S&DS 230 Final Project

Introduction

The video game industry is one of the largest in the world, generating around 150 billion dollars in revenue each year. While there are a wide variety of games in terms of genre, rating, time period, console, etc., we know that there are extreme differences in video game sales. For example, the most popular games in the world make sales in the hundreds of millions, while some games will only make a few thousand dollars. This dataset from Kaggle has video game global sales over thousands of observations with many more interesting factors to analyze and compare. In addition to basic attributes like the ones mentioned, this dataset also includes critic score and user score according to Metacritic, a renowned website that rates and ranks video games (kind of like IMDB or Rotten Tomatoes but for video games instead of movies). Our goal in this project is to analyze all of the interesting factors from this dataset, checking for different correlations and interesting patterns that might arise between a game’s characteristics and its success in terms of global sales. We will also conduct other potentially interesting analyses that are unrelated to sales. Hopefully this information will give us a deeper look and new perspective to the video game industry!

Data

Here is the list of relevant variables and descriptions based on our cleaned data:

1. Name: categorical, character vector, provides the game’s name
2. Platform: categorical, character vector, provides the game’s platform
3. Year_of_Release: continuous, numeric, provides the game’s release year
4. Genre: categorical, character vector, provides the game’s genre
5. Publisher: categorical, character vector, provides the game’s publisher
6. Developer: categorical, character vector, provides the game’s developer
7. Global_Sales: continuous, numeric, provides the game’s global sales in thousands of dollars
8. Critic_Score: continuous, integer, provides the game’s Metacritic score on a scale of 1-100
9. Critic_Count: continuous, integer, provides the number of critic reviews a game received
10. User_Score: continuous, numeric, provides the game’s Metacritic user score on a scale of 1-10
11. User_Count: continuous, integer, provides the number of user reviews a game received
12. Rating: categorical, character vector, provides the game’s ESRB rating
13. log_User_Count: continuous, numeric, provides the log of User_Count
14. log_Critic_Count: continuous, numeric, provides the log of Critic_Count
15. log_Global_Sales: continuous, numeric, provides the log of Global_Sales

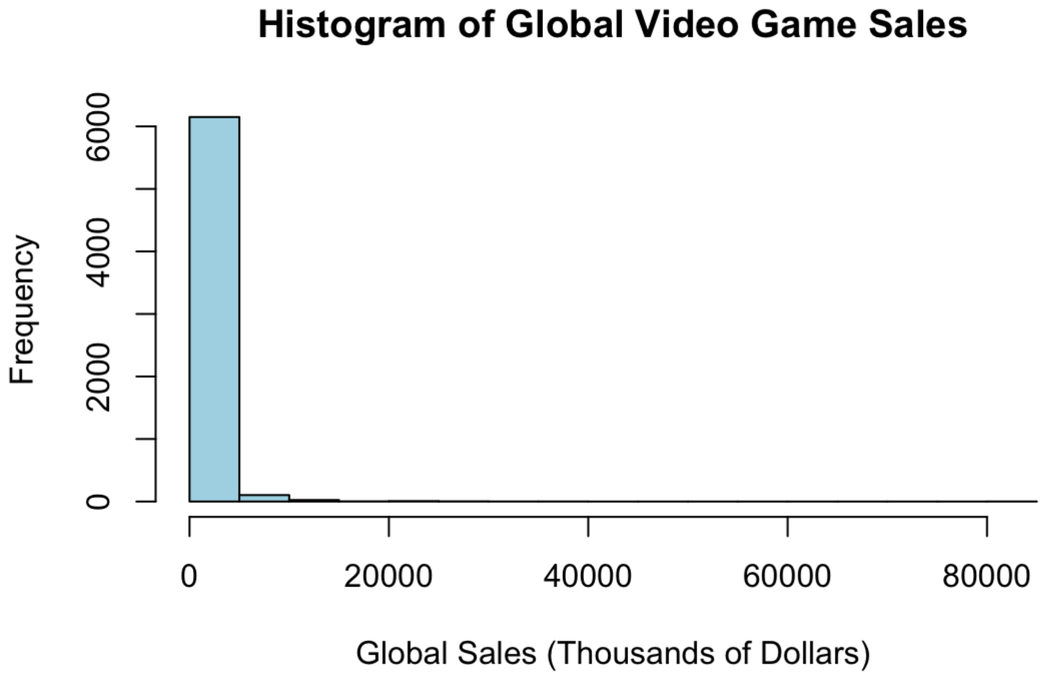
Data Cleaning

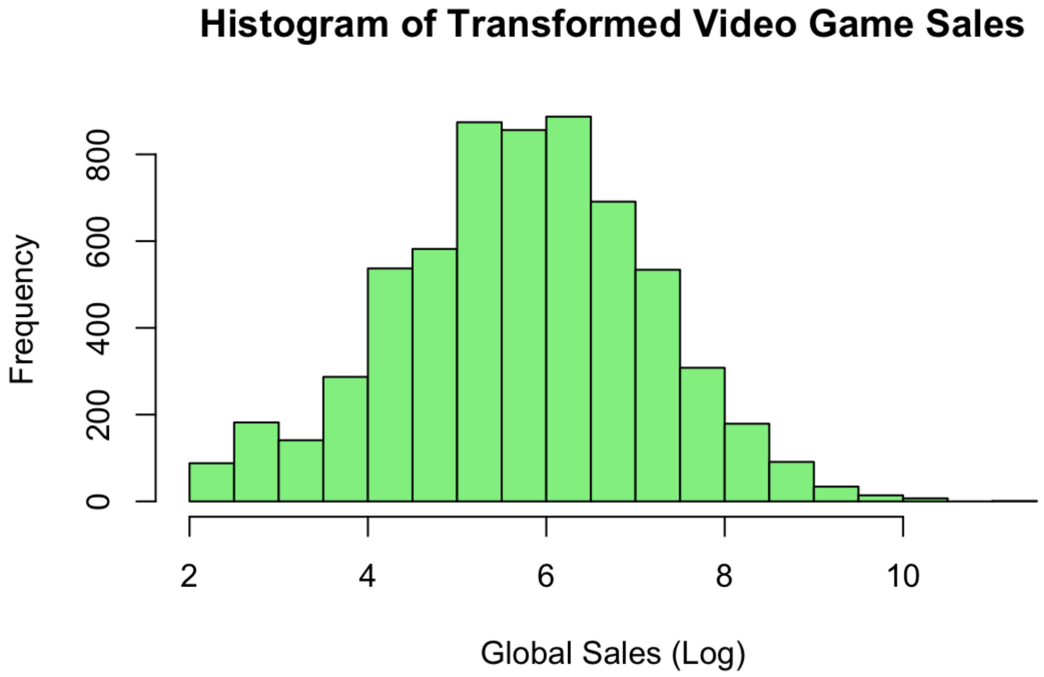
The first thing we did with the raw data was remove any duplicate rows and remove any rows that had incomplete (NA) fields. This was okay because although many rows were missing data, we were still left with over 6,000 observations on which we could perform our analyses. Next, we converted our global sales from millions of dollars into thousands of dollars as this would allow us to see whole numbers instead of decimals. After this, we analyzed the ESRB Ratings and made some consolidations. For example, we combined “Adults Only” and “Mature” into one category and combine “E for Everyone” and “Kids-Adults” into one category. We also removed “Rating Pending” as there were only 2 of these and they are essentially an incomplete field. After these adjustments, we had 4 ratings leftover: “Everyone”, “Everyone10+”, “Teens”, and “Mature” which can be seen in the ESRB graphic *below. We also renamed some of these ratings to make them more descriptive. We took a similar approach to genres, combining “Puzzle” and “Strategy” into one category, and “Action” and “Adventure into one category. This left us with just 10 genres. Next, we consolidated many of the platforms. For example, we combined all of the Play Stations (1, 2, 3, etc.) into just one category. Similarly, we combined the various Xbox groups into one category. This left us with 8 different platform groups. We then took a look at Publisher and Developer. We made no adjustments just yet, but will alter them later in the report. We decided to leave critic score and user score alone, but made some adjustments based on critic count and user count. We decided that for a game to be considered, it must have at least 4 critic reviews and at least 6 user reviews, so we removed any observations that did not meet these requirements. For Year of Release, there were empty values labeled”N/A" instead of “NA” so they weren’t caught the first time around. We now removed these rows and then converted the year into a number as it was previously a character vector. Next, we created log transformed versions of critic count, user count, and global sales as these will come in handy later on in the report. Finally, we deleted the rest of the sales columns as we are only concerned with global sales in this analysis. We were left with nicely cleaned data that has over 6,000 observations and 15 columns. Judging which groups to consolidate and how to approach deletion of certain observations were the two most challenging parts of this data cleaning process.*

