STAT 302 Homework 2

Board Game Data Exploration

Jaiden Atterbury

Due: 04-23-23 at 11:59 PM

Exploratory Data Analysis of Board Game Data:

Context:

Board games are a greatly forgotten about past time that has been linked with building memory, decision making, problem solving, and overall logical skills in kids, teens, and young adults. Board games also provide a way to decompress from the stressful realities of day-to-day life. In this exploratory data analysis, we will go over trends, features, and certain aspects of board games from the past 75 years through the use of the data visualization package `ggplot2`.

Description of the data:

In particular, we will be exploring and visualizing board game data from boardgamegeek.com. This data set is comprised of $90,400$ different board games and 15 different variables. These variables are:

- `game_id`: This variable represents the board games unique numerical identifier.
- `game_type`: This variable represents the particular type of the board game. The two types of board games include: the normal game (`boardgame`) and an expansion of the normal (`boardgameexpansion`).
- `maxplayers`: The maximum number of players allowed to play the game.
- `minplayers`: The minimum number of players allowed to play the game.
- `maxplaytime`: The maximum playing time of the game (in minutes).
- `minplaytime`: The minimum playing time of the game (in minutes).
- `playingtime`: The average playing time of the game (in minutes).
- `minage`: The minimum age allowed to play the game.
- `name`: The name of the board game.
- `year_published`: The year in which the game was published.
- `category`: The type(s) of game that the board game is. These types include but are not limited to `Economic`, `Card Game`, `Fighting`, etc.
- `mechanic`: The type(s) of ways players interact with the game and each other. These types include but are not limited to `Auction / Bidding`, `Trick-taking`, `Area Control`, etc.

- `users_rated`: The number of users who rated the game.
- `average_rating`: The average rating of the game across all of the users ratings on a 0-10 scale.
- `average_complexity`: The average complexity of the game across all of the users ratings on a 0-5 scale.

Description of analysis:

In this exploratory data analysis, we will primarily answer three main questions/focus on three main topics. These questions/topics include: How do the two different game types compare? How do certain aspects of board games change over time? And lastly, what are the key differences of the top 5 game categories and mechanics?

Most of, but not all, of the below analysis will start on the basis of a simple graph and then proceed to dive deeper into the underlying phenomenon.

Filtering and splitting the data:

In this analysis, we are only interested in games published in 1950 or later with at least 25 ratings, below we will filter the data to meet these requirements:

```
bgame_filtered <- bgame_data %>%
  filter(yearpublished >= 1950 & users_rated >= 25)
```

Now that we have the filtered data, we must notice that the `category` and `mechanic` columns are messy. Thus we will use the `cSplit()` function from the `splitstackshape` package to fix these. We will create a data frame using both `direction = long` and `direction = wide` since they both might be useful, depending on what plot we are generating.

```
# cSplit wide:
wide <- bgame_filtered %>%
  # get rid of "," in category and mechanic
  cSplit(splitCols = c("category", "mechanic"),
         sep = ",",
         direction = "wide"
  )

# cSplit long:
long <- bgame_filtered %>%
  # get rid of "," in category and mechanic
  cSplit(splitCols = c("category", "mechanic"),
         sep = ",",
         direction = "long"
  )
```

Now that we have cleaned up the data to our liking, we can begin with the first exploratory topic: differences between board game types.

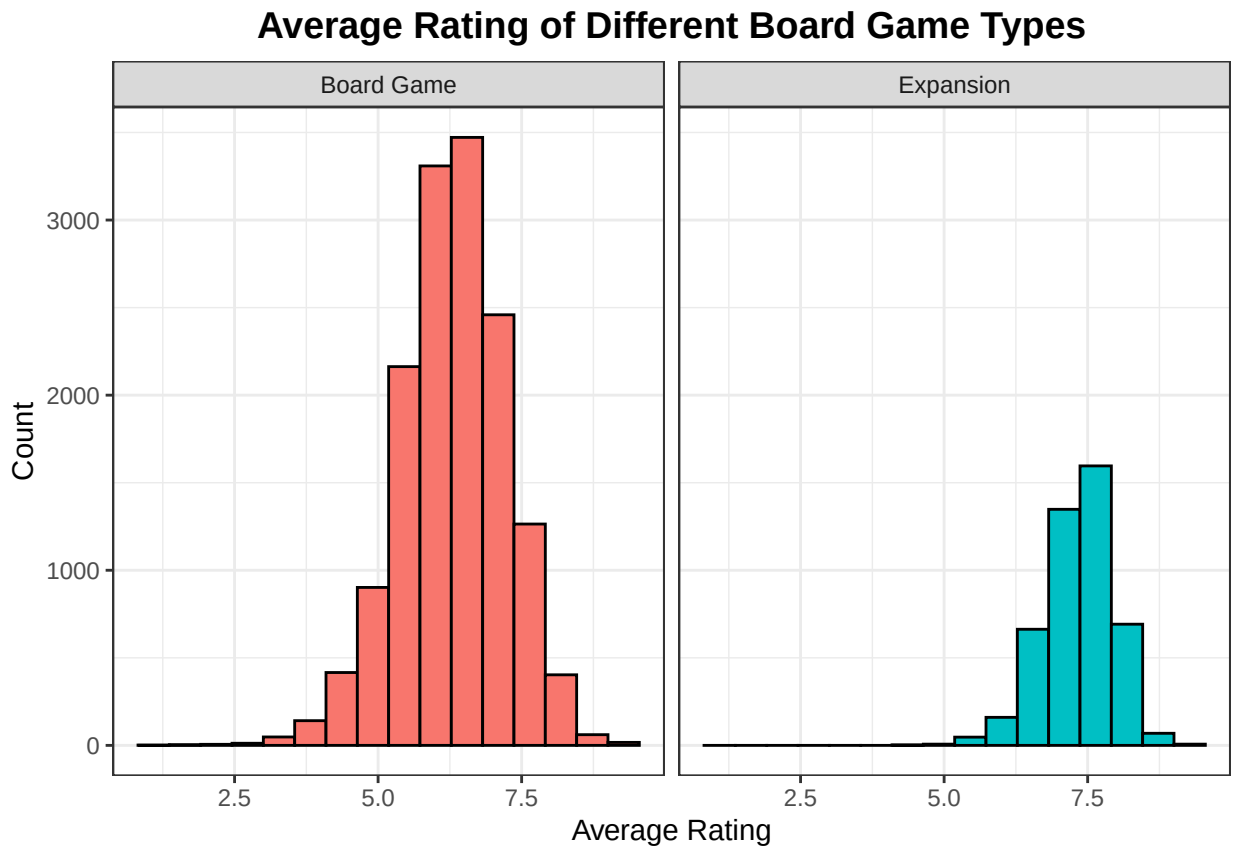
Exploratory topic 1: How do the two types of board games compare?:

Since there are only two types of board games in this datset, we will only focus on simple comparisons between these two types. Namely, we will only focus on comparing the average rating and the average complexity between these two types.

Comparing the average rating between the two types:

```
# Create a new data frame with more legible values:
new_type_labs <- bgame_filtered %>%
  mutate(game_type = recode(game_type, "boardgame" = "Board Game",
                                       "boardgameexpansion" = "Expansion"))

# Create a histograms of average rating of different board game types:
ggplot(data = new_type_labs,
       mapping = aes(x = average_rating, fill = game_type)) +
  geom_histogram(bins = ceiling(log(length(bgame_filtered$game_type), base = 2) + 1),
                 color = "black") +
  facet_wrap(~game_type) +
  labs(x = "Average Rating",
       y = "Count",
       title = "Average Rating of Different Board Game Types") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
        legend.position = "none")
```



Caption:
The pair of histograms above displays the counts on the y-axis for the different average ratings on the x-axis for each board game type (`game_type`). The two types of board games are the original board game (`Board Game`) colored red and expansions for board games (`Expansion`) colored blue.

