

## Capstone Project 1: Milestone Report

BoardGameGeek is an online community of board game players. Classic games, such as “Monopoly”, are listed; however, a myriad of lesser known games such as “Council of 4” are also available. Information is provided for each game including the game’s ratings, number of players, description, type, category, etc. Some games have links to video tutorials on how to play. Within this community, you can buy and trade games from other users or find links to websites where you can purchase them.

Membership to this community is free, however, donations from users are encouraged. There are also ads on the margins of the webpages. One way to increase both these forms of revenue is to increase traffic to the website. A recommender system can help do this. It can also drive up sales of board games for which the website can charge a fee.

A recommender system predicts what rating a user would give an item or product. This information would then be used to suggest the higher rated items to the user for sales or other purposes. Recommender systems are used by many companies such as Spotify, Pandora, Netflix, and YouTube to create music or video playlists. Facebook, Instagram, and Twitter use recommender systems to suggest content and ads to their users. Amazon and eBay use it to suggest items that the user may want to purchase. In this capacity is how a recommender system can help BoardGameGeek increase traffic and sales. A user may not be aware of a game that they would be interested in. The recommender system would predict that the user may like the game and bring it to the user’s attention. This increases the likelihood that the game will be purchased by the user. It also increases the likelihood that the user would return to the website for future board game purchases.

In order to build an accurate recommender system, data is required on all the board games listed on the website. Information about the users, the games, the games’ ratings, etc. is provided by BoardGameGeek’s API. Using Python code, the data will be downloaded from the

API and wrangled into a dataset that can be used for analysis. This will be accomplished using the “xml.etree.ElementTree” Python tool. Later, sklearn tools will be used to build the recommender system from the dataset. The code and markup used to create the recommender system will be available on GitHub.

BoardGameGeek’s API was used to scrape the board game data from the website. Each board game has a page with its characteristics, descriptions, users that rated the game, and the ratings given. Each page holds 100 ratings. Any additional ratings are recorded in subsequent pages. Some pages contain information of items that are not board games, such as expansion packs. These pages were ignored. Data for the first 50,000 pages was scraped for all the relevant board game data.

The API uses XML, therefore, xml.etree was used to parse the data. The following information was obtained for each board game: ID number, name, year published, minimum players, maximum players, playing time, minimum playing time, maximum playing time, board game category, board game mechanic, and board game designer. Once this information was obtained, the code assesses if there are ratings available for the board game. If there are, the code then assesses how many pages worth of ratings there are. Next, the code iterates through all those pages and retrieves all the ratings and users for that board game.

This process is repeated until the code reads through the first 50,000 ID numbers, where each ID number corresponds to a unique item. Only those that for board games are stored in individual JSON files on the hard drive. This is wrangled data is now ready to be put in a data frame for analysis.

Exploratory data analysis, EDA, is one of the initial analyses that are performed on a dataset when searching for main trends, summarizing the data’s characteristics, and using visual methods to get a better understanding of the data. John Tukey defined EDA as "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier,

more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

Since the goal of this recommender system is to provide a user with suggestions on similar games that are also favored by other users, the average rating for each game was the logical place to start when comparing board games. I checked to see if there were any games that had no ratings at all. There turned out to be 7,206 games that had no ratings. This came out to be almost 20% of the dataset. These board games were ignored when obtaining the mean and median average rating. The data then showed that even though 80% of the board games had ratings, a vast majority only had a handful or less ratings per game. Figure 1 below illustrates this finding.

```
Board games with 0 ratings: 7206 , 19.10 percent of the data
Board games with less than 2 ratings: 11792 , 31.25 percent of the data
Board games with less than 3 ratings: 15097 , 40.01 percent of the data
Board games with less than 4 ratings: 17427 , 46.19 percent of the data
Board games with less than 5 ratings: 19130 , 50.70 percent of the data
Board games with less than 10 ratings: 23731 , 62.90 percent of the data
Board games with less than 20 ratings: 27220 , 72.14 percent of the data
Board games with less than 30 ratings: 28890 , 76.57 percent of the data
Board games with less than 50 ratings: 30800 , 81.63 percent of the data
Board games with less than 100 ratings: 32978 , 87.41 percent of the data
```

