Project Proposal: Board Game Data

Lauren Grove, Crystal Zhang, Nate Ren

Data Overview

This is data from BoardGameGeek used in TidyTuesday in 2019. This data contains every board game released from 1950 to 2016 that have at least 50 ratings in order to follow along the FiveThirtyEight article about board games which mentions the golden age of board games started after 1950. This data frame has 10,532 rows with 22 variables.

Data Exploration

Data Structure

```r
board_games_data <- read.csv("./board_games.csv")
str(board_games_data)
```

```
'data.frame':   10532 obs. of  22 variables:
 $ game_id       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ description   : chr  "Die Macher is a game about seven sequential political races in diffe
 $ image         : chr  "//cf.geekdo-images.com/images/pic159509.jpg" "//cf.geekdo-images.com
 $ max_players   : int  5 4 4 4 6 6 2 5 4 6 ...
 $ max_playtime  : int  240 30 60 60 90 240 20 120 90 60 ...
 $ min_age       : int  14 12 10 12 12 12 8 12 13 10 ...
 $ min_players   : int  3 3 2 2 3 2 2 2 2 2 ...
 $ min_playtime  : int  240 30 30 60 90 240 20 120 90 60 ...
 $ name          : chr  "Die Macher" "Dragonmaster" "Samurai" "Tal der Könige" ...
 $ playing_time  : int  240 30 60 60 90 240 20 120 90 60 ...
 $ thumbnail     : chr  "//cf.geekdo-images.com/images/pic159509_t.jpg" "//cf.geekdo-images.c
 $ year_published: int  1986 1981 1998 1992 1964 1989 1978 1993 1998 1998 ...
 $ artist        : chr  "Marcus Gschwendtner" "Bob Pepper" "Franz Vohwinkel" NA ...
 $ category      : chr  "Economic,Negotiation,Political" "Card Game,Fantasy" "Abstract Strat
 $ compilation   : chr  NA NA NA NA ...
 $ designer      : chr  "Karl-Heinz Schmiel" "G. W. \"Jerry\" D'Arcey" "Reiner Knizia" "Chri
 $ expansion     : chr  NA NA NA NA ...
 $ family        : chr  "Country: Germany,Valley Games Classic Line" "Animals: Dragons" "Asia
 $ mechanic      : chr  "Area Control / Area Influence,Auction/Bidding,Dice Rolling,Hand Mana
 $ publisher     : chr  "Hans im Glück Verlags-GmbH,Moskito Spiele,Valley Games, Inc." "E.S.
 $ average_rating: num  7.67 6.61 7.44 6.61 7.36 ...
 $ users_rated   : int  4498 478 12019 314 15195 73 2751 186 1263 6729 ...
```

There are a total of 10523 observations, each representing a different board game. The dataset consists of 22 variables, which includes the types integer, string, and numerical. The integer variables are: game_id, max_players, max_playtime, min_age, min_players, min_playtime, playing_time, year_published, and users_rated. The character variables are: description, image, name, thumbnail, artist, category, compilation,’designer, expansion, family, mechanic, and publisher. The only numerical variable is average_rating. More information on variables can be seen in the Variable List

Data Summary

```
summary(board_games_data)
```

```
   game_id       description          image            max_players
Min.   :     1   Length:10532       Length:10532       Min.   :  0.000
1st Qu.:  5444   Class :character   Class :character   1st Qu.:  4.000
Median : 28822   Mode  :character   Mode  :character   Median :  4.000
Mean   : 62059                                         Mean   :  5.657
3rd Qu.:126410                                         3rd Qu.:  6.000
Max.   :216725                                         Max.   :999.000
 max_playtime         min_age        min_players     min_playtime
Min.   :    0.00   Min.   : 0.000   Min.   :0.000   Min.   :    0.00
1st Qu.:   30.00   1st Qu.: 8.000   1st Qu.:2.000   1st Qu.:   25.00
Median :   45.00   Median :10.000   Median :2.000   Median :   45.00
Mean   :   91.34   Mean   : 9.715   Mean   :2.071   Mean   :   80.88
3rd Qu.:   90.00   3rd Qu.:12.000   3rd Qu.:2.000   3rd Qu.:   90.00
Max.   :60000.00   Max.   :42.000   Max.   :9.000   Max.   :60000.00
    name            playing_time       thumbnail         year_published
Length:10532       Min.   :    0.00   Length:10532       Min.   :1950
Class :character   1st Qu.:   30.00   Class :character   1st Qu.:1998
Mode  :character   Median :   45.00   Mode  :character   Median :2007
                   Mean   :   91.34                      Mean   :2003
                   3rd Qu.:   90.00                      3rd Qu.:2012
                   Max.   :60000.00                      Max.   :2016
   artist            category         compilation          designer
Length:10532       Length:10532       Length:10532       Length:10532
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character



 expansion            family            mechanic          publisher
Length:10532       Length:10532       Length:10532       Length:10532
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character



average_rating   users_rated
Min.   :1.384   Min.   :   50.0
1st Qu.:5.830   1st Qu.:   85.0
Median :6.393   Median :  176.0
Mean   :6.371   Mean   :  870.1
3rd Qu.:6.943   3rd Qu.:  518.0
Max.   :9.004   Max.   :67655.0
```

