What makes a movie financially successful ?

Group WAT
Ambroise Aigueperse, Thomas Zamblera, Wesley Nana Davies

Motivation

- 🎬 An industry valued at **103 billion USD** in 2023

- 💸 Even if a movie can involve hundreds of millions in investment, its financial **success** remains **highly uncertain**

- 🎯 Movie's success influenced by :

  - **Technical** factors (eg : genre)
  - **Social** aspects (eg : audience sentiment)
  - **Economic** elements (eg: budget)

=> Good example of a **techno-socio-economic system** where it is interesting to apply rigorous DS methods to **uncover insights** beyond intuition

Dataset Characteristics

1) TMDB dataset:

45000 movies with the following infos:

- General **metadata** (title, budget, genre, runtime, revenue, production country)
- **Credits** with information about the cast (name, gender, role)
- **Sequels**
- Overview (**plot**) of the movie

2) Rotten Tomatoes (RT) dataset:

17000 movies with the following infos:

- General **metadata** (similar to TMDB dataset)
- **Reviews** with both a **grade**, a **written critic** and a boolean indicating if the critic is made from a professional

Dataset Preprocessing and Limitations

Basic preprocessing :

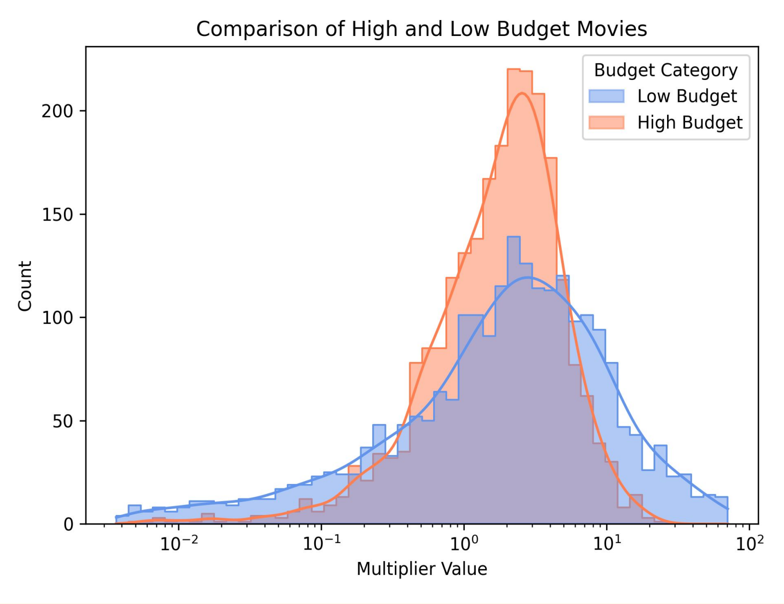
- **Clean** (remove missing values, duplicates and too extreme values like 0 budget)
- **Standardize** (normalize the grading scales from 0-10, convert budgets in USD)
- **Join datasets** (keep RT movies that are subset of TMDB movies)
- **Feature engineering** (details later)

Dataset limitations:

- **Sampling bias** (indie/popular, Hollywood/rest)
- **Temporal bias** (older movies judged differently)
- **Lack of data** (some movies have too few reviews)

Budget and Revenue characterization :

- **Low budget** is first 40%, **high budget** is top 40%
- **Multiplier** = revenue/budget



Plan of Action

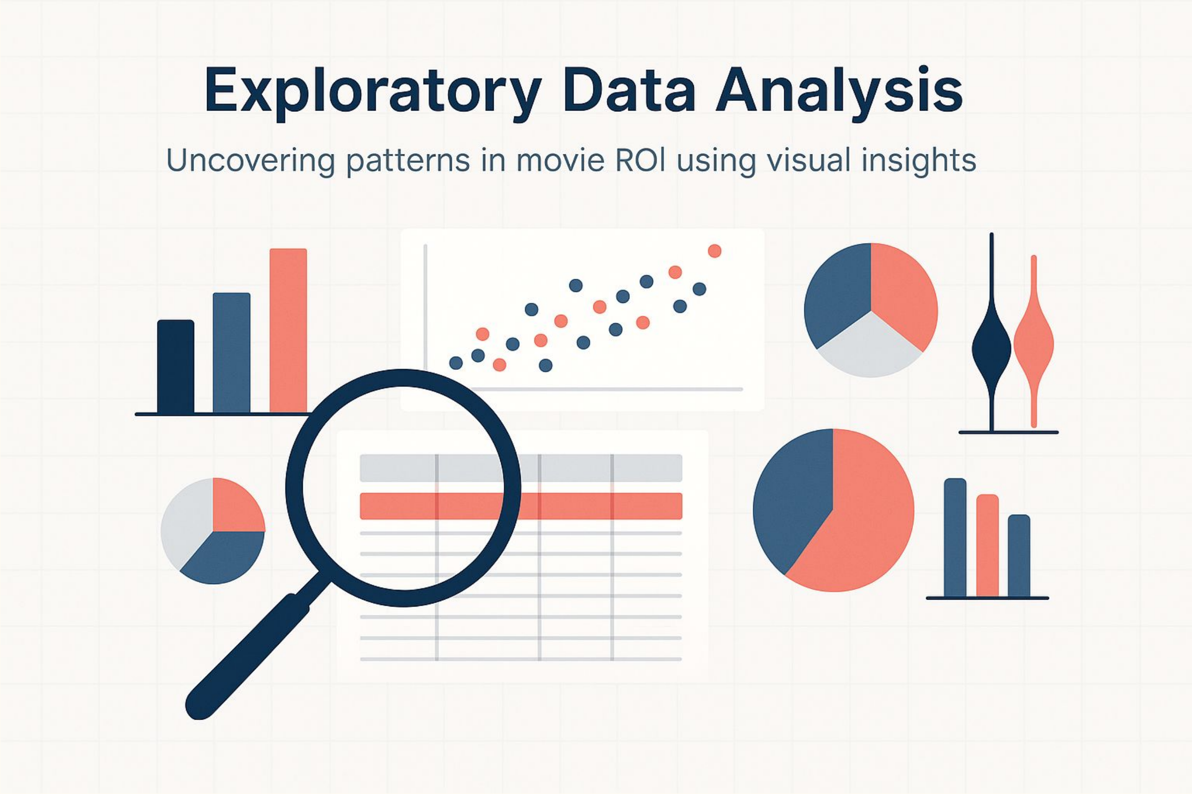
1) 📊 Basic Data Analysis

2) ⭐ Popularity & Financial Success: Are They Linked?

3) 📝 Mining Meaning — What Text Tells Us About Performance?

4) 🤖 Feature Importance & Prediction — Can Machine Learning Techniques Give Us Insights?

Part 1 : Data analysis: What basic analysis can tell us?

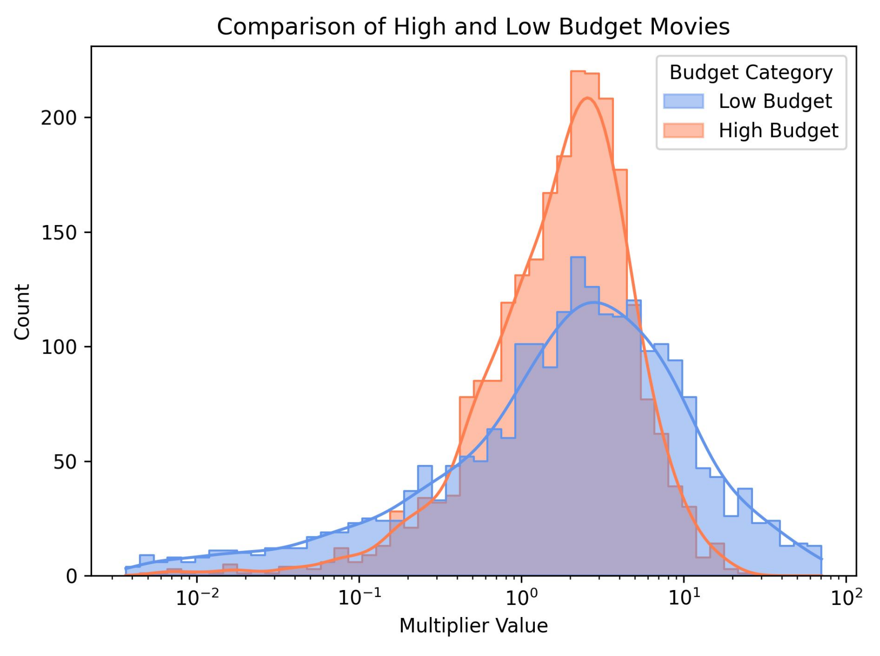


1.1 : Distribution of movies along Multiplier

	Average Multiplier	Variance Multiplier
low	5.77	68.59
high	2.61	5.76

<u>High budget movies:</u> smaller variance, better idea of what works and does not

<u>Low budget movies</u>: more **flippy**, can get really high but also really low multiplier



