ANALYSIS AND UNDERSTANDING ON MOVIE PERFORMANCE: VOTES, RATINGS, REVENUE & GENRE BREAKDOWN

Data Analyzed by:

Name: Stephen Muchai Kimani

Email: Stephen.Muchai@student.moringaschool.com

Business Problem

_ Overview

- The film industry is highly competitive, with millions spent on production, marketing, and distribution — yet many movies fail to deliver a return on investment. Studios and streaming platforms need deeper insights into the characteristics of high-performing films — both critically and commercially. This project aims to explore historical movie data to identify what makes a film successful.

Movie Business Pain Points

- Unpredictable movie performance: Studios invest heavily without clarity on what drives high ratings or revenue.
- Weak alignment between audience preferences and production decisions: Films may achieve high critical acclaim but underperform financially.
- Lack of data-driven greenlighting: Decisions about which movies to produce or acquire are often based on instinct or precedent rather than data.
- Limited understanding of genre and studio performance trends over time.

◈ Key Data Questions

- Which movies received the highest average ratings and why?
- Is there a strong relationship between IMDb ratings and domestic_gross?
- What genres tend to generate higher revenue or attract more votes?
- How has box office revenue trended over the last 10 years?
- Do certain studios consistently produce better-performing films (in terms of revenue and ratings)?
- Is there a correlation between number of votes and movie success?

◈ Why These Questions Matter (Business Impact)

- Strategic planning: Perception into past performance help guide future production and investment decisions.
- Market targeting: Knowing what genres and runtimes resonate with audiences can improve marketing and distribution.
- Risk reduction: Understanding key success factors lowers the risk of releasing underperforming titles.
- Studio benchmarking: Reveals which studios consistently perform well and why — useful for partnerships and acquisitions.

Data Understanding

◈ Data Sources

The dataset used has combines data from three key sources:

These sources were merged and cleaned.

1. IMDb Ratings (title.ratings.tsv.gz) Containing average ratings and number of votes for each title.
2. Box Office Mojo Gross Data (bom.movie_gross.tsv.gz), Providing domestic and foreign box office revenue by title and studio.
3. IMDb Basics (title.basics.tsv.gz) Including movie titles, genres, runtime, release year, and type.

◈ Data Representation

- Each row in the dataset represents a movie title that:

1. Is listed on IMDb
2. Has audience ratings and vote count
3. Has revenue data (domestic gross)

- The dataset includes:

1. Movies from different genres
2. Produced by various studios

◈ Sample Overview

- Observations: The dataset has 13,762 movies
- Time span: The dataset has recent movies
- Geography: Global, though separated into domestic (U.S.) and foreign

Target Variable

The target variable depends on the analysis goal. For this project, we consider two primary targets ie audience preference(AvarageRating) and revenue

