# What issues make or break your game?

**William Liu**
Stanford University
bob9001@stanford.edu

## Abstract

Every game studio has asked itself this question at one point or another: Given the remaining development time, which issues in the game are the most critical to fix before release? This paper analyzes over 100,000 Steam reviews, classifying them across 10 common issue categories: Bugs, Onboarding, Controls, Balance, Gameplay Loop, Content, Polish, Story, Support, and Paywalls. Three success metrics are analyzed in context of these issues: Steam review score, Metacritic score, and estimated revenue. Findings show notable differences in issue frequency in indie games, early access games, and across genres. While most issues have negligible impact on the three success metrics, notable exceptions include Bugs and Lack of Polish, which have slight negative impacts on Steam review score and Metacritic score. By quantifying which issues players care most about and which can be tolerated, this work aims to inform smarter prioritization of development resources to create the best outcome for both players and developers.

## 1 Introduction

In the past decade, it has become increasingly common for video games to launch in a buggy, unfinished, or even unplayable state. *Fallout 76* (2018), *Cyberpunk 2077* (2020), and *Battlefield 2042* (2021) were some of the most notable disasters in AAA games in recent history, where each title suffered from game-breaking bugs and a myriad of other issues upon launch. One could reasonably assume that such failures cause long-lasting reputational damage and discourage the playerbase from purchasing future titles, which should be sufficient incentive for studios to do everything in their power to avoid such outcomes.

While the aforementioned failures are not in the interest of any party involved, we begin to see dissonance between what is considered a great game by players and what is a profitable game for the shareholders when we examine some of the largest video game franchises. *Call of Duty* (2003-2024), *EA SPORTS FIFA* / *EA SPORTS FC* (1993-2024), and *EA SPORTS Madden NFL* / *EA SPORTS College Football* (1993-2024) are all franchises with annual releases, attracting millions of players despite high price tags in the range of $60-$80. Miraculously, demand for these games has held steady despite having some of the lowest review scores among AAA games, which can be attributed largely to their predatory monetization schemes and lack of innovation from year to year.

While these examples represent opposite extremes, it is important for all game studios to understand what types of issues are minor nuisances for players versus what types of issues can singlehandedly ruin the prospects of a game. With such an understanding, limited development time and resources can be more efficiently allocated to create a better experience for players and increased revenue for shareholders.

We can see some hints of why different types of issues can result in drastically different experiences by examining game design theory. The MDA framework (Hunicke et al., 2004) is a popular framework for understanding video games on three levels: mechanics, dynamics, and aesthetics. As part of the framework, the authors outline eight types of fun that games can aesthetically create for players: sensation, fantasy, narrative, challenge, fellowship, discovery, expression, and submission. Under this perspective, we can interpret the impact of different types of issues as obstructing the realization of each type of fun. For example, a poorly-designed input control system may

severely degrade the fun from the challenge in a game, while it might not affect the player's ability to enjoy the narrative or be immersed in the fantasy.

We can make similar hypotheses using a popular player categorization framework, Bartle's taxonomy of player types (Bartle, 1996), which proposes four categories of players along two axes: one axis runs from players to world and the other runs from acting to interacting. Killers are players who enjoy acting on other players; achievers are players who enjoy acting on the world; socializers are players who enjoy interacting with other players; and explorers are players who enjoy interacting with the world. With this framework, we can contextualize different types of issues as degrading the experience of different types of players. For example, an MMO would be completely unappealing to killers or socializers if features like in-game chat frequently went offline, while explorers might be able to enjoy the well-crafted game world without any regard for the issues.

As we live in a world in which fixing every bug and design flaw before release is practically impossible, decisions are being made within the production cycle of every studio on what issues to prioritize fixing, what issues are acceptable to ship, and what issues are severe enough to justify delaying the release of the game. At the extreme, some are also making decisions regarding intentional degradations of the player experience, if that would mean increased profitability.

This paper aims to quantify the effects of different types of issues on the success of a game by categorizing a large quantity of Steam reviews along 10 common issue types. These issues are then correlated with three measures of success: user review score, critic score, and estimated revenue. By providing decision makers with information on the impact of different issues, we hope to help narrow the gap between what players want and what developers think they want, benefiting consumer satisfaction and the game industry as a whole.

## 2 Related Work

Steam is one of the most popular PC video game platforms today. While competing PC game platforms like Epic Games Store and GOG exist—along with console game platforms like PlayStation Store, Xbox/Microsoft Store, and Nintendo eShop—an analysis of games on Steam are likely representative of the game industry overall due to the prevalence of cross-listing games on multiple stores. Below, we examine several past studies analyzing user reviews of games on Steam.

### 2.1 Review Trends

One such study (Lin et al., 2019) examines numerous aspects of Steam reviews, including how long players play before posting a review; differences in content between positive and negative reviews; differences in content between reviews of free versus paid, indie versus non-indie, and early access versus non-early access games; and impact of Steam sales and developer news updates on reviews. Reviews were divided into six categories: *not helpful*, *pro*, *con*, *video*, *suggestion*, and *bug*, using a classifier trained on a small subset of manually labeled reviews. Key findings include: 1) while negative reviews tend to be longer and contain more valuable feedback, positive reviews can also contain suggestions and bug mentions so they should not be ignored; 2) negative reviews tend to be posted after a significantly shorter period of play than positive reviews, especially when the game contains severe bugs or is unengaging; and 3) many more reviews contain issues relating to game design rather than bugs, suggesting either a lower prevalence or lower impact of bugs than other types of issues.

