Predicting Board Game Average Ratings

Statistically Significant - Darli Seranaj, Jacob You, Tuna Korkmaz

2025-04-28

Introduction and Data

Context

Board games have experienced significant growth in popularity, with thousands of new titles published annually. A key determinant of a game's success is its average user rating (AvgRating), reflecting player engagement and satisfaction. Understanding which attributes contribute to higher ratings can be valuable for game designers, publishers, and retailers.

Research Question:

Can we predict a board game's average rating (AvgRating) based on game attributes such as Number of Expansions, game complexity, number of copies owned and user engagement metrics?

Motivation:

1. Game designers can use these insights to enhance engagement and create better-received games.
2. Publishers can refine marketing and production decisions based on influential attributes.
3. Retailers and distributors can optimize inventory by understanding factors that contribute to popularity.

Hypothesis:

We expect that a combination of game complexity (GameWeight), player engagement metrics (NumOwned, NumUserRatings), and wanting or wishing (NumWant, NumWish) to have a game will be strong predictors of an Average Rating for a board game (AvgRating). Games with deeper strategic depth, and strong community engagement are likely to receive higher ratings. We also will explore if themes and mechanics can explain average rating of a game.

Dataset Information

The dataset was sourced from BoardGameGeek (BGG), a leading database for board games containing . The data curator used BGG's API to compile the data into different sheets, and published them on Kaggle.

We focus on three tables:

1. games.csv: Contains 21,925 board games with key attributes including ratings, complexity, player count, and user engagement metrics.
2. themes.csv: 217 binary indicator variables representing thematic categories (e.g., Sci-Fi, Adventure, Fantasy).
3. mechanics.csv: 157 binary indicator variables representing game mechanics (e.g., Dice Rolling, Deck-Building, Worker Placement).

BGG primarily gets its data from its users, who can submit board game information through BGG's API. BGG then takes this data and aggregates it, generating overall rankings, average rankings, playtime, themes, and other data, using moderators to ensure accuracy and consistency.

Exploratory Data Analysis

Key response variable:

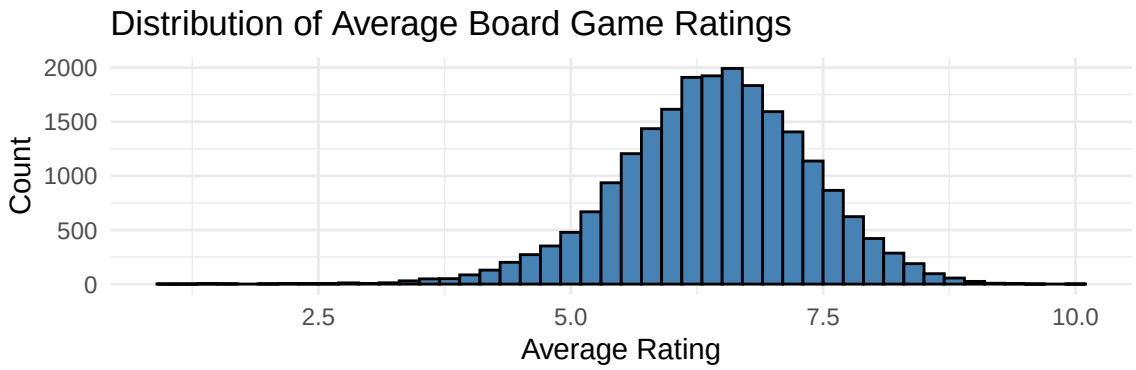
- AvgRating: The average rating given by users on BoardGameGeek.

Key potential predictor variables:

- NumUserRatings: Number of users who have rated the game.
- NumOwned: Number of users who own the game.
- NumWant / NumWish: Measures of user interest in having the game.
- MinPlayers / MaxPlayers: Range of players the game supports
- MfgPlaytime: Manufacturer-stated play time.
- GameWeight: An indicator of how difficult or complex the game is.
- NumExpansions: Shows the number of expansions a game has
- LanguageEase: How easy is it to understand the game.
- Theme: Encoded thematic categories (e.g., fantasy, sci-fi).
- Mechanic: Encoded mechanics categories (e.g., deck-building).

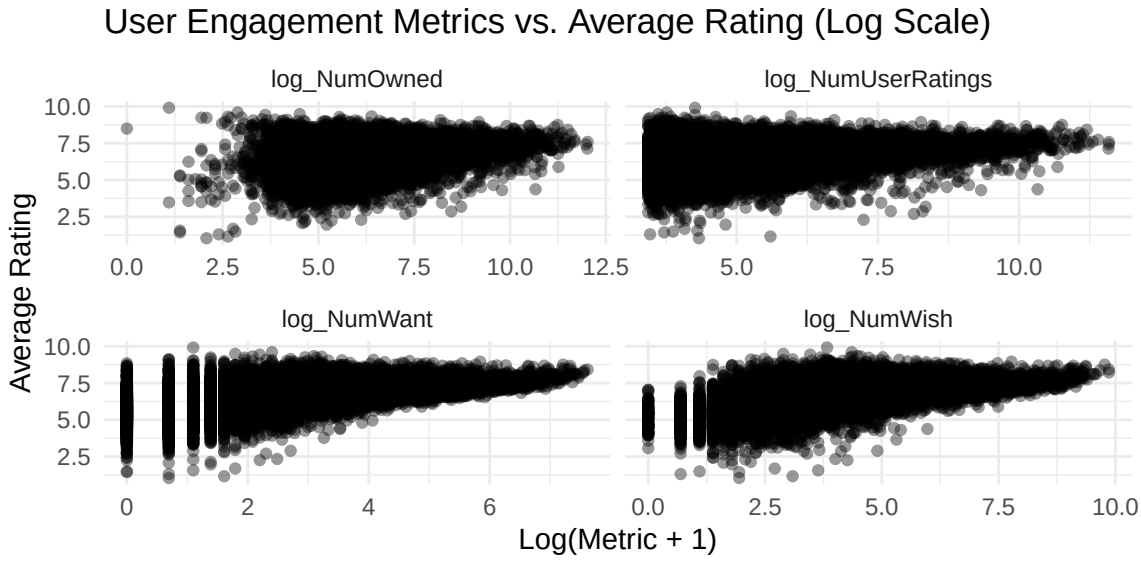
Univariate EDA

Response EDA - Average Rating



The distribution of AvgRating is approximately normal, centered around a mean of 6.425 and median of 6.454, indicating symmetry. Most ratings lie between 5.8 (Q1) and 7.1 (Q3), with relatively few games receiving very low or very high scores. Overall, this suggests that most board games on BoardGameGeek are rated favorably, with relatively few games receiving extremely low or extremely high ratings. The near-normal distribution indicates that user ratings are fairly consistent, with a slight tendency toward higher scores.

