Lab 2 Final Report

Team GitHub Repository lab-2-r-fan-club

Jenny Park, Justin Nhan, Peter Liu, Ryan Castillo

April 20, 2025

Introduction

Video games are one of the most influential forms of entertainment worldwide, with the global market valued at $184 billion and serving over 3.2 billion gamers worldwide.1 But what makes a video game so captivating?

There are many aspects people enjoy about video games, whether it be the escapism into new worlds, the ability to connect with friends or strangers, the thrill of overcoming challenges and unlocking achievements, or even to have a sense of agency in their video game life. Previous research has pointed to factors like sociability and in-game achievements as important predictors of how long players stick around for a session.2

In this analysis, we aim to better understand our primary research question: **What factors are associated with engagement in entertainment-style video games?**

Description of Data Source

For this analysis, we used Steam Games Dataset, which is publicly available on Kaggle. Steam is one of the largest digital distribution platforms for PC games, and contains a wide variety of information on over 97,000 titles. The data was collected using web scraping methods to extract game metadata either directly from Steam via their Web API, or SteamSpy, a third-party platform that estimates sales of games offered on Steam. The code used to compile this dataset can be found here, and was last compiled in October of 2024. This dataset contains individual games as the units of observation - with each row representing a single listing on Steam - and includes key features such as the game's title, release date, median playtime, genres, price, estimated owners, meta-critic scores, and number of user recommendations. The breadth of attributes in the dataset make it well-suited for exploratory analysis, with the goal of describing the relationships of some of these variables against playtime.

Data Wrangling

The Kaggle API was used to download the dataset via CSV. We engineered three new features: percent_positive (percentage of positive reviews), days_since_released, and a categorical re-binned owners column. Percent_positive was derived by summing the positive and negative counts for the game, and dividing by the total count of positive and negative values. Days_since_released was derived by taking the difference between the present date and the game's release date. Games with more than 5 millions owners were consolidated into a single bin to manage skewness, creating the owners feature. Finally, the cleaned dataset was split into a 70% training set and a 30% exploration set. The exploration set was used for initial analysis, and the training set for confirming the final model.

Conceptualization and Operationalization

To evaluate a video player's level of engagement, overall playtime was selected to serve as the key proxy for measure. Since average playtime is influenced by outliers (few players with extremely high playtimes), the median playtime (mins.) feature was selected as our primary metric instead. A conscious effort was made to only focus on features in the dataset for which video game creators have direct or indirect influence over (Price, # Achievements, % Positive Reviews, etc.), ensuring our model could provide interpretable and actionable insights.

The data was then filtered to remove games with zero estimated owners, zero playtime, and a release date within the past two years. Doing so reduced our dataset from 97,410 records to 14,665, effectively removing 82,745 observations. This was done to ensure our data only included games with a measurable level of engagement and games with enough time beyond being released to establish a reasonable steady-state of playtime. Based on our exploratory analysis of the dataset, many educational-style video games identified were not relevant to our focus. Thus, 13 non-entertainment genres were discarded. This was accomplished through manual analysis of all genres, categorizing genres as entertainment (Action, RPG, etc.) or non-entertainment (Accounting, Audio Production, etc.). This step removed 229 observations, and the remaining genres were one-hot encoded for further analysis.

Model Specification

Our first approach was to select a base variable to build a simple, credible model. We then theorize correlations with additional variables (See Table 4), considering collinearity and statistical significance in explaining variance, as detailed in the Model Results and Interpretation sections. Many variables are highly right-skewed with zero values, so we applied a log-plus-one transformation to normalize the data and improve "human" interpretability. For example, a $10 difference in a $20 game feels more significant than in a $100 game.

Model Assumptions

The OLS regression performed for the above model requires several assumptions to be met to ensure that coefficient estimates are unbiased and efficient. With a dataset of over 14,000 observations, we note that we can invoke the Central Limit Theorem and rely on asymptotic properties. This relaxes the requirement for normally distributed errors and heteroskedasticity (provided robust standard errors are used). Remaining assumptions are assessed below:

1. **I.I.D. data -** A violation of i.i.d would leave these observations fundamentally challenged - with standard errors being potentially understated and coefficients not being reflective of true relationships. The assumption of i.i.d. may be challenged due to both social influence and complex dynamics between variables. For instance, our friend's extended playing time may convince us to start playing, introducing interdependence among observations. Additionally, variables such as recommendations, ownership, and playtime may not be unidirectional - games with higher playtime may cause more recommendations, which drive further engagement. Also, quality content may increase playtime, generating revenue and causing the

studio to further improve the game. Finally, there may be a temporal element - with gaming behaviour changing pre vs post-pandemic resulting in different distributions of playtimes between periods.

