ETC5521 Assignment 1

Ruimin Lin Rahul Bharadwaj

2020-08-27

This assignment is for ETC5521 Assignment 1 by Team Echidna comprising of Ruimin Lin and Rahul Bharadwaj.

Introduction and Motivation

Board Game has been a type of leisure that people have enjoyed from a very long time even before computers and video-games existed and has gone through enormous evolution ever since its inception. Board Games enables a way for people to socialize, reducing stress under such a fast-moving society, and paves way for an extensive brain exercise. Being a popular choice of leisure, what makes board games great? What is the reason for Board Games to have survived in a world of Virtual Reality games? In other words, what are the common characteristics of top ranked board games? What are the best board games in terms of average rating?

The original board games data used in this report is obtained from the Board Game Geek database, and is cleaned and shared by Thomas Mock.

The tidy dataset consists of 22 columns and 10532 rows, in which there are 22 variables and 10532 observations. It consists of data such as max/min playtime, max/min players, min age of players that can play, game designer, game publisher, mechanics of the game and a lot more. One thing to notice is that even though the data set is tidy, we still find observations in variables like `category`, `family`, `mechanic` to be messy and repetitive, which may limit our ability to explore these variables.

Data Description

The aim of this exploratory analysis is to find out what factor affects the average rating of board games. This would give insights as to what board games are most popular and the characteristics these board games share. Therefore, we have articulated the following questions to help us with further exploration of the board games data.

Primary Question:

What are the common characteristics of top ranked board games?

Secondary Questions:

1. What are the top 10 ranked board games?
2. How do variables like min/max playtime, min/max players, or min_age affect the average rating?
3. Which game designer was most successful in producing popular games? Which publisher published the most popular games?

The variables included in the data are as follows:

```
## Rows: 10,532
## Columns: 22
## $ game_id        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1...
```

```
## $ description    <chr> "Die Macher is a game about seven sequential politic...
## $ image          <chr> "//cf.geekdo-images.com/images/pic159509.jpg", "//cf...
## $ max_players    <dbl> 5, 4, 4, 4, 6, 6, 2, 5, 4, 6, 7, 5, 4, 4, 6, 4, 2, 8...
## $ max_playtime   <dbl> 240, 30, 60, 60, 90, 240, 20, 120, 90, 60, 45, 60, 1...
## $ min_age        <dbl> 14, 12, 10, 12, 12, 12, 8, 12, 13, 10, 13, 12, 10, 1...
## $ min_players    <dbl> 3, 3, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 3, 2, 3, 2, 2...
## $ min_playtime   <dbl> 240, 30, 30, 60, 90, 240, 20, 120, 90, 60, 45, 45, 6...
## $ name           <chr> "Die Macher", "Dragonmaster", "Samurai", "Tal der Kö...
## $ playing_time   <dbl> 240, 30, 60, 60, 90, 240, 20, 120, 90, 60, 45, 60, 1...
## $ thumbnail      <chr> "//cf.geekdo-images.com/images/pic159509_t.jpg", "//...
## $ year_published <dbl> 1986, 1981, 1998, 1992, 1964, 1989, 1978, 1993, 1998...
## $ artist         <chr> "Marcus Gschwendtner", "Bob Pepper", "Franz Vohwinke...
## $ category       <chr> "Economic,Negotiation,Political", "Card Game,Fantasy...
## $ compilation    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "CAT...
## $ designer       <chr> "Karl-Heinz Schmiel", "G. W. \"Jerry\" D'Arcey", "Re...
## $ expansion      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "Elfengold,Elfen...
## $ family         <chr> "Country: Germany,Valley Games Classic Line", "Anima...
## $ mechanic       <chr> "Area Control / Area Influence,Auction/Bidding,Dice ...
## $ publisher      <chr> "Hans im Glück Verlags-GmbH,Moskito Spiele,Valley Ga...
## $ average_rating <dbl> 7.66508, 6.60815, 7.44119, 6.60675, 7.35830, 6.52534...
## $ users_rated    <dbl> 4498, 478, 12019, 314, 15195, 73, 2751, 186, 1263, 6...
```

The explanation of variables and variable types are provided to enable a better understanding of the variables in board games data set.

- game_id: ID of a particular game, the game_id should be a character vector(categorical) instead of a double vector mentioned in the table above.
- description: Game description, a character vector.
- image: URL image of the game, a character vector.
- max_players/min_player: maximum/minimum number of recommended players, double vectors.
- max_playtime/min_playtime: maximum/minimum recommended playtime, double vectors.
- min_age: recommended minimum player age, double vectors.
- name: name of the game, a character vector.
- playing_time: average playtime of a game, a double vector.
- thumbnail: URL thumbnail of the game, a character vector.
- year_published: year the game was published, a double vector.
- artist: artist for game art, a character vector.
- category: categories of the game, a character vector.
- compilation: name of compilation, a character vector.
- designer: game designer, a character vector.
- expansion: name of expansion pack (if any), a character vector.
- family: family of game - equivalent to a publisher, a character vector.
- mechanic: how game is played, a character vector.
- publisher: company/person who published the game, a character vector.
- average_rating: average rating from 1 to 10 on the website(Board Games Geek), a double vector.

- users_rated: number of users rated the game, a double vector.

To ensure the reliability of the board game ratings, the data is limited to games with at least 50 ratings and for games between 1950 and 2016. The site's database has more than 90,000 games with crowd-sourced ratings.

The original board games data set consists of 90400 observations, and 80 variables. Therefore, data cleaning and wrangling is necessary to enable better analysis procedure. Thomas has replaced long and complicated variable names like `details.description` in original data to `description` using `janitor::clean_names` and `set_names` function, which avoids messy code writing. In addition, he has eliminated around 50 variables using the `select` function and that leaves 27 variables at this stage.

The data set is then filtered to board games published from 1950 to 2016, with at least 50 users rated. 'NA' values in variable `year_published` is also omitted. Thomas then excludes variables that may not be useful for the analysis, such as `attributes_total`, `game_type` etc., which ultimately, leaves us with a tidy data set (22 variables and 10532 variables) that is relatively concise and convenient for further exploration.

