Board Game Analysis

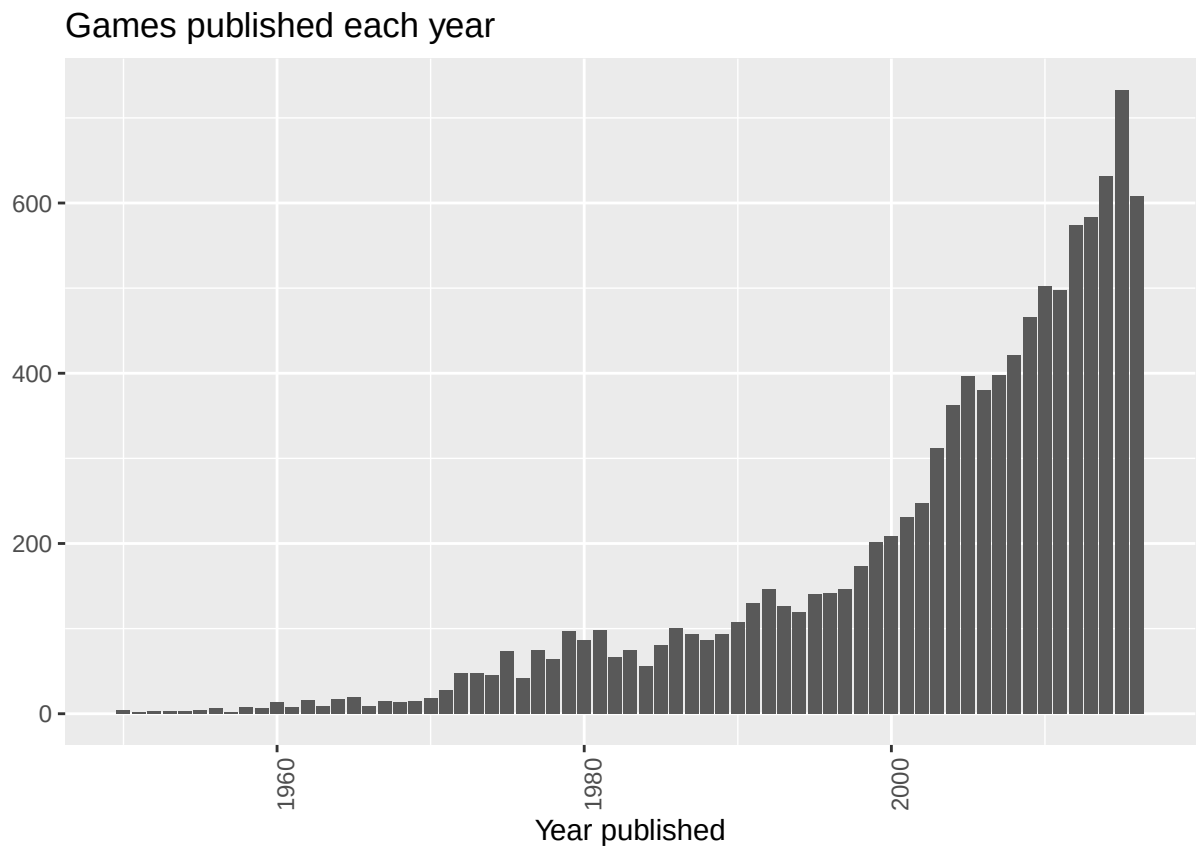
Initialise the libraries and dataset

```
library(tidyverse)
library(reshape2)
board_games_raw <- read_csv('board_games.csv')
```

