**Project Proposal**

Dylan Tucker, dst833, 11235055

**ABSTRACT**

In this paper, I discuss a dataset of steam reviews and genre tags for video games. I will construct a model for determining the probability of a game being recommended by users based on its genre tags. I will review and examine the statistical methods that will be used for this analysis.

**CCS Concepts**

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

# INTRODUCTION

The dataset I will be using was sourced from the website Kaggle under a CCO: public domain license. This dataset was originally used to predict user recommendations based on user reviews. The data is split into a training dataset and an overview dataset. The training dataset contains a list of reviews for video games in the format of review ID, title, year of release, user review and a user recommendation. The first data contains many duplicate columns with the only differing values being the user reviews and user suggestion. A second dataset is also provided which contains an overview of each game featured in the first dataset. The second dataset is formatted so that each row contains a title, the name of the developer, the name of the publisher, genre tags, and an overview/description. Every row and column in the second dataset is unique. A linear regression will be performed to predict a game’s probability of being recommended by users based on its genre and content tags.

# Formatting

The data needs to be formatted before it can be analyzed. The first step is to group the training dataset by title and aggregate the mean of the user suggestions for each title. The result will be a data frame with two columns; the title column will have removed each duplicate video game title and will only contain unique values. The user suggestion column will contain the mean of every user suggestion for each title. For example, PlanetSide 2 has 472 reviews, 423 of those reviews are positive, which the resulting data frame will have a single row of a title (PlanetSide 2) and a user suggestion mean of 0.896186.

## Formatting

The next step is to combine this new data frame from the previous formatting step with the overview data set. The resulting data frame will again be grouped by title and unnecessary columns such as the developer, publisher, year, overview, user review and review ID will be dropped from the dataset. The final data frame will have a title column with all unique entries, a mean of all user suggestions for each title and a set of tags for each title.

## Formatting

The last step is to encode the title column to a numeric value, as knowing the exact title for each game is not necessary, provided each numeric value is unique and the genre tags are preserved for each title. Encoding the titles will help organize and visualize the data. The tags column will be tricky to encode as it is a list of strings, where each entry has a unique combination of strings. It may not be necessary to encode the tags, but doing so may also help with data visualization.

## Formatting

Steam orders its tags based on weight and prominence, which means the first 5 tags of each game should give the clearest picture of what that game will be like to play.[[1]](#footnote-1) Many of the games in this data set contain a long list of tags. There are 64 games in the dataset and 61 of them contain the top 3 tags in the data. The number of tags for each game is likely to muddy the data and not likely to produce valid results. The solution is to use the first 5 tags for each game and ignore the rest.

## Formatting

The training dataset does not feature all games present in the overview dataset. A third data set was also provided which contains several games which are not in the training data set. Unfortunately, this third data set does not feature a user suggestion column, which means that there will be entries in the final data frame which will have NaN (Not a Number) values in the user suggestion column. This data will be removed and stored in a separate data frame, as it could still be useful test input for the regression model.

# Methods

A linear regression model will be constructed to determine how user suggestions and genre tags are linearly related. The final goal will be to determine if the model can accurately predict user suggestion from genre tags alone. Once the model is constructed and trained, a confusion matrix can be made to compare the predicted and observed values in the model.

## Methods

Clustering may also be done to better visualize and compress the data.

The input for the model will be the training dataset, which will consist of the average of the user recommendations for a game and its genre/content tags. User recommendations are a binary value of one for recommend, or zero for do not recommend. The average of a user recommendation will be a value between zero and one, with a value closer to one being recommended by more users than a value closer to zero. The output will be a series of predictions using test data that will attempt to predict the average of user recommendations from the test data. The test data is a small portion taken from the overall dataset. This task will be performed using linear regression as the genre will be the only variable with impact on user suggestion that will be considered by the model. The model’s prediction score will represent how accurate the model was at predicting the average user suggestion. An R2 score will also be provided to identify how well the regression line approximates the data. Various graphs and data visualizations will also be constructed to represent the model and its data.

## Data Sample

Figure . The top 5 highest recommended games in the dataset.

Figure 2 shows a range of different games paired against the average of their user suggestion. Planet side 2 is a massively multiplayer online (MMO) first-person shooter, while EverQuest II is an MMO role playing game (RPG). The aim for this analysis will be to determine if there is a trend amongst the user suggestions for first-person shooters or RPG’s that can be used to predict the average user score for a game within the same genres.

## Proposed Results

Figure was made using fabricated or synthetic data that represents a sample of the proposed results for this analysis.

# Conclusion

Understanding how genre can affect player reception of a video game could help developers decide which genre of games are likely to be reviewed more favorably. Building a model that could make accurate predictions on how it will be received by players could help achieve this goal. If the model is not accurate, it could show that genre tags have little or no affect on how a game is received by players.

1. https://tinyurl.com/3jkx3utf [↑](#footnote-ref-1)