An Empirical Analysis of Book Prices Using Web-Scraped Online Retail Data

Team Members:
Qiyan Chen (Individual Project)

1. Project Overview and Research Question

This project is titled *An Empirical Analysis of Online Book Prices Using Web-Scraped Retail Data* and was completed as an individual project by Qiyan Chen. The primary objective of the project is to examine how book prices are distributed in an online retail environment and to explore whether observable pricing patterns emerge across different product categories.

The central research question guiding this project is: *How are book prices distributed across categories in an online bookstore, and what descriptive patterns can be observed in terms of price dispersion and variability?* While the scope of the project is exploratory rather than causal, it aims to demonstrate how web-scraped data can be transformed into structured datasets suitable for statistical analysis and visualization.

More broadly, this project serves as an end-to-end demonstration of a data pipeline implemented entirely in Python. It integrates data collection through web scraping, systematic data cleaning, descriptive statistical analysis, and visual presentation of results. Emphasis is placed on reproducibility, transparency, and adherence to best practices in data processing.

2. Data Collection Process and Data Description

The data used in this project were collected from Books to Scrape (http://books.toscrape.com/), a publicly available website specifically designed for practicing web scraping techniques. The website consists of static HTML pages, making it well-suited for automated data extraction using standard Python libraries.

Data collection was implemented using the requests library to retrieve webpage content and BeautifulSoup to parse relevant HTML elements. The scraping script programmatically iterated through all catalog pages, ensuring comprehensive coverage of the available inventory. For each book listing, the script extracted the book title, product category, listed price in British pounds, availability information, and a timestamp indicating when the data snapshot was taken.

A total of 2,000 individual book records were collected, with each observation representing a unique book product at the time of scraping. The raw data were saved in JSON format within the data/raw/ directory, preserving the original structure of the scraped information prior to any cleaning or transformation. Importantly, no pre-downloaded datasets or external APIs were used; all data were obtained directly through Python-based web scraping, fully satisfying the project's data collection requirements.

During the development of the project, an initial attempt was made to scrape data from a commercial retail website. However, access restrictions and repeated HTTP 403 errors made consistent data retrieval unreliable. To address this challenge, the data source was revised to Books to Scrape, which ensured stable access while maintaining the core methodological goals of the project. This adjustment represents the primary deviation from the original plan and highlights a common real-world challenge in web-based data collection.

3. Data Cleaning and Preparation

Following data collection, a comprehensive data cleaning process was conducted to convert the raw scraped output into a structured and analysis-ready format. The cleaning procedures were implemented using pandas and focused on ensuring data consistency, accuracy, and usability.

HTML artifacts and non-numeric characters were removed from price fields, and prices were converted to numeric values. Observations with missing or invalid price or title information were identified and removed. Duplicate records were also eliminated based on a combination of product identifiers and snapshot timestamps to prevent repeated entries from affecting the analysis.

Timestamp variables were standardized and converted to consistent datetime formats to ensure compatibility with downstream processing and serialization. The cleaned dataset was then saved in both JSON and CSV formats within the data/processed/ directory. This dual-format output supports flexibility for future analysis while maintaining a clear separation between raw and processed data.

The cleaning process resulted in a finalized dataset of 2,000 observations with consistent variable definitions and no missing critical values, providing a reliable foundation for statistical analysis.

4. Data Analysis and Statistical Findings

The analytical component of the project focuses on descriptive statistics and exploratory analysis of book prices. Using pandas and numpy, summary statistics were computed to characterize the overall price distribution, including the mean, median, standard deviation, minimum, maximum, and interquartile range.

The results indicate that book prices on the platform exhibit moderate dispersion, with most prices clustered within a relatively narrow range. The proximity of the mean and median suggests that the distribution is approximately symmetric, with limited skewness. While a small number of higher-priced books exist, extreme outliers are rare, indicating relatively standardized pricing across the marketplace.

In addition to overall price statistics, prices were aggregated by product category to compute category-level average prices. This analysis reveals that certain categories tend to be priced slightly higher on average, suggesting that genre or subject matter may play a role in pricing strategies. However, the analysis remains descriptive, and no causal claims are made regarding pricing determinants.

