Final Project

Sebastián Soriano Pérez [ss1072]

12/10/2019

Analyzing Video Game Sales

- **Summary**

By analyzing the data on 869 newborn male babies and their families, a model was created with stepwise selection using BIC as a comparison parameter to interpret and associate the variables that were found to be significant with the response variable of a birth being premature (< 270 days of gestation). Afterwards, the model's accuracy, sensitivity and specificity were compared to a model including the variable mht. The new model improved these values marginally, so it was selected for the data analysis.
The final model estimates that only the variable of mracewhite is significant, but the rest of the mrace variables as well as med, mpregwt_c, smoke, and mht were included because they improve the model overall. The specific coefficient values can be found in the "Model" section.

- **Introduction**

This document presents a model to interpret the impact of several variables on a newborn's chances of being premature. A dataset was analyzed considering the available data in order to find the best model to explain the association between the predictive variables and the response variable through an initial exploratory data analysis (EDA), and later with a stepwise selection in R a logaritmic regression to estimate the probability of being premature. The main focus of this document is to find whether or not smoking during pregnancy had an impact in the chances of having a pre-term birth, and if this chances differ by race.

- **Data**

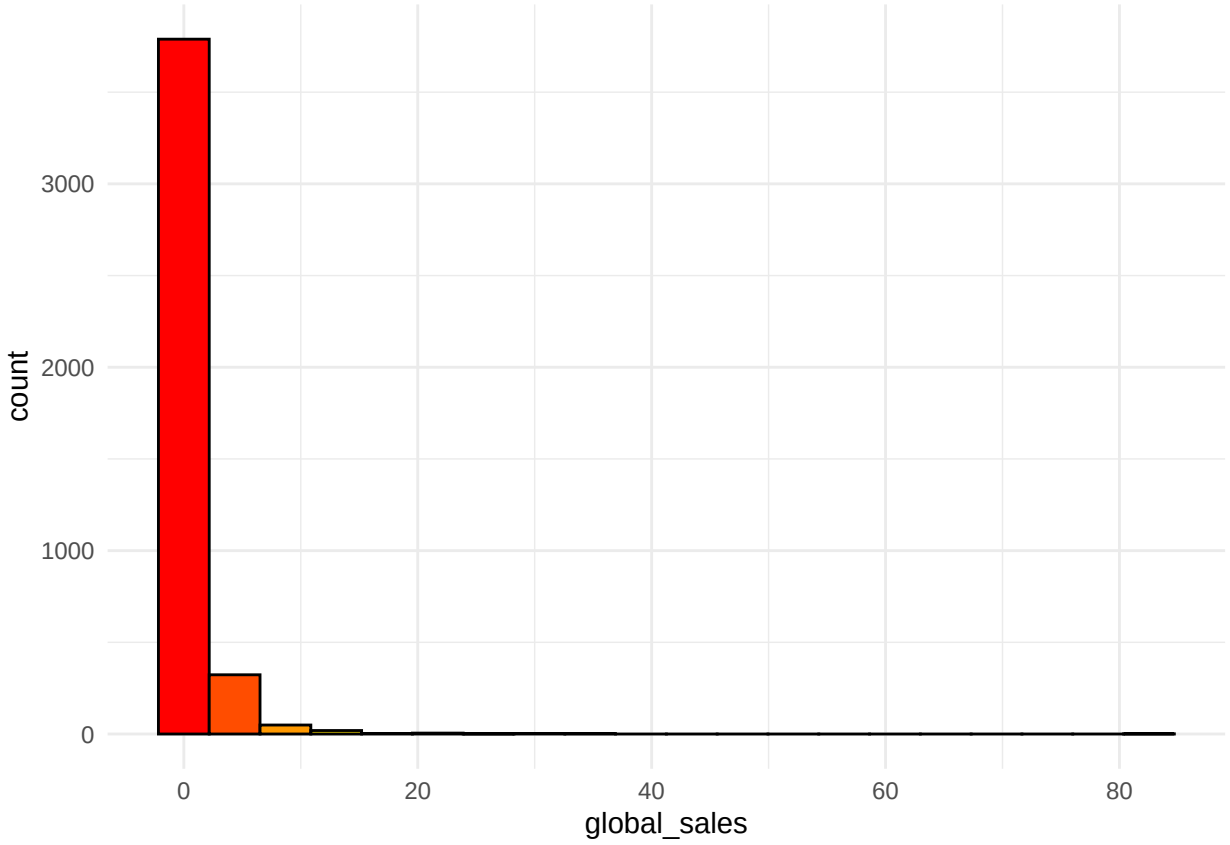
The Child Health and Development Studies research was one of the first to collect data to understand and quantify the risk of smoking during pregnancy to the baby's health. The data was collected from 1960 to 1967, and a subset of that data is being analyzed in this document (the variables related to the father's information are neglected for this analysis). 869 cases of newborn male babies who lived at least 28 days are being analyzed (data set smoking.csv). The purpose of this document is to present a statistical model to interpret and understand the correlation between several variables and the chances of having a pre-term birth (< 270 days). The variables being considered for building the model, in association to the response variable for a logarithmic regression model of the probability of having a pre-term birth (premature), are the following:
- Total number of mother's previous pregnancies (parity) (numeric)
- Mother's race or ethnicity (mrace) (categorical)
- Mother's age in years at pregnancy termination (mage) (numeric)
- Mother's education level (med) (categorical)
- Mother's height in inches (mht) (numeric)
- Mother's pre-pregnancy weight in pounds (mpregwt) (numeric)
- Family yearly income in 2500-increment categories (inc) (categorical)
- Indicator for the mother's smoking (smoke) (categorical)

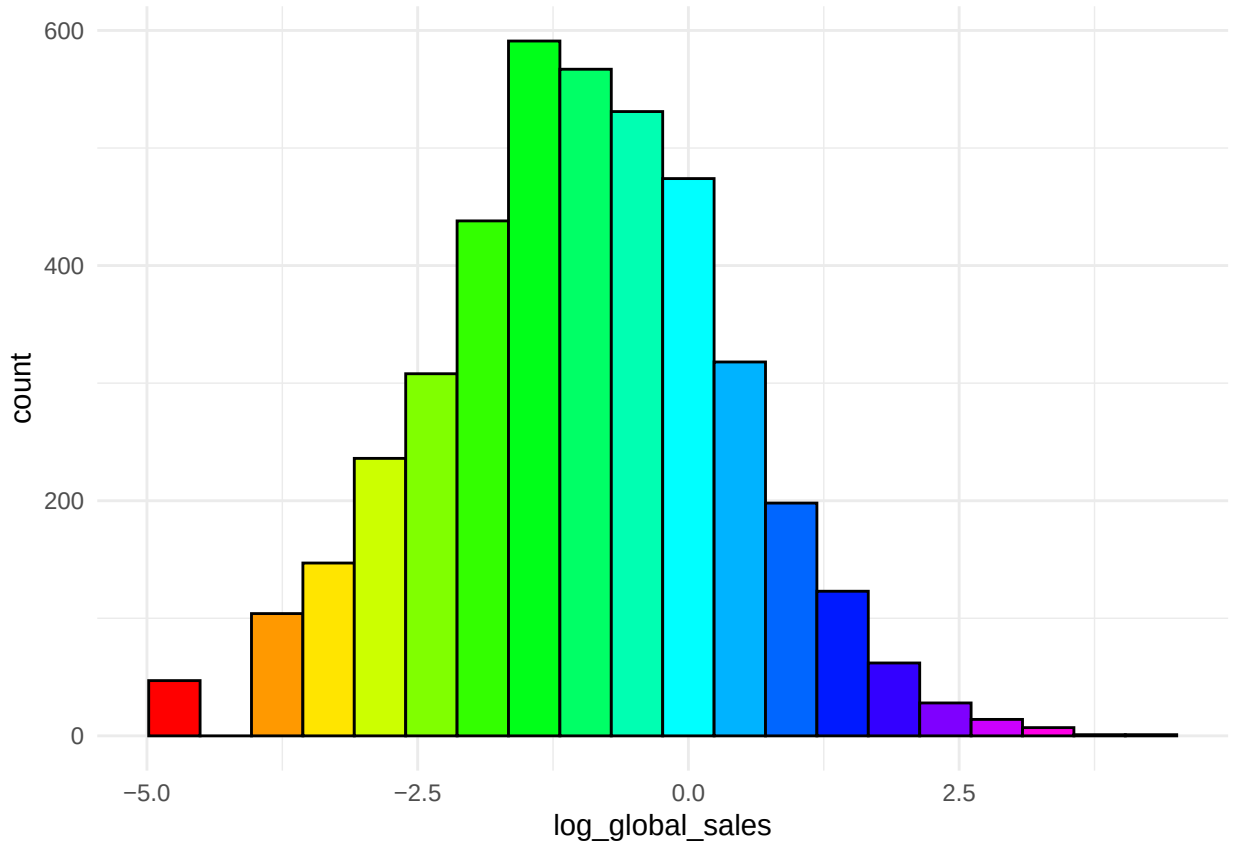
A summary of the data variables being analyzed can be found in Annex 1.1. An exploratory data analysis for all variables and plots for their interactions can be found in Annex 1.2.

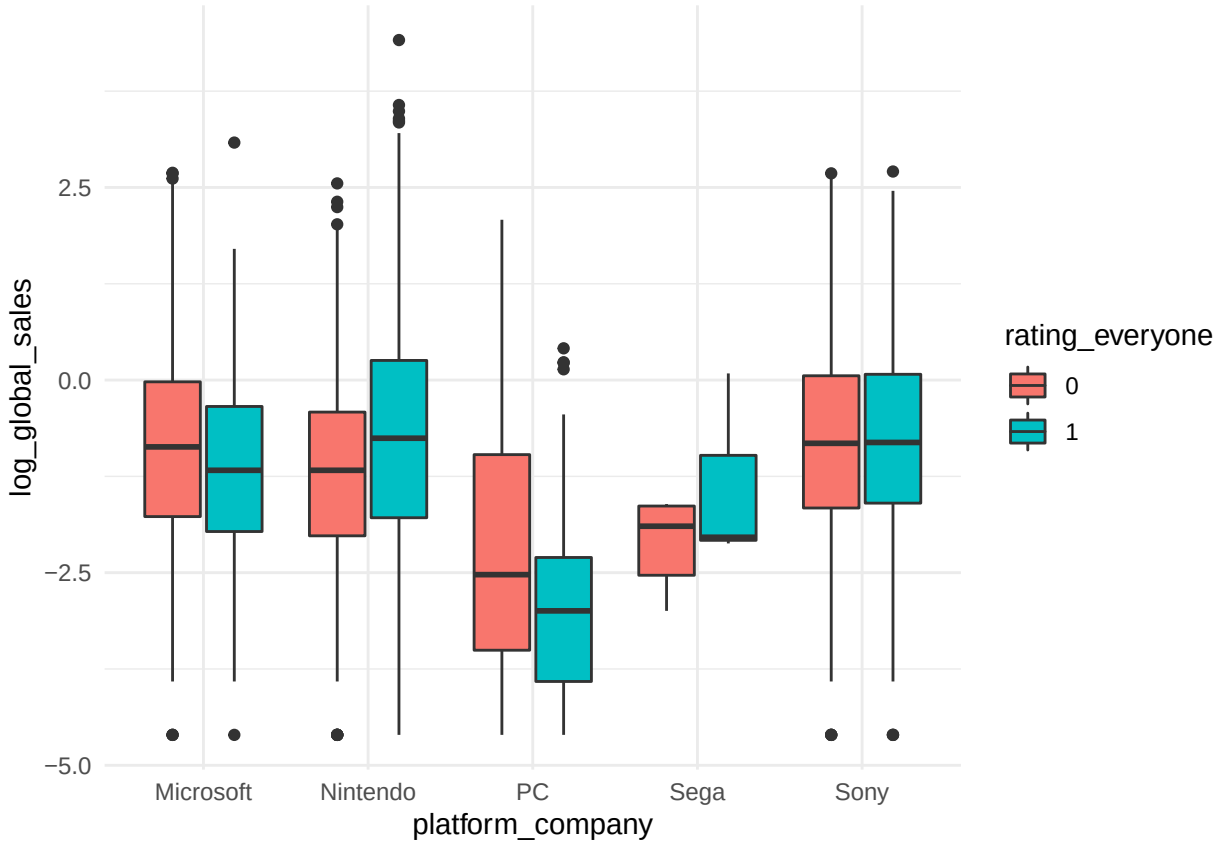
The EDA suggests none of the numerical variables have a clear association with premature as the boxplots for premature = 0 and premature = 1 do not have noticeable differences. For the categorical varibles, there are more interesting results in the conditional probability tables for each variable and their association with premature. This suggests that the categorical variables should be included in the model to evaluate their significance. The numerical variables do not need any obvious transformations as all of

