Exploring the Key Factors That Drive Success in the Video Game Industry*

Sakhil Goel

April 19, 2024

In the rapidly evolving world of video games, understanding the elements that contribute to a game's success is essential for developers and marketers aiming to capture the attention of diverse and dynamic player bases. This study utilizes data from the RAWG Video Games Database API to examine how various factors—namely platform availability, genre diversity, playtime, and release year—affect the success of video games, as measured by Metacritic scores. games available on more platforms, longer playtimes and more recent release years, generally achieve higher success. These findings emphasize the need for strategic platform coverage and timely, engaging content as key to maximizing a video game's success in a competitive market.

Table of contents

*Code and data are available at: https://github.com/Sakhil-Goel/Game-Success.git.

1 Introduction

In the fast-growing video game industry, identifying the factors that lead to a game's success is vital for developers and publishers aiming to optimize their impact on the market. As gaming technology advances and player preferences change, it is increasingly important to understand the strategies that contribute to high Metacritic scores and strong player engagement. While many studies have explored different aspects of video game success, there is still a gap in understanding how multiple factors such as platform availability, genre diversity, playtime, and release timing interact to affect success.

This paper addresses this gap by examining the influence of these variables on game success using data from RAWG (2024). Through Bayesian logistic regression, we analyze how platform diversity, genre count, playtime, and the timing of game releases correlate with high Metacritic scores. Our study expands beyond typical metrics to include variables like playtime and release year, capturing broader trends that affect game performance.

Our results show that games available on more platforms generally perform better, while having a wide range of genres does not necessarily improve performance. Additionally, games with longer playtimes and those released more recently tend to be more successful, indicating that players prefer engaging and up-to-date content. These findings highlight the importance of strategic platform distribution and focused content development for achieving success in the video game industry.

In the subsequent sections of this paper, we outline our research process, findings, and implications as follows: Section 2 dives into our methodology, as well as our data cleaning and measurement techniques. This section provides the foundation for understanding the variables and data structure used in our analyses. In Section 3, we detail the modeling approach, including the statistical methods employed to assess the impact of platform availability, genre

diversity, playtime, and release timing on game success. Following this, Section 4 presents our findings, using graphs and tables that clarify the statistical relationships uncovered in our analysis. This section aims to provide a clear presentation of how different factors contribute to the success of video games. Finally, Section 5 explores the implications of our findings, discussing their relevance to the video game industry. Here, we also address any surprising results, acknowledge the study's limitations, and suggest directions for future research. This section intends to contextualize our findings within ongoing industry challenges and debates.

By clarifying which factors are most effective in enhancing a video game's success, this paper provides valuable insights for industry stakeholders and supports the development of more targeted game production and marketing strategies.

2 Data

2.1 Data Source

The primary dataset for this study was obtained from the RAWG Video Games Database API RAWG (2024), which is an extensive repository of video game information. The RAWG API provides detailed data on video games, including their titles, genres, platforms, release dates, and user ratings such as Metacritic scores. Access to this data was facilitated through an API key obtained by registering on the RAWG.io website, which allowed for unrestricted access within the limits prescribed by the API guidelines.

2.1.1 Variables and Selection Criteria

The variables selected for analysis were specifically chosen for their potential influence on a video game's success. These include:

Name: The title of the video game. Genres: The categories or genres assigned to the game (e.g., Action, Adventure). Platforms: The gaming platforms on which each title is available (e.g., PC, Xbox). Release Date: The official release date of the game. Playtime: The average gameplay time reported by users. Metacritic: The Metacritic score of the game, used as a proxy for critical success. Each of these variables was deemed essential for analyzing the factors that could predict a video game's success, with the Metacritic score serving as the dependent variable in subsequent modeling.

2.1.2 Data Collection Process

Data was collected via a scripted series of GET requests to the RAWG API. The requests were designed to paginate through the API's response, ensuring comprehensive data retrieval. Each request fetched data on 40 games per call—the maximum allowed by the API—with the

script automatically handling pagination by incrementing the page number until all available data was downloaded.

