Analyzing IMDb Ratings and Box Office Success: Insights from Top-Grossing Movies

Nathan Lokenvitz

Github Repository: https://github.com/NateLokenvitz/data-analytics-template

1. Introduction

The relationship between a movie's critical reception and its box office performance has long intrigued filmmakers and analysts alike. This project explores this connection by analyzing the IMDb ratings, Metascores, and worldwide lifetime gross revenue of top-grossing movies. Using data from IMDb's Top 100 Movies and a dataset of the Top 1,000 Highest-Grossing Movies, this analysis aims to uncover key insights into how audience ratings and critical acclaim relate to financial success. Specifically, the study addresses three primary research questions: the correlation between IMDb ratings and box office revenue, the relationship between IMDb ratings and Metascores, and whether higher IMDb ratings translate to higher lifetime gross revenue. Through data cleaning, filtering, and exploratory visualizations, this report provides a comprehensive overview of trends and relationships in movie success metrics.

2. Data

This project combines two primary datasets: IMDb's Top 100 Movies and Kaggle's Top 1,000 Highest-Grossing Movies, to explore the relationship between audience ratings, critical acclaim, and box office success.

2.1 IMDb Top 100 Movies Data

The IMDb Top 100 dataset was created by writing a custom web crawling and scraping script to collect the desired information from IMDb's webpages. The script extracted the movie title, year of release, runtime (duration), Metascore, and star rating for each movie. These details were saved in a data frame and exported as a CSV file, which is located in the raw folder under the data tab on Github. Minor inconsistencies in this dataset, such as extra spaces in movie titles and inconsistent year formats, were addressed during the cleaning process to ensure compatibility with the second dataset.

2.2 Top 1,000 Highest-Grossing Movies Data

The Kaggle dataset contains information on the top 1,000 highest-grossing movies of all time, including fields such as movie title, year of release, genre, IMDb user rating, and worldwide lifetime gross revenue. This data is saved in under the raw folder in the data tab on Github. Like the IMDb Top 100 dataset, this data was cleaned to ensure consistency, with unnecessary columns removed

and essential fields such as Worldwide Lifetime Gross and Genre retained for analysis. Additional cleaning steps included standardizing the case of movie titles and ensuring release years were formatted consistently as four-digit strings.

2.3 Data Cleaning and Merging

To create a unified dataset for analysis, the two datasets were merged using an inner join on Movie Title and Year of Release. This step ensured that only movies present in both datasets were included in the final analysis. After merging, the following fields were retained: movie title, year of release, duration, Metascore, IMDb rating, genre, gross, and worldwide lifetime gross. These columns were renamed for clarity, such as renaming Duration_x to Duration and Star Rating to IMDb Rating. The cleaning and merging steps were performed in Python, and the resulting code is located in the final folder under the data tab on Github.

2.4 Additions During Exploratory Data Analysis

During the exploratory data analysis (EDA) phase, two additional columns were added to better address the research questions. The Decade column grouped movies into their release decades (e.g., 1980s, 1990s) and was used to analyze revenue trends over time. The Rating Band column categorized IMDb ratings into bands (<7, 7-8, 8-9, 9+) to assess how ratings correlated with financial performance. These enhancements were implemented in and made into the final dataset which is located in the final folder under the data tab.

2.5 Final Dataset

The final cleaned dataset contains 36 rows, including the header row, and is stored in the final folder under the data section of my Github repository. It includes key fields required for analysis, such as movie title, year of release, runtime, Metascore, IMDb rating, genre, worldwide lifetime gross, decade, and rating band. While additional metrics such as the number of votes from IMDb's dataset were considered, they were excluded to simplify the analysis and focus on audience and critic scores. This final dataset forms the basis for the visualizations and analyses performed to answer the research questions.

Data Dictionary

Column	**Type**	**Description**
Movie Title	Text	The name of the movie.
Year of Release	Numeric	The year the movie was released.
Genre	Text	Categories where the movie belongs.
Duration	Numeric	Movie running time in minutes.

Gross	Numeric	Gross earnings in U.S. dollars.
Worldwide LT Gross	Numeric	Worldwide Lifetime Gross (International + Domestic totals).
Metascore	Numeric	Weighted average of many reviews from reputed critics (on a scale of 0 to 100).
IMDb Rating	Numeric	Ratings given by IMDb registered users (on a scae of 1 to 10).
Decade	Text	The decade in which the movie was released (ex. 1980s 1990s, etc.).
Rating Band	Text	IMDb ratings categorized into bands (<7,7-8,8-9,9+).

