IMDb Movie Success: An Analysis of Movie Release Factors and Box 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 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. 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 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.

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.

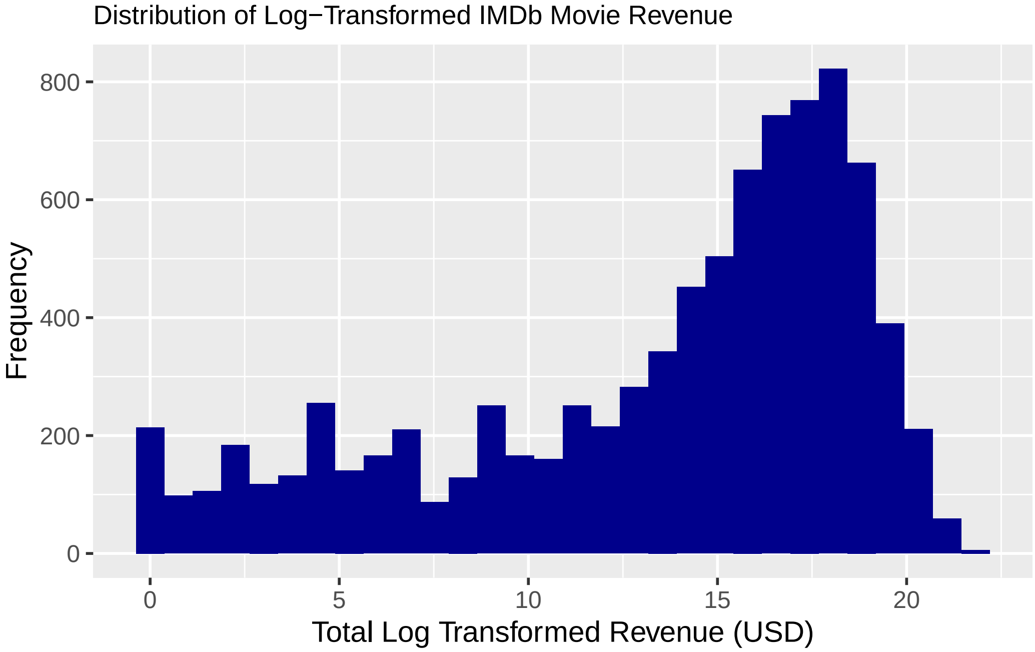
Due to the large size of the original data set and for the purpose of uploading this to our repository, we intentionally drop some characteristics, including `id`, `original_title`, `tagline`, `production_companies`, `overview`, `keyword` from the data set due to their redundancy, irrelevance to the research question and presence of a significant number of null values.

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. 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. Furthermore, a release_years_since_2000 variable was created that calculates the number of years of a movie’s release since 2000, excluding all movies that were released pre-2000.

We chose to examine movie revenue as our response variable.

