Stats Final Project

AUTHOR
Ken Shimokawa

PUBLISHED
June 9, 2024

1.) Preparing and describing the data

Professor Vossen is looking for advice for his new board game: Academia the Board Game. The goal of this board game is to simply make our professor the most money possible. To help advise him the data set Board Game Review Rating Predictor is given. This data set is produced by extracting data from the website BoardGameGeek. The goal of this analysis will be to advise him in what kind of board game he should make in order to achieve his goal.

A quick inspection of the data show a few features that may be good candidates for measuring the success of a board game: owned (meaning the number of people who own the board game), wanting (the number of people who want to trade another board game for the one listed) wishing (the number of people who have that board game in their wishlist) and trading (the number of people who want to trade away their copy).

When considering the goal of making the most money possible it is clear that the feature wanting is irrelevant. This is because when a person trades for our board game from somebody else Professor Vossen does not receive any revenue. In addition wishing is an unsuitable feature because this simply measures the number of people who potentially want to buy a board game but for whatever reason hasn't. Meaning that the producers of the board do not receive any revenue.

Excluding wanting and wishing we use the features owned and trading as measures of success. While this may initially seem counter intuitive to include the feature since owned does not qualify for means of acquisition (meaning they may have traded for the board game) and a higher number of trading means a higher number of people that want to trade away their board game in actuality higher numbers of both means more copies of the board game sold. Even though the customer may have had a terrible experience with the game and traded it away, what happened after the initial purchase and revenue for the producer for the purposes of this analysis is irrelevant.

Additionally, there are a number of variables that could be potentially used to inform game design choices. minimum age, playing time and minimum players. While there are a few more game design features the ones listed lower the barrier to entry for potential customers to play the game. While other features wouldn't limit the market for board games at all.