**Fig 1.** Percentage of data per number of ratings

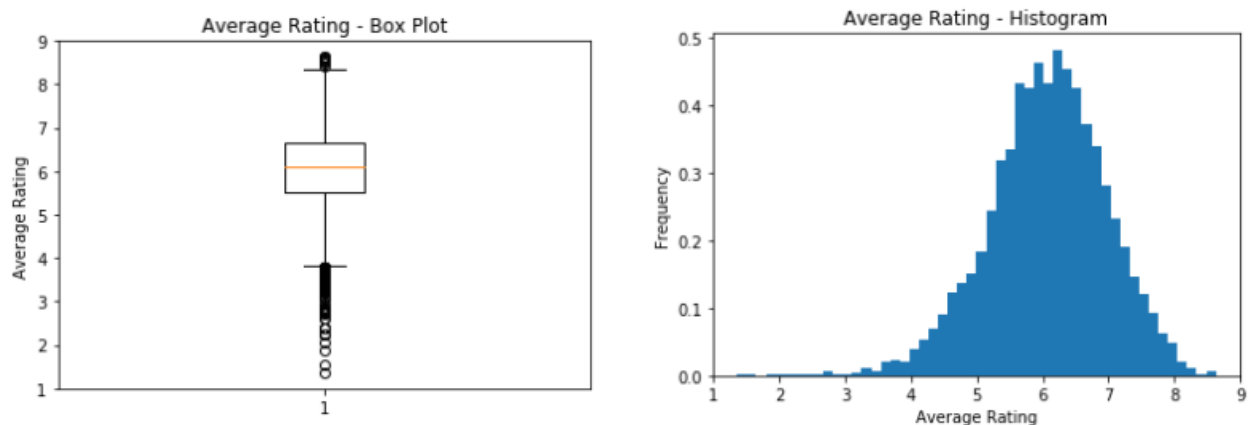
Since so much of the data had such few ratings. I compared the mean and median average rating and the mean and median number of ratings for each grouping. The data was grouped into games that had 1 or more ratings, less than five ratings, less than ten ratings, and less than 30 ratings. Figure 2 below shows the summary results.

Board games with 1 or more ratings:	Board games with less than 10 ratings:
Mean average-ratings: 5.455	Mean average-ratings: 5.873
Median average-ratings: 5.61	Median average-ratings: 5.93
Mean number of ratings: 223.182	Mean number of ratings: 482.67
Median number of ratings: 8.0	Median number of ratings: 48.0
Board games with less than 5 ratings:	Board games with less than 30 ratings:
Mean average-ratings: 5.742	Mean average-ratings: 6.068
Median average-ratings: 5.82	Median average-ratings: 6.11
Mean number of ratings: 364.914	Mean number of ratings: 754.42
Median number of ratings: 26.0	Median number of ratings: 113.0

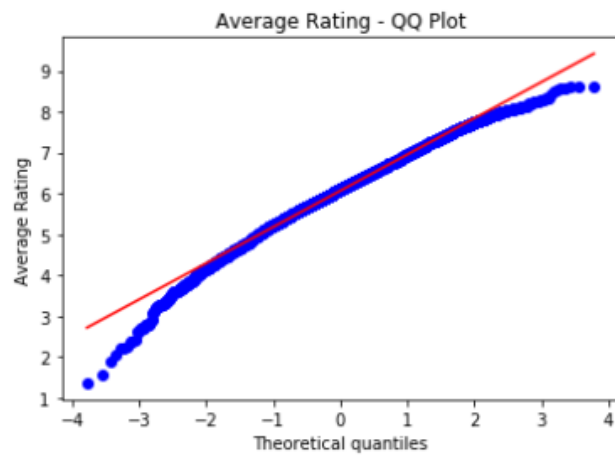
**Fig 2.** Summary results per group

The mean and median of the average ratings are close for all four groups. Since this is the case, it can be said that the dataset of board games with at least 30 ratings per game would reasonably reflect the games with fewer ratings.

