Applied Statistics

Final Presentation Written Report

Andy Brandt

December 11, 2025

Board Game Complexity Analysis

Abstract:

The motivation behind this analysis is to understand if there is any connection between different aspects of a board game and its release date. Are board games on average rated above 6.5 out of 10? Is there a correlation between a game's release date and how long it takes to play said game? Are newer games more complex than older games? Is there a correlation between a game's complexity and how long it takes to play, or its average rating?

Introduction:

Data was downloaded from the website Kaggle, with the raw data being collected from the website Board Game Geek, the main database on board games. Though the website houses over 125,600 board games, the dataset used contains about 20,000. Each board game has a description containing when it was first published or recorded, its advertised play time, and the site allows users to review games they have played or own, along with rating the game's complexity out of five and the game overall out of ten.

Methodology:

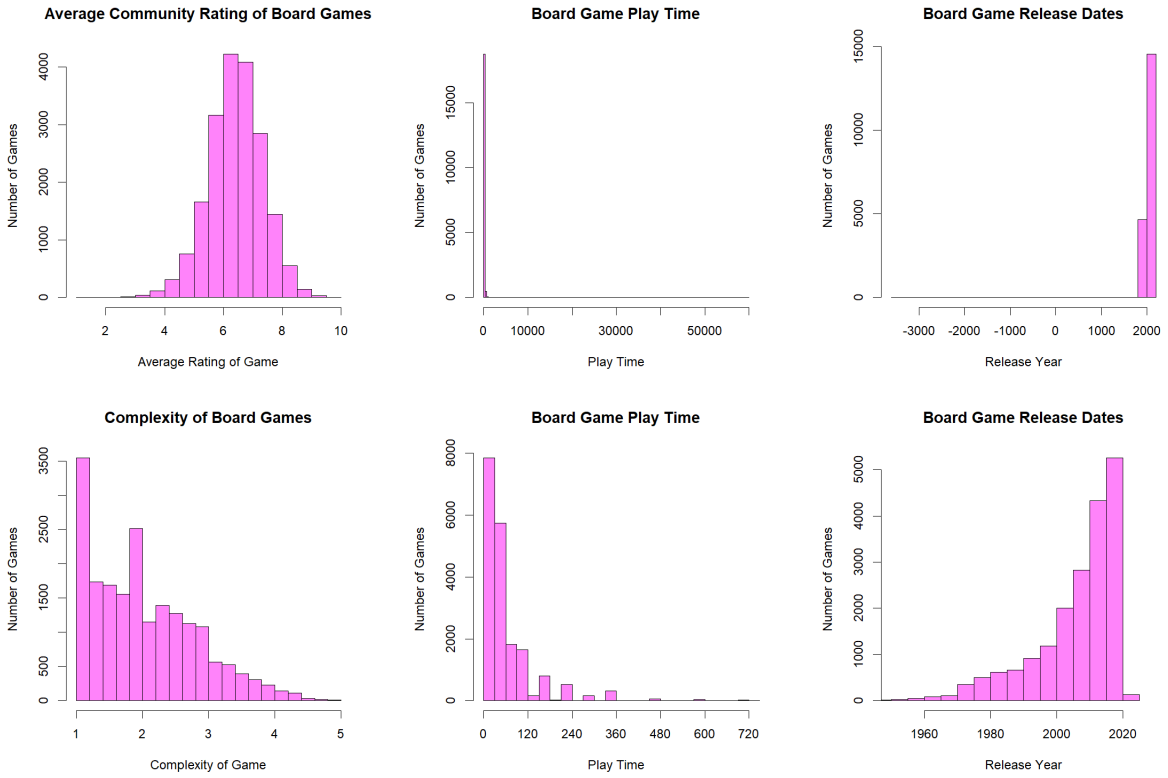
I) Graphical Representation and Summary Statistics

The original data set included 20,343 data points, but when reviewing the data, I discovered that many data points were incomplete, with missing data defaulting to zero. You can see this in the 'BGG_Data_Set.csv' file when sorting the data by complexity, as the data should be on a scale from one to five. This issue was resolved by creating a copy of the dataset and deleting any rows with incomplete information in the Year Published, Play Time, Rating Average, and Complexity columns, as those are the columns that are of concern for this analysis. The file I will be using for this analysis is called 'BGG_Data_Set_MODIFIED.csv' with 19,256 data points.

