Amazon Top 50 Bestselling Books (2009–2019) Analysis

Yuvika Shendge

2024-12-10

Contents

1 Abstract

This report explores a dataset of Amazon's Top 50 Bestselling Books from 2009 to 2019. By analyzing trends in genres, pricing, ratings, and reviews, this study identifies factors contributing to the popularity of books. Key findings include the dominance of Fiction, significant relationships between user ratings and review counts, and the negligible impact of pricing on user engagement. Future work includes extending the dataset and integrating Goodreads ratings for deeper insights.

2 Introduction

Over the past decade, online platforms like Amazon have reshaped the book industry. By examining Amazon's Top 50 Bestselling Books dataset (Sootersaalu 2020), this report uncovers insights into reader preferences and market trends. The primary objectives include:

1. Identifying trends in book genres over time.
2. Analyzing the relationships between pricing, ratings, and reviews.
3. Using statistical and clustering methods to understand factors that influence book popularity.

This analysis provides actionable insights for authors, publishers, and marketers aiming to optimize strategies for reaching readers utilizing a dataset of information from the largest book retailer globally (Girolino 2023).

3 Data

The dataset, sourced from Kaggle (link), contains 550 entries spanning a decade (2009–2019). Key variables include **Name**: Title of the book, **Author**: Author of the book, **Genre**: Fiction or Non-Fiction (as categorized using Goodreads metadata), **User Rating**: Average rating (1–5 stars), **Reviews**: Total number of reviews on Amazon, **Price**: Price in USD, **Year**: Year of bestseller status.

3.0.1 Preprocessing

- **Genre Categorization**: Fiction and Non-Fiction tags were added later based from Goodreads.
- **Data Cleaning**: Checked for duplicates, formatted numeric columns, and ensured consistent entries.
- **Derived Variables**: Added categories for deeper segmentation (e.g., price ranges).

3.0.2 Limitations

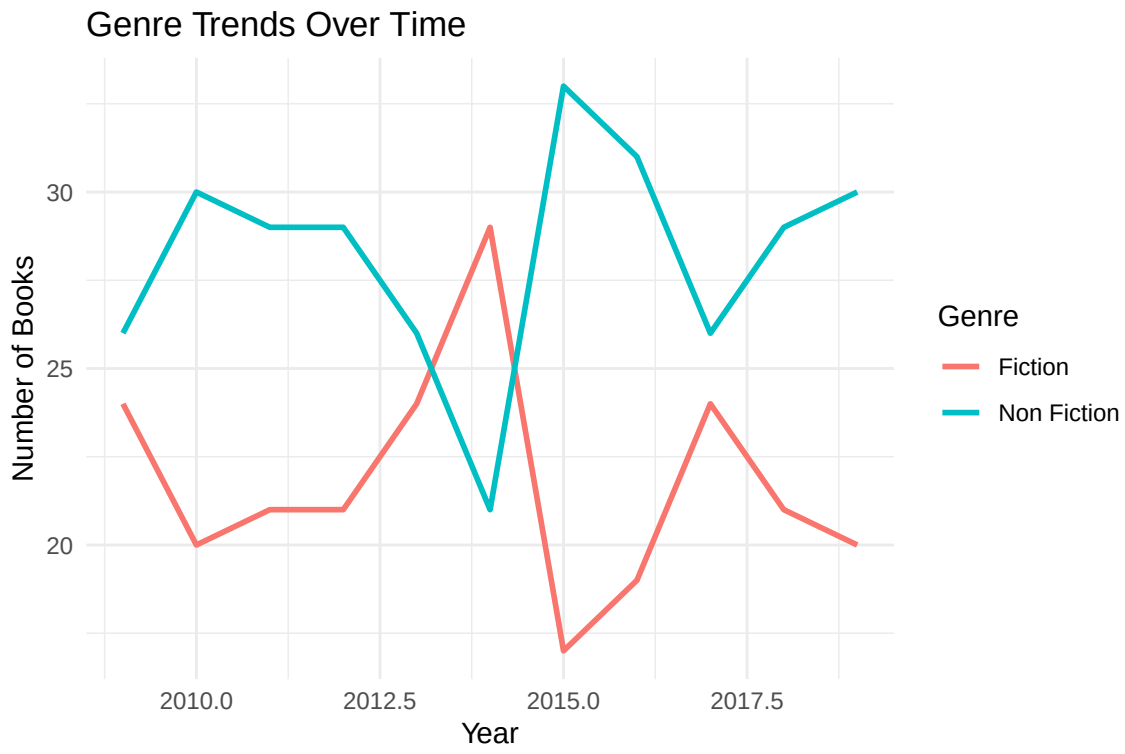
- Excludes regional and global sales data.
- Lacks Goodreads-specific reviews and user ratings.

