

OPTIMAL PRODUCT PRICE ANALYSING SYSTEM

A PROJECT REPORT

Submitted by

AKSHAT GOEL [Reg No: RA1811030010069]
SACHIN VERMA [Reg No: RA1811030010085]

Under the Guidance of

DR. S. PRABAKERAN

(Assistant Professor, Department of Networking and Communications)

In partial fulfillment of the requirements for the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE ENGINEERING
with specialization in Cyber Security



DEPARTMENT OF NETWORKING AND COMMUNICATIONS

COLLEGE OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR - 603203

MAY 2022

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR–603 203**

BONAFIDE CERTIFICATE

Certified that this B.Tech project report titled “**OPTIMAL PRODUCT PRICE ANALYSING SYSTEM**” is the bonafide work of **AKSHAT GOEL [Reg No: RA1811030010069]**, **SACHIN VERMA [Reg No: RA1811030010085]** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

SIGNATURE

DR. S. PRABAKERAN
GUIDE
Assistant Professor
Department of Networking and
Communications
SRM Institute of Science &
Technology, KTR

Signature of the Internal Examiner

SIGNATURE

DR. ANNAPURANI PANAIYAPPAN .K
HEAD OF THE DEPARTMENT
Professor
Department of Networking and
Communications
SRM Institute of Science &
Technology, KTR

Signature of the External Examiner



Department of Network and Communications

SRM Institute of Science & Technology

Own Work* Declaration Form

This sheet must be filled in (each box ticked to show that the condition has been met). It must be signed and dated along with your student registration number and included with all assignments you submit – work will not be marked unless this is done.

To be completed by the student for all assessments

Degree/Course : B.Tech in Computer Science Engineering with spl. in Cyber Security

Student Name : Akshat Goel, Sachin Verma

Registration Number : RA1811030010069, RA1811030010085

Title of Work : Optimal Product Price Analysing System

I/ We here by certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism**, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

- Clearly references/ listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalized in accordance with the University policies and regulations.

DECLARATION:

I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.

RA1811030010069 RA1811030010085

If you are working in a group, please write your registration numbers and sign with the date for every student in your group.

ACKNOWLEDGEMENTS

We express our humble gratitude to **Dr C. Muthamizhchelvan**, Vice-Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to Dean-CET, SRM Institute of Science and Technology, **Dr T.V.Gopal**, for his invaluable support.

We wish to thank **Dr Revathi Venkataraman**, Professor & Chairperson, School of Computing, SRM Institute of Science and Technology, for her support throughout the project work.

We are incredibly grateful to our Head of the Department, **Dr K. Annapurani Panaiyappan**, Professor, Department of Networking and Communications, SRM Institute of Science and Technology, for her suggestions and encouragement at all the stages of the project work.

We want to convey our thanks to our program coordinators **Dr. M. B. Mukesh Krishnan**, Program coordinator, and Panel Head, **Dr. S. Prabakeran**, Assistant Professor, Department of Networking and Communications, SRM Institute of Science and Technology, for their inputs during the project reviews and support.

We register our immeasurable thanks to our Faculty Advisor, **Dr. M. Uma**, Associate Professor, Department of Networking and Communications, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to my guide, **Dr. S. Prabakeran**, Assistant Professor, Department of Networking and Communications, SRM Institute of Science and Technology, for providing me with an opportunity to pursue my project under his mentorship. He provided me with the freedom and support to explore the research topics of my interest. His passion for solving problems and making a difference in the world has always been inspiring.

We sincerely thank the Networking and Communications Department staff and students, SRM Institute of Science and Technology, for their help during our project. Finally, we would like to thank parents, family members, and friends for their unconditional love, constant support, and encouragement.

AKSHAT GOEL
SACHIN VERMA

ABSTRACT

The problem statement we are attempting to answer is to determine the best pricing for a product that is being sold online. Online businesses are increasingly simple to start and offer low start-up costs all over the world. For a variety of reasons, many prefer to open an online store, including lower taxes, less crowds, a larger assortment, and faster updates. As a result, with the rapid growth of e-commerce websites, online shopping has become the norm these days. Though online shopping is convenient, determining which online site has the greatest price and offers is a tedious and time-consuming task.

As the number of customers grows, they are more likely to use e-commerce services to save a substantial sum of money, time spent deciding on price, review rating, product features, and delivery time. Comparing products and screening them from each online site takes a long time for a shopper. This report employs Web Scrapping techniques, as well as Python libraries like as Beautiful Soup, requests, and Matplotlib, to determine the best prices and product deals for customers from various online websites.

TABLE OF CONTENTS

| No. | Title | Page |
|----------|---------------------------------|-------------|
| | ABSTRACT | V |
| | LIST OF FIGURES | VIII |
| | LIST OF GRAPHS | IX |
| 1 | INTRODUCTION | 1 |
| | 1.1 Domain of Project | 1 |
| | 1.2 Challenges faced | 3 |
| | 1.3 Motivation | 4 |
| 2 | LITERATURE SURVEY | 6 |
| | 2.1 Research gap | 6 |
| | 2.2 Inference | 9 |
| 3 | SYSTEM ANALYSIS | 12 |
| | 3.1 Problem Definition | 12 |
| | 3.2 Proposed Solution/ Method | 13 |
| | 3.3 Algorithm | 14 |
| | 3.4 Data Extraction & Analysis | 15 |
| | 3.5 System Requirements | 18 |
| 4 | SYSTEM DESIGN | 20 |
| | 4.1 Architecture Diagram | 20 |
| | 4.2 Use Case Diagram | 21 |
| | 4.3 Class Diagram | 22 |
| | 4.4 Sequential Diagram | 24 |
| | 4.5 Data Flow Diagram | 25 |
| | 4.6 Modules | 26 |
| | 4.7 Implementation | 29 |
| 5 | RESULTS & DISCUSSION | 31 |
| | 5.1 Results | 31 |
| | 5.2 In-Depth Analysis of Data | 36 |
| | 5.3 Conclusion | 40 |
| | 5.4 Further enhancements | 41 |

| | | |
|----------|----------------------|-----------|
| 6 | REFERENCES | 42 |
| 7 | APPENDIX - I | 45 |
| 8 | APPENDIX - II | 50 |

LIST OF FIGURES

| Figure | Title | Page |
|---------------|---------------------------------|-------------|
| | SYSTEM ANALYSIS | |
| Figure 3.1 | Steps of Data Extraction | 15 |
| Figure 3.2 | How to Find Product URL | 16 |
| Figure 3.3 | Data Element – XPath | 17 |
| Figure 3.4 | Extraction of Raw Data | 17 |
| Figure 3.5 | Visualization of Filtered Data | 18 |
| | SYSTEM DESIGN | |
| Figure 4.1 | Architecture Diagram | 20 |
| Figure 4.2 | Use Case Diagram | 21 |
| Figure 4.3 | Class Diagram | 23 |
| Figure 4.4 | Sequential Diagram | 24 |
| Figure 4.5 | Data Flow Diagram | 25 |
| | RESULTS & DISCUSSION | |
| Figure 5.1 | Extracted Raw Data | 32 |
| Figure 5.2 | Filtered Data | 33 |
| Figure 5.3 | Best Price Analysis | 34 |
| Figure 5.4 | Best Rating Analysis | 34 |
| Figure 5.5 | Best Rating Count Analysis | 35 |
| Figure 5.6 | Amazon Rating Count Analysis | 36 |
| Figure 5.7 | Amazon Price Analysis | 36 |

LIST OF GRAPHS

| Figure | Title | Page |
|---------------|---|-------------|
| Figure 5.8 | Apple iPhone 12 (128GB) – Green (AMAZON RATING COUNT) | 37 |
| Figure 5.9 | Apple iPhone 12 (128GB) – Green (FLIPKART RATING COUNT) | 37 |
| Figure 5.10 | Iphone 20W Charger (AMAZON RATING COUNT) | 38 |
| Figure 5.11 | Iphone 20W Charger (FLIPKART RATING COUNT) | 38 |
| Figure 5.12 | Apple iPhone 12 (128GB) – Green (FLIPKART RATING COUNT) | 39 |

