Linear Regression: Goodreads Book Rating Predictions

Apoorva Grampurohit

Table of Contents:

- Client?
- Story
- Data Scraping and EDA
- Prediction Model
- Takeaways
- Future work?

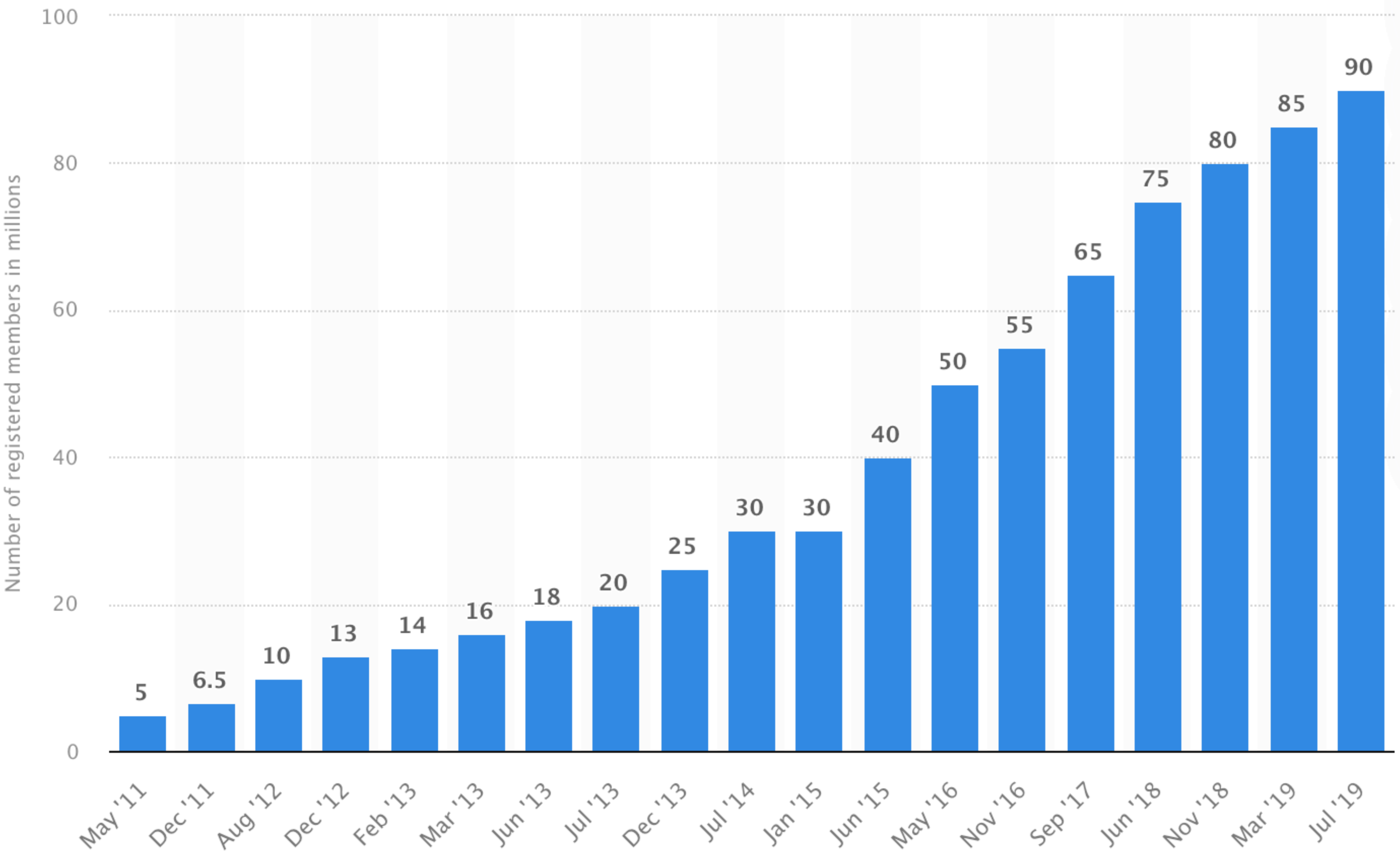
Client:

Barnes and Nobel

- Fortune 1000 company and the bookseller with the largest number of retail outlets in the United States.
- Received a new set of books and want to arrange them based on their ratings.
- Objective: Build a regression model that predicts average ratings of book.
- Business Need: Placing books in a way that readers have better accessibility in turn, increase their revenue.

Why Goodreads?

- goodreads.com : social networking website for book lovers
- Helps you keep track of books you are reading
- Write and read reviews
- Rate books and get recommendations.

