Exploratory Data Analysis of 2021 Steam Video Game Reviews

Mert Atay
linkedin.com/in/mertatay

Abstract

The main goal of this project is to concentrate on two main V's of big data: Volume and Value. This is achieved through an exploratory data analysis that centers around examining trends and user behavior within a vast dataset consisting of video game reviews. The ultimate aim is to offer valuable insights to the video game industry, which can be beneficial for various aspects such as design, development, community building, and public relations.

1 Introduction

With the massive expansion in the recent years, video gaming industry has become one of the largest markets in the globe, surpassing traditional entertainment venues like movies and music[1]. This growth presents a significant opportunity for investors and developers seeking to enter this thriving market. However, as the supply of video games continues to increase, it becomes crucial to accurately discern trends and preferences across different regions and user groups, enabling the design of games that align with market demands. In this context, leveraging big data analytics encompassing both volume and value aspects can yield valuable insights to address our research questions effectively.

To fulfill this need, an ideal source of data lies in the form of video game reviews directly provided by players themselves. Steam[2], a digital distribution platform catering to both video game enthusiasts and developers, has been operational since 2003 and boasts approximately 30 million daily active users. For this project, a sizable dataset of 8.2 GB comprising video game reviews from the Steam platform in 2021 is utilized to capture the voluminous aspect of our big data analytics endeavor.

To encompass the value aspect, an exploratory data analysis is conducted on this extensive dataset. The analysis comprises two key components. Firstly, we aim to identify trends and video game preferences for the year 2021. Secondly, we delve into an examination of user behavior when composing reviews for video games.

The analysis follows a structured approach that involves formulating research questions and subsequently addressing them through data manipulation techniques. By conducting this analysis, I can summarize my contributions and findings as follows:

- Identified the most and least popular, loved, and played titles among all users, top consumers, and top reviewers.
- Conducted a thorough comparison to provide valuable insights for game designers and developers.
- Observed distinct trends and preferences among different user groups and geographical regions, emphasizing the importance of selecting target markets and audiences.
- Noted that users tend to review gifts more positively, suggesting that offering free content can enhance public relations.
- Identified the most helpful and entertaining communities, enabling community managers to study these titles and cultivate supportive communities for their video games. Overall, users share a common sense of humor, as the most popular titles have the funniest reviews.

2 Related Work

To best of my knowledge, there is currently no comprehensive data analysis or academic work that directly focuses on video game trends and preferences using a large-scale dataset. While some basic data analyses and machine learning applications utilizing similar Steam video game reviews datasets can be found on platforms like Kaggle, they neither encompass as much data as my analysis nor possess the capability to address our specific research questions.

3 Methodology

3.1 Dataset

For this project, I used the publicly available Steam Reviews Dataset 2021, which can be accessed on Kaggle[3]. The dataset encompasses a staggering 21,747,371 video game reviews authored by 12,406,560 unique users. These reviews are written in 28 different languages, resulting in a dataset size of approximately 8.2 GB. Within the dataset, there are 23 columns, with 16 columns deemed relevant to our research. During the analysis, irrelevant columns are dropped, and the names and relevancy information of the remaining columns are as follows (note that an asterisk indicates irrelevancy):

[1] nasdaq.com | This Opportunity for Investors Is Bigger Than Movies and Music Combined

[2] steampowered.com/about

[3] kaggle.com/datasets/najzeko/steam-reviews-2021

Column Name	Data Type
index	int
app_id	string
app_name	string
review_id	string
language	string
review	string
timestamp_created	datetime
timestamp_updated*	datetime
recommended	boolean
votes_helpful	int
votes_funny	int
weighted_vote_score*	float
comment_count*	int
steam_purchase	boolean
received_for_free	boolean
written_during_early_access*	boolean
author.steamid	int
author.num_games_owned	int
author.num_reviews	int
author.playtime_forever (mins)	float
author.playtime_last_two_weeks (mins)*	float
author.playtime_at_review (mins)*	float
author.last_played*	datetime

3.2 Technologies And Libraries

The analysis is conducted using Python language in the form of IPython Notebooks. For big data analytics, PySpark library is used with two main imports: Pyspark.sql and Pyspark.pandas. Analysis is performed via data manipulation using DataFrames. The results are presented in various chart visualizations for this task, Matplotlib library is used.

