Final Project

Sebastián Soriano Pérez [ss1072]

12/10/2019

ANALYZING VIDEO GAME SALES

```
vgsales <- read.csv('vgsales.csv')
vgsales$Year_of_Release <- year(as.Date(as.character(vgsales$Year_of_Release), format = '%Y'))
colnames(vgsales) <- c('name', 'platform', 'year_of_release', 'genre', 'publisher', 'na_sales',
                       'eu_sales', 'jp_sales', 'other_sales', 'global_sales', 'critic_score',
                       'critic_count', 'user_score', 'user_count', 'developer', 'rating')
vgsales[vgsales$user_score == '',]$user_score <- 'tbd'
vgsales[vgsales$user_score == 'tbd',]$user_score <- NA
vgsales <- vgsales[!is.na(vgsales$critic_score),]
vgsales <- vgsales[!is.na(vgsales$user_score),]
vgsales$user_score <- as.numeric(vgsales$user_score)
vgsales$log_global_sales <- log(vgsales$global_sales)
vgsales$critic_score_c <- vgsales$critic_score - mean(vgsales$critic_score)
vgsales$critic_count_c <- vgsales$critic_count - mean(vgsales$critic_count)
vgsales$user_score_c <- vgsales$user_score - mean(vgsales$user_score)
vgsales$user_count_c <- vgsales$user_count - mean(vgsales$user_count)
vgsales$hit <- 0
vgsales[vgsales$global_sales > 1,]$hit <- 1
vgsales$platform_company <- 'Sony'
vgsales[
  vgsales$platform == '3DS'
  | vgsales$platform == 'DS'
  | vgsales$platform == 'GB'
  | vgsales$platform == 'GBA'
  | vgsales$platform == 'GC'
  | vgsales$platform == 'N64'
  | vgsales$platform == 'Wii'
  | vgsales$platform == 'WiiU',
  ]$platform_company <- 'Nintendo'
vgsales[vgsales$platform == 'DC',]$platform_company <- 'Sega'
vgsales[vgsales$platform == 'PC',]$platform_company <- 'PC'
#vgsales[vgsales$platform == 'WS',]$platform_company <- 'Bandai'
vgsales[
  vgsales$platform == 'X360'
  | vgsales$platform == 'XB'
  | vgsales$platform == 'XOne',
  ]$platform_company <- 'Microsoft'
vgsales$platform_company <- as.factor(vgsales$platform_company)
vgsales$rating_everyone <- 0
vgsales[vgsales$rating == 'E',]$rating_everyone <- 1
vgsales$rating_everyone <- as.factor(vgsales$rating_everyone)
summary(vgsales)
```

```
##                                           name         platform
##  Madden NFL 07                              :   9   PS2    :1161
##  LEGO Star Wars II: The Original Trilogy    :   8   X360   : 881
##  Need for Speed: Most Wanted                :   8   PS3    : 791
##  Harry Potter and the Order of the Phoenix  :   7   PC     : 692
##  LEGO Batman: The Videogame                 :   7   XB     : 581
##  LEGO Indiana Jones: The Original Adventures:   7   Wii    : 492
##  (Other)                                    :6905   (Other):2353
##  year_of_release          genre                            publisher
##  Min.   :2000    Action      :1666   Electronic Arts           : 961
##  1st Qu.:2004    Sports      : 972   Ubisoft                   : 512
##  Median :2007    Shooter     : 884   Activision                : 505
##  Mean   :2008    Role-Playing: 708   Sony Computer Entertainment: 322
##  3rd Qu.:2011    Racing      : 591   Nintendo                  : 309
##  Max.   :2016    Platform    : 402   THQ                       : 309
##                  (Other)     :1728   (Other)                   :4033
##     na_sales         eu_sales          jp_sales        other_sales
##  Min.   : 0.000   Min.   : 0.0000   Min.   :0.00000   Min.   : 0.00000
##  1st Qu.: 0.060   1st Qu.: 0.0200   1st Qu.:0.00000   1st Qu.: 0.01000
##  Median : 0.150   Median : 0.0600   Median :0.00000   Median : 0.02000
##  Mean   : 0.395   Mean   : 0.2378   Mean   :0.06494   Mean   : 0.08351
##  3rd Qu.: 0.385   3rd Qu.: 0.2100   3rd Qu.:0.01000   3rd Qu.: 0.07000
##  Max.   :41.360   Max.   :28.9600   Max.   :6.50000   Max.   :10.57000
##                                                                        
##   global_sales      critic_score    critic_count      user_score
##  Min.   : 0.0100   Min.   :13.00   Min.   :  3.00   Min.   : 5.00
##  1st Qu.: 0.1100   1st Qu.:62.00   1st Qu.: 14.00   1st Qu.:66.00
##  Median : 0.2900   Median :72.00   Median : 25.00   Median :76.00
##  Mean   : 0.7814   Mean   :70.14   Mean   : 29.02   Mean   :72.59
##  3rd Qu.: 0.7400   3rd Qu.:80.00   3rd Qu.: 40.00   3rd Qu.:83.00
##  Max.   :82.5300   Max.   :98.00   Max.   :113.00   Max.   :97.00
##                                                                   
##    user_count                          developer        rating
##  Min.   :    4   Electronic Arts           : 614   T      :2380
##  1st Qu.:   11   Ubisoft                   : 305   E      :2106
##  Median :   27   Konami                    : 148   M      :1448
##  Mean   :  174   Capcom                    : 132   E10+   : 948
##  3rd Qu.:   89   Sony Computer Entertainment: 107          :  66
##  Max.   :10665   Nintendo                  :  85   RP     :   2
##                  (Other)                   :5560   (Other):   1
##  log_global_sales  critic_score_c    critic_count_c     user_score_c
##  Min.   :-4.6052   Min.   :-57.137   Min.   :-26.016   Min.   :-67.586
##  1st Qu.:-2.2073   1st Qu.: -8.137   1st Qu.:-15.016   1st Qu.: -6.586
##  Median :-1.2379   Median :  1.863   Median : -4.016   Median :  3.414
##  Mean   :-1.2509   Mean   :  0.000   Mean   :  0.000   Mean   :  0.000
##  3rd Qu.:-0.3011   3rd Qu.:  9.863   3rd Qu.: 10.984   3rd Qu.: 10.414
##  Max.   : 4.4132   Max.   : 27.863   Max.   : 83.984   Max.   : 24.414
##                                                                        
##   user_count_c           hit          platform_company rating_everyone
##  Min.   : -169.96   Min.   :0.0000   Microsoft:1628    0:4845
##  1st Qu.: -162.96   1st Qu.:0.0000   Nintendo :1818    1:2106
##  Median : -146.96   Median :0.0000   PC       : 692
##  Mean   :    0.00   Mean   :0.1873   Sega     :  11
##  3rd Qu.:  -84.96   3rd Qu.:0.0000   Sony     :2802
##  Max.   :10491.04   Max.   :1.0000
##
```

- **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.

