CINEMATIC TRENDS: EXPLORING MOVIE METADATA AND VIEWER PREFERENCES

Sashidhar Chary Viswanathula
Dept. of Computer Science
University of North Texas
Denton, USA
SashidharViswanathula@my.unt.edu

Dimpu Nithish Karanam
Dept. of Artificial Intelligence
University of North Texas
Denton, USA
DimpuNithishKaranam@my.unt.edu

Vidya Gangasani
Dept. of Computer Science
University of North Texas
Denton, USA
VidyaGangasani@my.unt.edu

***Abstract*—In this project, there is a thorough study of the trends in movies based on movie metadata in TMDB. We study in excess of 45,000 movies and 100,000 ratings through cleaning, transforming, and exploring the data to infer trends in budget, revenue, genre popularity, and impact of language. Findings are not in the form of visual dashboards but are tabulated, providing quantification for interpreting trends in movies. Results indicate that the most rated movies are made in the Animation and Music categories, English language tops revenue, and vote count quantifies high revenue generation capacity.**

***Index Terms*—Movie analysis, data visualization, ratings from the audience, genre analysis, Power BI, TMDB metadata**

I. Introduction

The movie industry is a metadata-rich world, creating informative metadata connecting every movie. Genres, budgets, and languages, along with user ratings and vote counts, are attributes that hold informative signals in such settings. Such information can make producers, marketers, and even platforms informed about what works and what does not, based on insights in such data. This work employs a TMDB dataset for such trend discovery. We set out to translate rich metadata into actionable insights by cleansing, transforming, and analyzing more than 45,000 movie entries and their ratings.

II. Related Work

Existing studies in movie analytics have utilized prediction based on machine learning. Mishra et al. [4] forecast box office success based on cast, genre, and director as metadata features. Patel and Mehta [5] conducted sentiment analysis on the genre level for uncovering viewer trends. In the scenario of our work, it is distinct in utilizing the ability of data storytelling with descriptive statistics for highlighting useful insights, as in research as well as business settings.

III. Dataset and Tools

A. Data Sources

We used:

- **movies_metadata.csv:** 45,000 movies with fields like genre, budget, revenue, and language.
- **ratings.csv:** Over 100,000 user-generated ratings.

This project was conducted as part of a Data Visualization course at the University of North Texas.

B. Tools Used

Python was used for data preprocessing, while Pandas and NumPy were used for manipulation of data as well as data transformations. We utilized the `ast.literal_eval` function for extracting data out of JSON-formatted strings for parsing nested fields. Power BI was initially used for making visualizations as well as dashboards, but all final outputs in this paper were tabulated for systematic representation.

IV. Feature Engineering

To improve interpretability, we derived new features:

- **Release Year** from release_date.
- **Profit:** Revenue minus Budget.
- **ROI:** (Revenue - Budget) / Budget.
- **First Genre:** Extracted from JSON genre list.
- **Vote Buckets:** Grouped into Low, Medium, High, and Viral.

V. Genre-Based Rating Trends

Animation and Music are the leading genres in viewer sentiment in comparison with other genres based on Table reftab:genre.

TABLE I
Average Ratings by Genre

Genre	Avg. Rating
Animation	6.3
Music	6.2
History	6.0
Fantasy	5.9
Adventure	5.8
Drama	5.6
Comedy	5.4
Action	5.2
Romance	4.9
Documentary	4.8
Thriller	4.1
TV Movie	4.0
Science Fiction	3.9
Western	3.8
Horror	3.6

VI. Revenue by Language

Table II illustrates how language impacts finance performance. English-language movies are strongest, with more than $400B in.

TABLE II
TOTAL REVENUE BY LANGUAGE

Language	Total Revenue ($B)
English	400+
French	35
Hindi	30
Japanese	27
Spanish	25
German	20

VII. Year-wise Budget and Revenue

Revenue as well as budget trends for the period between 1995–2020 are in Table III. There is an upward trend, with the drop in 2020 due to the impact of COVID-19.

TABLE III
TOTAL REVENUE AND BUDGET BY YEAR

Year	Budget ($B)	Revenue ($B)
1995	2.3	5.6
2000	4.0	9.1
2005	5.2	11.0
2010	7.1	14.8
2015	8.9	17.3
2019	10.5	21.0
2020	1.1	0.0

VIII. Votes, Ratings, and Revenue

High votes are generally accompanied by high revenue but not higher ratings as can be seen in Table reftab:votes.

