Decoding Movie Success

A Cross-Sectional Study Research of the Factors Driving Movie Success in 2018

[redacted] | [redacted] | Hugo Alves [redacted] | [redacted]

INTRODUCTION

This study focuses on evaluating the economic success of mainstream movies, emphasizing gross income as a key metric. The decision to analyze movies released until the end of 2018 is based on the identification of 2018 as the most financially profitable year for the film industry in the last decades, with global revenues exceeding 40 billion US dollars (1,2). This choice is supported by a prior analysis of databases on worldwide box office revenue per year. Although 2019 was similarly profitable, it was not considered due to the greater availability of information in 2018 (1,2).

The proximity of 2018 to the emergence of the Covid-19 pandemic is another factor influencing the choice of the study period. The subsequent restrictions on theater attendance due to the pandemic raised questions about whether pre-epidemic box office values, as seen in 2018, will ever be reached again. This makes 2018 a compelling year for analysis, representing a time when streaming services were not as dominant in our lives. The study aims to provide insights into audience engagement, marketing effectiveness, and industry trends through the lens of gross income, considering the unique circumstances surrounding 2018 and the subsequent impact of the Covid-19 pandemic on the film industry.

PROBLEMS AND OBSTACLES

- The analysis faces challenges in data collection, with industry secrecy about certain information, such as marketing budgets, and difficulties in obtaining metrics like the popularity of the cast.
- Release dates discrepancies arise from different sources considering either worldwide releases or the first public screening in the production country or in the US. In this study we chose the former.
- While superstar cast analysis was initially considered (3), practical challenges led to simplification, focusing only on lead actors' gender. Similar simplification was applied to directors and writers.
- The study also recognizes the significance of studio types (3), but we opted for a simplified approach, categorizing studios as major (e.g., Walt Disney Studios, Warner Bros. Pictures) or non-major, given the complexity of detailed discrimination.

METHODOLOGY

Data was collected from two main sources: GitHub and Kaggle (4,5). No pre- or post-2018 information was collected, in order to preserve the nature of the study. However, less than 10 films had their metascore filled with information collected in 2023 since they were missing, but in this case, we will consider it as collected at the same time as the remaining data, since film critic ratings are issued when the release of the film (remain unchanged).

All preprocessing, data cleaning and feature engineering were carried out in Excel (Power Query) from CSV files extracted from the sources mentioned above.

Included features: release date, movie name, age rating, country, metascore (critic score), IMDb score (audience score), number of IMDb votes, gender of leading actor/actress, director and writer, budget, gross income, runtime, genre (action, adventure, animation, biography, comedy,crime drama, family,fantasy, horror, mystery, thriller, war, romance,musical, sci-fi,western and sport) and total genres.

On our approach and investigation, we used **ANOVA**, **Pearson correlation test**, **3 different Multiple Linear Regression Models**, **Variation Inflation Fator** (to assess for multicollinearity), **Joint Significance testing** (F-test), **Breusch-Pagan and White Special Test** (Heteroskedasticity assessment), **White heteroscedasticity-robust standard errors estimator** (to account for Heteroskedasticity)**, t-test on the mean of residuals** (assess if it is very different from 0) and finally **RESET** (Ramsey Regression Equation Specification Error Test).

RESULTS

Test	p-value
ANOVA (gross income VS month of release)	0.504
Joint Significance Test (F-test)	2.2e-16 ***
Breusch-Pagan Test	0.06457
White Special Test	0.001786
RESET	0.136
t-test (residuals)	1

Table 1 - Performed test and respective p-values

Final Multiple Linear Regression Model	Coefficient estimate	p-value
metascore	0.018	0.01017 *
log(budget)	0.946	1.56e-11 ***
crime	-0.796	0.00409 **
war	-2.760	6.00e-06 ***
Adjusted R^2=0.5835	p-value: < 2.2e-16	

Table 2 – Multiple Linear Regression Model (with statistically significant predictors and respective coefficients, goodness-of-fit value and p-value for the model)

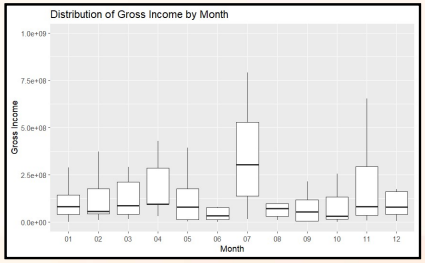

Figure 1 – Box-plots for gross income in each month of 2018 (to simplify analysis all the outliers in each month where removed)

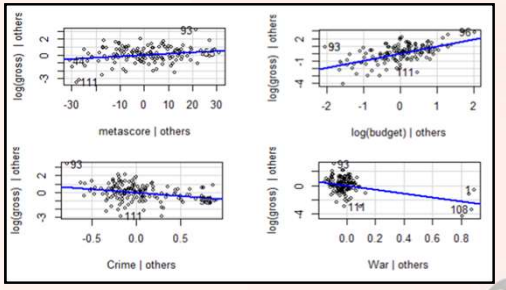

Figure 2 – Partial Regression plots for the statistically significant variables (metascore, log(budget), crime and war). Each partial regression plot helps visualize the relationship between each important independent variable and the dependent variable while holding other variables constant.

CONCLUSIONS

After analyzing the regression model, we understood that there were 4 statistically significant variables: **Metascore**, **log(Budget)**, **Crime** and **War genres**. Through the values of the coefficients, we concluded that: **if there is one-unit increase in Metascore, it is expected that the gross income will increase by around 1.81%; if there is a 1% increase in Budget, the gross income is expected to increase by approximately 0.95%; if the movie is within the Crime or the War genre it is expected to occur a decrease in gross income (around 54.91% and 93.67%, respectively).** Even though all these results are statistically significant, only the **crime and war genres are economically significant**, since they are the only variables to have a substantial impact on gross income (negative impact). Another interesting aspect was the relationship between **month of the release and gross income**, which it was statistically insignificant in the ANOVA test and marginally significant in the t-test, but nevertheless had distribution of box office success in accordance with what would be expected: **worst performing months/dump months being January, February, August and September, while July and November had movies premiering with the highest gross income** (6) . Perhaps, to obtain statistical evidence, a larger sample or some attention to dealing with outliers could be necessary. It is also worth mentioning that the **model we produced can explain 58.35% of the variability in log(gross income**).

When accounting for heteroskedasticity, we got the same coefficient values that we achieved initially. This can mean that either heteroscedasticity might not be a significant concern or that the heteroskedasticity-robust standard errors are not altering the parameter estimates greatly. Since the Breusch-Pagan test suggested that there was no strong indication of heteroskedasticity, **we assumed homoskedasticity** to simplify (but further exploration would be advised). When applying RESET test, there was **no evidence for the need of nonlinear terms**.

REFERENCES

1. https://www.statista.com/statistics/264429/global-box-office-revenue-by-region/
2. https://www.boxofficemojo.com/year/?area=XWW
3. https://swer.wtamu.edu/sites/default/files/Data/137-148-89-338-1-PB.pdf
4. https://github.com/danielgrijalva/movie-stats
5. https://www.kaggle.com/datasets/miazhx/metacritic-movie-reviews
6. https://academic-accelerator.com/encyclopedia/dump-months