Board Ga

Team Maximus

Harrison Peloquin, Hunter Sawyer, Chris Dinkins, Jackson Hanchek

2022-11-21

Library Installations

Explanation of Data source

https://www.kaggle.com/datasets/andrewmvd/board-games?resource=download

Our dataset is maintained on kaggle.com, a platform for data science related tools and discussion. Our data set, titled "Board Games", was uploaded to the site by a senior data scientist from the Hospital Israelita Albert Einstein, the highest prestige hospital in Latin America. This data set, however, does not involve medical data as it was a leisure project. The data set, which contains over twenty thousand different board games, was taken from the website BoardGameGeek in February of 2021. BoardGameGeek is currently the largest board game ranking platform, and contains many different stats about over one hundred thousand board games. The twenty thousand board games in the data set are a subset of the larger hundred thousand. They are ranked in popularity, which requires user accounts to vote on them. Alongside popularity and name, the data includes publishing year, minimum and maximum player count, play length, suggested age, and rating from a scale of one through ten.

Initial Observations

When our group first attempted to handle our dataset, we realized the number of rows was so large that our computers had difficulty processing it. In fact, the dataset's size was gargantuan enough to give us issues when trying to create tables using pander. We have decided the best course of action to solve this predicament was to take a smaller portion of the data and use that to represent the dataset. By choosing the first 1,000 entries, we eliminate the bias that would come from handpicking certain data.

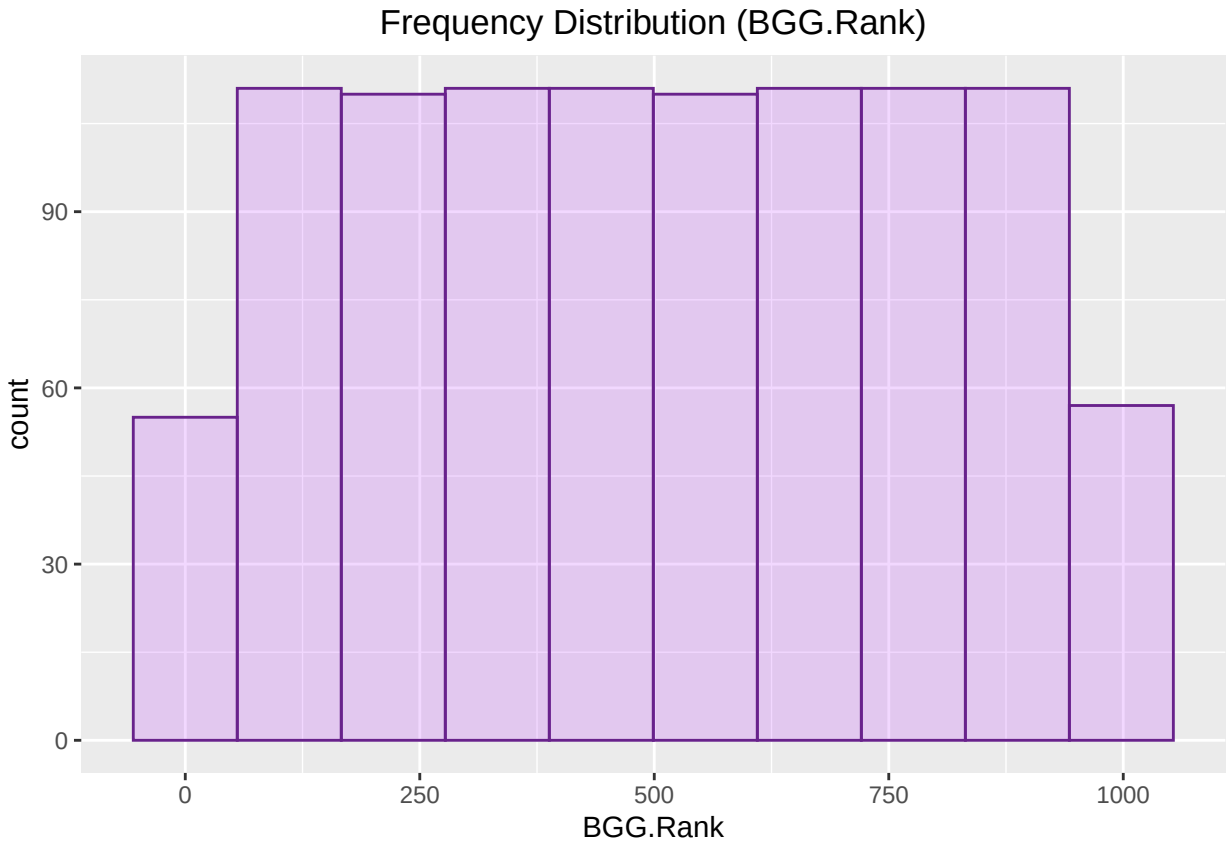
After observing our data closely, we have noticed that there are a few cases of missing data. One such case can be found in the 'BoardGamesGeek ID', where board games can be found that do not have any IDs entered at all. We have decided that these null values do not justify throwing the entire rows away as the 'BoardGamesGeek ID' does not directly affect any way that we are analyzing or wanting to use this data. Instead, we have decided to leave these values empty.

