Explaining Box Office with Minimal Information

Can a Movie's Success Be Explained by Basic Metadata?

Isabelle Huang

December 15, 2025

Movie box office performance is an important topic in both industry and academic research, informing investment decisions, marketing strategy, and our understanding of audience behaviour. This study examines how well simple movie attributes can explain box office outcomes without relying on complex models. Using a dataset of 6,897 films released between 2006 and 2015, we fit linear regression and mixture-of-regressions models to explain box office performance using only genre, distributor, MPAA rating, release time, and a small set of title-based indicators. Our best model achieves an R^2 of 0.68, suggesting that basic metadata captures much of the variation in the data but is not sufficient for accurate explanation. Richer information such as budget, marketing, franchise status, and word-of-mouth effects is likely essential for more accurate analysis.

Table of contents

1 Introduction

Movie theatres are a common destination for families and friends to gather and briefly escape from daily life at a relatively low cost. Yet thousands of films are released worldwide each year, and only a small fraction are widely watched or remembered. What determines which movies audiences choose? Prior work has studied determinants such as budget, sequels, star power, and reviews, and has used these variables to model or forecast revenue [@{Hao2023FactorsFilmRevenue, Scott2019ReturnRegressions}]. Many studies also apply machine learning methods to improve predictive performance (Jange Zarate and Aragon Encarnacion 2025). This raises a natural question: how much of a movie's box office can be explained using only easily available information and simple models? If strong performance is possible with minimal predictors, decisions could be made at lower cost, both in data collection and in analysis. In this project, we investigate the explanatory power of basic movie metadata under simple regression models.

We first fit a linear regression model to explain ticket sales using only genre, distributor, MPAA rating, release year and month, and a small set of indicator variables capturing whether the title contains certain frequent words. We then fit a mixture-of-regressions model to test whether a more flexible specification can substantially improve fit using the same limited feature set. Our main estimand is the log expected number of tickets sold for a movie with given basic attributes. Let Y_i denote the log number of tickets sold for film i, and let X_i denote its features. The target regression function is

$$m(x) = \mathbb{E}[\, Y_i \mid X_i = x \,]$$

Overall, these simple variables explain box office outcomes to a meaningful extent, though not necessarily sufficient. After tuning, our best model achieves an R^2 of approximately 0.68, leaving roughly one-third of the variation in ticket sales unexplained. Among the predictors considered, genre has the strongest association with ticket sales, while title-based indicators contribute the least. Although the resulting models are not accurate enough for practical forecasting, they may serve as useful baselines and as intermediate proxies when richer covariates are unavailable. Future work could either extract more signal from limited data or incorporate additional predictors to capture more complex drivers of performance.

The remainder of this paper is organized as follows. Section 2 describes the dataset and cleaning procedures. Section 3 presents the modeling approach and empirical results. Section 4 discusses findings and limitations and outlines directions for future work.

2 Data

2.1 Overview

We use the statistical programming language Python (Python Software Foundation, n.d.) together with the data library Pandas (The pandas development team 2025) to clean and analyze our dataset. The data come from the Data and Story Library (Data and Story Library, n.d.), an open-source repository of real-world datasets for teaching and practice. Our dataset contains 6,897 major films released between 2006 and 2015, with variables including title, genre, distributor, MPAA rating, release date, and gross revenue.

2.2 Outcome variables

There are two variables in the dataset that measure box office performance: gross revenue and the number of tickets sold. Gross revenue is an aggregate outcome that depends on realized attendance, ticket prices, and the length of the theatrical run, reported in USD. The number of tickets sold is an approximate count based on information provided by theatres and production companies. Because gross revenue can be derived from ticket sales, we focus our analysis on the number of tickets sold. Since this response is numeric, linear regression is a natural starting point.

However, ticket sales are highly right-skewed: a small number of blockbusters sell millions of tickets, while many films sell only a few thousand or fewer. The mean number of tickets sold is about 1.77 million, but the median is only 22,962, indicating that the mean is heavily influenced by a small set of extreme values. Consistent with this, the standard deviation is large at 5.45 million, reflecting substantial dispersion in ticket sales across films.

We visualize films with the highest and lowest ticket sales in the following graphs.

