Ramona Herz
Sonnerain 6c
6024 Hildisrieden
r.herz@posteo.de
Lara Nonis
Maispracherstrasse 24,
4312 Magden
lara.nonis01@gmail.com

Data Science Project

Board games' rating predictor

Conceptual Design Report

31 October 2023

Abstract

The BoardGameGeek ratings count as the reference values in the board game scene. The rating mainly reflects the popularity of a board game based on the number and level of ratings. Using supervised and unsupervised methods, we would like to develop a model to predict the rating of a newly developed product or one in planning, based on the game characteristics as an indicator of expected sales success. The data will be collected using web scraping from the BoardGameGeek platform [1]. This conceptual design report (CDR) describes the used methods, data flows and models for this project.

Table of Contents

1 Project Objectives

Target of this project will be to develop a model able to predict user rating for newly developed board games based on representative game features. Thought as a complementary tool for a game developer, the model will be trained to predict the expected rating by the users and will allow to fine tune a game in development for a successful launch into the market. The expected rating may be then further used by the game developer company, to plan the resources meant to be invested in the new developed game in terms, for example, of maintenance or sponsoring The project will be split in two main phases:

- Phase I: to define the game features (or combination of features) that contribute the most to the game's rating. The features may include i.e. number of players, mechanism, play time.

- Phase II: use the findings from Phase I to build a model able to predict a new game's user rating.

The project will require a combination of unsupervised machine learning algorithm (phase I) and supervised learning techniques (phase II). The performance will be evaluated by a cost function. Due to the amount of data and from an evaluation of the rating variability over time, a batch learning algorithm will be used, frequency of the regular re-training, in order to prevent data drifting, is to be assessed with the first phases of the project but could be estimated as fortnightly or monthly.

2 Methods

The computational part of the project for both Phase I and Phase II will be performed using Python as programming language in Jupyter Notebook and JupyterLab through Anaconda®Distribution GUI [2]. Also will the project be documented and presented by means of Jupyter Notebook. Within the iPython environment at least the following libraries will be needed:

- Pandas: for data analysis and management.
- Numpy: for data analysis, data cleaning and, in combination with pandas, preparation for a machine learning library.
- Matplotlib/seaborn: for data plotting and analysis
- SciPy: for statistical evaluation on the data.
- Sicktlearn: for component analysis, model development and testing.

3 Data

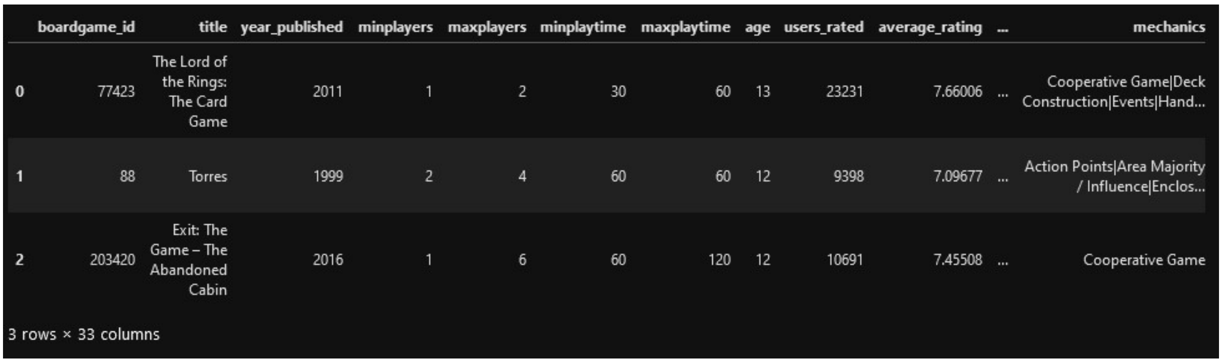
Data will be web scraped from a popular board games platform: BoardGameGeek [1] through its API [3]. At the time of the report a preliminary exploratory data analysis (EDA) has been performed on a feather dataset web scraped from the same platform on February 2023 and uploaded on Kaggle [4]. The newly web-scraped dataset will present the same columns as the inspected dataset, allowing an additional comparison between the two EDAs and therefore on the variability of users' preferences with time. This information could be used to define the frequency of model re-training in order to prevent data drift.

