MIS710 – Assignment 1

MACHINE LEARNING IN BUSINESS

BA HUY HOANG LE

s224309594

Table of Contents

Executive Summary

This report seeks to improve PQC's business performance by analysing factors influencing game ratings, thereby increasing revenue and profits. It examines the business context to enhance business understanding. It also analyses the dataset to uncover valuable insights, choose an appropriate predictive model, and make recommendations for future directions.

Key insights:

1. Games for those above 12 generally receive higher ratings.
2. The more challenging a game is rated, the higher the rating.
3. PremiumGame games generally receive higher ratings than BaseGame
4. The more time games spend on the game, the higher the rating likely to receive

A multilinear regression model was selected to predict the average rating based on age category, game complexity, average play time, and game type. The model could only provide a foundational understanding of these relationships. It only explains a small percentage of the ratings' variations. Despite a relatively low error magnitude, the model provides limited insights into the complete range of influences on game ratings.

The current model is not suitable for predicting average ratings of future games. It is recommended that PQC employ more advanced predictive modelling, record new features such as Game genres, and remove features that provide little to no insights into average ratings.

Business Understanding (BACCM)

Problem

Current data shows that PQC's game ratings are only average. Many games also have extremely low ratings, reflecting low customer satisfaction and/or overall gaming experience. This underlines a core problem in designing and delivering games with top-notch gaming experience in the industry.

Value

PQC can enhance the quality of its games, attract a larger client base, and ultimately boost its financial performance.

Changes

Addressing the issue will make PQC the top game provider, offering high quality and an exceptional gaming experience.

Solution

PQC should allocate additional resources to enhance game quality and deliver games that surpass user expectations. Some effective strategies involve enhancing comprehension of issues that influence game ratings, incorporating feedback mechanisms, and consistently improving game design and features.

Context

After the COVID-19 pandemic, the gaming industry has experienced intense competition (López-Cabarcos et al, 2020). A surge in demand has led to higher market saturation and accelerated digital adoption. Growing competition has elevated the importance of game ratings and customer satisfaction. High ratings and favourable feedback imply great gaming experiences, crucial for standing out in a crowded industry. As a result, gaming businesses are attempting to innovate and win market share amid fast-changing consumer expectations (Modgil et al., 2022)

Stakeholders

PQC's top management provides the dataset to analyse the factors that influence average ratings and extract valuable insights from the analysis of each feature in the dataset.

Game players who rate the game and evaluate how good the game is.

The analysts who analyse and evaluate key findings to provide actionable insights and facilitate future decision-making for PQC's game business.

Data Understanding

1. Data Overview and Preparation

The provided data has 24813 inputs and 17 features. Only seven inputs have missing values, and the feature with the missing value is “Game_Name.”

“Game_ID" and “Game_Name” were excluded

- "Game_ID" is a unique identifier for every game; it offers no valuable insights to explore relationships or patterns within the dataset.
- "Game_Name" is a categorical text variable that does not contribute significantly to quantitative analysis. Including this feature could make the analysis more complicated without adding valuable insights.

To facilitate further analysis and development of the machine learning model, new features were added to the data by converting some of the original features:

- “Release_Year" was converted into a categorical variable with four categories: 1970s, 1980s, 1990s, 2000s, 2010s, and 2020s and beyond
- “Age_Category" was converted into a numerical variable. “Under 5,” “5 to under 12,” “12 to under 18,” “18 to under 21,” and “21 and over”, became 0, 1, 2, 3, 4, respectively. The new variable name is “Age_Category_N”
- “Game_Type" was also converted into a numerical variable. “BaseGame" and “PremiumGame" became 0 and 1, respectively. The variable is “Game_Type_N”

2. Data Exploration and Visualisation

a. General Information and Game Configuration

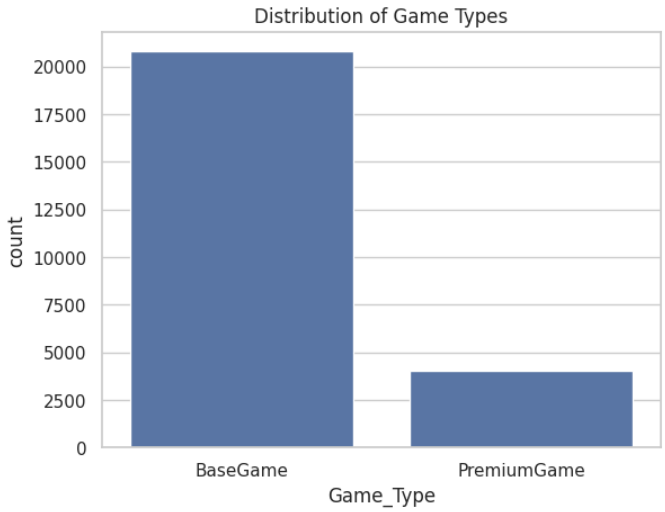


Figure 1 shows that the BaseGame type dominates the dataset, accounting for 20,7669 games. In contrast, the PremiumGame type only accounts for 16.2% of the total, with 4,017 games.

Figure 1. Distribution of Game Types

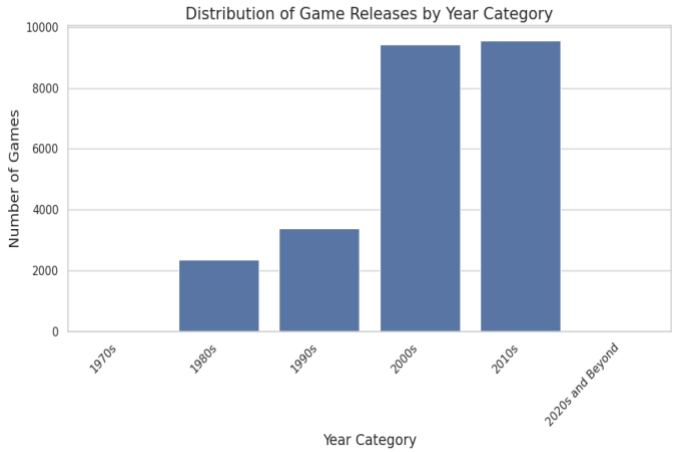


Figure 2 shows that most games in the dataset were released in the 2010s and 2000s, with 9559 and 9425 games, respectively.

