Factors Affecting Movies Gross Revenue

Richa Patel
Jun Yan

Department of Statistics
University of Connecticut

December 15, 2023

Abstract

The expanding economy and changing cultural expectations are driving the film industry's steady expansion. The study examines important factors such income, popularity, runtime, and vote average using a dataset that was obtained from the Kaggle website and included information on over 6820 movies. Techniques for visual analysis are used to comprehend how they are interdependent. The results demonstrate that three important factors influence a film's financial performance at the box office: directors, budget, and votes. The statistically substantial correlations between them are shown using linear regression modeling. This study offers useful insights into the elements that contribute to a film's box office success, direction for stakeholders, and doable recommendations to improve each specific film's performance. The report, which is supported by an in-depth understanding of the variables influencing box office revenue, presents a positive assessment of the film industry's future course.

Keywords: Linear Regression; Factors; Movie Revenue; Data Analysis; Key Factors of Movie Sucess; Impact on Revenue

1 Introduction

Since the dawn of cinema, a lot has evolved in terms of effects and sound design. Silent or black-and-white films are no longer common. However, there had been no change in the variables that are used to predict revenues. The stars, directors, authors, budget, corporation, rating, and scores are among those elements. With varying ratings and genres, movies are meant to amuse their viewers.Since IMDB is the most popular website for movie reviews and ratings, it could be interesting to look at user reviews and find out what people liked and didn't like, which might help us determine whether or not the movie has satisfied viewers, which could have a positive or negative effect on the movie's box office performance [3]. However, the writers' and directors' perspectives on those films, together with the actors whose acting transforms the director's vision into a live action, are what stimulates their interest in those films the most. Sometimes stars with a lot of popularity may also have a big impact on the direction of movies. In certain cases, obtaining agreements from distributors, exhibitors, producers, investors, and sponsors even serves as the ”green light” [2]. However, the budget also has a significant impact because it sets the initial outlay for the movie and, to some extent, predicts the movie's quality, as the initial outlay of funds may be a good indicator of the movie's total earnings [1]. That being said, movies deliver their viewers those things. The most significant measure of a movie's popularity among viewers is its box office earnings, Similarly, the best measure of a film's earnings is its box office performance [1]. The gross revenue, however, is what really matters most as it determines how well the film performed at the box office.

A similar study on the factors determining revenue was carried out in 2023 by Bingyu Hao of York University in Canada. Hao's study concentrated on the TMDb Movies dataset, which contains details about more than 10,000 films and their attributes including budget, runtime, revenue, popularity, and vote average, among others. He made use of the revenue distribution and scatter graph in relation to budget, runtime, vote average, and popularity, among other factors for EDA analysis. Then, he employed fitted regression and linear regression as his models. Then ultimately came to the conclusion that the last factors influencing the ability to anticipate movie revenue are popularity, vote average, and budget.

The data set in this study provides details on the variable found in Section 2. It is followed by a brief explanation of the methodology used in Section 3, which describes the study that was conducted to anticipate the variables influencing movie income. Results are presented in Section 4, which also includes a graphic representation explaining the analysis's conclusions. Finally, there is the conclusion section, Section 5, which contains the conclusion and further expectations for this paper's consideration.

