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].

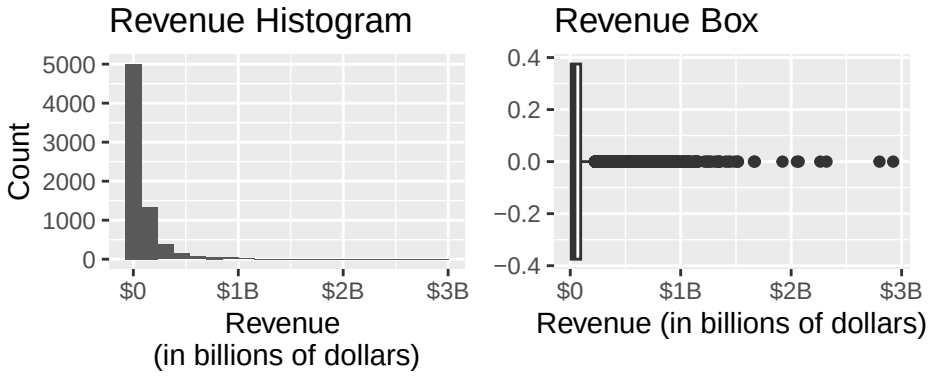
There were 953629 observations in the original dataset, each representing a single film. There were 24 columns in this dataset, with some of the most important being the name (an identifier), release year, runtime, genres, rating, votes, and revenue.

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.

Below is basic initial EDA of our untransformed response variable, revenue.



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.

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."

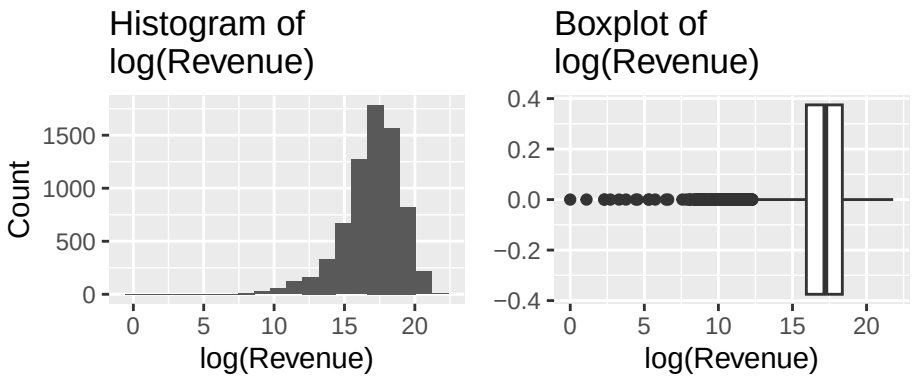
Method of approach:

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 that revenue is extremely skewed in its original form.

Data dictionary

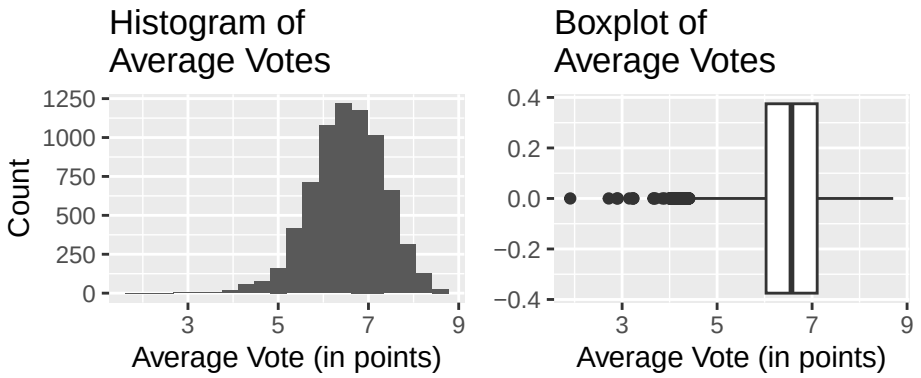
The data dictionary can be found here: data/README.md

