IMDb Movie Insights

Basic Info: https://jbarne14.github.io/

David Huston-Hakey — djhusto@clemson.edu —C10473094

Jalen Barnes — jalen5@clemson.edu - C78236975

Nikki Chen — nchen3@g.clemson.edu - C51938622

Background and Motivation:

Movies are a big part of our culture. We decided to dive into the IMDb movie database to understand more about the world of movies. We were motivated to use a dataset that was relatable to everyone in our group. Movies are major sources of entertainment and revenue for people and countries around the world.

Project Objectives:

1. Explore the correlation between a movie's budget and its global gross income. Does a higher production budget generally equate to higher earnings?
2. Revenue Trend Over Time: Look into the trend of worldwide gross income over time
3. Analyze movie genre trends: Which genres are most popular, and has the film industry been focusing on particular genres in recent years?

Data: https://www.kaggle.com/datasets/chenyanglim/imdb-v2

Our data comes from a result collected from a research regarding IMDb, a popular movie database. This dataset gives us a lot of details about movies. It tells us the movie's name, when it was released, its genre, how long it is, which country it's from, the languages spoken in it, who directed and wrote it, who acted in it, a short description, its average vote, how many votes it got, its budget, how much it earned in the USA and worldwide, and user reviews.

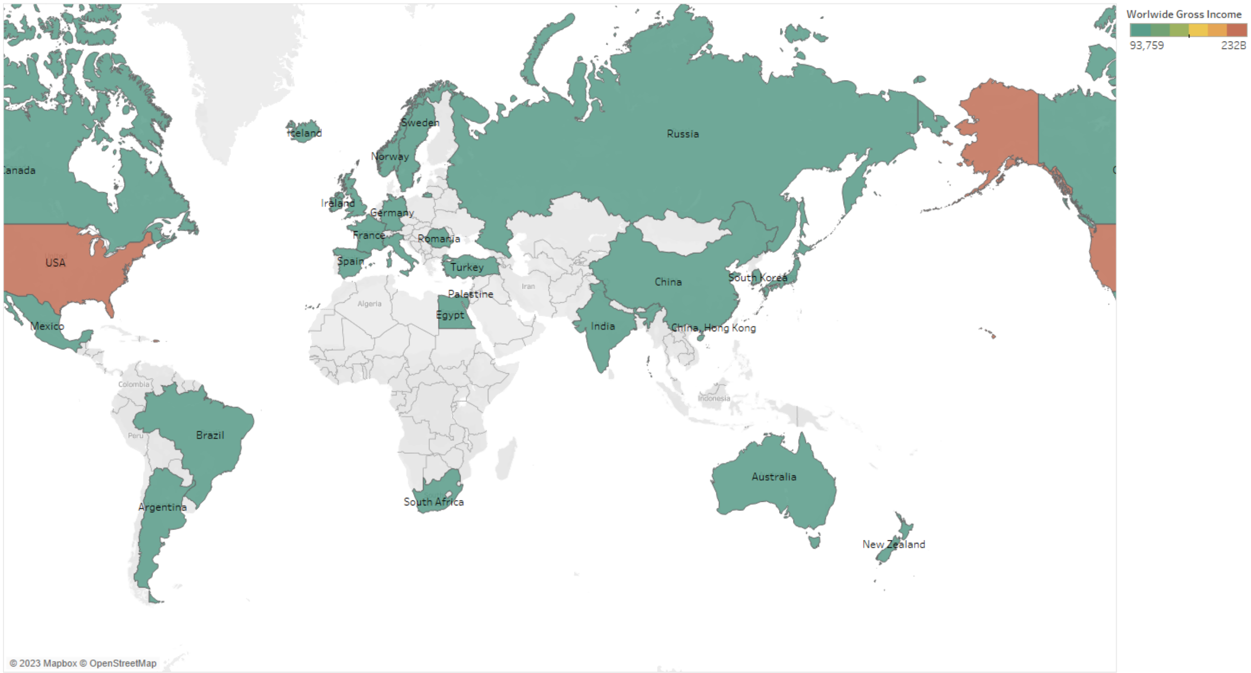
Data Processing:

The dataset in use is an enhanced version of the IMDb database, focusing on movies released from 2000 to 2020. Our dataset is a cleansed and modified version specifically tailored for a content-based recommendation engine. With a repository of 5,487 movies, the dataset encompasses a wide array of attributes, including but not limited to year of release, genre, duration, language, cast details, directors, ratings, and votes. Unique feature transformations, like "actors_f2" and "desc35," which represent the first two actors and the initial 35 characters from the movie description, respectively, have also been incorporated. It's noteworthy that the data has undergone rigorous rounds of error-checking, with several corrections being manually implemented as of May 2021 to address discrepancies, fill in missing details, and ensure the dataset's robustness. To enhance relevance for the recommendation engine, movies with fewer than 10,000 votes have been treated as less popular, leading to their exclusion from our dataset. Our objective in terms of data processing is to capitalize on the already cleaned dataset while ensuring minimal interference. This approach allows us to maintain an authentic representation of the data. Our core analyses will encompass correlations between movie budgets and revenues, evaluation of movie duration trends, and tracking the rise and fall in the popularity of movie genres over the two-decade period.

Visualization Design:

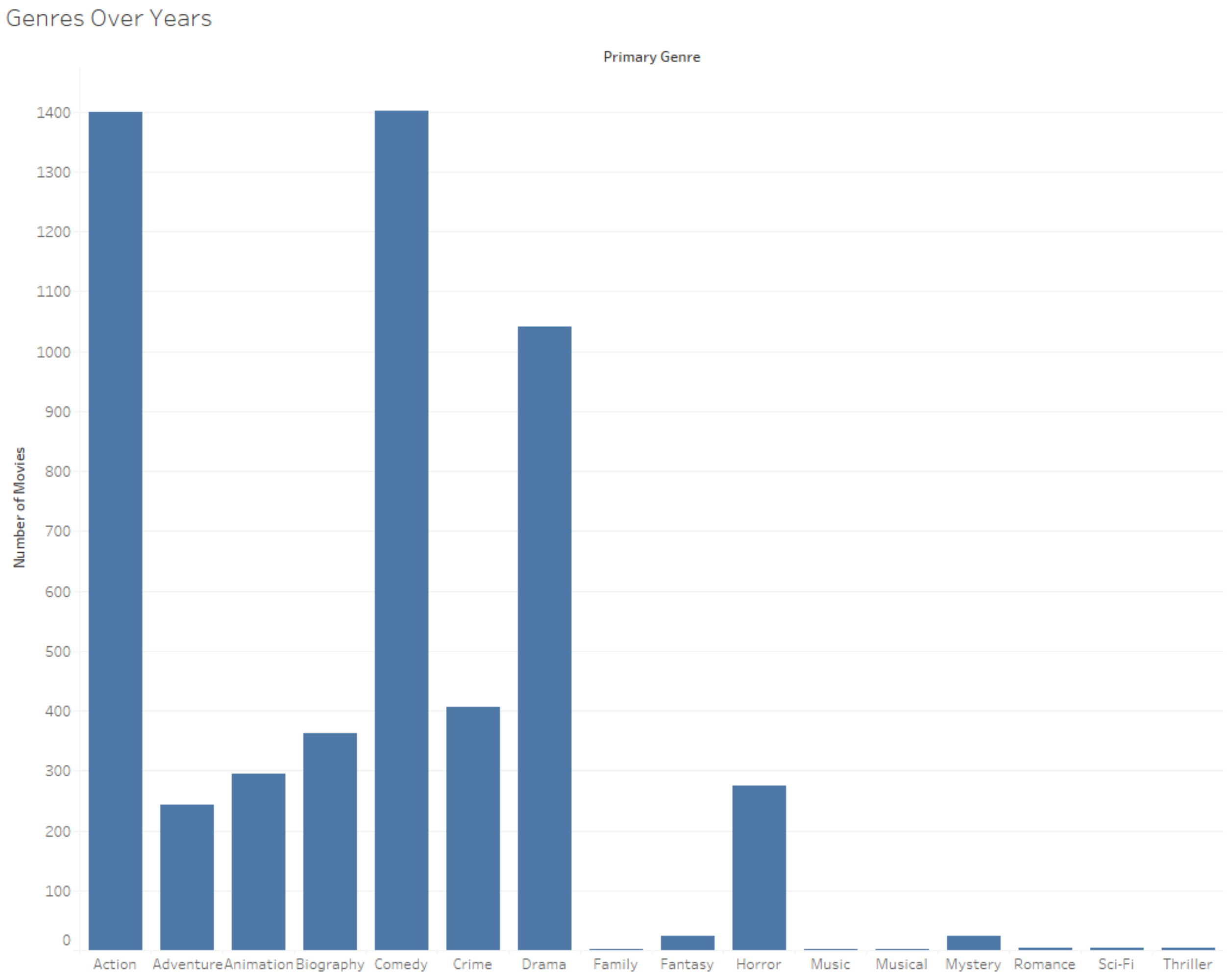
1. Budget vs Earnings: We'll use a world map to see if there's a connection between how much a movie costs (budget) and how much it earns (worlwide_gross_income). This will help us understand if spending more money on making a movie generally means it will earn more. We can also see how different movies thrive in different economic conditions. In countries like the U.S., we can afford to have large budget movies. The U.S. relies heavily on the movie and entertainment industry. We can also see a smaller economy like Egypt competes in the movie industry with smaller budgets. The graph below uses the sum of the budget to filter out countries and then the sum of the gross income to highlight which countries made the most with that specific budget range.

