Columbia School of Engineering
AI Boot Camp
PROJECT #1

FANTASTIC 4
(Team 4):
Jennifer Leone
James O’Brien
Osita Igwe
Giancarlo Ocasio
DoraMaria Abreu

02.14.24

Executive Overview

This data analysis and visualization project aims to investigate and identify critical attributes influencing the commercial success of movies. Leveraging data sourced from IMDb, the project aims to uncover patterns and correlations between various factors and a movie's box office performance. Through meticulous analysis and visualization techniques, the project seeks to provide actionable insights to stakeholders in the film industry, enabling them to make informed decisions to enhance the financial viability of their cinematic endeavors.

Our exhaustive analysis of a dataset encompassing over 10,000 films from 1903 to 2023 offers critical insights into the dynamics shaping box office performance. This rich dataset spans 60 countries, 54 languages, and 19 genres, providing an in-depth analysis of factors that drive movie popularity, revenue generation, and audience preferences. Equipped with these insights, executives can gain a deeper understanding of audience preferences and make well-informed decisions regarding future projects that are profitable and determining returns on investments (ROI).

KEY QUESTIONS

Are there any correlations between movie budgets and box office performance?

- Explore whether higher budgets lead to higher box office earnings.

Are there any outliers or anomalies in the data?

- Look for unusual or unexpected patterns in the data that may require further investigation.

Which actors have appeared in the most movies?

- Identify prolific actors and actresses in the IMDb database.

Data Collection, Cleanup & Exploration

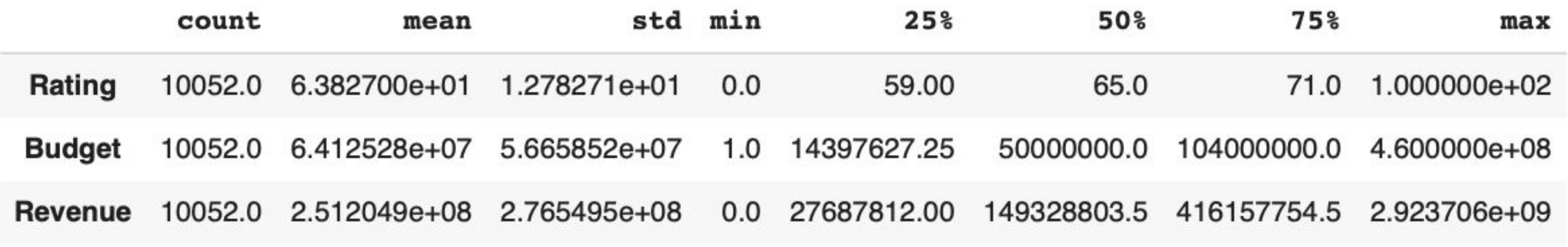
- **Source:** Kaggle (IMDb subset), cleaned by dataset creator using machine learning for missing values.
- **Format:** csv file
- **Contents:** 10,000+ entries, 12 columns including movie names, genres, crew, budgets, and revenues.
- **Data Clean-up:** Converted date formats, standardized numeric fields, and removed incomplete entries, resulting in 10,052 rows for analysis. Renamed columns and reset index.
- **Analysis Focus:** Trends in ratings, budgets, genres, and revenues, using categorical and numerical, single variate and multivariate analysis to identify key industry patterns.

DESCRIPTIVE STATISTICS

- Total films: 10,052
- Time period covered: 1903- 2023
- Total number of unique countries: 60
- Total number of unique languages: 54
- Total number of unique genres: 19

Descriptive Statistics Summary:

- **Ratings:** Ranges from 0 to 100, with a mean of approximately 63.8 and a standard deviation of 12.78. This suggests a moderate variation in movie scores.
- **Budget:** Shows a wide range, from as low as 1 to as high as 460 million, with a mean of approximately 64.12 million. The standard deviation is large (56.65 million), indicating significant variability in movie budgets.
- **Revenue:** Also varies widely, from 0 to about 2.92 billion, with a mean of approximately 251.20 million and a high standard deviation (27.65 million) reflecting substantial variability in movie revenues.