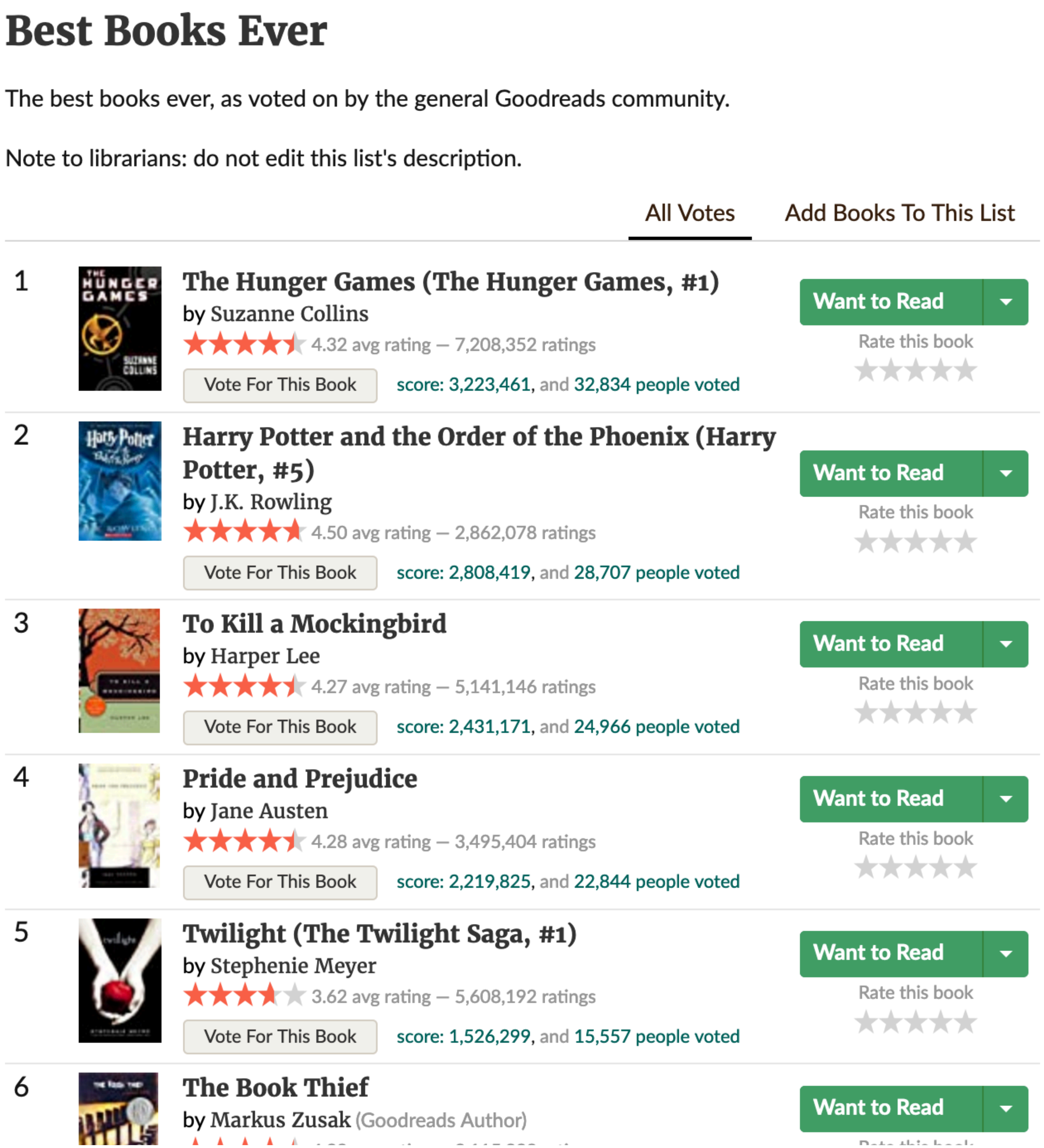


Number of Registered Members on goodreads

Source: www.statista.com

Methodology:

- Data scraped from goodreads.com using BeautifulSoup Python library
- Exploratory Data Analysis
- Building a Linear Regression Model



Best Books Ever

The best books ever, as voted on by the general Goodreads community.

Note to librarians: do not edit this list's description.

1 **The Hunger Games (The Hunger Games, #1)**
by Suzanne Collins
4.32 avg rating — 7,208,352 ratings
score: 3,223,461, and 32,834 people voted

2 **Harry Potter and the Order of the Phoenix (Harry Potter, #5)**
by J.K. Rowling
4.50 avg rating — 2,862,078 ratings
score: 2,808,419, and 28,707 people voted