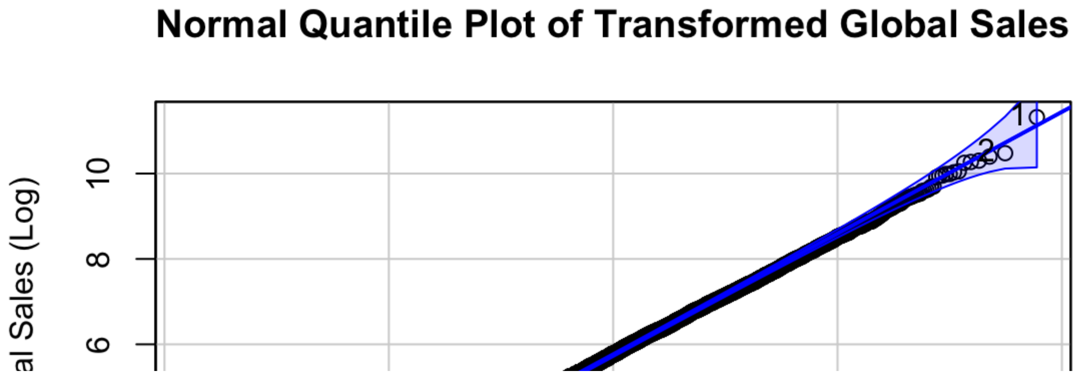
Summary Info and Basic Plots

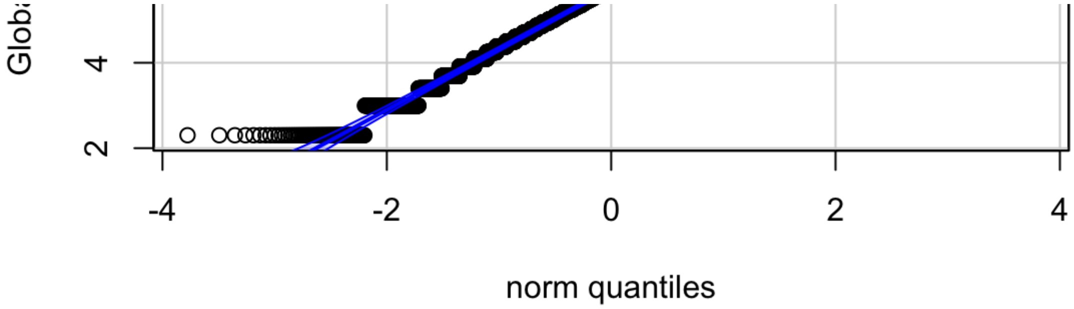
```
##      Name             Platform         Year_of_Release    Genre
##  Length:6293        Length:6293        Min.   :1985    Length:6293
##  Class :character   Class :character   1st Qu.:2004    Class :character
##  Mode  :character   Mode  :character   Median :2007    Mode  :character
##                                        Mean   :2008
##                                        3rd Qu.:2011
##                                        Max.   :2016
##   Publisher          Global_Sales      Critic_Score  Critic_Count
##  Length:6293        Min.   :   10.0   Min.   :13    Min.   :  4.00
##  Class :character   1st Qu.:  120.0   1st Qu.:63    1st Qu.: 15.00
##  Mode  :character   Median :  320.0   Median :73    Median : 26.00
##                     Mean   :  823.6   Mean   :71    Mean   : 30.19
##                     3rd Qu.:  820.0   3rd Qu.:81    3rd Qu.: 41.00
##                     Max.   :82530.0   Max.   :98    Max.   :113.00
##    User_Score      User_Count       Developer           Rating
##  Min.   :0.500   Min.   :    6.0   Length:6293        Length:6293
##  1st Qu.:6.600   1st Qu.:   14.0   Class :character   Class :character
##  Median :7.600   Median :   31.0   Mode  :character   Mode  :character
##  Mean   :7.235   Mean   :  189.2
##  3rd Qu.:8.200   3rd Qu.:  102.0
##  Max.   :9.600   Max.   :10665.0
##  log_User_Count  log_Critic_Count log_Global_Sales
##  Min.   :1.792   Min.   :1.386    Min.   : 2.303
##  1st Qu.:2.639   1st Qu.:2.708    1st Qu.: 4.787

##  Median :3.434   Median :3.258    Median : 5.768
##  Mean   :3.765   Mean   :3.181    Mean   : 5.746
##  3rd Qu.:4.625   3rd Qu.:3.714    3rd Qu.: 6.709
##  Max.   :9.275   Max.   :4.727    Max.   :11.321
```

First, our summary information interestingly shows that the minimum sales in our dataset was merely $10,000 while the maximum was over $80,000,000. We can also observe that for both User_Score and Critic_Score, no game got the highest or lowest score possible. The means and medians were in the 7s and 70s respectively which is quite high. We also observe that there aren't too many reviews for either critics or users as the medians of both are only in the 20s and 30s. However, the mean of User_Count is 189 which is much higher and might indicate that some games gets a very large number of user reviews.


Histogram of Global Video Game Sales
Frequency
0
2000
4000
6000
0
20000
40000
60000
80000
Global Sales (Thousands of Dollars)


Histogram of Transformed Video Game Sales
Frequency
0
200
400
600
800
2
4
6
8
10
Global Sales (Log)

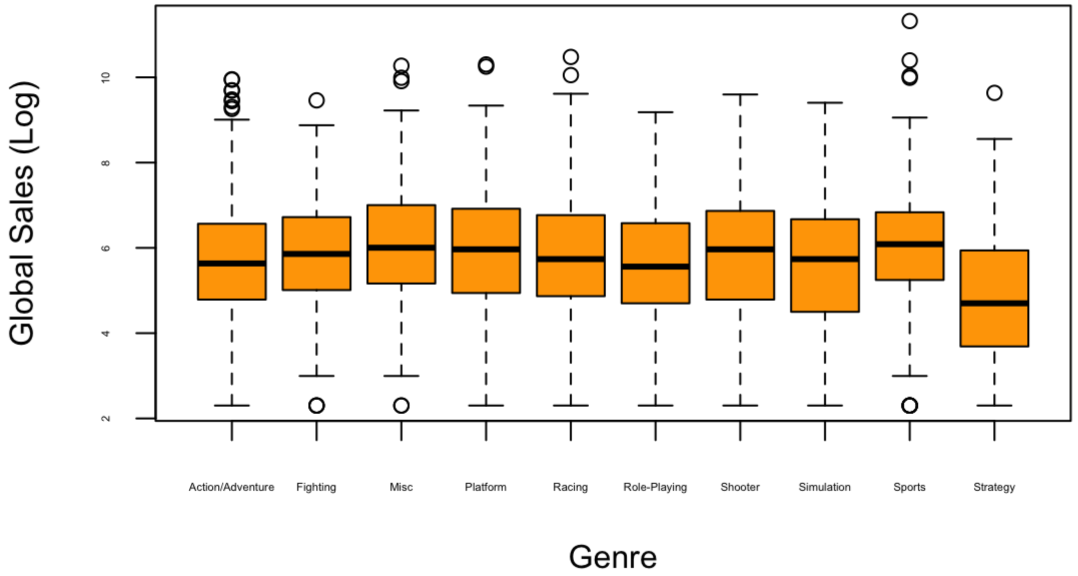

Normal Quantile Plot of Transformed Global Sales
al Sales (Log)
6
8
10
1
2



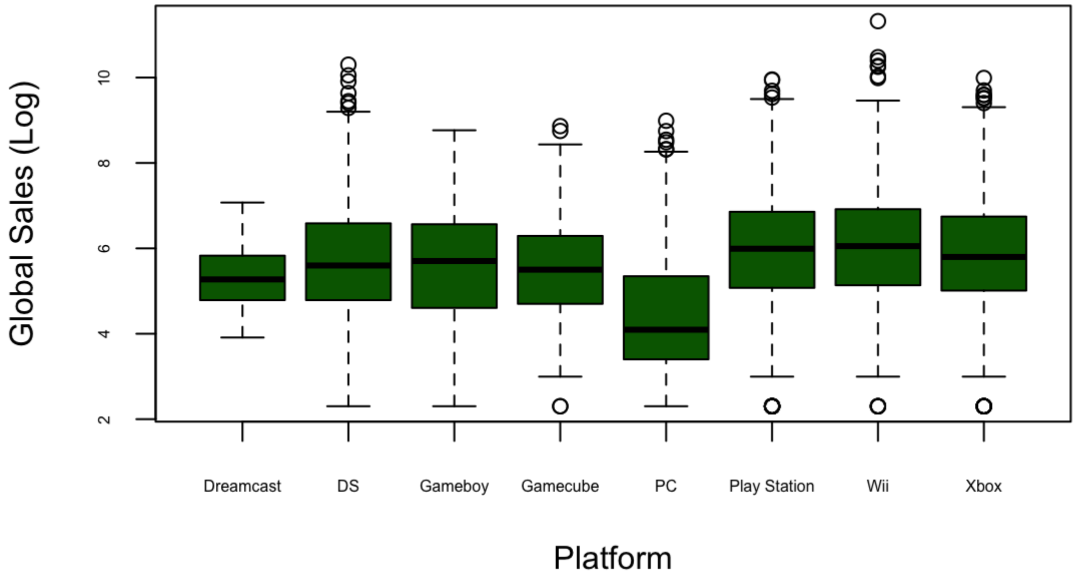
```
## [1] 1 2
```

We made two histograms. Using the raw global sales data, we can see that the shape of the distribution is extremely right skewed. This makes sense as there are many games with little sales but a few exceptions (the most popular games in the world) that generate tens of millions of dollars. We then used our log transformed global sales to make another histogram. The shape of this distribution was much better and we then made a normal quantile plot to check its normality. We can see that most of the dots fall within the bounds, but there are clearly many observations that fall outside towards the bottom left of the normal quantile plot. This indicates that even with the log transformation, global sales is still not very normal and remains somewhat right skewed.

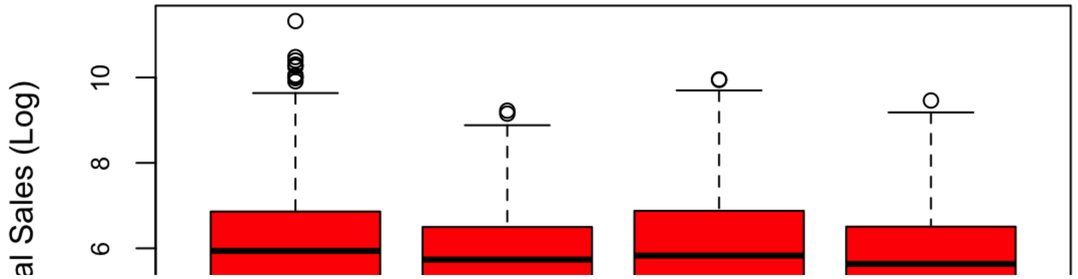
Boxplot of Transformed Global Video Game Sales by Genre