The dataset inspected [4], Figure 1, contains 2000 entries, being the most highly rated board games on the platform by February 2023 and 39 columns. After cleaning, the dataset was reduced to 28 columns. Two different rating are available within the website and were exported:

- Average rating: the mathematical average of the user's rate
- Bayesian rating (Bayes_average_rating): a more elaborated rating, using the average_rating and other parameters (not disclosed by the platform), ultimately used for the game's ranking.
  After the EDA the authors suspect, among others, a relationship between this value and the number of people owning the game (correlation coefficient 0.58), comparable with correlation coefficient of this value with the average rating (0.55).

The two average values are comparable among them and could be both potentially used for project Phase I and II.

For the scope of the analysis some additional columns were created within the data frame (5 additional columns) for a total of 33 columns in the cleaned dataset. Categorical variables with multiple combined entries (i.e. mechanics, categories) were split in single instances only in the plotting single algorithms through dictionaries to prevent long machine time related to the data frame weight.



	boardgame_id	title	year_published	minplayers	maxplayers	minplaytime	maxplaytime	age	users_rated	average_rating	...	mechanics
0	77423	The Lord of the Rings: The Card Game	2011	1	2	30	60	13	23231	7.66006	...	Cooperative Game|Deck Construction|Events|Hand...
1	88	Torres	1999	2	4	60	60	12	9398	7.09677	...	Action Points|Area Majority / Influence|Enclos...
2	203420	Exit: The Game – The Abandoned Cabin	2016	1	6	60	120	12	10691	7.45508	...	Cooperative Game

