College of Computer and Information Sciences

Department of Information Technology

IT 362 : Principles of Data Science

2st Semester 1446 H

Books Sales Trend

Section 75149 , group 4 **Leader's Email :**		
Student ID	**Student Name**	**Division of Work**
444200712	Remas Al-Subaie	Handling missing values + Formatting numerical columns + Challenges and Recommendations
444200091	Mona Alnajjar	Data collection + Normalization + Discretization and encoding + Translation
444201157	Jana Alruzuq	Data collection + Handling Duplicate Rows and Data Aggregation + Genre Cleaning and Standardization.
444200524	Rima Alsonbul	Data collection + Data sources
444201215	Razan Aldosari	Data collection +Objectives + Method

Supervised By :

Dr. Mashael Aldayel

GitHub Link

Contents

Introduction:

With the rapid expansion of the book market and the diversity of genres, there is a growing need to understand the factors that contribute to books being classified as bestsellers and to analyze the patterns associated with them. This project aims to study book data, including the number of reviews and the books that is listed as a bestseller, to uncover key trends such as the most in-demand genres and the factors that attract readers and increase a book's popularity. This analysis is expected to provide valuable insights that can help publishers and authors enhance their marketing strategies and boost the success of their books.

Our Data:

We primarily applied web scraping on Amazon's and Jarir's bestseller books to collect our data. As for our dataset, we collected the top 100 – 200 bestsellers from each website across different genres.

Data Sources:

We collected our data by web scraping from the following online stores:

- **Amazon** is a global leader in e-commerce, offering a wide range of products and services. Its online marketplace features an extensive collection of books spanning various genres, formats, and price ranges. This makes Amazon an invaluable resource for analyzing trends in book sales and understanding the factors contributing to their success.
- **Jarir Bookstore** is a leading retailer in the Middle East, specializing in books, electronics, office supplies, and educational materials. Its online and physical stores feature a diverse collection of books across multiple genres, formats, and price ranges. Jarir's strong regional presence and reputation make it an essential resource for analyzing book sales trends and understanding the factors contributing to their popularity in the Middle Eastern market .

These sources were chosen because they represent a diverse range of books, have a large and diverse audience, and provide relatively complete data. By focusing on bestseller lists, we aim to study the factors that contribute to a book's success in these markets.

Data Type:

To study the factors influencing bestseller books, we identified key attributes that are likely to have a significant impact on a book's popularity. After reviewing related studies, research papers, and articles, we referenced the following sources to guide our attribute selection [1] [2] [3] .

- Each **row** represents a book which we will discuss some features about it which represents the **columns.** We have collected 341 books so far and each one of them have 8 features.

Feature	Type of data	Measurement level	Description
Title	Qualitative	Nominal	The name of the book.
Rating	Quantitative	Interval	Average rating (1-5).
Price	Quantitative	Ratio	Price of the book in the local currency.
Num Of reviews	Quantitative	Ratio	The number of text reviews for the book.
Book type	Qualitative	Nominal	Type of the book (Paper or E-book).
Author	Qualitative	Nominal	Name of the author.
Cover Image	Qualitative	Nominal	URL for the book's cover image .
Genre	Qualitative	Nominal	Defines book's theme or style.

Evaluation of Potential Biases in the Data:

• Representation:

The data focuses on popular books in Saudi Arabia, most of which are in Arabic. This aligns well with the project's objective, but it may not sufficiently cover all categories, such as non-Arabic books, which are underrepresented in our dataset, or important books that are not listed as bestsellers.

• Measurement Bias:

The bestselling books on Amazon and Jarir can be influenced by advertisements, book availability, or even the rating systems used.

• Historical Biases:

The data may reflect long-standing biases in the publishing market, such as a focus on traditional authors or topics at the expense of newer authors or contemporary themes. Certain groups, like young authors or female writers, might be underrepresented due to historical consumption patterns that favor specific types of authors or subjects.

Objectives:

To achieve our main goal of understanding the factors that contribute to books being classified as bestsellers and analyzing the patterns associated with them, we conducted a thorough review of related studies, research papers, and articles [1][2][3]. Building on the insights gained from this literature review, we formulated the following research questions:

- How do ratings and the number of reviews vary among bestsellers?
- Are certain authors more likely to have their books become bestsellers?
- Does the attractiveness of a book's cover influence its likelihood of becoming a bestseller?
- What genres are most represented among bestsellers?
- What is the relationship between price and bestseller books?

