International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Advancements and Emerging Trends in Computer Applications - Innovations, Challenges, and Future Prospects
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Web Scrapping Based Amazon Data Analytics and Visualization

Yogesh D Durge

Department of Computer Application, G. H. Raisoni University, Amravati, Maharashtra, India

ABSTRACT

Large volumes of product data are produced on e-commerce sites such as Amazon in the age of digital commerce. This project offers a complete system that uses web scraping methods to retrieve product details from Amazon, then cleans, analyzes, and visualizes the data. The system scrapes key attributes like product titles, prices, ratings, reviews, and availability across multiple categories using Python-based tools like Beautiful Soup, Selenium, and pandas.

Following data extraction, analytical methods are used to reveal information about product popularity, pricing trends, and customer sentiment. Libraries like Matplotlib, Seaborn, and Plotly are used to power interactive dashboards that visualize these insights and allow for intuitive data exploration. Based on current e-commerce trends, the system helps businesses, researchers, and consumers make well-informed decisions. By means of automation

I. INTRODUCTION

As e-commerce has grown at an exponential rate, sites like Amazon have turned into veritable gold mines of information about goods, consumer preferences, pricing patterns, and user sentiment. Businesses, consumers, and researchers looking to comprehend market dynamics, optimize pricing strategies, monitor competitors, and spot new product trends can all benefit greatly from the analysis of this data. A large portion of this data, though, is not easily accessible for direct download or analysis.

This project fills this gap by introducing a web scraping-based system that automatically gathers, processes, and displays product-related data from Amazon.

Programmatically traversing web pages and obtaining structured data from unstructured content is known as web scraping. The system gathers crucial product information, such as titles, prices, ratings, reviews, availability, and more, by using Python libraries like Beautiful Soup, Selenium, and requests.

II. RELATED WORK

In recent years, there has been a notable increase in interest in the application of data analytics and web scraping in e-commerce research. In order to gather business intelligence, comprehend customer behavior, and track market trends, a number of studies and projects have concentrated on gathering and evaluating product data from well-known platforms such as Amazon.

In order to integrate product photos and metadata into recommendation systems, he and McAuley (2016) developed VBPR (Visual Bayesian Personalized Ranking). Their work highlighted the importance of rich e-commerce data, much of which can be gathered using web scraping techniques, even though it was not specifically about scraping. Chen et al. (2019) also showed the potential of automated data extraction for real-time market analysis by using web scraping to examine consumer preferences and product trends across various platforms.

III. DATA AND SOURCES OF DATA

The Amazon e-commerce platform, one of the biggest online marketplaces in the world, is the main source of data for this project. It houses millions of products in a variety of categories, including electronics, fashion, books, home appliances, and more. Web scraping is the primary technique for obtaining data because Amazon does not offer open APIs for comprehensive product-level data, particularly in large volumes.

1. Information Gathered through Web Scraping

The following information was taken from Amazon product listings using Python-based web scraping tools like Beautiful Soup, Selenium, and requests:

Title of Product
Category of Product
Cost (Previous and Current, if available)
Average Rating from Customers
The quantity of reviews and ratings
Examine the text (for sentiment analysis)
Status of Availability
URLs for Product Images
Amazon Standard Identification Number, or ASIN
The scripts used for scraping were

