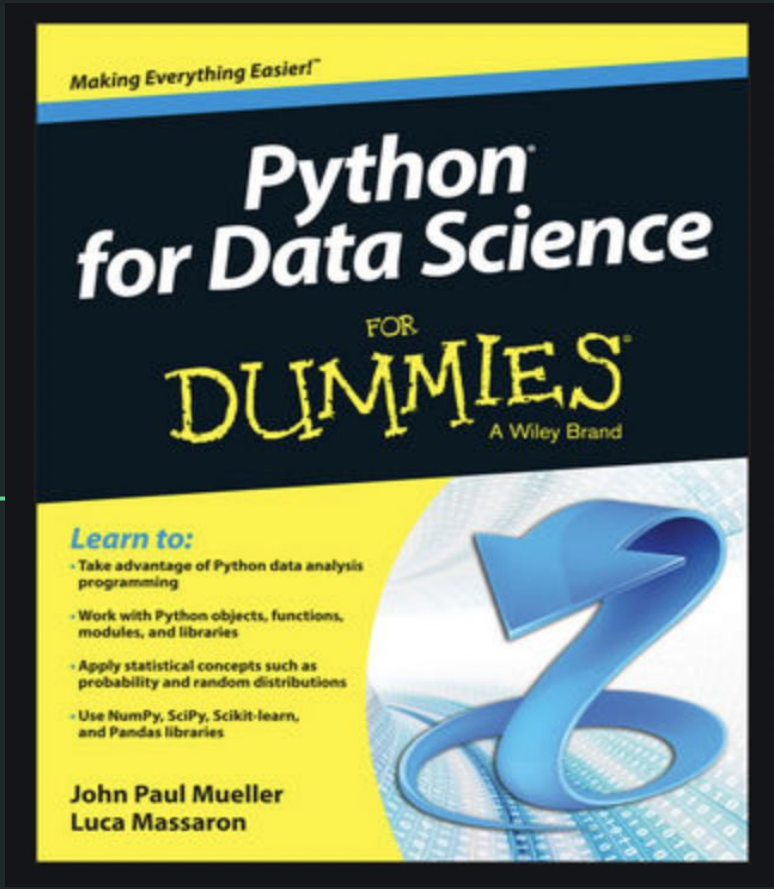
Judging a Book by… its Characteristics!

Linear Regression / Web Scraping

Metis Bootcamp
Project 2
Brandon McNeil



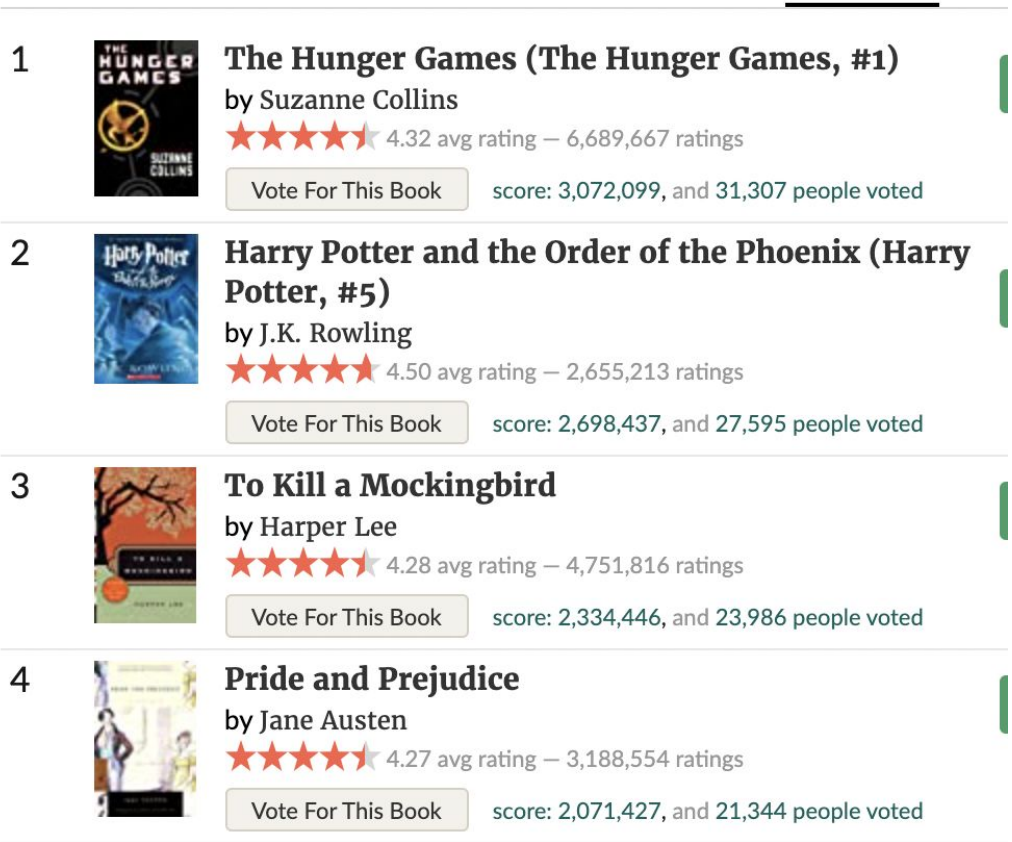
Predicting User Rating from GoodReads.com:

Context:

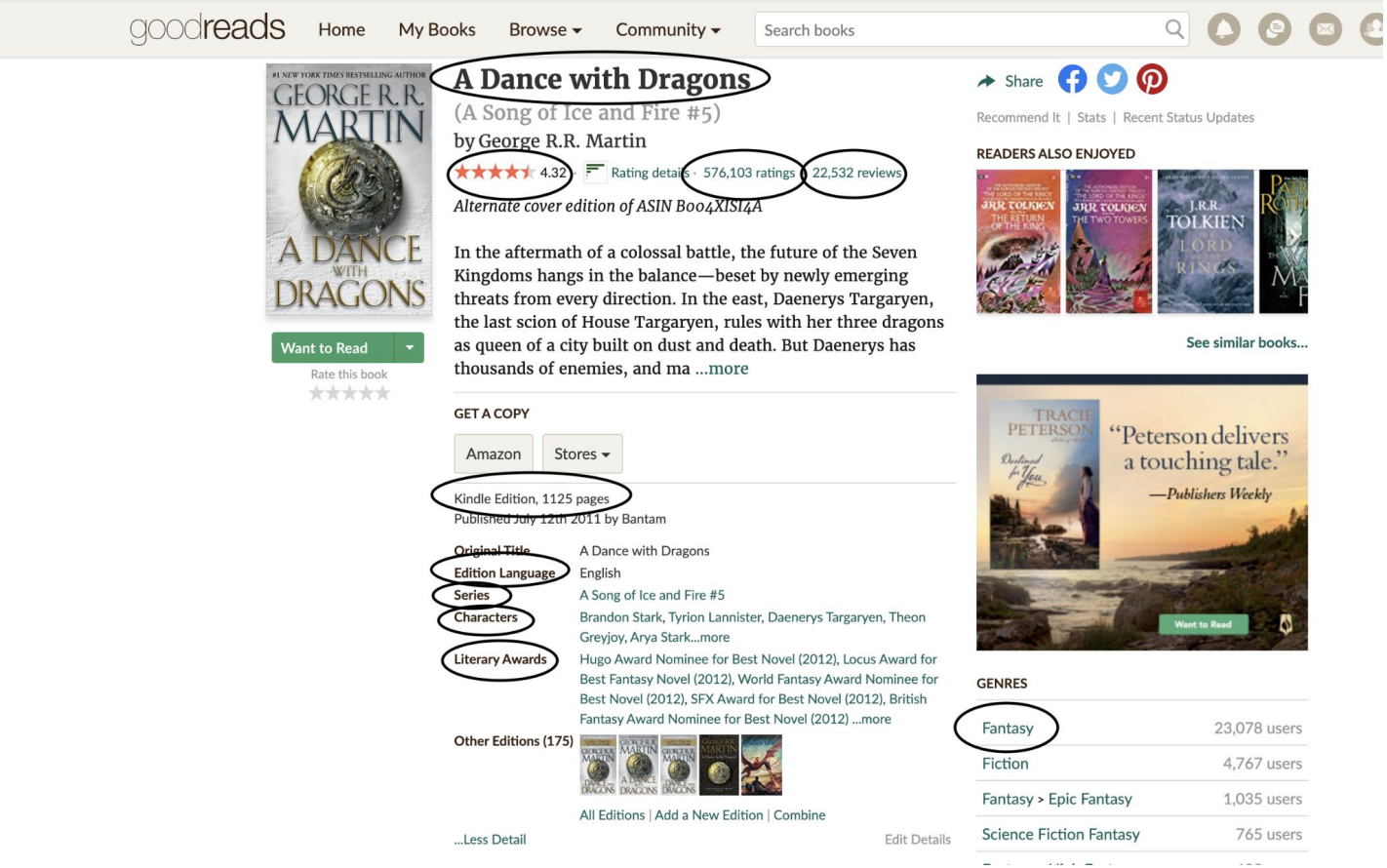
Goodreads.com is a free online platform where users can rate books on a scale of 1 to 5.

