What Makes A Best-Selling Book

Lindsey Henyan

Final Project Write-Up

Data Wrangling - DS2500

Prof. Marina Kogan

April 23, 2024

What Makes A Best-Selling Book

Over 2.2 million books are published every year and there are over 158 million unique books in the world, but what makes a book a cult classic and a best-seller? Many factors can play a role in the sales and ratings of a book. Whether it's from a well-known author, a new release, popularity, and possibly the language that it is in. The question posed is the relationship between the average rating of books, considering factors such as authors, publication date, language, copies sold, and ratings count This research and analysis will illustrate what factors play the biggest role in the average rating of a book. To formulate a conclusion I studied scatterplots, regression models, heat maps, and an analysis of variance test.

Working with the Data

Data Sets

The Best-Selling Books data set contains a list from Wikipedia of best-selling books and book series. This data set includes the title, author, language, year, copies sold, and the genre. However, the genre had missing data which led me to drop this data. The Goodreads data set is a list of books from the application, Goodreads. This app allows users to rate and review books they have read and also provides recommendations based on these reviews. This data set includes the bookID, title, authors, average rating, ISBN, ISBN13, language code, number of pages, number of ratings, number of text reviews, publication date, and publisher. These data sets allow me to look at features that can compare the average ratings of books to other variables such as number of pages, authors, published year, and other factors.

Data Cleaning

To join the two data sets together, I performed an inner join where the sets are connected based on the titles of the books. From the combined data I am able to take into consideration the

variables in the best-selling books with the average rating and the variables in the Goodreads data sets. To clean the data, I ordered the data based on the sum of missing values in each column, where genre was the only column to have missing data. As a result, I dropped the column since I could not predict the genre based on the values given in the data set.

Testing and Visualizations

Visualizations

The heatmap, figure 1A, illustrates the correlation between multiple variables, however, there are only two with some impact. Ratings count and copies sold have a positive relationship. Average rating and year have a negative relationship. As shown, the average rating has a slight positive relationship with the number of pages at 0.11, however, that is not a significant pull. There is very little correlation between the quantitative variables in this data set with the average rating.

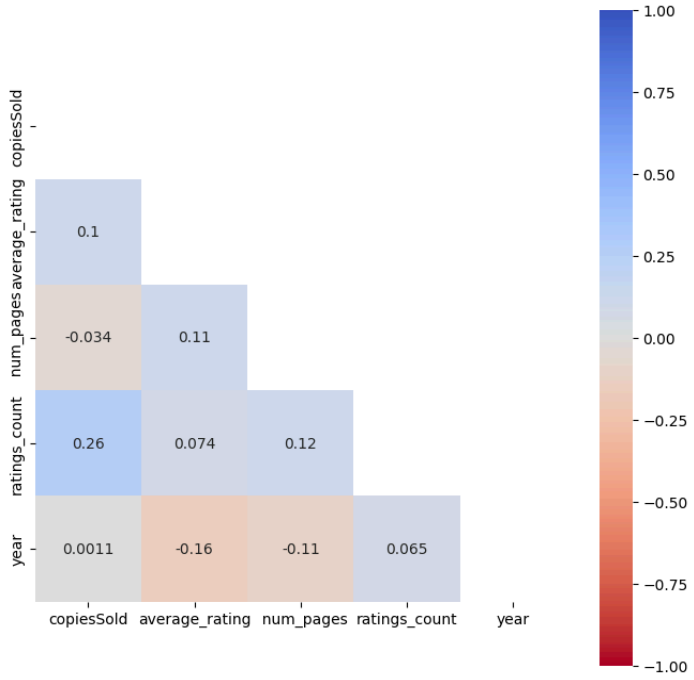

FIGURE 1A.

The scatterplot, figure 2A, models the number of pages versus the average rating. A majority of the books are under 600 pages with the lowest rating being about 3.65. The chart seems to have a slightly positive correlation. Since the correlation is not significant, it coincides with the results of the heatmap where the number of pages has very little effect on the average rating.

