LAB 2
Allie, Marque, Sarah, and Waqas

What makes a successful movie?

Table of Contents

Section 1: Introduction

The film industry is a multi-billion dollar industry that has only become more popular with time. Each generation improves the quality and creativity of what we can see on the big-screen, whether it was with the addition of sound to motion films or the CGI features of today. The film industry continues to push the boundaries for what audiences can enjoy; however, there are so many factors that can influence which films audiences choose to watch.

Many members of the film industry are working to produce the next big hit, which prompted our curiosity into which combination of film features will make a film the most successful. Though we acknowledge that there are some films created purely for the art and are not focused solely on profits, the majority of films released are looking to be as successful as possible and define their success based on their profits. As a result, we determined that the best way to judge the success of a film was through its gross income.

In the following report we focus on which genre will produce the highest gross income for films released in the United States after 1928. The data points used were sourced from IMDb as well as a supplemental Oscar director and writer list. It is important to note that for our data and in practice, genres are not specific to just a singular concept - many movies span multiple genres - romantic comedy, action adventure sci-fi, etc. This variation led us to also investigate which combination of genres produce the most successful films. We used other factors that could affect the gross income of a film as controls throughout our models, for example duration of a film and the year it was released. We completed this analysis through a causal model.

The remainder of this report is structured as follows. Section 2: Data and Research Design, we will discuss how we obtained and cleaned our data as well as how we determined our variables. Section 3: Model Building Process, we will discuss our models and how we determine which variables to add. Section 4: Results, we will discuss our key findings based on the final model determined in the previous section. Section 5: Model Limitations, we will discuss any parts of our model that could be cause for questions. Section 6: Conclusion, we will summarize our findings and discuss how this topic can be studied further.

Section 2: Data and Research Design

Our main goal for this research is to examine what variables make a film successful. For data on movies, we decided to begin with data collected from the Internet Movie Database (IMDb). This data was collected via IMDb's API and made available on Kaggle.com. This dataset consists of 70 variables across 85,855 movies spanning from 1911 to 2020. The ratings of movies are collected in a quantitative research method as users are invited to rate movies on a likert scale of 1 to 10. These totals are then converted into a weighted-mean rating to minimize a user from voting on a movie several times and having an outsized impact. The dataset is aggregated to one row per movie with movie details such as genre, directors, production companies, budget as well as review scores broken out overall and by various demographics(age, gender).

During our initial review of this dataset, we decided to narrow our dataset down to specifically what makes a film successful in the US. We limited the dataset to only movies that had one of the languages as English and one of the countries it was released in was the US. We also limited our dataset to movies where the budget and US gross income were reported in US dollars. After these conditions were applied, our dataset was limited to 6,498 movies. We also decided to reduce the number of variables under consideration. Many of the rating columns were subsets of the overall rating. Since we are not focusing on a specific demographic or age, we eliminated 40 rating variables. After we eliminated those columns we were left with the following variables:

- US Gross Income:Total film earning in the US
- Year: Year that the movie was released
- Genre: Categories that the film falls under (i.e comedy,drama,romance)
- Duration: Length of film in minutes
- Budget: Costs related to the development,production, and post-production of a film
- US Voters Vote: Number of votes from US voters

For our models, we wanted to define success based on US gross income. To that end, we established US gross income as our dependent variable throughout all of our models.

One of the enhancements that we wanted to make to the dataset is adding an indicator to show if a film contained an Oscar winning director or Oscar winning writer. We utilized a list of best director winners from the Oscars going back to 1928 and a list of best screenplay winners going back to 1957. We felt that best screenplay winners were appropriate because it awards a movie for not being based on previously published material. Because our Oscar dataset only went back to 1928, we limited our IMDb dataset to only movies created 1928 or later.

Additionally, we wanted to measure the impact of how popular a production company would have on the measure of success, US gross income. We added an additional variable, studio popularity, to measure the frequency of how often a production company appeared within our data. We also felt it was important to be able to derive genre specific models, so we created several genre variables. For example, if a movie was classified as comedy and sci-fi, both the comedy and sci-fi variables for this movie would be 1 while the rest of the genres would be 0.

As previously mentioned, our dependent variable is US gross income. To understand what makes a movie successful, we want to evaluate the relationship between income and variables such as genre,year and budget. We felt that an ordinary least square regression would be the most appropriate statistical process to evaluate our research question.

Section 3: Model Building Process

Section 3.1 Data Understanding

With all said and done, our end goal when building a model is to make it reproducible and accurate. Since our dataset is large enough, 6,498 observations, we decided to split our dataset into a training set and testing set. Training set will consist of 70% (n=4548) of the data, whereas the testing set will contain 30% (n=1950). We will perform all our model explorations and build on our training set and predict on the testing. From that, we will measure our model accuracy and how well it does on different datasets.

To help us answer what makes a successful movie, we decided to operationalize movie success by how much money it makes. In our data, we use a variable called "USA Gross Income" as our dependent variable. We will start off by exploring this variable.

Figure 1A below shows the distribution of US gross income for the movies in our dataset. Data seems to be skewed to the right so we will apply a transformation to help resolve

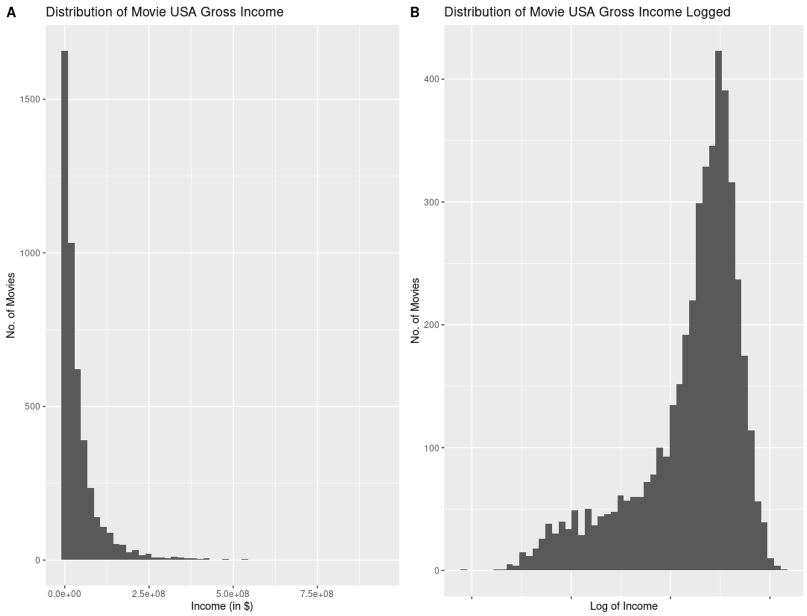


Figure 1. Histogram of USA Income, original (A) and transformed (B).

Figure 1B shows the income variable when log transformed. The data became more normally distributed, which we will use as our dependent variable when we build our model. No major outliers seem to be present as well.

Next, we will look at our independent variables individually and also observe the relationship between those and the dependent variable, gross income. There are several independent variables we would like to explore. First, we will look at our control variables; movie release year, movie duration and each genre type. We picked these as our controls because we think that these are the factors that mainly influence someone's decision on whether to see a certain movie or not in our data. We will also be looking at our movie budget, Oscar winning director, Oscar winning writer, movie studio popularity and US votes.

First up, we will look at our budget variable, which is how much budget a certain movie has. Figure 2A below shows the distribution of this variable in its original form. We see that it's skewed to the right. When we apply log transformation, this issue is mainly resolved and data is more normally spread, shown in Figure 2B. There seem to be some outliers, but nothing significant.