Our Analysis Goal:

Can we predict a book's average user rating on **Goodreads.com** based on the book's characteristics?



Data Collection via Web Scraping:



Data Features Scraped:

1) Title
2) Rating
3) Num of Ratings (target feature)
4) Num of Written Reviews
5) Book Type
6) Num of Pages
7) Published Language
8) Is it a Series?
9) Character Count
10) Literary Awards
11) First Listed Genre

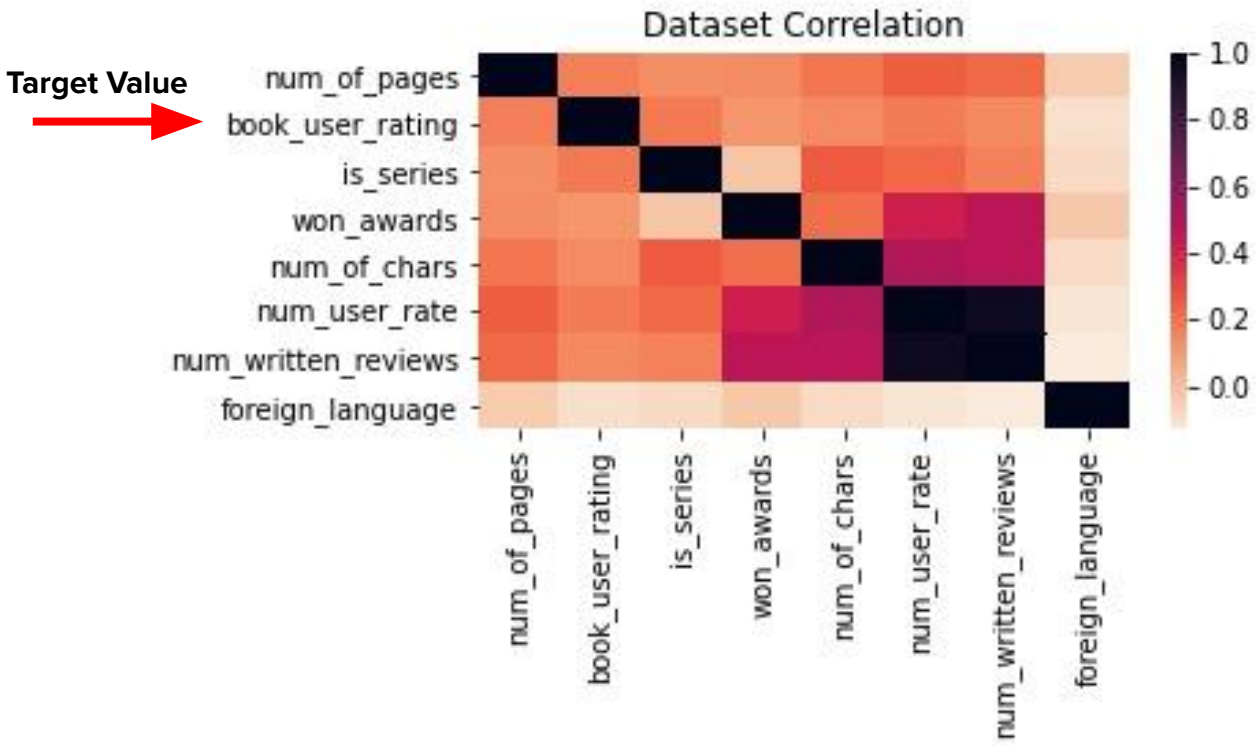
13,000 + records collected!.

Initial Dataset Correlation Findings:

- Using a heatmap, we observed that our features did not have an strong correlation to our target value.

- First modeling on raw data yielded a R-Square of **0.12**.

