Analysing the Factors That Influence Player Reviews on Steam

Shenjun Lu

31st March 2025

1 Overview

Player reviews are crucial because they influence potential consumers' purchasing decisions and serve as an indicator of game quality on platforms such as Steam. This report analysed the various factors influencing player review scores on Steam, with a particular focus on comparing ratings from the general public to those from professional critics (Metacritic), as well as analysing how user ratings vary by game type.

To gauge an overall picture of the different opinions between players and critics, I visualised the distribution of player reviews against Metacritic scores. Then, I evaluated the significance of the difference in mean scores using a paired t-test. I observed that players rate games more favourably overall, whereas the professional critics tended toward safer, mid-to-high scores.

Further, I investigated the factors impacting the number of player reviews a game garners. A Poisson regression was conducted to analyse how various aspects of the game relate to the total number of player reviews. My findings suggest that the number of concurrent users has the most substantial effect on the total reviews a game receives, while the price alone appears to exert a relatively minor influences. This implies that quality games are more likely to attract and retain active players, leading to an increase in player reviews.

Additionally, I explored how review scores varied across the four primary game types: Puzzle, Shooter, Strategy, and RPG.[1] The analysis reveals that puzzle games tend to receive more positive reviews compared to the others.

2 Introduction

Context and motivation Steam is the largest digital distribution platform for PC games[2], holding a substantial influence over the gaming industry. It offers a vast selection of thousands of games across multiple languages, platforms, and genres, which not only enhances its appeal but also fosters a rich diversity of player opinions and experiences.

In the realm of consumer behaviour, social influence plays a crucial role in shaping consumer choices. Many consumers rely heavily on high-quality reviews from trusted sources, as these evaluations can significantly impact their decisions. [3] As a result, game reviews have become an essential element in determining the success or failure of a game in such a competitive market.

Given the vast database available on Steam, analyses can be conducted to assess the impact of social influence, including the volume of review. This report will delve into the game data to explore these effects.

Previous work Previous research on movie reviews [4] highlighted the polarisation between critical and audience evaluations. The authors examined four categories of review scores: Rotten Tomatoes critics, Rotten Tomatoes audience (fans), IMDb, and Metacritic. They showed that although all score categories were interrelated, the correlations were significantly stronger amongst critic scores compared to the correlations between critic and audience scores.

In a blog post [5], Aila reflected on her personal reading preferences, noting that she tends to be more critical when evaluating contemporary novels as opposed to fantasy books. This observation suggests that varying types of games may similarly influence game reviews based on the reviewer's perspective.

Objectives The objective of this report is to analyse the variations in types of reviews and to determine the factors influencing player reviews. I will investigate the various game attributes that influence the volume of player reviews, explore the distinctions between player reviews and Metacritic scores, and assess how the genre of a game impacts player reviews.

3 Data

Data provenance The dataset was acquired from the GitHub repository developed by NewbieIndieGameDev [6]. The information in this repository was extracted from Steam's publicly available data using the Steam Web API and is provided under its terms of use [7]. Additionally, the dataset includes data from SteamSpy, which can be accessed through their API. SteamSpy collects publicly accessible information, which is permissible[8].

Data description Data for games was split into eight .csv files, where each dataset contain a unique game identifier, 'app_id'. For this analysis, I utilised four files: game details, tags, reviews, and SteamSpy insights.

The Steam data includes records for 140,082 distinct games, and SteamSpy provides insights for 140,077 games. For the analysis, only a subset of the available variables has been employed. In the game data, 'type' (indicating whether a game is a demo or full version) and 'app_id' are used.

From the reviews dataset, 'positive_reviews' reflects the number of positive reviews, whereas 'total_reviews' represents the overall number of reviews, and 'metacritic_score' provides the critic score for a game on Steam. Other information from this dataset has not been used.

The SteamSpy data includes the number of concurrent users recorded the day before the data collection as 'concurrent_users_yesterday'. I've also used 'price' (final price in USD), 'initial_price' (original price) and 'discount' (discount percentage).

The tags dataset associates multiple tags with each 'app_id', used for the identification of game types in later analysis.

