Final Project

Tao Guo

2022-11-29

Abstract

Steam is a digital distribution platform for video games that is developed and operated by Valve Corporation. It is one of the largest game platforms in the PC gaming market, offering a wide variety of games from many different publishers.

Price is an important factor for many game players when deciding whether or not to purchase a game. However, there is no direct relationship between price and game quality, and many lower-priced games can have impressive quality. As a result, game players may be interested in what factors influence the price of games on Steam.

One way to study this question is to use a multilevel model, which allows us to account for the hierarchical structure of the data. In this case, the publisher of the game could be considered a group-level variable, with individual games as the lower level. The multilevel model could then be used to analyze the relationship between game price and other factors, such as release date and rating.

The results of the multilevel model may show that release date and rating are significant factors related to game price. Additionally, the model may reveal that among the top 30 publishers, Square Enix has the highest average game price. However, it is important to note that these results are specific to the data and model used in this study, and may not generalize to other data sets or models.

Introduction

It is true that rating is often used as a standard for evaluating the quality of a game, and higher-rated games are generally considered to be of higher quality. However, it is also true that there is no direct relationship between price and game quality, and that some good games can be inexpensive while some poor-quality games can be expensive. This can be partially attributed to the subjectivity of game ratings, as different players may have different standards for evaluating a game's quality.

In order to better understand the relationship between game rating, price, and other factors, it may be useful to use a multilevel model. This type of model allows researchers to account for both fixed effects, such as the publisher of a game, and random effects, such as the individual player's standards for evaluating a game. By using a multilevel model, researchers can better understand the factors that influence game ratings and prices and identify any potential patterns or trends.

For example, the multilevel model may reveal that the publisher of a game is a significant predictor of its price, with some publishers having higher average prices than others. In particular, the model may show that Yo Yo Games Ltd. has an average game price of 154 pounds, while the average game price for many other game companies is only 2 pounds. These results can provide insight into the factors that influence game prices and help game players make informed decisions about which games to purchase.

Method

Data Processing

I found the data from Kaggle website (https://www.kaggle.com/datasets/nikdavis/steam-store-games).

This data set contain the steam game information from 1997 to 2019. This data set dovetail describe the several information of game, such as ID, name, developer, publisher, rating, etc.

column names	explanation
appid	Unique identifier for each title
name	Title of app (game)
release_date	Release date of game
developer	Name (or names) of developer(s)
publisher	Name (or names) of publisher(s)
platforms	supported platforms. windows;mac;linux
required_age	Minimum required age, 0 are unrated
categories	game categories,single-player;multi-player
genres	game genres, e.g. action;adventure
achievements	Number of in-games achievements
positive_ratings	Number of positive ratings
negative_ratings	Number of negative ratings
average_playtime	Average user playtime
owners	Estimated number of owners. like 20000-50000)
price	Current full price of title in GBP (pounds)

Exploratory Data Analysis

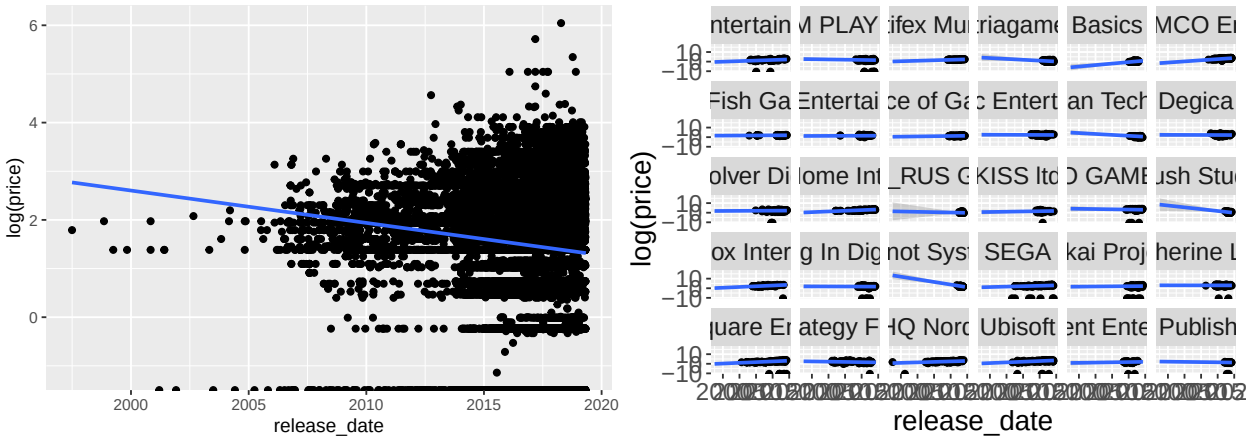


Figure 1: relationship between price and release date

Figure 1 illustrates the relationship between price and release date. The left figure shows as the release date increase, the price of the game decrease totally. In right figure show the relationship between the price of the top 30 sales publishers and their game release date. Meanwhile, the left figure also indicates changing to a different publisher, the slope of lines almost does not change for the most plot, so the publisher maybe not have the random effect of the release date.