Number of games published each year

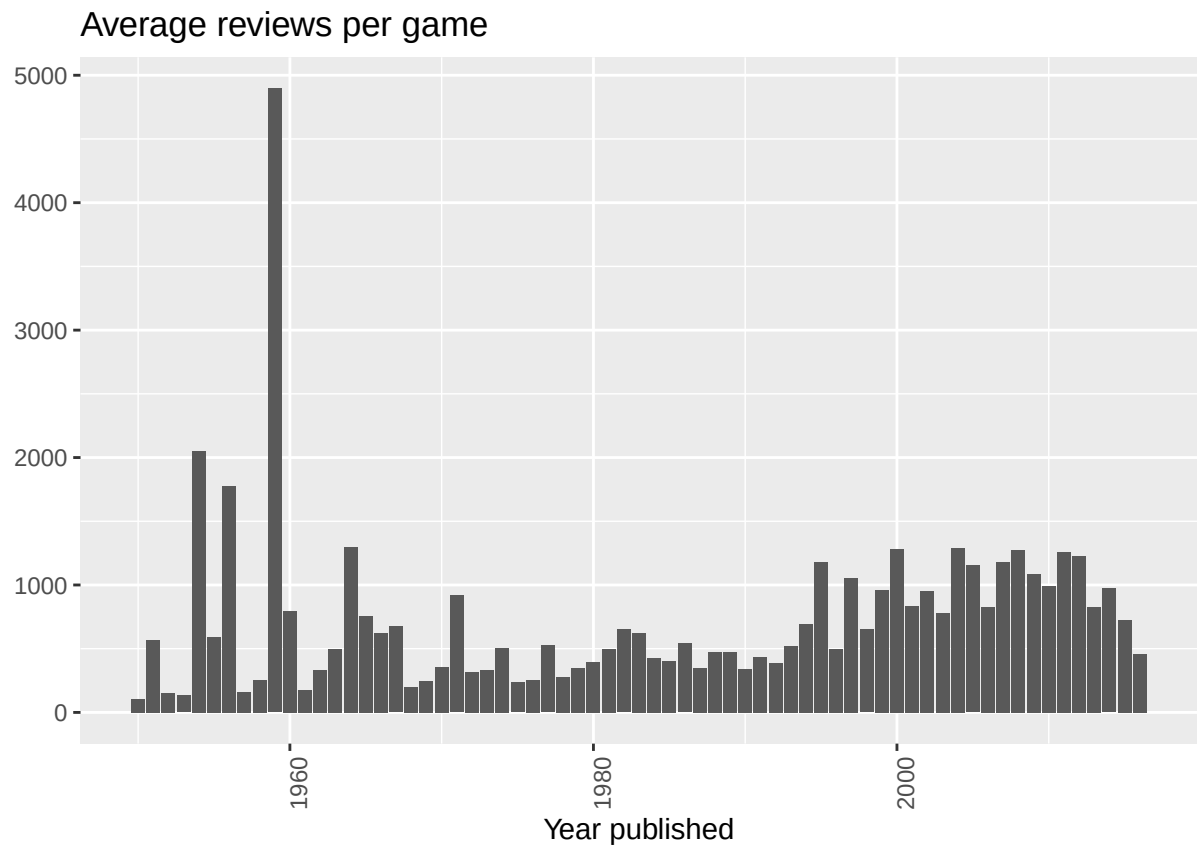
There are 10532 games in the dataset, published over many years. First assessment is to see how many games are in each year and how many reviews have been carried out

```
p <- ggplot(board_games_raw, aes(x=year_published)) +
  geom_bar(stat = "count") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle('Games published each year') +
  xlab('Year published') + ylab('')
p
```




```
p <- ggplot(board_games_raw %>%
              group_by(year_published) %>%
              summarise(avg_reviews_per_game=mean(users_rated))) +
  aes(x=year_published, y=avg_reviews_per_game) +
  geom_bar(stat = "identity") +
```

