Historically Best Selling and Amazon Success

GitHub Repository:

1. Introduction

With a major shift to online retail, the success of historically best-selling books is influenced by factors like customer reviews, ratings, and sales ranks. This project explores how these past best-sellers perform on Amazon today. Examining whether their historical success translates into high ratings, sales ranks, and reviews on Amazon. We will also identify books that have few sales on Amazon so there's an opportunity to optimize their product listings to revive their online presence.

2. Data

The first source of data is the 'Best Selling Books' dataset from Kaggle, which contains information on the best-selling books from the past. It includes the book, author(s), original language, first published, and approximate of overall sales. This data set provides a foundation for best-selling books from the past, so we have a baseline to compare how books perform online.

The second source of data will come from scraping Amazon's website. I plan to gather information on average ratings, number of reviews, and sales rank. Gathering this information will allow me to evaluate how well a book performs in the digital marketplace.

Each dataset will require some data cleaning. For best selling books, I will start by making my column names snake case and then I will delete the "genre" column because many of our data points are missing that column. For my scraped dataset, I will also start by making my column names in snake case. From there I will need to clean amazon_rating, amazon_sales_rank and amazon_total_reviews by removing any extra words in the data. An example of this is our amazon_rating was scraped as "4.4 out of 5" when we would really only want the "4.4".

As discussed with Mr. Colbert, I manually added each link to both datasets to crawl between pages. The book did not end up being a good primary key because the Amazon listing names did not match that of our Best-Selling Books dataset. Instead, I got the Amazon link for every book and added it to both datasets manually. This worked as the way to crawl between pages and became the primary key to merge the two datasets.

Data Dictionary:

Column	Type	Source	Description
book	Text	Best Selling Books	Title of the book
author(s)	Text	Best Selling Books	The author(s) of the book

original_language	Text	Best Selling Books	The language that the books were originally written
first_published	Numeric	Best Selling Books	The year the book was published
approximate_sales_in_millions	Numeric	Best Selling Books	The approximate lifetime sales of the book
amazon_rating	Numeric	Amazon	The average rating out of 5 stars
amazon_sales_rank	Numeric	Amazon	How well the book sells on Amazon in reference to others in the category (1 is best)
amazon_total_reviews	Numeric	Amazon	The total numbers of reviews for the book
amazon_link	URL	Both	URL for each books Amazon listing

3. Analysis

Using seaborn I was able to create some graphics to help get more insight into our merged dataset.