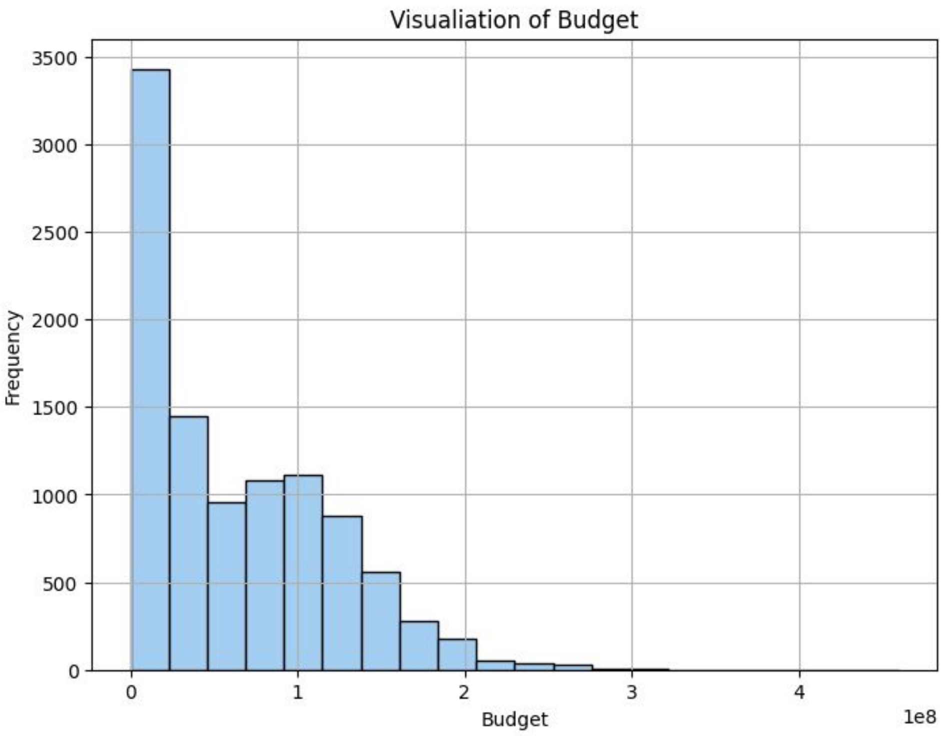


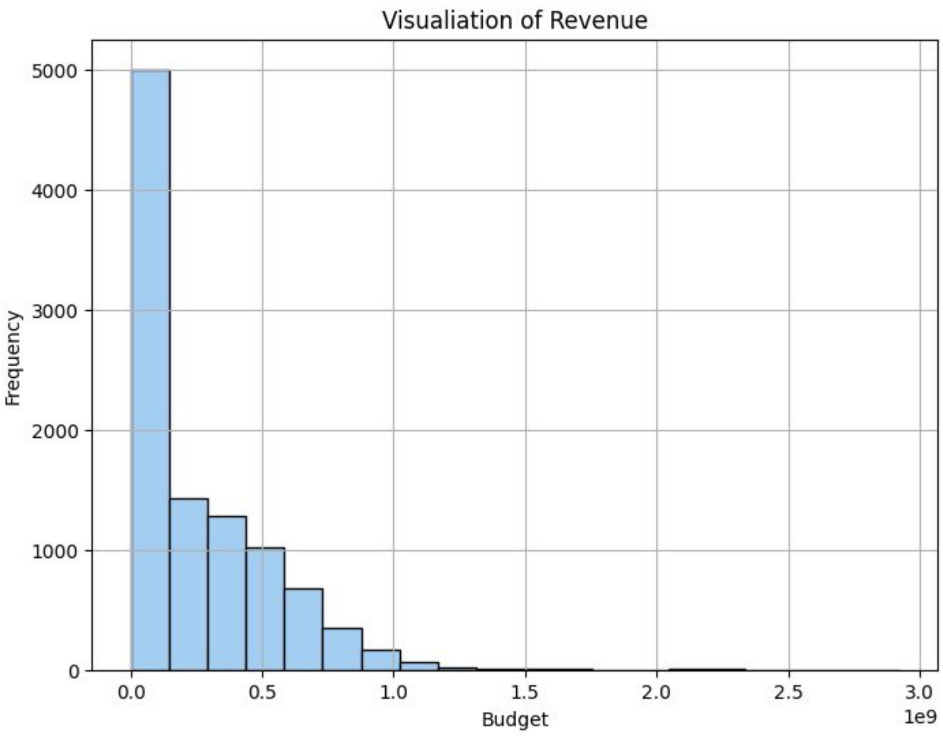
	count	mean	std	min	25%	50%	75%	max
Rating	10052.0	6.382700e+01	1.278271e+01	0.0	59.00	65.0	71.0	1.000000e+02
Budget	10052.0	6.412528e+07	5.665852e+07	1.0	14397627.25	50000000.0	104000000.0	4.600000e+08
Revenue	10052.0	2.512049e+08	2.765495e+08	0.0	27687812.00	149328803.5	416157754.5	2.923706e+09

BUDGET & REVENUE

DATA VISUALIZATIONS

BUDGET and REVENUE





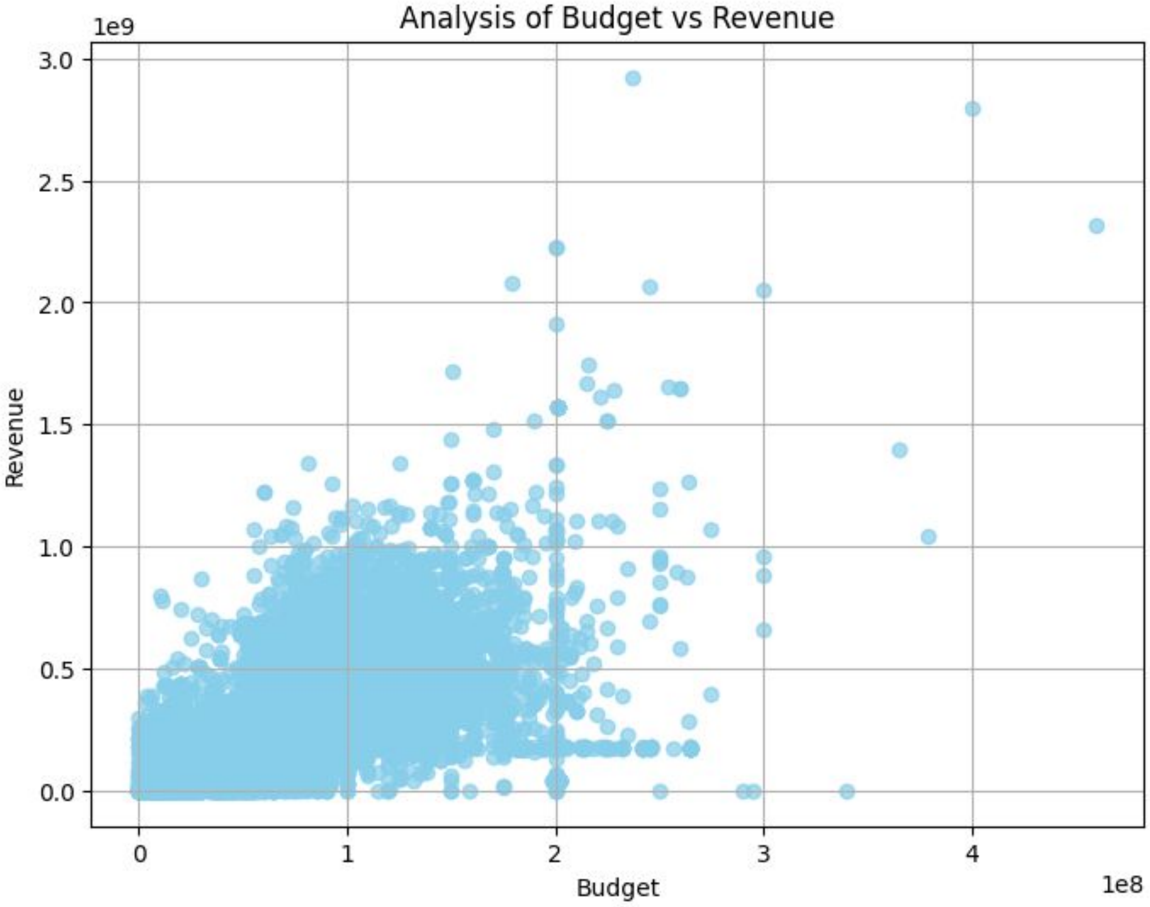
BUDGET vs. REVENUE

Lower Budget, Lower Revenue: Predominantly, films with modest budgets correspond to lower revenues, highlighting the indie and low-budget segment's contribution.

