Exploratory analysis of the 'Board Games' data set

João Conde

31st January 2020

Load libraries and data set

```
library(scales)
library(gridExtra)
library(tidyverse)
boardgames_tbl <- read_csv('boardgames.csv')
```

Explore data set

```
head(boardgames_tbl)
```

```
## # A tibble: 6 x 20
##       id type  name  yearpublished minplayers maxplayers playingtime minplaytime
##    <dbl> <chr> <chr>          <dbl>      <dbl>      <dbl>       <dbl>       <dbl>
## 1  12333 boar~ Twil~           2005          2          2         180         180
## 2 120677 boar~ Terr~           2012          2          5         150          60
## 3 102794 boar~ Cave~           2013          1          7         210          30
## 4  25613 boar~ Thro~           2006          2          4         240         240
## 5   3076 boar~ Puer~           2002          2          5         150          90
## 6  31260 boar~ Agri~           2007          1          5         150          30
## # ... with 12 more variables: maxplaytime <dbl>, minage <dbl>,
## #   users_rated <dbl>, average_rating <dbl>, bayes_average_rating <dbl>,
## #   total_owners <dbl>, total_traders <dbl>, total_wanters <dbl>,
## #   total_wishers <dbl>, total_comments <dbl>, total_weights <dbl>,
## #   average_weight <dbl>
```

```
glimpse(boardgames_tbl)
```

```
## Observations: 81,312
## Variables: 20
## $ id                  <dbl> 12333, 120677, 102794, 25613, 3076, 31260, 124...
## $ type                <chr> "boardgame", "boardgame", "boardgame", "boardg...
## $ name                <chr> "Twilight Struggle", "Terra Mystica", "Caverna...
## $ yearpublished       <dbl> 2005, 2012, 2013, 2006, 2002, 2007, 2012, 2011...
## $ minplayers          <dbl> 2, 2, 1, 2, 2, 1, 2, 1, 2, 2, 2, 2, 2, 1, 1, 3...
## $ maxplayers          <dbl> 2, 5, 7, 4, 5, 5, 2, 4, 4, 6, 6, 5, 4, 4, 5, 4...
## $ playingtime         <dbl> 180, 150, 210, 240, 150, 150, 45, 150, 90, 200...
## $ minplaytime         <dbl> 180, 60, 30, 240, 90, 30, 45, 150, 30, 60, 120...
## $ maxplaytime         <dbl> 180, 150, 210, 240, 150, 150, 45, 150, 90, 200...
## $ minage              <dbl> 13, 12, 12, 12, 12, 12, 14, 14, 12, 14, 12, 0,...
## $ users_rated         <dbl> 20113, 14383, 9262, 13294, 39883, 39714, 15281...
## $ average_rating      <dbl> 8.33774, 8.28798, 8.28994, 8.20407, 8.14261, 8...
```

```
## $ bayes_average_rating <dbl> 8.22186, 8.14232, 8.06886, 8.05804, 8.04524, 8...
## $ total_owners         <dbl> 26647, 16519, 12230, 14343, 44362, 47522, 2438...
## $ total_traders        <dbl> 372, 132, 99, 362, 795, 837, 680, 367, 215, 27...
## $ total_wanters        <dbl> 1219, 1586, 1476, 1084, 861, 958, 627, 1116, 9...
## $ total_wishers        <dbl> 5865, 6277, 5600, 5075, 5414, 6402, 3244, 5427...
## $ total_comments       <dbl> 5347, 2526, 1700, 3378, 9173, 9310, 3202, 2861...
## $ total_weights        <dbl> 2562, 1423, 777, 1642, 5213, 5065, 1260, 1409,...
## $ average_weight       <dbl> 3.4785, 3.8939, 3.7761, 4.1590, 3.2943, 3.6160...
```

Searching for missing column values and ranges

Game types

```
## [1] "boardgame"          "boardgameexpansion"
```

Negative valued publishing years

```
## [1] 24
```

Games with less than 10 reviews

```
## [1] 57178
```

Columns with missing values (NA)

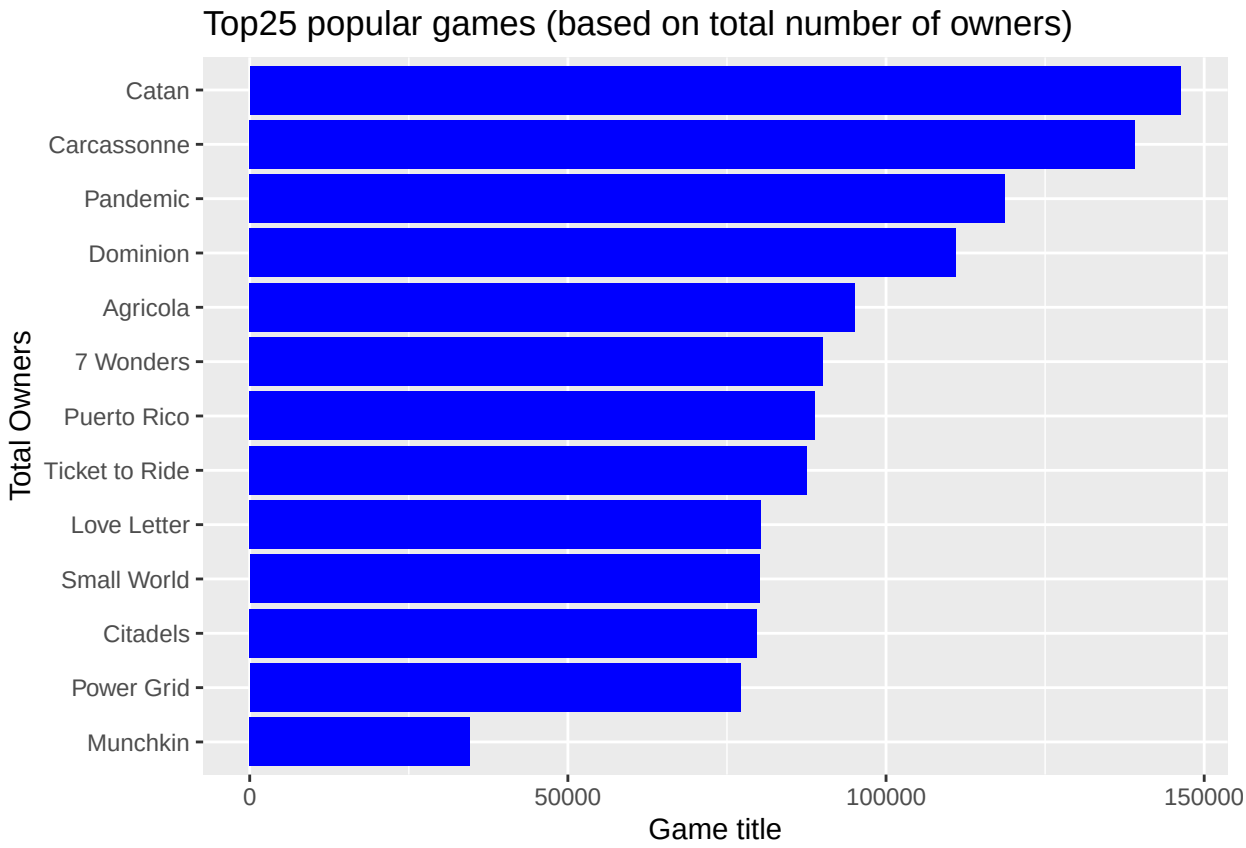
```
## [1] "name"          "yearpublished" "minplayers"    "maxplayers"
## [5] "playingtime"   "minplaytime"   "maxplaytime"   "minage"
```

Filtering dataset

```
boardgames_clean_tbl <-
  boardgames_tbl %>%

  # ignoring expansions
  filter(type == "boardgame") %>%

  # filter negative years while selecting a resonable interval
  filter(yearpublished >= 1960 & yearpublished <= 2019) %>%

  # eliminate games with too few reviews as the rating can be biased (e.g. one 5-star vote)
  filter(users_rated >= 10)
```

