Project 4

Farida Sondo

2023-05-05

R Markdown

Linear Regression

1. Data set story

This data was collected from the Steam reviews of Elden Ring from January 29th, 2023, to March 7th, 2023. Steam is a digital video game distribution service created by Valve in 2003. It was originally created as a way for Valve to send automatic updates for their own games but began selling third-party software in 2005. Since then, Steam has become the most popular platform to buy and download games digitally, giving users the ability to rate and review games, create downloadable content for certain games, and make discussion boards for questions and guides. Elden Ring is an action-adventure role-playing game (RPG) developed by FromSoftware and distributed by Bandai-Namco. FromSoftware and Elden Ring's writer, Hidetaka Miyazaki, have previously produced games such as Demon's Souls, the Dark Souls series, and Bloodborne, all of which are the same style of game as Elden Ring, commonly called "Souls-like" games. Elden Ring was critically acclaimed, with it being nominated for and winning multiple game of the year awards. This dataset contains variables such as:

- The content of the review
- The date and time the review was posted
- Whether the review was positive or negative
- The hours played by the reviewer at the time of the review and overall
- How many games the reviewer owns
- How many reviews the reviewer has written
- How many "likes" the review has received
- How many comments the review has received • Whether the game was received for free.

```r
library(readr)
elden_ring_steam_reviews <- read_csv("C:/Users/somde/Downloads/elden_ring_steam_reviews.csv")
```

```
## Rows: 9794 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): language, review
## dbl  (8): id, votes_up, comment_count, author_num_games_owned, author_num_re...
## lgl  (4): voted_up, steam_purchase, recieved_for_free, written_during_early_...
## dttm (2): created, author_last_played
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(elden_ring_steam_reviews)
```

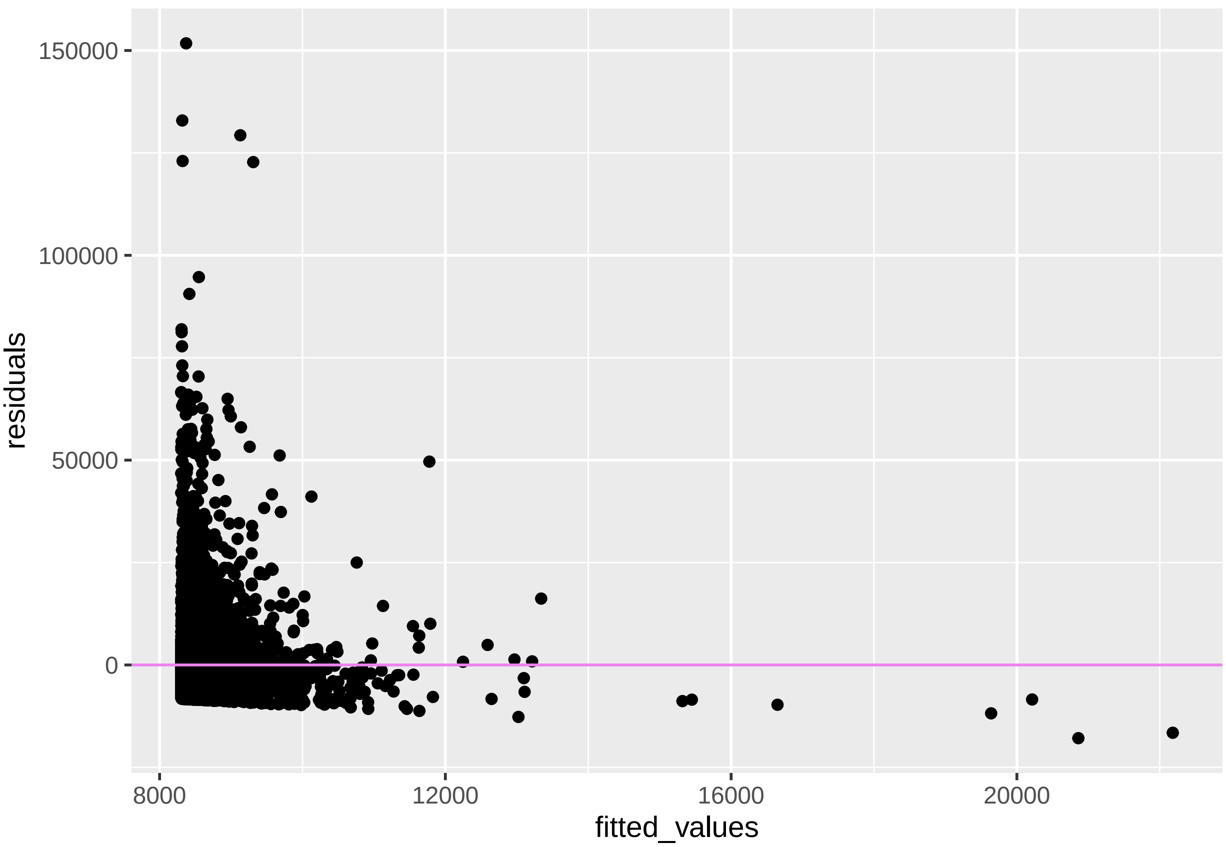
3. I choose “author_num_games_owned” as my predictor variable (x) and “author_playtime_forever” as the response variable (y). The reason why I chose “author_num_games_owned” as the predictor variable (x) is because it is plausible that the number of games the person own could be possibly related to the amount of time a person spends playing games. People who own more games may be more likely to spend more time playing games. People who own more games may have a wider range of options to choose from and therefore spend more time playing.

