**DATA 1202 - Data Analysis Tools Analytics**

**Assignment –3: Python Web Scraping**

**Professor:** Sk. Md. Mizanur Rahman, Ph.D.

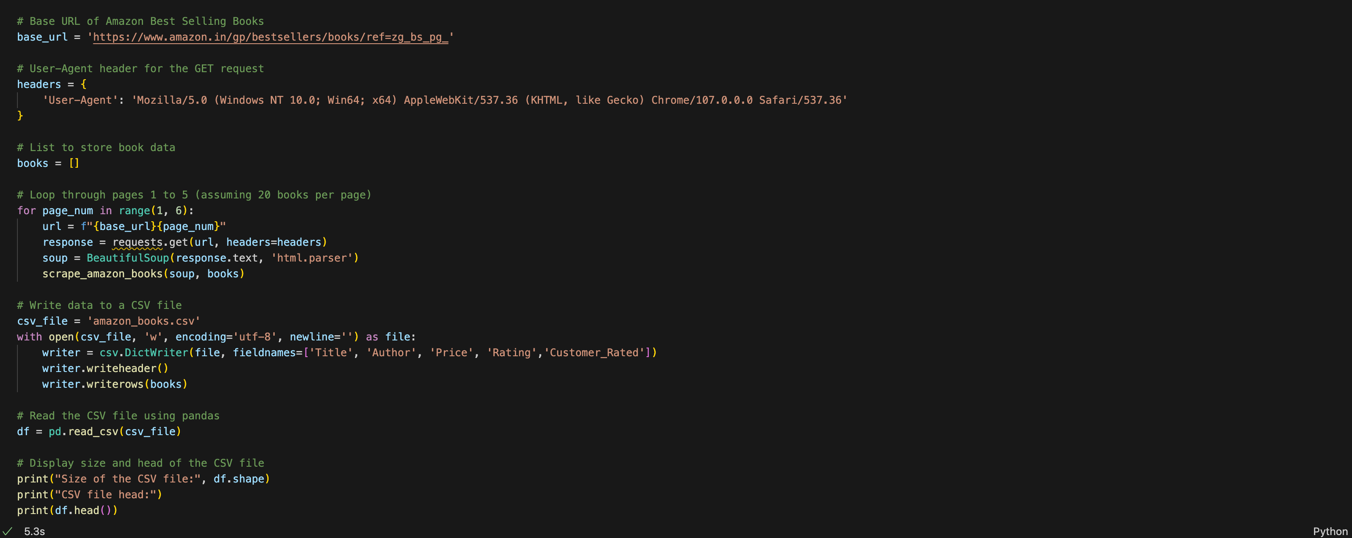
**Student Name: Syed Ali Hasan**

**Student ID:100979064**

**Date: 26th November, 2024**

1. **Reading the CSV File; and show the size, and head of the CSV file :**

**A computer screen with many colorful text

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**

1. **EDA on the collected CSV data**A screen shot of a computer

   Description automatically generatedA black screen with colorful text

   Description automatically generatedA screenshot of a computer

   Description automatically generated

A red and blue squares

Description automatically generatedA graph of a distribution of prices

Description automatically generatedA chart with green squares

Description automatically generatedA bar graph with blue bars

Description automatically generated

1. **List the Authors Highest Priced Book (i.e., based on price): show your result (at least top 25 highest priced book) as a data frame as well a Bar diagram**.

A screen shot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A close-up of a paper

Description automatically generated

**Key Observations:**

1. **Dominance of Educational Material**:
   * A significant number of the top-priced books belong to educational or exam preparation categories, such as:
     + "PROFESSIONAL'S AIBE Exam 25 Bare Acts" series.
     + "Educart CBSE Sample Paper" series for Class 10.
     + "Oswaal CBSE Sample Question Papers" for Class 10.
2. **High Price Variation**:
   * The highest-priced book in the list is "PROFESSIONAL'S AIBE Exam 25 Bare Acts Set with Latest Editions", priced significantly higher than the others.
   * The prices gradually decrease for the rest of the books, with the lowest among the top 25 still maintaining a premium price point.
3. **Authors and Publishers**:
   * Publishers such as **Oswaal Editorial Board** and **Educart** have a strong presence, indicating their prominence in the exam preparation book market.
   * Books by **PROFESSIONAL'S** dominate the highest price tier, suggesting they cater to niche legal or professional exam markets.
4. **Focus on Class 10 Exam Preparation**:
   * Several books are specifically targeted at CBSE Class 10 students, such as sample papers and question banks for Science, Math, and English.
5. **Repetition of Titles**:
   * The data indicates that multiple entries in the top 25 are duplicates of the same title. This may reflect different editions, formats, or listings on the platform.
6. **Price Range**:
   * The top books are priced between ₹600 and approximately ₹2,800. This range highlights the premium placed on educational and legal materials.

