Project 1

Marques Chacon

2/1/2020

Characteristics of Board Games Based on the Interaction of Multiple Variables

I have a natural curiosity for identifying trends and inherent characteristics, and an analysis of board games has led me to consider some interesting questions. Which variables remain constant over a period of time? Can we conclude that certain variables like average user rating of a certain game is affected at least partially by playing time? These types of cross-examinations help me understand the interactions between various factors and determine whether or not there is an inherent trend or dependency between different variables.

For example, let's analyze a particular trend – the number of board games released each year.

We can see from this graph that the number of board games from the website increases exponentially as the year increases. This raises some very interesting questions – does BoardGameGeek record all board games that have been released in a given year or is there a bias towards newly-released games (perhaps games that have been released since the website was created)? If the former is true, then the data suggests that board games are in the midst of a golden age as it stands. However, if the latter is true, then I would expect games to have reached a plateau after a certain point, since the website was founded in 2000, and it eventually would have been able to track any newly-released game. Thus, we can say with relative certainty that the exponential increase in the number of board games is due to an actual increase in the number of games released into the market.

We can further analyze this trend by stratifying the board games into distinct categories. For the sake of simplification, I took the 6 most recorded categories from the BoardGameGeek database and used it to perform analyses based off the stratified data.

The next visualization depicts the number of games released per category over 4 selected years, as a snapshot

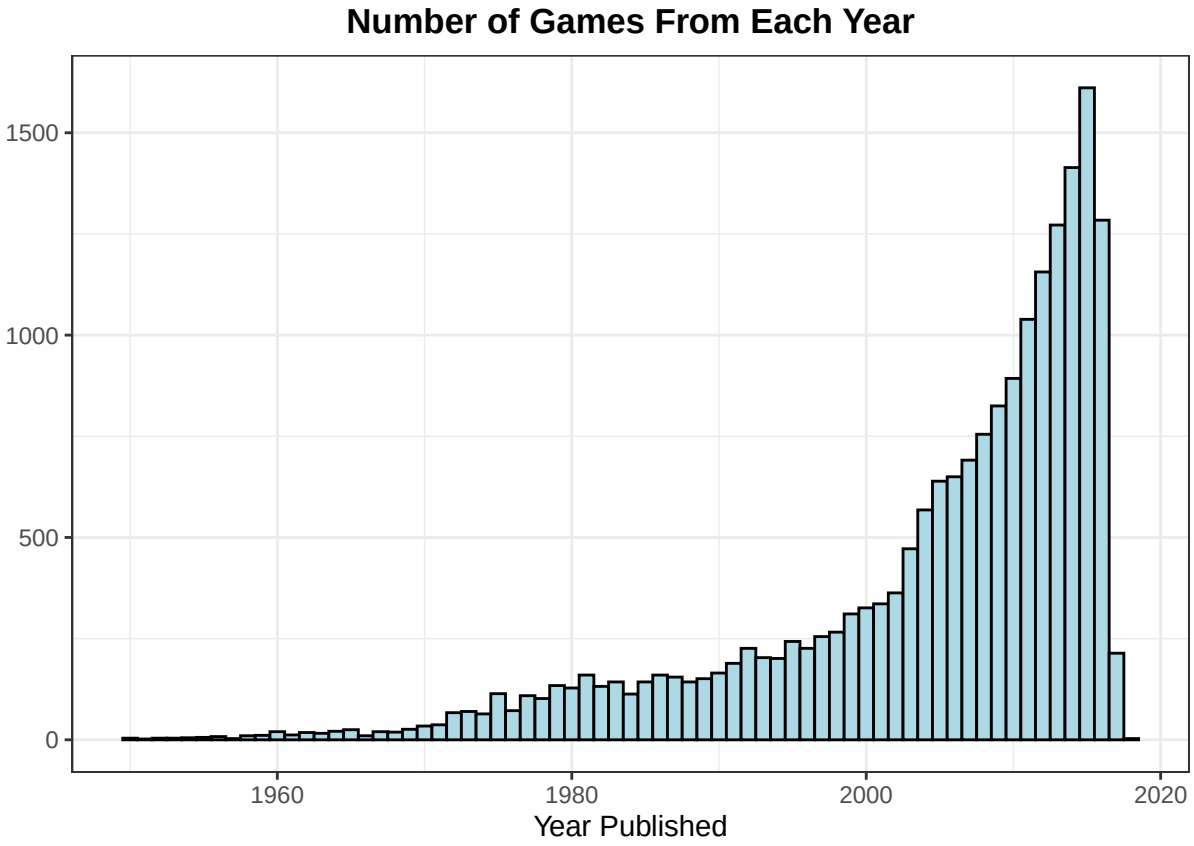
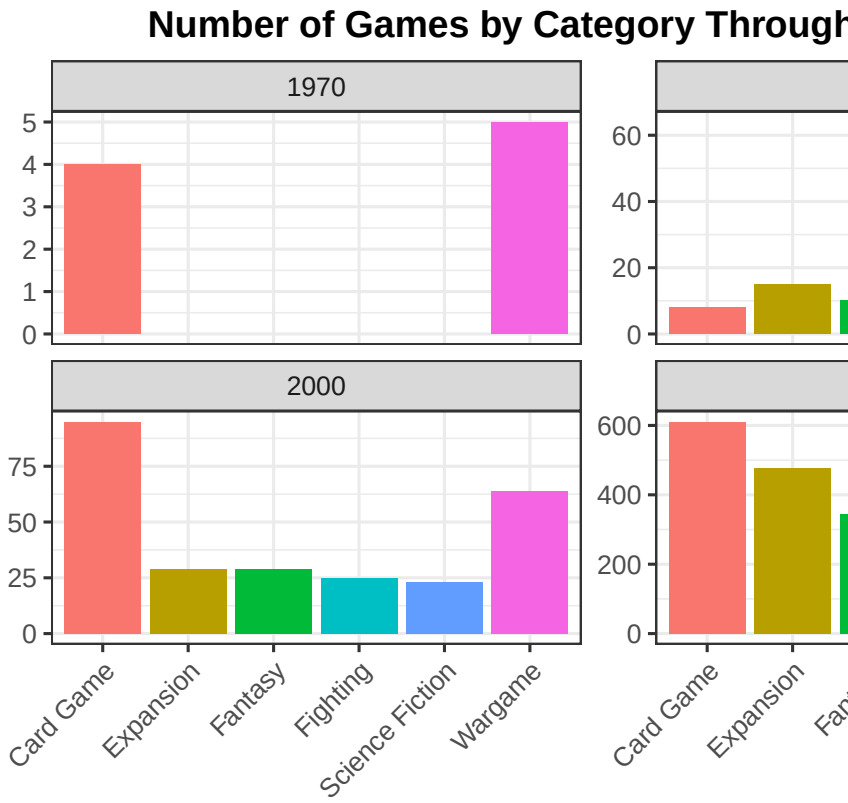