Figure 2. Distribution of Game Releases by Year Category

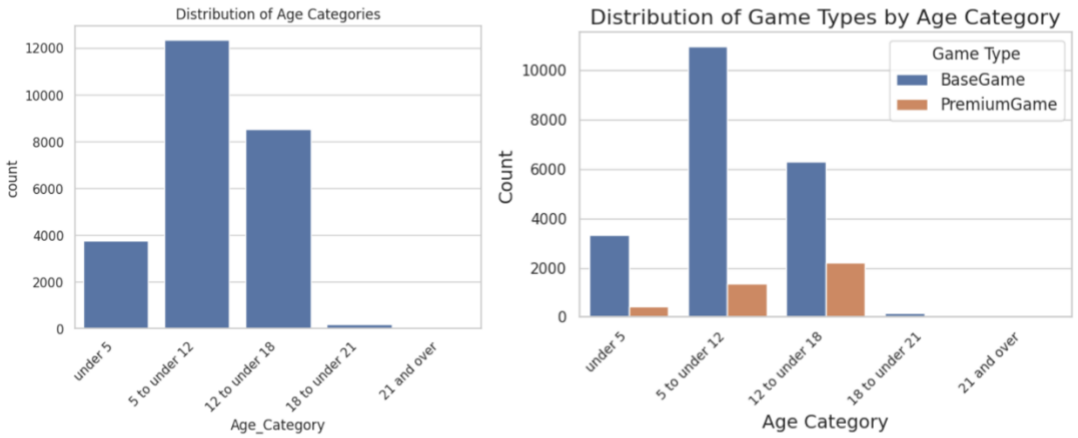


Figure 3. Distribution of Age category and each group choice of game types

From Figure 3, more than 99% of games are recommended for players under 18. Notably, 12,335 games—approximately 50% of the total—are aimed at children aged 5 to 12. In contrast, only 183 games target gamers aged 18 to 21, with only 24 games recommended for those aged 21 and over.

BaseGame is the most popular game type across all age groups, especially for the 5 to under 12 group (10979 compared to 1356 choosing PremiumGame). PremiumGame is more common among 12 to under-18-year-olds than the other age groups, with 2203 games.

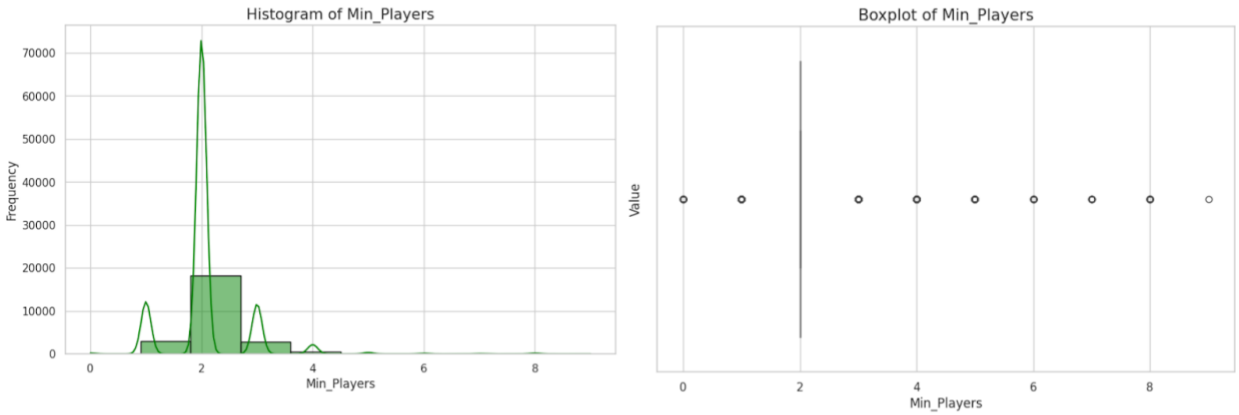


Figure 4. Overview of the Minimum number of players required

Figure 4 illustrates that most games require at least two players, with 18,177 games falling within this category. The number of games lowers dramatically when there are higher player numbers, with only one game in the dataset requiring a minimum of nine players. The boxplot reveals that the minimum number of players typically ranges between 1 and 3.

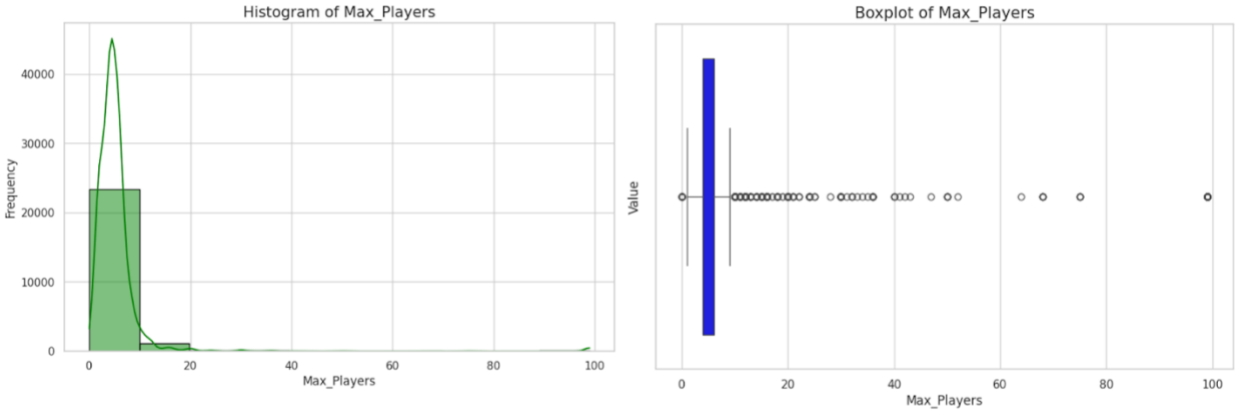


Figure 5. Overview of Maximum number of players allowed

The histogram and boxplot in Figure 5 show that most games are built for up to four players, with a considerable number tolerating up to six. This indicates a strong market focus on smaller group experiences. The distribution shows a gradual decline in the number of games as the maximum number of players increases; only a select few games support very large player counts, like 75 or 99.