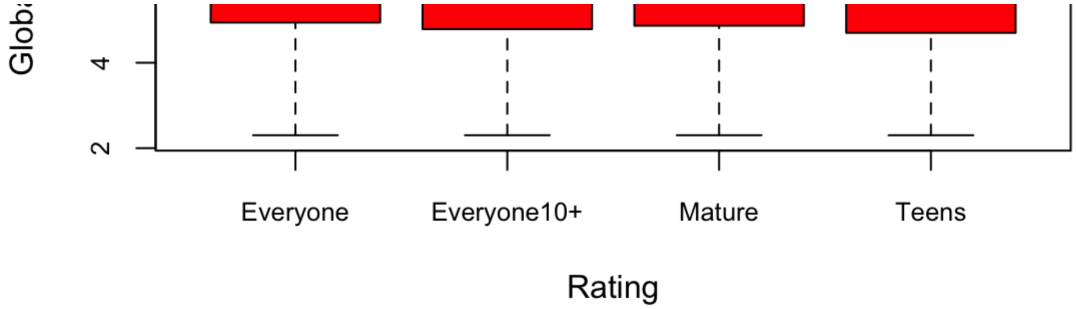


Boxplot of Transformed Global Video Game Sales by Platfor

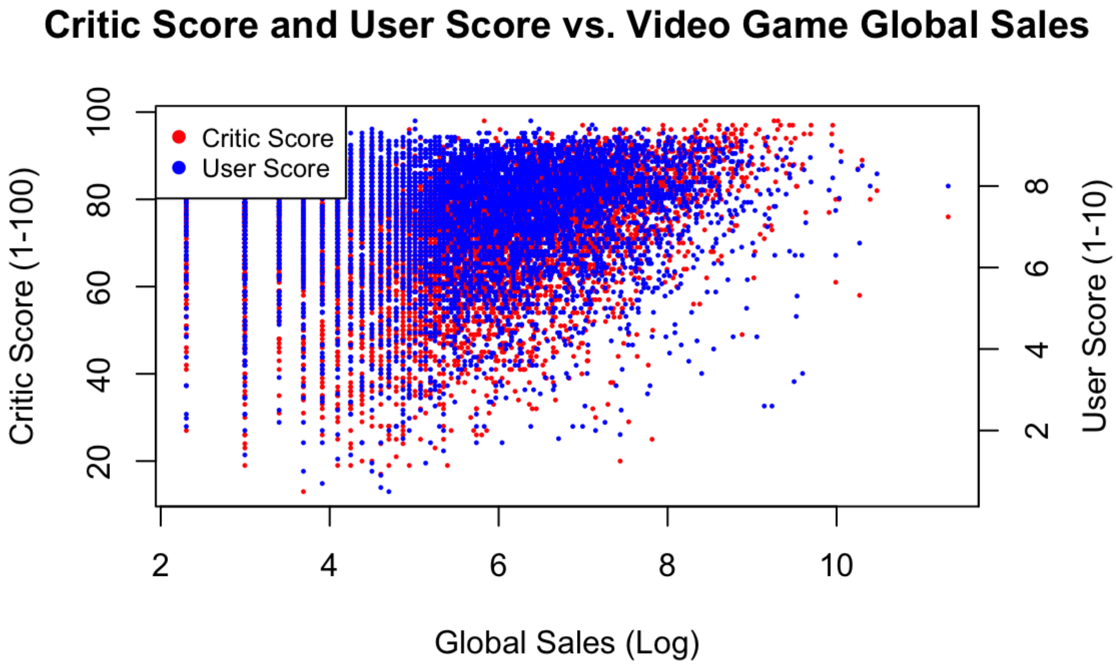


Boxplot of Transformed Global Video Game Sales by Ratinc





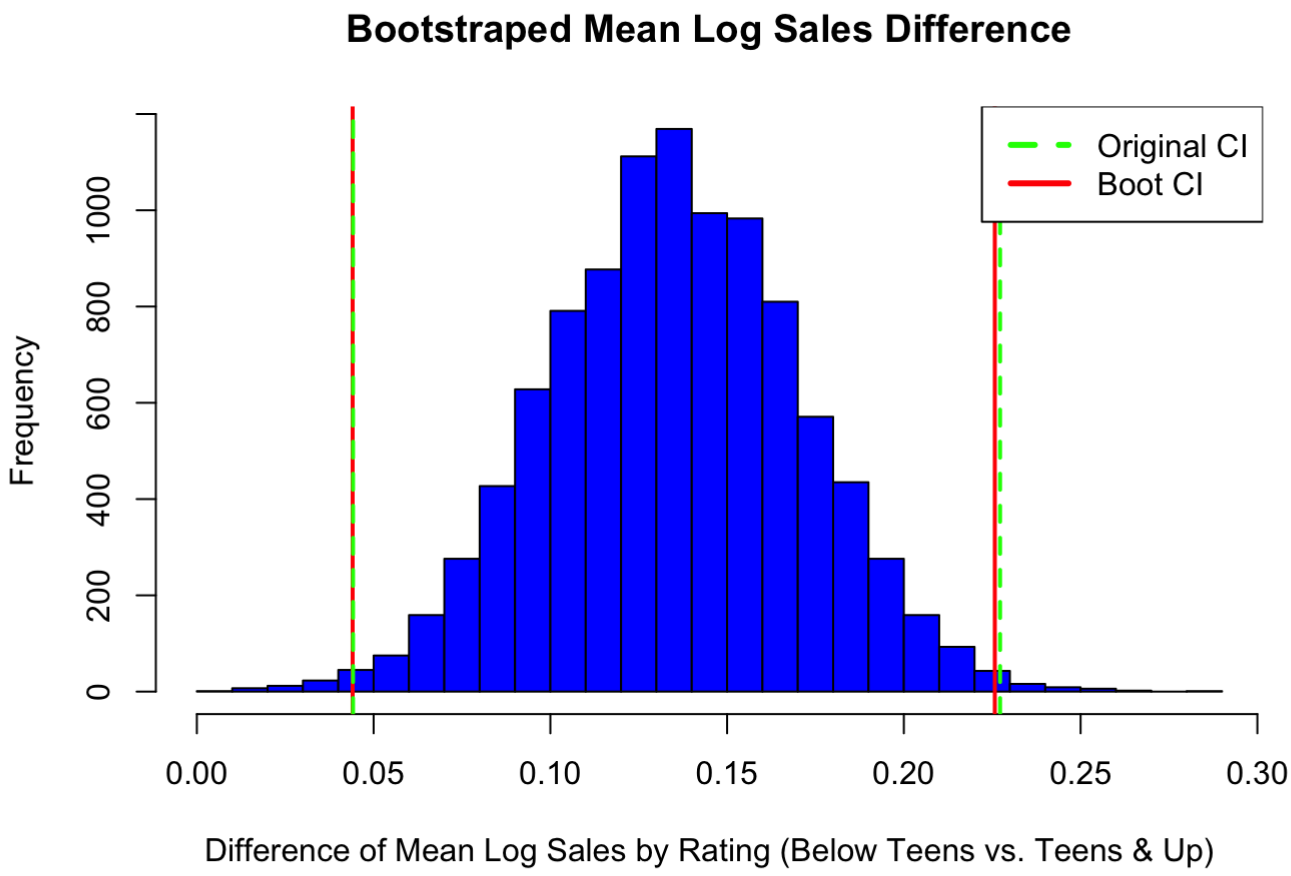
We made three interesting box plots. The first (orange) shows transformed global sales by genre. While many of the genres seem similar (are not obviously different), we can observe that the strategy genre (which also includes puzzle games) generally has lower sales. Some genres such as sports, racing, and action/adventure also seem to have more higher outliers (games that did extremely well). The second plot (green) is transformed global sales by platform. Here, we can see that PC has seemingly lower sales than other platforms as its box is noticeably lower. The Dreamcast also has very low spread as compared with most of the other platforms which all have a fair amount of outliers. Finally, our third (red) plot shows transformed global sales by ESRB Rating. There is no clear difference between groups which is fairly surprising as we might expect games that everyone could play to sell more. However, the Everyone category does appear to have more high outliers (games that did very well) as compared to the other ratings.



Finally, we made a two-way scatterplot of user scores and critic scores vs. transformed global sales. As shown in the graph, there definitely seems to be some similarity between the two scores (high critic scores have high user scores and vice versa). There also appears to be a positive relationship between both of the variables and the log global sales. Specifically, higher scores (of both types) seems to correspond with higher sales and vice versa. This is quite interesting but not entirely unexpected as a better rated game would probably sell more.

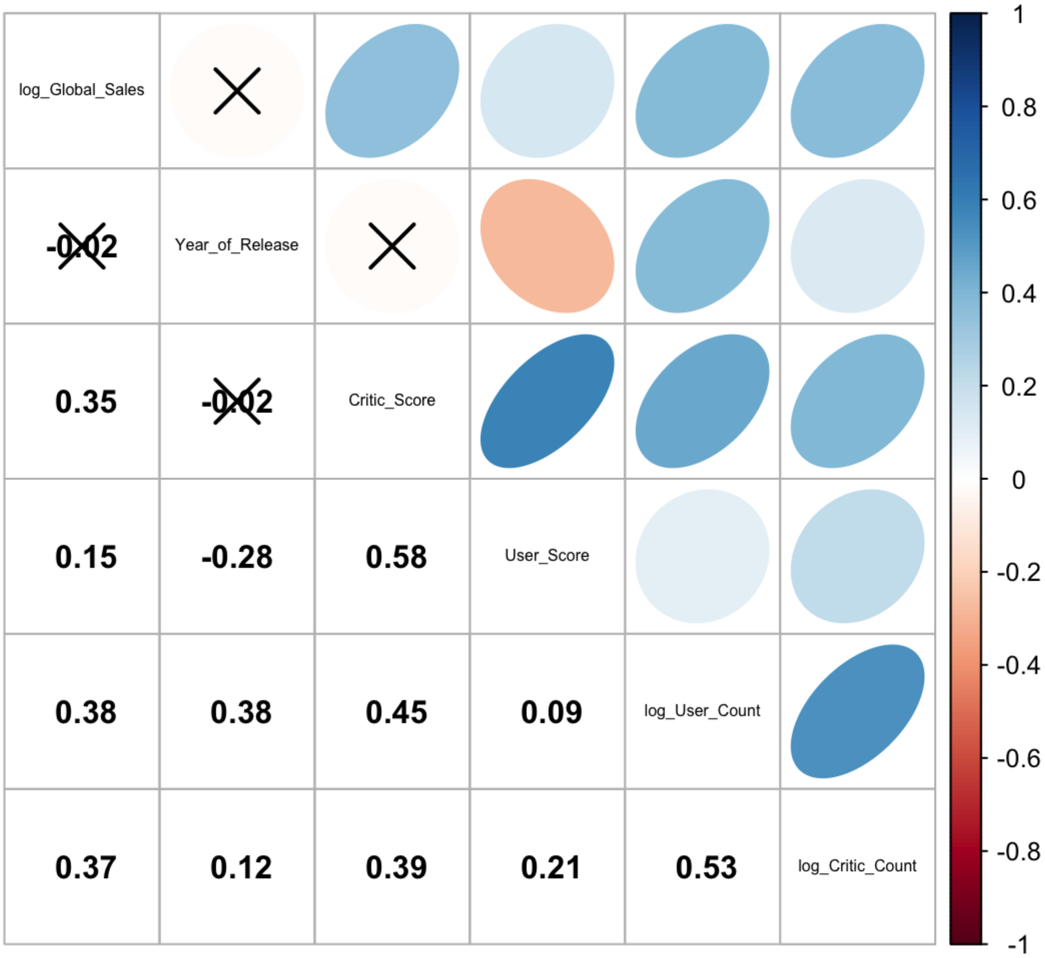
Analysis

T-test and Bootstrap CIs



The bootstrapped CI is extremely close to the theoretical CI on both ends. Thus, just as the theoretical CI did not include 0, neither does the bootstrapped CI. Therefore, both our CIs support the conclusion that the difference between the mean log Global Sales of video games rated for Everyone or Everyone10+ and those rated Teens or Mature is statistically significantly different than 0. Further, since both CIs are above 0, we are 95% confident that the video games rated Everyone or Everyone+ is larger than video games rated Teen or Mature. We merged the 4 ratings into these 2 categories to facilitate the t.test.