Univariate EDA



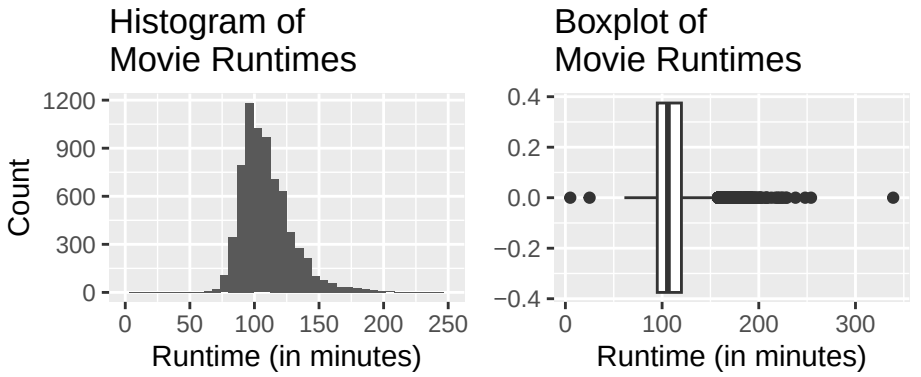
min	q1	median	q3	max	mean	sd	IQR
0	15.936	17.22	18.364	21.796	16.933	2.13	2.428

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.



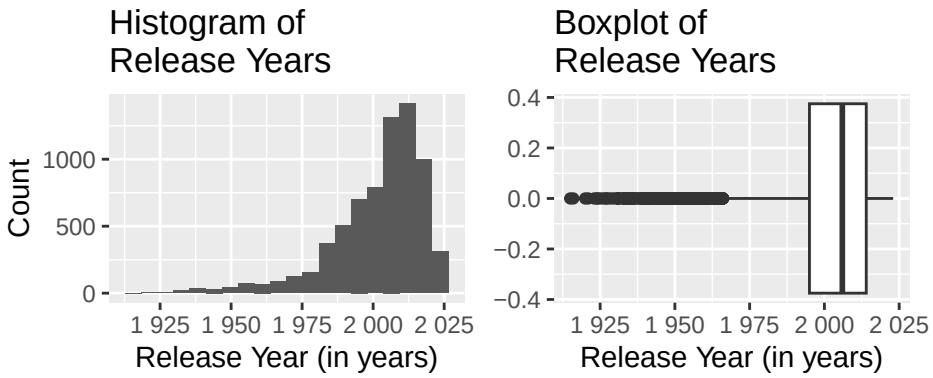
min	q1	median	q3	max	mean	sd	IQR
1.908	6.031	6.566	7.107	8.707	6.54	0.794	1.076

The predictor variable, vote_average, which measures the average rating given to a movie by users, is roughly normally distributed and is unimodal. It has a few outliers on the right side, and has many on the left. The median is 6.57 points (out of ten), and the IQR is 1.08.



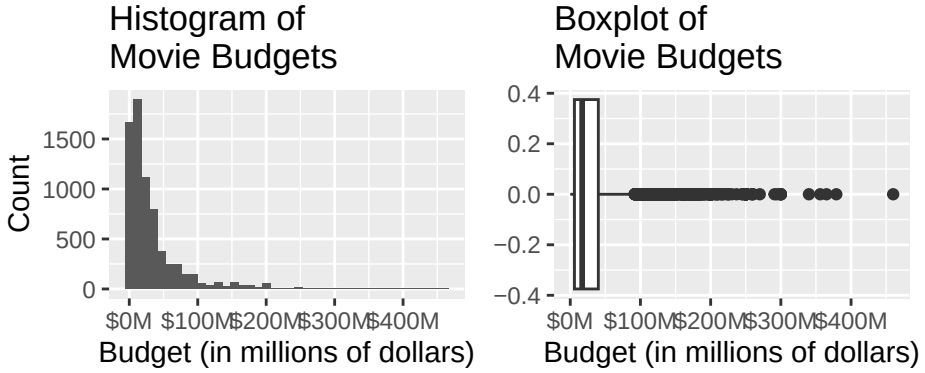
min	q1	median	q3	max	mean	sd	IQR
5	95	106	120	339	109.721	20.679	25

The predictor variable movie_runtime, is roughly normally distributed and unimodal, however it has a right skew. There are numerous outliers on both the left and right, including one very large outlier of 333 minutes. The median is 106 minutes, and the IQR is 25.



