1 Introduction

In the video gaming industry, there is a conceived disconnect between critic reviews and user reviews. Users believe that critics are biased toward certain games and developers and that these biases in turn produce inaccurate reviews. This bias is believed to be caused by relationship critics and game developers. In order to review a game, upwards of 20-30 hours are needed. Thus, critics are incentivized to get early-release copies of games to release early reviews (which generates more audience for the critic). However, early-release copies are provided at the behest of the video game developer. If a developer sees that a certain critic rates games too harshly, they might be less inclined to provide that critic with early releases, which in. turn motivates critics to be softer with their reviews. The relationship between game critics and users is the aim of this project, where we will study and analyze video game critics and user scores. We will compare these scores and try to find any correlations and patterns between them. Finally, we create regression models to predict one score based on the other.

2 Dataset

The data used in this project has 16 variables containing information about games. The dataset will be extracted from data on game review scores by critics and users.

- Name: This represents the game names in the data set. It is not important to our project.
- Platform: This represents the different gaming platforms that each game is on. There are 31 unique platforms, such as Wii, PC, and PS.
- Year of release: This represents the year each game was released in. The years start

from 1980 until 2020.

- Genre: This represents the genre of the game. There are 12 genres with Miscellaneous being the 12th one. The game genres are Action. Adventure, Fights, Platform, Puzzle, Racing, Role-Playing, shooting, simulation, sports, strategy, and Miscellaneous.

- Publisher: This represents the game's publisher which can be one of 582 publishers, such as Nintendo, Zushi Games, and Rockstar Games.

- NAsales, EUsales, JPsales, Other sales, GlobalSales: These five represent the sales numbers of a game across different regions. These regions can be North America, European Union, Japan, other countries, and aggregated global sales.

- Critic score: This represents the aggregated critics' review score for a game. The scores range from 13 to 100, some have NA but have been excluded during the cleaning process.

- Critic count: This data represents the number of critics' review score counts which ranges from 3-115. Each review score is then aggregated into the final critic score for a game.

- UserScore: This represents the aggregated critics' review score for a game. The scores range from 0 to 10. Some have NA but have been excluded during the cleaning process.

- User Count: This data represents the number of user review score counts. Each review score is then aggregated into the final critic score for a game.

- Developer: This represents the companies that developed a game, there are 1000 unique developers such as Maximum Games, and Crafts & Meister.

- Rating: This represents a game's age rating. These are given by a special company (dependent on region) that specifies the age rating of each game.

  The two most important data columns are user scores and critic scores. In order to fairly compare the two, user scores have been multiplied by 10 to have the same value ranges. Developer, Rating, Genre, and Platform are all categorical data that can be used in our regression models.

3 Preliminary Plots and Testing

Since our dataset is extremely large, plotting and analyzing it will give results that are hard to interpret. Thus, we have done uniform sampling (100 samples) and have used the sampled dataset when it simplifies the analysis. All our preliminary plots are centered around user and critic review scores as these data points are what our project will focus on.

3.1 Mean and Distribution

We first test and plot whether the means and distribution of the scores are similar.

