1. Cleaning

Before I was able to do any exploration, I needed to do some cleaning of the data. As mentioned in my literature review, there were around 80 different columns, many of which were not useful for analysis (for example, the url of each board game on the site). The full cleaning script can be found in Cleaning.R, attached at the end of this document, but I will summarize the main efforts I made before doing exploration. First, I did an initial removal of any columns that were clearly not useful for analysis (such as url). There were quite a few columns that were average scores based on category, which were not necessary given average score and category were already two separate fields. I also identified columns that I didn't think would be useful but wanted to examine further. Once the initial columns were removed, I checked for constant columns, and double/bijection columns, identifying one column to remove. I then returned to the flagged columns and looked at the amount of null rows present in them as well as the number of unique value/counts of these values for these columns. Sometimes, I found that there while most of the column was not of interest (for example there were many game publishers), a specific value was of interest (for example, "self-published") that could be extracted into a new variable. Once I had cleaned up the variables more, I made the decision to only keep games with more than 50 ratings. While in class, we have discussed not getting rid of any rows, all the literature I read made this cutoff and I feel like it is necessary to get a true measure of the game "goodness". A game with only a few ratings is not representative of the true score of the game as one person could have a very high or very low opinion of one game for some reason. While the amount of reviews to be considered valid is not concrete, I am following the literature and going with the 50 ratings threshold.

One major transformation I had to make of the values was converting the family, category and mechanic columns into useable fields. In the raw dataset, each of these columns could have multiple values, for example, one entry in the family column might be "cards, fantasy, dice" while another could be "racing, animals". I was able to find a library that separated each of these comma separated items into its own column, and then I used both my research questions and the prevalence of certain categories, families, or mechanics, to decide which ones to keep.

To support my research question about if the range of players matters, I also created a new variable called player range. I also created my binary predictor variable by score by assigning "No" to a game with a score less than 7 and "Yes" to a game with a score greater than 7.

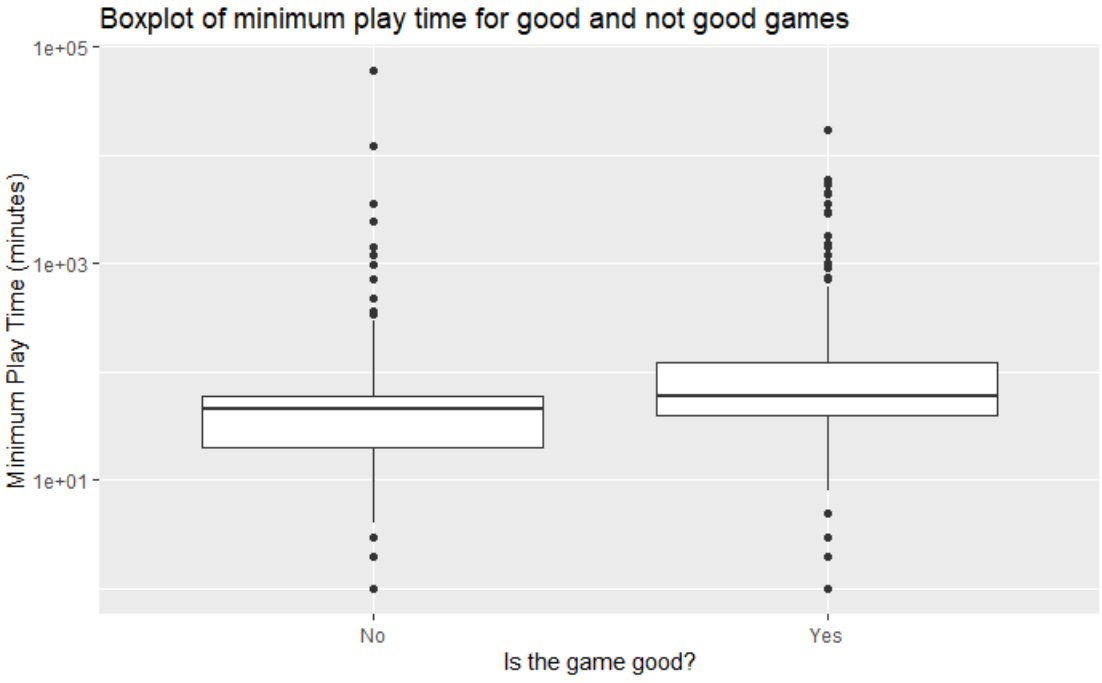
2. Exploration

Once this initial cleaning was complete, I ended up with 34 columns. This is still more than I would probably want to put into the model, which is where the exploration comes into play. I wanted to make sure I examined all the questions I asked in my literature review and other questions that would arise based on the data I had cleaned. The questions I asked were specific to values, but for this exploration I want to look at the general case/trends as well.

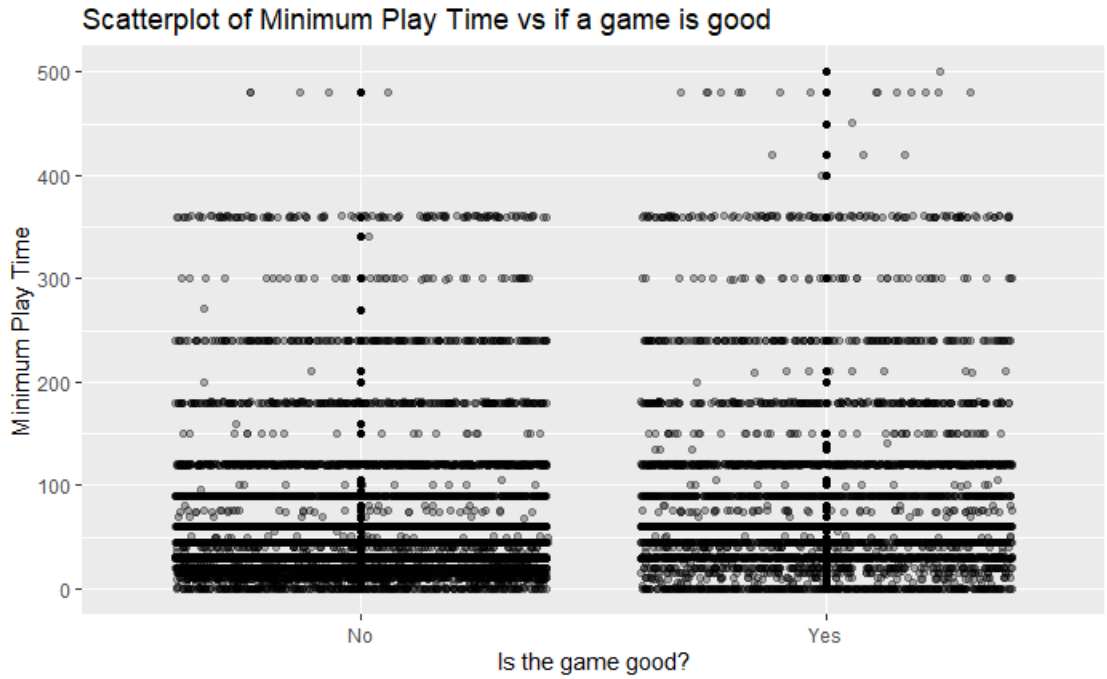
2.1. Research Questions

2.1.1. Are longer games generally classified as "good"? Specifically, how does the prediction of a game being "good" for a min length of 15 minutes compare to a game with a min length of 45 and 90 minutes?

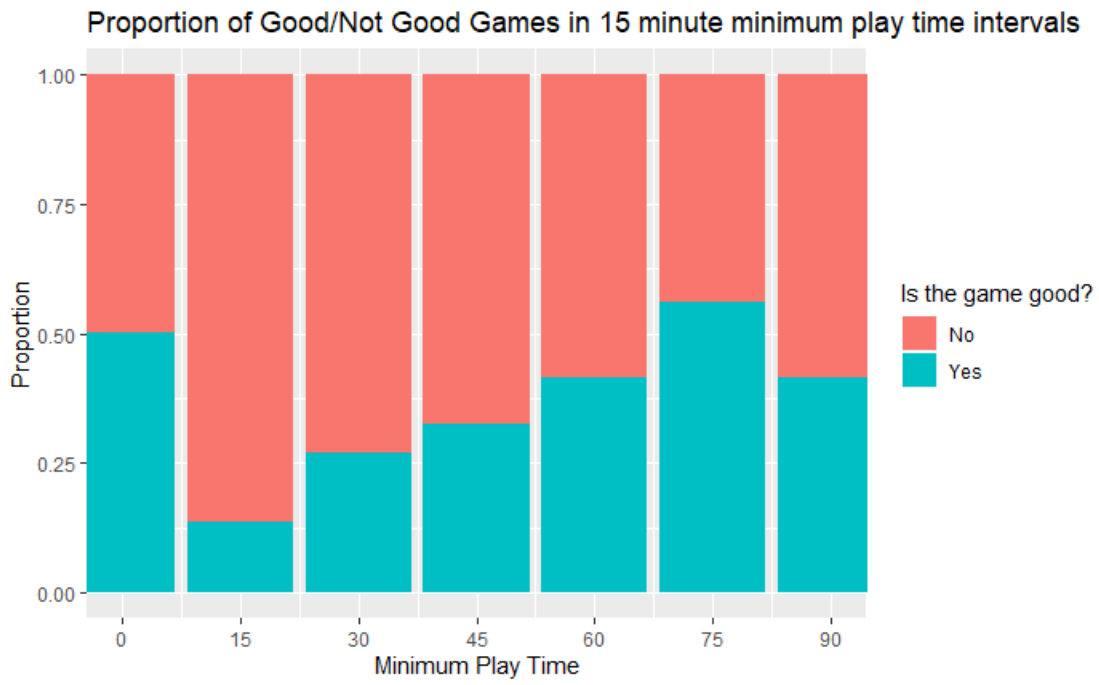
First, I looked at a boxplot of the good vs not good games for lengths. The boxplot is especially useful since there some extreme outliers in the dataset (this is also why I had to use a log scale). The boxplots showed that the median playing time for good games is slightly higher than that of not good games (about 60 minutes versus 90 minutes). Interestingly, the 3rd quartile for not good games is also below the median of the not good games.