min	q1	median	q3	max	mean	sd	IQR
1915	1995	2006	2014	2023	2001.884	16.481	19

The predictor variable year has a strong left skew, and is unimodal. There are a significant number of outliers on the left. The median release year is 2006, and the IQR is 19.



min	q1	median	q3	max	mean	sd	IQR
0	6	17	40	460	31.683	41.876	34

The budget is unimodal with a strong right skew. There are numerous outliers on the right side. The median budget is 17 million dollars, with an IQR of 34 million.

multi_language	n	proportion
FALSE	4871	0.69

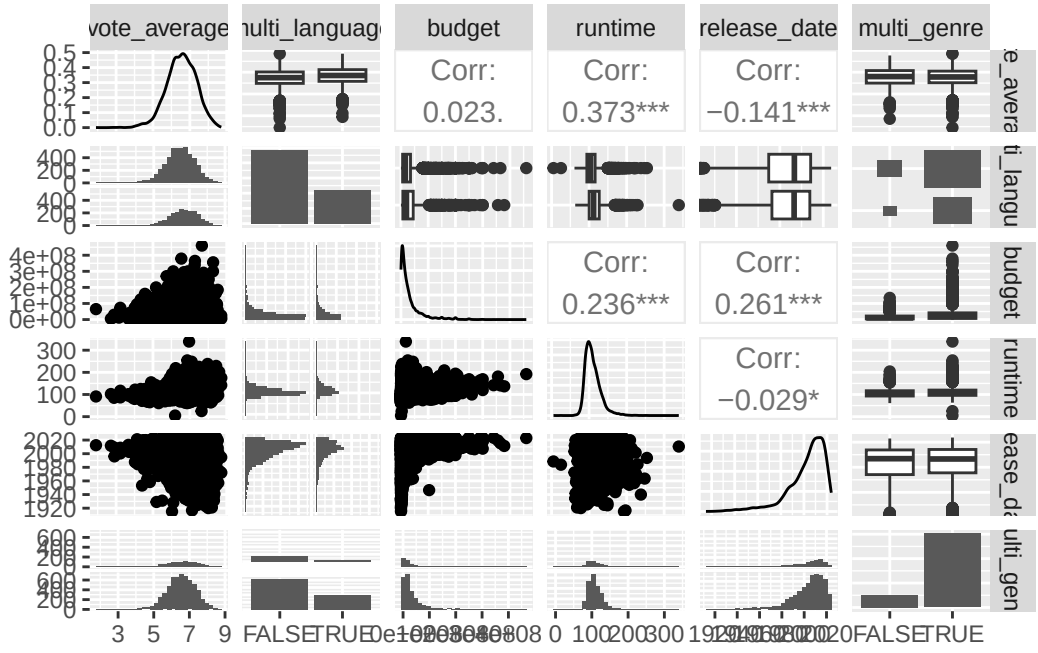
multi_language	n	proportion
TRUE	2187	0.31

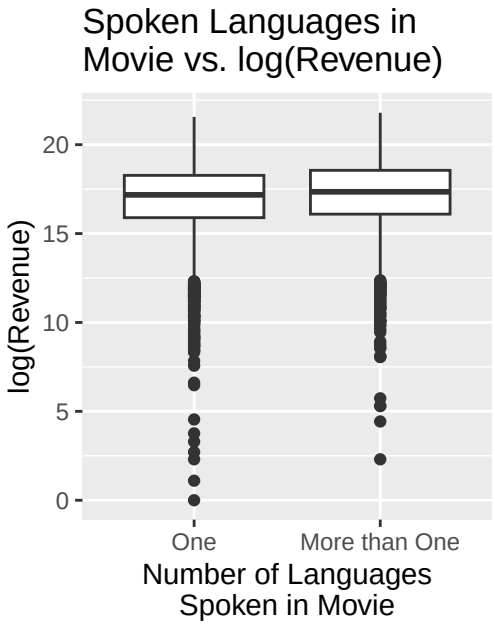
About 69 percent of the data is in just one language, and 31 percent features other languages.