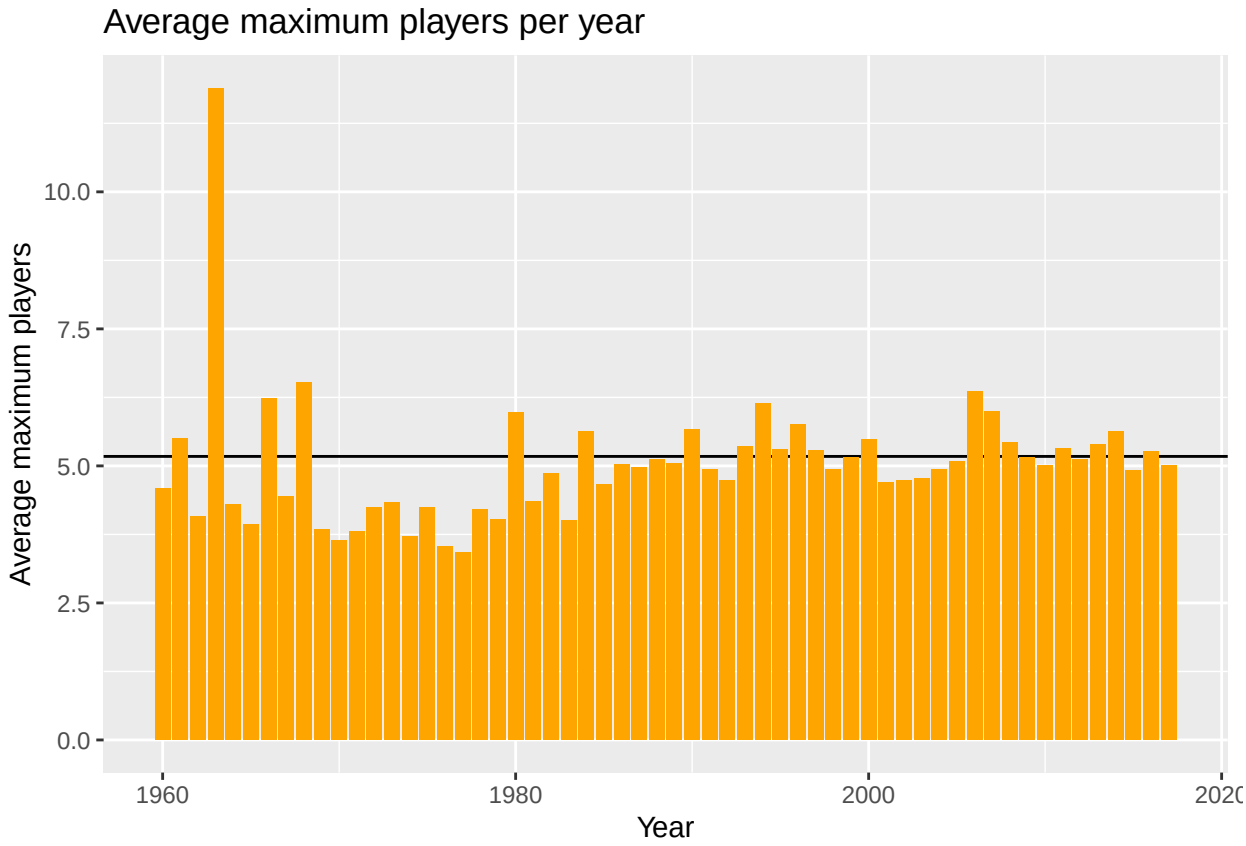
Top25 popular board games (based on total number of owners)

The graph below lists the top 25 games according to their popularity, based on total number of owners.



Maximum players average per game throughout the years

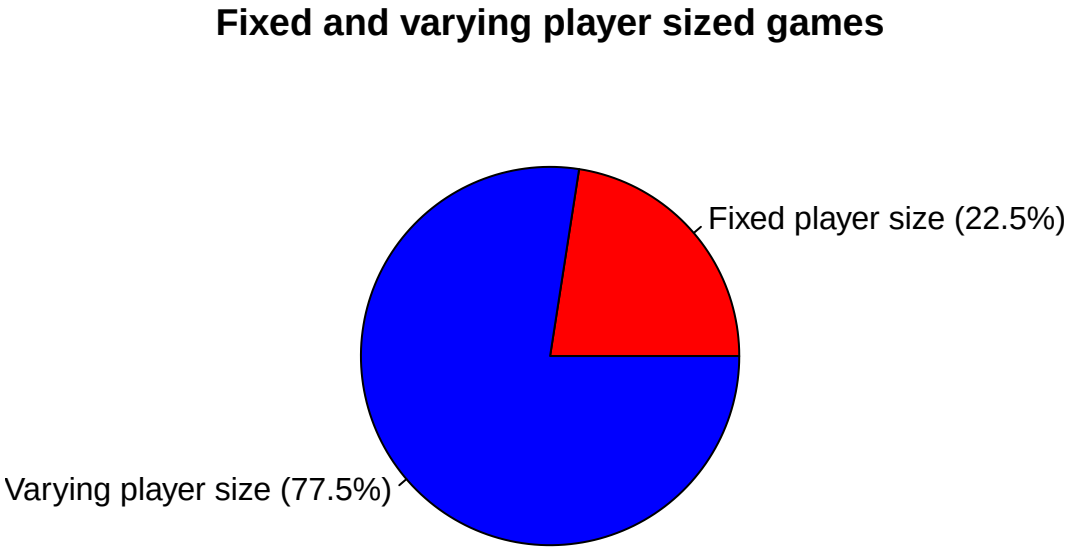
In order to evaluate if over the years games had become more multiplayer oriented, we plotted the graph below, with an average of the maximum number of players from all the games released per year and an horizontal black line as the all time average.



However, it seems board games are as multiplayer oriented as before, having very similar number of maximum players per year (excluding the outlier of 1963 which is far above the average for maximum player limit).

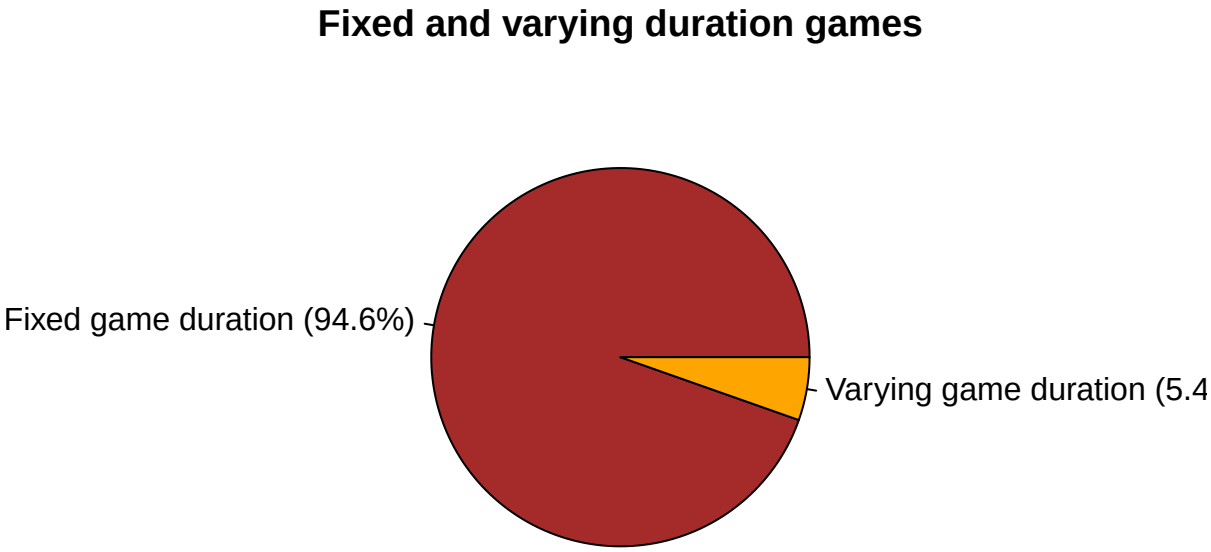
Fixed and variable number of players

A simple pie chart to illustrate the percentages of games that have an exact number of players and games that allow for a variable number of players (within a range).



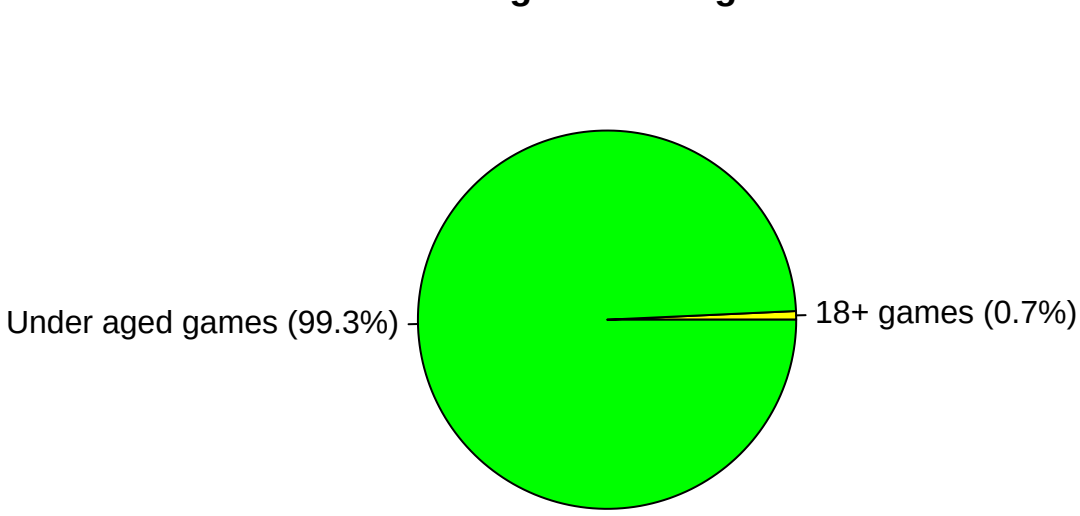
Fixed and varying duration games

Similarly to the last one, this pie chart illustrates the percentage of games with fixed and varying durations.



