MOVIES GLORIOUS MOVIES

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

PROCESS BOOK

December 2, 2020

TABLE OF CONTENTS

1. <u>Project Proposal</u>

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 companies 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

Our data is collected from Kaggle. A combination of the following two data sets will answer our questions relating to movies, release dates, distributors, viewer ratings, genres, gross income, and more.

"Highest Holywood Grossing Movies.csv" SANJEET SINGH NAIK: [Top 1000 Highest Grossing Movies | Kaggle](#)

"Movies(1986-2016).csv" DANIEL GRIJALVA: [Movie Industry | Kaggle](#)

We found an additional dataset that contains historical company stock information. We believe it will be interesting to compare the distributor stock information to its movie release information but may or may not use it. This data was also found on Kaggle.

"NASDAQ Historical Data" OLEH ONYSHCHAK: [NASDAQ dataset](#)

Data Processing

Most of the data we require is in the datasets we found on kaggle. As an optional feature, we were considering comparing the movie gross income to its distribution company's stock data. We would compare the success of a movie to its distributor's company success as a whole. We found a dataset that contains historical stock data of companies over a range of years. To get the stock data, we would need to collect all the movie distribution companies off our list of 1,000 highest grossing movies and retrieve their Nasdaq keys from the Nasdaq symbols metadata set. Then we could download the individual company stock csvs which contain the company stock data per year and compare them to our movie data. This would be the only data we would have to spend some time to acquire. Otherwise we do not expect to spend much time cleaning data.

Between the two movie datasets, we plan to use the following:
- Movie Name
- Distributor Name
- Original Release Date
- World Sales
- Main 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.

Project Schedule

Oct 21: Project Proposal Due
- Turn in project proposal and determine individual responsibilities

Oct 28:
- Data ready to use
- Webpage set up with svgs holding our visualizations

Nov 4:
- Individual graphs working but not interacting
- Process book set up

Nov 11: Project Milestone
- Work towards class milestone submission
- Update process book
- Completed Data Acquisition
- All Data structures in place
- Working visualization prototype

