MOVIES GLORIOUS MOVIES

By Don Harrison, Julia Ma, Sara Nurollahian
{u1131452, u0918404, u1217653}@utah.edu
CS 5630/6630 Final Project

PROCESS BOOK

December 4, 2020

TABLE OF CONTENTS

1. Project Proposal

Background and Motivation

The rise and fall, as well as acquisition and merging of media companies is of particular interest to us. We all like movies and this dataset provides interesting features to explore, particularly budget, revenue earned by a movie, its company, and the review score. We hope that visualizing the success of movies as well as information about the success of movie distributors will give us a clue about why companies succeed, fail, get bought out, or merged.

We also are interested in seeing what drives a movie's popularity over time. Is it purely because of the budget spent on the film? Does a movie's success financially correlate with the reviews it gets online? These two questions are what motivated us to choose this project.

Project Objectives

As we all are interested to learn more about movies and their companies' success, we use this project as an opportunity to first understand which movies and their companies were successful over time and how different factors play a role in their success. Some of the questions we are interested to examine are:

- What genre was more successful at what time?
- What distributors were more successful?
- The relationship between reviews and the monetary success of movies?
- How did movie distribution companies grow over time?

Data

We are using a combination of two data sets, "Highest Hollywood Grossing Movies.csv" SANJEET SINGH NAIK: Top 1000 Highest Grossing Movies | Kaggle and nice!"Movies(1986-2016).csv" DANIEL GRIJALVA: Movie Industry | Kaggle to answer. We originally had an idea to use a NASDAQ data set, containing historical company stock information to compare the distributor stock information to its movie release information. After our design process, we realized this data was going to be a lot of work for not much gain, so we decided to not use that data set.

Between the two movie datasets, we used the following attributes:

- Movie Name
- Distributor Name
- Original Release Date
- World Sales
- Genre
- IMDb Score

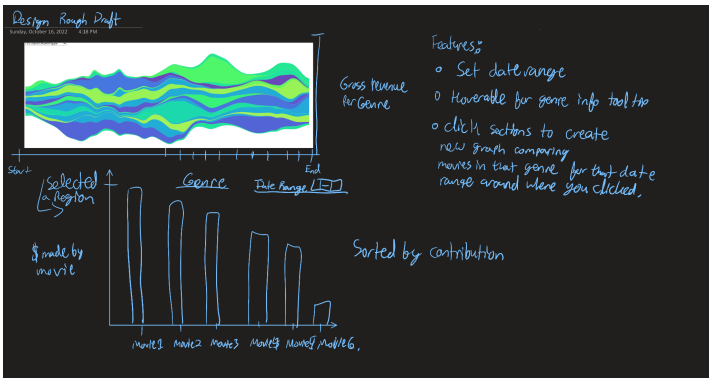
Must-Have Features

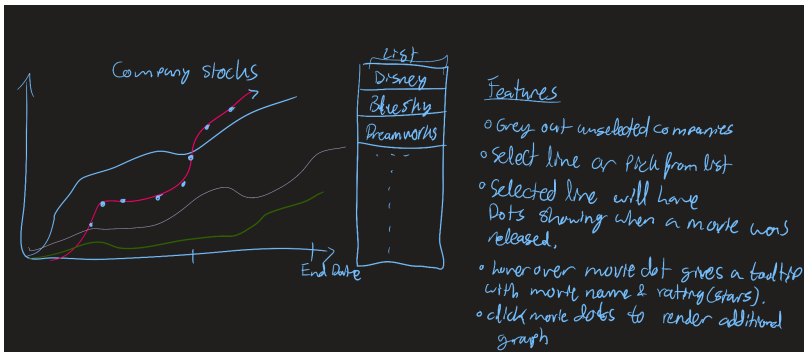
- **Line Chart (Company Success/Movie Chart)**
- **Bubble Chart (Movie Rating/Revenue Chart)**
- **Stream Chart (Revenue/Genre)**
    - These are the visualizations driving our data and without any one of these we feel the visualization will be lacking in painting a whole picture.
- **Tooltips on selection giving more info**
    - Since each movie contains so much info, we are using three separate but interconnected visualizations to portray our data. To help with the interconnectedness, we want mouseover and click event tooltips to appear on various elements on all the graphs. These tooltips will portray missing data on the visualization and connect the three graphs.
- **Click event on one graph updates every graph**
    - We want a selection of a movie on one graph to select the movie on all the graphs and highlight the movie across the whole page. Again, it connects the visualizations and allows viewers to see the movie's information (company success, rating, revenue, and genre) across the whole page.

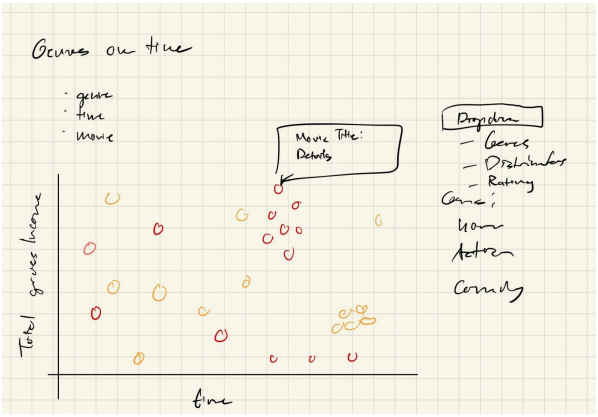
Optional Features

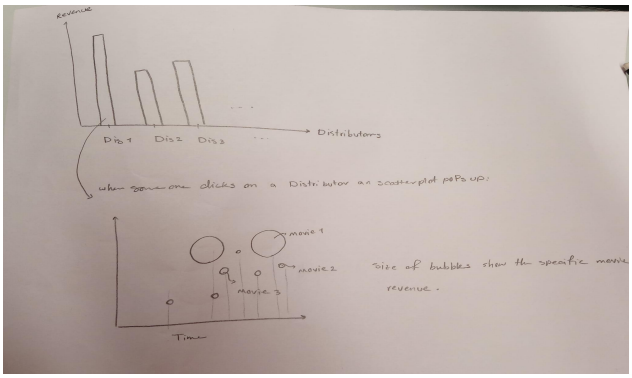
- **Search bars on tables**
    - Since there are many companies to choose from on the Company Success/Movie Chart, we want to add a search bar to the graph so you can go ahead and search for the line to select on the chart.
- **Add transitions to dots on Company Success/Movie Chart**
    - When a company line is selected, dots which show the movies published by that company would grow on the line.
- **Table sorts on different selections.**
    - We want to add a table with additional information about the movies in our data set, such as director, country produced in, main actors, etc. We want this table to sort by certain fields based on what you've clicked on in the page. For example, if you select a company line on the Company Success/Movie Chart, it will sort the table by Company, so you can see the extraneous information about the movies produced by that company in the table.

