BAIS: 3250 Data Wrangling

Behind the Numbers: Do Ratings, Runtimes, and Star Power Predict Box Office Success?

Ben Alger, Mason Even, Owen Pullam

GitHub: https://github.com/mieven/BAIS-3250-Final-Project

Introduction

Each year, the movie industry showcases a mix of critically acclaimed films and major blockbusters, yet many highly rated movies go unnoticed by general audiences. Despite earning strong reviews on platforms like IMDb, Metacritic, and Rotten Tomatoes, these films often underperform at the box office when compared to bigger-budget productions with average ratings. This disconnect raises key questions: Do higher ratings actually influence financial success? Are critically celebrated movies less commercially viable than widely marketed ones? In an era where digital reviews shape public perception, understanding the true impact of ratings, runtime, and star power on movie performance has become increasingly relevant.

This project explores three central questions to uncover the drivers behind a film's commercial and critical outcomes. First, we examine how movie runtimes have evolved over the decades and whether longer runtimes correlate with higher earnings. Second, we assess the relationship between ratings across major platforms and a film's total gross revenue using both correlation analysis and machine learning models. Lastly, we investigate whether the presence of award-winning actors and directors contributes to box office success or critical acclaim. By analyzing these patterns, this study aims to identify which data points (if any) can reliably predict a movie's financial and critical impact.

Data

For our first dataset, we decided to work with Movie Budgets And Revenues because it specifically carries the release date, budget, and both domestic and foreign gross for over 6,400 different movies. The movie budget metric helps to indicate the amount of money that was spent on each film and find the correlation between more spending leading to more revenue. Domestic and foreign gross will give insight into how movies perform both in and out of the United States. Overall, the Kaggle dataset *Movie Budgets And Revenues* provides insights to expand on our web scraping ratings and other findings. The only problem we had to address to merge with our scraped data was the discrepancies in movie names between our Kaggle dataset and our saved web scraping. To fix this we created a function to find similar movie titles and append them to a separate data frame. We then uploaded this data frame to an excel file and manually sifted through each movie title to find whether it was an encoding error or a different movie title. Finally, we replaced all the titles in the Kaggle data set before merging again.

When choosing a website to scrape it was critical to find one that includes multiple ratings at once. This saves us the hassle of scraping and crawling each of the different ratings websites such as, IMDb, Rotten Tomatoes, and Metacritic. We found Movie Rankings to hold all three of the common rating scores.

Movie Rankings is a website that provides rankings, reviews, and ratings for movies across different genres. The website also includes key attributes such as movie titles, release years, genres, each rating we are looking for, user and critic ratings, and box office performance. We collected the data from *Movie Rankings* by scraping the publicly available movie listings before saving our data as a CSV to integrate with our Kaggle dataset. *Movie Rankings* was built using mainly HTML with some elements of JavaScript which made it easier to navigate when looking for the key attributes to scrape from each page and movie.

Citations

Kaggle - https://www.kaggle.com/datasets/dahvid/movie-budgets-and-revenues

Website - https://www.movierankings.net/

Scraping Logic & Cleaning

While scraping *Movie Rankings* we started on the home page before navigating to the full rankings page. This page stores over 6,000 movies. We then scraped the first 120 movies from the first page of full rankings for the initial project proposal scrape. The script first ensures that all movies on the rankings page are fully loaded by continuously scrolling down until no new elements appear. Once loaded, it locates and extracts movie elements using class names, retrieving details like titles, links, and ratings. After compiling the initial dataset, the script iterates through each movie link, loading individual movie pages to extract director, cast, runtime, box office revenue, genres, and awards. Randomized delays between requests help mimic the human browsing behavior to reduce the risk of detection and blocking.

After scraping the data from the website, we changed the data types of ‘Budget’, ’Domestic Gross’, and ‘Worldwide Gross’ to integers. This helps to make the numbers easier to work with in analysis as opposed to the original object data type. To ensure consistent formatting and analysis of rotten tomatoes, IMDb, and Metacritic scores, we altered the values, so they are displayed out of 100. This will allow us to do mathematical operations with the column and derive more impactful insights.

We ended up dropping the ‘Number’ and ‘Year Release’ in the final integrated data frame because they were repetitive and irrelevant to any analysis we would perform.

Smaller changes we also made to the data such as replacing (minutes) in ‘Runtime’, the $ sign, and commas/dashes throughout the data with a blank space (“”).

After scraping and cleaning the movie details from *Movie Rankings*, we converted this information to a pandas data frame, aligning its structure with the existing movies and revenues dataset. The integration relied on a common key, the movie title, to merge the two datasets accurately. We merged the Kaggle dataset to the movies pandas data frame on ‘Title’ using ‘inner’. Once merged, the integrated dataset provides a more complete view of each movie, combining rankings, critic scores, and revenue figures, which allows for deeper analysis of financial success relative to reviews and awards.