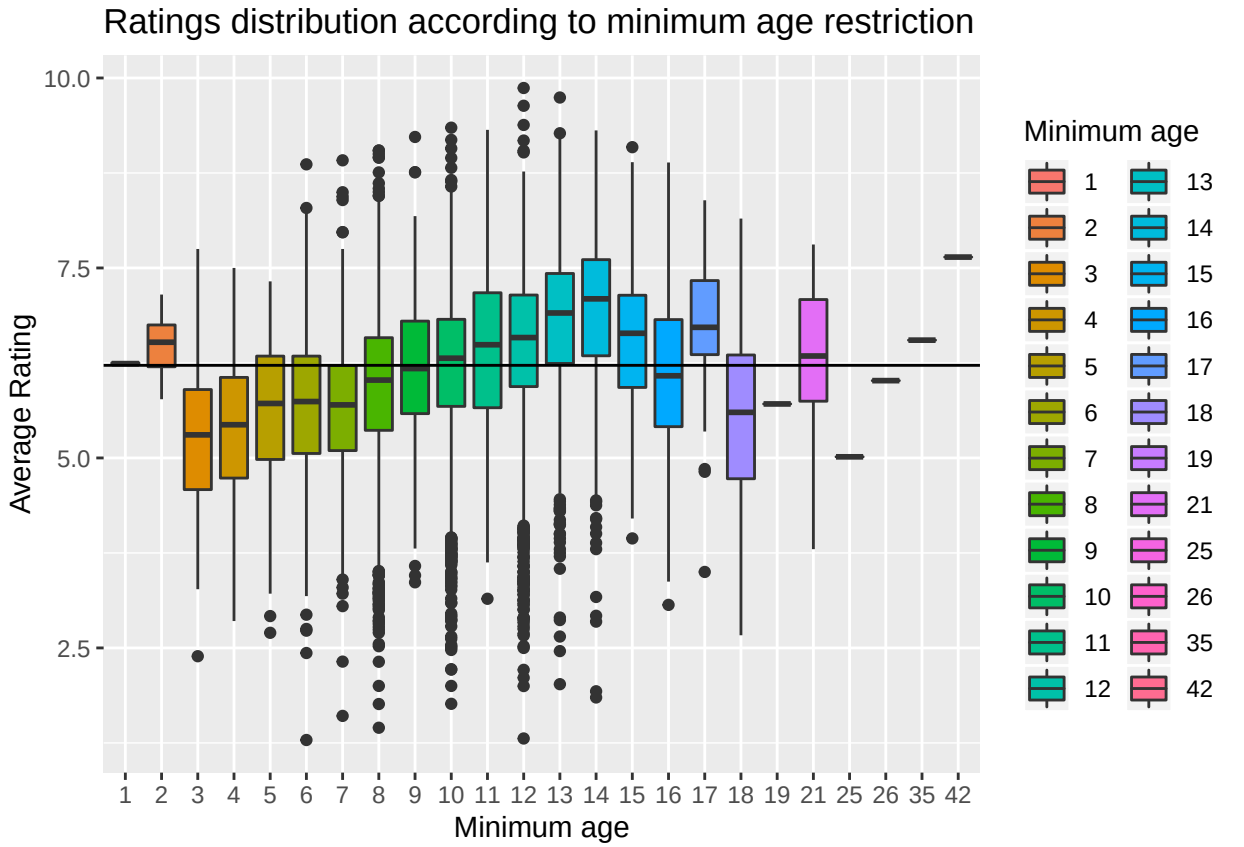
Games for adults (18+ years old)

The number of board games that requires players to be over 18 is, as expected, extremely low.



However, could the minimum age imposed by each board game attract different player bases hence having different ratings?

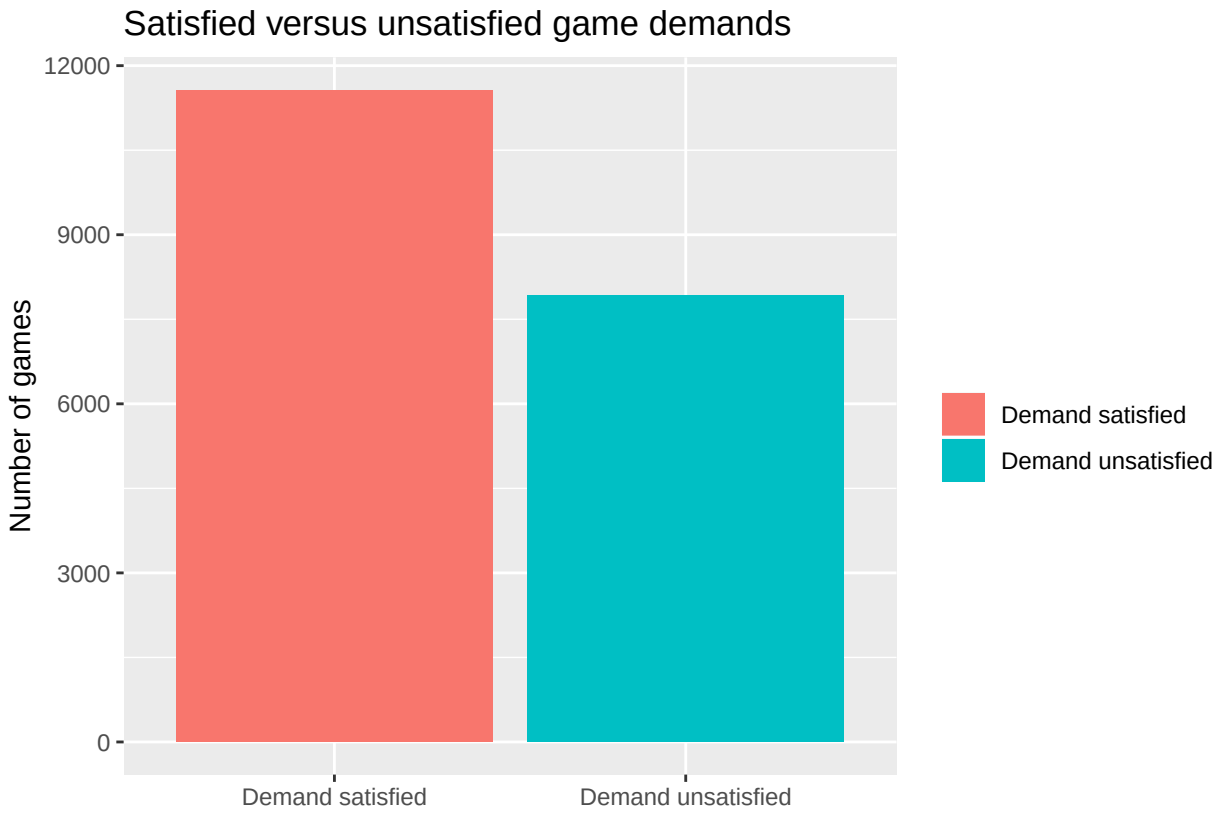
```
boardgames_clean_tbl %>%
  filter(minage > 0) %>%
  mutate(minage_discrete=as.factor(minage)) %>%
  ggplot(aes(minage_discrete, average_rating, fill=minage_discrete)) +
  geom_boxplot() +
  geom_hline(yintercept=mean(boardgames_clean_tbl$average_rating), color="black")+
  labs(x = "Minimum age", y = "Average Rating") +
  scale_fill_discrete(name = "Minimum age") +
  ggtitle("Ratings distribution according to minimum age restriction")
```



With the boxplot above we can conclude that there is not really a correlation between the minimum age required and the rating of a game.

