Milestone 1

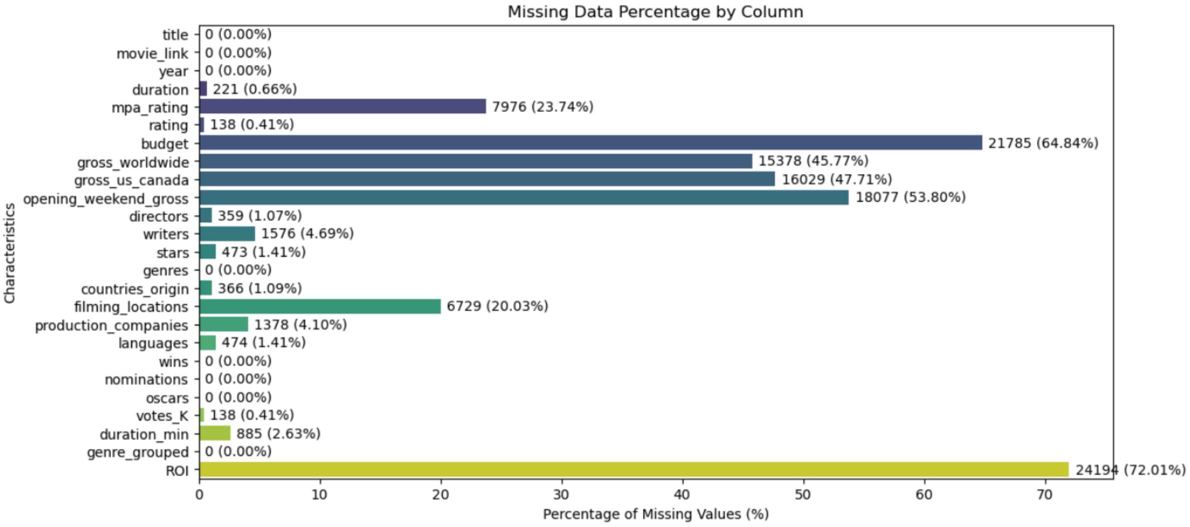
Dataset

The dataset we selected from Kaggle, IMDB Movies From 1960 to 2024, provides annual data on budget, worldwide gross, duration, IMDb rating, and more for the most popular 500–600 movies per year from 1960 to 2024, extracted from IMDb. It includes over 30,000 movies spanning more than 60 years, offering valuable insights into long-term trends in the film industry.

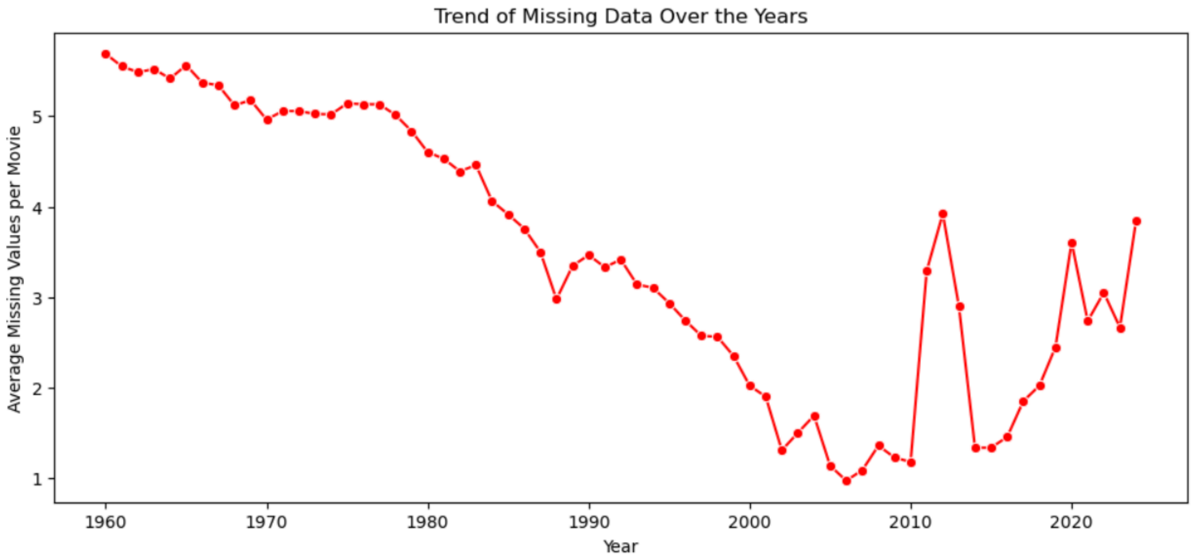
Additionally, using the title and release year, we can retrieve movie posters, which can be valuable for visualization purposes.

