ZKG INTERNATIONAL

ISSN: 2366-1313

WEB SCRAPING USING SELENIUM

[1]N. PAPARAYUDU, [2]P. MAHESH, [3]G. PADMA REDDY, [4]M. ABHIRAM

[1]Assistant Professor, Dept.of IT, TKR College of Engineering & Technology, Meerpet, Hyderabad, paparayudu.nagara@gmail.com

[2]BTech student, Dept.of IT, TKR College of Engineering & Technology, Meerpet, Hyderabad, patilmahesh4357@gmail.com

[3]BTech student, Dept.of IT, TKR College of Engineering & Technology, Meerpet, Hyderabad, paddureddygaddam@gmail.com

[4]BTech student, Dept.of IT, TKR College of Engineering & Technology, Meerpet, Hyderabad, Abhiramsonu0007@gmail.com

Abstract: *Selenium is a powerful tool in data science. One of its uses is to automate the collection of publicly available data from websites. With Selenium in Python, we can automate web browsers to access data on the website, collect and store in MySQL, CSV file, etc. Web scraping is a data mining technology that is commonly used for extracting unstructured data from different online sources and restructuring and converting acquired data into a structured form that can be further stored and analysed in a database. The benefit of a well-designed web scraper is that it automatically sifts through targeted data sources and form valuable information into a comprehensive dataset. There are different forms of web scraping including copy and pasting, text grabbing, HTML parsing, and others. A benefit of web scraping is that it simulates human interaction with a web page and can obtain attribute data from the web page itself. This is beneficial because it brings in pertinent information that is relevant to the topic assigned to look for and not scraping for erroneous information.*

Keywords: *Selenium, web browsers, Web scraping, text grabbing, HTML parsing*.

I. INTRODUCTION

Machine learning is operating today's technological marvels, including powerless vehicle driving, spaceflight, photography, and the popularity of speech. However, a data science expert may want a large number of statistics to create a robust and reliable machine domain version of such business problems. Data mining or information accumulation is a very primitive step in the IT knowledge life cycle. Depending on business

ISSN: 2366-1313

requirements, you may also need to collect statistics from sources such as SAP servers, logs, databases, APIs, online repositories, or the Internet. Particularly fast time. Web scraping can help us extract a great deal of information about customers, products, people, stock markets, etc[1].

Internet contain various information from various websites, where it is categorized, interlinked and freely available for everyone. Some data that is available on the web is presented in a format that makes it easier to collect and use it. For extracting relevant data from websites, it is very tedious task and time consuming to manually extracting it. For formatting this Web Scraping is used. Web Scraping is process of extracting relevant data from websites. This technique of automating the process of navigating through links, and then navigating and collecting the data from relevant websites. After automation, instead of manually coping the data from websites, Web Scraping will replicate the same task within a fraction of time. The various technique used for Web Scraping are Text pattern matching, HTTP programming, DOM parsing, Text Grepping, Vertical aggregation, Semantic Annotation recognizing, computer vision web-page analysis etc [2]. Most of required data is unstructured data in HTML format which is then converted into structures data in a spreadsheet or a database so that it can be used in various applications. Internet contain various information from various websites, where it is categorized, interlinked and freely available for everyone. Some data that is available on the web is presented in a format that makes it easier to collect and use it. For extracting relevant data from websites, it is very tedious task and time consuming to manually copy-paste it. For this Web Scraping is used. Web Scraping is process of extracting relevant data from websites. This technique of automating the process of navigating through links, and then navigating and collecting the data from relevant websites. After automation, instead of manually coping the data from websites, Web Scraping will replicate the same task within a fraction of time. The various technique used for Web Scraping are Text pattern matching, HTTP programming, DOM parsing, Text Grepping, Vertical aggregation, Semantic Annotation recognizing, computer vision web-page analysis etc. Most of required data is unstructured data in HTML format which is then converted into structures data in a spreadsheet or a database so that it can be used in various applications [3].

ISSN: 2366-1313

Web Scraping, i.e., the automated and targeted extraction of data, is a traditional technique to retrieve Web content at scale. A multitude of frameworks and Application Programming Interfaces to develop customized scrapers, as well as configurable ready-to-use scraping tools exist.

We also theorize that having an artificial random delay when scraping and randomizing intervals between each visit to a website would counteract some of the anti-scraping measures. Another, smaller aspect of our research was the legality and ethicality of scraping. Further thoughts and comments on potential solutions to other issues have also been include.

OBJECTIVE

It is a process of automating the extraction of data in an efficient and fast way. With the help of web scraping, one can extract data from any website, no matter how large is the data, on your computer. Tools for web scraping like Selenium can scrape a large volume of data such as text and images in a relatively short time. Web scraping allows you to acquire non-tabular or poorly structured data from websites and convert it into a usable, structured format, such as a .csv file or spreadsheet. With the help of web scraping, one can extract data from any website, no matter how large is the data, on your computer. extracting unstructured data from different online sources and restructuring and converting acquired data into a structured form that can be further stored and analysed in a database. The benefit of a well-designed web scraper is that it automatically sifts through targeted data sources and form valuable information into a comprehensive.

II. LITERATURE SURVEY

The survey on deduplication work with various algorithms tabulated them on the basis of algorithm, objective criteria, environment to which the works being performed. From the literature survey it is clear that, lot of work had been done already in deduplication but still it needs further development. (i.e) Deduplication need to establish with high level security and minimum space wastage.

