

Translating Box Office Outcomes

Basswood Project Group,
Erdős Fall 2022 Cohort



Table of Contents

01



ABOUT US

Introducing the Fall 2022
Basswood group

02



PROJECT BACKGROUND

Box office revenue
prediction

03



GOALS & STRATEGY

Using The Movie
Database metadata

04



METHODOLOGY

Use of machine learning
techniques in Python

05



TRANSLATION

Interpretation of results

06



KPI OVERVIEW

Why we actually don't
recommend exclusive focus on
box office revenue

About Us

01



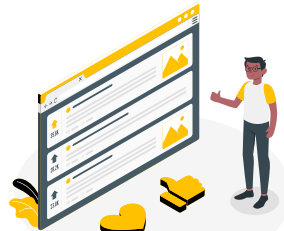
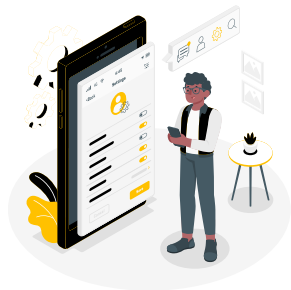
xF Team Science

Cross functional team science came to life with the Basswood project, a collaboration between PhD candidates in different departments with different coding experience

Sridhar Venkatesh, Math Ph.D., University of Michigan, algebraic geometry, problem solving

Kaitlin Engelbrecht, Political Science Ph.D, Ohio State University, insights, experience, and marketing research

Reebhu Bhattacharyya, Math Ph.D, University of Michigan, geometry and topology, visualization





Project Background

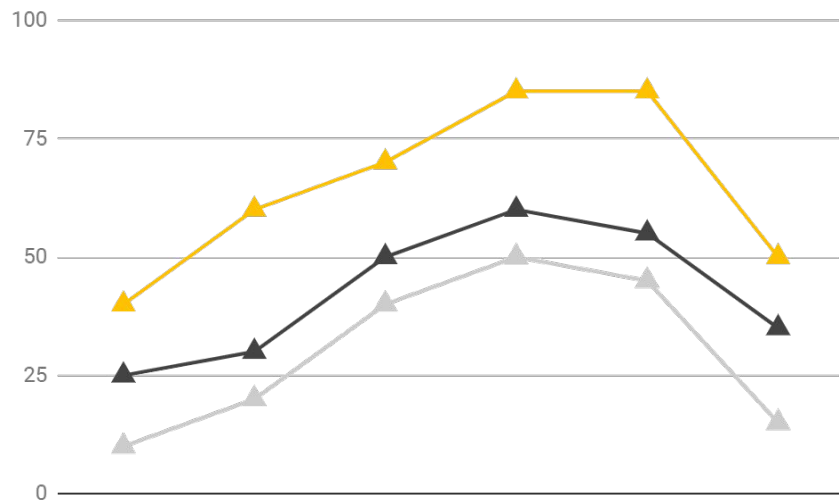
Using The Movie Database and machine learning techniques, we have built a model predicting films' box office revenues

We find that budget (how much money is invested in the production and promotion process) is most predictive of box office success, with important caveats

02

Goals and Strategy

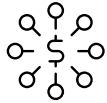
03



Our job as data scientists is not just to analyze data but to provide the strategic framework prior to data analysis (data cleaning, variable selection, model fit) and to contextualize the findings post analysis

Predicting box office revenue for past films

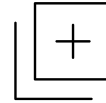
Aspects of metadata we were able to use, as
well as accounting for inflation in our analyses
of budgets and revenue



