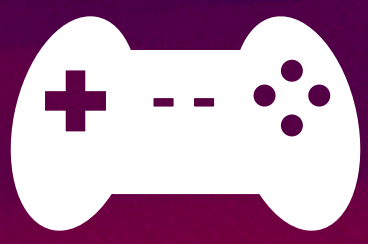


Data analysis lab/experiment & discovery

# Steam games data report

*steamgames.csv*

-Peter Van Beever

Student data engineer

**Outline:**

**Intro**

**Methodology**

**Analysis and discovery**

**Graphs/Tables/Visuals**

**Conclusion**

# Intro

## Data Source

The provided `steamgames.csv` file was supplied by an educational instructor as part of a data analysis assignment. To conduct the analysis, JupyterLab was utilized within a browser environment, allowing for an interactive exploration of the dataset. Custom charts and visualizations were created with the assistance of ChatGPT, which helped in tailoring the visual representation of the data to meet specific analytical needs.

## First attempt at complex data

`steamgames.csv` dataset was selected after a thorough evaluation of the complexities associated with presidential polling data, which can often be intricate and challenging to interpret. The decision to use the `steamgames.csv` reflects a strategic choice to start with a more straightforward dataset, allowing for a focused and manageable introduction to data analysis techniques. This report is the first by the author, to build upon their understanding of more complex data sets in future projects.

# Methodology

## Approach and references

The goal was to explore the dataset with a creative and innovative approach to uncover insights. Leveraging familiarity with JupyterLab and Python with Pandas for data manipulation, as well as experience with MS Charticulator and preliminary trials with Tableau, Chart.js, and other charting software, the focus was on building dynamic and interactive visualizations rather than coding static charts. This approach aimed to engage with the data more deeply, using advanced features and tools to facilitate a more exploratory analysis.

## Quantizing and “stepping” the data

Standard charts were utilized in conjunction with more specialized visualizations to provide a comprehensive view of the data. Logarithmic scales were chosen to handle data with a wide range of values, ensuring that smaller variations were visible alongside larger ones. Binning was employed to group data points, simplifying analysis and revealing trends. Scatter plots were used to explore relationships between variables, while Sankey diagrams illustrated flow and connections. Rollover effects were implemented to enhance human interactivity, enabling users to uncover deeper insights through visual "aha" moments. This multifaceted approach aimed to maximize the effectiveness of experimentation.

## Question for analysis

The analysis question was all about discovery—just poking around to see what interesting patterns or insights might pop up. The goal was to explore the data with an open mind and see what stood out.

# Process and calculations

**Initial Exploration**: I began by examining the dataset, starting with the total number of rows. I then viewed the first few rows (head) and the last few rows (tail) to get an initial sense of the data's structure.

**Column Analysis**: I reviewed the column titles and found a total of 50,751 unique game titles. Similarly, I counted the unique ratings and other relevant fields.

**Release Trends**: I plotted a simple bar chart to visualize game release dates. This revealed a steady but low number of releases before 2010, a noticeable spike around 2015, and a sharp increase in games up to 2020.

**Review Analysis**: I examined the types of user reviews, which totaled 92,812,148. A bar chart illustrated that most reviews were positive.

**Platform Distribution**: A bar chart was also used to analyze the platforms, showing a higher number of Windows-based games compared to Mac or Linux.

**Price Examination**: I explored game prices and discovered a significant gap: most games were free, while a few were priced between $200 and $300. To better understand this, I used a scatter plot to examine the correlation between the positive review ratio (scaled from 1-100) and the number of reviews. Due to the outlier distribution of reviews, I applied a logarithmic scale for better visualization.

**Further Exploration**: With these insights, I proceeded to explore multiple views of the data to uncover additional patterns and insights.

# Head and Tail of the data set

```
df.head(3)
```

| | app_id | title | date_release | win | mac | linux | rating | positive_ratio | user_reviews | price_final | price_original | discount | steam_deck |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 13500 | Prince of Persia: Warrior Within™ | 2008-11-21 | True | False | False | Very Positive | 84 | 2199 | 9.99 | 9.99 | 0.0 | True |
| **1** | 22364 | BRINK: Agents of Change | 2011-08-03 | True | False | False | Positive | 85 | 21 | 2.99 | 2.99 | 0.0 | True |
| **2** | 113020 | Monaco: What's Yours Is Mine | 2013-04-24 | True | True | True | Very Positive | 92 | 3722 | 14.99 | 14.99 | 0.0 | True |

```
df.tail(3)
```

| | app_id | title | date_release | win | mac | linux | rating | positive_ratio | user_reviews | price_final | price_original | discount | steam_deck |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **50869** | 1402110 | Eternights | 2023-09-11 | True | False | False | Very Positive | 89 | 1128 | 30.0 | 0.0 | 0.0 | True |
| **50870** | 2272250 | Forgive Me Father 2 | 2023-10-19 | True | False | False | Very Positive | 95 | 82 | 17.0 | 0.0 | 0.0 | True |
| **50871** | 2488510 | FatalZone | 2023-10-23 | True | False | False | Very Positive | 88 | 144 | 4.0 | 0.0 | 0.0 | True |

