crime	68%	English	$78,324,220
Paradise City	Crime, Action, Thriller	63%	English	$0*
The Woman King	Action, Drama, History	79%	English	$91,000,000
Emily the Criminal	Crime, Drama, Mystery, Thriller	69%	English	$0*
The Minute You Wake Up Dead	Thriller, Crime	49%	English	$0*
Lyle, Lyle, Crocodile	Comedy, Family, Music	77%	English	$79,761,937

The focal point of our project is the movie table. This is the main entry to our interactions and visualizations for our project. We have the movie's title, genre, rating, language, and revenue of each movie. As shown the color on the row will change as the mouse hovers over the different rows in the table, and a little bit brighter color will show on the row of the movie that is currently selected, and shown on the movie detail card.

The Quintessential Quintuplets Movie	Animation, Comedy, Romance	90%	Japanese	$0*
Teasing Master Takagi-san: The Movie	Animation, Comedy, Romance	88%	Japanese	$0*
The Godfather	Drama, Crime	87%	English	$245,066,411
20th Century Girl	Romance, Drama	87%	Korean	$0*
The Shawshank Redemption	Drama, Crime	87%	English	$28,341,469

Hex	Action, Horror, Thriller	43%	English	$0*
The Minute You Wake Up Dead	Thriller, Crime	49%	English	$0*
Corrective Measures	Science Fiction, Action	49%	English	$0*
Nice Sister-In-Law 2	Romance	50%	Korean	$0*
The Friendship Game	Science Fiction, Horror	52%	English	$0*

Rating is our first visualization. It's a small radial bar chart, where the percentage rating 0-100% is encoded as the bar. We've also encoded the percentage to a color for certain thresholds. 0-59% is a movie with not good reviews encoded as red, 60-74% is a movie with ok reviews, and 75-100% is a movie with good reviews. This allows users to very quickly see which movie fits into those three categories. For example, when looking at top rated movies, we can see they all have good reviews as implied by the selection. Additionally we can sort the column by rating, seeing the best and worst rated movies, just by clicking on the "Rating" table header.

Movie Name	Genre	Rating	Language	Revenue
Spider-Man: No Way Home	Action, Adventure, Science Fiction	80%	English	$1,916,050,941*
The Lord of the Rings: The Return of the King	Adventure, Fantasy, Action	85%	English	$1,118,888,979*
The Dark Knight	Drama, Action, Crime, Thriller	85%	English	$1,004,558,444*
Jurassic World Dominion	Adventure, Action, Science Fiction	70%	English	$1,001,000,000*
The Lord of the Rings: The Two Towers	Adventure, Fantasy, Action	84%	English	$926,287,400

Revenue is our second visualization, encoding the revenue of the movie as a barchart. Users can compare revenues amongst the list, but even better with our dotplot, but we will get there later. A user can click on the Revenue header to sort the rows by revenue, showing the movies that made the most at the box office. Within our data lots of smaller indie films, and films on streaming platforms don't have revenue data, which is denoted as "$0*".
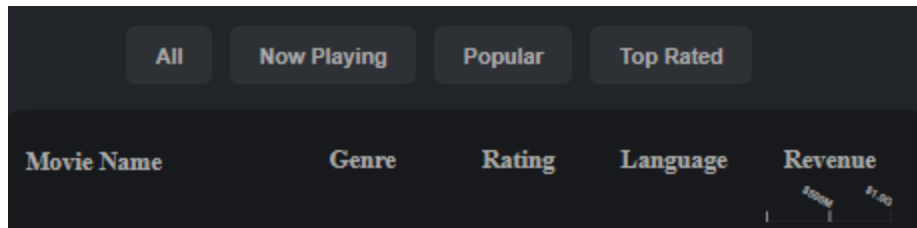
Lastly we have our main filter buttons. These buttons will filter the movie table data as well as the data in the barchart, and dotplot. This allows users to quickly browse between these 4 dominant categories of movies.