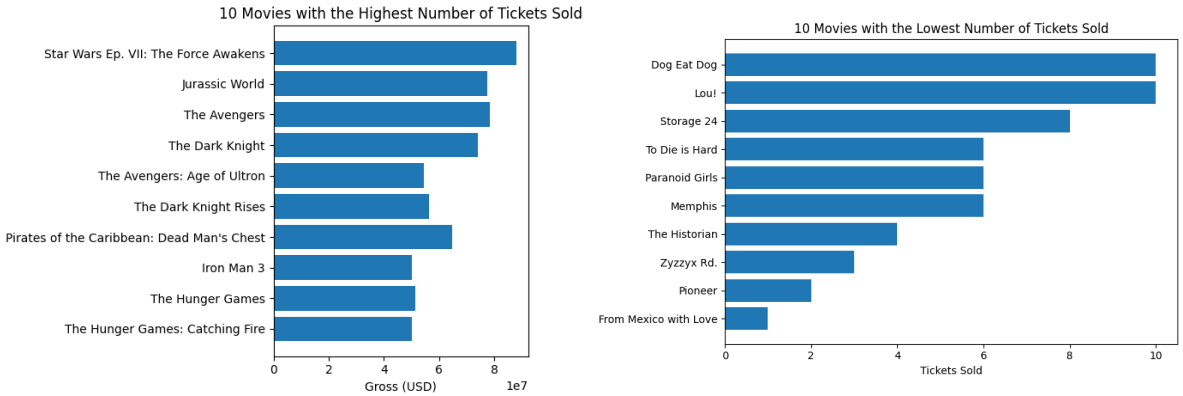

Figure 1: Ticket sales for the ten best- and worst-selling movies.

Because of the strong clustering near zero, we work with a base-10 log transformation of tickets sold instead. This compresses the scale and reduces the influence of extreme values, producing a distribution that is more symmetric and closer to normal. The mean of log tickets sold is about 7.38, the median is about 10.0, and the standard deviation is about 2.87. The transformed distribution is still slightly bimodal, with one peak around 4 and another around 7.

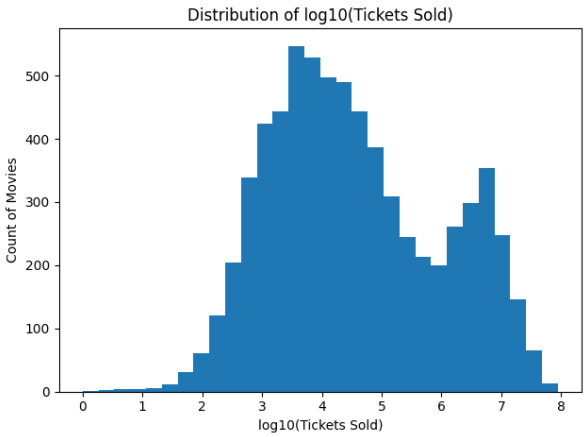

Figure 2: Number of tickets sold on the log base 10 scale

2.3 Predictor variables

We treat all remaining variables in the dataset (excluding gross revenue) as potential predictors of ticket sales. Our primary predictors of interest are Genre and Distributor. Genre is a categorical variable describing the film's primary genre such as Action and Comedy. In our sample, Comedy and Adventure are among the most popular genres in recent years, while Drama is the most frequently produced genre (Figure 3). Distributor is also categorical and identifies the company responsible for releasing the film to theatres. Between 2006 and 2015, IFC and Warner Bros appear most often in the dataset, with roughly 300 films each.

The remaining variables are expected to have weaker effects on box office performance, but we include them for a more complete picture. MPAA rating captures the film's content rating (e.g., G, PG, PG-13, R). Release year and release month record when the film was released and may capture timing effects such as holiday-season boosts.

In addition, we incorporate information from movie titles by extracting simple text-based features. We test whether including certain keywords in a title is associated with higher ticket sales, under the idea that recognizable or appealing words may attract audience attention. We identify the most frequent words appearing in titles—"man", "love", and "life"—and create three indicator variables for whether each word appears in a given title. These variables are then merged into the main dataset.