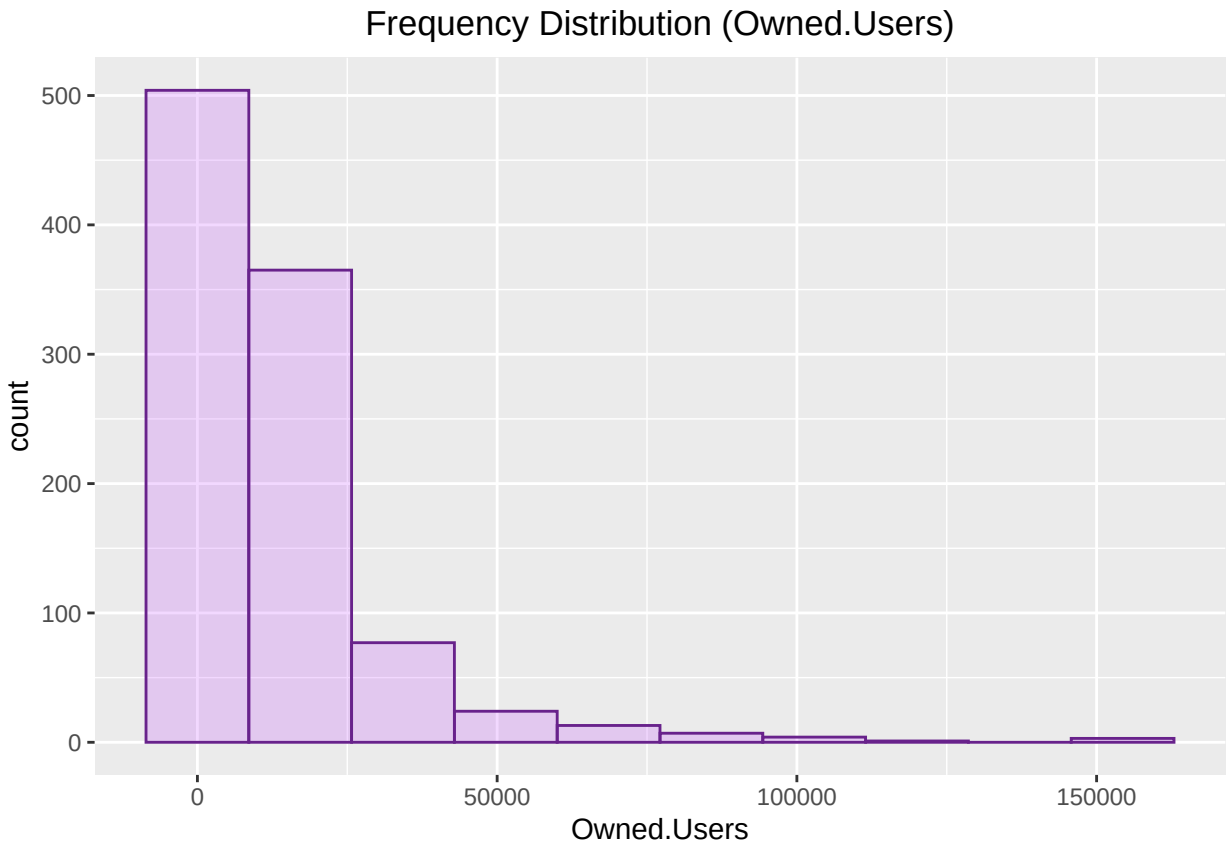
One of the most prominent predicaments we have come across in the early stages of observing our data is how to handle noisy data. The first instances we noticed were in the 'Min Players' and the 'Max Players' columns. Here, we saw board games hold the value of 0, which seems meaningless and does not make much sense at first. However, considering the context of what the minimum players and the maximum players means, we theorize that a 0 value means the board game did not specify a minimum or a maximum number of players. With that said, we have decided insert what we have decided. Similarly, we can see the 'Min Age' and the 'Play Time' column also have values of 0, so we plan to insert what we plan to.

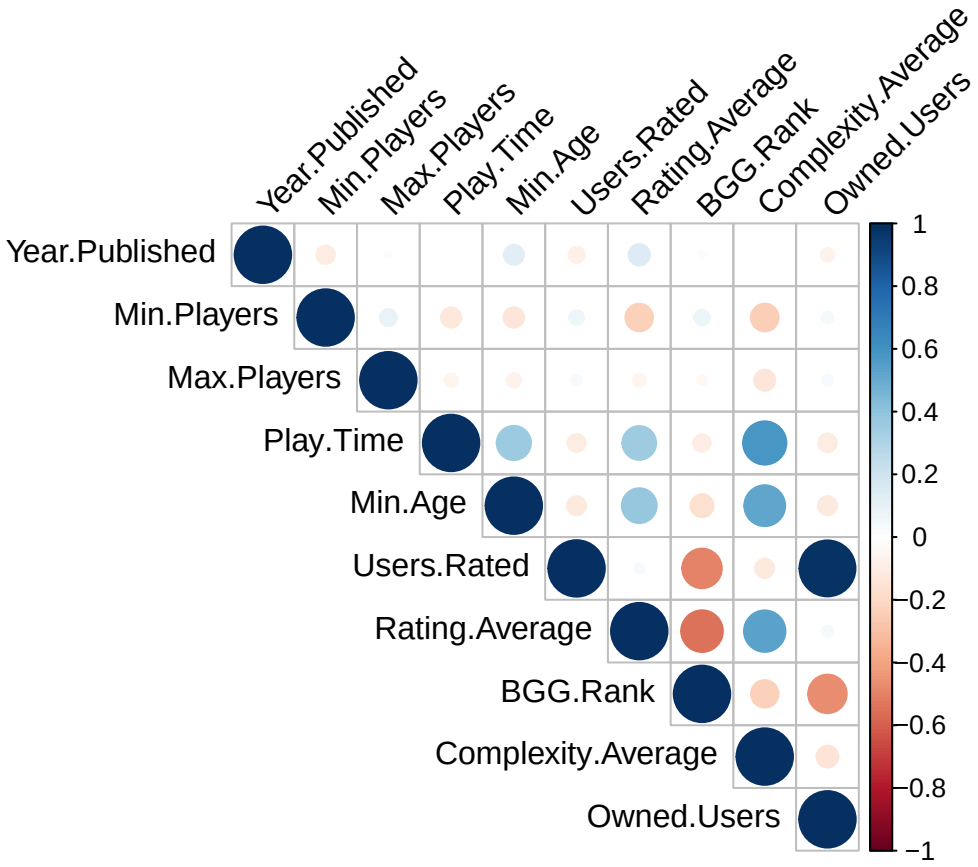
We have also seen that the 'Year Published' column has negative values in it, which is completely illogical given the context. At first, we considered that the negative signs were a mistake, and we could take the absolute value. However, values reach -3500 so an absolute value would result in a year that has not yet occurred at the time of observing the data. Perhaps negative values represent years from BC as opposed to AD. This could be the case, but it is difficult to imagine humans in 1300 BC passing time by playing a friendly match of Tic-Tac-Toe, so it would probably be in our best interest to consult a domain expert before making assumptions.

