**Predicting Game Sales and Minimizing Risks:**

**An Analysis of Video Game Publishers’ Successes and Failures**

**1. Abstract**

The growth of the video game industry in the past decade, both in terms of popularity and business, shows increasing potential for the companies that comprise the industry. The evolving hardware that allows video games to be more realistic and engaging plays a role in the industry’s growth; however, it is the software, or the games themselves, that drives the sales of the hardware, and thus, plays the most important role in the growth of gaming industry. Publishing and developing video game software is a cooperative endeavor that allows both publishers and developers to attain financial success, while not compromising the creative process of game development that produces a more entertaining and visceral experience. The publisher’s goal is to generate the highest profit possible on the games they finance. To achieve this goal, publishers must make decisions on the developers they hire and fund. This paper discusses the prediction of global sales (in units sold) from other characteristics of video game software using the R statistical programming language. From this analysis, publishers could uncover which aspects of video game software have the most impact on sales and alter their business decisions, accordingly.

**2. Background**

It is important to understand the structures and roles of both publishers and developers, and what changes in the video game industry have made the largest impact on their game-making decisions.

**2.1 The Role of the Developer**

Video game development studios are the creative soul of the video game industry and are responsible for ultimately producing the games in their entirety. The structures of these companies vary wildly by their size and their relationship to the console platforms they develop games for. Game developers can range from a single person who handles the development process alone, to a large team that splits the process among each other based on the specializations of its team members. The video game development process requires a variety of jobs that create the game’s story, art assets, music, game design, and game mechanics. While it could take a small development team a few years to complete a low-budget project, a large development studio could complete a multi-million-dollar game in the same time-frame.

Game developers also vary by their relationship to the console platform they create games for and are usually one of four distinct types: first-party, second-party, third-party, or indie. First-party developers are a part of the company that manufactures a console platform and exclusively develops games for that company. Second-party developers are studios that are partially owned or have exclusivity contracts with console platform-holders which dictates that they develop games only for that platform. Third-party developers are studios that develop games without exclusivity to any console platform. While most third-party developers are also contracted by publishers to develop a game, it is not uncommon for a third-party developer to publish their own games as well. Finally, indie game developers are smaller studios that will usually self-publish their own low-budget projects. Most indie games rarely gain much notoriety but will sometimes be featured on multiple console platforms.

**2.2 The Role of the Publisher**

Video game publishers are the driving force of the industry and are responsible for all the logistics involved in making and selling games. Game publishers are usually large companies that can handle the financial burden of video game development and are willing to take business risks. Three of the largest game publishers are also the manufacturers of console platforms: Nintendo with the Switch, Sony with the PlayStation 4, and Microsoft with the Xbox One. Big publishers like these will often have internal development studios, or first-party developers, that exclusively make games for that company’s platform. However, publishers do more than just finance game development, they are also responsible for: manufacturing the physical copies of the games, marketing the game through research and advertising, and handling all the costs associated with licenses and localization to other regions. The financial risks associated with game development and marketing are constantly increasing as game development becomes more expensive with the evolving technology and consumer expectations. Simply, as video games become more realistic, they become more labor-intensive and costlier for developers to create.

Along with heavy competition from other publishers, there are financial risks associated with capturing and keeping the attention of consumers with large marketing campaigns. Publishers decide how and when to market their upcoming games and if the marketing does not entice consumers to purchase their game, then the publisher will inevitably lose potential profit. Along with the risk associated with marketing, there is a risk that developers will “slip” or fail to meet their projected completion date for games. This is a problem for publishers because, if not warned enough in advance, they could launch a costly marketing campaign prematurely. An early marketing campaign for a game could cause consumers to lose excitement for the release of the game, resulting in a loss of potential sales. The financial risks that publishers face will continue to grow as the need to catch the interest of consumers and meet the expectations of consumers continue to increase. All together these costs could add up to millions of dollars for a single game title; therefore, publishers do the most they can to make sure their games are profitable for themselves and the developers they work with.

**2.3 The Current State of the Industry**

The state of the gaming industry evolves with innovations in hardware and software development, as well as with changes in consumer tastes and behavior. Generally, technological innovations in game development causes games to improve over time. As games improve, graphically and mechanically, consumer expectations for future games will continue to rise, as well. As a result, publishers and developers will try to satisfy these expectations, which causes the cost for game development to increase. This development in the gaming industry caused costs for games from the largest publishers to average 60-70 million dollars, which is ten times as expensive as the decade prior. Indirectly, these development costs affect the way games are made as publishers are forced to mitigate the financial risks of an unsuccessful game by creating games formulaically. This trend in game development creates an oversaturation of games that feel like each other. Unfortunately, some consumers feel negatively about this trend as they believe publishers and developers are only concerned with making a profit with minimum creativity or diversity.

Similarly, the availability of the internet has changed the landscape of the game industry immensely by molding the behavior and tastes of consumers. Websites like Metacritic allow consumers to be more informed about their game purchases by reading reviews from game critics and allow consumers to become critics of those games as well by posting their own reviews. Access to this information allows consumers to be more selective of their purchases, which means publishers and developers are pressured more than ever to create games that will be well reviewed. Metacritic is now so integral to the gaming industry that it even affects the publisher-developer relationship. It is now a common practice for development contracts to include clauses that tie a developer’s pay to the scores their game receives on Metacritic. As a result, developers are pressured to forego some of their ingenuity in game development for fear of a negative reception online.

