IMDB Movie Analysis: Exploring Factors Behind Movie Success

Sidda Patel and Mary Dorenbos

3250: Data Wrangling Fall 2024

12/17/2024

1.Introduction

The film industry plays a critical role in shaping global entertainment and culture. For over a century, movies have captured the imagination of audiences, driven by their creative narratives, production quality, and marketing strategies. Understanding the factors that contribute to a movie's success has profound implications for producers, marketers, and streaming platforms in an ever-evolving entertainment landscape.

Our project aims to analyze movies using two datasets: one from Kaggle, which provides comprehensive details on 1,000 top-rated IMDB movies, and another scraped from IMDB, which offers supplementary information. By merging these datasets and performing exploratory data analysis (EDA), we will investigate relationships between runtime, rating, genres, and other variables, uncovering trends in viewer preferences and industry practices.

This report documents our methodology, data preparation steps, and initial visualizations, providing a foundation for deeper insights.

2.Data Sources

2.1. Kaggle Dataset

The Kaggle dataset includes information on 1,000 top-rated movies and TV shows from IMDB. It contains 16 columns with details such as Series_Title, Released_Year, Runtime, Genre, IMDB_Rating, and Gross. This dataset provides a robust foundation for analyzing movie attributes and their impact on success metrics. In order to obtain this data, we did a deep dive into the Kaggle website to find the most comprehensive IMDB data set that would provide us with the needed data for us to derive certain insights. We came across this specific dataset and decided to move forward with it. We downloaded the csv file and uploaded it into our jupyter notebook as a dataframe titled ‘imdb_kaggle’.

2.2. Scraped IMDB Dataset

The scraped dataset supplements the Kaggle data with 92 movies, including details such as Title, Duration (mins), Content Rating, Genre, Rating, and Vote Count. These additional records

enrich the analysis and provide an alternative perspective on movie success. To obtain this data, we wrote a web scraping script in Jupyter Notebook that allowed us to extract the necessary details from the IMDB website. By analyzing the HTML structure of the pages, we identified the precise locations of the required data within the HTML elements. Using this approach, we successfully gathered and compiled the information into a DataFrame titled 'imdb_scraped'.

2.3. Combining Kaggle and Scraped

To enrich the analysis, the scraped IMDB dataset was merged with the Kaggle dataset. While the Kaggle dataset included 1,000 movies and the scraped dataset added 92 new records, the merge focused on combining all common columns available details into a single, comprehensive dataset. Prior to merging we had to drop and rename all columns that weren't common in both data frames to be able to vertically merge the two.

After merging, the combined dataset underwent a rigorous cleaning process to ensure uniformity and accuracy. Key cleaning steps included:

- Handling Missing Values: We ran a check for all columns to find any missing values in our final merged data frame and found none.
- Converting Data Types: Columns such as Duration (mins) and Rating were converted to numeric formats for consistency across the dataset.
- Standardizing Genres: Genre was condensed into just the first genre listed, rather than a list of all genres the movie categorizes into.

The cleaned dataset was exported as IMDB_Combined.csv for use in further analysis. A detailed data dictionary for the final dataset is provided in Table 1.

Table 1: Data Dictionary

Variable	Type	Description
Title	Text	Title of movie
Duration (mins)	Float	Length in minutes of movie
Genre	Text	Main genre of movie
Rating	Float	IMDB rating of movie
Vote Count	Integer	Total vote count of movie

