Final Project

Sebastián Soriano Pérez [ss1072]

12/10/2019

Analyzing Video Game Sales

Summary

The video game industry has been growing consistently during the last two decades, and in 2017 it was worth more than 78 billion USD worldwide. Video game software sales account for around 80% of total revenue. There are several factors that influence whether a video game will be successful or not, such as the developer studio, the critics rating, the user rating, among others. I will build a model to analyze what factors can help us determine how successful will a video game be in terms of global sales.
I analyzed a total of 4195 video game software releases across the world by 50 different developers. A hierarchical linear model on a logarithmic transformation of the global sales, with random slopes by developer, was selected using a manual stepwise selection process using BIC and conditional R-squared as selection criteria. The variables on the final model were found to be significant in predicting the global sales for a video game.

Introduction

This document presents a model to identify which factors can help determine the global sales for a particular video game release. The goal of this project is to find what are the most significant predictors for a video games success, measured as the number of sales around the world. I use a hierarchical linear regression model to explain the number of sales a particular videogame has. Considering that the developing studio plays a major role on a customer's decision to buy a new game, and due to the fact that accounting for every single studio included in the dataset (444 total) would make the interpretation too complicated, I used a random sample of 50 developers and built an appropriate hierarchical model with random intercepts effects for each one of them.

Data

The data was obtained from Kaggle (https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings). It contains data from video games with 100,000 or more global sales from 1976 to 2016. The data contains 16719 rows with the following columns:

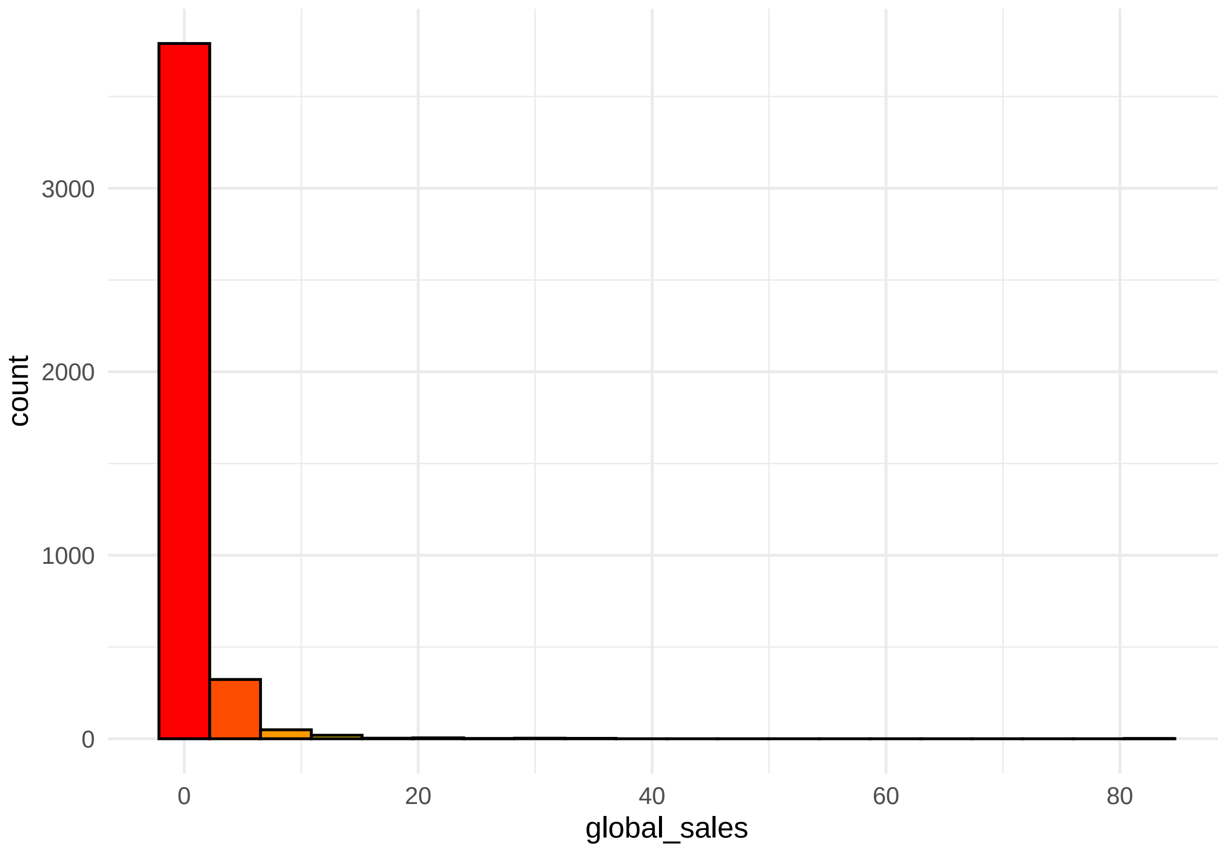
- *name* (categorical): Name of the video game
- *platform* (categorical): PLatform or console for which the video game was released
- *year_of_release* (categorical): Year of first release
- *genre* (categorical): Genre of the video game
- *publisher* (categorical): Publishing company
- *na_sales* (numerical): Units sold in North America
- *eu_sales* (numerical): Units sold in Europe
- *jp_sales* (numerical): Units sold in Japan
- *other_sales* (numerical): Units sold in the rest of the world
- *global_sales* (numerical): Total units sold worldwide
- *critic_score* (numerical): Average score (from 0 to 100) according to critics from other media aggregated by Metacritic
- *critic_count* (numerical): Number of critics taken into account for the Metacritic critic score
- *user_score* (numerical): Average score (from 0 to 100) according to Metacritic users
- *user_count* (numerical): Number of user scores on Metacritic
- *developer* (categorical): Video game developing company

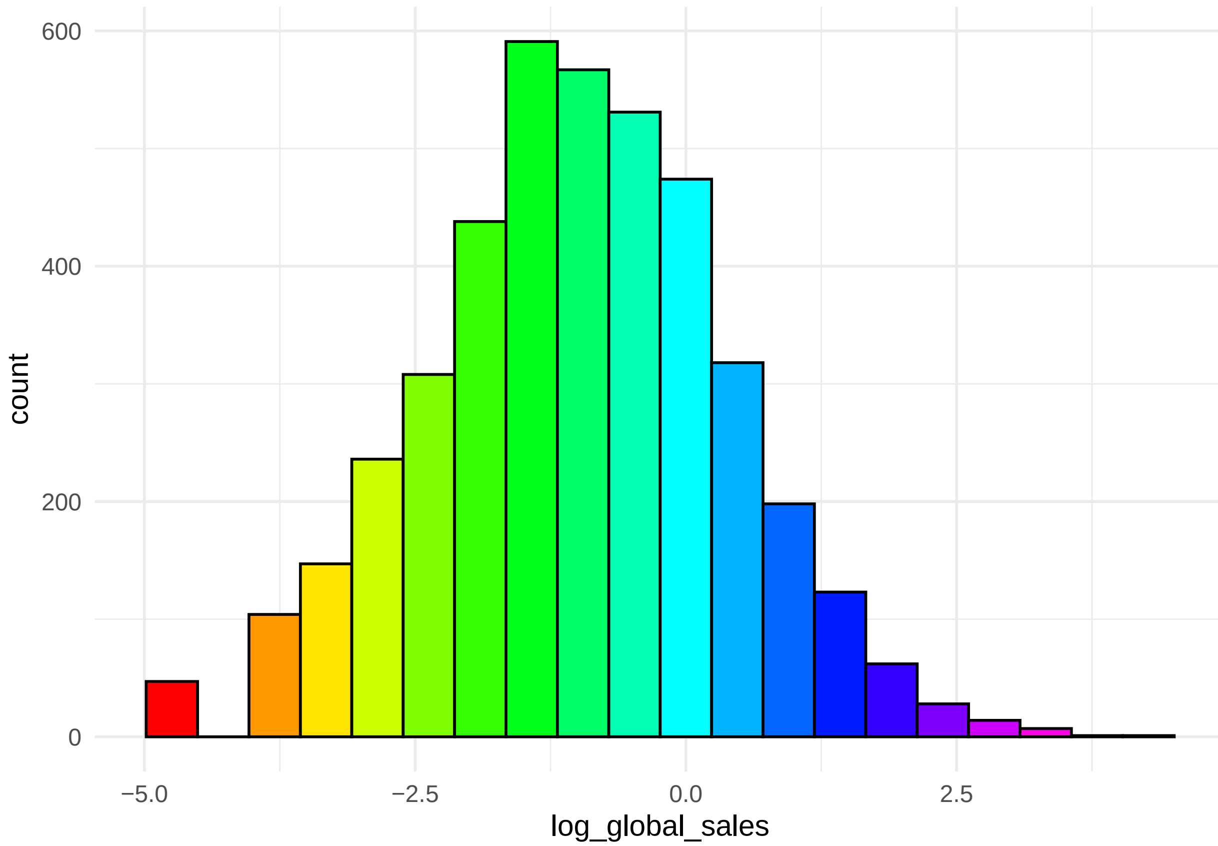
- *rating* (categorical): Video game rating according to the ESRB that indicates the appropriate audience

