Video Game Sales Analysis

Jorge

3/12/2019

This project attemps to quantify to what extent certain variables affect video game sales. I will be using data gathered from vgchartz, metacritic, and igdb scraped from the web using python.(The code for the scraper is available at https://github.com/Jorgelopez1992/vg_sales_regression

There were 54,853 games with data available on vgchartz.com, out of those only 4,189 games have both sales data and review scores on metacritic. Unfortunately the sales numbers are in incriments of 10,000 and it is only physical sales data, so there are quite a few limitations on the data, but nevertheless there is enough data available to establish some significant relationships. The data cleaning and manipulation code can be found at https://github.com/Jorgelopez1992/vg_sales_regression

A glimpse of the processed data shows the 10 variables I will be examining.

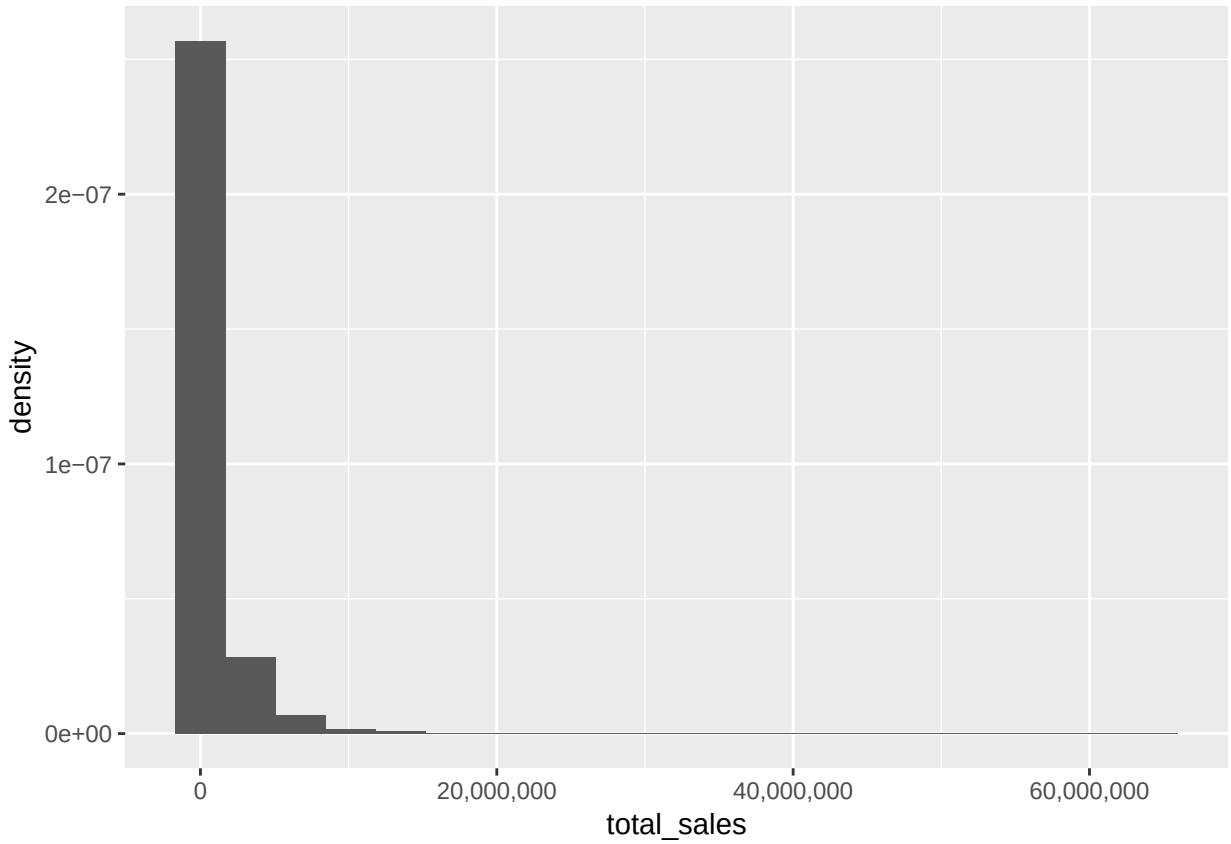
game	total_sales	release_date	genre
Grand Theft Auto V	64290000	2013-09-17	Action
Grand Theft Auto: Vice City	16190000	2002-10-27	Action
Grand Theft Auto III	13110000	2001-10-22	Action
Grand Theft Auto IV	22530000	2008-04-29	Action
Grand Theft Auto: Liberty City Stories	11260000	2005-10-24	Action

