NATIONAL UNIVERSITY OF SINGAPORE

FACULTY OF SCIENCE

AY 2022/2023, SEMESTER 2
ST4248: Statistical Learning II

Term Paper

Title:
Analysis of Video Game Sales and Ratings

Matric Number:
A0201825N

Summary:

In this analysis, we explored the determinants of video game sales using a multifaceted approach. A linear mixed-effects model captured the hierarchical data structure and assessed platform and genre effects while accounting for publisher variability. Furthermore, we predicted sales through regression methods, incorporating critic and user scores, and employed multiple regression to directly examine the publishers' impact on sales. This combination of methods offered a comprehensive understanding of factors influencing video game sales, yielding valuable insights for strategic decision-making in the gaming industry.

Table of contents

Introduction

The video game sector has witnessed substantial expansion in recent years, emerging as a prominent player in the worldwide entertainment industry (Pashkov, 2021). As the volume of games launched across diverse platforms and genres continues to grow, it becomes increasingly important to identify the factors driving sales and ratings. Both sales and ratings act as key determinants of a game's triumph and its acceptance among the intended audience. Analyzing the relationship between game features, critical assessments, user opinions, and sales achievements empowers industry participants to make well-informed choices, refine promotional tactics, and ultimately deliver superior gaming experiences for enthusiasts across the globe.

Description of Data

For this paper, we will be using a Kaggle dataset - "Video Game Sales with Ratings" (Kirubi, 2016), containing information on various aspects of video games released from 2007 to 2016, including their sales performance, ratings, and general characteristics. The dataset encompasses 16,719 games released across different platforms, including PC, PlayStation, Xbox, and Nintendo consoles. Each game entry includes information on the game's name, platform, release year, genre, publisher, sales figures in North America, Europe, Japan, and other regions, as well as global sales. Additionally, the dataset features Critic_Score and User_Score from Metacritic, along with the ESRB[1] rating for each game. From this dataset, we will explore the factors that influence video game sales and examine relationships between game characteristics, ratings, and sales performance.

[1] The Entertainment Software Rating Board (ESRB) functions as a self-governing body responsible for assigning age and content classifications to consumer video games within the United States and Canada. (Wikipedia, 2022)

Initial Data Cleaning and Exploratory Data Analysis (EDA)

After dropping observations with missing values in certain columns, we are left with 7,017 games with fully-filled rows to perform our analysis with. This is done especially to ensure that all the models can be run as accurately as possible without ambiguity when considering all of the variables in the dataset. We then plot a correlation plot to measure the collinearity of the numeric variables in the dataset, as shown below:

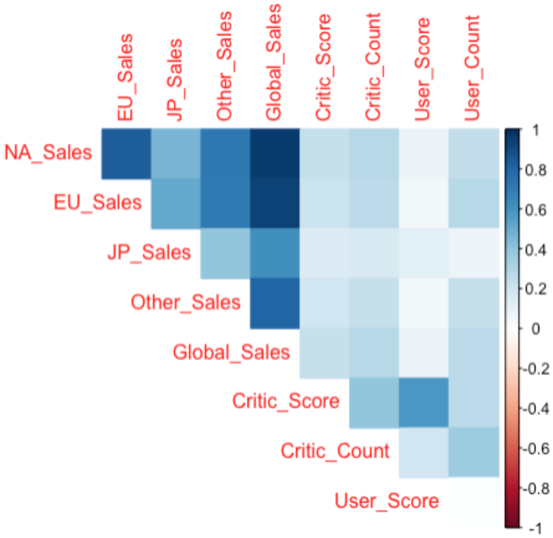


Figure 1: Collinearity plot of numerical variables in cleaned dataset

As evident above, there appears to be very high collinearity between Global_Sales and the various (location)_Sales variables, which is understandably obvious as the sum of these variables add up to provide the Global_Sales value for each game. Another unique finding would be the high collinearity between Critic_Score and User_Score, confirming that professional critics and consumers alike generally have similar grading preferences when evaluating the enjoyment of a video game, despite certain anomalies with polarizing reviews such as Call of Duty: Infinite Warfare and Mass Effect 3 (Phillips, 2012), where consumers in particular have varied their opinions of these games - especially when comparing the contexts of their initial release and a few years afterwards.