On the back half of our wrangling, we ended up creating a couple of new columns to be used for analysis. We created ‘remake’ as a binary to show if the movie listed was a remake of the original and then we could account for it appropriately. ‘Oscar Status’ and ‘Globes’ statues further break down the award categories. ‘Cast Strength’ and ‘Director Strength’ consider the score/rank of the cast or director based upon the awards earned. Finally, the total score combines each of the IMDb, Metacritic, and Rotten Tomatoes scores to be one score out of 300.

<u>Data Dictionary</u>

Field Name	**Data Type**	**Description**
Title	object	Title of the movie.
Oscar Status	int64	Indicator of Oscar recognition (possibly 1 = Won, 0 = Nominated, or similar).

Globes Status	int64	Indicator of Golden Globes recognition (same logic as above).
Decade Released	object	Decade during which the movie was released (e.g., "2010s").
Runtime (minutes)	int64	Duration of the movie in minutes.
Box Office	float64	Box office revenue (may include both domestic and international, in USD).
Genre	object	Primary genre classification.
Sub-Genre	object	More specific genre category.
Studio	object	Studio responsible for producing/distributing the movie.
Rotten Tomatoes (out of 100)	float64	Rotten Tomatoes critic score out of 100.
IMDb (out of 100)	float64	IMDb rating scaled to 100.
Metacritic (out of 100)	float64	Metacritic score out of 100.
Director	object	Director of the film.
Starring	object	Main cast members.
Release Date	object	Official release date.
Budget	int64	Estimated budget in USD.
Domestic Gross	int64	Domestic earnings (likely U.S.).
Worldwide Gross	float64	Global earnings.
Total Gross	int64	Sum of domestic and possibly international gross.
Net Gross	int64	Total Gross minus Budget.
remake	int64	Binary flag (1 = remake, 0 = original).
Cast Strength	int64	Numeric score or rank indicating popularity or quality of cast.
Director Strength	int64	Numeric score or rank for director reputation or prior success.
Total Status	int64	Combined status from awards (maybe cumulative of Oscar/Globes/etc.).

total score	float64	Possibly an aggregate score combining all ratings or metrics.

Analysis

Question 1: Has the average runtime of movies changed over different decades and do movies with longer runtimes tend to have higher or lower box office earnings?

We set out to explore two key questions for this analysis question: whether the average runtime of movies has changed over the decades, and whether a movie's runtime has any significant impact on its box office success. We began by examining average runtimes over time. Descriptive statistics revealed that the overall average runtime in the dataset is approximately 113 minutes (just under two hours), with runtimes ranging from 56 to 233 minutes. When grouping movies by decade, we found that the highest number of films were released between 2010 and 2019 (1,059 movies), while the lowest counts came from decades prior to 1960. To statistically evaluate differences in average runtimes across decades, we performed a one-way ANOVA test. The test yielded an F-statistic of 14.63, indicating that the variation between decade averages is 14.63 times greater than the variation within each decade. The accompanying p-value of 1.733e-16 falls well below the 0.05 alpha threshold, allowing us to reject the null hypothesis and confirm that average runtimes differ significantly across decades.

To visualize this trend, we created a boxplot (shown below), which provided several key insights. Runtime has gradually increased over the years, though not dramatically. Earlier decades (such as the 1940s through 1960s) display shorter medians and tighter interquartile ranges. From the 1990s onward, median runtimes have shifted upward, with more films approaching or exceeding 110–120 minutes. Additionally, recent decades exhibit greater variability in runtime, likely reflecting the growth of blockbuster cinema and experimentation in storytelling formats. Outliers are present across nearly all decades, emphasizing the diversity in film lengths. These patterns are clearly illustrated in the visualization below.

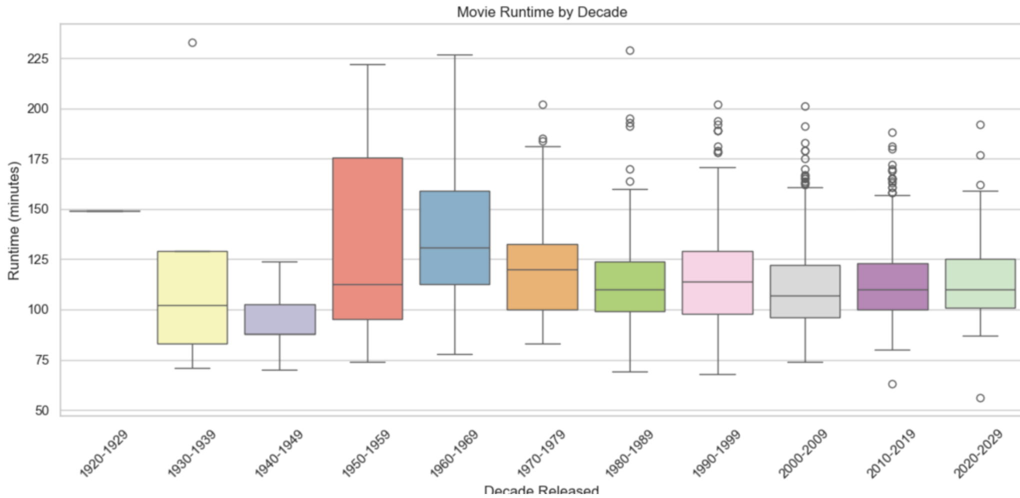


Figure 1 – Question 1: Has average runtime changed over the decades?

