Steam Game User Rating Analysis

Nguyễn Hoàng Phúc

Introduction to Data Science

Project Objectives

- Analyze user ratings of Steam games to understand trends and influential factors.
- Derive meaningful insights to benefit developers and understand the Steam market.

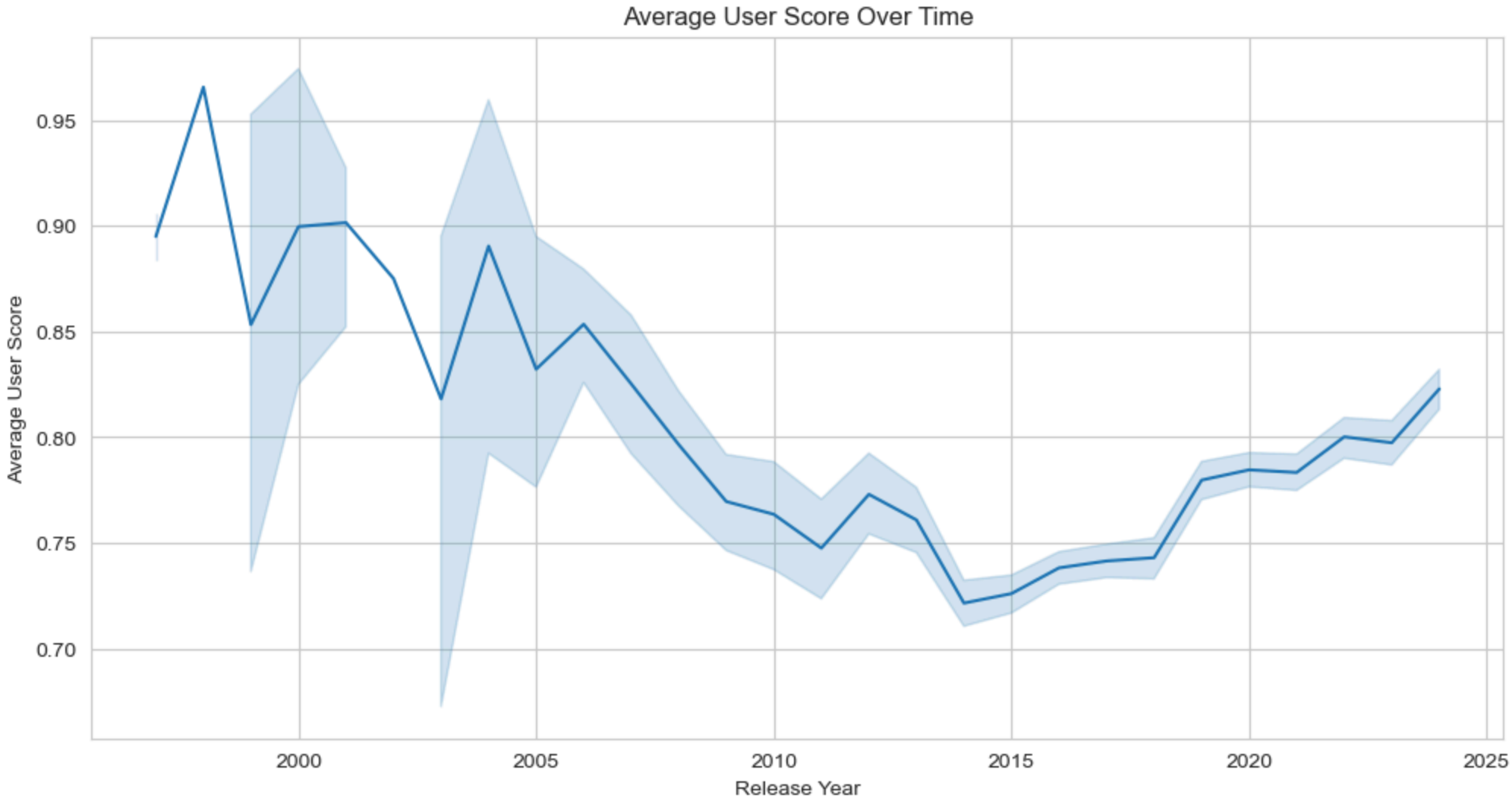
Data Collection & Preprocessing

- Collected data from SteamSpy and Steam API.
- Merged and cleaned the datasets.
- Engineered features by converting genre information into binary features using one-hot encoding; normalize concurrent user and other skewed features through logarithmic transformation.
- Final dataset: 18137 games with 15 features (22 including engineered features).

