# Analyzing Features of Board Games from BGG Dataset

Computer Science 466: Knowledge Discovery From Data

Professor Alexander Dekhtyar

March 20, 2025

Content created, written, and displayed by:

Lucas Summers (lsumme01@calpoly.edu)
Xiuyuan Qiu (xiqiu@calpoly.edu)
Braeden Alonge (balonge@calpoly.edu)
Nathan Lim (nlim10@calpoly.edu)

## 1. Abstract

This study investigates the Board Games Dataset on Kaggle, which contains data from BoardGameGeek, covering over 20k board games and ratings from more than 400k users. We first apply collaborative filtering techniques on the user ratings present in the dataset to predict individual user ratings for board games they have not yet rated based on rating data from similar users. We also apply clustering on the attributes to see if we can discover any insights from meaningful clusters of the various games. Finally, we also examine the impact of game attributes on average user ratings by implementing both linear regression and random forest regressor models. Using these three main analytical methods, this study provides a comprehensive overview of the data and seeks to explore the relationships between board game attributes and user preferences to benefit game designers and players alike.

## 2. Introduction

Our main goals in our analysis were to understand what types of attributes tend to lead to the most successful games, if there were hidden patterns in user ratings we could uncover through collaborative filtering, and if we can successfully build a model that can predict the success of a game (measured by rating) based on its attributes. We hoped that along the way, we would also gain insights into the relationships between the data. The overall approach we took was a divide-and-conquer approach, where we used the multiple techniques that we learned in the course to explore different aspects of the dataset. We focused on three approaches: collaborative filtering, clustering, and regression.

## 3. Dataset

The dataset was found on kaggle.com, and it contains data sourced from BoardGameGeek, an online forum for board gaming hobbyists. The feature-rich data encompasses basic metadata about 22k board games like rating, publish year, number of comments, and more. It also includes more details in other tables like artists, designers, publishers, themes, subcategories, and mechanics of the games. The largest file in the dataset is a 400MB CSV containing ~19 million user ratings. All data is stored as CSV files, and the dataset is 673MB in total.

## 4. Research Questions

From the dataset above, we developed the following research questions:

1. Which attributes of a game tend to produce the highest user ratings?
2. Can we cluster board games into meaningful categories based on mechanics, themes, and categories?

3. How effectively can we predict a “value” metric (e.g., average rating) for a board game from its attributes?
4. Given a user's past game ratings, can we predict their rating of new games based on ratings of similar users?

# 5. Methods

## 5.1 Collaborative Filtering

Our collaborative filtering methods mainly involved the use of the user_ratings data. The user_ratings.csv data file includes each individual rating as a row, with the username, the game being rated, and the rating value itself as features. To perform collaborative filtering on this data, we needed to convert this table into a format where each row represents a unique user, and each column of the table represents a game. Then, the item in each row represents that specific user’s rating for that specific game. We combined this data with the games' data only to get the actual name of the game (as the user_ratings data only included the board game ID).

### 5.1.1 Reducing Dataset Size

The dataset has a list of 18,942,215 user ratings, with 411,375 users and 21,925 games. Building a matrix of this data capable of performing collaborative filtering (where each row is a user, each column is a game, and each item is that user’s rating of the game) would require a size of 41, 1375 x 21,925, or a size of 9,019,396,875 cells total. This massive matrix would also be very sparse, as there are a very small fraction of users that have rated more than 100 games.

To combat the heavy memory requirements of working with the entire dataset and to simplify our methods, we decided to heavily filter these ratings before applying our collaborative filtering algorithm. In order to keep our matrix within a reasonable size and ensure the matrix was as non-sparse as possible, we filtered our data to only include the top 1% of popular games. In this context, we measured popularity by calculating the amount of ratings the game had—the higher the number of user ratings, the higher the popularity.

After removing duplicates, this brought our matrix down to a size of 332506 x 219, which is much more manageable on its own but still very sparse. In order to bring the size down even further and limit the issues of sparsity, we filtered out all users who didn’t have at least 50 of those top 1% popularity games rated. This brought our matrix size down to 35096 x 219. At 7,686,024 cells total (2,836,563 populated ratings), this was a matrix that we were happy with.

### 5.1.2 Filtering

For the actual collaborative filtering algorithm, we decided to use a slightly modified version of our Lab 5 implementation with a K-nearest neighbors approach using adjusted weighted sum and cosine similarity.

To measure the success of our model, we looked at Mean Absolute Error (MAE) with randomly selected sampling and tests in the same manner that we implemented in lab 5. Specifically, we performed 10 samples with 100 randomly selected tests per sample, where for each test, a randomly selected user-item pair was selected from our data matrix to perform a test on. For each test, that user's rating of the item to be predicted was removed, and a prediction was made using our collaborative filtering model. Once a prediction was made, we examined the error between the prediction and the actual user rating.

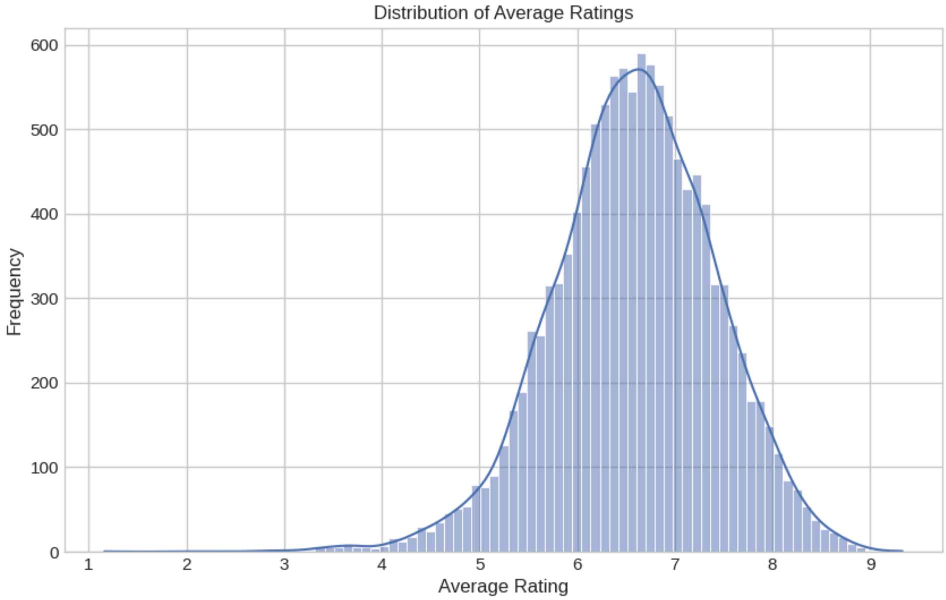
### 5.1.3 Applet

In addition, we developed a command-line applet to use our filtering algorithm to get suggestions for new board games. The applet gives the user a link to 10 games randomly selected from the 100 most popular games in the dataset, and then runs CF to predict the 3 games it thinks the user will rate the highest.

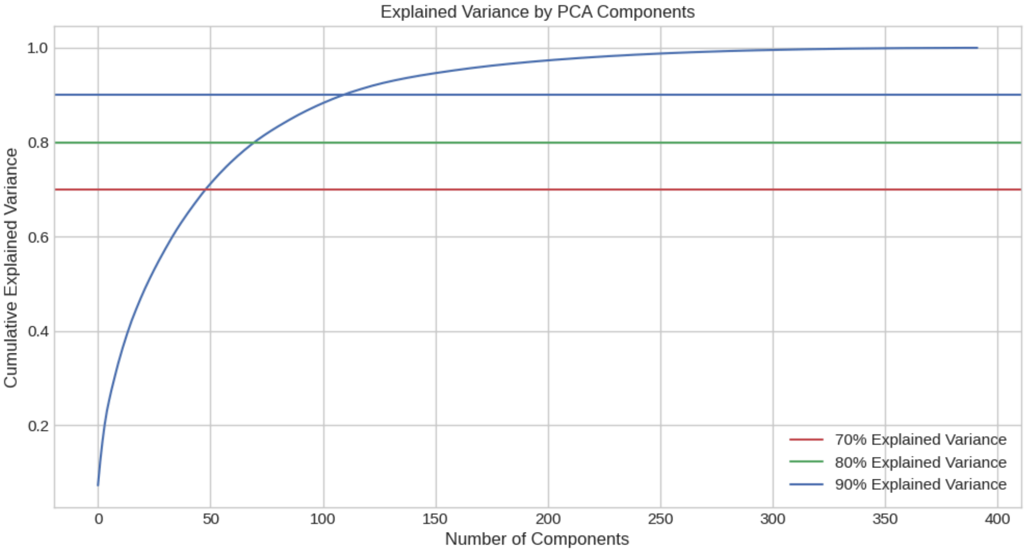
Quantitative and qualitative evaluation of our model is in section 6.1 below.