Nov 18:
- Start interaction
- Update process book

Nov 25:
- All graphs interacting
- Update process book

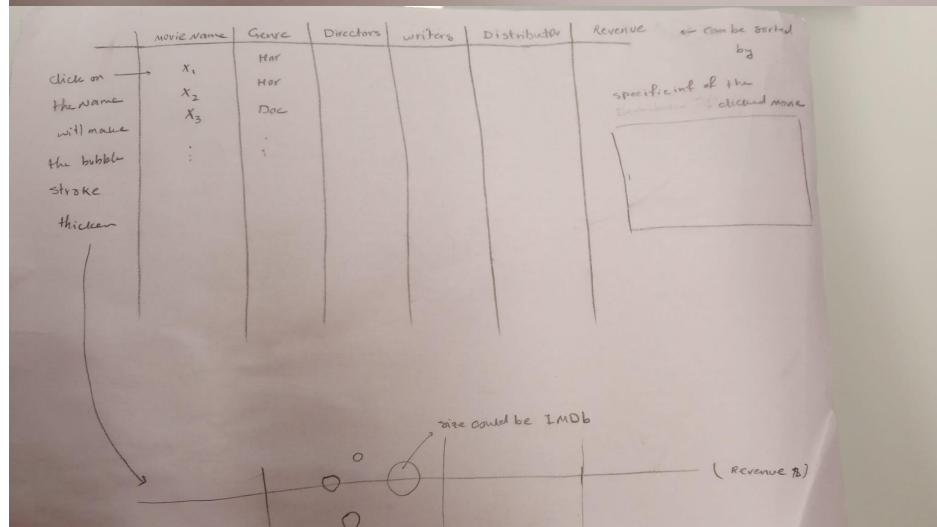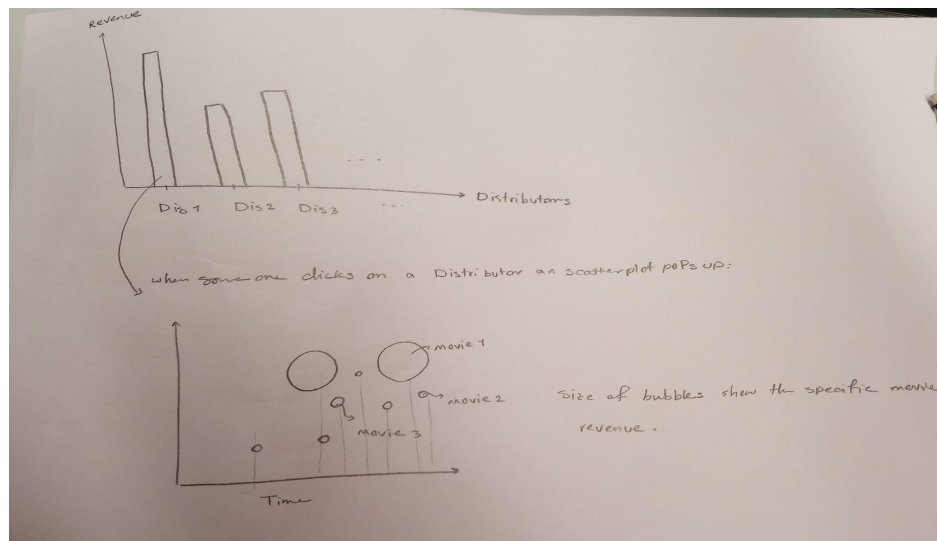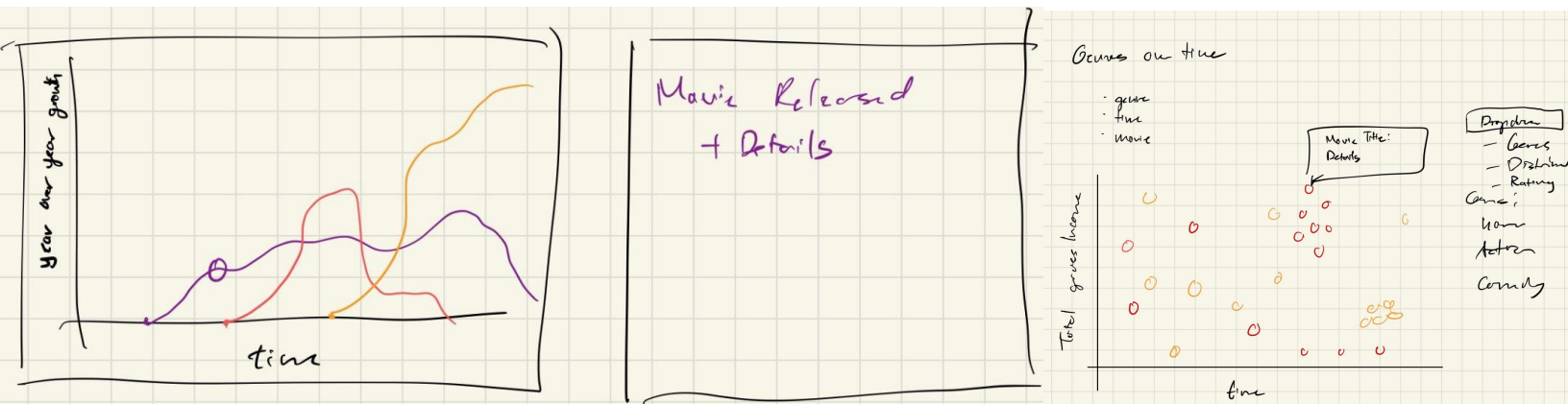**Dec 2: Final Project Due**
- Fix bugs
- Make minor fixes
- Complete process book
- Process book
- Code
- Project Website
- Project Screen Cast

# Visualization Design Sketches

Brainstorm:

Initial Designs:

We decided to use a line chart to show the cumulative company revenue with the gross income of the movies belonging to the company. After brainstorming, we scrapped scatter plots, since we have a lot of data points and felt it was not a good visual representation of our data. We want to display genres in some way and settled on a stream graph to do so. We also liked using the bubble chart to compare the IMDb score vs. movie revenue. Finally, we also will use a table to provide extra information for possible questions that a user may want to examine.

