INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Books to Scrape: Web Scraping and Exploratory Data Analysis Using Python

Presented By :- Pradnya Vikas Shinde

About me

- **Name :- Pradnya Vikas Shinde**
- **Education** :- Bharti Vidyapeeth deemed University, Pune.
- **Why I Want To Learn Data Analyst** :-
  I enjoy understanding patterns behind data and explaining them clearly, which is why Data Analytics perfectly matches my problem-solving and presentation skills

  Connect With Me :-

  LinkedIn - https://www.linkedin.com/in/pradnya-shinde-0b05a02a9/

  GitHub - https://github.com/pradnya2506

Business Objectives

- To **collect structured book data such as price, rating, and availability** from(Datset ss?) an online source using automated web scraping.
- To **analyze pricing and rating patterns** to understand customer preferences and market trends.
- To identify popular and **highly rated books** that can support better inventory and recommendation decisions.
- To transform **raw web data into actionable insights** through exploratory data analysis (EDA).
- To demonstrate an **end-to-end data analytics** workflow from data extraction to business-ready insights.

Web-Scrapping : Details

- Data was collected from the **Book To Scrap website.(**
- Books information was extracted using **web scraping.**
- The dataset represents real-time, publicly available online book listings collected from the *Books to Scrape* website.
- Scraped attributes include:
  - Title
  - Price
  - Ratings
  - Availability
  - Category
  - Stock
  - Url

Data Summary

Database Schema

- **Data Source:** *Books to Scrape* website (web scraped)
- **Data Format:** Structured tabular data (CSV/DataFrame)

Total Records	**400 Books**
Total Features	**8-9 (Columns)**
Categories Covered	**50+**
Rating Levels	**5 (One to Five Star)**
Price Range	**£10-£60**

Data Quality & Readiness

- Minor missing values due to web scraping
- Duplicate records identified and removed
- Data cleaned, standardized, and prepared for EDA

Data Cleaning

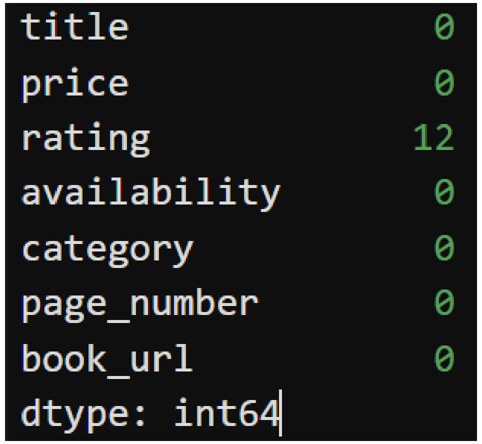
Missing Value Handling

- Identified missing values in **price, rating, and availability** columns
- Handled missing records using **removal and logical imputation techniques**
- Missing values occurred due to **incomplete HTML tags during web scraping**

Data Type Standardisation

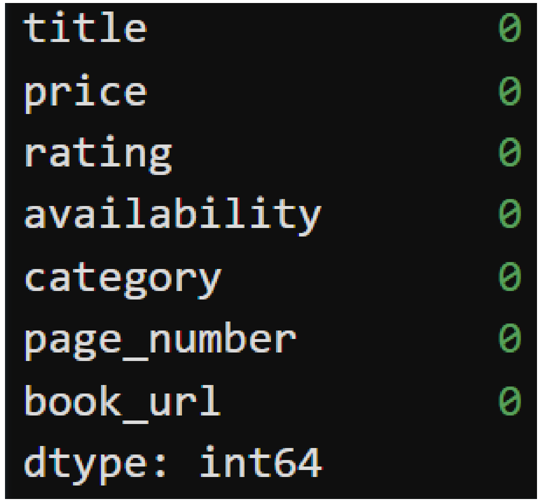
- Converted **price** from text to numeric format
- Transformed **rating** from text labels (e.g., “Three”) to numeric values
- Removed **currency symbols (£)** and unnecessary characters