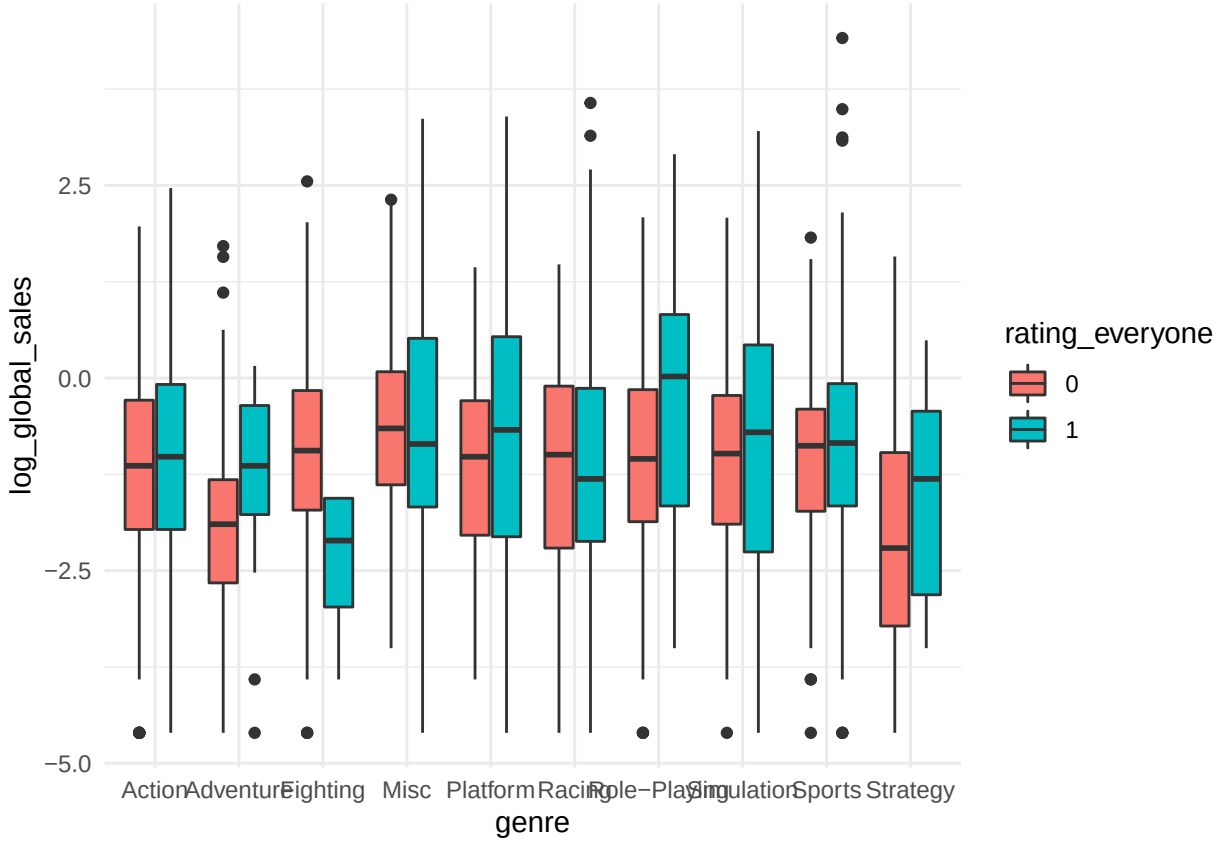
them suggest linear trends. The interactions parity_c:mage_c, parity_c:mpregwt_c, mage_c:mpregwt_c, mht_c:mpregwt_care being considered as those predictors have the largest correlations as seen in Annex 1.1's correlation table.

```
#Selecting 50 sample publishers
publishers <- unique(vgsales$publisher)
set.seed(2163386)
sample_publishers<- sample(publishers, 50)
sample_data <- vgsales[vgsales$publisher %in% sample_publishers,]
```

