W203 Lab 2

Mark's Minions Group 1 Sec 3
Abel Ninan, Collin Chee, Dan Nelon, Sunny Shin

1. Importance and Context

There are many aspects of what makes a movie rise to the top of the box office. Genre is one aspect that can be considered by viewers when deciding on a movie. The genre creates a certain level of expectation for the audience, opting for a specific experience. Consequently, the size of the potential audience may vary depending on the genre. The global movie industry generates hundreds of billions of revenue each year and our client, Minions Production, posed us with a question of whether there is a relationship between the genre of a movie and the revenue.

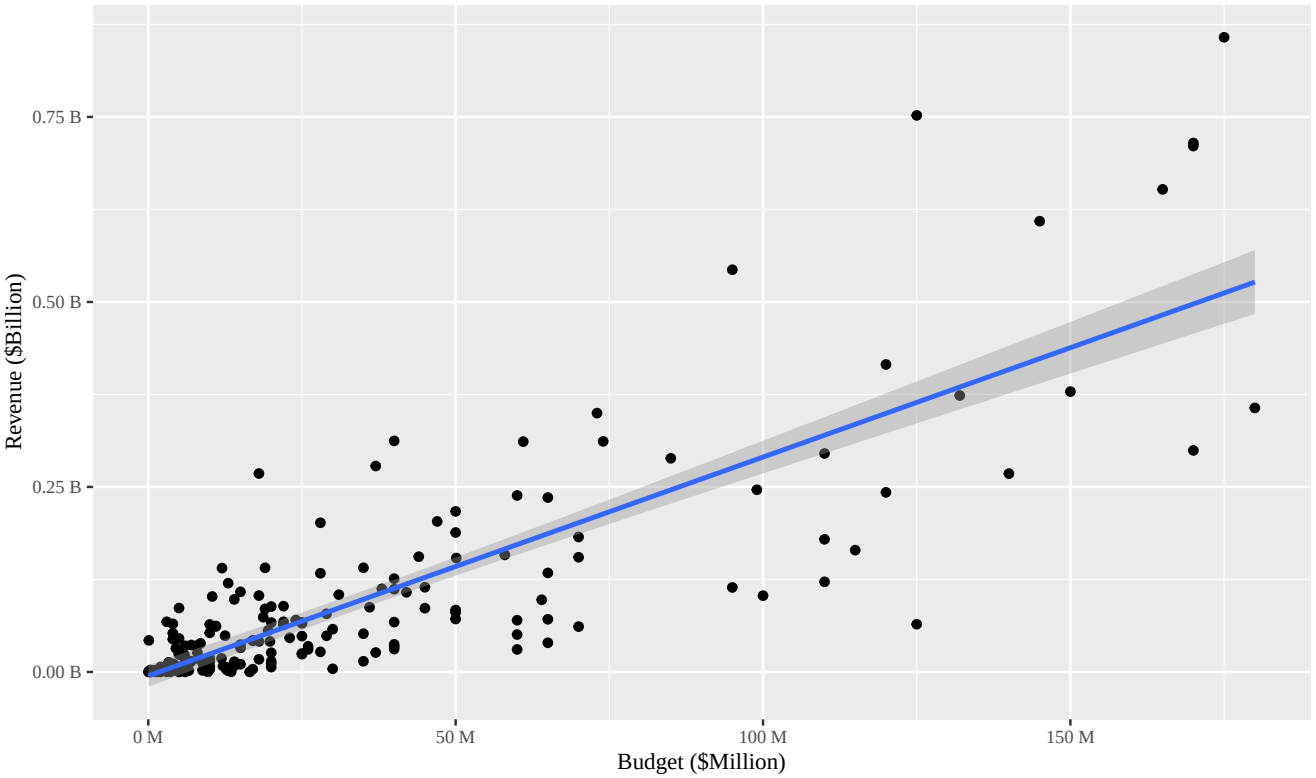
Our study seeks to provide an answer to this question by exploring the effect of various factors on revenue using statistical models. In particular, we build linear regression models to examine the relationships between revenue and genre while controlling for movie budget and release year. Answering this question will play a huge role in the next phase of movie productions for our client. If a certain genre of a movie is showing a positive trend, there may be an opportunity for investment in that specific genre, in turn maximizing revenues.