The five-number summary can be viewed here. Again, Min.Players, Max.Players, Min.Age, Users.Rated, BGG.Rank, and Owned.Users can be ignored as their data will not be used for this analysis. Year.Published and Play.Time have some

```
       ID              Name           Year.Published    Min.Players       Max.Players        Play.Time
 Min.   :     1   Length:19256       Min.   :-3500   Min.   : 0.000   Min.   :  0.000   Min.   :    1
 1st Qu.: 10187   Class :character   1st Qu.: 2001   1st Qu.: 2.000   1st Qu.:  4.000   1st Qu.:   30
 Median : 86003   Mode  :character   Median : 2011   Median : 2.000   Median :  4.000   Median :   45
 Mean   :106980                      Mean   : 2002   Mean   : 2.019   Mean   :  5.597   Mean   :   95
 3rd Qu.:191886                      3rd Qu.: 2016   3rd Qu.: 2.000   3rd Qu.:  6.000   3rd Qu.:   90
 Max.   :331787                      Max.   : 2022   Max.   :10.000   Max.   :999.000   Max.   :60000
 NA's   :13
    Min.Age        Users.Rated       Rating.Average     BGG.Rank     Complexity.Average  Owned.Users
 Min.   : 0.000   Min.   :    30.0   Min.   :1.05   Min.   :    1   Min.   :1.000      Min.   :     3
 1st Qu.: 8.000   1st Qu.:    60.0   1st Qu.:5.84   1st Qu.: 4827   1st Qu.:1.350      1st Qu.:   155
 Median :10.000   Median :   132.0   Median :6.44   Median : 9764   Median :2.000      Median :   336
 Mean   : 9.706   Mean   :   884.9   Mean   :6.42   Mean   : 9935   Mean   :2.037      Mean   :  1478
 3rd Qu.:12.000   3rd Qu.:   416.0   3rd Qu.:7.03   3rd Qu.:15004   3rd Qu.:2.560      3rd Qu.:   926
 Max.   :25.000   Max.   :102214.0   Max.   :9.54   Max.   :20344   Max.   :5.000      Max.   :155312
```

extraordinary outliers which skew their mean quite aggressively. Since Board Game Geek has a comprehensive list of every game recorded, it also includes games from ancient history like Senet which were preserved well enough to sell copies today. 95.8% of board games advertise their play time to be less than 240 minutes, but the 4.2% of games who are above 240 minutes boast such grandiose play times cause the mean to be more than double the median. The enormous play times come from a category of games called campaign games, where the game is designed to be played over multiple sessions.



The four histograms can be seen above. The Average Community Rating and Complexity histograms remain unchanged, but the Play Time and Release Date histograms were illegible due to the outliers. This was solved by zooming in on the majority of the results. The unmodified graphs can be seen above the modified version.

With this we can see that the average rating over all board games is about 6.5 with no skew. And the most common complexity for games is 1 with a hard right skew. The play time and release date have a right and left skew respectively, with most games ending in 0-30 minutes and more games coming out every year. The last bar on the Release Date histogram is due to the dataset's newest games coming out in 2022, with each bar lasting 5 years.

II) Statistical Analysis and Inferences

a) <u>One-Mean T-Test</u>

I would like to believe that most board games are good. After all, no one would publish a game if they thought it was poorly made or had bad mechanics. This can be shown if, on

average, a board game is ranked at least 6.5 out of 10. I will also test this at a 5% significance level. I chose 25 at random by using a random number generator and finding the game with that corresponding ID. Their average ratings can be seen on the right, while the name of the chosen games can be found on the slides attached to this assignment.

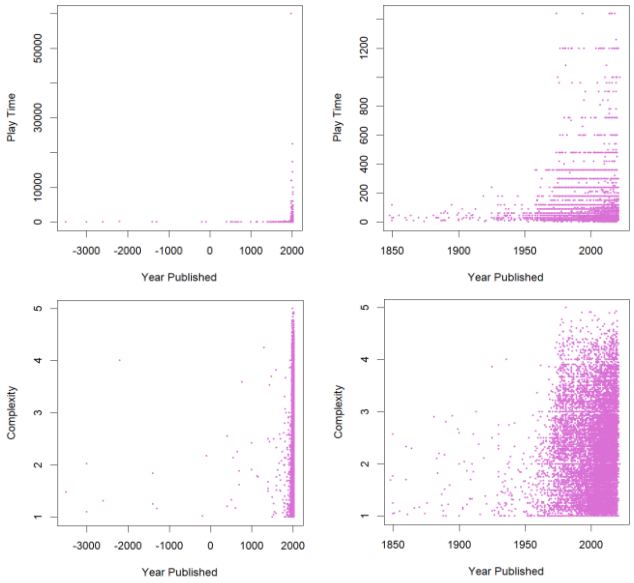
5.24	7.55	5.39	7.49	8.05
8.12	7.38	7.24	5.65	6.67
5.28	7.05	6.51	5.07	5.84
6.00	5.20	6.13	5.80	5.56
5.59	8.00	6.65	6.17	7.52

With this we can find the sample mean to equal 6.446 and the sample standard deviation to be 0.9979. Our $H_0 = 6.5$, the average and our $H_a > 6.5$. With this we can find

$$t_o = \frac{\bar{x} - \mu 0}{\frac{s}{\sqrt{n}}} = \frac{6.5 - 6.446}{\frac{0.9979}{\sqrt{25}}} \approx 0.2706$$

and our P-score to equal 5.750052e-07. Since the p-value < 0.05, we reject H_o at a 5% significance level. The data provides sufficient evidence to conclude that the average game's community ranking is greater than 6.5/10.

