Analyzing and Identifying Unique Archetypes in Reviews from *Grand Theft Auto:Vice City*

BINF5003 Data Mining, Modeling and Biostatistics

Professor Arastonejad

Sam Lenet, Abrar Faruque, Brenda Soljic

December 14, 2025

Introduction

Marketing and user perception strongly influence whether a video game becomes commercially successful. Online review platforms, such as Steam, provide rich behavioural information about the individuals who choose to write the reviews. By understanding the characteristics of these reviewers, it can provide value insight for developers, marketers, and consumers.

Grand Theft Auto: Vice City offers an ideal dataset for this type of analysis. As part of the long running *Grand Theft Auto (GTA)* franchise developed by Rockstar Games, the title attracts a wide and highly engaged player base with its story-driven adventure game and online multi-player roleplay elements that are targeted toward the late-teenage and early-adult audience (17+), selling over 215 million units worldwide as of August 2025 (Statista Research Department, 2025). The GTA:Vice City dataset contains 51, 801 Steam reviews posted between June 27th, 2023, and February 1st, 2024, providing a diverse sample of players with buried playtime histories and recommendation behaviours. In addition to recommendation status and review text, the GTA dataset includes variables such as vote direction, comment count, Steam purchase status, total playtime, and the number of games owned. This offers a broader view of reviewer engagement and platform activity.

This analysis aims to determine what characteristic groups exist among individuals who write reviews for GTA: Vice City. By examining engagement, writing effort (e.g. unique word count), and general platform activity, this study seeks to uncover natural behavioural clusters.

Research Question

What are characteristic behaviours in individuals who play Grand Theft Auto: Vice City based on their reviews?

Through this analysis, we aim to provide insight into how reviewing behaviour reflects different types of players. By understanding these behavioural patterns, it can help consumers

make more informed decisions about which reviews to trust when purchasing a new video game, and it can also aid game developers to better understand the characteristics and engagement styles of their player base.

Analysis

The dataset, summarized below (Table 1), contains 51, 801 reviews, each of which is tied to a unique reviewer, with no missing data across any variable. The most frequent review entries were short text snippets including, “good,” “fun,” and “yes,” reflecting a tendency toward low-effort reviews. Over 90% of all reviews consisted of one of the top ten repeated comments of similar, brief text. The review sentiment was strongly positive, with 90.7% of the reviews marked as *voted_up = TRUE.* This suggests that most players who chose to leave a review had a positive impression of *Grand Theft Auto: Vice City.* Interestingly, when measuring review interaction, *comment_count* and *votes_up* were near zero for most entries, which indicates that individual reviews rarely receive community engagement.

Reviewer activity metrics showed extremely wide variability and exhibited engagement levels ranging from only a few minutes played to over 1.4 million minutes, and many having no recent activity at all. Review effort was generally low, with a median of only 3 words. A small minority wrote highly detailed reviews, reaching over 1,800 words. Together, these patterns indicate a highly diverse reviewer population. This variability in engagement and writing effort provides a strong foundation for examining whether distinct behavioural groups exist among reviewers.

Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing Values
id\ [character]	[51801 others]		51801 (100.0%)	0\ (0.0%)
review\ [character]	1\. good(enter) 2\. yes(enter) 3\. good game(enter) 4\. fun(enter) 5\. .(enter) 6\. nice 7\. good 8\. gg(enter) 9\. cool(enter) 10\. good game [35281 others]	\ 1337 (2.6%)\ \ 687 (1.3%)\ \ 649 (1.3%)\ \ 502 (1.0%)\ \ 487 (0.9%)\ \ 383 (0.7%)\ \ 351 (0.7%)\ \ 167 (0.3%)\ \ 162 (0.3%)\ \ 156 (0.3%)\ \46920 (90.6%)	51801\ (100.0%)	0\ (0.0%)
created_date\ [Date]	min : 2023-06-27\ med : 2023-10-22\ max : 2024-02-01\ range : 7m 5d	220 distinct values	51801\ (100.0%)	0\ (0.0%)
voted_up\[logical]	1\. FALSE\ 2\. TRUE	\ 4830 (9.3%)\ \46971 (90.7%)	51801\ (100.0%)	0\ (0.0%)
votes_up\ [integer]	Mean (sd) : 0.6 (22.4)\ min < med < max:\ 0 < 0 < 3942\ IQR (CV) : 0 (38.2)	102 distinct values	51801\ (100.0%)	0\ (0.0%)
Comment_count\ [integer]	Mean (sd) : 0 (0.4)\ min < med < max:\ 0 < 0 < 46\ IQR (CV) : 0 (19)	18 distinct values	51801\ (100.0%)	0\ (0.0%)
author_num_games_owned\[integer]	Mean (sd) : 20 (127.8)\ min < med < max:\ 0 < 0 < 18126\ IQR (CV) : 6 (6.4)	662 distinct values	51801\ (100.0%)	0\ (0.0%)
author_num_reviews\ [integer]	Mean (sd) : 5.8 (12.6)\ min < med < max:\ 1 < 2 < 440\ IQR (CV) : 5 (2.2)	184 distinct values	51801\ (100.0%)	0\ (0.0%)
Author_playtime_forever (min)\ [integer]	Mean (sd) : 11199 (24777.6)\ min < med < max:\ 5 < 5052 < 1470929\ IQR (CV) : 9539 (2.2)	21087 distinct values	51801\ (100.0%)	0\ (0.0%)
author_playtime_last_two_weeks(min)\ [integer]	Mean (sd) : 342.7 (1028.8)\ min < med < max:\ 0 < 0 < 19849\ IQR (CV) : 151 (3)	3745 distinct values	51801\ (100.0%)	0\(0.0%)

author_playtime_at_review(min)\ [integer]	Mean (sd) : 8224.4 (22873.6)\ min < med < max:\ 5 < 2422 < 1275576\ IQR (CV) : 6655 (2.8)	17266 distinct values	51801\ (100.0%)	0\ (0.0%)
word_count\[integer]	Mean (sd) : 11.7 (41.8)\ min < med < max:\ 1 < 3 < 1883\ IQR (CV) : 7 (3.6)	441 distinct values	51801\ (100.0%)	0\ (0.0%)
Unique_word_count\ [integer]	Mean (sd) : 9.2 (23.3)\ min < med < max:\ 1 < 3 < 590\ IQR (CV) : 7 (2.5)	302 distinct values	51801\ (100.0%)	0\ (0.0%)
Author_last_played_date\ [Date]	min : 2015-04-22\ med : 2024-01-11\ max : 2024-02-01\ range : 8y 9m 10d	928 distinct values	51801\ (100.0%)	0\ (0.0%)

Table 1. ‘ID’ refers to a unique 9-digit code which correlates to an individual player. Players have the ability to leave multiple reviews, however, no individual review was >1% of total reviews and thus no one individual could influence the data significantly. ‘Review’ consists of the content of the review left by a reviewer, note that the most common comment often included a simple good, or a ‘good with an enter’ some left simply left an enter as their review. ‘Voted_up’ refers to the reviewer’s approval of the game. When ‘voted_up” is TRUE, it is assumed that the review was meant to be positive; if FALSE, then the review was meant to be negative. ‘Votes_up’ and ‘comment_count’ refer to the review themselves; others are able to leave comments on reviews or ‘vote_up’ a comment if they agree with the content of the review. ‘Author_playtime_forever’, ‘author_playtime_last_two_week’ and ‘author_playtime_at_review’ are all recorded in minutes as of the time when this dataset was updated, which was in October (3 months ago).