Realization Design:
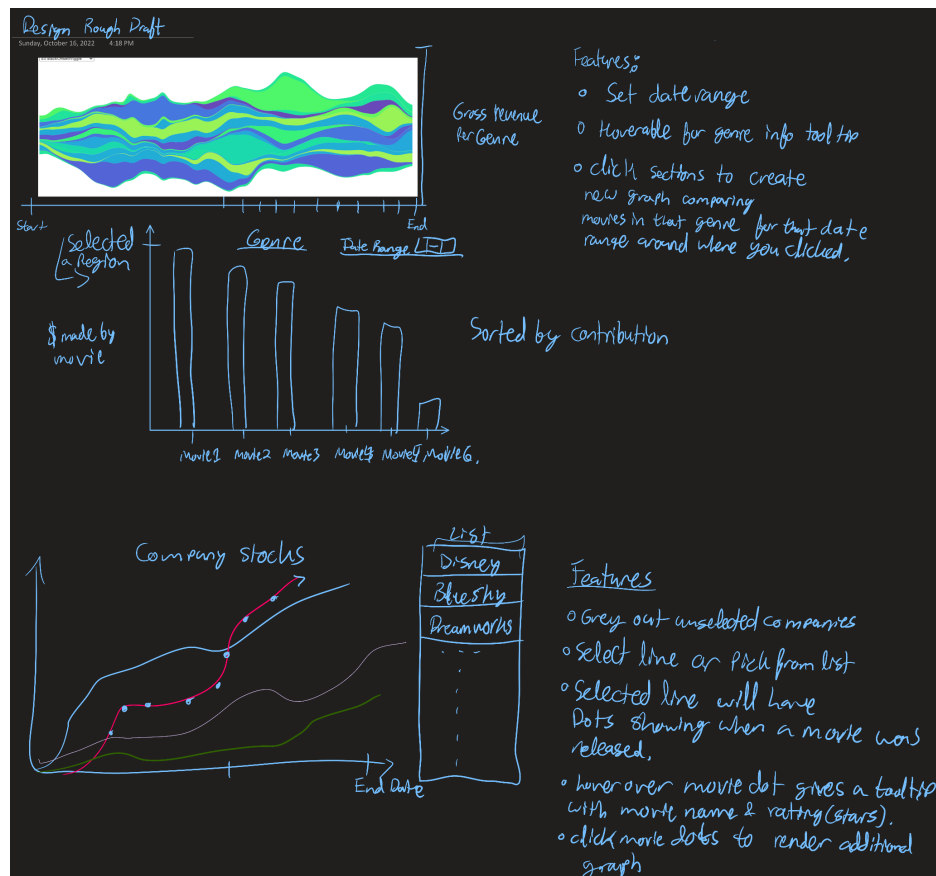
Our final design will feature three separate visualizations that all interact with each other.
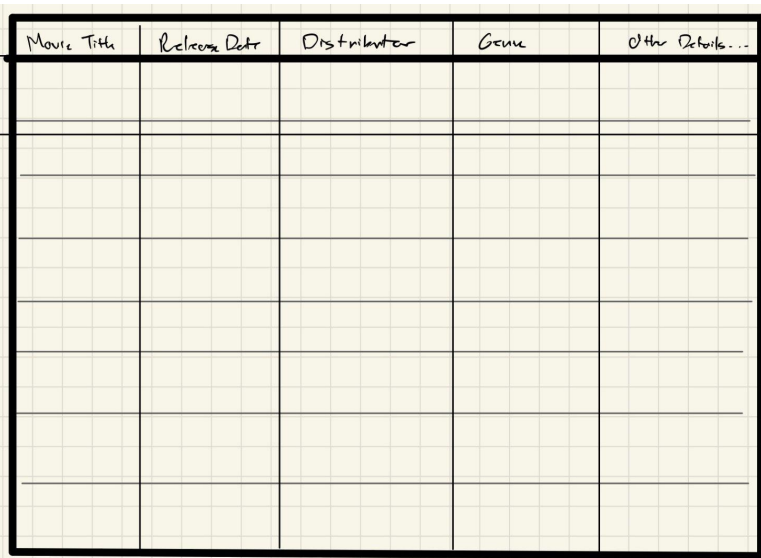
First, is a line chart that displays movie distributor data. The line is the cumulative movie income by each company. When a line gets selected, points get added on the line, each representing a movie. Additionally, a list of movies gets displayed alongside the list of distributors. Every other line gets grayed out.
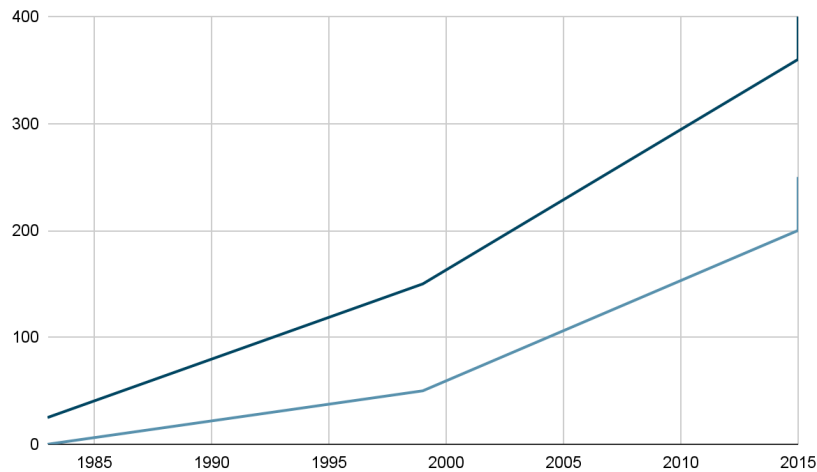
As someone clicks on a single distributor, a bubble plot will display with all the movies of that distributor. It will show the average, first, and third Quartile revenue and the position of bubbles will show that movie's comparative contribution to the distributor revenue. Size of the bubbles show the movie IMDb score. This way we can