IMDb Movie Success: The Link Between Movie Release Factors and Box Revenue

The Four-mula: Amy Duan, Hellen Han, Lila Rogers, Eva Aggarwal

2025-04-28

Introduction:

This project aims to examine key factors influencing IMDb movie revenues to understand what drives a movie's box office success. Revenue, a widely used measure of success in the film industry, reflects audience demand and commercial viability. Revenue is defined as the total amount of money a given movie generates from all sources related to the film. Existing research has highlighted various influences, such as star power, genre, and marketing, but the relative importance of these factors remains debated. Thus, the central research question our project aims to answer is: **What production and release factors have the greatest impact on a movie's total revenue?** By addressing this question, we aim to not only deepen the understanding of revenue drivers and gain valuable insights into IMDb movie success, but also improve predictive models for movies' box office performance in the future. We hypothesize that IMDb movies with higher budgets, higher average ratings, and higher vote counts tend to have higher revenues.

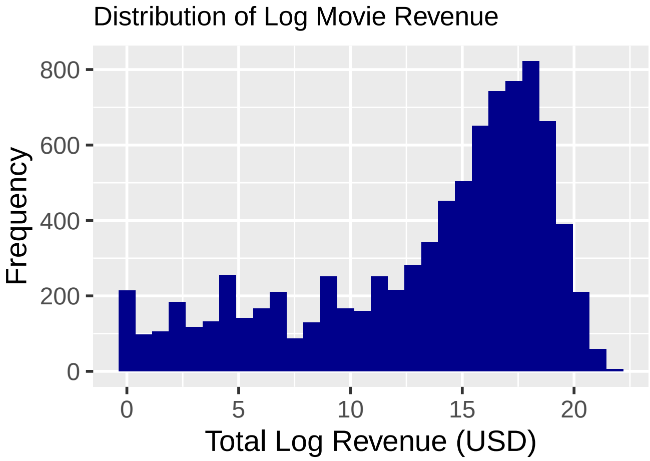
Exploratory Data Analysis

We obtained our data set from Kaggle, an online data science platform with a collection of community-developed open data sets. This data was collected by **Anand Shaw** from the **IMDb website** using various IMDb sites, and converted into a .csv file. The data was updated on a daily basis until 2 months ago. To get more indepth into our data set, the data set we used collected information available on the IMDb website for different movies, such that each observation describes characteristics of a specific movie. In general, the characters being measured follow basic information about the movie, various classifications of the movie's popularity and rating, and the monetary values associated with the movie.

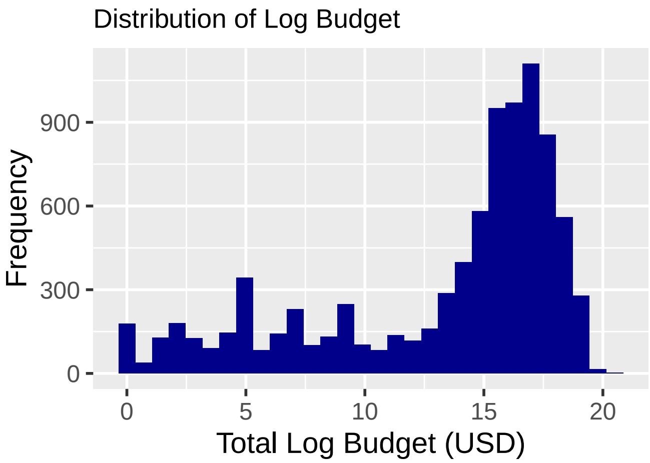
Due to the large size of the original data set, we intentionally drop some characteristics, including `id`, `original_title`, `tagline`, `production_companies`, `overview`, `keyword` from the data set due to their redundancy, irrelevance to the research question, and presence of a significant number of null values. Additional data cleaning was performed for the scope of this project. To obtain informative interpretations of the distribution of the revenue response variable in terms of recency, we analyze IMDb movies released from 2000-present. Missing values and unnecessary fields were removed, and variables names and units were standardized. Furthermore, we log transform revenue and budget, as the revenue generated and budget allocated across the IMDb movies are heavily skewed.

Out of the variables in our dataset, we chose to examine movie revenue as our response variable. Our predictors include: average user rating, runtime, whether the movie is classified as adult content, budget, popularity rating, and total number of movie ratings.

