Steam Games Analysis

Group no. 5: Davinder Singh, Alan Duran, Alnagdi Mohsen, Socheath Ok

California State University, Stanislaus CS 3520 - Data Visualizations

Professor: Dr. Martin

May 16, 2025

Executive summary

Video games are one of the most popular forms of entertainment today, with player engagement often influenced by factors such as user reviews, game pricing, genre, and game mode. On platforms like Steam, reviews play a key role in shaping public perception and game popularity. What we have found is that moderately priced games tend to perform best, while free games thrive in competitive multiplayer genres. The process taught us to go beyond simple assumptions and look at how multiple factors interact to shape user experience.

Main Research Questions:

How do price, genre, and game mode (single-player vs. multiplayer) influence user reviews of Steam games? Do free games receive significantly different user review scores compared to paid games on Steam? What factors, price, genre, or multiplayer functionality, are most strongly associated with positive user reviews on Steam?

What do the visuals say in the end? The “Top 10 Most Reviewed Steam Games” tells us that free games can reach a large audience; the top-reviewed game on all of Steam, “Counter-Strike” is free. The second thing this graph tells us is that paid games control the reviews, although none of the paid games on the list can match the outstanding reach of Counter-Strike, the remaining games on the list are all paid. This tells us that paid games tend to be more popular, but with the right support and features, free games can still be very successful.

The “Price vs Review Score” tells us that most paid games in the 10-40 dollar range have scores above 7. This is also where the biggest clusters of games are (represented by larger bubbles). This shows that developers have honed in on moderately priced games and can produce games that are often well-received. There is also an evident downward trend in review scores as the price of games increases, This leads us to believe that as the price increases, so do the expectations of the consumers.

The “Total Positive Reviews by Genres: Free vs Paid Games” provides some key insights. Firstly, it can be seen that in just about every genre, paid games receive more positive reviews. This indicates a preference for paid games across the entire Steam marketplace. Whether this preference is fueled by better reception or the games just being more popular. This graph also shows that free games have the best chance of succeeding in the shooter and BR(Battle Royale) genres. This shows that games with repetitive competitive environments based around multiplayer game modes are particularly successful when free.

Background information and summary of the data

The data that was used in this project came from Kaggle. Kaggle has several saved Steam datasets that can be found online on their website. The data is structured in Excel format, which can be converted to comma-separated values. The data is in its original form, meant to be used to compile all of the Steam review games data for a project such as this one. In its raw form, there is not much that can be interpreted from the data. The original data had 100,000+ observations of 21 columns. Only 14 of those columns were chosen as relevant to this project.

Data and Cleaning

The data was cleaned by first filtering to only the top 50 most reviewed games on free games. From there, the data was looked through to see if there were any errors in the data, such as duplicates, missing datapoints, etc. The data was then saved as a comma-separated value file. This process was repeated for paid data as well. The data

was then uploaded into R and renamed as 2 separate data frames with the names “Free_Data” and “Paid_Data”. The final result was 49 observations of 14 columns of data for the top 50 most reviewed free games, and 49 observations of 14 columns of data for the top 50 most reviewed paid games. Once in separate data frames, a combined data frame was made, which resulted in 98 observations of 14 columns of data. The columns are as follows: Name, Developers, Publishers, Multiplayer?, Genres, Total reviews, Total positive reviews, Total negative reviews, Review score (average), Review score desc, Positive percentual, Is_free, Price(USD), and Type(free or paid). The visuals we then created were those that we found to be the most useful and insightful regarding discovering patterns within the data. Many visuals were made with a variety of types, but the 3 final visuals are those discussed in this report. Visuals can be referenced in the appendix of this report.

Explanation of design choices

We chose a grouped bar chart for the “Total Positive Reviews by Genres: Free vs Paid Games” because it makes it easy to compare free vs paid games within each genre. The height of the bars shows clearly which has more positive reviews. It works well for comparing categories. We found that this graph upheld the characteristics of being truthful, aesthetically pleasing, and insightful for the following reasons: -Truthful - The graph shows the raw totals of positive reviews. -Aesthetically Pleasing - The graph looks clean, uses good color combinations, and space. -Insightful - It reveals key genre differences like action and RPG games.

For “Top 10 most Reviewed Steam Games” we chose a horizontal bar chart because: game titles are long, and horizontal bars keep the text readable. It’s easy to compare games by review count. This visual clearly shows the most popular games based on total positive reviews. This visual upheld the following characteristics: -Truthful - The graph shows the actual totals of positive and negative reviews. -Aesthetically Pleasing - Clean layout with horizontal bars for readability. -Insightful - Highlights which games attract the most attention from players with positive reviews. -Elegant - It’s simple and focused.

