Lights, Camera, Action! - Predicting Gross Revenue of Movies

Rated R: Carlie Scheer, Christina Lee, Jerry Lin, Samir Travers

2024-10-27

Introduction

Movies have been at the heart of pop culture for nearly a century. And while movies have changed a lot over the years, varying in length, theme, due to technological advances, and more, one thing has never changed–movies have remained as a dominant part of the entertainment industry. Yet, ever since the COVID-19 pandemic, the North American box office for movies has not recovered–it seems that people just aren't going to movies as much as they used to. As mentioned in (Box Office Mojo 2024), domestic box office gross revenue in 2019 reached nearly $11.4 billion, while the yearly total from 2023 topped off at $8.9 billion after a very slow two years in 2020 and 2021. With moviegoers back on the rise, the film industry and box offices are hopeful to end the decade on a high note and return back to their original state. As such, we are interested in investigating the trends in gross revenue for top movies and how various elements impact the amount of money that a movie makes. More formally, what are the key factors and metrics that most significantly influence a movie's gross revenue?

As a group of movie enthusiasts, movie popularity has always been a large source of curiosity. We've seen terrible movies whose popularity we find puzzling, and unknown masterpieces that do not get the attention they deserve. What is it that separates the two? Is it primarily the "goodness" of the movie that determines the money it brings in, or are there other factors that are equally influential? Is the 1.4 billion dollar success of Alvin and the Chipmunks (an objectively pretty bad movie) just an outlier, or reflective of a deep and complex relationship between a movie's characteristics and its gross profit? Through our analysis, we hope to get to the bottom of these questions, and gain a deeper understanding of what truly drives a movie's success or failure in the box office.

We hypothesize movies that have high popularity scores and ratings, substantial budgets, and broad audience appeal will be more likely to generate higher gross revenue. We also hypothesize movies with more than one genre will be more likely to generate higher gross revenue. Movies with a runtime between 1.5-2 hours will be more likely to generate higher gross revenue.

Data description

The source of the data set is taken from TMDb (The Movie Database), which is an online, public database for movies that contains information such as ratings, runtime, cast, director, and box office performance. We discovered this data set on Kaggle, and it is updated daily by user asaniczka, who takes the data from TMDb (Asaniczka 2023).

There were 953629 observations in the original data set, each representing a single film. There were 24 columns in this data set, with some of the most important being the name (an identifier), release year, runtime, genres, rating, votes, and revenue.

Exploratory Data Analysis

We filtered the original data set outside of this project first because the original data set of almost 1 million rows was far too large to be able to run efficiently in RStudio; as such, we filtered for where the movies had more than 100 votes, meaning more than 100 people submitted ratings for the movie. After filtering, there are now 18086 observations, and we were able to import this filtered data set into Rstudio (The file name was "final_vote_filtered_movie_data.csv").

We are going to transform the variables spoken_languages and genres. For spoken_languages, we want to make a categorical variable based on whether more than one language is spoken in the movie or not (True = more than one language, False = just one language). For genres, we want to take a similar approach to see whether the movie is tagged for more than one genre (True = more than one genre, False = just one genre). Additionally, we want to transform the month, date, year format of the release date to just be the year. Lastly, before we begin EDA and further analysis, we are going to remove all 0s in the revenue and budget columns because we are not able to accurately impute the values; we have evidence to suggest that the original author of the data set just put 0s instead of NAs.

Analysis Approach

Variables of interest:

vote_average: The average rating given to the movie by viewers. We expect this to be correlated with the response, as good ratings will encourage more people to go see the movie.

spoken_languages: The number of different languages spoken in the movie. We hypothesize that movies with more languages will appeal to broader audiences, earning greater revenue. We created a new variable from this variable called "multi_language" that categorizes movies with multiple languges as "TRUE" and movies with one language as "FALSE."

budget: The amount of money spent on producing the movie. We expect movies with a larger budget to earn more revenue, as the production company has more resources to put into the movie.

runtime: This is the length of the movie. We think that people may be more willing to see shorter/average length movies, as it is less of a time commitment for them.