Using Multiple Regression to Investigate Impact of Publishers on Global Sales

In this section, we shall investigate the impact of the various publishers on global video game sales. For this, we first perform data cleaning of the original dataset, where we remove observations with missing values for Critic_Score, User_Score, Rating and Publisher. With 8,137 games remaining in the dataset, we next perform one-hot encoding on the Publisher variable, before fitting a multiple regression model to identify the top publishers based on their regression coefficients:

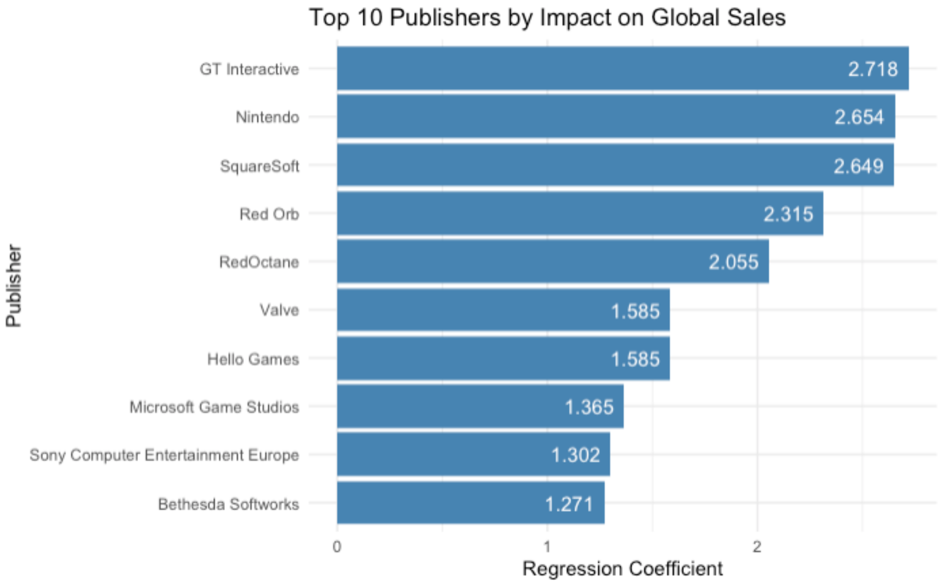


Figure 2: Bar chart representing top 10 publishers by impact on global sales, based on regression coefficient

From the above, we can identify the top 10 publishers with the greatest impact on Global_Sales, including popular publishers today such as Valve, Sony, and Microsoft. In particular, Nintendo's dominance as a top publisher can be accredited to the successful launch of the Wii console, with Wii Sports unanimously being the top contributor in Global_Sales during the late 2000s.

Using Various Regression Methods to Predict Global Sales using Critic Score and User Score

Using the cleaned dataset from the EDA section, we will attempt various regression methods: namely multiple linear regression (MR), random forest (RF) and XGBoost - in order to assess if Critic_Score and User_Score are indeed good predictors of Global_Sales. This can be done by comparing the performances of the 3 models, assuming that 10-fold cross-validation is applied, with a table of summarized results below:

Evaluation Metric/Model	MR	RF	XGBoost
Mean RMSE	1.766241	**1.730558**	1.76862
Mean R-squared	0.07863555	**0.1105488**	0.06202084
Mean MAE	0.760865	**0.6995031**	0.7022947

Based on the above, it appears that the random forest model performs the best across all three evaluation metrics. However, the R-squared values for all three models are relatively low, indicating that these models do not explain a substantial portion of the variance in Global_Sales. To improve model performance, we shall normalize the predictors and make use of a 80-20 train-test split before running the 3 models again with 10-fold cross validation:

Evaluation Metric/Model	MR	RF	XGBoost
Mean RMSE	1.496797	**1.684257**	1.476058
Mean R-squared	0.05374856	**0.0286819**	0.0888855
Mean MAE	0.717222	**0.7204104**	0.6358971

It appears that although there is no significant impact on the multiple regression and random forest models, normalizing the predictors did improve the XGBoost model, as evident in the improvements in RMSE and R-squared scores. Despite Critic_Score and User_Score not being strong predictors, random forest is still the optimal model to use, especially without normalization.

Using Mixed-Effects Model to Investigate Impact of Publishers, Platform and Genres on Global Sales

In this analysis, we used a linear mixed-effects model to account for the hierarchical nature of the data, with video games nested within publishers. The model included fixed effects for platform and genre, as well as random intercepts for publishers. Our results revealed significant effects of platform and genre on global video game sales. Particularly, the platforms PS, PS2, PS3, PS4, Wii, and X360, as well as the shooter and strategy genres, were found to have notable influences on sales. In particular, strategy games have a negative effect on global sales, presumably because they were not as popular globally in the 1990s and 200s compared to today. This model allowed us to account for variability in sales performance across different publishers. Model diagnostics, including residual plots and Q-Q plots, suggested that our linear mixed-effects model was an appropriate choice for the data, with only minor deviations from normality in the tails of the residual distribution.

Conclusion

In conclusion, we conducted a comprehensive analysis of the video game sales dataset using three distinct approaches. First, we implemented a multiple regression model to investigate the impact of publishers on global sales directly, highlighting the importance of accounting for publisher-level effects when modeling video game sales. Second, we used different regression methods to predict global sales based on critic scores and user scores, allowing us to explore the relationship between these variables. Lastly, we employed a linear mixed-effects model to account for the hierarchical nature of the data, capturing the variability in sales performance across different publishers while accounting for platform and genre effects. By combining these approaches, we gained a deeper understanding of the factors that drive video game sales and provided valuable insights that can inform strategic decisions within the gaming industry.

References

Pashkov, S. (2021). Video game industry market analysis: Approaches that resulted in industry success and high demand. Retrieved April 11, 2023, from https://www.theseus.fi/bitstream/handle/10024/497979/e1700994ThesisRevised.pdf?sequence=2

Rush Kirubi. (2016). Video Game Sales with Ratings. Retrieved April 11, 2023, from https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings

Wikipedia contributors. (2022). Entertainment Software Rating Board. In Wikipedia, The Free Encyclopedia. Retrieved April 11, 2023, from https://en.wikipedia.org/wiki/Entertainment_Software_Rating_Board

Phillips, T. (2012). Why Do Gamers and Critics Disagree on Game Reviews? Retrieved April 11, 2023, from https://gamerant.com/game-review-player-critic-gap-metacritic/

Appendix

Column Name	Column Type	Description
Name	String	The title of the video game.
Platform	String	The gaming console or system on which the game is available (e.g., PC, PlayStation, Xbox).
Year_of_Release	Integer	The year the game was released.
Genre	String	The game's genre (e.g., action, adventure, sports, strategy).
Publisher	String	The company responsible for publishing the game.
NA_Sales	Float	Total sales of the game in North America, in millions of units.
EU_Sales	Float	Total sales of the game in Europe, in millions of units.
JP_Sales	Float	Total sales of the game in Japan, in millions of units.
Other_Sales	Float	Total sales of the game in other regions, in millions of units.
Global_Sales	Float	Total worldwide sales of the game, in millions of units.
Critic_Score	Float	Average score given by professional critics, as aggregated Metacritic (range: 0-100).
Critic_Count	Integer	Number of critics who reviewed the game.
User_Score	Float	Average score given by users, as aggregated by Metacritic (range: 0-10).
User_Count	Integer	Number of users who reviewed the game.
Developer	String	The company responsible for developing the game.
Rating	String	The ESRB rating assigned to the game (e.g., E for Everyone, T for Teen, M for Mature).