investigate if the IMDb scores are correlated with the movie revenue and also we can see what movie(s) contribute to most revenues/ score
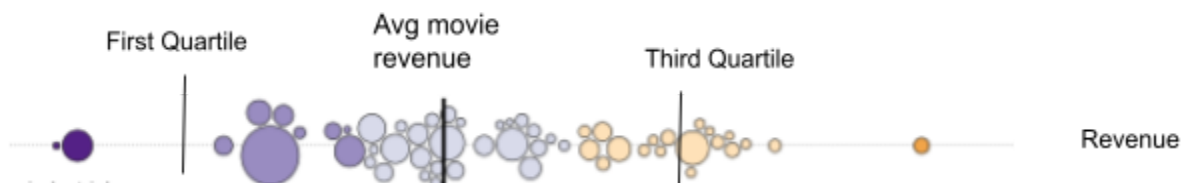
Then, a stream graph will be used to display the success of each genre over time. Each color will represent a genre over time with the y axis being revenue. The selection of a genre will result in a bar chart displaying 10 of the most successful movies of that genre sorted by revenue.
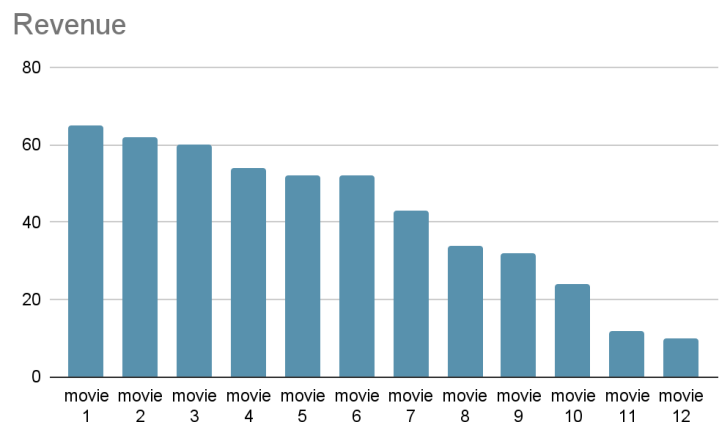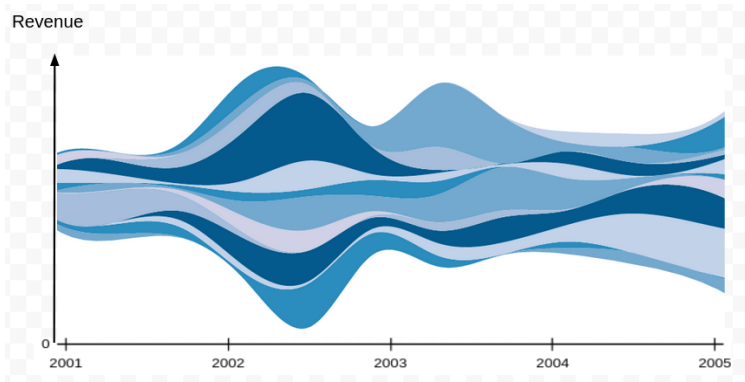
Finally, a table at the bottom of the page will display all extra data. It will get sorted on any event click on the page to display the most relevant data at the top. For example, when a genre is clicked, all movies with that genre will be displayed at the top of the table. When a singular movie is selected, the table will highlight that movie.