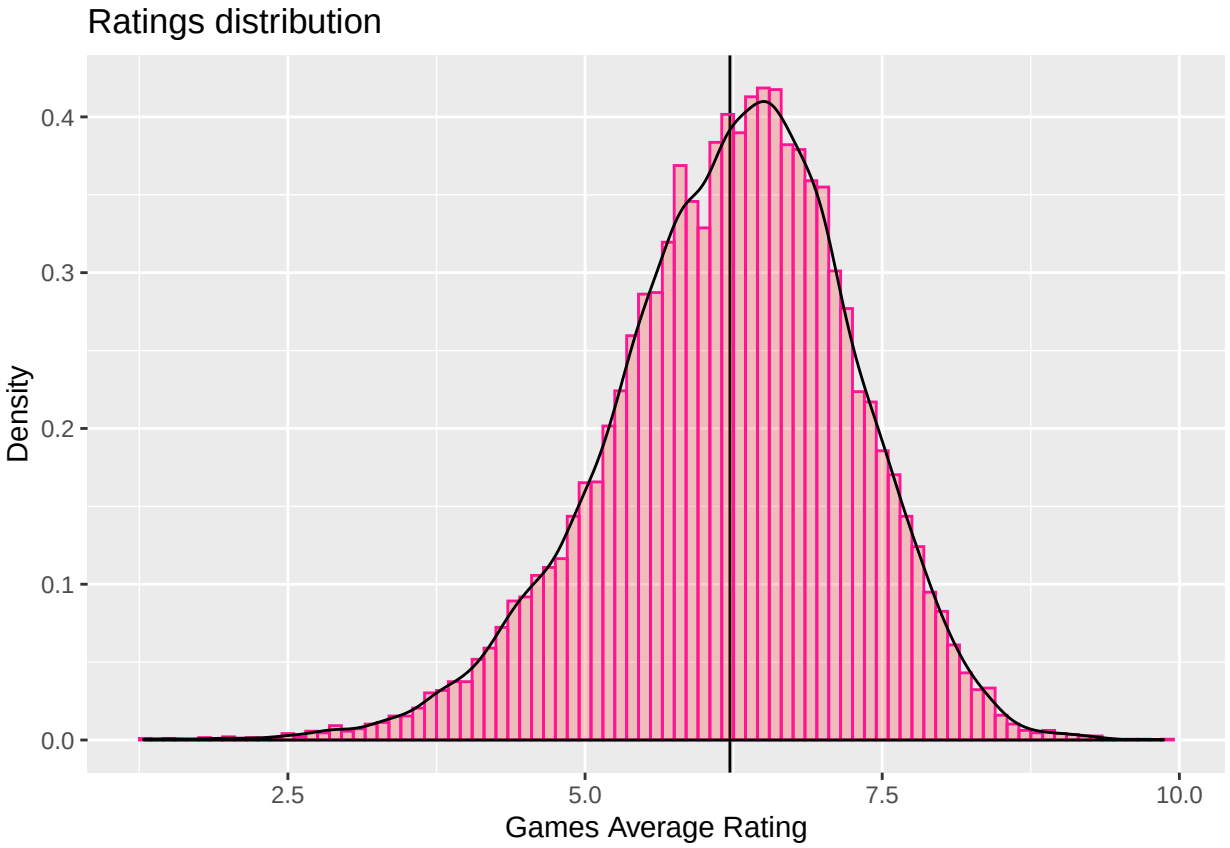
Traders and wanters

In the bar chart below we present the number of games where their demand is satisfied - *i.e.* there are more traders than wanters - and the number of games where their demand is unsatisfied - *i.e.* there are more wanters than traders.



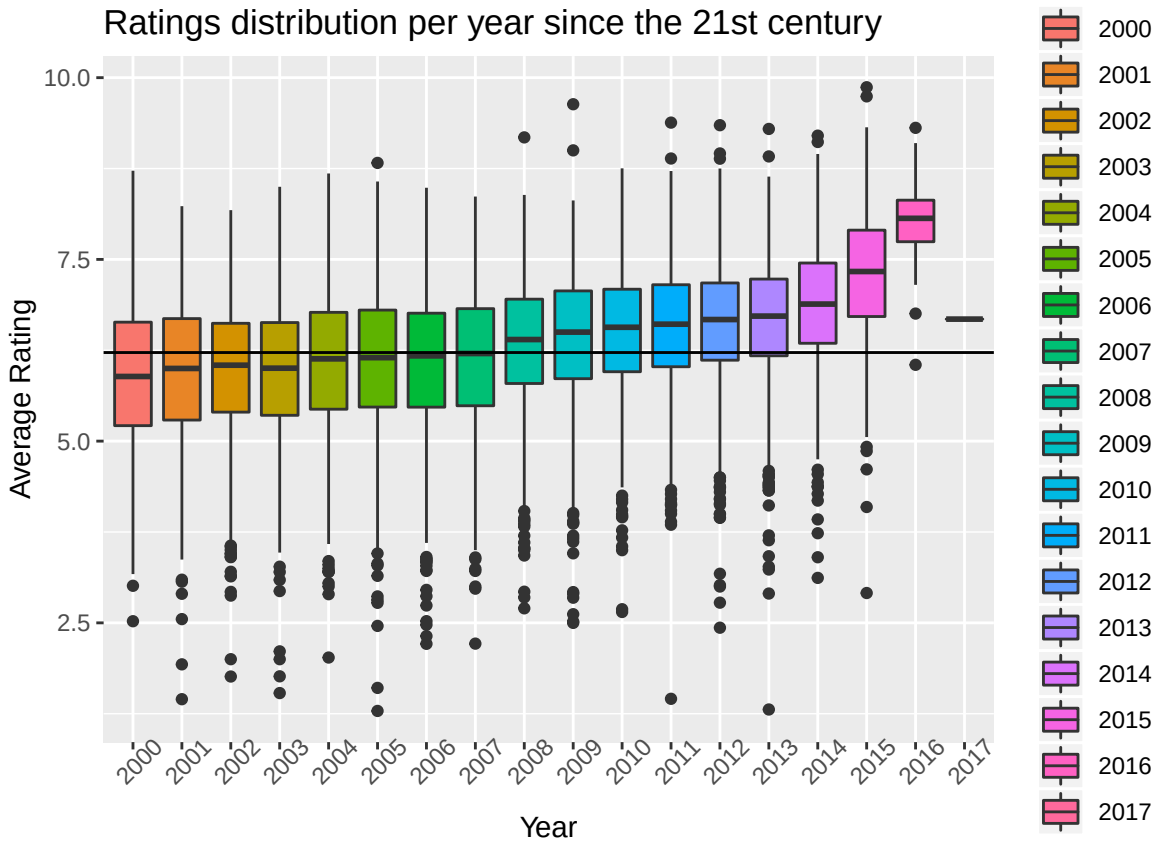
Rating distribution

The graph below shows us that ratings follow a gaussian distribution. The vertical black line represents the average of ratings



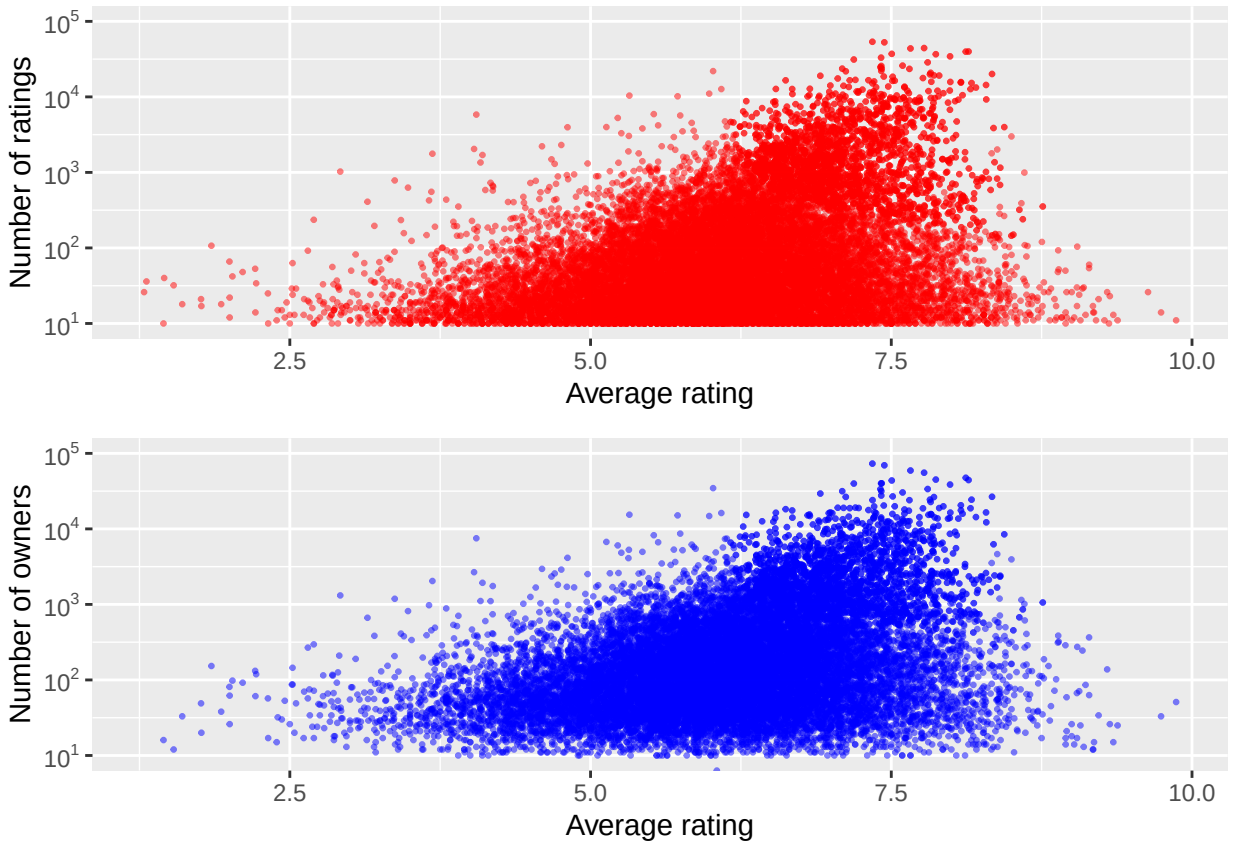
Ratings distribution per year since the beginning of the 21st century

The graph below shows the distribution of ratings per year since the beginning of the 21st century. The horizontal black line shows the all time average of those ratings.



