## Business Problem

I would like to see if book covers are related to average ratings for books.

## Background

Goodreads.com is a website that catalogues published books of all formats. It allows users to add books to personalized shelves, give star ratings, and write reviews, among other functions. The data used here has been collected from Goodreads.com and stored in a dataset on Kaggle.com.

## Data Explanation

The original data set contained 100,000 records. To get a data set of a more reasonable scale for the question I want to answer, I narrowed down the genres only to sci-fi and fantasy. This reduced the number of records to 4724 records.

Here are the variables that were included with the original data set:

| **Variable** | **Description** |
|---|---|
| Author | A text value containing the name of the author |
| Bookformat | The format of the book (e-book, mass market, hardback, etc.) |
| Desc | A text value that usually contains the cover blurb (book description found on the back of most physical books), or other summary of the book's content |
| Img | An https link to the image of the cover |
| Isbn | A standard unique number identifying the book |
| Isbn13 | A standard unique number identifying the book that contains 13 characters |
| Link | A link to the Goodreads.com page containing the book |
| Pages | An integer describing the number of pages in a book |
| Rating | A float (0-5) containing the average of all user ratings that existed at the time the data set was collected |
| Reviews | An integer representing how many reviews a book had at the time the data set was collected |
| Title | A text value containing the title of the book |
| Totalratings | An integer reflecting the number of individual ratings a book had at the time the data set was collected |
| Genre | A string of genres that could apply to the book separated by commas |

# Methods

I implemented a few data science techniques to clean and process the data prior to making predictions.

## Data Cleaning

To clean the data, I first reduced the number of records to make for easier processing. I dropped any records that had null values for the ‘img’ column, which decreased the number of records from 100,000 to 96,955. Next, I split out the “genre” variable on the commas into multiple columns with the prefix “genre_”. I then gradually limited down the books to records that contained science fiction and/or fantasy for the genre entries. This reduced the number of records to 10,586.

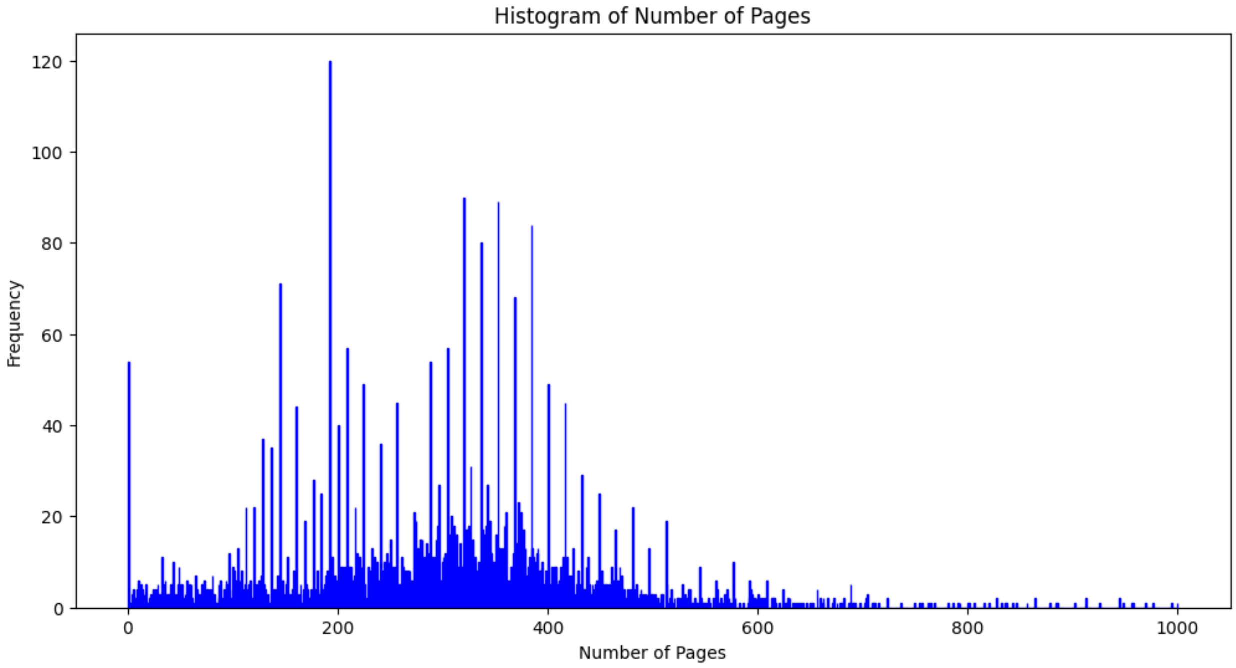
There were some entries that contained book descriptions (in the “desc” column) that were in languages other than English. I created a function to detect the language of in that column and removed all but those that were in English. Then I tokenized the language for that column, in case I wanted to use that column to help predict book ratings. The number of records was reduced to 9,676.