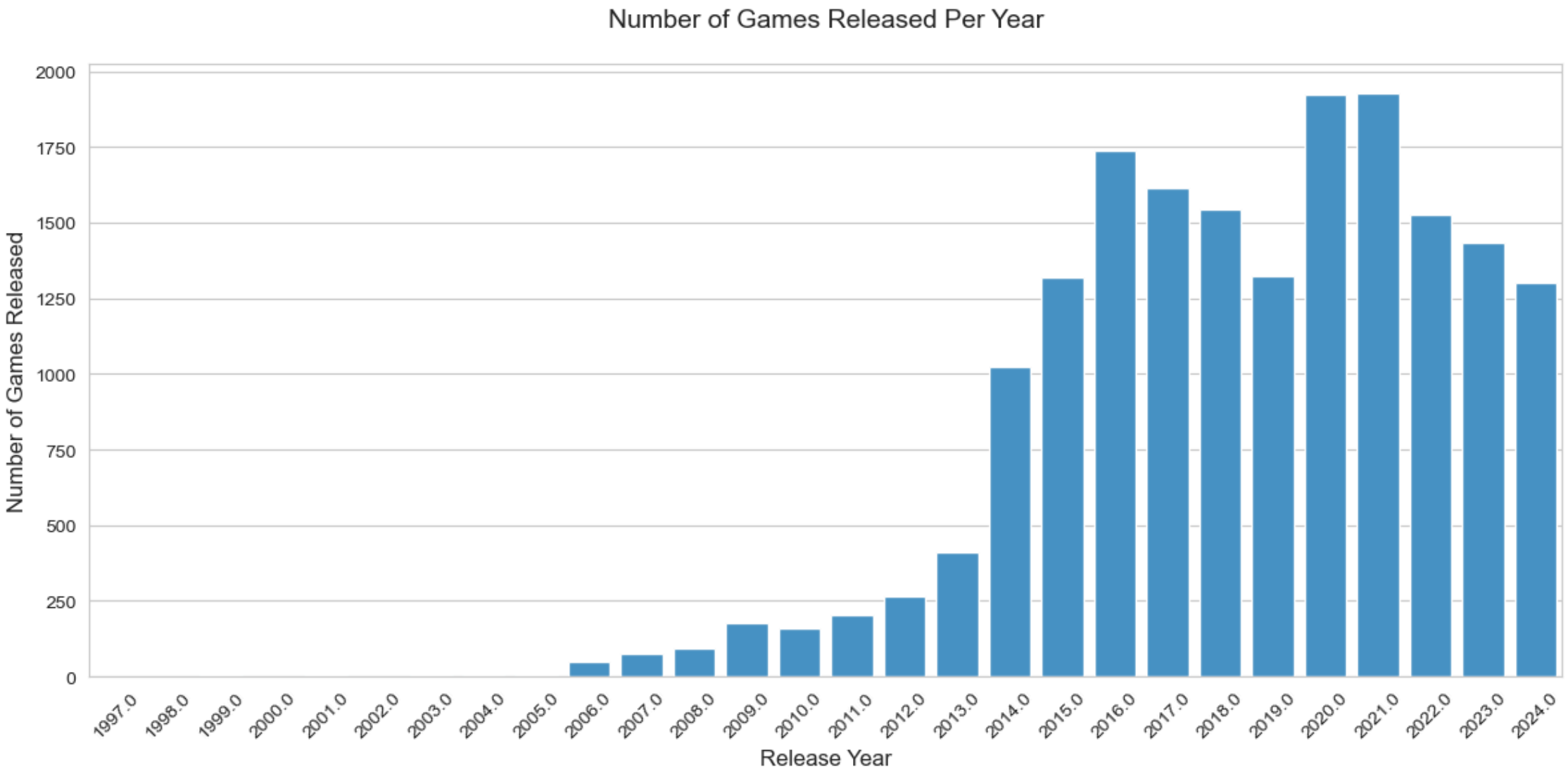
Data Analysis & Modeling

Q1: How has user reception evolved over time?

Q1: How has user reception evolved over time?



Q1: How has user reception evolved over time?



Q1: How has user reception evolved over time?

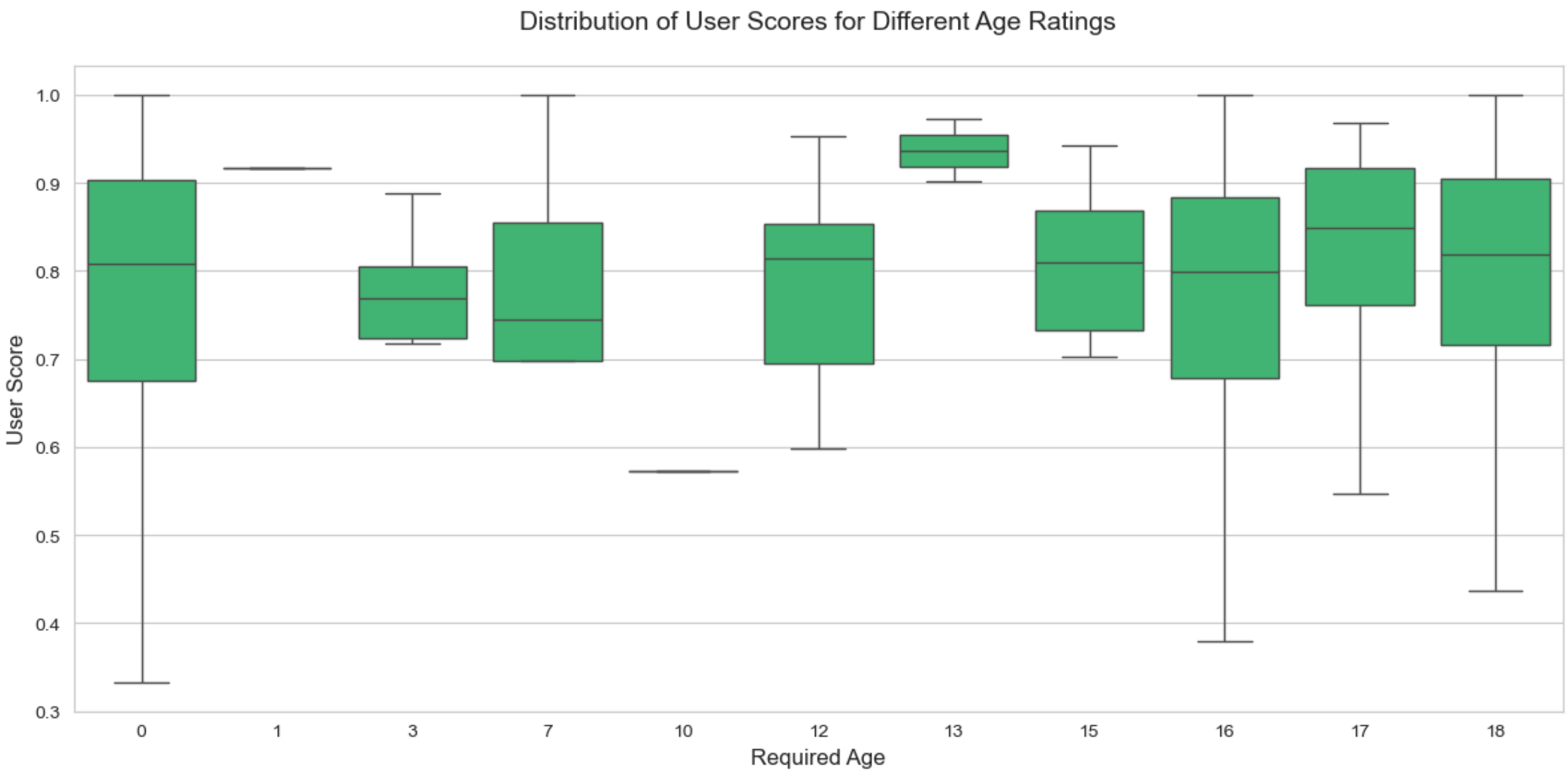
- **Findings:**
- Average user scores show a downward trend from 2000 to 2024.
- The number of games released has significantly increased.
- Negative correlation (-0.59) between the number of releases and average user score.

Q1: How has user reception evolved over time?

