Final Project S&DS 230

Analysis of Sales and Rating Data (1980-2020)

A11 Squad

Introduction:

The purpose of this statistical analysis was to gain insights regarding trends occurring within the video game industry. The dataset utilized for this study was downloaded (https://www.kaggle.com/datasets/akshatchowdhary/kagglevideogames) and cleaned before being statistically assessed using various procedures ranging from t-tests and ANOVA to multiple regression and ANCOVA. The variables consistently explored throughout this analysis included the sale of different video game units in various markets and the ratings given these video games by both users and professional critics. Dependent variables considered throughout this investigation included the type of console for which the video game was intended, the genre of the video game, and the rating of the video game in terms of intended age demographic. Putative relationships between user and critic ratings and sales in various markets were of particular interest, and this was the central question addressed throughout this analysis. Interactions between variables were also noted in the construction of a final model containing predictors for the success of a video game in global markets. Such an analysis to determine predictors of market success could be used to assess the utility of focus groups of either critics or users before making a large investment in distributing a video game globally. Furthermore, this analysis could be used by video game companies to inform which video game genres, platforms, or ratings (proxy for age groups) to target in order to strategically recoup their investment.

Data Included in Statistical Analysis:

Categorical Variables:

- **Platform (Console):** Wii, DS, Xbox 360, PlayStation (1-4, Vita, and Portable), Xbox, 3DS, PC, Xbox One, Wii U, GameCube, and Game Boy Advance
- **Genre:** Sports, Racing, Platform, Misc., Action, Puzzle, Shooter, Simulation, Adventure, Strategy, Fighting, and Role-Playing
- **Rating (Age Demographic):** Everyone, Everyone 10+, Teen, and Mature 17+
- **Console Family:** Nintendo, Xbox, PlayStation, and PC

Continuous Variables:

- **North American Sales** in millions of units (NA_Sales)

- **European Sales** in millions of units (EU_Sales)
- **Global Sales** in millions of units (Global_Sales)
- **Critic Score** on scale of 1-10 as voted on by professional video game critics (Critic_Score)
- **User Score** on scale of 1-10 as voted on by video game users (User_Score)

Data Cleaning Procedures:

Step 1: Handling of Missing and Irrelevant Data

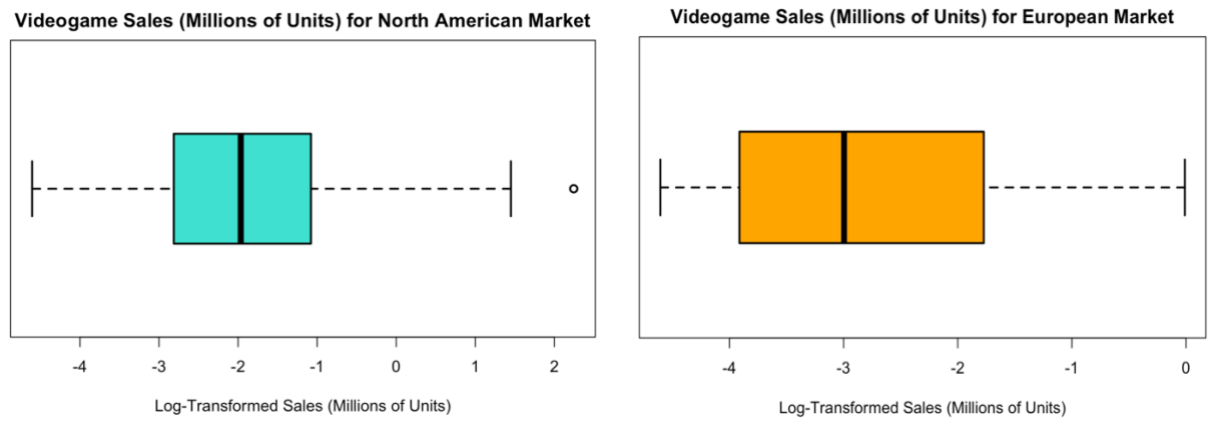
The dataset was first imported using the read.csv() function the aforementioned variables were included in a subset referred to as "clean_data". Missing values in the dataset were then addressed (e.g. blank values in Genre were set to "Misc." and "Rating Pending" values in Rating were set to NA). Furthermore, video games rated as "adult only" or designed for the SEGA Dreamcast console were infrequent enough to reasonably remove from "clean_data". All rows containing missing (NA) values were then omitted from "clean_data" using the na.omit() function. This preliminarily cleaned data was also queried to ensure that no duplicates existed in a process known as matching.

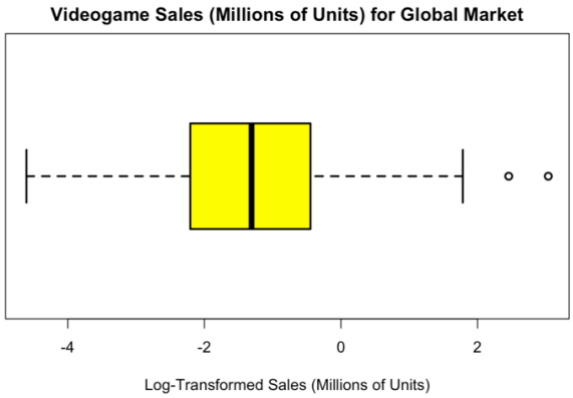
Step 2: Grouping of Data into Broader, Useful Categories

The video games were then assorted into the broader group of "Console_Family" (PlayStation, Xbox, or PC) and "Console_Type" (Portable or Home console). The names of the various consoles and ratings were also manually recoded with more intuitive, useful names. "Critic_Score", initially on a continuous scale of 1-100, was divided by 10 in order to provide a more direct comparison to "User_Score", initially on a continuous scale of 1-10. Further data cleaning, such as text character replacement, was not necessary due to the relatively clean nature of the initial data (consistent use of capitalization and other such conventions throughout). The final dimensions of the cleaned dataset were 6930 rows and 11 columns (including video game name and year of release).

Statistical Results:

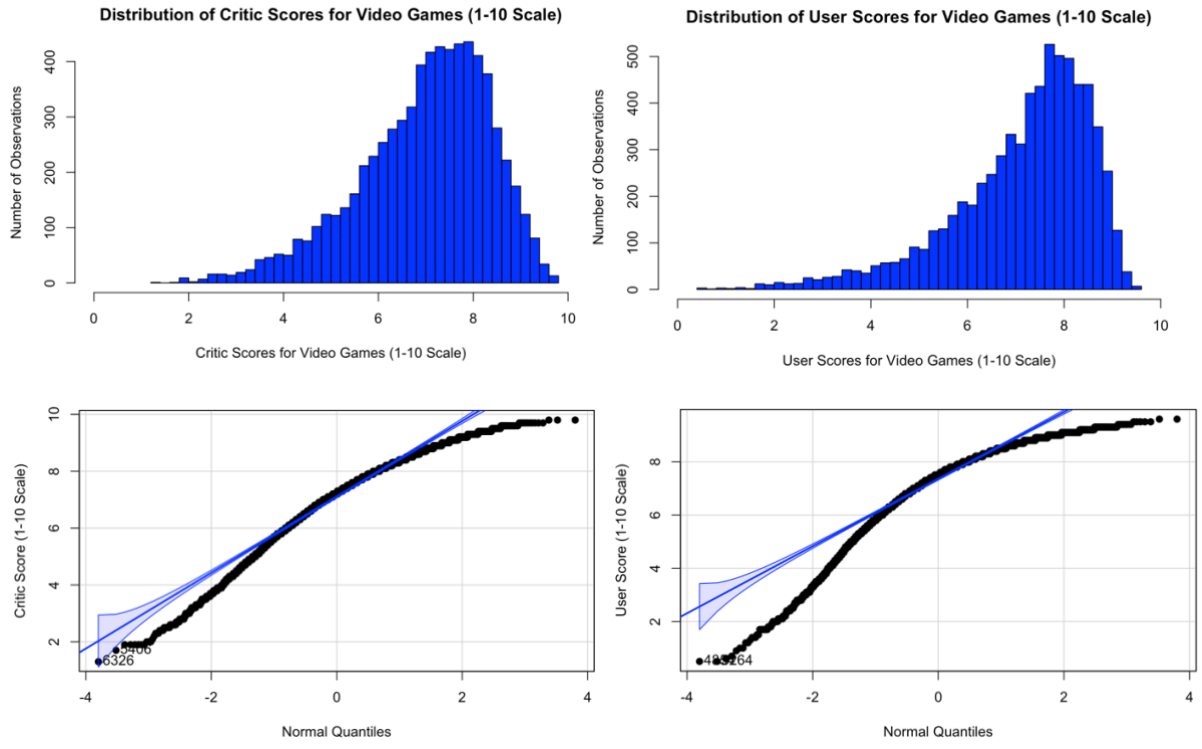
Boxplots of Video Game Sales in North American, European, and Global Markets:





The sales data (in millions of units) for the European, North American, and Global markets are displayed on a log-transformed scale to address the inherent right-skew of this data, which stems from the existence of some blockbuster video games that far exceed the vast majority of the competition. Additionally, the relatively larger market for video games in North America compared to Europe is demonstrated. Following the log-transformation, the data appears fairly normally distributed although, perhaps, slightly right-skewed. There remain only a few large outliers (greater than $1.5 \times IQR + 3Q$). These displays were generated using the boxplot() function and "car" library.

