Book Sales & Ratings

Project Category: Exploratory Data Analysis

Tools & Technologies Used: Python, NumPy, Pandas, Matplotlib, Seaborn, Jupyter Notebooks

1. Introduction

This data analysis project explores the sales, ratings, and publication trends in the book industry. Using a comprehensive dataset containing detailed information about books — including sales figures, author ratings, genres, and more — the project aims to extract insights that can help publishers, bookstores, and data enthusiasts understand market patterns, author performance, and evolving reader interests.

We aim to answer questions like:

- Which genres and authors generate the most revenue and reader engagement?
- How do pricing and author ratings influence a book's commercial success?
- What trends can be observed in publishing patterns and reader tastes over the years? By exploring these questions, this project offers actionable insights for stakeholders in the book industry—helping them make smarter decisions on publishing, promotion, and stocking strategies.

2. Project Overview

Dataset Description:

Books_Data_Clean.csv

- **Source**: Provided CSV file
- **Size**: 1070 rows × 15 columns
- **Features Include**:

- Publishing Year
- Book Name
- Author
- Language Code
- Author Rating
- Book Average Rating
- Book Ratings Count
- Genre
- Gross Sales
- Publisher Revenue
- Sale Price
- Sales Rank
- Units Sold
- Publisher

Preprocessing Steps:

- Dropped rows with missing Publishing Year and corrected values < 0.
- Cleaned column names and categorical inconsistencies (e.g., merged duplicate genres, standardized publisher name field).
- Removed rows with missing values in key columns.
- Outliers removed using the IQR method for: Book Ratings Count, Gross Sales, Publisher Revenue, and Sale Price.

3. Data Exploration

Key Steps:

- Identified 3 columns with missing data: Publishing Year (1), Book Name (23), Language Code (53).

- No duplicate rows were found.
- Summary statistics showed skewed distributions in many numerical features (e.g., Sales, Ratings Count).
- Removed 22 entries with "Unknown Title."
- Cleaned and grouped genres (e.g., merged "genre-fiction" and "fiction").
- Cleaned language codes (e.g., merged en-US, en-GB, etc. into 'eng').
- Re-categorized Author Ratings into ordinal groups.

4. Analysis Methods

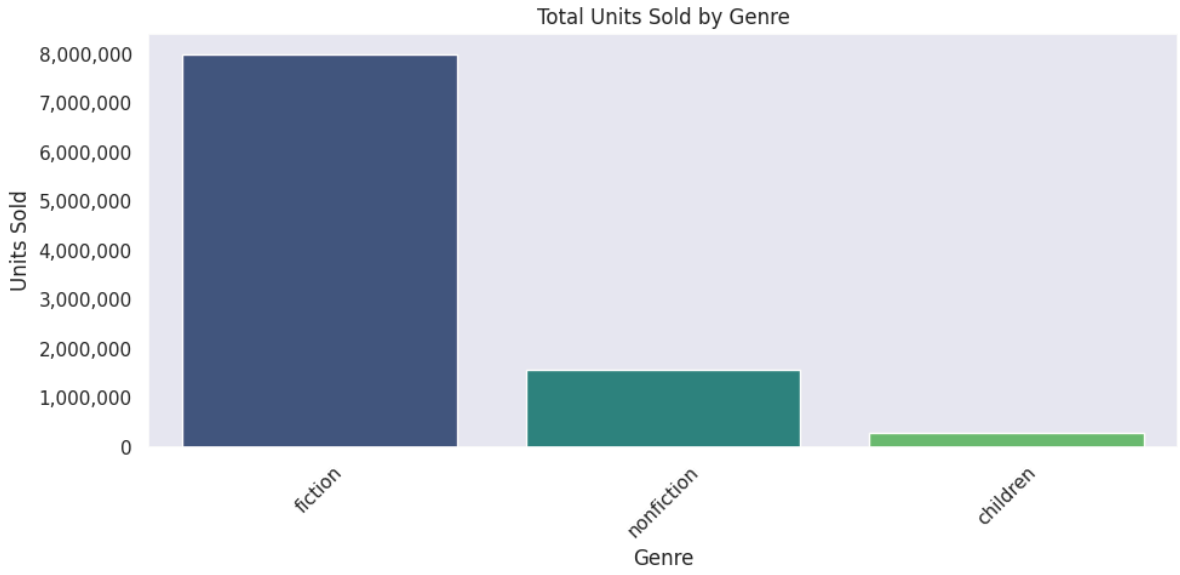
The analysis involved:

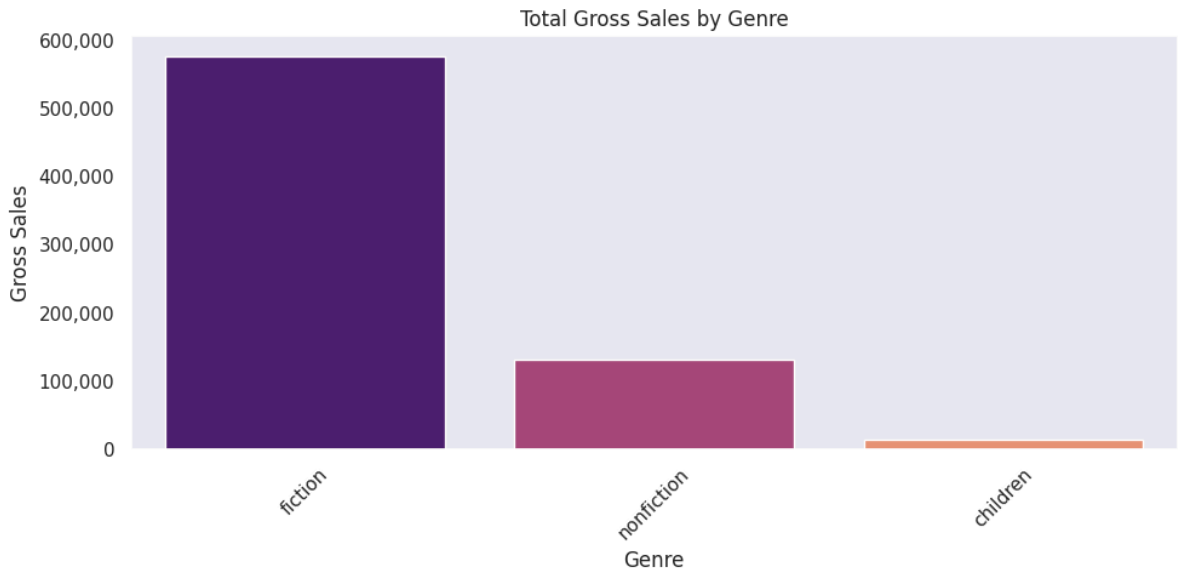
- **Univariate Analysis**: Distribution analysis of individual features.
- **Bivariate & Multivariate Analysis**: Relationships between variables such as Author Rating vs Book Rating, Sales vs Units Sold, etc.
- **Aggregation Techniques**: Grouping by genres, publishers, authors to compute means, sums, and ranks.
- **Visualization**: Used `seaborn` and `matplotlib` to create bar charts, histograms, scatter plots, and line graphs.