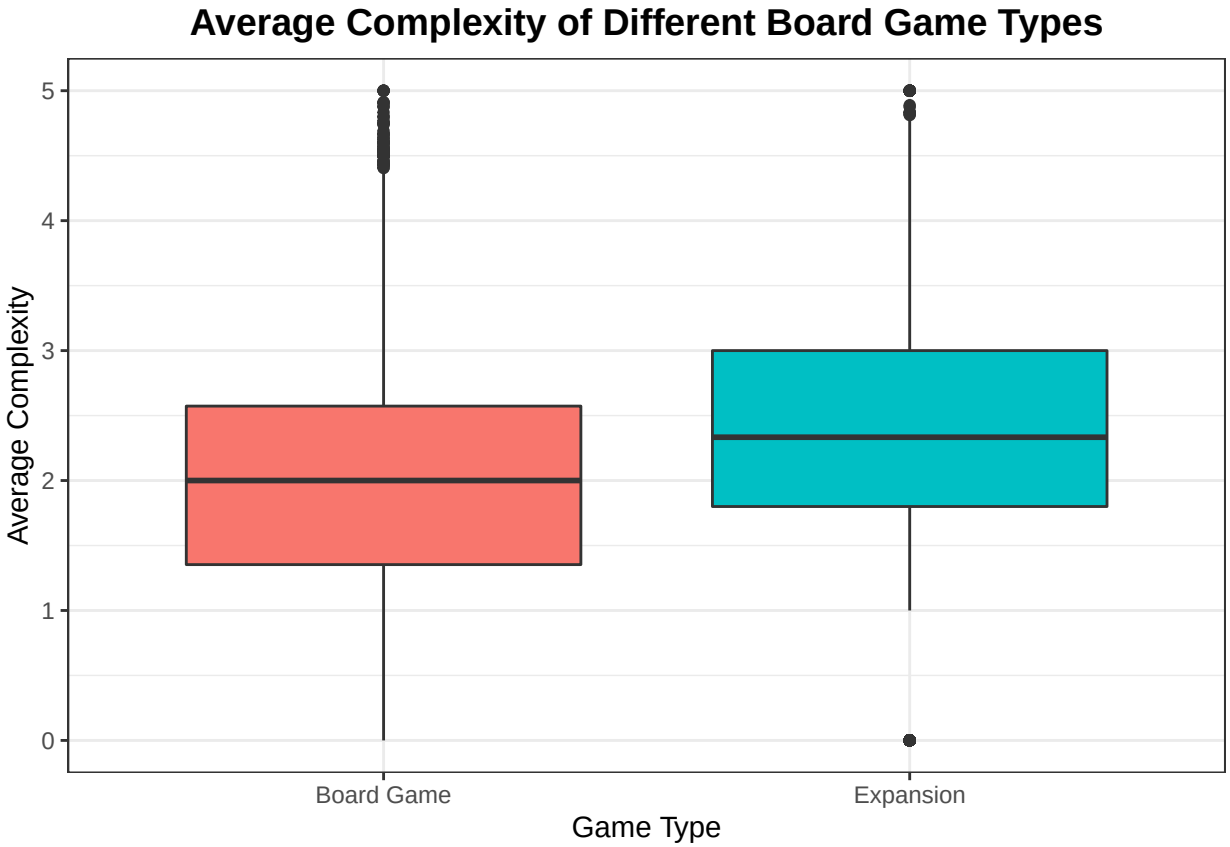
Description:

As can be seen from the above histograms, the board game type called `Board Game` has a slight left skew in its distribution but can be considered approximately normal with its center/mean around an average score of 6.25 with the associated count being around 3500. Also, the board game type called `Expansion` has a slight left skew in its distribution but can be considered approximately normal with its center/mean around an average score of 7.5 with the associated count being around 1500. Both of the histograms have a relatively similar spread.

Insights:
From the above histograms and accompanying description, we can see that there are a lot less expansions than there are actual board games. Furthermore, we can see that, on average, the average rating for expansions are much higher than those for the normal board games. This makes sense intuitively because in order to buy an expansion you must already like the game itself, and expansions are there to enhance the game play of the original game.

Comparing the average complexity between the two types:

```
# Create boxplots of average complexity of different board game types:
ggplot(data = new_type_labs,
       mapping = aes(x = game_type, y = average_complexity, fill = game_type)) +
  geom_boxplot() +
  labs(x = "Game Type",
       y = "Average Complexity",
       title = "Average Complexity of Different Board Game Types") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
        legend.position = "none")
```



Caption:
The pair of boxplots above displays the different average complexities on the y-axis for each board game type (`game_type`) on the x-axis. The two types of board games are the original board game (`Board Game`) colored red and expansions for board games (`Expansion`) colored blue. Each boxplot shows the first, second, and third quartiles of the average complexity data as well as any outliers.

Description:
As can be seen from the above boxplots, the board game type called `Board Game` has a slight right skew in its distribution but can be considered approximately normal with its center/median around an average complexity of 2. Furthermore, the boxplot for the original board games has several high outliers. Also, the board game type called `Expansion` has a very slight right skew in its distribution but can be considered approximately normal with its center/median around an average complexity of 2.4. Furthermore, the boxplot for the board game expansions has a few high outliers, and a single low outlier way below the lower fence. Both boxplots have a similar spread, however the boxplot for the `Board Game` type has more density in its lower tail than that of the `Expansion` board game type.

Insights:
From the above boxplots and accompanying description, we can see that, on average, the average complexity for expansions is slightly higher than those for the normal board games. This makes sense intuitively because expansions are meant to enhance/add more elements to the game, which in essence makes the game more complex and difficult.

Overall conclusions for part 1:

Overall, we can see that board game expansions, on average, have higher ratings and are more complex than the original board games themselves. However, it is important to note that we have not shown that any

significant difference persists between the two, we have merely shown that there is a clear pattern in the data and presented it through visualizations.

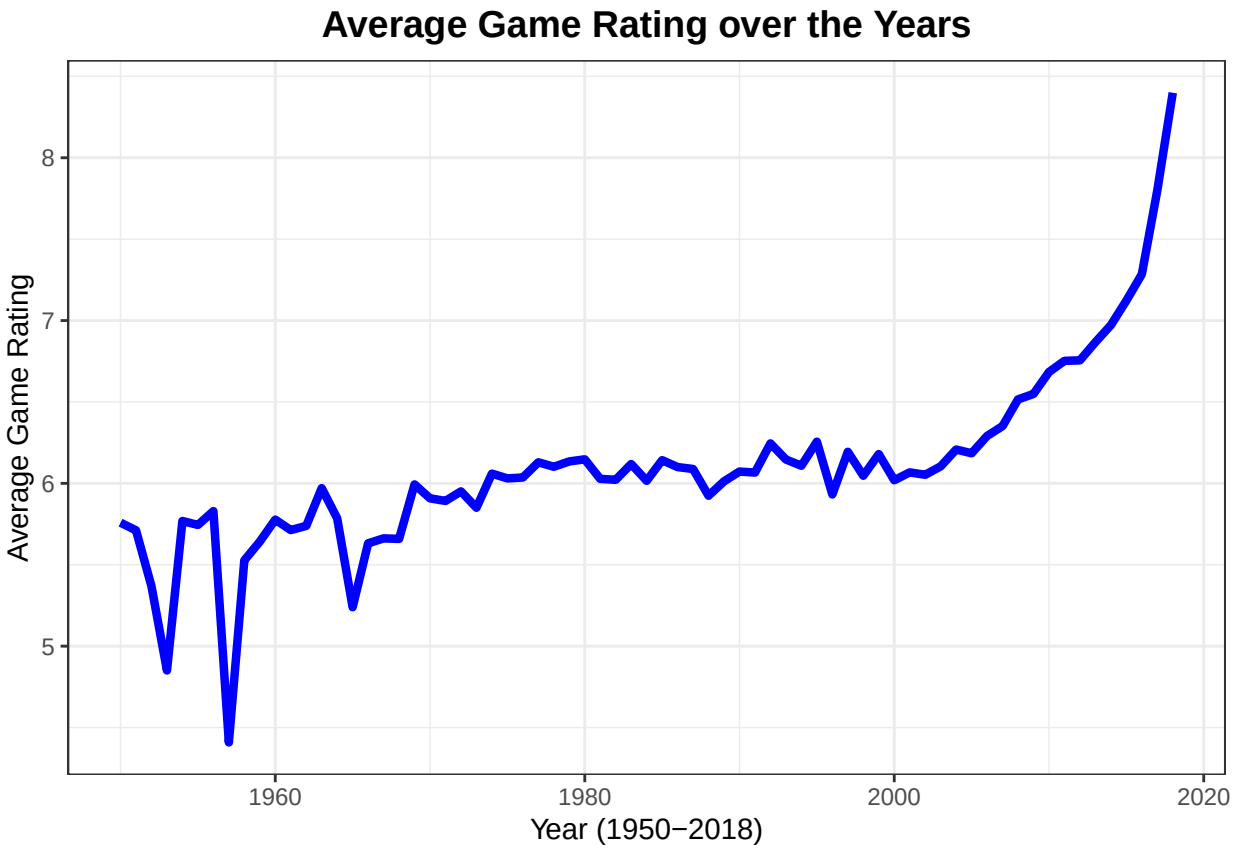
Exploratory topic 2: How have aspects and features of board games changed over time?:

Since the board games that we will focus on in this datset span over 75 years, we will have a lot of flexibility and options to look for trends and insights in the data. In particular we will focus on how the average rating and complexity change over time, break these two features up based on certain categories, and lastly look at how users have interacted with these games over the years.