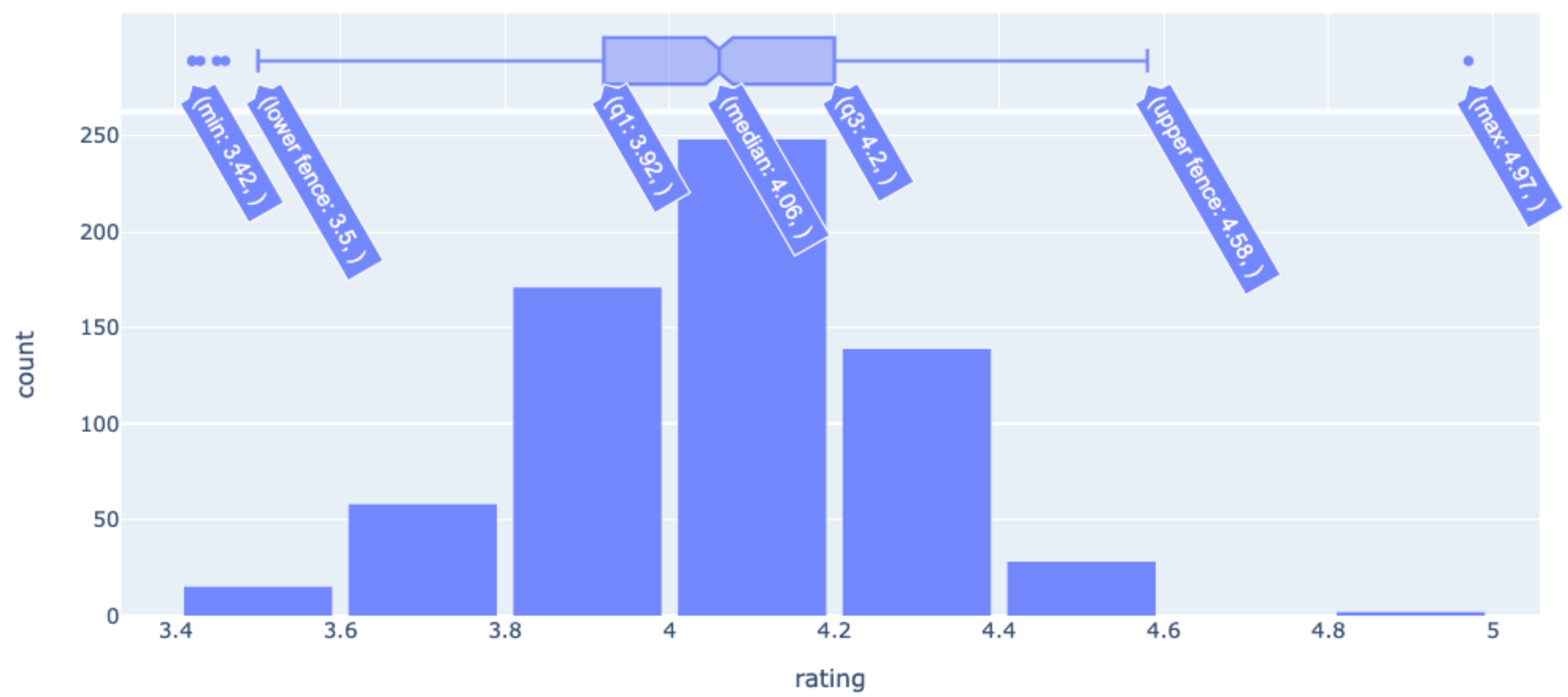
3 **To Kill a Mockingbird**
by Harper Lee
4.27 avg rating — 5,141,146 ratings
score: 2,431,171, and 24,966 people voted

4 **Pride and Prejudice**
by Jane Austen
4.28 avg rating — 3,495,404 ratings
score: 2,219,825, and 22,844 people voted

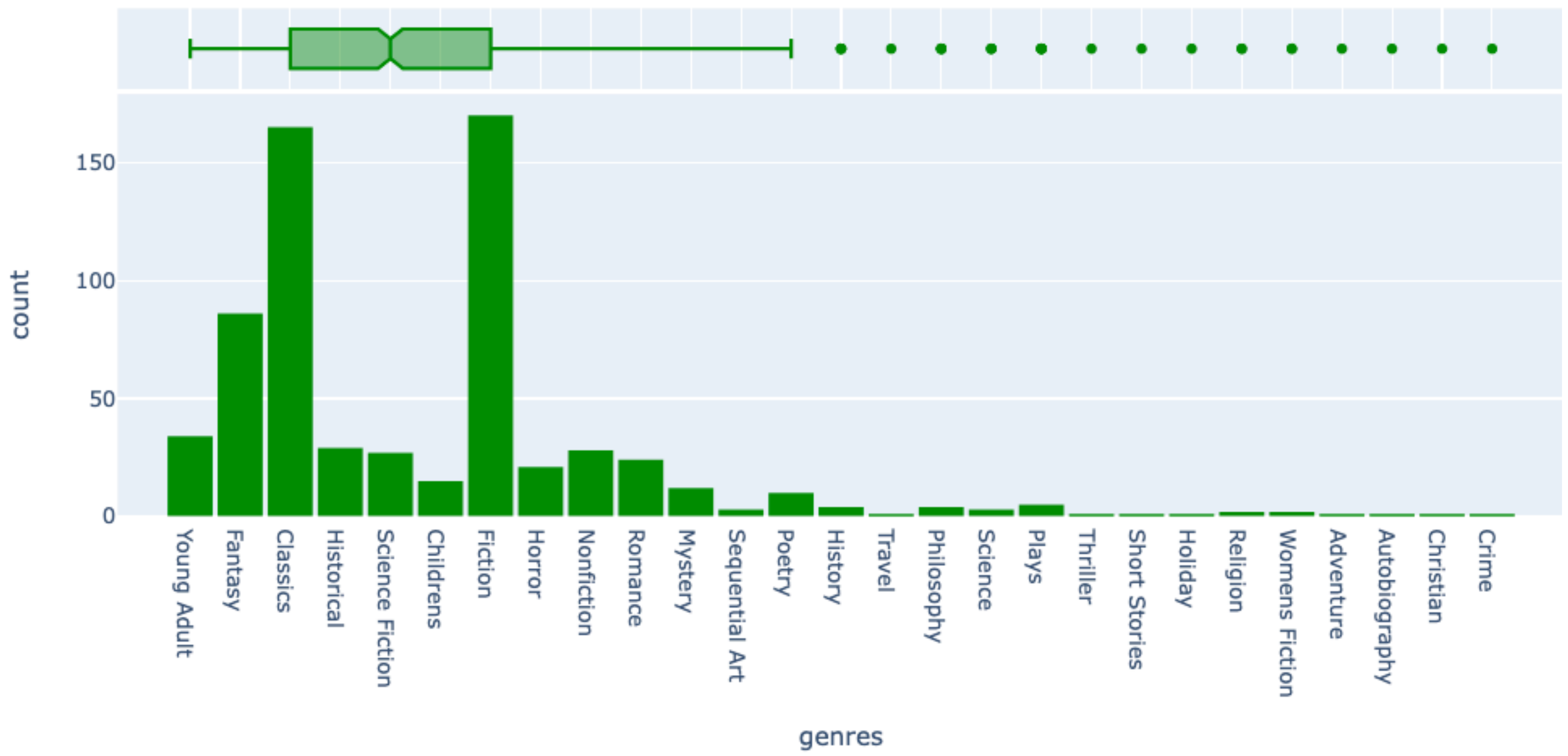
5 **Twilight (The Twilight Saga, #1)**
by Stephenie Meyer
3.62 avg rating — 5,608,192 ratings
score: 1,526,299, and 15,557 people voted