Movie Detail Card

A centerpiece for our data, that represents a visual representation of each selected movie. The main attraction for our movie detail card is the movie poster. A user can click on a movie from the movie table and the poster and other information will appear on this card. Users can quickly see genre, runtime, revenue, budget, and a text description of the movie.
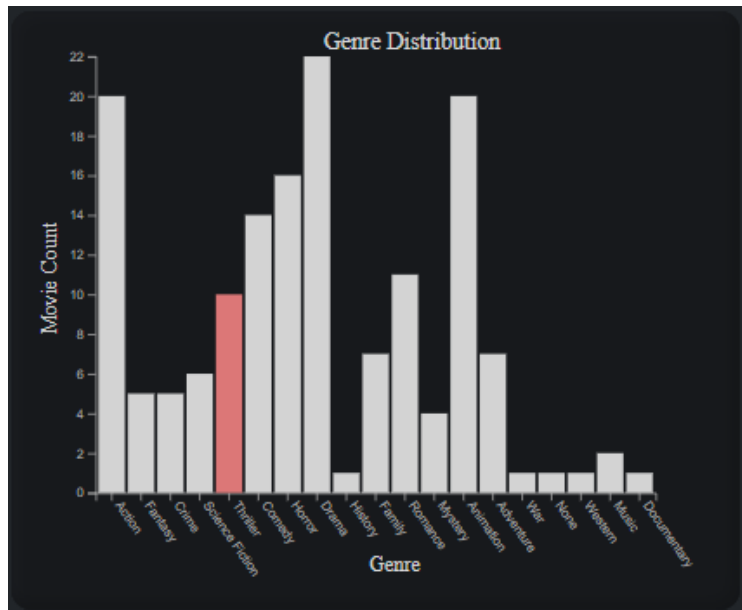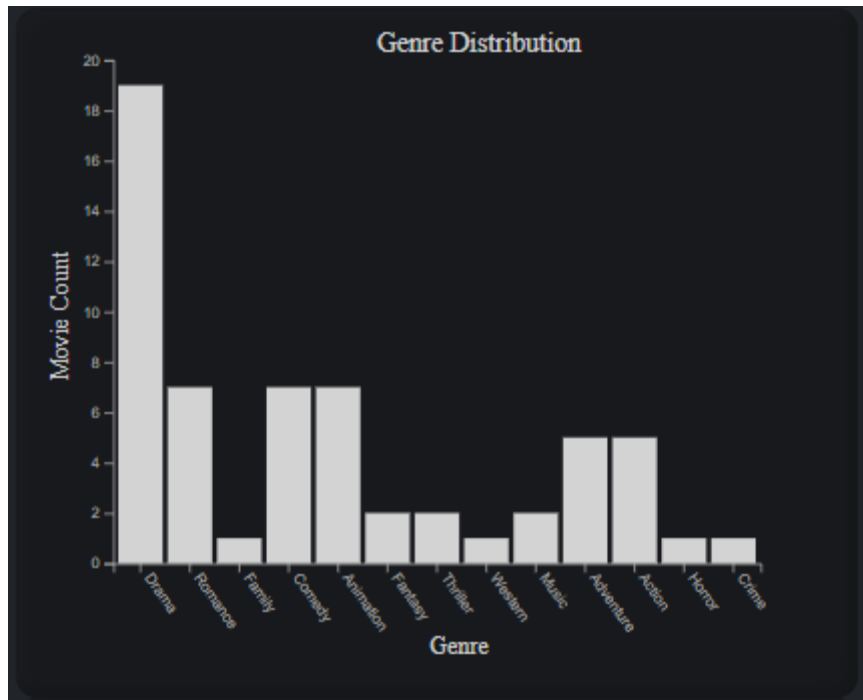
Genre Distribution Chart
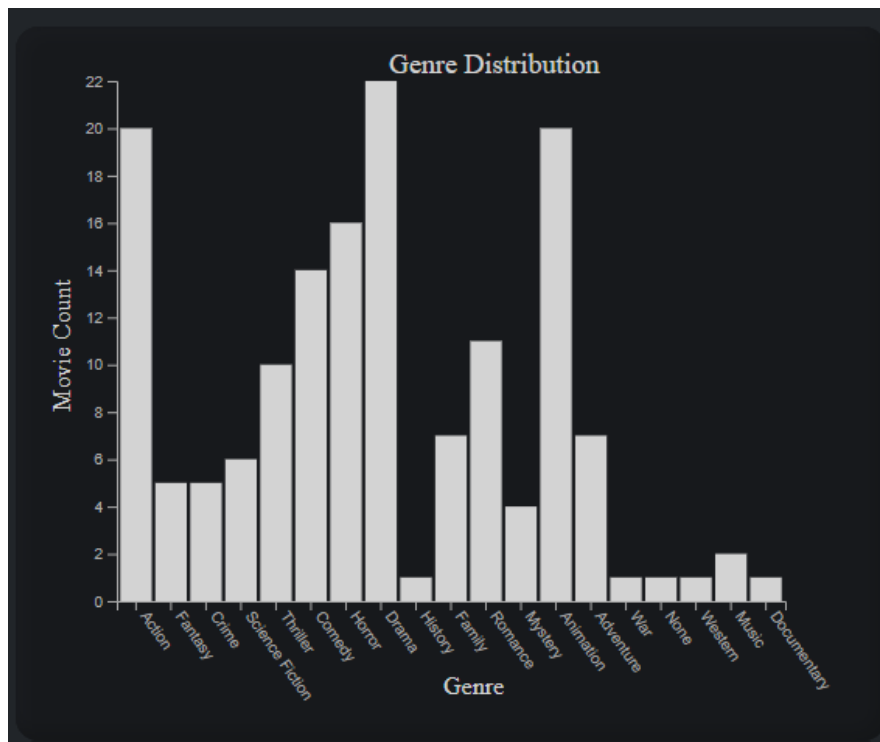
Here we can see a general distribution of genres of movies. Users can determine what is the most popular genre, what genre is more prominent