To ensure the dataset was well-prepared for analysis and visualization, we performed several data-cleaning steps, which can be found in this Jupyter Notebook. These included: merging all the csv into one, renaming columns for consistency, handling duplicates and missing values, converting data types, grouping genres for better categorization …

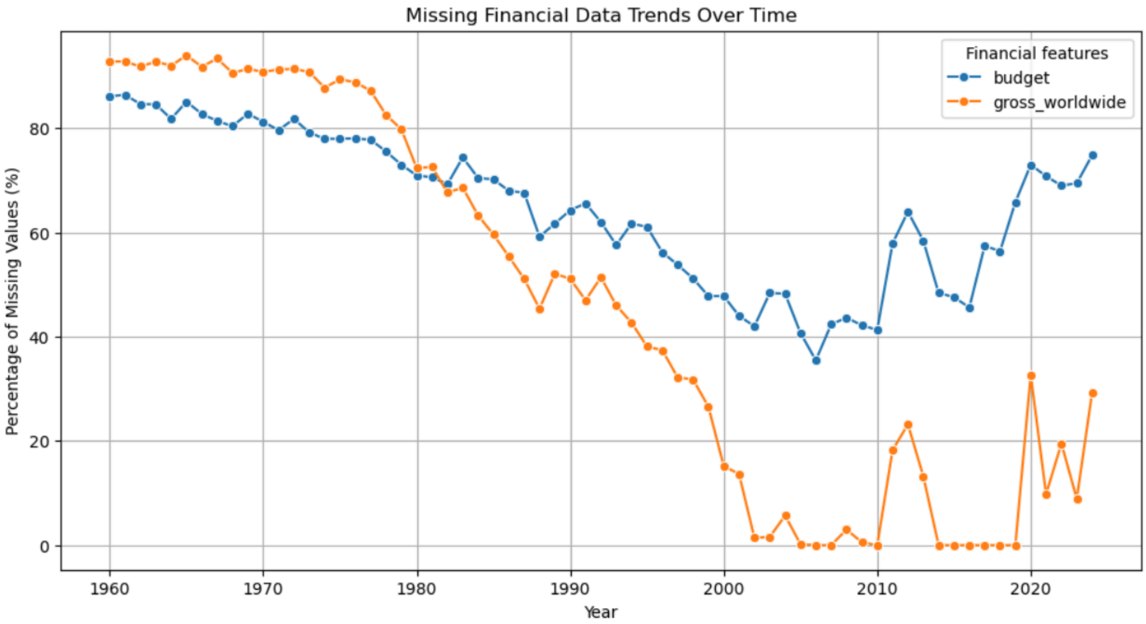
However, this dataset contains a significant amount of missing values, particularly in key financial fields such as budget and worldwide gross. Since these fields are crucial for analyzing box office performance and profitability, we must be cautious when drawing conclusions.



Looking at trends over time, the dataset consistently includes 500–600 movies per year. Between 1960 and 2010, the number of missing values gradually decreased, but a noticeable spike occurred in the early 2010s. Since then, missing data has been on the rise, affecting the reliability of more recent entries. This makes it crucial to be mindful when analyzing years with a high proportion of missing values.



Despite the overall increase in missing values in recent years,worldwide gross data remains relatively well-covered. However box office analyses for earlier decades should be interpreted with caution.



We could enhance this dataset by incorporating additional information from the following sources, which might be useful for obtaining actor names, category-based ratings or tv-shows:

- **IMDb Top 1000**: A collection of 1,000 movies, each featuring details about the four main actors.
- **IMDb Dataset**: Includes 5,200 movies and approximately 100 TV series, listing the top four actors for each title along with poster links.
- **MovieLens 20M**: 27,000 movies with 12 million relevance scores across 1,100 tags.