High Budget Successes: Notable outliers in high budget and revenue areas underscore blockbuster and franchise movies' commercial triumphs.

Low Budget, High Revenue Outliers: Several films defy expectations with high returns on low investments, illustrating remarkable success stories.

Moderate Budget Performers: Films with balanced investments and returns demonstrate that substantial success doesn't always require huge budgets, often attributed to strong narratives or effective marketing.



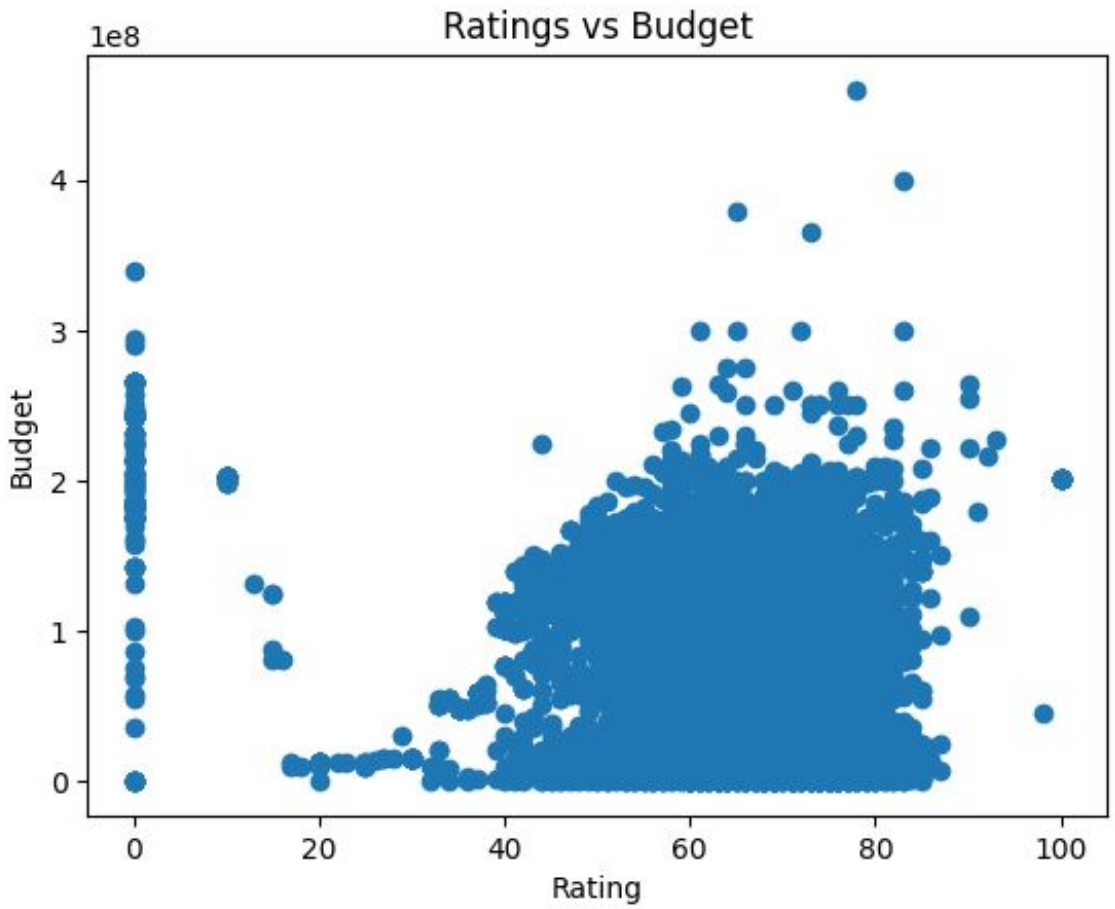
BUDGET vs. RATING

High Ratings, Lower Revenues: Many films with higher ratings don't necessarily earn more, often being critically acclaimed yet not widely popular.

Middle Range Popularity: Most movies score between 60-80, indicating a broad appeal across various genres and audiences.

Rare High Performers: Few films achieve both high ratings and revenues, marking them as unique successes in both critical and commercial aspects.

Low Ratings and Revenues: A minority of films with low ratings also see minimal financial success, highlighting challenges in attracting audiences.