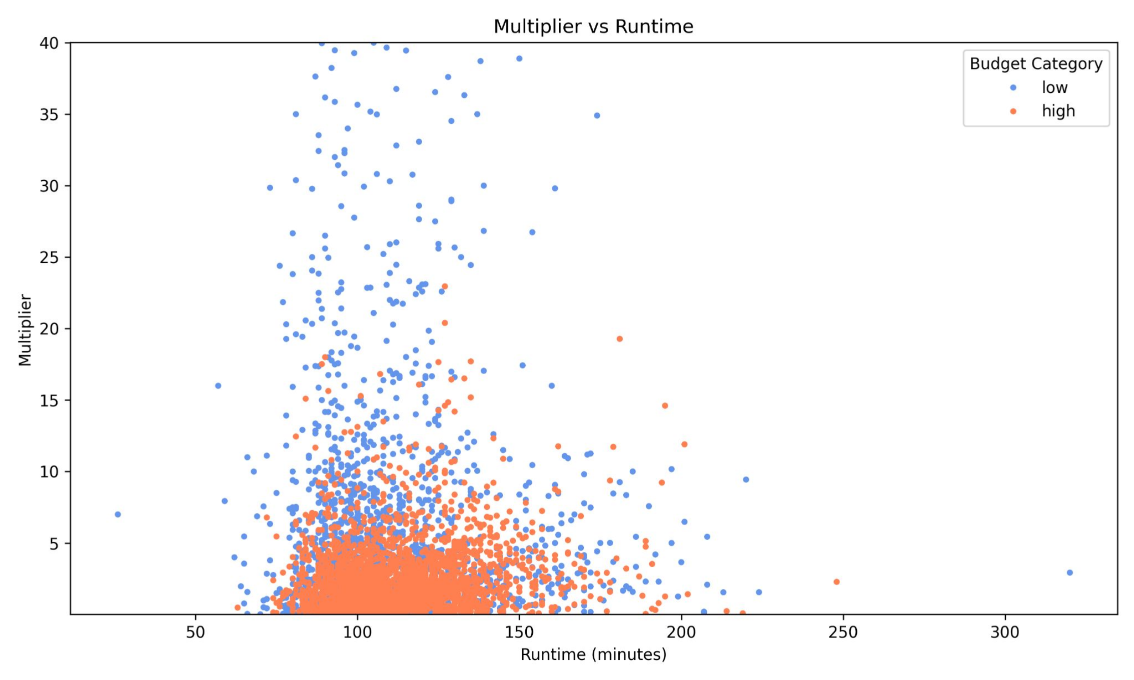
P-value: 3.97e-42

1.2 : Runtime of movies

<u>**H0:**</u> Movies' runtime has no direct impact on Multiplier

Most movies have the same runtime (few outliers)

No particular **evidence** of direct influence



Corr. coef: -0.02
P-value: 0.18

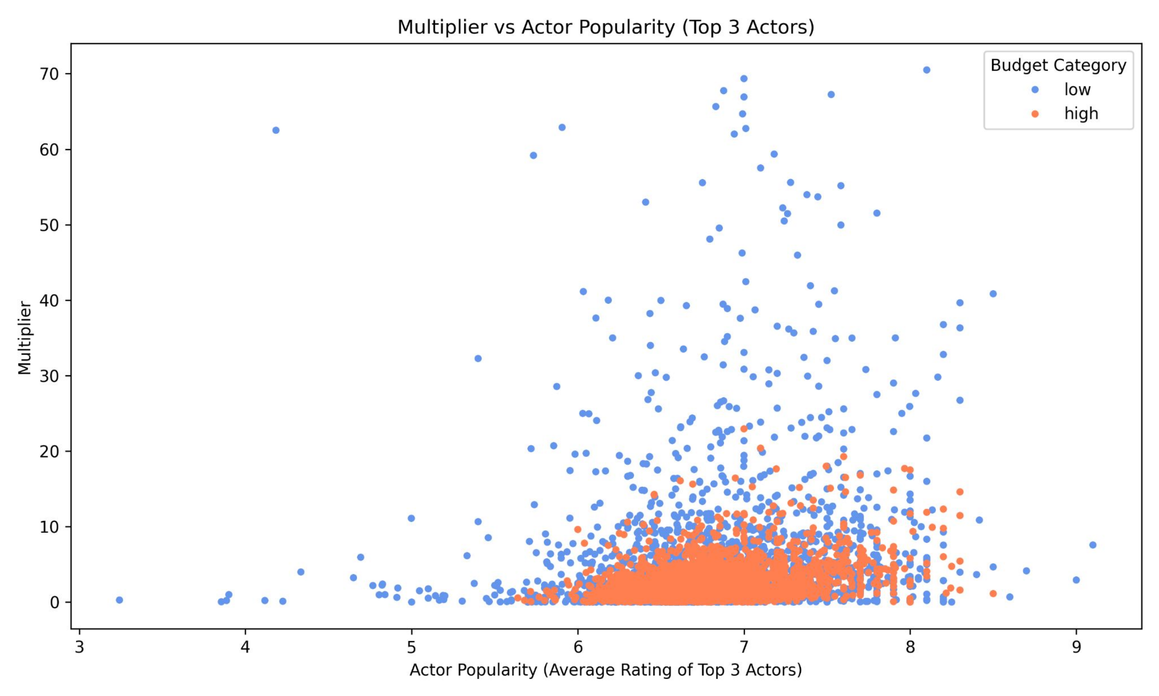
=> **no direct impact**

1.3 : Influence of actors on a movie's success

For each movie, average of ratings of top 3 actors' ratings (among the films they took part in)

Small trend showing that really successful movies tend to have great actors

Impact on audience targeted



Corr. coef: 0.17
P-value: 6.54e-13

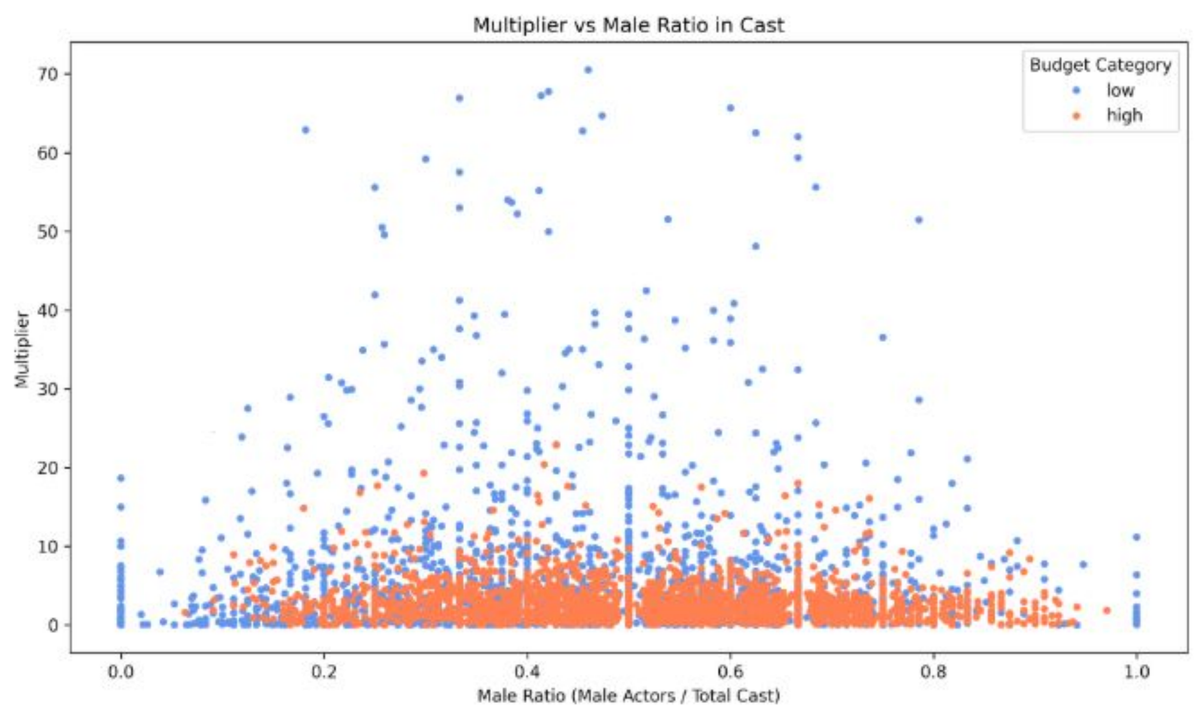
=> **trend confirmed**

1.4 : The effect of gender diversity

Is having a well balanced cast beneficial?