## 5.2 Clustering

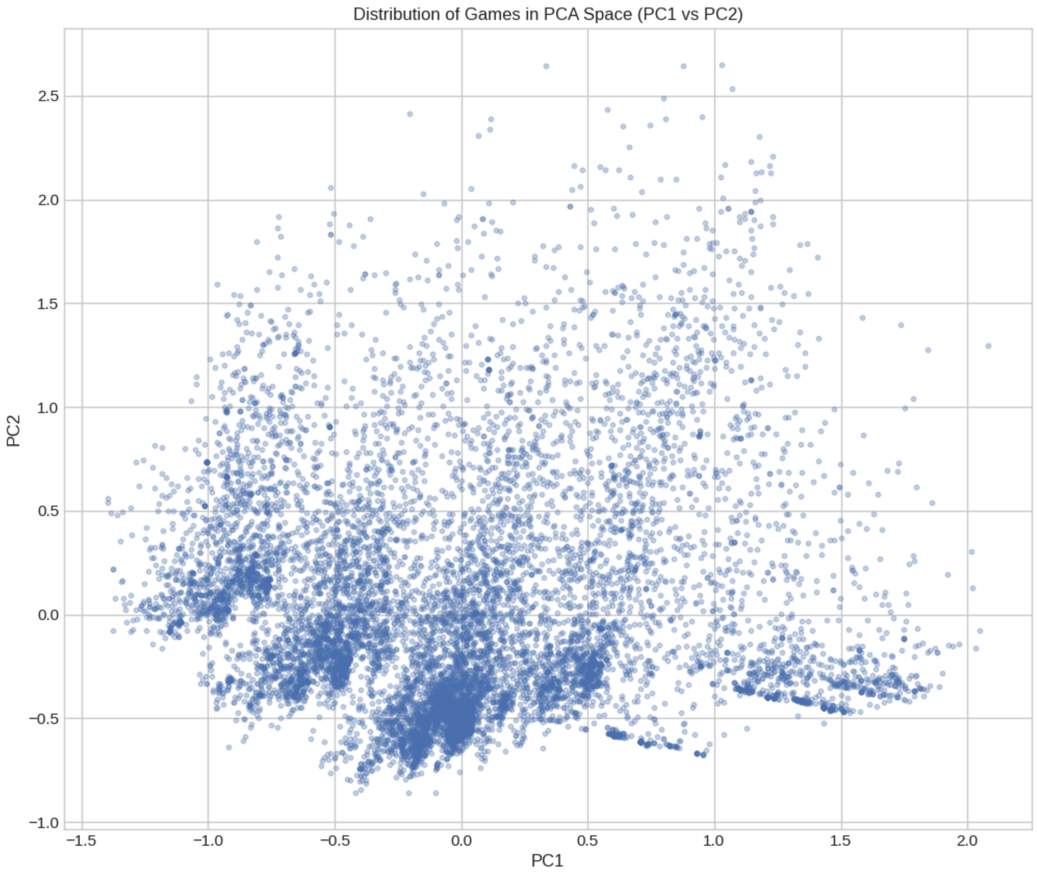
For clustering, we used the main **games** table and enhanced it with more features by merging it with the **mechanics**, **themes**, and **subcategories** tables. This gave us all the games along with all possible attributes that the game could have. Note that the three added tables only contain binary attributes, meaning that the possible values for the attribute are either 0 or 1 (1 the game has the feature, 0 it doesn't). We then decided to filter to board games with at least 100 user ratings in order to remove any possible outliers or niche games, which reduced the rows in the dataset from 21925 to 12239 games.



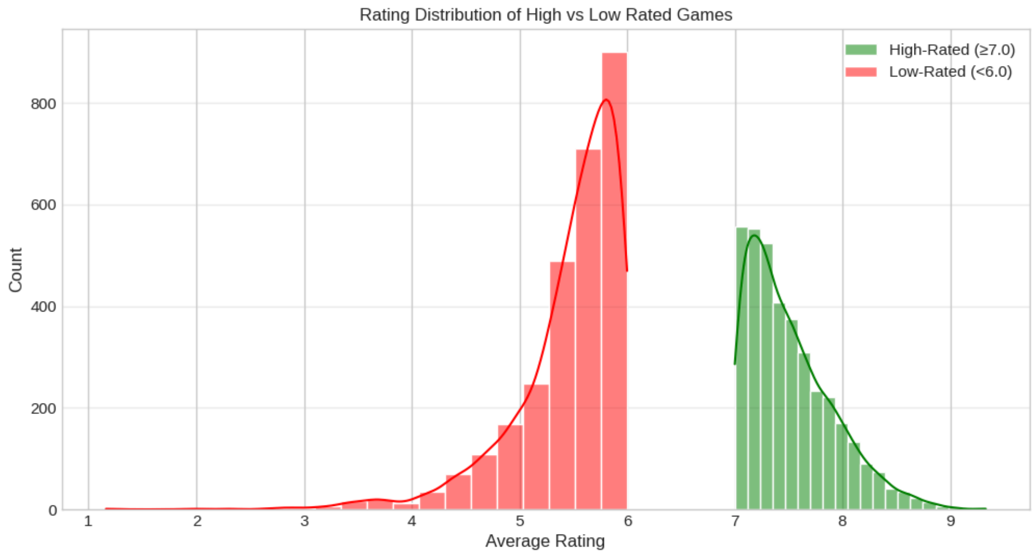
We can see the distribution of averaging ratings follows a normal distribution, centered around a mean of 6.58 (note that ratings are on a scale of 1-10). Then we decided to just focus on binary features (which are most of the features) to keep the data on the same scale and thus make PCA reduction and clustering more effective, which lead to a total of 392 features.



We can see from the graph above, 80% of the variance is explained by 71 components and 90% of the variance is explained by 110 components. Because of diminishing returns, we decided to apply PCA with 71 components to reach 80% variance.



Plotting PC1 vs PC2, we can already see some clusters naturally forming in the space (note this is only the first two components), but there is still a lot of overlap and noise. Because of this and initial poor clustering results on the whole dataset, we ultimately decided to split the dataset in half (around the mean) to try and remove "average" games. Our thinking is that points around the mean incorporate many "good" and "bad" attributes together (leading to an average rating), and thus are responsible for a lot of the overlap/noise that occurs. Clustering "good" games and "bad" games separately can also help us better reveal specific attribute combinations that make a board game high/low rated.



First, we started with only taking games with an average rating greater than 7, reducing it to 3771 games (30.8% of the original dataset). Next, we took the games with an average rating less than 6, reducing the number of points to 2791 games (22.8% of the original dataset). Thus, we further removed 5677 games that fell between an average rating of 6 and 7, giving us games with overall higher ratings (mean of 7.51) and lower ratings (mean of 5.46). Now we can start to cluster each of these data sets separately.

We chose to use agglomerative hierarchical clustering as our main model. From the exploration of the data, we thought kmeans would struggle with non-spherical clusters and not be able to deal with all the many overlaps of games, as well as dbscan would have difficulty with varying density clusters, and it could be hard to determine the appropriate hyperparameters to get clusters to form with minimal noise points. Thus, agglomerative clustering provided the best model for us in many ways. First, board games often have hierarchical structures that agglomerative could naturally reveal. Second, agglomerative works better, with irregular shaped clusters, which can already be seen in the data. Last, agglomerative clustering traditionally works better with datasets that aren't extremely large, and each of the data frames are only a few thousand games.

For grid search, we decided to use predefined cluster sizes instead of a size threshold, using an n_clusters range of 5-15. We wanted to prevent lots of small clusters in order to make it easier to inspect and compare the clusters later. We also experimented with different linkage options (ward, single, complete, average) as well as distance/similarity metrics (Euclidean, Manhattan, cosine) to make sure we were getting the best results. While we recorded Silhouette score, Calinski-Harabasz score, and

Davies-Bouldin score, we ultimately decided to optimize our search with Calinski-Harabasz score, as it provided the best results in terms of finding meaningful clusters. This is because its approach handles feature-rich data well, is less sensitive to non-spherical cluster shapes and varying cluster densities, and overall works better with hierarchical clustering from our experience.

We also chose to do PCA reduction again, but separately on each of the two data sets to reduce dimensionality in our data and achieve faster grid search runtimes. Ultimately, we settled on either using 50 components or the number of components needed to explain 80% of the variance, which ever provided fewer components.

Lastly, inspecting the makeup of clusters was achieved in two ways. First, we simply examine feature prevalences in each cluster for each feature type (mechanics, themes, categories, subcategories). Next, we also trained a small Decision Tree Classifier (max depth of 3)  to distinguish the particular cluster from the others. On top of visualizing the tree, we also used sklearn's functionality to extract the feature importances from the trained model to get the top features that the model uses to distinguish the cluster from the others.

## 5.3 Linear Regression and Random Forest

The objective of these methods is to generate a model that can predict well a game will be rated given numeric attributes assigned to each game from the BGG dataset. Attributes that were not numeric were filtered out. Only games with 100 or more ratings were selected for the models. Doing so reduced the number of games to examine to 12239 games. This reduction ensured that our analysis focused on games with reliable user feedback.

