

Final Project

Sebastián Soriano Pérez [ss1072]

12/10/2019

Analyzing Video Game Sales

Summary

A few sentences describing the inferential question(s), the method used and the most important results.

Introduction

A more in-depth introduction to the inferential question(s) of interest.

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. I built a hierarchical linear model on a logarithmic transformation of the global sales, with random slopes by developer, 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.

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

You should describe the data in this section: how you obtained the data, the variables included, dealing with missing/erroneous values, exploratory data analysis etc.

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): PPlatform 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

Since the categorical variables have many levels (See Appendix 1.1 for a summary of the dataset *vgsales*) and in order to center the numerical variables to avoid multicollinearity issues, I created the following additional columns:

- *critic_score_c* (numerical): Values of *critic_score* minus the column's mean value
- *critic_count_c* (numerical): Values of *critic_count* minus the column's mean value
- *user_score_c* (numerical): Values of *user_score* minus the column's mean value
- *user_count_c* (numerical): Values of *user_count* minus the column's mean value
- *platform_company* (categorical): Company that manufactures the video game's platform
 - 'Nintendo' when *platform* is '3DS', 'DS', 'GB', 'GBA', 'GC', 'N64', 'Wii', or 'WiiU'
 - 'Sega' when *platform* is 'DC'
 - 'PC' when *platform* is 'PC'
 - 'Microsoft' when *platform* is 'X360', 'XB', or 'XOne'
 - 'Sony' when *platform* is 'PS', 'PS2', 'PS3', 'PS4', 'PSP', or 'PSV'
- *rating_everyone* (binary): Indicates if there is not an age restriction for the video game release
 - '1' when *rating* is 'E'
 - '0' when *rating* is not 'E'

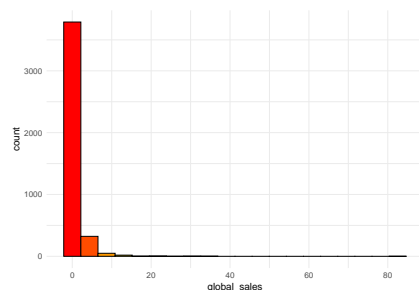
To consider the variable *publisher* into the model, I took a random sample of 50 different publishing companies from the dataset and built a sample dataset to work with. I observed at the distribution of the response variable *global_sales* and noticed that its distribution is not normal. After applying a logarithmic transformation, the distribution is normal, so I added the following column to the dataset:

- *log_global_sales*: Logarithmic transformation of *global_sales*

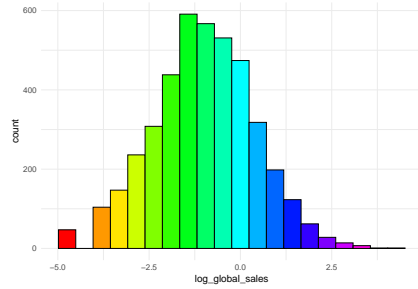
#Selecting 50 sample publishers

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

ggplot(data = sample_data, aes(x = global_sales)) +
  geom_histogram(color = 'black', fill = rainbow(20), bins = 20) +
  theme_minimal()
```

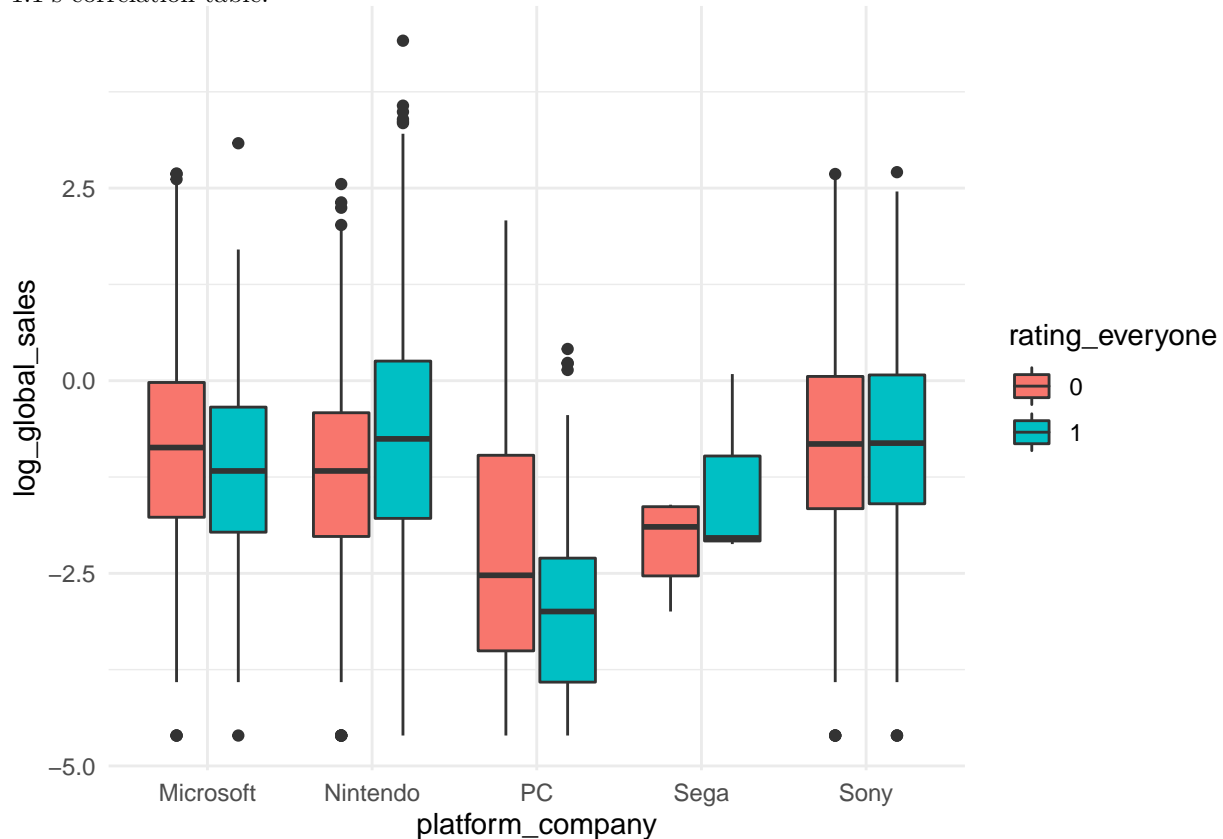


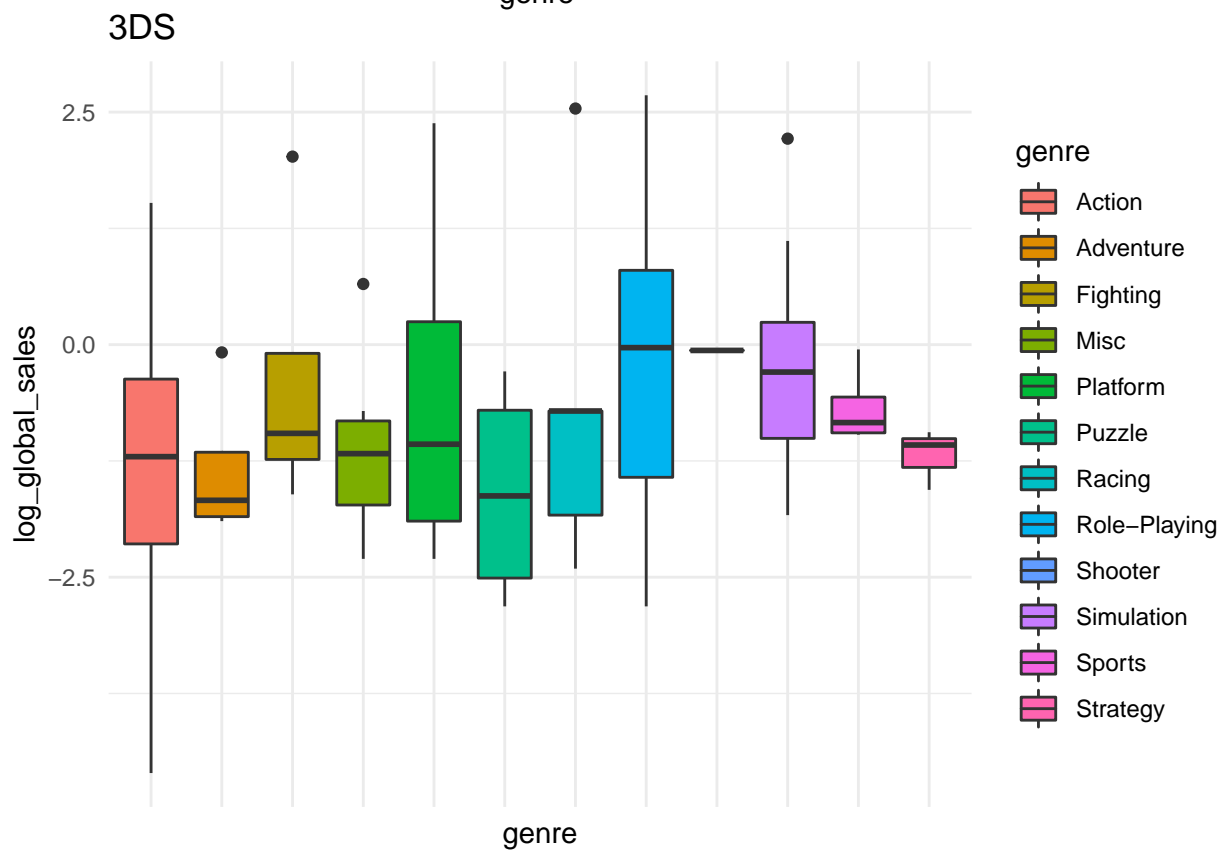
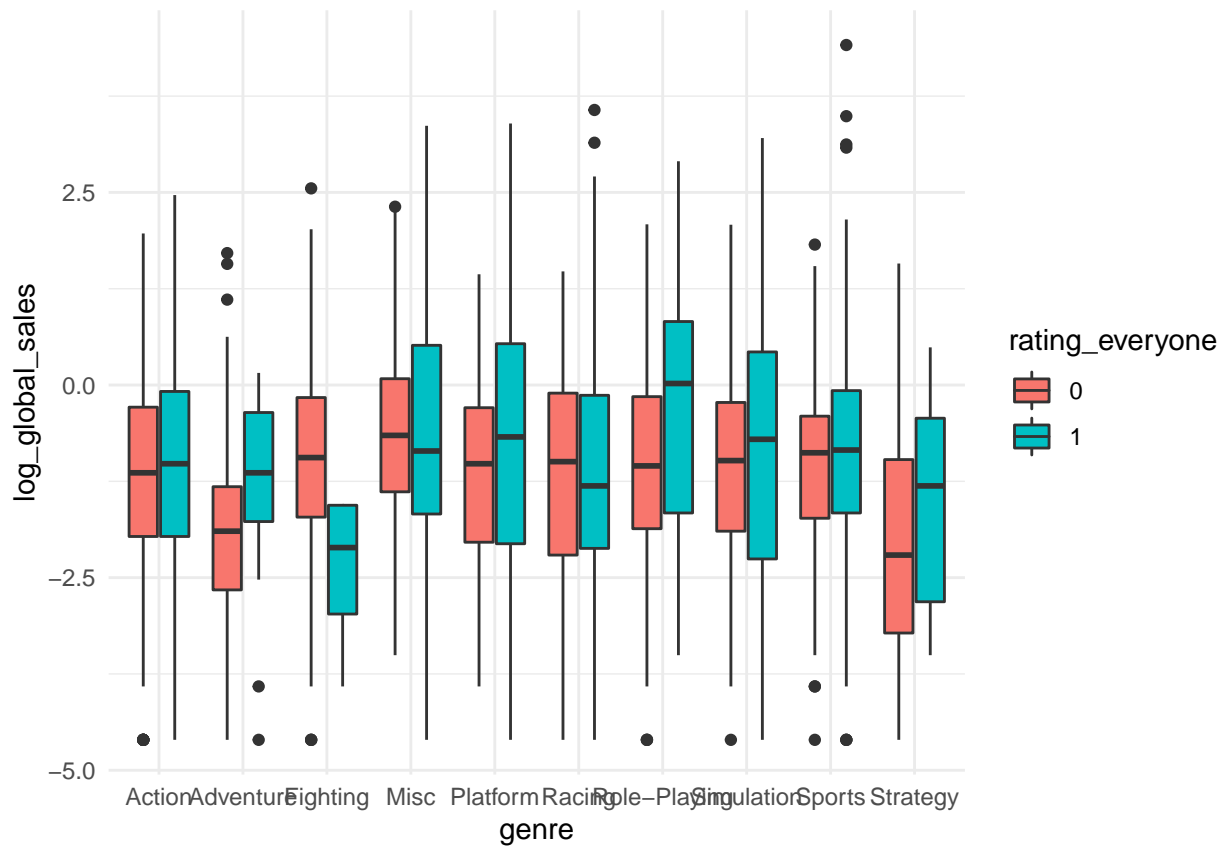
```
ggplot(data = sample_data, aes(x = log_global_sales)) +
  geom_histogram(color = 'black', fill = rainbow(20), bins = 20) +
  theme_minimal()
