Movie Revenue Prediction Through Regression Analysis

Student's Name

Institutional Affiliations

Course Title

Professor's Name

Date

Movie Revenue Prediction Through Regression Analysis

Introduction

The movie industry in the USA and other major economies like China has experienced significant growth. In the USA, the company has an average investment of $60 million per film and produces approximately 500 movies annually (Ahmad et al., 2020). Despite this exponential rise in the production budget, this industry's revenue production or profitability has remained largely uncertain. Statistical analysis of the historical data of Chinese movies has revealed that only a fraction of the films released in the first half of 2013 were profitable. Such uncertainties may compound the risks associated with investing in this sector (Ahmad et al., 2020). Therefore, designing machine learning models to predict movie revenues may benefit investors, producers, and theatres. The current study uses multiple linear regression analysis to determine the significant predictors of movie revenues.

Research Question: What factors are significantly associated with movie revenue?

Methodology

Research Design

The current study employs a secondary research design. It uses pre-existing data from reliable and credible sources to answer the research question. The main objective is to determine the factors significantly associated with movie revenue.

Data

The dataset (Full TMDB Movies Dataset 2023) used in this study was downloaded from Kaggle, one of the most reliable data repositories for machine learning models. This dataset contained 15 variables with 45463 observations (rows). The data set includes features like revenue, budget, popularity, votes, release dates, and runtime, which can be statistically analyzed to answer this study's research question. While the dataset was downloaded from Kaggle's database, it is originally from the Movie Database (TMDb), known for its longitudinal collection and storage of accurate movie or film information. TMDb collects information from various sources, primarily relying on user contributions. It operates on a crowdsourced platform where movie producers, developers, and enthusiasts submit movie-related data. These contributions are edited and reviewed to ensure data accuracy and quality. For this reason, TMDb's data is relatively reliable and valid. However, due to its crowdsourced nature, there may be missing data, inaccuracies, and variations in data that should be removed or validated before using the data.

Therefore, the dataset was cleaned and transformed before creating a linear regression model to predict movie revenue. Since the linear regression model uses continuous (numerical) data, all the nonnumerical features (characters and factors) are dropped from the dataset. These variables include movie ID, adult, title, status, original language, and others. The remaining variables are further processed to remove missing values. Since revenue is the target variable and the budget for producing a movie cannot be $0, all the zero revenue and budget observations are dropped from the dataset. The zero rows for revenue and budget do not contribute meaningful information to the prediction. Removing them helps to reduce redundancy and noise in the data. It also balances the data and allows the model to identify and predict revenue. All missing values in the resulting dataset are omitted (deleted) listwise. The dataset was filtered to get movie records produced between 2000 and 2020. The dataset dimension was reduced to 3419 rows and 8 variables.

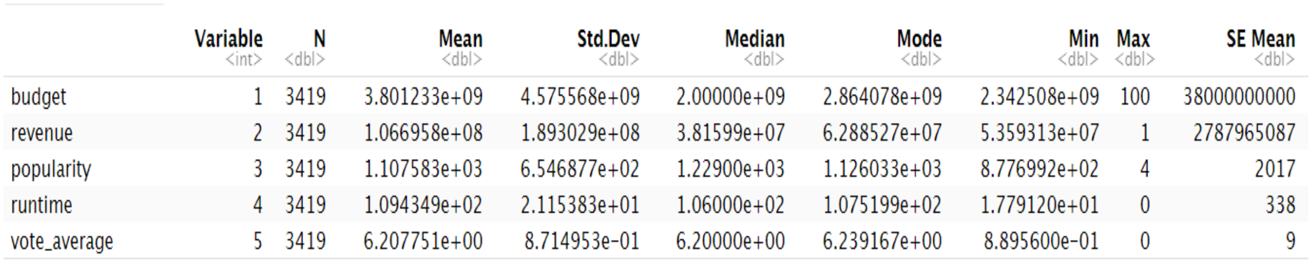
Data Analysis

Descriptive statistics summarize the data to examine the characteristics and uncover hidden patterns and trends. Descriptive statistics like mean, median, standard deviations, and others are calculated to understand the distribution of different variables. Inferential analysis (a linear regression model) is used to determine if there are significant associations between the target variable and its predictors. The dataset is split into training and testing sets, with the training set constituting 70% and the testing set 30%. A regression model is trained with revenue as the dependent variable and the other variables as the predictors. The model is used to predict revenue utilizing the test data. The model is evaluated based on the coefficient of determination (R-Square), the beta coefficients of the predictors, the p-values associated with their t-statistics, and the F-test statistic. Normal probability plots (Q-Q plots) and scatter diagrams are used to visualize the model's performance.

