Unveiling Cinematic Insights: Exploring The Influencing Economic Factors

dcroft@g.clemson.edu, jrhuggi@g.clemson.edu, mtolchi@g.clemson.edu
C51164114, C19989917, C72356453
https://github.com/JonahRileyHuggins/DataVizDashboard

Overview and Motivation

At the formation of the project the investigators identified a united interest in understanding the economics behind the film industry over time. This led to the evaluation of a multiplicity of factors (gender, ratings, genre, revenue, etc.) and presentation of the following project. The Movies Dataset presented a rich source of information that was leveraged for predictive purposes based on available metrics. This dataset, sourced from Kaggle.com, encompassed a compilation of movie data spanning 1930 to 2017. Attributes detailed budget, gross revenue, genre, directorship, votes, as well as other accessory material. This dataset was particularly captivating due to its capacity to address a wide spectrum of inquiries, such as correlations between revenue and votes or release date. This offered valuable insights into the film industry's evolution over the years.

The project was motivated by a collective fascination with the world of cinema and the desire to delve deeper into the intricacies of movie data. There was a collective interest in gaining a more profound understanding of the dynamics within the film industry. Further, we were keen on expanding our knowledge of JavaScript and Tableau; tools we were all familiar with but had yet to utilize for dashboard creation. We aimed at creating interactive dashboards that provided in-depth insights into the movie domain. Here, our objective was to build a dashboard that not only engaged users but also served as a comprehensive repository of information related to the world of movies, encompassing facets like gross revenue, genre trends, actor influence, and directorial impact.

Related Work

This work was primarily inspired by personal experiences with reviewing revenue numbers of films over the years, and the desire to explore a long-term view at the industry side of film using techniques learned in class. Academic papers on data visualization methodologies and techniques served as foundational references, offering insights into effective approaches for conveying complex information. Additionally, online platforms specializing in data-driven analyses of the film sector provided valuable benchmarks and inspiration for visual representation. We also drew inspiration from visualizations discussed in relevant classes, extracting key principles and strategies for conveying trends and patterns in dynamic datasets. These diverse sources collectively contributed to shaping the conceptual framework and design considerations for the interactive data dashboard project.

Questions

Our project's primary objective was to create an interactive dashboard that provides a comprehensive overview of the movie industry. For each question posed, we formalized a number of potential visualizations. We aimed to achieve this by addressing the following questions:

1. How has the gross revenue of the film industry changed over time with regards to genre, actor, or director?

It was our belief that this curiosity motivated the entire project, stemming from our primary objective (**Fig. 1A**). However, as the project progressed, we found that the general trend of revenue was directly related to the number of films released (**Fig. 1D**). Therefore we pivoted towards the number of films released over time, allowing the following visualizations to be motivated by revenue.

2. How does the profitability of a film relate to the viewer ratings of the film?

Profitability could be considered the underlying driver of modern cinema[1]. Therefore, it seemed apt to explore how ratings, a subjective metric on how professionals and the general public favor a film, impacts the profitability of a film (Fig. 2A). During our exploration however, we found that ratings were influenced by a myriad of factors. We thus thought to explore whether there was a disparity between gender in ratings and the revenue drawn in, as that has been a major focus in films for the better part of the 21st century[2]. While there was considerable difference in the revenue brought in by male leading actors to that of female leading actresses, it seemed less influential than genre, which seemed a critical factor towards ratings (Fig. 2B). We thus posited that genre was a more influential factor to explore in relation to ratings and revenue. This left only an understanding of the correlation between revenue and budget to have a firm understanding of the economics of the film industry.

3. What qualitative factors affect the profitability of a film?

Here we posited that the correlation between budget and revenue would describe profitability of the film industry, and decided that the impact factors driving the prior questions would be a better investigative point(Fig. 3A). While this investigation proved largely correct, we further wanted to explore the genre-based impact on profitability, to determine whether general aggregates within the data might be misleading towards the actual trends culminating in the profit for each genre.

