Analysis of IMDb Movie Revenue

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

2025-03-20

Introduction:

This project aims to examine key factors influencing IMDb movie revenues to understand what drives box office success. Revenue, the most widely used measure of success in the film industry, reflects audience demand and commercial viability. Prior research has highlighted various influences, such as star power, genre, and marketing, but the relative importance of these factors remains debated.

One 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 deepen the understanding of revenue drivers and improve predictive models for box office performance.

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.

Exploratory Data Analysis:

To get more indepth into our data set, the data set we used collects 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. It measures the following **15 characteristics** per movie:

`title`: The name of the movie.

`vote_average`: The average rating the movie has received from users (on a scale, typically from 0 to 10).

`vote_count`: The total number of votes or ratings submitted for the movie.

`status`: The current state of the movie (e.g., "Released," "Post-Production").

`release_date`: The date when the movie was officially released.

`revenue`: The total earnings the movie made (usually in USD).

`runtime`: The duration of the movie in minutes.

`adult`: Indicates whether the movie is classified as adult content (e.g., "True" or "False").

`budget`: The total cost of producing the movie (usually in USD).

`imdb_id`: The unique identifier for the movie on IMDb (Internet Movie Database).

`original_language`: The language in which the movie was originally produced (e.g., "en" for English).

`popularity`: A metric indicating how popular the movie is (typically based on views, searches, or ratings).

`genres`: The categories or genres the movie belongs to (e.g., Action, Comedy, Drama).

`production_countries`: The countries where the movie was produced.

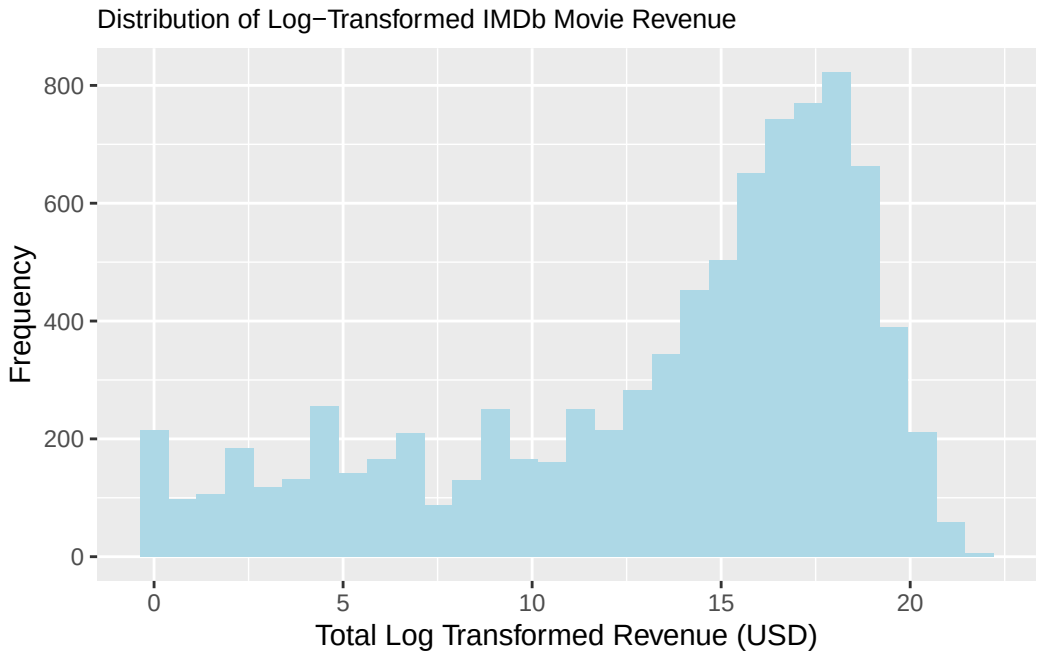
`spoken_languages`: The languages spoken in the movie.

Due to the large size of the original data set and for the purpose of uploading this to our repository, 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.

To obtain informative interpretations of the distribution of the revenue response variable, we analyze IMDb movies released from 2000 and beyond that have a revenue greater than 0 (non-missing values in this dataset). Furthermore, we log transform revenue, as the revenue generated across the IMDb movies is heavily right-skewed. (hellen could you include this in the description of the variables section, maybe write a short blurb for log transformation of log transformation for budget too).

