**Problem:**

The problem that I’d like to solve is if and how we can predict book ratings from the book description. The book description (along with the cover) can literally make or break a book’s sales. It is a form of marketing. The publisher has this one chance to impress the audience. It is giving the public its best pitch on why it should read this book. While sales for books are difficult to find, Goodreads is a website that provides data on books, with one of the core features being the book’s average rating. The public votes for their rating of the book, and I intend to develop a model that can show which words are associated with the highest ratings.

**Dataset:**

The Goodreads data set consists of roughly 30,000 observations. The target variable is the book’s average movie rating, which is a continuous numeric value and ranges between 0.0 and 5.0. The primary independent variable is the book description, which is a paragraph or more long. Since it is a column consisting of text, it will need to be pre-processed.

I attempted to clean the data set before I settled in on which variables were to be included within the model. Since there were only 19 columns total, I figured that this would not be a strung-out task. The text data--coming from the columns corresponding to the book’s description and the book’s title--were pre-processed using code sourced from my mentor DJ’s Github repository. He has code that defines functions meant to remove stopwords, lemmatize words, lower case words, expand contractions, and remove special characters. Using these functions, I pre-processed the text data and created new columns to correspond with each.

There were other columns consisting of text data that were not cleaned, such as hometown, author name, publisher, and book genre. While the researcher would have liked to have included these in the model, it would have been much too cumbersome to try to clean. For instance, the hometown variable contained many missing values, and often the hometown label was inconsistent. Sometimes naming a city, state, province, country, etc. There are too many author names to add much value. The same goes for the publisher. The book genre variable was an absolute mess. It seemed like it was input entirely by the public. Many words were hyphenated with little value added (e.g., “currently-reading”). A single cell for a book could contain dozens of words within the “genre” variable, so it was decided to not use it.

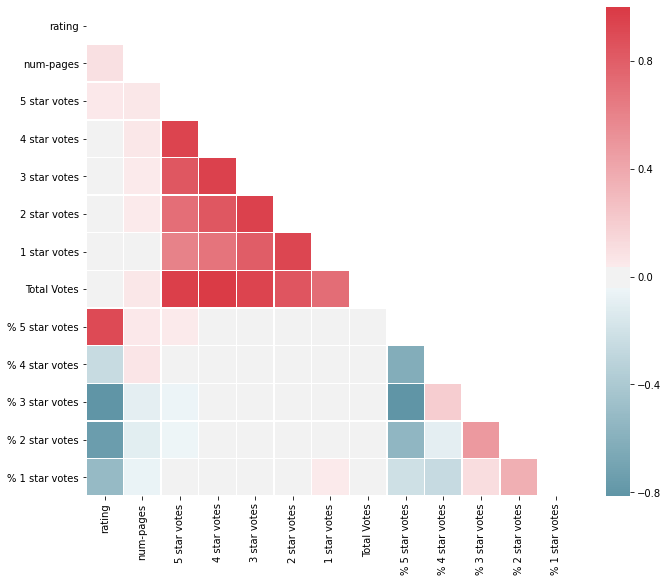
Other features of the data set that will most likely go into the model are author gender, a binary variable indicating whether the book was turned into a movie or not, the total number of book pages, total votes, and rating distribution. The gender variable consisted of many genders beyond the usual “male”/”female” labels, such as “undecided,” “fluid”, etc. These were not many, but if “gender” is included in the model, the other labels will simply have to be discarded. The rating distribution variable consisted of a book’s ratings for every vote. In other words, it had the total 5-star votes, total 4-star votes, total 3-star votes, total 2-star votes, and total 1-star votes. These were eventually separated into five separate columns. Another five separate columns were made where I took the percentages of these X-star votes. This was calculated by dividing the total X-star votes with the total votes from the “Total Votes” variable.

I had to discard any observations that had text that was missing. The reason for this is because the pre-processing text functions will not work on missing data. Fortunately, there were only a few thousand observations that had to be discarded out of the 30,000 or so.

**Exploratory Data Analysis:**

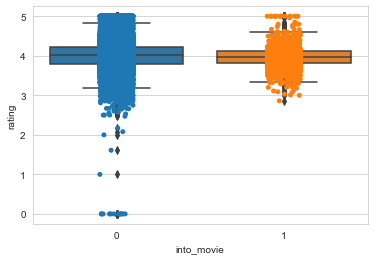
When exploring the data, a few trends stood out to me involving the three different types of variables in my data set: numerical, categorical, and text.