Sudhir Kumar Patnaik et al. [4] Data are crucial to the growth of e-commerce in today's world of highly demanding hyper personalized consumer experiences, which are collected using advanced web scraping technologies. However, core data extraction engines fail because they cannot adapt to the dynamic changes in website content. This study investigates an

ISSN: 2366-1313

intelligent and adaptive web data extraction system with convolutional and Long Short-Term Memory (LSTM) networks to enable automated web page detection using the You only look once (Yolo) algorithm and Tesserac LSTM to extract product details, which are detected as images from web pages. This state-of-the-art system does not need a core data extraction engine, and thus can adapt to dynamic changes in website layout. Experiments conducted on real-world retail cases demonstrate an image detection (precision) and character extraction accuracy (precision) of 97% and 99%, respectively. In addition, a mean average precision of 74%, with an input dataset of 45 objects or images, is obtained.

D. T. Meyer et al. [5] Web scraping is a process of extracting valuable and interesting text information from web pages. Most of the current studies targeting this task are mostly about automated web data extraction. In the extraction process, these studies first create a DOM tree and then access the necessary data through this tree. The construction process of this tree increases the time cost depending on the data structure of the DOM Tree. In the current web scraping literature, it is observed that time efficiency is ignored. This study proposes a novel approach, namely UzunExt, which extracts content quickly using the string methods and additional information without creating a DOM Tree. The string methods consist of the following consecutive steps: searching for a given pattern, then calculating the number of closing HTML elements for this pattern, and finally extracting content for the pattern. In the crawling process, our approach collects the additional information, including the starting position for enhancing the searching process, the number of inner tags for improving the extraction process, and tag repetition for terminating the extraction process. The string methods of this novel approach are about 60 times faster than extracting with the DOM-based method. Moreover, using this additional information improves extraction time by 2.35 times compared to using only the string methods. Furthermore, this approach can easily be adapted to other DOM.

Belen Vela et al. [6] The growing amount of data on the Internet has led to a situation in which it is essential to process these data to generate new services with the specific aim of improving people's daily living conditions. Transport data is of the utmost importance, since everyday people have to move around to perform some daily tasks, such as going to work,

ISSN: 2366-1313

studying and shopping, and this means that the number of journeys by public transport grows daily. People with special needs make a large number of these trips, but they do not have sufficient information about the accessibility of the routes they want to take. Although there are numerous websites and applications that provide information on public transport services, most do not provide detailed information on the accessibility of the routes. We are, therefore, developing a technological framework for the processing, management, and exploitation of open data to promote accessibility to urban public transport. This is taking place within the framework of the Access@City project. This paper specifically focuses on the data extraction and processing of the existing information on the web concerning public transport and its accessibility for the generation of an open data repository in which to store this information.

In 2019 [7], COVID-19 quickly spread across the world, infecting billions of people and disrupting the normal lives of citizens in every country. Governments, organizations, and research institutions all over the world are dedicating vast resources to research effective strategies to fight this rapidly propagating virus. With virus testing, most countries publish the number of confirmed cases, dead cases, recovered cases, and locations routinely through various channels and forms. This important data source has enabled researchers worldwide to perform different COVID-19 scientific studies, such as modeling this virus's spreading patterns, developing prevention strategies, and studying the impact of COVID-19 on other aspects of society. However, one major challenge is that there is no standardized, updated, and high-quality data product that covers COVID-19 cases data internationally. we developed a toolset using cloud-based web scraping to extract, refine, unify, and store COVID-19 cases data at multiple scales for all available countries around the world automatically. The toolset then publishes the data for public access in an effective manner, which could offer users a real time COVID-19 dynamic dataset with a global view. Two case studies are presented about how to utilize the datasets. This toolset can also be easily extended to fulfill other purposes with its open-source nature.

Clairie Lauer et al. [8] We discuss the collection and coding processes, and demonstrate how they might be replicated with web scraping and machine coding. Results/discussion: We found that web scraping demonstrated an obvious

ISSN: 2366-1313

advantage of automated data collection: speed. Machine coding was able to provide comparable outputs to hand coding for certain types of data; for more nuanced and verbally complex data, machine coding was less useful and less reliable. Conclusions: Our findings highlight the importance of considering the context of a particular project when weighing the affordances and limitations of hand collecting and coding over automated approaches. Ultimately, a mixed-methods approach that relies on a combination of hand coding and automated coding should prove to be the most productive for current and future kinds of technical communication work, in which close attention to the nuances of language is critical, but in which processing large amounts of data would yield significant benefits as well.

III. PROPOSED SYSTEM

Web scraping is a data mining technology that is commonly used for extracting unstructured data from different online sources and restructuring and converting acquired data into a structured form that can be further stored and analysed in a database. The benefit of a well-designed web scraper is that it automatically sifts through targeted data sources and form valuable information into a comprehensive dataset. There are different forms of web scraping including copy and pasting, text grabbing, HTML parsing, and others. A benefit of web scraping is that it simulates human interaction with a web page and can obtain attribute data from the web page itself. This is beneficial because it brings in pertinent information that is relevant to the topic assigned to look for and not scraping for erroneous information. For example, Weng and his colleagues applied web scraper techniques to collect large-scale datasets of horticultural products information to predict the trend of price fluctuation with Auto Regressive Integrated Moving Average (ARIMA) and integrated recurrent neural network (RNN) model. Pawar and colleagues implemented a web scraper to search medicinal plants and relevant diseases in the India Ayurvedic system. Web scraping is widely used by epidemiological research and public health studies. By scraping and analysing text-based data from the Internet, researchers can successfully detect diseases and food hazards, as well as predict potential pandemics. For example, Pollett and colleagues used a web scraper as a tool to scrape unstructured Internet newswire data to timely detect outbreaks and epidemics from vector-borne diseases. Walid and his team scraped worldwide Twitter data for 2 years. By applying

ISSN: 2366-1313

sentiment analysis and natural language processing on Walid's data, they built a model to detect and predict cancer. In addition to diseases detection, web scraping has been adopted in food hazards detection and dissemination. By scraping the events related to food hazards from news and social media, Ihm and colleagues built a system to prevent and control food hazards in Korea. In addition, Majumder et al. utilized web scraped data collected by HealthMap coupled with Google Trend time series data to calculate the R0 and predict the outbreak level of Zika virus in 2015. Beyond scraping text-based data from Internet resources, images have been scraped as a valuable dataset to support public health research. For example, Li et al. scraped illicit drug dealer-related photos and posts from Instagram. With 3 different deep learning models applied, they detected 1129 drug dealers successfully.

