IMDB Movie Ratings Analysis: Final Report

Project Completed By:

Dawit Ashenafi Getachew – 3752264

Chizaram Ikpo - 3760059

Course: CS*2704 – Data Analytics using Python

Instructor: Dr. Jong-Kyou Kim

GitHub Repo: https://github.com/Dash79/CS2704-Project

---

1. Introduction

Movies have long been a staple of global entertainment, captivating diverse audiences with stories of every genre and scale. However, one enduring question is whether a film's production budget heavily influences its critical reception. This project aims to address that question by exploring **how various factors—including genre, year of release, and revenue—relate to IMDB movie ratings**, with a particular focus on budget's impact.

2. Research Questions

Building on the above motivation, this report explores several targeted questions:

1. **Which movie genres receive the highest IMDB ratings?**

2. **How do IMDB ratings trend over time?**

3. **Is there a correlation between a movie's budget and its IMDB rating?**

4. **Primary Hypothesis:** Does a **high budget** lead to a statistically significant higher rating compared to a **low budget**?

3. Dataset

• **Source:** Kaggle's IMDB Movie Dataset

• **Number of Movies:** ~838

• **Key Columns:** Title, Year, Genre, Rating, Budget (Millions), Revenue (Millions), Director, Actors

Following cleaning and preprocessing (detailed below), we saved the refined data as **Cleaned_IMDB_Movie_Data.csv** for further analysis.

4. Methodology

4.1 Data Cleaning & Preprocessing

1. **Missing Values:**

• Removed rows with missing data in key columns (e.g., **Budget (Millions)** and **Rating**).

2. **Relevant Columns:**

• Kept **Title**, **Year**, **Genre**, **Rating**, **Budget (Millions)**, **Revenue (Millions)**, **Director**, **Actors**.

3. **Data Formatting:**

• Ensured **Budget** and **Revenue** columns were converted to numeric types to facilitate analysis.

4. **Output:**

• Final cleaned dataset saved for subsequent visualization and statistical tests.

4.2 Exploratory Data Analysis (EDA)

1. **Genre-Based Averages:**

• Computed mean ratings by genre (e.g., Sci-Fi, Drama, Thriller).

2. **Rating Trends Over Time:**

• Assessed how average ratings shift from older films (pre-2000) to modern releases (post-2010).

3. **Budget vs. Rating and Revenue vs. Rating:**

• Generated scatterplots and correlation measures to explore potential linear relationships.

4.3 Hypothesis Testing

1. **Correlation Analysis:**

• Calculated both **Pearson** and **Spearman** correlations between budget and IMDB rating to gauge strength/direction of the relationship.

2. **Two-Sample T-test:**

• Split the dataset into **High Budget** vs. **Low Budget** groups using the median budget as a threshold.

• Compared mean ratings between the two groups.

• Determined statistical significance with a threshold of $p < 0.05$.

5. Results & Key Findings

5.1 Highest-Rated Genres

• **Sci-Fi, Drama, and Thriller** consistently show higher average ratings.

• **Action** and **Comedy** display broader variability—some outliers perform well, but overall means trail behind top genres.

5.2 Rating Trends Over Time

• **Older films** (pre-2000) tend to maintain **consistently strong** IMDB ratings.

• **Post-2010** releases exhibit **wider variability**, possibly reflecting changing audience preferences and industry growth.

5.3 Budget vs. IMDB Ratings

• **Correlation:** A weak but statistically significant correlation (~0.15) suggests that a larger budget **does not** strongly guarantee a high rating.

• **Two-Sample T-test:**

• Mean rating of high-budget films: **~6.8**

• Mean rating of low-budget films: **~6.5**

• $p < 0.05$, implying a real difference—though small (0.3 points on a 10-point scale).

• **Interpretation:** While higher-budget movies tend to earn slightly higher ratings, the effect size is modest. Story, cast, marketing, and other creative factors likely play a significant role.

6. Conclusion

Overall, the data suggest that **genre** and **time period** can strongly influence IMDB ratings. **Budget** does have a **mild** positive effect on ratings, but not enough to claim it is the sole driver of a film's success. Movies with more modest budgets can still achieve robust critical reception—especially if other creative or narrative elements resonate well with audiences.

7. Future Work

• **Refined Budget Brackets:**

• Instead of a high/low split, use multiple brackets (e.g., low, medium, high) for an ANOVA test.

• **Director/Actor Influence:**

• Investigate the extent to which renowned directors or casts affect ratings, potentially via regression.

• **Award/Streaming Data:**

• Incorporate critical awards, film festival accolades, or streaming platform data to explore additional predictors of high ratings.

8. References & Resources

1. **IMDB Movie Dataset (Kaggle)**: Original dataset used for the project.

2. **imdb_analysis.py**: Main Python script including data cleaning, EDA, correlation, and hypothesis testing.

3. **Cleaned_IMDB_Movie_Data.csv**: Final cleaned dataset.

4. **Final_IMDB_Analysis.csv**: Extended dataset with additional columns (e.g., High/Low budget group).

5. Getachew, D. A. (2025). **Project Proposal: IMDB Movie Ratings Analysis**. (Unpublished course proposal).