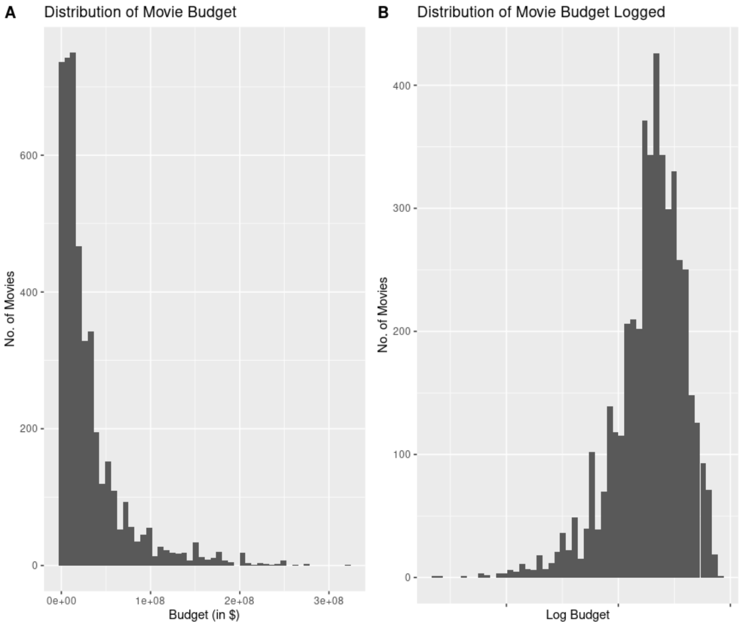


Figure 2. Histogram of budget variable, original (A) and transformed (B).

Next up, we will look at our movie duration variable. Figure 3A below shows the distribution. We see that it's normally distributed. When we apply a log transform, Figure 3B, we don't see any visible change so we will keep this variable as is. Additionally, no major outliers seem present.

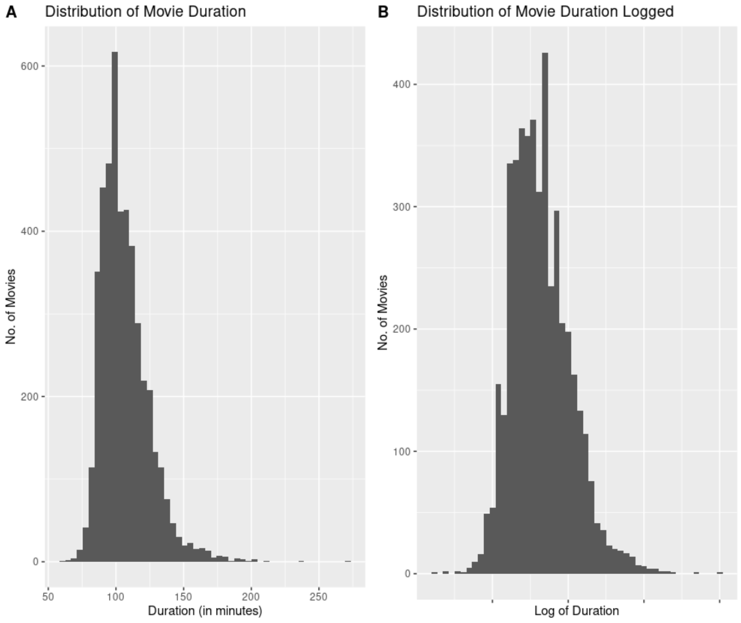


Figure 3. Histogram of movie duration, original (A) and transformed (B).

Studio popularity is another interesting variable we want to consider. We created this based on our data. We define studio popularity as the number of movies a particular movie studio has made. For example, for the movie Frozen, Walt Disney Animation Studios is the production company in charge. So we look into all our data and see how many movies Walt Disney Animation has made, and populate the studio popularity for the movie Frozen with that number. We did this for all the movies in our data. Figure 4A below shows the distribution of this variable.

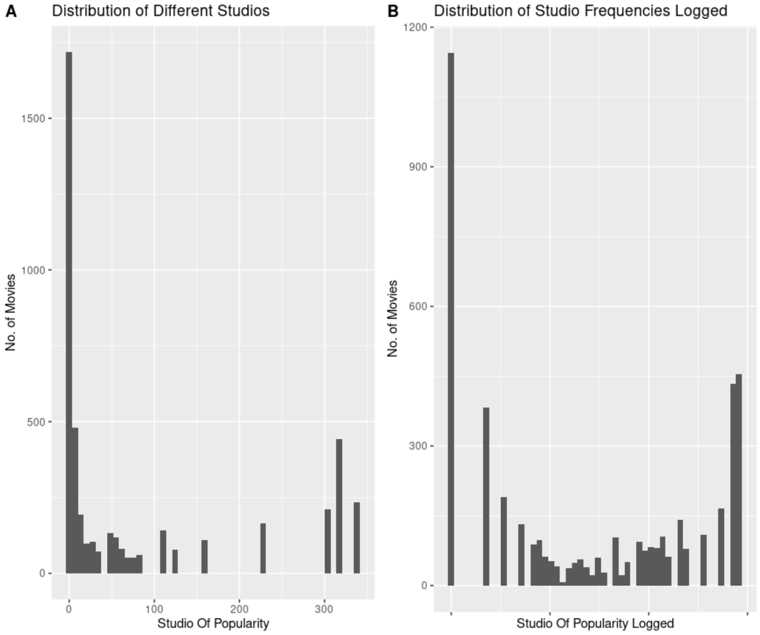


Figure 4. Histogram of studio popularity, original (A) and transformed (B).

From the histogram, we can see that the data is not evenly distributed. When we applied a log transformation, we can see that we get a little more spread in the middle, Figure 4B. It’s also clear we have a lot of studios that only made one movie, which were more of the independent, smaller companies.

Figure 5A below shows the distribution of the movie release year. The data is not skewed substantially but movies seem to be dated from 1975 on. Total income increases around in the 2000’s and starts decreasing. Inflation can also be playing a role in the increasing trend. There seems to be no difference in the distribution when we apply log transformation, shown in Figure 5B. Thus we will keep the year as is as well. There also seem to be a few movies made before 1960, but we will leave them in our data as our goal is to look at movie data from 1928 onward.

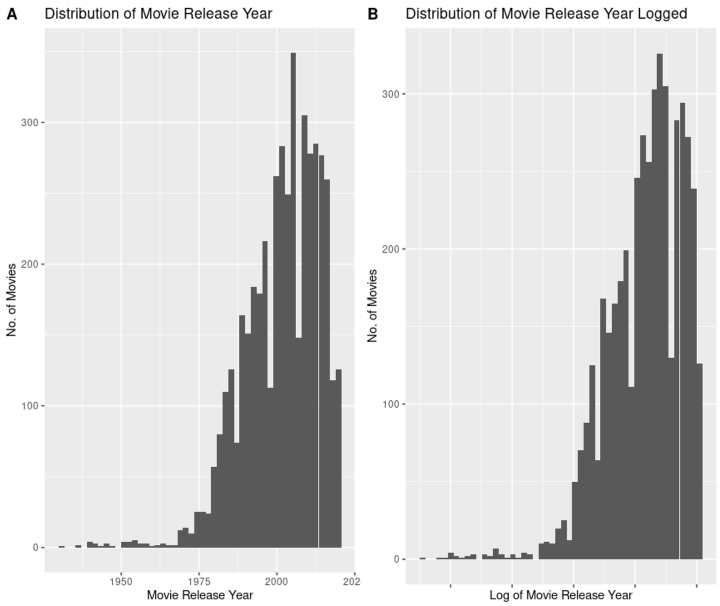


Figure 5. Histogram of studio popularity, original (A) and transformed (B).