CHAPTER 1

INTRODUCTION

1.1 Domain of Project

The optimal product analyzing System is used to find the lowest price available of a product from Ecommerce websites such as amazon, Flipkart, Snapdeal, aytm, roma, and so It entails the application of modern technologies such as online scrapping, data science, and data analytics[3]. By using the optimal product analyzing System, we will be able to not only determine the lowest price of a product, but also analyse the product data available on the internet. It covers a wide range of important techniques, including analysing, modelling, visualisation, and evaluating. It has a wide range of application perspectives in areas including price monitoring, news monitoring, trend recognition, and lead generation[7]. Digital marketers are becoming increasingly easy to start and provide minimal startup expenses around the world. Many people prefer to create an online store for a range of reasons, including lower taxes, fewer crowds, a wider selection, and faster updates. As a result of the rapid application of e websites, online shopping has now become the standard. While internet buying is handy, determining which website has the best pricing and offers is a time-consuming and laborious effort.

When looking for information on the internet, most people utilise a web browser. Browsers make it simple to view and navigate different websites. Large volumes

of unstructured data are found on websites[9]. There is a ton of stuff mixed in together with useful material on a website. It's an automated method for swiftly and easily retrieving big amounts of data. Large volumes of information are gathered and stored in a structured fashion (such as .SV files or database files). Just several commercial dynamic web administrators around the world consider web scrapping as lawful, and others do not

The legality of using online scripting is entirely up to the website administrators. Users would be able to obtain information on a particular website if they consent. Data is changed from that of an unorganized to a structured manner via Web Scraping. Around the world, internet businesses are simple to create and operate. Users love to sell online for a range of reasons, including reduced taxes, fewer crowds, greater selection, and quicker updates.

Customers are increasingly likely to spend a large amount of time deciding on pricing, rating, product characteristics, and delivery time as the number of e-commerce services expands. Users looking for information about products or administrations utilise 54 percent of the Internet, 48 percent for academic reasons, 40 percent for research and patient research, 28 percent for job-seeking activities, and 24 percent for data. This research looks at one way for collecting information from e-commerce webpages and providing it to consumers on a screen that helps them to sort through with a vast number of unnecessary information.

It is entirely up to the website administrators to determine whether or not employing online scripting is legal. If users agree, they will be able to access information on a certain website. Web Scraping transforms data from an unstructured to a structured state. Internet businesses are easy to start and run all over the world. Users choose to sell online for a variety of reasons, including lower taxes, fewer crowds, more variety, and faster updates.

Online scrapping can be done, including ython, Node.js, H, Perl, ++, and others. This paper leverages the Python language for online Scrapping as Python is more adoptive to further information analysis; it is easy to implement; and it contains libiries such as The website administrators are solely responsible for determining whether or not using online script is legal. People will be able to examine data on another website if they approve. Web extraction transforms unstructured data into a structured format.

Online businesses are increasingly simple to start and offer low start-up costs all over the world. For a variety of reasons, many prefer to open an online store, including lower taxes, less crowds, a larger assortment, and faster updates. Customers are more likely to utilise e-commerce services as the number of them grows. Customers are more willing to spend money and effort on price, review rating, product qualities, and delivery time as the number of them grows[11].

1.2 Challenges faced

We made sure to keep our goals straight. Alongside, we faced multiple challenges and Problems doing the project and some of them had become too difficult to overcome or at least minimize but ultimately were sorted out with different approaches and right guidance.

Some of the major challenges that we faced are as follows:

- Bots are restricted on a marketplace.
- Blocking IP addresses from a marketplace.
- A marketplace's loading speed is slow and unreliable.

- Login require at the time of accessing products.
- Dynamic content of the webpage.
- Dynamic XPath locator of the webpage.
- Collecting of input dataset is a time consuming task.
- Handling of larger dataset.
- No Such element-based errors.
- Connection with webpage based errors.
- Data extracted from many webpages contains trash data.

1.3 Motivation

Pricing Optimization - At any one time, we frequently find different prices for the same goods on multiple marketplaces such as Amazon and Flipkart. Then there are sales and special deals, which influence price changes. As a result, we made the decision to keep an eye on product prices and try to figure out what the current trend is. We frequently find various rates for the same items on multiple marketplaces, such as Amazon and Flipkart, at any one time. Then there are sales and special offers, which have an impact on price fluctuations. As a result, we decided to keep a close eye on product prices in order to determine the current trend.

Competitor Monitoring - The e-Commerce marketplace has risen substantially in the last decade. The online retailing landscape will look to advance as technological gadgets become more interwoven into our lives and our buying patterns change. The profitable sector is easy to break into, but retailer competition will only grow, minimizing the potential for newcomers to flourish. What strategies do you use to keep your traditional retail afloat? You must

conduct research on your competitors. If you understand your opponents and yourself, you will never again be defeated. In business, this is also true. As technology gadgets grow more integrated into our lives and our purchasing behaviours alter, the internet commerce scene will continue to evolve. The lucrative area is simple to enter, but retailer competition will only increase, reducing the possible profit.

Product Optimization - We all know that reading online reviews before making a purchase is a good idea. Customers' purchasing decisions can be influenced by reviews in a predictable way. As a result, we can assess how they perceive us in order to meet their expectations. Assume your product team is getting ready to launch a new product. You're worried because you don't know if it'll work. Collecting client feedback is critical for cross-examining your product and making adjustments. Customers' attitudes, whether favourable, neutral, or negative, are studied using the sentiment analysis technique. However, in order for the analysis to work, it requires a large volume of text data from a variety of websites. Web scraping allows you to automate the extraction process, saving you a lot of time and work. As a result, we may evaluate how they see us in order to suit their needs. Assume that your product team is preparing to release a new product. You're concerned because you're unsure if it will work. Collecting consumer feedback is essential for re-evaluating and improving your offering. The sentiment analysis technique is used to investigate customers' attitudes, whether positive, neutral, or negative.

CHAPTER 2

LITERATURE SURVEY

2.1 Research gap

A textual review is a critique of previously published works on a certain topic. A whole academic document or a portion of an expert work such as a book or article might be referred to as a dissertation.

In any event, book reviews should provide a general summary of the facts contained in the issue being addressed to both the researcher and the audience.

A thorough examination of the literature can help you formulate the right research question and choose the right theoretical framework and/or research approach. Book reviews are accurate because they situate current findings in the context of relevant literature and give the reader perspective.

In this scenario, the evaluation usually happens before the phases of performance and the results of those stages. Book review writing is often part of the work of graduate and graduate students, which includes preparing a thesis, dissertation, or journal article.

This technology would allow data to be scraped from a wide range of websites, reducing human interaction, saving time, and boosting data relevancy quality.

It will also help the user collect data from the site, save it according to their needs, and use it as they see fit ^[1].

The scribbled information can be used for database construction, research, and

other related tasks. This technique should be viewed as a gift that must be carefully managed if human races are to advance. Nand Saurka discovered the most recent approach, known as Web Scraping ^[2].

A thorough examination of the literature can help you formulate the right research question and choose the right theoretical framework and/or research approach. The process of transforming unstructured material on the internet into structured data is known as web cleaning ^[3]. Scrapping formed structural data, which was then gathered and evaluated in central database spreadsheets. In this essay, the authors explained the fundamentals of web processing ^[5].

They focused on web-scaffolding techniques. The study finishes with a review of the various technology options now available on the market for efficient online scrapping. Federico Olidoro et al. focused on the repercussions of web scrapping evaluation strategies with particular orientation to user electronic services and goods across the sector ^[9].

