Springboard capstone project 1 milestone report

Board Game Geek Blog

Board Game Geek (BGG) is the biggest board game forum for hobbyist and contains a database of reviews, images, and ratings. People or companies may join BGG for free and join in conversations about different games, post questions, or rate their favorite games. There is a lot of data on the site about each game.

Problem

With hundreds of new games coming out each year it is hard to not have a game get lost in all the noise and fanfare. How can a fan find games that they will enjoy but haven't received the hype and publicity that other games have gotten? Can you predict a game's popularity by the mechanics, theme, and player count?

Data

The data that will be used is from Kaggle. (https://www.kaggle.com/mrpantherson/board-game-data). It was collected from BBG in March of 2017.

The board game data has 20 columns and 5000 rows. Each column is a characteristic of a board game and each row is a unique board game. There seems to be three types of columns:

1. Qualitative: mechanic, category, designer
2. Quantitative: min_players, max_ player, avg_time, min_time, max_time, year, avg_rating, geek_rating, num_votes, age, owned, weight
3. unique rank, bgg_url, game_id, names, image_url

The columns and what data they contain are:

- rank: each game is ranked by popularity from 1 to 4999.
- bgg_url: url of game on BBG

- game_id: each game is given a unique id.
- names: name of board game
- min_players: minimum number of players
- max_player: maximum number of players
- avg_time: average time the game takes to play
- min_time: minimum time to play game
- max_time: maximum time to play game
- year: year board game was released
- avg_rating: mean of all board game user ratings
- geek_rating: BGG rating based on avg_rating but altered with secret BBG Stats
- num_votes: number of votes by users
- image_url: url to picture of board game cover
- age: recommended starting age
- mechanic: list of board game mechanics
- owned: number of users who stated they own game
- category: list of themes or categories of the board game
- designer: list of designers
- weight: average rating of how hard a game is to understand given by users

One problem with the data is some categories are user given such as weight, and rating. Others will be on the board game box but could be entered wrong. These are things that must be taken into account when exploring the data. It isn't full of hard facts but we must assume most categories are correct.

The following columns were removed because the data would not help a model and were unique:

1. bgg_url - url to the board game on BBG.

2. image_url - url to a picture of the board game on BBG.
3. rank - coresponds to geek_rating

For columns min_players, max_players, avg_time, min_time, max_time having a 0 as a value doesn’t make sense. A game has to have at least one player and time has to be greater than 0 to be played. They are only 127 games that have these missing values so they will be removed.

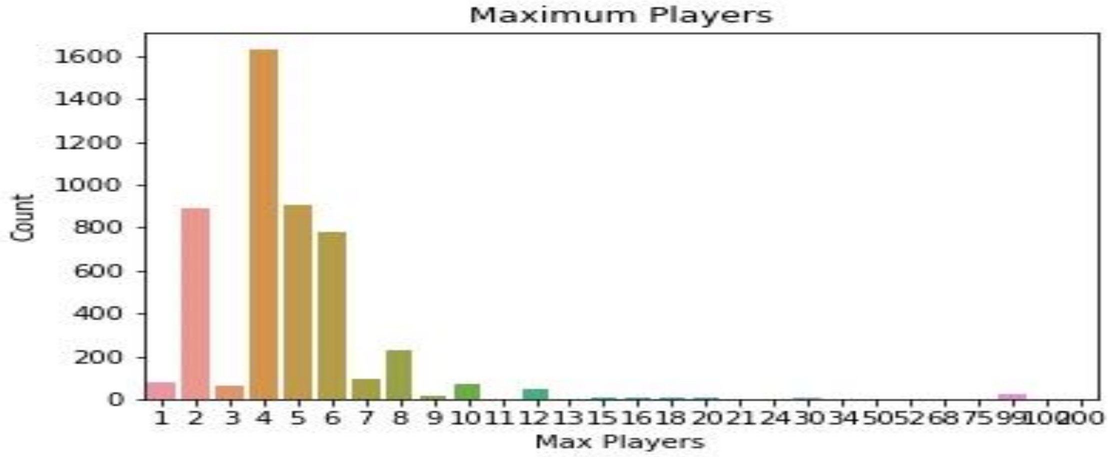


Figure 1

Most games have a maximum player count of 4 with 2, 5, and 6 being the next most common. In the figure 1 there seems to be a a significant amount of games at 99 players. There are only 120 games that have a max of 11 players or more and most times you will not have more than 10 people at a board game party. If I assume that I will also remove any board games that have a minimum player count of more than 10 players. I will remove those.