Problematic : Understanding Movie Success Over Time

Overview & Motivation

The film industry is one of the most influential and profitable entertainment sectors worldwide. Understanding movie trends, audience preferences, and financial performance is crucial for filmmakers, production companies, and marketers. This project aims to analyze a dataset spanning over 60 years of cinematic history, providing insights into key movie characteristics, financial success, and audience reception.

Why this Subject?

The movie industry has evolved significantly over the decades, with changes in genres, production budgets, audience expectations, and technological advancements. By analyzing historical data, we can identify trends in movie success factors through a central question:

With success in cinema being measured through multiple lenses—audience reception, box office revenue, and industry recognition—what are the key factors influencing these metrics, and how have they evolved over the decades?

Objectives

The main goals of this project are:

- Analyze key characteristics influencing movie performance.
- Compare box office revenue, IMDb ratings, and awards to uncover patterns.
- Identify how success factors evolved due to changes in audience tastes, filmmaking techniques, and industry practices.

Target audience

This project is intended for:

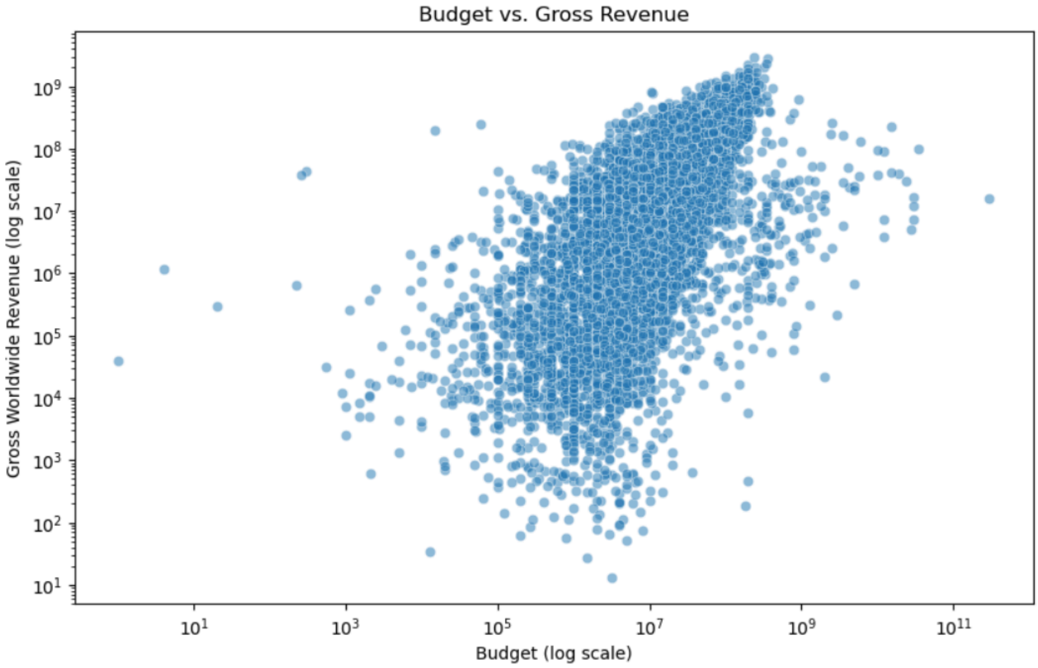
- Film enthusiasts and critics curious about the factors that define great films over time
- Film industry professionals (producers, directors, screenwriters) who want to understand what contributes to a movie's success
- Investors & marketers who seek data-driven insights into movie profitability and audience engagement

Exploratory Data Analysis

In this Jupyter Notebook we conducted an exploratory data analysis to uncover key patterns in the dataset. Here are some key insights we gained from this analysis:

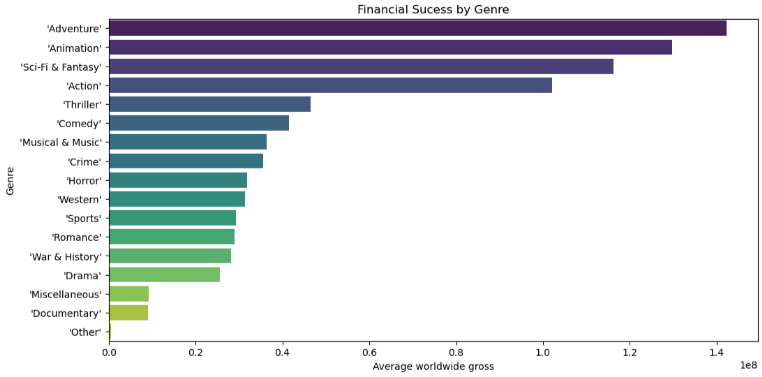
Budget vs Gross

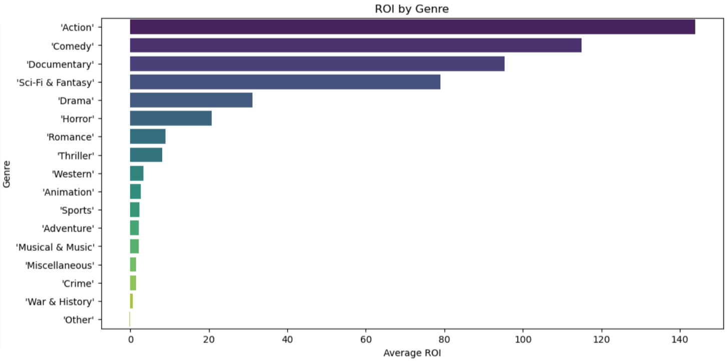
The scatter plot reveals a positive correlation between a movie's budget and its gross worldwide revenue. Higher financial investment generally leads to greater box office success.

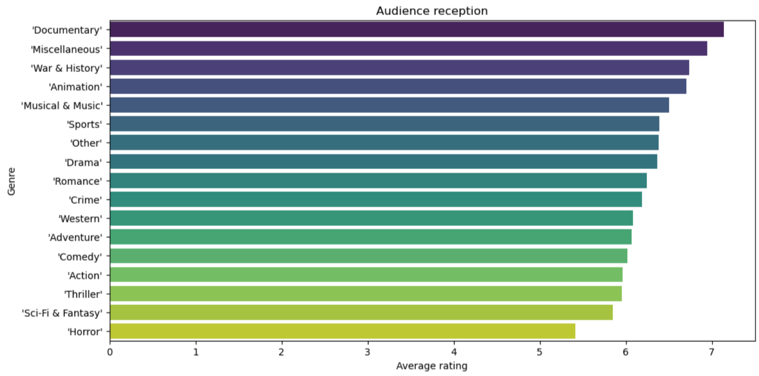


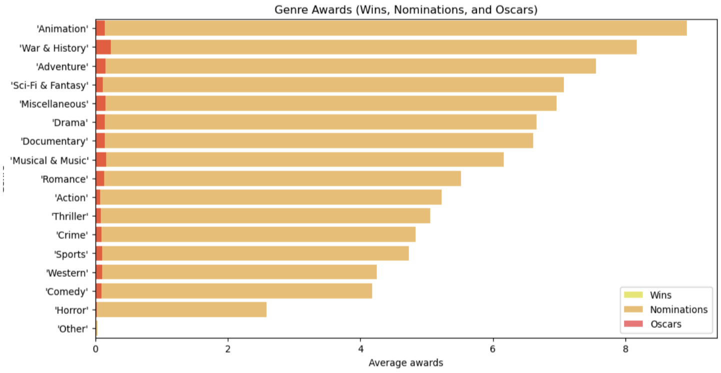
Genre Analysis

Focusing on movie genres, we performed an in-depth analysis to determine how different genres correlate with success metrics like worldwide gross, IMDb rating and awards.Our analysis reveals that Adventure, Animation, and Science Fiction dominate the box office, while Documentary, War, and History receive strong critical acclaim despite lower earnings. Interestingly, audience ratings and industry recognition do not always align with financial success. This highlights that success in the film industry can be measured in multiple ways.