Analysis and Findings

Initial Data Analysis

- Initial Data Analysis is a process which helps one get a feel of the data in question. This helps us have an overview of the data and gives insights about potential Exlporatory Data Analyis (EDA).
- Initial data analysis is the process of data inspection steps to be carried out after the research plan and data collection have been finished but before formal statistical analyses. The purpose is to minimize the risk of incorrect or misleading results.
- IDA can be divided into 3 main steps:
  - Data cleaning is the identification of inconsistencies in the data and the resolution of any such issues.
  - Data screening is the description of the data properties.
  - Documentation and reporting preserve the information for the later statistical analysis and models.
- The plot above @ref(fig:visdatData) clearly visualizes the distribution of data types in our dataset with column in x-axis and number of observations on the y-axis. This gives a concise overview of the data and what columns are useful for analysis. This plot hints that we can use all the numeric columns along with `designer` and `publisher` columns for our analysis.
- The above plot @ref(fig:vismissData) shows the percentage of missing values and where exactly they are missing with x-axis showing columns and the y-axis showing the corresponding observations. We can also observe that each column has a percentage of missing values mentioned which come in handy while deciding what columns not to pick for analysis.
- It is evident that the following columns have missing values and are not of much use for the analysis:
  - `compilation` - 96.11% missing
  - `expansion` - 73.87% missing
  - `family` - 26.66% missing
  - `mechanic` - 9.02% missing
- This is a limitation of the dataset and we frame our questions keeping this in mind.