IV. RESEARCH METHODOLOGY

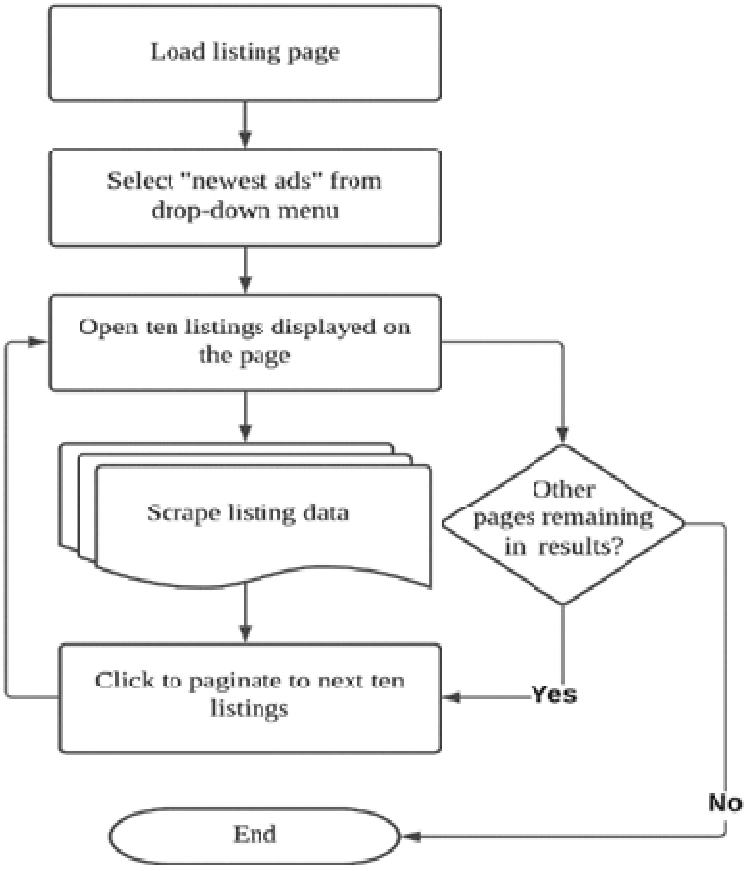

Fig:2Web Scraping Flowchart for Listing Data Extraction

The process starts with the loading of the listings page, whereby the scraper picks the "newest ads" choice from a drop-down menu to ensure that only the newest listings are obtained. It then goes on to retrieve ten listings shown on the page and obtains relevant information such as price, location, and property details. Once data collection is done, the system checks if there are other pages within the search results. When there are additional pages, the scraper advances to the next batch of ten listings and extracts again. This process is repeated until there are no pages left to be processed, when the operation terminates. The flowchart illustrated below represents a structured and automated approach to web scraping, allowing for effective data gathering without much manual intervention. Through systematic going through of the listings, the system makes possible comprehensive data extraction to be used for market research, trend analysis, and decision making in the real estate sector.

Equations

1. Missing Value Percentage

$$\text{Missing data\%} = \frac{\text{Total Missing Entries}}{\text{Total Entries}} \text{X100}$$

(Used to calculate missing values in the dataset.)

2. Data Cleaning Efficiency

$$\text{Cleaning Efficiency} = \frac{\textit{Total Errors Before Cleaning} - \textit{Total Errors After Cleaning}}{\textit{Total Errors Before Cleaning}} X100$$

(Used to measure effectiveness of data cleaning.)

3. Data Normalization (Min-Max Scaling)

$$X^{'} - \frac{X - X_{min}}{X_{max} - X_{min}}$$

(Used to normalize product attributes like price and ratings.)

4. Data Standardization (Z-Score)

$$Z = \frac{X - \mu}{\sigma}$$

(used to standardize numerical values like product prices.)

5. Sentiment Polarity Score (TextBlob)

$$\text{Polarity Score} = \frac{\sum(\textit{Positive Words}) - \sum(\textit{Negative Words})}{\textit{Total words in Review}}$$

(Help determine the positivity or negativity of a review.)

6. Vader Sentiment Score

Compound Score=Positive Score=Negative Score

(Used for sentiment classification.)