# Mean, median, mode, summarized

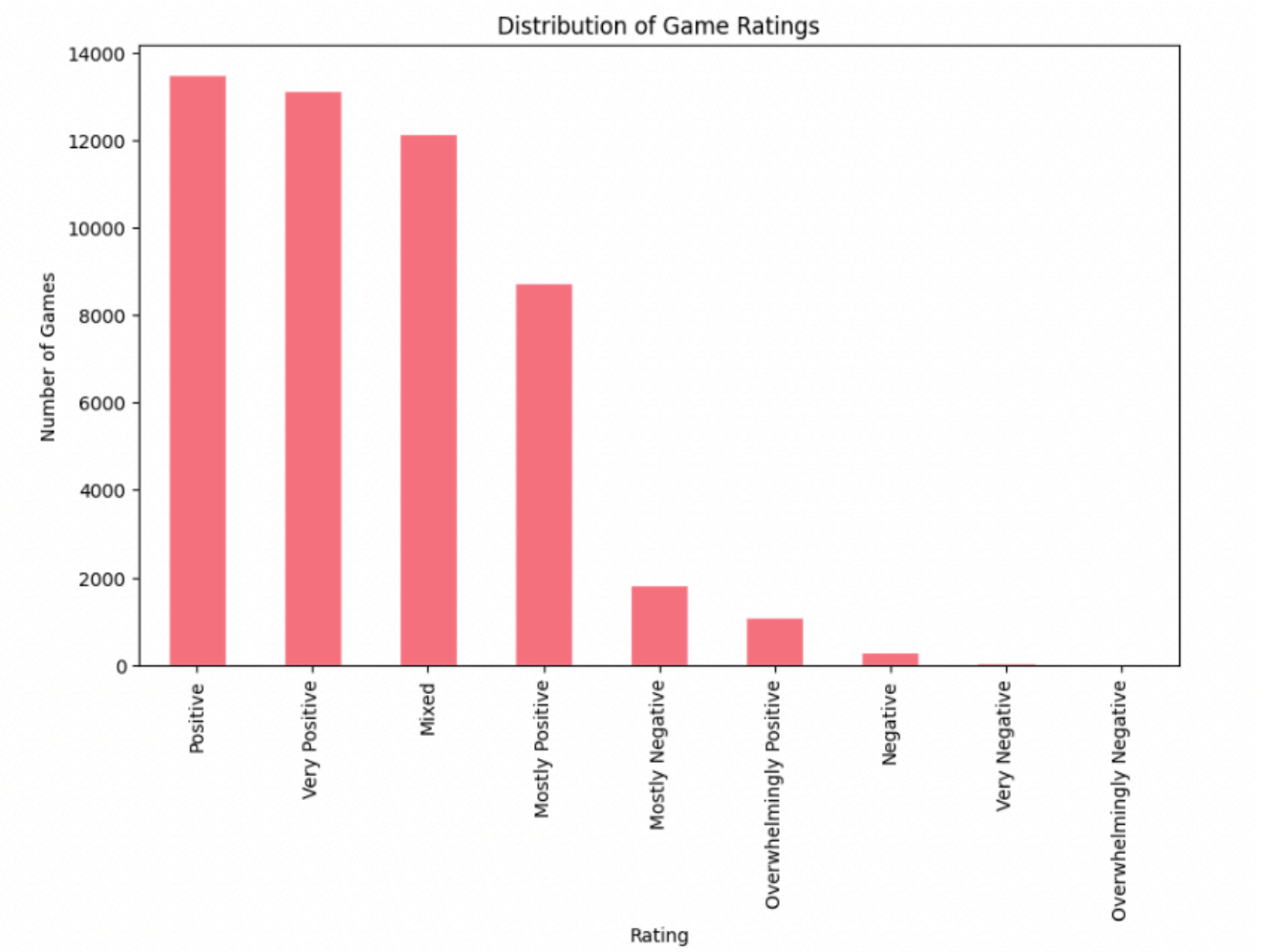
| | app_id | positive_ratio | user_reviews | price_final | price_original | discount | release_year_bin | price_bin | positive_ratio_bin |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 50872.00 | 50872.00 | 50872.00 | 50872.00 | 50872.00 | 50872.00 | 50872.00 | 50872.00 | 50872.00 |
| **mean** | 1055223.81 | 77.05 | 1824.42 | 8.62 | 8.73 | 5.59 | 3.63 | 0.00 | 2.50 |
| **std** | 610324.95 | 18.25 | 40073.52 | 11.51 | 11.51 | 18.61 | 0.57 | 0.06 | 0.71 |
| **min** | 10.00 | 0.00 | 10.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **25%** | 528737.50 | 67.00 | 19.00 | 0.99 | 0.99 | 0.00 | 3.00 | 0.00 | 2.00 |
| **50%** | 986085.00 | 81.00 | 49.00 | 4.99 | 4.99 | 0.00 | 4.00 | 0.00 | 3.00 |
| **75%** | 1524895.00 | 91.00 | 206.00 | 10.99 | 11.99 | 0.00 | 4.00 | 0.00 | 3.00 |
| **max** | 2599300.00 | 100.00 | 7494460.00 | 299.99 | 299.99 | 90.00 | 4.00 | 3.00 | 3.00 |

## Discovery 1

Price:
The majority of game prices averaged $8.62.
The lowest price of a game was FREE.
The highest price of a game was $299.99

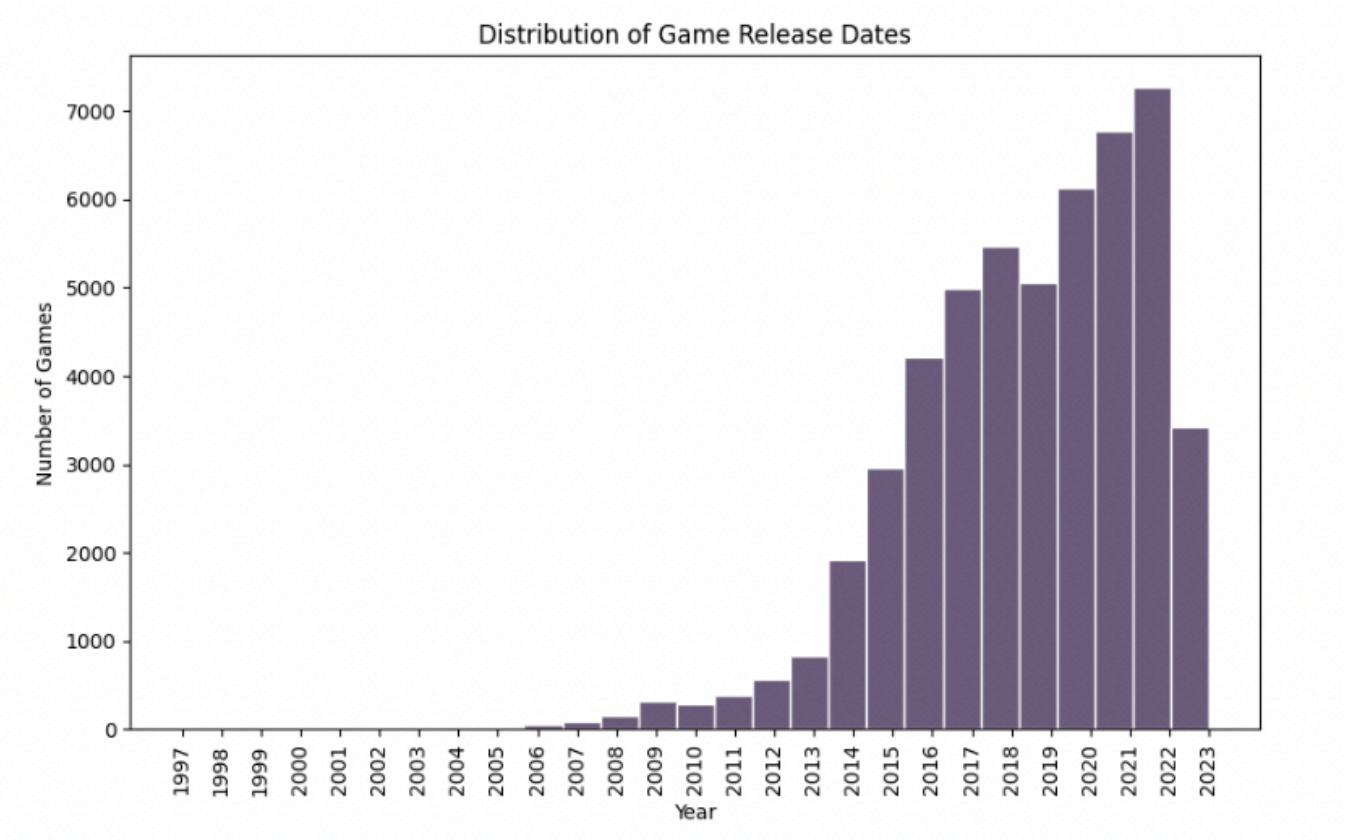
## Discovery 2

Positive reception and reviews:
The majority of game reception was 77% positive.
The average number of reviews per game was approximately 1,825 reviews.