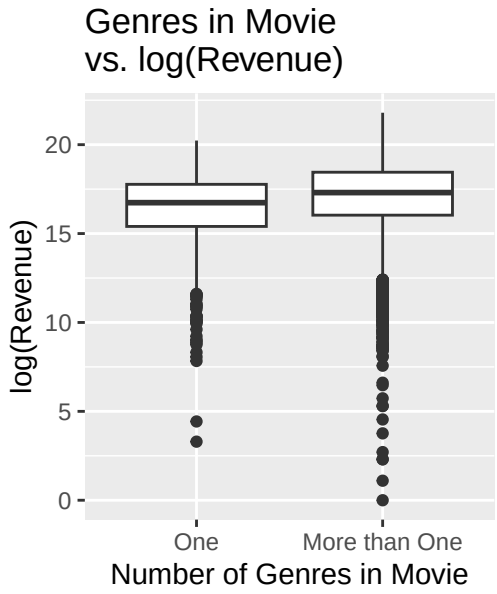
multi_genre	n	proportion
FALSE	1001	0.142
TRUE	6057	0.858

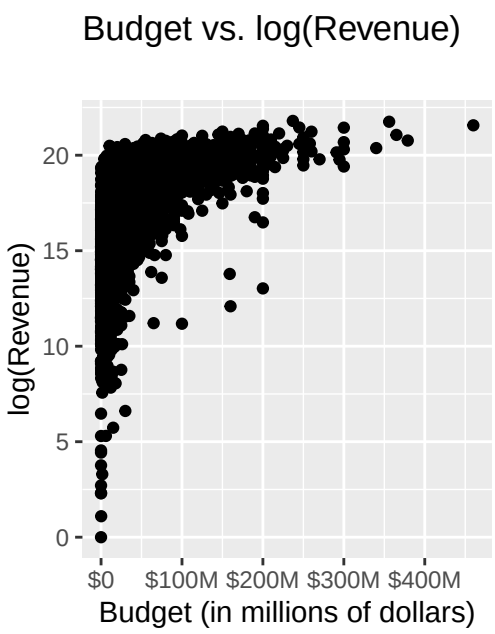
About 86 percent of the movies are classified as more than one genre.