Data

Our data was entirely derived and downloaded from The Movies Dataset on Kaggle.com[3]. The dataset comprises metadata for 45,000 movies listed in the Full MovieLens Dataset, encompassing information such as cast, crew details, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts, and vote averages. The dataset also includes 26 million ratings from 270,000 users for the same set of movies, obtained from the official GroupLens website, with ratings on a scale of 1-5. The dataset is organized into several CSV files, including movies_metadata.csv (main movies metadata), keywords.csv (plot keywords), credits.csv (cast and crew information), links.csv (TMDB and IMDB IDs), links_small.csv (subset of IDs for a smaller dataset), and ratings_small.csv (subset of ratings from 700 users on 9,000 movies).

A substantial amount of cleaning needed to be performed for the visualizations answering questions 1 and 2. Foremost, there were a number of redundant entries, as well as missing, incomplete, and improperly filed entries. Each visualization handled data cleaning and preparation separately. For visualization 1, the film-level granularity in movies_metadata.csv file was suitable for the visualization needs, but several features had to be modified to suit the needs of the year-on-year evaluation. Since genre is the primary focus of filtering on our dashboard, it was necessary to reformat the genre field from a stringified JSON to a simple string list of comma separated genres. In this form, the genres could simply be parsed using a string splitting function in the JS file. As well, a small subset of films had a different release date format, which had to be reformatted to conform to the standard JS date format. After reformatting the movies_metadata.csv file, the data was loaded in JS code, and using D3, grouped by movie count by genre per year into a map table. Finally, the data was reformatted from a map to a standard JS list, which facilitated using it in many D3 functions.

To answer the second question, a number of variables needed to be cut and multiple datasets needed to be conjoined. Due to the disjointed nature between the metadata surrounding each film and the ratings, the movies_metadata.csv and ratings.csv files had to be conjoined by the shared identifier each dataset contained for every individual entry. Next, all columns not pertaining to genre, gender, rating, revenue, and unique identifier were excised from the dataframe. Following this, redundant entries were dropped on the basis of non-unique identifier numbers within the dataset. The data was then grouped according to rating, leaving a table of rating, multiple genre, gender, and revenue. The genre column consisted of multiple genres for each entry, therefore only the leading genre was used as the primary genre for the film. This same approach was taken for gender as well. Finally, the average revenue for each

rating was calculated and the table was transposed to give each rating as an entry, columns of genre, their respective revenue, and gender. Thus providing the final dataset to answer question 2.

Finally, data for the third question was processed in a similar manner to question 1. The film-level granularity of movies_metedata.csv was useful for accessing the name of every movie when it is hovered individually on the plot, but some genre wide averages were necessary for the unfiltered visualization. As before, each film's genres were converted from JSON to a comma separated string list, then filtered using a JS string-splitting function. Then, average budget and revenue were calculated and joined over each genre via D3 for use in the unfiltered visualization. In addition, a subset of the entire dataset was stored for each genre, which is accessed when the visualization is filtered.

Exploratory Data Analysis

Initial data analysis was conducted in Tableau to generate proof of concept visuals aimed at answering these questions. To answer question one, three preliminary visualizations were generated: a line graph plotting the revenue over the years (**Fig 1A)**, a scatter plot of revenue versus year with a trend line (**Fig1B)**, and a bar graph displaying the revenue for each year (**Fig 1C)**. While revenue over time was insightful, it was correlated with the number of films that were released. Since the following questions were heavily concerned with profitability and revenue, we decided to pivot towards the number of films released to allow revenue and other factors to be explored by downstream exploratory analysis (**Fig1D)**.