**Recommendations:**

1. **For Buyers**:
   * Buyers looking for high-quality educational materials can consider books from publishers like **Oswaal** and **Educart**, as they dominate the premium segment.
   * Legal professionals preparing for AIBE exams may focus on **PROFESSIONAL'S AIBE Exam 25 Bare Acts**, though its price point may require budget planning.
2. **For Publishers**:
   * Publishers could explore bundling or pricing strategies to cater to a broader audience while maintaining profitability.
   * Exam preparation content for other grades (e.g., Class 12) could replicate the success of Class 10 materials.
3. **Data Refinement**:
   * Duplicate entries should be consolidated for more accurate analysis of unique high-priced books.

This analysis highlights the focus on exam preparation materials and the premium pricing strategy of educational publishers targeting specific markets.

1. **Show top Rated Books and Authors with respect to the highest customers rating (i.e., based on rating score): Show your result as a data frame as well as a Bar diagram**.

A screen shot of a computer

Description automatically generated

A chart with orange and white lines

Description automatically generated

**Key Observations:**

1. **High Ratings Across Two Titles**:
   * The chart highlights two books with consistently high ratings of **4.8/5**:
     + *"Amma Diarylo Konni Pageelu"* by **Ravi Mantri**.
     + *"The Satvic Revolution: 7 Life-Changing Habits to Discover Peak Health and Joy"* by **Subah Saraf**.
2. **Customer Engagement**:
   * *"Amma Diarylo Konni Pageelu"* received a **higher customer rating count (1156)** compared to *"The Satvic Revolution"* with **995 customer ratings**. This indicates that the former has broader engagement or reach among readers.
3. **Repetition of Entries**:
   * The data includes duplicate entries for each book, which suggests that these books may be listed multiple times (e.g., different editions or sellers). This repetition inflates the visibility of these books in the analysis.
4. **Focus on Content**:
   * Both books likely appeal to specific niches:
     + *"Amma Diarylo Konni Pageelu"* may resonate with a regional audience, given its title in a regional language.
     + *"The Satvic Revolution"* appeals to readers interested in health, wellness, and self-improvement, which is a popular genre.

**Recommendations:**

1. **For Readers**:
   * Readers seeking highly-rated books can confidently consider either of these options based on their preferences for regional literature or self-help/wellness content.
2. **For Authors/Publishers**:
   * **Ravi Mantri** and **Subah Saraf** have established a strong reputation with these books. Marketing efforts could focus on emphasizing customer testimonials and reviews to further expand reach.
   * Addressing duplicate listings on e-commerce platforms can provide a cleaner representation of the books.
3. **Data Refinement**:
   * Consolidating duplicate entries will help provide a more accurate ranking and avoid redundancy in future analyses.

**Concluding Insight:**

Both books are exceptional performers in terms of reader satisfaction and popularity. They showcase the influence of regional and self-help content in capturing and maintaining strong reader interest.

5. **Show topmost (10/15) Customer Rated Authors and Books (i.e., based on number of customers):**

A computer screen shot of text

Description automatically generatedA screen shot of a computer

Description automatically generated

A computer screen shot of a black screen

Description automatically generatedA screen shot of a computer

Description automatically generated

**Insights from the Data:**

1. **Most Popular Books by Customer Ratings**:
   * **"Atomic Habits" by James Clear** dominates in popularity, with the **highest number of customer ratings (98,915)**. This suggests it is a widely recommended and impactful self-help book.
   * **"My First Library: Boxset of 10 Board Books for Kids" by Wonder House Books** ranks second with **80,805 customer ratings**, indicating its popularity among parents and educators for early childhood learning.
   * **"The Psychology of Money" by Morgan Housel** follows with **67,786 customer ratings**, reflecting its strong appeal in the personal finance and self-improvement genre.
2. **Duplicate Entries**:
   * The data contains duplicate records for each book, possibly due to multiple sellers or formats (e.g., hardcover, paperback, e-books). These duplicates should be merged for more accurate insights.
3. **Genre Popularity**:
   * The most popular books fall under distinct but impactful genres:
     + **Self-help and Habit Formation** (*"Atomic Habits"*).
     + **Children's Educational Material** (*"My First Library"*).
     + **Personal Finance** (*"The Psychology of Money"*).
   * This highlights the strong reader preference for practical, actionable, and educational content.
4. **Wider Reach and Influence**:
   * **"Atomic Habits"**' massive customer engagement indicates its global reach and its role as a cornerstone in the self-help category.
   * The strong performance of *"My First Library"* signifies a high demand for accessible and affordable early learning resources.
   * *"The Psychology of Money"*’s popularity underscores a growing interest in understanding financial behaviors and planning.