A scatterplot sheds more light on this, showing that while similar numbers of games that have higher play times are classified as good/not good, there is a higher density of low playing time games that are not good.

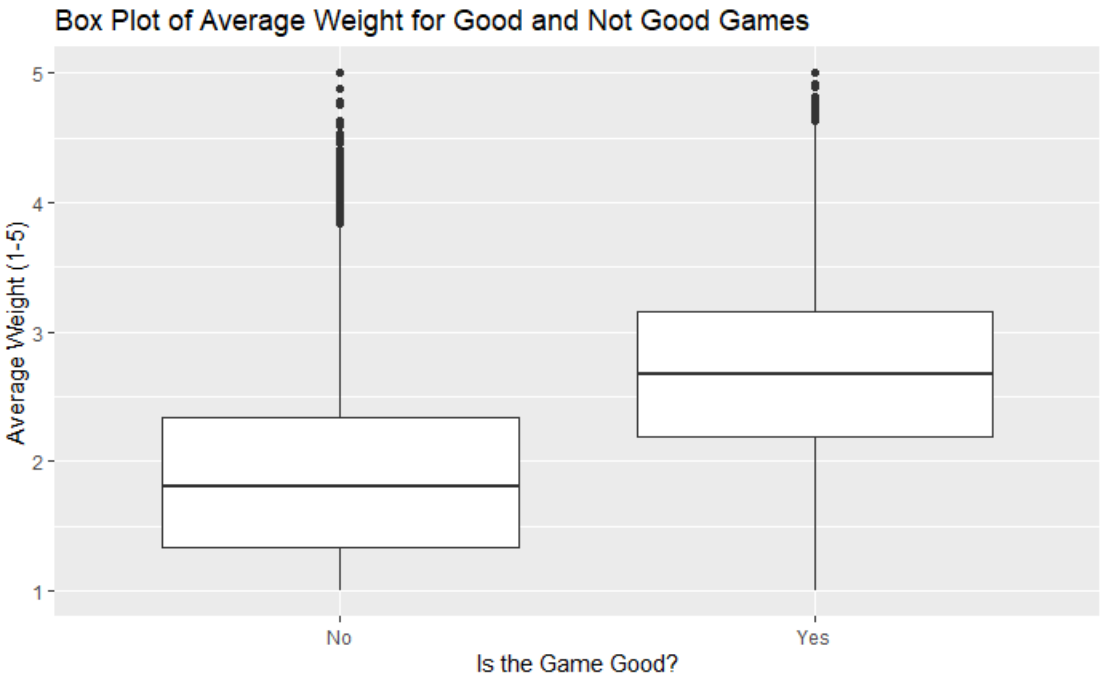


Finally, I wanted to look closer at my original question, so I zoomed in on games with a play time between 0 and 90 minutes. The graph below is broken up in 15 minute intervals and shows the proportion of games that are classified as good or not good in each category. Interestingly, a large portion of games between 0-15 minutes are good as well as a large portion of games between 75-90 minutes. However, games between 15-30, 30-45, 45-60 and 60-75 have less good games (increasing in proportion from the 15-30 category to the 60-75 one). This suggests to me that if a game is going to be quick, it is more favorable for it to be very quick, otherwise longer games are more favorable.

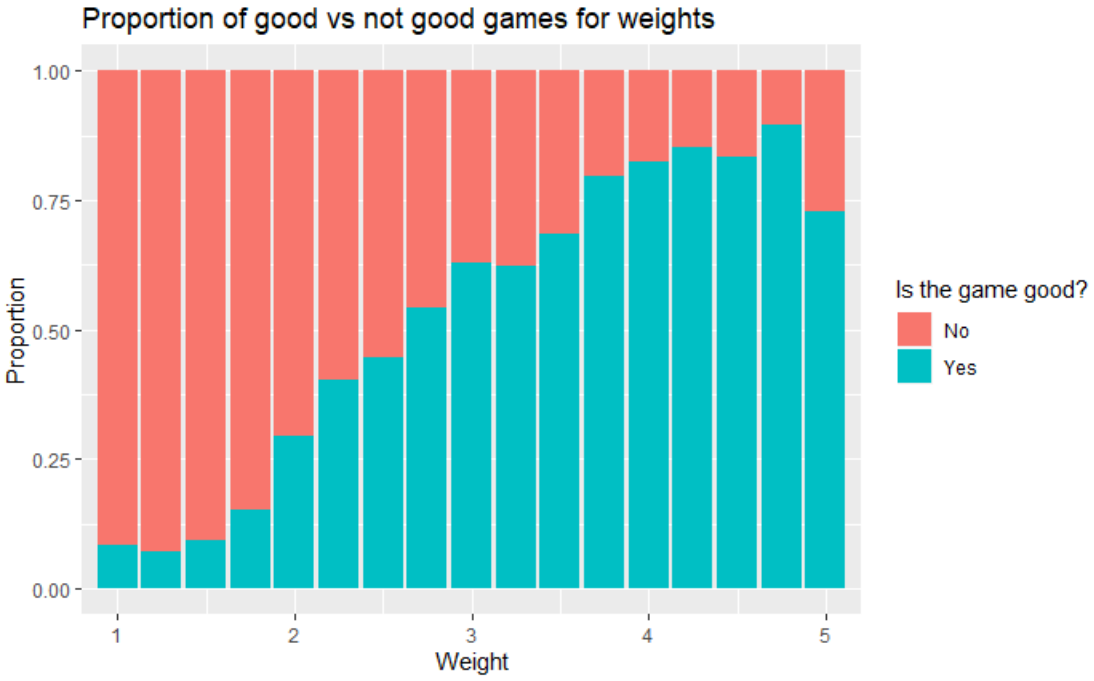


2.1.2. Are heavier games generally classified as "good"? Specifically, how does a game with a weight of 1.5 compare to a weight of 4?

When I started to look at this column, I first noticed that despite the weight scale going from 1-5, there were several entries that had a weight of zero. I looked up a few games with a weight of zero on board game geek to understand the dataset better, and saw that games with a weight "0" are actually games that have not been given a weight, so these values should be changed to NA. I will have to figure out how to deal with these missing values in my model in later phases. I changed this in my cleaning script and will have to keep this in mind when I look at weight. I took a similar approach for weight of the game as I did for minimum play time, first looking at a box plot comparison.

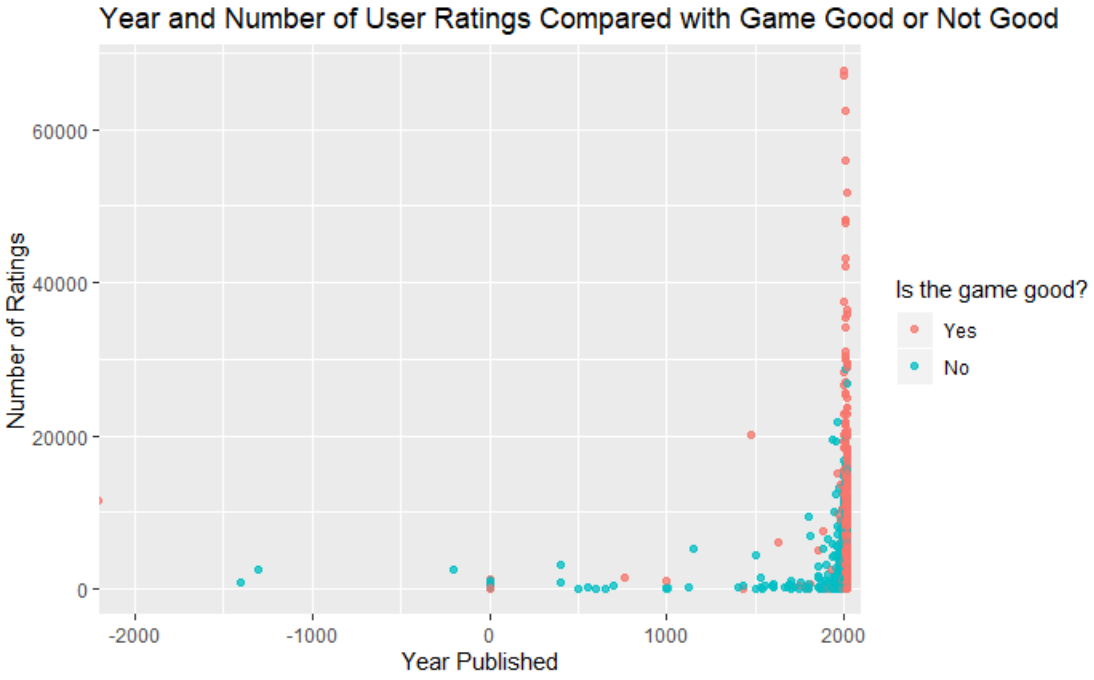


Here, there is a very clear distinction in the weights of games that are good/not good. The 3rd quartile for not good games is very close to the 1st quartile of the good games, and the median for not good games is around 2.75 vs around 3.5 for good games. Looking at another proportional histogram, there is a general trend that the proportion of good games goes up as the weight goes up. Based on this histogram, it seems a game with a weight of 1.5 is less likely to be good than a game with a weight of 4.