**2.4 Motivation for Analysis**

The motivation for this analysis comes from a surge of news from the gaming industry regarding issues with publishers and developers having to lay off large quantities of employees due to games failing to meet sales quotas. Creating a model that can accurately predict sales figures of games could mitigate this issue. This analysis follows the point of view of game publishers and explores the impact of the decisions they make on the sales figures of their games. As previously stated, video game publishing comes with large financial risks that publishers attempt to mitigate during the time of development. While the financial risks are important to the success of new video games, this analysis will focus on features and trends of the individual games themselves to predict how well the games will sell. Because the analysis will not contain features regarding marketing or development issues, it will focus on the publisher’s decision on the developers they finance and the type of games the developers create.

**3. The Data Set**

The data set used in this analysis is “Video Games Sales with Ratings” by Rush Kirubi on kaggle.com: <https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings>. Part of the data set was originally collected by another user on Kaggle by performing a scrape of the website VGChartz. The data set was then extended to its current iteration to include more variables by performing a scrape of the website Metacritic to improve the predictability of the model. The data set is comprised of a list of video game titles before 2017 that have sold at least 10,000 copies. In total, the data set contains 16 variables of 16,718 observations with some missing values. A table of important variables and their descriptions are given below as figure 1.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Global\_Sales | Total worldwide sales. (Dependent variable) |
| Name | Name of the game. |
| Platform | Platform of the game’s release. |
| Publisher | Publisher of the game. |
| Developer | Developer of the game. |
| Year\_of\_Release | Year of the game’s release. |
| Genre | Genre of the game. |
| Rating | The Entertainment Software Rating Board (ESRB) rating of the game. |
| Critic\_Score | Aggregate score compiled by Metacritic staff. |
| User\_Score | Aggregate score by Metacritic subscribers. |

*Figure 1: List of important variables*

Other variables, such as sales from specific regions, were not included in the analysis because of their high correlation to the dependent variable, Global\_Sales. An initial concern with the data set is that it contains many missing values for some key variables, such as the ESRB rating and the scores from critics and users. Furthermore, the variable for developers contains nearly 1700 unique character strings, which could cause overfitting if included in the final model. Finally, another concern with the data set is that it does not contain specific game features that could aid in predicting sales, such as, game modes, online multiplayer functionality, and whether it has downloadable content or in-game purchases. The data set requires a fair amount of cleaning to be ready to analyze.

**4. Data Wrangling**

This data wrangling section does not include later instances where, upon exploration of the data set, it was discovered that it would be beneficial to create variables to include in the model selection process.

**4.1 Read in the data set and Identify Issues**

First, the initial data set (Video\_Games\_Data) csv file was loaded, viewed, and summarized to identify any noticeable outliers or missing values. There were a substantial number of missing values for key variables in the analysis that needed to be cleaned. Furthermore, upon inspection of a handful of observations with missing values, it was discovered that some of this data should now be available on the website it was originally scraped from. Cross-reference of Metacritic.com revealed that certain web pages did have data for the missing values.

**4.2 Replace any Missing Values**

With the issues identified, the first step was a secondary web scrape of the Metacritic website to gather as many of the missing values for the critic score and the user score variables as possible. The web scrape resulted in a csv file with the variables Name, Platform, Critic\_Score, and User\_Score, which was then loaded into the workspace (Meta\_Critic\_Data). Both data frames were joined together into a new data frame (Clean\_Video\_Games\_Data) which contained all observations and variables from the original data set and includes the updated values for the critic and user scores from the web scrape. Finally, the original critic and user score variables were replaced by the updated variables of the same name using the ifelse() function. Then, the extraneous variables from the added from the join function are removed from the data.

**4.3 Remove Remaining Missing Values**

The last issues that needed to be cleaned were the missing values. First, the user scores had a special character called “tbd” which means “to be determined” on the Metacritic website and is given to games that were not rated at the time. All instances of the character “tbd” were treated as missing values. Next, the rest of the missing values of key variables in the data set were filtered out. Filtering out the missing values using the filter() and is.na() functions from User\_Score, Critic\_Score, and Rating is enough to eliminate all the missing values from the key variables.