2. **Linear Conditional Expectation** - If violated, the linear model becomes a mis-specified approximation of the true relationship, leading to biased or misleading coefficient estimates. However, the above relationship between fitted relationships and the residuals appears roughly linear. We do note a slight curvature in the graph indicating a slight underprediction towards the lower playtimes and overprediction towards the higher playtimes. (See Figure 1)

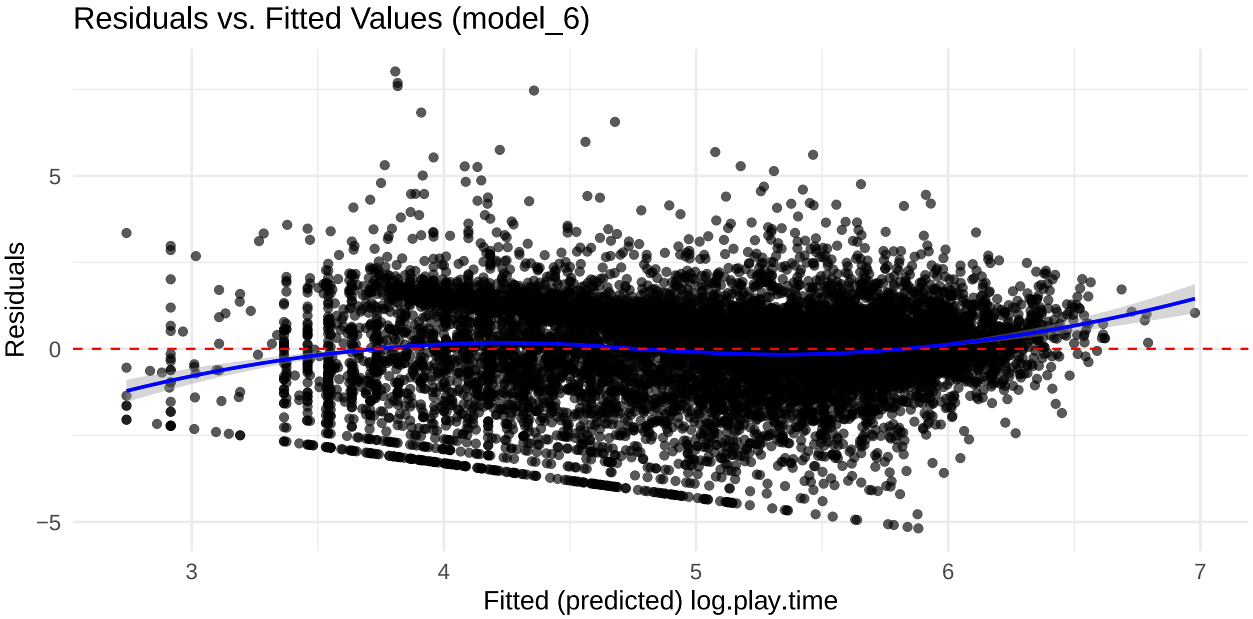

Figure 1: Model Residuals Fits

3. **BLP Exists & No Perfect Collinearity** - A violation of this assumption may lead to strange balancing between coefficients (one a huge positive, one may be hugely negative), or a splitting of effects between variables. P-Values will also be inaccurate. However, our model retained all factors - suggesting no perfect collinearity. Additionally, VIF (See Table 1) was calculated for each variable; with all being substantially below 5 confirming collinearity is not a major concern in the model. This supports the existence of a BLP.

Model Results and Interpretations

As per the model specification, the final model was chosen with select genres to improve interpretability. The model resulted in an adjusted R^2 of 0.207/0.216 (train/test), suggesting strong generelization. An F-Statistic of 143, and an ANOVA comparison (See Table 2) between models 1 and 6 yielded an F-Statistic of 198.1 ($p < 2.2e-16$) supports us with more confidence in the model.

