Project Report
IMDb's Influence on the Box Office

1. Introduction:

In 2024, over 3,000 English-language films will be released, according to IMDb's advanced title search[1]. In a landscape where our attention and dollars are limited, many utilize IMDb to sift through content and find the diamonds in the rough. IMDb allows users to quickly gain insight into how much the general audience and critics enjoy a given movie, but do high review scores always mean more tickets sold? Specifically, do movies with high earnings at the end of their theater runs benefit from high reviews from critics and the audience? And based on that question, are critic or user scores better predictors of box office sales? With attention spans presumably getting shorter, one can also wonder if a movie's running length affects revenue. Fortunately, IMDb offers more information than just reviews; one may also investigate parameters like runtime to analyze the box office.

In this project, I collected data from IMDb on Kaggle[2] to explore the correlation between a movie's scores and financial performance, through web scraping Box Office Mojo[3]. I also took a look at duration times on IMDb and to determine whether general audiences prefer shorter movies. Joining these two datasets allowed me to gain insight into IMDb's relevance in the movie industry and analyze whether these factors are vital for a movie's success.

2. Data:

This project uses two primary sources of data: IMDb's movie critic and audience review scores of different movies and Box Office Mojo data containing box office information.

2.1 Box Office Details

[1] IMDb Advanced Title Search

[2] Kaggle IMDb Dataset

[3] Box Office Mojo

In this part of the project, I focused on scraping data from Box Office Mojo, targeting worldwide, domestic, and foreign box office earnings of the top 200 movies from 2020 to 2023. Utilizing the R packages: rvest, dplyr, and readxl, I extracted relevant data from the Box Office Mojo website.

The scraping process began by identifying the table containing the box office information on the website. Using Google Chrome's inspect tool, I examined the HTML structure to locate the table's class name, which was "a-bordered." This specific class contained the data I needed. To ensure all table rows and columns were captured, I set the fill = TRUE argument. This flexibility is useful when scraping different years, as table structures might vary slightly.

After scraping data from 2020 to 2023, I combined the individual data frames using bind_rows(). The combined data frame was stored in the boxofficemojo_data variable. Given that the raw data sometimes contains inconsistent or placeholder values like "-", I replaced such values with NA for easier handling using the mutate_all() function with an ifelse() condition. To ensure consistency, I also trimmed leading and trailing whitespace from movie names with trimws(). Additionally, I converted any textual numeric data (like percentages) to proper numeric formats using mutate() to facilitate further analysis.

2.2 IMDb Ratings

For this project, the IMDb dataset from Kaggle provided crucial information on movies released between 2020 and 2023. This information included data points such as movie title, IMDb rating, vote count, Metascore, and duration for each movie. Before integrating the IMDb data with my scraped Box Office Mojo dataset, I cleaned it using Excel. I removed unnecessary columns like genre, PG rating, cast, and director, which were irrelevant to my analysis. Additionally, I trimmed the rows to focus only on movies released between 2020 and 2023. To ensure consistency and facilitate a seamless merge with other datasets, I renamed the columns to have a uniform naming convention.

During a project check-in, feedback prompted me to convert the duration column from an hour-and-minute format to a purely minute-based format using Excel for easier interpretation.

With the IMDb dataset now cleaned and prepped, I was ready to proceed with data integration, combining it with the cleaned Box Office Mojo data for comprehensive analysis.

2.3 Combining Box Office and Ratings

Merging the two datasets was completed using the merge() function with a left join. This strategy keeps all records from the Box Office Mojo dataset, even if there are no corresponding matches in the IMDb data. After merging, I noticed that some fields, particularly Meta.Score, contained missing values (NA). To address this, I calculated the mean of the existing Meta.Score values and used mutate() to fill in the gaps with this mean value. This approach helps maintain consistency in the dataset by providing a statistically grounded method for imputing missing values.