Table 1: Schema of “Video Game Sales with Ratings” dataset, sourced from Kaggle

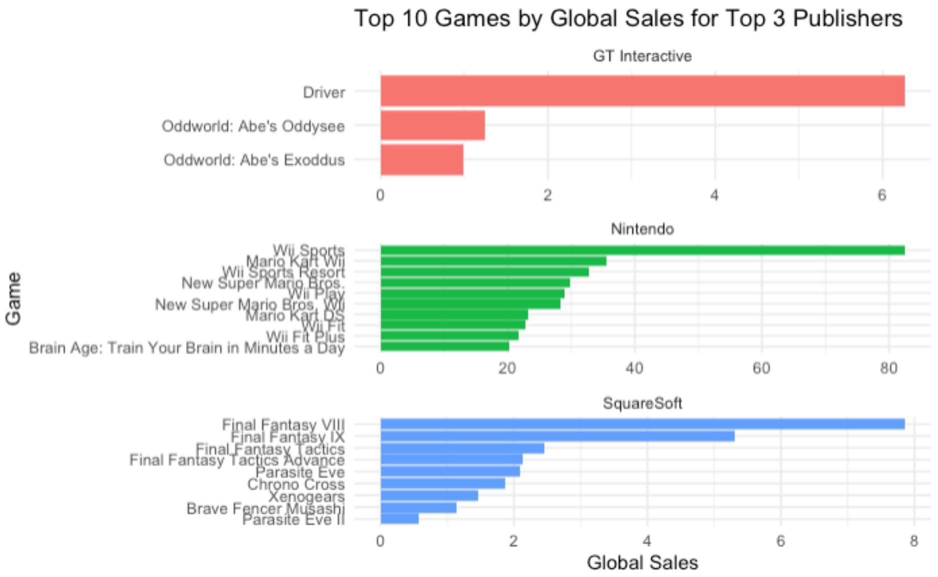


Figure 3: Faceted bar chart representing top 10 games by global sales for Top 3 most impactful publishers, based on regression coefficient

