COMPSCI 571

Process Book @ Final Phase

Soniya Gaikwad	Manan Parikh	Shiva Sravani Mudiyanur
32927398	34968148	34965596
sgaikwad@umass.edu	mananrajeshb@umass.edu	shivasravani@umass.edu

Project Title: Montage - Exploring Movie Trends Over Time

GitHub Repository: https://github.com/soniyagaikwad/movie-trends-data-vis

1 Overview & Motivation

Our team decided to pursue "Montage: Exploring Movie Trends Over Time" because of our love and interest in entertainment, especially movies. Also, some of us have projects revolving around media, so we wanted to continue looking into this industry and its crossover with data visualization.

2 Related Work

Some related works that inspired us to work on "Montage: Exploring Movie Trends Over Time" include the visualization from Rotten Tomatoes from Homework 1 and the PokeData Final Project from the University of Utah shown in class. We wanted to combine aspects we liked from both visualizations to create a data visualization website for movie-related data.

3 Project Objectives and Questions

The primary questions we aimed to answer with our visualization include:

- While looking at a specific (movie/genre/MPAA rating/budget), how much revenue was generated?
- What are the relationships between the MPAA ratings, budget, gross revenue, release dates, genres, runtimes, and ratings the movies have received?
- What are the highest and lowest revenue-generating movies?
- What are the highest and lowest-budget movies?
- What are the ratings for the movies depending on their genres?
- How does profitability vary across movie ratings?
- Which genre has the most interest over the years?

We would like to learn how these pieces of data related to movies correlate with one another and theorize how they potentially play out in the entertainment industry. Some benefits of learning and accomplishing these aspects include helping us and others understand what factors could be critical to the movie industry for potential box office success and how the public interacts with these movies based on these factors.

4 Data

We are using a dataset from Kaggle by Yashwanth Sharaff called “Movies Performance and Feature Statistics: Analyzing Box Office Performance, Rating and Audience Reactions” [1] to visualize our project.

https://www.kaggle.com/datasets/thedevastator/movies-performance-and-feature-statistics

5 Data Processing

In the dataset “Movies Performance and Feature Statistics: Analyzing Box Office Performance, Rating and Audience Reactions,” we gain access to a large amount of data related to movies, such as their titles, MPAA ratings, budget, gross revenue, release dates, genres, runtimes, ratings they have received, actor id, and many more. Based on our initial designs, we focus on each movie’s title, MPAA ratings, budget, gross revenue, genres, and summaries.

In terms of data processing and cleaning, we process it using Javascript. We start importing our csv file, which is kept inside the data folder. We use d3’s csv function to read the csv file, and only consider the first 510 rows as only they are required, then we consider these columns only to proceed further: title, MPAA rating, budget, gross revenue, genre, runtime, rating, and release date.

Moreover, there is a column named 'release date' in our dataset, which we don't intend to use in its entirety. Based on our brainstorming, we only extract the release year from it. In addition to the release year as a derived attribute, another derived attribute would be the profit, in which we take the revenue - budget.

To support our visualizations, we created specific utility functions in JavaScript. One such function, groupGrossByRating, groups movies based on their MPAA rating and collects their gross revenues into sorted sets. This helps in preparing data for boxplots, where revenue distribution by rating can be visualized more effectively.

Another function, getMoviesByRatingAndYear, filters the dataset based on a given MPAA rating and release year. It returns a simplified list of objects containing only the movie title and its gross revenue, making it useful for focused comparisons or detailed listings in visual outputs.

6 Exploratory Data Analysis

To begin exploring our data and planning our visualization tool, we created a basic table to display a list of movies along with their key details like title, MPAA rating, genre, release year, budget, revenue, profit, and rating. With these details, we decided to use a combination of simple bar charts, pie charts, scatter plots, and box plots.