Crew Info



Cast Info



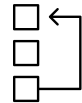
Release Dates



**Genres and
Run Times**

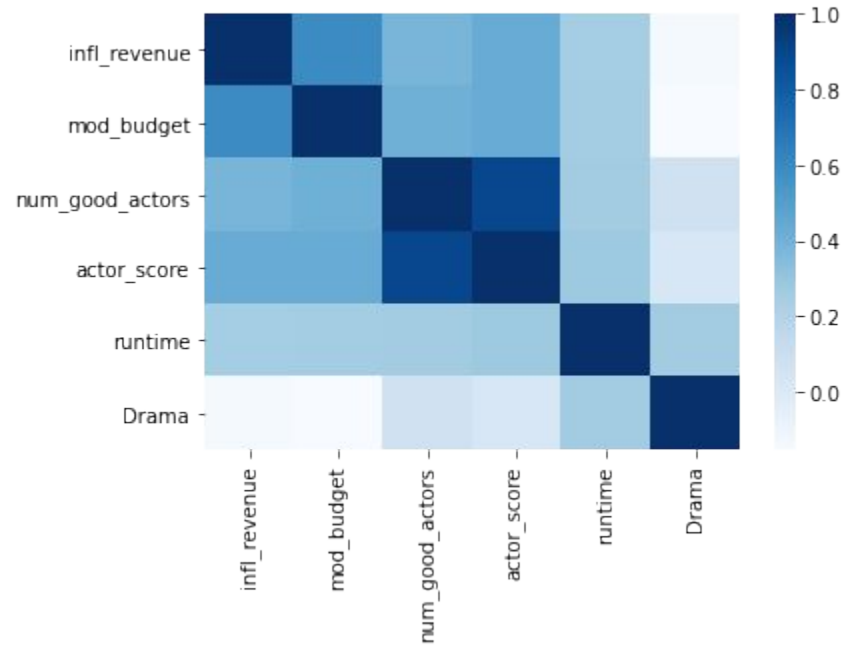


Ratings

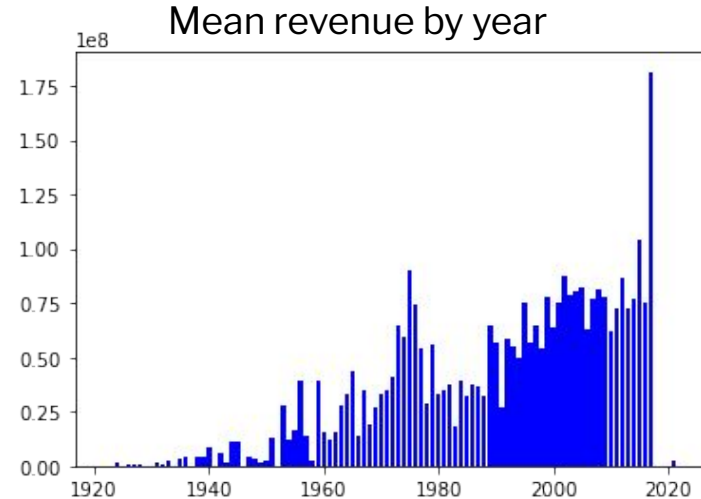
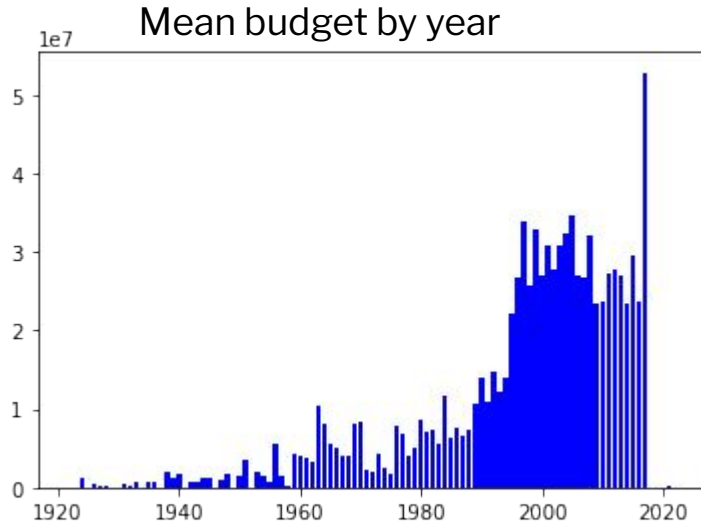


**Overall
Investment**

Budget is most useful predictor



Budget Comparisons



Film budgets have steadily increased in the last century of filmmaking, but especially in the last two decades. Revenues have followed a similar trajectory: there is much money to be spent and money to be made in films.