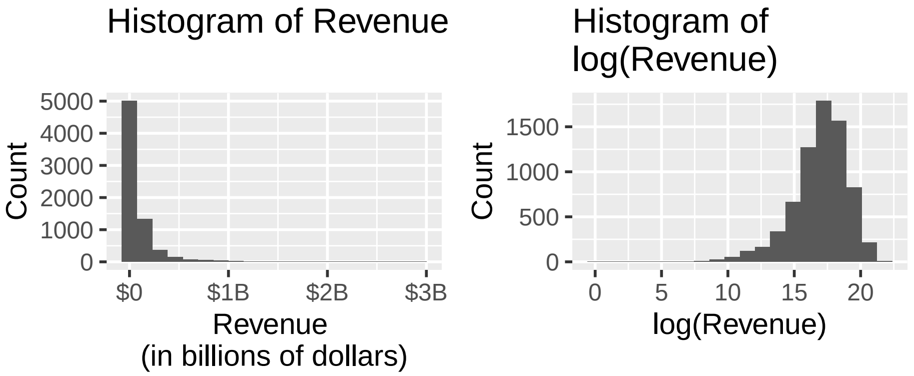
release_date: This indicates the day, month, and year that the movie was released. We expect that the year will greatly impact gross revenue because, for instance, movies released right before or during the pandemic will have very low gross revenue as no one was able to go to the theaters. We created a new variable from this variable called "year" that only accounts for only the year (yyyy) that the movie was released in instead of the original date (yyyy/mm/dd).

genres: The genre or list of genres associated with the movie. We expect movies tagged with more than one genre to appeal to wider audiences, driving up the number of people who go to see the movie and increasing gross revenue. We created a new variable from this variable, "multi_genre", that categorizes multiple genres movies as "TRUE" and single genre movies as "FALSE."

Method of approach:

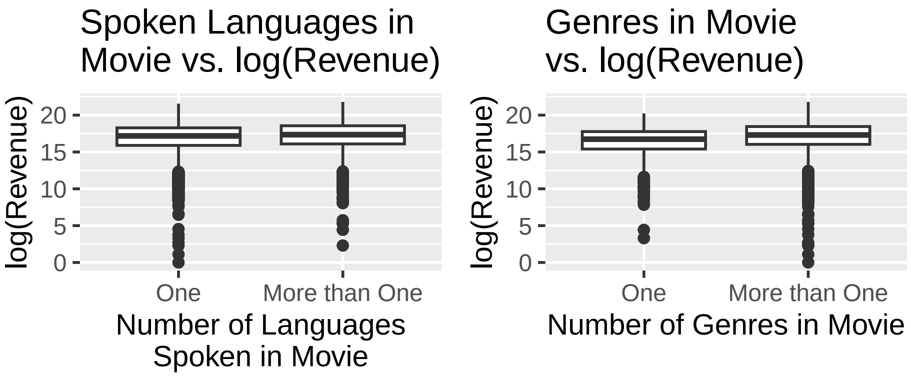
We plan to use multiple linear regression to predict gross revenue, with gross revenue log-transformed as we can see from our EDA (below) that revenue is very skewed in its original form.

Univariate EDA



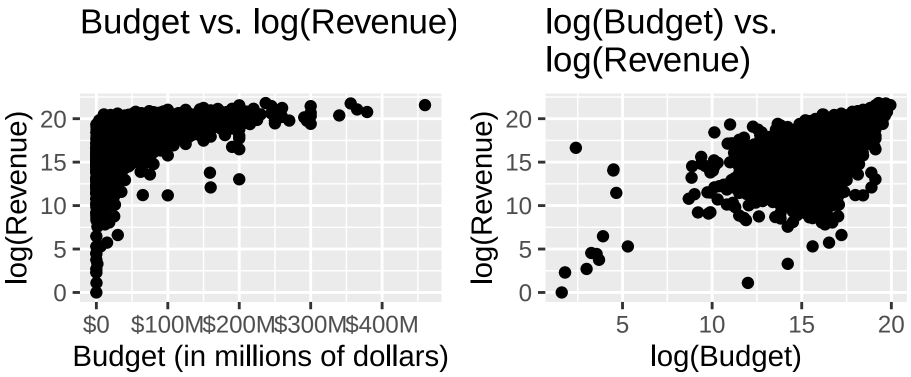
The distribution of revenue is strongly skewed to the right. The boxplot suggests that we have a significant amount of outliers. The median is $30,099,904 and our interquartile range is $86,123,799. Following a log transformation, our response variable, logged revenue, has a slight left skew but is unimodal and very roughly normal. It has a significant number of outliers on the left side, and log(Revenue) has a median of 17.22 and an IQR of 2.43 (Appendix Figure 1).

Bivariate EDA



Based on the box plots for the number of spoken languages in a movie, it seems that whether one language was spoken or multiple were does not have much of an impact on the logged revenue. This is evident based on the overlapping of the two box plots.