```

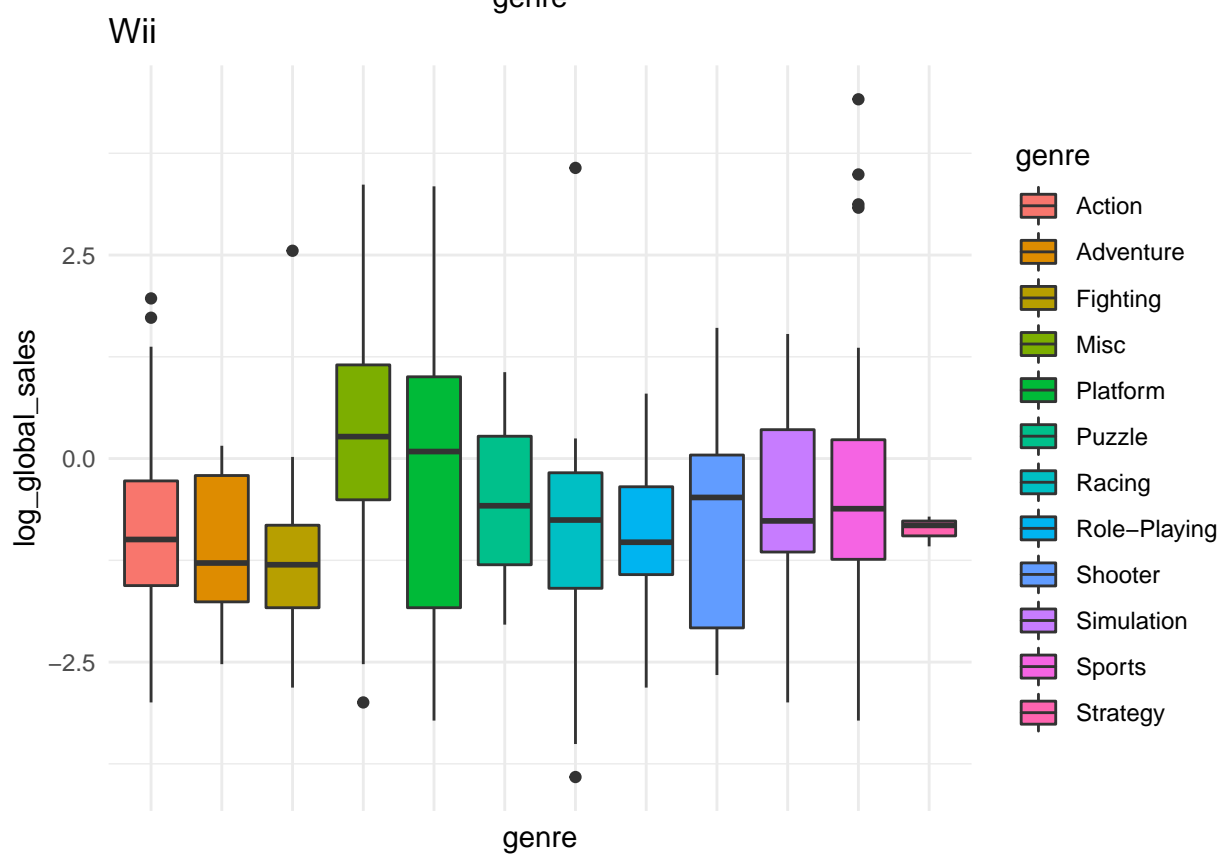
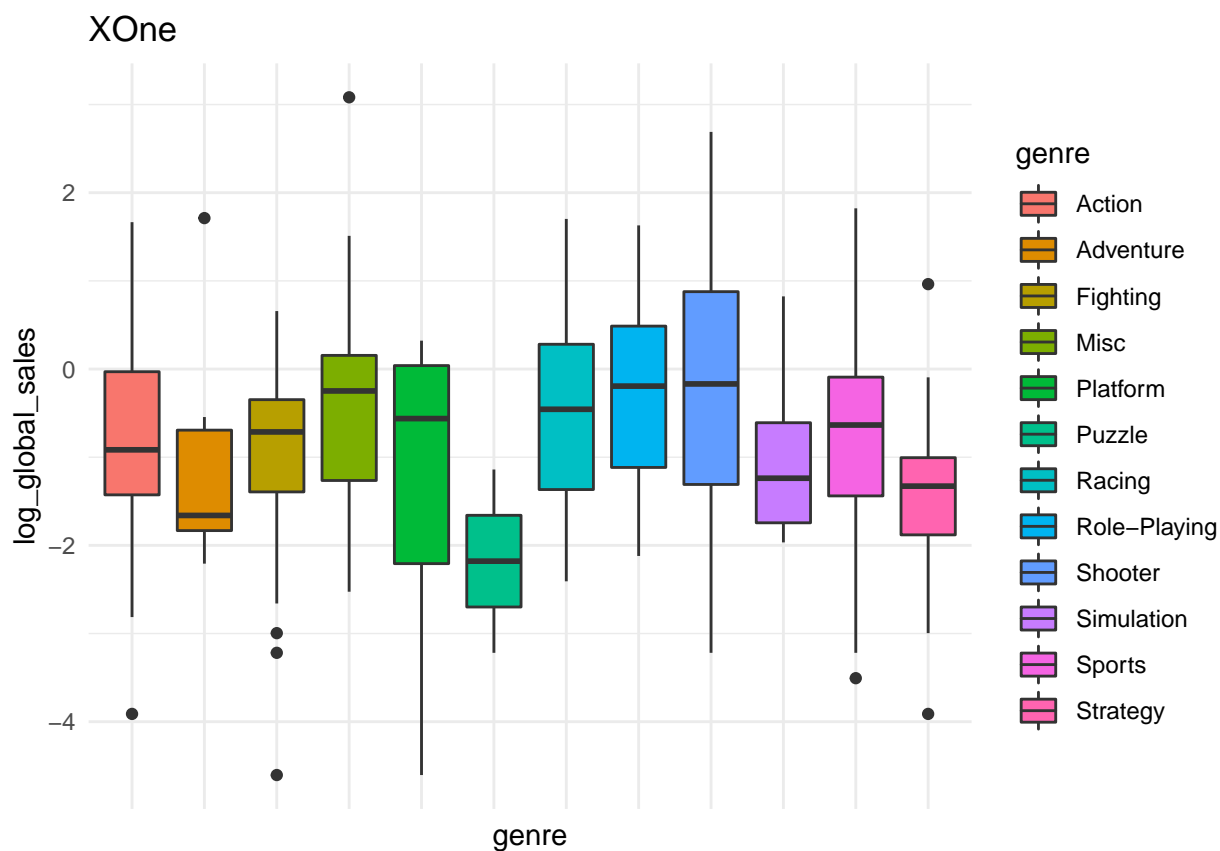


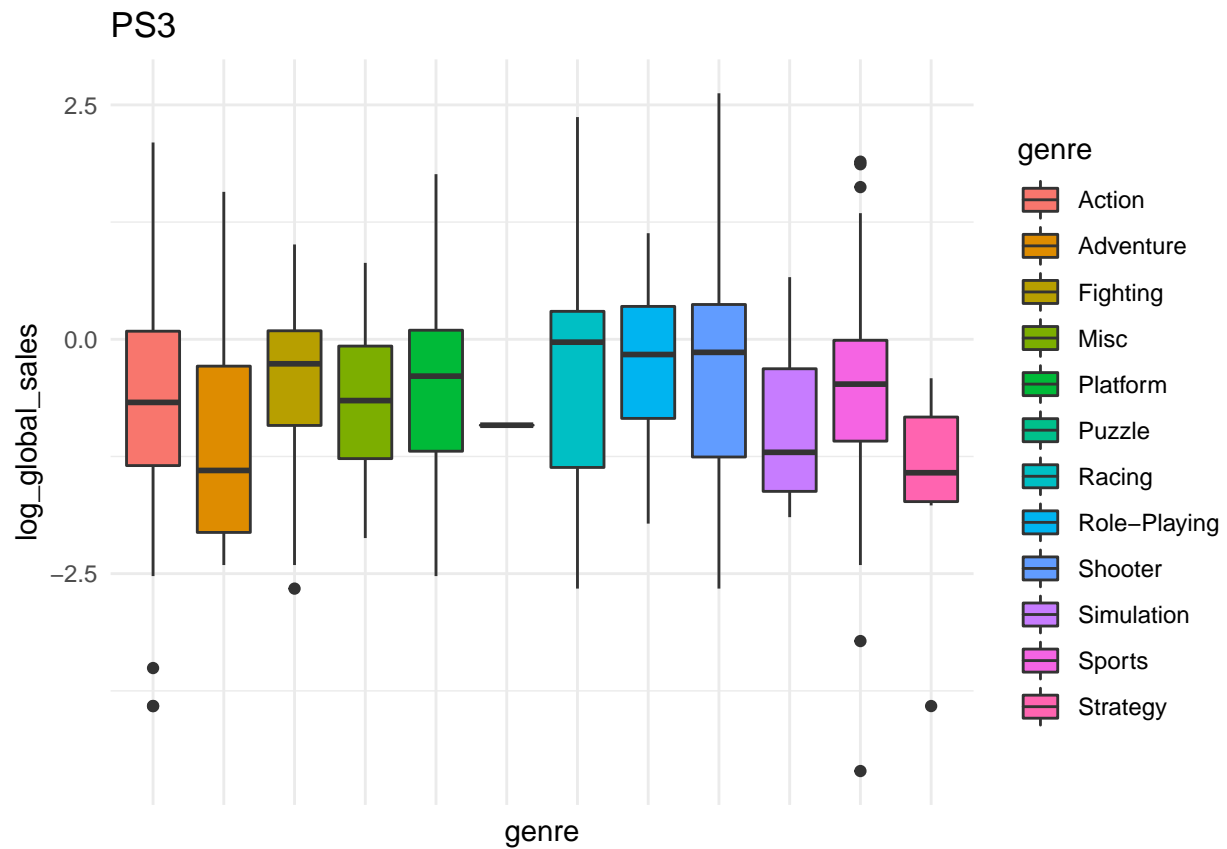
A summary of the data variables being analyzed can be found in Annex 1.1. An exploratory data analysis for all variables and plots for their interactions can be found in Annex 1.2.