2. Design Process





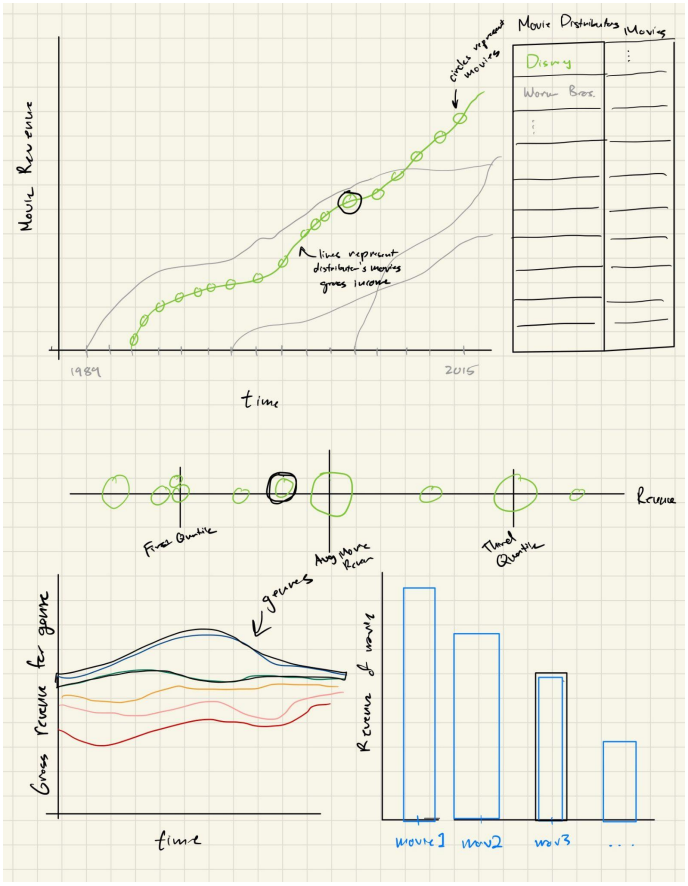




Sketches from initial brainstorms/designs

The questions we want to answer are too different from each other to be answered in one visualization. Our brainstorm contained many different visualizations such as scatterplots, line and bar charts, stream graphs, and beeswarm plots. We had a difficult time tying the charts all in together but an even harder time trying to find one visualization that could do the job. We decided the legibility of individual graphs was more important so we tried to find different visualizations that could work well together. We quickly scrapped the idea of scatter plots as a primary feature since we had so many data points.

We decided to feature three visualizations that each focus on a different aspect but interact with each other. A line chart will be used to show the cumulative company revenue with the gross income of the movies belonging to the company. A bubble chart will compare the IMDb score vs. movie revenue. To display genres, we will use a stream graph over time. This graph will be paired with a bar chart to display movies associated with a selected genre. Finally, we also will use a table to provide extra information for possible questions that a user may want to examine.



Movie Title	Release Date	Distributor	Genre	Other Details...

Final Sketch.

3. Feedback

Peer Feedback

We were reviewed by Milena Belianovich, Tark Patel, and Xiaoya Tang. In summary, they felt our proposal and initial idea was strong and had few notes. Their primary feedback is, “{they’re} not sure if the program is innovative or not, but the data is interesting and the approach to the visualization is good.” Our main takeaways from their feedback is that since we utilize basic, foundational data set types, we need to capitalize on other aspects to make our visualization interesting. This would come in the form of interaction, color, and added features to build on our graphs.

We already have numerous interactions planned as every graph will update in response to mouse events like mouseover and click. We will include nice transitions with these features to make the visualization more aesthetically satisfying and pleasing. Color was an aspect we had not talked about during the proposal process and through their feedback we are ensuring we are more conscious about color by assigning color palettes and making decisions on color in our drafts. Finally, as added features, we initially thought search bars and certain tooltips would be optional but feel they should fall under must-haves as without those features our visualization would suffer; it would have a bare and unfinished quality without those added features.

TA Feedback

We were assigned Pranav Rajan as our TA throughout this process. He gave us similar feedback as our peers in our description and proposal is complete and interesting. Additionally, he suggested adding widgets to update and interact with our visualization.

Milestone Review Feedback 11/10

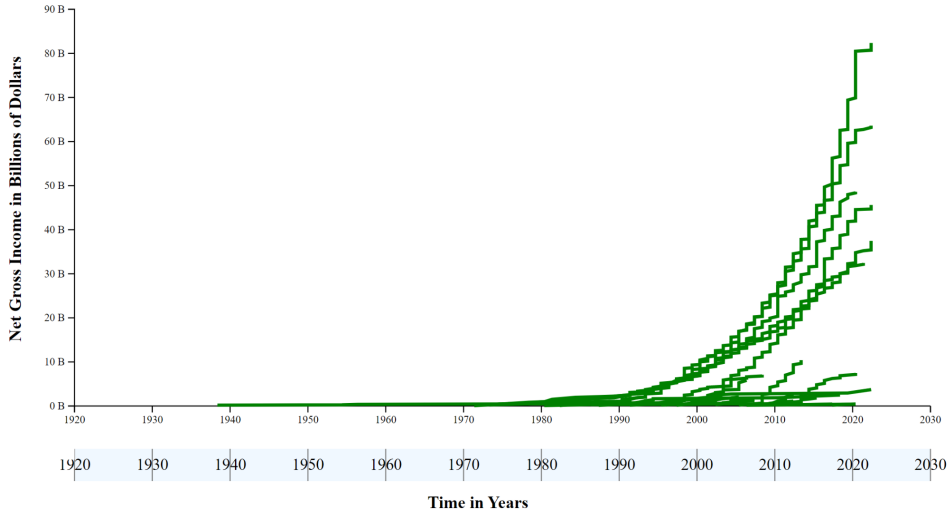
At this point, our TA’s main concern is that the bigger picture of our visualization was not clear. Since we each worked on different parts of the visualization, the interaction had not been implemented yet and acted as three separate data visualizations. He warned us to be careful of a lot of scrolling and scaling and told us to keep in mind spacing. Besides small details in individual charts, he gave us positive feedback and applauded us on our equal partition of work.