Despite the fact that the research was conducted in a short amount of time, it was able to gather useful but not exhaustive information. In actuality, engaging with this viewpoint demands a review of the current survey architecture, which does not require or only selectively authorises the use of huge data approaches within the existing sample framework K.Kambatla and colleagues presented a web-based framework for precise and quantified ecosystem mining for development ^[18].

| Year | Author | Title | Publication | Result |
|------|---|---|--|---|
| 2020 | Erin J. Farley and Lisa Pierotte | An Emerging Data Collection Method for Criminal Justice Justice Researchers | Justice Research and statistics association | It focuses on issues related to the causes and consequences of crime, delinquency, and victimization. |
| 2019 | Jan Kinne and Janna Axenbeck | Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study | International Journal on Future Revolution in Computer Science & Communication Engineering | It derive firm-level information from the extracted web data. |
| 2018 | Sameer Padghan, Satish Chigle and Rahul Handoo | Web Scraping-Data Extraction Using Java Application and Visual Basics Macros | Journal of Advances and Scholarly Researches in Allied Education | Web scraping using the Java language and built a functional scraper using the simple but powerful JSoup library. |
| 2018 | Anand V. Saurkar, Kedar G. Pathare and Shweta A. Gode | An Overview On Web Scraping Techniques And Tools | International Journal on Future Revolution in Computer Science & Communication Engineering | It used to automatically extract large amounts of data from websites and save it to a file or database. |
| 2018 | Kaushal Parikh, Dilip Singh, Dinesh Yadav and Mansingh Rathod | Detection of web scraping using machine learning | Open access international journal of Science and Engineering | It used to create advanced scraping algorithms, as it is well suited for the task of generalizing. |
| 2017 | Chaulagain, Ram Sharan, Subarna Shakya | Cloud-based web scraping for big data applications | IEEE International Conference on Smart Cloud | By using cloud servers we can automate our bot to work at any particular time & interval. |
| 2017 | Madhusudan, Lambhate Poonam, D | Deep Web Crawling Efficiently using Dynamic Focused Web Crawler | International Research Journal of Engineering and Technology | It derive firm-level information from the extracted web data. |
| 2016 | Zhao, F., Zhou, J., Nie, C., Huang, H. and Jin, H | SmartCrawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces | IEEE transactions on services computing | By using multiple proxies we can distribute the requests made from user-end |
| 2015 | Federico Polidoro, Riccardo Giannini, Stefano Mosca | Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation | Statistical Journal of the IAOS | Boilerpipe and nutch modules are used for extracting the source data. |
| 2015 | Renita Crystal Pereira and Vanitha T | Web Scraping of Social Networks | International Journal of Innovative Research in Computer and Communication Engineering | It acquire non-tabular or poorly structured data from websites and convert it into a usable, structured format, such as a .csv file or spreadsheet. |
| 2020 | Nadikattu, Rahul Reddy | Data Science, Data Analytics and Big Data | INTERNATIONAL JOURNAL OF ENGINEERING, SCIENCE | Data Analytics seeks to provide operational insights into complex business situations. |
| 2015 | M. K. Khakhani, S. Khakhani, and S. R. Biradar | Research challenges in big data analytics | International Journal of Application or Innovation in Engineering and Management | It used to create advanced scraping algorithms, as it is well suited for the task of generalizing. |
| 2012 | A. Gandomi and M. Haider | Beyond the Hype: Big Data Concepts, Methods, and Analytics | International Journal, 2012 | Conceptual Analysisx of data proxies |
| 2008 | C. Lynch | Big data: How do your data grow? | Indonesian Journal of Science and Technology | Appending of large data |
| 2015 | X. Jin, B. W. Wah, X. Cheng, and Y. Wang | Big Data Research: Significance and Challenges | Open access international journal of Science and Engineering | Important information gathering |
| 2014 | R. Kitchen | Big Data, New Epistemologies, and Paradigm Shifts, Big Data Society | Engineering and Applied Science Research | New technologies and Paradigm Shifting |

Table 2.1: Literature Survey

| | | | | |
|------|---|---|--|---|
| 2014 | C. L. Philipp, Q. Ckhen and C. Y. Zxang | Data-intensive applications, challenges, approaches, and technologies: A survey on big data, Information Sciences | IEEE transactions on services computing | Applications under Data-Intensive |
| 2014 | K. Kambatla, G. Kollias, V. Kumar and A. Gram | Trends in big data analytics, Journal of Parallel and Distributed Computing | Indonesian Journal of Science and Technology | Latest trends in big data to consider for project |
| 2014 | S. Del Rio, V. Lopez, J. M. Bentez, and F. Herrera | Information Sciences, on the application of mapreduce for imbalanced big data utilising random forest | Journal of Advances and Scholarly Researches in Allied Education | Data collection and information Science |
| 2014 | MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki, and D. K. Grunwell | Health big data analytics: current perspectives, difficulties, and prospective solutions, International Journal of Big Data Intelligence | Open access international journal of Science and Engineering | Data analytics to collect web-scraped material |
| 2018 | Mitchal, Rayn | Online Extraction with .py: Collecting data available from the modern web pages | Indonesian Journal of Science and Technology | It used to automatically extract large amounts of data from websites and save it to a file or database. |
| 2019 | Toby Michale, H.K. Wanhnga | Web Extraction & Valuation using Python: Evaluating data available from the marketplace websites | Engineering and Applied Science Research | Valuation using python |
| 1976 | T. J. McCabe | A complexity measure. IEEE Transactions on Software Engineering | IEEE transactions on services computing | Array and decimal arrangements(Pricing) |
| 2009 | Andreas Mehlführer | Web scraping: A tool evaluation | Technische Universität Wien | Ways of Scraping data from web |
| 2016 | Yolande Neil | Web scraping the easy way | Georgia Southern University | Scraping through shortcuts |
| 2016 | Joacim Olofsson | Evaluation of webscraping tools for creating an embedded webwrapper | KTH, School of Computer Science and Communication (CSC) | Multiple tools for evaluating data after scraping |
| 2015 | S. Sirisuriya | A comparative study on web scraping | International Research Conference, KDU, 8:135–139, 11 | Study between different types of extracted data |
| 2005 | Andrew S Tanenbaum and Albert S Woodhull | Operating Systems Design and Implementation (3rd Edition) | Prentice-Hall, Inc., Upper Saddle River, NJ, USA | Different |
| 2015 | N. Mishra, C. Lin, and H. Chang | A cognitive adopted framework for iot large data management and knowledge discovery perspective, International Journal of Distributed Sensor Networks | Journal of Advances and Scholarly Researches in Allied Education | It derive firm-level information from the extracted web data. |
| 2012 | X. Y. Chen and Z. G. Jin | Research on key technology and applications for the internet of things, Physics Procedia | Open access international journal of Science and Engineering | IOT based study |
| 2015 | M. D. Assuno, R. N. Calheiros, S. Bianchi, M. a. S. Netto, and R. Buyya | Web Extraction & analysis of emotional sentiment | Journal of Parallel and Distributed Computing, 79 | It used to create advanced scraping algorithms, as it is well suited for the task of generalizing. |

Table 2.2: Literature Survey

2.2 Inference

Web scraping is quite important methodology used to produce structured data based on the unstructured data available on the internet. Scraping formed structural data, which is then collected and evaluated in spreadsheets in a central

database.

The proposed research system enables for an integrated, less expensive simulation of whole business communities, that could be conducted more efficiently and in relatively short time periods when compared. The web scrapping strategies employed in the growth analysis will provide exposure to a greater volume of data accessible in the present data set, with the potential to increase the growth estimate. There are numerous technological resources available in the business for effective online scrapping.

It also helps in comparison of sales of different types of products from different marketplace. Data extraction of product can help a business or startup to find a trending product, so, they can also boost their sales. The amount of data available on the internet is steadily expanding. Now this data can be stored in local data servers or computer for ease of retrieval.