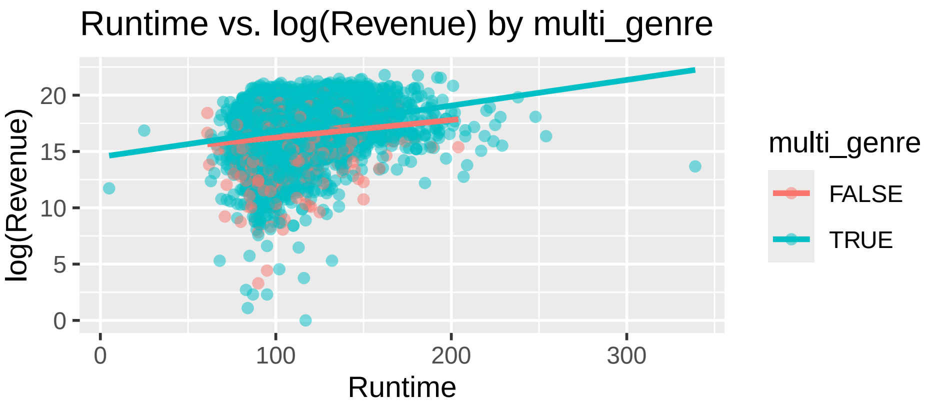
Based on the box plots for the number of genres a movie is tagged for, it also appears that whether the movie is grouped into one genre or multiple genres does not have much of an impact on the logged revenue. Again, this is evident based on the overlapping of the two box plots.



Because of the extreme curvature present in the graph of budget vs. log(revenue), it makes sense to also do a log transformation on Budget in order to get the graph log(budget) vs. log(revenue) which is seen on the bottom right. The graph of log(budget) vs. log(revenue) has far less curvature and appears to take more of a linear shape than budget vs. log(revenue).

Interaction

The main interaction we want to investigate is between the genre and runtime. It is possible that the effect of movie runtime on log(revenue) differs based on if it is a multi-genre movie or if the movie is only of a single genre.



The slope of each seems to differ slightly, and it seems that the effect of runtime on log(Revenue) is marginally stronger for multi-genre movies. We intend to do further analysis later in this project to investigate these interaction effects as well as other interactions of interest.

We additionally investigated the interaction between genre and vote average, however, the slopes of the lines appear to be very similar (Appendix Figure 3) . We will still do further analysis to ensure that the interaction is not useful in our final model, but we will likely not be using it.

Methodology and Analysis

Base Model

We fit a base model *without* the interaction terms first.

term	estimate	std.error	statistic	p.value
(Intercept)	23.542	2.391	9.844	0.000
vote_average	0.623	0.026	24.217	0.000
log_budget	0.925	0.013	70.568	0.000
runtime	-0.003	0.001	-2.733	0.006
year	-0.013	0.001	-10.543	0.000
multi_genreTRUE	0.035	0.054	0.646	0.518
multi_languageTRUE	-0.056	0.041	-1.375	0.169

Based off of our output for the base model (no interaction terms), it looks as if there is significant statistical evidence to suggest that vote_average, log_budget, runtime, and year all are useful predictors for log(revenue), as the p-values for each of these coefficients are all very close to 0. However, there are potential issues with using the multi_genre and multi_language because their coefficients have p-values on the larger end, suggesting that they may not be useful variables to add to our model.

We were also surprised that the year coefficent has a negative value, as that would imply that as the movie is more recent, the log(revenue) decreases, which is interesting. We are now interested in analyzing whether the added interaction terms are worth adding in the model to use as predictors of the log(revenue).

We tested for multicollinearity by using VIF; all of our predictor variables have a VIF of less than 2. Thus, it seems as if there are no variables that raise issues of multicollinearity (Appendix Figure 4).

Investigating Multi-Genre/Runtime Interaction

term	estimate	std.error	statistic	p.value
(Intercept)	24.771	2.409	10.283	0.000
vote_average	0.628	0.026	24.416	0.000
log_budget	0.926	0.013	70.692	0.000
runtime	-0.013	0.003	-4.689	0.000
year	-0.013	0.001	-10.623	0.000
multi_genreTRUE	-1.198	0.315	-3.796	0.000
multi_languageTRUE	-0.060	0.041	-1.455	0.146
runtime:multi_genreTRUE	0.012	0.003	3.965	0.000

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
7051	17124.30	NA	NA	NA	NA
7050	17086.19	1	38.11	15.725	0

When adding the interaction effect between multi_genre and runtime to our model, our results differ from the base model in that, now, adding multi_genre as a predictor does seem to be useful and, additionally, the interaction between multi_genre and runtime is also statistically significant (both indicated by p-values of nearly 0). Still, we see evidence to suggest that multi_language may not be a useful predictor.