Data processing To ensure successful data loading, I skipped bad lines when reading the CSV files and filtered out demo games, which were irrelevant to my analysis. Several variables were renamed for clarity, and I excluded games with fewer than 25 reviews to avoid outliers. I converted the price values from cents to dollars by dividing, and removed games with inconsistent pricing and unavailable values. All relevant data, including game details, reviews, and SteamSpy information, was merged into a single dataframe based on 'app_id'.

Additionally, I added 'player_review_score' variable, which represents the percentage of positive reviews for each game. This change was implemented to create a more continuous data structure compared to the discrete categories found in the 'review_score', enhancing its suitability for numerical analysis.

4 Exploration and analysis

4.1 What Drives Review Volume? A Poisson Regression Approach

From the cleaned data, only a small number of numerical factors can be analysed. A random sample of 1,000 values was drawn for each variable, and a pair-plot was generated to ascertain correlation. Logarithmic transformations were applied to both the number of concurrent users and the total number of reviews.

To unravel the correlations, I calculated the correlation matrix for relating the numerical values across the dataset. This was then visually represented with a heatmap, assisting in the selection of variables for the Poisson regression.

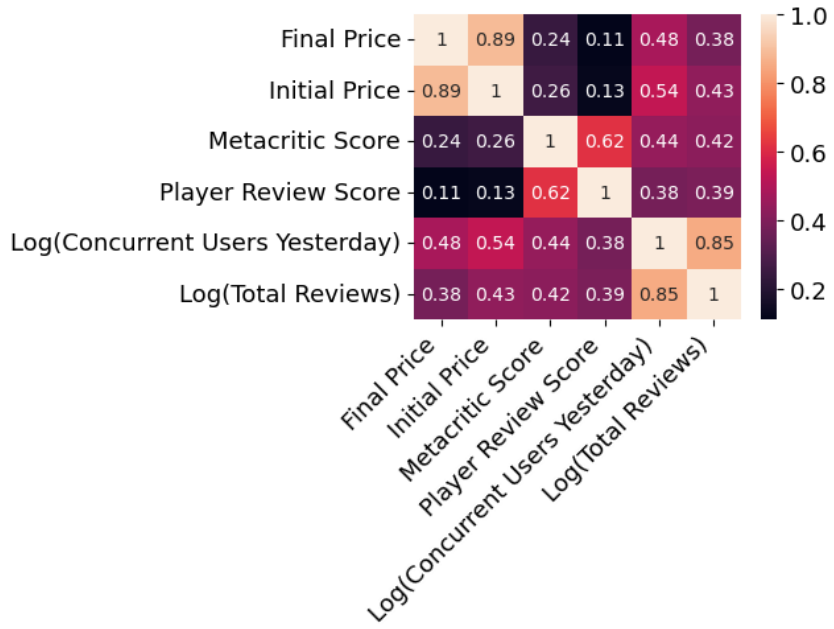

Figure 1: Heatmap of correlation matrix to show relationships between key numerical game features.

A strong correlation is anticipated between the initial and final prices of the games, and incorporating both variables in a regression model may lead to multicollinearity.

The player review score demonstrates a moderate to high correlation with the Metacritic score, this relationship is explored in greater detail later in the report.

Notice that the logarithm of total reviews displays a strong positive correlation with the logarithm of concurrent users from the previous day, suggesting that player retention is a key determinant iin the volume of player reviews received. Additionally, the price and the rating of the game exhibit a moderate correlation with the number of game reviews.

To identify and quantify the influence of correlated variables on the number of total reviews, a Poisson regression [9] was chosen.

For each predictor, a hypothesis test was performed. The Null Hypothesis (H_0) is the predictor does not have an influence on the logarithm of total reviews, this means the coefficient is zero. The Alternative Hypothesis (H_1) is that the predictor does affect the outcome, so the coefficient is not zero.

The regression results indicate that only the player review score and the logarithm of concurrent users have a statistically sufficient effect with the logarithm of the total number of reviews at 5% significance level.