We also noticed that the average rating and average complexity were in decimal format with ,'s instead of decimal points. This is common practice in Brazil, where this data was collected. we changed the comma's to decimals to make performing calculations easier

Here are the histograms for the bgg ranks and owned users

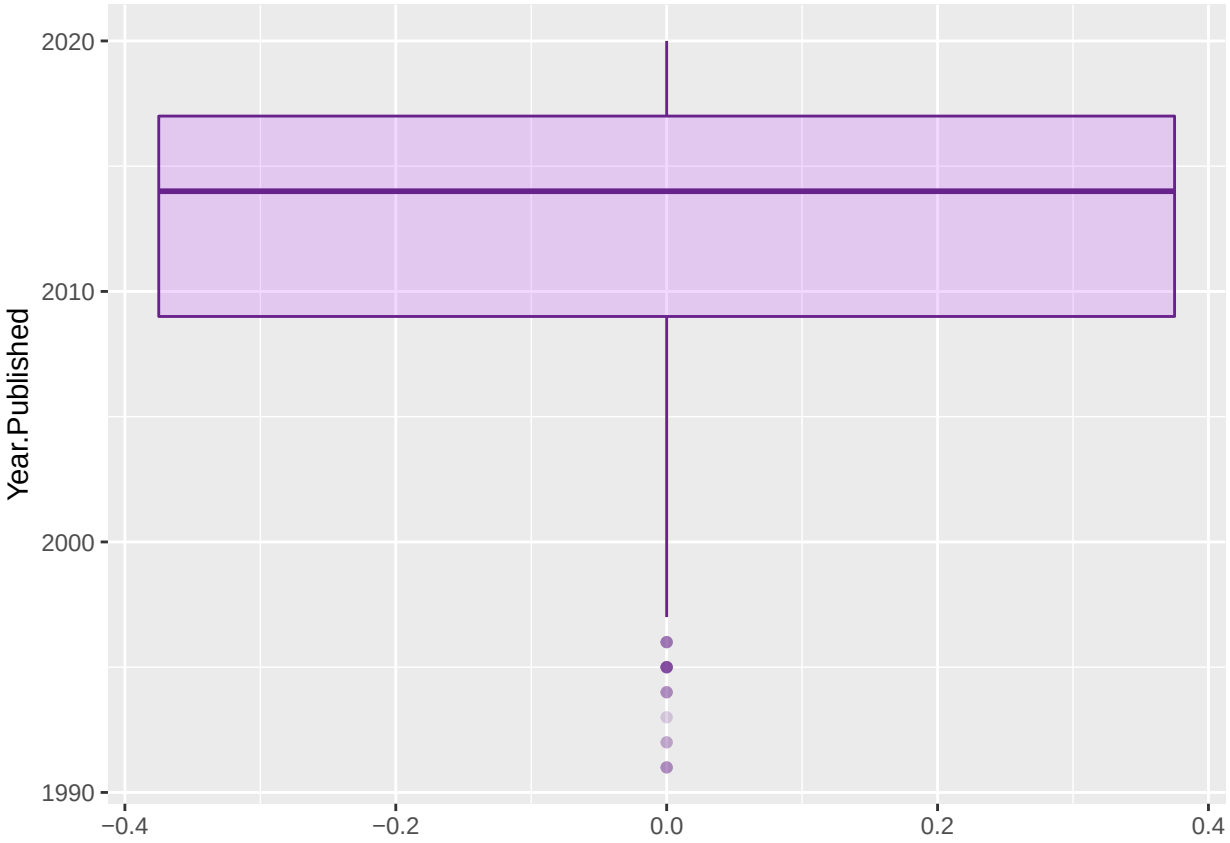


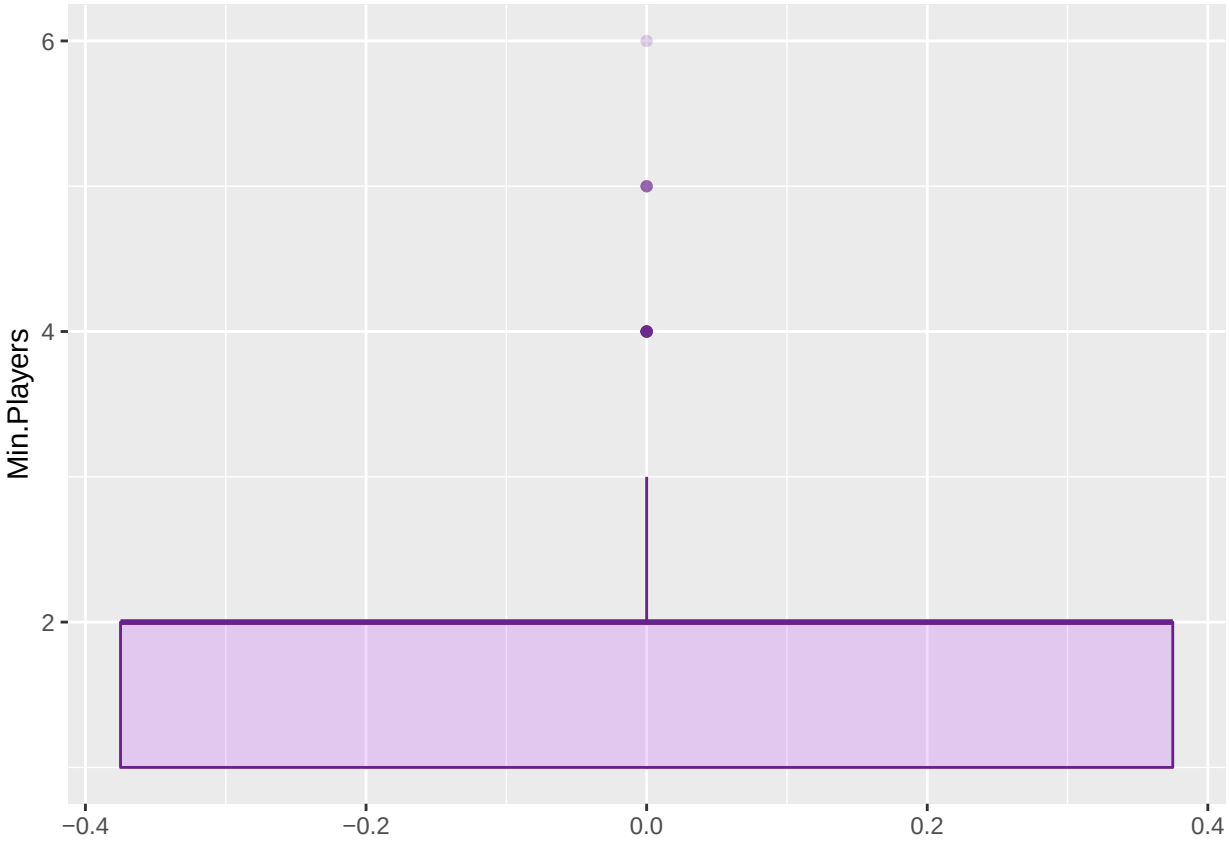

Frequency Distribution (Owned.Users)
count
500
400
300
200
100
0
0
50000
100000
150000
Owned.Users