4 Analysis and Results

4.1 Genre Trends Over Time

```
genre_trend <- data %>%
  group_by(Year, Genre) %>%
  summarize(Count = n())

ggplot(genre_trend, aes(x = Year, y = Count, color = Genre)) +
  geom_line(size = 1) +
  labs(title = "Genre Trends Over Time",
       x = "Year", y = "Number of Books",
       color = "Genre") +
  theme_minimal()
```

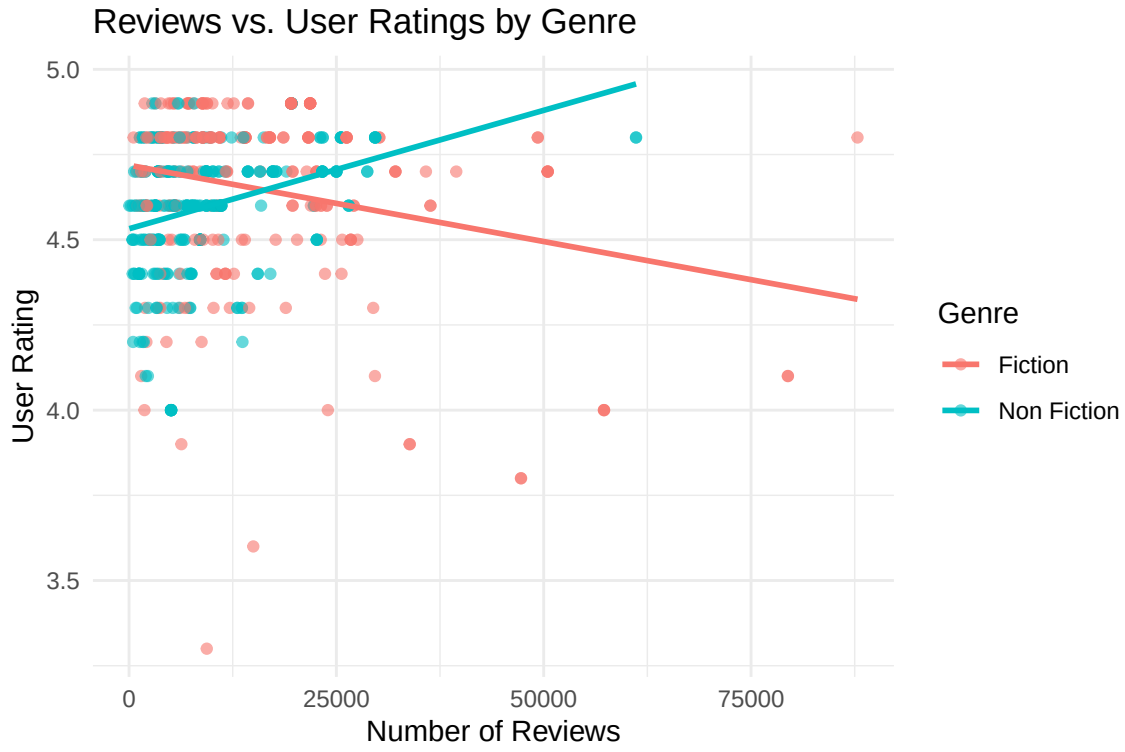


4.1.1 Observation

The plot illustrates fluctuating trends for both Fiction and Non-Fiction over time. While Non-Fiction consistently dominates from 2010 to early 2013, there is a sharp decline in Non-Fiction entries around later in 2013, coinciding with a significant increase in Fiction. Post-2013, Non-Fiction regains prominence and maintains higher counts through 2019. This dynamic shift may indicate changing consumer preferences influenced by external factors such as market trends or significant cultural events during those periods.

