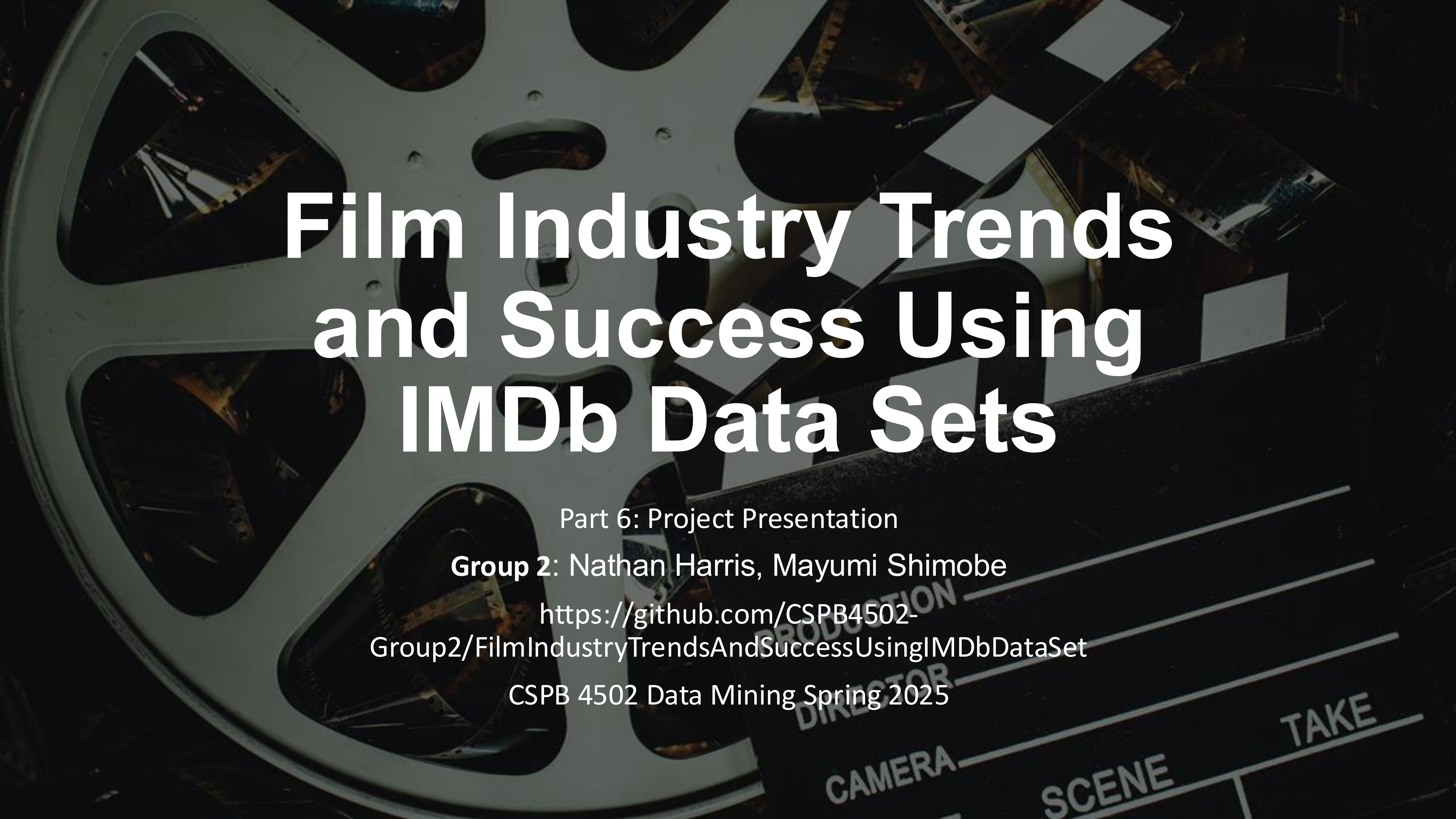


Film Industry Trends and Success Using IMDb Data Sets

Part 6: Project Presentation

Group 2: Nathan Harris, Mayumi Shimobe

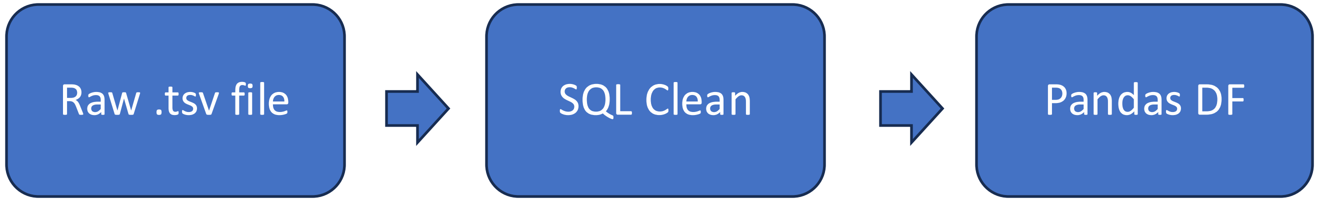
https://github.com/CSPB4502-Group2/FilmIndustryTrendsAndSuccessUsingIMDbDataSet

CSPB 4502 Data Mining Spring 2025

Questions We Asked

- Which movie attributes drive audience ratings?
- Can we predict a film’s rating or classify “success”?
- How do linear vs. ensemble models compare?

Data Preparation



- Loaded 1.5 M IMDb titles & ratings via MySQL RDS
- Cleaned: convert “\N” → NULL, remove duplicates
