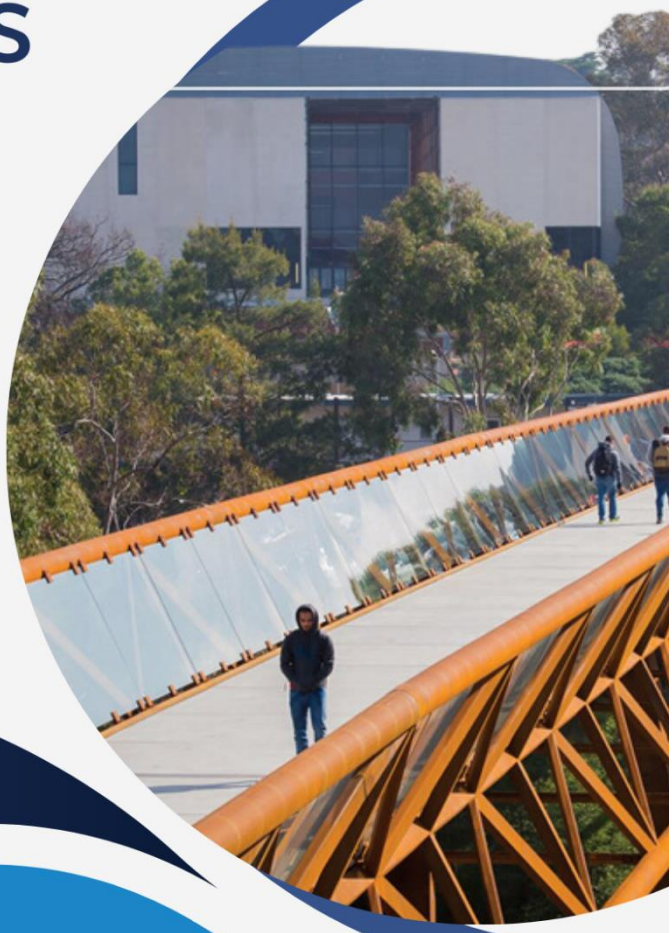
# MIS710 – Machine Learning in Business

# BUSINESS REPORT

## TRIMESTER 1

Presented By
Student name: Huy Hung Pham
Student number
s224212292
Word count: 2121



## Contents



# I. Executive Summary

This project aimed to assist Play Quest Conquer (PQC) in understanding the key factors that influence game ratings, which is essential for optimizing their game development, acquisition, and marketing strategies. We developed a multiple linear regression model to analyse the relationship between game ratings and features such as Average Complexity, High-Interest Number, and Game Type. The model identified these factors as significant predictors of game ratings, although it explained only a portion of the variance.

In order to enhance the model's accuracy, interaction terms should be incorporated, non-linear relationships should be explored, and the PQC team should implement regularization techniques to mitigate overfitting. Continual updates to the model with fresh data will enable PQC to yield more precise game ratings, facilitating informed decision-making and elevating user satisfaction. This model establishes a framework for further enhancement and furnishes actionable insights for PQC's strategic planning.

# II. Business understandings

## 1. Business problem

Play Quest Conquer (PQC), a globally based online gaming platform in Australia, needs help understanding the key factors influencing game ratings. Despite offering various games and maintaining a steady user base, the company needs help to figure out why certain games consistently receive higher ratings. This lack of insight hamper PQC’s ability to strategically develop, acquire, and promote games that match user preferences. As a result, this may lead to poor investments, reduced user engagement, customer dissatisfaction, and a potential decline in revenue and market share.

## 2 Business Analysis Core Concept Model (BACCM)

The PQC company aims to identify the key factors affecting game ratings to make better decisions about developing, acquiring, and promoting games with future potential. Understanding customer preferences and satisfaction in the online gaming can provide a competitive edge and significantly impact business outcomes. The proposed solution involves creating a machine learning model to predict game ratings based on various game characteristics, such as game type, average playtime, and complexity. This model will help PQC to identify the most influential factors affecting ratings and use these insights to make strategic decisions. The value of addressing this challenge lies in optimizing PQC's game offerings to align with user preferences, leading to increased user engagement, higher customer satisfaction, and increased revenue. Additionally, this approach will help PQC allocate resources effectively by focusing on games with more significant market potential.

Implementing this solution represents a shift towards data-driven decision-making, allowing PQC to make strategic decisions that resonate with the needs and preferences of its user base. Key stakeholders include PQC's management team, developers, marketing teams, and users. Management and developers will use the insights to design and select games more likely to receive high ratings. PQC must adapt to technological advancements and user preferences in the fast-paced online gaming industry. Understanding the determinants of game ratings is critical for sustaining success.