IMPLEMENTATION

METHOD OF IMPLEMENTATION

- Data Scraping
- Making Data Set
- Data Processing
- Data Cleaning
- Data Analysis
- Uni Variate Analysis
- Bi Variate analysis
- Multi Variate analysis

Data Scraping:

Here we import some of the python libraries such as pandas, seaborn, matplotlib, numpy, Re,

Selenium, selenium web driver. we are going to need to install Selenium along with a few other packagesTo manage our webdriver, we will use webdriver-manager. You can use Selenium to control most popular web browsers including: Firefox, Internet Explorer, Opera, Safari, and Chrome. I will be using Chrome.

Making Data Set :

Here we will prepare a data set using some python libraries like Panadas. We will make a table format data

Data Processing :

Data processing is an iterative process and may require multiple rounds of cleaning and analysis before the data is in a usable format. The specific steps involved in data processing may vary depending on the type of data being scraped and the requirements of the project.

Data Cleaning :

Data cleaning is an important step in web scraping as it involves cleaning and

ISSN: 2366-1313

transforming the data obtained from the web into a usable format. Here are some common steps involved in data cleaning in web scraping.

Data Analysis :

1. Exploratory Data Analysis (EDA): Start with a preliminary analysis of the data, such as calculating summary statistics, creating visualizations, and identifying patterns in the data.

2. Data visualization: Create visualizations such as histograms, scatter plots, bar charts, etc. to help identify trends and patterns in the data.

Uni Variate Analysis

Univariate analysis is a method of analyzing a single variable in a dataset in web scraping. The goal of univariate analysis is to understand the distribution of the variable and identify any patterns, trends, or outliers.

Bi Variate Analysis :

Bivariate analysis is a method of analyzing the relationship between two variables in a dataset in web scraping. The goal of bivariate analysis is to understand the relationship between the two variables, including any patterns, trends, or correlations.

Multi Variate Analysis :

Multivariate analysis is a method of analysing the relationships between multiple variables in a dataset in web scraping. The goal of multivariate analysis is to understand the complex relationships between the variables and identify any patterns, trends, or correlations.

IV. RESULTS

ARCHITECTURE DIAGRAM

An architectural diagram is a diagram of a system that is to abstract the overall outline of the developing software system and the relationships, constraints and boundaries between components. It is important tool as it provided an overall view of the physical deployment of the software system and its evolution roadmap.

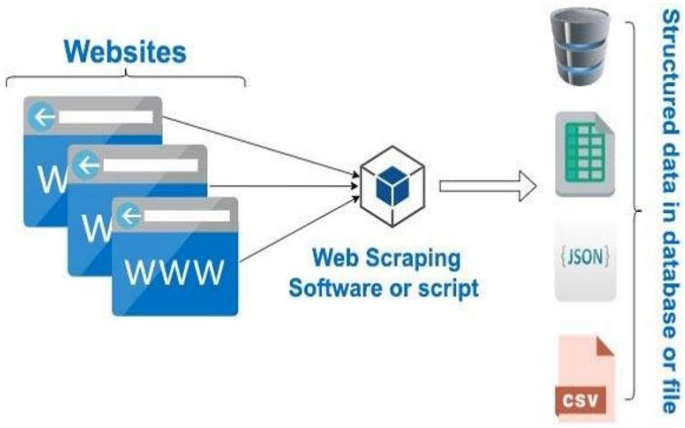

Fig.1 Web Scraping Structure

ISSN: 2366-1313

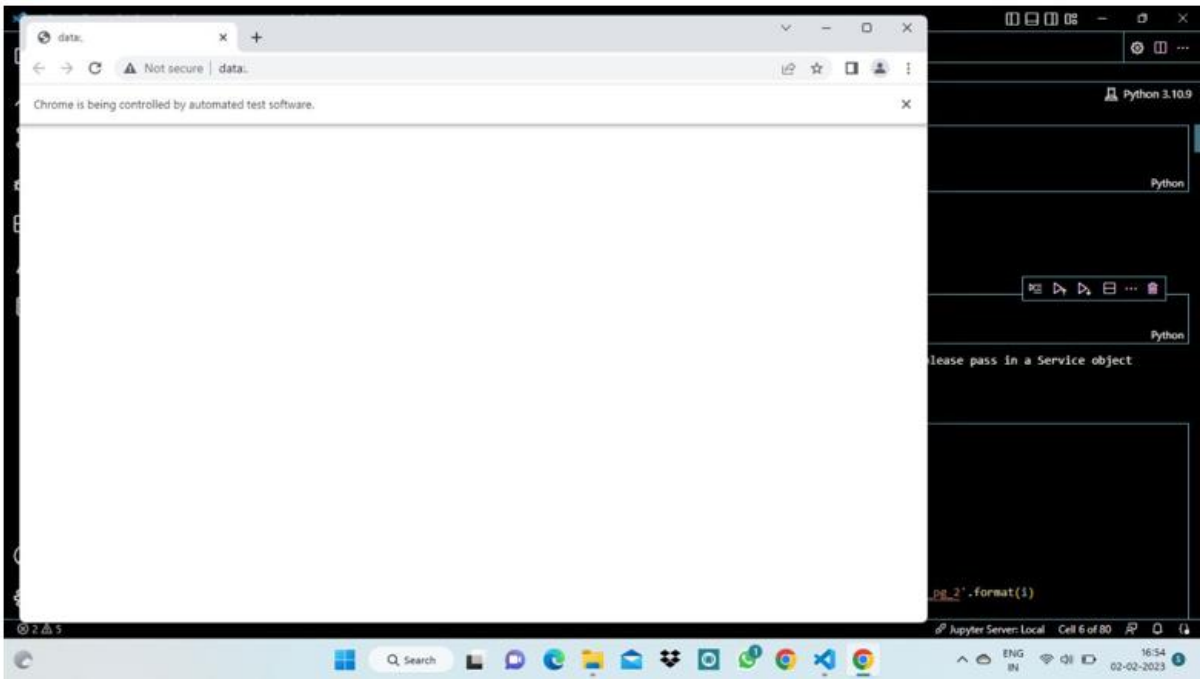

Fig.2 first output screen of the project is automation of the opening chrome browser

