WEB SCRAPING AND DATA ANALYSIS USING PYTHON

Assignment Report

Submitted by

Name: Vasavi Gangisetty

Roll Number: 23471A05DO

Branch: CSE

Section: C2

College Name: Narasaraopeta Engineering College

Web Scraping of water bottle data

A Flipkart Product Analysis Case Study

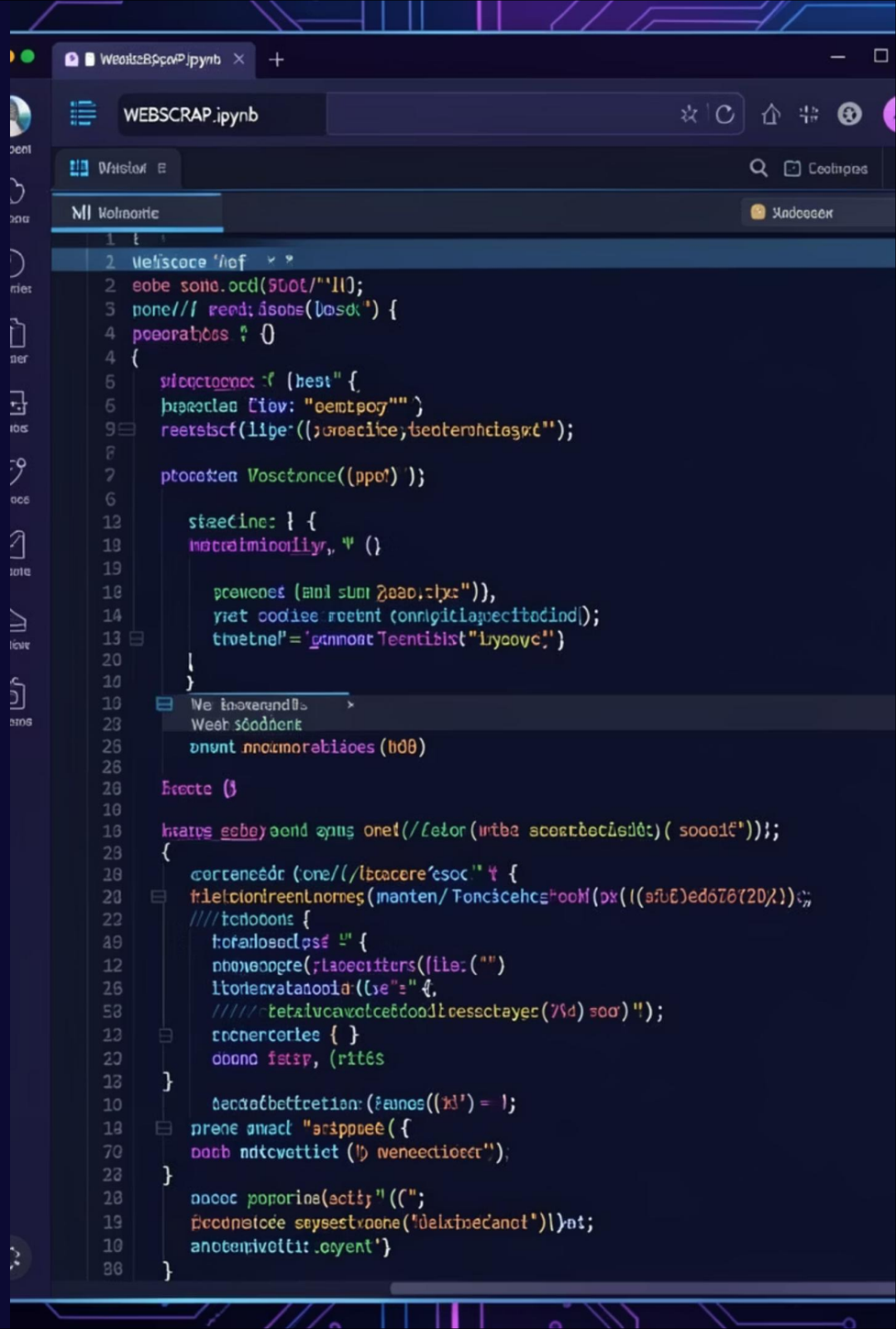
This presentation delves into a comprehensive web scraping and data analysis project, focusing on Flipkart product data. We'll explore the process from initial data extraction to insightful visualisations, all powered by Python.