# II. Data understanding, preparation, exploration, visualization, and insights gained.

## 2.1 Game information and game configuration

The dataset provided by PQC's market research team contains 24,813 entries across 18 columns, each representing various attributes of the games offered on their platform. The data includes numerical and categorical variables, offering a comprehensive overview of the games, configurations, and user interactions.

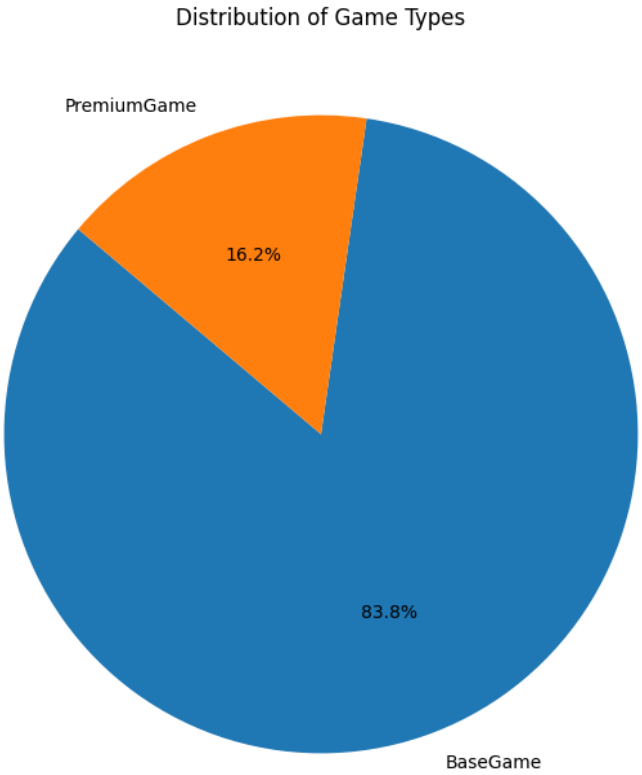



Figure 1: A pie chart to show the distribution of game types (Base game vs Premium games)

According to PQC research, most of the games they distribute are base games (83.8% of the dataset), and the rest of the Premium Games account for 16.2% of the dataset). This distinction is crucial as it might influence the ratings and user engagement metrics.

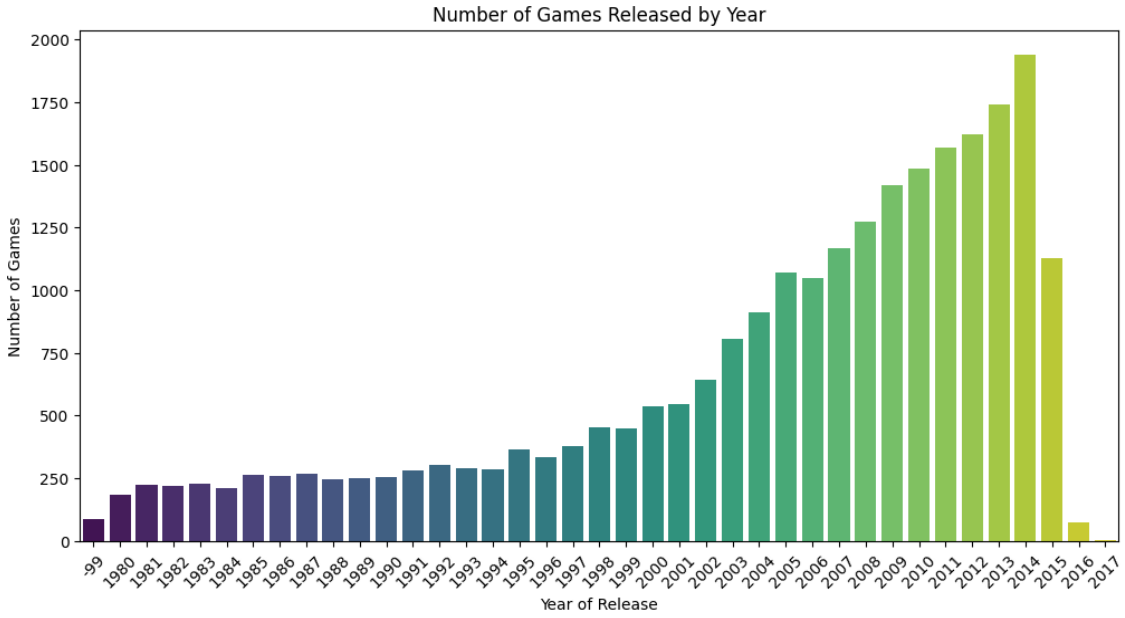



Figure 2: A bar plot displaying the number of games released yearly.

