Factors that Impact Video Game Sales

This paper explores trends in past video game sales by fitting a linear regression. It is meant to give a general idea of what factors may be contributing to higher sales, which could lead to a better understanding of the market overall. Specifically, it looks for what kind of variables had impact on past global sales and if those influential variables change when the focus is shifted to sales of three specific regions. It also looks at video game that did outstandingly well in one particular region but not as well in the others, and looks for influential variables that stand out and are different. North America and Europe tend to have more similar interests than Japan in both genre and gaming platforms. In addition, for the analyses of games popular in certain regions but not in others, shooting, sports and role-playing games were found to be popular in North America, Europe and Japan respectively. Results are contrasted to sales over time.

I looked at trends in video game sales to see what variables had impacts on past sales, and if they changed when the data was divided up into regional sales. I also looked at games that did outstandingly well in one region in comparison to the others, and if there were specific variables that stand out. The data included information on the video game's Title, Platform, Year of Release, Genre, Publisher, Sales in North America, Europe, Japan and "Other" regions in addition to Global Sales (in millions), Metacritic Critic Scores, Number of Critics giving scores, Metacritic User Scores, Number of Users giving scores, Developer and ESRB Rating. From the original data set, I omitted any rows with missing values, which gave me 6,947 observations. In the future, I would look into the whole data set, as taking out these observations simply due to missing data introduces a bias, especially taking out older and unpopular games.

I fit a linear regression to see the impacts of different variables in global sales. In fitting the model, I took out Name, Publisher and Developer because there were too many levels and made the model too complicated to interpret. To ensure best fit, I transformed variables and took out insignificant (p-value > 0.05) variables and refitted the model. As a result, I got a model that explained about 45% of the variability. I also checked the conditions necessary for this model (linearity, independence, normality and equal variance) and the model seemed to be a pretty good fit despite a few outliers and some minor concerns in equal variance. However, I proceed with caution, keeping the limitations in mind. I fit linear regression models for *North American, European* and *Japanese* sales as well, again checking conditions (which had very similar results as the above) and proceeding with caution. These models explained about 34% to 38% of the variability.

Figure 1 shows the significance of variables across regions ("Global" will be referred to as a region for simplicity). Platform, Genre, Critic Score, Critic Number, User Score, User Count and Rating were all significant. Higher values in Critic Score, Critic Count and User Count all had positive effects on sales across regions. However, Year of Release was only significant for video games in *North America* and *Japan*; in both regions, the more recently the game was released, the fewer in sales it had. Finally, the trends seen for User Scores came as a surprised because higher user scores had a negative effect for *Global*, *North American* and *European* sales although it had a positive effect for *Japanese* sales. This would require further investigation, as usually one would expect higher scores to lead to more sales. There could be a confounding variable or perhaps a distrust in user scores by the community.

For the effects of different Platforms in **Figure 2**, there were more differences seen between regions than in **Figure 1.** The reference group for this variable was 3DS. PC, PS Vita and Xbox were all negative in comparison to the 3DS across all regions. Interestingly, every significant platform for Japan had a negative impact – including the Wii, which had a positive impact for all other regions. Otherwise, the PlayStation series seemed to be popular globally overall and in Europe. In looking at this outcome, it wasn't surprising that the impacts of other platforms on Japanese sales were negative given the popularity of the reference group, 3DS, in Japan. On the other hand, the North American and European market seem to favor video games that are not handheld (e.g. Wii, PlayStation, etc.).

The significance and effects of Genre also varied between regions as shown in **Figure 3.** Both North America and Europe found higher sales for Miscellaneous games, which is not as helpful of a factor given that they can cover a wide-range of video games; Japan found higher sales for Fighting and Role-Playing. Keeping in mind that the reference group is Action, a relatively popular genre, it makes sense that many other genres have negative impacts in comparison to the reference group.