4. Implementation

Overall Layout

We used grid styling to arrange all of our graphs. The line chart and distributor table will be displayed together since they both describe movie distributors. Below it is the beeswarm chart, which shows movie incomes compared against each other. Next, the stream graph and bar chart are displayed together since both are concerned with genres. Finally, at the bottom of our page, is a table displaying all the string information we were unable to show in our visualizations. We display this data for users who are interested in things like directors and writers.

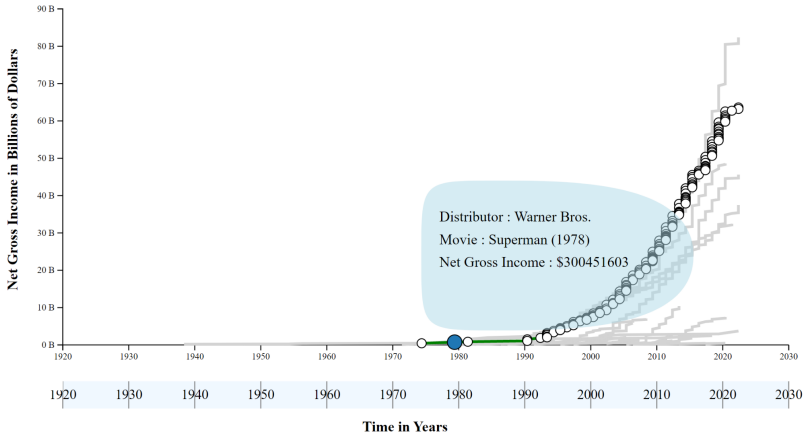
Linechart



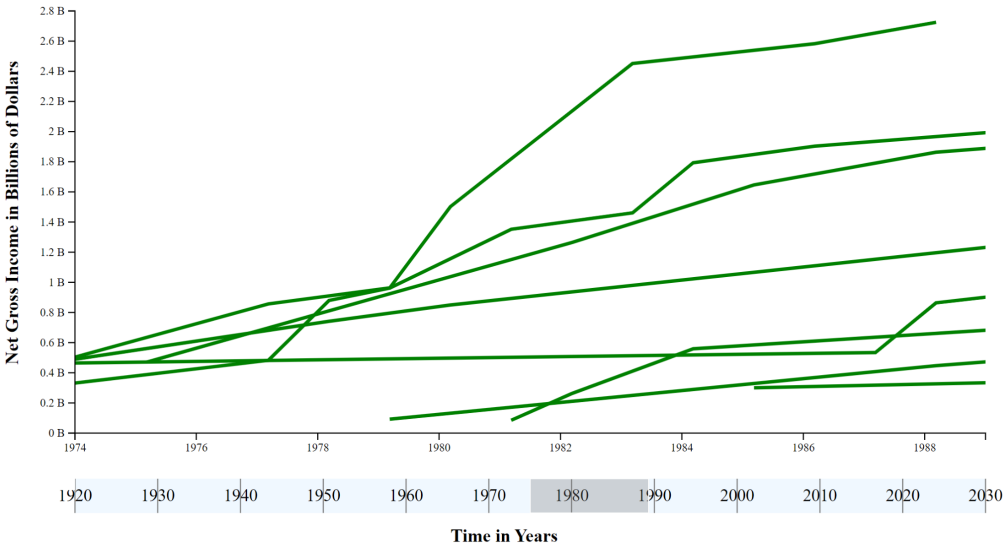
The line chart shows distributors' net gross income by the top 1,000 highest grossing movies over time. The data was filtered by distributor and a cumulative income was calculated for every movie sorted by time, to give the line chart its x and y values. The paths increment in steps because it updates in years rather than months and dates. We tried to convert it to using the entire release date data column, instead of just the years, but quickly found many of the release date entries are empty. After consulting with our TA regarding if we should look for the data and fill the columns, he told us we should not alter our original data set so we settled on the stepping lines.

The chart has hover interactions that will highlight a path and a tooltip will display what distribution company that path is describing. On selection, circles will populate the line representing every movie belonging to that distribution company. Hover events for the circles will show the movie title and additional information. Selection of a movie will color it according to genre and highlight the movie in the distributor table, beeswarm chart, and update the genre

graph with the genre of the selected movie. A click anywhere on the chart other than the paths or circles will reset the table.



After implementation, we observed the cumulative data is exponential due to concentration in recent years with outliers dating back 100 years. This results in most of the graph being empty space and not very interesting to look at. To combat this, we have a brushable, dynamic timeline, to allow users to decide what window of time they would like to look at. The y axis scales alongside it to prevent the chart from being difficult to read. A single click on the timeline will reset the year selection.



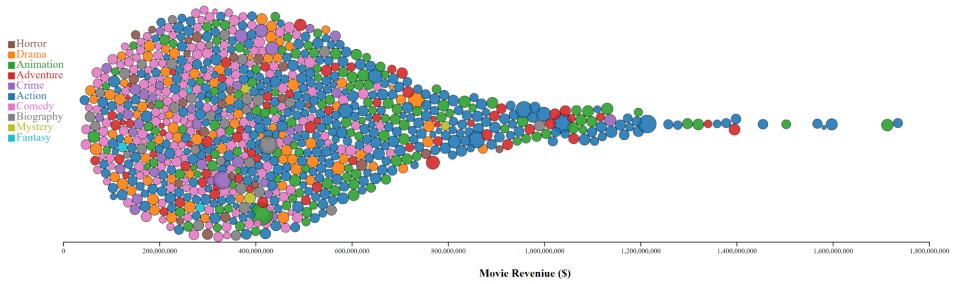
Distributor Table

Distributors
Walt Disney Studios Motion Pictures
Twentieth Century Fox
Sony Pictures Entertainment (SPE)
Paramount Pictures
Universal Pictures
Warner Bros.
DreamWorks Distribution
Lionsgate
DreamWorks
New Line Cinema
Newmarket Films
Summit Entertainment

The table shows the same data used by the line chart but lists out all the distribution companies for users to read. A selection on the table will highlight the line in the line chart and display a second table. This table lists all the movies put out by the selected distributor in our data set. A selection of a movie will highlight that movie in the line chart and beeswarm chart.