5. Results and Interpretation

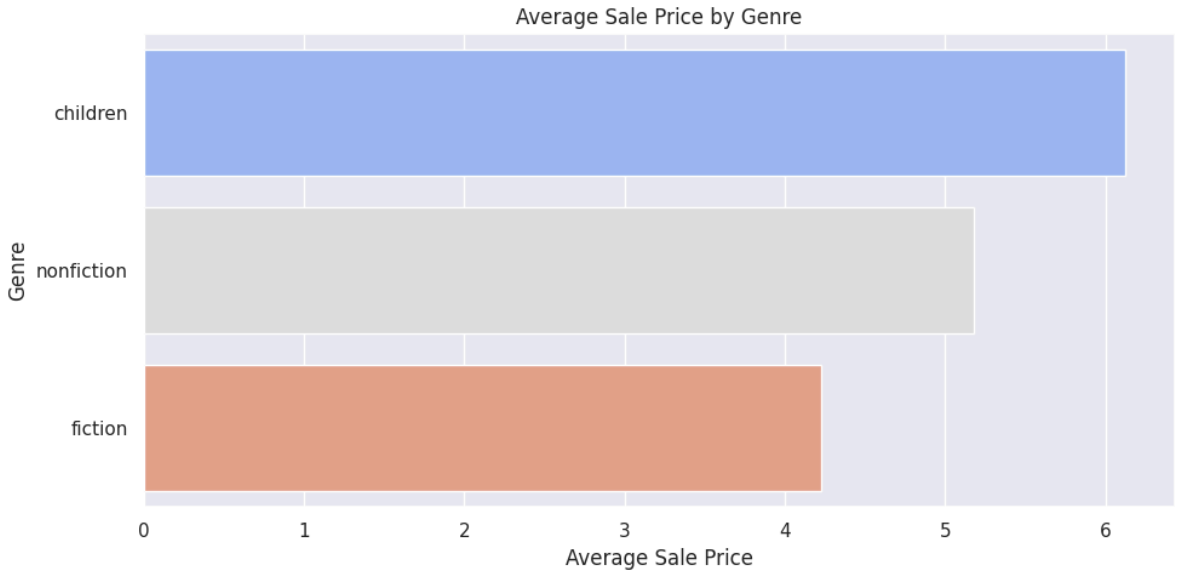
Market Analysis:

- **Fiction** genre sells the most units and earns the most revenue.

