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.

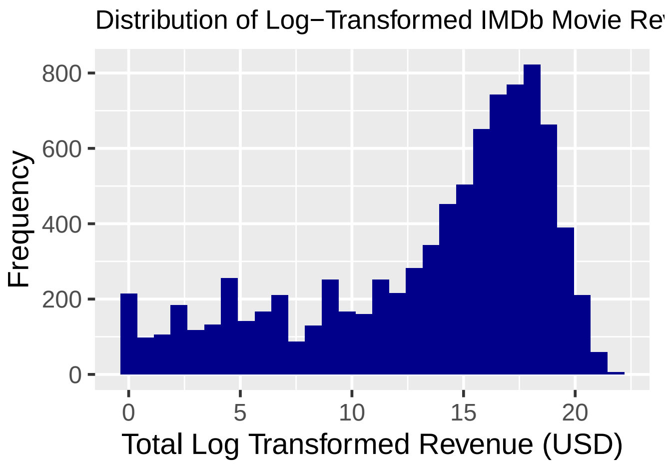
Exploratory Data Analysis

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.

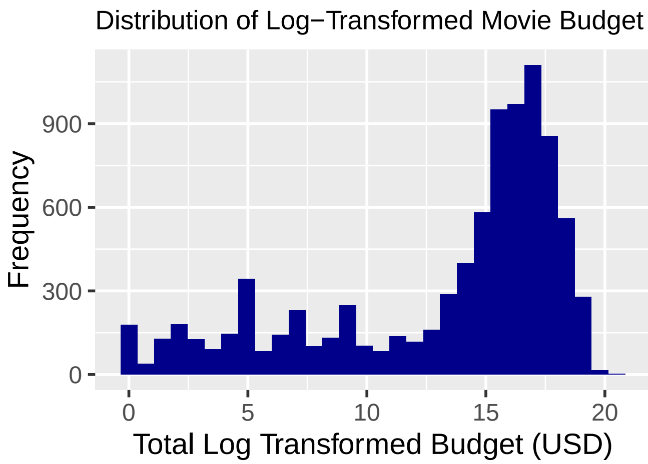
Due to the large size of the original data set, we intentionally drop some characteristics, including `id`, `original_title`, `tagline`, `production_companies`, `overview`, `keyword` from the data set due to their redundancy, irrelevance to the research question, and presence of a significant number 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

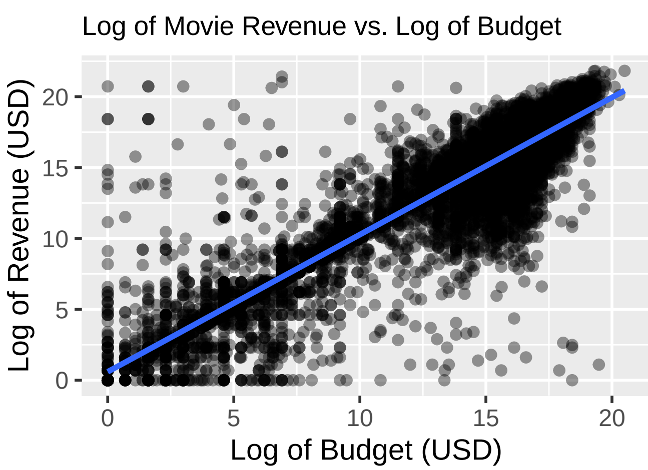


The histogram shows the distribution of the log transformed IMDb movie revenues response variable, which is left-skewed and unimodal, indicating that on average, movies listed on IMDb tend to generate higher revenues. Furthermore, the distribution of the log transformed IMDb movie revenues has a center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

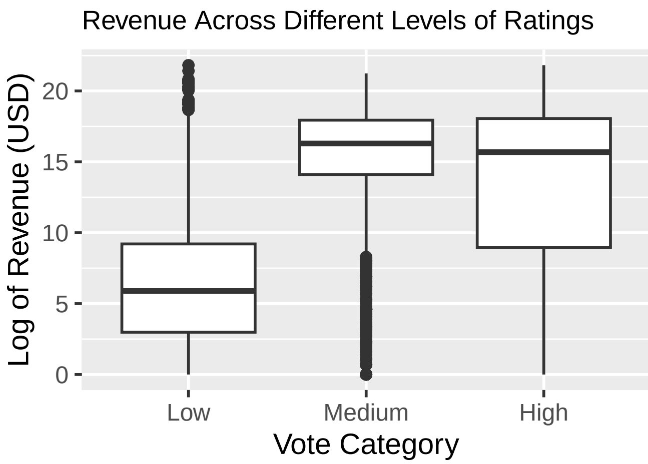


The histogram shows the distribution of the log transformed IMDb movie budget predictor variable, which is left-skewed and unimodal, indicting that although IMDb movie budget varies greatly, on average, movies listed on IMDb tend to have higher budgets. Furthermore, the distribution of the log transformed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