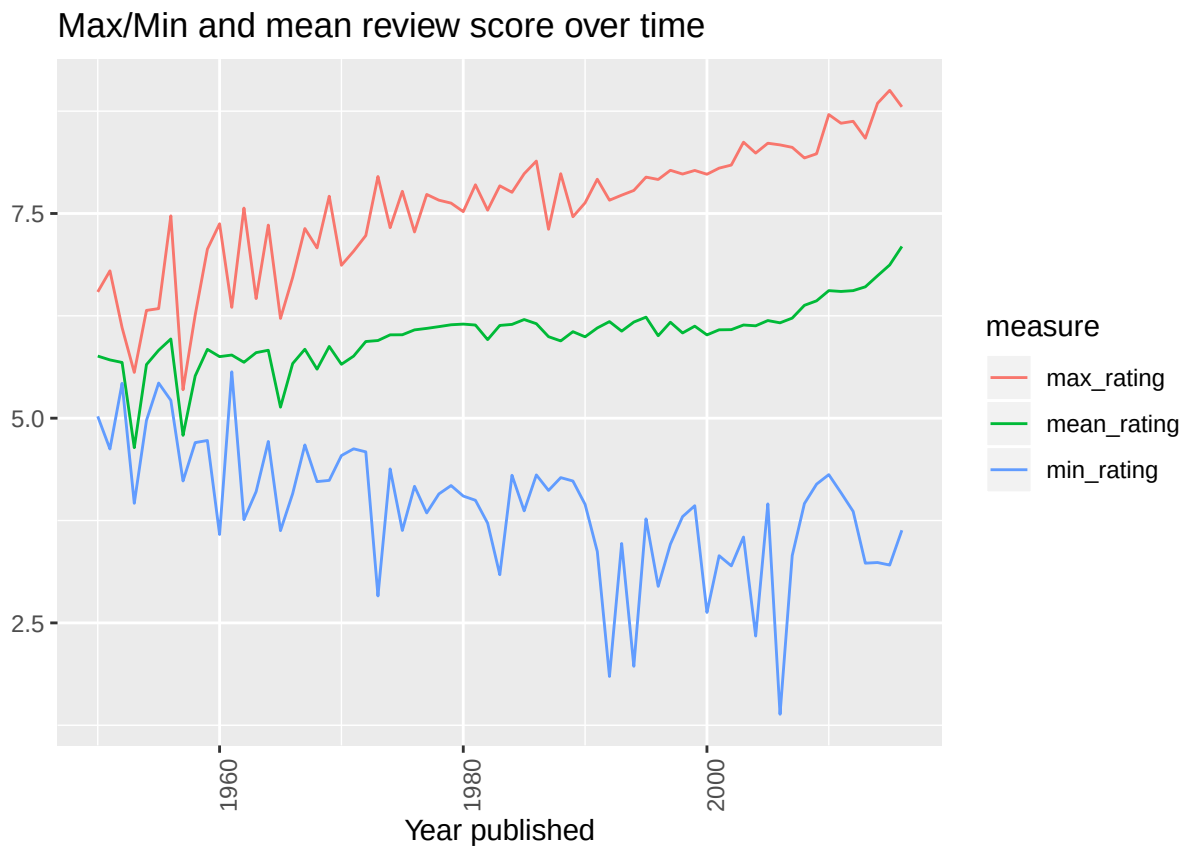
```
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle('Average reviews per game') +
  xlab('Year published') + ylab('')
p
```



There are a healthy number of reviews of games from all years, although later games have roughly twice as many reviews per game as earlier years. A big outlier pre 1960 to be investigated