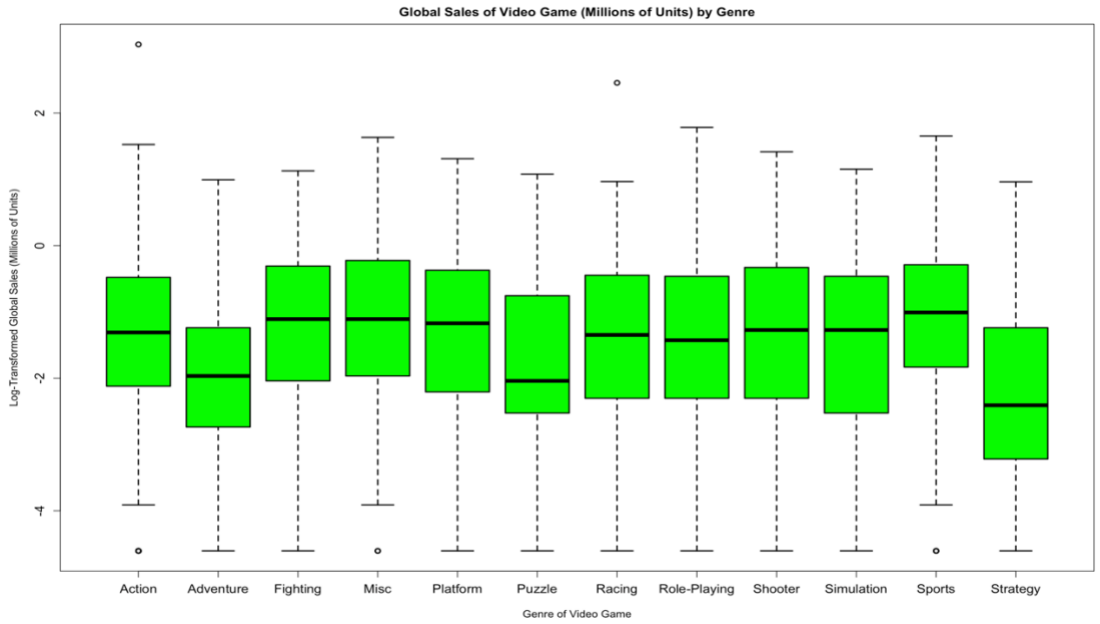
Histograms and Normal Quantile Plots of Critic and User Scores (1-10 Scales) for Video Games:

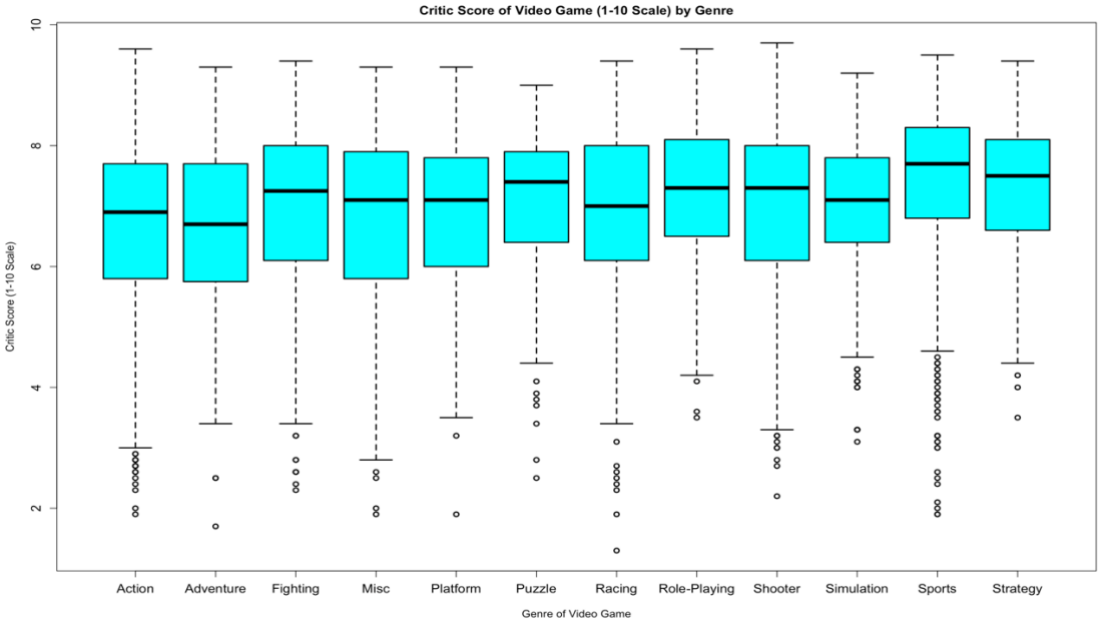


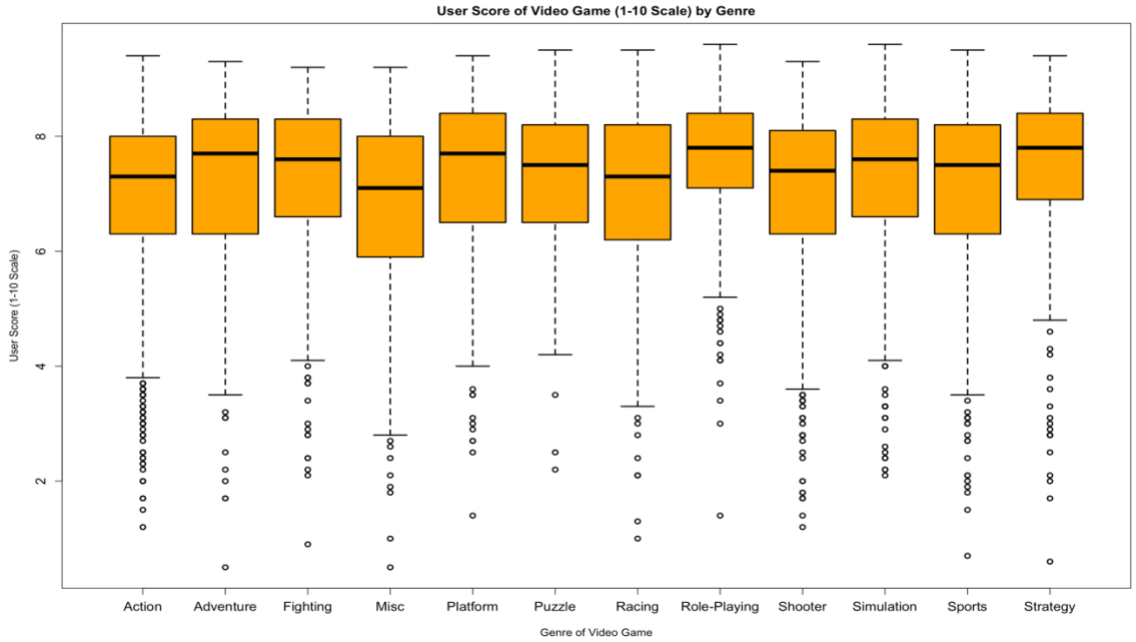
The upper panels contain histograms displaying the distribution of the critic and user assigned scores for the various video games in the cleaned dataset. Both illustrate a clear left-skewed pattern (mean/median well above theoretical mean/median of 5 implied by a

1-10 scale). This left-skewed pattern is stronger for the user scores, suggesting more generous scores compared to those assigned by the professional critics. These tail-end deviations from normality amongst both the user and critic assigned scores are conveyed by the normal quantile plots (lower panels), which strongly deviate from the confidence bands, especially at the tail ends. The hist() and qqPlot() functions were utilized to construct these displays.

Boxplots of Global Video Game Sales and Critic/User Scores as a Function of Genre:

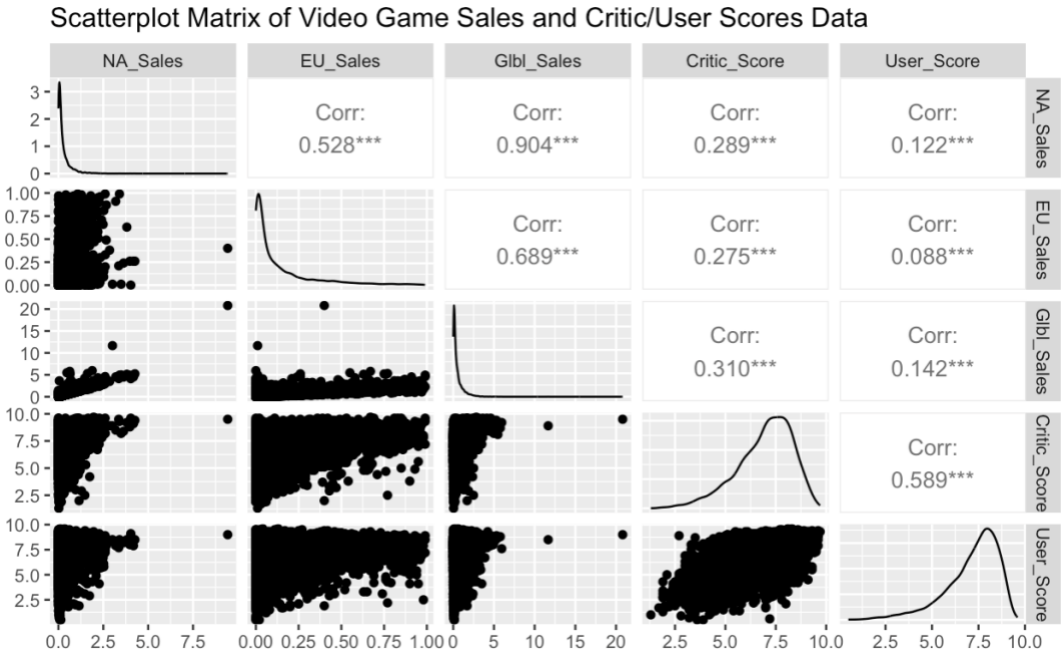






The uppermost boxplot displays the spread and mean of log-transformed sales (millions of units) for video games as a function of genre while the middle and lower panels display the spread and median of critic and user scores (1-10 scales), respectively, as a function of genre. These latter two boxplots contain many low outliers, again demonstrating the left-skewed nature of this data. At a simple, observational level, it appears that the user and critic scores are often fairly different, and neither appears to be a particularly clear predictor of which genres will be best-sellers on the global market. These displays were generated using the boxplot() function.

