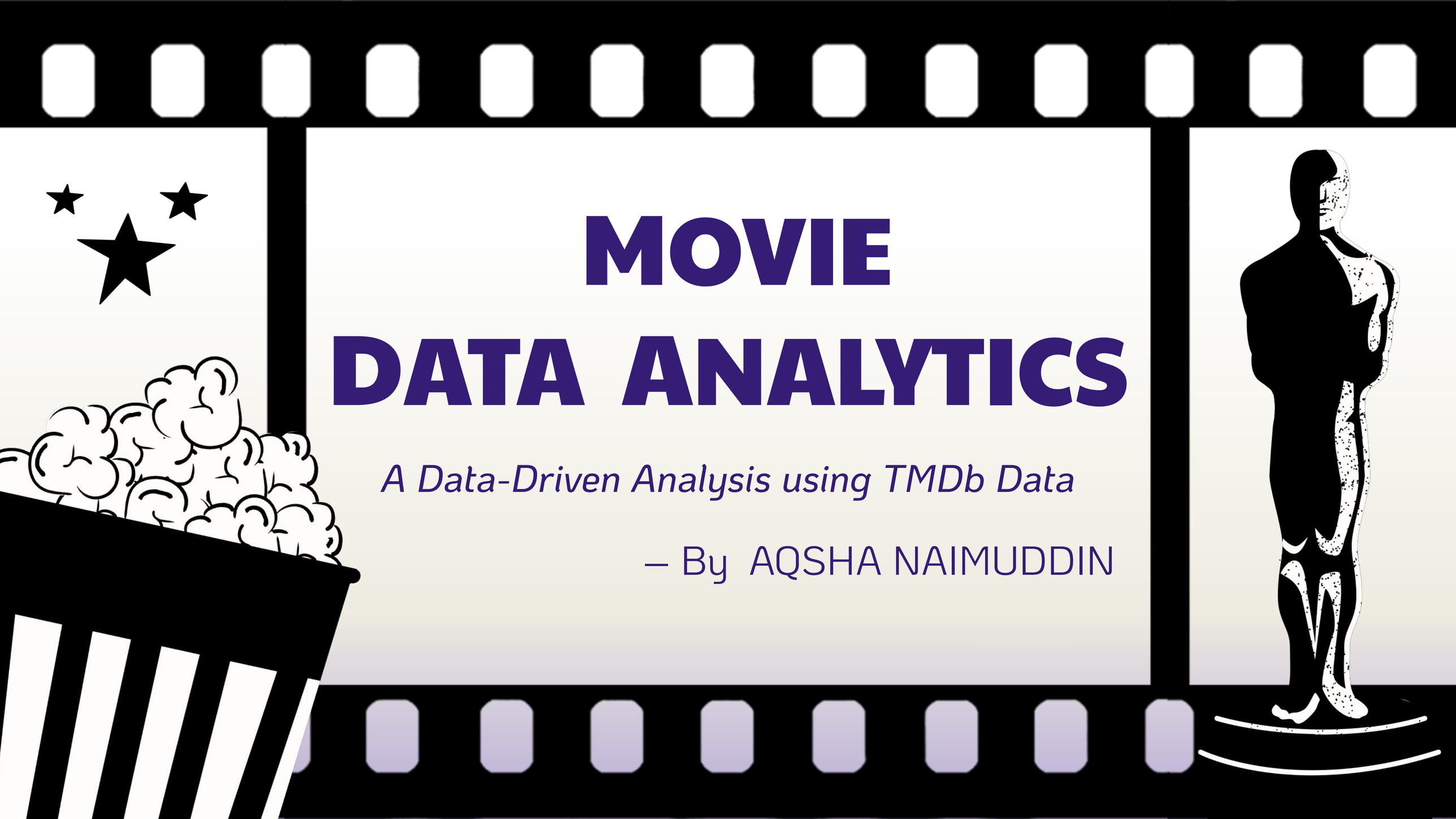
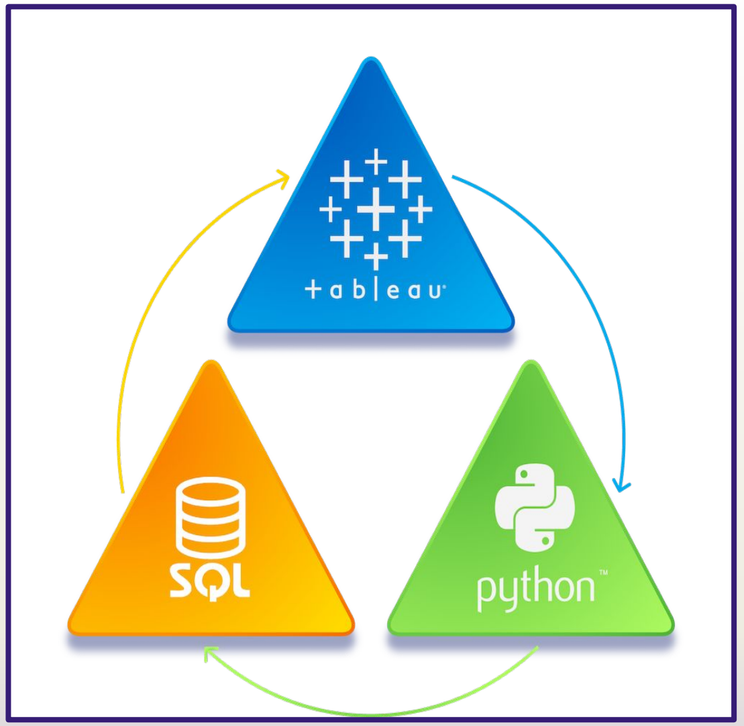
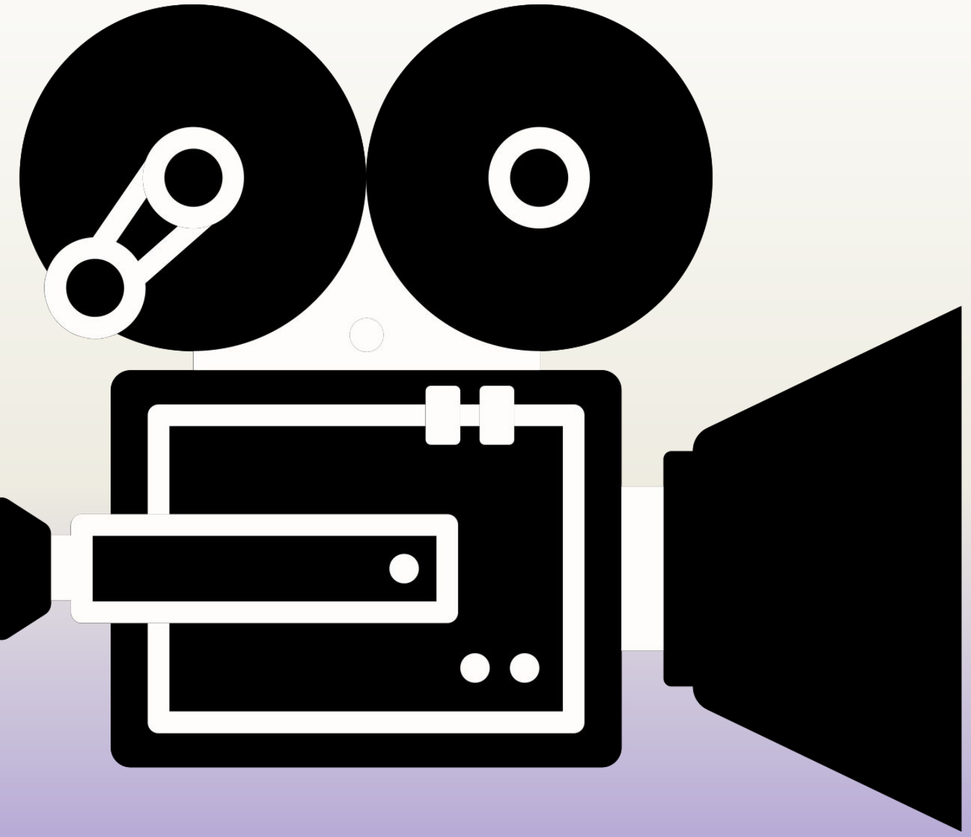

MOVIE
DATA ANALYTICS
A Data-Driven Analysis using TMDb Data
– By AQSHA NAIMUDDIN

Our Journey in this Project

- Data acquisition, cleaning and database creation
- Exploratory Data Analysis and descriptive insights
- Statistical validation of key business questions
- Predictive and clustering models to understand drivers of success
- Interactive dashboards for decision makers
- Recommendations and future roadmap