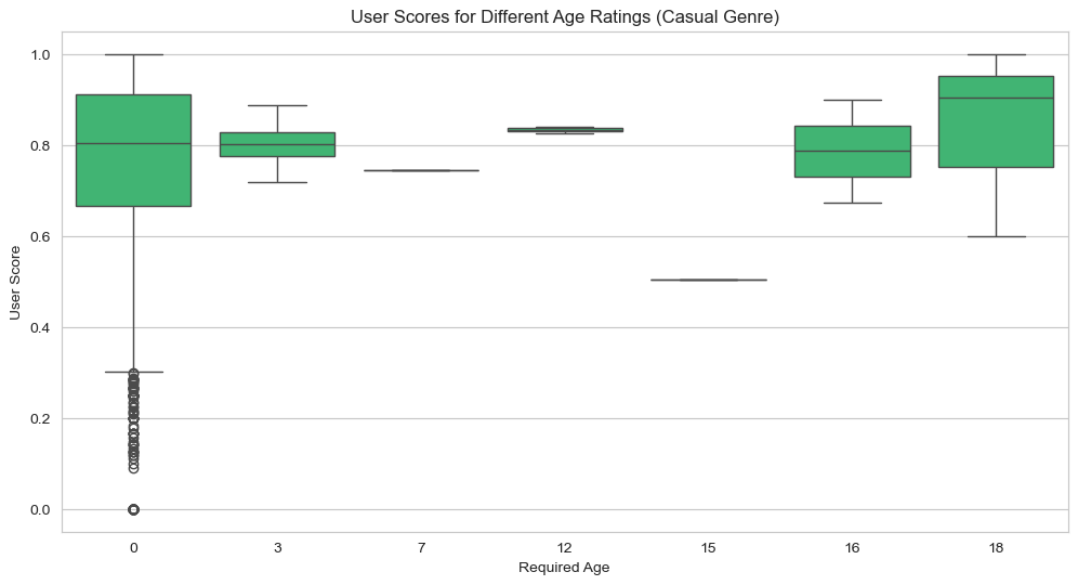
- **Conclusion:** User reception, as measured by average user scores, has shown a slight downward trend over the years on Steam. Years with more game releases tend to have slightly lower average scores, possibly due to increased competition or a wider range of game quality.

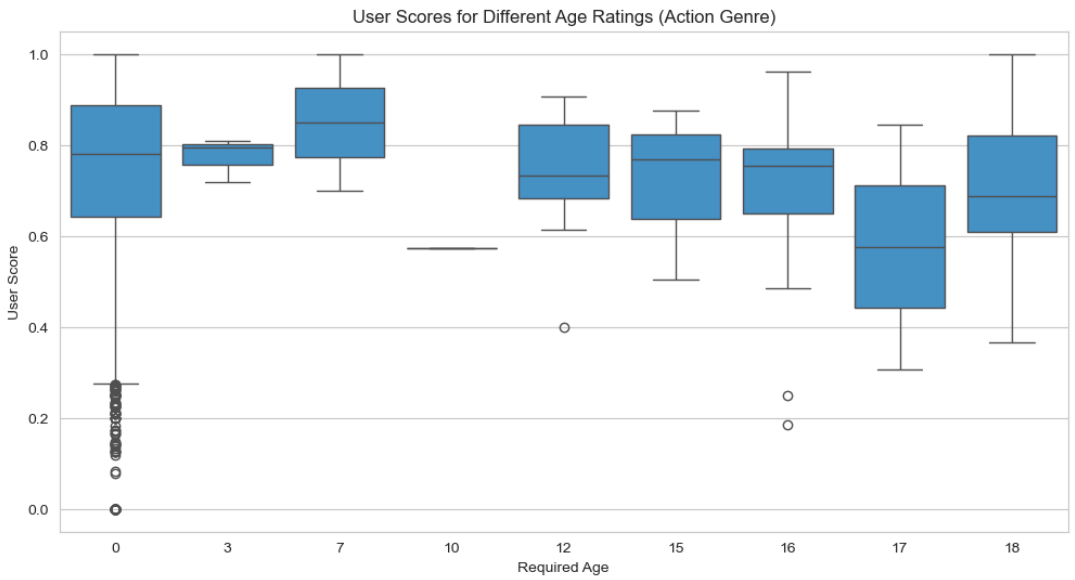
Q2: Do mature-rated games tend to have different user scores compared to those with lower age ratings, and how does this intersect with various genres?

Q2: Age Ratings and User Scores



Q2: Age Ratings and User Scores





Q2: Age Ratings and User Scores

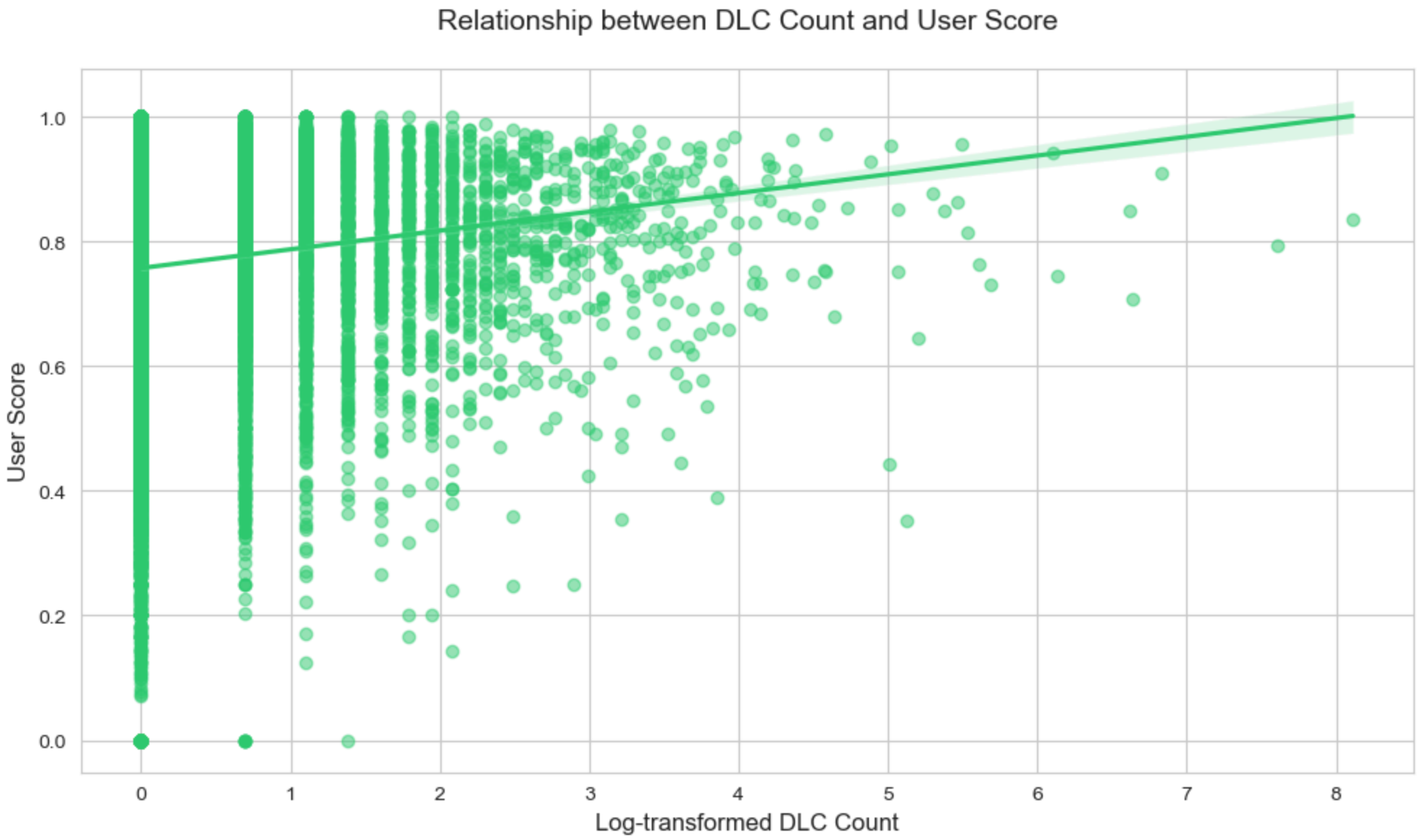
- **Findings:**
  - No statistically significant differences in user scores between different age ratings, either overall or within the Casual and Action genres.
  - Age rating does not appear to be a major factor influencing user reception in these genres.