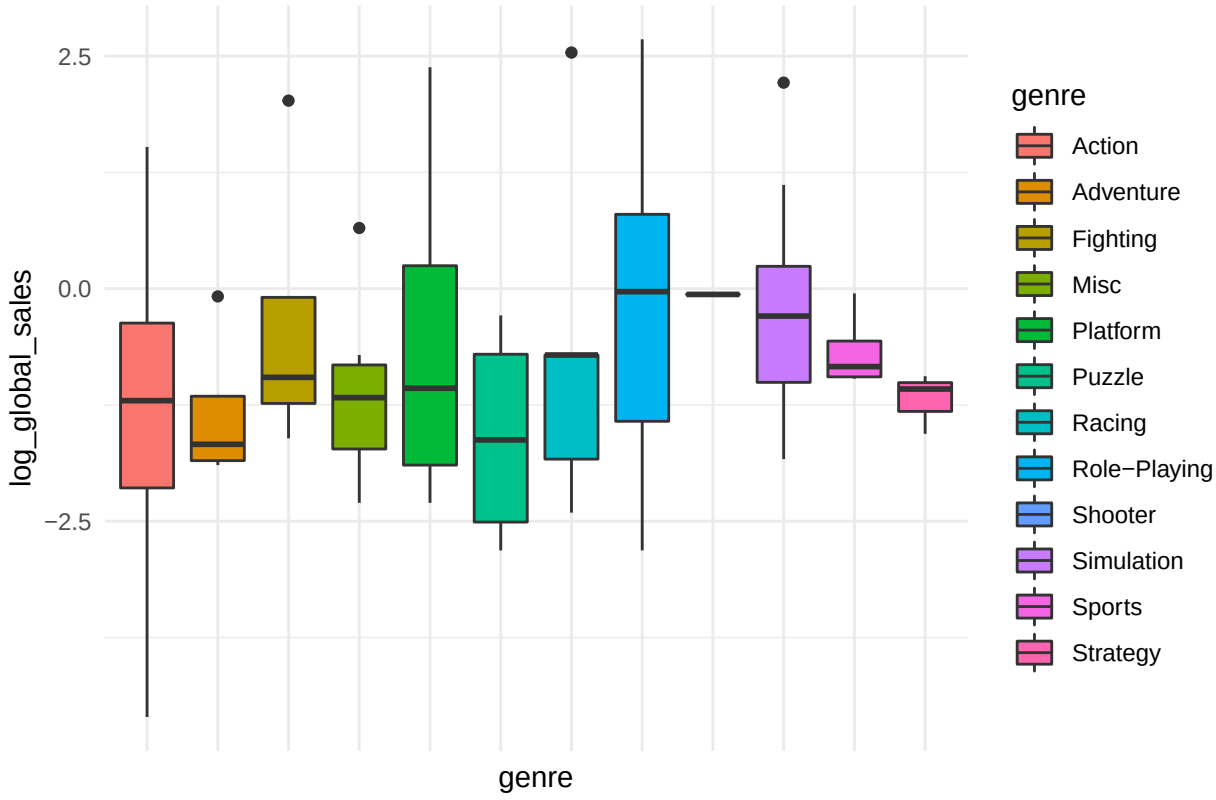


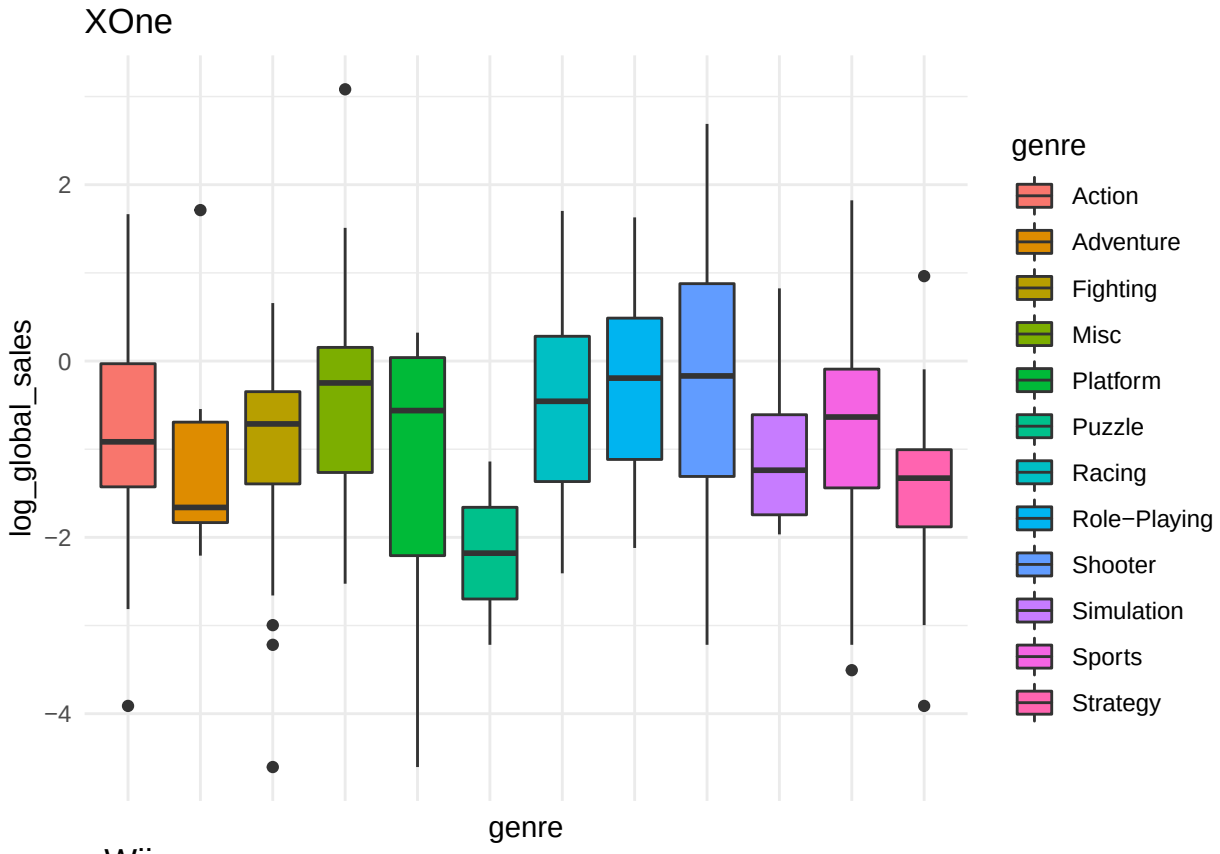

600
400
200
0
count
−5.0
−2.5
0.0
2.5
log_global_sales

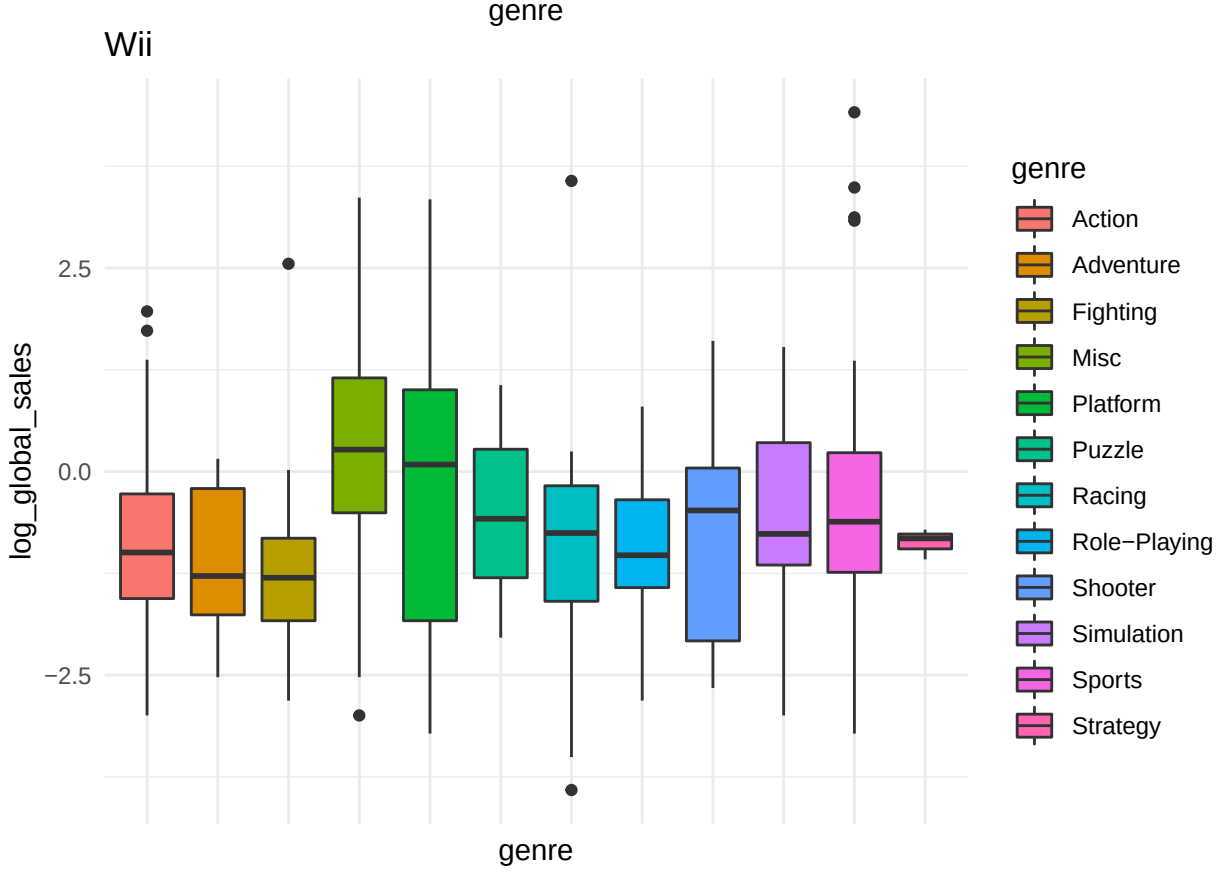

2.5
0.0
−2.5
−5.0
log_global_sales
Microsoft
Nintendo
PC
Sega
Sony
platform_company
rating_everyone
0
1

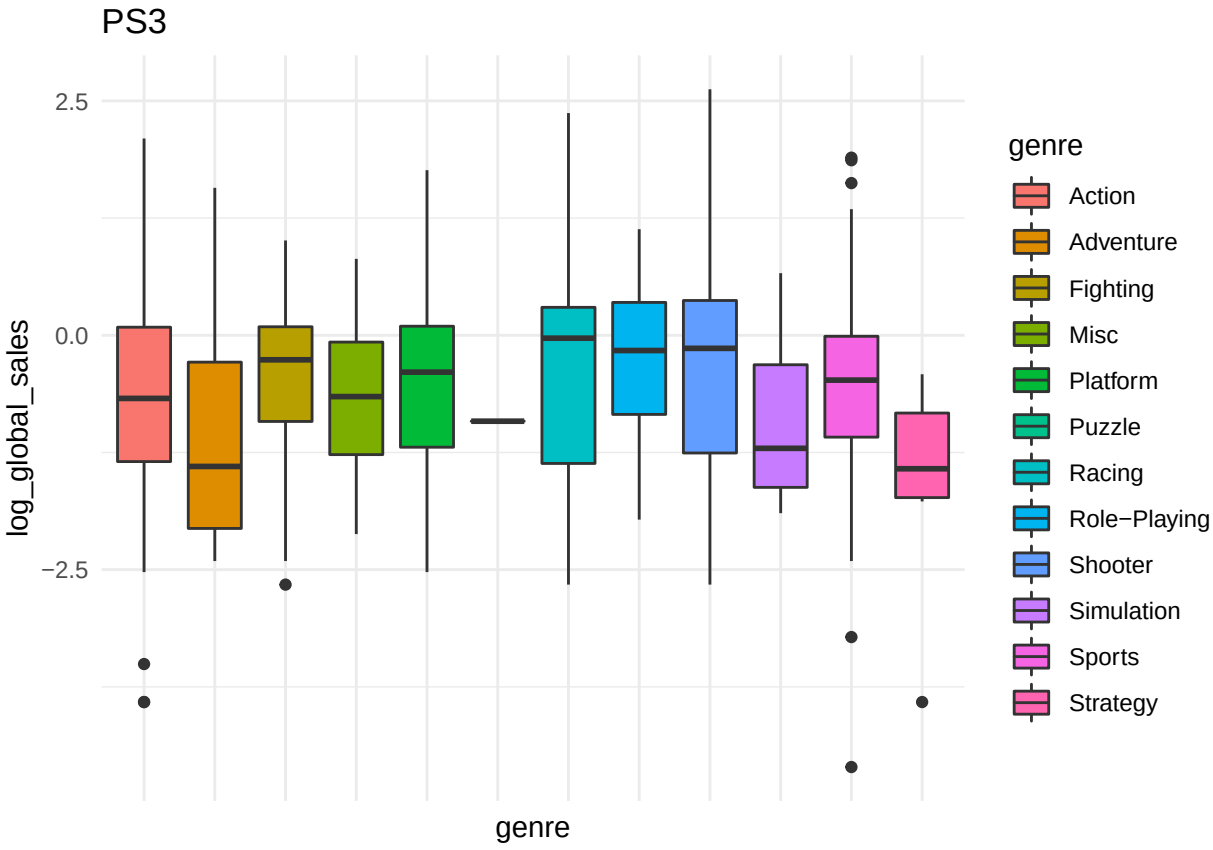

log_global_sales
2.5
0.0
−2.5
−5.0
Action
Adventure
Fighting
Misc
Platform
Racing
Role−Playing
Simulation
Sports
Strategy
genre
rating_everyone
0
1

3DS

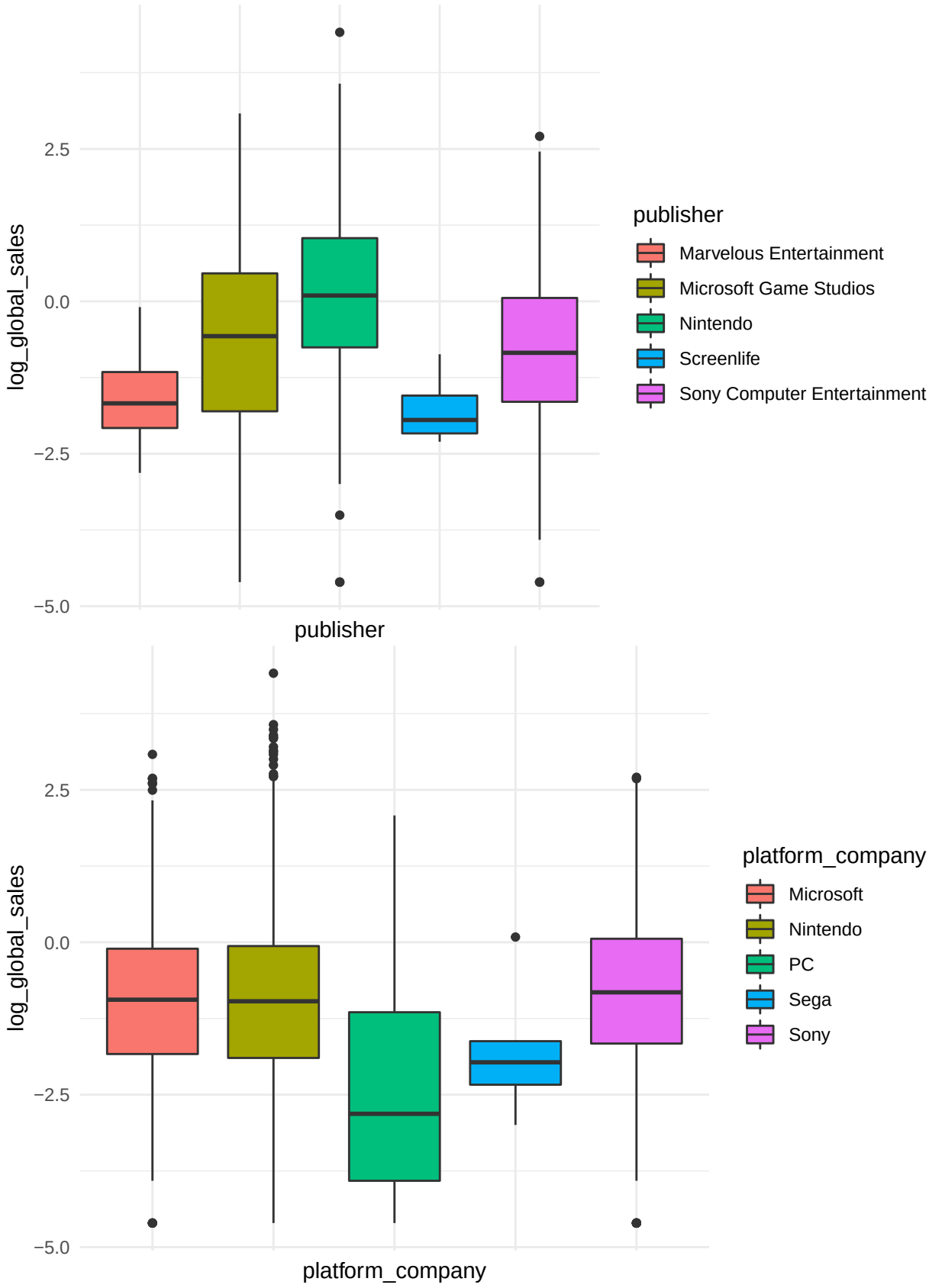

log_global_sales
2.5
0.0
−2.5
genre
genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role−Playing
Shooter
Simulation
Sports
Strategy