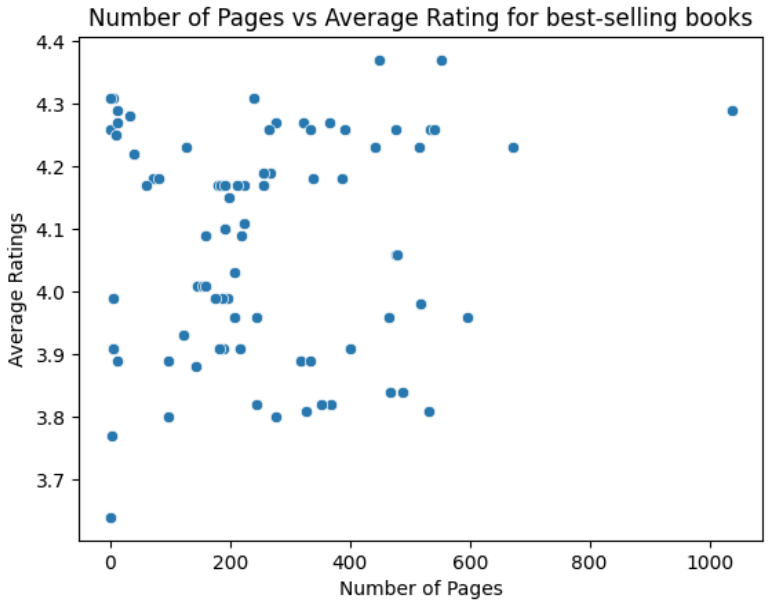

FIGURE 2A.

This next scatterplot, figure 3A, compares the published year with the average rating. Most of the books have been released in the last 100 years with the earliest being before 1825. The graph appears to have a slight positive correlation since the points increase from left to right. There is one point that has a strong effect on the distribution which is the earliest book since it has a rating of 4.25.

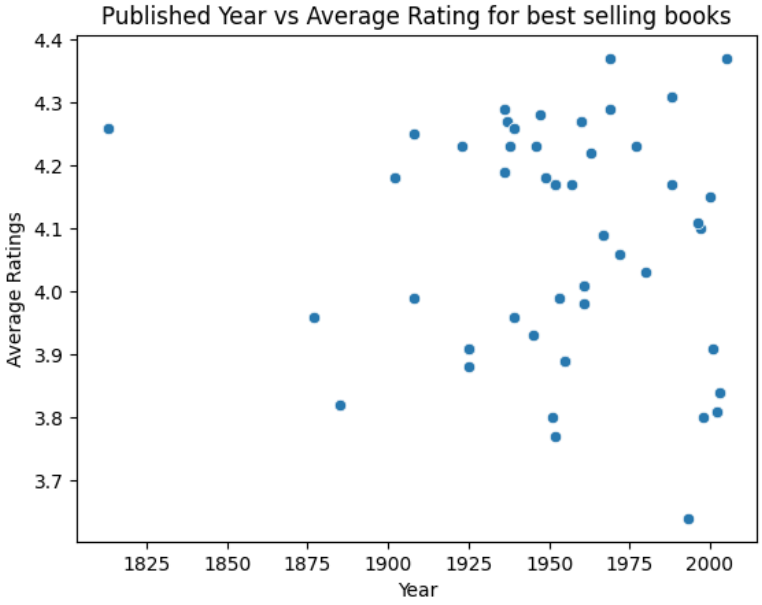

FIGURE 3A.

Figure 4A is a scatterplot that compares copies sold to the average rating. There seems to be a negative relationship among this data set, where most of the points are in the top left of the graph. However, there are two who have over 100 million copies which is the most out of any other novel with a rating of about 4.25. These points can cause a significant impact by skewing the correlation among these variables.

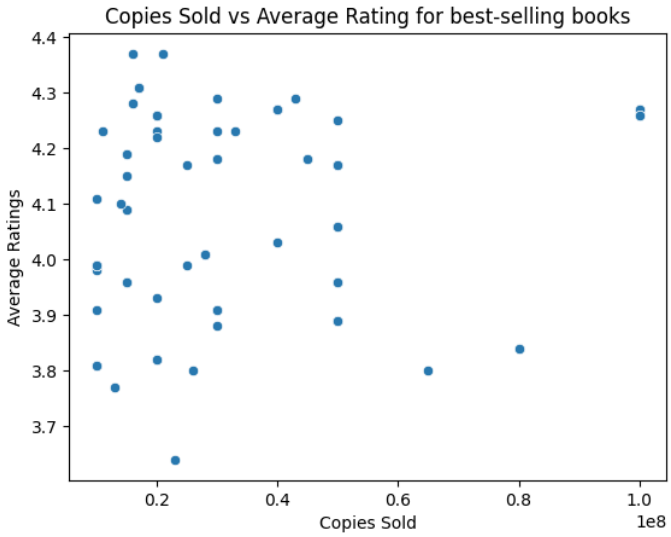

FIGURE 4A.

Do Authors Impact the Rating?

