Lab 2: What Makes a Movie Successful? Fall 2021

w203: Statistics for Data Science

November 1, 2021

Introduction

What makes a movie "one of the greats"? While that question may elude both data analysts and artists for years to come, we are looking to find some effective indicators of box office success for films, in the hopes of maintaining a thriving movie industry for years to come. This project will focus on several commonly cited reasons for movie fiscal success and analyze how much of an impact each factor has.

The client for this analysis could be any major movie studio or its parent company, e.g. NBCUniversal or Walt Disney Studios. This is a significant issue, because box office sales decreased by 60% in 2020 during COVID-19 pandemic, leaving the movie theaters empty for months. This analysis will provide clues for what a studio can do to speed up the recovery and account for lost revenue.

The main dataset we'll use for this analysis comes from The Movie Database (TMDB - https://www.themoviedb.org/), as well as several assembled lists from online polls (e.g. "Greatest Actors/Actresses of All Time" - Ranker.com).

We're planning on using profitability as the response variable for analysis in this study - using R to clean the empty and irrelevant cells, a "profitability ratio" can be determined for each movie based on its box office sales divided by budget (and controlling for foreign films and currencies).

In order to build our models, we decided to focus on some factors that are usually linked to a movie's financial success, such as budget, time of release, its cast's popularity and director's name brand recognition. We will also take into account some control variables such as runtime and genre.

The products of this analysis will include the code and final datasets, as well as a written summary of the significant results.

Data, Research Question, Underlying Model and Study Design

Data

As mentioned before our main dataset comes from data extracted from The Movie Database (TMDB - https://www.themoviedb.org/). This dataset was compiled by Kaggle for a Machine Learning competition a few years ago.

The original dataset consists of 22 variables for over 7000 films, from the 1920s to 2017. This dataset consists of objective data for each of these movies such as budget, revenue, cast, crew, release date, production company, original language among others.

We also added two more datasets consisting of two lists with a ranking of the "Top 100 movie actors of all time", as well as the top "Top 50 directors". Unlike the previous dataset, these ranks is subjective, but we plan on using these lists to help us create a variable that could stand in for an actor's or director's popularity.

We performed some data cleaning operations on the main dataset in order to have it be ready for our analysis. These included:

- Eliminating any entries that had some vital information missing (such as revenue, budget, cast, crew and genres)

- Parsing certain fields from a stringified JSON.
- Creating the "Actors" column from the "Cast" column
- Extracting the Director from the "Crew" column

After completing this process we still had over 2000 entries.

Research Question

This study intends to find significant relationships between factors known before a movie release and its box office success. We will look at both its profitability ratio (defined as box office divided by its budget) and its box office as ways of measuring a movie's success.

The results of this study could be used by film makers to better inform their investment decisions.

Study Design

We will use the gathered data to answer our research question by finding possible causal relationships between our selected variables and a movie's box office revenue.

We will first conduct an exploratory data analysis on each of our variables, as well as commonly cited relationships, in order to decide which variables could potentially have a causal relationship with a movie's revenue.

Since we are trying to find causal relationships we will focus solely on factors that could be known or decided on before a movie is released, so variables like Rotten Tomatoes rating, word of mouth, or awards will not be considered.

Underlying Model

As mentioned in the introduction, we plan on using commonly cited reasons for movie's financial success to create our models, starting with a simplified model with just the key variables, and then add more variables we hypothesize could have an effect on a movie's financial success to the following models

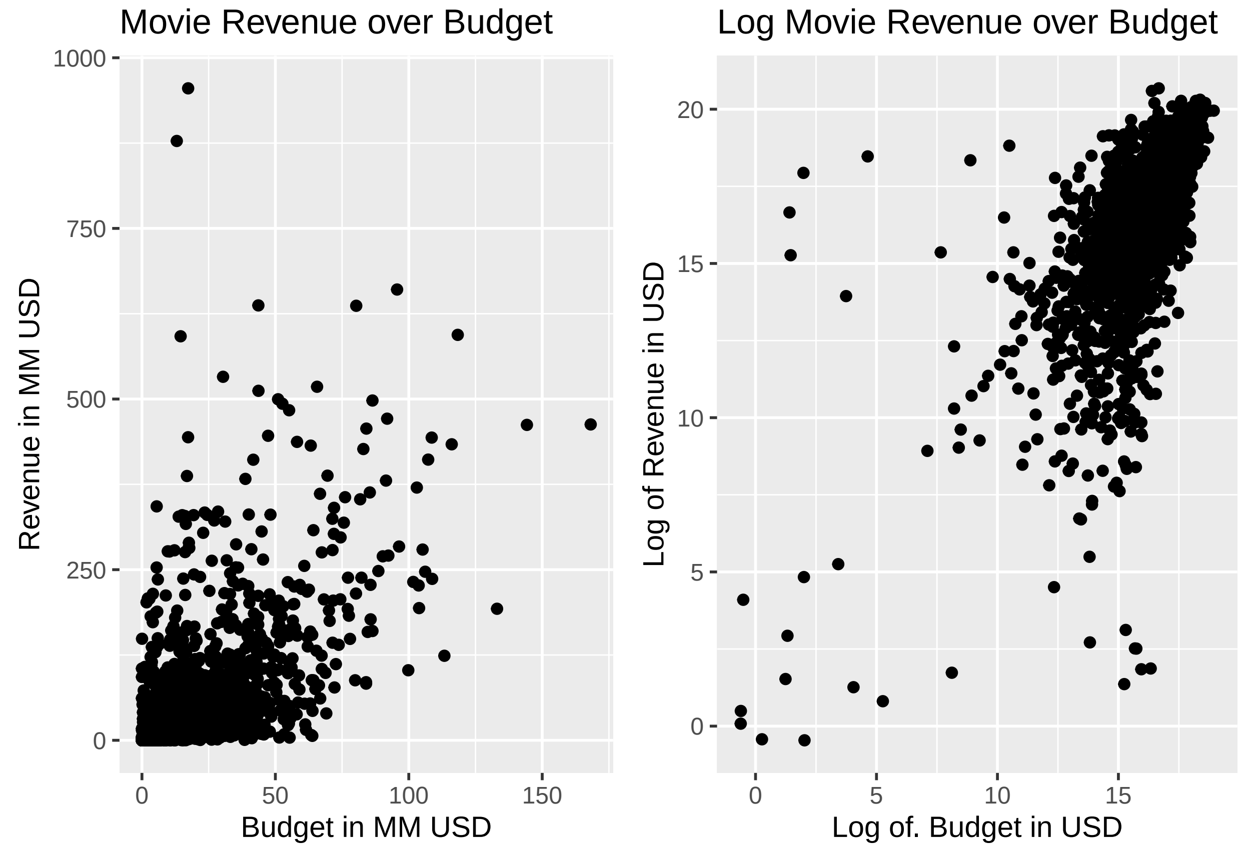
Our outcome variable - Revenue Given our research question, having revenue be our outcome variable was a natural choice. We considered using profitability ratio, as that would allow us to skip some necessary adjustments (Mainly inflation), however we realized that interpreting the results for profitability ratio made it more complicated than it needed to be.