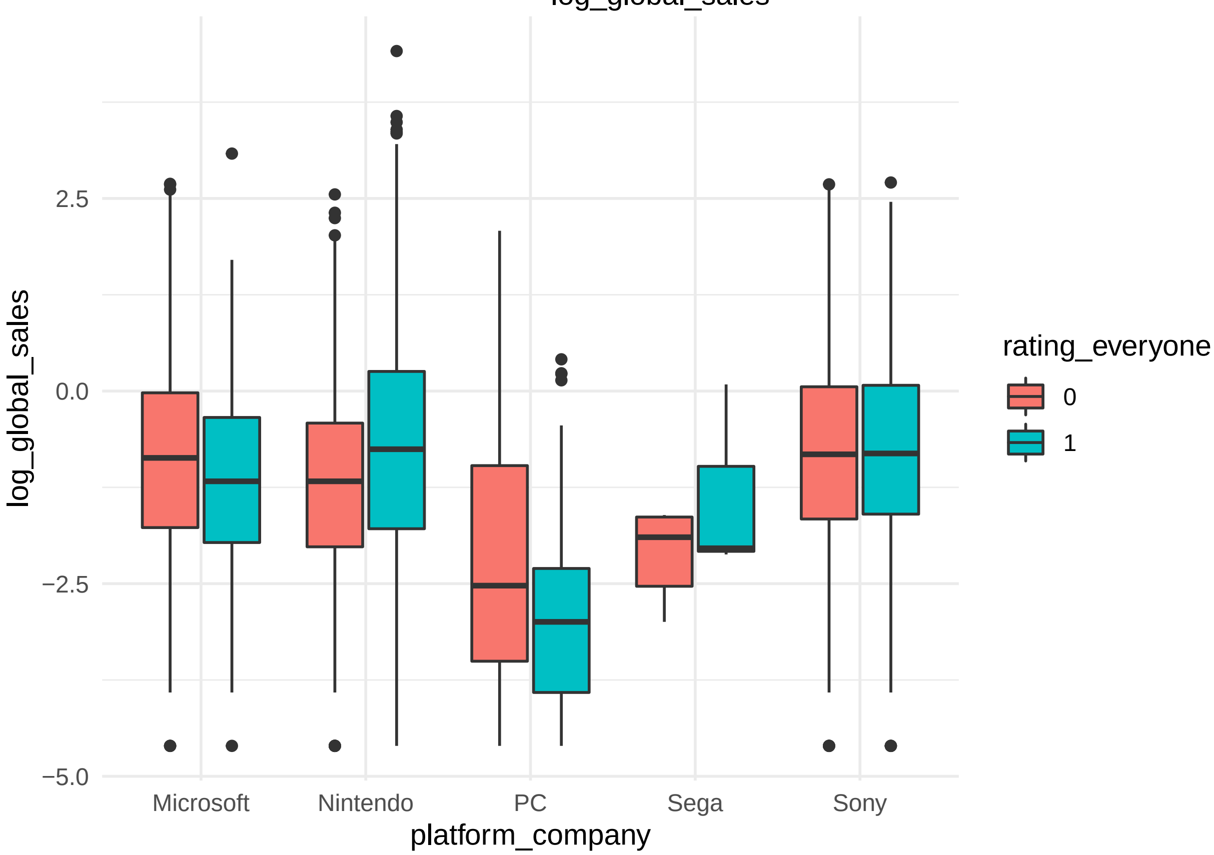
A summary of the data variables being analyzed can be found in Annex 1.1. An exploratory data analysis for all variables and plots for their interactions can be found in Annex 1.2.

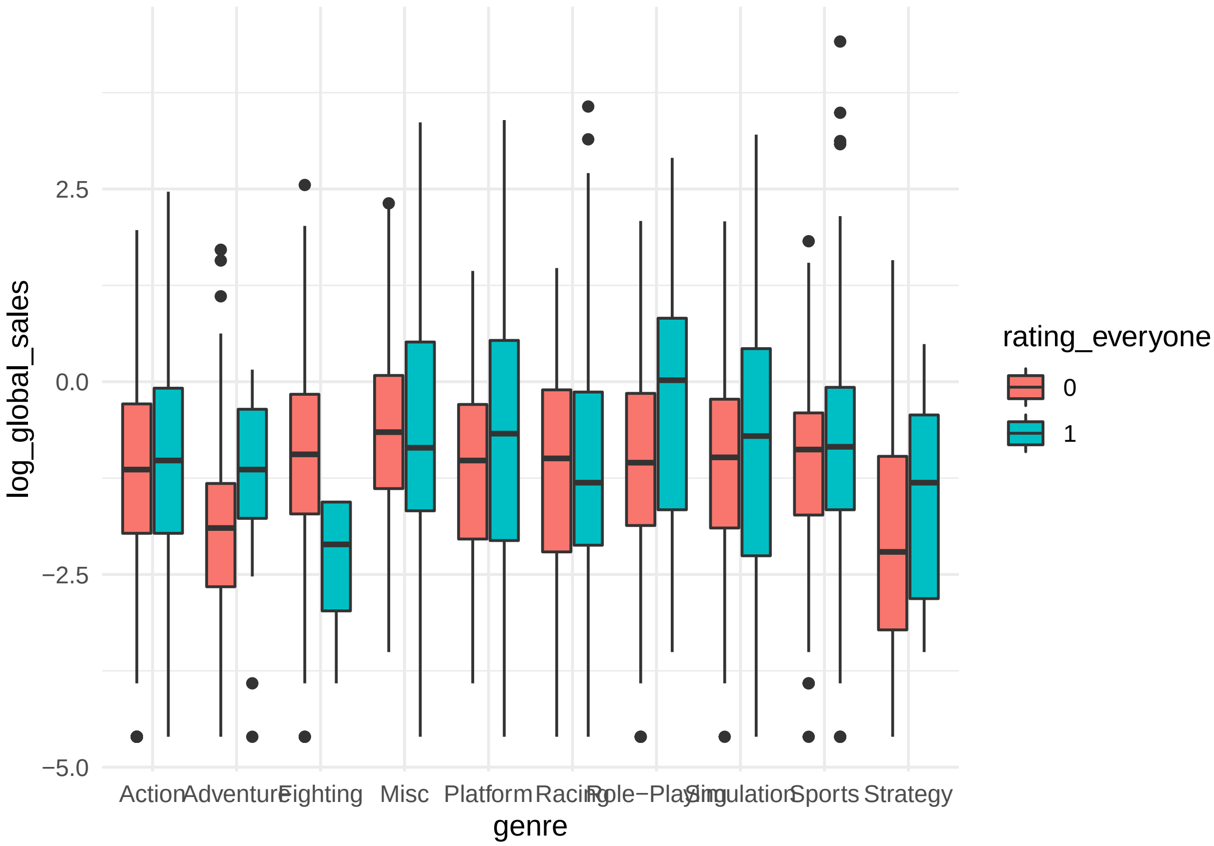
The EDA suggests none of the numerical variables have a clear association with premature as the boxplots for premature = 0 and premature = 1 do not have noticeable differences. For the categorical varibles, there are more interesting results in the conditional probability tables for each variable and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