3. Analysis

3.1 Is there a correlation between IMDb Ratings and Box Office Revenue?

To determine if there is a correlation between IMDb ratings and worldwide box office revenue, I began by creating a scatterplot of IMDb Rating on the x-axis and Worldwide LT Gross on the y-axis, called Figure 1. The initial correlation coefficient was calculated as **-0.075**, indicating a very weak negative correlation. This suggested little to no linear relationship between audience ratings and box office revenue. The scatterplot further reinforced this conclusion, as the data points appeared scattered without any clear upward or downward trend.

However, the presence of extreme outliers, particularly movies with exceptionally high revenues, skewed the visualization and correlation results. To address this, I applied the **Interquartile Range (IQR)** method to filter out outliers from the Worldwide LT Gross column:

1. The 25th percentile (Q1) and 75th percentile (Q3) of the Worldwide LT Gross were calculated.
2. Outliers were defined as values outside **1.5 times the IQR** below Q1 or above Q3.
3. A filtered dataset was created, removing these extreme values to improve the robustness of the analysis.

Using this filtered dataset, I calculated the updated correlation coefficient, which increased to **0.138**. While still weak, the positive value indicates a slight upward trend, suggesting that movies with higher IMDb ratings may, on average, generate higher box office revenue. This trend is visually represented in the second scatterplot, Figure 2, where a regression line highlights a modest positive relationship. This visual representation of the relationship between IMDb ratings and box office revenue shows that having a higher IMDb rating may actually have a slight positive impact on the

worldwide lifetime gross value of a movie. This is in line with common reason, as one would think that the better a movie is perceived by the audience and critics, the more money it has the chance to make.

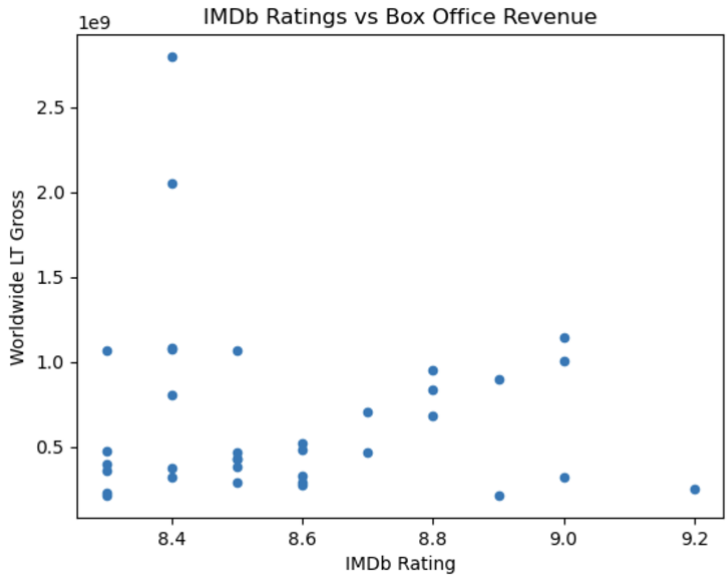

Figure 1

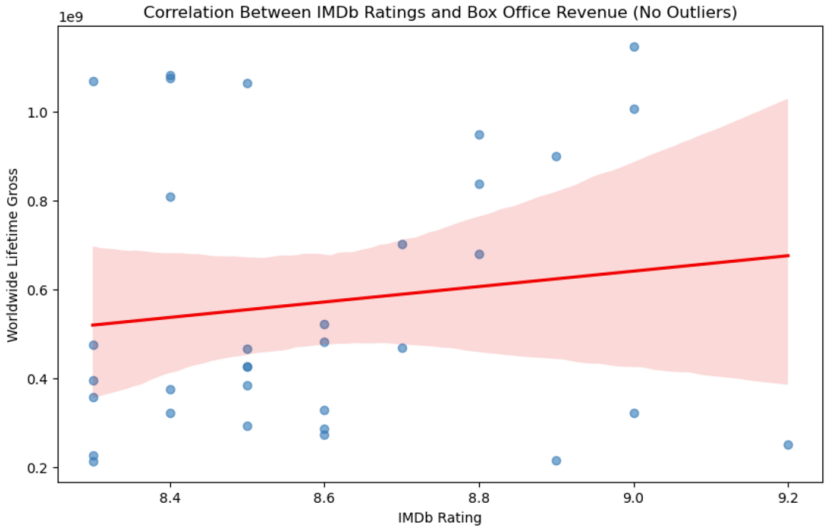

Figure 2

The initial correlation of **-0.075**, which included outliers, indicated no meaningful relationship between IMDb ratings and box office revenue. However, after removing outliers, the correlation