As mentioned before, if we were to use revenue, we would need to account for inflation. Our dataset goes back almost 100 years, which means that the box office of a movie in the 1920s would be compared to a box office in the 2010s. This would skew our results towards more modern films, which would, all things equal, have a higher box office just because of inflation.

We decided to use the Consumer Price Index for All Urban Consumers: All Items in U.S. City Average (CPIAUCNS) from the Federal Reserve Economic Data website. It uses the years 1982 to 1984 as their base.

We applied this index to both our budget and revenue since those are the only variables that are in USD.

Simplified model The most simple and straight forward way of predicting a movie's box office would be its budget. The relationship is pretty clear. The more money that is put into a project the better it is expected to do. One would also expect the increase budget would go into improving the quality of the movie that in turn should increase its box office performance.



As seen in the previous charts, there is a clear positive relationship between budget and revenue, both for with and without a logarithmic transformation. This contributes to our hypothesis than a higher budget leads to a better box office.

Seasonality Just like for a lot of other industries, seasonality in the movie industry is a well documented phenomena. [1]

Certain times of the year tend to attract more moviegoers, such as the summer, the holiday season and certain weekends that coincide with important holidays (Christmas, 4th of July weekend, Memorial Day weekend, among others). This increase demand would lead to a better box office performance.

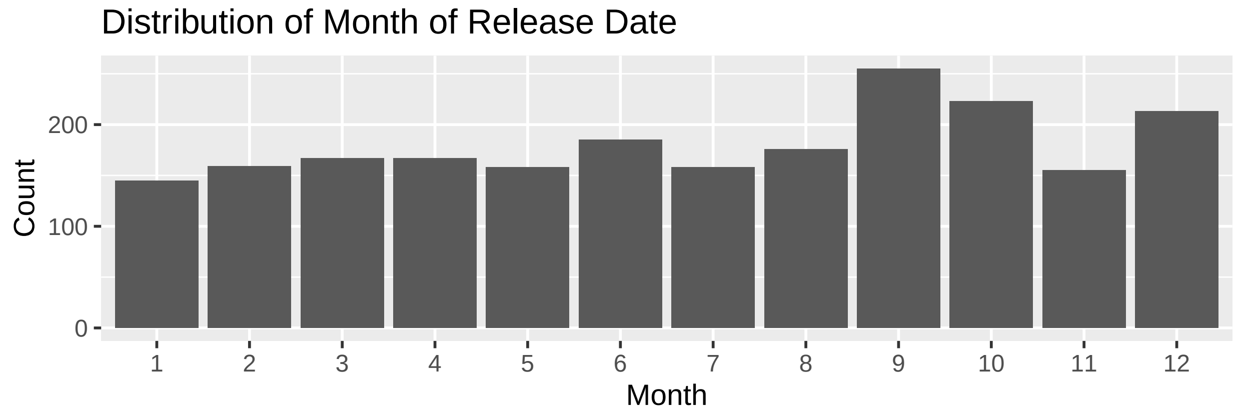
On the other hand, supply is also affected by seasonality. Big budget movies are more likely to be released during a time of the year that could help them maximize their potential revenue.

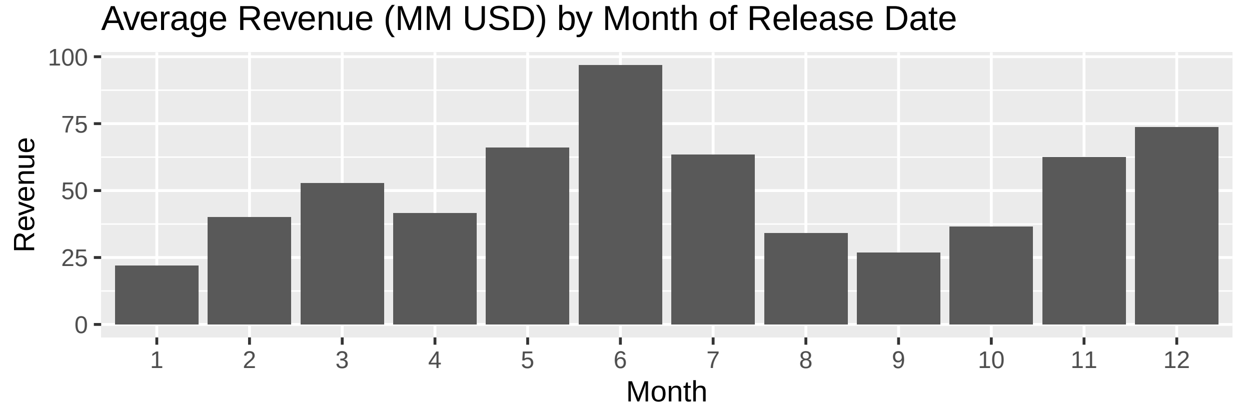
This self-selection on the supply side is likely to have an effect on the demand as well, as a greater number of quality movies during certain times would attract more movie goers.

Our model will control for this effect by adding the month of release to the model. We would expect the summer months, as well as December to have a significant positive effect on revenue.

If we just look at the distribution, we'll notice that more movies come out in the fall, however if we instead look at the average revenue per month, we see a peak during the summer, and a smaller peak towards the end of the year.

[1] Einav, L. (2007). Seasonality in the U.S. Motion Picture Industry. The RAND Journal of Economics, 38(1), 127–145. http://www.jstor.org/stable/25046296





The Cast and Director's Name Recognition It shouldn't be a surprise to anyone that famous actors attract crowds, which in turn increase the movie's box office performance. Audiences want to see their favorite actors in their movies, for a variety of reasons.

A recent example that comes to mind is the 2021 movie adaptation of Dune. The actress Zendaya was a big part of the movie's marketing, despite the fact that she only briefly appears in the movie. A lot of movie goers were upset after seeing the movie because they expected her to have a bigger role.[2].