5. Visualization and Interpretation

To complement the numerical analysis, several visualizations were created using the matplotlib library. A histogram of book prices illustrates the overall distribution and highlights the concentration of prices in the lower-to-mid range. This visualization provides an intuitive understanding of price frequency and variability.

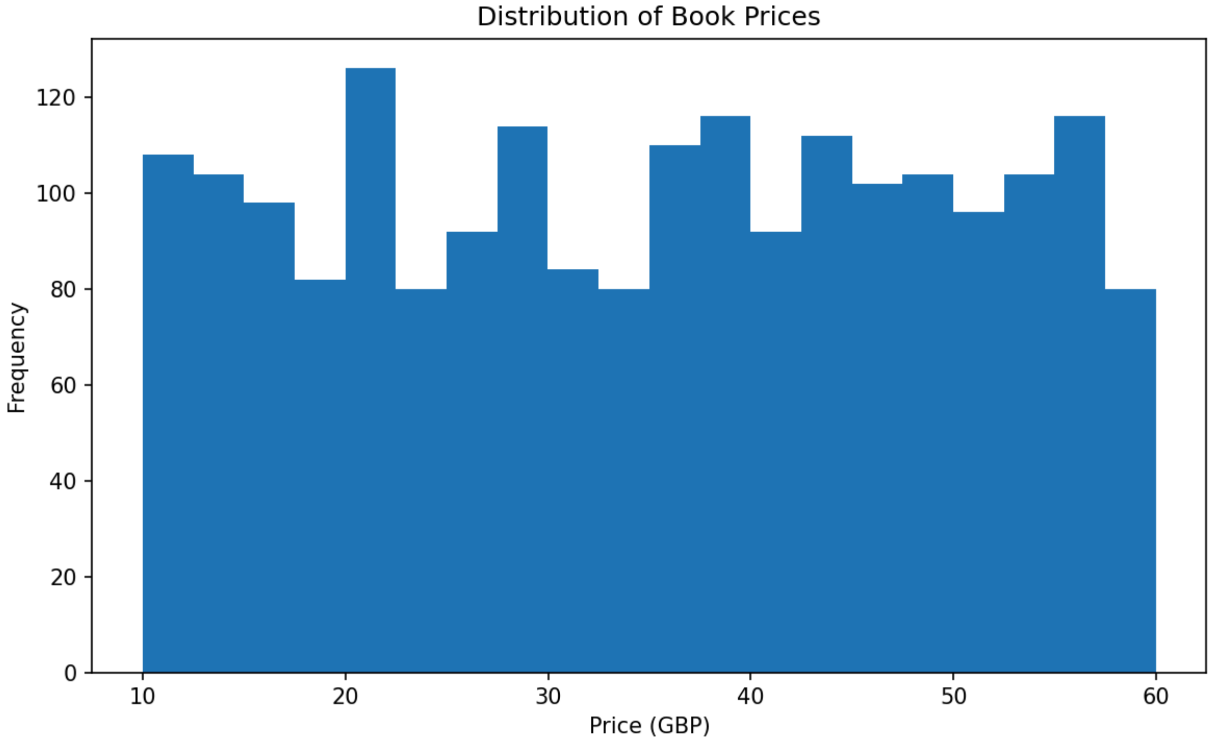

Figure 1.Histogram — Distribution of Book Prices

To further examine price dispersion and identify potential outliers, a boxplot of book prices was constructed.A boxplot was also generated to summarize the distribution using quartiles, offering a clear visual representation of the median, interquartile range, and potential outliers. This figure reinforces the conclusion that price dispersion is limited and that extreme values are uncommon.