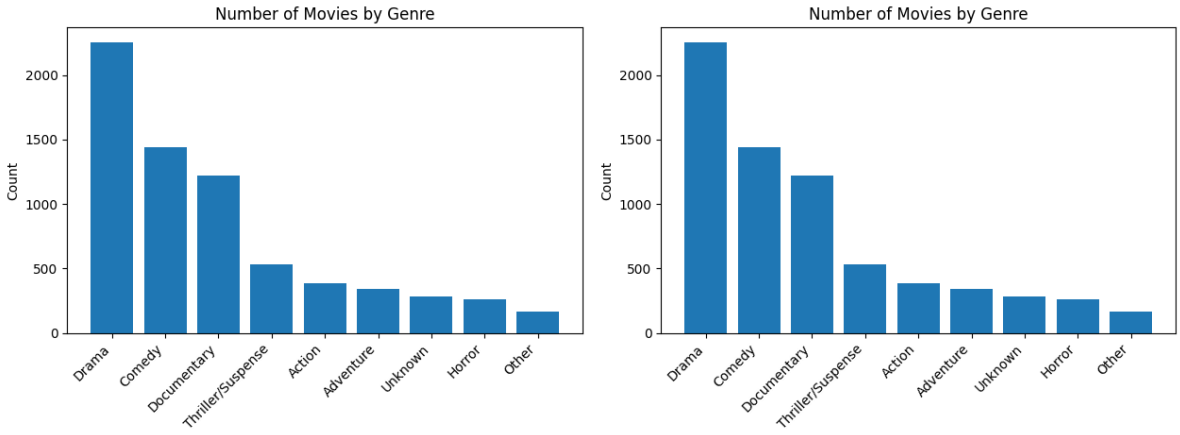

Figure 3: The best selling genres and most produced genres of movies.

Rows with missing values were removed from the dataset prior to analysis, leaving 6,337 observations. The dataset is relatively well curated, so no further cleaning is required. That said, because this is an observational dataset of commercially released films, some measurement error may be present, for example, in approximate ticket counts or genre assignments. We treat the recorded values as given throughout.

3 Model

Based on the exploratory analysis, it appears reasonable to model the base-10 log of tickets sold as a linear function of the predictors. In addition, motivated by the bimodal shape of the transformed distribution, we also fit a mixture-of-regressions model to test whether allowing two latent groups can better capture the data and improve model fit.

3.1 Model set-up

3.2 Linear regression model

To estimate the regression function $m(x)$ we first fit Linear Regression model for the logged tickets sold using Ordinary Least Square Estimators. For each movie $i \in \{1, \dots, n\}$, let

- Y_i be logged number of tickets sold,
- $x_{i,\text{year}}$ and $x_{i,\text{month}}$ be release year and month of the movie,
- G_i, D_i, and M_i denote the categorical variables Genre, Distributor, and MPAA rating,
- $z_{i,\text{man}}, z_{i,\text{love}}, z_{i,\text{life}} \in \{0, 1\}$ be the three title indicators.

Then the full linear model can then be written as

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_1 x_{i,\text{year}} + \beta_1 x_{i,\text{month}} \\ & + \sum_g \gamma_g \mathbb{1}\{G_i = g\} + \sum_d \delta_d \mathbb{1}\{D_i = d\} + \sum_m \eta_m \mathbb{1}\{M_i = m\} \\ & + \alpha_{\text{man}} z_{i,\text{man}} + \alpha_{\text{love}} z_{i,\text{love}} + \alpha_{\text{life}} z_{i,\text{life}} + \varepsilon_i \end{aligned}$$

We implement this model in Python using Seabold and Perktold (2010) Linear Rregression makes several assumptions, including linearity, independence, homoscedasticity, and normality of errors and these assumptions are roughly satisfied in our case.

3.3 Mixture Regression Models

We also consider a finite mixture and linear regression for Y_i with the same predictors that we previously considered. Let $Z_i \in \{1, 2\}$ be an unobserved component indicator such that

$$\Pr(Z_i = k) = \pi_k, \qquad k = 1, 2,$$

with $\pi_k > 0$ and $\pi_1 + \pi_2 = 1$. Conditional on $Z_i = k$ we assume a standard linear regression model

$$Y_i \mid (Z_i = k, \mathbf{X}_i) \sim \mathcal{N}(\mathbf{X}_i^\top \beta_k, \sigma_k^2), \qquad k = 1, 2,$$

where β_k is the vector of regression coefficients and σ_k^2 is the error variance for component k. So this allows us to model the relationship specific to the two groups of movies that we suspect exist in the data.