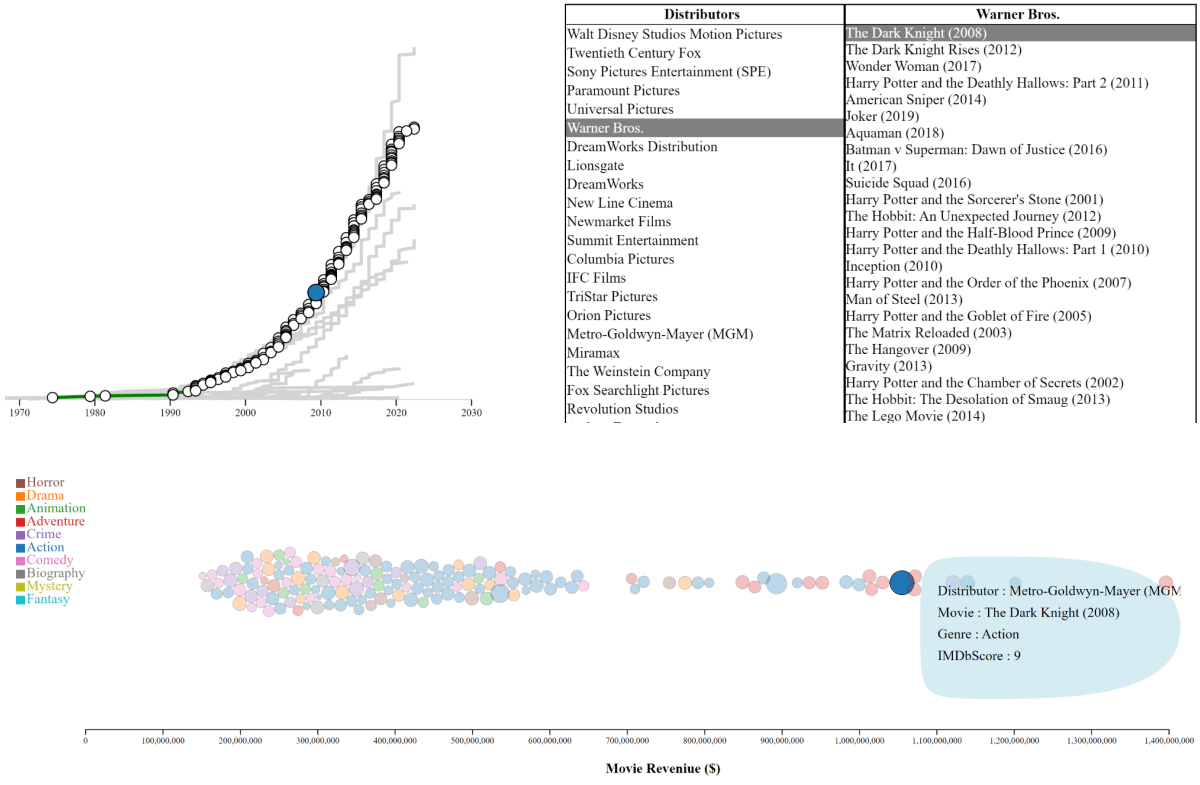
Distributors	Walt Disney Studios Motion Pictures
Walt Disney Studios Motion Pictures	Star Wars: Episode VII - The Force Awakens (2015)
Twentieth Century Fox	Avengers: Endgame (2019)
Sony Pictures Entertainment (SPE)	Black Panther (2018)
Paramount Pictures	Avengers: Infinity War (2018)
Universal Pictures	The Avengers (2012)
Warner Bros.	Star Wars: Episode VIII - The Last Jedi (2017)
DreamWorks Distribution	Incredibles 2 (2018)
Lionsgate	The Lion King (2019)
DreamWorks	Rogue One: A Star Wars Story (2016)
New Line Cinema	Star Wars: Episode IX - The Rise of Skywalker (2019)
Newmarket Films	Beauty and the Beast (2017)
Summit Entertainment	Finding Dory (2016)
Columbia Pictures	Frozen II (2019)
IFC Films	Avengers: Age of Ultron (2015)
TriStar Pictures	Toy Story 4 (2019)
Orion Pictures	Captain Marvel (2019)
Metro-Goldwyn-Mayer (MGM)	Pirates of the Caribbean: Dead Man's Chest (2006)
Miramax	The Lion King (1994)
The Weinstein Company	Toy Story 3 (2010)
Fox Searchlight Pictures	Iron Man 3 (2013)
Revolution Studios	Captain America: Civil War (2016)
Artisan Entertainment	Frozen (2013)
Sony Pictures Classics	Guardians of the Galaxy Vol. 2 (2017)
United Artists	Finding Nemo (2003)
Screen Gems	The Jungle Book (2016)
USA Films	Inside Out (2015)
20th Century Studios	Aladdin (2019)
	Zootopia (2016)
	Alice in Wonderland (2010)
	Guardians of the Galaxy (2014)

Beeswarm Plot



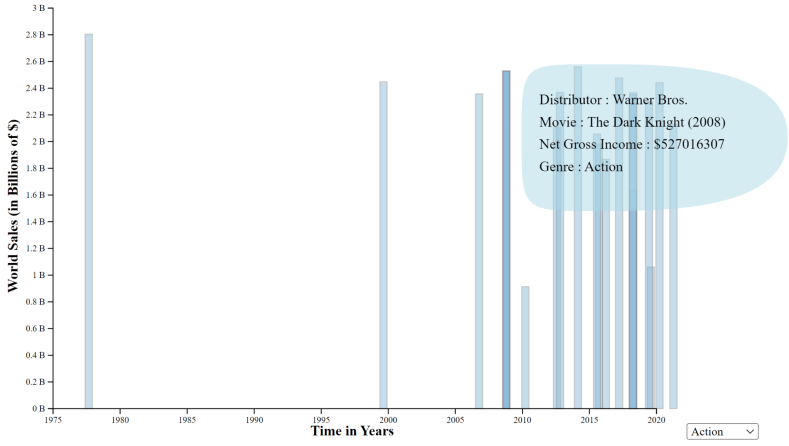
The beeswarm plot visualizes movies' revenue against each other. The size of the bubbles reflects its IMDb score (on a scale of 1-10) and the color shows its genre. As a default, the plot shows all 1,000 movies from the data. The selection of a distributor above will update the beeswarm plot to only plot that distributor's movies.