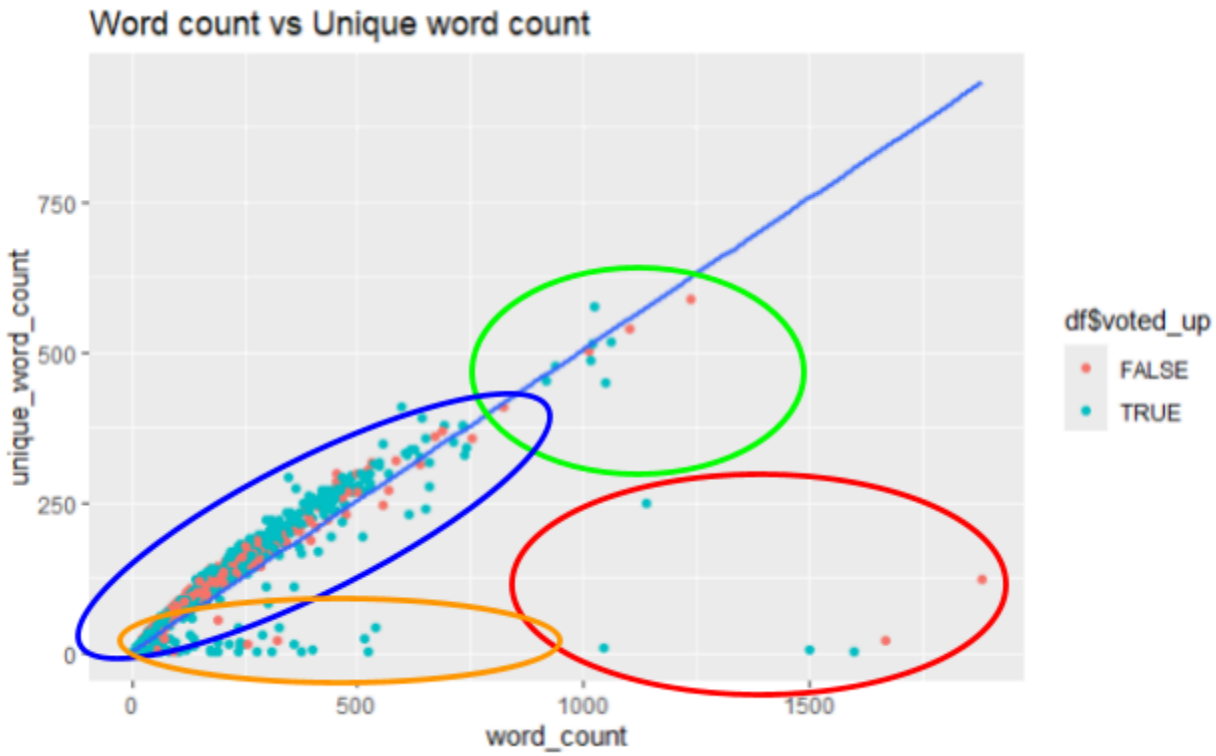


Figure 1. The orange circle (subgroup 1) shows reviews with <1000 words and <100 unique words, blue circle (subgroup 2) shows reviews with 1-500 words and less than <1000 unique words, green circle (subgroup 3) shows reviews with >=1000 words and >500 unique words, and red circle (subgroup 4) shows reviews with >1000 words and <500 unique words. There was no significant pattern between the four groupings and 'voted_up' TRUE or FALSE.