For example, to visualize the budget and revenue of each film, we decided to use bar charts. To observe the different movie details and how each aspect potentially relates to another aspect, we decided to work with different graphs that fit the attributes appropriately. For example, while digging deeper into the data, we noticed there could be a potential pattern between MPAA ratings and gross revenue, which we haven’t explored yet in other data visualization tools. So, to explore and communicate this relationship clearly, we decided to use a box plot, which effectively shows how revenue varies across different rating categories.

However, we noticed that if we placed all these charts on a single webpage, it would look cluttered

and overwhelming. To solve this, we decided to create an interactive, dynamic interface where users can select their own X and Y axes, allowing them to generate the specific visualizations they're interested in. This makes the experience more user-friendly and customizable.

These insights helped shape our final design.

7 Design Evolution

At our Project Proposal phase, we considered the following design that integrated all the aspects we wanted to use from the alternate designs each member created.

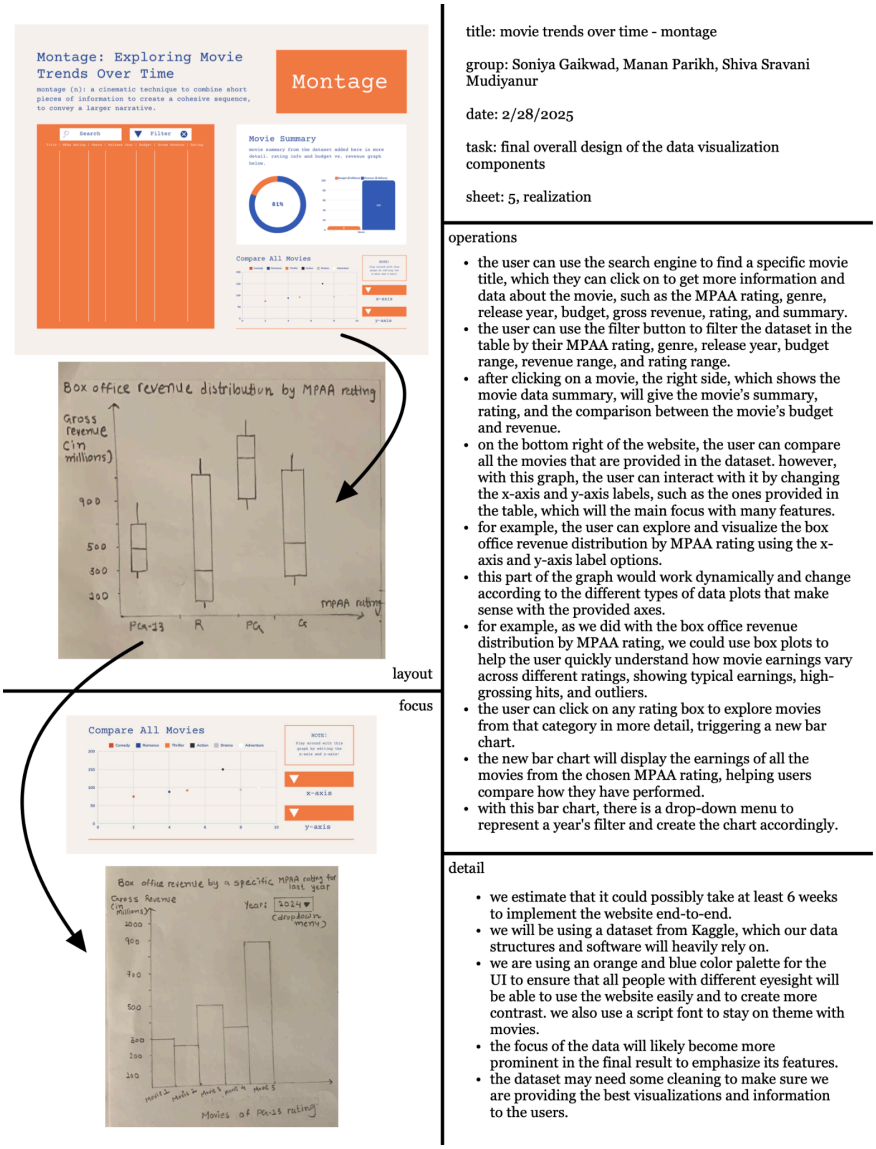

Figure 1: Sheet 5, Realization, of the 5 Design Sheet Methodology (/project-proposal)