Another study (Guzsvinecz and Szűcs, 2023) found similar patterns in review length and playtime differences between positive and negative reviews. The authors also used sentiment analysis to observe the emotional valence over the course of reviews, finding that reviews tend to have a structure that begins with more positive emotions and ends with more negative emotions. In addition, the analysis was divided by Steam's 11 top-level genres: Strategy, Puzzle, Adventure, Simulation, Action, Casual, RPG, Sports, Racing, Tabletop, and Experimental. They find that Action game reviews contained the most anger sentiment; Tabletop game reviews contained the most anticipation sentiment; Racing and Sports game reviews contained the most disgust sentiment; Action game reviews contained the most fear sentiment; Experimental game reviews contained the most joy, sadness, and surprise sentiments; and Tabletop and Simulation game reviews contained the most trust sentiment. The sentiment of Steam reviews has also been found to be predictive of user churn

(Abdul-Rahman et al., 2024).

### 2.2 Success Metrics

This paper aims to contextualize reviews relative to three success metrics of a game: Steam user review score, Metacritic score, and estimated revenue.

Steam reviews contain a written description along with a positive or negative vote (hereafter referred to as positive or negative reviews, not to be confused with whether the text contains positive or negative sentiment). Steam user review score—the percentage of reviews that are positive—is the most straightforward metric. This is likely also the most relevant metric for a majority of players, as this score reflects the sentiments of the playerbase, or at least a vocal subset of the playerbase. As this score is featured prominently on the store page of each game, maintaining a Very Positive (80%+) or Overwhelmingly Positive (95%+) score likely bolster the brand image of the game and drive further sales. Such a correlation can be observed in Figure 4 in the Appendix.

Metacritic is one of the most referenced professional review sites in the entertainment industry. A study (Park and Byun, 2016) analyzed the correlation between user and professional critic scores, finding insignificant differences in scores for Indie and Casual games, but larger discrepancies among all other top-level genres, where the Metacritic score tends to be higher than the user score. This suggests the Steam user score and the Metacritic score should be examined as related but independent metrics. As only games with a certain level of popularity are likely to be reviewed on Metacritic, an analysis of Metacritic scores can provide us with insights on many AAA games but paint an incomplete picture of indie games.

Unfortunately for researchers, revenue data of a game is not publicly available. However, studies using private estimation algorithms (Zhang, 2022) or leaked data (VG Insights, 2021) find there is a high correlation between the number of reviews and the number of sales. To estimate the number of sales of a game from the number of reviews, a multiplier known as the "Boxleiter ratio" (Birkett, 2015) is commonly used. VG Insights estimates the Boxleiter ratio has steadily declined from an average of 60 for games released before 2014 down to an average of 30 for games released after 2020 (VG Insights, 2021), suggesting increased consumer behavior for reviewing games. This is in large part due to changes in the Steam platform itself that encourage users to write more reviews, leading firms like The Multiplayer Group (The Multiplayer Group, 2022) to develop empirically more accurate estimation heuristics that take into account not only the number of reviews, but also a game's year of release, price, and review score. Due to the lack of release time in the dataset we used, a Boxleiter ratio of 30 is used as a rough estimate.

Unfortunately, there is also no reliable method to estimate the revenue of a game from the number of copies sold, as games frequently go discounts which result in a significant portion of players obtaining the game for cheaper than the base price. Even though there are websites that track the prices of games over time, such as SteamDB.info or IsThereAnyDeal.com, we lack temporal information regarding how many sales were made at what price point. On the other hand, some games also distribute keys on alternative game stores like Fanatical.com or HumbleBundle.com. Reviews written by players who used a game key are not taken into account to the Steam user review score, which creates another source of bias in the data. Thus, while we calculate estimated revenue in this paper, such results should be interpreted with caution. Results of our estimation can be found in Figure 5 in the Appendix.

## 3 Methodology

### 3.1 Steam Scrape

The Steam website contains endpoints that allow us to easily retrieve information in JSON format. Three such endpoints were used:

- `api.steampowered.com/ISteamApps/GetAppList/v2` for retrieving a list of all apps.
- `store.steampowered.com/api/appdetails` for retrieving information about each app, including type (game, dlc, demo, or music), name, initial price (i.e., current undiscounted price), metacritic score (if available), and genres.
- `store.steampowered.com/appreviews` for retrieving total number of positive and negative reviews and contents of individual reviews. Only a subset of reviews were returned from this endpoint.

Data obtained is up to date as of May 2025.

### 3.2 Filter for Helpfulness

We first use a simple classifier to determine a binary variable *helpful* or *not helpful*. Similar to the *not helpful* category used by (Lin et al., 2019), we wish to filter out the many reviews that may only contain short comments (e.g. "Great game"), attempts at humor using in-game references (e.g. "Survivors have zero brain cells, but at least I can hit them when they acting up"), and all other types of reviews that are not particularly informative to prospective buyers. A *helpful* review is one that contains a detailed description of the player's in-game experience, suggestions, frustrations, and/or bugs.

Given the finding of (Lin et al., 2019) that positive reviews may also frequently contain useful information, there is little reason to use categories such as *pro*, *con*, *suggestion*, and *bug* since a review may contain multiple types of information, and categorizing them into mutually exclusive buckets may lose information.