b. Summary of Average Playtime

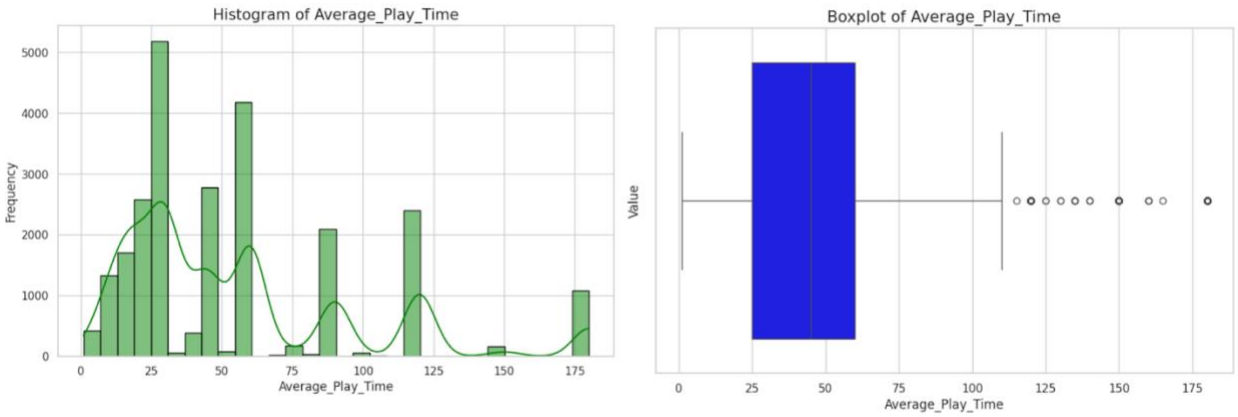


Figure 6. Overview of Average playtime

Statistics	Overall Average Play Time (min)	Outliers Average Play Time (min)
Count	24813	3672
Mean	55	139
Std	43	27
Min	1	115
1st Quarter	25	120
Median	45	120
3rd Quarter	60	180
Max	180	180

Table 1. Summary statistics of Average playtime

Analysis from Figure 6 and Table 1 indicate that the average playtime for most games is 55 minutes, with significant variations between titles. Specifically, 3,672 games had extended playtimes, ranging from 115 to 180 minutes and averaging 139 minutes. This shows that, while most games have shorter playtimes, many are meant for longer periods, reflecting a wide range

of gaming experiences and highlighting opportunities to tailor products to different gaming preferences.

c. Relationship between Playing Time and Average Ratings

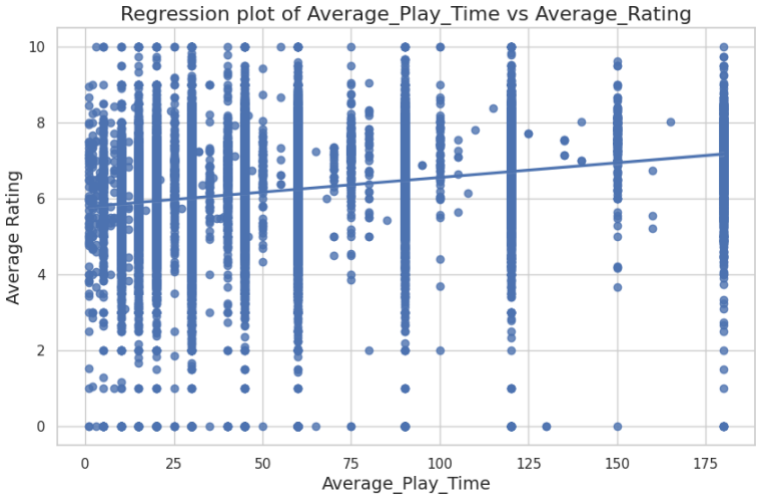


Figure 7 indicates a moderate positive relationship between average play time and average rating, with a correlation coefficient 0.22 (Appendix 1). As players devote more time to a game, its average rating tends to be higher.

Figure 7. Regression plot of Average playtime and Average rating

d. Relationship between Level of Game Complexity and Average Ratings

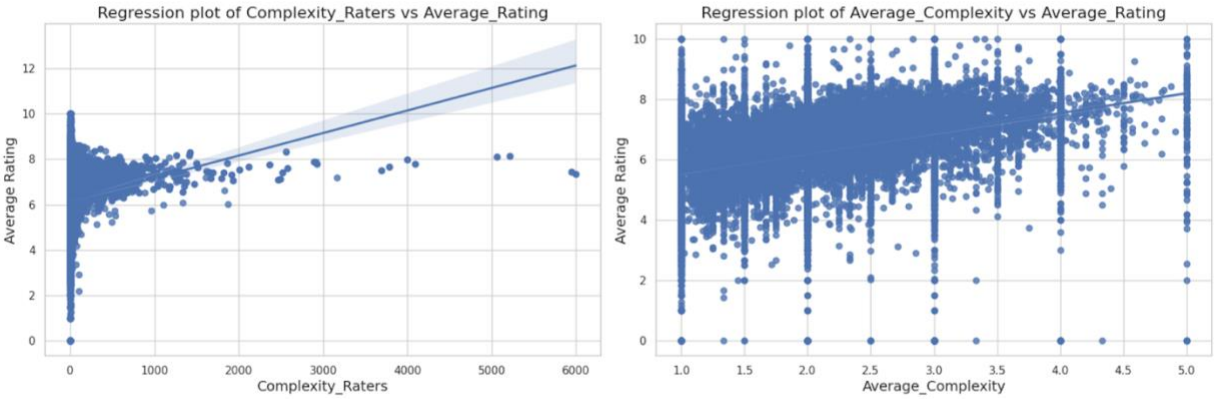


Figure 8. Regression plot of Complexity raters and Average complexity against Average ratings

Figure 8 illustrates a relatively strong positive relationship between the game's average complexity and average ratings, with a correlation coefficient of 0.36 (Appendix 1). A positive but weaker relationship exists between the number of people rating the game's complexity and the average rating, with a correlation coefficient 0.12 (Appendix 1). While more complex games tend to have higher ratings, the volume of complexity ratings alone has little impact on how high the game is rated.

