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 [@boxofficemojo2024], 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.

Research Question: What are the key factors and metrics that most significantly influence a movie's gross revenue?

Motivation: 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.

Hypotheses: 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 dataset 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 dataset on Kaggle, and it is updated daily by user asaniczka, who takes the data from TMDb [@asaniczka2023].

There were 953629 observations in the original dataset, each representing a single film. There were 24 columns in this dataset, 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 dataset outside of this project first because the original dataset 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 dataset 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 dataset just put 0s instead of NAs.

Analysis Approach

Variables of interest:

vote_average: This is the average rating given to the movie by viewers. We expect this to be correlated with the response variable, as good ratings will encourage more people to go see the movie, and if someone likes the movie they will be more likely to recommend it to their friends.

spoken_languages: This is the number of different languages spoken in the movie. We hypothesize that if a movie has more languages in it, it will appeal to a broader audience and therefore

earn 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: This is 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.

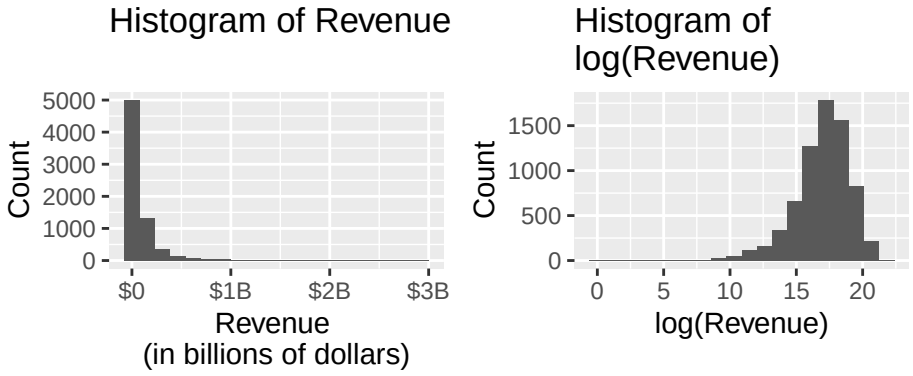
release_date: This indicates the day, month, and year that the movie was released. We expect that the year will greatly impact gross revenue because, for instance, movies released right before or during the pandemic will have very low gross revenue as no one was able to go to the theaters. We created 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: This is a genre or list of genres associated with the movie. We expect movies tagged with more than one genre to appeal to a wider audience, thus driving up the number of people who go to see the movie and increasing gross revenue. We created a new variable from this variable called "multi_genre" that categorizes movies with multiple genres as "TRUE" and movies with one genre as "FALSE."

Method of approach:

We plan to use multiple linear regression to predict gross revenue, but we will log transform gross revenue because we can see from our EDA (below) that revenue is extremely skewed in its original form.

Univariate EDA

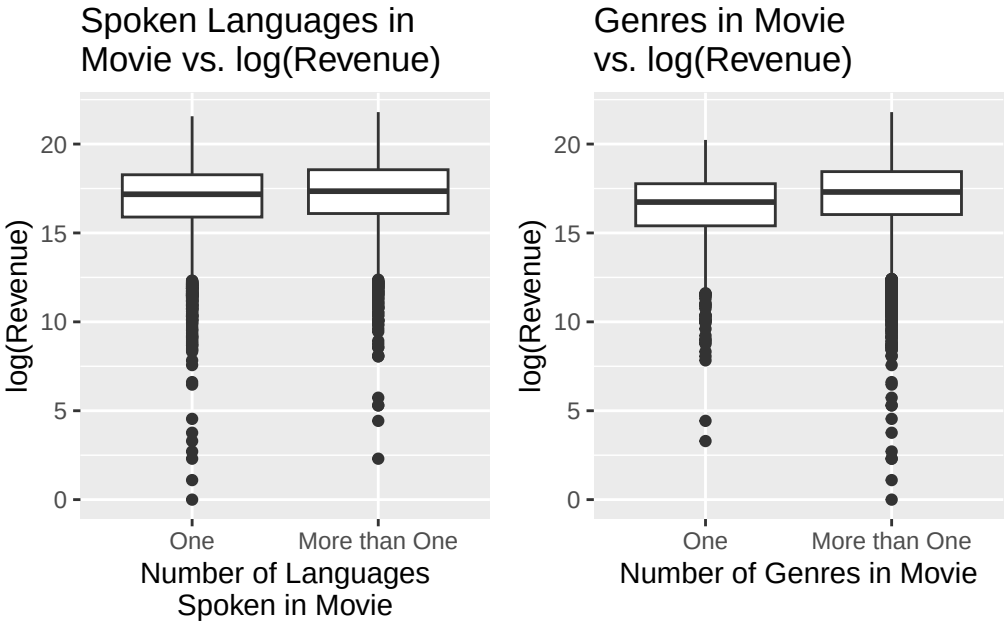


min	q1	median	q3	max	mean	sd
1	8334856	30099904	94458655	2923706026	91717069	178960324

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.