Worldwide Income and Budget for Movies around the World

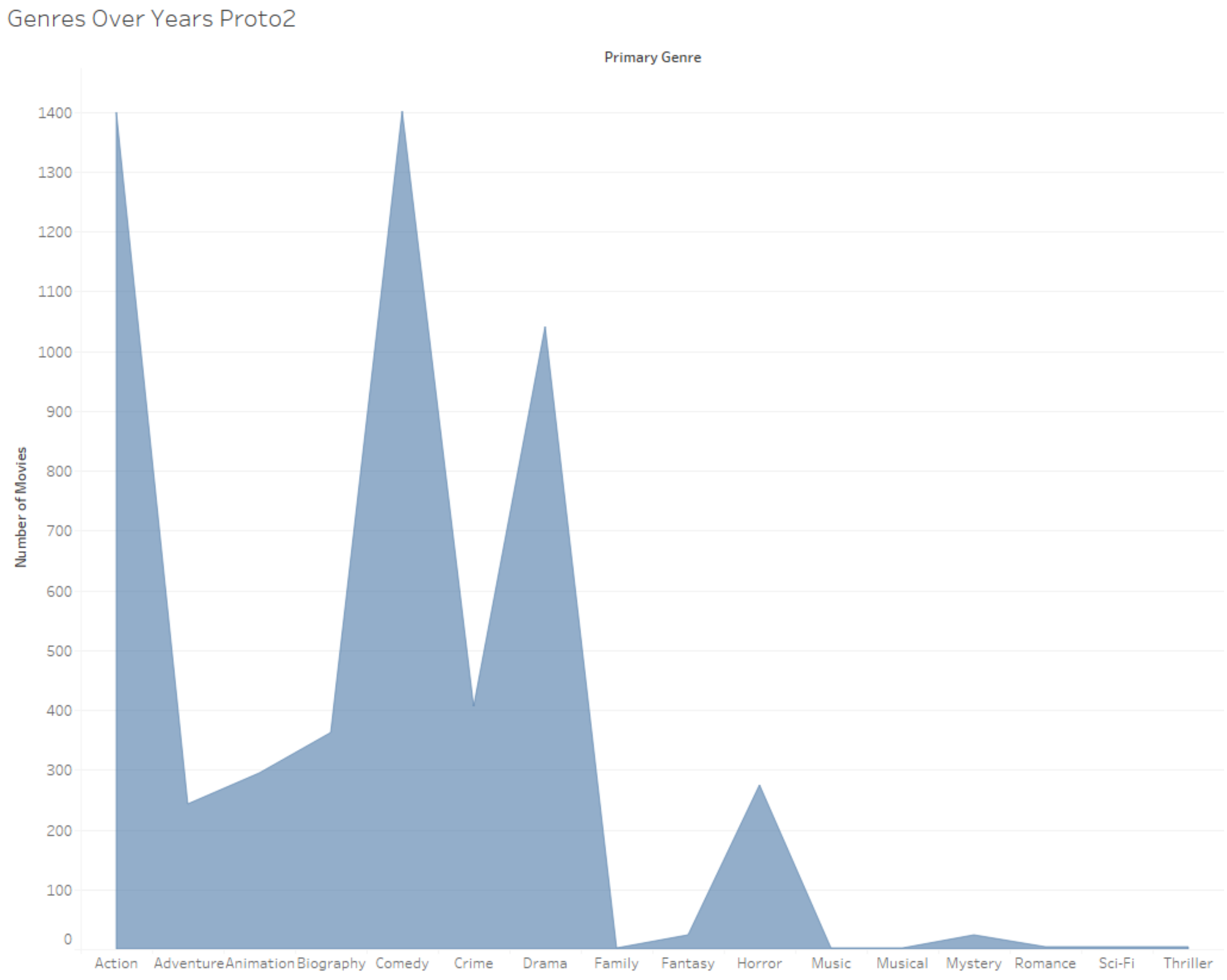


Map based on Longitude (generated) and Latitude (generated). Color shows sum of Worlwide Gross Income. The marks are labeled by Country. The data is filtered on sum of Budget Integer, which ranges from 230,000 to 75,515,313,000.

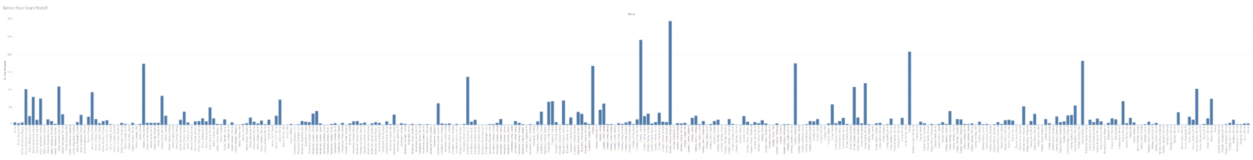
2. Revenue Trend Over Time: We'll create a line chart to show the trend of worldwide gross income over time. This will help us see if the movie industry's revenue is growing.
3. Genre Popularity: Movies come in many types or genres. We want to see how popular each movie type (genre) has been over the years (year). We'll use a bar chart with a slider scale for this. This will show us if some movie types are becoming more popular while others are less watched.



(Img1)