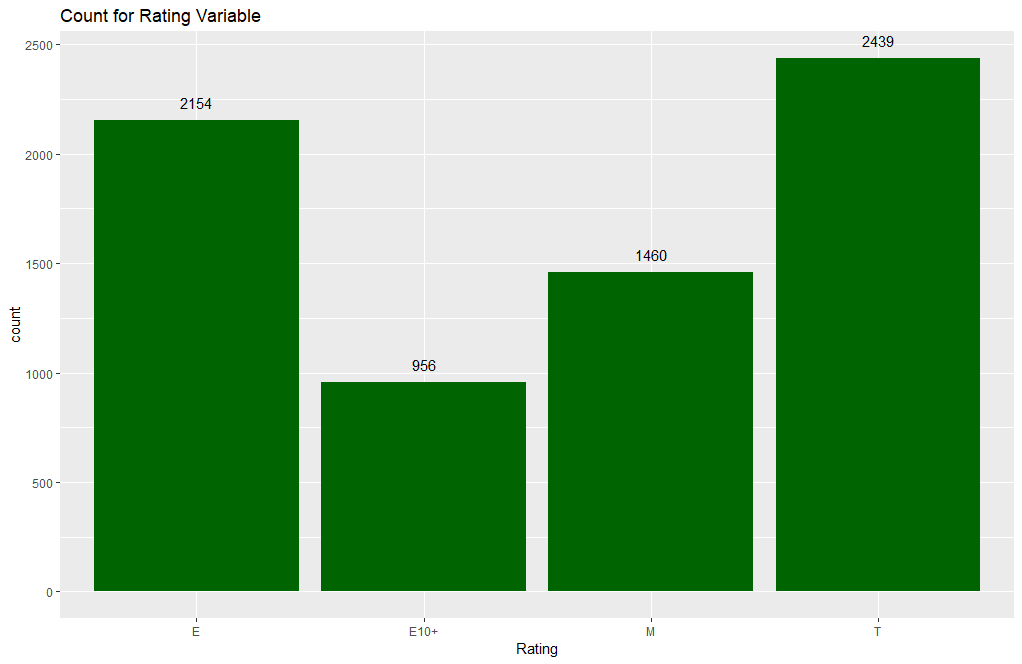
**4.4 Rename or Remove Outliers**

Observations in the rating variable with “K-A” or “AO” ratings were changed to “E” and “M” respectively due to special circumstances that concern changes in the ESRB. Also, observations in the rating variable with a “RP” rating were considered outliers and were removed from the data set. Lastly, the global sales variable had some large outliers that could affect the scope of future plots in the exploratory and regression analysis sections. The data set was filtered to exclude instances where global sales were more than 20.0, or twenty million units sold. The final cleaned data set now contains 7,009 of the original 16,718 observations.

**5. Exploratory Analysis**

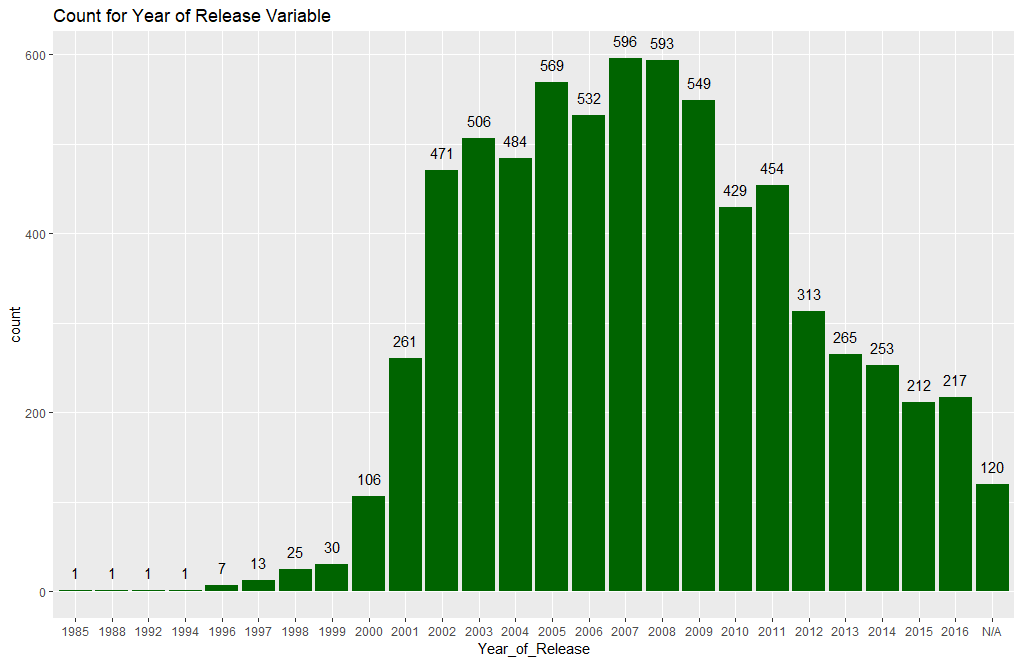
**5.1 Counts of Categorical Variables**

To start the exploratory analysis, counts of categorical variables were made through bar plots to show how the data is divided according to each. These bar plots are mostly to gain a sense for the categorical variables that could become important for predicting sales figures. The bar plot for the “Rating” variable, shown below in figure 2, shows that most games in the data set are either rated T for teens (ages 13 and up) or E for everyone (any age). This result was expected as most video games that have these ratings have the largest possible audience.



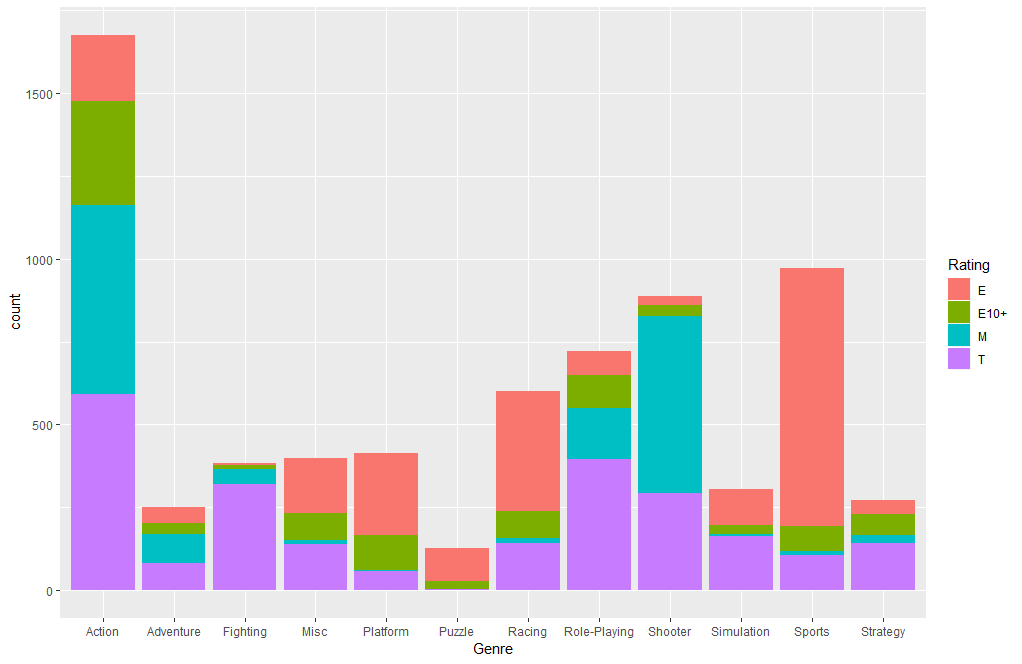
*Figure 2: Count for Rating variable*

Next, the bar plot for the “Year\_of\_Release” variable, shown below as figure 3, shows a spike of games in the early 2000’s until 2010, where it decreases. This could possibly show the expansion of the consumer base for games due to the new generation of console hardware that took place during the time, which started to excel graphically. Also, the decline from 2010 and onwards could reflect the development of consumer behavior to be more critical with their purchases based on reviews from websites like Metacritic. Another factor that could explain this decline could be the insurgence of video games on handheld devices, such as phones and tablets, which are not included in the data set.