4. Two diagnostic plots

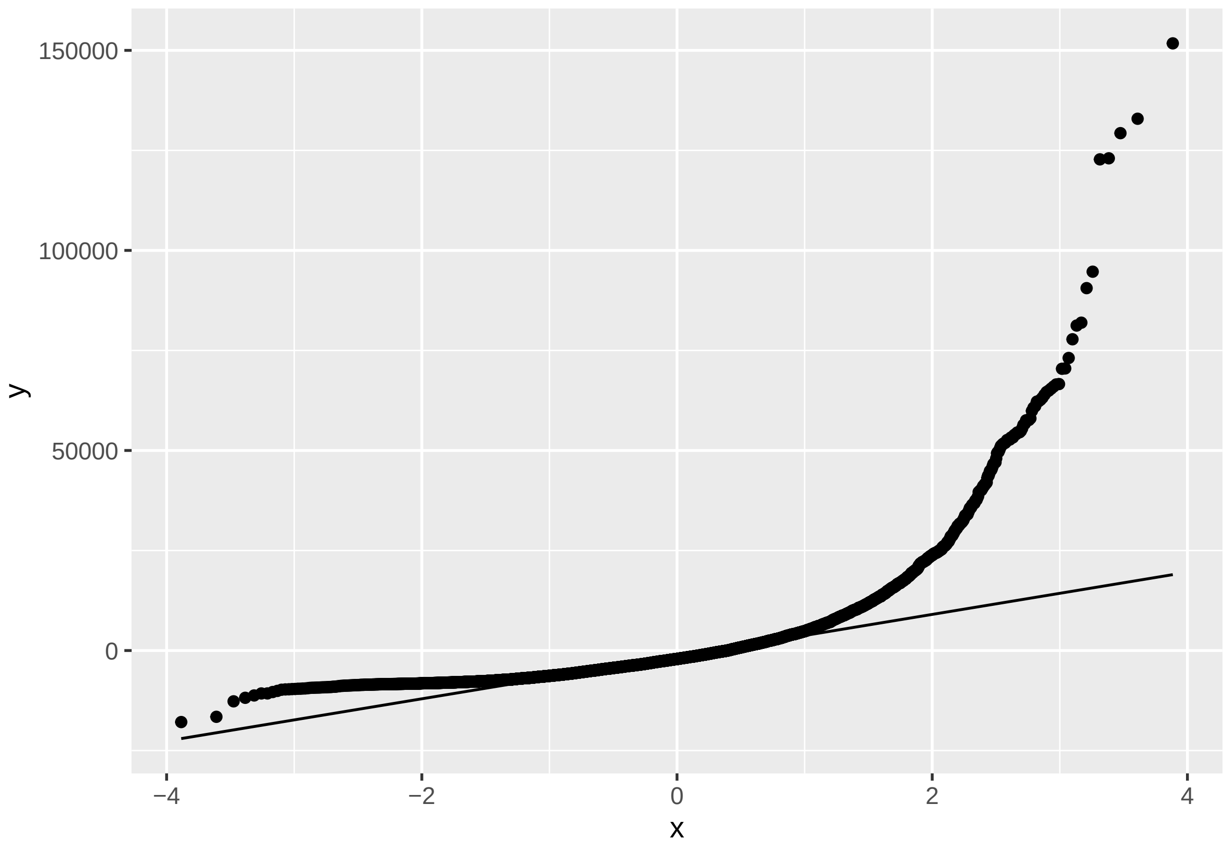
```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v dplyr   1.0.10
## v tibble  3.1.8      v stringr 1.5.0
## v tidyr   1.2.1      v forcats 0.5.2
## v purrr   1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
new_lm <- lm(author_playtime_forever ~ author_num_games_owned, data = elden_ring_steam_reviews)


fitted_resid_df <-
  tibble(fitted_values = fitted(new_lm),
         residuals = resid(new_lm))


fitted_resid_df %>%
  ggplot(aes(x = fitted_values, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "violet")
```