Linear regression was first implemented. This assumes a direct, linear relationship between input features and the target variable. Its simplicity allows for easy interpretation of model coefficients and therefore highlighting individual contributions of each feature. However, its simplicity also limits its ability to interpret more complex results and input.

To account for potential non-linear relationships, we employed a Random-Forest Regressor. Predictions were aggregated from multiple trees in an ensemble of decision trees to improve accuracy and reduce overfitting. In recognition of potential intricacies of game attributes, we anticipated that Random Forest would perform better. We evaluated the models using standard metrics including Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination ($R^2$), which indicates how well a model performed.

To optimize performance, we conducted a grid search algorithm (using sklearn) over a set of parameters, including:

1. n_estimators: The number of trees in the forest
2. max_depth: The maximum depth of the forest, and
3. min_samples_split: The minimum number of samples required to split an internal node.

Using 5-fold cross-validation, we evaluated 36 different parameter combinations (totaling 180 fits). The best of these results allowed us to maximize the $R^2$ score.

# 6. Results

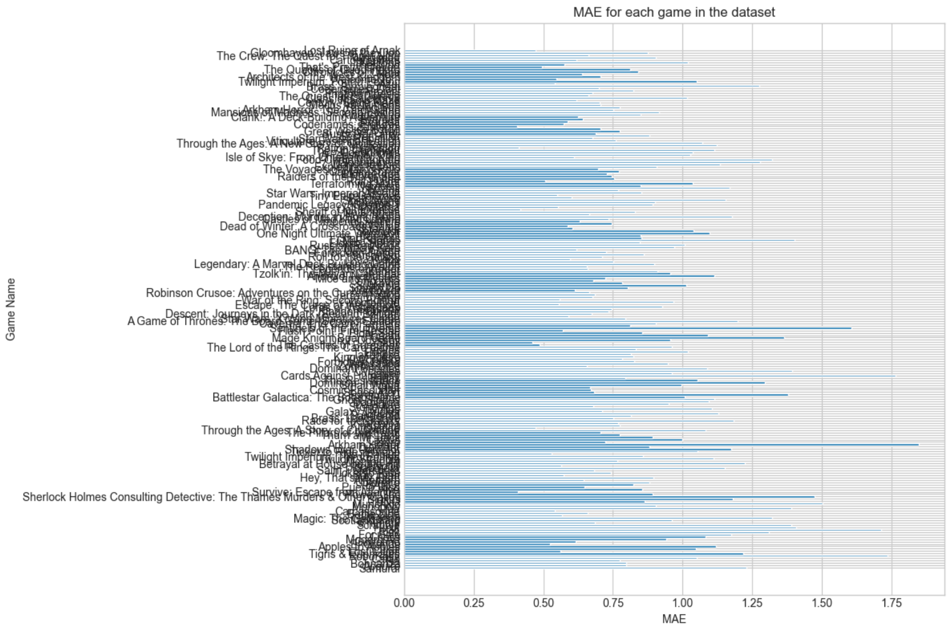
## 6.1 Collaborative Filtering

### 6.1.1 Collaborative Filtering Results

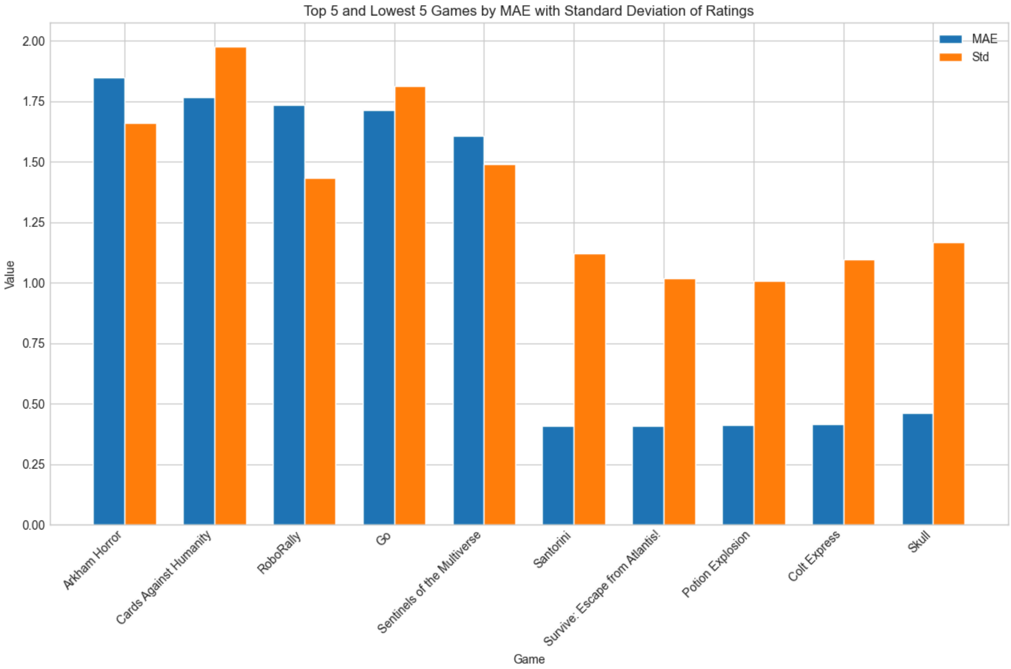
Our collaborative filtering model seemed to be very successful. With our testing of 10 samples of 100 randomly selected tests each, our model achieved an average MAE of 0.91952735 with a standard deviation of 0.079634.

When we performed evaluation on a per-game basis to see how our model differed in performance with different games, we got a range of around ~1.5 MAE (~0.4 to ~1.9 MAE). Due to time constraints and performance issues, we ran 10 tests for each of the 211 games of our filtered dataset. The tests were from randomly selected users.

The graph below shows the various MAE's for each game—the Y axis is cluttered, but the important information is not the name of the game but rather the various spreads in MAE's themselves, which are clearly visible.

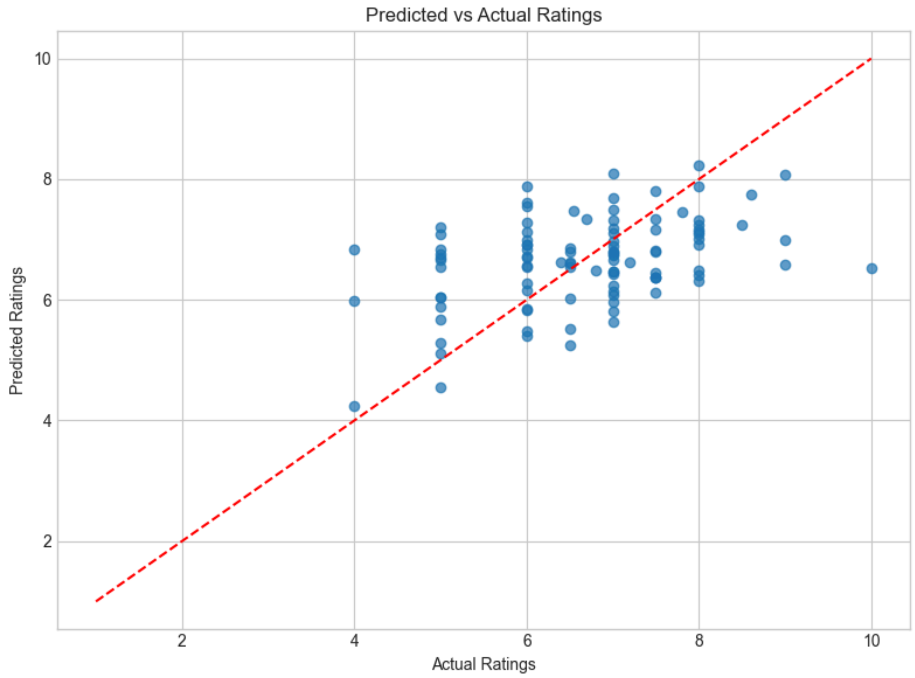


With this data, we wanted to give closer examination to the performance on a per-game basis. We decided to take a look into how the MAE of our collaborative filtering algorithm for a specific game was correlated with the standard deviation of ratings for that specific game, if at all. We took the top 5 and top 5 lowest games by MAE in our testing and graphed them next to the game's standard deviation of ratings. Here were the results:

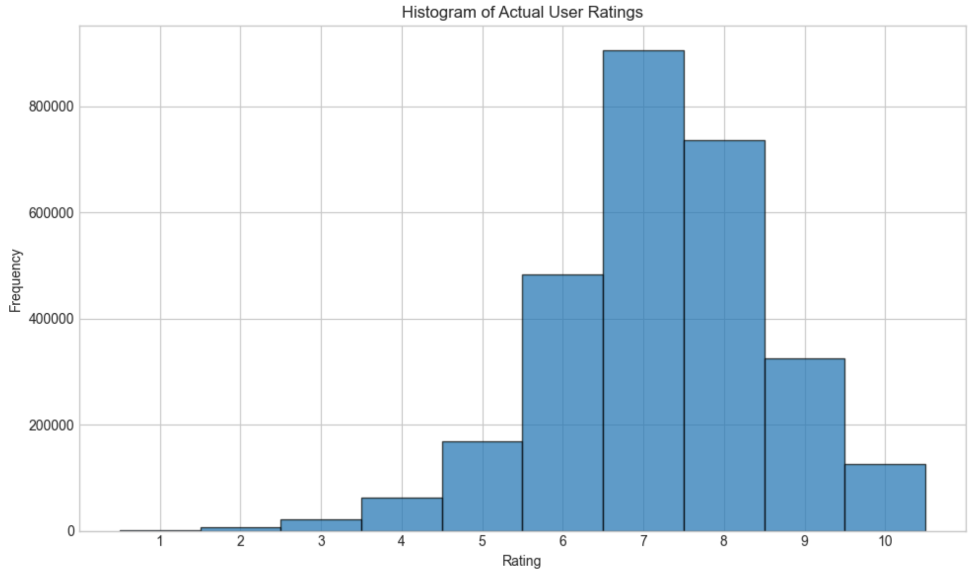


While the difference is small, there is a very noticeable drop in standard deviation of ratings between the games with the highest and lowest performance in terms of collaborative filtering rating prediction. This goes to show that variability in user perception of specific games directly correlates with model performance. While this makes sense intuitively, it is interesting because our model makes its predictions based on the K most similar users; even amongst these groups of similar users, these results show that variability in user perception causes disagreement or unexpected behavior (in the model's eyes, at least) for the end user ratings.

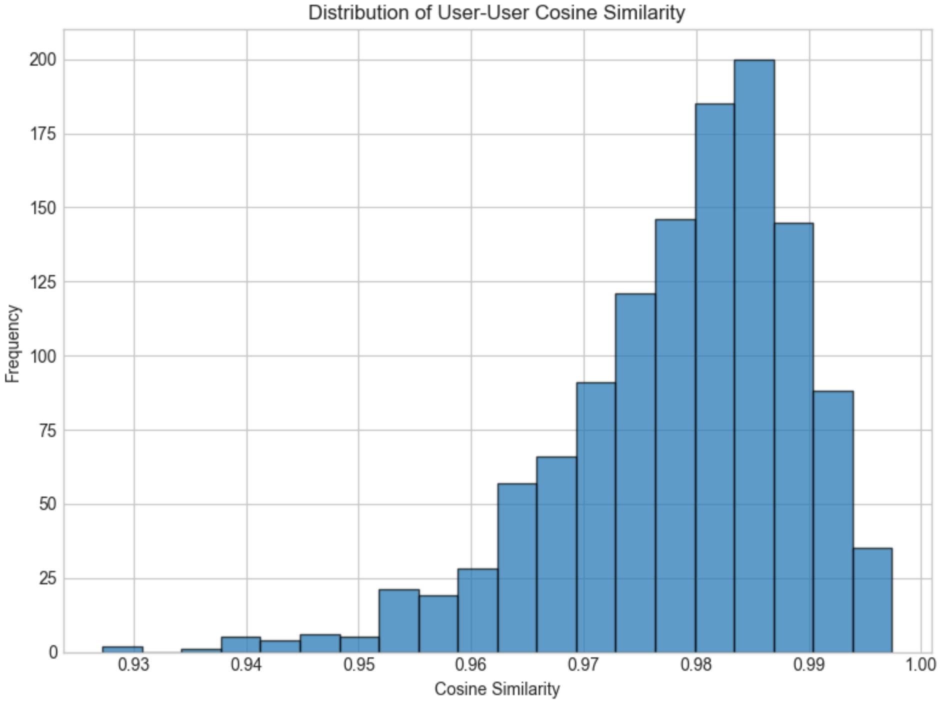
### 6.1.2 Additional Analysis of Collaborative Filtering Results



When we examine this graph of actual user rating in the filtered dataset versus our model's predicted ratings from a series of random tests, showing a clear concentration of 7/10 rating predictions. We were curious about why the model was concentrating these guesses, and although we already suspected that it was because of a very high concentration of user ratings at 7/10, we wanted to take a look at the actual data.



Taking a look at a histogram of the actual user rating distribution shows what we expected. This is the data for our small subset filtered data, but it is evident that users tend to pool their ratings at 7/10 (the perceived "average" mark) despite the rating system being from 0-10.



Taking a look at the distribution of user-user cosine similarity in the filtered dataset, it becomes evident that many users seem to stick to the same rating pattern (pooling their ratings around an average of 7/10). This same rating pattern is what leads to such a high average similarity between users.

### 6.1.3 Baseline comparison

We run a comparison between our approach and a baseline approach of guessing the average (~7.1) score for every game and every user. This was evaluated on the smaller user set of 2,836,563 ratings.

| Method | MAE ↓ | RMSE ↓ |
|---|---|---|
| Baseline (Average Scoring) | 1.069 | 1.54 |
| Collaborative Filtering | 0.879 | 1.15 |

*Collaborative filtering performs better than the baseline*

We get ~20% lower MAE and ~30% lower RMSE from our CF approach than the baseline!

### 6.1.4 Applet Qualitative Evaluation

As described above, the applet asks the user to rate 10 games and suggests 3 new games it thinks the user will like the most. To evaluate this, we rated games from the perspective of 4 different (fake) users. We're reporting the first result where the 10 rated games contain more than one game that the user likes.

| The user likes | # of suggested games that fit (out of 3) |
|---|---|
| minifigures | 2 (War of the Ring: Second Ed., Twilight Imperium: Fourth Ed.) |
| card games | 1 (Food Chain Magnate) |
| social deduction games | 1 (Skull) |
| history | 0 |

For use as a board game recommender, CF is hit-and-miss, sometimes suggesting really cool games for users, and sometimes failing on all three counts.

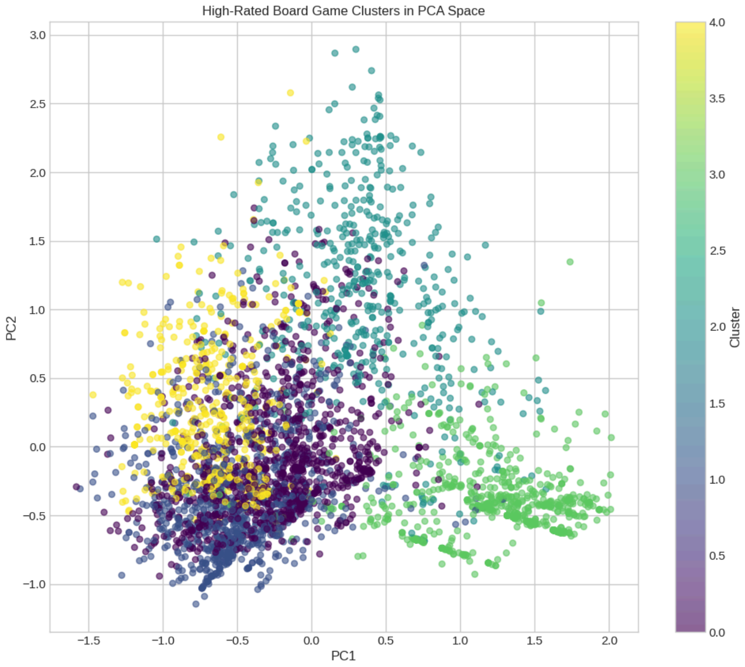
### 6.1.5 Key Insights from Collaborative Filtering

Collaborative filtering seemed to perform very effectively to predict board game ratings, especially when performing with our filtered dataset. The dominant rating pattern across users, with a very significant pooling around the 7/10 mark, significantly influenced the model's predictions and performance. Predictions seemed to default around ratings near 7, and even though the model seemed to have a high overall accuracy with pretty low MAE, it looked like the model was barely adjusting away from that average 7/10 mark. High similarity scores among the users in this filtered dataset further demonstrate that many users tend to stick to the mindset of "7 is the average" scoring system. It also appeared evident that variability in individual user perceptions of specific games (which we observed through the standard deviation of a game's ratings) directly correlated with fluctuations in model accuracy. This is what we expected, but it reflects the fact that more divisive games created a bigger challenge for predicting ratings even when comparing among similar users (as our algorithm did).