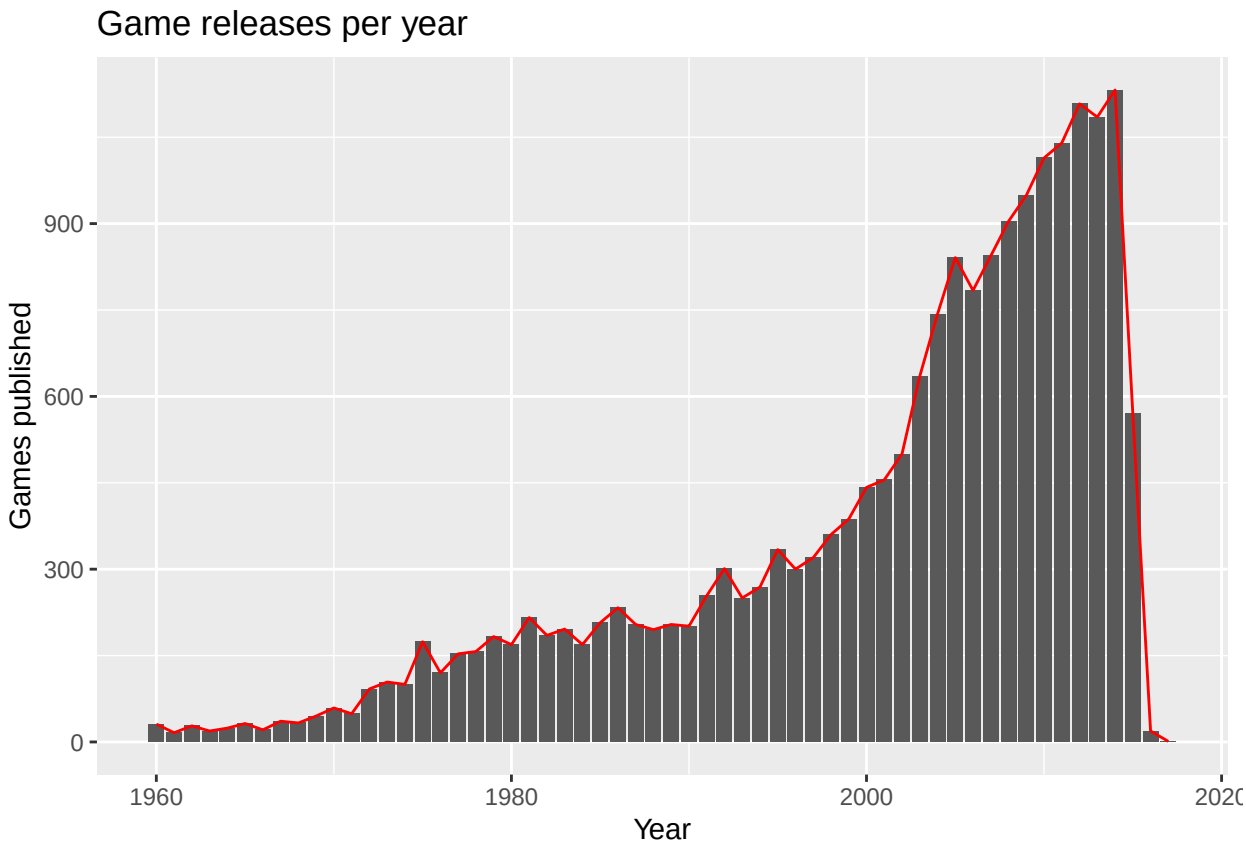
Popularity and ratings

In the graph below we observe a positive correlation between the total number of owners and the number of ratings and their influence on the average rating of a game. Hence, one can make the assumption that people in the community tend to prefer high-rated games. It is fair to assume that highly rated games are the most popular thus selling better.



Game releases over time

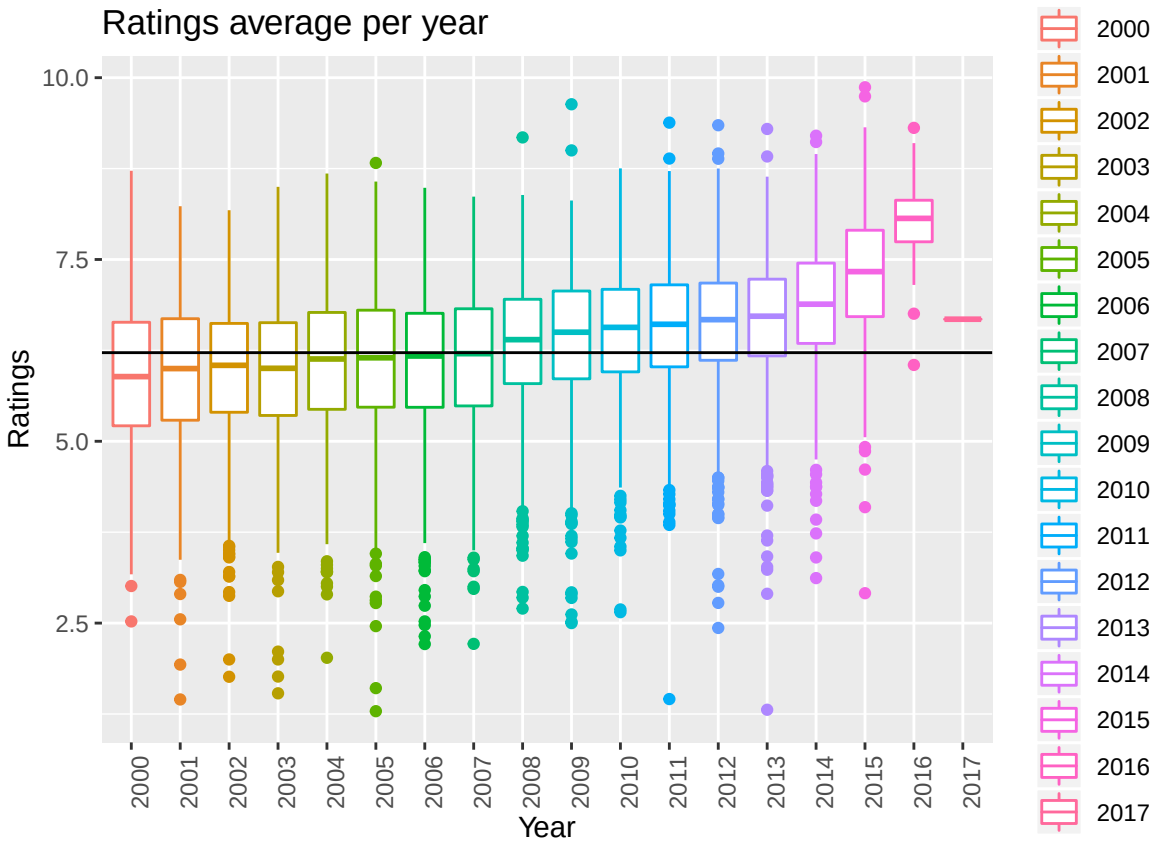
The following bar chart and line demonstrate that since 1960 untill 2014 the amount of games released per year kept increasing. However, from 2015 onwards the amount of games released per year fell drastically.



Game ratings since the beginning of the 21st century

The boxplot graph below shows the average ratings of games released since the beginning of the 21st century. The horizontal black line represents the all time ratings average.

We can observe that the average of ratings has been steadily increasing.