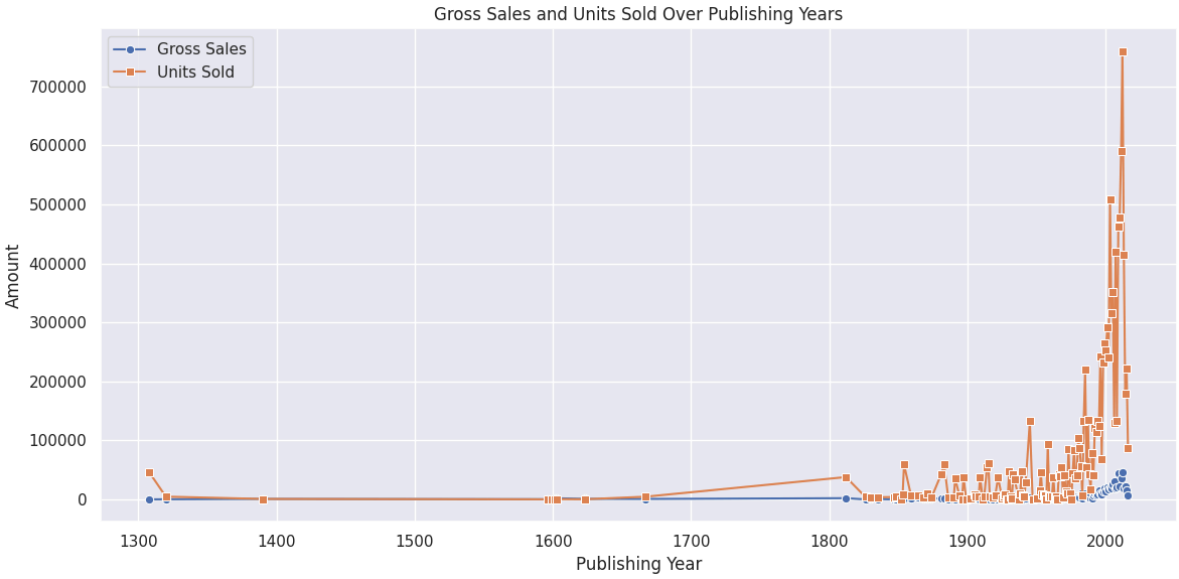




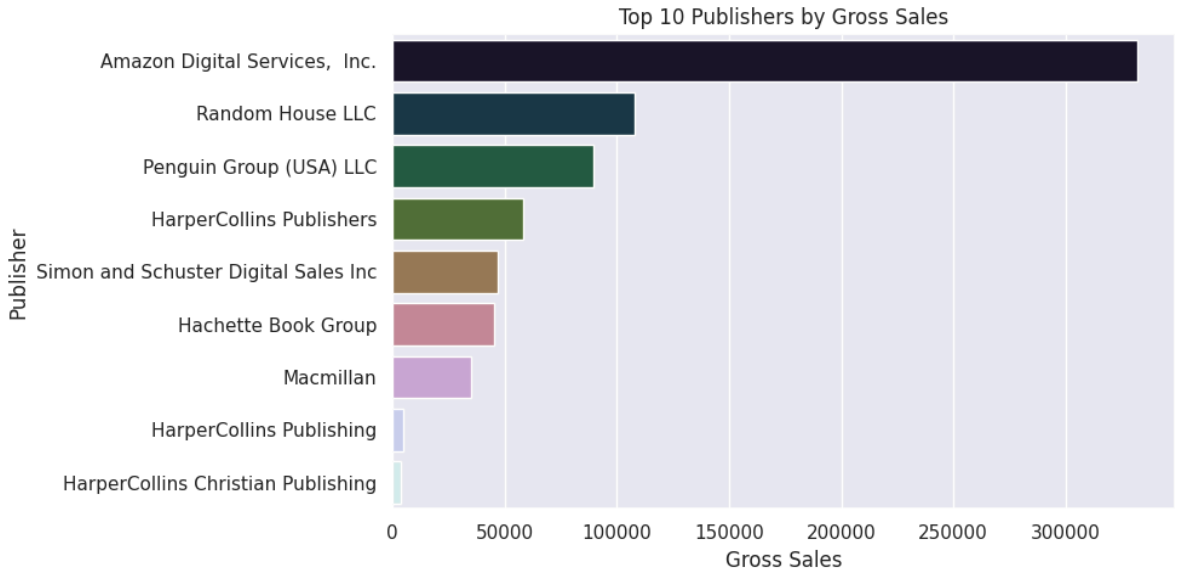
- **Children's** books have the highest average sale prices.



- **Newer** books (post-2000) sell more units but may not always result in higher gross sales.



- Top publishers by revenue are also those selling the most units.