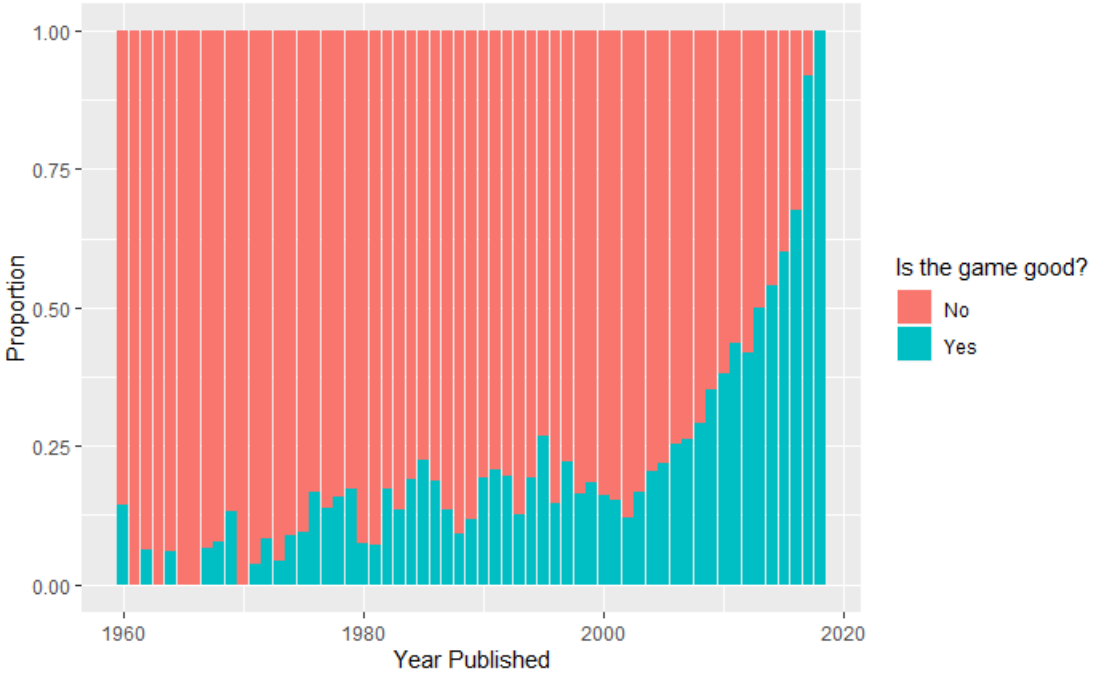


2.1.3. Are older generally classified as "good"? Specifically, how does a game from 2005 compare with a game from 2019?

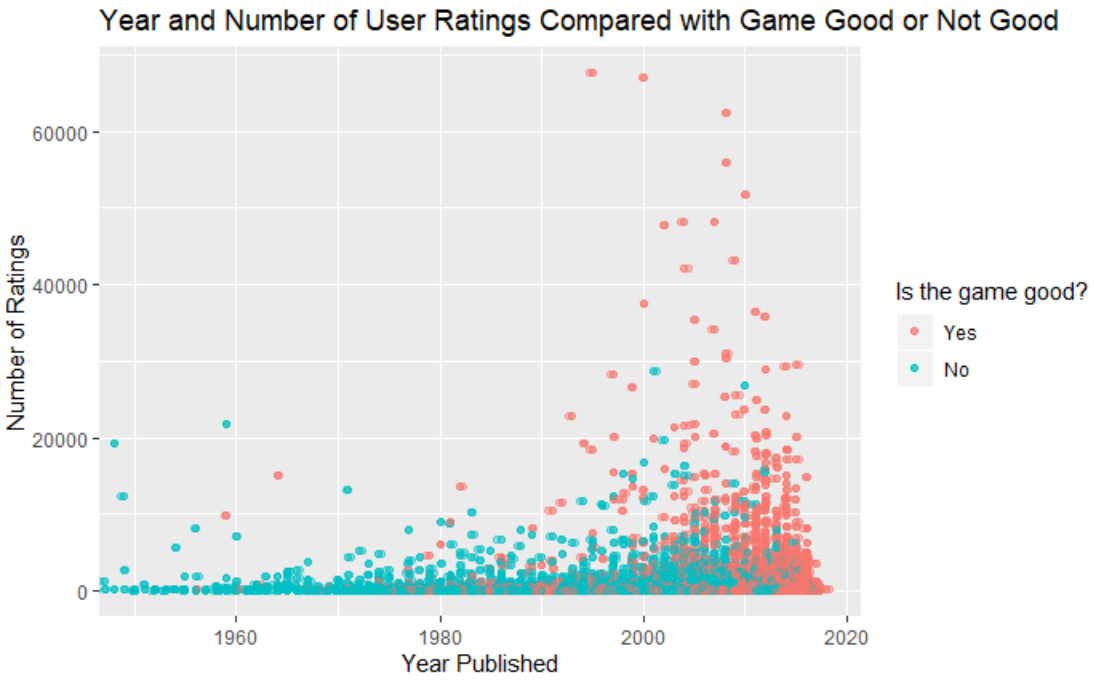
At first, I looked at games with a range of -2000 to 2019. However, most games are from 1960 or later and most games before 1960 are considered not good.



Zooming in from 1960 or later makes it easier to see other trends in the data. Looking at a proportional graph of the years and good/not good games, there is a general trend that more recent games tend to be more highly rated than older games, especially for games after 2010. This suggests that a game from 2005 would likely not be as highly rated as a game from 2019.

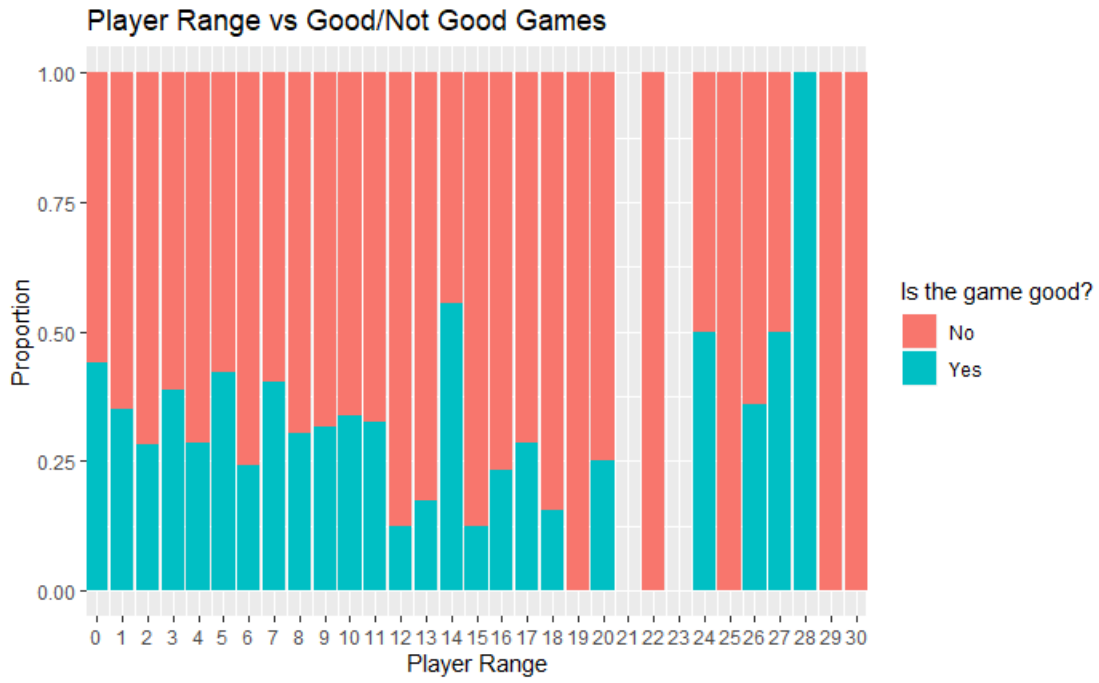


Looking at a scatterplot from 1950 onward shows a similar story, with most “good” games being from later years. I was also interested in comparing the number of ratings versus the year published, since I felt that new games might have more ratings. This scatterplot shows that newer games do typically have more ratings in addition to being more likely to be rated as “good”.

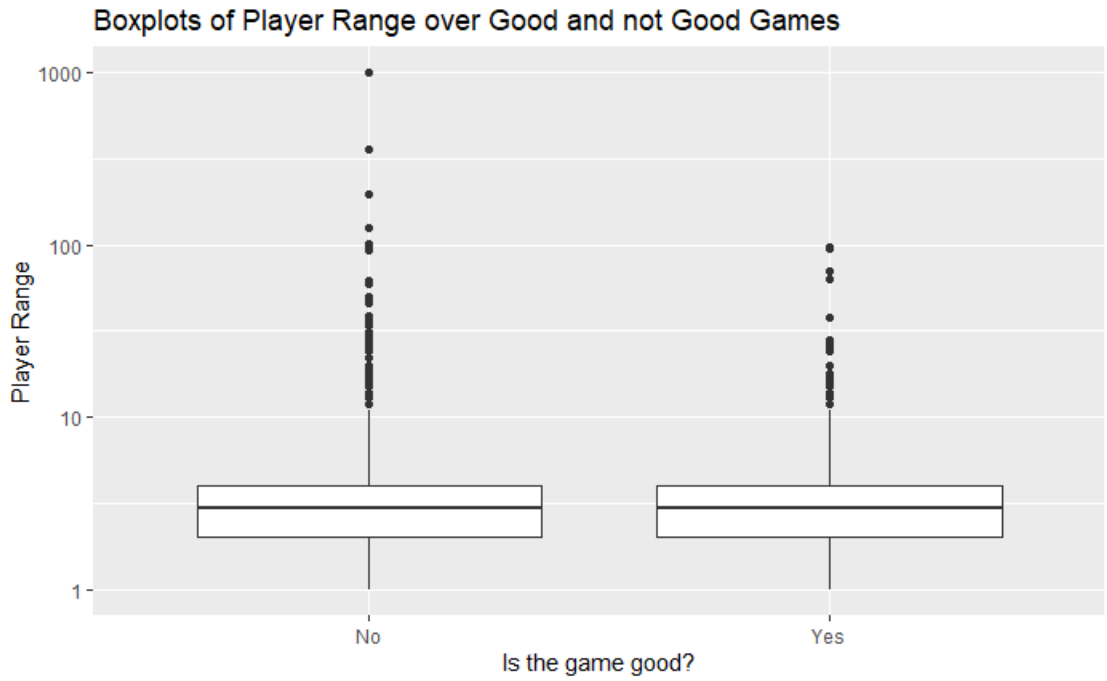


2.1.4. Are games with a larger player range more likely to be classified as “good”? Specifically, does a game with only 2 possible plater amounts have a higher probability of being “good” than a game with 4 possible player amounts?

There doesn’t seem to be much of a relationship between player range and if a game is good or not based on the proportional bar graph.