When running a nested F-test, our output shows that the more complex model, which includes the interaction effect between multi_genre and runtime, does a significantly better job of predicting movie revenue than the base model which does not contain it—this is again indicated by a resulting p-value of nearly 0. Additionally, using model comparison diagnostics, we see that the new model (including interaction effect) has a slightly higher adjusted R^2 value of 46.56% as compared to the base model's adjusted R^2 of 46.45% (Appendix Figures 5 and 6). Similarly, the new model's BIC drops by nearly 7 points (from 26356.40 to 26349.53), suggesting an improvement. We are focused on using BIC because it penalizes more harshly for additional parameters than AIC, so it will allow us to find the model which best fits our data.

In effect, we will add the interaction effect between multi_genre and runtime to our final model. We now look to see if the other interaction effect we are interested in, multi_genre and vote_average, will be a useful additional predictor.

Investigating Multi-Genre/Vote Average Interaction

Since we already added the multi_genre and runtime interaction, we now test the model with BOTH interaction terms against the model with ONLY the multi_genre and runtime interaction (Appendix Figure 7).

From the F-test results alone, we have strong evidence to suggest that adding the interaction effect between multi_genre and vote_average does not significantly improve our model—this is indicated by the F-test p-value of 0.577. This proves our hypothesis from the EDA that this interaction term would not be signfiicant. Additionally, both the R^2 value and BIC are worse for the model with the added interaction effect between multi_genre and vote_average, as compared to the model from the previous section without it (Appendix Figure 8).

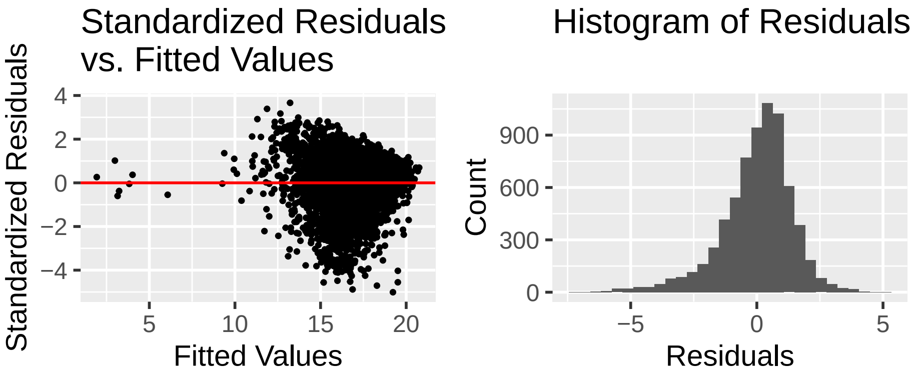
Our output also continues to prove that multi_language is not a useful predictor variable, so we will remove it from future consideration; also, we will not put the interaction between multi_genre and vote_average in our final model.

Final Variables and Data Entry Errors

Therefore, below we have decided on the following variables for our final model: vote_average, log_budget, runtime, year, multi_genre, and the interaction between multi_genre and runtime.

As we were testing this model and checking assumptions for linear regression, we noticed some concerning evidence of data entry errors. We investigated points with high standardized residuals and began to look for a pattern in the data. After not finding anything indicating a certain genre, language, or runtime length may be causing some sort of abnormality, we looked to log_revenue and log_budget. Interestingly, we found that the ratio between log_revenue and log_budget tended to be abnormally extreme for points with high standardized residuals, as compared to the average of the entire data set. As such, we looked at a few specific data points to see if there were potential issues; we found data entry errors were present. For instance, the movie *Kiss of the Spider Woman* has a reported budget of $11, but according to the American Film Institute, the actual budget was north of $1.5 million. When briefly looking at a few other data points, this was the case for multiple other movies with extreme log_revenue to log_budget ratios as well.