6 **The Book Thief**
by Markus Zusak (Goodreads Author)

Distribution of ratings



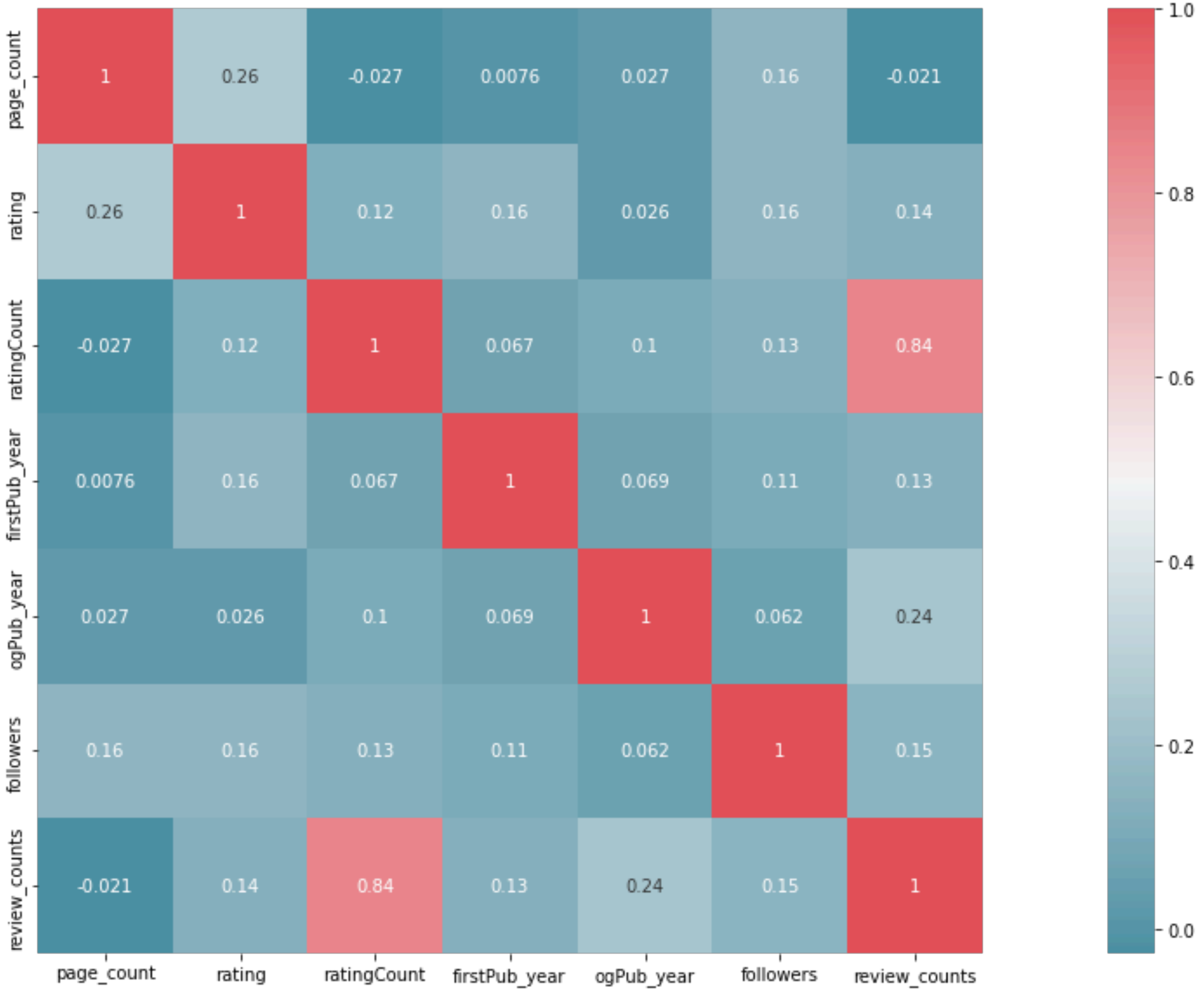
Distribution of genres



Correlation with Ratings:

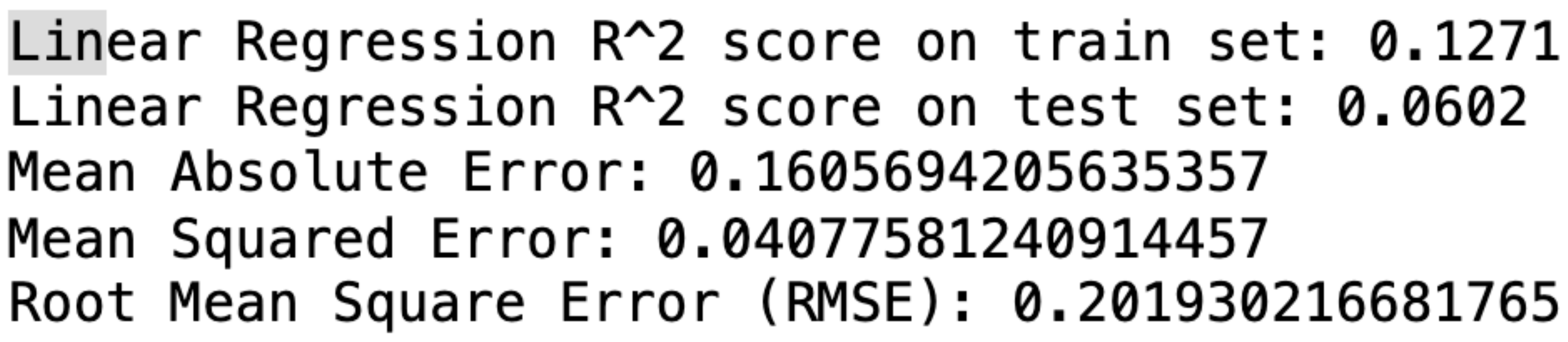
- page_count, review_counts, rating counts, followers, firstPub_year has a better correlation