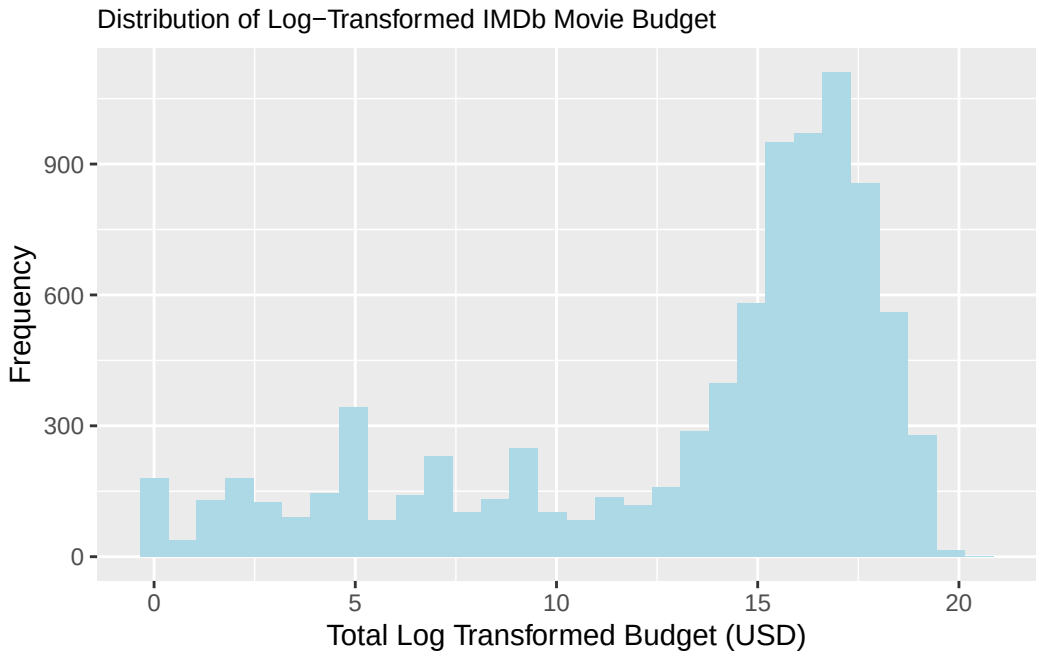
Univariate EDA

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   9.933  15.323  13.387  17.646  21.822
```

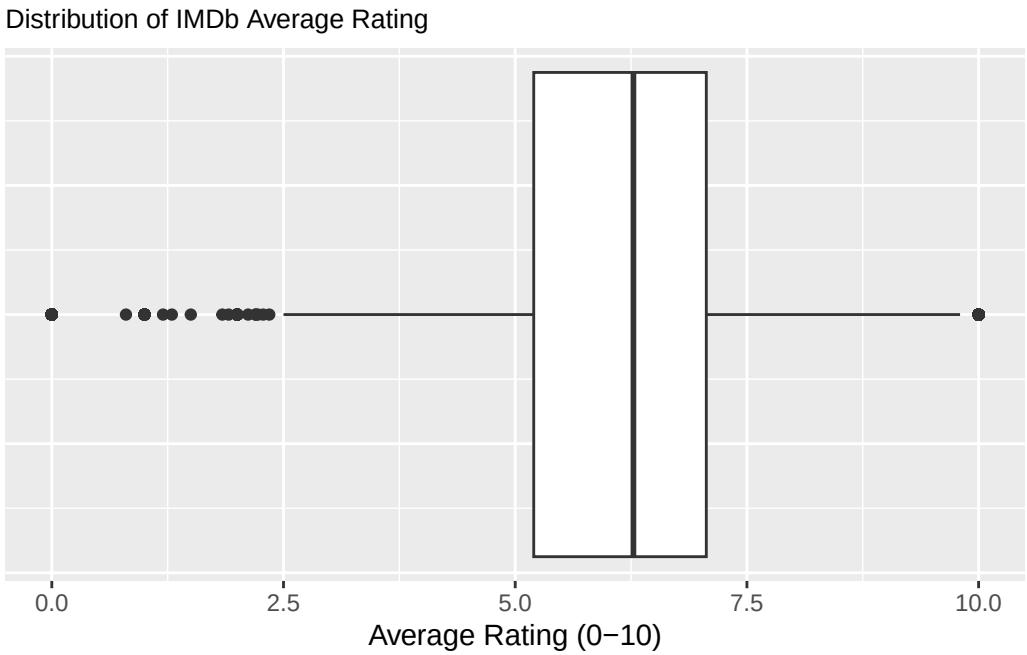


The histogram shows the distribution of the log transformed IMDb movie revenues response variable, which is left-skewed and unimodal, indicating that on average, movies listed on IMDb tend to generate higher revenues. Furthermore, the distribution of the log transformed IMDb movie revenues has a center of approximately 1,800,000 US dollars, described by the median, and a spread of approximately 18,600,000 US dollars, described by the interquartile range. The majority of IMDb movies generate between approximately 24,100 and 18,600,000 US dollars.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   9.629  15.425  13.222  17.034  20.500
```

