Movie Performance & ROI Analysis

Overview

This project analyzes movies from multiple datasets to identify patterns in ratings, financial performance, and audience reception. The goal is to generate actionable recommendations for film production teams on what to produce, when to release, and how to maximize returns.

Business Understanding

•We worked with a production team seeking to answer key business questions that drive investment and strategy decisions

Stakeholders include;

- Film Producers
- Marketing Executives
- Distribution Strategists.

Business Questions

i.What genres deliver consistently high ratings and ROI?

ii. When is the best time of year to release a film for financial success?

iii. How well does popularity reflect quality or profitability?

iv. Are there lesser-known but high-return film types being overlooked?

Data Understanding & Analysis

Data Sources

Dataset	Description
movie_basics (IMDb)	Genre, release year, title
movie_ratings (IMDb)	Ratings and vote counts
tmdb.movies (TMDb)	Popularity scores, vote metrics
tn.movie_budgets (The Numbers)	Budget, revenue, release dates

Final Dataset had a Shape of (46287, 21) rows, and columns respectively.
At the end of the merging process, we obtained a consolidated dataset ready for exploratory analysis and business recommendations
Observations from movies_metadata.info()

- Core variables (e.g. title, genres, vote_average_x) are available for ~36K movies
- tmdb features (e.g. popularity, language) are present for ~11K movies
- Financial data (e.g. production_budget, worldwide_gross) exists for ~6K movies
- Several columns are duplicated or suffixed (e.g. _x, _y) due to merges.

Target Variables for Analysis

Based on our business objectives, the following columns are central to our analysis:

- production_budget
- domestic_gross, worldwide_gross
- distributor
- vote_average, vote_count
- genres
- release_date (will be used to extract release_month)
- popularity

Data Cleaning Steps

- Standardized column names for consistent merging
- Converted date columns to datetime format
- Calculated ROI (worldwide_gross - production_budget) / production_budget
- Created log_vote_count for better skew handling
- Filtered movies with valid vote counts, budgets, and years (2015–2025)
- Removed duplicate records

Due to combining multiple datasets with different coverage (e.g., tmdb, im.db, and tn.movies), the final merged dataset contains missing values in several fields.

Instead of dropping rows, we'll apply targeted filtering based on the specific recommendation.

For example, ROI analysis only used rows where financial data was available, while genre-based recommendations relied on ratings and genre fields.

For EDA we used univariate, bivariate and multivariate analysis to observe and understand how variables related to each other.

Multivariate analysis yielded the solution to our problems as we shall see below.