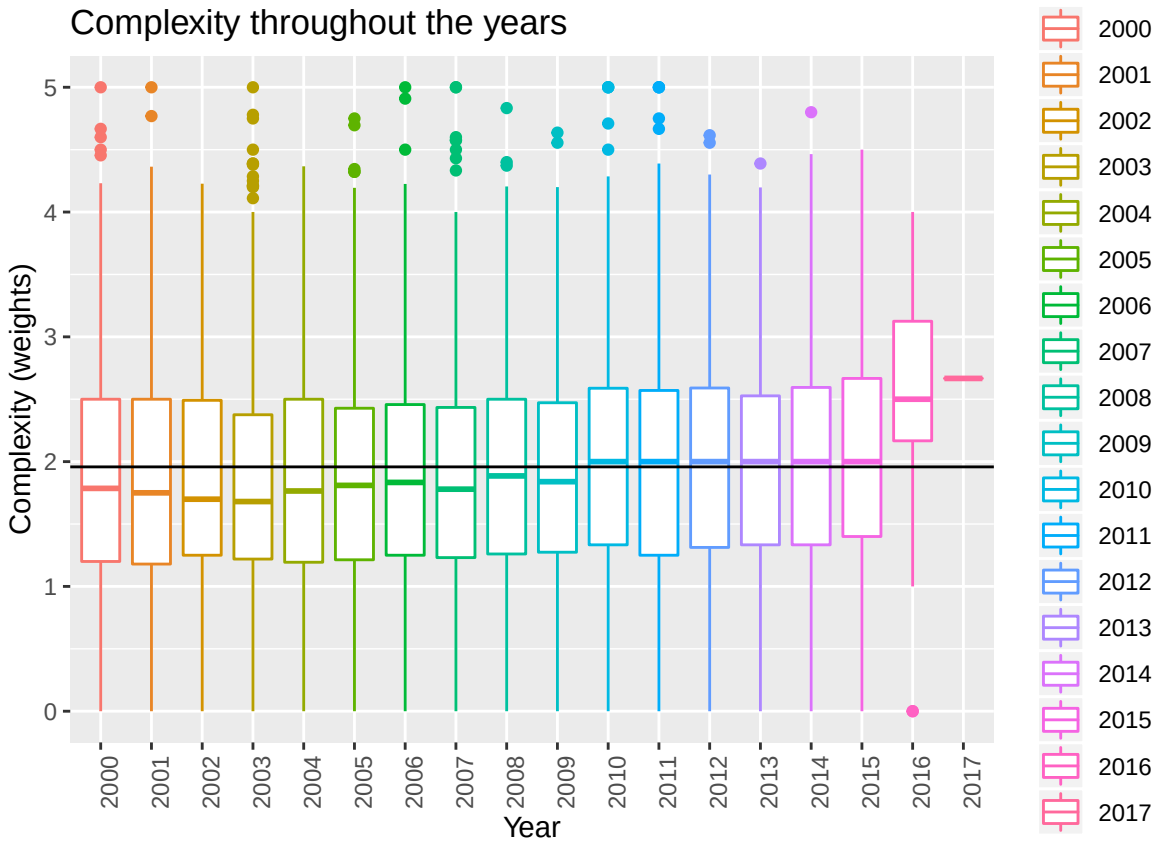


Complexity of games since the beginning of the 21st century

Game complexity is hard to defined and measure. However, the BoardGames data set has a 'weight' attribute, stored in the columns 'total_weights' and 'average_weight'. It is a score given by users that provides a reduced sense of the complexity of a game, based on their perception.

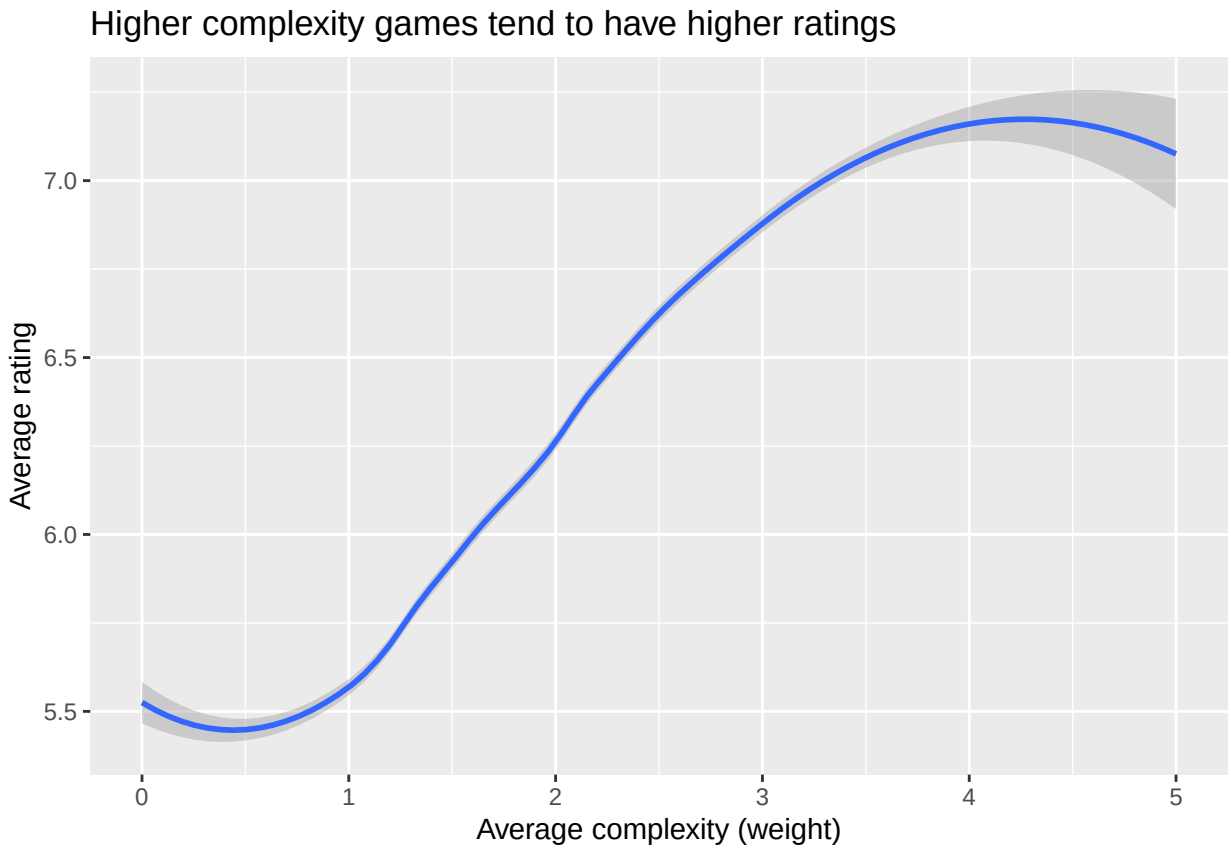
The boxplot graph below shows the average weight (or complexity) of games released since the beginning of the 21st century. The horizontal black line represents the average of the weights.

We can observe that the complexity of games has been steadily increasing.



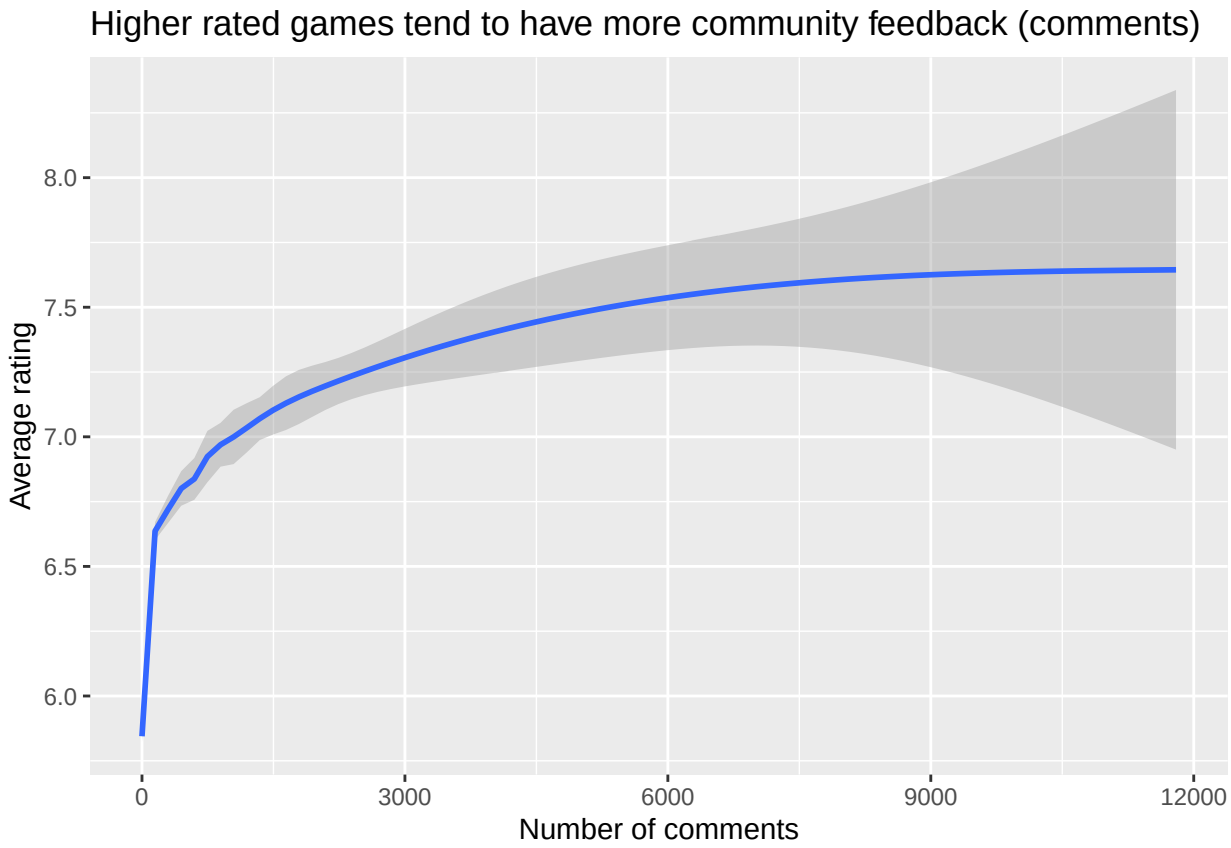
Complexity and rating

The line below shows that higher complexity games usually have an higher rating as well. This does not hold for extremely simple and extremely complex games.