Q2: Age Ratings and User Scores

- **Conclusion:** There is no strong evidence to suggest that mature-rated games have significantly different user scores compared to lower-rated games, at least within the Casual and Action genres. Developers in these genres might not need to be overly concerned about age ratings having a major impact on user reception.

Q3: How does the presence and quantity of downloadable content (DLC) relate to user scores, and has this relationship changed over the years?

Q3: DLC and User Scores



Q3: DLC and User Scores

- **Findings:**
  - Weak positive correlation (0.11) between DLC count and user scores.
  - The relationship varies across time periods, with a slightly stronger correlation in the Mid period (2006-2015).
  - The mere presence of DLC might be associated with slightly higher scores, but the amount of DLC is not a primary driver.

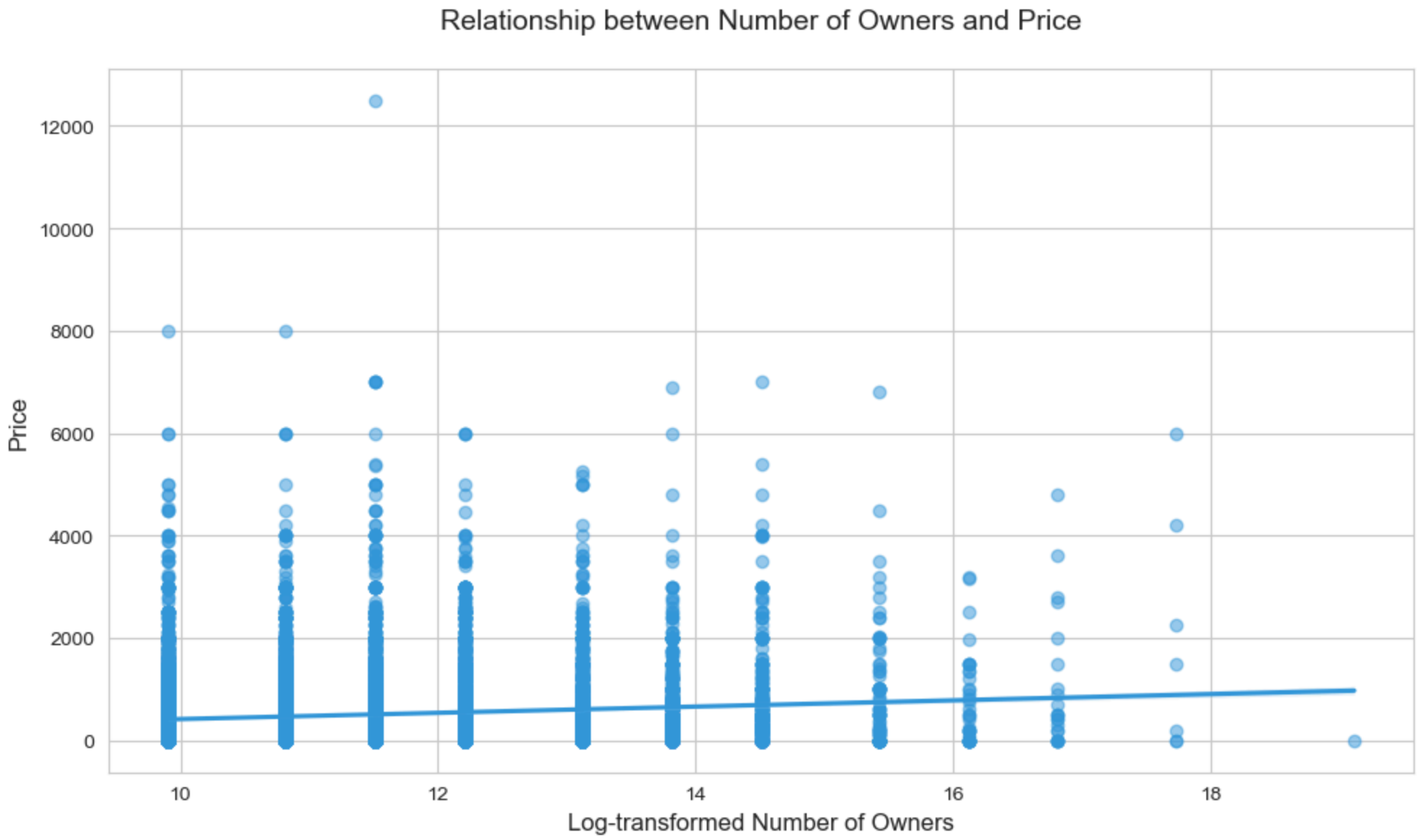
Q3: DLC and User Scores

- **Conclusion:** There's a weak positive relationship between DLC and user scores. While games with more DLC tend to have slightly higher scores, the impact of DLC seems to vary across different time periods and is not a primary driver of user reception. Developers should consider DLC as one factor among many and focus on quality and value rather than just quantity.

Q4: What is the relationship between a game's price, its user score, and its commercial success (estimated by the number of owners)?