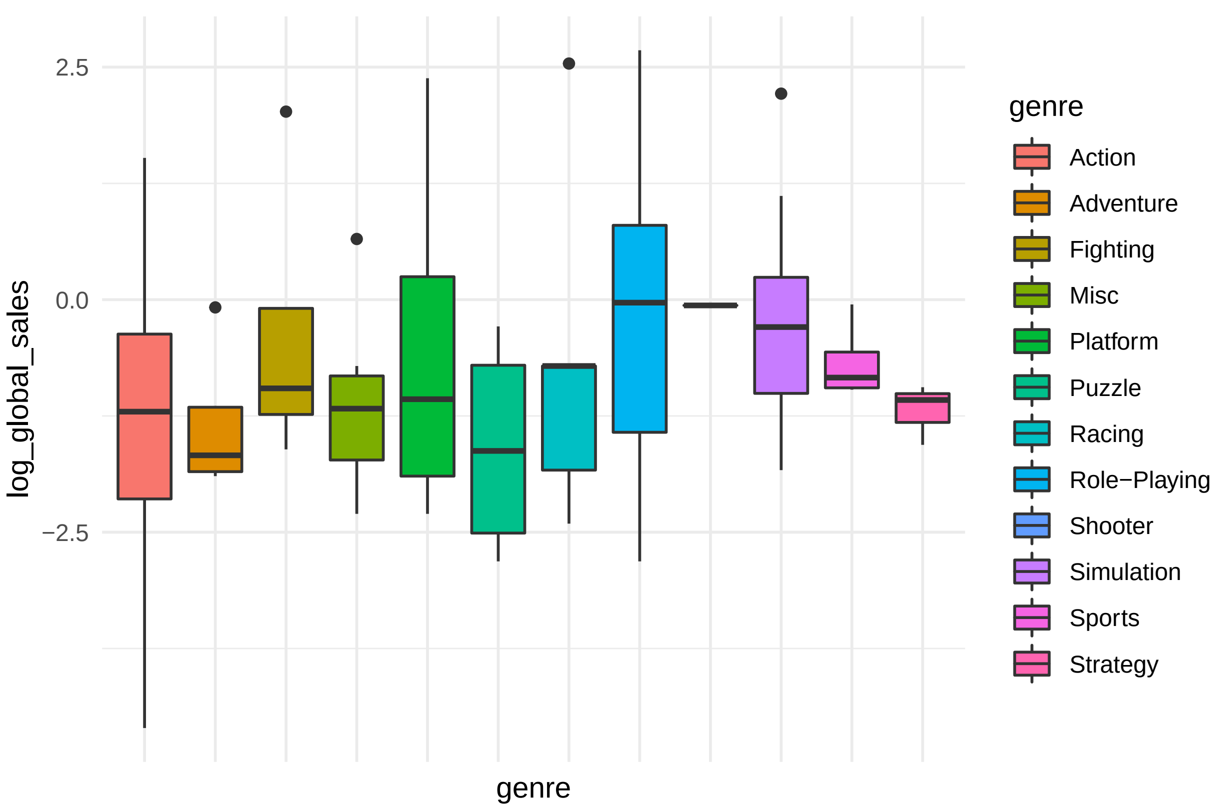


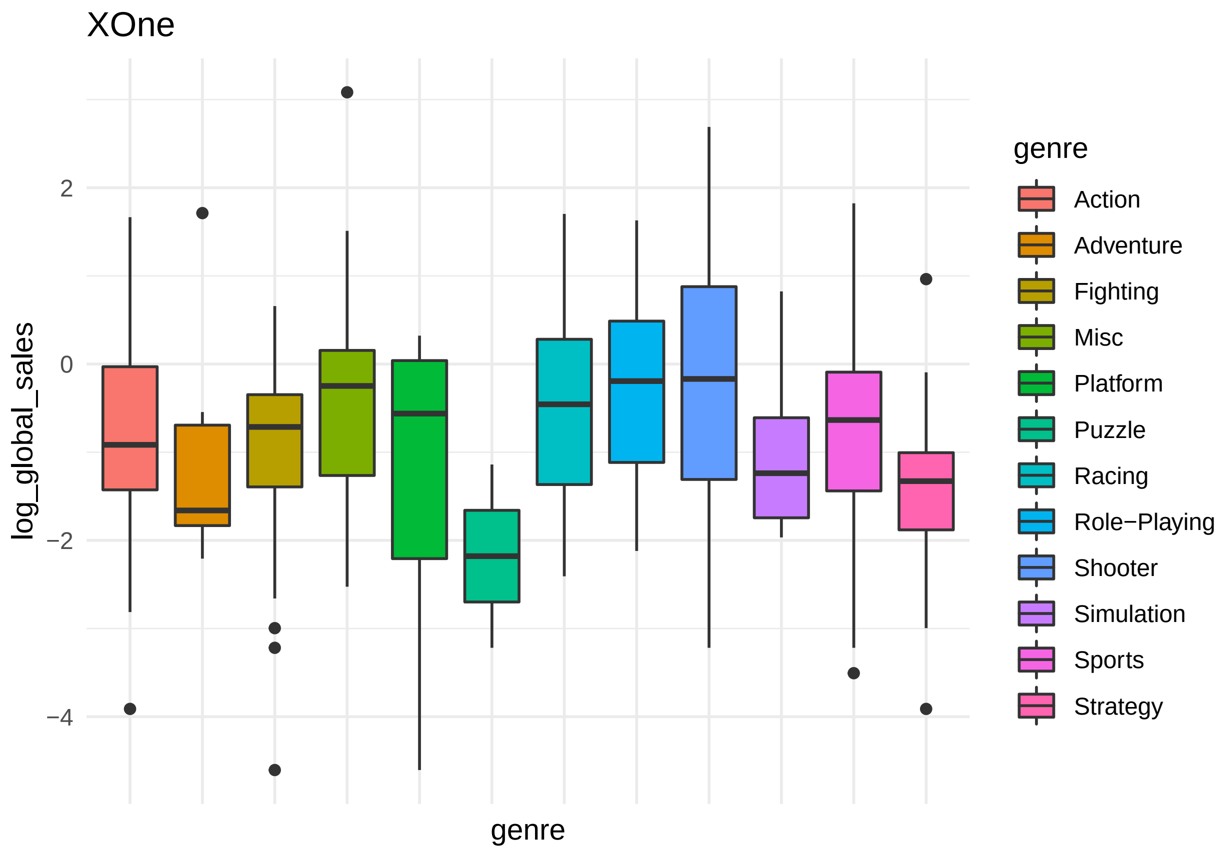

600
400
200
0
count
−5.0
−2.5
0.0
2.5
log_global_sales

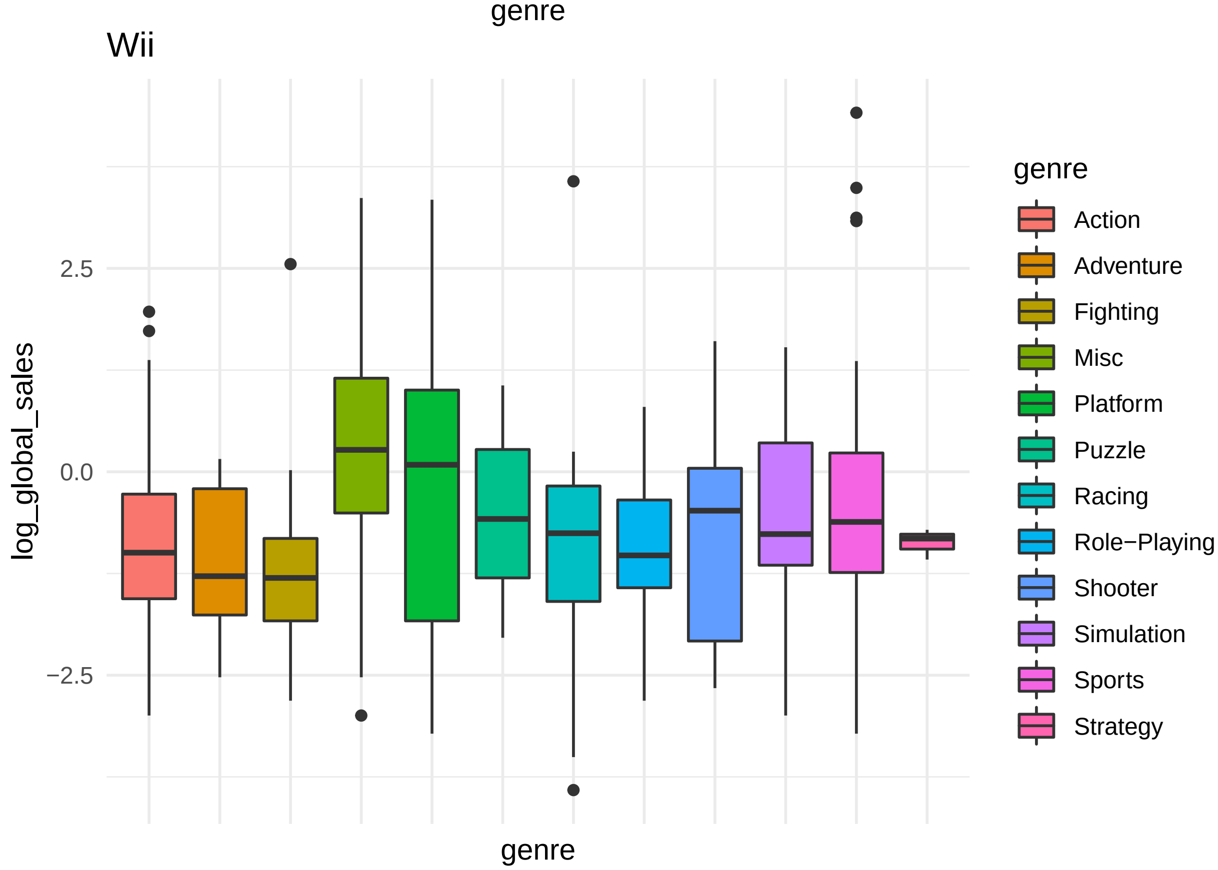

2.5
0.0
−2.5
−5.0
log_global_sales
Microsoft
Nintendo
PC
Sega
Sony
platform_company
rating_everyone
0
1

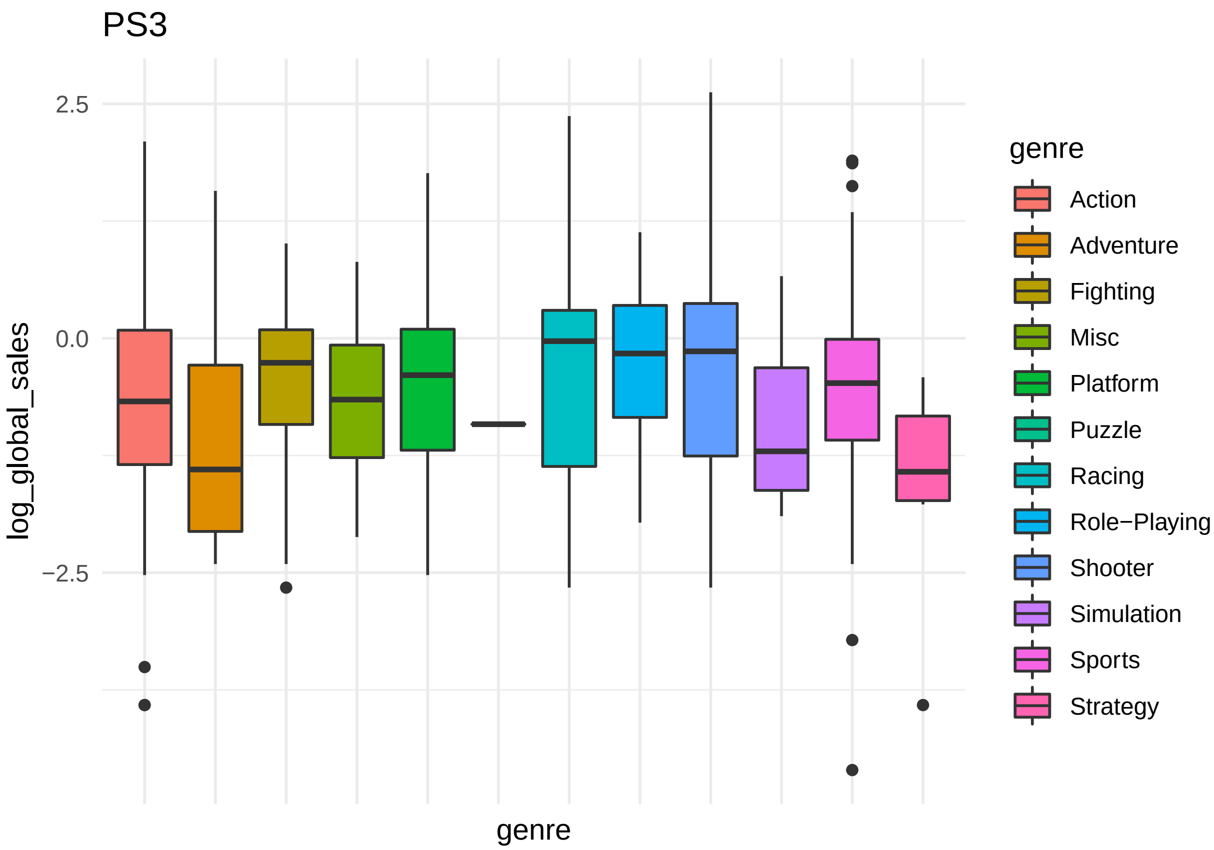

log_global_sales
2.5
0.0
−2.5
−5.0
Action
Adventure
Fighting
Misc
Platform
Racing
Role−Playing
Simulation
Sports
Strategy
genre
rating_everyone
0
1

3DS

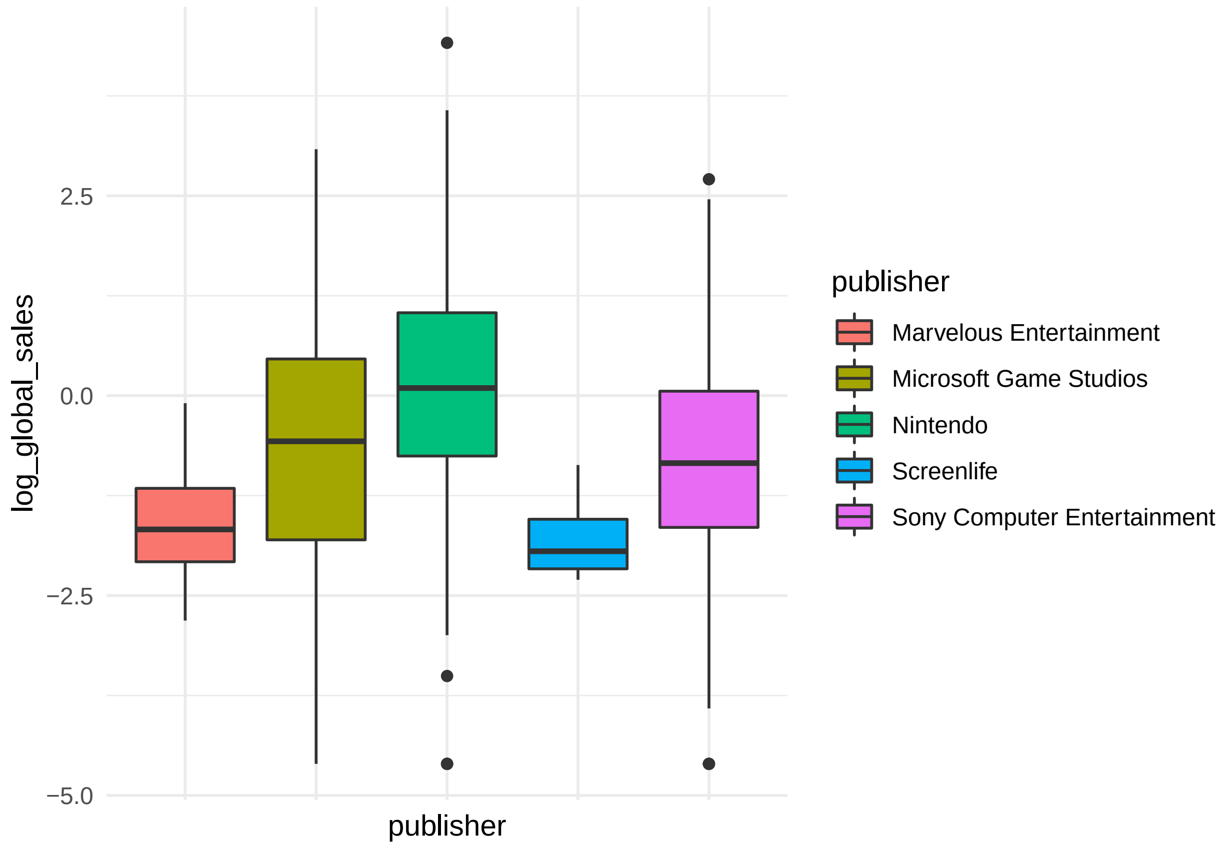

log_global_sales
2.5
0.0
−2.5
genre
genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role−Playing
Shooter
Simulation
Sports
Strategy