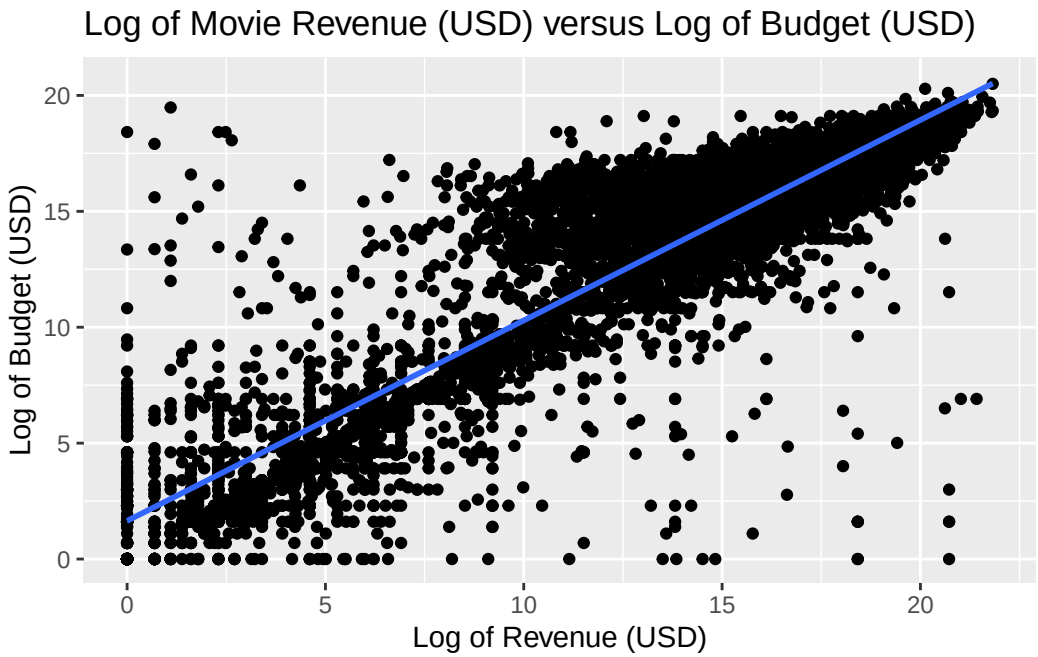


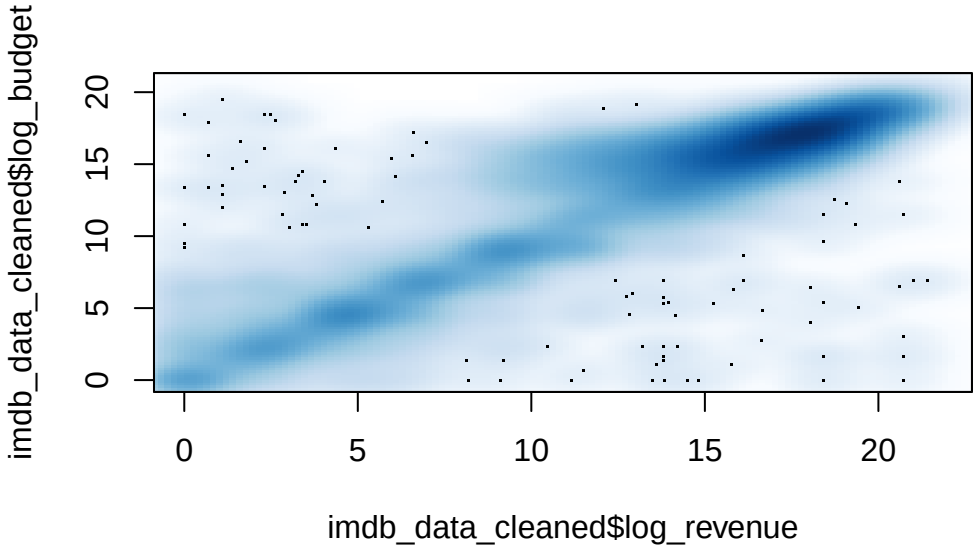
The histogram shows the distribution of the log transformed IMDb movie budget predictor variable, which is left-skewed and unimodal, indicting that although IMDb movie budget varies greatly, on average, movies listed on IMDb tend to have higher budgets. Furthermore, the distribution of the log transformed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