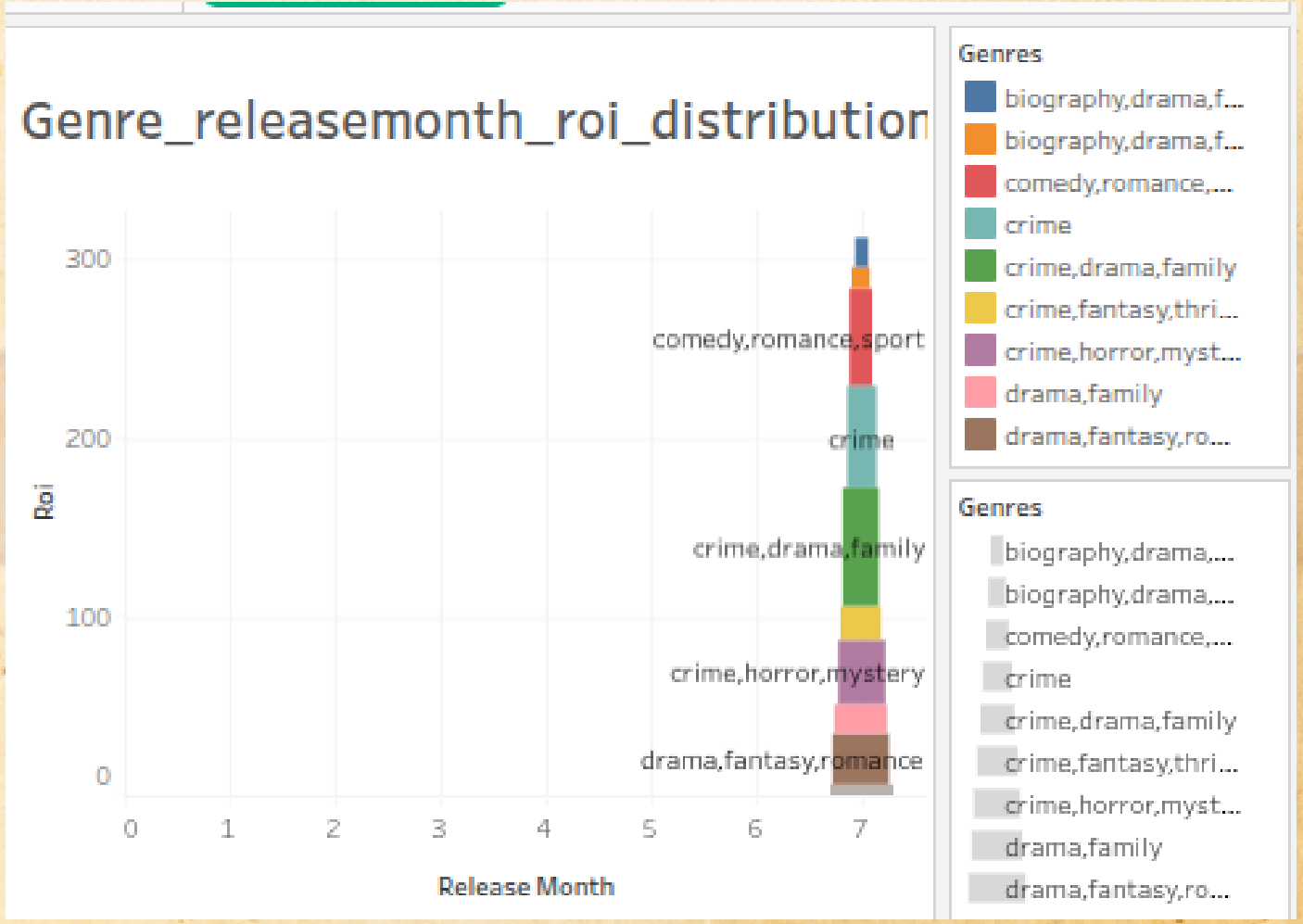
Visualizations

All visualizations were created in Python (Seaborn + Matplotlib) and Tableau for clarity and presentation versatility.

1. ROI vs Production Budget by Genre

- Insight - Low-budget films in genres like Horror and Thriller often yield very high returns.
- A high ROI isn’t necessarily tied to popularity or votes, so studios might explore lower-cost niche genres that surprise at the box office.
- Useful for - Budget allocation strategies.

See below:

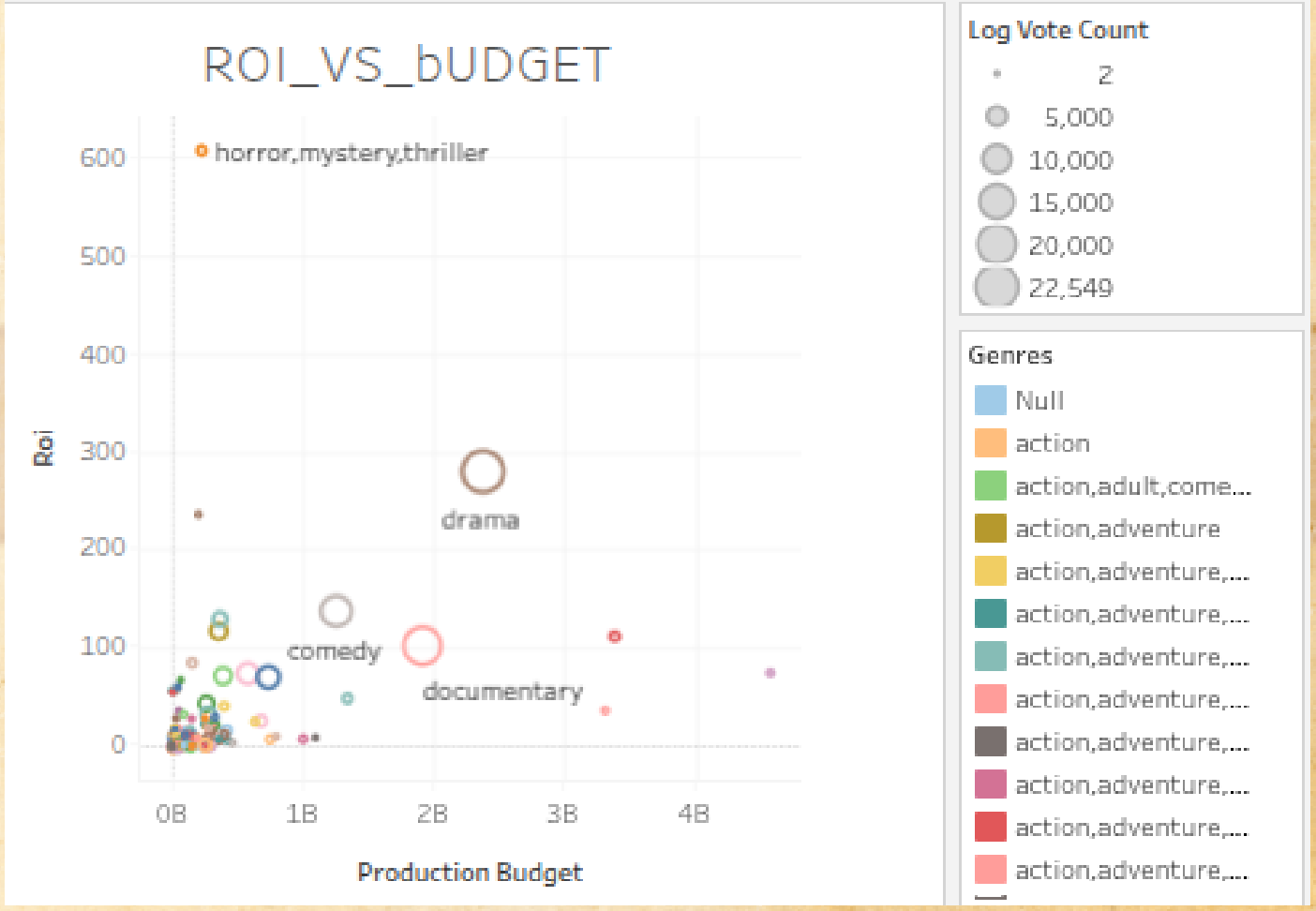

Genre_releasemonth_roi_distribution
300
200
100
0
Roi
0
1
2
3
4
5
6
7
Release Month
comedy,romance,sport
crime
crime,drama,family
crime,horror,mystery
drama,fantasy,romance
Genres
biography,drama,f...
biography,drama,f...
comedy,romance,...
crime
crime,drama,family
crime,fantasy,thri...
crime,horror,myst...
drama,family
drama,fantasy,ro...
Genres
biography,drama,...
biography,drama,...
comedy,romance,...
crime
crime,drama,family
crime,fantasy,thri...
crime,horror,myst...
drama,family
drama,fantasy,ro...

2. ROI by Release Month and Genre

-Insight: Movies released in July, October, and December show the highest ROI spikes. Aligning genre with optimal release windows could improve ROI , e.g., releasing family or adventure films during summer or holiday periods.

-Useful for: Release date planning.

See below:

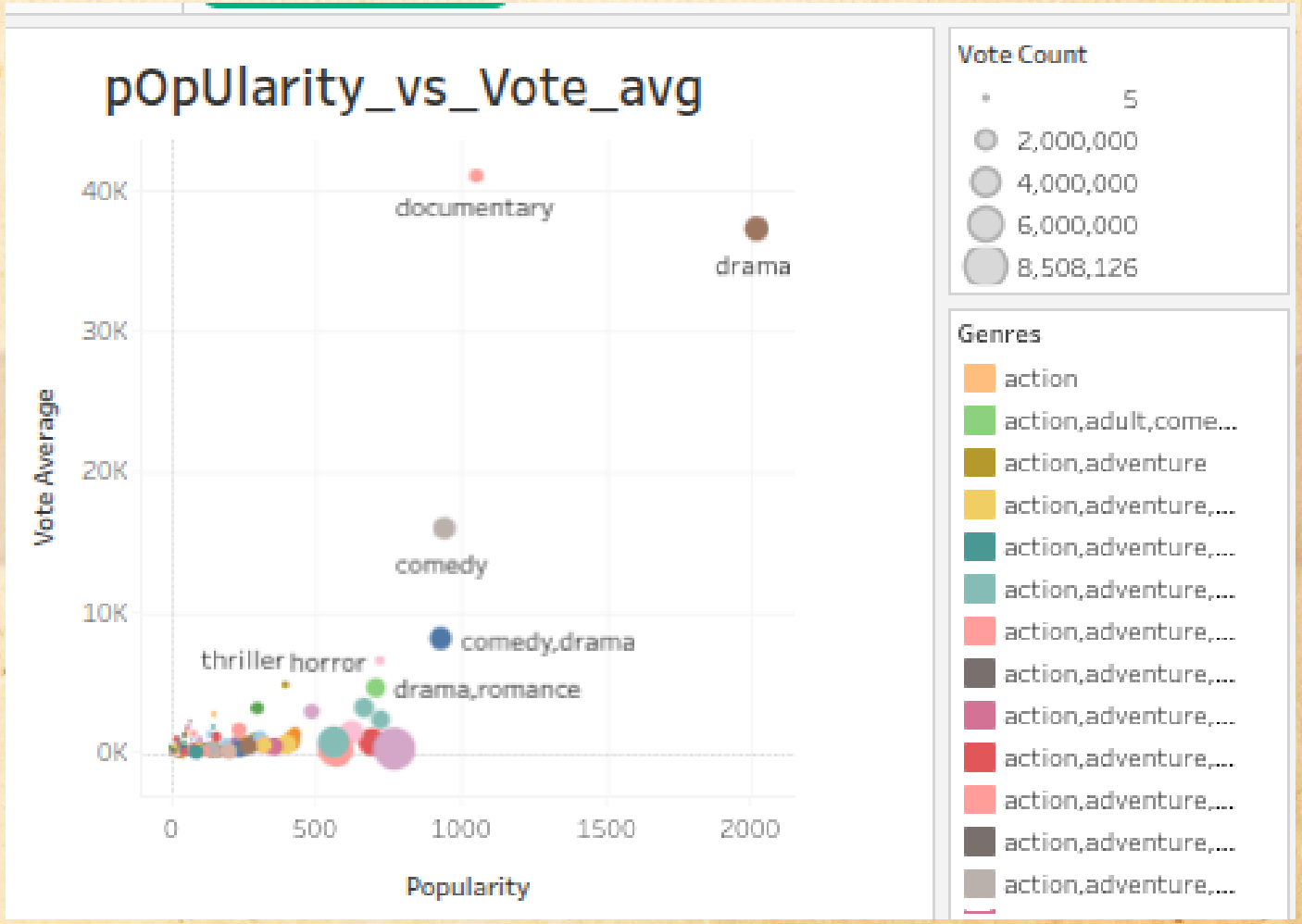

ROI_VS_bUDGET
horror,mystery,thriller
drama
comedy
documentary
Roi
600
500
400
300
200
100
0
0B
1B
2B
3B
4B
Production Budget
Log Vote Count
2
5,000
10,000
15,000
20,000
22,549
Genres
Null
action
action,adult,come...
action,adventure
action,adventure,...
action,adventure,...
action,adventure,...
action,adventure,...
action,adventure,...
action,adventure,...
action,adventure,...
action,adventure,...

3. Popularity vs Ratings by Genre

- Insight: Highly popular films do not always have the highest ratings.

- Useful for: Balancing critical and commercial appeal.

See below:This chart allows non-technical stakeholders to quickly spot genre-based clusters of success and outliers worth further review.


pOpUlarity_vs_Vote_avg
40K
30K
20K
10K
0K
Vote Average
documentary
drama
comedy
comedy,drama
thriller
horror
drama,romance
0
500
1000
1500
2000
Popularity
Vote Count
5
2,000,000
4,000,000
6,000,000
8,508,126
Genres
action
action,adult,come...
action,adventure
action,adventure,...
action,adventure,...
action,adventure,...
action,adventure,...
action,adventure,...
action,adventure,...
action,adventure,...
action,adventure,...
action,adventure,...
action,adventure,...