Average rating of board games over time:

```
# Group the data by year and find the average rating of games in that year:
bgame_year <- bgame_filtered %>%
  group_by(yearpublished) %>%
  summarise(mean = mean(average_rating))

# Create a line chart to show how average rating changes over time:
ggplot(data = bgame_year, mapping = aes(x = yearpublished, y = mean)) +
  geom_line(color = "blue", size = 1.5) +
  labs(x = "Year (1950-2018)",
       y = "Average Game Rating",
       title = "Average Game Rating over the Years") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14))
```



Caption:
The above line chart shows how the average rating of the games released in each year has changed over the past 75 years. In particular, the y-axis shows the mean of the `average_rating` variable for all of the games released in each year from $1950 - 2018$, and the x-axis shows the year represented by the `yearpbulished` variable from $1950 - 2018$.

Insights:
The above line chart shows that despite an up-and-down trend throughout the 1950s-1960s, the average rating of games released has been on the rise since the 1970s. In particular, ever since the dawn of the 21st century, the ratings for games released in each respective year has increased considerably. One thing that could explain this rapid rise of ratings in the 21st century is the sheer amount of technological advances have allowed for more creativity from game creators, as well as more opportunities to create never before seen concepts and mechanics.

In the next visualization, we will see this same data but broken up into different categories. Since there are 85 unique game categories in this dataset, that would be far too many categories to visualize. Thus, we will focus on only the top 5 categories based on frequency.

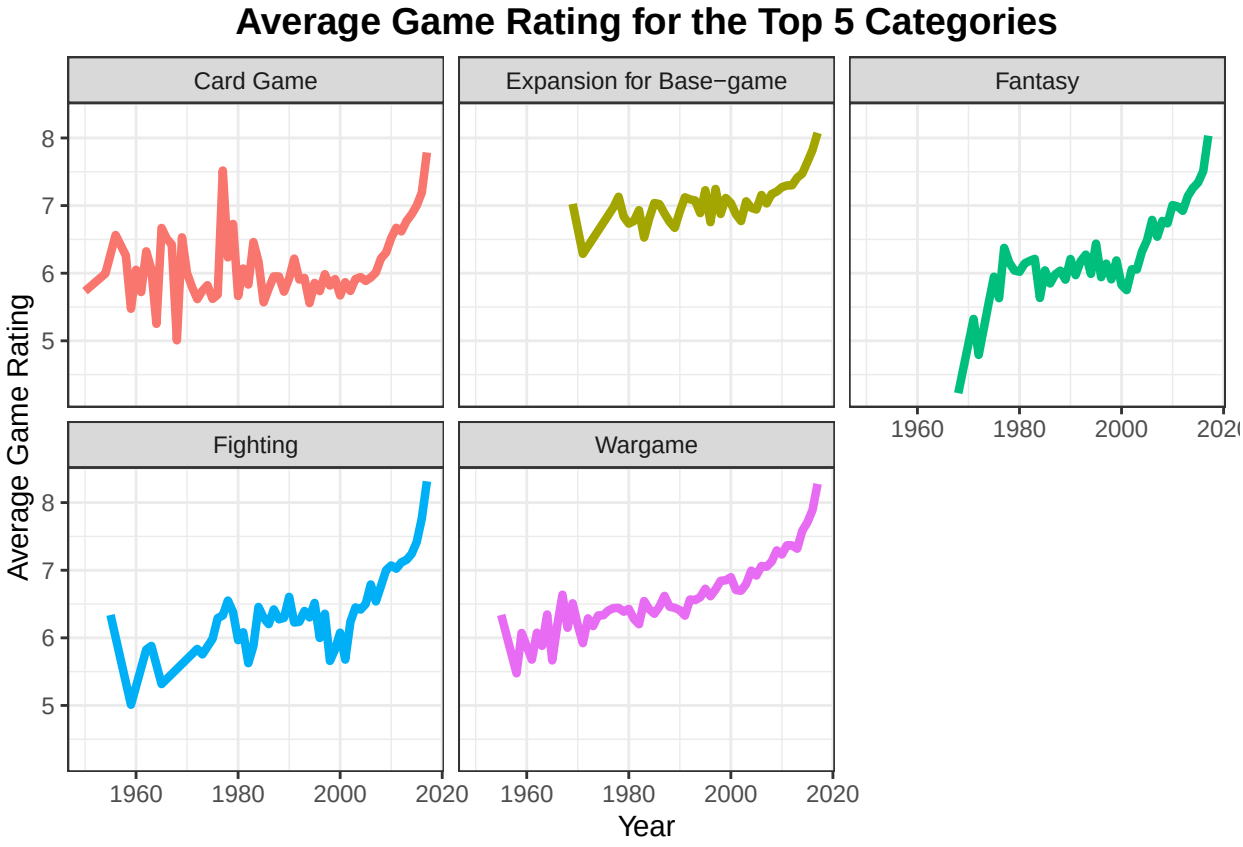
Average rating of board games over the top 5 categories:

```
# Find the top 5 game categories in terms of number of games:
top_cat <- long %>%
  filter(category %in% names(sort(table(long$category), decreasing=T)[1:5]))

# Find the average rating and complexity for the top 5 game categories in every
# year:
```

```r
avg_cat_years <- top_cat %>%
  filter(yearpublished != 2018) %>%
  group_by(yearpublished, category) %>%
  summarise(mean_rating = mean(average_rating),
            mean_complex = mean(average_complexity))

# Plot how the average rating of games changes over time for each of the top 5
# categories:
ggplot(data = avg_cat_years,
       mapping = aes(x = yearpublished, y = mean_rating, color = category)) +
  geom_line(size = 1.5) +
  facet_wrap(~category) +
  labs(x = "Year",
       y = "Average Game Rating",
       title = "Average Game Rating for the Top 5 Categories") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
        legend.position = "none")
```



Caption:
The above line charts show how the average rating of the games released in each year has changed over the past 75 years for each of the top 5 most occurring game categories (`category`). These top 5 categories are: `Card Game`, `Expansion for Base-game`, `Fantasy`, `Fighting`, and `Wargame`. In particular, the y-axis shows the mean of the `average_rating` variable for all of the games released in each year from 1950 – 2018, and the x-axis shows the year represented by the `yearpbulished` variable from 1950 – 2018.

