Analyzing Board Game Ratings

Insights from User Engagement and Game Features

Jheronimus Academy of Data Science

June 2024

Analyzing Board Game Ratings - Insights from User Engagement and Game Features

Paulo Henrique da Silva Mota | SNR 2134751

Summary

Data Preparation and Description

Data Source

The data source pertains to BoardGameGeek Reviews and is composed of two datasets. The first dataset contains details about various board games, and the second dataset includes reviews from users for these games, linked by an ID. All reviews are from users, and their comments are included.

`Details Dataset:`

`Number of observations: 21631`

`Number of columns: 23`

The Board detail dataset contains information about many board games, where which obeservation represent a board game.

Table 1: Board game details dataset

variable	class	description
num	double	Game number
id	double	Game ID
primary	character	Primary name
description	character	Description of game
yearpublished	double	Year published
minplayers	double	Min n of players
maxplayers	double	Max n of players
playingtime	double	Playing time in minutes
minplaytime	double	Min play time
maxplaytime	double	Max play time
minage	double	minimum age
boardgamecategory	character	Category
boardgamemechanic	character	Mechanic
boardgamefamily	character	Board game family
boardgameexpansion	character	Expansion
boardgameimplementation	character	Implementation
boardgamedesigner	character	Designer
boardgameartist	character	Artist
boardgamepublisher	character	Publisher
owned	double	Num owned
trading	double	Num trading
wanting	double	Num wanting
wishing	double	Num wishing