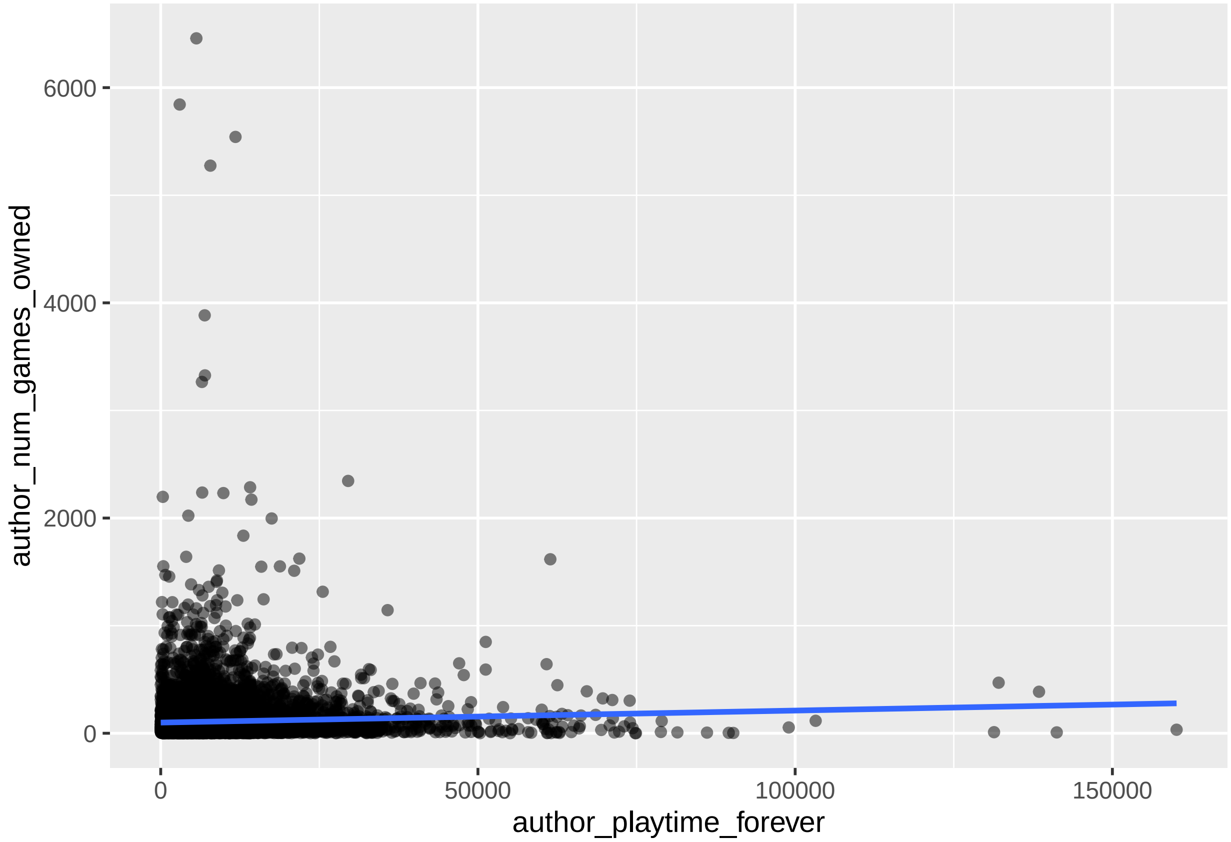


```r
fitted_resid_df %>%
  ggplot(aes(sample = residuals)) +
  geom_qq() +
  geom_qq_line()
```



