EDA Mini Project -Board Game Analysis

Group 2 - Chiu Wing Fung, Aaron; Wong Shing Fung, Dov

Project Objective

This project aims to analyze the elements of board game and the relationship between them. After that, we would like to find a formula that how can we make a high rating and/or bestselling board game.

Data Description

The data set is downloaded from Kaggle which extracted from BoardGameGeek (BGG) which contain 20343 entries and 16 columns.

Categorical Data	Numerical Data	
ID (Nominal)	Year Published (Discrete)	User Rated (Discrete)
Name (Nominal)	Min Players (Discrete)	Rating Average (Continuous)
Mechanics (Nominal)	Max Players (Discrete)	BGG Rank (Discrete)
Domains (Nominal)	Play Time (Discrete)	Complexity Average (Continuous)
	Min Age (Discrete)	Owned Users (Discrete)

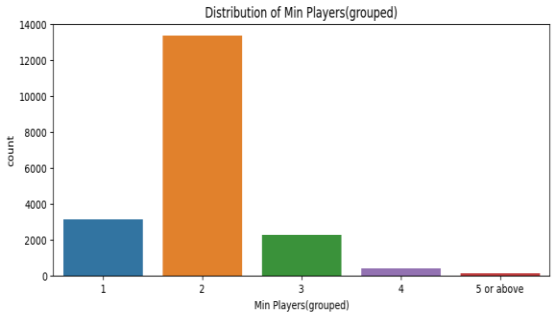
Methodology (Data Cleansing)

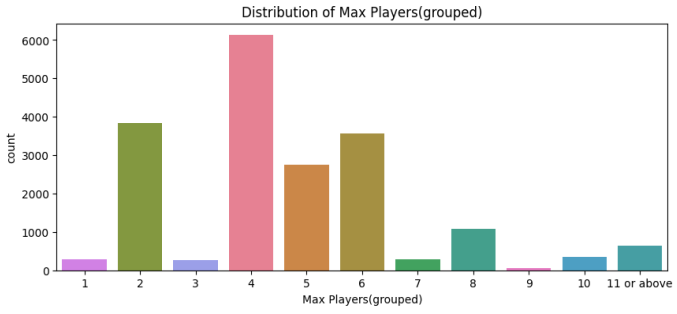
After cleansing the rows with missing value and unreasonable zero values (e.g. Min Players: 0), there are 19264 entries remained (94.70%). As there are too much missing rows (1401 rows after first step of cleansing) for the column of “Mechanics”, we decide to keep it to ensure as much as entries. Instead, we will not use it for further analysis with other columns.

"Owned Users" is an important indicator for bestselling board games. However, the data for this metric is not accurate, as it is collected based on BGG users rather than directly from manufacturers. Therefore, we will give up the objective of analyzing how to make a bestselling game.