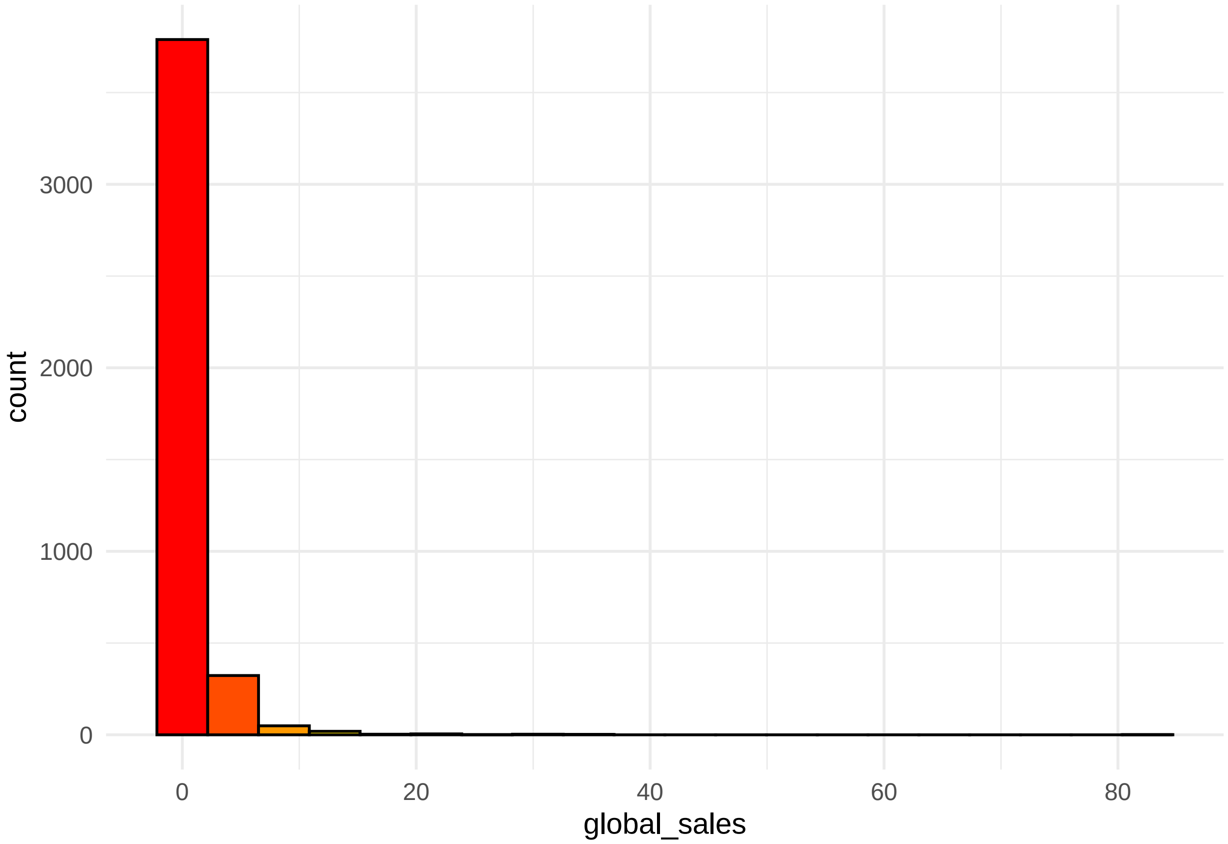
- **Data**

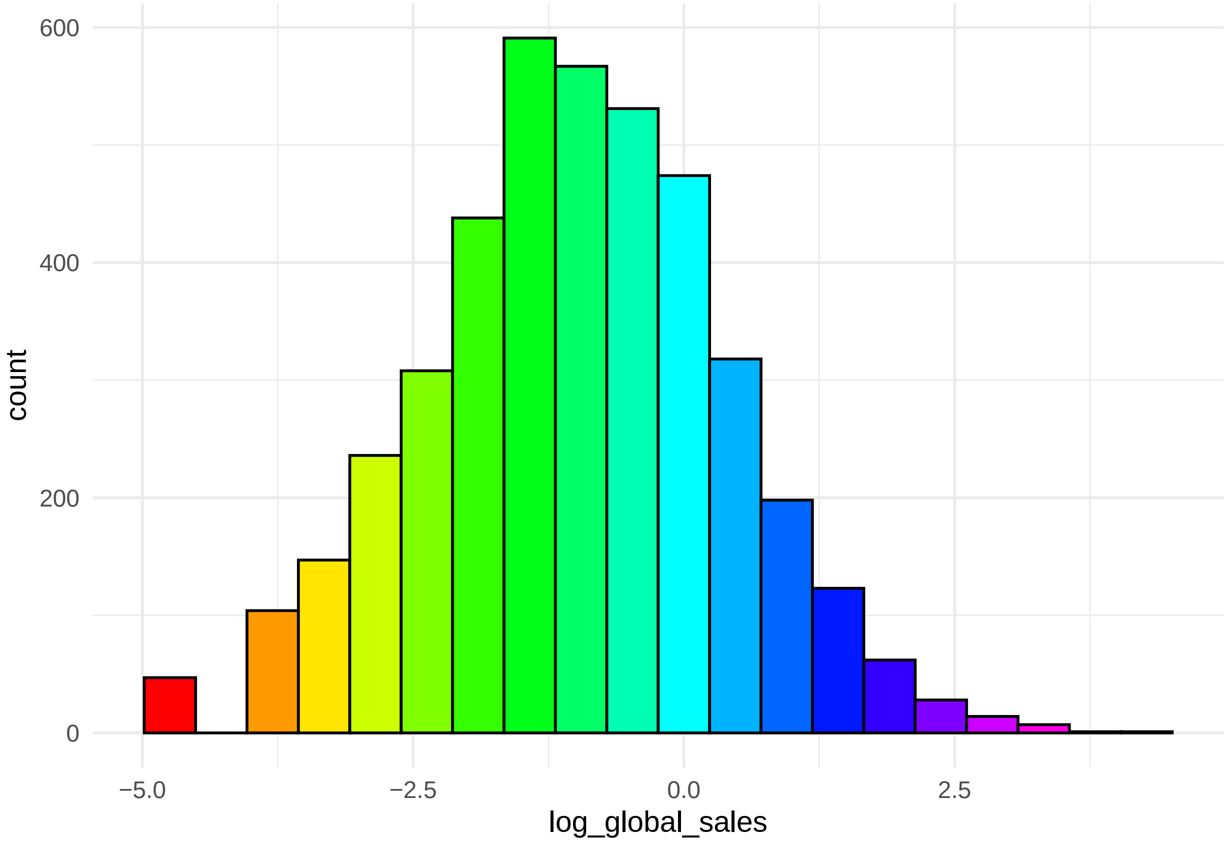
The Child Health and Development Studies research was one of the first to collect data to understand and quantify the risk of smoking during pregnancy to the baby's health. The data was collected from 1960 to 1967, and a subset of that data is being analyzed in this document (the variables related to 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)

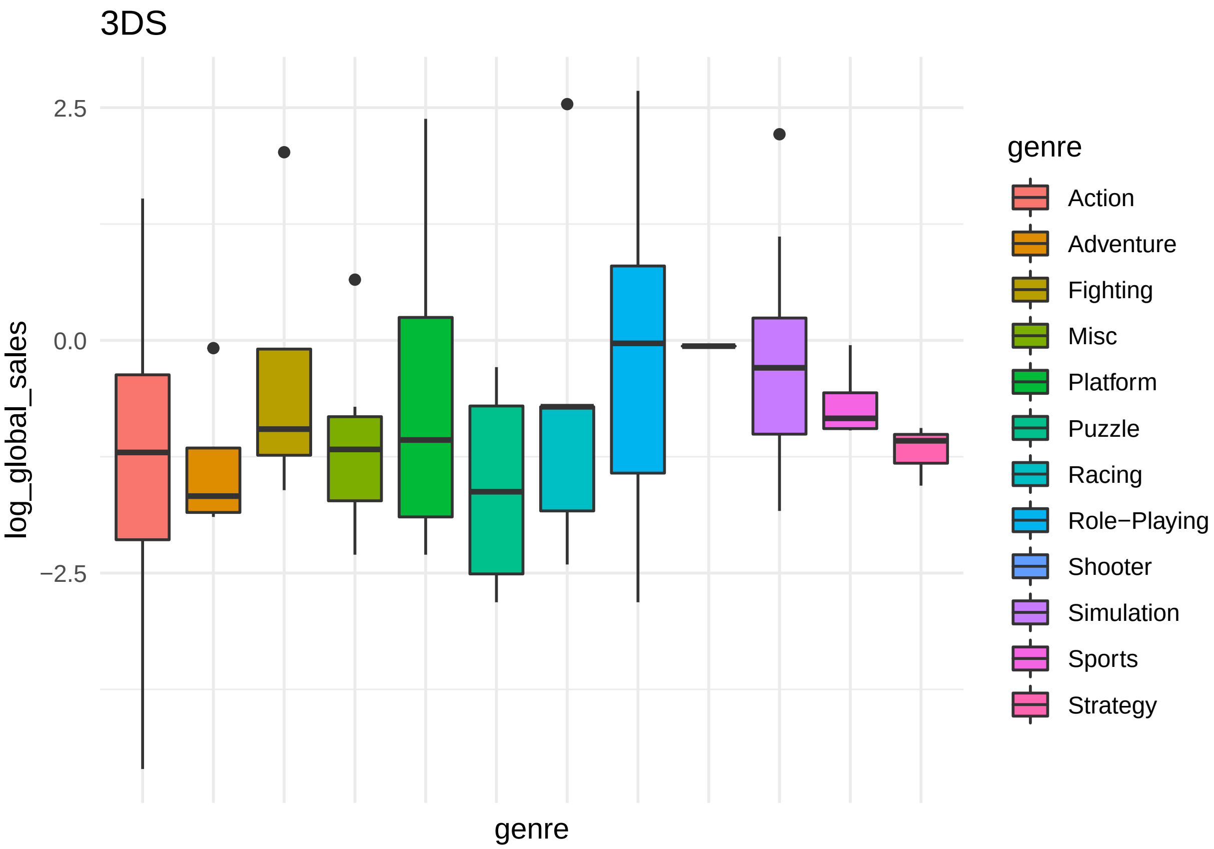
A summary of the data variables being analyzed can be found in Annex 1.1. An exploratory data analysis for all variables and plots for their interactions can be found in Annex 1.2.

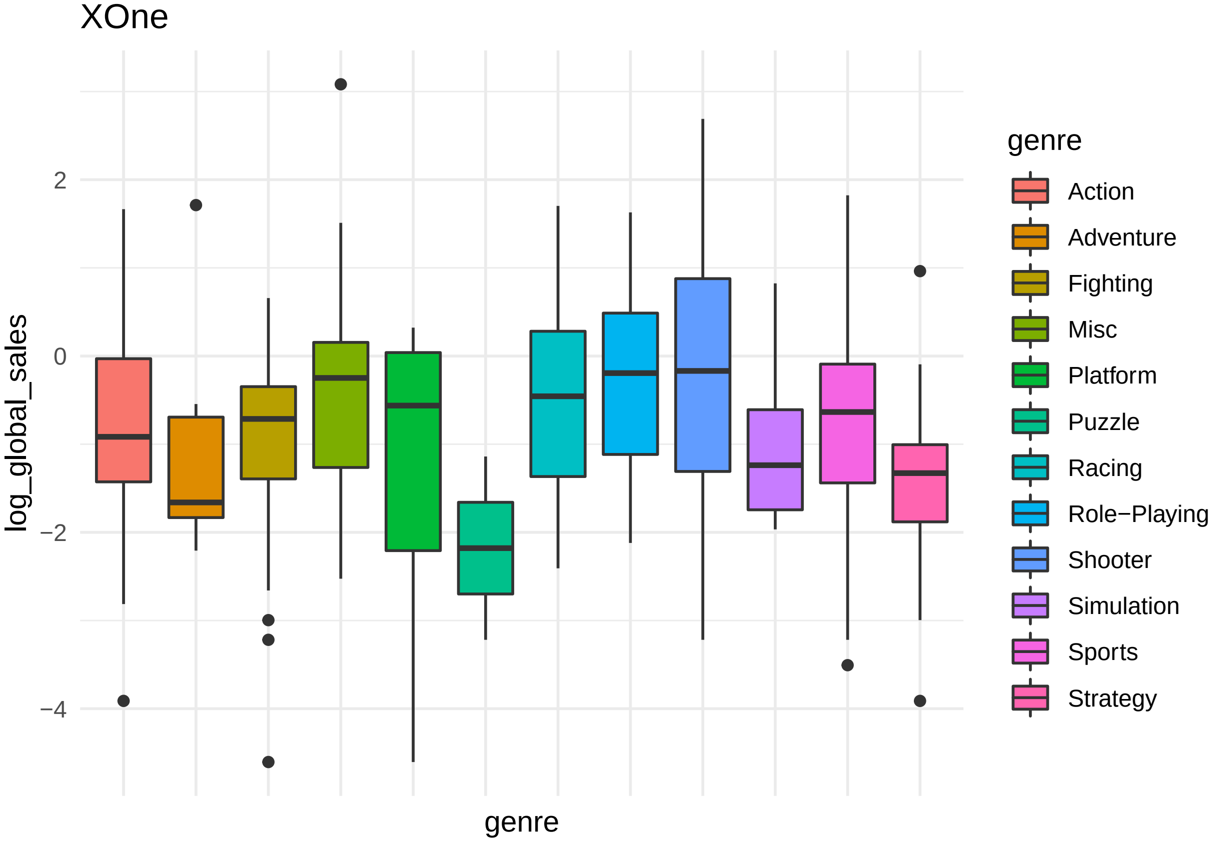
The EDA suggests none of the numerical variables have a clear association with premature as the boxplots for premature = 0 and premature = 1 do not have noticeable differences. For the categorical varibles, there are more interesting results in the conditional probability tables for each variable and 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.

