The Effects of Pre-Release Information on Public Reception

1 Introduction

As the market becomes more and more saturated with videogames, those who play them may find it difficult to figure out what games are worth buying. This issue is important as video games on release are on average 60$ and no one wants to spend that much money on something they will get bored of after only playing for a few hours. By drawing relationships between these variables and the reception of the gamer population, we can provide consumers with a better understanding on how to decide what game to purchase. I aim to answer the following questions through this model:

* Does the publisher, critic reception, genre, and developer affect the user reception of a video game individually?
* Do these same variables produce a greater or lesser effect when their impact is viewed together?

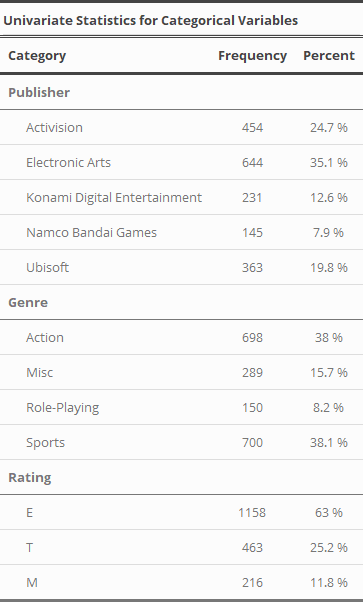
I believe that the variables discussed should all affect the user score individually due to how public perception of a game and expectations of it can vastly change its reception. Therefore, the combination of these variables could produce a stronger effect than they could individually.

2 Data

         The data for this analysis was originally compiled together by user Gregory Smith and made publicly available on Kaggle[[1]](#footnote-1). User Rush Kirubi found this set and was inspired by it, choosing to expand upon the data by adding additional variables and republished this expanded data set to his Kaggle[[2]](#footnote-2). The unit of observation for this dataset is the videogame name and Smith initially created the dataset through non-probability sampling, only adding games that sold 100,000 copies or more to his dataset. Kirubi expanded upon this initial collection by conducting a web scrape of the website Metacritic, a popular media review site that has an extensive section for video games. Due to the differences in games listed on each site the limitations of this data show itself in how there are many missing observations in the expanded data set. Of the 16,719 entries Kirubi states that there are about 6,900 entries that have an observation for every variable. From this 6,900 we are only going to be analyzing those that come from the 5 most popular publishing companies, the 5 most popular developers, and fall under the 4 most popular genres; Action, Sports, RPGs, and Miscellaneous. Ideally, with this data set we should be able to identify statistically significant relationships to predict public reception to a video game.

2.1 Variables

For this analysis the variables that I will be using to create our models will be the Metacritic score, the publishing company, the developers, and the genre of the game.