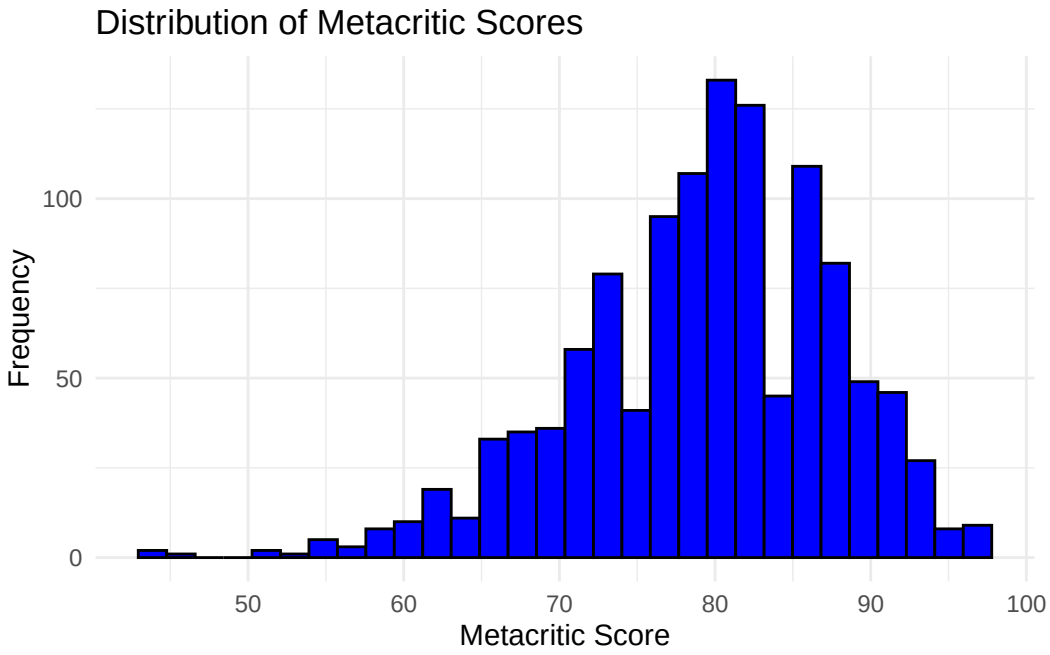

Figure 1: Distribution of Metacritic Scores

2.2 Methodology

Data analysis is performed using the statistical programming language `R` (R Core Team 2023), with packages `lubridate` (Grolemund and Wickham 2011), `tidyverse` (Wickham et al. 2019), `dplyr` (Wickham et al. 2023), `tidyr` (Wickham, Vaughan, and Girlich 2023), `knitr` (Xie 2014), `janitor` (Firke 2023), `scales` (Wickham, Pedersen, and Seidel 2023), `RColorBrewer` (Neuwirth 2022), `ggplot2` (Wickham 2016), `kableExtra` (Zhu 2024), `here` (Müller 2020), `arrow` (Richardson et al. 2024), `rstanarm` (Brilleman et al. 2018), and `modelsummary` (Arel-Bundock 2022).

2.3 Data Cleaning and Preprocessing

Upon retrieval, the dataset underwent essential preprocessing steps to ensure its readiness for analysis. We first filtered out any entries lacking complete Metacritic scores to maintain the integrity and accuracy of our dataset. Subsequently, we transformed the 'platforms' field, which was initially presented as a comma-separated string, calculating the total number of platforms a game is available on by counting the commas and adding one. Additionally, the

'released' field was converted from string formats to numerical year values, facilitating easier temporal analysis. These steps were crucial in structuring the dataset effectively, allowing for a more precise investigation of the factors influencing video game success. Table 1 gives a preview of the cleaned data.