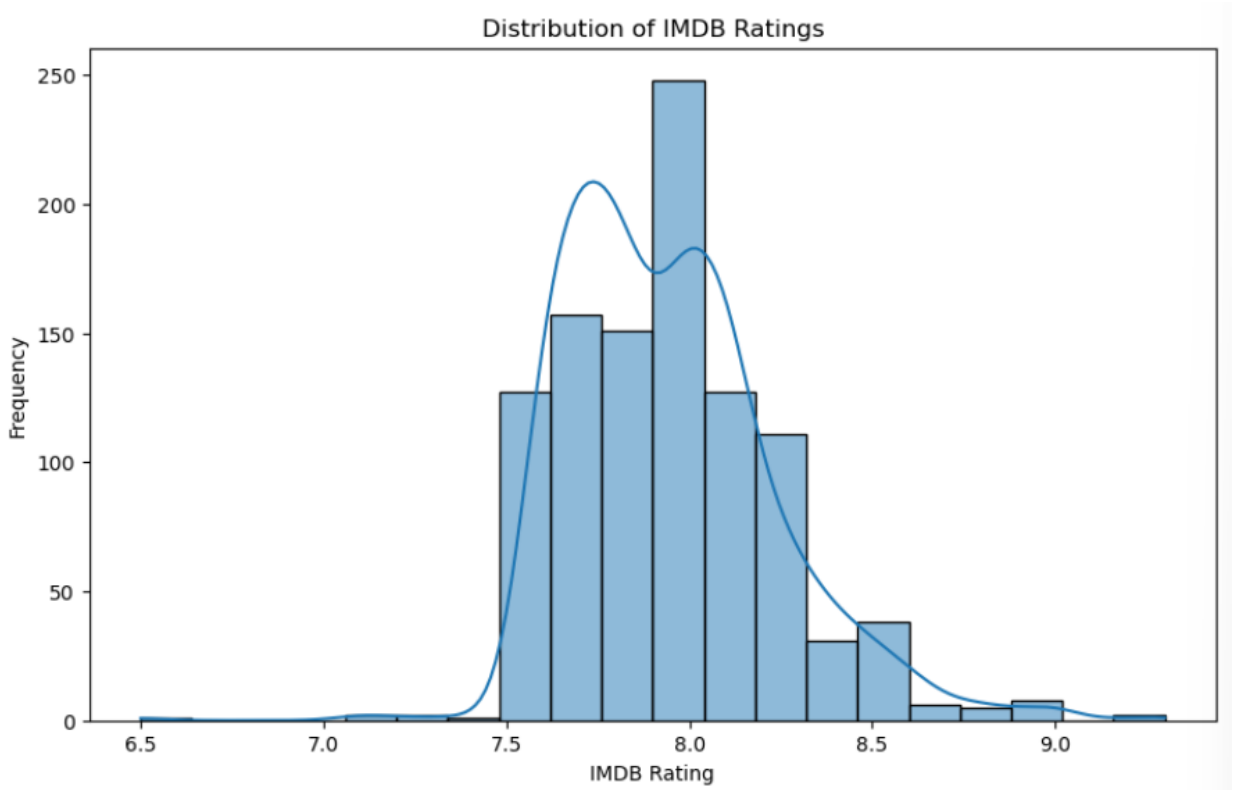
This comprehensive dataset ensures that all necessary information is consolidated and standardized, setting the stage for meaningful analysis of movie trends and success factors.

3. Analysis

In this analysis, we aim to explore several key questions that could offer valuable insights into how different movie attributes correlate with their success in terms of both audience reception and box office performance. First, we focus on understanding how IMDB ratings are distributed across the movies in the dataset. By examining the spread of ratings, we can determine whether most movies tend to receive higher or lower ratings, and whether the distribution is balanced or skewed in a particular direction. This will help us better understand the overall reception of movies in the dataset. Next, we investigate the relationship between movie rating and gross vote count. Specifically, we want to explore whether there is a correlation between the rating of a movie and how many votes it earns. By exploring the spread of genres, we can determine which genres high rated movies tend to be, thus providing the movie industry with deeper insights into what the public likes to watch. Next, we want to take a deeper dive into the relationship between rating and runtime to see if longer movies or shorter movies tend to have a higher or shorter rating. Knowing this may provide insights for the perfect runtime for a movie. Moving on, we want to investigate which genres have the highest rating. This information can help determine which genres are dominating in rank or have the potential to become more popular. Lastly, we analyze whether there is a significant relationship between movie runtime and IMDB ratings. This question aims to assess whether the length of a film has any impact on how audiences rate it, potentially offering insights into how runtime might influence audience satisfaction. Together, these questions aim to uncover important relationships between movie characteristics and their financial and critical performance, providing a deeper understanding of what drives success in the movie industry.