Kind of pyramidal shape, so test for quadratic relationship

Movie cast more based on **creative aspects** than pure financial considerations

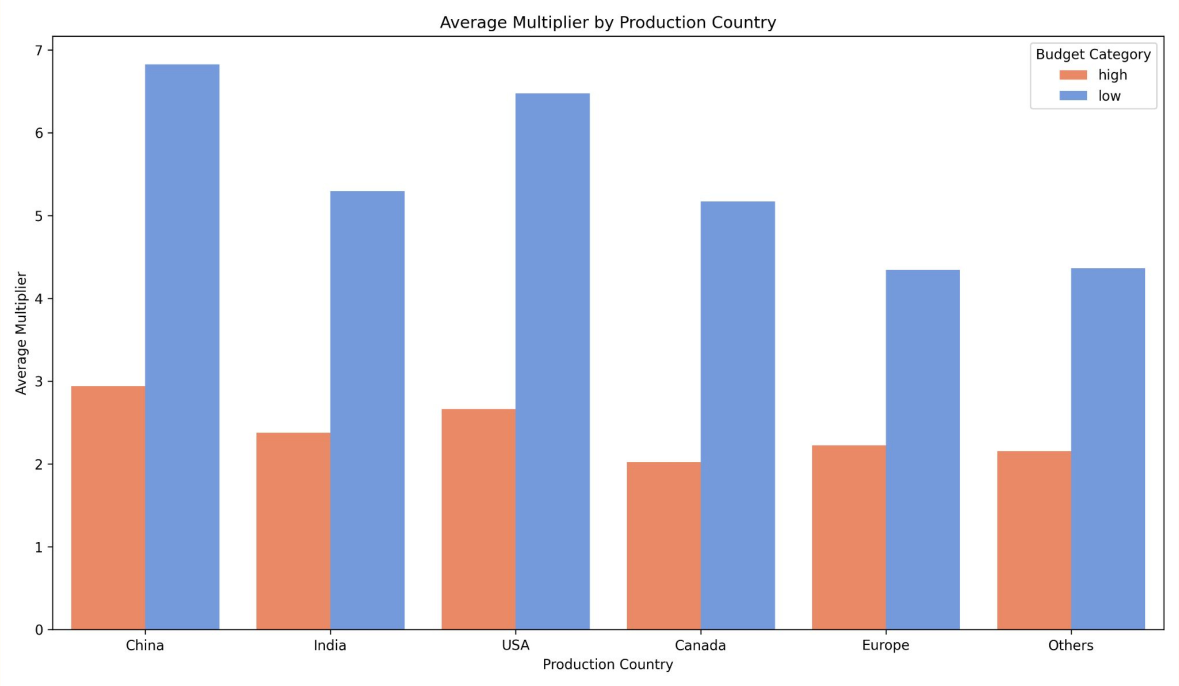


P-value: 0.68 => no basic relationship

1.5 : Does the production country affect the multiplier?

Looks like **bigger population => better Multiplier**

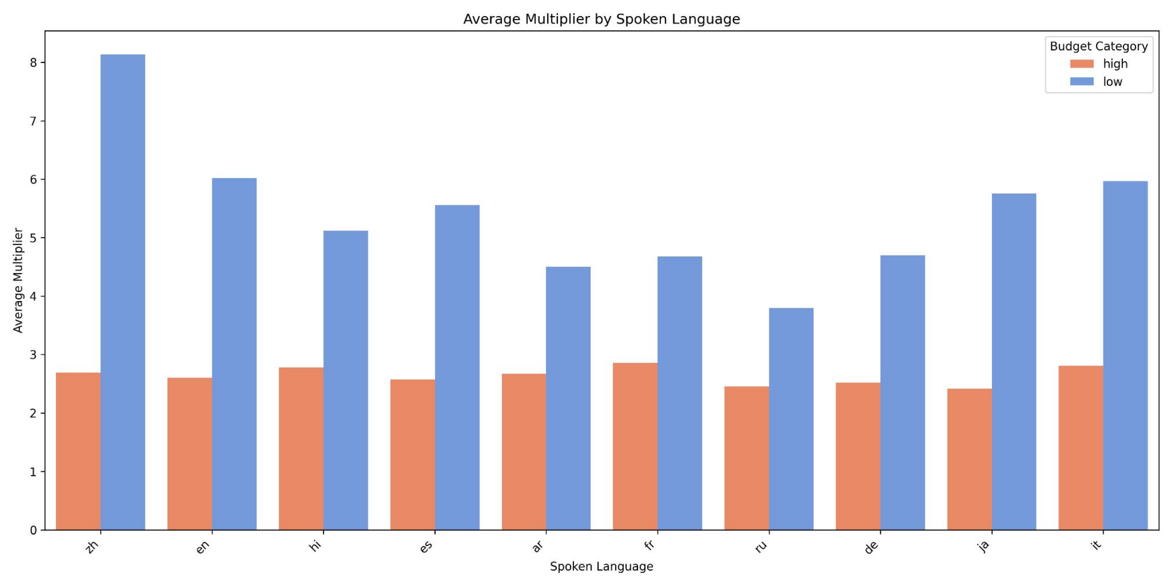
Control hypothesis with **spoken language** in the movie, should see **similar behavior**



1.6 : What about the language of the movie?

Follows the **same scheme**

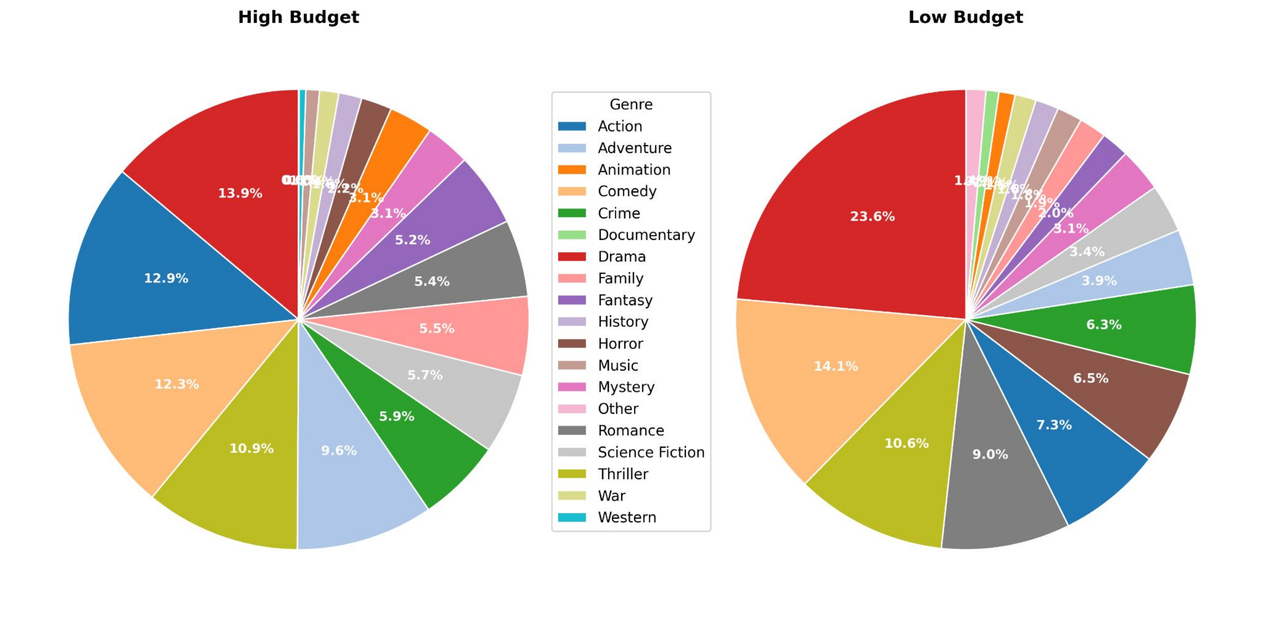
Intuitive because **bigger audience**



1.7 : Are selected genres to be preferred?