A scatter plot was selected for the Price Vs Review Score graph because it helps show the relationship between two continuous variables, price and review score. The graph upheld the following characteristics: -Truthful - Displays raw data without filtering -Aesthetically Pleasing - Simple colours and layout help to focus on the pattern. -Elegant - Clear and simple.

Methodology

We collected data from Steam on a sample of games, focusing on user review scores, price, free or paid, genre, and multiplayer or single-player features. We then analyzed average review ratings across these groups. We used descriptive statistics such as average positive reviews and grouping by price brackets to visualize the results with charts and graphs to see if we could identify patterns and correlations between game type and user reviews.

Finding your story

Our initial assumption was that price alone determined review scores; however, this didn’t hold up. Both free and paid games had highly reviewed and poorly reviewed titles, so we realized we needed to consider other factors like genre and multiplayer mode. We also tried using the entire Steam dataset at first, but it was too large and messy, with duplicates and missing data. This led us to focus on a cleaner, more manageable subset: “the top 50 most reviewed free and paid games”. Some visualization attempts didn’t work out either, those being pie charts and stacked bars, which made genre comparisons confusing. We eventually found that grouped bar charts were the most effective way to show differences between free and paid games by genre. We also considered sentiment analysis of user reviews, but didn’t pursue it due to time and tool limitations. Instead, we focused on clear, quantitative variables. In the end, our most meaningful insight was that moderately priced games tend to perform best, while free games thrive in competitive multiplayer genres. The process taught us to go beyond simple assumptions and look at how multiple factors interact to shape user experience.

Conclusion

This study examined how factors like price, genre, and game mode influence user reviews on Steam, and our findings show that paid games, especially those priced between $5 and $20, tend to receive more reviews and higher positive ratings than free games. However, price alone doesn’t guarantee better reviews, as highly rated free games often succeed due to strong multiplayer features and active communities.Overall, user satisfaction appears to be driven more by gameplay experience and social features than by cost. These insights can help developers and players better understand what makes a game well-received on platforms like Steam.

References

Kaggle: Steam Games Dataset 2025. https://www.kaggle.com/datasets/srgiomanhes/steam-games-dataset-2025 (https://www.kaggle.com/datasets/srgiomanhes/steam-games-dataset-2025)

Appendix is on the following page

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readr)
 library(dplyr)
Free_Data <- read_csv("Steam Games Free_Data (1).csv")
```

```
## Rows: 49 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): Name, Developers, Publishers, Multiplayer?, Genres, Review score desc
## dbl (6): Total reviews, Total positive, Total negative, Review score, positi...
## lgl (1): is_free
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
 Paid_Data <- read_csv("Steam Games Paid_Data (1).csv")
```

```
## Rows: 49 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): Name, Developers, Publishers, Multiplayer?, Genres, Review score desc
## dbl (6): Total reviews, Total positive, Total negative, Review score, positi...
## lgl (1): is_free
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Free_Data$type <- "Free"
 Paid_Data$type <- "Paid"
```

```
colnames(Free_Data)
```

```
##  [1] "Name"                "Developers"          "Publishers"
##  [4] "Multiplayer?"        "Genres"              "Total reviews"
##  [7] "Total positive"      "Total negative"      "Review score"
## [10] "Review score desc"   "positive percentual" "is_free"
## [13] "Price (USD)"         "type"
```