We used a TF-IDF vectorizer with unigrams and bigrams, as well as two additional features: the number of words and average word length. A random forest classifier was used, with 250 manually labeled reviews for training and validation.

### 3.3 Issue Categories

For easier aggregation of results, we wish to discretize each review into a vector of binary variables indicating the presence or lack of each issue, so we first come up with a set of issues to check for. For our purposes, we provided GPT-4.1 with 200 randomly sampled negative *helpful* reviews and asked it to propose 10 categories of the most common issues. This list was then manually edited to minimize overlap between categories.

The issue categories chosen for analysis are as follows:

1. Bugs and Technical Problems
   Crashes, freezes, progression-blocking bugs, broken features, or severe frame drops. Performance and optimization issues may also be present even on supported or high-end hardware.

2. Confusing or Inadequate Onboarding
   Tutorials or onboarding may be missing, unhelpful, or unclear. New players may struggle to understand core mechanics, navigate menus, or get up to speed without frustration.

3. Poor Controls and Input Issues
   Clunky, unresponsive, or poorly implemented controls. Keybinding may be missing, controls may be counter-intuitive, lack controller support, or input methods may make gameplay frustrating or inaccessible.

4. Poor Balancing
   Frustrating balance issues such as oppressive matchups, difficulty spikes, or mechanics that either trivialize or stall progression. Meta-progression may override skill or core gameplay.

5. Shallow or Grindy Gameplay Loop
   Excessive grinding, artificial time-gating, forced repetition, or mechanics that pad game length without meaningful engagement.

6. Lack of Content
   Insufficient depth, variety, or progression. Gameplay may feel repetitive, lack replay value, or offer too little content to maintain long-term interest.

7. Lack of Polish
   The game may feel unfinished, with placeholders or incomplete systems. Missing features, rough edges, or lack important quality-of-life improvements.

8. Poorly Written Story and Characters
   Boring, predictable, shallow, unoriginal, or inconsistently written narratives. Dialogue may feel unnatural, overly verbose, awkward, or disconnected from the game world.

9. Abandonment or Poor Developer Support
   Unfinished or unsupported games. Ongoing bugs, missing features, infrequent updates, or unresponsive support may create the impression of neglect.

10. Paywalls or Microtransactions
    Progression or content gated behind purchases. Can lead to unfair advantages, imbalance in multiplayer, or diminished experience for free or budget-conscious players.

### 3.4 Issue Categorization

Each *helpful* review was processed using GPT-4.1-mini with the 10 issue categories as context. The

LLM was asked to return a structured output of 10 booleans indicating the existence of each issue within the review. Each query to the OpenAI API was made independently without additional context.

## 4 Results and Analysis

### 4.1 Steam Scrape

Data for 250,065 apps were obtained, of which 141,777 (56.7%) were games.

Games collectively received 1,033,077 reviews, of which 839,650 (81.3%) were positive and 193,427 (18.7%) were negative.

Only 4,182 (2.9%) of games had an associated Metacritic score.

### 4.2 Filter for Helpfulness

We manually labeled the helpfulness of 250 reviews randomly sampled from the dataset, which included 87 (34.8%) *helpful* reviews and 163 (65.2%) *not helpful* reviews. For the classifier, 200 (80%) reviews were used as the training set while the other 50 (20%) reviews were used as the validation set.

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| *helpful* | 1.00 | 0.25 | 0.40 |
| *not helpful* | 0.74 | 1.00 | 0.85 |

Table 1: Validation results of the random forest classifier.

The random forest classifier classified 1,032,827 unlabeled reviews into 111,176 (10.8%) *helpful* reviews and 921,651 (89.2%) *not helpful* reviews. The per-class breakdown is shown in Table 1, with an overall classification accuracy of 0.76. As can be seen from the high precision with low recall in the *helpful* class and low precision and high recall in the *not helpful* class, the classifier is overly conservative in predicting *helpful* reviews. This suits our purposes well, as this filtering step is primarily intended to reduce the cost of querying the LLM. Since false positives will likely be categorized as containing no issues, and false negatives tend to echo common themes in true negatives, classification errors here are unlikely to impact results significantly.

### 4.3 Issue Categorization

The 111,263 *helpful* reviews were classified by the LLM along the 10 issue categories, with results shown in Table 2. We can observe that Lack of Content, Lack of Polish, and Bugs and Technical Problems are the most common issues mentioned within reviews.

| Issue | Frequency |
|---|---|
| Content | 32.5% |
| Polish | 30.1% |
| Bugs | 29.5% |
| Balance | 22.8% |
| Controls | 21.8% |
| Gameplay Loop | 19.3% |
| Story | 15.7% |
| Onboarding | 12.4% |
| Support | 8.7% |
| Paywall | 3.5% |
| Total | 111,263 |

Table 2: Issue mention frequency in *helpful* reviews across all games.

| Issue | Indie | Non-Indie |
|---|---|---|
| Content | 33.9% | 31.4% |
| Polish | 31.0% | 30.6% |
| Bugs | 28.3% | 28.6% |
| Balance | 23.2% | 24.5% |
| Controls | 21.7% | 20.2% |
| Gameplay Loop | 18.6% | 19.6% |
| Story | 15.3% | 16.2% |
| Onboarding | 13.8% | 10.4% |
| Support | 8.5% | 9.1% |
| Paywall | 2.3% | 5.4% |
| Total | 66,459 | 44,804 |

Table 3: Issue mention frequency in *helpful* reviews for Indie versus non-Indie games.