Another element to visualize is whether the average rating score is dependent on the author. To visualize this, I created a bar chart with the number of times an author appears on the list illustrated in Figure 5A. The two leading authors on the list F. Scott Fitzgerald and Roald Dahl, have 6 entries in the data set. The second graph, figure 5B, is a histogram looking at the frequency of authors for each average rating. The data is left skewed with the peak being at 12 authors at about a 4.3 rating. This shows that a majority of authors have above a 4.0 rating.

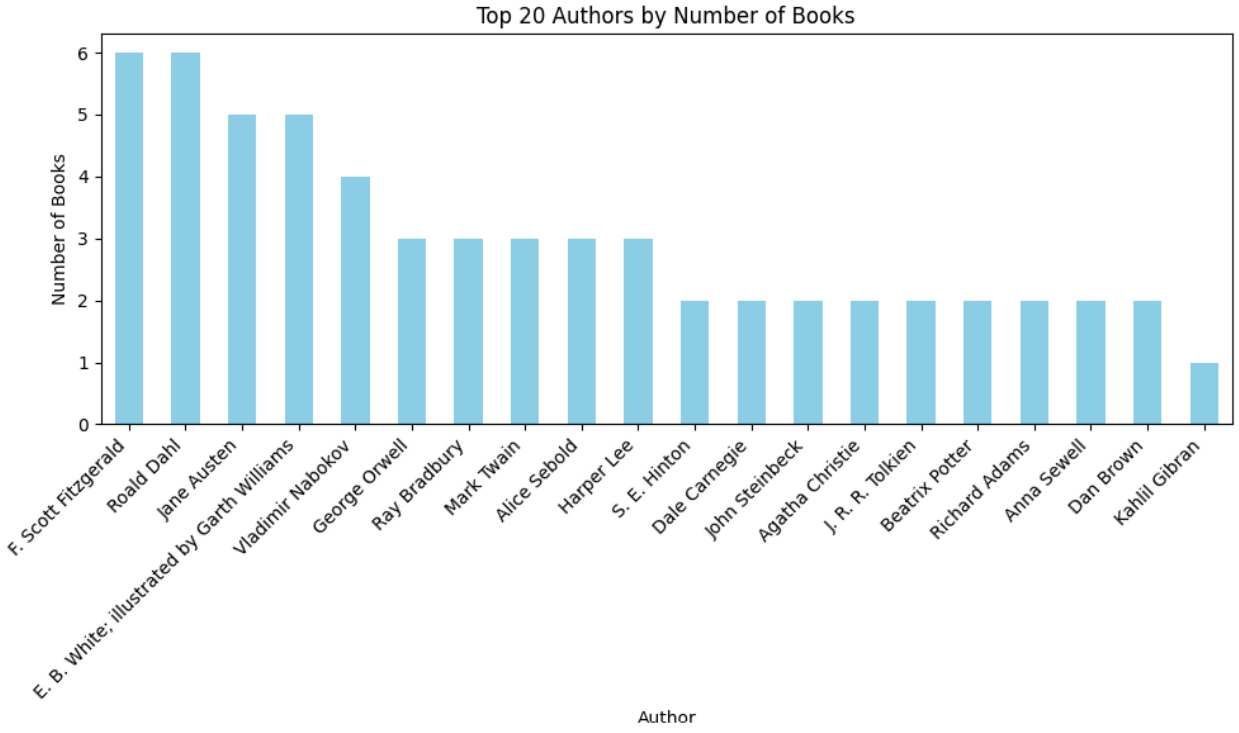

FIGURE 5A.