The selection of a movie above will highlight the corresponding bubble so users can compare that movie to the success of other movies by the distribution company.



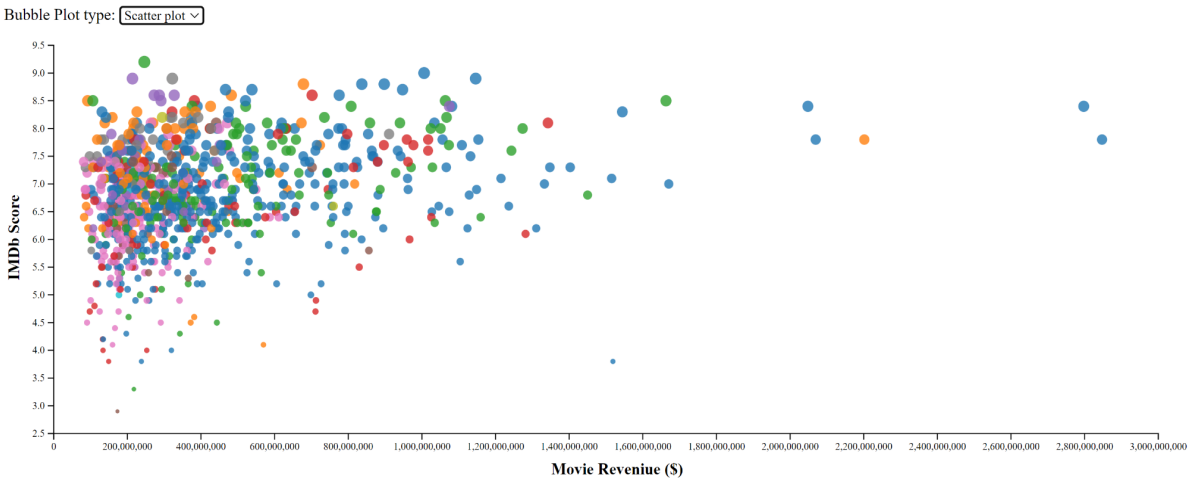
Hover events on the beeswarm plot display the movie the hovered circle represents as well as additional movie information. The selection of a movie on the bubble plot updates the barchart

based on the genre of selected movie. The bar chart updates to show the top twenty contributors in income of that genre.



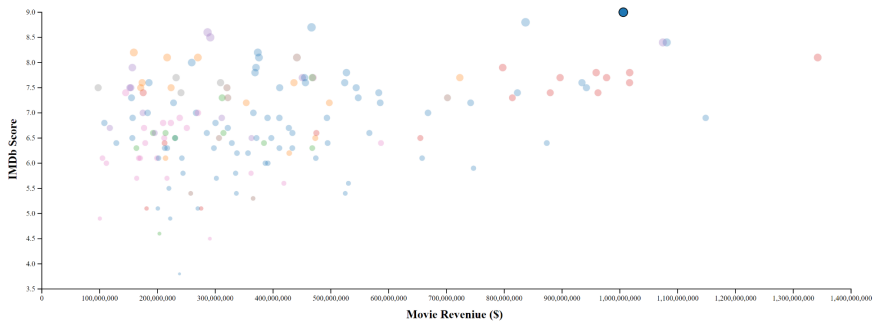
After implementation, we were concerned with how concentrated the bubbles were on one side of the plot and considered altering the plot or axes. Then we realized, when all movies are displayed on the beeswarm plot, the majority of movies gain less than $800 million and less than 10% of movies have an income of $1 billion or more. We kept the plot as it was since it offers insightful information about the average monetary success of the movies.

Scatter Plot

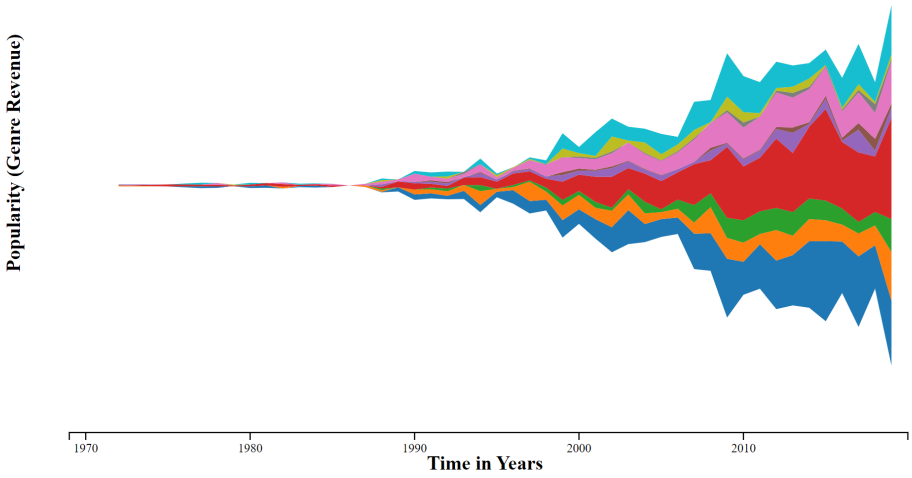


A dropdown is provided allowing users the option to display the beeswarm plot as a scatter plot. The x axis still shows revenue but the y axis encodes the IMDb score. We included this because although the score is shown in the radius of the bubbles, it is difficult to compare sizes of all the data points. Here, it is clear that all variation in movie revenues can not be justified with IMDb score alone, as some low scoring movies have greater revenues than high scoring movies.