3.3 Exploratory Data Analysis

The data analysis follows a structured approach of identifying research questions and subsequently providing corresponding answers. The research questions for each section are as follows:

- Identifying trends in video game preferences in 2021
  - Which video games/genres were the most/least popular?
  - Which video games/genres were loved/hated (i.e., which video games have the largest fan base)?
  - Which video games/genres have the most playtime (i.e., which video games are more addictive)?
  - Answer these three questions again but this time focus on the following user groups: Top reviewers and top consumers.
  - Is there a difference in video game preferences across various geographical regions?
  - Which geographical regions purchase more video games?
- Observing user behavior when writing reviews for video games
  - Do users write reviews more when they love a video game or hate a video game?
  - For the video games they did not purchase, do users write positive reviews or negative reviews more?
  - Which video game/genre fans are more helpful in reviews?
  - Which video game/genre fans are funnier in reviews?

4 Experiment Results

Results and the respective research question are as follows:

4.1 Which video games/genres were the most/least popular?

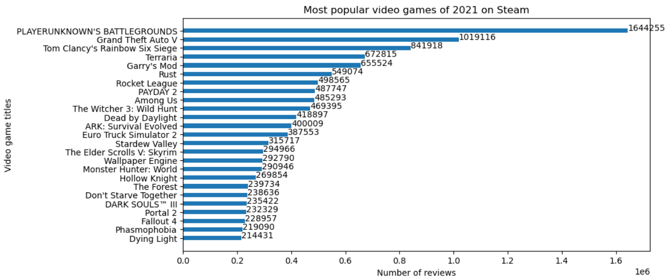

Figure 1. Most popular video games of 2021 on Steam (All Users)

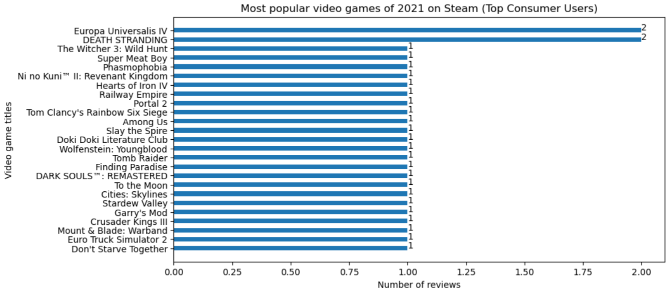

Figure 2. Most popular video games of 2021 on Steam (Top Consumer Users)

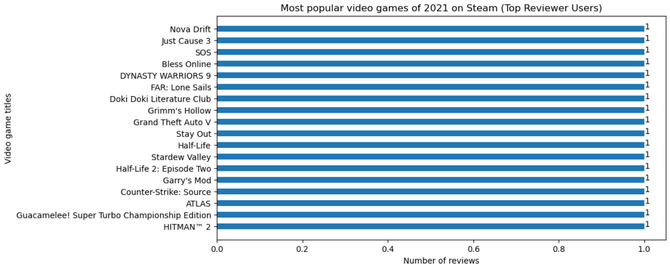

Figure 3. Most popular video games of 2021 on Steam (Top Reviewer Users)

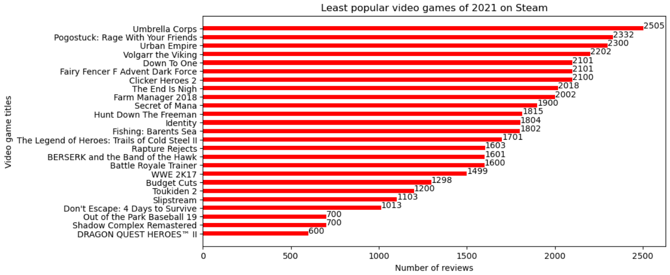

Figure 4. Least popular video games of 2021 on Steam (All Users)