We can further break down the results by comparing Indie games versus non-Indie games, as shown in Table 3. Notable differences include a higher prevalence of reviews mentioning Lack of Content (+2.5%) and Confusing or Inadequate Onboarding (+3.4%) in Indie games; they also had fewer mentions of Paywalls or Microtransactions (-3.1%). These findings align with our expectations of Indie games, which tend to have less development resources to create adequate content or playtesting capabilities to identify onboarding issues.

For Early Access games versus non-Early Access games as shown in Table 4, notable differences include a higher prevalence of reviews men-

| Issue | Early Access | Non-EA |
|---|---|---|
| Content | 38.0% | 32.2% |
| Polish | 35.9% | 29.0% |
| Bugs | 35.2% | 28.7% |
| Balance | 24.5% | 22.6% |
| Controls | 21.5% | 21.8% |
| Gameplay Loop | 19.2% | 19.3% |
| Story | 16.8% | 17.1% |
| Onboarding | 14.1% | 12.2% |
| Support | 5.4% | 7.6% |
| Paywall | 3.0% | 3.6% |
| Total | 13,581 | 97,682 |

Table 4: Issue mention frequency in *helpful* reviews for Early Access versus non-Early Access games.

tioning Lack of Content (+5.8%), Lack of Polish (+6.9%), Bugs and Technical Problems (+6.5%), and Confusing or Inadequate Onboarding (+1.9%) in Early Access games; they also had fewer mentions of Abandonment or Poor Developer Support (-2.2%). These results also align with our expectations of Early Access games, being primarily intended to offer a preview of the game while players have low expectations of it being complete, polished, or bug-free.

Furthermore, drastic differences in issue prevalence can be seen when dividing games by genre, as shown in Figure 1. Notable observations include:

- Casual games contain the most mentions of Lack of Content as well as the least mentions of Poor Balancing.
- Action games contain the most mentions of Poor Balancing.
- Adventure games contain the most mentions of Poorly Written Story and Characters; as well as the least mentions of Abandonment or Poor Developer Support.
- Simulation games contain the most mentions of Confusing or Inadequate Onboarding.
- Sports games contain the most mentions of Poor Controls and Input Issues; the most mentions of Lack of Polish; as well as the least mentions of Shallow or Grindy Gameplay Loop.
- Racing games contain the most mentions of Bugs and Technical Problems; the most mentions of Abandonment or Poor Developer Support; as well as the most mentions of Paywalls or Microtransactions.

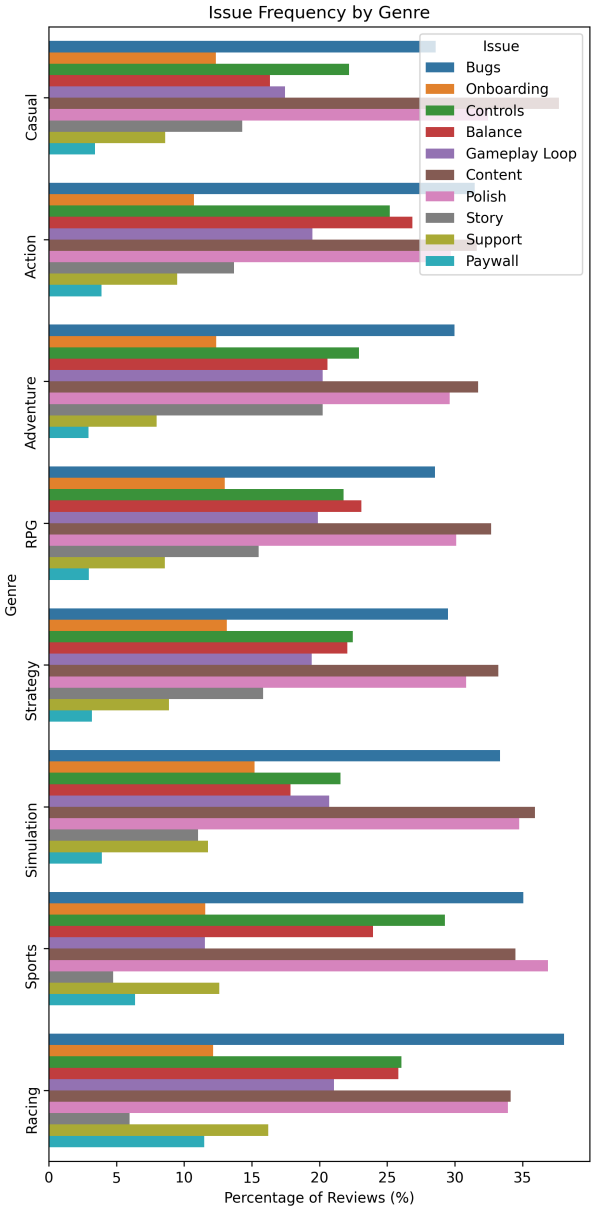


Figure 1: Bar chart of issue mention frequency in *helpful* reviews by genre.