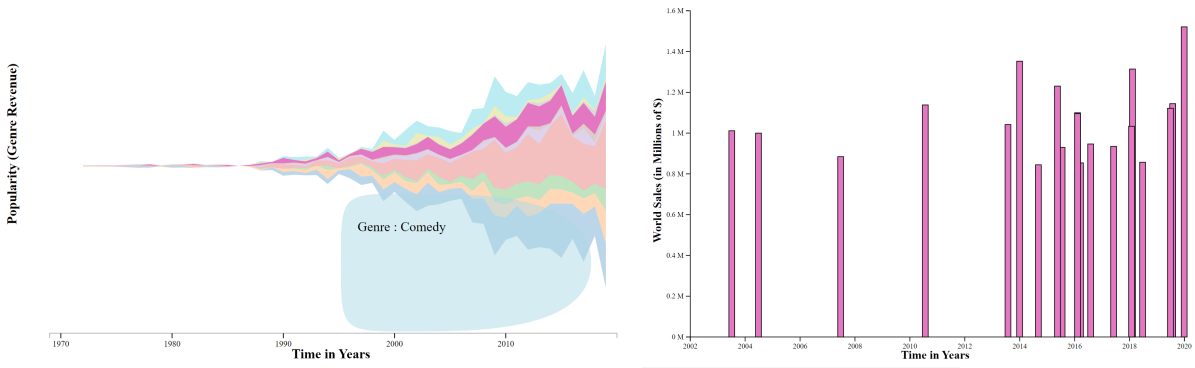
Similar to the beeswarm plot, the selection of a movie above will highlight the corresponding circle on the scatter plot to allow comparisons against other movies.



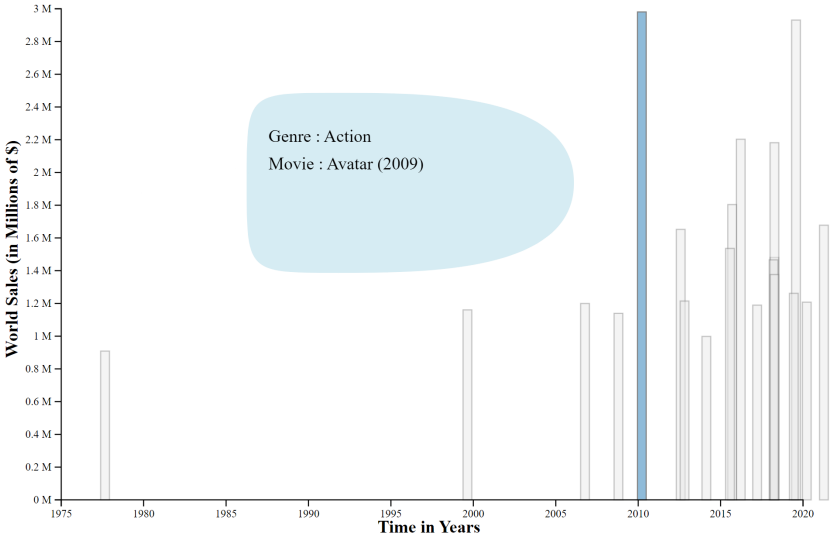
Stream Graph



The Stream graph sums the total income of movies of a genre in a given year. This lets us visualize the “popularity” of a genre by seeing what genres were successful over time. The genres are colored with the same scale seen across the page, with the key by the beeswarm plot. Hovering over a segment will highlight that genre. Clicking a segment will leave it highlighted while you are hovering over the chart so you know which segment is selected as the barchart is populated with the movies for that genre. The tooltip remains visible as you hover over each segment so you know what genre is highlighted. We removed the y-axis because it isn’t very helpful since it would be difficult for the user to get any useful numbers from an axis. Because of the nature of the visualization, the width of a segment is the most important part and we wanted the user to focus more on the size of the sections at a given time, compared to the other segments, rather than numerical data.

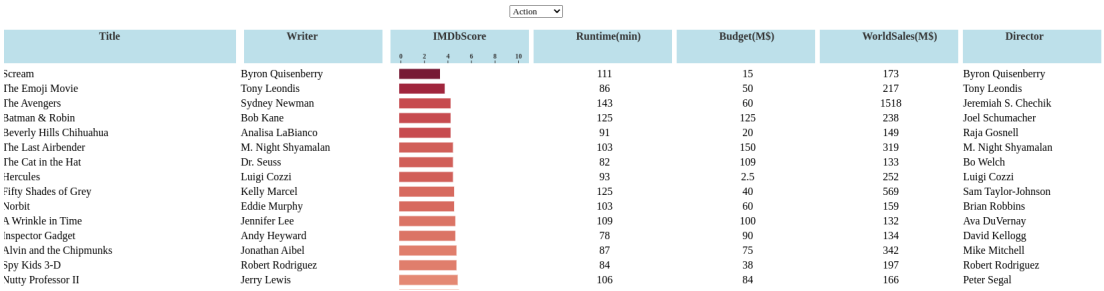


Bar Chart



This chart returns the top 20 grossing movies for a selected genre by world sales. On hover, the bar will gray out the other bars for added focus and the tooltip will inform you of the movie you selected. As you move your mouse around the bar chart the tooltip remains visible to inform you of the genre you are exploring.

Info Table



The final table provides additional text information to the user. Here, a user can find the writer and director of a movie. The table header sorts the data ascending or descending when clicked. IMDb score is displayed with bars rather than numbers for easier comparison. The color of the bars diverge from red to blue, scaling from 1-10. A score of 5 is gray. Below 5 is a redder hue and above 5 is a bluer hue. We double encoded this aspect so when a user is viewing the bars away from the header, the IMDb success is obvious.