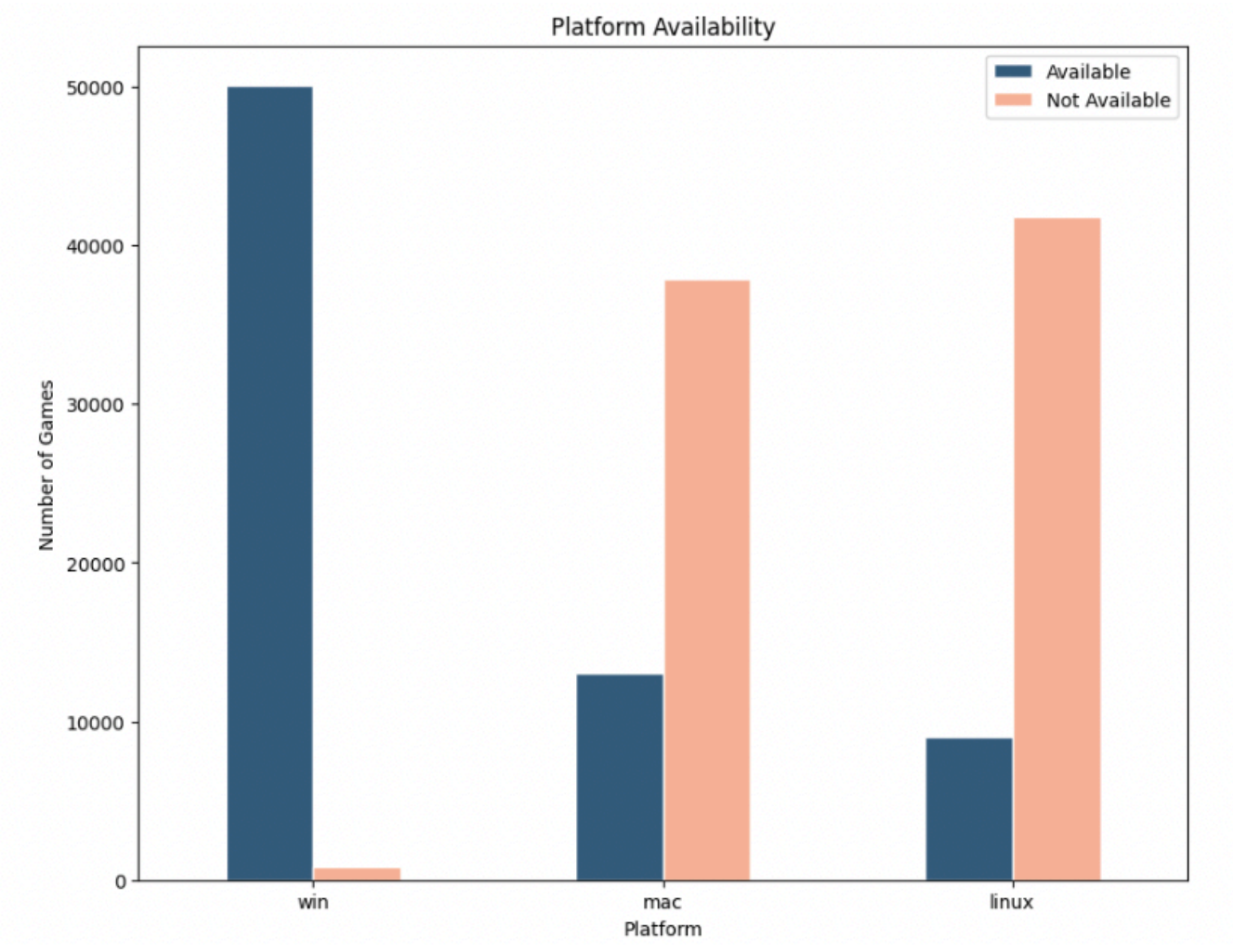

## Discovery 3

Release date:
The majority of games were released between 2014-2022



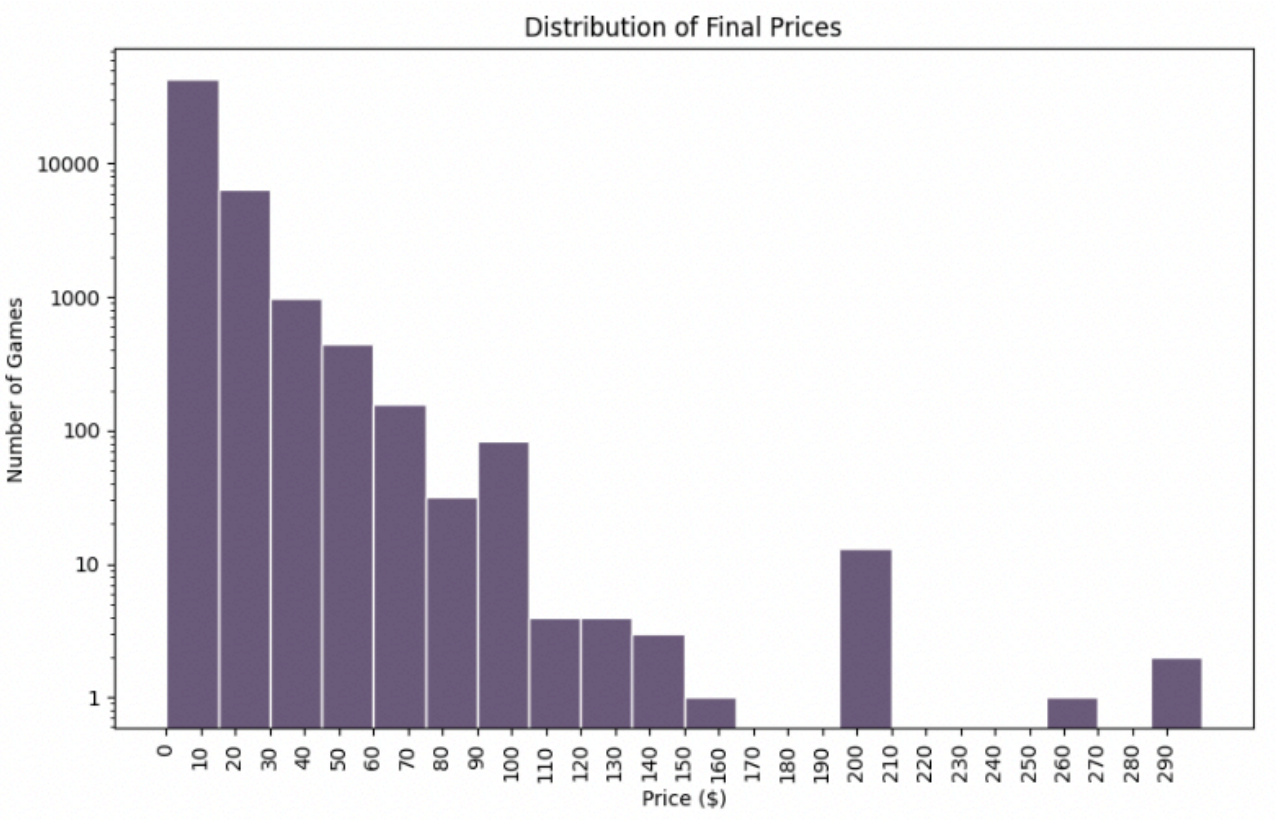
# Discovery 4

Platform availability:
The majority of games are not available for Mac or Linux.
Windows platform has the most availability.




