Project Proposal

Crazy Eights - Kate, Bella, Shreyans, Daniel

Introduction

In an increasingly digital world, board games have facilitated a renewed interest in face-to-face communication, serving as a common human leisure activity that combines entertainment, social interaction, and intellectual challenge. In a 2018 study, Gonzalo-Iglesia and colleagues examined how Spanish students in Communication and Biochemistry studies engaged with commercial board games and discovered that engaging with board games promoted learning, communication, decision-making, and teamwork (Gonzalo-Iglesia et al., 2018).

Having understood the benefits of board games as a learning tool and beyond, the motivation for our research question originates from the rapidly increasing sales of board games, as enthusiasts often find themselves navigating through an abundance of unique gaming choices. This led us to wonder: If a person walks through an aisle of board games, what catches his or her attention first? More importantly, what makes a board game successful and enjoyable? How do various factors influence a game's reception among players?

In this analysis, we will focus on a singular broad research question: What qualities constitute a good board game? We'll be looking at elements like game themes, categories, mechanics, popularity, and game structure where a higher average user ranking denotes a better board game.

Before our exploratory data analysis, we crafted some hypotheses based on the context of board games. (1) Games that are owned more and have higher ranks will have a positive relationship with average user ranking because popularity typically increases rating. (2) Board games from the fighting category may have a positive relationship with average user ranking because of dedicated fan bases. (3) While different players might prefer more different levels of difficulty, we hypothesize that difficult games will have a negative relationship with average user ranking, on average, because it takes away from the fun and leisureness of board games.

The goal is to provide insights that can help both gamers and game creators navigate the world of board games more effectively.

Data description

This boardgames dataset involves data collected from BoardGameGeek(BGG), an online board game forum and database that stores information and ratings for over 125,600 different board games. The data was originally collected by web-scraping BGG's website using BGG's XML API, but it was sourced from Kaggle for the purpose of this analysis.

The initial boardgames data was split among several different files, but through data cleaning, merging, and variable selection, we reduced the size of the data significantly [see Section 3: EDA].

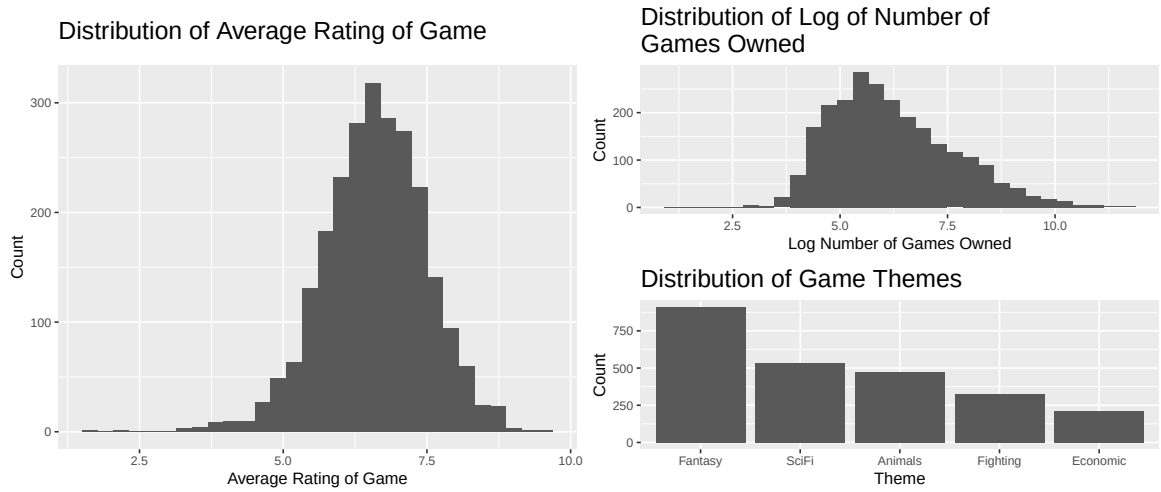
The final games dataset we use in our analysis has 1878 observations and 28 columns where 2 columns are identifiers (BBGid and name). There are characteristics that describe the structure of the game such as number of players, recommended time to play, and recommended age that are determined by the game designers. There are also more subjective characteristics that reflect the public opinion of each board game such as number of games owned/wanted, average rating, difficulty level, and average rank. Lastly, we also have a set of categorical variables that highlight what theme and subcategory each board game belongs to which gives more insight into the type of game and intended audience.