The variable "*Released_Year*" denotes the year of a game's release, encompassing titles dating back to 1980. Analysis of the distribution reveals a heightened concentration of game releases in more recent years, particularly within the 2000 to 2014 timeframe.

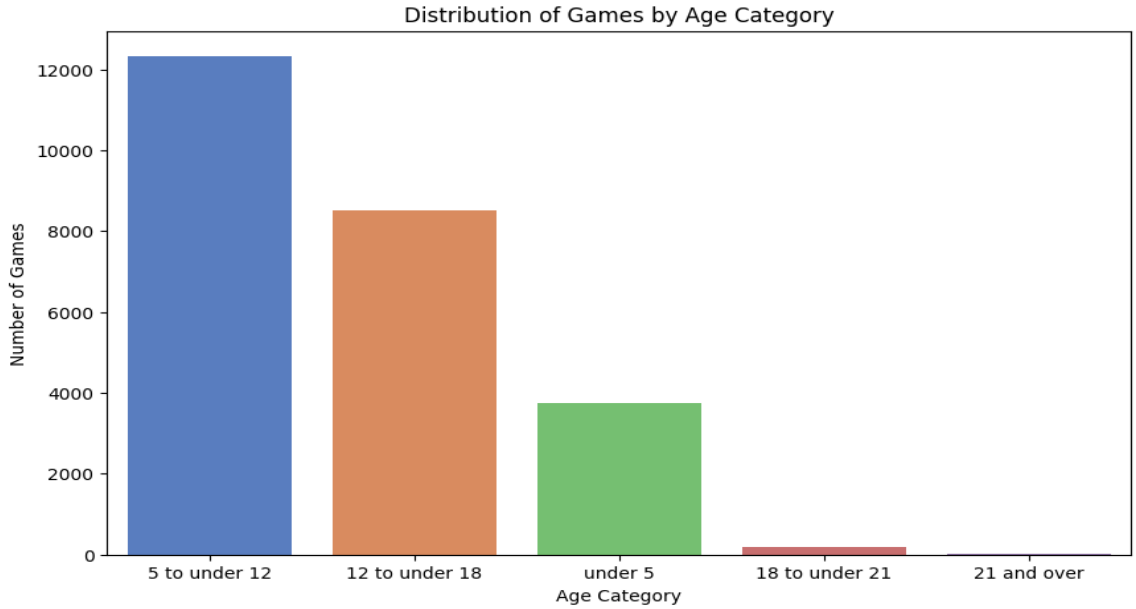



Figure 3: A bar plot to show the distribution of games across different age categories.

The age Category variable describes the target age group for each game, with the majority of games targeted at the "5 to under 12" age group (49.7%) and the "12 to under 18" group (34.3%). There are relatively few games for older audiences, with only 24 entries targeted at "21 and over."

