**Prediction of Game Downloads**

### Narendranath Edara

### Missouri University of Science and Technology, Rolla, Missouri, USA

**Abstract**

This paper explored the subject of prediction of a video game’s success. A preliminary data modeling was done before collecting and cleaning data and forming a model to test the hypothesis using R. The endogenous and exogenous variables were found and identified. Data was collected, cleaned and formatted for processing. Training and testing data sets were formed and linear cases were built. Lastly, a regression diagnostics was formed through linear models, a statistical estimation and data visualization.

**Keywords: Video Games, Ratings, User Ratings**

**Table of Contents**

1. **Introduction**
2. **Literature review**
3. **Data Acquisition**
4. **Data Preparation**
5. **Data Description**
6. **Experiments**
7. **Conclusion**
8. **Bibliography**
9. **Introduction**

In today’s world, the video game industry is one of the largest industries in the world. In the year 2020, the PC gaming market reached a worldwide estimation of almost 37 billion U.S. dollars. Within the United States alone in the same year, the video game market reached an estimated 77 billion dollars. By March 2022, the hardware revenue of the United States video game industry reached 515 million dollars, while the value of the video game market reached a total estimate of 85.86 billion dollars. [[1]](https://www.statista.com/topics/868/video-games/#editorialPicks)

Video content creation by civilians and private users have now created an emergence of influencers and micro-celebrities of which their contents, views and reach with followers creates a significant income. Reviews from various influential internet personalities and users on top live streaming services for gaming such as Twitch have great influence over millions of online users, casting a wide net potential market influence.

The objective of this study is to project the total estimated amount of money the video game Poppy Playtime will make through analyzing the content provided on these various websites.

**2. Lit Review (Related Work)**

The emergence of video games started in the 1970s. Now, more than 50 years later, an entire generation of first generation gamers have grown up and have significant consumer market power. It was found that within gaming habits of internet users 18 years or older within the U.S., 76% of those sampled played video games and almost 18% spent more than 6 hours per week playing video games. [[1]](https://www.statista.com/topics/868/video-games/#editorialPicks)

Within the video game industry, it has grown and changed over the decades, starting with arcade games and console games to now the most popular market of PC and mobile games. These massive gaming companies are continuing to strive to publish and produce financially successful games to stay competitive within the massive, multi-billion dollar market. Alpert (2007) summarized his account of how massive, valuable and competitive this market is: *“Finally, consider how well games companies are doing financially. Many of the early game companies have shrunk or disappeared (e.g., Infocom, Broderbund games), but the largest modern games company is gigantic by any standard. In 1994 Electronic Arts (EA) was a company with $500 million in sales and a stock market capitalization of $1.8 billion (all dollar figures in this paper are US$). By 2004, it had become a $3 billion in sales company with a market capitalization of $15 billion. This places an exclusively games company as the fourth largest capitalized software maker in the world, behind Microsoft, Oracle and SAP Corporation (Lowenstein 2004).” (Alpert, 2007)* [*[2]*](https://core.ac.uk/download/pdf/15035023.pdf.)In the year 2022, the economic power behind the consolidation of the video game industry has created lasting changes within the market. These massive companies have made acquisitions to follow the wave of what is now seen as an unstoppable market of big tech. Microsoft acquired Activision Blizzard, another gaming company for $68 billion on January 18, 2022 while Amazon acquired Twitch for $970 billion in August 2014. [[3]](https://blog.twitch.tv/en/2016/07/13/game-creator-success-on-twitch-hard-numbers-688154815817/)

The nature of gaming itself has changed how users purchase video games through subscription options that allow users to access an entire library of games at their disposal, of which new games are constantly added into. This has opened doors for gamers and consumers to reach a much wider net within the market.

Youtube is a popular and widely used service among gamers to consume gaming content. Producers of gaming content also provide a massive following and reach a wide audience through their content of reviews, first impressions and gameplays. Many content producers reach millions of users per video on Youtube and maintain that consistent viewership through their following that could be impacted by the reach and popularity of the games itself. Through Twitch, a massive video service that became a popular streaming platform for gamers to share real-time gameplay, it has also proven to be a source of large appeal reaching millions with influencers of their own. This platform allows users to share their gameplay with anyone in the world in real time.

Twitch conducted a study between the connection of Twitch views to sales on Steam, another similar popular website that collects video game data, claimed that up to 25% of sales on Steam came from users who watched a game on Twitch and subsequently bought it within the next 2 hours. Furthermore, players who watched a given game within the same week the user had played it showed a 5% increase in retention. However, there are flaws within this study based on the small sample size of only 0.53% of viewers that connected their Twitch account to Steam. [[4]](https://www.statista.com/topics/868/video-games/.) For future research, more studies regarding the connection between Twitch view to Steam accounts should be delved deeper into with a much higher sample size as well as more current data. Another weakness in this study was the aspect of direct bias and dubious intent, as Twitch conducted the study about their own company and most likely omitted unfavorable data that reflected poorly about their company’s power of influence.