The inferential tests were performed with a confidence level of 95% $(level\ of\ confidence,\ \alpha = .05)$. All calculations, data transformations, cleaning, and modeling are done through R programming (RStudio).

Results

Descriptive Statistics



	Variable <int>	N <dbl>	Mean <dbl>	Std.Dev <dbl>	Median <dbl>	Mode <dbl>	Min <dbl>	Max <dbl>	SE Mean <dbl>
budget	1	3419	3.801233e+09	4.575568e+09	2.00000e+09	2.864078e+09	2.342508e+09	100	38000000000
revenue	2	3419	1.066958e+08	1.893029e+08	3.81599e+07	6.288527e+07	5.359313e+07	1	2787965087
popularity	3	3419	1.107583e+03	6.546877e+02	1.22900e+03	1.126033e+03	8.776992e+02	4	2017
runtime	4	3419	1.094349e+02	2.115383e+01	1.06000e+02	1.075199e+02	1.779120e+01	0	338
vote_average	5	3419	6.207751e+00	8.714953e-01	6.20000e+00	6.239167e+00	8.895600e-01	0	9

Table 1: Summary statistics for the variables

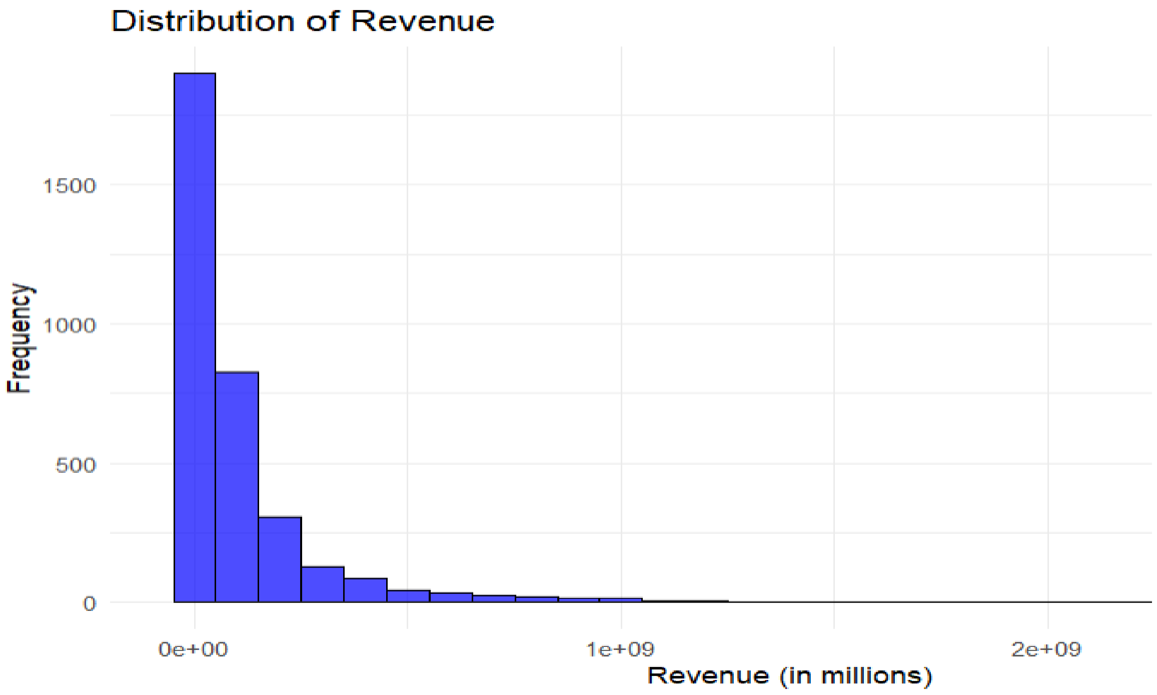


Figure 1: The distribution of revenue

Table 1 above shows the summary statistics of all the variables included in this study. The average revenue that the movie industry accrued from 2000 to 2017 is $(M = 106695800, SD = 189302900)$. Most movies made $107 million in revenue between 2000 and 2020. However, there are considerable variations in the revenues, as indicated by the relatively large standard deviation. While some movies earned huge revenues, some might have made dismal revenues. Such inconsistencies or variations may be associated with many factors. The median income between 2000 and 2017 was $(Mdn = 38159900)$, implying that half of the movies produced between 2000 and 2017 earned more than $38 million, while the other half earned less than $38 million in revenues. Since the median revenue is significantly less than the average, the revenue is positively skewed. Figure 1 above shows the distribution of revenue. The budgets for these movies also vary significantly. For the same period, an average of

$(M = 3801233000, SD = 4575568000)$ USD was spent on producing films. These statistics show that the budget for movie production has been relatively high.

