Project Deliverable 1: Data Collection & Preprocessing

Instructions

- Submit your project iteration through Google Classroom by the given deadline. Late submissions will not be accepted.
- Submit your code as a single Jupyter/Google Colab notebook, following the filename format:
  roll_number1 & roll_number2.ipynb/.py
  Also submit a PDF report named:
  roll_number1 & roll_number2.pdf
- The report must clearly mention names, roll numbers, and sections of both group members.
- Ensure your notebook includes proper comments and visible output for all major steps.
- Only one member of the pair should upload the project iteration.
- Plagiarism is strictly prohibited. Any detected plagiarism will result in zero marks for the entire project (worth 10 absolute marks).
- You must fully understand your code and implementation, as a viva will be conducted during evaluation.

Objective:
For your first deliverable, you are required to collect and preprocess data according to your selected project. The specific tasks differ based on the project type.

If you are working on the *Smart Price Recommender* project:

You must perform **data scraping** in this deliverable.

Your submission should include:

1. **Code for Data Scraping:**
   - Well-structured, documented, and executable code clearly showing the scraping process.
2. **Scraped Data:**
   - The dataset obtained through your scraping efforts.
3. **Documentation (1–2 pages):**
   - **Approach:** Describe your scraping method (e.g., BeautifulSoup, Scrapy, API extraction).
   - **Sources:** Mention all sites scraped (e.g., Jalal Sons, Al-Fatah, Imtiaz, Carrefour, other local stores that are easily scrapable).
   - **Rationale:** Explain why you chose this approach—considering ease, scalability, and data accessibility.

For all other projects:

You are required to **explore, understand, and preprocess** your assigned dataset.
If your dataset is too large to handle practically, you may **work on a representative subset** instead.

Your submission should include:

1. **Code File:**
    - Clean, well-commented, and organized code implementing appropriate preprocessing steps.

2. **Report:**

    - A concise report covering the following sections:

        - **Dataset:** Description and source
        - **Preprocessing Techniques:** Methods used and their intuition
        - **Visualizations:** Basic insights or distributions
        - **Future Work:** Planned next steps for model implementation