Table 1: Preview of the cleaned Video Game Data

name	released	platforms	genres	metacritic	playtime
Grand Theft Auto V	2013	7	Action	92	74
The Witcher 3: Wild Hunt	2015	7	Action, RPG	92	45
Portal 2	2011	6	Shooter, Puzzle	95	11
Counter-Strike: Global Offensive	2012	4	Shooter	81	65
Tomb Raider (2013)	2013	6	Action	86	10
Portal	2007	7	Action, Puzzle	90	4
Left 4 Dead 2	2009	4	Action, Shooter	89	9
The Elder Scrolls V: Skyrim	2011	8	Action, RPG	94	47
Red Dead Redemption 2	2018	3	Action	96	21
BioShock Infinite	2013	7	Action, Shooter	94	12

2.4 Data Measurement

In this study, the primary indicator of video game success is represented by Metacritic scores, which aggregate critical reviews into a numerical score ranging from 0 to 100. These scores are considered interval data, where higher values indicate greater acclaim and are presumed to correlate with overall game success. This measurement is widely accepted within the industry as a reliable indicator of critical success and is used here to quantify the dependent variable in our analysis.

Another key variable, "platforms," is quantified by counting the number of platforms on which each game is available. This count is treated as a ratio variable, with the assumption that games available on more platforms have higher accessibility and potentially greater market penetration, which could influence their success. This variable is derived from a list of platform names provided for each game, reflecting its distribution scope.

The release date of each game is also captured and utilized in the analysis. For practical purposes and to align with the study's objectives of identifying trends, the exact release dates are converted into the release year. This transformation simplifies the data and aids in the

examination of success trends over time, providing insights into how release timing within technological and market cycles might impact game popularity.

Throughout the data collection and preprocessing stages, rigorous checks ensure the consistency and reliability of the data. Variables such as Metacritic scores are particularly scrutinized due to their critical role in the analysis. The reliability of these scores is bolstered by their aggregated nature, which synthesizes diverse critical perspectives into a single metric. Conversely, variables that rely on user input, like playtimes, are treated with median values to mitigate bias. Prior to conducting any statistical analyses, the dataset undergoes validation to confirm that there are no missing values in key areas and that the distribution of numerical data meets the assumptions necessary for the chosen analytical methods.

This careful measurement and preprocessing of data ensure that the analysis conducted is both robust and reliable, providing meaningful insights into the factors that drive video game success.