BUSINESS OBJECTIVE



1. Help studios identify which genres, budgets, and talents drive revenue and ratings

2. Build data-driven dashboards to track performance and profitability

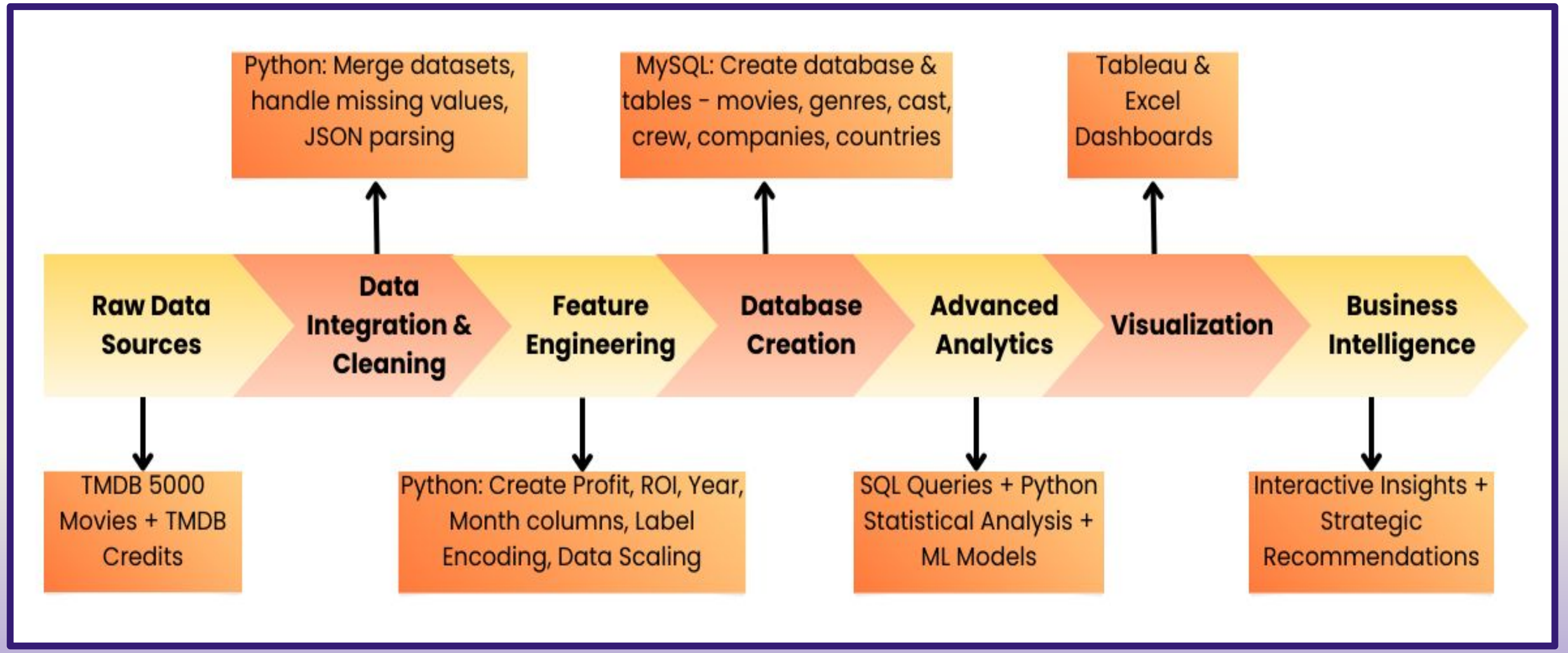
3. Deliver predictive models for revenue, profit, and rating classification

4. Recommend investment and marketing strategies backed by evidence

KEY FOCUS AREAS

🎬 PERFORMANCE ANALYSIS	Release trends by year, month & genre
💰 REVENUE & PROFITABILITY	Revenue, Profit & ROI across genres, companies, directors
⭐ CAST & CREW IMPACT	Top actors/directors by revenue & ratings
🌍 GLOBAL PRODUCTION TRENDS	Production country mapping, budgets by year, production company
⏱️ RUNTIME & RATINGS	Average runtime, rating distributions, stratified sampling
📊 PREDICTIVE MODELS	Revenue prediction, rating classification, profit prediction

Data Preparation Pipeline

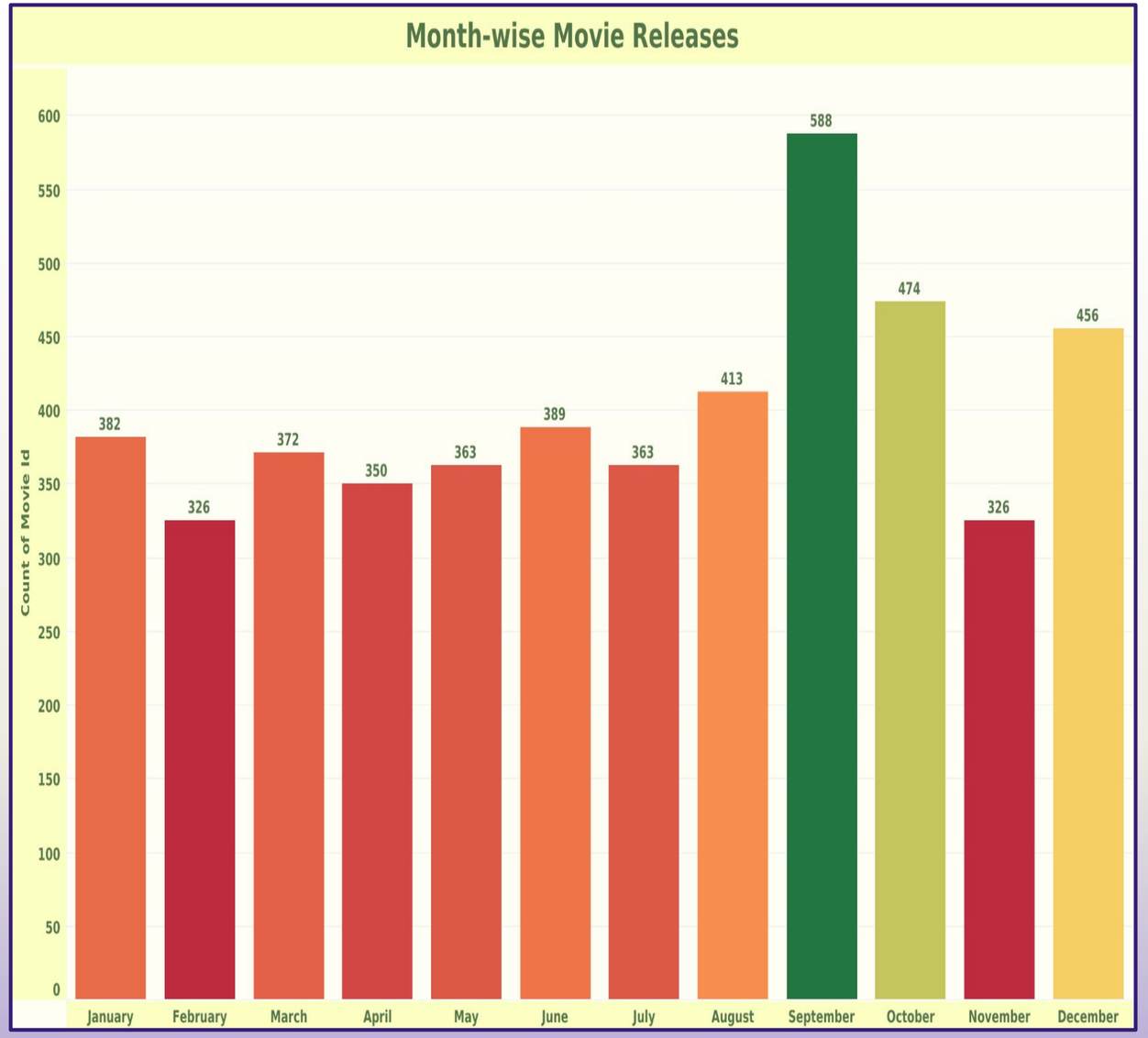


01

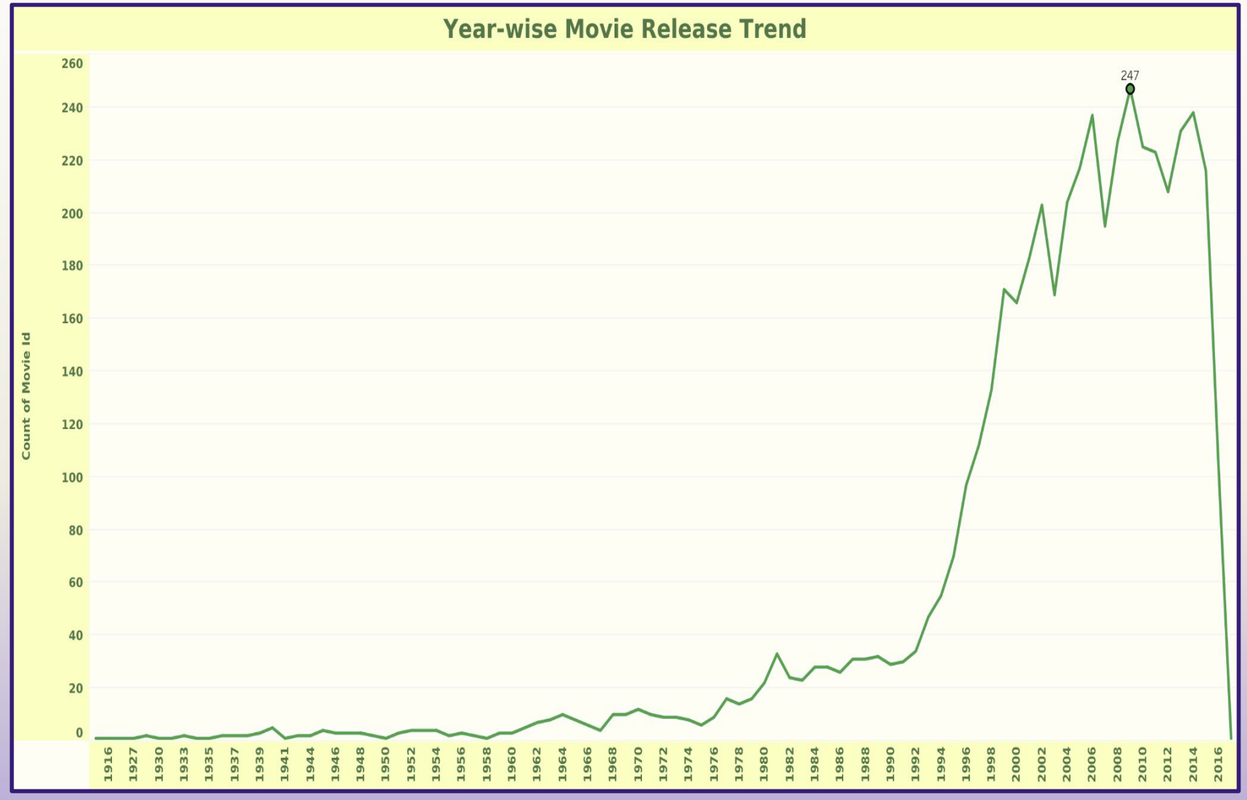
EXPLORATORY INSIGHTS

Movies Release Patterns Over Time

(RELEASE TRENDS)