Next, I checked statistics for the number of pages using the describe function in Python. Given that the third quartile was at 350, I decided to remove records that had greater than 1000 pages. I checked the same for reviews and total number of ratings. I removed all books that had fewer than 1000 ratings to increase the reliability of the average rating.

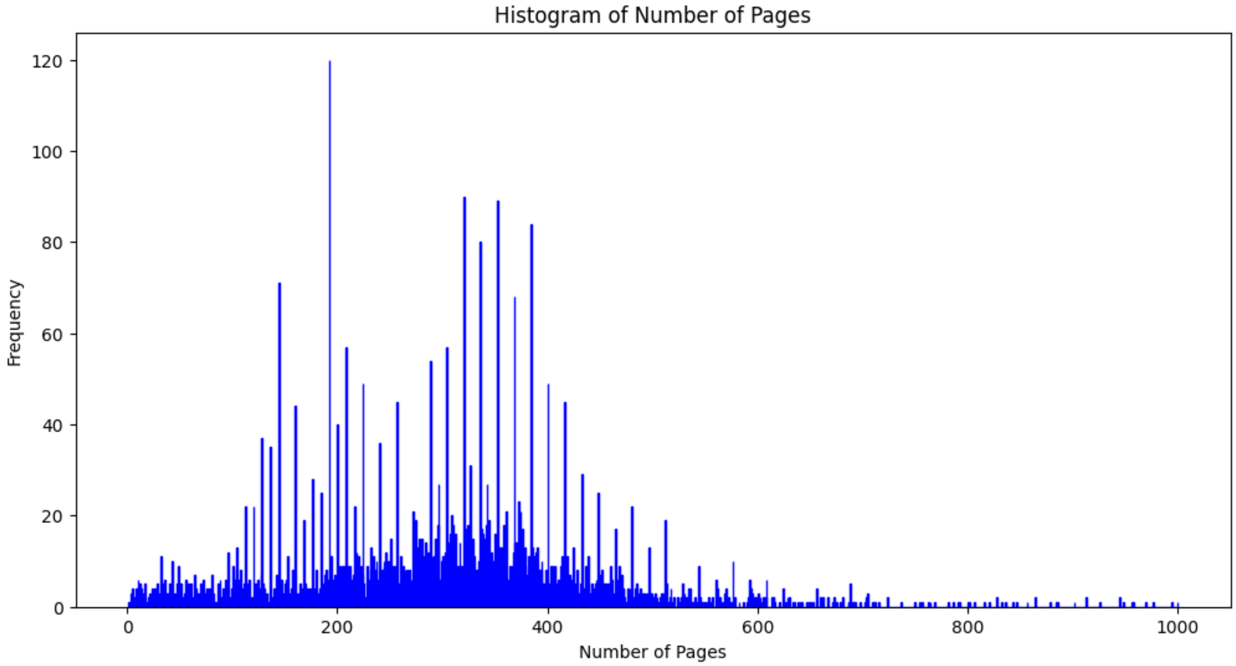
Then, I validated that the “img” column only contained one link for each record. Following that, I created a function to request the image based on the link in the “img” column and download it to my local machine and place in the data frame as jpegs. I also standardized the size and type of the images. Finally, I wrote the data frame to a CSV file to use later.