To clean the data further, I used rowSums(is.na(merged_data)) to identify rows with any missing values and retained only those with no NA values. This step ensured that the final dataset was complete and consistent for further analysis. Additionally, I removed the duration column with the outdated format and the ranked column, which were deemed unnecessary for the final output.

Finally, I saved the cleaned data to a new CSV file, merged_data_clean.csv. This action completed the data integration and cleaning process. A description of each variable from the final dataset is shown in Table 1.

Table 1 Data Dictionary

Column	Type	Source	Description
Movie.Name	text	Both	The title of the movie
Worldwide	numeric	Box Office Mojo	Combined domestic & foreign box office sales
Domestic	numeric	Box Office Mojo	The box office sales in the United States
Domestic.Percentage	numeric	Box Office Mojo	The percentage of total sales, U.S.
Foreign	numeric	Box Office Mojo	The box office sales internationally

Foreign.Percentage	numeric	Box Office Mojo	The percentage of total sales, international
Rating	numeric	Kaggle	The IMDb user rating of the movie
Votes	numeric	Kaggle	The number of ratings a given movie has on IMDb
Metascore	numeric	Kaggle	The review score as decided on by the critics
Year	numeric	Kaggle	The movie's year of release
Duration	numeric	Kaggle	The length of the movie

3. Analysis

3.1 Movie Ratings and Box Office

The purpose of this analysis is to investigate whether there is a strong correlation between a movie's IMDb rating and box office performance. To explore this question, I calculated the correlation coefficients among several key metrics in a movie dataset, including IMDb ratings, box office revenues, and other related variables.

The correlation between a movie's IMDb rating and its worldwide box office revenue is 0.222, indicating a relatively weak positive relationship. This result suggests that while higher-rated movies may tend to perform better at the box office, the correlation is not particularly strong. A similar trend is observed with domestic box office revenue, where the correlation with IMDb rating is 0.258, also suggesting a modest positive relationship. Meanwhile, the correlation between box office revenues and other variables, such as domestic and foreign revenue, is notably stronger. For example, the correlation between worldwide box office revenue and domestic revenue is 0.960, pointing to a very strong relationship. This high correlation might be due to the considerable proportion of domestic revenue in total box office earnings.

Overall, these findings suggest that the relationship between a movie's IMDb rating and its box office performance is not particularly strong. While there may be a slight positive correlation, other factors likely play a more significant role in determining a movie's commercial success. These factors might include marketing efforts, star power, or production budgets. The data indicates a weak to moderate positive correlation, suggesting that while higher ratings might

contribute to better box office performance, they are not definitive predictors. Further research into other contributing factors could provide additional insights into what drives box office success.