2. Data and Methodology

Our analysis leverages data from Kaggle dataset pulled from The Movie Database (TMD) web API. The data was collected by GroupLens Research that specializes in recommender systems, online communities, mobile and ubiquitous technologies, digital libraries, and local geographic information systems. The dataset contains 45,466 movies from 1989 through 2017 with basic information such as title, budget, revenue, release dates, languages, IMDB movie ID, status, production countries and companies. Knowing the fast paced nature of the movie industry, our team had determined to narrow down the dataset to only more recent years to better fit to our client's interest. Additionally, past movies produced with little to no budget would not help answer our client's questions as they were looking for investment opportunities and therefore those movies were also removed from consideration.

Based on our exploratory data analysis (EDA) such as plotting the distribution of some predictor variables such as genre, year and budget, we discovered that 2014, 2015 and 2016 have the most reliable data, which helped narrow down our data set to approximately 660 observations. Given the data size and the number of different genres, we decided to begin by splitting the dataset into a 30% exploratory set and a 70% confirmation set. We used the exploratory dataset to discover any patterns, spot anomalies and check assumptions to gather as many insights from it.

Figure 1. EDA − Revenue vs. Budget



To operationalize the budget, we used the total amount of money listed under budget. As shown in Figure 1 indicating budget vs revenue, we can see there is a visual linear trend between the two variables. Therefore, we did not have to perform any transformations to the budget when we incorporate it into our regression model.

During the EDA phase, we also learned that a film can belong to multiple genres, so we instead assigned an indicator variable of just two values: 0 and 1 one hot encoding method to indicate the absence or presence of a discrete genre. In our observations, there were the following 17 genres: *Drama, Comedy, Fantasy, War, Western, Science Fiction, History, Romance, Family, Mystery, Animation, Crime, Thriller, Action, Horror, Documentary, and Adventure.*

Finally, we built a linear regression model of the response variable revenue vs. genre and budget to explore the relationship between the predictor and response variables. We will specifically use an ordinary least squares regression instead of a classical linear model due to our large sample size. The original regression model contained all 17 different genres, budget, and release years.

After some initial testing, we found that only budget and the animation genre produced significant coefficients.

Additionally, considering the different levels of budget required to produce certain movies (due to additional production costs, such as special effects), an interaction term between budget and genre was considered. The linear regression was as follows. $\hat{Revenue} = \beta_0 + \beta_1 Budget + \beta_2 Animation + \beta_3 (Animation * Budget)$ \end{center}

3. Results

Table 1: OLS model of revenue based on total budget and animation genre

	Dependent variable:		
	Revenue		
	(1)	(2)	(3)
Budget ($)	2.96***	2.84***	3.76***
	(0.14)	(0.15)	(0.14)
Animation		60,293,601.00**	55,734,381.00
		(23,706,346.00)	(44,044,819.00)
I(Budget∗Animation)			0.23
			(0.49)
Constant	−5,267,642.00	−5,595,362.00	−15,103,926.00*
	(7,544,067.00)	(7,442,603.00)	(8,798,104.00)
Observations	199	199	461
R^2	0.68	0.69	0.66
Adjusted R^2	0.68	0.69	0.66
Residual Std. Error	82,412,363.00 (df = 197)	81,291,772.00 (df = 196)	148,244,676.00 (df = 457)
F Statistic	426.04*** (df = 1; 197)	222.17*** (df = 2; 196)	299.27*** (df = 3; 457)
Note:			*p<0.1; **p<0.05; ***p<0.01

As Table 1 shows above, the budget was highly statistically significant for every model for all three models. Point estimates range from 2.84 to 3.76, which indicates that for every $1.00 we invest in a movie's budget, we could predict that we would be able to generate anywhere from $2.84 to $3.76 in revenue, with all else held constant. After testing all 17 different models, we concluded that animation was the only genre that showed statistical significance. Thus we narrowed down to just the animation index in the second model. The coefficient indicates that if we decide to create an animation movie as opposed to other movie genres, we should be able to generate an additional 60 million dollars in revenue, assuming budget stays constant. However, the variance suggests that this could range from about 37 to 83 million dollars. The third model is the revenue regressed on budget, animation and the interaction between the budget and animation. The interaction variable was included to test whether the changes in budget and animation together will impact the statistical significance of the model overall. The results indicate that it is not statistically significant, meaning the two predictor variables together do not necessarily have an impact on the response variable.