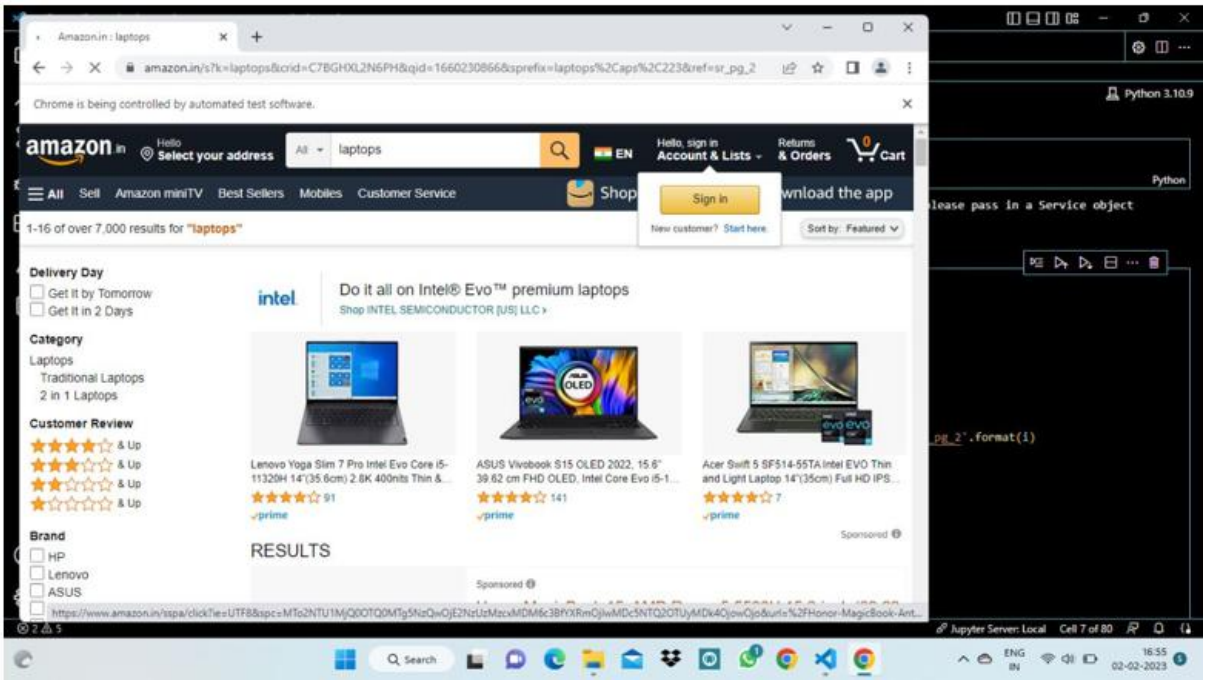

Fig.3 Then you will get below screen by taking amazon url link and searching for the product laptops. with the help of selenium elements, we will take the path of the particular element and can access through the selenium automation tool.

