Data-Driven Movie Business Decisions

By Alondra Castro-Valadez, Jingyi Zhu, Fanny Cao

Introduction

Movies are an investment. From production costs, actors' million dollar salaries, and marketing campaigns, producing a movie can cost anything from $10,000 for an indie-film and up to $400,000,000 for a Hollywood blockbuster. Therefore, for business stakeholders', identifying any factors that could potentially influence their production strategies is crucial and could lead to millions of dollars in saving cts. In our project, we will be using data analysis and visualization methods to find trends, identify relationships, and compare data with an end goal of identifying areas of potential investment and cost savings in the movie industry. Some examples of our areas of observation are: exploring correlations across different types of production factors vs revenue, comparing profitability across genres, and identifying any seasonal trends in the industry. We will be using online data which will require initial cleaning and processing to ensure accuracy and efficiency of our analysis and visualizations. Overall, our project will consist of using various data visualization methods, identifying any controllable factors in the movie production process that could possibly affect the success of a film.

Two reference papers that maybe related to your topic (10 pts)

Paper 1: Paul, C., & Das, P. K. (2022). Predicting movie revenue before committing significant investments. *Journal of Media Economics*, *34*(2), 63–90. https://doi.org/10.1080/08997764.2022.2066108

The authors of this paper aim to study how different factors in movie production can influence an investor's financial risk in investing in that movie. In this study, they made a predictive model to estimate the amount of revenue a specific movie will make based on the cast, crew, genre, and distinctive characteristics of the movie.

Paper 2: Wisnefsky, Zachary, "Breaking Down the Box Office: An Analysis of Film Profitability Factors" (2023). Honors Scholar Theses. 968. https://opencommons.uconn.edu/srhonors_theses/968

In this paper, Zachary Winefsky, a student at the University of Connecticut, analyzes ~2000 films from 2007-2019 to find patterns between qualitative film factors and their possible impacts on profitability. Wifensky developed three hypotheses to test and used the empirical model to test each.

Main Dataset

The dataset we used is called **TMDB 5000 Movie Dataset** which contains metadata on 4800 movies from Kaggle. This dataset contains information from The Movie Database (TMDb) API. This dataset contains 20 features for data like budget, revenues, genres, popularity, production company, which includes both categorical (genre) and continuous (budget) data. We will be using this to analyze how

different features of movie production might contribute to the overall success/returns of a movie. Since it is in a csv file, we will be loading it using Pandas for cleaning and manipulating for data visualization.

Link for the dataset: [Movie_Dataset_Project]

Two static images related to your topic

Plan and tasks for later analysis

Plans: Preparation work with dataset

Initial data processing and cleaning if there is any missing values in critical columns like budget, revenue, genres, release_date

- Merging the two csv files (movies and credits) to add crew and cast information to the dataset
- Converting data types like release_date to datetime and budget and revenue to numeric
- Converting currencies for budget and revenue to USD for comparability
- Converting columns with dictionaries into dataframes
- Calculating the statistics we will use for comparisons: ROI (return of income), rates, etc.
- Filter for movies released after 1990 to make the dataset more reliable and up-to-date

Tasks for analysis

1. Explore correlations between production factors (budget, cast size, company, genre) and financial outcomes (Revenue and ROI)
2. Identify seasonal or temporal trends in movie releases
3. Compare profitability across different genres and production companies
4. Examine how the power of cast and crew can influence on the revenue generation

Planned Data Visualizations (Static and Interactive)

Tools and libraries we are going to use: Altair, D3, Seaborn

1. **Scatter plot with diagonal reference line (break-even point analysis) - D3 interactive**
    - **Purpose:** to show the investment and return relationship and what are some of the successful movies from the business perspective
    - **Glance of the figure:** Each movie is represented by a single point, with the x-axis being the budget, y-axis being the revenue, and color representing the genre. The size of the

point will depend on the popularity of the movie. There will be a line showing the break-even point (where X=Y, means budget equals to revenue)

2. **Side-by-side Boxplot: ROI comparison across genres - Altair**
    - **Purpose:** To help the investors choose the potential successful genres which have the top level of return of income based on this visualization
    - **Glance of the figure:** Graph will display median, quartiles and outliers for each genre. The axis will be: movie genre on the x-axis and ROI percentage on the y-axis. The genre color will be consistent to the color in scatterplot.

3. **Release time scroller (Line graph):** movie release patterns by year and month to identify seasonal trends (users can specify/adjust the time)
    - **Purpose:** Line graph that will allow users to identify seasonal release trends to help aid their movie release strategies (when to release film in theatres).
    - **Glance of the figure:** The x-axis is the date and y-axis is the number of movies released.

4. **Scatterplot:** Potential relationships between cast size and revenue (different size per point)
    - **Purpose:** Analyze if a larger cast size leads to a higher revenue which will help stakeholders determine whether to invest on a larger cast size and determine price distributions.
    - **Glance of the figure:** The x-axis is the cast-size and the y-axis is the revenue. Each movie will be plotted as a point. The color of the point will depend on cast size to help visualize any existing relationships relative to cast size.

5. **Grouped Bar chart (D3):** the average budget by major studios across different countries
    - **Purpose:** Compare average budget across different groups (leading movie companies) ordered by country. This will help compare budget spend across firms, informing business stakeholders on whether they should spend more to remain competitive.
    - **Glance of the figure:** The x axis on the figure will be Category (country) and within those categories we will have groups that have a different color (movie company). The Y axis will be an average budget (in millions USD).

Describe each group member's duties.

All the team members will rotate responsibilities throughout the project to ensure equal distribution:

- All members will collaborate on data cleaning and the pre-processing phase.
- Each member will develop at least one visualization and work on the website and dashboard together.
- All members will contribute to analysis, interpretation, future report and presentation slides.
- Every decision-making and problem-solving will be handled collectively through regular group meetings.