The design decisions we made using the perceptual and design principles we've learned in class include the following important features:

- User-Friendly Interface
    - A clean, intuitive design with easy navigation.
    - As described in Figure 1 and seen in Figure 2, we are using an orange and blue palette for the UI to ensure the visualization tool is accessible to all users.
- Movie Search Functionality
    - Users can search for movies by name to retrieve relevant details.
- Movie Details Display
    - Essential information such as title, MPAA rating, budget, revenue, release date, genre, runtime, number of ratings, and summary should be presented.
- Filtering Options
    - Allow users to filter movies by genre, release year, MPAA rating, and ratings.
- Clickable Links
    - Users can click on a movie to get the information summary accordingly, such as the movie summary, rating percentage, budget, and gross revenue.
- Trends over Time
    - Graphs or visualizations showing revenue trends over time.

In addition to these important features, we had some other ideas we thought would be nice to have, but were not critical to our project such as:

- User Ratings & Reviews
    - Users can rate and review movies.
- Social Sharing
    - Allow users to share movie details on social media.
- Streaming Availability
    - Indicate where the movie is available for streaming (Netflix, Disney+, etc.).
- Movie Recommendations
    - Suggest similar movies based on selected titles or genres.

However, we had to deviate from these features due to the lack of quick and accessible data.

For the majority of our visualization tool design, we continued to stick to the design decisions we originally made; however, to improve the design, we decided to enlarge the “Compare All Movies” aspect, so that the data plots are easier to interact with and analyze, as we continue to allow the user to play around with the x-axis and y-axis with different attributes and plots that fit these attributes, as seen in Figure 2, as well as enlarge the “Movie Revenue Distribution by MPAA Rating For All Movies” aspect to reduce clutter and increase clarity.

8 Implementation

In Figure 2, we have our key design of our website, “Montage: Exploring Movie Trends Over Time.”

On the left side of the website in Figure 2, we have a table that will list all the movies in the dataset with their appropriate details, such as their title, MPAA rating, genre, release year, budget, gross revenue, profit, and rating. In this table, the user will have the ability to search for a specific movie or filter the movie based on their attributes.

If the user selects a specific movie in the table, the selected movie’s data summary will show on the window to the right side of the table on the website. In addition to the movie’s summary, there will be two plots shown; one of which provides the rating information using a circle percentage graph and another providing information about the budget and revenue using a bar chart.

Below the table and movie summary section, the user has the opportunity to compare all the movies in the dataset. With this graph, the user can interact with it by changing the x-axis and y-axis labels by using the data attributes from the table, such as their MPAA ratings, genres, release years, budgets, gross revenues, profits, and ratings, and hover over the data points to learn more information about the movie. This part of the graph works dynamically and changes according to the different types of data plots that make sense with the provided axes.

Below the “Compare All Movies” section, the user has the opportunity to observe the correlations between MPAA ratings and the movie revenue distribution for all movies. With this box plot, the user can interact with it by hovering over the boxes, and learn more about the quartile information and ranges for all movies in the dataset.