Marginally, the conditional density of Y_i given $\mathbf{X}_i$ is

$$f(y_i \mid \mathbf{X}_i, \theta) = \sum_{k=1}^{2} \pi_k \, \phi(y_i \mid \mathbf{X}_i^\top \beta_k, \sigma_k^2),$$

where $\theta = (\pi_1, \pi_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$ and $\phi(\cdot \mid \mu, \sigma^2)$ denotes the univariate Normal distribution. The observed-data likelihood for θ is therefore

$$L(\theta \mid \{y_i, \mathbf{X}_i\}_{i=1}^n) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{2} \pi_k \, \phi(y_i \mid \mathbf{X}_i^\top \beta_k, \sigma_k^2) \right\},$$

and can be maximised using an expectation–maximization (EM) algorithm which is built in to the gmm model provided by ScikitLearn (Pedregosa et al. 2011).

3.4 Model justification

As discussed earlier, our response variable is numeric, making linear regression a natural starting point. Linear regression has four assumptions: linearity, uncorrelated errors, homoscedasticity, and normal errors. Our exploratory analysis suggests that after the log transformation, the response is roughly symmetric and close to normal, which supports the normality assumption. Moreover, because our focus is model fit and overall explanatory power rather than precise inference on individual coefficients, mild departures from normality are less concerning.

Residual diagnostics do not show strong systematic patterns, and the spread of residuals is fairly stable across fitted values, suggesting that the linearity and homoscedasticity assumptions are at least approximately reasonable. One potential issue is multicollinearity–using many indicator variables for high-cardinality categorical predictors such as Genre and Distributor, can introduce strong correlations among predictors and make individual coefficient estimates unstable. Since our primary goal is the overall performance rather than interpreting specific coefficients, we treat this as a secondary concern.

On the other hand, the bimodal shape of the log-transformed response suggests the sample may reflect two latent groups. This motivates the use of mixture models. Accordingly, we also apply a Gaussian mixture model (GMM) to partition films into two components, and then fit separate linear regression models within each component to allow the relationships between predictors and ticket sales to differ across groups.

4 Results

The linear regression on the log scale achieves an in-sample R^2 of about 0.68 and a root mean squared error (RMSE) of approximately 0.82 in log10 units. Thus, while the model explains substantially more variation than the unlogged version (which had $R^2 \approx 0.35$), there is still considerable unexplained heterogeneity in box office performance.

The coefficient estimates suggest that distributor and genre are among the strongest predictors of log ticket sales. Large studio distributors such as Walt Disney, Universal, 20th Century Fox, and Warner Bros. are associated with substantially higher expected log ticket sales than the baseline distributor, conditional on the other covariates. Many smaller distributors, in contrast, are associated with lower expected sales, consistent with the idea that well-known distributors have greater reach and marketing capacity. Across genres, Adventure and Action are associated with higher ticket sales relative to the reference genre, whereas Documentary and Drama are associated with lower sales. MPAA ratings also show a pattern that PG and PG-13 films tend to have higher expected ticket sales than the baseline rating, while Not Rated and G films are associated with lower sales. However, since the majority of films in the dataset are Not Rated, this estimate may be a result of differences in representation across rating categories rather than an actual effect. Finally, the simple title indicators (whether the title contains “man”, “love”, or “life”) have small coefficients and do not materially improve model fit.

Overall, the log-transformed model captures broad, interpretable associations between basic metadata and demand.

Predictor	Estimate	Std Error	p-value
Genre = Adventure	0.097	0.072	0.178
Distributor = 20th Century Fox	2.835	0.862	5.9e-09
MPAA = T.PG-13	0.395	0.092	1.7e-05
Genre = Documentary	-0.233	0.061	0.000
T.Universal	3.062	0.862	0.000

In terms of the mixture regression, we fit a two-component Gaussian mixture model to the log-transformed ticket sales to split the data into two groups, as suggested by the bimodality of response. yielding a "low–box-office" group (component 1, $n = 4{,}484$) and a "high–box-office" group (component 0, $n = 2{,}196$).

We then fit separate linear regression models within each component. The resulting in-sample fit is weaker than the single pooled regression. The component-wise R^2 values drop to approximately 0.38 and 0.52. A likely explanation is that splitting the data reduces the effective sample size within each group while retaining a large number of indicator predictors, leading to noisier and less stable estimates. Overall, this mixture-based strategy does not substantially improve our ability to explain variation in ticket sales using the limited feature set.

That said, the component-specific regressions recover broadly similar qualitative patterns. Action and Adventure genres and major studio distributors are associated with higher log ticket sales, whereas more niche genres and non-major distributors tend to be associated with lower sales, conditional on the remaining covariates.