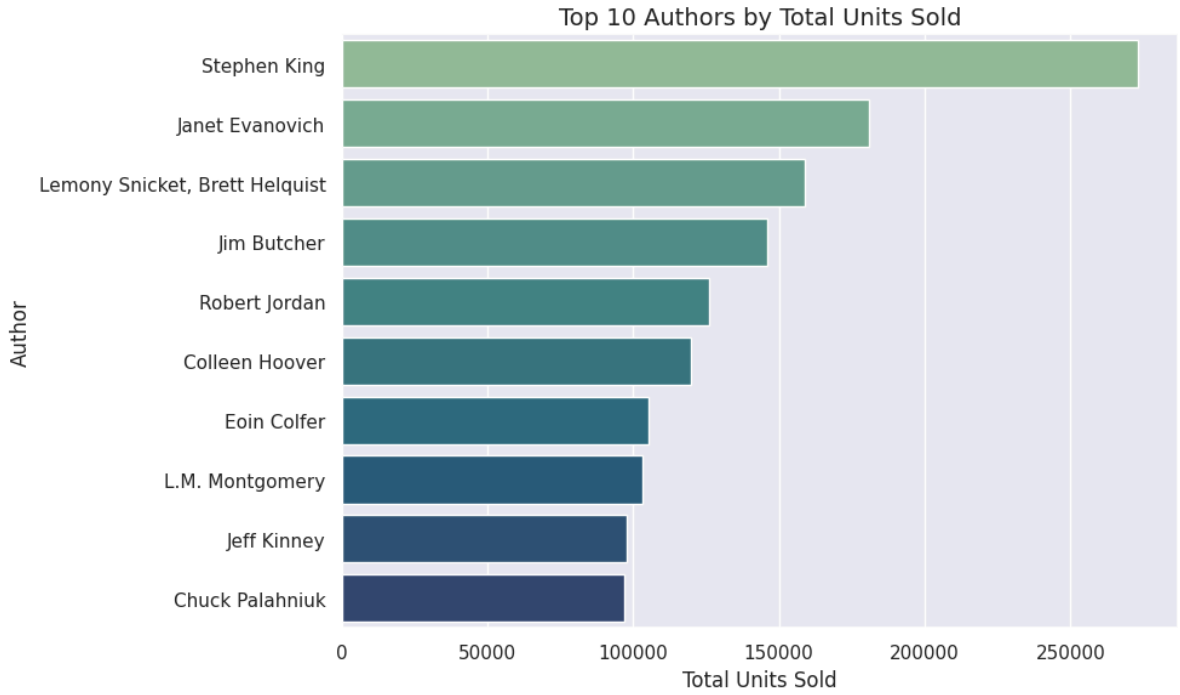


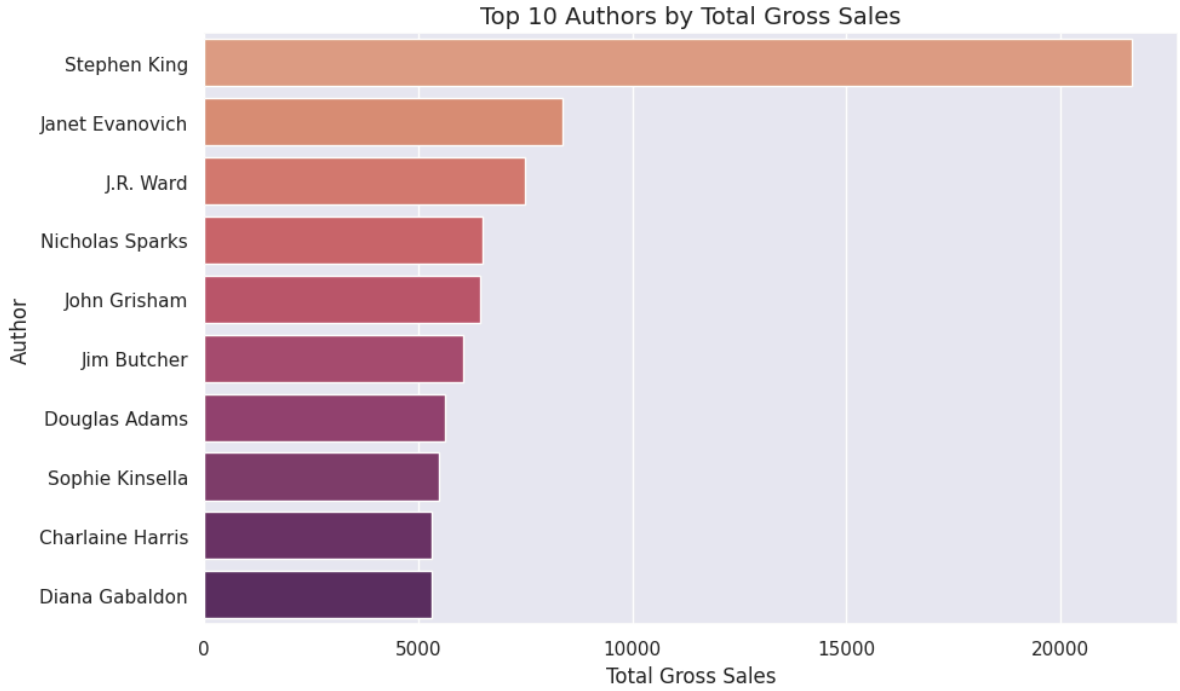
- Top 3 titles per genre helped identify standout performers.

Author Performance Evaluation:

- Higher Author Ratings tend to align with higher Book Average Ratings.

- Not all popular books (based on ratings) result in high sales — indicating a possible gap between critical acclaim and commercial success.
- Some authors consistently rank high across multiple metrics: units sold, revenue, ratings.



