DATA SCIENCE INTERNSHIP ASSIGNMENT REPORT

ALINTA BIJU
26-08-25

Introduction

The objective of this assignment was to collect product details from one of the e-commerce websites listed in the provided instructions.From the available options, I selected Limeroad as the website for data collection.

Data Collection Process

- **Limeroad** was chosen as the Indian fashion e-commerce site for this assignment.
- Since Limeroad uses infinite scrolling to display products, Selenium was used to automate scrolling and load additional products dynamically.
- The main listing page did not contain complete details (brand and product title). Therefore, product URLs were collected from the main page.
- Using the collected URLs, Selenium visited each product's detail page to extract the required attributes.
- A total of **310** product details were collected during this process.
- The collected product details include **Product name, Brand, Category, MRP, Discounted price, and Product URL**.
- The number of reviews attribute was also mentioned in the provided instructions but Limeroad does not contain this information.
- The extracted details were stored in a structured format using a Pandas DataFrame and saved to a csv file.
- This dataset was loaded into Pandas for preprocessing like standardizing brand names,converting MRP and discount price to numeric formats etc.
- After cleaning,dataset is saved again to a csv file.

Key findings from Data Analysis

- **Descriptive statistics**:A descriptive statistics analysis was conducted to summarize the key features of the product prices and ratings, with the results, including the mean, median, minimum, and maximum values, presented in Table.

	MRP	Discounted Price	Rating
Count	310	310	310
Mean	1652.48	675.52	4.16
Std	720.12	275.35	0.46
Min	299	236	2.5
25%	999	449.25	4
50%	1624	679	4
75%	2398	849	4.5
Max	3040	1579	5

- **Brand analysis**:The top five brands by number of products collected are Showoffff with 78 products, Ketch with 67 products, V-Mart with 30 products, Flick By Vmart with 23 products, and Beyou Fashion with 17 products.

- **Discount Analysis**:Among all the brands,**Ambi** offers highest average discount of approximately 82.9%,followed by **All Ways You** with an average discount of approximately 81.5%.

Challenges Faced

- Unlike many e-commerce websites that use paginated product listings, Limeroad uses infinite scrolling,which made scraping challenging.Traditional methods could not capture all products at once.So I implemented automated scrolling in Selenium to continuously load more products until sufficient data is collected.This process required extra effort and time compared to a normal page-based scraping method.

- Another challenge was extracting the product brand and title, as these details were not available on the main listing page despite using automated scrolling. They were only accessible within each product's individual page. To address this, I collected product URLs and used Selenium to visit each page to extract the title and brand.

Conclusion

This assignment provided hands-on experience in web scraping and data cleaning. Despite challenges such as infinite scrolling and extracting details from product detail pages, the required dataset was successfully collected, processed, and stored in a CSV file for further analysis.