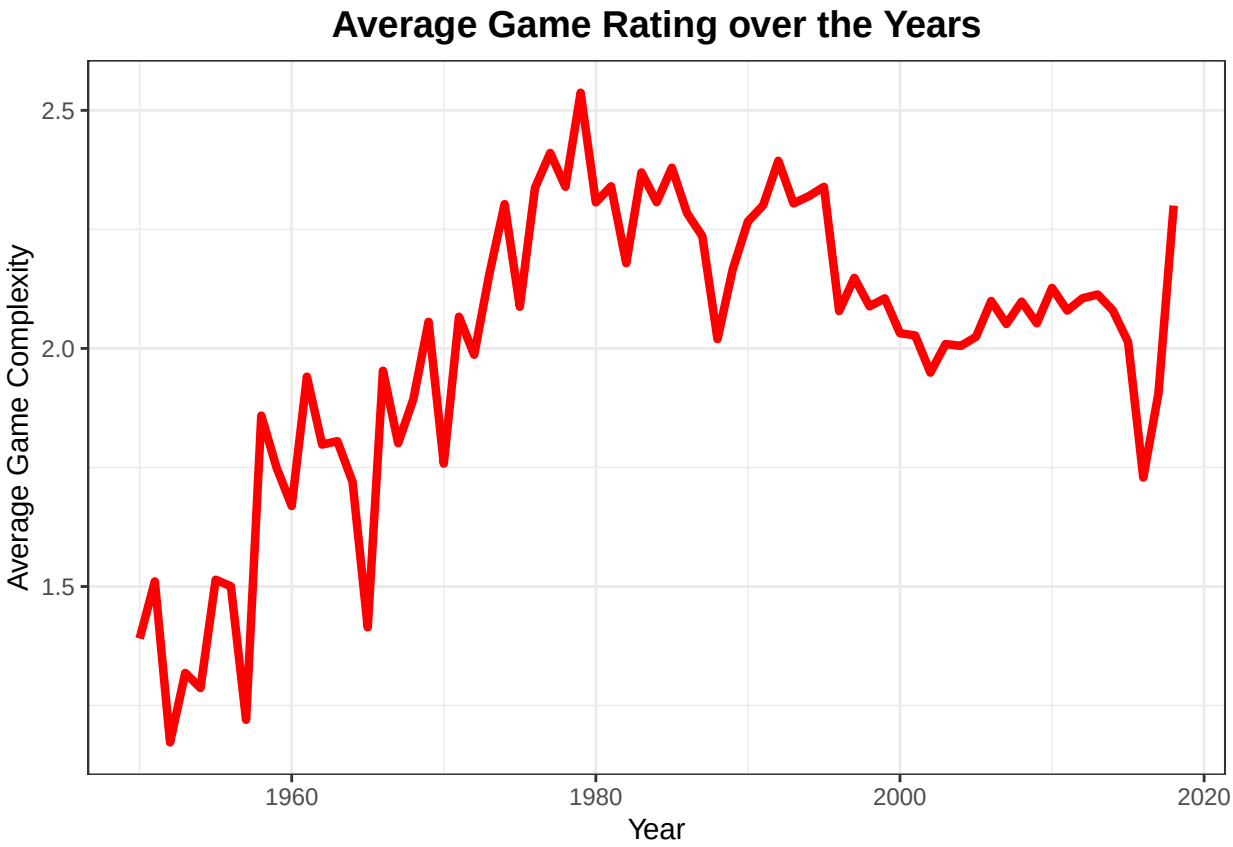
Insights:
As can be seen from the above line charts, each of the top 5 categories of board games has seen an increase of average rating from their first inception, each peaking with a max average rating of about 8 out of 10. The `Fantasy` category has seen the most growth in average rating, starting at an average rating of about 4.5 in the 1970s to its peak average rating of about 8 in the late 2010s. On the contrary, the `Card Game` category has had the most variable growth as it has seen many ups and downs since the 1950s but still managed to hit its peak average rating of around 8 in the late 2010s. The `Expansion for Base-game`, `Fighting`, and `Wargame` categories all followed a similar increasing trend throughout the given time frame. One other important thing to notice about this graph is that the different categories start at different points in time. For example, the `Expansion for Base-game` category didn't have its first game published until the late 1960s, while the `Card Game` category has been around since the beginning of our allotted time frame. One reason which might explain why the other categories performed better than the `Card Game` category over all of the years, is that card games have been around for centuries, thus the creation of new categories of games was more exciting to game enjoyers and thus led to higher average ratings for the newer categories, and a stagnating average rating for the `Card Game` category.

Now we will shift gears and do the same analysis but for the `average_complexity` variable.

Average complexity of board games over time:

```
# Group the data by year and find the average complexity of games in that year:
bgame_year <- bgame_filtered %>%
  group_by(yearpublished) %>%
  summarise(mean = mean(average_complexity))

# Create a line chart to show how average complexity changes over time:
ggplot(data = bgame_year, mapping = aes(x = yearpublished, y = mean)) +
  geom_line(color = "red", size = 1.5) +
  labs(x = "Year",
       y = "Average Game Complexity",
       title = "Average Game Rating over the Years") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14))
```



Caption:
The above line chart shows how the average complexity of the games released in each year has changed over the past 75 years. In particular, the y-axis shows the mean of the `average_complexity` variable for all of the games released in each year from $1950 - 2018$, and the x-axis shows the year represented by the `yearpbulished` variable from $1950 - 2018$.

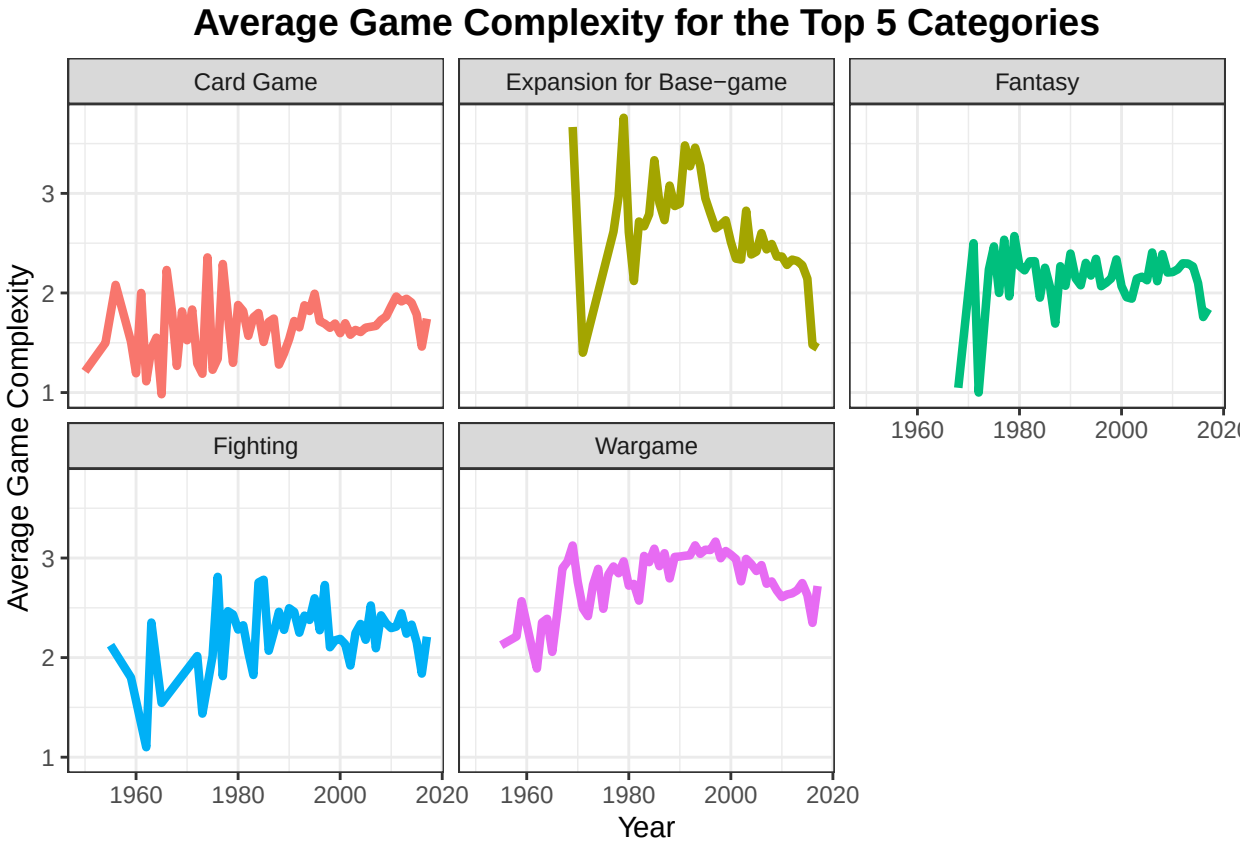
Insights:
The above line chart shows that from 1950-1980 the general trend was a linear increase in the average game complexity. In particular, in the year 1980 the max average game complexity of 2.5 out of 5 was reached. However, after this peak, the general trend flipped from increasing to decreasing. Since 1980 the highest average game complexity was around 2.3 in 1993, while the low average game complexity was around 1.75 in 2015. One explanation for this decrease after 1980 can be explained by the previous line graph plotting average game ratings over time. In that plot we saw that game scores slowly increases until 1980 where they seemed to stagnate. This may have occurred due to the fact that the average game was too complicated for the average player and thus the target audience was narrower than it was before. Thus, in order to increase overall ratings, simpler games were released to appease a wider audience.