For our model we will give each movie an "Actor Score" which will be just a counter of how many well known actors are part of their cast.

We have also added a "Director Score" which will reflect a movie's director ranking in the following manner:

- Top 10 - 5 points
- Top 20 - 4 points
- Top 30 - 3 points
- Top 40 - 2 points
- Top 50 - 1 point
- Not on the ranking - 0 points

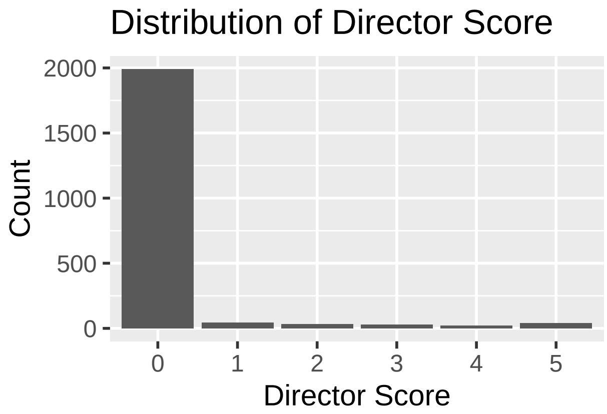
Our initial exploratory data analysis showed us that the vast majority of movies in our dataset were not made by a "Top Director", however, once we look at the average revenue by director score we notice an interesting shape.

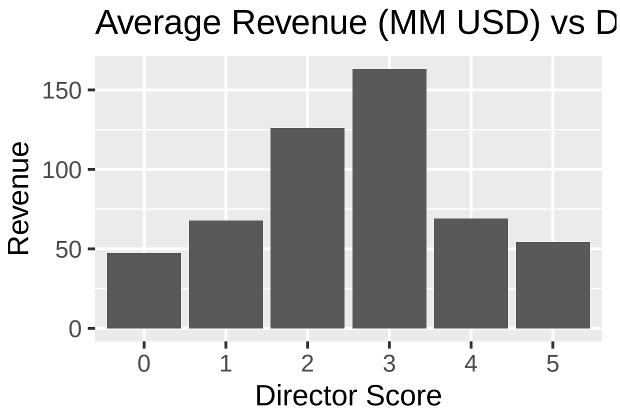
Having a "Top Director" does seem to have a higher average revenue, but what's quite interesting is that the highest revenue is found with the "mid-tier" Top Directors. One explanation could be that the "Best

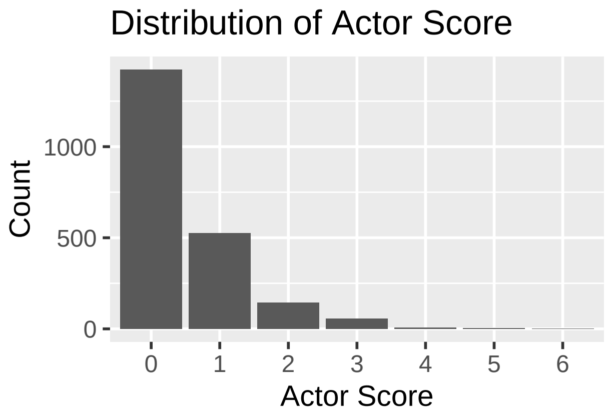
[2] https://nypost.com/2021/10/26/fans-outraged-that-zendaya-is-in-dune-for-only-7-minutes/

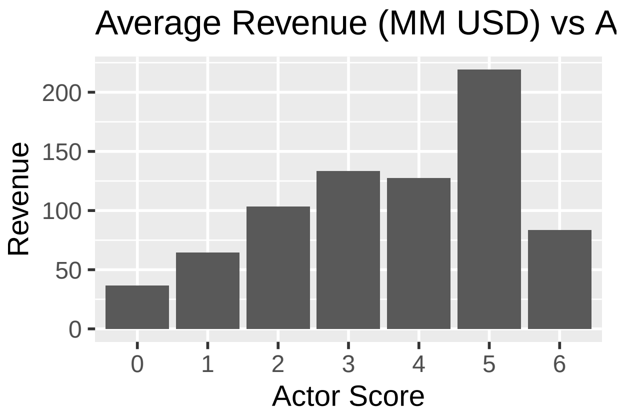
Directors" do not always do films that appeal to mainstream audiences, while the ones towards the middle of the list are both great directors and direct movies that appeal to a wider audience.

A similar thing happens when we look at the "Actor score" variable. Movies with "Top Actors" tend to have a higher average revenue, with this maxing out at 5 "top actors".