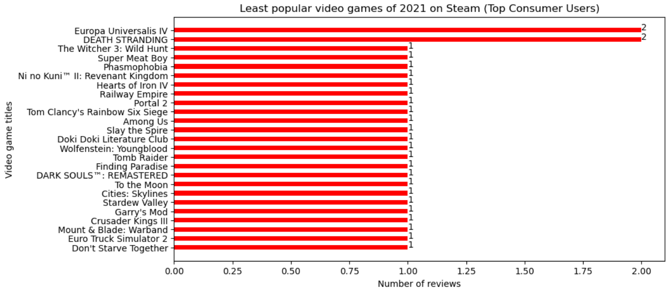

Figure 5. Least popular video games of 2021 on Steam (Top Consumer Users)

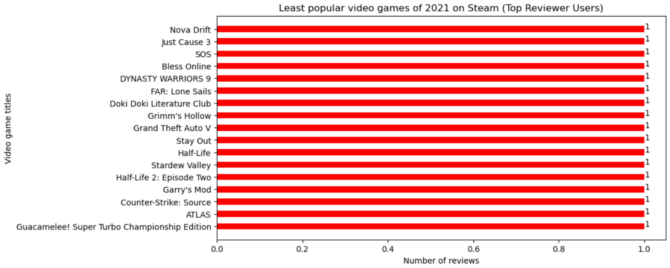

Figure 6. Least popular video games of 2021 on Steam (Top Reviewer Users)

Looking at the results, we can identify the most popular titles among different user groups. It is evident that distinct trends and preferences exist among these groups. Therefore, it is essential for designers and developers to be mindful of their target audience. Additionally, they can delve deeper into the analysis of the most popular titles among all users to pinpoint the specific features that contribute to their popularity. A preliminary observation reveals that these titles tend to incorporate either multiplayer functionality or captivating storylines.

Regarding the least popular titles, no unanimous consensus emerges. Some of these titles are either significantly outdated, could be plagued with bugs, or simply lack any form of appeal. Nonetheless, developers can still conduct further analysis on these titles to identify particular mechanics or methods that should be avoided.

4.2 Which video games/genres were loved/hated?

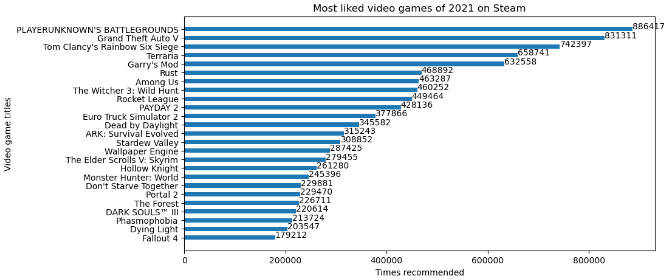

Figure 7. Most liked video games of 2021 on Steam (All Users)

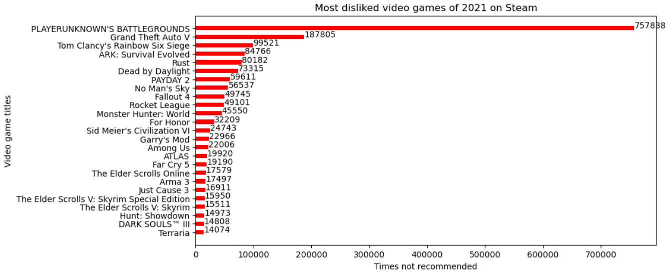

Figure 8. Most disliked video games of 2021 on Steam (All Users)

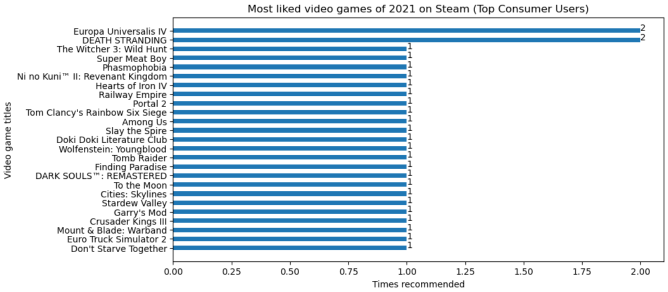

Figure 9. Most liked video games of 2021 on Steam (Top Consumer Users)