Table 2: Budget - Animation ANOVA results

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	459	1018860815264197427 2.00				
2	458	1004814001533671424 0.00	1	14046813730526003 2.00	6.40	0.0117

The ANOVA table shown above also indicates that during our initial training, there was an indicator showing that animation was significantly improving our explanation of revenue variance (Model 1: Revenue ~ Budget vs. Model 2: Revenue ~ Budget + Animation). Lastly, our model which includes an interaction indicates that there is no statistically significant interaction between budget and the animation genre.

4. Limitations

When creating our explanatory model, large sample model assumptions require an assumption of independent and identically distributed observations. This assumption is reasonably met since the characteristics and financial performance of one movie do not tell us anything about how another movie might fare at the box office. However, it must be noted that this assumption may be slightly violated due to the fact that some movies are sequels and prequels to other movies. Although every movie is taken from the same distribution of each year, we can definitely not rule out that some movies' budgets may be influenced by the contents of other movies. This is why we originally added the release date to account for not only movies having a longer time to generate revenue, but also to account for related films. Another noteworthy point that was acknowledged was the presence of the same actor (or numerous actors) in multiple movies. While this situation does indeed exist in the industry, data for this is challenging to collect and so we assumed that it would not uniquely interfere with our model since a single actor (or numerous actors) cannot work on a large number of movies simultaneously due to timing constraints.

Large sample model assumptions also state that the population distribution is described by a unique best linear predictor (BLP). In accordance with this assumption, our team does not see any evidence of highly skewed distributions in any diagnostic plot. Therefore, we can conclude that there is a finite amount of variance and that a unique BLP does in fact exist. To be safe, we looked at the variance inflation factor and saw that it is 1.025, indicating that there is no perfect collinearity among the predictors.

When addressing structural limitations, there are multiple omitted variables that could potentially influence our estimates. Monetary investment being divided across movie elements such as casting, graphics, set, marketing campaigns (typically excluded from production budget completely), filing for rights, and other expenses when bringing a movie to the box office results in a few examples of variables that can impact our model. Specifically, the marketing budget allotted to a film is an example of an omitted variable in the context of this study. The marketing budget of a film would be positively correlated with both the revenue generated and the overall budget of said film. Therefore, the omitted variable bias would be positive in nature. Furthermore, when taking into consideration that the overall budget is also positively correlated with the revenue, we end up with the direction of bias as moving away from 0. A similar analysis holds for other variables such as actors, film directors, and production companies.

5. Conclusion

This study sought to predict the financial revenue generated for a movie at the box office through the genre that it is classified as. Of the genres examined, only animation was determined to be a statistically significant feature when explaining the variance of revenues generated by movies. We were also able to determine that the budget for a movie was a statistically significant feature that was present in all of our proposed models. Higher budget movies do tend to attract larger audiences and have larger marketing budgets, therefore leading to larger revenues since they can afford to drive up demand, have top talent casting, and have high fidelity graphics.

Future research may desire to collect data on movies exclusively released in other countries (besides the USA) around the world. A similar analysis like the one done in this study could help determine which genres generate higher revenues in those countries. Ultimately, companies who release movies globally could focus their marketing efforts on countries where the most influential genre mirrors that of the movie that is about to be released. This would allow for maximized revenue from a film and incentivize companies to continue releasing films of a similar nature and subsequently marketing them to interested countries. Additionally, the research into this topic can be further advanced if one were to study the profitability (revenue minus expenses) of films across different genres with an added correction for monetary inflation over time. The statistically significant features of these explanatory models could prove to be vital information for a production studio seeking to gain a substantial competitive edge in the film industry.