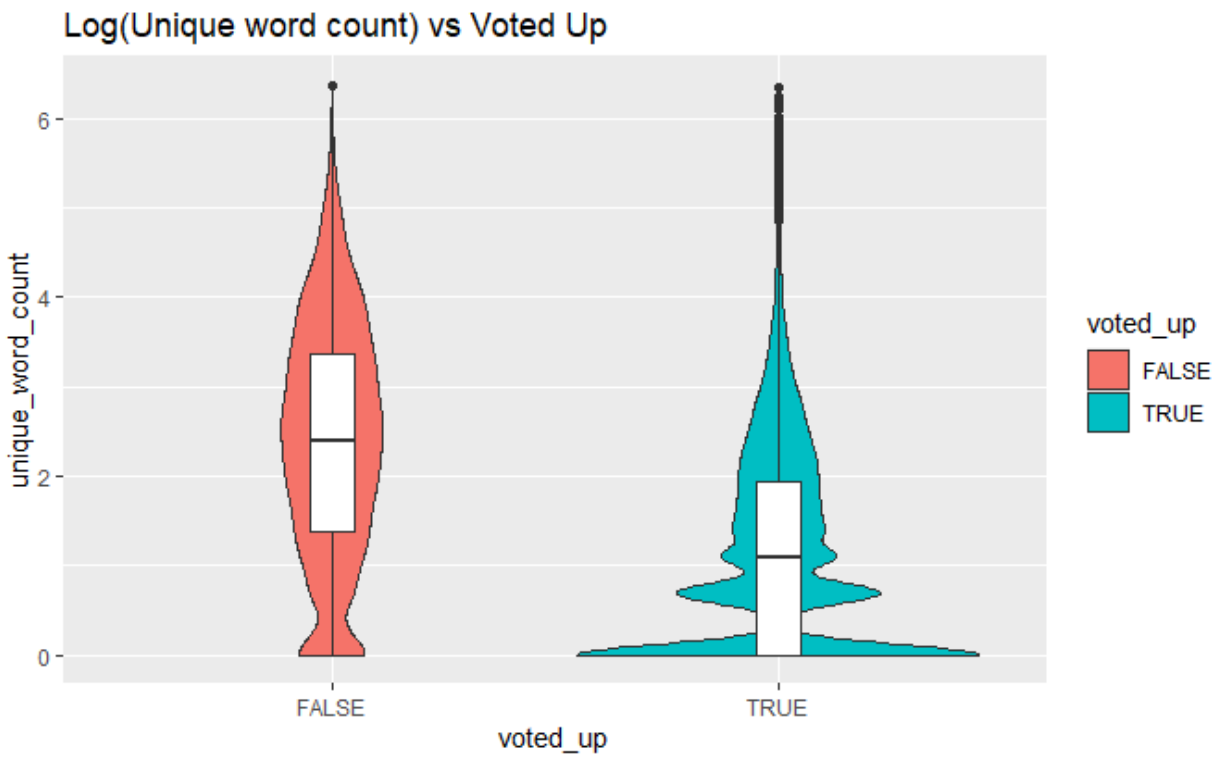


Figure 2. Violin+box plot showing the log(Unique word count) categorized by whether the review was positive or negative (voted_up, TRUE and FALSE, respectively). The Log(unique word count) mean was 1.165279 and 2.378818, standard deviation was 1.134085 and 1.34633, for voted_up TRUE and FALSE respectively. Voted_up FALSE reviews tended to follow a normal distribution for log(unique word count) (Sup. figure 1), with a noticeable group of reviews that had a small word count, which is reflected in the grouping of Figure 1. Voted_up TRUE reviews don't follow a normal distribution until roughly the average Log (unique word count) mark (Sup. figure 2). Unique word count of 1 was so prevalent in voted_up TRUE reviews that it made-up ~1/3 (32.95%) of all voted_up TRUE reviews.