ISSN: 2366-1313

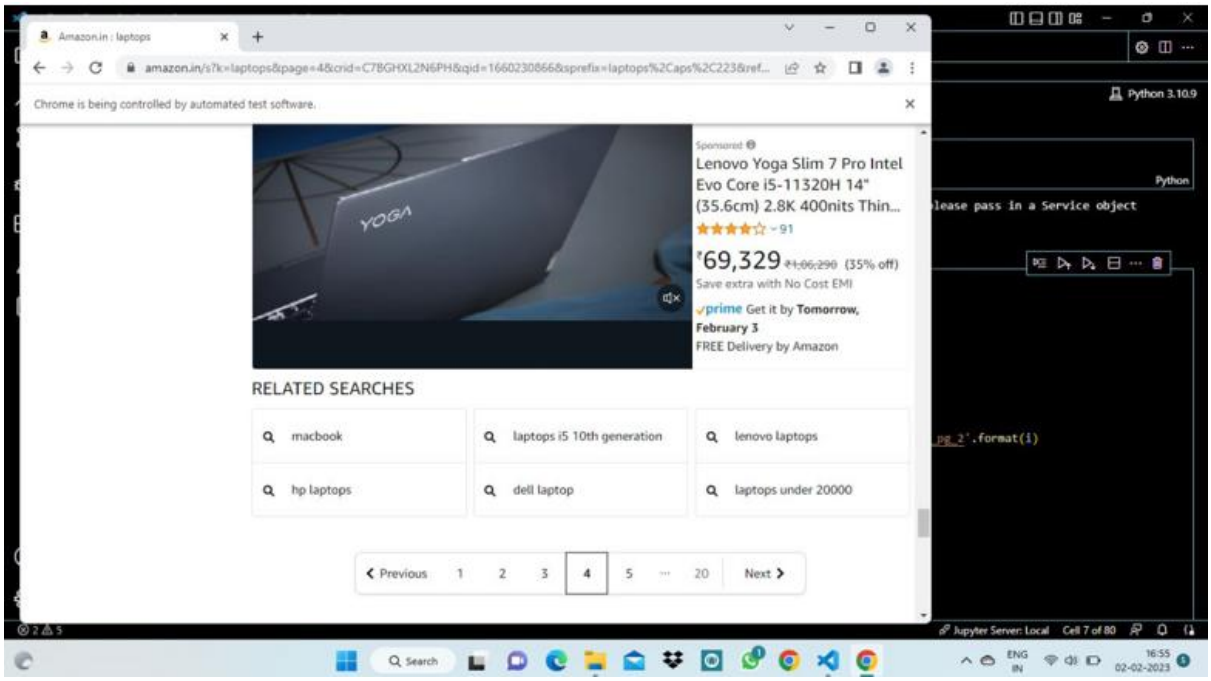

Fig.4 In this screen we can see that selenium tool automatically access the dataset through then amazon website. Here we can see that it will take the data upto 20 pages and collect the data

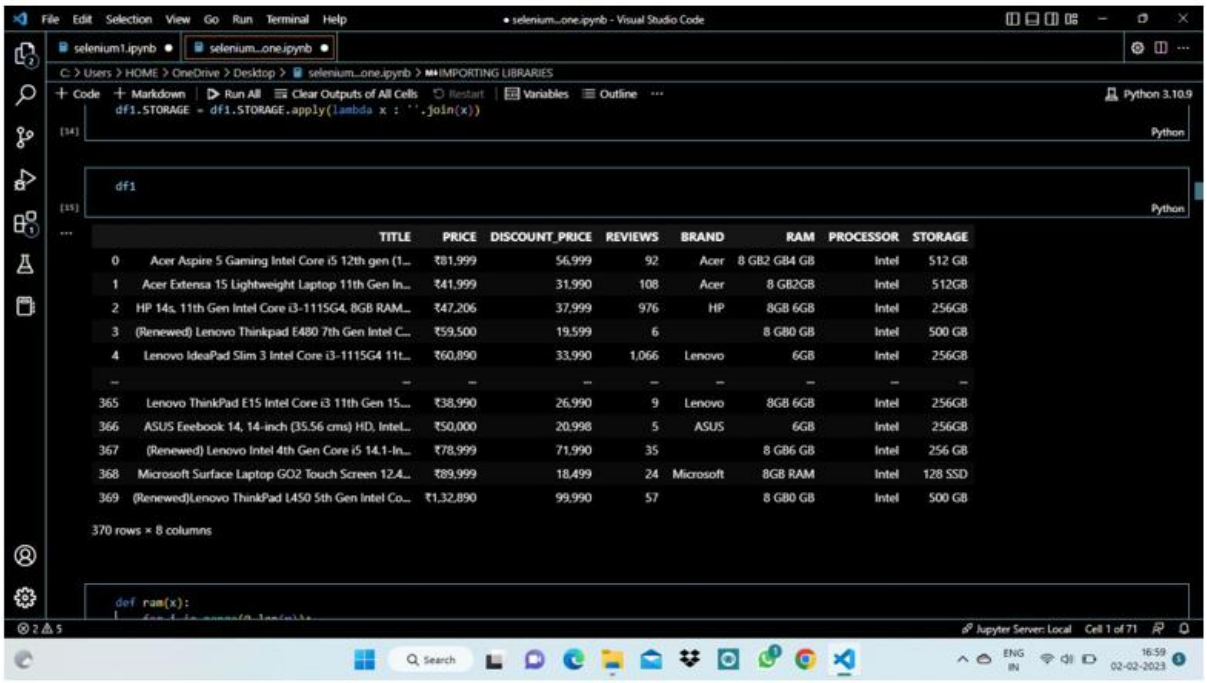

Fig.5 Here we can see that making of the dataset through python libraries that nis pandas.

