Assignment 1: Report on Web Scraping

1. Problem Definition:

The Assignment aims to perform web scraping on the Best Buy website to extract product information such as titles, prices, and ratings. The Assignment includes web scraping, data extraction, and saving the data to a CSV file.

Future Scope:

i. **Recommendation Systems:** By understanding which products are frequently purchased together (e.g., through association analysis), we can build recommendation systems that suggest related products to customers, improving their shopping experience.
ii. **Customer Sentiment Analysis:** The ratings and reviews data can be used for sentiment analysis to understand customer satisfaction and identify areas for product improvement.
iii. **Machine Learning:** The scraped data can be used to build predictive models, such as regression models to predict product prices based on various factors, or classification models to predict product ratings based on product features.
iv. **Natural Language Processing (NLP):** NLP techniques can be applied to customer reviews to extract valuable insights, perform topic modeling, and identify common issues or features that customers mention.

2. Objective:

The objective of this project is to scrap the data from a popular e-commerce platform, Best Buy, for further analysis. The goal is to collect detailed information about these products, including titles, prices, and ratings. This data will be further utilized for market research, analysis, and potential insights for pricing strategies, customer sentiment, and product recommendations for the Bestbuy website.

3. Methodology:

The methodology involves web scraping Best Buy's website using Python and the BeautifulSoup library to extract product data. It includes identifying and parsing relevant HTML elements, handling exceptions, and iterating through product pages.

Environment Setup:

The project uses Python environment with necessary libraries like BeautifulSoup, pandas, NumPy and requests.

Tools Used:

Python: The primary programming language used for web scraping and data manipulation.
BeautifulSoup: A Python library for parsing HTML and XML documents, used for extracting data from web pages.
Requests: A Python library for making HTTP requests to retrieve web pages.
Pandas: A powerful data manipulation library used for structuring and analyzing the scraped data.
NumPy: Used for numerical operations and data handling.

4. Solution:

Handling of HTML and CSS:

BeautifulSoup is utilized to navigate and extract data from the HTML structure of Best Buy's web pages. It allows for the identification and retrieval of specific elements like product titles, prices, and ratings based on their HTML tags and attributes.

Error Handling and Rate Limiting:

The code likely incorporates error-handling mechanisms to address issues like network errors, timeouts, or unexpected HTML changes on the website.

Data Storage Format:

The scraped data is stored in a structured format for further analysis. The data is saved in CSV (Comma-Separated Values) format with file name "bestbuy_data.csv."

As the data is values and not images, I have choose CSV over format. Also, CSV naturally represents data in a tabular structure, like a spreadsheet. This makes it suitable for datasets with rows and columns, such as database exports, Excel spreadsheets, or data collected from web scraping.

5. Challenges and Outlook

Handling road blockers:

I chose Best Buy over Amazon and Walmart for web scraping primarily because Amazon and Walmart had robust firewall and robot protection systems in place, posing significant challenges during data extraction. While I could have used Selenium to bypass these barriers, I opted for Best Buy due to its website's accessibility and feasibility for scraping without encountering such blockers. This choice allowed for a smoother and more straightforward scraping process, aligning better with my project's requirements, and avoiding potential legal complications associated with Amazon and Walmart's stricter terms of service.

Better Solution for Future:

Continuous Monitoring and Adaptation: Regularly monitor the target websites for changes in their structure or security measures. Develop adaptive scraping scripts that can adjust to alterations in the site's design or security protocols.

Scalability: Design the scraping solution to be scalable, capable of handling larger volumes of data efficiently. Consider using cloud-based solutions or distributed systems to manage increased workloads.

Data Quality Control: Implement data validation and cleansing processes to ensure the quality and accuracy of the scraped data. Address issues like duplicate entries or incomplete information.

Screenshot:

	title	price	rating
0	True Seating - Ergo Electric Height Adjustable...	$369.99	User rating, 4.2 out of 5 stars with 5 reviews.
1	Walker Edison - Industrial Modern End / Side T...	$46.99	User rating, 4.7 out of 5 stars with 27 reviews.
2	Sauder - Cottage Road Storage Coffee Table - B...	$270.99	User rating, 5 out of 5 stars with 3 reviews.
3	Aluratek - Adjustable Ergonomic Laptop Cooling...	$49.99	User rating, 4.6 out of 5 stars with 639 reviews.
4	X Rocker - Ocelot Gaming Desk - Black, Red, Blue	$139.99	User rating, 4.8 out of 5 stars with 17 reviews.
5	WorkSmart - Resin Table - Gray	$65.99	User rating, 4.6 out of 5 stars with 12 reviews.
6	SD Gaming - Overlord Curved Table - Black	$158.99	User rating, 4.8 out of 5 stars with 35 reviews.
7	Walker Edison - Huntsman Wood Dining Table - B...	$460.99	Be the first to write a review
8	Walker Edison - Round Rustic Coffee Table - Sl...	$96.99	User rating, 4.8 out of 5 stars with 6 reviews.
9	Walker Edison - 72" Rectangular Solid Pine Woo...	$419.99	User rating, 4.3 out of 5 stars with 4 reviews.
10	Walker Edison - Coffee Table with wicker stora...	$97.99	User rating, 4.6 out of 5 stars with 8 reviews.
11	Simpli Home - Hayward Side Table - Natural	$149.99	Be the first to write a review
12	Walker Edison - 48" Wood Modern Coffee Table -...	$158.99	User rating, 5 out of 5 stars with 2 reviews.
13	Serta - Harton Rustic Expandable C Side Table ...	$69.99	User rating, 5 out of 5 stars with 1 review.
14	Walker Edison - Rectangular Farmhouse Wood Din...	$1,249.99	Be the first to write a review
15	Walker Edison - Modern Rectangle End/Side Tabl...	$52.99	User rating, 4.8 out of 5 stars with 4 reviews.
16	Walker Edison - Rectangular Farmhouse Wood Din...	$1,159.99	Be the first to write a review
17	Walker Edison - Modern Bridge-Leg Dining Table...	$573.99	Be the first to write a review