Can a Movie's Budget Be Used to Predict its Success?

Richard Glennon, Michael Lerch, and Brandon Adams
Math 193 A Intro Statistics / Data Science

Abstract

When you think of the most successful movies recently, you tend to think of massive marvel movies or Disney movies with huge budgets. What if we could use budget data from movies to predict the success?

This brings the next question, how do we measure success of a movie? The two obvious choices we thought of were overall revenue, and ratings.

The dataset we are using is "The Movies Dataset" from Kaggle. It includes 5 data tables, the one we are focusing on is "movies metadata" with data including cast, crew, keywords, budget, revenue, posters, release dates, languages, production companies, countries, vote counts, vote averages, and more. There is also over 26 million reviews included.

Introduction / Data Cleaning

While we looked into many other relationships, we believed budget would be one of the best predictors. Other variables like title length, and runtime seemed to have no effect on either form of success.

We first had to join our 5 data tables together to be able to compare all of the information easily.

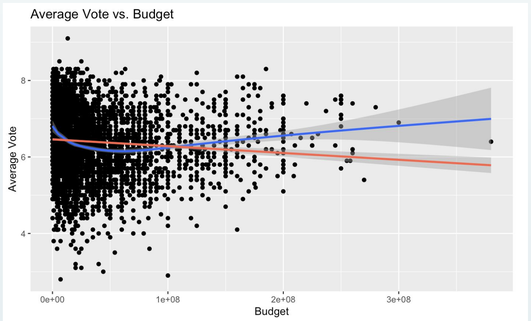
When we were cleaning our data, we came across movies that had a revenue or budget of zero. We think that Keggle put in zeros in places where they did not have data. We replaced these improbabilities with "NA"s.

In addition to the zeros, we had to filter out some of the data that we did not think would have a great impact. We filtered out some of the smaller genres such as Adult films, or movies with low vote counts. We did this so the vote averages would be more trustworthy.

For this analysis, we are using a single linear model. In the future we may look into multiple regression, and nonlinear models to try and more accurately model the relationship.

Budget as a Predictor for Ratings

When graphing budget and average vote, there does not appear to be any correlation. there are almost no patterns in the data, and the blue LOESS curve and the red fitted linear model are almost horizontal.



cor
<dbl>
-0.08975617

The equation for the linear model shown is: $\hat{y} = 6.46 + 0x$

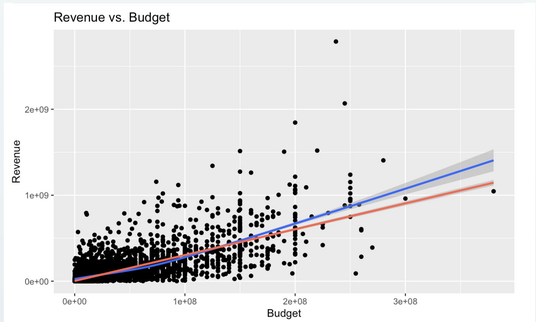
This gives us an r value of -0.0898, and an r^2 value of 0.00805617, meaning 0.8% of the variation in vote average is explained by budget. This is very low, and means budget explains almost none of the variation. It is clear that budget cannot be used to predict vote average well.

There is little proof to prove our hypothesis, that movies with higher budgets have higher average votes.

Maybe we should move onto a different variable: revenue. Movies need to at least make back their budget to make money, so we thought maybe they would have a relationship.

Budget as a predictor for Revenue

When graphing budget and revenue, the variables appear to have a moderate positive correlation. The blue line shown is a LOESS curve, while the red line is our fitted linear model.



cor
<dbl>
0.7091052

The equation for the linear model shown is: $\hat{y} = 2998573.384 + 3.007x$

This gives us an r value of 0.7091052, and an r^2 value of 0.5028302, meaning 50.2% of the variation in revenue is explained by budget. This is fairly high, meaning budget is a fairly good predictor of revenue.

This is good evidence for our hypothesis, that movies with higher budgets have higher revenue.

This is just like we thought, proving our hypothesis correct.

Conclusion

To summarize, you cannot accurately use a movie's budget to predict its average rating, but it is much more possible to use a movie's budget to predict its revenue.

This is not surprising, movies with larger budgets can typically pay more for advertising, special effects, and more famous actors, which more people will buy tickets to see.

From this dataset, we were able to conclude that if a movie wanted to optimize its revenue, it should be released in English during the summer of 2020 with a higher budget.

We also deduced that if a movie wanted to optimize their ratings, the movie should be released in Japanese in any month and the budget would not matter. The directors would also have to create a time machine to optimize ratings because the time to have the best ratings would be in the past.

In the future, we should explore other models than a linear model to see if they are a better predictor, but for now linear does a good enough job. We will also look into multiple regression, using more than just the budget to predict these metrics. Maybe that is the extra piece we need to predict vote average.

Another issue we should look into is inflation. Adjusting for inflation, many movies in the past have made astounding amounts of money. Currently, our analysis has bias from movies budgets and revenue growing together over time due to inflation.

Acknowledgements

The movies Dataset:
https://www.kaggle.com/rounakbanik/the-movies-dataset