Summary of Recommendations

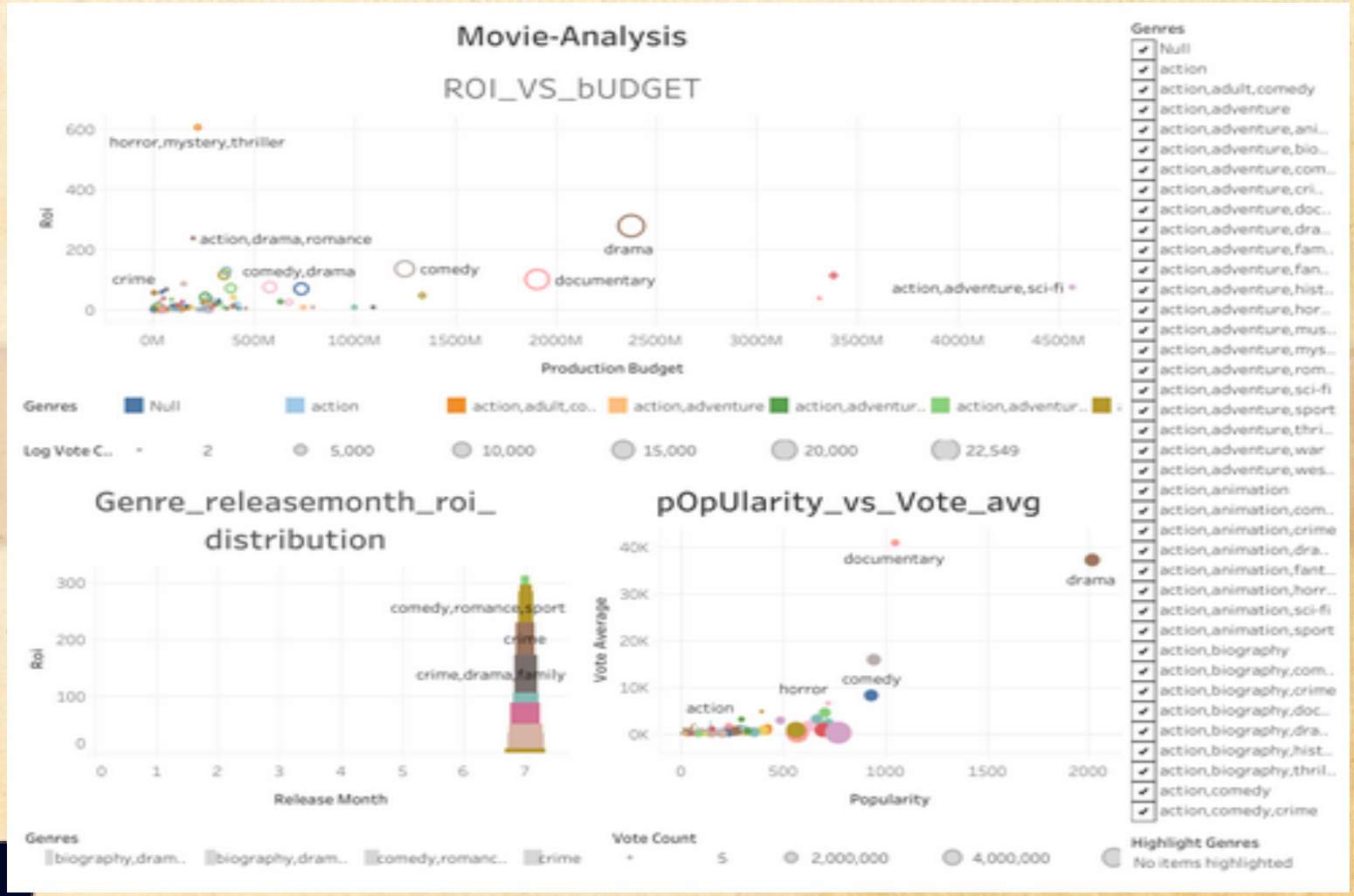
1. Company should invest in high-ROI genres like Horror and Documentary - small budgets can yield large returns.
2. Time your movie releases for Q3 and Q4 (July, October, December) - these months show strong revenue returns.
3. Don’t equate popularity with quality - viewer satisfaction (ratings) and profitability aren’t always aligned.
4. Focus on mid-range popularity movies with high ROI potential — hidden gems that outperform expectations.

Conclusion

The data shows that budget and release timing are just as critical as content. By combining viewer ratings, vote counts, popularity, and ROI.

We derived actionable strategies to guide production and release decisions. From Tableau we have interactive dashboards where one can assess how they all relate.

Its demonstrated below:


Movie-Analysis
ROI_VS_bUDGET
horror,mystery,thriller
action,drama,romance
crime
comedy,drama
comedy
drama
documentary
action,adventure,sci-fi
Roi
Production Budget
Genres
Null
action
Log Vote C..
2
5,000
10,000
15,000
20,000
22,549
Genre_releasemonth_roi_ distribution
comedy,romance,sport
crime
crime,drama,family
Release Month
pOpUlarity_vs_Vote_avg
documentary
drama
comedy
horror
action
Vote Average
Popularity
Vote Count
5
2,000,000
4,000,000
Highlight Genres
No items highlighted

Next Steps

1. Further investigate distributor-specific trends.
2. Explore actor/director influence using the principals and persons tables from IMDb.
3. Incorporate review sentiment analysis to refine quality scoring.
4. Develop a predictive model to estimate ROI based on movie attributes.

This project is accompanied by:-

I. notebook.ipynb (full analysis with Python code)

II. Tableau Public dashboard (interactive visuals)

https://public.tableau.com/app/profile/rosemary.wanjiru/viz/MovieAnalysis_17544072088810/Movie-Analysis

III. movies_metadata_cleaned.csv (final dataset)