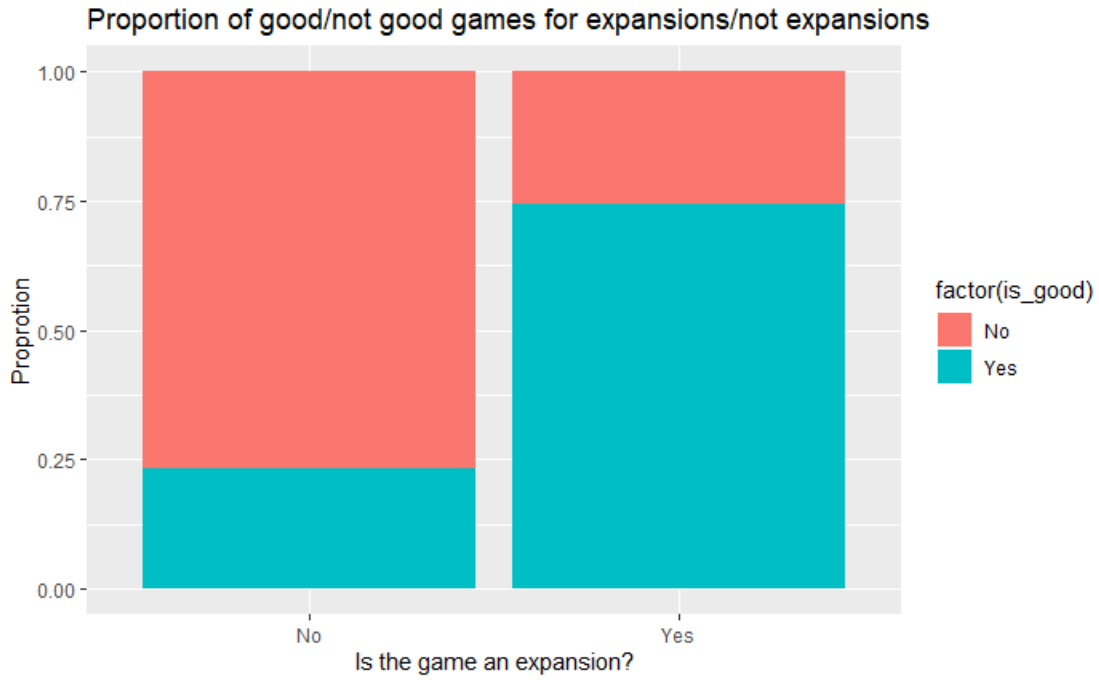


Likewise, the boxplot shows the distribution of good/not good games for player ranges are very similar. This doesn't seem to be a good predictor.



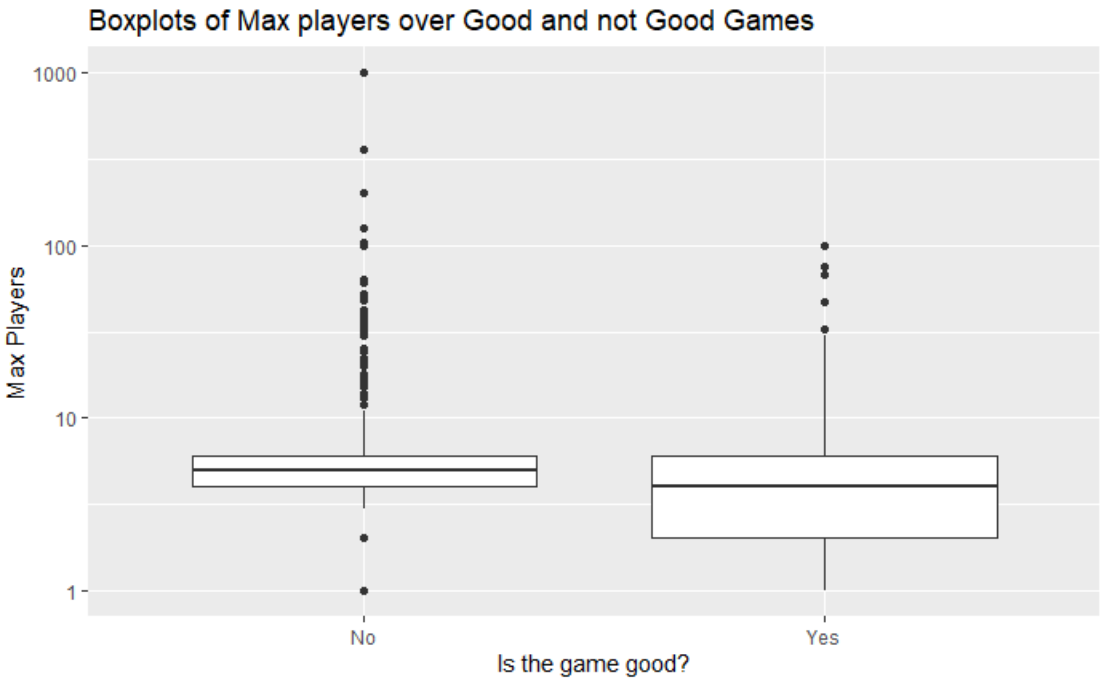
2.1.5. Do expansions have a lower probability of being "good"?

There is a distinct proportional difference between good/not good games for expansions. Almost 75% of expansions are considered "good" while only about 25% of base games are considered good. In a way, this makes sense, since generally games that get expansions are games that initially sold very well. I think this will be a useful factor for predicting good vs not good games.

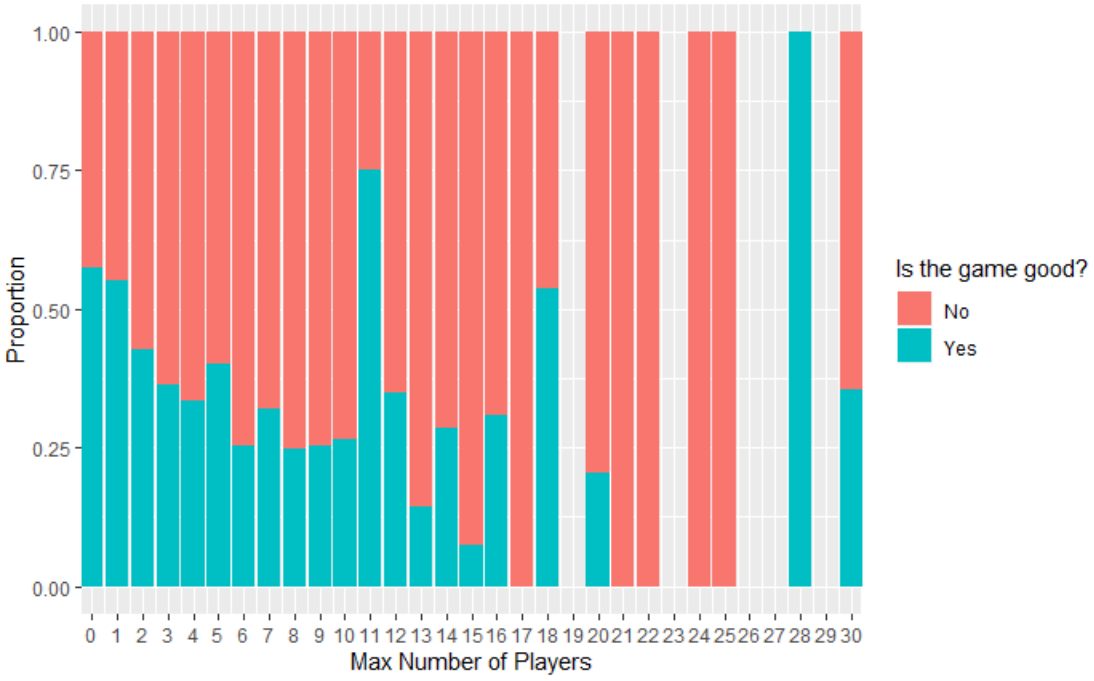


2.1.6. Are games with a lower max player count more likely to be "good"? How does a max 2 player game compare to a max 6 player game?

While player range didn't show much of a difference, there does seem to be a difference for max player count, with the 1st-3rd quartiles for not good games all being at/above the median for good games. There are a lot more outliers for not good games with max number of players, however.

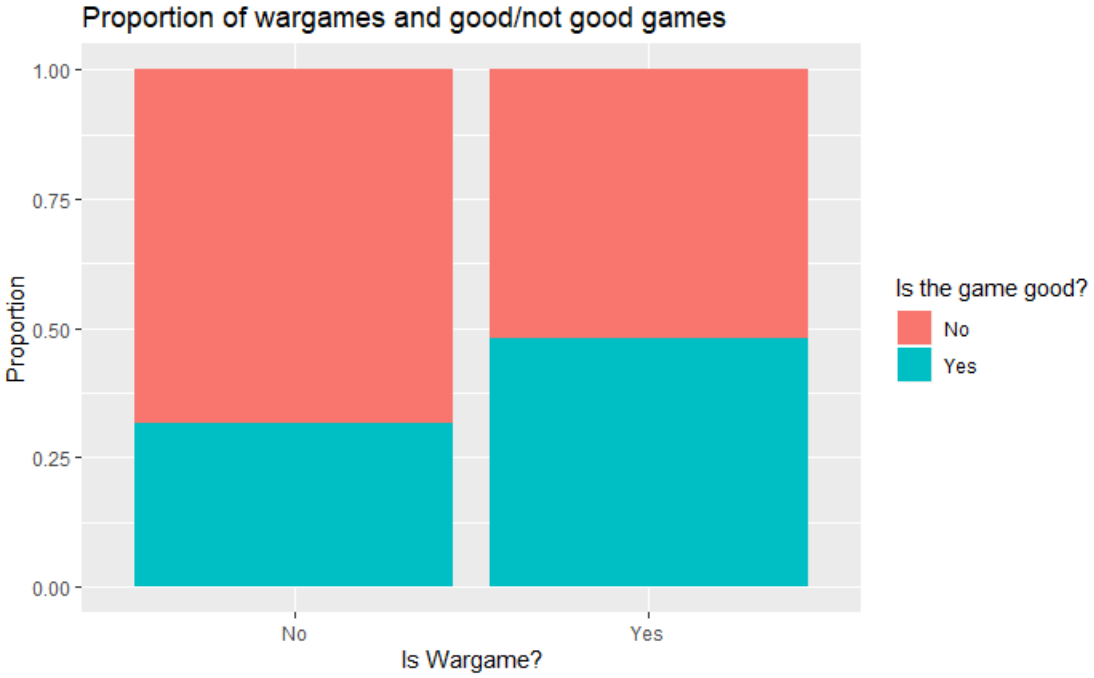


The proportional box plot shows a similar story, with games with 1 or 2 players having higher proportions of good games than 3-10 players. I do notice that there are a large number of games that have a max player count of "0" which seems odd. Its possible this is a similar scenario for weight where it should actually be NA. Also, some games with a very high number of max players have high proportions of "good" games, but some of this could be do to the fact that there aren't many games with high numbers of max players.