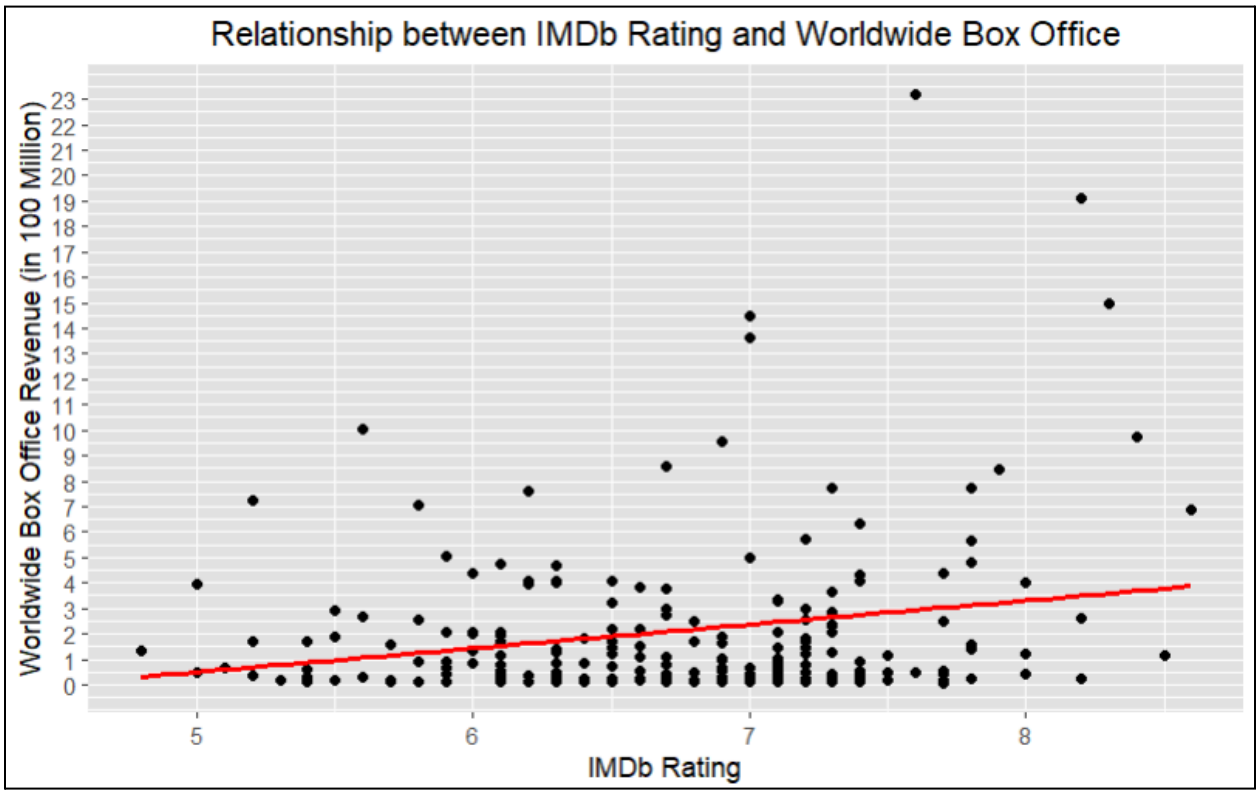

Figure 1: Linear Regression Model

The linear regression model I created between IMDb ratings and worldwide box office revenue depicted in Figure 1 shows a statistically significant relationship but with a weak correlation. The positive slope of the regression line suggests that higher IMDb ratings are generally linked to higher worldwide box office revenue. However, the adjusted R-squared value of 0.045 indicates that IMDb ratings explain only a small fraction of the variability in box office revenue, suggesting that other factors have a more significant impact on a movie's financial success. The high residual standard error of about 311.5 million reflects considerable dispersion in the data. Despite the statistical significance (p-value of 0.00126), the wide range of residuals indicates that the relationship between IMDb ratings and box office revenue is inconsistent, emphasizing the role of other contributing factors.

3.2 IMDb Rating vs. Metascore

Although we have established that there is generally a weak correlation between reviews and box office success, a deeper comparison of the influence of IMDb ratings and Metascore on box office performance can offer valuable insights. IMDb ratings typically reflect the collective opinion of audiences, while Metascore is derived from professional critics' reviews. By analyzing how these two distinct forms of review impact box office earnings, we can better understand the role that critical and audience acclaim play in driving revenue. The intention is to explore whether one set of reviews has a more significant impact on box office success than the other.