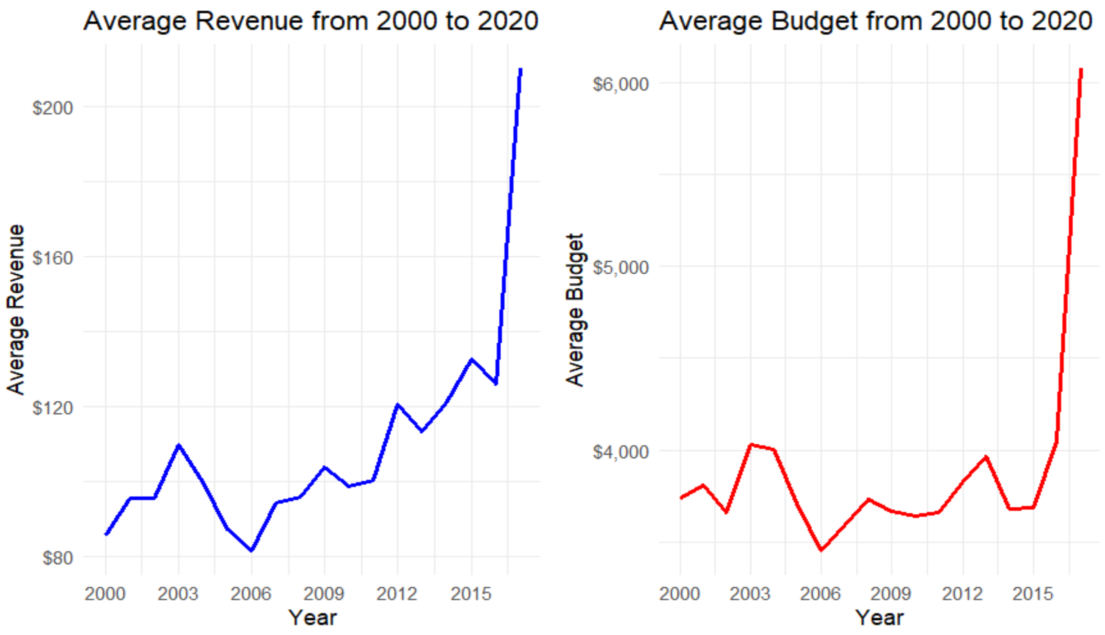
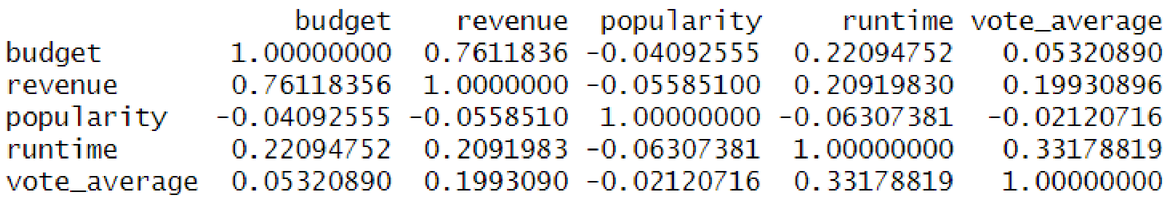