increased to **0.138**, suggesting a small but noticeable positive trend: movies with higher IMDb ratings tend to achieve slightly higher box office revenues. Despite this improvement, the scatterplot highlights significant variability, with movies of similar IMDb ratings showing vastly different revenue figures. This suggests that while IMDb ratings may play a role in a movie's financial success, other factors such as marketing efforts, genre appeal, or star power likely have a larger influence. By filtering the data and recalculating the correlation coefficient, this analysis provides a clearer and more accurate understanding of the relationship between audience ratings and box office performance.

3.2 What is the relationship between Metascore and IMDb Star Ratings?

To explore the relationship between Metascore (critic ratings) and IMDb star ratings (audience ratings), I created two visualizations: a **scatterplot** and a **heatmap**. Together, these visualizations provide good insight as to how critic and audience scores are related.

I started off with the visualization being a scatterplot that plots Metascore on the x-axis and IMDb Rating on the y-axis. This plot provides a direct comparison of the two ratings for each movie in the dataset. The correlation coefficient between Metascore and IMDb Rating was calculated as **0.453**, indicating a moderate positive correlation. This means that as critic scores (Metascore) increase, audience ratings (IMDb Rating) tend to increase as well.

From the scatterplot:

- Movies with **higher Metascores (80-100)** generally have IMDb Ratings clustered between **8.5 and 9.2**, suggesting that movies highly rated by critics are also well-received by audiences.
- However, there is noticeable spread for movies with **lower Metascores (below 70)**. Some movies with lower Metascores still achieve relatively high IMDb Ratings, which may indicate a disconnect between critical reviews and audience perception for certain films.
- There are no strong outliers, and most points align with the positive trend, reinforcing the moderate correlation.

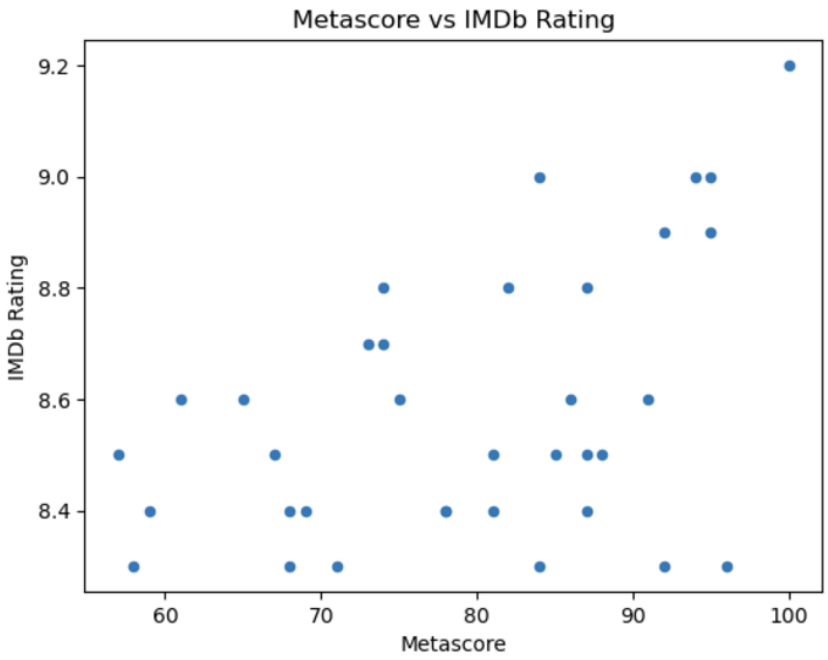

Figure 3 Metascore vs. IMDb Rating Scatterplot

To complement the scatterplot and better understand the density of ratings, I created a **heatmap** using a kernel density estimate (KDE). This visualization highlights where the majority of the Metascore and IMDb Rating data points are concentrated.