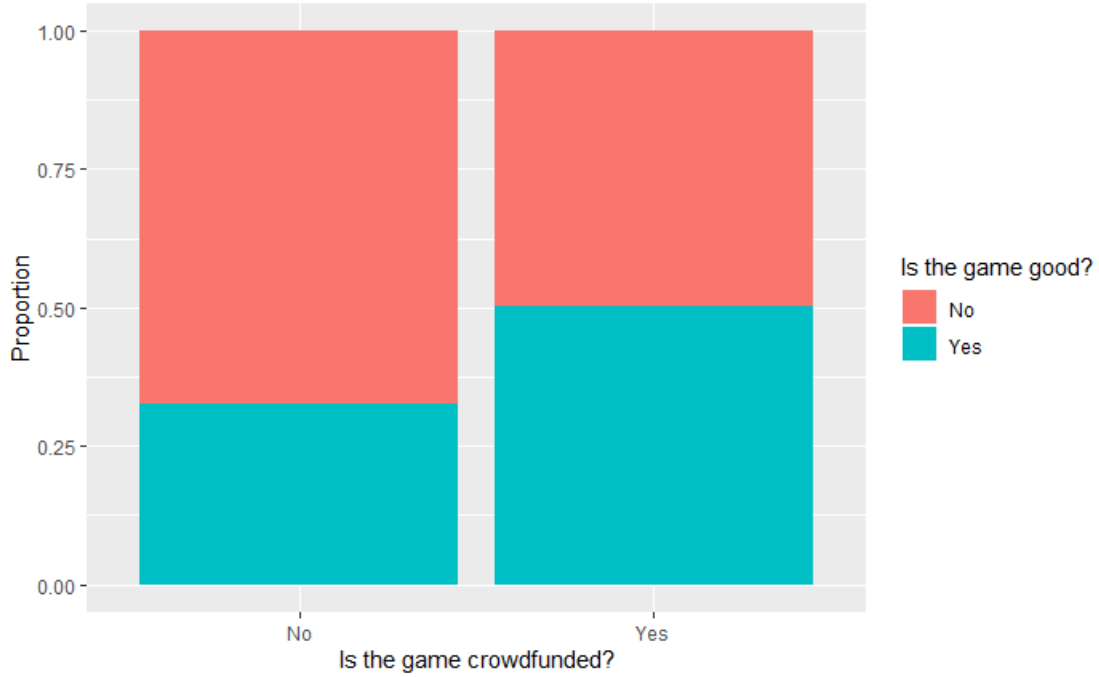
2.1.7. Are "Wargames" more likely to be classified as "good"?

There does seem to be some difference in the proportion of good/not good games for wargames. Almost 50% of wargames are classified as good, while about 30-35% of games that are not wargames are classified as good.



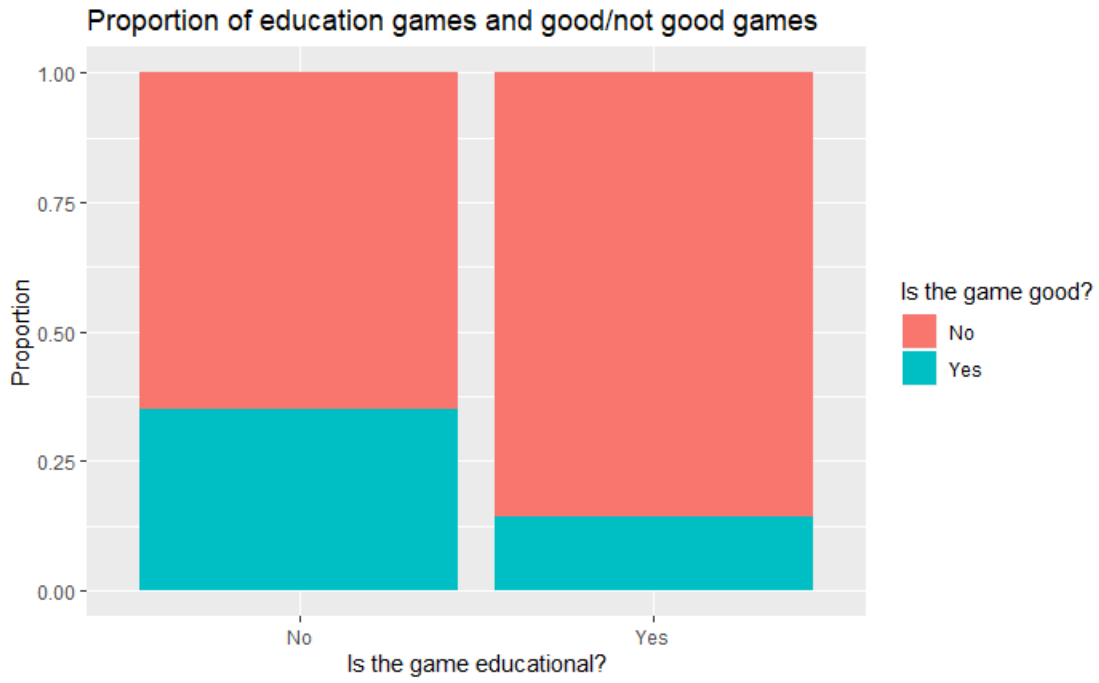
2.1.8. Are “Crowdfunded” games more likely to be classified as “good”?

Like wargames, crowdfunded games seem more likely to be good, with about 50% of crowdfunded games being good and only about 30-35% of non-crowdfunded games classified as good.



2.1.9. Are educational games less likely to be classified as “good”?

Educational games seem less likely to be good based on the proportions. About 37% of games that are not education are classified as good, but only about 15% of games that are educational are classified as good.



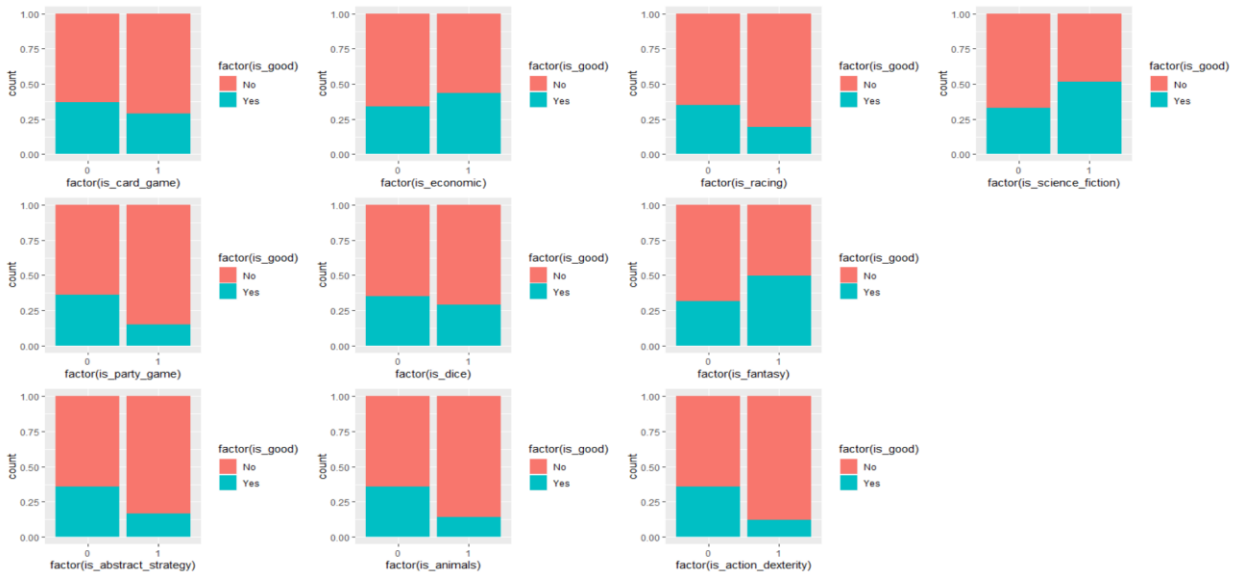
2.1.10. Are press your luck games less likely to be classified as "good"?

While there is a difference in the proportion of press your luck games for good/not good games, it is not as distinct as some of the other categories. About 30-35% of non-press your luck games are classified as good, while about 25-30% of games classified as press your luck are good. This probably isn't as good of as predictor as other categories.

2.2. Other Questions/Exploration

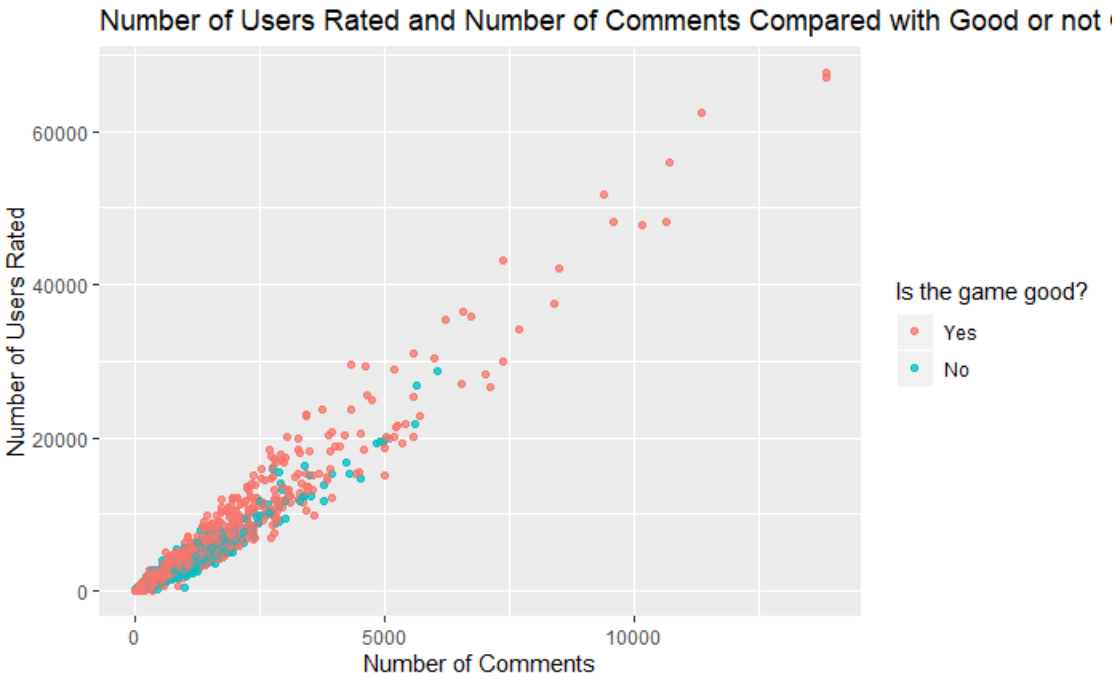
2.2.1. Are there any other "family" categories that are more likely to be "good" than "not good"?

