Introduction

After cleaning the data and doing an initial visual exploration of the data, I am ready to showcase any findings that will help with the further analysis of the data. I will go through the variables I am using, their significance and if any of the statistical analysis showed further insights.

Data Insights

Starting with the variable “authors”, some interesting insights were discovered. Firstly, of the nearly 10,000 books in the dataset there are only 4355 authors. This means that each author roughly wrote two books in this dataset. However, upon further inspection there is actually a great difference between the amount of books per author. Fig 1. There are two explanations for this. One, some authors are very prolific writers such as Stephen King whereas others write a series or several series of books (A typical series is any sting of books longer than three).

Next is “Books Count” which shows the amount of number of times a book has been tagged or labelled by individual users.

Book “Year” is also included but doesn’t show any statistical significance. Though most books on the list are after 1750, there isn’t any correlation between book ratings and book year. Fig 2.

The “Ratings Count” measures the number of ratings per book. There doesn’t seem to be much of a correlation between ratings count and book count. In fact, it seems that the books that have the most ratings are outliers. Most books have less than 150,000 ratings. Fig 3.

“Title” refers to the title of each book. This is useful with identifying individual books or series of books. Some books have a very high amount of ratings, much more than the average. Fig 4.

Analysis

I tested for correlations between many of the numerical data in order to see if further analysis would be required. I first tested if there was any correlation between book counts and ratings counts. I used the Spearman correlation coefficient (Spearman) because both categories contain outliers. The null hypothesis is there isn’t any correlation between the two groups. After completing the test and calculating a p-value of 0.0, it is within reason to say that the null hypothesis was failed to be rejected. There isn’t a correlation between the two groups and further analysis isn’t required. Stat 1

Next I tested to see if there was a correlation between average rating and ratings count per book. My null hypothesis is there is not a correlation between the two variables. I chose the Spearman test again since I am testing correlation between an ordinal and interval variables, both of which contain outliers. My findings show there is a correlation between the two, although not the strongest. The null hypothesis is rejected. Stat 2

Based on the descriptive statistics and graphs, the average rating is very skewed to the right, with the vast majority of books rated above 3.5. This leads to speculation that there is a bias in reviews. That is to say, most people will review a book only if they really like it or really hate it. Since this dataset is from Goodreads Kaggle competition, it is more than likely that these books were already the most popular books on the site and as such their average ratings will be higher than the expected normal distribution.

Without having the statistical tools at the moment, I concur there is a text correlation that can be used to further analyze the data. In the next section of the project, machine learning algorithms will be applied to further test this theory.

Conclusions

Though only a few correlations were found within the statistical and graphical analysis of the data, the correlations that do appear will be instrumental with the machine learning algorithm in order to solve the original problem of producing a list of books based on the selection of a book. These correlations along with the text data will still aid in finding additional correlations. After this exploratory data analysis, I will then split the data into training and test data.