Figure 2: The trend of movie revenue and production budgets from 2000 to 2017

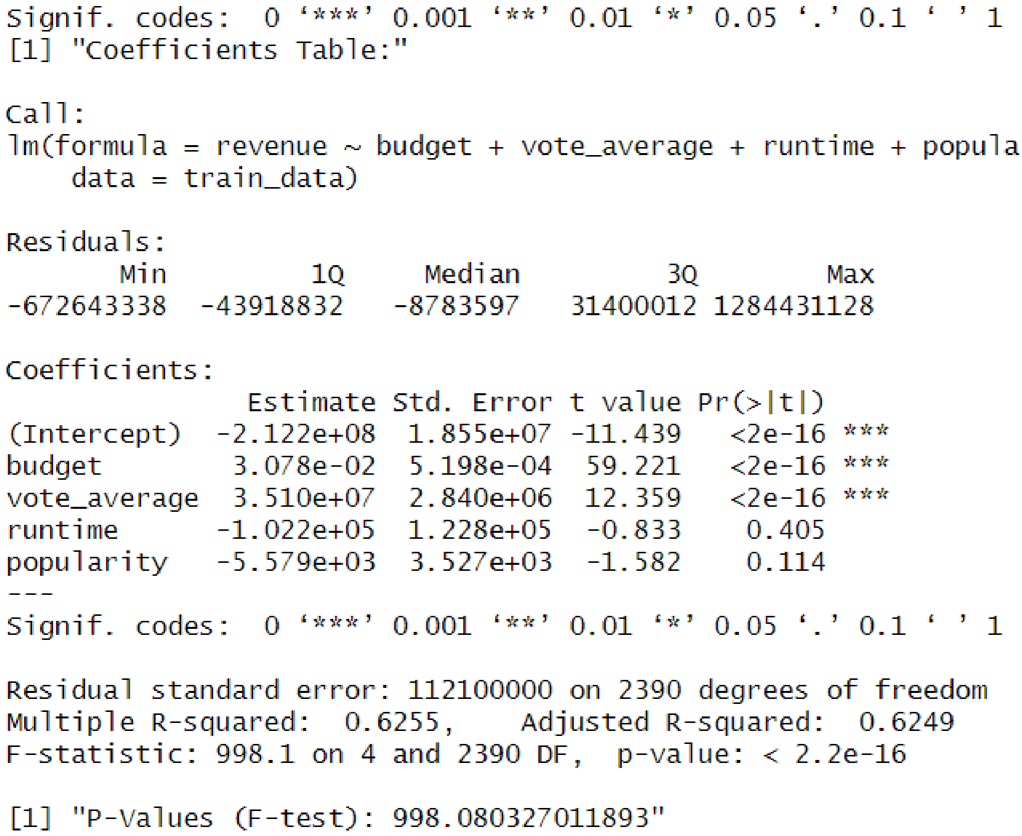
Figure 2 above shows the trend of the annual average costs of producing movies and the revenues earned between 2000 and 2017. The movie production budgets and revenues have exhibited somewhat similar tendencies. While the budgets and revenues dwindled between 2003 and 2006, they trended seasonally between 2006 and 2012. From 2015, the budget and revenues skyrocketed. The correlation matrix in Table 2 below shows a strong positive association between income and budget. As the budget increases, the revenue also increases. Revenue is also positively correlated with runtime and vote average. However, these associations are relatively weak. Revenue has a weak negative association with popularity.



```
                 budget    revenue  popularity     runtime vote_average
budget       1.00000000  0.7611836 -0.04092555  0.22094752   0.05320890
revenue      0.76118356  1.0000000 -0.05585100  0.20919830   0.19930896
popularity  -0.04092555 -0.0558510  1.00000000 -0.06307381  -0.02120716
runtime      0.22094752  0.2091983 -0.06307381  1.00000000   0.33178819
vote_average 0.05320890  0.1993090 -0.02120716  0.33178819   1.00000000
```

Table 2: Correlation matrix of the variables

Regression Model



```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Coefficients Table:"

Call:
lm(formula = revenue ~ budget + vote_average + runtime + popularity, 
    data = train_data)

Residuals:
       Min         1Q     Median         3Q        Max 
-672643338  -43918832   -8783597   31400012 1284431128 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.122e+08  1.855e+07 -11.439   <2e-16 ***
budget        3.078e-02  5.198e-04  59.221   <2e-16 ***
vote_average  3.510e+07  2.840e+06  12.359   <2e-16 ***
runtime      -1.022e+05  1.228e+05  -0.833    0.405    
popularity   -5.579e+03  3.527e+03  -1.582    0.114    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 112100000 on 2390 degrees of freedom
Multiple R-squared:  0.6255,    Adjusted R-squared:  0.6249 
F-statistic: 998.1 on 4 and 2390 DF,  p-value: < 2.2e-16

[1] "P-Values (F-test): 998.080327011893"
```