e. Correlation of Game Configuration, Popularity, and Interest with Average Ratings

i. Relationship between Age Category and Average Rating

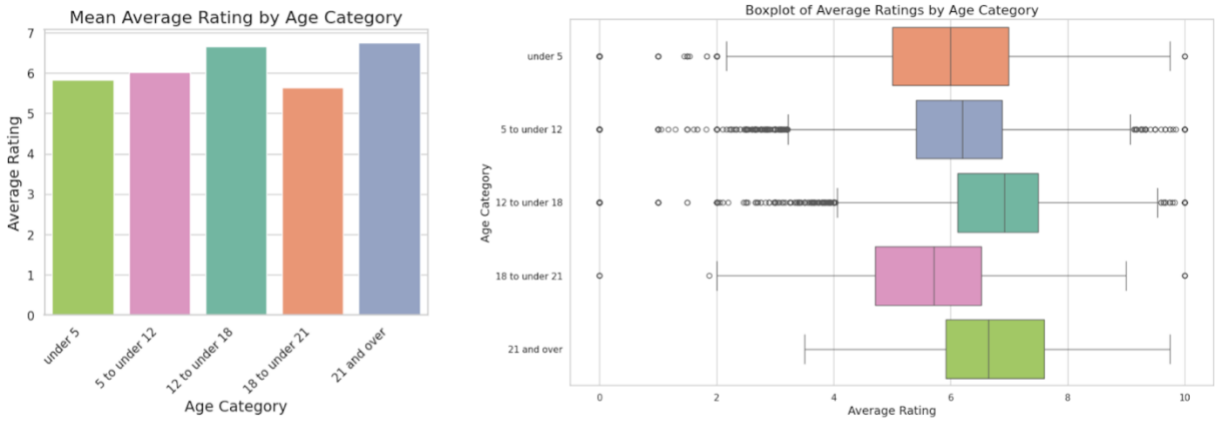


Figure 9. Bar chart and boxplot of Age category against Average rating

Age Category	Mean of Average rating
Under 5	5.824
5 to under 12	6.033
12 to under 18	6.660
18 to under 21	5.641
21 and over	6.744

Table 2. Summary statistics of mean of Average rating among each Age category

Figure 9 and Table 2 show that Games for older players (12 and above) generally receive higher average ratings, with a marginal mean of 6.35 compared to 5.93 for younger players. While there are some variations within these age groups, the average game rating tends to be higher at games for older players.

ii. Relationship between Minimum Players and Average Rating

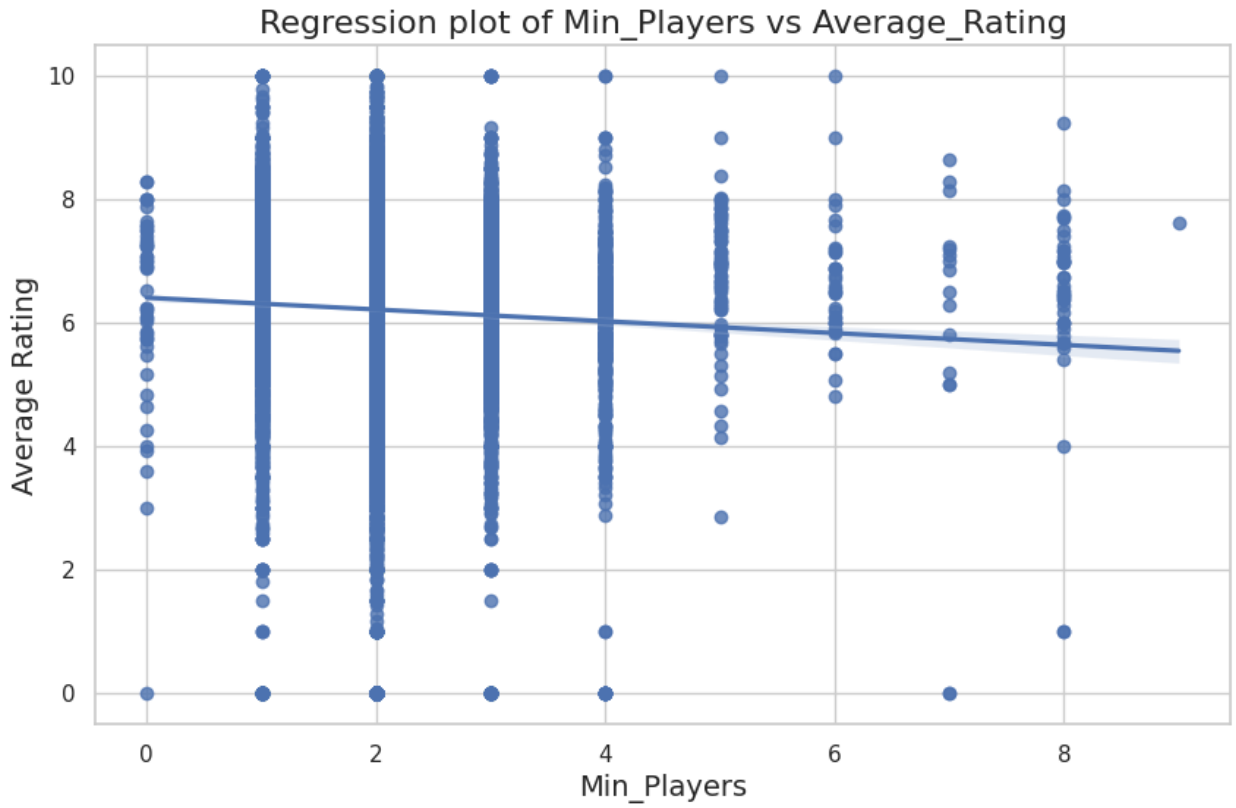


Figure 9. Regression plot between Minimum players required and Average ratings

Figure 9 shows a slight negative relationship between the minimum number of players required to play a game and its average rating, with a correlation coefficient of -0.04 (Appendix 1). This implies that, while not strongly related, games requiring fewer players are marginally more popular with players, as demonstrated in their higher average ratings.