*Figure 3: Count for year of release variable.*

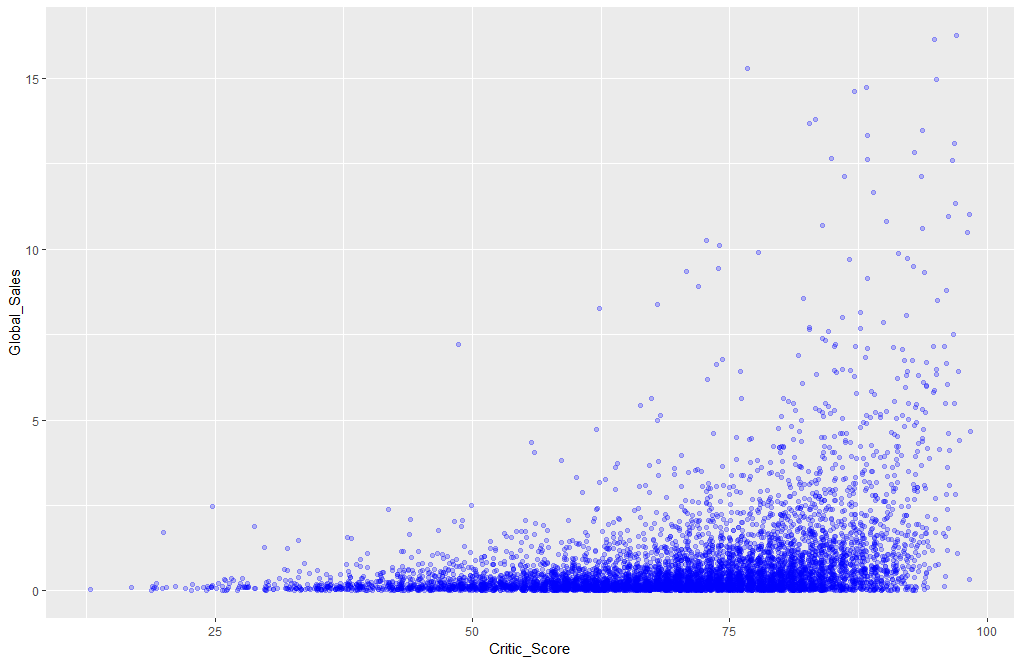
The bar plot of the “Genre” variable showed that the action genre is by far the most abundant in the data set, followed by sports, shooter and role-play. This could mean that consumers tend to buy games with a faster pace or contain a competitive online component. When this bar plot is filled by the “Rating” variable, shown below as figure 4, the ratios of each genre are shown. An interesting observation in this plot is that each genre tends to be mostly comprised of one or two ratings for most of its games. For example, 75% of the action genre is either rated T (13+) or M (18+) which are the two highest maturity ratings available in our data set. This could be an indication of what the rating for games of a certain genre are preferred and sell the most.

****

*Figure 4: Count for Genre variable filled with Rating variable*

**5.2 Plotting Global Sales by Critic Score**

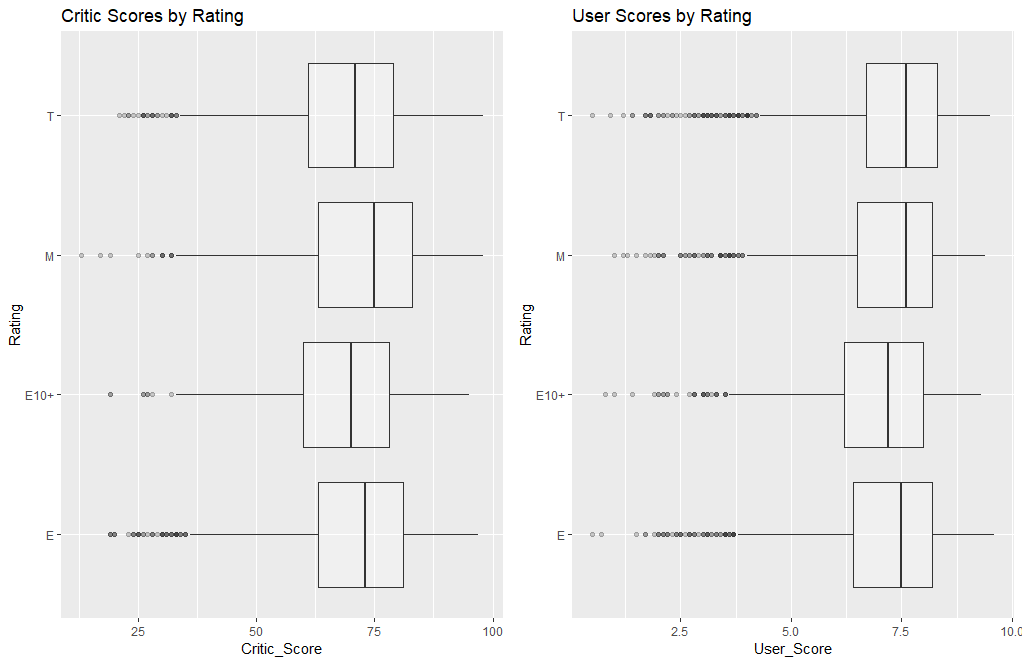
A scatterplot of the critic scores, which give the scores of professional video game critics out of 100, and global sales is produced using the ggplot, shown below as figure 5. The plot shows that a clear majority of observations are well below 2 million units sold in global sales, however, the observations that are above 2 million in global sales are also mostly all above a critic score of 60. This observation suggests that games with a higher critic score tend to sell more copies, which was expected.



