Steam Sales Project

Andrew Dai

12/10/20

Abstract

Growing interest in the video game industry has attracted significant attention from investors, developers and players alike. Certain titles are known to be more popular among players than others and price plays an important role when players choose the games they buy. Steam is an online platform where users can purchase and play downloadable video games. Data was collected from the Steam API and records Steam game sales from May of 2019. This report will use a multilevel linear model to see how a game's ownership and price effect it's rating score, accounting for random effects between genres. The results found that there was no relationship between ownership levels and rating score and a small positive relationship between price and rating score.

Introduction

The video game industry has developed significantly since the rise of the home computer. One market in particular which has sprung onto mainstream audiences is the access to downloadable computer games. Valve is a game distributor, among many others, from whom players can purchase and download games. The proliferation of Valve is attributed to their ability to connect individuals on their online platform, Steam, and enable interaction between players.

Developers publishing their work onto Steam can charge a one time fee for the indefinite access to their game. This access is typically restricted to the player's Steam account and intended for a single person.

With corporations such as Valve, Blizzard, Epic and Riot propping up digital content markets with their own respective platforms, they provide audiences with an enormous library of computer games- a selection process which forces players to be more discriminating towards different titles. Not every game can be treated equally by the average gamer- the truth is that a majority of titles will probably go untested.

As the industry grows, it attracts more attention from investors. There is a growing interest in what makes certain titles so popular among audiences. One group who is interested in what makes a game popular is the advertising industry who target the video gaming demographic to purchase products that they are likely to buy.

This report will investigate the question of what makes a game popular on Steam. We will use a linear mixed model with random effects between game genres. The dataset was taken from the Steamspy API by Nik Davis who published it on Kaggle and performs Exploratory Data Analysis on his blog (Davis, 2019). Davis has provided valuable insights through his visualization and this report intends to follow-up by 1) checking his EDA with our own and 2) fitting the data to a statistical model.

Methods

Variable Selection

Price is one consideration players may have before they choose to purchase a game. When we invest ourselves in a game, we dedicate some amount of resource into actually playing the game through. With a free title, the only downside could be an hour or two of wasted time. However the more expensive a game is, we expect less players to be willing to invest in such a costly risk and so we expect smaller ownership of that title.

Number of players is important for players who value multiplayer games. The most-owned games on Steam are typically massive multiplayer games where a large number of players contribute to the game's activity. More players often leads to a better experience as there is less waiting to begin a match. With a large playerbase, developers are also more likely to push new content and patch existing problems. As a result, games with large ownership tend to be higher rated than games with small ownership.

Ratings are important to a prospective game buyer. Players are more likely to try a game with positive ratings and pass on a game with negative ratings. Ratings can reflect the value of the game, factoring in quality as well as cost. We can expect games with a large number of positive ratings to be 1) inexpensive and 2) with large ownership.

Data selection

The raw dataset contains 27,000 observations. For our analysis, we removed 1 game with price greater than 100, 2,390 games with less than 20,000 owners, and 23,113 games with more than one genre tag or genre tags that were not our groups of interest. The primary interest of this report will be to show the effect of price and ownership on game rating with random effects between genres. Games with price under 100, more than 20,000 owners and only one genre tag will allow our analysis to be more meaningful to the majority of games. The ten genre groups we have chosen are Action, Indie, Strategy, RPG, Racing, Casual, Adventure, Sports, Simulation, and Free to Play.

Additionally a rating score was computed based on SteamDB's method (SteamDB Team, 2017):

$TotalReviews = PositiveReviews + NegativeReviews$

$ReviewRatio = \frac{PositiveReviews}{TotalReviews}$

$RatingScore = ReviewRatio - (ReviewRatio - 0.5) * 2^{-log(TotalReviews+1)}$

This method pulls ratings towards 50%, meaning that a game with 0 reviews will have a Rating Score of 50. Because we do not know how players would rate the game, we assign it the average value.

We begin our analysis with an initial Exploratory Data Analysis to see generally what is going on in our subset data.