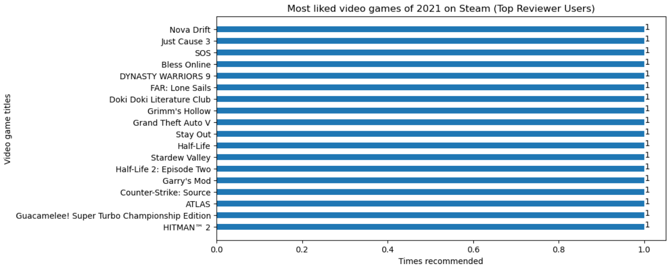

Figure 10. Most liked video games of 2021 on Steam (Top Reviewer Users)

Upon initial observation, it becomes apparent that the most popular titles are typically the ones that receive the highest number of likes. Therefore, a comparison between the most liked and disliked titles can provide a more comprehensive understanding by examining the like/dislike ratio. Notably, "PLAYERUNKNOWN'S BATTLEGROUNDS" stands out as a title that garners both the most likes and dislikes, indicating mixed reviews. Additionally, there are titles such as "Portal 2" that receive overwhelmingly positive reviews, while others like "No Man's Sky" tend to have predominantly negative reviews. By conducting further analysis on these titles and their development processes, designers and developers can discern effective strategies to adopt and pitfalls to avoid. For instance, the negative reviews of "No Man's Sky" can be attributed to its controversial launch.

Furthermore, discernible differences in preferences among various user groups continue to emerge. It is worth noting that top consumer users and top reviewer users exclusively provide reviews for the titles they enjoy, reinforcing the notion of user selectivity when it comes to providing feedback.

4.3 Which video games/genres have the most playtime?

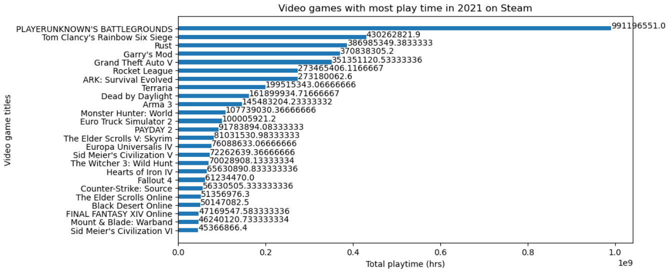

Figure 11. Video games with most playtime in 2021 on Steam (All Users)

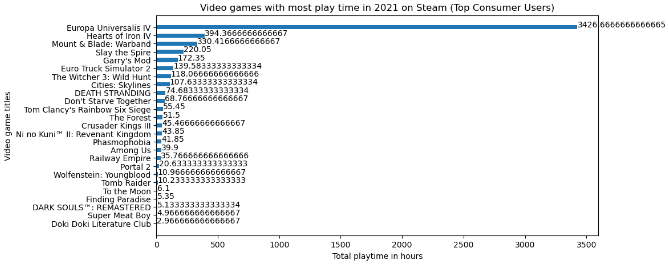

Figure 12. Video games with most playtime in 2021 on Steam (Top Consumer Users)

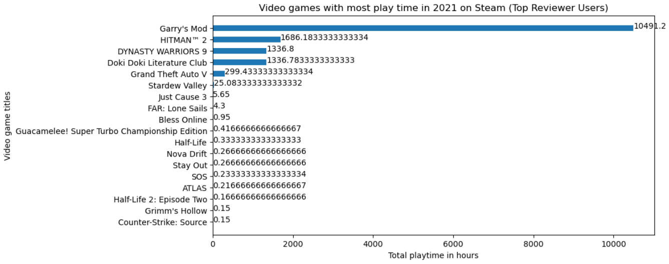

Figure 13. Video games with most playtime in 2021 on Steam (Top Reviewer Users)

Analyzing titles with the highest playtime can provide insights into the factors that make these video games addictive. For all users, it appears that multiplayer titles dominate the list. However, when examining the preferences of top consumer users and top reviewers, we find that they are hooked on different video game titles, including some single-player experiences.