Scatterplot Matrix of Continuous Variables to be Assessed in this Study:



The ggpairs() function from the "GGally" library was used to create a scatterplot matrix, allowing for visual assessment of the relationships between sales (millions of units) in various markets (North American, European, and Global) and the critic and user scores (1-10 scales). The correlations and associated level of statistical significance (number of asterisks next to Pearson r coefficient) are also displayed for each of these relationships. Of particular interest are the apparently linear relationships between the user and critic scores, the lack of clear concordance (linearity) between North American and European sales figures, and the seemingly stronger relationship between sales (in all three markets) and critic score as opposed to user score.

Welch's Two-Sample t-Tests Comparing Means of Continuous Variables:

Global vs. European Sales:

```
        Welch Two Sample t-test

data:  clean_data$Global_Sales and clean_data$EU_Sales
t = 42.515, df = 7549.6, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 0.3505123 0.3957365
sample estimates:
mean of x mean of y
0.5071775 0.1340531
```

Global vs. North American Sales:

```
        Welch Two Sample t-test

data:  clean_data$Global_Sales and clean_data$NA_Sales
t = 23.321, df = 10650, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 0.2037958 0.2544136
sample estimates:
mean of x mean of y
0.5071775 0.2780728
```

European vs. North American Sales:

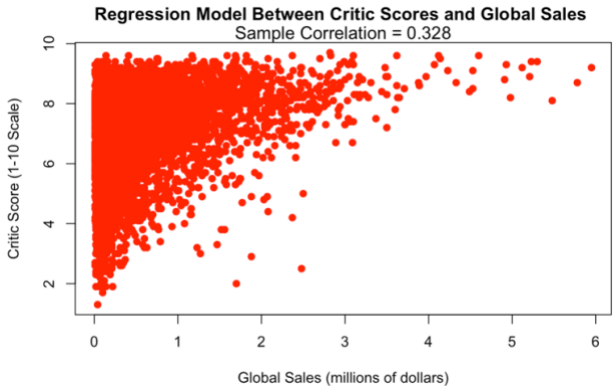
```
        Welch Two Sample t-test

data:  clean_data$EU_Sales and clean_data$NA_Sales
t = -26.294, df = 9269.3, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -0.1581313 -0.1299081
sample estimates:
mean of x mean of y
0.1340531 0.2780728
```

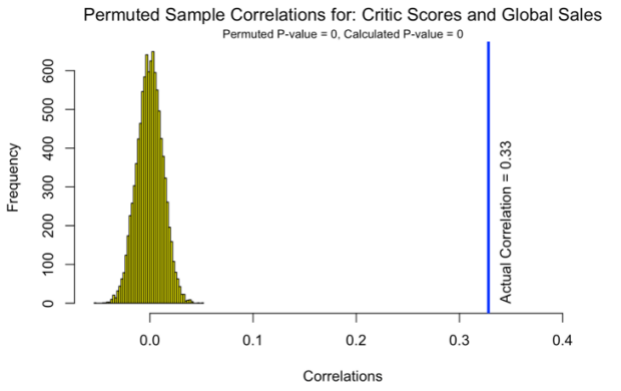
Critic Score vs. User Score:

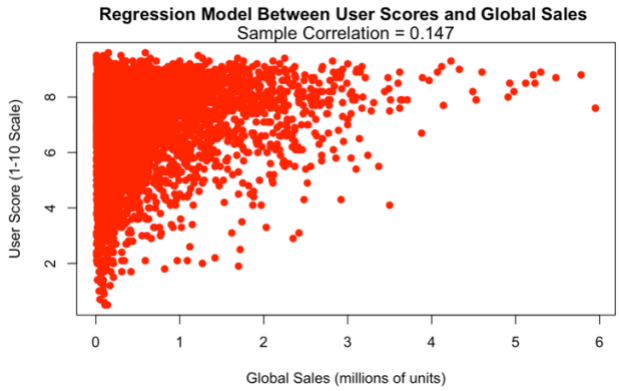
```
        Welch Two Sample t-test

data:  clean_data$Critic_Score and clean_data$User_Score
t = -8.2365, df = 13143, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -0.2658224 -0.1391548
sample estimates:
mean of x mean of y
 6.956783  7.159272
```

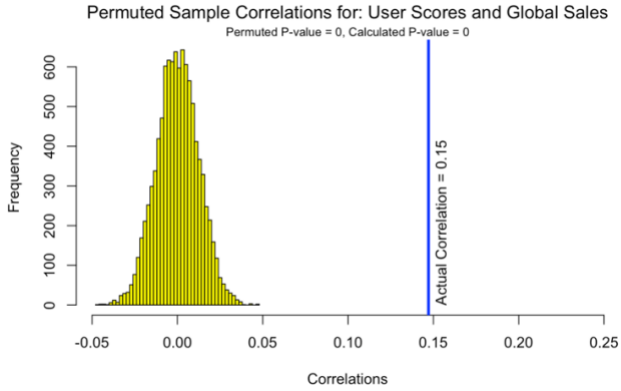
Each of the t-tests performed produced a statistically significant result at the 99% confidence level ($p < 0.01$), providing sufficient evidence to reject the null hypothesis (no statistically significant difference in means of two tested groups) in each case. Therefore, none of the 99% confidence intervals calculated by these t-tests contained the value zero, indicating that each of the groups had distinct means at a highly statistically significant level. This conclusion indicates that none of the tested variables are redundant, and that sales in each of the three markets (North American, European, and Global) and both scores (User and Critic) should be further investigated using other statistical tests. This does not, however, indicate that these variables are not collinear (this is investigated later) despite demonstrating they are not redundant. The t.test() function was utilized for this analysis.

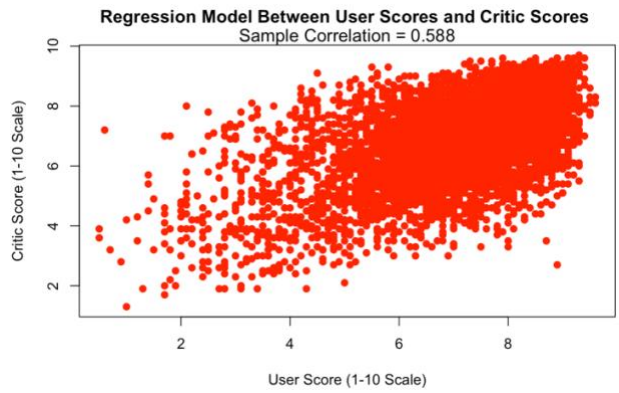
Correlation Analysis and Permutation Test (Sample Correlations) for Continuous Variables:

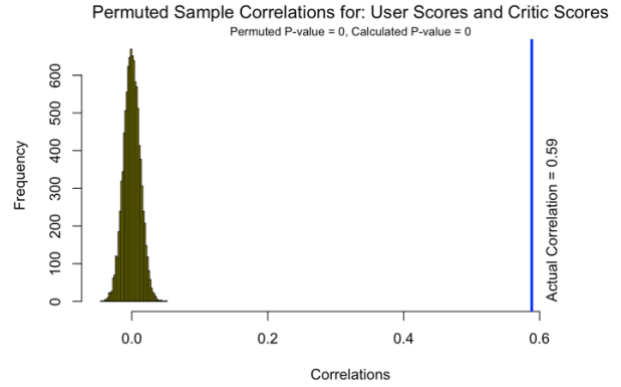










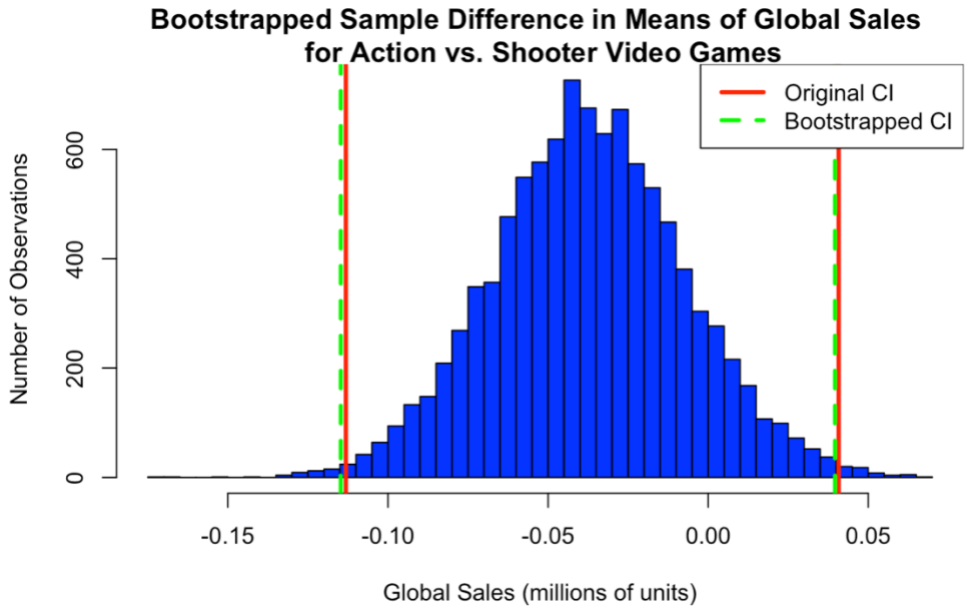


Large outliers for Global sales were first removed in order to avoid false linearity (can stem from influential points with disproportionate impact on outcome of regression). Correlations were then calculated between critic/user scores and global sales and between the user and critic scores. Critic scores were found to be a stronger predictor (r = 0.328) of global sales than user scores (r = 0.147). A fairly high correlation (r = 0.588) was found between the user and critic scores, suggesting a basic degree of agreement between these video game rating metrics. The non-linear shape the critic/user scores and global sales stems from a disproportionately high amount of video games having very low global sales (strongly right-skewed), which can not effectively be addressed by a logarithmic or square root transformation in the correlation analysis (both resulted in minimal changes in r

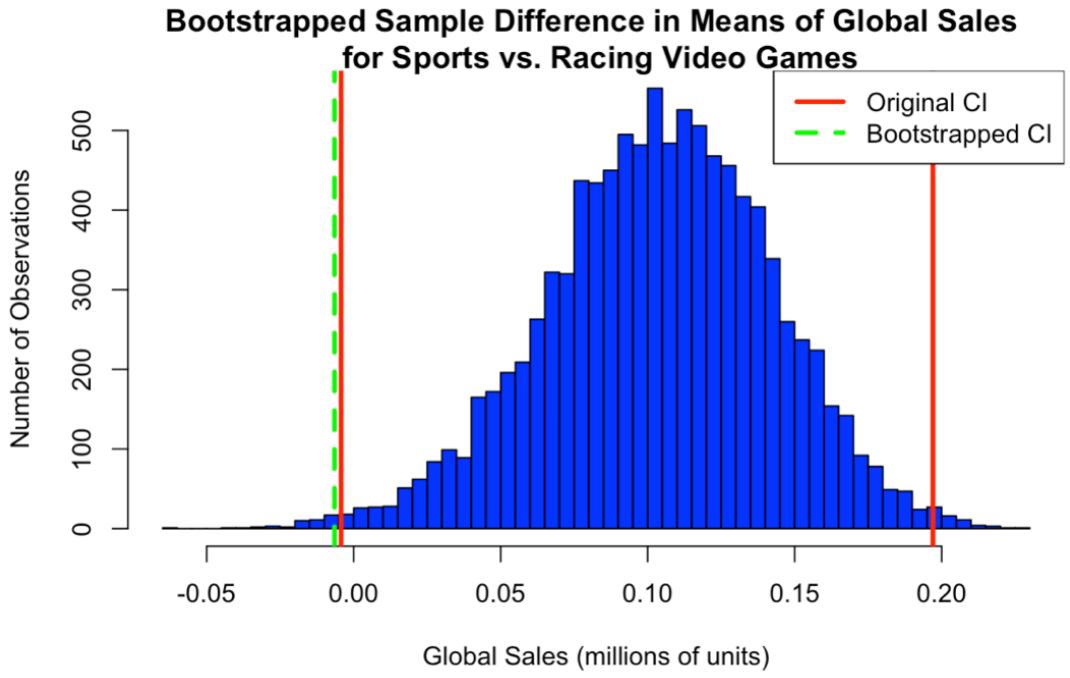
value). This lack of linearity would suggest that a simple correlation may not be the most useful way to characterize the relationships between user/critic score and global sales. The corresponding permuted sample correlations for each of these correlation analyses on the right-side panels. These histograms of calculated sample correlations do not overlap with the actual calculated correlation coefficients in any of the three cases (p-value essentially equal to 0 to at least seven decimal places), indicating that these correlations are statistically significant in that they likely did not result from random chance. This is most true for the permuted correlations of user vs. critic scores and least true for user scores vs. global sales. These displays were generated using modified versions of the myCor() and permCor() functions.

Bootstrap Analysis of Relationship Between Selected Genres and Global Sales:

```
        Welch Two Sample t-test

data:  clean_data$Global_Sales[clean_data$Genre == "Action" | clean_data$Genre == "Shooter"] by
clean_data$Genre[clean_data$Genre == "Action" | clean_data$Genre == "Shooter"]
t = -1.2553, df = 1915.4, p-value = 0.2095
alternative hypothesis: true difference in means between group Action and group Shooter is not
equal to 0
99 percent confidence interval:
 -0.11475443  0.03960361
sample estimates:
 mean in group Action mean in group Shooter
            0.4966963             0.5342717

 0.5% 99.5%
-0.11  0.04
```



```
        Welch Two Sample t-test

data:  clean_data$Global_Sales[clean_data$Genre == "Sports" | clean_data$Genre == "Racing"] by
clean_data$Genre[clean_data$Genre == "Sports" | clean_data$Genre == "Racing"]
t = -2.754, df = 1139.9, p-value = 0.005981
alternative hypothesis: true difference in means between group Racing and group Sports is not
equal to 0
99 percent confidence interval:
 -0.200230711 -0.006524714
sample estimates:
mean in group Racing mean in group Sports
           0.4897853            0.5931630

 0.5% 99.5%
  0.0   0.2
```



Welch's two-sample t-tests were performed to discern the impact of genre (specifically for the pairs of action-shooter and sports-racing, deemed to likely be similar) on global sales. The 99% confidence interval for the action-shooter pair was not statistically significant (p = 0.2095) and overlapped with zero while the 99% confidence interval for the sports-racing pair was statistically significant (p = 0.005981) and did not overlap with zero. In both instances, this theoretical confidence interval calculated using Welch's t-test was compared to 10,000 bootstrapped sample differences (with replacement) in the means of global sales as a function of the respective genres (histograms). In both instances, the bounds were roughly equivalent, suggesting a relatively high degree of certainty in the results generated via the t-tests. These displays were generated using the hist(), abline(), t.test() functions.

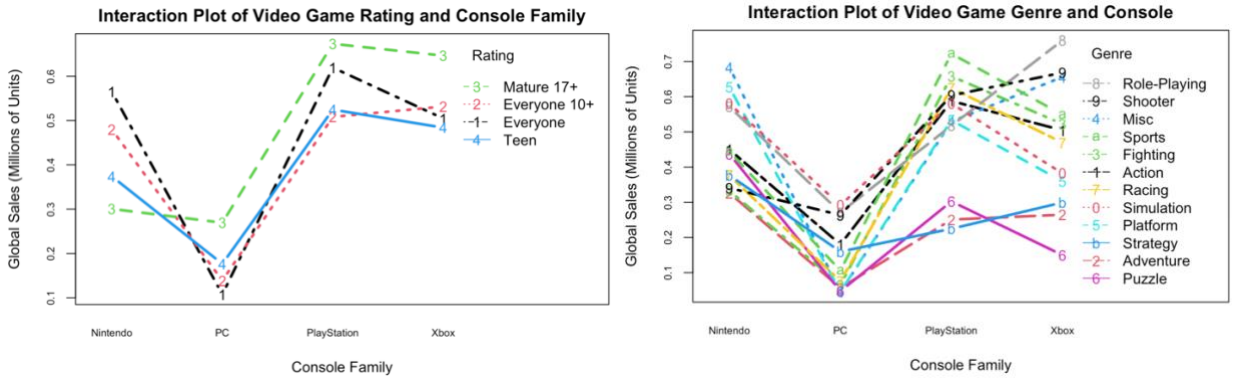
Multiple Regression to Determine Optimal Predictors of Video Game Global Sales:

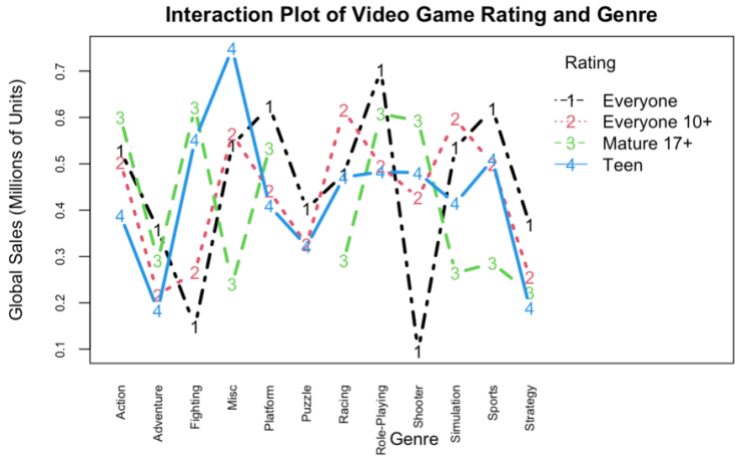
Plan:

Statistically significant interactions between categorical variables (console,

rating/age appropriateness, and genre) were first investigated via interaction plots. The statistical significance of these interactions was then quantitatively determined via type III ANOVAs. Backwards stepwise regression was then performed in which all potentially statistically significant categorical and continuous variables in addition to any statistically significant interaction effects (per interaction plots and ANOVAS) were implemented into an initial multiple regression model (e.g. Critic/User scores, Console, Rating, and Genre)and then individually removed in reverse order of statistical significance until an optimal model was determined. The criteria used to assess each iteration of the multiple regression model was the p-value of the given effect, and no main effects would be removed before all interaction effects had been removed.