Cast Quality

- Each actor given rating, driven by IMDb ratings
- Scores are based on the popularity and IMDb rating of movie
- To account for discrepancies, we gave a weight based on how many movies of the actor appeared in the dataset

	MoviesN	AvgPop	AvgIMDB	Score
Samuel L. Jackson	30	0.049898	0.673251	0.486245
Robert De Niro	30	0.031322	0.683951	0.488162
Morgan Freeman	27	0.059705	0.673983	0.489699
J.K. Simmons	25	0.042024	0.675556	0.485496
Liam Neeson	25	0.041218	0.660741	0.474884

Strategy

Being a sequel or part of a highly anticipated series is lucrative

We cannot simply advise that a production company produce sequels (promising high predicted revenues), as they will still need to produce the initial film, so being part of a collection is not the most useful predictor for revenue generation

We opt for regression analyses which are sensitive to outliers, because the inclusion of collection films means we have extreme outliers with respect to budget (cost of production)

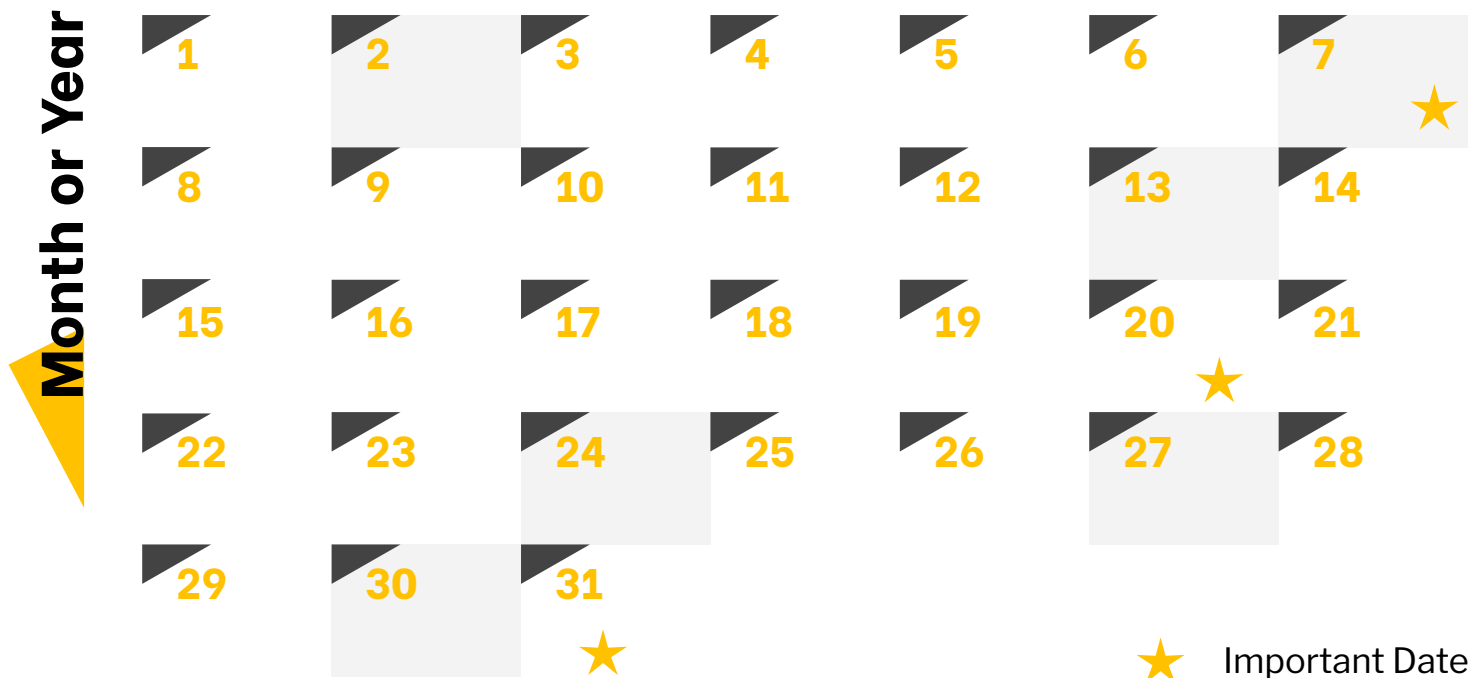
Top 10 highest budget films (est.)