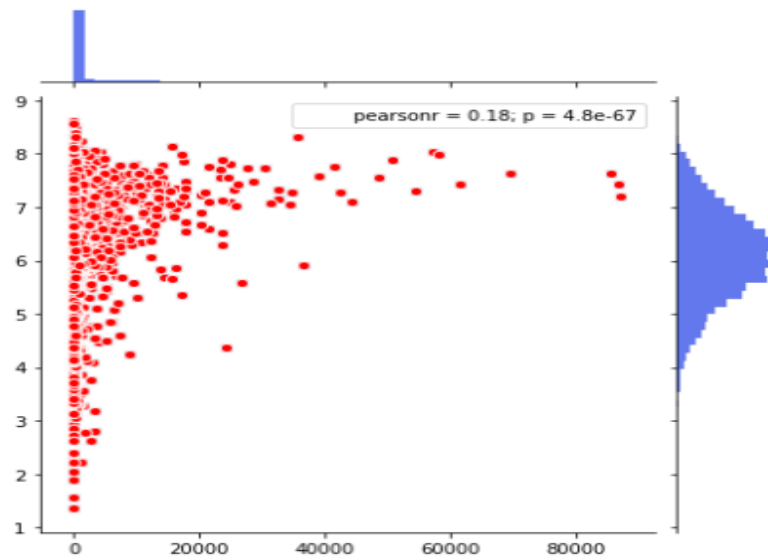
Data visualization tools were used to illustrate characteristics of the data. The average rating box plot below, Figure 3, shows that most of the average ratings users gave the games are between 3.5 and 8, with a mean average rating of about 6. Outside these bounds there are many outliers, however, most of the games are rated within this range. The histogram in Figure 3 shows that the data is normally distributed. The QQ plot, in Figure 4 below, shows this as well. However, it can be seen in the histogram that the data is slightly left skewed. This left skewness is more clearly seen in the QQ plot.



**Fig 3.** Box plot and Histogram



**Fig 4.** QQ plot



**Fig 5.** Joint plot

The joint plot in Figure 5 is a combination of a Number of Ratings vs. Average Ratings scatter plot, a histogram of the average ratings, and a bar graph of the number of ratings. The scatter plot shows the number of ratings vs the average rating of a board game. The plot shows that games with a higher number of ratings tend to receive a higher average rating. The histogram on the right demonstrates that the average, average-rating is around 6. The Bar graph on top illustrates that a vast majority of the board games have few ratings. The joint plot also contains a correlation analysis. The two variables are shown to have an  $r$  of .18 and a  $p$ -value of zero.

The correlation was then analyzed independent of the joint plot. First, the correlation between the average rating and the number of ratings was taken for games with at least 30 ratings. It was found to have a positive correlation of .182. The  $p$ -value for the correlation coefficient is zero, meaning that it is statistically significant. The correlation between the two variables was then taken without considering the number of ratings. Here, the correlation coefficient decreases slightly to .105.

The top twenty games in terms of average rating and number of ratings are shown in Figure 6 and 7 respectively. The top twenty includes only games that had at least thirty ratings.