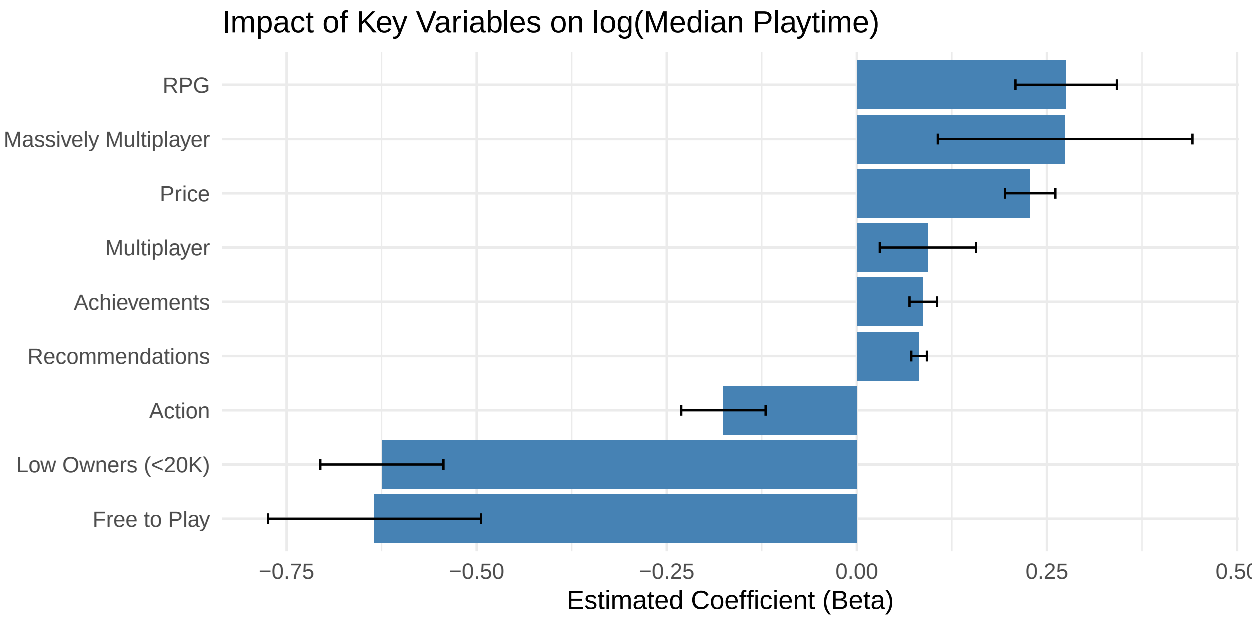

Figure 2: Model Coefficients

Coefficients & Interpretation

As the outcome variable median_playtime is logged, each of these variables can be interpreted as the percentage change in median playtime. A number of themes emerged when analysing the coefficients (See Figure 2)

Quality games: Price remained both statistically and practically significant. As the variable is logged, the coefficient can be interpreted as a % change in price driving a % change in median playtime - a 10% change in price (See Table 3) translates to a roughly 2-3% bump in playtime***. Caution must be applied to this variable's interpretation, as price generally does not incentivize others to play the game (save for some 'guilt' of having to play a game after paying for it). Rather, we believe that price serves as a proxy for the financial resources put into a game - a higher price my translate to more talented developers, longer build-time etc. Similarly, recommendations both practically and significantly improve median playtime - a 10% increase in recommendations translates to a 1% increase***. We have interpreted recommendations as community perception of a game's quality. Thus, we propose that high quality games get played longer.

Social Incentive: Multiplayer games are associated with a 1% improvement in median playtime***. Similarly, one of the largest coefficients is Massively Multiplayer (MMO), which is both statistically and practically significant (played ~30% longer than others***). MMO games such as World of Warcraft possess an element of companionship as thousands of players can participate in the world together. Conversely, we see games with a low amount of owners appear to get played 50% less***. Though we acknowledge that games with low owners tend to also be of poorer quality, we also propose that there is a social element which incentivizes game time.

Genre Impact & Achievements: A game's genre appears to have the largest impact on median playtime - for example roleplaying (RPG) games are played ~30%*** longer. Both MMOs and RPGs tend to focus on levelling up and immersion, which drives longer playtimes. Also associated with levelling up are in-game achievements which players may seek for a sense of satisfaction.

(An achievement may involve beating a certain level) The model suggests that a 10% increase in achievements increases playtime by ~1%***) Conversely, Action and Free-To-Play games are designed for more 'quick-hit' style gameplay, the latter of which is associated with a 50% drop in play-time***. It is worth noting that because these games are free, they also have a lower barrier to entry, incentivizing gamers to try the game even if not fully invested, also factoring in to the low play-times.