Last but not least, we looked at the US votes variable, which was the number of total US Voters that voted on a particular movie on IMDB.

Figure 6A below shows the distribution of the US votes. We see that data is heavily skewed to the right. However, when we apply a log transformation, seen in Figure 6B, the distribution clearly improves and results in a normal spread.

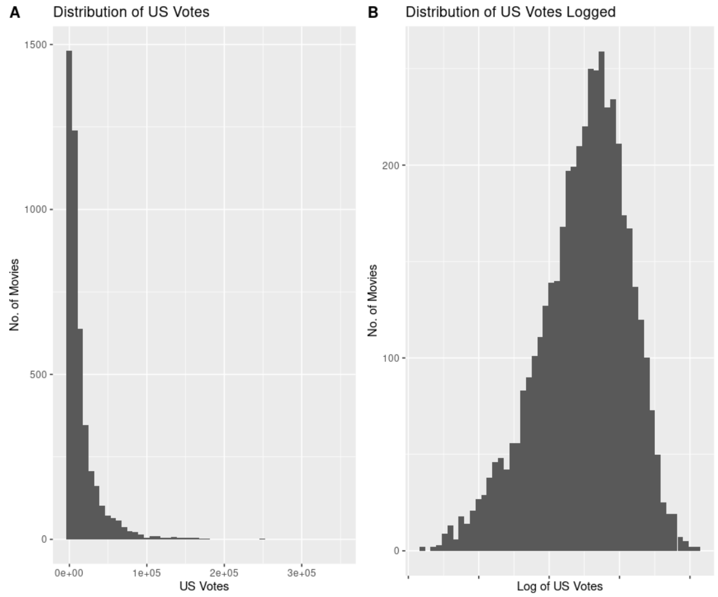


Figure 6. Histogram of US Votes, original (A) and transformed(B).

We also had to explore the relationship between the independent variables mentioned above with our dependent variable, USA Gross Income, all shown in Figure 7 below.

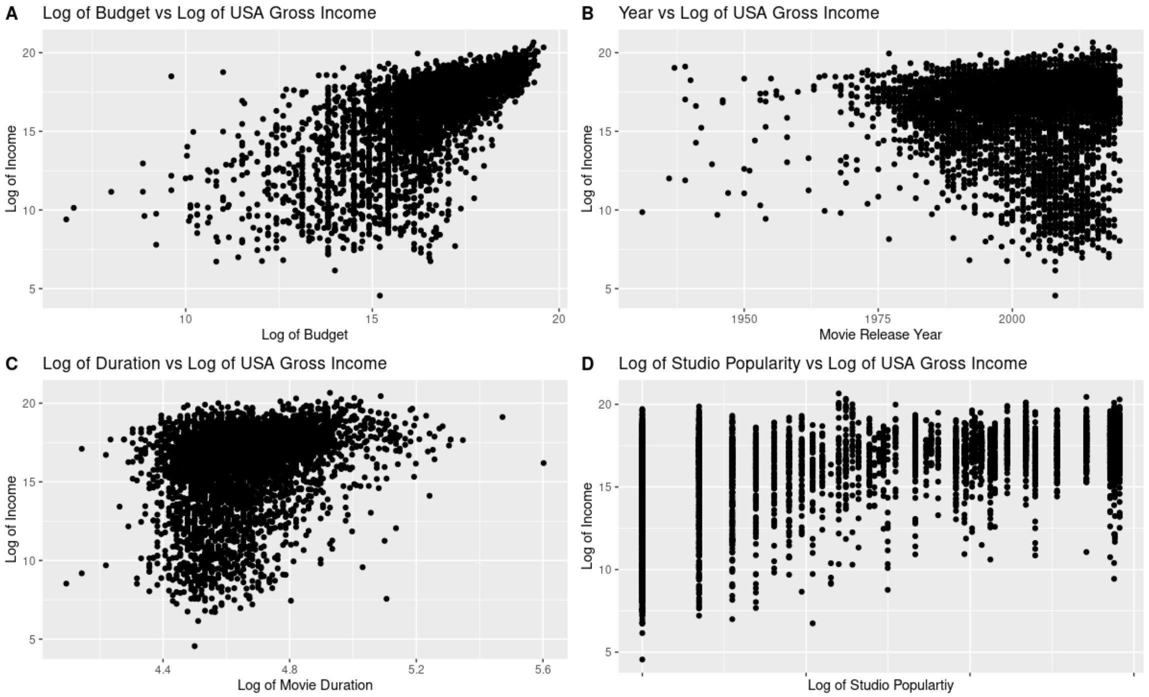


Figure 7. Scatterplots of the independent variables versus the dependent variable.

Figure 7A shows that movie budget and movie income, both with log transformation, have a strong linear relationship. This is expected as a movie that has a higher budget would most likely have a higher gross income. Infact, these two variables have a very strong positive correlation, .67. Figure 7B shows the relationship between movie release year and income, which exists a slight linear trend. Overall, more movies tend to have been made after 1975, with the majority of them increasing in total gross as years progressed. However, there are still a substantial amount of movies that grossed on the lower side, even as years passed. The correlation between these variables is also positive but not as strong, .12. Figure 7C shows the relationship between income and movie duration. Although not as linear as budget, it is more linear than the relationship between year and gross income. As movies become longer, they also increase in overall gross income. However, there are a substantial number of shorter movies as well with higher gross income. The correlation between duration and income is .29. Last but not least, Figure 7D shows the relationship between gross income and studio popularity. There appears to be a slight linear relationship, the more popular a studio, the more gross income their produced movie brings. This makes sense in theory because studio popularity tends to be a major influence on how much money a film makes. For example, a Universal Pictures film will be expected to bring in more money than a small independent studio. The correlation between studio popularity and income is also positive, .27.

We also looked at US gross income for each genre. We noticed that the distributions were highly skewed towards the left and they were spread out at a very wide range. In this case, taking log of US gross income was a natural choice after which the distribution looked normal for most of the genre as can be seen in the Figure 8 below:

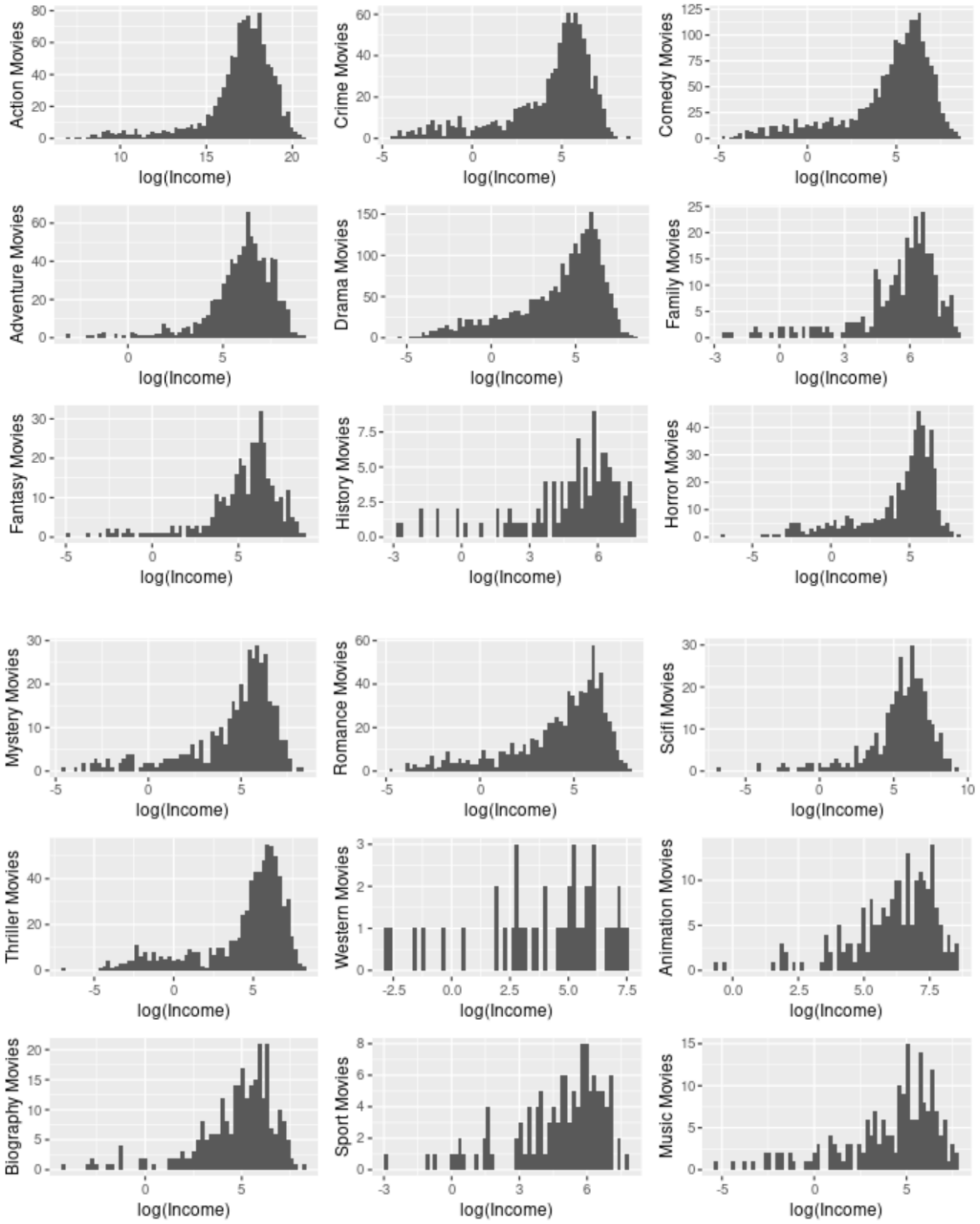


Figure 8. Histogram of log of movie gross income for each genre.

The average rate of return for a movie ranges from approximately 100% to 170% and it varies depending upon the genre of the movie. The lowest rate of return is for ‘War’ movies whereas the highest is for ‘Musical’ movies. Average budget, average income, and rate of return for each genre of the movie is plotted in Figure 9 below:

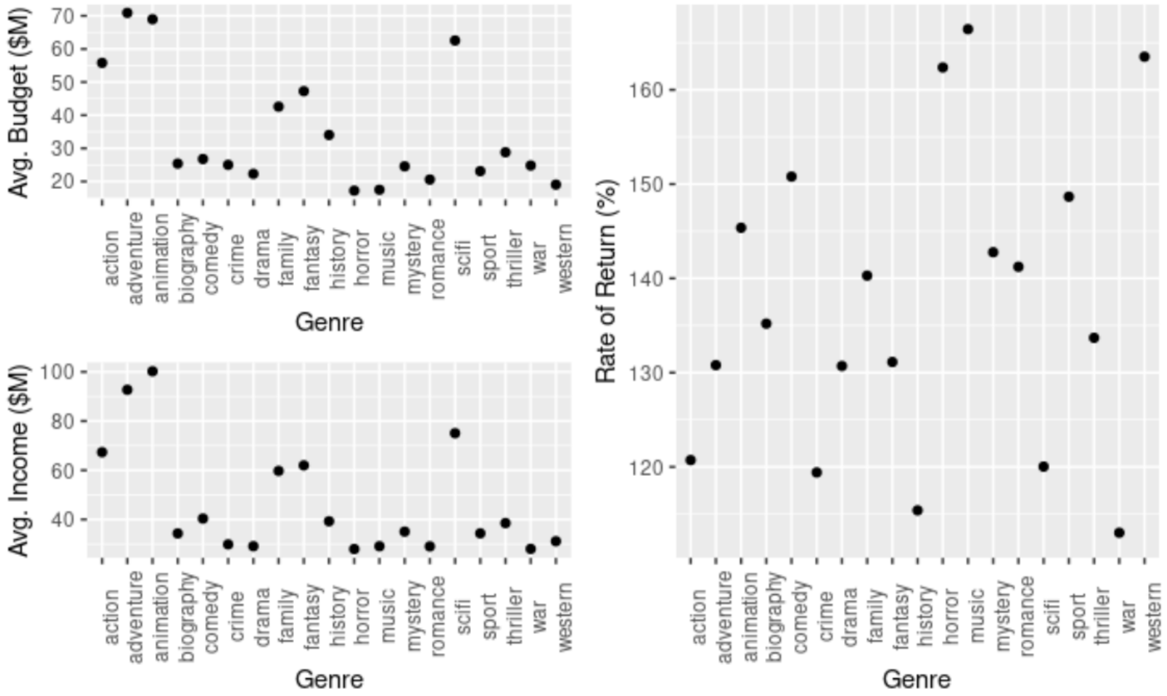


Figure 9. Average budget, average income and rate of return for different genres of the movies. Rate of return varies from genre to genre and ranges from 100-170%.

And finally, before we began with our model building, we were interested in seeing if there was a difference in gross income in movies where an Oscar winning director or writers were behind it. We did see that movies that were directed by Oscar winning directors made almost twice as much gross income than movies that were not, $60,271,602 and $39,993,43, respectively. However, when it came to having an Oscar winning writer or not, we didn't see that great of a difference, $40,927,918 and $48,180,652, respectively. Because of these differences in income between Oscar winning directors and writers, we wanted to include both factors in our model.

All these different variables i.e. genre along with the control variables like budget, year, duration etc. are drawn in the causal diagram below in Figure 10. We are hypothesizing that all these variables have a causal pathway towards our dependent variable i.e. gross income. In this causal diagram we are showing a single node for the genre just for simplicity, however in the way we have modeled it, there are 19 nodes: one for each genre type and they all have a causal pathway to the gross income. We have also shown other dependencies that we think might be present for example studio popularity might affect the votes and votes might affect the studio popularity.

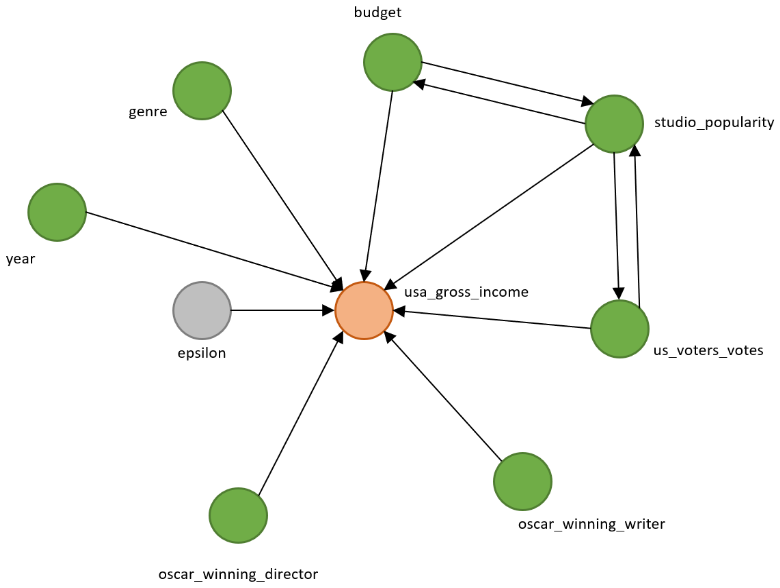


Figure 10. Causal Graph

Section 3.2 Model Building

Building our model was a process and we investigated several different versions.