### 4.4 Issue Correlation with Success

We plot issue mention frequency against Steam review score, Metacritic score, and estimated revenue for each of the 10 issues in Figures 6, 7, and 8 in the Appendix. Weak to non-existent explainability is observed among the plots. For Steam user review scores, Bugs and Technical Problems

($R^2 = 0.10$); Lack of Polish ($R^2 = 0.08$); Abandonment or Poor Developer Support ($R^2 = 0.08$); and Paywalls or Microtransactions ($R^2 = 0.05$) have the most explainable variation from issue frequency. For Metacritic scores, Lack of Polish ($R^2 = 0.06$); Bugs and Technical Problems ($R^2 = 0.04$), and Poor Controls and Input Issues ($R^2 = 0.04$) have the most explainable variation from issue frequency. Virtually none of the variation in estimated revenue was explainable by issue frequency.

The low explanatory power from Confusing or Inadequate Onboarding; Poor Balancing; Shallow or Grindy Gameplay Loop; Lack of Content; and Poorly Written Story and Characters across all three of the plots are also notable. Despite their non-negligible mention frequencies of 12.4%, 22.8%, 32.5%, and 19.3% respectively, these issues seem to have the virtually no influence on the Steam review score or Metacritic score. One possible interpretation is that these four issues are the least impactful to the player experience; another possible interpretation is that players are willing to give the developers most slack for the existence of these issues.

## 5 Conclusion

This study provides a data-driven lens into the prioritization of issue types in game development, revealing that while issues such as Bugs and Lack of Polish have slight impacts on user review scores, most other issue types show little statistical impact on success metrics. This finding affirms the intuitive assumption that not all types of issues are equally impactful to the player experience, however slight these differences may be. Instead, it suggests game studios that use a more strategic approach to resource allocation may be able to gain an advantage in this hypercompetitive industry.

The lack of any highly predictive issue should not be taken as evidence that these issues are unimportant—player experiences would certainly suggest otherwise. The low predictivity is moreso an indication of the fact that video games are incredibly complex endeavors where no single factor can determine its success or failure.

Aspects of this paper that could be improved upon by future research include the poor performance of the *helpful* versus *not helpful* classifier and inaccurate estimation of revenue. Additional research directions could explore temporal analysis of reviews written at different points of a game's lifecycle, such as reviews written during Early Access versus full release; incorporate active player count as an additional metric of success; or correlate existing results with the development costs associated with resolving each issue. Nonetheless, this work lays a foundation for more evidence-based game development prioritization, ultimately benefiting both studios and players.

## Acknowledgments

Thank you to Christina Wodtke for supporting this research project.

## References

Shuzlina Abdul-Rahman, Muhamad Faidi Akif Md Ali, Azuraliza Abu Bakar, and Sofianita Mutalib. 2024. Enhancing churn forecasting with sentiment analysis of steam reviews. *Social Network Analysis and Mining*, 14(1):178.

Richard Bartle. 1996. Hearts, clubs, diamonds, spades: Players who suit muds. *Journal of MUD research*, 1(1):19.

Jake Birkett. 2015. How to estimate how many sales a steam game has made.

Tibor Guzsvinecz and Judit Szűcs. 2023. Length and sentiment analysis of reviews about top-level video game genres on the steam platform. *Computers in Human Behavior*, 149:107955.

Robin Hunicke, Marc LeBlanc, Robert Zubek, et al. 2004. Mda: A formal approach to game design and game research. In *Proceedings of the AAAI Workshop on Challenges in Game AI*, volume 4, page 1722. San Jose, CA.

Dayi Lin, Cor-Paul Bezemer, Ying Zou, and Ahmed E Hassan. 2019. An empirical study of game reviews on the steam platform. *Empirical Software Engineering*, 24:170–207.

Hyemin Park and Haewon Byun. 2016. Correlation analysis: Game professional score and user score on steam. *International Journal of Multimedia and Ubiquitous Engineering*, 11(12):237–246.

The Multiplayer Group. 2022. Using steam reviews to estimate player numbers: An intuitive method.

VG Insights. 2021. How to estimate steam video game sales.

Haocheng Zhang. 2022. The establishment of multivariable linear regression in steam sales. In *2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)*, pages 853–856. Atlantis Press.

## Appendix

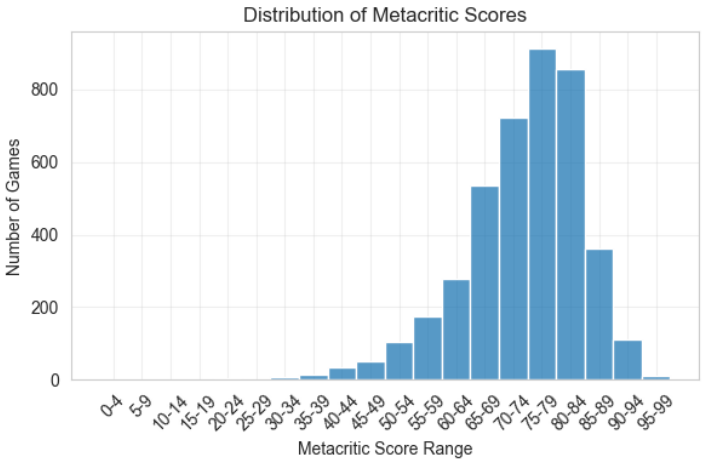


Figure 2: Histogram of Metacritic scores.