The majority of IMDb movies' average ratings are concentrated between a rating of 4.0 to 8.0. A few IMDb movies have a low rating (average rating below 2.5), contributing to the left-skewness of the distribution of IMDb movies' average rating predictor variable. Furthermore, a few IMDb movies have an extremely high rating (average rating of 10.0).

Bivariate EDA



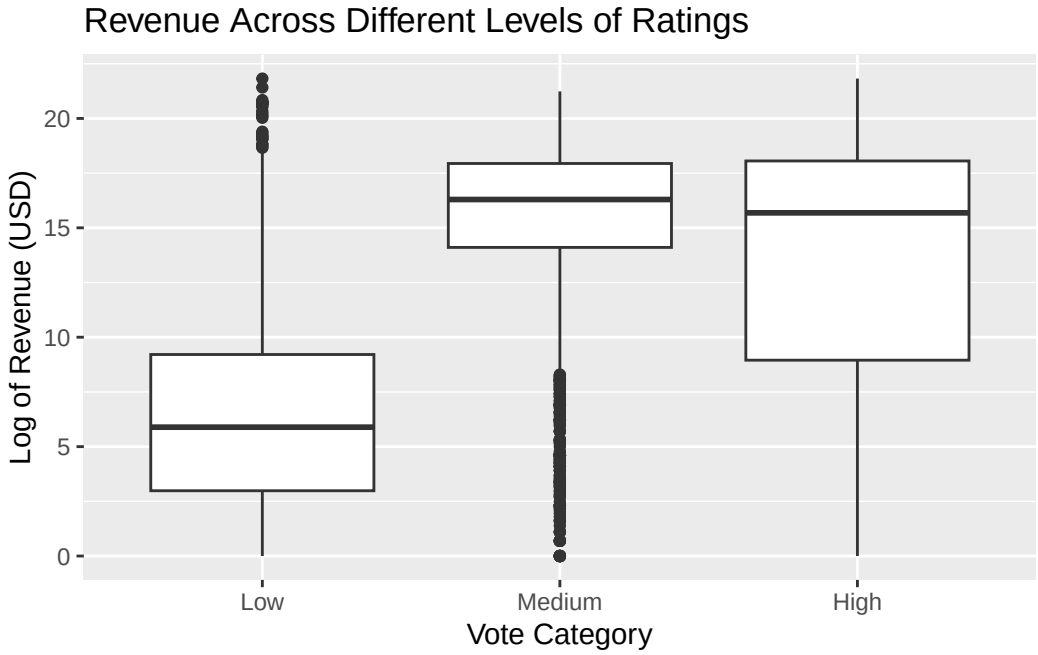



```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
```

```
1 (Intercept)    0.595   0.0649       9.16 6.08e-20
2 log_budget     0.967   0.00456    212.   0
```

```
[1] 0.8365862
```

From our visualization plotting the log of movie revenue versus the log of movie budget as well as our relativley high r-squared value associated with a linear regression model, we can identify that there is a strong linear correlation between log of movie revenue and log of movie budget. Our p-value is effectively 0, further exemplifying that there is a linera relationship. We expected this, as high budget movies are often more anticipated and therefore more people buy tickets. Because our data set is so large, we decied to also display this relationship using a smoother scatter visualization, which illustrates point density seen in our scatter plot.



In this visualization, we transformed the vote_average variable into a categorical variable taking 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 log revenue for Medium rated movies (around 16) is actually slightly higher than those with High ratings (15.5). The median log revenue for Low rated movies is significantly lower (6). The IQR of movies with high ratings is also much larger than for Medium and Low rated movies.

Potential Interaction Effects

To check interaction effects between some of our predictors to ensure we don't have issues with multicollinearity in our models, we checked interaction effects between some variables that are likely to have some collinearity. For 3 pairs of predictor variables, we fit an interaction model with revenue as the response variable to see if there is a statistically significant impact of the interaction term on the model.