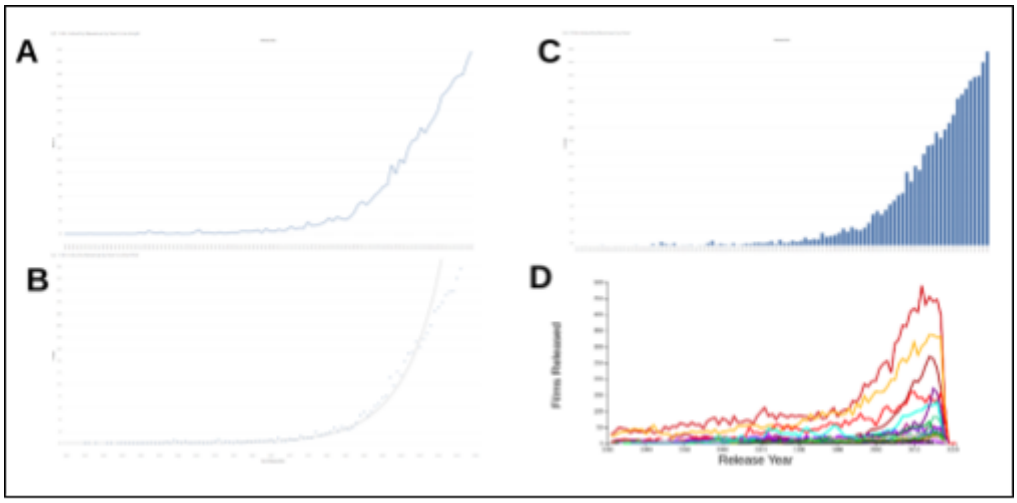

Figure 1: A. A line graph of revenue by year. **B.** A scatter plot of revenue vs year, with a trend line. **C.** A bar chart of revenue by year. **D.** The final figure, displaying films released by year, further sorted by genre (line color).

To answer the question "how does profitability of a film relate to the film's ratings?" We explored three different preliminary visuals. The first was a bubble plot where revenue (y-axis) and budget (x-axis) are displayed; the size of each bubble is related to the viewer ratings (**Fig. 2A)**. We hoped that visualizing this association would be clearly indicative of ratings association with profitability, however, the resulting plot was chaotic and poorly displayed the intended association. The second plot we created is a histogram that shows how many profitable movies are in each rating "star" category (**Fig. 2B)**. However, this proved unfruitful, as there was trivial difference in the bars of the histogram. Having to explore granular visualizations would diminish impact. In tandem, we explored visualizing the question in terms of a scatterplot of revenue (y-axis) plotted against budget (x-axis) (**Fig. 2C**). Here, the plot left little to explore in terms of dimensionality. We finally evaluated a barchart, displaying rating (x-axis) versus average revenue as an aggregate (y-axis) (**Fig. 2D**). This allowed a distinction between the ratings and revenue which we felt enabled an interpretation of the data. Further, we were able to explore further dimensionality through stacking genres into each bar, so the average revenue at each rating was visualizable.

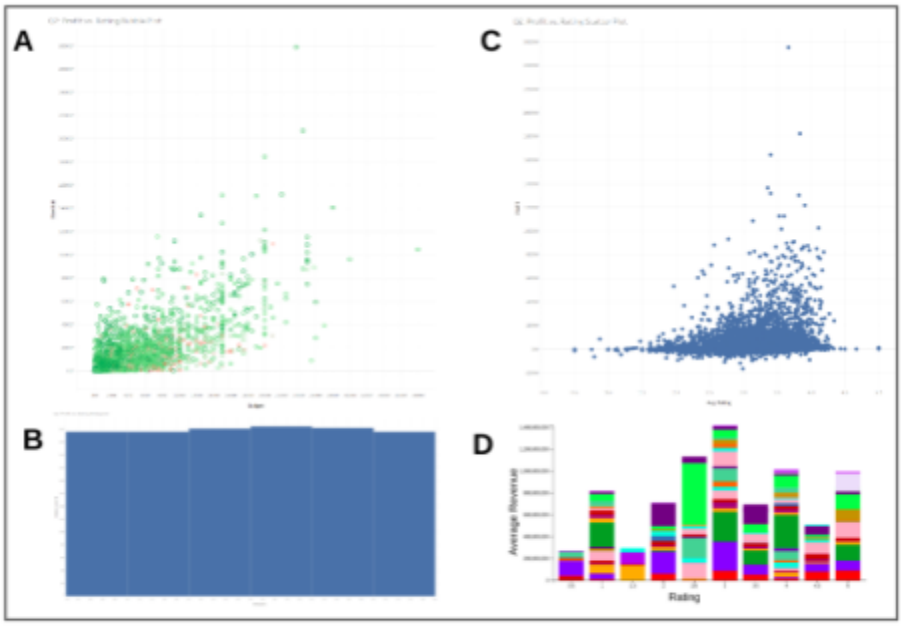

Figure 2: A. A bubble plot of revenue vs budget, with the size of the marks related to film viewer ratings. **B.** A histogram of rating "stars" (0-1 stars, 0-2 stars, etc.) with a count of profitable films (films with a revenue/budget ratio of at least 2). **C.** A scatter plot of profit (revenue-budget) vs viewer ratings of a film.