iii. Relationship between Maximum Player and Average Rating

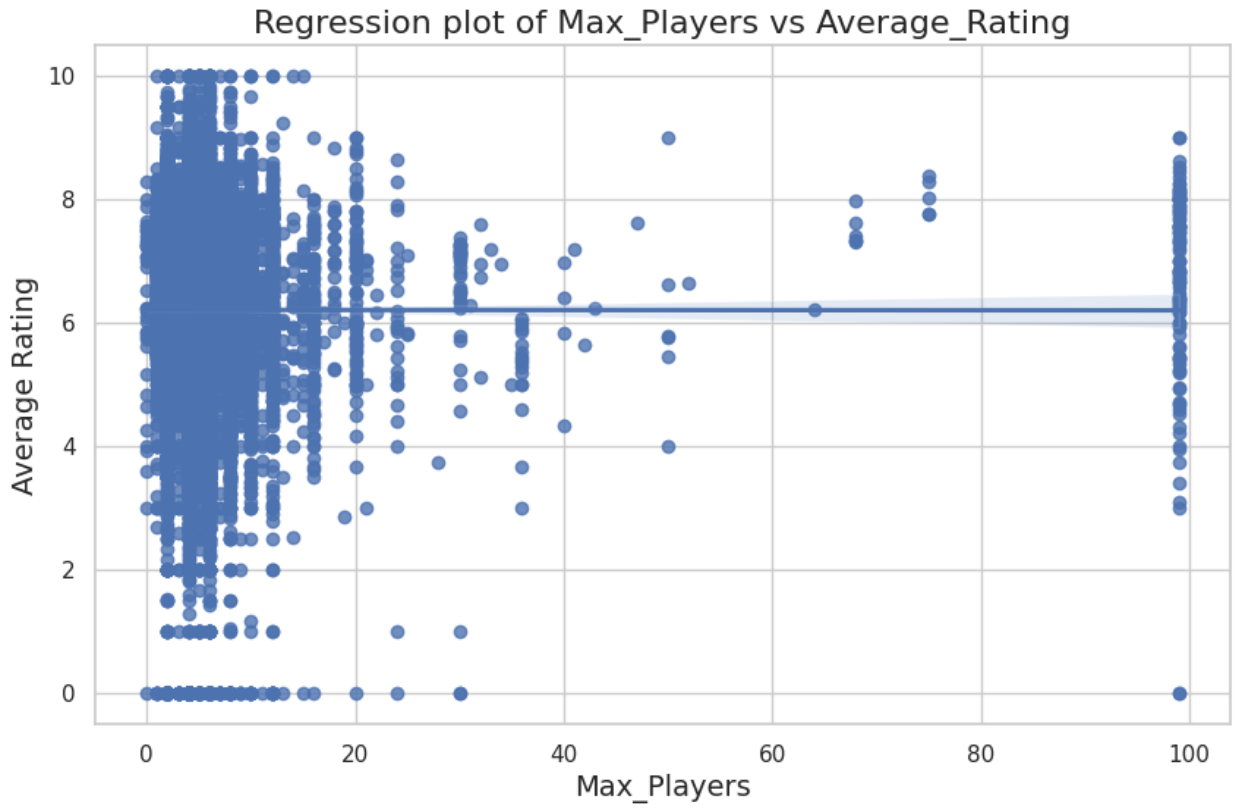


Figure 10. Regression plot between Maximum number of players allowed and Average rating

Almost no correlation exists between the maximum number of players a game can support and its average rating. This is demonstrated by a correlation coefficient of approximately 0, demonstrating that the number of players a game can support has little to no effect on the game's average ratings.

iv. Relationship between Number of Game Ownership and Average Rating

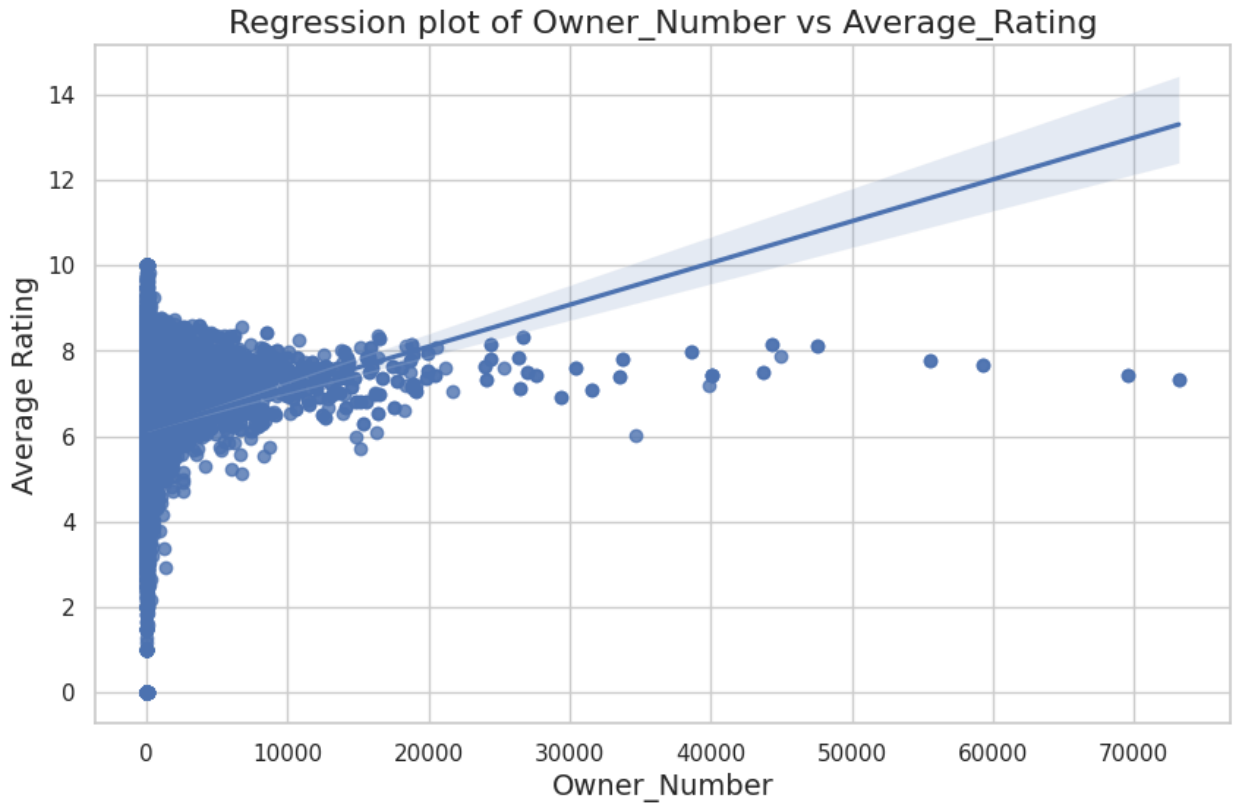


Figure 11. Regression plot between Game ownership and Average rating

Figure 11 indicates a positive correlation between game ownership and average rating, with a coefficient of 0.16 (Appendix 1). While games with a larger owner base tend to have higher ratings, the relationship is not strong.