In comparing the three regions, *North America* and *Europe* seem to follow more similar trends than *Japan*. This makes sense given the sizes of the regions analyzed. (The difference in the regions compared is an aspect that should be looked at in the future, given that it is not as ideal to compare such differently-sized regions (especially between Japan and the two other regions). *North America* and *Europe* may also be more culturally similar. Thinking about these

similarities and differences, however, I next explore what characteristics video games that did exceptionally well in certain regions had.

On a side note, I thought it was interesting that Year of Release had a negative impact on sales, so more recent games have fewer sales than in earlier years. Since a time series is much more complicated and usually does not have a linear trend across time, I plotted the time series of video game sales in **Figure 4.** As suspected, the number of sales doesn't go down the entire time from 1985 to 2016, and all three regions' sales seem to follow a very similar trend despite some differences, with the largest peak in video game sales happening in 1996 which, for example, was when the first of the *Pokémon* series was released.

In answering the second set of questions about sales that did outstandingly well in one region against the rest and specific variables that stand out, I normalized the data with the equation, [Sales-min(Sales))/(max(Sales)-min(Sales)], to compare values by performance within their own region. Next, I created a new variable that indicated which country had the greatest normalized sales value for each row. Finally, I created three new target data sets called "target_NA", "target_EU" and "target_JP", in which each game's normalized sales were compared across regions and the observation was added to the "target_" data set of the region with the highest normalized score. Then, I further filtered each "target_" data set for cases in which the target region's normalized sales had a difference of more than 0.1 with at least one other region's normalized sales. (e.g. in "target_NA", an observation in which North America had the highest normalized value and had a normalized value that was at least 0.1 greater than Europe or Japan's normalized value for that video game will be included.) This enabled me to find the video games in each region that did exceptionally well in that particular region when it wasn't the case with the others. Therefore, an internationally popular game will be expected to have high sales throughout the three regions, so it will likely not make it into any of the "target_" data sets because the normalized sales values will be similarly high throughout the three regions. However, a game that is very specific to a certain region's tastes may have extremely high normalized values in one region but not as high in the rest. In looking at the new "target_" data sets, all of them contain multiple games from the same series (e.g. Call of Duty). Since the popularity of one game can lead to sales in the next in the series, this may impact results.

Figure 5 uses the "target_" data set to compare the popular games' genres by region. Looking at each region's distribution, there are some trends. Role-Playing seems to be very popular in *Japan*. This seems to be the case for Sports in *Europe* and Shooter in *North America*.

In **Figure 6**'s breakdown of each region's popular game publishers, the fact that Nintendo is the most popular for *Japan* is not surprising, given the earlier finding about the popularity of 3DS games in Japan. There is also a clear divide by region. For example, *North America* and *Europe* share some bars, but when *Japan* has a considerable portion of a bar, the other two regions are non-existent. This may be due to issues not only related to cultural trends but also availability (e.g. if it is released in Japanese, it may not be sold in *North America* or *Europe,* whereas if it is released in English, it may be sold in both *North America* and *Europe*).

Finally, **Figure 7** shows the distribution of the games in all "target_" data sets over time. This is interesting because F**igure 4** showed a decrease in sales over time, but this graph implies that video games that did particularly well in one region have come out in more recent years, particularly for *Europe.* Although this may be due to omitting rows with missing values (which introduces a bias for newer games), could there be a correlation between decreasing sales overall but popularity in specific regions? Interestingly, *Japan* has a similar trend as **Figure 4** in which most data points are around 2010 when sales for each region peaked.

This study showed which variables might have an effect on global video game sales overall, and which platform and genre were more popular in certain regions than in others. In addition, it supported the general assumption of higher critic reviews leading to more sales, although aspects such as user reviews leading to fewer sales could be looked into more. There were also differences in taste between regions that stood out, and could be further researched.

This study should serve as a guideline in obtaining an idea of video game sales, and further research should be done before making any claims. For one, the model was fit on a data set that is only a smaller portion of a larger data set and the models used only explain around 30-50% variability. In addition, it compares two continents and a country, which is not ideal. In the future, I would work on finding other types of models which may be of better fit. Including data past 2016 and validating the model may also be helpful, as my current model is telling me that the more recent the release date is, the fewer sales there will be (but that would not be an accurate model past a certain date since a *negative* video games sales does not make sense).