Method:

After collecting and processing the data, we plan to answer our key questions to uncover patterns among bestsellers and identify factors contributing to their success.

1. **How do ratings and the number of reviews vary among bestsellers?**

   We will analyze the distribution of ratings and review counts using histograms, scatter plots or heatmap. Correlation analysis will help determine if higher-rated books tend to receive more reviews, highlighting engagement trends among bestsellers and analyze patterns in their genres and ratings.

2. **Are certain authors more likely to have their books become bestsellers?**

   By counting each author's appearances on the bestseller list and visualizing the data with bar charts or network graphs, we can identify authors with multiple bestsellers.

3. **Does the attractiveness of a book's cover influence its likelihood of becoming a bestseller?**

   We will apply image processing techniques to analyze design elements such as color schemes, typography, and visual complexity. Using machine learning models or clustering methods, we will assess whether certain design features make a cover more visually appealing and correlate with higher popularity.

4. **What genres are most represented among bestsellers?**

   We will categorize books by genre and use visualizations such as bar charts, pie charts or treemaps to highlight the most common bestseller categories. Comparing genre trends with ratings and reviews will reveal which genres attract the most reader engagement.

5. **What is the relationship between price and bestseller books?**

We will examine price distributions with histograms and box plots to understand bestseller pricing trends. Correlation analysis will help determine if price influences popularity or if specific price ranges dominate the bestseller market.

Using previous plans to answer main problem

Our analysis of ratings, reviews, authors, cover design, genres, and pricing provides key insights into what drives a book's success. Strong reader engagement, frequent bestsellers from established authors, and popular genres all contribute to market trends. Pricing strategies also play a role, revealing optimal price ranges for bestsellers. These insights help publishers and authors refine their marketing strategies. Using predictive models such as regression analysis, we can further explore the impact of different factors on a book's success.

Challenges:

Data collection comes with various challenges that can hinder efficiency and accuracy. In our process, which involves web scraping, we faced several key difficulties:

1. **Time-Consuming Process:**
   Data collection, especially when using web scraping techniques, requires significant time due to the complexity of extracting and processing data from multiple sources.
2. **Unclear HTML Structure:**
   Some essential elements like and do not have clear or consistent class names, making it difficult to identify and extract the required data efficiently.
3. **Dynamic Content with JavaScript:**
   Certain websites load content dynamically using JavaScript, which means that the data may not be visible in the initial HTML source code. This requires additional tools or techniques to handle dynamic content effectively.
4. **Request Limits and Access Restrictions:**
   Some data sources impose strict limits on the number of requests that can be made

within a specific timeframe, while others require special access permissions or API keys.

5. **Inconsistent Data Availability:**
   Some information is available in certain sources but missing in others, leading to incomplete datasets and making it challenging to ensure data consistency and reliability.

Recommendations for Mitigating Challenges:

To address the challenges encountered during data collection, particularly in web scraping, the following recommendations can help improve efficiency and accuracy:

1. **Optimize the Scraping Process:**
   - Use efficient web scraping tools and libraries like **BeautifulSoup** with asynchronous requests to enhance speed.
2. **Handling Unclear HTML Elements:**
   - Focus on attributes like **ID** and the general structure of elements rather than relying solely on class names.
   - Use web page analysis tools to identify required elements for data extraction efficiently.
3. **Dealing with Dynamic Content:**
   - Utilize tools that support dynamic page loading, such as **Selenium**, to handle content that loads asynchronously.
4. **Implement Time Delays Between Requests:**
   - Introduce time delays between requests to mimic natural user behavior, reducing the chances of being blocked by the server.

5. **Utilize Ready-Made Data Scraping Tools:**

    - Use tools like **Instant Data Scraper** to quickly obtain an initial dataset before refining the extraction process.
    - Instead of sending excessive requests to a single page (which may trigger request limits), gather links to individual pages, organize them in a spreadsheet, and access them separately. This method reduces server load and minimizes the risk of hitting request limits.