v. Relationship between Trader Number and Average Rating

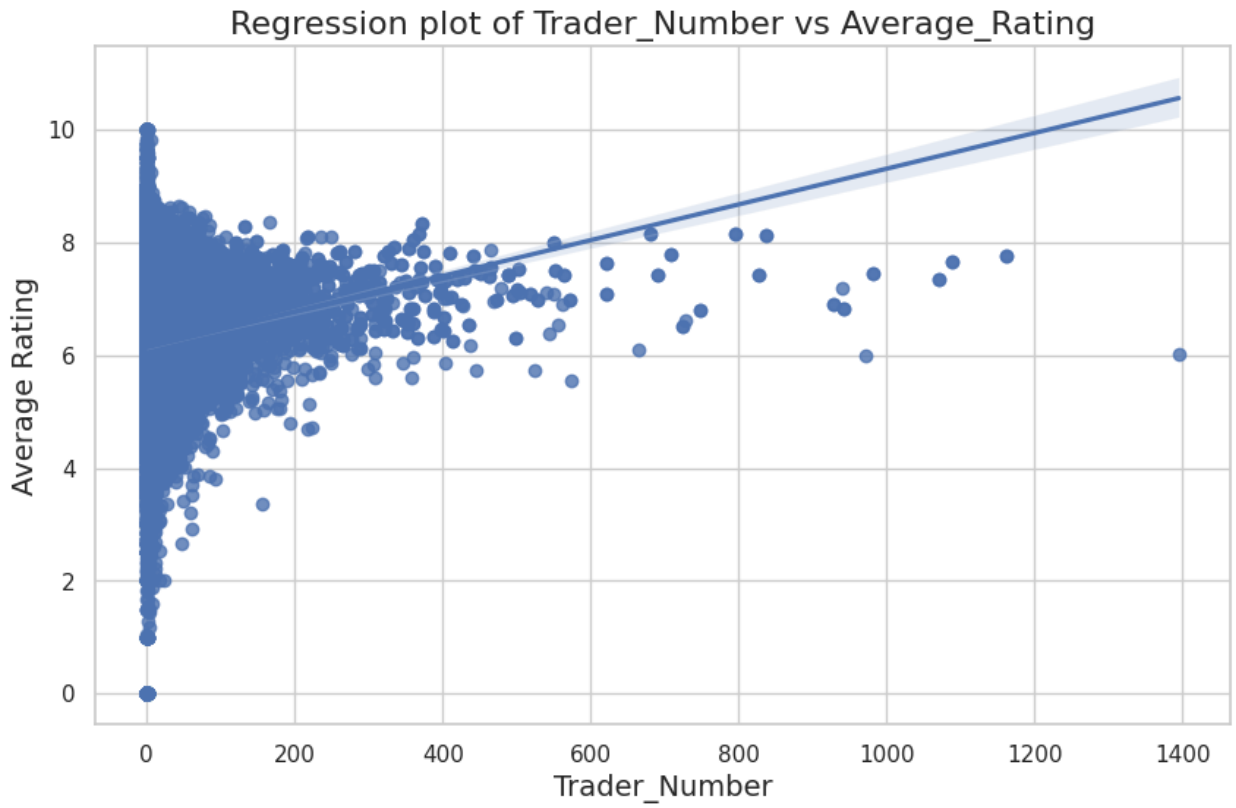


Figure 12. Regression plot between Trader number and Average rating

Similarly, the number of traders that traded games with others correlates positively with a game's average rating. While games with higher trading volumes have slightly higher ratings, their impact on overall ratings is minor.

vi. Relationship between Game Interest and Average Rating

Figure 13. Relationship between Game interest and Average rating

The analysis discovered similar patterns regarding interest in the game and its influence on its average rating. The more players indicate that they are generally or very interested in playing the game, the higher the average rating tends to be. However, the effect is moderate, with coefficient correlation values of 0.23 and 0.20, respectively.

vii. Further Exploration of the Data

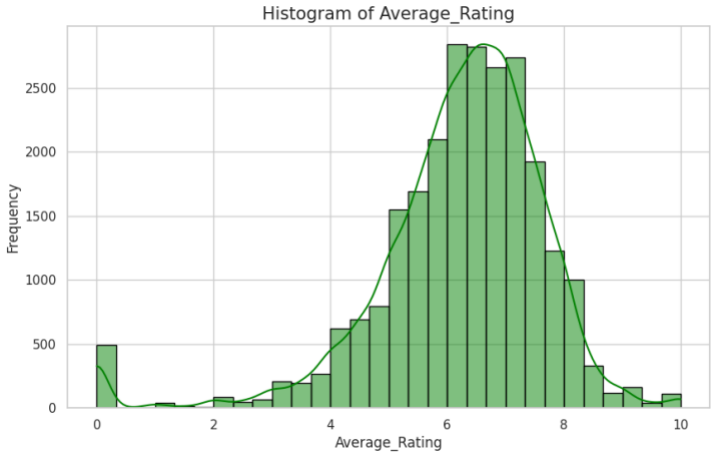


Figure 14 shows that game players are generally satisfied with their games, with average and median ratings nearly equal, 6.24 and 6.42, respectively. However, there are many games with an average rating of nearly 0.