## 6.2 Clustering Results

### 6.2.1 Clustering High Rated Games

Best Agglomerative parameters: n_clusters=5, linkage=average, metric=cosine
Silhouette: 0.0541, CH: 222.5, DB: 3.0067



Plotting PC1 vs PC2, we can still see a lot of overlap (as expected) but clusters seem to overall occupy different spaces with some separability. We can clearly see Cluster 3 forming a distinct group on the right side, Cluster 2 appearing mostly in the upper right region, as well as Clusters 0, 1, and 4 showing some overlap in the central and left regions\

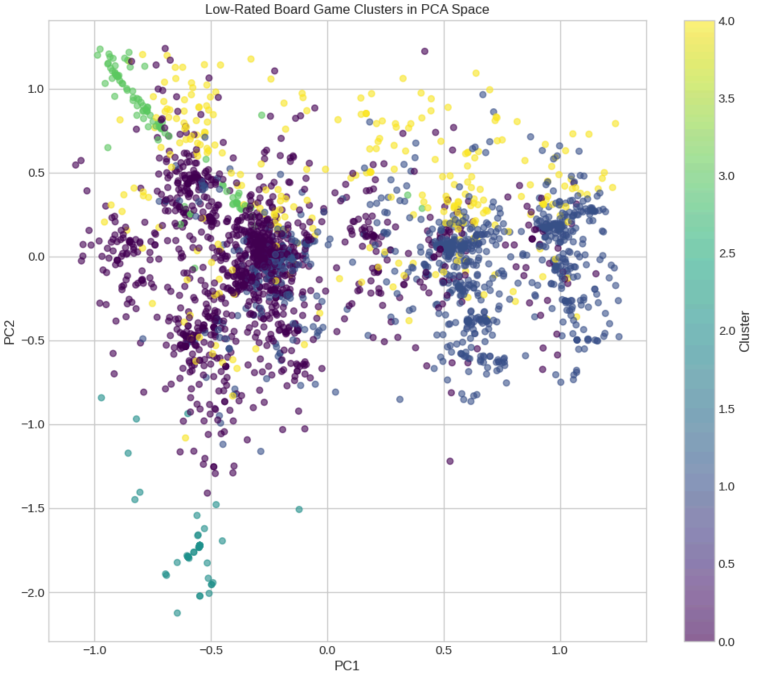
**Cluster Details:**

| Cluster | Description | % of high-rated games | Avg. Rating | Mechanics | Themes | Categories/ Subcategories | Top Decision Tree Features | Example Game |
|---|---|---|---|---|---|---|---|---|
| 0 | Card Games | 21.1% | 7.43 | Hand management (61.7%), drafting (31.6%), deck building (28.2%) | Fantasy (35.4%), fighting (20.2%), Science fiction (8.9%) | Card games (78.3%), strategy (22%) | Card game (0.839), war category (0.042), deck building (0.042) | Aeon's End: Outcasts (8.76) |
| 1 | Classic Strategy Games | 31.4% | 7.49 | Area majority/influence (31%), tile placement (29.1%), worker placement (26.5%) | Economic (33.3%), city building (15.5%), fantasy (13.4%) | Strategy (65.5%), Territory building (11.8%) | Strategy category (0.561), tile placement (0.154), card game (0.107) | Anachrony: Infinity Box (9.08) |
| 2 | Dice-Based Adventure Games | 14.5% | 7.64 | Dice rolling (78%), variable player powers (68%), cooperative gameplay (49%) | Fantasy (43.9%), adventure (41.1%), fighting (39.3%) | Thematic (51.6%), Miniatures (58.3%), Exploration (24.6%) | Miniatures (0.593), adventure theme (0.226), cooperative game (0.081) | The Fantasy Trip: Legacy Edition (9.33) |
| 3 | War Simulation Games | 20% | 7.63 | Dice rolling (76.2%), simulation (63.3%), hexagon grid (58%) | World War II (40.5%), modern warfare (9.6%), aviation/flight  (9.2%) | War games (93.9%), thematic (5.2%) | War category (0.812), miniatures (0.093), strategy category (0.033) | Wings of the Motherland (9.29) |
| 4 | Party & Deduction Games | 13% | 7.39 | Deduction (28.9%), cooperative gameplay (27%), dice rolling (26.8%) | Science fiction (13.4%), murder/mystery (9.1%), humor (7.3%) | Family (15.9%), party games (14.4%), card games (13%) | Deduction (0.493), dexterity (0.197), abstract category (0.129) | Star Trek: Alliance – Dominion War Campaign (8.95) |

Overall, the diversity of features among the clusters demonstrates that high-rated games don't follow any particular pattern, but that there are multiple features, that when combined in the right way, can lead to higher ratings. Cluster 0 shows card games are successful when paired with strategic elements like deck building and hand management. Cluster 1 is strongly strategy based, more complex economic and city building themes with complex mechanics like tile placement. Cluster 2 shows that lots of dice usage is still successful when balanced with other heavily strategic elements like variable player powers and cooperation, which still allows for player agency. Cluster 3 demonstrates the successful integration of themes (WWII) with mechanics (hexagon grid, simulation, etc.) to create good player experiences. Cluster 4 shows successful integration of party and family games with mechanics that promote social interaction along with fun themes.

### 6.2.2 Clustering Low Rated Games

Best Agglomerative parameters: n_clusters=5, linkage=ward, metric=euclidean
Silhouette: 0.0586, CH: 96.8, DB: 3.0570



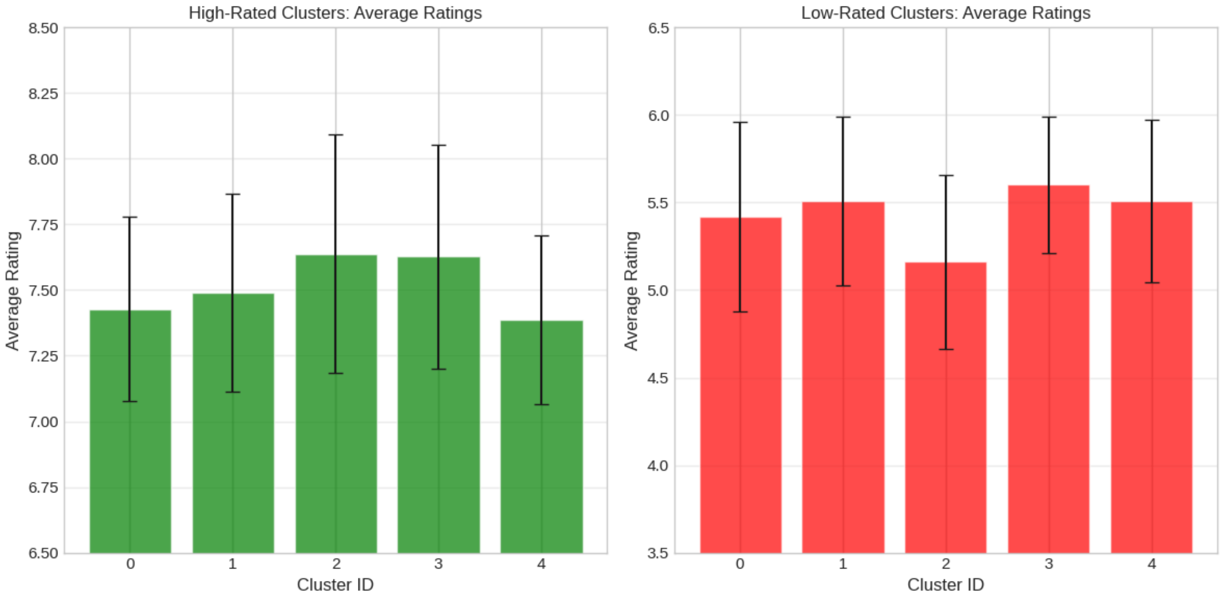
Visually, clusters are less well-defined here than in the high rated games, reflected in the much lower CH score for the best model. Again, the grid search found the minimum of 5 clusters provided the best results.