4.2 Relationship Between Reviews and User Ratings

```
ggplot(data, aes(x = Reviews, y = User.Rating, color = Genre)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Reviews vs. User Ratings by Genre",
       x = "Number of Reviews", y = "User Rating") +
  theme_minimal()
```

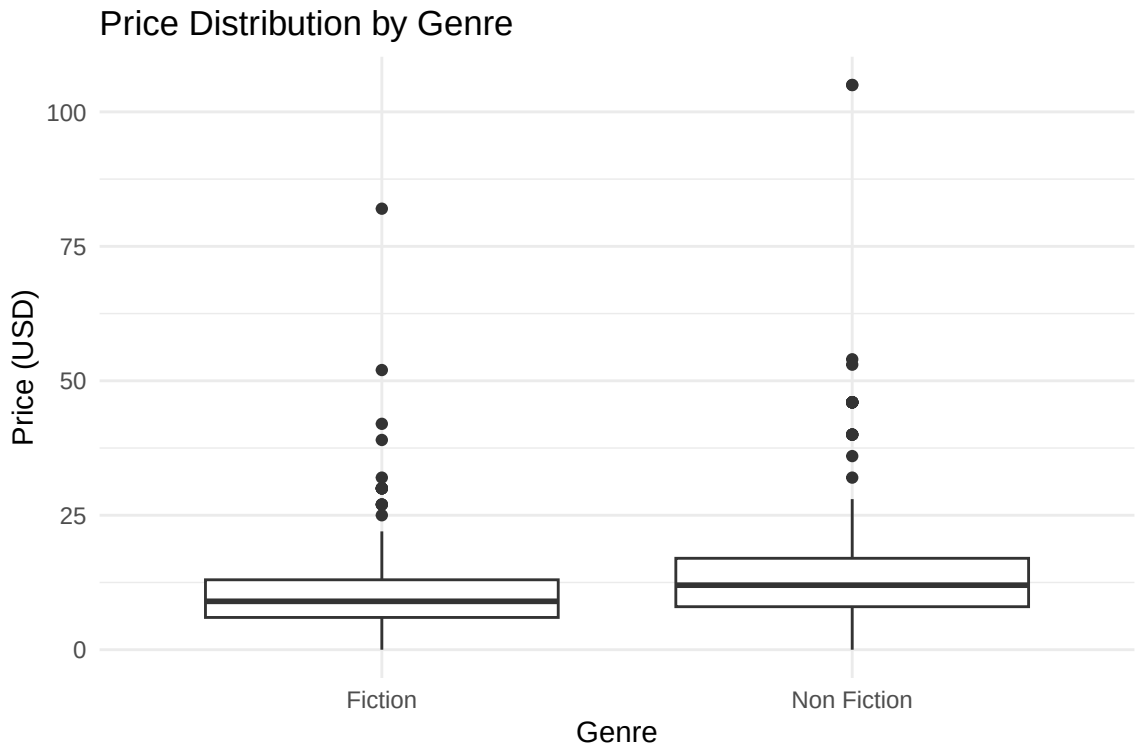


4.2.1 Observation

The scatter plot reveals contrasting trends between Fiction and Non-Fiction in terms of the relationship between reviews and user ratings. For Non-Fiction, there is a positive trend, where books with a higher number of reviews generally have higher user ratings. In contrast, the regression line for Fiction trends slightly downward, indicating that as the number of reviews increases, the user rating slightly decreases. This could suggest that highly reviewed Fiction books receive more critical scrutiny, or it may reflect the broader appeal of Fiction attracting a wider and more diverse audience with varying opinions.