3 rows × 33 columns

Figure 1 - Head section of the dataset

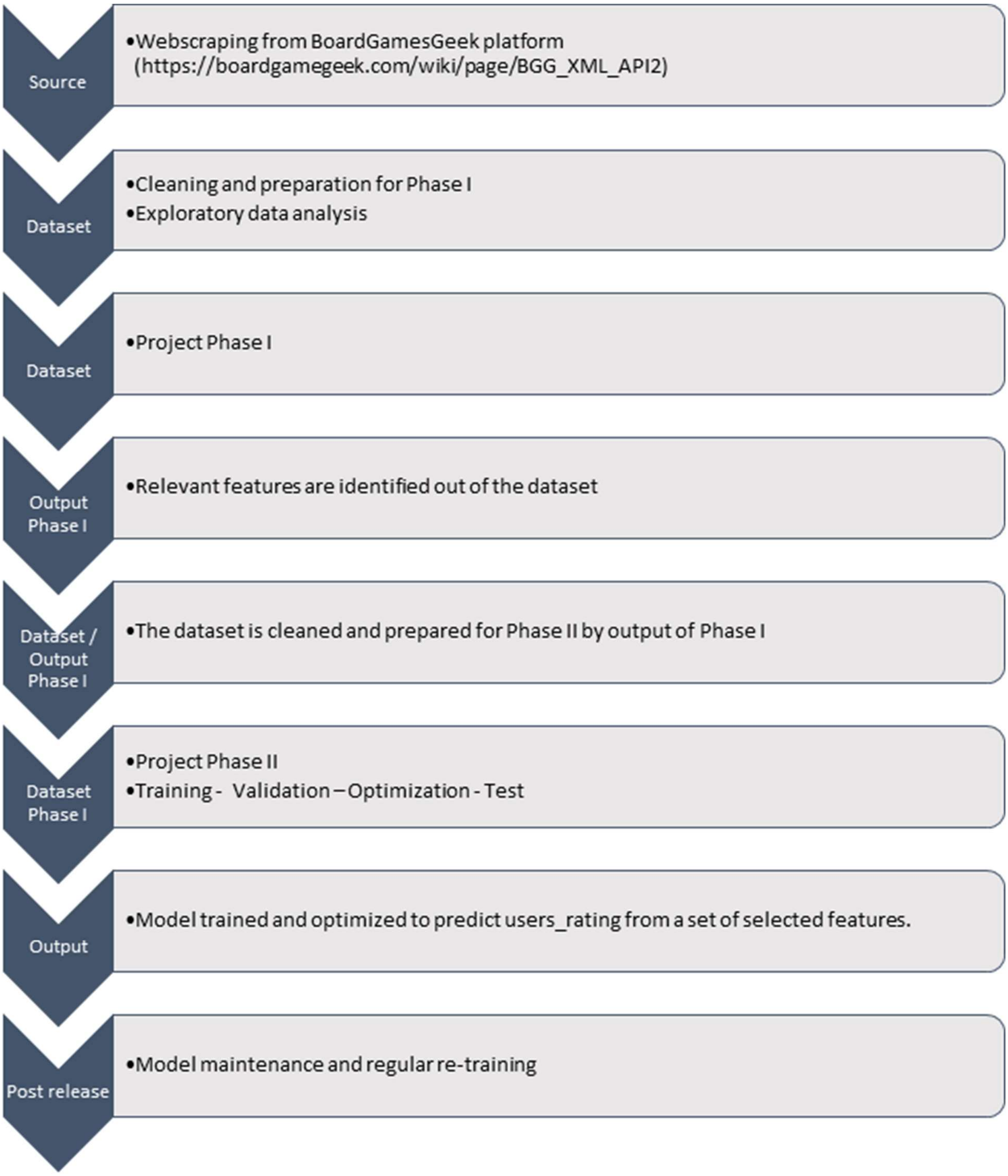
4 Metadata

The game information will be collected via scraping between 01.11.2023 and 10.11.2023 from the BoardGameGeek platform using our personal laptops. The available data for each game includes: game ID, game name, minimal number of players, maximal number of players, rated weight, average rating, bgg rating, number of expansions, year published, designers, artists, publishers, mechanisms, categories, sculptors, game family and game subtype. The game ID is used as identification for every included object. The metadata will be stored on our personal CAS github depository, where it will be publicly accessible.

5 Data Quality

The data quality depends on the information provided by boardgamegeek.com and it is yet unclear if we get a well balanced dataset, with a similar number of games to be included in the dataset from ratings between 1 to 10. Not having enough data or an unequal number of games for the different rating levels would definitely have an impact on the model quality. The handling of missing values may have to be reconsidered, if after exclusion of ID's with missing values other than sculptors leads to a high data loss. A compromise would be to include all games with the values game ID, game name, number of players, rated weight, year published, average rating and bayesian rating. For sculptors we will neglect missing data, as this variable depends on the game design - most of the games will not have a sculptor). There is the possibility that we are missing one of the most important information influencing the rating, the price of a game. This information could also be included subsequently through amazon.com if the model does not have the necessary quality.

6 Data Flow



Source

- Webscraping from BoardGamesGeek platform (https://boardgamegeek.com/wiki/page/BGG_XML_API2)

Dataset

- Cleaning and preparation for Phase I
- Exploratory data analysis

Dataset

- Project Phase I

Output Phase I

- Relevant features are identified out of the dataset

Dataset / Output Phase I

- The dataset is cleaned and prepared for Phase II by output of Phase I

Dataset Phase I

- Project Phase II
- Training - Validation – Optimization - Test

Output

- Model trained and optimized to predict users_rating from a set of selected features.

Post release

- Model maintenance and regular re-training

7 Data Model

Conceptual:

The data, exported and pre-treated as described in in Chapter 3, will be imported in the pandas environment, cleaned from duplicates, analyzed for missing values, all non useful variables will be dropped, the categorical and numerical variables will be analyzed, if necessary by statistical means. All categorical variables will be encoded (or one-hot-encoded) to be used in project Phase I.