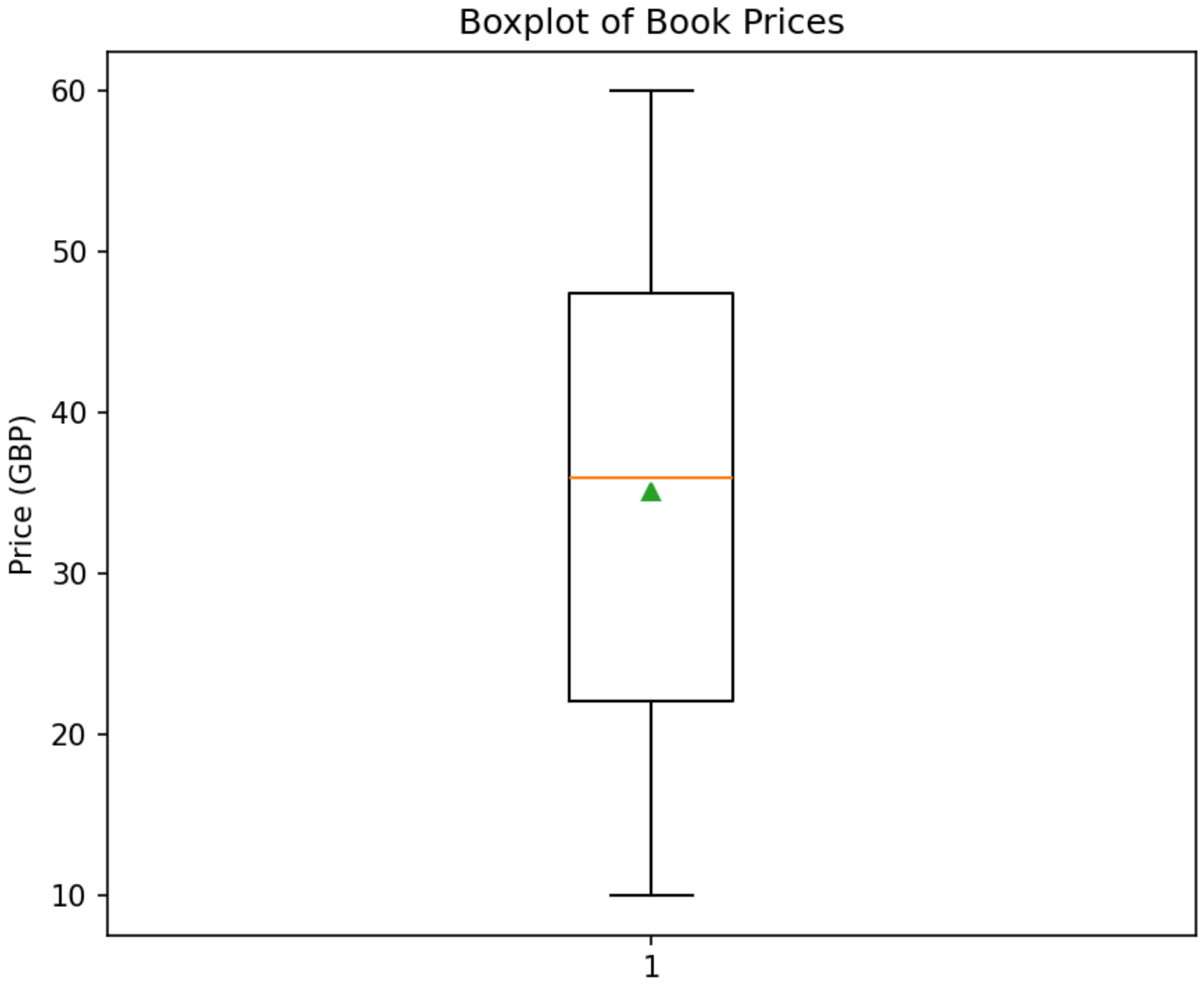

Figure 2.Boxplot — Boxplot of Book Prices

Finally, a bar chart displaying the ten most expensive books was produced to examine the upper tail of the price distribution. By explicitly labeling individual titles and prices, this visualization provides insight into how high-priced items compare to the broader market.