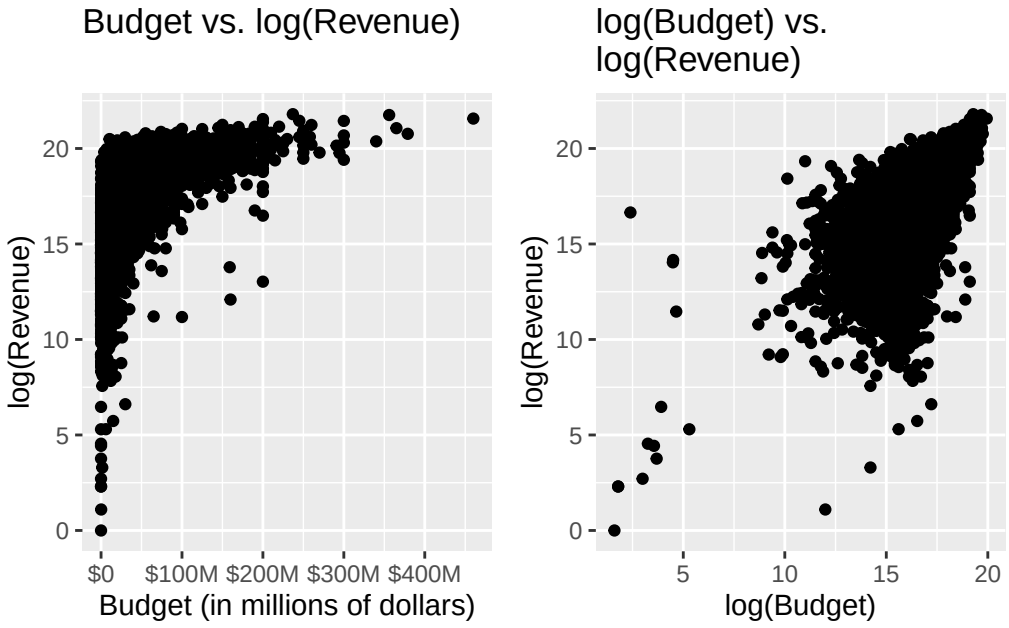
Following a log transformation, our response variable, logged revenue, has a slight left skew but is unimodal and very roughly normal. It has a significant number of outliers on the left side, and log(Revenue) has a median of 17.22 and an IQR of 2.43.

