## Final Project Report: Analyzing the Relationship Between Critical Acclaim and Box Office Success in Films

**1. Introduction**

The film industry frequently balances artistic quality with commercial viability. While some films gain both critical praise and high box office returns, many excel in only one area. Examining the link between movie ratings (like IMDb scores) and financial performance (box office gross) can offer valuable insights for studio investments and creative choices. This project investigates how IMDb ratings and genre classifications correlate with global box office performance for top-rated films. Key questions include: Do higher-rated films consistently generate more revenue? Do specific genres tend to be more profitable? Has critical acclaim changed over time? This analysis aims to identify patterns connecting artistic merit and financial outcomes by integrating and analyzing data on top-rated and top-grossing films.

**2. Data**

**2.1 Data Sources**

This project utilizes two primary data sources:

1. **Box Office Mojo Top Worldwide Grossing Films (Web-Scraped):**
   * *Source:* Data was obtained by scraping 5 pages (1000 movies total) from Box Office Mojo's "Top Lifetime Grosses" chart using Python libraries requests and BeautifulSoup. *(Described in MWRoper\_Project.ipynb)*.
   * *Description:* This dataset contains the top 1,000 movies ranked by their lifetime global gross earnings, including Rank, Title, Lifetime Gross, and Year. The raw scraped data was saved to box\_office\_mojo\_top\_worldwide.csv.
2. **IMDb Top 1000 Rated Movies (Pre-Existing Dataset):**
   * *Source:* Downloaded as a CSV file (imdb\_top\_1000.csv) from Kaggle.
   * *Description:* This dataset includes metadata for IMDb’s top-rated films. The columns used for this analysis were Series\_Title, IMDB\_Rating, Genre, Runtime, Certificate, and Released\_Year. Unnecessary columns were initially dropped. *(Loading described in MWRoper\_Project.ipynb)*.

**2.2 Data Cleaning and Integration**

Several cleaning and integration steps were performed using pandas *(as shown in MWRoper\_Project.ipynb)*:

* **Box Office Mojo Data Cleaning:**
  + The 'Lifetime Gross' column was cleaned by removing '$' and ',' characters and converting the data type to float.
  + Movie titles ('Title') were stripped of leading/trailing quotes.
  + 'Year' was converted to a numeric type, handling potential errors.
* **IMDb Data Cleaning:**
  + The 'Runtime' column was cleaned to extract only the numeric minute value using regular expressions and converted to float.
  + 'Genre' strings were standardized using title case and stripped of extra whitespace.
  + Movie titles ('Series\_Title') were stripped of leading/trailing quotes.
  + 'Released\_Year' was converted to a numeric (integer) type, handling potential errors.
* **Integration:**
  + The cleaned IMDb DataFrame (imdb\_df) and the cleaned Box Office Mojo DataFrame (box\_office\_df) were merged.
  + A **left merge** was performed, keeping all rows from the IMDb dataset and adding matching data from Box Office Mojo.
  + The merge keys were the movie title (Series\_Title, Title) and the release year (Released\_Year, Year). Rows with missing years in either dataset were dropped before merging to ensure accuracy.
  + The resulting merged DataFrame (merged\_df) was saved to cleaned\_movie\_data.csv.
* **Feature Engineering (Genre Dummies):**
  + The 'Genre' column in the merged data (containing comma-separated strings) was processed using str.get\_dummies(sep=', ') to create binary (0/1) dummy variables for each unique genre present in the dataset (e.g., 'Action', 'Drama', 'Comedy', 'Fantasy', etc.).
* **Final Dataset Preparation:**
  + Redundant columns resulting from the merge ('Title' and 'Year' from the Box Office data) were dropped.
  + The DataFrame containing the original merged data and the new genre dummy columns (final\_df) was saved as movies\_with\_genre\_dummies.csv. This is the dataset used for the final analysis.

**Data Dictionary (Final Dataset: movies\_with\_genre\_dummies.csv)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Type** | **Description** | **Source** |
| Series\_Title | Text | Movie title | Both |
| IMDB\_Rating | Numeric | IMDb user rating (0-10) | IMDb |
| Genre | Text | Original genre string (e.g., "Action, Drama") | IMDb |
| Runtime | Numeric | Duration in minutes | IMDb |
| Certificate | Text | Content rating (e.g., "A", "UA") | IMDb |
| Released\_Year | Numeric | Release year (Integer) | Both |
| Rank | Numeric | Movie's rank by global gross earnings (contains NaNs) | Box Office Mojo |
| Lifetime Gross | Numeric | Worldwide earnings in dollars (Float, contains NaNs) | Box Office Mojo |
| *Genre Dummies* | Numeric | Binary (0/1) columns for each genre (e.g., 'Action', 'War') | Derived |

*Note: The left merge on the IMDb dataset resulted in 999 rows. 'Rank' and 'Lifetime Gross' have significant missing values (only 173 non-null values), as only a fraction of the top 1000 IMDb movies were also in the top 1000 Box Office Mojo grossing list.*

**3. Analysis**

The following analysis was performed using the final movies\_with\_genre\_dummies.csv dataset *(results from MWRoper\_DW\_Final.ipynb)*.