Correlation, Cor-Test, and More Bootstrap CI



```
## [1] "Critic Score v Log Sales 95% Parametric Corr CI: (0.3293, 0.3726)"
```

```
## [1] "Critic Score v Log Sales 95% Bootstrap Corr CI: (0.329, 0.3717)"
```

```
## [1] "User Score v Log Sales 95% Parametric Corr CI: (0.1272, 0.1755)"
```

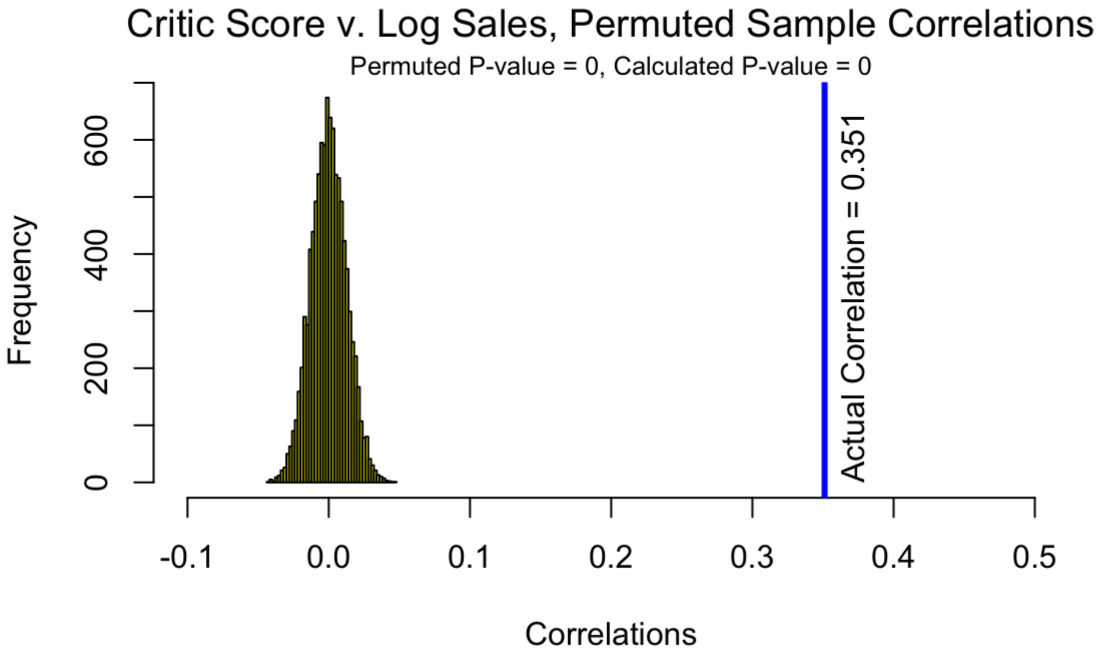
```
## [1] "User Score v Log Sales 95% Bootstrap Corr CI: (0.1299, 0.1728)"
```

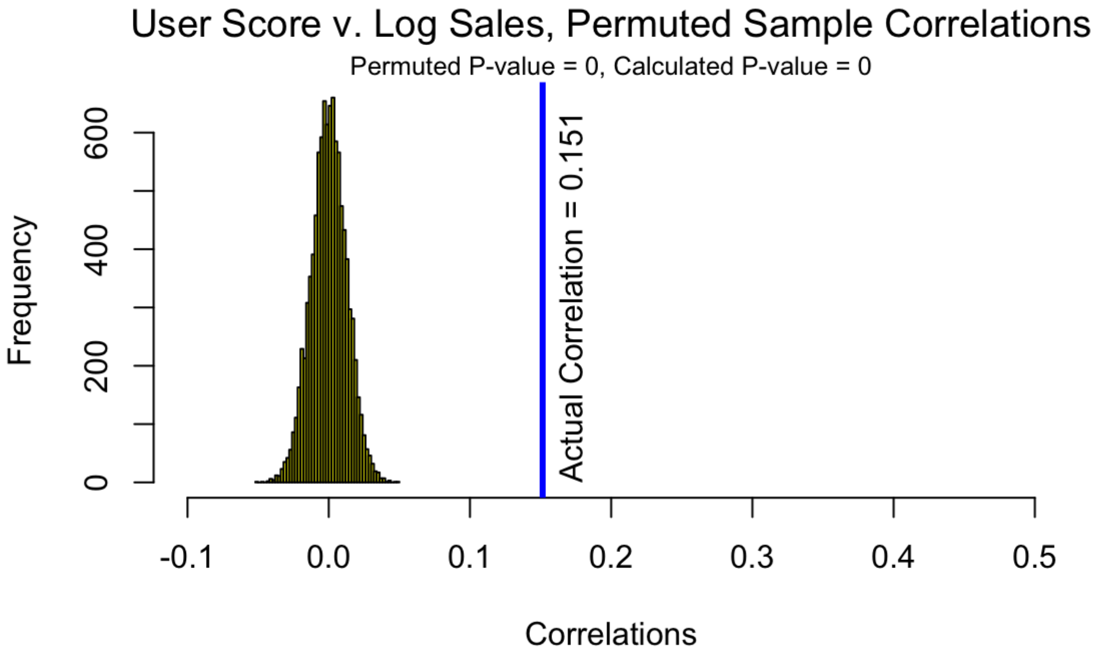
Looking at the matrix generated by corrplot.mixed, we can see that there is a statistically significant correlation between log global sales and critic score, and log global sales and user score. However, we should also note the statistically significant correlation between user score and critic score, as this collinearity could affect our conclusions.

We used parametric tests to determine a 95% confidence interval for the true value of correlation between log global sales and critic score and between log global sales and user score. These confidence intervals are [0.3293, 0.3726] and [0.1272, 0.1755], respectively.

We also used bootstrapping to accomplish the same thing. Taking 1000 samples, we calculated 95% confidence intervals of [0.3290, 0.3717] and [0.1299, 0.1728] for log global sales and critic score, and log global sales and user score, respectively. These are very similar to our parametric confidence intervals, which is what we want.

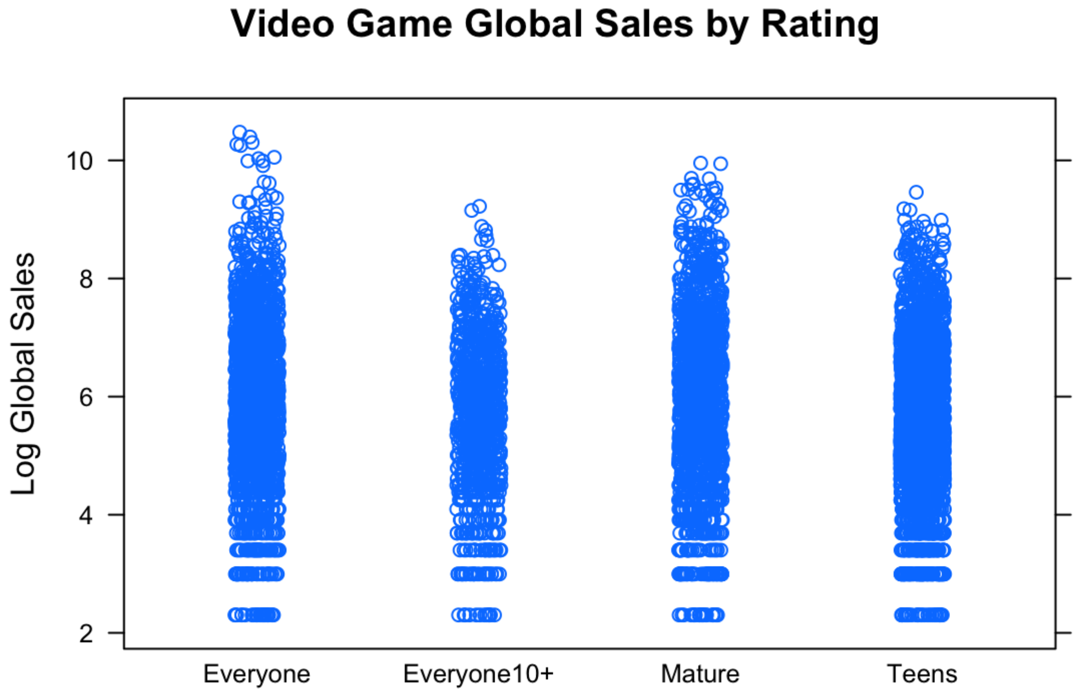
Permutation Test





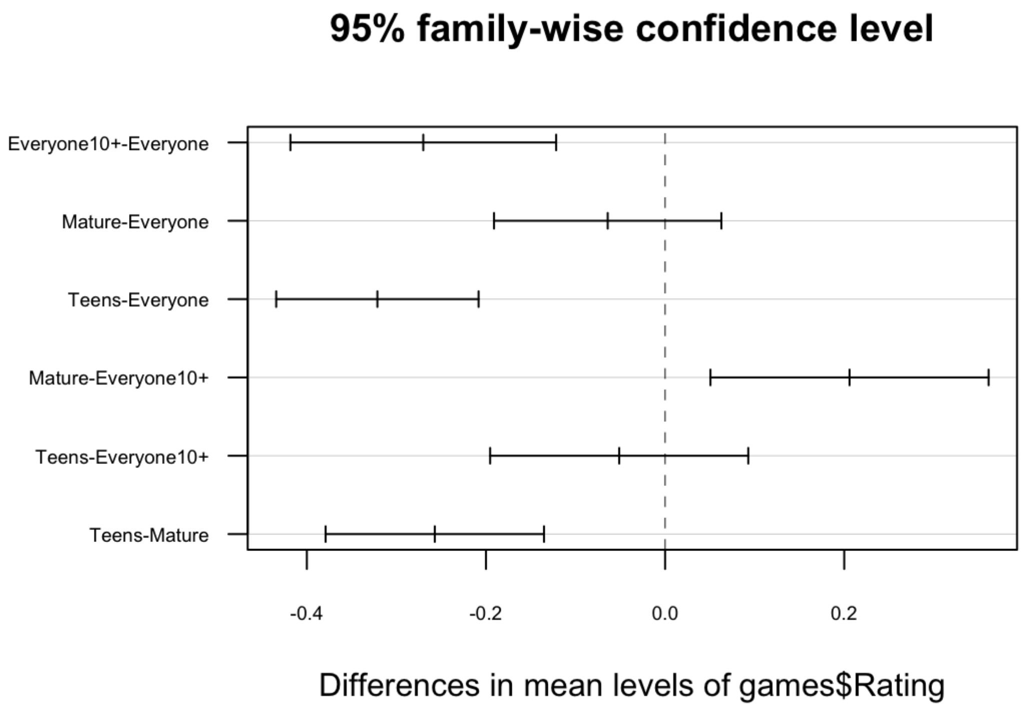
We calculated a sample correlation of 0.351 for Critic Score and Log Global Sales. Running a permutation test, we get a very small p-value (almost 0), and we can see that our sample correlation is way off to the right on the permutation test histogram. The same is true for our sample correlation between User Score and Log Global Sales, which had a calculated value of 0.151.