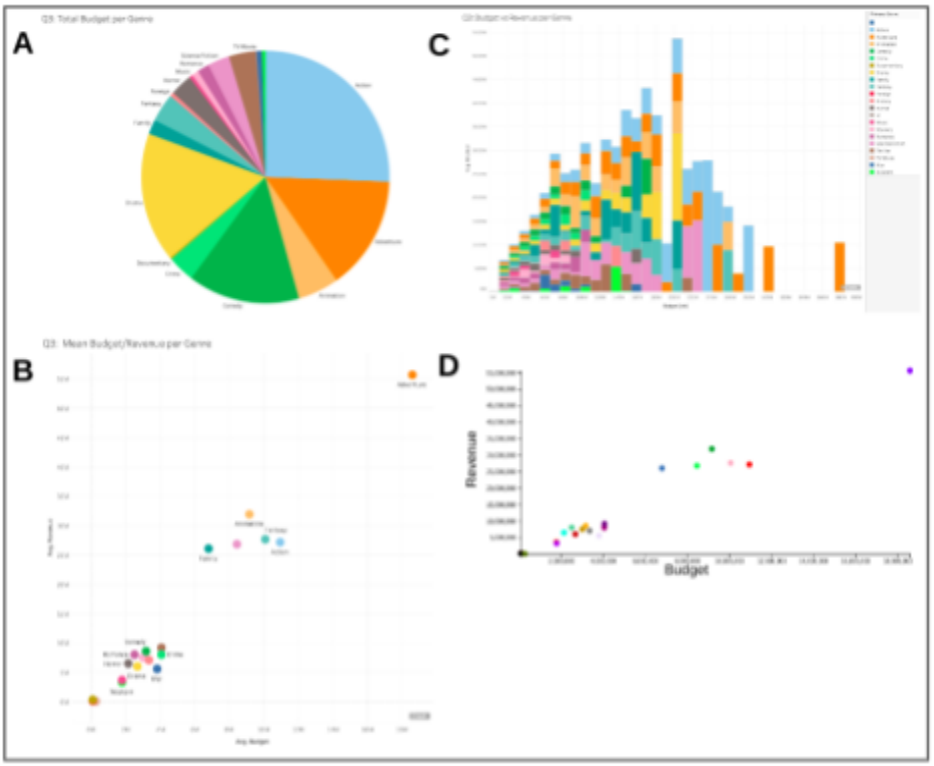

Figure 3: A. A pie chart of total budget and revenue per genre. **B.** A plot of average revenue vs average budget per genre. **C.** A histogram of average revenue vs budget per genre. **D.** A scatterplot displaying budget versus revenue, further categorized by genre.

Finally, our last visualization was primarily focusing on addressing profitability as it relates to genre. Our first inclination was to use a pie chart to display each genre's budget (**Fig, 3A**), however, this took away from the scale of revenue and genre's influence. In tandem, we generated a scatterplot displaying average revenue versus average budget (**Fig. 3B**), along with a stacked bar chart displaying the same comparison (**Fig. 3C**). However, the latter was largely redundant, displaying congruent characteristics of the stacked barchart answering question 2. Therefore, we moved forward with the scatterplot (**Fig. 3D**). Further categorizing gender here allowed us to later categorize each genre into separate scatter plots for further analysis.

Evolution & Implementation

Genre became a central theme in regards to budget, revenue, number of films, and ratings. Therefore, we aimed at making this a fixture point among all of the visuals we displayed. Each visual has an element of genre, denoted by a shared color palate for each visual (**Fig 4B-D**). For simplicity and clarity, the legend is in a grid pattern (**Fig. 4A**). While the legend itself is meant as a guide to the data, clicking on any legend point, as a button, will filter the dashboard to display visuals for that genre (**Fig. E**). A functional clear button has been implemented to allow a user to return to the

aggregated overview visualizations first present on loading the dashboard.