The EDA suggests none of the numerical variables have a clear association with premature as the box-plots for premature = 0 and premature = 1 do not have noticeable differences. For the categorical variables, 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_c are being considered as those predictors have the largest correlations as seen in Annex 1.1's correlation table.

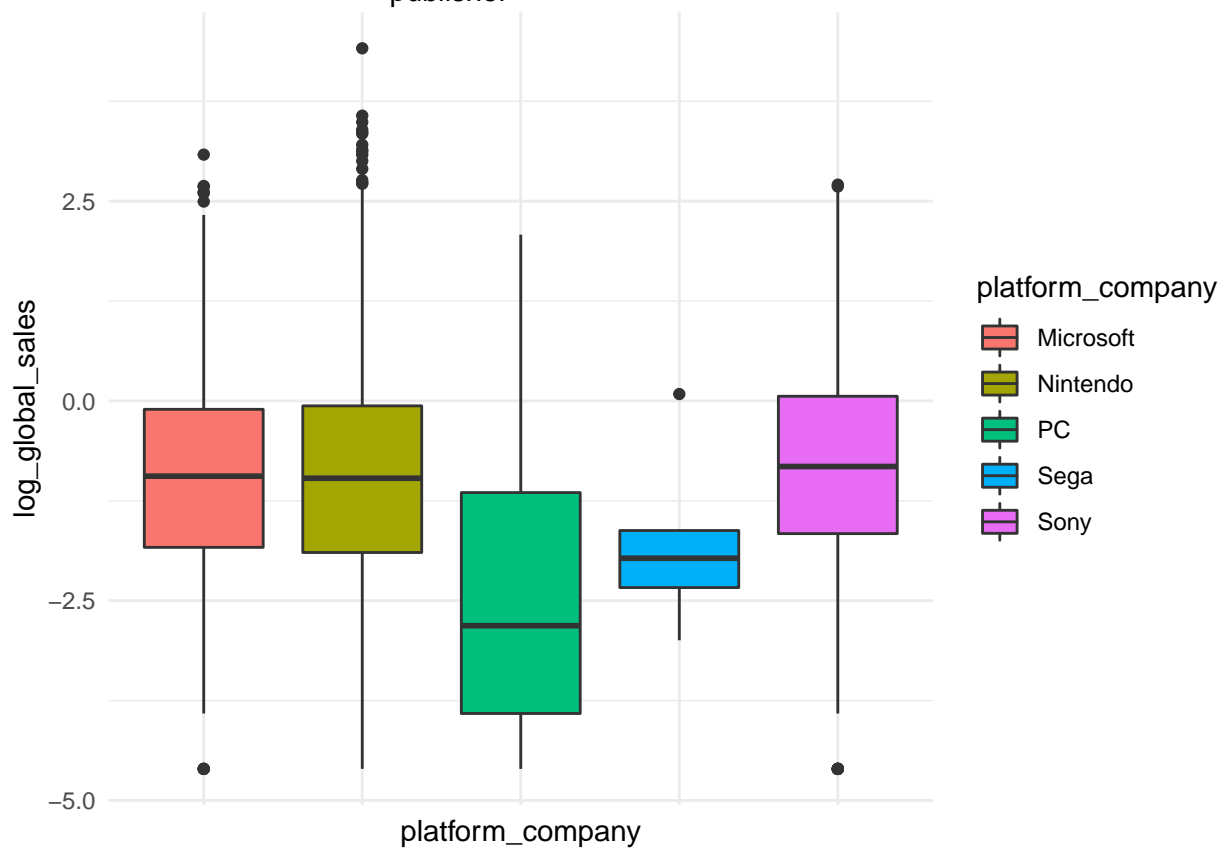
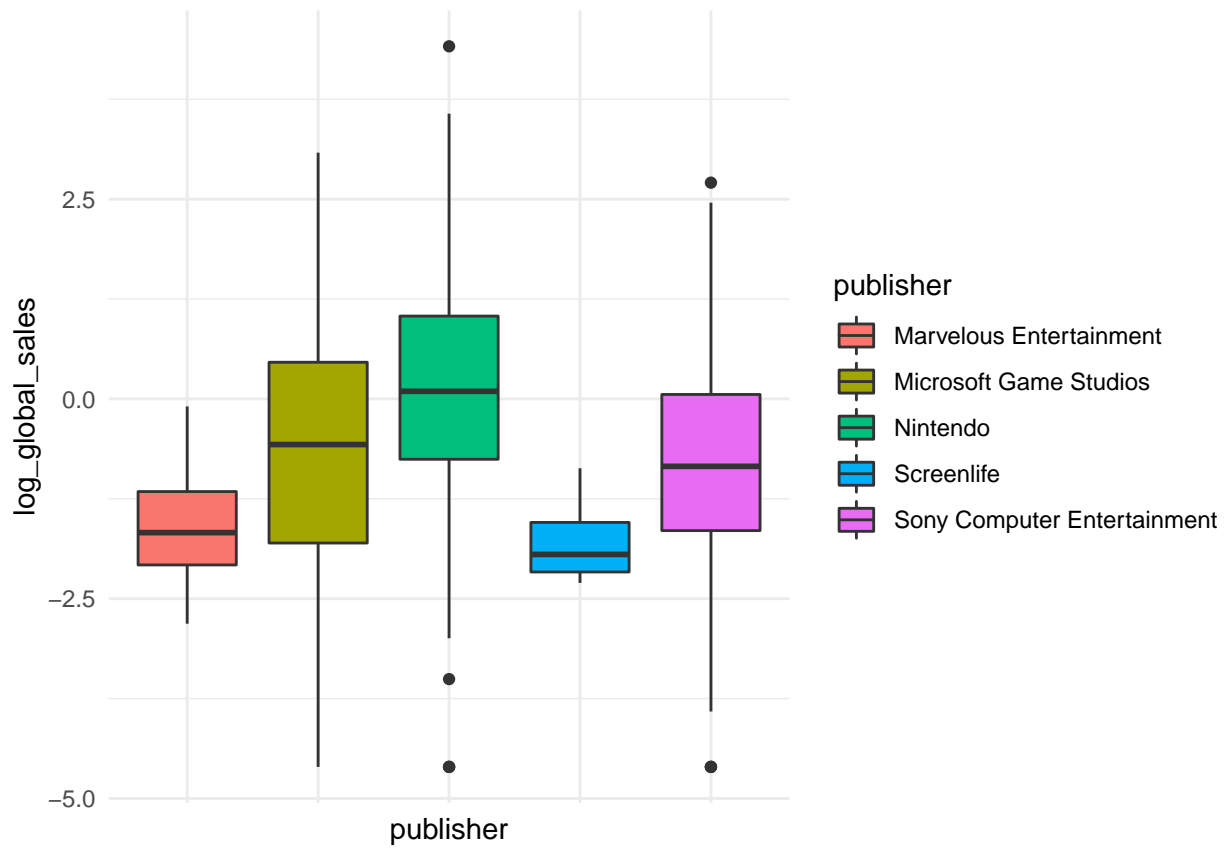


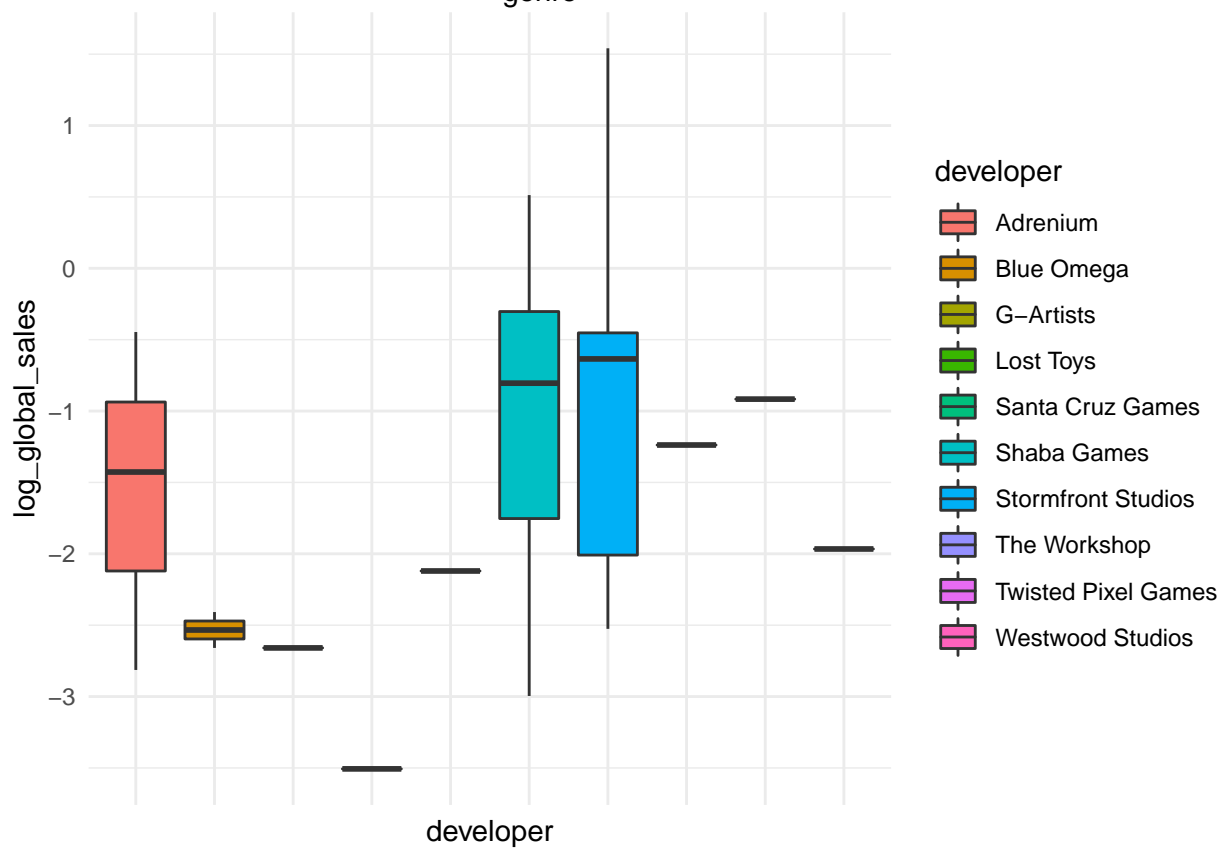
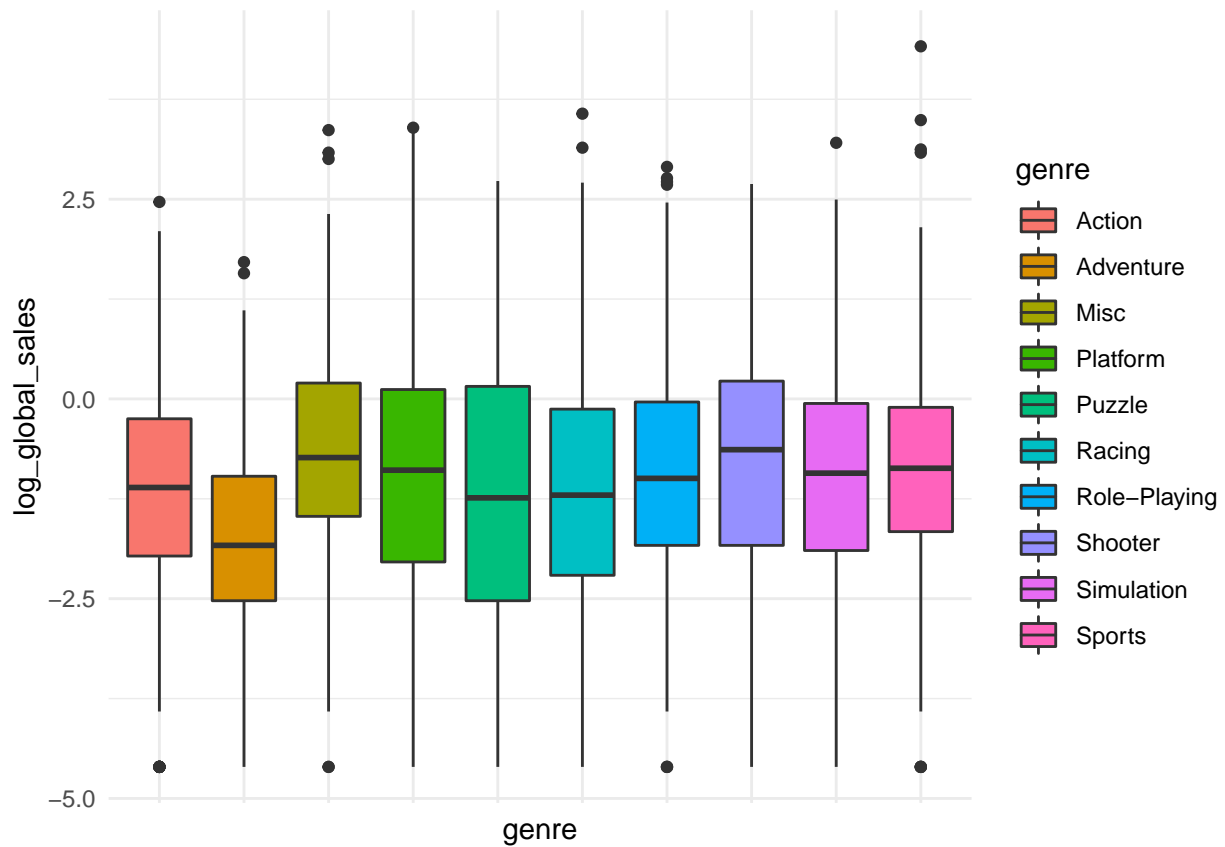


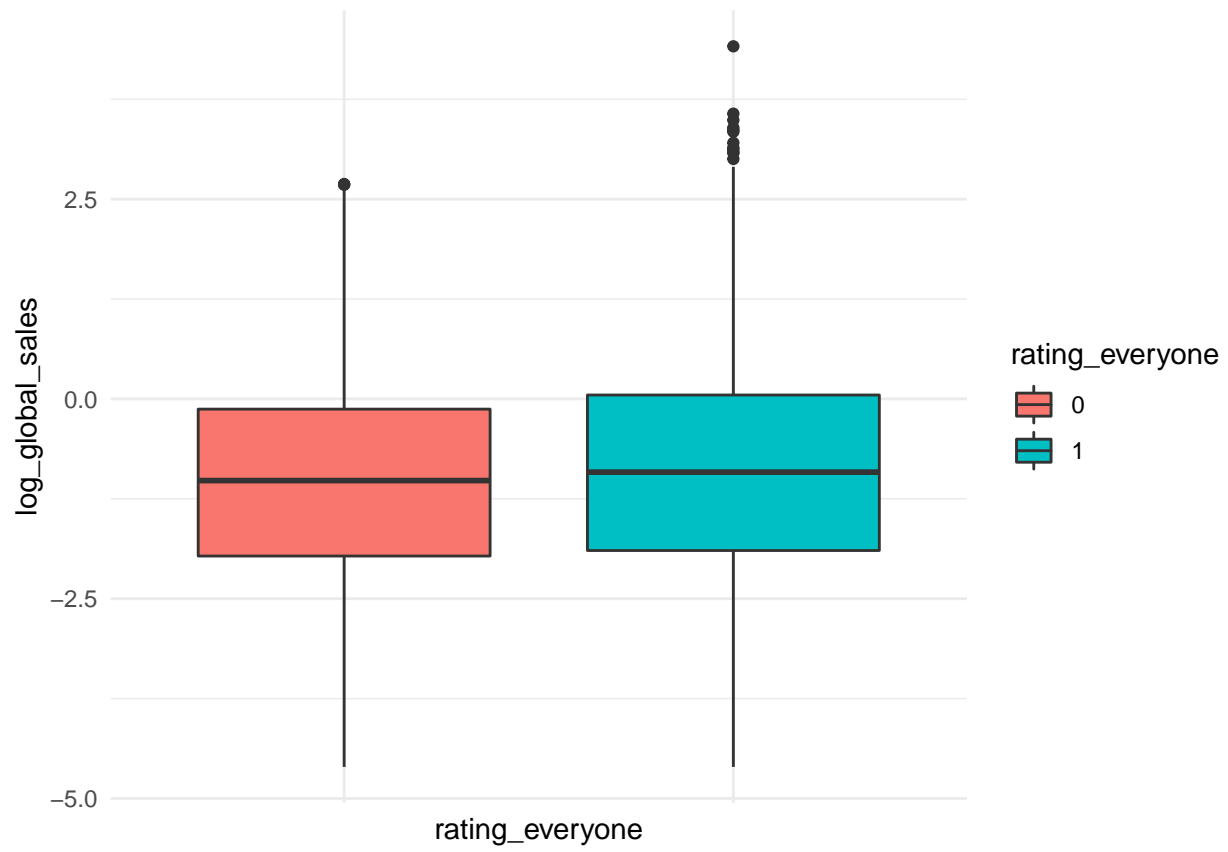