<u>High budget:</u> the most dominant categories (in number) are Drama, Action and Comedy

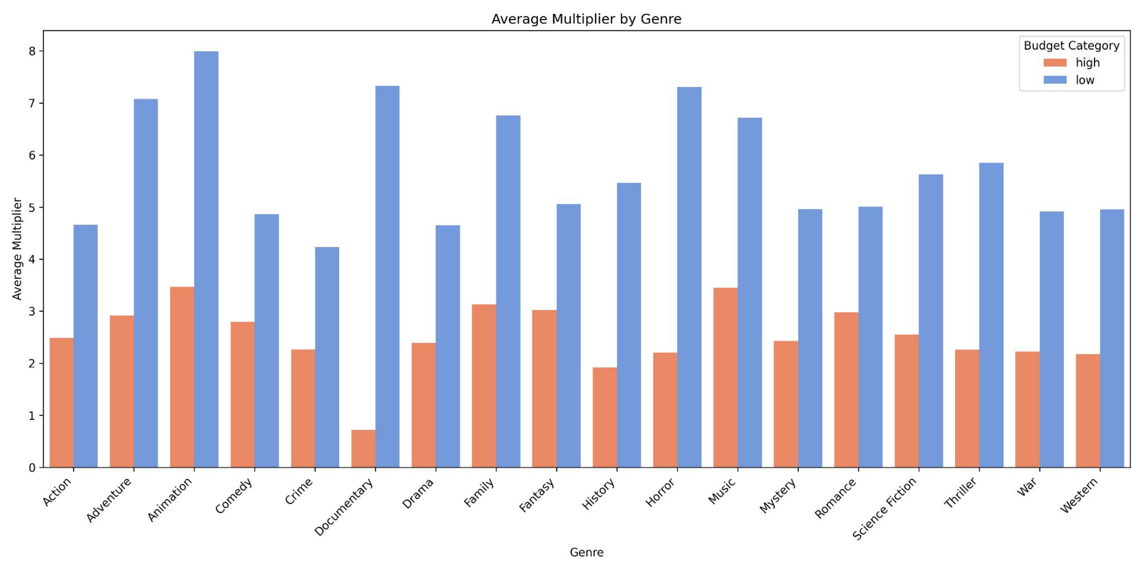
<u>Low budget:</u> Drama, Comedy but Action took a big drop



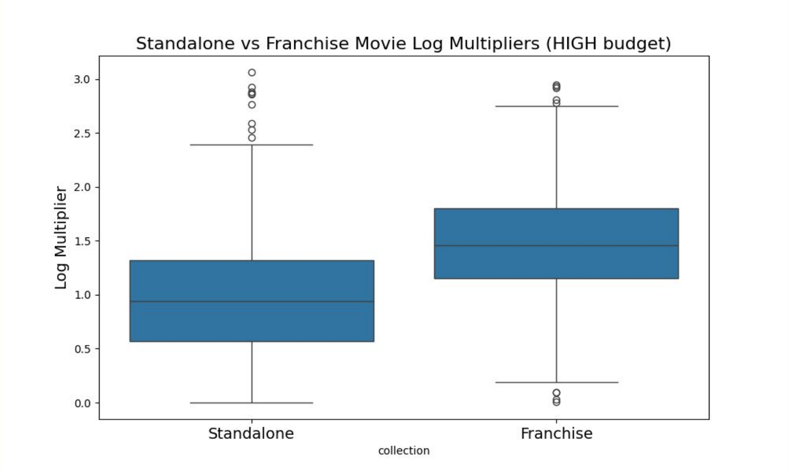
1.7 : Are selected genres to be preferred?

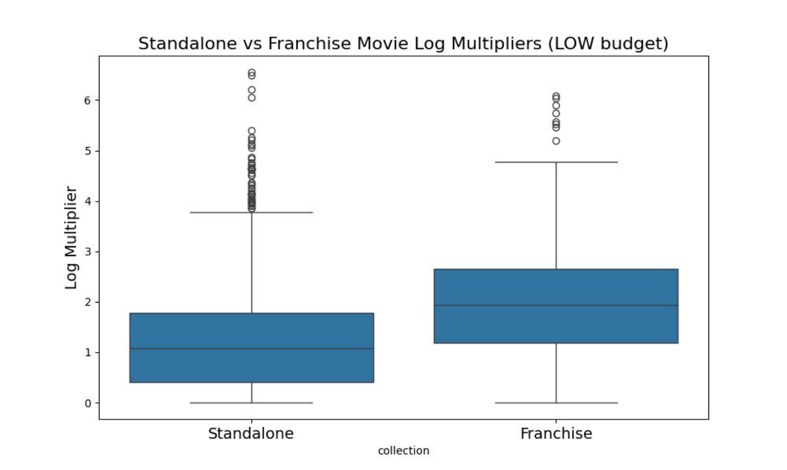
=> most Action movies require a higher budget
And **higher budget** movies => more **conservative multipliers**

Contrarily, documentaries are **cheap** & target big **audience** => **higher multipliers**



Bonus : Are franchise movies financially worth it?





H0: There is no difference in mean multiplier between franchise and standalone movies

H1: There is a difference in mean multiplier between the 2 groups

Budget Type	Standalone Mean Multiplier	Franchise Mean Multiplier	p-value	Conclusion
High Budget	2.06	3.98	3.46e-67	✅ Franchise movies perform better
Low Budget	7.18	18.43	3.91e-09	✅ Franchise movies perform better

Bonus : Are franchise movies financially worth it?

-CAREFUL : Just comparing standalone vs franchise is biased ❗

👉 *Studios may only create a sequel if the first movie was financially successful*

🔄 So instead, we analyze:

Correlation between a movie's financial performance and the next one in the franchise

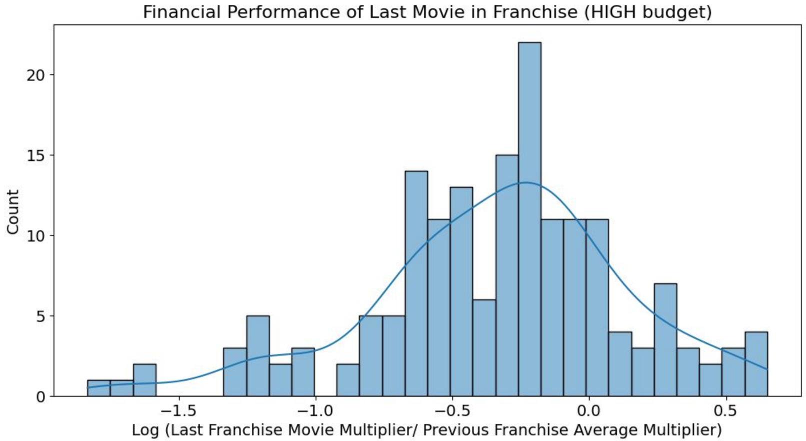
Budget Type	Correlation Coefficient	Conclusion
High Budget	0.52	🔗 Strong correlation → Prior success predicts sequel success
Low Budget	0.30	🔗 Moderate correlation

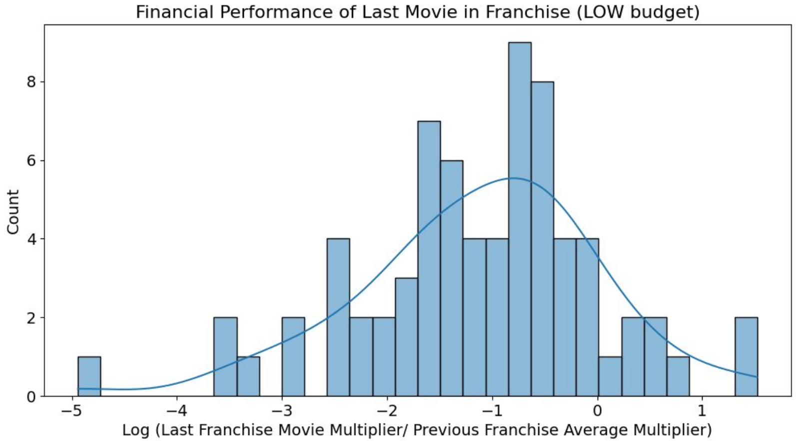
Why lower correlation for low-budget ?

- less marketing => less brand loyalty
- low budget sequels vary more in cast quality, distribution scale, audience expectation

Bonus : Are franchise movies financially worth it?

- If prior success predicts sequel success, do sequels keep paying off...forever??🎬💸