Rank	Title	Estimated Cost in Millions	Part of collection
1	Pirates of the Caribbean: On Stranger Tides	\$379	Yes
2	Avengers: Age of Ultron	\$365	Yes
3	Avengers: Endgame	\$356	Yes
4	Avengers: Infinity War	\$325	Yes
5	Pirates of the Caribbean: At World's End	\$300	Yes
	Justice League	\$300	Yes
7	Solo: A Star Wars Story	\$275	Yes
	Star Wars: The Rise of Skywalker	\$275	Yes
9	John Carter	\$264	Planned collection (cancelled)
10	Batman v Superman: Dawn of Justice	\$263	Yes

Comparison of films part of collection and not

- Average revenue for movies belonging to collection is: \$79 million
- Max revenue for movies belonging to collection is: \$1.53 billion
- Average revenue for movies not belonging to collection is: \$281 million
- Max revenue for movies not belonging to collection is: \$2.95 billion

Release Date Relevance



Methodology

All analyses conducted in Python

Primary methods are linear models

Multiple linear regression employed because of relevance of outliers

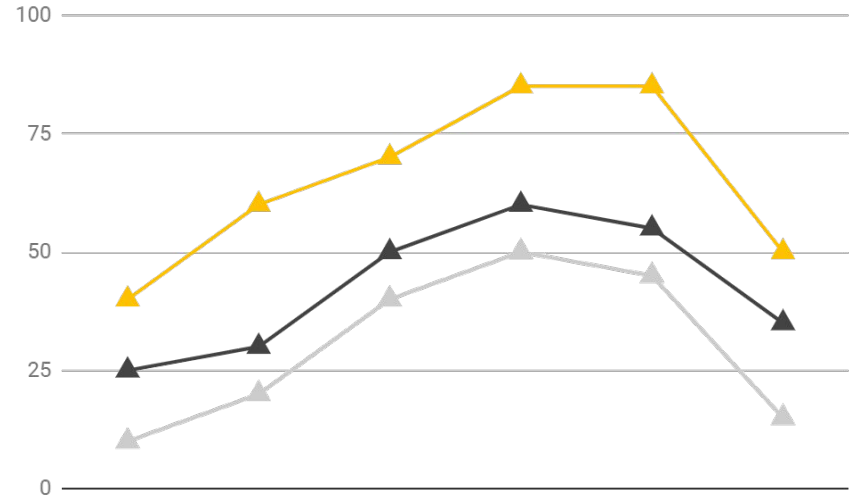
04

Missing values & other considerations

- TMDb is a useful database but often incomplete
- Approx 800 movies in dataset have no budget recorded
- Film budgets are not always public knowledge, but can be estimated with high degree of accuracy
- For our purposes, defined budget to be average of nearby released movies in the same genre