In the next visualization, we will see this same data but broken up into different categories. Just like the average ratings plots, since there are 85 unique game categories in this dataset, that would be far too many categories to visualize. Thus, we will focus on only the top 5 categories based on frequency.

Average complexity of board games over the top 5 categories:

```
# Plot how the average complexity of games changes over time for each of the top
# 5 categories:
ggplot(data = avg_cat_years,
```

```
      mapping = aes(x = yearpublished, y = mean_complex, color = category)) +
geom_line(size = 1.5) +
facet_wrap(~category) +
labs(x = "Year",
     y = "Average Game Complexity",
     title = "Average Game Complexity for the Top 5 Categories") +
theme_bw() +
theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
      legend.position = "none")
```



Caption:
The above line charts show how the average complexity of the games released in each year has changed over the past 75 years for each of the top 5 most occurring game categories (`category`). These top 5 categories are: `Card Game`, `Expansion for Base-game`, `Fantasy`, `Fighting`, and `Wargame`. In particular, the y-axis shows the mean of the `average_complexity` variable for all of the games released in each year from $1950 - 2018$, and the x-axis shows the year represented by the `yearpbulished` variable from $1950 - 2018$.

Insights:
As can be seen from the above line charts, each of the top 5 categories of board games has seen a relatively stagnant growth in their average complexity from their first inception, each peaking in the upper 2s and lower 3s in terms of complexity. Notice however, that the `Expansion for Base-game` category has seen the most variable and large changes in average complexity, and unlike the rest of the categories, saw its peak average complexity at around 3.5 out of 5. We also see that the `Card Game` category has stayed relatively simple in terms of complexity throughout the entirety of the time period. The reason why `Expansion for Base-game` is much more difficult on average than the other categories, is that expansions add new elements to the game. So if a game is already complex, the expansion only makes it more complex. Furthermore,

`Card Games` have always been relatively simple because these categories of games are meant to be quick and have a more general target audience than `Fantasy` or `Fighting` games do.

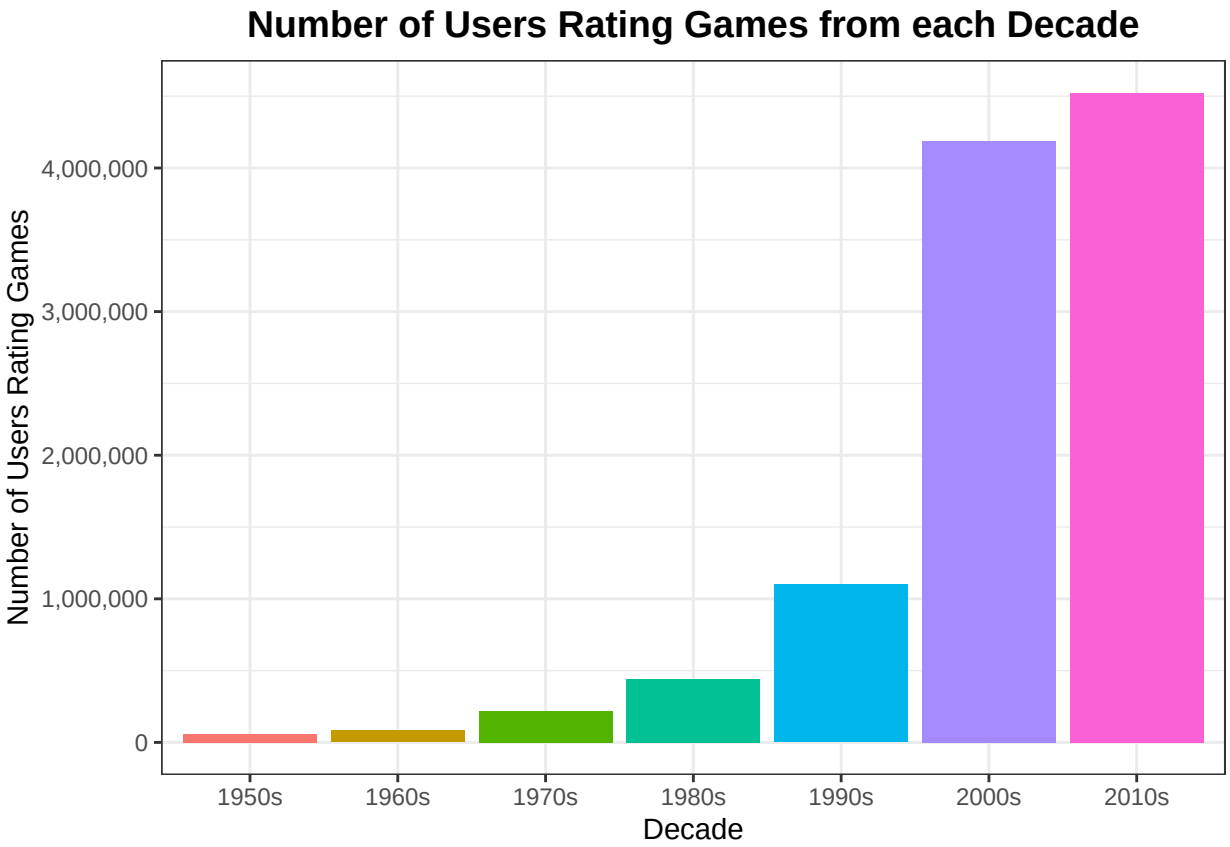
Lastly, we will change gears and see how the number of user ratings have changed over the years.

Number of users rated per decade:

```
# Recode the data into a new column named decade which maps the corresponding
# publish year into the decade it was published:
bgame_decade <- bgame_filtered %>%
  mutate(year = yearpublished,
         decade = case_when(year >= 1950 & year < 1960 ~ "1950s",
                            year >= 1960 & year < 1970 ~ "1960s",
                            year >= 1970 & year < 1980 ~ "1970s",
                            year >= 1980 & year < 1990 ~ "1980s",
                            year >= 1990 & year < 2000 ~ "1990s",
                            year >= 2000 & year < 2010 ~ "2000s",
                            year >= 2010 & year < 2020 ~ "2010s"))

# Group the data by decades and count the total number of users who submitted
# a rating from that decade:
users_decade <- bgame_decade %>%
  group_by(decade) %>%
  summarise(sum = sum(users_rated))

# Plot a bar chart of the total number of users rating games from each decade:
ggplot(data = users_decade, mapping = aes(x = decade, y = sum, fill = decade)) +
  geom_bar(stat = "identity") +
  labs(x = "Decade",
       y = "Number of Users Rating Games",
       title = "Number of Users Rating Games from each Decade",
       fill = "Decade") +
  scale_y_continuous(label=comma) +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
        legend.position = "none")
```



Caption:
In the above barplot, the number of users rating games in from decade is shown. On the y-axis, the number of users rating games, which is the sum of the `users_rated` variable throughout all games published in a given decade, is displayed. On the x-axis, the decades from the 1950s to the 2010s are shown. These decades were found by recoding the `yearpublished` variable.

Insights:
As can be seen from the above barplot, the number of users rating games from each decade has increased dramatically from each decade to the next. The most dramatic rise was from the 1990s to the 2000s, where an astounding 3 million more ratings were received from games released in the 2000s than were received for game from the 1990s. Despite such a jump from the 1990s to 2000s, the rise from the 2000s to the 2010s wasn't nearly as big, with only a rise of around 250,000 more ratings. The reason for such a disparity between the amount of ratings for games published in the 21st century versus game published in the 20th century boils down to one simple cause: the rise of technology in the early 21st century. Due to the fact that boardgamegeek.com didn't even get founded until 2000, most of the major fans of the games from the 1950s-1980s were not living in a technological age in which they could rate games online. Thus when the site was created, they were past the point where they would feel inclined to rate a game. On the contrary, for every game released in the 21st century, there was an easily accessible place to rate games, and hence why we see an astounding increase of ratings in these decades.