Some games will last longer than there is time in a day according to the play time listed. There are 1440 minutes in a 24 hour day. These games are probably campaign style games that can be left set up and come back to. Most people can't or won't devote that much time to them.

I will only be looking at games that are less than 8 hours to play. That is a full day of gaming.

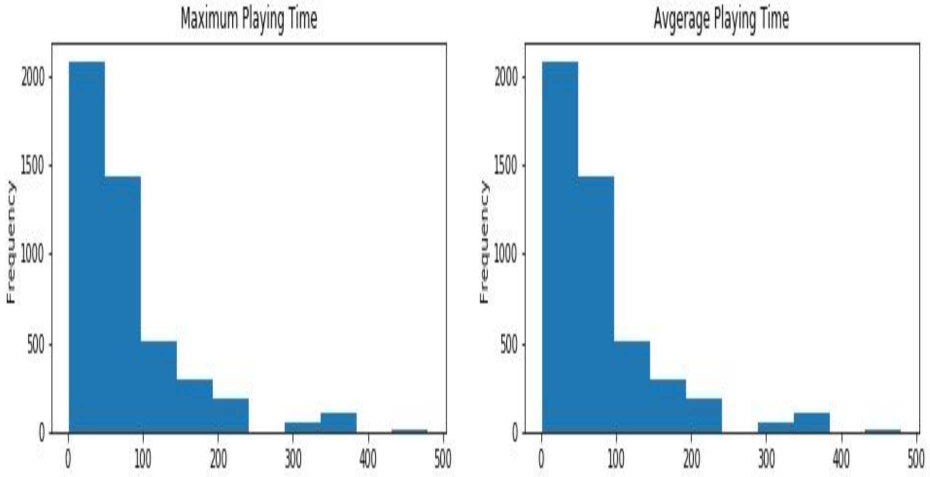


Figure 2

Looking at average time vs. max time in figure 2 we see that they are very similar. If they are it would be redundant to have both average time and max time. We will keep max time.

When we look at the data in the year column 97.5% of board games were released after 1974. I will be using that has the cut off for my analysis.

Exploratory Analysis

There are two different ratings in the data.

One is avg_rating thatis the average of all ratings from registered BGG users that the game has recieved, calculated by adding up all individual ratings.

The second is geek_rating that the rank column is based on. It is based on the avg_rating but the number is altered by BGG using an algorithm and beysian statistics to prevent games with relatively few votes climbing up to the top of the BGG ranting.