```
colnames(Paid_Data)
```

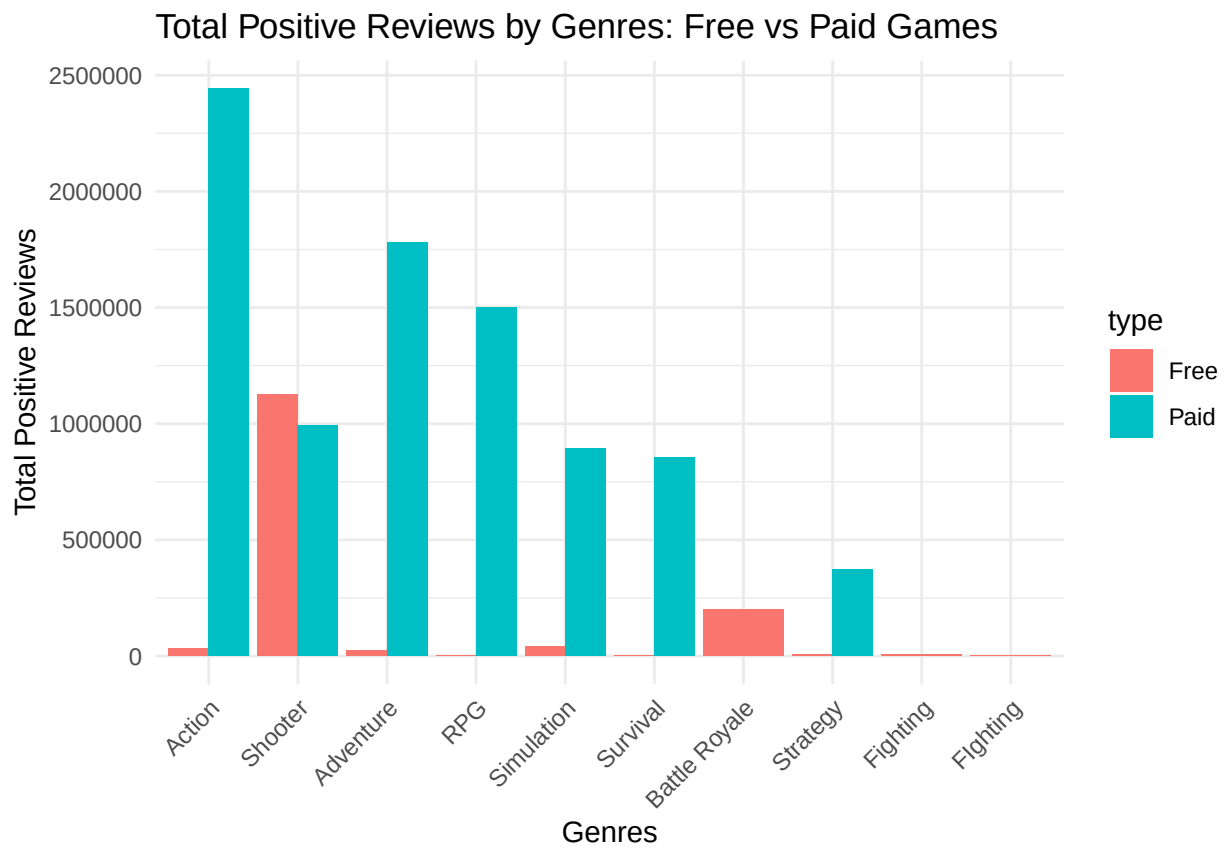
```
##  [1] "Name"                "Developers"          "Publishers"
##  [4] "Multiplayer?"        "Genres"              "Total reviews"
##  [7] "Total positive"      "Total negative"      "Review score"
## [10] "Review score desc"   "positive percentual" "is_free"
## [13] "Price (USD)"         "type"
```

```r
combined_data <- bind_rows(
  Free_Data %>% select(Name, Genres, `Total positive`, type),
  Paid_Data %>% select(Name, Genres, `Total positive`, type)
)
```

```r
colnames(combined_data)[3] <- "Total_Positive"
```

```r
genre_summary <- combined_data %>%
  group_by(Genres, type) %>%
  summarise(Total_Positive = sum(Total_Positive, na.rm = TRUE), .groups = "drop")
```

```r
ggplot(genre_summary, aes(x = reorder(Genres, -Total_Positive), y = Total_Positive, fill = type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Total Positive Reviews by Genres: Free vs Paid Games",
    x = "Genres",
    y = "Total Positive Reviews"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
Combined_Data <- bind_rows(Free_Data, Paid_Data)
```

```r
colnames(Combined_Data)
```

```
##  [1] "Name"                "Developers"          "Publishers"
##  [4] "Multiplayer?"        "Genres"              "Total reviews"
##  [7] "Total positive"      "Total negative"      "Review score"
```

```
## [10] "Review score desc"   "positive percentual" "is_free"
## [13] "Price (USD)"         "type"
```

```
top_10_games <- Combined_Data %>%
  arrange(desc(`Total reviews`)) %>%
  slice_head(n = 10) %>%
  select(Name, `Total positive`, `Total negative`, type)
```

```
top_10_long <- top_10_games %>%
  pivot_longer(cols = c(`Total positive`, `Total negative`),
               names_to = "Review_Type",
               values_to = "Count")
```