Figure 2 shows the positive correlation between positive ratings and price. The figure 4 shows different publishers have different slopes. Therefore, I will use publishers as random effect of positive ratings. This

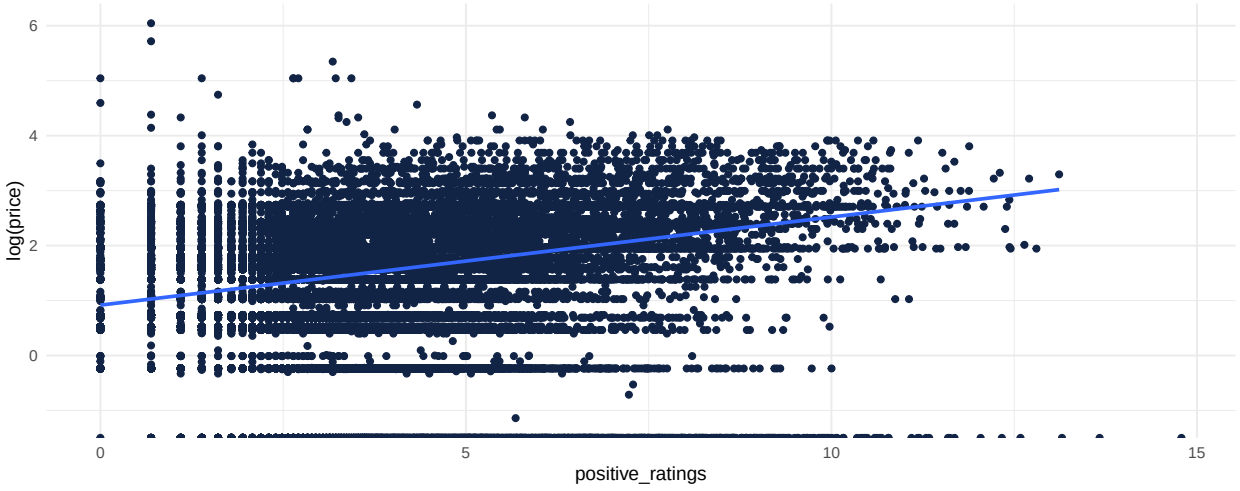


Figure 2: relationship between price and positive rating

result is relatively make sense because better game should have better price, and player may have different standard for rating different publishers.

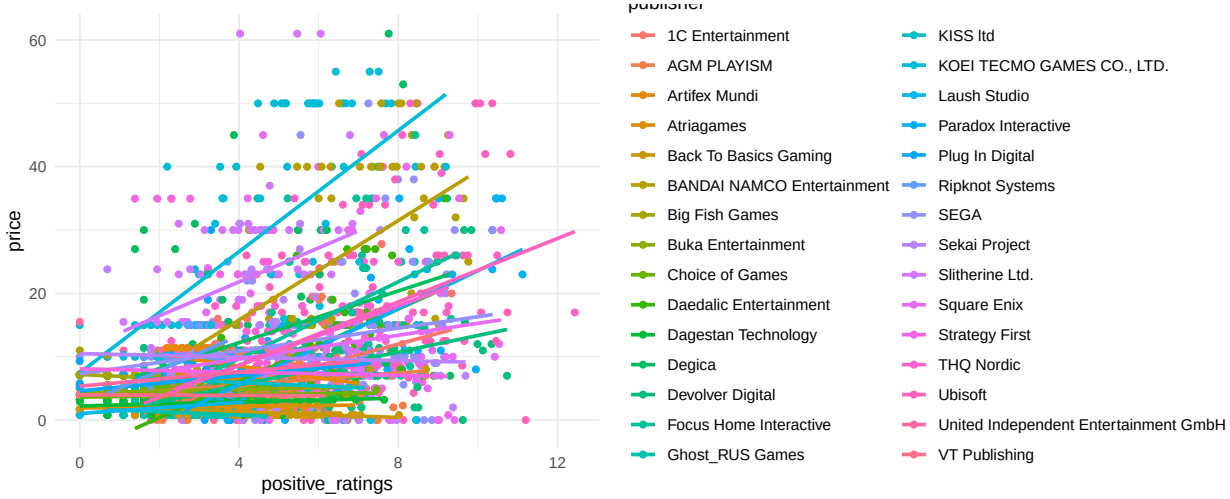


Figure 3: relationship between positive_ratings and price

The result from figure 3 is a little wired because, in a common idea, the worst game should not be worth to higher price, but facts may come from a different direction. For instance, this is expensive and worst so players give more negative feedback to this game. The left figure in figure 5 shows both intercept and slop are different compared to different publishers. Therefore, the publishers are also the random effect by negative rating

Figure 4 shows the positive relationship between average playing time and price. On my opinion, more playing time for game mean this have more content than others, the company may also have higher cost to provide a longer game experience for player, as the cost increase, the price of game also increase. Figure 5 expresses obious vary intercept for different publishers, so the publisher also will the random effect for average playtime in my model.