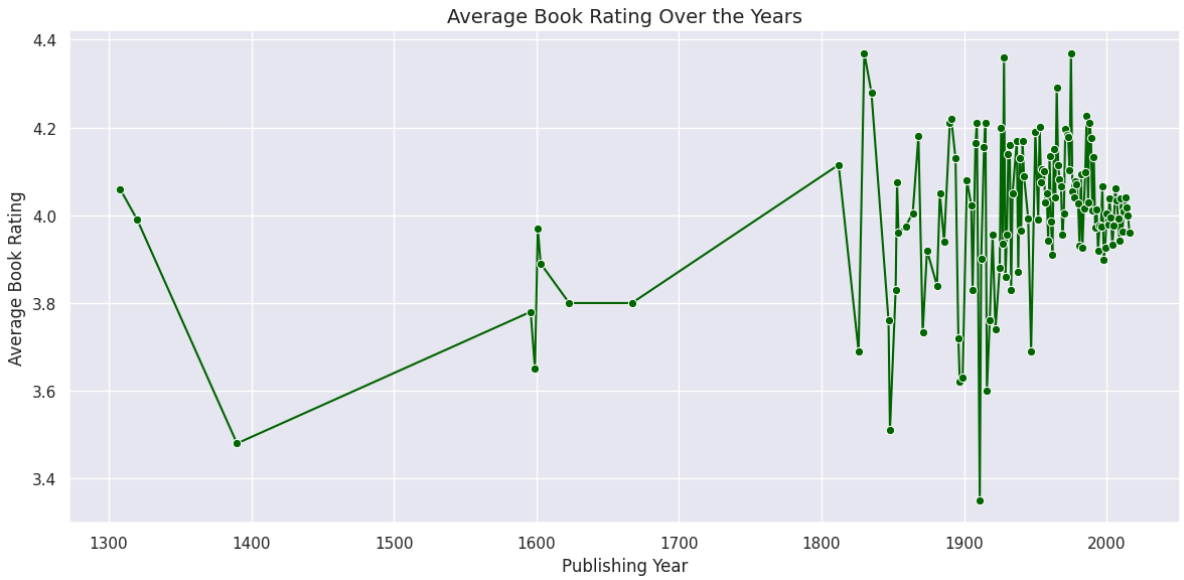


- Authors with higher average sale prices may target niche audiences.

Trend Analysis:

- Surge in book publications post-2000 with peak around 2010-2012.

- Fiction genre has been a dominant force consistently.
- Reader ratings remained relatively stable over time (~4.0 average).



6. Conclusion

This analysis of the Books dataset provided comprehensive insights into:

- Genre and publisher market performance
- Author influence on sales and ratings
- Evolving trends in book publishing and sales

These findings can help guide publishers, marketers, and readers in making data-driven decisions.

7. Recommendations

- **Focus on Fiction**: Since fiction drives the highest units sold and gross revenue, publishers should continue investing heavily in this genre.
- **Targeted Pricing Strategy**: Children's books command the highest average prices — optimizing price elasticity across genres could improve profitability.
- **Leverage Author Popularity**: Promote authors with strong average sale prices or consistent unit sales across titles.

- **Data-Driven Inventory Management**: Use genre, price, and sales rank insights for better stock planning and promotions.
- **Modernize Publishing Portfolios**: Given the increase in post-2000 book success, newer publications should remain a strategic focus.

8. Limitations and Future Work

Limitations:

- Dataset may not cover all global markets equally.
- Some values (e.g., Publishing Year < 0) indicated possible data entry issues.
- Zero entries in revenue fields suggest missing or incomplete financial tracking.

Future Work:

- Deeper NLP analysis of book titles for sentiment/category alignment.
- Time series forecasting of sales trends.
- Modeling to predict best-seller potential using combined features.

9. References and Citations

- Dataset Source: **Download Dataset**
- Python Libraries Used: `pandas`, `numpy`, `matplotlib`, `seaborn`
- Kaggle - https://www.kaggle.com/code/faresabbasai2022/books-sales-and-ratings-eda

10. Code Documentation

- All analysis steps were conducted in Python Jupyter Notebook.
- Each cell is documented with markdown explanations for clarity.
- Instructions:
  - Required: Python ≥3.7, Jupyter Notebook
  - Install required packages with: `pip install pandas numpy matplotlib seaborn`

11. Materials & Resources

IPYNB Notebook: https://colab.research.google.com/drive/1qts71CzLuJlalFSmwJ3BOb3TB7cHmVAx?usp=sharing

GitHub Link: https://github.com/Rudrajit12/Book-Sales-Ratings

Credits:

Author: Rudrajit Bhattacharyya

Email ID: rudrajitb24@gmail.com

LinkedIn: https://www.linkedin.com/in/rudrajitb/

GitHub: https://github.com/Rudrajit12