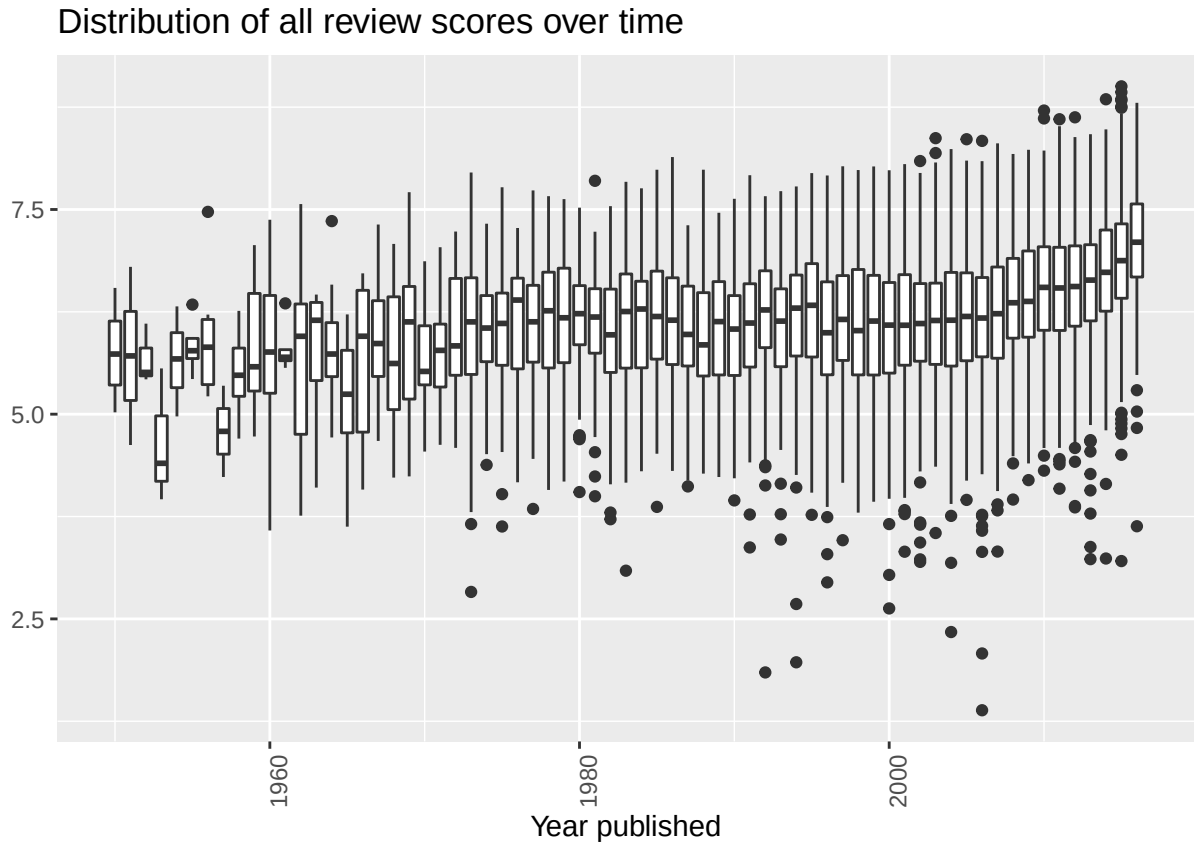
Next look at the average rating over time.

```
p <- ggplot(board_games_raw %>%
              group_by(year_published) %>%
              summarise(mean_rating=mean(average_rating),
                        max_rating=max(average_rating),
                        min_rating=min(average_rating)) %>%
              gather('measure','value', 2:4)) +
  aes(x=year_published, y=value, colour= measure) +
  geom_line() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle('Max/Min and mean review score over time') +
  xlab('Year published') + ylab('')
p
```



Clearly the newer games are rated more highly, but looks like there is also a greater range of scores. Maybe a boxplot would show these differences