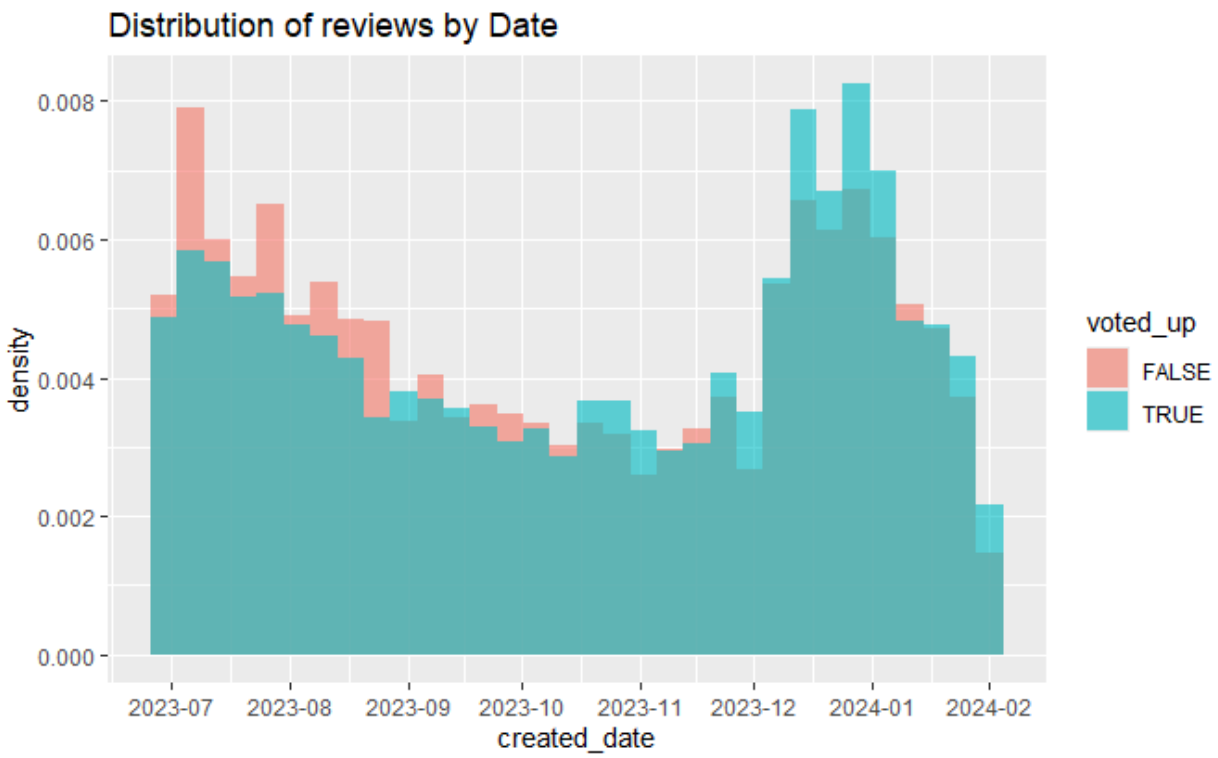


Figure 3. Distribution of reviews by date from June 27th 2023 to February 1st 2024 and if the review was voted_up TRUE or FALSE. From 2023 June to October, there is a general trend of more negative reviews than positive reviews, after October voted_up TRUE overtook the number of voted_up FALSE reviews. There is a noticeable spike in the number of reviews specifically after December of 2023, with more than 2 (2.08) times more reviews in December than the month prior (November 2023).

Cluster Analysis of Numerical Data from the Dataset (UMAP)

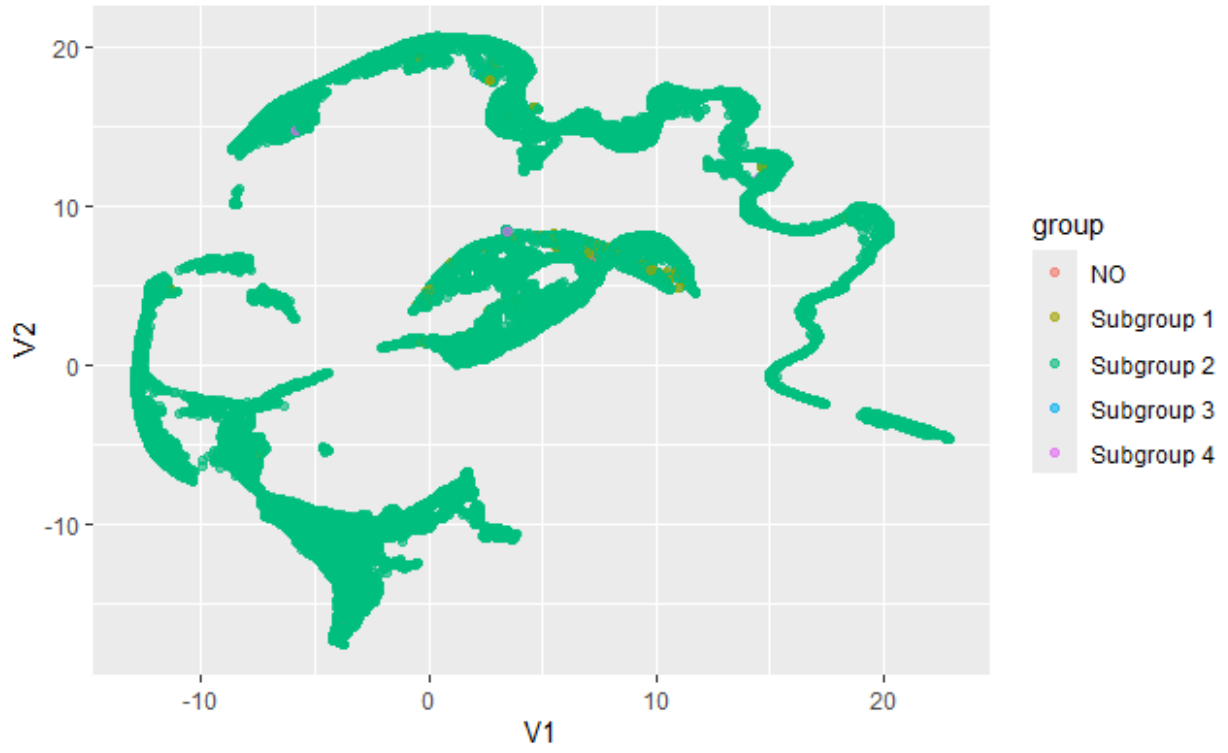


Figure 4. UMAP shows some distinct clustering, however, due to the sheer number of reviews that fall under group 2 (refer to Figure 1), any subtle groupings that would have been made from the other groups is overwhelmed, further investigation is required to identify any meaningful patterns in the clustering data.

