

The Analysis of the Factors that Influence the Film Revenue

Bingyu Hao

York University, 4700 Keele St, Toronto, ON, Canada M3J 1P3

icey7@my.yorku.ca

Abstract. As the rapid development of the economy in this era, in response to the needs of people in modern society, people are also demanding more and more movies. the film industry is also in constant development. Firstly, summarizing the results of other domestic and foreign scholars' research on the analysis of factors influencing box office revenues by reviewing national and international academic websites for information related to the topic of this paper. Afterwards, in this paper, analysis of the factors that influence movie revenue with respect to the data set on movies provided by Kaggle website which contains the information of more than 10,000 movies. By choosing the characteristics from the data set which are revenue, popularity, runtime and vote average followed by performing visual analysis showing the relationship between the response variables and independent variables and building the linear regression model. The result shows that the popularity, budget and vote average have significant effects on the revenue. Finally, the suggestions are made on the factors influencing each film's box office revenue and also expressed positive thoughts on the future development of the film industry.

Keywords: Movie box office revenue, factors, linear regression model, visual analysis.

1. Introduction

In this era of rapid development, with the rapid development of the global economy, meanwhile, the film industry is also growing in leaps and bounds. Statistics show that China's box office in the first half of 2022 has reached 17.186 billion, already close to the total annual box office in 2020, and so far, the total box office of 2022 has reached 25 billion. Depending on the rapid development of society and the continuous improvement of living standards, people are more pursuing to improve their quality of life after working. Thereupon, watching films has become a form of leisure. People's demand for movies is also growing, some people just regard it as leisure, and some people will seriously evaluate it after watching it. It is precisely because of the large demand, and at the same time, the revenue generated by movies have been very considerable in recent years. Box office revenue is the biggest indicator of audience recognition of a film. Likewise, the box office is the biggest indicator of a movie's revenue. Therefore, evaluating the factors that influence a film's revenue is crucial for future investment and production. At the same time, the analysis and forecast of box office revenue is of great significance to the future direction of film development, company investment decisions, and even economic development. Also, as an important part of the cultural industry, film also bears the burden of spreading the national culture of the country. In fact, the United States has done a very good job in this regard. In this regard, by analyzing the influencing factors of box office revenue, we can measure the future development direction, lead the development of the cultural industry and represent the national culture of our country to the world. For these reasons, in recent years, many domestic and foreign scholars have also studied this case.

Hang et al. chose the variables of word-of-mouth, release date, film genre, attention, preliminary box office and film length were selected and to construct multiple linear regression and random forest models. The result shows that the random forest-based model for analyzing and predicting the influence factors of movie box office is practical [1]. Hu et al. used multiple linear regression and General Linear Model Univariate to analyze the factors that influence the box office and the degree of the impact [2]. This paper indicated that the most significant factor is the director, followed by remake the famous IP, actor or actress and shooting the sequels. Films' type, release schedule and the producer do not have that deep clout compared with formers. Wang demonstrated that the popular slot releases have a boost to movie box office, online ratings have a prominent contribution

to movie box office, and the increase in per capital disposable income also contributes to movie box office by using Eviews [3]. Ni treated the multiple regression model to analyze the effect of the various film genres to the box office. The conclusion shows that the popular IP, drama and youth film have a big effect on the box office revenue [4]. Gao built the Fully Connected Neural Network, FNN base on the dataset to predict the box office, and suggested that to adjust the proportion of the film genres, increase the genres that audience like (e.g., comedy, action movie), choose the suitable director and actor and arrange the release time reasonably [5].

Sehwan, et al. researched about sharing the movie trailers on the social media platform whether it will have effect of the box office revenue by collecting the statistics of movie trailers from Youtube and corresponding with the box office revenue and applying a panel simultaneous equation model [6]. Apte, et al. analyzed the previous movies revenue, by using this approach they have been able to predict the expected revenue for a movie. They carried out the linear regression, K-means clustering, weighted linear regression and polynomial regression to predict and improve the prediction accuracy [7]. Yang evaluated the substitution effect on pirated movies by establishing a dataset include the information on online piracy, theater showings, and revenues, and using the structural estimation to show that the average piracy loss is 64% and the DID regression resulted that although the anti-piracy actions has been shut down on the website, the box office revenue of movies would increase 60% during the short period [8]. Jehoshua, et al. built a forecasting and risk management tool based on the movie script. They extracted three levels of textual information from the movie script and incorporated these and budget as predictors into a BART-QL model and received the posterior predictive distributions. Their predictive distributions of box office revenues can help movie producers to determine how their movies should be produced [9]. Andy built linear regression models to examine the relationships between gross revenue, length of movie, and movie budget. By using this approach, the research received the result that there is a positive relationship between gross revenue and budget and the low correlation between the movie runtime and revenue [10]. In this paper, I will use the linear regression model to evaluate the characteristics of movies' revenue

2. Methodology

2.1. Data

The dataset of this paper is from Kaggle website which is called TMDb Movies dataset which contains the information of more than 10,000 movies, the characteristics including budget, cast, revenue, genres, release date, runtime, vote average etc.