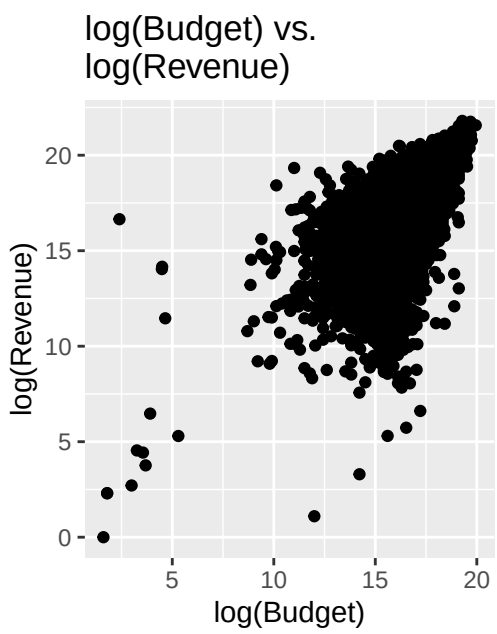
Bivariate EDA









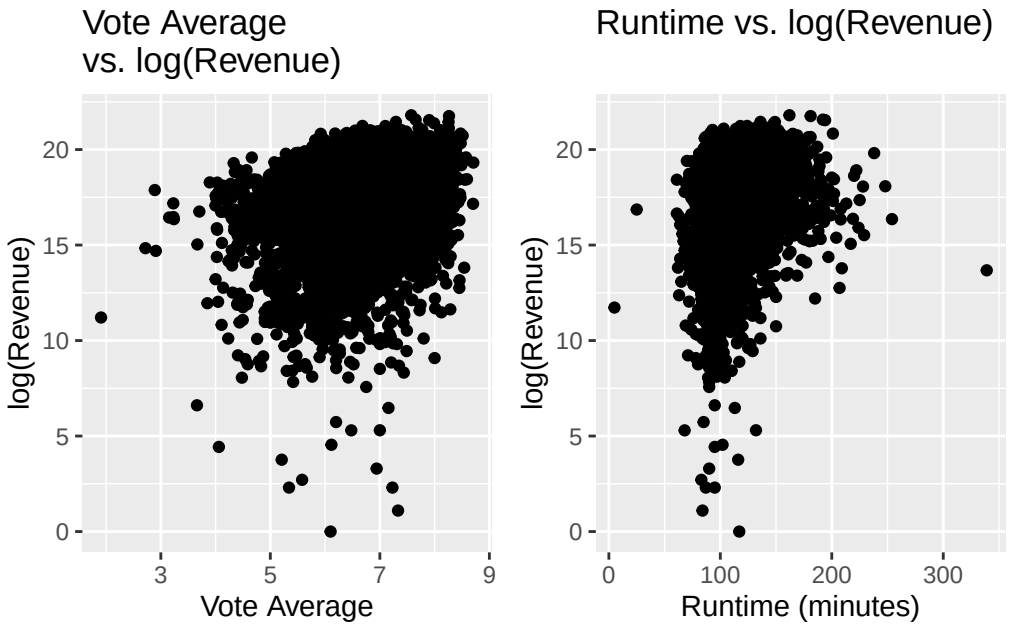


Based on the boxplots for the number of spoken languages in a movie, it seems that whether one language was spoken or multiple were does not have much of an impact on the logged revenue. This is evident based on the overlapping of the two boxplots.

Based on the boxplots for the number of genres a movie is tagged for, it also appears that

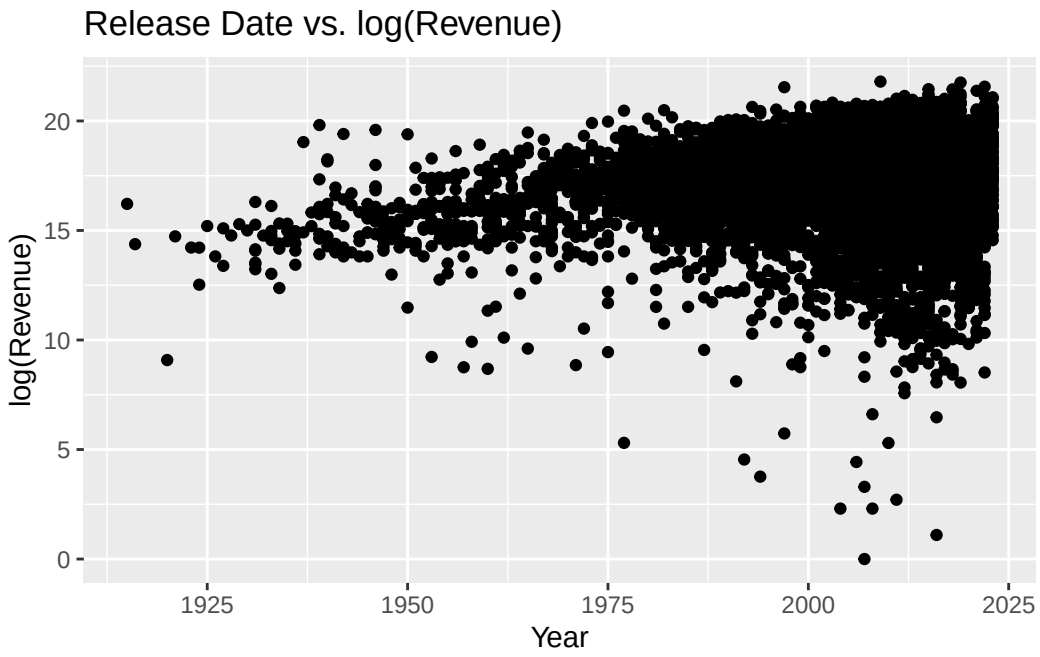
whether the movie is grouped into one genre or multiple genres does not have much of an impact on the logged revenue. Again, this is evident based on the overlapping of the two boxplots.

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 vote average and logged revenue, there is a weak positive correlation. Movies with higher ratings tend to have slightly higher revenues. Most of the data points are also clustered between vote averages of 5 and 8.