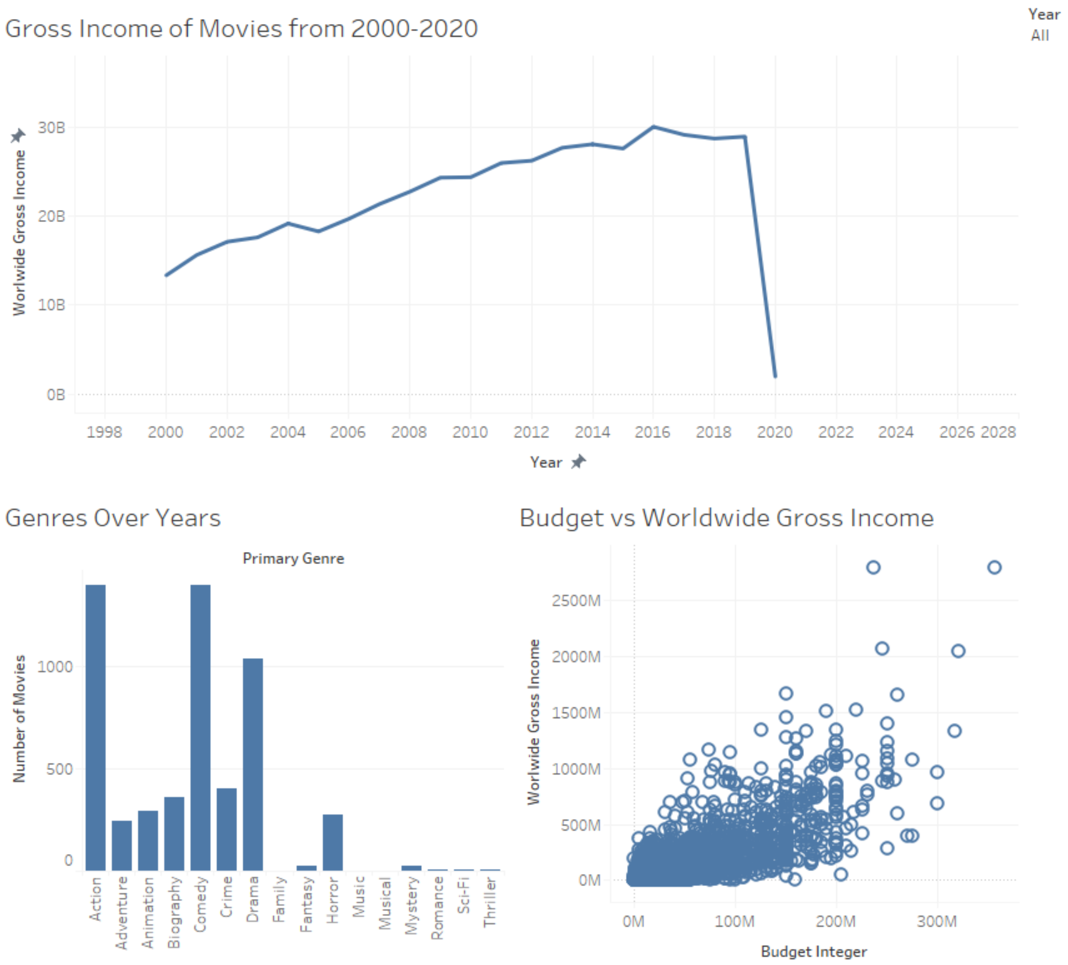
(Img2)



(Img3)

These are the three visualizations made in reference to genre popularity over the years. The item is the movie. The attributes associated with the movie are year and genre in this case. Year is a quantitative attribute and genre is a categorical attribute. I believe the best visualization is the first (Img1). The problem with the third one(Img3) is there is too much data and too many categorical attributes associated with the item and it makes the graph unreadable. The problem with the second one(Img2) is that it makes it seem like there is a continuous trend with the data that in reality, each movie is different and not continuously related.

Dashboard:



This is the current state of our dashboard with the year slider on the end so it will sort by year and give stats based on each year and associated attributes.

Must-Have Features:

- Our above visualizations
- Interactive Filters:

- Genre Selector: A dropdown menu enables users to concentrate on one or multiple genres, refining the data to showcase only their chosen categories. This function can significantly change the landscape of the displayed visualizations, catering to individual genre interests.
- Popularity Threshold Slider: With this feature, users can set a vote threshold, ensuring the dashboard visualizes only movies that surpass the selected vote count. This allows for a balance between sheer volume and perceived quality or popularity.
- Year Range Slider: This adaptable slider lets users focus on movies released within specific years or spans, updating all visualizations dynamically based on the selected range.

Optional Features:

- Information about the salaries writers, produces, and actors

Project Schedule:

Task	Complete Date	Contributor
Set up the github page	09/30	Jalen Barnes
Data Acquisition + Cleaning	10/05	David Huston-Hakey
Visualization Prototyping	10/10	Nikki Chen
Website Integration / Project Prototype	11/05	All Members
Refinement & Feedback	11/19	All Members
Final Presentation Prep	11/19 - 12/5	All Members