Budget and Popularity

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 0.48, indicating that the effect is not statistically significant at the 0.05 level, meaning budget's impact on revenue does not significantly change based on popularity.

term	estimate	std.error	statistic	p.value
(Intercept)	-4711433.669	1415969.481	-3.327	0.001
budget	2.975	0.031	95.869	0.000
popularity	84395.635	28855.053	2.925	0.003
budget:popularity	0.000	0.000	0.706	0.480

Vote Average and Vote Count

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. So, 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

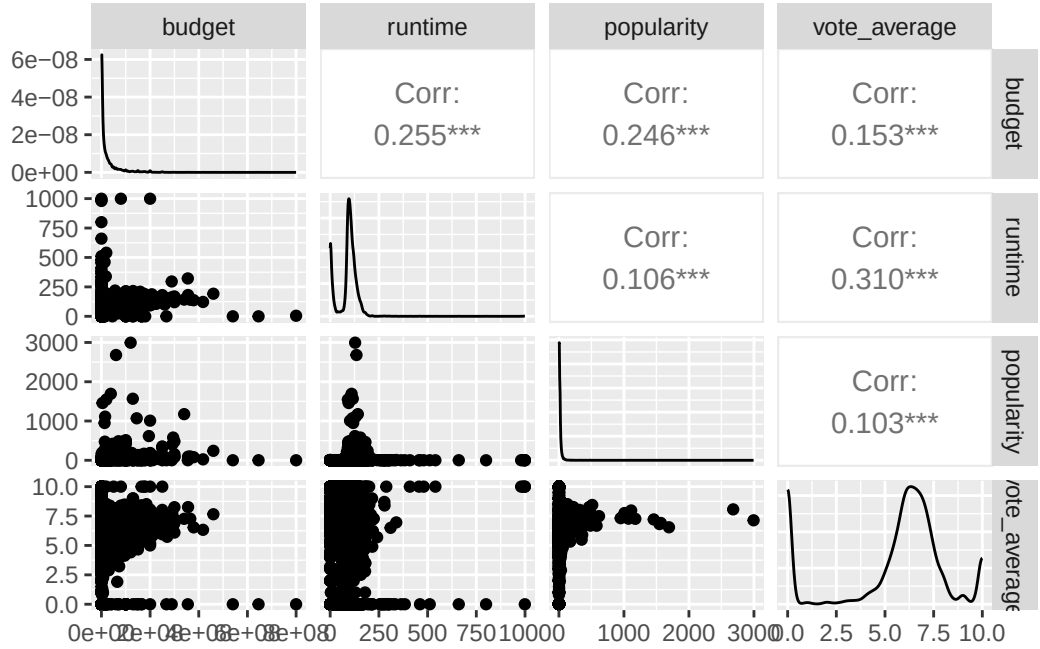
Action Movies and Runtime

Action movies have longer runtimes typically compared to other genres due to storytelling, long fight sequences, and action scenes. We thought that there may be an interaction effect between being an action movie and runtime when predicting revenue. Longer runtimes could allow for larger audiences and increasing box office earnings. However, excessively long movies

may also reduce viewers due to time constraints. However, from the p-value of the interaction term, which is 2.991935e-15, we can see that this term is quite significant at the 0.05 level. This implies that there is some interaction between a movie having the 'action' genre and longer runtimes, or the effect of runtime on revenue is different for action movies compared to non-action movies. This could suggest that longer runtimes contribute positively to revenue for action films, possibly because audiences expect more elaborate fight sequences and storytelling in the genre, making them more willing to engage with longer films.

term	estimate	std.error	statistic	p.value
(Intercept)	13916205.3	3864760.77	3.601	0.000
action_movie	-892455.3	9145331.04	-0.098	0.922
runtime	372326.6	38529.42	9.663	0.000
action_movie:runtime	630145.9	79709.91	7.905	0.000

We also explored the correlation between all possible combinations of the variables budget, runtime, popularity and vote_average, to see how they interact with each other. The results indicate that the correlations between these variables are relatively low, with none exceeding 0.310. This suggests that there is no strong linear relationship between them, reducing the likelihood of multicollinearity issues when including these variables in a predictive model. As a result, we can confidently incorporate them into our analysis.



! Important

Before you submit, make sure your code chunks are turned off with `echo: false` and there are no warnings or messages with `warning: false` and `message: false` in the YAML.