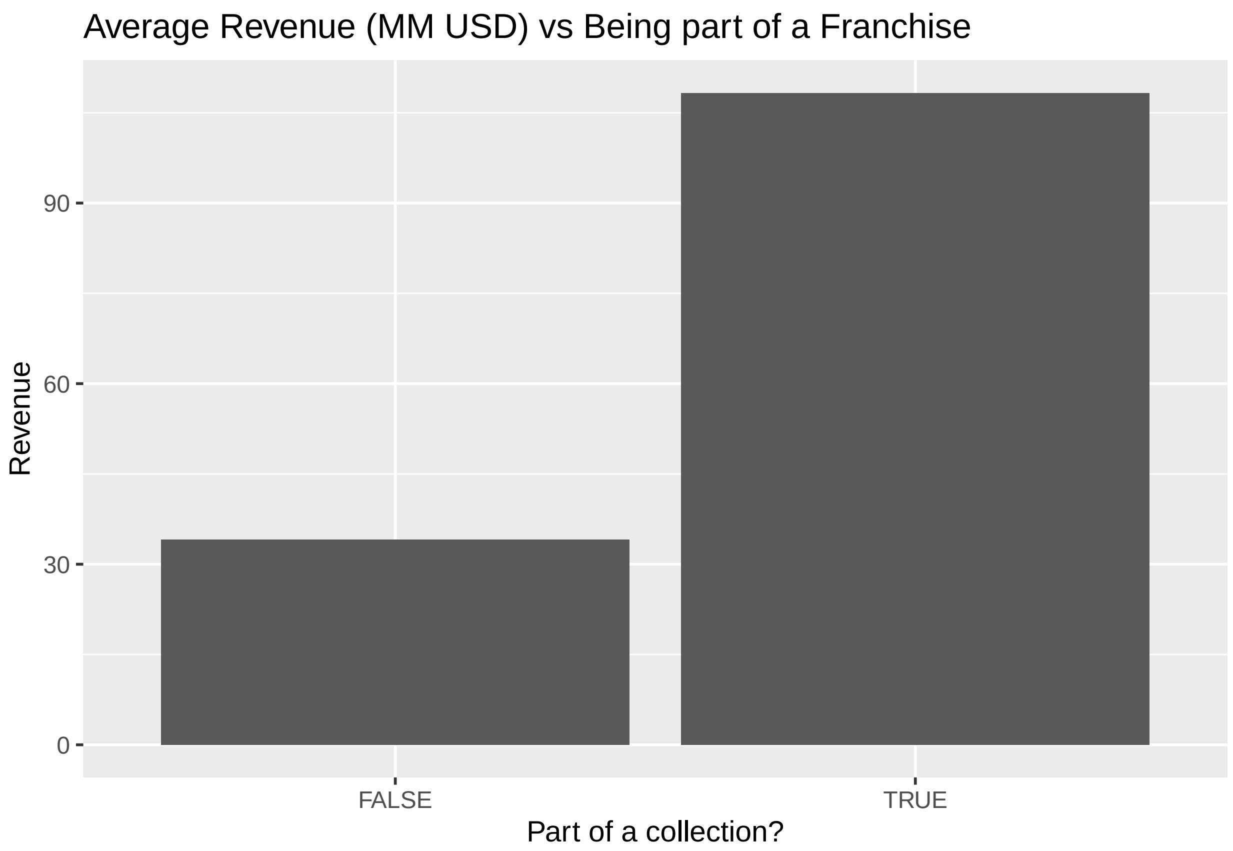




Genre There are a lot of different movie genres, and a movie can be considered to have multiple of those. We do not have very specific expectations regarding which genres may have a significant effect on revenue.

Is it part of a franchise/collection? There have always been movies that have been part of something bigger. Successful movies tend to spawn sequels, or even entire series. Just being associated with something greater may help a movie do better in the box office.

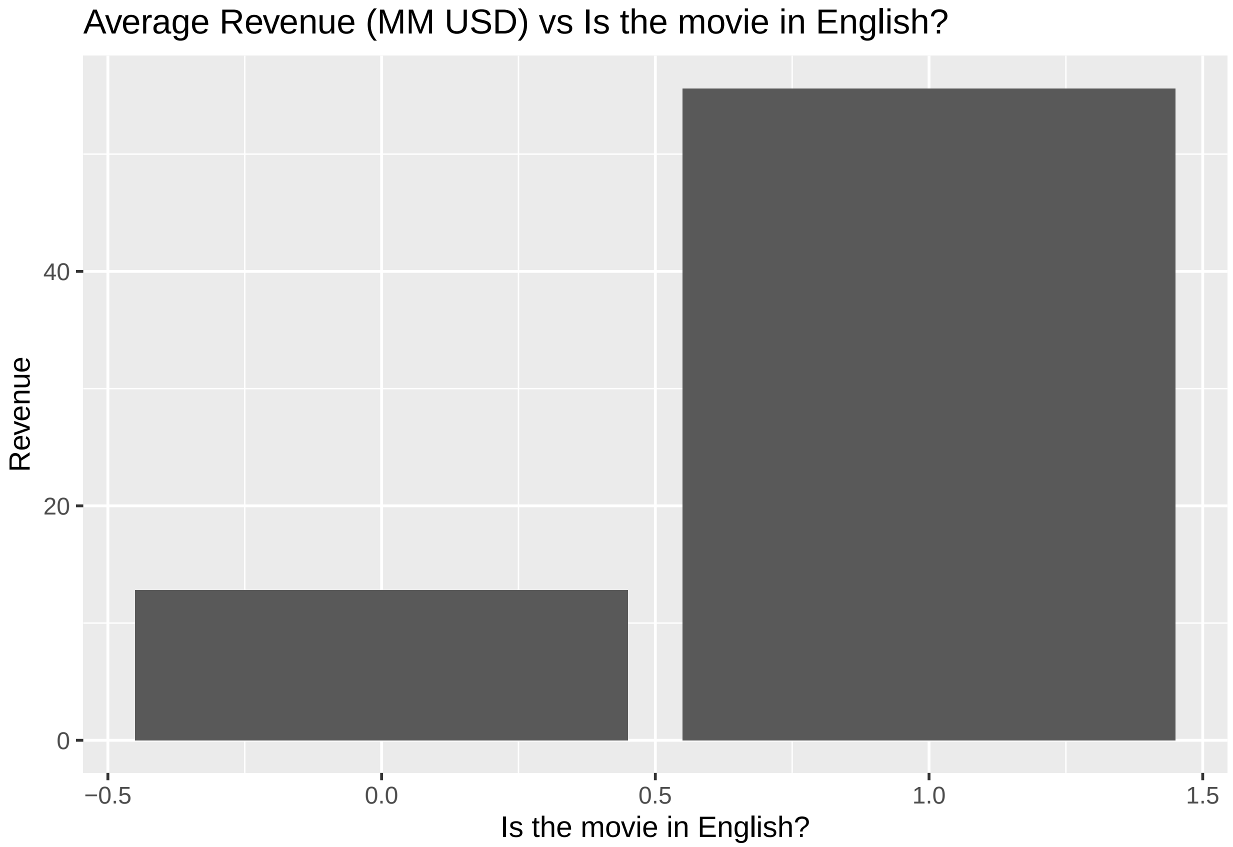
An initial EDA shows that movies that belong to a collection have a significantly higher average revenue.



Where is the movie from? The global box office is not evenly distributed. While this uneveness has decreased over the years, the American audience still makes up the biggest chunk of the global box office.