Q4: Price, User Score, and Owners

Q4: Price, User Score, and Owners



Q4: Owners vs. User Score & Price

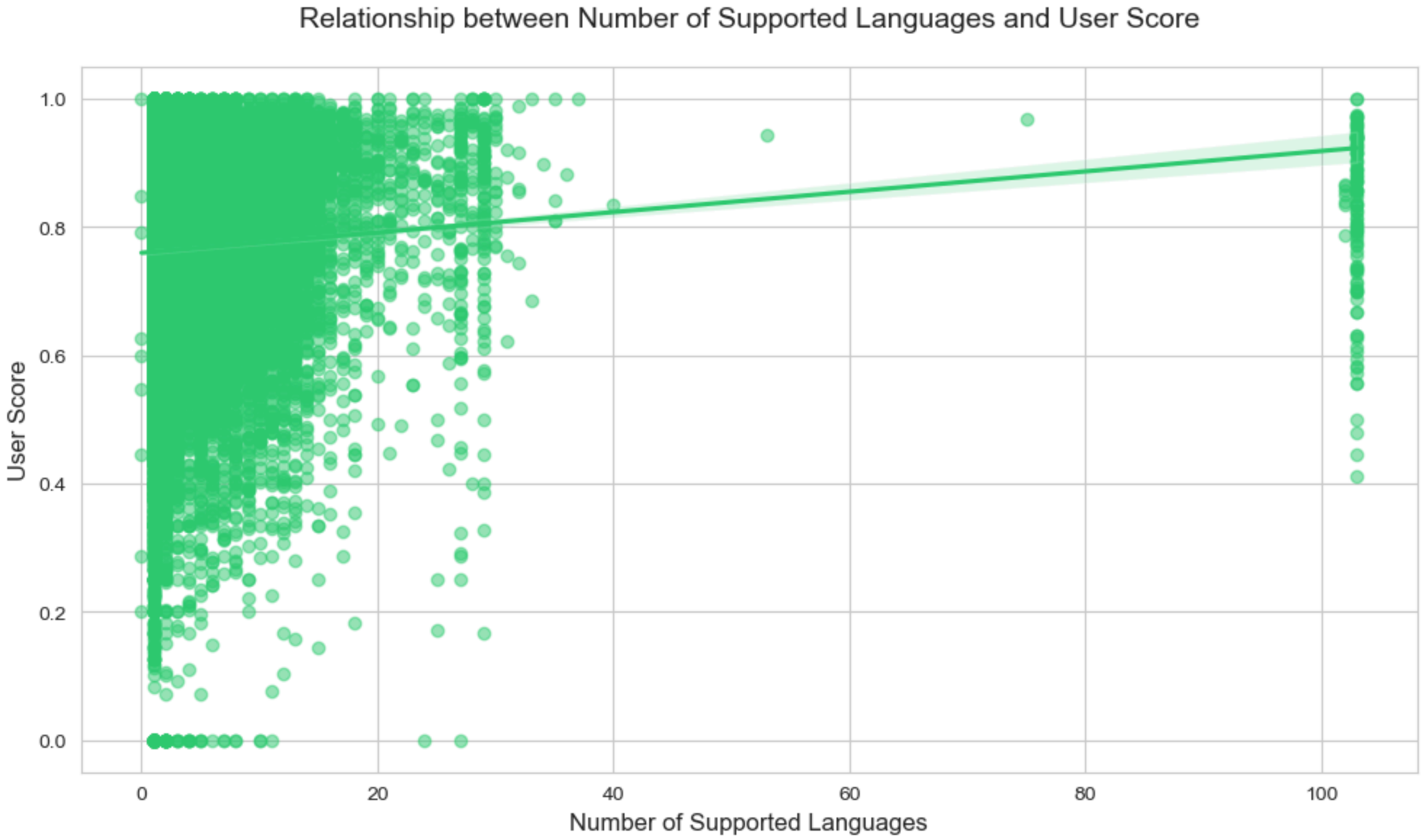
- **Findings:**
  - Weak positive correlations between:
    - Price and user score (0.12).
    - Number of owners and user score (0.14).
    - Number of owners and price (0.11).
  - Price alone doesn't guarantee high user scores or high sales.

Q4: Owners vs. User Score & Price

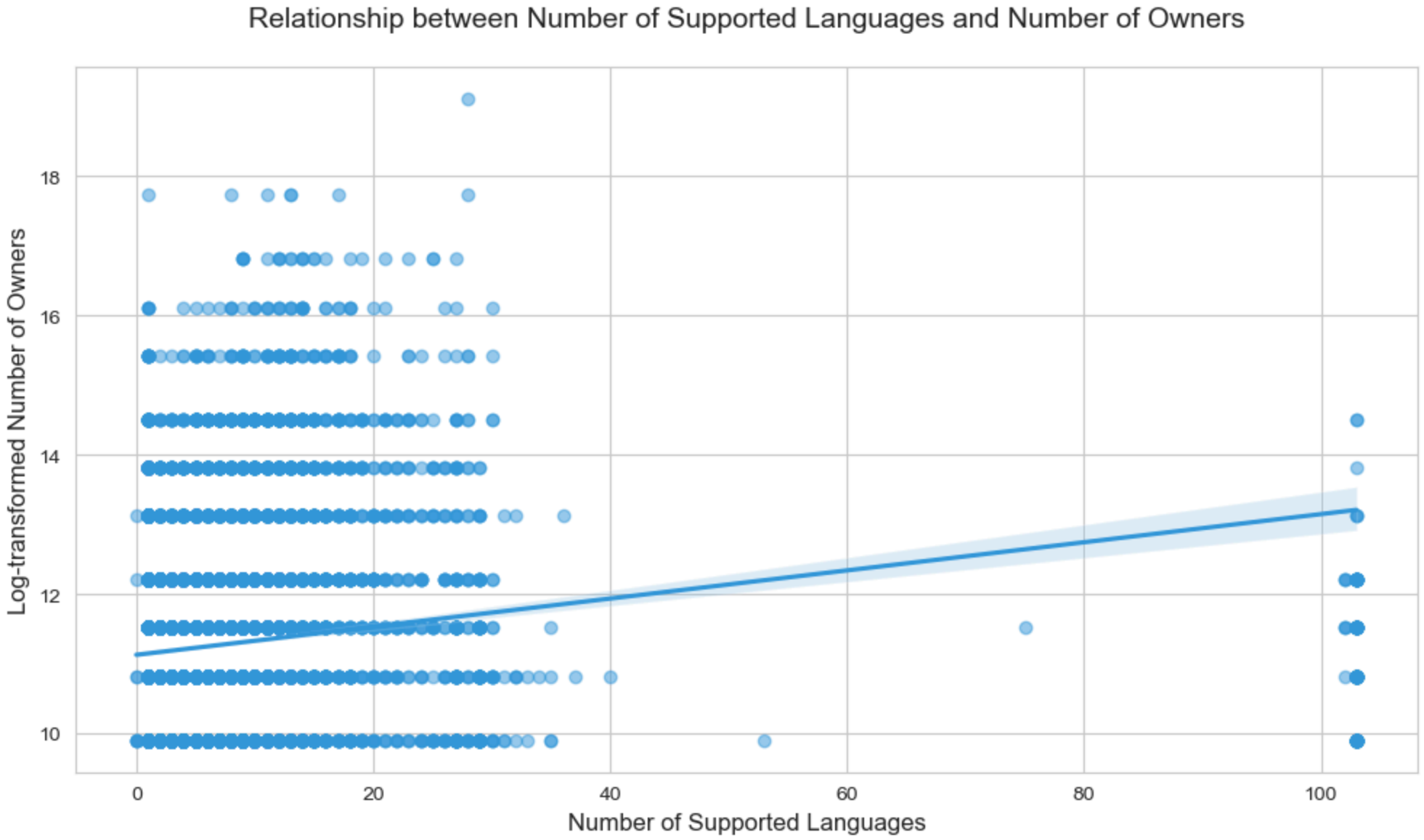
- **Conclusion:** The analysis reveals weak positive relationships between a game's price, user score, and the number of owners. However, price alone is not a strong determinant of user reception or commercial success. Developers need to carefully balance pricing with game quality, marketing, and target audience expectations.