Data Visualizations

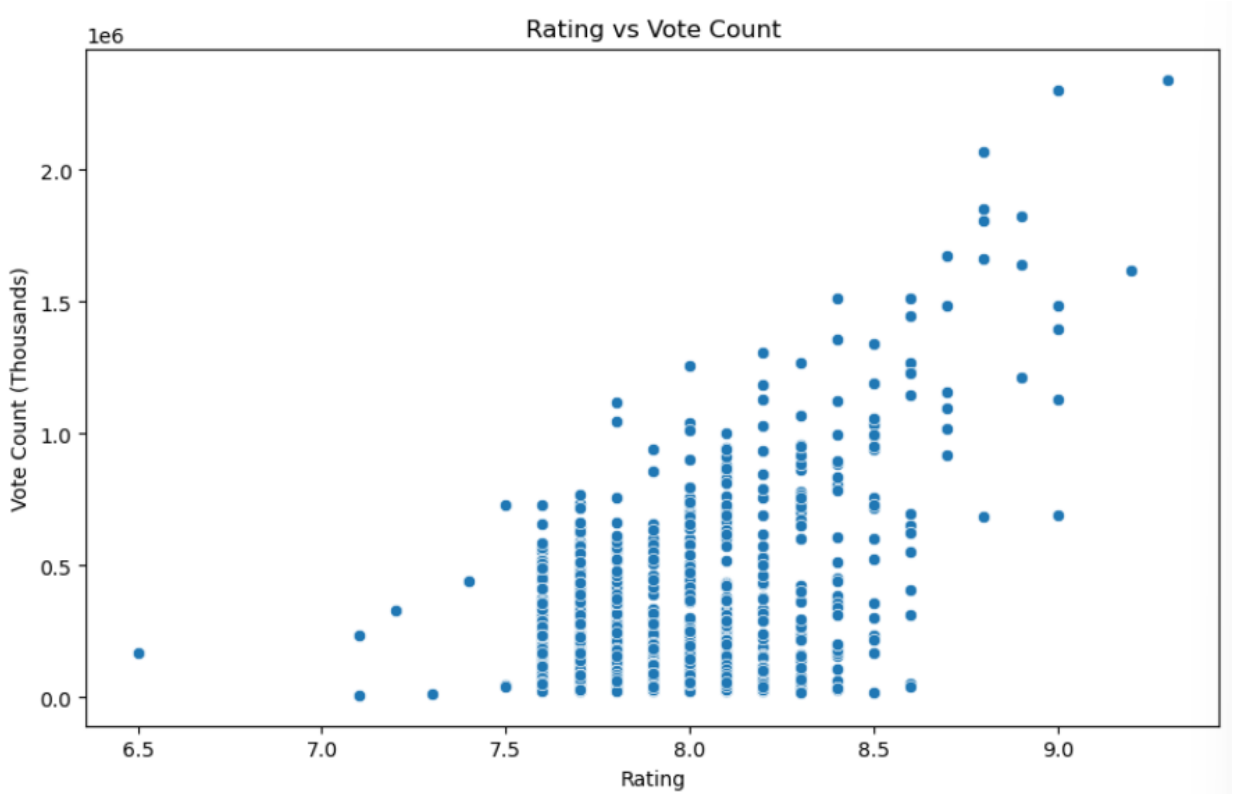
3.1. IMDB Rating Distribution



To explore how IMDB ratings are distributed across the movies in our dataset, we created a histogram of IMDB ratings. We used **Seaborn's histplot()** function to visualize the frequency of each different rating values. The histogram includes a kernel density estimate (KDE) to understand the smooth distribution curve of ratings. By doing this, I was able to identify any potential skew in the ratings and assess whether most movies in the dataset fall within a certain rating range.

The histogram is skewed to the right as there is a high frequency of movies with ratings ~8.0, fewer movies with ratings 8.5 and above, and close to none below 7.5. The interpretation of this graph makes sense as the data is collected from the highest rated movies, and we would expect them to all be decently rated.