Univariate EDA

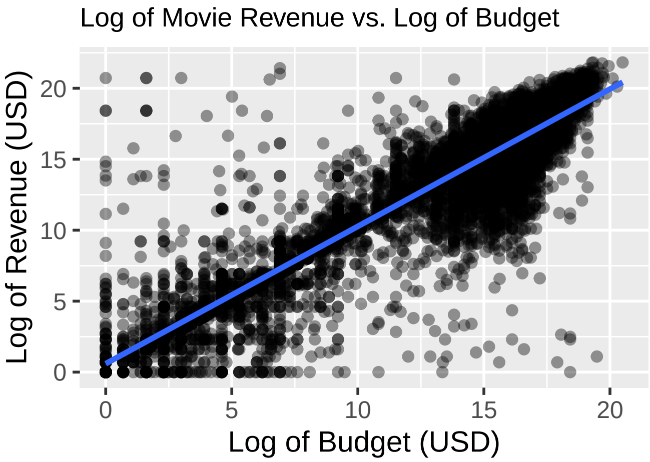


The histogram shows the distribution of the log transformed IMDb movie revenues response variable, which is left-skewed and unimodal, indicating that on average, movies listed on IMDb tend to generate higher revenues. Furthermore, the distribution of the log transformed IMDb movie revenues has a center of approximately 1,800,000 US dollars and a spread of approximately 18,600,000 US dollars.

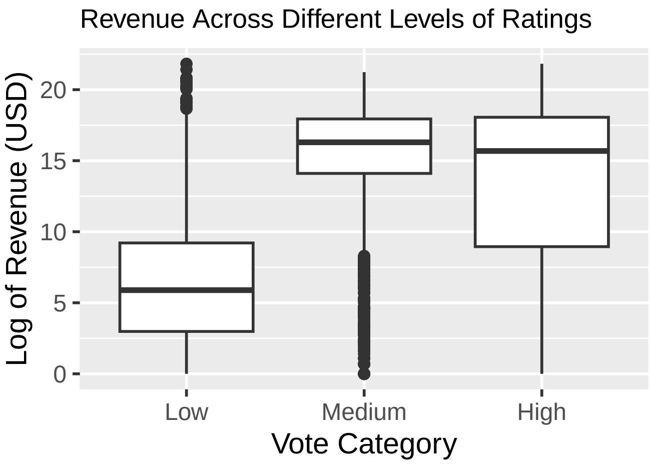


The distribution of the log transformed IMDb movie budget predictor variable is left-skewed and unimodal, indicting that although IMDb movie budget varies greatly, on average, movies listed on IMDb tend to have higher budgets. Furthermore, the distribution of the log transformed IMDb movie budget has a center of approximately 5,000,000 US dollars and a spread of approximately 25,000,000 US dollars.

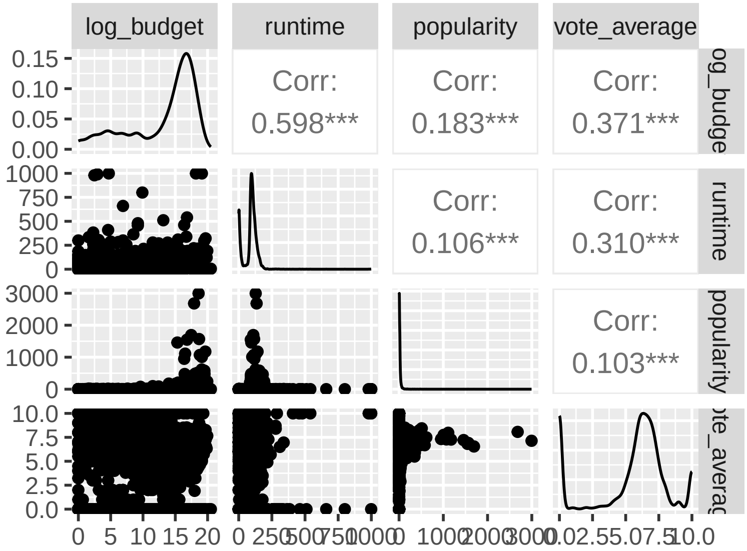
Bivariate EDA



From our visualization of log-transformed revenue versus log-transformed budget, along with the high R² value (0.8366) from the linear regression model, we observe a strong linear relationship between these variables. The p-value (approximately 0) further confirms there is a linear relationship between the log transformed movie revenue and the log transformed movie budget. This aligns with expectations, as higher-budget films typically generate greater audience anticipation and ticket sales.



A categorical vote_category variable was created from the vote_average variable to reflect the relative average rating category for the movie, with levels “Low” for ratings 0-3, “Medium” for ratings 3-7, and “High” ” for ratings 7-10. The boxplot shows that the median revenue for Medium rated movies (approximately 8,880,000 dollars) is slightly higher than those with High ratings (approximately 5,400,000 dollars). The median revenue for Low rated movies is significantly lower (approximately 400 dollars). The IQR of movie revenues with High ratings is also much larger than for Medium and Low rated movies.