Next, we explored whether a movie's runtime has any significant impact on its box office success. Descriptive statistics revealed that the average box office earnings per movie in the dataset were approximately $179 million, with values ranging from as low as $18,000 to as high as $2.79 billion. To better understand how runtimes are distributed, we created a bar chart showing the number of movies across different runtime intervals. Most films clustered between 90 and 140 minutes, reflecting a strong central tendency around the 2-hour mark. The peak density occurred between 110 and 130 minutes, marking this as the most common movie length. Short films under 80 minutes and long ones over 160 minutes were relatively rare, reflected in the thinner tails of the distribution. The accompanying KDE curve highlighted a unimodal and slightly right-skewed shape, suggesting that while most movies fall within a standard range, a few significantly longer films contribute to the extended tail.

We then conducted a similar analysis by examining a bar chart of box office earnings versus the number of movies. This visualization revealed that the majority of films earn less than $200 million, with a high concentration of titles clustered below this point. The distribution is heavily right-skewed, indicating that only a small fraction of movies achieves exceptionally high box office returns. Blockbuster films earning over $500 million represent clear outliers, significantly extending the distribution's upper tail. The KDE curve further illustrates this pattern, showing a sharp decline after the peak, which highlights the stark contrast between typical movie earnings and the rare financial successes.

We then examined the relationship between runtime and box office earnings using a scatter plot (shown below). The plot revealed no strong linear relationship—films of varying lengths can achieve both high and low box office performance. There is a wide vertical spread across most runtime intervals, particularly between 90 and 150 minutes, where earnings range from modest to exceptionally high. Notably, a few blockbuster outliers earning over $1 billion appear across diverse runtimes, indicating that extreme financial success is not tied to a specific film length. Overall, the analysis suggests that runtime alone is not a reliable predictor of box office performance.

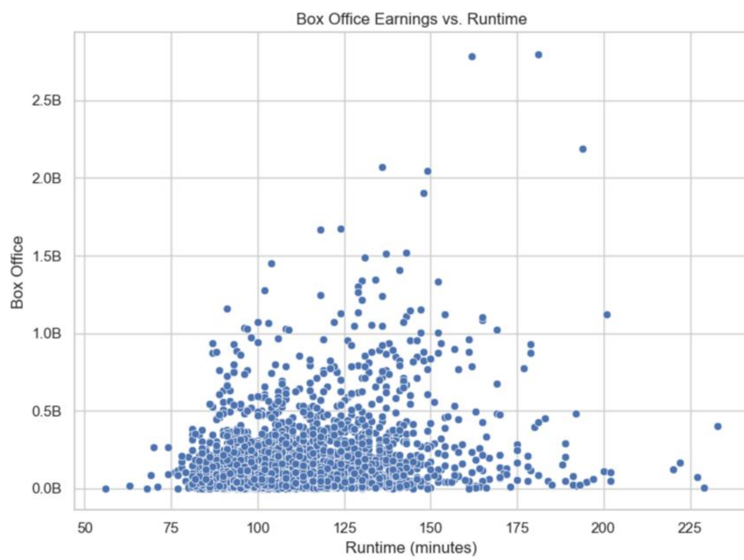


Figure 2 – Question 1: Does Runtime Affect Box Office

We then conducted a Pearson correlation test to evaluate the relationship between runtime and box office earnings, resulting in a correlation coefficient of 0.245. This indicates a weak but positive association—on average, longer movies tend to earn slightly more at the box office. We also calculated a p-value of 1.242e-36, which is effectively zero and far below the standard significance threshold of 0.05. This confirms that the correlation is statistically significant and unlikely to be due to chance. While the scatter plot shows considerable variability, a slight upward trend is evident, particularly among higher-earning films. However, despite this significance, the strength of the relationship remains weak, suggesting that runtime alone is not a reliable predictor of box office performance.

We then prepared the data for linear regression and evaluated the model using Mean Absolute Error (MAE) and R^2. The results were an MAE of $151,696,860 and an R^2 of 0.066. This tells us that while the model's average prediction error is substantial, it also explains only about 6.6% of the variance in box office earnings based on runtime. In other words, runtime has minimal predictive power when it comes to forecasting box office performance, reinforcing the conclusion that other factors play a much larger role in determining a movie's financial success.