The first visualization displays 12 overlaid line charts with each line corresponding to a genre (**Fig. 4B**). Hovering over any line displays the genre associated with that particular line chart(**Fig. 5A**). Further clicking the visual filters each plot displayed to represent its intended topic for the genre selected, and scales the axis accordingly to fill the whole graph. (**Fig 5B**). Originally, this plot displayed average revenue compared to release year, but this became redundant with other visuals, and was directly correlated with films released per year. To create diversity in the information displayed in our visuals, we opted to change this visualization to films released per year. The intention behind this visualization was to prompt curiosity in the user, to notice chronological patterns in the film industry and investigate them for themselves.

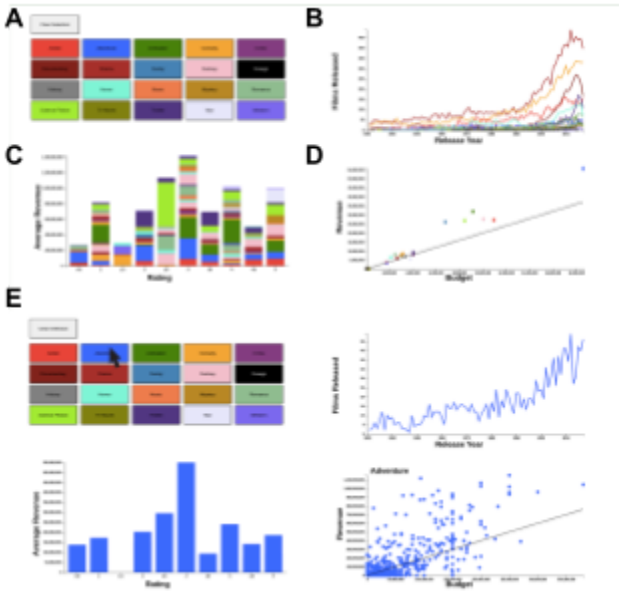

Figure 4: A. Interactable legend displaying each genre as a filterable element. **B.** Overlaid line chart displaying the number of films released per year by genre. **C.** Stacked bar chart displaying the average revenue compared to film ratings on a 5-point scale. Individual stacks correspond to average revenue for the specific genre at a particular rating. **D.** A scatter plot describing budget versus revenue of each film genre. A line of best fit describes a positive trend. **E.** Clicking on a legend item filters the dashboard for that specific genre.

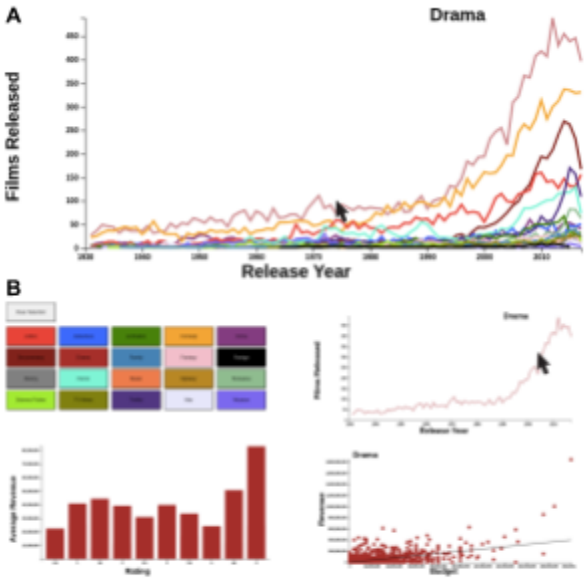

Figure 5: A. A mouse cursor hovering over the a line chart. A genre label rests in the upper right corner. **B.** Further clicking the line chart displays a visual for each genre.

For the second visualization, it was highly evident that a simple aggregate of ratings wasn't indicative of the entire industry. Some genre types are rated higher than others on average, while other genres are not rated highly ever. Therefore, the barchart was made so that each stack pertains to a genre (**Fig. 6A**). Hovering over any element in the stacked barchart will give the revenue garnered for that genre at that particular rating. For example, **Figure 6A** demonstrates that hovering over a stack, such as "Science Fiction" at the 2.5 rating scale (out of 5) displays an average revenue of all science fiction movies that received a 2.5 star rating; 512 million. Further, by clicking on the bar, the dashboard will filter all visualizations to the specific genre of interest.

For our last visualization, we directly answer how genre impacts the profitability of a film (**Fig. 7A**). Each genre is plotted as a dot within a scatter plot, where average budget is displayed against average revenue. Further, a trend line denotes the relationship between revenue and budget. Hovering over any genre displays the specific genre's name, as well as increases the size of the dot. Once clicked, the dashboard filters to the data of individual movies of that specific genre (**Fig. 7B**). Further, hovering over a point on the visualization once a genre is selected displays the film that dot corresponds to.