`Ratings Dataset:`

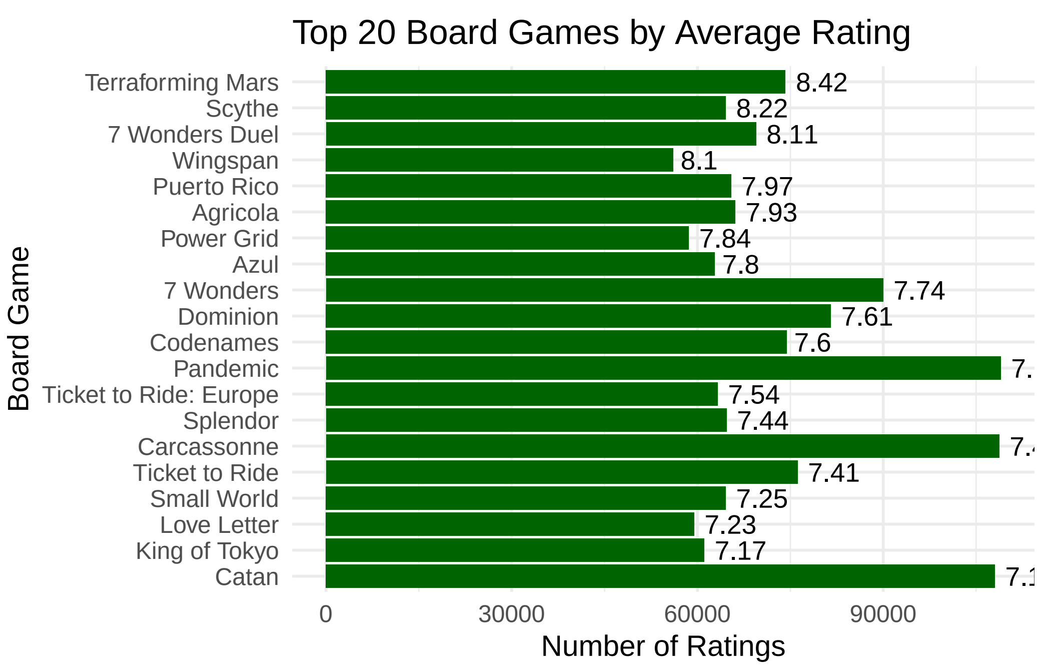
`Number of rows: 21831`

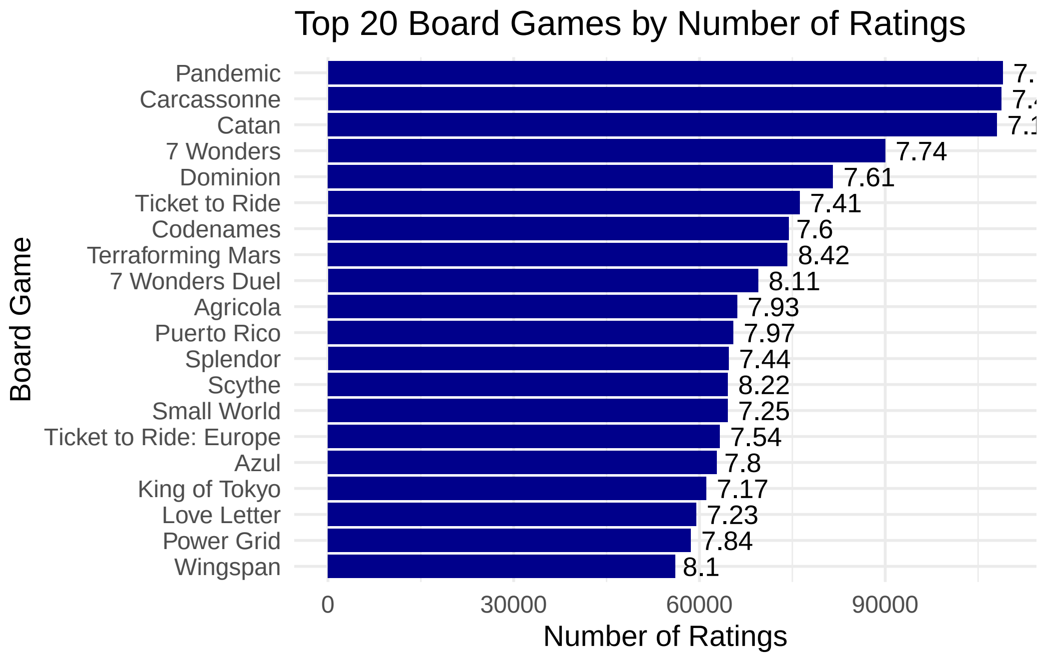
```
Number of columns: 10
```

Table 2: Ratings dataset

variable	class	description
num	double	Game number
id	double	Game ID
name	character	Game name
year	double	Game year
rank	double	Game rank
average	double	Average rating
bayes_average	double	Bayes average rating
users_rated	double	Users rated
url	character	Game url
thumbnail	character	Game thumbnail

The first columns that catches my attention was the number of ratings and the rating average values, and I decided to make a trend to compare which other to see the results of the rating score changes if we have different number of ratings.





Data Preparation

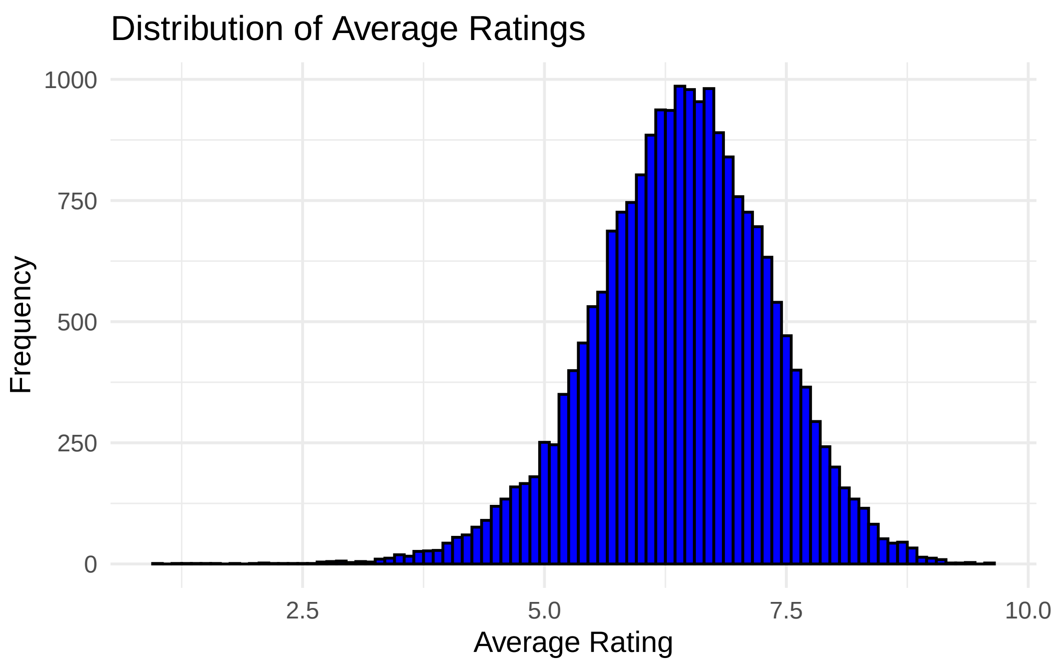
Detail the steps taken to clean and preprocess the data in R. Include code snippets where appropriate.