Missing Values

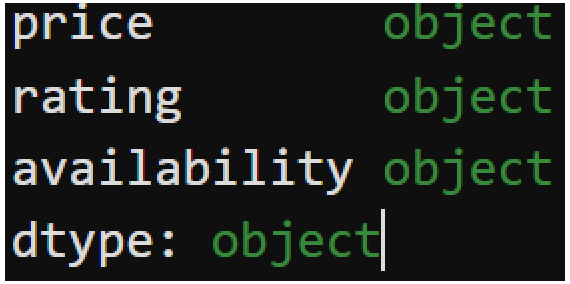


```
title           0
price           0
rating         12
availability    0
category        0
page_number     0
book_url        0
dtype: int64
```

After handling missing values

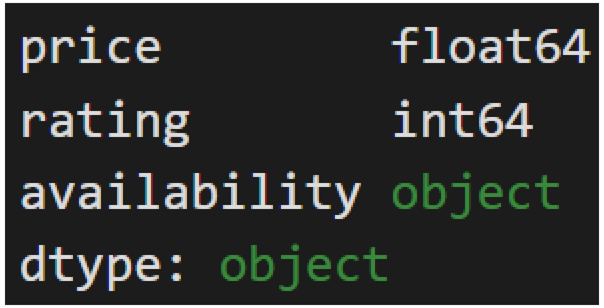


```
title           0
price           0
rating          0
availability    0
category        0
page_number     0
book_url        0
dtype: int64
```



```
price           object
rating          object
availability    object
dtype: object
```

Before Converting data types

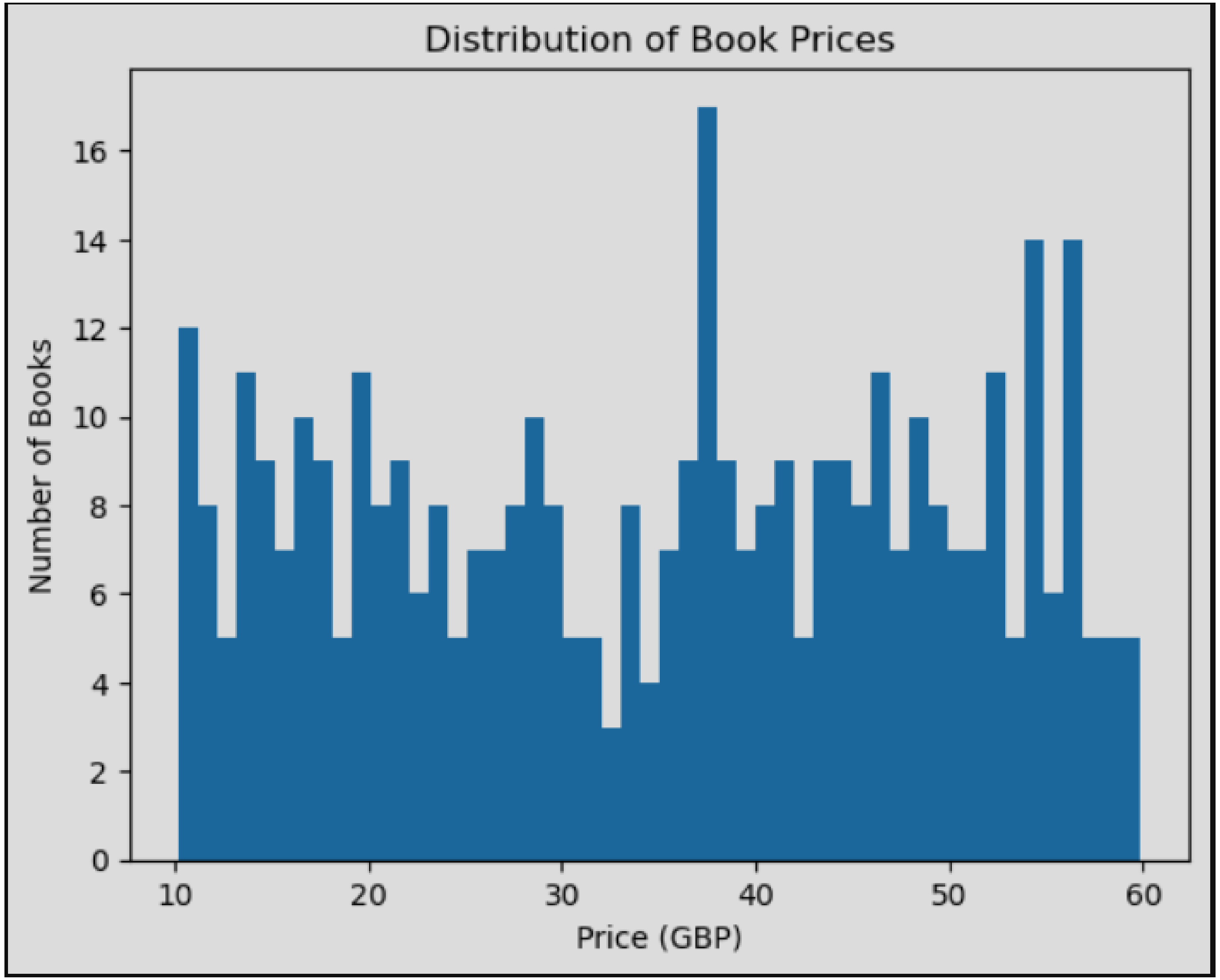


```
price           float64
rating            int64
availability     object
dtype: object
```