2.2. Characteristics

The raw data includes 21 characteristics. This paper chooses to focus on the revenue, popularity, budget, runtime, vote average to analyze the factors that influence the revenue. In substance, the most important indicator to measure the economic performance of a film is box office revenue. Furthermore, the popularity can effectively reflect the public's attention to the film, so that its potential impact on the final box office revenue can be analyzed. Meanwhile, the budget is also an influential factor because of the budget determines the upfront investment of the film and can also infer the quality of a film to a certain extent. The quality of a film also affects the audience's expectations before its release and viewing experience during the movie time. The amount of upfront investment can predict the final revenue of the movie potentially. Furthermore, the runtime of a movie should be considered as well, due to the runtime will affect whether the content is full or not, the content of a movie will cause what if people would like to pay for the ticket. Many viewers will score the movie after watching it, and the vote average is the average of all the viewers' scores for a movie. The vote average infers that whether they will choose this specific movie to watch or recommend to others. Nowadays people like to read reviews, are more cautious and have some influence on whether they will eventually choose to see them. If a movie has very few people voting

for it, then of course people will rightfully assume that the movie isn't worth watching. As table 1 shows.

Table 1. Variables description

Variables	Variables Explanation
Revenue	Specific amount of box office
Popularity	The popularity score of TMDB
Budget	The investment of the movie
Runtime	Duration of a movie (by min)
Vote Average	The average that people whether recommend

After the selection of the characteristics of the analysis for box office revenue, therefore base on these 5 variables to build a linear regression model to analyze the factors that influence the box office revenue. Choosing the columns including Revenue, Popularity, Runtime and Vote average that the factors chosen before to make a new data set, after that, importing the data set to the R studio. Conducting the exploratory data analysis with plotting the relationship between the revenue and other variables to do some visual analysis like making the scatter plots of the variables to find their relationship by using R studio. Accordingly, to build the linear regression model of the response variable which is Revenue and independent variables that are Popularity, Budget, Runtime and Vote average. The model of the linear regression is: $\text{Revenue} = \beta_0 + \beta_1 * \text{popularity} + \beta_2 * \text{budget} + \beta_3 * \text{runtime} + \beta_4 * \text{vote average}$. By applying the backwards elimination method and adjusting the p-values to drop the redundant variables. Finally receive the coefficient of these independent variables and obtain the regression function.

3. Results and discussion

During the exploratory data analysis, after the pre-processing of the data set, by using R. Figure 1 shows the distribution of the revenue. The distribution of revenue is the right-skewed distribution, with the mean is 3.982×10^7 and the peak is the $2,782 \times 10^9$.

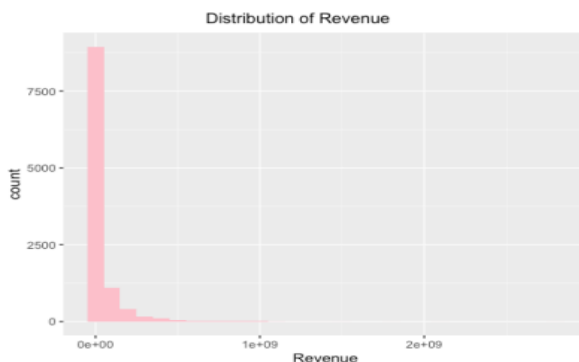


Figure 1. The Distribution of Revenue

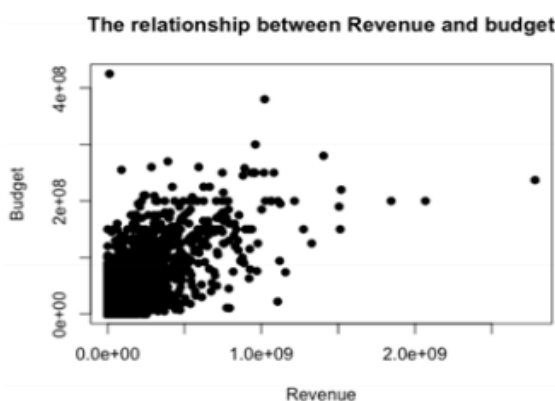


Figure 2. The Relationship between Revenue and Budget

Figure 2 shows the relationship between Revenue and Budget, the scatter plot illustrates there is a positive and strong strength of relationship between Revenue and Budget as the correlation is 0.7349. Therefore, it can be inferred that as the budget increases, the revenue would increase as well. The budget is also an important indicator of the quality of the film. Nowadays, people pay more and more attention to the content of the film, the acting skills and the details of the special effects, etc. Thus, with a high budget, the crew can enhance the quality of the movie.