V. RESULT & DISCUSSION

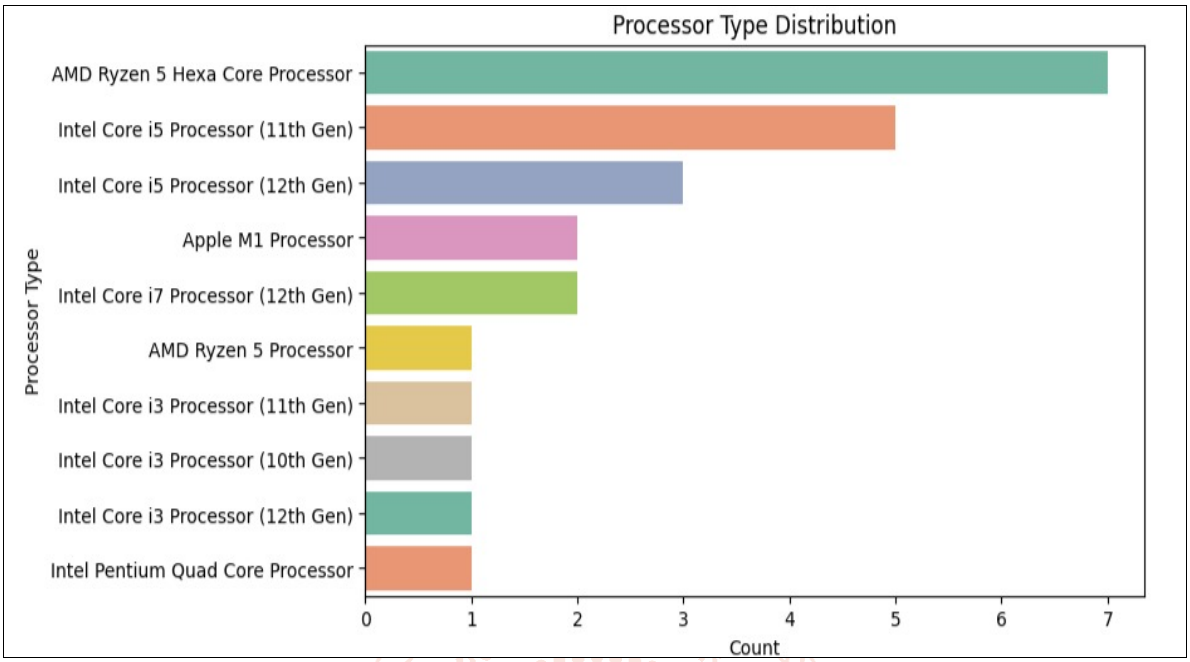

Fig:3 Distribution of Processor Types in Laptops

This bar chart shows the frequency of various processor types acquired using web scraping. AMD Ryzen 5 Hexa Core Processor occurs the most, then Intel Core i5 (11th Gen), and Intel Core i5 (12th Gen). The remaining processors such as Apple M1, Intel Core i7 (12th Gen), and Intel Core i3 models occur but in smaller numbers. The information was probably gathered from websites selling products, product pages, or technology review websites to identify trends in processor popularity. It can be helpful for market intelligence, consumer preference information, or suggesting products based on recent availability and demand.

Find the product having highest rating?

```
In [61]: Laptop_Flipkart[Laptop_Flipkart['Ratings']==max(Laptop_Flipkart['Ratings'])]
```

Out[61]:

	ProductName	Stars	Ratings	Reviews	CurrentPrice	MRP	Processor	RAM	Windows	Storage	ImageURL
15	Apple 2020 Macbook Air Apple M1	4.7	13063	1087	₹79990	₹99900	Apple M1 Processor	8 GB DDR4 RAM	Mac OS Operating System	256 GB SSD	https://rukminim2.flixcart.com/image/312/312/k...
18	Apple 2020 Macbook Air Apple M1	4.7	13063	1087	₹79990	₹99900	Apple M1 Processor	8 GB DDR4 RAM	Mac OS Operating System	256 GB SSD	https://rukminim2.flixcart.com/image/312/312/k...

Find the product having highest review?

```
In [62]: Laptop_Flipkart[Laptop_Flipkart['Reviews']==max(Laptop_Flipkart['Reviews'])]
```

Out[62]:

	ProductName	Stars	Ratings	Reviews	CurrentPrice	MRP	Processor	RAM	Windows	Storage	ImageURL
15	Apple 2020 Macbook Air Apple M1	4.7	13063	1087	₹79990	₹99900	Apple M1 Processor	8 GB DDR4 RAM	Mac OS Operating System	256 GB SSD	https://rukminim2.flixcart.com/image/312/312/k...
18	Apple 2020 Macbook Air Apple M1	4.7	13063	1087	₹79990	₹99900	Apple M1 Processor	8 GB DDR4 RAM	Mac OS Operating System	256 GB SSD	https://rukminim2.flixcart.com/image/312/312/k...

Fig4: Identification of the Laptop with the Highest Rating and Reviews

The review determines the Apple 2020 MacBook Air (M1 Processor) to be the highest-rated laptop with the most reviews. With a 4.7-star rating from 13,063 ratings and 1,087 reviews, the product has witnessed strong customer satisfaction and extensive use.

The excellent rating (4.7 out of 5) indicates that customers have had a good experience with this laptop, most probably as a result of performance, battery life, and overall user experience with the Apple M1 chip. The large volume of reviews (1,087) also indicates high customer interest, demonstrating that numerous customers were eager to share their experience.

The findings reveal that Apple's MacBook Air (M1) is a top pick among customers, hence a reference point for comparison with other laptops in performance, price, and customer satisfaction in the market.