2.1.1 Publishers

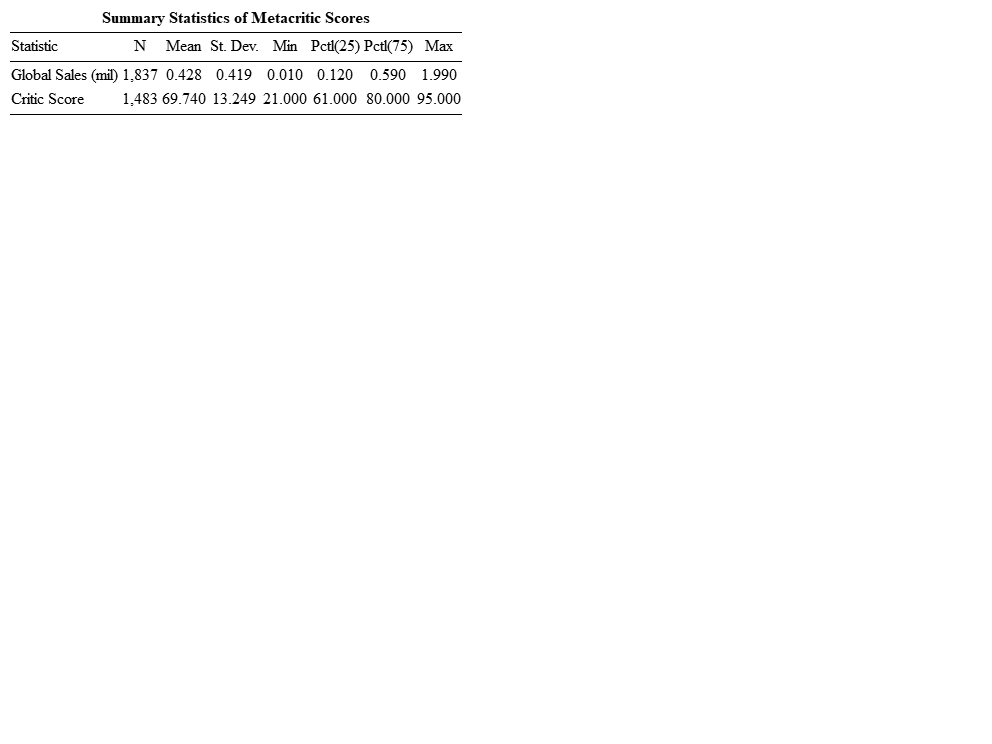
A publishing company is in charge of financing the development of a videogame, securing contracts for distribution, marketing, and handling the localization of these games to regions outside of the one it was created in. In the dataset we have taken only the top 5 publishing companies. Electronic Arts makes up the highest proportion of games, publishing 35.1% of the games in our subset. On the other side of the spectrum Namco Bandai Games only published 7.9% or 145 videogames.

 2.1.2 Genre

Genre is a variable that is similar to those used by movies and TV shows. For videogames it is used as a way to categorize media based on its gameplay challenges and is not defined by its visual style or narrative. From our univariate table, the genres that make up the largest proportions are the Action genre (38%) and the Sports genre (38.1%). Meanwhile, games that fall under the “Miscellaneous” genre make up 15.7% and RPGs make up 8.2%.

 2.1.3 ESRB Rating

Ratings are assigned to games by the ESRB according the type of content on display. For example, the presence of adult content such as excessive violence, nudity, and the presence of profanity will result in a M or ‘Mature’ rating. Over time the ESRB have used different labels for similar classifications. For this reason, the ‘E’ group encompasses the following labels: E, E10+, EC, and K-A. Similarly, the ‘M’ group encompasses the ‘Adult Only’ rating as well as the ‘M’ rating. From the univariate table shows us that E games are the most prevalent making up 63% of the sample while T and M games make up 25.2% and 11.8% of the data respectively.

 2.1.4 Metacritic Score

Metacritic is a website that reviews videogames and they rate them using two scores; the User Score and the Metacritic score. The Metacritic score is calculated through calculating a weighted average of critic ratings. Each critic is assigned a value that give their reviews more or less weight and this value is based on the critics quality and stature. From the summary table we can see that the average Metacritic score is 69.74 with a standard deviation of 13.24. Additionally, the IQR for Metacritic scores is 19 ranging from 61 to 80.

2.1.5 Global Sales

The final variable involved in this analysis is the global sales variable. The global sales variable is a measurement of units sold by a videogame in all regions. This metric is also measured in millions of units meaning that 1.2 would equal 1,200,000 copies sold. From the summary statistic table for this variable we can see that there exists standard deviation of around 419,000 copies and a mean of 428,000 copies. Interestingly, the top selling 25% of videogames make up the range of 590,000-1.9 million copies sold. This shows us that there exist some extreme outliers in our data.

3 Results and Interpretations

A screenshot of a video game

Description automatically generated The first thing that we did was to visualize the relationships between global sales and the predictors. From the arrangement of graphs in Figure 1 we can get an initial idea of these relationships. First, a high Metacritic score seems to correlate with higher global sales numbers. Second, genre seems to play a significant role as the 75th percentile of Role-Playing and Sports games is significantly higher than that of Miscellaneous and Action games. Third, Rating seems to have a direct influence on higher sales numbers as Mature games have much better earning potential than that of the other two ratings. Finally, the differences in sales between publishers seems to suggest the existence of a relationship.