Questions of Interest

1. What are the top 10 ranked board games?

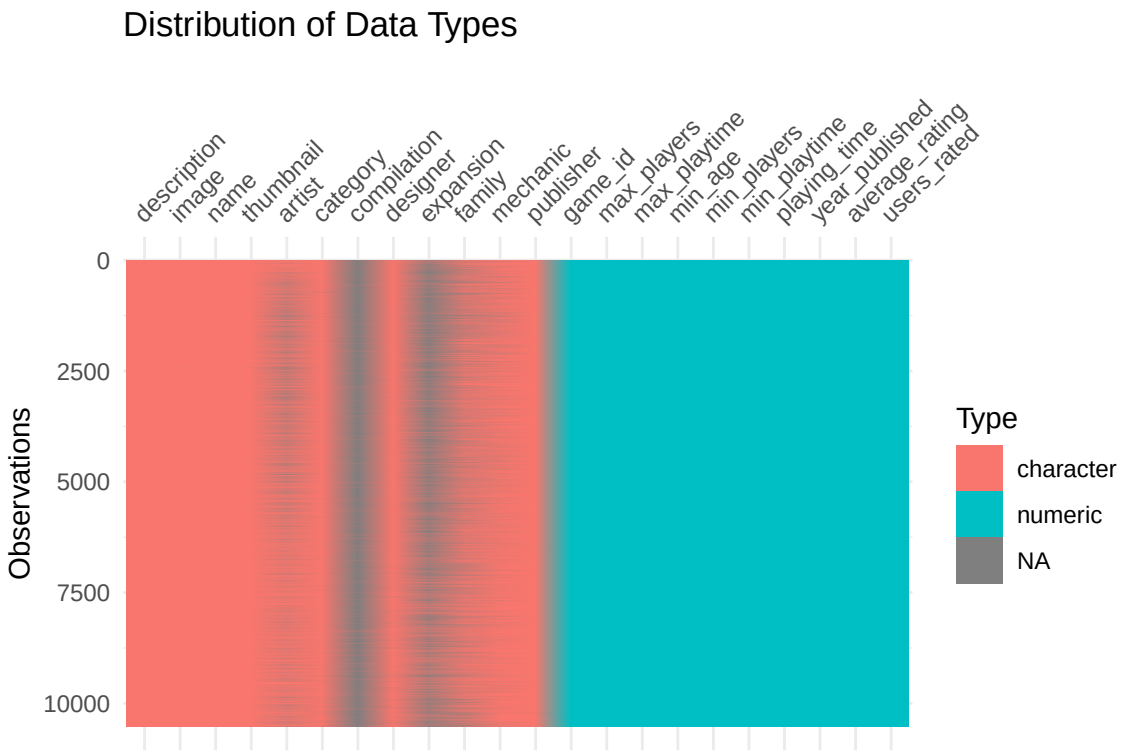


Figure 1: Visualization of Data Types

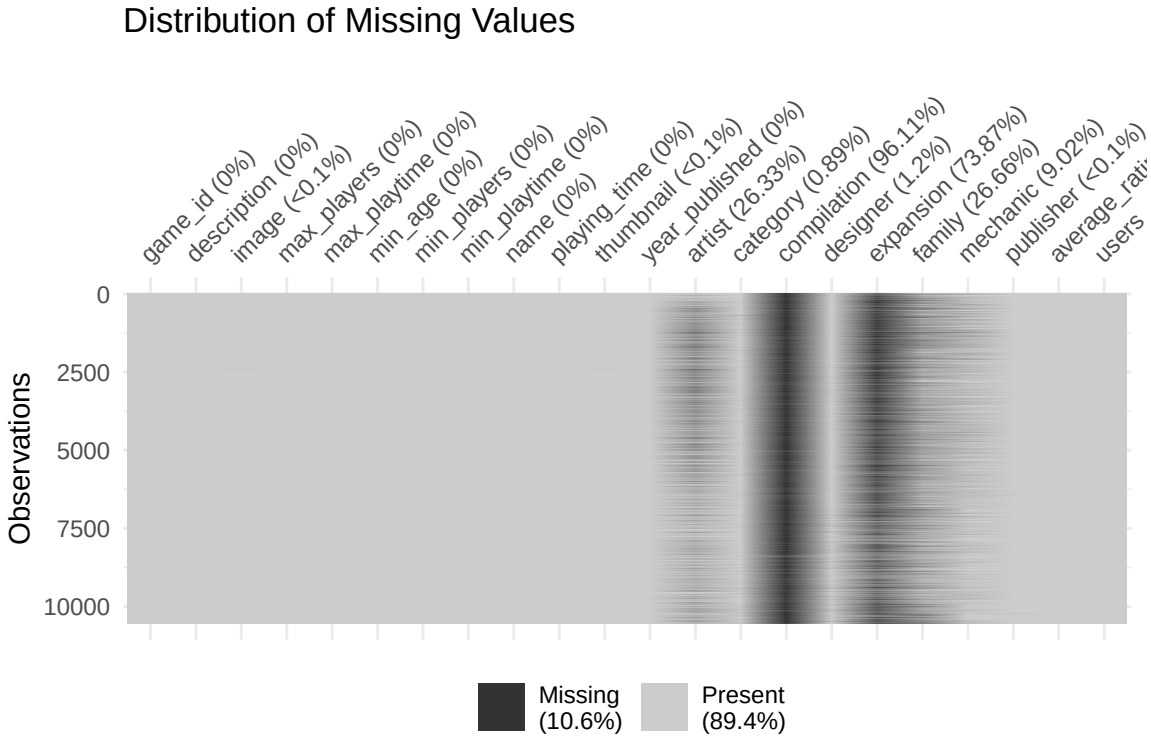


Figure 2: Visualization of Missing Values

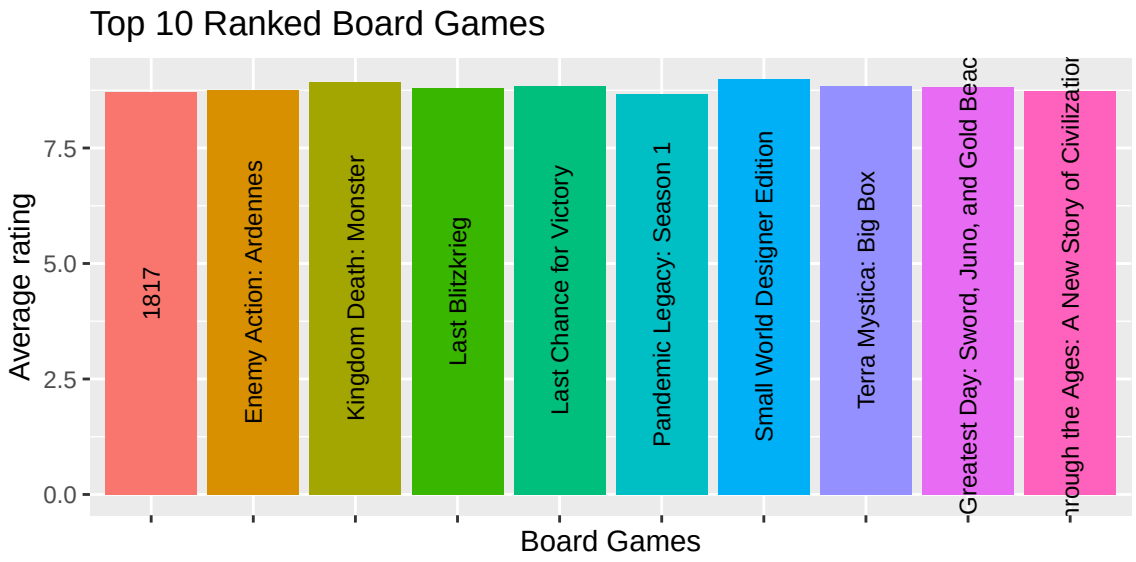


Figure 3: Top 10 ranked board games

name	average_rating	max_playtime	min_playtime	max_players
Small World Designer Edition	9.00392	80	40	6
Kingdom Death: Monster	8.93184	180	60	6
Terra Mystica: Big Box	8.84862	150	60	5
Last Chance for Victory	8.84603	60	60	2
The Greatest Day: Sword, Juno, and Gold Beaches	8.83081	6000	60	8
Last Blitzkrieg	8.80263	960	180	4
Enemy Action: Ardennes	8.75802	600	0	2
Through the Ages: A New Story of Civilization	8.74235	240	180	4
1817	8.70848	540	360	7
Pandemic Legacy: Season 1	8.66878	60	60	4

2. How do variables like min/max playtime, min/max players, or min_age affect the average rating in these top-ranked board games?

- The above plot @ref(fig:visdatTop50) shows a distribution of Data Types in our Top 50 Games dataset with x-axis showing column names and y-axis its corresponding observations.
- It is evident that our selection of columns is appropriate and there are no missing values in our data. Hence, we need not check for missing values through `vis_miss()` function. We can use all these columns for an effective analysis of our questions of interest.