One study published in April 2019 examined the efficacy of predicting future video-game sales based on past sales using data from about 8,000 different games across a variety of platforms. Data was collected from the website Kaggle from the specific dataset entitled “Video Game Sales with Ratings” released in 2016 as well as the review website of Metacritic. The independent variables collected were platform, genre, North American Sales, Japanese sales, other sales, critic score, number of critics scoring, user scores, number of users scoring, and rating. The qualitative values were 17 various popular video game platforms. The qualitative values collected by genre were adventure, role playing, action, sports, stimulation, puzzle, shooter, racing, strategy, playform, fighting, and miscellaneous. Sales in millions of units were collected and divided into 4 geographical categories of North American, European, Japanese and other sales (the rest of the world excluding the three previously mentioned). The fifth and final sales variable used as the dependent/ output variable was global sales. Furthermore, critic score and critic count, user score and user count and the qualitative variable of user rating was collected. The major hypothesis was that users buy games based on their overall aspects and features, not genre classification. [[5]](https://www.questjournals.org/jrbm/papers/vol7-issue3/I07036064.pdf)

This study helps point towards the potential positive connection between global sales and higher critic scores and the negative connection between global sales and higher user scores. This data suggests that potential game purchasers and reviewers are skeptical of the ratings of other users while they tend to believe users who have more professional credibility. Concerning the data we collected for our study, the results of this study supports the pattern seen by our dataset which suggests a possible positive correlation between followers of certain video game “influencers” and number of positive reviews of a particular game.

**Research Method**

**3. Data Acquisition**

Numerical statistical data was collected from the websites Steam, Twitch and Youtube. First, we decided to collect data with games based on the 9 different genres within the Steam website: action, adventure horror, sports, stimulation, RPG, Arcade, Sandbox and puzzles. From Steam, we collected and analyzed the price of different games, the developer, the release date, all reviews until the release date, the number of recent reviews, all positive reviews, all negative reviews. From Twitch, we collected metrics on the average gain or loss on overall players by looking at the average number of players. Finally, we tracked whether or not each game was playable on Mac OS, Windows OS or both. For each video game, data was collected via Steam with the name of the game, price, developer, release date, number of all reviews, recent reviews, positive and negative reviews of the individual variables.

We collected data from a popular Youtube video for each game by picking the video review created by popular Youtube gamers or “influencers”.

The number of downloads and purchases on the website Steam was estimated through using the New Boxleiter number, since the data of how many copies of games that are sold on steam is not public. However, by using the public data for the number of player reviews on Steam, one can calculate the average number of sales per review as well as the median number of sales per review.

**4. Data preparation**

First global options were set to apply to every chunk in the file. The following linear models were used: logistic regression, gradient boosting machine and one for data visualization and data cleansing.

The following libraries were imported: caret, ggplot2, tidyr, dplyr and MLmetrics.

To begin, first we cleaned the data by selecting the data frame and converted all character values to be numeric for each category.

Then we displayed the internal structure of the data using str(data). We manipulated the categorical variables set with factor levels.

To deal with any missing values from our data set, labeled “N/A”, we used the parameter “na RM” to remove the missing values and check NA values which let us know if the return was true or false. After it was run, there were no missing values in the data set found.

To create the response variable, we used mutate to add “downloads,” which is a product of the difference of days between the release date of the game and April 30, 2022 as well as “all\_rev”.

**5. Data visualization**

To visualize the data, we first used ggplot() to create a scatter plot with a regression line of Twitter followers versus Twitch followers by the number of days since the game was released. The plot revealed that there seems to be a positive correlation between their values.

**Data description**

To track whether or not the game could be played on Mac or Microsoft windows or both, the variables 1 and 0 were used. If a game could be played on one of these platforms or both, it was labeled with a 1. If not, it was marked with a 0. Data was collected from twitter by extracting from a popular tweet about the game that included a video within it. The number of twitter views of that video was collected, labeled under the variable “Twt\_view” while the number of followers of the author of that particular tweet was collected.

**Experiments**

*Correlation plot*

A correlation plot was created to extract and transform the objects. The function “corr() was used to calculate the correlation. The results revealed that variables all\_rev, rec\_rev to downloads and twt\_view to diff\_days are highly correlated. Other variables like ytc\_view to ytc\_likes to ytc\_com, twt\_view to diff\_days and twi\_flw to ytc\_view have a strong correlation.

*Grouped boxplot*

A grouped boxplot was run to visualize the data categories of Youtube channel, subscriptions, likes, and comments. First, the reshape package was called. The data frame was melted to create a boxplot. Then the title of the graph as well as x and y axises were labeled and the title was added. We found that the boxplots ytc\_view, ytc\_likes and ytc\_com have more outliers as compared to ytc\_subs which have just 1 outlier. The data in ytc\_subs and ytc\_likes seem to have a similar range of values and median. The data in ytc\_view and ytc\_com look the same with a minor difference in ytc\_view.