From the heatmap:

- The darkest region (highest density) appears around **Metascores of 75-90** and **IMDb Ratings of 8.4-9.0**. This indicates that most movies in the dataset receive strong critic reviews (Metascores) and are also highly rated by audiences.
- As the Metascore decreases below 70, the density of IMDb Ratings spreads out, confirming the variability seen in the scatterplot. Lower Metascores do not always translate to poor audience ratings.
- Interestingly, there are no high-density regions where Metascores are extremely high (above 90) and IMDb Ratings are low, reinforcing the notion that critically acclaimed movies are typically well-received by audiences.

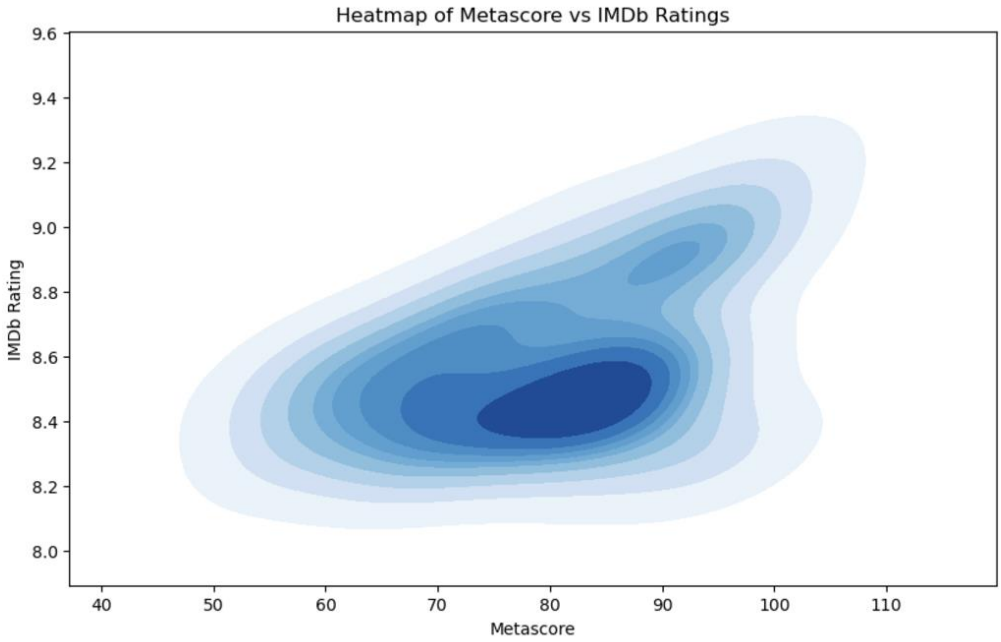

Figure 4 Heatmap of Metascore vs. IMDb Ratings

The correlation coefficient of **0.453** confirms a moderate positive relationship between Metascore and IMDb Ratings, suggesting that movies rated highly by critics tend to be appreciated by audiences as well. The scatterplot highlights this upward trend, with a clear clustering of movies that have both high Metascores and high IMDb Ratings. This pattern is further reinforced by the heatmap, which shows that the majority of movies fall within the range of **75-90 for Metascore** and **8.4-9.0 for IMDb Ratings**, indicating strong agreement between critics and audiences for most top-performing movies. However, variability observed at lower Metascores suggests that while critic reviews play an influential role, audience ratings may sometimes reflect other factors such as entertainment value, star power, or genre appeal. Overall, the analysis demonstrates a positive correlation between Metascore and IMDb Ratings, though the relationship remains moderate. Further exploration into specific movies or genres could provide deeper insights into the discrepancies seen at lower Metascores.

3.3 Does a higher IMDb Star Rating translate to a higher Lifetime Gross?

To investigate whether higher IMDb star ratings lead to higher worldwide lifetime gross revenue, I analyzed the data using two visualizations: a **bar chart** and a **boxplot**. These charts provide both aggregated averages and the spread of revenue across different IMDb rating bands.

The first visualization is a bar chart, Figure 5, that groups movies into IMDb rating bands and calculates the average Worldwide Lifetime Gross for each group. The IMDb ratings were divided into four bands: <7, 7-8, 8-9, and 9+.

From the bar chart (Figure 5):

- The **8-9 rating band** shows the highest average lifetime gross, significantly surpassing all other bands. Movies in this category have an average revenue close to **$700 million**.
- Interestingly, movies in the **9+ rating band** have much lower average revenue compared to the 8-9 band. This could be due to the rarity of movies achieving such high ratings and the possibility that these movies are critically acclaimed but may not appeal to broader audiences for commercial success.