game	genre	avg_critic_score	avg_user_score	esrb_rating	multiplayer
Grand Theft Auto V	Action	97.00000	8.175000	M	yes
Grand Theft Auto: Vice City	Action	94.50000	8.800000	M	no
Grand Theft Auto III	Action	95.00000	8.400000	M	no
Grand Theft Auto IV	Action	95.33333	7.366667	M	no
Grand Theft Auto: Liberty City Stories	Action	83.00000	7.750000	M	no

game	multiplayer	number_platforms	series
Grand Theft Auto V	yes	4	yes
Grand Theft Auto: Vice City	no	2	yes
Grand Theft Auto III	no	2	yes
Grand Theft Auto IV	no	3	yes
Grand Theft Auto: Liberty City Stories	no	2	yes

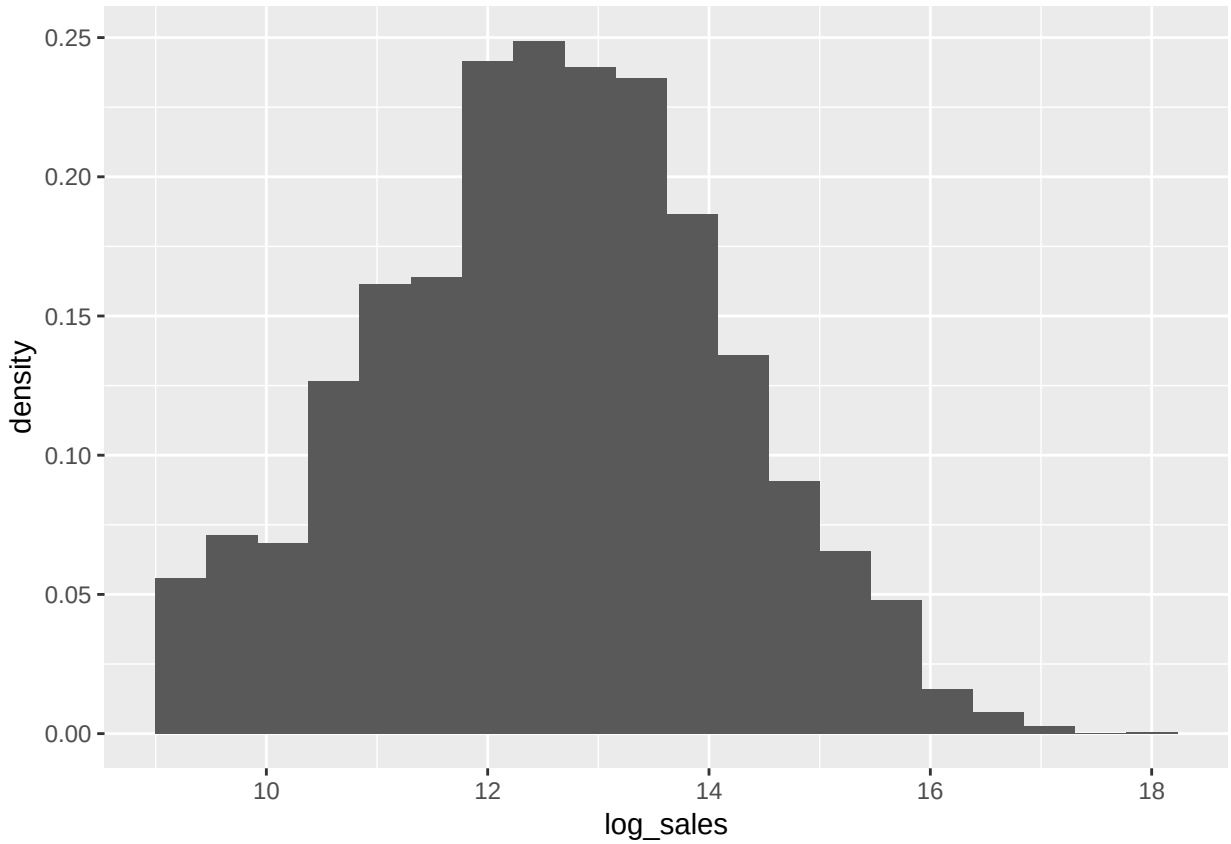
Data Distribution