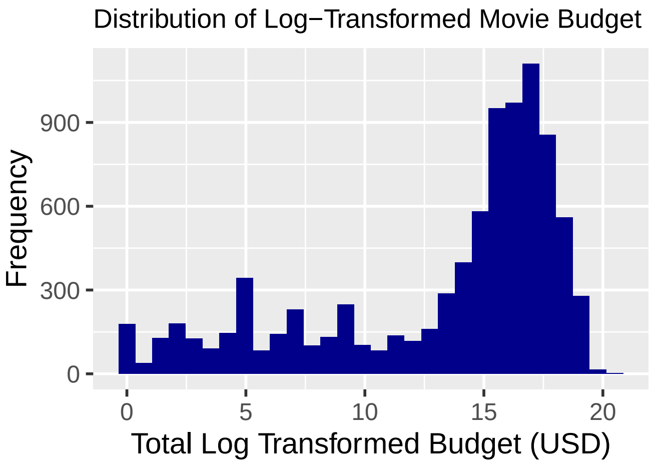
Univariate EDA

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

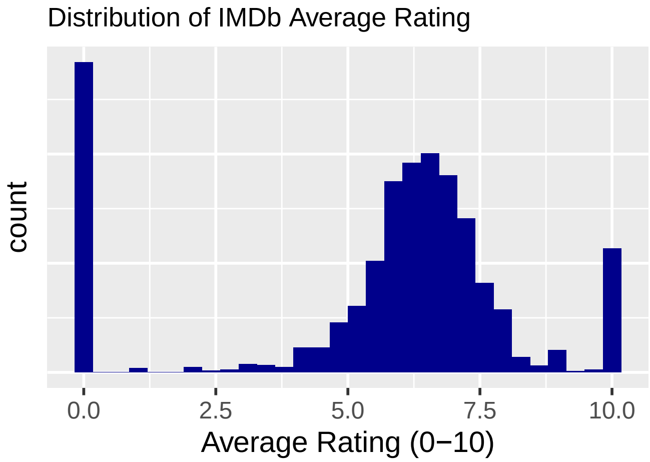


The histogram shows the distribution of the log transformed IMDb movie revenues response variable, which is left-skewed and unimodal, indicating that on average, movies listed on IMDb tend to generate higher revenues. Furthermore, the distribution of the log transformed IMDb movie revenues has a center 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 
```

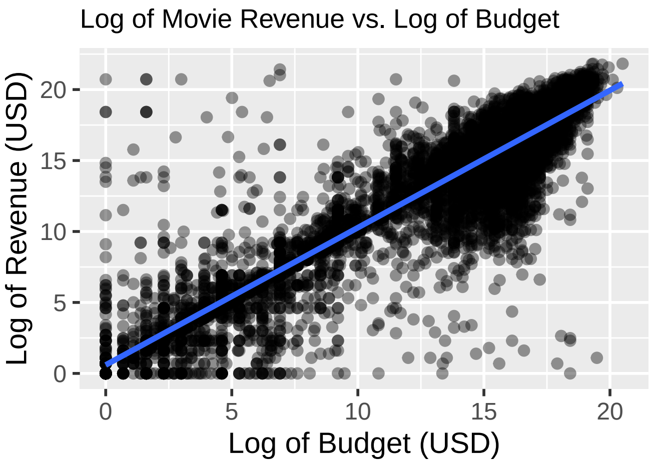


The histogram shows the distribution of the log transformed IMDb movie budget predictor variable, which is left-skewed and unimodal, indicting that although IMDb movie budget varies greatly, on average, movies listed on IMDb tend to have higher budgets. Furthermore, the distribution of the log transformed 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).

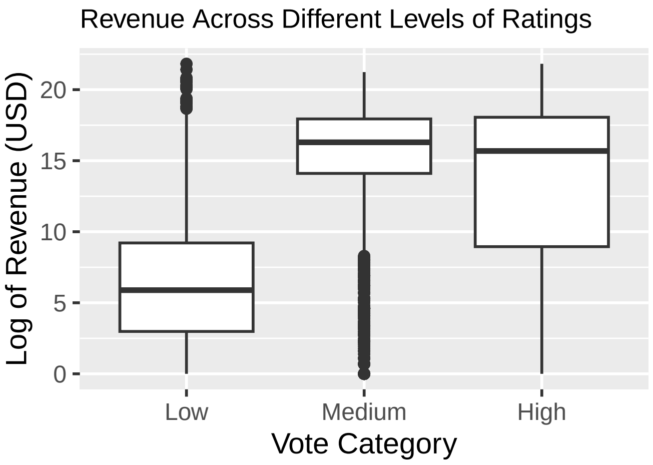
Bivariate EDA



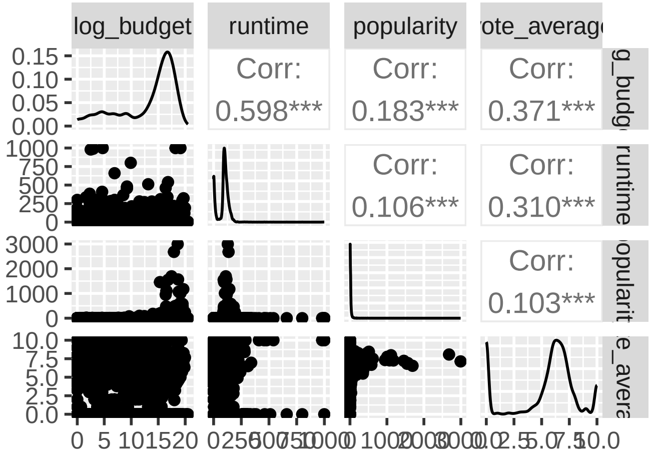
```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    0.595   0.0649       9.16 6.08e-20
2 log_budget     0.967   0.00456    212.   0
```