Bivariate EDA



User engagement metrics (NumUserRatings, NumOwned, NumWant, and NumWish) showed weak but positive associations with AvgRating. Log-transformations clarified these trends, revealing that higher ownership and desire modestly correspond to higher average ratings, supporting the idea that popularity and anticipation reflect perceived quality.

Final Decisions of EDA

After completing exploratory analysis (more EDA is available in Appendix), we performed the following data cleaning and transformation steps to prepare the data for modeling:

- Removed games with inconsistent player count data (where MinPlayers > MaxPlayers) or missing values (0 players).
- Limited the dataset to games with fewer than 15 maximum players to focus on standard board games.
- Excluded games with a GameWeight of 0, which likely indicated missing complexity ratings.
- Log-transformed skewed variables (NumUserRatings, NumOwned, NumWant, NumWish, MfgPlaytime, LanguageEase, MfgAgeRec) using log1p to stabilize variance and create linearity.
- Created categorical groupings for:
  - MinPlayers (1-2 players vs. 3+ players)
  - MaxPlayers (1-3 players vs. 4+ players)
  - NumExpansions (0, 1–10, and 10+ expansions)
- Themes and Mechanics were explored but not included in final modeling due to high sparsity and variability across games.

These transformations ensured a clean, interpretable dataset, better suited for linear modeling. We also explore potential correlation between predictors. More in depth information regarding this can be found in Appendix however we deal with relevant correlation in our Methodology section.

Methodology

Initial Model Building (Model 1)

We used multiple linear regression to model average user rating (AvgRating) based on gameplay attributes, user engagement metrics, and structural characteristics of board games. Throughout the modeling process, we systematically improved model fit, reduced multicollinearity, and checked model assumptions to ensure robustness and interpretability.

Table 1: Model 1 Summary: Raw Engagement Metrics

term	estimate	std.error	statistic	p.value
(Intercept)	5.615	0.030	188.027	0.000
log_NumOwned	-0.063	0.011	-5.800	0.000
log_NumWant	0.028	0.008	3.394	0.001
log_NumWish	0.473	0.009	55.349	0.000
GameWeight	0.282	0.008	36.706	0.000
LogPlaytime	0.014	0.005	2.853	0.004
MinPlayers_Group3+	0.032	0.013	2.400	0.016
MaxPlayers_Group4+	-0.171	0.011	-14.884	0.000
log_NumUserRatings	-0.234	0.011	-21.123	0.000

Table 2: Model 1 VIF Values

Variable	VIF
log_NumOwned	11.31
log_NumWant	8.22
log_NumWish	9.66
GameWeight	1.90
LogPlaytime	1.56
MinPlayers_Group3+	1.06
MaxPlayers_Group4+	1.19
log_NumUserRatings	13.04

Model 1 included log-transformed engagement metrics (NumOwned, NumWant, NumWish, NumUserRatings), GameWeight, LogPlaytime, and grouped player count variables. Although $R^2 = 0.519$, high multicollinearity among engagement variables (VIF > 10) prompted us to reduce redundancy before improving explanatory power.

Reducing Multicollinearity and Improving Interpretability (Model 2)

Table 3: Model 2 VIF Values

Variable	VIF
log_DesireScore	3.20
GameWeight	1.82
LogPlaytime	1.54
MaxPlayers_Group4+	1.09
log_NumOwned	2.87

In Model 2 we excluded log_NumUserRatings and replaced log_NumWant and log_NumWish with log_DesireScore (defined as the mean of log-transformed NumWant and NumWish),

retaining log_NumOwned separately to represent actual game ownership. The VIF values dropped significantly across predictors, all falling below 4, suggesting that the multicollinearity problem had been substantially mitigated.

Expanding Predictor Set (Model 3)

After achieving acceptable multicollinearity, we sought to improve predictive power by adding additional meaningful predictors. Based on additional exploratory data analysis and logical reasoning, we added: `LogLanguageEase` (language complexity) and `NumExpansions_Group` (grouped into 0, 1-10, and 10+ expansions).

Model 3 included these additional variables and achieved a higher R^2 of 0.550. Importantly, VIF values remained low, confirming that the model remained stable despite the inclusion of extra predictors. We briefly considered adding MfgAgeRec (minimum recommended age) but found that it did not improve model performance and slightly increased overfitting risk.

Focusing on Popular Games (Model 4 and Final Model)

Recognizing that extremely niche games (with very low ownership) could introduce noise and bias ratings, we restricted our dataset to games with at least 1000 owners. This cutoff ensured that our model predictions would be based on well-established games with a more reliable volume of user ratings.

Model 4 was re-fitted on this filtered dataset. Interestingly, `LogPlaytime` became statistically insignificant ($p = 0.29$), possibly because popular games tend to converge on similar playtime ranges. When we removed LogPlaytime, our final model achieved an adjusted R^2 of 0.657, the highest among all models (equal to model 4), with acceptable multicollinearity (all VIFs < 3).