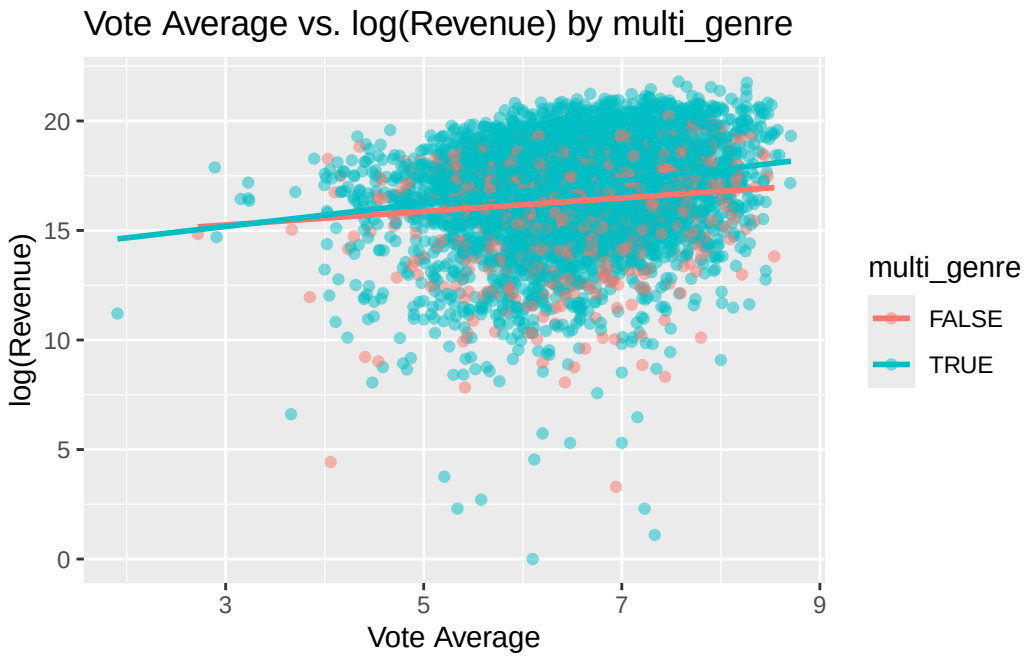
According to the graph comparing runtime and log(revenue), there is no clear positive or negative correlation. There are many data points clustered between 100 and 200 minutes, which makes sense considering the average length of most movies.



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.

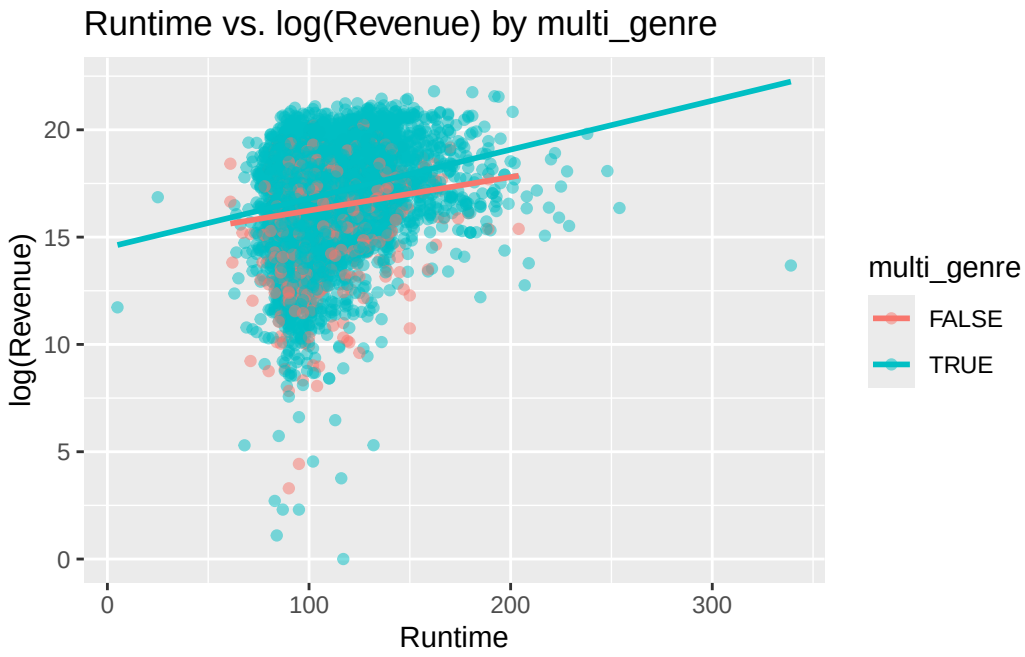
Interaction

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.