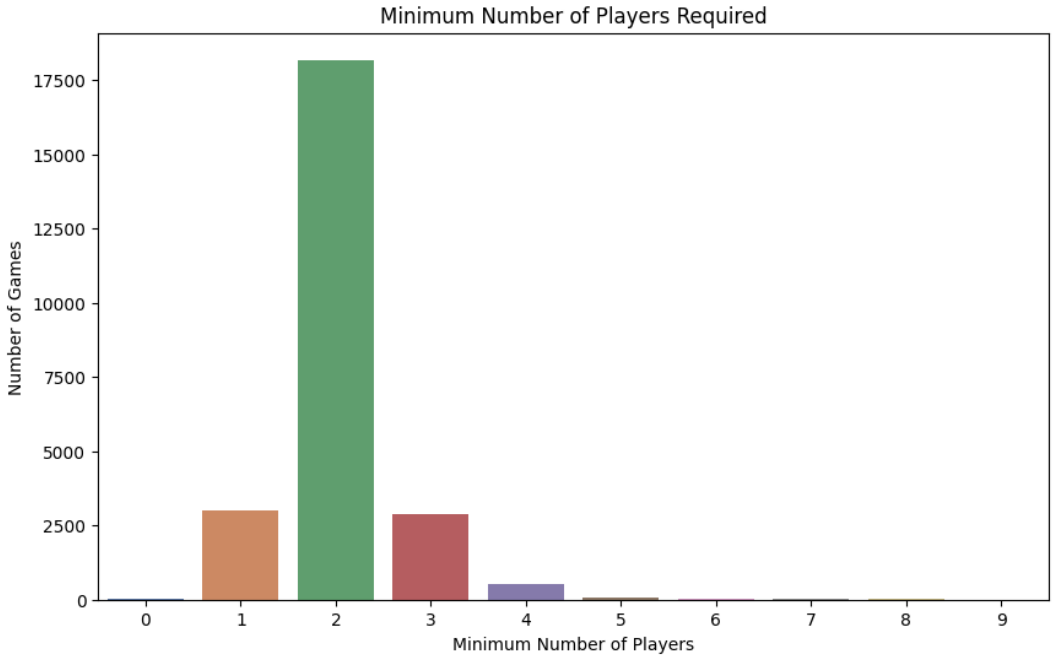



Figure 4: A bar plot showing the distribution of the minimum number of players required for the games.

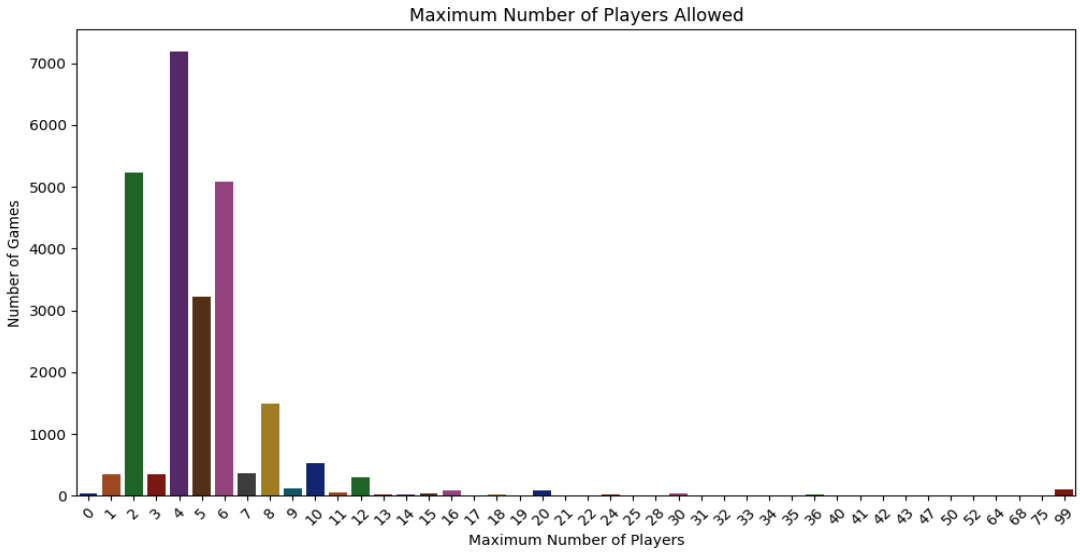



Figure 5: A bar plot showing the distribution of the maximum number of players allowed for the games.

Min Players and Max Players are numerical variables that indicate the minimum and maximum number of players required for the game. Most games accommodate two players as the minimum requirement (73.3%) and up to four players as the maximum (28.9%). Games commonly accommodate 4 to 6 players, with 7,186 games allowing a maximum of 4 players and 5,083 allowing six players. Some games cater to huge groups, with some allowing up to 99 players.