Initial exploratory data analysis

We performed a lot of data wrangling in order to prepare our data for exploratory data analysis and modeling. We started with three separate datasets (games, themes, subcategory) where the "games" dataset was our target dataset we wanted to tidy/merge to based on "BGGid". Our methodology is detailed below:

1. Select variables of interest from "games" dataset, drop the remaining.
2. To remove the number of dimensions and reduce future model bias from underrepresented categories and themes, we decided to only include board games in the top five populated themes and categories from each dataset. In order to make categorical distinctions more clear, we also removed all board games that belonged to multiple categories.
3. We performed an inner-join to merge all three datasets while also dropping observations with any NA's, reducing our observations to 1878.
4. Columns were renamed to remove spacing and special characters.
5. Since our themes and categories were represented as indicator variables, we created a new factor column for each that classified each board game to aid in graphing.

Distribution of response variable (AvgRating), quantitative predictor variable (log of NumOwned), and categorical variable (Theme)



The distribution of Theme shows that the most popular game type is Fantasy by a significant margin. There at least 200 more Fantasy games than the next most popular game theme of Science Fiction. The remaining categories from most to least popular are Animals, Fighting, and Economics.

Summary Statistics for Histogram of Average Rating

mean	sd	q0	q25	median	q75	q100	range
6.593	0.896	1.562	6.019	6.614	7.198	9.483	7.922

The distribution of AvgRating is approximately normal and unimodal. Thus, the best measure of the center is the mean, which is 6.639 points. The standard deviation of 0.844 indicates that there is little variability in the data, so the data is closely clustered around the mean. There are not any outliers shown in the histogram.

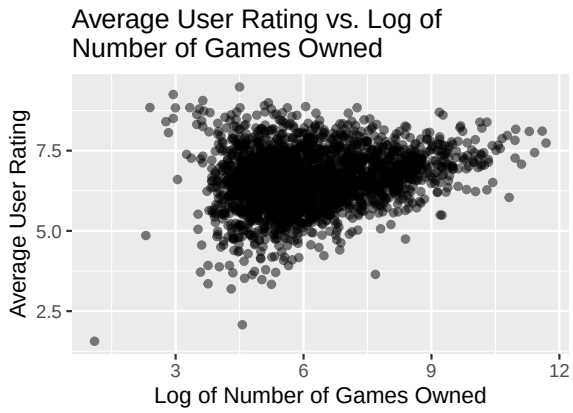
Summary Statistics for Histogram of Log of Number of Games Owned

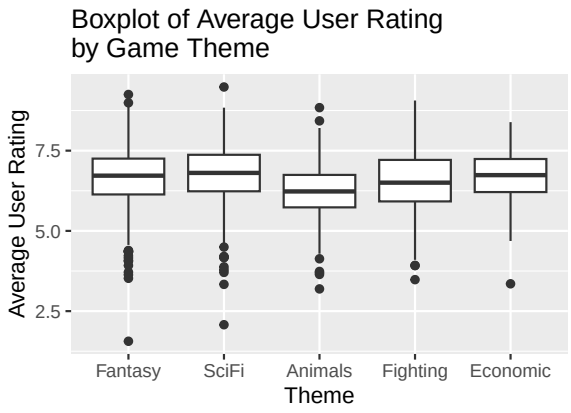
mean	sd	q0	q25	median	q75	q100	range
6.206	1.452	1.099	5.128	5.965	7.112	11.689	10.59

The distribution of the log of NumOwned is approximately normal and unimodal. We chose to take the log of the variable NumOwned, because distribution of the variable before the log transformation was left skew, with most points around 0. However, there were a few games where the NumOwned was over 50,000 games. Taking the log of NumOwned produced a distribution that was approximately normal. Thus, the best measure of the center is the mean,

which is 6.497 points. The standard deviation of 1.479 indicates that there is a moderate amount of variability in the data. There are not any outliers shown in the histogram.