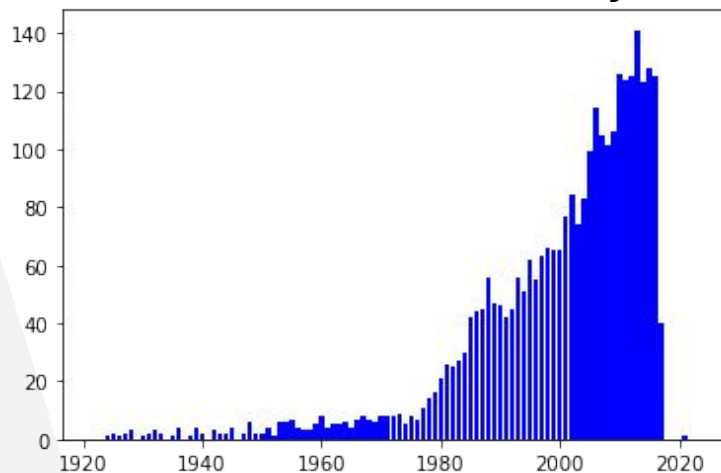
Interpretation of Results

05

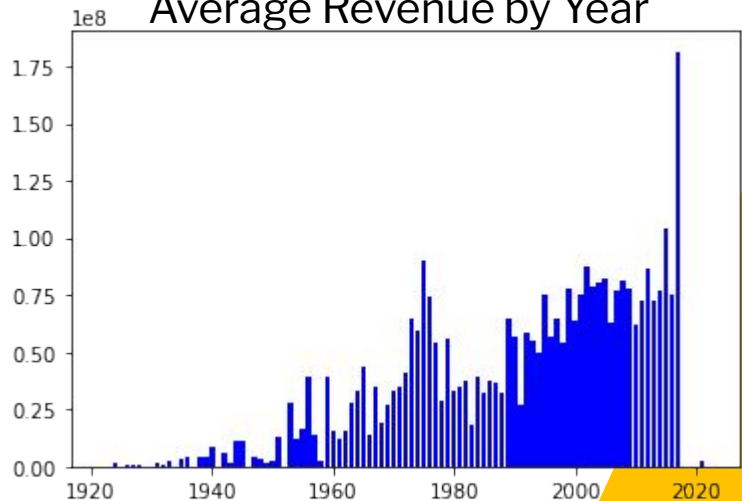


Film production has steadily increased over time, and revenues have followed

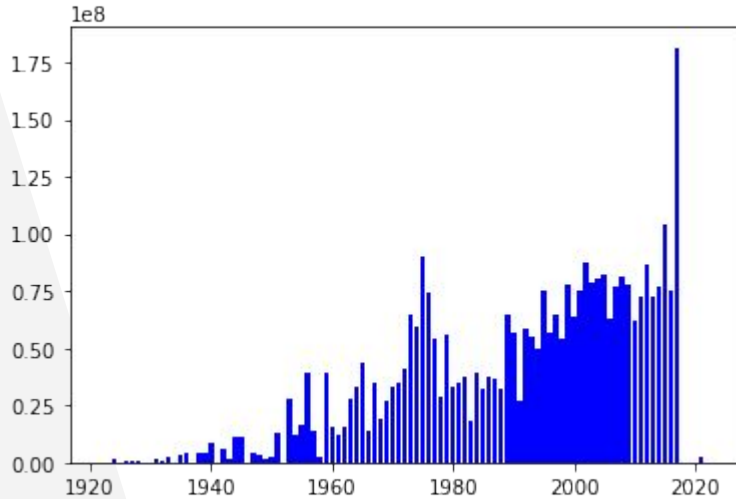
Count of Films Produced by Year



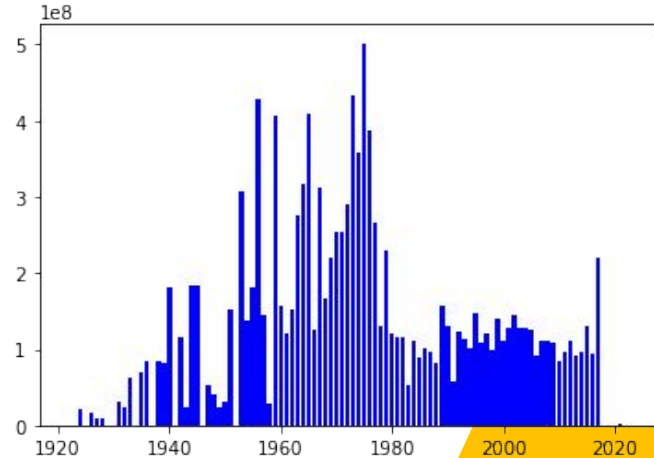