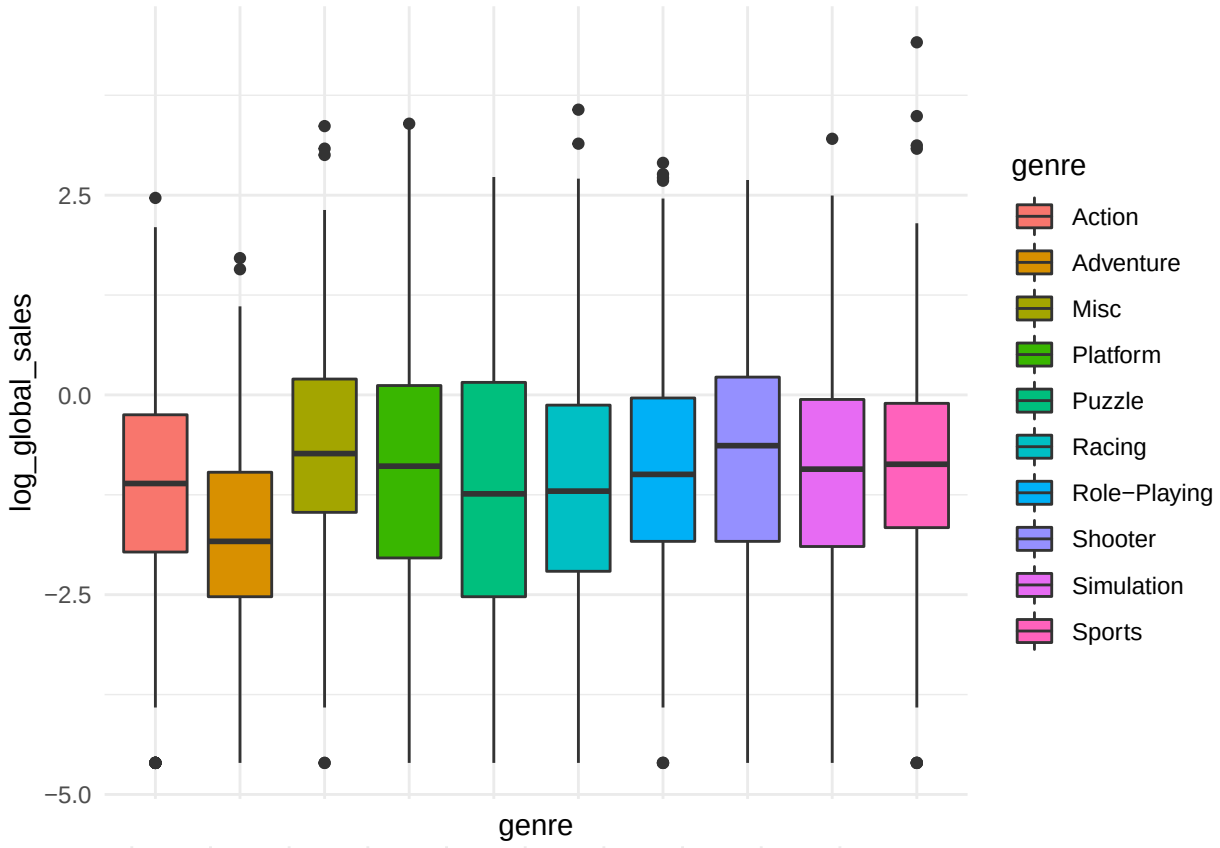

XOne
log_global_sales
2
0
−2
−4
genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role−Playing
Shooter
Simulation
Sports
Strategy

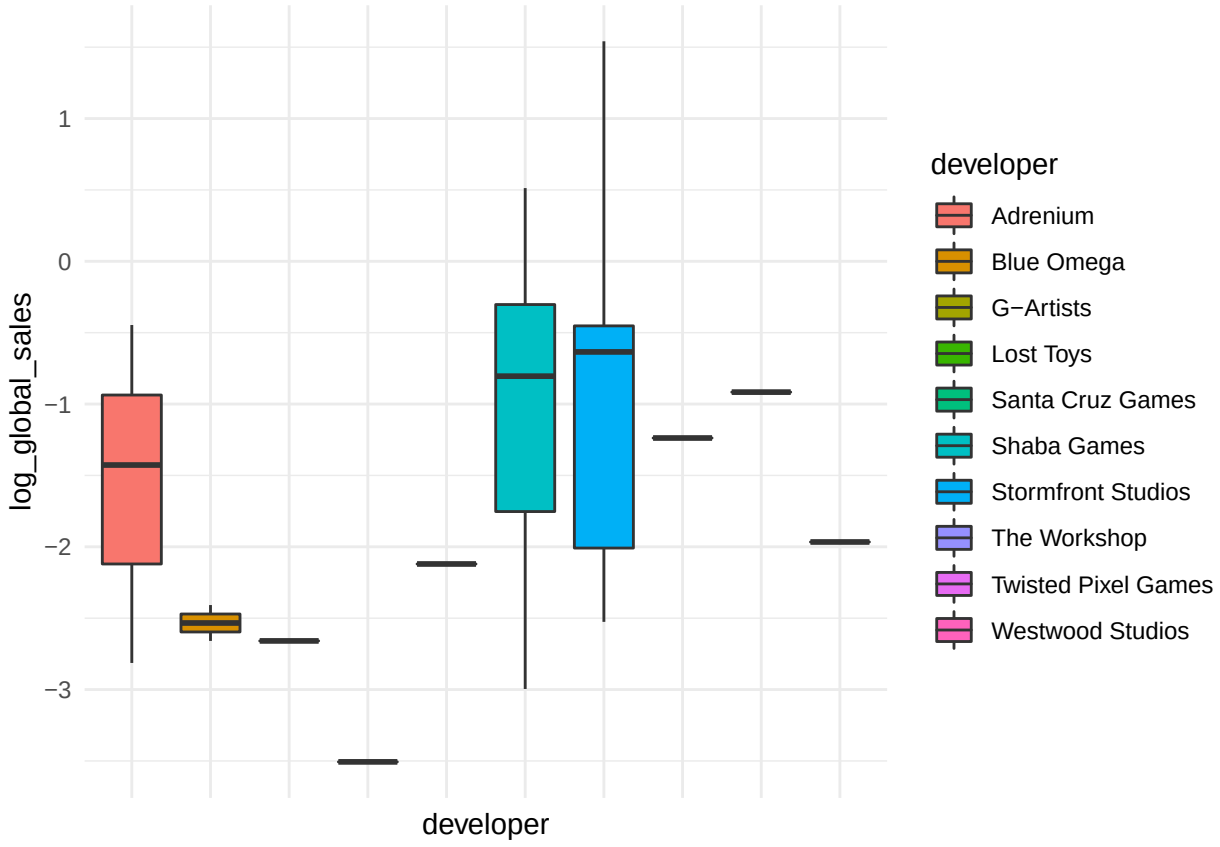

Wii
log_global_sales
2.5
0.0
−2.5
genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role−Playing
Shooter
Simulation
Sports
Strategy

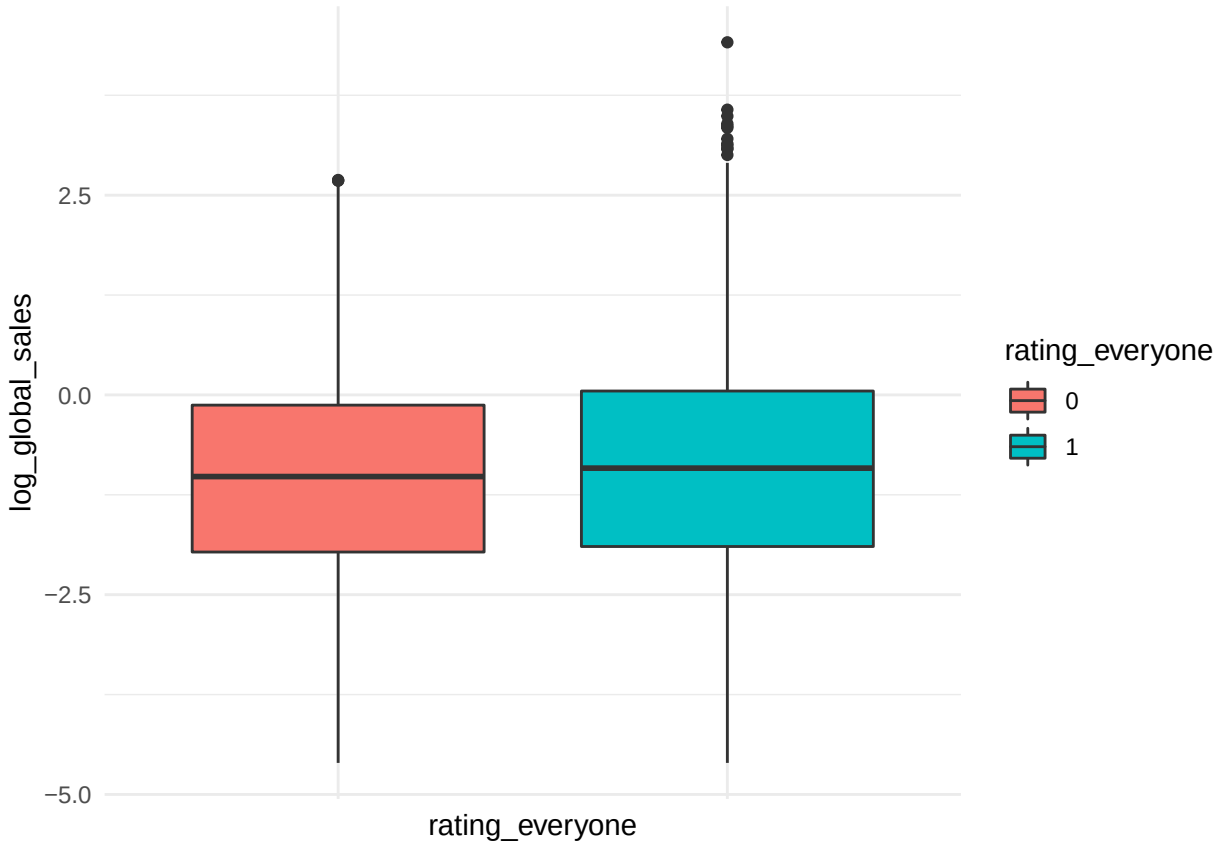



```
## [1] Nintendo                      Screenlife
## [3] Sony Computer Entertainment Microsoft Game Studios
## [5] Marvelous Entertainment
## 444 Levels: 10TACLE Studios 1C Company 2D Boy 2K Sports 3DO ... Zushi Games
```


log_global_sales
2.5
0.0
−2.5
−5.0
publisher
publisher
Marvelous Entertainment
Microsoft Game Studios
Nintendo
Screenlife
Sony Computer Entertainment
log_global_sales
2.5
0.0
−2.5
−5.0
platform_company
platform_company
Microsoft
Nintendo
PC
Sega
Sony