```
[1] 0.8365862
```

From our visualization plotting the log transformed movie revenue versus the log transformed movie budget, as well as our relatively high R^2 value 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. Because our data set is so large, this relationship is also displayed using a smoother scatter visualization, which illustrates point density seen in our scatter plot.



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.

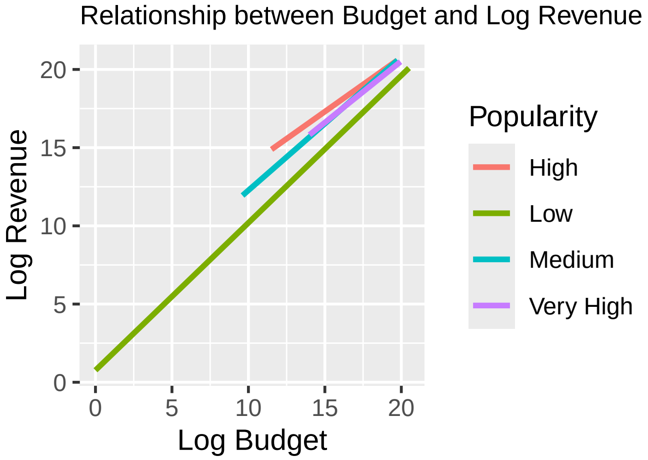


In order to assess potential multicollinearity among predictors, we also created a correlation plot to identify which predictors have high correlations. We can see the highest correlation existing between runtime and budget and vote average and budget. Because of these correlations, we will take special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

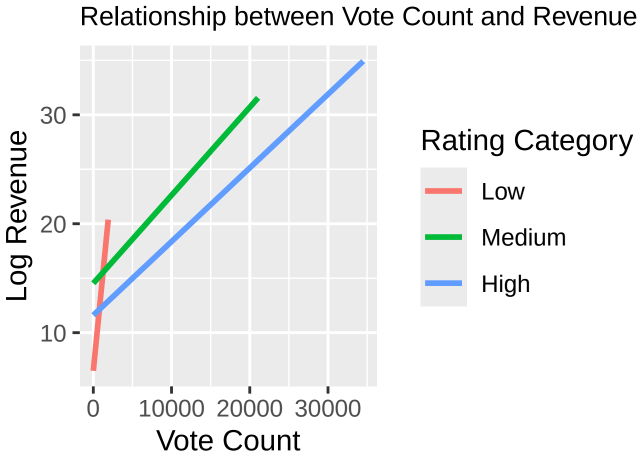


From this visualization, we can see that there is a slight interaction effect between budget and popularity, since although the relationship between budget and revenue is generally positive and linear, as we noted in our univariate visualizations, its strength is influenced by popularity of the film. 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.

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, 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. 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

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.

term	df.residual	rss	df	sumsq	p.value
log_revenue ~ vote_average_mc + runtime + adult + log_budget_mc + popularity_mc + vote_count	8771	40529.73	NA	NA	NA
log_revenue ~ vote_average_mc + runtime + adult + log_budget_mc + popularity_mc + vote_count + log_budget_mc * popularity_mc	8770	40372.77	1	156.963	0

term	df.residual	rss	df	sumsq	p.value
log_revenue ~ vote_average_mc + runtime + adult + log_budget_mc + popularity_mc + vote_count	8771	40529.73	NA	NA	NA
log_revenue ~ vote_average_mc + runtime + adult + log_budget_mc + popularity_mc + vote_count + vote_average_mc * vote_count	8770	40053.69	1	476.035	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 (vote_average), movie runtime, adult movie status, log-transformed budget (mean-centered), popularity (mean-centered), and an interaction term between the centered log-budget and popularity.

The model explains approximately 66.9% of the variability in log-transformed revenue, as reflected by the adjusted R-squared value of 0.669. The RMSE of the model was approximately 1.765, indicating relatively low residual error on a log scale.

Several predictors emerged as statistically significant contributors:

- Vote Average: A one-unit increase in average IMDb rating is associated with an increase in log-revenue, holding other factors constant. This supports our hypothesis that higher-rated movies tend to generate more revenue.
- Log-Budget (Mean-Centered): 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.
- Popularity (Mean-Centered): Popularity, which proxies online engagement, also significantly predicts higher revenue. This suggests a strong link between digital attention and commercial performance.
- Interaction between Log-Budget and Popularity: The positive and significant interaction term indicates that the effect of popularity on revenue is stronger for high-budget movies. In other words, popularity boosts revenue more when a movie also has a large budget — possibly because high-budget movies can better capitalize on attention through marketing and distribution.