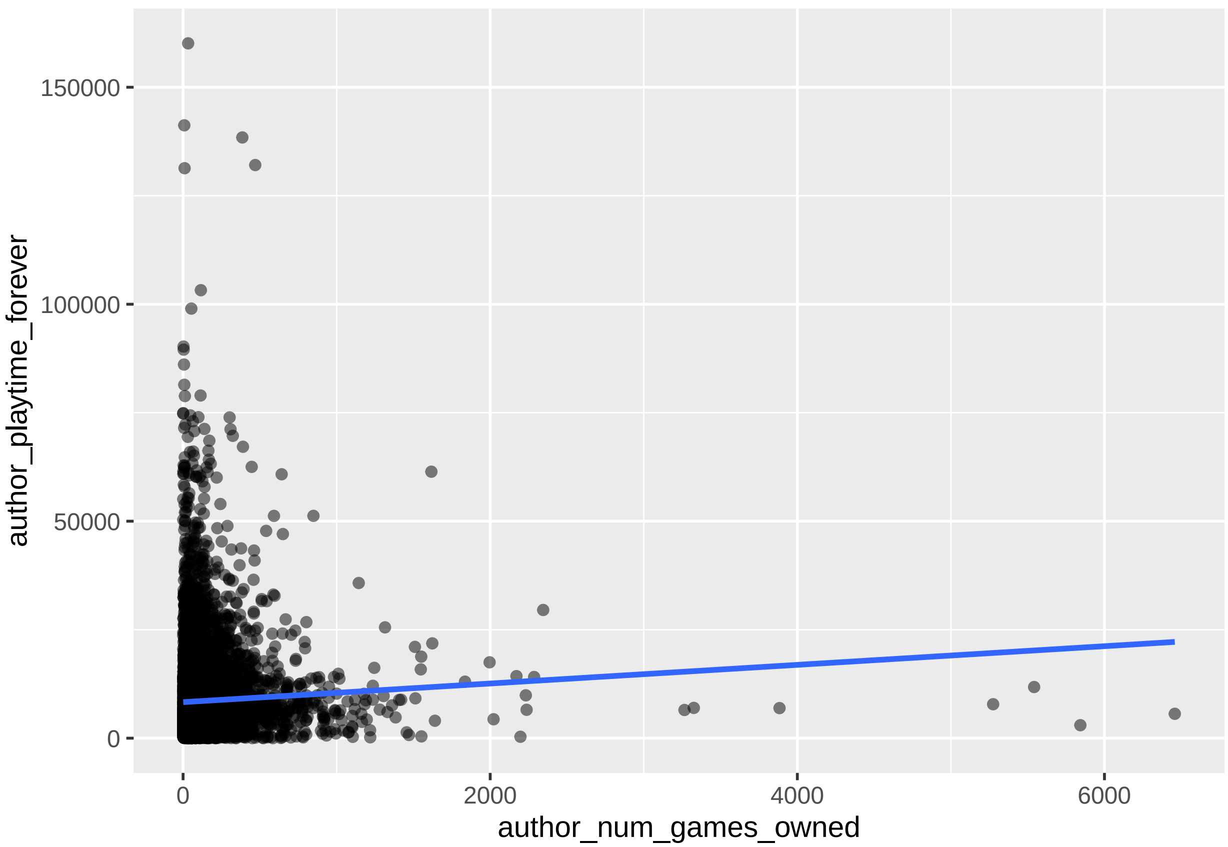
```
elden_ring_steam_reviews %>%
ggplot(aes(x = author_playtime_forever, y = author_num_games_owned)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

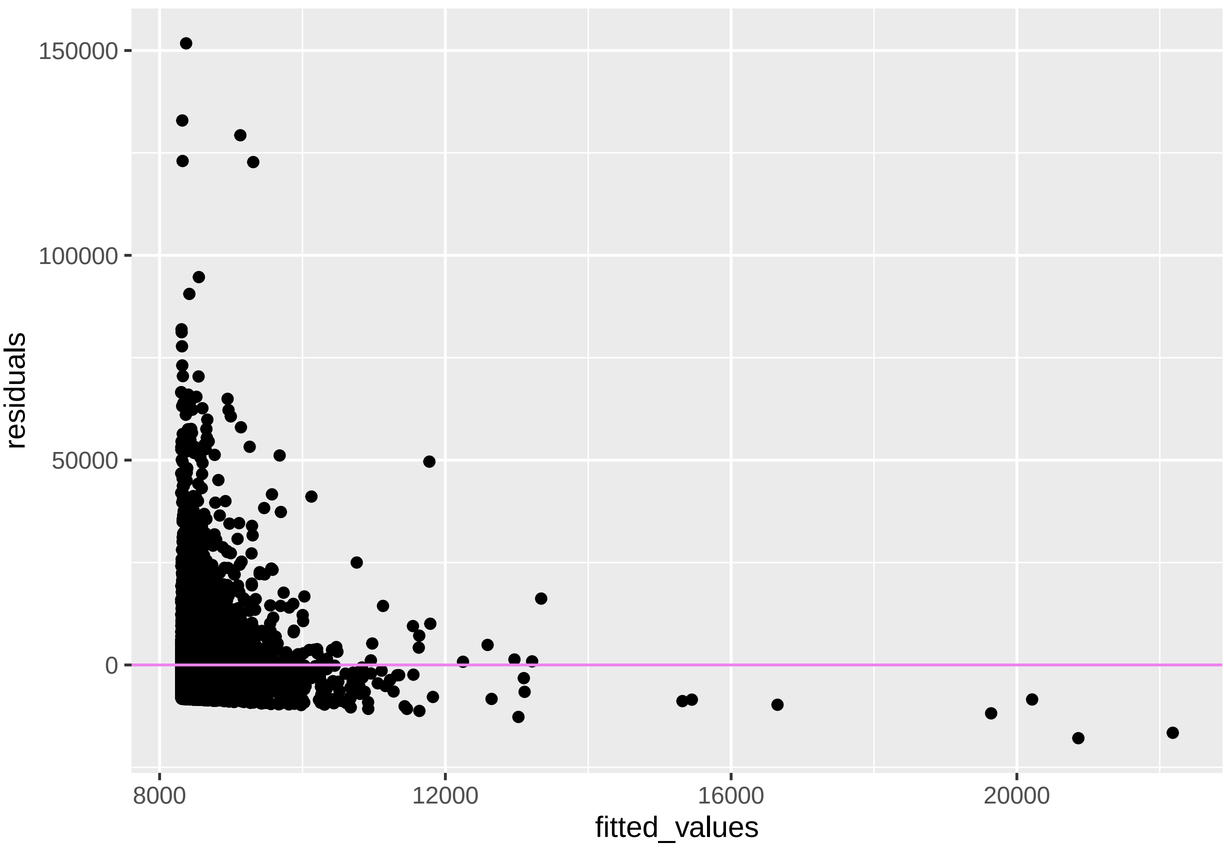