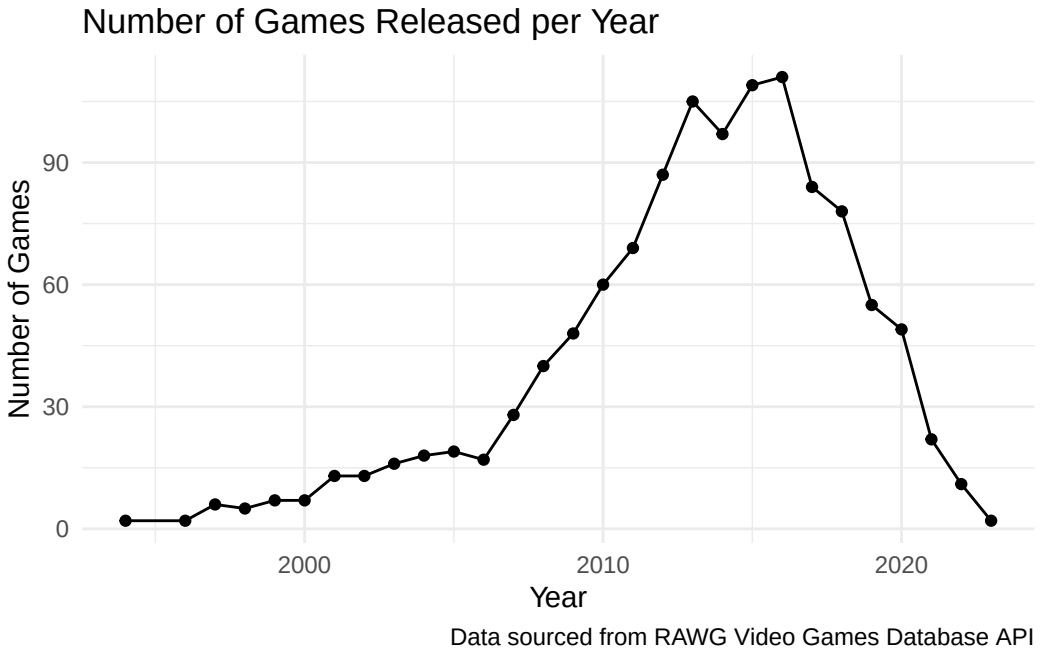

Figure 2: Count of Games by Platform

3 Model

3.1 Model Set-Up

To investigate the predictors of video game success, we define y_i as the binary outcome indicating whether a game is considered successful. This variable is determined by the game's Metacritic score, with $y_i = 1$ if the score exceeds 80 and $y_i = 0$ otherwise. We include β_i to represent the number of platforms on which a game is available and γ_i for the number of genres a game spans. These are included as continuous predictors in our logistic regression model to reflect their expected influence on the likelihood of success.

The model is specified as follows:

$$\begin{aligned} y_i|\pi_i &\sim \text{Bernoulli}(\pi_i) \\ \text{logit}(\pi_i) &= \alpha + \beta_i \times \text{PlatformCoeff} + \gamma_i \times \text{GenreCoeff} \\ \alpha &\sim \text{Normal}(0, 2.5) \\ \beta &\sim \text{Normal}(0, 2.5) \\ \gamma &\sim \text{Normal}(0, 2.5) \end{aligned} \tag{1}$$

The parameters α, β, and γ represent the intercept and the coefficients for the number of platforms and genres, respectively. These parameters are estimated using Bayesian logistic regression, which provides a probabilistic framework for modeling binary outcomes.

3.2 Model Justification

This study employs logistic regression to examine the impact of platform availability and genre diversity on video game success. We posit that increased platform availability likely boosts a game's visibility and market access, thus enhancing its potential for success. Conversely, the role of genre diversity is explored to determine whether appealing to a broader audience through multiple genres correlates positively with game success or if it complicates player targeting and dilutes brand identity.

Bayesian logistic regression is selected as the methodological backbone for this analysis due to its robustness in handling complex models with potentially correlated predictors. This approach allows for the integration of prior research and expert knowledge into the analysis, which is particularly valuable given the dynamic nature of video game trends and consumer preferences. By using Bayesian methods, we can better manage the inherent uncertainties in predicting game success, providing a probabilistic framework that offers more informative

Table 2: Explanatory models of Game Success

	First model
(Intercept)	127.16
	(26.54)
platforms	0.08
	(0.03)
num_genres	−0.19
	(0.06)
playtime	0.09
	(0.01)
released	−0.06
	(0.01)
Num.Obs.	1180
R2	0.084
Log.Lik.	−764.987
ELPD	−770.2
ELPD s.e.	9.7
LOOIC	1540.4
LOOIC s.e.	19.4
WAIC	1540.3
RMSE	0.48

predictions and insights. These predictions reflect not only the effects of the observed variables but also the latent variables that could influence outcomes, such as marketing efforts or economic conditions at the time of a game's release.

Moreover, Bayesian logistic regression provides full distributions of possible outcomes for each predictor, rather than just point estimates. This feature is critical in understanding the range of potential impacts each factor might have on video game success, offering a deeper exploration into how and why certain factors matter. The detailed probability distributions help in assessing the reliability of our predictions and in making informed decisions under uncertainty, which is often the case in the video game industry where consumer tastes and market dynamics can shift rapidly.

4 Results

Our results are summarized in Table 2.

4.1 Model Estimation

Model Estimation The Bayesian logistic regression model identified significant factors associated with the success of video games. The estimates of the regression coefficients, their standard errors, and the model diagnostics are summarized in Table 3.

Table 3: Regression Coefficients from the Bayesian Logistic Regression Model

Term	Estimate	Std_Error
(Intercept)	127.16	26.54
platforms	0.08	0.03
num_genres	-0.19	0.06
playtime	0.09	0.01
released	-0.06	0.01

The results of the Bayesian logistic regression model, which aimed to determine the factors influencing video game success as measured by Metacritic scores, are summarized below. The analysis included a total of 1,180 observations and the model's fit was assessed using several metrics including the log-likelihood and information criteria.