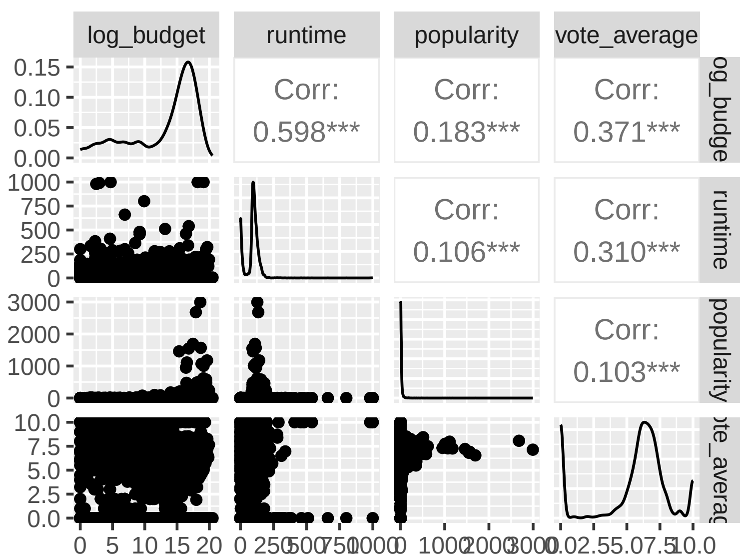
Bivariate EDA



From our visualization plotting the log transformed movie revenue versus the log transformed movie budget, as well as our relatively high R^2 value (0.8366) associated with the linear regression model, we can identify that there is a strong linear correlation between the log transformed movie revenue and log transformed movie budget. The p-value is effectively 0, further exemplifying that there is a linear relationship between the log transformed movie revenue and the log transformed movie budget. This is expected, as high budget movies are often more anticipated and therefore, more people tend to purchase tickets.



A categorical vote_category variable was created from the vote_average variable to reflect the relative average rating category for the movie, creating levels "Low" for ratings 0-3, "Medium" for ratings 3-7, and "High" " for ratings 7-10. From our boxplot, we can see that the median revenue for Medium rated movies (approximately 8,880,000 dollars) is actually 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.

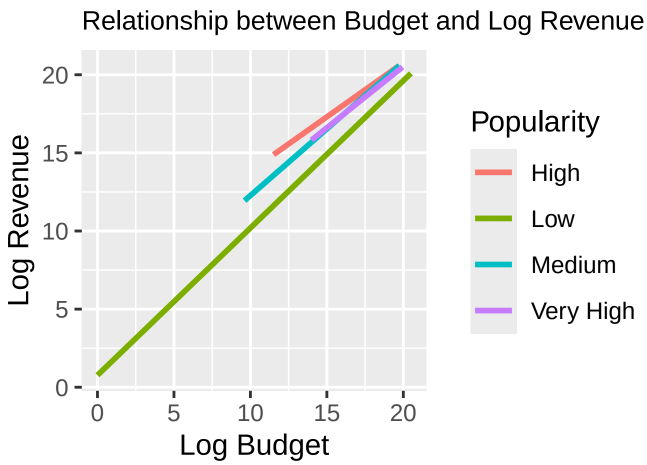


In order to assess potential multicollinearity among predictors, we also created a correlation plot to identify which predictors have high correlations. We can see the highest correlation existing between runtime and budget and vote average and budget. Because of these correlations, we will take special 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" with the following thresholds:

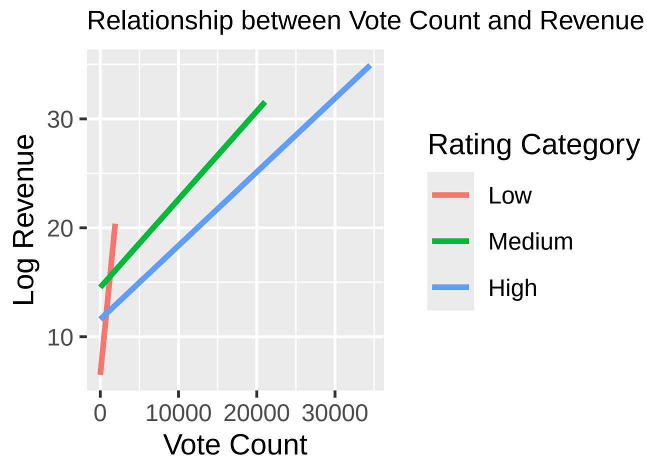
Category	**Popularity Range**
Low	0–30
Medium	30–90
High	90–150
Very High	150+



From this visualization, we can see that there is a slight interaction effect between budget and popularity, since although the relationship between budget and revenue is generally positive and linear, as we noted in our univariate visualizations, its strength is influenced by popularity of the film. We can see that the slope tends to be higher for more popular films, ie, popular movies yield a stronger positive return on budget compared to less popular ones.