Variable	**Coef.**	**Std. Err**	**z**	**P>|z|**	**[0.025, 0.975]**
Final Price	-0.0011	0.001	-1.043	0.297	[-0.003, 0.001]
Initial Price	0.0004	0.001	0.411	0.681	[-0.002, 0.002]
Metacritic Score	0.0005	0.001	0.634	0.526	[-0.001, 0.002]
Player review score	0.0019	0.059	3.218	0.001	[0.074, 0.303]
Log(Concurrent Users Yesterday)	0.0840	0.003	27.336	0.000	[0.078, 0.090]
intercept	1.6021	0.051	31.623	0.000	[1.503, 1.701]

Table 1: Poisson Regression Model Results

The output of the regression reveals a high p-value for game prices, suggesting a negligible impact on the number of reviews generated. This implies price alone is insufficient to predict the number of reviews, and can infer players are willing to invest in a game irrespective of its cost, given high-quality content.

Conversely, player review scores possess a positive coefficient in the regression model, with a statistically sufficient p-value below 0.05. The coefficient proposes as the player review score increases by 1%, the value for 1+ total reviews is expected to be multiplied by $e^{0.0019} \approx 1.0019$. Thus, a favourable player experience is anticipated to exert a small positive influence on review volume.

Among the considered variables, the most substantial factor related to the total review count is the number of concurrent users yesterday. One unit increase in $\log(1+\text{Concurrent Users Yesterday})$ is associated with an increase of 0.0840 in $\log(1+\text{Total reviews})$. This emphasises the critical role of user retention in game reviews, as long engagement with a game enhances the likelihood of players leaving a review.

4.2 Comparison of Player Review Scores and Metacritic Score

Next, I compared the characteristics between the player review score and the Metacritic score. The Kernel Density Estimation (KDE) plot was chosen because it produces a less cluttered and more interpretable plot [10], and a boxplot is used for comparative clarity.

The distribution analysis reveals that the mean player review score is notably higher than the critics. The player reviews (in blue) exhibit a rightward skew, suggesting that players often give more generous ratings, but with greater variability shown by both the range and the interquatile range. On the other hand, the Metacritic scores (in pink) present a distribution with a lower mean, indicating that critics adopt a more conservative approach in their assessments and are less willing to express extreme opinions.

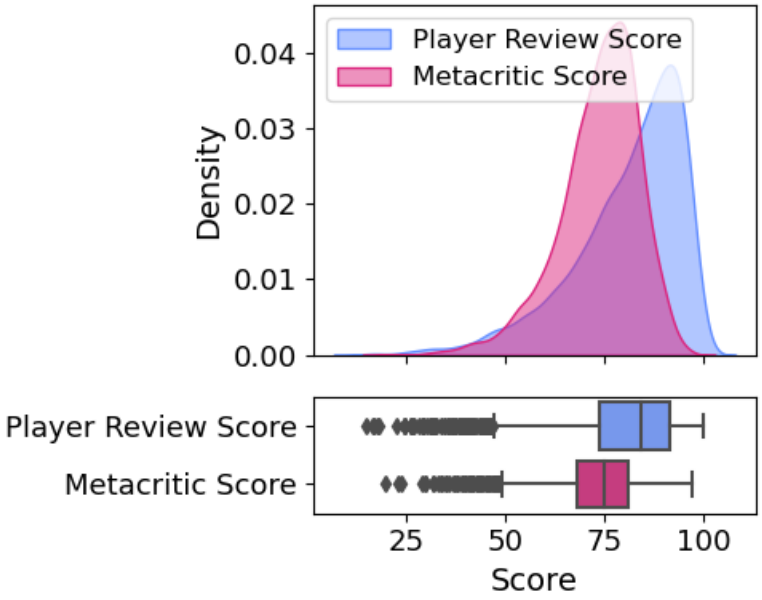

Figure 2: Distribution of Player Review and Metacritic Score shown in KDE plot and boxplot with shared x-axis.

However, despite the players' tendency to produce more positive ratings, they also tend to issue more extreme negative ratings. This phenomenon may be due to the influence of negative critiques: players who read the negative feedback can rate the game significantly lower than players who encounter positive feedback. [11] Hence, as a game receives more negative reviews, it can experience a snowball effect, contributing to heightened variability in player ratings overall.