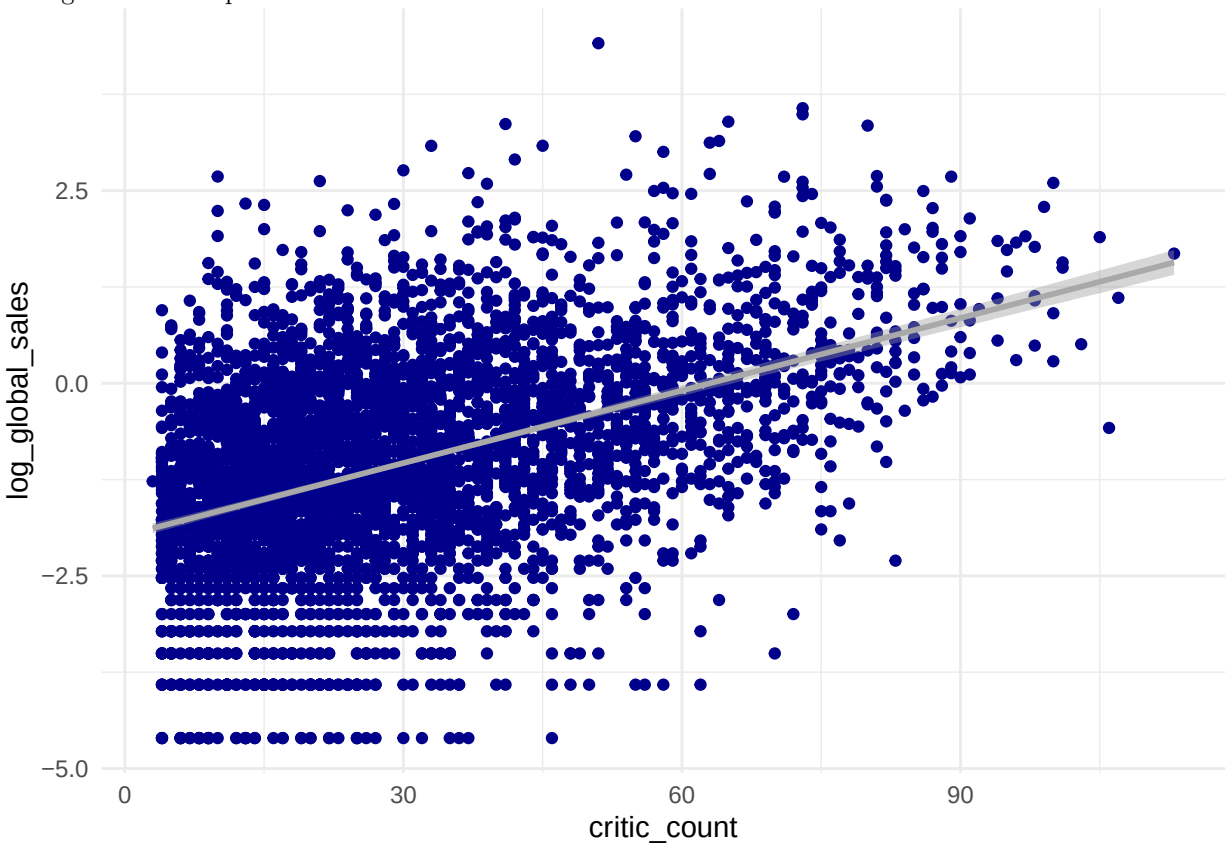

log_global_sales
2.5
0.0
−2.5
−5.0
genre
genre
Action
Adventure
Misc
Platform
Puzzle
Racing
Role−Playing
Shooter
Simulation
Sports

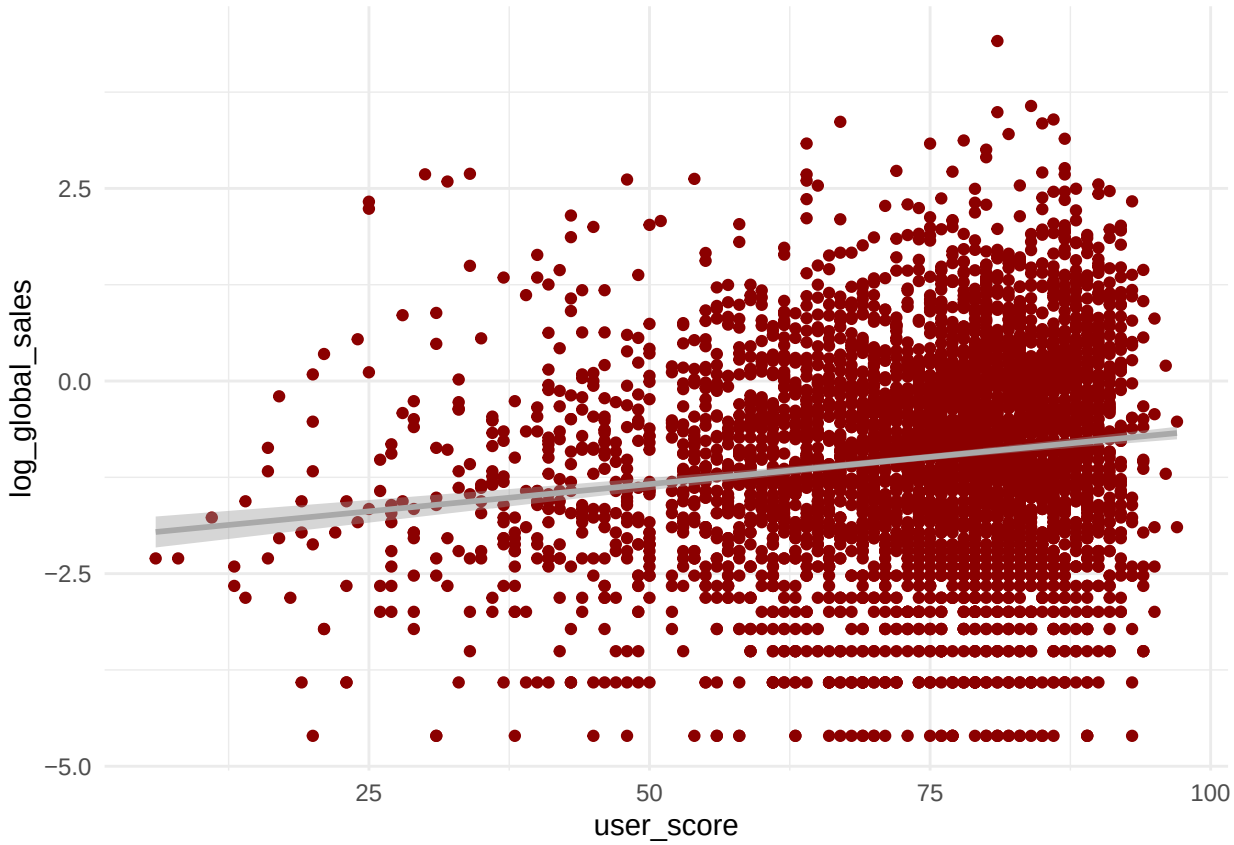

log_global_sales
1
0
−1
−2
−3
developer
developer
Adrenium
Blue Omega
G−Artists
Lost Toys
Santa Cruz Games
Shaba Games
Stormfront Studios
The Workshop
Twisted Pixel Games
Westwood Studios

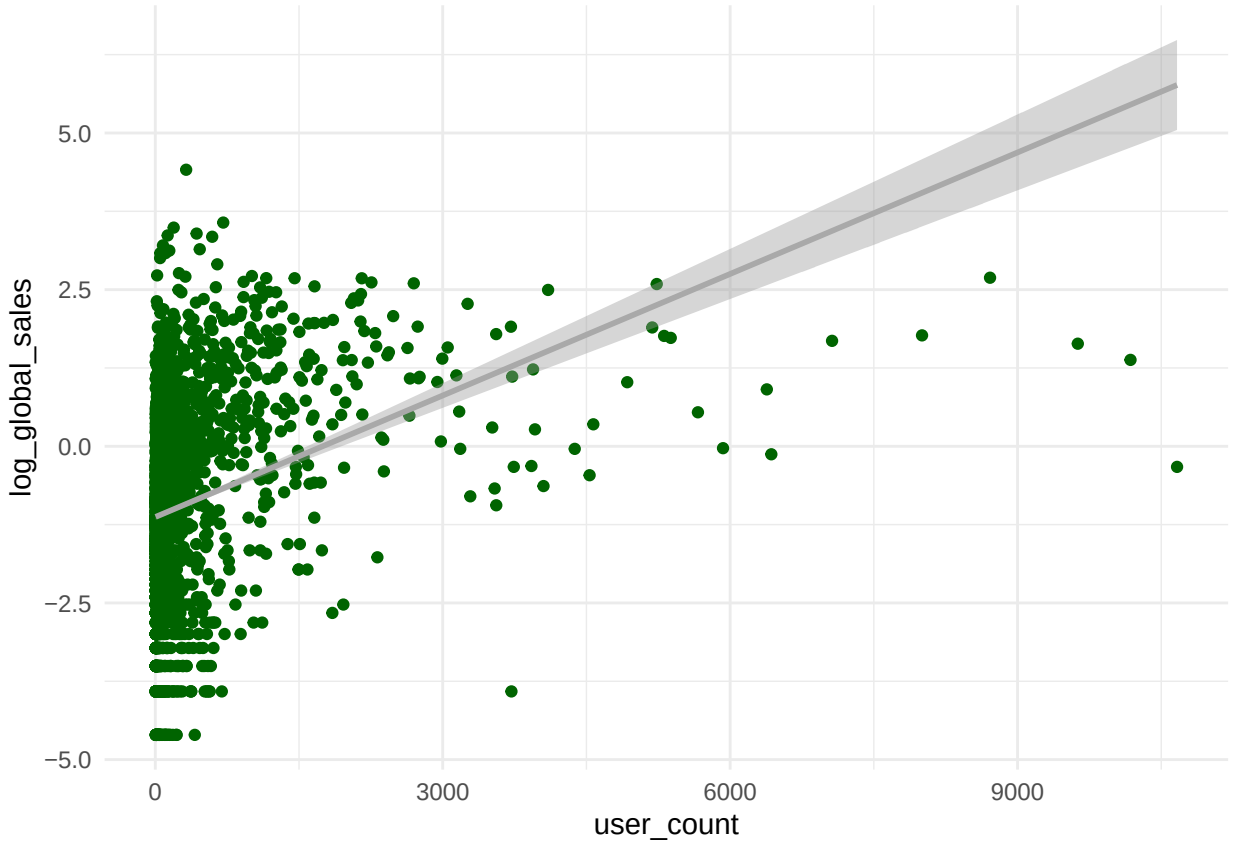


Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code

that generated the plot.