Budget and popularity are likely to have interaction effects because higher-budget movies often receive more marketing, leading to increased visibility and higher popularity. However, the p-value for the interaction term (budget:popularity) is 0, indicating that the effect is statistically significant at the 0.05 level, meaning budget's impact on revenue does significantly change based on popularity.

Vote Average and Vote Count



From the visualization, it seems as though there is a strong interaction effect between vote count and vote average (ratings). We observed an interaction between vote count and rating category in predicting revenue. As shown in the plot, the relationship between vote count and revenue differs across rating groups. Movies with higher ratings generally exhibit a stronger positive relationship between vote count and revenue, compared to lower-rated movies. This suggests that ratings modify the impact of audience engagement (vote count) on financial success of a movie.

Since movies with higher vote counts often also have higher vote averages, there could be a multiplicative effect on revenue. If an interaction is present, it means that vote count alone does not fully explain revenue — its impact depends on the vote average. And, the p-value for the interaction term (vote_average:vote_count) is 2.331652e-45, indicating that the effect is extremely significant at the 0.05 level, meaning there is very strong evidence that the relationship between vote count and revenue depends on the vote average. Thus, there is a very strong interaction effect between these two variables.

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.

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

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

Model 2 outperforms Model 1 on both key metrics. Model 2 has a slightly lower RMSE (2.132 for Model 2 versus 2.149 for Model 1), indicating better predictive accuracy, and a higher adjusted R-squared value (0.852 for Model 2 versus 0.850 for Model 1), 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 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.

Thus, the final model is reflected by 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

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 approximately 85.2% of the variability in log-transformed revenue, as reflected by the adjusted R-squared value of 0.852. The RMSE of the model was approximately 2.132, indicating relatively low residual error on a log scale. The intercept 12.66 represents an expected revenue of approximately 315,000 USD when all predictors are zero. However, since it's not feasible for all predictors to be zero in a real-world context, the intercept doesn't represent a meaningful value in this scenario. Furthermore, several predictors emerge as statistically significant contributors in explaining revenue:

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.

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.

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, the coefficient for adult movies is the greatest among all other predictors, and thus most greatly positively influencing the movie reveue. Contrary to expectations that adult movie status would narrow the movie audience, adult movie status, on average, increases movie revenue.

For the 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.

For the 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.

For the interaction between budget and 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.

Although the predictors for vote count and the interaction term between average user rating and vote count have coefficients of 0.000 (small, near zero coefficients), these predictors are statistically significant, thus suggesting that these predictors have a small, but meaningful effect on movie revenue.

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

We began this project with the intention to answer our research question asking what production and release factors have the greatest impact on a movie's total revenue. We ultimately achieved this through an exploratory analysis, fitting linear regression models, and conducting model comparison tests. Our analysis indicates that average rating, runtime, adult status, budget and popularity are all strong predictors of revenue, confirmed by low p-values in our linear regression model statistics. We also proved that there was a statistically significant interaction existing between the variables of popularity and budget and between the variables of average rating and vote count, also using p-values. After fitting several models, we found that a model including the interaction effect between budget and popularity and interaction effect between vote average and vote count yielded the lowest RMSE value and the highest adjusted R-squared value (2.13, 0.852). To add to the strength of our model, we dramatically reduce the variance inflation factor among our interaction terms by mean centering popularity and log-transformed budget, as well as mean centering vote average. This effectively reduced our VIF value from approximately 264 to 19. Despite this success, a VIF over 10 is still a risk of multicollinearity, and so this is an inherent limitation of our model. Lastly, we conducted a drop-in deviance test on our final model and determined that the interaction effects were significant, which is important to include in our model. Limitations to our model include our risk of multicollinearity, which we were able to diminish the risk of but not entirely eliminate. We also were forced to drop several columns from our initial data set including "keyword" and "productions companies", due to the large size of our dataset, possibly eliminating other relevant predictors to predict movie revenue. To further shrink our dataset, we also limited production years, which would prove to be a limitation as it only considers more recent IMDb movies. Our final model yielded an R-squared value of 0.852, indicating that it explains 85.2% variability in revenue of a given film. This is relatively high, but future work could potentially increase this further by including dropped columns discussed in limitations.