All the categories have a difference in proportions, but the party game, action dexterity and animal categories all have about at 25% difference in the proportion, so they seem to be the most distinctive.

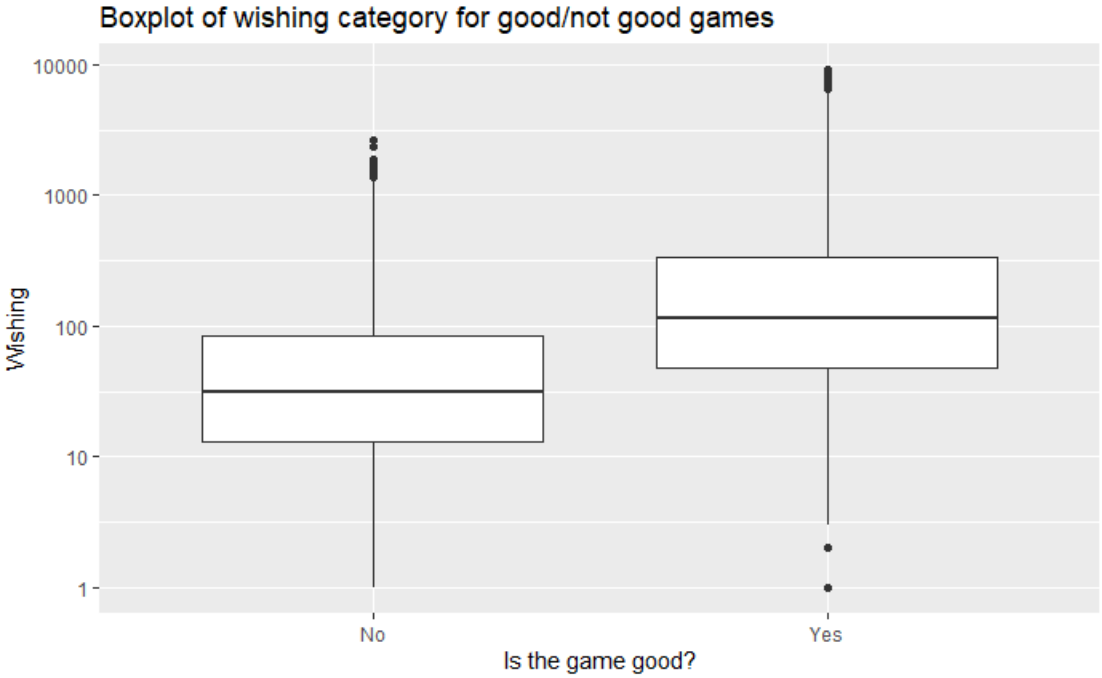


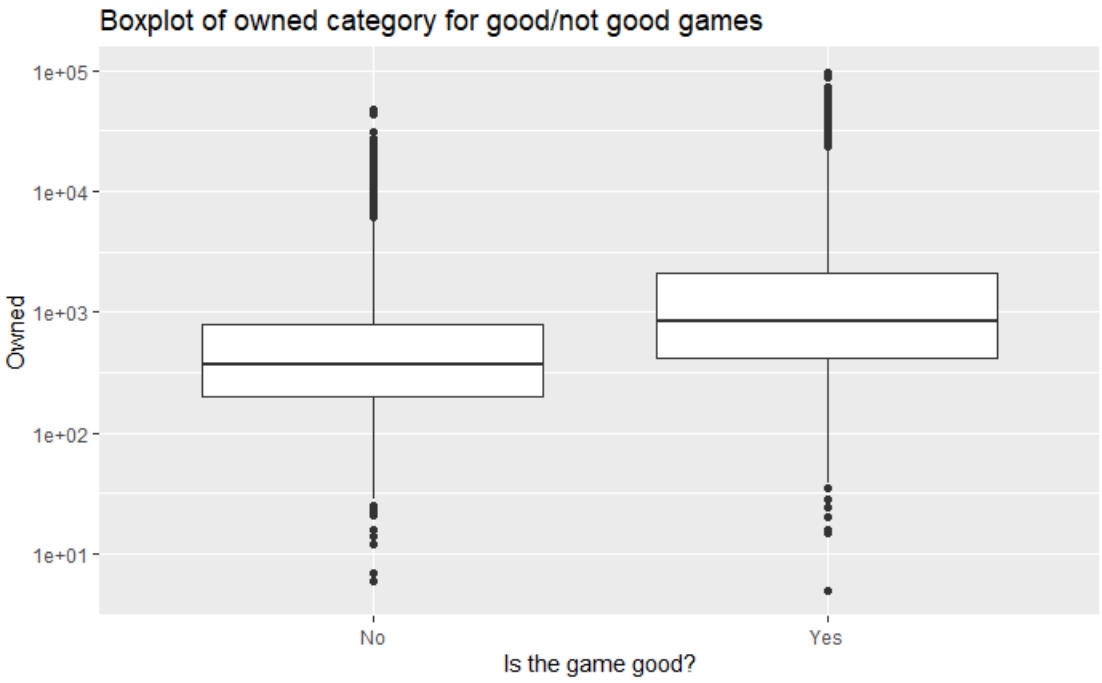
2.2.2. Are games with a high amount of activity (high number of ratings, comments, wanting, wishing, owned, trading, etc.) more likely to be "good"?

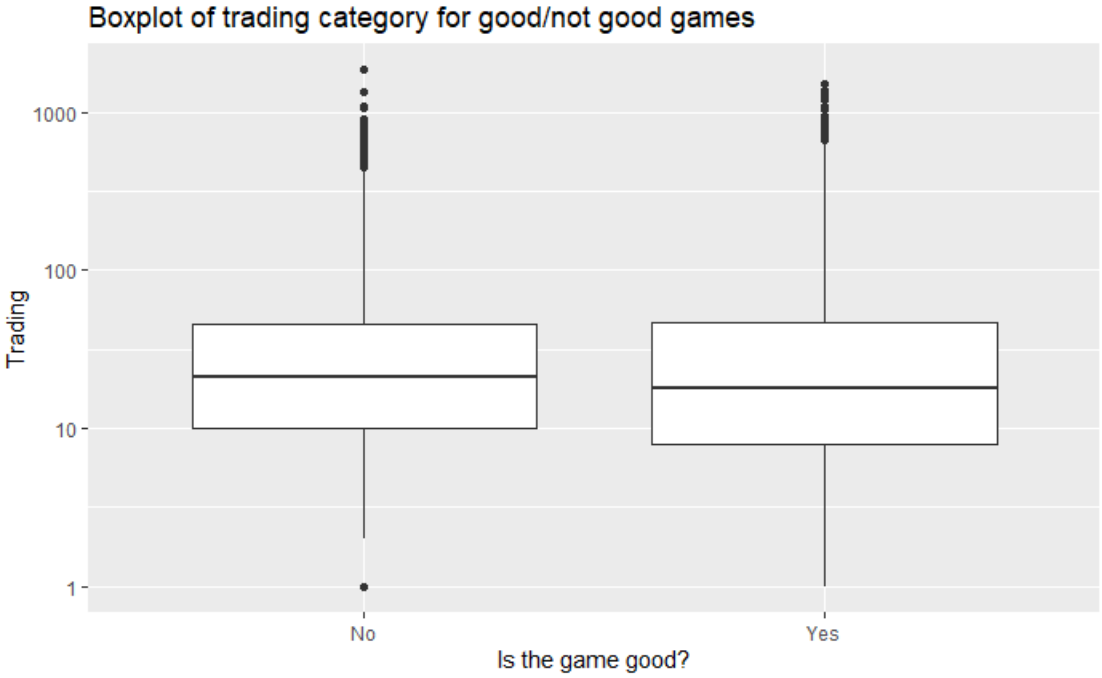
First, I wanted to see the relationship between number of user ratings and number of comments. I expected that the number of user ratings and number of comments are related, since usually people leave comments when they rate the game. I also color coded the points based on if the game is good or not. There is a clear linear relationship between number of users rated and number of comments. Additionally, games with more comments seem to have a higher proportion of good games (after a certain point of about 6000 comments, all the games were classified as good). Additionally, games with higher number of users rated seem more likely to be good.