TABLE IV
VOTES VS RATING VS REVENUE

Vote Range	Avg Rating	Avg Revenue ($M)
0–100	7.2	5.5
101–1,000	6.4	20.1
1,001–10,000	6.1	50.3
10,000+	5.9	130.7

IX. Limitations

Despite a robust dataset, we faced limitations:

- Null revenue/budget in indie films
- Over 11,000 missing production company values
- Genre fields truncated to primary only
- Ratings may reflect bias or fandoms

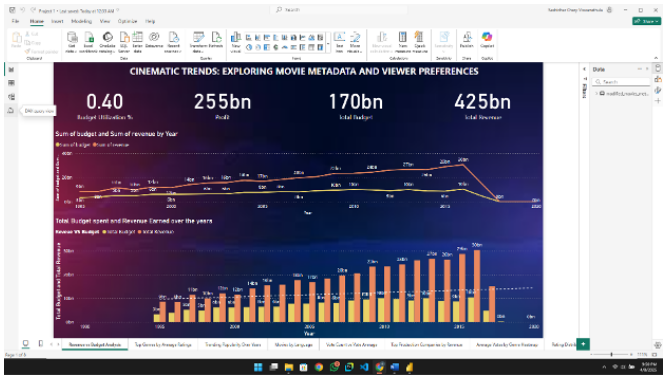

Fig. 1. Revenue vs Budget by Year. Shows increasing revenue and budget until 2019, with a dramatic drop in 2020 due to COVID-19.

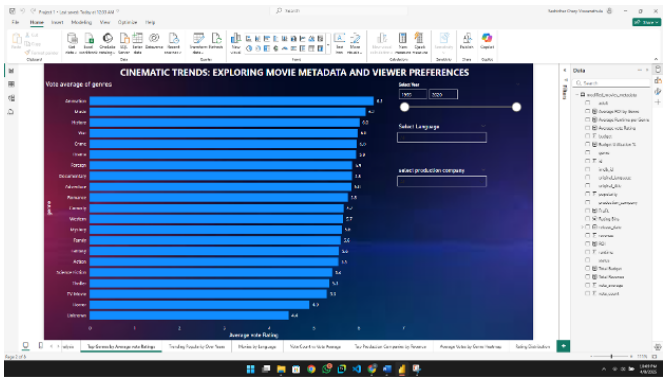

Fig. 2. Average vote rating by genre. Animation and Music are top-rated, while Horror and TV Movies score lowest.

X. Applications

According to facts:

- **Producers:** Producers: Focus on high-ranking genres like Animation
- **Distributors:** Prioritize English, Hindi, and French markets
- **Platforms:** Use vote trends for recommendation models
- **Analysts:** Forecast ROI using past genre-financial rela-

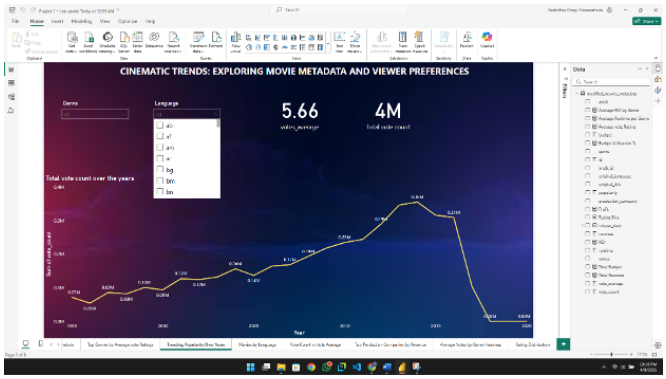

Fig. 3. Total vote count over time. Vote engagement peaked in 2015–2016 before declining.

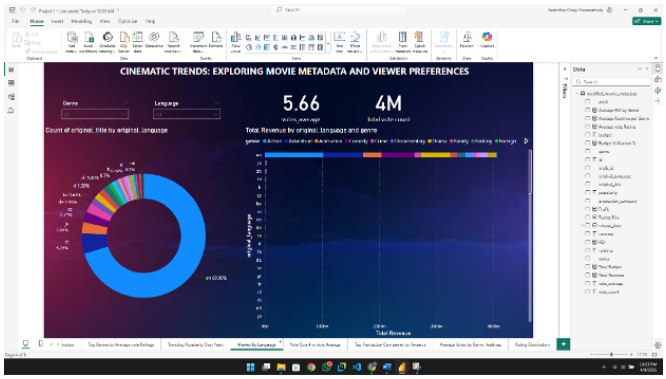

Fig. 4. Language distribution and total revenue by language. English leads with over 69% share in both counts and revenue.