In terms of data cleaning, I should also think more about how to clean data for video games that are the same but offered on different platforms depending on region. For example, one of the "Danganronpa" games was released in Japan for PSP but later in North America for PS Vita. Therefore, there are two rows, in which North America has 0 sales for PSP. It also becomes evident that the PSP "Danganronpa" game (Japan) is not in the cleaned data set because it has missing values such as reviews on Metacritic. Thus, although in the final data (i.e. PS Vita version only) North American sales are 27.3% and Japanese sales are 45.5% of the global sales, if the two rows were to be combined since they are the same game, sales would be 15.8% and 68.4% respectively. That is an enormous difference! A dilemma to combine these two rows, however, would be the difference in platforms.

In addition, the way I created the "target_" data sets should also be reconsidered. The sample sizes varied widely across the regions, with Japan having the highest sample size. This is most likely the case because *Japan* had a small range in sales to begin with (6.5 million in comparison to North America's 41.36), so normalizing the score may have not been the best method. This is because I was unable to detect data points in which certain region's sales were 0 because the video game was unavailable in that region instead of when sales were actually 0 despite availability. This led to all regions' minimum sales values to be 0 by default, even though the minimum sales value effects normalized scores greatly and not having an accurate one would impact the normalized values. This should be further considered in the future. Additionally, if I had more time to work on this project, I would choose a better method in determining the difference in normalized sales with the other regions (i.e. 0.1), instead of merely choosing what I thought was an optimal difference from looking at the different sample sizes.

This time I only spoke about the positive and negative impacts of different aspects of video games for the sake of simplicity; however, I could also explore the degree to which these sales are impacted (e.g. per every one value increase in critic scores, critic count, user scores and user count, *Global* sales are expected to increase by 30 thousand, 20 thousand, -90 thousand and 450 games respectively (**Figure 8**)). Therefore, keeping in mind that critic scores are on a scale of 100 and user scores are on a scale of 10, the impact to which a ten point increase in critic scores would have on sales is a higher degree than a one point increase in user scores). I could also look into the individual categories (e.g. platforms) and explore questions specific to it. For example, for titles in which the same game is offered on two different platforms, to what degree do sales change depending on the platform? "Grand Theft Auto V" is interesting in that the same game for *Xbox 360* was more popular in *North America* whereas *PS3* was for *Europe* (even though the game was available for both platforms in the two regions). Could this be representative of a platform that is more popular in one region over another? In fact, earlier findings did say that PS3 had a positive impact on European sales in **Figure 2,** whereas it was insignificant for North America. Finally, earlier the Sports genre seemed to have a negative impact on sales (**Figure 3**) but when games that did particularly well in Europe were looked at, a large portion was Sports. Could this mean that specific games are doing very well but the genre overall isn't? What could be distinguishing them?

Further research and model improvement could lead to a better understanding of video game sales, which could be used to not only predict how well a game may sell, but also help in deciding key factors of creating new games to fit the wants of the current market.

Appendix.

Variable		Global	N. America	Europe	Japan
Name	*unique names: 6493*	n/a	n/a	n/a	n/a
Platform	*levels: 17*	√~	√~	√~	√~
Year of Release	*range: 1985-2016*	X	√-	X	√-
Genre	*levels: 12*	√~	√~	√~	√~
Publisher	*levels: 263*	n/a	n/a	n/a	n/a
Sales in North America in millions	*range: 0-41.36*	n/a	n/a	n/a	n/a
Sales in Europe in millions	*range: 0-28.96*	n/a	n/a	n/a	n/a
Sales in Japan in millions	*range: 0-6.5*	n/a	n/a	n/a	n/a
Sales in Other Regions in millions	*range: 0-10.57*	n/a	n/a	n/a	n/a
Global Sales in millions	*range: 0.01-82.53*	n/a	n/a	n/a	n/a
Critic Score Aggregation by Metacritic Staff	*scale: 0-100*	√+	√+	√+	√+
Number of Critics giving Scores	*range: 3-113*	√+	√+	√+	√+
User Scores by Metacritic's subscribers	*scale: 0-10*	√-	√-	√-	√+
Number of Users giving Scores	*range: 4-10665*	√+	√+	√+	√+
Developer	*levels: 1297*	n/a	n/a	n/a	n/a
ESRB Rating	*levels: 7*	√~	√~	√~	√~