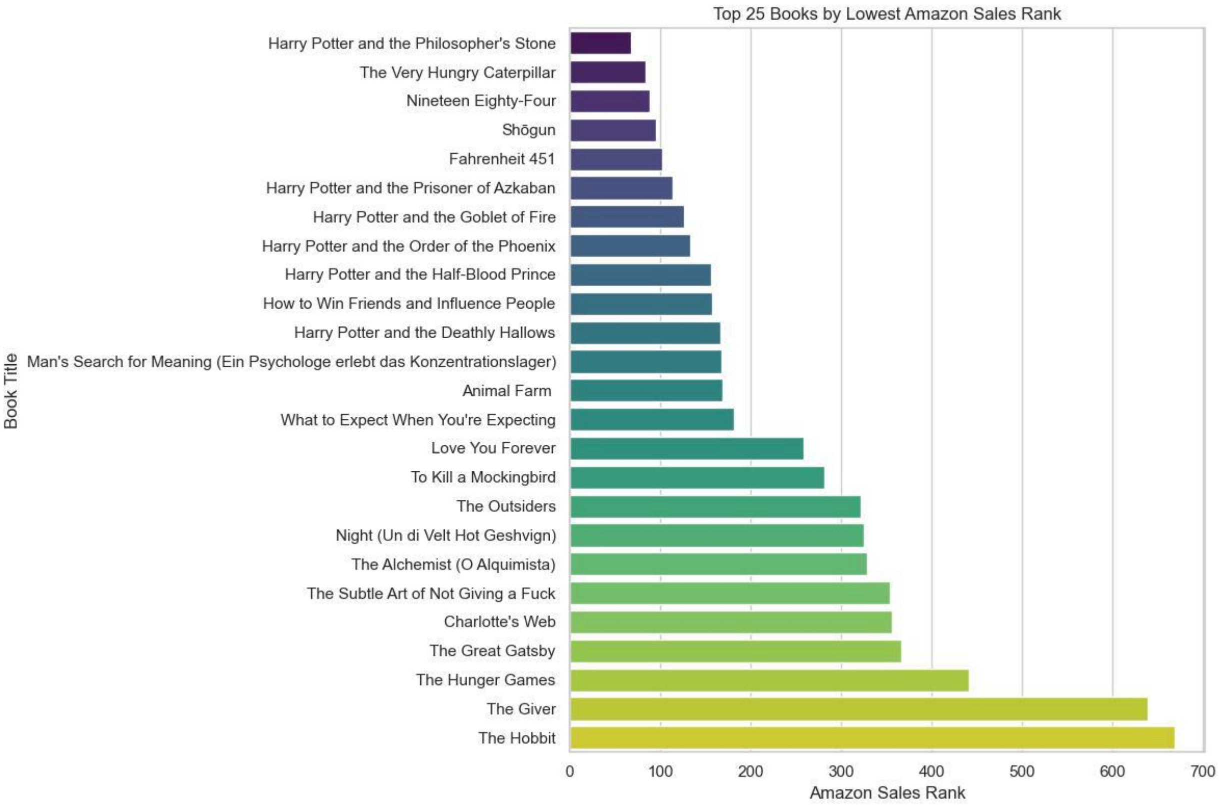

Figure 1 above displays the 25 books with the lowest sales ranks, which represent the best-selling books on Amazon. It's important to note that a lower sales rank indicates better performance, with a rank of 1 signifying the top-selling book on the platform.

Notably, multiple Harry Potter books appear on this list, reinforcing their position as some of the best-selling titles.

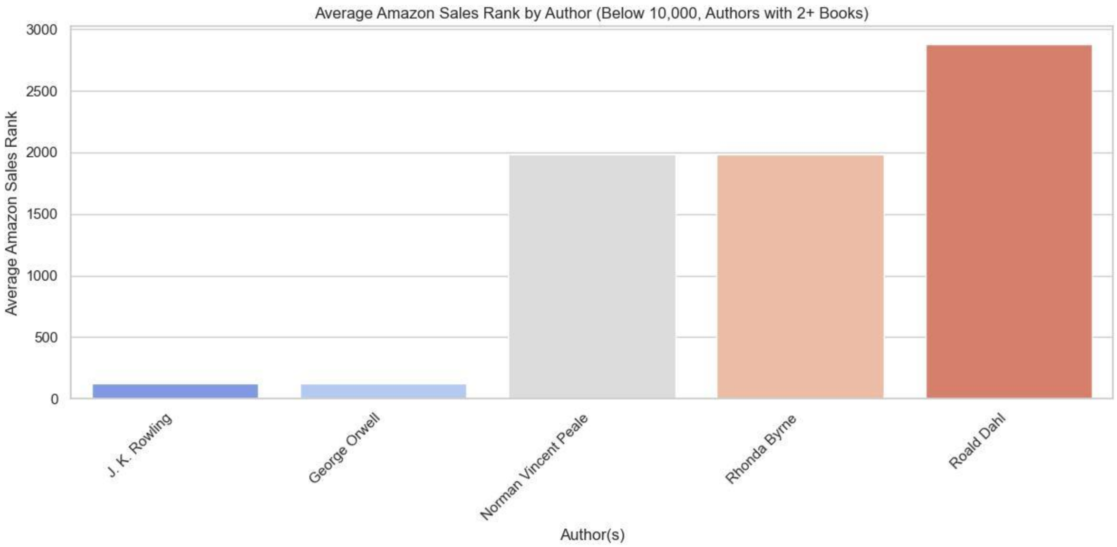

Figure 2 above shows the top 5 best-selling authors with multiple books by the lowest sales rank. Similar to the graph above, we know Harry Potter has sold historically well on Amazon since the author, JK Rowling has the lowest average sales rank. George Orwell is a very close second.