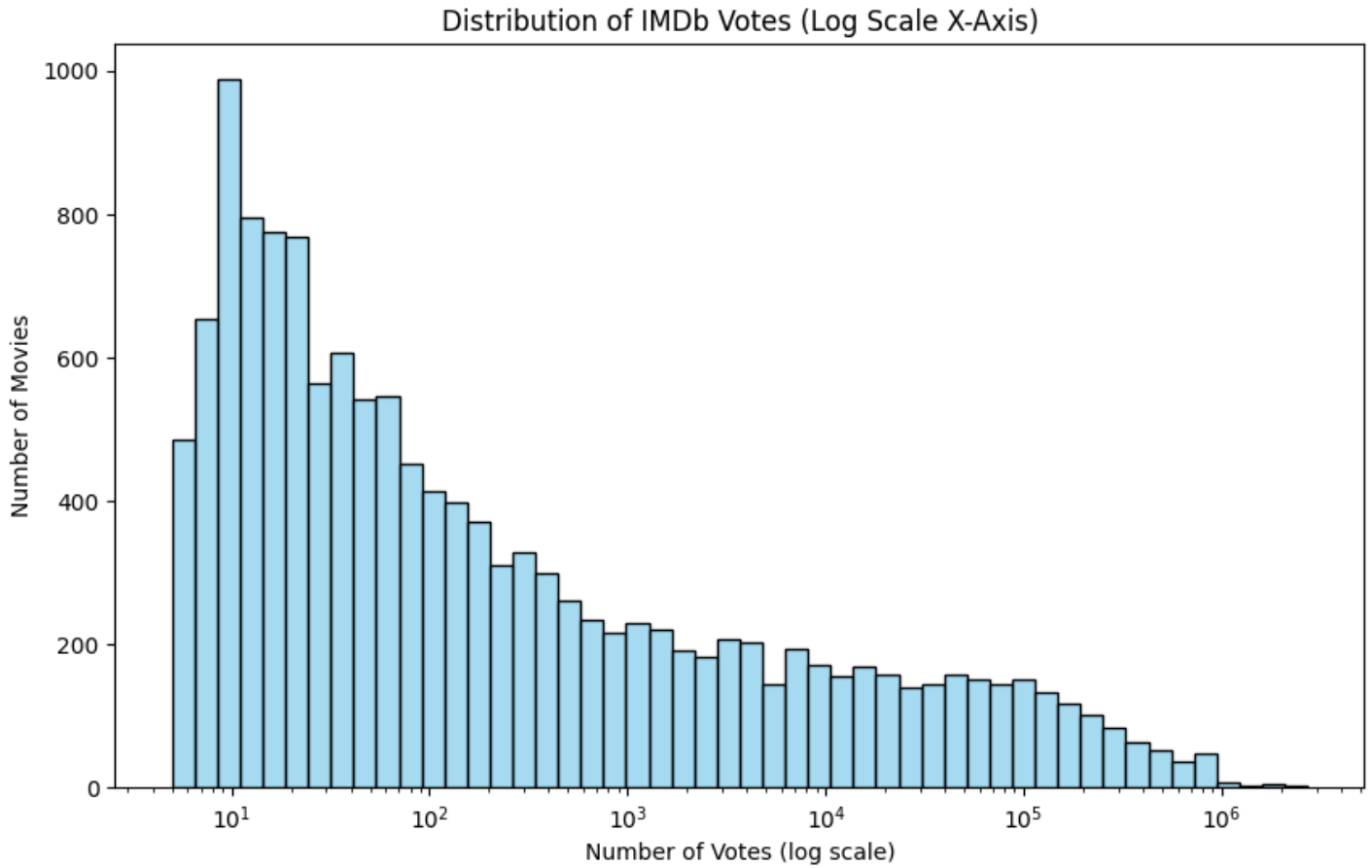
Key Variables

Variable	Description	Type
primaryTitle	Movie name	Object
averageRating	IMDb average user rating (0–10 scale)	Float
numVotes	Number of user votes on IMDb	Integer
domestic_gross	U.S. box office revenue	Float
foreign_gross	International box office revenue	Object*
year	Release year	Integer
genres	One or multiple genre tags	Object
studio	Producing or distributing studio	Object
runtimeMinutes	Movie duration	Object*