Overall conclusions for part 2:

Overall, we can see that as time has moved on, the average ratings of games and the number of users rating games has increased, with the increases for average rating carrying over to the top 5 board game categories. However, since the 1980s, the average complexity of games has gone down and has only seen a recovery in the average complexity since 2015, this same sentiment has carried over for the most part to the top5

categories. Although in these categories, the decline/growth of the average complexity has been a lot more stagnant than the overall trend over time.

Exploratory topic 3: How do aspects and feature of games differ between the top categories and mechanics?:

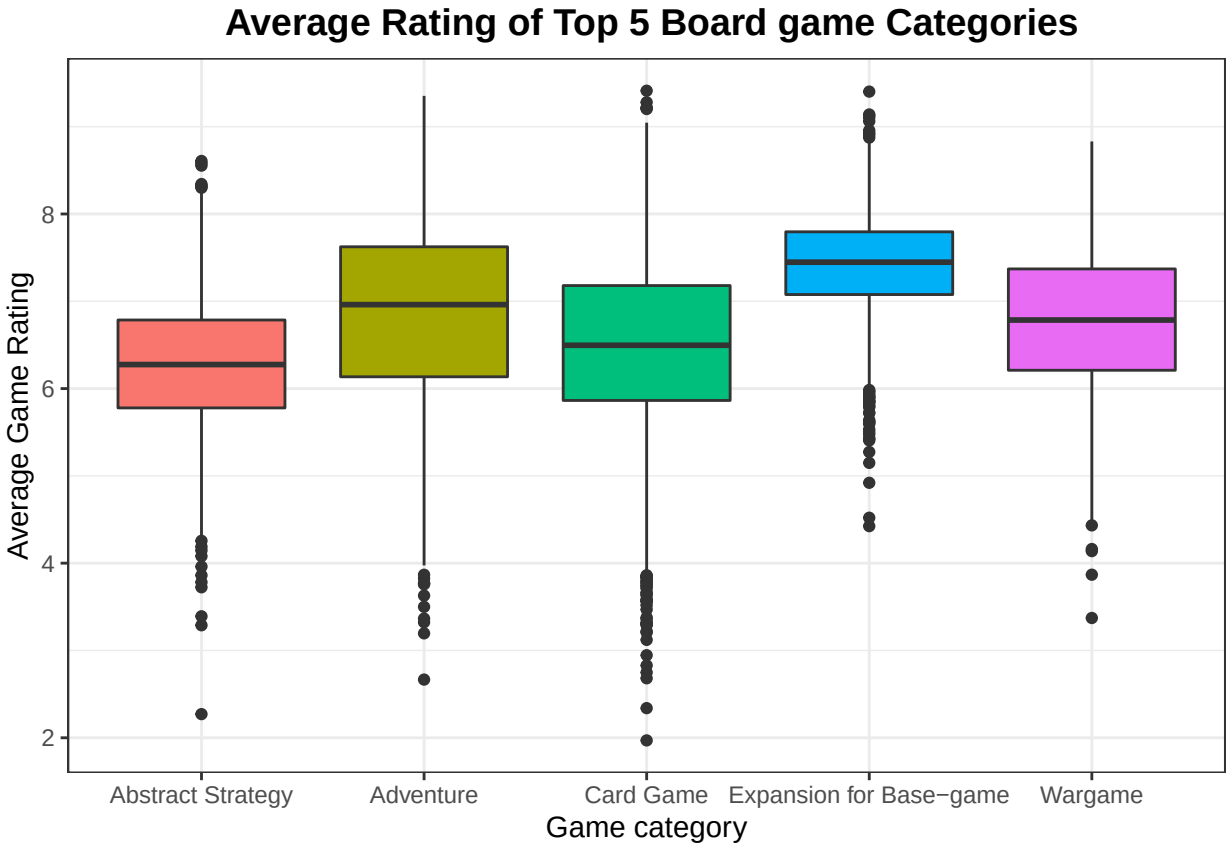
We will now dig deeper into the top 5 categories of game frequency and mechanics. It is important to note that each game has multiple categories, and in the last part we didn't account for these overlaps since we were looking at changes over time. Hence, unlike the last topic, we will only be looking at the top 5 categories with no overlap. Meaning we will only consider the top category of a board game even if they are considered a part of another category. These top categories and mechanics are stored in the `category_01` and `mechanic_01` columns of the `wide` dataframe, respectively. Namely, we will look at the average rating of the top 5 categories, the average playing time of the top 5 mechanics, and lastly we will dive into two scatterplots of the average complexity and average ratings of board games.

Average rating for games in each category:

Just like the plots in part 2, since there are 85 unique game categories in this dataset, that would be far too many categories to visualize. Thus, we will focus on only the top 5 non overlapping categories based on frequency.

```
# Finding the top 5 unique game types:
top_cat_uniq <- wide %>%
  filter(category_01 %in% names(sort(table(wide$category_01), decreasing=T)[1:5]))

# Plot boxplots to show the average rating for each of the top 5 unique game
# types:
ggplot(data = top_cat_uniq,
       mapping = aes(x = category_01, y = average_rating, fill = category_01)) +
  geom_boxplot() +
  labs(x = "Game category",
       y= "Average Game Rating",
       title = " Average Rating of Top 5 Board game Categories") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
  legend.position = "none")
```



Caption:
The above boxplots displays the different distributions of average rating for the top 5 non-overlapping board game categories. These categories include: `Abstract Strategy`, `Adventure`, `Card Game`, `Expansion for Base-game`, and `Wargame`. On the y-axis the average rating for each respective category is displayed, while on the x-axis the different board game categories of interest are displayed. Each boxplot shows the first, second, and third quartiles of the average rating data as well as any outliers.

Description:
As can be seen from the above boxplots, all of the distributions of the average rating for the top 5 non-overlapping categories show very symmetric/approximately normal tendencies. Furthermore, all of the distributions have outliers, with the `Expansion for Base-game` distribution having the most outliers out of all of the categories. All of the categories have a center/median average rating between 6 and 7.5, again with the `Expansion for Base-game` category having the highest at around 7.5 out of 10. The `Abstract Strategy`, `Adventure`, `Card Game` categories all have a similar spread, while the `Expansion for Base-game` has the smallest spread out of all of the categories.

Insights:
As can be interpreted from the above plot and the accompanying description, all of the categories have similar average ratings other than the `Expansion for Base-game` category, which has the highest average rating by around half of a rating point. It also turns out that the `Expansion for Base-game` category is the most centered around its mean value, meaning that it strays away from its average less than the other categories. As has been explain numerous time in the previous two parts, the reason why the `Expansion for Base-game` performs so much better than the other categories on average is because in order to get an expansion, you must already like the original game to begin with. Thus, you are more likely to enjoy an expansion than you are a game you've never played before.

We will now take a break from looking at the different game categories and instead look at game mechanics.