Figure 1: Figure 1: Histogram Showing the Number of Board Games Released Each Year (Each bin corresponds to one year)



into the distribution of games throughout time.

The game data for 1970 is sparse, as it only includes Card Games and Wargames. However, in 1985 there includes data for all six categories and we can make some notable observations. The Wargame genre dominates, having as much games released as the other 5 genres combined. However, in 2000, Card Games leap-frogged every other genre, while Wargames slid into a solid second. Each other genre had around 20-30 games released that year. One thing to note is that the number of games for each genre increased from 1985, with the exception of Wargames. This could signal a turning point for the popularity of Wargames, as it was no longer the most popular genre. In fact, 2015 shows that it became the least popular genre (out of our 6 selected genres). There also became an established hierarchy among the different game genres. For instance, Card Games remained on top, followed by Expansion games, and then Fantasy games. However, it should be noted that the number of games for each genre drastically rose compared to their counts in 2000. This is in line with our first visualization which shows that counts for all games increased exponentially.

Continuing this category-based analysis, we now focus on the average user rating over the years. To gain further insight into this question, I created two visualizations. The first tracks the distribution of the average rating over the years for each category. The second visualization tracks how these ratings trends compare with other categories.

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## `geom_smooth()` using formula 'y ~ x'
```

We notice from the first visualization that the variance in the average ratings of a game for any given year is consistent. Furthermore, it appears that the ratings trendlines for each category goes up. This is, in fact, confirmed in the second visualization. This might allude to the idea that games makers have learned how to appeal to consumers over the years. We also notice that the average rating varies widely by per category.

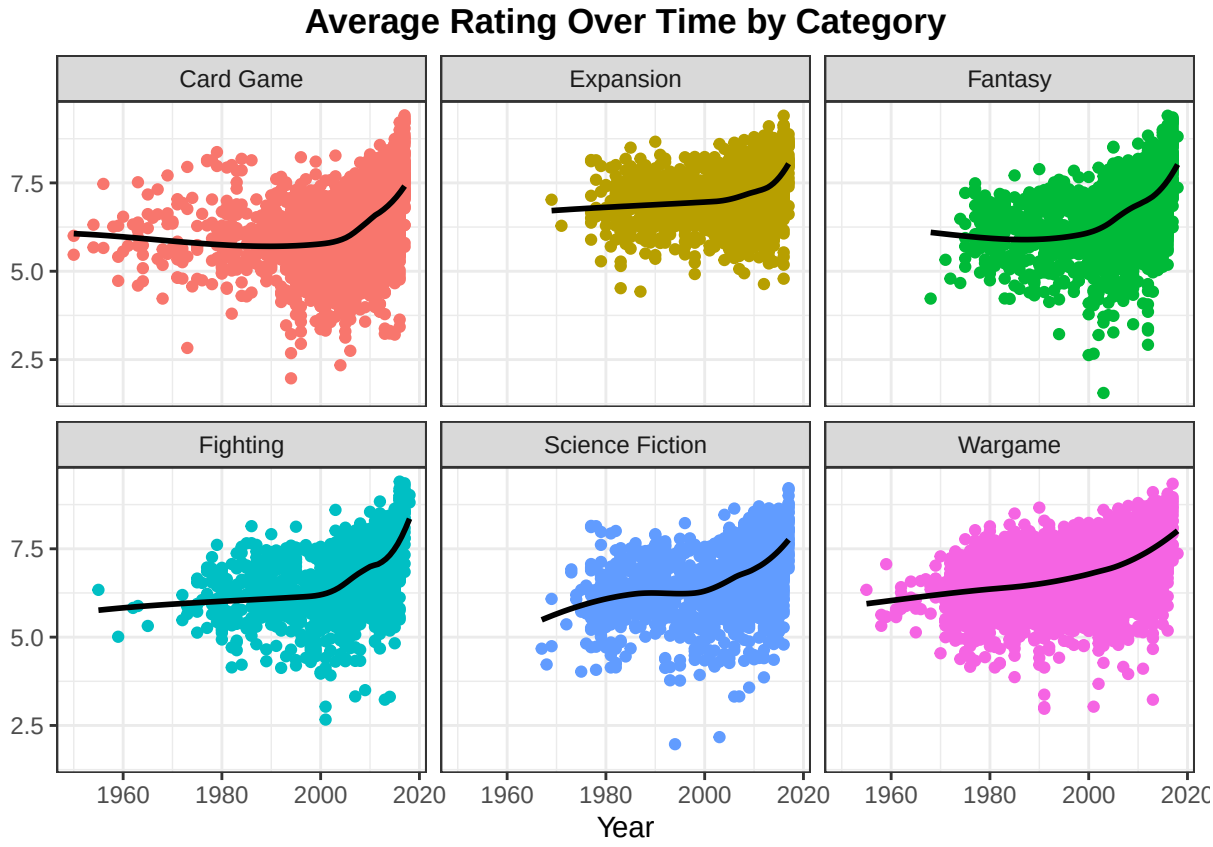


Figure 2: Figure 3: Each Scatterplot Tracks the Average User Rating for a Game, for Each Category. The black lines are trend lines based off the data

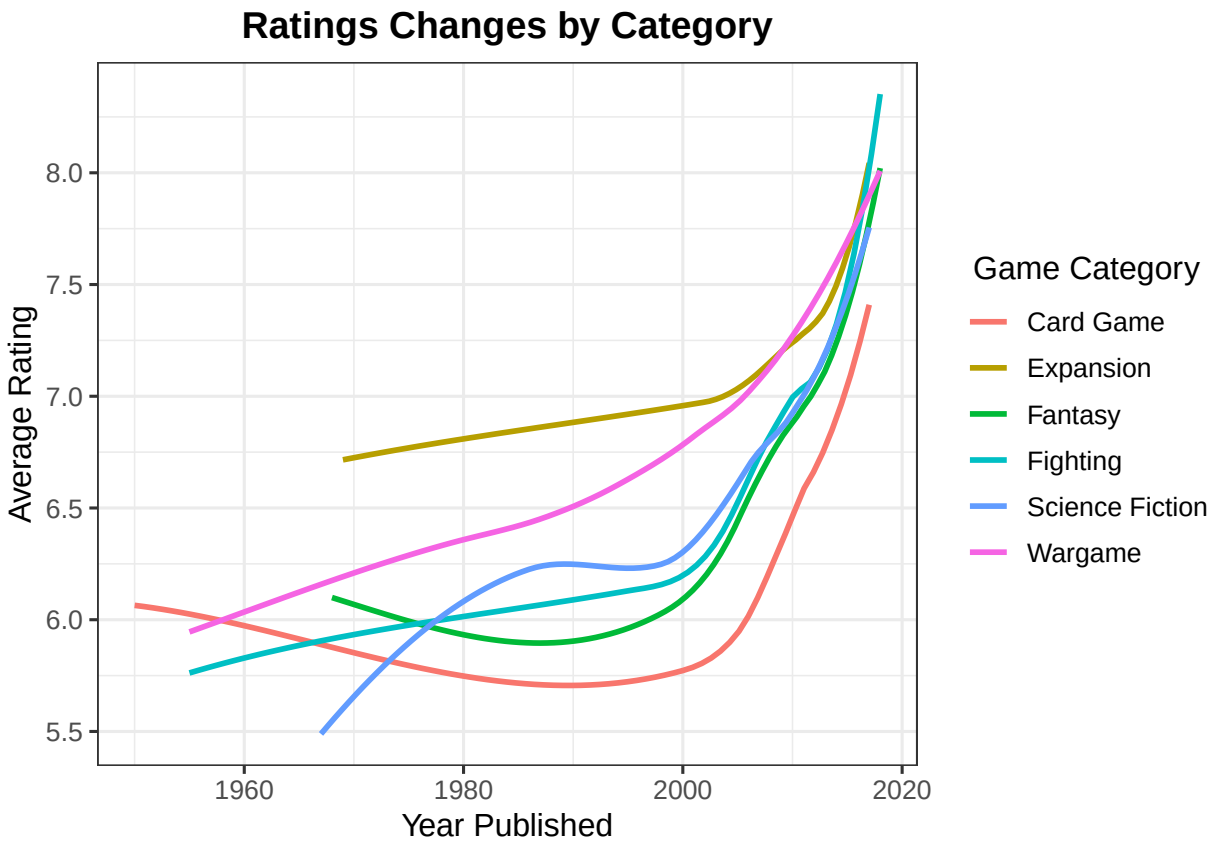
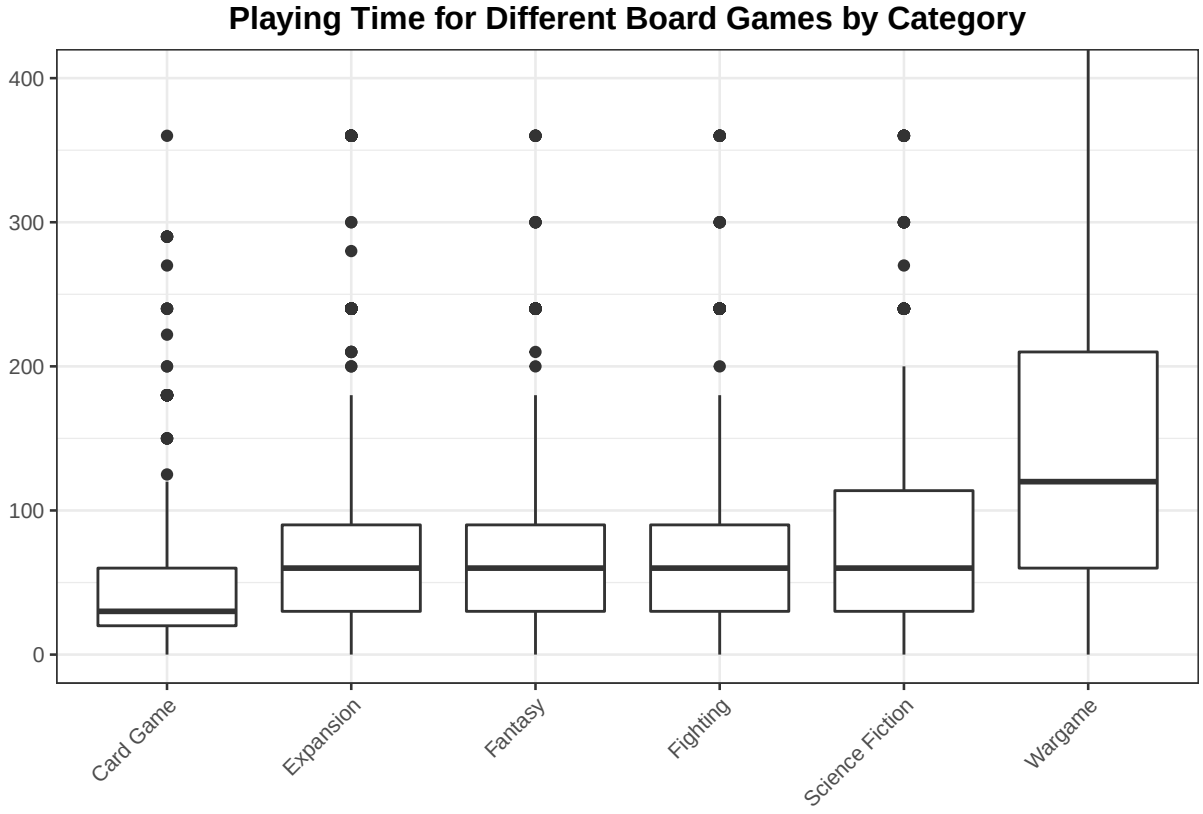


Figure 3: Figure 4: Ratings Trendlines For The Top 6 Game Categories

The line plot shows that Expansion games had the highest average user ratings up until around 2010, where it appears most other categories started to catch up. One other interesting observation is that card games are consistently the lowest rated games, while Fighting games are now the highest rated games. This could reflect a change in consumer habits over the years.

We can further study the categorical data by identifying any inherent characteristics for each game category. In this case, we are interested if the Playing Time significantly varies between categories.



From this data, it appears that Card Games are generally short, which makes sense as most Card Games are relatively simplistic and meant for passing time. Expansion, Fantasy, Fighting, and Science Fiction games are relatively similar with each other, suggesting that there is nothing inherent about these particular categories which affect playing time. On the other hand, Wargames have notably higher playing times. This might be due the fact that Wargames are generally more complex and strategic, requiring patience on the part of the user in order to have success in these games.

Shifting away from the analysis of the different game categories, we now focus on the interaction of various quantitative variables. One type of interaction that interested me was how the playing time would affect the average rating of a game. My hypothesis was that there was a sweet spot with regards to the length of a game that would generate the highest likely rating. The rationale for this was simple – really short games are often too simplistic for the user to enjoy, while really long games would wear out most users. The dotplot below shows a different conclusion:

```
## `geom_smooth()` using formula 'y ~ x'
```

Note that this plot excludes games with a playing time greater than 250 as the majority of the data falls within this range. Although the trend line shows a positive correlation, the association between the two variables is very weak. Thus, there doesn't seem to be any discernible conclusions to make from this graph.

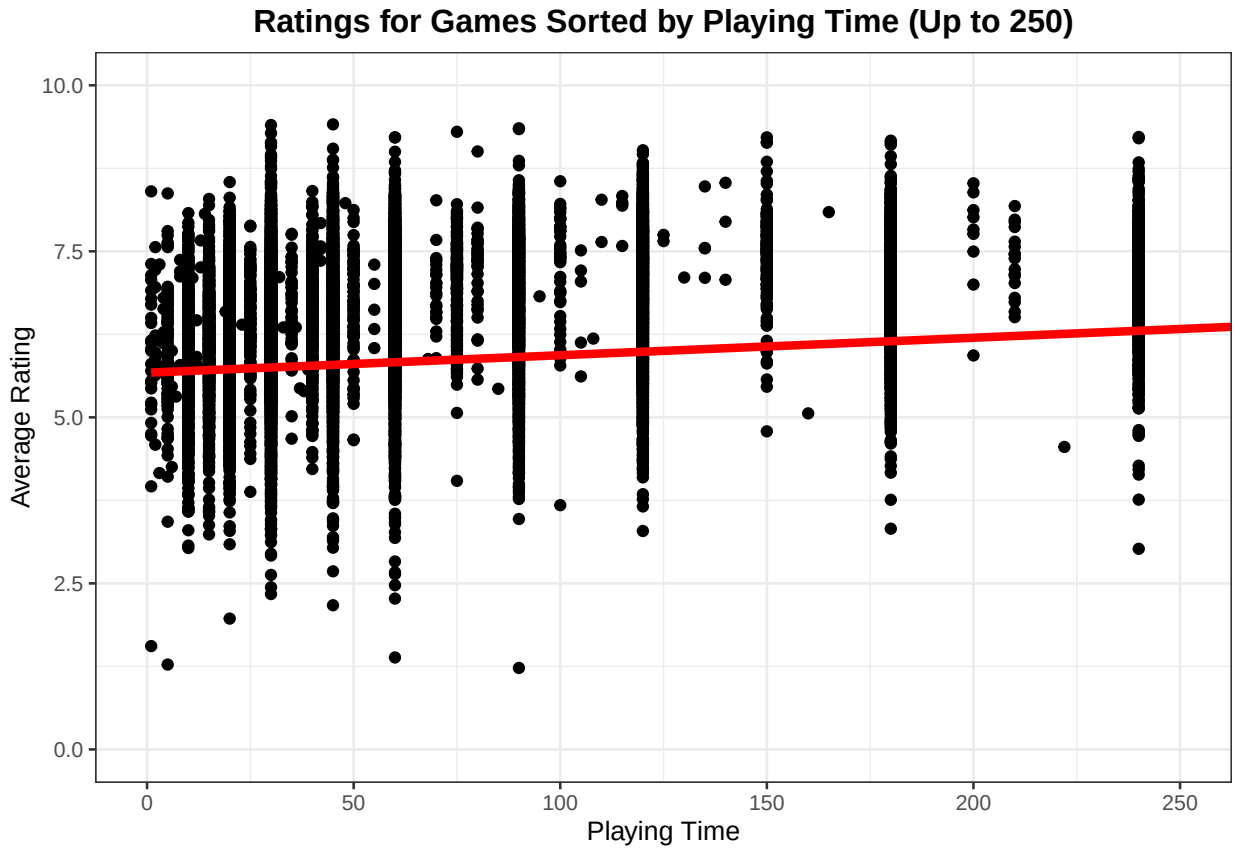


Figure 4: Figure 6: Dotplot Depicting The Association Between Playing Time and Average Rating (Games Listed With a Playing Time greater than 250 were omitted)

This suggests that the average rating of a game is not affected by the playing time. At the very most, playing time has a marginal impact.

We continue our quantitative analysis by studying the relationship between the minimum age necessary to play a game and its complexity. I tackled this question expecting a straightforward answer – surely, the higher the minimum age requirement, the higher the complexity, right? Well according to the scatterplot below, that is true up to a certain point:

```
## `geom_smooth()` using formula 'y ~ x'
```

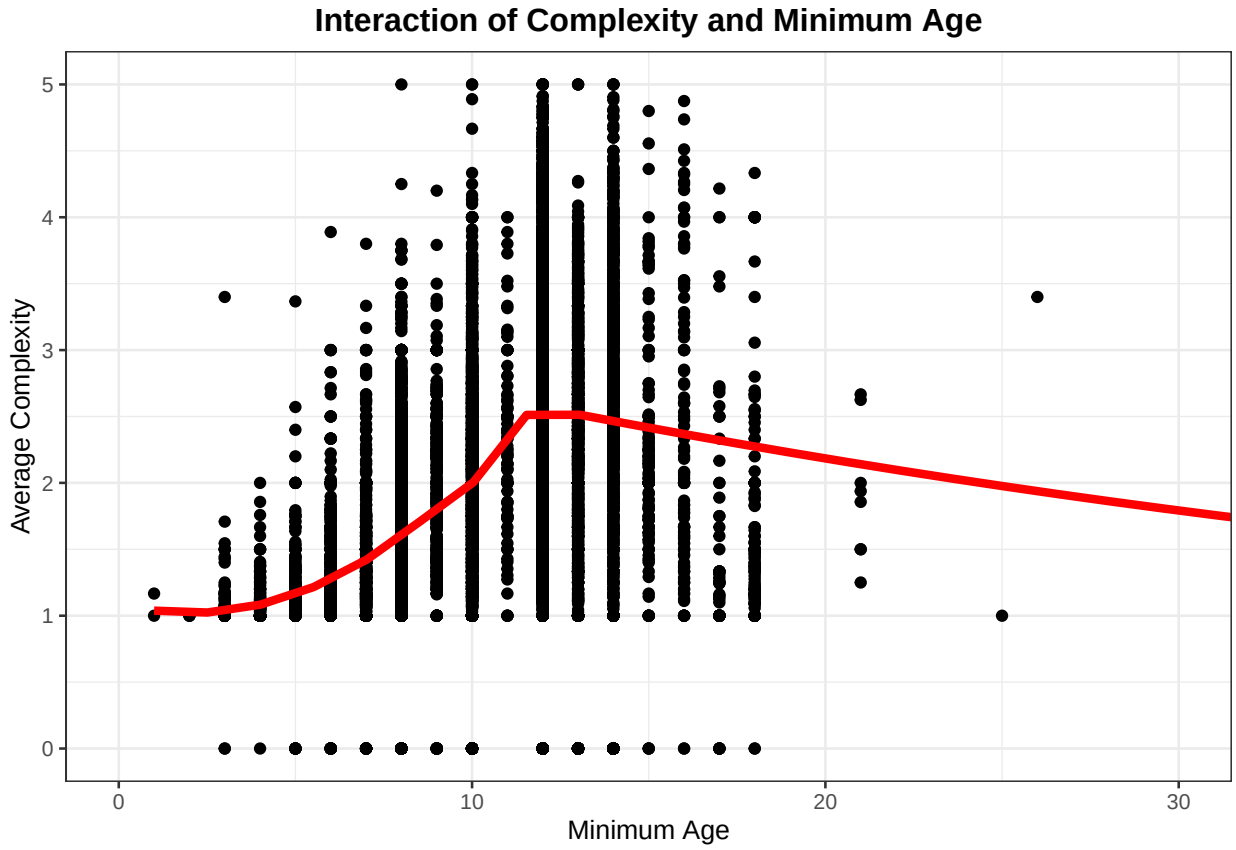
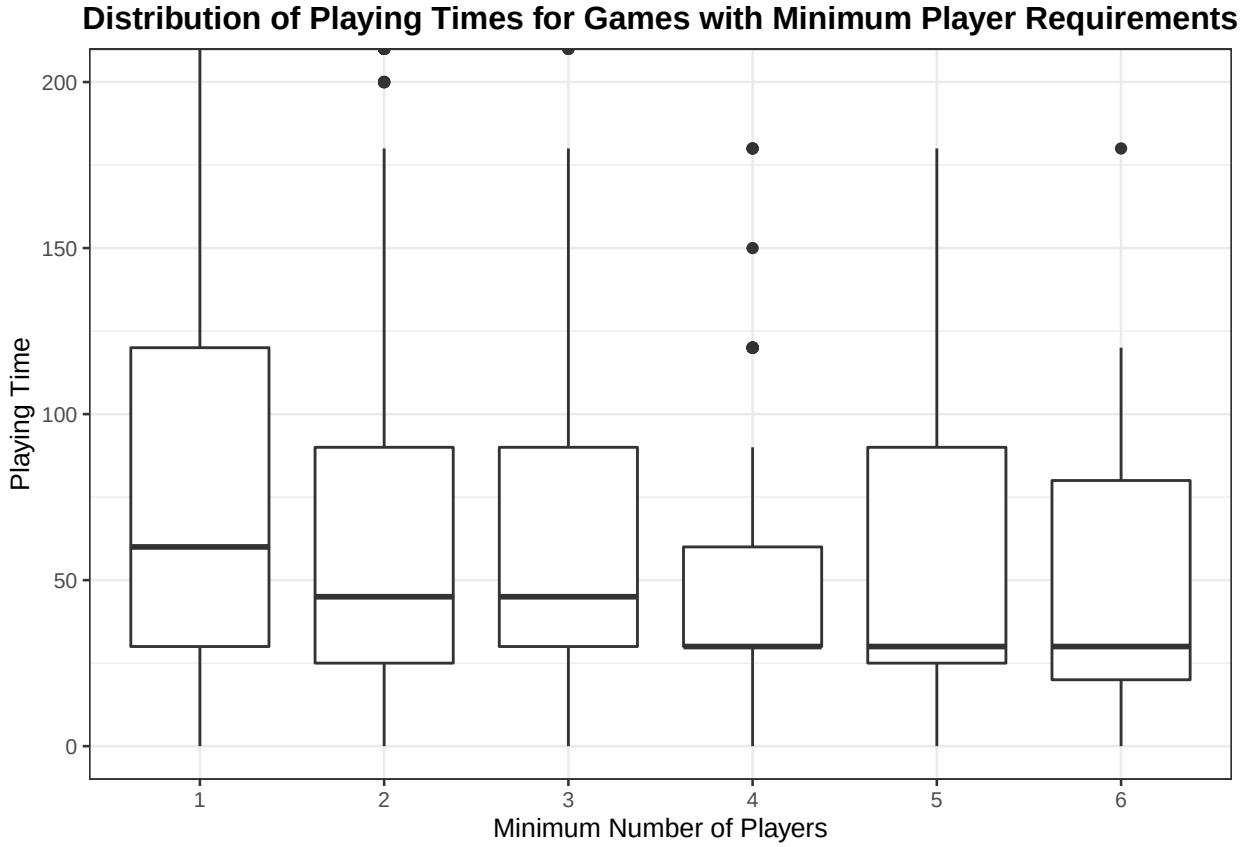


Figure 5: Figure 7: Dotplot Depicting The Association Between Minimum Age and Complexity (Games Listed With a Minimum Age greater than 30 were omitted)

This graph shows that the complexity increases along with age up until the ages of 12 and 13, at least according to the trendline. However, the data suggests that the average complexity is likely to decrease for games with a minimum age higher than that. The explanation for complexity increasing along with age up until age 13 is simple: Games that are more complex require more maturity and patience, while games that are accessible for young children are likely to be simplistic to accommodate their demographic. The explanation for the relative decline in complexity for games with a minimum age higher than 13 is less clear. This could simply be due to a lack of data for games listed with these age requirements. An alternative explanation would be that games with higher age requirements aren't necessarily more complicated, but rather may have elements deemed inappropriate for younger people. This would make sense according to the data.

There was one other question that I was curious to investigate with this data. Does the minimum number of players needed for a game give any indication towards the playing time of a game? My hypothesis was that

with more players needed, a game would take longer, as more players often means more turns needed and more planning, which would affect the playing time. Well according to the boxplot below, the opposite seems to be



the case:

The first thing to note about this graph is that I excluded games with a minimum player requirement of greater than 6 people, as the data becomes much more sporadic when considering games that require more than 6 players (In reality, there were around 50 different values for Minimum Players. However, I chose to focus on games with a minimum player requirement between 1 and 6, which also happened to be the six most common values in our dataset). The graph suggests that the playing time of a game is inversely related to the number of players needed to play the game, since the median playing time for a 1-player game is higher than that of a 2- or 3-player game, which themselves are higher than the median playing time of 4-, 5-, or 6-player games. My hypothesis about this relationship turned out to be the opposite of what the actual conclusion was. The question is, why is that the case? A possible explanation could be that games that require more players are party-oriented, which means that it would be in the best interest of game-makers to have a cap on the playing time (or risk a making a dull party game). Meanwhile, 1-player games allow game-makers more liberty in devising games with longer playing times, as games oriented towards one player are meant to hook said player into the game, and hence require more time commitment.

The various analyses of the board game data highlight some interesting conclusions. We find that the number of games released each year increases exponentially, yet the number of games is not equally distibuted among the top six game categories. We find that the ratings for each game category has fluctuated over time, with the average ratings for each category catching up with each other in recent years. We also find correlations between variables such as minimum age vs. complexity, and minimum players vs. playing time. Ultimately, this analysis has enlightened me of the various trends and inherent characteristics of certain variables for the board game data. Future analyses may focus on the interactions of more variables, or whether the ratings for a certain category is also predicted on the playing time of games from that category.

Appendix