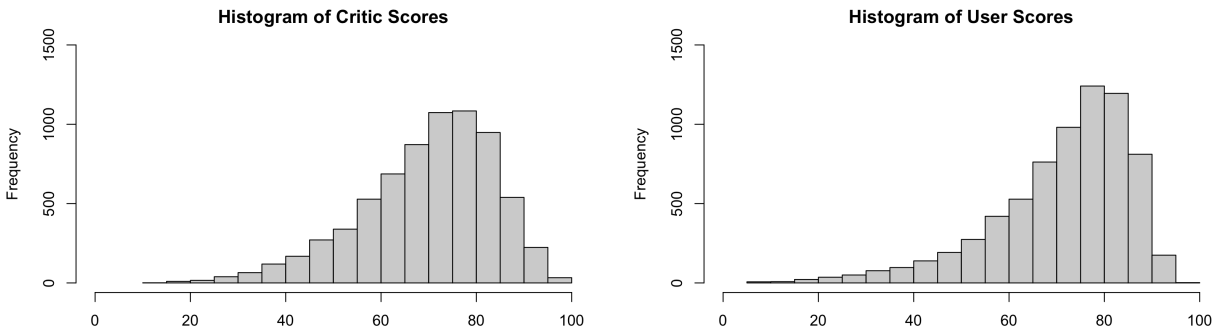

Figure 1: Histogram of the full dataset for critic and user score

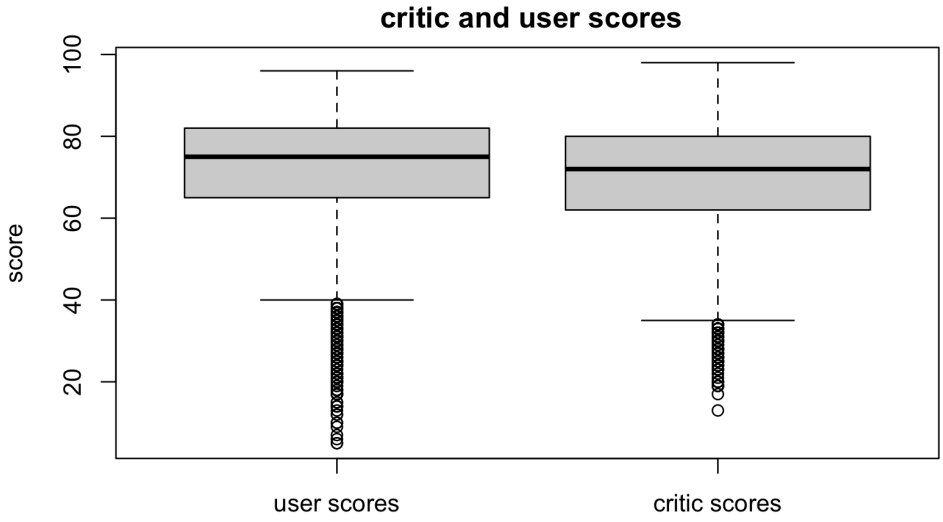

Figure 2: Boxplot of the full dataset for critic and user score

As the boxplot and Histogram show, both the user score and critic score have very similar distribution and mean. To confirm this, we used the Wilcoxon rank sum test with continuity correction and the Welch Two Sample t-test on the sampled dataset. The resultant P values are 0.1554 and 0.1841, meaning we cannot reject the null hypothesis that the two means for user and critic scores are equal. This, however, does not mean that the two datasets are correlated. Their means and distributions being similar can be the result of working on the same range (0-100) and human bias tending towards the same values (in the case of scores, both center around 70-80) while having very different scores for the same game.

3.2 Normality

Since the means and distribution are indeed similar, we now try to find the distribution type by testing for normality.

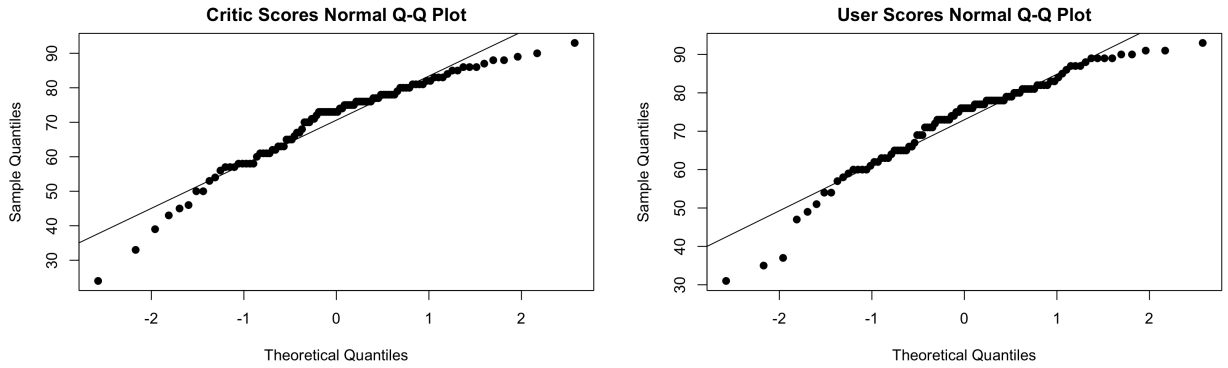

Figure 3: Normality test of the full dataset for critic and user score

As the plots show, both scores are not normal, this is confirmed using the Shapiro-Wilk normality test on both samples. The resultant P values are less than 0.001. Meaning we can reject the hypothesis that the samples (and by extrapolation the full dataset) are normally distributed.

3.3 Correlation

Since both datsets have similar means and distribution, we want to test how well correlated they are. If the correlation is high then we can conclude that user and critic reviews are equivalent.