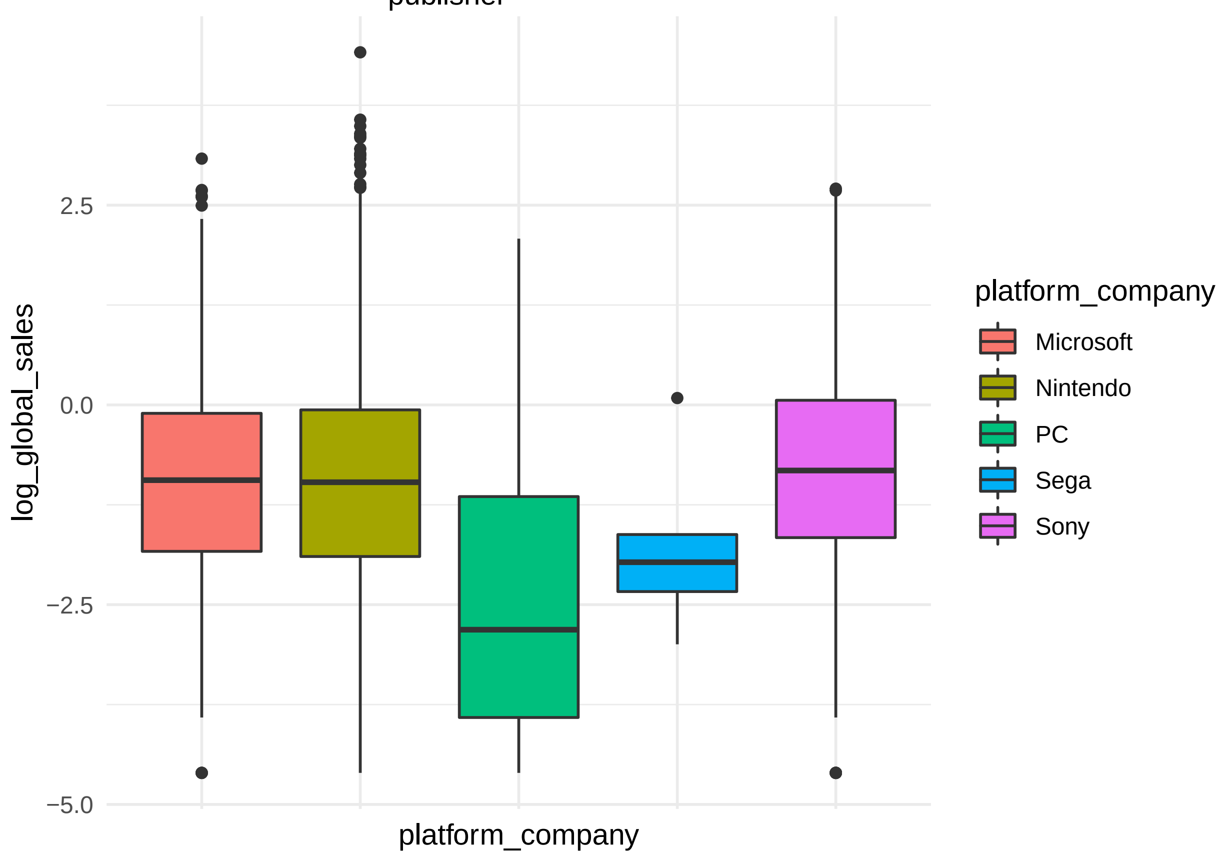

XOne
log_global_sales
2
0
−2
−4
genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role−Playing
Shooter
Simulation
Sports
Strategy

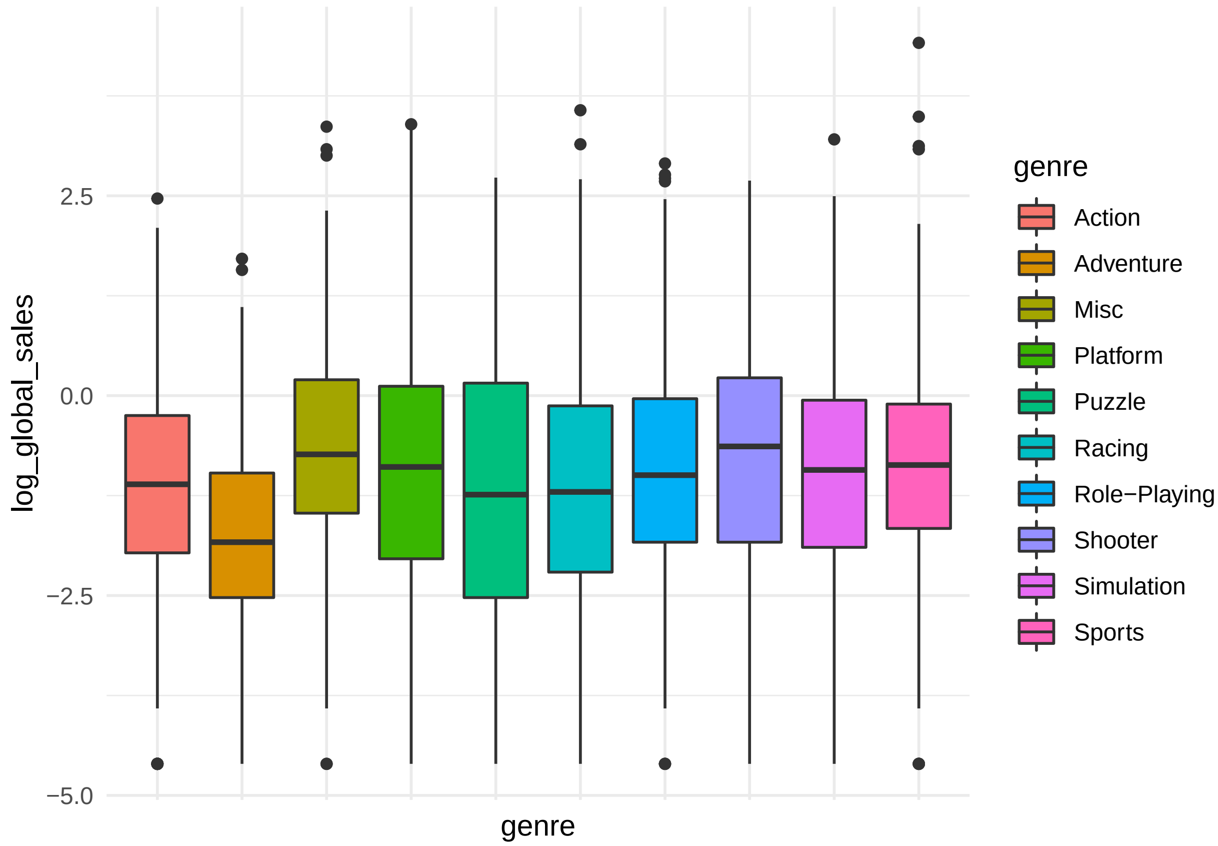

Wii
log_global_sales
2.5
0.0
−2.5
genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role−Playing
Shooter
Simulation
Sports
Strategy