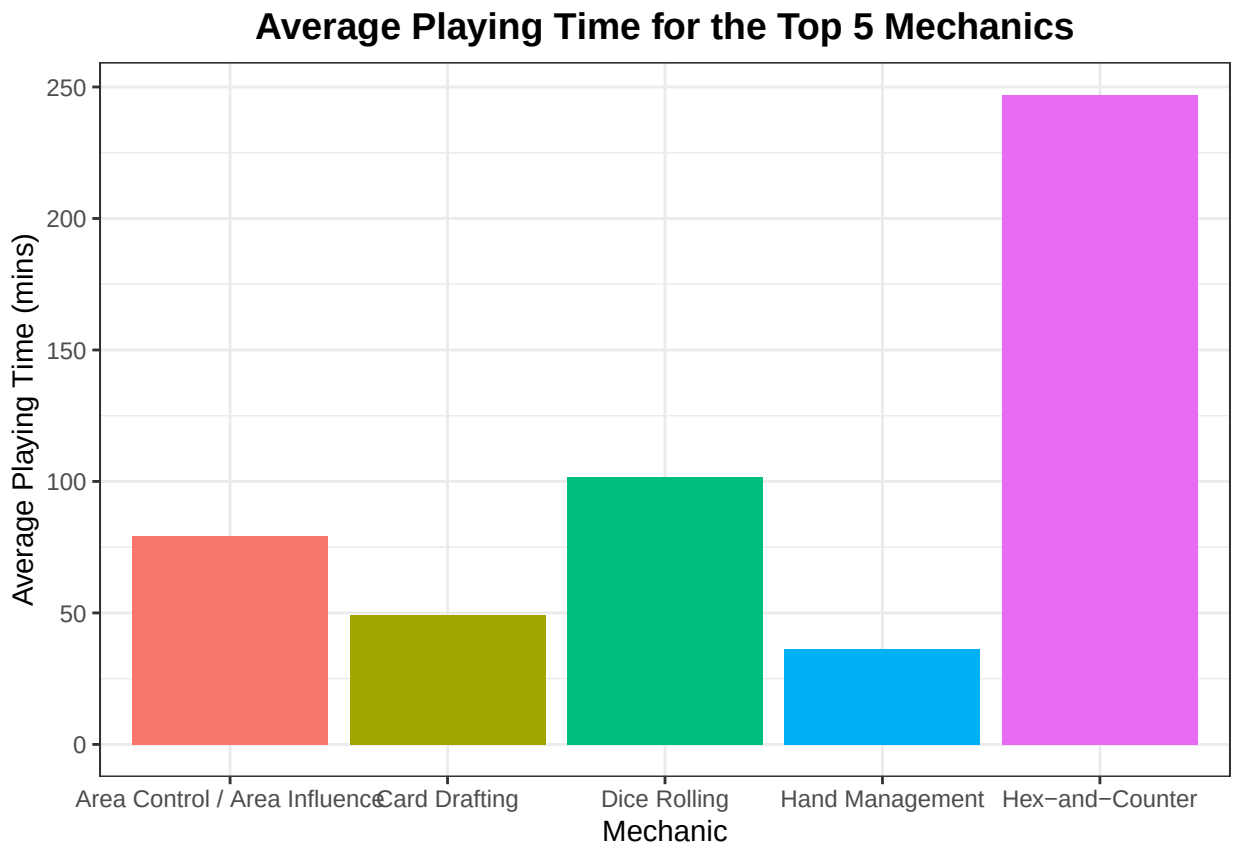
Average playtime for top 5 mechanics:

In the next visualization we will see playing time data broken up by different game mechanics. Since there are 51 unique game mechanics in this dataset, that would be far too many categories to visualize. Thus, we will focus on only the top 5 non-overlapping mechanics based on frequency.

```
# Find the top 5 unique game mechanics:
top_mech_uniq <- wide %>%
  filter(mechanic_01 %in% names(sort(table(wide$mechanic_01), decreasing=T)[1:5]))

# Group by the top 5 unique mechanics and find the average playing time:
avg_time <- top_mech_uniq %>%
  group_by(mechanic_01) %>%
  summarise(avg_play = mean(playingtime))

# Plot a barplot to show the average playing time for each of the top 5 game
# mechanics:
ggplot(data = avg_time,
       mapping = aes(x = mechanic_01, y = avg_play, fill = mechanic_01)) +
  geom_bar(stat = "identity") +
  labs(x = "Mechanic",
       y = "Average Playing Time (mins)",
       title = "Average Playing Time for the Top 5 Mechanics") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
        legend.position = "none")
```



Caption:
The above barplot shows the average playing time in minutes for the top 5 non-overlapping game mechanics. These game mechanics include: `Area Control / Area Influence`. `Card Drafting`, `Dice Rolling`, `Hand Management`, and `Hex-and-Counter`. On the y-axis the average playing time in minutes across all games of each respective mechanic is displayed, while on the x-axis the top 5 non-overlapping mechanics are displayed.

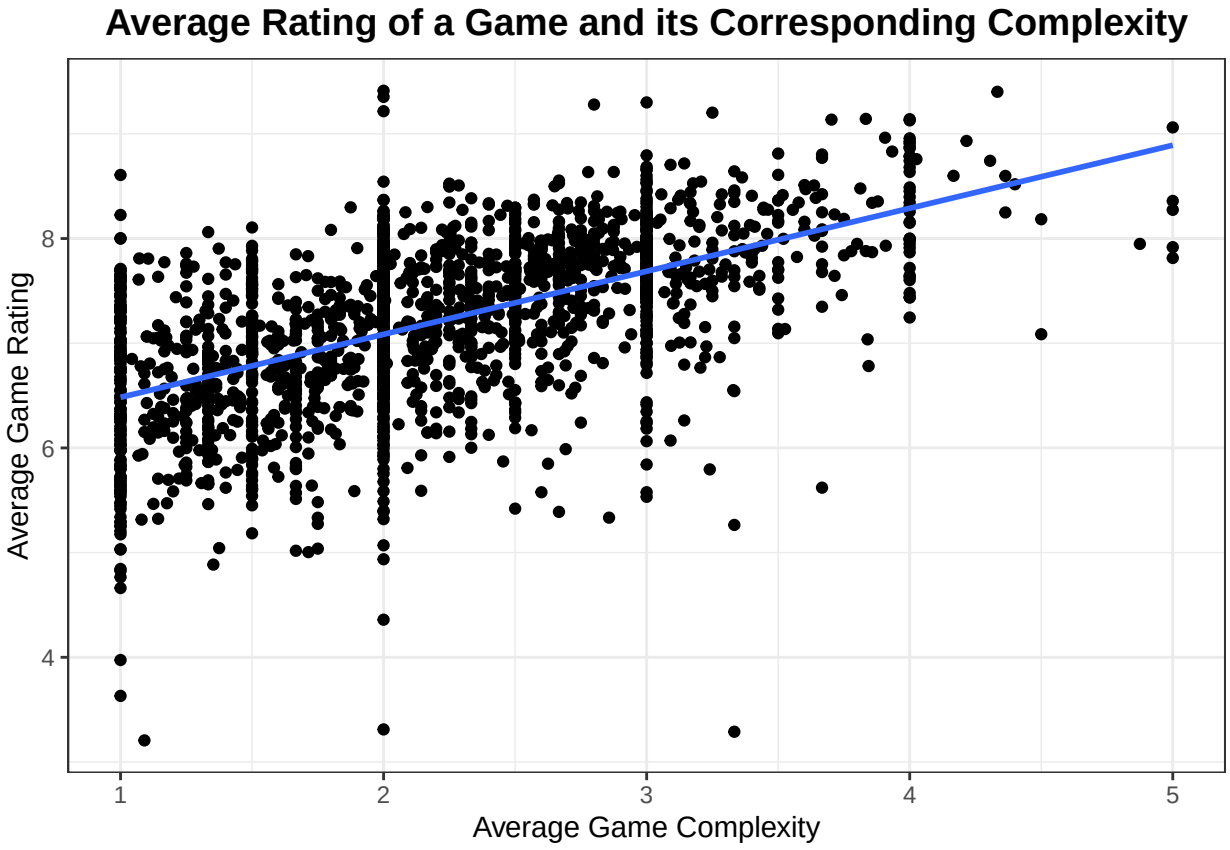
Insights:
As can be seen by the above barplot of the average playing time for the top 5 mechanics, we can see that there are vastly different playing time for the different mechanics. In particular the the mechanics that deal with cards: `Card Drafting` and `Hand Management`, take the least amount of time with an average playing time of around 45 minutes respectively. On the contrary, strategy games that follow the `Hex-and-Counter` game mechanic take an astounding 250 minutes to complete on average. There are many reasons why these trends may be occurring. To start us off, the reason why we may be seeing the card mechanic games taking the least amount of time is that card games are usually simpler than other games with shorter turn lengths and thus can be learned and played quicker than their more complicated counterparts. On the other hand, hex-and-counter games usually have long and complicated rules and turns and thus take much longer to learn and play when compared to simpler games. One concrete reason why the `Hex-and-Counter` has such a high average is because the game "The Struggle For North Africa", which is a hex-and-counter game, takes over 40 days to complete! Thus this ridiculously high number skews the data upwards. It is important to note that even without the inclusion of this game, the `Hex-and-Counter` mechanic still has a longer playing time on average.