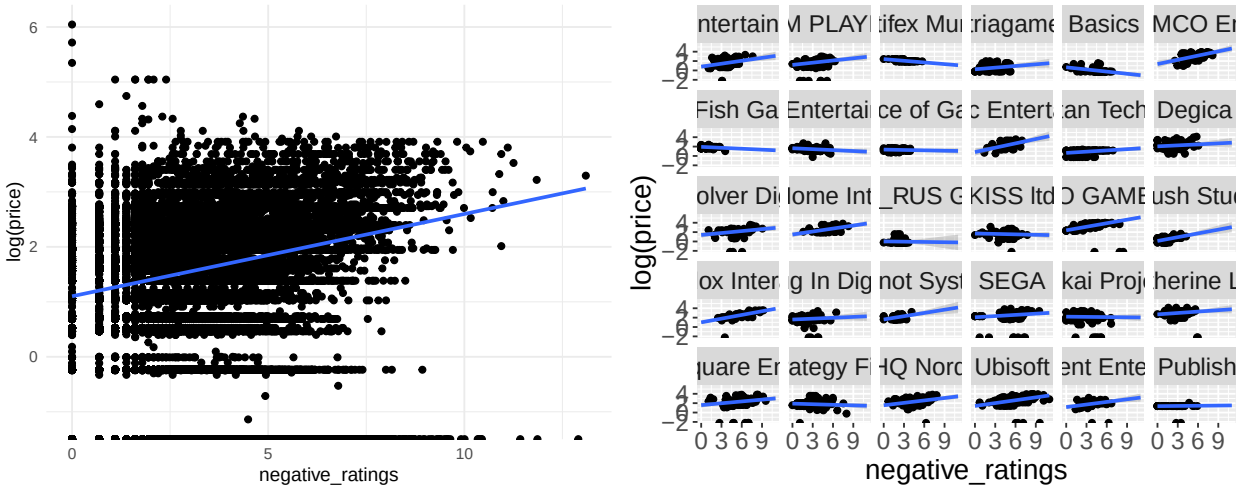


Figure 4: relationship between negative_ratings and price

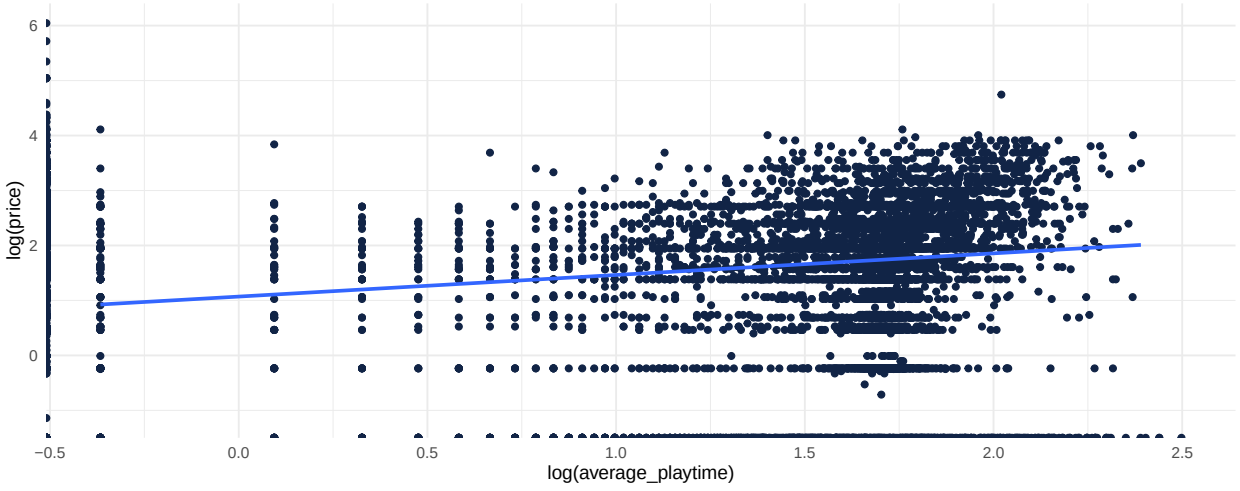


Figure 5: relationship between average play time and price

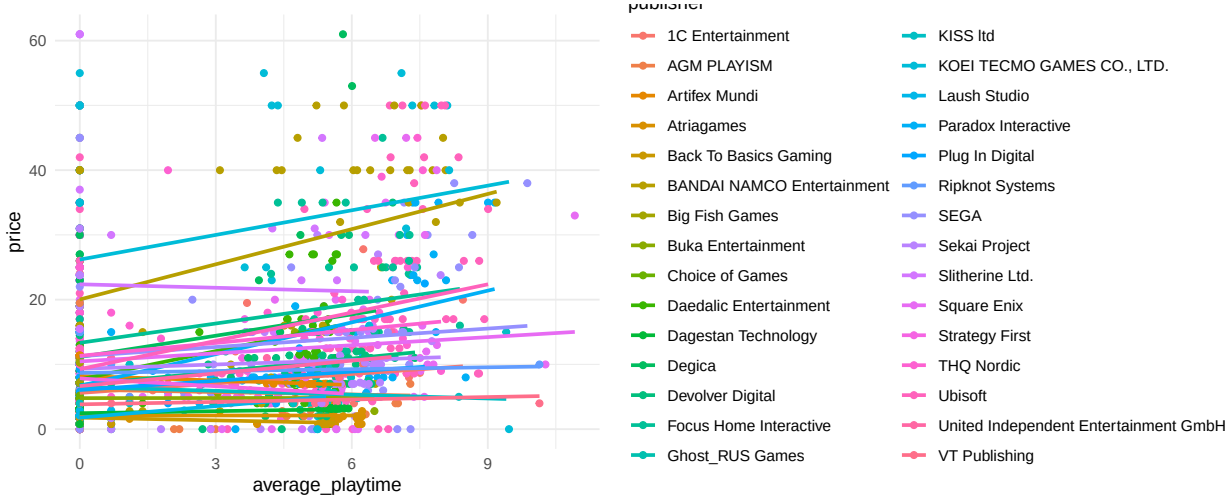


Figure 6: relationship between average play time and price

Model Fitting

Since different publishers have quite large impact on the model,I determine to use multilevel model to fit steam data sets. For variables, this fixed effect is release date, positive ratings, negative ratings, and average playing time. The publishers is the random effect. Here is the model:

$$price = (1|publisher) + release\ date + average\ playtime + positive\ ratings + negative\ ratings + (0 + average\ playtime + positive\ ratings + negative\ ratings|publisher)$$

