IMDB Movie Ratings Analysis – Project Proposal

Course: CS*2704 – Data Analytics using Python
Instructor: Jong-Kyou Kim, PhD
Student Name: Dawit Ashenafi Getachew
Student ID: 3752264
GitHub Repo: https://github.com/Dash79/CS2704-Project

---

1. Introduction

Movies are a significant part of global entertainment culture, and understanding what makes a movie successful can be valuable for filmmakers, producers, and analysts alike. This project focuses on analyzing IMDB movie ratings to determine factors that contribute to a movie's success or critical reception. The dataset is sourced from Kaggle and includes key information such as title, year of release, genre, budget, revenue, and IMDB rating.

---

2. Primary Hypothesis

"Movies with higher budgets tend to have higher IMDB ratings."

We propose this hypothesis for both theoretical and practical reasons. From a theoretical standpoint, the notion that bigger budgets lead to better movies aligns with the idea that higher financial resources allow for improved production values, advanced special effects, greater marketing reach, and possibly more renowned actors or directors. Practically, movie studios often invest heavily in blockbuster films, which can garner significant attention and potentially inflate viewer ratings.

However, it is also plausible that certain low-budget or independent films can outperform big-budget counterparts on IMDB. Factors such as strong storytelling, talented casts, or critical acclaim at film festivals can drive up ratings. Therefore, our hypothesis must be rigorously tested with real data to determine whether large budgets truly correlate with higher average ratings.

---

3. Research Questions

1. Which movie genres receive the highest IMDB ratings?
2. Has the average IMDB rating changed over time?
3. Is there a correlation between a movie's budget and its IMDB rating?
4. (Key Question) Does a high budget lead to a statistically significant higher rating compared to a low budget?

---

4. Dataset

- Source: IMDB Movie Dataset (Kaggle)
- Data Points: Title, Year, Genre, Rating, Director, Actors, Budget (Millions), Revenue (Millions), etc.
- Size: Approximately 800+ movies spanning various years and genres.

---

5. Methodology

5.1 Data Cleaning

- Remove missing values, especially in the budget and rating columns.
- Filter relevant columns: Title, Year, Genre, Rating, Budget (Millions), Revenue (Millions), Director, Actors.
- Convert Budget and Revenue to numeric types as needed.

5.2 Exploratory Data Analysis (EDA)

- Summarize average ratings by genre.
- Observe rating trends over the years.
- Visualize revenue vs. rating, budget vs. rating.

5.3 Statistical Analysis

- Correlation: Use Spearman or Pearson correlation to see if there is a relationship between Budget and Rating.
- Hypothesis Testing:
  - Split movies into “High Budget” vs. “Low Budget” using a chosen threshold (e.g., median budget).
  - Perform a two-sample T-test to compare average ratings between these two groups.
  - Evaluate statistical significance (p-value < 0.05) and consider effect size.

---

6. Visualization

- Bar plots for ratings by genre.
- Line plots for ratings over the years.
- Boxplots/scatterplots for budget vs. rating to illustrate distribution differences.

---

7. Expected Outcomes

- Clear identification of top-rated and lower-rated genres.
- Visual display of rating trends across different release years.
- A quantitative conclusion on whether big-budget movies truly score higher on IMDB on average.
- Insights on whether budget alone, or in combination with other variables, significantly drives IMDB ratings.

---

8. Tools & Resources

- Python Libraries: Pandas, NumPy, Matplotlib, Seaborn, SciPy (stats)
- Environment: VS Code

---

9. Timeline

Task	Deadline
Dataset Selection & Proposal	Week 1
Data Cleaning & Preprocessing	Week 2
Exploratory Data Analysis (EDA)	Week 3
Visualizations & Hypothesis Testing	Week 4
Report & Presentation Preparation	Week 5
Final Revisions & Submission	Week 6

---

10. Conclusion

By coupling robust data cleaning with exploratory analysis and formal hypothesis testing, this project aims to illuminate whether higher production budgets genuinely impact viewer ratings. Beyond the primary question, analyzing genre, time trends, and potential correlations with revenue will provide a well-rounded view of movie success on IMDB.