among top rated movies, and what genre is most prominent among the most popular movies. For example when filtering for top rated movies, we can see that Drama has the highest count.
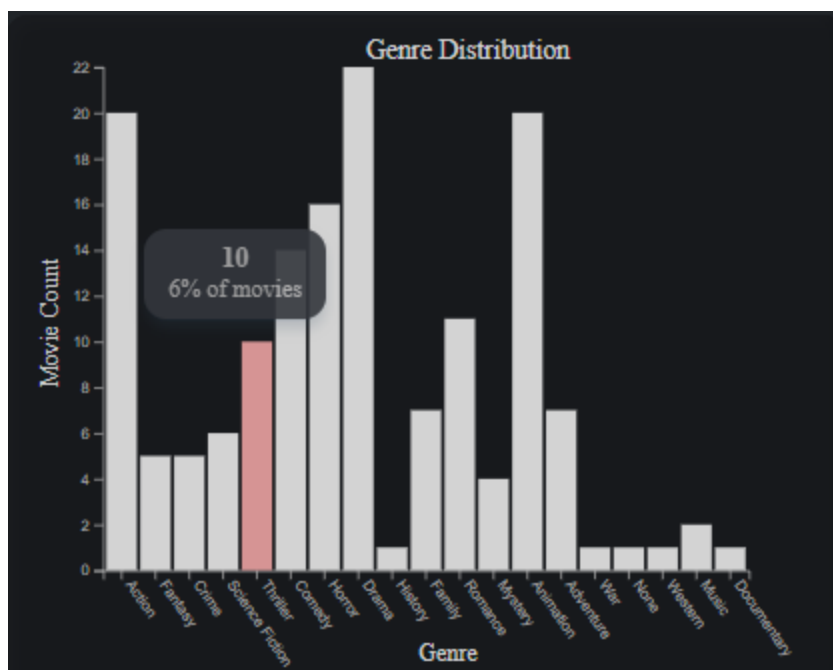
Our barchart has a tooltip for hover interactions. Users can quickly see how many movies fall in that genre category and what percentage of the current display of movies is that genre. We used a light red color to draw attention to the barchart a user is hovering on, so there's no confusion of what bar the tooltip is for. The tooltip is also positioned right above the barchart so users don't have to look around for more information.