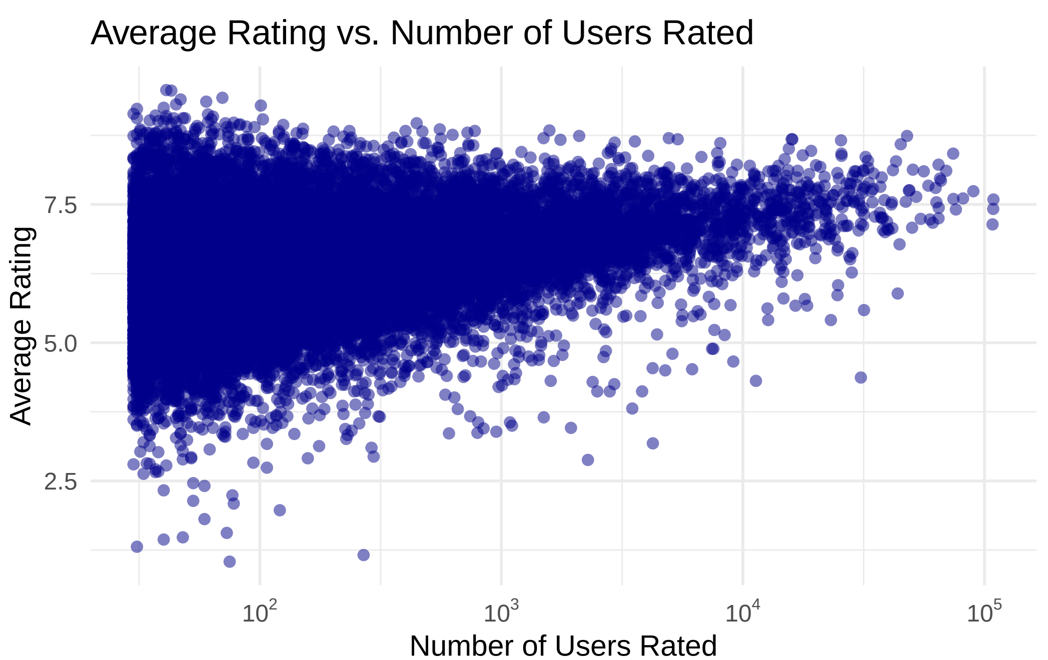
In order to have a better dataset to make our analysis, first I merged details with ratings by the column ID, following the instructions in the dataset documentation, with my dataset merged I will remove all games with no ratings information. After that I pliot the correlation matrix to identify possibly correlation with my variables.

Performance Measures

In the context of profitability within a board game, two critical performance metrics stand out: the average rating, represented by the 'average' column, as higher ratings typically signify superior market reception; and the number of reviews, captured by the 'users_rated' variable, as a greater volume of reviews tends to correlate with increased sales and heightened popularity.

The charts below illustrate the distribution of the performance metrics:


Distribution of Average Ratings
Frequency
1000
750
500
250
0
2.5
5.0
7.5
10.0
Average Rating


Average Rating vs. Number of Users Rated
Average Rating
7.5
5.0
2.5
10^2
10^3
10^4
10^5
Number of Users Rated