To explore the trends of the numerical variables (that is, the variables composed of continuous or integer numbers) I made a correlational heatmap using the Seaborn package. This will be a visual aid in helping me to decide which variables will go into the model. Ideally, I want to choose predictor variables that are correlated with the target variable yet are not correlated with each other:



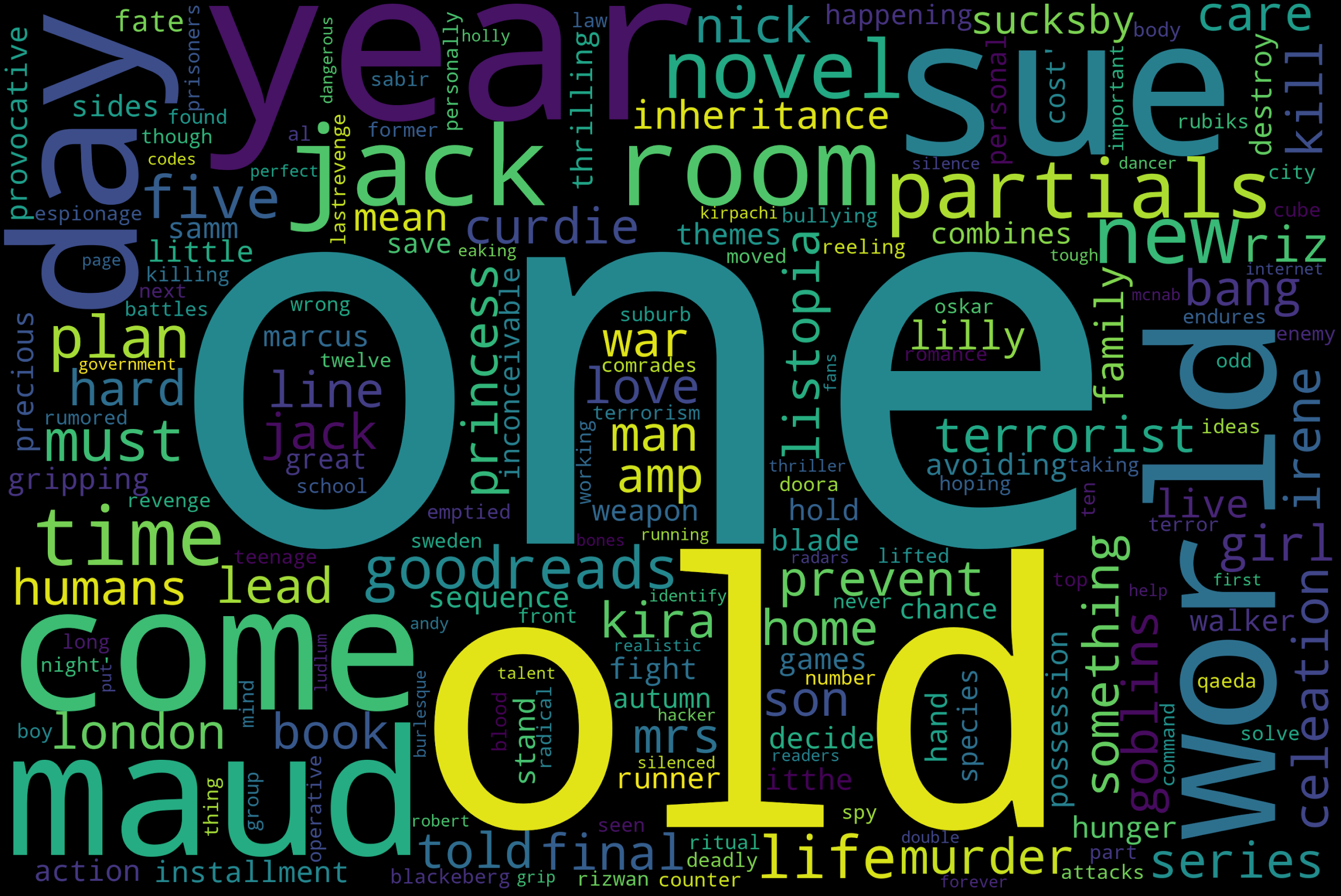
“Rating” is to refrain our target variable. From the heat map, the variables “% 5-star votes,” “% 2-star votes”, and “% 3-star votes” are highly correlated with the target variable. The former positively, and the latter two negatively. None of the others are particularly correlated. Interestingly, none of the X-star votes were correlated with their corresponding % X-star votes. Thus, going by this heat map and remembering our “ideal” feature selection, I’d choose the following variables for model inclusion: “% 5-star votes,” “num-pages,” “Total Votes,” and “% 1-star votes.”