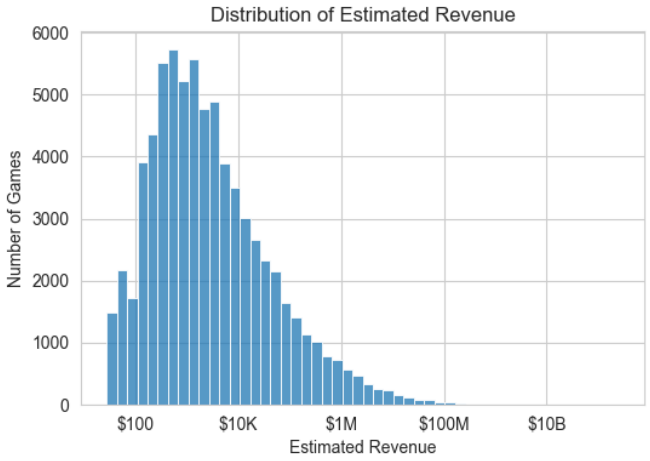


Figure 3: Histogram of estimated revenue.

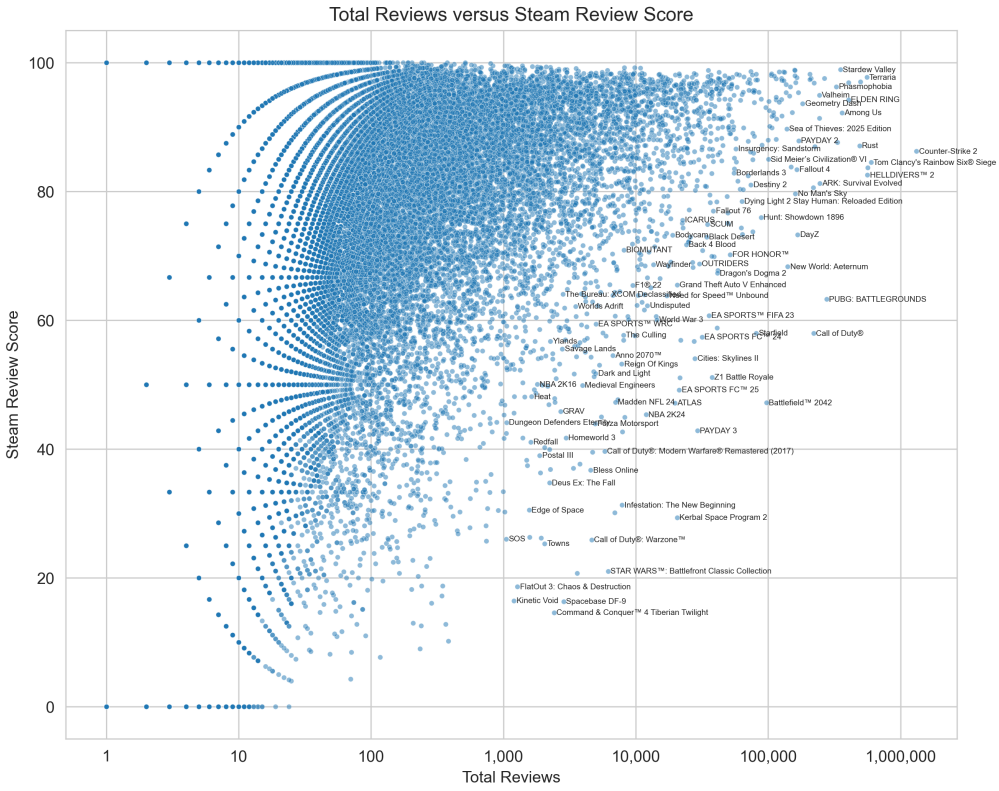


Figure 4: Scatter plot of Steam user review score versus number of reviews.


Figure 5: Scatter plot of Steam user review score versus estimated revenue.