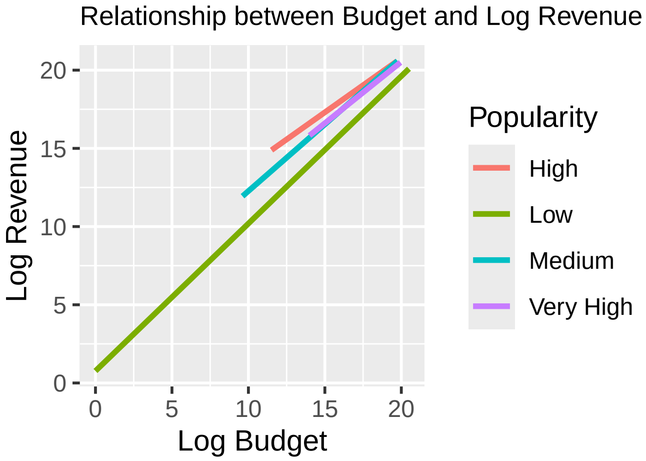


In order to assess potential multicollinearity among predictors, we also created a correlation plot to identify which predictors have high correlations. We can see the highest correlation existing between runtime and budget and vote average and budget. Because of these correlations, we will take special care to check the VIF values in our final model and assess how this multicollinearity should be addressed.

Potential Interaction Effects

Budget and Popularity

After exploring popularity, it is observable that its values range from 0-2994 with an extremely heavy right skew. Thus, popularity is binned into four categories, “Low”, “Medium”, “High”, and “Very High” corresponding to popularity ranges 0-30, 30-90, 90-150 and 150+ respectively.

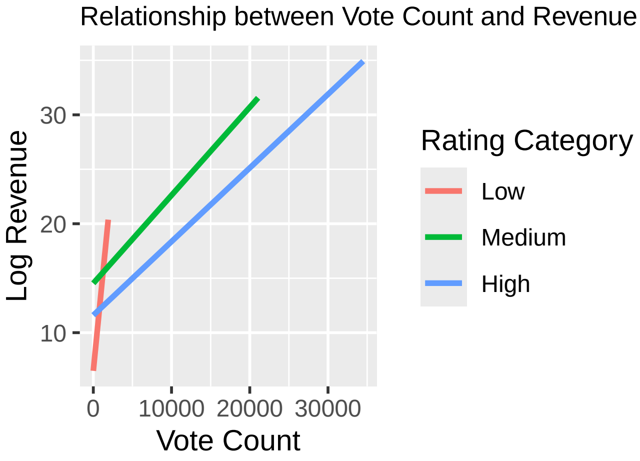


From this visualization, we observe a slight interaction effect between budget and popularity. While the relationship between budget and revenue remains generally positive and linear (as seen in our univariate visualizations), the strength of this relationship varies with a film’s popularity. Specifically, the slope is steeper for more popular films, indicating that popular movies show a stronger positive return on budget investment compared to less popular ones.

This interaction between budget and popularity likely occurs because higher-budget movies typically benefit from greater marketing efforts, which increases their visibility and boosts popularity. The p-value for the budget:popularity interaction term is 0, confirming this effect is statistically significant at the 0.05 level. This means budget’s influence on revenue does indeed vary significantly depending on a film’s popularity.

term	estimate	std.error	statistic	p.value
(Intercept)	-59897669.4	4562623.37	-13.128	0
log_budget	9290256.3	323627.11	28.707	0
popularity	-8099479.2	408355.55	-19.834	0
log_budget:popularity	474931.9	22682.92	20.938	0

Vote Average and Vote Count



From the visualization, we observe a strong interaction effect between vote count and vote average (ratings). The plot shows that the relationship between vote count and revenue varies across rating categories. Specifically, higher-rated movies exhibit a stronger positive association between vote count and revenue compared to lower-rated ones. This indicates that ratings moderate the effect of audience engagement (vote count) on a movie's financial success.

