STA302 Final Project Proposal*

Ziqi Zhu Author 2 Author 3

October 11, 2024

Table of contents

1 Introduction (350)

The gaming industry has experienced exponential growth in recent years, with millions of players engaging in various games on platforms like Steam. Understanding what influences average playtime in games is crucial for game developers, publishers, and marketers. Average playtime is a key indicator of player engagement and game success, and investigating factors that impact it can provide valuable insights into game design, pricing strategies, and marketing efforts. So our research aims to explore the factors that influence the average playtime of

*Code and data are available at: https://github.com/zzq20010617/sta302Paper games available on Steam using a dataset we get from Kaggle. Specifically, we will investigate how variables such as price, user ratings, and other game characteristics contribute to player engagement, as measured by playtime. From some previous papers, we learned that both positive and negative reviews have a significant correlation with playtime, but the effect is more pronounced for positive reviews(Brodschneider and Pirker (2023)). Another paper found out that game pricing significantly affects player engagement, with lower-priced games generally seeing higher playtime.(Luisa et al. (2021)). The third paper did an empirical study that also examined the influence of game reviews on playtime, focusing on how the number and sentiment of reviews can affect player engagement. It concludes that games with more positive reviews see increased playtime, especially if those reviews highlight the game's quality(Lin et al. (2019)). We found this problem suitable for linear regression because it allows us to model the relationship between average playtime (the dependent variable) and various predictors (price, ratings, developer, etc.). This approach helps quantify how each factor influences playtime, enabling a clear understanding of the impact of different game attributes. By fitting a linear regression model, we can determine which factors have statistically significant effects on playtime, providing actionable insights for game developers and marketers.

2 Data description (300)

We get the Steam games data(Davis (2023)) from Kaggle. The R programming language (R Core Team (2023)) and packages readr(Wickham, Hester, and Bryan (2024)), httr(Wickham (2023)), jsonlite(Ooms (2014)) were used to download the data, dplyr(Wickham et al. (2023)) and lubridate(Grolemund and Wickham (2011)) were used to clean the data. Data was originally gathered from the Steam Store and SteamSpy APIs around May 2019. This table Table 1 shows the first 3 entries of cleaned data. We are going to take average_playtime as the response variable, it measures the mean time (in minutes) that players spend on a game, a summary of this variable is shown in Table 2. This variable captures overall player engagement and serves as a good indicator of how immerse or entertaining a game is. It is suitable for a linear regression model because it is continuous and quantitative. The predictors selected are price, positive ratings, negative ratings, developer, and number of owners. All are numeric except for the number of owners, which is categorical since it represents estimated ranges. Price may influence playtime, as higher costs could lead to longer engagement. Positive ratings and negative ratings are reviews of players of that game and it can only be positive or negative. Positive ratings likely correlate with greater playtime, while negative ratings could indicate the opposite. Developer refers to the development company of the game, which could affect playtime, with established studios often producing longer, high-quality games. Lastly, the number of owners is an estimated number of owners, containing lower and upper bounds (like 20000-50000), it could signal popularity, where more owners may suggest higher average playtime. Summary of numerical predictors can be find in table Table 3.

Table 1: Steam Games Data

release_month	english	developer	publisher	achievements	positive_ratings	negative_ratings	average_playtime	median_playtime	owners	price
Nov	1	Valve	Valve	0	124534	3339	17612	317	10000000-20000000	7.19
Apr	1	Valve	Valve	0	3318	633	277	62	5000000-10000000	3.99
May	1	Valve	Valve	0	3416	398	187	34	5000000-10000000	3.99

Table 2: Summary of Average Playtime

Mean	Median	Std_Dev	Min	Max
657.37	222	3783.67	1	190625

Table 3: Summary of Predictors

mean_price	median_price	mean_positive_ratings	median_positive_ratings	mean_negative_ratings	median_negative_ratings
7.47	4.99	4181.49	408	858.13	113