- Most releases in 2009, sharp growth after 2000s
- Peak release months: September, October and December (holiday season)
- Seasonal patterns visible across decades



genre_name	avg_rating
History	6.719796954314720
War	6.7138888888888900
Drama	6.38859381802349
Music	6.355675675675670
Foreign	6.352941176470590
Animation	6.341452991452990
Crime	6.274137931034490
Documentary	6.238181818181810
Romance	6.207718120805380
Mystery	6.183908045977020
Western	6.1780487804878100
Adventure	6.15696202531646
Fantasy	6.096698113207550
Family	6.029629629629630
Thriller	6.0109890109890100
Science Fiction	6.005607476635510
Action	5.98951473136915
Comedy	5.945586527293840
TV Movie	5.6625
Horror	5.626589595375720

GENRE PERFORMANCE

TOP RATED GENRES : 1. HISTORY 2. WAR 3. DRAMA

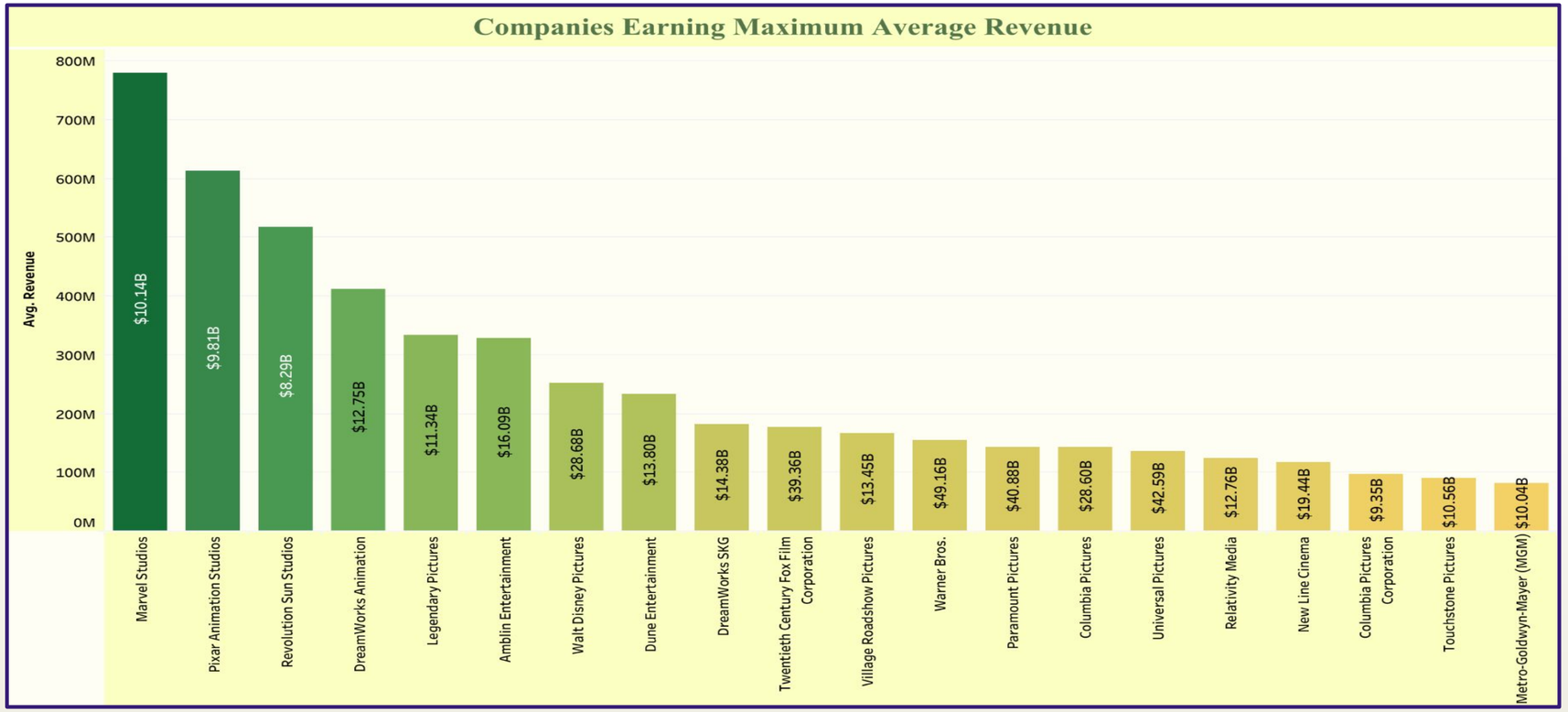
HIGHEST REVENUE : 1. ACTION 2. ADVENTURE 3. FANTASY

- *Drama* & *Comedy* dominate in count, but Action & Adventure lead in high-revenue potential
- Genre mix changed over time with rising share of Sci-Fi and Fantasy showing changing audience preferences.
- Drama has remained the most consistent genre for decades, and is still going strong.

Studios Driving The Industry

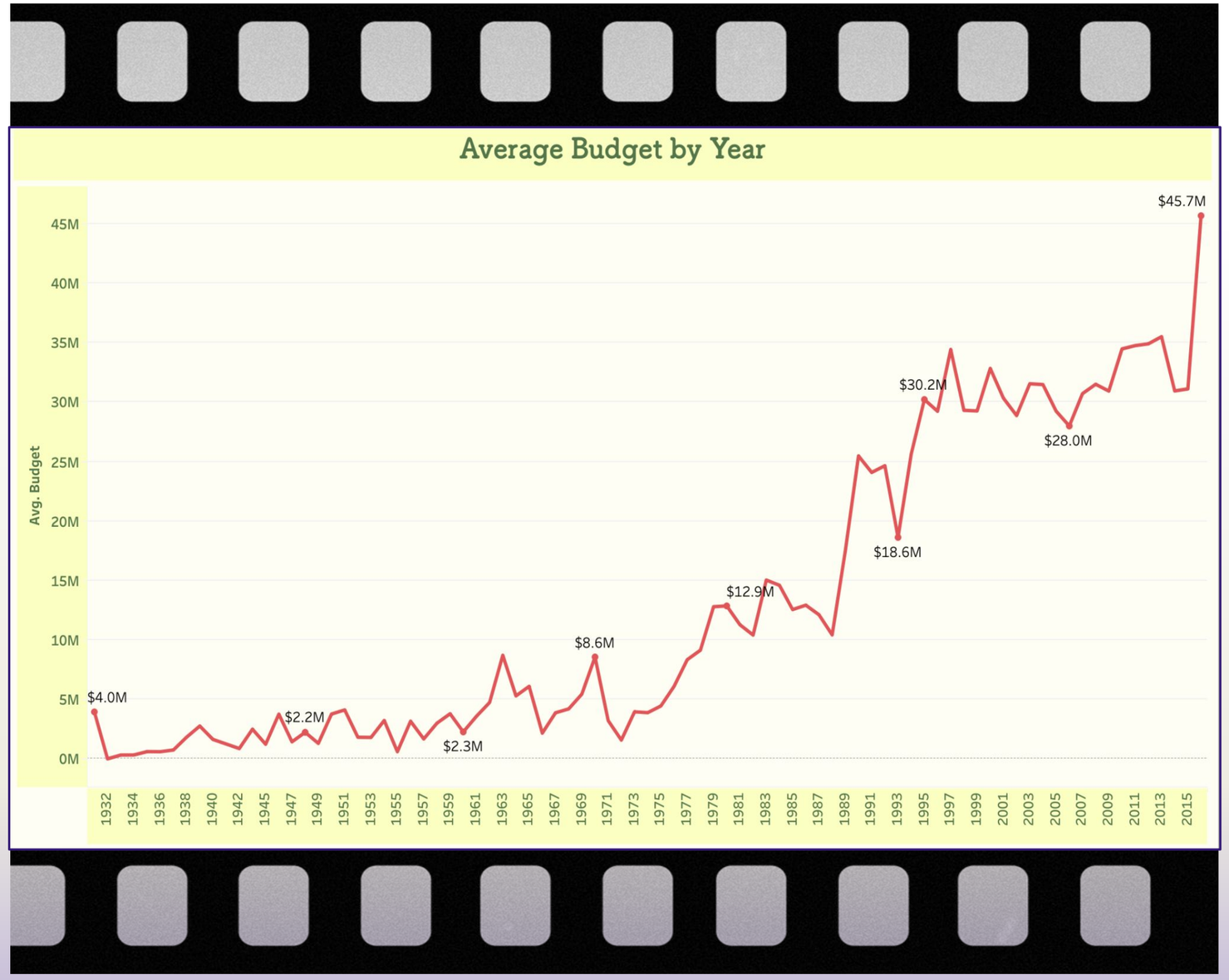
(PRODUCTION COMPANY INSIGHTS)

company_name	movie_count
Warner Bros.	319
Universal Pictures	311
Paramount Pictures	285
Twentieth Century Fox Film Corporation	222
Columbia Pictures	201
New Line Cinema	165
Metro-Goldwyn-Mayer (MGM)	122
Touchstone Pictures	118
Walt Disney Pictures	114
Relativity Media	102
Columbia Pictures Corporation	96
Miramax Films	94
Village Roadshow Pictures	81
DreamWorks SKG	79
United Artists	75



- *Warner Bros.* leads in movie count, followed by *Universal Pictures*.
- High-average revenue companies, like *Marvel Studios* & *Pixar Animation Studios* often produce fewer but bigger hits.
- The contrast between “high volume” studios and “high revenue per movie” studios highlights different business models and risk strategies.