Bivariate EDA

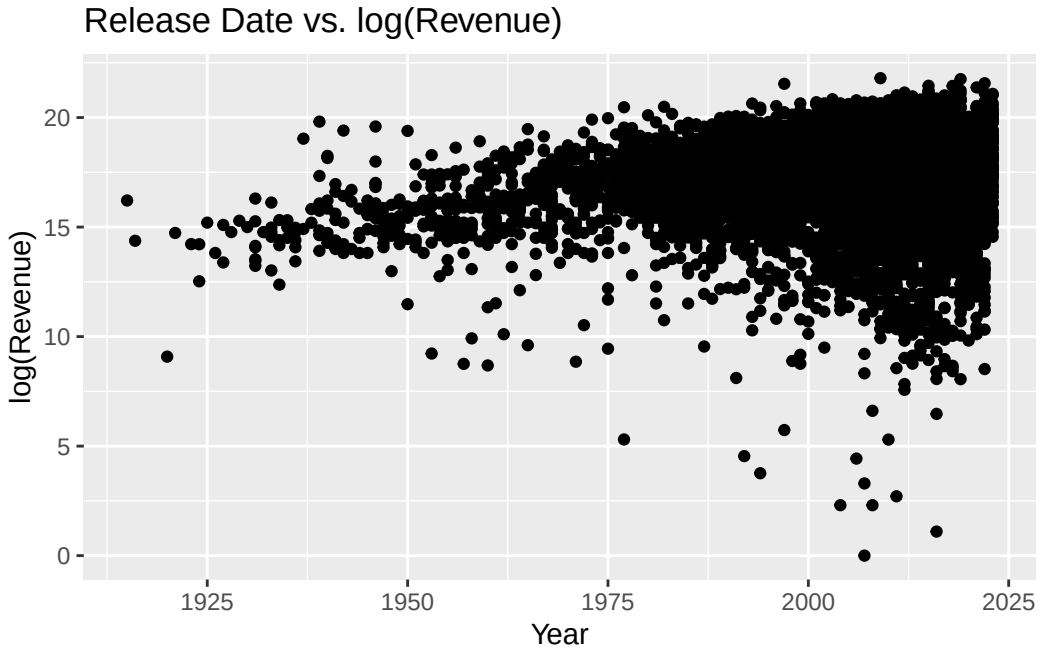


Based on the box plots for the number of spoken languages in a movie, it seems that whether one language was spoken or multiple were does not have much of an impact on the logged revenue. This is evident based on the overlapping of the two box plots.

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.



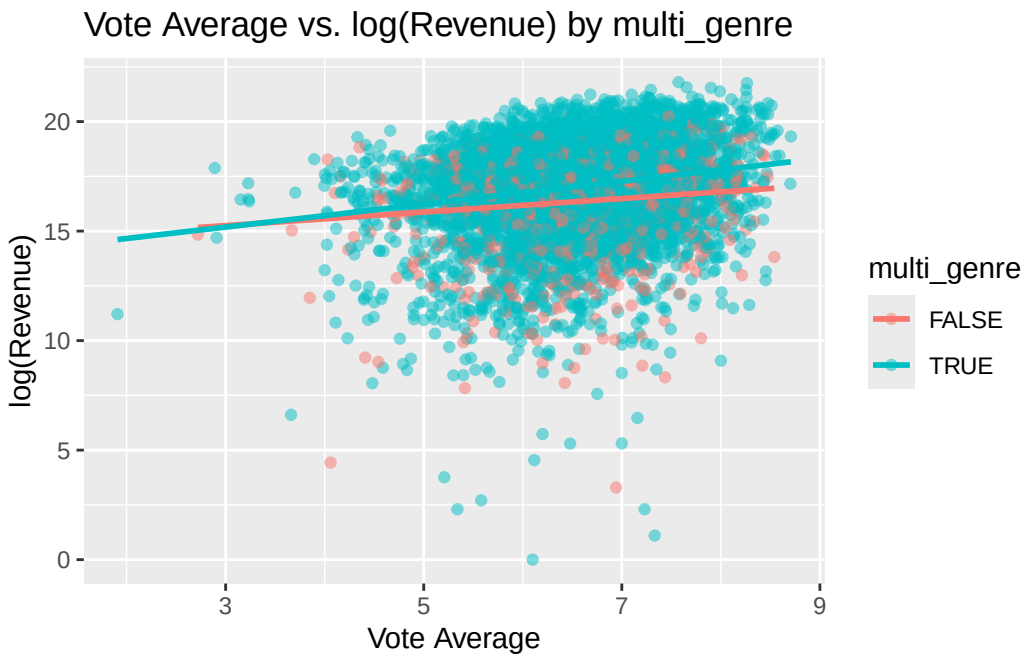
Because of the extreme curvature present in the graph of budget vs. log(revenue), it makes sense to also do a log transformation on Budget in order to get the graph log(budget) vs. log(revenue) which is seen on the bottom right. The graph of log(budget) vs. log(revenue) has far less curvature and appears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



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.