- Feature engineering was necessary!

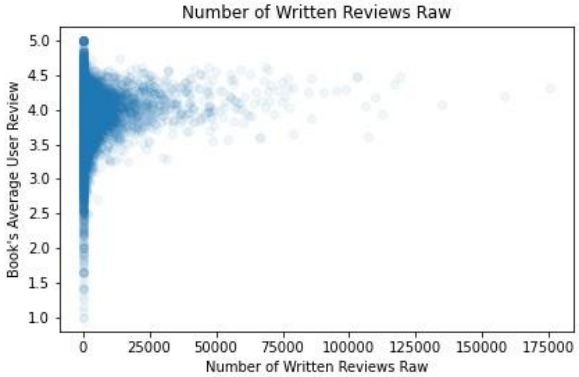


Feature Engineering:

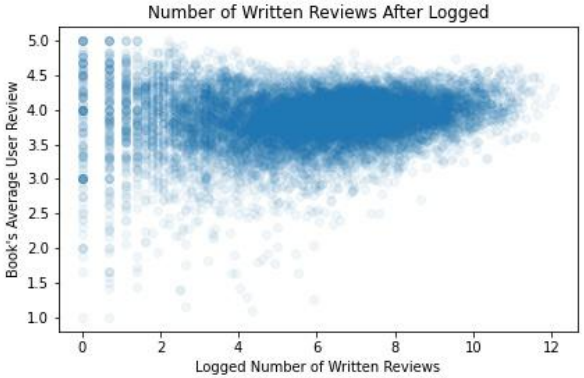
Important Callouts:

- Removed outliers from data that skewed our data (ex. books with over 2000).
- Logged **Number of Written Reviews** to fix distribution issues.
- Created Dummy variables for a **Book's listed Genre** which created around 50 new variables.

Before Log Function



After Log Function



Cross Validation:

- Initial testing indicated Simple Linear Regression and Ridge Regression performed the best on data.

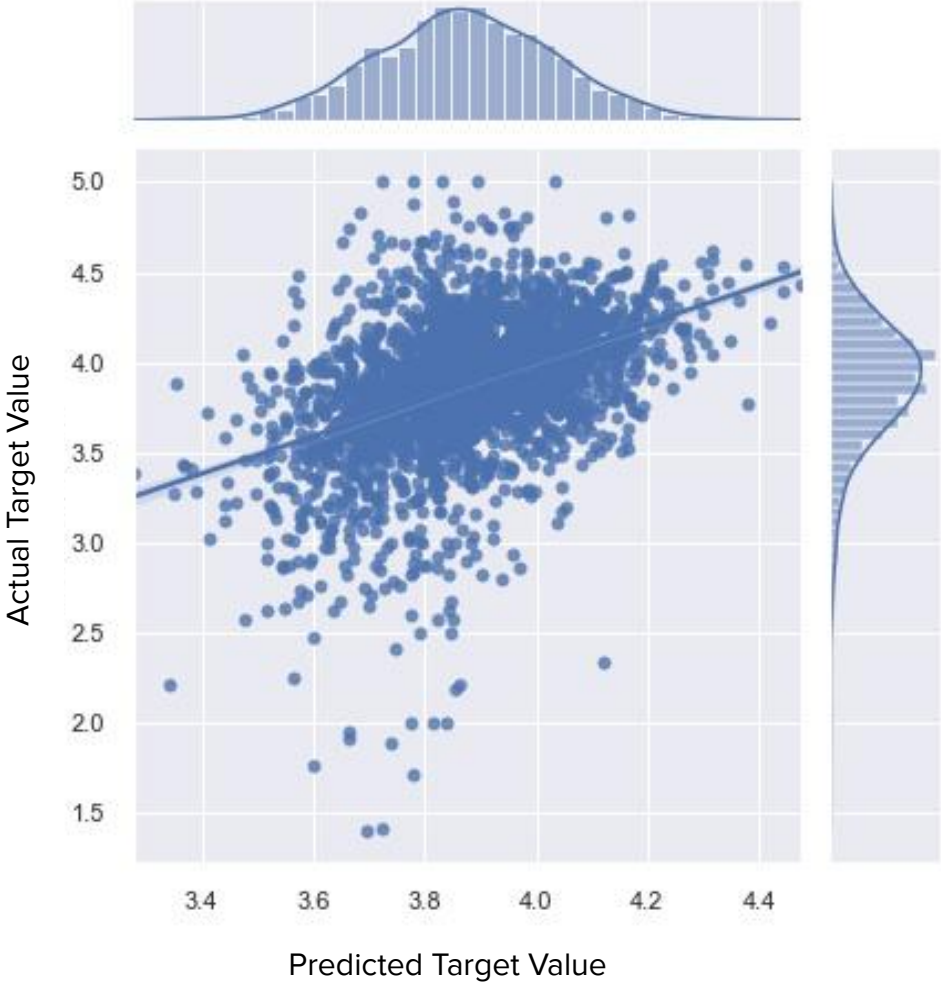
- After running optimal models through KFold cross-validation, both received the same scores - so we kept things simple!

```
Simple mean cv r^2: 0.1725 +- 0.0128
Ridge mean cv r^2: 0.1725 +- 0.0128
```