BUDGET EVOLUTION



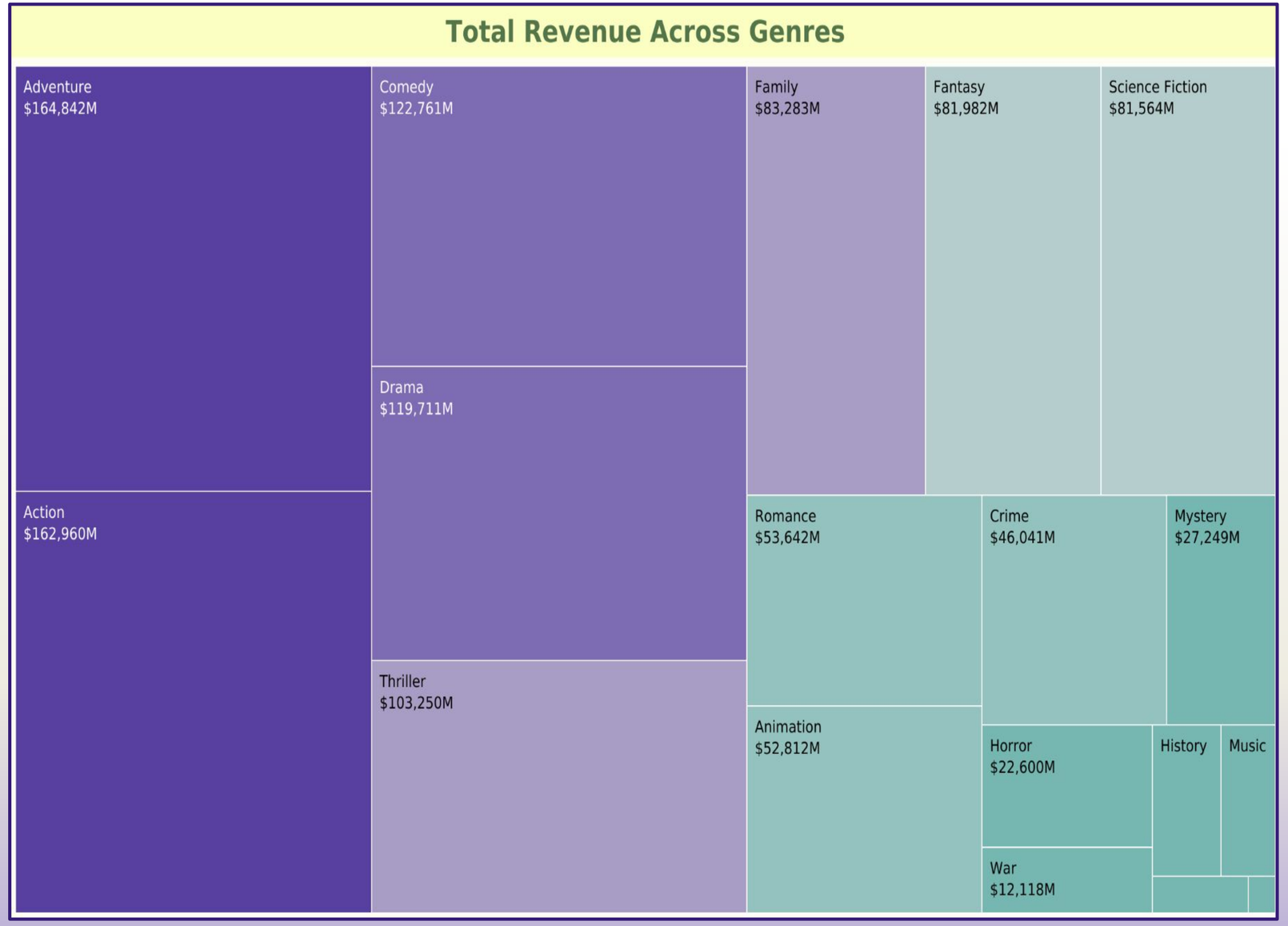
- Steady low budgets dominated early decades (1930s–1960s) , then a significant jump in the 1980s–1990s.
- Peak observed at $45.7M in 2016, indicating modern high-cost filmmaking.
- Rise reflects greater competition, global markets, and franchise-driven production strategies.
- The trend also mirrors advances in technology, marketing spends, and audience expectations.

Revenue & Profit Drivers

HIGHEST REVENUE :
AVATAR ($2.79B)
STUDIOS : 20th Century Fox, Dune Entertainment, Ingenious Film Partners & Lightstorm Entertainment.

LOWEST REVENUE :
A CHRISTMAS STORY ($19.29M)
STUDIOS : Christmas Tree Films, MGM Studios

- Adventure & Action genres dominate top revenue brackets
- Profitability insights reveal ROI not always tied to high budgets, with Drama and Comedy having maximum ROI.



Talent Shaping Success

(CAST & CREW IMPACT)

crew_name	Average_Rating
Simon Scotland	10
Danny Hambrook	10
Gary Sinyor	10
Paul Simpkin	10
Nigel Savage	10
Rohit Jugraj	9.5
Milton S. Gelman	9.3
Paul L. Newman	9.3
Joseph Brad Kluge	9.3
Pancho Alcaine	9.3

- Top 10 actors drive significant box-office revenue
- Directors with highest average ratings often specialize in niche genres & deliver critically acclaimed films, enhancing studio credibility.

actor_name	Total_Revenue
Stan Lee	17364063582
Samuel L. Jackson	14806065788
Frank Welker	11614837160
John Ratzenberger	11038044745
Hugo Weaving	10822190781
Cate Blanchett	9726416776
Ian McKellen	9710670395
Jess Harnell	9633458775
Morgan Freeman	9275477679
Tom Cruise	8993387534

- Key departments like Production, Sound, and Arts employ the largest crews, reflecting their influence on final quality.

02

STATISTICAL VALIDATIONS

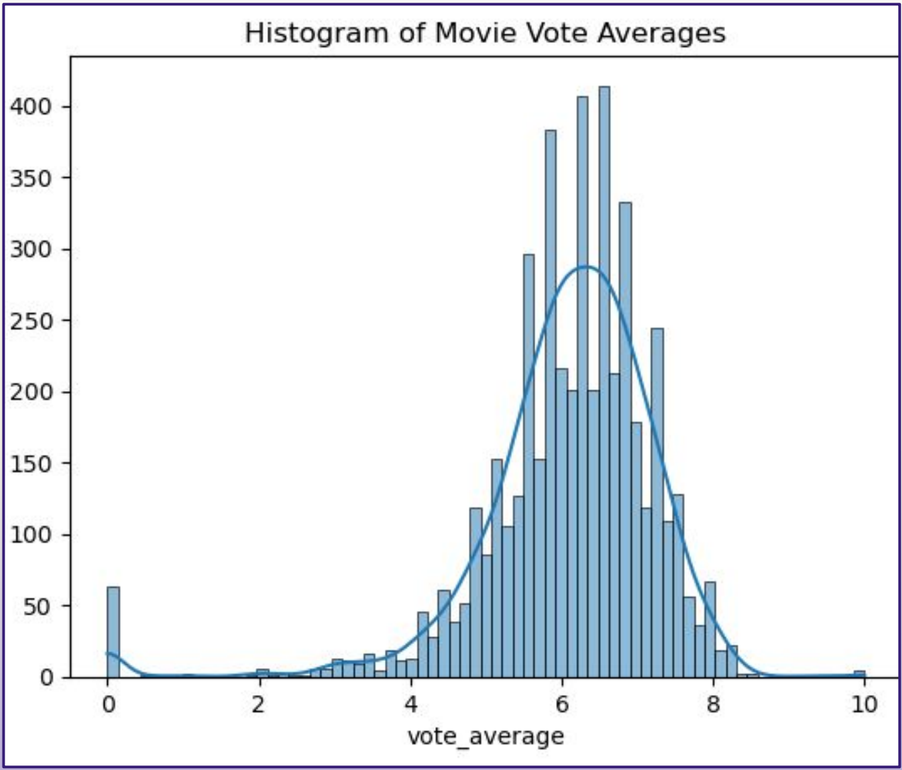
```
Total Action Movies: 1154
Action Movies with Revenue > 500M: 80
Conditional Probability: 0.07
```

```
Sample Mean Budget: 29045039.88
Z-Statistic: -120.76
P-Value: 0.0000
Z-critical value (two-tailed): 1.96
```

Statistical Conclusion

i. p-value = 0.00 < alpha = 0.05 : REJECT Null Hypothesis

ii. Z-Critical = 1.96 < |Z-Statistic| = 120.76 : REJECT Null Hypothesis



Testing Our Assumptions

- Conditional probability of Action movies earning >$500M: 7%
- Average budgets significantly differ from $100M (Z-test p<0.05). The industry's typical production budget is either much higher or lower than the assumed benchmark.
- Ratings (Vote Averages) distribution is not perfectly normal (Skew=-1.96, Kurtosis=7.79).There's a heavier left tail and a higher peak than a true normal curve.

Validating Sampling Accuracy

- Compared population vs stratified sample average runtimes per genre.
- Sample means closely matched population means (<5% variance).
- Variance in niche genres (like Western or Music) highlighted smaller sample sizes but still showed meaningful trends.
- Confirms reliability of stratified sampling for further studies.

genre_name	Population Avg Runtime	Sample Avg Runtime
Action	110.544194	107.466667
Adventure	111.332911	104.833333
Animation	89.923077	90.533333
Comedy	100.030197	97.600000
Crime	109.666667	106.300000
Documentary	92.963636	104.066667
Drama	113.265564	110.433333
Family	97.298246	94.166667
Fantasy	107.278302	107.033333
Foreign	110.617647	107.833333
History	135.989848	135.566667
Horror	95.949904	96.066667
Music	109.924324	113.633333
Mystery	109.591954	107.900000
Romance	109.379195	108.200000
Science Fiction	107.478505	107.800000
TV Movie	85.625000	85.625000
Thriller	107.544741	102.466667
War	131.833333	130.900000
Western	117.353659	126.833333