**Cluster Details:**

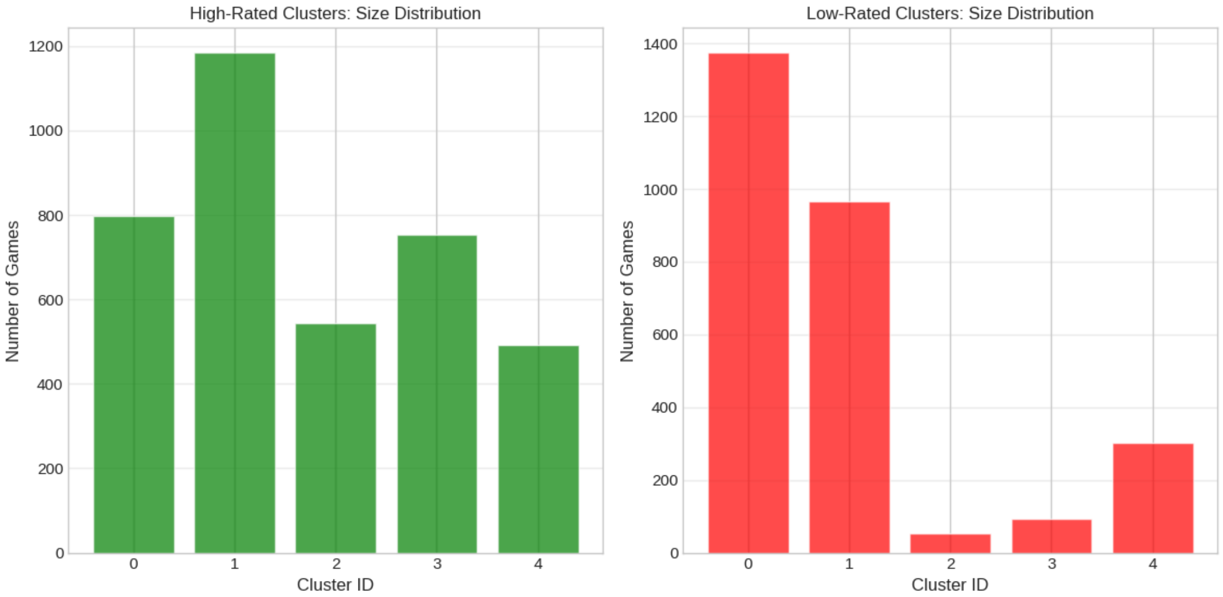
| Cluster | Description | % of low-rated games | Avg. Rating | Mechanics | Themes | Categories/ Subcategories | Top Decision Tree Features | Example Game |
|---|---|---|---|---|---|---|---|---|
| 0 | Family & Children's Games | 49.2% | 5.42 | Dice rolling (27.8%), roll/spin and move (21.5%), dexterity (12.8%) | Movies/TV (10.1%), economic (8.7%), trivia (8.7%) | Family (19.2%), children's games (14.6%), abstract (10.8%) | Card game (0.717), War (0.133), fantasy (0.085) | Astro Drive (2018) |
| 1 | Card Games | 34.6% | 5.51 | Hand management (33.1%), set collection (22%), betting/bluffing (16.9%) | Humor (15%), animals (12.4%), fantasy (5%) | Card games (73.4%), family (20.2%), children’s games (9.8%) | Card game (0.714), dice rolling (0.096), deduction (0.082) | Cat Town (2016) |
| 2 | Monopoly Games | 1.9% | 5.16 | Roll/spin and move (100%), negotiation (98.1%), trading (98.1%), auction/bidding (96.2%) | Economic (100%), movies/TV (32.1%), video game (13.2%) | Family (24.5%), electronic (11.3%), children’s games (7.5%) | Roll/spin and move (0.520), negotiation (0.464), player elimination (0.014) | Monopoly: Pokémon Kanto Edition (2014) |
| 3 | War Simulation Games | 3.4% | 5.60 | Dice rolling (74.5%), hexagon grid (68.1%), simulation (61.7%) | World War II (29.8%), modern warfare (22.3%), science fiction (14.9%) | War games (100%), territory building (7.4%), miniatures (4.3%) | War (0.819), hexagon grid (0.131), simulation (0.050) | Pegasus Bridge: The Beginning of D-Day (1988) |
| 4 | Themed Fantasy Games | 10.9% | 5.51 | Dice rolling (44.9%), hand management (25.7%), variable player powers (14.9%) | Fantasy (51.5%), fighting (37.3%), humor (16.2%) | Thematic (34.7%), card games (51.5%), exploration (9.6%) | Fantasy (0.525), fighting (0.307), thematic (0.103) | Barbarian Kings (1980) |

Overall, the predominance of dice rolling and roll/spin move mechanics across Clusters 0, 2, 3, and 4 shows how players tend to value strategy and making choices over simply depending on luck.  Cluster 0 seems to focus on media/TV themes mostly for family and children, with no significant gameplay mechanics or strategy to back it up. Cluster 1 is also mostly card games like Cluster 0 in the high rated games, yet seems to focus on family/children and fun themes over more strategic elements like deck building and drafting, making it less received overall. Cluster 2 is very small (1.9% of the data) and seems to interestingly encompass mostly Monopoly-style games. It could be that reusing mechanics and gameplay while slightly changing themes isn't received very well by gamers. Interestingly, Cluster 3 has very similar characteristics as Cluster 3 in the high rated games, while being rated much lower overall. It's hard to tell, but it could be due to it being such a small cluster (only 3.4% of the data), and thus individual games skew the average rating much more. It could also be that it lacks the mechanics and strategy that integrate well with war games. Cluster 4 focuses on popular themes like fantasy and fighting, yet it's also hard to tell if the lack of compelling mechanics or strategic elements is the reason for lower ratings.

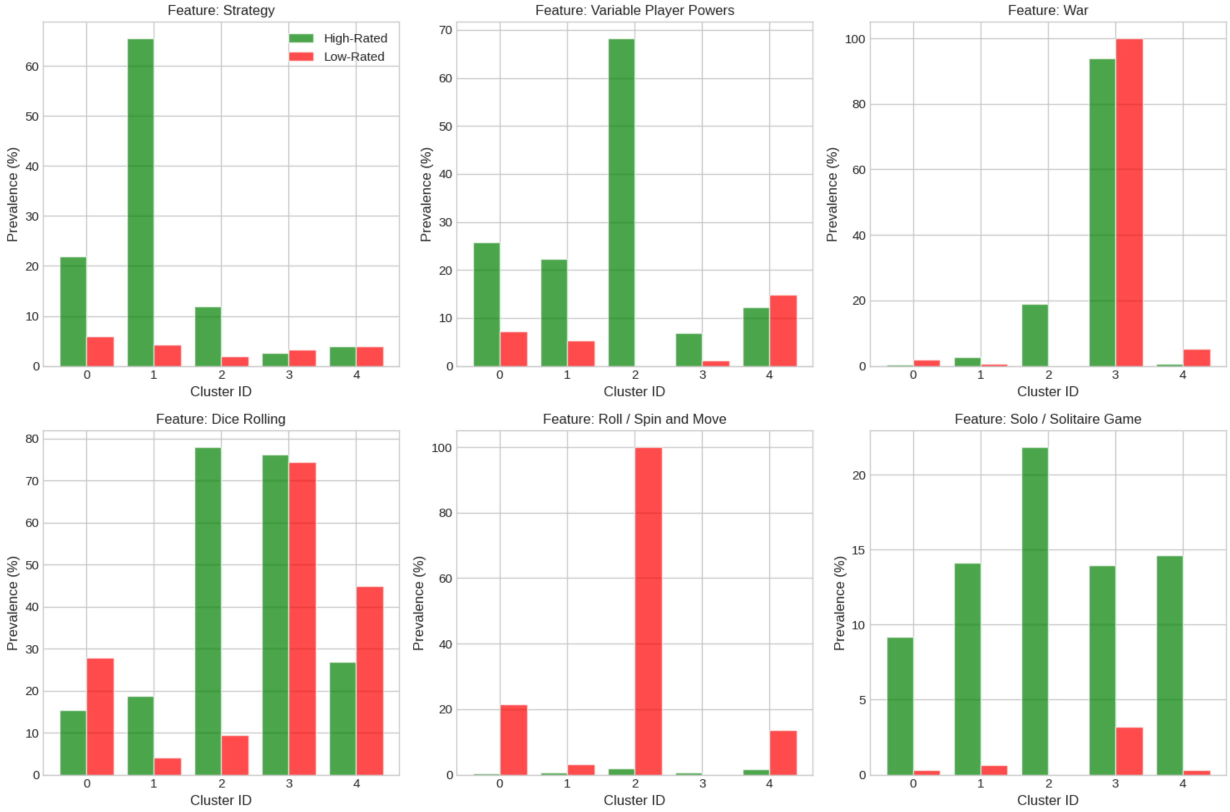
### 6.2.3 High/Low Clustering Comparison



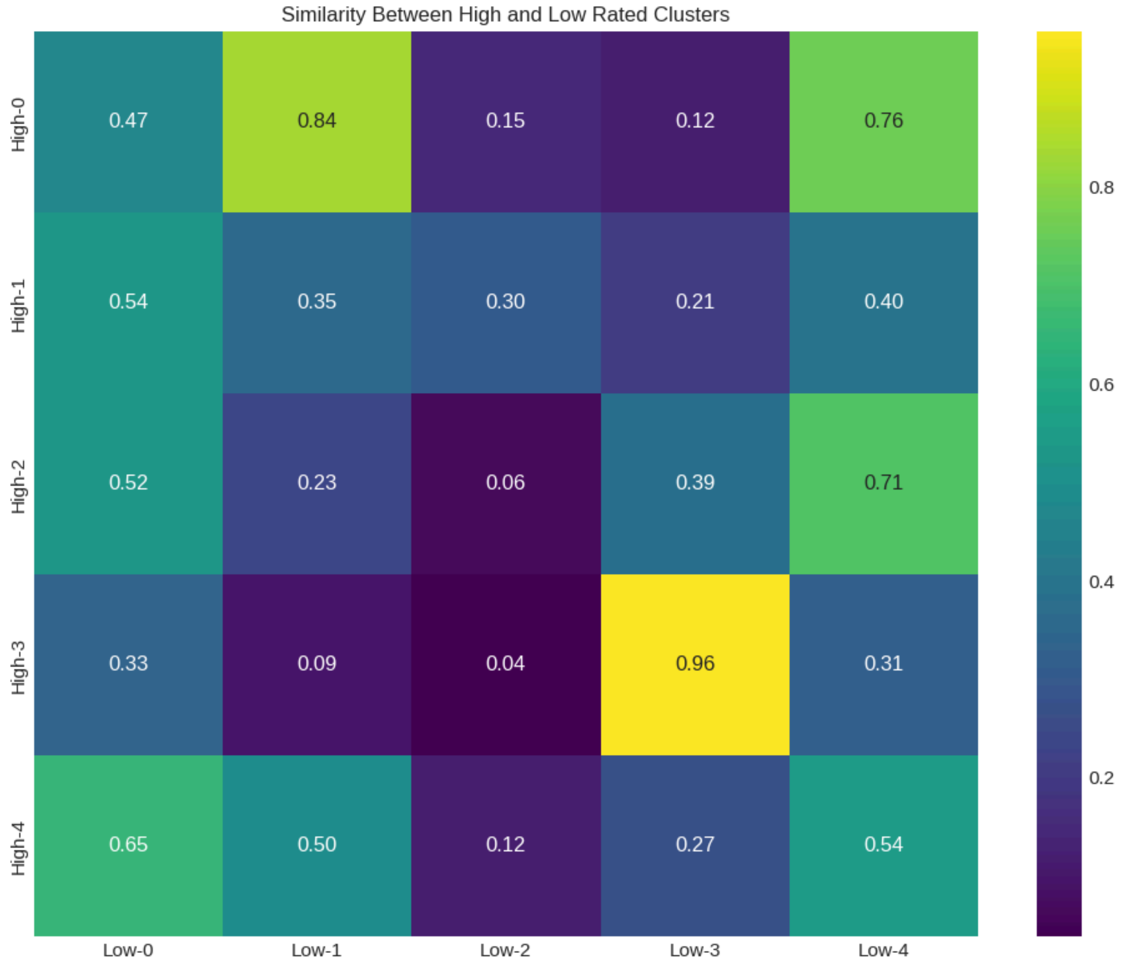
Average ratings within high-rated clusters range from 7.4 to 7.65, with clusters 2 and 3 achieving the highest average ratings and cluster 4 having the lowest. The low-rated clusters display averages between 5.15 and 5.65, with cluster 3 having the highest average and cluster 2 the lowest. Note that the error bars also indicate consistent ratings' variability across all clusters, high or low rated, indicating consistent clustering techniques.