In order to perform a linear regression on the data I will be taking the natural log of total sales. There are two reasons why I chose to do natural log, first is the impact that the network effect has on games, especially multiplayer games. People like to play games that their friends are playing, meaning that the popularity of a game is exponential. The other reason is that taking a look at the distribution of the data its clear that the distribution is very skewed and thus needs to be normalized.



Once the log of total sales is taken the data resembles a normal distribution.

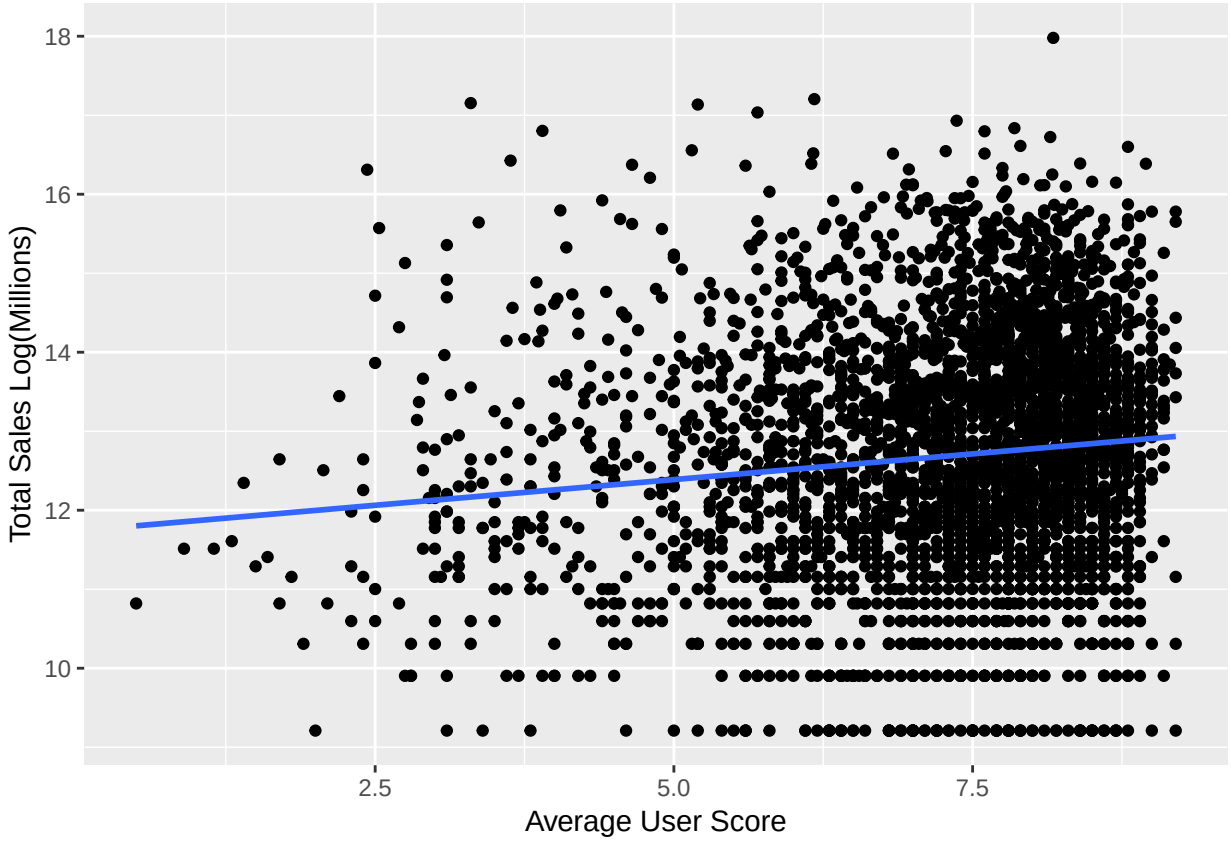
```
full_data_clean$log_sales<-log(full_data_clean$total_sales)
```