The main issue was that we have no way of parsing through the entire dataset and checking which ones had actual entry errors. As such, we decided to filter out these points using a sort of proxy-filter and create a new data set to use for our final model. The filter was to remove points that were +/- 3 standard deviations away from the mean ratio between log_revenue and log_budget. Although this won't be an entirely accurate filter (and may even take out some legitimate data points, such as movies that made much more than they spent), we decided this was the best and most efficient way to attempt to control for inaccurate data, as many data entry errors did indeed result in extreme ratios. In total, 61 movies were removed. We checked linear regression assumptions for the full data set with these points still included, which can be found in the appendix (Appendix Figures 10 and 11). The assumptions analysis for the updated dataset is below. The key difference from removing these data entry errors is that constant variance was not satisfied before, but it appears to be much closer to being satisfied now.



Based on the new residual vs. fitted value plot above, the constant variance condition appears to be satisfied to a much better extent compared to this same plot using the original data set before the data entry error removals, and the plot is much less skewed than the constant variance plot for the model with obvious data entry errors present (Appendix Figure 10). We'll assume our model meets the linearity assumption as we can't predict with strong accuracy if a residual will be positive or negative. Additionally, our model meets the normality assumption. According to the histogram of residuals, we can see the residuals are approximately normally distributed, following a bell-curve shape centered at 0. Additionally, we did a check for independence by segmenting movies by their individual production companies. As there are over 7,000 different production companies, we chose to check independence in three that we know tend to have a lot of sequels/franchises: Marvel, Warner Bros., and Disney. Each plot for residuals vs. fitted values showed no pattern, so it seems that the individual residuals are independent of one another (Appendix Figure 12).

As a whole, we will assume that our model meets **all** requirements: linearity, constant variance, normality of residuals, and independence.

Results

Final Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	24.689	2.224	11.103	0	20.330	29.048
vote_average	0.607	0.024	25.589	0	0.560	0.653
log_budget	0.973	0.013	76.615	0	0.948	0.998
runtime	-0.013	0.003	-4.947	0	-0.018	-0.008
year	-0.013	0.001	-11.740	0	-0.015	-0.011
multi_genreTRUE	-1.075	0.292	-3.685	0	-1.647	-0.503
runtime:multi_genreTRUE	0.010	0.003	3.851	0	0.005	0.016

When fitting our final model based on the data without the likely data entry error points present, we saw that our R^2 value increased by over 4% (from 46.556% to 50.650%). In terms of the BIC, this model experienced a nearly 1500 point decrease from the model using the data set with potential entry errors (Appendix Figure 9 vs 13). We believe this is more than enough evidence to suggest that our model does a better job of predicting movie revenue when movies with extreme log_revenue to log_budget ratios are not present in the data, because they are likely the result of data entry errors. It is important to note that ultimately, as mentioned in the analysis, that removing these movies based off the filter we used is not perfect, as it is entirely possible that a movie may have a very low budget and very high revenue, or vice versa. However, the size of our dataset being so large means that these removals of legitimate data points may not be a cause of significant concern.

The p-values for every coefficient are near-zero, suggesting that all these coefficients are statistically significant for predicting the log(revenue).

- When vote_average is 0, budget is 1, runtime is 0, year is 0, multi_genre is FALSE, we would expect the movie to make exp(24.689) dollars. However, this interpretation is not relevant because there are variables like year and log_budget that would never be zero, particularly since we filtered out movies that were below a nonzero threshold for both these variables.
- For a one point increase in the average score given to the movie, we would expect the revenue to multiply by a factor of 1.835, holding all other variables constant.

- For a 10 percent increase in budget, we would expect the revenue to multiply by 1.09, holding all other variables constant.
- For a ten minute increase in runtime, we would expect the revenue to multiply by 0.878 (exp(-0.013*10), holding all other variables constant.
- For a one year increase in production year, we would expect the revenue to multiply by .99, holding all other variables constant. According to the confidence interval of [-0.015, -0.011], 0 is not included which means it is statistically significant. However, in the context of the data, for every one year increase in year, we can expect the revenue to decrease by $10 per $1,000 of revenue (1,000(1 − 0.99)), holding all other variables constant. A $10 decrease is not practically significant because it is a very small amount in terms of the greater amounts of revenue being made.
- We would expect a movie with more than one genre to make .341 times the revenue of a single genre movie, holding all else constant.
- For multi-genre movies, we would expect a ten minute increase in runtime to multiply the revenue by .119 more than if the movie was single genre, holding all other variables constant.

Conclusion

Our goal for this project was to predict and understand movie box office revenue by examining various factors. In a post-COVID world, this topic feels particularly interesting because of how the entertainment industry has changed, with fewer people going to theaters due to the rise of streaming platforms, changes in viewer habits, and other trends. Through this project, we aimed to better understand what drives a movie's success or failure at the box office.