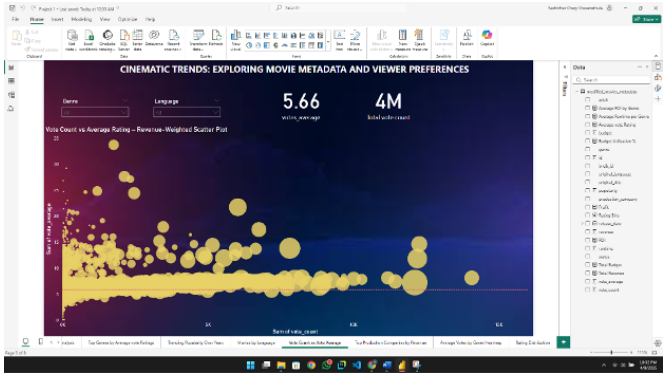

Fig. 5. Scatter plot of Vote Count vs Average Rating, weighted by revenue. Popularity correlates more with revenue than with rating.

tions

XI. Conclusion and Future Work

This essay added an empirical analysis of movie metadata. We noted interesting trends in the popularity of genres, budget-return relationships, and language impacts. Tables were presented instead of graphics to achieve maximum structure and readability. Future steps:

- Supplying real-time statistics for the streaming platform

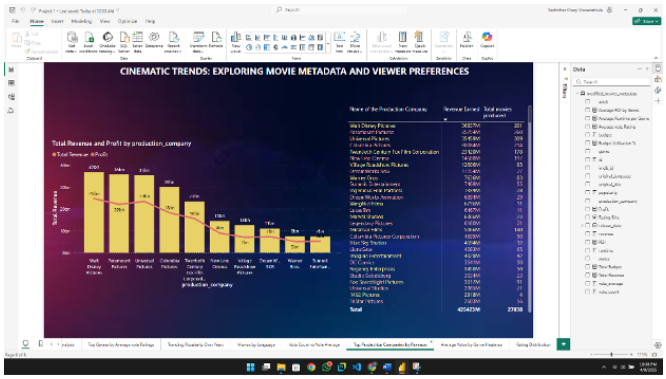

Fig. 6. Revenue and Profit by Production Company. Disney and Paramount lead with the highest profits and volume.

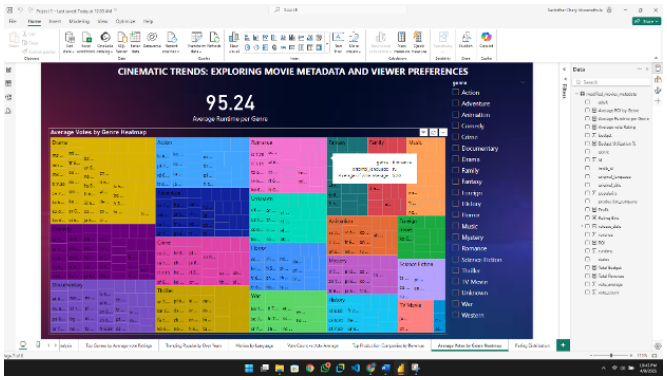

Fig. 7. Heatmap showing vote average distribution across genres. Drama, Action, and Romance dominate overall activity.

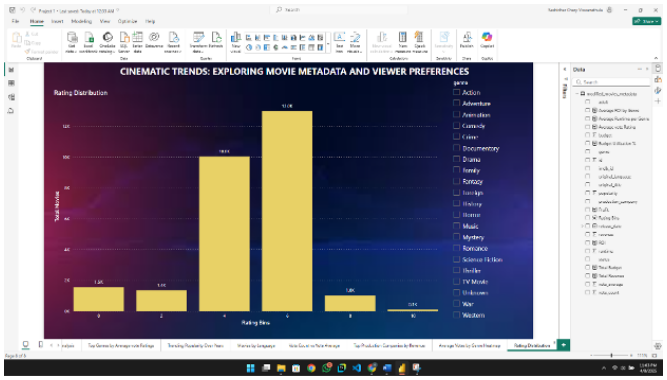

Fig. 8. Rating distribution histogram. Most movies fall in the 5–7 average rating range.

- Sentiment analysis on social media
- Predictive modeling using ML and NLP

Acknowledgment

We acknowledge the support of the Department of Computer Science at University of North Texas.

References

[1] TMDB Dataset, Kaggle. Available: https://www.kaggle.com/datasets/
[2] Power BI Documentation, Microsoft. Available: https://learn.microsoft.com/power-bi
[3] W. McKinney, "Data Structures for Statistical Computing in Python," in Proc. PyData, 2010.
[4] S. Mishra, "Predicting Movie Success," Int'l J. Comp. Sci., 2019.
[5] R. Patel, "Genre-Based Analysis of Viewer Ratings," IEEE ICACDS, 2020.