Table 4: Final Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	5.683	0.076	75.058	0
log_DesireScore	0.948	0.016	57.976	0
GameWeight	0.238	0.009	25.703	0
log_NumOwned	-0.270	0.011	-24.423	0
MaxPlayers_Group4+	-0.204	0.017	-11.761	0
LogLanguageEase	0.013	0.004	3.576	0
NumExpansions_Group1-10	0.086	0.014	6.293	0
NumExpansions_Group10+	0.264	0.025	10.439	0

Table 5: Final Model VIF Values

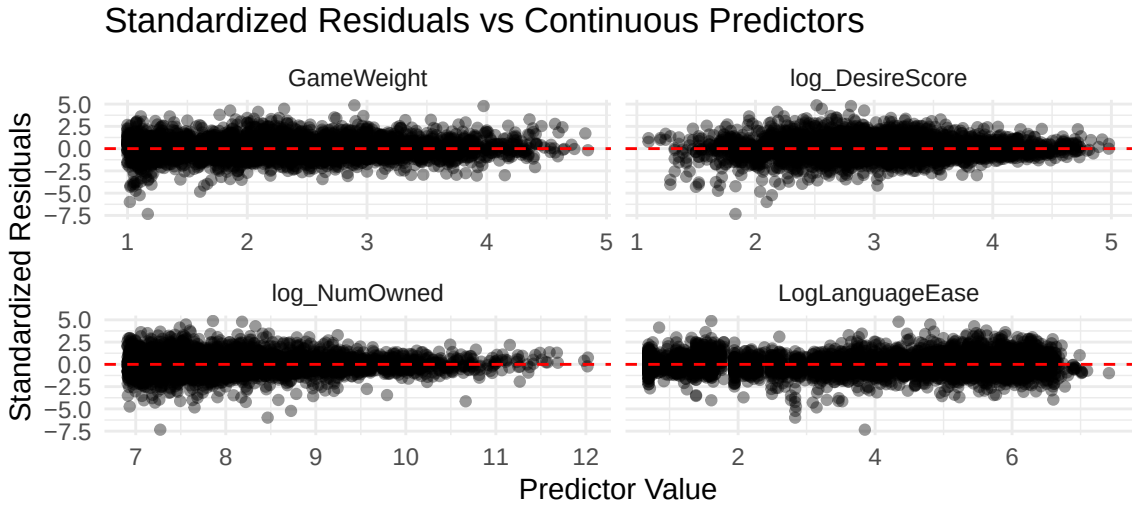
Variable	VIF
log_DesireScore	2.83
GameWeight	1.37
log_NumOwned	2.68
MaxPlayers_Group4+	1.03
LogLanguageEase	1.19
NumExpansions_Group1-10	1.18
NumExpansions_Group10+	1.22

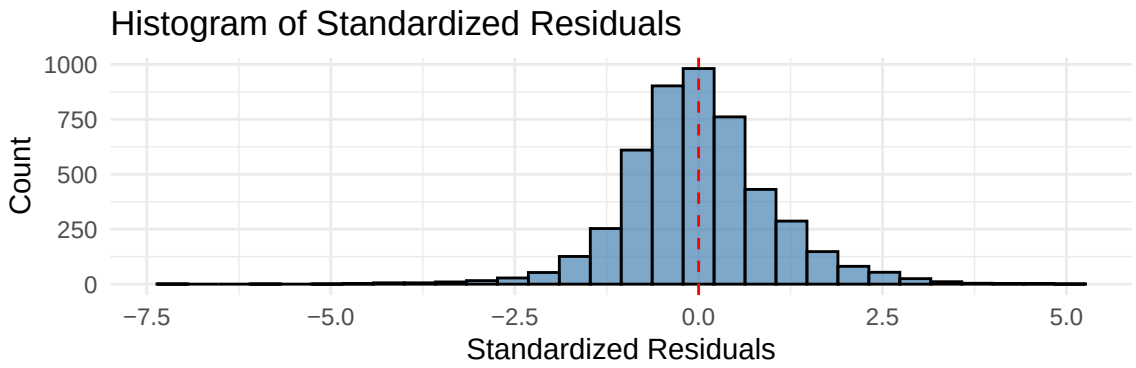
We also considered interaction terms. We saw no real interactions and no improvement in R^2 or $Adj.R^2$ and they were not statistically significant based on 0.05 p-value threshold, therefore we did not include any interaction terms in our final model in order to reduce complexity. We considered interactions between Game complexity and Max players group and Desire Score and Num Expansions Group which are intuitive considerations.

Table 6: Summary of R-Squared and Adjusted R-Squared for Each Model

Model	R_Squared	Adjusted_R_Squared
Model 1	0.518	0.517
Model 2	0.504	0.504
Model 3	0.544	0.544
Model 4: With LogPlaytime	0.657	0.657
Final Model: Without LogPlaytime	0.657	0.657

Checking Model Assumptions





Linearity

The residuals were randomly scattered around zero with no discernible patterns or curves. This suggests that the relationship between each predictor and the response variable (AvgRating) is approximately linear, and no major non-linear effects were left unmodeled.

Independence

Since the data points (individual board games) are inherently independent of one another — each game is a distinct entity with its own ratings and attributes — the assumption of independence is reasonable. Additionally, no time-based or grouped structure (such as repeated measures) exists in the data that would threaten independence.

Constant Variance

The spread of the standardized residuals remained relatively constant across the range of each predictor. There was no evident funneling, fanning out, or other constant variance violation pattern in the residual plots (some slight funneling on NumOwned but given the size of data not concerning).

Normality of Residuals

The histogram of standardized residuals displayed a symmetric, bell-shaped curve centered around zero. Although a perfect normal distribution is not required for regression validity, the residuals' distribution was sufficiently close to normal.

Results

Our final model achieved an R^2 of 0.657 and an adjusted R^2 of 0.657, meaning it explains about 65.7% of the variability in board game average ratings. The most important predictors identified were **log_DesireScore**, **GameWeight**, **NumOwned**, **MaxPlayers_Group**, **LogLanguageEase**, and **NumExpansions_Group**, along with the interaction between log_DesireScore and GameWeight.

Interestingly, we found that a **10% increase** in the number of copies owned (NumOwned) is associated with a slight **0.026-point decrease** in average rating with other variables held constant. This may suggest that more popular games are exposed to a broader audience and thus subject to more critical reviews. User interest (log_DesireScore), reflecting the average of NumWant and NumWish, was associated with an increase in average rating, suggesting that board games who are wanted are most liked.

Holding all variables constant, a **one-unit increase** in GameWeight (strategic complexity) leads to a **0.238-point** rise in average rating, confirming that more complex games tend to be more highly rated.

Regarding player count, holding other variables constant, we found that games that support **4 or more players** have an lower average rating when compared to games supporting **1–3 players**, suggesting that smaller-group games are slightly better received. Ease of language also showed a positive effect: a **10% improvement** in LanguageEase corresponds to an approximate **0.008-point increase** in average rating holding all else constant, indicating that more accessible games may have a modest advantage. Finally, games with expansions tend to perform better as each category with expansion led to a higher average rating. This highlights that ongoing content support (through expansions) strongly correlates with higher user satisfaction.