Movie Name	Distributor	Genre	Writer	IMBD rating	Revenue	…
Movie 1	Disney					
Movie 2	…					
Movie 3						
…						

2. 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.

3. <u>Getting Started</u>

We created a group repository and set up a project board to organize tasks. We set up a system on how we are creating branches to protect our main source code and allow us to easily make updates.

After this we added all necessary documents to our repo (html, css, and js files) and began processing our data. At this point we decided the NASDAQ data was going to be a lot of work for not much gain, so we decided to not use that data set. We settled on creating an object to hold both our data sets and using the movie titles as keys when we need data from each set. As we work towards the project milestone, we assigned each member a visualization to focus on and work on getting all visualizations up and running without interaction first.

4. <u>Milestone Review Meeting 11/10</u>

Met with Pranav to discuss milestone expectations and to get an opinion of what we have so far. His main concern is that the bigger picture of our visualization is not clear. Since we each worked on part of our visualization, the interaction has not been implemented yet and as of now seem like three separate data visualizations.

LineChart
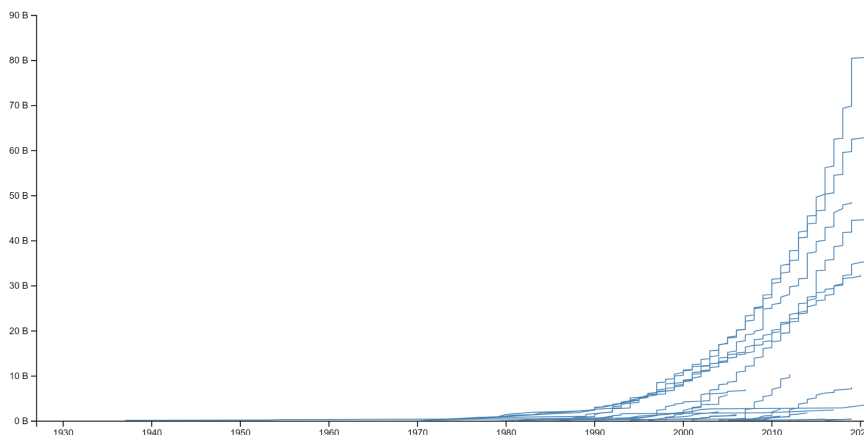
For the line chart, the main concern was the data was all concentrated in recent years but spans nearly 100 years because of a few outliers. This results in most of the graph being empty space and not very interesting to look at. To combat this, we are adding a scrollable,

dynamic x axis, to allow users to decide what window of time they would like to look at. The y axis will also scale to prevent the chart from being difficult to read.

The table displaying distributor data is on the right track, but to make the movie data appear after the distributor is selected, we are going to create a second movie data table which we can turn on and off through opacity. The table is going to update with the selection of a distributor and display that distributor's movies.