03

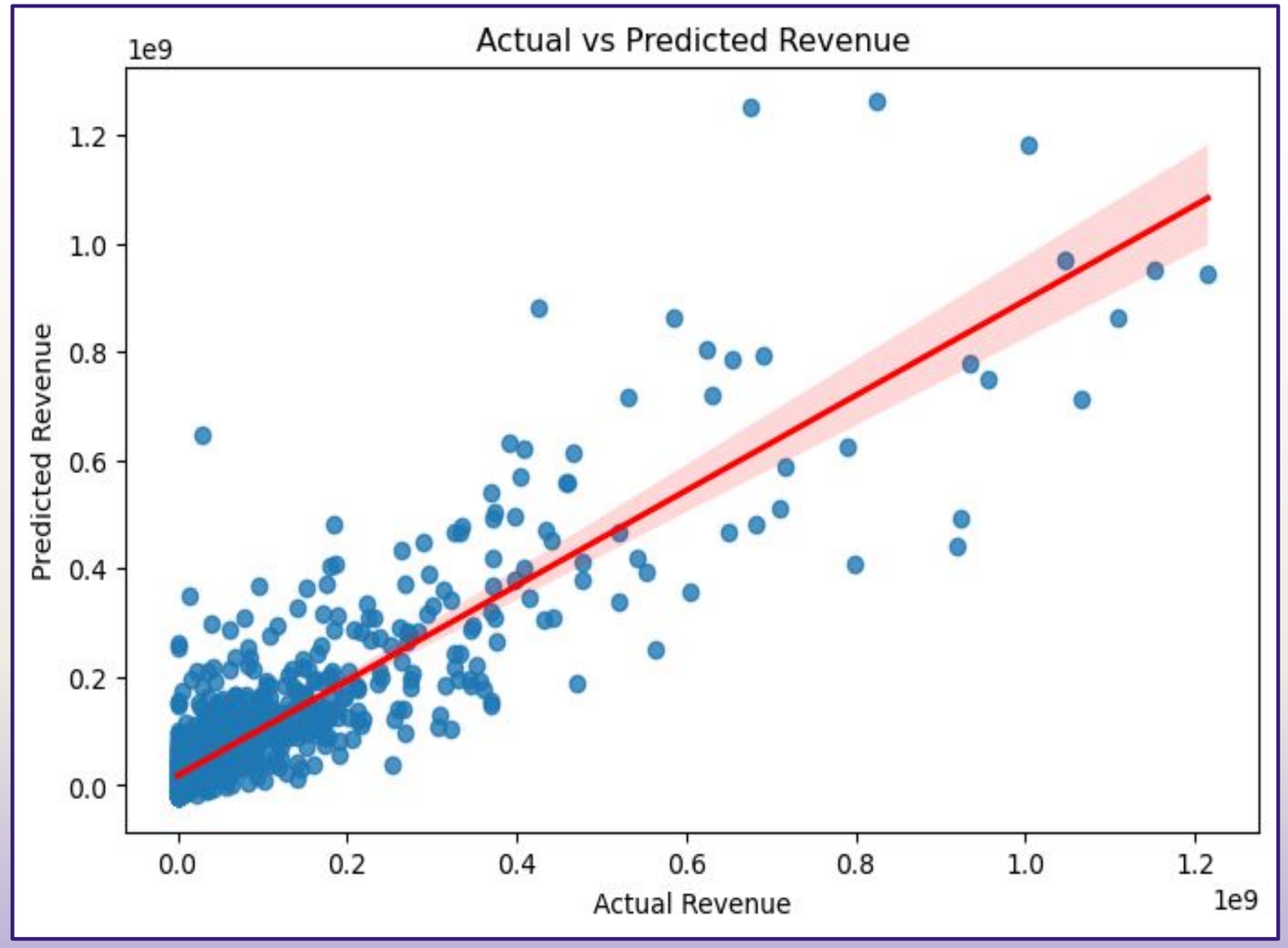
PREDICTIVE & CLUSTERING MODELS

1. Predicting Revenue from Key Drivers

(MULTIPLE LINEAR REGRESSION)

```
R² Score: 0.75
MAE: 46451512.66
RMSE: 80743724.62
```

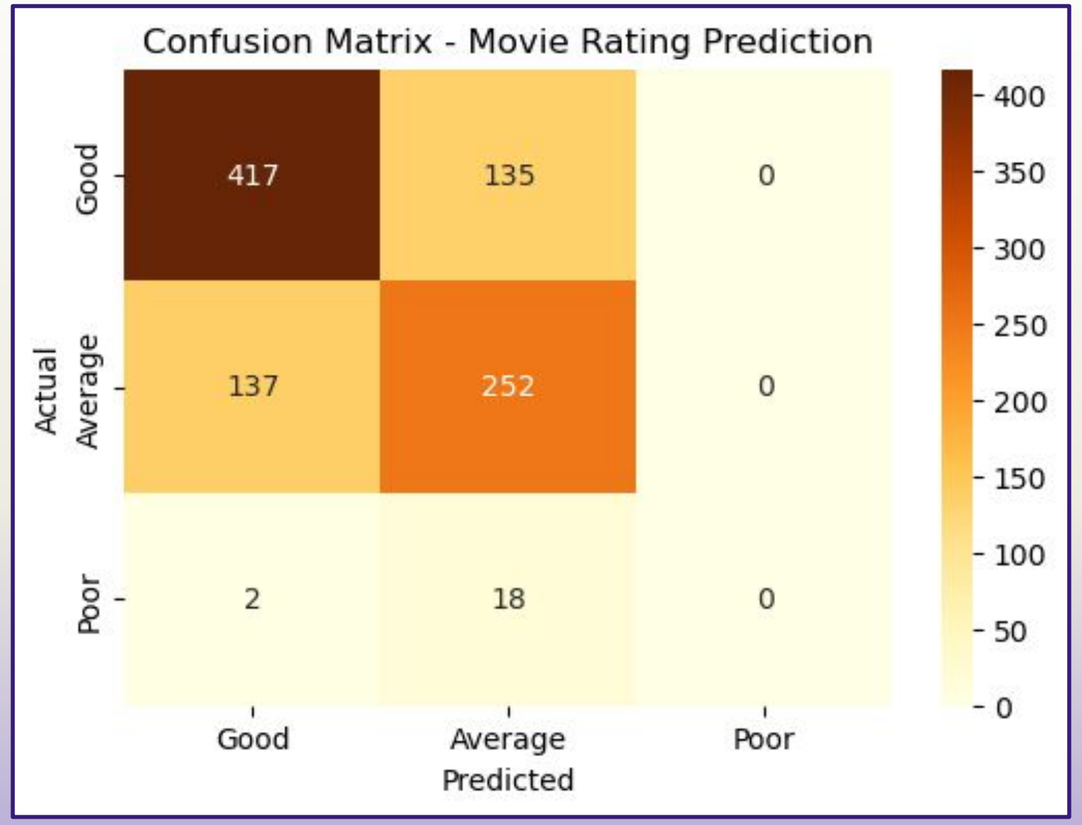
- Features: budget, runtime, vote_average, vote_count, popularity
- Model explains 75% of variance (R^2=0.75)
- Budget emerges as strongest predictor, followed by vote_average and runtime.



2. Classifying Movies into Good, Average, Poor

(LOGISTIC REGRESSION MODEL)

```
Accuracy: 0.70
Precision: 0.68
Recall: 0.70
```



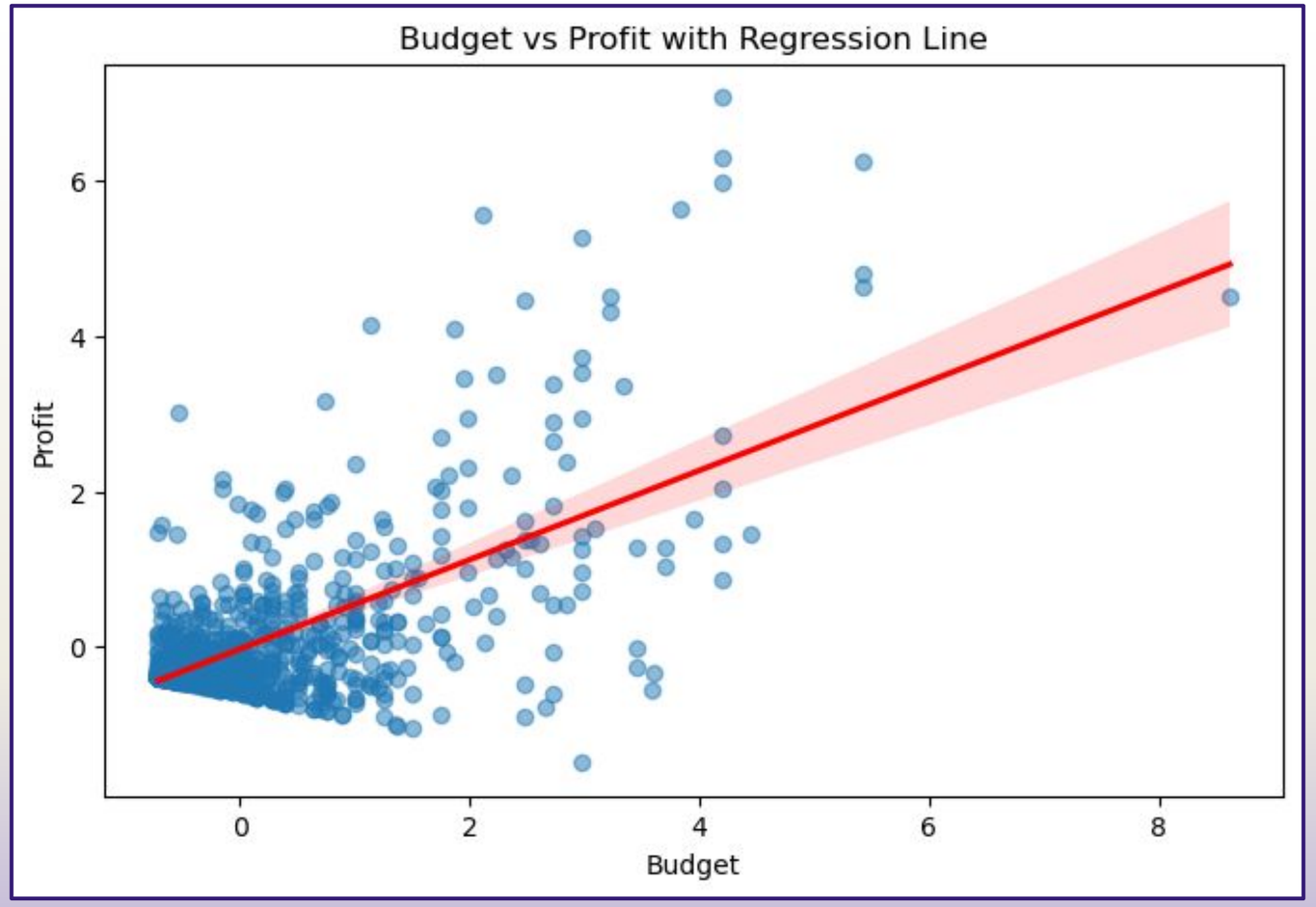
- Model correctly predicts 70% of movies' ratings, showing a reasonably good performance for a simple model using only budget, runtime, and vote count.
- About 68% of predicted ratings are correct, & 70% of actual ratings are correctly retrieved.
- Performs well for Good/Average classes, needs balancing for Poor
- Offers early indicator of likely critical reception