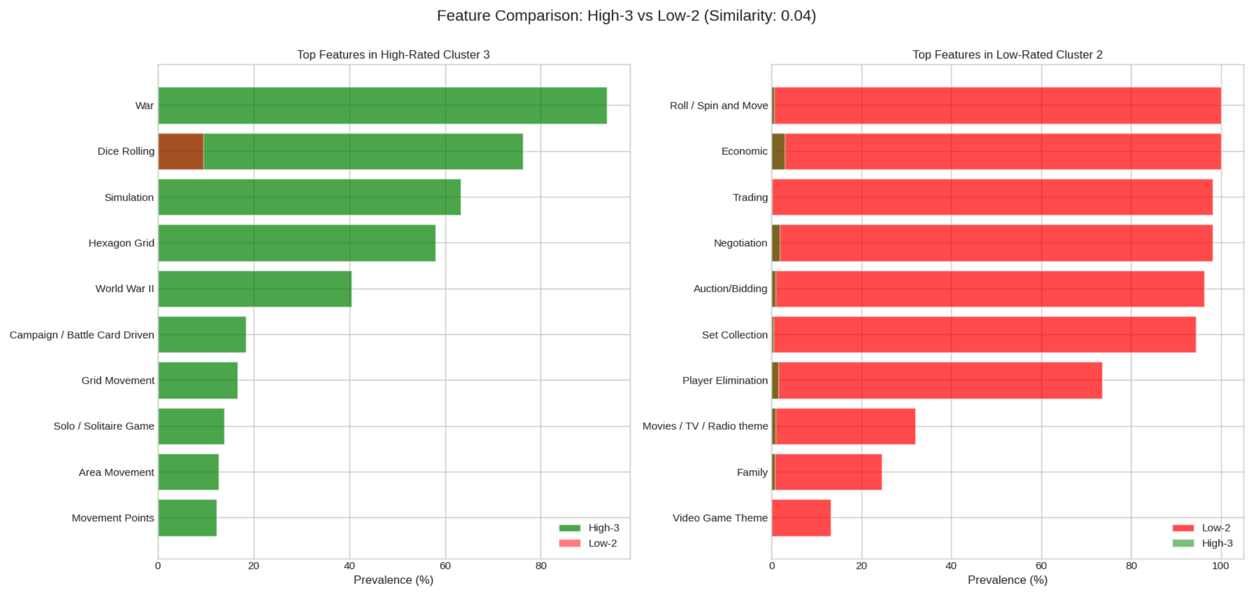
High-rated games form more balanced clusters in terms of size, with the largest cluster having almost 1200 games and other clusters having 600 to 800 games. Low rated clusters have much more imbalanced sizes, with two large clusters accounting for most of the games (around 2400 games) along with three much smaller clusters (50 to 300 games).



These graphs show the prevalence of top distinguishing features across both high and low rated games in the clusters. We can see that strategy, variable player powers, and solo/solitaire game are almost always in the high rated clusters only, while roll/spin and move is almost always in the low rated clusters only. These are features that can mostly stand alone when predicting ratings. On the other hand, dice rolling and war, as discussed before, can be in both high rated and low rated clusters, and thus are not good predictors of rating by themselves.



This visual represents the similarity matrix between high and low-rated clusters, which reveals how similar feature vectors are in each cluster by using cosine similarly. The high similarity between certain high and low-rated clusters, particularly High-3/Low-3 (war games) and High-0/Low-1 (card games), suggests that implementation quality and effective integration/balance of features is more important than the mere presence of specific features alone. Despite this, we also see clusters with very low similarity, particularly High-3/Low-2, High-2/Low-2, and High-3/Low-1.



This shows the prevalence of top features in the most dissimilar cluster pairs (war simulation games vs. monopoly-style games). Interestingly, these are also the highest rated cluster and lowest rated cluster out of all the clusters. The high rated cluster has strong war themes, lots of strategic depth, diverse grid/movement systems, and options for solo gameplay. The low rated cluster has the dreaded roll/spin move feature, economic and trading mechanics, player elimination, as well as movies/TV/video game themes.

Overall, this analysis demonstrates that while certain features do correlate with higher ratings, successful board games excel through thoughtful integration of categories, mechanics, themes, and player experience rather than simply including popular features.

## 6.3 Linear Regression

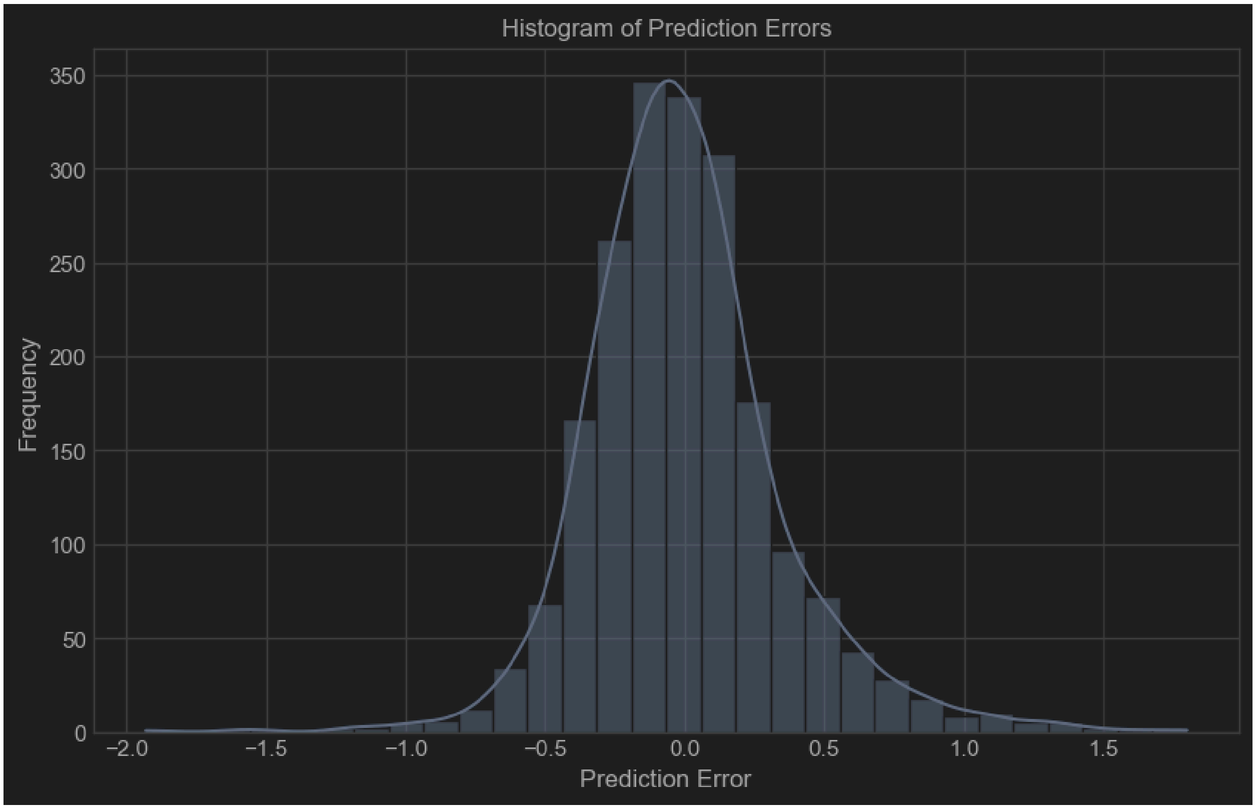
### 6.3.1 Linear Regression (LR) Results

After applying the linear regression model to the filtered dataset, we assessed the model using Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination ($R^2$). The results are as follows:

***Model Performance:***

Mean Squared Error (MSE): 0.1219
Mean Absolute Error (MAE): 0.2557
$R^2$ Score: 0.8166

Note that the $R^2$ value ranges from values 0 to 1, and a *lower* MSE and MAE value is indicative of a better model, while a *higher* $R^2$ value is indicative of a better model. These results indicate that the linear model explains approximately 82% of the variance in the average ratings, but the error shows that there is certainly room for improvement.