Results - Predicted vs Actual:

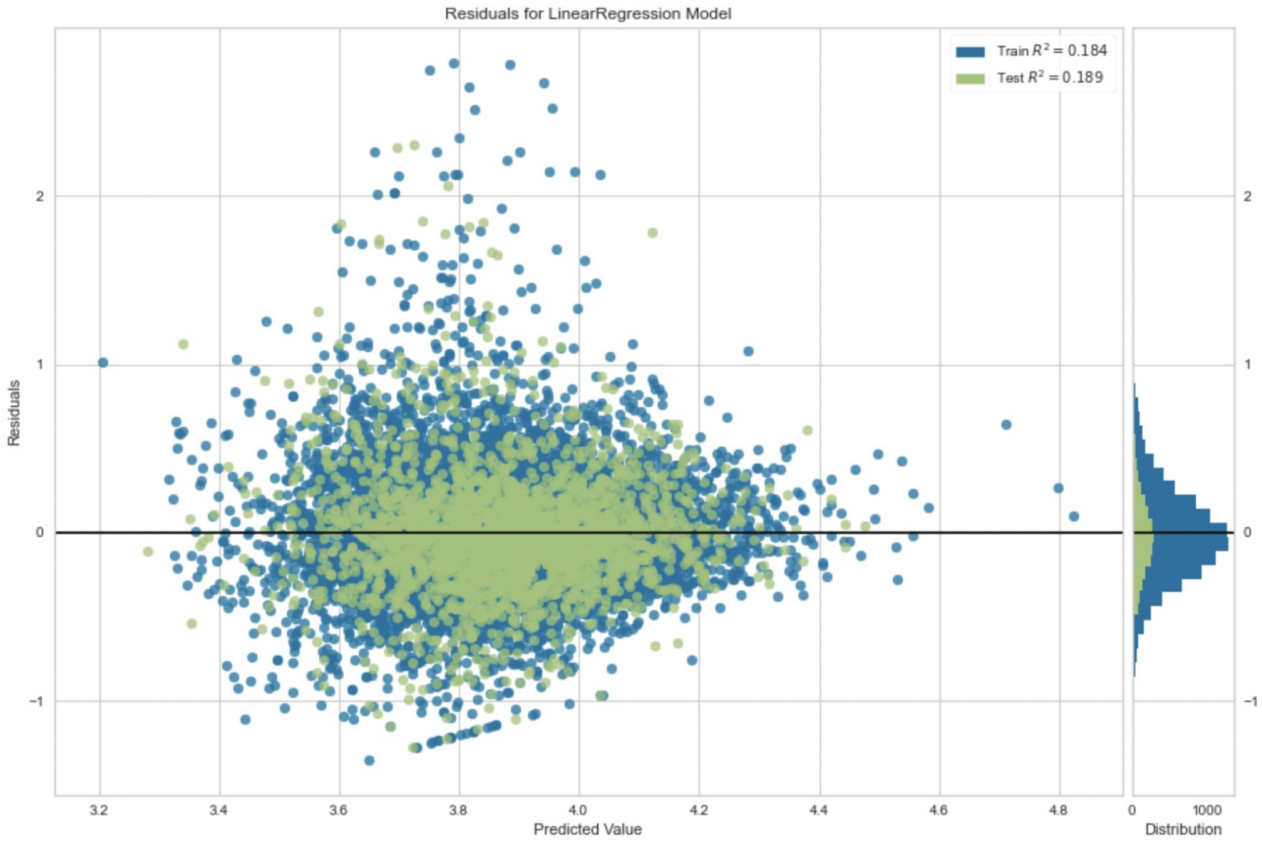
Final R-Squared: 0.184

Predicted and Actual targets are not symmetrical around the line & do not align neatly to it.



