Project Proposal - Predicting Board Game Average Ratings

Team name - Statistically Significant

```
library(tidyverse)
library(tidymodels)
library(readr)

games <- read_csv("data/games.csv")
themes <- read_csv("data/themes.csv")
mechanics <- read_csv("data/mechanics.csv")
```

Section 1: Introduction

Board games have seen a resurgence in popularity, with thousands of new games published each year. A critical factor influencing the success of a game is its average user rating, which reflects player satisfaction and engagement. This project aims to predict a board game's average rating based on various characteristics, helping game designers and publishers understand what factors contribute to higher-rated games.

- Research Question: Can we predict a board game's average rating using its attributes, such as mechanics, themes, complexity, and user engagement metrics?
- Motivation: Understanding the key factors that influence board game ratings can help game designers create more engaging and well-received games. Board game publishers and designer can use these results and insights to refine what game mechanics they want, target ideal complexity levels, market effectively, and more. Retailers and distributors can use data on what drives high ratings and popularity and can use these decisions to help them decide which games to stock.
- Hypothesis: We expect that a combination of game complexity, player engagement (measured through ownership and ratings), and thematic/mechanical elements will be strong predictors of a game's average rating. Games with deeper mechanics, strong community engagement, and strategic depth are likely to receive higher ratings.

Section 2: Data Description

- Source: 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.
- Data Collection: 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. The data curator pulled this board game data from the API three years ago, and the dataset contains data of board games released as recently as 2021.
- Data Desciption: The curator separated their pulled data into nine files. We will only use 3. Games is a general information table with the most information about the game. Themes and mechanics tables containing the themes and mechanics for each game through indicator variables.

Observations and Characteristics:

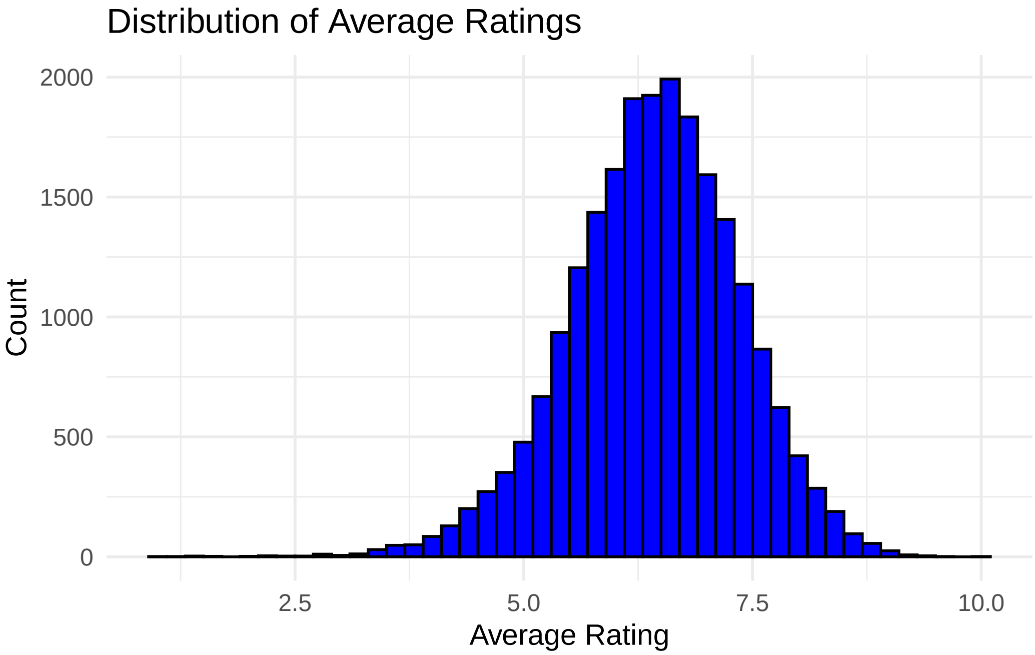
The dataset contains 21,925 board games with games no more recent than 2021. We intend to use the games, themes, and mechanics tables, as we feel that these tables contain the information that would best predict rating. `games.csv` consists of 47 features, describing aspects such as player count, recommended age, user ratings, category, playtime, and rankings. `themes.csv` contains 217 different indicator variables about the different themes of the game, such as adventure, science fiction, medical, etc. `Mechanics` contains 157 different indicator variables about the different ways the games play, such as alliances, dice rolling etc.

Section 3: Data Processing

- Themes & Mechanics Processing: join binary theme/mechanic flags with id on `games`.
- Categorical Encoding: Convert the Cat: binary variables into a single categorical feature.
- Dropping Unnecessary Columns
- Handling Missing Data and Outliers by removing or imputing missing values in critical variables such as ratings and user engagement metrics.

Preliminary Exploration:

```r
ggplot(games, aes(x = AvgRating)) +
  geom_histogram(binwidth = 0.2, fill = "blue", color = "black") +
  labs(title = "Distribution of Average Ratings",
       x = "Average Rating",
       y = "Count") +
  theme_minimal()
```



```
summary(games$AvgRating)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.041   5.837   6.454   6.425   7.052   9.914 
```

The distribution of average board game ratings is shown in the histogram. The ratings are approximately normally distributed, with most games receiving a rating between 5 and 7. The peak of the distribution occurs around 6 to 7, indicating that the majority of games tend to have above-average ratings.

The minimum rating is 1.041, while the maximum rating is 9.914. The median rating is 6.454, which is close to the mean of 6.425, suggesting a fairly symmetric distribution. The first quartile (Q1) is 5.837, meaning that 25% of games have a rating below this value. The third quartile (Q3) is 7.052, meaning that 75% of games have a rating below this value.

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.

Section 4: Analysis Approach

Potential Predictor Variables of Interest:

- NumUserRated: Number of users who have rated the game which is a proxy for popularity.
- NumOwned: Number of users who own the game which can be used as another engagement metric.
- NumWants / NumWishes: Measures of user interest in having the game.
- MinPlayers / MaxPlayers: Range of players the game supports
- MfgPlaytime: Manufacturer-stated play time, which can affect a game's accessibility or appeal.
- Complexity: An indicator of how difficult or deep the game is.
- Theme: Encoded thematic categories (e.g., fantasy, sci-fi, abstract).
- Mechanic: Encoded mechanics categories (e.g., deck-building, worker placement, cooperative).

Modeling Strategy:

Our response variable AvgRating is continuous so we will employ a multiple linear regression model. This method will allow us to estimate how various predictors will influence a game's average rating overall. To build the model we will first fit a baseline multiple linear regression model using the predictors we've mentioned above. The next step will be to look for multicollinearity by figuring out if there are any correlated predictors and if so we will consider removing or combining correlated variables. After several tweaks, when we have the finalized version of our model, we will interpret the regression coefficients to understand how each variable contributes to changes in AvgRating. This will be done by looking at evaultion metrics. Some of these will be:

- R-squared: Measures how well the model explains variation in AvgRating.
- Root Mean Squared Error (RMSE): Evaluates prediction accuracy.
- Feature Importance Analysis: Determines which variables contribute most to predicting game ratings.

Data dictionary

The data dictionary can be found here.