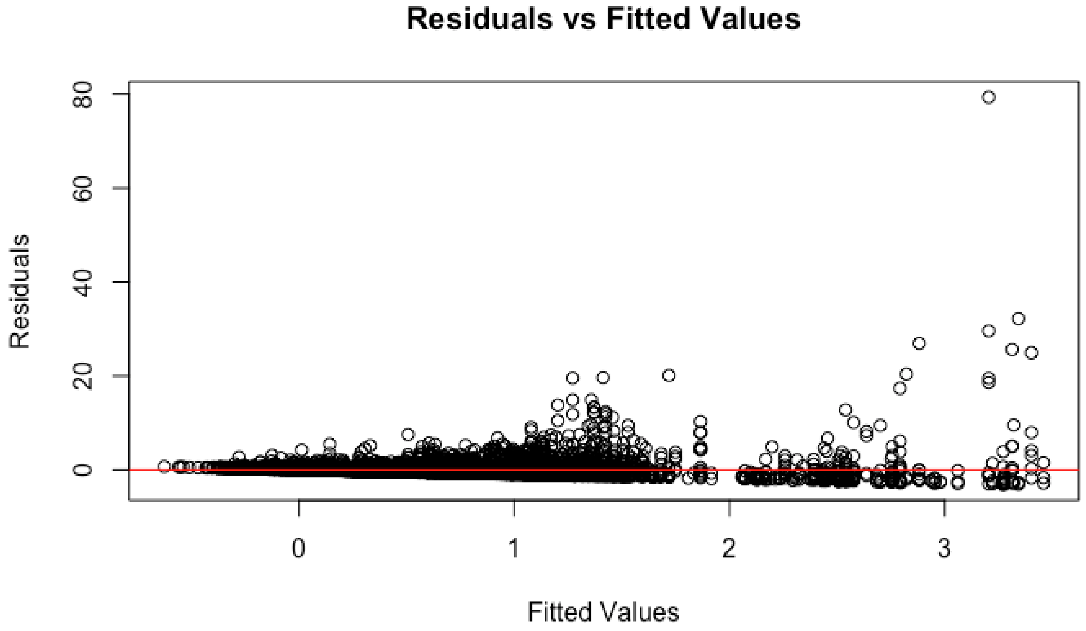


Figure 4: Residual plot to assess linear and homoscedasticity of the mixed model