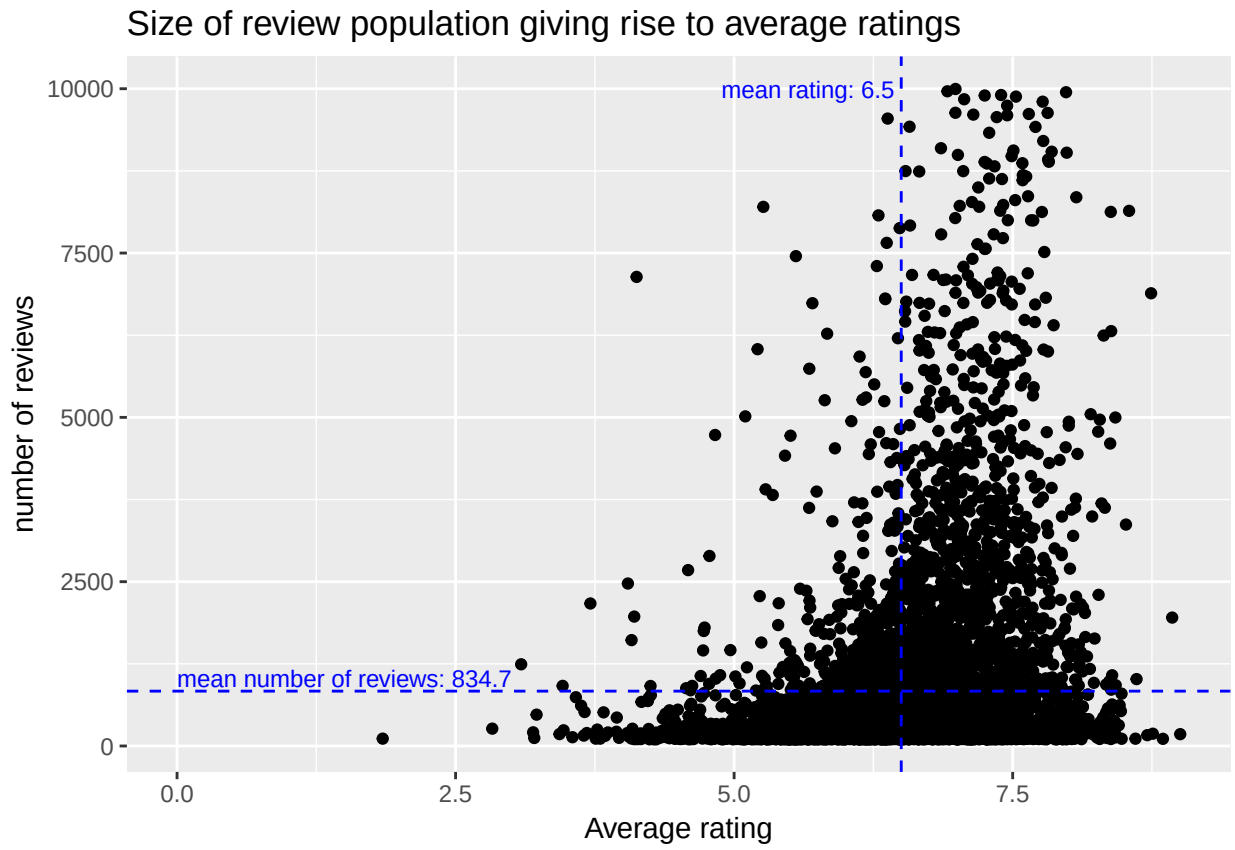
```
p <- ggplot(board_games_raw ) +
  aes(x=year_published, y=average_rating, group=year_published) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle('Distribution of all review scores over time') +
  xlab('Year published') + ylab('')
p
```



Finally, it's worth investigating if there is a relationship between the number of reviews and the average review score

```
# remove a few outliers that have a very high number of reviews
user_reviews <- board_games_raw %>%
                filter(users_rated>100, users_rated<10000)

mean_review_score <- round(mean(user_reviews$average_rating),1)
max_reviews <- max(user_reviews$users_rated)

mean_user_reviews <- round(mean(user_reviews$users_rated),1)



p <- ggplot(user_reviews ) +
  aes(y=users_rated, x=average_rating, group=year_published) +
  geom_point() +
  ggtitle('Size of review population giving rise to average ratings') +
  xlab('Average rating') + ylab('number of reviews') +
  geom_vline(xintercept = mean_review_score, linetype = "dashed", colour="blue") +
  annotate("text", label=paste0("mean rating: ", mean_review_score, " "),
           x=mean_review_score,  y=max_reviews,
            hjust=1, colour="blue", size=3, colour="black") +
  geom_hline(yintercept = mean_user_reviews, linetype = "dashed", colour="blue") +
  annotate("text", label=paste0("mean number of reviews: ", mean_user_reviews, " "),
           x=0,  y=mean_user_reviews+200,
            hjust=0, colour="blue", size=3, colour="black")
```

```
p
```



It would seem that popularity could be defined by a combination of the number of reviews and the average rating. There are quite a few games who have a high avergae rating and a large number of reviews.