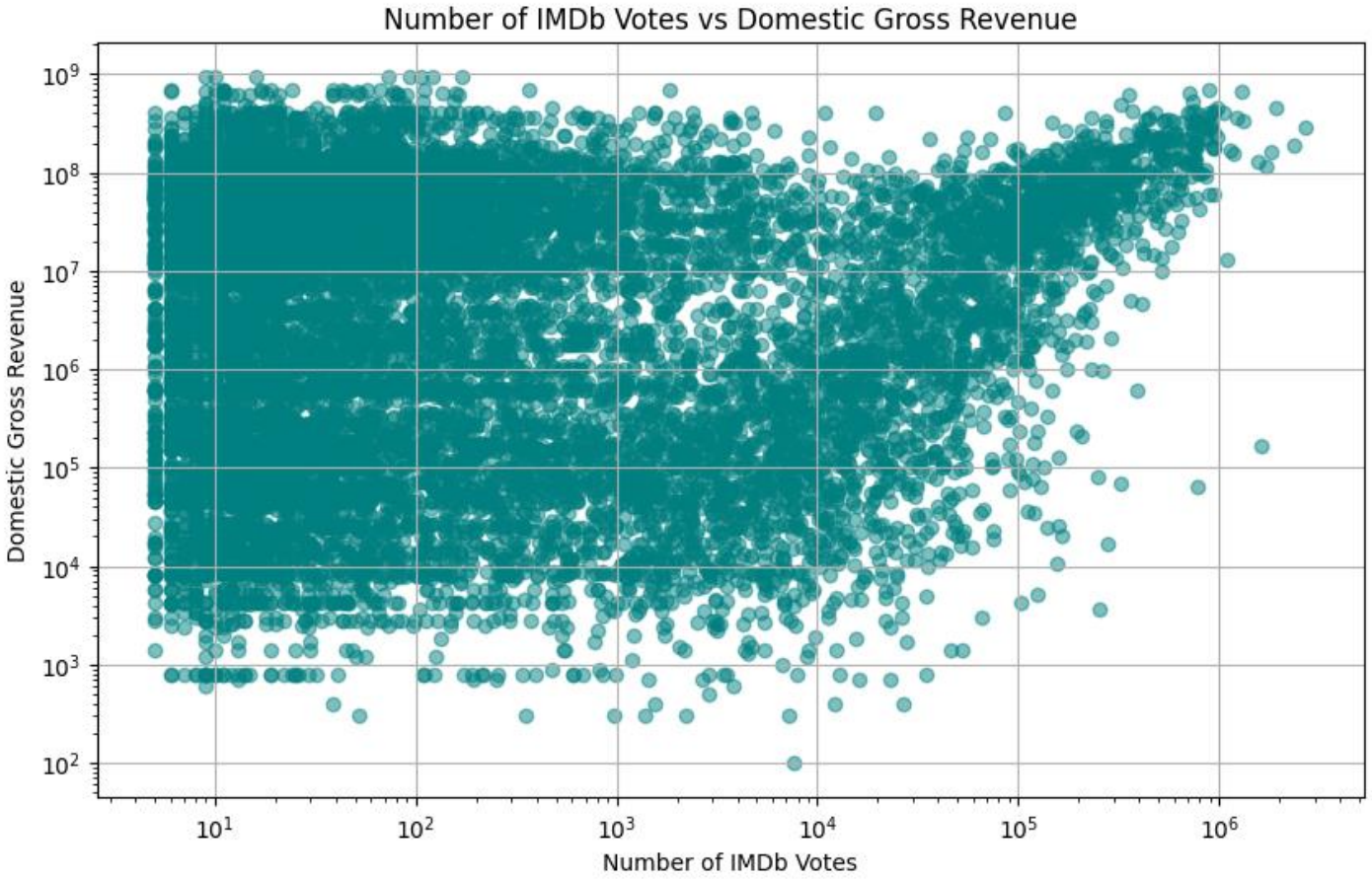
◈ Properties of the Main Variables

- avarageRating: Continuous, skewed slightly right (most movies fall between 6 and 8).
- numVotes: Discrete, long-tailed distribution (few movies get extremely high vote counts).
- domestic_gross: Continuous, highly skewed (many low earners, few blockbusters).
- Year: Discrete, useful for trend analysis.
- genres: Categorical, multi-label

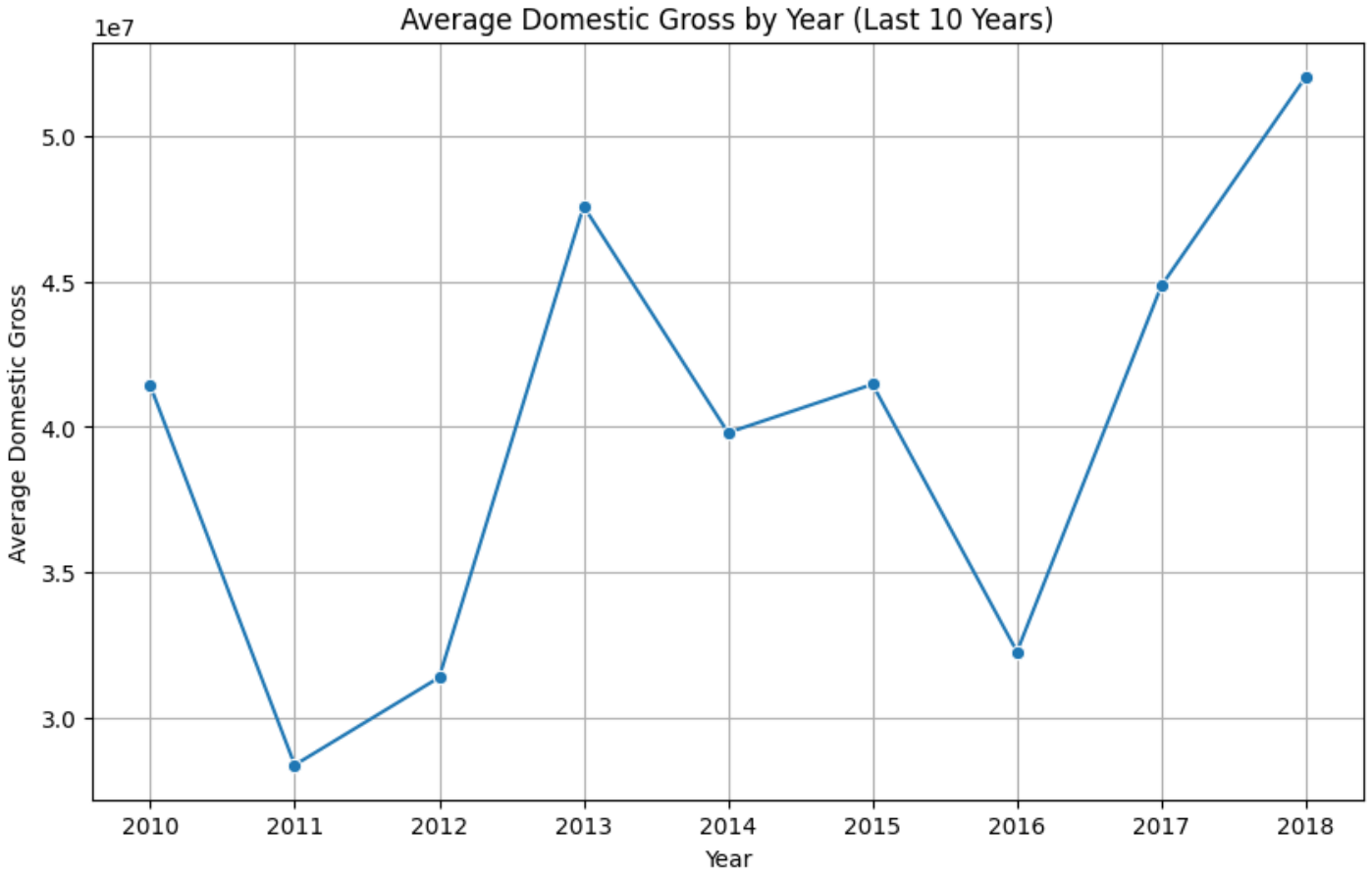
The histogram showing the distribution of IMDb votes across movies