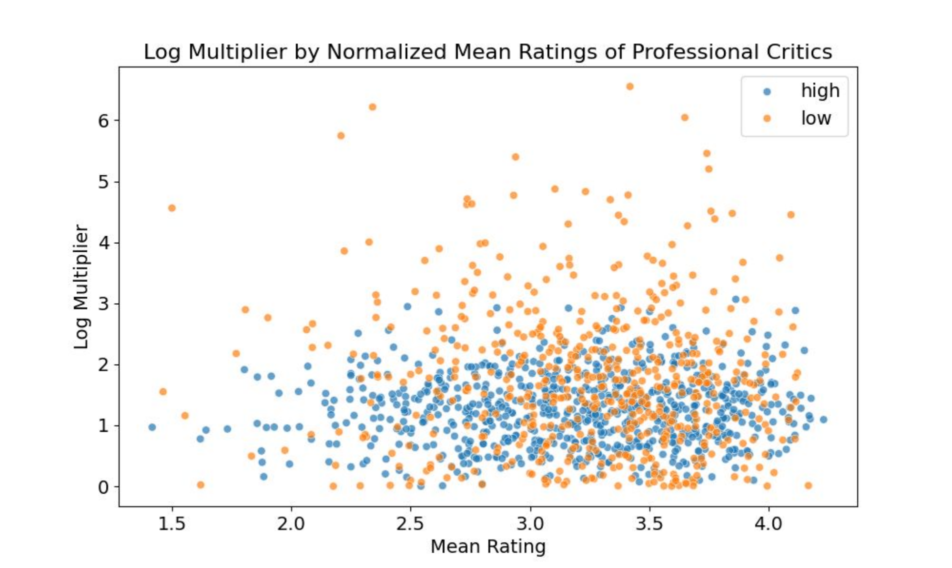


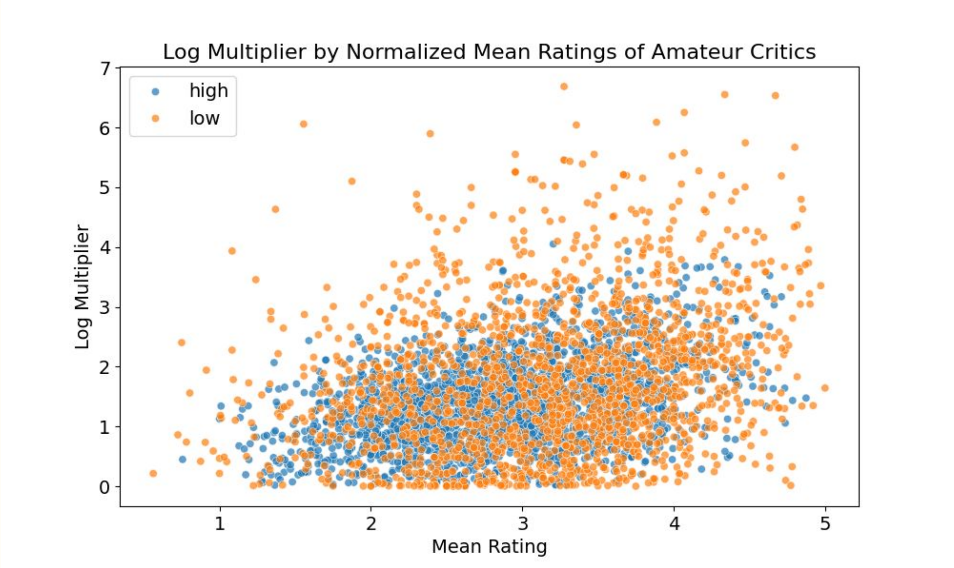
=> In both high and low budget cases, the **last movie** in a franchise tends to **underperform** compared to earlier entries

=> Studios seem to **stretch** the franchise **until the return drops**

Part 2 : Popularity & Financial Success: Are They Linked?

2.1 : Professional vs Amateur critics =>what matters the most ?





- **No clear trend** : high critic ratings don't strongly relate to financial success
- **Higher ratings** tend to match with **higher multiplier : public opinion** matters

2.1 : Professional vs Amateur critics =>what matters the most ?

<u>H0</u>: No linear correlation between mean rating and multiplier

<u>H1</u>: Linear correlation between mean rating and multiplier

<u>Professional Ratings</u>:

Budget	Correlation	p-value	Interpretation
High Budget	0.04	0.2779	Very weak, not significant.
Low Budget	-0.05	0.2371	Also weak and not significant.

<u>Amateur Ratings</u>:

Budget	Correlation	p-value	Interpretation
High Budget	0.32	< 0.0001	Significant positive correlation.
Low Budget	0.12	< 0.0001	Weak but statistically significant correlation.

2.1 : Professional vs Amateur critics =>what matters the most ?

<u>Conclusion</u> : Amateur ratings matter more!

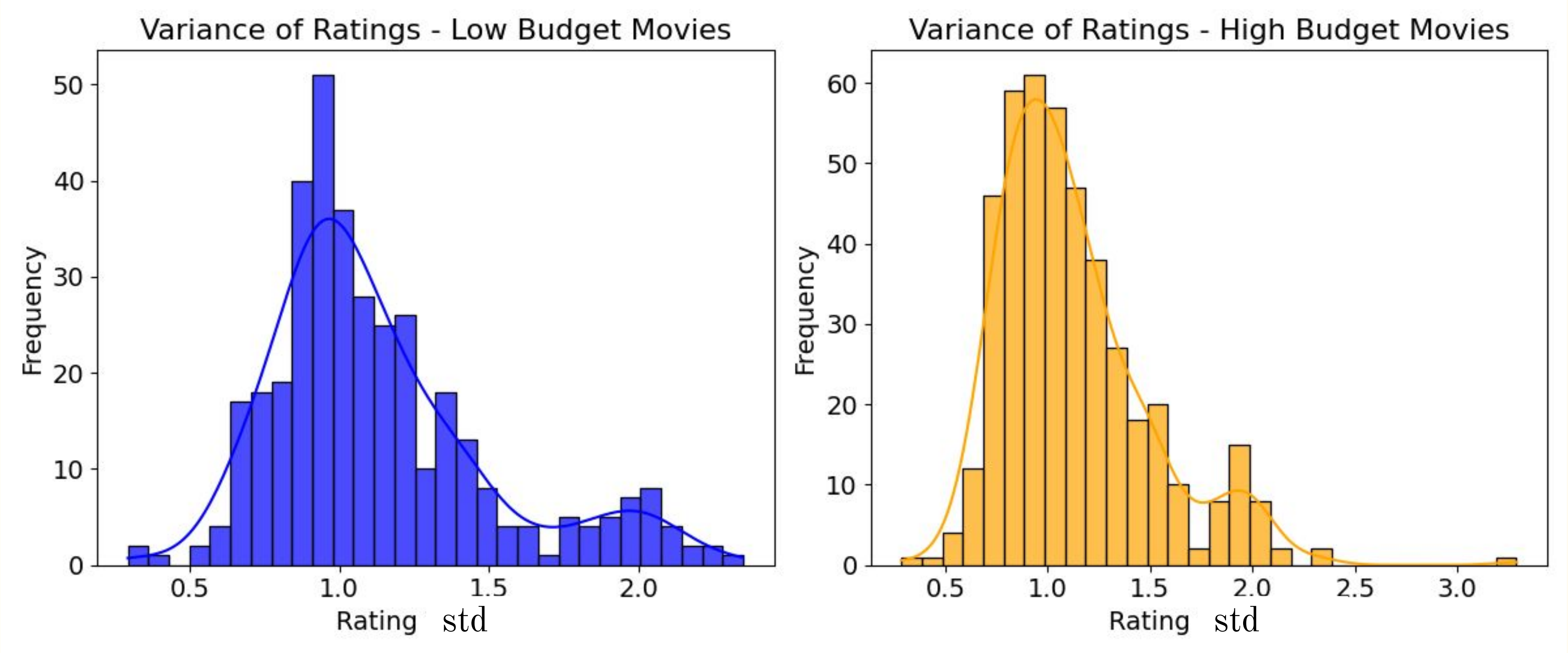
<u>Explanatory Hypothesis</u> :

- **Mass appeal** matters

- **Critics $\neq$ consumers**

- **Word of mouth/virality**

2.2 : What about the actual rating distributions of financially successful movies? (focused on amateurs)



=> in both cases there seems to be 2 clusters. Can we investigate them ?

2.2 : What about the actual rating distributions of financially successful movies?

Use **K-Means clustering** with 2 clusters and engineered features on distribution shape

<u>Low Budget Movies:</u>

Cluster	Mean Rating	Std Rating	Harsh Reviews	Very Good Reviews	Skewness	Kurtosis	Avg Cluster Multiplier	Std Cluster Multiplier
0 (Clivant)	3.17	1.97	0.258	0.402	-0.09	3.53	**12.84**	**18.13**
1 (Safe)	3.58	0.96	0.102	0.516	-0.73	0.85	**22.38**	**7.16**

<u>High Budget Movies:</u>

Cluster	Mean Rating	Std Rating	Harsh Reviews	Very Good Reviews	Skewness	Kurtosis	Avg Cluster Multiplier	Std Cluster Multiplier
0 (Clivant)	3.25	1.86	0.201	0.434	-0.04	2.96	**4.32**	**3.93**
1 (Safe)	3.57	0.96	0.107	0.511	-0.76	0.85	**5.67**	**2.47**

2.2 : What about the actual rating distributions of financially successful movies?

Conclusion :

- Successful movies fall into two camps: **“safe crowd-pleasers”** and **“polarizing wildcards”**
- **Polarizing films** carry **more risk**, but offer **greater upside** potential
- High budget movies, more stable, make these 2 groups less distinct

Explanatory Hypothesis :

- Polarizing content creates buzz => riskier
- High-budget movies prioritize consistency

2.3 : What about the temporal distribution of reviews ?

H0: Early buzz does not matter

H1: Good early ratings do impact financial success

High Budget Movies:

Rating Bin	Multiplier for High Initial Mean Rating	Multiplier for Low Initial Mean Rating	p-value
2–2.5	3.01	3.11	0.7986
2.5–3	3.36	2.56	0.0653
3–3.5	3.78	3.22	0.0564
3.5–4	3.60	3.18	0.1196
4–4.5	3.99	2.55	0.0491

Conclusion : Early buzz seems to matter in general for high budget movies

Possible explanation :

- Opening weekend performance for big movies
- Early ratings create momentum

2.3 : What about the temporal distribution of reviews ?