## 2.2 Game Engagement

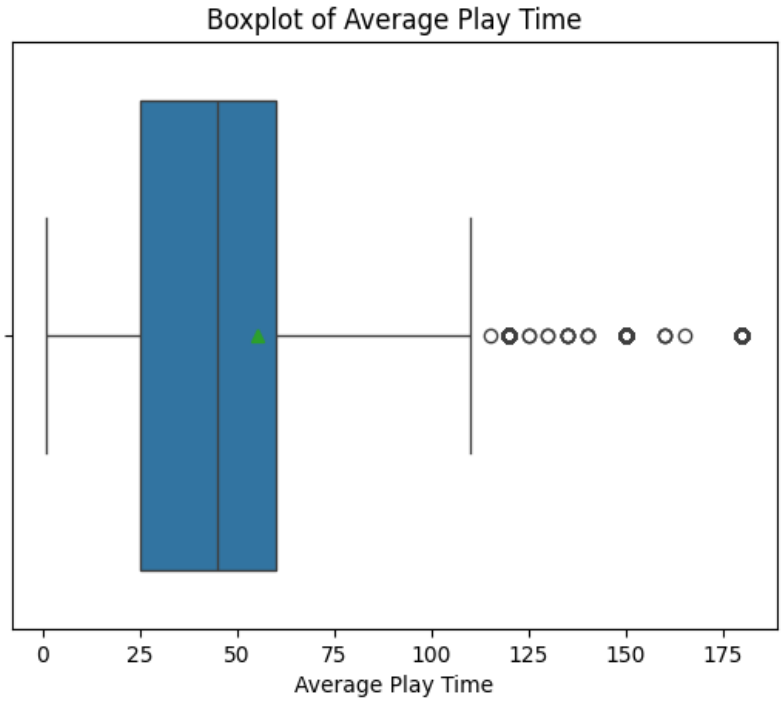



Figure 6: A box plot showing the distribution of average playtime across games

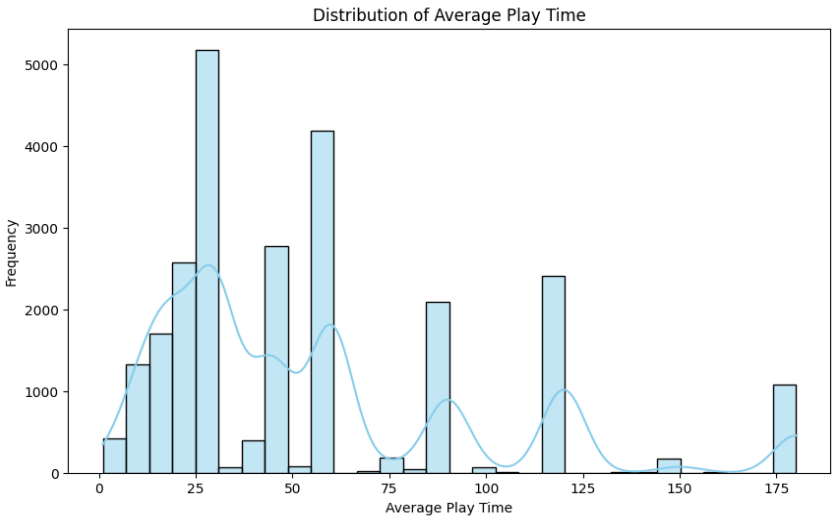



Figure 7: Distribution of average playtime

Most games have an average playtime of about 25 minutes, indicating that many games are designed for short play sessions. However, the variance may be attributed to different types of games with varying complexities and levels of engagement. Additionally, there are games with much longer playtimes, with an additional peak of around 175 minutes, likely representing more complex or highly engaging experiences

## 2.3 Game engagement and rating

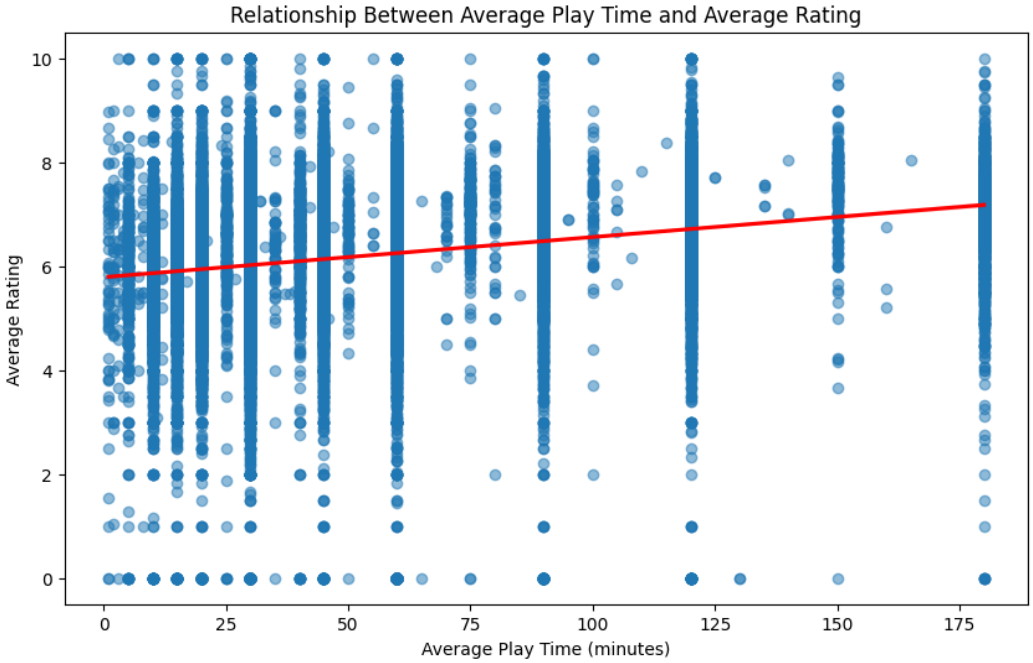



Figure 8: The scatter plot with the regression line visually represents the relationship between Average Play Time and Average Rating

The scatter plot with the regression line indicates a slight positive correlation between average playtime and average rating Games with longer play times tend to receive slightly higher ratings, but the relationship could be stronger. Significant variability in ratings for games with similar play times suggests that playtime alone is not a strong predictor of rating. Outliers exist, representing games that are either exceptionally well-designed or poorly received, independent of playtime.