After Converting data types

INNOMATICS RESEARCH LABS

Distribution of Book Prices

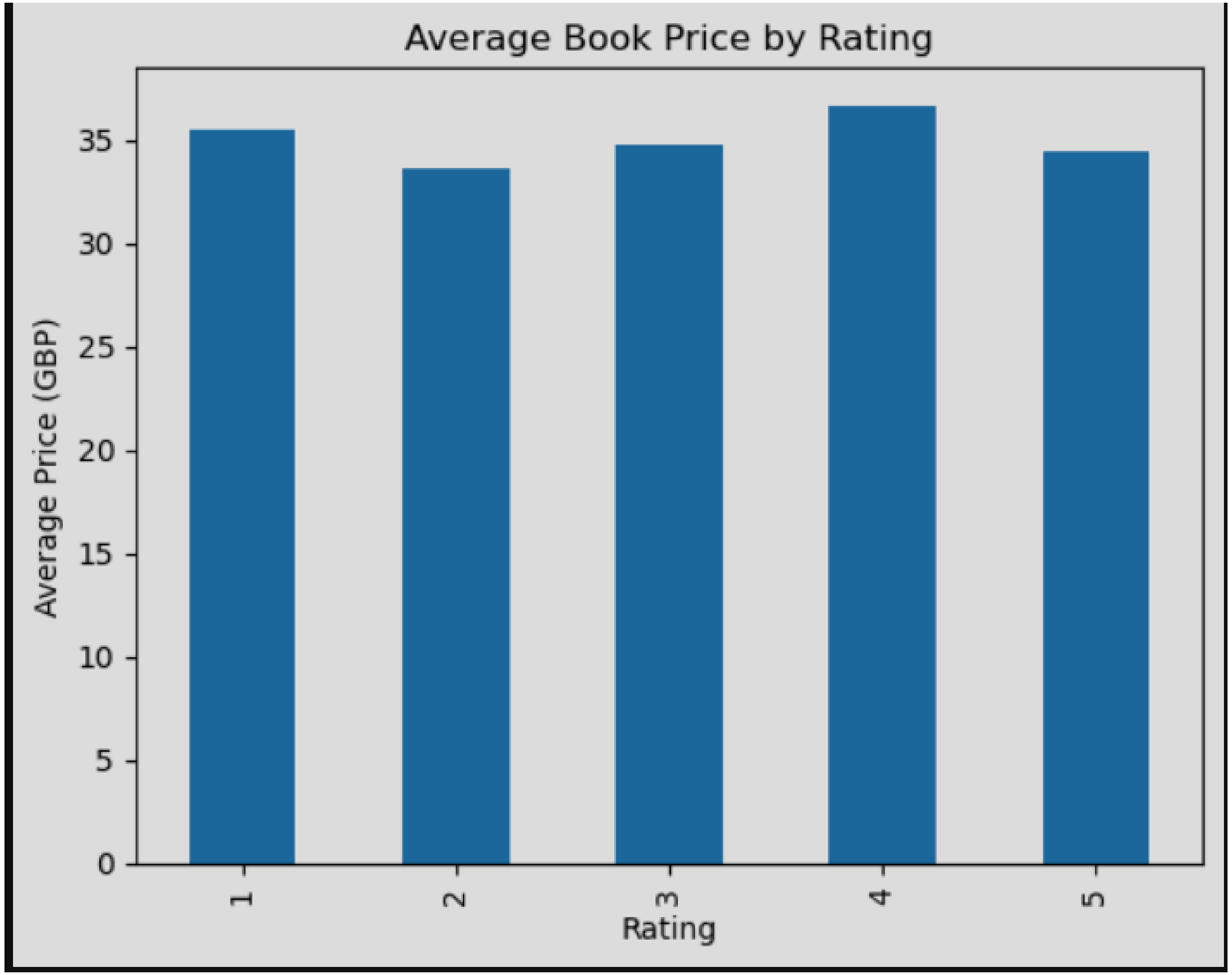


Insights :-

- The highest concentration of books lies in the **£20–£40 price range**
- This shows that **most books are moderately priced**
- Very few books fall in the extremely low or high price ranges
- This graph shows the distribution of book prices.
  Most books fall in the mid-price range, which indicates a focus on affordable pricing