Project Phase I will be an unsupervised learning task, the target will be to find patterns and define the most relevant features. This part will complete the brief exploratory data analysis performed manually and statistically on the dataset prior to Phase I. This first section may include clustering and dimensionality reductions algorithms.

After Phase I a second cleaning will be performed on the data to keep only the most relevant features and prepare the data for Phase II.

Project Phase II will be supervised learning task; a multiple regression problem where the system will use multiple features to make a prediction (the user rating), but also univariate regression problem (only one rating will be predicted for the new game).

Logical:

Output column will be bayes_average_rating (number), as described in Chapter 3, based among the rest on the average_rating for its calculation and used for the games' ranking.

The input columns will be all those of the cleaned dataset for Phase I, columns dropped were represented by columns with many missing values or with information coded within other columns (i.e. only 'expansions' column with 'yes/no' entries removing additional columns with specific information on the game's expansion).

For Phase II the input columns will be those suggested from the results of Phase I and from the EDA (number of players, playing time, weight, category).

Physical:

A regular laptop will have enough computational power to run the machine learning algorithms for both Phase I and Phase II. Data will be saved in a GitHub repository.

8 Documentation

During the project we will take into account the best practices, so the project (data / code) stays FAIR (findable, accessible, interoperable, reproducible) and ethical. The code and the data will be uploaded to GitHub, to our CAS GitHub repository. We do not use sensitive data, there is no anonymization needed. All the data is permanently publicly available already before the project starts. In addition to inline comments, applied naming conventions, consistency and meaningful variable names, a readme file will ensure that others are able to understand the key elements of the project (transparency).

9 Risks

Some variables are highly related - it is extremely important to recognize those relations before the model training. Further there are several variables we are not knowing beforehand during the developing phase of a game besides the rating (example number of raters). Such variables also have to be excluded for our purpose.

As there is a limitation of requests per time (1 request every 5 seconds) for the use of the API we may not be able to get all the data from the website in the necessary time. As the identificator is the game ID and we will start with the highest, so the newest games, we may have a bias and no older games included. If we recognize we are getting short in time, we could stop to follow this backward scraping in the timeline but take a range of IDs from older games, so they are included in the model too.

The comparison between results of our own data and the Kaggle dataset will be interesting, but there is a big difference in those 2 sets: the Kaggle set only includes highly rated games - so we do not really have an impression about the lower rated games. We need to be careful with an interpretation of this comparison. Lower rated games may also not have all the information needed to include them into the training dataset, which could lead to a situation where we have an unbalanced dataset with much more highly rated games included than lower ones. But this will only be seen after the scraping and should be taken into account when the dataset gets splitted into training and test data.

10 Preliminary Studies

A preliminary analysis on the available data (Figure 2) showed a range of publication date for the most appreciated board games spanning from the last 25 years. Only few games had an older publication date. Some were showing an anomalous number as publication date, these games were therefore excluded from this first evaluation but used for the following ones not affected by this value.

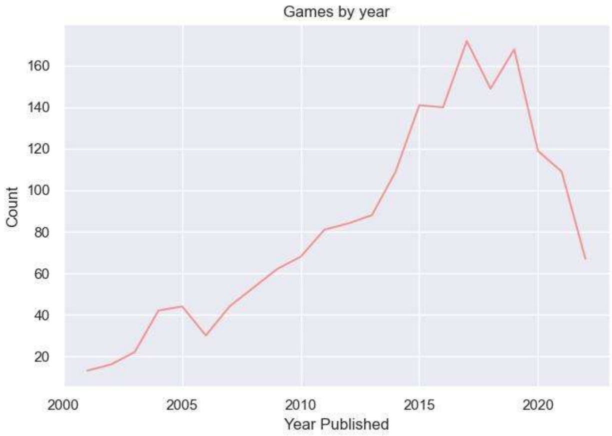


Figure 2 - Board games published by years

Correlation between the two rating values discussed above (average_rating and bayes_average_rating) versus all the numerical features of the dataset was explored to assess potential relationship by using a correlation matrix:

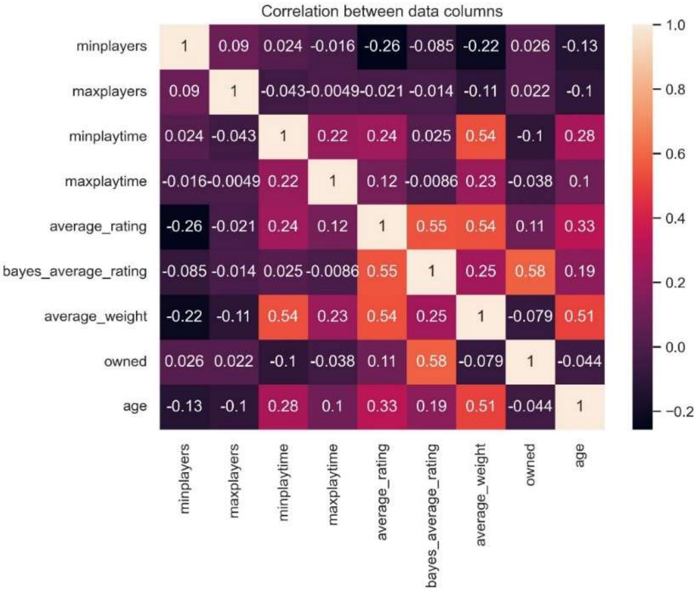


Figure 3- All versus all correlation

From the heatmap (Figure 3) some interesting relationship could be observed: as expected the minimum play time was correlated with the weight of the game (more complicate games require longer time to be completed), also expected was the correlation between the average_rating and the bayes_average_rating (where the average rating is used together with other factors to assign the rank). On this parameter, the authors suspect the number of users owning the game (owned) to be a parameter in the calculation of the bayes_average_rating as the correlation coefficient is comparable with the one of the correlation with the average_rating (known to be a parameter used for this value determination). Interestingly the games' weight was also connected with the rating (more complicate games tended to have higher ratings). Still involving the weight but more intuitive was instead the correlation between the age of the players and the weights (older people tends to appreciate more complicate games).

Common features in board games are the number of players, the time required for the game, the weight of the game, as mentioned above being its difficulty (in a scale from 1-5 from easy to hard), and the game's category. All these features were explored in details:

a) b)

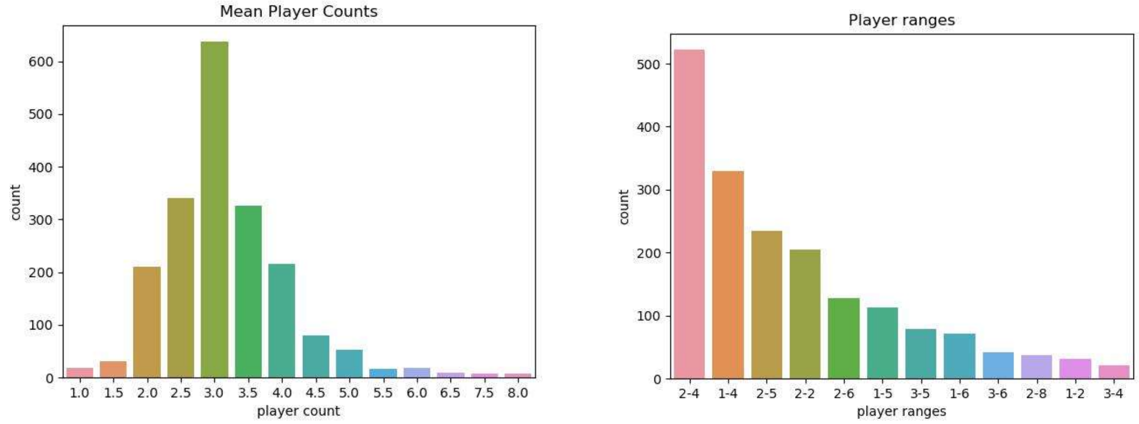


Figure 4 - Average number of players (a) and distribution of player ranges (b) for the games of the inspected dataset

By using the average only (Figure 4, a) the games with 3 players seemed to be the most appreciated. By moving from average number to ranges (Figure 4, b), the higher number of games within the most rated was 2-4 player games, followed by 1-4 players games. These ranges were in accordance with the averages.