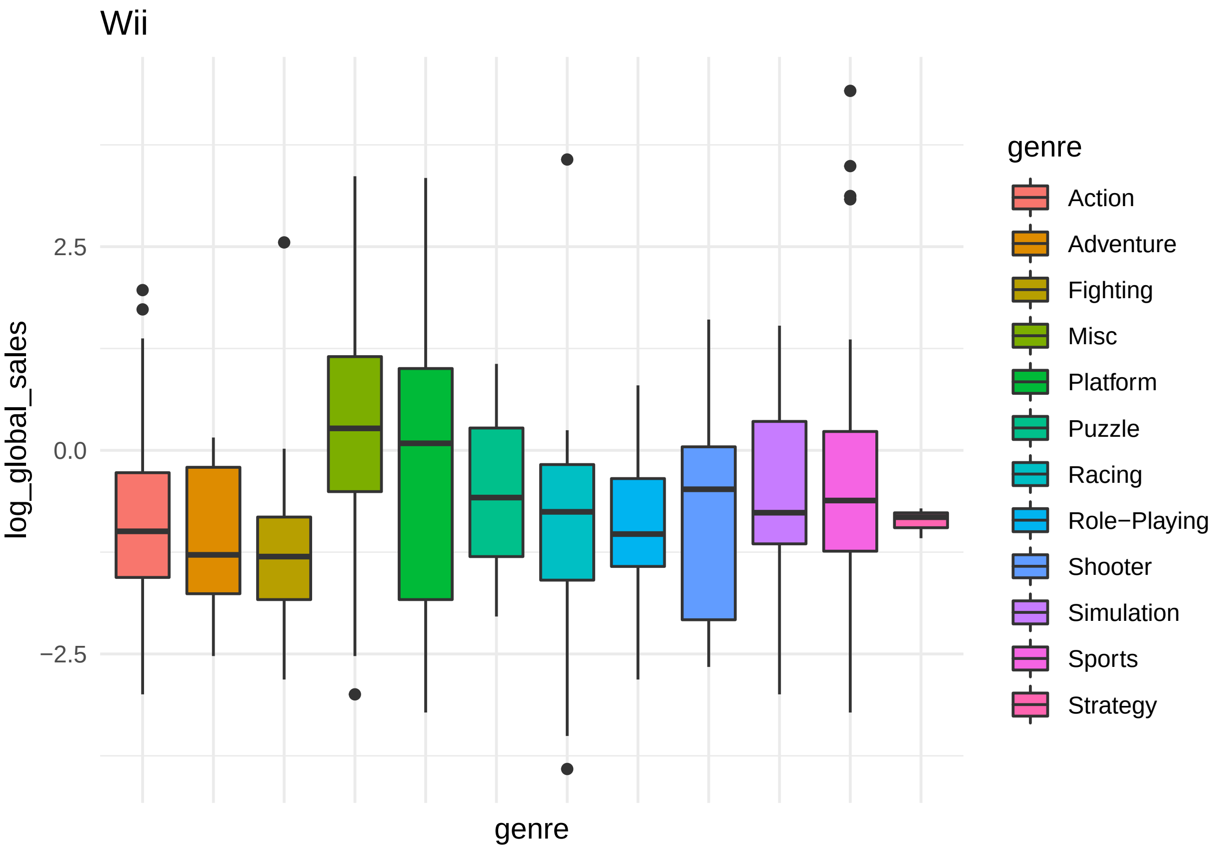
```r
#Selecting 50 sample publishers
publishers <- unique(vgsales$publisher)
set.seed(2163386)
sample_publishers<- sample(publishers, 50)
sample_data <- vgsales[vgsales$publisher %in% sample_publishers,]
```

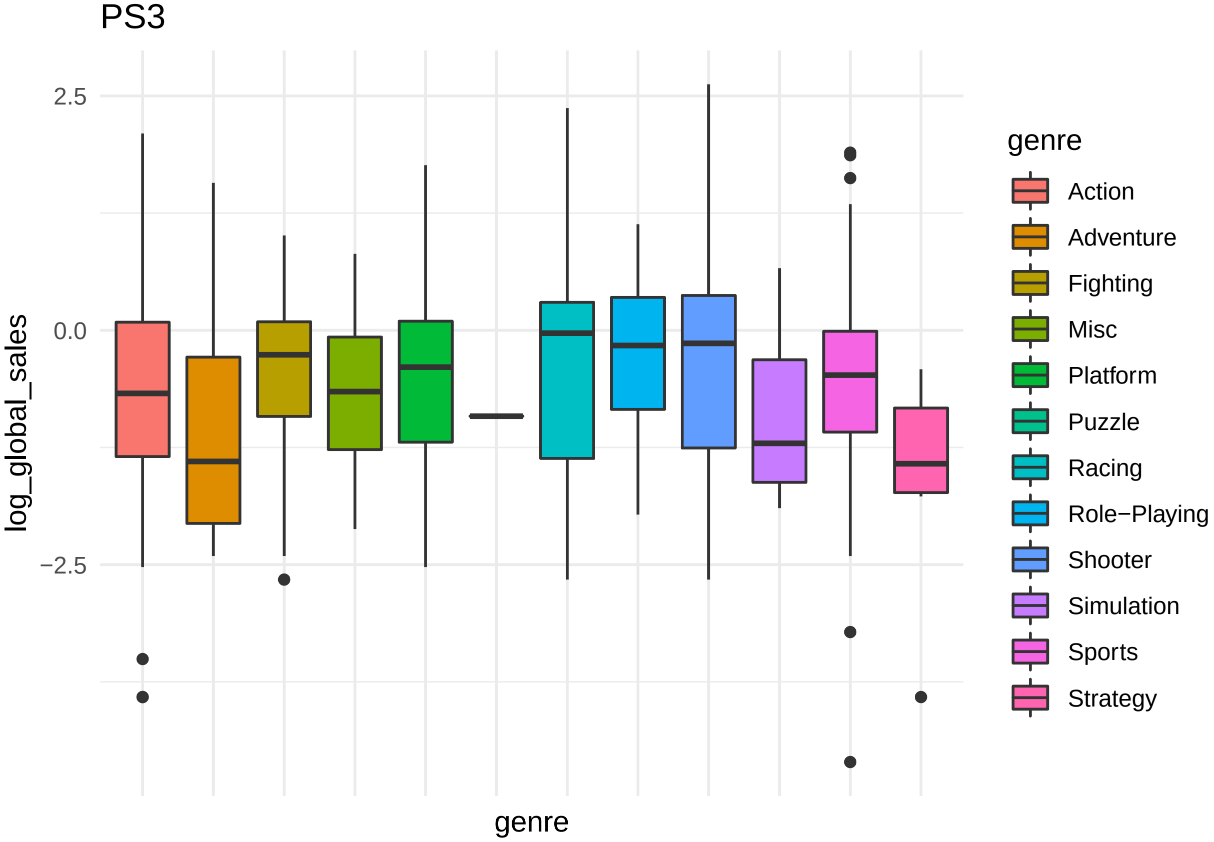

count
3000
2000
1000
0
0
20
40
60
80
global_sales


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

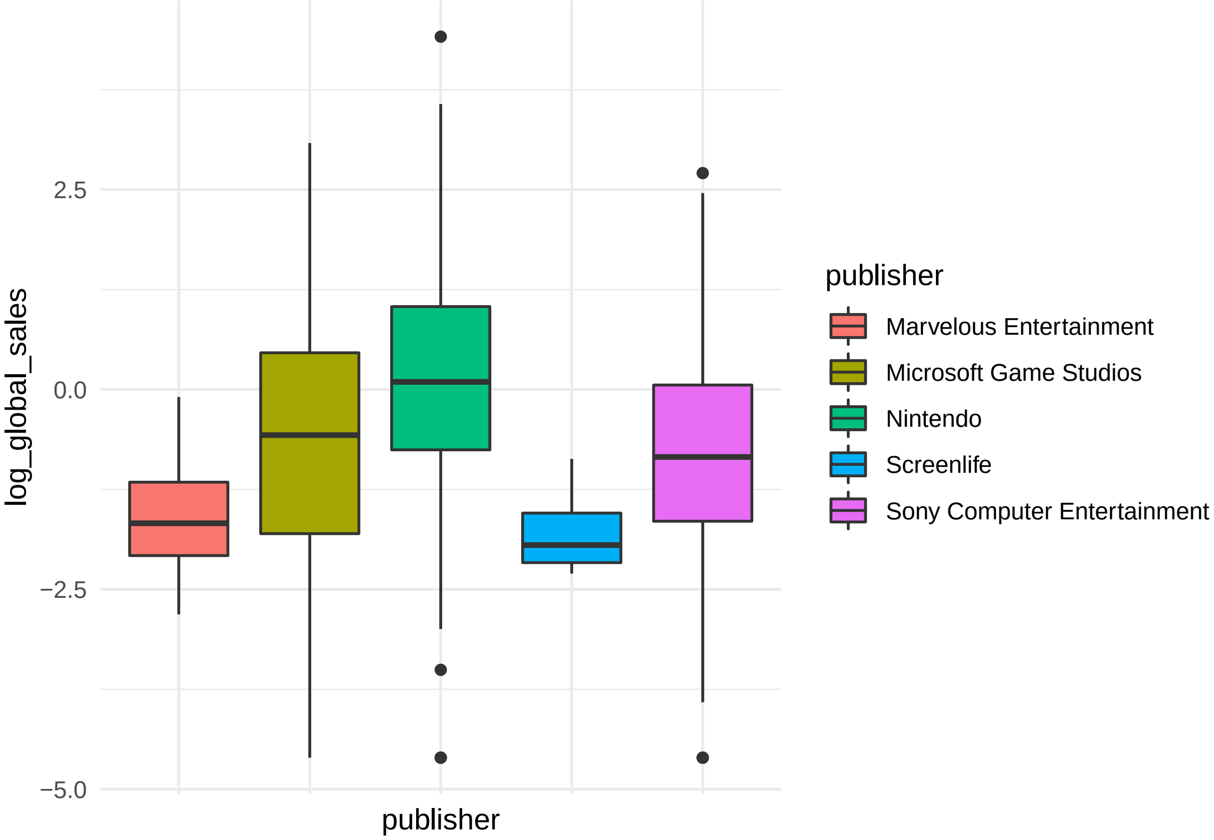

3DS
log_global_sales
2.5
0.0
−2.5
genre
genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role−Playing
Shooter
Simulation
Sports
Strategy

