**Project Documentation: E-Commerce Data Analysis**

**Overview**

This project, the **E-Commerce Data Analysis Script**, is designed to collect and analyze product information from a single online marketplace. The script performs a one-time data scrape to capture a snapshot of current product prices, brands, discounts, and ratings. The primary goal is to use this data to perform a competitive analysis focused on brand pricing, discount patterns, and reliability, rather than tracking historical price changes.

**Problem Statement**

In today’s competitive e-commerce environment, businesses face challenges in understanding how their products are positioned relative to competitors in terms of **price, discounting strategy, and perceived quality**. Without accurate and structured data, identifying low-price leaders, recognizing brands that rely heavily on promotions, and assessing brand reliability can be difficult.  
This project aims to bridge that gap by **automating data collection and analysis** from an online marketplace to provide actionable insights that support strategic decisions in marketing, pricing, and product positioning.

**Questions to Answer**

* **Which brand has the lowest average product price?**  
  Helps identify low-price leaders in the marketplace, crucial for competitive positioning and understanding market segments.
* **What is the average discount percentage across all products in the dataset?**  
  Provides a high-level view of the overall discount strategy on the marketplace, useful for gauging market price sensitivity.
* **Which brands have the highest average discount percentage?**  
  Reveals which brands are most reliant on promotional pricing, helping in counter-strategy development or understanding competitors’ marketing approaches.
* **Is a product's rating a good indicator of its price?**  
  By plotting price against rating, determines if higher-priced items are generally perceived as higher quality or if there’s a disconnect between price and customer satisfaction.
* **What is the reliability of each brand based on its average rating?**  
  Highlights which brands consistently deliver quality products, informing purchasing or marketing decisions.

**Methodology**

The analysis is performed using a Python script leveraging several popular libraries, following these steps:

1. **Web Scraping with requests and BeautifulSoup**  
   The script sends an HTTP GET request to a specified Jumia URL using requests, retrieves the HTML content, and parses it with BeautifulSoup to isolate individual product items.
2. **Nested Data Extraction**  
   Extracts product details like name and current price from the main listing page. For additional details (e.g., rating), the script follows each product’s link to its detail page.
3. **Data Cleaning and Structuring**  
   Converts text-based values (e.g., prices, ratings) to numerical data types. Extracts brand names from titles and organizes all details into a list of dictionaries.
4. **Data Analysis and Visualization**  
   Converts the structured data into a pandas DataFrame for analysis. Uses matplotlib to create bar charts, a pie chart, and a scatter plot to visually answer the research questions.
5. **Output**  
   Prints the final DataFrame to the console and can optionally save the data to a CSV file for future use.

**Tech Stack**

* **requests** – For making HTTP requests.
* **Beautiful Soup** – For parsing HTML and extracting data.
* **pandas** – For data manipulation and analysis.
* **matplotlib** – For creating visualizations.