# Discovery 5 using *log* scale vs linear

Price:
Due to the gap in prices, a log scale was used to visually see the other prices of games instead of seeing only the majority of prices (Free)

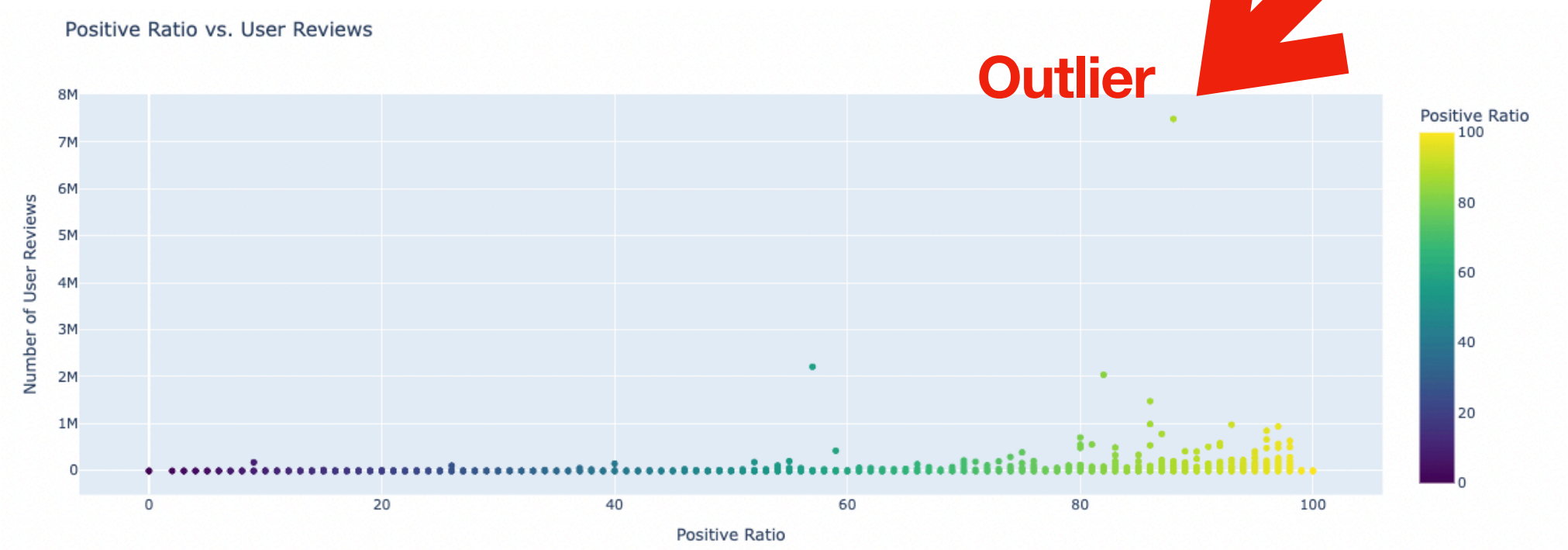


# Visual Exploration Part 1

**Correlation**:

Is there correlation between number of reviews and and positive reviews?