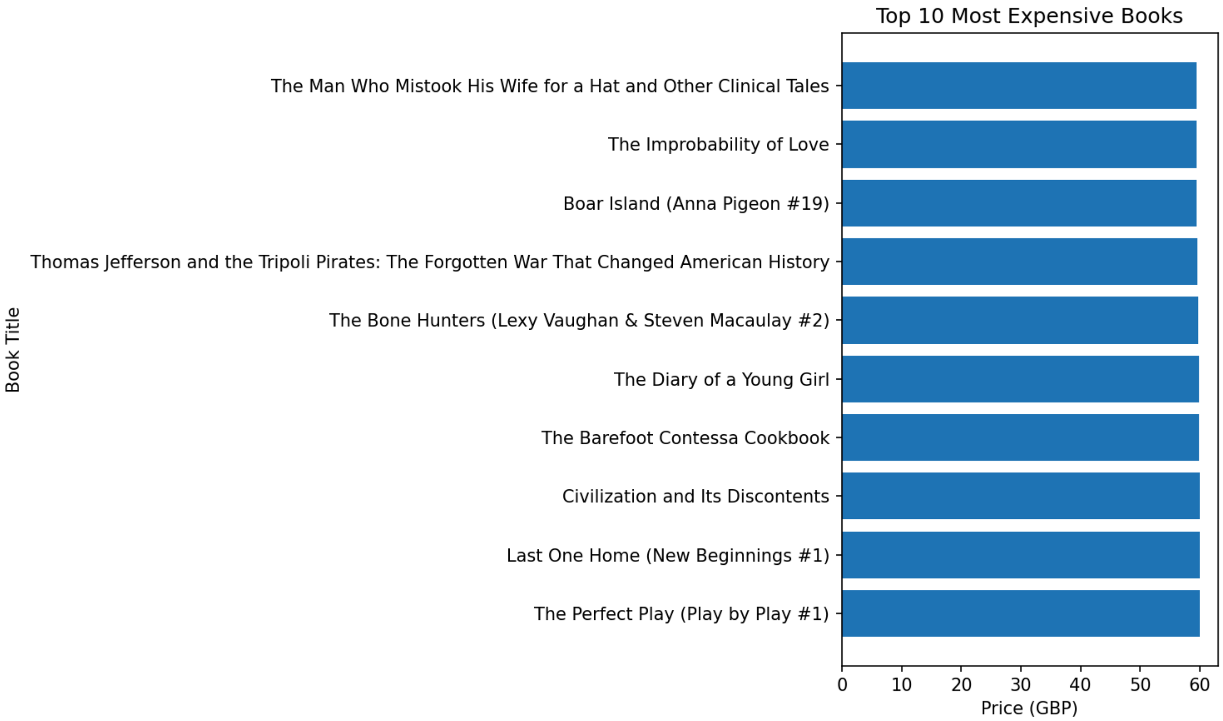

Figure 3.Bar Chart — Top 10 Most Expensive Books

All visual outputs were saved to the results/ directory, ensuring reproducibility and clear organization of analytical outputs.

6. Conclusions and Implications

The findings from this project suggest that online book prices on the analyzed platform are relatively stable and tightly distributed. Category-level differences in average prices indicate some degree of pricing differentiation, but overall variability remains limited. These results align with expectations for standardized consumer goods in competitive online markets.

Beyond the substantive findings, the primary contribution of this project lies in demonstrating a complete and reproducible data workflow. The project illustrates how web-scraped data can be systematically cleaned, analyzed, and visualized using widely adopted Python tools. As such, it provides a template for future projects involving online retail data or similar web-based sources.

7. Future Work

Given additional time and resources, the project could be extended in several meaningful directions. Collecting data at multiple points in time would enable the analysis of price dynamics and temporal trends. Incorporating additional variables such as customer ratings or review counts could support richer analyses of price-quality relationships. Finally, expanding the data sources to include multiple retailers would allow for comparative studies across platforms.

In summary, this project fulfills all technical and analytical requirements by implementing an end-to-end data pipeline grounded in reproducible Python code. Through careful data collection, cleaning, analysis, and visualization, the project demonstrates both methodological rigor and practical applicability in the context of online price analysis.