This Table is the summary of fixed effect of my model. All variables are considered to significant at $\alpha = 0.5$. In order to more clear show the fixed to fixed effect, the next figure is also correspond to summary table.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-34.08	1.208	23730	-28.2	0.00 ***
release_date	0.0021	6.8e-05	23630	31.3	0.00 ***
log_average_playtime	-0.05	0.025	2145	-1.96	0.0497 *
log_positive_ratings	0.30	0.044	13510	14.68	0.00 ***
log_negative_ratings	0.53	0.05	23640	4.975	0.00 ***

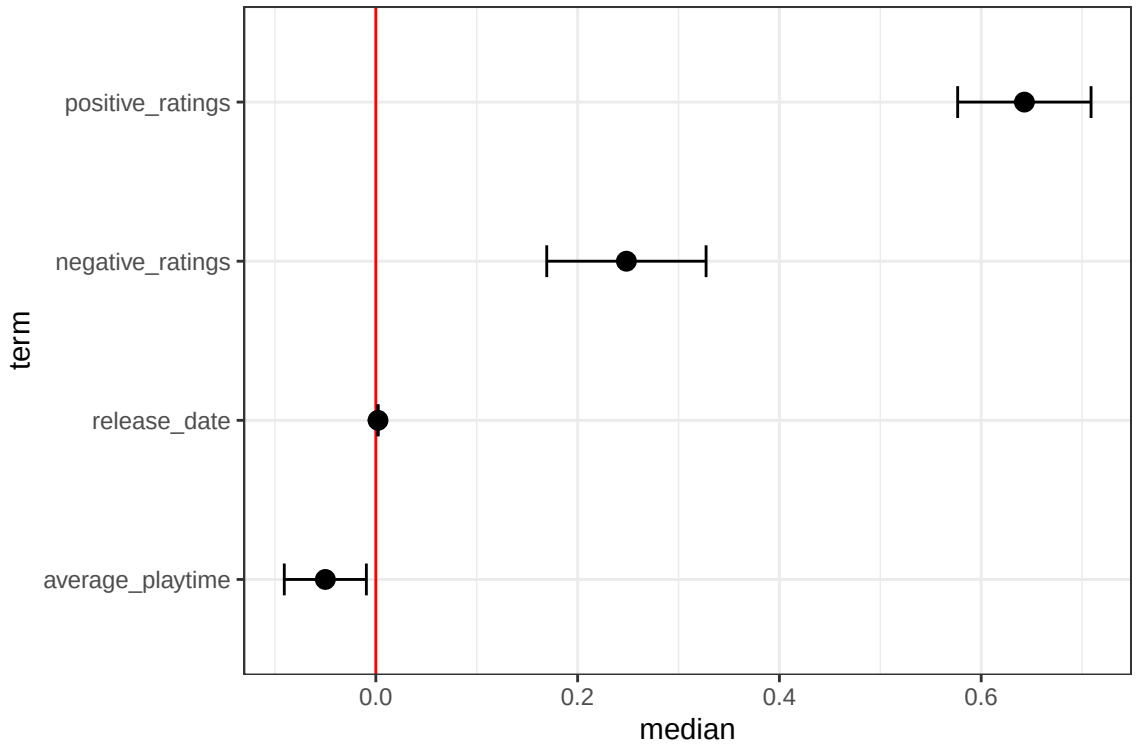


Figure 7: Fixed Effect of Steam Model

The next figure shows the random effect of top 30 publishers to each variables. From this figure, price baseline of some publisher is higher than others. Meanwhile, I found the publishers did not effect so much on average playing time. Fot positive and negative rating, the random effect work well, which obviously shows some publishers have higher positive ratings than others.

Result

In order to interpret my model, lets use one of famous publisher Square Enix as example. Firstly, I build up following formula of fixed effect

$$price = -34.08 + 0.0021 \times release\ date - 0.05 \times log(1 + average\ playtime) + 0.643 \times log(positive\ ratings)$$
$$+0.247 \times log(negative\ ratings)$$

After adding the random effect of Square Enix, the formula is:

$$price = -32.1 + 0.0021 \times release\ date - 0.071 \times log(1 + average\ playtime) + 0.021 \times log(positive\ ratings)$$
$$+0.957 \times log(negative\ ratings)$$

In this formula, only average playtime are negative. Other parameters are positive. This parameter of release date indicates the price of Square Enix's game increased by 0.0021 pounds every day, which does not correspond to my EDA plot, but it is reasonable because this parameter may reflect the inflation between 1994 to 2019. The estimator of average playtime means every unit increase on the log scale of average playtime, the price will drop by 0.071 pounds. This result is different from my previous thought. In general, more playtime means the game has more content, which affects the cost of the game. I also can use different aspects to explain this result. people may easily give positive ratings when they only play for a few hours,