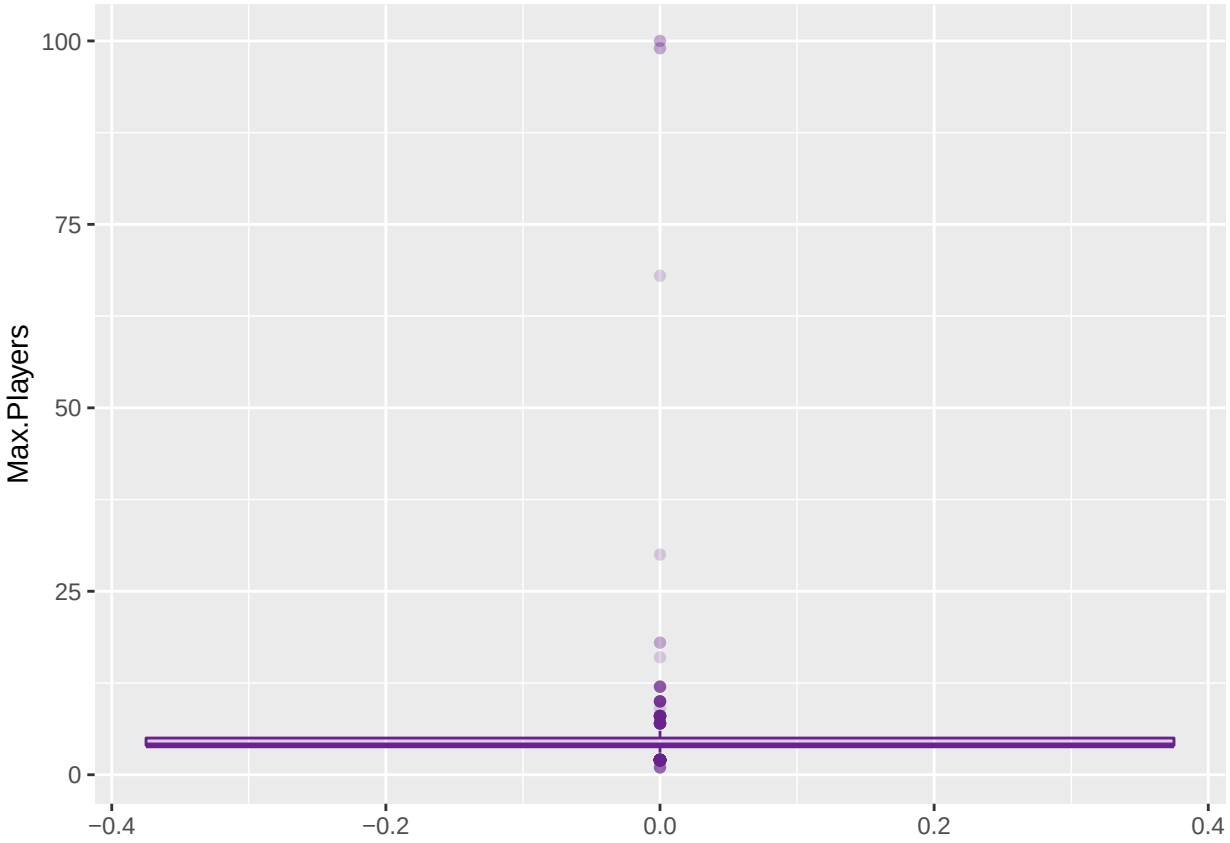


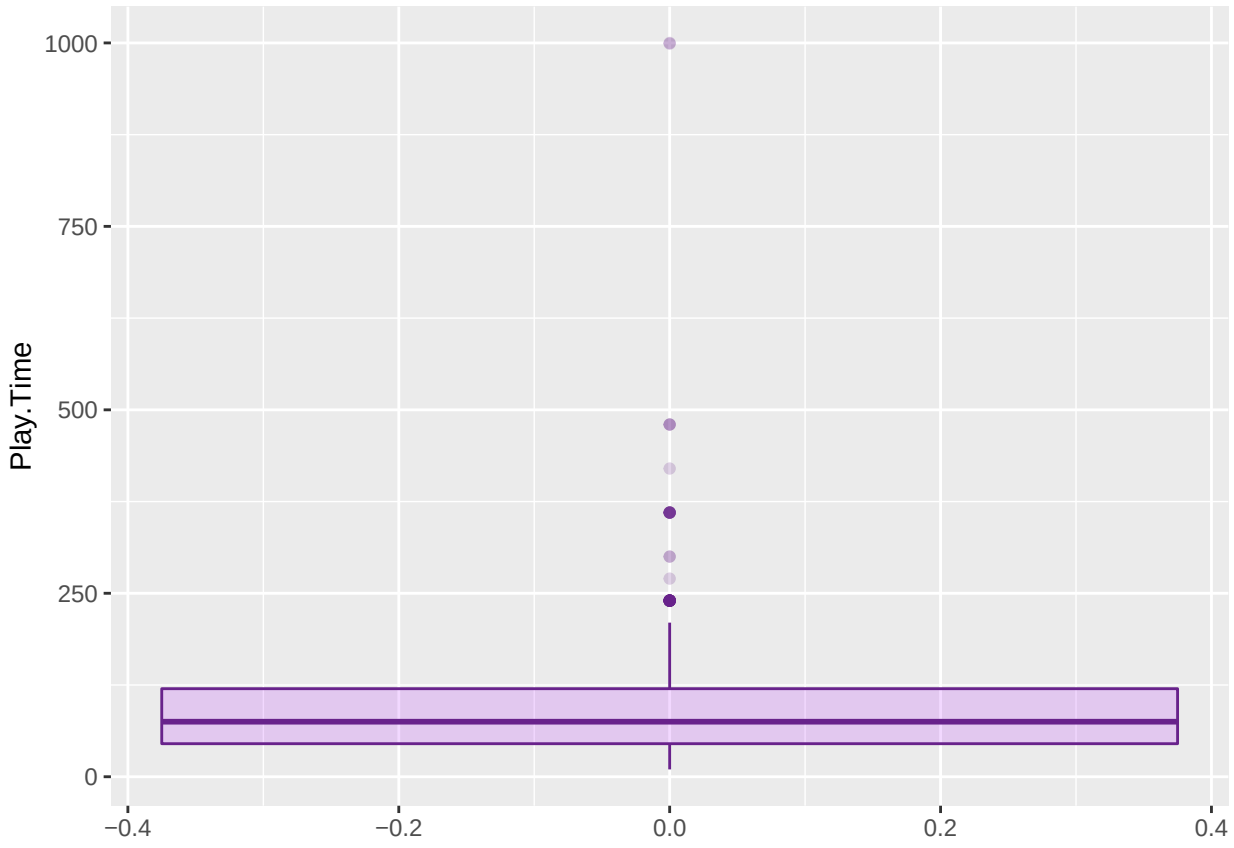
box plots

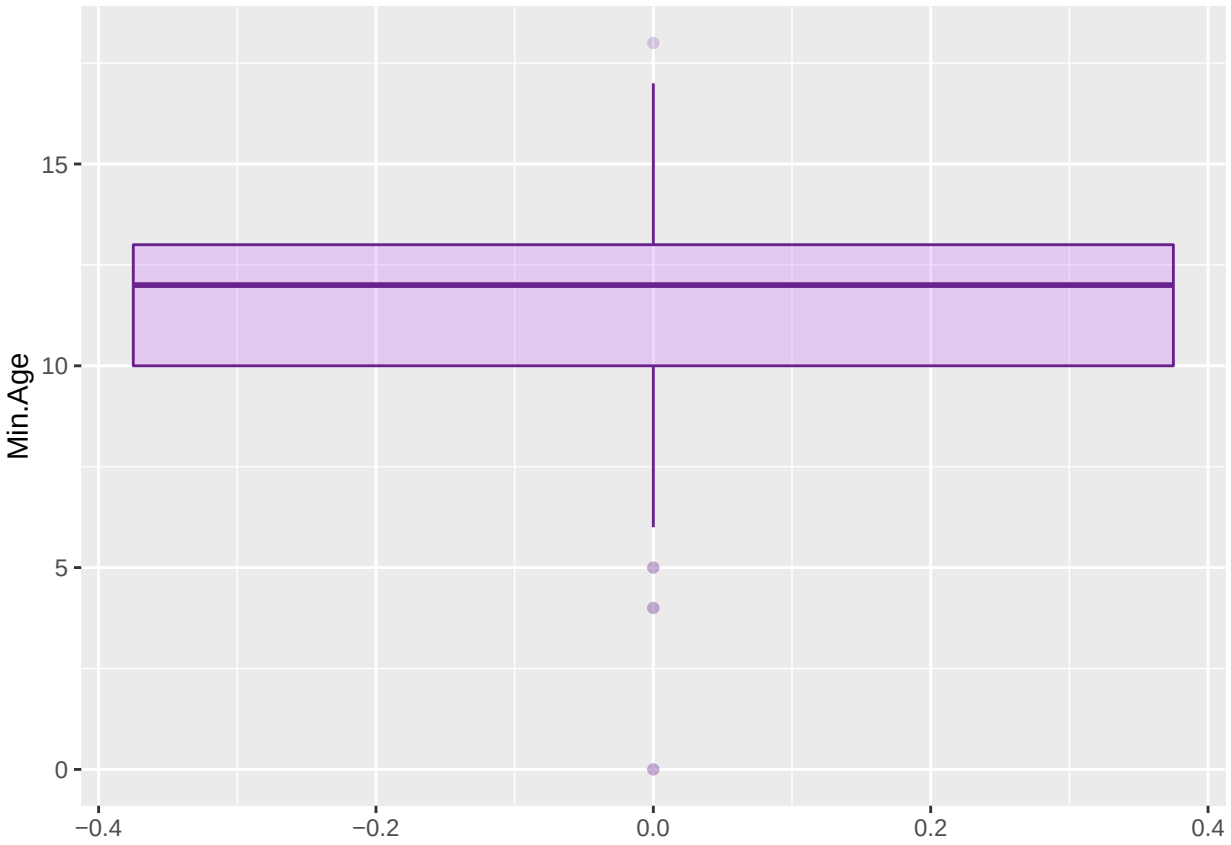
```
for(var in numeric_variables){
  group <- "color"
  print(
    ggplot(above_2000_df, aes_string(y=var))
    + geom_boxplot(colour="darkorchid4", fill="darkorchid1", alpha=0.2)
  )
}
```

