Game Recommendations on Steam

By Dani DiPinto

General Assembly DS-SEA-06

1. **Introduction**

I am a big fan of using the Steam platform for gaming. I thought it would be fun to do a project using data from Steam. Steam has some open source data available.

1. Problem statement and hypothesis

Problem:

Gaming Recommendations aren’t always good. They sometimes just promote their own games, or games that are generally popular.

Hypothesis:

Based on games purchased by Steam users, we can create game recommendations based on what a player is actually playing. I would like to use play hours and games purchased to cluster the user with other users with similar habits to recommend new games.

1. Description of your data set and how it was obtained

I found this dataset on Kaggle (<https://www.kaggle.com/tamber/steam-video-games>). It contains 200,000 rows of user data. The data includes a unique Steam user ID, the game’s name, a behavior (purchase or play), play hours (or 1.0 if linked with purchase), and a numeric value of zeros. The shape is 200000, 5.

1. **Cleaning the Data**
2. Description of your data set and how it was obtained

This is the step I have spent the most time on so far. As mentioned above, this dataset came from Kaggle. There are 200,000 rows and 5 columns. Not all of this information is set up usefully, or in a productive manner. Luckily, there are not any null values.

With my initial look at the data, I knew I would have to remove the fifth “0 column” since it does not contain any information. Then, the data behavior row has one row for “play” which gives the hours that the user has played that game, and one row for “purchase” which is always a ‘1.0’ to signify that the game was purchased. Taking a deeper look, this “purchase” behavior actually signifies downloads. Some of the games listed are free, so they were not purchased, however, we’ll just say that for this dataset, purchase is the same as download.

One of the biggest challenges is the format. I decided to separate the data into two data frames, one with the purchase = 1.0 data, and one with the play = hours played data. Then, I took both and merged them into a single data frame.

1. What you learned from exploring the data, including visualizations

After merging the play and purchase data frames, the new one runs much more quickly. The amount of rows has significantly decreased. Looking at the new table, I don’t have a lot of options for features. It limits the type of model I can use significantly. It looks like I’ll have to work with clustering.

1. **Creating a Model**
2. How you chose which features to use in your analysis

Since I don’t have a lot of features, I don’t have many options here. I’ll have the User ID, Game, and Hours Played to work with. I’m going to start making a clustering model with UserID and Game. If possible, I’ll try and add in hour played, to make the model more accurate based on what player enjoy playing.

1. Details of your modeling process, including how you selected your models and validated them
2. Your challenges and successes
3. **Real Life Application**
4. Possible extensions or business applications of your project
5. **Conclusions and key learnings**