The graph above shows the distribution of predictions, showing that the majority of predictions have low error. A bell-shaped curve is indicative of a stronger model, although again, there is certainly room for improvement.

### 6.3.2 Random Forest Regressor (RFR) Results

To better capture nonlinear interactions, we implemented an RFR. A grid search was conducted with 5-fold cross-validation over 36 parameter combinations, totaling 180 fits. The optimal parameters that we found were as follows:

Best parameters: {'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 200}

The best cross-validation $R^2$ score during grid search was **0.95617**. When evaluated on the test set, the Random Forest model achieved:

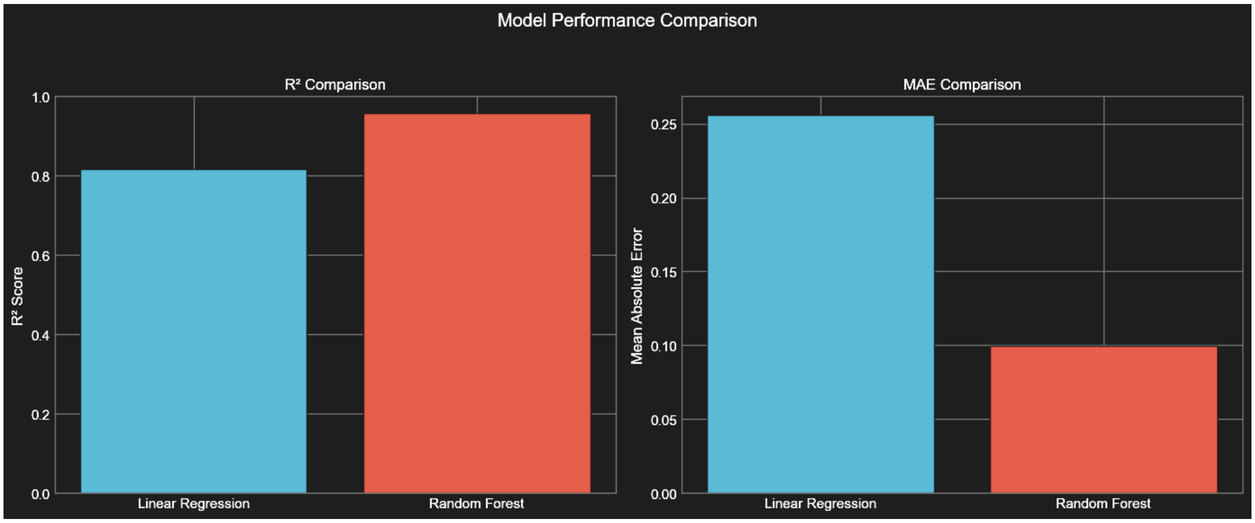
- **Mean Squared Error (MSE):** 0.02795
- **Mean Absolute Error (MAE):** 0.09931
- **R² Score:** 0.95797

These results were very promising, as they were extremely indicative of a well-trained model.

### 6.3.3 LR & RFR Model Performance Comparison

The Linear Regression model explains around 82% of the variance in the average ratings, whereas the Random Forest Regressor explains roughly 96%. The RFR proved substantial improvement over the linear regressor, which makes sense taking into account that the attributes of games are most likely more non-linear than the latter.

With an MAE of approximately 0.26, the LR model's predictions are on average off by about a quarter of a rating point, which is suboptimal. The RFR model, with MAE of 0.10, demonstrates a much better prediction accuracy. Similarly, the significantly lower MSE for the Random Forest model reflects its ability to limit larger errors.



The clearly better-performing model, RFR, indicates that board game ratings are influenced by complex interactions among features. While the linear regression model does provide a useful baseline, its linear assumption inhibits its ability to predict more complex datasets. The RFR on the other hand is much more effective at modeling these intricate patterns, making it the preferred method to predict game ratings from numeric attributes.  Overall, we showed that we can in fact find a correlation between game attributes and the overall rating of the game, and we can build a model to accurately predict a value metric, in this case average rating, given game attributes.

## 7. Conclusion

In this project, we explored a complex dataset of board games sourced from BoardGameGeek and explored three different key methods of uncovering relationships and predictions within the data. The analytical methods we covered were collaborative filtering, clustering, and predictive modeling using both linear regression and random forest models. These three approaches each produced their own insights into how the board games relate to each other, how users are similar to one another, and the factors on what makes a board game successful. To provide a brief summary of the answer to our research questions (which were answered comprehensively in our results sections):

1. Which attributes of a game tend to produce the highest user ratings?
    a. It seemed like players preferred games with strategic depth; freedom of choice matters to players, and excessive reliance on luck-driven mechanics tended to lead to lower ratings. Even so, it seemed like the execution, quality, and effective integration of features was sometimes more important than just the features alone.
2. Can we cluster board games into meaningful categories based on mechanics and complexity?
    a. Yes—five distinct clusters clearly emerged in the high and low-rated segments of the data that we examined with decent separability. It was easy to see that categories like card games, classic strategy games, dice-based games, war games, etc. were clearly separated in the clusters.
3. How effectively can we predict a "value" metric (e.g., average rating) for a board game from its attributes?
    a. We explored both the use of linear regression and random forest regressors. Both models, but specifically the random forest model, performed extremely well.
4. Given a user's past game ratings, can we predict their rating of new games based on ratings of similar users?
    a. Collaborative filtering on this dataset worked very well. We constructed a user-item matrix and used a collaborative filtering approach that used k-nearest neighbors with adjusted weighted sum and cosine similarity.

### 7.1 What we learned/discovered

Analyzing data is not a one-and-done process. It is not something that can be treated as a singular problem. Throughout the project, we realized that our initial findings were never the main takeaways; we only discovered our most significant findings after diving deeper, asking "why" whenever we had results, and trying to push our analytical methods further.

In our clustering analysis, for example, we first observed overlap and lots of noise in our initial clustering methods. By deciding to split the dataset into high and low-rated subsets, we were able to reduce some of the noise and uncover much more distinct groupings that led to better analysis. When analyzing our clusters, we also realized that "why" is more revealing than "what"; identifying the clusters was not enough to show us the relationships in the data. The card games clusters in both the low and high rated subsets, for example, appeared to be very similar at first, but after closer examination of the features of those games, it became very evident that the two clusters were very different. In the high rated subset, those card games were more likely to involve strategic elements (like deck building and drafting) compared to more luck-based elements in the low rated games.

Continually asking more questions when faced with initial answers is what drove our studies to excellence, and by employing as many differing methods as we did for the data, we were able to paint a better overall story for the dataset as a whole.

## 7.2 Societal Impact

Analyzing board games extends beyond technical modeling; it has real-world implications for designers, consumers, and anyone with an interest in the ins and outs of board game curiosity. Our study highlights several key societal impacts that our research may allude to.

We have shown correlations between game attributes and game ratings. Game designers can utilize this fact (and our prediction models when tuned to their specific genre of game making)  to help them create engaging games for their consumers. However, this generalization of game creation may be at risk of creating many games that are too homogeneous. In other words, an over-reliance on the data predicted by the models may lead developers to create too many games that are similar to each other, creating formulaic designs that will eventually bore the consumer. Also, the reliance on the data results may result in a lack of, or loss of, creativity in the development of games.

The ratings and metadata reflect community biases and preferences. The algorithms that are trained on these mainstream/biased data will reinforce existing trends. This means that mainstream games will gain more visibility, and games that are more niche will become even more hidden. It is important to address bias as a key to keeping inclusivity and representation of all types of games in the board game market.

Board games are not only there for entertainment, but they contribute to local and global economies. With these modeling techniques, these data-driven insights will help to shape marketing, promotions, and investments. Popular titles will benefit, but industry inequalities may increase, shoving the less popular, more niche games down into the ground.

In conclusion, while our technical analysis allows us powerful predictive abilities, we must also consider the real-world impact in terms of universal societal responsibilities. Ethical data practices, awareness of biases, and balanced innovation and standardization are important to be considered and taken into account to ensure that data analytics enrich the community as a whole without, in the process, limiting creativity.