Discussion and Conclusion

Determining Effort

There is no specific metric used to determine effort; however, for this analysis we assumed that reviews with higher word counts and greater unique word diversity reflect greater effort, a reasonable approximation given that longer reviews tend to be more helpful than very short or empty reviews (Mellinas & Leoni, 2024).

As expected, word count and unique word count were strongly correlated. Four general clusters emerged from this relationship (Figure 1). Most reviews fell within a near 2:1 ratio of total words to unique words (subgroups 2 and 3), representing brief to moderately expressive reviews. Two smaller subgroups (subgroups 1 and 4) showed unusually low unique word

diversity relative to length, suggesting repetitive or low-information content. However, after further investigation, many of these low-diversity reviews consisted of the same repeated words or phrases several times. While the intent behind these types of reviews cannot be determined from the data alone, their structure is consistent with spam-like or low-effort posting. This pattern suggests the presence of a reviewer subtype that contributes minimal or repetitive content, a behaviour commonly associated with “spamming/trolling” in online communities. Importantly, writing effort did not correlate and is not a predictor of whether a review voted the game up or not.

Another notable trend is the difference of positive (voted_up = TRUE) and negative (voted_up = FALSE) reviews. Positive reviews tend to require less effort, whereas negative reviews generally contain more unique words and greater elaboration (Figure 2.). This matches findings from a recent study of Steam reviews across videogame genres, which found that negative reviews tend to be longer than positive ones (Guzsvinecz & Szűcs, 2023). That being said, longer reviews should not be assumed to be more helpful, a study done by Lutz et al. argues that quality of argumentation matters more than length. They state that longer reviews help provide better points about product quality and previous experiences playing the game (Lutz et al., 2019).

In our data, positive reviews generally required less writing effort as players who enjoyed the game often express approval with a brief statement (e.g. “good”) without feeling the need to justify further, compared to longer negative reviews where an individual is more likely to elaborate on why they think the game is bad. This aligns with the observation that over 30% of positive reviews consist of only a single unique word. From this pattern, two additional reviewer archetypes emerge, the “simpleton,” leaving minimal but positive feedback, and the “critic,” who provides a detailed comment when expressing their dissatisfaction.

Influence of Canonical Events on Reviews

Plotting the reviews in canonical order (Figure 3) revealed a general decline in the frequency of reviews following a large canonical event. Interestingly, despite the overall dominance of positive reviews, the amount of negative reviews outnumbered the positive ones from June 2023 to November 2023. The trend shifted after December 2023, corresponding with a significant game update resolving several previously reported issues. Research done by Zhong and Xu in 2022 on player behaviour supports the idea that major updates can lead to renewed engagement, they found that there was a 11-49% increase in player activity following new updates to a game (Zhoung & Xu, 2022). This would support the fact that around this time where we see the spike, there was an announcement of Grand Theft Auto 6, renewing an interest in the franchise and prompting the players to revisit Vice City. Although these temporal patterns are noteworthy, there was no significant correlation between when the person purchased the game (author_playtime_forever) and the time of their review (author_playtime_at_review) (Supplementary table 1).

Overall, people who purchase video games should look at the content and clarity of reviews, not just their length. And in turn, game developers should focus less on word count and more on what players are actually saying.