```
## [1] Nintendo                               Screenlife
## [3] Sony Computer Entertainment Microsoft Game Studios
## [5] Marvelous Entertainment
## 444 Levels: 10TACLE Studios 1C Company 2D Boy 2K Sports 3D0 ... Zushi Games
```



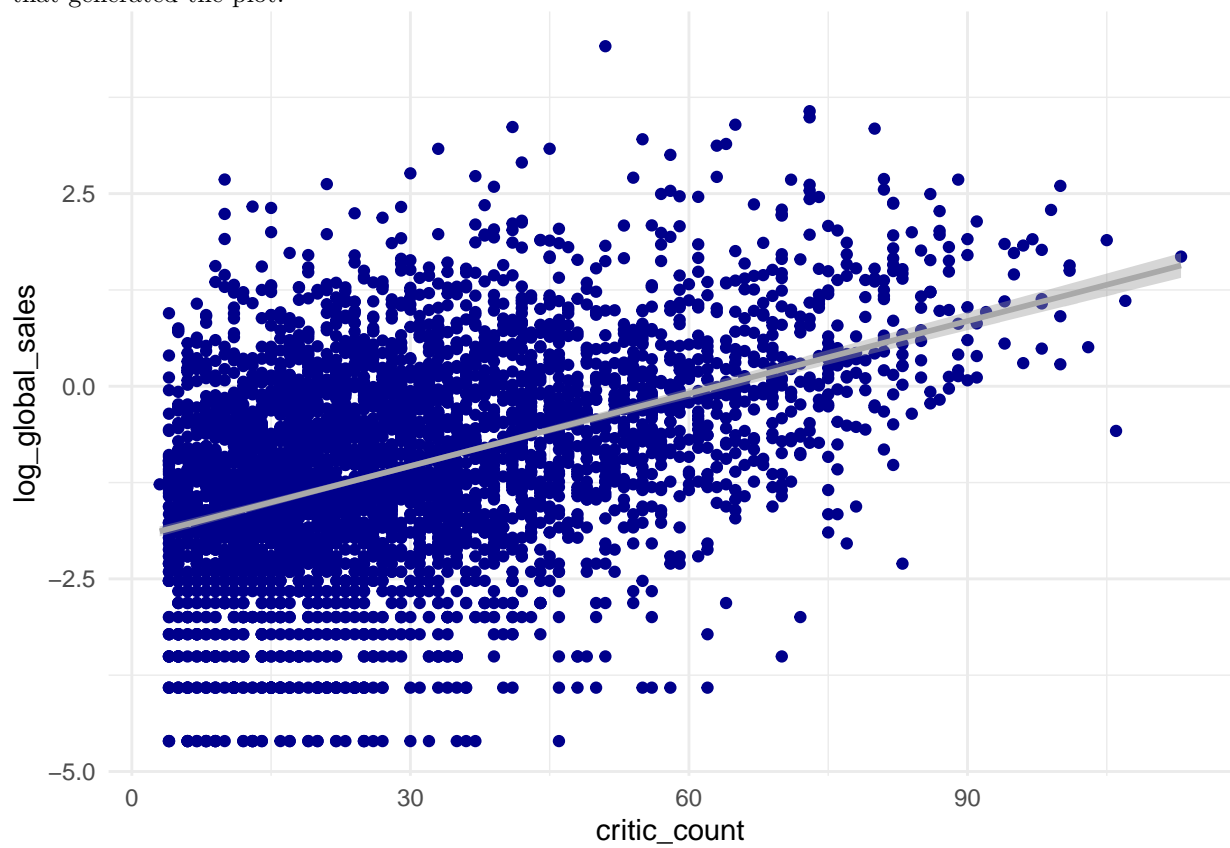


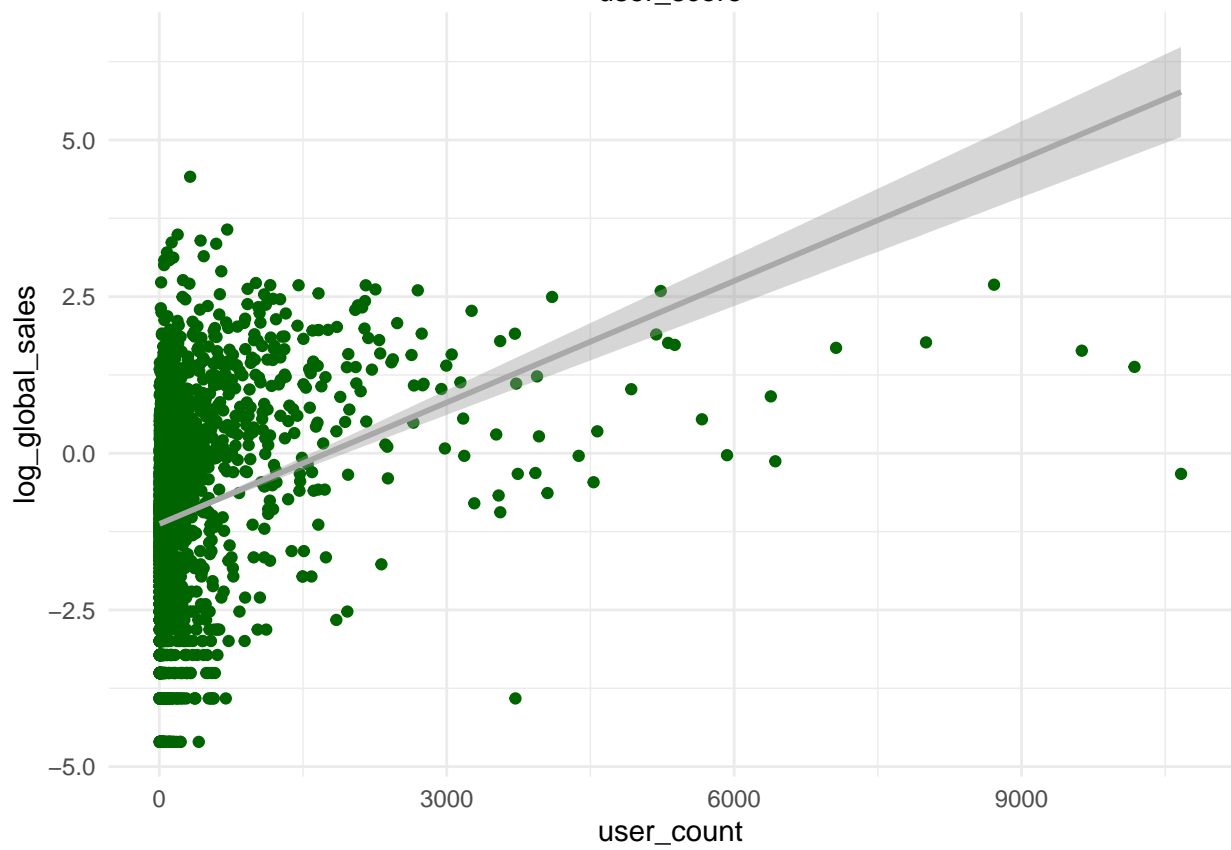
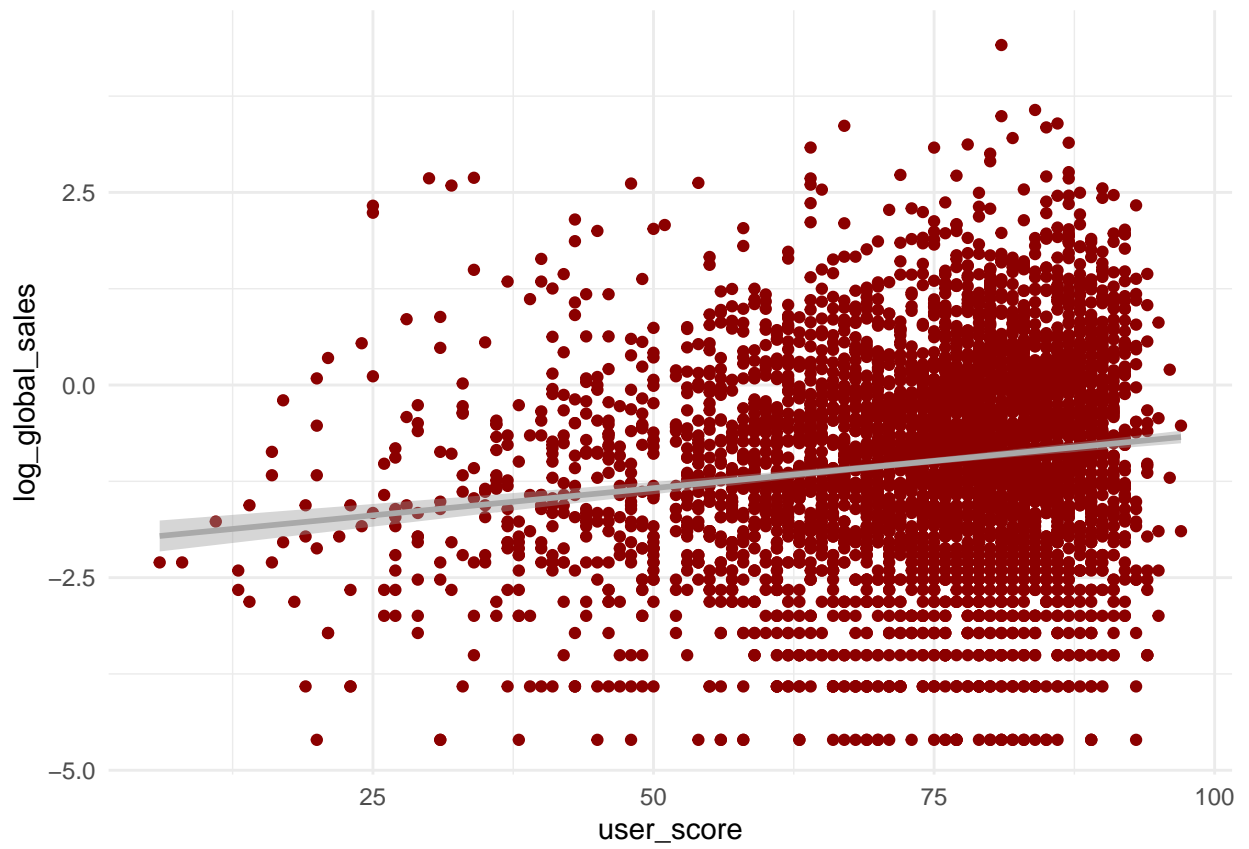


Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code



that generated the plot.





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:

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

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

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