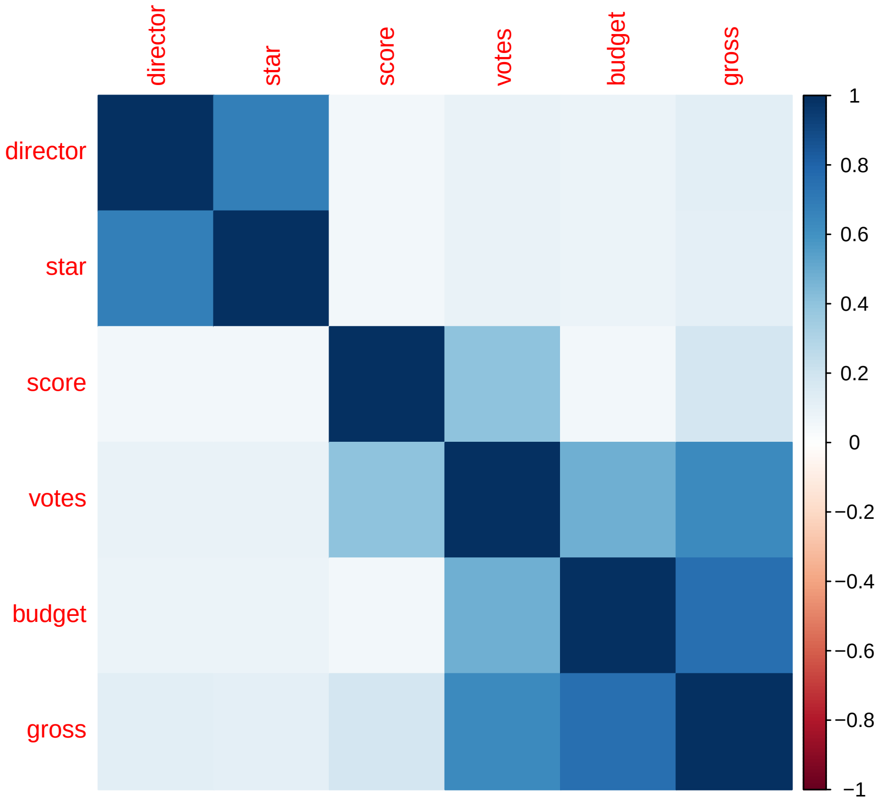
2 Data

The data of this paper is taken from Kaggle . The data includes 6820 films from 1986 to 2016 (about 220 films annually). It has fifteen features: budget, runtime (the amount of time the film is on screen), company or productions, country, director, genre, gross (the amount of money the film brings in), name of the film, rating (R, PG, etc), date of release, socre (IMDb ratings), votes (user votes), actors starring in the film, year of release, and writers. I will be analyzing the variables that affect gross revenue in this paper by concentrating on the following variables directors, stars, scores,votes and budget.

3 Methods

After selecting the 5 varibles to focus on gross revenue of movies, I would be using liner regression model in my analysis medthod. But Before getting started into detailed regression model, I converted the nomial columns to numerical columns. As the data contained numerical data that is score, budget, votes, runtime and nominal data that are directors, stars, writers, ratings, year of realease, genre, company, country. Afterwards, I looked for any missing values or NA values and replaced them with 0.

For EDA analysis as it is to gaining insights and understanding the characteristics of



In Figure 1, the plots show the Correlation Graph.

the data I will be using Correlation graph. Allowing to identify patterns, dependencies, or potential relationships among variables.

As we can see in Figure 1 the correlation between each axis's variables is displayed by each square. From -1 to +1 is the correlation range. A linear trend between the two variables is shown by values that are closer to zero. Positive correlation indicates that the closer one is to 1, the greater the link; that is, as one grows, so does the other. When one variable decreases while the other grows, the correlation is comparable when it is closer to -1. Because the squares there are perfectly correlated, each variable is correlated with itself, which is why all of the diagonals are 1/dark blue. Regarding the other variables, a greater connection is indicated by a larger number and a darker hue. The plot is also symmetrical about the diagonal since the same two variables are being paired together in those squares.

The data was then divided into training and test sets so that a model could be trained on one set and its performance assessed on the other, with a 7:3 split ratio adjustment made.

And then Linear regression is applied to the data set.

The model of the linear regression is:

$$Revenue = \beta_0 + \beta_1 * directors + \beta_2 * stars + \beta_3 * scores + \beta_4 * votes + \beta_5 * budget \quad (1)$$

4 Results

Table 1: Table 1: Linear Regression Model

	Estimate	Std.Error	t value	Pr (>—t—)
intercept	-4.745e+07	9.255e+06	-5.127	3.05e-07
director	9.393e+03	2.153e+03	4.363	1.31e-05
star	-3.136e+02	2.245e+03	-0.140	0.8889
score	3.391e+06	1.441e+06	2.353	0.0187
votes	3.178e+02	9.622e+00	33.028	2e-16
budget	2.501e+00	3.848e-02	65.010	2e-16

The linear regression model's coefficients are displayed in Table 1 for our observation. Additionally, each variable's p-value is has to beless than 0.05 in order to be significantly inpacting Revenue. The data indicates that neither the variable star (p-value of 0.8886) nor the score (p-value of 0.0186) are statistically significant in predicting the result. Not as significant as other factors, even if the score's p-value is 0.0186 which is somewhat below the usual significance level of 0.05.