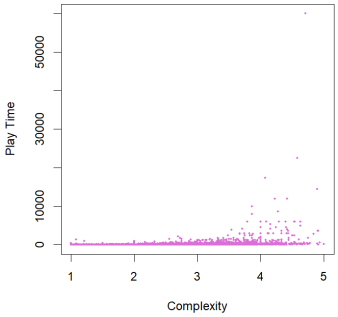
b) Correlation

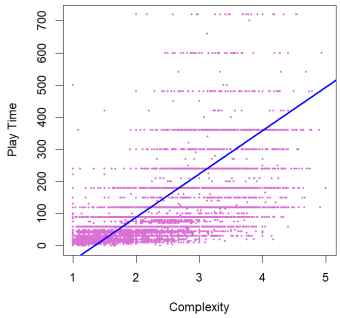


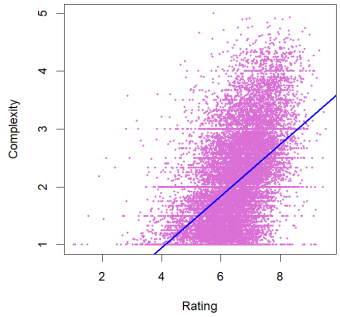
Since both correlation tests are using either Play Time or Year Published (or both) I have supplied additional graphs that crop out the outliers. However, both of these tests have came back negative. The correlation between Play Time and Year Published has a correlation coefficient of -0.0008690, and the correlation between Complexity and Year Published resulted in a correlation coefficient of 0.006162. This is due to the fact that board games are not a science. With such a broad subject, games are never going to line up nicely. That being said, I would have assumed that there would be at least a little correlation between how complex a game is and when it was published. It seems that games are only getting more and more complex, but this view might come from me looking for more and more complex board games.

c) Linear Regression

Once again, I have supplied a legible variant of the Play Time x Complexity graph. The lines of points shown on that graph come from the fact that games prefer to advertise a clean number in increments of 15 minutes. This is also why the five-number summary for Play Time is in multiples of 15.







In the Play Time x Complexity graph, the correlation coefficient is 0.1929. Using R, we can calculate the y-intercept to be -178.8 and the scaler to be 134.4. Putting this together we can construct the linear line $Y = -178.8 + 139.4x$. When the complexity increases by one, the play time will also increase by about 140 minutes. This is obviously not true, but it does show how a few one bad outlier can change the results of multiple tests.

In the Complexity x Rating graph, the correlation coefficient is 0.5116. Using R, we can calculate the y-intercept to be -0.8354 and the scaler to be 0.4474. Putting this together we can construct the linear line $Y = -0.8354 + 0.4474x$. When the rating of a game increases by one, you can expect the complexity to also increase by about 0.45. This seems much more realistic than the previous test. After all, people will rate a game higher if it has complex mechanics that take time to master rather than a game that becomes boring after the first few playthroughs.

Results and Conclusions:

After running the analysis using R-studio software the following was concluded. The average rating of all board games in the sample is above 6.5. This makes sense because publishers would not allow bad games to be published, as that would now result in copies being sold. We also discovered that when a game was published has no factor on a game's complexity or how long it takes to play. This can be extrapolated to say that, given a game complexity, you could not guess the age of the game. The same can be said for a game's play time.

There is correlation between a game's complexity and a game's play time. This correlation can be seen better if we remove the outliers in the play time, as the existence of campaign games makes the bet fit line look faulty. There is also a correlation between a game's complexity and its rating. This has the strongest correlation and is the most logical. People will naturally a complex game like Catan much higher than a game like Go-fish or War.

References:

Monfared, Melissa. (2024) A Data Driven Review of Board Game Design and Interactions of Their Mechanics.