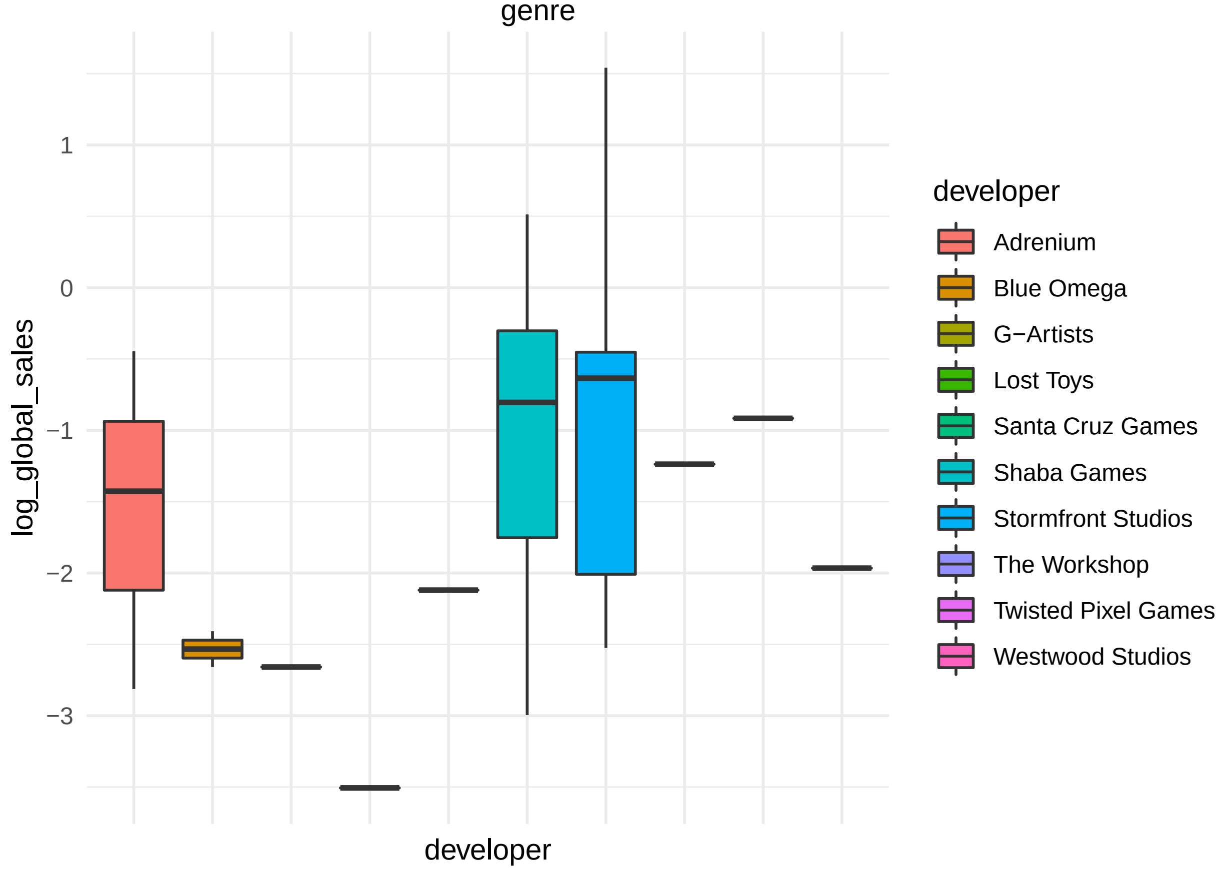


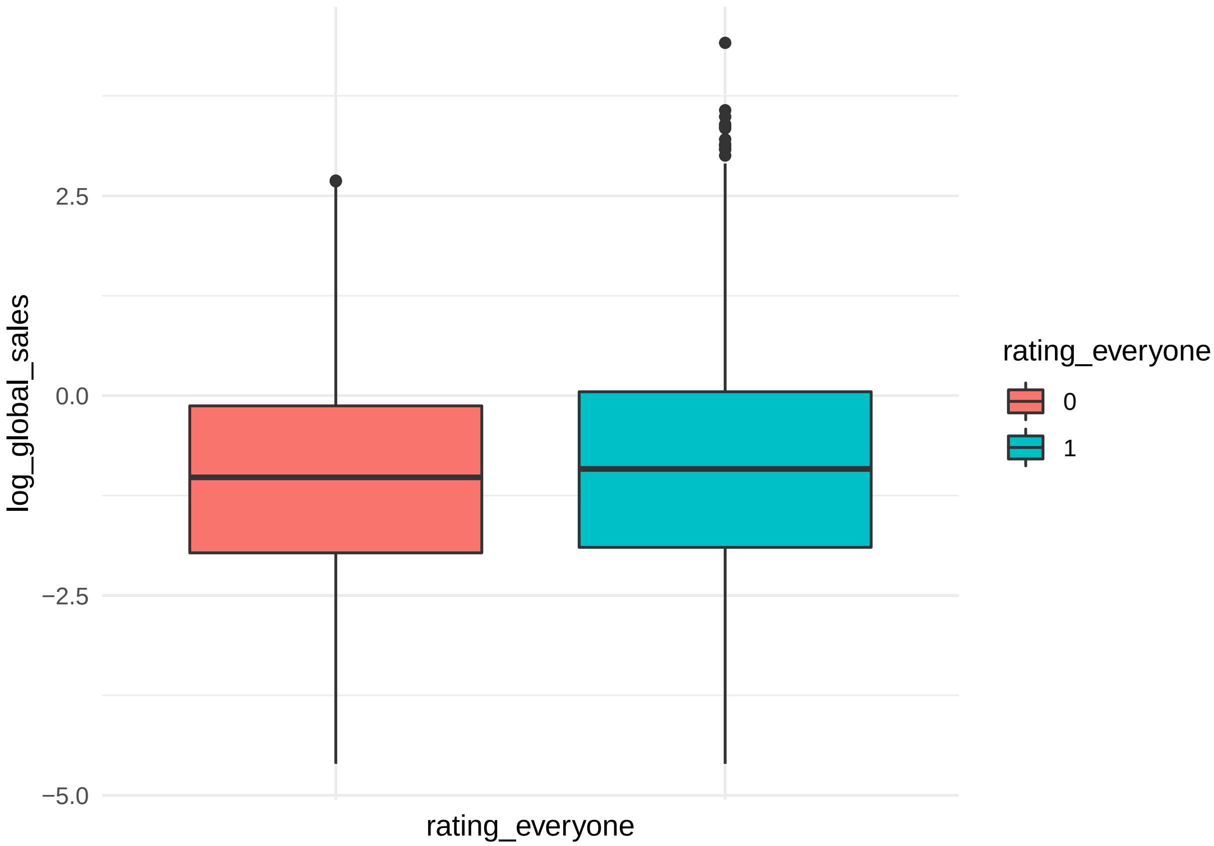

```
## [1] Nintendo                     Screenlife
## [3] Sony Computer Entertainment Microsoft Game Studios
## [5] Marvelous Entertainment
## 444 Levels: 10TACLE Studios 1C Company 2D Boy 2K Sports 3DO ... Zushi Games
```


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

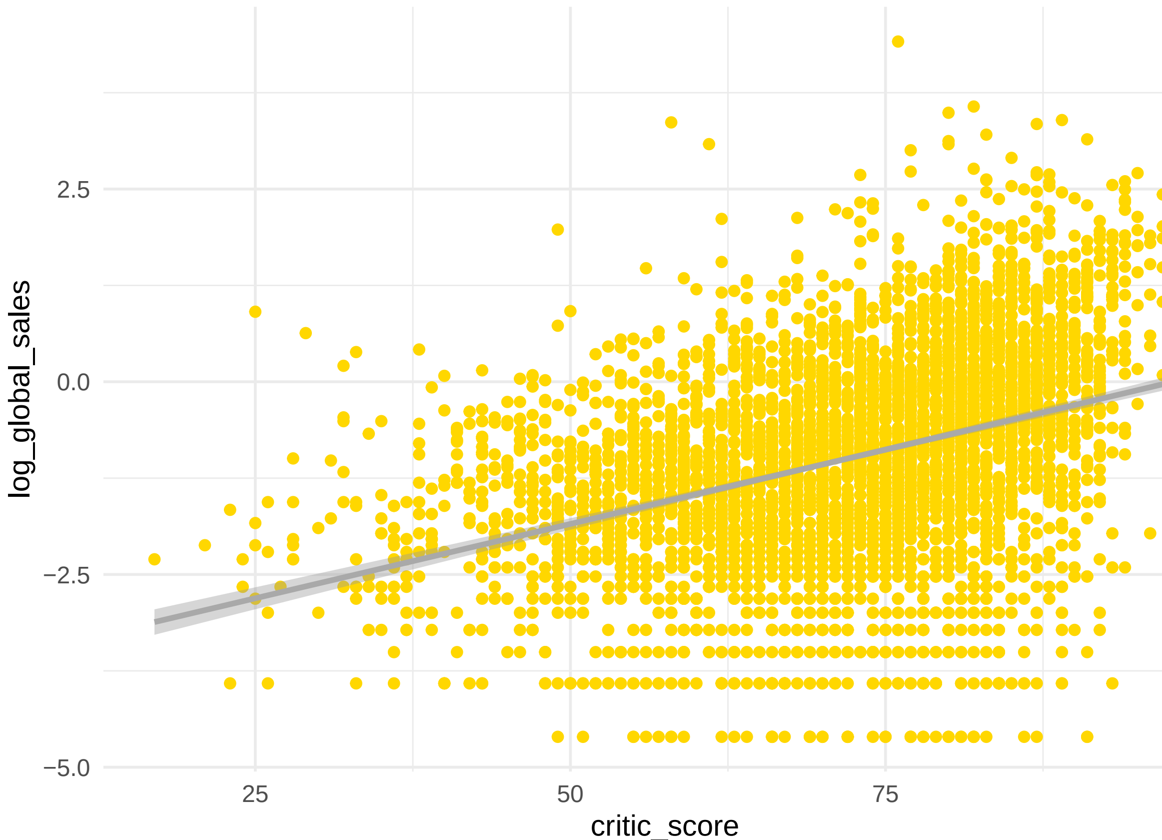

log_global_sales
2.5
0.0
−2.5
−5.0
platform_company
platform_company
Microsoft
Nintendo
PC
Sega
Sony


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

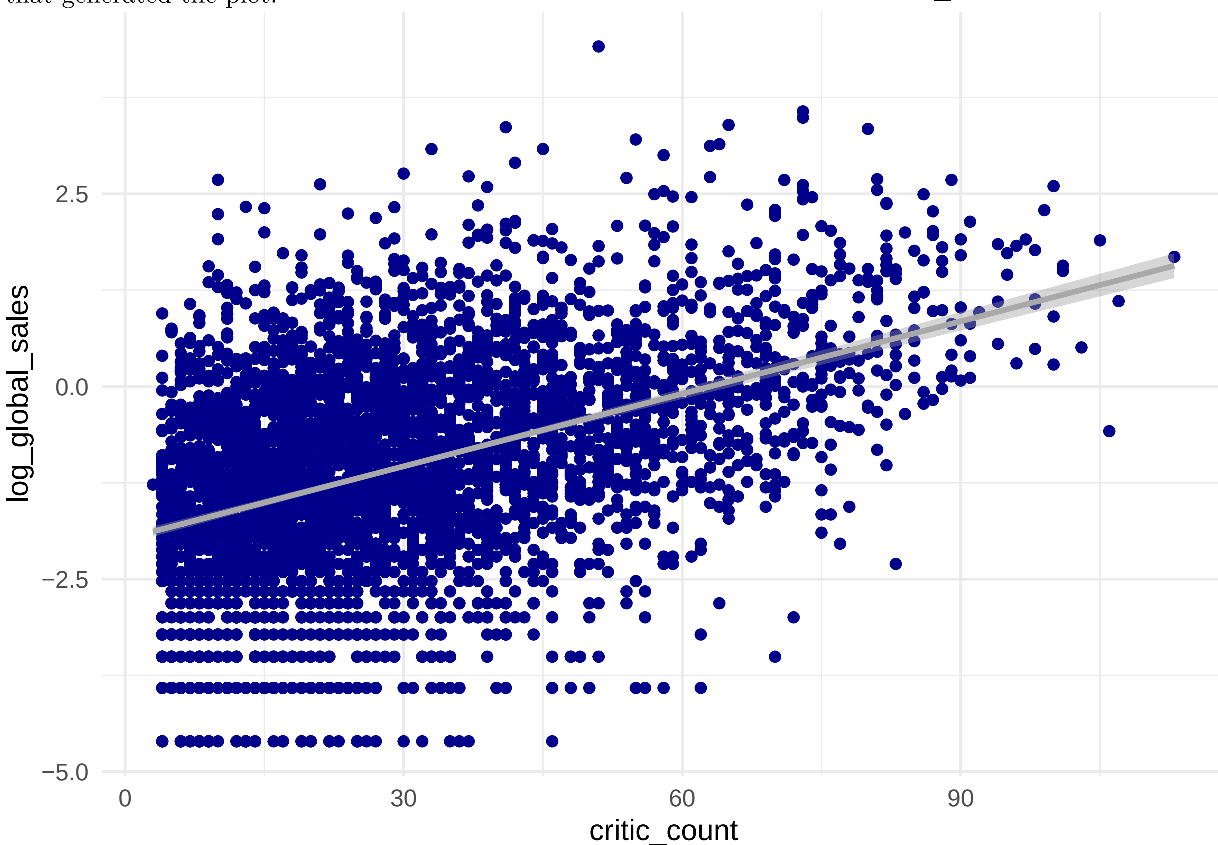

log_global_sales
1
0
−1
−2
−3
developer
developer
Adrenium
Blue Omega
G−Artists
Lost Toys
Santa Cruz Games
Shaba Games
Stormfront Studios
The Workshop
Twisted Pixel Games
Westwood Studios