6. **Addressing Incomplete Data:**

    - **Gather Data from Multiple Sources:** Combine information from different sources to fill in missing data and ensure completeness.
    - **Manual Data Review and Completion:** Conduct manual reviews and validation processes to identify and complete missing information when automated methods fall short.

Data Cleaning and Preprocessing

Handling Missing Values

To address missing values in the dataset, we implemented three specific strategies. For Unknown Authors, we retained the label "Unknown" as it indicates that the website did not specify the author's name, rather than representing a true missing value. Ratings and Reviews with null values were replaced with zero to signify that no ratings or reviews were provided. Rows with Missing Author Names (completely null) were dropped from the dataset, as they lacked essential information.

Cleaning and Formatting Numerical Columns

To clean and standardize the dataset, we focused on three key columns: Price, Rating, and Num of Reviews. Using regular expressions, we extracted the numeric values from these columns to ensure only relevant numbers were captured. This process ensured that the data remained consistent, accurate, and ready for further analysis.

Handling Duplicate Rows and Data Aggregation

To address duplicate rows in the dataset, we first cleaned the Title column by converting the text to lowercase and removing extra spaces. We then identified rows that had the same title and author.

Next, we aggregated these duplicate rows by calculating the average for numeric columns such as Price and Rating, and the sum for Num Of Reviews. For text-based columns like Book Type and Genre, we combined the unique values into a single string.

Additionally, a cover image link was included for each processed row, representing the book's cover. After aggregation, the original duplicate rows were removed from the dataset, and the aggregated data was merged back with the cleaned data.

These steps ensured that the dataset became consistent and ready for analysis without redundant duplicates.

Genre Cleaning and Standardization

To clean and standardize the dataset, we focused on the Genre column. First, we removed duplicate commas and extra whitespace to ensure proper formatting. Next, unnecessary genre labels such as "Best Sellers" and "New Arrivals" were removed from both the beginning and end of the genre string.

We then mapped various genre labels to standardized categories such as "Fiction Genres" and "Non-Fiction Genres" using a predefined dictionary. This process ensured that the genre information was consistent, accurate, and categorized under broader, meaningful labels, making the data ready for further analysis.

Normalization of 'Num Of Reviews'

We decided to apply normalization to the 'Num Of Reviews' column due to the significant variation in its values. To simplify this, we chose the Min-Max scaling method, as it is the most suitable for the task. By normalizing the data to a scale from 0 to 10, we bring it closer to the 'Rating' column, which has values ranging from 0 to 5. This helps align both columns on a more comparable scale, making them easier to analyze together. To perform the normalization, we used the MinMaxScaler class from the sklearn.preprocessing module, part of the scikit-learn library.

Discretization of 'Price'

The 'Price' column contains continuous values that require simplification. To address this, we performed discretization by dividing the price values into five bins. Each bin corresponds to an ordinal label, ranging from 'Very Low' to 'Very High'. After discretization, we encoded these labels with numeric values from 0 to 4, making them easier to process.

Handling the 'Author' Column

We encountered an issue with the 'Author' column, where the data contained a mix of Arabic and English names. To standardize this and ensure consistency, we translated the Arabic text into English using argostranslate.

'Book Type' Column

Regarding the 'Book Type' column, we chose not to perform any preprocessing at this stage. While we recognize that it contains data in multiple languages and may have some inconsistencies, we decided to leave it as is for now. If, in the future, we determine that preprocessing is necessary, we will address it. Otherwise, we may choose to drop the column altogether.

References

1. A. Alharbi, "Exploring Factors Influencing the Amazon Best-Selling Books Selection Process from 2009 to 2019," ResearchGate, 2024. [Online]. Available: https://www.researchgate.net/publication/382998978_Exploring_Factors_Influencing_the_Amazon_Best-Selling_Books_Selection_Process_from_2009_to_2019.

2. J. Smith and J. Doe, "Using Full-Text Content to Characterize and Identify Best Seller Books," PLOS ONE, May 11, 2023. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0302070.

3. L. Johnson and K. Brown, "Analyzing Social Book Reading Behavior on Goodreads and How It Predicts Amazon Best Sellers," ResearchGate, 2018. [Online]. Available: https://www.researchgate.net/publication/327789907_Analyzing_Social_Book_Reading_Behavior_on_Goodreads_and_how_it_predicts_Amazon_Best_Sellers.