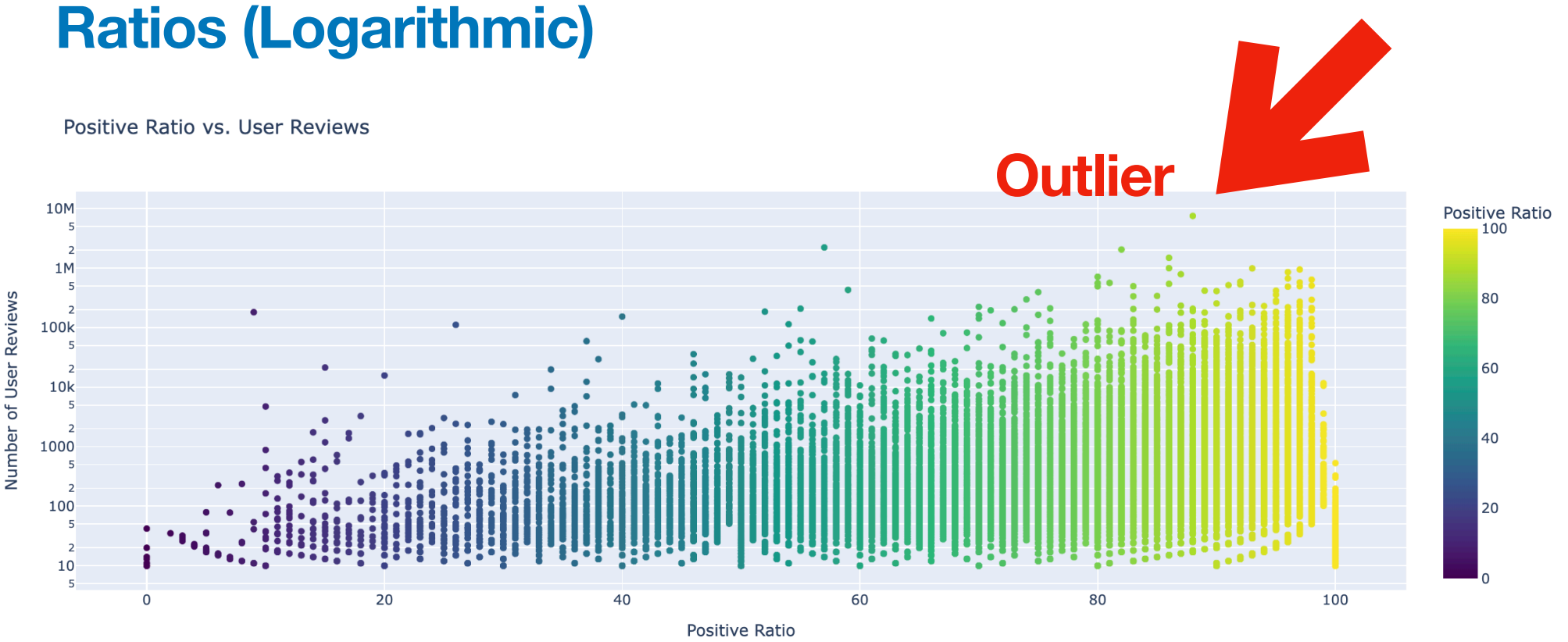
## Linear




**Logarithmic Scale**:

By using a logarithmic scale, the differences between lower review counts become more apparent, and the distribution of review counts is more evenly represented, allowing for better analysis of review patterns.

## Ratios (Logarithmic)



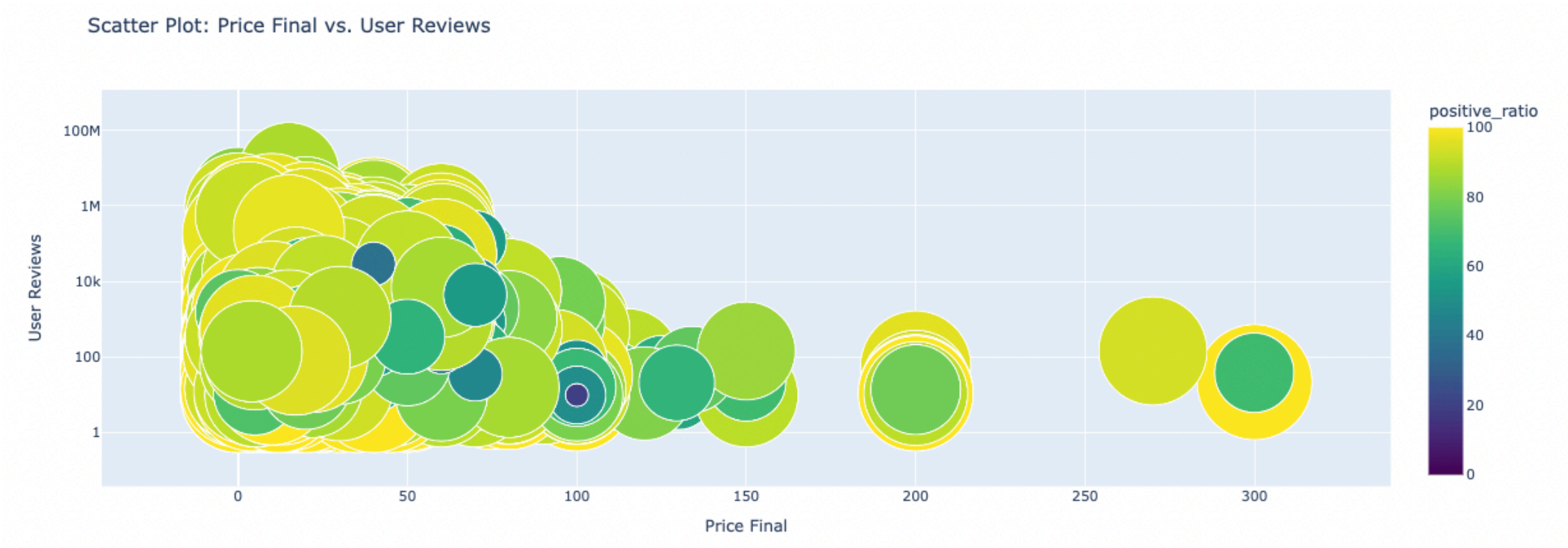

**Correlation**:

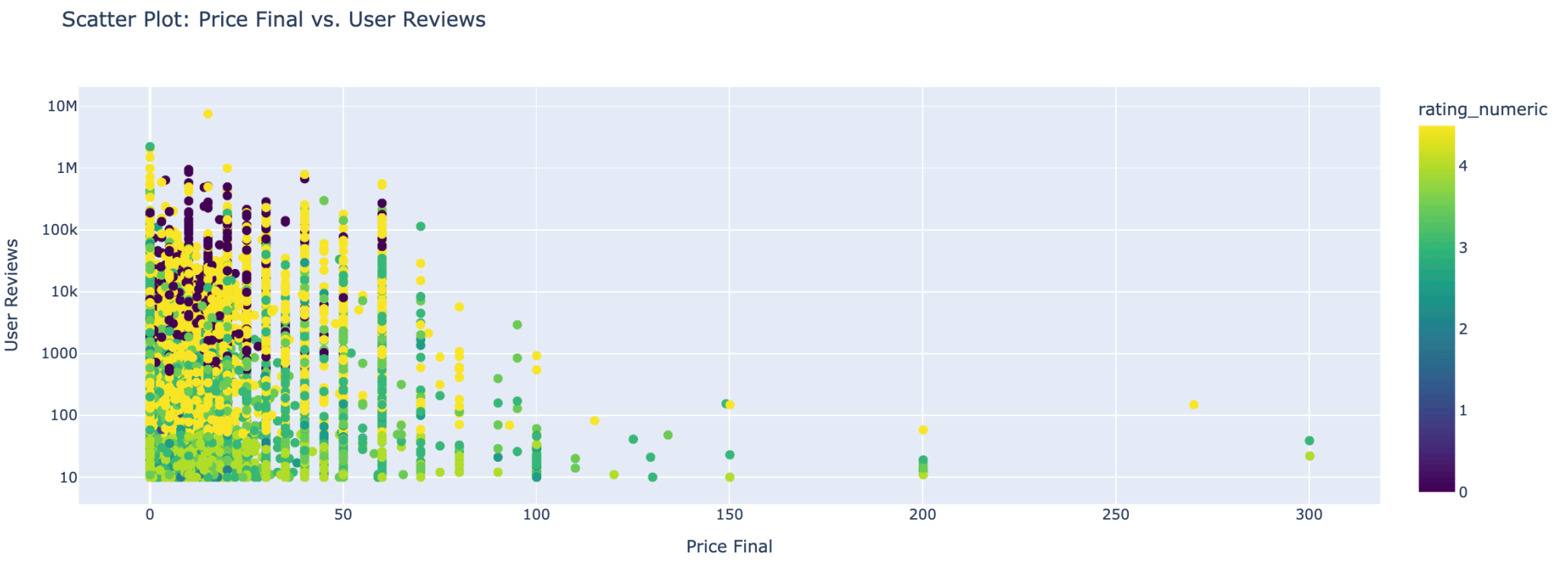
It seems the number of reviews increases towards positive reviews.