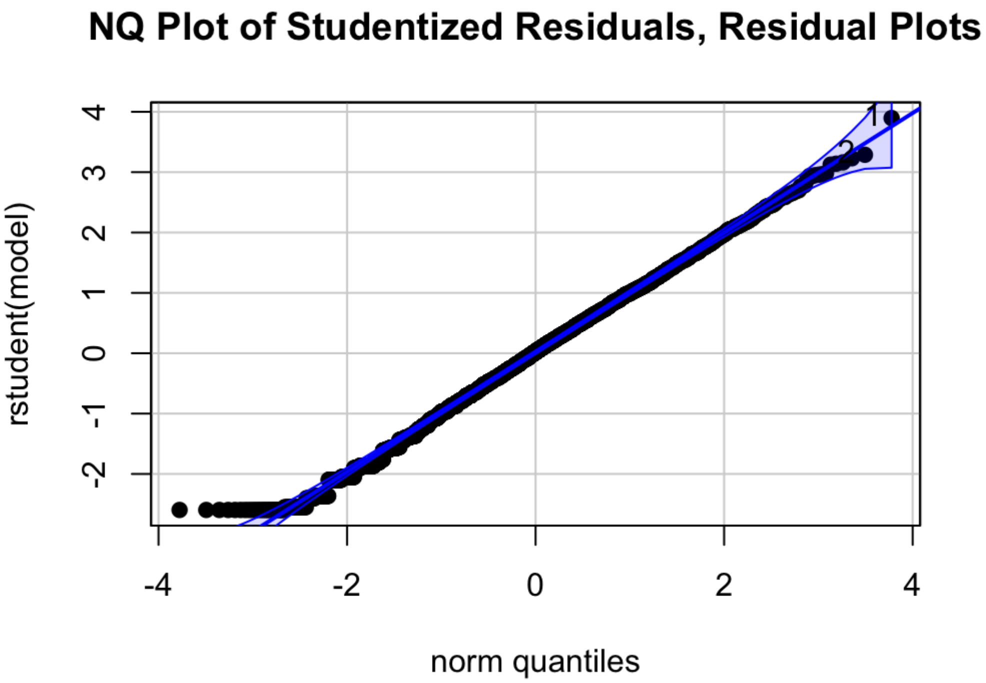
ANOVA

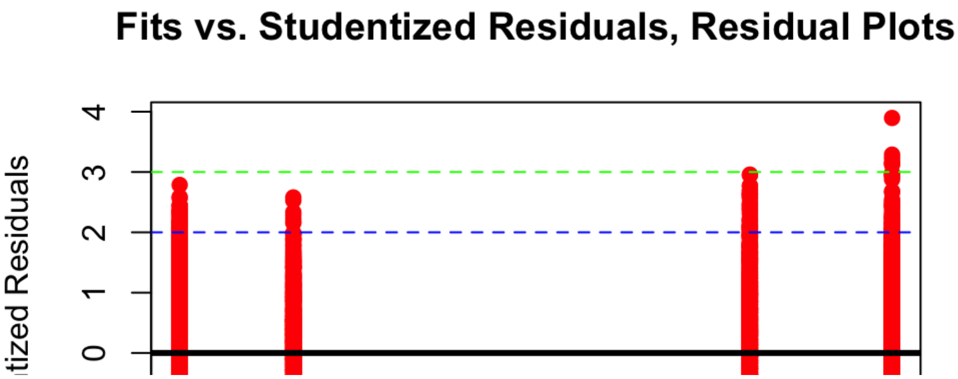


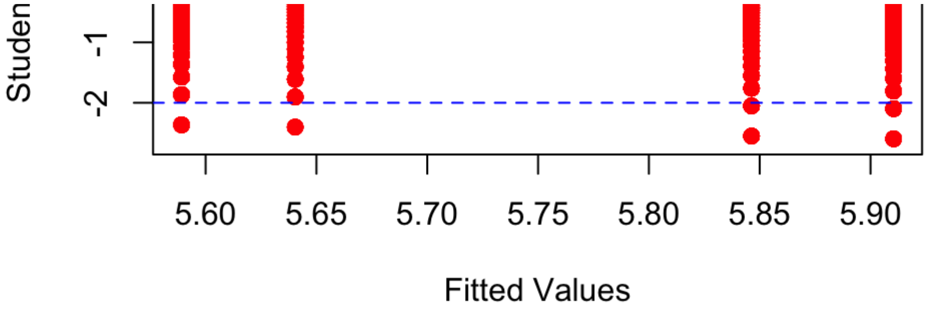
```
## [1] 1.148916
```

```
##                   Df  Sum Sq Mean Sq F value              Pr(>F)    
## games$Rating     3    127   42.47   21.96 0.0000000000000382 ***
## Residuals     6289  12161    1.93                               
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


95% family-wise confidence level
Everyone10+-Everyone
Mature-Everyone
Teens-Everyone
Mature-Everyone10+
Teens-Everyone10+
Teens-Mature
-0.4
-0.2
0.0
0.2
Differences in mean levels of games$Rating


NQ Plot of Studentized Residuals, Residual Plots
rstudent(model)
4
3
2
1
0
-1
-2
-4
-2
0
2
4
norm quantiles
1
2


Fits vs. Studentized Residuals, Residual Plots
tized Residuals
4
3
2
1
0



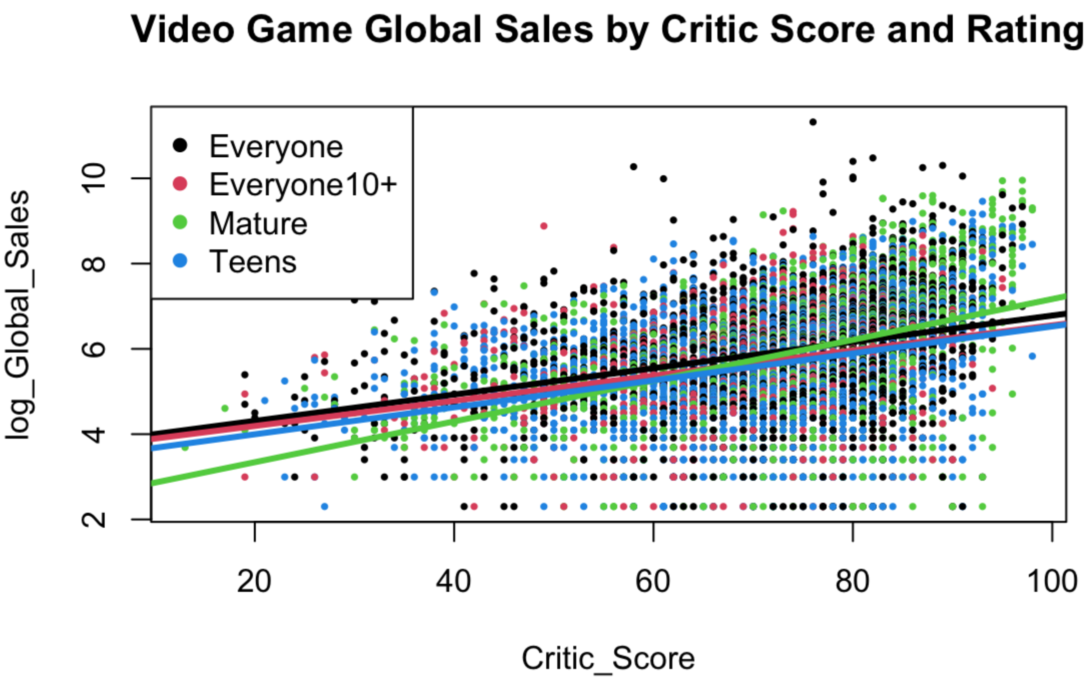
One way ANOVA test for log(global sales) and game ratings. First we remove extreme outliers, of which there is at least one in our data. Nintendo Wii Sports really obliterated the competition so we're gonna have it sit this one out for the safety of our results. As predicted earlier, raw global sales data is not suitable for ANOVA analysis due to the severe right skew of the data. Having prior performed a log transformation, we examine our data visually using a strip plot. From this, we see that standard deviations across categories seem roughly the same and the data seems to be normally distributed within each rating category. Although we cannot tell for certain due to the high concentration of points in the strip plot, our earlier boxplots also seem to indicate that standard deviations are roughly the same and that the data is normally distributed. We also confirm numerically that the maximum standard deviation to minimum standard deviation ratio is less than 2. With these results, we can confidently proceed with ANOVA analysis. Our one way ANOVA test produces a p value of 3.82e-14, indicating that global sales is statistically significantly different across game ratings. Our degrees of freedom (3) also make sense, because we have n = 4 categories (ratings), and degrees of freedom are calculated as n-1. The Tukey Comparison graph plots the 95% confidence interval of mean differences between pairs of different ratings. Through this, we see that log(global sales) between Rating:Mature and Rating:Everyone and between Rating:Teen and Rating:Everyone10+ are NOT statistically significantly different. In the plot, the 95% CI mean difference for these two categories overlap with 0. All other pairings of categories are statistically significantly different (do not overlap with 0 in the plot). Last is to check our assumption for the ANOVA test. Our residual plots indicates that our residual data is normally distributed. Our fits vs residuals plots shows minimal signs of heteroskedsticity and has relatively few outliers. These plots indicate that our assumptions for ANOVA hold and that our analysis using that method is sound.

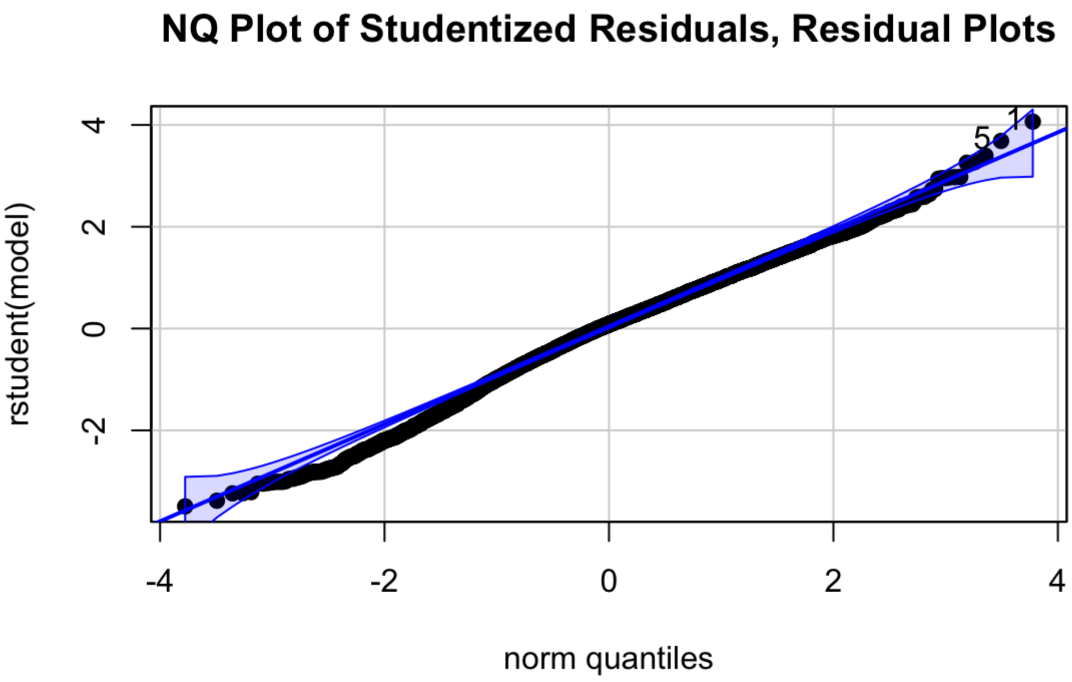
ANCOVA

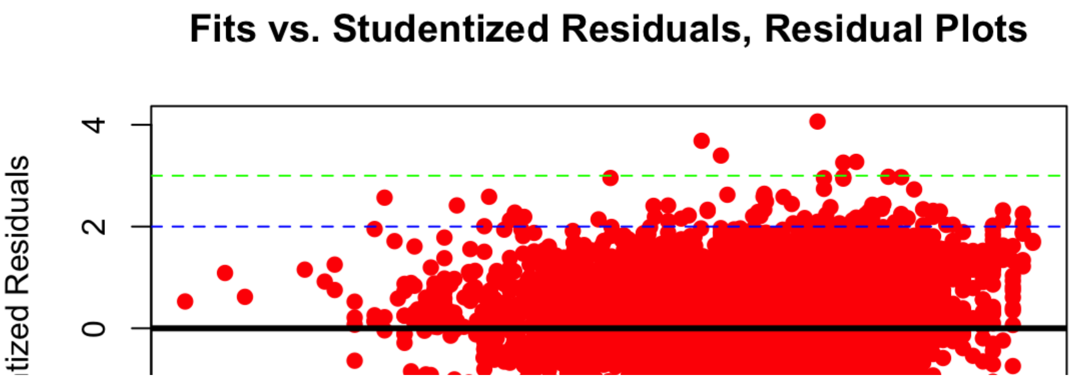
```
## Anova Table (Type III tests)
##
## Response: games$log_Global_Sales
##                                  Sum Sq   Df  F value                Pr(>F)
## (Intercept)                       887.4    1 524.637 < 0.00000000000000022 ***
## games$Critic_Score                331.7    1 196.085 < 0.00000000000000022 ***
## games$Rating                       58.4    3  11.509         0.00000016039 ***
## games$Critic_Score:games$Rating    63.8    3  12.580         0.00000003383 ***
## Residuals                       10630.2 6285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