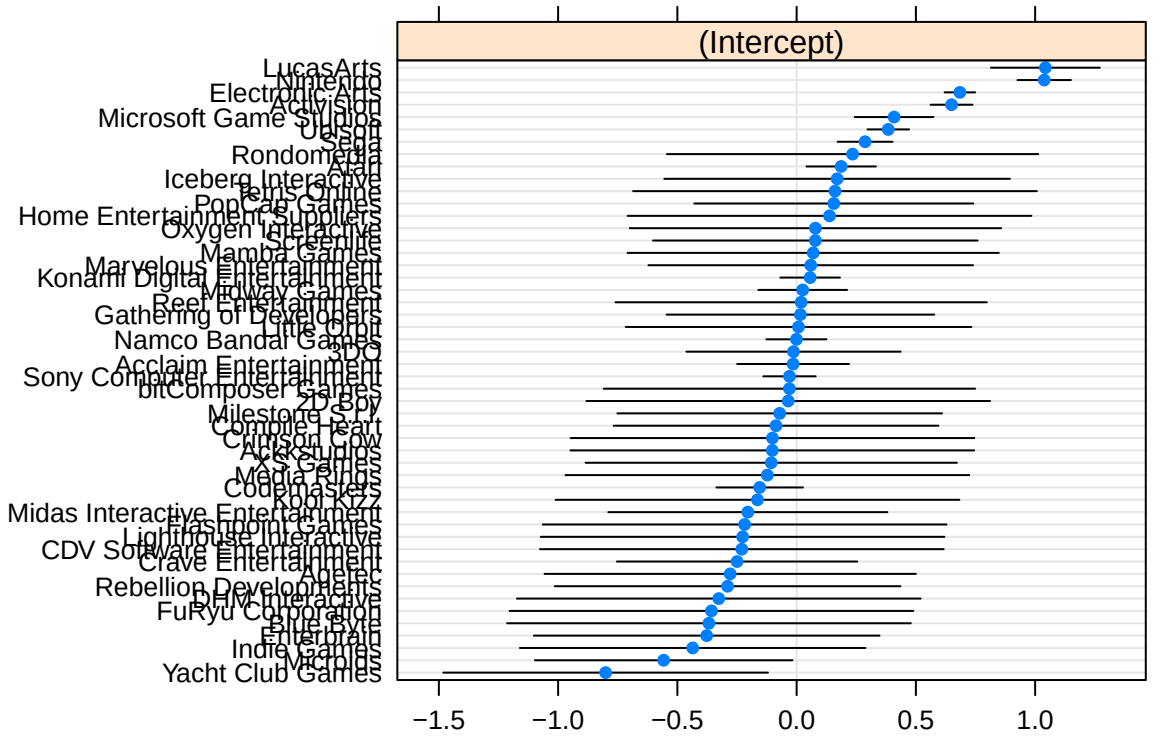

```
## Warning: 'r.squaredGLMM' now calculates a revised statistic. See the help
```

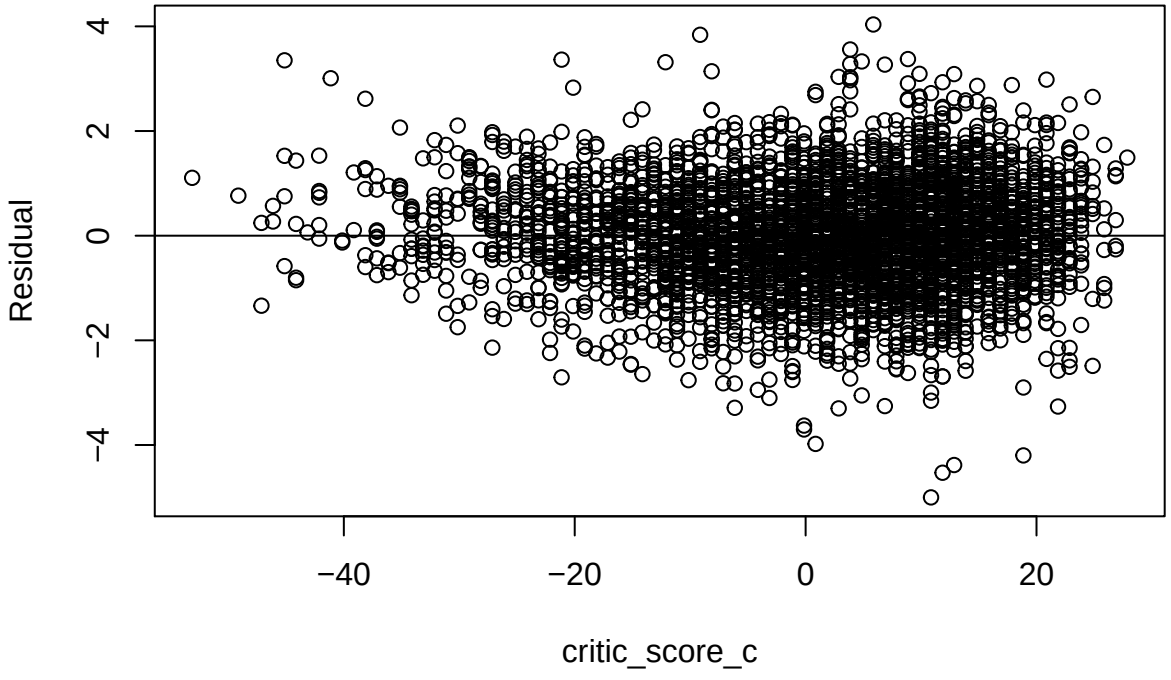
```
## page.
##           R2m       R2c
## [1,] 0.379874 0.4928372
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: log_global_sales ~ platform_company + genre + rating_everyone +
##     critic_score_c + critic_count_c + user_count_c + platform_company:rating_everyone +
##     genre:rating_everyone + (1 | publisher)
##    Data: sample_data
##
## REML criterion at convergence: 12162.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.9433 -0.6211  0.0006  0.6455  3.9891
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  publisher (Intercept) 0.228    0.4775
##  Residual              1.024    1.0117
## Number of obs: 4195, groups:  publisher, 50
##
## Fixed effects:
##                                           Estimate Std. Error         df
## (Intercept)                             -1.705e+00  1.065e-01  5.669e+01
## platform_companyNintendo                 8.224e-02  6.274e-02  4.159e+03
## platform_companyPC                      -1.438e+00  7.871e-02  4.157e+03
## platform_companySega                    -6.584e-01  3.919e-01  4.127e+03
## platform_companySony                     4.591e-01  5.219e-02  4.155e+03
## genreAdventure                          -3.381e-01  1.129e-01  4.074e+03
## genreFighting                            3.061e-01  8.488e-02  4.150e+03
## genreMisc                                5.342e-01  8.907e-02  4.149e+03
## genrePlatform                            5.899e-02  1.022e-01  4.143e+03
## genrePuzzle                             -2.041e-01  2.776e-01  4.160e+03
## genreRacing                              1.522e-01  8.912e-02  4.141e+03
## genreRole-Playing                       -1.871e-01  7.416e-02  4.160e+03
## genreShooter                             1.077e-01  5.869e-02  4.143e+03
## genreSimulation                          4.263e-01  9.425e-02  4.148e+03
## genreSports                              1.364e-01  9.419e-02  4.150e+03
## genreStrategy                           -4.983e-01  1.113e-01  4.158e+03
## rating_everyone1                         1.635e-01  1.192e-01  4.137e+03
## critic_score_c                           2.347e-02  1.374e-03  4.154e+03
## critic_count_c                           1.995e-02  1.058e-03  4.160e+03
## user_count_c                             4.565e-04  3.062e-05  4.132e+03
## platform_companyNintendo:rating_everyone1  2.343e-01  9.788e-02  4.142e+03
## platform_companyPC:rating_everyone1      -4.455e-02  1.567e-01  4.159e+03
## platform_companySega:rating_everyone1     3.611e-01  7.038e-01  4.123e+03
## platform_companySony:rating_everyone1     1.260e-01  8.833e-02  4.139e+03
## genreAdventure:rating_everyone1          -1.271e-01  2.486e-01  4.153e+03
## genreFighting:rating_everyone1           -7.776e-01  5.245e-01  4.132e+03
## genreMisc:rating_everyone1               -2.309e-01  1.583e-01  4.135e+03
## genrePlatform:rating_everyone1           -1.483e-01  1.598e-01  4.145e+03
## genrePuzzle:rating_everyone1             -4.005e-01  3.244e-01  4.158e+03
```