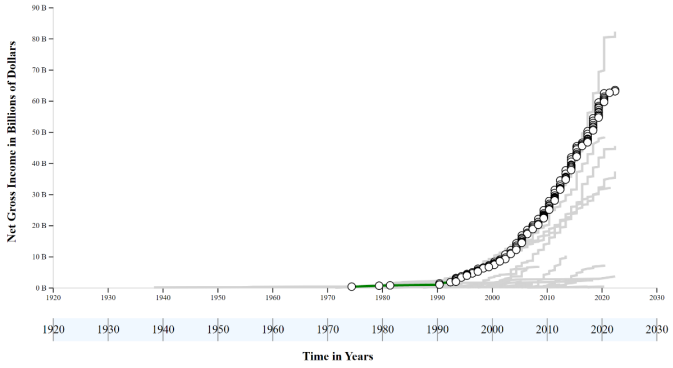
5. Evaluation

Data Findings

With our visualizations, we are able to view the movie data in numerous diverse ways. Through our page, we discovered something interesting regarding the distributors. *Walt Disney Studio*s is the dominating distributor in net gross revenue. This is due to a number of factors we know about the company; it has movies dating a hundred years and has absorbed a number of other large companies (*Marvel* and *Lucasfilms* contributing a lot of revenue to *Disney* in its later years). Aside from *Disney*, there are certain movies well known for holding the record for highest grossing films of all time: Titanic (*Paramount Picture*s), Avatar (*Twentieth Century Fox),* and most recently Avenger: End Game (*Walt Disney Studios Motion Pictures*). One would assume

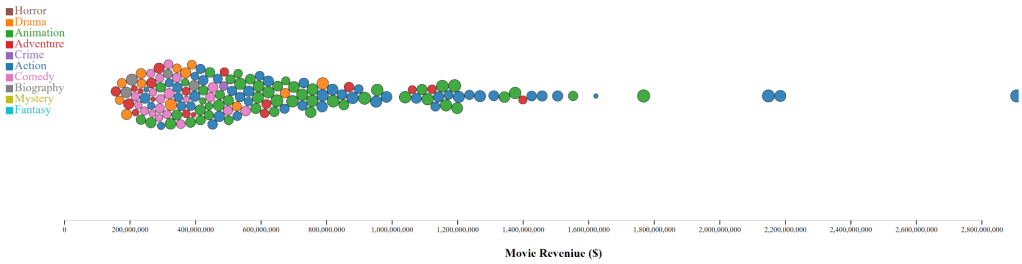
the next dominating distributor would be one of these studios, however there is another clear winner for the title after Disney, *Warner Brothers*.

When looking at *Warner Bros.'* lineup of highest grossing movies, you will see their highest grossing movie is The Dark Knight, a fan favorite but not internationally loved like the top three highest grossing movies. In fact, none of *Warner Bros.*' movies fall in even the top 10 highest grossing movies. So then what contributes to their success over the other companies?

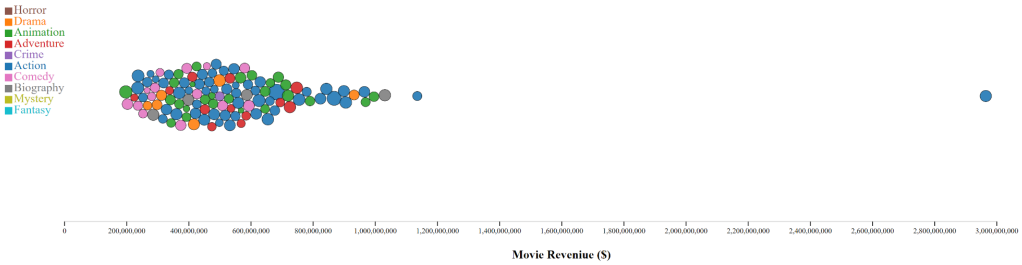


Distributors	Warner Bros.
Walt Disney Studios Motion Pictures	The Dark Knight (2008)
Twentieth Century Fox	The Dark Knight Rises (2012)
Sony Pictures Entertainment (SPE)	Wonder Woman (2017)
Paramount Pictures	Harry Potter and the Deathly Hallows: Part 2 (201
Universal Pictures	American Sniper (2014)
Warner Bros.	Joker (2019)
DreamWorks Distribution	Aquaman (2018)
Lionsgate	Batman v Superman: Dawn of Justice (2016)
DreamWorks	It (2017)
New Line Cinema	Suicide Squad (2016)
Newmarket Films	Harry Potter and the Sorcerer's Stone (2001)
Summit Entertainment	The Hobbit: An Unexpected Journey (2012)
Columbia Pictures	Harry Potter and the Half-Blood Prince (2009)
IFC Films	Harry Potter and the Deathly Hallows: Part 1 (201
TriStar Pictures	Inception (2010)
Orion Pictures	Harry Potter and the Order of the Phoenix (2007)
Metro-Goldwyn-Mayer (MGM)	Man of Steel (2013)
Miramax	Harry Potter and the Goblet of Fire (2005)
The Weinstein Company	The Matrix Reloaded (2003)
Fox Searchlight Pictures	The Hangover (2009)
Revolution Studios	Gravity (2013)
Artisan Entertainment	Harry Potter and the Chamber of Secrets (2002)
Sony Pictures Classics	The Hobbit: The Desolation of Smaug (2013)
United Artists	The Lego Movie (2014)
Screen Gems	I Am Legend (2007)
USA Films	The Blind Side (2009)
	The Hobbit: The Battle of the Five Armies (2014)
	The Hangover Part II (2011)
	Batman (1989)

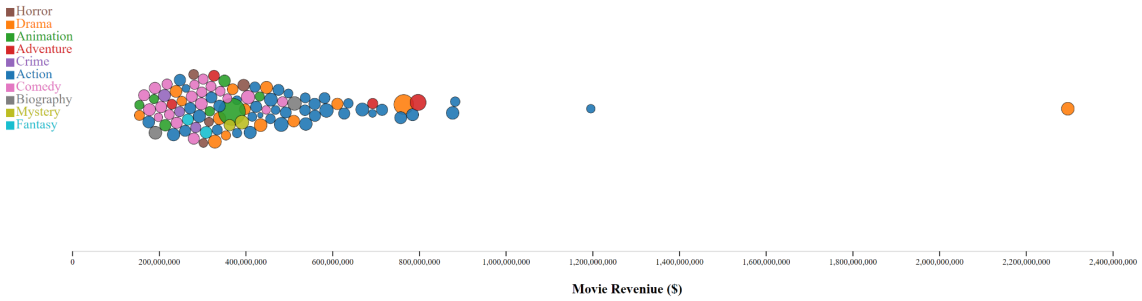
If you observe the same four companies in the beeswarm plot, the answer becomes clear. Although *Disney*, *Fox*, and *Paramount* put out the three highest grossing movies of all time, those movies are outliers for their companies. Outside of those three movies, these companies put out movies that on average make less than 700 Million. If you observe *Warner Bros.*, you will see a smoother range and better distributed chart. Thus, it is observed, releasing a highest grossing movie of all time does not contribute to a company's success if everything else they released is subpar. A company will be more successful if they are able to produce consistently good movies rather than focussing on a single smash hit.