INNOMATICS RESEARCH LABS

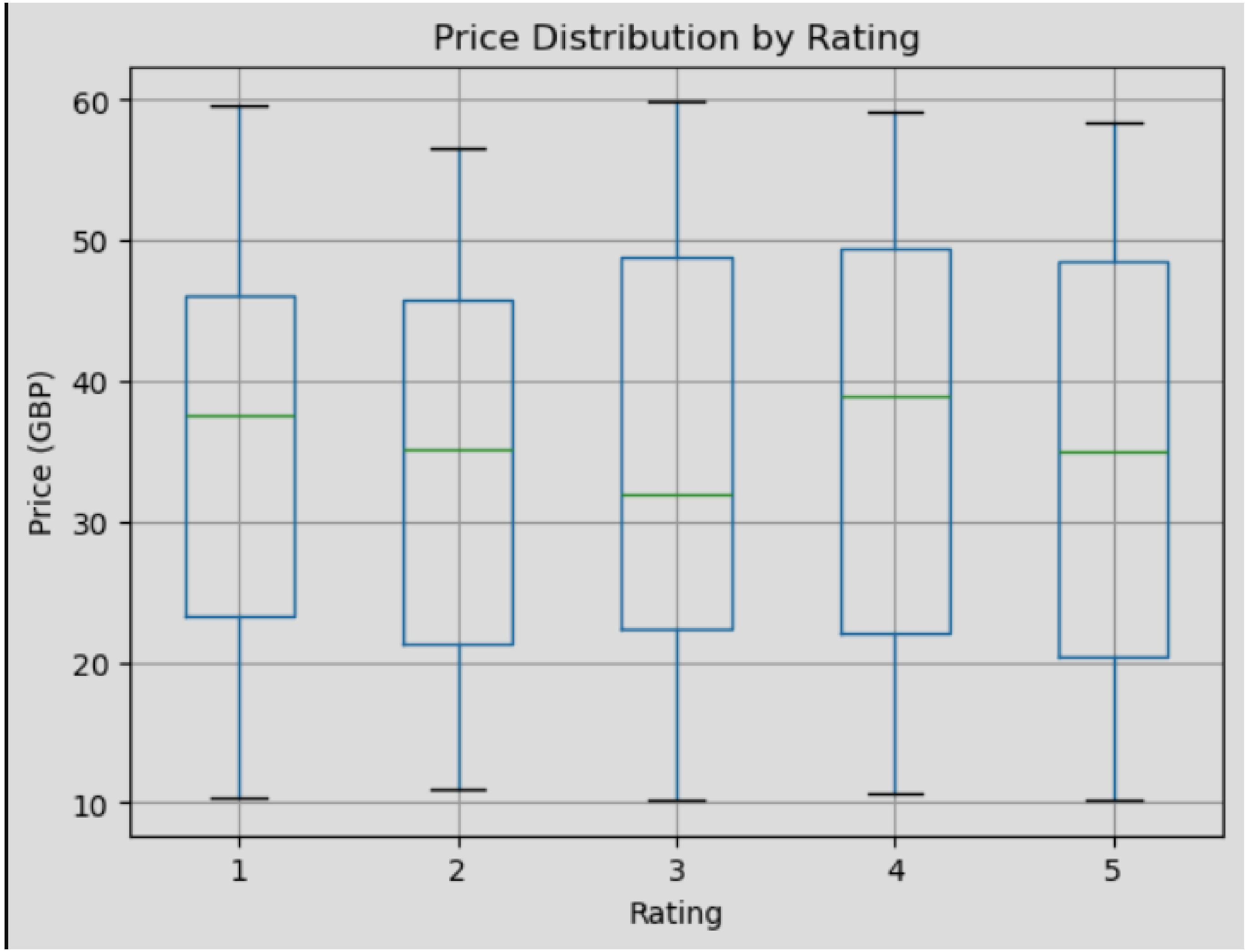
Average Book Price by Rating



Insights :-

- Higher-rated books (4★) tend to have slightly higher average prices.
- Price differences across ratings are small, showing pricing is fairly consistent.