## Exploratory Data Analysis

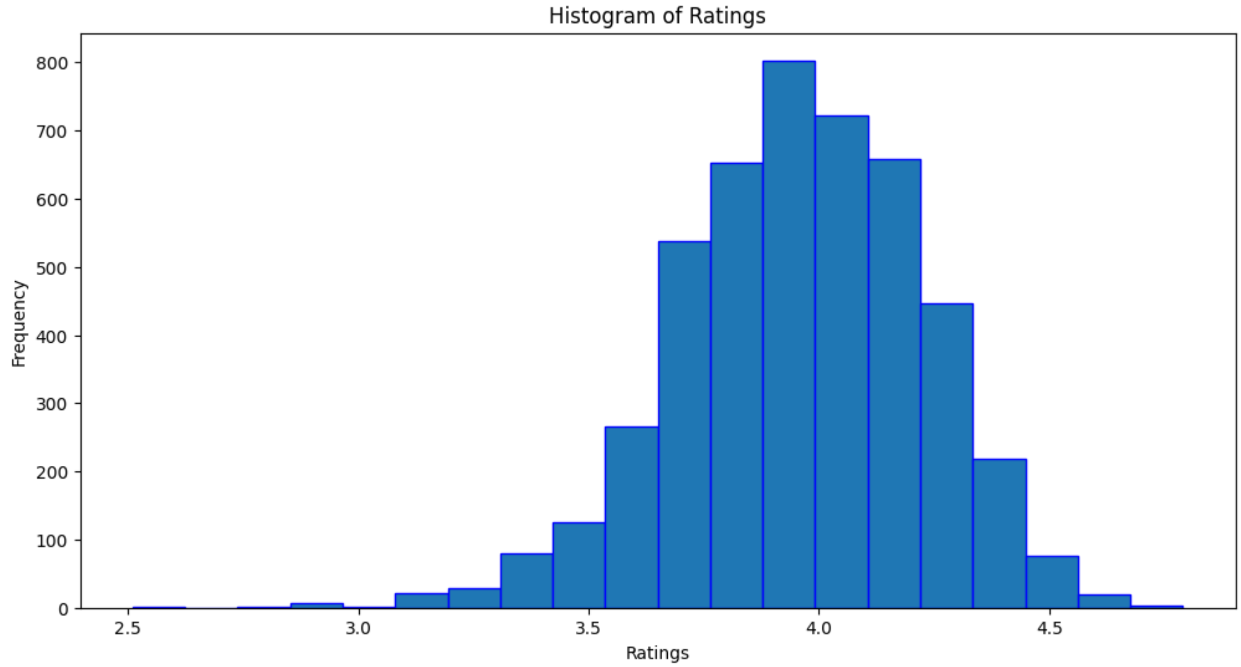
When exploring the data and before modeling, I decided to drop all the genre columns to simplify the data set. I also realized that there were records that contained zero for the number of pages.



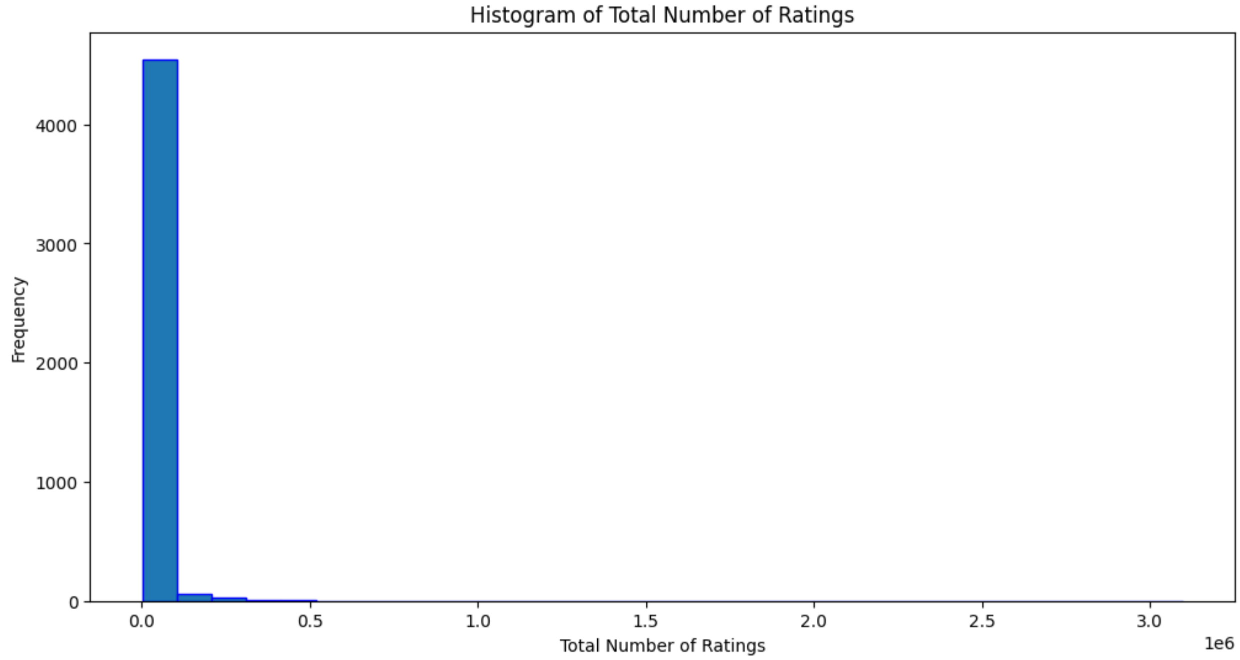
Looking at the chart, I realized that there were way too many books with zero pages, so I dropped the records that had 0 pages. Here is the new visualization:



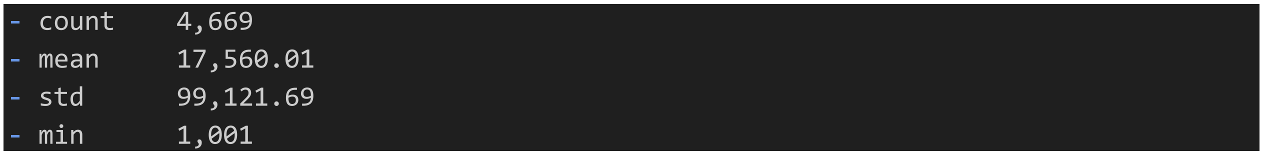
I also checked the distribution of ratings. We can see that most records contained ratings between 3.5 and 4.5 stars.



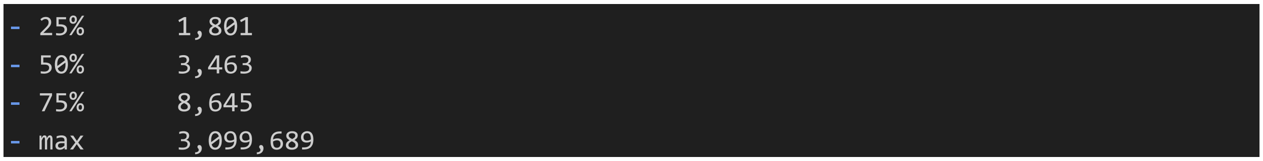
Next, I checked the column containing the total number of ratings for each book.



We can see that some books received an astronomical number of ratings, leading to drastically right-skewed data. I checked the statistics of this attribute with the describe function.