## 2.4 Game complexity and rating

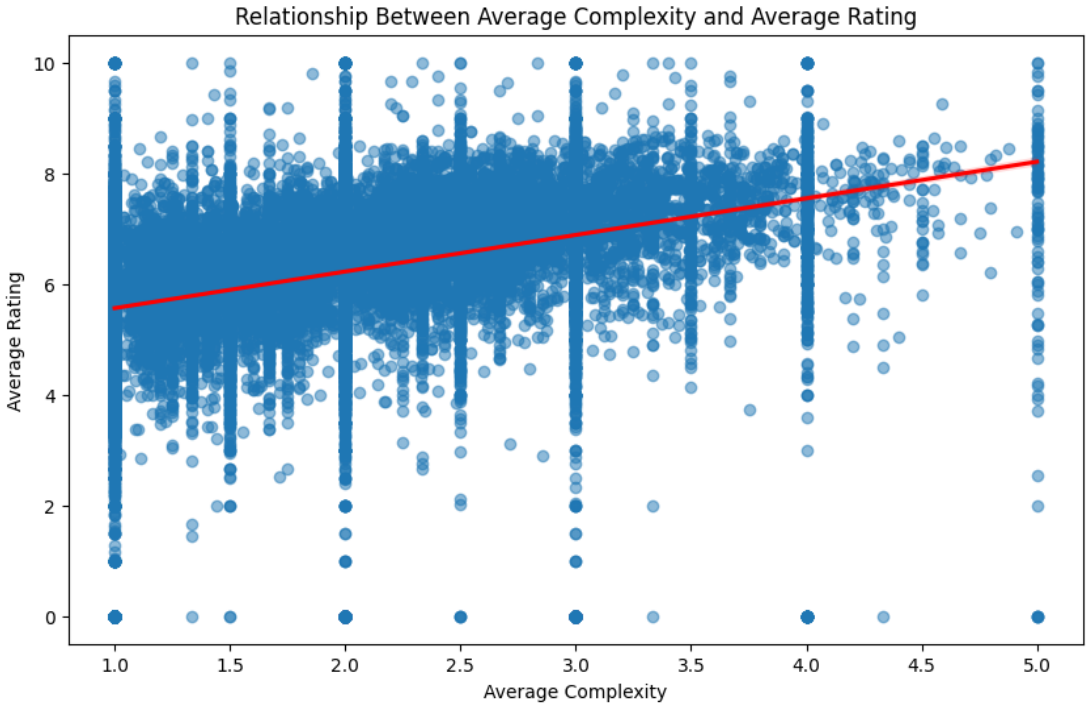



Figure 9: The scatter plot with the regression line visually represents the relationship between Average Complexity and Average Rating

The scatter plot with the regression line shows a positive correlation between average complexity and average rating of games. While higher complexity tends to result in slightly higher ratings, there is significant variability in ratings for games with similar complexity levels. Games with complexity ratings around 3.0 and higher are more likely to have higher average ratings. Outliers, representing unique cases, suggest that factors like game design and user preferences have a more substantial impact on ratings than complexity.