We all know that reading online reviews before making a purchase is a good idea. Customers' purchasing decisions can be influenced by reviews in a predictable way. As a result, we can assess how they perceive us in order to meet their expectations. The online retailing landscape will look to advance as technological gadgets become more interwoven into our lives and our buying patterns change.

The profitable sector is easy to break into, but retailer competition will only grow, minimizing the potential for newcomers to flourish.

When compared, the proposed research method enables an integrated, less expensive simulation of whole business communities that may be carried out more efficiently and in relatively short time periods. The web scrapping tactics used in the growth analysis will expose the user to a larger volume of data than is now available in the data set, perhaps increasing the growth estimate. For

effective online scrapping, there are several technological tools available in the industry. We use two marketplace/e-commerce websites to extract product pricing for mobile phones and electrical goods such as cameras, printers, and chargers for our research. We must determine the XPath of the data to be extracted. We may achieve this by inspecting the web page by right-clicking it and selecting inspect element, then copying the XPath.

We must examine the web page to determine which element holds the data we wish to extract. We scrape data from a marketplace using Python.

CHAPTER 3

SYSTEM ANALYSIS

3.1 Problem Definition

The Problem that we are going to focus on is the price fluctuation in different marketplace such as Amazon, Flipkart, or any other online website selling products. A consumer/ Customer wants to find the best price available for a product, In addition to this, he/ she also wants to buy a product from a genuine seller which can also provide further services at affordable rates. This is a customer-based scenario.

We're attempting to solve is choosing the appropriate pricing for an online-only product. Small retailers are becoming increasingly easy to start and provide minimal startup expenses around the world. Many people prefer to create an online store for a multitude of reasons, including lower taxes, fewer visitors, a wider selection, and faster updates. As a result of the rapid development of e websites, online shopping seems to have become the standard. While internet buying is convenient, identifying which website has the best pricing and offers is a time-consuming and laborious effort.

We also want to focus on the business-based scenario which can help other startups or businesses who are dealing with some issues in selling their products online. We want to study/ research the life cycle of different types of products. This also helps in extracting the factors which cause price fluctuations over a

year. This also helps businesses to keep an eye on their competitors' products.

3.2 Proposed Solution/ Method

We can solve the proposed problem by using simple techniques such as data extraction, data mining, data analysis and data visualization.

The process of importing information from a website into a Excel/CSV or a local file saved on your computer is known as data extraction, also known as online scrapping. It's one of the most effective ways to collect data/information from the web and, in some situations, to send that information to another website. Data scraping is commonly used for the following purposes:

Web content/business intelligence research, Travel booking portals/price comparison sites are priced differently, By crawling public data sources, you can get sales leads and conduct market research, Sending product information from one e-commerce site to another.

And that's only the tip of the iceberg. Data scraping has a wide range of uses; it may be used in almost any situation where data needs to be transferred from one location to another. The fundamentals of data extraction are simple to learn. Let's look at how to use Excel to create a simple data extraction operation.

The information gathered here is not appropriate for immediate use. It must go through some sort of cleansing process before we can utilise it. For this, methods such as string manipulation and regular expressions can be employed. It's worth noting that extraction and modification can both be done in one process. We can also add a scheduler using cloud computing, which is used to fetch the product URL stored in the database and extract the data such as product price, product rating, product name, etc.

Its main function is to call the URL at a particular interval of time to avoid IP blockage for the server-end. The visual display of data or information is known as data visualization. The purpose of data visualization is to clearly and effectively communicate facts or information to readers. A chart, info graphic, diagram, or map is typically used to visualize data. This is only the top of the iceberg. Data scraping can be utilised in practically any case where data needs to be transmitted from one area to another. Learning the foundations of data extraction is simple. Let's look at how to make a simple data extraction process in Excel. The data collected here is not suitable for immediate use. Before we can use it, it must go through some sort of purification procedure.

Data visualization is a field that blends data science and art. While a data visualization might be artistic and pleasant to the eye, it must also be effective in terms of visual data transfer.

3.3 Algorithm

Product Price Analyzing System is a web automation system which is used to extract as well as analyze the optimal product price.

It is divided into 5 steps:-

Basic Data - Collection of basic data such as Product URL/ Link & data element to be extracted

Fetching Raw Data – With the help of Basic data we fetch the product page & extract the raw data from the web page

Example - ₹99,450 is the raw price of a product. This represents a string data, we need to convert it into integer, so that we can apply mathematical formulas on this data.

Applying Filter – We need to remove unwanted/ garbage data from the raw data such as ‘₹’, ‘,’ , etc. we need to convert string into integer or float.

AVE of Data – AVE of data includes Analyzing, Visualization and Evaluating of data which helps the end user to get optimal information in more refined manner. And also it helps the strategist or data concluder to take further steps in optimizing the product.

Data Concluder – Data concluder is a process which is used to predict the outcome of a certain event. It is used to take measured steps for making a process more reliable or error free.

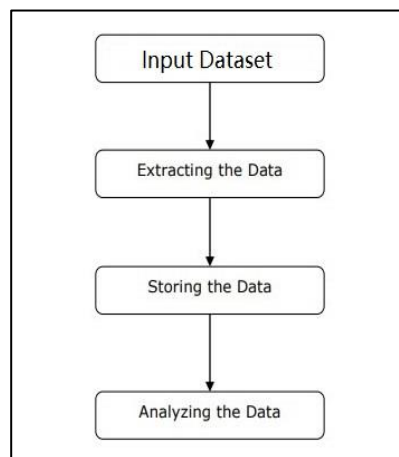


Figure 3.1: Steps of Data Extraction

3.4 Data Extraction & Analysis

1. As depicted in the figure 3.1 above, First, user need to collect the basic information such as product type, product name, and websites on which product is available. This is one time process.

Example –

Device name – Blue - iPhone 12 (64GB)

Product Type – Mobile Phone

Product URL –

<https://www.amazon.in/New-Apple-iPhone-12-64GB/dp/B08L5WHFT9/>

Data element –

`//*[@id="corePrice_desktop"]/div/table/div/tbody/tr[4]/span[1]/span[2]/d[3]/`

We can see this in the figure 3.2 and 3.3 given below :



Figure 3.2: How to Find Product URL

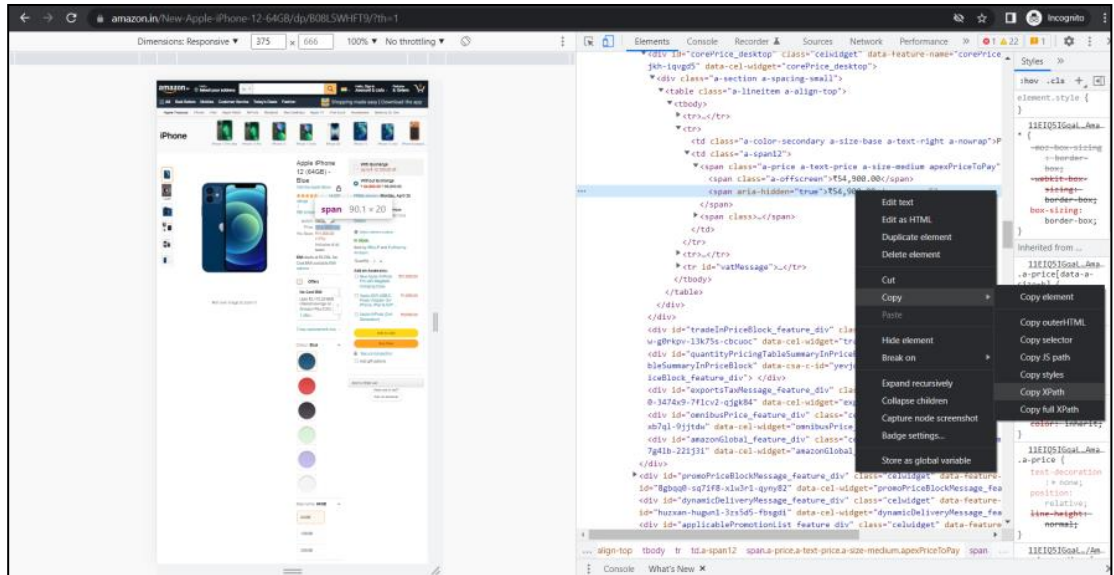


Figure 3.3: Data Element – XPath

2. As shows in the figure 3.4 below, The next step is to collect raw data of the product, which is the structured web data from E-commerce websites such as Amazon, Flipkart etc.