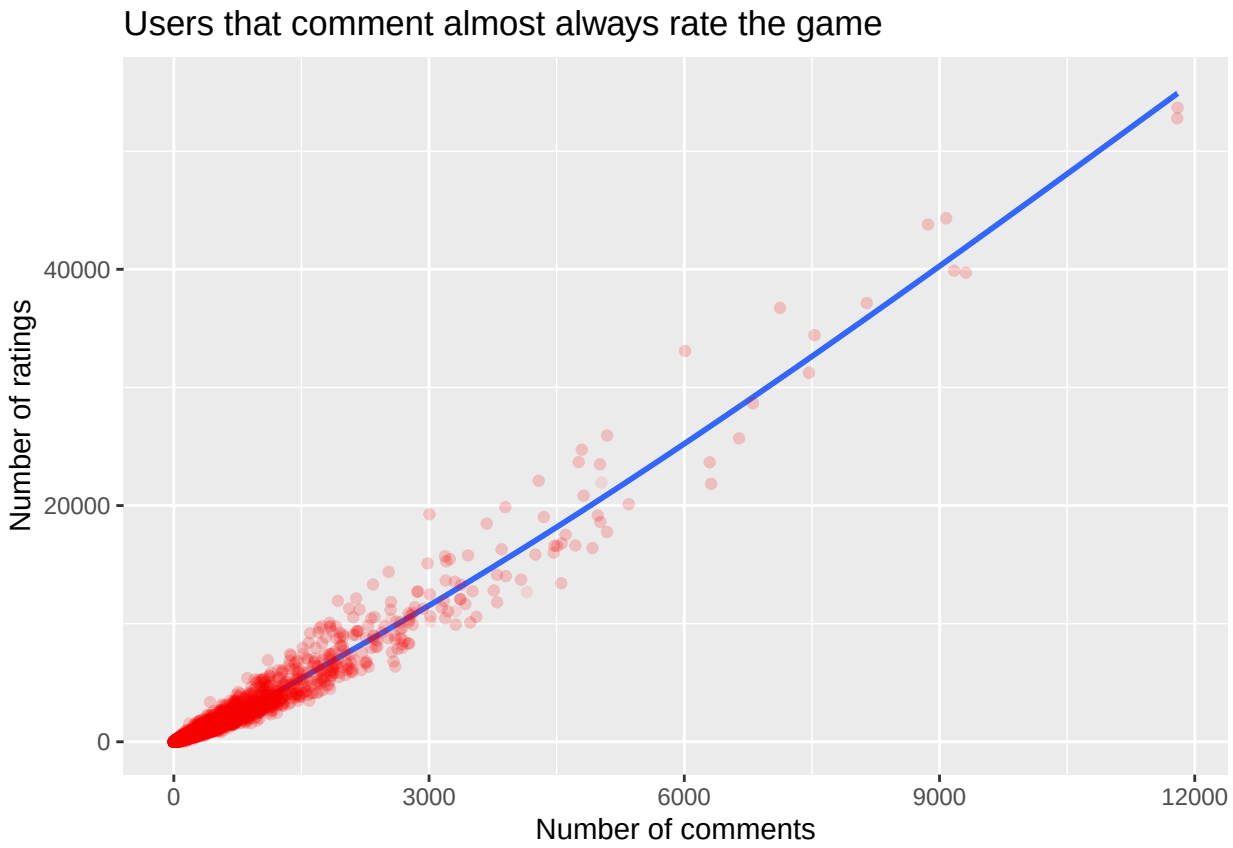


Game comments and rating

Higher rated games appear to have more comments, hence spiking more engage from the board game community.



Also, as expected, the number of comments follows the number of ratings closely, as most likely each person comments once and rates once.

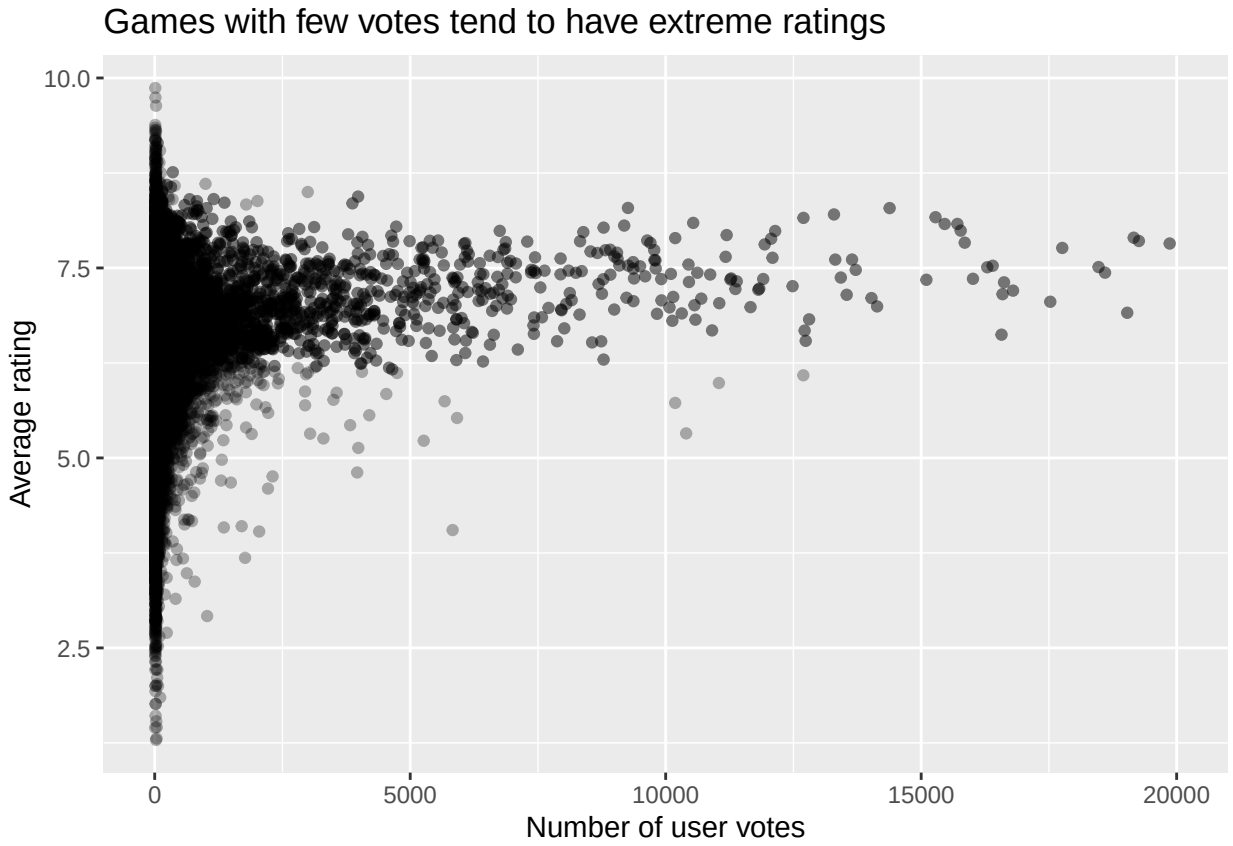


Future improvements

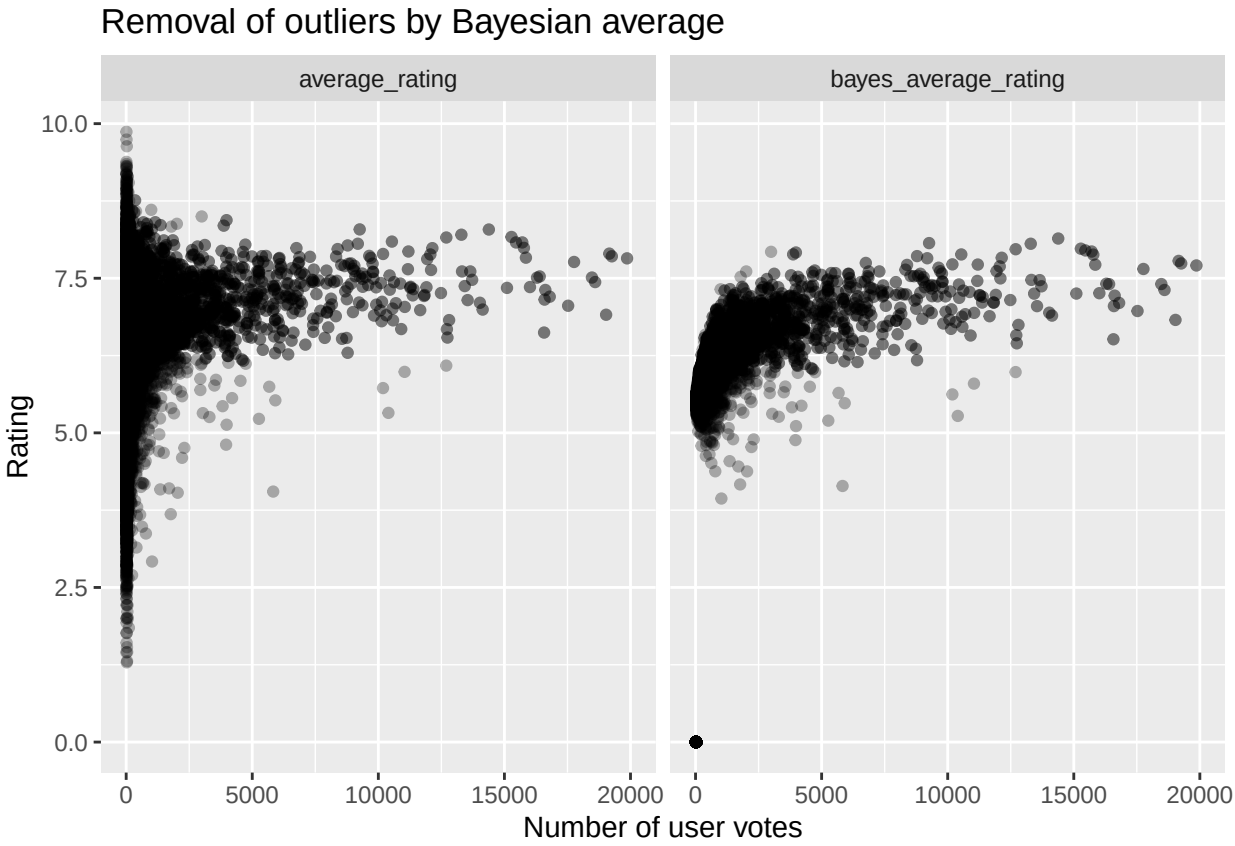
A non explored concept was that of bayesian average. Bayesian average was later discovered during this project development's phase. Future exploratory analysis should be made using this feature.