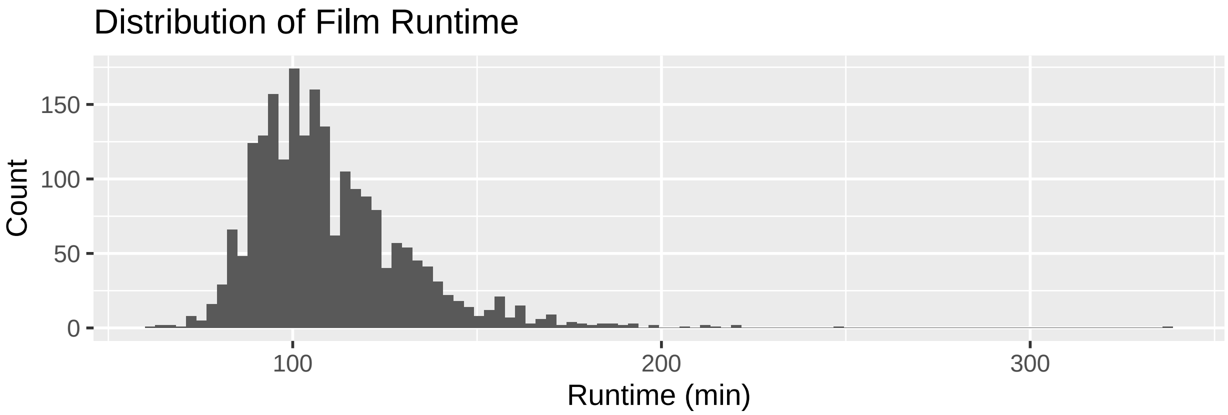
Because of this we decided to add a dummy variable that uses a movie's original language. If the movie's original language is in English, then it will have a value of 1, and if it isn't, then 0.

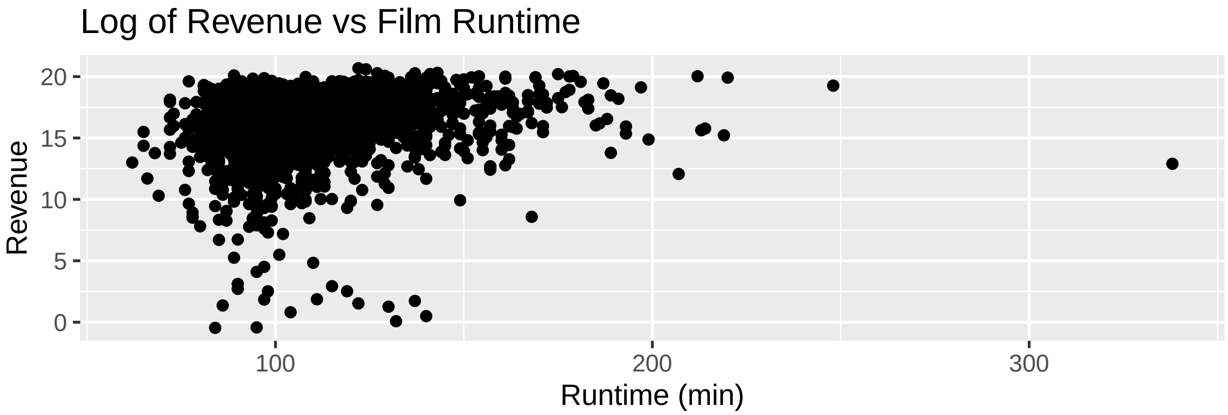
Our initial EDA shows us that there is indeed a clear difference between movies with english as their original language and those that have another language.



How long is the movie? The final covariate that we are considering adding to the movie is runtime. As seen in the distribution chart, most movies tend to be around the 100 minute mark.

If we look at the relationship between log of revenue and runtime, we'll notice that movies with a longer runtime tend to be more grouped together, and their average revenue seems to be higher. While movies closer to the 100 minute mark have a wider range of revenue.



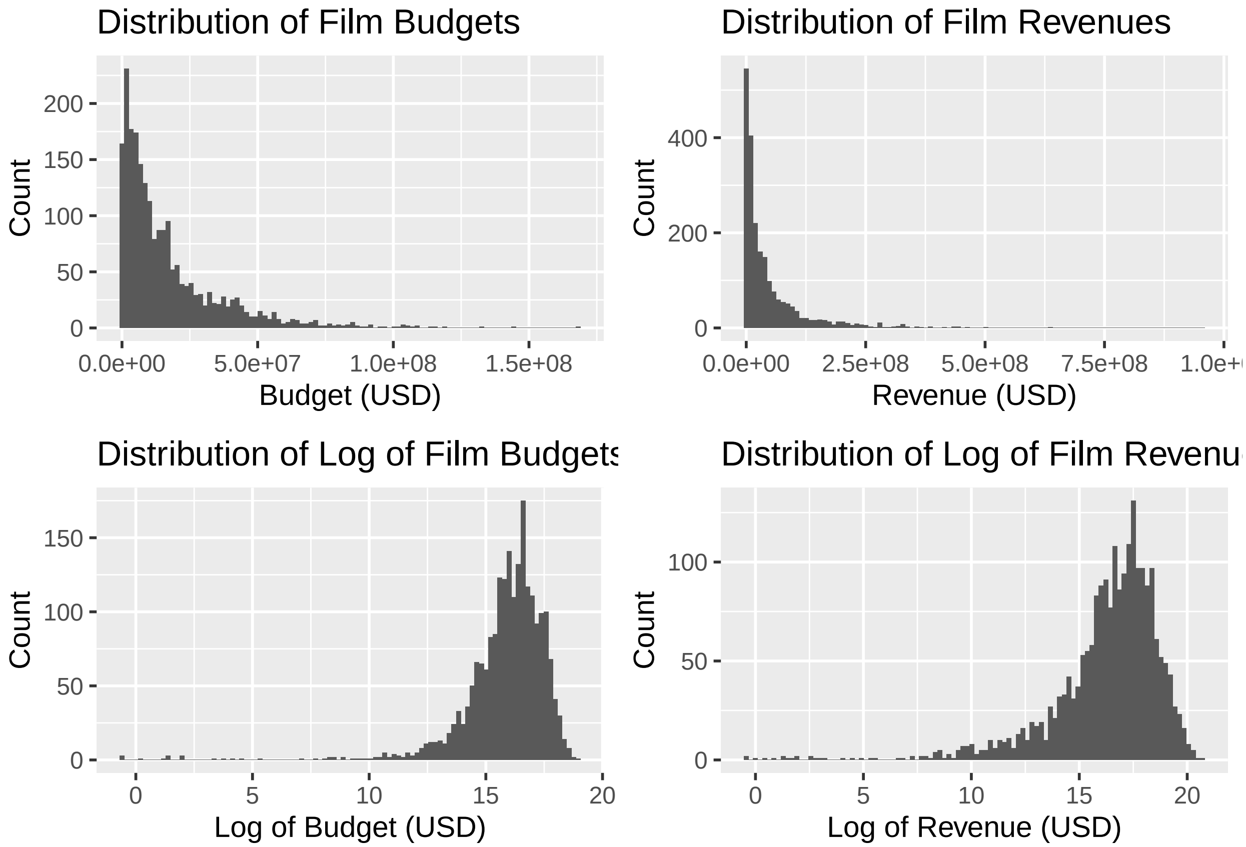


Statistical Model

We're looking to see which model and variables will be a good approximation for future films' natural logarithm of revenue(in U.S. dollars). We will be building 4 predictive models to help assess our research question of which variables play a more significant role in box office revenues to help the industry recover from the decline in the last year. We start with a simple model with just the `budget` variable.