Regression analysis

Review Scores

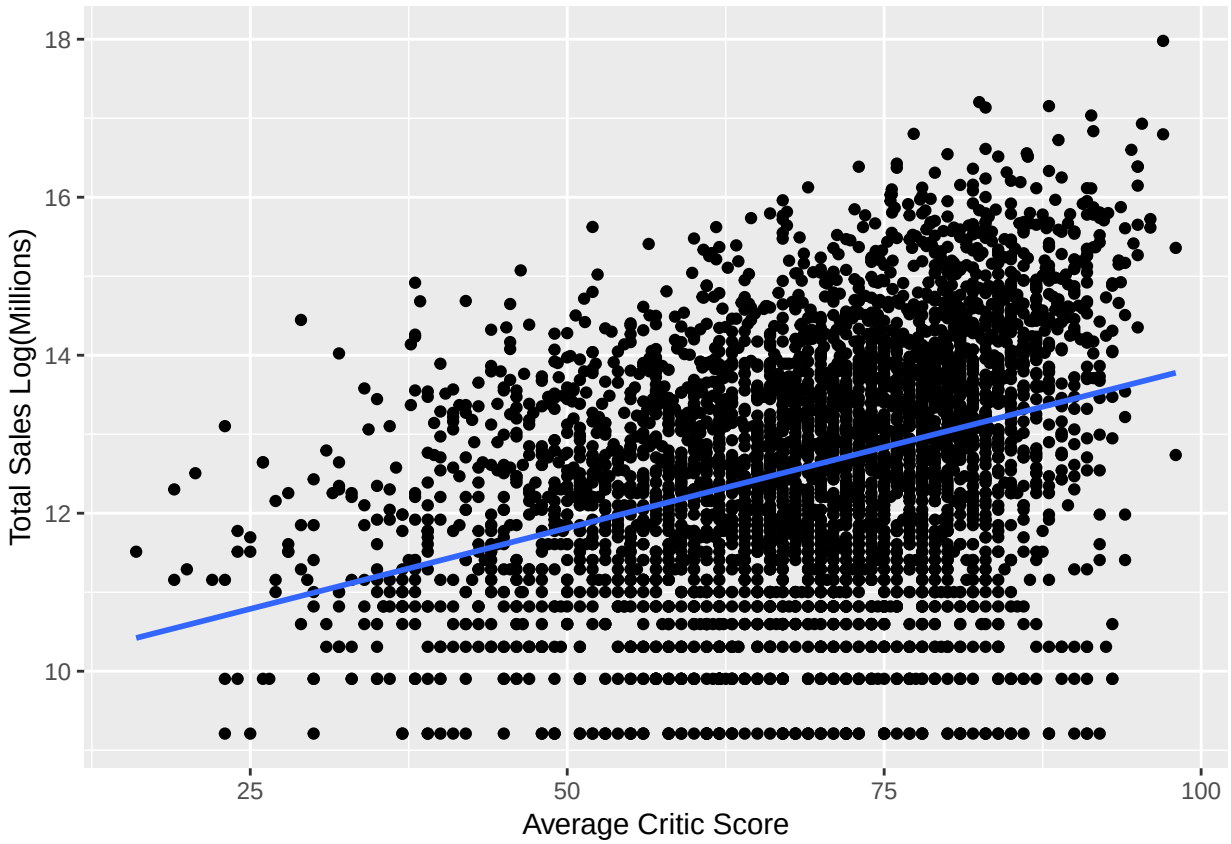
Review scores are the first variables that I chose to take a look at. Since there are many debatable factors on what makes game quality "good" aggragate critic/user scores should be a good proxy. At first I assumed user scores would have a stronger correlation than critic scores but I was mistaken.




```
##
## Call:
## lm(formula = log_sales ~ avg_user_score, data = full_data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7240 -1.0439 -0.0044  1.0070  5.1779
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.7373     0.1397  84.020  < 2e-16 ***
## avg_user_score   0.1301     0.0190   6.849 8.71e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.531 on 3606 degrees of freedom
##   (581 observations deleted due to missingness)
## Multiple R-squared:  0.01284,    Adjusted R-squared:  0.01257
## F-statistic: 46.91 on 1 and 3606 DF,  p-value: 8.712e-12
```