References

Guzsvinecz, T., & Szűcs, J. (2023, December). *Length and sentiment analysis of reviews about top-level video game genres on the steam platform*. ELSEVIER. https://www.med.upenn.edu/pmi/events/https-www-sciencedirect-com-science-article-abs-pii-s1047847720300046-via-3dihub

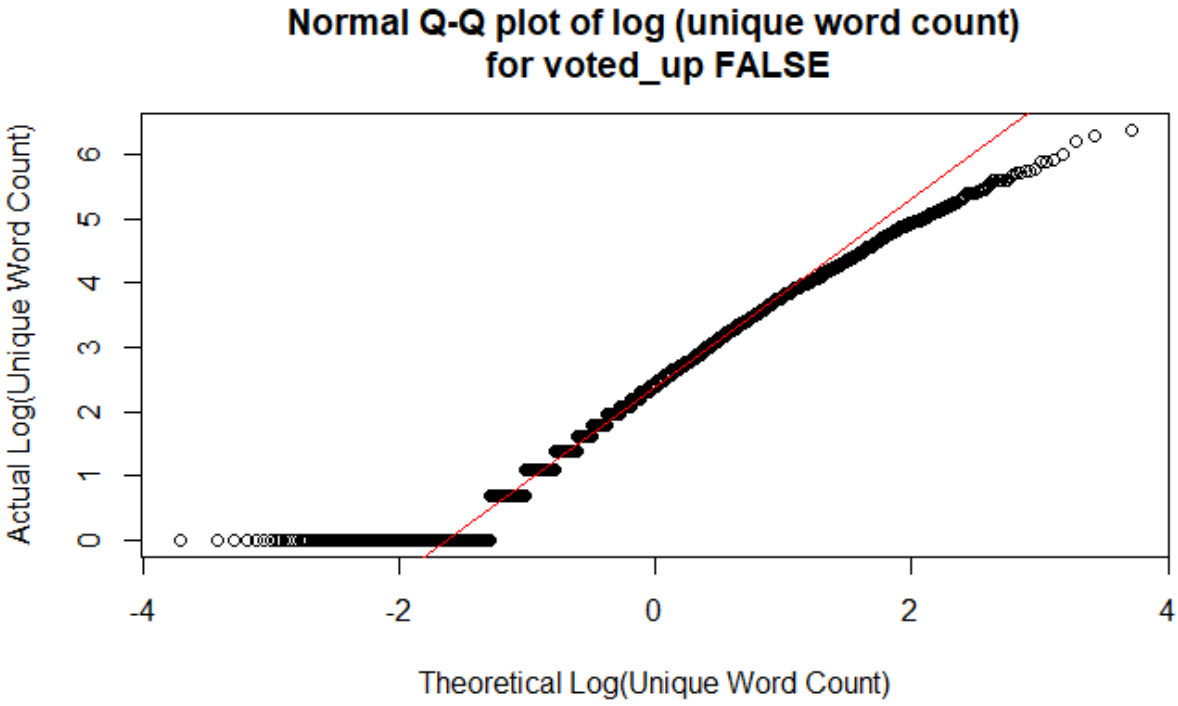
Lutz, B., Prollochs, N., & Neumann, D. (2019, September). *The Longer the Better? The Interplay Between Review Length and Line of Argumentation in Online Consumer Reviews*. arXiv. http://arxiv.org/pdf/1909.05192

Mellinas, J. P., & Leoni, V. (2024, September). *Beyond words: unveiling the implications of blank reviews in online rating systems*. SpringerLink. https://link.springer.com/article/10.1007/s40558-024-00300-4