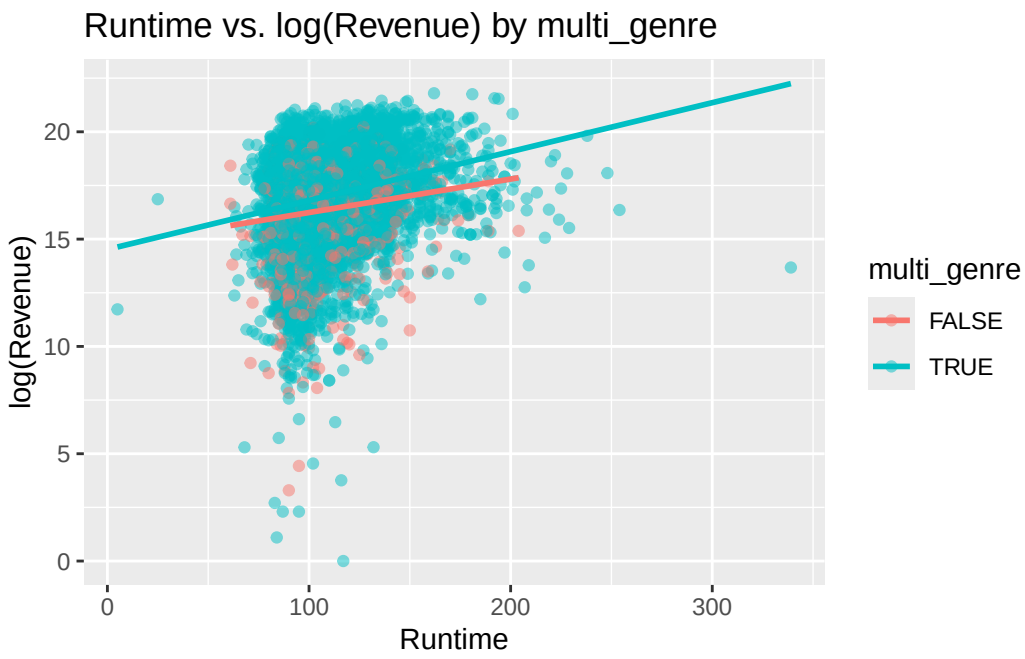
Interaction

First interaction we want to investigate is the genre and vote average. It is possible that the effect of vote_average 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 line differs by a bit, so it is possible there is evidence of interaction, but further analysis is needed.

Second interaction we want to investigate is 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, but 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.

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

```
# A tibble: 1 x 4
  r.squared adj.r.squared    AIC    BIC
      <dbl>         <dbl>  <dbl>  <dbl>
```

```
1     0.465         0.465 26302. 26356.
```

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).

Multicollinearity Analysis

Then, we will take care of the issue of multicollinearity by using the VIF to see which predictor variables could have correlation.

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

Based off of our VIF function, it seems as if there are no variables that raise issues of multicollinearity because all of our variables have a VIF much lower than 10.

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

```
# A tibble: 1 x 4
  r.squared adj.r.squared    AIC    BIC
      <dbl>         <dbl>  <dbl>  <dbl>
1     0.466         0.466 26288. 26350.
```

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.

In effect, 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%. 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 will now test the model with BOTH interaction terms against the model with ONLY the multi_genre and runtime interaction.

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

term	estimate	std.error	statistic	p.value
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

```
# A tibble: 1 x 4
  r.squared adj.r.squared    AIC    BIC
      <dbl>         <dbl>  <dbl>  <dbl>
1     0.466         0.466 26289. 26358.
```

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

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. 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 (**see appendix**). **SHOULD WE JUST MOVE MODEL OUTPUT TO APPENDIX TO SAVE SPACE?**

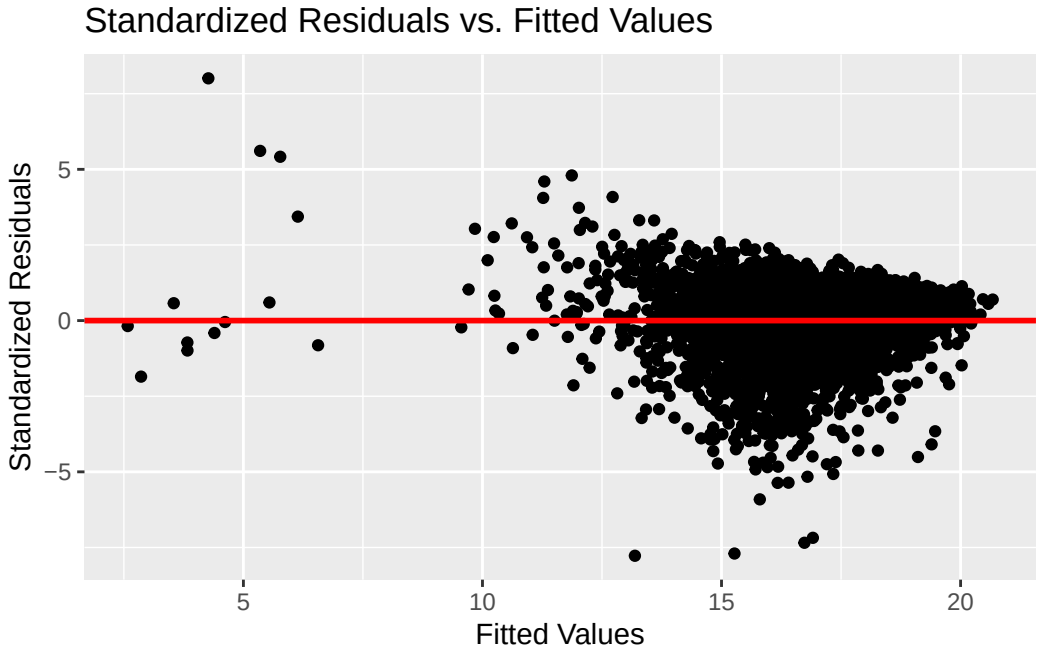
Our output also continues to prove that multi_language is not a useful predictor variable, so we will remove it from future consideration; also, we will not put the interaction between multi_genre and vote_average in our final model.