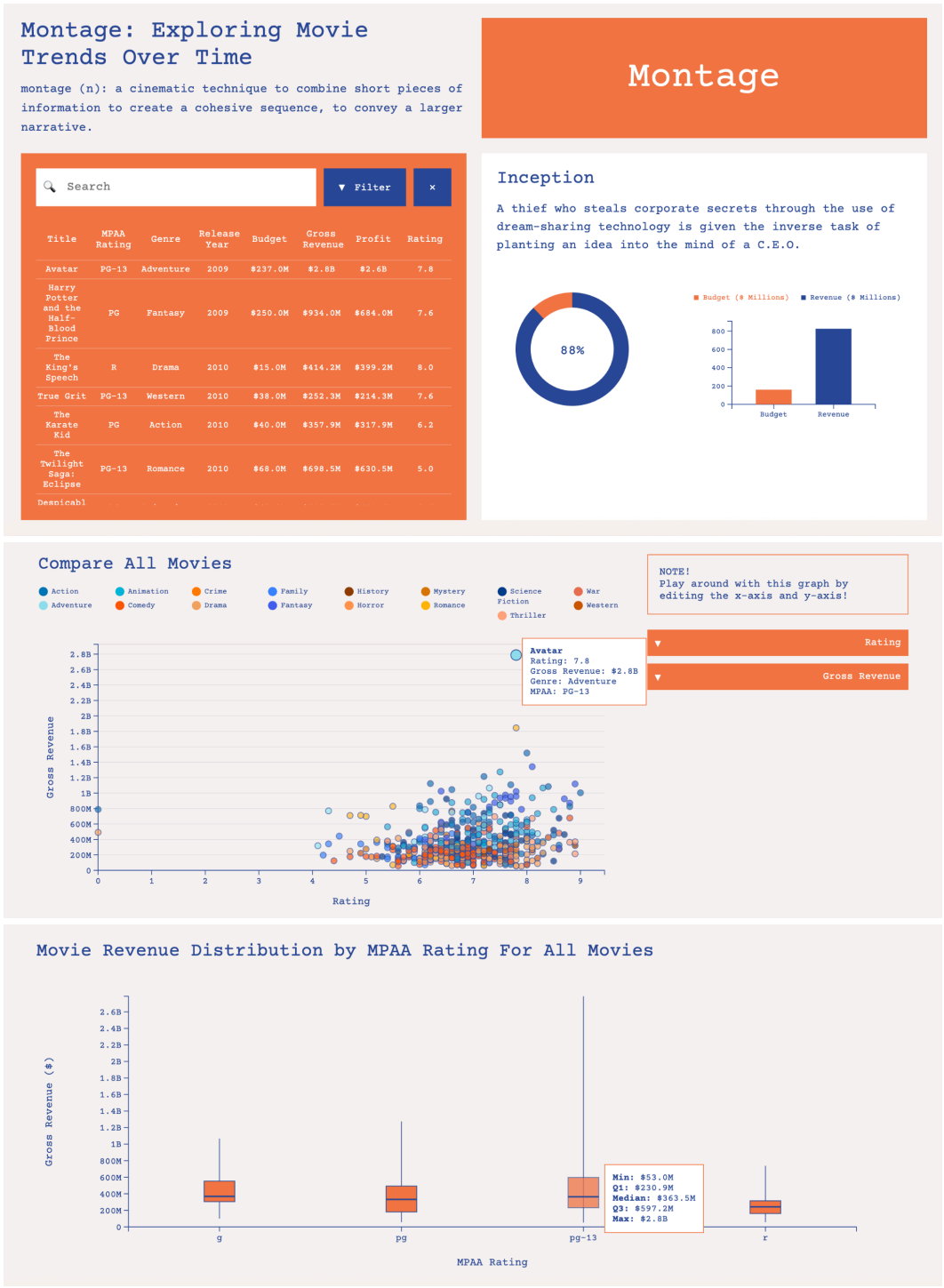

Figure 2: UI Implementation of the Visualization (montage.html and montage-design.css)

9 Evaluation

Table & Movie Summary

In the table, we list 510 movies from the dataset with their details, such as their title, MPAA rating, genre, release year, budget, gross revenue, profit, and rating. These details helped answer questions, such as "while looking at a specific movie, how much revenue or profit was generated?"

When the user selects a specific movie in the table, the selected movie's data summary shows two plots; one of which provides the rating information using a circle percentage graph, and another providing information about the budget and revenue using a bar chart.