  This bar chart effectively highlights the trend that movies with IMDb ratings between **8 and 9** tend to generate the highest average lifetime gross. It also is worth noting that in the final dataset, there are only 35 movies out of the top 100 ranked on IMDb that made it into the top 1000 highest grossing movies of all time. Out of those 35 none have a rating below 8.3, meaning that movies with higher IMDb ratings are more likely to achieve both critical acclaim and significant box office success. This suggests that for a movie to rank among the top-grossing films of all time, it typically needs to be well-received by audiences, as reflected by an IMDb rating of 8.3 or higher.

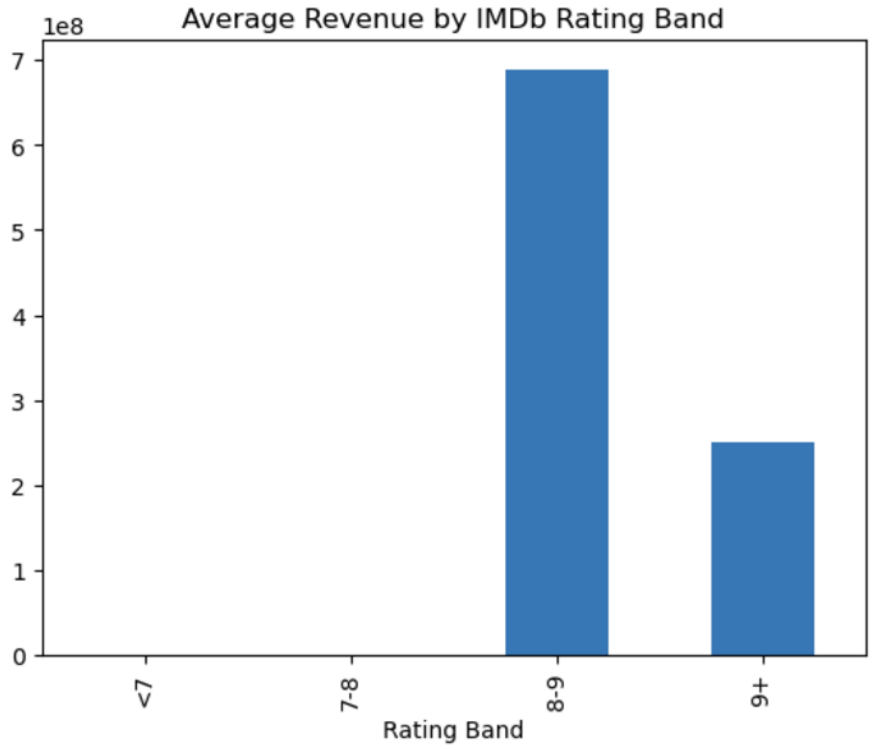

Figure 5 Average Revenue by IMDb Rating Band

The second visualization, Figure 6, is a boxplot that shows the distribution of worldwide lifetime gross across the same IMDb rating bands. Unlike the bar chart, the boxplot reveals additional details about the spread, median, and presence of outliers.

From the boxplot:

- The **8-9 rating band** not only has the highest median revenue but also a relatively wide spread, indicating that movies in this group achieve a range of revenue figures, with several movies surpassing $1 billion.
- Outliers are observed in the **8-9 band**, with a couple of movies achieving exceptionally high lifetime gross revenues. These outliers likely include blockbuster hits that skew the average revenue upward.
- The **9+ band** shows a much lower median and little variability, which confirms the trend from the bar chart. Movies with such high ratings tend to have modest box office performance despite their critical acclaim.