From our analysis, the most influential factors were the average audience rating (vote), budget, runtime, release year, and whether the movie was categorized as having multiple genres. These metrics provided meaningful insights into what contributes to a movie's box office performance. However, it is also important to recognize what is missing from our data. For example, not all movies are competing at the box office—streaming platforms like Netflix and Hulu offer viewers the convenience of watching from home. Additionally, by removing points with extreme revenue to budget ratios in order to reduce the influence of data entry errors on our final model, we potentially got rid of valid data points as well.

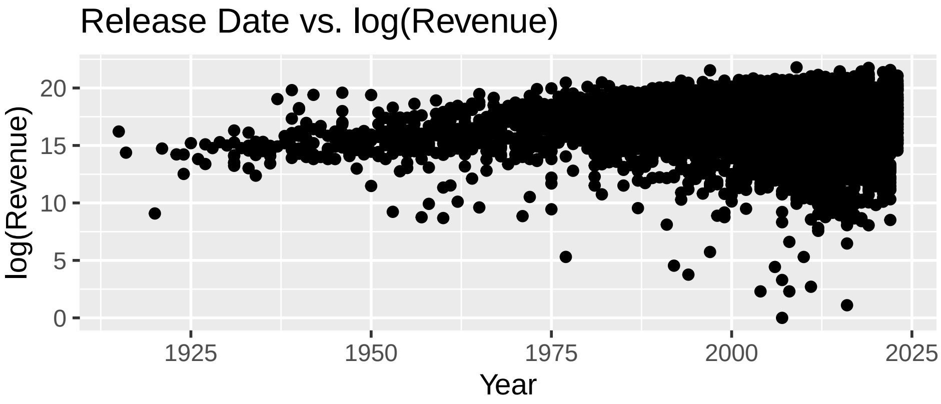
Ultimately, understanding what influences box office revenue is a powerful tool for producers and studios. By focusing on factors like how well people like the movie, how much money is spent making it, and how many genres it covers, producers can make more informed decisions during production. At the same time, they should stay mindful of emerging trends in the entertainment industry, like streaming services, that continue to reshape what makes a movie successful. Overall, with success no longer defined solely by ticket sales, producers can explore new ways to measure and achieve it.

Appendix

(1)

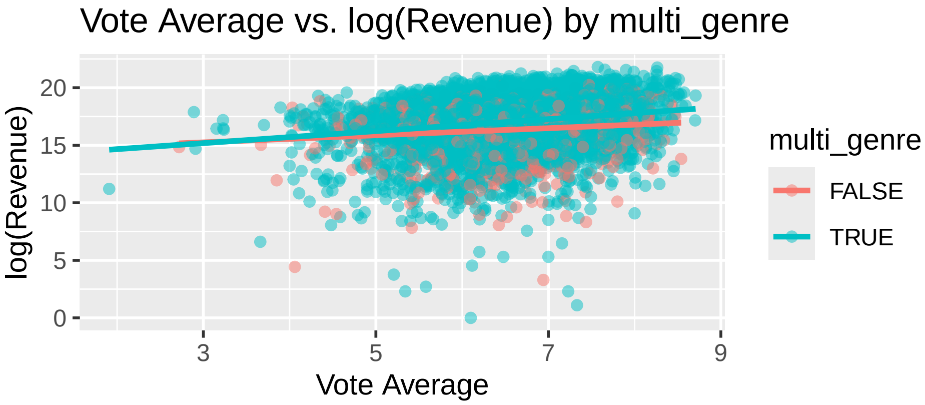
min	q1	median	q3	max	mean	sd
1	8334856	30099904	94458655	2923706026	91717069	178960324

(2)



According to the graph comparing year and log(revenue), there is a positive, weak correlation. Movies with more recent years tend to have a slightly higher logged revenue, which makes sense in the context of inflation. There are many data points clustered between 1975 and 2023.