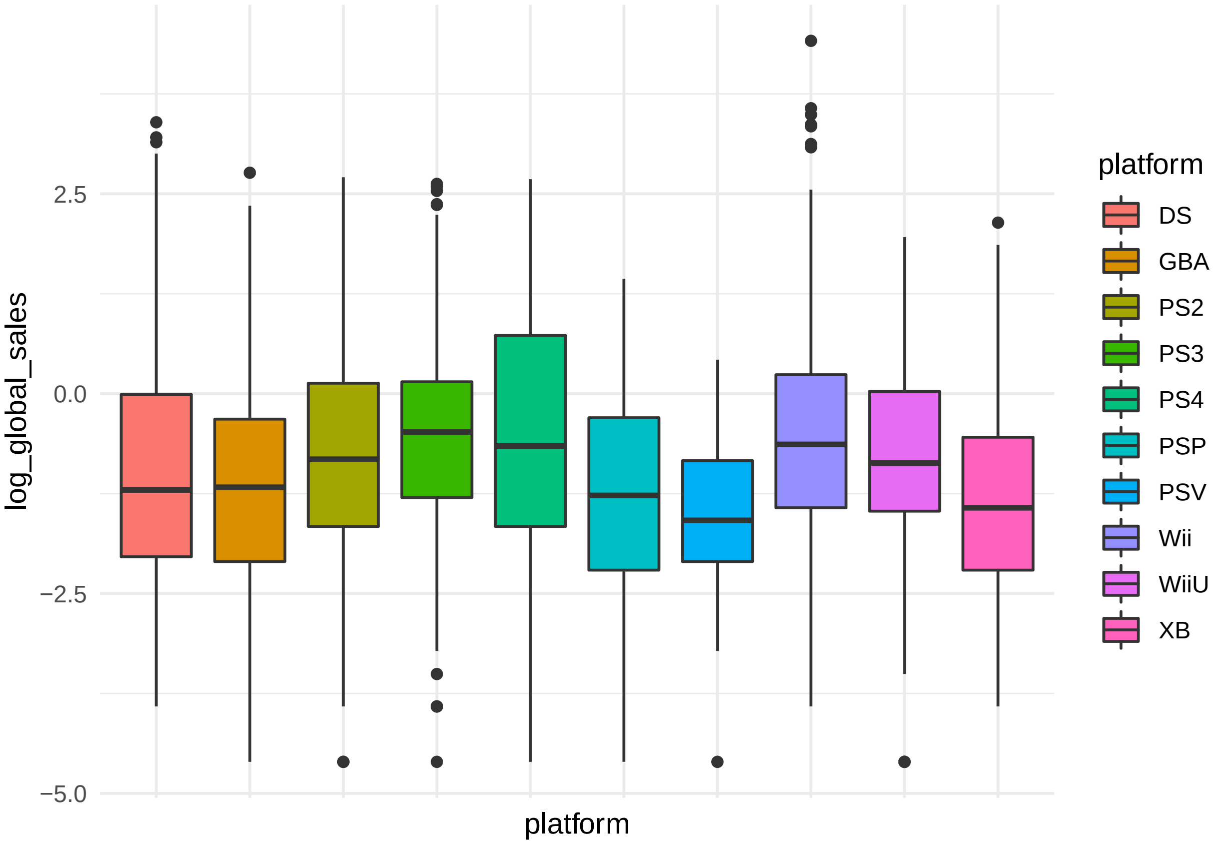

XOne
log_global_sales
2
0
−2
−4
genre
genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role−Playing
Shooter
Simulation
Sports
Strategy

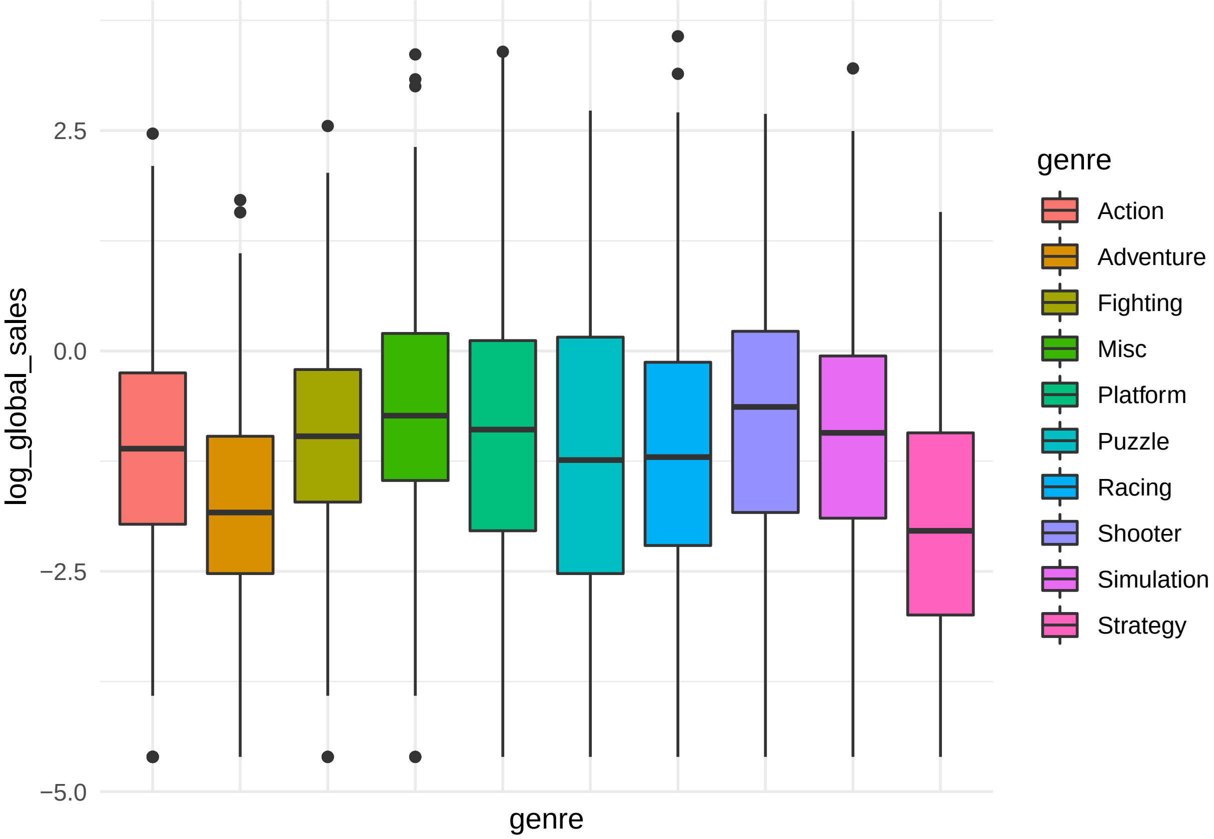


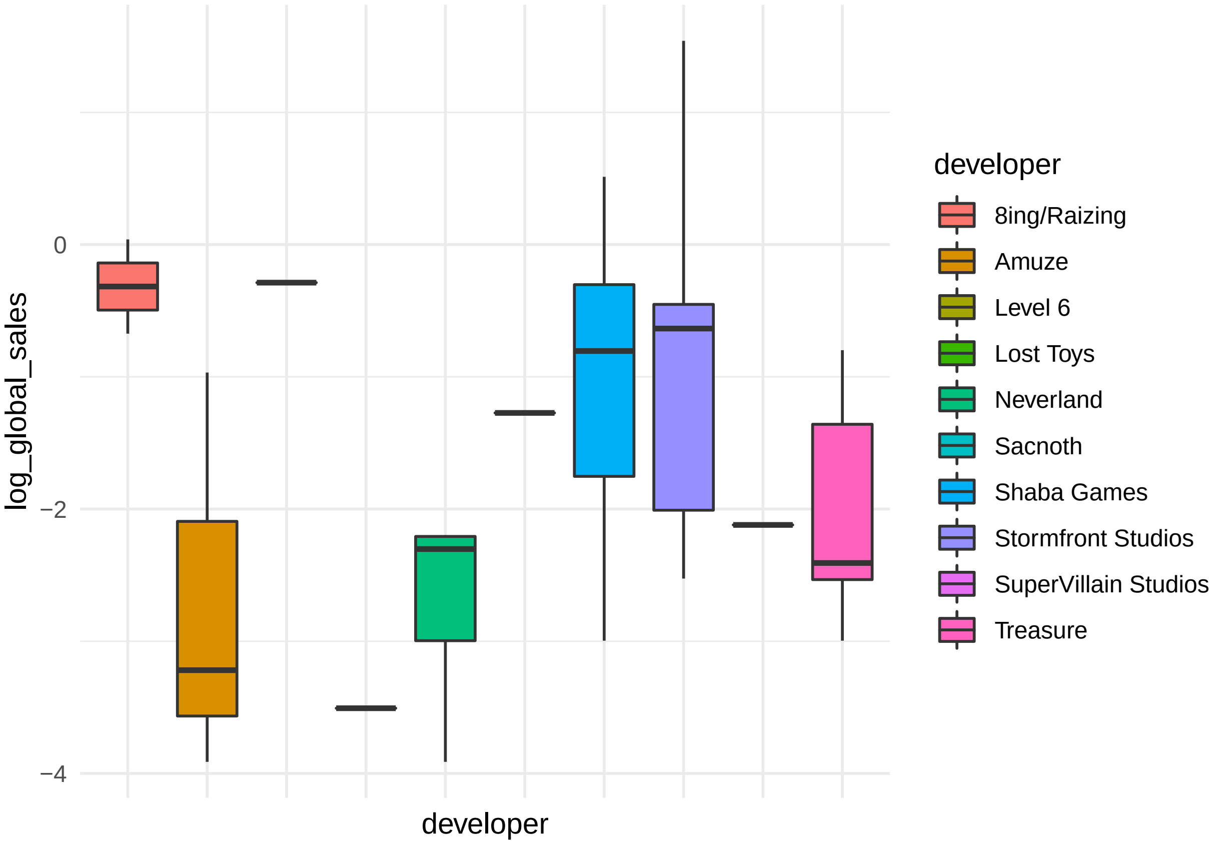


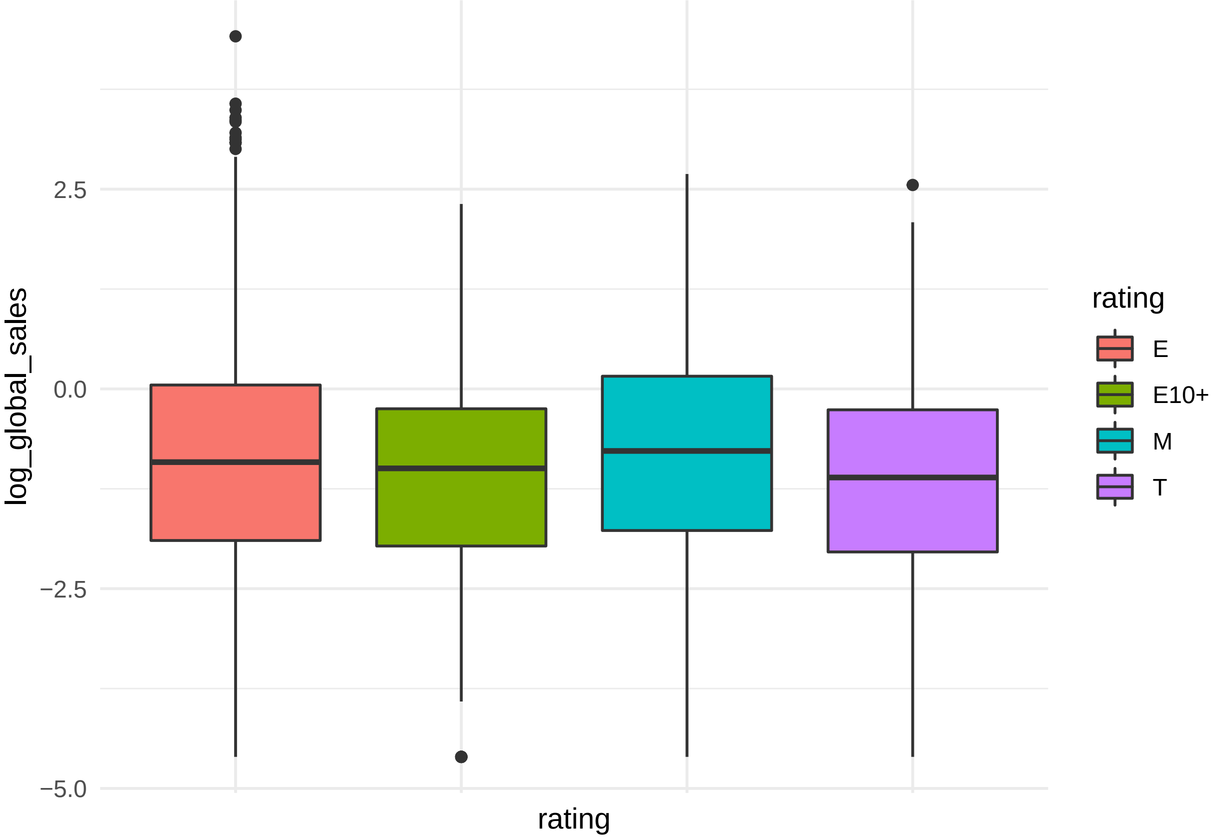
```
## [1] Nintendo                          Screenlife
## [3] Sony Computer Entertainment Microsoft Game Studios
## [5] Marvelous Entertainment
## 444 Levels: 10TACLE Studios 1C Company 2D Boy 2K Sports 3DO ... Zushi Games
```



