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

Sakhil Goel

April 19, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

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

1 Introduction

You can and should cross-reference sections and sub-sections. We use R Core Team (2023) and Wickham et al. (2019).

The remainder of this paper is structured as follows. Section 2….

2 Data

2.1 Data Source

The primary dataset for this study was obtained from the RAWG Video Games Database API, 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.

2.2 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.

2.3 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.

2.4 Data Cleaning and Preprocessing

Upon retrieval, the data underwent several preprocessing steps to ensure its suitability for analysis:

Cleaning: Data were cleaned to remove any entries with missing or incomplete information, particularly in the fields of Metacritic scores and release dates. Transformation: Release dates were converted from string formats to date formats. Additionally, the 'genres' and 'platforms' fields, initially received as comma-separated strings, were transformed. For 'genres', each game was encoded into binary variables representing the presence or absence of each genre. For 'platforms', a count variable was created to indicate the number of platforms a game is available on. New Variables: Two new variables were calculated: num_genres: The total number of genres associated with each game. num_platforms: The total number of platforms on which each game is available.

Table 1: Preview of Cleaned 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

Bills of penguins

2.5 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.

The variable "genres" is initially presented as a list of comma-separated values for each game, reflecting the multiple genres a game might belong to. To facilitate quantitative analysis, these genre labels are transformed into binary variables through a process known as one-hot encoding. Each genre is represented as a separate column in the dataset, with a value of 1 if the game is associated with that genre and 0 otherwise. This approach allows for the exploration of the impact of various genres on a game's success, accommodating the complex nature of genre classification in the gaming industry.

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.

Talk way more about it.

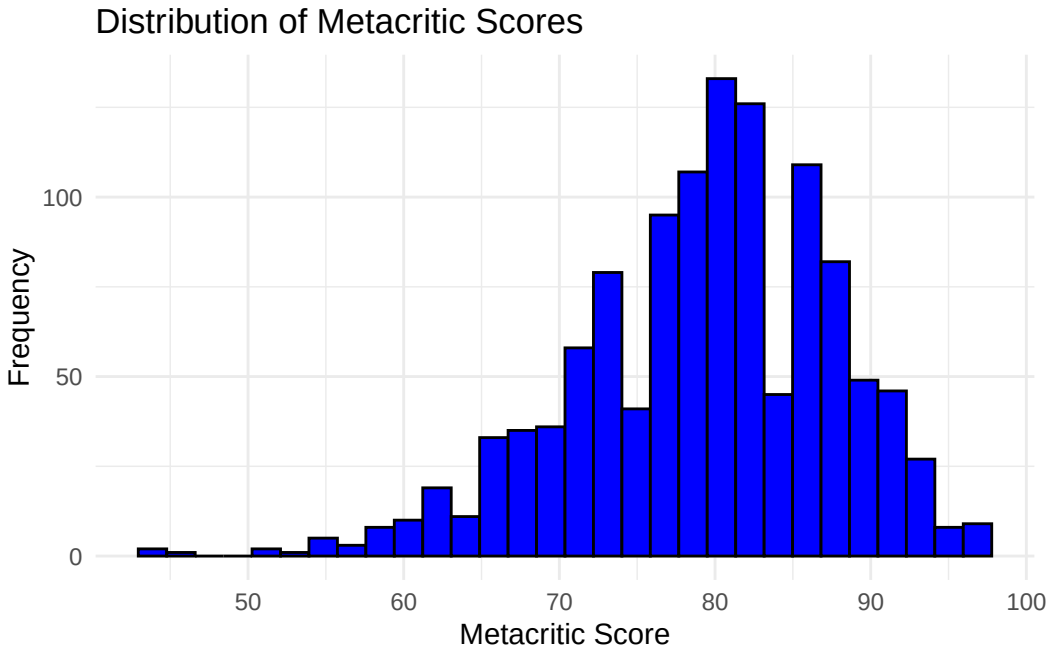

Figure 1: Distribution of Metacritic Scores

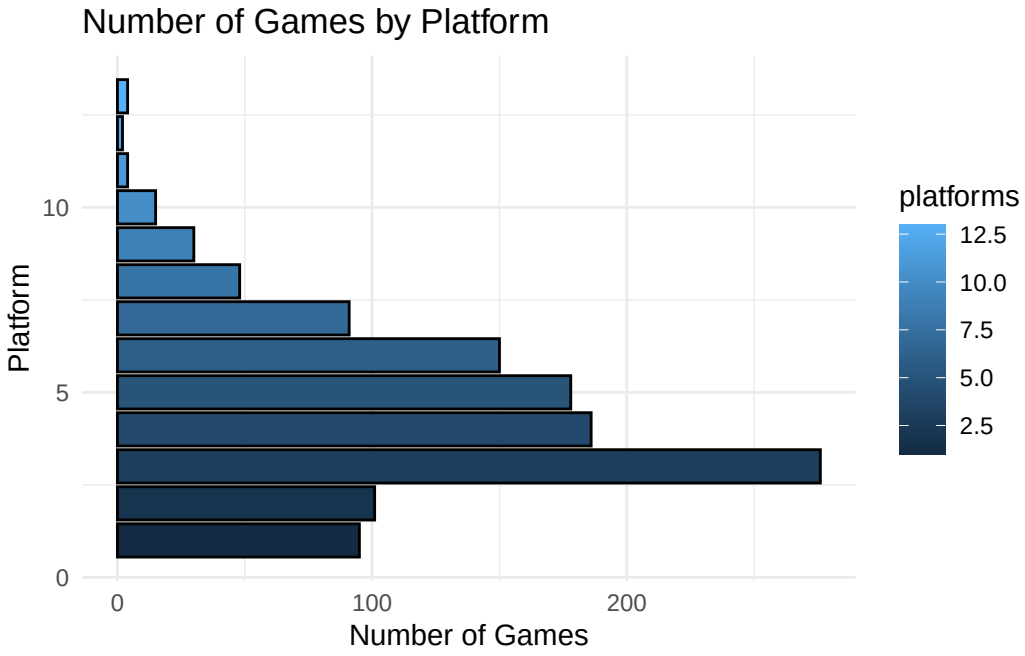

Figure 2: Count of Games by Platform

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table 2.

Table 2: Explanatory models of flight time based on wing width and wing length

	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

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected by, the data

Figure 3: **?(caption)**

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algorithm

Figure 4: **?(caption)**

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

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