- To have a better idea on the common characteristics of top-ranked board games and ensuring the reliability of the results, we have widened the range to top 50.
- In plot @ref(fig:MaxPlaytime) we can see that there are a few obvious distinct values present, which are:
  - The Greatest Day:Sword, Juno, and Gold Beaches with 6000 minutes max. playtime and an average rating of 8.8308
  - Axis Empires: Totaler Krieg! with 3600 minutes max. playtime and average rating of 8.4194
  - Beyond the Rhine with 3000 minutes max. playtime and average rating of 8.5979
- It is difficult to examine the trend or common characteristics with these outliers presents, therefore,

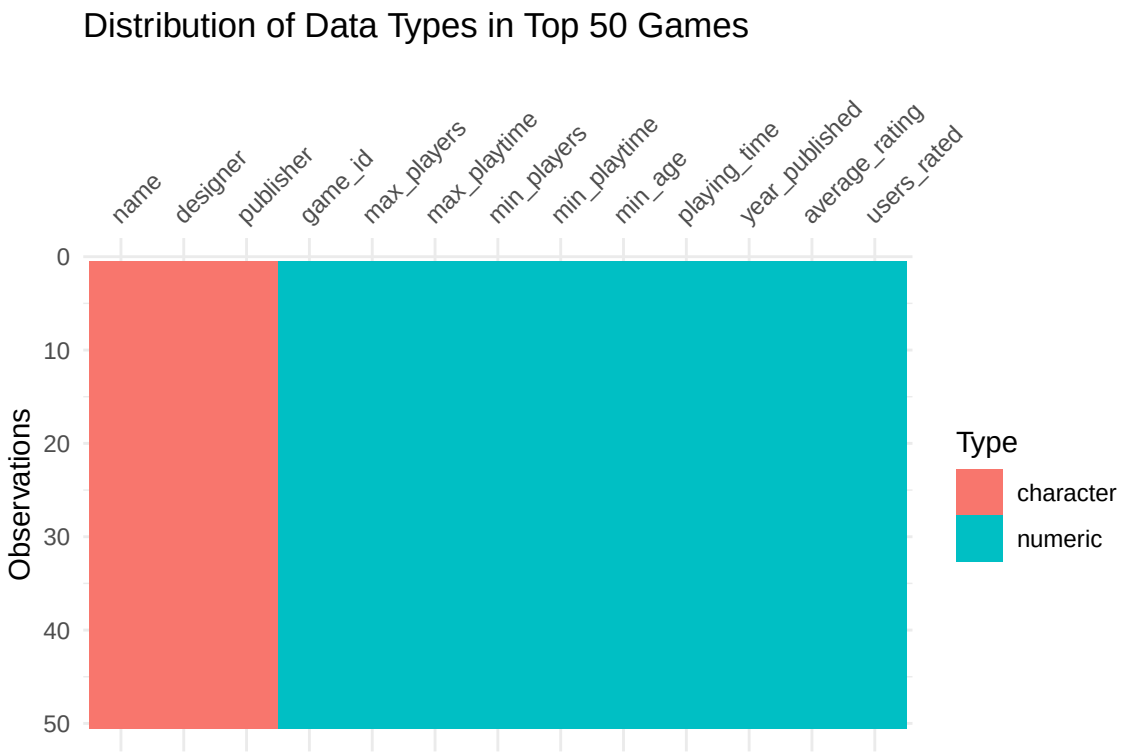


Figure 4: Visualization of Data Types in Top 50 Games

we have limited the maximum playtime to less than xx minutes using the IQR outliers formula. (Q1 - 1.5IQR and Q3 + 1.5 IQR)

```
## # A tibble: 1 x 6
##   minimum    q1 median  mean    q3 maximum
##     <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1       0  82.5   142.  461.   345    6000
```

```
## # A tibble: 1 x 2
##   lower_range upper_range
##         <dbl>       <dbl>
## 1       -311.        739.
```