- Runtime: Runtime also showed a small but significant positive effect, suggesting that longer movies may be associated with higher revenues, possibly due to being perceived as more substantive or event-like.
- Adult Status: While statistically significant, adult status had a relatively small negative effect on revenue, implying that non-adult films tend to generate more revenue — likely due to broader audience appeal.

Overall, the model supports our hypothesis that budget, rating, and popularity are key predictors of movie revenue, and it additionally highlights the importance of interaction effects, particularly between budget and popularity.

Discussion

From our analysis, we determined that 6 of the predictors tested are statistically significant in their influence on the revenue of an IMDB movie. However, we were especially interested in the impact of predictors "popularity" and its relationship to "budget". As we expected, there exists an interaction effect between the two predictors, as the popularity of a movie has a greater influence over the revenue when budget is higher. In practice, we could potentially account this to marketing strategies associated with high budget films. In addition to our two predictors of speical interest, we also determined significance associated with "runtime" and "adult status". We chose to hold off further analysis on these factors because of their relatively small effect on revenue. One aspect of interest among these predictors, however, is the fact that "adult stats", a binary variable indicating whether or not a movie contains adult content, has a slight negative effect on revenue. This was the only negative effect we found. We hypothesize that this may be due to the limited group of viewers deemed as "appropriate" for these mature films.

In regards of our research question, inquiring about what factors most strongly influence a movie's box office revenue, our analysis indicates that runtime, adult status, budget and popularity are all predictors revenue, with an interaction between the variables as popularity and budget well. This was determined by the fact that in fitting linear models, we oberverd low p-values. We then fit several models, and found that our model including an interaction between budget and popularity outperformed that without. This was determined using metrics such as RMSE and adjusted r-squared. However, from this model we also detected potential multicollinearity between the two predictors indicated by high VIF values that exceeded the threshold of 10. This aspect of our analysis will require futher analysis, and is certianly a limitation to our model. This could potentially be improved by finding a metric to combine the two predictors, because we feel crucial information would be lost if we were to simply remove one.

For our future anlysis, it would be ideal to derive a model with a higher r-squared, as our model still leaves much to be accounted for. We hope to test other predictors, or potentially created new metrics from the existing ones to try and understand the trend in revnue better.

In this section you'll include a summary of what you have learned about your research question along with statistical arguments supporting your conclusions. In addition, discuss the limitations of your analysis and provide suggestions on ways the analysis could be improved. Any potential issues pertaining to the reliability and validity of your data and appropriateness of the statistical analysis should also be discussed here. Lastly, this section will include ideas for future work.

Conclusion

❗ 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 theAML.