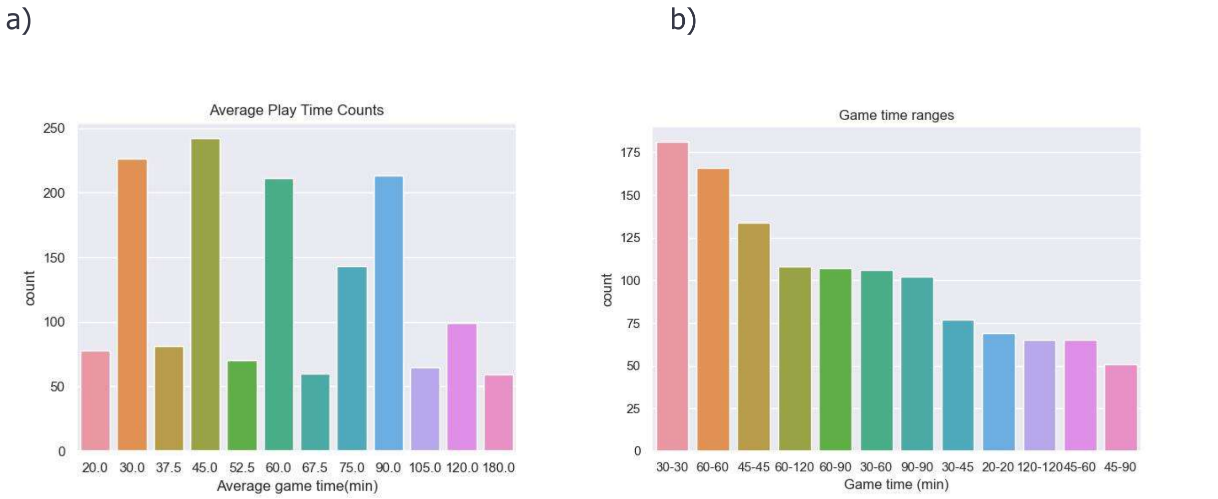


Figure 5- Average time required to complete the name (a), and range of time required (min, max) (b) for the games of the inspected dataset

In a similar manner also the game time was analyzed. The average (Figure 5, a) and the ranges (Figure 5, b) were in accordance on the conclusion that a relevant number of games within the most rated were relatively short games (30-60 minutes required), with the shortest (30 min games) being the most represented.