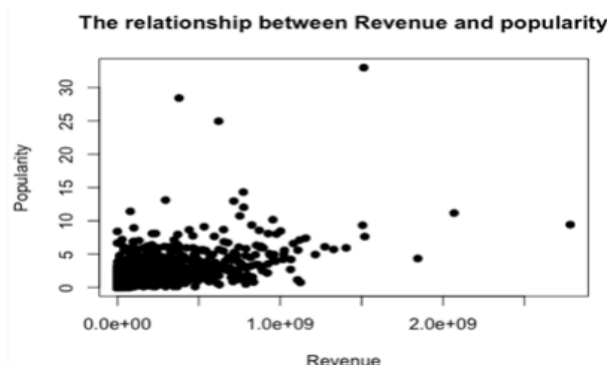


Figure 3. The Relationship between Revenue and Popularity

The relationship between Revenue and Popularity was shown in Figure 3. Revenue and Popularity also has a positive relationship, and the correlation is 0.6634, so it is a moderate strength of positive relationship. Popularity also has some impact on the revenue. Having popularity is proof of publicity and exposure, a way to make more people aware about this movie. People often chase popularity nowadays, and as long as more people know about the movie, box office revenues will go up.

Figure 5. The Relationship between Revenue and Vote Average.
The relationship between Revenue and Runtime

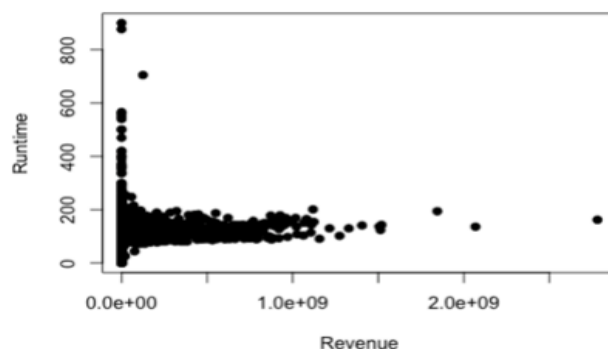


Figure 4. The Relationship between Revenue and runtime

Seeing the Figure 4, the scatter plot does not show any association between runtime and revenue, there is a very weak relationship between runtime and revenue as the correlation is just 0.1628. Therefore, the runtime in this data set does not have any effect on revenue.

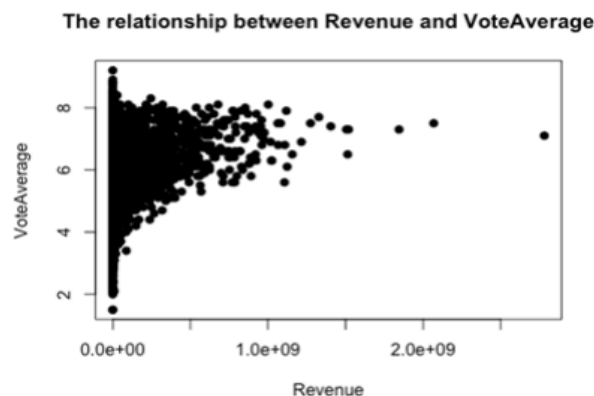


Figure 5. The Relationship between Revenue and VoteAverage

From the Figure 5, the relationship between revenue and vote average has been shown that they have a weak relationship because the correlation is 0.1726. But there is a little bit of positive association that can be seen from the scatter plot. How exactly we still have to explore by building a linear regression model.

In this paper, building the linear regression model after the pre-processing of the data set by using R. After calling the full model, the result is Table 2. From the Table 2, it is obvious that the variable which is runtime has the significant p- value, the others have the insignificant p-value. Statistical results show that the impact of runtime on revenue is not significant. Therefore, this paper dropped the variable runtime that has the most significant p-value by using the backwards elimination, and refitting a new regression model.

Table 2. Linear Regression Model

	Estimate	Std. Error	t value	Pr (> t)
(Intercept)	-5.776e+07	4.659e+06	- 12.399	<2e- 16
Popularity	4.233e+07	8. 188e+05	51.575	<2e- 16
Budget	2.018e+00	2.627e-02	76.831	<2e- 16
runtime	8.499e+03	2.208e+04	0.385	0.7
Vote average	6.678e+03	7.436e+05	8.981	<2e- 16

Table 3. Refitting Regression model

	Estimate	Std. Error	t value	Pr (> t)
(Intercept)	-5.716e+07	4.385e+06	- 13.035	<2e- 16
Popularity	4.233e+07	8. 187e+05	51.587	<2e- 16
Budget	2.020e+00	2.5992-02	77.721	<2e- 16
Vote average	6.718e+06	7.363e+05	9.124	<2e- 16
R-Square		0.6408		

The refitting model has been shown on Table 3. From Table 3, they do not have an explanatory variable that has the significant p-value, so a parsimonious model has been attained. Also, the Table 3 demonstrates that the adjusted R-square is 0.6408 which means that the model has a relatively good fit. The p-value of all independent variables is less than 0.05, it indicates that when the significance level $\alpha = 0.05$, the independent variables popularity, budget and vote average are significant. Hence, the variables Popularity, Budget and Vote average can explain the response variable Revenue very well.