# Visual Exploration Part 2

**Marker Size**:

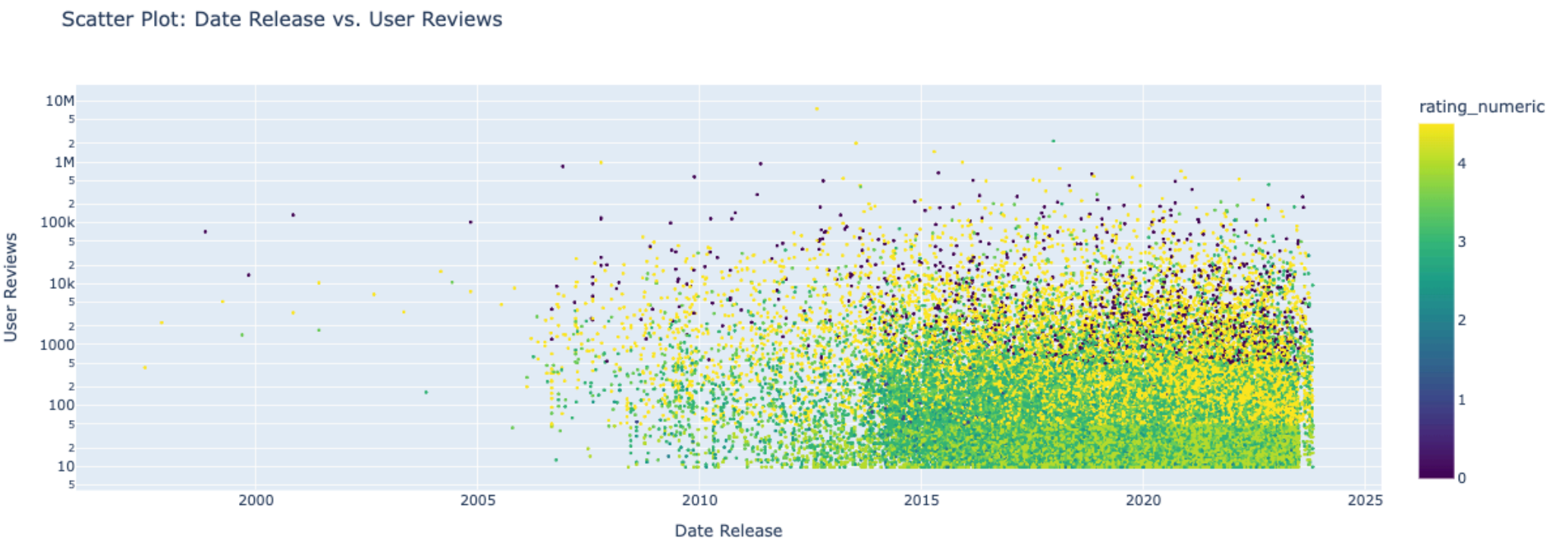
**Description**: The size of the markers are scaled according to a specific column's values, in this case they are set to the Positive Ratio (of reviews). Larger markers represent higher values, while smaller markers indicate lower values.
**Chart**: The scatter plot shows positive reviews as large circles, small circles indicate a lower score of review. It appears that a higher price game does not get negative reviews but do get less reviews due to lower number of players.



**Controls**:

**Description:** This chart demonstrated the various views of comparing user reviews with year release, indicating reviews increased after 2014. The quality of positive review % shows (by color) what types of reviews there are.
It appears there are always “peppered” negative reviews or low reviews for games from 200 to 2025.

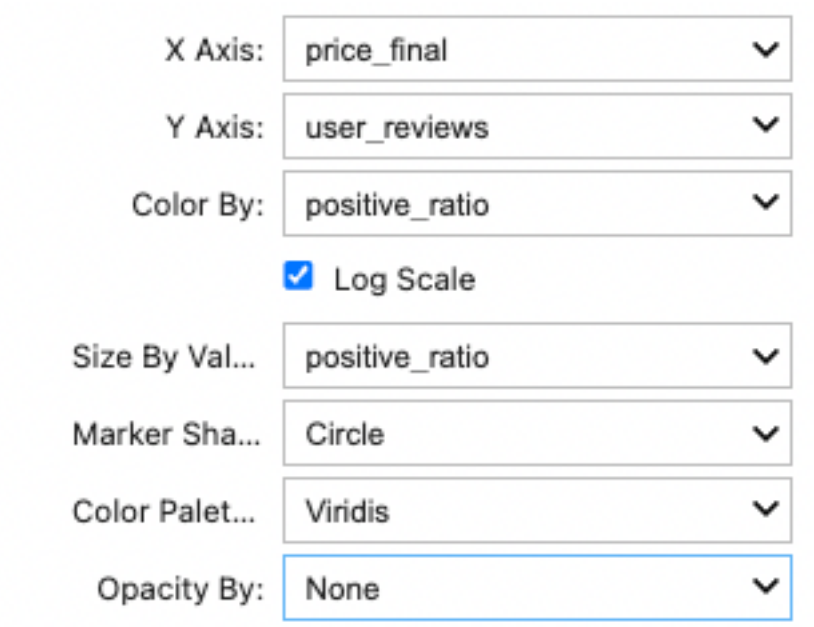



## Further Questions for exploration

An observation is whether the positive reviews are allowed to accumulate to increase exposure to games, and are negative reviews filtered for an unknown reason such as profanity, usefulness or critique, or by the game makers themselves.

In other industries, a service is allowed to dispute a negative review from a non-client.

The user controls implemented for exploration.



**Binning**:

**Description**: Binning groups continuous variables into discrete ranges. In the Sankey diagram, release years, prices, and positive ratios are binned into defined intervals to simplify the data and reveal patterns.
**Example**: Release years are divided into bins such as "2010-2012" and "2013-2015". This transforms the continuous release year data into discrete categories, which helps in tracking trends across different periods.