A screenshot of a cell phone

Description automatically generated After examining these relationships, we created the first linear model from it we can observe multiple things. The coefficient of the intercept tells us that with the Metacritic score at 0 the videogame will sell -472,000 copies and since this is not possible this value is not meaningful. The coefficient of the Metacritic score tells us that for every 1.0 increase in Metacritic score 13,000 copies of that game is sold. From here we look at the p-value of the f-test and r-squared values which are <.001 and .17 respectively. This tells us that this model fits our data better than one without independent variables would and that Metacritic score explains 17% of the variance in global sales.

A close up of a map

Description automatically generatedAfter evaluating the values in Model 1 we move to evaluate the assumptions. The first assumption of independence of errors can be evaluated as there may be some curvilinear relationship between these variables. Next, we look at the plot of residuals and from it we can see that there is a large amount of deviation in normality in the upper tail so we should evaluate for influential outliers. This check for influential outliers returns that observations 220, 57, and 66 are influential outliers. By removing those outliers and squaring the predictor variable, we can address the violation of independent errors and influential outliers. After checking the assumptions again, we conclude that there is no noticeable difference and add variables to our analysis to create Model 2.

A picture containing meter, drawing

Description automatically generatedA screenshot of a cell phone

Description automatically generated The intercept coefficient in Model 2 is similarly insignificant as it is negative as games cannot sell negative copies. First, the miscellaneous genre is the only significant predictor and estimates an increase of .160. Second, all the other ratings are significant predictors when compared to the reference with ‘T’ linked to an increase of .071 and ‘M’ estimating an increase of .203 in global sales. The publishers Konami and Ubisoft are considered significant predictors and estimate a decrease in sales of .206 and .113 in global sales respectively. This model is also significant so we must compare Model 2 to Model 1 to see if more variables gives us a more accurate model. From Table 1 we can see that the p-value is less than the selected alpha of .05, allowing us to conclude that the bigger model, Model 2, is preferred. After confirming this we moved to evaluate the assumptions for Model 2.

As it can be seen in the plot of residuals, the extreme deviation in normality in the upper tail seen in Model 1 is still present in Model 2, thus the residuals are not normally distributed. Next, we check for multicollinearity by evaluating the Variance Inflaction Factor and find that no VIF value exceeds the determined safe threshold, with the highest value being 2.04. A screenshot of a map

Description automatically generatedAdditionally, we note that the cone like shape of the fitted residuals graph would be grounds to say the assumption of constant variance has been violated and after running a Breuch-Pagan test it can be concluded that in fact it has. Finally, in looking for influential outliers we find many and note to remove them to refit the model.

A screenshot of a cell phone

Description automatically generatedA picture containing meter, drawing

Description automatically generatedThe final model that we created has dropped the publisher variable, removed the influential outliers seen in Model 2, and squared the Metacritic score values. We decided to drop the publisher variable because it seemed to be affecting the analysis negatively. After conducting an ANOVA analysis between Models 2 and 3 we can see that the p-value is less than the alpha and therefore the bigger Model 2 is more accurate than Model 3. Checking the assumptions shows us that all of the problems that were present in the past models are still there and there are even more influential outliers.

A screenshot of a map

Description automatically generated4 Conclusion

Overall, we were unable to solve the problems plaguing our models and our final model is not a good predictor for the global sales of videogames. Each iteration of the models suffered greatly from influential outliers and a violation of constant variance. Attempts to remove the influential outliers were made but each time more appeared in the following refit. The final model violates the assumption of equal variance, does not have normally distributed and independent errors, and contains many influential outliers.

1. <https://www.kaggle.com/gregorut/videogamesales> [↑](#footnote-ref-1)
2. <https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings> [↑](#footnote-ref-2)