ISSN: 2366-1313

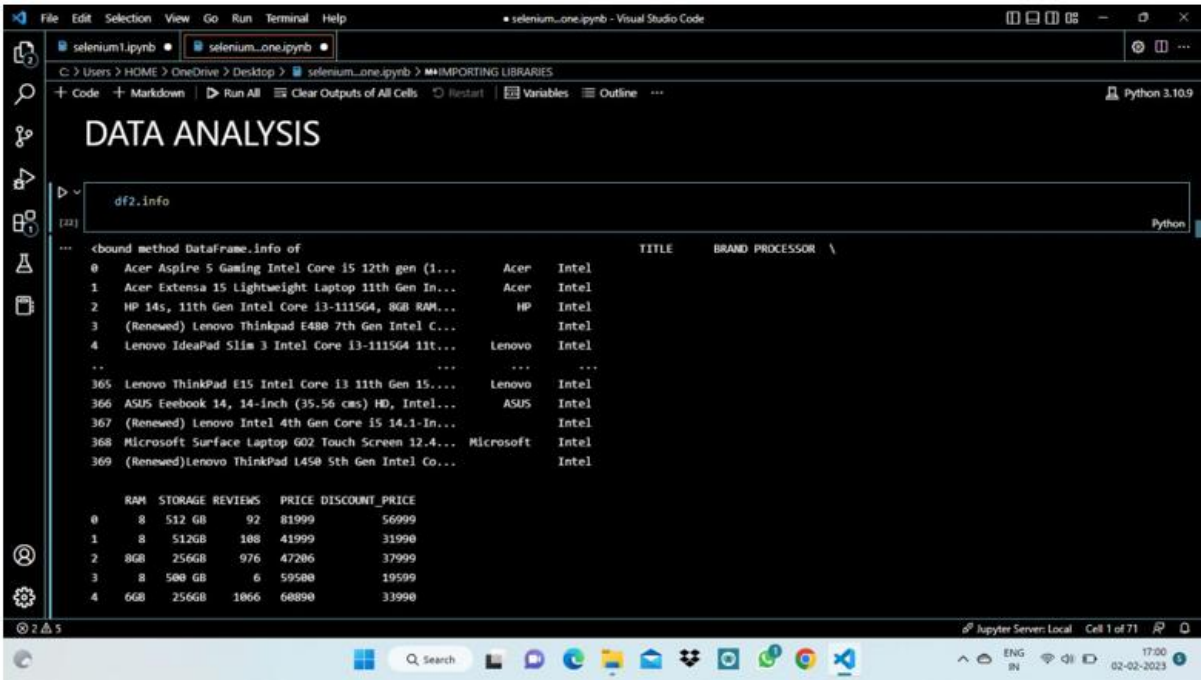

Fig.6 In this screen we can see that data analysis

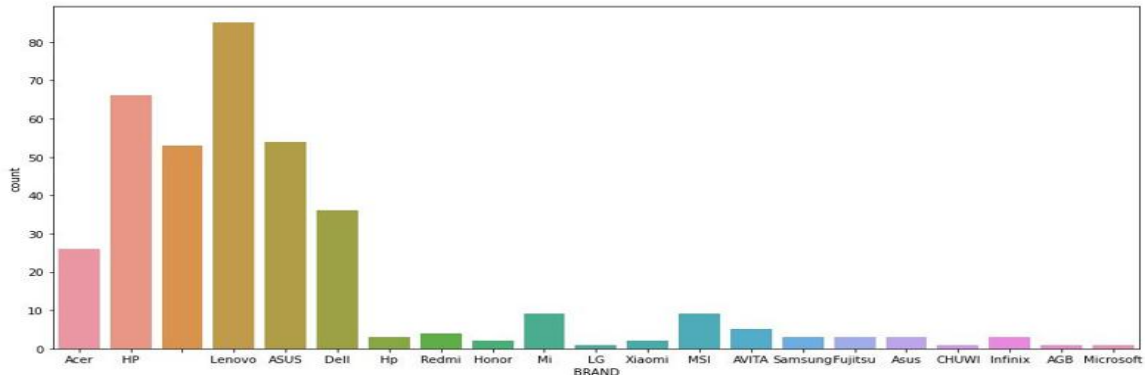

Fig.7 In this screen we can see that data analysis through Uni-Variate analysis graph COUNT PLOT

ISSN: 2366-1313

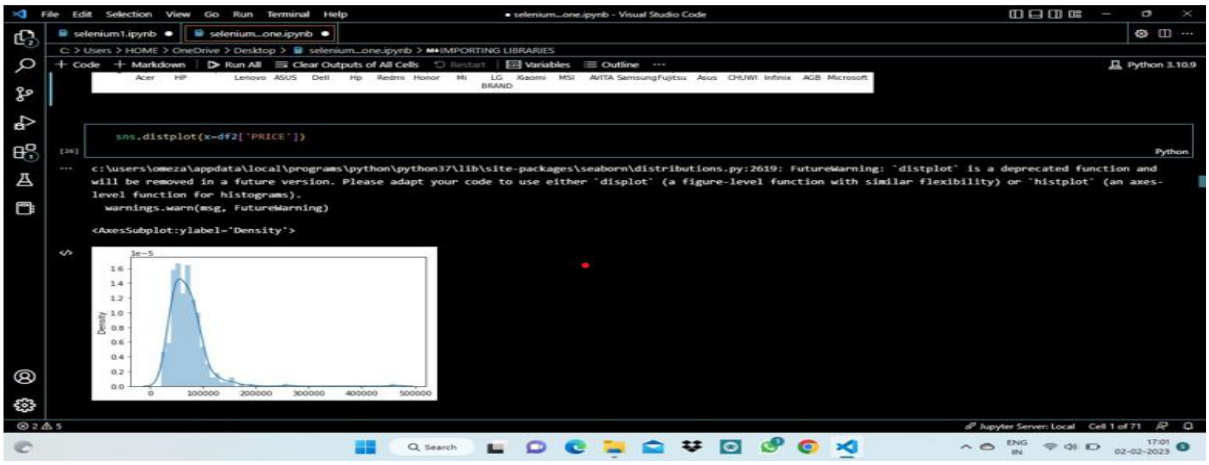

Fig.8 In this screen we can see that the Uni Variate analysis using graph DIST PLOT