BubblePlot

For the bubble chart our plan was to visualize the movie's IMDb score as the bubble size and movie revenue as the x position for each bubble(movie). First we plotted the whole 1000 movies and our bubble chart looks like a spoon, most of the movies focused at the left side of the plot. As the x axis is the movie revenue (M $) we can see most movies' revenue is less than 500 million dollars. However, we can also see that some movies managed to gain even more than 1 billion dollars. The color of the bubbles represents the movies' genre.
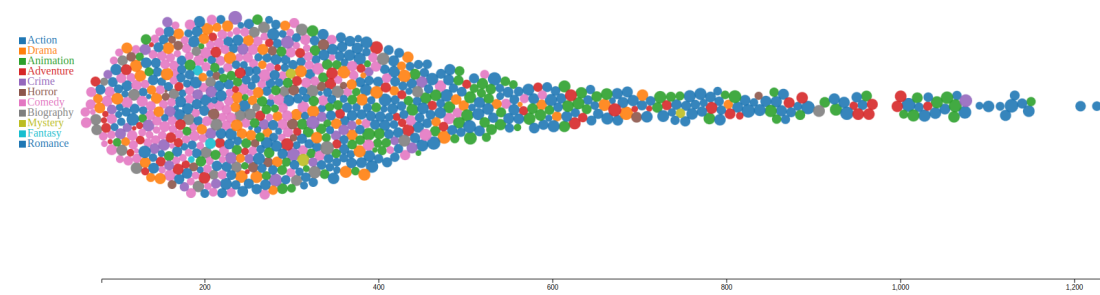
To make the plot more intuitive, we also allow users to explore a specific subset of data (to filter the data) by revenue and by movie company. Using two text boxes, we allow users to enter a company and observe all the movies made by that specific company. With the second textbox, we allow users to filter out the data based on movie revenues. The user chooses a revenue between the dataset minimum and maximum revenue and the js code filters the data to the movies having a revenue less than the user's input. This allows user to more clearly observe and compare movie revenues that were kind of similar and were not distinguishable in the first plot.
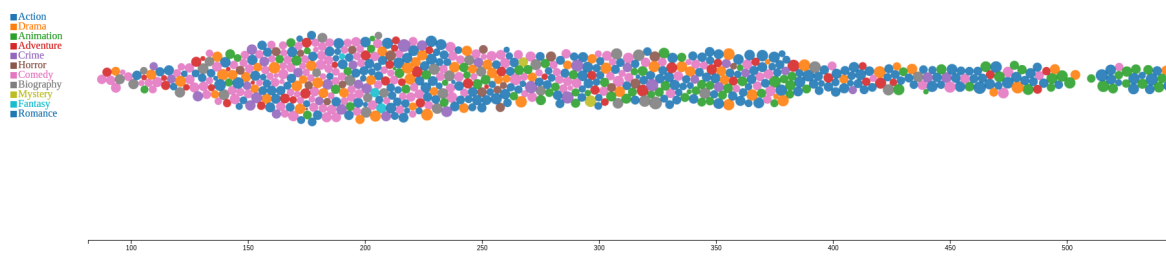
We will make our plots shorter in width, so then the user does not need to scroll to the right to see the whole picture. Further, we planned to use scatter plots to also benefit from our y axes. To make the IMDb scores more observable we planned to use a non linear scale.

{write about issues and next steps for stream chart}

- Need to add some interactions between charts for cohesion.
- Need to merge individual sections into main branch
- Change scaling on bubblechart circle radius-imbd score so that the difference between closeness to 5 stars and lower scores is clearer
- Need to fix data structure of genre-revenue total for genre to render streamchart
- Need to fix any styling issues after merge
- Need to populate movie table after selection of a line on linechart

Stream Chart
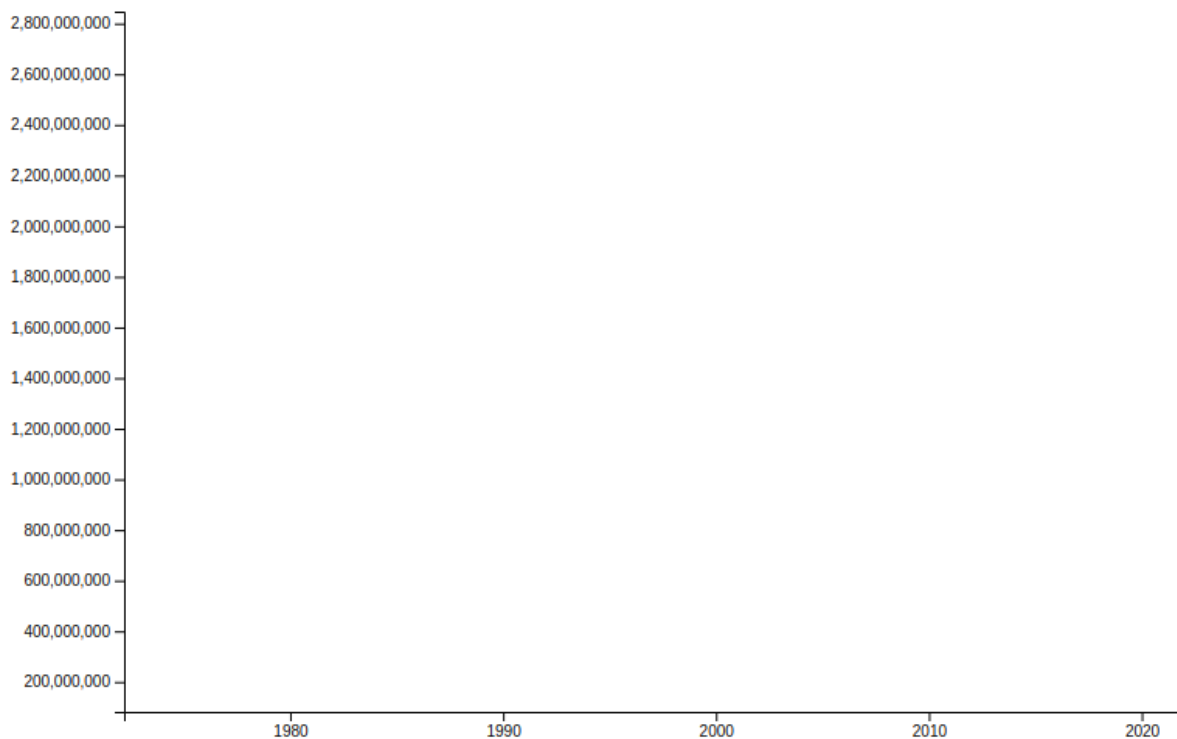
Was able to find some mock code to guide me on how to do this visualization since its a little more complex than I'm used to. Currently have the axes being labeled correctly. The final