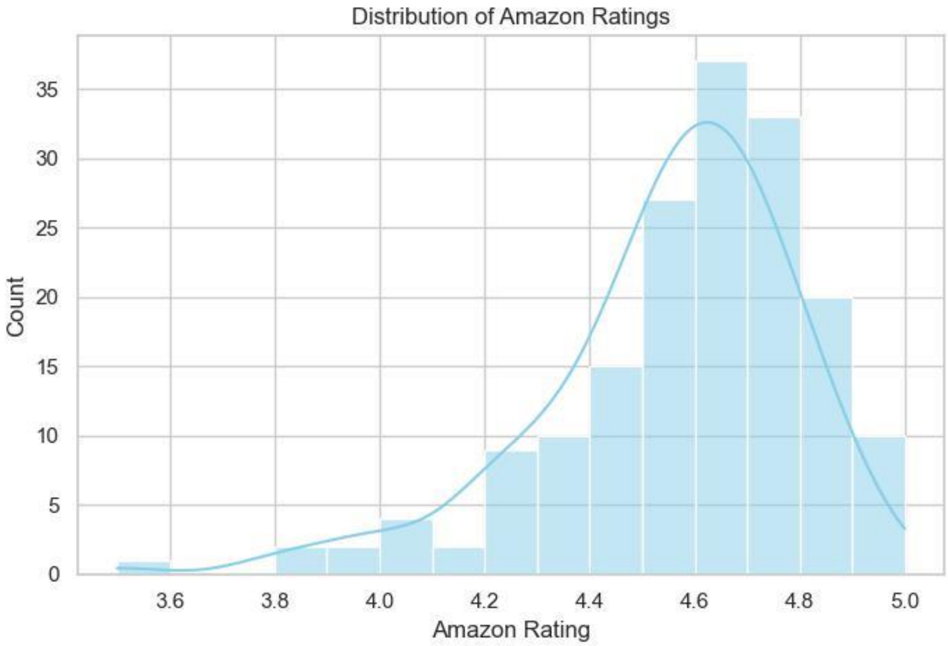

Figure 3 above shows the distribution of Amazon ratings on their platform. This helps give a baseline as to what's a good rating and what's not. Based on the graph we can conclude that about 4.55 is average with the median score being 4.7.

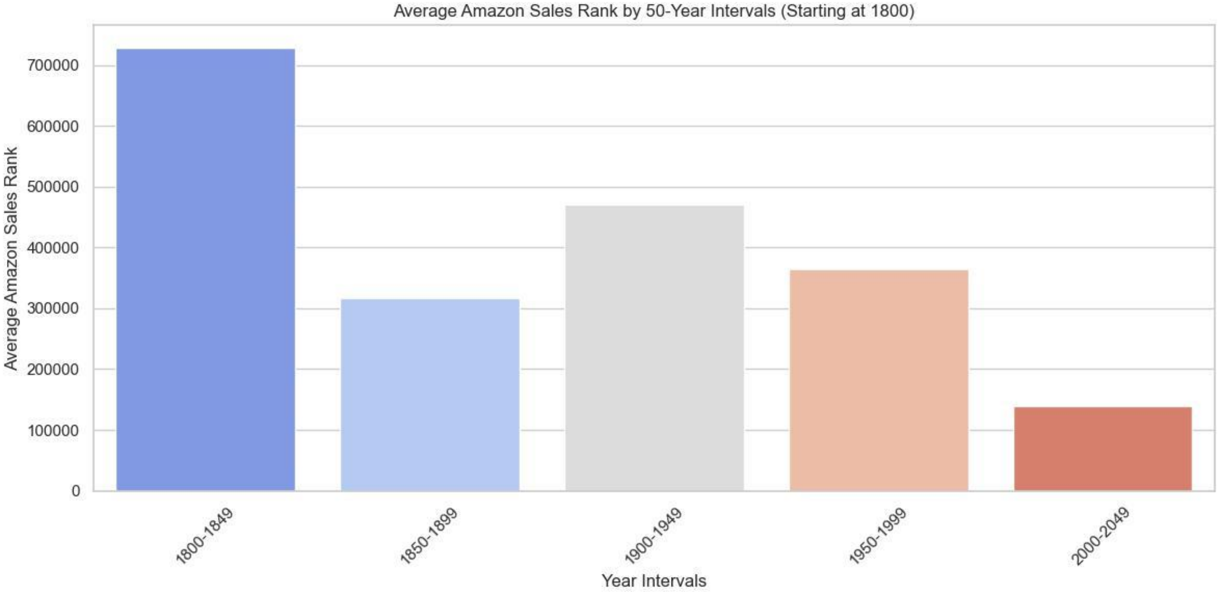

Figure 4 above shows the average sales ranks of books by period of time. It's clear that the more recent a book has been released, the better it will sell on Amazon. Books in the 1800-1849 period rarely sell anymore.