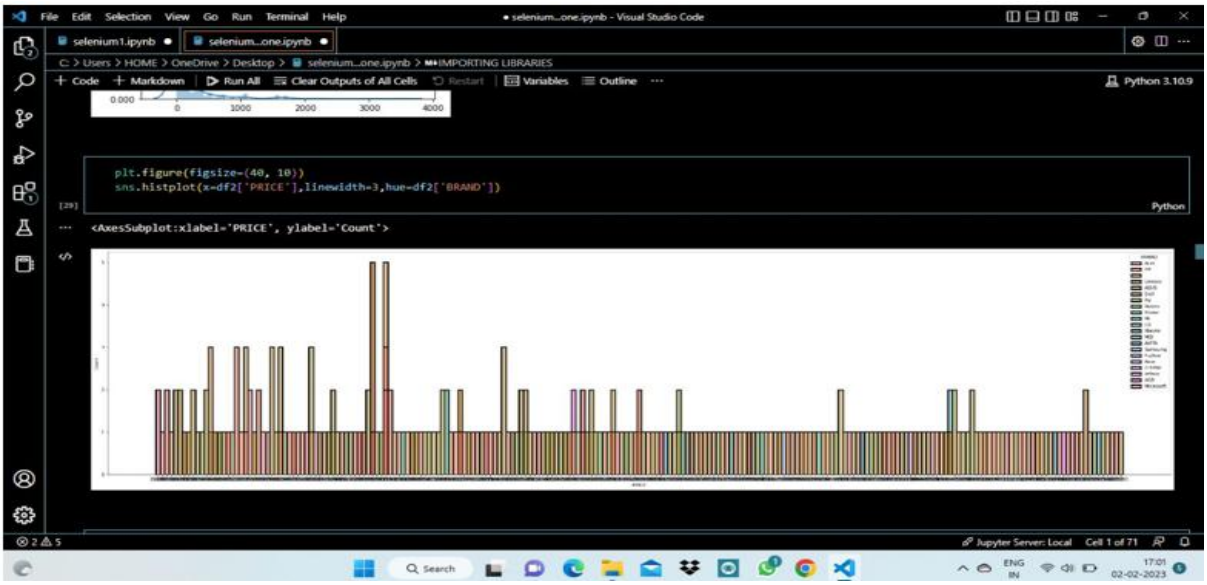

Fig.9 In above screen we can see the Uni Variate analysis through using then graph HIST PLOT

ISSN: 2366-1313

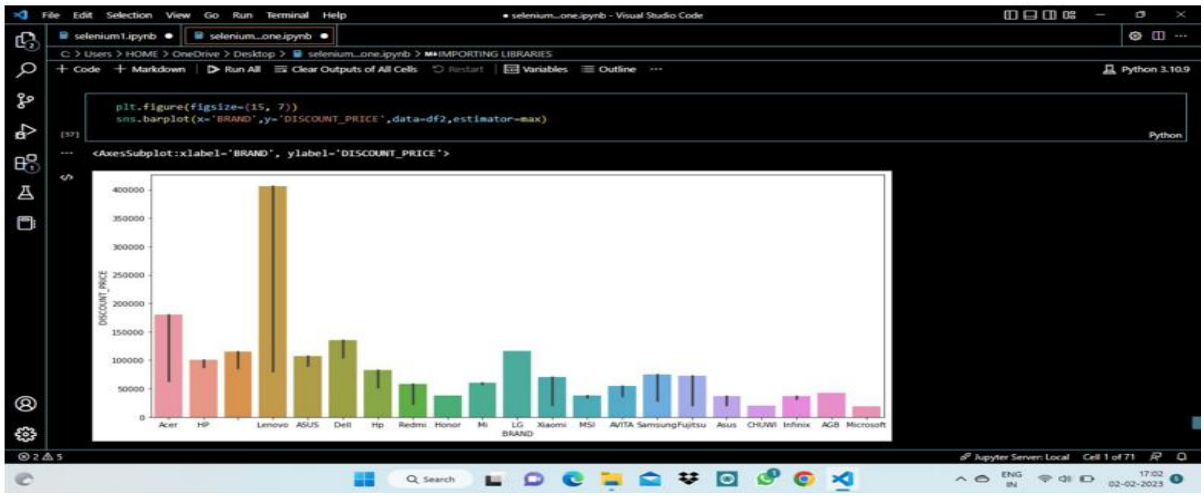

Fig.10 In this screen we can see that Uni Variate analysis using the graph BAR PLOT

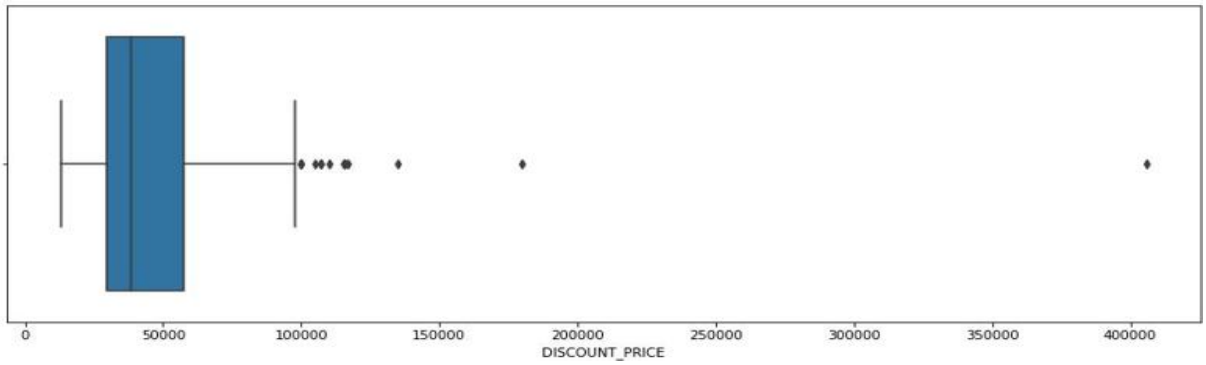

Fig.11 In this screen we can see that Uni Variate analysis using then graph BOX PLOT