4.4 Is there a difference in video game preferences across various geographical regions?

Although our dataset lacks explicit geolocation data, we can leverage language as a proxy indicator for regional grouping. However, it's important to note that languages such as English, Spanish, and Russian are not included in this grouping, as they may span multiple countries across the globe. The distribution of data among languages is as follows:

Language	Review Count
english	9635437
schinese	3764967
russian	2348900
brazilian	837524
spanish	813320
german	752596
turkish	635868
koreana	613632
french	541751
polish	495529
tchinese	218203
czech	133980
italian	133307
thai	127503
japanese	81754
portuguese	81386
swedish	80226
dutch	77555
hungarian	71001
latam	70103
danish	55915
finnish	54712
norwegian	36797
romanian	32730
ukrainian	21169
greek	14472
bulgarian	10454
vietnamese	6580

Regional mapping of languages is as follows:

- **Europe:** German, Turkish, French, Polish, Czech, Italian, Portuguese, Dutch, Hungarian, Romanian, Ukrainian, Greek, Bulgarian
- **Asia:** Schinese, Korean, Tchinese, Thai, Japanese, Vietnamese
- **Latin America:** Brazilian, Latam
- **Scandinavia:** Swedish, Danish, Finnish, Norwegian

In order to observe the differences, we can focus on popularity.

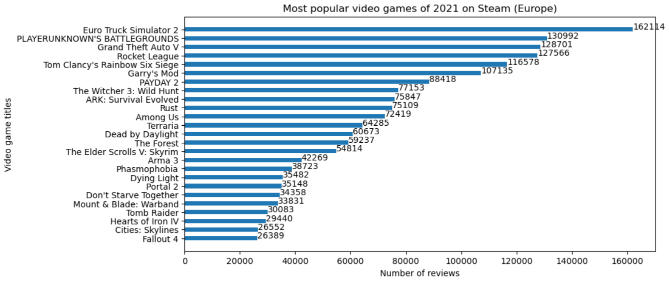

Figure 14. Most popular video games of 2021 on Steam (Europe)

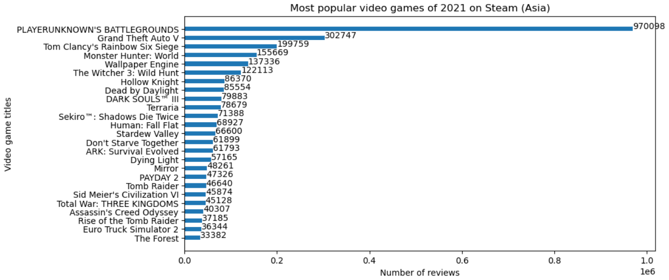

Figure 15. Most popular video games of 2021 on Steam (Asia)

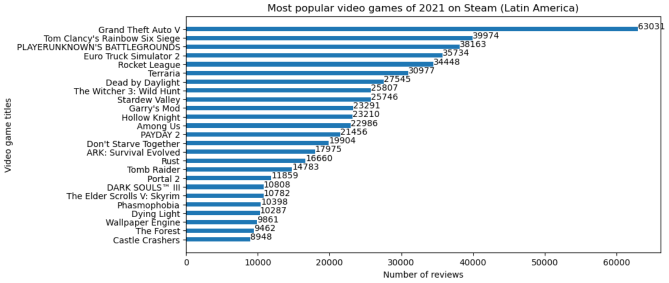

Figure 16. Most popular video games of 2021 on Steam (Latin America)