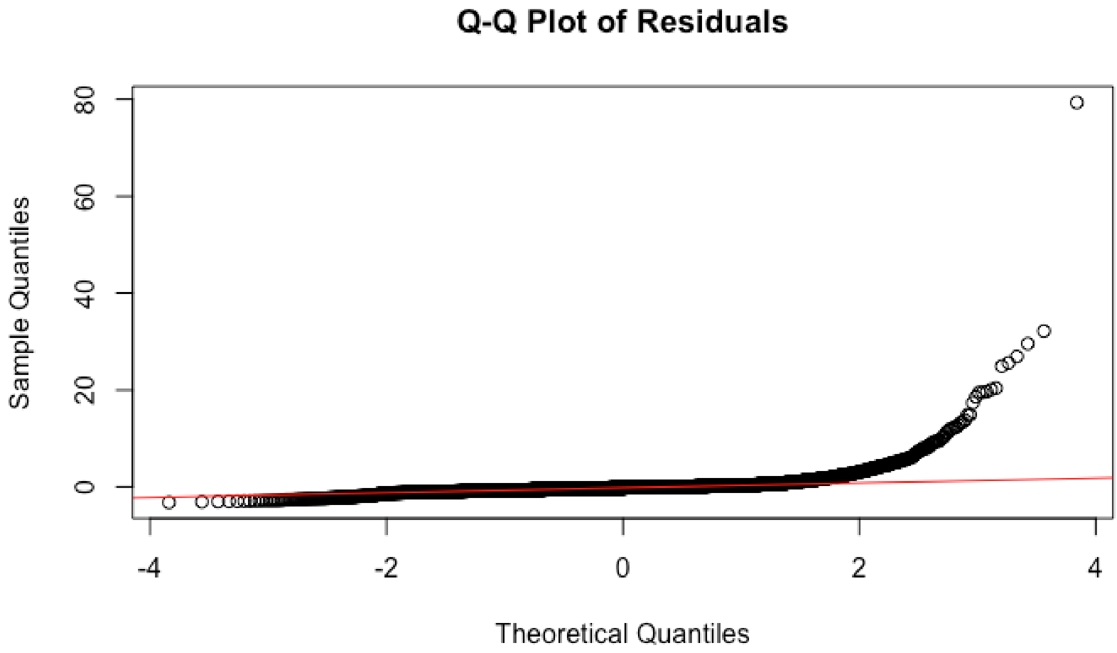
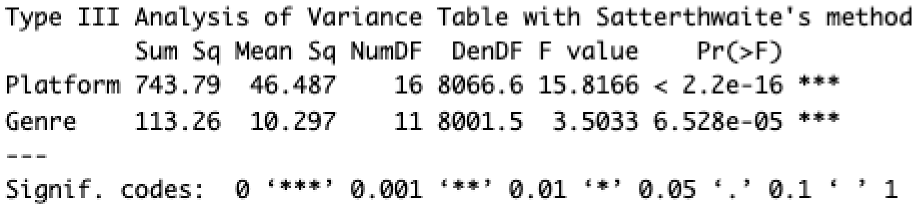
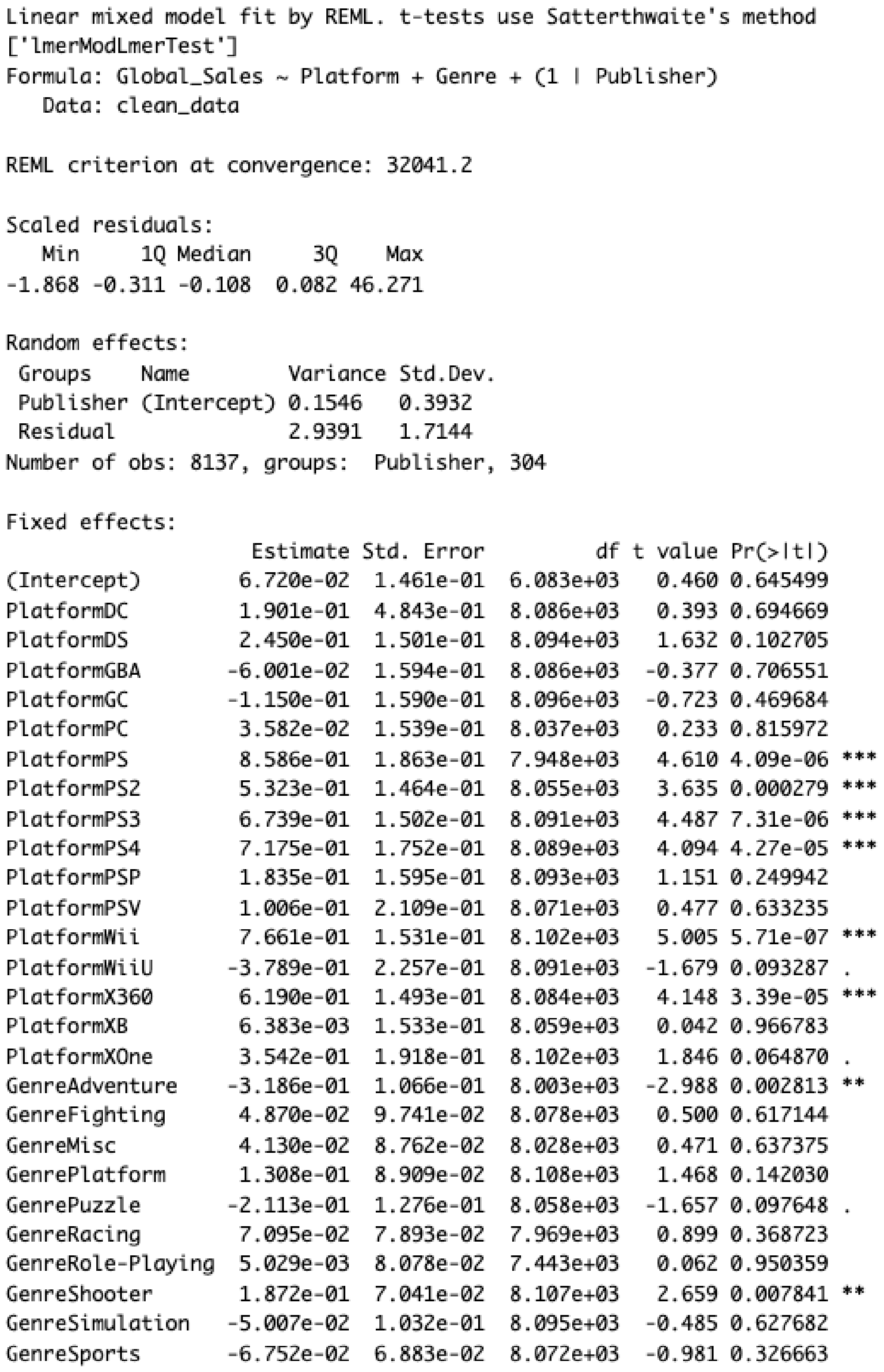