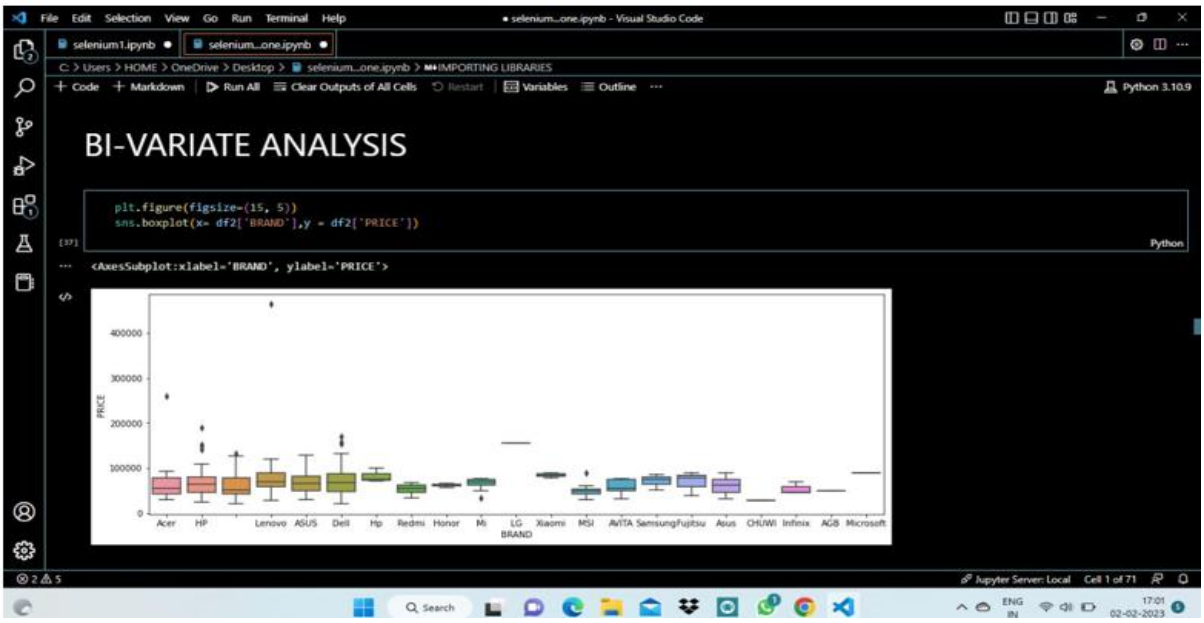

Fig.12 In this screen we can see that Uni Variate analysis using then graph BOX PLOT

ISSN: 2366-1313

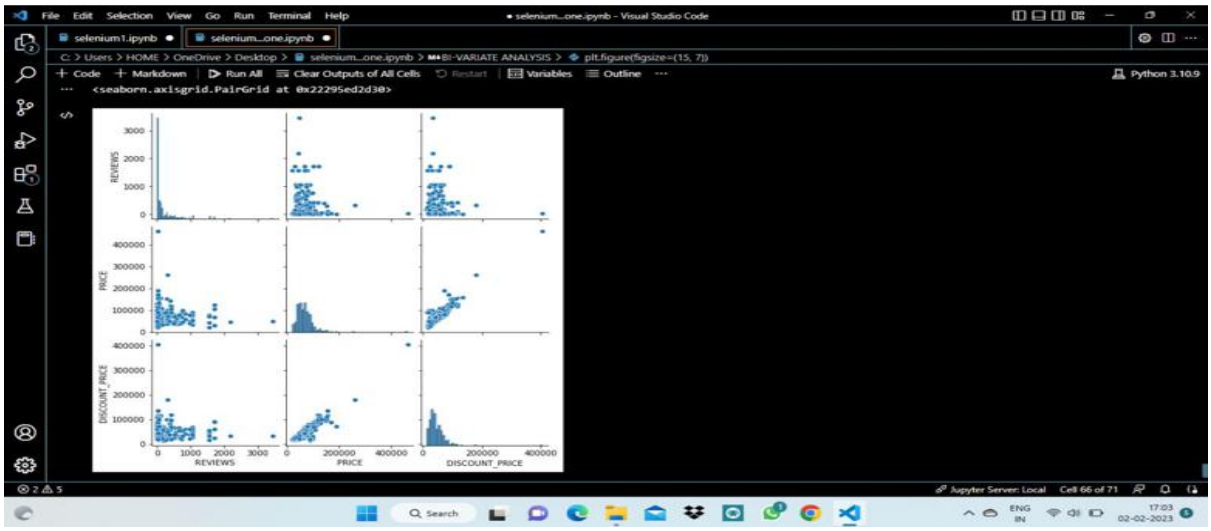

Fig.13 In this screen we can see that Multi Variate analysis using the graph PAIR PLOT

V. CONCLUSION

Web scraping can help us extract an enormous amount of data about customers, products, people, stock markets, etc.One can utilize the data collected from a website such as e-commerce portal, Job portals, social media channels to understand customer's buying patterns, employee attrition behavior, and customer's sentiments and the list goes on.

REFERENCES

[1] h. Song, D. Rawat, S. Jeschke, and C. Brecher, Cyber-Physical Systems: Foundations, Principles and Applications, 1st ed., ser. 1. Elsevier, 2016, academic Press.

[2] Y. Maleh, M. Shojafar, A. Darwish, and A. Haqiq, Cybersecurity and Privacy in Cyber Physical Systems, 1st ed., ser. 1. CRC Press, 5 2019.

[3] H. Song, G. Fink, and S. Jeschke, Security and Privacy in CyberPhysical Systems, 1st ed. Wiley Online Library, 2017.

[4] Y. Zhang, C. Xu, H. Li, K. Yang, J. Zhou, and X. Lin, "Healthdep: An efficient and secure deduplication scheme for cloud-assisted ehealth systems," IEEE Transactions on Industrial Informatics, vol. 14, no. 9, pp. 4101–4112, 2018.

[5] S. Garg, K. Kaur, N. Kumar, G. Kaddoum, A. Y. Zomaya, and R. Ranjan, "A hybrid deep learning-based model for anomaly detection in cloud datacenter networks," IEEE Transactions on Network

ISSN: 2366-1313

and Service Management, vol. 16, no. 3, pp. 924–935, 2019.

[6] S. S. Gill, S. Tuli, M. Xu, I. Singh, K. V. Singh, D. Lindsay, S. Tuli, D. Smirnova, M. Singh, U. Jain et al., "Transformative effects of iot, blockchain and artificial intelligence on cloud computing: Evolution, vision, trends and open challenges," Internet of Things, vol. 8, p. 100118, 2019.

[7] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Annual international conference on the theory and applications of cryptographic techniques. Springer, 2013, pp. 296–312.

[8] David Mathew Thomas, Sandeep Mathur ,Amity Institute of Information Technology ,Amity University (AUUP), Sec-125, Noid