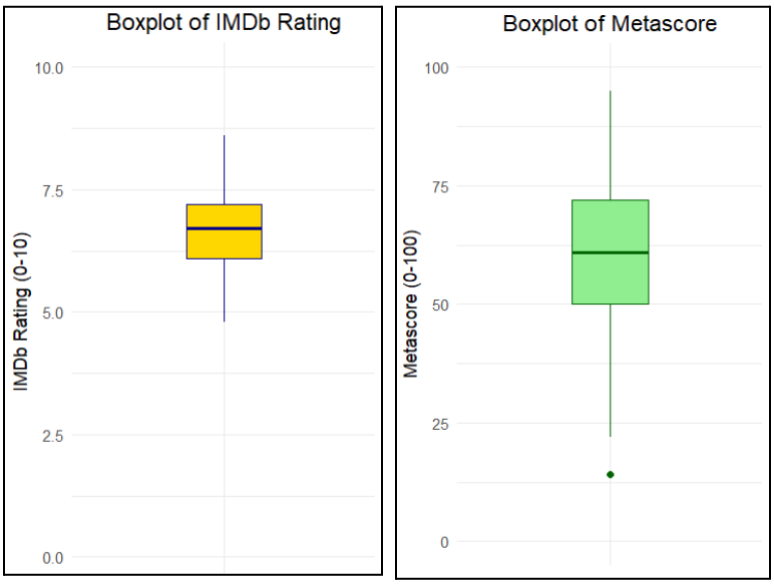

Figure 2: Review Box Plots

The boxplot of Metascore in Figure 2 shows a wider range of values compared to IMDb ratings, suggesting more variability in critic reviews. This wider distribution can impact correlation scores with other variables, such as Worldwide box office revenue. Metascore's broader spread reflects the diverse criteria critics use to review movies, potentially including factors like directing, acting, and storytelling. This diversity can lead to a weaker correlation with box office revenue because the critical scores may not consistently align with commercial performance.

Conversely, the boxplot for IMDb ratings shows a narrower range, indicating greater consistency among audience reviews. This consistency often results in stronger or more predictable

correlations with other variables. Referring to the correlation matrix I created, the correlation between IMDb ratings and Worldwide box office revenue is 0.222, suggesting a weak positive relationship. In contrast, the correlation between Metascore and Worldwide box office revenue is much lower at 0.054. The narrower range of IMDb ratings, representing a more unified perspective, might be why it shows a slightly stronger correlation with box office revenue.

3.3 Duration and Box Office

With people's attention spans seemingly getting shorter in the digital age, it's worth exploring whether the running length of a movie affects its revenue. To understand this dynamic, let's examine whether movie duration has a measurable impact on box office revenue. In this part of the analysis, we'll look at the correlation between movie duration (measured in minutes) and worldwide box office revenue.