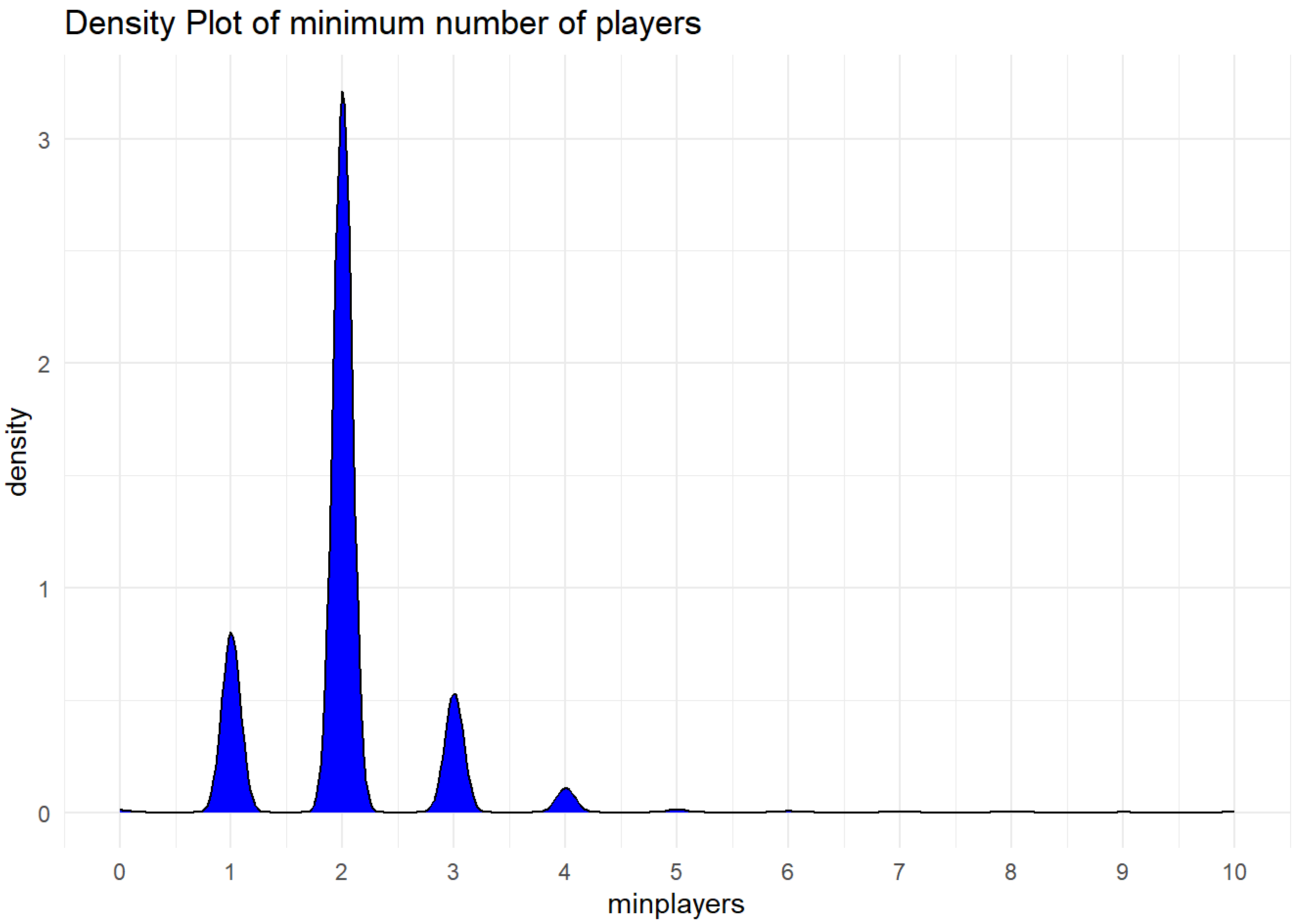


Figure 1 Density plot of minimum number of players

```
    0     1     2     3     4     5     6     7     8     9    10 
   44  3675 14834  2462   503    59    24    11    16     1     2 
```

figure 2 Distribution table of minimum number of players required to play the game

Here the density plot and distribution table show that most of minimum players required to play a board game is below 3. This can lead to numerous problems that result from having a small variance such as reduced discriminatory power, increased risk of multicolinearity, difficulty in detecting relationships and inadequate representation of population.

Discriminatory power

Discriminatory power refers to the ability of a variable to distinguish between different groups or outcomes. When variance is low, the values of the variable are very similar. This lack of variation makes it difficult to differentiate between different categories or groups based on that variable.

Multicolinearity

Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, leading to unreliable estimates of their effects. Low variance in one predictor can make it highly correlated with other predictors that also have low variance or are measuring similar constructs. When there is Multicollinearity its hard to decipher which of the colinear effects are being measured

Detecting Relationships

Detecting relationships involves identifying meaningful associations or correlations between variables. With low variance, it is hard to detect any existing relationships because the range of values is too narrow to reveal meaningful patterns. In the case of board games since most of our board games require 2 players to play it becomes more difficult to measure the success of board games that don’t require 1 or more than 2 players.

Inadequate representation of other board games

As stated previously having such a small variance in our independent variables make it difficult to measure the effects of the minimum number of players that aren’t two. If the board games listed on boardgamegeek aren’t an accurate representation of the broader population and the distribution for minimum players is more evenly distributed our model would be obsolete.

An important factor to consider that initially may seem counter intuitive is the disregard of ratings. While high ratings may be an indicator of board game sales for this analysis which informs on the type of board game to create since make a “good board game” is obvious advise and not a type of board game this feature will be disregarded.

Finally an important variable that we should control for is release year. Release year is an important factor to control for when considering longitudinal data or data that is collected over a long period of time. In the case of board games it could be that older games that have been in the market more sell more copies just because consumers have had a longer time to purchase them. This is most likely not an issue for unpopular board games but among the successful board games older games would have a longer time to sell copies.