Since movies with higher vote counts tend to also have higher average ratings, there may be a multiplicative effect on revenue. If an interaction exists, it implies that vote count alone does not fully determine revenue—its impact depends on the vote average. The p-value for the interaction term (vote_average:vote_count) is 2.33 × 10 , which is highly significant ($p < 0.05$). This provides extremely strong evidence that the relationship between vote count and revenue is indeed influenced by vote average. Therefore, we conclude there is a very strong interaction effect between these two variables.

term	estimate	std.error	statistic	p.value
(Intercept)	10071337.816	2855919.960	3.526	0.000
vote_average	-1527973.810	467221.817	-3.270	0.001
vote_count	102244.575	4248.822	24.064	0.000
vote_average:vote_count	-8149.997	573.312	-14.216	0.000

Methodology

With the EDA analysis done above, we decided to first fit Models 1-6 with individual predictors, including the average rating, runtime length, adult movie status, log-transformed budget, popularity rating, and vote count. After testing a linear regression model with each of the individual predictors to predict the log-transformed budget, since the p-value for each of these individual predictors is less than 0, we can reasonably conclude that there is a significant linear relationship between each of these individual predictors and the log-transformed revenue. Next, we test the significance of these relationships in tandem to predict movie revenue.

Final Model

With the analysis done above, we decided to first fit a Model 1 with the significant individual predictors (average rating, runtime length, adult movie status, log-transformed budget, popularity rating, and vote count). Then, we want to compare to Model 2 where we modify the model to account for potential interaction effects that we explored in EDA.

Model 1:

term	estimate	std.error	statistic	p.value
(Intercept)	0.594	0.070	8.477	0.000
vote_average	0.076	0.009	8.412	0.000
runtime	0.005	0.001	9.279	0.000
adultTRUE	0.832	0.248	3.359	0.001
log_budget	0.880	0.006	148.620	0.000
popularity	0.001	0.000	3.620	0.000
vote_count	0.000	0.000	22.435	0.000

Metric	Value
R-squared	0.850
RMSE	2.149
AIC	38355.357
BIC	38411.997

Model 2:

term	estimate	std.error	statistic	p.value
(Intercept)	0.628	0.070	8.991	0
vote_average	0.079	0.009	8.839	0
runtime	0.005	0.001	10.160	0
adultTRUE	0.860	0.246	3.500	0
log_budget	0.864	0.006	142.106	0
popularity	0.035	0.006	6.257	0
vote_count	0.001	0.000	12.742	0
log_budget:popularity	-0.002	0.000	-6.031	0
vote_average:vote_count	0.000	0.000	-10.321	0

Metric	Value
R-squared	0.852
RMSE	2.132
AIC	38219.308
BIC	38290.108

Model 2 outperforms Model 1 on both key metrics. Model 2 has a slightly lower RMSE, indicating better predictive accuracy, and a higher adjusted R-squared value, indicating a better fit and more variance explained by the model. Furthermore, since Model 2 has slightly lower AIC and BIC values, we select Model 2 as the better model. However, looking at the VIF values for the predictors in Model 2, popularity and the interaction term between the log-transformed budget and popularity have extremely high VIF values. Furthermore, vote_count and the interaction term between the vote average and vote count have extremely high VIF values, thus indicating a multicollinearity issue such that these predictors provide redundant information or they are highly linear dependent/related. Thus, to address this issue, we mean center popularity and log-transformed budget, as well as mean center vote average to transform these predictors and reduce the VIF to create Model 3.

Model 3:

term	estimate	std.error	statistic	p.value
(Intercept)	12.660	0.059	213.843	0
vote_average_mc	0.079	0.009	8.839	0
runtime	0.005	0.001	10.160	0
adultTRUE	0.860	0.246	3.500	0
log_budget_mc	0.833	0.008	104.950	0
popularity_mc	0.010	0.002	6.742	0
vote_count	0.000	0.000	18.607	0
log_budget_mc:popularity_mc	-0.002	0.000	-6.031	0
vote_average_mc:vote_count	0.000	0.000	-10.321	0

Metric	Value
R-squared	0.852
RMSE	2.132
AIC	38219.308
BIC	38290.108

Model 3 performs the same as Model 2 across all key metrics, including RMSE, R-squared, and AIC/BIC. However, looking at the VIF values for the predictors in Model 3, popularity and the interaction term between the log-transformed budget and popularity, as well as vote count and the interaction term between the vote average vote count, have significantly lower VIF values. Although the VIF values still exceed the threshold of 10 for popularity_mc and log_budget_mc:popularity_mc, indicating potential multicollinearity concern, it's crucial to understand the context behind the interaction term, as it is both theoretically justified and statistically significant, improving the model's predictive power overall. Furthermore, although the regression coefficients for vote_count and vote_average_mc:vote_count appear to be zero, since their effect is statistically significant, as indicated by the p-value, these predictors are likely impactful in a larger magnitude when scaled, and thus isn't accurately represented by the following model output.

Furthermore, we conduct a drop-in deviance test to determine whether the interaction effects we examined above are statistically significant. For both the interaction between popularity and log-transformed budget and vote average and vote count, since the p-value is less than 0 for both interactions, it is likely that these interaction effects are significant in predicting the log-transformed revenue. This decision to include the interaction effects in the final model is supported by the fact that in Model 3, the p-value for these interaction effects are below 0.

