CineMetrics: Building the Ultimate Cinematic Dataset

A Multi-Source Data Analysis on Film Performance & Strategic Forecasting

Vulcan Variance

August 2025

1. Executive Summary

CineMetrics merges fragmented movie data sources into a clean, unified dataset that provides a strategic lens on the cinematic landscape. This project empowers production houses, streaming platforms, and investors to make data-driven decisions based on cost structures, audience response, and revenue outcomes.

Key Highlights:

- Horror and thriller genres deliver the highest return on investment (ROI) up to **6:1**
- Final dataset contains **1,927 enriched movie records**, with **56.9% verified budget data**
- Comparative analysis of IMDB vs TMDB ratings reveals platform bias

Business Impact:

- Informs genre funding strategies
- Supports ROI forecasting pre-release
- Highlights director/cast performance trends for casting optimization

2. Project Background

The entertainment industry is evolving rapidly, yet decision-making often remains reactive or gut-driven due to fragmented data. This project bridges the data divide by consolidating major industry sources into a single, analysis-ready dataset.

Scope:

- Focused on theatrical releases (2000–2024)
- Excluded streaming-only content (phase 2)
- Prioritized budget, revenue, genre, ratings, cast, and popularity metrics

Objective:
Create a reliable base dataset that aligns creative potential with commercial viability.

3. Methodology

3.1 Data Sources

- **IMDB:** Core metadata – genre, cast, runtime, and user ratings
- **TMDB:** Popularity metrics, rating counts, alternate audience scores
- **Box Office Mojo:** Verified domestic and international gross earnings
- **The Numbers:** Estimated or reported production budgets

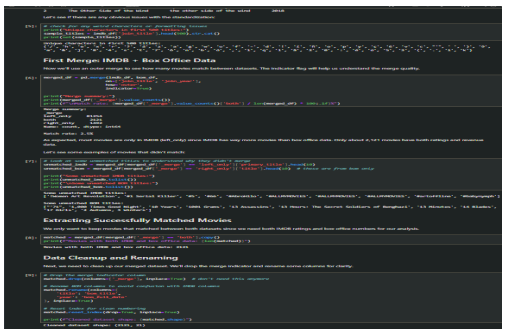
3.2 Data Cleaning & Standardization

- Unified inconsistent naming conventions using fuzzy matching
- Removed duplicates and outliers
- Harmonized rating scales and genre categories

3.3 Merging Strategy

- **Merge Keys:** Title + Year

- **Order:** IMDB → Box Office Mojo → TMDB → Budgets



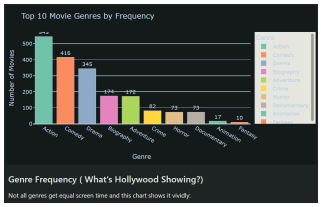
Final Output:

- 1,927 movies
- Multi-source field coverage for revenue, rating, genre, and budget

4. Exploratory Data Analysis (EDA)

4.1 Genre Distribution

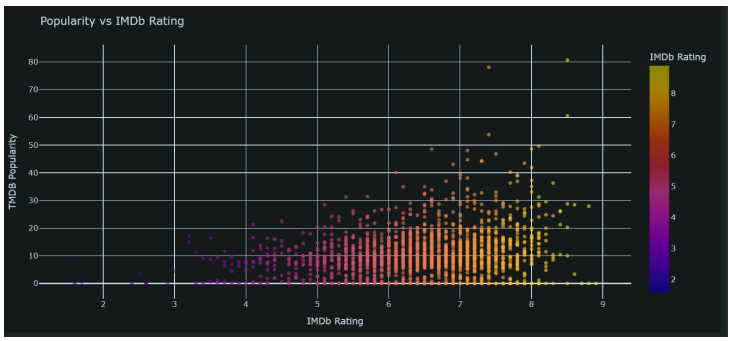
- Action, Drama, and Thriller dominate the landscape
- Documentaries are underrepresented



4.2 Rating Analysis

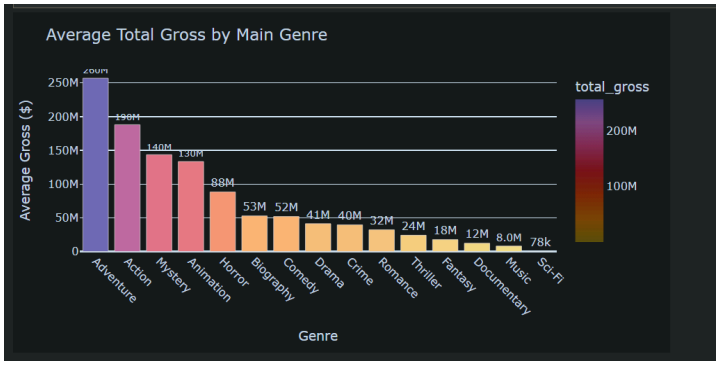
- IMDB scores average between 5.4 and 7.2

- TMDB ratings skew slightly higher



4.3 Budget Insights

- Budgets range from under $1M to $250M+
- Budget imputation applied for missing values (genre median logic)