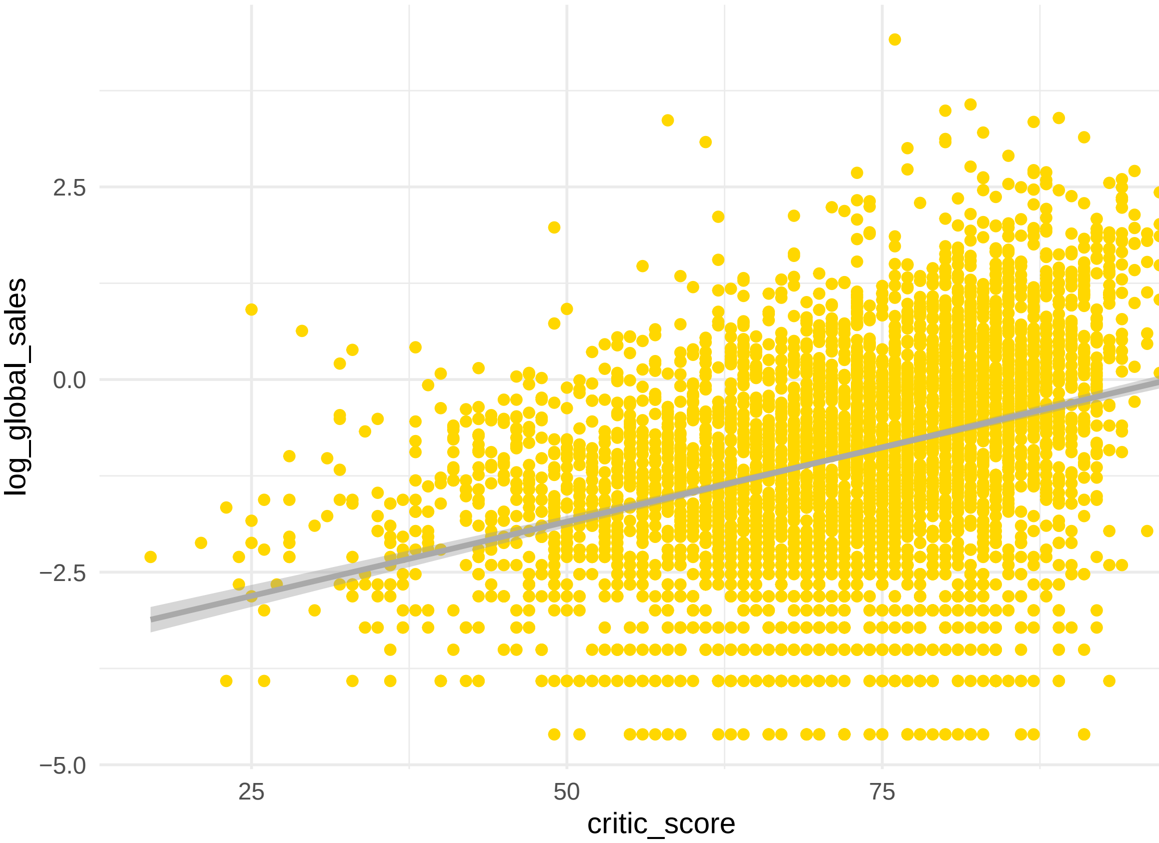

log_global_sales
2.5
0.0
−2.5
−5.0
platform
platform
DS
GBA
PS2
PS3
PS4
PSP
PSV
Wii
WiiU
XB


log_global_sales
2.5
0.0
−2.5
−5.0
genre
genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Shooter
Simulation
Strategy