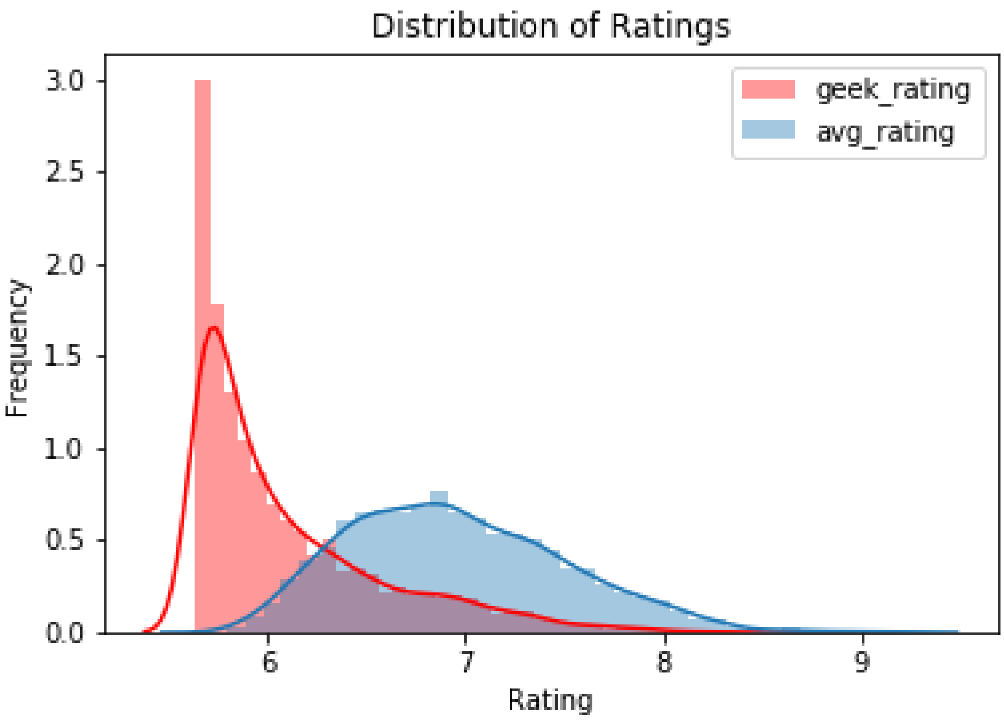


Figure 3

The distribution of the two ratings we see in figure 3 that geek_rating is skewed towards the left. This makes sence since the geek_rating is a ranking system that BGG uses. The avg_rating has much more of a normalized distribution.

An important part of any board game is how many people can play. There are 2 different columns that describes this: min_players and max_players. I added a new column by combining the min and max players to form a player range. The most popular player ranges is 2-4 players with 2-2 player games as the surprise. I don't play a lot of 2-2 player games but they are popular to make.

The average rating of 1-1 player games are the best rated but also have the fewest board games with 50 or more counts. 2-4 player games which have the greatest amount have a wide range of ratings.

Time is the ultimate currency for a board game. Are we going to spend all night to play one board game or are we going to play lots of shorter games to maximize the amount of exposure.

Play time ranges for most games are less than 2 hours. Also a lot of the games have the same min_time and max_time.

The weight column has an average rating by users on how complex a game is. It is a scale from 0-5. We need to keep in mind that BGG is user driven data and that the data might be skewed towards people who are involved in hobby. The hobbyist might not think a game is complex and rate is lower. But we will take the weight as a good measure. The weight distribution seems to be a normal distribution skewed a little to the left.

Looking at the relationship between age and weight we see that there are a lot of games that have an age of 0 but have a wide range of weight. This is weird. Babies will not understand a game with complexity over 4. I think these are missing values. We will replace all 0 in the age column with the mean of the age column

The owned column has the same shape as num_votes. There is probably a strong correlation between the two. This makes sense because people vote on the games they own.

We can see that there are 52 different Mechanics. We combined all mechanics of board games of 250 or less into a column of other_mechanic

96% designers haven't made more than 10 games. It seems that the designer category will have very little influence on predicting the rating of a game and will be dropped from the data frame.

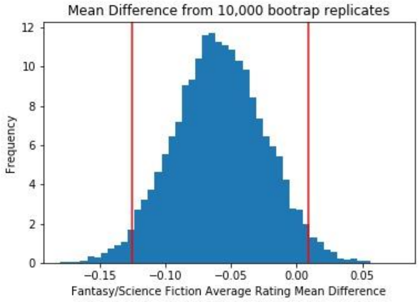
There are a lot of different themes of board games. Two of the most popular themes are Science Fiction and Fantasy. The question that I have is if there is a statistical difference between the two themes average score or geek score.

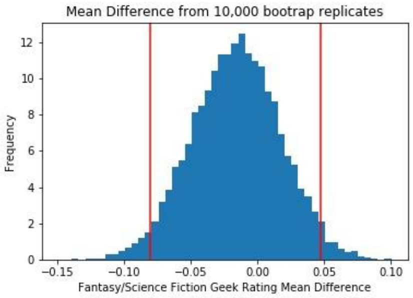
We see that the avg_rating and geek_rating between Fantasy and SciFi board games are very close. Using bootstrap replicates we will see if there is a statistical difference between the two.

The null hypothesis is that there is no difference between the two themes of board games while hypothesis claims there is a difference.

H0 = fantasy ratings = scifi ratings OR fantasy ratings - scifi ratings = 0

Ha = fantasy ratings != scifi ratings





The above histograms is the mean difference in bootstrap deviation of 10,000 samples. The red lines mark the 95% confidence interval. We see that the 95% confidence interval does contain 0 on both avg_rating and geek_rating. This means we can accept the null hypothesis that the avg_rating and geek_rating of the themes Fantasy and SciFi have no statistical difference.