Average Revenue by Year



Importance of accounting for inflation



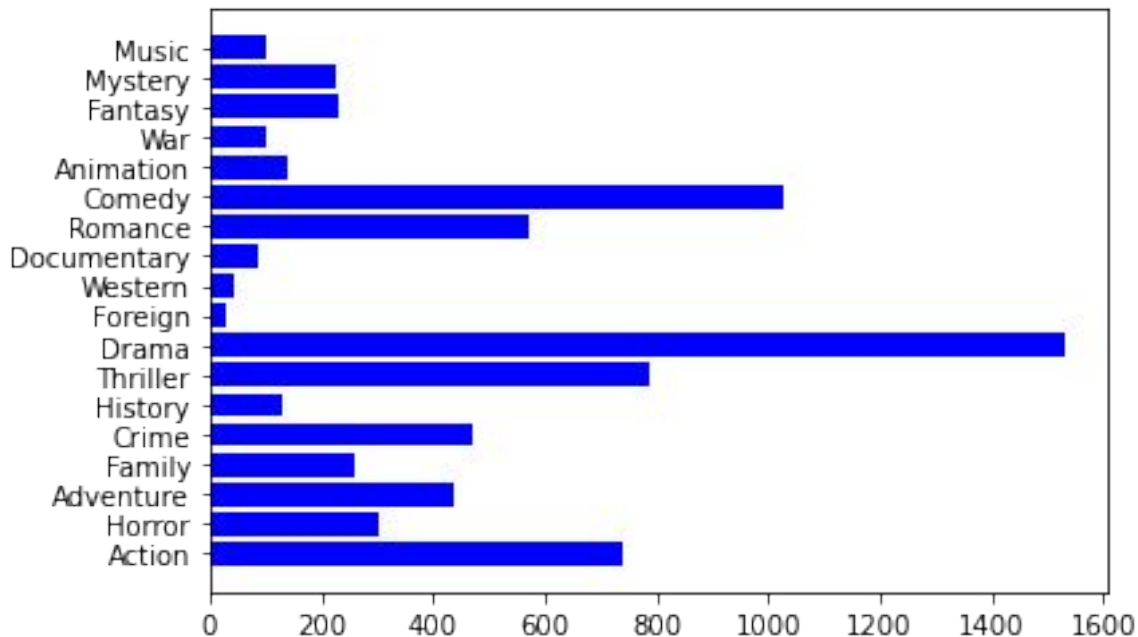
Avg Revenue by Year



Avg Revenue by Year adjusted for inflation

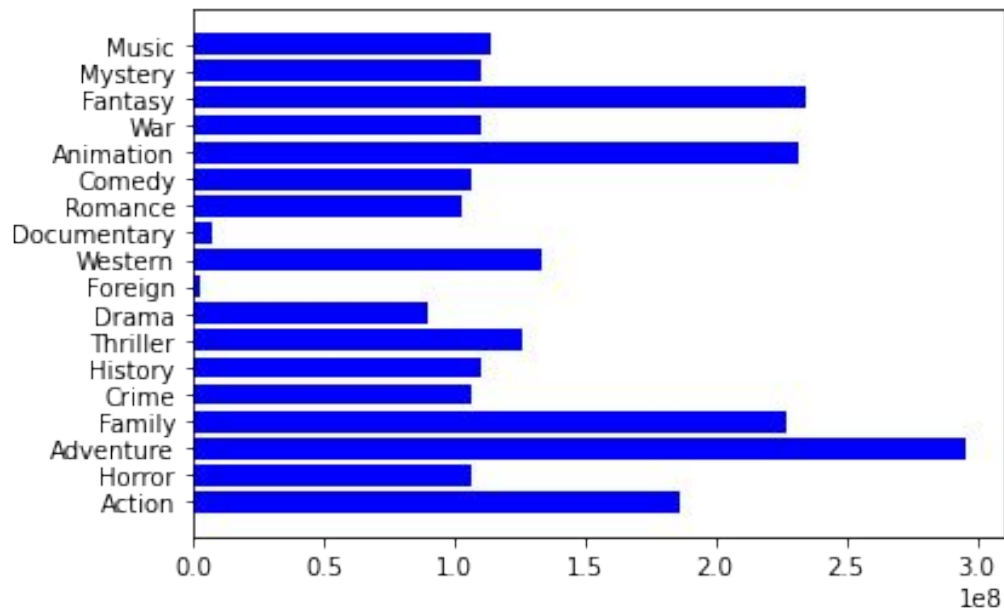
Dramas are the most produced genre, but they are a great gamble