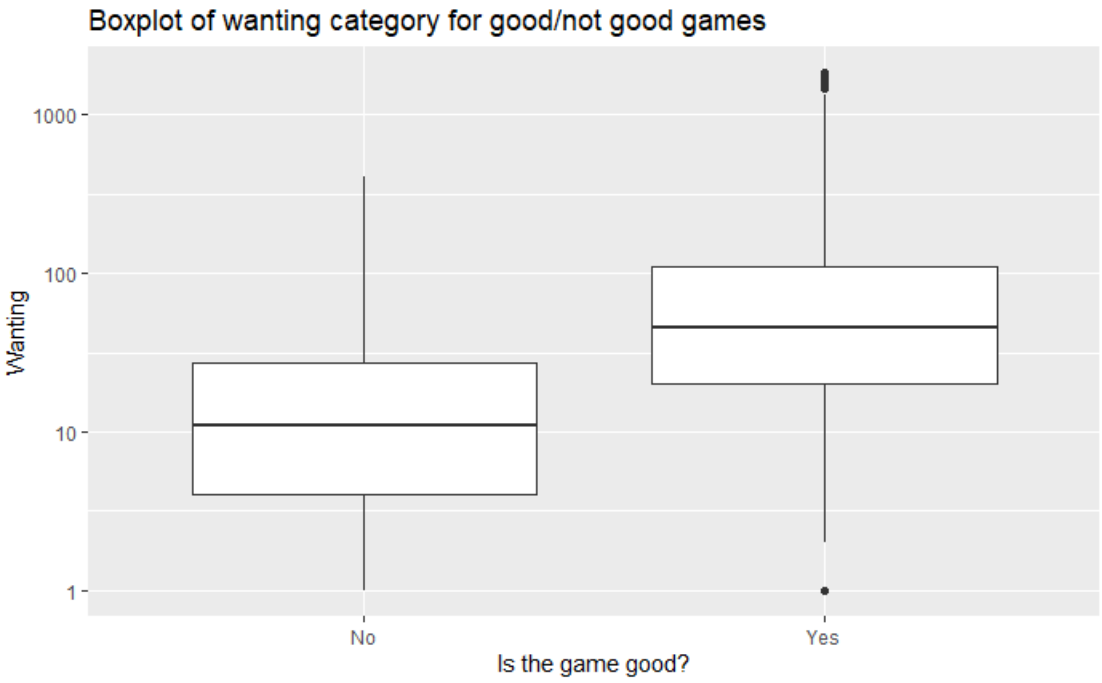


Box plots of some of the other "activity" categories show that good games have a higher median of people wanting, owning, and wishing for the games. However, trading doesn't seem to be as related to if the game is good or not good as much as the other categories.



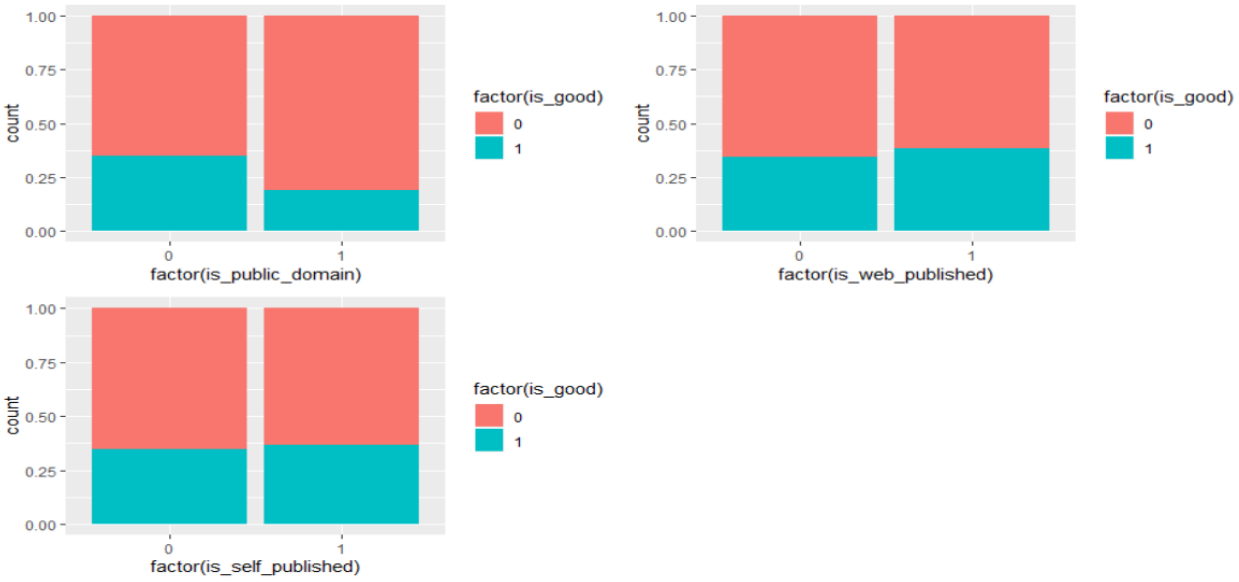

Boxplot of owned category for good/not good games
1e+05
1e+04
1e+03
1e+02
1e+01
Owned
No
Yes
Is the game good?


Boxplot of trading category for good/not good games
1000
100
10
1
Trading
No
Yes
Is the game good?



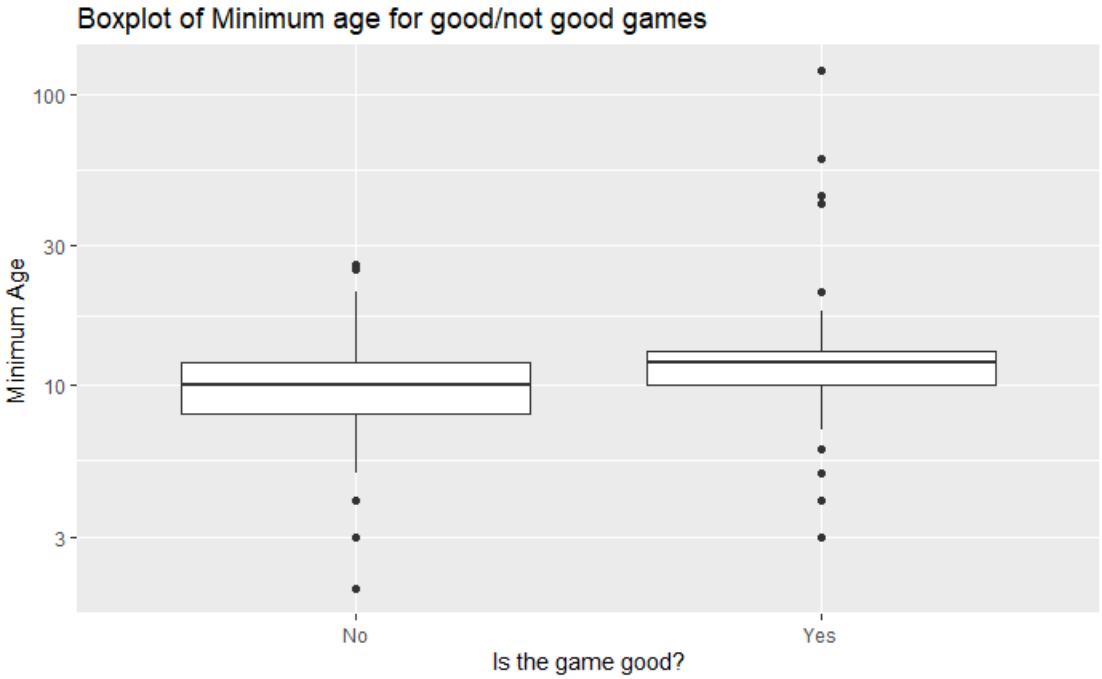
2.2.3. Are games that are self or web published or public domain less likely to be "good"?

There doesn't seem to be much of a difference for good/not good games in the web published/self-published categories, but the public domain category has about a 20% difference between games in the public domain and not in the public domain. It seems games in the public domain are less likely to be good.

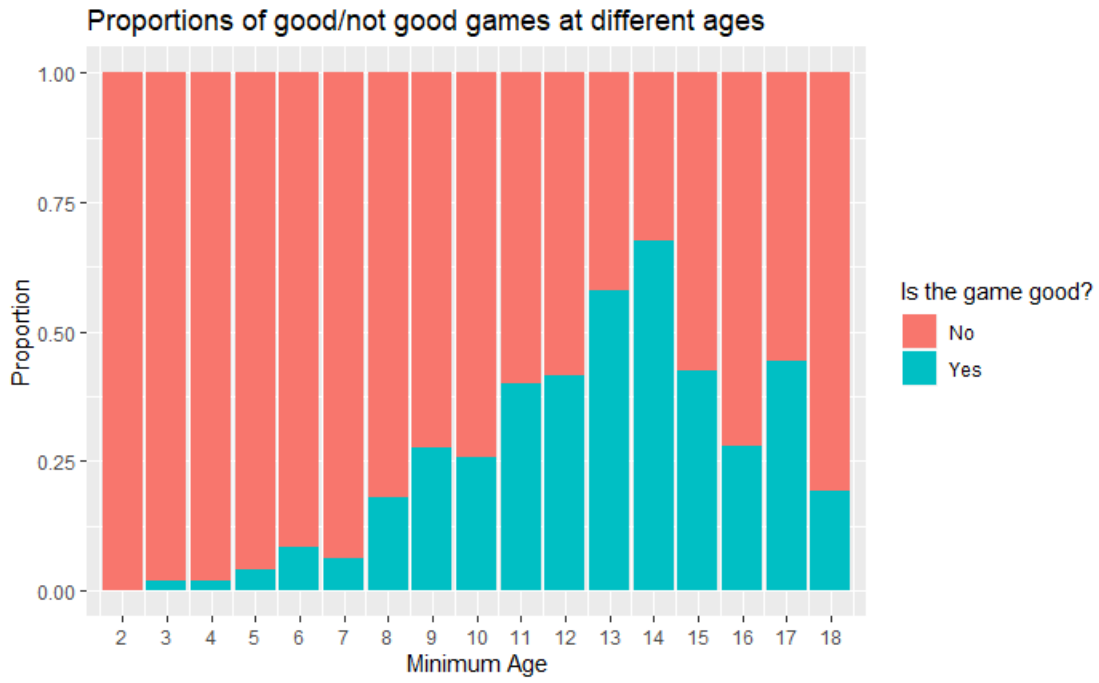


2.2.4. Are games with a lower minimum age less likely to be good?

The boxplots of the good/not good games look similar; however, good games have a higher median age.