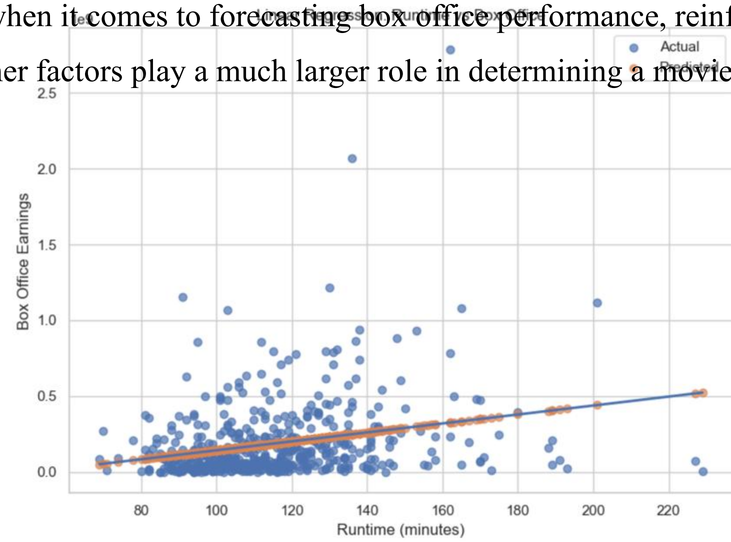


Figure 3 – Question 1: Does Runtime Affect Box Office

<u>Machine Learning</u>

We then applied three machine learning classification techniques to analyze the relationship between runtime and box office performance: Logistic Regression, K-Nearest Neighbors (KNN), and Decision Tree Classifier. Logistic Regression achieved the highest accuracy at 74.61%, followed by KNN at 72.09%, and Decision Tree Classifier at 64.34%. These results indicate that Logistic Regression was the most effective model for classifying movies as high or low box office earners based on runtime, though overall accuracy levels suggest that runtime alone is still a limited predictor.

<u>**Question 2**: What is the correlation between movie ratings on platforms like IMDb, Metacritic, and Rotten Tomatoes and total gross earnings?</u>

In our first step of this question, we analyzed the relationship between ratings from these major review platforms and the box office earnings of various films. Surprisingly, the correlation coefficients across all platforms were consistently low (close to zero) indicating virtually no linear relationship between critical or audience reception and commercial success. IMDb showed the highest correlation at only ~0.064, while Metacritic and Rotten Tomatoes followed closely at around ~0.054. Even an aggregated Total Score, combining ratings from multiple sources, failed to significantly strengthen the correlation. These findings suggest that higher ratings do not reliably predict greater box office performance, pointing to the influence of other factors such as marketing, star power, and franchise recognition.

Taking a deeper look into Total Score, to account all separate ratings, the correlation between a movie's Total Score (an aggregate of ratings from platforms like IMDb, Metacritic, and Rotten Tomatoes) and its total gross earnings is statistically significant (with a p-value of 0.00257, well below the conventional threshold of 0.05) the actual strength of this relationship is extremely weak. The correlation coefficient, approximately 0.0635, indicates that while the relationship is unlikely due to random chance, it lacks meaningful predictive power. This weak correlation is visually supported by the scatter plot, which shows data points widely dispersed and mostly concentrated at lower Total Score values, with no discernible linear trend. Furthermore, the presence of a few high-grossing outliers with high Total Scores does not reflect the broader dataset and likely skews any perceived association.

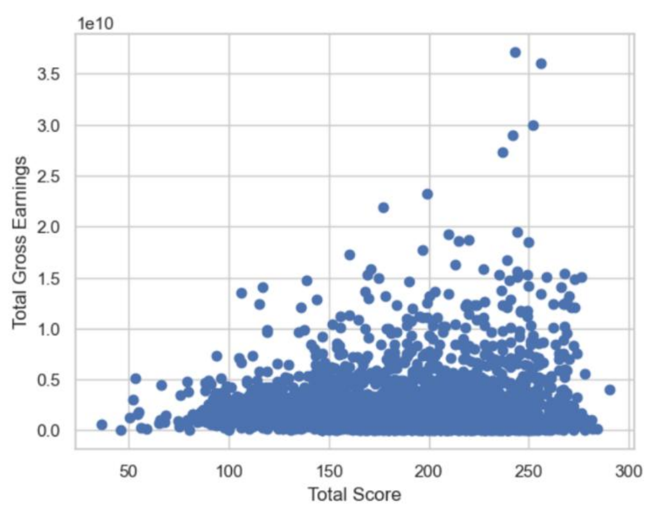


Figure 1 – Question 2: Total Score vs. Total Gross

We then aimed to focus on predicting the Total Gross based off of scores, again Total Score is used to showcase all different rating platforms. The prediction results, ranging narrowly between approximately 2.52 and 2.97 billion, indicate that the model produces low variability across different Total Scores. This mirrors the findings from individual rating platforms and reinforces that even the combined Total Score has limited explanatory power in predicting Total Gross.

The distribution of Total Gross earnings is remarkably similar across all rating sources, with comparable medians and interquartile ranges, indicating that no single platform effectively distinguishes high-earning movies. Each group also contains numerous outliers (films with exceptionally high box office returns) regardless of their ratings, while most movies cluster below $500 million, reflecting modest revenue overall. Notably, the aggregated Total Score behaves much like the individual ratings, showing that combining scores does not meaningfully alter the distribution or enhance predictive insight. The box plot below showcases this distribution and the outliers of those having a much higher gross than the mean.