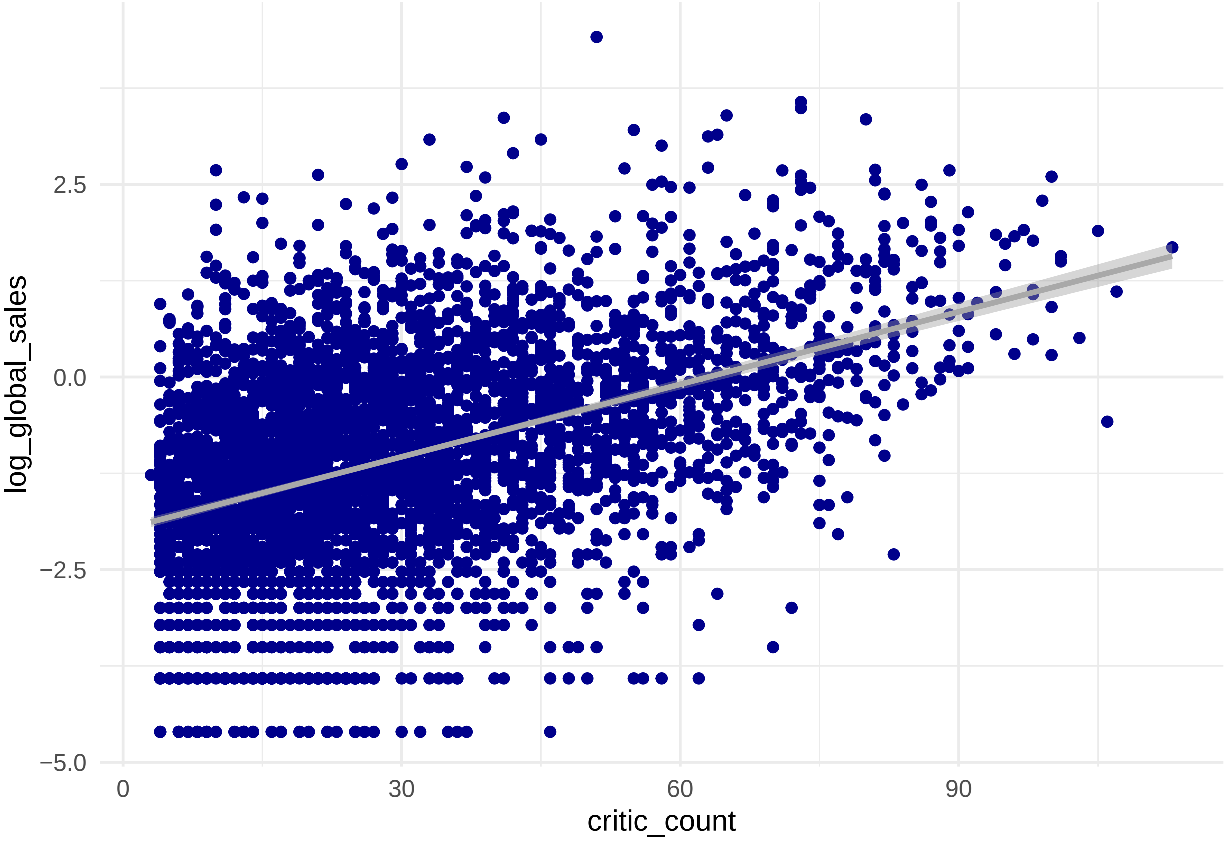


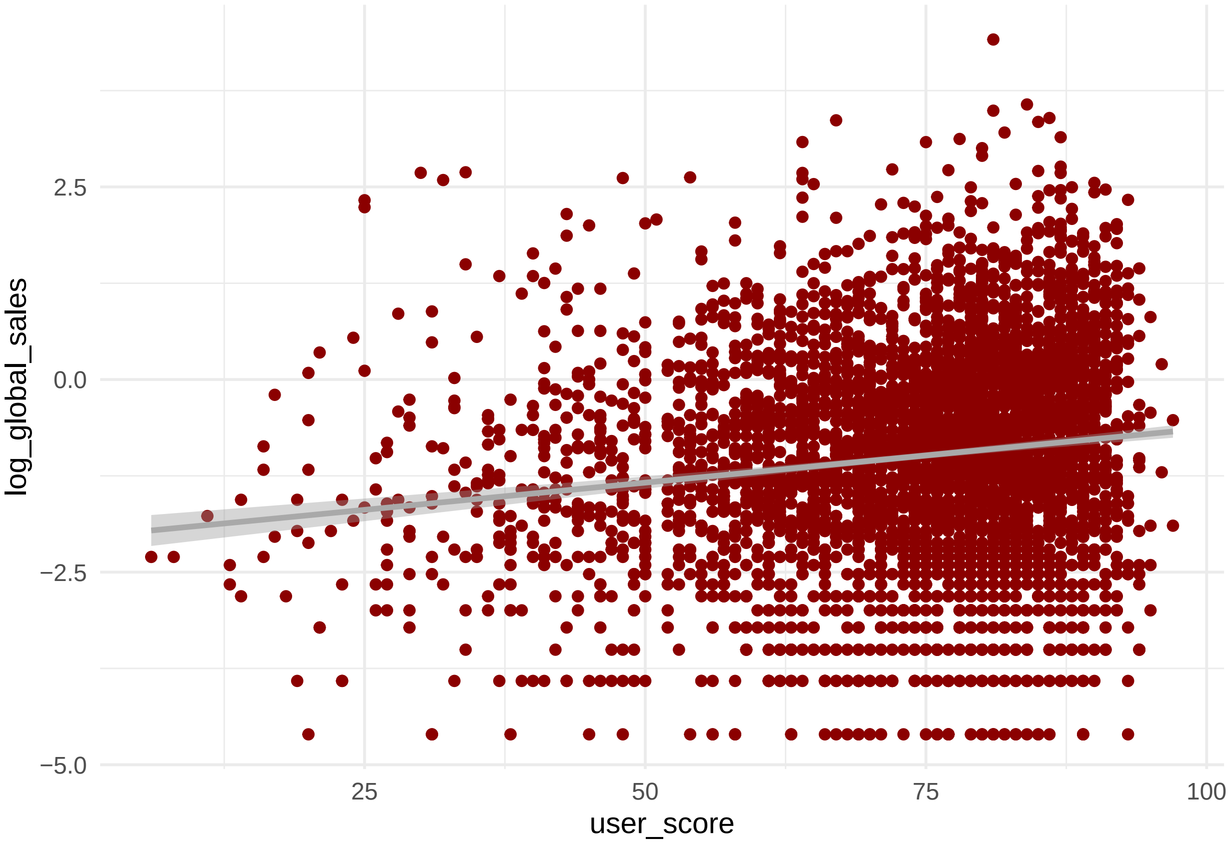


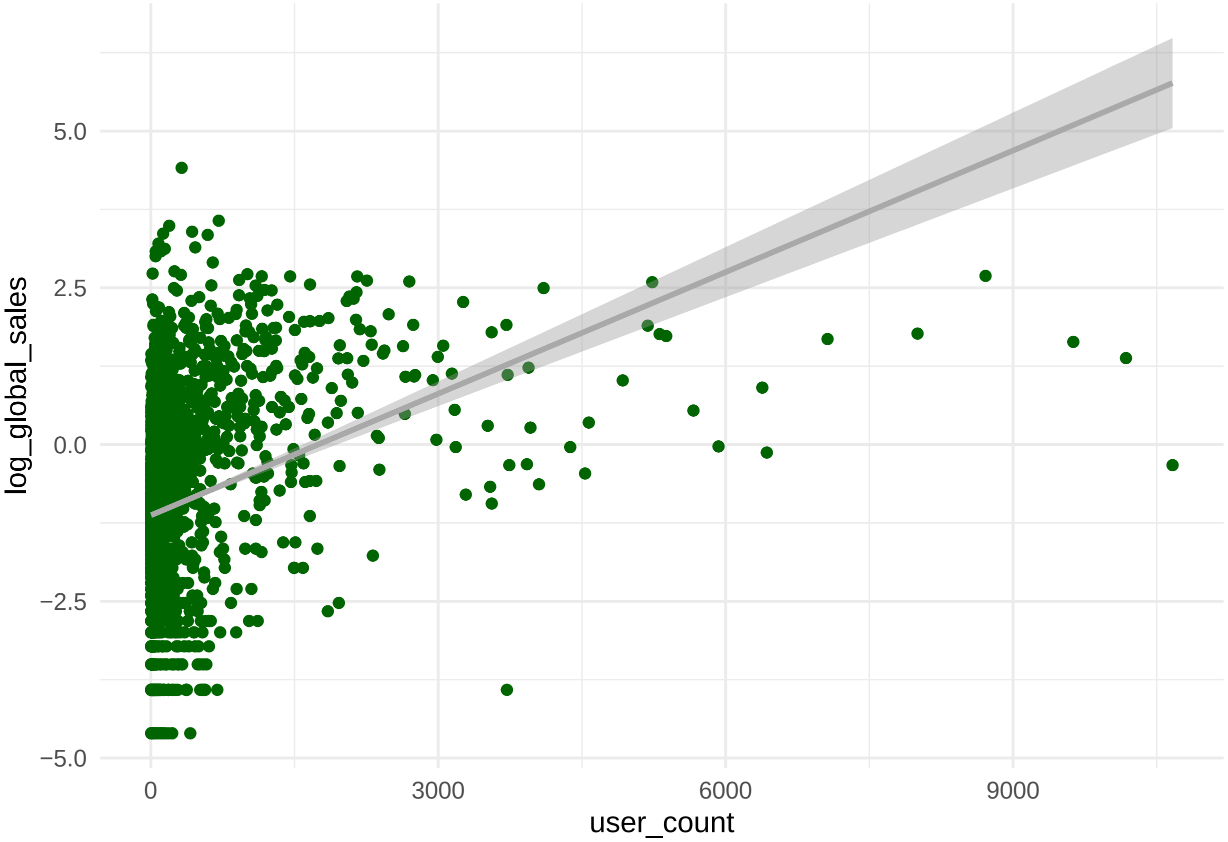
Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code



that generated the plot.







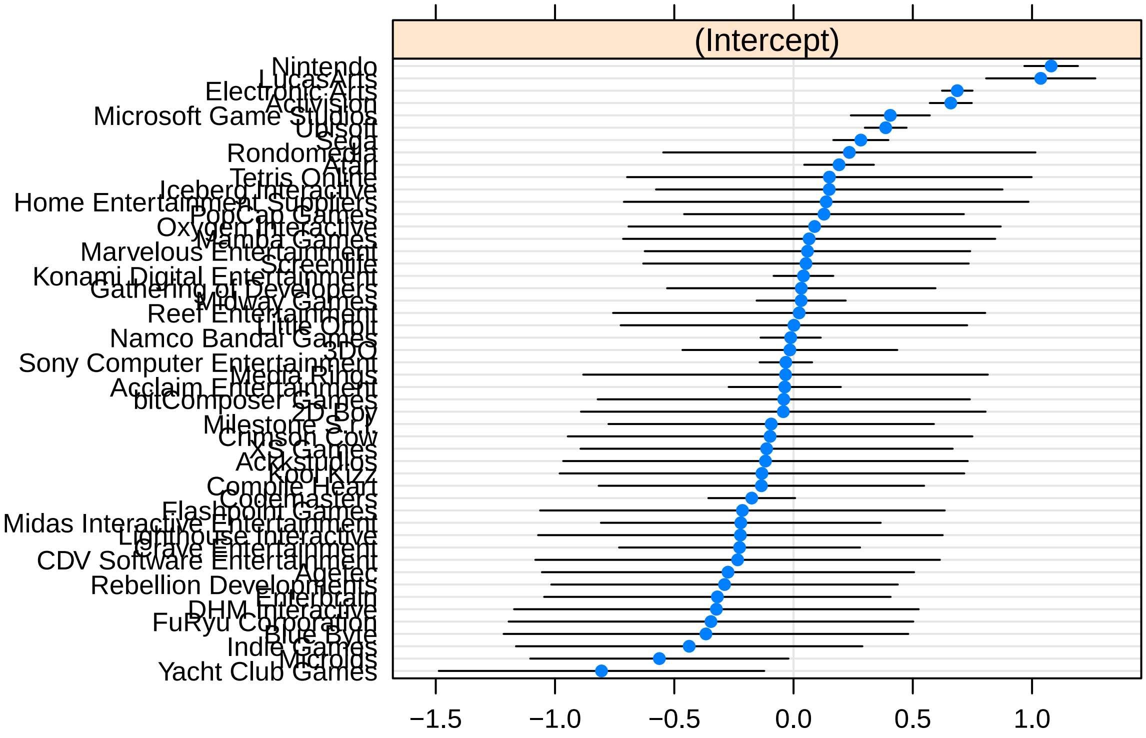
```
## Warning: 'r.squaredGLMM' now calculates a revised statistic. See the help
## page.
```

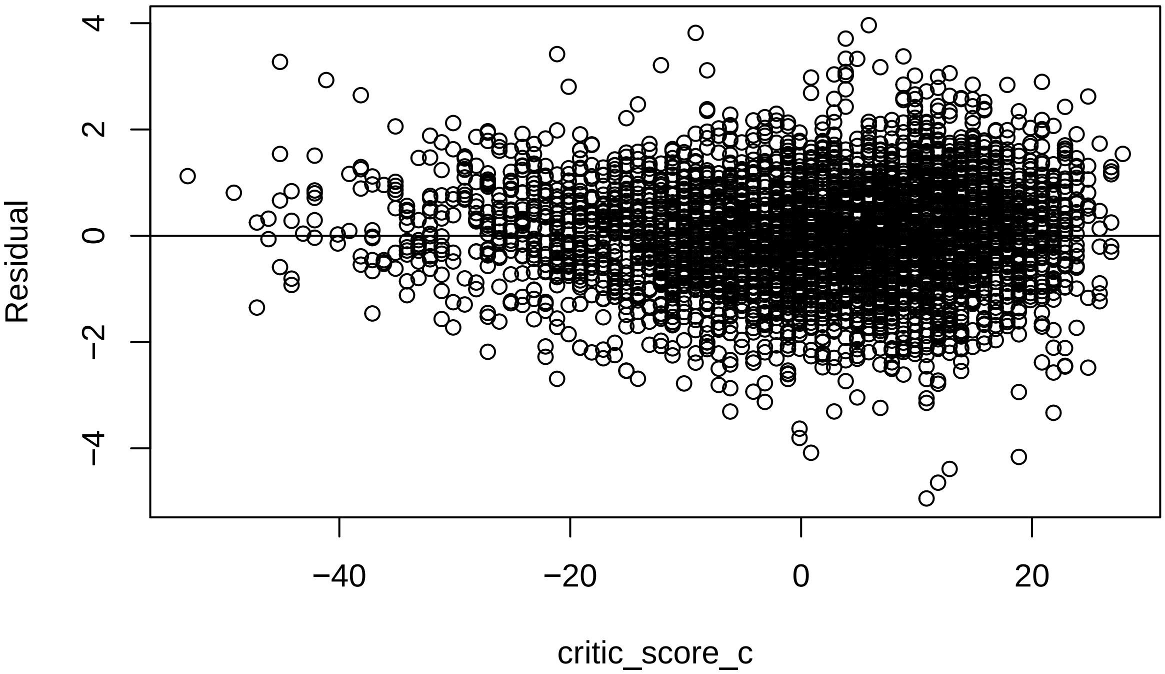
```
##            R2m       R2c
## [1,] 0.3747508 0.4884808
```