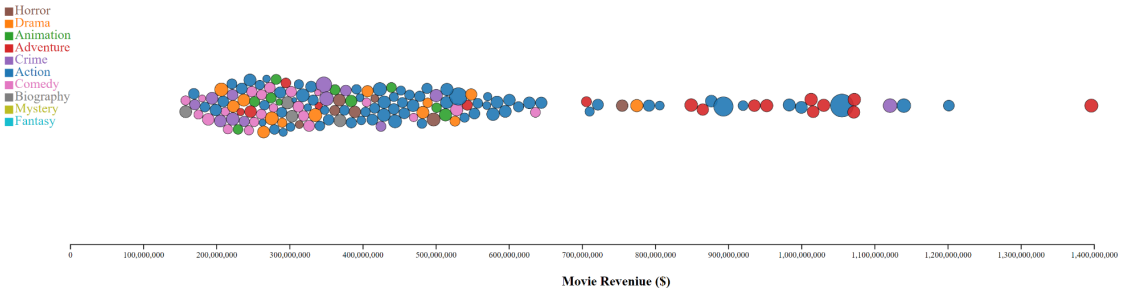
Walt Disney Studios Motion Pictures (Avengers: End Game)



Twentieth Century Fox (Avatar)



Paramount Pictures (Titanic)

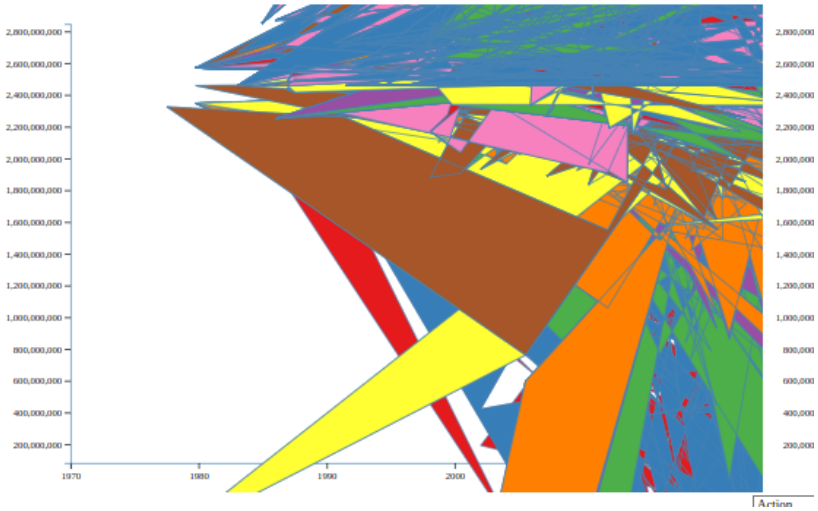


Warner Brothers (The Dark Knight)

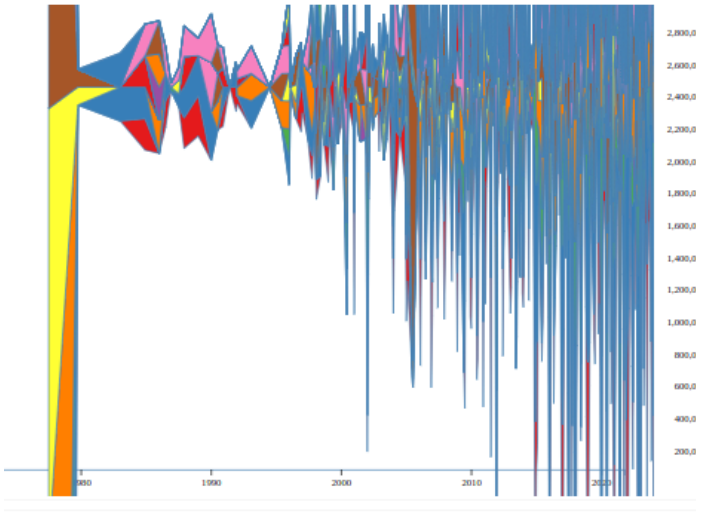
Reflection

Our visualization is successful in answering the questions we posed. It is able to display a very large set of data with multiple differing attributes. We struggled with connecting the charts and finding a streamlined way to implement all the different visualizations. The styling, colors, and interactions is what we focused on to tie the visualizations together. This turned out to be a struggle since we all imagined a different implementation for each of these channels. However, we were able to pair-program through it and come to a consensus on these decisions.

The most difficult visualization was the stream graph. The biggest issue was how to manipulate the data into a format that D3's stack function could parse and return the data for the render. Initially, the data was grouped by month and year, and after parsing the data through the stack function, the render looked like the following image.



Instead of using months, we tried dates which resulted in the following image.



At this point, we grouped by year to minimize noise and we are using those final results in our visualization. Grouping by dates or months would have been more accurate but we felt it was more important to have a legible visualization.

The visualization works how we intended it to but there are some bugs in interactions. It succeeds in its purpose but in communicating with all other graphs at all times, minor bugs or lagging arises. To further improve it we would like to add more finesse in the details. We would

want to fix the lagging from the force simulation of the beeswarm plot and from the charts calling other charts to update. Our visualization has an object that all the visualizations share and when a mark is selected, the data is passed through the object then calls a function to redraw others. We did this to minimize redundancy but we would be interested in finding a faster way to do this. Additionally, we would want to add smoother transitions and add animations.