We will now switch our frame of focus onto seeking the relationship between the average rating and average complexity variables.

Normal scatterplot for average complexity vs average ratings:

```
# Filter the data to only get games that have a complexity that is not zero and
# that were released in the past 10 years:
cleaned_comparison <- top_cat_uniq %>%
  filter(average_complexity > 0) %>%
  filter(yearpublished > 2013)

# Plot a basic scatterplot of average game rating and average complexity:
ggplot(data = cleaned_comparison,
       mapping = aes(x = average_complexity, y = average_rating)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Average Game Complexity",
       y = "Average Game Rating",
       title = "Average Rating of a Game and its Corresponding Complexity") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
        legend.position = "none")
```



Caption:
The above scatterplot attempts to show the relationship between the average complexity (`average_complexity`) and the average rating (`average_rating`) variables of all board games published in the past 10 years that had an average complexity above 0. On the y-axis, the average game rating for games released in the past 10 years are displayed. While on the x-axis the average game complexity for games released in the past 10 years are displayed.

Description:
As can be seen in the above scatterplot, there is a moderate increasing linear positive association between the average complexity and average rating of board games released in the past 10 years. Due to the moderate correlation between these two variables, there are many outliers that stray far away from the line of best fit, especially near the lower end of the average rating variable. Also, there are distinct lines at the integer values of 1, 2, 3, and 4 because individuals are more likely to rate something as a whole number than a very specific floating point rating.

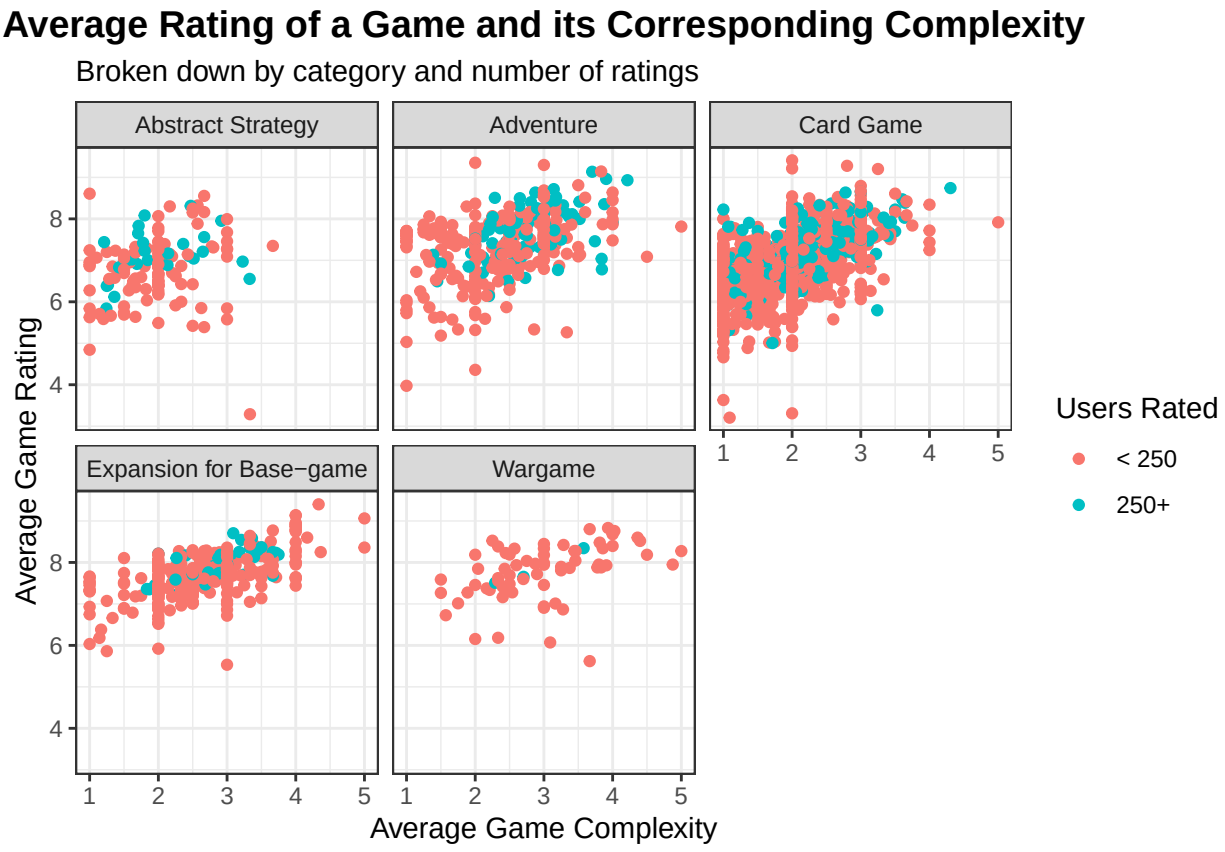
Insights:
Based on the above scatterplot and accompanying description, we can see that, in general, as the average complexity of a game increases, the average rating of the game also increases. This can mainly be explained due to the nicheness of board games themselves. Since board games aren't the most popular past time among other things like video games, movies, TV shows, etc. they have become more of a past time for certain audiences. Thus those specialized audiences enjoy more complex games in order to challenge themselves and not get sick of board games in general. Hence, on average, when a game gets more difficult, avid board game fans accept the challenge and thus have higher ratings for the game.

In the next visualization, we will see this same data but broken up into different categories. Since there are 85 unique game categories in this dataset, that would be far too many categories to visualize. Thus, we will focus on only the top 5 non-overlapping categories based on frequency.

Scatterplots for average complexity vs average rating broken down:

```
# Create a new column that enforces the rating criteria and makes labels
# easier to read:
mutated_comparison <- cleaned_comparison  %>%
  mutate(properly_rated = users_rated >= 250,
         prop_rated = case_when(properly_rated == TRUE ~ "250+",
                                properly_rated == FALSE ~ "< 250"))

# Plot scatterplots of average game rating and average complexity split up
# based upon categories and number of users rated:
ggplot(data = mutated_comparison,
       mapping = aes(x = average_complexity, y = average_rating, color = prop_rated)) +
  geom_point() +
  facet_wrap(~category_01) +
  labs(x = "Average Game Complexity",
       y = "Average Game Rating",
       title = "Average Rating of a Game and its Corresponding Complexity",
       subtitle = "Broken down by category and number of ratings",
       color = "Users Rated") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14))
```



Caption:
The above scatterplots attempt to show the relationship between the average complexity (`average_complexity`)

and the average rating (`average_rating`) variables of all board games published in the past 10 years that had an average complexity above 0, for each of the top 5 non-overlapping game categories. These categories include: `Abstract Strategy`, `Adventure`, `Card Game`, `Expansion for Base-game`, and `Wargame`. Furthermore, all of the data points in all of the top 5 categories are broken up into two types: those that have over 250 ratings, and those that have under 250 ratings. On the y-axis, the average game rating for games released in the past 10 years are displayed. While on the x-axis the average game complexity for games released in the past 10 years are displayed.

Description:
As can be seen from the above scatterplots, all of the individual categories have the same moderate linear positive association between the average game rating and average complexity variables. The two game categories that have the strongest positive linear association are the `Expansion for Base-game`, and `Wargame` categories. While the `Card Game` category had the weakest association between these two variables. Lastly, on average, the points that had the more ratings usually were higher rated than there lesser rated counterparts.

Insights:
Based on the above scatterplots and accompanying description, we can see that, in general, as the average complexity of a game increases, the average rating of the game also increases. This is the same conclusion we came to in the single scatterplot and it indeed holds, for the most part, for all of the top 5 non-overlapping categories. As has been a common theme in this report, the reason why the `Expansion for Base-game` category follows this trend the best is because if you are already passionate about a board game then when the expansion is released you're more likely to enjoy that as well. On the same token, this expansion also adds new features to the game and thus makes it more difficult.

Overall conclusions for part 3:

Overall, we can see that as more complicated Mechanics increase the average playing time of a board game, and that across all game categories, the `Expansion for Base-game` category has the highest average rating and the least variability from this average. Lastly, in general, regardless of the underlying category, as the average complexity of a board game increases, the average rating of the game also increases.