In the Log Revenue over Budget graph earlier in the report we can see a relationship between budget and film success measured in revenue. For the 2 following models we add covariates to control for additional variables that we expect to be significant. And the final 4th model will include only the variables that proved to be statistically and practically significant.

Variable that we're most interested in is budget β_1, which measures a percent increase or decrease in revenue for every 1% increase in budget. We applied a natural logarithm transformation to both variables because the distribution was skewed towards lower variables(shown in the graph below). The transformation spread out the values, and the results show that the original data has a more log-normal distribution, making this transformation a good choice to improve our model.



Two of the covariates that we will be adding are movie genre and release month that will have multiple categories within each variable. Since we will be assessing multiple t-test results we may want to account for p-values that may show significance only because we have numerous categories that we're testing against. We will be multiplying the p-value by the number of genres or months in the model.

1) Model 1: Natural logarithm of revenue on a constant and natural logarithm of budget

$$log(\text{revenue}) = \beta_0 + \beta_1 log(\text{budget})$$

2) Model 2: Natural logarithm of revenue on a constant, natural logarithm of budget and genre

$$log(\text{revenue}) = \beta_0 + \beta_1 log(\text{budget}) + \beta_2\text{genre}$$

3) Model 3: Natural logarithm of revenue on a constant, natural logarithm of budget, genre, runtime, actor_score, director_score,release month, indicator variable of native english film and belongs to a collection(if the film is part of a sequel)

$$log(\text{revenue}) = \beta_0 + \beta_1 log(\text{budget}) + \beta_2\text{genre} + \beta_3\text{runtime} + \beta_4\text{actor score} + \beta_5\text{director score} + \beta_6\text{native english film} + \beta_7\text{release mo}$$

4) Model 4: Natural logarithm of revenue on a constant, natural logarithm of budget, significant genre, runtime, actor_score, director_score, significant release month, indicator variable of native english film and belongs to a collection(if the film is part of a sequel)

$$log(\text{revenue}) = \beta_0 + \beta_1 log(\text{budget}) + \beta_2\text{genre}_s + \beta_3\text{runtime} + \beta_4\text{actor score} + \beta_5\text{director score} + \beta_6\text{native english film} + \beta_7\text{release m}$$

- The subscript s for significant variables

Results

We're presenting a regression table below to compare results and find variables that prove to be significant.

Table 1:

	Dependent variable:			
	log(revenue)			
	(1)	(2)	(3)	(4)
log(budget)	0.881***	0.846***	0.754***	0.763***
	(0.022)	(0.023)	(0.024)	(0.024)
comedy		−0.031	0.113	
		(0.108)	(0.105)	
drama		−0.288***	−0.172*	−0.217**
		(0.103)	(0.101)	(0.093)
family		0.283	0.243	
		(0.182)	(0.176)	
romance		0.124	0.188*	0.259**
		(0.117)	(0.113)	(0.106)
thriller		−0.208*	−0.125	
		(0.113)	(0.108)	
animation		−0.093	0.178	
		(0.230)	(0.227)	
adventure		0.494***	0.198	0.227**
		(0.131)	(0.128)	(0.115)
horror		0.516***	0.440***	0.307**
		(0.152)	(0.149)	(0.138)
music		0.321	0.348	
		(0.251)	(0.240)	
crime		−0.006	−0.058	
		(0.125)	(0.121)	
sci_fi		−0.239*	−0.247*	
		(0.144)	(0.138)	
action		0.123	0.094	
		(0.114)	(0.110)	
war		−0.063	−0.097	
		(0.232)	(0.222)	
western		−0.321	−0.185	
		(0.360)	(0.345)	
fantasy		−0.065	−0.079	
		(0.161)	(0.154)	
foreign		−2.248***	−2.175***	−2.118***
		(0.604)	(0.582)	(0.574)
mystery		0.147	0.155	
		(0.162)	(0.155)	
history		0.122	0.032	
		(0.211)	(0.209)	
documentary		−0.571	−0.317	
		(0.483)	(0.463)	
runtime			0.010***	0.008***
			(0.002)	(0.002)
actor_score			0.226***	0.213***
			(0.057)	(0.056)
director_score			0.169***	0.175***
			(0.047)	(0.047)
native_english_film			0.386***	0.345**
			(0.145)	(0.141)
february			0.251	
			(0.217)	
march			0.198	
			(0.216)	
april			0.398*	0.257*
			(0.214)	(0.154)
may			0.267	
			(0.219)	
june			0.541**	0.431***
			(0.213)	(0.148)
july			0.720***	0.600***
			(0.218)	(0.158)
august			0.289	
			(0.211)	
september			0.049	
			(0.198)	
october			−0.077	
			(0.203)	
november			0.150	
			(0.220)	
december			0.599***	0.482***
			(0.208)	(0.141)
belongs_to_collection			1.093***	1.106***
			(0.107)	(0.106)
Constant	2.298***	2.889***	2.199***	2.437***
	(0.355)	(0.387)	(0.451)	(0.407)
Observations	2,161	2,161	2,161	2,161
R^2	0.421	0.441	0.495	0.489
Adjusted R^2	0.420	0.436	0.486	0.486
Residual Std. Error	1.990 (df = 2159)	1.963 (df = 2140)	1.873 (df = 2124)	1.874 (df = 2145)
F Statistic	1,567.282*** (df = 1; 2159)	84.396*** (df = 20; 2140)	57.824*** (df = 36; 2124)	136.943*** (df = 15; 2145)
Note:				*p<0.1; **p<0.05; ***p<0.01

All models are overall significant when looking at F-statistic which means that our linear regression model provides a better fit to the data than a model that contains no independent variables. As we're adding more predictors into the model we want to also evaluate the Adjusted R2 since that metric will account for predictors that improve the overall model more than expected by chance. Across the model we're seeing an increase in Adjusted R2 from most simple model to the most complex, which means that independent variables are capturing the variation in the dependent variable quite well and the addition of covariates is improving the model.