Results (Interaction Plots and ANOVAs):





The interaction plots for the categorical variable pair rating-console demonstrated minimal interaction effects (few intersections) with respect to global sales while the pairs genre-console and rating-genre each showed some potential interaction effects (some intersections). Intuitively, one might expect certain genres to be more popular on certain consoles (e.g. shooter on PC or simulation on Nintendo) and certain ratings to be more aligned with certain genres (e.g. fighting and mature 17+ or sports and everyone).

ANOVA for Console-Rating Interaction:

```
Anova Table (Type III tests)

Response: Global_Sales
               Sum Sq   Df  F value                 Pr(>F)
(Intercept)    250.63    1 548.7034 < 0.00000000000000022 ***
Console         23.31    3  17.0132      0.00000000005275 ***
Rating          14.14    3  10.3214      0.00000089494138 ***
Console:Rating  15.01    9   3.6511              0.000145 ***
Residuals     3002.84 6574
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA for Genre-Rating Interaction:

```
Anova Table (Type III tests)

Response: Global_Sales
             Sum Sq   Df F values      Pr(>F)
Genre         19.04   11   3.7459 0.000023124 ***
Rating        12.35    3   8.9111 0.000006867 ***
Genre:Rating  28.99   32   1.9605   0.0009749 ***
Residuals   3023.17 6543
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA for Console-Genre Interaction:

```
Anova Table (Type III tests)

Response: Global_Sales
               Sum Sq   Df  F value                 Pr(>F)
(Intercept)     82.96    1 183.3157 < 0.00000000000000022 ***
Console         19.56    3  14.4039        0.000000002375 ***
Genre           20.61   11   4.1399        0.000004106198 ***
Console:Genre   45.14   33   3.0227        0.000000015013 ***
Residuals     2960.67 6542
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVAs intended to interrogate the significance of the interactions between the categorical pairs (console-rating, genre-rating, and console-genre) displayed highly statistically significant results (p approximately equal to 0) for all the main and interaction effects, indicating that each of these effects should be tested in subsequent backwards stepwise regression.

Results (Manual Backwards Stepwise Backwards Regression):

Initial Multiple Regression Model:

```
m1 <- lm(Global_Sales ~ Critic_Score + User_Score + Console + Rating + Genre + Console*Rating + Console*Genre + Critic_Score*User_Score + User_Score*Genre +
Critic_Score*Genre + Console*User_Score + Console*Critic_Score + Rating*Critic_Score + Rating*User_Score + Rating*Genre)
```

```
Anova Table (Type III tests)

Response: Global_Sales
                        Sum Sq   Df F values                 Pr(>F)
Critic_Score              2.40    1   6.3189              0.0119702 *
User_Score               10.64    1  28.0405         0.00000012273 ***
Console                   8.65    3   7.5946         0.00004563106 ***
Rating                    5.48    3   4.8142              0.0023774 **
Genre                     5.98   11   1.4331              0.1504278
Console:Rating           14.95    9   4.3763         0.00001021780 ***
Console:Genre            29.48   33   2.3542         0.00001982833 ***
Critic_Score:User_Score  29.42    1  77.5272 < 0.00000000000000022 ***
User_Score:Genre          9.44   11   2.2618              0.0095910 **
Critic_Score:Genre       14.05   11   3.3653              0.0001180 ***
User_Score:Console        8.01    3   7.0322              0.0001021 ***
Critic_Score:Console     15.10    3  13.2676         0.00000001243 ***
Critic_Score:Rating      11.62    3  10.2098         0.00000105241 ***
User_Score:Rating         3.38    3   2.9707              0.0305567 *
Rating:Genre             16.38   32   1.3487              0.0907709 .
Residuals              2451.88 6461
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.616 on 6461 degrees of freedom
Multiple R-squared:  0.2132,    Adjusted R-squared:  0.1976
F-statistic: 13.67 on 128 and 6461 DF,  p-value: < 0.00000000000000022
```

Final Multiple Regression Model:

```
m2 <- lm(Global_Sales ~ Critic_Score + User_Score + Console + Rating + Genre + Console*Rating + Console*Genre + Critic_Score*User_Score + Console*User_Score +
Console*Critic_Score + Rating*Critic_Score)

Anova Table (Type III tests)

Response: Global_Sales
                         Sum Sq   Df F value                Pr(>F)
(Intercept)               18.02    1 47.1053  0.000000000007348211 ***
Critic_Score               4.02    1 10.5171              0.001189 **
User_Score                23.82    1 62.2623  0.000000000000003501 ***
Console                    9.48    3  8.2615  0.000017504726601758 ***
Rating                    16.32    3 14.2174  0.000000003117275497 ***
Genre                     18.40   11  4.3736  0.000001447556019853 ***
Console:Rating            15.90    9  4.6183  0.000004133472070275 ***
Console:Genre             32.42   33  2.5680  0.000002190834377742 ***
Critic_Score:User_Score   32.04    1 83.7424 < 0.00000000000000022 ***
User_Score:Console        12.65    3 11.0237  0.000000323618575877 ***
Critic_Score:Console      17.28    3 15.0526  0.000000000922719289 ***
Critic_Score:Rating       11.49    3 10.0093  0.000001406245473780 ***
Residuals               2493.47 6518
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6185 on 6518 degrees of freedom
Multiple R-squared:  0.1998,    Adjusted R-squared:  0.1911
F-statistic: 22.92 on 71 and 6518 DF,  p-value: < 0.00000000000000022
```

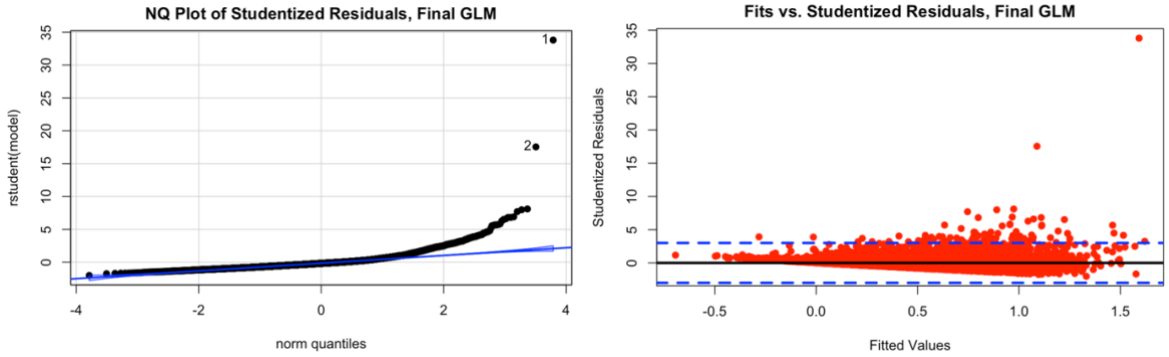
Between the initial and final multiple regression models, the interaction effects between the pairs rating-genre, user_score-rating, critic_score-genre, and user_score-genre due to their lack of statistical significance ($p > 0.01$ or 99% confidence level). This process involved removing only one interaction-effect at a time as each removal of an effect impacted the p-value of all the other effects. Additionally, no main effects were removed as main effects should only be removed after all interaction-effects are removed in backwards stepwise regression. The multiple R-squared value actually decreased from the initial model (0.2132) to the final model (0.1998), but this often results from the selection of the most parsimonious, and therefore generalizable, linear model. The final model could thus be said to characterize 19.98% of the variability in global sales of the video games with the included predictors and interactions. The final model was also highly significant with a p-value approximately equal to zero.

Parameter Directions in Final Multiple Regression Model:

```
Coefficients:
                                            Estimate  Std. Error t value            Pr(>|t|)
(Intercept)                                1.08513799  0.15810666   6.863  0.0000000000073482 ***
Critic_Score                              -0.08917719  0.02749829  -3.243           0.001189 **
User_Score                                -0.18287884  0.02317666  -7.891  0.0000000000000035 ***
ConsolePC                                  0.09286584  0.21763393   0.427           0.669607
ConsolePlayStation                        -0.55397495  0.12312667  -4.499  0.0000069384439200 ***
ConsoleXbox                               -0.31356516  0.13824903  -2.268           0.023355 *
RatingEveryone 10+                        -0.17092962  0.13232732  -1.292           0.196501
RatingMature 17+                          -0.87706721  0.13533414  -6.481  0.0000000000979370 ***
RatingTeen                                -0.27936584  0.10823254  -2.581           0.009868 **
GenreAdventure                            -0.14046844  0.07396951  -1.899           0.057608 .
GenreFighting                             -0.07694562  0.07561463  -1.018           0.308905
GenreMisc                                  0.15826628  0.06297549   2.513           0.011990 *
GenrePlatform                              0.00535926  0.05805999   0.092           0.926458
GenrePuzzle                               -0.19699049  0.07780806  -2.532           0.011373 *
GenreRacing                               -0.14315930  0.07011375  -2.042           0.041210 *
GenreRole-Playing                          0.01846703  0.05527689   0.334           0.738328
GenreShooter                              -0.08800901  0.06672429  -1.319           0.187217
GenreSimulation                            0.03124780  0.07534943   0.415           0.678371
GenreSports                               -0.21793727  0.05945801  -3.665           0.000249 ***
GenreStrategy                             -0.21484878  0.09279932  -2.315           0.020633 *
ConsolePC:RatingEveryone 10+               0.11095910  0.11869138   0.935           0.349898
ConsolePlayStation:RatingEveryone 10+      0.10530947  0.06270079   1.680           0.093092 .
ConsoleXbox:RatingEveryone 10+             0.15075687  0.07463974   2.020           0.043446 *
ConsolePC:RatingMature 17+                 0.35767962  0.12551964   2.850           0.004391 **
ConsolePlayStation:RatingMature 17+        0.45364876  0.08532092   5.317  0.0000001090166291 ***
ConsoleXbox:RatingMature 17+               0.41139935  0.09370891   4.390  0.0000115043557101 ***
ConsolePC:RatingTeen                       0.27336532  0.09917768   2.756           0.005862 **
ConsolePlayStation:RatingTeen              0.26089355  0.05758693   4.530  0.0000059912988414 ***
ConsoleXbox:RatingTeen                     0.25873525  0.06749162   3.834           0.000127 ***
ConsolePC:GenreAdventure                   0.07859049  0.15499216   0.507           0.612129
ConsolePlayStation:GenreAdventure         -0.18340543  0.10015392  -1.831           0.067111 .
ConsoleXbox:GenreAdventure                -0.09847608  0.13087797  -0.752           0.451822
ConsolePC:GenreFighting                   -0.16118733  0.32701701  -0.493           0.622098
ConsolePlayStation:GenreFighting           0.10459655  0.09241859   1.132           0.257773
ConsoleXbox:GenreFighting                  0.07459557  0.10481558   0.712           0.476686
ConsolePC:GenreMisc                       -0.17747610  0.32279105  -0.550           0.582464
ConsolePlayStation:GenreMisc              -0.24745840  0.08712611  -2.840           0.004522 **
ConsoleXbox:GenreMisc                     -0.02665488  0.09778115  -0.273           0.785171
ConsolePC:GenrePlatform                   -0.05481130  0.21472726  -0.255           0.798530
ConsolePlayStation:GenrePlatform          -0.04838696  0.08460485  -0.572           0.567398
ConsoleXbox:GenrePlatform                 -0.09313577  0.11197328  -0.832           0.405570
ConsolePC:GenrePuzzle                      0.11024150  0.33467506   0.329           0.741865
ConsolePlayStation:GenrePuzzle            -0.15873710  0.15492868  -1.025           0.305599
ConsoleXbox:GenrePuzzle                   -0.04254652  0.32334226  -0.132           0.895318
ConsolePC:GenreRacing                      0.04978614  0.13741613   0.362           0.717138
ConsolePlayStation:GenreRacing             0.18262622  0.08661307   2.109           0.035023 *
ConsoleXbox:GenreRacing                    0.09787352  0.09507410   1.029           0.303309
ConsolePC:GenreRole-Playing                0.01579910  0.10568854   0.149           0.881174
ConsolePlayStation:GenreRole-Playing      -0.14987739  0.06990996  -2.144           0.032081 *
ConsoleXbox:GenreRole-Playing              0.14713819  0.09540564   1.542           0.123064
ConsolePC:GenreShooter                     0.11172507  0.10327167   1.082           0.279357
ConsolePlayStation:GenreShooter            0.05388603  0.08012775   0.673           0.501288
ConsoleXbox:GenreShooter                   0.17727008  0.08326174   2.129           0.033286 *
ConsolePC:GenreSimulation                  0.11153872  0.12389575   0.900           0.368014
ConsolePlayStation:GenreSimulation        -0.04396260  0.10723267  -0.410           0.681839
ConsoleXbox:GenreSimulation               -0.14151753  0.11951522  -1.184           0.236418
ConsolePC:GenreSports                      0.11189430  0.16968262   0.659           0.509641
ConsolePlayStation:GenreSports             0.26408798  0.07716668   3.422           0.000625 ***
ConsoleXbox:GenreSports                    0.16005958  0.08449907   1.894           0.058240 .
ConsolePC:GenreStrategy                    0.21544417  0.12512627   1.722           0.085151 .
ConsolePlayStation:GenreStrategy          -0.15926083  0.12165728  -1.309           0.190549
ConsoleXbox:GenreStrategy                 -0.02377851  0.14369949  -0.165           0.868576
Critic_Score:User_Score                    0.02939043  0.00321169   9.151 < 0.0000000000000002 ***
User_Score:ConsolePC                      -0.09677518  0.02527608  -3.829           0.000130 ***
User_Score:ConsolePlayStation             -0.02553244  0.01756531  -1.454           0.146113
User_Score:ConsoleXbox                    -0.09480402  0.01941070  -4.884  0.0000010637366493 ***
Critic_Score:ConsolePC                    -0.00006627  0.03249306  -0.002           0.998373
Critic_Score:ConsolePlayStation            0.09662025  0.01889303   5.114  0.0000003242927060 ***
Critic_Score:ConsoleXbox                   0.11788499  0.02066076   5.706  0.0000000120923370 ***
Critic_Score:RatingEveryone 10+            0.01067353  0.01912236   0.558           0.576747
Critic_Score:RatingMature 17+              0.08174893  0.01676364   4.877  0.0000011051023572 ***
Critic_Score:RatingTeen                    0.00578523  0.01473270   0.393           0.694569
```

The results of the model in terms of parameter direction were at times slightly counterintuitive. For example, “Critic_Score” and “User_Score” both had negative coefficients, indicating an inverse relationship between user/critic ratings of a video game’s quality and total global sales. Other statistically significant, interesting observations include that games with a sports genre were strongly correlated to lower global sales, PlayStation games led to significantly lower sales than Xbox or PC games, and games with a rating of mature 17+ sold far less units than more mild games.

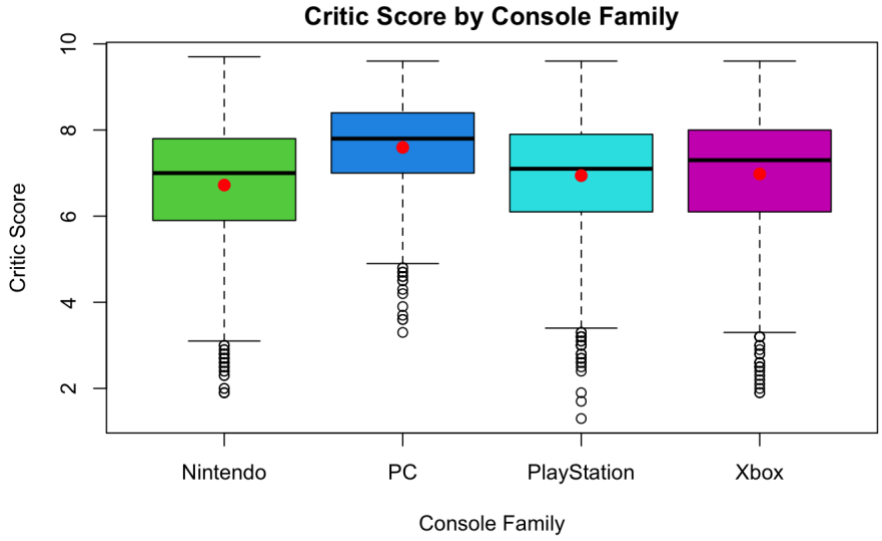
Residual Plots from Multiple Regression:



The normal quantile plot of the final generalized linear model (GLM) model (result of backwards stepwise multiple regression) demonstrated normality with the exception of fairly significant deviation from normality at the high tail end. Additionally, the plot of model fits for the GLM vs. the studentized residuals displays some evidence of heteroskedasticity as the variance and frequency of outliers increases as the fitted values grow in magnitude. These results suggest that the GLM fails to fit data in the higher range of global sales, indicating that this model may not be completely generalizable. However, the results of the residual plots do suggest that the GLM does provide a reasonably accurate prediction of global sales overall.

One-Way ANOVA to Investigate Relationship Between Critic Score and Console:

Boxplot of Critic Scores as Function of Console Family:



The above boxplot illustrates some potential differences in the mean (red dot), median (black line), and spread of the critic scores across video game groups defined by the

console family. The left-skewed nature of the critic score data is also clear from this boxplot across all four console families.

Ratios of Maximum to Minimum Standard Deviations for Critic Score across Console Families:

The ratio of the maximum to minimum standard deviations for the critic score within each group (console family) was calculated in order to satisfy the condition of performing a one-way ANOVA that this ratio not exceed two. As seen below, this condition for a one-way ANOVA was met by this data:

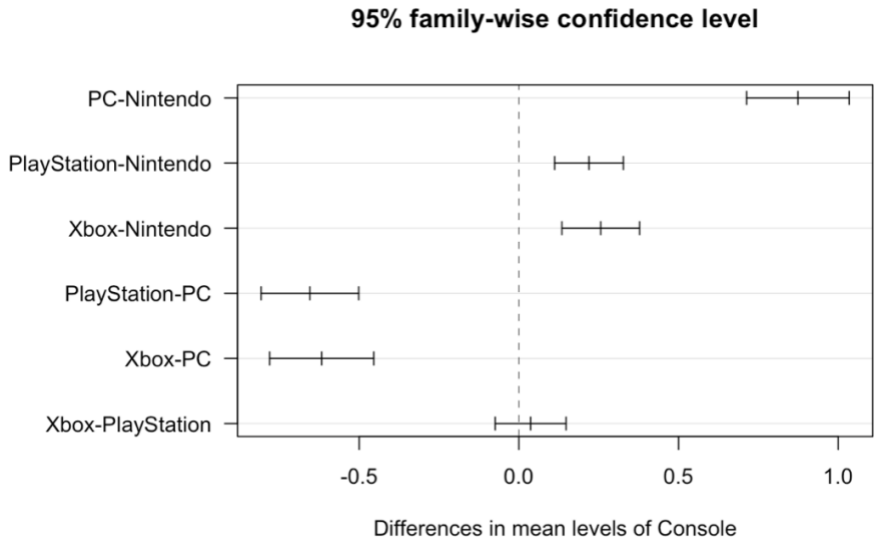
```
Nintendo          PC PlayStation        Xbox
1.426707    1.106982    1.303879    1.446515
```

One-Way ANOVA and Pairwise Comparisons (Holm P-Values) for Critic Score and Console Family:

```
             Df Sum Sq Mean Sq F value             Pr(>F)
Console       3    359  119.77   65.29 <0.0000000000000002 ***
Residuals  6586  12080    1.83
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        Pairwise comparisons using t tests with pooled SD

data:  Critic_Score and Console

           Nintendo             PC                   PlayStation
PC         < 0.0000000000000002 -                    -
PlayStation 0.00000032          < 0.0000000000000002 -
Xbox        0.00000019          < 0.0000000000000002 0.39

P value adjustment method: holm
```

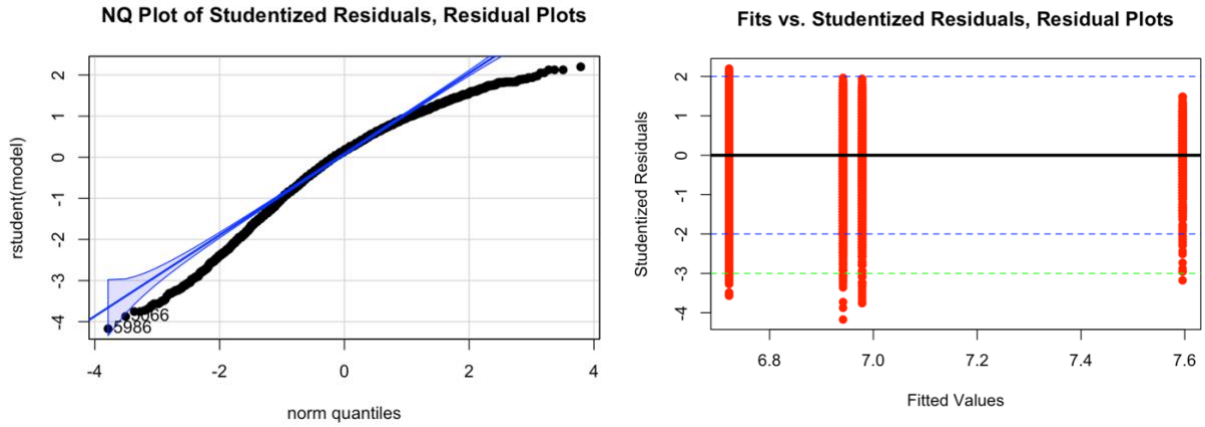
The one-way ANOVA, utilizing the Holm p-value adjustment method, demonstrated clear statistical significance (p approximately equals zero) for each of the groups (with the exception of Xbox and PlayStation, p-value = 0.39), indicating that there is a statistical difference between most groups (console families) with respect to critic score.

Tukey Simultaneous 95% Confidence Intervals for Differences in Critic Score:



Tukey simultaneous 95% confidence intervals were calculated, representing the range of possible differences in mean critic score between two groups of video games (grouped together by console for which they were intended). Each pair of consoles had statistically significantly different means for critic score (interval did not overlap with zero) with the exception of Xbox and PlayStation (does intersect with zero). Thus, one can conclude that Xbox and PlayStation video games are not significantly different in their ratings amongst professional video game critics. It is important to note, however, that this Tukey plot should not be extrapolated to compare distinct pairs with each other. Ultimately, the conclusions reached using the Holm p-values identical to those reached here with the Tukey intervals.

Residual Plots for One-Way ANOVA Model:

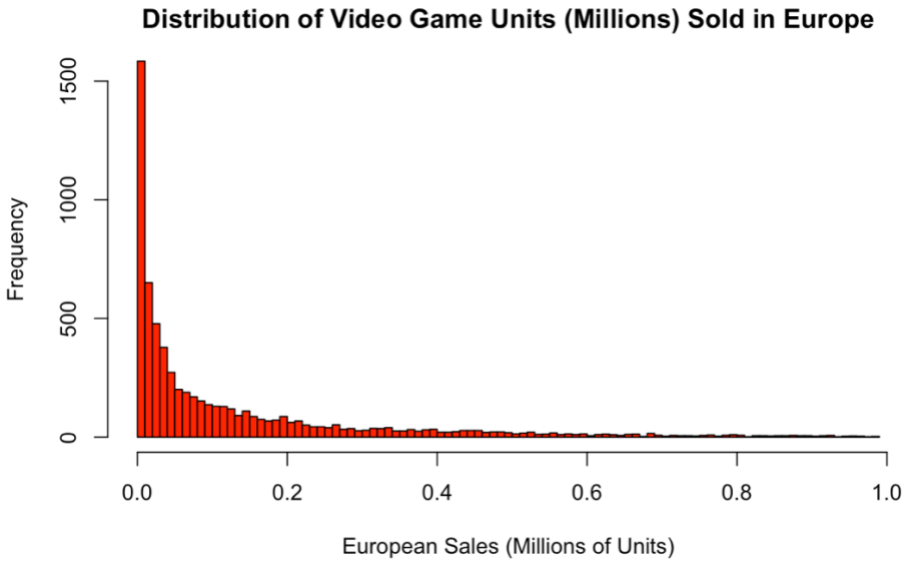


The normal quantile plot demonstrates deviations from normality at the tail-ends of the distribution. This is a problem for the reliability of the one-way ANOVA performed since one-way ANOVAs assume a normal distribution amongst the residuals of data within groups. However, the plot of fitted values vs. studentized residuals shows reasonably few

significant outliers and little evidence of heteroskedasticity, suggesting a Box-Cox transformation is not completely necessary. These residual plots do, however, suggest that confidence in the results of the above one-way ANOVA is not complete, and various potential transformation could be explored (not included in report since not effective when performed in analysis).

ANCOVA for Investigation of Relationship Between European Sales and Critic Score:

European Sales Data for Video Games with Under 1 Million Units Sold:



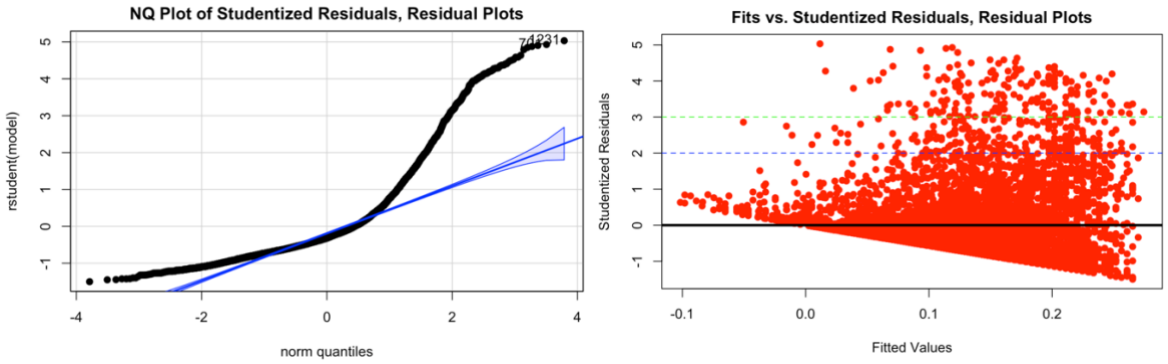
This histogram of European sales for video games with under 1 million units sold in Europe is heavily right-skewed, indicating that most video games sell relatively few units in Europe (same for North America or globally). An ANCOVA was performed next in order to elucidate potential predictors of European sales.

Type-III ANOVA of Linear Model Including Critic Score and Console Family:

```
Call:
lm(formula = EU_Sales ~ Console_Family + Critic_Score + Console_Family *
    Critic_Score, data = clean_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26537 -0.11022 -0.05585  0.04310  0.88837

Coefficients:
                                   Estimate Std. Error t value          Pr(>|t|)
(Intercept)                       -0.038625   0.020548  -1.880           0.06019 .
Console_FamilyPC                  -0.152757   0.052592  -2.905           0.00369 **
Console_FamilyPlayStation         -0.121917   0.027705  -4.400 0.000010971077227 ***
Console_FamilyXbox                -0.148464   0.030175  -4.920 0.000000886067506 ***
Critic_Score                       0.021890   0.002990   7.320 0.000000000000277 ***
Console_FamilyPC:Critic_Score      0.017244   0.006980   2.470           0.01352 *
Console_FamilyPlayStation:Critic_Score  0.023419   0.003983   5.880 0.000000004311787 ***
Console_FamilyXbox:Critic_Score    0.024964   0.004308   5.795 0.000000007143122 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.177 on 6582 degrees of freedom
Multiple R-squared:  0.09669,   Adjusted R-squared:  0.09573
F-statistic: 100.6 on 7 and 6582 DF,  p-value: < 0.00000000000000022
```

This linear model investigated the effectiveness of console family and critic score (and their interaction effect) in predicting European video game sales. This model had an overall p-value essentially equal to zero, indicating strong statistical significance. However, the multiple R-squared value of this model (0.09669) indicates that only 9.669% of variability in European sales of video games is accounted for by the model's included variables. Additionally, each of the main effects of console type and critic score and each of the interaction effects between console type and critic score was at least statistically significant at the alpha value of 0.05.

Residual Plots for Type-III ANOVA Predicting European Sales with Console Family and Critic Score:



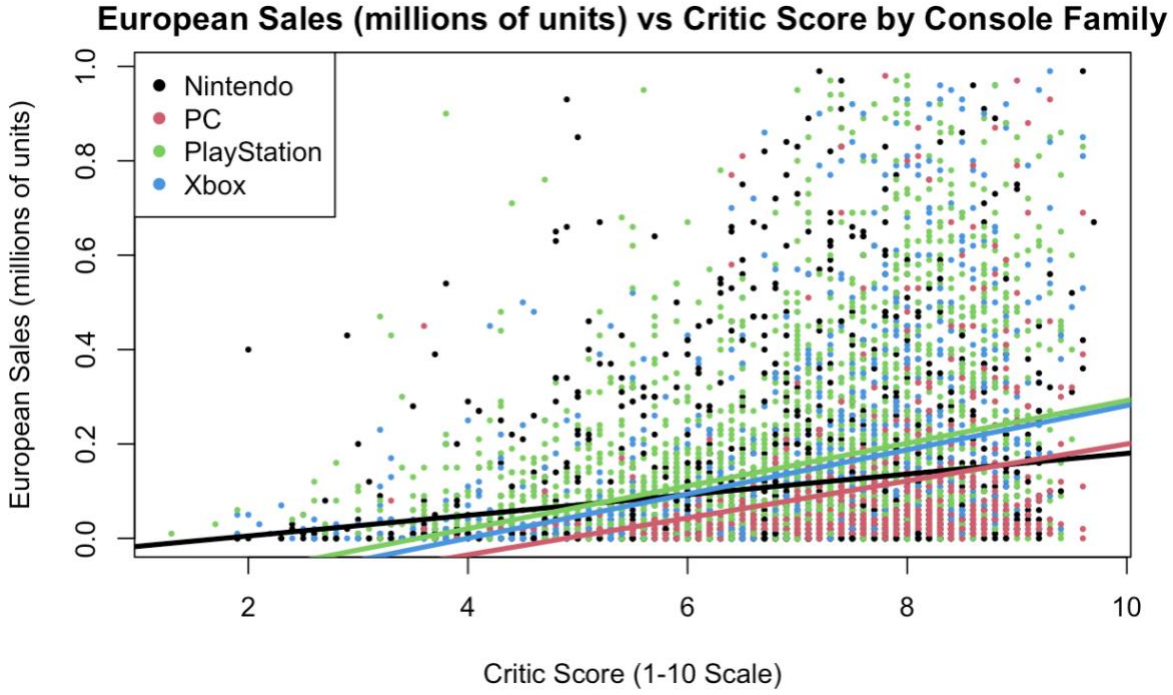
The normal quantile shows significant deviation of the residuals from normality throughout, and the plot of fitted values vs. studentized residuals displays clear heteroskedasticity. Thus, this data is not well suited for an ANOVA (lack of normality and clear heteroskedasticity), providing the rationale for performing an ANCOVA instead in order to test the main and interaction effects on European video game sales. ANCOVAs are generally used when covariates (e.g. console type on the relationship between critic score and European sales) are potentially present in a model.

ANCOVA Determining Extent to Which Console Family Mediates the Relationship Between Critic Score and European Sales of Video Games:

ANCOVA Model:

```
clean_data <- clean_data[clean_data$EU_Sales<1,]
plot(EU_Sales ~ Critic_Score, data = clean_data, col = factor(Console_Family), main =
       "European Sales (millions of units) vs Critic Score by Console Family",
     xlab = "Critic Score (1-10 Scale)",
     ylab="European Sales (millions of units)", pch = 16, cex = .5)
legend("topleft", col = 1:4, legend = levels(factor(clean_data$Console_Family)), pch = 16)
coefs<- coef(m3)
#nintendo, pc ,play station, xbox
abline(a=coefs[1], b =coefs[5], lwd=3, col=1)
abline(a=coefs[1]+coefs[2], b =coefs[5]+coefs[6], lwd=3, col=2)
abline(a=coefs[1]+coefs[3], b =coefs[5]+coefs[7], lwd=3, col=3)
abline(a = coefs[1] + coefs[4], b = coefs[5] + coefs[8], lwd = 3, col = 4)
```

This ANCOVA model explores the relationship between critic score and European sales of video games (millions of units) as a function of the potential covariate console family. The abline() functions is used to indicate the slopes of each of these relationships (stratified by console type) with the corresponding coefficients being taken from the Type-III ANOVA above.



This ANCOVA model demonstrates the relationship between European video game sales and critic score as a function of console type. The trends for PlayStation and Xbox are nearly identical with regards to slope and intercept, suggesting there is little difference in popularity for either console in Europe and that, on average, critics rate Xbox games very similarly to PlayStation games. PC and Nintendo video games seem to have similar levels of sales in Europe when at a critic score above 8, but Nintendo appears to clearly outperform PCs at lower critic scores (could also be that Nintendo has wider range of critic scores than PCs). Overall, this ANCOVA model provides some basic insights into the way console type interacts with the variables critic score and European sales.

Conclusion and Summary of Statistical Analysis of Video Game Data (1980-2020):

Throughout this analysis various statistical tests and data visualization techniques were utilized in order to understand the sales in different markets and ratings (both of users and critics) of video games. Data cleaning procedures were first implemented in order to effectively handle missing or useless data and to group data together into simple logical categories that could be used to provide clear analyses. Simple graphic visualizations of key data were then produced (e.g. boxplots, histograms, and scatterplots

of sales and critic/user scores) by performing the necessary transformations and grouping the data in ways that could provide useful insights (e.g. sales or critic/user score as a function of video game genre). Basic Welch's t-tests were then performed in order to calculate 99% confidence intervals that represented the sample mean difference between pairs of variables (sales of video games in different markets and user vs. critic scores). This allowed for the conclusion that none of the variables tested were redundant (all were statistically significantly different), even those that were highly correlated to each other. Regression models (correlations) were then produced for user vs. critic scores and user/critic scores vs. global sales, and accompanying permuted correlations were performed to test the statistical significance of the calculated correlations. In order to effectively investigate these correlations, large outliers (potential influential points) were removed to improve how accurately the correlations represented the majority of the data. Bootstrap confidence intervals were also calculated and compared to t-test derived confidence intervals in order to investigate the effect on global sales of selected genres of video games (action vs. shooter and sports vs. racing).

Backwards stepwise multiple regression was then produced in order to rigorously assess which variables (categorical and continuous) could most effectively contribute to a GLM capable of predicting the global sales of video games. In order to accomplish this, interaction plots were assessed for the categorical variables, and ANOVAs were used to determine the significance of any interaction effects. All these potential main and interaction effects were then included in a preliminary multiple regression model, and this model was made more parsimonious in a stepwise manner by eliminating the least statistically significant interaction effect until a final model was achieved. A one-way ANOVA was then performed to demonstrate the presence of a relationship between critic score and the console. Statistical significance for this relationship was shown via both the Holm p-value adjustment method and Tukey simultaneous 95% confidence intervals. Finally, an ANCOVA was performed to investigate any relationship between European sales of video games and critic scores, and the specific nature of this interaction was fairly rigorously characterized. The results of this statistical analysis can provide insights into the relationship between various video game attributes (e.g. genre, user score, critic score, console, and rating) and overall sales in three markets (North America, Europe, and global). Furthermore, this analysis has provided a more detailed perspective on the exact nature by which these variables interact with each other to mediate larger impacts on the independent variables of sales and rating scores.