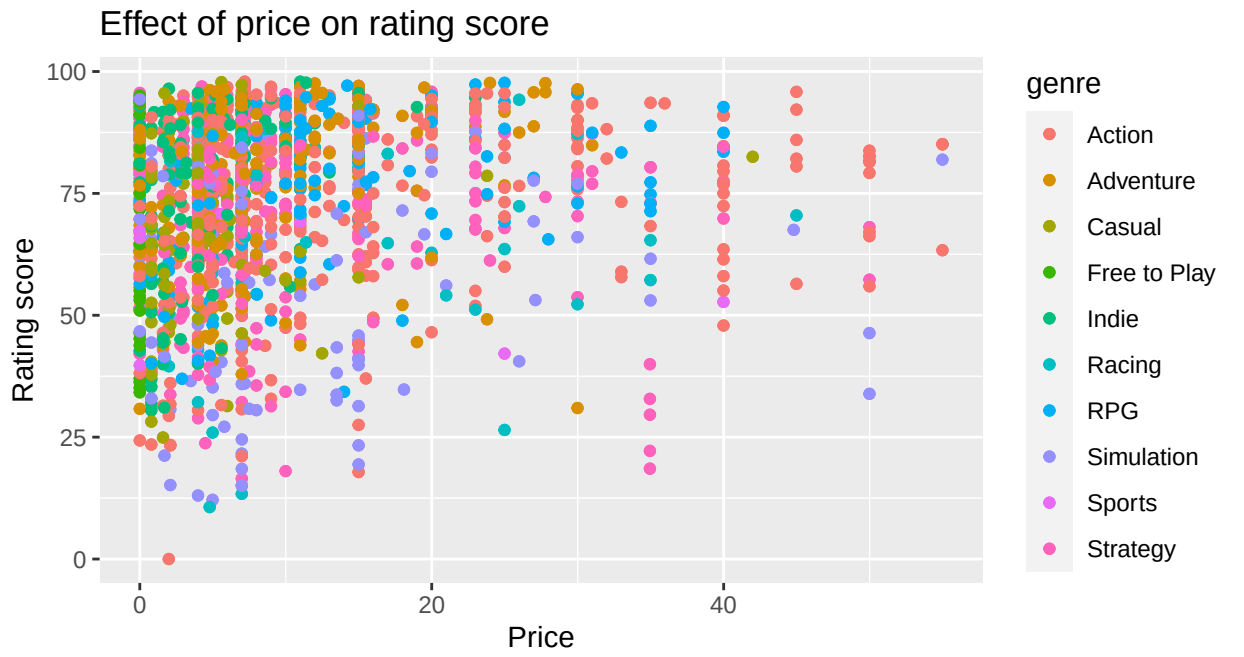


Figure 1: the effect of price on rating score, grouped by genre. A majority of our data exists with a rating score above 50 and price less than 30.