**3.1 Basic Descriptive Statistics & Visualizations**

* **IMDB Ratings:** The distribution of IMDb ratings for these top 1000 films is concentrated at the higher end, with a mean rating around 7.95 and a standard deviation of only 0.28. Ratings range from 7.6 to 9.3.

A graph with a line going up

AI-generated content may be incorrect.

* **Lifetime Gross:** Analysis was performed on the 173 films with available data. Lifetime gross earnings are heavily right-skewed (mean: $575M, range: $201M - $2.9B). The log-transformed distribution is more normal.

A graph of different sizes and numbers

AI-generated content may be incorrect.

* **Average Gross by Genre:** For the subset of films with earnings data, genres like Fantasy, Adventure, Action, and Sci-Fi showed the highest average gross earnings. Genres such as Sport, Horror, and History showed lower averages in this group.

A graph of a number of different colored bars

AI-generated content may be incorrect.

* **Rating Trends Over Time:** A time series analysis of average IMDb ratings per year (from 1950-2023) showed ratings generally hovering around 8.0. There were peaks in the late 1950s and dips in the mid-1950s within this dataset of top-rated films. A 5-year moving average indicated relative stability with minor fluctuations.

A graph showing the number of films

AI-generated content may be incorrect.

**3.2 Correlation Between Ratings and Earnings**

The relationship between IMDb Rating and Lifetime Gross was examined for the 173 films with complete data.

* A scatter plot did not reveal a strong linear pattern.

A graph of blue dots

AI-generated content may be incorrect.

* Correlation coefficients were calculated:
  + Pearson Correlation: 0.069 (p-value: 0.37)
  + Spearman Rank Correlation: 0.088 (p-value: 0.248)

**3.3 Hypothesis Test: Higher ratings correlate with higher earnings.**

* **H₀:** No correlation between IMDb rating and lifetime gross (ρ = 0).
* **H₁:** Positive correlation between IMDb rating and lifetime gross (ρ > 0).
* **Test:** Spearman Rank Correlation (one-tailed).
* **Result:** ρ = 0.088, p-value = 0.124.
* **Conclusion:** At a significance level of α = 0.05, we fail to reject the null hypothesis (H₀). There is insufficient statistical evidence in this dataset subset (n=173) to conclude that higher IMDb ratings are positively correlated with higher lifetime gross earnings.

**3.4 Machine Learning: Predicting Earning Classification**

To explore predictors of financial success, 'Lifetime Gross' was categorized into 'Low', 'Medium', and 'High' earner classes based on tertiles for the 173 films with data (Cutoffs: <$333M, $333M-$589M, >$589M). The distribution showed a relatively even split between classes.

A chart of different colors and sizes

AI-generated content may be incorrect.

Two classification models were trained (80/20 split, stratified) using IMDb Rating, Runtime, Rank, and Genre dummy variables as features to predict the 'Earner\_Class'.

* **Logistic Regression (Multinomial):**
  + Accuracy: 97%
  + Performance: High precision and recall across classes, with one misclassification (High predicted as Medium).
* **Random Forest Classifier:**
  + Accuracy: 97%
  + Performance: High precision and recall, with one misclassification (Medium predicted as Low).
* **Feature Importance (Random Forest):** 'Rank' was the most important feature (≈0.56), followed by 'Runtime' (≈0.10), 'Adventure' genre (≈0.07), and 'IMDb\_Rating' (≈0.06). Other genres had minimal impact.

The high accuracy suggests these features can predict the earnings *tier*, but the dominance of 'Rank' indicates that prior box office performance is the strongest factor in this context. IMDb rating played a minor role.

**4. Conclusion**

This project investigated the connection between critical acclaim (IMDb Rating) and commercial success (Lifetime Gross) by integrating data from IMDb's top 1000 rated movies and scraped Box Office Mojo top grossing data.

* **Data Acquisition & Integration:** Data was successfully scraped from Box Office Mojo and merged with a pre-existing IMDb dataset based on title and year. Cleaning involved data type conversions and string manipulation. Genre features were one-hot encoded.
* **Ratings vs. Earnings:** The core hypothesis that higher ratings correlate positively with higher earnings was tested on the subset of 173 films appearing on both lists. Neither visual inspection (scatterplot) nor statistical correlation tests (Pearson, Spearman) found a significant positive relationship (p > 0.05).
* **Genre & Temporal Trends:** Analysis of the subset showed genres like Fantasy and Adventure had higher average grosses. Time series analysis of IMDb ratings for the top 1000 films showed overall stability around the 8.0 mark over the decades.
* **Predictive Modeling:** Classification models (Logistic Regression, Random Forest) accurately predicted earnings tiers (Low, Medium, High) with 97% accuracy using Rank, Runtime, Rating, and Genre features. However, feature importance revealed that 'Rank' heavily dominated the predictions, with 'IMDb Rating' having limited influence.

**Limitations:** The primary limitation is the significant amount of missing box office data ('Lifetime Gross', 'Rank') for the majority of the top-rated IMDb films (only 173/999 had this data). This restricts the generalizability of the findings regarding the rating-earnings correlation and the machine learning models, as they are based on a small, potentially biased subset (films successful enough to be on both lists).