We found that quite a few variables in our consideration proved to be statistically significant. Budget, our main variable in question in the first model is significant because the p-value is below 0.05 threshold and is showing practical significance with a high coefficient. Since both variables were transformed using natural logarithms we can conveniently interpret the results in percents. For every percent increase in budget will have 0.81% increase in revenue. Since this could mean that for every additional dollar in the budget the return will be 80 cents. Which is why we want to interpret additional variables to understand what is the differentiating factor between profitable and not profitable movies.

In the second model when we include genre we have to interpret the statistical significance differently to avoid getting a significant genre by chance since we're evaluating a large number of genres. To make the significance stricter we will multiply the p-value for genre by 19, since we have 19 coefficients for genre. We conclude that drama may only be significant by chance because we ran multiple t-tests. And we can say that other variables, adventure, horror and foreign genres are significant predictors of revenue.

Table 2: Model 2 p-values

	P-value
(Intercept)	0.0000000
log(budget)	0.0000000
comedy	0.7744494
drama	0.0051260
family	0.1210612
romance	0.2897918
thriller	0.0658728
animation	0.6860743
adventure	0.0001771
horror	0.0006893
music	0.2007542
crime	0.9614200
sci_fi	0.0984315
action	0.2769806
war	0.7850598
western	0.3722254
fantasy	0.6847735
foreign	0.0002019
mystery	0.3620167
history	0.5625699
documentary	0.2368716

We considered evaluating the release months in the same way to make sure we're not proving a variable to be significant by chance. We'll multiply each p-value by 12 because there are 12 month variables in the model. Which leaves us with July and December as a significant variables, where June doesn't meet that level of significance and only July and Dcember are proving to be significant.

Table 3: Model 3 p-values

	P-value
(Intercept)	0.0000011
log(budget)	0.0000000

	P-value
comedy	0.2832102
drama	0.0897364
family	0.1666124
romance	0.0949195
thriller	0.2482645
animation	0.4325523
adventure	0.1224839
horror	0.0032266
music	0.1475914
crime	0.6317319
sci_fi	0.0746733
action	0.3926721
war	0.6614127
western	0.5918472
fantasy	0.6102344
foreign	0.0001889
mystery	0.3174618
history	0.8777906
documentary	0.4929422
runtime	0.0000582
actor_score	0.0000736
director_score	0.0003385
native_english_film	0.0077521
february	0.2468821
march	0.3573659
april	0.0631721
may	0.2226595
june	0.0110183
july	0.0009918
august	0.1715622
september	0.8049626
october	0.7026378
november	0.4948905
december	0.0040271
belongs_to_collectionTRUE	0.0000000

Some additional variables that proved to be a good predictor of revenue are Director Score, Actor Score and Native English Film, Belongs to Collection as well as Runtime with a caveat that runtime is not practically significant with a very low coefficients which means that there will be insignificant increase to revenue with every additional minute of runtime. On the other hand we want to pay attention to variables that are statistically significant and have a high practical significance, such as belonging to a collection of movies, having a higher budget.

Our suggested final model will be model 4 that will have practically and statistically significant variables.

Limitations

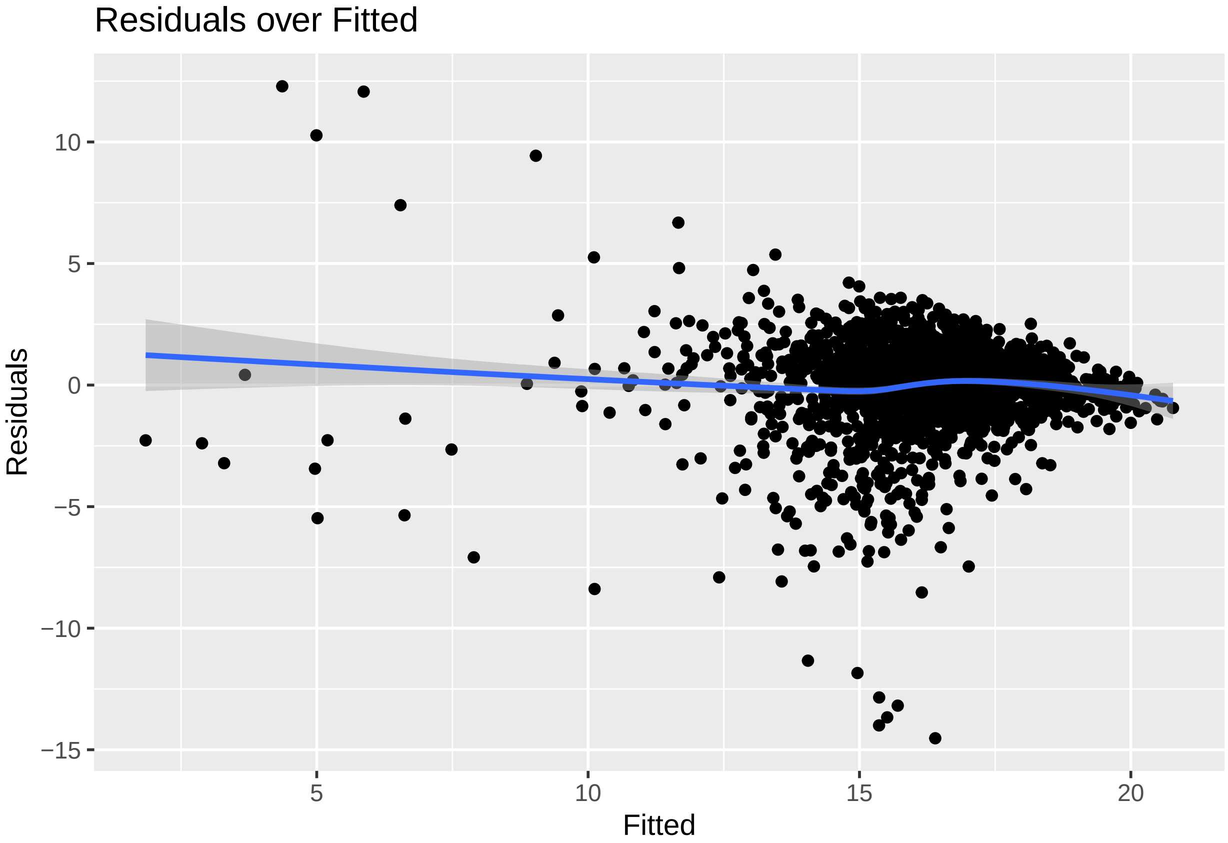
Statistical limitations

We will evaluate this model against 2 large model assumptions since we have a sample of well above 100. The first is IID and presence perfectly coliniear variables.