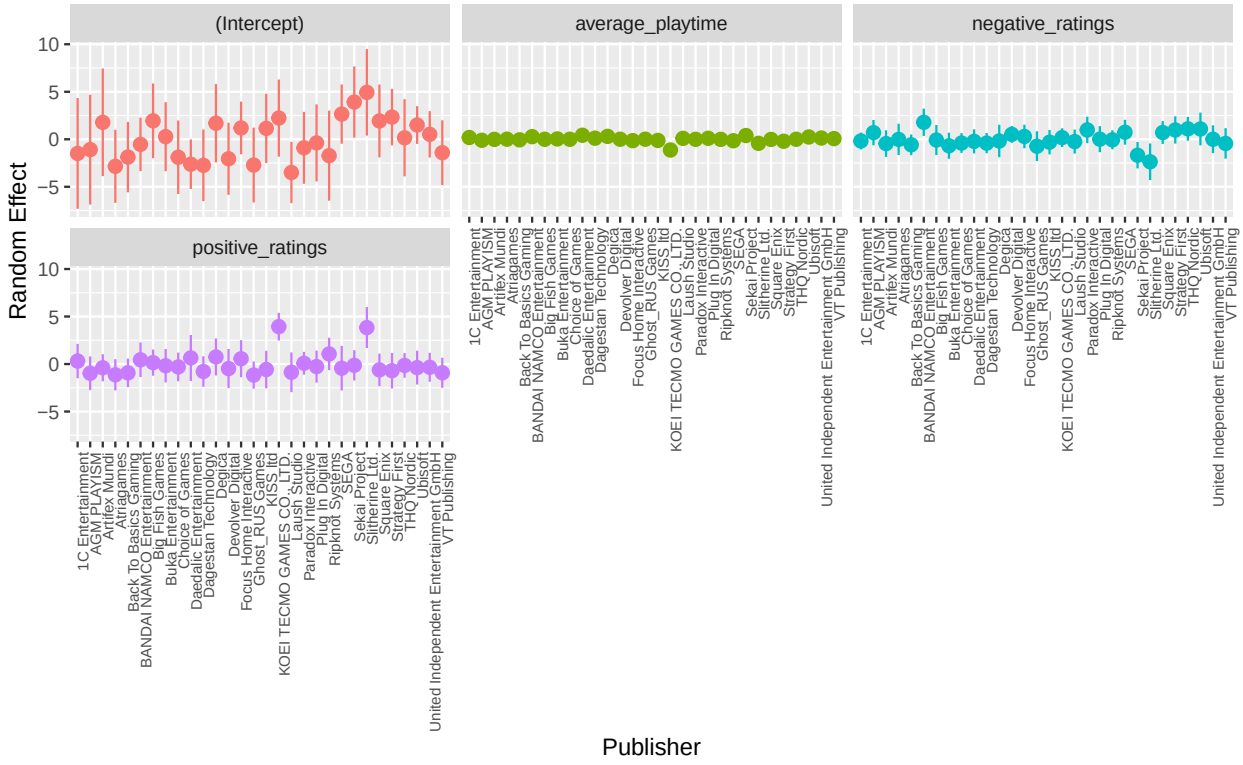


Figure 8: Random Effect of Model

but when they play for a long time, they will have a more comprehensive understanding of the game to change their opinion to negative. The parameter of positive rating means one unit increase in the log scale of positive rating the price will increase by 0.021 pounds. The parameter of negative rating means one unit increase in the log scale of negative rating the price will increase by 0.957 pounds. These results are the same to my original thought, and the negative rating has more effects on price than positive ratings

Model Checking

The next figure is th residual plot and qq plot. The mean of residuals is close to 0, but the variance of residuals increase, which indicate some correlation I do not find from data. For qq plot, the middle parts of plot is closer to normality, but two two tail diverge from normality. Based on the density plot, my model is over dispersion at 0 to 10, which overestimates price of game.