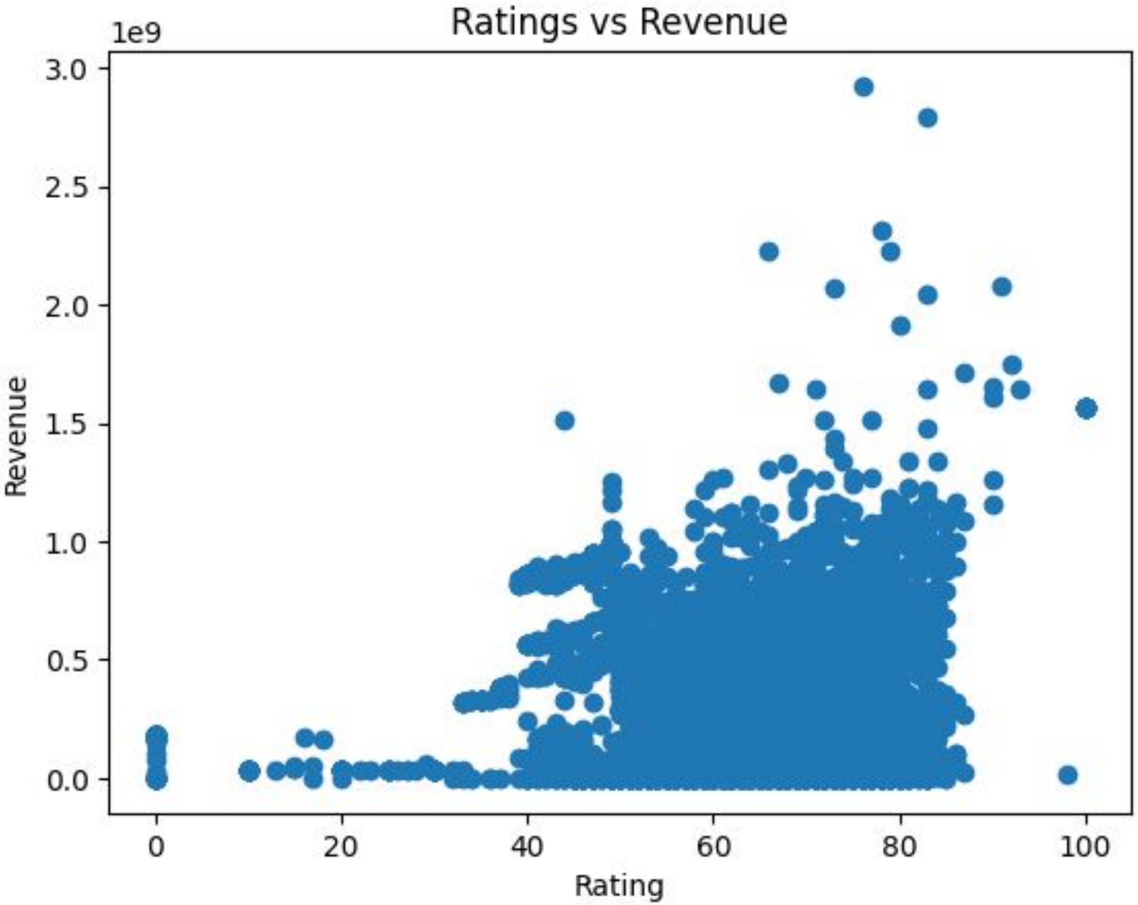
RATING vs. REVENUE

High Ratings, Lower Revenues: Many films with higher ratings don't necessarily earn more, often being critically acclaimed yet not widely popular.

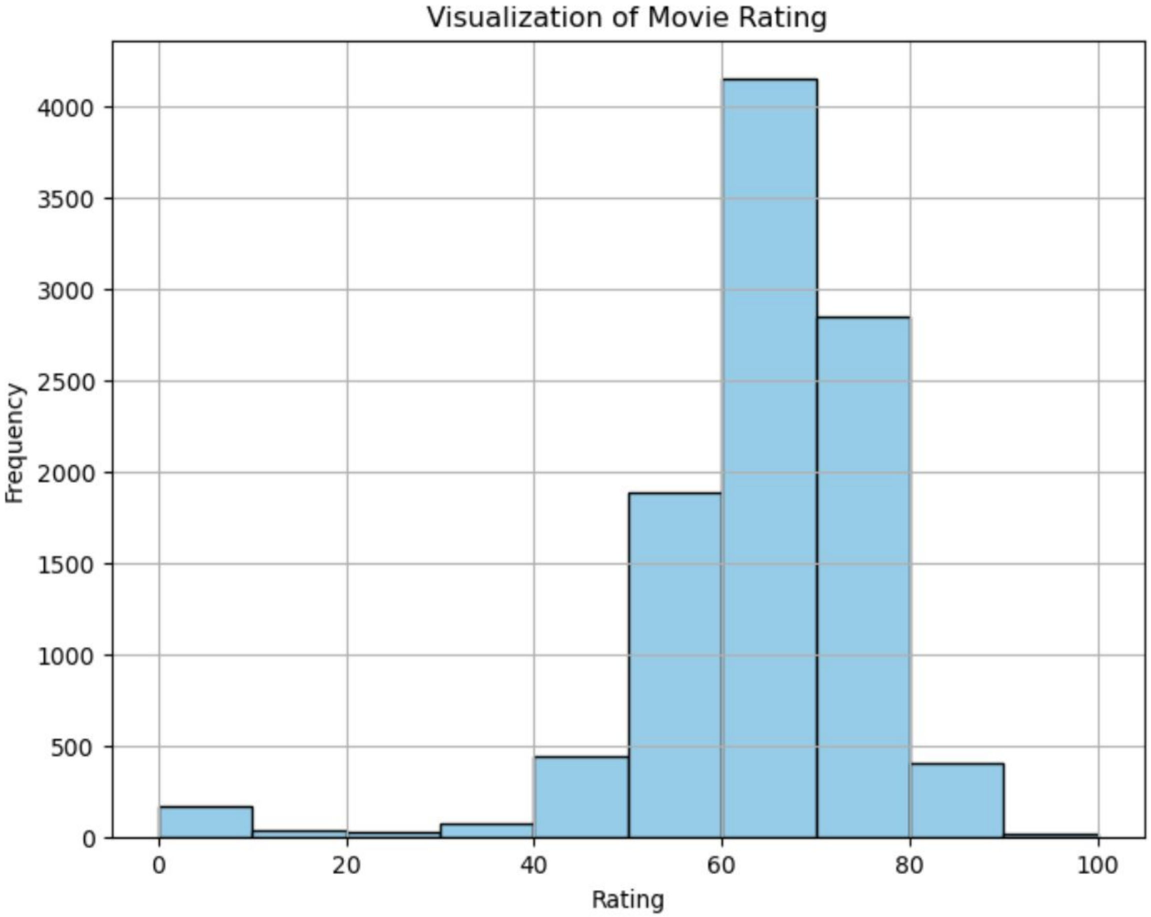
Middle Range Popularity: Most movies score between 60-80, indicating a broad appeal across various genres and audiences.

Rare High Performers: Few films achieve both high ratings and revenues, marking them as unique successes in both critical and commercial aspects.