*Figure 5: Scatterplot of global sales and critic scores*

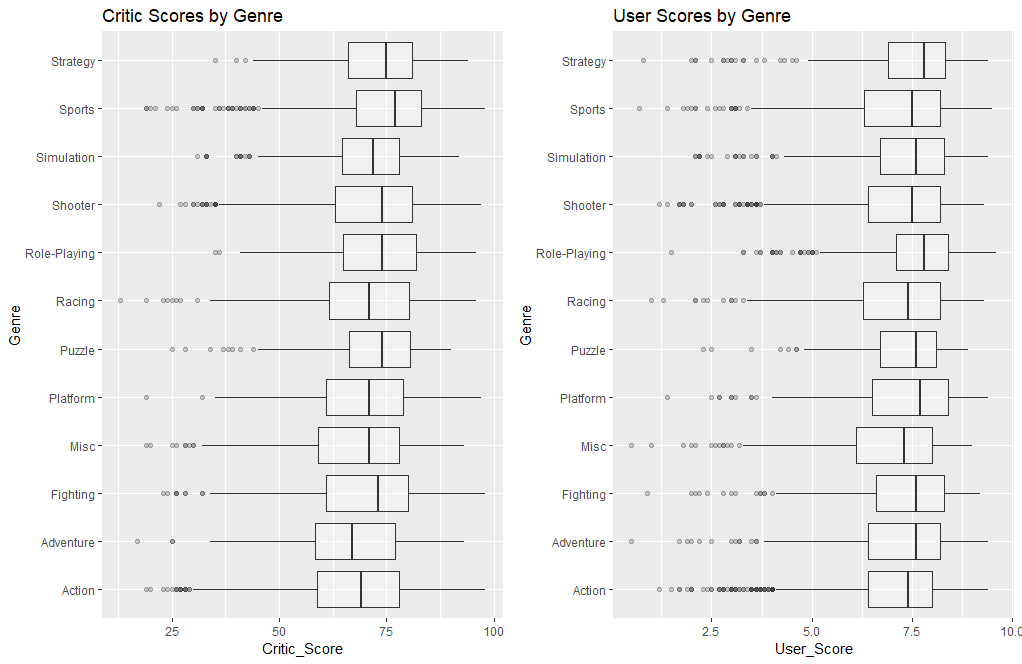
**5.3 Boxplots of Critic and User Scores**

Two boxplots were made using “Rating” with “Critic\_Score” and “User\_Score” to determine the spread of the data, shown below as figure 6. Both boxplots show that critics and users tend to score games most games between 60 and 80 out of 100, with the medians being near 70 to 75 for all ratings. This can help publishers establish a sense of how well games are scoring on Metacritic and create target scores to reach to stand out among most games.



*Figure 6: Boxplots of Critic Scores and User Scores by Rating*

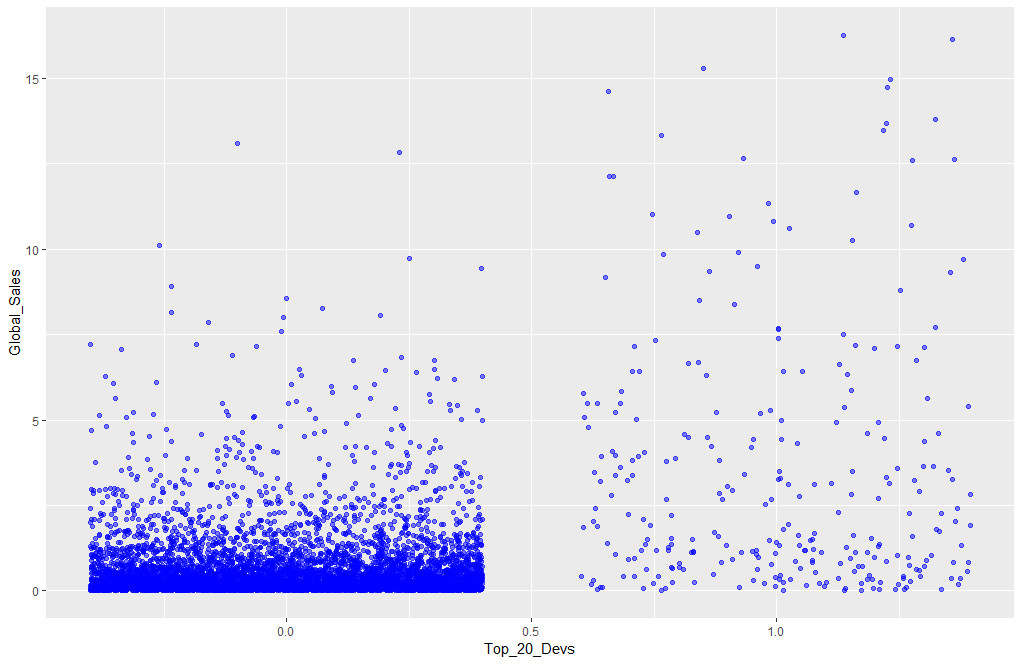
Similarly, two more boxplots were made using critic and user scores to measure the scores by genre, shown below as figure 7. The first boxplot shows that the three highest scoring genres for critics are sports, strategy, and puzzle, according to the medians. The second boxplot shows that the three highest scoring genres for users are role-play, strategy, and platform, according to the medians. These plots show that while genres such as puzzle, strategy, and platform are well reviewed, their sales aren’t proportional. Interestingly, the action genre scored relatively low in both plots although it is the most prevalent genre in the data set by far. This could suggest that many gamers are predisposed to purchasing games with a fast pace, even if it isn’t well reviewed.