- Flagged low-vote titles (< 10 votes), log-transformed votes
- One-hot encoded genres for 20+ categories

Data Warehouse & Cubes

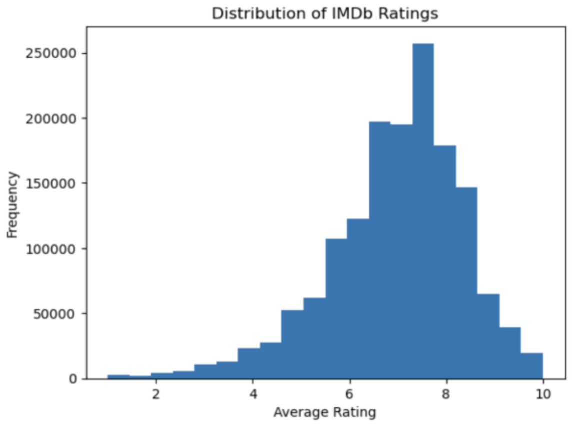
- Built imdb_integrated view (basics+ratings)
- Star schema with fact table + genre, year dimensions
- Precomputed “yearly_votes” & “genre_performance” cubes

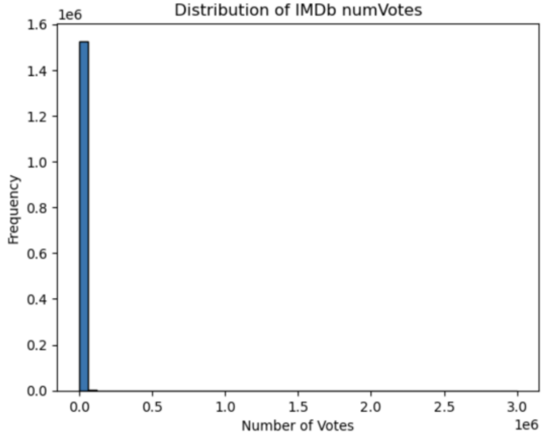
Tools & Environment

- Database: MySQL on AWS RDS (+ SQLAlchemy, %sql magic)
- Analysis: Python / Jupyter, Pandas, NumPy, SciPy
- Modeling: scikit-learn (Linear, Logistic, RF, HGB)
- Visualization: Matplotlib and Seaborn