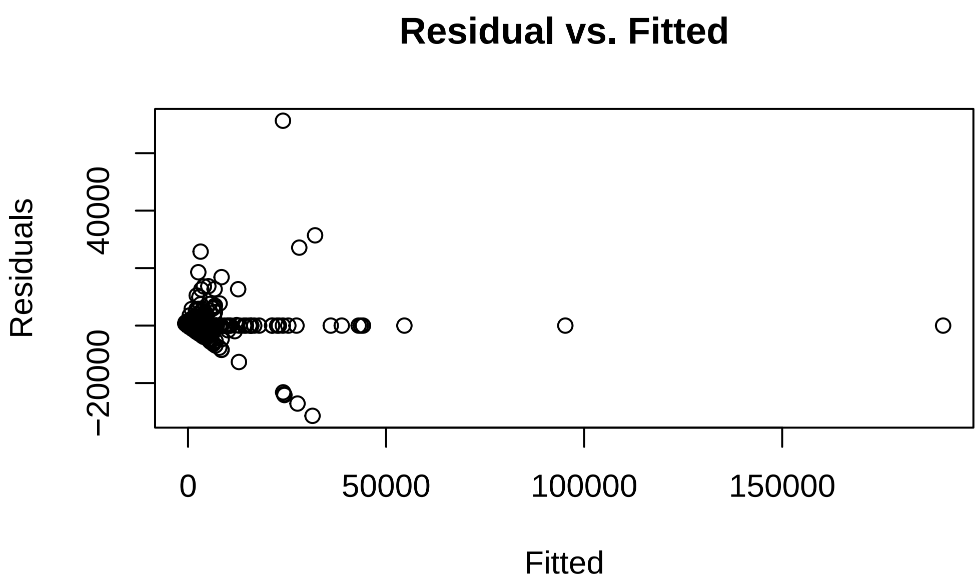
3 Ethics discussion (100-200)

The Steam Games dataset(Davis (2023)) used in this research is a collected dataset, not simulated, comprising information on over 97,000 games published on the Steam platform. The metadata is comprehensively filled out, including details such as game titles, release dates, genres, and user ratings, which enhances the dataset's usability and reliability. The source of the data is clearly described, originating from the Steam store pages and Steam API, ensuring transparency and traceability. Additionally, the dataset is hosted on Kaggle, a reputable platform for data science and machine learning, which implies a level of vetting and popularity within the data science community. This widespread use and accessibility suggest that the dataset is well-regarded and trusted by third parties, further validating its credibility. There are no ethical concerns to report.

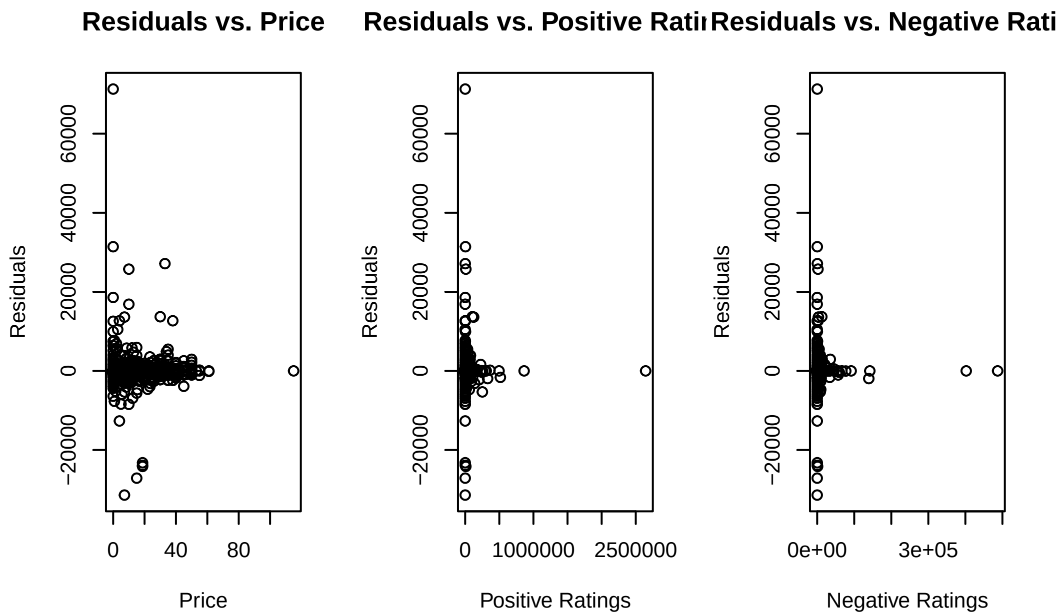
4 Preliminary results (300)