Figure 1. Significance of Variables by Region

** Irrelevant or omitted variables are highlighted in gray as they are not part of the model.*

** "✓" represents significance for that variable, and the symbol after it represents the effect ("+" for positive, "-" for negative, "~" for significance in only certain levels). "X" represents insignificance.*

Platform	Global	N. America	Europe	Japan
3DS	reference	reference	reference	reference
Dreamcast	X	X	X	X
DS	X	X	√-	X
Gameboy Advanced	X	X	X	X
Game Cube	X	X	√-	X
PC	√-	√-	√-	√-
PS	√+	X	√+	X
PS2	√+	X	√+	X
PS3	√+	X	√+	√-
PS4	√-	√-	X	√-
PSP	X	√-	X	√-
PS Vita	√-	√-	√-	√-
Wii	√+	√+	√+	√-
WiiU	√-	X	X	√-
Xbox 360	X	X	X	√-
Xbox	√-	√-	√-	√-
Xbox One	X	X	X	√-

Figure 2. Significant Levels of Platforms by Region (3DS is the reference level.)

** "✓" represents significance for that variable, and the symbol after it represents the effect ("+" for positive, "-" for negative, "~" for significance in only certain levels). "X" represents insignificance.*

Genre	Global	N. America	Europe	Japan
Action	reference	reference	reference	reference
Adventure	✓-	✓-	✓-	✗
Fighting	✗	✗	✗	✓+
Misc	✓+	✓+	✓+	✗
Platform	✗	✗	✗	✓-
Puzzle	✓-	✓-	✓-	✗
Racing	✓-	✓-	✗	✓-
Role-Playing	✓-	✓-	✓-	✓+
Shooter	✗	✗	✗	✓-
Simulation	✓+	✗	✗	✗
Sports	✗	✗	✓-	✓-
Strategy	✓-	✓-	✓-	✓-

Figure 3. Significant Levels of Genre by Region (Action is the reference level.)
** "✓" represents significance for that variable, and the symbol after it represents the effect ("+" for positive, "-" for negative, "~" for significance in only certain levels). "✗" represents insignificance.*