(3)



(4)

```
      vote_average          log_budget             runtime                year 
          1.212047            1.267771            1.296056            1.160600 
   multi_genreTRUE multi_languageTRUE 
          1.031545            1.043336 
```

(5)

r.squared	adj.r.squared	AIC	BIC
0.465	0.465	26301.5	26356.4

(6)

r.squared	adj.r.squared	AIC	BIC
0.466	0.466	26287.78	26349.53

(7)

term	estimate	std.error	statistic	p.value
(Intercept)	24.958	2.432	10.262	0.000
vote_average	0.594	0.067	8.899	0.000
log_budget	0.926	0.013	70.672	0.000
runtime	-0.012	0.003	-4.080	0.000
year	-0.013	0.001	-10.632	0.000
multi_genreTRUE	-1.379	0.453	-3.047	0.002
multi_languageTRUE	-0.059	0.041	-1.448	0.148
runtime:multi_genreTRUE	0.011	0.003	3.357	0.001
vote_average:multi_genreTRUE	0.040	0.072	0.558	0.577

r.squared	adj.r.squared	AIC	BIC
0.466	0.466	26289.46	26358.08

(8)

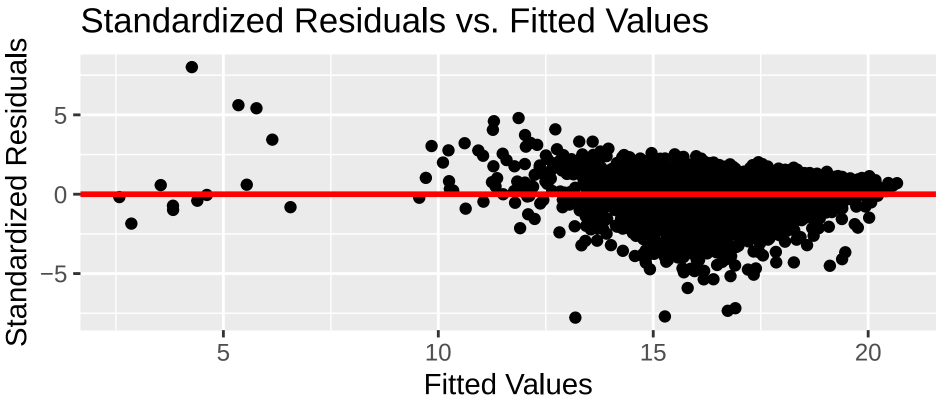
Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
7050	17086.19	NA	NA	NA	NA
7049	17085.43	1	0.755	0.312	0.577

(9)

term	estimate	std.error	statistic	p.value
(Intercept)	24.785	2.409	10.288	0
vote_average	0.626	0.026	24.372	0
log_budget	0.925	0.013	70.671	0
runtime	-0.013	0.003	-4.743	0
year	-0.013	0.001	-10.620	0
multi_genreTRUE	-1.193	0.315	-3.780	0
runtime:multi_genreTRUE	0.011	0.003	3.937	0

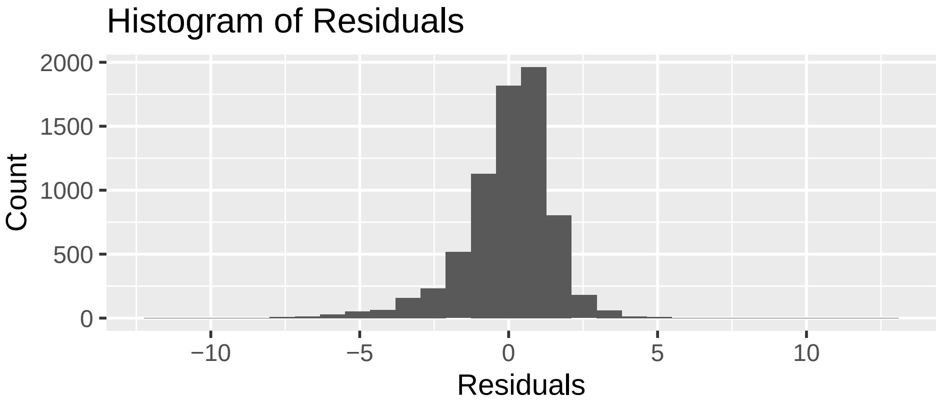
r.squared	adj.r.squared	AIC	BIC
0.466	0.466	26287.89	26342.79

(10)