RPGQuest: Greek Mythology  
The Battle of Fontenoy: 11 May, 1745  
RPGQuest  
Connection Games  
Prague: The Empty Triumph  
Sports Action Canadian Pro Football  
Crusade and Revolution: The Spanish Civil War,...  
Axis Empires: Totaler Krieg!  
1844: Switzerland  
Twilight Struggle  
The Penguin Book of Card Games  
Where Eagles Dare  
RPGQuest: Oriental Adventures  
Funkenschlag: EnBW  
D-Day at Omaha Beach  
DAK2  
Case Blue  
Manassas  
International Cricket  
Face To The Mat

**Fig 6.** Top 20 Board games with the highest average rating

Game	Average Rating (approx.)
Catan	4.8
Carcassonne	4.7
Pandemic	4.6
Dominion	4.5
Ticket to Ride	4.4
Agricola	4.3
Puerto Rico	4.2
Small World	4.1
Power Grid	4.0
Ticket to Ride: Europe	3.9
Citadels	3.8
Dixit	3.7
Race for the Galaxy	3.6
Stone Age	3.5
Munchkin	3.4
Twilight Struggle	3.3
Arkham Horror	3.2
Bohnanza	3.1
Lost Cities	3.0
The Resistance	2.9

**Fig 7.** Top 20 rated games

The list of games that received the greatest number of ratings has games that are widely popular and can be found at most toy and bookstores. It is interesting to note, however, that none of those games made it to the top twenty list of games by average rating. Evidently the number of ratings is not the only factor related to a game's average rating.

The correlation coefficient between average rating and number of ratings was .182. The correlation between the other attributes in comparison to average rating was calculated to search for any attributes that may be highly correlated. Figure 8 shows the attributes that had a positive or negative correlation greater than .1.

Correlation between a boardgame's average rating and bg mechanic Route/Network Building = 0.121  
 Correlation between a boardgame's average rating and bg mechanic Set Collection = -0.102  
 Correlation between a boardgame's average rating and bg mechanic Campaign / Battle Card Driven = 0.105  
 Correlation between a boardgame's average rating and bg mechanic Chit-Pull System = 0.100  
 Correlation between a boardgame's average rating and bg mechanic Hex-and-Counter = 0.278  
 Correlation between a boardgame's average rating and bg mechanic Area Control / Area Influence = 0.113  
 Correlation between a boardgame's average rating and bg mechanic Roll / Spin and Move = -0.279  
 Correlation between a boardgame's average rating and bg mechanic Dice Rolling = 0.159  
 Correlation between a boardgame's average rating and bg mechanic Simulation = 0.228  
 Correlation between a boardgame's average rating and bg category World War II = 0.215  
 Correlation between a boardgame's average rating and bg category Action / Dexterity = -0.104  
 Correlation between a boardgame's average rating and bg category Civil War = 0.108  
 Correlation between a boardgame's average rating and bg category Movies / TV / Radio theme = -0.181  
 Correlation between a boardgame's average rating and bg category Napoleonic = 0.133  
 Correlation between a boardgame's average rating and bg category Wargame = 0.367  
 Correlation between a boardgame's average rating and bg category Party Game = -0.123  
 Correlation between a boardgame's average rating and bg category Miniatures = 0.148  
 Correlation between a boardgame's average rating and bg category World War I = 0.239  
 Correlation between a boardgame's average rating and bg category Trivia = -0.160  
 Correlation between a boardgame's average rating and bg category Card Game = -0.115

**Fig 8.** Correlations greater than .1/ less than -.1

The list above shows that there is a high, positive correlation between wargames, World War I, and World War II games. There is a strong, negative correlation between a roll-spin-move games and that game's average rating. Hex-and-counter games, on the other hand, have a strong positive correlation. From the correlation results, it can be concluded that games that simulate war tend to receive higher ratings. In addition to this, games that do not rely on luck or “the roll of the die” tend to be more well-liked by users.

Board games that are categorized as “Wargame” have the highest correlation with the average rating. A two-sided t-test was conducted to see if the difference in the mean of the average rating between games categorized as “Wargame” and those that are not is statistically significant. The p-value of the two-sample t-test was far below .05 in the Frequentist Approach. Hence, we reject the null hypothesis that the mean of the average ratings of board games in the “Wargame” category is equal to that of board games not in the “Wargame” category. There is evidence to suggest that board games in the “Wargame” category do indeed receive higher average ratings than board games that are not in the category.