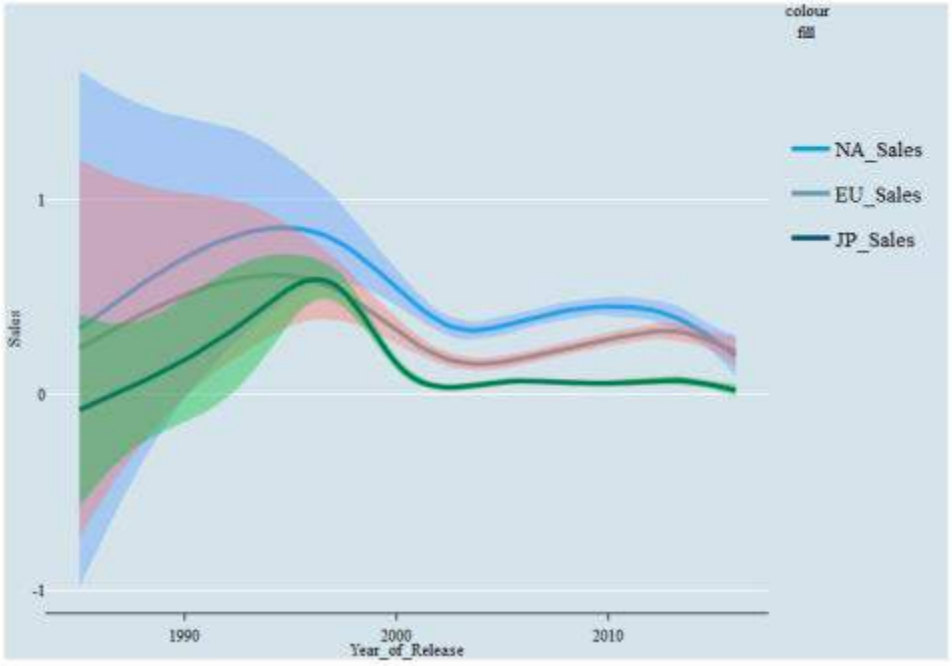

Figure 4. Time Series of Video Game Sales 1985-2016 by Region. Click here for the interactive version!

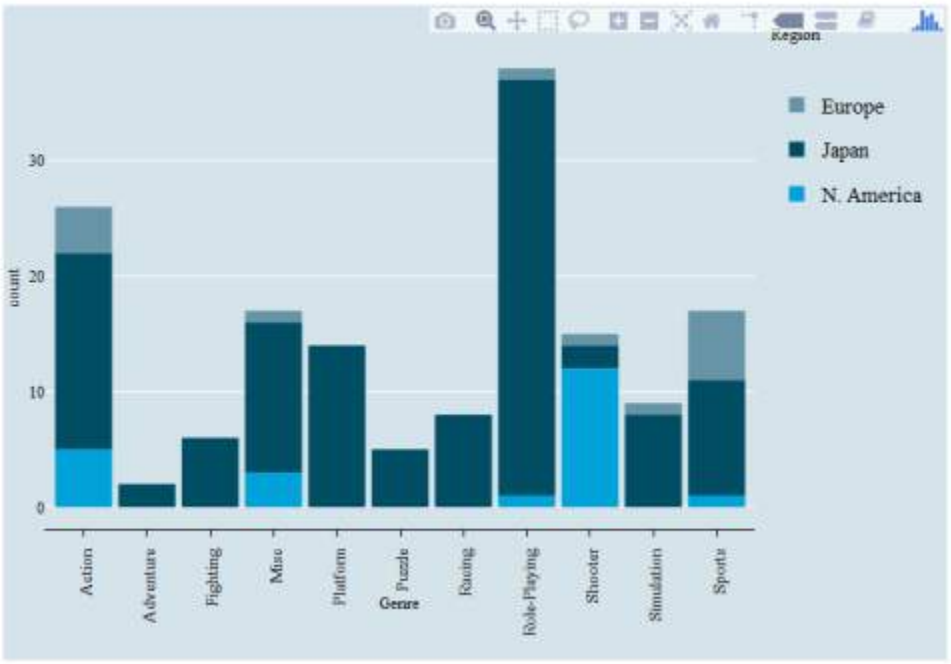

Figure 5. Breakdown of Top Video Game Sales in each Region by Genre. Click here for the interactive version!