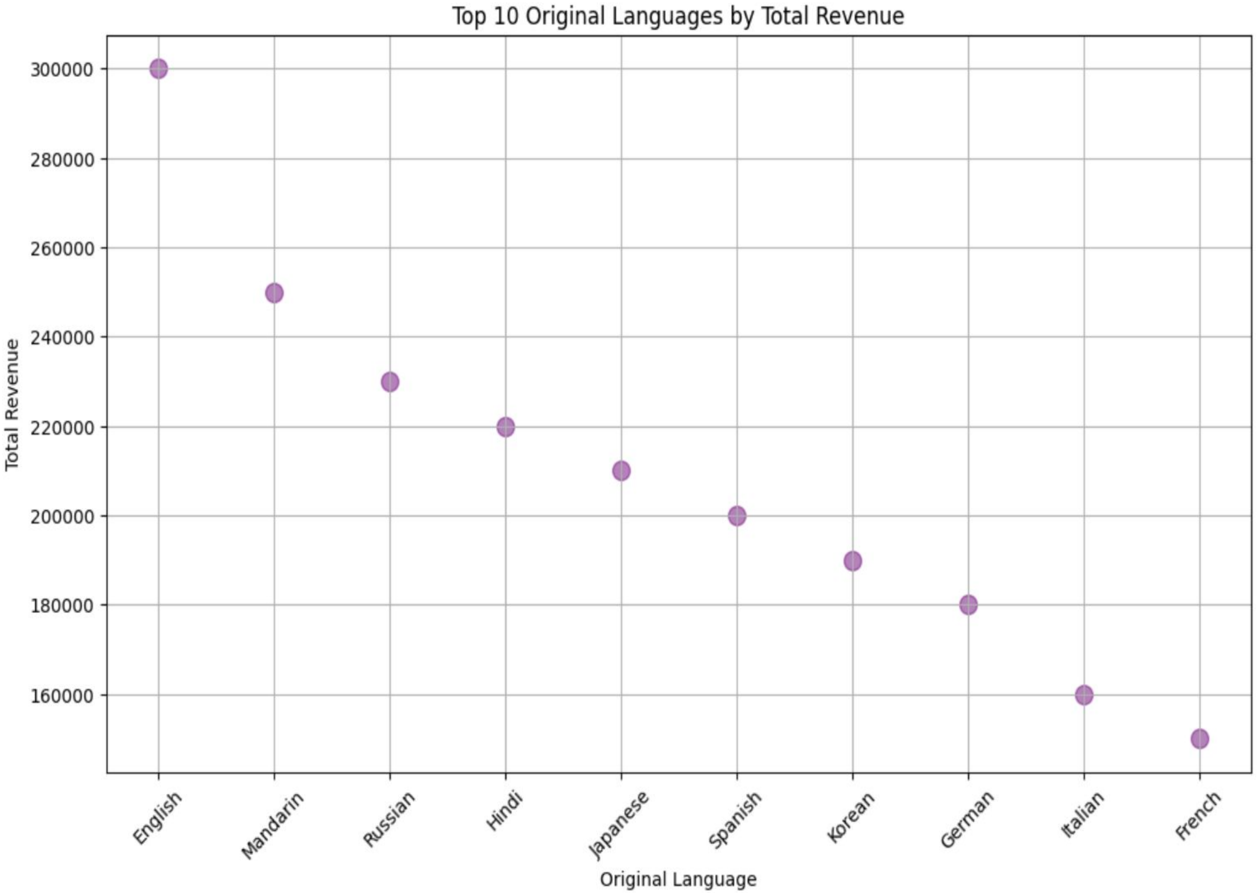
Low Ratings and Revenues: A minority of films with low ratings also see minimal financial success, highlighting challenges in attracting audiences.