Evaluation

Our initial belief was that the project's core motivation was to understand how gross revenue in the film industry changed over time concerning genres, actors, or directors. However, as the project progressed, we identified a strong correlation between the general trend of revenue and the number of films released. Consequently, our visualizations shifted towards exploring the number of films released over time. This allowed us to examine revenue alongside other factors, providing a more nuanced understanding of the industry's dynamics.

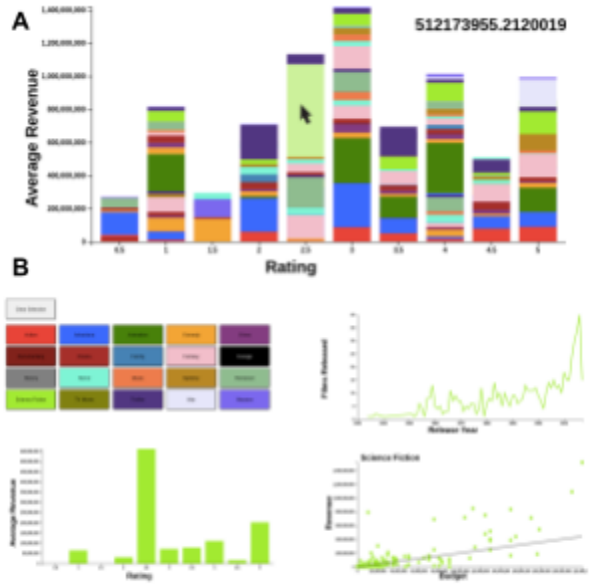

Figure 6: A. A mouse cursor hovers over a stacked bar within a barchart. The average revenue for that genre at a particular rating is displayed in the top-right corner. **B.** Clicking on a selected bar filters the total dashboard to display data pertaining to the genre of interest.

A large takeaway from the dashboard was that profit is highly correlated with genre. Analyzing **Figure 7A** shows that the general trend of revenue is positively correlated with budget. Yet, when filtering each genre, the results vary drastically on budget and revenue spent. For instance, filtering by "action" yields films exceeding 2.8 billion in revenue, whereas lower grossing genres, such as "foreign", has a maximum revenue of 22 million.

However, this trend does not extend to the number of films released over the years. The highest grossing genres (adventure and action) only produced upwards of 210 films in their combined peak years; a number more than doubled by the "Drama" category (**Fig. 5A**). While a small few films within drama do gross revenue close to that of action, (Namely, "Titanic"), the average revenue grossed by the drama category is only 5 million.

What might account for the large difference in the number of films produced in the highest earning categories? One potential explanation could be that the budget spent might be considerably lower, which would increase the profit received from these films. To reinforce this idea, **Figure 8B** displays the average films released per year under the drama category. At its peak year 450 films were released under drama. If we look at **figure 8C**, we can see the average revenue, regardless of rating, was above 20 million dollars. Using **Figure 6B** as a comparison, we can see that while there are fewer films released in the category, they make a larger revenue, while costing minimal budget to produce.