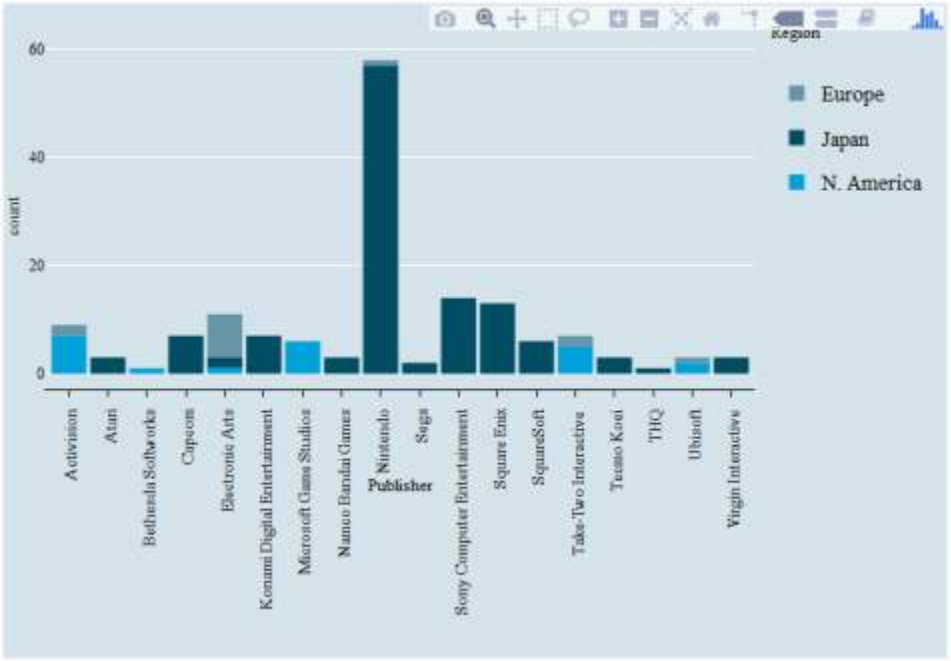

Figure 6. Breakdown of Top Video Game Sales in each Region by Publisher. Click here for the interactive version!