Before we began any testing, we set some specifications for our models to follow:

1) Use USA Gross Income Logged as our dependent variable
2) Use movie duration, movie release year and genre as our control variables
3) Reject the null hypothesis with alpha level of .05.
4) Use a two tailed alternative.
5) Use F-test to determine if certain independent variables are necessary if they improve our model performance.
6) Use min-max accuracy value of greater than equal to .90 to assess how well a model is predicting
7) Use adjusted R^2 greater than or equal to 0.6 to evaluate the amount of variability in our model.
8) Use MSE to assess error rates.

Table 1 below shows how we assessed the model. All models contain the controls, the additional variables are specified in the table.

Model	Variables Included	Accuracy	Adjusted R^2	MSE	F test
Mod1	-	0.895	.24	5.20	-
Mod2	log(budget)	0.915	.50	3.45	Budget does improve model, continue with mod 2
Mod3	log(budget)+log(studio_popularity)	0.918	.54	3.18	Studio_popularity does improve model, continue with mod 3
Mod4	log(budget)+log(studio_popularity)+log(us_voters_votes)	0.933	.69	2.15	US Voters does improve model, continue mod 4
Mod5	log(budget)+log(studio_popularity)+log(us_voters_votes)+Ocar_Winning_Writer	0.933	.69	2.15	Oscar winning writer doesn't improve model accuracy, we fail to reject null. Go with mod 4
Mod6	log(budget)+log(studio_popularity)+log(us_voters_votes)+Ocar_Winning_Director	0.933	.69	2.15	Oscar winning director doesn't improve model accuracy, we fail to reject null. Go with mod 4

Table 1: Model variations without any interaction terms.

From the table above, it appears as model 4 seems to be the best one. It has the highest accuracy rate (93%), highest adjusted R squared (.69) and a MSE of 2.15. Additionally, this is the model where the most difference was made with the additional covariates.

We also wanted to see if there are any interactions between the different variables. To do this we applied domain knowledge and some descriptive statistics. We think that budget may have some relationship with studio popularity. From research, we learned that film financing can be acquired through several different ways two which including film studios and entertainment companies. So how popular a studio may be, may have strong interaction with their overall budgets.

It was also mentioned earlier that our data initially consisted of a genre variable. This genre did tend to have more than 1 type for each movie. We saw over 400 unique combinations from these. Because of this, we decide to create 19 separate dummy variables to account for each genre to make it easier and interpretable in our regression models. But when we ran summary statistics on the genre variable, we noticed that comedy, drama and romance combinations composed most of the movies. Due to that we want test two way interactions between comedy/drama, comedy/romance, drama/romance and one term with all 3 (comedy*drama*romance) to see if there is something statistically significant about these genres when it comes to bringing in a high gross income.

We will test the interaction terms with the model we had chosen previously and see whether adding them in makes a difference to our predictions. Table 2 below shows the different interaction term models ran and the value they brought to our models.

Model	Variables Included	Accuracy	Adjusted R^2	MSE	F test
Mod4	log(budget)+log(studio_popularity)+log(us_voters_votes)	0.933	.6864	2.15	
Mod4_i	Mod 4 + ((log)studio_popularity * log(budget))	.934	.6899	2.13	Interaction term made an improvement to our model. Continue with model mod4_i
Mod4_ii	Mod4_i+(comedy*drama)	.934	.6899	2.13	Interaction term didn't make an improvement to our model, stick with mod4_i
Mod4_iii	Mod4_i+(comedy*romance)	.934	.6899	2.13	Addition of comedy*romance interaction didn't improve our model, stick with mod4_i
Mod4_iv	Mod4_i+(drama*romance)	.934	.6902	2.13	Interaction term made an improvement to our model, continue with mod4_iv
Mod4_v	Mod4_iv+(drama*romance*comedy)	.934	.6901	2.13	Addition of comedy*romance*drama interaction didn't improve our model, stick with mod4_iv

Table 2: Model variations with interaction terms.

The table above shows that mod4_iv is the best model with the interaction terms. Out of the interactions we initially thought were relevant, only movie budget/studio popularity and movies that were romance dramas and comedy dramas made the most impact. The other terms, comedy*romance interaction, comedy*drama and comedy*drama*romance were not significant. This was a surprise to us, as we assumed romantic comedies would be significant due to their overall popularity in the movie industry.

As another way to validate our interaction terms, we ran mod4_i against individual subsets of data. We decided to create subsets of our dataset based on the genres included - one subset for all films containing action as a genre, one subset for all films containing adventure as a genre, and so on. This allowed us to have the genre that the subset is for as the base case and then we can see the direct effect of adding a new genre in combination with the original genre.

Moving forward, we will proceed with mod4_iv as the best model, which includes the following variables: Y ~ movie release year + movie duration + [action + adventure + comedy + crime + drama + family + fantasy + history + horror + mystery + romance + scifi + thriller + western + animation + biography + sport + music war] + log(budget) + log(studio_popularity) + log(us_voters_votes)*log(budget) + log(studio_popularity) + log(us_voters_votes) + (drama*romance)

