Report Final Project

Aya Tallah Medhat Ahmed	23010122
Sara Ashraf El Sayed	23010094
Shahd Ashraf Hassan	23011091
Mariam Tamer Shawky	23011149
Salma Ayman Khamis	23011081

Project Idea

Project Title: Book Data Analysis & Insights from Web Scraping.

The project involves extracting data , an available website created for practicing web scraping techniques. It simulates an online bookstore and includes multiple page listing a wide variety of books.

Types of Data to be collected (Extracted):

- Title
- Price
- Availability
- Rating (in text format such as "One", "Two", ...)
- Product page link

PROJECT GOALS

1 The goal of this project is to **Scrape Book Data** from multiple pages of an e-commerce site and perform **data cleaning, exploration, and various analysis** including clustering and statistical summaries.

2 *The analysis aims to:*

- **Understand pricing trends and rating distributions.**
- **Detect outliers and best deals (cheap high-rated books).**
- **Explore whether there's a relationship between price and rating**
- **Group books into clusters based on price and rating for pattern discovery.**

Tools

Project Research and Approach

1.Data Extraction:

- Used requests and BeautifulSoup to iterate over 49 pages of the website.
- Extracted required fields from HTML content and stored them in a list of dictionaries.

2.Data Cleaning:

- Converted prices from string to float.
- Mapped ratings from words to numbers using a dictionary.
- Checked for duplicates, null values, and outliers using IQR.

Tools

Follow Approaches

3.Data Analysis:

Performed several types of analysis:

- **Descriptive statistics (mean, min, max).**
- **Most common ratings.**
- **Average price per rating.**
- **Price comparison between high-rated and low-rated books.**
- **Frequency of book price ranges.**
- **Word frequency in book titles.**
- **K-Means clustering: Grouped books into 3 clusters based on standardized Price and Rating.**
- **Best deal detection: Identified cheapest books with 5-star ratings.**

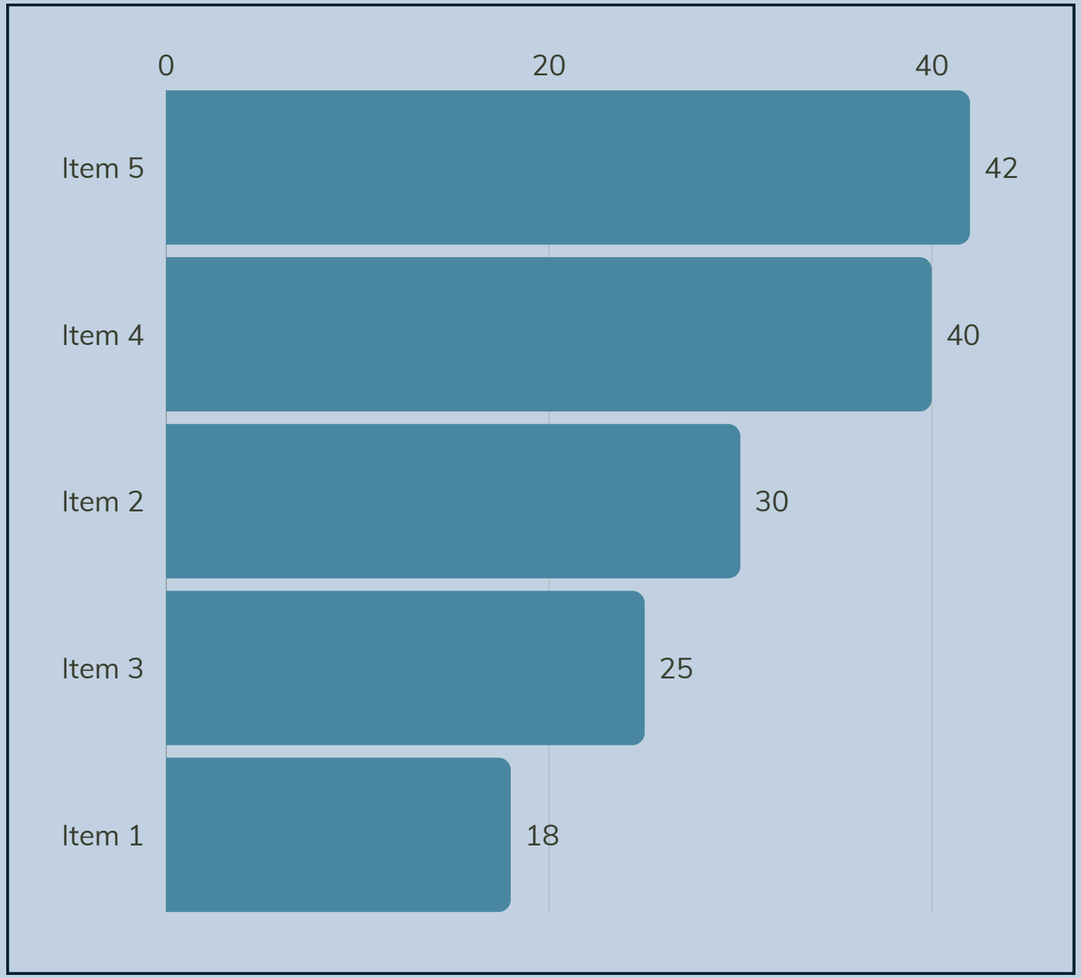
4.Visualization:

Using matplotlib & seaborn:

- **Price distribution (Histogram.)**
- **Price vs. Rating (Scatterplot).**
- **Boxplot by availability.**
- **Violin plot for price & availability.**
- **Count of books by rating.**

RESULT

Data Analysis



Sample Results from Data Analysis:

1. Average Price of Books: £35.03
2. Maximum & Minimum Prices:
   - Highest Price: £59.99
   - Lowest Price: £10.00
3. Most Common Rating: ⭐ 1 Star — appears in 221 books
4. Number of 5-Star Books: 191 books have the highest rating.
5. Books with High Rating & Low Price (Best Deals):
   - An Abundance of Katherines — ⭐ 5 stars, £10.00
   - Greek Mythic History — ⭐ 5 stars, £10.23
6. Average Price per Rating:
   - 1 star: £34.33
   - 5 stars: £35.44
   - → Slight increase in price with higher ratings
7. Price Ranges Distribution:
   - 0–10£: 1 book
   - 10–20£: 194 books
   - 40£ and above: 396 books
   - → Most books are in the higher price range
8. Most Common Words in Titles:
   - "life", "girl", "love", "vol", "you"
9. Clusters Discovered by K-Means: Books were grouped into 3 clusters based on price and rating, indicating different pricing tiers and perceived quality levels.

CONCLUSION

In this project, we successfully scraped book data from a simulated online bookstore and conducted a comprehensive analysis of the collected dataset. Using Python libraries such as requests, BeautifulSoup, pandas, and seaborn, we extracted, cleaned, and analyzed data related to book titles, prices, ratings, availability, and links.

Overall, the project demonstrates how web scraping and data analysis can be combined to generate actionable insights from unstructured online data. This approach can be extended to real-world scenarios in e-commerce, marketing, or recommendation systems to better understand product trends and customer preferences

Tasks

Data Extraction & Data Cleaning, Processing, and Regular Expressions:

Sara Ashraf & Shahd Ashraf

Data Analysis: Aya Tallah

Data Visualization: Salma Ayman & Mariam Tamer

THANK YOU