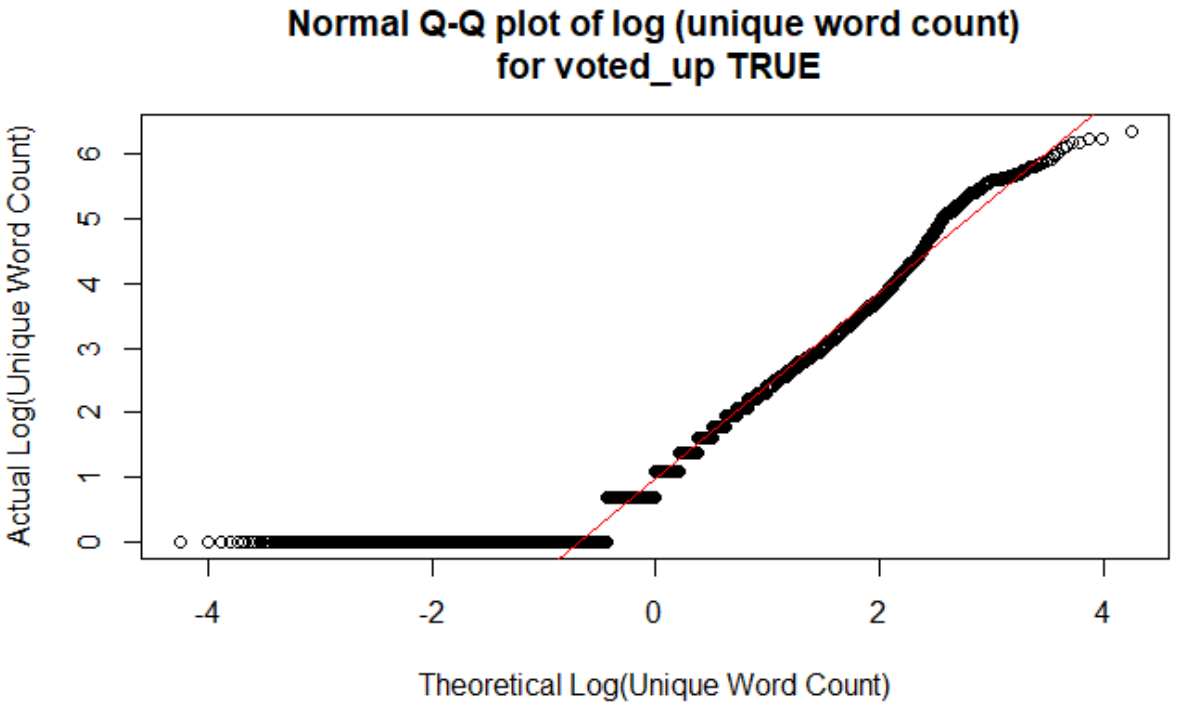
Statista. (2025, August). *Grand Theft Auto V Total Unit Sales 2025*. Statista. https://www.statista.com/statistics/1247955/gta-v-unit-sales-worldwide-total/#statisticContainer

Zhong, X., & Xu, J. (2022, June). *Measuring the effect of game updates on player engagement: A cue from DOTA2*. Research Gate. https://www.researchgate.net/publication/361091263_Measuring_the_effect_of_game_updates_on_player_engagement_A_cue_from_DOTA2

Supplementary Figures



Supplementary figure 1. Q-Q plot of log(unique word count) for voted_up FALSE reviews



Supplementary figure 2. Q-Q plot of log(unique word count) for voted_up FALSE reviews

	votes_up	comment_count	author_num_games_owned	author_num_reviews	author_playtime_forever	author_playtime_last_two_weeks	author_playtime_at_review	word_count	unique_word_count
votes_up	1	0.247577625	0.03244518	0.028370068	0.010954348	-0.002363583	0.01215683	0.02641987	0.03382086
comment_count	0.247577625	1	0.01785066	0.033160487	0.044846807	0.009388418	0.04311027	0.04961258	0.05958347
author_num_games_owned	0.03244518	0.017850662	1	0.303477984	0.055151656	-0.012149823	0.06425121	0.09699401	0.11810162
author_num_reviews	0.028370068	0.033160487	0.30347798	1	0.005699799	-0.026804472	0.01529067	0.12085607	0.14671054
author_playtime_forever	0.010954348	0.044846807	0.05515166	0.005699799	1	0.219384902	0.967435	0.08304388	0.09461056
author_playtime_last_two_weeks	-0.002363583	0.009388418	-0.01214982	-0.026804472	0.219384902	1	0.10968487	0.01392542	0.01398292
author_playtime_at_review	0.012156833	0.043110272	0.06425121	0.015290672	0.967434997	0.109684875	1	0.08789817	0.10025415
word_count	0.026419873	0.049612583	0.09699401	0.120856067	0.08304388	0.013925422	0.08789817	1	0.90232516
unique_word_count	0.033820861	0.059583467	0.11810162	0.146710541	0.094610559	0.013982924	0.10025415	0.90232516	1

Supplementary Table 1. The correlation matrix of all numeric values excluding date review was made ‘created_date’ and the last time the author had played the game ‘author_last_played’, a cut off of 0.9 was used. The two relations that fall above the cut-off were ‘author_playtime_forever - author_playtime_at_review’ and ‘word_count - unique_word_count’.