Figure 14. Histogram of Average rating

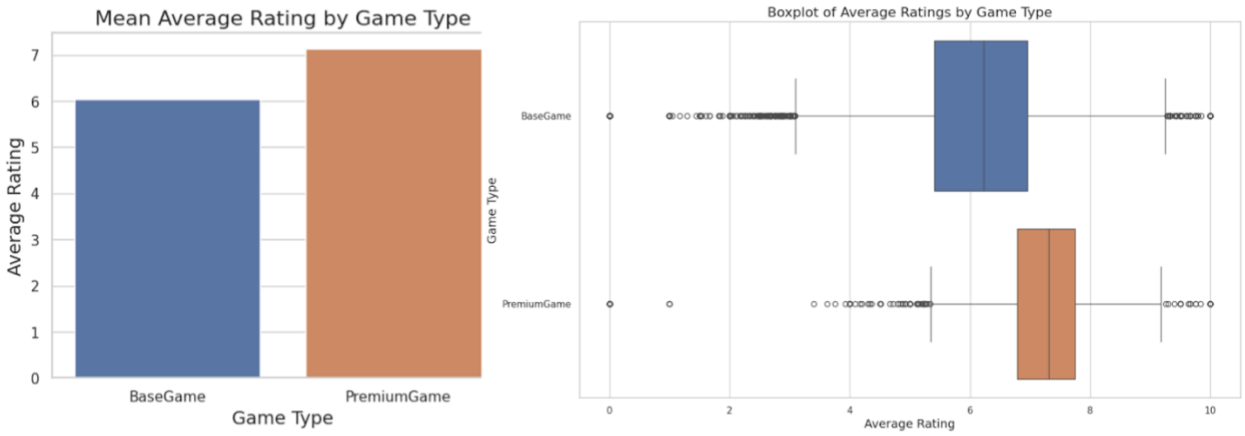


Figure 15. Average ratings between two Game types

Figure 15 shows that players generally give higher ratings for PremiumGame. The mean average rating for the premium game is 7.13, and 6.037 for BaseGame. This demonstrates that players have a more favourable attitude toward PremiumGame, suggesting a larger market appeal.

3. Estimating Average Rating through Machine Learning Model

a. Feature Selection

Using the heatmap data, the following features were selected because their correlation coefficient values are above 0.2, indicating a moderately strong relationship with our target variable - Average Rating:

- Game_Type_N
- Average_Complexity
- Average_Play_Time
- HighInterest_Number
- Interest_Number

b. Model Selection and Preparation

To estimate the average rating of each game, this report employs a multilinear regression model because the target variable is a labelled continuous numerical variable.

The original data was split, with 70% of the data used to train the model and the remaining 30% used to test the model.

c. Model Result and Interpretation

The multi-linear regression result was

Estimated Average_Rating = 5.01 + (-0.00)*Interest_Number + (0.00)*HighInterest_Number + (0.00)*Average_Play_Time + 0.48*Average_Complexity + 0.8*Game_Type_N

- When the effects of all predictors are absent, 5.01 is the baseline predicted average rating of a game.
- Given that all other variables remain constant, each unit rating increase in game complexity increases the estimated average rating by 0.48 rating points.
- Given that all predictors remain the same, the estimated rating for a PremiumGame type is expected to be 0.8 points higher than that for a BaseGame type.
- The effects of game interest are virtually negligible under this model. Given that all factors remain constant, an increase in the number of players generally interested or very interested in the game would not change the average ratings.

d. Model Performance and Evaluation

The R-Square is low (0.21). The model can only explain 21% of the variation of the average ratings by the predictors chosen in the model. In contrast, 79% of the results cannot be interpreted from the model. The low R-square could be attributable to a variety of factors. First is the insufficient number of predictors in the regression model. Second, many features were not selected because they had correlation coefficient values. Third, the chosen predictors only have moderate correlation strength. The high VIF values for Interest_Number (12.67) and HighInterest_Number (12.78) indicate multicollinearity issues, which might skew the model's coefficient estimations and limit its capacity to explain variation in average ratings.

The MAE value of 0.86 is relatively low, meaning predicted average ratings are generally off by 0.86 points from the actual rating. This relatively small error suggests that the model's predictions and true ratings are fairly close.

The RMSE value (1.30) is higher than the MAE, suggesting there are some large discrepancies between actual and predicted average ratings due to outliers.