OVERVIEW

File Overview: WEBSCRAP.ipynb

The `WEBSCRAP.ipynb` Jupyter Notebook is a powerful tool designed to scrape water bottle product data from Flipkart across multiple pages (1 to 10). It's a full-stack solution for data acquisition and preliminary analysis.

- **Extracts product information using robust web scraping techniques.**
- **Cleanses and structures raw data for optimal use.**
- **Stores data efficiently in Pandas DataFrames.**
- **Combines data from various pages into a single, cohesive dataset.**
- **Performs extensive exploratory data analysis (EDA).**
- **Answers 20 analytical questions with statistics and visualisations.**
- **Exports the refined dataset as a CSV file for further applications.**



OBJECTIVES

Key Objectives of This Analysis

Master Web Scraping

Demonstrate advanced web scraping techniques without relying on Selenium.

Collect Real-World Data

Gather authentic e-commerce data for practical analysis.

Analyze Product Metrics

Examine pricing, ratings, discounts, and availability patterns.

Practice Data Skills

Refine data cleaning, transformation, and visualisation capabilities.

Create Usable Datasets

Generate a clean dataset suitable for dashboards or Machine Learning tasks.

TECHNOLOGY STACK

Foundation: Python's Role in Data Science

Python serves as the backbone of this project, offering a versatile ecosystem essential for various data-related tasks. Its extensive libraries make it the go-to language for:

Web Scraping: Efficiently extracting data from websites.

Data Analysis: Performing complex computations and statistical analysis.

Visualisation: Creating compelling charts and graphs.

LIBRARIES

Requests Library: Fetching Web Content

Purpose

To send HTTP requests to web servers and retrieve raw webpage content. It acts as the bridge between our Python script and the target website.

Usage Example

```
import requestspage = requests.get(url)
```

This simple code snippet demonstrates how to initiate a GET request and store the response.

Definition

`Requests` is a Python library that simplifies making HTTP requests, allowing developers to programmatically interact with web services and download HTML for further processing.

LIBRARIES

BeautifulSoup (bs4): Parsing HTML

Purpose

To parse HTML and XML documents, enabling easy extraction of data by navigating the document as a tree structure.

Usage Example

```
from bs4 import BeautifulSoupsoup = BeautifulSoup(page.text, 'html.parser')
```

This transforms raw HTML into a navigable object.

Definition

`BeautifulSoup` is a Python library for pulling data out of HTML and XML files. It provides methods for searching and modifying the parse tree, making data extraction intuitive.

LIBRARIES

Regular Expressions (re): Text Cleaning

1

Purpose

To define and apply patterns for advanced text searching, manipulation, and cleaning, crucial for standardising scraped data.

2

Usage Example

```
import rere.findall(r"^\w+", text)
```

This example finds all word characters at the beginning of a string.

3

Definition

`re` is Python's built-in module for regular expressions, enabling powerful pattern matching operations like searching, splitting, and replacing strings based on complex rules.

LIBRARIES

NumPy & Pandas: The Data Handling Duo

NumPy

Purpose: Essential for high-performance numerical operations, especially with arrays and matrices. It forms the mathematical foundation for many data science libraries.

Usage Example: `import numpy as np`**Used for efficient calculations and handling large datasets.**

Pandas

Purpose: The cornerstone for data manipulation and analysis, providing powerful DataFrame and Series objects for structured data.

Usage Examples: `import pandas as pdpd.DataFrame()pd.concat()`**Used for creating, combining, and querying tabular data.**

LIBRARIES

Matplotlib: Visualising Insights

1

Purpose

To create static, animated, and interactive visualisations in Python, making complex data interpretable through various plots.

2

Usage Example

```
import matplotlib.pyplot as plt
```

The primary interface for creating plots.

3

Charts Created

- **Scatter plots to show relationships.**
- **Histograms for data distribution.**
- **Rating distributions to understand user feedback.**

4

Definition

`Matplotlib` **is a comprehensive library for creating static, animated, and interactive visualisations in Python, providing a vast array of plotting functions.**

DATA SCHEMA

Key Data Extracted: Product Features

For each water bottle product scraped from Flipkart, the following crucial details are extracted and stored:

Brand	Product brand name
Rating	User rating (1–5 stars)
Price	Selling price
Original_Price	MRP before discount
Stock	Availability status
Ratings_Reviews_Count	Number of ratings and reviews
Discounts	Discount percentage

Thank You