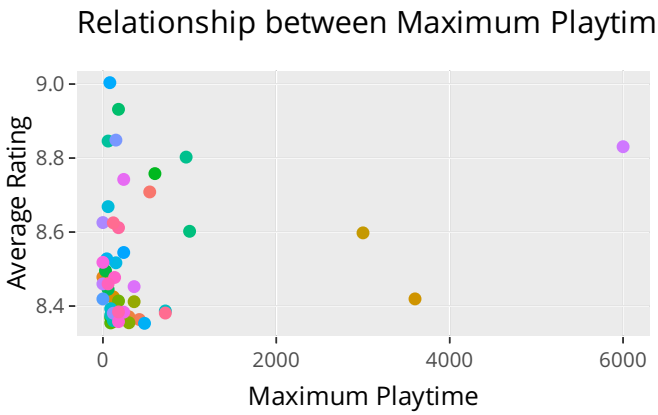
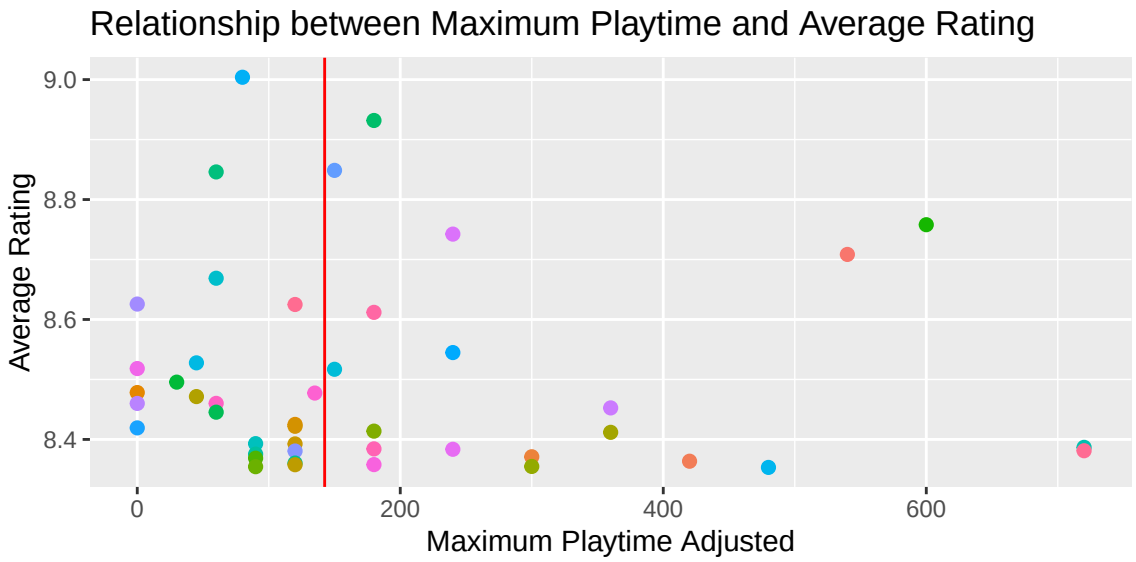
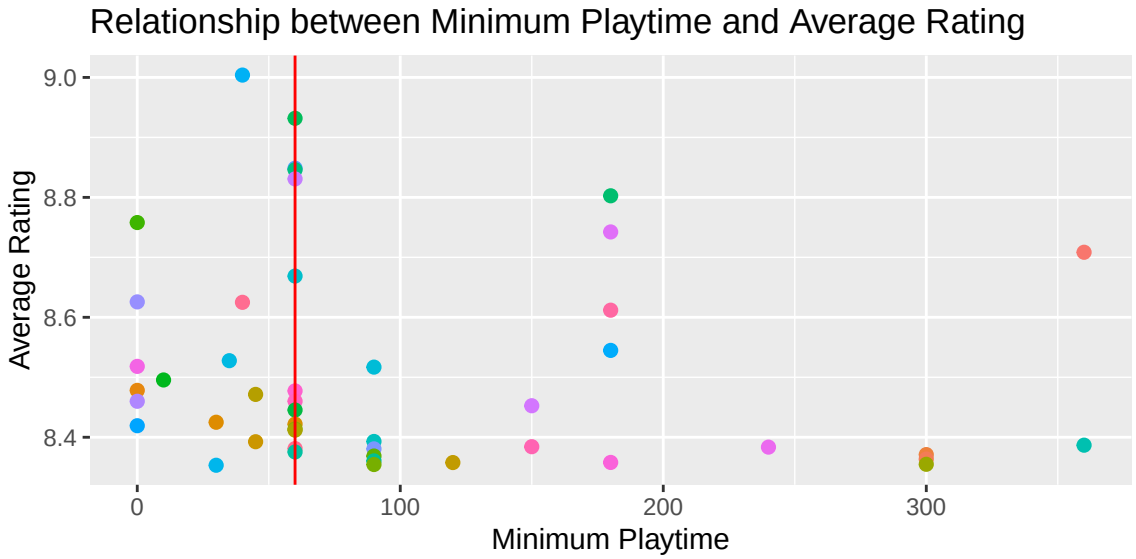


Figure 5: Relationship between Maximum Playtime and Average Rating



Now we can have a clearer picture of where majority of top-50 ranked board games lie in the graph of average rating against maximum playtime. Which, majority of board games lie within the range of 200 minutes of maximum playtime, the highest ranting board game also lies within the range, around 100 minutes of maximum playtime. Another thing to notice is that, for board games that have maximum playtime longer than 600 minutes, the rating is comparatively lower.

Nearly half of high rating board games are crowded in the range of 0-200 minutes, suggesting that people tend to play board games that does not occupy too much leisure time.