The variables that I used yielded the same findings as those that Hao used in his linear regression. Additionally comparable are the variables having p-values less than 0.05. Instead of using population, budget, runtime, and vote average as my variables, I utilized directors, star, score, vote, and budget.

The factors star and score were eliminated because of greater p-values than 0.05, which indicated that they had little to no impact on revenue. In order to have a more thorough understanding of the variables influencing revenue, an altered linear regression model was developed. Table 2 shows a p-value of less than 0.05 for each of the variables—director,

Table 2: Table 2: Fitted Linear Regresion Model

	Estimate	Std.Error	t value	Pr (>—t—)
(Intercept)	-2.649e+07	2.312e+06	-11.454	< 2e-16
director	9.300e+03	1.568e+03	5.931	3.2e-09
votes	3.276e+02	8.675e+00	37.765	< 2e-16
budget	2.486e+00	3.791e-02	65.576	< 2e-16

votes, and budget—indicating that these factors have a considerable influence on revenue.

The budget, vote average, and popularity are significant independent variables with p-values less than 0.05, according to Hao's prediction, which has a significance p-value of 0.05. Therefore, in order to analyze movie earnings, he looked at votes, budget, and population. The variables that were utilized are identical to the ones I had used for my analysis, and after the fitted regression, the variables with p-values less than 0.05 are also similar. However, I used directors as my variable rather than pupulation.

Final Model:

$$Revenue = -2.649e + 07 + 9300 * directors + 327.6 * votes + 2.486 * budget \qquad (2)$$

As in this model, directors, budget and votes are the most influential variables. Since the coefficient of directors is 9300, it may be inferred that selecting a director for a unit boosts revenue by 9300. Comparably, the votes coefficent of 327.6 means that a unit increase in votes corresponds to an indirect gain in income of 237.6. The budget's coefficient of 2.486 indicates that a unit increase in the budget will result in a 2.486 increase in revenue.

5 Conclusion

Intriguing findings were revealed by this study, which examined important variables affecting box office receipts. After thorough examination, it became clear that budget, votes, and directors all had a significant impact on a movie's ability to make money. The variable with the greatest influence was the director, indicating their critical function in generating money.

And the second gretest factor being votes. This shows that people are now paying more and more attention to what other people think of a movie, and people are easily influenced by what others say about amovie to see it or not [1]. In contrast to traditional assumptions, stars and ratings showed low statistical relevance in revenue forecast.

The results highlight how directors have the power to shape the sector and call for a reassessment of accepted wisdom. Despite the longstanding importance of stars and scores, their relatively little influence calls into question business norms. Analyses in comparison with previous research, particularly that of Hao, reveal similar results, but slightly different in focus. One should interpret the dataset cautiously due to its scope limitations and its only concentration on numerical measures. To supplement quantitative analysis, future research can examine qualitative elements that capture complex audience attitudes.

According to this study, those involved in the film business may utilize the information gathered from it to help them make strategic decisions. It highlights the significance of directors in predicting income and proposes that production decisions should be made with careful consideration for director selection, budgetary constraints, and audience feedback. Subsequent investigations may encompass qualitative information, sophisticated machine learning methods, and delicate factor integration to enhance the precision of revenue forecasting.

References

[1] Anita Elberse. The power of stars: Do star actors drive the success of movies? *Journal of Marketing*, 71(4):102–120, 2007.

[2] Bingyu Hao. The analysis of the factors that influence the film revenue. *Highlights in Science, Engineering and Technology*, 47:154–159, 05 2023.

[3] Latika Kharb, Deepak Chahal, and Vagisha. Forecasting movie rating through data analytics. In Usha Batra, Nihar Ranjan Roy, and Brajendra Panda, editors, *Data Science and Analytics*, pages 249–257, Singapore, 2020. Springer Singapore.