High Box Office Group (Component 0):

Predictor	Estimate	Std Error	p value
Distributor = Sony Pictures Classics	-0.859	0.062	0.000
Distributor = Roadside Attractions	-0.8956	0.094	0.000
Distributor = Paramount Pictures	0.0453	0.053	0.393
MPAA = PG	0.154	0.065	0.018
Genre = Drama	-0.3106	0.040	0.000

Low Box Office Group Coefficients(Component 1):

Predictor	Estimate	Std Error	p value
Distributor = 20th Century Fox	0.6007	0.707	0.396
Genre = Adventure	0.1209	0.101	0.232

Predictor	Estimate	Std Error	p value
MPAA = Not Rated	-0.3786	0.120	- 0.002
Distributor = Walt Disney	0.5205	0.705	0.460
Genre = Drama	0.1609	0.066	0.015

5 Discussion

5.1 Summary of analysis

This report investigates how much of movie box office performance can be explained using simple models and easily obtained data. We analyze a dataset of 6,897 major releases from 2006–2015. Based on exploratory analysis, we adopt linear regression to model ticket sales. Because ticket sales are extremely right-skewed, we model the log-transformed number of tickets sold as a function of release year and month, distributor, genre, MPAA rating, and a small set of title indicators.

Motivated by the bimodality of the log-transformed response, we also consider a mixture-based approach. We first fit a two-component Gaussian mixture model to the response to identify two latent groups. We then fit separate regression models within each component to assess whether explanatory power improves through grouping, and whether the relationships between predictors and outcomes differ between films with high ticket sales and those with low ticket sales.

The models were then evaluated using in-sample R^2 and RMSE. Assumptions were also checked using residual diagnostics.

5.2 Key Findings

From this analysis, we find that basic movie attributes can explain a moderate amount of variation in box office performance. Our linear regression achieves an R^2 of 0.68, which is non-trivial. This type of model can serve as a useful intermediate step for understanding broad patterns in how observable attributes relate to ticket sales, especially when computation or data-collection capacity is limited.

The results suggest that films released by major studio distributors and belonging to commercially oriented genres, such as Adventure and Action, tend to sell more tickets than films from smaller distributors and narrower genres. In contrast, genres like Documentary and Drama which are less light-hearted are associated with lower sales. This aligns with the intuitive idea that institutional backing and genre positioning matter for commercial success and also that the audience tend to prefer something relaxing when heading to the theatres. Although we initially hypothesized that certain keywords in a title might attract audiences, the title-based indicators

we tested have minimal association with ticket sales. Overall, with minimal information and simple models, we were able to generate several meaningful inference in relationship between basic movie attributes and box office performance.

Moreover, the mixture analysis suggests that the movie market can be partitioned into at least two groups: a large group of films with low-to-moderate ticket sales and a smaller group of films with very high ticket sales. A two-component Gaussian mixture fits the log-ticket distribution noticeably better than a single Normal, with one component centered at much lower sales and the other corresponding to blockbuster-level outcomes.

When we fit separate regressions within these groups, basic attributes appear more informative for high-performing films. The regression within the high-sales component achieves a higher R^2 and shows clearer, more stable associations for distributor and genre, whereas the low-sales component has weaker fit and greater residual variability. This is seems to suggest that once a film reaches an upper commercial tier, its performance is more systematically tied to structural factors such as distributor reach and genre positioning. In contrast, outcomes in the lower tier appear more idiosyncratic and less predictable from simple metadata. These results also point to mixture-based frameworks as a promising direction for future work, particularly if combined with additional covariates to better characterize the mechanisms that differentiate film types and drive box office outcomes.

5.3 Limitations and future work

However, this analysis has several important limitations. From a data perspective, the dataset is observational and restricted to major releases from 2006–2015. As one decade has passed, audience preferences and industry structure may have great shifts since then. Patterns that were explainable with simple attributes in this period may not generalize to today's market. In addition, focusing on major releases limits external validity for smaller independent films. Measurement error is also possible, for example, genre and MPAA categories can be coarse and may group heterogeneous content together.

Methodologically, our models face multicollinearity due to the large number of dummy variables created from multi-level categorical predictors, especially distributor and genre, which can make coefficient estimates unstable. Although the log transformation improves model behavior, the linearity assumption is still only approximate, so some coefficient bias is possible. We also did not consider interaction effects in attempts to avoid further inflating collinearity, even though interactions may be important.