Average ratings and outliers (few extreme votes)

In the graph below we observe that board games with few votes tend to have overall better ratings. A massive amount of points representing ratings are concentrated in the region of few voted games.



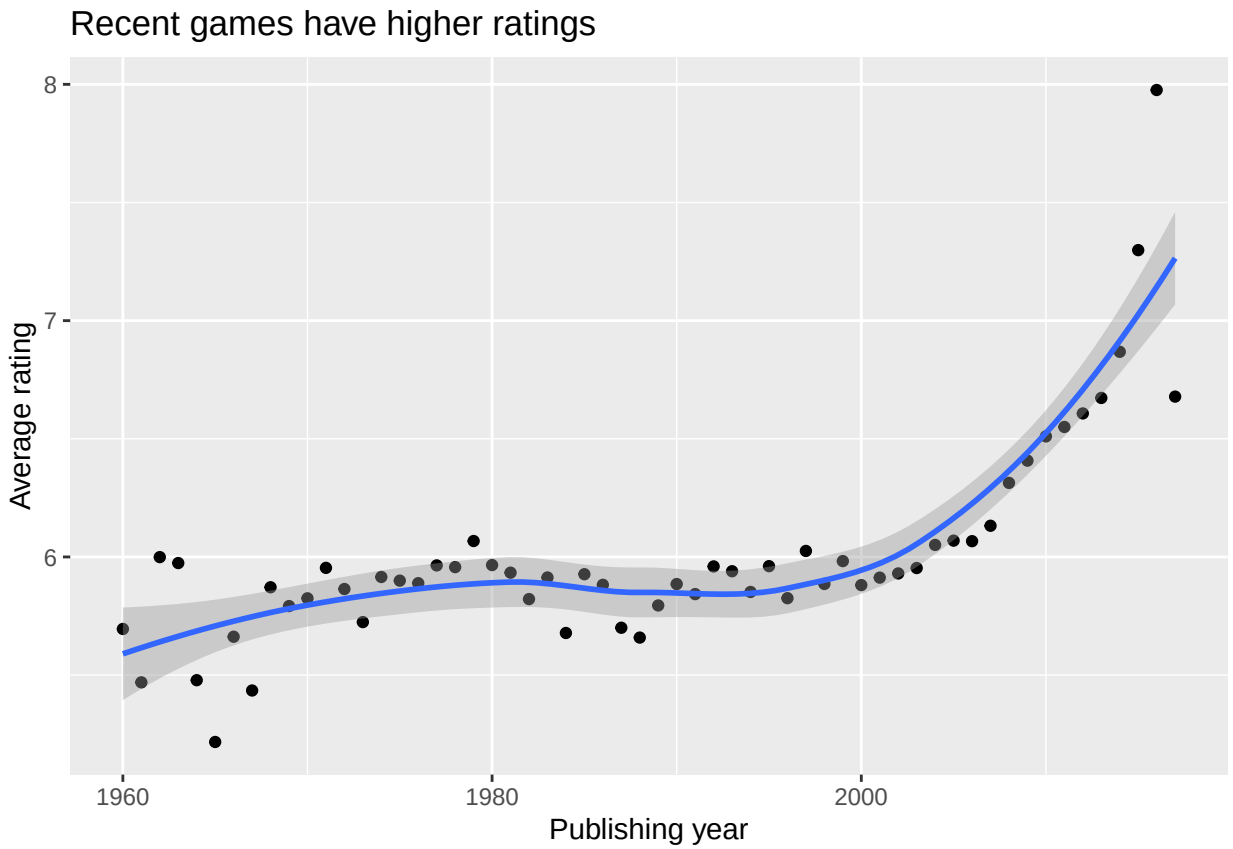
Even though we filtered games with 10 or less votes, there is a *"smoother"* way of dealing with outliers. That is done using the Bayesian average. In Bayesian statistics we start out with a prior that represents our a *"priori"* assumptions. When evidence comes in we can update this prior, computing a so called posterior that reflects our updated belief. Thus if we have an unrated game we assume its average. If not, the ratings will have to convince us otherwise. As seen below, this removes outliers fairly well.



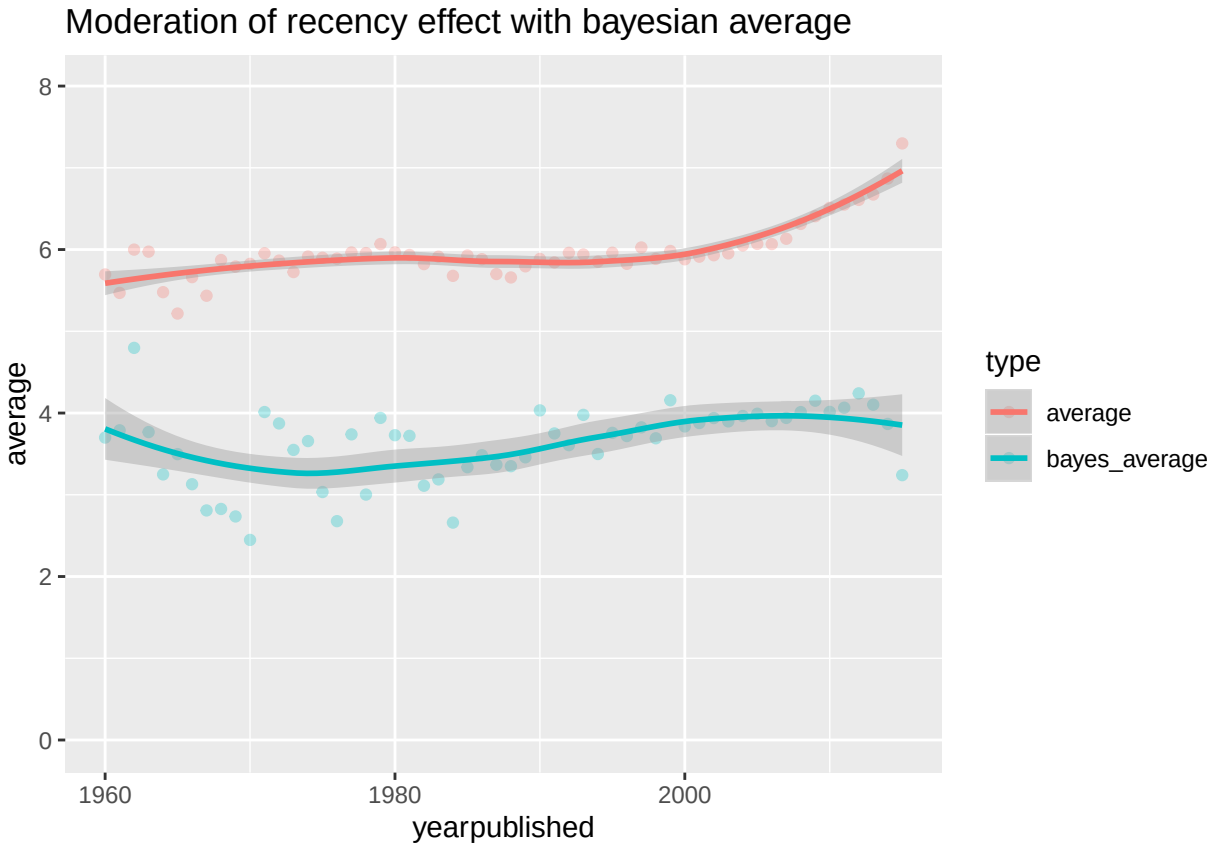
In the graph above we see that higher rated games are better distributed across different number of votes when using the bayesian average rather than the average rating.

Recent games bias

Newer games tend to have higher ratings, due to their populatiry as seen in the graph below.



Using Bayesian average will also prove beneficial here since it also *smooths* the recency effect, as seen below.


Moderation of recency effect with bayesian average
8
6
4
2
0
average
1960
1980
2000
yearpublished
type
average
bayes_average