## 2.5 Game configuration, popularity, and Interest

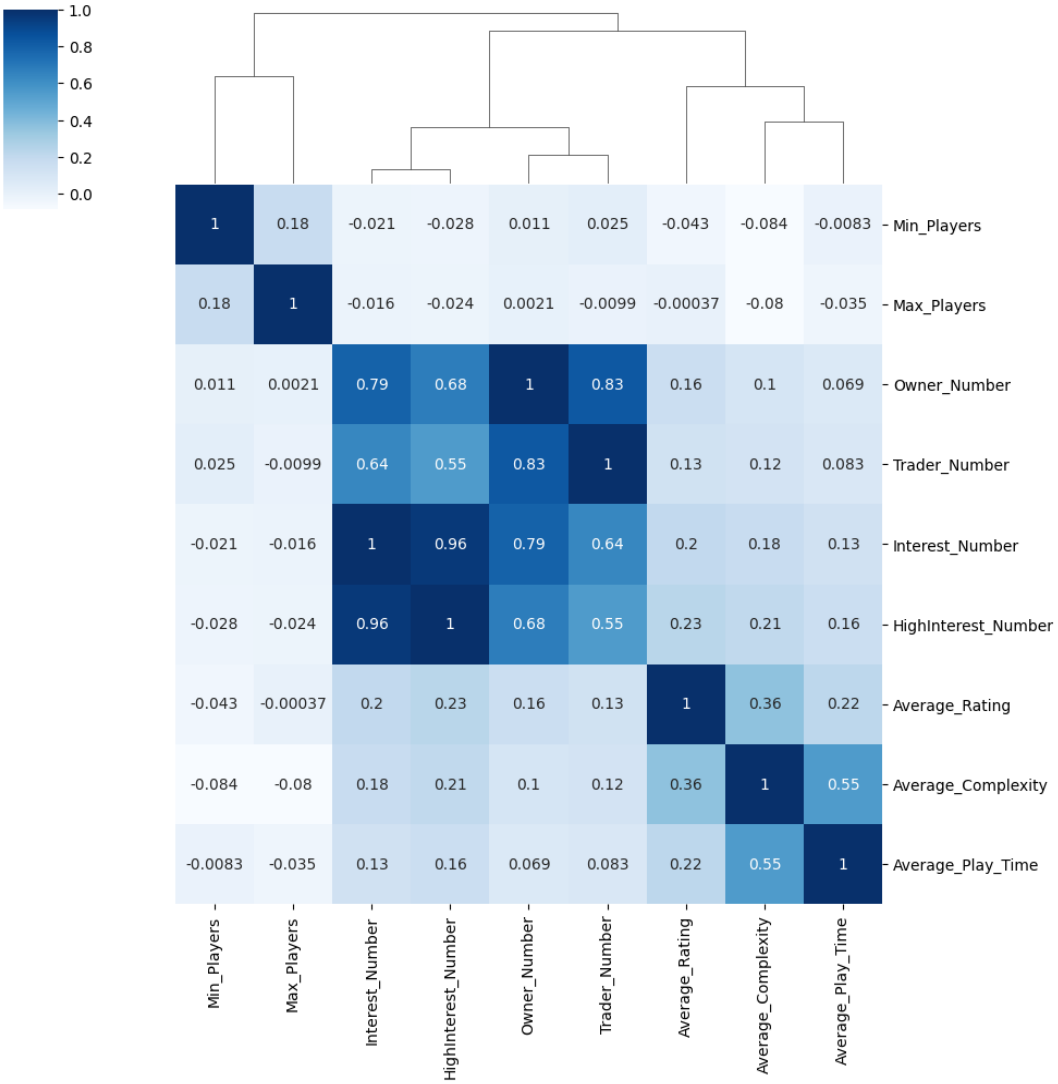



Figure 10: Correlation Matrix of Game Configuration, Popularity, Interest, and Rating

According to the factors considered in the heatmap, a game's complexity has the highest positive correlation with average ratings, indicating that more complex games are generally rated higher. Additionally, high interest in a game shows a moderate positive correlation with average ratings, suggesting that games with higher interest are more likely to receive higher ratings. Similarly, longer play times and general interest in a match also correlate positively with ratings, and the number of owners has a moderate positive correlation with the average rating, indicating a relation between popularity and higher game ratings.

### 2.6 Additional insights on data quality, other variables, and relationships.

The "High_Interest_Number" and "Interest_Number" variables are strongly correlated, suggesting that individuals with a high interest in a game contribute to the overall interest. Similarly, "Owner_Number" and "Trader_Number" are also highly correlated, indicating that widely owned games are frequently traded. Additionally, there is a strong correlation between the number of owners and the level of interest in a game.
The dataset have some missing values, particularly in the *'Game_Name'* column. Outliers in '*Average*_Play_Time' and highly skewed variables like'*High_Interest_Number'* , '*Owner_Number'* and *'Trader_Number'* can affect the performance of machine learning models, particularly linear models. The high correlations between variables can lead to multicollinearity issues in regression models, affecting their performance.

# III. The Machine Learning Approach Undertaken

A multiple linear regression model predicted Play Quest Conquer platform ratings for games. This model helps us understand how factors influence game ratings, providing valuable insights for game development and marketing strategies.