We could define the four quadrants as follows:

- *popular* higher average rating and higher number of reviews than mean
- *disappointing* higher number of reviews of mean but lower average rating
- *niche* higher average rating but lower number of reviews
- *unpopular* lower average rating and lower number of reviews than mean

This column can be added to the raw dataset

```
# add a popularity categorisation column
board_games_raw  <- board_games_raw %>%
                mutate(popularity = if_else(average_rating>mean_review_score & users_rated > mean_user_
                                    if_else(average_rating>mean_review_score & users_rated < mean_user_
                                    if_else(average_rating<mean_review_score & users_rated > mean_user_
                                            'unpopular'))))
```

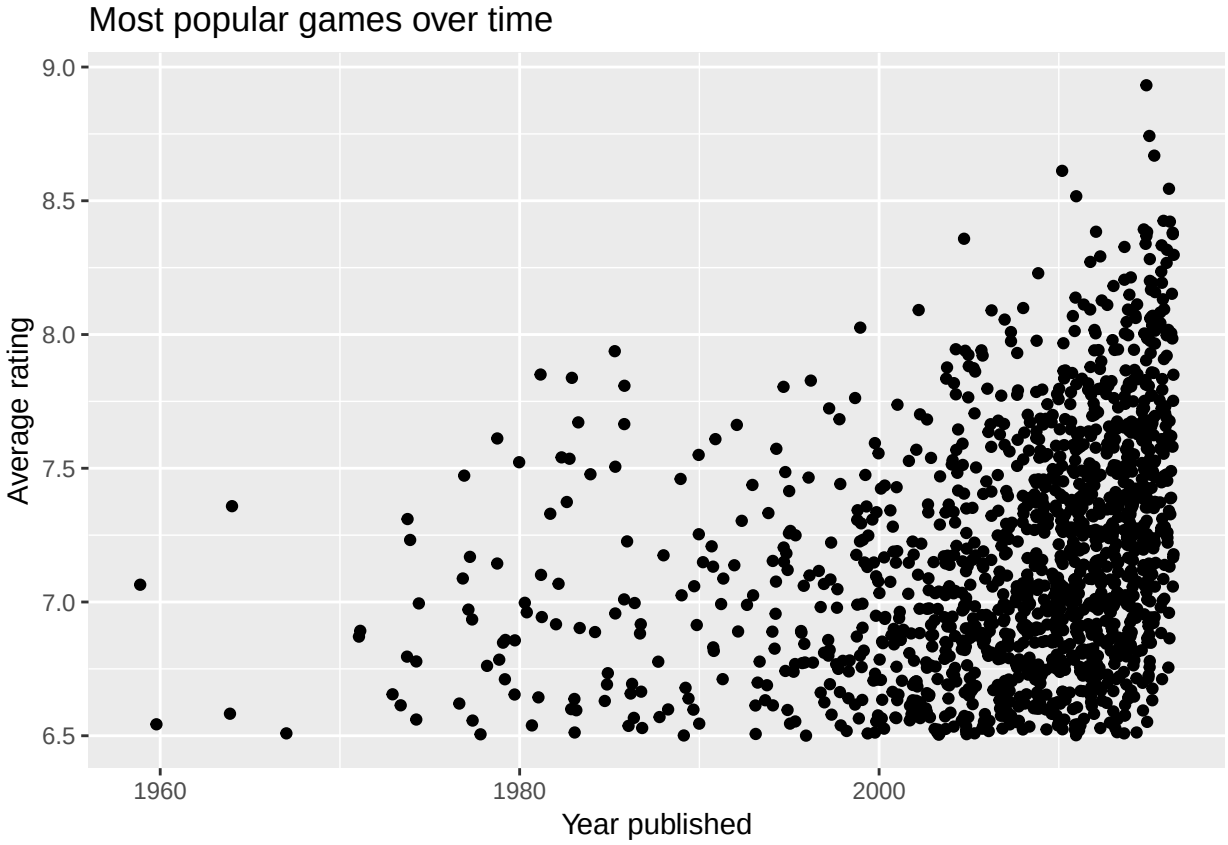
which produces the following classification

```
##
## disappointing         niche       popular     unpopular
##           452          3280          1413          5387
```

Popular games over time

We can filter on this new classification and re-examine the average scores over time of the most popular games

```
p <- ggplot(board_games_raw %>%
              filter(popularity=='popular') ) +
  aes(x=year_published, y=average_rating) +
  geom_jitter() +
  ggtitle('Most popular games over time') +
  xlab('Year published') + ylab('Average rating')
p
```



The most popular games are more recent but there is a reasonable number that have had appeal for decades.

Categories of game

Each game is assigned one or more categories. We can use this to assess