H0: Early buzz does not matter

H1: Good early ratings do impact financial success

Low Budget Movies:

Rating Bin	Multiplier for High Initial Mean Rating	Multiplier for Low Initial Mean Rating	p-value
2–2.5	12.43	9.47	0.7506
2.5–3	7.86	9.21	0.3262
3–3.5	12.00	5.28	0.0942
3.5–4	8.55	21.65	0.4846
4–4.5	8.81	3.20	0.6490

Conclusion : Early ratings do not significantly affect financial success for low budget

Possible explanation :

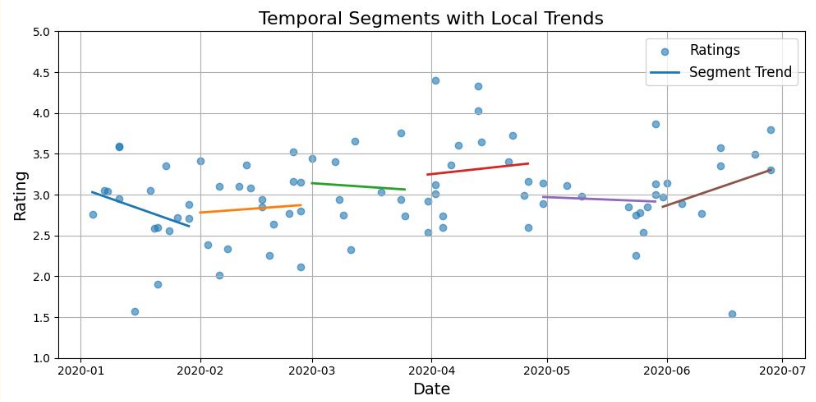
- longer discovery cycle for indie movies (after release via word of mouth)
- don’t rely on big opening weekends (less cinema exposure)
- niche slowly attracts a dedicated fan base that rediscovers older movies

2.3 : What about the temporal distribution of reviews now ?

=> Measure the impact of ratings' fluctuation

<u>Method</u> :

- Filter the movies with enough data
- For each movie split reviews into 30 days segments
- Fit a ridge regression for each of these segments
- Compute the std of the local slope coefficients



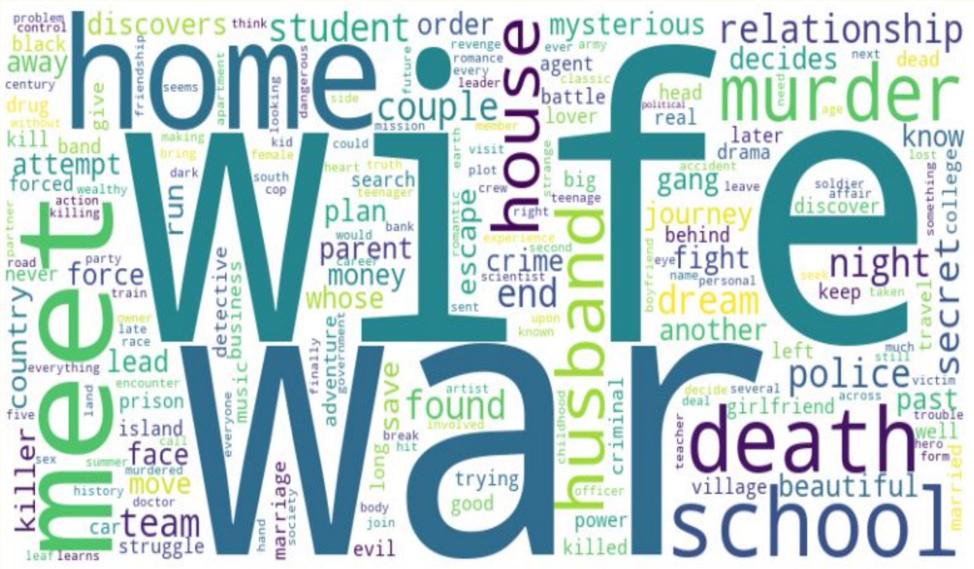
<u>H0</u>: There is no difference in return if you have high or low ratings' fluctuation

<u>H1</u>: There is a meaningful difference in returns between the 2 groups

Budget Type	High Fluctuation Mean Multiplier	Low Fluctuation Mean Multiplier	p-value	Conclusion
High Budget	3.18	3.33	0.646	✕ No statistical difference
Low Budget	22.09	45.43	0.319	✕ No statistical difference

Part 3 : Mining Meaning — What Text Tells Us About Performance?

3.1 What topics extracted from movies' summaries tell us ?



Cloud of words with the highest TF-IDF scores over the entire corpus

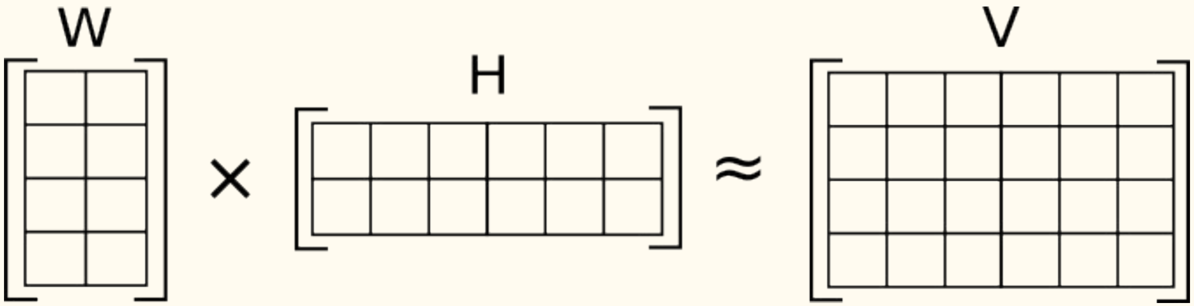
TF-IDF scores measure the importance of a word in a document

TF-IDF(t,d)=TF(t,d) x IDF(t)

- TF(t, d) being term frequency of t in document d
- IDF(t) being the inverse document frequency, reducing importance of word appearing across all documents

3.1 What topics extracted from movies' summaries tell us ?

Method: Non-negative matrix factorization (NMF)



This algorithm finds two matrices W and H which best approximates the TF-IDF scores matrix V the best in the Frobenius norm sense.

Shapes of matrices:

- W → num_movies x k
- H → k x voc_size
- V → num_movies x voc_size

k is the hyperparameter to choose which defines the number of topics / dimensionality of the embeddings

Makes it possible to identify topics and find the most relevant associated words and movies

3.1 What topics extracted from movies' summaries tell us?

Results

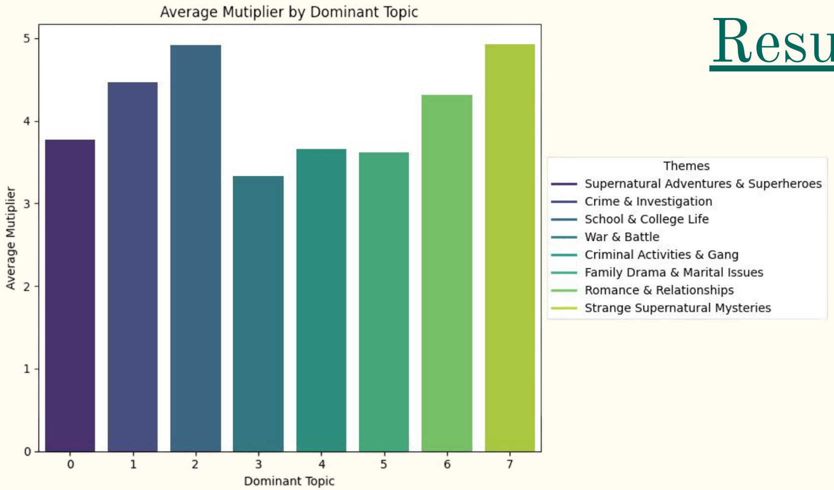
After some tuning and manual labeling, we opted for k=8:

- **Topic 1: "Supernatural Adventures & Superheroes"**
  Top Words: evil, power, save, fight, face, plan, team, ...
  Top Movies: Men in Black III, Mercury Man, Interceptor Force 2, ...
- **Topic 2: "Crime and Investigation"**
  Top Words: murdered, guilty, investigate, suspect, investigation, clue, body, ...
  Top Movies: But who killed Pamela Rose?, The Key to Reserva, Werewolf in a Women's Prison, ...
- **Topic 3: "School & College Life"**
  Top Words: school, student, college, class, popular, university, teenager, ...
  Top Movies: Fear No Evil, Election, Monster High: Fright On!, The Flying Classroom, ...
- **Topic 4: "War & Battle"**
  Top Words: war, soldier, civil, army, battle, military, prisoner, ...
  Top Movies: Escalation, The Biggest Battle, Enemy at the Gates, ...
- **Topic 5: "Criminal Activities & Gang"**
- **Topic 6: "Family Drama & Marital issues"**
- **Topic 7: "Romance & Relationship"**
- **Topic 8: "Strange Supernatural Mysteries"**

3.1 What topics extracted from movies' summaries tell us ?

Null Hypothesis (H_0): There is no statistically significant difference in the revenue-to-budget ratio across different topics; that is, no specific topic provides a consistent advantage in generating a higher multiplier.