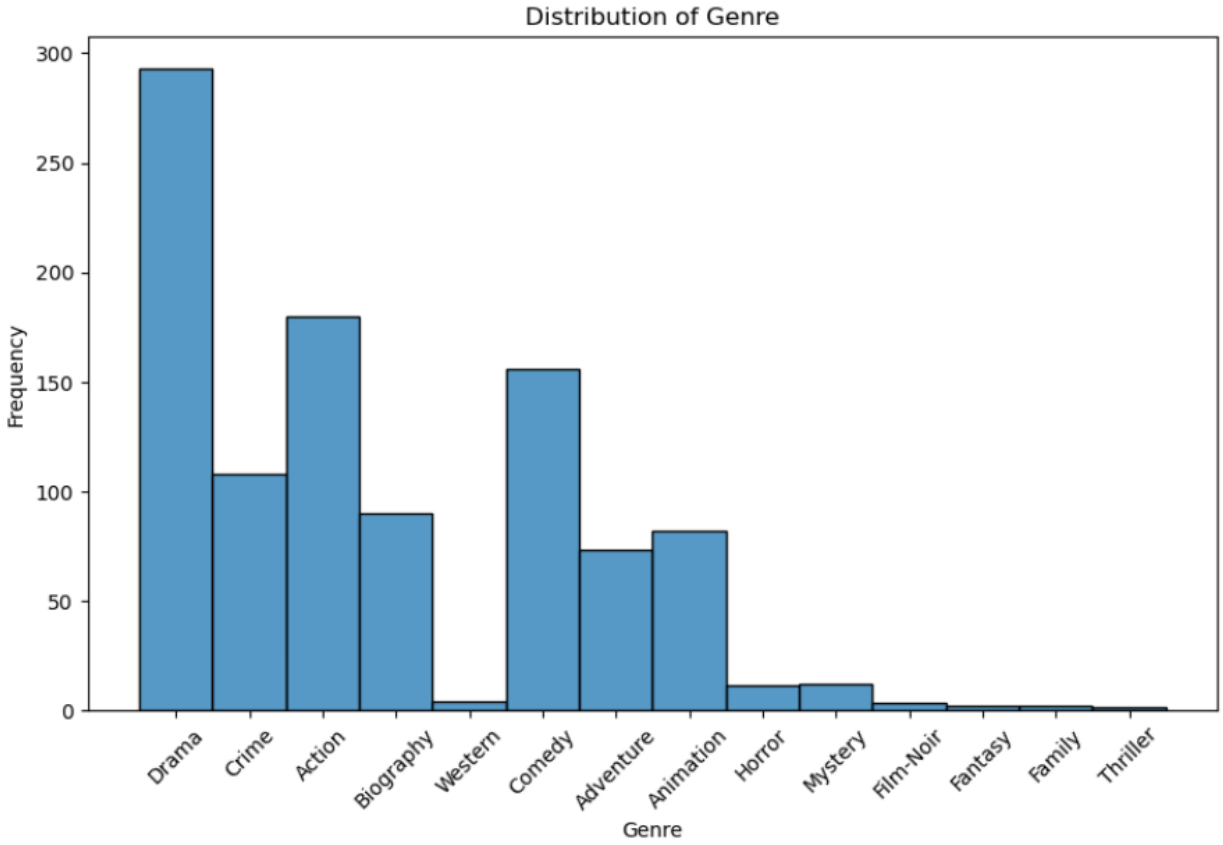
3.2. Rating vs Vote Count



To examine the relationship between movie rating and vote count, we used a scatter plot. The x-axis represents the rating of each movie, while the y-axis represents its total vote count. This plot allows us to visually assess whether there is any noticeable trend between higher rated movies and higher vote counts. It helps determine whether rating is a contributing factor to a movie's total vote count which can lead to more exposure that impacts the success of the movie.

The scatter plot demonstrates a positive linear correlation between these two variables. Meaning as rating increases, voter count increases as well. This is interesting because it shows that higher rated movies are voted on by more people. Thus, inferring that higher rated movies are seen by a larger audience than lesser rated ones.

