Methods

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 the dataset and how the data will be prepared before analysis. I will also discuss the statistical methods that will be used for this analysis.

CCS Concepts

• Information • Data Analytics •Statistics • Linear Regression

1. INTRODUCTION

The dataset I will be using was sourced from the website Kaggle under a CCO: public domain license. 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 training dataset contains many duplicate columns with the only differing values being the user reviews and user suggestions. 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. A linear regression will be performed to predict a game's probability of being recommended by users based on its genre and content tags.

2. Formatting (removing duplicates)

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 average 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 average user suggestion for that title. For example, PlanetSide 2 has 472 reviews, 423 of those reviews are positive, which means the resulting data frame will have a single row of a title (PlanetSide 2) and a user suggestion average of 0.896186.

2.1 Formatting (grouping data)

The next step is to combine this new data frame from the previous formatting step with the overview dataset. 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, an average of all user suggestions for each title, and a set of tags for each title.

2.2 Formatting (encoding)

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 list has a unique combination of strings. It may not be necessary to encode the tags, but doing so may also help with data visualization.

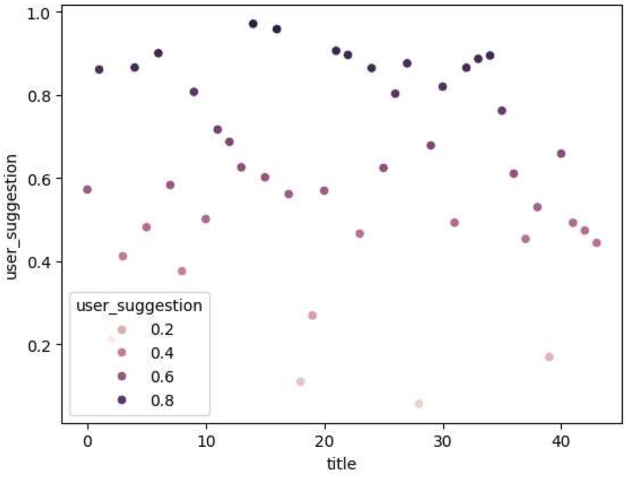

Figure 1. An overview of the entire data set after the titles of each game have been encoded into numeric values. Each data point is a video game plotted against its user suggestion average.

2.3 Formatting (extra steps)

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] 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 could potentially muddy the results of the analysis and may not produce interesting or useful results. It might be beneficial to limit the number of tags for each title to the first n-tags in the dataset.

2.4 Formatting (extra data)

The training dataset does not feature all games present in the overview dataset. A third dataset was also provided which contains several games which are not in the training dataset. 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 a useful test input for the regression model.

[1] Valve. Steam Tags. Retrieved November 8, 2023 from https://tinyurl.com/3jkx3utf

3. Methods

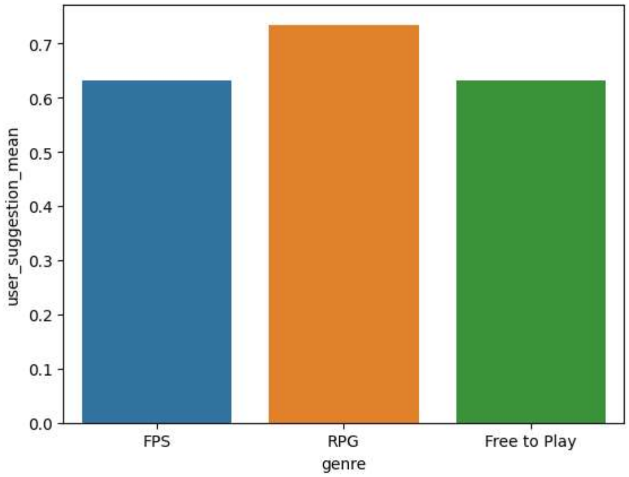

Figure 2 is a sample of what the results of this analysis might resemble. The data in figure 2 is genuine data sourced from the data set outlined in this paper.

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. K-means clustering may also be done to better visualize and compress the data.

3.1 Methods

The measure of evaluation will be a coefficient of determination and a prediction value to determine the accuracy of the model's classifications. The regression model itself will be evaluated based on its mean squared error (MSE). The MSE will be used to determine the difference between the predicted and expected values. Many data visualizations will also be provided, such as bar and scatterplots.

4. Conclusion

A linear regression model will be used to predict user suggestions from genre tags. Before the model can be made, the data will need to be formatted by grouping training data by title and combining it with genre tags found in the overview data set. The title column of the resulting data frame will then be encoded into numeric values. The mean squared error and coefficient of determination will be the metric to evaluate the analysis.