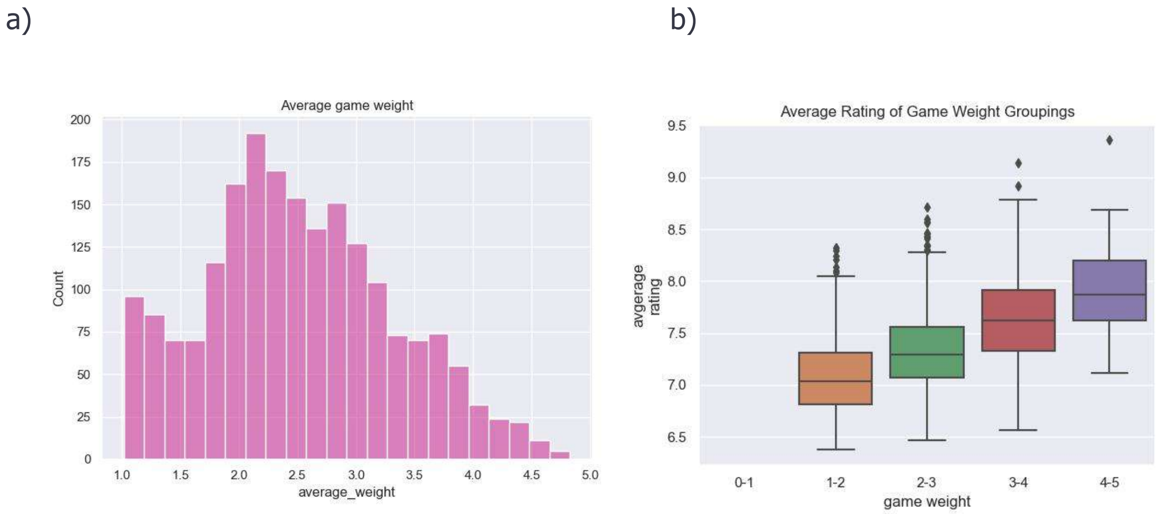


Figure 6- Average game weight (a) and game weight distribution against average game rating (b) for the board games of the inspected dataset

In accordance with what was observed through the heat map (Figure 3), very difficult games has generally higher ratings among the highly rated games, while easier games (weight 1-2) has the relatively lower ratings (Figure 6, b). This conclusion, together with the previous could suggest that in the development of a board game, a very difficult game could aim to higher rating from the users (boxplot) but this result is difficult to achieve (the number of games with height weight among the most highly rated is low, see Figure 6, a). It should be taken into account though, that it is not known here the absolute number of 'easy' games in the platform, vs 'difficult' games.

Where the first number is very high compared to the second this could be a possible explanation to the neat prevalence in Figure 6, a.

Lastly the game category was inspected:

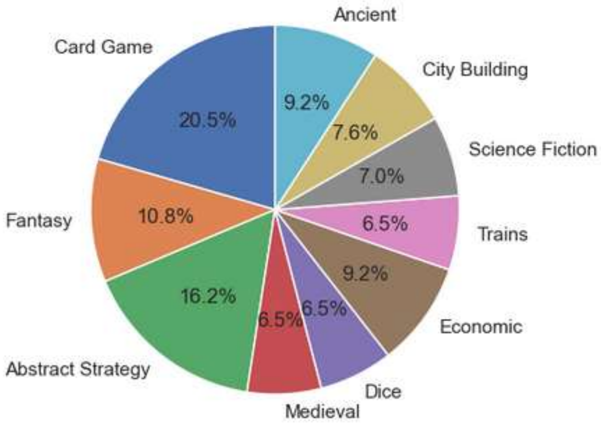


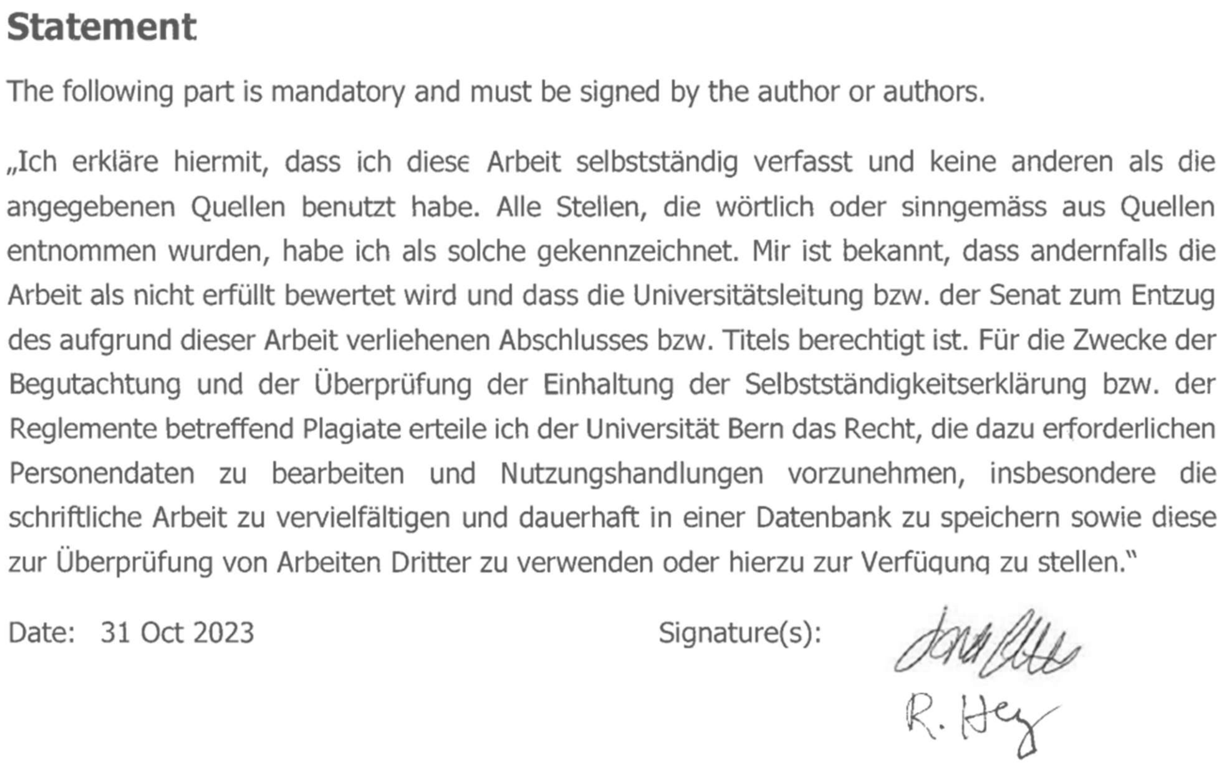
Figure 7- Board games by category

Among the most rated games, the bigger amount (20.5%) belonged to Card Games category, followed by Abstract strategy (16.2%) and Fantasy (10.8%) (Figure 7).

Basing on this preliminary EDA a board game, in order to aim for high rating should be: for 2-4 player, with a playing time between 30-60 minute, of average difficulty (2-3) and should be a card game. Additional exploration of the dataset may be required for more deep and specific requirements. Furthermore, should be highlighted that the choice of the parameter to be inspected was subjective and based on the features that intuitively could play a role in the appreciation of a board game. A machine learning algorithm may be able potentially to identify more, without the subjective bias, or combination of factors that play a role and could be used, eventually, as decision maker for a developer (i.e. 30 min vs 60 min games).

11 Conclusions

At this stage, there are still many uncertainties. Some of them will have to be discussed again after the data has been obtained and analyzed, especially the number of games being part of the dataset, information influencing the rating the most and the handling of missing values. On the basis of the preliminary studies carried out on the available data set and the existing resources, there is a reasonable likelihood that the project will be feasible. The accuracy of the ranking estimation from Phase II will be strongly dependent on the results from Project Phase I. If no significant result will be output from Project Phase I, field knowledge and relevant parameters from similar data analysis [5][6][7] will be used as predictive features.



Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen."

Date: 31 Oct 2023

Signature(s):

R. Herz

References and Bibliography

[1] BoardGameGeek. (2023). BGG. https://boardgamegeek.com

[2] Anaconda Software Distribution. (2020). *Anaconda Documentation*. https://docs.anaconda.com

[3] BoardGamesGeek API, https://boardgamegeek.com/wiki/page/BGG_XML_API2

[4] Kaggle, Top 2000 Board Games Ratings, Nikita Fedorov, https://www.kaggle.com/datasets/nfedorov/top-2000-board-games-ratings

[5] Kaggle, Board Game Analysis, 2016, mrpantherson https://www.kaggle.com/code/mrpantherson/board-game-analysis/notebook

[6] Kaggle, Board Games Dataset Exploratory Analysis, 2017, Gabriele Baldassarre https://www.kaggle.com/code/gabrio/board-games-dataset-exploratory-analysis/report

[7] Kaggle, Data Viz. What to focus on as a game dev?, 2019, hkhoi https://www.kaggle.com/code/hkhoi91/data-viz-what-to-focus-on-as-a-game-dev/notebook