The barchart also features selection interaction. A user can select one of the genres, and the movie list and dot plot will update. This allows users to see best rated by genre, most revenue by genre, distribution of genre by budget and revenue, and more. We use a slightly brighter red for the bar selection so users are aware only movies in that genre are being shown in the movie list and dot plot. A user can click the barchart again to remove the selection.

The movie list filter buttons will also update the barchart. When filtering by top rated movies, we can see drama is the winner, and there are lots of genres absent from the distribution with no top rated movies.

Dot plot
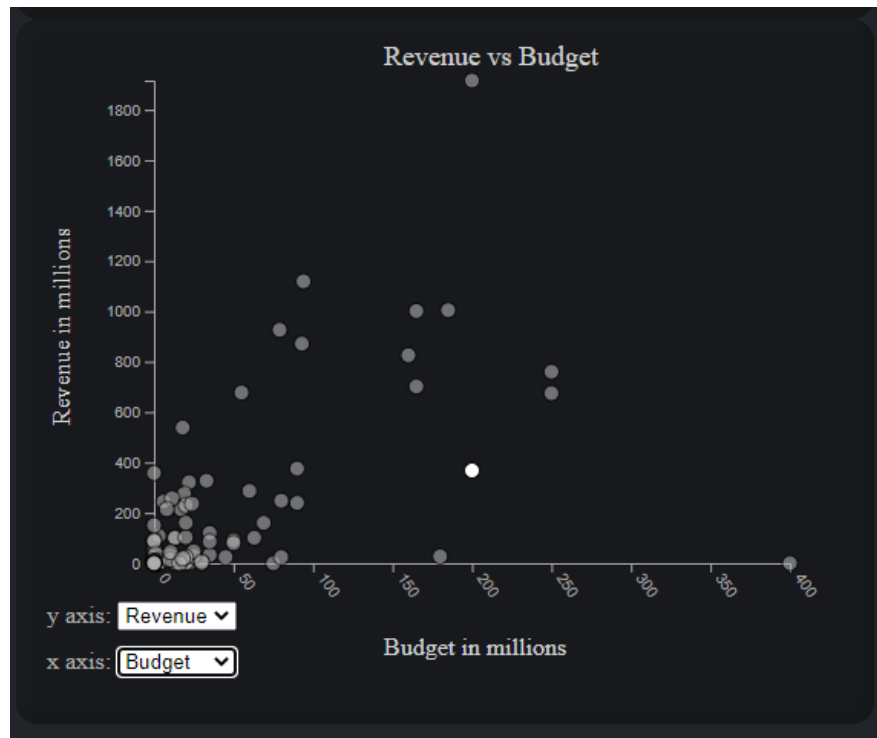
Our dot plot is our most versatile visualization. It features selection, tooltips, brushing, and variable axes. A user can see lots of different distributions here. We have set the default as budget vs revenue which we think is most interesting to see which movies made less or more than their budget. The current selected movie which is displayed on the movie detail card is shown as a brighter white dot here. Opacity was a strong encoding for our dark site. Selecting another dot will also update our movie detail card to the new selected movies.
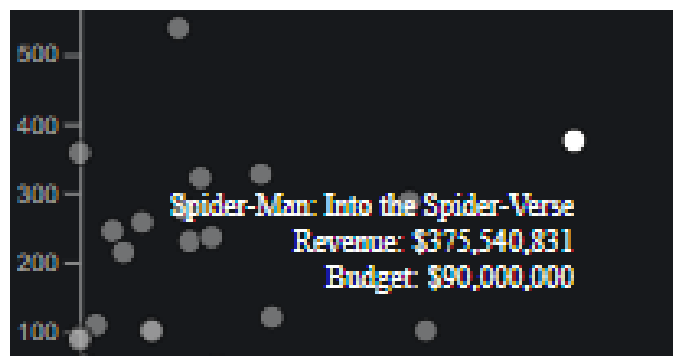

Here we can see our tooltip. The tooltip appears on hover and provides the movie title and exact numbers of the movie based on the axis selection. Since our current axis is Budget vs