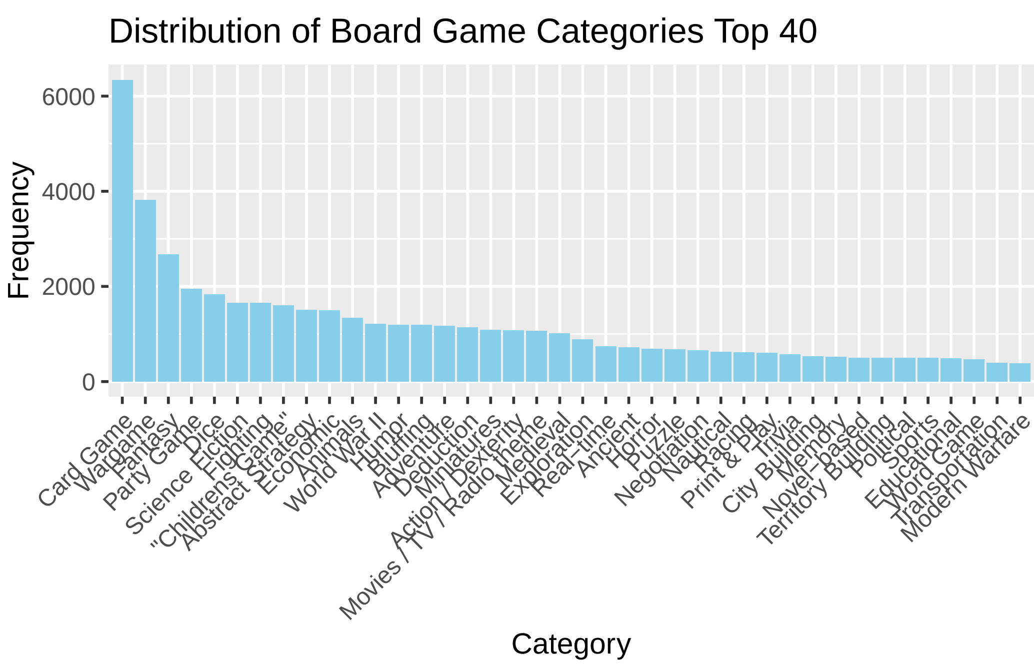
Key Variables

The variables influencing game design decisions within our dataset are closely intertwined with the intrinsic traits of board games. These variables encompass Category and Mechanic where specific genres may demonstrate stronger financial performance, and Play Time, which profoundly shape a game's attractiveness to potential audiences. Moreover, MaxPlayers act as indicators of the potential player base, potentially aligning with increased review counts.

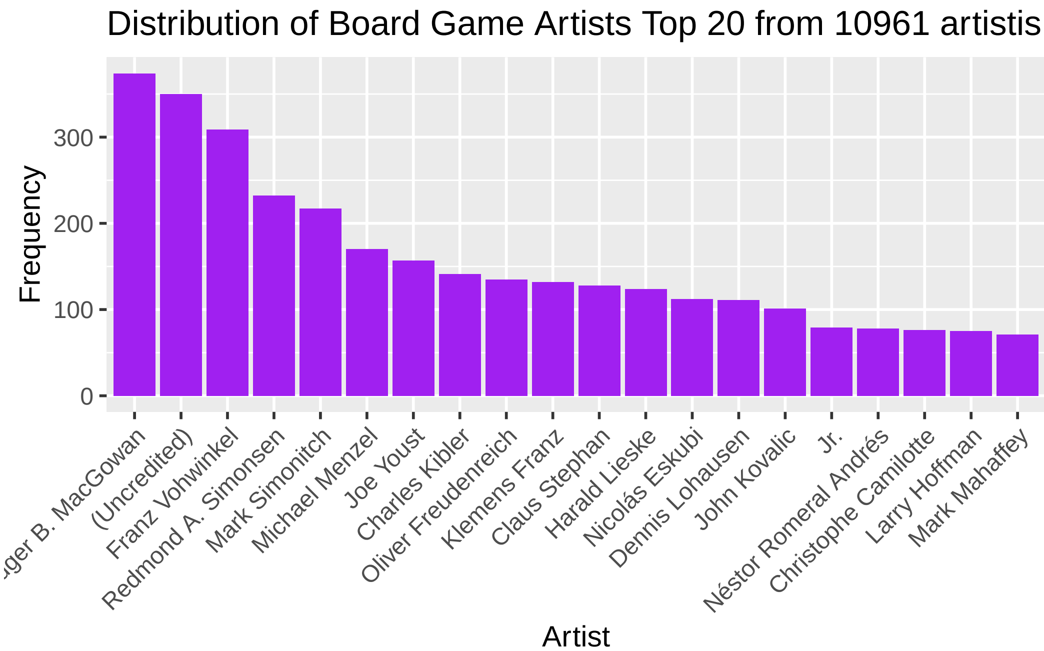
In addition to these variables, it's crucial to consider certain control variables to ensure a comprehensive analysis. Control variables such as Year of Release, represented by "Year Published," are vital as newer games may experience different market dynamics compared to older ones. However, factors like Artist, Designer, and Publisher reputation play significant roles and could impact the outcomes of success.