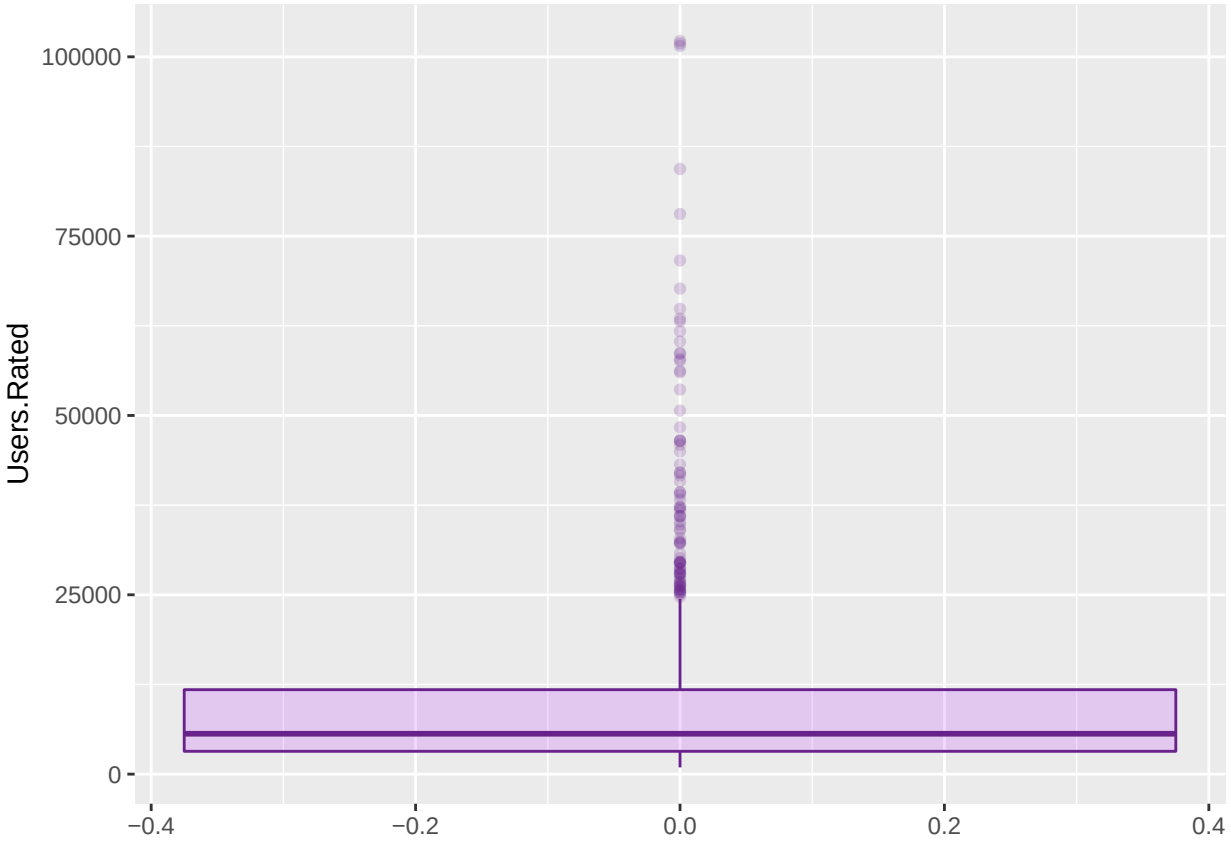

Year.Published
2020
2010
2000
1990
−0.4
−0.2
0.0
0.2
0.4

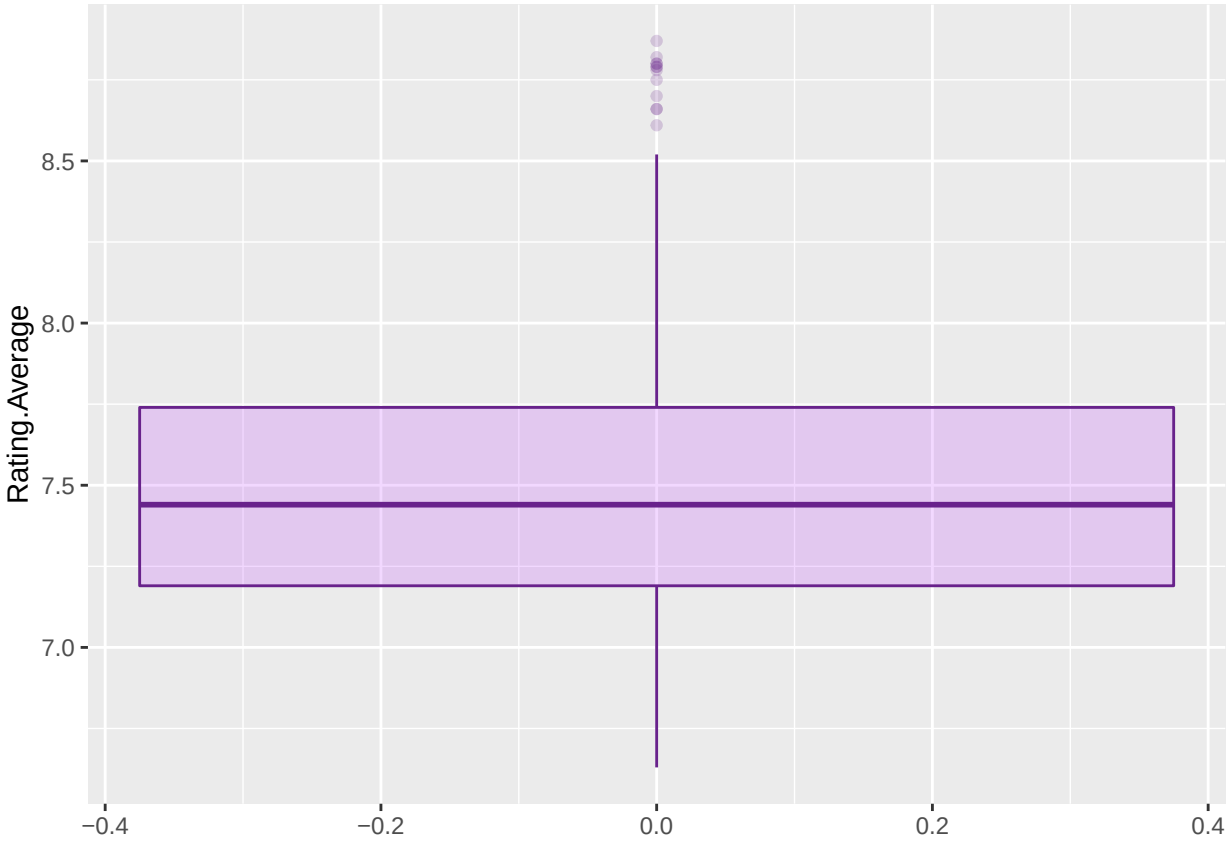

Min.Players
6
4
2
−0.4
−0.2
0.0
0.2
0.4

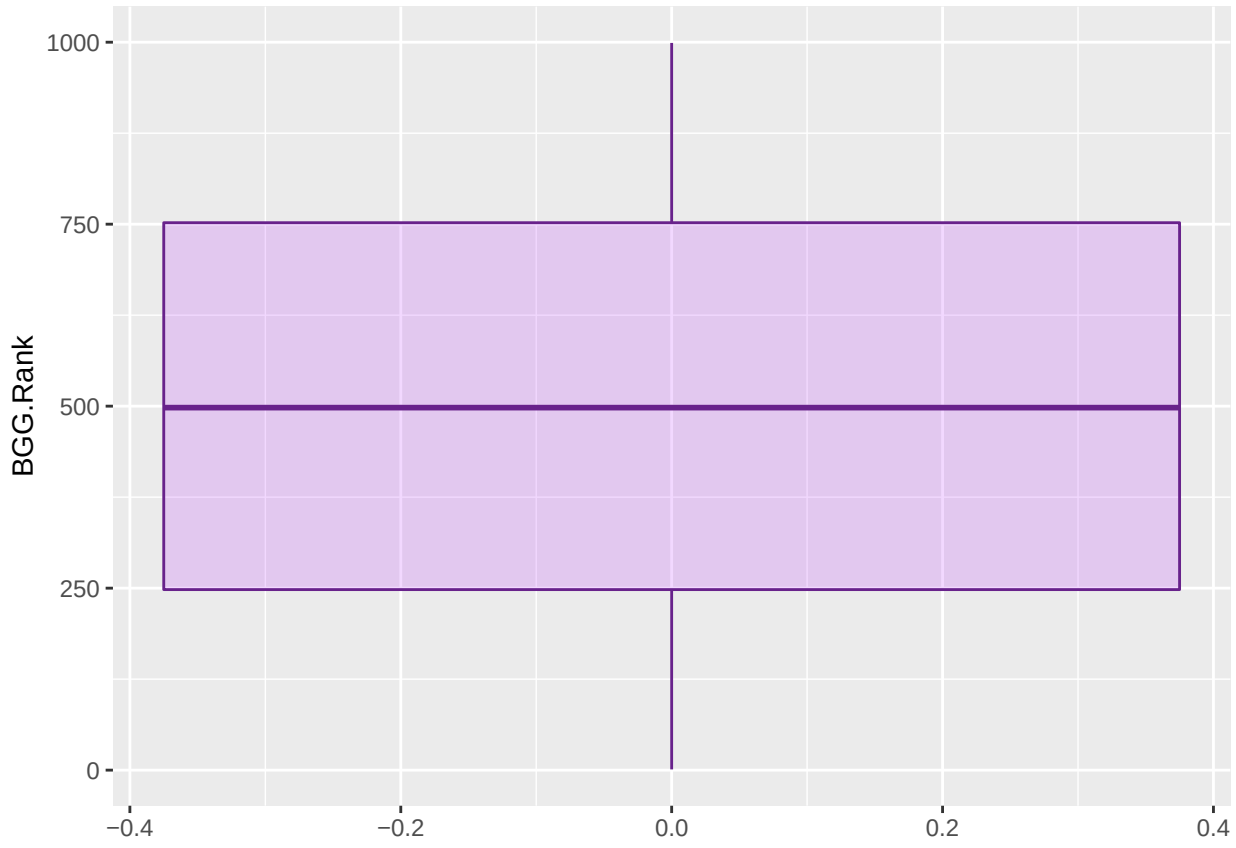

Max.Players
100
75
50
25
0
−0.4
−0.2
0.0
0.2
0.4