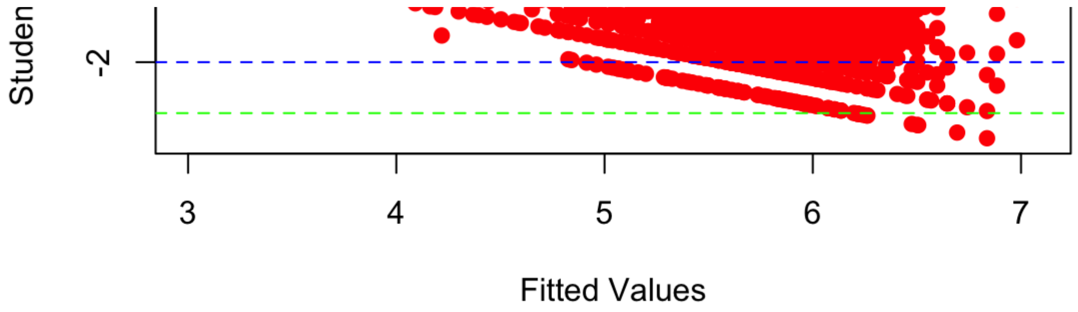
```
## 
## Call:
## lm(formula = games$log_Global_Sales ~ games$Critic_Score * games$Rating)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.5341 -0.7918  0.1088  0.8834  5.2783 
## 
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                 3.6932089  0.1612406  22.905
## games$Critic_Score                          0.0309132  0.0022076  14.003
## games$RatingEveryone10+                    -0.0917179  0.2912785  -0.315
## games$RatingMature                         -1.3085018  0.2383946  -5.489
## games$RatingTeens                          -0.3291026  0.2193931  -1.500
## games$Critic_Score:games$RatingEveryone10+ -0.0013789  0.0040992  -0.336
## games$Critic_Score:games$RatingMature       0.0169579  0.0032465   5.224
## games$Critic_Score:games$RatingTeens        0.0007187  0.0030320   0.237
##                                                        Pr(>|t|)    
## (Intercept)                                < 0.0000000000000002 ***
## games$Critic_Score                         < 0.0000000000000002 ***
## games$RatingEveryone10+                                   0.753    
## games$RatingMature                                  0.000000042 ***
## games$RatingTeens                                         0.134    
## games$Critic_Score:games$RatingEveryone10+                0.737    
## games$Critic_Score:games$RatingMature               0.000000181 ***
## games$Critic_Score:games$RatingTeens                      0.813    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.301 on 6285 degrees of freedom
## Multiple R-squared:  0.135,  Adjusted R-squared:  0.134 
## F-statistic: 140.1 on 7 and 6285 DF,  p-value: < 0.00000000000000022
```

```
##                                (Intercept) 
##                                     3.6932 
##                         games$Critic_Score 
##                                     0.0309 
##                    games$RatingEveryone10+ 
##                                    -0.0917 
##                         games$RatingMature 
##                                    -1.3085 
##                          games$RatingTeens 
##                                    -0.3291 
## games$Critic_Score:games$RatingEveryone10+ 
##                                    -0.0014 
##      games$Critic_Score:games$RatingMature 
##                                     0.0170 
##       games$Critic_Score:games$RatingTeens 
##                                     0.0007 
```


Video Game Global Sales by Critic Score and Rating


NQ Plot of Studentized Residuals, Residual Plots


Fits vs. Studentized Residuals, Residual Plots



Now let's fit an ANCOVA model predicting log(Global Sales) based on Rating, Critic Score, and the interaction of Rating and Critic Score. We see that both predictors (Rating and Critic Score) are statistically significant and that the interaction between the two predictors is statistically significant. The p values are, respectively, 0.00000000000000022, 0.00000016018, and 0.00000003378. Reminder that our significance levels here are measuring OVERALL significance between the predictors, and that once we stratify into individual pairings, the significance may not hold. From the summary of our ANCOVA model, we see that our overall model is statistically significant, with a p value of 0.00000000000000022. It has an adjusted R-squared value of 0.134, which means that it predicts 13.4% of the variance in games global sales. Unfortunately, this does not appear to be a particularly strong model. We see that of Everyone10+, Teen, and Mature, only sales data for Mature is statistically significantly different from that of Rating: Everyone (the default reference), with a p value = 0.000000042. We also see that 1) Critic Score and Rating: Everyone10+ do not have a statistically significant interaction (p value = 0.737), 2) Critic Score and Rating: Mature have a statistically significant interaction (p value = 0.000000181), 3) Critic Score and Rating: Teen do not have a statistically significant interaction (p value = 0.812). The ANCOVA function measures if these pairings behave statistically significantly different from the default, which is Rating: Everyone. We see that games with a rating of mature have global sales most strongly predicted by critic scores, with the highest coefficient in the model (0.04787). This indicates that the model predicts a 0.04787 increase in sales for every increase of 1 in critic scores. To see this visually, we graph a scatterplot of global sales by critic scores, with lines for each linear model (split into each of the four rating categories) overlayed on top. These lines predict log(global sales) based on critic score for each of the four rating categories. We see the line representing mature games (green line) has the steepest slope. We also see that it has the lowest intercept. This means that at low critic scores, Mature games sell the lowest relative to other ratings, and at high critic scores, Mature games sell the highest relative to other ratings. Intuitively, this does make sense, because adults (the target consumers of mature games) are more likely to be reading and informing their purchasing decisions from reviews. The lines for the other three ratings (Everyone, Everyone10+, and Teen) are all noticeably less steep but are very similar to each other. The order of steepness (high to low coefficient) is Teen, Everyone, and Everyone10+. As we go down the list, log(global sales) is less and less affected by critic scores. Lastly, we use myResPlots2 to verify that ANCOVA assumptions hold. Overall, our results looks pretty good. The normal quantile plot is linear so errors are normally distributed. In the plot of fits vs studentized residuals, we see mild evidence of heteroskedasticity. We do see some large outliers, but this is fine and unavoidable because there are so many games (data points).

GLM and Residual Plots

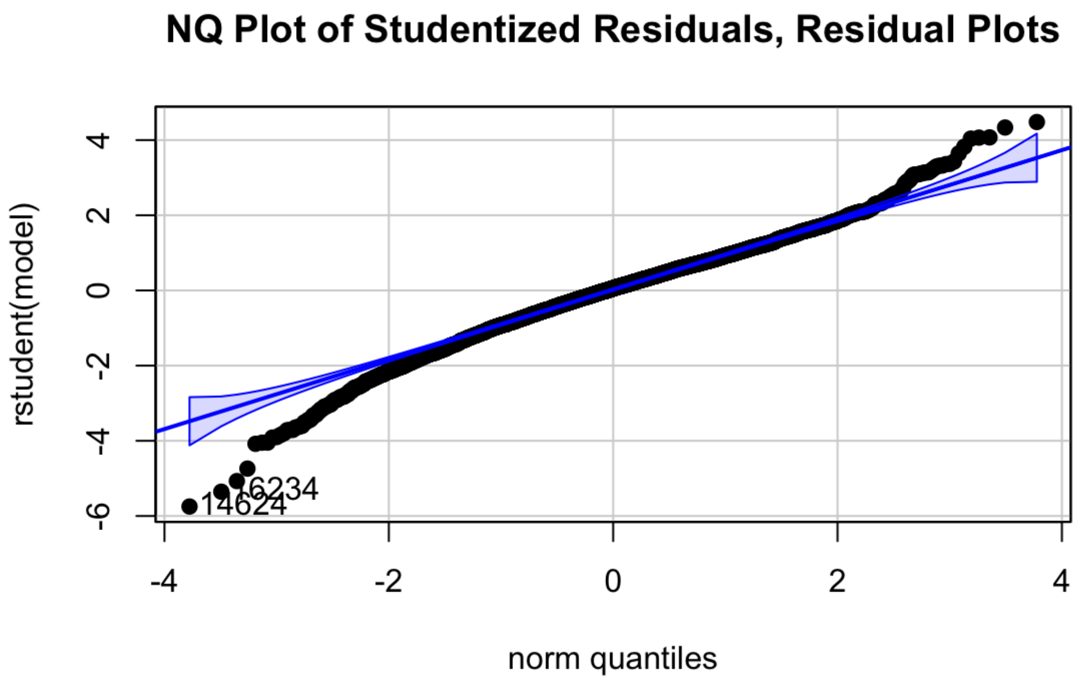
The plan for this multiple regression is to use the continuous predictors Critic Score, User Score, log of Critic Count, log of User Count and Year Released along with categorical predictors Rating, Publisher, Developer, Genre, and Platform to predict the log of Global Sales. We are not using NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales, User_Count, or Critic_Count in an attempt to reduce multicollinearity. The first step was to convert all the string predictors to factors in order to use them in our linear model, as well as condensing in order to reduce factor levels. The initial plan was to use best subsets regression but this ran into the issue that each level of each factor was a separate predictor and it was not computationally feasible to evaluate each subset of the *resultant 40+ predictors. Also, even after limiting the number of factors, any suggested model would be impossible to create without a new dataframe that had a column for each level of each factor. For these reasons we decided to make a GLM.*