Finally, our mixture-regression uses hard assignment of films to mixture components and then fits regressions conditional on those assignments. This two-stage approach does not propagate uncertainty from the mixture step into the regression stage, which can understate uncertainty in component-specific results.

This analysis should be viewed as a first step rather than a complete model of box office performance. If we want to continue pursue the path of simple models and minimal predictors, a natural next step is to develop methods that better extract signal from limited data. This is an active research area, with many possible directions that could improve performance in our setting.

However, if the goal is stronger explanation and prediction more broadly, richer data and more flexible models will likely be helpful. Future work could expand the feature set to include production budgets, sequel and franchise indicators, star metrics, and measures of pre- and post-release attention such as critic scores, online ratings, and social media activity. With these covariates, it would be useful to compare regular linear models, tree-based methods, and hierarchical models that allow distributor and genre effects to vary across time or markets. On the mixture side, a finite mixture of regressions fit jointly-estimating both component membership and regression parameters—could provide a principled framework for studying group-specific relationships.

Appendix

A Additional data details

Here are a few additional plots to further illutrate the data.

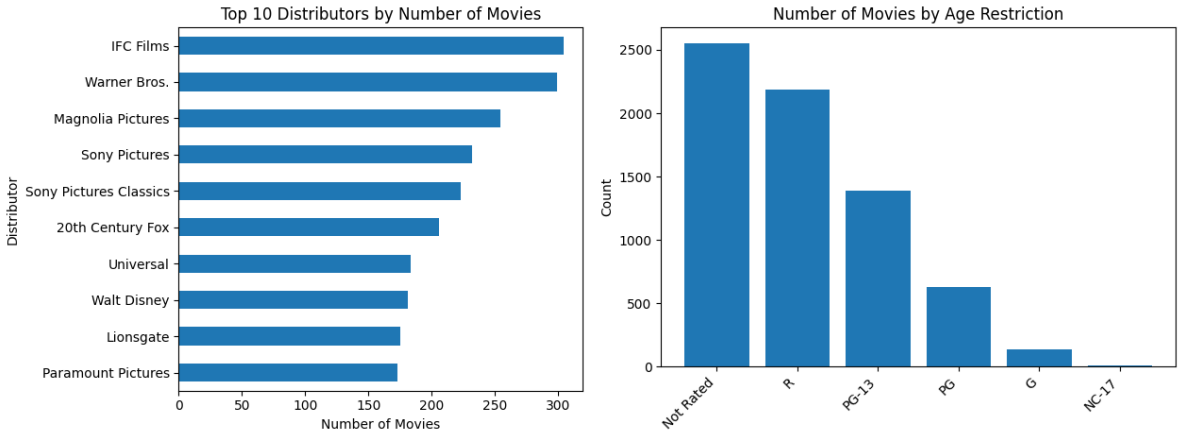

Figure 4: Distributors with the most movies and MPAA age restrictions overview of movies contained in the dataset.

B Model details

On top of the two models considered in the main text, we also consider the following two specifications as part of exploration.

Ordinary linear regression on raw ticket sales

Let $i = 1, \ldots, n$ index movies and define the outcome as the number of tickets sold, Y_i = the number of tickets sold for movie i. For each movie i, we observe:

- Year_i: release year,
- Month_i: release month $(1, \ldots, 12)$,
- G_i: genre (categorical),
- D_i: distributor (categorical),
- M_i: MPAA rating (categorical),

- $z_{i,\text{man}}, z_{i,\text{love}}, z_{i,\text{life}} \in \{0,1\}$: indicator variables for whether the title contains the words "man", "love", or "life".

G_i, D_i, and M_i are categorical variables represented with dummy variables.

The linear regression model on the original ticket scale is

$$\begin{aligned} Y_i = \beta_0 &+ \beta_{\text{year}}\,\text{Year}_i + \beta_{\text{month}}\,\text{Month}_i \\ &+ \sum_g \gamma_g\,\mathbb{1}\{G_i = g\} + \sum_d \delta_d\,\mathbb{1}\{D_i = d\} + \sum_m \eta_m\,\mathbb{1}\{M_i = m\} \\ &+ \alpha_{\text{man}} z_{i,\text{man}} + \alpha_{\text{love}} z_{i,\text{love}} + \alpha_{\text{life}} z_{i,\text{life}} + \varepsilon_i, \end{aligned}$$

where one level of each factor (genre, distributor, MPAA) is taken as the reference category, and the error terms satisfy

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \qquad i = 1, \ldots, n.$$

The parameters are estimated by ordinary least squares. However, this model did not perform well, and only acheived an R^2 of about 0.35, therefore was not included in the main text.