Appendices:

```
bgg <- BGG_Data_Set_MODIFIED


summary(bgg)


sd(bgg$Rating.Average)
sd(bgg$Complexity.Average)
sd(bgg$Play.Time)
sd(bgg$Year.Published)
  # Calculated in Excel, resulted in SD = 199.6


tsample <- c(5.24, 8.12, 5.28, 6.00, 5.59,
             7.55, 7.38, 7.05, 5.20, 8.00,
             5.39, 7.24, 6.51, 6.13, 6.65,
             7.49, 5.65, 5.07, 5.80, 6.17,
             8.05, 6.67, 5.84, 5.56, 7.52)
mean(tsample)
sd(tsample)
pt(mean(tsample), 24, lower.tail = FALSE)


hist(bgg$Rating.Average,
     main="Average Community Rating of Board Games",
     xlab="Average Rating of Game",
     ylab="Number of Games",
     col="orchid1")
```

```
boxplot(bgg$Rating.Average,
        main="Community Ratings Box Plot",
        ylab="Board Game Community Rating",
        col = "orchid1")

hist(bgg$Complexity.Average,
     main="Complexity of Board Games",
     xlab="Complexity of Game",
     ylab="Number of Games",
     col="orchid1")

boxplot(bgg$Complexity.Average,
        main="Complexity Box Plot",
        ylab="Board Game Complexity",
        col = "orchid1")

# Looking only at rows with Play.Time <= 12 hours
hist(bgg$Play.Time,
     main="Board Game Play Time",
     xlab="Play Time",
     ylab="Number of Games",
     breaks = seq(0, 60000, l=2001),
     xlim = c(0, 720),
     xaxp = c(0, 720, 6),
     col="orchid1")

hist(bgg$Play.Time,
```

```
    main="Board Game Play Time",
    xlab="Play Time",
    ylab="Number of Games",
    breaks = seq(0, 60000, l=181),
    xlim = c(0, 60000),
    xaxp = c(0, 60000, 6),
    col="orchid1")

playTimeSD = sd(bgg$Play.Time)
playTimeMean = mean(bgg$Play.Time)
table(bgg$Play.Time >= (1 * playTimeSD) + playTimeMean)
table(bgg$Play.Time >= (2 * playTimeSD) + playTimeMean)
table(bgg$Play.Time >= (3 * playTimeSD) + playTimeMean)
table(bgg$Play.Time >= (4 * playTimeSD) + playTimeMean)
table(bgg$Play.Time >= (5 * playTimeSD) + playTimeMean)
table(bgg$Play.Time >= (6 * playTimeSD) + playTimeMean)
table(bgg$Play.Time >= (7 * playTimeSD) + playTimeMean)
table(bgg$Play.Time >= (8 * playTimeSD) + playTimeMean)

hist(bgg$Year.Published,
    main="Board Game Release Dates",
    xlab="Release Year",
    ylab="Number of Games",
    breaks = 250,
    #xlim = c(1950, 2030),
    #xaxp = c(-3500, 2022, 6),
    col="orchid1")
```

```
boxplot(bgg$Year.Published,
        main="Complexity Box Plot",
        ylab="Board Game Complexity",
        col = "orchid1")

plot(bgg$Year.Published, bgg$Play.Time,
     pch=16,
     cex=0.4,
     xlab="Year Published",
     ylab="Play Time",
     col="orchid")
plot(bgg$Year.Published, bgg$Play.Time,
     pch=16,
     cex=0.4,
     xlab="Year Published",
     ylab="Play Time",
     xlim=c(1850, 2022),
     ylim=c(0, 1440),
     col="orchid")
abline(lm(bgg$Play.Time ~ bgg$Year.Published), col = "blue", lwd = 2)
cor(bgg$Year.Published, bgg$Play.Time)

plot(bgg$Year.Published, bgg$Complexity.Average,
     pch=16,
     cex=0.4,
     xlab="Year Published",
     ylab="Complexity",
     col="orchid")
```

```
plot(bgg$Year.Published, bgg$Complexity.Average,
    pch=16,
    cex=0.4,
    xlab="Year Published",
    ylab="Complexity",
    xlim=c(1850, 2022),
    col="orchid")
abline(lm(bgg$Complexity.Average ~ bgg$Year.Published), col = "blue", lwd = 2)
cor(bgg$Year.Published, bgg$Complexity.Average)

plot(bgg$Complexity.Average, bgg$Play.Time,
    pch=16,
    cex=0.4,
    xlab="Complexity",
    ylab="Play Time",
    col="orchid")
plot(bgg$Complexity.Average, bgg$Play.Time,
    pch=16,
    cex=0.4,
    xlab="Complexity",
    ylab="Play Time",
    ylim = c(0, 720),
    col="orchid")
abline(lm(bgg$Play.Time ~ bgg$Complexity.Average), col = "blue", lwd = 2)
cor(bgg$Play.Time, bgg$Complexity.Average)
summary(lm(bgg$Play.Time ~ bgg$Complexity.Average))

plot(bgg$Rating.Average, bgg$Complexity.Average,
```

```
    pch=16,
    cex=0.4,
    xlab="Rating",
    ylab="Complexity",
    col="orchid")
abline(lm(bgg$Complexity.Average ~ bgg$Rating.Average), col = "blue", lwd = 2)
cor(bgg$Rating.Average, bgg$Complexity.Average)
summary(lm(bgg$Complexity.Average ~ bgg$Rating.Average))
```