*Figure 7: Boxplots of Critic Scores and User Scores by Genre*

**5.4 Create a Dummy Variable for the Top 20 Developers**

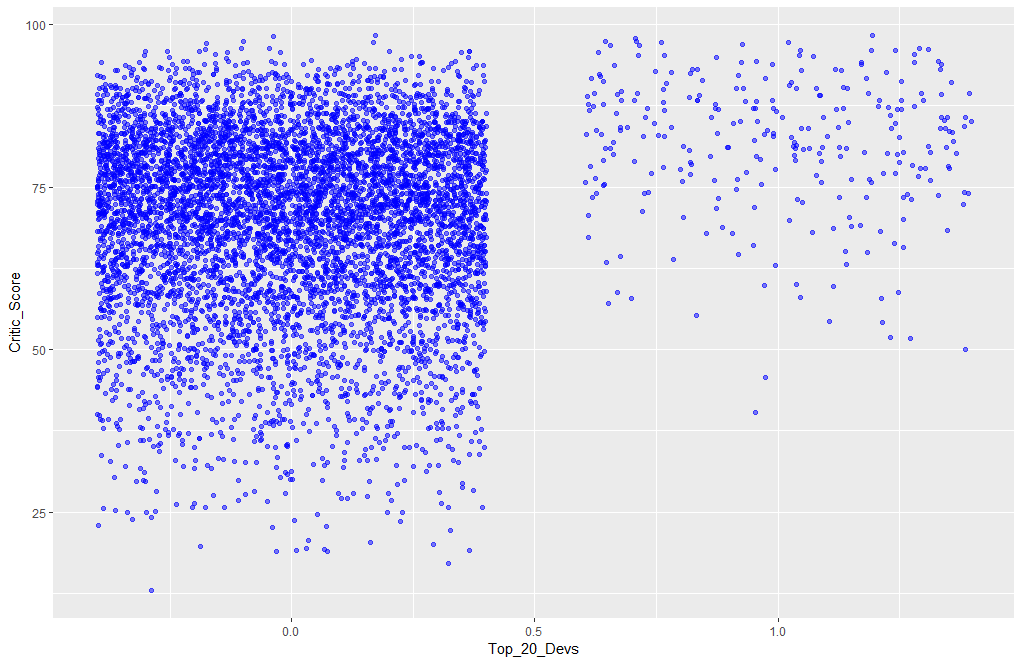
It is important to explore developer trends and records because producers have a vested interest in knowing whether to finance or hire a developer. To start, a dummy variable was created to indicates whether a game was developed by a studio that is within the top 20 developers according to average global sales of their games. These top 20 best-selling developers only account for 4% of all games represented in the data set. A scatterplot is made using global sales and critic scores with the dummy variable as the color aesthetic. The scatterplot shows that most of the observations for developers the top 20 have at least a critic score of 50 and are among the observations with the highest global sales. To further this point, another scatter plot is made with the global sales and the top 20 developers dummy variable, shown below as figure 8. While the observations in the top 20 developers are dispersed thoroughly among global sales, observations from other developers seems to be clustered at or below the 1 million copies mark (a mild to moderate success for game sales). A count reveals that 68% of games by developers in the top 20 sells at least one million copies, while only 16% of games made by other developers do the same. This result is expected as the dummy variable was made up of the developers whose games sold more copies on average. Also, this plot could mean that most of the most successful games are made by a handful of large developers and that there are many small or newly established developers in the data set.



*Figure 8: Global sales of the top 20 developers (right) and all other developers (left)*

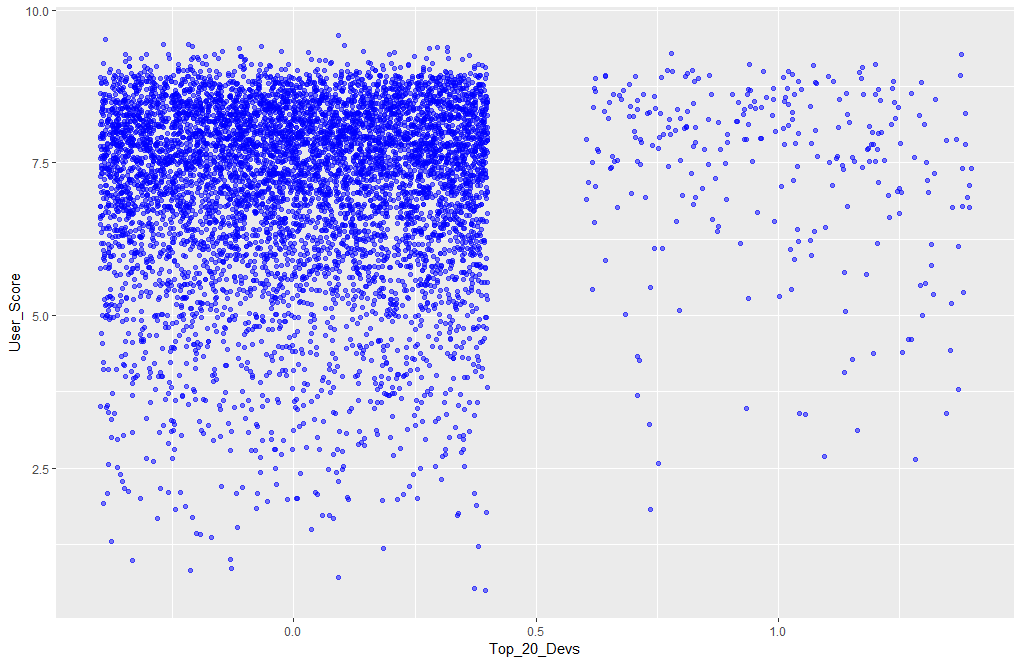
**5.5 Critic and User Scores of the Top 20 Developers**