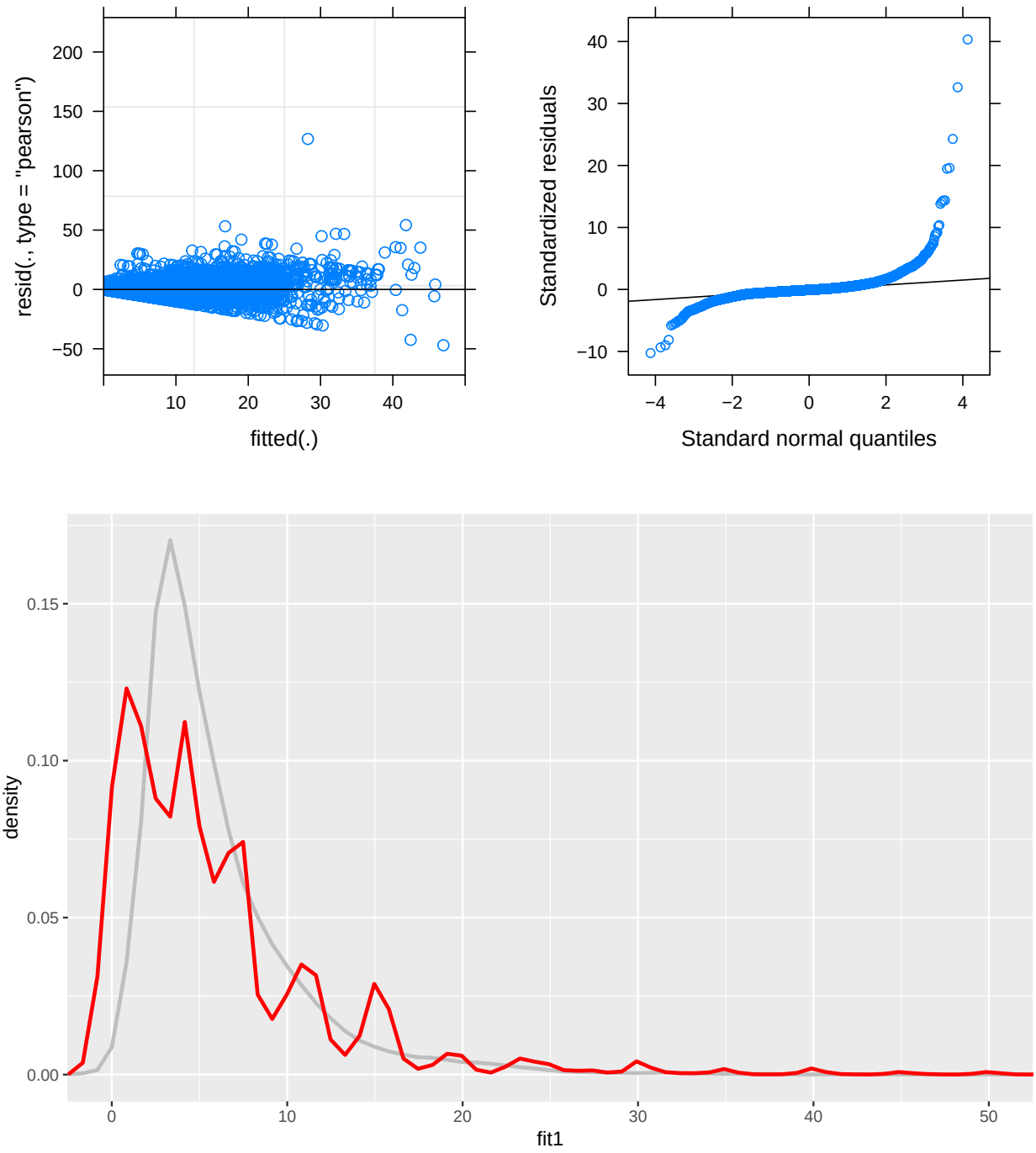


Figure 9: Model Fitting check

Discussion

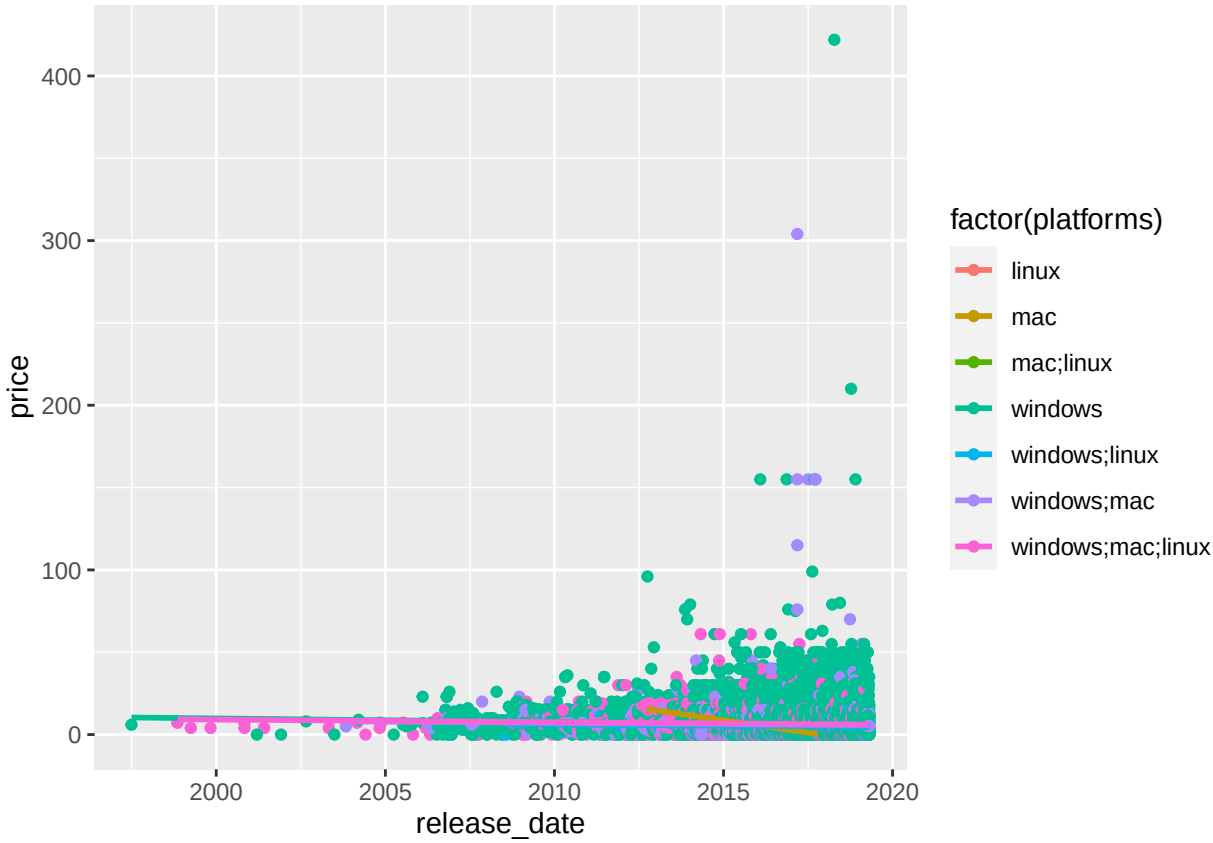
In my report, I would like to figure out what factors may influence the price of a game. I built this multilevel model. From the fixed effect table, every variable is significant to the game price. Although the random effect publishers are not related to each variable, which proves different publishers have different price baselines

and players also have different attributes for different publishers. However, there are many limitations to my model. Firstly, I should be more careful to select data sets because many variables are useless to my analysis. Due to the model checking, I should consider more about the random effect, I may ignore some correlations in the data sets. For the next steps, I will carefully select the data and considering the potential for random effects can improve the accuracy and reliability of your model. Additionally, incorporating additional variables that may be relevant to the game price can provide a more complete picture of the factors that influence the price..