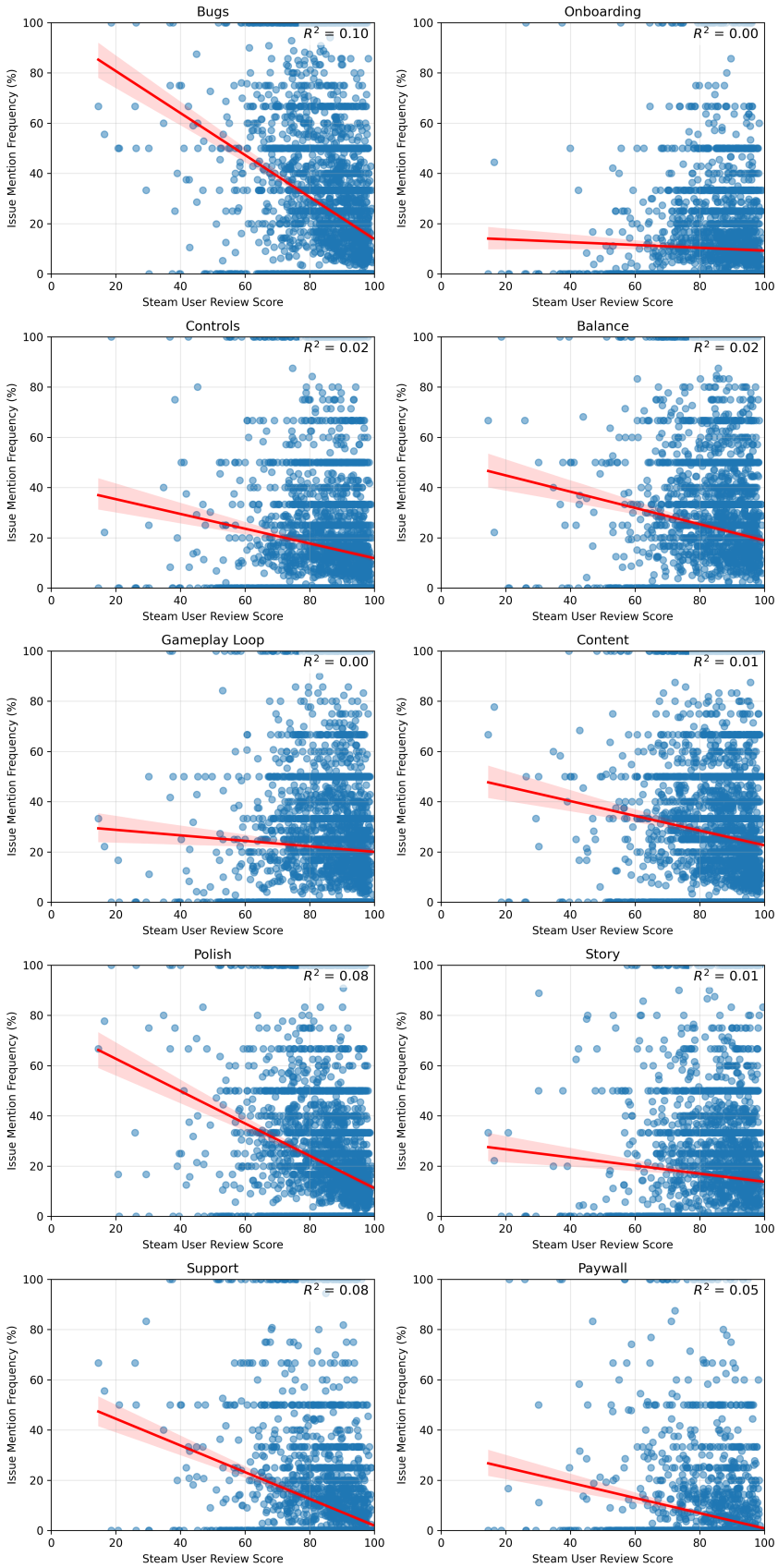


Figure 6: Scatter plots of issue mention frequency in *helpful* versus Steam review scores for games with at least 1,000 *helpful* reviews.