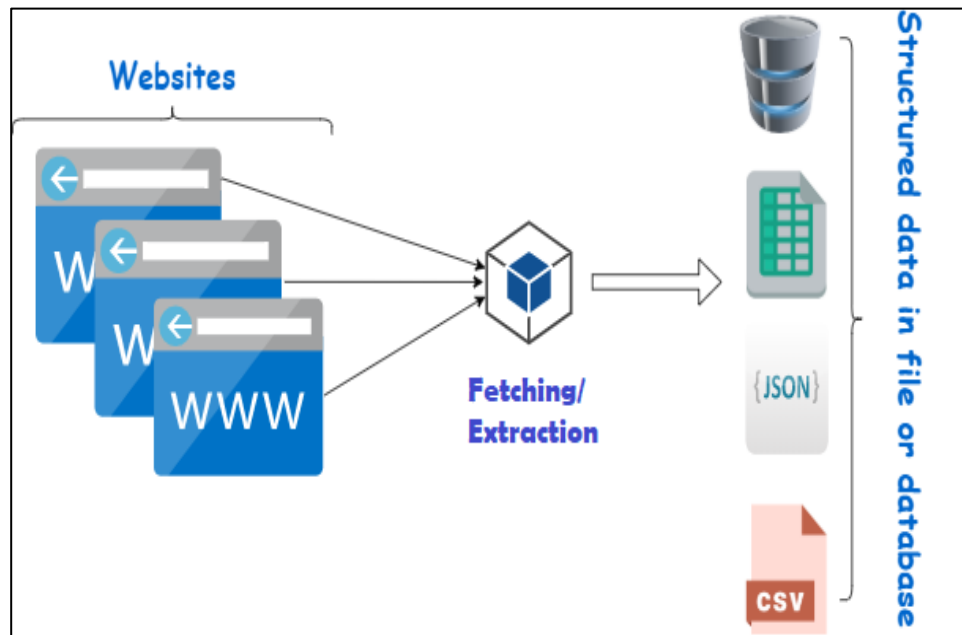


Figure 3.4: Extraction of Raw Data

3. There are times when collected data may contains useless information, then we need to apply some filters to exclude those information.
4. AVE of data includes Analyzing, Visualization and Evaluating of data which helps the end user to get optimal information in more refined manner. And also it help the strategist or data concluder to take further steps in optimizing the product. We can see the different forms for visualizing the data in figure 3.5 below:



Figure 3.5: Visualization of Filtered Data

3.5 System Requirements

- ✓ Hardware Requirements -
 - 8 GB RAM minimum for fast performance

- 25GB space minimum
 - i3 processor or above
 - Stable Internet Connection
-
- ✓ Software Requirements -
 - Anaconda IDE
 - Jupyter Notebook
 - Python Language version 3.8.0
 - Window 10 with 64-bit compatibility
 - Pip package for downloading Python Modules

CHAPTER 4

SYSTEM DESIGN

4.1 Architecture Diagram

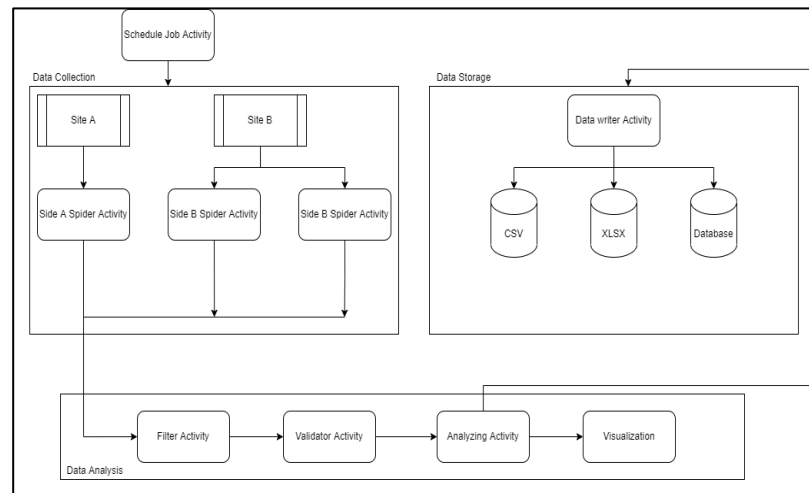


Figure 4.1 : Architecture Diagram

A. Data Collection

As we can see in the figure 4.1 above, firstly, The web extractor module, an essential component of a web scraper, is often used to browse the web address by making Https or Http requests to the URLs. The crawler downloads complex data (HTML contents) and sends it to the extractor, which is the next module. The extractor takes the fetched Html elements and converts it into semistructured data. This is also known as a parser module, and it works by

utilising several parsing techniques such as Regular expressions, HTML parsing, DOM parsing, and Artificial Intelligence.

B. Data Analysis

The information gathered here is not appropriate for immediate use. It must go through some sort of cleansing process before we can utilise it. For this, methods such as string manipulation and regular expressions can be employed. It's worth noting that extraction and modification can both be done in one process.

C. Data Storage

We must keep the data when it has been extracted, according to our needs.

The data will be be output in a standard format, such as JSON or CSV, which may be saved in a database.

4.2 Use Case Diagram

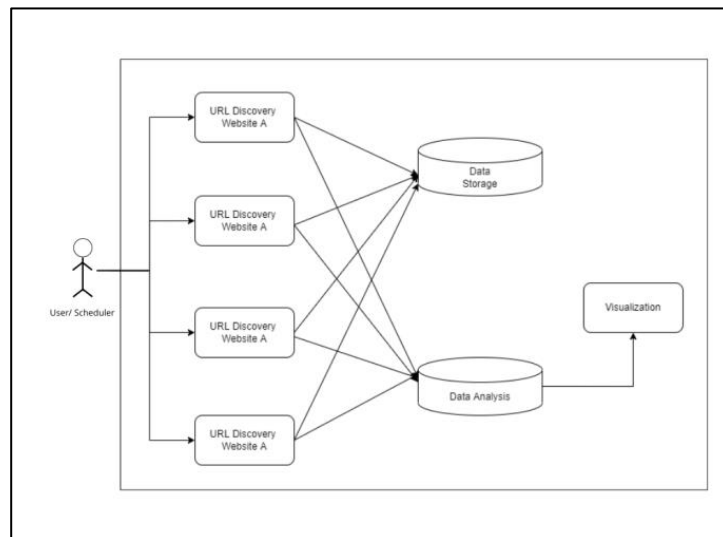


Figure 4.2: Use Case Diagram

From figure 4.2 of Use case Diagram above we can further explain each of the components:

Scheduler - A scheduler is used to fetch the product URL stored in the database and extract the data such as product price, product rating, product name, etc. Its main function is to call the URL at a particular interval of time to avoid IP blockage for the server-end.

Data Storage - We must keep the data when it has been extracted, according to our needs. The data will be output in a standard format, such as JSON or CSV, which may be saved in a database.

Data Analysis - The information gathered here is not appropriate for immediate use. It must go through some sort of cleansing process before we can utilise it. For this, methods such as string manipulation and regular expressions can be employed. It's worth noting that extraction and modification can both be done in one process.