- Dropped ogPub_year

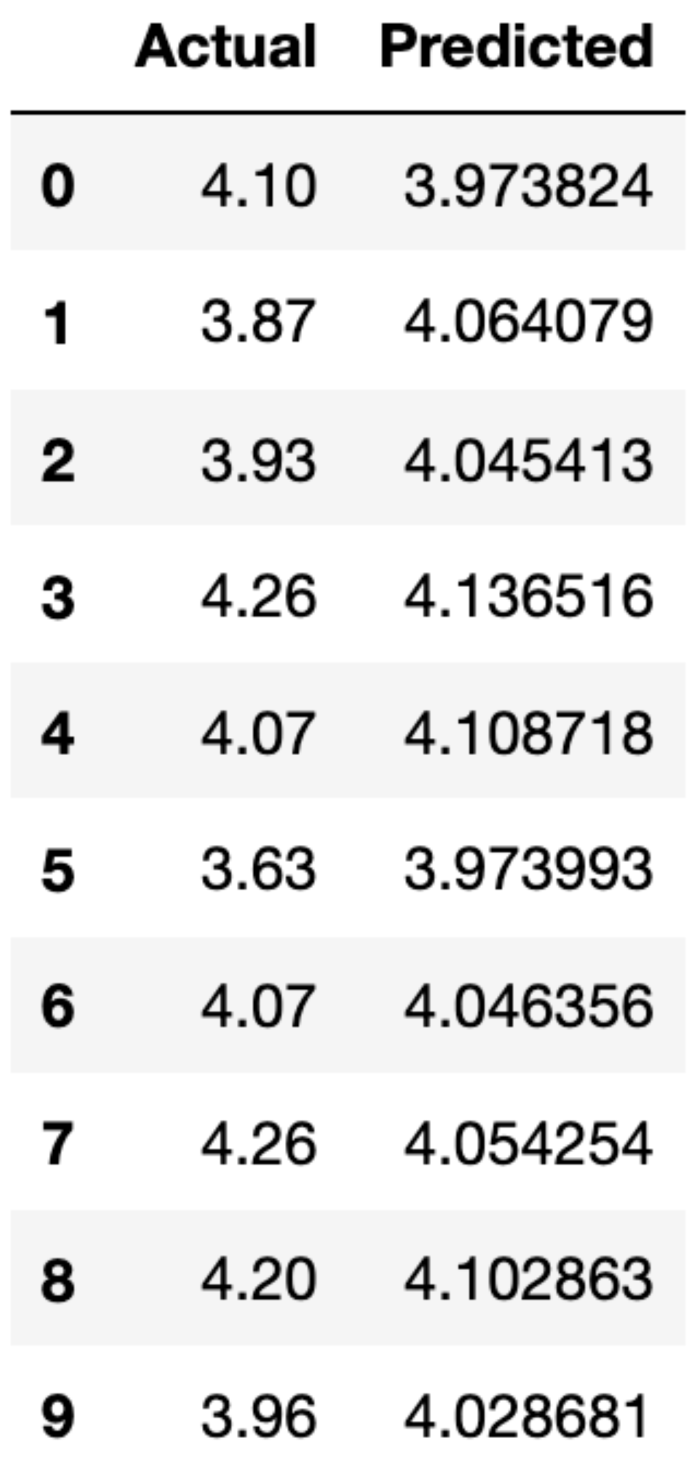


	page_count	rating	ratingCount	firstPub_year	ogPub_year	followers	review_counts
page_count	1	0.26	-0.027	0.0076	0.027	0.16	-0.021
rating	0.26	1	0.12	0.16	0.026	0.16	0.14
ratingCount	-0.027	0.12	1	0.067	0.1	0.13	0.84
firstPub_year	0.0076	0.16	0.067	1	0.069	0.11	0.13
ogPub_year	0.027	0.026	0.1	0.069	1	0.062	0.24
followers	0.16	0.16	0.13	0.11	0.062	1	0.15
review_counts	-0.021	0.14	0.84	0.13	0.24	0.15	1

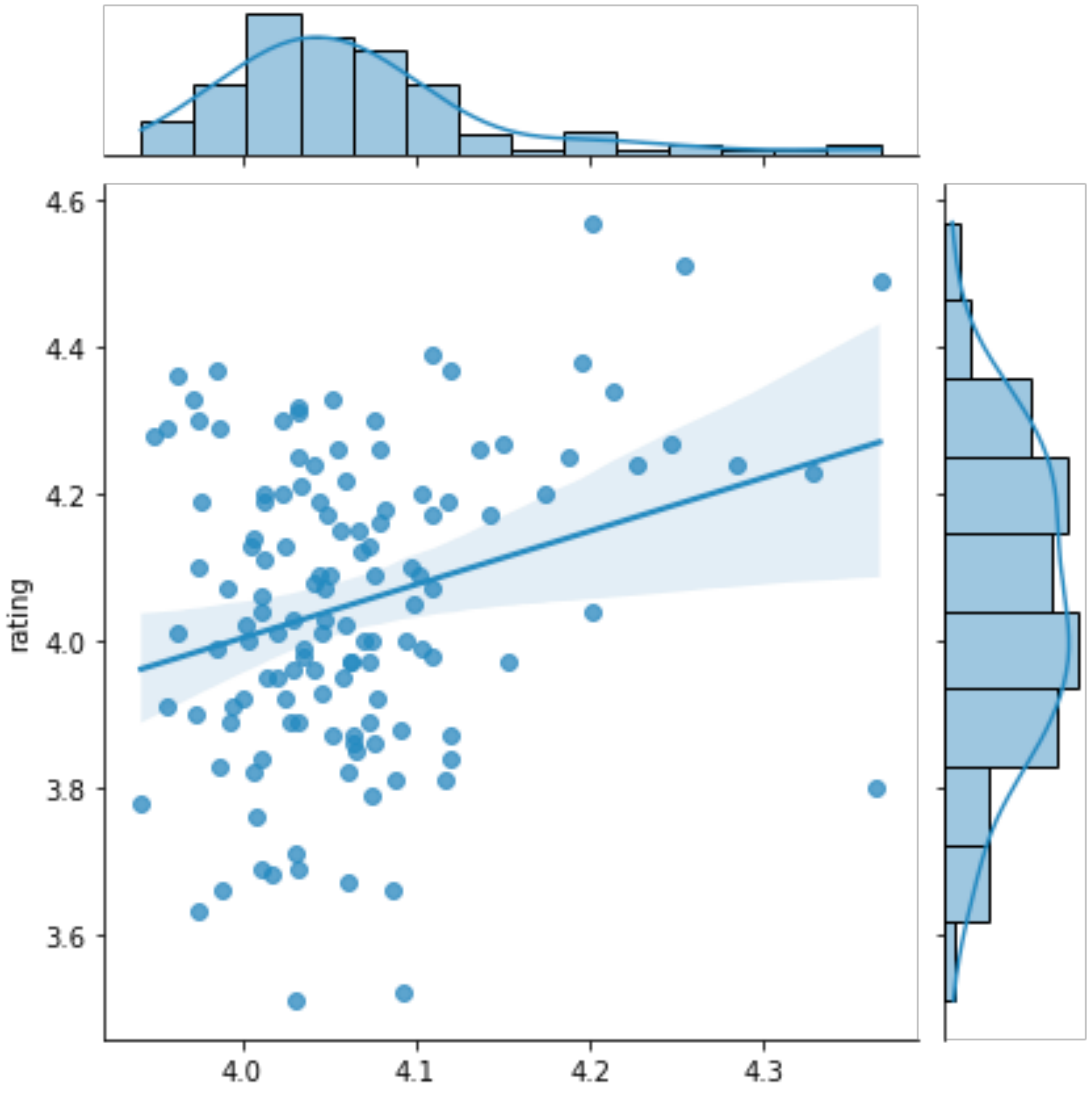
Linear Regression Model:



```
Linear Regression R^2 score on train set: 0.1271
Linear Regression R^2 score on test set: 0.0602
Mean Absolute Error: 0.1605694205635357
Mean Squared Error: 0.04077581240914457
Root Mean Square Error (RMSE): 0.201930216681765
```



