Web Scraping & Data Analysis for E-Commerce Insights

Objective:

Candidates will extract, clean, and analyze data from e-commerce websites to generate meaningful insights. This project simulates real-world data tasks in a structured manner, developing web scraping, data processing, and visualization skills.

Project Scope:

The project consists of five phases:

1. Web Scraping – Extract product details from free-to-scrape sources.
2. Data Cleaning – Prepare the data for analysis by handling missing values and formatting inconsistencies.
3. Exploratory Data Analysis (EDA) – Identify key trends and insights.
4. Insights & Recommendations – Summarize findings in a report.
5. Presentation & Submission – Share findings via GitHub and a final report.

Week 1: Data Collection & Preparation:

Phase 1: Web Scraping (Day 1-3)
Candidates will collect data from at least one of the following sources:

- Books to Scrape (Books, prices, ratings, availability)
- Fake Store API (Electronics, clothing, furniture – JSON format)
- ScrapeMe (Gaming products, accessories)
- E-Commerce Dummy JSON (Simulated e-commerce data)
- Best Buy Web Scraper (Requires API key)

Tasks:

- Use Python with BeautifulSoup, Scrapy, or Requests.
- Extract product name, price, rating, availability, and review count (if available).
- Save data in a structured format (CSV or JSON).

Deliverable: A Python script that extracts and saves product data.

Phase 2: Data Cleaning (Day 4-5)
Tasks:

- Remove missing or duplicate values.
- Convert price columns to numerical values.
- Normalize ratings (e.g., convert star ratings to numbers).
- Handle outliers if applicable.

Deliverable: A cleaned dataset ready for analysis.

Week 2: Data Analysis & Presentation

Phase 3: Exploratory Data Analysis (Day 6-7)
Tasks:

- Analyze price distribution using histograms.
- Identify top-rated products.
- Determine correlations between price and rating.
- Summarize key trends (e.g., do higher-priced products have better ratings?).

Deliverable: A Jupyter Notebook with charts and summary findings.

Phase 4: Insights & Recommendations (Day 8-9)
Tasks:

- Identify average product prices.
- Determine the best-rated price range.
- Detect any seasonal pricing trends.
- Highlight outliers (extremely high or low-priced products).

Deliverable: A short report summarizing key findings.

Phase 5: Presentation & Submission (Day 10-12)
Tasks:

- Upload project files to GitHub with a README.md file.
- Present findings using tables and visualizations.
- Ensure all deliverables are well-documented.

Deliverable: A GitHub repository with the project code, dataset, and report.

Bonus Challenge (Optional - 3 Extra Points)

- Deploy a Streamlit dashboard for interactive data visualization.
- Automate scraping using Scrapy and schedule updates with Cron Jobs.

Submission Requirements

☑ GitHub Repository Link
☑ Dataset (CSV/JSON)
☑ Jupyter Notebook with Analysis
☑ Summary Report (PDF or Markdown)

Evaluation Criteria

- Web Scraping Accuracy (30%) – Successfully extracting relevant data.
- Data Cleaning & Processing (30%) – Properly handling missing or dirty data.
- Insightful Analysis (20%) – Extracting meaningful patterns and trends.
- Presentation & Communication (20%) – Clearly explaining findings.

This two-week project ensures a step-by-step learning experience, from data extraction to analysis and presentation.