```r
library(tidyverse)

new_lm <- lm(author_playtime_forever ~ author_num_games_owned, data = elden_ring_steam_reviews)


fitted_resid_df <-
  tibble(fitted_values = fitted(new_lm),
         residuals = resid(new_lm))

elden_ring_steam_reviews %>%
ggplot(aes(x = author_num_games_owned, y = author_playtime_forever)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE)
```

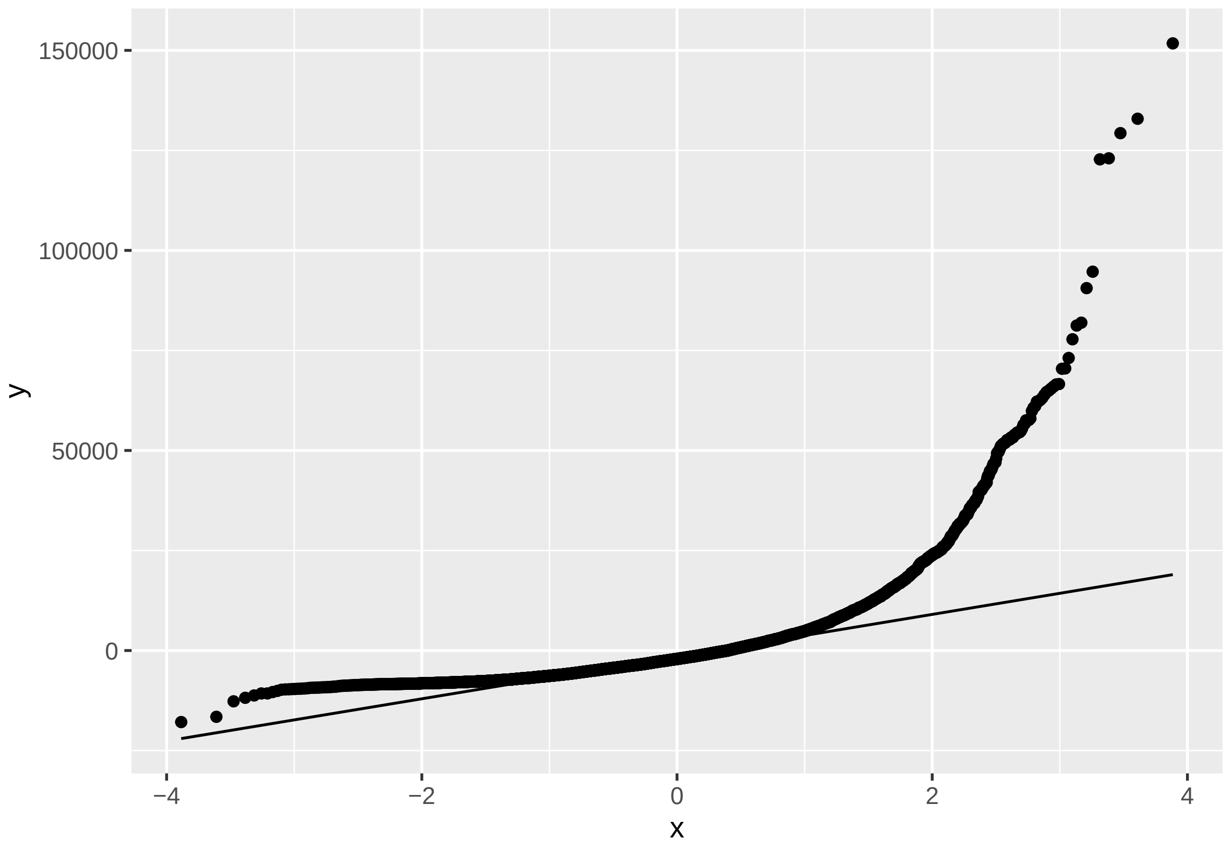
```
## `geom_smooth()` using formula = 'y ~ x'
```



```
fitted_resid_df %>%
  ggplot(aes(x = fitted_values, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "violet")
```



```
fitted_resid_df %>%
  ggplot(aes(sample = residuals)) +
  geom_qq() +
  geom_qq_line()
```



5. Effectiveness of the linear model

It seems that the linear model is not suitable for my data. This is likely due to a number of data points that deviate significantly from the trend line, which suggests that the relationship between the variables being modeled is more complex than a simple linear relationship. Furthermore, the lack of a clear pattern in the trend line suggests that a linear model may not be appropriate for my data. Therefore, it is necessary to explore other types of models that can better capture the complexity of the relationship between the variables.

6. Explanation of my prediction findings

```
new_x <- 500
predicted_y <- predict(new_lm, newdata = data.frame(author_num_games_owned = new_x))
predicted_y
```