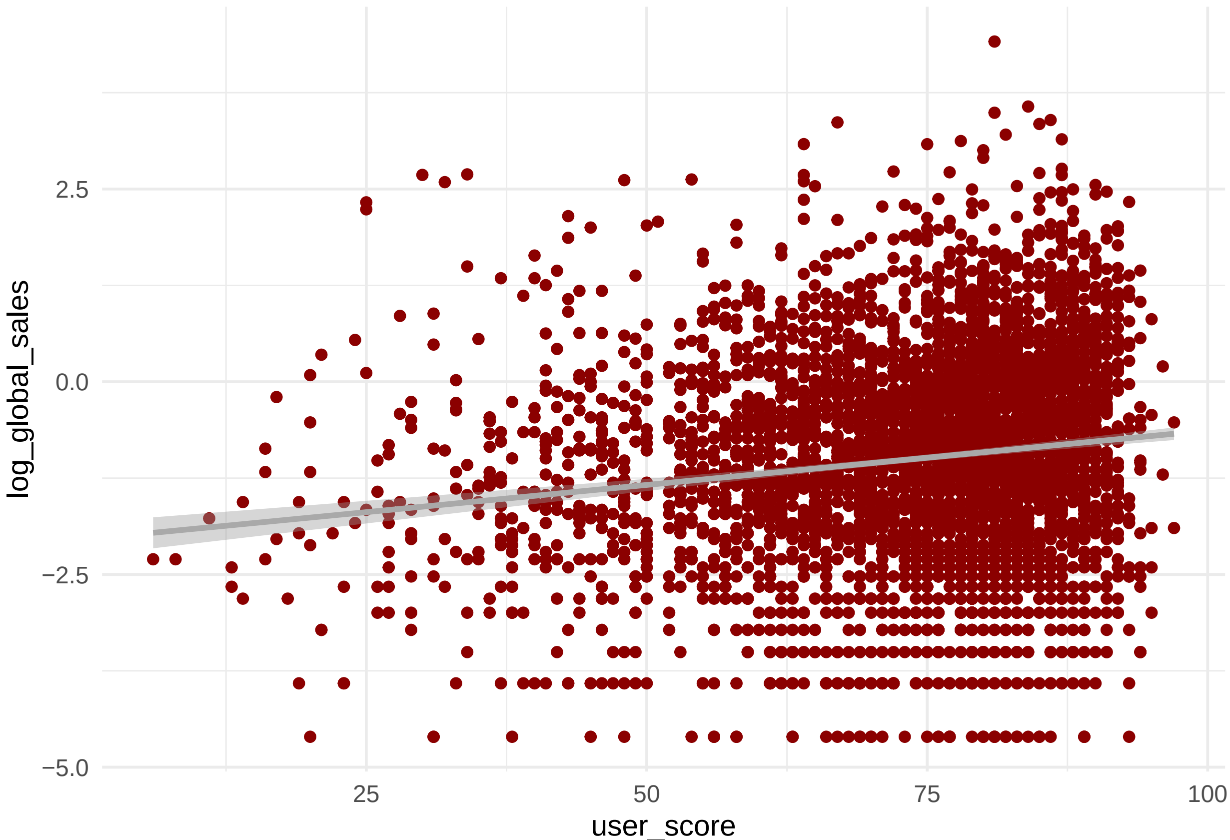


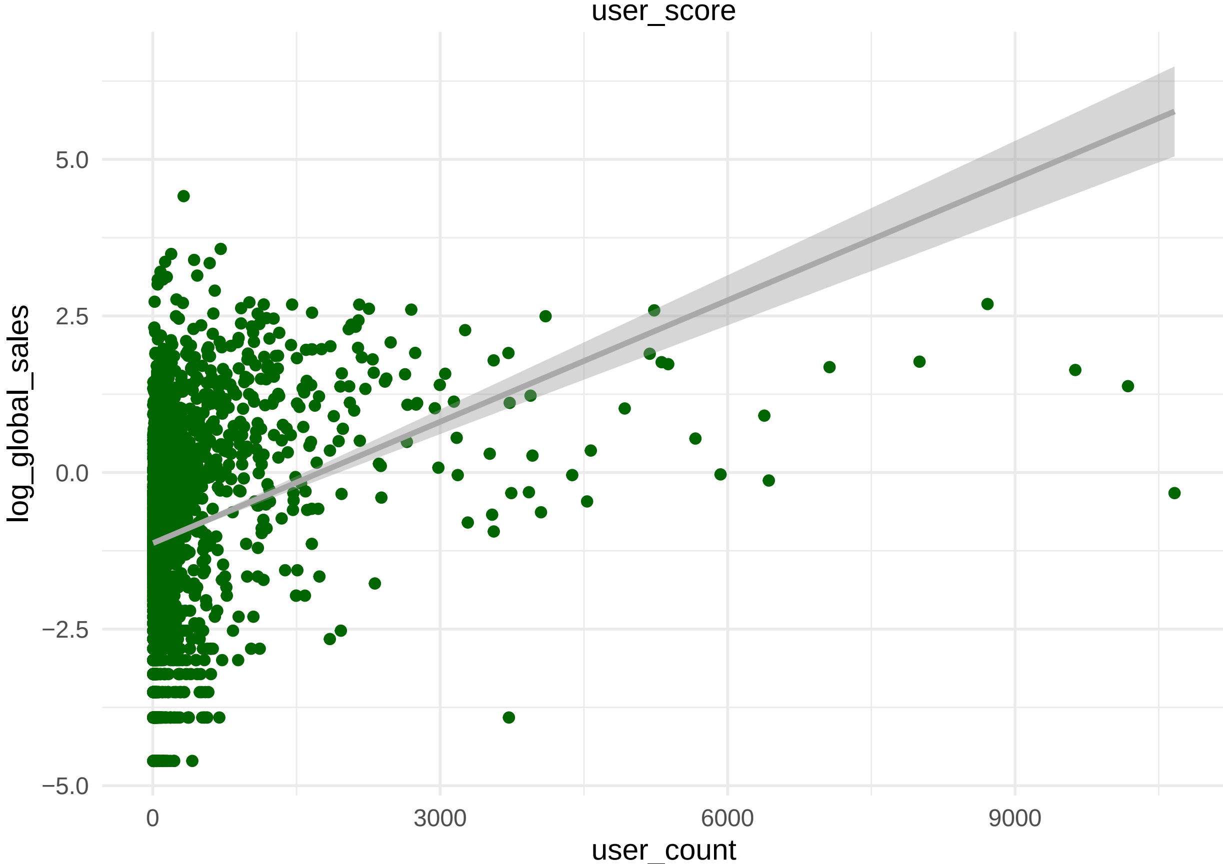
Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code



that generated the plot.




log_global_sales
2.5
0.0
−2.5
−5.0
25
50
75
100
user_score


log_global_sales
5.0
2.5
0.0
−2.5
−5.0
0
3000
6000
9000
user_count

Model

In order to obtain a final model for the response variable (the probability that someone from a certain demographic group will turn out to vote) various methods for model selection were tested and interactions between predictors were considered as part of the full model. Since the counties in our data are just a sample of the total county population, we decided to include random intercept effects for *county* in the final model. The rest of the categorical variables are included in the model as predictors. We also included the *sex:party* interaction term to see if it is significant.

We used a manual stepwise approach in R to compare multilevel logistic models with binomial outcomes using AIC as the selection criteria. We tried to include as many variables as possible and came up with a model using *age*, *party*, *sex*, *race*, *ethnic*, *sex:party*, and random intercepts by *county*. The final model has an AIC value of 137406.4, and the following formula:

$$\Pr\left[\frac{total\_voters_{i,C}}{reg\_voters_{i,C}}=1\right]=\pi_{i,C} \quad \text{and} \quad \Pr\left[\frac{total\_voters_{i,C}}{reg\_voters_{i,C}}=0\right]=1-\pi_{i,C};$$

$$\log\left(\frac{\pi_{i,C}}{1-\pi_{i,C}}\right)=\beta_{\text{(Intercept)}}+\gamma_{\text{(Intercept)},C}+\sum_{A\in\{'\text{Age 26}-40','\text{Age 41}-65','\text{Age Over 66}'\}}\beta_A\left(age_A\right)_{i,C}$$

$$+\sum_{P\in\{\text{LIB,REP,UNA}\}}\beta_P\left(party_P\right)_{i,C}+\sum_{S\in\{\text{M,U}\}}\beta_S\left(sex_S\right)_{i,C}+\sum_{R\in\{\text{B,I,M,O,U,W}\}}\beta_R\left(race_R\right)_{i,C}$$

```
## 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
```