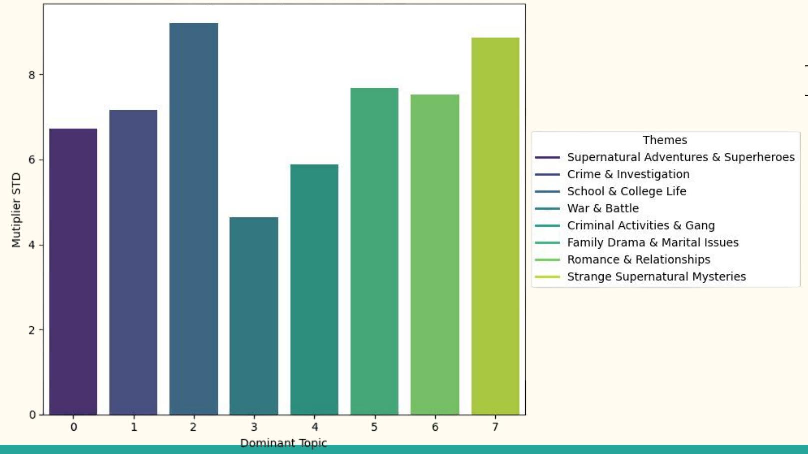
3.1 What topics extracted from movies' summaries tell us ?



Results

We can see most **profitable themes** are:

- Strange Supernatural Mysteries
- School & College Life
- Romance & Relationships
- Crime & Investigation



<u>ANOVA test</u>:

- p-value $= 0.0032 < 0.05$ **really unlikely** to observe such results under the null hypothesis and we **can reject it 'safely'**

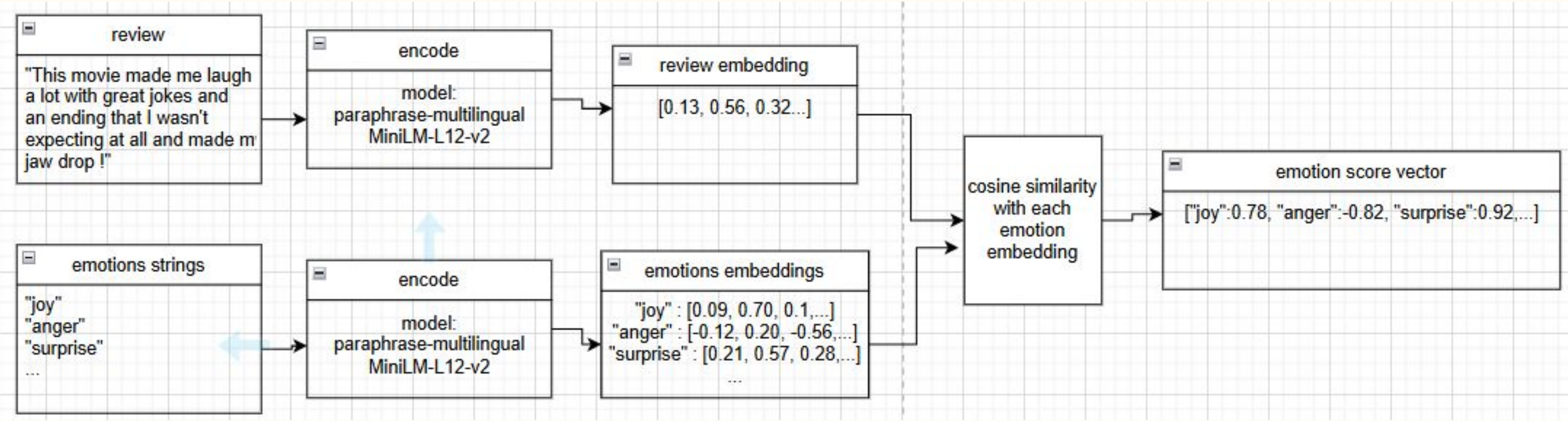
Multipliers for these movies have **higher variance: Relates to the idea of high-risk high reward /mean variance tradeoff**

<u>Explanation</u>: Such movies target a large audience and often require a lower budget compared to other movies (superheroes/war)

3.2 : Feelings & Finances — Do reviews' emotions explain movie success?

Method:

- Embedding model (paraphrase-multilingual-MiniLM-L12-v2) to **encode reviews and emotion** string
- Compute **cosine similarity** between review embeddings and emotion embeddings
- Represent **each movie with an emotion vector** averaged across all its reviews
- Compute **correlations**

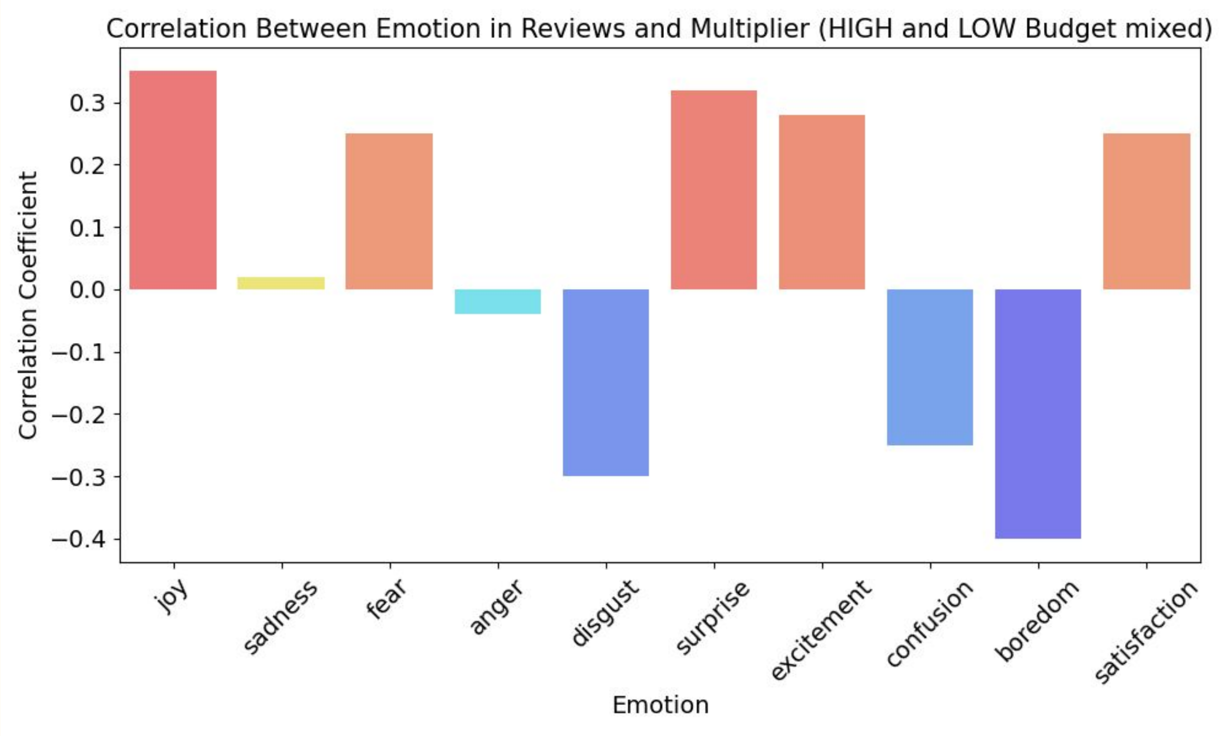


3.2 : Feelings & Finances — Do reviews' emotions explain movie success?

<u>Results:</u>

There seems to be 3 groups :

- positively correlated
- negatively correlated
- neutral/weakly correlated



Part 4 : Feature Importance & Prediction — Can Machine learning Techniques Give Us Insights?