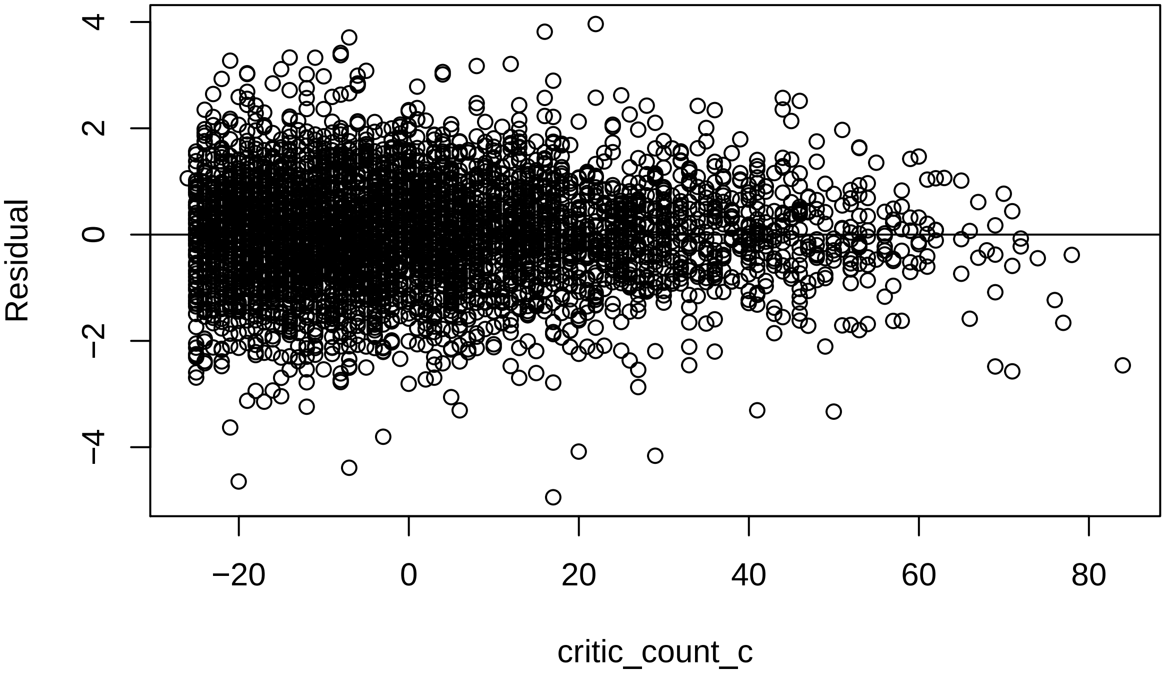
```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: log_global_sales ~ platform_company + genre + rating_everyone +
##     critic_score_c + critic_count_c + user_count_c + platform_company:rating_everyone +
##     (1 | publisher)
##    Data: sample_data
##
## REML criterion at convergence: 12183.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.8705 -0.6314  0.0026  0.6406  3.9042
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  publisher (Intercept) 0.2289   0.4785
##  Residual              1.0297   1.0147
## Number of obs: 4195, groups:  publisher, 50
##
## Fixed effects:
##                                             Estimate Std. Error         df
## (Intercept)                               -1.677e+00  1.060e-01  5.701e+01
## platform_companyNintendo                   6.336e-02  6.240e-02  4.169e+03
## platform_companyPC                        -1.461e+00  7.816e-02  4.169e+03
## platform_companySega                      -7.114e-01  3.926e-01  4.139e+03
## platform_companySony                       4.508e-01  5.227e-02  4.166e+03
## genreAdventure                            -3.583e-01  1.010e-01  4.137e+03
## genreFighting                              2.753e-01  8.339e-02  4.161e+03
## genreMisc                                  4.669e-01  7.198e-02  4.159e+03
## genrePlatform                              2.587e-02  7.469e-02  4.157e+03
## genrePuzzle                               -4.509e-01  1.370e-01  4.116e+03
## genreRacing                                7.007e-02  6.778e-02  4.171e+03
## genreRole-Playing                         -1.105e-01  6.860e-02  4.171e+03
## genreShooter                               7.236e-02  5.719e-02  4.155e+03
## genreSimulation                            4.164e-01  8.114e-02  4.158e+03
## genreSports                                5.543e-02  6.126e-02  4.162e+03
## genreStrategy                             -4.753e-01  1.000e-01  4.169e+03
## rating_everyone1                           1.604e-03  7.904e-02  4.159e+03
## critic_score_c                             2.349e-02  1.370e-03  4.166e+03
## critic_count_c                             1.966e-02  1.058e-03  4.171e+03
## user_count_c                               4.560e-04  3.068e-05  4.144e+03
## platform_companyNintendo:rating_everyone1  2.906e-01  9.567e-02  4.156e+03
## platform_companyPC:rating_everyone1       -2.282e-04  1.537e-01  4.170e+03
## platform_companySega:rating_everyone1      4.087e-01  7.053e-01  4.135e+03
## platform_companySony:rating_everyone1      1.338e-01  8.840e-02  4.151e+03
##                                           t value Pr(>|t|)
## (Intercept)                               -15.824  < 2e-16 ***
## platform_companyNintendo                    1.015 0.310036
## platform_companyPC                        -18.692  < 2e-16 ***
## platform_companySega                       -1.812 0.070067 .
## platform_companySony                        8.624  < 2e-16 ***
## genreAdventure                             -3.546 0.000395 ***
## genreFighting                               3.302 0.000970 ***
## genreMisc                                   6.486 9.82e-11 ***
## genrePlatform                               0.346 0.729115    
## genrePuzzle                                -3.293 0.001001 ** 
## genreRacing                                 1.034 0.301271    
## genreRole-Playing                          -1.611 0.107329    
## genreShooter                                1.265 0.205844    
## genreSimulation                             5.132 3.00e-07 ***
## genreSports                                 0.905 0.365564    
## genreStrategy                              -4.752 2.09e-06 ***
## rating_everyone1                            0.020 0.983804    
## critic_score_c                             17.151  < 2e-16 ***
## critic_count_c                             18.584  < 2e-16 ***
## user_count_c                               14.862  < 2e-16 ***
## platform_companyNintendo:rating_everyone1   3.037 0.002403 ** 
## platform_companyPC:rating_everyone1        -0.001 0.998815    
## platform_companySega:rating_everyone1       0.579 0.562321    
## platform_companySony:rating_everyone1       1.514 0.130075    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

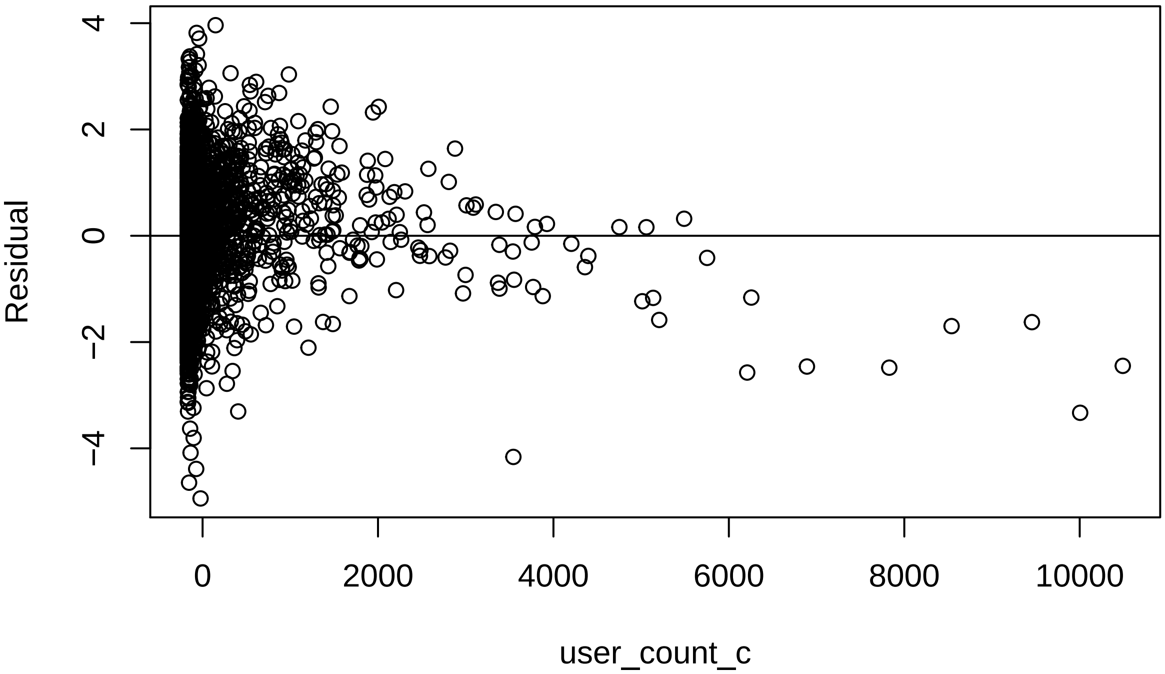
```
## 
## Correlation matrix not shown by default, as p = 24 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

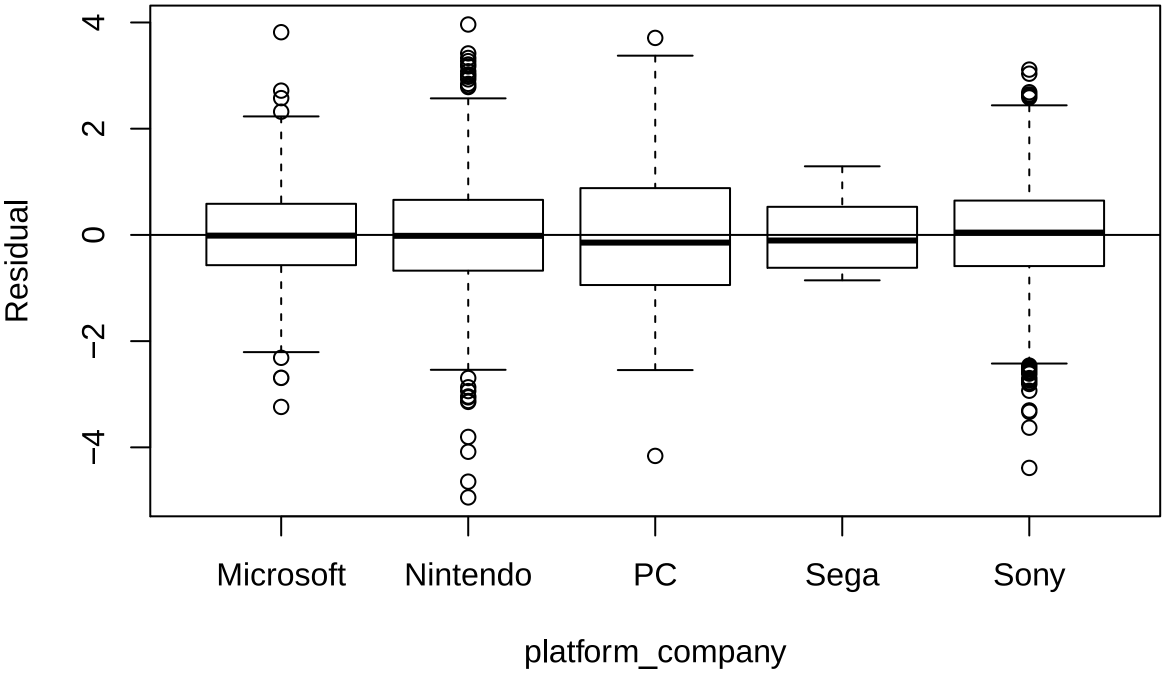
```
##                 platform_companyNintendo 
##                                 2.336460 
##                       platform_companyPC 
##                                 1.863758 
##                     platform_companySega 
##                                 1.455973 
##                     platform_companySony 
##                                 2.129665 
##                           genreAdventure 
##                                 1.106568 
##                            genreFighting 
##                                 1.179787 
##                                genreMisc 
##                                 1.255714 
##                            genrePlatform 
##                                 1.269283 
##                              genrePuzzle 
##                                 1.098215 
##                              genreRacing 
##                                 1.425356 
##                        genreRole-Playing 
##                                 1.293752 
##                             genreShooter 
##                                 1.428571 
##                          genreSimulation 
##                                 1.222977 
##                              genreSports 
##                                     2.011325
##                                genreStrategy
##                                     1.153686
##                           rating_everyone1
##                                     5.137204
##                               critic_score_c
##                                     1.272156
##                               critic_count_c
##                                     1.577106
##                                 user_count_c
##                                     1.330988
## platform_companyNintendo:rating_everyone1
##                                     3.450623
##       platform_companyPC:rating_everyone1
##                                     1.573152
##     platform_companySega:rating_everyone1
##                                     1.434711
##     platform_companySony:rating_everyone1
##                                     3.272168
```