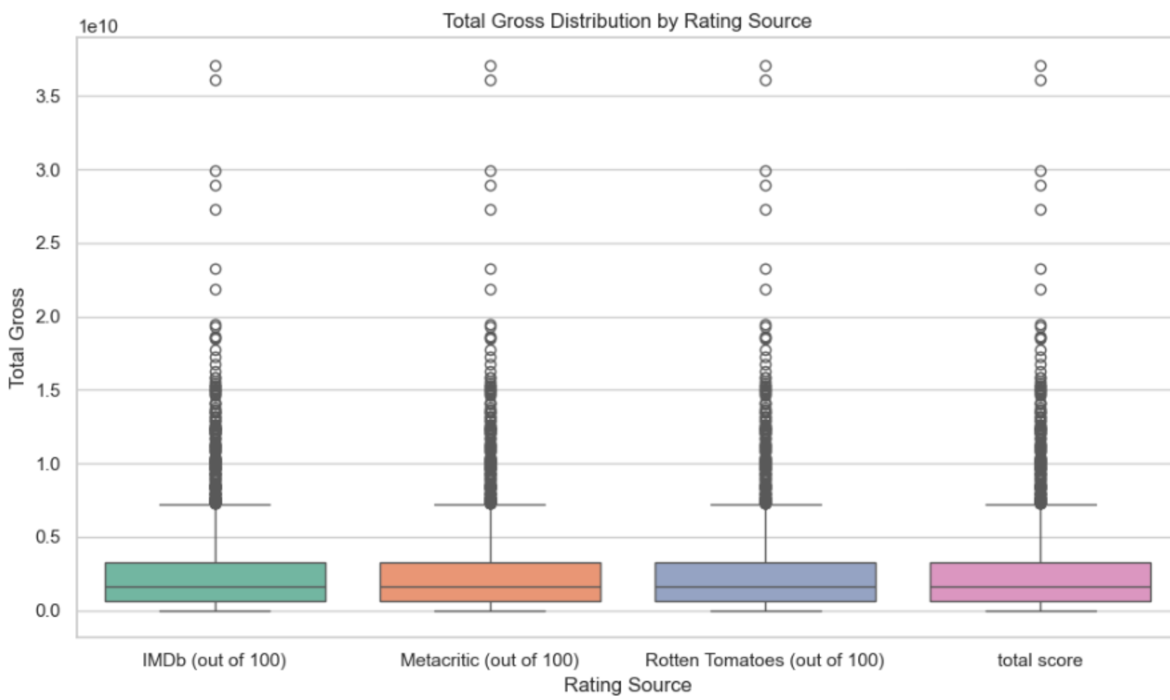


Figure 2 – Question 2: Total Gross Distribution by Rating Source

There is high internal consistency among the rating metrics, with strong correlations ranging from 0.91 to 0.96 between IMDb, Metacritic, and Rotten Tomatoes scores—suggesting they capture similar evaluative signals. The Total Score also shows strong correlations (0.70–0.96) with each individual rating, confirming it effectively aggregates the underlying data. However, all rating metrics, including the Total Score, exhibit very weak correlations (~0.05–0.06) with Total Gross, reinforcing the conclusion that ratings are poor predictors of box office performance. This heatmap displays the correlation matrix between movie rating metrics Total Gross earnings, showing strong internal correlations among the rating platforms (up to 0.96), but very weak correlations (around 0.05–0.06) between any rating metric and Total Gross.

