**Results**

Dylan Tucker, dst833, 11235055

**ABSTRACT**

In this paper, I discuss a dataset of steam reviews and genre tags for video games and the analysis I conducted. I will review the results from this analysis and evaluate the model’s effectiveness at determining user suggestion averages from genre tags. I will also discuss the statistical methods that will be used for this analysis.

**CCS Concepts**

• **Information** • **Data Analytics** • **Statistics** • **Linear Regression**

# INTRODUCTION

[Figure 1 here]

The goal of this analysis was to build a linear regression model to predict user suggestion averages using genre tags for video games in a dataset. There are a total of 44 games and 65 unique genre tags in the dataset. The data was split into 2 subsets which were the training and test data respectively. The training data made up 80% of the original dataset while the test data was the remaining 20%. Both datasets contain 3 columns consisting of a title, user suggestion average, and genre tags. Both the genre tags and the titles were encoded into numeric values to better visualize and organize the data. This paper will explain the results of a linear regression model to determine the relationship between genre tags and user suggestions.

# Main Results

[Figure 2 here]

The initial test had a mean squared error (MSE) value of 2.56 and a coefficient of determination of 0.2. This initial split of the data showcases promising results as a 0.05 MSE value is close to zero. The coefficient of determination is less promising at 0.9, which indicates that the regression line may not fit the data and that genre tags may have little or no correlation with user suggestions.

## Main Results

Steam tags are organized based on prominence and relevance to the games content and genre[[1]](#footnote-1). The genre tags were originally limited to the first 5 to attempt to decouple the data as many games contain the same tags. I decided to perform the same regression without limiting the number of tags for each title. In this analysis, there were 140 unique tags, as opposed to 65 in the original regression. Despite adding more data, the outcome was the same with an MSE of 0.02 and an R2 score of -0.11. Again, this indicates that there may be little correlation between genre tags and user suggestions.

# Cross Validation

[table of trials and results here]

There was a total of 5 more tests conducted, each using a different split of the data for training and testing. Many of the R2 scores are very low, or even negative, which suggests that the regression line does not fit the data. An R2 score of 0.07 suggests that only 7% of the variance from the dependent variable can be explained by the independent variable. The MSE values are also low, which should mean that the model is accurate in its predictions. However, seeing both a low MSE and low R2 score likely means that there is little or no correlation between genre tags and user suggestions.

# Roadblocks

The biggest roadblock to this analysis was the lack of titles in the dataset. The dataset only includes 44 games once the titles with no user suggestions were removed. Approximately 8 of the 44 titles, around 20% of the data, was reserved for the test dataset. The dataset was originally used to determine user suggestions based on game reviews; the quantity of titles was less important when compared to the quantity of reviews per title.

## Some speculation

Most steam games, at least the ones featured in this dataset, contain many genre tags. I originally believed that this would muddy the results, but the opposite may have been true. The steam user suggestion is binary; either the user gives a thumbs up (1) or a thumbs down (0), which corresponds to recommend and not recommend respectively. Most of the reviews in this dataset were positive with 2442 1s than 0s in the dataset. The games which featured a unique content tag that no other game in the dataset contained, may not have been useful training or test data for the model, due to their lack of recurrence.

## Diversity in the dataset

Perhaps the second largest roadblock was the homogeneity of the data. All 44 games in the dataset contain the ‘Free to Play’ tag and 43 games contain the ‘Multiplayer’ tag, but only 1 game contains the ‘Racing’ tag or the ‘Stealth’ tag. The genre tags with only one title were likely not useful to the model as they did not recur. There could also be a type of survivorship bias, as the number of games featuring ‘Free to Play’ and ‘Multiplayer’ tags could itself be an indication that these are the types of games users enjoy. Having a larger and more diverse dataset of games would likely produce more interesting results by providing better training and test data for the model.

# Conclusion

The goal with this analysis was to determine if there is a correlation between genre tags and user suggestions. The results showed low mean squared error values and R2 scores which were occasionally negative. Limiting the number of genre tags did not have a significant impact on the results. There were many reasons that may have contributed to the outcome of this analysis, such as the lack of diversity in the dataset, or the lack of titles which led to few genre tags or recurring tags. Regardless, the results of this analysis show that genre tags have little effect on user suggestions.

1. Valve. Steam Tags. Retrieved November 17, 2023 from https://tinyurl.com/3jkx3utf [↑](#footnote-ref-1)