Revenue, those are the values we see in the tooltip. These update as the different axes are changed.
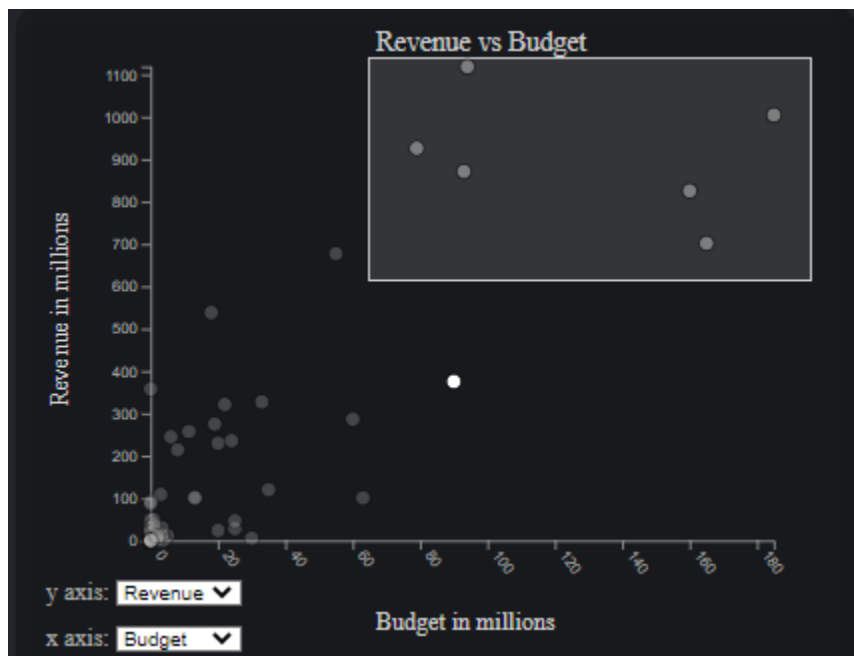
With brushing a user can select certain parts of the data. We can brush here for movies that had big budgets and made the most revenue. This brush will update the movie list to only show the movies in the brush layer. A user can see the top rated movie in the selection.
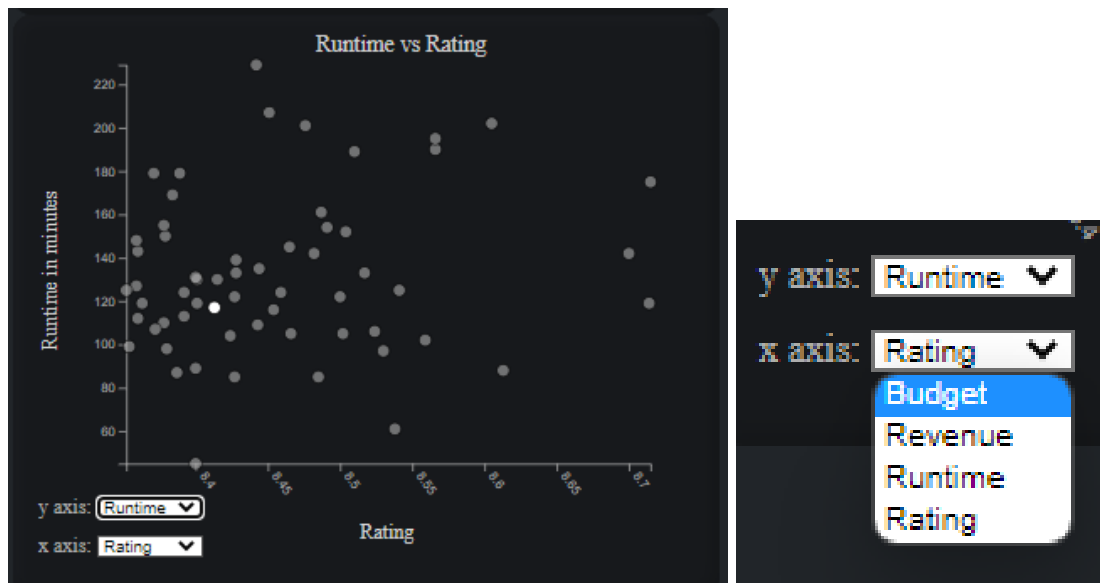

Variable axes give the user more control. They can pick from these four options what distributions they want to see. More freedom, more power to the user. Having this update one

dot plot with smooth transitions doesn't distract the user from having to look at three plots at once, and instead updating one how they like, with minimal eye movement.

Evaluation

We learned a lot from our data, however it's important to note that our data was collected from October to the end of November which impacts which movies would be popular at that time and which movies would be playing at that time.