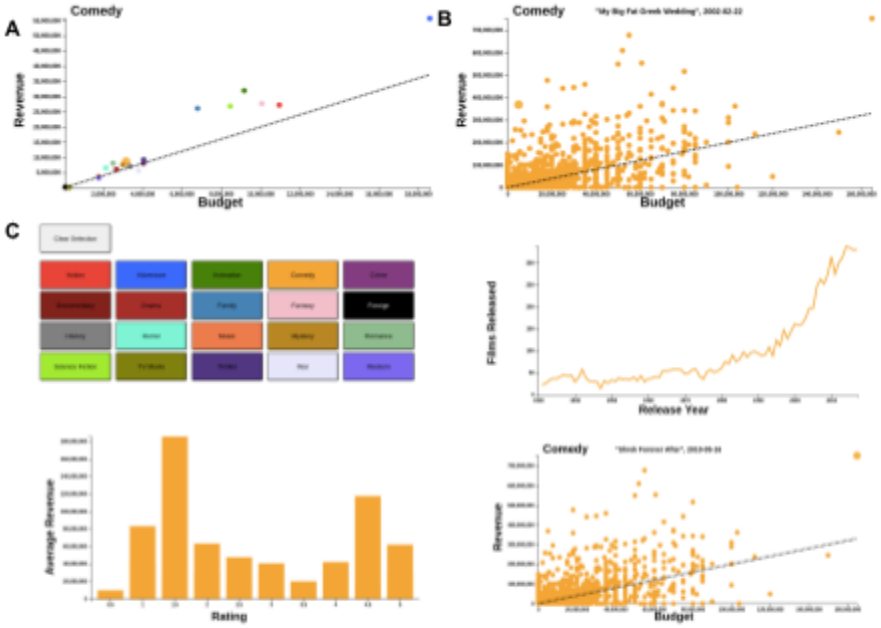

Figure 7: A. A scatter plot displaying various genre positioned by their budget versus revenue. **B.** Clicking any attribute displays the distribution of individual films per that attributes genre. Further, the trend line is updated to reflect the filtered data. Hovering over a specific point within the filtered distribution displays the associated film's title. **C.** Filtering by clicking on the element filters the entire dashboard to display a specific genres information.

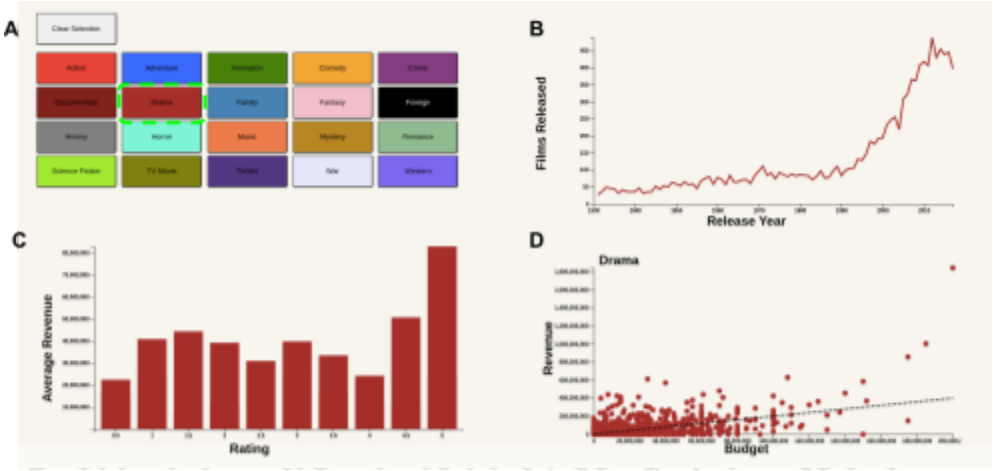

Figure 8: A. A green box drawn around the Drama selector, indicating its selection. **B.** Drama films released per year. **C.** Barchart of average revenue for each rating (out of 5 stars). **D.** Scatter plot of films labeled as "Drama and the correlation between budget and revenue.

This analysis provides a clear narrative of economic drivers in the film industry. Inherently, this industry is driven by profit, which consists of making a higher revenue than the budget spent. Where films cost an exorbitant amount to produce, often we see few films released within the year, and a higher box office revenue. However, if the revenue for a film genre is lower, there is a compensatory action, where more films are released at a lower budget to produce. Moreover, despite ratings being the voice and opinion of the audience, they seemingly have little impact on the revenue grossed by a film. This could be because opinion is subjective, making running a business model on subjectivity alone difficult. This drives a concluding ideology that the myriad of factors controlling a profitable film actually depends on production budgets significantly, rather than subjectivity of the audience.

References:

1. "Driving Economic Growth." *Motion Picture Association*, 23 Jan. 2023, www.motionpictures.org/what-we-do/driving-economic-growth.
2. "2020 Film: Historic Gender Parity in Family Films." *Geena Davis Institute*, 29 Sept. 2023, seejane.org/research-informs-empowers/2020-film-historic-gender-parity-in-family-films/.
3. "2020 Film: Historic Gender Parity in Family Films." *Geena Davis Institute*, 29 Sept. 2023, seejane.org/research-informs-empowers/2020-film-historic-gender-parity-in-family-films/.