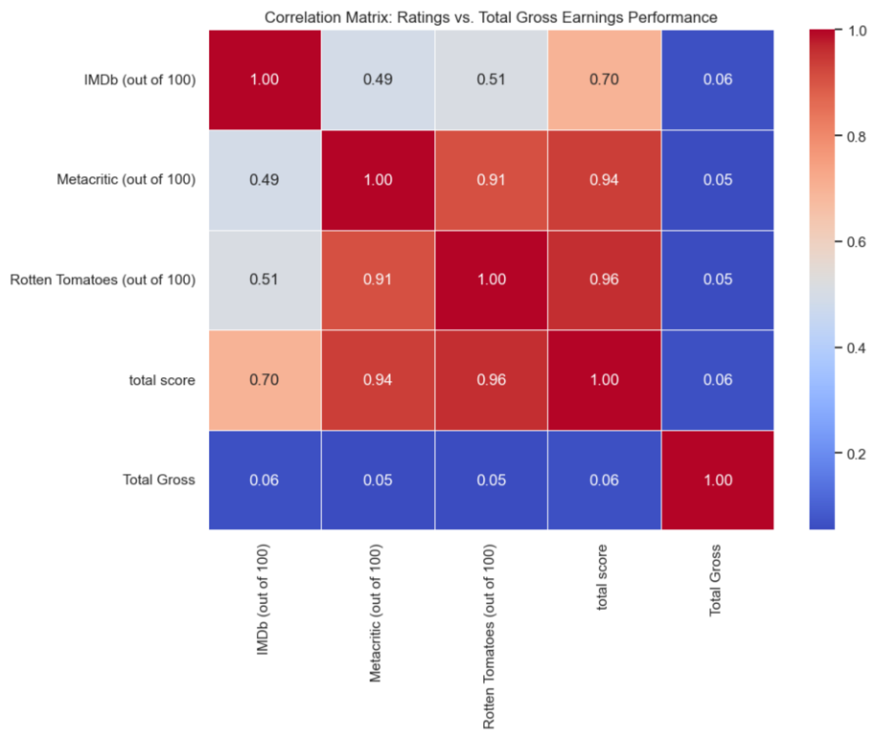


Figure 3 – Question 2: Heat Map of Ratings and Total Gross

Question 3: Does the presence of award-winning actors/directors influence box office or total score?

To explore this question, we had to create a separate data frame called "awards," that listed each person's name, their role (actor or director), the movie title and the release date for each movie an individual participated in that was either nominated for an award or won an award. Then, we classified each nomination to be worth 1 point and each award won to be worth 2 points. This is because there is often controversy amongst nominated awards that "should have" won. Ultimately, we wanted to represent these highly talented actors as fairly as possible. Finally, we counted the total score for each director and actor up to the release date of a future movie, essentially showing the prestige of the cast and director leading up to the release of any given movie.

Our null hypothesis was that highly awarded directors and actors (director strength and cast strength) would have a positive correlation with total score. First, we charted the score of movies over time, seeking any insights or trends to performance, or any indicators of skewed data.

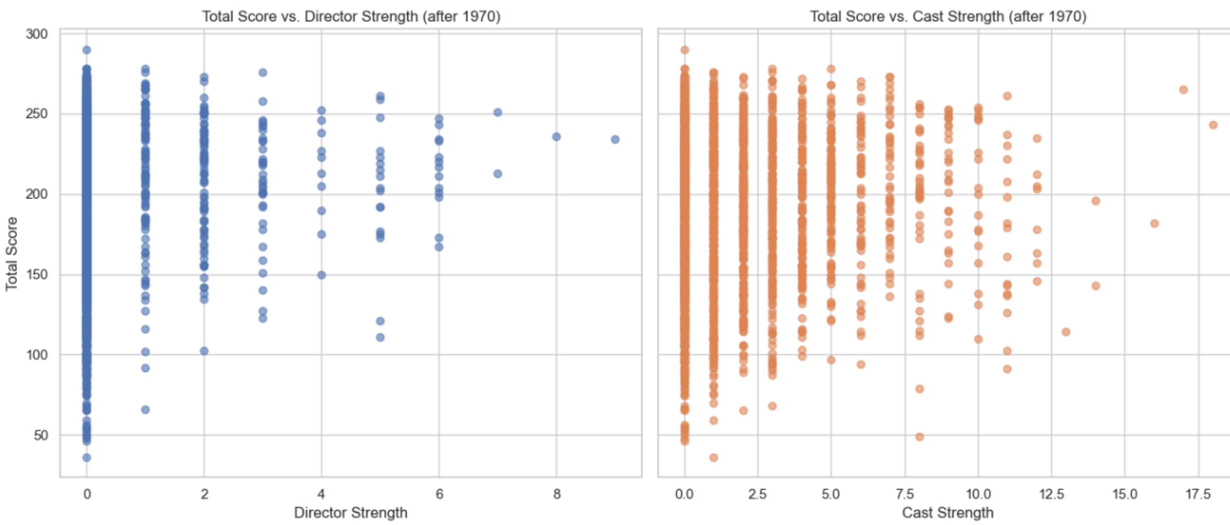
Decade Released	count	mean	median
1920-1929	1	278.000000	278.0
1930-1939	5	264.800000	269.0
1940-1949	7	261.557143	273.0
1950-1959	6	237.400000	247.0
1960-1969	22	230.645455	242.5
1970-1979	48	230.250000	241.5
1980-1989	118	213.588983	224.0
1990-1999	279	199.098925	205.0
2000-2009	749	182.412951	186.0
2010-2019	990	189.558586	194.5
2020-2029	23	188.400000	196.0

Figure 1 – Question 3: Count, Mean and Median Total Scores per Decade

From this chart, we see that the median score and average score of movies steadily decreased overtime, along with most movies in the dataset being released from 2000 to 2020. This could be a product of many factors, most likely the availability of historical data. While less movies were released prior to 2000, the datasets we used were likely only stored the most prevalent ones, as there wasn't data of the lesser-known movies of the time. For this reason, we decided to use data after 1970 in our analysis to be more representative of modern movie culture.

It's also important to note how 86 movies that had a cast strength or director strength greater than 0 were not included in this analysis question, as they did either have a null total score, box office revenue, or both. Roughly 68% of the movies being removed were released after 2020, showing how the website we used did not have box office or score data collected at the time of scraping.