Figure 5: Q-Q plot to assess normality of residuals of the mixed model



```
Type III Analysis of Variance Table with Satterthwaite's method
         Sum Sq Mean Sq NumDF  DenDF F value    Pr(>F)    
Platform 743.79  46.487    16 8066.6 15.8166 < 2.2e-16 ***
Genre    113.26  10.297    11 8001.5  3.5033 6.528e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: ANOVA table of mixed model suggesting platform and genre’s high importance in explaining variance in global sales, after accounting for random effect of publishers



```
Linear mixed model fit by REML. t-tests use Satterthwaite's method
['lmerModLmerTest']
Formula: Global_Sales ~ Platform + Genre + (1 | Publisher)
   Data: clean_data

REML criterion at convergence: 32041.2

Scaled residuals: 
   Min     1Q Median     3Q    Max 
-1.868 -0.311 -0.108  0.082 46.271 

Random effects:
 Groups    Name        Variance Std.Dev.
 Publisher (Intercept) 0.1546   0.3932  
 Residual              2.9391   1.7144  
Number of obs: 8137, groups:  Publisher, 304

Fixed effects:
                   Estimate Std. Error        df t value Pr(>|t|)    
(Intercept)       6.720e-02  1.461e-01  6.083e+03   0.460 0.645499    
PlatformDC        1.901e-01  4.843e-01  8.086e+03   0.393 0.694669    
PlatformDS        2.450e-01  1.501e-01  8.094e+03   1.632 0.102705    
PlatformGBA      -6.001e-02  1.594e-01  8.086e+03  -0.377 0.706551    
PlatformGC       -1.150e-01  1.590e-01  8.096e+03  -0.723 0.469684    
PlatformPC        3.582e-02  1.539e-01  8.037e+03   0.233 0.815972    
PlatformPS        8.586e-01  1.863e-01  7.948e+03   4.610 4.09e-06 ***
PlatformPS2       5.323e-01  1.464e-01  8.055e+03   3.635 0.000279 ***
PlatformPS3       6.739e-01  1.502e-01  8.091e+03   4.487 7.31e-06 ***
PlatformPS4       7.175e-01  1.752e-01  8.089e+03   4.094 4.27e-05 ***
PlatformPSP       1.835e-01  1.595e-01  8.093e+03   1.151 0.249942    
PlatformPSV       1.006e-01  2.109e-01  8.071e+03   0.477 0.633235    
PlatformWii       7.661e-01  1.531e-01  8.102e+03   5.005 5.71e-07 ***
PlatformWiiU     -3.789e-01  2.257e-01  8.091e+03  -1.679 0.093287 .  
PlatformX360      6.190e-01  1.493e-01  8.084e+03   4.148 3.39e-05 ***
PlatformXB        6.383e-03  1.533e-01  8.059e+03   0.042 0.966783    
PlatformXOne      3.542e-01  1.918e-01  8.102e+03   1.846 0.064870 .  
GenreAdventure   -3.186e-01  1.066e-01  8.003e+03  -2.988 0.002813 ** 
GenreFighting     4.870e-02  9.741e-02  8.078e+03   0.500 0.617144    
GenreMisc         4.130e-02  8.762e-02  8.028e+03   0.471 0.637375    
GenrePlatform     1.308e-01  8.909e-02  8.108e+03   1.468 0.142030    
GenrePuzzle      -2.113e-01  1.276e-01  8.058e+03  -1.657 0.097648 .  
GenreRacing       7.095e-02  7.893e-02  7.969e+03   0.899 0.368723    
GenreRole-Playing 5.029e-03  8.078e-02  7.443e+03   0.062 0.950359    
GenreShooter      1.872e-01  7.041e-02  8.107e+03   2.659 0.007841 ** 
GenreSimulation  -5.007e-02  1.032e-01  8.095e+03  -0.485 0.627682    
GenreSports      -6.752e-02  6.883e-02  8.072e+03  -0.981 0.326663    
```

Figure 7: Summary of mixed model indicating significant platforms and genres based on t-value