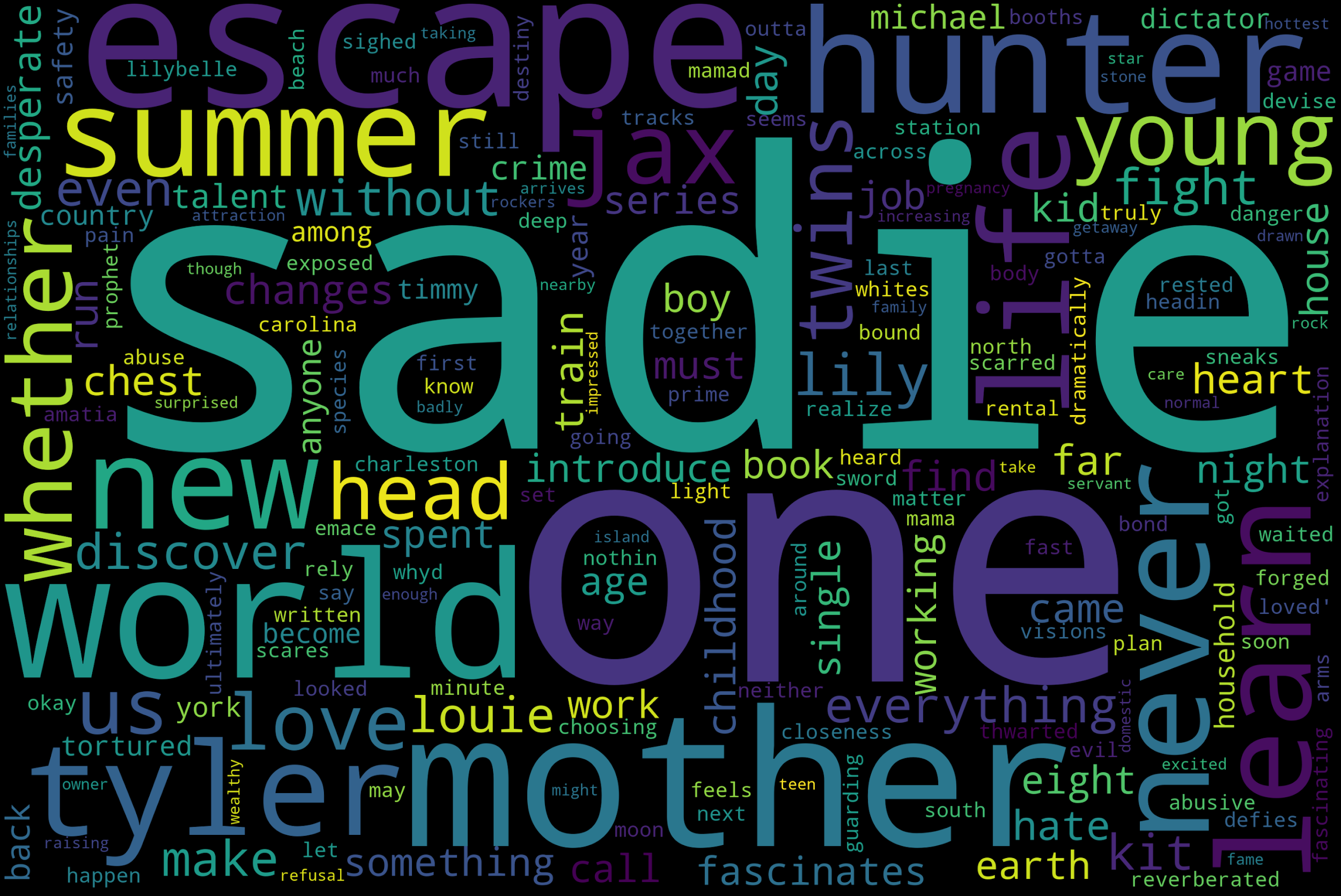
Returning to the categorical variables, there appears to be a pattern between the variable “made-into-movie” and “rating,” as shown by the following box plots:



Books that *were not* made into a movie had more extreme outliers than books that *were* made into a movie. The former had outliers where the average rating was awful, falling below 2.0. Interestingly though, the box plots show the books that weren’t made into movies had on average higher ratings.

For the text data, I made two Word Cloud charts. The first shows the difference in the frequency of words with books that averaged ratings of 4 stars or more, while the second chart shows the frequency of words with books that averaged ratings of less than 4 stars.





Amazingly, there is definitely SOME difference. The word “one” is more prevalent in the 4 stars or more chart than the other chart. Somehow the name “Sadie” is in a lot of books with a rating of less than 4. “Sue” is prevalent in books with a rating of 4 or more.

To summarize my findings, there are definitely some patterns within the data set that indicate that the model could be successful in predicting book ratings. Variables such as “% 5-star votes,” “num-pages,” “Total Votes,” and “% 1-star votes” generally fit the ideal features. The “into movie” variable shows an unusual difference between the two book groups (whether the book was or was not made into a movie) in terms of ratings. And the word chart shows that some words are more prevalent than others when the books are split between those with high ratings (4 or more) versus lower ratings (less than 4).