- We have implemented the same method to omit the outliers as done previously, the graph demonstrates that in top-50 ranked board games, most of them have a minimum playtime less than 100 minutes.
- In the scatterplot for average rating against minimum players, we observed that most top 50 board games have at least 2 players.
- In the scatterplot for average rating against maximum players, we observed that most top 50 board games have a maximum of 4 or 5 players.
- The figure @ref(fig:MinPlayersPlot) and @ref(fig:MaxPlayersPlot) indicates that majority of high rating

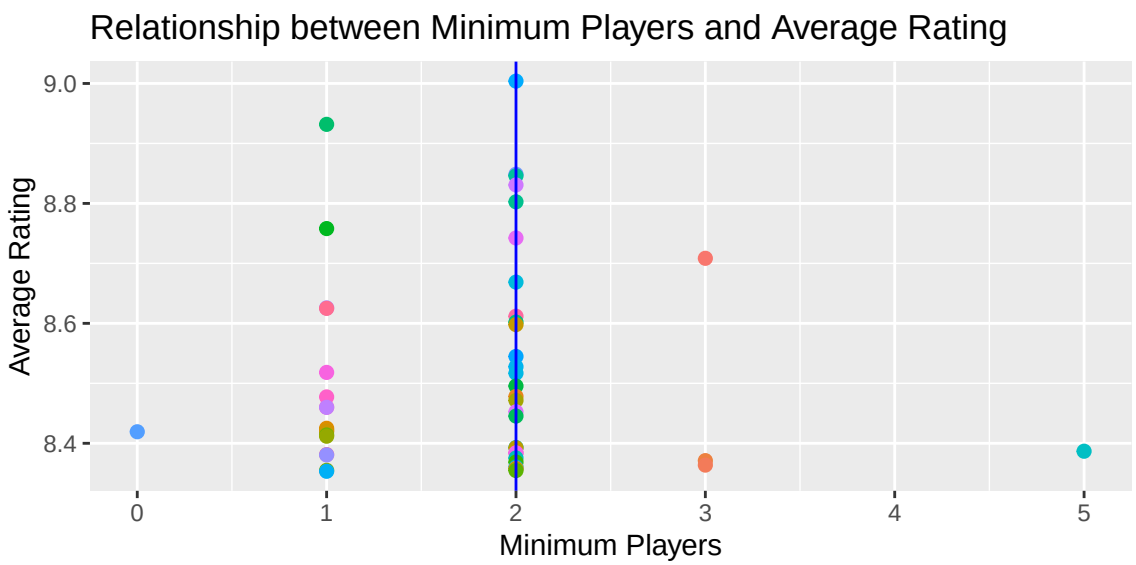


Figure 6: Relationship between Minimum Players and Average Rating

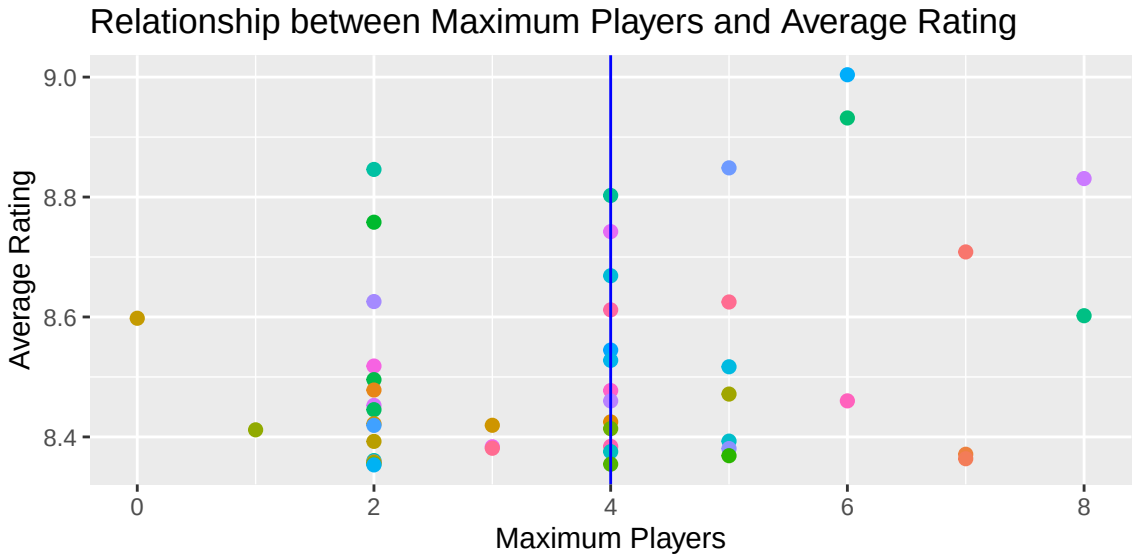
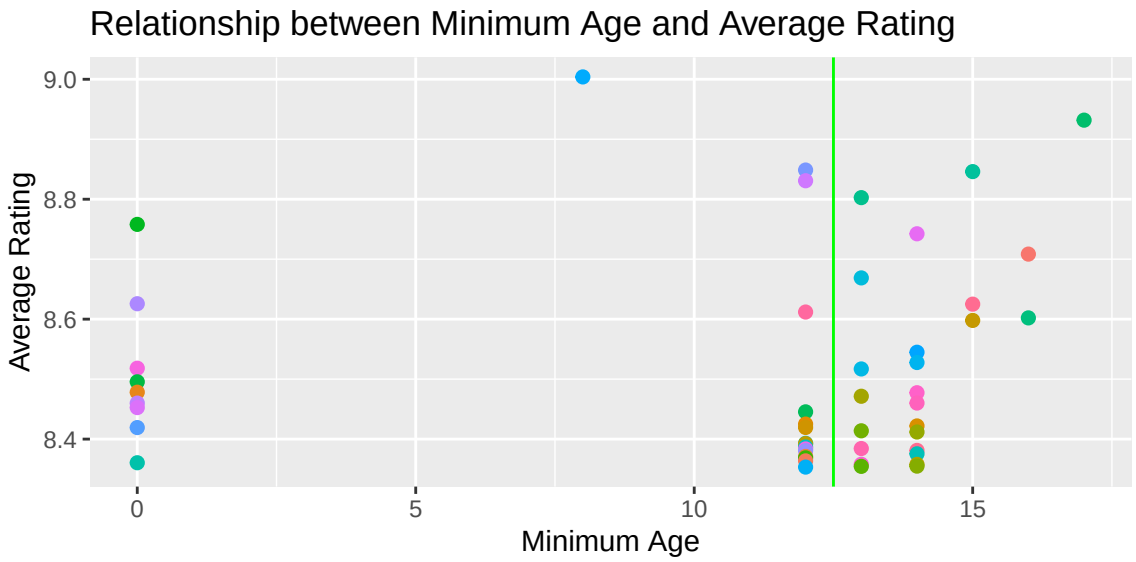


Figure 7: Relationship between Maximum Players and Average Rating

board games have set the players to between 2 and 4/5 players. The limitation of players suggest that people tend to play board games that fulfills their sense of participation, for example, a board game of 8 players may not be as attractive as a board game of 2 players, because a 2-player game has little downtime than a 8-player game, and satisfies each players' sense of participation in the board game.

- On the other hand, it is easier to gather a group of 2-4 people interesting in play board games at leisure time than gathering a group of 8 or more people.



- In the scatterplot for average rating against minimum age of players, we observed that the minimum age set by majority of board games are between 10 - 15.
- All the insights for the top 50 popular games are summarized in the boxplots above as follows:
  - A **maximum of 4 players** and **minimum of 2 players** is most popular in the top 50 games.
  - The **maximum and minimum playtime** seem to be almost close and **range between 60-150 minutes** for top 50 games.
- The above plot @ref(fig:smooth) shows a trend for different attributes against average rating on x-axis. We can get a better idea using this pattern.
- We can observe the following trend for the top 50 rated games as average rating increases -
  - The Minimum Players tends to be around 2 players. The Maximum Players tends to be around 4 and increases up to 6.
  - The Minimum Playtime tends to vary between 60-500 minutes. The Maximum Playtime tends to vary between 150-1000 minutes.