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

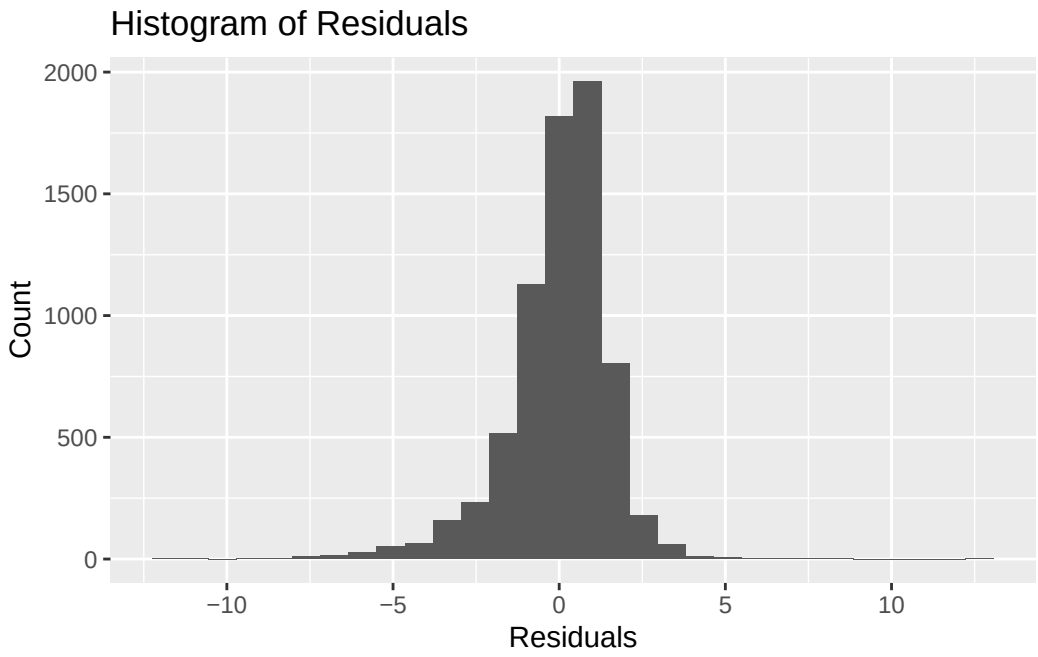
```
# A tibble: 1 x 4
  r.squared adj.r.squared    AIC    BIC
      <dbl>         <dbl>  <dbl>  <dbl>
1     0.466         0.466 26288. 26343.
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
7051	17091.32	NA	NA	NA	NA
7050	17086.19	1	5.13	2.117	0.146

SAMIR TRAVERRSSSS - removing multi language; F test and see it is not significant; technically our final model



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. Our model meets the linearity assumption because we cannot predict with strong accuracy if a residual will be positive or negative.



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 meet the independence assumption because we cannot distinguish movie sequels or 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 or movies produced by the same company (e.g., Marvel movies), which creates a dependency between observations.

- need to interpret coefficients for final model

SAMIR TRAVERSSSSSSS - interpret all coefficients (remember it is log logged –> gotta think)

Investigating Influential Points (IGNORE FOR NOW)

We dove deeper into our potential influential points and began to look for a pattern in the data. After not finding anything to indicate that 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 much more extreme for our influential points as compared to the average of the entire data set (see calculations in appendix). 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.

When fitting our final model based on the data without these points present, we saw that our R^2 value increased by nearly 4%. 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, removing all these movies is not perfect, as it is entirely possible that a movie may have a very low budget and very high revenue, or vice versa.