Data Visualization - The filtered data is ready to visualize and it also contains the capability to predict further data. It helps in giving lively view to the numerical data. It also helps in predicting the data trend.

4.3 Class Diagram

Figure 4.3, is the class diagram of optimal product price analysing system. Here, we have a url dataset which basically contains the input url or product url which act as an input data. Input data also contains xpath locator of price, rating and rating count.

so, first it checks weather the product url is valid or not by using `requestConnection()` and if the product url is valid then it will search for element in the HTML with the help of `noSuchElement()` function. It tries to find text data

in the element using Xpath locator. This comprises of urlSet.

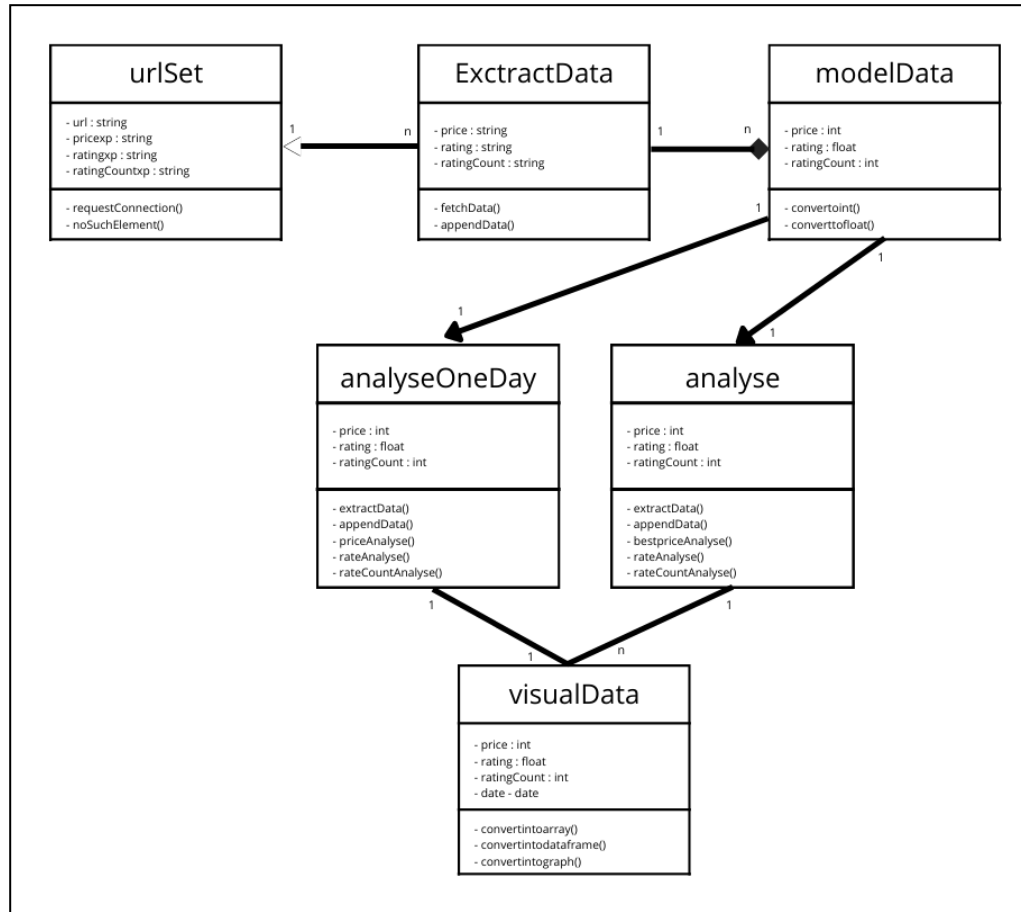


Figure 4.3 : Class Diagram

Now, when the data is verified and valid, it uses `fetchData()` function to fetch the data. It extract the price, rating, rating count and store it as string datatype. Then, with the help of `appendData()` function it data the fetched data in master file/ storage, which already had past extracted data. Afte, storing the data into storage like database or local server we need to clean the data to remove unnecessary information and then convert the string into integer or float data by using `convertoint()` and `converttofloat()` respectively. we generally prefer price and rating count as of integer datatype and rating as float datatype.

After, cleaning of data, we can analyse the filtered data as analysis of one day data and analysis of collective data. Analysis of one day data is the analysis of data collect in a one day. It can be the best price available on that day, what is the rating count on that day. We can also use `rateCountAnalyse()` function to analyse rating count of that day. On the other hand collective data analysis is analysis of previous collected data. It provides a user with the best optimal price available for a product at a particular time in past. We can also use `rateCountAnalyse()` function to find the number of product sold in a given interval of time. Higher the change in rating count higher will the be the product sold.

Now, this analysed data is converted into arrays using `convertintoarray()` function to visualise into graphs. We use `plot()` function to plot differt types of graphs such as line and bar. This whole process is done using `visualData`.

4.4 Sequential Diagram

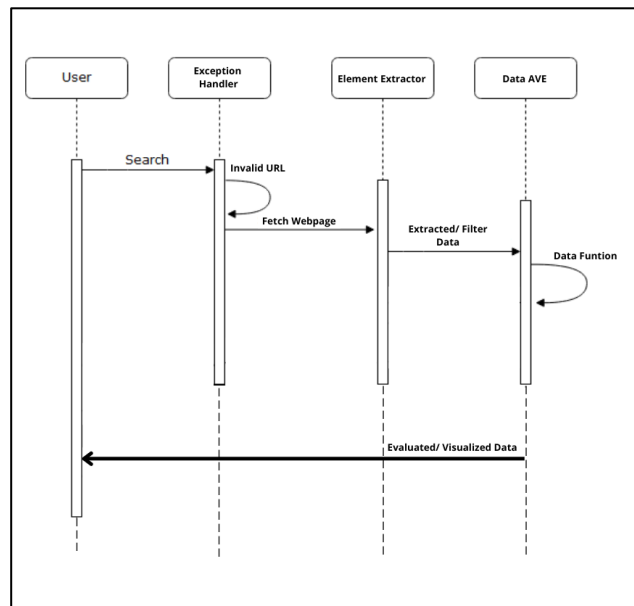


Figure 4.4: Sequential Diagram

From the figure 4.4 above, we conclude that User or Scheduler is used for the execution of program, they are also used to fetch the URL/ Product Link form the info. Dataset. The URL then passed through an exceptional handler and when it's a valid URL then the program is responsible for fetching the webpage. And, successfully getting the HTML content, we use element extractor for extracting the product information. Element extractor such as `findElementbyID()`, `XPath Locator`. And the element extractor is passed through exceptional handler to check whether the element is present or not in the HTML content.

If, the element is present in the HTML, it fetches the element & return the text data from that element. Then this data is converted into integer or float for mathematical operations. Different types of mathematical functions can be applied on this data to get the meaningful insight from the data. The collected data & the analysed data be stored in local datasever or filetype such XML, CSV etc.

4.5 Data Flow Diagram

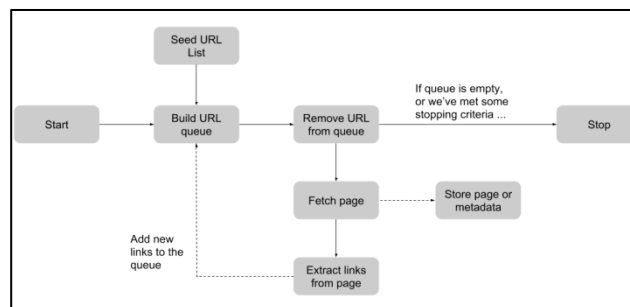


Figure 4.5: Data Flow Diagram

From the figure 4.5 above we can say that The internal functioning of a common web scrapper are depicted in the diagram above. A webpage scrapper is a

programme that crawls the internet and indexes webpages in a methodical manner. Scraping bot is the most well-known application of online scraping.