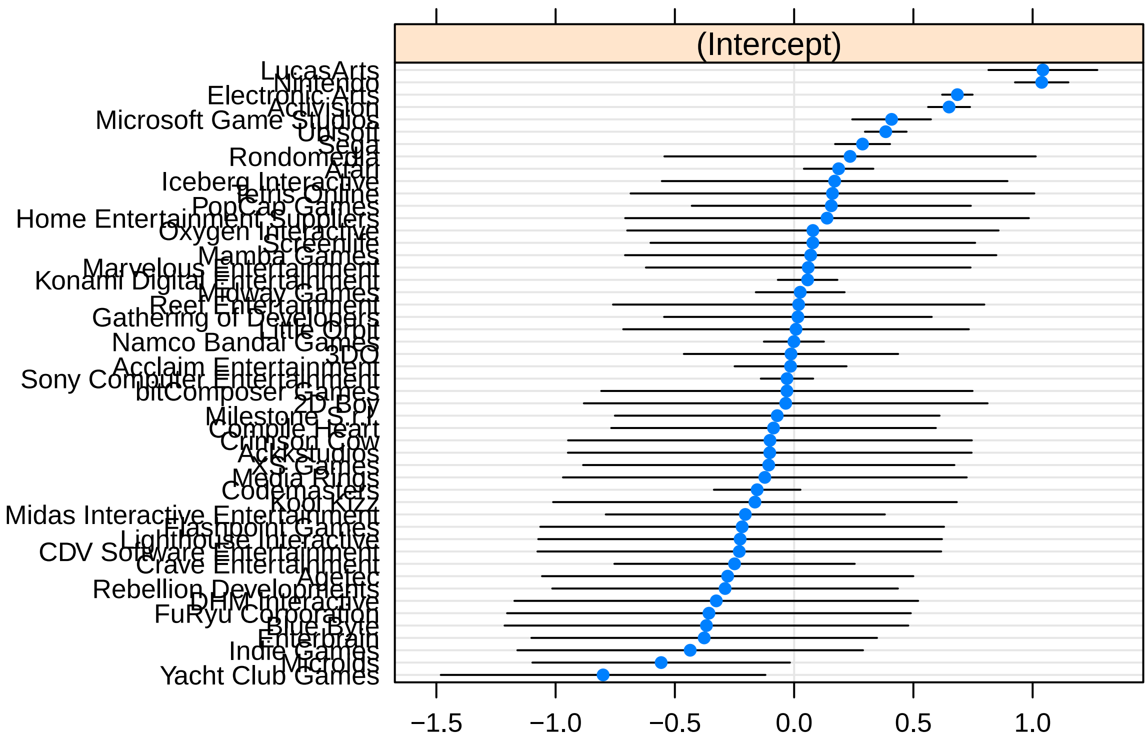
```
## genreAdventure                             -3.381e-01  1.129e-01  4.074e+03
## genreFighting                               3.061e-01  8.488e-02  4.150e+03
## genreMisc                                   5.342e-01  8.907e-02  4.149e+03
## genrePlatform                               5.899e-02  1.022e-01  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
```



Conclusions

Voting tendency of different demographic groups could be analyzed from the model we built. Baselines of our model are Age 18 - 25, Democratic Party, Female, Asian, and Hispanic/Latino and we could see the following interesting facts.

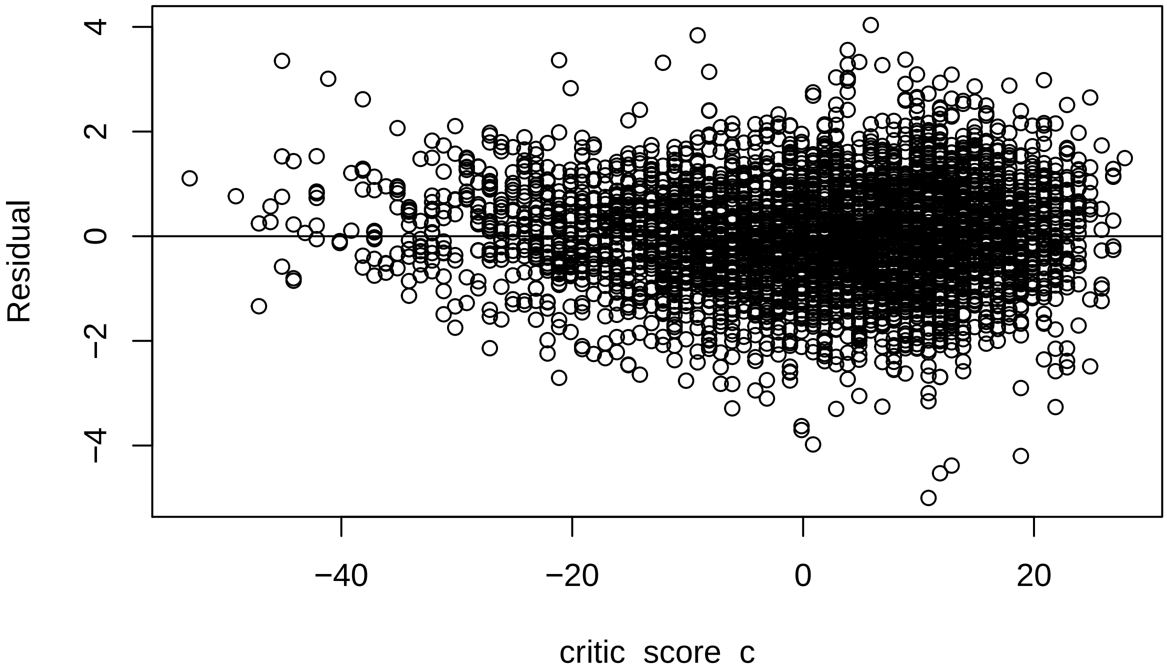
- Most of variables are significant.
- All county random intercept effects are significant (except for Jackson), which means the odds of voting differed by county in 2016.
- Except mixed race, most of races are less likely vote as compared to the baseline.
- People aged above 40 are approximately 150% more likely to vote than younger people.
- Libertarians are 75% more likely to vote than other parties although they only make a small fraction of the population
- Males belonging to the Republican Party, Libertarian Party or even if they are Unaffiliated are more likely to vote as compared to the baseline. But, overall males are 23% less likely to vote as compared to overall females.
- The Male supporting Democrats are the group that bring the odds down.

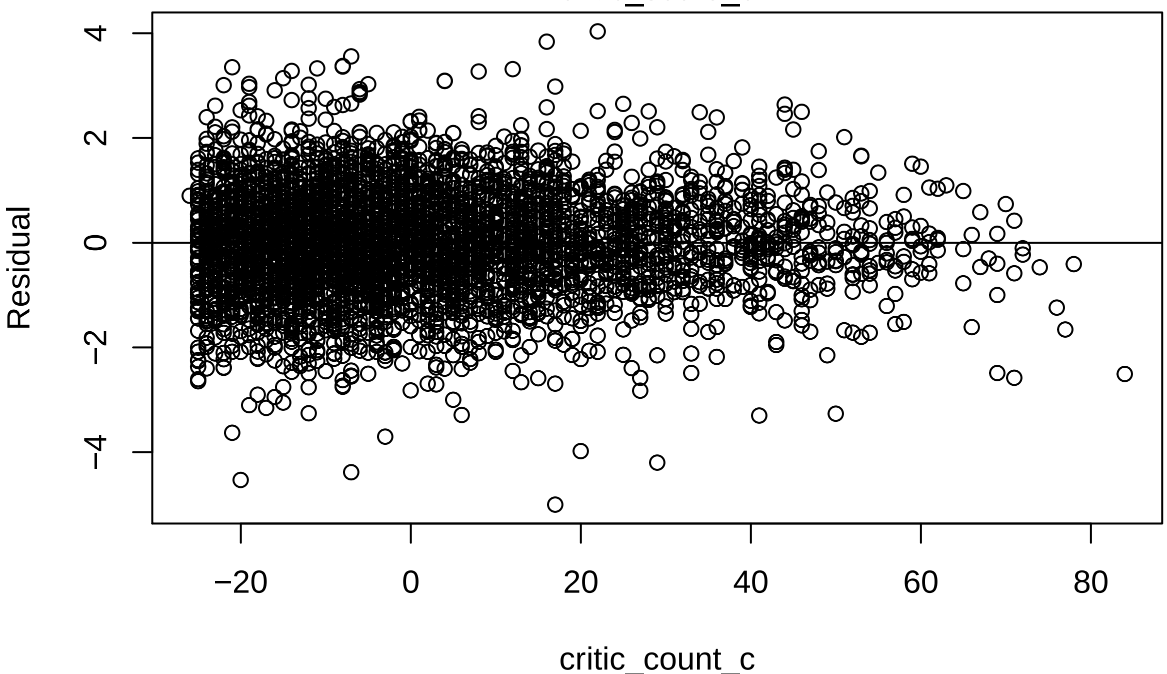
Through the multi-level logistic model from dataset from the North Carolina State Board of Elections (NCSBE), we could reveal the voting tendency of different demographic groups in North Carolina in 2016 and predict the turnout rate of each subgroup. From the analysis, we could find that age, party, race, sex, ethnic, county, etc., are significant factors to decide their tendency of voting.

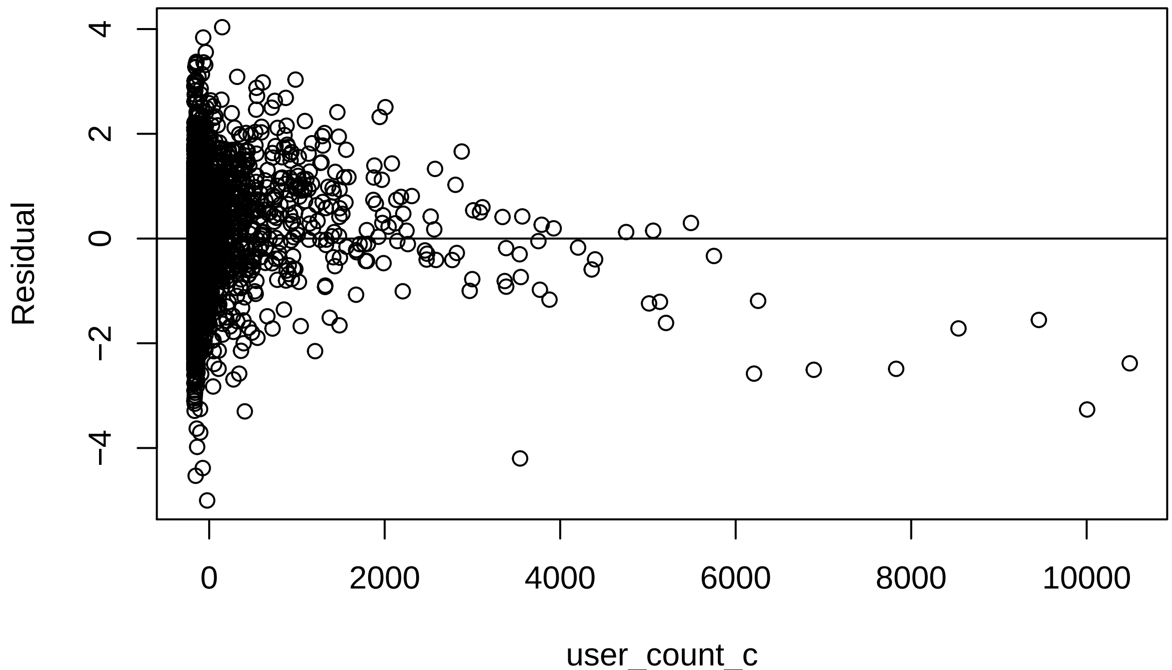
Although, we could predict turnout rate of voting and voter's tendency, categorical age and unidentified gender, race, ethnic would be clarified to get more precise results.