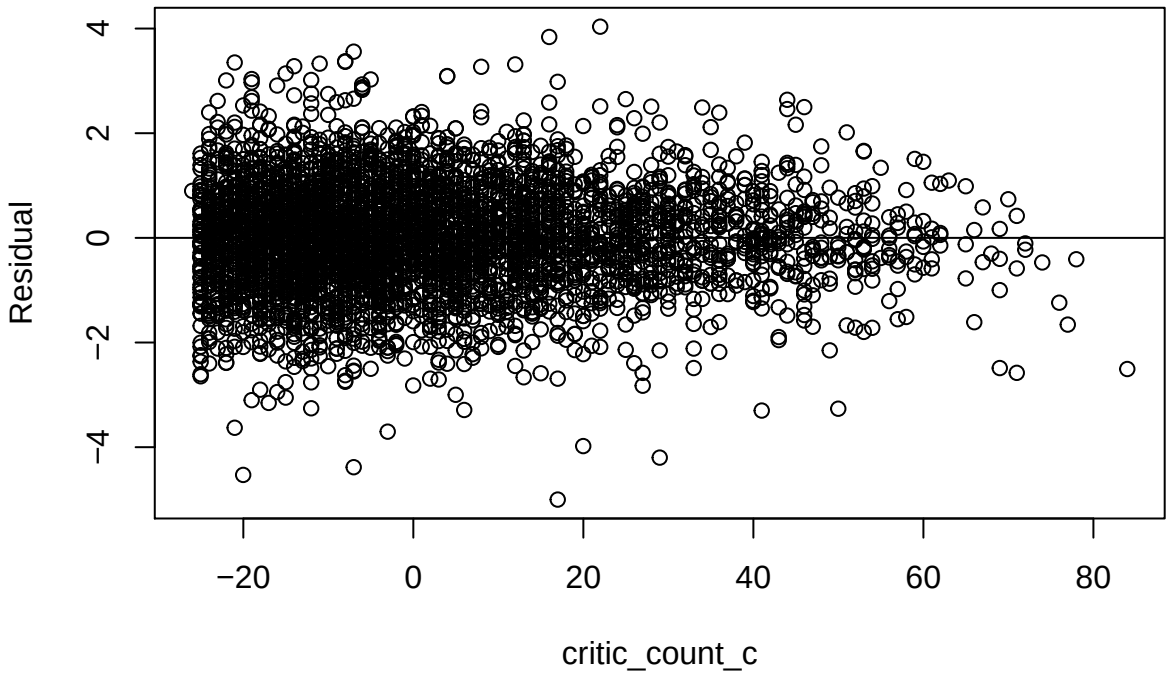
```
## genreRacing:rating_everyone1              -2.561e-01  1.434e-01  4.142e+03
## genreRole-Playing:rating_everyone1         4.731e-01  1.825e-01  4.154e+03
## genreShooter:rating_everyone1             -1.678e+00  4.283e-01  4.125e+03
## genreSimulation:rating_everyone1          -9.791e-02  1.850e-01  4.144e+03
## genreSports:rating_everyone1              -2.154e-01  1.364e-01  4.135e+03
## genreStrategy:rating_everyone1             6.813e-02  2.483e-01  4.160e+03
##                                           t value Pr(>|t|)
## (Intercept)                               -16.010  < 2e-16 ***
## platform_companyNintendo                    1.311 0.189973
## platform_companyPC                        -18.271  < 2e-16 ***
## platform_companySega                       -1.680 0.092999 .
## platform_companySony                        8.798  < 2e-16 ***
## genreAdventure                             -2.995 0.002762 **
## genreFighting                               3.607 0.000313 ***
## genreMisc                                   5.997 2.18e-09 ***
## genrePlatform                               0.577 0.563804
## genrePuzzle                                -0.735 0.462190
## genreRacing                                 1.708 0.087739 .
## genreRole-Playing                          -2.524 0.011656 *
## genreShooter                                1.835 0.066650 .
## genreSimulation                             4.523 6.26e-06 ***
## genreSports                                 1.448 0.147742
## genreStrategy                              -4.479 7.70e-06 ***
## rating_everyone1                            1.372 0.170099
## critic_score_c                             17.078  < 2e-16 ***
## critic_count_c                             18.849  < 2e-16 ***
## user_count_c                               14.910  < 2e-16 ***
## platform_companyNintendo:rating_everyone1   2.393 0.016746 *
## platform_companyPC:rating_everyone1        -0.284 0.776238
## platform_companySega:rating_everyone1       0.513 0.607957
## platform_companySony:rating_everyone1       1.426 0.153810
## genreAdventure:rating_everyone1            -0.511 0.609194
## genreFighting:rating_everyone1             -1.483 0.138257
## genreMisc:rating_everyone1                 -1.458 0.144839
## genrePlatform:rating_everyone1             -0.928 0.353350
## genrePuzzle:rating_everyone1               -1.235 0.217037
## genreRacing:rating_everyone1               -1.786 0.074201 .
## genreRole-Playing:rating_everyone1          2.592 0.009566 **
## genreShooter:rating_everyone1              -3.919 9.04e-05 ***
## genreSimulation:rating_everyone1           -0.529 0.596703
## genreSports:rating_everyone1               -1.578 0.114550
## genreStrategy:rating_everyone1              0.274 0.783789
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation matrix not shown by default, as p = 35 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
##                  platform_companyNintendo
##                                  2.375819
##                       platform_companyPC
##                                  1.901742
##                     platform_companySega
```

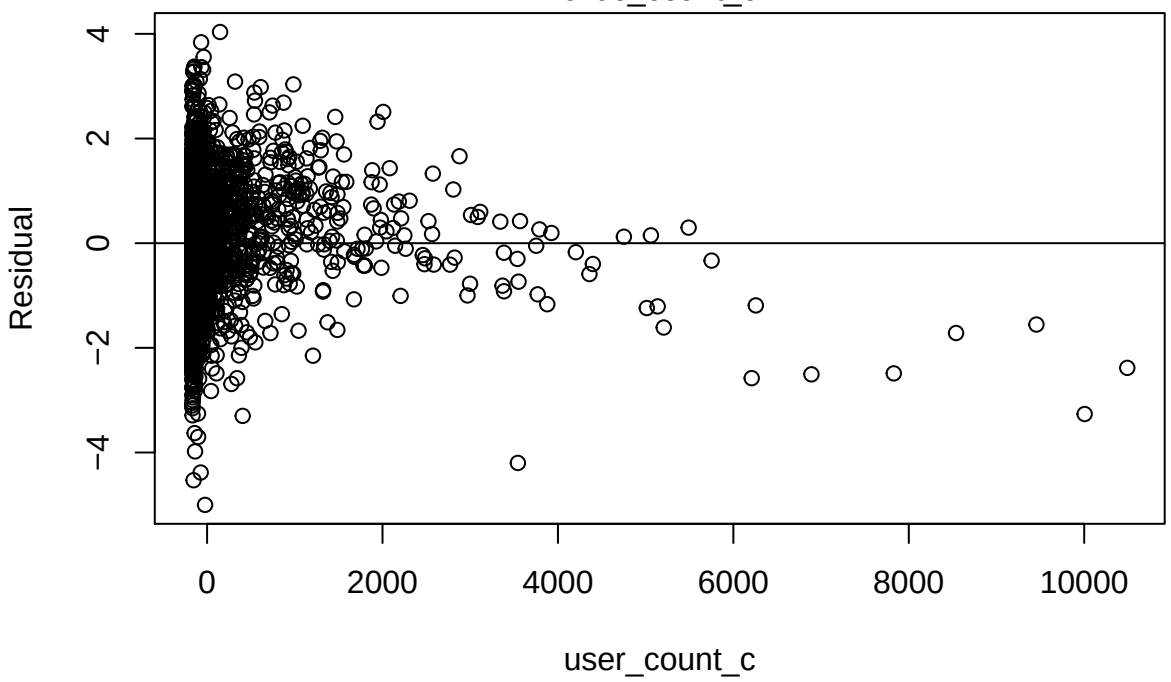
```
##                                   1.459059
##                         platform_companySony
##                                   2.135685
##                               genreAdventure
##                                   1.389655
##                                genreFighting
##                                   1.229708
##                                    genreMisc
##                                   1.934193
##                                genrePlatform
##                                   2.390652
##                                  genrePuzzle
##                                   4.537605
##                                 genreRacing
##                                   2.479311
##                             genreRole-Playing
##                                   1.521184
##                                genreShooter
##                                   1.513517
##                              genreSimulation
##                                   1.659963
##                                 genreSports
##                                   4.783893
##                               genreStrategy
##                                   1.435584
##                             rating_everyone1
##                                  11.753879
##                               critic_score_c
##                                   1.288739
##                               critic_count_c
##                                   1.587880
##                                 user_count_c
##                                   1.333754
## platform_companyNintendo:rating_everyone1
##                                   3.633627
##       platform_companyPC:rating_everyone1
##                                   1.644967
##     platform_companySega:rating_everyone1
##                                   1.436984
##     platform_companySony:rating_everyone1
##                                   3.286729
##           genreAdventure:rating_everyone1
##                                   1.477025
##            genreFighting:rating_everyone1
##                                   1.058247
##                genreMisc:rating_everyone1
##                                   2.599116
##            genrePlatform:rating_everyone1
##                                   3.330940
##              genrePuzzle:rating_everyone1
##                                   4.882269
##              genreRacing:rating_everyone1
##                                   3.918297
##        genreRole-Playing:rating_everyone1
```

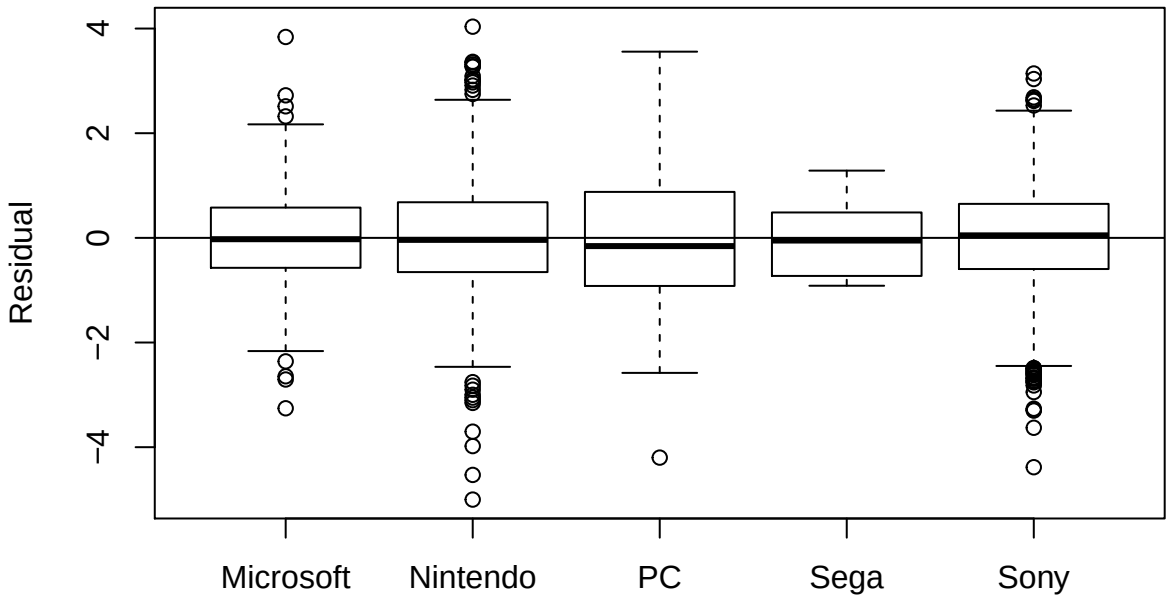
```
##                                    1.603394
##               genreShooter:rating_everyone1
##                                    1.065803
##            genreSimulation:rating_everyone1
##                                    1.948509
##                genreSports:rating_everyone1
##                                    8.517019
##              genreStrategy:rating_everyone1
##                                    1.469225
```