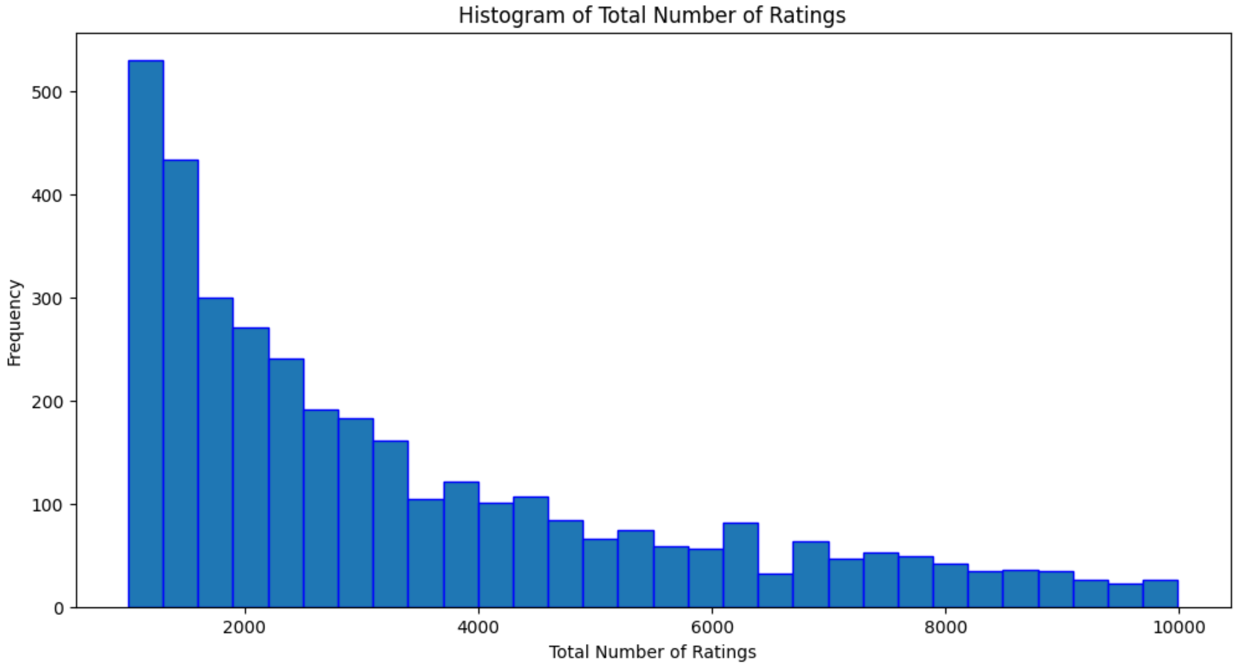


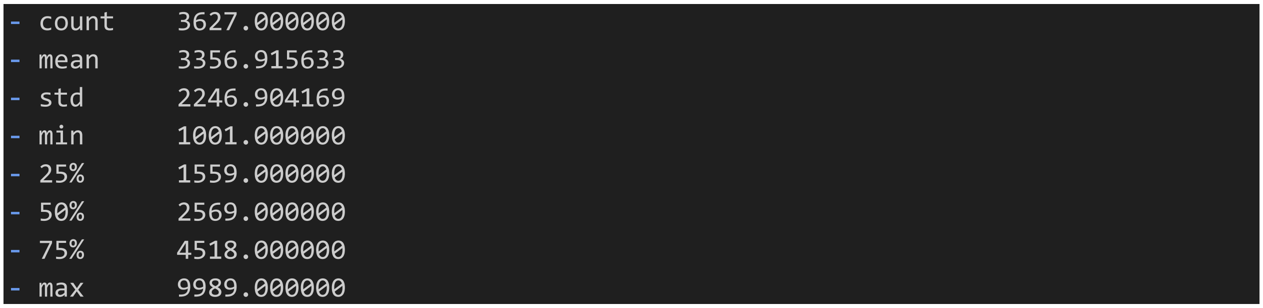
```
- count    4,669
- mean     17,560.01
- std      99,121.69
- min      1,001
```



```
- 25%      1,801
- 50%      3,463
- 75%      8,645
- max      3,099,689
```

The third quartile was at 8645, while the maximum value was over 3,000,000. I decided to limit the number of ratings to be less than 10,000 ratings to get rid of outliers.





```
- count    3627.000000
- mean     3356.915633
- std      2246.904169
- min      1001.000000
- 25%      1559.000000
- 50%      2569.000000
- 75%      4518.000000
- max      9989.000000
```

The new distribution of ratings was much more legible and reasonable.

The total number of records was now reduced to 3627.

## Data Preparation

When exploring the data and before modeling, I decided to drop all the genre columns to simplify the data set. I also removed records with 0 pages based on what I found during exploratory data analysis, and limited the records to those with less than 10,000 reviews. I also dropped the Title, Link, ISBN, ISBN13, IMG, DESC, and Author columns. At some point in the future, I would like to explore the description column more, but for now I am removing it to simplify analysis.

I decided that I wanted to create columns with numerical representations and color histograms for the cover images. I created functions to do both those, flattened the resulting numerical and histogram columns, then dropped the Cover Image column that contained the JPEGs. I also created dummy columns out of the book format column, and changed the column label type to string. Lastly, I wrote the resulting data frame to a new CSV file.