Discusssion and Conclusion

Our final model provides a strong and interpretable foundation for understanding what makes well-known board games well-received. It highlights that not just popularity, but **desire**, **complexity**, **accessibility**, and **ongoing community engagement** matter most in predicting board game success on BoardGameGeek. Games with **more owners** were surprisingly rated **lower**, likely due to inviting more critical reviews from potentially casual players.

Limitations

BoardGameGeek ratings are **user-submitted** and voluntary, meaning the dataset tends to reflect the views of self-selected players who are often hobbyists or enthusiasts and can introduce **selection bias**. As a result, the ratings may not generalize to casual consumers or the broader population of board game buyers.

By filtering the dataset to only games with greater or equal to **1,000 owners**, we ensured more stable ratings and more well known games, but this also introduces **survivorship bias**. Newer, niche, or indie games are excluded, limiting the model's applicability to predicting ratings for emerging or less-established games.

Although the dataset included detailed **theme and mechanic tags**, the sparsity of these categories, with some categories only having one or two games, led us to exclude them from final modeling. As a result, **genre or mechanic-specific effects** that may exist are **not captured**, such as the potential difference between economy games vs. storytelling games, or dice rolling games vs. card games.

Practical Implications and Future Work

Our findings could have important implications for the board game industry. For **designers**, the results emphasize that strategic complexity and accessibility both positively impact reception. **Publishers** can leverage the strong relationship between community desire and ratings by investing early in community engagement strategies, while also continuing to seek out opportunities to release expansions. **Retailers and distributors** can use ownership metrics and early user desire signals to better forecast which games are likely to become fan-favorites.

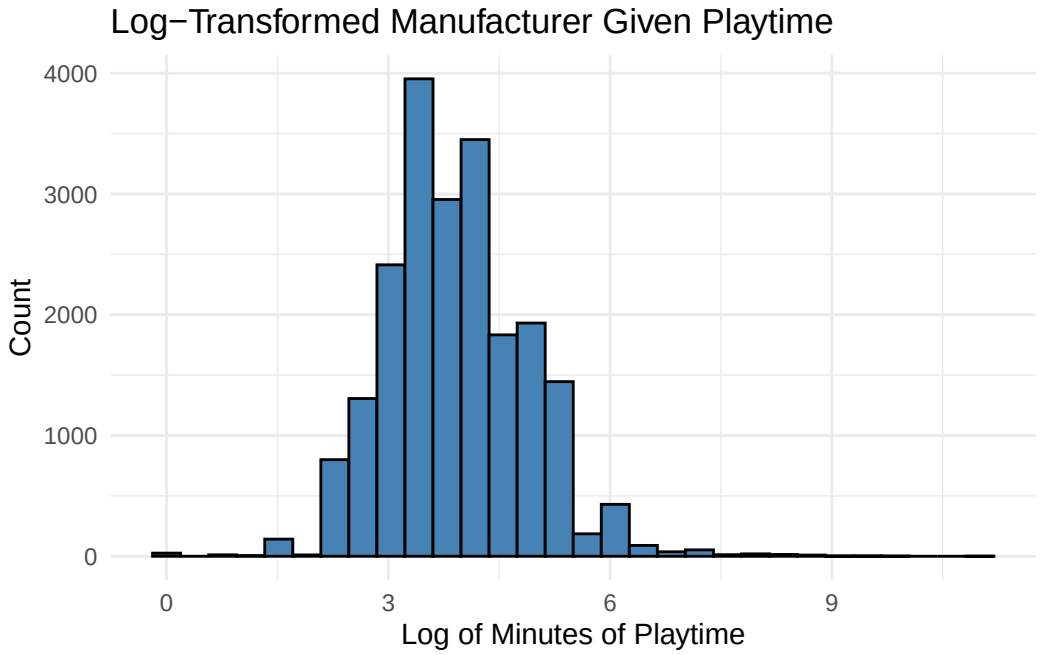
Future studies could explicitly model **rating trajectories over time** by incorporating data on the release year, first rating date, and changes in average rating over time. This would provide a more dynamic view of game reception. Future research could use **multi-platform data** such as ratings from Amazon, Target, or other game review sites to test whether the same predictors hold outside of the highly engaged BGG community. This could provide external validity and increase the population of our consumers and critics.

Appendix

(Additional EDA)

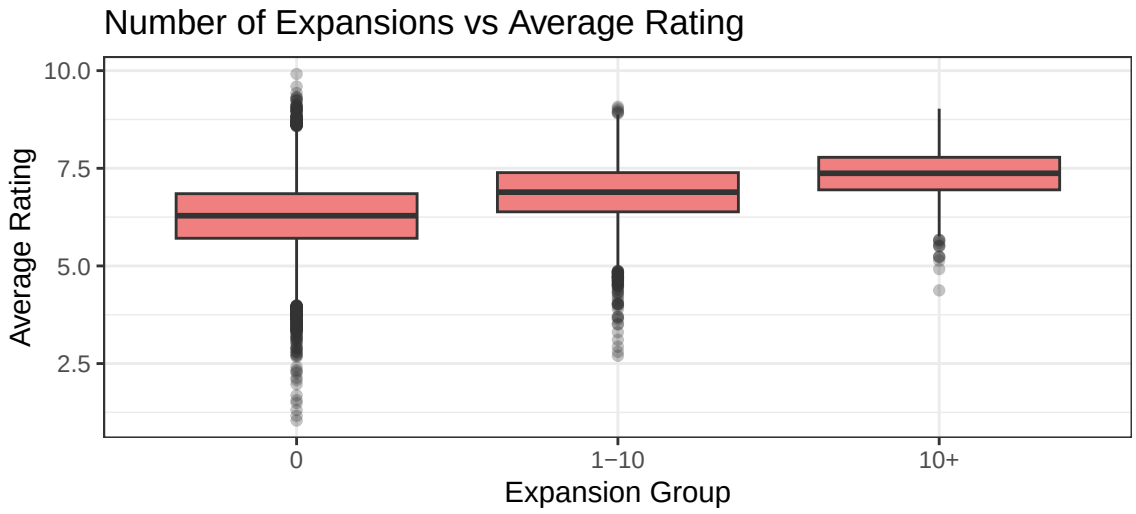
Predictor Univariate EDA before Bivariate EDA

```
# A tibble: 5 x 2
  Name                                                  MfgPlaytime
  <chr>                                                       <dbl>
1 The Campaign for North Africa: The Desert War 1940-43       60000
2 Case Blue                                                   22500
3 1914: Offensive à outrance                                  17280
4 Atlantic Wall: D-Day to Falaise                             14400
5 Empires in Arms                                             12000
```



