Exploratory Data Analysis (EDA) Report

Project: Book Reviews Analysis

Task 2 Duration:
Objective: Explore, clean, and analyze the scraped book dataset to uncover patterns, trends, and anomalies that inform visualization and sentiment analysis tasks.

---

1. Dataset Overview

- **Total records:** 100 books
- **Initial features:** 4 columns
  - `title`
  - `price`
  - `rating`
  - `availability`

Initial Observations

- No missing values were found in any column.
- All columns were initially stored as `object` data types.
- `availability` had only one unique value (`In stock`), offering no analytical value.
- `price` contained currency symbols and encoding issues.
- `rating` was represented as text values (One–Five).

---

2. Data Cleaning Summary

The following cleaning steps were applied:

- Removed currency symbols and encoding artifacts from `price` and converted it to `float`.
- Mapped textual ratings (`One`–`Five`) to numeric values (1–5).
- Dropped the `availability` column due to lack of variance.
- Checked for duplicates (none found).
- Detected outliers using the IQR method (no extreme outliers identified).

Final Cleaned Dataset

- **Rows:** 100
- **Columns:** 3 (`title`, `price`, `rating`)
- Dataset saved as: `data/eda-cleaned_books.csv`

---

3. Statistical Analysis & Distributions

Price Statistics

- **Mean:** £34.56
- **Median:** £34.78
- **Standard Deviation:** £14.64
- **Range:** £10.16 – £58.11

Insight: Mean and median prices are nearly identical, indicating a fairly symmetric price distribution with moderate variability.

Rating Statistics

- **Mean:** 2.93
- **Median:** 3
- **Mode:** 1

Insight: Ratings are evenly distributed across the 1–5 scale, suggesting no strong positive or negative bias in reviews.

Rating Distribution

- Ratings are well balanced, with each rating level (1–5) appearing a similar number of times.

Price Distribution (Binned)

- Books are spread across low, mid, and high price ranges.
- The lowest (£10–£20) and highest (£48–£58) price bins contain the most books.

---

4. Trends, Patterns & Anomalies

Price vs Rating Relationship

- **Correlation coefficient:** -0.122

Insight: There is a very weak negative relationship between price and rating, indicating that higher-priced books do not necessarily receive higher ratings.

Average Price by Rating

- 5-star books have the **lowest average price**.
- 1–2 star books are among the **most expensive on average**.

Insight: Reader satisfaction does not increase with price; affordable books often receive higher ratings.

Anomaly Detection

- **Expensive & Low-Rated Books:** 11
  - Example: *Tipping the Velvet* (£53.74, rating 1)
- **Cheap & High-Rated Books:** 10
  - Example: *Sophie's World* (£15.94, rating 5)

Insight: These anomalies highlight potential overpricing and strong value-for-money books, respectively.

---

5. Key Insights & Conclusions

- The dataset is clean, balanced, and suitable for further analysis.
- Price is not a strong predictor of book rating.
- High-value books (low price, high rating) exist and are worth highlighting.
- Several overpriced, poorly rated books present interesting anomaly cases.

---

6. Next Steps

- **Task 3:** Create visualizations (histograms, boxplots, scatter plots) to visually communicate findings.
- **Task 4:** Apply sentiment analysis to review text and compare sentiment with numeric ratings.

End of Task 2 – Exploratory Data Analysis