Table 3: Regression model coefficients, r-squared, and f-test results

The model indicates that the budget ($\beta_{budget} =.0.031,\ t(2390) = 59.22,\ p <.001$) and vote average ($\beta_{vote\_average} = 35100000,\ t(2390) = 12.36,\ p <.001$) are significant positive predictors of movie revenues. If the movie budget increases marginally (by one million dollars), the revenue will increase by 0.031 dollars *ceteris paribus* (the other factors remaining constant). The effect size of the production budget seems small, but it may be practically significant, since the budget is measured in millions of dollars. Similarly, if the vote average increases marginally, the revenue will increase by 35100000, *ceteris paribus*. However, popularity ($\beta_{popularity} =- 5579,\ t(2390)\ =- 1.58,\ p\ =.114$) and runtime ($\beta_{runtime} =- 102200,\ t(2390)\ =-.83,\ p\ =.405$) as the p-values associated with their beta coefficients are not greater than the significance level. Therefore, movie revenue can be predicted as follows:

$Revenue\ =\ 0.031budget\ \ +\ 35100000average\_vote\ -\ 212200000$

The variables in the model explain about 63% of the variation in the movie revenue, $F(4,\ 2390)\ =\ 998.08,\ p\ <.001,\ R^2\ =.62$. The model fits the data well.

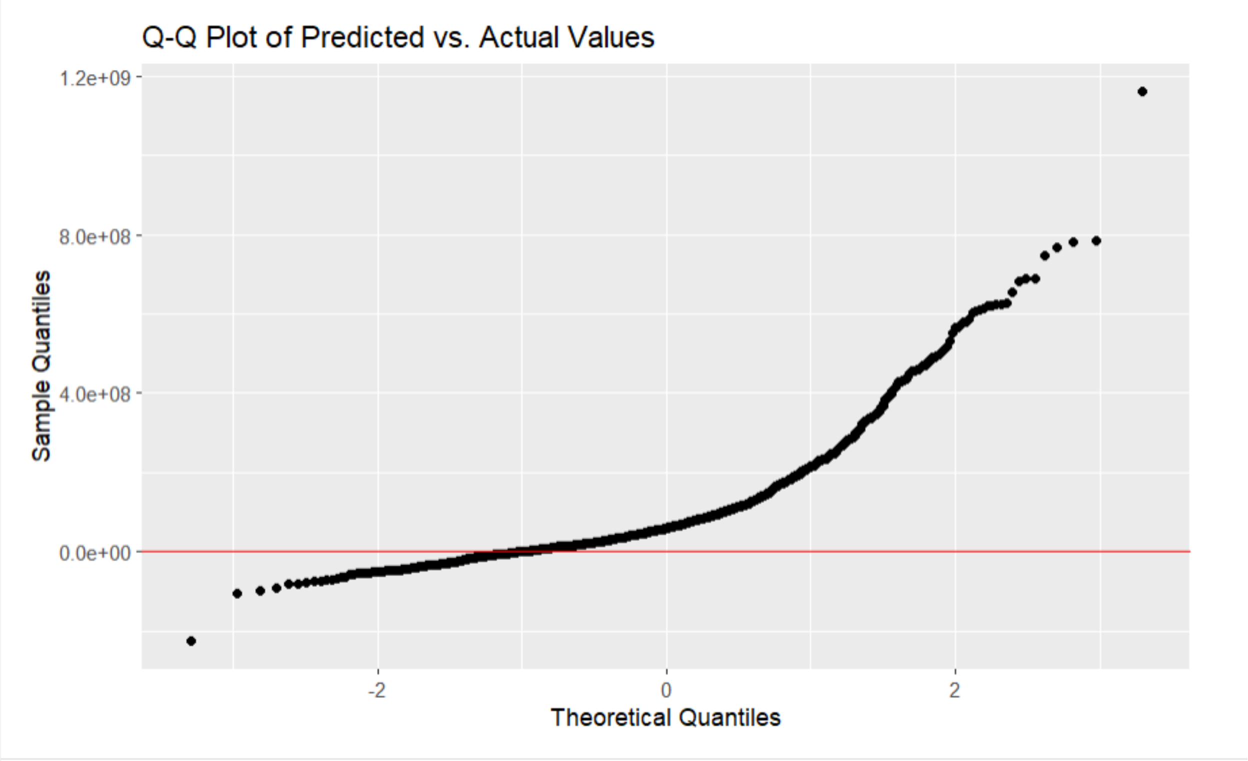


Figure 3: Q-Q Plot showing the distribution of residuals

Figure 3 above shows data quantile deviations from the straight line. It is inferable that the quantiles deviate significantly from the straight line, meaning that the normality assumption was not met.