3. Linking Budget to Profitability

(SIMPLE LINEAR REGRESSION)

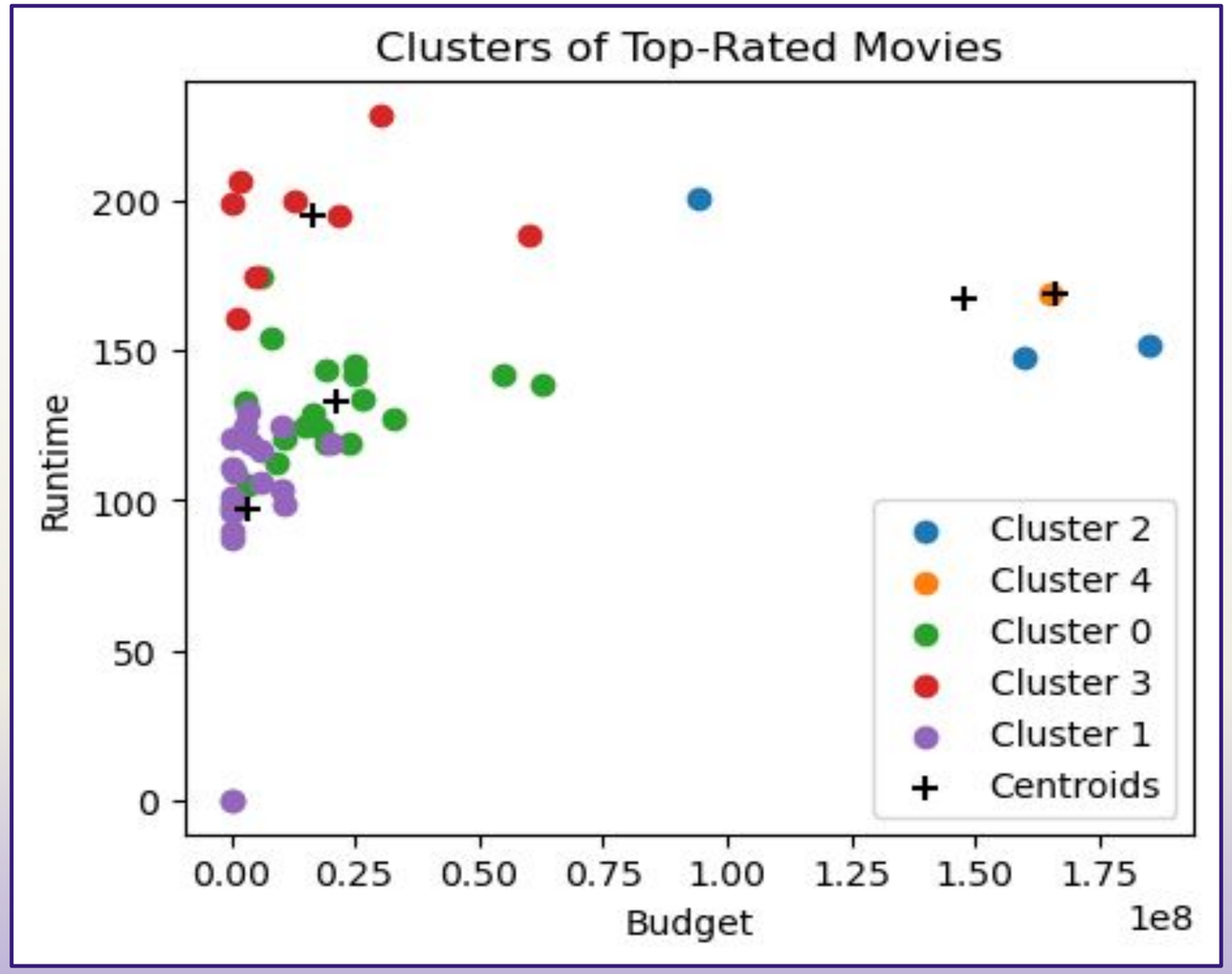
```
R2 Score: 0.43
MAE: 0.40
RMSE: 0.71
```

- The scatter plot shows a positive relationship indicating that as movie budget increases, profit generally increases too.
- Moderate predictive power, indicates other factors at play.
- Useful baseline for ROI forecasting.



4. Discovering Movie Archetypes

(K - MEANS CLUSTERING)



- K-Means identified 5 clusters of top-rated movies
- Clear patterns by budget and runtime (e.g., Indie vs Blockbuster)
- Supports tailored marketing and investment strategies

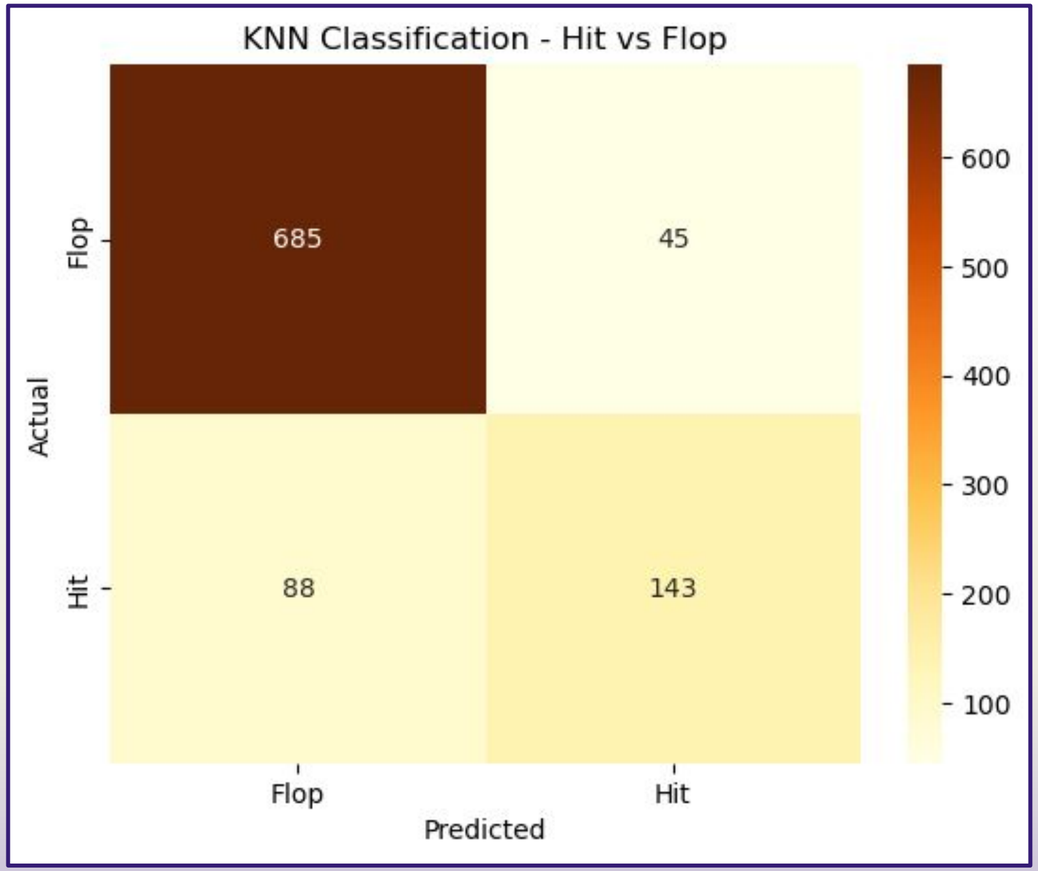
5. Movie Success Classification

(K-NEAREST NEIGHBOURS)