**Sankey Diagram Flow Mapping**:

**Description**: The Sankey diagram visualizes flow between different categories (nodes) using links. Each link represents the flow of user reviews between categories, such as between release years, operating systems, price ranges, and positive ratio bins.

**Example**: A link from the "2010-2012" year bin to the "Windows" node, and then to a specific price range and positive ratio bin, shows how user reviews flow through these categories. The width of each link corresponds to the volume of reviews.

**Normalization for Color Mapping**:

**Description**: Link values (user reviews) are normalized to map them onto a color scale. This ensures that colors represent the magnitude of the flows relative to the minimum and maximum values.
**Example**: If the range of user reviews is between 100 and 10,000, links with higher review counts will be assigned colors that indicate greater magnitude, while those with lower counts will use colors indicating lesser values.

**Bar Intensity Adjustment**:

**Description**: The thickness of the Sankey diagram bars can be adjusted to enhance visual clarity. This affects how prominently different flows are displayed.
**Example**: Increasing the bar intensity makes the flows more pronounced, making it easier to compare the volume of reviews between different categories.

# Visual Exploration Part 3

Sankey Diagram and flows of reviews.

Year Bins: 5

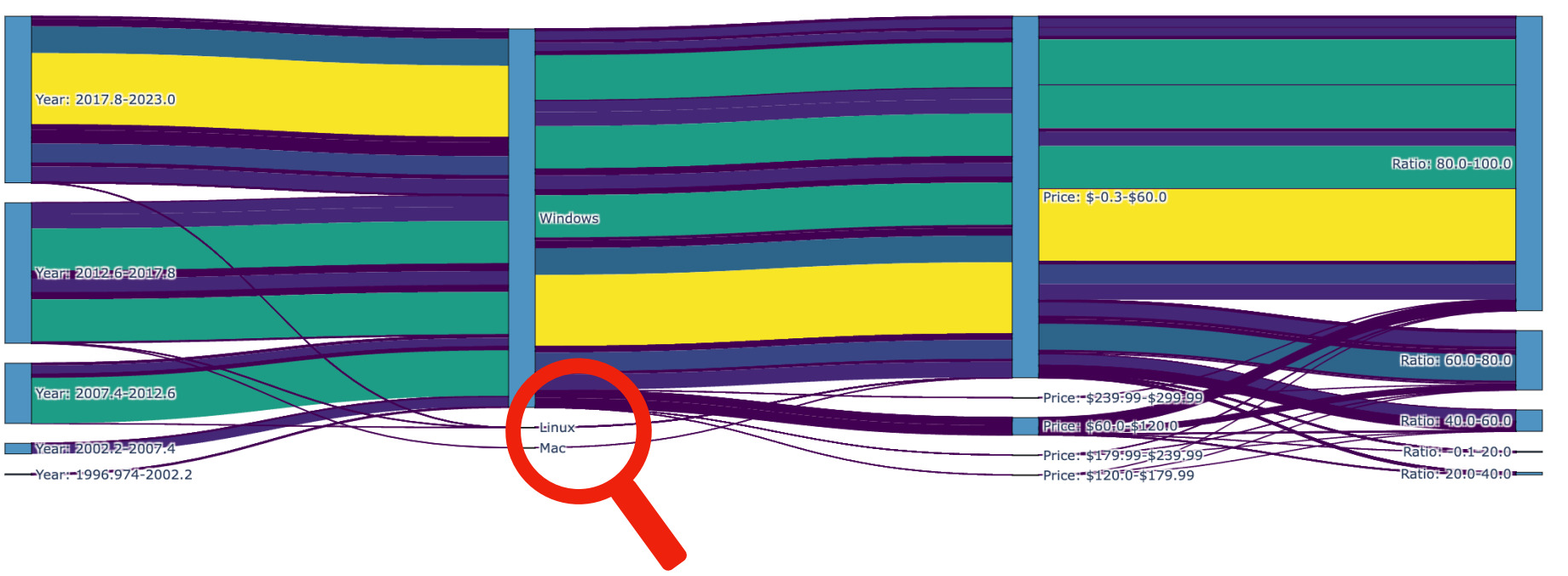
Price Bins: 5

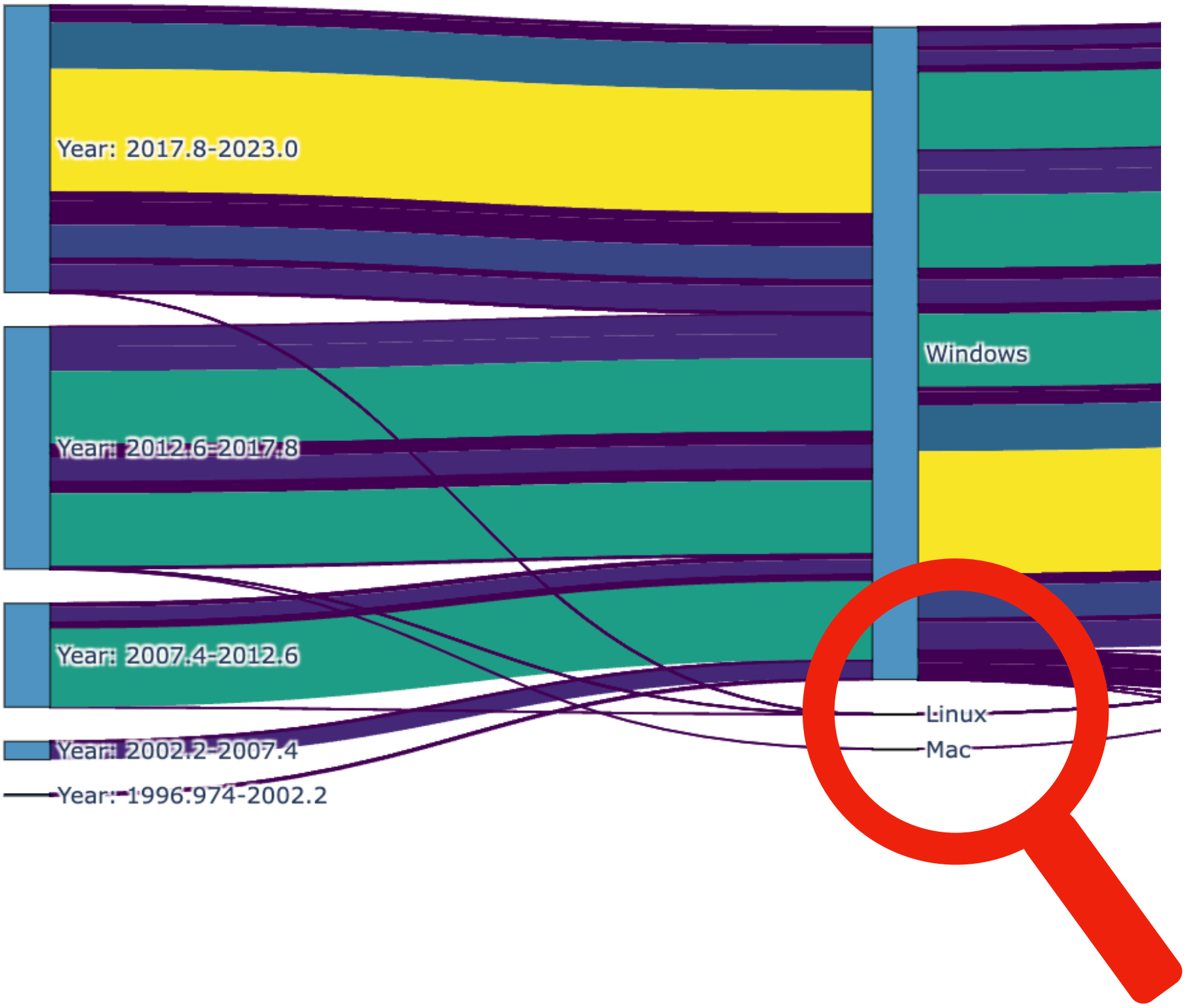
Ratio Bins: 5

Color Scale: Viridis

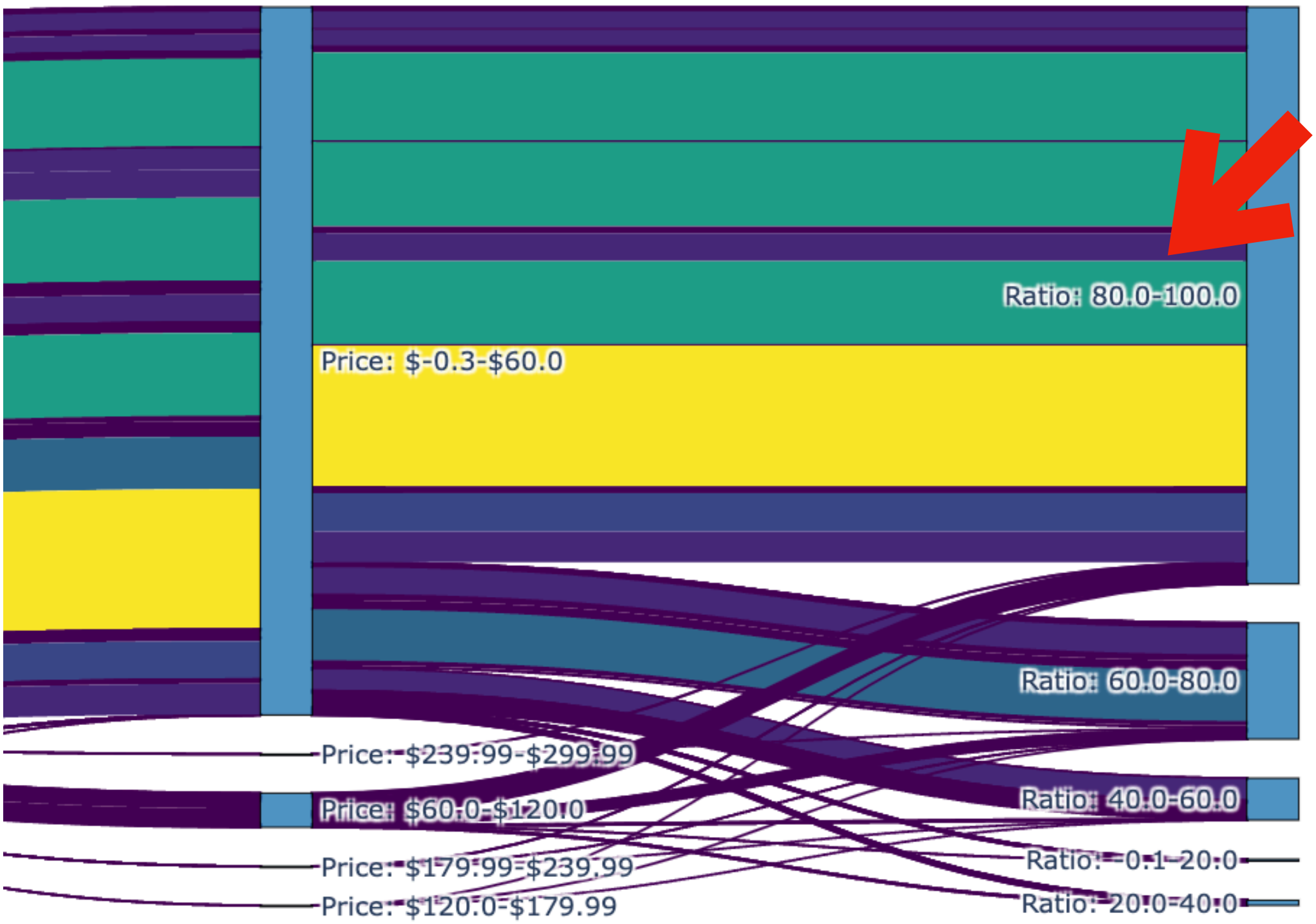
Bar Intensity: 1

Sankey Diagram