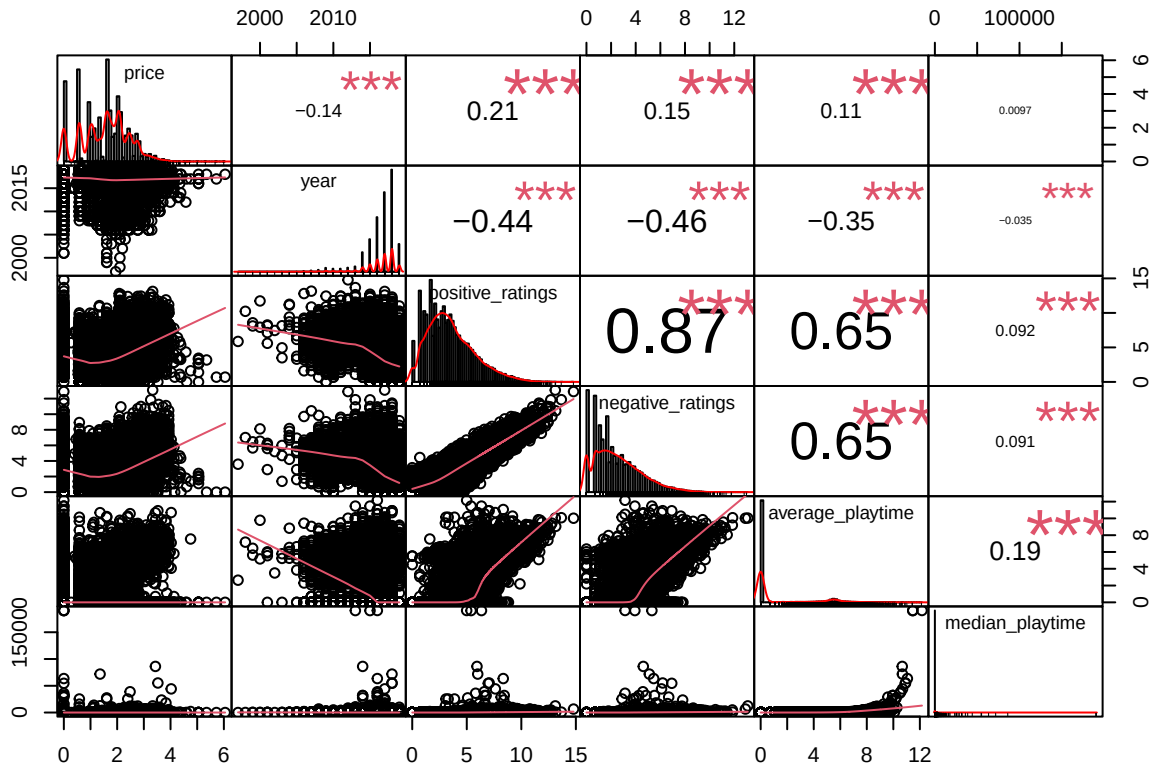
Reference

Frank Zhang https://github.com/BU-Franky/MA678_midterm

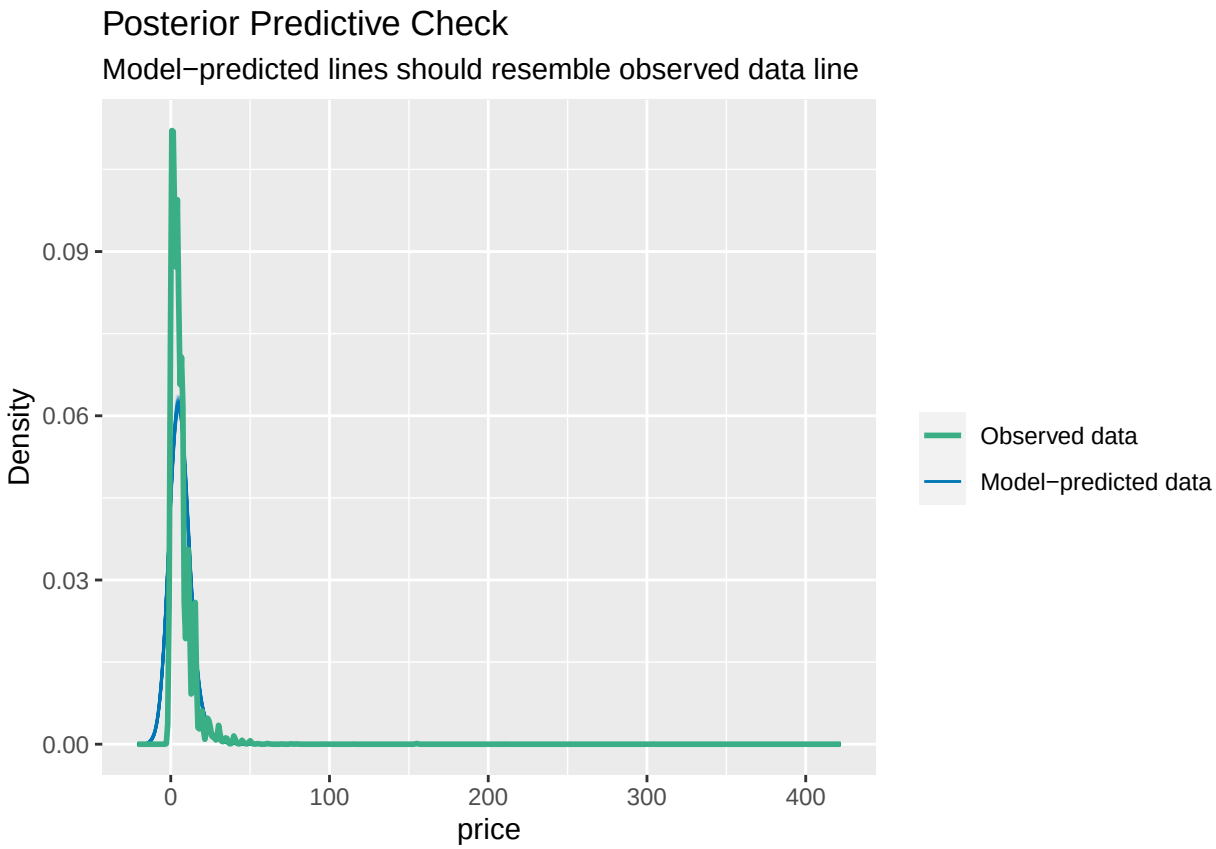
Appendix

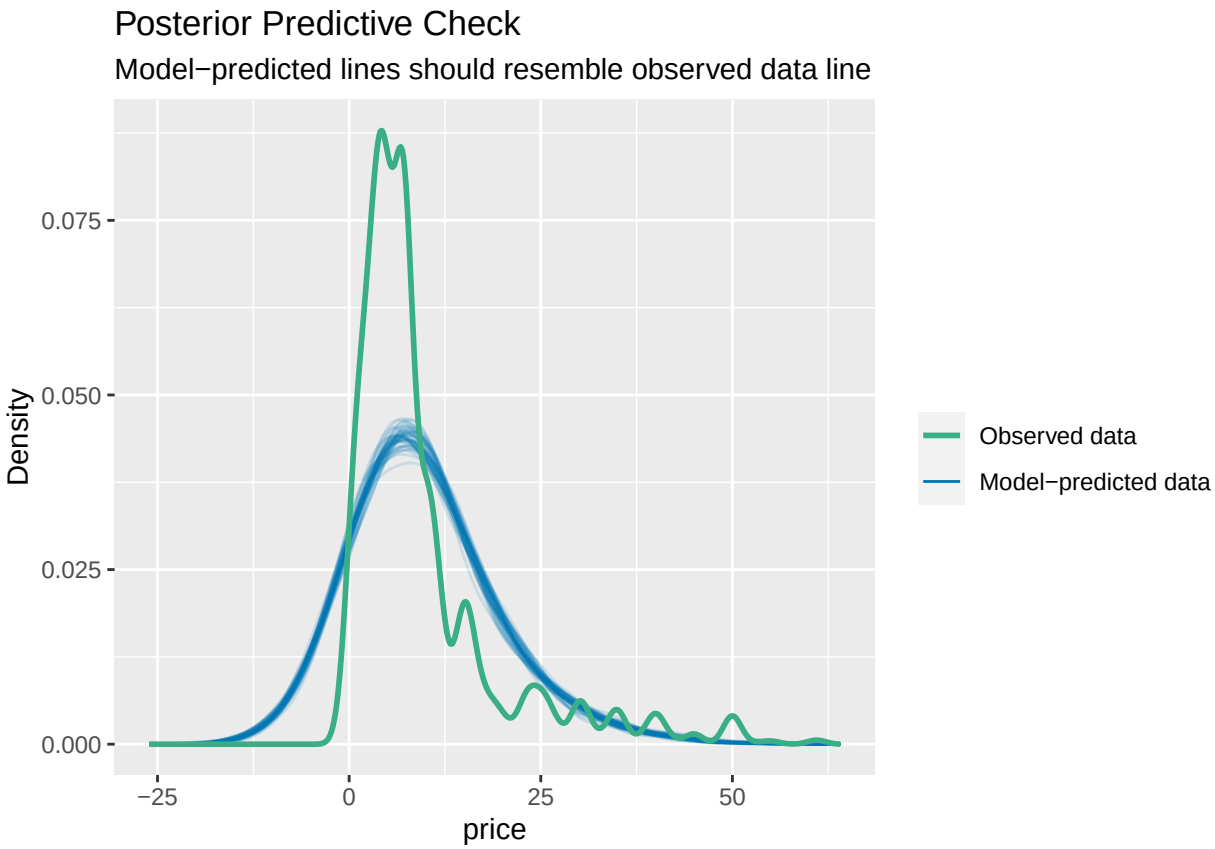



```
b <- steam1
b$price <- log(1+b$price)
b %<>% separate(release_date, into=c("year","month","day"), sep = "-")
b$year <- as.numeric(b$year)
chart.Correlation(b[, c(23,3,18:21)], histogram=TRUE, pch=20)
```




```
## Warning: Maximum value of original data is not included in the
##   replicated data.
##   Model may not capture the variation of the data.
```


Posterior Predictive Check
Model−predicted lines should resemble observed data line
Density
0.09
0.06
0.03
0.00
0
100
200
300
400
price
Observed data
Model−predicted data


Posterior Predictive Check
Model−predicted lines should resemble observed data line
Density
0.075
0.050
0.025
0.000
−25
0
25
50
price
Observed data
Model−predicted data