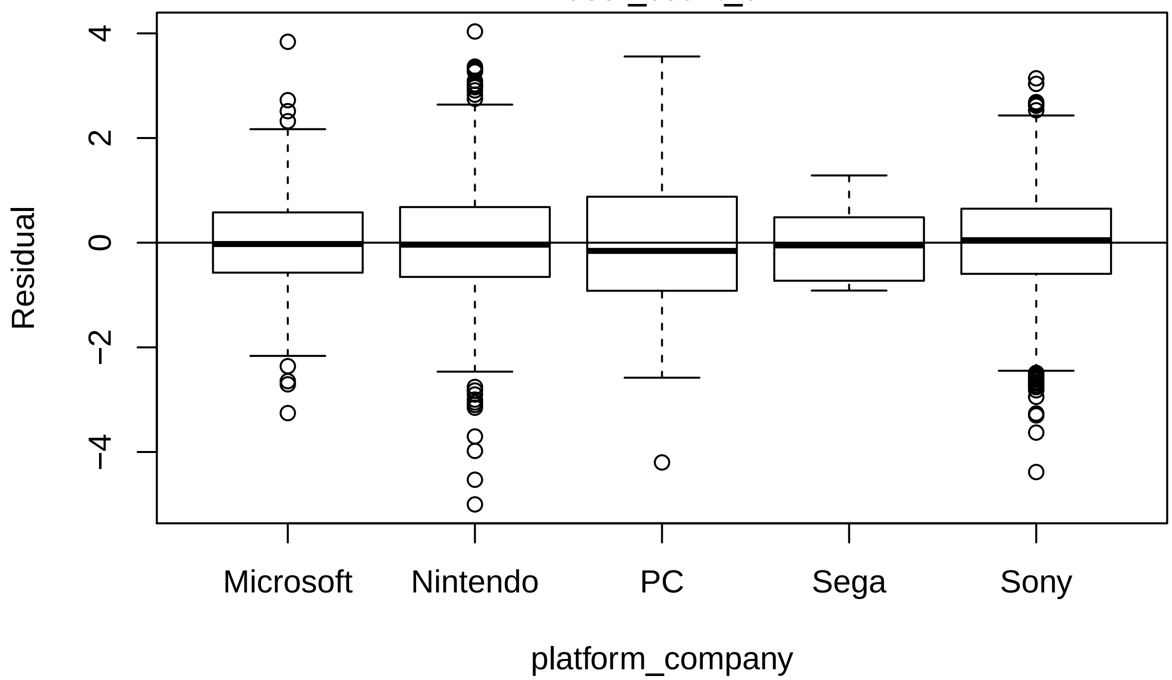
Appendix

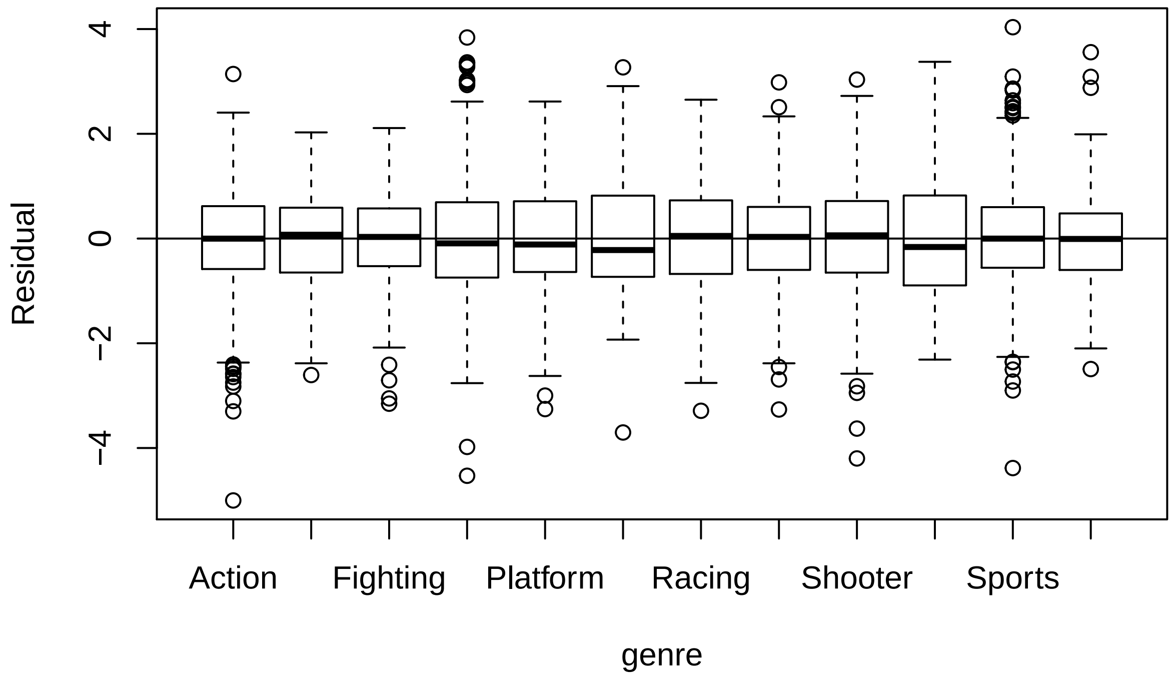
1.1 EDA

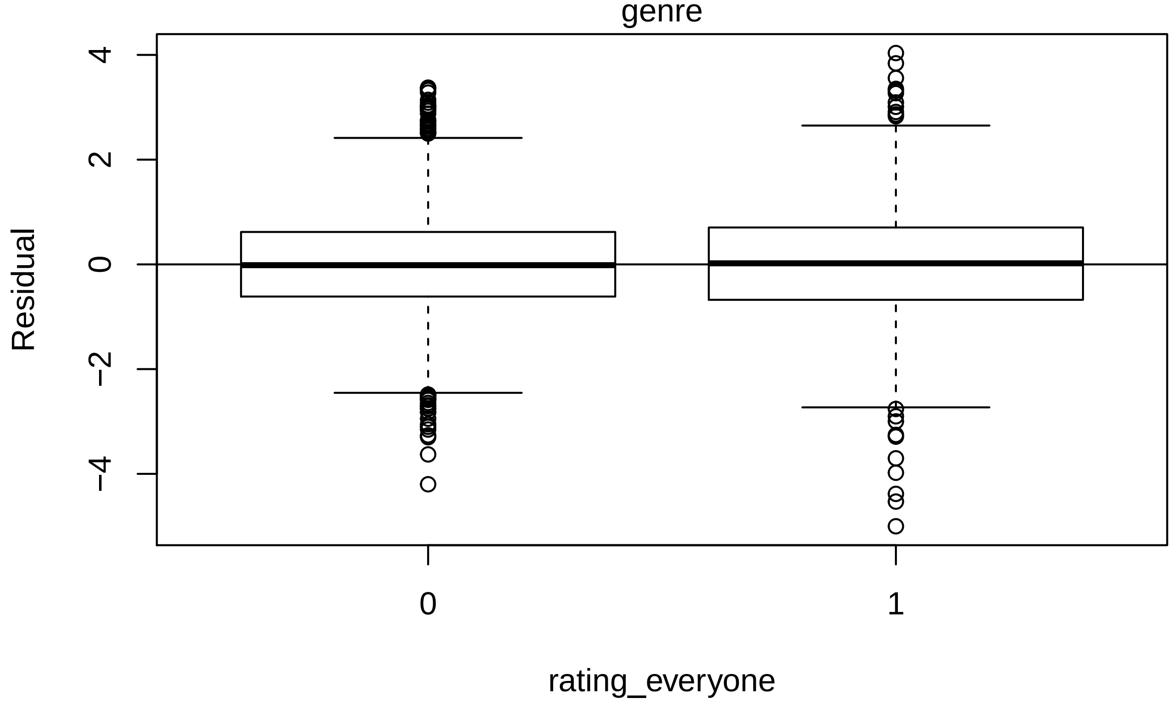


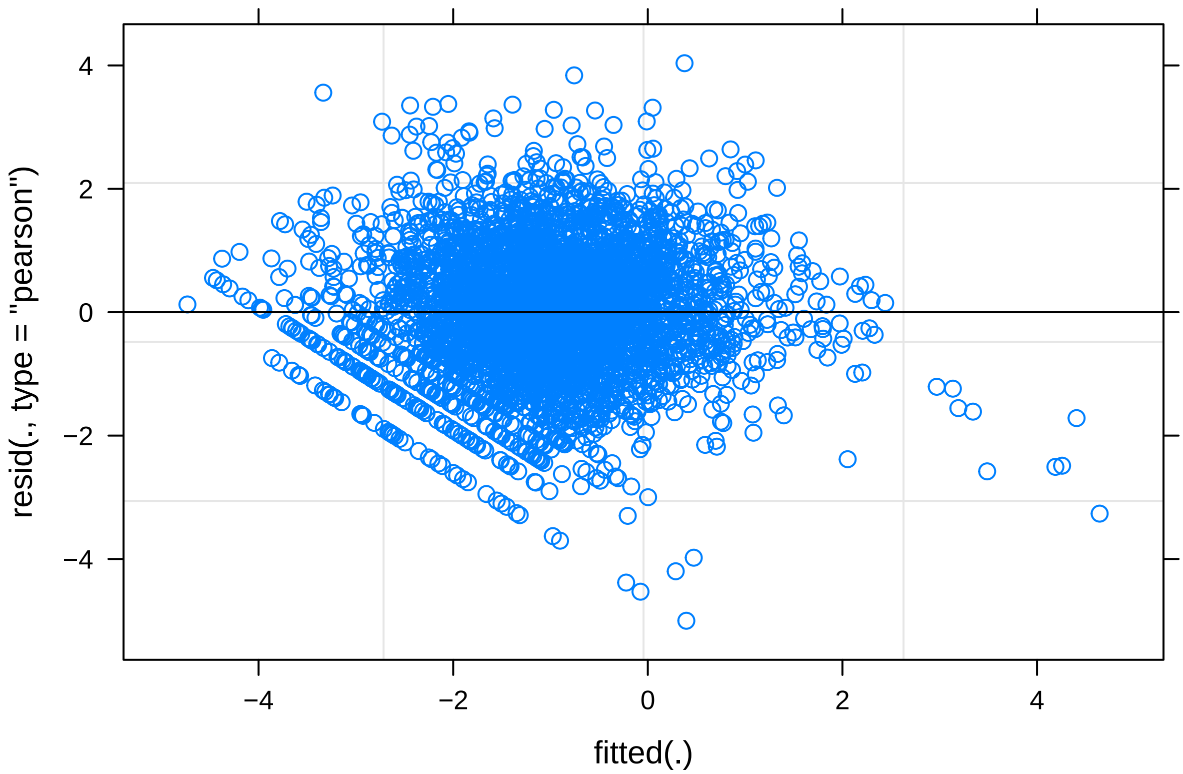



Residual
4
2
0
−2
−4
0
2000
4000
6000
8000
10000
user_count_c


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



Normal Q−Q Plot