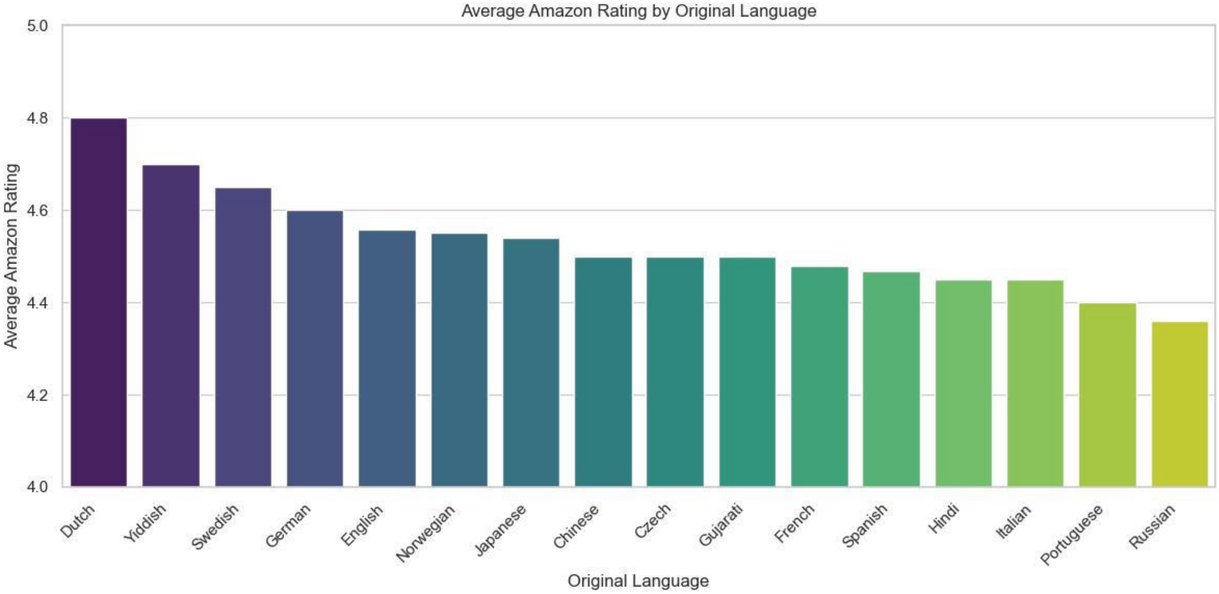

Figure 5 above shows the average Amazon rating by original language of publication. It shows that Dutch has the highest average rating and Russian has the least. English has the 5th highest average rating.

4. Conclusion

In this project, I was able to scrape 3 different things, amazon_rating, amazon_sales_rank, and amazon_total_reviews to get a good picture of how well past popular books have sold online. Looking back at some of the analysis questions from before the project, I found the following results.

- **How does the published language affect its sales rating on Amazon?**

Published language did have some impact on the sales rating with Dutch coming in at first and Russian coming in at last. The variance wasn't extremely high overall.

- **Do books from specific authors have consistently higher average ratings and sales ranks on Amazon compared to others?**

Yes, we found that our best-selling authors were J.K Rolling, George Orwell, and Norman Vincent Peale.

- **What is the sales rank for books across different publication decades, and does the trend show any patterns?**

Yes, books that were more recently published had a higher sales rank and overall sales. Books from 2000 and after had the best sales rank while books before 1850 had the worst sales rank.

This project has several limitations including the lack of understanding of some of our data. We can't be sure how our original dataset from Kaggle calculated the approximate sales in millions or what formula Amazon uses to calculate sales rank. It would be improved if we could get an exact number of sales through the Amazon platform for each book. To improve this project, we could data from other sites like Ebay and cross reference the approximate sales to other data sources to ensure that they're accurate.