Q5: How does the extent of a game's localization (number of supported languages) relate to its user score and its reach (estimated by the number of owners)?

Q5: Localization and Game Success



Q5: Localization and Game Success



Q5: Localization and Game Success

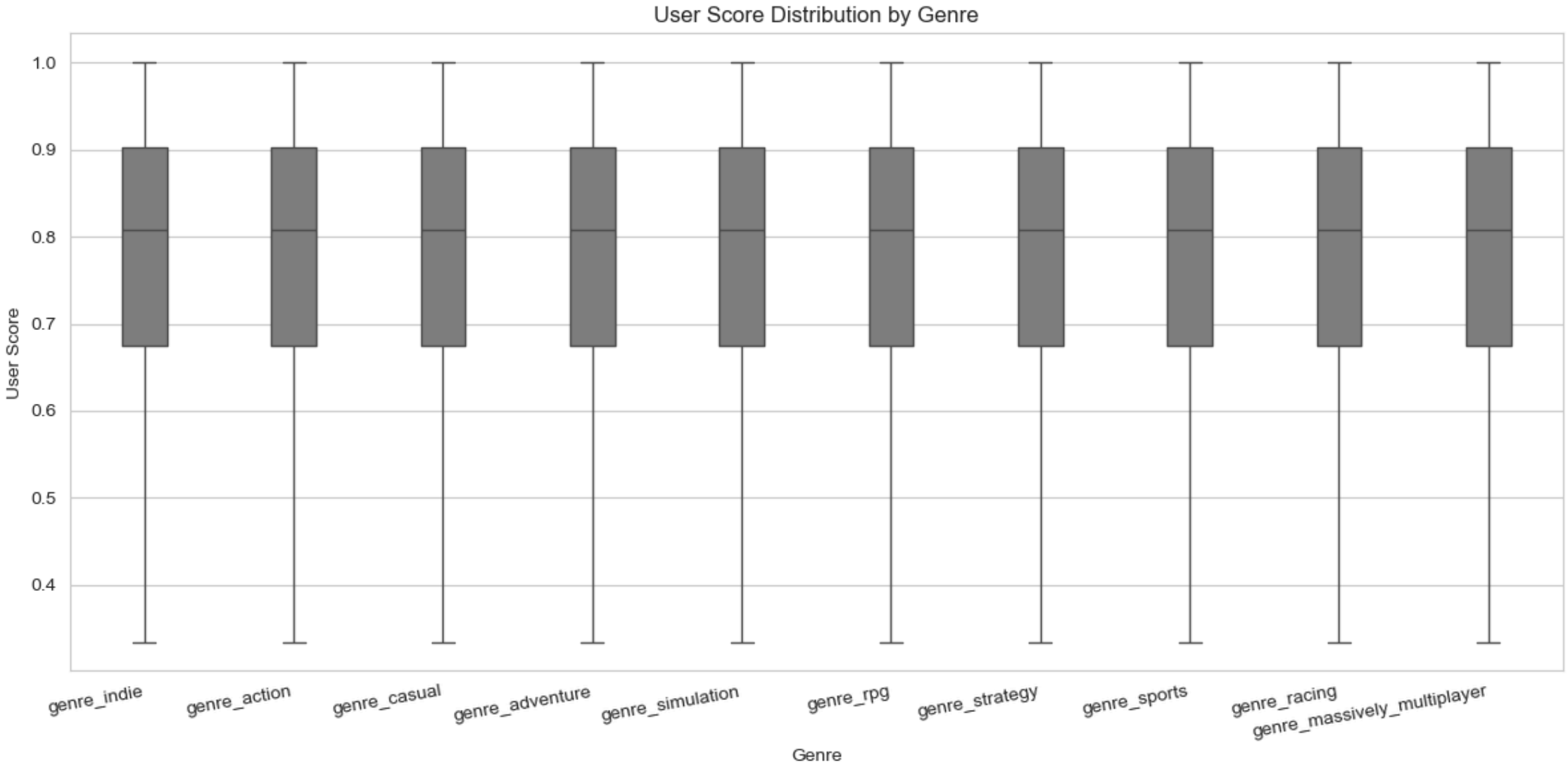
- **Findings:**
  - Very weak positive correlation (0.08) between language count and user scores.
  - Positive correlation (0.15) between language count and the number of owners.
  - Localization might not be a primary driver of user satisfaction but is important for reaching a global audience.