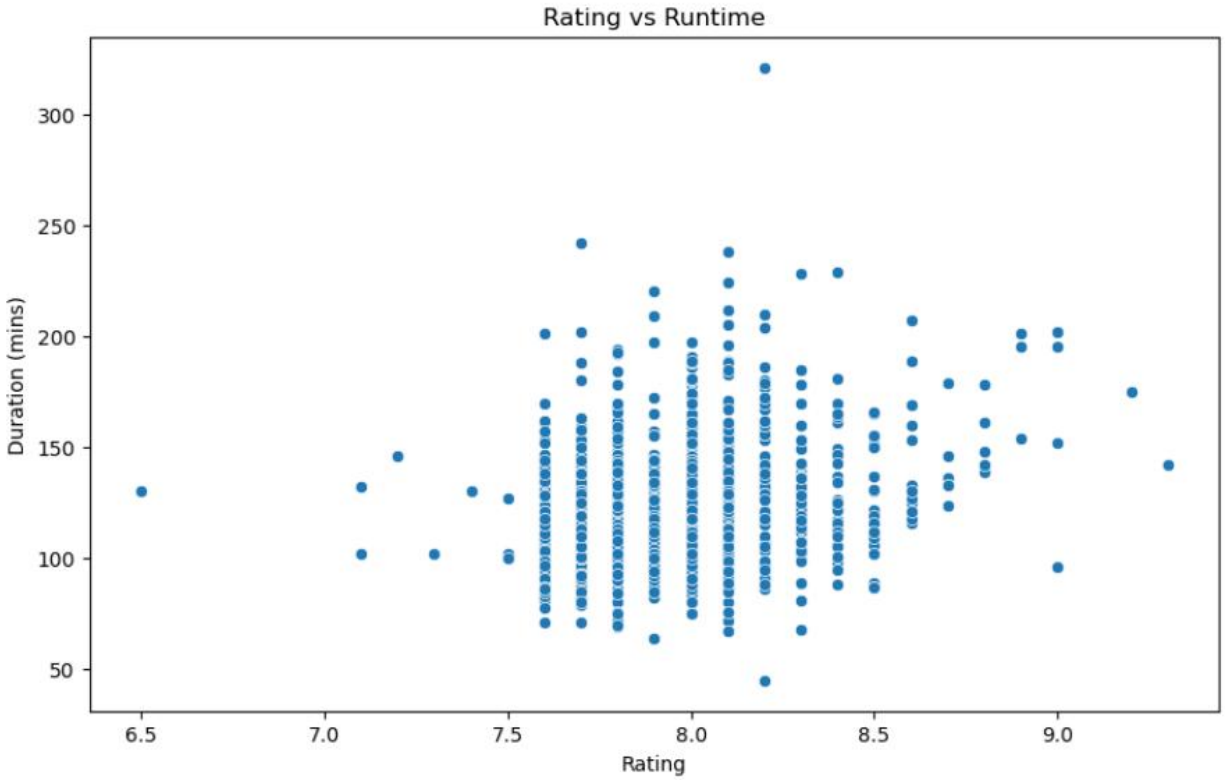
3.3. Distribution of Genre



We created a histogram to explore which genres occurred the most in our data. We used Seaborn's histplot() function to visualize the frequency of each different rating values. In this plot, the x-axis represents the genre of movie, while the y-axis represents its frequency in the data frame. The histogram helps us explore if certain movie genres are more popular than others.

This histogram demonstrates the most popular genre is drama, followed by action and comedy. The middle ground genres are crime, biography, adventure, and animation, with all other genres trailing behind. This information is useful for