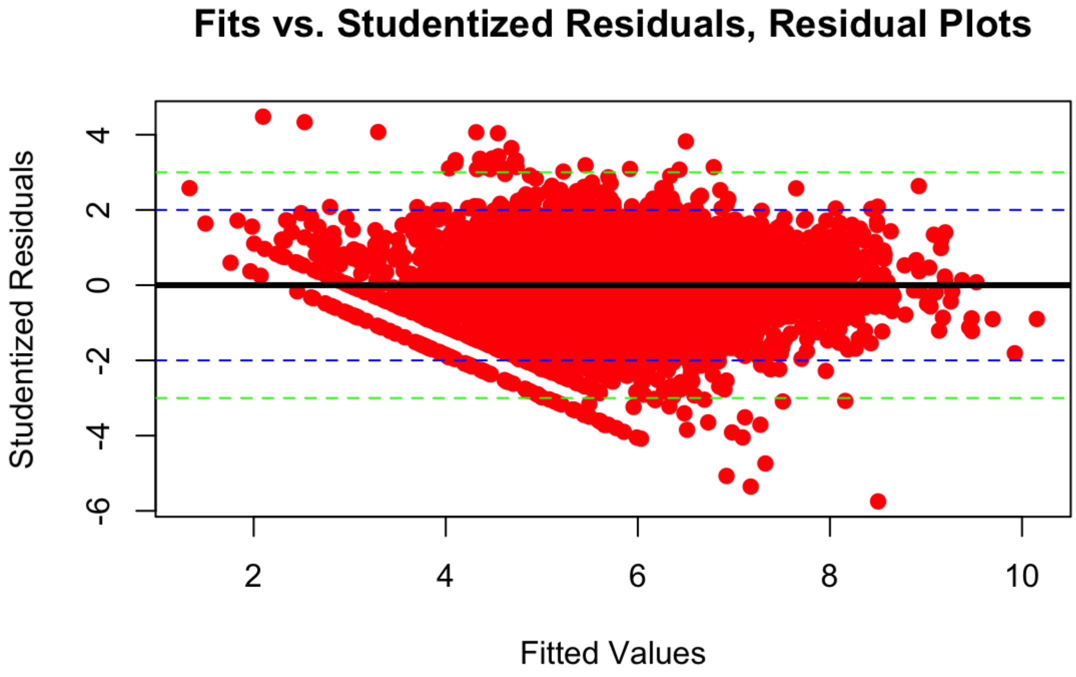
```
## Anova Table (Type III tests)
##
## Response: log_Global_Sales
##                   Sum Sq   Df   F value                 Pr(>F)    
## (Intercept)        426.5    1  505.6291 < 0.00000000000000022 ***
## Platform          2508.7    7  424.8305 < 0.00000000000000022 ***
## Year_of_Release    411.1    1  487.2852 < 0.00000000000000022 ***
## Genre               81.2    9   10.6991 < 0.00000000000000022 ***
## Publisher          286.2   12   28.2756 < 0.00000000000000022 ***
## Critic_Score        87.9    1  104.1925 < 0.00000000000000022 ***
## log_Critic_Count    43.3    1   51.2810    0.0000000000008929 ***
## User_Score          32.7    1   38.8202    0.0000000004951218 ***
## log_User_Count    1663.9    1 1972.3385 < 0.00000000000000022 ***
## Developer          112.6   12   11.1207 < 0.00000000000000022 ***
## Rating              85.9    3   33.9293 < 0.00000000000000022 ***
## Genre:Rating       100.3   27    4.4055    0.0000000000002342 ***
## Residuals         5244.7 6217                                  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

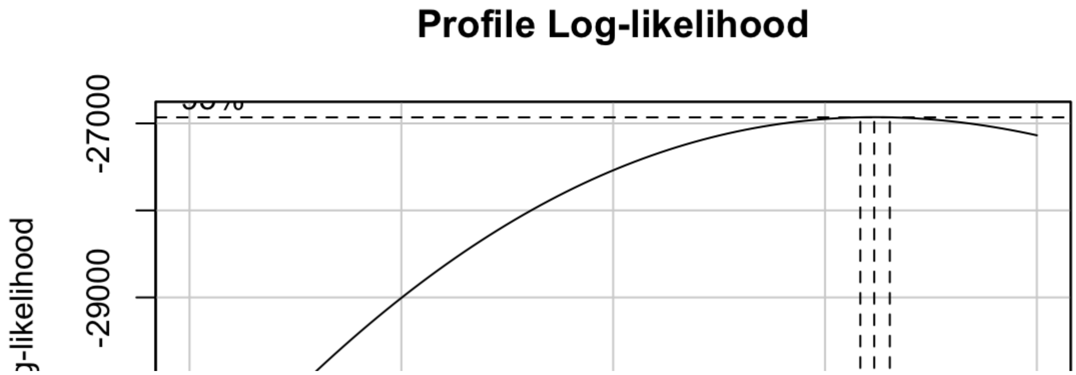
```
## 
## Call:
## lm(formula = log_Global_Sales ~ Platform + Year_of_Release + 
##     Genre + Publisher + Critic_Score + log_Critic_Count + User_Score + 
##     log_User_Count + Developer + Rating + Genre * Rating, data = games_mod)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.1002 -0.5484  0.0500  0.5938  4.0970 
## 
## Coefficients:
##                                         Estimate Std. Error t value
## (Intercept)                           157.629497   7.010056  22.486
## PlatformDS                              1.467347   0.256632   5.718
## PlatformGameboy                         1.100152   0.261323   4.210
## PlatformGamecube                        0.916588   0.257205   3.564
## PlatformPC                             -1.092423   0.255143  -4.282
## PlatformPlay Station                    1.403122   0.252649   5.554
## PlatformWii                             1.574354   0.256475   6.138
## PlatformXbox                            1.154043   0.253910   4.545
## Year_of_Release                        -0.077310   0.003502 -22.075
## GenreFighting                          -1.032189   0.466862  -2.211
## GenreMisc                               0.270706   0.102720   2.635
## GenrePlatform                          -0.034781   0.091995  -0.378
## GenreRacing                            -0.099540   0.084706  -1.175
## GenreRole-Playing                      -0.140740   0.130419  -1.079
## GenreShooter                           -1.238034   0.227854  -5.433
## GenreSimulation                         0.416370   0.115338   3.610
## GenreSports                             0.049678   0.079899   0.622
## GenreStrategy                          -0.416917   0.109912  -3.793
## PublisherAtari                         -0.519530   0.087407  -5.944
## PublisherCapcom                        -0.846899   0.103234  -8.204
## PublisherElectronic Arts               -0.047175   0.062302  -0.757
## PublisherKonami Digital Entertainment  -0.722454   0.091473  -7.898
## PublisherNamco Bandai Games            -0.542839   0.079238  -6.851
## PublisherNintendo                      -0.196988   0.084043  -2.344
## PublisherOther                         -0.610379   0.049334 -12.372
## PublisherSega                          -0.421705   0.073940  -5.703
## PublisherSony Computer Entertainment   -0.558132   0.072830  -7.663
## PublisherTake-Two Interactive          -0.296277   0.077312  -3.832
## PublisherTHQ                           -0.082154   0.074030  -1.110
## PublisherUbisoft                       -0.525458   0.070829  -7.419
## Critic_Score                            0.013441   0.001317  10.207
## log_Critic_Count                        0.155811   0.021758   7.161
## User_Score                             -0.071037   0.011401  -6.231
## log_User_Count                          0.588759   0.013257  44.411
## DeveloperEA Canada                     -0.441297   0.147595  -2.990
## DeveloperEA Sports                     -0.270776   0.151357  -1.789
## DeveloperEA Tiburon                    -0.235480   0.164739  -1.429
## DeveloperElectronic Arts               -0.116149   0.169500  -0.685
## DeveloperKonami                         0.042242   0.173205   0.244
## DeveloperNintendo                       0.487387   0.171455   2.843
## DeveloperOmega Force                    0.140785   0.166061   0.848
## DeveloperOther                         -0.261619   0.113454  -2.306
## DeveloperTraveller's Tales              0.586227   0.172711   3.394
## DeveloperUbisoft                        0.238369   0.157874   1.510
## DeveloperUbisoft Montreal               0.202557   0.159923   1.267
## DeveloperVisual Concepts                0.019537   0.172935   0.113
## RatingEveryone10+                       0.072867   0.088162   0.827
## RatingMature                           -0.571041   0.079232  -7.207
## RatingTeens                            -0.385843   0.077890  -4.954
## GenreFighting:RatingEveryone10+         0.479925   0.540226   0.888
## GenreMisc:RatingEveryone10+             0.148817   0.161670   0.921
## GenrePlatform:RatingEveryone10+        -0.410018   0.142492  -2.877
## GenreRacing:RatingEveryone10+          -0.118144   0.147154  -0.803
## GenreRole-Playing:RatingEveryone10+    -0.391998   0.168688  -2.324
## GenreShooter:RatingEveryone10+          0.479941   0.286219   1.677
## GenreSimulation:RatingEveryone10+      -0.298443   0.233239  -1.280
## GenreSports:RatingEveryone10+          -0.471322   0.147630  -3.193
## GenreStrategy:RatingEveryone10+        -0.283045   0.160689  -1.761
## GenreFighting:RatingMature              1.375330   0.488767   2.814
## GenreMisc:RatingMature                 -0.273594   0.298108  -0.918
## GenrePlatform:RatingMature              0.495201   0.542396   0.913
## GenreRacing:RatingMature               -0.249274   0.257229  -0.969
## GenreRole-Playing:RatingMature          0.083982   0.155233   0.541
## GenreShooter:RatingMature               1.316852   0.235027   5.603
## GenreSimulation:RatingMature           -0.076762   0.430803  -0.178
## GenreSports:RatingMature                0.300753   0.291901   1.030
## GenreStrategy:RatingMature              0.317358   0.225772   1.406
## GenreFighting:RatingTeens               1.277225   0.471458   2.709
## GenreMisc:RatingTeens                   0.305064   0.141552   2.155
## GenrePlatform:RatingTeens              -0.354969   0.161242  -2.201
## GenreRacing:RatingTeens                 0.356899   0.125716   2.839
## GenreRole-Playing:RatingTeens          -0.077907   0.143606  -0.543
## GenreShooter:RatingTeens                1.230479   0.238304   5.163
## GenreSimulation:RatingTeens             0.109514   0.143194   0.765
## GenreSports:RatingTeens                 0.251350   0.130698   1.923
## GenreStrategy:RatingTeens               0.114157   0.140934   0.810
##                                                    Pr(>|t|)    
## (Intercept)                          < 0.0000000000000002 ***
## PlatformDS                           0.000000011299131999 ***
## PlatformGameboy                      0.000025905262314149 ***
## PlatformGamecube                                 0.000368 ***
## PlatformPC                           0.000018833426971946 ***
## PlatformPlay Station                 0.000000029137521407 ***
## PlatformWii                          0.000000000884769448 ***
## PlatformXbox                         0.000005595097991031 ***
## Year_of_Release                      < 0.0000000000000002 ***
## GenreFighting                                    0.027078 *  
## GenreMisc                                        0.008425 ** 
## GenrePlatform                                    0.705386    
## GenreRacing                                      0.239987    
## GenreRole-Playing                                0.280566    
## GenreShooter                         0.000000057372386174 ***
## GenreSimulation                                   0.000309 ***
## GenreSports                                       0.534123    
## GenreStrategy                                     0.000150 ***
## PublisherAtari                        0.000000002936135062 ***
## PublisherCapcom                       0.000000000000000281 ***
## PublisherElectronic Arts                          0.448959    
## PublisherKonami Digital Entertainment 0.000000000000003328 ***
## PublisherNamco Bandai Games           0.000000000008053253 ***
## PublisherNintendo                                 0.019115 *  
## PublisherOther                        < 0.0000000000000002 ***
## PublisherSega                         0.000000012291155753 ***
## PublisherSony Computer Entertainment  0.000000000000020865 ***
## PublisherTake-Two Interactive                     0.000128 ***
## PublisherTHQ                                      0.267155    
## PublisherUbisoft                      0.000000000000134079 ***
## Critic_Score                          < 0.0000000000000002 ***
## log_Critic_Count                      0.000000000000892877 ***
## User_Score                            0.000000000495121789 ***
## log_User_Count                        < 0.0000000000000002 ***
## DeveloperEA Canada                                0.002802 ** 
## DeveloperEA Sports                                0.073666 .  
## DeveloperEA Tiburon                               0.152937    
## DeveloperElectronic Arts                          0.493215    
## DeveloperKonami                                   0.807331    
## DeveloperNintendo                                 0.004489 ** 
## DeveloperOmega Force                              0.396589    
## DeveloperOther                                    0.021147 *  
## DeveloperTraveller's Tales                        0.000692 ***
## DeveloperUbisoft                                  0.131127    
## DeveloperUbisoft Montreal                         0.205349    
## DeveloperVisual Concepts                          0.910056    
## RatingEveryone10+                                 0.408545    
## RatingMature                          0.000000000000638879 ***
## RatingTeens                           0.000000747560537226 ***
## GenreFighting:RatingEveryone10+                   0.374372    
## GenreMisc:RatingEveryone10+                       0.357346    
## GenrePlatform:RatingEveryone10+                   0.004022 ** 
## GenreRacing:RatingEveryone10+                     0.422085    
## GenreRole-Playing:RatingEveryone10+               0.020168 *  
## GenreShooter:RatingEveryone10+                    0.093626 .  
## GenreSimulation:RatingEveryone10+                 0.200749    
## GenreSports:RatingEveryone10+                     0.001417 ** 
## GenreStrategy:RatingEveryone10+                   0.078211 .  
## GenreFighting:RatingMature                        0.004910 ** 
## GenreMisc:RatingMature                            0.358776    
## GenrePlatform:RatingMature                        0.361284    
## GenreRacing:RatingMature                          0.332547    
## GenreRole-Playing:RatingMature                    0.588523    
## GenreShooter:RatingMature             0.000000021973196958 ***
## GenreSimulation:RatingMature                      0.858584    
## GenreSports:RatingMature                          0.302897    
## GenreStrategy:RatingMature                        0.159877    
## GenreFighting:RatingTeens                         0.006765 ** 
## GenreMisc:RatingTeens                             0.031190 *  
## GenrePlatform:RatingTeens                         0.027740 *  
## GenreRacing:RatingTeens                           0.004541 ** 
## GenreRole-Playing:RatingTeens                     0.587490    
## GenreShooter:RatingTeens              0.000000249914429854 ***
## GenreSimulation:RatingTeens                       0.444422    
## GenreSports:RatingTeens                           0.054509 .  
## GenreStrategy:RatingTeens                         0.417967    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.9185 on 6217 degrees of freedom
## Multiple R-squared:  0.5732, Adjusted R-squared:  0.5681 
## F-statistic: 111.3 on 75 and 6217 DF,  p-value: < 0.00000000000000022
```

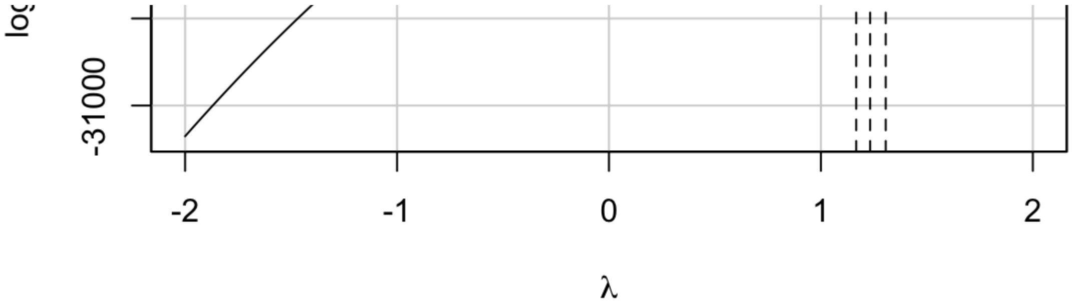
We can see that each predictor in our model is statistically significant since each has a p-value well below any reasonable alpha. However, this does not mean that each level alone of each categorical predictor is statistically significant, and in fact we observe many levels such as the Racing level of Genres, which is not. The R-squared of the model is 0.5661 so it is able to explain 56.61% of the variance in log global sales and the model as a whole is statistically significant since it has a p-value that is well below any reasonable alpha. We chose this as our final model as it includes all reasonable possible predictors (avoiding collinearity between say global sales and log global sales, and avoiding vastly reducing degrees of freedom with name as a factor) and all of these predictors are statistically significant. The interaction we choose to include was based on the hunch that it would be significant, in other words it seemed likely to use that certain genres would be better with certain ratings.

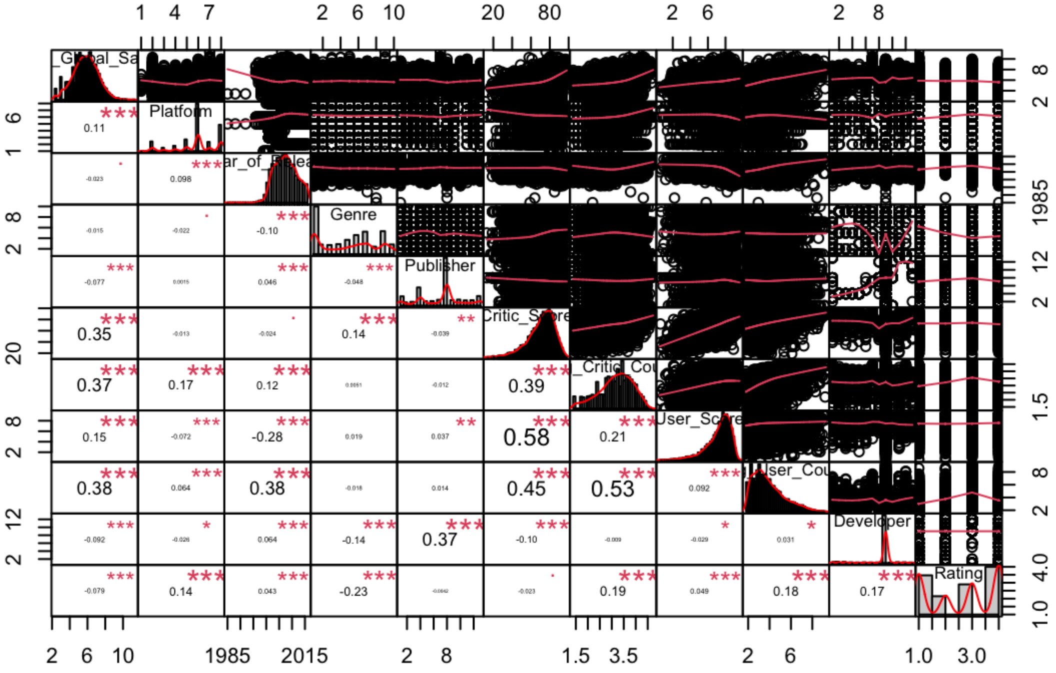
It is not feasible to go through each categorical predictor level individually, but reported in the summary of the model is an estimate by how much this predictor effects the value of log of Global Sales. For example, since Rating Mature has the value -0.571041, it suggests the log of the global sales of a game with the mature rating will be 0.571041 less than that of a game with with a Everything rating. If this value was positive, as for the rating Everything10+, it would suggest a relative increase to log of Global Sales. For our continuous predictors, all Critic Score and log Critic Count had a slight positive effect on log of Global Sales, User Score and Year of Release had a slightly negative effect, and User Count had a significant positive effect. The interaction between Genre and Rating is insignificant for many levels, but for some, it is significant. For example, the interaction between the Genre of Shooters and the Rating of Teen suggests Shooters are much more successful for games rated Teens than they are for games rated Everyone. This follows common sense.


NQ Plot of Studentized Residuals, Residual Plots


Fits vs. Studentized Residuals, Residual Plots


Profile Log-likelihood





The studentized residuals plot reveals that log global sales is not quite linear. This violates an assumption in our model and questions the validity of our results. Box cox suggests a transformation of 1.232323 which is within the range around 1 such that we can assume no further exponential transformation would solve our non-linearity. The fits vs studentized residuals plot displays mild signs of possible heteroskedasticity. The values take a form of a blob around 0 for the most part, but in the bottom left corner there is a clear abnormality. This abnormality is due to the prevalence of exactly equal values in the low range of log of global sales. We believe this is inherent in the data set due to a flaw in the data collection process. The studentized residuals plot is not abnormal enough for serious concern. It is clear in our last plot that this model has some serious multicollinearity. There are multiple statistically significant correlations between our predictors. This reduces the precision of our estimated coefficients and in this manner reduces the statistical power of our model.

Web Scraping