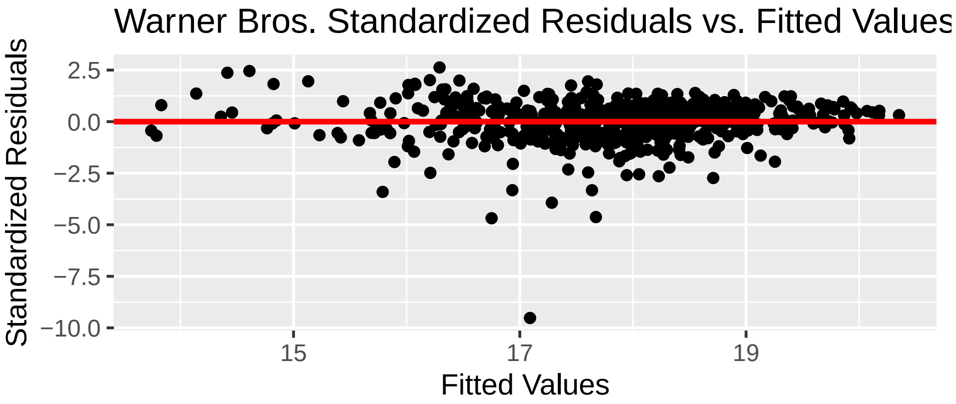
Our model does not satisfy the constant variance assumption, which indicates that the variability in our data could be improved. In the scatter plot of Standardized Residuals vs. Fitted Values, the points are not randomly or evenly distributed around the red line. Specifically, there are outliers that are noticeable near the top left, and many points are densely clustered below the red line on the right side of the plot.

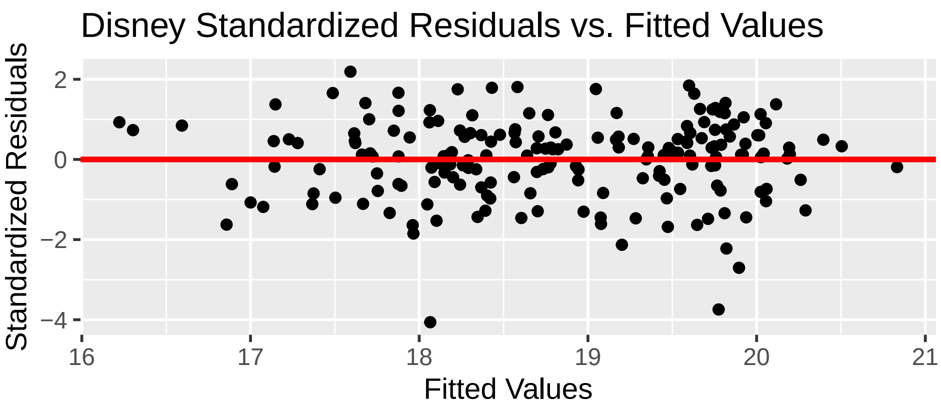
(11)

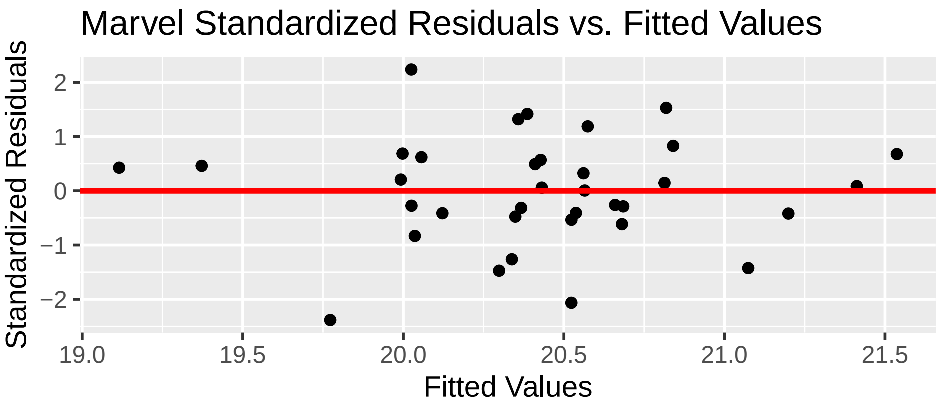


Our model meets the normality assumption. According to the Histogram of Residuals, we can see that the residuals are approximately normally distributed, following a bell-curve shape centered around 0. Our model does not necessarily meet the independence assumption because we cannot confidently distinguish movie sequels or easily identify whether movies belong to the same company or franchise. This limitation could introduce bias into our data since viewers often tend to watch sequels, movies within franchises, or movies produced by the same company (e.g., Marvel movies have a certain audience that tend to watch all their movies to keep up with the franchise, contributing to the huge revenue they make), which could create a dependency between observations.

(12)







(13)

r.squared	adj.r.squared	AIC	BIC
0.507	0.506	24822.73	24877.55

Asaniczka. 2023. "TMDb Movies Dataset (2023): 930K Movies." https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies.

Box Office Mojo. 2024. "Domestic Yearly Box Office." https://www.boxofficemojo.com/year/.