4.1 Fit multiple linear regression model

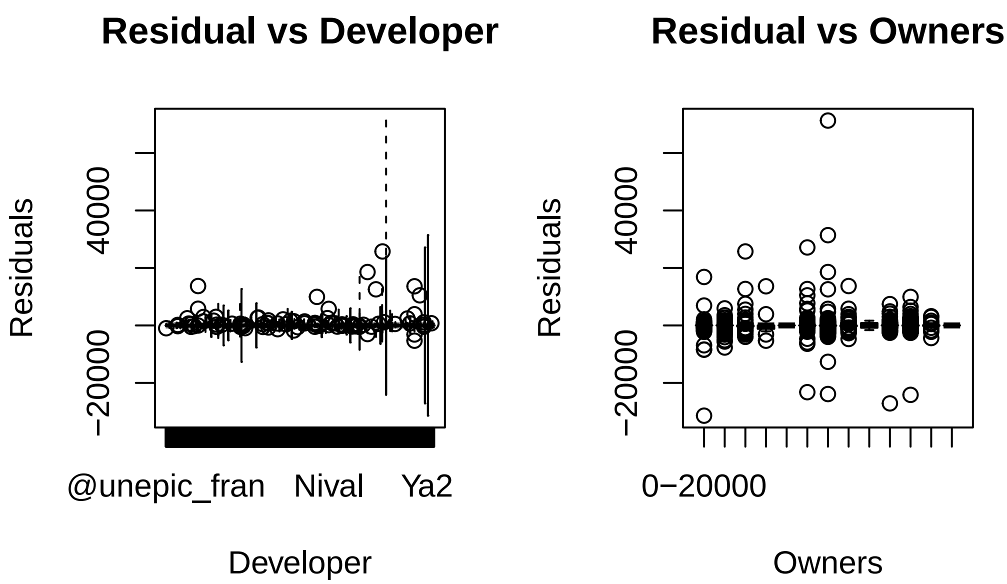
4.2 Create Residuals vs Fitted Scatterplot to check Linearity, Uncorrelated Errors, and Constant Variance



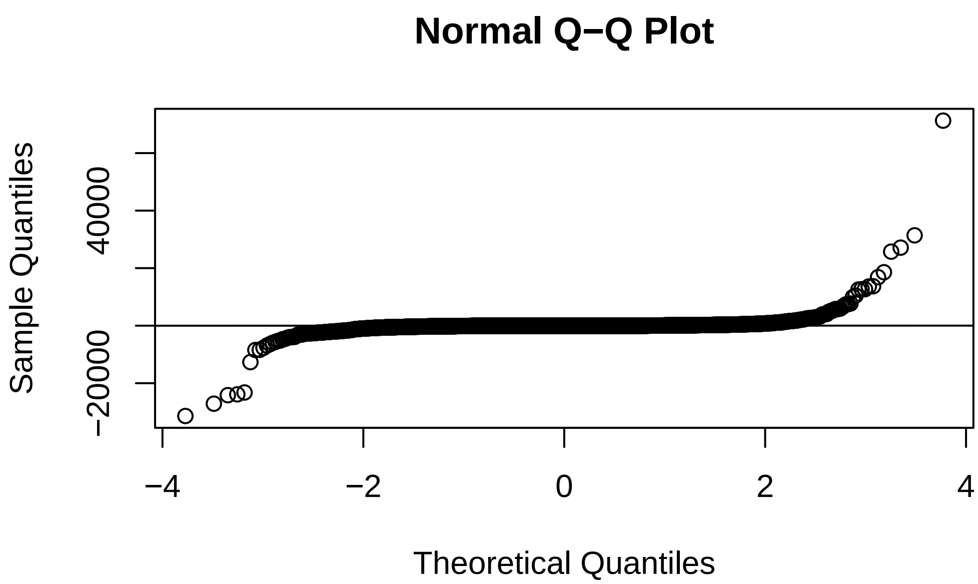
4.3 Create Residuals vs each Quantitative Predictor scatterplot to check Linearity, Uncorrelated Errors, and Constant Variance



4.4 Create BoxPlot to check assumptions for Categorical Predictors



4.5 Create Normal QQ Plot to check Normality assumption



References

Brodschneider, Vinzenz, and Johanna Pirker. 2023. "On the Influence of Reviews on Play Activity on Steam - a Statistical Approach." *ResearchGate*. https://www.researchgate.net/publication/376227235_On_the_Influence_of_Reviews_on_Play_Activity_on_Steam_-_A_Statistical_Approach.

Davis, Nik. 2023. "Steam Store Games (Clean Dataset)." https://www.kaggle.com/datasets/nikdavis/steam-store-games/data.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Lin, Dayi, Cor-Paul Bezemer, Ahmed E. Hassan, and Ying Zou. 2019. "An Empirical Study of Game Reviews on the Steam Platform." In *Empir Software Eng*, 24. https://doi-org.myaccess.library.utoronto.ca/10.1007/s10664-018-9627-4.

Luisa, Andraž De, Jan Hartman, David Nabergoj, and Samo Pahor. 2021. "Predicting the Popularity of Games on Steam." *ResearchGate*. https://www.researchgate.net/publication/355110719_Predicting_the_Popularity_of_Games_on_Steam.

Ooms, Jeroen. 2014. "The Jsonlite Package: A Practical and Consistent Mapping Between JSON Data and r Objects." *arXiv:1403.2805 [Stat.CO]*. https://arxiv.org/abs/1403.2805.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2023. *Httr: Tools for Working with URLs and HTTP*. https://CRAN.R-project.org/package=httr.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.