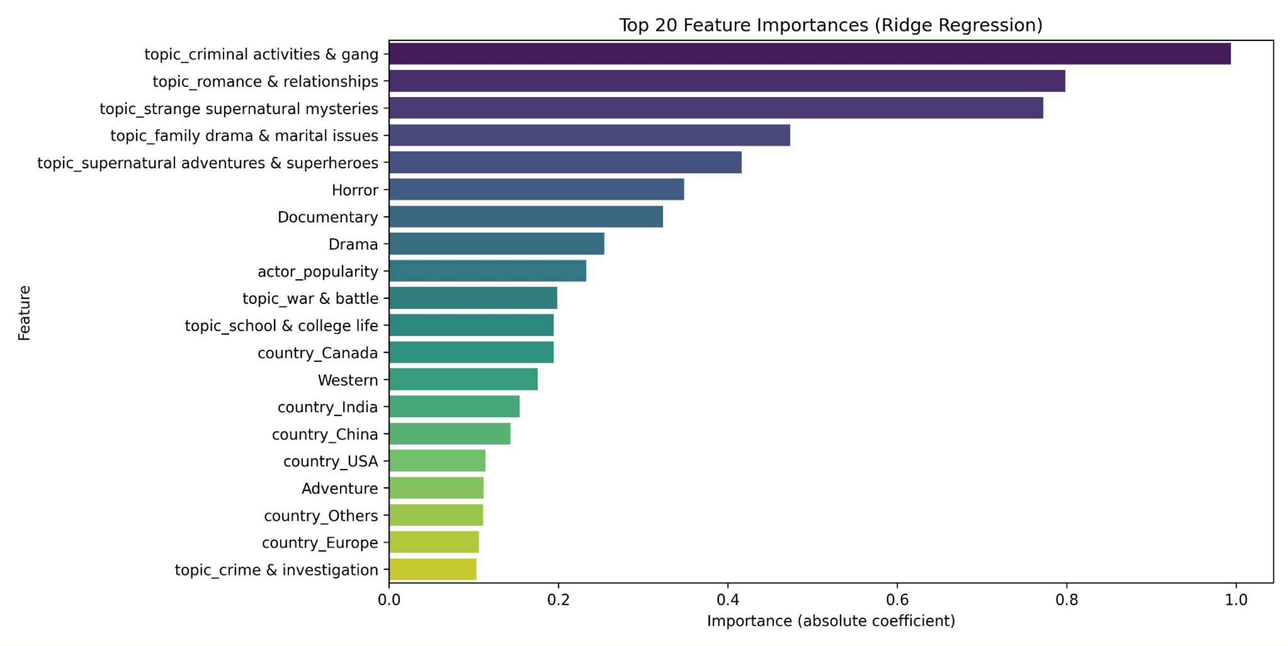
4.1 : Feature importance (linear model)

Dominance of **topic** features => audience appeal

Genres => known trends (e.g Horror low budget high revenue)

Actor popularity **notable**

But can **only** reveal **linear relationships**

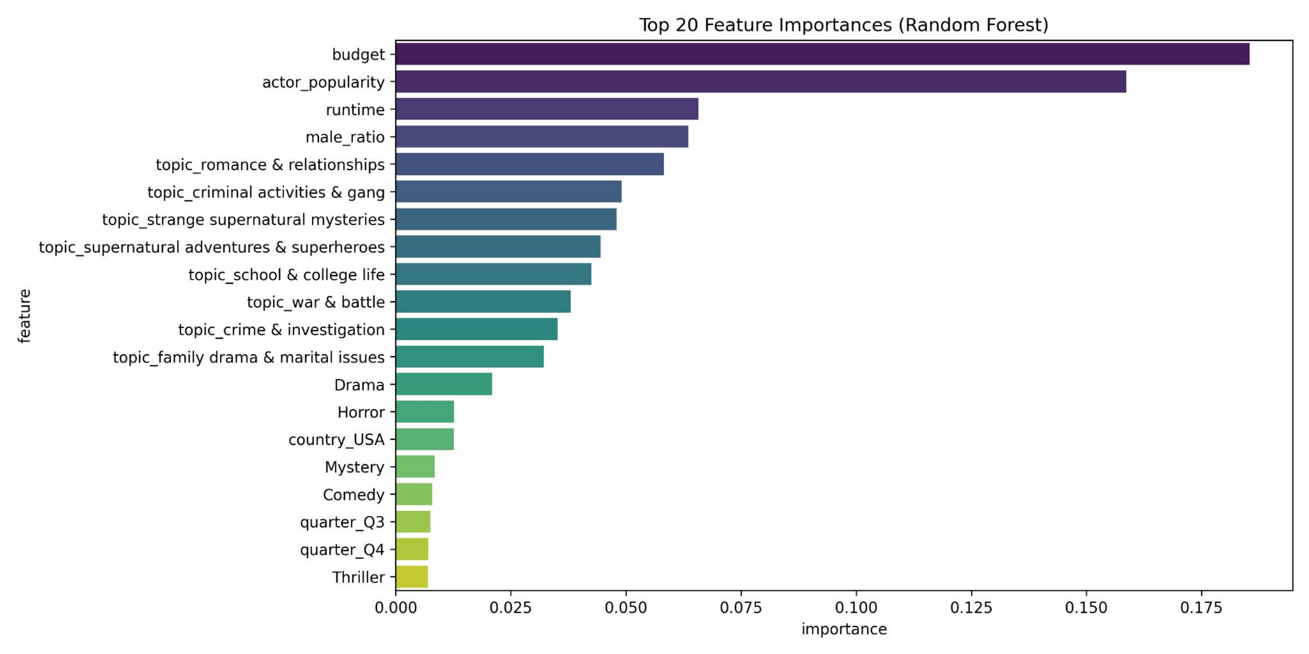


4.1 : Feature importance (non-linear model)

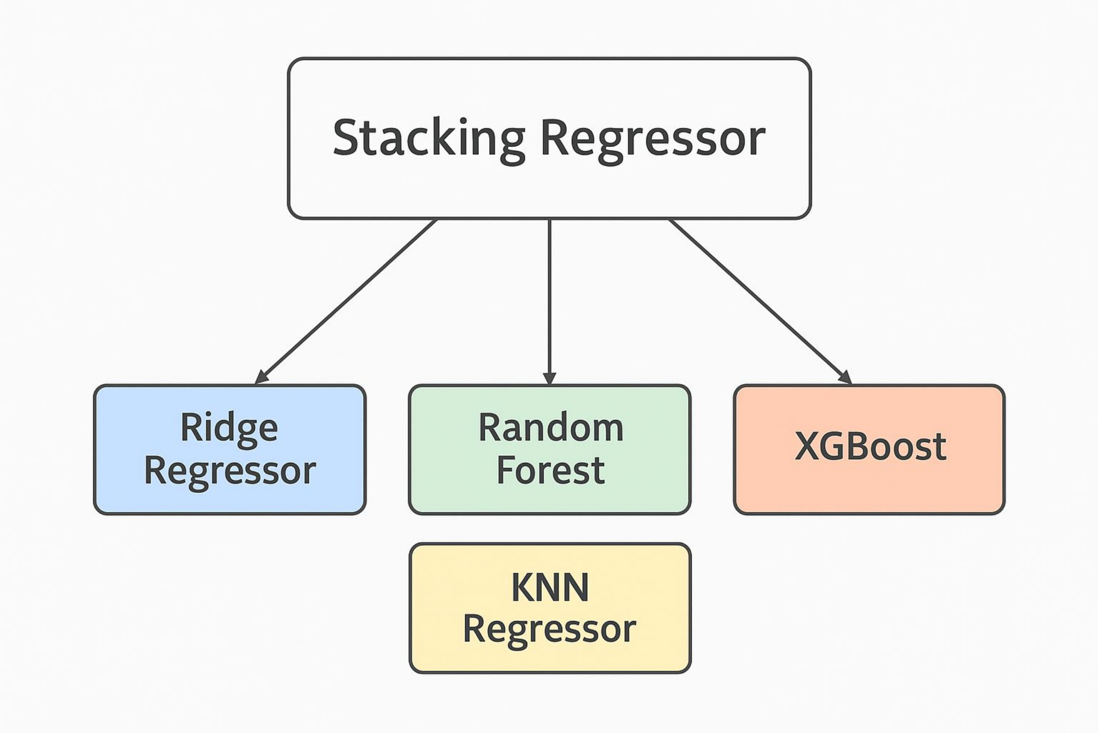
Can model **arbitrary** relations

Budget => Multiplier = Revenue/Budget

Actor popularity, runtime, male ratio => intuitive link



4.2 : Prediction model architecture — Ensemble methods



- Combine strength of diverse models
- Improved predictive accuracy
- Captures more complex relationships
- Reduces overfitting risks
- Interpretable weighting via meta-learner

4.2 : Can we predict a movie's financial success using ML? — Results

- ML models still captures some rather small signal (**$R^2 \approx$ 0.27** , "only" 27% variance explained)
- Models perform relatively **poorly individually**
- Ensemble **improves performance** modestly

Model	RMSE	MAE	R^2
Ensemble	0.753	0.564	0.265
XGBoost	0.776	0.580	0.251
Random Forest	0.779	0.588	0.244
KNN	0.857	0.638	0.086
Ridge	0.857	0.623	0.086

Majority of variance in Multiplier remains unexplained due to:

- External factors (**marketing** budget, timing, social trends)
- Creative factors (movie script **quality**, direction, **performance** of actors)
- And many other aspects

Conclusion

- Possible to identify trends and uncover some facts (e.g: amateur critics matter more, some genres/topics are more profitable but riskier, etc...)

- Still really hard to do predictions based on available data (external factors marketing budget, social, actors' performance)

References

Dataset:

https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?

https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset

Methods:

https://medium.com/@quindaly/step-by-step-nmf-example-in-python-9974e38dc9f9

https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

https://scikit-learn.org/stable/modules/ensemble.html

https://xgboost.readthedocs.io/en/release_3.0.0/

https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/