## Analysis

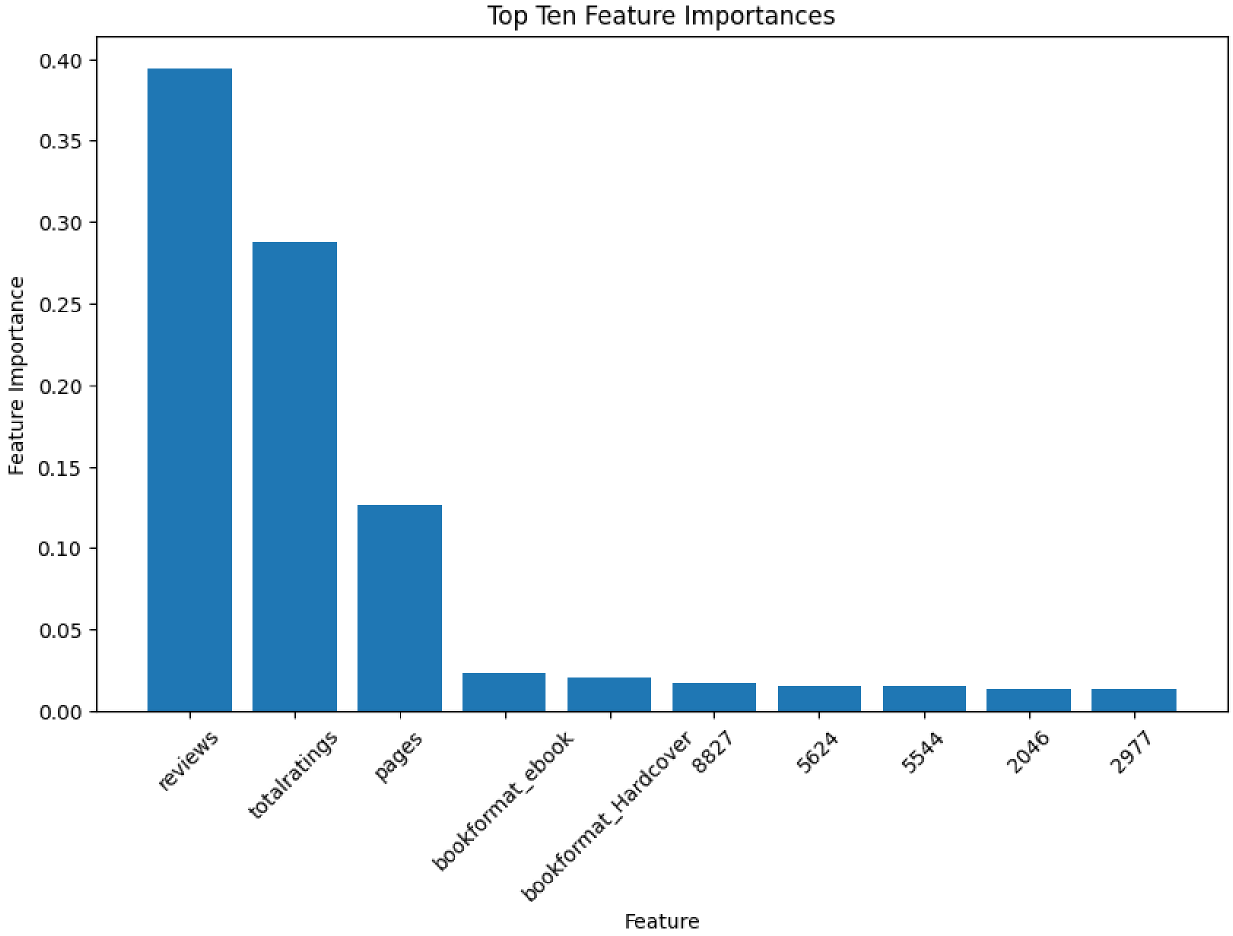
I used three different models on this data set, a Decision Tree Regressor, a Random Forest Regressor, and a Gradient Boosting Regressor. The R2 scores were as follows:

| **Model** | **R2 Score** |
|---|---|
| Decision Tree Classifier | 0.1855482811043503 |
| Random Forest Regressor | 0.22779588918474214 |
| Gradient Boosting Regressor | 0.2354524597871297 |

The Gradient Boosting Regressor was the most accurate based on these scores. We will also see below that the same three features were of the top importance in all three of these models.