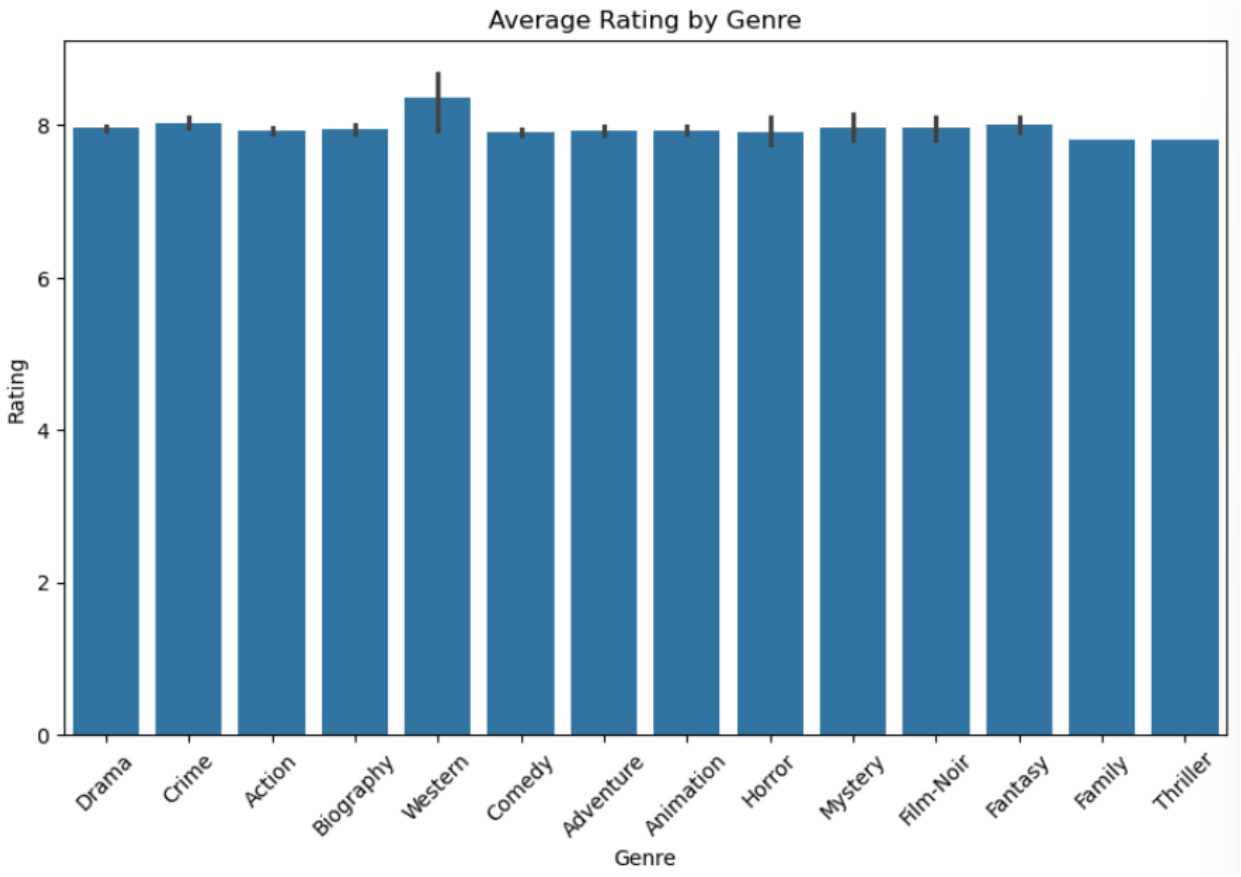
3.4. Rating vs Runtime



We used another scatter plot to visualize the relationship between a movie's rating and its duration or runtime. The x-axis represents the rating, while the y-axis represents runtime. This plot enables us to assess whether the length of a film has a direct influence on the rating it receives.

As we can see from the upwards trend in the scatterplot, as rating increases, so does runtime. This positive linear relationship helps to indicate that movies longer in length tend to receive higher ratings. The information gathered by this graph can help the movie industry make decisions that impact movies' performances.

3.5. Average Rating by Genre



We created a bar chart to visualize which genres in this dataset of most highly rated movies on average receives a higher rating than other genres. While most of the genres perform about the same, one outperforms the rest. Western movies are the highest rated genre on average. This is interesting because looking back to our genre distribution (3.3) we see that western movies have the lowest frequency in this dataset containing the most highly rated movies. This indicates that the few western movies that the public enjoy are rated exceptionally better than other genres. Thus, providing us with insight into the movie market. As we know, most good movies nowadays tend to be drama, action, or comedy. Producers can use this information to corner the market by creating unique movies that outperform other genres and stand out from the rest.

3.6. Correlation of Rating and Runtime

```
Correlation between runtime and IMDB rating: 0.25
Pearson correlation coefficient: 0.25
P-value: 0.000
The relationship between runtime and IMDB rating is statistically significant.
```

Finally, I calculated the Pearson correlation coefficient to quantify the strength of the relationship between IMDB movie rating and duration in minutes. Additionally, I performed a statistical Pearson correlation test to determine whether the relationship is statistically significant.

Just like from our analysis shown in visualization 3.4, we see that this relationship is statistically significant, not strongly correlated, but still significant enough to address. Meaning these two variables, increase with one another and thus effecting the course of how well a movie will perform when taking into account how many minutes it is in length.

4.Conclusion

Our analysis of the film industry using comprehensive datasets from Kaggle and IMDB revealed several key insights. The IMDB ratings of top-rated movies are predominantly high, clustering around the 8.0 mark, demonstrating a skew towards quality content. We discovered a positive correlation between movie ratings and vote counts, indicating that higher-rated films attract more votes and likely a larger audience. Drama emerged as the most prevalent genre, followed by action and comedy, suggesting these genres' enduring popularity. Additionally, we identified a positive relationship between movie runtime and ratings, hinting that longer films tend to be rated higher.

Further, the analysis showed that Westerns, though fewer in number, received the highest average ratings, indicating a niche yet highly appreciated genre. The Pearson correlation coefficient confirmed a statistically significant relationship between runtime and ratings, emphasizing that movie length impacts audience reception. These findings provide valuable insights for industry stakeholders, helping producers, marketers, and streaming platforms tailor strategies to align with audience preferences and optimize movie attributes for better reception and performance.