Q5: Localization and Game Success

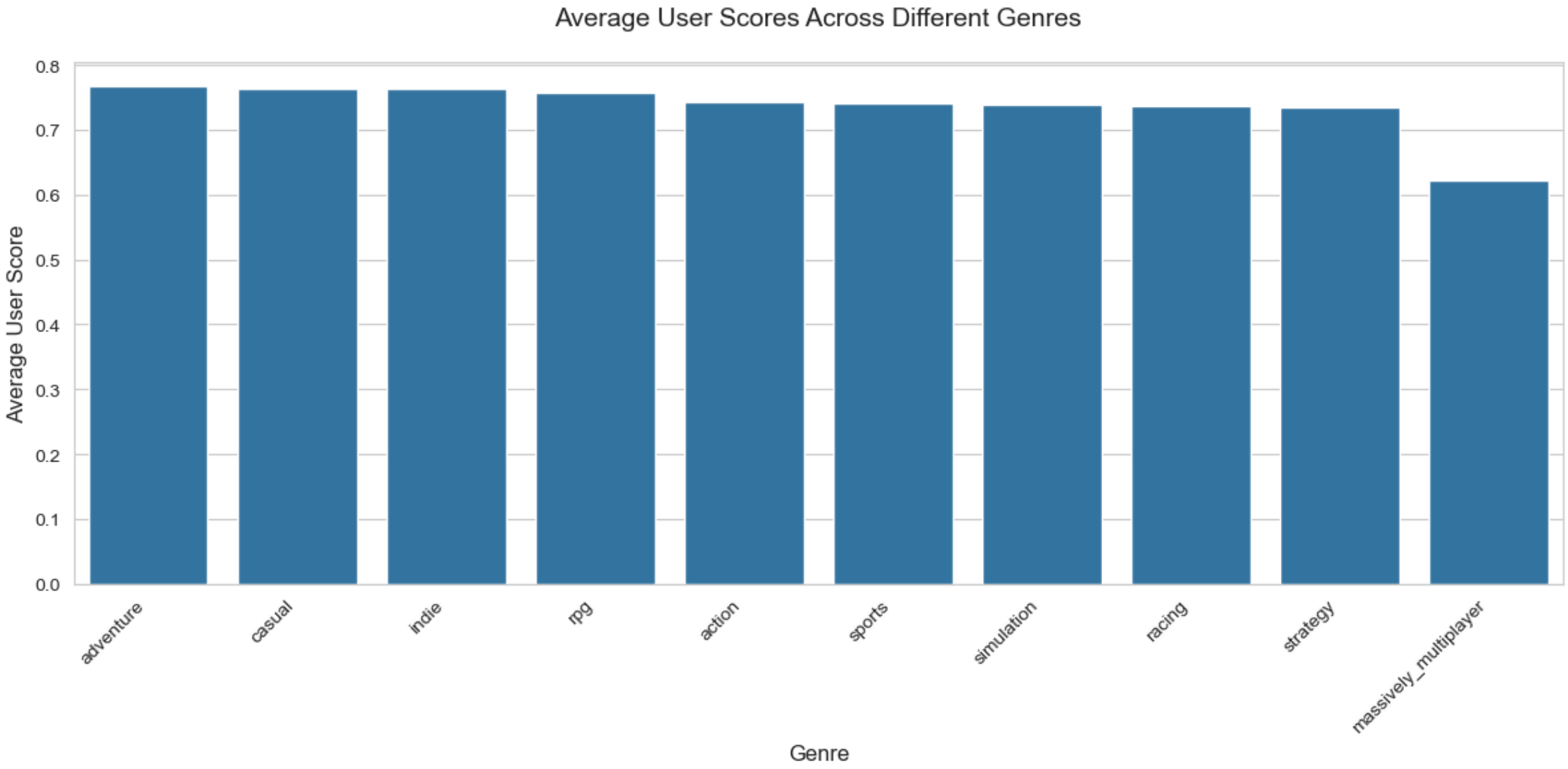
- **Conclusion:** While games with more language support tend to have slightly higher user scores and a wider reach, the correlation with user scores is very weak. Localization is likely more important for expanding a game's audience than for directly impacting user satisfaction.

Q6: How do user ratings vary across different game genres?

Q6: User Scores by Genre



Q6: User Score Distribution by Genre



Q6: User Score Distribution by Genre

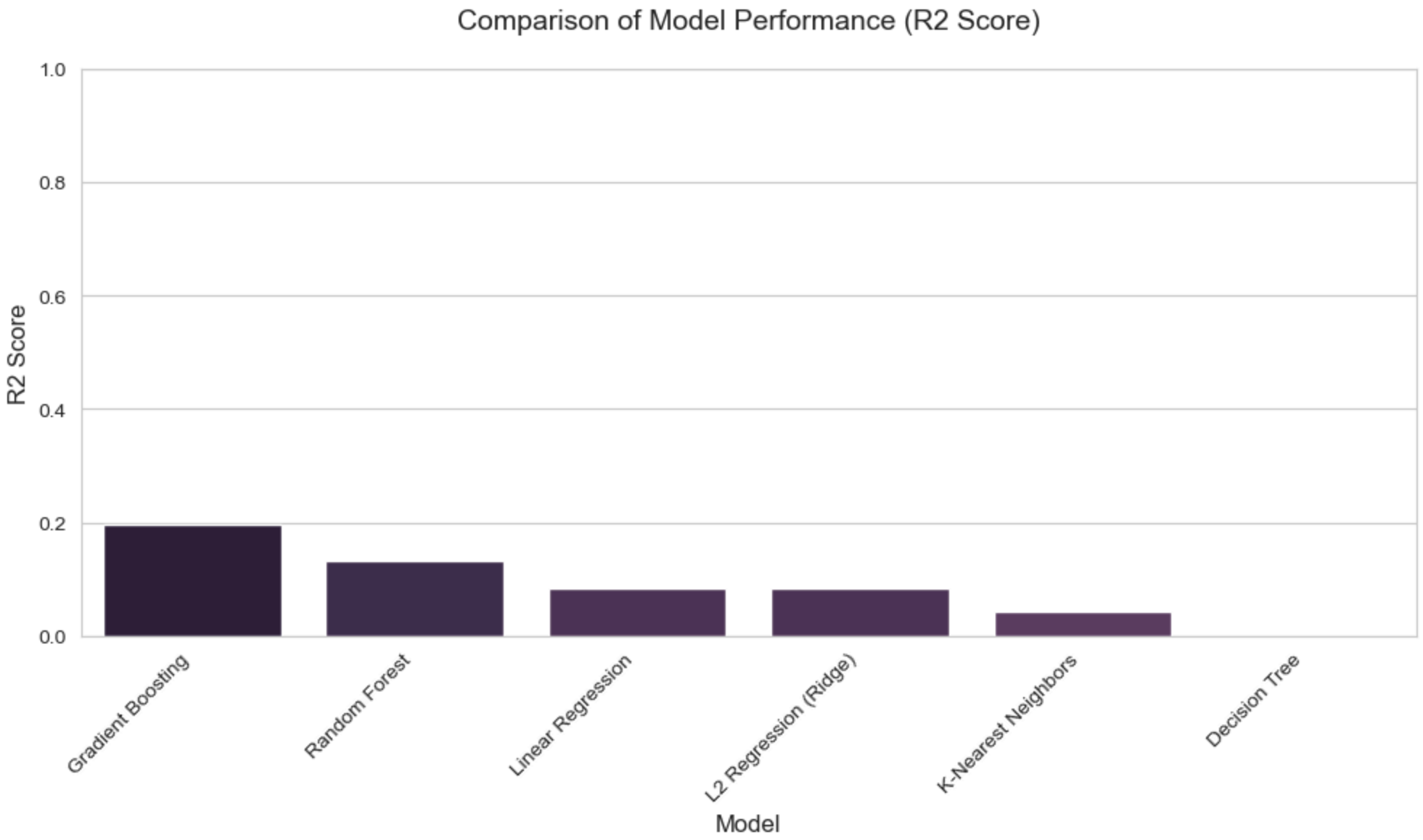
- **Findings:**
  - Average user scores are relatively close across genres, except for massively multiplayer games, which have a noticeably lower average score.
  - The distributions of user scores within each genre are similar.
  - No single genre significantly outperforms others in terms of user reception (except for the challenge in massively multiplayer).