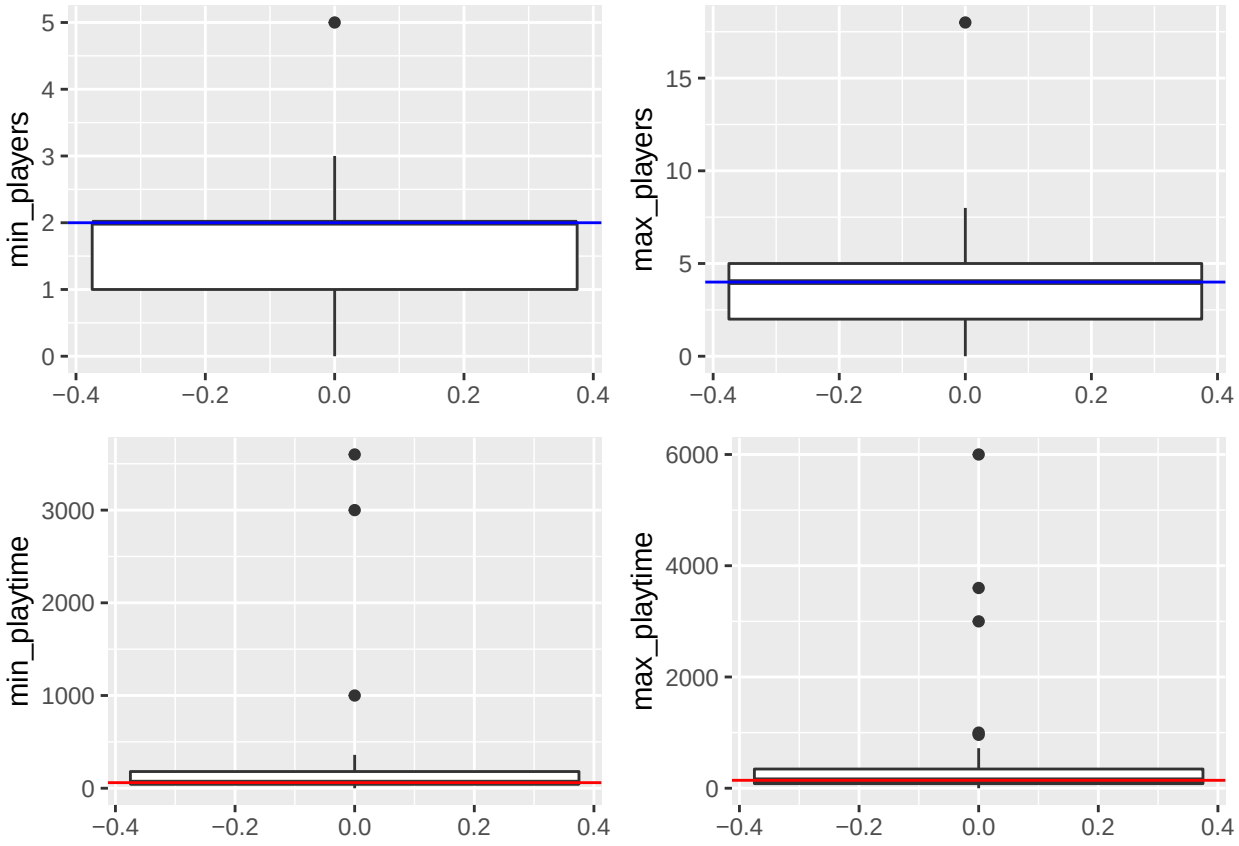


Figure 8: Summarizing all observations as Boxplots

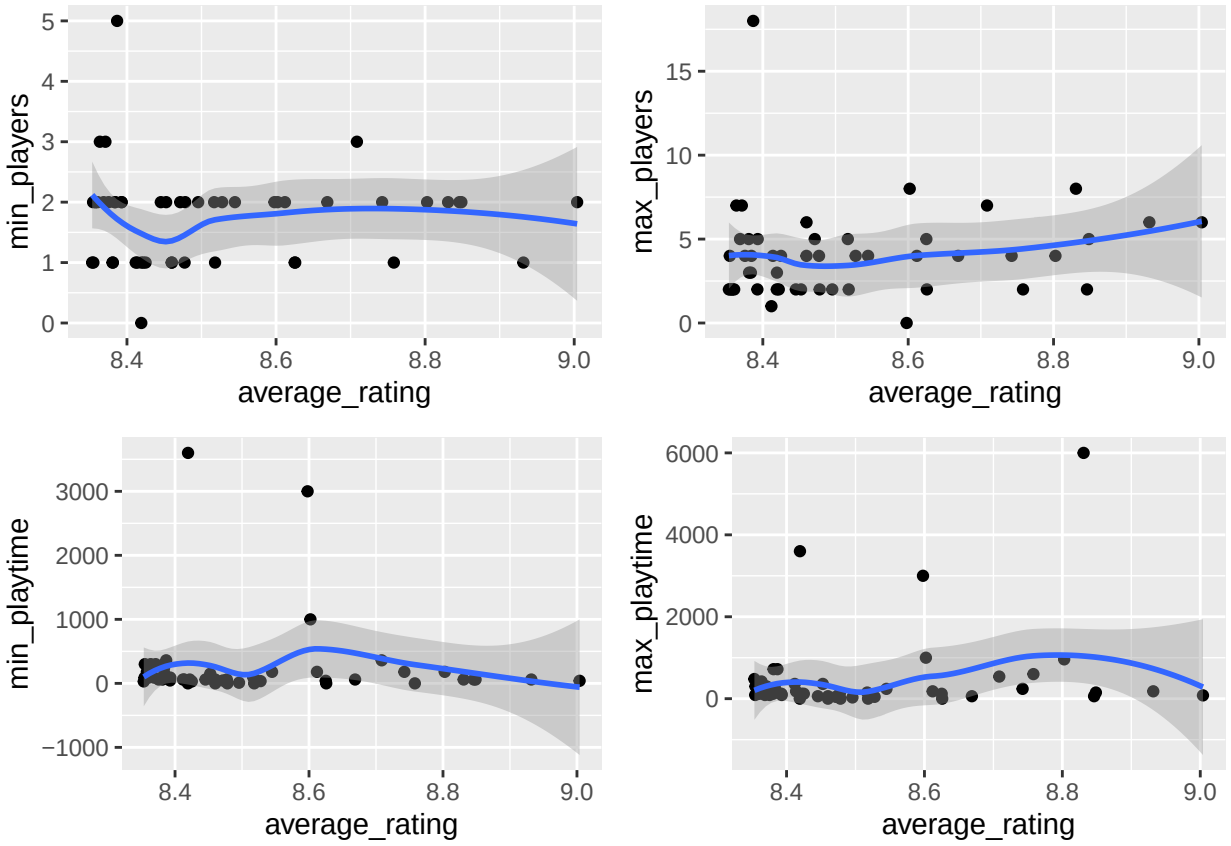
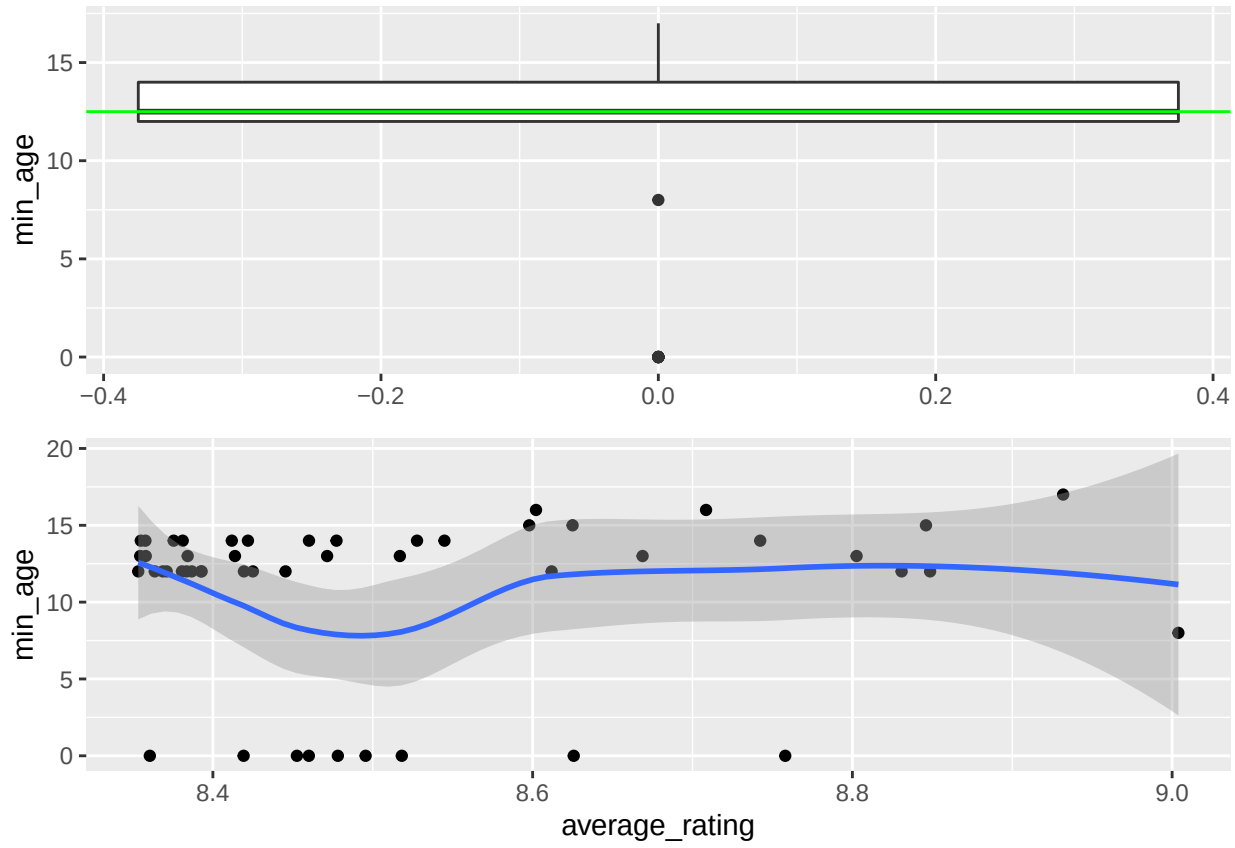


Figure 9: Relationship between Average Rating and other Attributes



- We can observe the following for the attribute Minimum Age -
  - Players of **age between 10-15 years** mostly play the top 50 games.
  - We can observe from the trend that games are more popular among **age group of 7-13 year olds**

3. Which game designer was most successful in producing popular games? Which publisher published the most popular games?

- The above scatter-plot @ref(fig:Designers) consists of average rating on x-axis and designer on y-axis. The black x-intercept represents the mean value of average ratings of the top 10 designers. The plot conveys that the mean average rating is around 8.82 with 5 observations on either side of the line.
- Philippe Keyaerts has the highest rated game at around 9+ followed by Vlaada Chvatil around 8.93 with all the other designers falling around the mean value. The lesser rated designer in the top 10 is Rob Daviau, Matt Leacock. We should note that Dean Essig has two games in the top 10.
- Who among these is the best is still a debatable question. Some might say it is Dean, while some might consider Philippe. Nevertheless, all of the designers in the plot are among the top 10 and have produced the most popular games.
- The above scatter-plot @ref(fig:Publishers) consists of average rating on x-axis and publisher on y-axis.
- The first thing that strikes from looking at this plot is that Multi-Man Publishing has 3 among the top 7 rated board games which hints that they are one of the best publishers.
- The top rated game was published by Days of Wonder and the lesser rated game in the top 7 was published by Compass Games. Again, it is debatable as to who is best but the publishers in the above plot have published some of the finest board games.