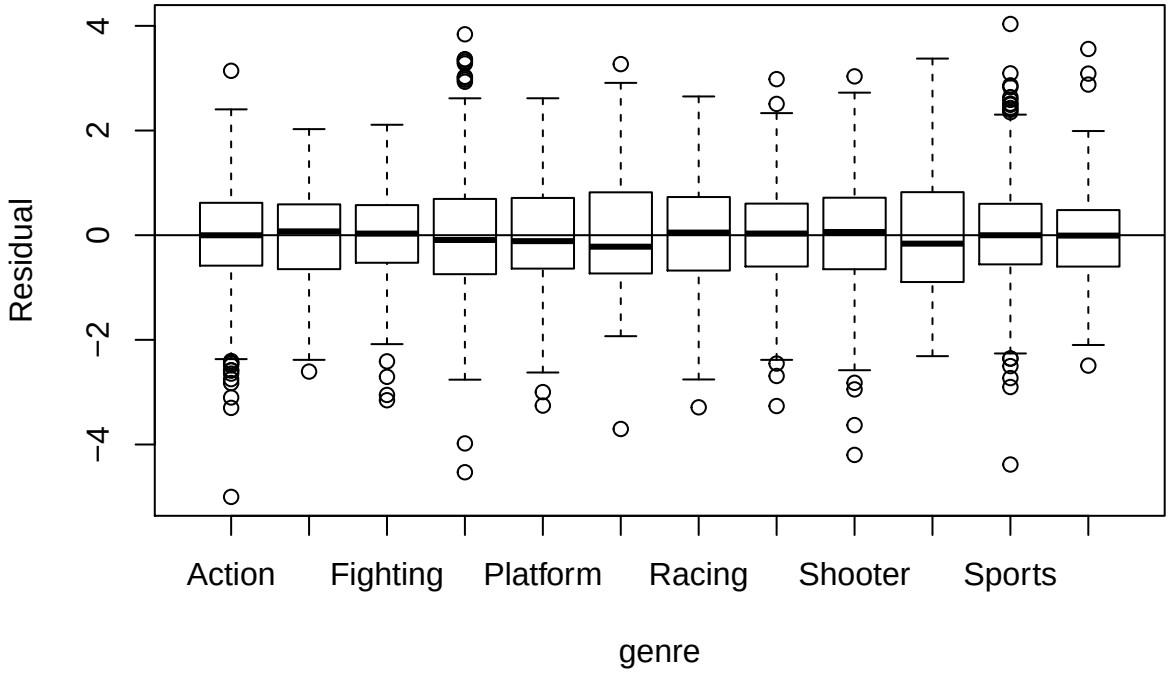


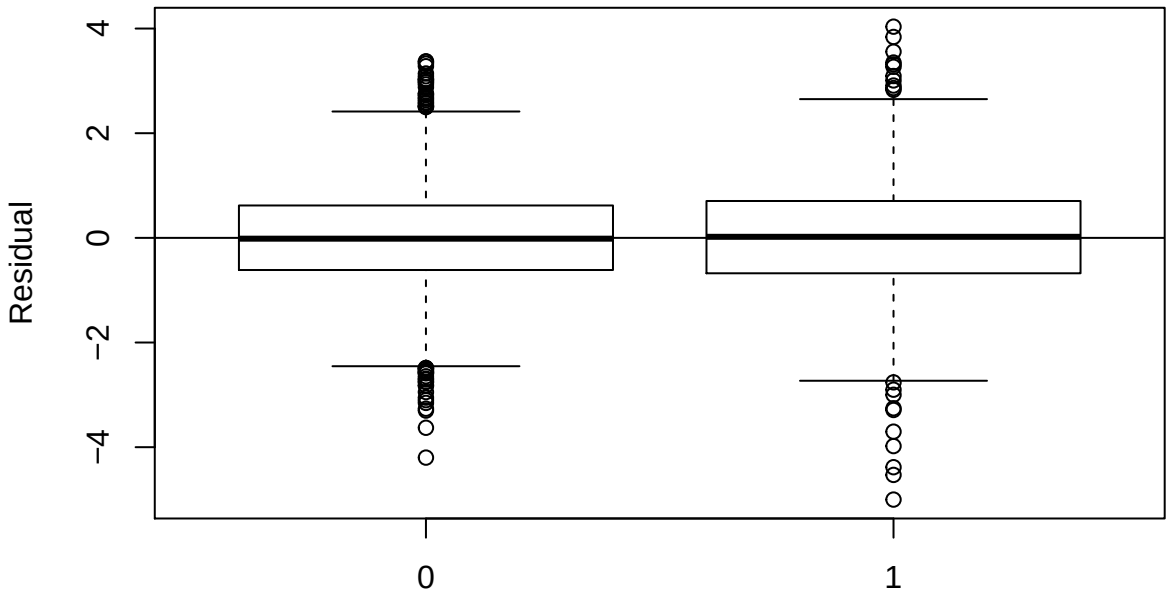


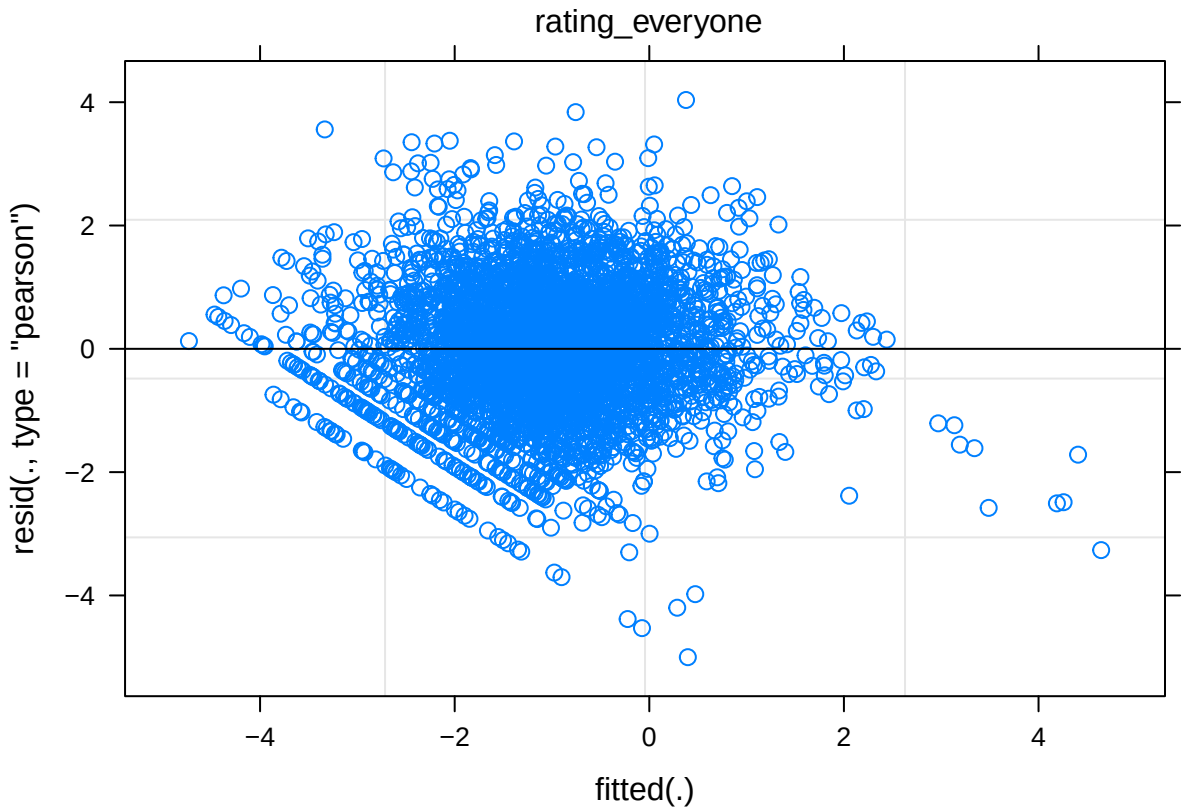

Residual
4
2
0
−2
−4
−20
0
20
40
60
80
critic_count_c

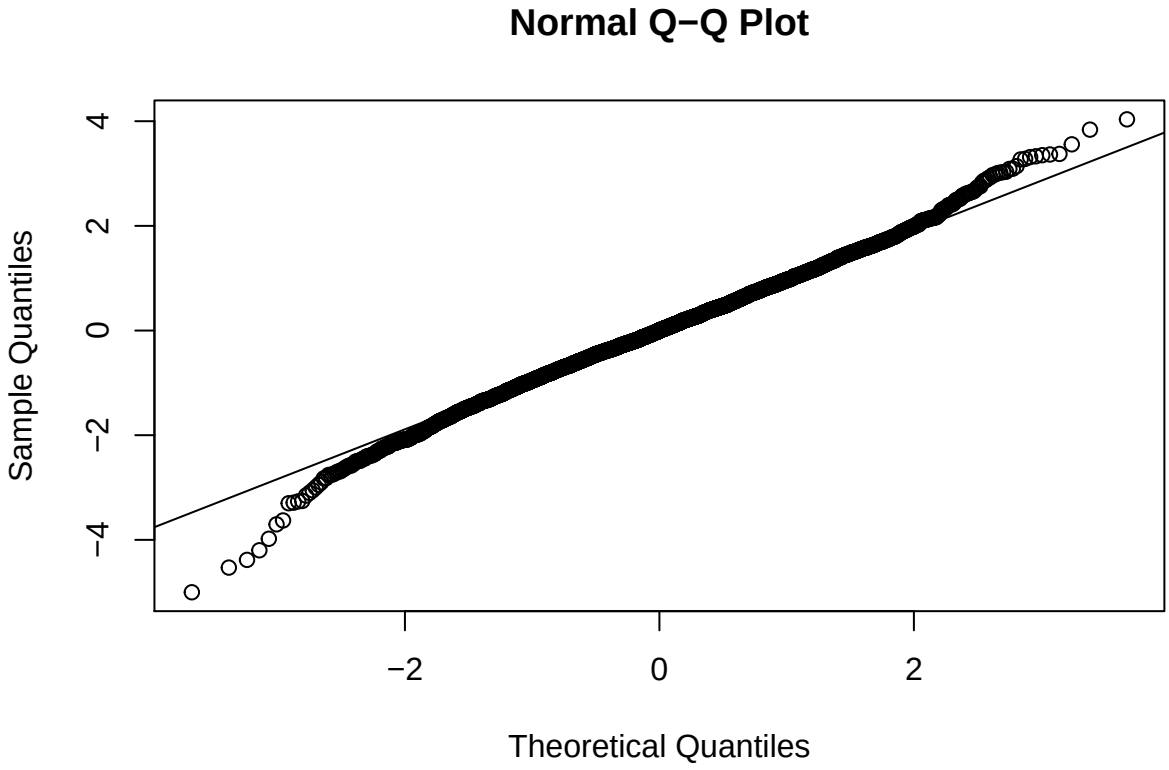

Residual
4
2
0
−2
−4
0
2000
4000
6000
8000
10000
user_count_c


Residual
4
2
0
−2
−4
Microsoft
Nintendo
PC
Sega
Sony
platform_company


Residual
4
2
0
−2
−4
Action
Fighting
Platform
Racing
Shooter
Sports
genre


Residual
4
2
0
−2
−4
0
1


rating_everyone
resid(., type = "pearson")
4
2
0
−2
−4
−4
−2
0
2
4
fitted(.)


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
4
2
0
−2
−4
−2
0
2