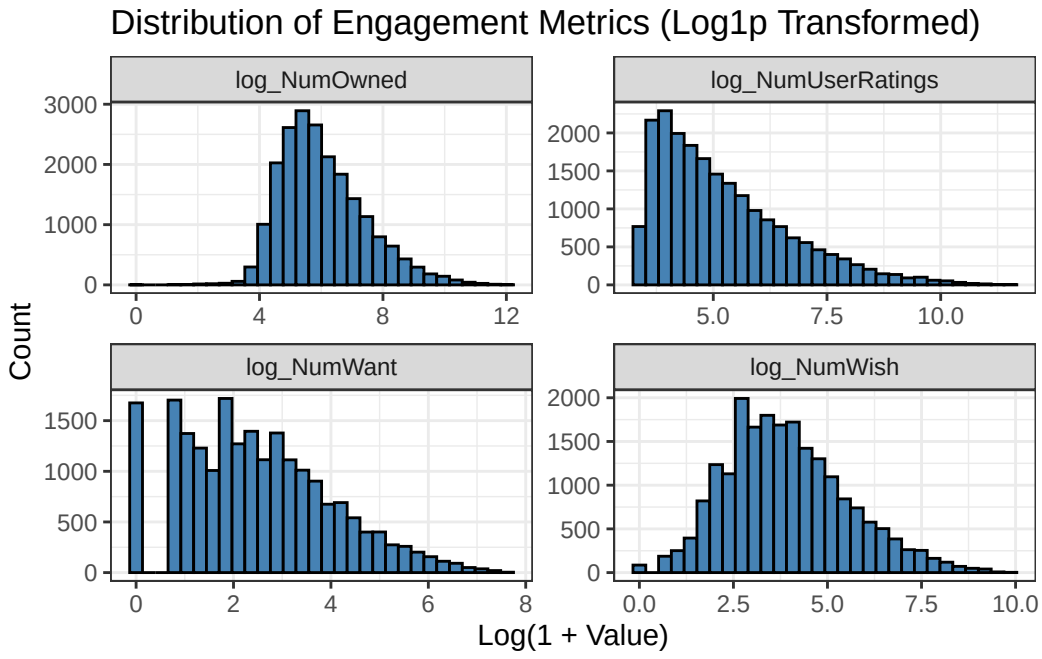
Games like The Campaign for North Africa exhibit extreme playtimes (up to 60,000 minutes), leading to a heavily right-skewed distribution for MfgPlaytime. A natural log transformation (excluding zeros) was applied to normalize this skew. Post-transformation, the central 50% of games fall between log-playtimes of 3.2 to 4.5, corresponding to roughly 25–90 minutes — a reasonable range for typical gameplay durations. Extreme outliers persist, but their influence is mitigated.

Number of Expansions vs. Average Rating

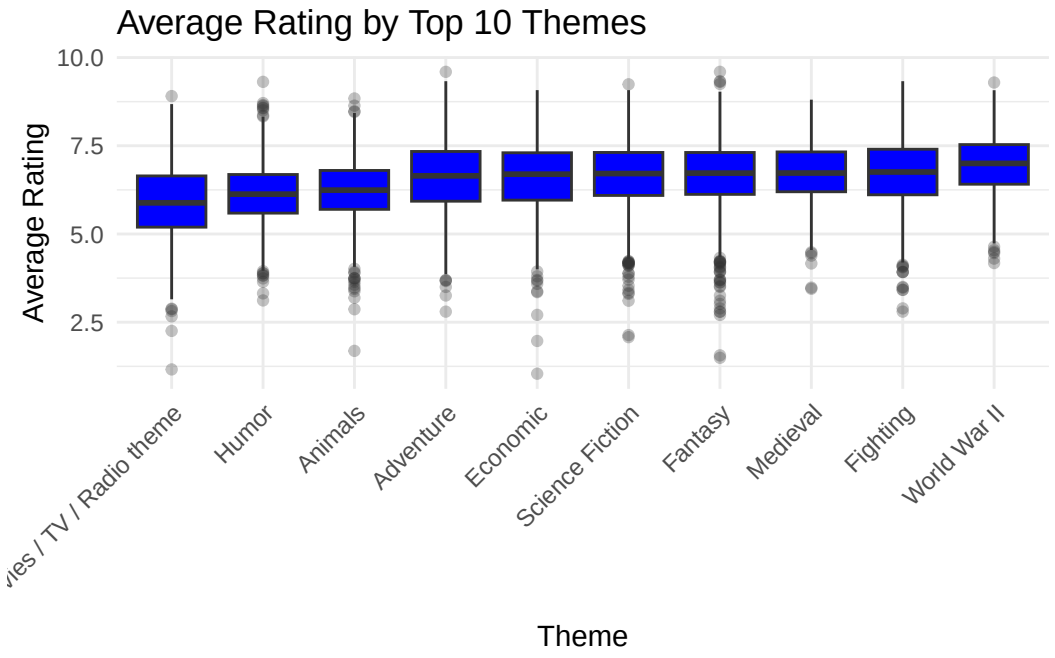


Games with more expansions tend to have higher average ratings, suggesting that expansions are a marker of quality and popularity. While overlap between groups exists, the median rating increases clearly with more expansions.