# Positive Ratio Reviews 100% = Best



# Conclusion

Standard charts were utilized in conjunction with more specialized visualizations to provide a comprehensive view of the data. Logarithmic scales were chosen to handle data with a wide range of values, ensuring that smaller variations were visible alongside larger ones. Binning was employed to group data points, simplifying analysis and revealing trends. Scatter plots were used to explore relationships between variables, while Sankey diagrams illustrated flow and connections. Rollover effects were implemented to enhance human interactivity, enabling users to uncover deeper insights through visual "aha" moments. This multifaceted approach aimed to maximize the effectiveness of experimentation.

The analysis of the Steam games dataset revealed several interesting insights into game release trends, pricing, and user reviews. The application of binning techniques allowed for a clearer examination of patterns over time and across different price ranges and rating ratios. The Sankey diagram effectively illustrated the flow of user reviews through various categories, highlighting how reviews are distributed across different release years, operating systems, and pricing bins. ChatGPT was instrumental in assisting with the creation of interactive dropdown menus and in fine-tuning the visualizations, particularly in crafting an experience similar to Tableau or MS Charticulator.

This prompt engineering approach enhanced the interactivity and depth of the data analysis, providing a comprehensive view of the dataset and uncovering significant patterns within the gaming industry. The use of advanced visualizations and interactive features allowed revealing key trends and helping to explore the dynamics of game releases and user engagement.