To verify the statistical significance of the mean difference between player and critic reviews, a paired t-test was used. This is a suitable test because individual values from both samples can be paired by game [12].The difference is represented as $D_i = X_i - Y_i$ where X_i, Y_i denote the random variables for player review score and Metacritic score for game i respectively. Given the aim was to test if the mean player review score is higher than the mean Metacritic score, a one-tailed test is used to test for the mean difference between the two, denoted by μ_D.

The Null Hypothesis is the two reviews have the same mean ($\mu_D = 0$). The Alternative Hypothesis is the mean player review score is greater than the mean Metacritic score ($\mu_D > 0$). The test statistic is given by

$$T = \frac{\bar{D} - \Delta_0}{\sqrt{S_D^2/n}} \sim t_{n-1}$$

where n denotes the number of paired observations, $\bar{D}$ and S_D^2 are the mean and variance estimator on the difference, respectively.

The results from the t-test yielded a t-statistic of 39.487777127358655 and a p-value of 8.361549204304257e-286. Since the p-value is significantly less than 0.05, we can reject the null hypothesis at 5% significance

level. This result strongly indicates that Steam users rate games significantly higher on average than professional critics (Metacritic).

4.3 The Influence of Game Type on Player Reviews

To evaluate the potential influence of game type on player reviews, I decided to organise games into four main types of games, namely, Puzzle, Shooter, Strategy and RPG. [1] A boxplot was created for each type to assess the distribution. The data illustrate a consistent pattern in Metacritic scores, revealing similar mean values across all classified types. This observation highlights the relatively narrow interquartile range associated with Metacritic scores from earlier distribution, indicating a uniform mean across games in general.

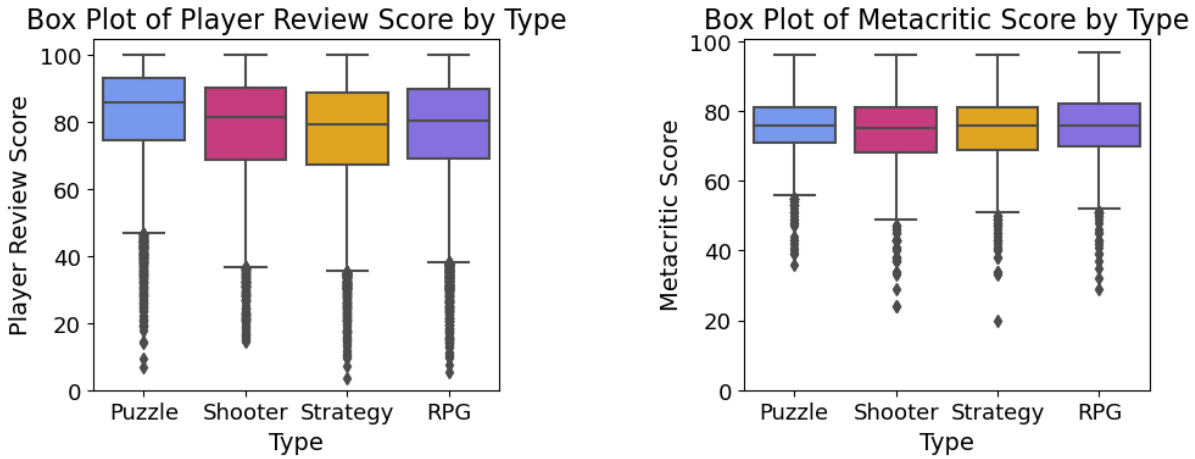

Figure 3: Box Plot showing the distribution of Player Review Scores and Metacritic Scores across four game types (Puzzle, Shooter, Strategy, and RPG).

It can be observed that the mean for player review scores of puzzle games is slightly higher than the rest. To formally test this hypothesis, I decided to employ a bootstrap sample to analyse and normalise the data. [13] A sample of 10,000 datapoints are created by bootstrap $d_1, \cdots, d_{10,000}$, where

$$d_i = x_i - y_i$$

where x_i denotes the bootstrapped mean derived from a random sample of 3,000 games tagged as Puzzle, while y_i is the bootstrapped mean of a random sample of 3,000 from the games that are not Puzzle.

A hypothesis test is performed, where the Null Hypothesis is that the mean review scores for Puzzle games and non-Puzzle games are equivalent, while the Alternative Hypothesis propose that the mean of the mean review score for Puzzle games is higher.

