Case Study

Amazon Top 50 Best Selling Books (2009–2019)

**Step 1: Ask**

**Business Task:**  
Identify trends and characteristics of best-selling books on Amazon between 2009 and 2019. What factors (genre, author, price, etc.) appear to influence a book's success?

**Key Questions:**

* Are fiction or nonfiction books more likely to be bestsellers?
* Which authors consistently appear among top sellers?
* Is there a relationship between price and popularity?
* How have trends changed over time?

**Step 2: Prepare**

**Dataset Info:**

* Source: Souter Saalu on Kaggle
* License: CC0 (open, free to use)
* Columns (typically include):
  + Name
  + Author
  + User Rating
  + Reviews
  + Price
  + Year
  + Genre
  + Rank

**Data Integrity Checks:**

* Check for missing values or duplicates
* Ensure Price is numeric
* Confirm Year ranges from 2009 to 2019
* **Duplicates**: There are **no duplicate rows** in the dataset.
* **Missing Values**: There are **no missing values** in any column.
* **Price Type**: The Price column is of type **integer** (int64), which is suitable for numerical analysis.
* **Year Range**: All years fall within the expected range of **2009 to 2019**.

**Step 3: Process**

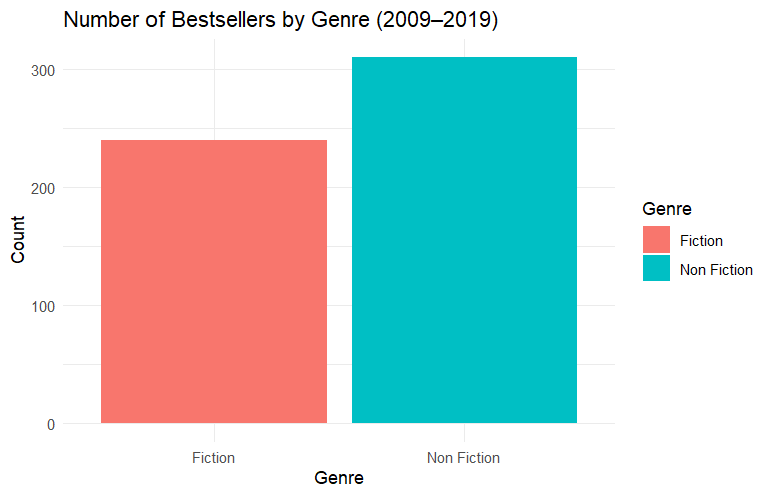
Clean the data:

* Convert Price, Reviews, and User Rating columns to numeric if necessary
* Drop duplicate entries
* Standardize text (e.g., lowercase author names)
* Create any new variables (e.g., fiction vs. nonfiction count per year)

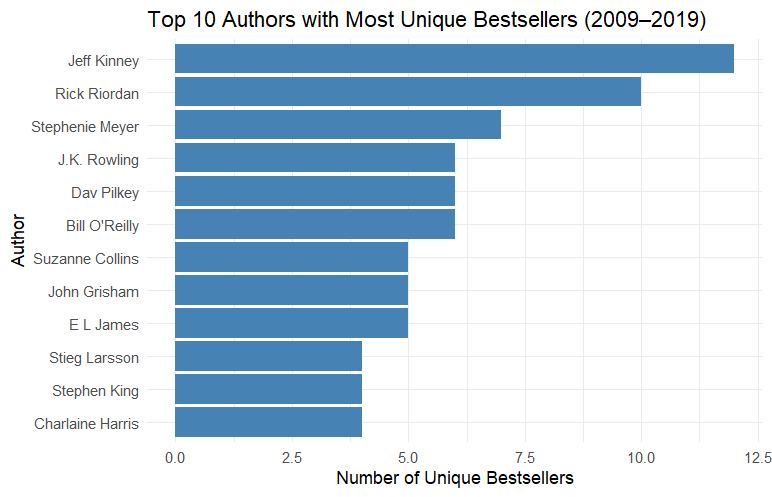
**Step 4: Analyze**

R to explore:

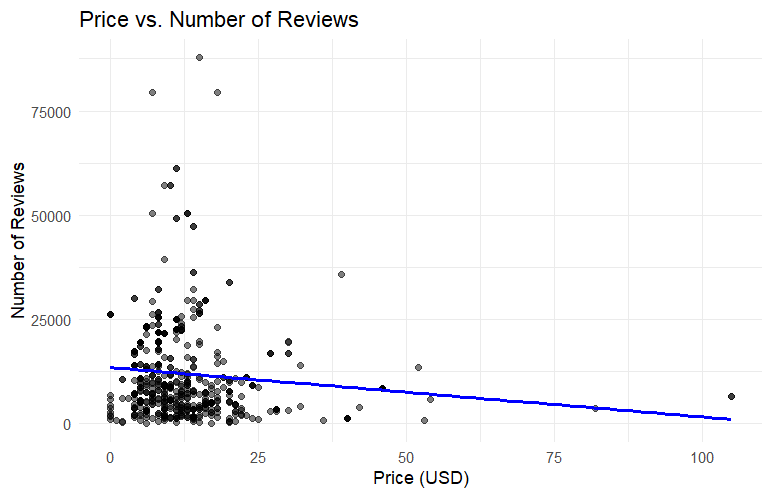
* Are fiction or nonfiction books more likely to be bestsellers?

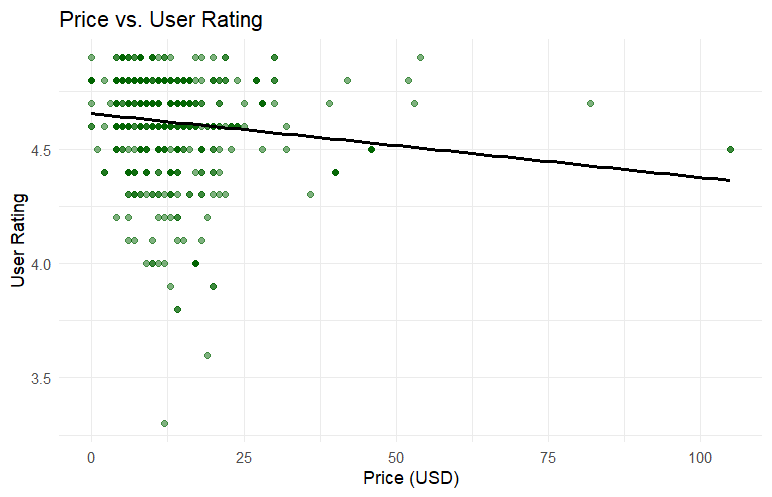


* Which authors consistently appear among top sellers?

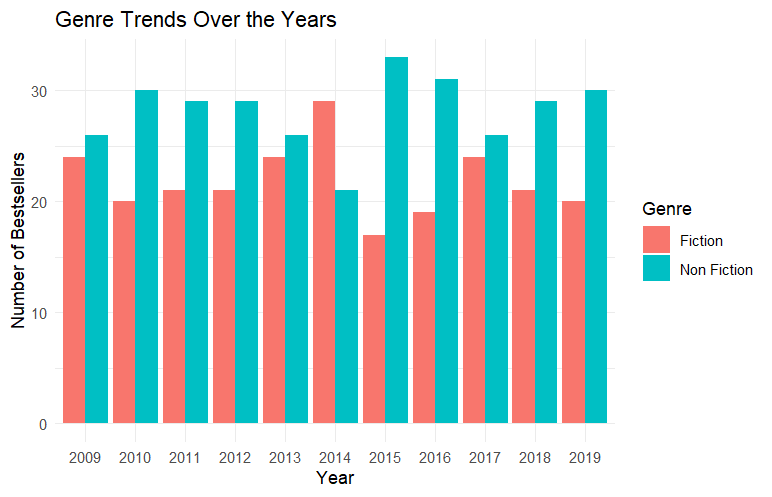


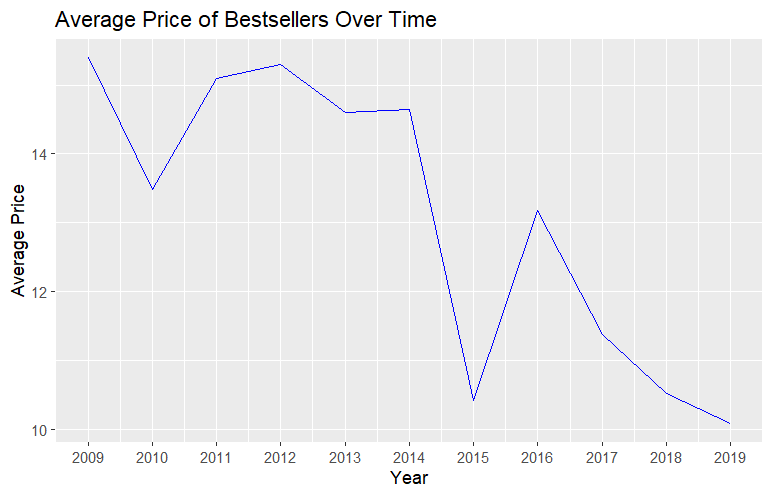
* Is there a relationship between price and popularity?