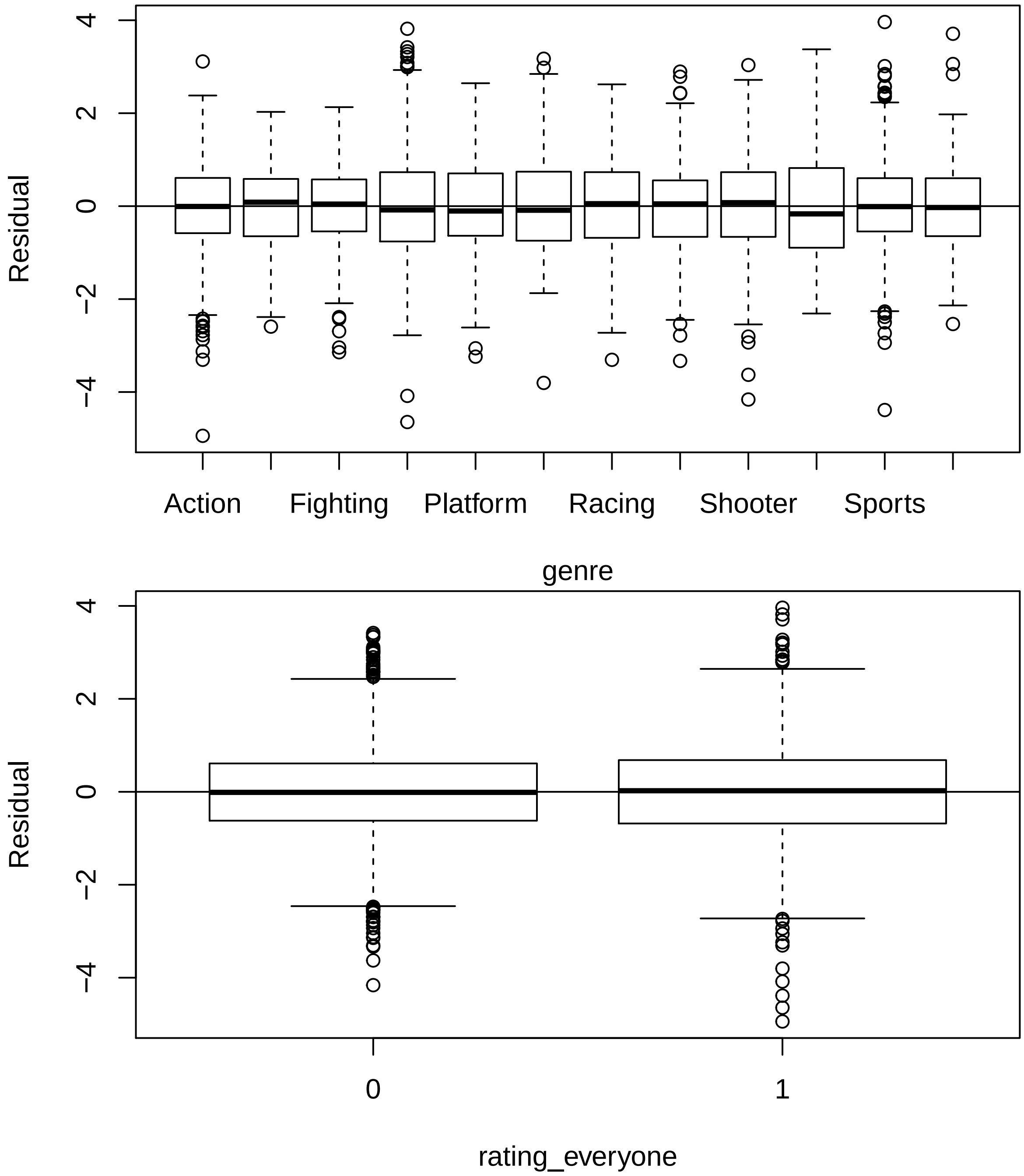


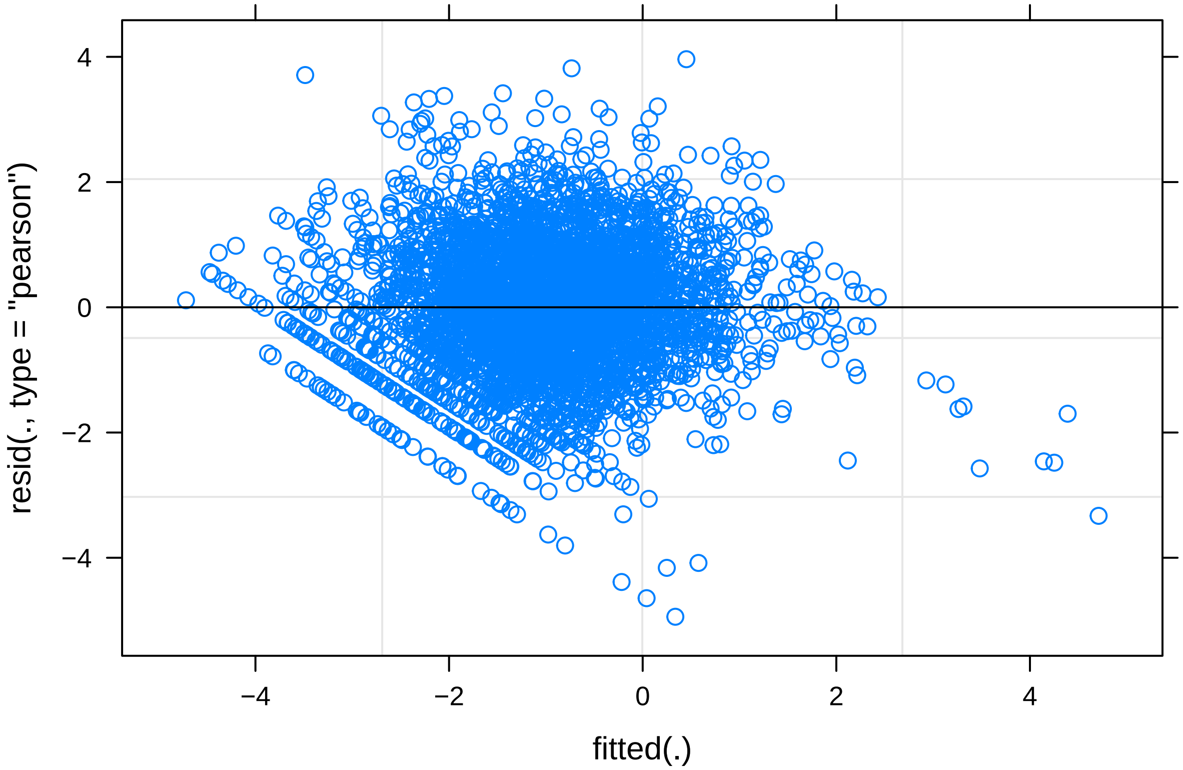

Residual
4
2
0
−2
−4
−40
−20
0
20
critic_score_c

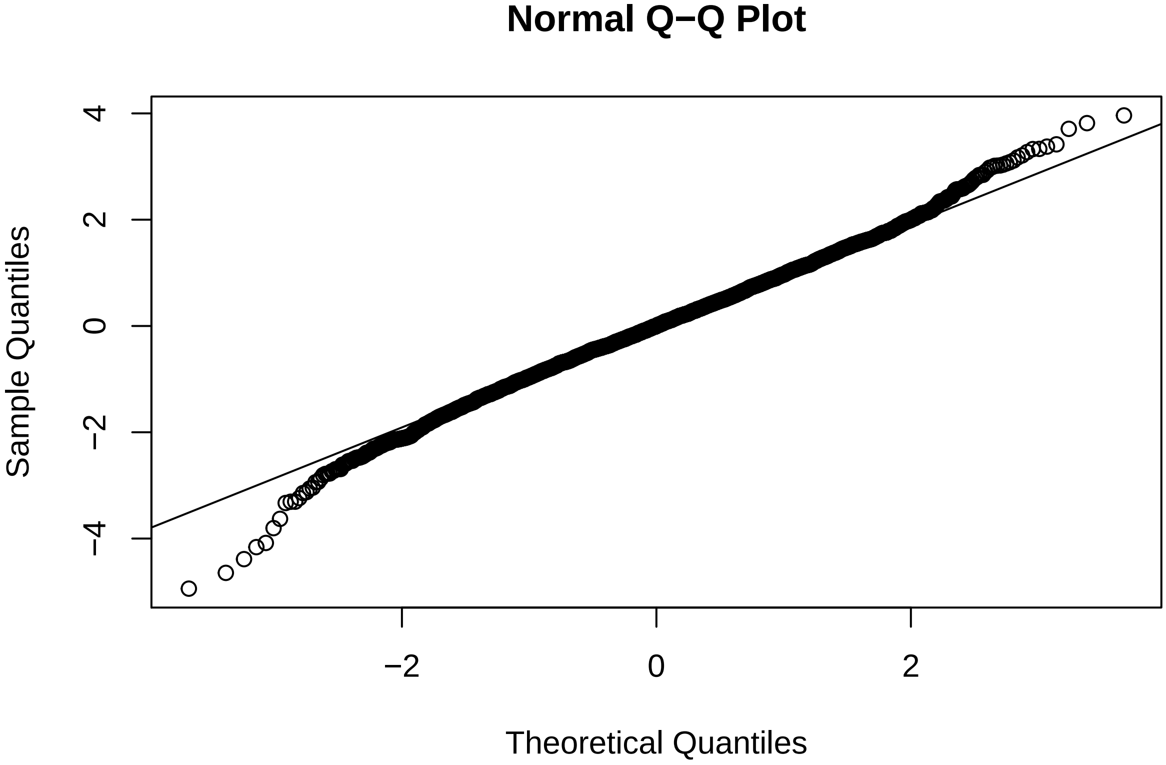

Residual
4
2
0
−2
−4
−20
0
20
40
60
80
critic_count_c


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")
fitted(.)
4
2
0
−2
−4
−4
−2
0
2
4


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