I partitioned the data into training and testing sets using randomized numbers via set.seed(). I randomly split the dataset into a train (80%) and a test set (20%). Then we set the rows for both the train and test sets and created a table. The table displays that out of 80 observations, the train-test-split of 80-20 assigned 64 observations for the training set and 16 observations for the test set.

Model 1: Multiple linear regression

The first multiple linear regression model created included selected predictors of all reviews, recent reviews, youtube channel subscribers, and twitch followers. The results concluded that the predictors are statistically significant.

The second multiple linear regression model concluded that the predictors of youtube comments, twitch followers and recent reviews are statistically significant as well.

The third multiple linear regression model with all predictors revealed that the time between the release date and April 30, 2022 and the number of twitch followers are statistically significant.

Model 2: Logistic regression

glm() was used to plot the logistic regression because our data has a non-normal distribution.

The predicted probability was calculated, then the probability of downloads was calculated using the “predict()” function. Predict() was used to find the first 10 values of test\_probs. The function logit() was used to model a logistic regression. Then we did a logit model prediction to calculate the predicted downloads. The results were plotted.

Model 3: Gradient Boosting Machine

The gbm() library was used to create a gbm model. Within this model, the data had a normal distribution which was set to “gaussian”, which is a bell shaped curve. We tuned the parameters by setting the number of trees (the number of gradient boosting iteration) to 1000. The cv.folds were set to 20, shrinkage to 0.01, n.minobsinnide to 10 (the minimum number of observations in the trees terminal nodes) and the number of trees (n.trees) to 1000.

**Limitations**

Some limitations that we encountered in our study was the absence of data altogether for Twitch followers. Some games were not present on the Twitch website, and therefore, a couple of games did not have data on the current number of Twitch followers. Some video games also did not release the data for all of the set metrics we were looking for.

**Conclusions**

After running all models on the data set to predict the number of downloads, the resultant performance figures do not favor the models, as their RMSE and MAE values seem to not convince them that their performance was optimal. Although comparatively GBM alone did a good job with its highest R2 value among all the models, but with the results of the model, we will not choose to go with any of the above supervised learning models to predict PC game downloads w.r.t data collected from Steam, Twitter, Youtube, and Twitch. However comparatively, the gradient boosting machine has somewhat performed better than others.

Data Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Meaning** | **Source** | **Notes** |
| gname | Name of the game | Steam page | gname |
| gprice | Price of the game | Steam page | gprice |
| dev | Developer of the game | Steam page | dev |
| rel\_days | Days since the release date to april 22nd 2022 | Steam page | rel\_days |
| all\_rev | All reviews until date | Steam page | all\_rev |
| rec\_rev | Recent Reviews | Steam page | rec\_rev |
| pos\_rev | All positive reviews | Steam page | pos\_rev |
| neg\_rev | All negative reviews | Steam page | neg\_rev |
| ytc\_name | Reviewer Youtube channel name | Youtube | ytc\_name |
| ytc\_subs | Reviewer Youtube subscribers | Youtube | ytc\_subs |
| ytc\_view | Reviewer Youtube video views | Youtube | ytc\_view |
| ytc\_likes | Reviewer Youtube video likes | Youtube | ytc\_likes |
| ytc\_com | Reviewer Youtube video comments | Youtube | ytc\_com |
| twt\_view | Twitter views for game trailer | Twitter | twt\_view |
| twt\_flw | Number of current followers on Twitter | Twitter | twt\_flw |
| twi\_Flw | Number of current followers on Twitch | Twitch.tv | twi\_Flw |
| platform\_mac | If game is playable on Mac OS | Steam page | if the game can be run on mac, then 1, if not then 0 |
| platform\_win | If game is playable on Windows | Steam page | if the game can be run on windows, then 1, if not then 0 |
| platform\_win+mac | If game is playable on both Mac OS and Windows | Steam page | if the game can be run on both windows and mac, then 1, if not then 0 |

Bibliography

1. Alpert, Frank (2007), “The Entertainment Software Industry: Suddenly Huge, Little Understood,” Asia-Pacific Journal of Marketing and Logistics, Vol. 19, No. 1, 87-100. Retrieved from https://core.ac.uk/download/pdf/15035023.pdf.
2. Z. Xu, Analysis of economic benefits and development trend of China's game live streaming market. Consumer Guide, 2019, no. 5, pp 126-127.
3. HERNANDEZ, Danny. *Game Creator Success on Twitch: Hard Numbers* [online]. 2016 [visited on 2017-03-21]. Available from: https://blog. twitch.tv/https-blog-twitch-tv-game-creator-success-on- twitch-hard-numbers-688154815817
4. Clement, J. “Topic: Video Game Industry.” *Statista*, 19 Nov. 2021, https://www.statista.com/topics/868/video-games/.
5. Alice Yufa" Predicting Global Video-Game Sales" Quest Journals Journal of Research in Business and Management, vol. 07, no. 03, 2019, pp 60-64