Project Proposal: Data-Driven Insights from IMDb Movies (2010–2023)

Introduction

The film industry is undergoing rapid transformation driven by the rise of streaming platforms, changing audience behaviors, and the evolving dynamics of movie production. Studios, filmmakers, and streaming services face challenges in understanding what types of films succeed, which runtimes keep audiences engaged, and how audience ratings and votes evolve over time.

This project leverages IMDb movie data (2010–2023) to uncover patterns, trends, and actionable insights that can guide data-driven decision-making in the film industry.

Problem Statement

The entertainment industry invests billions of dollars annually in content production. However, greenlighting films is often based on intuition, star power, or market trends—leading to high risk and inconsistent returns.

Key challenges include:
- Understanding which genres balance both critical acclaim and commercial success.
- Determining the ideal runtime that maximizes audience engagement.
- Tracking audience behavior shifts post-2020, especially with the streaming boom.
- Evaluating whether ratings and votes show stable or declining trends over time.

Without data-driven insights, decision-makers risk overinvesting in formats that may not align with evolving audience preferences.

Project Objectives

The main goal of this project is to analyze IMDb movie data from 2010–2023 to derive insights that can:
1. Identify top-performing genres by average ratings and popularity.
2. Analyze runtime distribution to uncover engagement sweet spots.
3. Track box office and ratings trends over the years.
4. Compare pre-2020 and post-2020 audience engagement (impact of streaming).
5. Provide data-backed recommendations for filmmakers, studios, and streaming platforms.

Dataset Description

The dataset was sourced from IMDb and covers the years 2010–2023. It is stored in the zippedData/ folder of the repository.

Key Columns:
- movie_id → Unique identifier
- primary_title → Movie title
- start_year → Release year
- runtime_minutes → Duration of the movie
- genres → Genres (Drama, Comedy, Action, etc.)
- averagerating → IMDb average rating (1–10)
- numvotes → Number of audience votes

Preprocessing Steps:
- Removed missing or inconsistent runtimes.
- Split multi-genre films to allow analysis per genre.
- Created new temporal features (year_minus_1, year_plus_1).

Methodology

The analysis will be conducted in Python (Jupyter Notebooks) using pandas, matplotlib, seaborn, and numpy.

Steps:
1. Exploratory Data Analysis (EDA)
2. Genre Analysis
3. Runtime Distribution
4. Yearly Trends
5. Audience Engagement
6. Visualization (professional charts saved in images/)

Expected Insights

- Top Genres: Drama dominates volume, while Documentary & Mystery lead in ratings.
- Runtime Distribution: Most films fall between 100–120 minutes, supporting audience attention span theory.
- Yearly Ratings: Ratings remain stable, indicating consistent quality.
- Audience Votes: Post-2020 films receive fewer votes, showing a shift to niche streaming audiences.
- Box Office Trends: Franchise films drive peaks in gross revenue, especially after 2015.

Recommendations for the Film Industry

Based on the analysis of IMDb data (2010–2023), here are actionable, human-centered recommendations for filmmakers, studios, and streaming platforms:

- ### Optimize Runtime (Sweet Spot: 100–120 minutes)

Why it matters: Our analysis shows a natural bell curve in runtimes, with most films clustering around 100–120 minutes.

Human insight: This length strikes the perfect balance—long enough to develop meaningful characters and plots, yet short enough to keep modern audiences engaged in a world full of distractions.

Actionable tip: Studios should avoid excessively long runtimes unless justified by epic storytelling (e.g., Avengers: Endgame). For streaming content, lean toward the shorter end of this range.

- ### Diversify Genre Portfolio

Why it matters: Drama dominates in volume, but genres like Documentary and Mystery score higher on average ratings.

Human insight: This suggests that while audiences flock to dramas for storytelling, they turn to documentaries and mysteries when they crave originality and critical depth.

Actionable tip: Studios should strike a balance—continue producing commercially safe dramas, but strategically invest in smaller, high-quality projects in niche genres that build brand prestige and attract loyal viewers.

- ### Leverage Streaming Trends Post-2020

Why it matters: After 2020, films became shorter and often received fewer votes, reflecting the rise of streaming-first releases.

Human insight: Viewers at home are less patient with long runtimes and more selective with what they choose to review or rate.

Actionable tip: Streaming platforms should design content around binge-friendly runtimes (90–110 min) and target niche audiences with personalized recommendations, since loyal fans drive most of the engagement.

- ### Prioritize quality over quantity

Why it matters: Ratings have remained relatively stable across years, even though more movies are being released.

Human insight: Audiences can spot rushed, low-quality productions. A smaller slate of polished, original projects will perform better in both ratings and word of mouth .

Actionable tip: Studios should focus budgets on fewer, well-researched scripts rather than flooding the market with formulaic films.

● Use Data-Driven Greenlighting

Why it matters: Data reveals clear sweet spots genres with consistent high ratings, runtimes audiences prefer, and trends linked to streaming behavior.

Human insight: Greenlighting films has always been a gamble, but data reduces the risk. By combining creative instinct with analytical evidence, studios can improve the odds of both financial success and critical acclaim.

Actionable tip: Establish data review checkpoints during the early development phase—before committing full budgets—to assess alignment with audience preferences.

Deliverables

- Student.ipynb: Main analysis notebook
- index.ipynb: Supporting exploration
- images/: Folder with saved charts
- README.md: Documentation of project findings
- Final Report/Presentation: Summarizing insights and recommendations

Technologies Used

- Python
- Jupyter Notebook
- Libraries: pandas, numpy, matplotlib, seaborn
- Git/GitHub

✓ Conclusion

This project provides a data-driven framework for understanding audience preferences, optimizing runtimes, and aligning production strategies with industry shifts.
With clear insights into genres, runtimes, ratings, and engagement trends, it empowers stakeholders to make smarter, evidence-based decisions in the fast evolving film industry.