**Recommendations:**

1. **For Publishers and Sellers**:
   * Leverage the popularity of these books by focusing on promotions during key sales events (e.g., holidays, back-to-school seasons).
   * Ensure clarity in listings to avoid duplication and confusion for buyers.
2. **For Authors**:
   * Authors in similar genres can study the success of these books to identify key themes and strategies, such as simplicity in writing, actionable insights, and audience targeting.
3. **Data Refinement**:
   * Consolidate duplicates in datasets to provide a more accurate ranking and analysis.
   * Add pricing and regional performance data to understand broader trends.

**Concluding Insight:**

The dataset reveals the dominance of actionable, practical, and educational books across diverse genres. These books’ high customer ratings demonstrate their value and effectiveness in meeting readers' needs, making them benchmarks in their respective categories.

1. **Complete the above tasks using at least three separate web craping packages and compare their performance (processing time).**

**Selenium**

A screen shot of a computer program

Description automatically generated

**Playwright**

**A screen shot of a computer program

Description automatically generated**

**compare their performance (processing time).A screenshot of a computer program

Description automatically generatedA screen shot of a computer program

Description automatically generated**

The performance comparison among three web scraping methods—Selenium, Playwright, and BeautifulSoup—shows nearly identical execution times, all completing the task within approximately 2 seconds.

* Selenium took slightly longer (2.01 seconds) than the other methods, likely due to its reliance on a browser-based execution environment, which can introduce overhead.
* Playwright and BeautifulSoup both completed in exactly 2.00 seconds. While Playwright uses a browser-driven approach similar to Selenium, its modern asynchronous architecture might explain its efficiency. BeautifulSoup, being a lightweight library, performs similarly well in this context.

This suggests that for simple tasks, all three methods are comparable in performance, but the choice should depend on the complexity of the scraping project:

* Use BeautifulSoup for simpler, static websites.
* Opt for Selenium or Playwright when handling dynamic, JavaScript-heavy pages. Playwright might be preferable for newer projects due to its asynchronous capabilities and faster execution.

7.**Conclusion from the EDA and Visualization:**

1. **Highest Priced Books**:
   * The most expensive books in the dataset are primarily educational materials, such as sample papers and question banks from publishers like Educart, Oswaal, and PW (Physics Wallah).
   * These books cater to CBSE and competitive exam preparations, indicating that high prices are justified by their niche audience and perceived value.
   * "PROFESSIONAL'S AIBE Exam 25 Bare Acts" series dominates the top-priced category, reflecting the high demand for legal reference materials.
2. **Top Rated Books**:
   * The highest-rated books include *"Amma Diarylo Konni Pageelu"* by Ravi Mantri and *"The Satvic Revolution"* by Subah Saraf, both scoring an impressive average rating of **4.8**.
   * These books are highly appreciated for their unique themes, ranging from health and wellness to personal experiences. They reflect diverse reader interests.
   * Despite their high ratings, the total number of customer ratings is moderate compared to popular mainstream titles.
3. **Popular Books (Customer Rated)**:
   * The most popular books, based on the number of customer ratings, include *"Atomic Habits"* by James Clear, *"My First Library"* by Wonder House Books, and *"The Psychology of Money"* by Morgan Housel.
   * These books have garnered thousands of customer reviews, indicating widespread popularity and mainstream appeal. For instance, *"Atomic Habits"* leads with over **98,000 ratings**, highlighting its global recognition and impact.

**Key Insights:**

* Educational and exam preparation books dominate the high-price segment, showcasing their importance in specialized markets.
* Highly rated books often represent niche genres or themes, catering to specific audiences with strong appreciation for quality content.
* Popular mainstream books, with their vast number of customer ratings, dominate sales across various categories, appealing to a broader audience.

This analysis demonstrates the diversity in the book market, where high-priced educational materials coexist with highly rated niche books and globally recognized mainstream titles.