Price Distribution by Rating



Insights :-

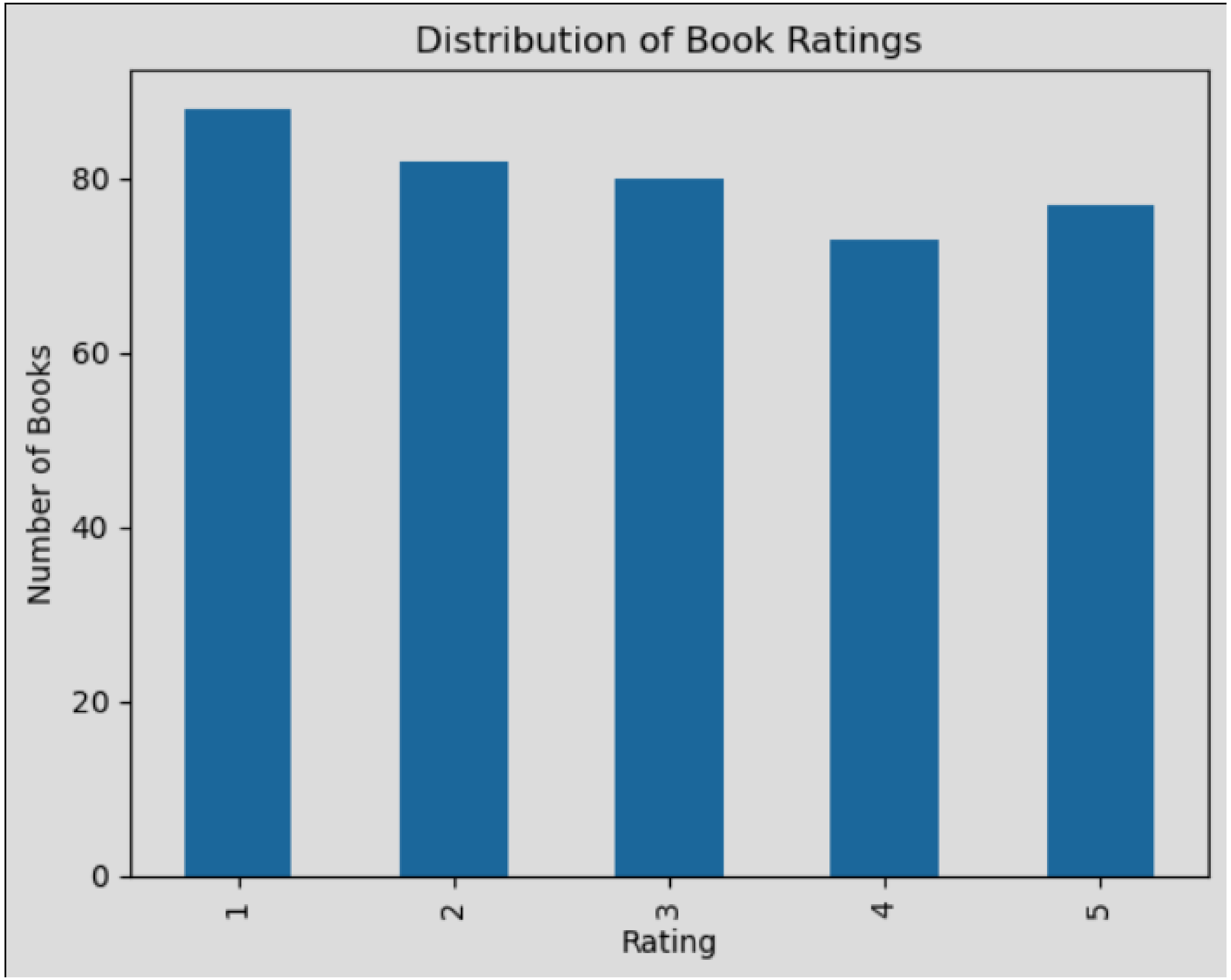
- **Higher-rated books (4★) show a higher median price** compared to lower ratings.

- **Price ranges overlap across all ratings**, indicating rating is not the only factor influencing price.

- This analysis compares prices across ratings.
  Higher-rated books tend to have slightly higher prices, but overall, price differences are small.
  This suggests that ratings alone do not strongly influence book prices

INNOMATICS RESEARCH LABS

Distribution of Book Ratings



Insights :-

- **Lower ratings (1★–2★) appear more frequently** than higher ratings.
- **Ratings are fairly evenly spread overall**, indicating diverse reader opinions across books.
- Here we can see how ratings are distributed.
  Ratings are fairly spread, showing diverse reader opinions

INNOMATICS RESEARCH LABS

Challenges

- **Dynamic and inconsistent HTML structure(attached ss for before data)** across pages made element selection difficult
- **Missing or incomplete data** due to unavailable tags during scraping
- **Text-based values** (prices, ratings) required additional cleaning and conversion.
- **Pagination handling** was needed to extract data from multiple pages
- **Ensuring data quality** while removing duplicates and invalid records

Conclusions and Recommendations

Conclusions :-

- Web scraping successfully extracted **structured book data** from the *Books to Scrape* website.
- The dataset revealed **balanced pricing** with most books in the mid-price range.
- **Ratings and prices show a weak relationship**, indicating price is not solely driven by ratings.

The cleaned dataset was **analysis-ready** and suitable for meaningful EDA.

Recommendations :-

- Focus on **mid-priced books**, as they represent the majority of listings.
- Use **ratings along with other factors** (category, popularity) for better recommendations.
- Automate periodic scraping to **track price and rating trends over time**.
- Extend analysis by including **category-wise and sentiment analysis**.

THANK YOU