The top 20 developers dummy variable can be used to show the difference in critical reception between the two groups of developers. A critic score of 80 or more (or user score of 8.0 or more) is considered a well-reviewed game by industry standards. A scatterplot is made using the top 20 developers and critic scores variables, shown as figure 9. The plot shows that more than half of the observations from the group in the top 20 developers are over a score of 80. Performing a count reveals that 64% of games by developers in the top 20 score an 80 or above with critics, while only 26% of games made by other developers score the same.



*Figure 9: Critic scores of the top 20 developers (right) and all other developers (left)*

Similarly, a scatterplot using the top 20 developers and user score variables, shown below as figure 10, shows a more varied spread of points. Performing a count reveals that 46% of games by developers in the top 20 score an 8.0 or above with users, while only 34% of games by other developers score the same. This a very interesting observation because users are more critical of games by the top 20 developers, while critics are more critical of games by other developers.



*Figure 10: User scores of the top 20 developers (right) and all other developers (left)*

**5.6 Create Other Variables**

The final section of the exploratory analysis includes the addition of other variables to the data set that could have some explanatory power in the final model. First, the developer variable had instances where more than a single developer name was listed for an observation, therefore, a co-developer variable was created using the separate() function. Separating the developer variable in this way reduced the number of unique developer names from about 1,700 to 1,119 and highlights the lead developers of each game. Also, the co-developer variable needed to reassign all its missing values to “None” so that the variable could be included in the model selection process in the future. Next, two variables were created that show counts for both the developer and co-developer variables. If included in the final model, these count variables could be used to indicate whether having more or fewer game titles as a developer could improve sales of a game. Finally, the variable Mod\_Developer was created as a mutation from the developer variable. Mod\_Developer shows the names of developers who have at least 10 game titles in the data set while displaying “Small/New Developer” if a developer has less than 10. Indicating which developers are newer/smaller studios greatly reduces the number of unique developer names from 1,119 to 117. For the remainder of this analysis, the Mod\_Developer variable will be used instead of the original Developer variable. Figure 11 below shows the list of variables that were added to the original data set and their descriptions. At the end of the exploratory analysis, the data set contains 7,009 observations of 21 variables.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Top\_20\_Devs | The top 20 developers according to highest average sales. |
| Co\_Developer | Developer that assisted the development of the games. |
| Dev\_Count | The number of each developer in the data set. |
| Co\_Dev\_Count | The number of each co-developer in the data set. |
| Mod\_Developer | Developer of the game (indicates “Small/New Developer”). |

*Figure 11: List of variables added to the data set.*

**6 Regression Analysis**

To start the regression analysis the data set was split into two groups with 80% of the data set becoming the training set and the remaining 20% becoming the test set.

**6.1 Model Selection**

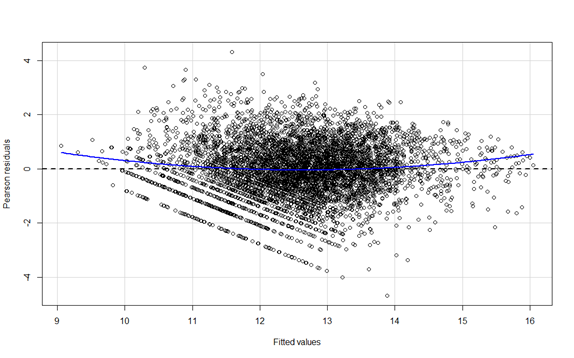
To predict global sales, a log-linear model was created through backward selection of 10 possible independent variables. A log-linear regression analysis was chosen because the outcome of the dependent variable was right-skewed. By taking the log of the dependent variable, observations became more symmetrically distributed. Also, the dependent variable, which was in decimal units, was changed to the integer value for the actual amount of games sold by multiplying the variable by a million. This improves interpretability further as the actual and predicted log values are not small decimal or negative values. The model was selected using an automatic selection process. The step() function selects the best possible model using an iterative process that removes variables, from those provided, from the function until an optimal model is chosen based on the Akaike Information Criterion (AIC). From the 10 independent variables possible, 8 were included in the final model selected through this process. Although the backward selection process chose the best model based on AIC, a common concern with a model containing these many variables is overfitting. For comparison, 5 other models were created using fewer independent variables. The model chosen through the automatic backward selection process had the highest adjusted R-squared value, shown in the table below as figure 12. Although the “Backward\_mod” model contains the most variables, it has the highest adjusted R-squared value and the lowest AIC value.

|  |  |  |
| --- | --- | --- |
| **Model** | **Variables** | **Adjusted R2** |
| Backward\_mod | Genre + Critic\_Score + User\_Score + Mod\_Developer + Rating + Top\_20\_Devs + Dev\_Count + Platform | 0.4623 |
| Games\_Mod | Genre + Critic\_Score + User\_Score + Mod\_Developer + Rating + Top\_20\_Devs + Dev\_Count | 0.2967 |
| Games\_Mod2 | Genre + Critic\_Score + User\_Score + Mod\_Developer + Rating + Top\_20\_Devs | 0.2931 |
| Games\_Mod3 | Genre + Critic\_Score + User\_Score + Rating + Top\_20\_Devs | 0.2305 |
| Games\_Mod4 | Genre + Critic\_Score + User\_Score + Mod\_Developer + Rating | 0.2656 |
| Games\_Mod5 | Genre + Critic\_Score + User\_Score + Mod\_Developer + Top\_20\_Devs | 0.2902 |

*Figure 12: Summary of regression model analysis.*

**6.2 Plot the Residuals against the Fitted Values**

While the AIC is a good estimator for the model selection process, the chosen model must be validated to assure its quality. Figure 13 shows the plot of the model’s residuals against the fitted values. The plot shows that the residuals appear to be dispersed about the range of fitted values and centered around zero with no clear patterns. This means that the residuals are randomly distributed, as they should be.