Therefore, after the transformation of the scientific notation, the final model is:

$$\text{Revenue} = 4323809 + 4233000 * \text{Popularity} + 2.02 * \text{Budget} + 6718000 * \text{Vote Average} \quad (1)$$

In this function, the popularity and the vote average can affect the revenue much more. The coefficient of popularity is 4233000 which means that with all other conditions being equal, the popularity increases one unit, the model predicts that the revenue will increase 4233000. This illustrates that movies are leading the way right now, and everyone tends to catch up with what's trendy at present. Secondly, the coefficient of budget is 2.02, it conveys that if other factors hold in constant, when the budget has been increased 1 unit, the revenue would increase 2.02 more. This proves that the bigger the investment and the higher the budget, the higher the probability of making a high-quality movie that audiences will desire to watch. The coefficient of vote average is 6718000 which represents that assuming that for every one unit increase of the vote average, the revenue will increase 6718000. This manifests in the audiences improving their aesthetics, paying more attention to the praise of movies, and will look ahead to see whether some other people support the movie before deciding to go see it, and they are increasingly demanding for movies.

4. Conclusion

This paper addresses the factors that influence the revenue so that to build the linear regression model to analyze the influence of popularity, budget, runtime and vote average. Based on the data set used in this paper, The model displays that the popularity, budget and vote average of a movie has a significant impact on the final box office revenue of a movie. The coefficient of the vote average has the biggest value. 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 a movie to see it or not. Meanwhile, the film industry is developing rapidly worldwide, and film producers, directors and actors are paying more and more attention to the word-of-mouth of their films. Nonetheless, the film industry is developing and people's aesthetic level is improving. On this basis, producers can only improve their own strength of creating the scripts by performing their imagination, studying diverse shooting techniques and looking for a new, diverse and strong acting ensemble. At the same time, the producers need to pay more attention to the objective evaluation of the movie by the audience. The factor with the second largest coefficient is the popularity. In this era, people are more willing about the pursuit of popularity. Under these circumstances, the action of promoting a movie before and during its release is particularly vital. The effect of publicity can be achieved through social networks, offline roadshows, TV shows, etc. There is also an approach which is to encourage audiences to actively post their thoughts and comments about the film on the Internet. The factor with the thirdly largest coefficient is the budget. From the regression model, the budget is directly proportional to the box office revenue. A good production cannot be separated from the upfront investment. Because the budget to some extent determines the director as well as the producer's room to play. The higher the budget, the more the producer or director can develop new technology, new special effects and hire better actors to attract more paying audiences. In this article, only one algorithm which is a linear regression model was used to analyze the factors influencing box office revenue, which is a great limitation. At the same time, the selection of factors will be more extensive, for example, film genre and release time are also possible influencing factors of box office revenue, so that the model of influencing factors of movie revenue can be more completed.

References

- [1] Huang, W. Q., et al., "Analysis and forecast of factors influencing movie box office," *Information Technology and Informatization*, 173- 178 (2019).
- [2] Hu X., Li, B., Wu, Z, P., "The Analysis of the Factors Which Influence Film Box Office," *Journal of Communication University of China (Science and Technology)*, 62-67 (2013).
- [3] Wang, Y. F., "Analysis of factors influencing China's movie box office," *Marketing Research*, 30-32 (2015).
- [4] Ni, S. M., "Analysis of factors influencing movie box office," *Reform and Opening*, 119- 120, 122(2016).
- [5] Gao, Q., "Analysis and prediction of factors influencing movie box office based on fully connected neural networks," *Qufu Normal University*, (2021).
- [6] Sehwan, O., et al., "Predictive value of video-sharing behavior: sharing of movie trailers and box-office revenue," *Internet Research* 27(3), 691-708 (2017).
- [7] Apte, N., Forssell, M., and Sidhwa, A., "Predicting movie revenue," *Stanford University*, (2011).
- [8] Yand, Y., "The effects of movie piracy on box-office revenue: an empirical analysis of the Chinese movie market," *Journal of Applied Economics* 23(1), 618-655 (2020).
- [9] Eliashberg, J., et al., "Green-lighting Movie Scripts: Revenue Forecasting and Risk Management," *NYU Stern*, (2010).
- [10] Chen, A. W., "A statistical analysis of gross revenue in movie industry," *International Journal of Business Management and Economic Research* 9(3), 1276- 1280 (2018).