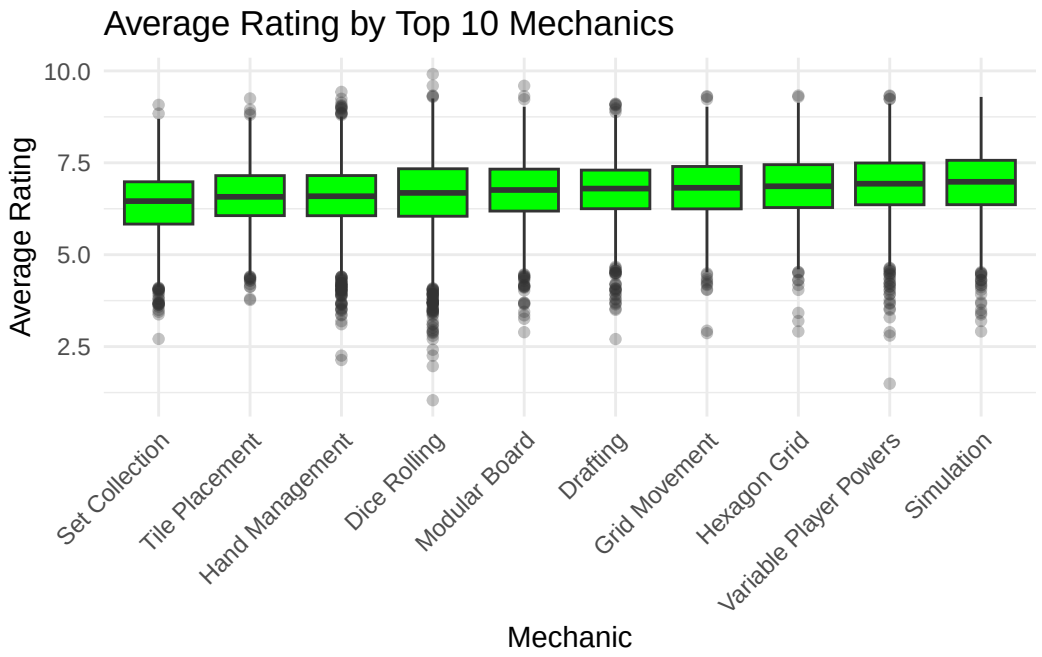
Bivariate EDA



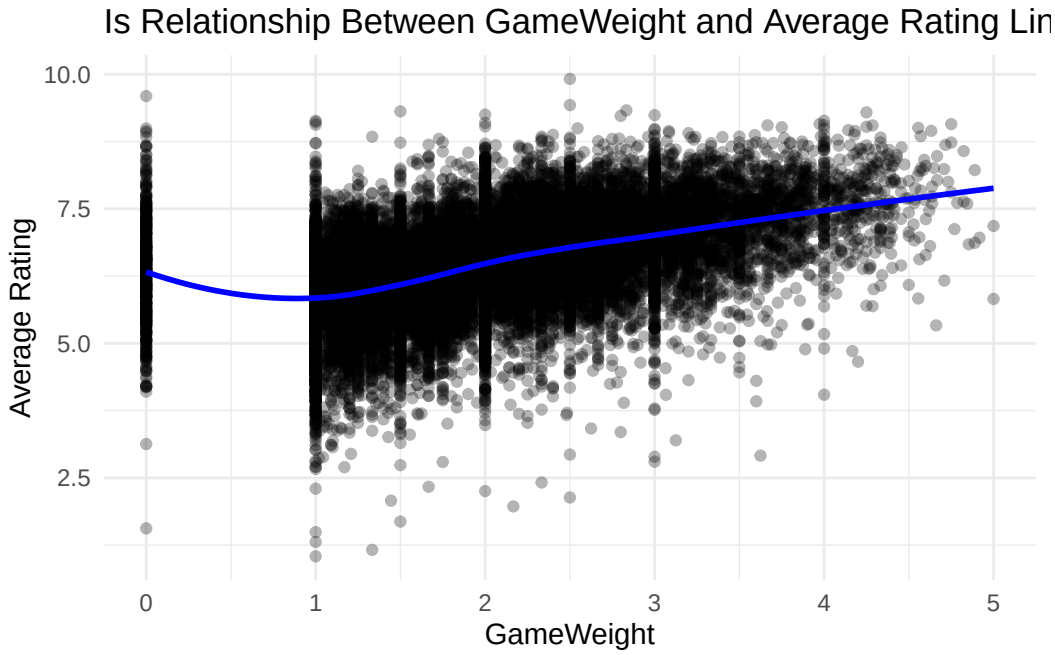
The faceted histograms show the log-transformed distributions of NumUserRatings, NumOwned, NumWant, and NumWish, representing different engagement metrics for board games. NumOwned and NumUserRatings follow a right-skewed but smoother distribution after transformation, making them more suitable for modeling. NumWish appears more balanced, suggesting a consistent spread of user interest. However, NumWant remains irregular, with a large spike near zero, indicating many games have few or no "want" requests. To address this, NumWant could be binarized (HasWant = 1 if NumWant > 0) or grouped into categories to reduce sparsity. We also saw before that we expect multicolinearity so we could also ignore num_want and use a combination of other variables.



Among the top 10 most frequent themes, Economic, Adventure, and Animals show slightly higher median ratings, whereas themes like Humor and Movies/TV lag behind. This variation implies that theme choice may influence perception, but popularity alone does not guarantee higher quality in user ratings.