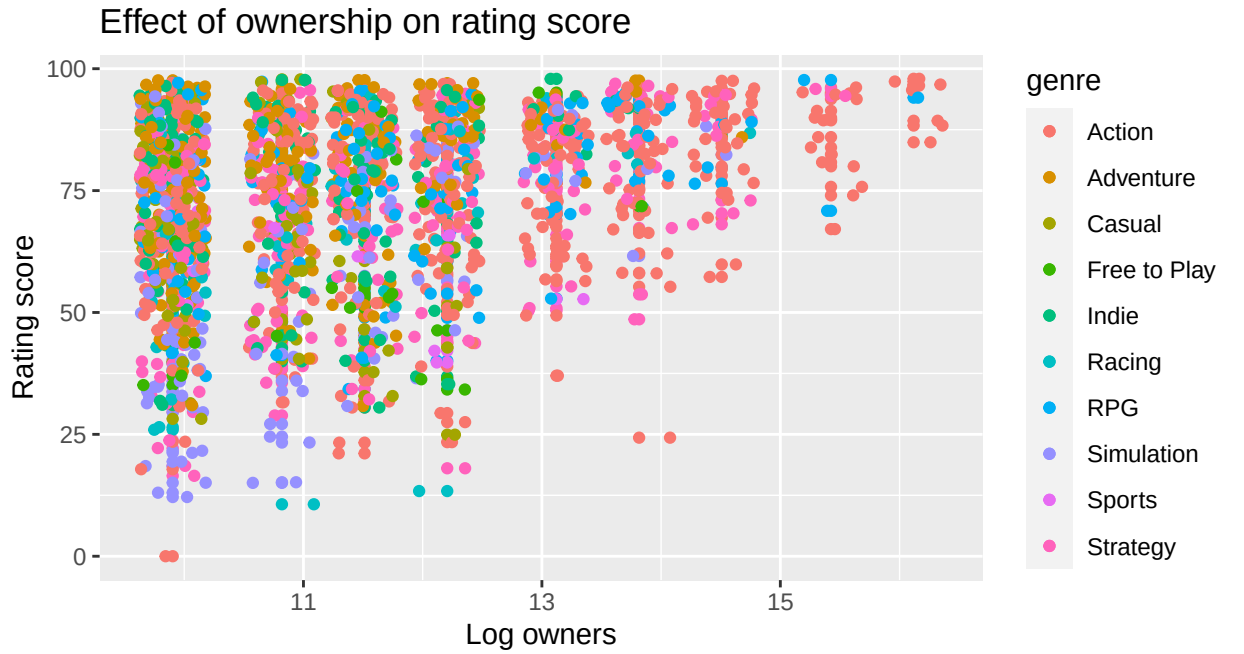


Figure 2: the effect of ownership on rating score, grouped by genre. Games with less owners have greater variation in their rating score while games with more owners tend to be rated highly with less variation. We also see that a majority of our data have relatively less owners.

Results

Our linear mixed model indicates a game with a price and ownership of 0 will have a rating score of 68 points. A dollar increase in price corresponds with a 0.2 increase in rating score. An increase of one owner

in a game has no corresponding change in rating score. From the figures below, we can see the different fits from complete pooling to no pooling. Overall, when we consider the multilevel linear model, there are random effects between genres that do not correspond with the trend of the overall data.

Model Checking

The distribution of our residuals centers around 0, however the variation exceeds our -2 to 2 standard error band.(Appendix A)

Figure 3: Price linear fit with complete pooling

Figure 4: Price linear multilevel fit with no pooling

Discussion

When predicting rating score as a function of price and ownership in our mixed linear model, there is zero effect of ownership on the rating score. Some genres, like Sports, have a negative relationship between rating score and our predictors. This means that as Sports games become more expensive and accumulate more players, their rating score decreases. For price, there is an effect size of 0.2 on a game's rating score.

The no relationship between ownership and rating score is unexpected, but understandable given the large variability of rating score in games with low ownership. The errors are too large to suggest a linear relationship between ownership and rating score, despite expensive games having positive ratings. (Appendix B and C)

The relationship between price and rating score is expected as we hypothesized that more expensive games tend to produce higher quality games and achieve better ratings.

Some limitations to this dataset in fitting our model is the low frequency of more expensive titles and low frequency of games with large number of owners. The variation of data appears to be heteroskedastic and the assumptions of a linear model could be violated. Additionally the genre "Free to Play" has no data points beyond a price of zero so the use of a linear model does not provide any meaningful insights to the genre. Some recommendations for a future analysis are to develop the partial to no pooling model as there are closer similarities in certain genres than others. For example, "Action" and "Adventure" are contextually more similar than "Simulation" is to "Sports".