Results: Residual Plotting

- While there are some floating points above 2 on the y-axis, most of the clusters are forming around 0-1
- Not a fully symmetrical distribution so improvement is needed!



Key Takeaways:

Predicting reviews based on book features is difficult!

Reviews are a subjective and therefore more elements are required to capture that subjectivity!

Future iterations of this project will look to analyze user's written reviews to aggregate common words to see if we can peel back the subjective blocker.

Appendix - Additional Results:

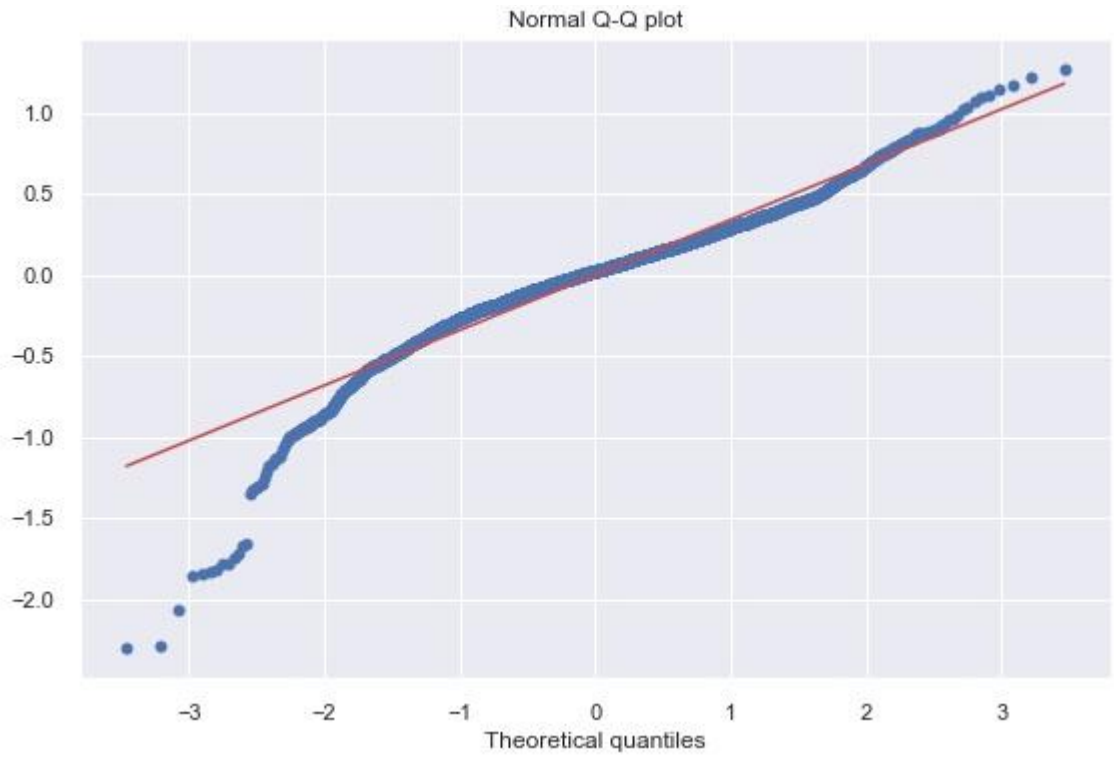
1. **Sum of Squared Error (SSE)** - 317.564

2. **Mean Absolute Error (MAE)** - 0.249

3. **Root Mean Squared Error (RMSE)** - 0.352

Appendix - Q-Q Plot of Residuals:

Our residuals do not fit the line as well as we'd like!



Appendix - Final Coefficients:

```
[('book_cover_type', 0.00035134685747900376),
 ('num_of_pages', -0.09070320517401742),
 ('foreign_language', 0.12298710079757218),
 ('is_series', 0.05619773788479662),
 ('won_awards', 0.007462186959562189),
 ('num_of_chars', 0.022623215338221148),
 ('genre_1', -0.14942867531347964),
 ('num_written_reviews', -0.11379753004007456)]
```