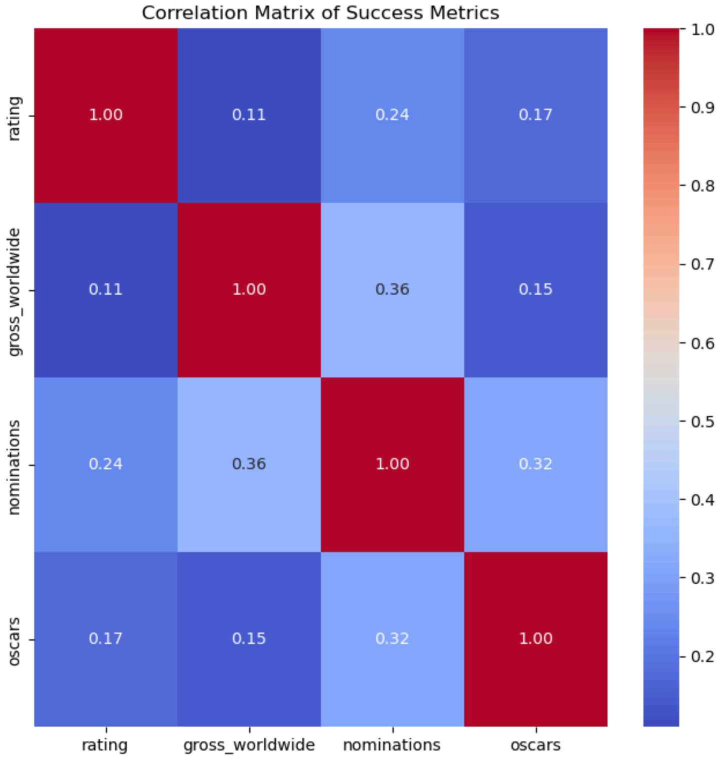






Measures of Success and Correlation

The genre analysis highlights the complexity of defining success in the film industry as commercial performance, critical reception, and cultural impact can each tell a different story. Thus we analyzed the correlation between key metrics such as worldwide gross, IMDb ratings, nominations, and Oscars.



The correlation between IMDb rating and worldwide gross is quite low (0.11), indicating that audience ratings have little influence on a movie's financial success. Similarly, the correlation between ratings and nominations is slightly stronger (0.24) but still weak. On the other hand, the correlation between worldwide gross and nominations is moderate (0.36).

These findings suggest that financial success, critical acclaim, and industry recognition operate somewhat independently, reinforcing the need for a multifaceted approach when evaluating movie performance.

Related Work

Several analyses have been conducted using IMDb data, focusing on various aspects of the film industry. For example, the Historical Movie Map visualizes movies based on their historical settings, allowing users to explore films by release year, IMDb score, and historical accuracy. Another project, IMDb TOP 10 Movies, presents an interactive visualization of top-rated movies, enabling users to filter by genre, release year, and ratings.

Originality of Our Approach

Our analysis distinguishes itself by examining the interplay between financial performance, critical reception, and industry recognition across different movie characteristics (genres, budget, languages,...) . While previous studies have focused on IMDb rating, our approach integrates multiple success metrics to provide a comprehensive understanding of what defines success in the film industry. By analyzing correlations between worldwide gross, IMDb ratings, nominations, and Oscars, we highlight the multifaceted nature of cinematic success.

Sources of Inspiration

We were inspired by several visualizations we found, such as the Spotify Webinar, which displays multiple characteristics side-by-side with dynamic filtering options (similar to a PowerBi Dashboard). We believe this feature would be effective for our project, enabling users to compare different metrics of success. Additionally, this visualization uses music to add another dimension to the experience, making it more immersive and enjoyable.

We also liked the Top 50 IMDB Movies: How many have you watched?, which showcases movie posters that reveal detailed information when you hover over them. We think this would be a fun way to display data, as using movie posters could make the experience more visually appealing and intuitive.

Finally, we were inspired by the scatter plot visualizations used in Beauty and the Beast. The clean and colorful scatter plots effectively represent the relationship between variables, and makes the data easy to interpret.