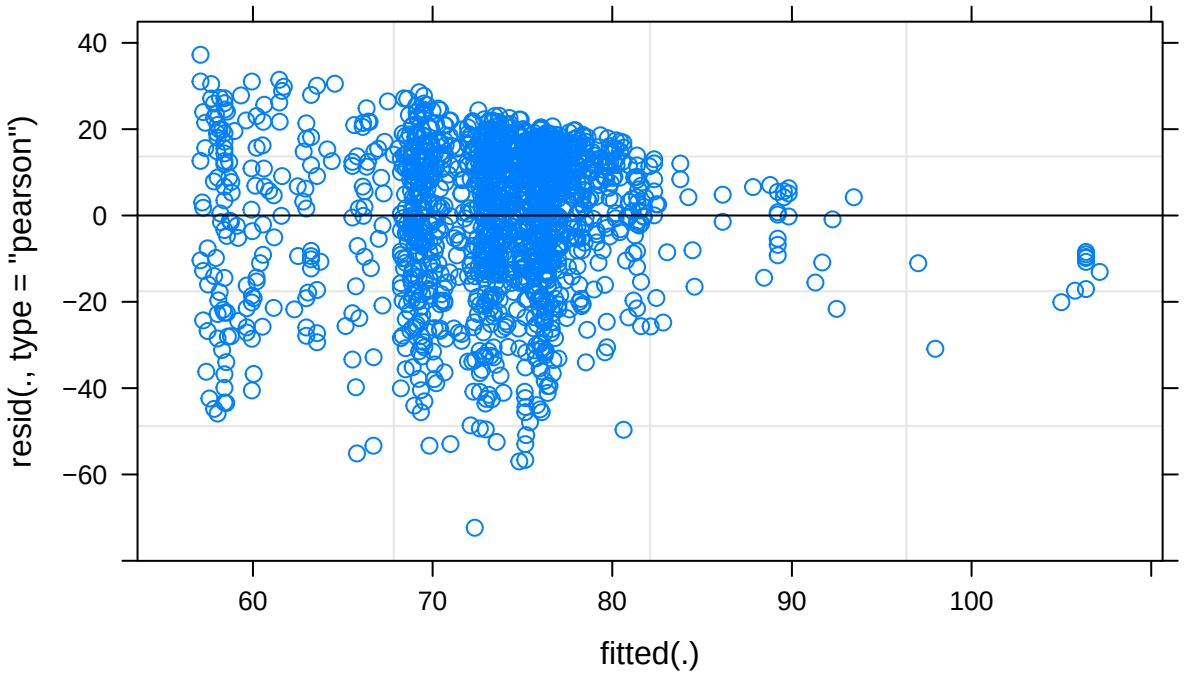
Bibliography

Bates, Douglas; Maechler, Martin; Bloker, Ben; Walker, Steve (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.

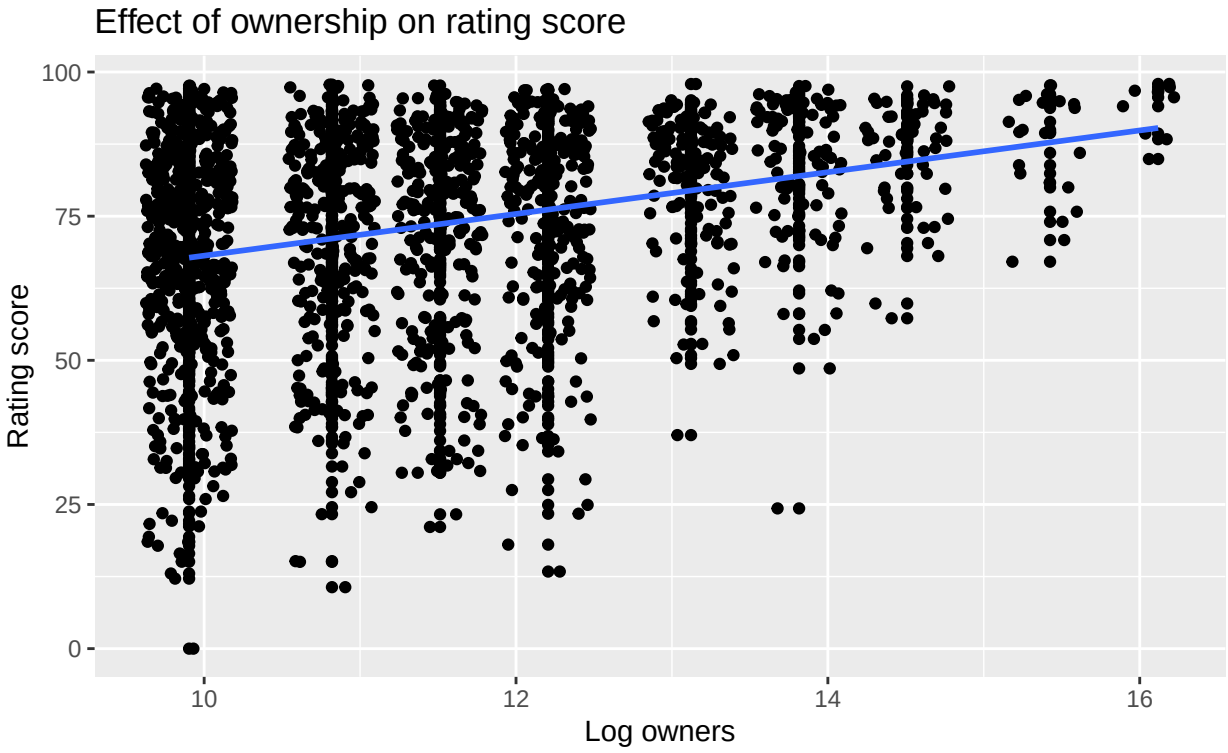
Davis, Nik (2019). Steam Data Exploration in Python. https://nik-davis.github.io/posts/2019/steam-data-exploration/