2.) Analysis

As Identified previously our dependent variables will be owned and trading. Our dependent variables shall be minimum age, playing time and minimum players. The model will also control for release year as discussed earlier. Since there is more than one dependent variables a multivariate regression will be used.

```
model <- lm(cbind(owned, trading) ~ minage + minplayers + playingtime + yearpublished, d

summary(model)
```

```
Response owned :

Call:
lm(formula = owned ~ minage + minplayers + playingtime + yearpublished,
    data = details)
```

```
Residuals:
   Min     1Q Median     3Q    Max 
 -3247  -1397  -1018   -315 167063 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    451.51809  369.42700   1.222 0.221641    
minage         120.45827   10.10025  11.926  < 2e-16 ***
minplayers    -201.13445   53.10167  -3.788 0.000152 ***
playingtime     -0.05841    0.06842  -0.854 0.393254    
yearpublished    0.14481    0.17480   0.828 0.407449    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5376 on 21626 degrees of freedom
Multiple R-squared:  0.007232,  Adjusted R-squared:  0.007048 
F-statistic: 39.38 on 4 and 21626 DF,  p-value: < 2.2e-16


Response trading :

Call:
lm(formula = trading ~ minage + minplayers + playingtime + yearpublished, 
    data = details)

Residuals:
    Min      1Q  Median      3Q     Max 
 -91.58  -38.12  -25.99   -2.25 2469.88 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -3.580956   6.979722  -0.513   0.6079    
minage         3.554120   0.190828  18.625   <2e-16 ***
minplayers    -0.554271   1.003270  -0.552   0.5806    
playingtime   -0.002100   0.001293  -1.624   0.1043    
yearpublished  0.007205   0.003303   2.182   0.0292 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 101.6 on 21626 degrees of freedom
Multiple R-squared:  0.01651,   Adjusted R-squared:  0.01633 
F-statistic: 90.76 on 4 and 21626 DF,  p-value: < 2.2e-16
```

Response owned:

The estimated intercept is 451.52, but it is not statistically significant (p = 0.221641). This suggests that when all independent variables are zero, the expected value of `owned` is approximately 451.52, though this value is not reliable. The coefficient for `minage` is 120.46, which is highly significant (p < 2e-16). This indicates that for each one-year increase in `minage`, the expected value of `owned` increases by 120.46. The coefficient for `minplayers` is -201.13, which is also highly significant (p = 0.000152). This suggests that for each additional player, the expected value of `owned` decreases by 201.13. The coefficient for `playingtime` is -0.05841 and is not statistically significant (p = 0.393254), implying that `playingtime` does not have a meaningful effect on `owned`. A negative coefficient also implies a negative linear relationship meaning as playing time increase the number of copies ownded decreases. The coefficient for `yearpublished` is 0.14481 and is not statistically significant (p = 0.407449), indicating that the year a game was published does not significantly affect the number of times it is owned. The multiple R-squared value is 0.007232, which means that only 0.72% of the variance in `owned` is explained by the model. The adjusted R-squared is similarly low at 0.007048, suggesting that the model does not fit the data well. The F-statistic is 39.38 with a p-value < 2.2e-16, indicating that the model is statistically significant overall, even though individual predictors (playingtime and yearpublished) are not.

Feature	Estimate	Significance (yes/no)
Intercept	451.52	No
Minage	120.46	Yes
Minplayers	-201.13	Yes
Playing Time	-0.05841	No
Year Published	0.14481	No

Response trading:

The estimated intercept is -3.58, which is not statistically significant (p = 0.6079). This implies that when all independent variables are zero, the expected value of `trading` is -3.58, though this value is not reliable. The coefficient for `minage` is 3.55, which is highly significant (p < 2e-16). This suggests that for each one-year increase in `minage`, the expected value of `trading` increases by 3.55. The coefficient for `minplayers` is -0.55, which is not statistically significant (p = 0.5806), indicating that the number of players does not significantly affect `trading`. The coefficient for `playingtime` is -0.0021 and is not statistically significant (p = 0.1043), implying that `playingtime` does not have a meaningful effect on `trading`. The coefficient for `yearpublished` is 0.0072, which is statistically significant (p = 0.0292). This suggests that the more recent the game is published, the slightly higher the expected value of `trading`. The multiple R-squared value is 0.01651, meaning that only 1.65% of the variance in `trading` is explained by the model. The adjusted R-squared is similarly low at 0.01633, indicating that the model does not fit the data well. The F-statistic is 90.76 with a p-value < 2.2e-16, suggesting that the model is statistically significant overall, despite the low R-squared value.