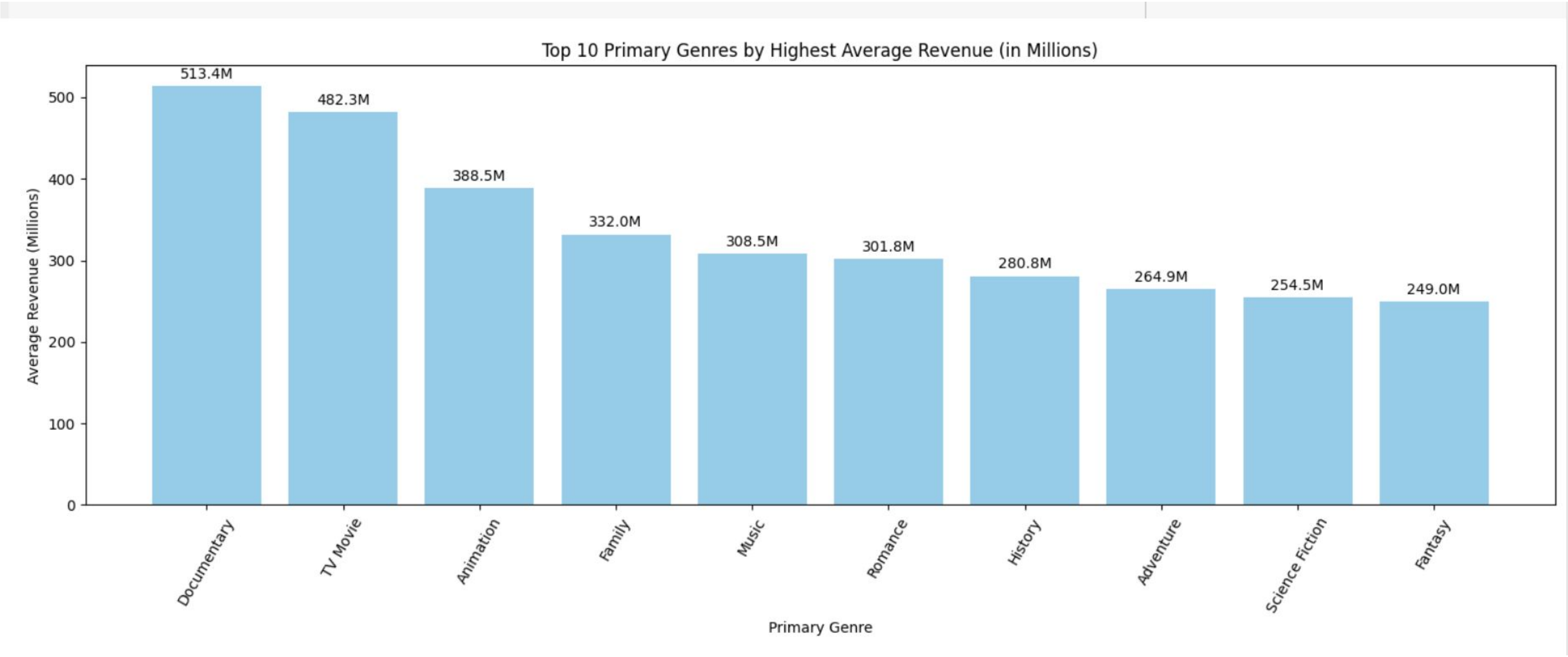
RATING DISTRIBUTION 1903-2020



REVENUE by LANGUAGE



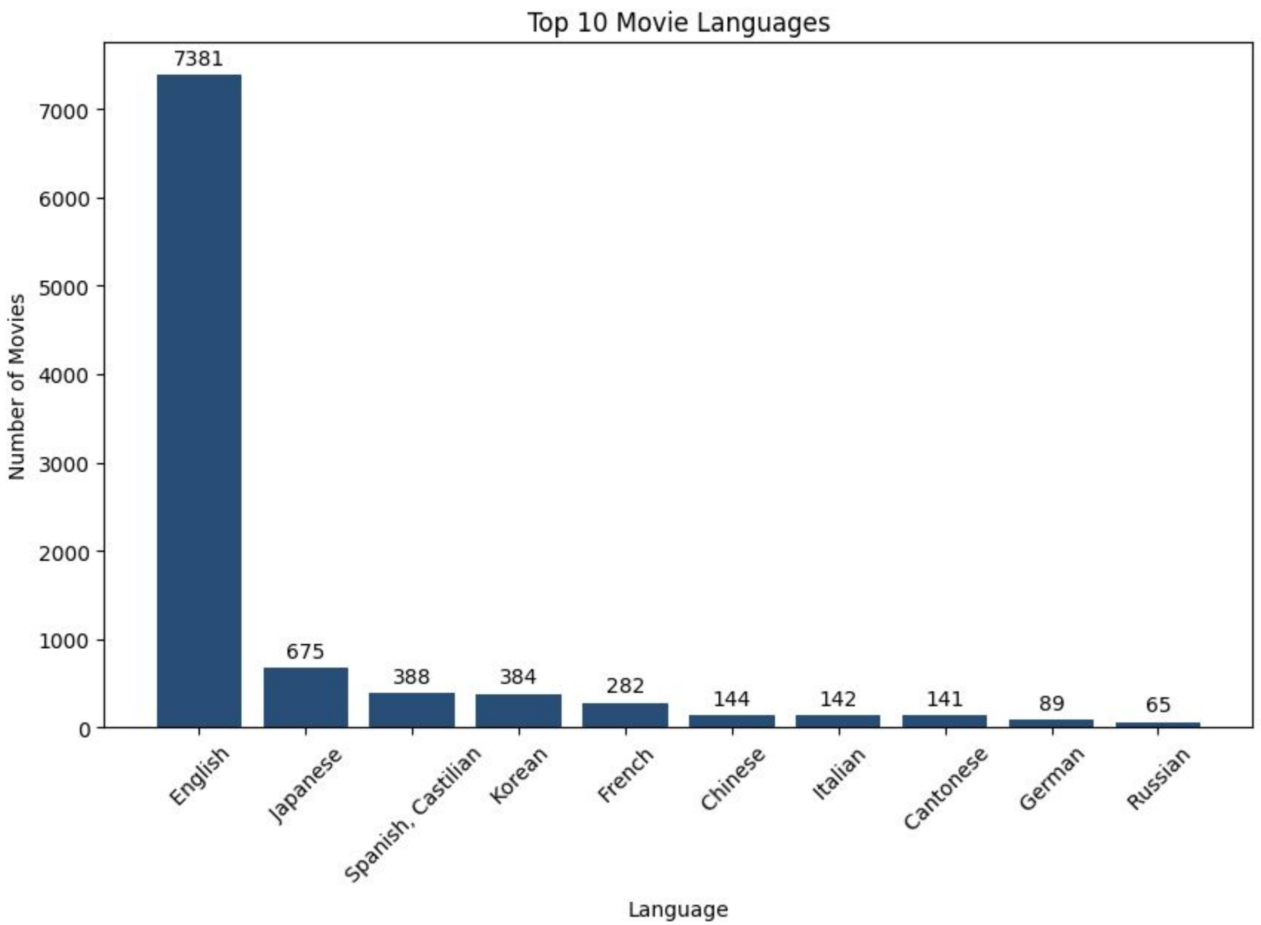
REVENUE by GENRE



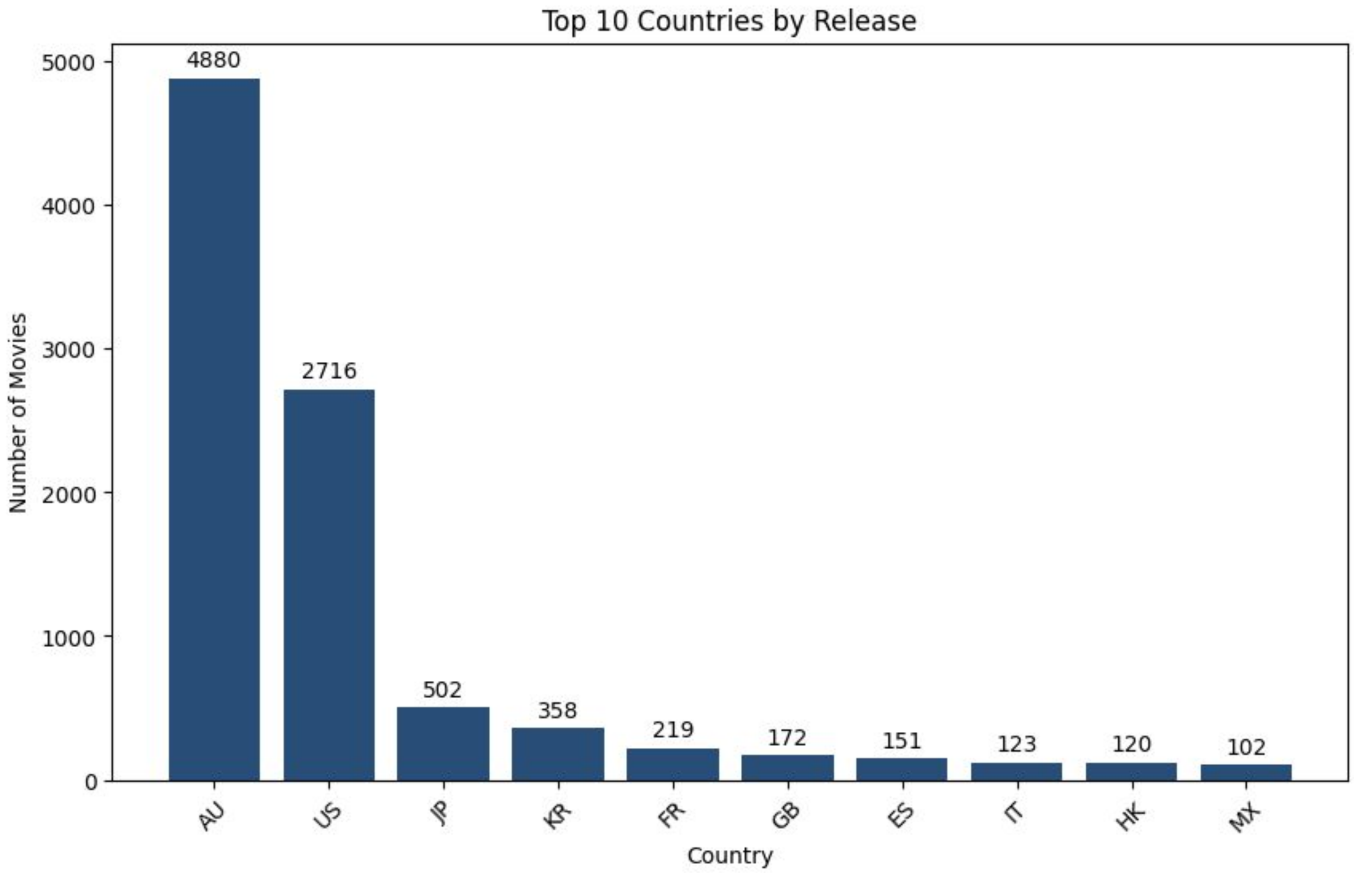
COUNTRY & LANGUAGE

DATA VISUALIZATIONS

TOP 10 MOVIE LANGUAGES



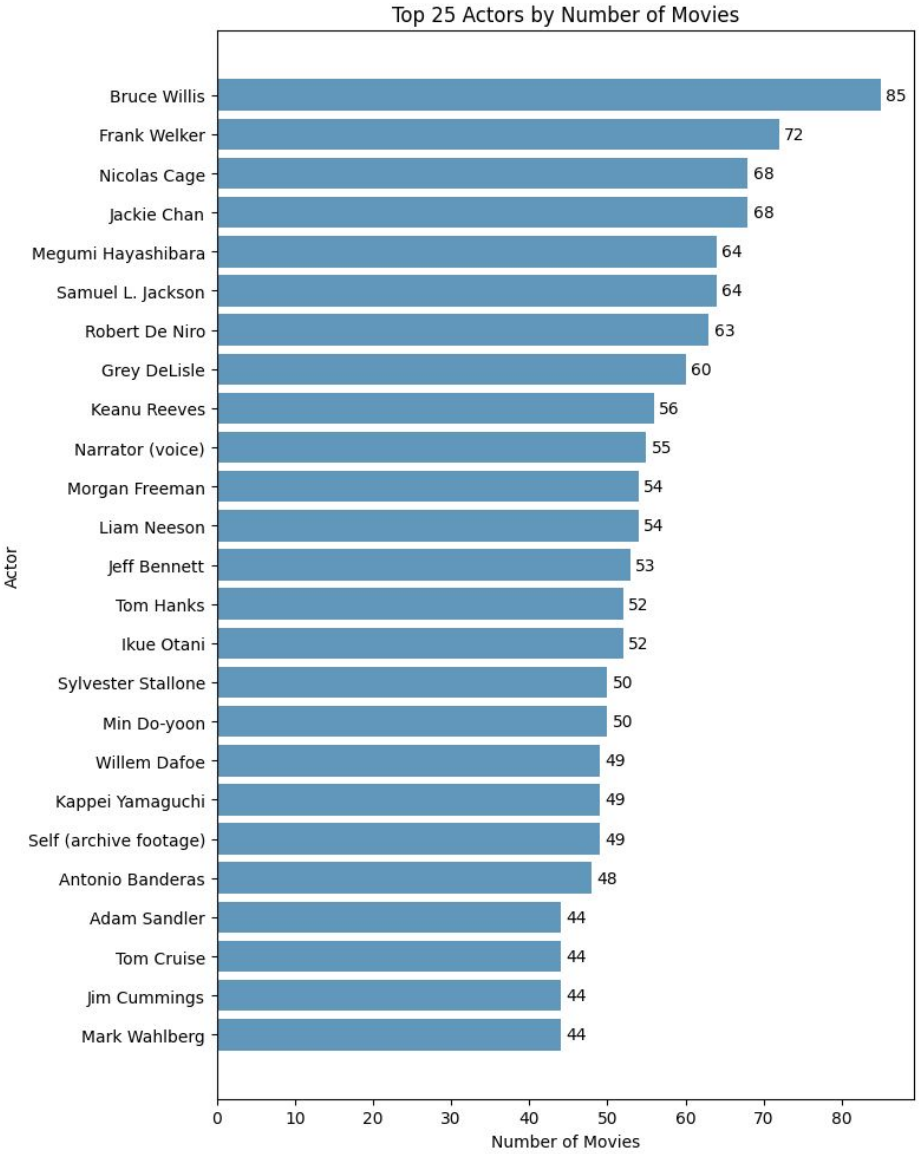
TOP 10 COUNTRY by RELEASE



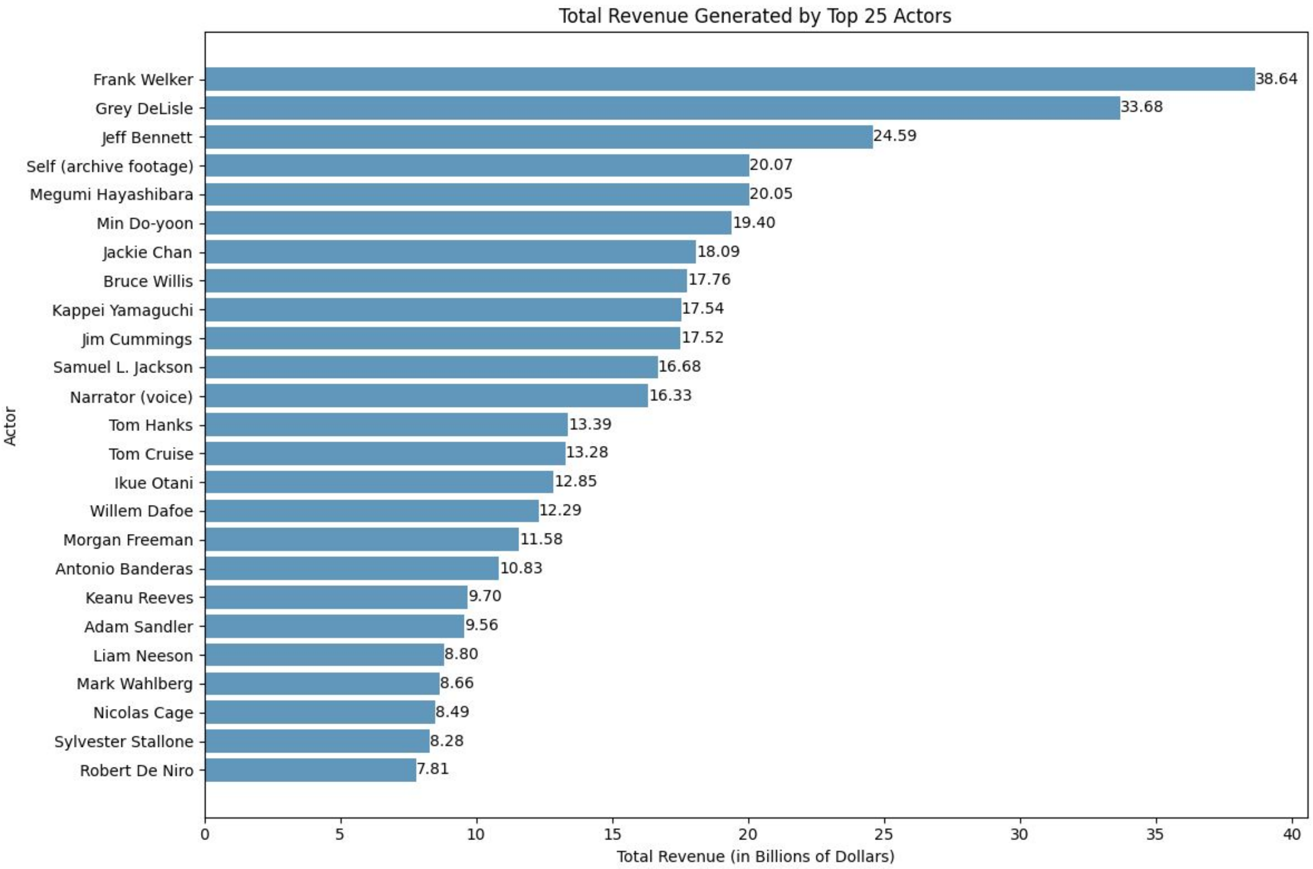
ACTORS

DATA VISUALIZATIONS

TOP 25 ACTORS BY NUMBER OF MOVIES



TOTAL REVENUE GENERATED BY TOP 25 ACTORS



And if you are wondering who Frank Welker is...

FINDINGS BASED ON DATA

KEY FINDINGS

Budget and Revenue Correlation: A strong positive correlation exists between movie budgets and revenues, highlighting the importance of strategic budget allocation.

Genre Popularity: Certain genres consistently engage audiences more, suggesting potential areas for investment.

Impact of Language: English-language films dominate the dataset, indicating a significant market preference.