The above-mentioned queue is known as the "frontier," and the URLs in it may be scored and ordered in a ready queue by "focused" or "topical" web scrappers. URLs may also be blocked out of the queue depending on their hostname or file type.

Web scrappers might be viewed as a nuisance from the perspective of a website developer. These internet bots impersonate genuine site visitors and therefore can request a large amount of pages from your website in a short period of time, putting a strain on your server.

Crawlers such as Google and Yahoo! analyze your material and offer it to interested individuals, resulting in more actual website visitors. Many domains have included a robots.txt file that specifies developers how they want online bots to communicate with their website in order to manage the amount and types of queries made.

4.6 Modules

The main modules that we use in Product Price Analyzing System:-

- **Info. Dataset** - A data set (also known as an input data) is a gathering of documents. A set of data equates to one or even more data blocks in the case of data tables, where each column of a table target a single variable and each establish a sense to a specific record of the set of data in question. It the

dataset of product items which are used to extract the raw information from the internet.

Example of a info dataset is shown in the table 4.1 below:-

| Amazon_link | Amazon_price_xpath | Amazon_rating_xpath | Amazon_rating_count_xpath |
|---|--|--|---|
| https://www.amazon.in/Apple-iPhone-13-512GB-Blue/dp/B09G9JJT7M/ | //*[@id="corePrice_desktop"]/div/table/div/tbody/tr[4]/td[3]/span[1]/span[2] | //*[@id="corePrice_desktop"]/div/table/div/div/tbody/tr[6]/td[2]/span[1]/span[2] | //*[@id="review_sMedley"]/div/div[1]/div[2]/div[2]/span |

Table 4.1: Info. Dataset

- **URL Crawler** - A webpage scrapper is a programme that crawls the internet and indexes webpages in a methodical manner. Scraping bot is the most well-known application of online scraping. URLs may also be blocked out of the queue depending on their hostname or file type. It is used to crawl the web pages of a domain to extract the URLs which are indexed on search engine.
- **XPath Locator** - Locators are used to identify elements on a Webpage. A locator can either be a basic attribute value, be an XPath query, identify an element from the DOM or CSS-based Locator or HTML5 based locator. We can use locators to find elements of a web page accurately. XPath Locator is to find the element in the webpage. Example of a XPath Locator of a price of a product is-

```
//*[@id="corePrice_desktop"]/div/table/div/div/tbody/tr[6]/td[2]/span[1]/span[2]
```

- **Exception Handler** – An exception handler is code that stipulates what a program will do when an anomalous event disrupts the normal flow of that program's instructions. An exception, in a computer context, is an unplanned event that occurs while a program is executing and disrupts the flow of its instructions. This module is used to check whether the URL & XPath is working or not.
- **Firefox Webdriver** - It is used to provide a link between test cases and the Firefox browser. We use GeckoDrive in this case.
- **Analyzing Module** - The information gathered here is not appropriate for immediate use. It must go through some sort of cleansing process before we can utilise it. For this, methods such as string manipulation and regular expressions can be employed. It's worth noting that extraction and modification can both be done in one process. It help the user with complex mathematical operations on arrays and matrices.
- **VE Module** - Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. It provide graphical plotting functionality to the interpreted data.

- **Concluder** - The final stage after analysing the data is to develop your findings. Conclusions state whether the results of the experiment or study support or refute the initial premise. To assist explain the results, teams should add crucial facts from their background research. It helps the end user in concluding with output data and also help in formulating strategic decisions.
- **Scheduler** - Schedule is in-process scheduler for periodic jobs that use the builder pattern for configuration. Schedule lets you run Python functions (or any other callable) periodically at pre-determined intervals using a simple, human-friendly syntax. It helps in executing a particular block of code at a given time on its own (i.e, without the involvement of user again-and-again).

4.7 Implementation

For our research, we use two marketplace/ e-commerce websites for extracting product prices such as mobile phones, electronic gadgets such as camera, printer, charger etc. We need to find the XPath of the data which we need to extract. We can do so by inspecting the web page by right clicking on the page and go to inspect element and then copy XPath.

We need to look into the web page to identify which element contains the information we want to extract. We use Python to scrape data from a marketplace.

We need to import some modules such as Requests and BeautifulSoup, as well as additional data analytic libraries such as numpy, Pandas & matplotlib.

The Requests Library/ module then generate HTTP requests to fetch the webpage link for the extraction of raw & unstructured data from the web elements. The HTML data is then parsed using the BeautifulSoup package. Now, the structured data can be analyzed and visualised with the help of matplotlib module and then it can be stored in database or in CSV or XLS format. Because the unstructured data is formatted and stored in a local data base or CSV file, it may be accessed easily at any time.

Any e-commerce site's data is constantly changing, which means that product pricing can change at any time, and certain items could be unavailable. This periodic change in data should be factored accordingly in the code. Big chunks of data are dynamically, fetched, extracted and saved in CSV format. For visualisation of the retrieved data, the information can be stored in CSV or XLS format and plotted using the matplotlib tool in Python.

Pseudocode Code:-

Step 1 Check for valid URL, If, Valid, call the URL.

Step 2 Check if element is present or not using XPath Locator.

Step 3 If, present, fetch text data, If, not present, go to Step 1

Step 4 Convert the fetch string data into integer/ float datatype

Step 5 Apply mathematical operations on the filtered data

Step 6 Convert the data into arrays, dataframe

Step 7 Visualise the data to get meaningful insight

Step 8 Store the data in local server/ database

CHAPTER 5

RESULTS & DISCUSSION

5.1 Results

Input data contains-

Let's take an example of a product to understand the input dataset-

Device Name - Blue - Iphone 12 (64GB)

Product Type - Mobile Phone

Product URL -

<https://www.amazon.in/New-Apple-iPhone-12-64GB/dp/B08L5WHFT9/>

Data element of Price -

`//*[@id="corePrice_desktop"]/div/table/div/tbody/tr[4]/span[2]span[1]/td[6]`

Now we will recall this URL and fetch the product price using the data element and then store it in a database. In our project we have extracted product price, rating & rating count. So, this is a dataset of a product from amazon website.

Example of Input dataset is shown in the table 5.1 below (Amazon):-

| Amazon_link | Amazon_price_xpath | Amazon_rating_xpath | Amazon_rating_count_xpath |
|---|--|--|---|
| https://www.amazon.in/Apple-iPhone-13-512GB-Blue/dp/B09G9JJT7M/ | //*[@id="corePrice_desktop"]/div/table/div/tbody/tr[4]/td[3]/span[1]/span[2] | //*[@id="corePrice_desktop"]/div/table/div/div/tbody/tr[6]/td[2]/span[1]/span[2] | //*[@id="review_sMedley"]/div/div[1]/div[2]/div[2]/span |