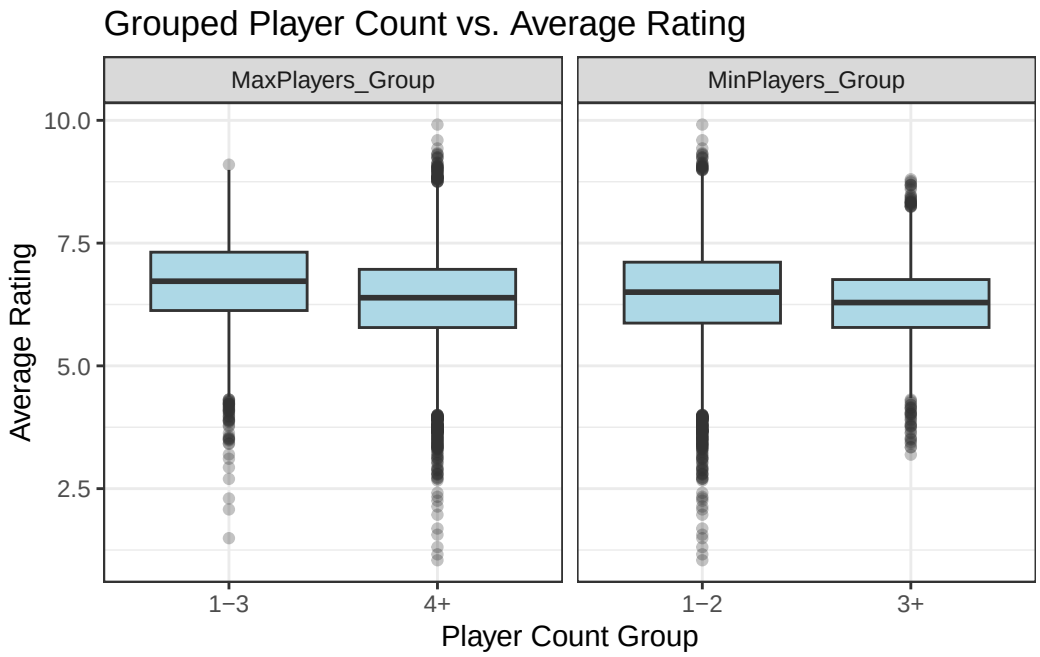


Mechanics such as Simulation, Variable Player Powers, and Hexagon Grid exhibit higher median ratings, potentially reflecting deeper strategic complexity or player agency. Conversely, simpler mechanics like Dice Rolling and Hand Management, while prevalent, are associated with slightly lower average ratings, possibly due to perceived randomness or simplicity.



We can see that there is some data error with game weight, is there is not supposed to be gameweights of 0. In order to stay consistent with our analysis and use gameweight as a useful predictor, we will drop data that has gameweight of 0 since this is not intended.

Player Count vs. Average Rating

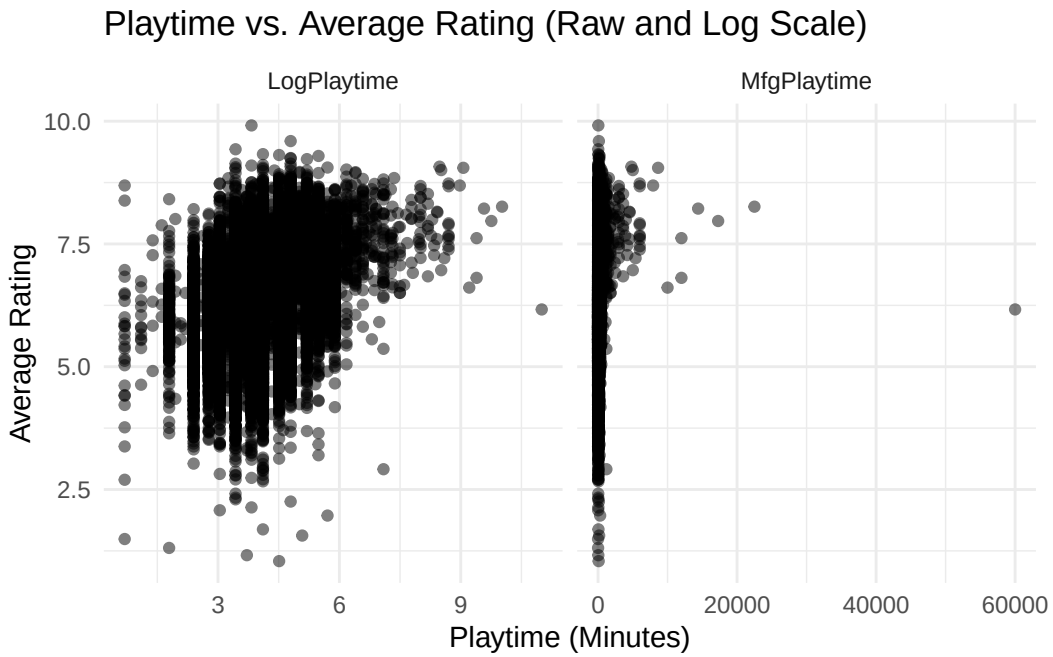


The boxplots above compare the average ratings of board games grouped by their minimum and maximum player count.

Games supporting 1–3 players (MaxPlayers_Group) and games requiring only 1–2 players (MinPlayers_Group) tend to have slightly higher median ratings than those accommodating larger groups. However, the differences are small, and the interquartile ranges overlap substantially.

Thus, while there may be a modest preference for games that support smaller groups, **player count alone does not appear to be a strong driver of average rating**. Other factors such as complexity, engagement, or theme likely have a more significant impact on perceived game quality.

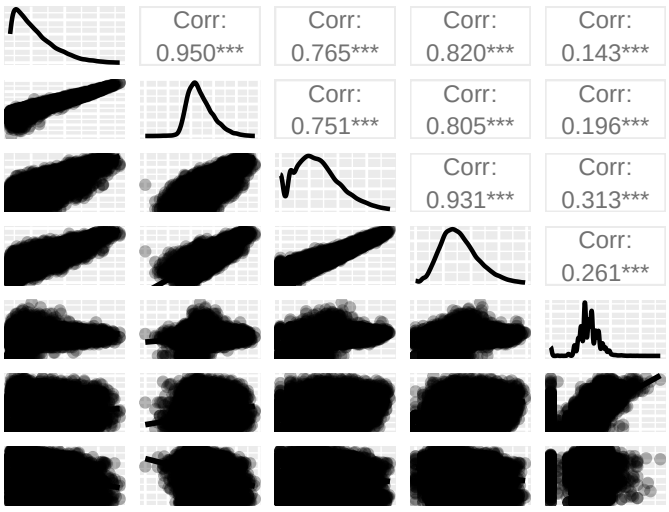
Playtime of a Board Game vs. Average Rating



The plots show the relationship between playtime and average rating, with most games clustered at shorter durations. While no strong linear trend exists, some highly rated games have longer playtimes, suggesting that complexity may enhance reception.

After log transformation, the distribution becomes more balanced, revealing a slight positive trend where games with 30–120 minutes of playtime tend to score higher. However, beyond this, diminishing returns appear, as excessively long games do not necessarily receive better ratings. Playtime alone does not determine success; a potential interactions with mechanics of games, likely play a more significant role in user ratings.