Data Summary

The chart below you can see the distribution form the categorical values and relaize that the top categories may have the best probability to be more sucessfull, for being high trended in the data. The same reason is applied to the board game mechanic.



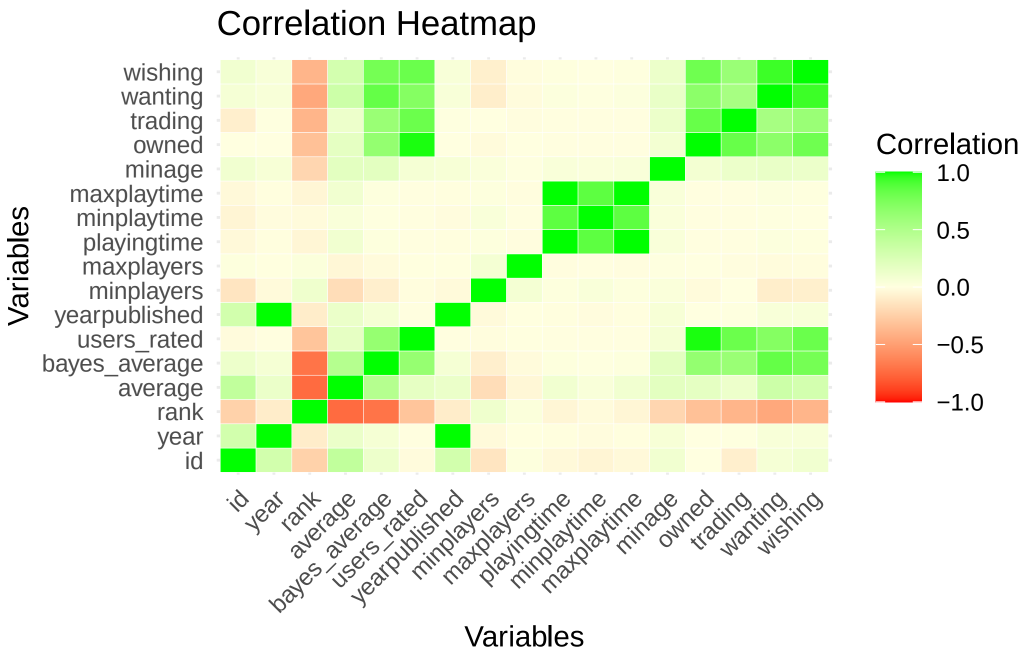
About the artist,design and publishers, it is more about reputation, some new fame with the right names could end up successful just because the best experience that those users had with a previous game. These people and companies become well now between the community with more credibility.



Data Analysis

Relationships with Performance Measures

Analyzing the correlation was my initial step in identifying potential variables to include in my linear regression model. This analysis helps to determine the relationships between the measured variables. We observed that the variables "owned," "wishing," and "wanted" are highly correlated with our measured variables. This suggests that by increasing the average rating and the number of user reviews, we can potentially improve the profitability of a board game.