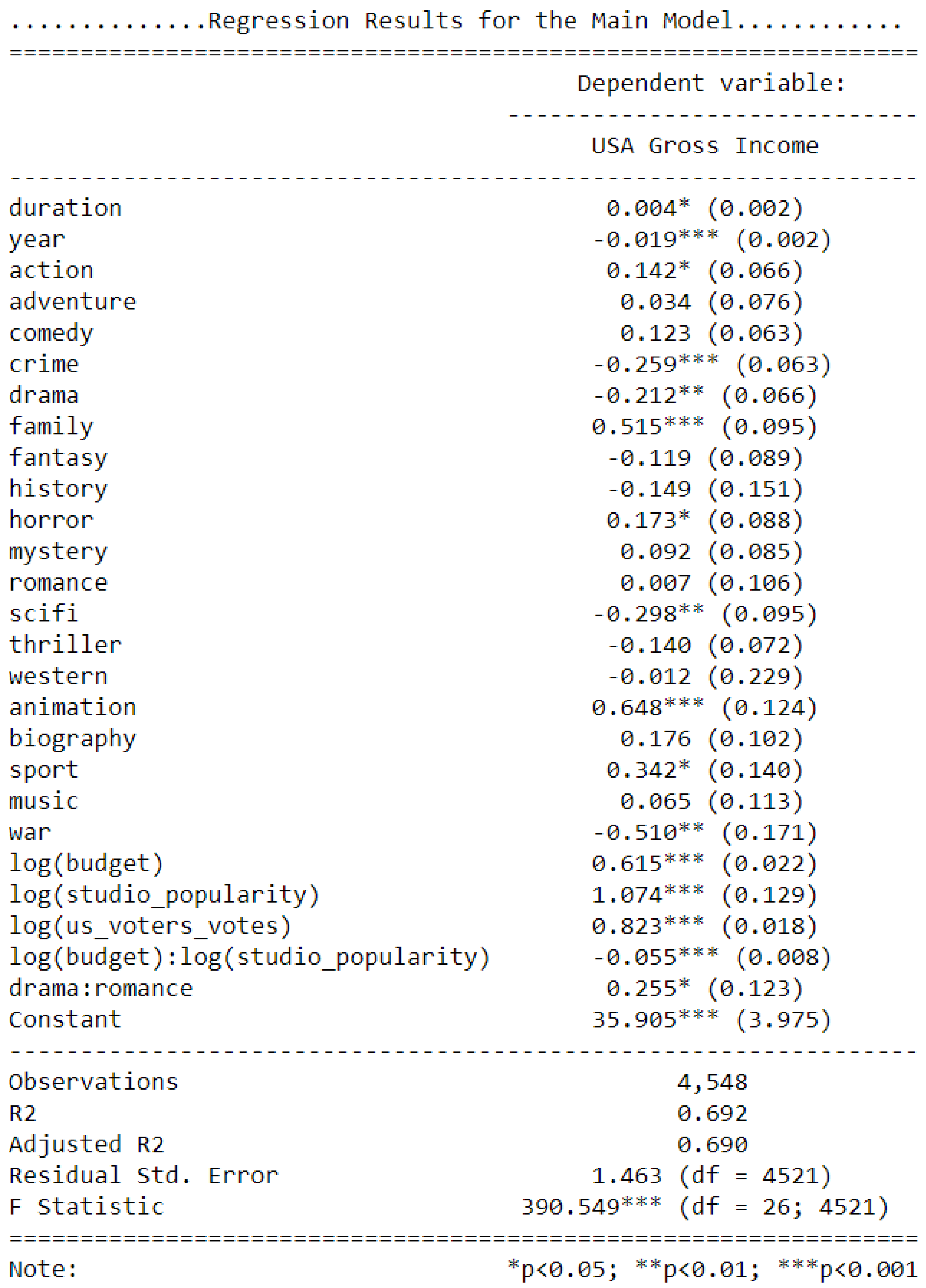
Section 4: Results

$$
\begin{gathered}
log(\text{usa\_gross\_income}) = \beta_0 + \beta_1\text{duration} + \beta_2\text{year} + +\beta_3\text{action} + \beta_4\text{adventure} + \\
\beta_5\text{comedy} + \beta_6\text{crime} + \beta_7\text{drama} + \beta_8\text{family} + \beta_9\text{fantasy} + \beta_{10}\text{history} + \beta_{11}\text{horror} + \\
\beta_{12}\text{mystery} + \beta_{13}\text{romance} + \beta_{14}\text{scifi} + \beta_{15}\text{thriller} + \beta_{16}\text{western} + \beta_{17}\text{animation} + \\
\beta_{18}\text{biography} + \beta_{19}\text{sport} + \beta_{20}\text{music} + \beta_{21}\text{war} + \beta_{22}log(\text{budget}) + \\
\beta_{23}log(\text{studio\_popularity}) + \beta_{24}log(\text{us\_voters\_votes}) + \\
\beta_{25}log(\text{budget}) * log(\text{studio\_popularity}) + \\
\beta_{26}log(\text{drama}) * log(\text{romance}) + e
\end{gathered}
$$

There are 26 coefficients in our final model and since we are running a linear regression, i.e. fitting a linear model to predict the outcome that is US gross income for movies, our best linear predictor (BLP) will track the mean of US gross income. We have structured our data in a way that we can see the effect of different genres on the US gross income and when we interpret our coefficients, our base case is not the mean US gross income for the movies without any genre rather when we look at the coefficient of action for example, what it tells us is that how does the US gross income would change (i.e. does it increase or decrease) compared to the US gross income for all the movies other than the action movies. For control variables, we see whether increasing or decreasing that variable causes higher or lower gross income than the mean gross income of all the movies (of all the genres). A detailed description of the interpretation of coefficients is given in table 3 below:

Variable	General Interpretation
Genre Type	How does the gross income vary in comparison to the gross income for all the movies excluding that particular genre type (i.e. action, comedy, etc)
Duration	How does the gross income vary, as we look at a short movie (e.g. an hour long movie) in comparison to a longer movie (e.g. a movie that is 2 hrs. long)
Year	How does the gross income vary, as we look at a new movie (e.g. a movie released 1 year ago) in comparison to an old movie (e.g. a movie released 10 years ago)
Budget	How does the gross income vary, as we vary the budget, in comparison to the mean gross income for all the movies of all genre type
Studio Popularity	How does the gross income vary as we look at a movie that comes from a popular studio in comparison to the gross income of the movies that come from a relatively less popular studio.
Votes	How does the gross income vary for the movie that has amassed high number of votes in comparison to the gross income of movies that have received relatively lower number of votes

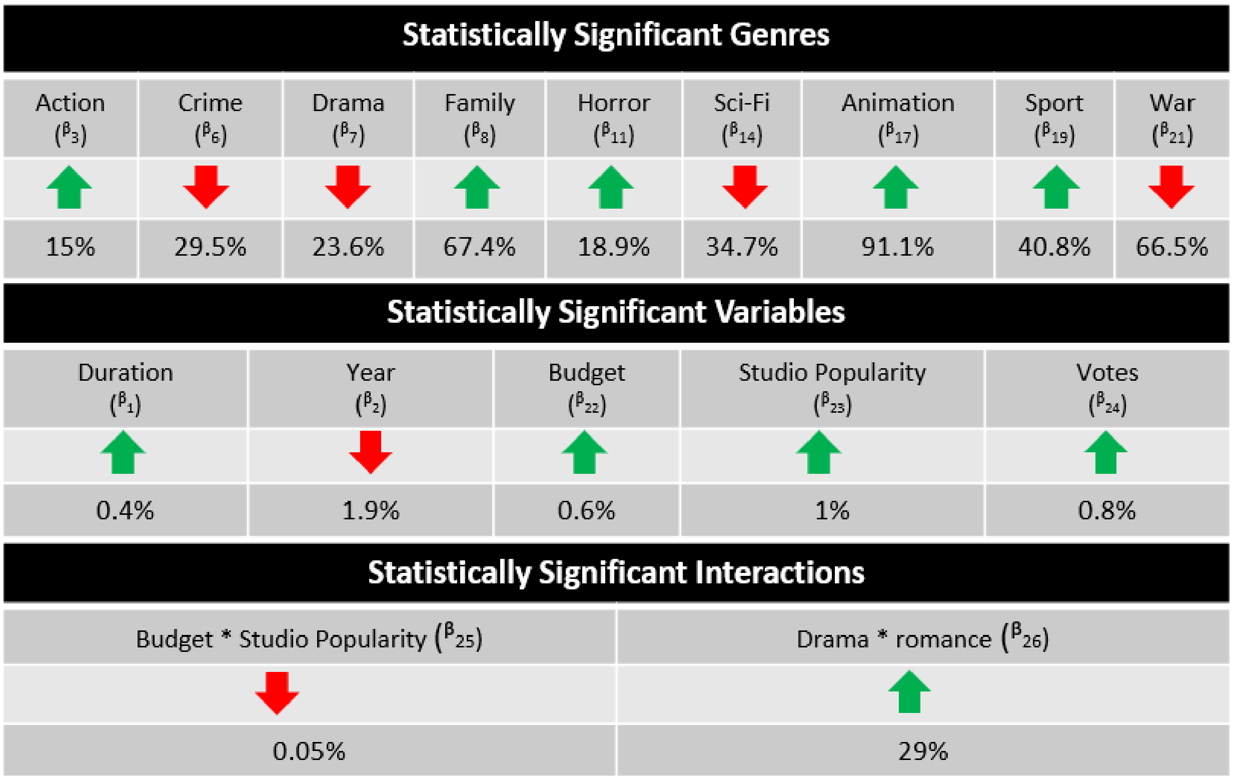
Table 3. Regression coefficients and their interpretation



```
..............Regression Results for the Main Model...........
================================================================
                                          Dependent variable:
                                   -----------------------------
                                           USA Gross Income
----------------------------------------------------------------
duration                                    0.004* (0.002)
year                                       -0.019*** (0.002)
action                                      0.142* (0.066)
adventure                                    0.034 (0.076)
comedy                                       0.123 (0.063)
crime                                      -0.259*** (0.063)
drama                                      -0.212** (0.066)
family                                     0.515*** (0.095)
fantasy                                     -0.119 (0.089)
history                                     -0.149 (0.151)
horror                                      0.173* (0.088)
mystery                                      0.092 (0.085)
romance                                      0.007 (0.106)
scifi                                      -0.298** (0.095)
thriller                                    -0.140 (0.072)
western                                     -0.012 (0.229)
animation                                  0.648*** (0.124)
biography                                    0.176 (0.102)
sport                                       0.342* (0.140)
music                                        0.065 (0.113)
war                                        -0.510** (0.171)
log(budget)                                0.615*** (0.022)
log(studio_popularity)                     1.074*** (0.129)
log(us_voters_votes)                       0.823*** (0.018)
log(budget):log(studio_popularity)        -0.055*** (0.008)
drama:romance                               0.255* (0.123)
Constant                                  35.905*** (3.975)
----------------------------------------------------------------
Observations                                     4,548
R2                                               0.692
Adjusted R2                                      0.690
Residual Std. Error                        1.463 (df = 4521)
F Statistic                           390.549*** (df = 26; 4521)
================================================================
Note:                              *p<0.05; **p<0.01; ***p<0.001
```