SteamDB Team (2017). Introducing Steam Database's new rating algorithm. https://steamdb.info/blog/steamdb-rating/

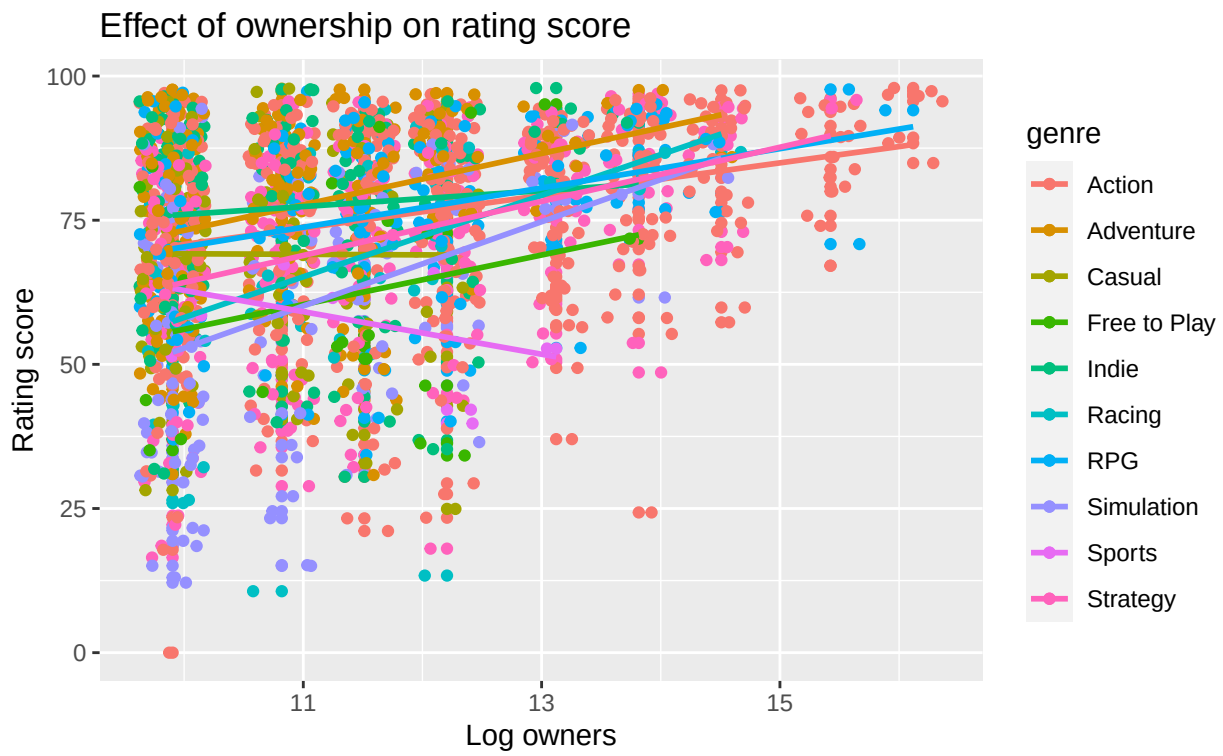
Appendix



Appendix A: Residual Plot of multilevel linear model



Appendix B: Owner linear fit with complete pooling



Appendix C: Owner linear multilevel fit with no pooling