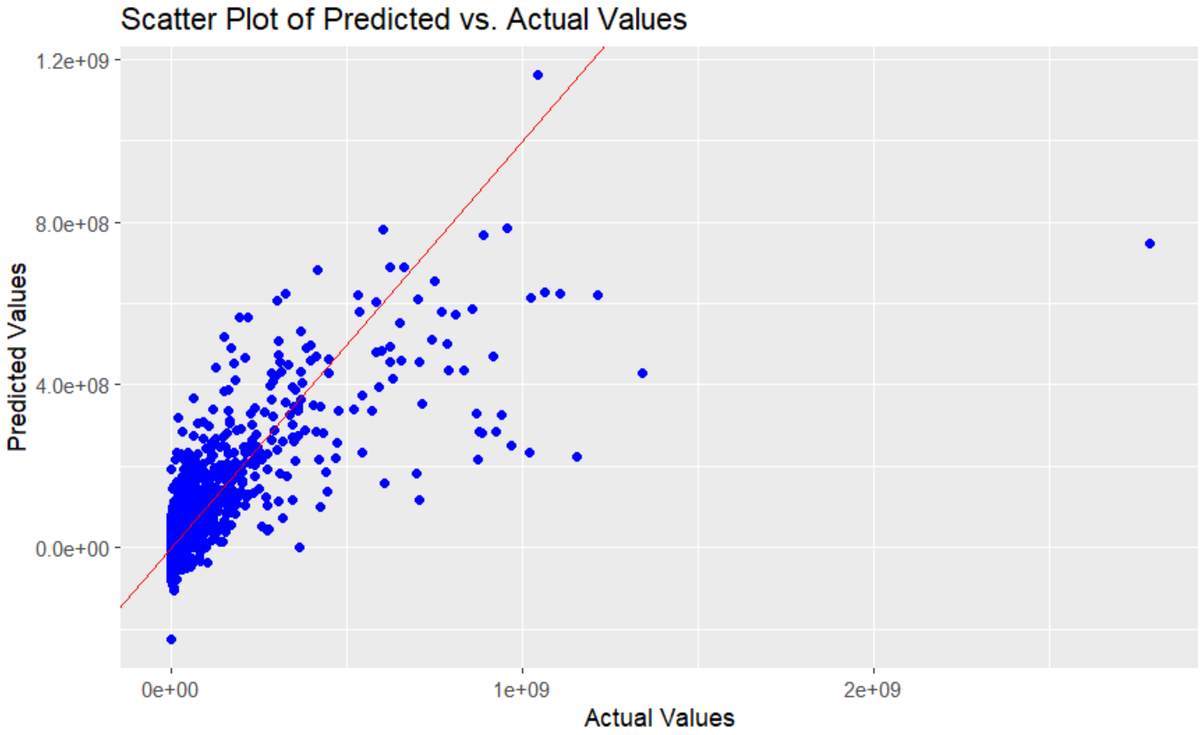


Figure 4: The predicted versus the actual values

Most predicted revenue values cluster around the straight line representing the actual revenue values. The model is relatively accurate in predicting the revenue values. However, there are some outliers in the data.

Discussion and Conclusion

The current study's findings show a positive correlation between movie revenue and both budget and vote average increases. These findings are consistent with the findings of previous studies. Ahmad et al. (2020) found that reviews and budgets positively affect movie revenue. However, this study has certain limitations. Firstly, using a multiple linear regression model makes assumptions about the linearity of relationships between predictors and the response variable, which may oversimplify the intricate factors influencing movie revenue.

To obtain more precise predictions, employing a robust machine learning model such as Random Forest could be beneficial, as it can handle categorical data and capture nonlinear relationships effectively. Additionally, the investigation could broaden its scope to examine potential modifications in the film industry's structure. This might involve analyzing the effects of technological advancements and economic disruptions like COVID-19 on revenue trends.

Moreover, enhancing the accuracy of predictions could be achieved by addressing data anomalies and refining assumptions within the model, such as normality. In conclusion, although this study aligns with previous research findings, overcoming these limitations and employing advanced machine-learning approaches is essential for a more nuanced understanding of factors influencing movie revenues.

References

Ahmad, I. S., Bakar, A. A., & Yaakub, M. R. (2020). Movie revenue prediction based on purchase intention mining using YouTube Trailer Reviews. *Information Processing &amp; Management*, *57*(5), 102278. https://doi.org/10.1016/j.ipm.2020.102278

Souza, T. L. e, Nishijima, M., & Pires, R. (2023). Revisiting predictions of movie economic success: Random Forest applied to profits. *Multimedia Tools and Applications*, *82*(25), 38397–38420. https://doi.org/10.1007/s11042-023-15169-4