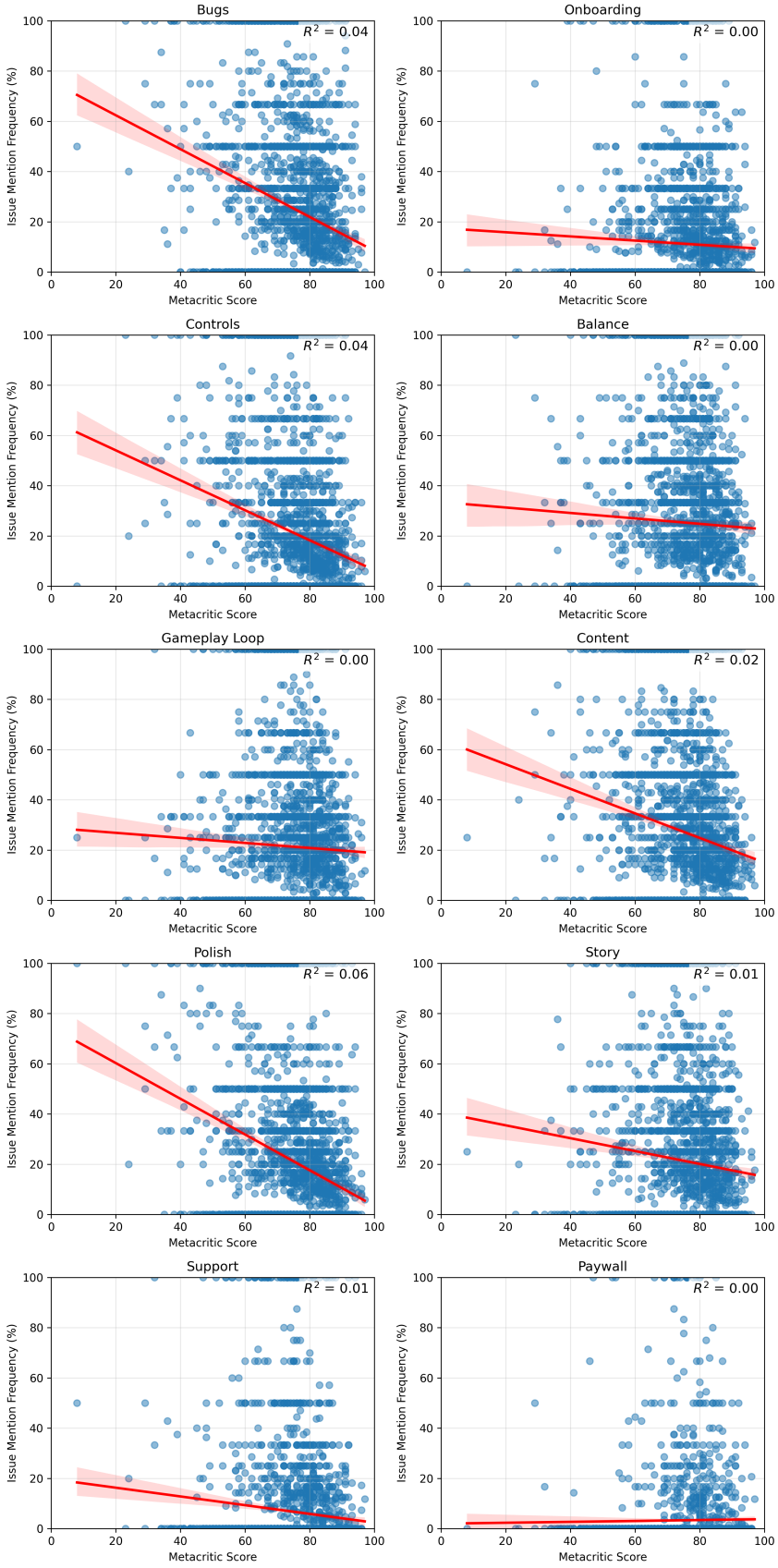


Figure 7: Scatter plots of issue mention frequency in *helpful* reviews versus Metacritic scores.

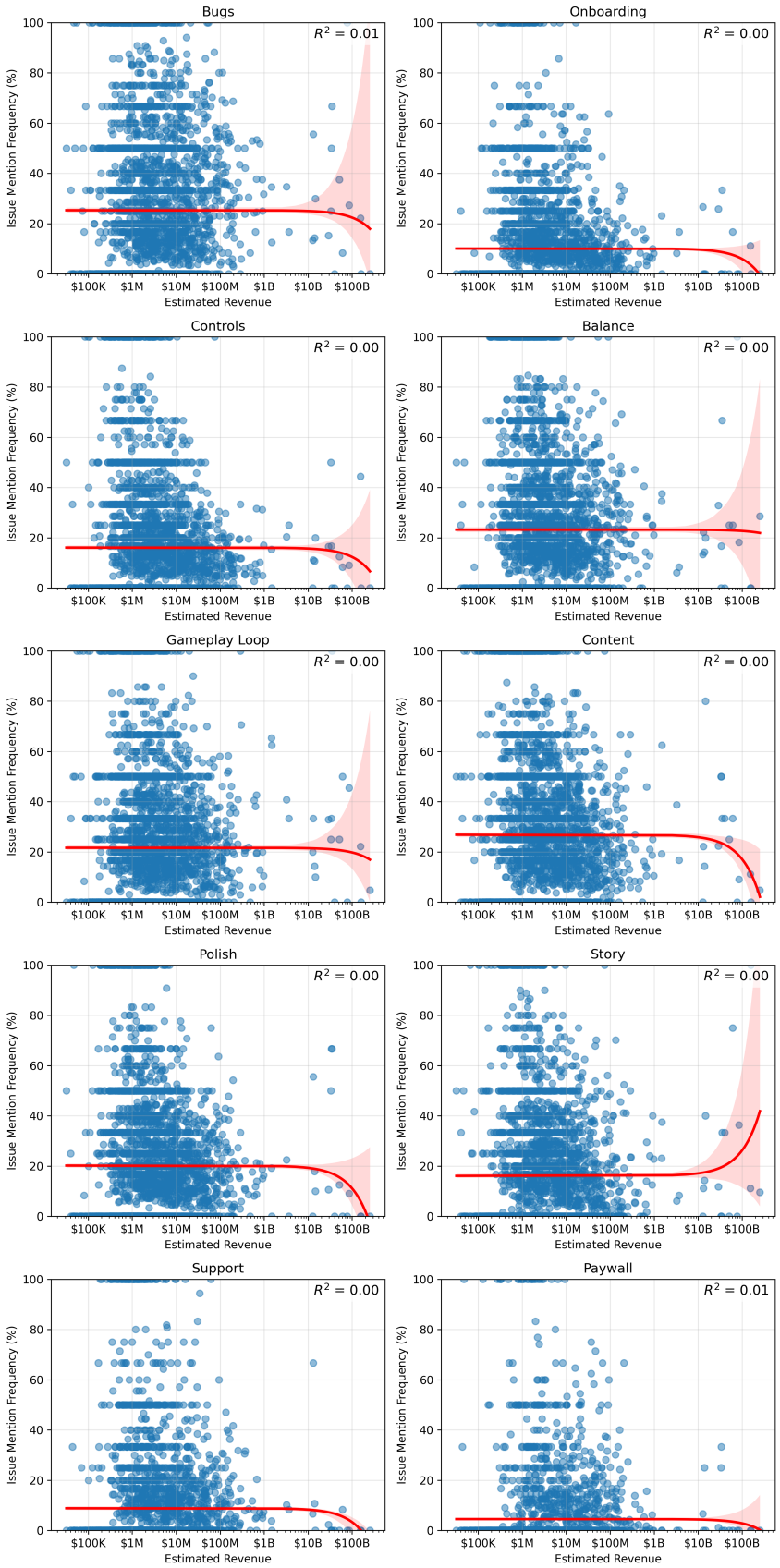


Figure 8: Scatter plots of issue mention frequency in *helpful* reviews versus estimated revenue for games with at least 1,000 *helpful* reviews.