Analysis

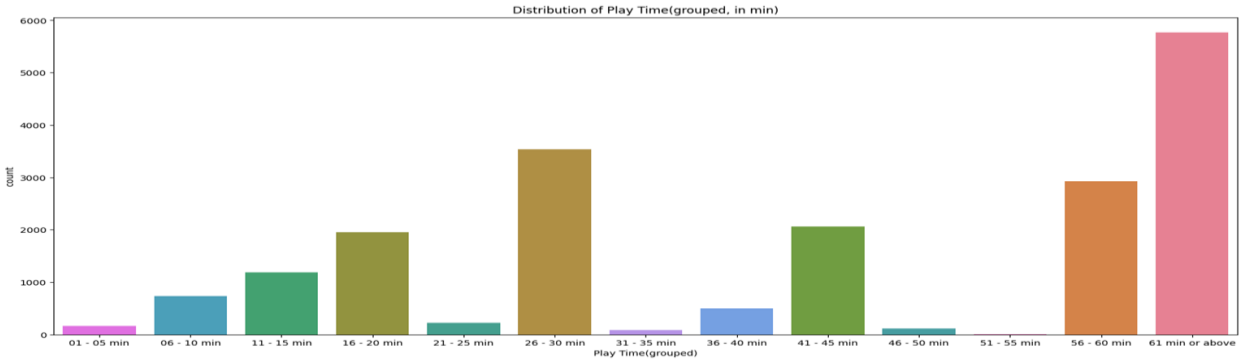
Univariate Analysis



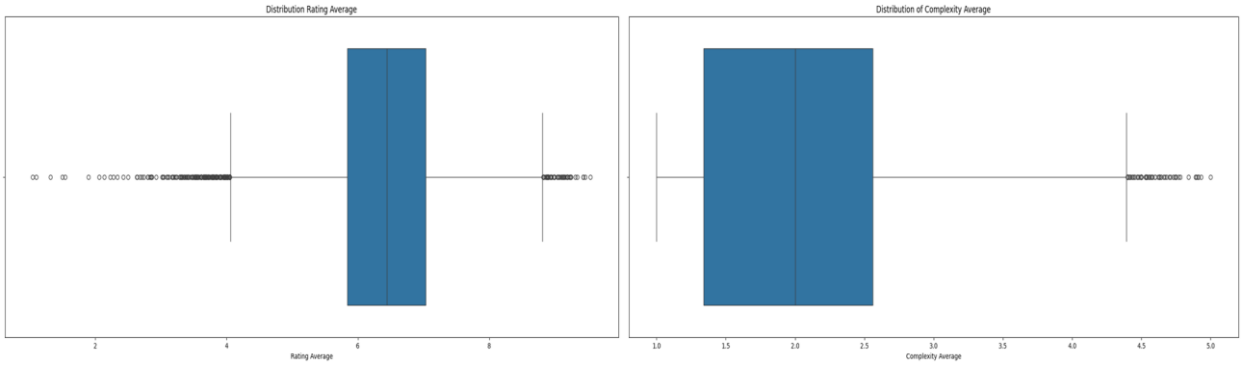


As the number of entries with 5 or more players in the "Min Player" feature is too small, we grouped them into a "5 or above" category. Most of the board games need at least 1 to 3 players to play, and 70% of them require 2 players.

We also did some grouping for "Max Players". Most games can only accept 4 to 6 players, and there are also nearly 4,000 games that can only accept 2 players.



Same as before, we did some grouping for the "Playing Time" feature. Most games need 56 minutes or more to play.

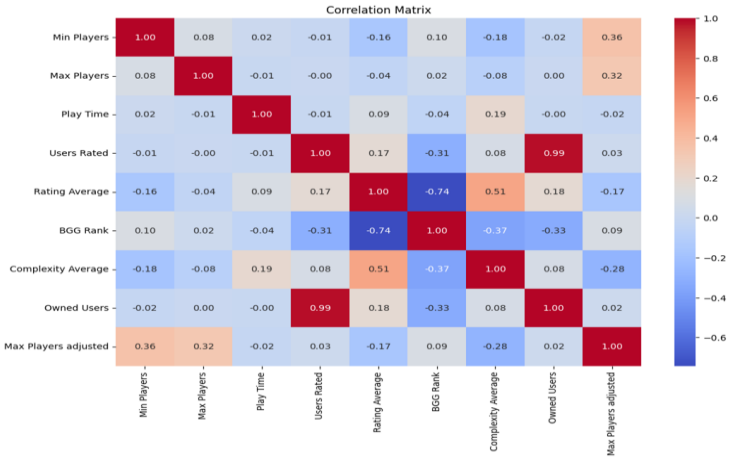


Excluding outliers, the range of the "Rating Average" feature is from 4.05 to 8.82. However, since the outliers with high ratings are important for our objective, we have decided to keep them.

Excluding outliers, the range of the "Complexity Average" feature is from 1.0 to 4.39. Since there are only 0.36% of outliers and their values are not too extreme, we have also decided to keep them.

From the analysis of the word cloud visualization, we can see that there are some common game mechanics with high frequency, such as Dice Rolling and Hand Management.

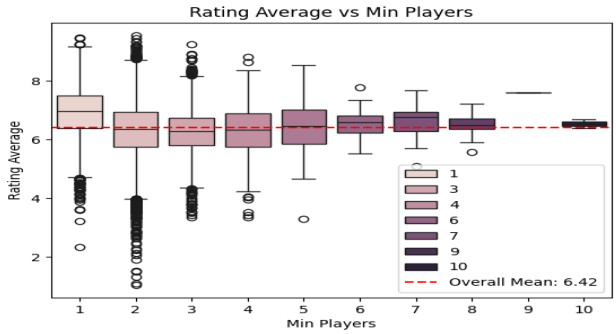
Bivariate Analysis

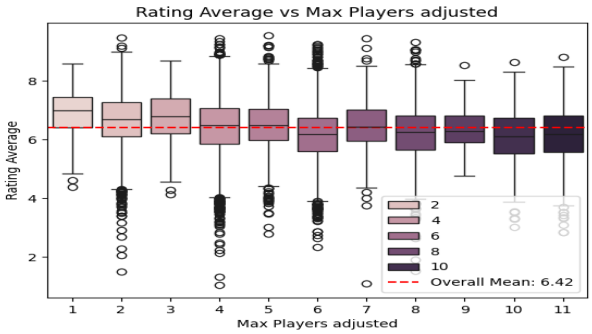


There are high correlation coefficients between various elements. Some of the factual relationships identified include:

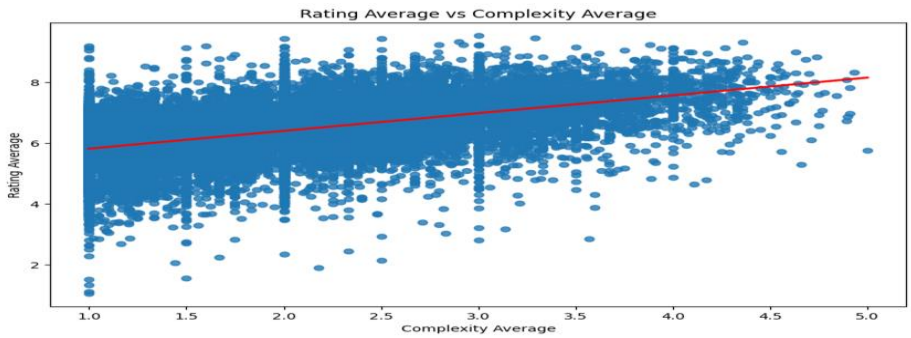
- "BGG Rank" and "Users Rated" (-0.31)
- "BGG Rank" and "Rating Average" (-0.74)
- "Owned Users" and "Users Rated" (0.99)
- "Owned Users" and "BGG Rank" (-0.33)

Furthermore, there is a positive correlation between "Complexity Average" and "Rating Average" (0.51). Also, there is a negative correlation between "Max Player adjusted" (amending values of 11 or more to 11 for analysis) and "Complexity Average" (-0.28).





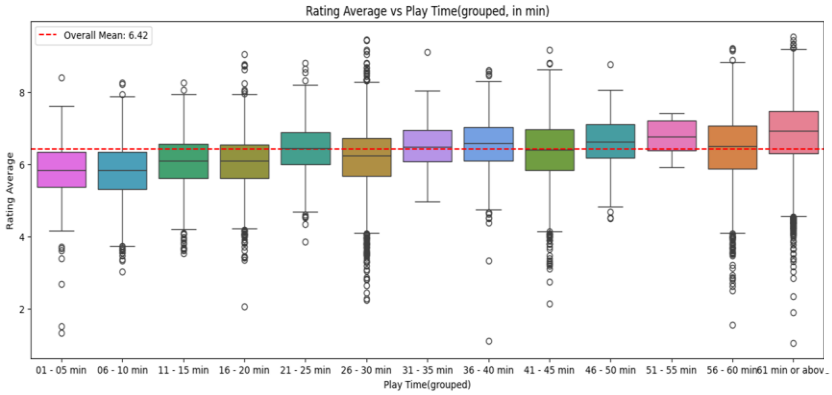
Board games with a "Min Players" value of 1 seem to have higher ratings. Board games with 1 - 3 "Max Players" seem to have higher ratings compared to games with higher maximum player counts. The ratings tend to decrease as the "Max Players" value increases.




```
slope: 0.58
intercept: 5.23
R-squared: 0.26
p-value: 0.0000
stderr: 0.01
```

From the result of the analysis, the p-value is 0.0000, which means there is a significant relationship between "Rating Average" and "Complexity Average". However, the relationship is not strong, as indicated by the R-squared value of 0.26. The equation derived from the analysis is:

- Rating Average = 5.23 + 0.58 x Complexity Average



The boxplots showed that more play time, higher rating. It is reasonable because complex games have higher rating and complex game need more play time, so board games with more play time will have higher rating.