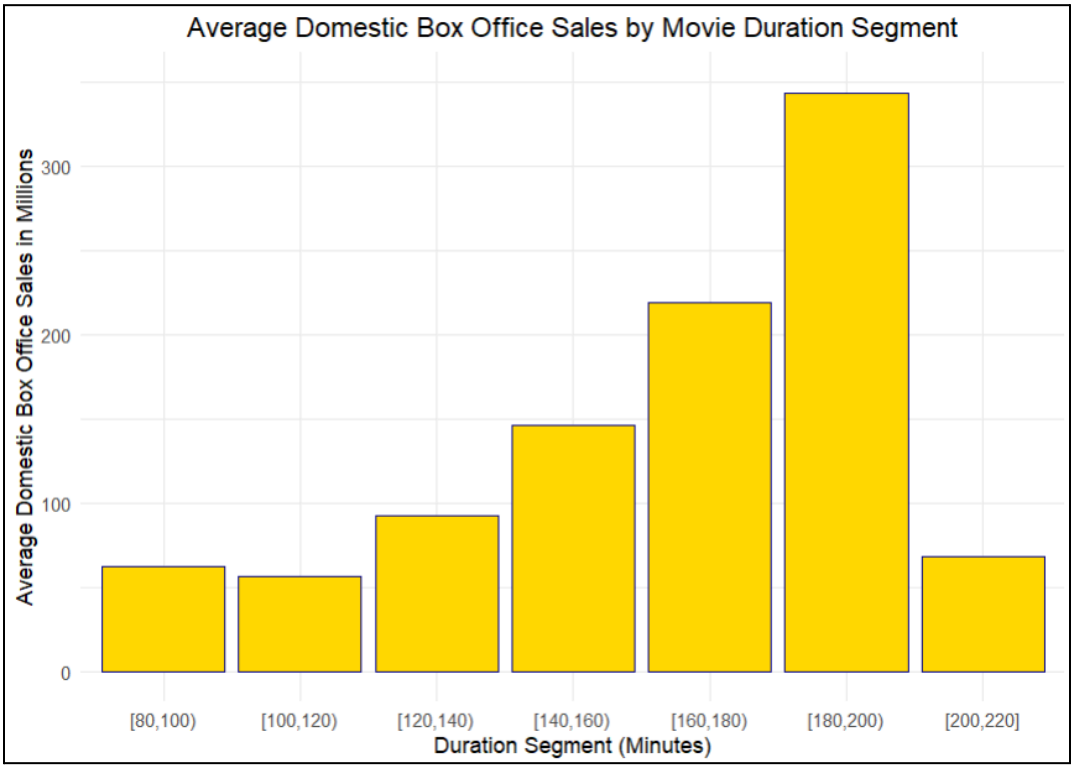

Figure 3: Duration Segment Bar Graph

Although shorter attention spans might suggest that longer movies struggle to perform at the box office, Figure 3 indicates otherwise. Among the various duration segments, movies that run between 180 and 200 minutes have the highest average domestic box office revenue.

To further investigate this trend, I conducted a correlation analysis and hypothesis testing to understand if there's a significant relationship between movie duration and Worldwide box office revenue. The correlation coefficient between Duration.Minutes and Worldwide box office revenue was approximately 0.37, indicating a moderate positive correlation. This correlation, along with the high average revenue for longer movies, reinforces the idea that longer films might perform better at the box office.

However, a correlation coefficient alone does not confirm statistical significance. In this case, the p-value from the correlation test was approximately 3.73e-08, significantly lower than the usual significance threshold of 0.05, indicating that this observed correlation is highly unlikely to be random. This provides strong evidence to reject the null hypothesis that there's no correlation between movie duration and box office revenue, suggesting that longer movies generally tend to perform better at the box office.

4. Conclusion

In this project, I analyzed the impact of IMDb ratings, Metascore, and movie duration on box office revenue. I used data from Box Office Mojo and IMDb to explore whether critical and audience reviews affect a movie's financial success, as well as the impact of movie duration on box office revenue. Below are the results from my analysis.

1. Is there a strong correlation between an IMDb rating and a movie's box office performance?

The correlation coefficient between IMDb ratings and Worldwide box office revenue is 0.222, indicating a weak positive relationship. Although there is a positive trend, the correlation is not particularly strong, suggesting that other factors besides user ratings contribute to a movie's box office success.

2. Is the critic score or the user score a better predictor of box office sales?

In comparing IMDb ratings (user scores) with Metascore (critic scores), the correlation between Metascore and Worldwide box office revenue is even lower at 0.054. This indicates that IMDb user ratings are a slightly better predictor of box office sales, as the correlation is stronger compared to critic scores.

3. Did the duration of the movie impact box office sales?

Yes, there is a statistically significant positive correlation between movie duration and Worldwide box office revenue, with a correlation coefficient of approximately 0.37. This suggests that longer movies tend to generate higher box office revenue. The p-value for this correlation is 3.73e-08, indicating strong evidence to reject the null hypothesis that there is no relationship between movie duration and box office revenue.

Despite the insights gained from this project, there were limitations. A significant constraint was data loss due to discrepancies between the Kaggle IMDb dataset and Box Office Mojo's list of the top 200 box office earners from 2020 to 2023. This inconsistency meant that some movies were included in one dataset but not in the other, leading to missing data that could affect the analysis's accuracy and generalizability.

4.1 Limitations and Future Work

The scope of analysis was also limited by the available variables. The dataset only provided the year of a movie's release, without specific dates, impacting the ability to study trends related to release timing and seasonality. Furthermore, the lack of information on marketing budgets and franchise affiliations restricted the analysis, as these factors can play a critical role in a movie's success at the box office. Future work could focus on incorporating additional variables, such as detailed release dates, franchise information, and marketing budgets, to build a more comprehensive model of box office success.