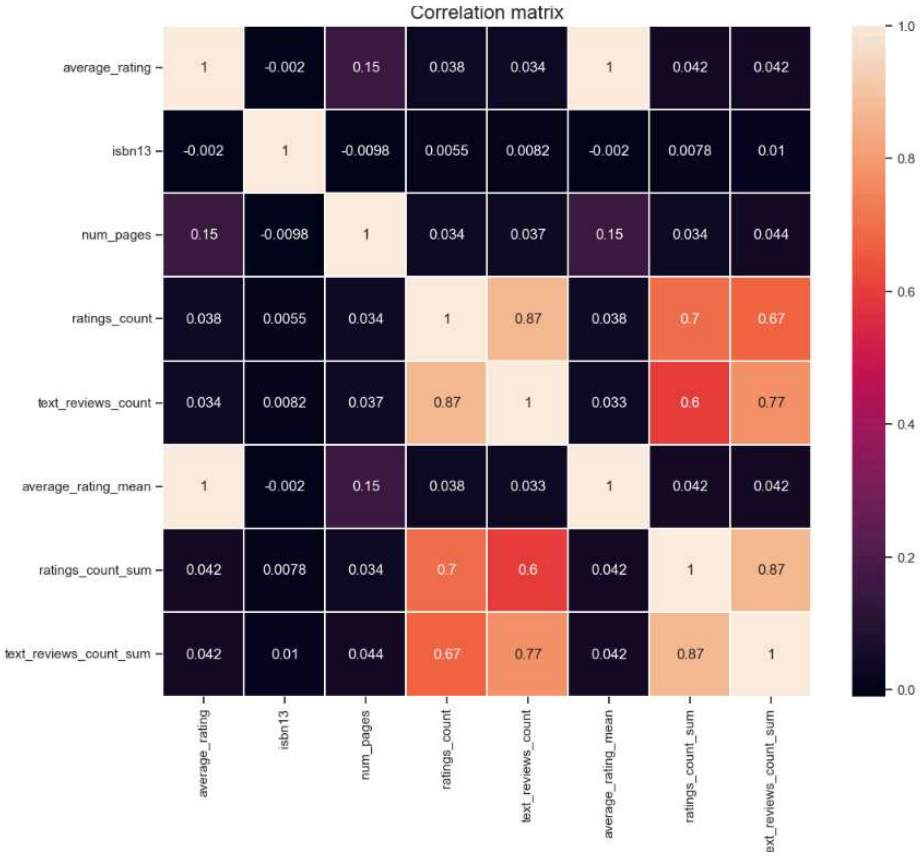
Capstone Final Report

Predict Average Book Rating with Linear Regression

My goal for this project was to build a model that would predict a book's average rating. I found a dataset on Kaggle that was sourced from goodreads.com. The data was complemented by the addition of a genres column.

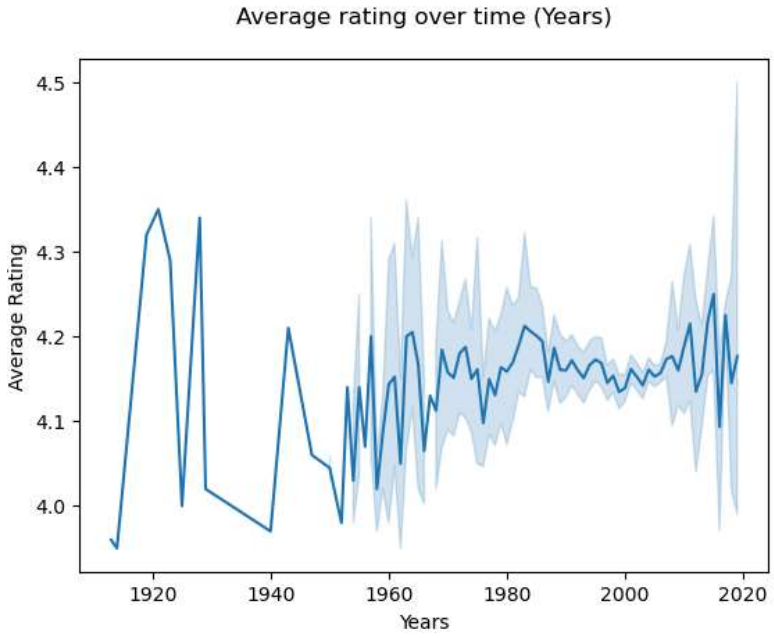
The data required minimal cleaning and had no missing values. Some titles appeared more than once but were just reprints and not duplicates. These titles were merged into one book and given a mean count of ratings and text reviews. Also not every book had a corresponding genre and those books were dropped.

The data needed more features and for some features to be encoded. There was not much correlation as seen from this heat correlation matrix.



Some features were encoded and some manipulated in preparation for model training. For example, just the year was kept from the publication date. I created the features ‘pages_per_review’ and ‘publisher_count’. The ‘title’, ‘author’, ‘publisher’,

‘genre’ and ‘language_code’ were encoded. With addition of these feature and encoding the correlation increased and improved model scoring.



Two models were built, Linear Regression and a Random Forest Regressor. A simple r-squared score linear regression was 60% and 85% for the random forest model. The random forest model was chosen for the modeling portion and it performed well. I chose a random sample of 1000 books or about 10% of the data. The mean difference between predictions and actual ‘average_rating’ was 0.008.

	Actual	Predicted	Difference
2238	4.50	3.485740	1.0143
792	4.67	4.150244	0.5198
7720	4.15	3.631667	0.5183
1578	3.90	3.408504	0.4915
3832	4.20	3.813366	0.3866

```
compare['Difference'].mean()
```

```
-0.008052499999999999
```

The goal stated for this model was to predict a book's average rating based on its features. The model achieves this with an average error of -0.008, which is quite

accurate. This predictive capability could be valuable for authors and publishers assessing potential book performance. It's worth noting the dataset was relatively small compared to the vast inventory of books available on a platform like Amazon, which lists tens of millions of titles. Further research could explore how increasing the dataset size and in-depth hyperparameter tuning might improve the models performance.