STA302 Final Project Proposal*

Ziqi Zhu Author 2 Author 3

October 8, 2024

Table of contents

1 Introduction (350)

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

*Code and data are available at: https://github.com/zzq20010617/sta302Paper

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

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 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)

4 Preliminary results (300)

References

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/.

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.