1. For the IID assumption we want to make sure that our samples are Independent and Identically

Distributed. It is hard to find a sample that is perfectly independent. Our data source is coming from The Movie Database (TMDB) which is a community built movie and TV database including international movies. It's one of the largest community managed database of movies, which should provide a good collection of movies including international releases. And we used a sample of approximately 2000 randomly selected movies that were compiled for a Kaggle competition. The randomized sample should help with the assumption of independence. And there are no signs to believe that the underlying distribution for samples in non-identical.

Although multiple factors could cause the movies to be dependent on each other. For example movies that are produced as part of the sequel do dependent on the revenue and success of the previous releases. There could be multiple other scenarios that can influence other movies in the sample, and we're seeing some pattern in the Residuals vs. Fitted plot below which also may indicate error terms not to follow a constant variance, but given random sample that is drawn from a large database we will consider this to satisfy the IID assumption.



2. The second model assumption is that there is no perfect colliniarity between variables which is proven by the fact that none of the variables were dropped from the model when we were fitting the model. In the presence of multicolliniariry the `lm` function in R would drop a variable and not display a coefficient value or significance.

We are meeting the 2 large model assumptions that are necessary to prove the model is reflecting a true information about the world and that we are reporting a results and coming to a conclusion that will be useful in the industry.

Structural limitations

Although we've considered a multititude of variables in our model there are some limitations, our actor and director scores are measured only for top 100 actors, actresses and directors at a time when in a movie industry and through social media the popularity of a cast and crew could evolve.

Omitted Variables

Some of the structural limitations of the model are omitted variables. If there is presence of another significant predictor that we don't measure or include in the model that influences the revenue. We suppose some of those variables are quality and promotion budget or strategy. It's harder to measure the true quality of a movie and in our model we use more of a proxy for quality through evaluating actor and director score as well as budget, which may not capture the whole measurement of quality. Another omitted variable is marketing budget since we only have the production budget in our dataset. Promotional strategy and budget could be a significant influence on revenue, since more popular movies have more box office revenues.

While we have covered a lot of possible variables for our model, it's almost certain that there are more variables that could explain a movie's financial success. Here we will cover the ones we consider to be the main omitted variables.

The Quality of the Movie So far we have discussed more objective measurements, such as the budget, release date, cast and crew. However, these are not the only factors that can affect a movie's success.

Word of mouth has been studied as an important factor in a movie's success [3].

Average movies tend not to create a lot of word of mouth, so they are unlikely to benefit from that effect, while excellent movies will get talked about and recommended.

We do run into a problem when trying to include it in our model. While we could use rankings and ratings as proxy variables, this research focuses on predicting a movie's revenue, and these kinds of rankings and ratings tend to come out after a movie is released, making including them on the research pointless.

We have decided that the best way to capture the effect of the quality of a movie would be through the Director Score variable. A movie by a great director is probably the safest bet someone can make when it comes to the movie's quality.

We would also expect some of its effect to be captured by the Actors score, and its budget, both are things are associated with a movie's quality.

Since the relationship is positive, and we assume that so is the Actor's score and the budget, then it would be safe to assume that omitting this variable is pushing these coefficients away from zero, thus overestimating their effect on a movie.

The Marketing strategy and budget Another possible omitted variable is the movie's marketing. If nobody hears about a movie, they are much less likely to go see it. A good marketing strategy can get people to the movie theatres, and it can also maximize the effect of an actor's appeal, as was the case with the Zendaya example.

Quantifying a "good" marketing strategy is complicated, but we could use the movie's marketing budget as a proxy variable. However, the marketing budget is not part of our dataset's budget column, as the marketing and distribution budget are not usually part of the movie's reported budget, which tends to focus on just production costs.

Having said that, it's very likely that a marketing budget will be heavily correlated with a movie's production budget. So by including the budget, we are including a proxy variable for a movie's marketing strategy.

Just like with the quality, we would expect a good marketing strategy to have a positive effect on the revenue, as well as a positive effect on related variables such as the Actors Score.

Conclusion and Discussion

Considering the limitations of our study, we found reasonable results in trying to answer whether we can build a predictive model to measure movie box office success. We used one of the largest movie databases and

[3] Liu, Yong. "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue." Journal of Marketing, vol. 70, no. 3, American Marketing Association, 2006, pp. 74–89, http://www.jstor.org/stable/30162102.

selected a random sample. We focused on the variables that could be obtained before movie production and release to justify using such a model in the movie consideration process. We identified a relationship between movie budget and revenue and were able to find additional significant covariates. Finally, we considered the limitations of the covariates and the model and have tested some of the underlying large sample model assumptions.

To conclude our findings and provide more insight into the movie industry. Although we established a clear relationship between movie budget and revenue, it may not be the only variable to produce a blockbuster. We're finding that movies that are part of a sequel tend to perform well, which may also be explained by the fact that films in a sequel have proven to be popular and were given a budget for more movies to be produced.

We confirmed the seasonality of movie releases when certain months perform better than others. Summer proved to be a good movie season, with July producing more considerable revenues than other months in a year, as well as December being a good month to release movies. However, as mentioned in the Data section, there may be some self-selection element with seasonality, with production companies choosing to release high budget movies during those months.

We also see that native English films tend to do better while foreign films may decrease future profitability.

We hypothesized that movie genre might be a good predictor and guiding factor for movie industry decision-makers. When considering movies in the horror or adventure genre, it could be a safer bet for return on investment.

And finally, the actor score and director scores for measures of popularity of the cast and crew have proved to be significant, with more popular actors bringing more revenue and followed by famous directors.

If we were to continue with this research, we consider that it would be interesting to separate our dataset into time periods, as movie audiences can vary a lot throughout the years, and as mentioned before, our dataset goes back almost 100 years. In particular, perhaps a research that focuses just on the last 20 years may be more practically significant than one that uses data from all of movie history. We would also need to come up with a way to categorize actors and directors from just this time period in include that variable in our analysis.

Another interesting point to consider would be focusing just on Opening Weekend Box Office. One limitation of our research was the omitted variable of "Quality". While we used other variables as proxys that we hoped would capture the effect of quality, it's still not the perfect solution.

However, if we were to use Opening Weekend Box Office as the dependent variable, instead of total Box Office revenue, then the "Quality" effect would likely be not as meaningful. We assume this would be the case since there wouldn't be enough time for word of mouth to get out and attract more movie goers.