Scatterplot of showing the relationship between Vote with Domestic_gross revenue



A line plot showing the average domestic gross by year for the last 10 years (2010–2018).



Data Modeling

◈ Analytical Approach

Exploratory data analysis (EDA) was used as the primary modeling strategy. This includes:

- Descriptive Statistics: For understanding central tendencies, variance, and distributions of key metrics such as avaragerating, numVotes and domestic_gross
- Correlation Analysis: To see the linear relationships between variables like ratings and revenue.
- Visualization Modeling: Histogram, scatterplot, bar chart and line charts were used.

◈ Modeling Iteration & Improvement

The following steps to refine the analysis:

- Data Cleaning: Removed adult titles and non-movie formats, converted runtime, gross, and year to proper numeric types , cleaned and standardized column names for usability.
- Outlier Detection: Identified and excluded extreme values where necessary to avoid skewed visuals (e.g., revenue outliers).
- Focused Time Window: Limited revenue trend analysis to the last 10 years to reflect recent viewer and market behavior.

◈ Why EDA Fit the Business Problem

EDA and segmentation provides clarity without overcomplication:

- Simple visualizations are ideal for stakeholders who may not have a technical background.
- Correlation analyses support actionable insights, such as

1. Which genres are consistently profitable
2. Do better-rated movies always earn more?
3. Which studios dominate the high-grossing films?

Evaluation

◈ Results Analysis

- A positive relationship between numVote and domestic_gross that indicate popular movies (more votes) tend to perform better financially.
- Certain genres and studios consistently outperform others in terms of ratings or revenue.
- Over the **last 10 years**, there are visible trends in domestic revenue, likely influenced by factors like streaming, franchise dominance, and global releases.

Conclusions

◈ Key Takeaways

Over 13,000 movie records combining ratings, genres, revenue were analyzed and concluded as below;

- Highly-rated movies tend to have strong audience engagement, but ratings alone do not guarantee financial success.
- Studios and genres significantly influence box office outcomes, with certain studios consistently outperforming others.
- Recent years show a shift in viewer behavior, possibly due to streaming, franchise fatigue, or changing content preferences.

◈ Recommendations

Based on the analysis, see below recommendations:

1. Invest in genres and studios that show consistent revenue performance.
2. Numvote is an early indicator of market interest.
3. Investigate underperforming genres with high average ratings for potential growth areas
4. Monitor the coming years , to align with evolving audience behavior.
5. More focus on foreign gross to gain a more complete global performance view.

◈ Future Improvements

- **Streaming data and viewer demographics** for more modern, holistic insights.
- **Benchmarking against the main competitor studios** to sharpen strategic decisions.
- **Conduct series analysis** to capture changing consumer patterns and taste over time.
- **Automate models** to forecast revenue and ratings

THANK YOU