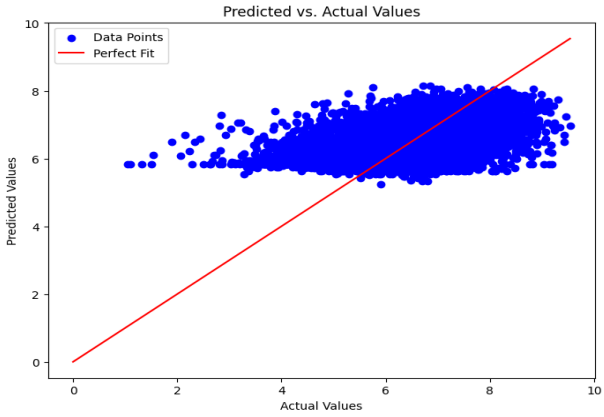
Multivariate Analysis

We will use linear regression to analyze with “Min Player”, “Max Player adjusted”, “Play Time adjusted” (amend value of 60 or more to 60 for analysis) and “Complexity Average” as independent variables and “Rating Average” as dependent variable.

```
df2["Play Time adjusted"] = df2["Play Time"].apply(lambda x: 60 if x >= 60 else x)

x = df2[['Min Players', 'Max Players adjusted', "Play Time adjusted", 'Complexity Average']]
y = df2['Rating Average']

model = LinearRegression()
model.fit(x, y)

print('Intercept:', model.intercept_)
print(f'Coefficients: {[round(c, 6) for c in model.coef_.tolist()]}')
print('R-squared:', model.score(x, y))
```

✓ 0.0s Python

```
Intercept: 5.464091694895461
Coefficients: [-0.097601, -0.001535, 7.6e-05, 0.567303]
R-squared: 0.2651588899041728
```



Although R-squared is not high (0.265), but we still get an equation: Rating average = 5.46 - 0.098 x Min Players - 0.0015 x Max Players + 0.000076 x Play Time (in min) + 0.57 x Complexity Average. Then, we use this equation to calculate the predicted rating and plot a graph to compare the difference between predicted values and actual values.

From the graph, we can see that the equation cannot predict the result accurately.

Challenges and Limitations

Challenges

1. It is hard to decide how to handle the missing data, zero values and outliers. It depends how we use those columns and the number of them.
2. As different analysis and graphs are required for different kinds of data. It is important to choose the appropriate analysis and visualization methods.
3. As we are both beginner in data engineering field, we must explore many resources to learn how to plot the graphs and perform the analysis. Debugging also take much time.

Limitations & Future Improvement

1. The dataset from BGG is limited to the player of certain countries and language. It cannot reflect the whole picture of the board game player population. We should obtain data from more platforms with various countries and language.
2. There is too much missing information especially mechanics. It is also one of the important factors to affect the rating of board game. The effect of single mechanic and combination of mechanics on rating average will be useful.
3. The data of "Owned Users" are inaccurate. It is an important factor in commercial perspective. If it is possible, we collect "Number of sales" from manufacturers directly although the sales number of pirated board games will not be counted.
4. We should perform more multivariate analysis with different combination to try to find out a better prediction model.

Future Work

We decide to give up any future work for this data set. As mentioned in the limitation, the sample of this dataset cannot reflect the population. However, BGG is the biggest and popular websites about board game. It is hard to get the dataset with similar size from other platforms.

Conclusion

To summarize, when "Min Players" or "Max Players" increase, the rating will decrease. 1 of "Min Player" and 1 – 3 of "Max Player" have the high "Rating Average".

On the other hand, when "Complexity Average" and "Play Times" increase, the rating will also increase. From boxplot form "Complexity Average", as values of 4.39 or above are outlier, so we should try to make a board game with complexity of about 4.39 will be more possible. Also, we should try to make it to be played for at least 60 minutes.

From the analysis, we obtain 2 equations:

1. Rating Average = 5.23 + 0.58 x Complexity Average
2. Rating average = 5.46 - 0.098 x Min Players - 0.0015 x Max Players + 0.000076 x Play Time (in min) + 0.57 x Complexity Average

However, our findings may not reflect the real situation. Not all board game players are BGG users, especially casual players. We believed that most of BGG users will be the experienced players and seems like experienced players prefer complex games. So, it is the reason why complex game get higher rating.

It is hard to gather many players to player a complex (not all people love playing complex game, they just want to relax) game and it may take ton of time for 10 or more player to play a complicated game. Therefore, the games for many players will be simpler and the games for few players can be very hard, such as Chess, GO etc.

In addition, high rating may not reflect the number of sales. In commercial perspective, game designers and companies want to design a bestselling game instead of high rating game.

Reference

https://www.kaggle.com/datasets/melissamonfared/board-games

https://www.fun.com/best-selling-board-games-all-time.html

https://boardgamegeek.com/

Distribution of Work

All works including data cleansing, analysis, PowerPoint and report are done together.