5. Key Findings & Discoveries

- **High-ROI Genres:** Horror/Thriller outperform all others in ROI despite modest budgets
- **Fantasy Films:** Surprisingly low ROI — often breakeven despite large investment
- **Platform Ratings Gap:** TMDB ratings tend to be 0.4–0.8 points higher than IMDB for the same title

6. Technical Implementation

6.1 Tools & Libraries

- **Python**, **Jupyter Notebooks**
- Libraries: Pandas, NumPy, Seaborn, Plotly, SciKit-learn, SciPy

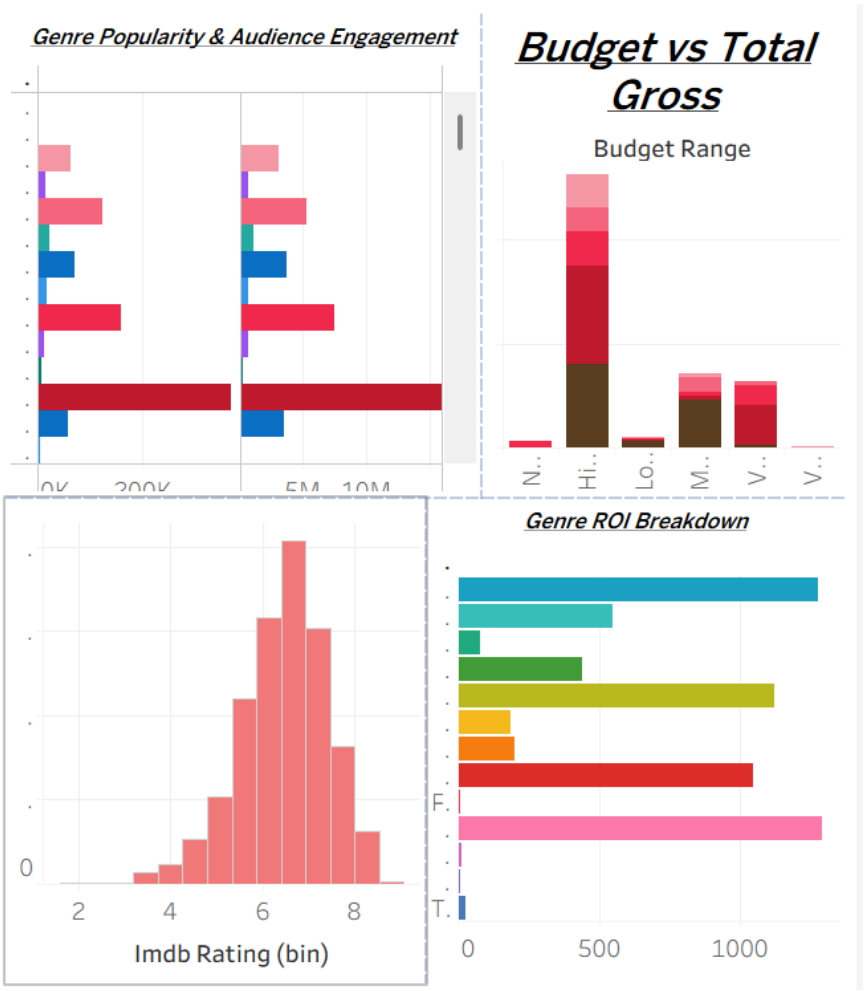
6.2 Techniques Used

- Genre-based imputation for missing budgets
- Min-max normalization for rating comparisons
- Exploratory correlation and regression modeling

7. Business Applications

This dataset enables:

- Predictive modeling of movie performance
- Strategic budget allocation across genres
- Cast/director performance tracking
- Portfolio planning for film studios and distributors



8. Limitations

- ~43% of entries lacked budget data pre-imputation
- Streaming-exclusive films excluded (to be addressed in future work)
- Ratings may reflect Western-centric viewing preferences
- Manual data merging limits real-time scaling

9. Next Steps

- Integrate streaming platforms like Netflix, Hulu, Prime
- Train ML models for revenue prediction based on pre-release metadata
- Launch Tableau dashboard for dynamic exploration
- Automate data ingestion via GitHub Actions or Airflow
- Extend dataset with NLP sentiment from reviews

10. Final Recommendations

- **Prioritize mid-budget horror/thriller investments** with proven ROI patterns
- **Track cast/director combos** to identify repeat success drivers
- **Invest in genre experimentation** using data-backed insights
- **Use the dataset as a foundational tool** for junior analyst training and C-suite decision dashboards

11. Appendices

- Data Dictionary: Column-level definitions
- Merge Strategy Logic: Matching and filtering criteria
- Code Snippets: Merge logic, normalization, imputation
- Visual Exports: All charts shown above in high resolution

12. About the Team

- **Felicity Muthoni**
- **Maureen Ngaire**

- **Brian Kiprop**
- **Crispus Wanene**
- **Henry Njoroge**
- **John Mungai**
- **Alvin Kipleting**

13. Acknowledgements & Contact

With thanks to:

- Data providers: IMDB, TMDB, Box Office Mojo, The Numbers
- TM

> “Data is our script. The world is our screen.” – *CineMetrics, 2025*

Contact Us:
Email: muthonifelicity4@gmail.com
GitHub:https://github.com/Annfelicty/VulcanVariance
Dashboard:
(https://public.tableau.com/views/phase2_17543566839470/Dashboard1?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link)