Accuracy: 0.86

	precision	recall	f1-score	support
0	0.89	0.94	0.91	730
1	0.76	0.62	0.68	231

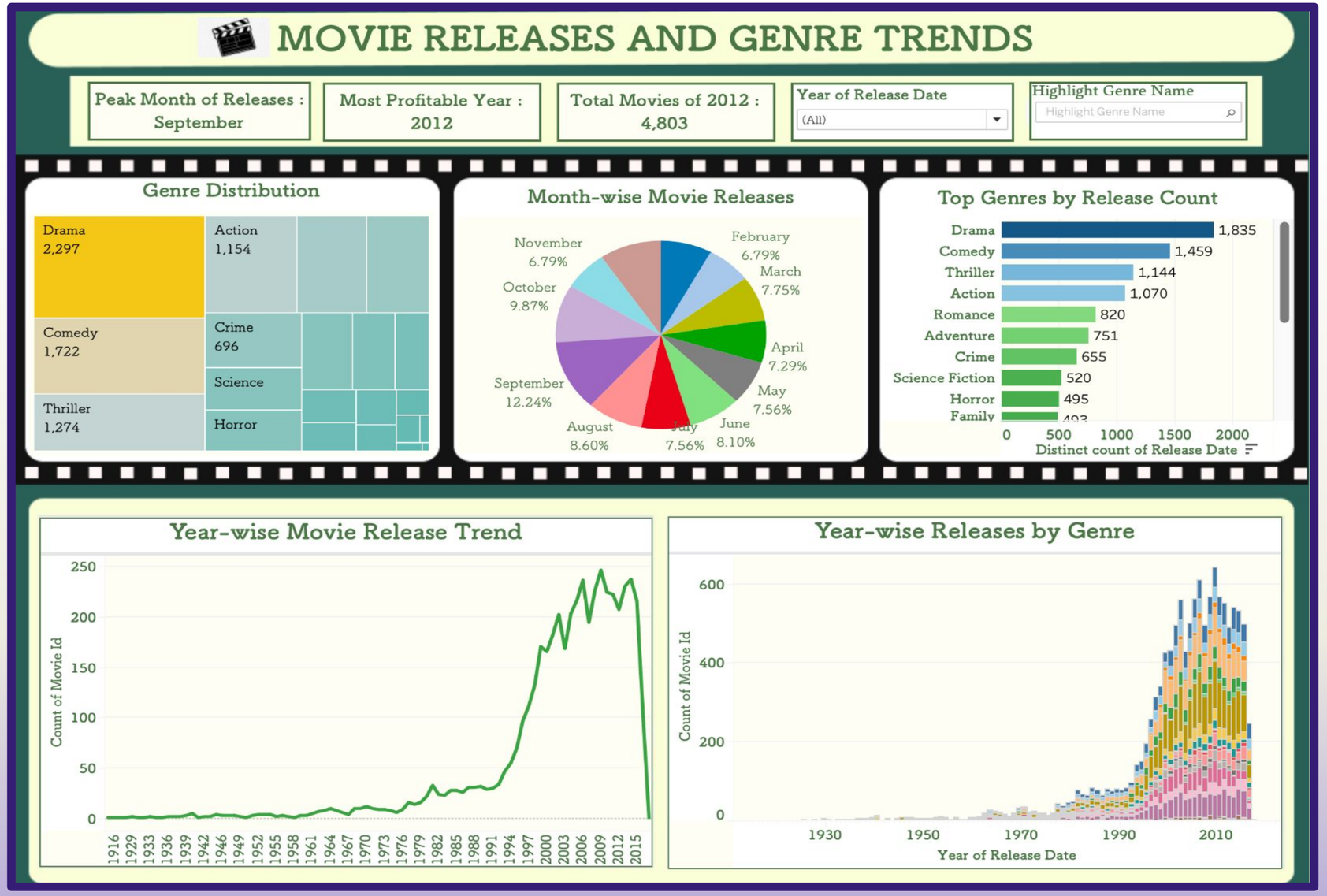
- 86% accuracy achieved in predicting movie outcomes.
- Flops predicted strongly (94% recall, 91% F1-score) – 685/730 correctly classified.
- Hits moderately captured (62% recall, 68% F1-score) – 143/231 identified.
- Model is biased toward flop detection, making it effective for risk filtering, but less reliable for hit prediction due to missing external factors (marketing, timing, competition).



04

DASHBOARDS

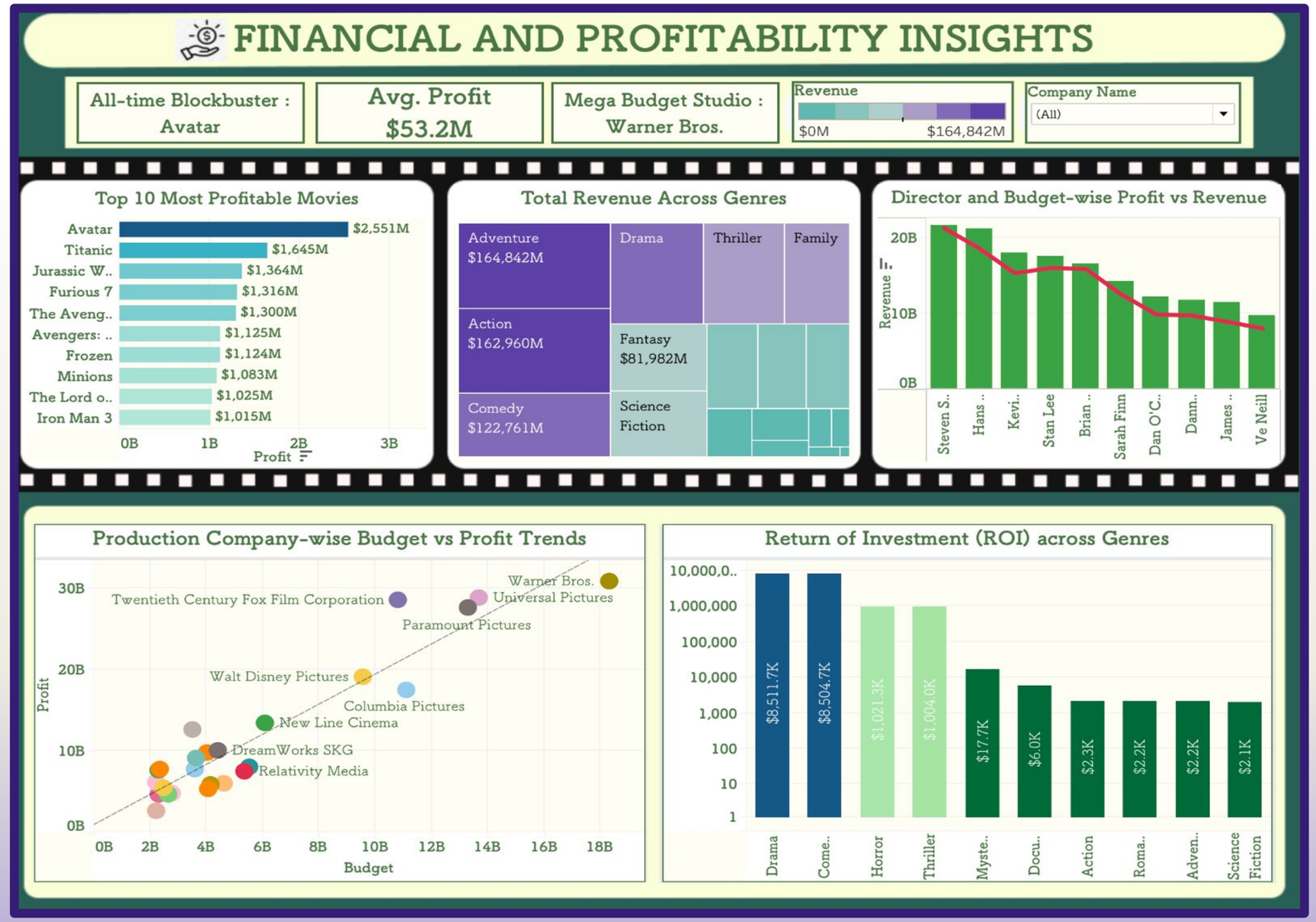
D1 : Release & Genre Trends



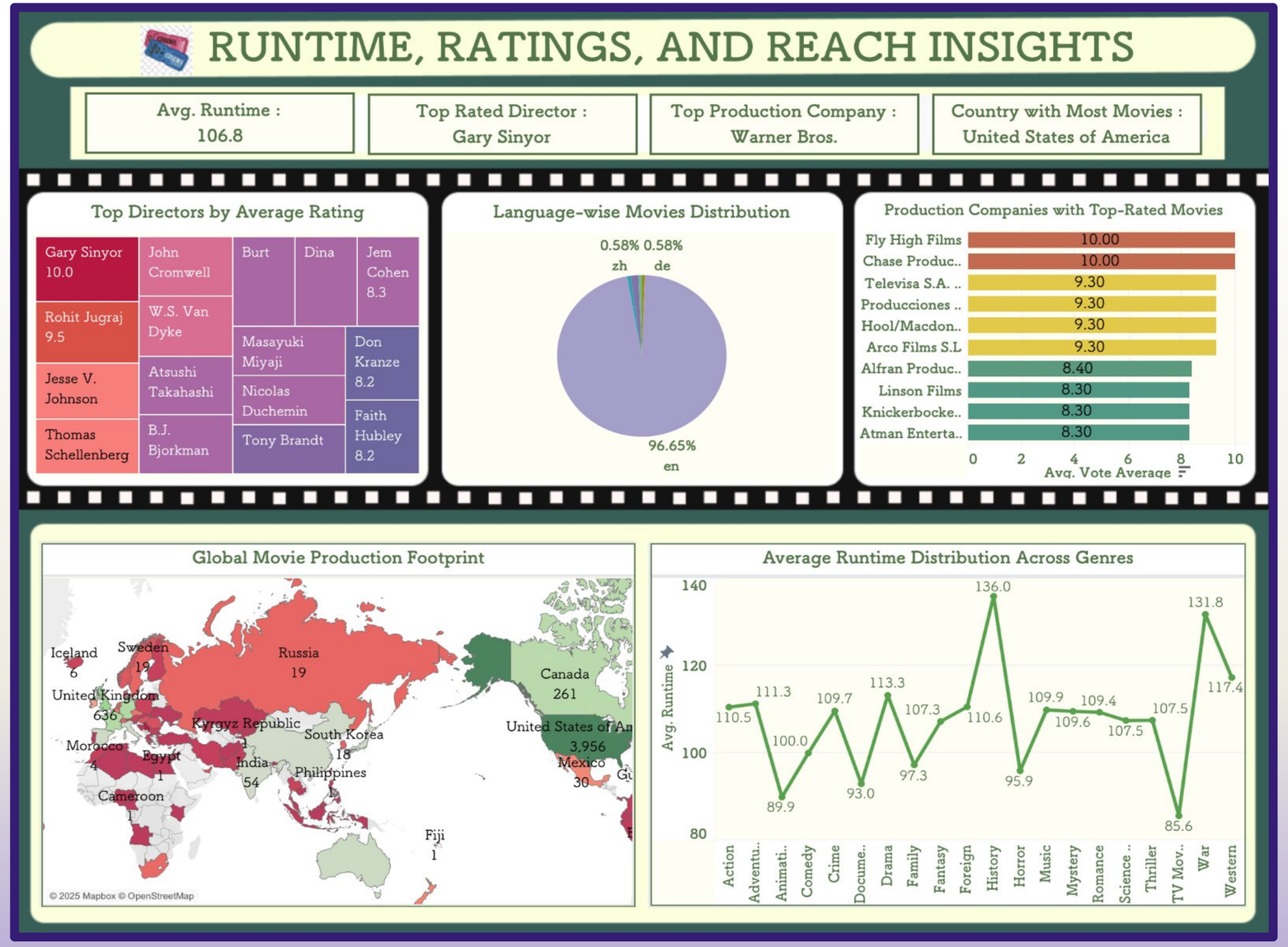
- Tracks yearly & monthly releases with genre breakdown
- Filters by year and genre for dynamic exploration
- Movie production shows exponential growth from minimal releases in 1930s to peak activity around 2010-2015.
- Drama leads with 2,297 releases (47.8% of total), followed by Comedy (1,722) and Thriller (1,274)

D2 : Financial & Profitability Insights

- Avatar leads as the most profitable movie ($2.55B profit) vs. industry average profit of $53.2M; Warner Bros. dominates mega-budget productions.
- Adventure ($164.8M) and Action ($163M) top total revenues, far ahead of Comedy ($122.7M) and other genres.
- Drama and Comedy show the highest ROI despite lower revenues; big studios like Warner Bros. and Universal see moderate ROI despite huge budgets.