Interaction	G_Statistic	df	p_value
log_budget × popularity	34.06	1	0
vote_average × vote_count	103.71	1	0

Thus, the final model is reflected by Model 3.

Results

Our final model (Model 3) provides valuable insights into which factors most strongly influence a movie's box office revenue on IMDb. This model included the following predictors: average user rating, movie runtime, adult movie status, mean-centered log-transformed budget, mean-centered popularity rating, total number of movie ratings, an interaction term between mean-centered log-transformed budget and mean-centered popularity, and an interaction term between mean-centered average user rating and total number of movie ratings.

The model explains about 85.2% of the variability in log-transformed revenue, as reflected by the adjusted R-squared value of 0.852. With an RMSE of 2.132, it shows relatively small prediction errors on the log scale. The intercept 12.66 represents an expected revenue of approximately 315,000 USD when all predictors are zero, though this scenario has no practical interpretation since zero values aren't meaningful for all predictors.

Furthermore, several predictors emerge as statistically significant contributors in explaining revenue:

1. Average IMDb Rating: For a one-unit increase above the mean average IMDb rating, the movie revenue is expected to have an 8.2% increase, holding all other variables constant. This supports our hypothesis that higher-rated movies tend to generate more revenue.
2. Runtime: For every additional minute increase in the movie runtime, the movie revenue is expected to have an approximate 0.5% increase, holding all other factors constant. Runtime shows a small, but statistically significant positive effect on movie revenue, indicating that longer movies may be associated with higher revenues.
3. Adult Movie Status: Regarding adult movies, adultTRUE has a coefficient of 0.860, indicating that adult movies are expected to have a movie revenue that is approximately 2.36 times the movie revenue of non-adult movies, on average, holding all else constant. Thus, adult films tend to generate more revenue. Surprisingly, this is the largest positive effect among all predictors, contradicting expectations that adult content might limit audience size.
4. Mean-centered log-transformed budget: for every one unit increase, the movie revenue is expected to be 130% increase, holding all other factors constant. As expected, budget is a significant predictor of revenue, with higher budget films tending to yield higher box office earnings. This aligns with industry trends where more resources generally enable wider distribution, better production value, and stronger marketing.
5. Mean-centered popularity: for every one unit increase in popularity above the mean popularity rating, the movie revenue is expected to have a 1% increase, holding all else constant. Popularity, which proxies online engagement, predicts higher revenue, thus suggesting a link between digital attention and commercial performance.
6. Budget × Popularity: for each additional unit of popularity, the positive effect of a higher budget on revenue diminishes. Specifically, each unit increase in popularity reduces the revenue change associated with budget by approximately 0.2%. This suggests that, among more popular films, the effect on revenue is lower for high-budget movies.
7. Vote Count & Rating × Votes Interaction: Though their coefficients appear as 0.000, these terms are statistically significant, indicating they have small but meaningful effects that may be more apparent when properly scaled.

Overall, the model supports the hypothesis that budget and average rating are key predictors of movie revenue, and additionally highlights the impact of various other predictors and interaction effects.

Conclusion

This project was initiated to determine which production and release factors most significantly influence a movie's total revenue. We addressed this research question through exploratory analysis, linear regression modeling, and model comparison testing. Our results demonstrate that average rating, runtime, adult status, budget, and popularity all serve as strong predictors of revenue, as evidenced by their statistically significant p-values in our regression analysis.

We further identified statistically significant interaction effects between popularity and budget, as well as between average rating and vote count, confirmed through their p-values. After evaluating multiple models, we found that including these interaction terms produced the optimal model with the lowest RMSE (2.13) and highest adjusted R-squared value (0.852). To strengthen our model, we mean-centered the popularity and log-transformed budget variables, along with mean-centering vote average. This approach substantially reduced the variance inflation factor for our interaction terms from approximately 264 to 19. While this represents significant improvement, VIF values exceeding 10 indicate some remaining multicollinearity, which constitutes a limitation of our model. We conducted a drop-in deviance test that confirmed the statistical significance of our interaction effects, supporting their inclusion in the final model. Additional limitations include the exclusion of certain variables such as keywords and production companies due to dataset size constraints, as well as our focus on more recent IMDb data which may limit historical applicability. Our final model achieves an R-squared value of 0.852, explaining 85.2% of the variability in movie revenue. While this represents strong predictive performance, future research could potentially improve upon these results by incorporating the excluded variables mentioned in our limitations.