Overall Effect

Overall, the analysis provided several themes which are associated with median playtime, which we propose may be used as a proxy for engagement. We find that higher quality games, indicated by the financial resources put into a game (reflected in price) and player perception (recommendations) tend to get played more. Second, social elements keep us engaged - we play games more with our friends (especially Massively Multiplayer Games). Finally, a game's genre and achievements help provide a greater sense of progression and keep gamers playing.

However, we note limitations in the assumption that median playtime is a proxy for engagement, as shorter length games also have the potential for engagement. We also caution against the interpretation that a game should be priced higher to be played more, and rather suggest that games with more financial resources put into them are played more.

As grounds for further study, we would implement interaction terms such as the interaction between owners and recommendations, and consider additional elements such as game production cost to capture more of the underlying effects of our price variable.

Next Step

Our current model focuses on price, based on the theory that it reflects content quality. We now aim to explore the question from the perspective of single-player vs. multiplayer game modes. Multiplayer may have more replay value due to regular social interaction, while single-player experiences can build long playtime through shorter, consistent sessions—similar to a Netflix-style binge watching.

As a next step, we propose using single vs. multiplayer as the primary variable, similar to how we started with price. We also suggest exploring alternative target measures, such as number of consecutive play days, to better capture different play patterns. Additionally, we plan to look further into interaction terms, especially between owners and recommendations, to understand how social proof and reach affect engagement. This follow-up can help uncover new perspectives on what keeps players coming back.

References

1. https://www.trade.gov/media-entertainment-video-games-sector
2. https://www.sciencedirect.com/science/article/abs/pii/S0747563216304563?via%3Dihub
3. https://github.com/mkearney/kaggler
4. https://www.kaggle.com/code/berent/r-api-for-kaggle-datase

Appendix

Table 1: VIF table

Variable	VIF
log1p(price)	1.697425
log1p(recommendations)	1.681103
genre_Free_to_Play	1.560597
genre_Massively_Multiplayer	1.263985
owners_low	1.254329
cat_multiplayer	1.194890
log1p(achievements)	1.095672
genre_Action	1.061924
genre_RPG	1.048998

Table 2: ANOVA test to compare nested models

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
10100	21891.94	NA	NA	NA	NA
10092	19231.65	8	2660.295	174.502	0

Table 3: Model 6 - Coeftest Robust SEs

term	estimate	std.error	statistic	p.value
(Intercept)	4.17569	0.04379	95.35908	0.00000
log1p(price)	0.22789	0.01693	13.46059	0.00000
log1p(recommendations)	0.08180	0.00525	15.59498	0.00000
log1p(achievements)	0.08734	0.00926	9.43153	0.00000
owners_low	-0.62474	0.04131	-15.12211	0.00000
cat_multiplayer	0.09343	0.03228	2.89412	0.00381
genre_Action	-0.17550	0.02834	-6.19328	0.00000
genre_RPG	0.27528	0.03404	8.08729	0.00000
genre_Massively_Multiplayer	0.27392	0.08543	3.20623	0.00135
genre_Free_to_Play	-0.63431	0.07146	-8.87661	0.00000

Table 4

[!htbp]
Table 5

	Dependent variable:				
	log(1 + Median Playtime)				
	Price	+ Recs	+ Achv.	+ Multi	+ All Genres
	(1)	(2)	(3)	(4)	(5)
log(Price)	0.458***	0.261***	0.244***	0.245***	0.228***
	(0.014)	(0.015)	(0.015)	(0.015)	(0.018)
log(Recommendations)		0.132***	0.125***	0.124***	0.082***
		(0.005)	(0.005)	(0.005)	(0.005)
log(Achievements)			0.082***	0.082***	0.087***
			(0.009)	(0.009)	(0.010)
Multiplayer					$-$0.625***
					(0.042)
Owners (low)				0.041	0.093**
				(0.030)	(0.034)
Genre: Action					$-$0.175***
					(0.030)
Genre: RPG					0.275***
					(0.035)
Genre: MMO					0.274**
					(0.086)
Genre: F2P					$-$0.634***
					(0.072)
Constant	4.134***	3.911***	3.786***	3.776***	4.176***
	(0.033)	(0.034)	(0.036)	(0.037)	(0.059)
Observations	10,102	10,102	10,102	10,102	10,102
R^2	0.109	0.175	0.182	0.183	0.217
Adjusted R^2	0.109	0.175	0.182	0.182	0.217
Note:			*p<0.05; **p<0.01; ***p<0.001		

[!htbp]
Table 6

TRUE