```
##        1
## 9375.163
```

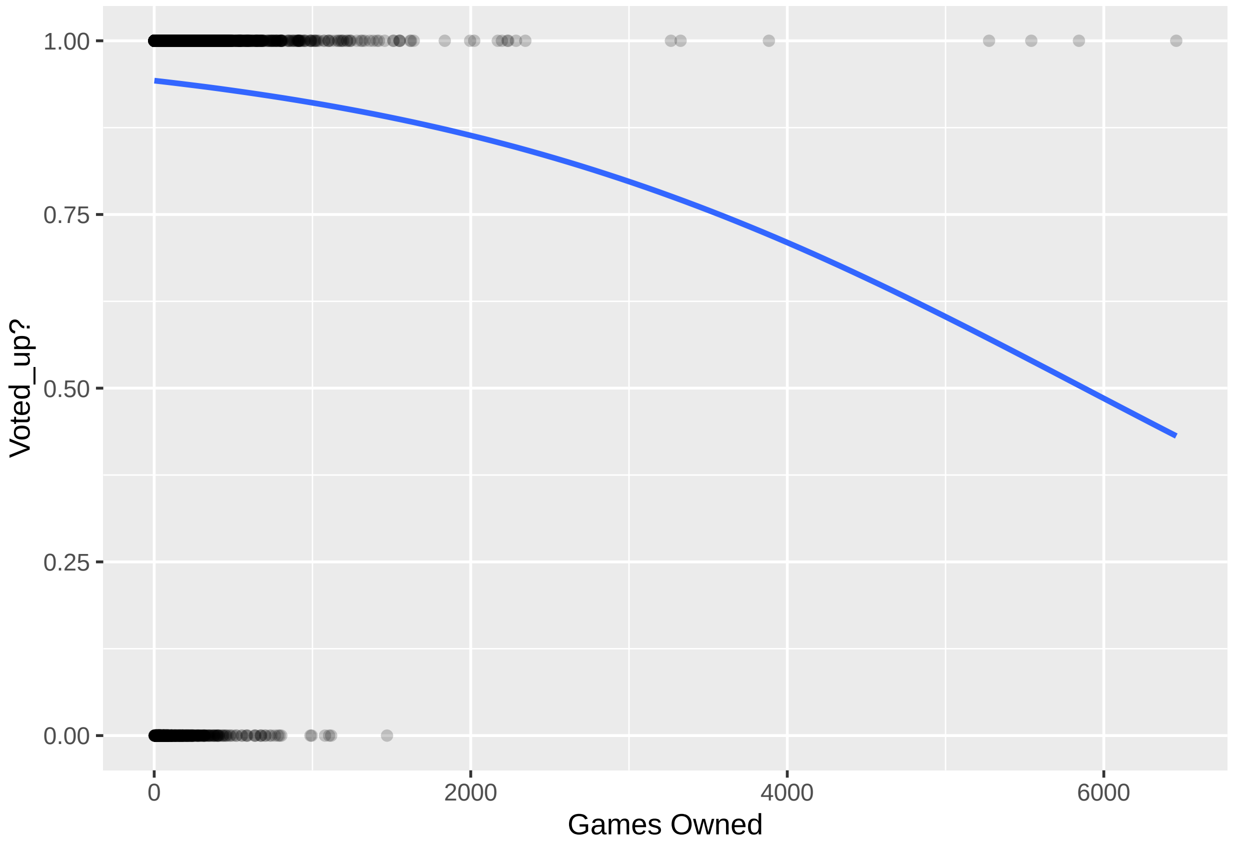
The model use "author_num_games_owned" as a predictor of the game review score or rating, which implies that the number of games owned by the author may have a significant impact on their review or rating of Elden Ring.This code predicts that an author who owns 500 games on Steam would have a author_playtime_forever value of 9375.163.

Logistic Regression

7. For the predictor variable (x), I will use "author_num_games_owned" from my data set, which represents the number of games the author has owned. For the response variable (y), I will use "voted_up," which is a binary variable indicating whether the author has recommended the Elden Ring game. I will a scatter plot of these two variables that includes an S-curve. I chose these variables because I am interested in exploring whether the number of games an author has owned influences their likelihood of recommending the Elden Ring game. I believe that the relationship between these variables could be nonlinear, which is why I chose to include an S-curve in the scatter plot.

```
elden_ring_steam_reviews %>%
  ggplot(aes(x = author_num_games_owned, y = as.numeric(voted_up))) +
  geom_point(alpha = 0.2) +
  labs(x = 'Games Owned', y = 'Voted_up?') +
  geom_smooth(method = 'glm', method.args = list(family = "binomial"), se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



8. The model isn't great due to the concentration of games owned for both positive and negative reviews being lower. There doesn't seem to be a very strong relationship between type of review(positive or negative) and number of games owned, although the model suggests that the more games owned predicts a more likely chance for a negative review.It appears that the logistic regression model is not be the most appropriate fit for the data I chose.

9. Prediction

```
elden_ring_steam_reviews_glm<- glm(voted_up ~ author_num_games_owned, data = elden_ring_steam_reviews, 
elden_ring_steam_reviews_glm %>%
  predict(tibble(author_num_games_owned = 1000), type = 'response')
```

```
##         1
## 0.9107477
```

The predicted probability of the binary outcome variable "voted_up" for the new data point based on the logistic regression model is 0.9107477. This prediction is not accurate as the logistic regression model itself is not.I think there is no clear link between the number of games owned and the type of review.