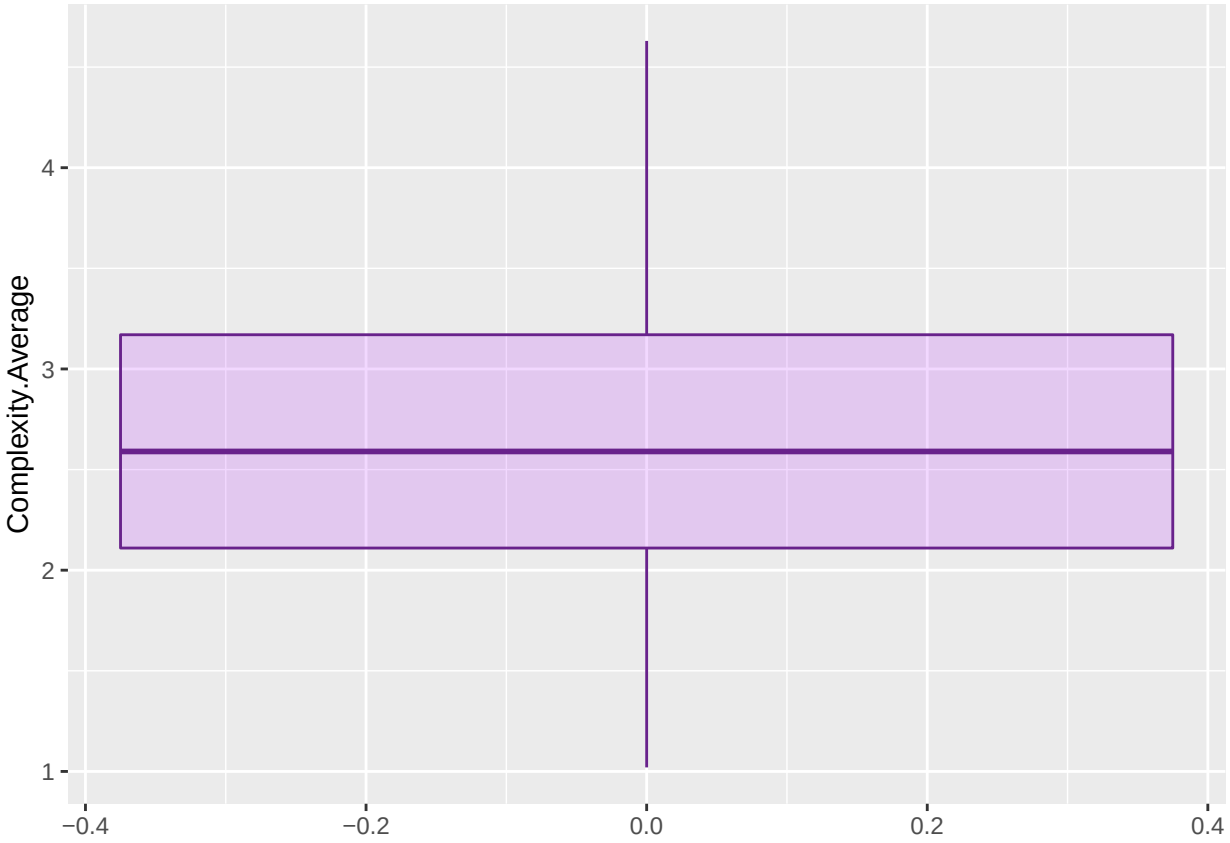

Play.Time
1000
750
500
250
0
−0.4
−0.2
0.0
0.2
0.4

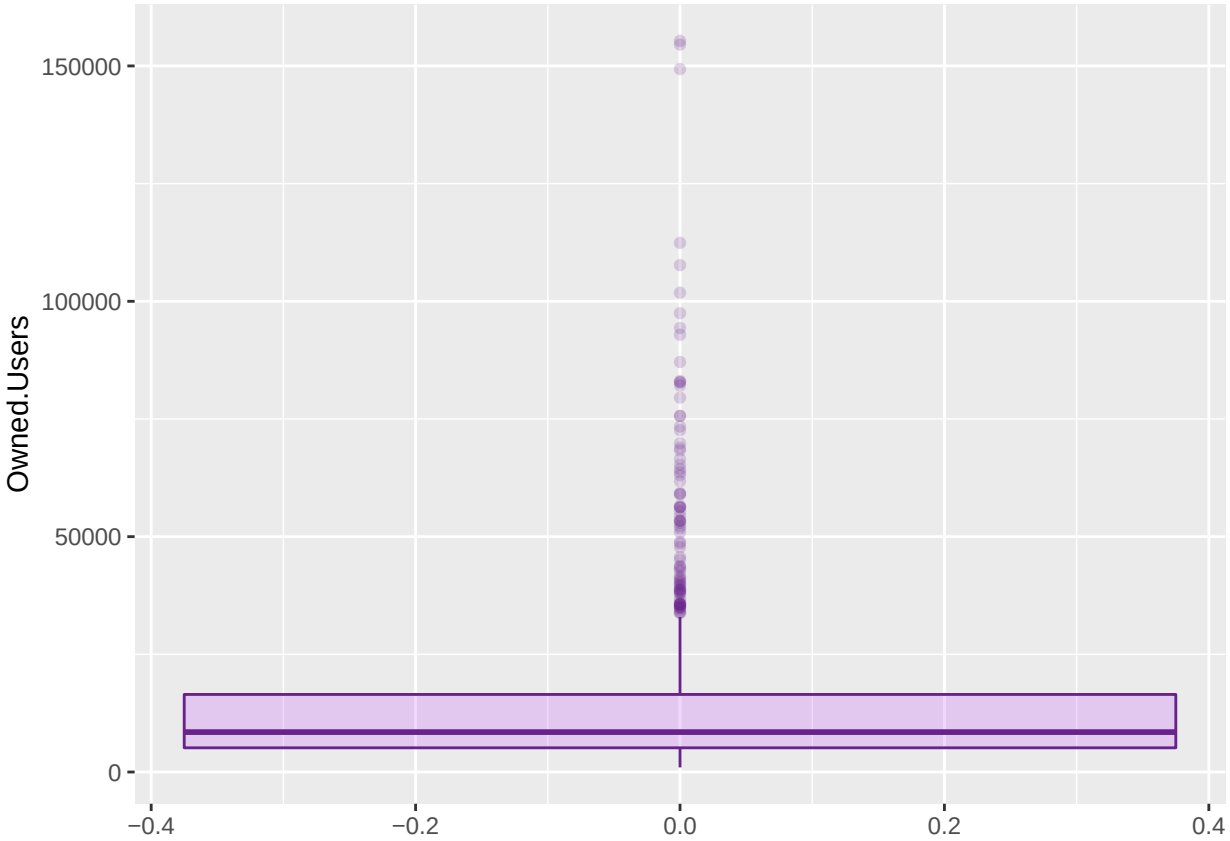

Min.Age
15
10
5
0
−0.4
−0.2
0.0
0.2
0.4


Users.Rated
100000
75000
50000
25000
0
−0.4
−0.2
0.0
0.2
0.4


Rating.Average
8.5
8.0
7.5
7.0
−0.4
−0.2
0.0
0.2
0.4


BGG.Rank
1000
750
500
250
0
−0.4
−0.2
0.0
0.2
0.4


Complexity.Average
4
3
2
1
−0.4
−0.2
0.0
0.2
0.4



What we want to do with the data

Citations

Dilini Samarasinghe, July 5, 2021, “BoardGameGeek Dataset on Board Games”, IEEE Dataport, doi: https://dx.doi.org/10.21227/9g61-bs59.