User Ratings: There's a slight negative correlation between budget and ratings, and a very weak positive relationship between revenue and ratings, suggesting that financial success is not solely determined by high budgets or high ratings.

RECOMMENDATIONS

Strategic Budget Allocation

- While higher budgets often lead to higher revenues, significant returns can also be achieved with moderate investments. Prioritize spending on aspects that enhance production value and audience appeal.

Leverage User Ratings and Feedback

- Engage actively with audience feedback to align future projects with viewer preferences, focusing on quality over quantity to boost ratings and success.

RECOMMENDATIONS

Data-Driven Decision Making

- Engage actively with audience feedback to align future projects with viewer preferences, focusing on quality over quantity to boost ratings and success.

Audience Engagement and Marketing

- Develop targeted marketing strategies based on genre, themes, and user ratings. Build a strong online presence to foster community and anticipation.

Further Analysis

Longitudinal Analysis of Genre Trends: Investigating the evolution of genre popularity and profitability over time can provide studios with actionable insights into shifting cultural trends and technological advancements, enabling them to stay ahead of market dynamics.

Assessing the Digital Landscape's Impact: Analyzing the influence of streaming platforms and social media on movie success, especially in comparison between traditional theatrical releases and digital premieres, can offer studios strategic insights into optimizing content distribution in the digital age.

Demographic-Driven Content Strategy: A deeper understanding of audience demographics can enable studios to tailor content and marketing strategies more effectively, enhancing viewer engagement and financial outcomes.

Predictive Analytics for Box Office Success: Employing machine learning to analyze pre-release data can help predict box office performance, allowing studios to make informed decisions on marketing and distribution strategies to maximize revenue potential.