Since the size of the sample for d_i is large, by the Central Limit Theorem, the sample is approximately normally distributed, thus a Z distribution can be used. So to test the hypothesis, a confidence interval is constructed for the mean of the difference μ_d. Since this is a one-tailed test, the confidence interval is given by

$$CI = [\bar{d} + z_{0.95}\frac{s_d}{\sqrt{n}}, \infty)$$

where $z_{0.95}$ is the 95th percentile of the standard normal distribution, s_d is the sample standard deviation of the difference, n is the number sample size of d.

The resulting confidence interval I computed was $CI = [0.3479, \infty)$. Since $0 \notin CI$, the null hypothesis is rejected at 5% significance level. Thus, there is sufficient evidence to suggest Puzzle games have higher mean review scores than the other three types of game analysed.

5 Discussion and conclusions

Summary of findings The number of concurrent users yesterday is a strong feature that correlates to the total number of reviews a game receives on Steam. This suggests that developing games with strong player retention strategies is an effective way to enhance a game's visibility and popularity, thereby encouraging players to recommend the game to others, such as by inviting friends. Another factor that is positively related to total review is the player review score. This was expected since players are willing to put time and effort into sharing their thoughts on well-crafted games.

The analysis of player reviews and Metacritic scores has shown a distinct review habit. The higher average in player reviews shows that most players respond positively to games and are willing to give high scores. Whereas critics tend to give a more uniform mid-high score. One possible explanation for this difference is that players typically choose to purchase and play games they believe they will enjoy, leading to generally favourable reviews. However, when a game fails to meet their expectations, it often results in stronger negative reviews, which accounts for the wider distribution of negative scores among players compared to those on Metacritic. Conversely, critics are likely to review games that are of higher quality, and with a more natural opinion than players, resulting in more consistent opinions and thus lower variance in their scores.

Additionally, the type of game has an affects on player review. In particular, I tested and concluded puzzle games are more likely to receive higher review scores on average. This suggests different types of games lead to varying levels of feedback from the player, likely due to the difference in player base; those who enjoy certain types of games may be less likely to appreciate others.

Evaluation of own work: strengths and limitations My report used visual representations and statistical tests to support the interpretations drawn from them. It delves into various aspects of user reviews. However, the most significant limitation was the lack of valuable time series data that could continuously monitor game performance, such as average playtime and number of user reviews over time. Therefore, the report can only address a narrow scope of what could be examined. Additionally, I found it challenging to identify appropriate areas of study due to the overwhelming amount of available data. I experimented with various techniques and analytical approaches before honing in on a more focused analysis, this was very time-consuming and delayed my progress.

Comparison with any other related work The result closely matches with Venkateshan who also compared User scores and Metacritic scores in his Linkedin article,[14] which both suggested user reviews have a higher mean and are more volatile than the Metacritic scores.

Studies have also been done in the paper The Edge of Glory: The Relationship between Metacritic Scores and Player Experience [15], and shown notable differences in how professional critics and regular users assign ratings.

The paper evaluating player experience by genre[16] also aligns with my finding of the genre. It suggested the existence of transferable in-game features that can accurately predict games of the same genre.

Improvements and extensions When evaluating the factors influencing user review, it is important to note that categorical data have not been utilised. A potential improvement to the evaluation is to incorporate categorical data, and analyse the significance using ANOVA. This could present interesting insights that may not have been previously considered.

An extension of this analysis could involve examining the release date of the game, and evaluating the effect of player and Metacritic review in the early stage of the game and later stage of a game's lifecycle. This can be conducted similarly to the investigation of factors that affect the amount of game reviews. Having a large number of user reviews in the early stage of the game can provide an advantage, and boost in popularity.

References

[1] Xiaozhou Li and Boyang Zhang. 'A Preliminary Network Analysis on Steam Game Tags: Another Way of Understanding Game Genres'. In: *Proceedings of the 23rd International Conference on Academic Mindtrek*. AcademicMindtrek '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 65–73. ISBN: 9781450377744. DOI: 10.1145/3377290.3377300. URL: https://doi.org/10.1145/3377290.3377300.