Feature	Estimate	Significance (yes/no)
Intercept	-3.58	No
Minage	3.55	Yes
Minplayers	-0.55	No

Feature	Estimate	Significance (yes/no)
Playingtime	-0.0021	No
YearPublished	0.0072	Yes

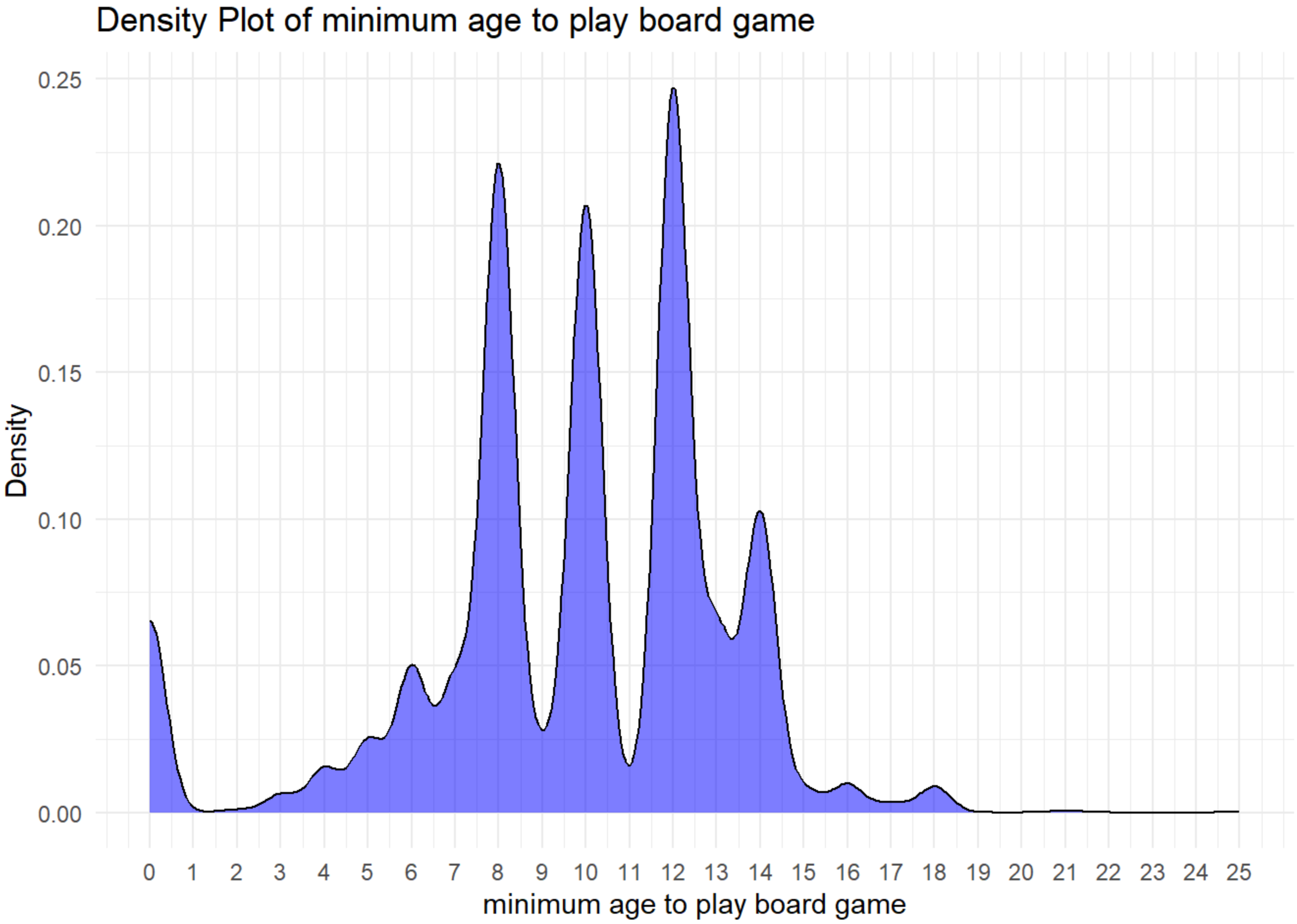
Both Models showed showed that `minage` was a significant variable while `minplayers` was only significant for `owned` but not `trading`. Additionally, `year published` was only significant for `trading`. Overall the model seems appropriate based on the output.