```
## 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

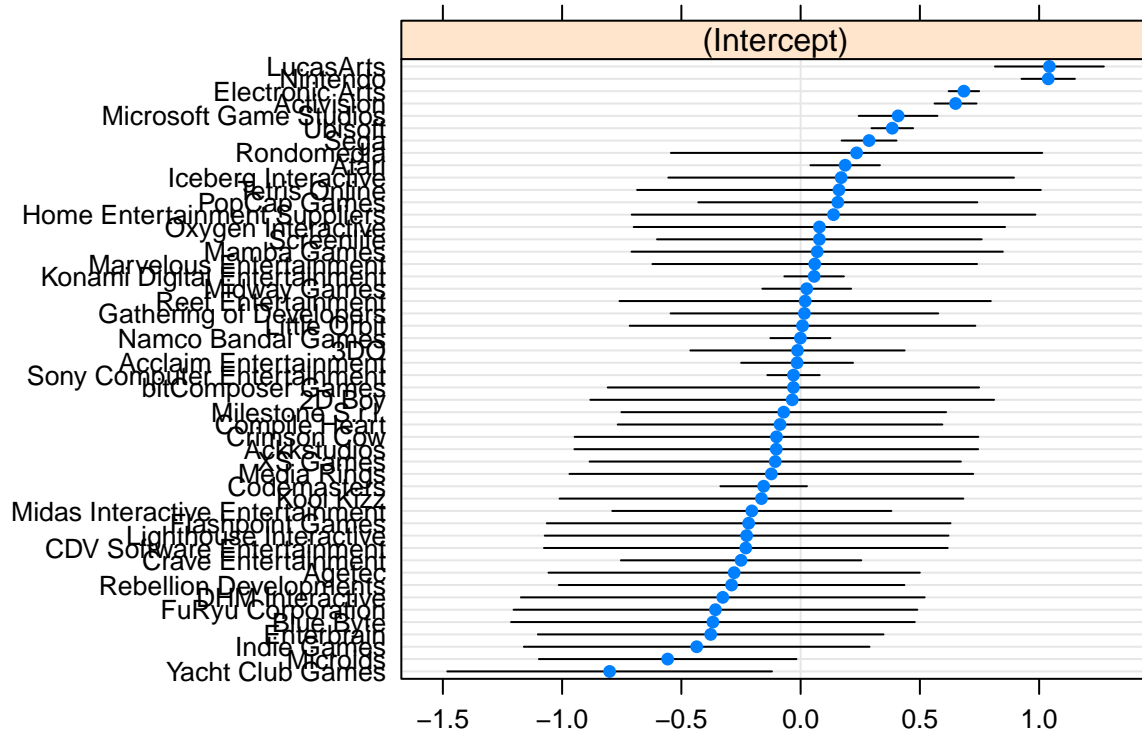
##           platform_companyNintendo
##                   2.375819
##           platform_companyPC
##                   1.901742
##           platform_companySega
##                   1.459059
##           platform_companySony
##                   2.135685
##           genreAdventure
##                   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

```
## [1] "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 platform_company rating_everyone
```

```

## Min. : -169.96 Microsoft:1628 0:4845
## 1st Qu.: -162.96 Nintendo :1818 1:2106
## Median : -146.96 PC : 692
## Mean : 0.00 Sega : 11
## 3rd Qu.: -84.96 Sony :2802
## Max. :10491.04
##

## [1] "sample_data"

## name platform
## Madden NFL 07 : 9 PS2 : 717
## LEGO Star Wars II: The Original Trilogy : 8 X360 : 516
## Need for Speed: Most Wanted : 8 PS3 : 478
## Harry Potter and the Order of the Phoenix : 7 XB : 391
## LEGO Indiana Jones: The Original Adventures: 7 PC : 362
## Madden NFL 08 : 7 Wii : 306
## (Other) :4149 (Other):1425
## year_of_release genre publisher
## Min. :2000 Action :899 Electronic Arts : 961
## 1st Qu.:2004 Sports :793 Ubisoft : 512
## Median :2007 Shooter :529 Activision : 505
## Mean :2007 Racing :396 Sony Computer Entertainment: 322
## 3rd Qu.:2010 Role-Playing:324 Nintendo : 309
## Max. :2016 Misc :275 Sega : 292
## (Other) :979 (Other) :1294
## na_sales eu_sales jp_sales other_sales
## Min. : 0.0000 Min. : 0.0000 Min. :0.00000 Min. :0.0000
## 1st Qu.: 0.0800 1st Qu.: 0.0200 1st Qu.:0.00000 1st Qu.:0.0100
## Median : 0.1900 Median : 0.0800 Median :0.00000 Median :0.0300
## Mean : 0.4842 Mean : 0.2998 Mean :0.08069 Mean :0.1014
## 3rd Qu.: 0.4800 3rd Qu.: 0.2600 3rd Qu.:0.01000 3rd Qu.:0.0900
## Max. :41.3600 Max. :28.9600 Max. :6.50000 Max. :8.4500
##
## global_sales critic_score critic_count user_score
## Min. : 0.0100 Min. :17.00 Min. : 3.00 Min. : 6.00
## 1st Qu.: 0.1500 1st Qu.:64.00 1st Qu.: 16.00 1st Qu.:66.00
## Median : 0.3700 Median :74.00 Median : 26.00 Median :77.00
## Mean : 0.9663 Mean :71.58 Mean : 30.77 Mean :73.13
## 3rd Qu.: 0.9300 3rd Qu.:81.50 3rd Qu.: 42.00 3rd Qu.:83.00
## Max. :82.5300 Max. :98.00 Max. :113.00 Max. :97.00
##
## user_count developer rating
## Min. : 4.0 Electronic Arts : 612 E :1493
## 1st Qu.: 12.0 Ubisoft : 305 T :1396
## Median : 29.0 Konami : 146 M : 732
## Mean : 181.4 Sony Computer Entertainment: 107 E10+ : 548
## 3rd Qu.: 96.0 Nintendo : 85 : 26
## Max. :10665.0 Codemasters : 66 AO : 0
## (Other) :2874 (Other): 0
## log_global_sales critic_score_c critic_count_c user_score_c
## Min. : -4.60517 Min. : -53.137 Min. : -26.016 Min. : -66.5857
## 1st Qu.: -1.89712 1st Qu.: -6.137 1st Qu.: -13.016 1st Qu.: -6.5857
## Median : -0.99425 Median : 3.863 Median : -3.016 Median : 4.4143
## Mean : -1.01178 Mean : 1.448 Mean : 1.749 Mean : 0.5435

```

```
## 3rd Qu.: -0.07257 3rd Qu.: 11.363 3rd Qu.: 12.984 3rd Qu.: 10.4143
## Max. : 4.41316 Max. : 27.863 Max. : 83.984 Max. : 24.4143
##
## user_count_c platform_company rating_everyone
## Min. : -169.960 Microsoft:1005 0:2702
## 1st Qu.: -161.960 Nintendo :1130 1:1493
## Median : -144.960 PC : 362
## Mean : 7.464 Sega : 10
## 3rd Qu.: -77.960 Sony :1688
## Max. :10491.040
##
```