As we observe in the table, circle percentage graph representing ratings, and scatter plot, the ratings typically range from 4.0 to 9.0, or also 40% to 90%, except for a few outliers that don't have a rating.

As we observe in the table, bar chart representing budget, revenue, and profit, and scatter plot, the budgets range from 60K to 380M dollars, the revenues range from 50M to 2.8B dollars, and the profits range from 18K to 2.6B dollars.

Some limitations that come with the table and movie summary section include not being able to find these ranges immediately. The user would have to actively observe the table and its values to verify these pieces of information, in comparison to using the scatter plot to compare all the movies in the dataset, especially since we have a large dataset of 510 movies, which can be overwhelming to look through.

Scatter Plot

The scatter plot visualization reveals several interesting patterns in the movie dataset like rating distribution across time, genre patterns, temporal trends. We can observe that there's a slightly higher concentration of highly rated films (7.5+) in the late 1990s and early 2000s and there's a noticeable increase in the diversity of genres over time.

The visualization helped to answer the research questions like the following:

- How do different genres perform in terms of ratings? By color-coding the dots by genre, we can see that certain genres (like drama) tend to receive higher critic ratings on average. The visualization allows for easy identification of outliers - exceptionally high or low-rated films within each genre.
- What are the relationships between MPAA ratings, budget, gross revenue, release dates, genres, runtimes, and ratings? The scatter plot directly answers this by allowing variable selection for both axes. Users can plot any two variables against each other (e.g., budget vs. revenue, rating vs. runtime). Using this we can answer many questions like, How does profitability vary across movie ratings? By setting the axes to Profit and Rating, this question is directly answered.
- Are there notable patterns in movie production over time? The clustering of certain genres in specific time periods reveals trends in film production, for example, more Science Fiction films in recent years.
- How do movie ratings correlate with release years? Visualization shows that highly-rated movies exist across all time periods, indicating that quality filmmaking isn't tied to a specific era.

Strengths of the visualization include interactive tooltips, flexible axis selection, and clear genre-based color coding. However, it faces limitations such as overplotting and the current implementation focuses on a few key variables (rating, year, budget, revenue) and more derived metrics like ROI (Return on Investment), rating-to-budget ratio, or popularity metrics can be

added. It can be difficult to find a specific movie within the scatter plot as there are many data points that may be stacked on top of each other.

Boxplot

Using the boxplot to visualize gross revenue distributions by MPAA rating revealed several key insights.

At first glance, PG-13 movies stood out with the highest maximum revenue. However, a closer analysis showed that G-rated movies actually had a slightly higher median gross revenue ($368.4M) than PG-13 movies ($363.5M), suggesting that G-rated films tend to perform more consistently at the box office.

Additionally, G-rated movies had the highest minimum gross revenue among all MPAA categories — meaning even the lowest-grossing G-rated film outperformed the lowest-grossing films in other categories. This indicates that G-rated content offers a more stable baseline of commercial success, likely due to its broad, family-friendly appeal.

These findings directly supported my research question about how profitability varies across movie ratings. The boxplot allowed me to go beyond surface-level observations and explore central tendencies and overall distribution, not just extreme values.

The boxplot worked well to visualize revenue spread, medians, and outliers across ratings. Interactive tooltips provided detailed insight on hover, helping users compare values with precision.

However, the visualization has limitations:

- No sample size shown: It's unclear how many movies are in each rating group, which could provide important context.
- No time filter: Trends may differ by year, but the current view aggregates all time periods.
- While this boxplot focused on revenue distribution across MPAA ratings, a similar approach could have been applied to other categorical variables — most notably, movie genres. Visualizing revenue distribution across genres would likely uncover equally compelling patterns, such as which genres tend to have more financial volatility, which have consistent median returns, and which rely heavily on a few blockbuster successes.

References

[1] TheDevastator. (2023). Movies performance and feature statistics: Analyzing Box Office Performance, Rating and Audience Reactions. Kaggle. https://www.kaggle.com/datasets/thedevastator/movies-performance-and-feature-statistics