Plots of Response Variable and Predictor Variables

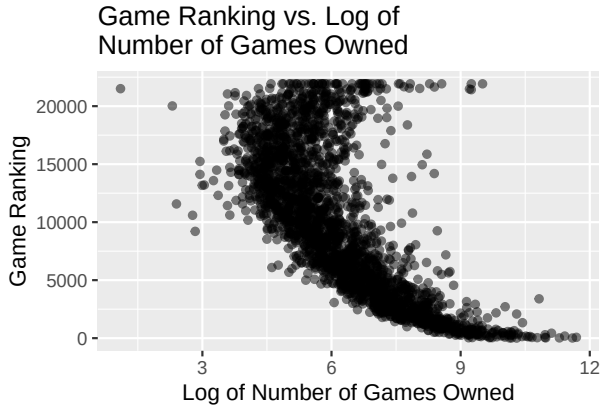


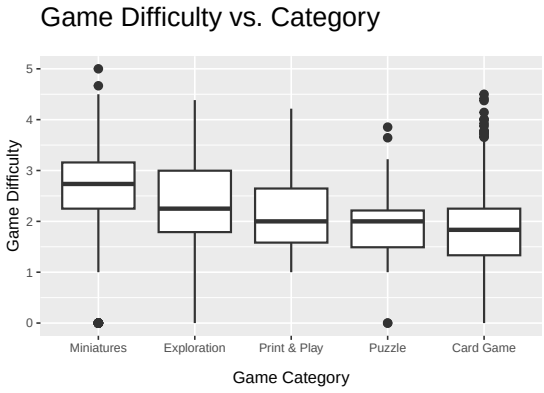


When plotting the number of games owned (continuous predictor) with the average user rating (response), we used a log transformation again. We can observe a moderate, positive, linear relationship between the log of number of games owned and user rating. However, there is a cluster on the top left, which suggests that the effect of the number of games owned on user rating is stronger when the number is very high.

When plotting the board game themes (categorical predictor) against the average user rating (response), we can observe that the median of average user rating for all themes fall between 6 and 7 with Fantasy, Scifi, and Economic themes above the 6.5 mark. Ranges and IQR's are pretty consistent as ratings typically fall between 5-9 with 50% of ratings in each theme within 0.75 points off the median. There are a few boardgame outliers with significantly lower ratings in each theme.

Visualizations of Potential Interaction Effects





Based on the context of board games, number of games owned could be correlated with game ranking (note: game ranking and user rating are not the same metric). The scatterplot above confirms this as on average, ranking number decreases (in other words ranking is better) when log of number of games owned increases. The relationship is strong, negative, and mostly linear but we would have to explore this effect in more depth.

It also seemed plausible that the category of game could correlate with game difficulty, another predictor. The boxplots above show that indeed some board game categories like Miniatures and Exploration have higher difficulty ratings on average. This interaction is good to investigate because we hypothesize that game difficulty will have a negative effect on user rating, and the presence of multicollinearity here will affect our model.

Analysis approach

The response variable is the average rating of a board game that ranges from a continuous score of 0 to 10.

The response variable will be predicted by 5 themes (Fantasy, Science Fiction, Fighting, Economic, and Animals) and categories (Card game, Miniatures, Exploration, Puzzles, and Print Play) along with the number of games owned by consumers, game difficulty, game ranking, and recommended play time. Predictors may be subject to change upon further exploration.

Interactions between the difficulty and categories will also be considered to assess whether the rating is affected if the mechanics of a game is harder to understand. Interactions with the ranking of the game and number of games owned will be assessed to determine if the rating a game is affected by consumer ranking given how popular a game is.

We will utilize a multiple linear regression model because we want to predict the average rating of a board game given the parameters of multiple predictors. Rather than saying whether a game is good or bad, we want to see how good a game is relative to others. This is why we utilize the rating system because all games are unique and therefore multiple games can be rated highly that would then be recommended for one's next game night.

Data dictionary

The data dictionary can be found here