Exploring Correlation and Interaction effects between potential predictors



The pairwise plot of continuous predictors reveals strong positive correlations among the engagement metrics, particularly between log_NumUserRatings and log_NumOwned (r = 0.952), and between log_NumWant and log_NumWish (r = 0.931). This suggests that games with higher user activity in one area (e.g., ownership) also tend to be highly desired or rated. The strong correlations highlight potential multicollinearity concerns if these predictors are included together in a regression model. Moderate positive correlations were also observed between GameWeight and LogPlaytime (r = 0.720), indicating that more complex games generally take longer to play. LogLanguageEase was negatively correlated with engagement metrics, though the relationships were weaker (r -0.2 to -0.3), suggesting that harder-to-understand games may slightly deter engagement. Overall, careful variable selection or the creation of composite metrics is necessary to minimize redundancy and ensure model stability.

Model 3 VIF and terms (after adding two more predictors)

Table 7: Model 3 Summary: More Predictors

term	estimate	std.error	statistic	p.value
(Intercept)	5.512	0.036	154.835	0
log_DesireScore	0.872	0.011	79.906	0
GameWeight	0.269	0.008	32.388	0
LogPlaytime	0.027	0.006	4.800	0
log_NumOwned	-0.256	0.006	-41.547	0

term	estimate	std.error	statistic	p.value
MaxPlayers_Group4+	-0.214	0.012	-18.323	0
LogLanguageEase	0.031	0.003	10.971	0
NumExpansions_Group1-10	0.139	0.012	11.662	0
NumExpansions_Group10+	0.382	0.029	13.286	0

Table 8: Model 3 VIF Values

Variable	VIF
log_DesireScore	3.54
GameWeight	1.94
LogPlaytime	1.61
log_NumOwned	3.38
MaxPlayers_Group4+	1.10
LogLanguageEase	1.07
NumExpansions_Group1-10	1.20
NumExpansions_Group10+	1.12

Interaction Terms in modeling

```
Call:
lm(formula = AvgRating ~ c_log_DesireScore + c_GameWeight + log_NumOwned +
    MaxPlayers_Group + c_log_DesireScore * NumExpansions_Group +
    LogLanguageEase + NumExpansions_Group, data = games_interaction)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1904 -0.2559 -0.0190  0.2306  2.0842

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               8.983709   0.096039  93.542  < 2e-16
c_log_DesireScore                         0.981480   0.018435  53.239  < 2e-16
c_GameWeight                              0.237496   0.009242  25.698  < 2e-16
log_NumOwned                             -0.257174   0.011385 -22.589  < 2e-16
MaxPlayers_Group4+                       -0.195449   0.017342 -11.270  < 2e-16
NumExpansions_Group1-10                   0.079288   0.013758   5.763 8.78e-09
NumExpansions_Group10+                    0.304021   0.027511  11.051  < 2e-16
LogLanguageEase                           0.011682   0.003519   3.320 0.000907
```

```
c_log_DesireScore:NumExpansions_Group1-10 -0.064327   0.022512  -2.857 0.004289
c_log_DesireScore:NumExpansions_Group10+  -0.191761   0.036632  -5.235 1.72e-07

(Intercept)                               ***
c_log_DesireScore                         ***
c_GameWeight                              ***
log_NumOwned                              ***
MaxPlayers_Group4+                        ***
NumExpansions_Group1-10                   ***
NumExpansions_Group10+                    ***
LogLanguageEase                           ***
c_log_DesireScore:NumExpansions_Group1-10 **
c_log_DesireScore:NumExpansions_Group10+  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4313 on 4793 degrees of freedom
  (147 observations deleted due to missingness)
Multiple R-squared:  0.6591,    Adjusted R-squared:  0.6585
F-statistic:  1030 on 9 and 4793 DF,  p-value: < 2.2e-16


                          c_log_DesireScore
                                   3.622699
                               c_GameWeight
                                   1.380358
                               log_NumOwned
                                   2.846109
                          MaxPlayers_Group4+
                                   1.043967
                    NumExpansions_Group1-10
                                   1.196727
                     NumExpansions_Group10+
                                   1.454222
                             LogLanguageEase
                                   1.202783
c_log_DesireScore:NumExpansions_Group1-10
                                   2.229331
 c_log_DesireScore:NumExpansions_Group10+
                                   1.684953


Call:
```

```
lm(formula = AvgRating ~ c_log_DesireScore + c_GameWeight + log_NumOwned +
    MaxPlayers_Group + c_log_DesireScore * MaxPlayers_Group +
    LogLanguageEase + NumExpansions_Group, data = games_interaction)

Residuals:
     Min       1Q   Median       3Q      Max
-3.15243 -0.25347 -0.01976  0.23070  2.10249

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         9.082339   0.094203  96.413  < 2e-16 ***
c_log_DesireScore                   0.960416   0.030125  31.881  < 2e-16 ***
c_GameWeight                        0.238151   0.009290  25.634  < 2e-16 ***
log_NumOwned                       -0.270280   0.011084 -24.385  < 2e-16 ***
MaxPlayers_Group4+                 -0.204398   0.017386 -11.757  < 2e-16 ***
LogLanguageEase                     0.012585   0.003516   3.580 0.000347 ***
NumExpansions_Group1-10             0.086772   0.013749   6.311 3.02e-10 ***
NumExpansions_Group10+              0.263954   0.025266  10.447  < 2e-16 ***
c_log_DesireScore:MaxPlayers_Group4+ -0.014997   0.029983  -0.500 0.616974
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4326 on 4794 degrees of freedom
  (147 observations deleted due to missingness)
Multiple R-squared:  0.6571,    Adjusted R-squared:  0.6565
F-statistic:  1148 on 8 and 4794 DF,  p-value: < 2.2e-16
```

```
                  c_log_DesireScore                         c_GameWeight
                           9.618351                             1.387037
                      log_NumOwned                   MaxPlayers_Group4+
                           2.682129                             1.043280
                     LogLanguageEase               NumExpansions_Group1-10
                           1.193987                             1.188379
               NumExpansions_Group10+ c_log_DesireScore:MaxPlayers_Group4+
                           1.219620                             8.342917
```

As we can see no significant improvement was made in explaining variability of board game average rating with these interaction terms, therefore we did not include them in our model.