```
# Loads the necessary libraries to perform the data visualizations and manipulations
library(tidyverse)
library(readr)
library(splitstackshape)

# Reads a the board game .csv file and saves it into a data subfolder
data <- read_csv("https://raw.githubusercontent.com/bryandmartin/STAT302/master/docs/Projects/project1/
write.csv(data, "./Data/data_output.csv")

# Filters the dataframe to only include games published in 1950 or later and with at least 25 ratings
modern_and_popular <- data %>%
  filter(yearpublished >= 1950 &
           users_rated >= 25)

# Stores and plots a histogram depicting the number of games per year
num_games <- ggplot(modern_and_popular) +
  geom_histogram(aes(yearpublished),
                 binwidth = 1,
                 color = "black",
                 fill = "lightblue") +
  theme_bw() +
  labs(title = "Number of Games From Each Year",
       x = "Year Published",
       y = "") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
num_games

# Filters the dataframe to only include games published in 1950 or later and with at least 25 ratings
# cSplit tidies the "category" column by denoting a separate row for each category listed (using commas
# Renames the "Expansion for Base-game" category to "Expansion"
modern_and_popular_category_long <- data %>%
  filter(yearpublished >= 1950 &
           users_rated >= 25)  %>%
  cSplit(splitCols = "category",
         direction = "long") %>%
  mutate(category =
           ifelse(category == "Expansion for Base-game",
                  "Expansion", as.character(category)))

# Stores a tibble depicting the 6 most commonly listed categories
cat_count <- modern_and_popular_category_long %>%
  count(category, sort = TRUE) %>%
  head(6)
# Filters the dataframe to only include games tagged under one of the six most commonly listed categori
six_category_limit <- modern_and_popular_category_long %>%
  filter(category %in% cat_count$category)

# Stores and plots six boxplots (one for each category) on one axis, depicting the distribution of play
playing_time_by_category <- ggplot(six_category_limit) +
  geom_boxplot(aes(category, playingtime)) +
  coord_cartesian(ylim = c(0, 400)) +
```

```
  labs(title = "Playing Time for Different Board Games by Category",
       y = "",
       x = "") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1))
playing_time_by_category

# Stores and plots six dotplots (one for each category) on separate axes, depicting the distribution in
year_vs_avg_ratings <- ggplot(six_category_limit) +
  geom_point(aes(yearpublished, average_rating,
                 color = category)) +
  geom_smooth(aes(x = yearpublished,
                  y = average_rating,
                  group = category),
              se = FALSE, size = 1,
              color = "black",
              method = "loess") +
  facet_wrap(~ category) +
  theme_bw() +
  labs(title = "Average Rating Over Time by Category",
       x = "Year", y = "") +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5, face = "bold"))
year_vs_avg_ratings

# Stores and plots a line plot comparing the ratings trendlines for each category (on one axis)
ratings_changes_by_cat <- ggplot(six_category_limit) +
  geom_smooth(aes(x = yearpublished, y = average_rating,
                  group = category, color = category),
              se = FALSE, size = 1,
              method = "loess") +
  theme_bw(base_size = 12) +
  labs(title = "Ratings Changes by Category",
       x = "Year Published", y = "Average Rating",
       color = guide_legend(title = "Game Category")) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        legend.title = element_text(hjust = 0.5))
ratings_changes_by_cat

# Filters the dataframe to only include games published in 1970, 1985, 2000, or 2015
selected_years <- six_category_limit %>%
  filter(yearpublished %in% c(1970, 1985, 2000, 2015))
# Stores and plots four bar graphs (one for each selected year) depicting the number of games published
num_per_cat <- ggplot(selected_years) +
  geom_bar(aes(category,
               fill = category)) +
  facet_wrap(~ yearpublished, scales = "free_y") +
  theme_bw(base_size = 12) +
  labs(title = "Number of Games by Category Throughout the Years",
       x = "", y = "") +
  theme(plot.title = element_text(hjust = 0.5, face ="bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
```

```
        legend.position = "none")

num_per_cat

# Filters for games with an explicit estimated playing time
games_with_play_time <- modern_and_popular %>%
  filter(playingtime > 0)
# Stores and plots a dotplot depicting the correlation between playing time and average rating
playing_time_vs_avg_rating <- ggplot(games_with_play_time) +
  geom_point(aes(playingtime, average_rating)) +
  labs(title = "Ratings for Games Sorted by Playing Time (Up to 250)",
       x = "Playing Time", y = "Average Rating") +
  coord_cartesian(xlim = c(0, 250)) +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
playing_time_vs_avg_rating

# Filters for games with an explicit minimum age requirement
games_with_min_age <- modern_and_popular %>%
  filter(minage > 0)
# Stores and plots a dotplot depicting the correlation between minimum age and complexity
min_age_vs_complex <- ggplot(games_with_min_age) +
  geom_point(aes(minage, average_complexity)) +
  coord_cartesian(xlim = c(0, 25)) +
  theme_bw(base_size = 10) +
  labs(title = "Interaction of Complexity and Minimum Age",
       x = "Minimum Age", y = "Average Complexity") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold")) +
  geom_smooth(aes(minage, average_complexity),
              se = FALSE, size = 1.5,
              color = "red",
              method = "loess")
min_age_vs_complex

# Filters for games with minimum player requirements from 1-6 players
num_player_filter <- modern_and_popular %>%
  filter((minplayers > 0) & (minplayers <= 6)) %>%
  mutate(minplayers = as.character(minplayers))
# Stores and plots 6 box plots (one for each player threshold) on one axis, depicting the distibution o
num_players_and_time <- ggplot(num_player_filter) +
  geom_boxplot(aes(minplayers, playingtime)) +
  coord_cartesian(ylim = c(0, 200)) +
  theme_bw(base_size = 10) +
  labs(title = "Distribution of Playing Times for Games with Player Limits",
       x = "Minimum Number of Players", y = "Playing Time") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
num_players_and_time
```