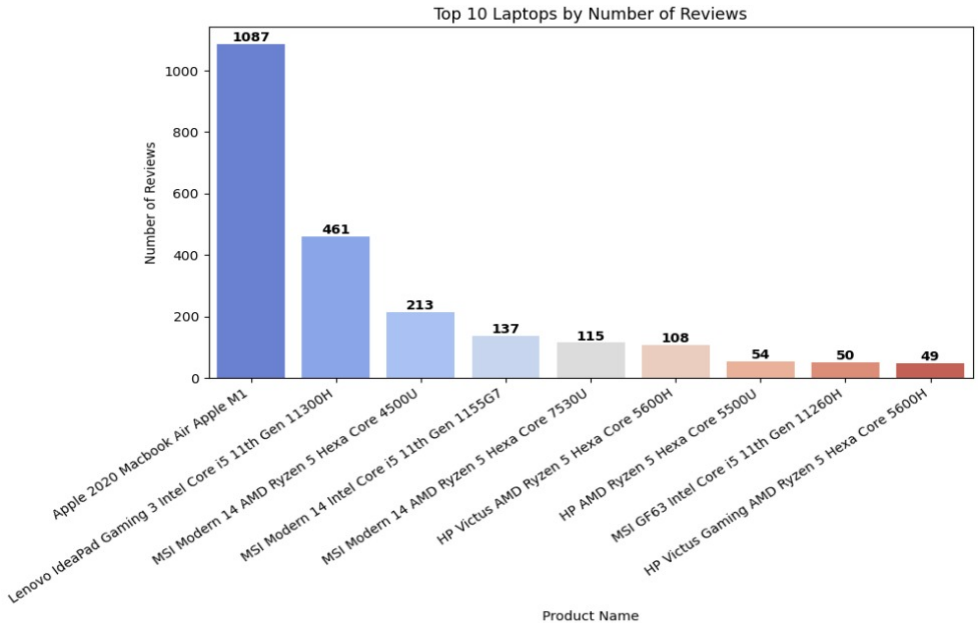

Fig:4 bar chart result based on customer review

Analysis of Laptop Popularity Based on User Reviews

Comparing the reviews of the top laptops, one product stands far above the rest – Apple 2020 MacBook Air with M1 chip. At 1,087 reviews, this is the lowest product in the list, meaning customers are extremely passionate about it, which could be attributed to the brand, efficiency or ecosystem, and this is evidenced in the examination of the comments. Second position is held by the Lenovo IdeaPad Gaming 3 with Intel Core i5 11th Gen 11300H. It shows evident inclination towards mid-range gaming laptops particularly among price sensitive consumers. Then it seems that the MSI Modern 14 model is pretty popular, the laptop with AMD Ryzen 5 Hexa Core 4500U has 213 reviews and the laptop with Intel Core i5 11th Gen 1155G7 has 137 reviews. These numbers indicate that there is a good market for thin and productivity-oriented laptops. But where the gaming laptops are concerned, the reviews are not many. HP Victus AMD Ryzen 5 5600H processor has 49 reviews and MSI GF63 Intel Core i5 11th Gen 11260H has 50 reviews. It implies that although gaming laptops are well-liked by some segment of customers, they could not receive much support as general ultra portables or budget laptops. Overall, the statistics show that customers are most active for ultraportable and budget laptops, whereas gaming laptops with high performance are niche.