Q6: User Score Distribution by Genre

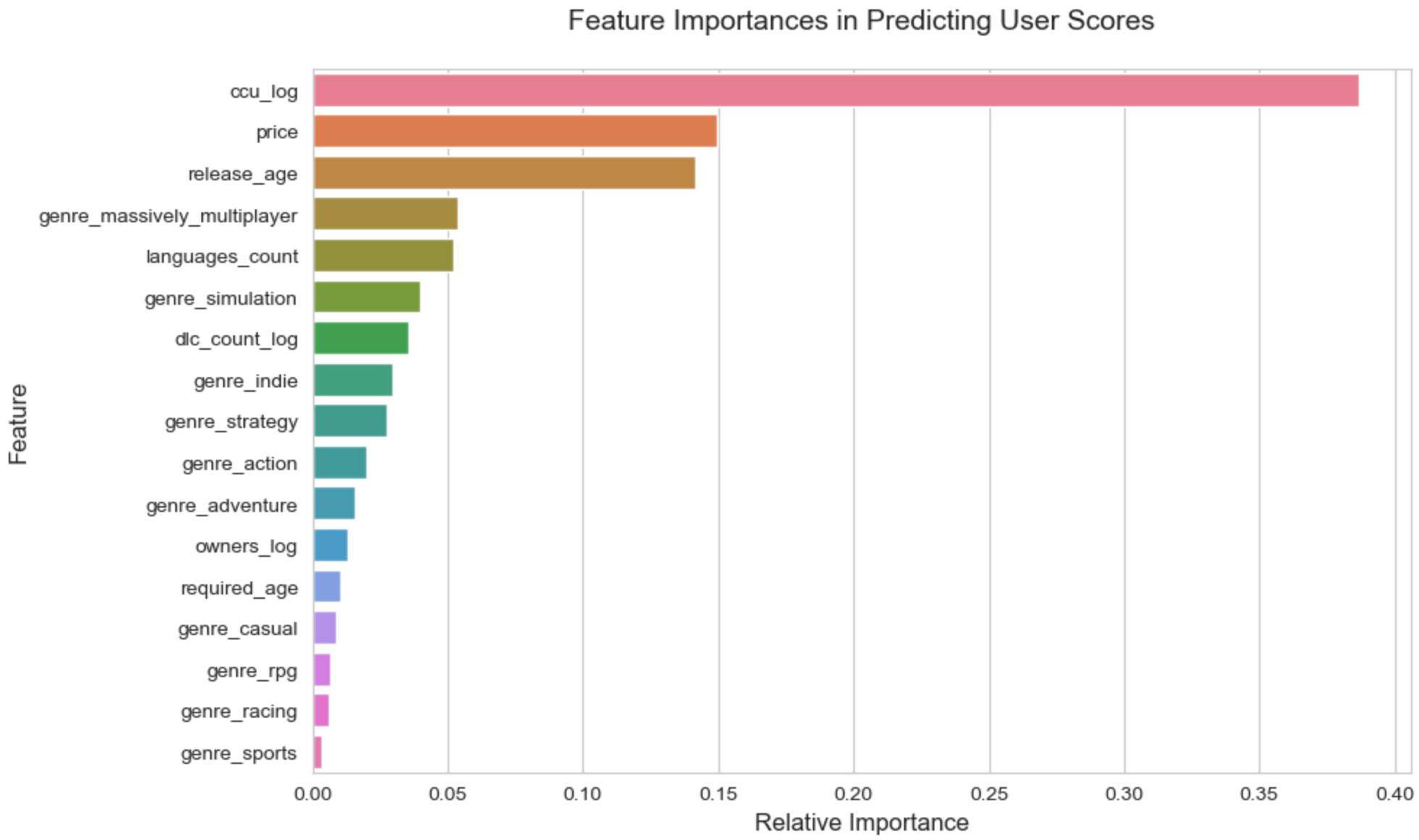
- **Conclusion:** User ratings do not drastically vary across most game genres on Steam. The lower average score for massively multiplayer games highlights the need for developers in that genre to carefully manage player expectations and address the unique challenges of online multiplayer experiences.

Q7: Which factors are the most influential in predicting user scores?

Q7: Factors Predicting User Scores



Q7: Factors Predicting User Scores



Q7: Factors Predicting User Scores

- **Findings:**
  - Strongest correlations with `userscore`: `ccu_log` (concurrent users), `owners_log`, and `price`.
  - Gradient Boosting model achieved an R-squared of 0.19.
  - Most important features: `ccu_log`, `price`, and then `release_age`.

Q7: Factors Predicting User Scores

- **Conclusion:** Concurrent users (`ccu_log`) and price (`price`) are the most influential factors in predicting user scores, based on the model, followed by release age (`release_age`). Developers should pay close attention to factors that contribute to a healthy concurrent player base and consider the potential impact of pricing on user perception.

Further Improvements

- Analyze sentiment in user reviews to gain deeper insights.
- Investigate the influence of marketing and promotional activities.
- Build more advanced predictive models with additional data.

References

- Teja, A. S., Hanafi, M. L. I., & Qomariyah, N. N. (2023). Predicting Steam Games Rating with Regression. *E3S Web of Conferences, 388*, 02001. https://www.e3s-conferences.org/articles/e3sconf/abs/2023/25/e3sconf_icobar2023_02001/e3sconf_icobar2023_02001.html