Modeling and Evaluation

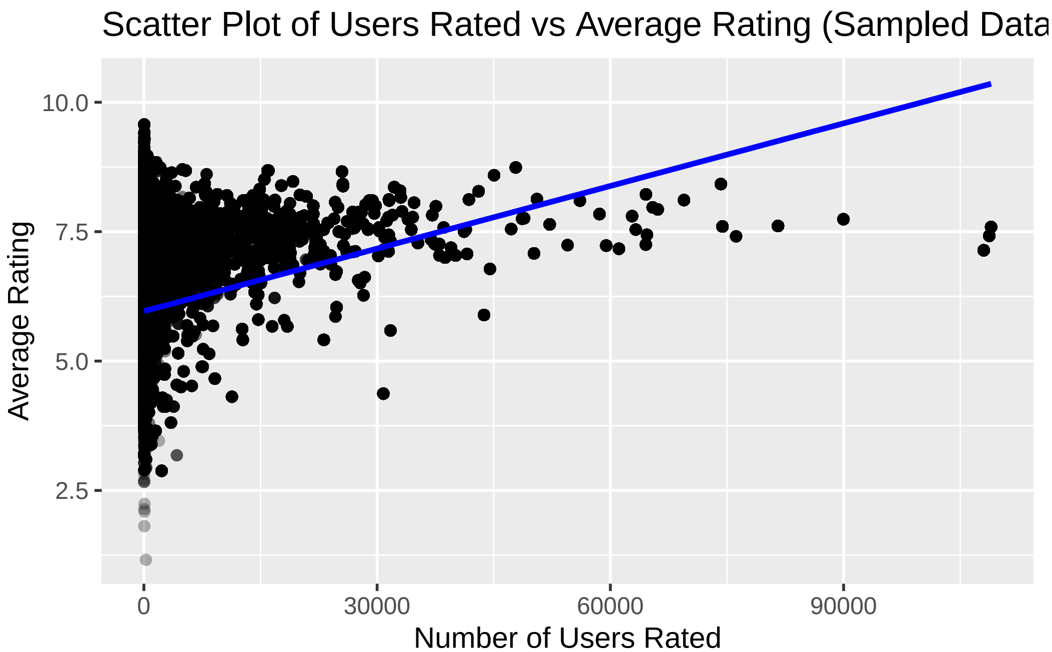
```
Call:
lm(formula = average ~ users_rated + maxplayers + playingtime +
    yearpublished + boardgamecategory + boardgamemechanic + boardgameartist +
    boardgamepublisher + boardgamedesigner, data = clean_data)

Residuals:
    Min      1Q  Median      3Q     Max
-35.133  -0.692  -0.075   0.735   6.633

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.666e+00  4.818e-03  760.99   <2e-16 ***
users_rated         2.698e-05  4.176e-08  646.02   <2e-16 ***
maxplayers         -2.180e-02  9.551e-05 -228.28   <2e-16 ***
playingtime         5.684e-04  2.150e-06  264.36   <2e-16 ***
yearpublished       1.050e-03  2.223e-06  472.21   <2e-16 ***
boardgamecategory   4.384e-03  2.436e-05  179.95   <2e-16 ***
boardgamemechanic  -2.080e-03  1.074e-05 -193.65   <2e-16 ***
boardgameartist    -1.327e-05  1.725e-07  -76.93   <2e-16 ***
```

```
boardgamepublisher -3.973e-05  2.862e-07 -138.81   <2e-16 ***
boardgamedesigner   2.109e-04  1.858e-07 1135.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.002 on 3383229 degrees of freedom
  (128350 observations deleted due to missingness)
Multiple R-squared:  0.507, Adjusted R-squared:  0.507
F-statistic: 3.865e+05 on 9 and 3383229 DF,  p-value: < 2.2e-16
```