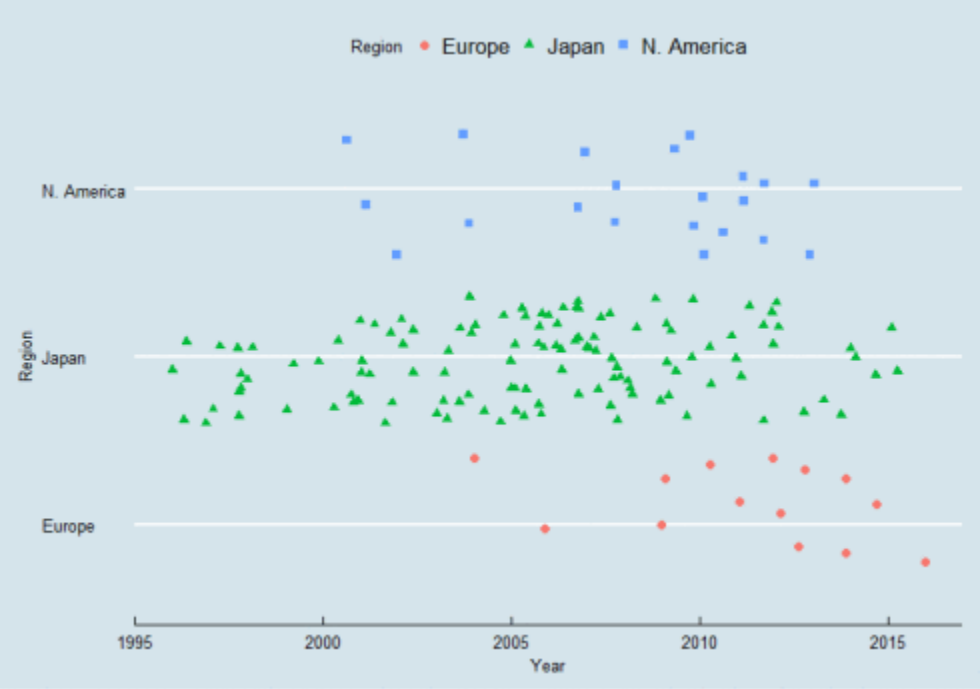

Figure 7. Distribution of Top Video Game Sales in each Region by Year of Release.
Note: **Figure 7's vertical position of points are caused by jitter and are not representative of the number of sales or any value other than the representative region around the white line.*

```
summary(m2)
```

```
##
## Call:
## lm(formula = Global_Sales_log ~ . - Name - Publisher - Developer -
##     Global_Sales - JP_Sales - EU_Sales - NA_Sales - Other_Sales -
##     Year_of_Release, data = games)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4132 -0.6645  0.0125  0.6770  4.1393
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.458e+00  1.046e+00  -2.350 0.018817 *
## PlatformDC        -4.299e-01  2.918e-01  -1.473 0.140787
## PlatformDS         1.231e-01  9.700e-02   1.269 0.204522
## PlatformGBA        2.825e-02  1.094e-01   0.258 0.796275
## PlatformGC        -1.662e-01  1.020e-01  -1.629 0.103341
## PlatformPC        -1.691e+00  9.698e-02 -17.440  < 2e-16 ***
## PlatformPS         9.295e-01  1.223e-01   7.603 3.29e-14 ***
## PlatformPS2        3.940e-01  9.090e-02   4.334 1.48e-05 ***
## PlatformPS3        2.555e-01  9.290e-02   2.750 0.005969 **
## PlatformPS4       -5.344e-01  1.086e-01  -4.922 8.79e-07 ***
## PlatformPSP       -6.909e-02  9.963e-02  -0.693 0.488052
## PlatformPSV       -4.933e-01  1.279e-01  -3.858 0.000115 ***
## PlatformWii        5.340e-01  9.664e-02   5.525 3.41e-08 ***
## PlatformWiiU      -2.850e-01  1.384e-01  -2.059 0.039577 *
## PlatformX360      -2.710e-02  9.250e-02  -0.293 0.769581
## PlatformXB        -4.642e-01  9.656e-02  -4.807 1.56e-06 ***
## PlatformXOne      -1.288e-01  1.194e-01  -1.079 0.280545
## GenreAdventure    -4.986e-01  7.116e-02  -7.006 2.68e-12 ***
## GenreFighting      1.021e-02  6.105e-02   0.167 0.867122
## GenreMisc          2.016e-01  6.116e-02   3.296 0.000985 ***
## GenrePlatform     -7.368e-02  6.151e-02  -1.198 0.231022
## GenrePuzzle       -6.846e-01  1.036e-01  -6.608 4.18e-11 ***
## GenreRacing       -1.105e-01  5.450e-02  -2.027 0.042662 *
## GenreRole-Playing -2.713e-01  4.803e-02  -5.649 1.68e-08 ***
## GenreShooter      -8.283e-02  4.526e-02  -1.830 0.067321 .
## GenreSimulation    2.194e-01  6.792e-02   3.231 0.001241 **
## GenreSports       -1.470e-02  5.166e-02  -0.284 0.776066
## GenreStrategy     -5.744e-01  7.156e-02  -8.027 1.16e-15 ***
## Critic_Score       3.178e-02  1.290e-03  24.633  < 2e-16 ***
## Critic_Count       2.274e-02  8.720e-04  26.077  < 2e-16 ***
## User_Score        -9.226e-02  1.134e-02  -8.137 4.79e-16 ***
## User_Count         4.523e-04  2.533e-05  17.854  < 2e-16 ***
## RatingE           -7.703e-01  1.040e+00  -0.741 0.458853
## RatingE10+        -9.131e-01  1.040e+00  -0.878 0.379966
## RatingK-A         -6.619e-01  1.472e+00  -0.450 0.652971
## RatingM           -1.061e+00  1.039e+00  -1.021 0.307348
## RatingRP          -5.169e-01  1.471e+00  -0.352 0.725209
## RatingT           -1.056e+00  1.039e+00  -1.016 0.309825
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.038 on 6788 degrees of freedom
##   (121 observations deleted due to missingness)
## Multiple R-squared:  0.453,  Adjusted R-squared:   0.45
## F-statistic:   152 on 37 and 6788 DF,  p-value: < 2.2e-16
```

Figure 8. Global Sales Linear Model Summary

References.

Kirubi, Rush. “Video Game Sales with Ratings.” *Kaggle*, Kaggle Inc., 1 Dec. 2016, www.kaggle.com/rush4ratio/video-game-sales-with-ratings.