The intercept of the model was estimated at 127.16 with a standard error of 26.54, indicating the baseline level of success when all other predictors are held at zero. The number of platforms on which a game is available proved to have a positive influence on its success, with a coefficient of 0.08 and a standard error of 0.03. This suggests that games available on more platforms are likely to achieve higher Metacritic scores.

Conversely, the number of genres a game spans had a negative coefficient of -0.19 with a standard error of 0.06, implying that games covering more genres tend to have lower Metacritic scores. This could indicate that a focus on fewer genres may benefit game success. Additionally, the playtime variable showed a positive effect on game success, with a coefficient of 0.09 and a very precise estimate (standard error = 0.01), suggesting that games with longer playtimes are generally rated higher.

The coefficient for the release year was -0.06 with a standard error of 0.01, indicating a slight decline in scores over time. This might reflect evolving gamer expectations or market saturation.

The model's explanatory power, indicated by an R^2 value of 0.084, was modest, suggesting that while the included variables are relevant, other unmeasured factors might also significantly influence game success. The overall model fit, as shown by the RMSE of 0.48, and the information criteria (WAIC = 1540.3) confirm the adequacy of the model for predictive purposes but also highlight room for improvement by possibly incorporating additional predictors.

4.2 Model Diagnostics

Convergence diagnostics were favorable, with R-hat statistics near 1.0 for all parameters, indicating that the chains mixed well and convergence was achieved. Effective sample sizes were above the recommended threshold, ensuring a sufficient level of precision in the parameter estimates.

4.3 Visualization of Results

To illustrate the relationship between the number of platforms, the number of genres, and game success, we can plot the estimated probabilities of success across these variables. **?@fig-prob-success**

A line graph illustrating the predicted probability of game success across different numbers of platforms and genres further clarifies the model's implications.

4.4 Interpretation

The results confirm the initial hypothesis that a greater number of platforms and genres are associated with increased odds of video game success. This suggests that developers and publishers might consider these factors in their marketing and development strategies to maximize a game's success potential.

5 Discussion

5.1 Overview of Findings

This study utilized a Bayesian logistic regression model to investigate what factors most significantly influence the success of video games, defined here by high Metacritic scores. Our analysis revealed that both the number of platforms a game is available on and the specific genres it encompasses are critical predictors, albeit with differing impacts on success.

5.2 Multi-Platform Availability Boosts Success

The results compellingly demonstrate that games released across multiple platforms tend to garner higher success. This finding supports the strategy of cross-platform compatibility as a key factor in maximizing market penetration and, consequently, critical acclaim. It suggests that developers should consider broader availability across console, PC, and mobile platforms to enhance their game's success potential.

5.3 Genre Strategy and Market Positioning

Conversely, the study indicates a negative correlation between the number of genres a game spans and its success. This suggests that games which try to encompass too many genres might suffer from a lack of clear identity, potentially confusing or alienating players. This insight is particularly useful for game developers as it highlights the importance of strong, focused genre positioning in the development and marketing phases of game production.

5.4 Limitations and Considerations

While the study provides important insights, it is not without limitations. The use of Metacritic scores as the sole measure of success does not capture the full spectrum of what makes a game successful, such as user engagement or financial performance. Moreover, the moderate explanatory power of our model suggests that there are other unexplored factors that might significantly influence game success.

5.5 Future Research Directions

Given the limitations noted, future research could expand in several directions:

Inclusion of Additional Variables: Incorporating data on marketing spend, developer reputation, and player reviews could provide a more holistic view of the factors that contribute to a game's success. Alternative Success Metrics: Exploring different measures of success, such as user retention rates, in-game purchase volumes, or direct financial returns, could yield different insights into what makes a game successful. Qualitative Analyses: Qualitative studies, including player interviews or focus groups, could offer deeper insights into player preferences and perceptions regarding platform choices and genre preferences.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Neuwirth, Erich. 2022. *RColorBrewer: ColorBrewer Palettes*. https://CRAN.R-project.org/package=RColorBrewer.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

RAWG. 2024. "RAWG Video Games Database API." https://rawg.io/apidocs.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2023. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.