Count of films by genre, all time



Dramas are the most produced genre, but they are a great gamble

Revenue of films by genre, all time



Revenue by Genre: Trends

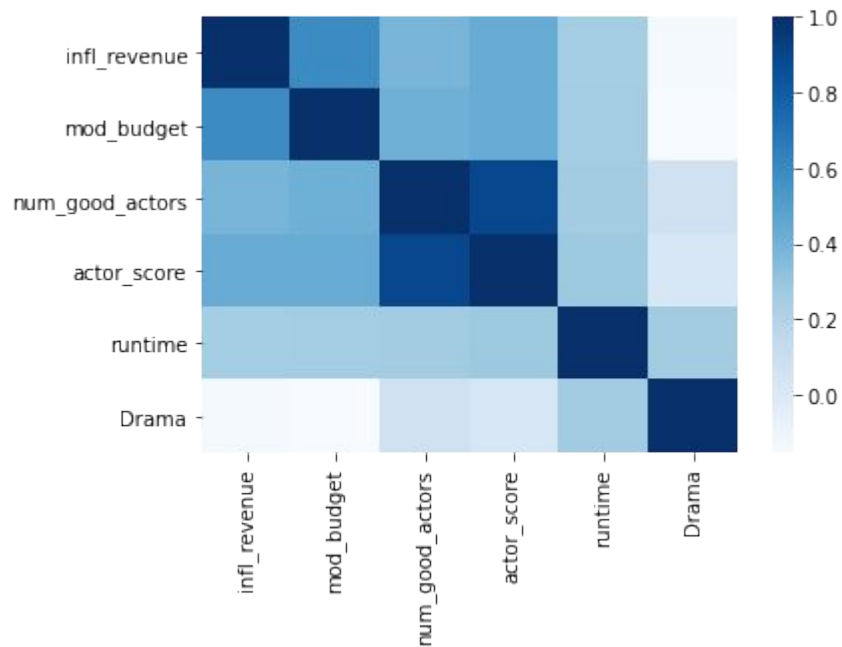
On average, dramas are the most poorly performing genre at the box office, and producing a drama is not predictive of box office success (high revenue)

Genres included in our analysis with TMDb:

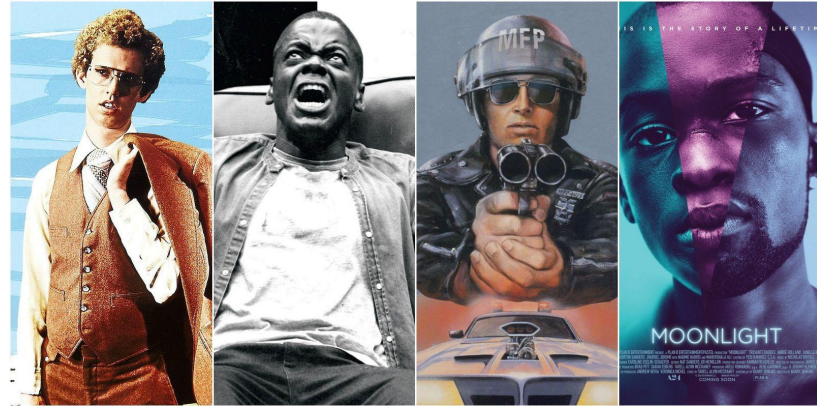
- Adventure
- Comedy
- Horror
- Mystery
- Drama
- Action
- Crime
- Family
- Romance
- Fantasy
- Thriller
- Documentary
- Foreign
- Western
- Animation
- War
- Music
- History

Fantasy and Adventure films perform best over time

Top Predictive Performers



Awards, Ratings, Reviews



What factors predict box office revenue?

Trends in films we find important:

BUDGET

GENRE OF FILM

SERIES/COLLECTION

CAST 'QUALITY'

RELEASE DATE

PROMOTION



Recommendations

1. **Budget is the best predictor** for revenue generation: stakeholders should know that they are more likely to increase revenue with investment in movie budget
2. Revenue prediction is tricky due to **noisy measures**, opaque reporting of budgets and other factors, missingness in available data
3. **Changing nature of film production and revenue generation:** Covid-19 and rise of streaming platforms

Thanks

We hope this was a useful look into the dynamics of film revenue and important KPIs for production companies and anyone interested in box office performance!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik** and illustrations by **Stories**

Sridhar Venkatesh, Math Ph.D.,
University of Michigan
Kaitlin Engelbrecht, Political Science
Ph.D, Ohio State University
Reebhu Bhattacharyya, Math Ph.D,
University of Michigan