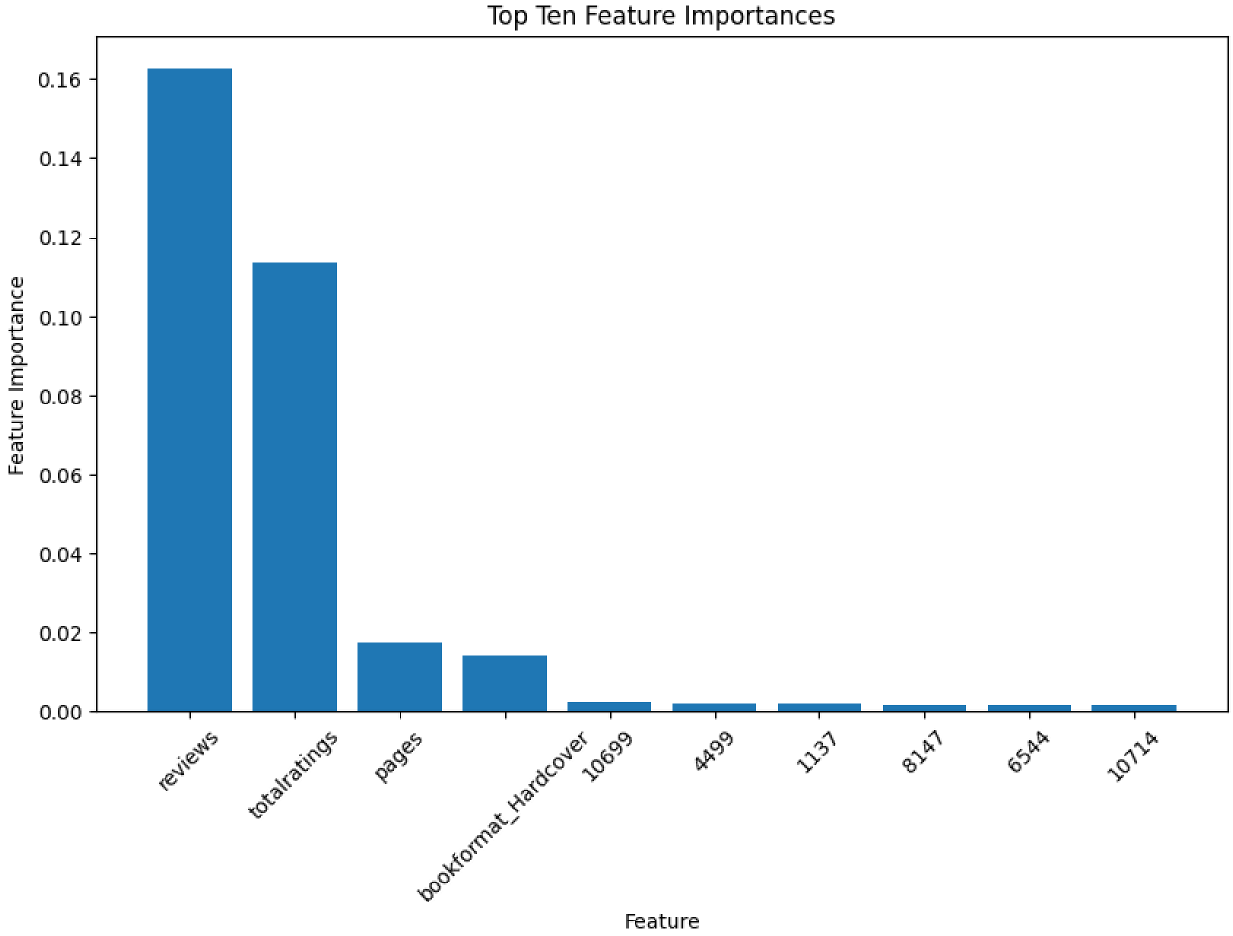
The models will all need tuning to see if the accuracy can be increased, but for now these are the results.