Now that we analyzed potential errors, we started analysis by comparing scatterplots of director strength and cast strength against box office, looking for any initial trends or insights




```
--------------------------------------------------
Correlation and P-value after 1970
Pearson Correlation Coefficient: 0.116
P-value for Director Strength vs. Total Score correlation test: 4.60129570566695e-09
Pearson Correlation Coefficient: 0.061
P-value for Cast Strength vs. Total Score correlation test: 0.002136890617815616
```

Figure 2 – Question 3: Director Strength and Cast Strength vs Total Score

From these charts, we noticed a positive skew down the x axis. It seemed that as strength increased, the range of scores decreased, signaling higher performance was more likely for these movies, especially those with a strong director. To confirm our hypothesis, we found the correlation coefficient and p-value of each chart. While the correlations were lower than our expectations (0.116 and 0.061 for director strength and cast strength respectively), the p-values indicated both relationships were statistically significant. Director Strength vs Total score has a lower p-value and higher correlation, indicating a stronger and more significant relationship. To get more insight into the distribution of scores, we created a box plot, shown below:

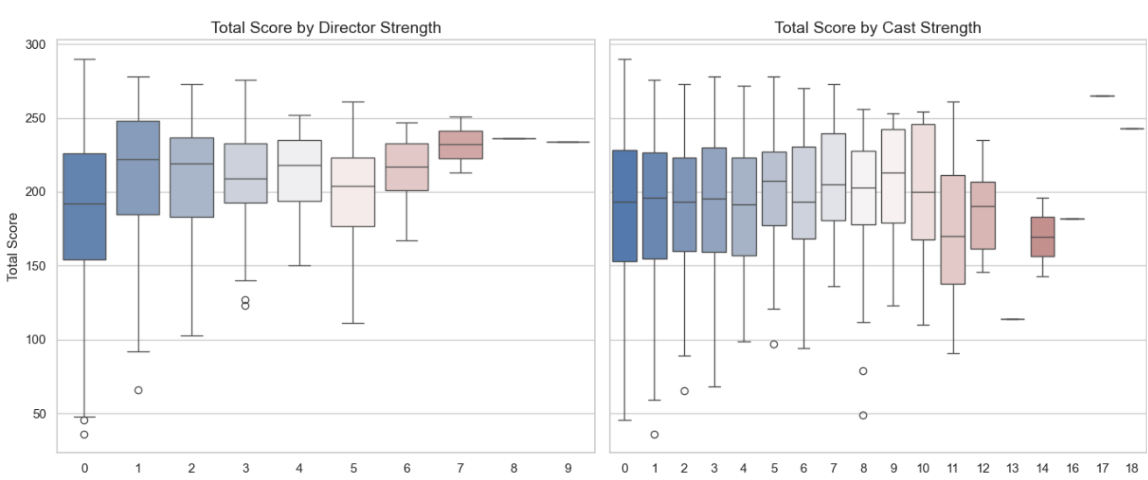
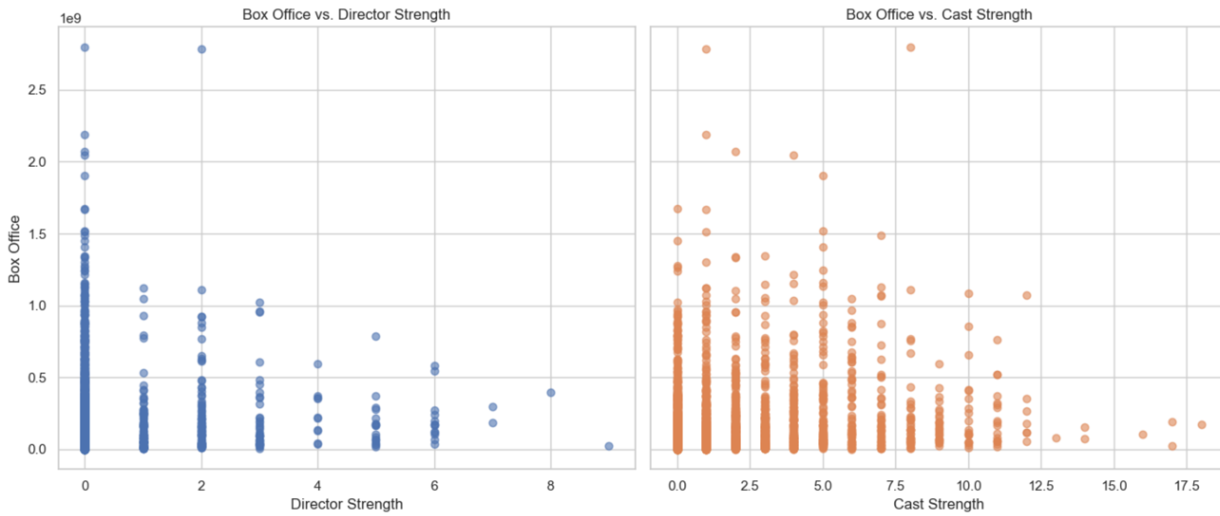


Figure 3 – Question 3: Score Distribution for Director and Cast Strength

From this chart, we can see that as Director Strength increases, median scores slightly increase and the range of scores shortens, indicating a higher and more predictable total score. For Cast Strength, we see median score stay constant around 200, with the range remaining wide for all levels of strength.