This is a basic summary of the dataset. The distribution of integer and numberical variables, such as max_playtime and min_players are demonstrated with the 5-number summary and the mean. For character variables, only length (which is the same as the number of observations), class, and mode (both of these show the data type) are present.

Variable List

- game_id = unique board game ID Number
  - Data Type = Integer
- description = short description of the game
  - Data Type = String
- max_players = the largest amount of players that can play the board game at a certain time

  - Data Type - Integer

- max_playtime = the longest estimated amount of time that it will take for people to finish playing the game. This is written in minutes

  - Data Type = Integer

- min_age = the minimum age a person needs to be in order to play the board game

  - Data Type = Integer

- min_players = the minimum amount of people that can play the board game at a certain time

  - Data Type = Integer

- min_playtime = the shortest estimated amount of time that it will take for people to finish playing the game. This is written in minutes.

  - Data Type = Integer

- name = the name of the board game

  - Data Type = String

- playing_time = the estimated average amount of time that it will take for people to finish playing the game. This is written in minutes

  - Data Type = Integer

- year_published = the year that the game was published

  - Data Type = Integer

- category = this is a string that says what categories each board game belongs to. If a board game has multiple categories all of them will be written down and will be separated by commas

  - Data Type = String

- compilation = if the game is part of a multi-game compilation is lists what compilation it is apart of

  - Data Type = String

- designer = this is the name of the person or people that designed the game. If there are multiple people that designed a game their names are split by commas in between each name

  - Data Type = String

- expansion = if the game has an expansion pack this variable has the name of th expansion pack
  - Data Type = String
- mechanic = this explains how the game is played and what the game uses to work. If a game uses multiple elements to run the elements are split by commas
  - Data Type = String
- publisher = this is the company or people that published the game. If a game has multiple publishers then the publishers are split by commas
  - Data Type = String
- average_rating = this is the average rating of the game on Board Game Geek. The rating system runs on a scale from 1 to 10
  - Data Type = Double
- users_rated = This is the number of user that rated the game on Board Game Geek.
  - Data Type = Integer

Data Cleaning Required

New Variable using `pivot_wider` to get game duration range

- Create a new variable game duration by finding the range of playing time and then use pivot_wider to have separate columns for each duration range.

Split `Cateogry` data by the commas to remove secondary categories

- Separate category data by splitting with commas to get the primary category

Combining two columns

- Combine artist and designer into one column and make a distinction with /

Removing Unnesccary columns

- Filter out image and thumbnail variables

Clean Variable Names and Values

Variable names

- Change variable names to camel case or some other case

Clean NA Values

- Variables artist, category, compilation, designer, expansion, family, mechanic, and publisher have NA values.
  - Remove rows with NA values in category and publisher

Change Variable Values

- Recode compilation and expansion values to Yes or No or Yes expansion or No expansion
- Recode max_players of 999 for game_id 2922 to something more reasonable, the game is for 1 or more players

Factorizing a Column

- Use string replace to remove secondary mechanics or categories to get only the primary mechanic or category for each game
- Change separated category variable from character to a factor

Research Question

Research Question 1

Is there a relationship between average_rating and average_playtime?

Variables Needed

- average_rating
- average_playtime

Methodology

We need to create a new variable average_playtime using mutate with max_playtime and min_playtime and then we can create a scatter plot to visualize this data.

Research Question 2

Between, playtime, number of users rated, and max players, what has the most impact on average rating.

Variables Needed

- max_playtime
- min_playtime
- users_rated
- max_players
- min_players
- average_rating

Methodology

Utilizing ggplot and faceting function to create box plots to plot the relationship between all the variables and average_rating.