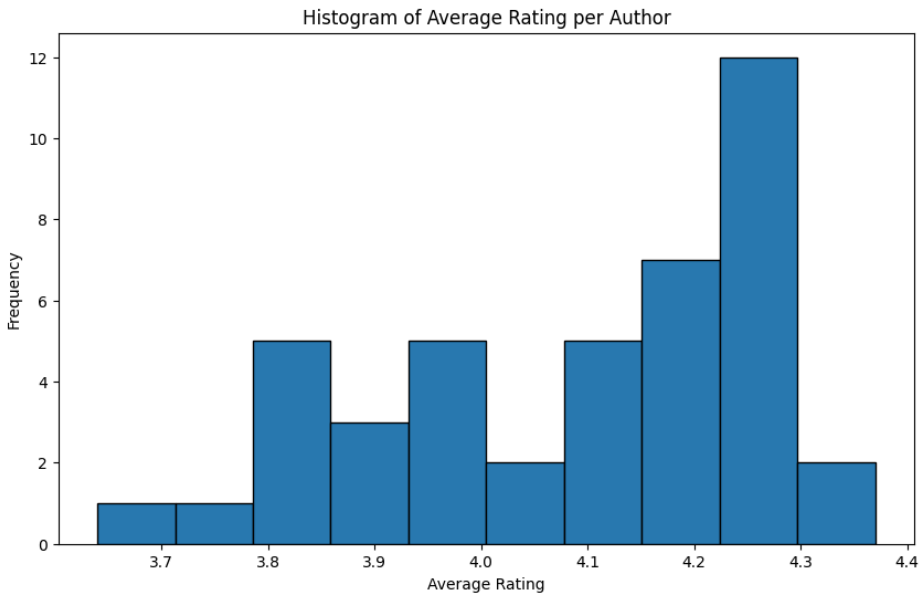

FIGURE 5B.

Does the Publisher Make a Difference?

After locating all of the different entries of "The Great Gatsby", seen in Figure 6A, I compared the copies sold, the average ratings, and the publication dates. The main difference I found is the average rating for Penguin Global is slightly lower at 3.88 than the rest of the publishers who get 3.91. In addition, some of the versions are listed to have 4 or 6 pages which are audiobooks, making the number of pages less. The histogram, figure 6B, displays the number of times the top 20 publishers were listed. Harper Collins has 20 novels on the list, with the rest of the publishers only having no more than three books. This illustrates that every publisher has the opportunity to create a best-selling book with a high average rating.

	title	author	language	year	copiesSold	bookID	authors	average_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publisher
29	The Great Gatsby	F. Scott Fitzgerald	English	1925	30000000.0	4673	Kathleen Parkinson/F. Scott Fitzgerald	3.88	0140771972	9780140771978	eng	144	557	28	11/25/2003	Penguin Global
30	The Great Gatsby	F. Scott Fitzgerald	English	1925	30000000.0	4674	F. Scott Fitzgerald/Tim Robbins	3.91	0060098910	9780060098919	eng	6	258	58	10/1/2002	Caedmon
31	The Great Gatsby	F. Scott Fitzgerald	English	1925	30000000.0	4675	F. Scott Fitzgerald/Alexander Scourby	3.91	1572702567	9781572702561	eng	4	63	13	3/13/2002	Audio Partners
32	The Great Gatsby	F. Scott Fitzgerald	English	1925	30000000.0	4677	F. Scott Fitzgerald	3.91	0140620184	9780140620184	eng	188	2729	245	1/13/1994	Penguin Books
33	The Great Gatsby	F. Scott Fitzgerald	English	1925	30000000.0	14235	F. Scott Fitzgerald	3.91	0891906797	9780891906797	eng	182	149	7	9/1/1925	Amereon Ltd
34	The Great Gatsby	F. Scott Fitzgerald	English	1925	30000000.0	27451	F. Scott Fitzgerald/Matthew J. Bruccoli	3.91	0684801523	9780684801520	eng	216	9844	1050	6/1/1995	Scribner

FIGURE 6A.