During the month of October with our first initial data we gathered, when looking at all the movie data, the thriller and horror genres were most popular. This makes sense with Halloween being in October. When we gathered more data towards the end of November we can now see that drama is the genre with the most movies.
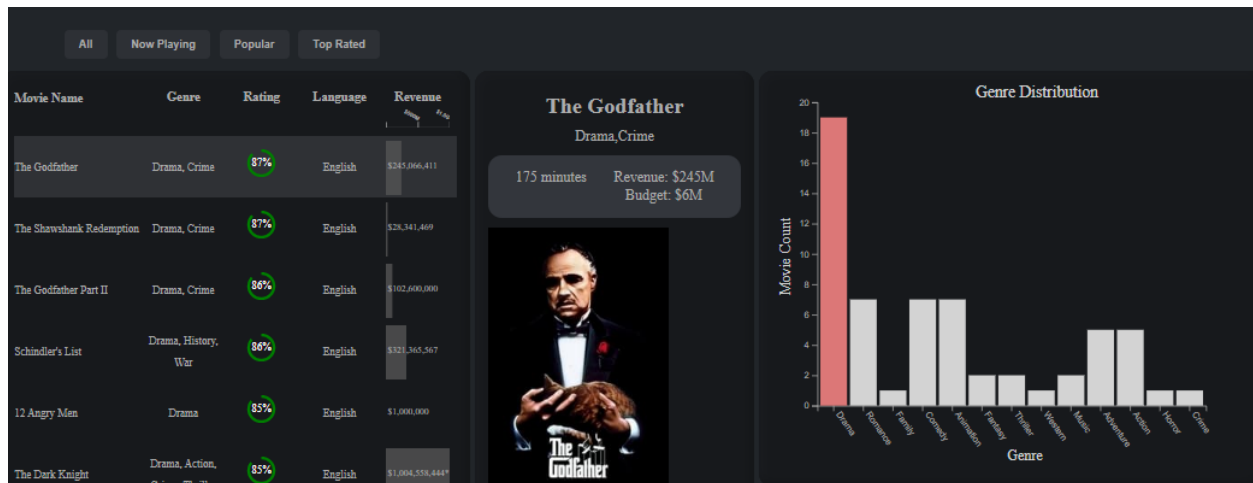
Drama was also the genre that had the most top rated films which included the likes of The Godfather, The Dark Knight, and Fight Club, with The Godfather having the best rating. This was answered by filtering the movie data by top rated, selecting the drama genre bar in the barchart, and then sorting by rating.
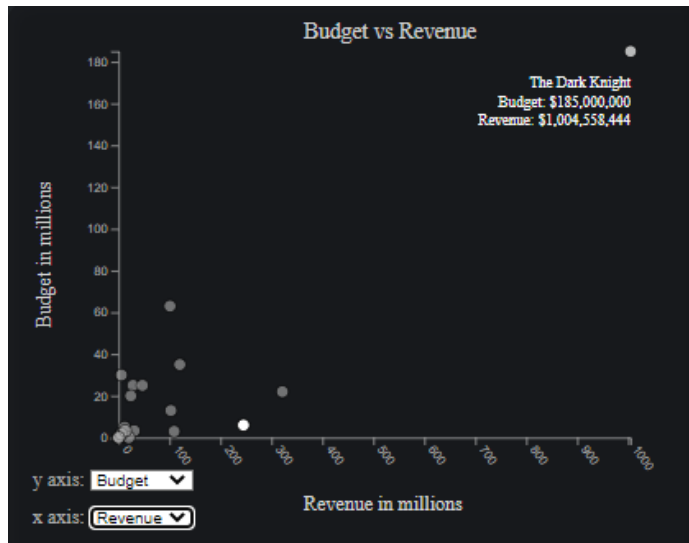
Also within the category of top rated dramas, The Dark Knight was outlier, having a much bigger budget and made a lot more revenue than its peers. We knew it was outlier by filtering by top rated, then selecting the drama bar in the barchart, and then looking at the dotplot comparing budget vs revenue.
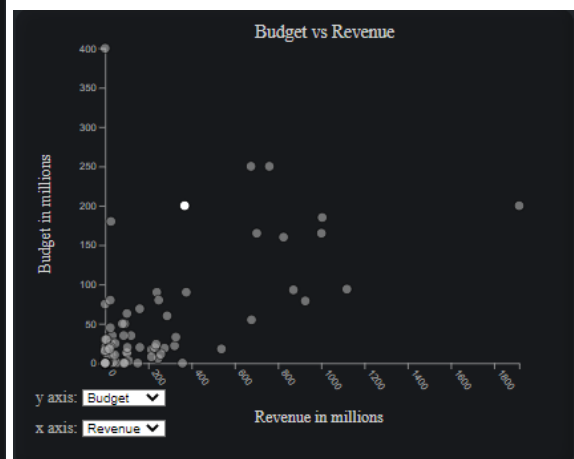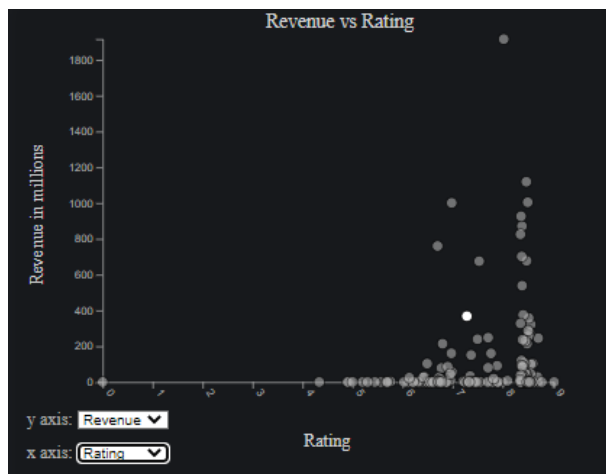
We learned that there's not exactly a sweet spot of total movie length(runtime) that correlates with movie rating. In fact a lot of the distributions didn't have strong correlations between runtime, rating, revenue, and budget. This visualization works best when brushing for outliers and groupings, but didn't work well in seeing general correlations between the data.