Figure 11. Regression results for the main model

As can be seen in Figure 11. above, out of 19 genres, 9 of these have a statistically significant causal relationship with the gross income. Action, Family, Horror, Animation and Sports movies make more income than the baseline and this is truly due to the genre of the movie. Whereas crime, drama, sci-fi and war movies make lower gross income than the baseline. An action movie makes 15% more gross income than mean gross income of all the movies excluding action movies whereas crime movies makes 29.5% lower gross income. For all statistically significant genres the impact on gross income is mentioned in the Figure 12 below.



Statistically Significant Genres								
Action (β_3)	Crime (β_6)	Drama (β_7)	Family (β_8)	Horror (β_{11})	Sci-Fi (β_{14})	Animation (β_{17})	Sport (β_{19})	War (β_{21})
⬆	⬇	⬇	⬆	⬆	⬇	⬆	⬆	⬇
15%	29.5%	23.6%	67.4%	18.9%	34.7%	91.1%	40.8%	66.5%

Statistically Significant Variables				
Duration (β_1)	Year (β_2)	Budget (β_{22})	Studio Popularity (β_{23})	Votes (β_{24})
⬆	⬇	⬆	⬆	⬆
0.4%	1.9%	0.6%	1%	0.8%

Statistically Significant Interactions	
Budget * Studio Popularity (β_{25})	Drama * romance (β_{26})
⬇	⬆
0.05%	29%

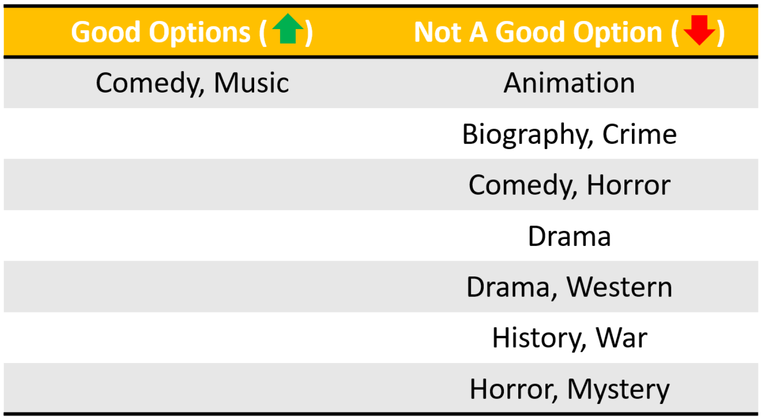
Figure 12. Statistically significant genres, control variables and interaction terms. A green upward arrow indicates that they have a positive impact on the dependent variable i.e. gross income whereas a downward red arrow indicates a negative impact of that variable on the gross income. For genres, the percentage change is w.r.t to the mean gross income for all other movies excluding that genre. For control variables and the interaction terms, the percentage change is the increase or decrease in gross income when the variable or interaction term is increased by 1%.

We also found that duration, budget, studio popularity and votes also have a causal impact on the gross income and they all push it above the baseline whereas year has a negative causal impact on the gross income meaning that older movies amass higher gross income than the newer ones. Finally, we found two statistically significant interactions, one is budget and studio popularity that negatively impacts gross income and the second one is drama and romance that has a positive impact on the gross income.

We divided our data into training and test data to check the reliability of our model. For our testing data we got an adjusted R^2 value of .689, accuracy of 93% and a Normalized RMSE of .095. Since these statistics are essentially the same values as we received while building our model on our training set (can be seen in Figure 11 above) , we can say our model was well reproduced.

We further looked at how it affects the gross income of a movie, let say an action movie, if we combine it with one or two other genres to answer questions like will an action movie be more successful or an action movie that is also comedy has a better chance of success than an action movie alone? To do this we built a model that was run on a subset of data which only included all those movies that had action as one of the genres. In this scenario our base case is just the mean gross income of all the action movies (with no other genres present) only.

REcommendations for action movies are summarized in Figure 13 below. We found that we can combine action movies with comedy and music because doing so will result in a more successful movie than an action movie alone. Whereas combining an action movie with other genres like animation, biography and crime, comedy and horror, drama, drama and western, history and war, and horror and mystery will result in a movie that is less successful than an action movie alone. We are including results for action movies only but this framework can be used the same way to find recommendations for other genres of the movies.



Good Options (⬆)	Not A Good Option (⬇)
Comedy, Music	Animation
	Biography, Crime
	Comedy, Horror
	Drama
	Drama, Western
	History, War
	Horror, Mystery

Figure 13. Recommendations for action movies. 'Good Options' means other genres that when combined with action movies result in more successful movies than action movies alone whereas 'Not A Good Option' are the genres that when combined with action movies results in a less successful movie than action movie alone.

Section 5: Model Limitations

Section 5.1: Statistical Limitations

Because our dataset contains more than 100 movies, we evaluated the large sample model assumptions. The two assumptions are a) data is collected in an independent and identically distributed manner (i.i.d) and b) a unique best linear predictor exists.

Assumption 1: Independent and Identically Distributed (I.I.D.)
First, to evaluate if the data is i.i.d we needed to examine the sampling process. Since this data is captured and provided by an individual their rating of a movie on a scale of 1 to 10, one individual does not provide insight into the rating of another, and thus consider this data independent and identically distributed.

Even though the requirement of i.i.d was met, there are some potential risks within our dataset. First, there is a potential risk of clustering where the ratings are not coming from a representative group of the US. To provide ratings you have to register for an account on IMDb's website. Second, there is a potential risk of subjectivity with our Oscar dataset as Oscar winners historically have not been diverse.Even though the Oscars is considered a gold standard in the film industry, it might not be an inclusive measure. Finally, we derived our studio popularity based on how often a production company appears in our data. We limited our dataset to movies where one of the countries it was released in the US and one of the languages is

English. With that restriction, we could potentially undercount studios that are popular outside of the US (or is a subsidiary of a larger production company) but have only released a small number of movies in the US.