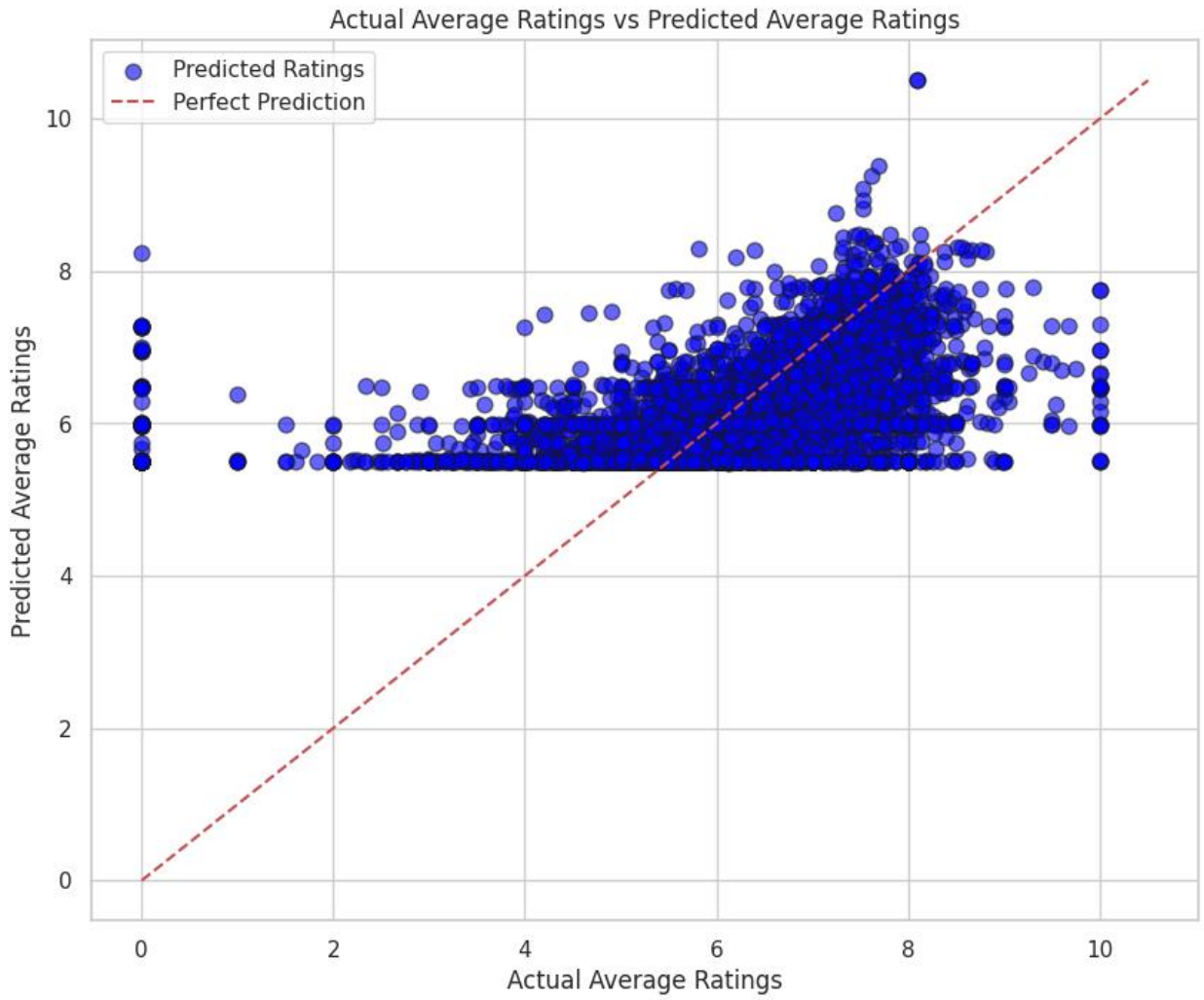


Figure 16. Actual and Predicted Average Ratings

Figure 16 supports the relatively low MAE and RMSE values. Most predicted average ratings cluster around the 6-7 rating point, with some predicted values with large discrepancies.

4. Business Solution and Recommendation

Despite the relatively low error magnitude, the existing model is unsuitable for generating accurate predictions that are valuable for strategic decisions.

For future improvement, this report recommends that

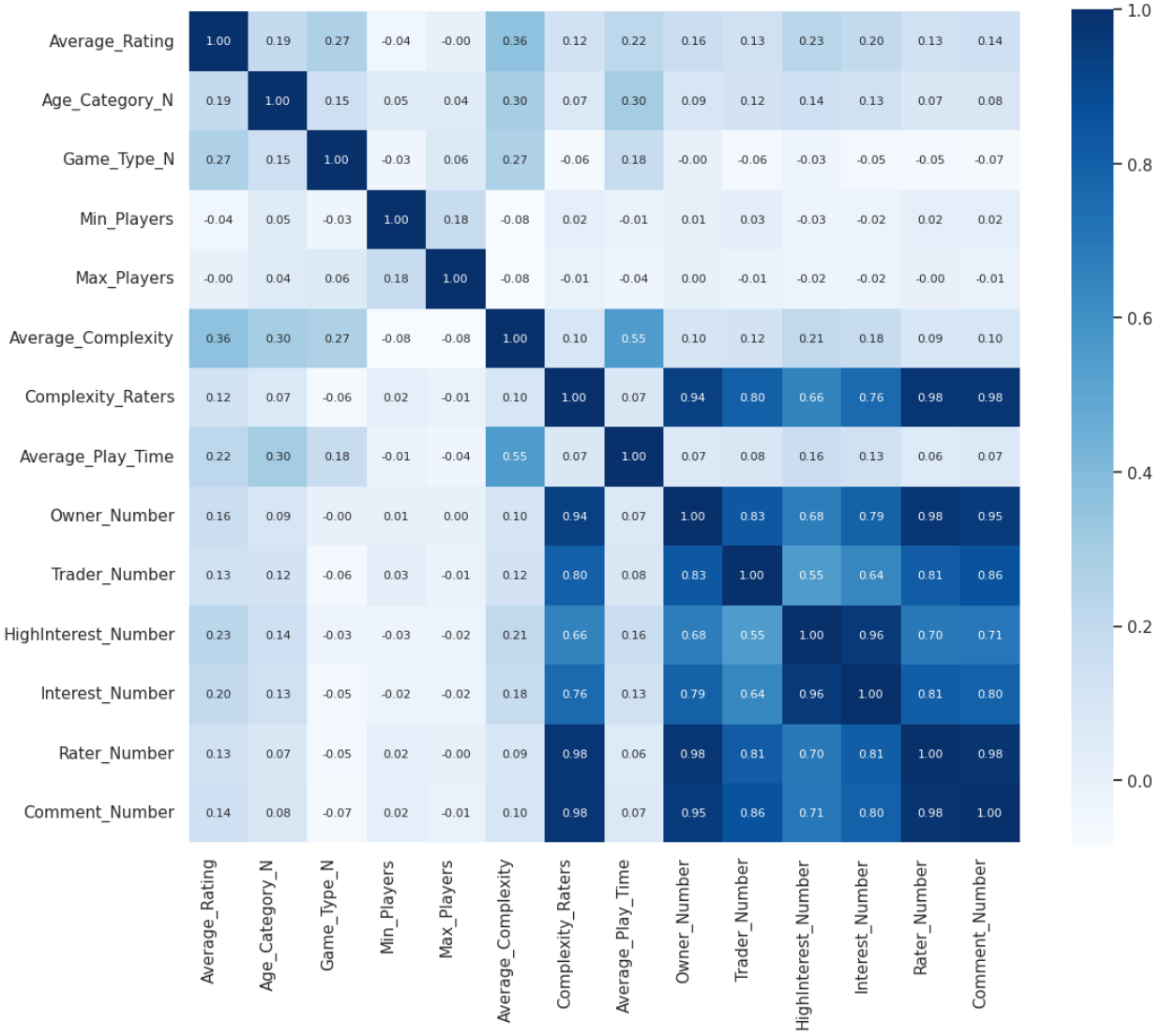
1. Consider adding additional predictors and excluding features with high multicollinearity and weak relationships with average ratings. Additional features, such as game genres, can potentially contribute to predicting the average ratings with higher accuracy. Features such as comment and interest numbers can be excluded from future datasets.
2. This report only employs one predictive model; more advanced techniques like Lasso regression and cross-validation could be investigated.

References

López-Cabarcos, M. Á., Ribeiro-Soriano, D., & Pineiro-Chousa, J. (2020). All that glitters is not gold. The rise of gaming in the COVID-19 pandemic. *Journal of Innovation & Knowledge, 5*(4), 289-296. https://doi.org/10.1016/j.jik.2020.10.004

Modgil, S., Dwivedi, Y. K., Rana, N. P., Gupta, S., & Kamble, S. (2022). Has Covid-19 accelerated opportunities for digital entrepreneurship? An Indian perspective. *Technological Forecasting and Social Change*, 175. https://doi.org/10.1016/j.techfore.2021.121415

Appendix



Appendix 1. Heatmap