Lasso regression on log-transformed ticket sales

For the penalized model, we work with a log-transformed outcome

$$Y_i = \log_{10}(\text{Tickets.Sold}_i)$$

using the same underlying predictors.

Let $\mathbf{x}_i \in \mathbb{R}^p$ denote the design vector for movie i after preprocessing, where:

- continuous predictors (such as year and month, and the binary title indicators) have been centered and scaled, and
- categorical predictors (distributor, genre, MPAA) have been expanded into dummy variables via one-hot encoding, with one reference level dropped for each factor.

The Lasso model assumes the linear relationship

$$Y_i = \mathbf{x}_i^\top \beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

with mean-zero errors ε_i, but estimates $\beta \in \mathbb{R}^p$ by solving the penalised least squares problem

$$\hat{\beta}_\lambda = \arg\min_\beta \left\{ \frac{1}{2n} \sum_{i=1}^n (Y_i - \mathbf{x}_i^\top \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\},$$

We select λ by K-fold cross-validation. This model was not presented since it slightly decreases the R^2 of our regression of log-transformed tickets sold.

B.1 Diagnostics

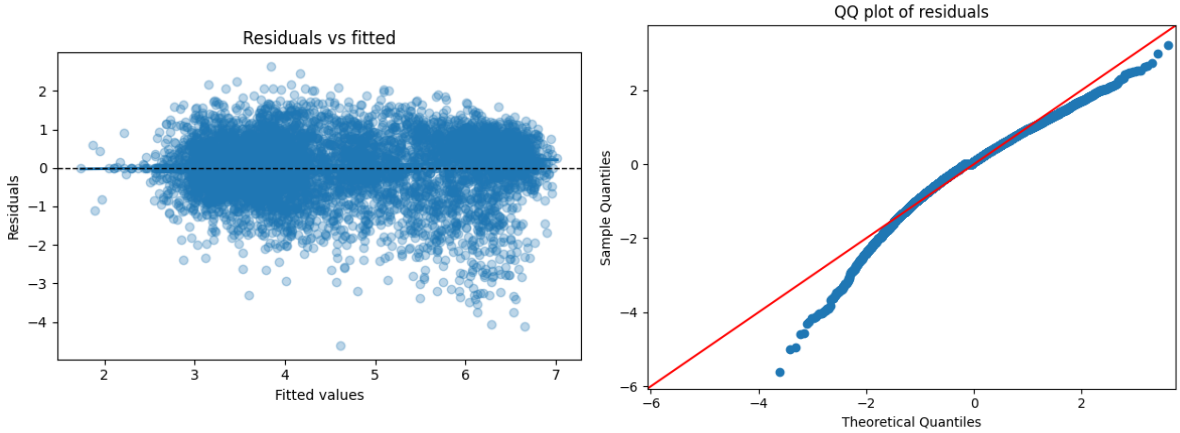

Figure 5: Diagnostic plots for linear regression model residuals. As mentioned in main text, there seems to be

C References

We use Python and several open-source packages for our analysis (**python?**; **harris2020numpy?**; **mckinney2010pandas?**; **pedregosa2011scikit?**; **seabold2010statsmodels?**)

Data and Story Library. n.d. "Movies Box Office Dataset." https://dasl.datadescription.com/.

Jange Zarate, Karla Jemima, and Michael Moises Aragon Encarnacion. 2025. "A Systematic Review on Forecasting for Box Office Success of a Movie Using Machine Learning Prediction Methodologies." In *Research Perspectives on Software Engineering and Systems Design: Proceedings of 8th Computational Methods in Systems and Software 2024, Volume 3*, edited by Radek Silhavy and Petr Silhavy, 1491:378–92. Lecture Notes in Networks and Systems. Cham: Springer Cham. https://doi.org/10.1007/978-3-031-96380-3_33.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Python Software Foundation. n.d. "Python Language Reference." https://www.python.org/.

Seabold, Skipper, and Josef Perktold. 2010. "Statsmodels: Econometric and Statistical Modeling with Python." In *Proceedings of the 9th Python in Science Conference (SciPy 2010)*.

The pandas development team. 2025. "Pandas-Dev/Pandas: Pandas (V2.3.3)." Zenodo. https://doi.org/10.5281/zenodo.17229934.