Assumption 2: A Unique Best Linear Predictor(BLP) Exists

To evaluate if a unique BLP exists, we need to determine if any of our predictor variables can be written as a linear combination of any of the other predictors.For our dataset, each predictor is unique and cannot be derived from another predictor. For example, you cannot determine what the budget of the film will be from any of our other predictor variables. Because of the variation between our variables and that none of our model coefficients increased suddenly, we determined this assumption is satisfied.

Section 5.2: Structural Limitations

Though our model takes into account many of the factors that could affect the gross income for a film, there are a few omitted variables for this model that should be addressed. We would also like to acknowledge that through our investigation we did not analyze the relationship between genre and the other variables, for example budget. We are aware that there could be a relationship between genre and other variables and therefore our report can only speak to the relationship between genre and gross income under the assumption that there is no dependency between genre and other variables in our model. Through the following discussion we will analyze each omitted variable and the effect it could have on our core model.

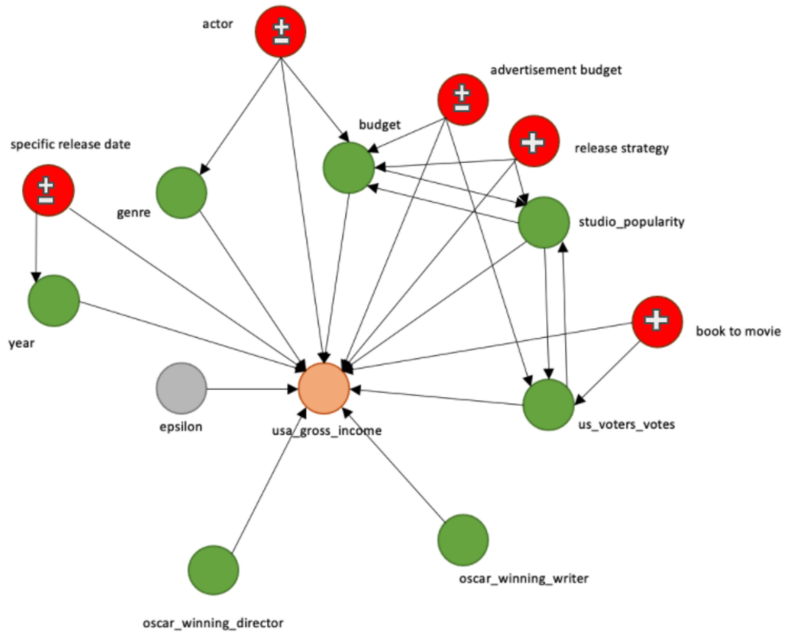


Figure 14. Causal Graph with potential 'Omitted Variables' that pose a threat to the findings of our model

The above causal graph, Figure 14, shows the interaction between the omitted variables in red and the variables included in our model in green as well as the US gross income.

Actor and Actress

Similar to directors and writers, actors and actresses can influence the success of a film and its gross income. For the most part, Americans have a favorite actor or actress and take pride in seeing their full body of work. Furthermore, audiences can feel strongly towards boycotting films that include specific actors or actresses. As a result of this influence, the gross income of a film can be directly affected either positively if it is an actor or actress audiences enjoy or negatively if it is an actor or actress that audiences disapprove of. This direction bias could be away from zero or towards zero. This omitted variable could have a direct effect on our core result as some actors and actresses build their filmography in specific genres.

Book to Movie

When a book gets a film deal, there is regularly a cult following that then gets absorbed when the film is released. Whether fans of the book series are excited for the film to come out or if they disagree with the casting, it can be assumed that they would go to film regardless as they would want to see their fantasy world brought to life. With this assumption we could assume that if it is a film adaptation of a book, there would be an increase in gross income which would result in a direction bias away from zero. Though this could have an effect on our core outcome, historical releases of film adaptations of books do not prove there would be a guaranteed effect as there have been both hits and flops in the theatre.

Advertisement Budget

Though we look at the overall budget of a film in our models above, the budget specifically associated with advertisement for a film could affect the gross income of the film. If the advertising budget for a film is very high, it will create interest in the film prior to the release and get audiences excited about the film and could influence them to go see a film they otherwise would have not. On the other hand, over advertising a film prior to its release can also make audiences lose interest in a film as they feel as though they already know the plot based on the previews they've seen. As a result of this difference, the direction bias for this omitted variable would be away from zero or towards zero and this would not have a substantial effect on our core findings.

Release Strategy

The release strategy of a film determines if a film is released solely to theaters, streaming platforms, or a combination of the two. This could affect the gross income of a film and would be dependent on what the agreements are with the streaming platforms for each film. Should a film be released to both theatres and streaming platforms there is the potential that this would increase the gross income of the film as opposed to if it was just released to one option. As a result, the direction bias would most likely be away from zero but we do not think this would have a substantial effect on our core findings.

Release Date

We currently account for the year for which a film is released in our model; however, we believe it would be more valuable to look at the month for which a film was released. Genres can be more popular during specific months, for example horror films during Halloween or romance films during Valentine's Day. We would expect that there is a direction bias both towards and away from zero for this omitted variable as when it is a month that a film genre is not as popular the gross income would decrease but if it was a month when a genre is more popular the gross income would increase. We believe this omitted variable could have an effect on our core findings.

Section 6: Conclusion

Section 6.1: Conclusion

Through our report we found that our causality model had significant results for the question of if genre has an effect on the gross income of a film. We ran our model against a data set of 6,498 films that were released in the US after 1928 and found that for this subset animation had the greatest positive effect on gross income. These results were both statistically significant and practically significant.

In summary, we found that the most successful movie genre, animation, led to a 91.1% increase in US gross income when compared to the mean US gross income of all movies excluding animation. Though as previously stated this result is under the assumption that there is no dependency between genre and our other variables; however, we still believe these results are statistically and practically significant because they proved to be reproducible and share further insight into the film industry that can help with future productions.

Section 6.2: Discussion

While creating this report, we discovered other areas that would be interesting to explore to better understand what affects a film's gross income and what makes it successful. A specific area that would be worth exploring is seeing how the Coronavirus affected gross income. As many people spent more time watching films during quarantine but on streaming platforms it would be interesting to see if gross income was affected substantially since there were so few movie theatre ticket purchases and in fact many movie theaters ended up closing as the result of the quarantine so it could also affect gross income in years to come.

Additionally, the film industry has many other factors that could affect gross income beyond genre. It would be interesting to do a more in depth analysis on which of the many elements of a film affect the gross income the most - if not genre, what drives a film's profits?

We ultimately believe our findings are statistically and practically significant; however, there are a multitude of areas that can continue to be explored within this industry to determine what factor of a film can bring the highest US gross income.

Section 7: References

1. Full GitHub Repository for this project: https://github.com/mids-w203/Lab2_Team_Daring
2. Kaggle. IMDb Movies Extensive Dataset. Retrieved November 15, 2021, from https: //www.kaggle.com/stefanoleone992/imdb-extensive-dataset/
3. Filmsite. Academy Awards Best Director Winners. Retrieved November 22, 2021, from https: //www.filmsite.org/bestdirs2.html/
4. Filmsite. Academy Awards Best Screenplay Winners. Retrieved November 22, 2021, from https: //www.filmsite.org/bestscreenplays5.html/