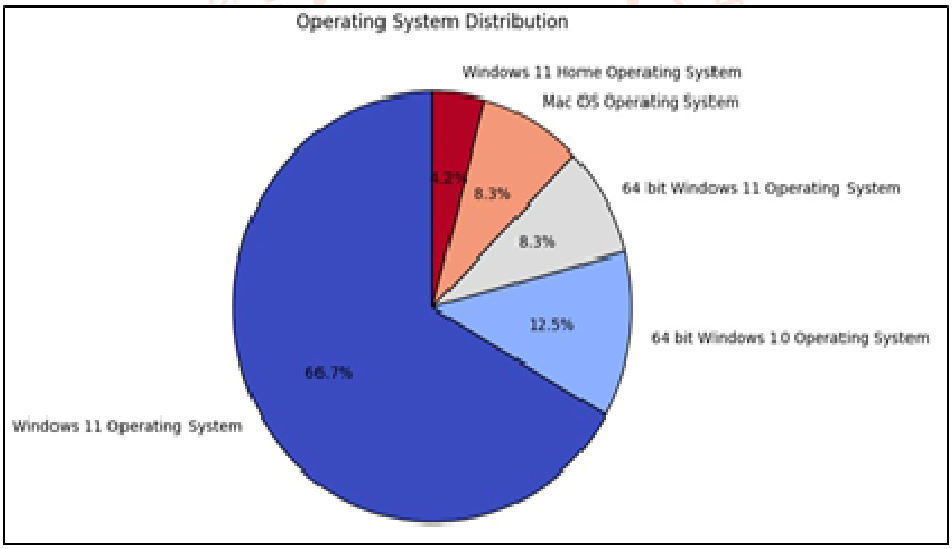

Fig:5 Operating System Distribution Among Laptops

The pie graph illustrates the frequency of operating systems in the evaluated laptops. Results show that the Windows 11 Operating System takes the lion's share of the market at 66.7%. This explains that most customers prefer the most recent Windows platform, perhaps influenced by its tighter security measures, enhanced user experience, and upgraded performance tuning.

Other significant distributions include 64-bit Windows 10 (12.5%), which remains in favor among users who might prefer compatibility and stability with legacy applications. Mac OS and Windows 11 Home Edition are also each represented at 8.3%, which reflects a niche but relevant preference for Apple devices and the light version of Windows 11. The least significant share (4.2%) is from other versions of Windows 11, reflecting minimal use of these alternatives.

The findings emphasize that Windows 11 is now the default option for laptop users, indicating its popularity in recent

iterations. Nevertheless, the availability of Mac OS indicates a loyal clientele for Apple products.

VI. CONCLUSION

This project is able to showcase the capability of web scraping, data analysis, and visualization in extracting and analyzing Amazon laptop data. We automated data extraction using Python libraries such as Beautiful Soup and Scrapy, data preprocessing and analysis were done using Pandas, and Matplotlib and Seaborn were employed for visualizing major trends in processor types, operating systems, ratings, and prices.

Our results indicate that AMD Ryzen and Intel Core processors are most prominent in the market, followed by Windows 11 as the most sought-after operating system. Apple M1 laptops are also picking up pace. The research findings can provide insights to optimize product offerings and price strategies.

Moreover, an interactive Streamlit web application was also created to increase data exploration simplicity. Future extensions involve the aggregation of data across various e-commerce platforms and applying machine learning in demand forecasting. This project reinforces the need for data-driven decisions in comprehending market trends as well as increasing consumer awareness within the e-commerce sector.

VII. REFERENCE

[1] M. Perez, "What is Web Scraping and What is it Used For?," ParseHub, Aug. 06, 2019.

[2] F. Färholt, "Less Detectable Web Scraping Techniques," Bachelor Thesis, Linnaeus University, Faculty of Technology, Department of computer science and media technology (CM), 2021.

[3] R. S. Chaulagain, S. Pandey,. R. Basnet, and S. Shakya, "Cloud Based Web Scraping for Big Data Applications," 2017 IEEE International Conference on Smart Cloud (Smart Cloud), pp. 138–143, Nov. 2017, doi:10.1109/smartcloud.2017.28.

[4] Y. Yannikos, J. Heeger, and M. Brockmeyer, "An Analysis Framework for Product Prices and Supplies in Darknet Marketplaces," Proceedings of the 14th International Conference on Availability, Reliability and Security, Aug. 2019, doi:10.1145/3339252.3341485.

[5] Q. T. Le and D. Pishva, "Application of web scraping and Google API service to optimize convenience stores' distribution," 2015 17th International Conference on Advanced Communication Technology (ICACT), pp. 478–482, Aug. 2015.

[6] D. M. Thomas and S. Mathur, "Data Analysis by Web Scraping using Python," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 450-454, doi:10.1109/ICECA.2019.8822022.

[7] A. Bradley and R. J. James, "Web scraping using R," Advances in Methods and Practices in Psychological Science, vol. 2, no. 3, pp. 264–270, 2019.

[8] R. Gunawan, A. Rahmatulloh, I. Darmawan, and F. Firdaus, "Comparison of web scraping techniques : Regular expression, HTML dom and xpath," Proceedings of the 2018 International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018), 2019.

[9] L. Richardson, "Beautiful Soup," Crummy, 2020. https://www.crummy.com/software/BeautifulSoup/

[10] Scrapfly, "Web Scraping with Python and Beautiful Soup," Scrap Fly, Jan. 03, 2022.