4.3 Price Analysis

```
ggplot(data, aes(x = Genre, y = Price)) +
  geom_boxplot() +
  labs(title = "Price Distribution by Genre",
       x = "Genre", y = "Price (USD)") +
  theme_minimal()
```



4.3.1 Observation

The boxplot illustrates that Fiction books are priced within a narrower and lower range, with a median price significantly below that of Non-Fiction. Non-Fiction books exhibit a broader price range, including several outliers on the higher end, reflecting a more premium pricing strategy. This disparity may be attributed to the specialized or professional nature of Non-Fiction content, as opposed to Fiction's more mass-market appeal.

4.4 Clustering Analysis

```
clustering_data <- data %>%
  select(User.Rating, Reviews, Price) %>%
  scale()

kmeans_result <- kmeans(clustering_data, centers = 3)
```

```r
data$Cluster <- as.factor(kmeans_result$cluster)

ggplot(data, aes(x = Reviews, y = Price, color = Cluster)) +
  geom_point(alpha = 0.7) +
  labs(title = "Clustering of Books by Reviews and Price",
       x = "Number of Reviews", y = "Price (USD)") +
  theme_minimal()
```

4.4.1 Observation

The clustering analysis identifies three distinct segments of books:

1. **Books with medium to high reviews and low prices**
   These books are likely aimed at mass-market audiences who prioritize affordability and widely appealing content. Their high review counts suggest strong reader engagement and popularity.

2. **Books with low reviews but high prices**
   This segment likely represents specialized or high-value content, often targeting niche audiences. The higher price points may reflect premium positioning or limited availability, catering to readers with specific preferences or professional interests.

3. **Books with low reviews and low prices**
   These books may appeal to cost-conscious readers or serve niche markets with limited demand. Their low pricing strategy could aim to boost accessibility despite limited visibility or popularity.

The clustering highlights distinct audience demographics and pricing strategies within the best-seller list, showcasing how books are tailored to diverse consumer preferences. This segmentation underscores the interplay between reviews, pricing, and reader demographics in shaping the book market's dynamics.

5 Conclusion

The analysis of Amazon's Top 50 Bestselling Books from 2009–2019 provides several insights into consumer behavior and trends in the book market:

1. **Genre Trends Over Time:**
   Fiction and Non-Fiction genres exhibit fluctuating popularity throughout the decade. The shift in dominance, especially the decline of Non-Fiction around 2013 followed by its resurgence, highlights how external factors such as cultural shifts or market trends can influence consumer preferences. Fiction's steady appeal reflects its broader mass-market reach, while Non-Fiction's varying trends may stem from its specialized and time-sensitive content.

2. **Reviews and User Ratings:**
   The relationship between reviews and user ratings varies significantly by genre. Non-Fiction books with higher review counts tend to achieve better ratings, suggesting strong reader approval. Conversely, Fiction books with more reviews show slightly lower average ratings, potentially indicating more critical scrutiny due to a broader and more diverse audience.

3. **Price Dynamics:**
   Pricing strategies differ markedly between genres. Fiction books are generally priced lower, appealing to a wide audience with accessible pricing. In contrast, Non-Fiction books command a higher median price with greater variability, likely due to their specialized content and premium positioning.

4. **Clustering Insights:**
   The clustering analysis uncovers three distinct book segments based on reviews, ratings, and pricing:

   - High reviews, low prices: Targeting mass-market readers.
   - Moderate reviews, mid-range prices: Serving niche markets.
   - Low reviews, high prices: Catering to specialized or high-value audiences.
     These segments highlight the diversity in book offerings and their alignment with distinct reader demographics and pricing strategies.

Future Work:
In the future it would be nice to include data from Goodreads or regional sales figures beyond just Amazon could provide deeper insights into global reader preferences. Additionally, exploring changes in sub-genres or author-specific trends could also further offer valuable perspectives.

6 Citations

Girolino. 2023. "The Complete Amazon Company History Timeline: A Guide to the e-Commerce Giant's Journey." https://www.girolino.com/the-complete-amazon-company-history-timeline-a-guide-to-the-e-commerce-giants-journey/.

Sootersaalu. 2020. "Amazon Top 50 Bestselling Books 2009–2019." https://www.kaggle.com/datasets/sootersaalu/amazon-top-50-bestselling-books-2009-2019?resource=download.