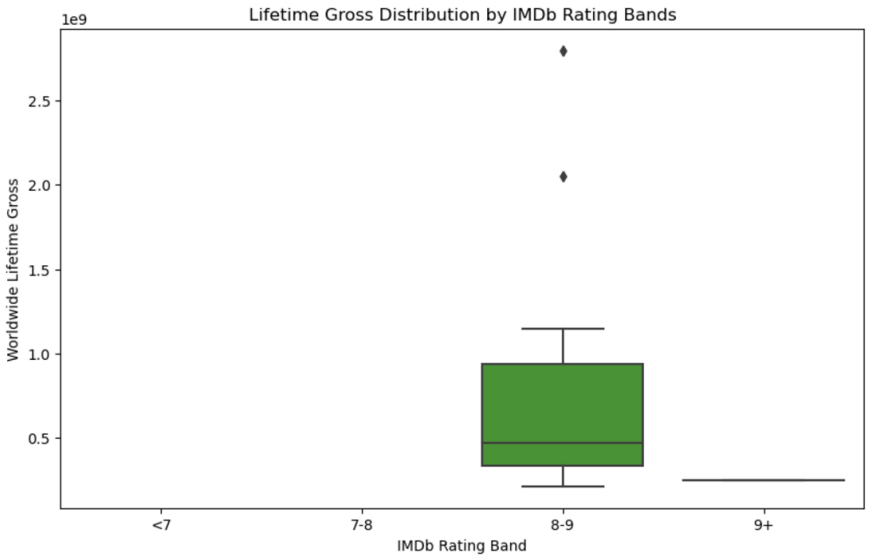

Figure 6 Lifetime Gross Distribution by IMDb Rating Bands

The combination of the bar chart and boxplot provides a clear answer to the research question: while higher IMDb ratings generally correspond to higher lifetime gross, this relationship is not perfectly linear. Movies in the **8-9 rating band** generate the highest average and median revenues, suggesting that this group strikes an optimal balance between critical acclaim and broad commercial appeal. In contrast, movies in the **9+ rating band**, though rare and highly rated, tend to achieve lower average revenues. This pattern indicates that extremely high ratings may reflect niche appeal rather than widespread popularity. Finally, lower-rated movies (**<8**) consistently underperform in terms of lifetime gross, both in their average revenue and overall distribution, reinforcing the idea that movies with lower audience ratings are less likely to achieve significant financial success.

4. Conclusion

In this project, I analyzed three key aspects of movie performance: the relationship between IMDb ratings and box office revenue, the relationship between critic scores (Metascore) and audience ratings (IMDb Rating), and how IMDb star ratings relate to a movie's lifetime gross revenue. From the research questions presented in my proposal, I found the following results:

1. **Is there a correlation between IMDb ratings and box office revenue?**
   There is a very weak relationship between IMDb ratings and worldwide lifetime gross revenue when outliers are included, as indicated by the initial correlation coefficient of **-0.075**. After removing outliers using the Interquartile Range (IQR) method, the correlation increased slightly to **0.138**, revealing a small but noticeable positive trend. This suggests that while higher IMDb ratings may contribute to a movie's financial success, other factors such as marketing, genre, or star power likely play a more significant role.

2. **What is the relationship between Metascore and IMDb star ratings?**
   The analysis revealed a **moderate positive correlation** of **0.453** between Metascore and IMDb Ratings. This suggests that movies highly rated by critics (Metascore) tend to also be appreciated by audiences (IMDb Ratings). The scatterplot showed an upward trend, while the heatmap highlighted that most movies clustered within the ranges of **75-90 for Metascore** and **8.4-9.0 for IMDb Rating**, indicating a general alignment between critic and audience scores. However, variability at lower Metascores suggests that audience ratings are influenced by additional factors beyond critical reviews, such as entertainment value or fan engagement.

3. **Does a higher IMDb star rating translate to a higher lifetime gross?**
   The analysis demonstrated that movies in the **8-9 rating band** generate the highest average and median revenues, as seen in both the bar chart and boxplot. This rating range appears to strike a balance between critical acclaim and broad commercial appeal. In contrast, movies with ratings in the **9+ band**, while rare and highly rated, tend to achieve lower average revenues, suggesting niche appeal rather than widespread popularity. Movies with ratings below 8 consistently underperform in terms of lifetime gross, reinforcing the trend that higher ratings generally align with better financial performance.

This project does have some limitations. The final dataset contains only 35 movies from IMDb's Top 100 list that also appear in the Top 1,000 Highest-Grossing Movies of All Time dataset. Additionally, none of these movies have an IMDb rating below 8.3, which may limit the generalizability of conclusions to lower-rated or less commercially successful films. Another limitation is the exclusion of additional variables, such as the number of votes cast (indicating popularity) and star power, which may provide deeper insights into box office performance.

Future work on this project could include expanding the dataset to incorporate movies beyond the Top 100 IMDb list or the Top 1,000 grossing movies, as well as incorporating additional variables such as production budget, genre, release season, and marketing spend. This would allow for a

more comprehensive understanding of the factors influencing both critical and financial success in movies.