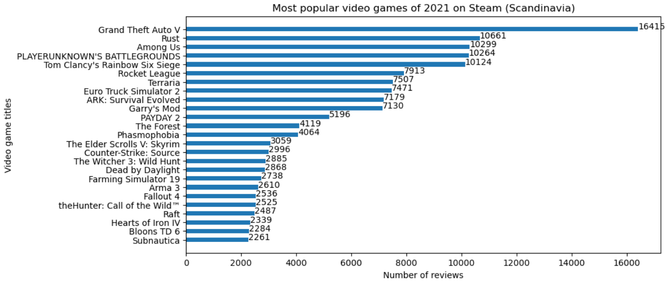

Figure 17. Most popular video games of 2021 on Steam (Scandinavia)

Upon analyzing the results, it becomes evident that different regions exhibit distinct video game trends. This underscores the importance for developers to be attentive to their target market and its preferences.

4.5 Which geographical regions purchase more video games?

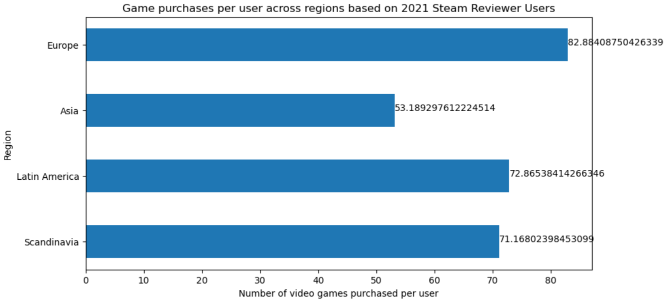

Figure 18. Game purchases per user across regions based on 2021 Steam reviewer users

Upon closer examination, it appears that Europe has the highest number of purchases per user. This finding highlights the significance of developers not only being mindful of the trends within their target markets but also carefully selecting their target markets.

4.6 Do users write reviews more when they love a video game or hate a video game?

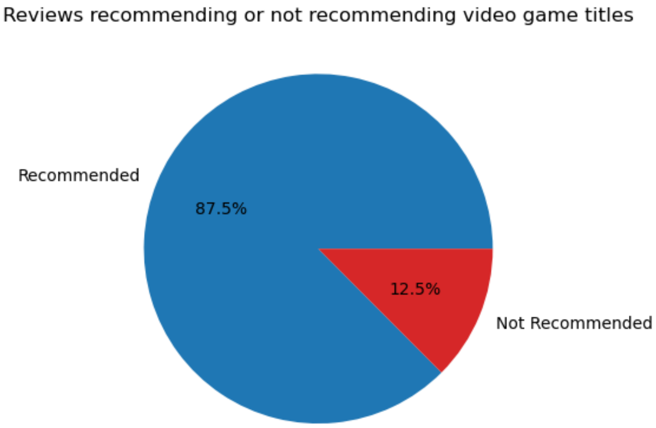

Figure 19. Reviews in 2021 recommending or not recommending video games

4.7 For the video games they did not purchase, do users write positive reviews or negative reviews more?

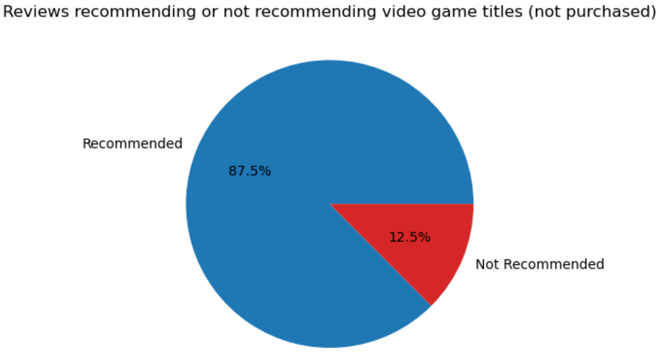

Figure 20. Reviews in 2021 recommending or not recommending video games not purchased