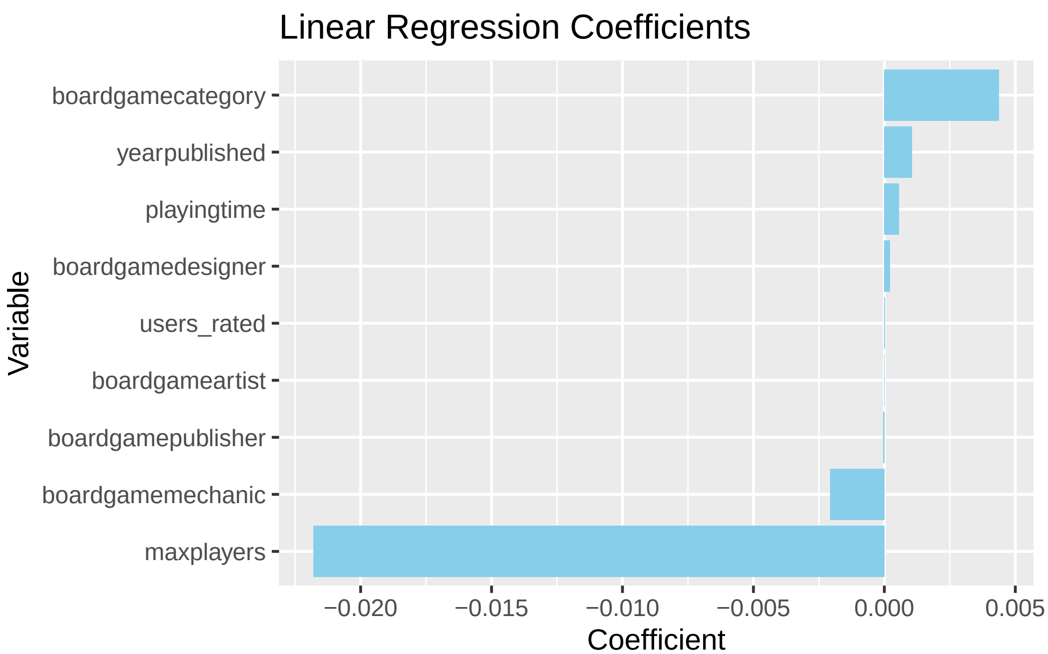


The call shows the command that I used to train my linear regression and below we can see different results and metrics from this object.

The linear regression model you've fitted to predict the average rating of board games based on various features such as the number of users who rated the game, maximum number of players, playing time, year published, and categorical variables including board game category, mechanic, artist, publisher, and designer. The coefficients indicate the estimated change in the average rating for a one-unit increase in each predictor variable, holding other variables constant. For instance, the positive coefficient for users rated suggests that as the number of users who rated a game increases by one, the average rating tends to increase by approximately 0.000027, all else being equal. Similarly, the negative coefficient for max players indicates that, on average, games with more maximum players tend to have slightly lower ratings.

The diagnostic results at the bottom provide insights into the overall performance and assumptions of the model. The residual standard error indicates the average amount by which the observed values deviate from the predicted values, providing a measure of the model's accuracy. The multiple R-squared value of 0.507 indicates that the model explains approximately 50.7% of the variability in the average ratings, suggesting moderate predictive power. The F-statistic tests the overall significance of the model, with a very low p-value indicating that at least one of the predictor variables is significantly related to the average rating. Overall, the model appears to be statistically significant and explains a substantial portion of the variability in average ratings, but further exploration and refinement may be beneficial.

Results Presentation



Strategy Recommendation

Based on the analysis, several factors significantly influence the average rating of board games. The number of users who rated the game, the year published, and specific categorical attributes like the game category, mechanics, and designer have substantial impacts. Therefore, focusing on increasing user engagement and incorporating popular game mechanics and categories can be beneficial. Designing games that cater to current trends and involving renowned designers can also enhance the game's appeal and improve ratings. To create a successful and profitable board game, it is essential to encourage more players to rate the game by enhancing user experience, promoting active participation, and implementing effective marketing strategies.

Additionally, developing games that align with popular categories and mechanics can attract a broader audience, and engaging well-known designers can boost the game's credibility and attractiveness. Staying updated with industry trends and adjusting game elements to match the preferences of contemporary players is also crucial.

Limitations

The analysis has certain limitations.The linear regression model simplifies complex relationships and may not capture non-linear interactions or other intricate patterns in the data. Furthermore, the model only includes a specific set of variables and may not account for other important factors influencing game ratings, such as marketing efforts, social media presence, or community engagement.Finally, all data is generated by users and could be bias to our analysis.

Additional Data and Analysis

To further inform the strategy and enhance the analysis, including marketing and social media data could provide deeper insights into the impact of marketing campaigns and social media interactions on game ratings. Collecting data on user demographics and feedback can help to build games to specific target audiences and improve user satisfaction. In addition, analyzing data on competitors' games and their performance can offer valuable strategic insights. Exploring non-linear models and interaction terms can also help capture more complex relationships and improve predictive accuracy.

Conclusion

The key findings highlight the importance of user engagement, game categories, mechanics, and designer reputation in determining the average rating of board games. By focusing on these aspects and incorporating additional data, game developers can design successful and profitable board games. Despite the limitations, this analysis provides a strong foundation for strategic decision-making in game design.

References

[1] 'How to Create Your First Board Game | Bits + Pieces', BoardGameGeek. Accessed: Jun. 09, 2024. [Online]. Available: https://boardgamegeek.com/blogpost/96723/how-to-create-your-first-board-game

[1] ‘Browse Board Games | BoardGameGeek’. Accessed: Jun. 09, 2024. [Online]. Available: https://boardgamegeek.com/browse/boardgame