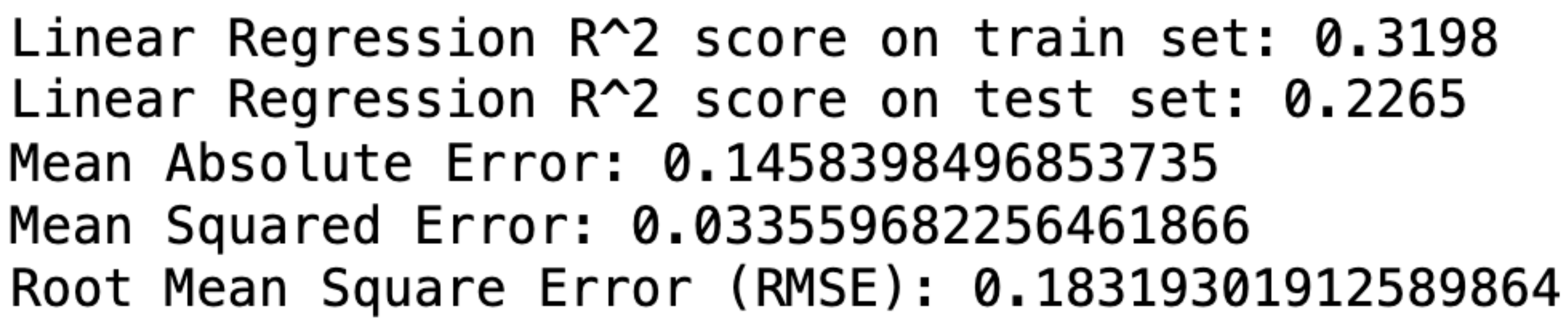
	Actual	Predicted
0	4.10	3.973824
1	3.87	4.064079
2	3.93	4.045413
3	4.26	4.136516
4	4.07	4.108718
5	3.63	3.973993
6	4.07	4.046356
7	4.26	4.054254
8	4.20	4.102863
9	3.96	4.028681



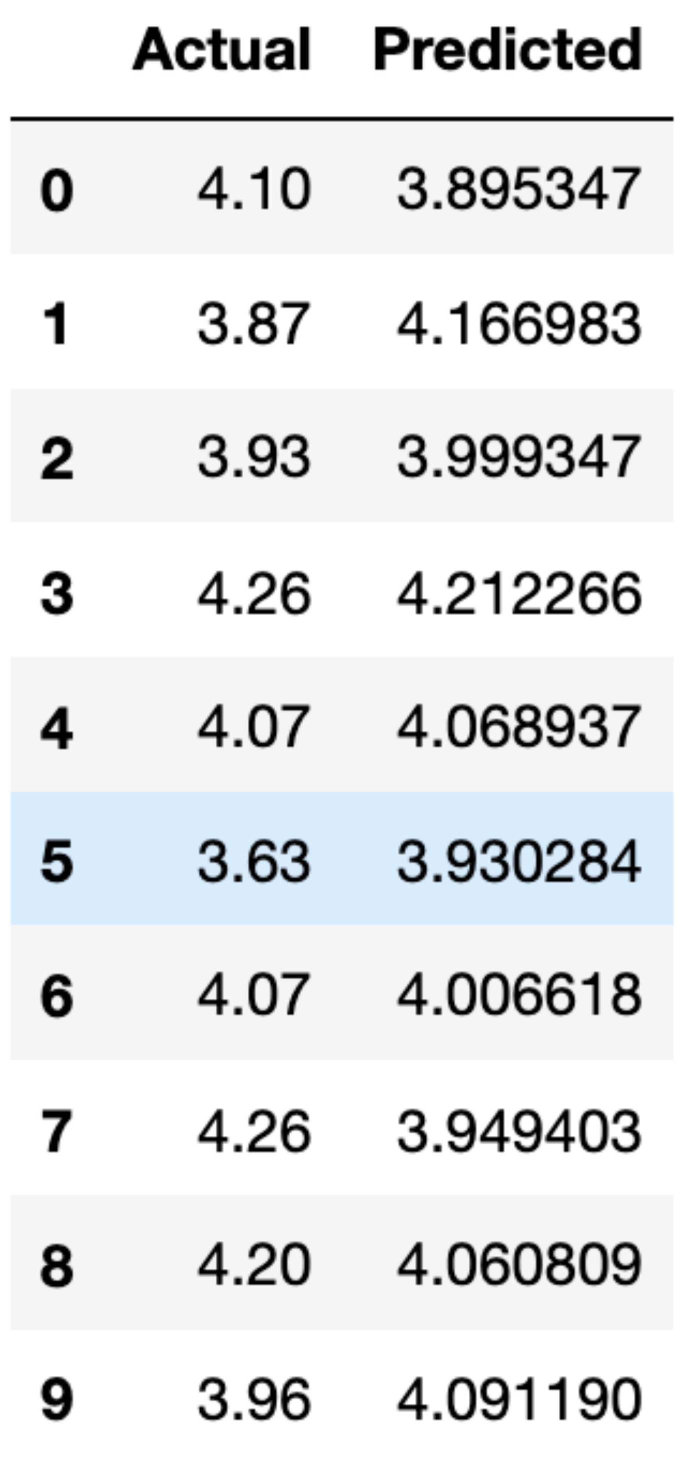
Just with Numerical features:

Linear Regression Model:

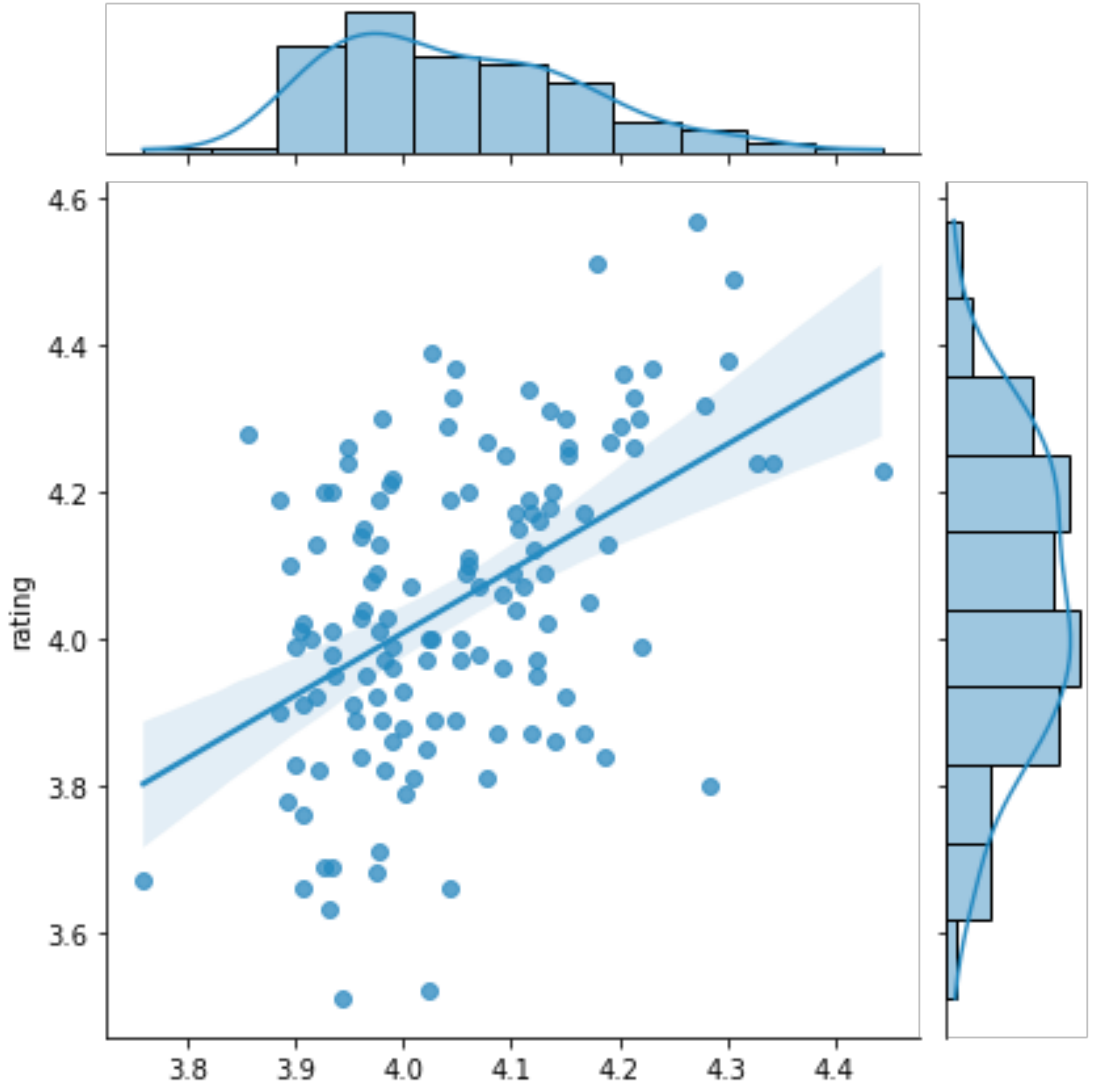
With categorical values



```
Linear Regression R^2 score on train set: 0.3198
Linear Regression R^2 score on test set: 0.2265
Mean Absolute Error: 0.1458398496853735
Mean Squared Error: 0.033559682256461866
Root Mean Square Error (RMSE): 0.18319301912589864
```



	Actual	Predicted
0	4.10	3.895347
1	3.87	4.166983
2	3.93	3.999347
3	4.26	4.212266
4	4.07	4.068937
5	3.63	3.930284
6	4.07	4.006618
7	4.26	3.949403
8	4.20	4.060809
9	3.96	4.091190



Model	R^2	MAE	MSE	RMSE
Linear Regression Numerical Features only	0.2265	0.1458	0.0335	0.1831
Linear Regression Numerical +Categorical	0.2210	0.1446	0.0339	0.1841
Polynomial Features	0.00526	-	-	-
Ridge Regression	0.1953	0.1495	0.0349	0.1868

Model Performance

Conclusion:

- The model built on just the numeric values can be used to predict the ratings.
- Adding Genres didn’t really make much difference.
- The important features are:
  - Number of pages
  - Review counts
  - Rating counts
  - number of followers an author has

Future Work:

- More data points and features
- More thorough data cleaning, specially handling outliers will result in accurate models.
- Data from multiple sources can increase the reliability and generalization of the models.

Thank you