The R squared using aggragate user scores are quite low and some of the best selling games had aggragate user scores of less than 50%.

The results using aggragate critic scores are much better.

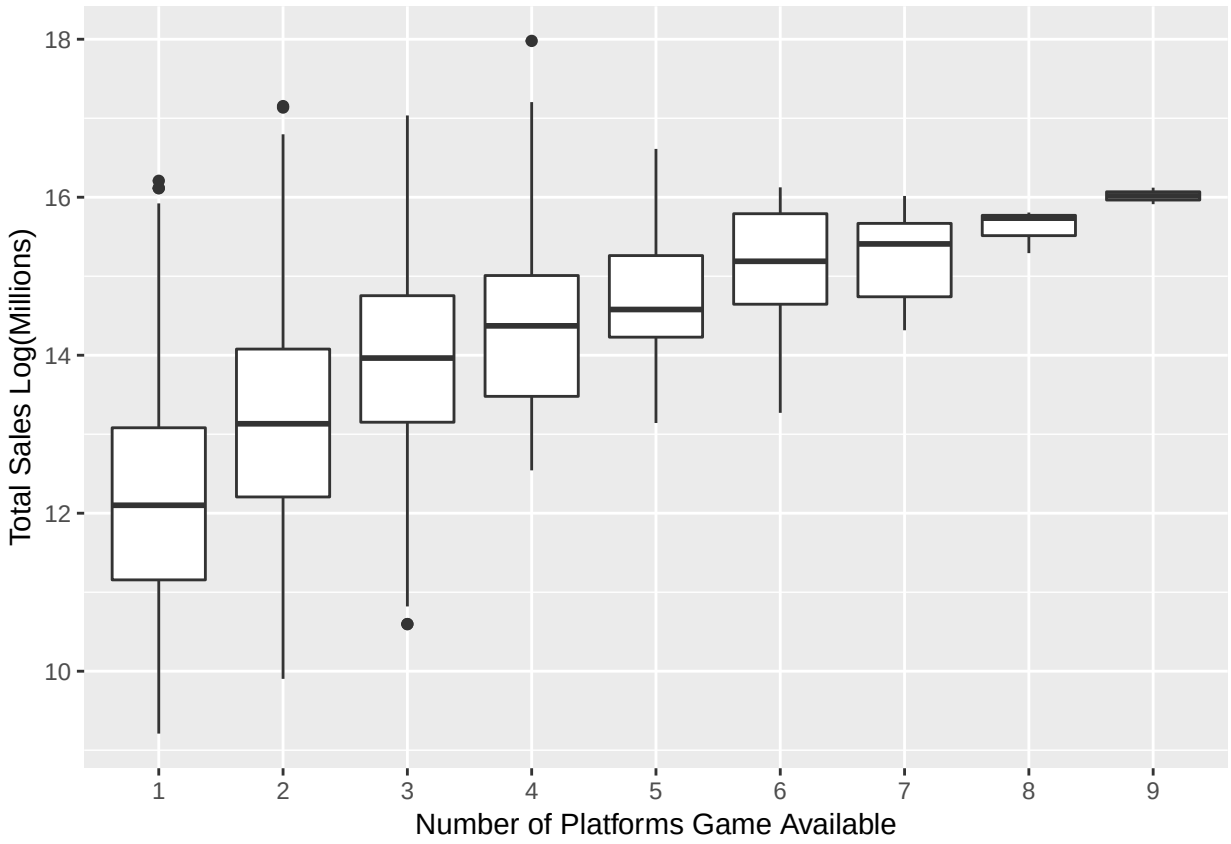



```
##
## Call:
## lm(formula = log_sales ~ avg_critic_score, data = full_data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3216 -0.9485  0.0574  1.0055  4.2423
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.765463   0.116058   84.14   <2e-16 ***
## avg_critic_score 0.040940   0.001654   24.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.444 on 4187 degrees of freedom
## Multiple R-squared:  0.1276, Adjusted R-squared:  0.1274
## F-statistic: 612.4 on 1 and 4187 DF,  p-value: < 2.2e-16
```