The steps involved in building the model were as follows:

1. *Data Collection and Preprocessing*:
    - The dataset underwent preprocessing to address missing values and ensure all features were formatted appropriately for modeling. Categorical variables like Game_Type and Age_Category were transformed into numerical form using encoding techniques, enabling their use as predictors in the model.
2. *Feature Selection:*
    - Key features were chosen for the model based on their potential impact on game ratings: Average Complexity, High-interest number, Average Play Time, etc.. These features offered a balanced perspective on game characteristics, player interest, and demographic considerations.
3. *Model Training:*
    - The data was split into training and test sets, with 80% of the data used to train the model and 20% reserved for testing.
4. *Model Evaluation:*
    - After training, the model was evaluated on the test set to assess its predictive performance. Key performance metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared ($R^2$), were calculated to gauge the accuracy and explanatory power of the model.

# IV. The Model and Performance Metrics

## 1. Model Evaluation Metrics Interpretation

The R-squared value is relatively low (0.20), which means that the model can explain 20% of the variation in game ratings, while the remaining 80% is due to factors not accounted for by the model.

The MSE (Mean Squared Error) is the average of the squared differences between the predicted and actual values. An MSE score suggests that there is a moderate level of error in the predictions (1.73).

The RMSE (Root Mean Squared Error) is the square root of the MSE and is measured in the same units as the dependent variable (in this case, game ratings). An RMSE result is relatively high (1.32), suggesting that, on average, the model's predictions are far from the actual ratings.

## 2. Feature Coefficients Interpretation

The coefficient for *'Average_Complexity'* is moderate, indicating the relationship of complexity with average game rating. The coefficient for *'Game_Type_Numeric'* is also highly positive, implying that certain game types are rated more favourably than others. Other variables have zero to negative relationships with game ratings, indicating an insignificant impact on the model.

## 3. Interpretation of VIF (Variance Inflation Factor) Results

The VIF values are all below 10, indicating that multicollinearity is not a significant issue in this model. This suggests that the model's coefficients are likely to be stable and reliable. The low VIF values suggest that the model is well-suited for making accurate predictions without the risk of inflated variance due to multicollinearity.

4. Applying regression models to improve the predictive model.

When building and evaluating models to predict Average Ratings, I tried different regressors approaches, starting with a Polynomial regression model, moving on to Lasso regression, and experimenting with Random Forest regression to capture potential non-linear relationships.

| | Training model | Testing model |
|---|---|---|
| MSE | 1.41 | 1.49 |
| MAE | 0.74 | 0.80 |
| R square | 0.396 | 0.288 |

Figure 11: The performance result of the Random Forest regression model.

Out of all the models we looked at, the Random Forest regressor showed the best balance between training and test performance. It had a moderate MSE on the training set and the test set, along with MAE. Although there was a slight drop in the R square metric from the training set to the testing set, which suggests some overfitting, the model still performed better overall compared to other models we evaluated, such as the Lasso and Polynomial regressors.

# V. Discussion of the Pros and Cons of the Model

1. Advantages of the predictive model

The multiple linear regression model is simplicity and interpretability, making it beneficial for stakeholders who need actionable insights. Moreover, the VIF values indicate low multicollinearity among predictors, ensuring reliable coefficients and strengthening interpretability.

2. The disadvantages of the predictive model

The model might only capture some of the underlying patterns due to missing features or its simplicity. Some features have a negligible impact on the model's predictions, indicating their limited relevance or that other factors overshadow their effects. Lastly, the model assumes a linear relationship between predictors and the dependent variable, which could lead to less accurate predictions if the actual relationships are non-linear.

# VI. Business Solutions and Recommendations for Implementation and Improvement

## 1. Business Solution

The data shows that game complexity and playtime have the most significant positive impact on game ratings. Therefore, Play Quest Conquer should prioritize developing and promoting more complex games for users looking for a more profound gaming experience. Furthermore, the slightly positive effect of high-interest games on ratings suggests that targeting marketing efforts toward them could boost engagement and satisfaction among specific user groups.

## 2. Recommendations for Improvement

A. Utilize Nonlinear Models:

Considering the low R-squared values, using nonlinear models like decision trees, random forests, or gradient-boosting models would be beneficial to capture complex relationships between features and game ratings.

B. Feature Engineering:

Enhancing the model's predicting power can be achieved by including additional features. For instance, integrating user engagement metrics, social influence, or marketing variables could offer a more comprehensive view of the factors influencing game ratings.

C. Validate Techniques:

Consider implementing cross-validation during model training to obtain a more robust assessment of the model's performance.

# Conclusion

The multiple linear regression model offers valuable insights into the factors influencing game ratings on the Play Quest Conquer platform. While the model performs reasonably well and generalizes effectively to new data, its limited explanatory power suggests room for improvement. By incorporating more sophisticated models, additional features, and regularization techniques, Play Quest Conquer can enhance the accuracy and usefulness of its predictions, leading to more informed business decisions and ultimately improving user satisfaction.