This could be improved upon by adding a line of correlation in the dot plot to show dots above or below the line. For example when comparing budget vs revenue, a line in the middle of the plot to show dots above it that made more money than its budget and vice versa.
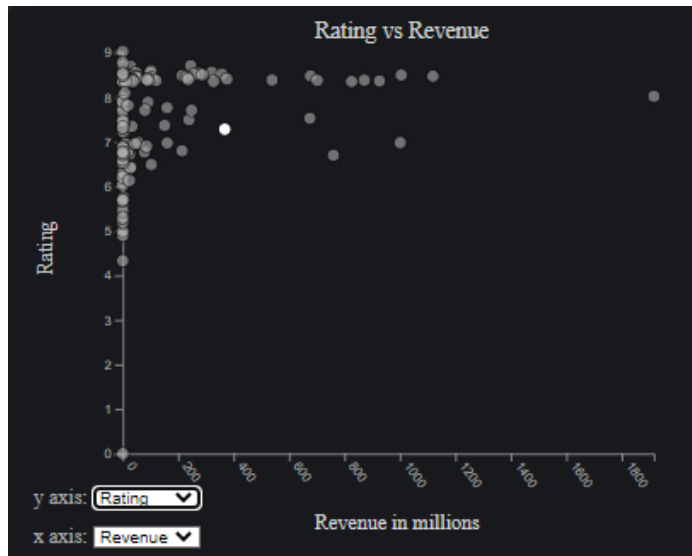
Another part to be improved is the data we show on the dot plot. There were lots of movies within our data that didn't have any revenue or budget reported. This skewed our distributions and we could have removed those data points from the dot plot. We can see this most prominently when comparing revenue and rating. We can see all the data along the y axis, when x(revenue) is 0.

This was a similar problem with our revenue bar on the movie table. This almost always sets the domain at 0 and leaves small bars with not a lot of variety in length.

Milestone Check In 11/15

After our milestone check in with our TA we have established two main takeaways for changes and updates on our project

1. Have less color with our visualizations and use color sparingly
   a. Color is one of the weaker encodings for data visualization, and should be used with intention. We will make changes to our barchart to not show color and possibly only use color to highlight selected bars.
2. Focus on Interaction
   a. Have interaction happen across all our visualizations where one click will affect all our charts.
   b. Add tooltips to barcharts, tables, and scatter plots
   c. Add hover functionality to highlight data, changing opacity