The R squared using aggragate critic scores is much better and the best selling games also had high critic scores.

Platform Availability

Another factor that also impacts how well a game sales is the number of platforms the game is available. Most gamers do not have multiple consoles and a game that releases on multiple platforms has a much higher potential customer number.




```
##
## Call:
## lm(formula = log_sales ~ number_platforms, data = full_data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0884 -0.9433  0.0218  0.9570  4.2276
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.40844    0.03936   289.9   <2e-16 ***
## number_platforms  0.75888    0.02150    35.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.357 on 4187 degrees of freedom
## Multiple R-squared:  0.2293, Adjusted R-squared:  0.2291
## F-statistic:  1246 on 1 and 4187 DF,  p-value: < 2.2e-16
```

The number of platforms a game is available accounts for nearly a quarter of the variation! Clearly there is a very strong relationship here.

Multiple Linear Regression

There are two other variables that I chose to include in a multiple linear regression: * **Genre:** Certain genres are more popular than others, visual novels for example are a genre that does not have as much mainstream appeal as other genres. * **Series:** 19 of the top 20 best selling games are part of a series, gamers may be more willing to purchase a game if they are familiar with the gameplay or story.

```
##
## Call:
## lm(formula = log_sales ~ avg_critic_score + number_platforms +
##     factor(genre) + factor(series), data = full_data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1655 -0.7696  0.0522  0.8255  3.5636
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    8.936388   0.106492  83.916  < 2e-16 ***
## avg_critic_score               0.035671   0.001432  24.917  < 2e-16 ***
## number_platforms               0.657793   0.019735  33.331  < 2e-16 ***
## factor(genre)Action-Adventure  0.085758   0.163382   0.525 0.599685
## factor(genre)Adventure        -0.362435   0.089431  -4.053 5.16e-05 ***
## factor(genre)Board+Game        0.625253   1.204545   0.519 0.603734
## factor(genre)Fighting         -0.050495   0.090421  -0.558 0.576575
## factor(genre)Misc              0.100591   0.093716   1.073 0.283172
## factor(genre)MMO               1.096929   0.540120   2.031 0.042328 *
## factor(genre)Music            -0.837174   0.280208  -2.988 0.002827 **
## factor(genre)Party             0.954451   0.696483   1.370 0.170640
## factor(genre)Platform         -0.139698   0.089070  -1.568 0.116860
## factor(genre)Puzzle           -0.594533   0.111812  -5.317 1.11e-07 ***
## factor(genre)Racing           -0.075938   0.077370  -0.981 0.326411
## factor(genre)Role-Playing     -0.261580   0.070920  -3.688 0.000229 ***
## factor(genre)Shooter          -0.001578   0.069938  -0.023 0.982002
## factor(genre)Simulation       -0.331132   0.094128  -3.518 0.000440 ***
## factor(genre)Sports            0.053453   0.071053   0.752 0.451909
## factor(genre)Strategy         -1.214430   0.097845 -12.412  < 2e-16 ***
## factor(genre)Visual+Novel     -2.393895   0.541199  -4.423 9.97e-06 ***
## factor(series)yes              0.501524   0.040445  12.400  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.203 on 4167 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3968, Adjusted R-squared:  0.3939
## F-statistic: 137.1 on 20 and 4167 DF,  p-value: < 2.2e-16
```

Due to the previously mentioned limitations of the data as well as well as information that is not publicily available (game budget/marketing being a big one) this model is fairly limited in modeling game sales, but the variables that we do have available can still explain a decent amount!