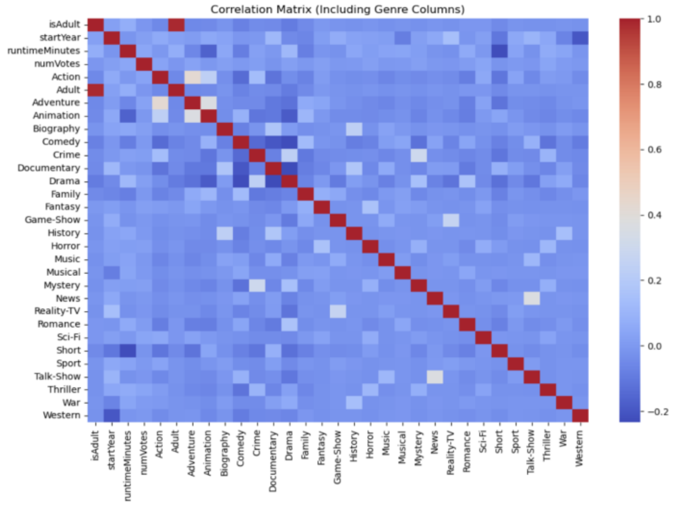
Exploratory Data Analysis

- Ratings histogram peaked at 6–7 (Gaussian-like)
- Votes histogram heavy-tailed (skew ≈ 57) → log^{1+} transform
- Correlation heatmap: nearly zero pairwise ρ → non-linear focus

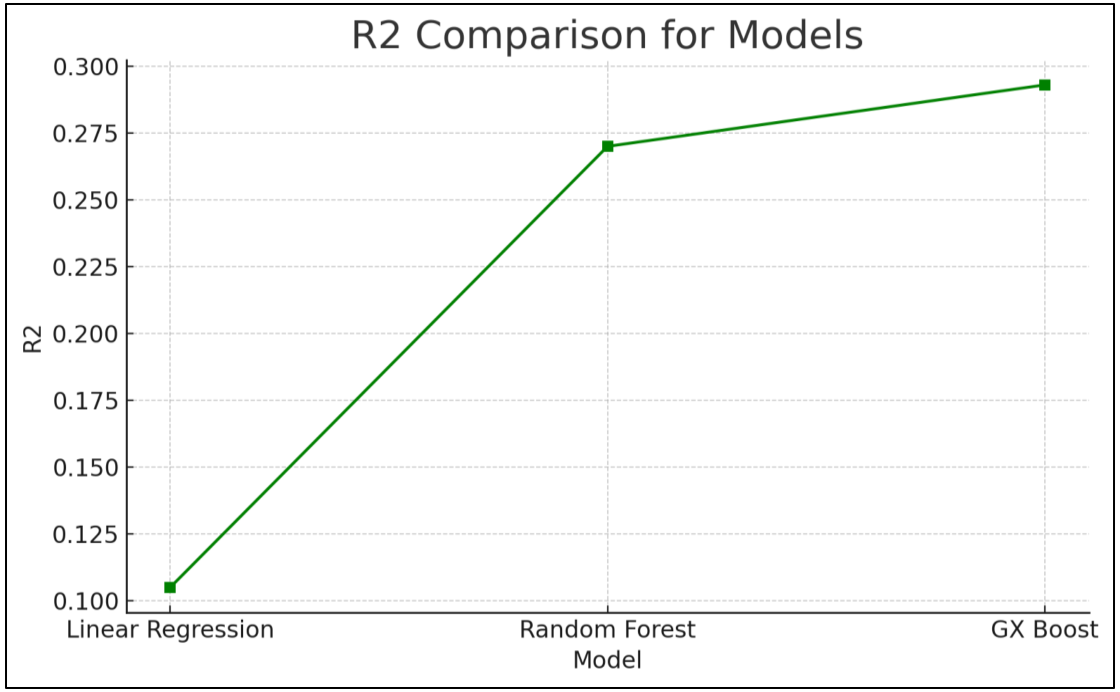




```
Skewness of numVotes: 57.31708778938491
```



Regression Models



Model	RMSE	R^2
Linear Regression	0.946	0.105
Random Forest	1.135	0.270
GX Boost	1.117	0.293

Classification Models

Model	Acc	Recall	Precision
Logistic Regression	0.79	0.00	0.50
Random Forest	0.56	0.87	0.30
GX Boost	0.80	0.07	0.67

Knowledge Gained

- Vote volume is critical but skewed → log transform
- Weak linear correlations justify non-linear / ensemble methods
- Class imbalance dominates classification performance

Applications

- **Studios** forecast ratings pre-release to guide budgets
- **Marketing** target campaigns to high-potential titles
- **Platforms** integrate scores into recommendation engines

Thank you!