[2] Wikipedia contributors. *Steam (service) – Wikipedia, The Free Encyclopedia*. Last accessed 29 March 2025. 2025. URL: https://en.wikipedia.org/wiki/Steam_(service).

[3] Young Ae Kim and Jaideep Srivastava. 'Impact of Social Influence in E-Commerce Decision Making'. In: *Proceedings of the Ninth International Conference on Electronic Commerce*. ICEC '07. New York, NY, USA: Association for Computing Machinery, 2007, pp. 293–302. ISBN: 9781595937001. DOI: 10.1145/1282100.1282157. URL: https://doi.org/10.1145/1282100.1282157.

[4] Kyle Day and Jong-Min Kim. 'Investigating Polarisation in Critic and Audience Review Scores via Analysis of Extremes, Medians, Averages, and Correlations'. In: *International Journal of Environment, Workplace and Employment* 7.1 (2023), pp. 3–12. DOI: 10.1504/IJEWE.2023.132409. URL: https://www.inderscienceonline.com/doi/abs/10.1504/IJEWE.2023.132409.

[5] Aila J. *Chatterbox: Do We Have Different Ratings for Different Genres?* Accessed: 2025-03-31. 2016. URL: http://www.happyindulgencebooks.com/2016/08/26/chatterbox-different-ratings-different-genres/.

[6] NewbieIndieGameDev. *steam-insights*. https://github.com/NewbieIndieGameDev/steam-insights. Commit: 4530e75befa1a01da33d77ffaab3d4e76fdaa766. Oct. 2024.

[7] Valve Corporation. *Steam Web API Terms of Use*. Accessed: 2025-03-29. 2010. URL: https://steamcommunity.com/dev/apiterms.

[8] Sergey Galyonkin. *Steam Spy*. Accessed: 2025-03-29. 2015. URL: https://steamspy.com/about.

[9] David Sterratt. *Re: CW2 help on steam data*. https://piazza.com/class/m06jk54upts3kh/post/162. Piazza class post, accessed on 2025-03-24. 2025.

[10] Michael Waskom and the seaborn development team. *seaborn.kdeplot — seaborn 0.11.2 documentation*. Accessed: 2025-03-29. 2021. URL: https://seaborn.pydata.org/generated/seaborn.kdeplot.html.

[11] Ian J. Livingston, Lennart E. Nacke and Regan L. Mandryk. 'The Impact of Negative Game Reviews and User Comments on Player Experience'. In: *Proceedings of the 2011 ACM SIGGRAPH Symposium on Video Games*. Sandbox '11. New York, NY, USA: Association for Computing Machinery, 2011, pp. 25–29. ISBN: 9781450307758. DOI: 10.1145/2018556.2018561. URL: https://doi.org/10.1145/2018556.2018561.

[12] School of Mathematics, University of Edinburgh. *MATH08051: Statistics (Year 2) Course Notes*. https://www.learn.ed.ac.uk/courses/1/MATH080512024-5SV1SEM2/content/_10952227_1/index.html. Version 2024.0, University of Edinburgh. 2024.

[13] Anonymous Student. *Edinburgh Cycle Hire in the Pandemic*. Internal course resource. Unpublished student report, University of Edinburgh, fds-project-option-3-pair-1. 2021.

[14] Karthik Venkateshan. *Are Steam User Reviews a Better Predictor of Video Game Sales than Metacritic Scores?* Aug. 2019. URL: https://www.linkedin.com/pulse/steam-user-reviews-better-predictor-video-game-sales-than-karthik/ (visited on 22/03/2025).

[15] Daniel Johnson et al. 'The Edge of Glory: The Relationship Between Metacritic Scores and Player Experience'. In: *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 141–150. ISBN: 9781450330145. DOI: 10.1145/2658537.2658694. URL: https://doi.org/10.1145/2658537.2658694.

[16] David Melhart, Antonios Liapis and Georgios N. Yannakakis. 'Towards General Models of Player Experience: A Study Within Genres'. In: *2021 IEEE Conference on Games (CoG)*. 2021, pp. 01–08. DOI: 10.1109/CoG52621.2021.9618902. URL: https://doi.org/10.1109/CoG52621.2021.9618902.