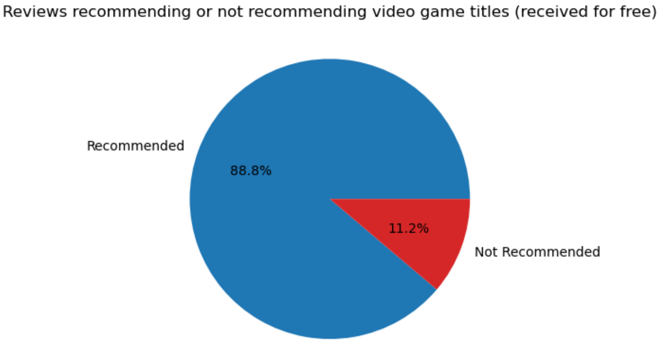

Figure 21. Reviews in 2021 recommending or not recommending video games gifted

Users seem to write reviews for video games they enjoy, which is an important finding in this analysis. A noteworthy takeaway is that we observe a 1% increase in positivity when users receive the video game as a gift. While this increase may appear modest, considering the scale of our extensive dataset, it corresponds to approximately 200,000 reviews. This suggests that users tend to rate gifted video games more positively. Community managers and developers can leverage this insight by offering free content to enhance the reputation of their video games.

4.8 Which video game/genre fans are more helpful in reviews?

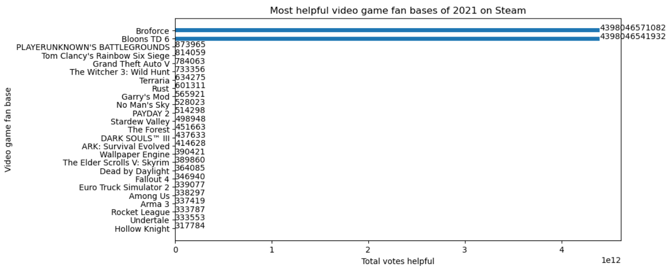

Figure 22. Most helpful video game fan bases of 2021 on Steam

Examining the results, we discover that two niche titles, Broforce and Bloons TD 6, receive an exceptionally high number of helpful votes. This highlights the potential for community managers and community builders to conduct further analysis on these communities. They can investigate the factors that contribute to the high level of collaboration within these communities, either specific video game mechanics that attract like-minded individuals or the effective strategies employed by the community managers of these titles.

4.9 Which video game/genre fans are funnier in reviews?

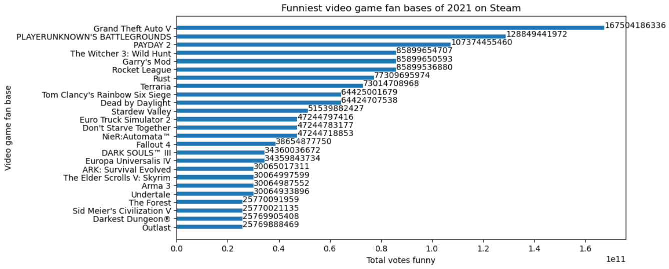

Figure 23. Funniest video game fan bases of 2021 on Steam

Analyzing the results, we observe that the majority of popular titles are associated with communities known for their humor, with a few exceptions like "Nier:Automata" and "Undertale". From this, we can deduce that there appears to be a shared sense of humor among overall users or a prevailing sense of humor that resonates within the most popular video games' communities.

5 Conclusion

In conclusion, big data analytics emerges as a valuable tool for the video gaming industry, offering valuable insights across various dimensions such as design, development, community building, and public relations. Through an exploratory data analysis of a large-scale dataset of video game reviews, I successfully explored the volume and value aspects of big data, enabling the identification of video game trends and preferences in 2021 across different user groups and diverse regions. These findings underscore the significance of carefully selecting target audiences and markets, taking into account their unique distinctions.

Furthermore, my analysis delved into user behavior and revealed a notable tendency for users to write reviews for video games they enjoy. Additionally, an intriguing discovery was made regarding the increased positivity observed for gift-received video games, indicating that offering free content presents a valuable opportunity for enhancing a video game's reputation.

Moreover, I found that more niche titles tend to foster the most helpful communities. Community managers and builders can further delve into studying these titles to gain insights on how to cultivate supportive communities for their own video games. Additionally, with a few exceptions, the majority of popular video games are associated with communities known for their humor, pointing to a prevailing sense of humor resonating within the most popular video games' communities.