We conducted the same analysis for box office revenue by replacing it with the total score as the Y axis, looking to see trends amongst box office revenue generated and a strong cast or director.




```
-----------------------------------------------
Correlation and P-value after 1970
Pearson Correlation Coefficient: 0.043
P-value for Director Strength vs. Box Office correlation test: 0.03224201379174098
Pearson Correlation Coefficient: 0.149
P-value for Cast Strength vs. Box Office correlation test: 4.3667850379269335e-14
```

Figure 4 – Question 3: Director Strength and Cast Strength vs Box Office

Unlike with total score, the graphs indicated a downward sloping relationship between box office and director strength. Surprisingly, the p-values for both relationships indicated a minimal but positive relationship, likely due to the proportion of movies having a low strength rating. Both charts also had p-values less than the alpha of 0.05, indicating a significant relationship. Unlike the comparison against total score, cast strength had a higher correlation coefficient and p-value, indicating a stronger and more significant relationship.

<u>Machine Learning</u>

To further prove our hypothesis that a strong cast and director impacted total score and box office revenue, we used machine learning by conducting 2 linear and 2 logistic regressions. Our first logistic regression used Total Score as the dependent variable and cast and director strength as independent variables. When training the data on a 20% split, we yielded an r-squared coefficient of just 0.012.

However, our next model used a logistic regression, splitting total score into 3 quantiles “bad,” “average,” and “good,” classifying the column as “score_level.” Using a score level of “good” (top 33% of total score ratings) as the dependent variable and the same independent variables before, we yielded a pseudo r-squared of 0.944. This indicates a strong and positive relationship, proving that a casts and directors that have previously won awards tend to output “good” movies.

We followed this same exact process with Box Office as the dependent variable. The second linear regression yielded an r-squared of 0.028, which was expected based on the results of the first linear regression. For the second logistic regression, we followed the same steps of bucketing box office into 3 quantiles and using “good” box office revenue as the dependent variable. Once again, the logistic regression yielded a high r-squared value of .945, indicating a strong and positive relationship.

Overall, conducting machine learning showed us the difference between fitting regressions to appropriate data types. The strength of a cast or director will be difficult to indicate fluctuations in the numeric score or box office. However, these variables do have a very positive impact on determining movies with a total score or box office better than 66% of the population.

<u>Text Analytics</u>

We also conducted text analytics to investigate if movies have similar titles. Using ‘the’, ’a’, and ‘and,’ as stop words, we created the word cloud below:

Figure 5 - Question 3: Word Cloud of Titles

Visually, the word cloud highlights dominant title motifs superhero “Man” films, “Last”/ “Day” adventures, and recurring themes around “World,” “Love,” and “Night.”

Conclusion

1. Has the average runtime of movies changed over different decades and do movies with longer runtimes tend to have higher or lower box office earnings?
   The average runtime of movies has significantly increased over the decades, as confirmed by statistical testing and visual analysis, with recent films showing greater variability and longer lengths. While there is a weak but statistically significant positive correlation between runtime and box office earnings, the relationship lacks practical strength. Both regression and classification models confirm that runtime alone is a poor predictor of box office success, suggesting that other factors play a far more influential role.

2. What is the correlation between movie ratings on platforms like IMDb, Metacritic, and Rotten Tomatoes and box office performance?
   Despite strong internal consistency among IMDb, Metacritic, and Rotten Tomatoes scores (with correlations as high as 0.96) and a well-aligned Total Score that effectively aggregates them, all rating metrics show very weak correlations (~0.05–0.06) with Total Gross earnings. This suggests that while ratings consistently reflect similar evaluative signals, they are not reliable predictors of box office success, likely due to the influence of other external factors.

3. Does the presence of award-winning actors/directors influence box office or total score?
   On the surface, there is a weak but significant correlation between the presence of award-winning actors/directors and box office and total score. But, when conducting logistic regression, we see award winning actors and directors each have a strong influence on producing movies with a good (top 33%) box office and total score. Directors have a stronger impact on total score and actors have stronger impact on box office, suggesting that the critics look to directors for producing a quality movie while the casual movie goer enjoys following the big names.

Overall, our analysis found that while factors like runtime, ratings, and the presence of award-winning talent show statistically significant relationships with box office earnings, none emerged as strong predictors on their own. Runtime and star power had weak positive correlations, and ratings showed almost no predictive power for financial success. This suggests that external elements like marketing, franchise popularity, and distribution strategy likely play a much greater role in determining a film's commercial outcome. In the future, exploring the impact of streaming platform releases, marketing budgets, and social media engagement on both box office success and critical reception, as these modern factors likely play a significant role in a film's performance.