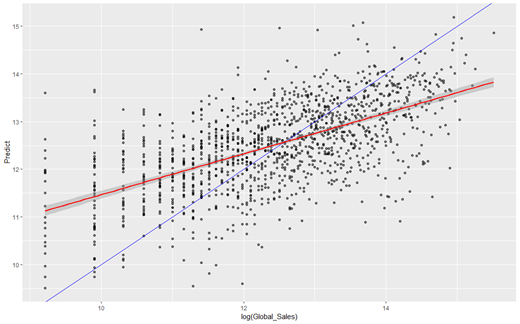


*Figure 13: Plot of the residuals against the fitted values.*

**6.3 Make Predictions and Analyze Plots**

Predictions for global sales are made using the predict() function with the model on the test set.

Figure 14 shows the plot of the global sales against the global sales predictions for those games in log of units sold.



*Figure 14: Predictions of global sales.*

Figure 14 shows two lines: one where observations from the test set are predicted accurately according to the actual sales data (blue line), and the other shows the line of the predictive model (red line). Based on this graph, it appears as though the predictive model, on average, tends to over- predict the global sales of games that sold below the log value 12.5, or below 268,000 units sold. This is shown in figure 14 when the line for the predictive model is above the accurate blue line. Conversely, it appears that the predictive model, on average, tends to under-predict the global sales of games that sold over the log values of 12.5. This is shown in figure 14 when the line for the predictive model is below the accurate blue line.

**6.4 Examples of Prediction**

The first example from the data set is a game that many would presume to sell well. The criteria for this example are that the game is from the most prominent genre, from the most prominent ESRB rating, and that the developer has created more than 40 titles in the data set. Therefore, the example will be an action game that is rated T for teens from a well-established developer. The title chosen was One Piece: Pirate Warriors 3. In this example, the model predicted that it would sell about 274,000 units, while it sold about 460,000. This prediction follows the previous statement that games that sold more than 268,000 units tend to be under-predicted by the model.

Interestingly, this result might have been anticipated further due to other factors that were not featured in the model. These other factors are that the game is based on a popular media franchise and that it is a sequel. Both factors tend to mean greater sales numbers for games as they have a previously established fanbase.

The second example is a game that many would presume to sell far less than most. The criteria for this example are that the game is from the least prominent genre, from the least prominent ESRB rating, and that the developer has created less than 10 titles in the data set. Therefore, the example will be a puzzle game that is rated E10+ from a relatively new developer. The title chosen was Puzzle Quest 2. In this example, the model predicted that it would sell about 143,000 units, while it sold 110,000. This prediction follows the previous statement games that sold less than 268,000 units tend to be over-predicted by the model.

The third example is a game that is between both extremes of the previous examples according to its prominence in the data set. The criteria for this example are that the game is from the seventh most abundant genre, from the second least prominent ESRB rating, and that the developer has created between 11 and 39 titles in the data set. Therefore, the example will be a game that is rated M for mature from a relatively established developer. The title chosen was Rare Replay. In this example, the model predicted that it would sell about 967,000 units, while it sold 790,000. The model over-predicted the global sales for this example, which did not follow previous assumptions for games that sold over 268,000 units. It is possible that games with less prominent attributes are more difficult for the model to predict.

The model struggled with accurately predicting any particular type of game as there are many factors to consider. Games that the model predicted accurately contained genres and ratings of each type and of various combinations, so it was difficult to determine a particular kind of game the model could accurately predict. Unfortunately, this means the model needs more comprehensive research to improve predictability of game sales.

**7.0 Conclusions and Recommendations**

One of the most important factors in the model to predict global sales of video games were the critic scores for games on Metacritic. As the gaming industry shifts toward being more competitive and expensive than it’s ever been, consumers will continue to rely on information found online to determine which games are worth their time and money to buy. Based on this analysis, publishers should:

* Establish or improve exclusive relationships with top third-party developers. The analysis shows that having a game from a top developer tends to increase sales dramatically, however, publishers do not have access to every developer. By developing exclusivity with well-established third-party developers, publishers gain a competitive edge as consumers tend to buy more games made by developers with a good record, according to the analysis.
* Set target critic score goals to above an 80. The analysis shows that games with a higher critic score, tend to sell more units. Also, critic scores on Metacritic are considered generally well-reviewed starting at a score of 75, with “universally acclaimed” scores begin at 90.
* Consider publishing games that are for mature audiences. The analysis shows that of all the ESRB ratings, the M (mature) rating had the most positive impact of global sales. This could suggest that consumers prefer games with more mature themes such as violence or horror. This could also suggest that games with more mature themes are of better quality than many other games with wider audiences.

Creating models to predict consumer behavior in an industry that demands constant change requires more comprehensive research. Further analysis using other features of video games, such as: game modes, online multiplayer functionality, downloadable content, and in-game monetization methods could prove to be important factors in predicting video game sales moving forward. Moreover, other factors relating to publisher costs, such as, a ratio between the cost of development and marketing, could improve the quality of the predictive model of video game sales.