```
## [1] "# of Games in the Top 100 List of Both IGN and Metacritic Scores"
```

```
## [1] 28
```

```
## [1] "# of Games in the Top 100 List of Both IGN and Metacritic User Scores"
## [1] 14
```

```
## [1] "# of Games in the Top 100 List of Both Metacritic Scores and User Scores"
```

```
## [1] 24
```

```
## [1] "Games in the Top 100 of IGN's Ranking, Metacritic Scores, and Metacritic User Sc
ores"
```

```
##  [1] "Burnout 3: Takedown"
##  [2] "Castlevania: Symphony of the Night"
##  [3] "Final Fantasy VII"
##  [4] "Half-Life"
##  [5] "Half-Life 2"
##  [6] "Metal Gear Solid"
##  [7] "Metroid Prime"
##  [8] "Resident Evil 4"
##  [9] "Star Wars: Knights of the Old Republic"
## [10] "The Last of Us"
## [11] "The Legend of Zelda: A Link to the Past"
## [12] "The Witcher 3: Wild Hunt"
```

As an interesting extra step, we were curious to see how Metacritic scores would compare to another video game ranking retrieved off the internet. So, we web scraped a ranking of the top 100 video games from an IGN article. We then retrieved the top 100 games from our data set by critic score and did the same thing for user score. It was interesting to see how much these three lists (IGN ranking, Metacritic critic ranking, and Metacritic user ranking) had in common. As shown in the output, out of the 100 games, IGN and Metracritic critics had 28 games in common. The critic and user rankings had 24 games in common, and the user ranking and IGN only had 14 games in common. Finally, we found the list of games that made the ultimate cut: they were featured in the top 100 IGN games, Metacritic ranking, AND user ranking. This list is shown in the output and features famous games like “Star Wars” and the “Legend of Zelda”.

Conclusion

Our analysis has resulted in mixed results. We were able to establish that games rated under Teen sold more than games rated Teen or above. This may be due to the fact that the lower the rating, the larger the potential audience of a game. We also found a significant correlation between our continuous variables. Critic and user, count and score, were all correlated positively with global sales. This made a lot of sense. Both these results were supported with bootstrapping. The correlations were firmly supported by permutation testing. A one way ANOVA test revealed that global sales are statistically significantly different across video game ratings. An ANCOVA test revealed that ratings and critic scores interact significantly when predicting global sales, and that Mature rating video games have global sales that are most affected by critic scores than other ratings. For both the ANOVA and ANCOVA test, we verified that our assumptions held true. We created a GLM which purported to significantly predict global sales with relatively large strength, but plots revealed that this model may not be entirely valid due to the non-linearity of log global sales, and some abnormalities among collection of the data.