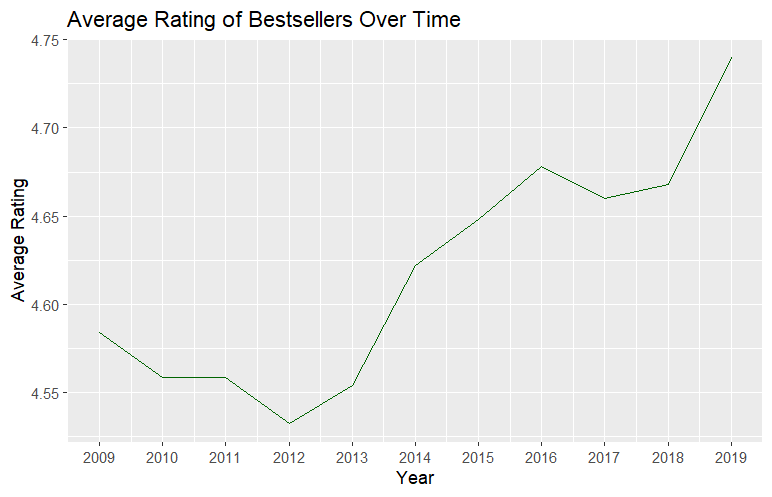


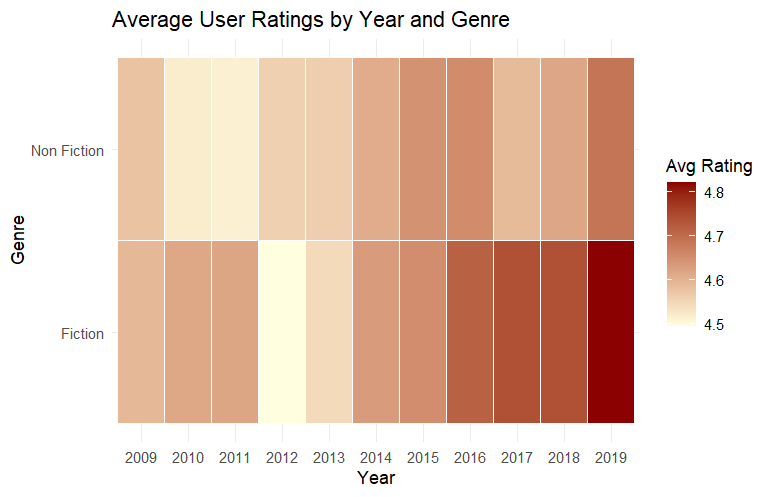


* How have trends changed over time?









**Step 5: Share**

**Key Visualizations to Include:**

* Bar chart of fiction vs. nonfiction over years
* Scatter plot of price vs. rank or rating
* Top authors (bar graph)
* Heatmap of average user ratings by year and genre

**Summary for Stakeholders:**

**Key Findings:**

* **Fiction dominates** the bestseller list overall, appearing more frequently than nonfiction across the years.
* **Consistent top authors** include Jeff Kinney, Rick Riordan, and Stephenie Meyer—each with multiple entries across several years, indicating strong reader loyalty and ongoing demand for their work.
* **Price is weakly correlated with popularity.**
  + There is a **small negative correlation** between price and the number of reviews (approx. -0.11), suggesting that higher prices might slightly discourage engagement.
  + A similarly weak negative correlation (~ -0.13) exists between price and user ratings, indicating that cheaper books do not necessarily receive lower ratings.
* **Genre trends over time**. While fiction remained a strong category, **nonfiction books consistently dominated** the bestseller lists across most years, especially during **2016–2018**, when nonfiction saw notable surges.
* **Average book prices** have stayed relatively stable, within the **$10–$16 range**, with no significant upward or downward trend observed.

While **average user ratings** remained high throughout the period, averaging above 4.5 out of 5.

* A **heatmap of user ratings by genre and year** confirms that fiction books consistently receive slightly higher ratings compared to nonfiction, though both genres perform well.

**Step 6: Act**

**Potential Recommendations:**

* Publishers may focus more on producing accessible nonfiction titles.
* Maintain Competitive Pricing Between $10–$16
* Highlight Proven Authors