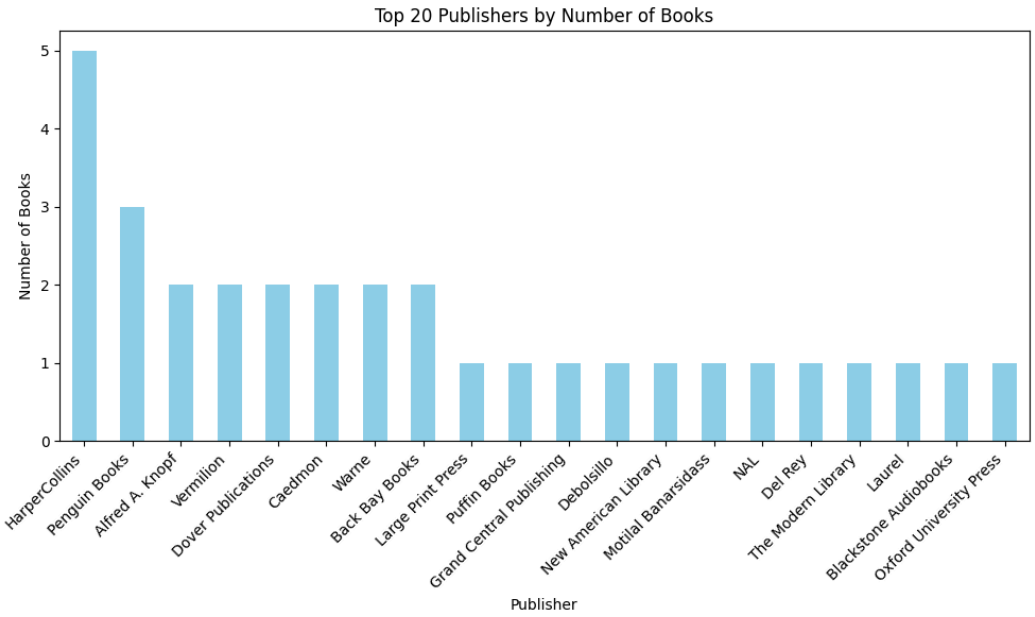

FIGURE 6B.

Testing

After creating a regression model, figure 7A, for the average rating by calculating the copies sold, number of pages, ratings count, and year published. This equation results in a very low R-squared suggesting that the average rating is only made up of 5.1% of the independent variables. All of the coefficients for the variables are close to zero with year being the only negative value.

OLS Regression Results

Dep. Variable:	average_rating	**R-squared:**	0.051
Model:	OLS	**Adj. R-squared:**	0.002
Method:	Least Squares	**F-statistic:**	1.044
Date:	Fri, 19 Apr 2024	**Prob (F-statistic):**	0.390
Time:	19:31:09	**Log-Likelihood:**	27.323
No. Observations:	82	**AIC:**	-44.65
Df Residuals:	77	**BIC:**	-32.61
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	5.2149	0.857	6.086	0.000	3.509	6.921
copiesSold	7.518e-10	9.28e-10	0.810	0.421	-1.1e-09	2.6e-09
num_pages	9.444e-05	0.000	0.893	0.374	-0.000	0.000
ratings_count	1.795e-08	4.14e-08	0.434	0.666	-6.45e-08	1e-07
year	-0.0006	0.000	-1.400	0.166	-0.001	0.000

Omnibus:	14.039	**Durbin-Watson:**	0.996
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	4.287
Skew:	-0.186	**Prob(JB):**	0.117
Kurtosis:	1.944	**Cond. No.**	1.72e+09

FIGURE 7A.

However, I also performed a model with how the different authors affect the average rating. The R-squared for this model was 100% and the adjusted was 99%, interpreting that the model explains all of the variability in the dependent variability. The author who had the largest coefficient was Jane Austen with 3.7. This regression model proves that the authors explain the variability among the average ratings. So, the author is in fact one of the leading factors in a book's rating.

OLS Regression Results

Dep. Variable:	average_rating	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	0.999
Method:	Least Squares	**F-statistic:**	2776.
Date:	Wed, 24 Apr 2024	**Prob (F-statistic):**	1.57e-53
Time:	22:38:55	**Log-Likelihood:**	359.39
No. Observations:	82	**AIC:**	-626.8
Df Residuals:	36	**BIC:**	-516.1
Df Model:	45		
Covariance Type:	nonrobust		

FIGURE 7B.

To further the analysis, I performed the statistical test, analysis of variance, for the average rating of each author to see if there was a significant difference between each group. After running the test, the F-statistic resulted in 13.049 which is the ratio of variance between the authors and the variance of ratings for the authors. Secondly, the p-value was zero making this test statistically significant leading to the authors having an impact on the average rating. This corresponds with the regression model using the author where the variability of average rating is explained by the authors.

Counterarguments and Weakness

Although the analysis did show some correlation between the authors and the average rating, there are several weaknesses and limitations with this data set. Book ratings often come from more than just the statistical information about a book instead they may come from more qualitative data. This can be found through diction, genre, relatability, and the overall story. An interesting aspect that could have been provided would have been user data from the Goodreads application about the books the user has read and their ratings of the books. This information would most likely create more significant results and lead to a more in-depth conclusion about what makes a high-rated book.

Results

After conducting this analysis, I have concluded that there are no quantitative variables that were provided in the data sets that have a strong effect on the average rating. However, the author of the book has proven to hold some significance towards the average rating of a book. This leads to the conclusion that many people will gravitate towards books from authors they know. Looking at the spread of the authors illustrates that many authors have several books on

their best-sellers list, which leads to them having a fan base who will continuously buy their books. Average rating can be dependent on so many other factors that were not provided such as personal interest, relatability, genre, and how the story is written. Therefore, the average rating regression model of just the quantitative variables did not produce significant evidence, but rather the author's regression model did since the rating is more of a personal opinion than a mathematical science.