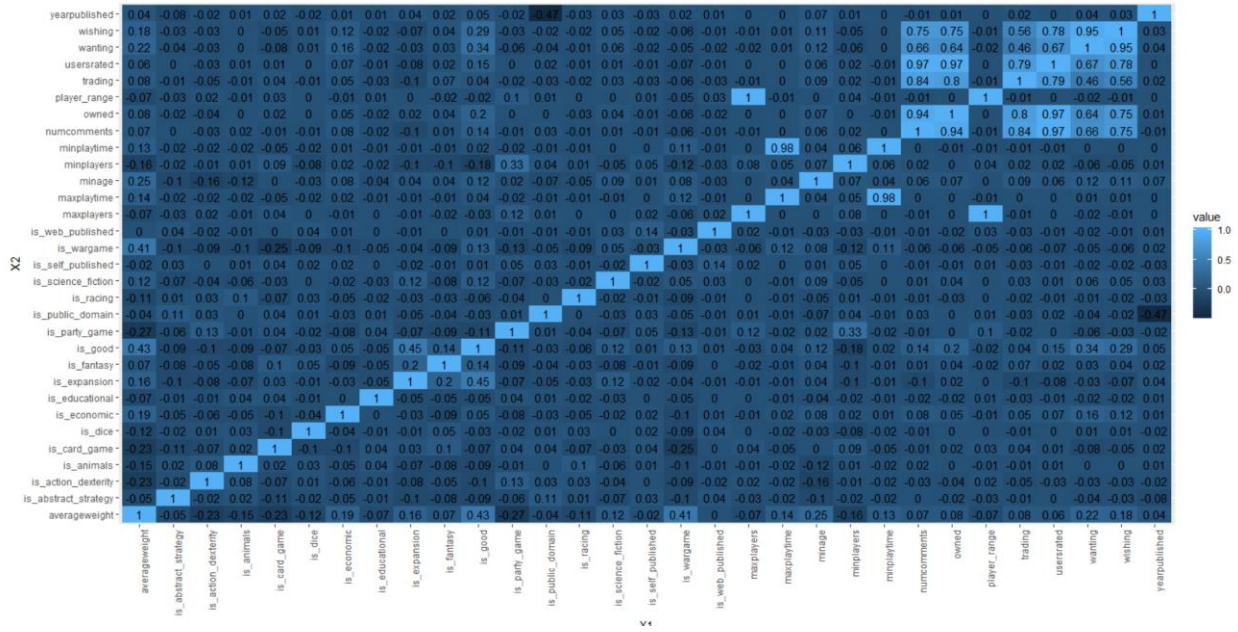


Likewise, there appears to be an increasing trend in the proportion of good games up to about age 14. After this, there is a decrease in the proportion of good games. Perhaps this suggests that people like games that are not so simple that they appeal to young children, but easy enough for teenagers to understand. I will be interested to see if age and weight are correlated in a future section.



2.2.5. Are there any categories with a high correlation to another category?

In a previous class, I learned that it is good to simplify your models and one way to do that is to remove predictor variables that are highly correlated. To do this, I looked at the R^2 values of most variables against each other. From this plot, I found that maxplaytime is highly correlated with minplaytime, and therefore can be removed. Number of comments is highly correlated with number of ratings, so it can also be removed. Finally, wishing is highly correlated with wanting, so it could be removed. I am also interested in if age and weight are correlated as well as if weight has any categories it is correlated heavily with (to potentially deal with the NA values in weight). Interestingly, there isn't a strong correlation between minimum age and weight as I was expecting. There also aren't many variables strongly correlated with average weight.

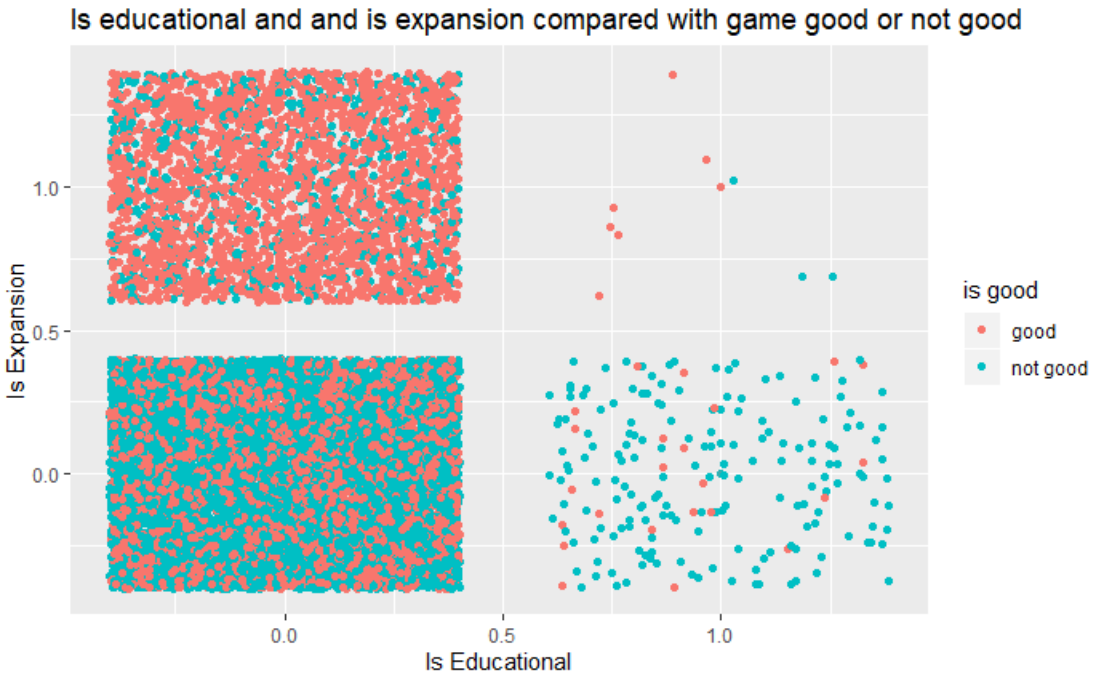


2.2.5. Any further trends when looking at two good predictors?

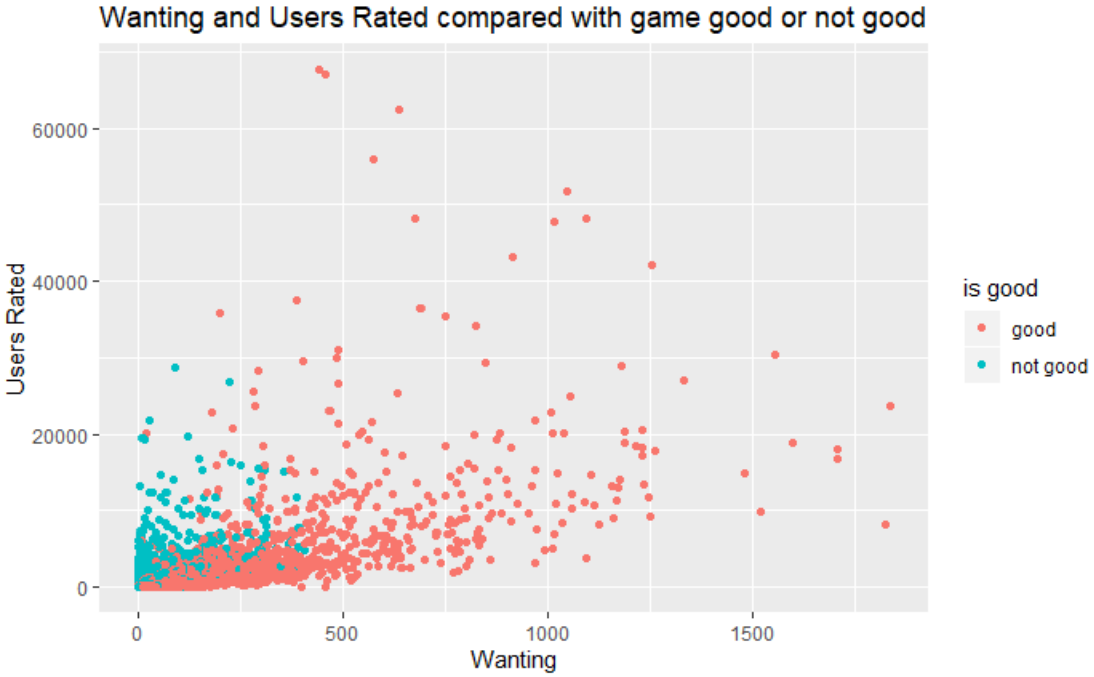
I have identified several strong predictors and I want to see if any stronger patterns emerge when I look at two predictors at the same time. I looked at many different comparisons between predictors but below are a few of the ones that seemed of note.

I thought the expansion and min age comparison was interesting because while clearly for games that aren't an expansion there are specific bands where age really seems to make a difference, for games that are an expansion, games are generally considered good regardless.

Expansion and educational categories were another interesting comparison since educational games tend to be not good and expansions tend to be good. I would expect games that are an expansion and not educational to be good. Interestingly, most educational games that are an expansion are still considered good. Note that some noise has been added to make it possible to see the points in the graph below.



Wanting and users rated showed that games with both high number of users rated and high amount of wanting are more likely to be good.



Finally, expansion and average weight showed most games that are an expansion and are higher weight are good. There was still the same trend of higher weight having more good games even for games that were not an expansion, however.

3. More Cleaning Based on Exploration

Based on the exploration, I think I can narrow down the variables I have to ones that based on the exploration seem like they would be good predictors. I can also remove the ones identified in the previous section as highly correlated to another variable. I will have to decide what to do with the NA weight values. One potential option could be to substitute in the average weight value, but I should consider other options as well. In particular I think weight will be one of the strongest factors based on my exploration. Additionally, factors like year published, if a game is an expansion or not, number of user ratings, minimum age, if the game is educational and wanting seem like the next strongest predictors. From there, is party game, is animal game, is action/dexterity game, owned, is wargame, max number of players, is public domain, and is crowdfunded seem like good candidates as well.

4. Scripts

Note that the exploration script is very messy and was more of a scratch pad for the paper above, but it is included for completeness.

Cleaning.R


Exploration.Rmd

5. Sources

The following sources were used for cleaning/exploration. The textbook, Data Mining for Business Analytics was most heavily used, but the dataPreparation article was used for cleaning and the multiple graphs on one page code was used to generate multiple plots for the family categories. The bivariate graphs article was used to try and think of different plots to create.

Anon. 2020.(February 2020). Retrieved October 28, 2020 from https://cran.r-project.org/web/packages/dataPreparation/vignettes/train_test_prep.html

Anon. Multiple Graphs on One Page (ggplot2). Retrieved October 28, 2020 from http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/

Galit Shmueli. 2017. *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*, New York, NY: John Wiley & Sons.

Rob Kabacoff. 2018. Bivariate Graphs. (2018). Retrieved October 28, 2020 from https://rkabacoff.github.io/datavis/Bivariate.html