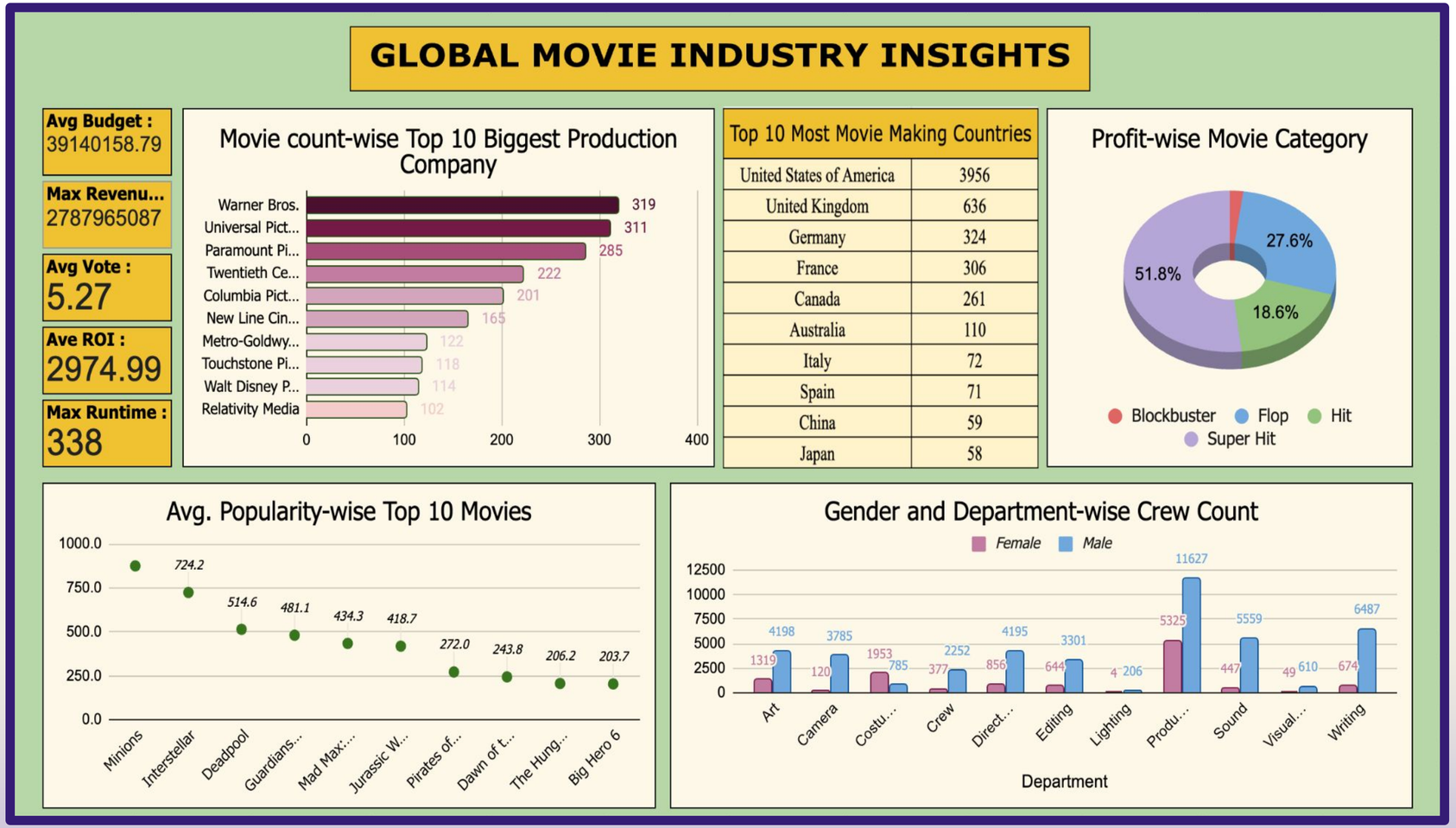
D3 : Audience & Geography



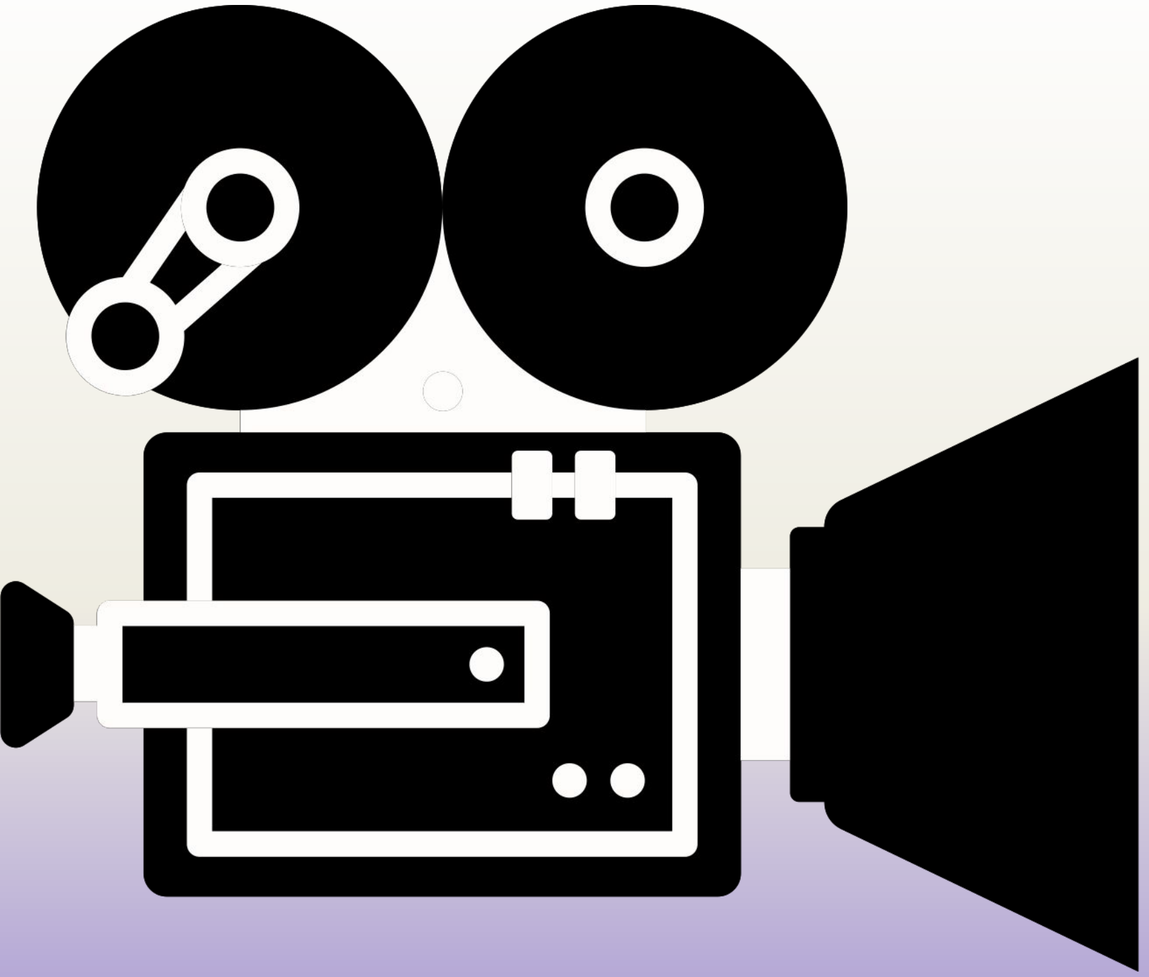
- Major studios dominate volume, but smaller studios (Fly High Films, Chase Productions) achieve perfect 10.0 ratings.
- USA leads with 3,956 movies; English dominates (96.65%); Gary Sinyor tops director ratings (10.0); average runtime ~106.8 min.
- History films longest (136 min) vs. TV Movies shortest (85.6 min), highlighting genre-specific storytelling styles.

D4 : Diversity & Category

- Majority of titles fall into “Super Hit” and “Hit” categories, with fewer “Blockbusters” revealing a pyramid of profitability.
- Minions achieving highest popularity score of 875.6, suggests audience engagement varies significantly from pure box office metrics.
- Gender and department analysis shows male dominance in technical areas like production and sound, while female presence is stronger in creative roles like costume and makeup.



BUSINESS RECOMMENDATIONS



1. **Balanced Portfolio:** Allocate 60% of budgets to Action/Adventure, 25% to high-ROI Drama/Crime, and 15% to emerging markets for diversified growth.

2. **Talent Optimization:** Use data-driven casting for big productions and nurture new talent in lower-risk Drama/Crime projects.

3. **Release Timing:** Target September (12.2% peak share) and apply predictive models (75% accuracy) to fine-tune budgets and marketing spend.

4. **Operational Excellence:** Close gender gaps in technical roles and adopt boutique studio practices to achieve higher quality ratings.

Conclusion & Way Forward

- Budget investment, genre selection, and strategic talent acquisition emerge as the strongest drivers of movie success.
- Statistical models achieved 75% accuracy in revenue forecasting, enabling better ROI planning and risk assessment.
- Interactive dashboards transformed raw data into actionable intelligence for production, marketing, and talent decisions.
- Current analysis excludes streaming metrics, marketing spend, and post-pandemic shifts, highlighting scope for refinement.
- Future Roadmap: Integrate real-time global data, audience sentiment, and regional patterns to enhance predictive models and guide next-gen entertainment strategies.

THANK YOU!