Phase 1 – Project Idea Submission

Project Title: Web Scraping & Analysis of Books from Books to Scrape

1. Website to Scrape

For our project, we've chosen to scrape data from the Books to Scrape website. This site provides a variety of books, including their ratings, prices, and categories, making it ideal for our analysis.

---

2. What Data Will Be Collected

From the Books to Scrape site, we will collect the following details for each book:

- **Book Title**
- **Price**
- **Star Rating**
- **Availability (In stock or not)**
- **Category/Genre**
- **Product Description**
- **UPC (Product Code)**
- **Number of Reviews**
- **Book URL** (optional for reference)

This dataset will allow us to explore various aspects of the books' market trends, ratings, and pricing.

---

3. Project Objective

The goal of this project is to analyze the books on the website to discover:

- The **relationship between price and star rating** for books.
- The **distribution of books by category/genre**.
- Which **categories** have the **highest-rated** or **most-reviewed** books.
- Identify the books with the **highest reviews or prices** (top N books).

These insights will be valuable for **book lovers** looking to discover trends in the market.

---

4. Data Cleaning & Processing

Once the raw data is extracted, we will clean and organize it using **Pandas**. The data will be converted into a structured **DataFrame**, and saved as a **JSON file** for easy future use.

We will also apply **Regular Expressions (Regex)** to handle text cleaning tasks like:

- Removing unwanted characters or formatting.
- Extracting and cleaning text fields (like descriptions).
- Standardizing date formats (if applicable).

The result will be a clean and ready-to-analyze dataset.

---

5. Data Analysis & Visualization

Using **Seaborn** and **Matplotlib**, we will perform both statistical and visual analysis to uncover meaningful trends. Our focus will be on answering questions like:

- What is the **relationship between price and star rating** for the books?
- What is the **distribution of books by category/genre**?
- Which **categories** have the **highest-rated** or **most-reviewed** books?
- Which books have the **highest reviews or prices**?

Planned Visuals

- **Bar Charts** for top genres, book categories, and rating distributions.
- **Scatter Plots** to show relationships between price and star ratings.
- **Pie Charts** to show the distribution of books across categories.
- **Word Clouds** to highlight frequent keywords from book descriptions or categories.
- **Box Plots** to examine price distributions across categories.

---

6. Data Storage

After processing, we'll save the final dataset as a **JSON file**. This format is easy to work with, portable, and can be reused in future projects, web apps, or databases.

---

7. (Optional) Web App Deployment with Streamlit

If time allows, we'd like to build a simple **Streamlit app** to showcase our findings interactively. Users could:

- Browse visual summaries and trends.
- Filter results by genre, price range, or rating.
- Explore the relationships between different book attributes in a dynamic way.

This would be a bonus step to improve the presentation and make the project more engaging.