```
library(scales)
```

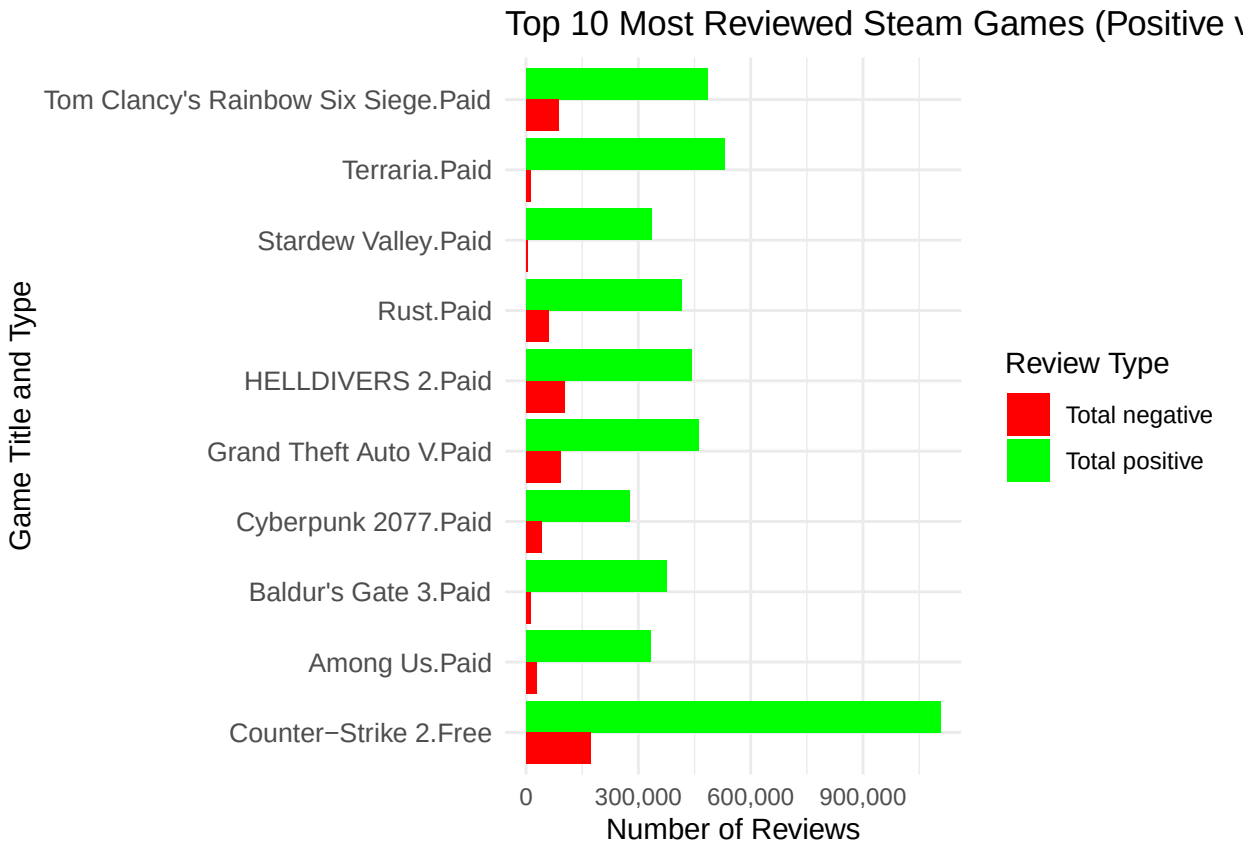
```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

Graph - Top 10 Most Reviewed Steam Games

```
ggplot(top_10_long, aes(x = interaction(Name, type), y = Count, fill = Review_Type)) +
  geom_col(position = "dodge") +
  coord_flip() +
  scale_y_continuous(labels = scales::comma) +
  labs(
    title = "Top 10 Most Reviewed Steam Games (Positive vs Negative by Type)",
    x = "Game Title and Type",
    y = "Number of Reviews",
    fill = "Review Type"
  ) +
  scale_fill_manual(values = c("Total positive" = "green", "Total negative" = "red")) +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10))
```



Graph- Price vs Review Score

```
Paid_Data_counted <- Paid_Data %>%
  count(`Price (USD)`, `Review score`, type, name = "n")


ggplot(Paid_Data_counted, aes(x = `Price (USD)`, y = `Review score`)) +
  geom_point(aes(color = type, shape = type, size = n),
             fill = "black",  # black dot fill
             stroke = 0.8,     # border thickness
             alpha = 0.8) +
  geom_smooth(method = "lm", se = TRUE, aes(group = 1),
              color = "black", linetype = "dashed") +
  labs(
    title = "Price vs Review Score",
    x = "Price (USD)",
    y = "Review Score",
    color = "Type",
    shape = "Type",
    size = "n"   # Legend for black dot sizes
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


```
## `geom_smooth()` using formula = 'y ~ x'
```