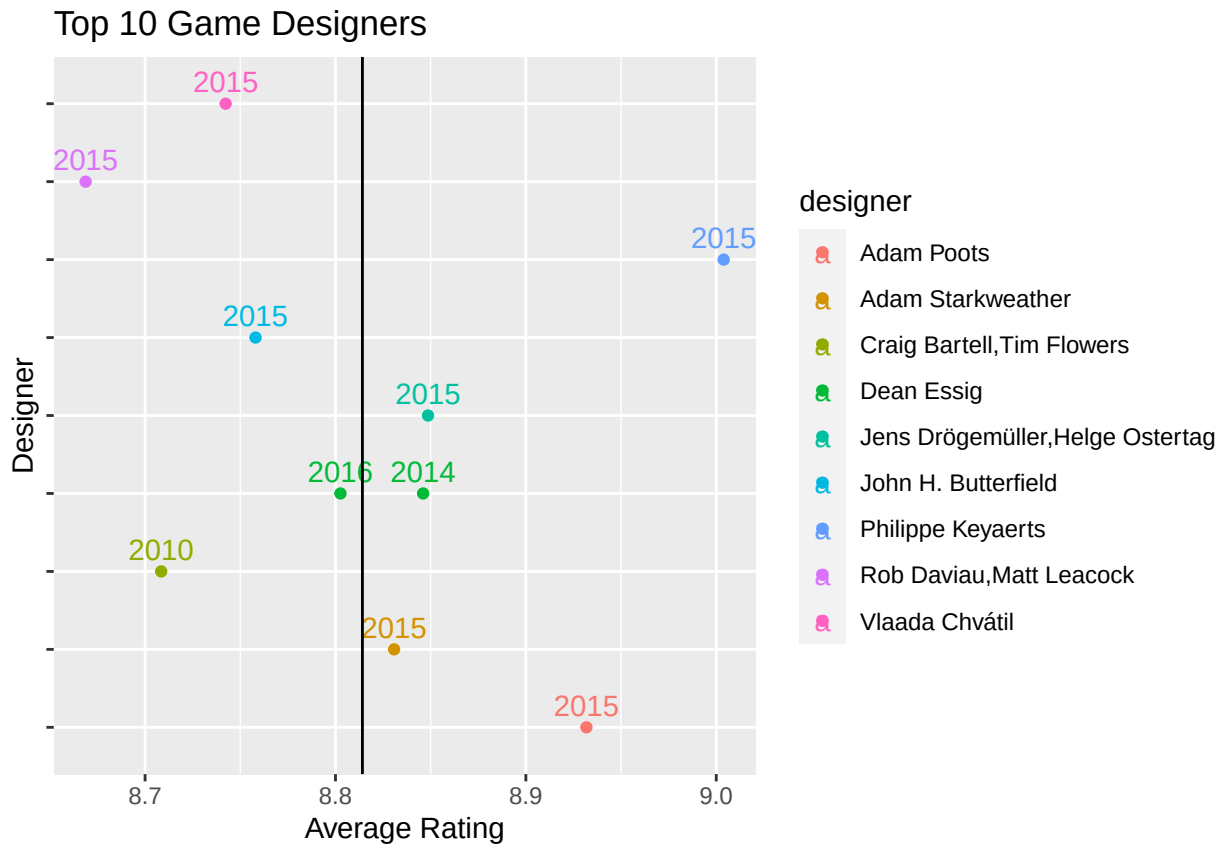


Figure 10: Top 10 Game Designers

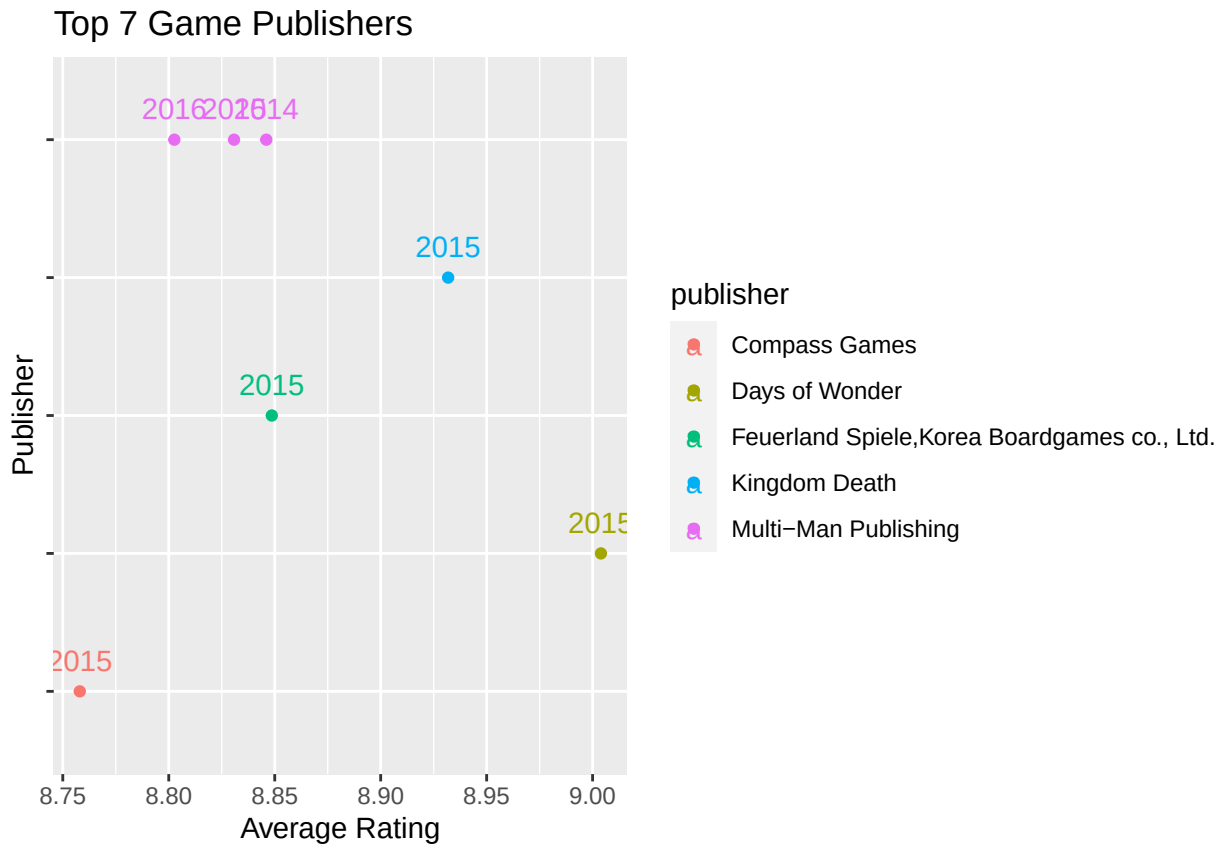


Figure 11: Top 7 Game Publishers

Bonus Insight - An interesting takeaway from the above two plots is that the best and **top rated board games were launched between 2010-2016** with most of the top rated games launched in the year **2015.**

References

Websites

- BoardGameGeek | Gaming Unplugged Since 2000. (2000). BGG. https://boardgamegeek.com/BoardGameGeek
- Huebner, M., Vach, W. and le Cessie, S., 2016. A systematic approach to initial data analysis is good research practice. The Journal of Thoracic and Cardiovascular Surgery, 151(1), pp.25-27.
- Thomas Mock, (2019). Tidy Tuesday. https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-03-12

R packages

- Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
- C. Sievert. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida, 2020.
- David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.0. https://CRAN.R-project.org/package=broom
- Hao Zhu (2019). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.1.0. https://CRAN.R-project.org/package=kableExtra
- Tierney N (2017). "visdat: Visualising Whole Data Frames." *JOSS*, *2*(16), 355. doi: 10.21105/joss.00355 (URL: https://doi.org/10.21105/joss.00355), <URL: http://dx.doi.org/10.21105/joss.00355>.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Yihui Xie (2020). bookdown: Authoring Books and Technical Documents with R Markdown. R package version 0.20.
- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.