An alternative way to test whether the relationships are significant could be by implementing a K-fold cross validation analysis. A K-fold Cross validation works by splitting the data into the training data set, validation training data set and testing data set and works very well for well balanced data sets. By using different folds in each iteration the model is tested on different subsets of data and can be used to assess the stability and significance of each data set. In addition using different folds would solve any issues of over fitting where the model is trained to fit the training data set too well where it would struggle in making accurate predictions for future data points.

3.) Strategy recommendations

Considering the output of our multivariate regression analysis, we identified minimum age as the most significant variable, followed by the minimum number of players and the year of publication. Interestingly, contrary to our initial hypothesis, minimum age is positively correlated with our success metrics. This indicates that board games with higher minimum age requirements tend to be more frequently traded or owned, suggesting that these games are more successful. At first glance, it seems counter intuitive that games with a higher barrier to entry (i.e., higher minimum age) are more successful.

One potential reason is the Children’s Product Certificate requirement by the United States Consumer Safety Commission. According to the CPSC, products designed for children 12 years or younger must undergo third-party testing and certification, which incurs additional costs . To avoid these costs, manufacturers might set the minimum age at 13, not necessarily reflecting the game’s actual complexity or content. Thus, the minimum age might not be an accurate indicator of the game’s characteristics but rather a strategic decision to avoid regulatory costs.



Here, a density plot of the minimum ages of each board game can be seen. There are three spikes in the minimum age of board games at: 8, 10 and 12.

```
Percentage of board game with the age limit under 13 : 82.50196 %
```

This number represents what percentage of the board games with the minimum age under 13. In addition to the density plot this figure shows that most board games are in fact not larger than 13 to avoid the additional costs due to the Children's Product Certificate. While it is not clear that the effect of the Children's Product Certificate does not exist the from the distribution of the minimum age suggest a small effect.

There are several other plausible explanations for this phenomenon. Based on anecdotal evidence (as official guidelines are not available), the age restriction could be due to factors such as the reading level required to understand the rules, the thematic content (e.g., games like Risk involving war), the complexity of game play, and the size of game pieces (which could pose a choking hazard for younger children). Among these, complexity seems to be a likely reason for the positive correlation between minimum age and success of a board game. This suggests that more complex games, which often have higher age restrictions, are more appealing and therefore measured as more successful on "BoardGameGeek".

It is plausible to consider that the data is subject to sampling bias. Sampling bias occurs when a sample selected for analysis or study is not representative of the larger population from which it is drawn. This bias can lead to inaccuracies or distortions in the results of the analysis, as the sample does not accurately reflect the characteristics of the population it is supposed to represent.

It is possible that the demographic composition of the site’s users skews towards an older age bracket, thereby potentially indicating a more mature taste among the user base, a characteristic that could manifest in the success metrics.

The negative coefficient for the minimum number of players indicates that games requiring fewer players tend to be more successful, aligning with our hypothesis that a lower barrier to entry leads to greater success. This finding supports the notion that games that are easier to start playing (i.e., do not require assembling a large group) are more popular.

The positive significance of the release year is also noteworthy. This suggests that the board game market has been growing over time, with newer games being more successful. This could reflect increasing interest and investment in board games in recent years.

However, our analysis has several limitations. Firstly, while our current features provided valuable insights, additional data on replay-ability, game mechanics, artwork quality, and social relevance would have enriched the analysis. Although the board game category feature partially covers these aspects, numerical measures for these factors would allow for a more detailed examination.

Secondly, the data-set lacks crucial economic variables such as production costs and retail prices of the board games. Since the goal of this analysis was to maximize profit, understanding profit margins is essential. Additionally, information on marketing budgets and the reputation of the publisher would be critical for a thorough analysis of consumer products. Including these variables would enable more specific and actionable recommendations aligned with the research objective.

In conclusion, while the current analysis provides a strong foundation, incorporating additional data and considering the mentioned limitations would lead to a more nuanced understanding of the factors driving the success of board games. This comprehensive approach would ultimately facilitate more informed decision-making and strategy development.

Reference

[1] “Children’s Product Certificate,” *U.S. Consumer Product Safety Commission*. https://www.cpsc.gov/Business–Manufacturing/Testing-Certification/Childrens-Product-Certificate