## Decision Tree Regressor Feature Importance



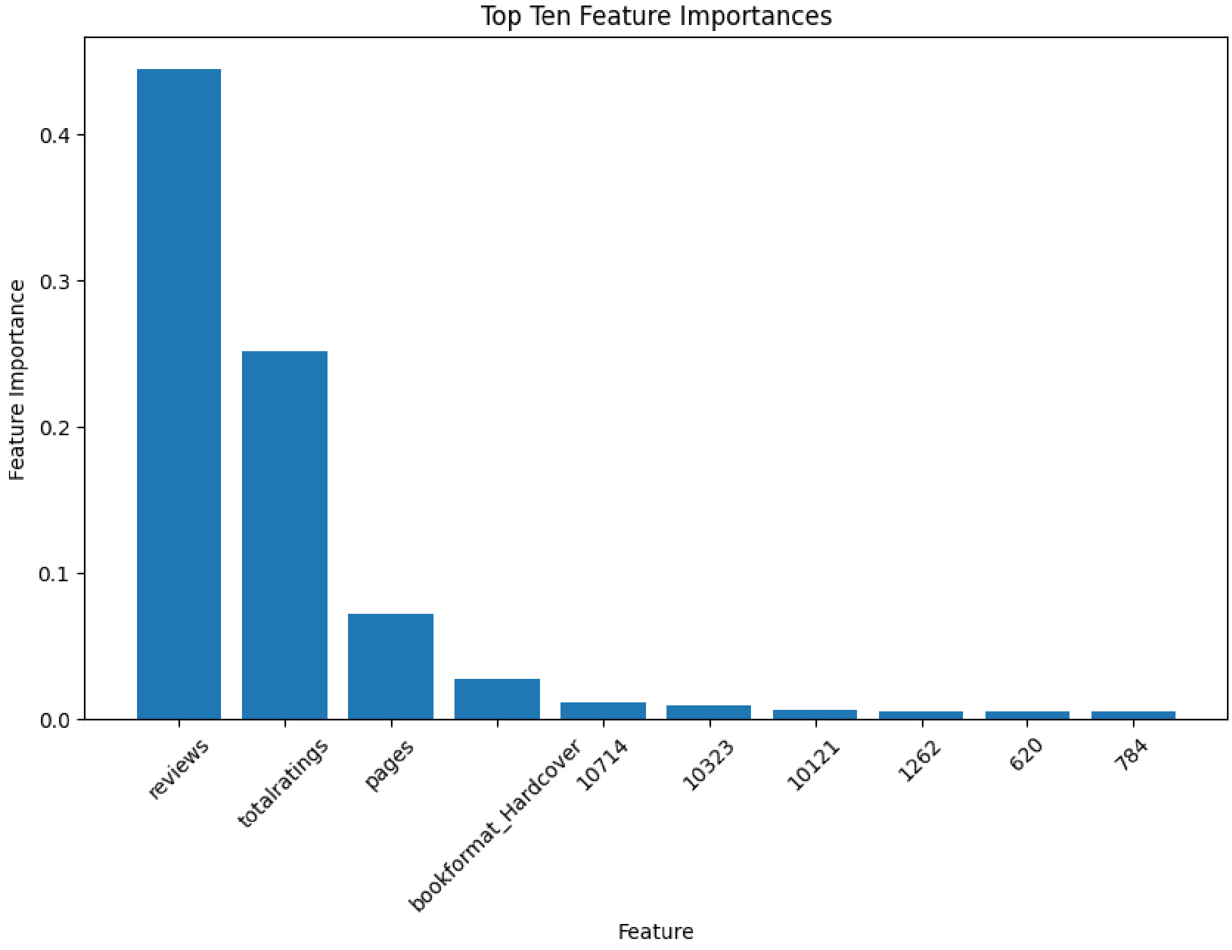
We can see that number of reviews, total ratings, and pages were the most important features in determining average ratings.

## Random Forest Regressor Feature Importance



As with the Decision Tree Regressor, we can see that number of reviews, total ratings, and number of pages were the most influential in determining average rating.

### Gradient Boosting Regressor Feature Importance



As with both the Random Forest Regressor and Decision Tree Regressor, we can see that number of reviews, total ratings, and number of pages were the most influential in determining average rating.

## Conclusion

After modeling and checking feature importance, I determined that the book covers did not have a strong impact, at least based on the way I represented them in my data set. The features that had the highest impact were number of reviews, number of ratings, and number of pages. This was consistent across the various models I tried. I suspect that adding more features, such as keywords in the description column, may increase model accuracy. Ultimately, it seems that covers are not big indicators of how highly a book is rated.

## Assumptions

For the purposes of this project, I assumed that the data was accurate at the time it was collected. Sometimes Goodreads can be misused and people will spam reviews on books they particularly like or dislike. I also assumed that the data cleaning and preparation I implemented produced an accurately limited set of just science fiction and fantasy books.

## Limitations

The processing power of my local machine was a big limitation on this project. With just text and numeral data for 100,000 records, it would likely not have been as challenging. However, when I attempted to process cover images, I had to  drastically reduce that number in order to have enough processing power to complete the function loop. Having more time would also have been beneficial, providing more opportunity for feature engineering.

## Challenges

Creating representations of the cover images that could be used by predictive models was difficult. I do wonder if flattening the numerical and histogram representations diluted the data, and ultimately was not very useful.

## Future Uses and Additional Applications

In the future, I would like to bring in the description feature and use it to determine if there are consistent words, descriptors, and/or topics that consistently bring in certain ratings. That would also be useful for anyone attempting to market a book, since cover blurbs are the primary way the audience is introduced to the book.

## Recommendations

I recommend exploring different methods of analyzing cover images in the future. I also recommend exploring the description column more to determine if there are certain key words that can be used to predict the average book rating. I also recommend seeing if more features could be added to the data set, such as indicating whether the book is part of a series.

## Implementation Plan

I believe that more work is needed on the modeling for this project prior to implementation, but eventually a new book could be plugged into the model to predict the ratings the book will likely have.

## Ethical Assessment

The data used in this project was free to access from the Goodreads website, and all data was public. There was also no personal information contained in this data set. So ultimately, the ethical risks of this project were very low.