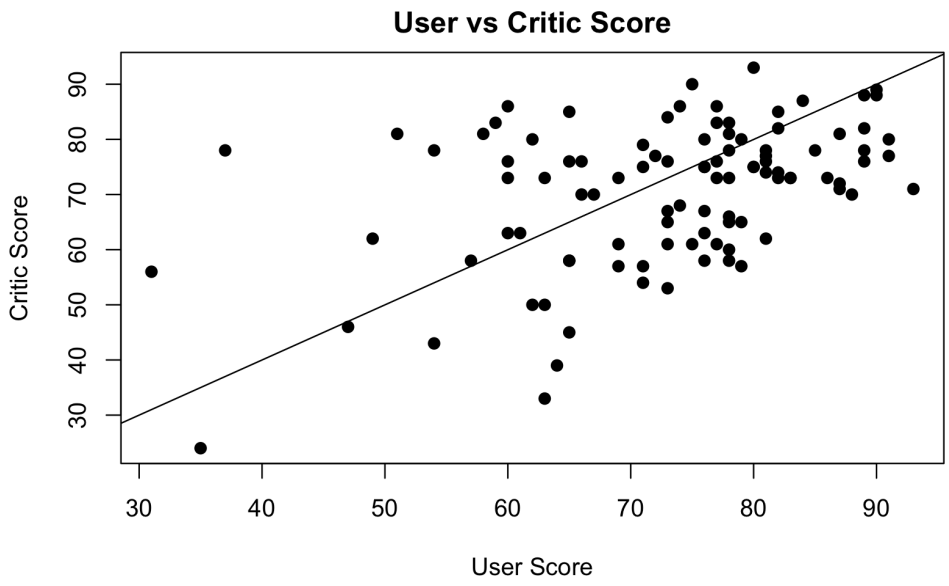

Figure 4: Correlation plot between user score and critic score

As the plot shows, there appears to be no significant correlation between user and score reviews. This is backed up by Pearson's product-moment correlation test and Kendall's rank correlation tau. Both tests give a P value less than 0.001. This means that user and critic scores are, most often than not, are at odds with each other.

3.4 Ordered Data

In this last plot, we sort user scores from lowest to highest. We then use the same order of indexes and apply it to critic scores. The end plots are user scores in increasing order, and critic scores ordered by user scores.

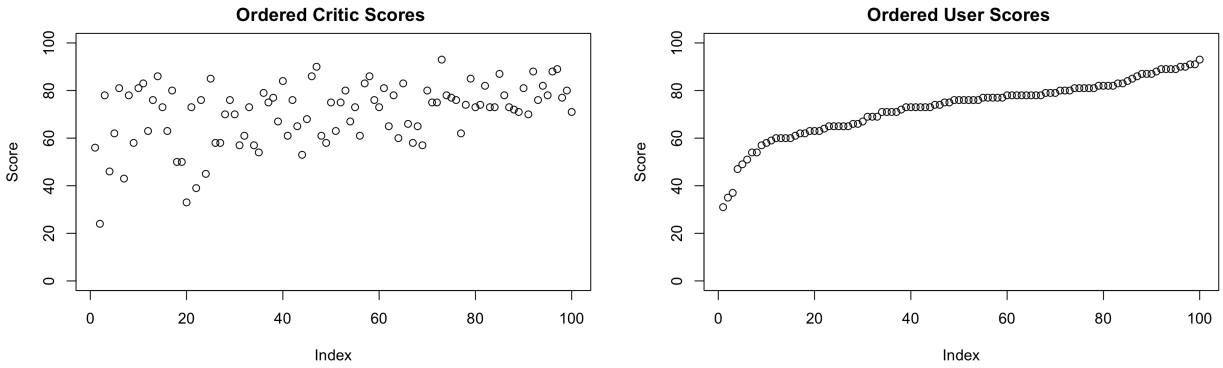

Figure 5: Plot of scores after ordering based on one

If both plots are indeed correlated, then both plots should show the same linear increase as we go up in the index. However, critic scores are not showing any specific pattern of increase or decrease, especially for the lower values. The higher values (i.e highly rated games), appear to be close in value for both critic and user scores. But for middle and lower-rated games, there is no strong correlation that can be shown between them.

4 Plan for Analysis and Modeling

This section outlines the group's plan for the analysis of the video game data. The process can be broken down into three main steps as follows.

4.1 Dataset Exploration

The first step is to explore trends and correlations in the dataset through analysis and testing. In the preliminary plots section, we applied some tests (including the t-test, and Pearson's correlation test) to examine the relation between the two score variables. This can be repeated with any pair of variables, with additional analysis tools, such as ANOVA or a simple linear regression. For example, applying a linear regression between user and critic

scores shows an R^2 of around 0.34, indicating some relationship, but failing to be a nearly adequate model. Another two simple linear regressions between scores and the year of the release show a correlation between a game's year of release with its user rating, but not its critics' rating.

4.2 Model Formulation and Testing

After identifying the relevant variables, we plan to formulate a model for scores using the multiple regression. We will then run diagnostic tests (using residual plots, QQ-plots, etc.) on the model to check for the correctness of assumptions. Depending on the results, we may have to apply a different model, perhaps a nonlinear one.

4.3 Discussion and Presentation

With the model finalized, we will then draw conclusions and explain the reasoning behind each found correlation, plot the results using the data representation techniques we covered in class, and prepare a report and presentation with our findings.