Problem Set 1 Note

In part one, I scrapped data on the page number, title, geek rating, average rating, and number of votes from the first 100 pages of board game geek. This scrapping gave me 10,000 entries with 5 pieces of data: page number, title, geek rating, average rating, and number of voters.

For the machine learning section of the project, I decided to do KNN classification to use the geek rating, average rating, and number of votes to predict which page a board game with certain characteristics would belong to. I wanted to see how well my model would predict how successful a board game would be. The first thing I did was create scatterplots to see how each dependent variable predicts the page number.

It was clear from these scatterplots that average ratings were used to sort the data. However, I decided to use all three of the characteristics to classify the data to increase the validity of the results.

I then used the GridSearchCV to identify the correct number of neighbors for the k-NN classifier. Using this program, I came up with the optimal number of neighbors as 1. I then use the K-fold cross validation to check the accuracy of my classification program. I separated my data into 5 folds so that I would have 5 test sets of data. Running the K-fold cross validation gives the following accuracy scores.

[0.0925 0.0825 0.082 0.097 0.0845]

cv\_scores mean:0.0877

I found that this data has a high accuracy for classifying the data into page numbers.