issue is just getting the data mutated to be fed into the stream chart code that's written to render it.
Need data formatted st. each row has a col for year, or probably month for finer paths. and then the total amounts from each genre for that year, or month in the other columns.

Bar Chart
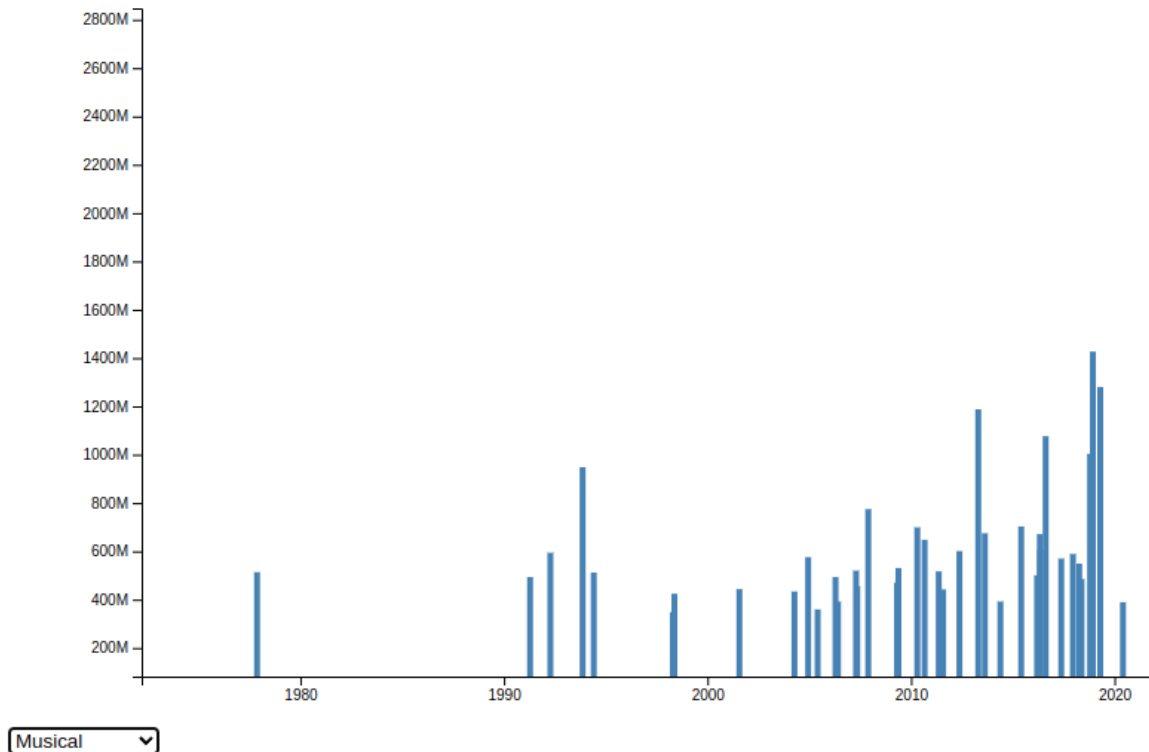
Bar chart rendering fine, issue with NaN from empty year columns, fixed that on one of the branches. Chart has a dropdown with each genre so you can select and see what the top 40 movies in international revenue there are for that genre and what years they came out.

Has basic transition, probably needs better animation for switching between genres.
Also could adjust axes based on the current list rather than all movies.