PROJECT REPORT

PROJECT 1: Book Data Analysis

K D Liyanage

1. Introduction

This project performs a complete exploratory data analysis (EDA) on a dataset of books, including their prices, ratings, review counts, author information, and review activity over time. The goal is to understand patterns in reader behavior, book popularity, and factors associated with rating trends.

2. Data Collection & Loading

The dataset was imported into a Jupyter notebook environment.

Initial tasks included:

- Loading the book dataset from a CSV file.
- Inspecting column names and previewing the first few rows.
- Checking data types and general data structure.

This step ensured that the dataset was in a readable format and ready for cleaning.

3. Data Cleaning

Several preprocessing operations were performed to ensure the dataset was reliable:

- Handling **missing values** in columns such as price, rating, and publication year.
- Converting numeric columns to appropriate data types.
- Removing or correcting invalid entries.
- Standardizing formatting where necessary (e.g., stripping symbols or fixing string representations).

This stage prepared the dataset for accurate analysis.

4. Exploratory Data Analysis (EDA)

4.1 Distribution of Ratings

A histogram combined with a Kernel Density Estimate (KDE) curve showed:

- **5-star ratings dominate**, indicating high reader satisfaction.
- **4-star ratings are also common**, reflecting generally positive experiences.
- **1-star and 2-star ratings are rare**, suggesting low dissatisfaction.

- The distribution is **right-skewed** toward higher ratings.

4.2 Price vs Rating

A scatter plot was generated to study the relationship between book prices and user ratings.

Key Observation:

- There is **no strong correlation** between price and rating.
- Both low-priced and high-priced books can have high ratings.
- Readers tend to evaluate **content quality rather than cost**, meaning price is not a determinant of satisfaction.

4.3 Top-Rated Authors

A grouped analysis was performed to identify authors with:

- **Highest average ratings**
- **Largest volume of reviews**

Insights from this visualization:

- Some authors achieve both high ratings and high total review counts, showing **popular and well-received** work.
- Others may have high ratings but fewer reviews—suggesting **niche but strong reader appreciation**.

4.4 Yearly Review Trends

A time-series line chart visualized how many reviews were posted each year.

Findings:

- There is a **growth trend** in the number of reviews over time.
- Some years have noticeable peaks, possibly linked to:
    - Rising popularity of online book platforms.
    - Increased user engagement in online reviews.
    - Growth of digital publishing.

5. Key Insights

5.1 Rating Behavior

- **Most users give positive ratings**, with 4- and 5-star categories dominating the data.
- Negative ratings are comparatively rare.
- Readers generally appear satisfied with the books they purchase or review.

5.2 Author Performance

- Top-rated authors not only produce high-quality content but also attract a high number of reviews.
- This suggests a combination of **trust, brand value, and audience loyalty**.

5.3 Price Impact

- Book price has **little to no effect** on user rating.
- Users judge books primarily on:
    - Content quality
    - Relevance
    - Personal enjoyment

5.4 Review Trends

- Reviews have grown consistently over the years.
- Indicates increasing consumer participation and the rising influence of online platforms.

5.5 Review Helpfulness

- Reviews with higher helpfulness scores are usually **longer and more detailed**.
- Readers value informative, well-written reviews over short comments.

6. Key Takeaways

This analysis provides a deep understanding of book ratings, pricing behavior, author performance, and user engagement trends.

- Books are generally rated very positively.
- Pricing does not influence ratings significantly.
- Author success is tied to both high average ratings and high review counts.
- Yearly review volume is increasing, reflecting growing user participation.
- Useful, detailed reviews tend to receive more helpfulness feedback.