Table 5.1: Sample Dataset of Amazon

5.1.1 Extracted Raw Data

| | Product | Amazon_price | Amazon_rating | Amazon_rating_count | Flipkart_price | Flipkart_rating | Flipkart_rating_count |
|----|---|--------------|---------------|-------------------------|----------------|-----------------|-----------------------|
| 0 | Apple iPhone 13 (512GB) - Blue | ₹1,05,900 | 4 out of 5 | 3 global ratings | ₹1,04,900 | 4.7 | 2,743 Ratings & |
| 1 | Apple iPhone 12 (64GB) - Blue | ₹53,999.00 | 4.6 out of 5 | 11,155 global ratings | ₹59,999 | 4.6 | 1,65,417 Ratings & |
| 2 | Redmi 10 Prime (128 GB) | ₹14,999.00 | 4.1 out of 5 | 32,108 global ratings | ₹15,450 | 4.2 | 2,449 Ratings & |
| 3 | boAt Rockerz 330 Bluetooth Wireless | ₹1,499 | 4.1 out of 5 | 63,781 global ratings | ₹1,449 | 4.2 | 2,80,802 Ratings & |
| 4 | OnePlus Buds Z2 Pearl White | 4,999 | 4.3 out of 5 | 2,308 global ratings | ₹4,949 | 4.3 | 4,526 Ratings & |
| 5 | Logitech G102 Light Sync Gaming Mouse | 1,495 | 4.6 out of 5 | 6,316 global ratings | ₹1,445 | 4.6 | 3,320 Ratings & |
| 6 | Dell Pro MS5120W Wireless mouse | 2,057 | 4.3 out of 5 | 89 global ratings | ₹2,149 | 4.3 | 43 Ratings & |
| 7 | Logitech MK215 Wireless Keyboard and Mouse Combo | ₹1,195.00 | 4.2 out of 5 | 21,044 global ratings | ₹1,099 | 4.3 | 17,456 Ratings & |
| 8 | Seagate Expansion 1TB External HDD | 3,899.00 | 4.4 out of 5 | 1,17,667 global ratings | ₹3,799 | 4.4 | 67,112 Ratings & |
| 9 | Apple 20W USB-C Power Adapter | 1,899 | 4.6 out of 5 | 34,967 global ratings | ₹1,149 | 3.9 | 58 Ratings & |
| 10 | Samsung Galaxy M52 5G (ICY Blue, 6GB RAM, 128G... | ₹24,999 | 4.2 out of 5 | 7,545 global ratings | ₹29,999 | 4.3 | 461 Ratings & |

Figure 5.1: Extracted Raw Data

From figure 5.1 above, we can see the Raw data which is extracted with the help of product link & xpath locator. Extracted data is of string data type, therefore, we cannot apply mathematical operations on the data, So, We need to filter this data to make it useful for the user.

5.1.2 Filtered Data

This is filtered data, is free from garbage value and unnecessary information. Data like ‘₹’, ‘ ’, ‘,’ are removed from the extracted data as shown in figure 5.2 below and now, the data is converted into integer or float, so that, the user can make use of this data. This data will serve as base dataset for analyzing system.

| | Product | Amazon_price | Amazon_rating | Amazon_rating_count | Flipkart_price | Flipkart_rating | Flipkart_rating_count |
|----|---|--------------|---------------|---------------------|----------------|-----------------|-----------------------|
| 0 | Apple iPhone 13 (512GB) - Blue | 105900 | 4.0 | 3 | 104900 | 4.7 | 2743 |
| 1 | Apple iPhone 12 (64GB) - Blue | 53999 | 4.6 | 11155 | 59999 | 4.6 | 165417 |
| 2 | Redmi 10 Prime (128 GB) | 14999 | 4.1 | 32108 | 15450 | 4.2 | 2449 |
| 3 | boAt Rockerz 330 Bluetooth Wireless | 1499 | 4.1 | 63781 | 1449 | 4.2 | 280802 |
| 4 | OnePlus Buds Z2 Pearl White | 4999 | 4.3 | 2308 | 4949 | 4.3 | 4526 |
| 5 | Logitech G102 Light Sync Gaming Mouse | 1495 | 4.6 | 6316 | 1445 | 4.6 | 3320 |
| 6 | Dell Pro MS5120W Wireless mouse | 2057 | 4.3 | 89 | 2149 | 4.3 | 43 |
| 7 | Logitech MK215 Wireless Keyboard and Mouse Combo | 1195 | 4.2 | 21044 | 1099 | 4.3 | 17456 |
| 8 | Seagate Expansion 1TB External HDD | 3899 | 4.4 | 117667 | 3799 | 4.4 | 67112 |
| 9 | Apple 20W USB-C Power Adapter | 1899 | 4.6 | 34967 | 1149 | 3.9 | 58 |
| 10 | Samsung Galaxy M52 5G (ICY Blue, 6GB RAM, 128G... | 24999 | 4.2 | 7545 | 29999 | 4.3 | 461 |

Figure 5.2: Filtered Data

5.1.3 Best Price Analysis

Best Price analysis helps in providing the best price available for a product and it also suggest the corresponding marketplace. It the lowest price available for a product at the particular time. We can see a comparison between two e-commerce websites in terms of best price analysis in the figure 5.3 given below:

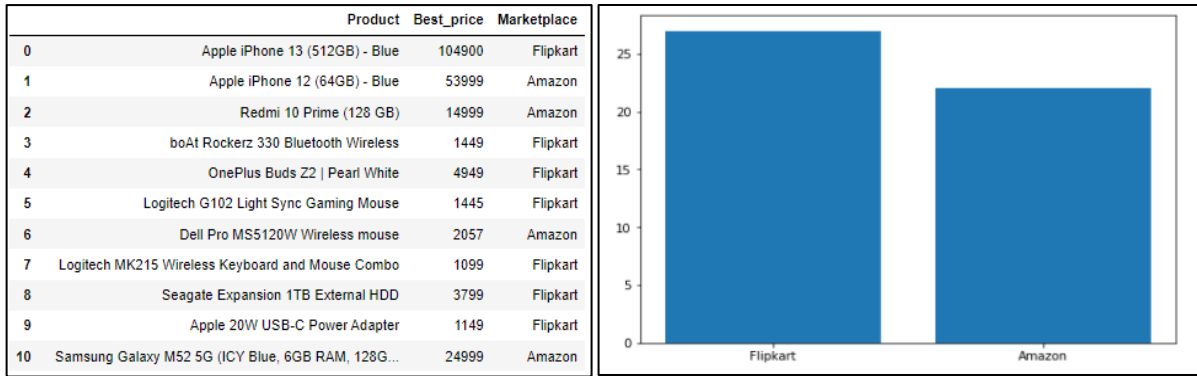


Figure 5.3: Best Price Analysis

5.1.4 Best Rating Analysis

Rating is the rate/ review given by the consumers of a product. It symbolizes that whether the product is liked by the consumers or not. It is usually measured out of 5. Best Rating analysis helps in providing the rating available for a product and it also suggest the corresponding marketplace as shown with a comparison if the figure 5.4 given below.

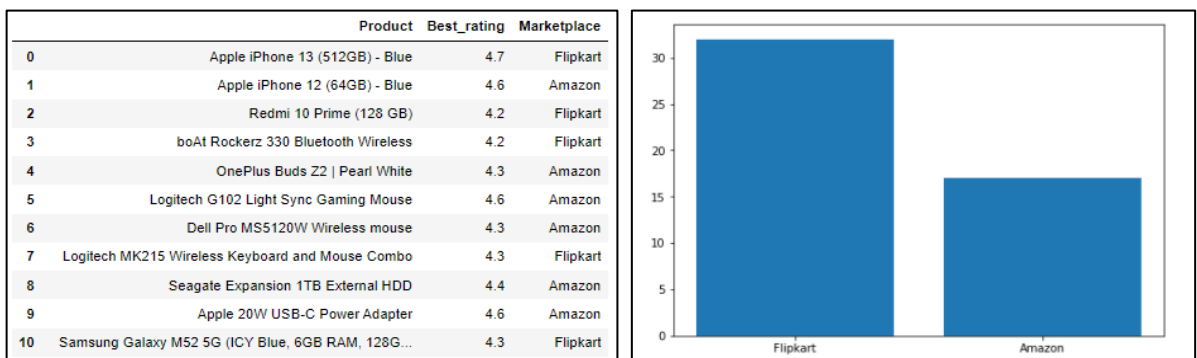


Figure 5.4: Best Rating Analysis

5.1.5 Best Rating Count Analysis

Rating count is the number of rates/ reviews given by the consumer of a product.

If 100 people buying a product that usually 10-15 people give the reviews/ rates. So, similarly we can say rating count is generally between 10-15% for a product. Rating count analysis tells us the best marketplace to buy a product or the marketplace users are more likely to buy a particular product. We can see the comparison in the figure 5.5 shown below:

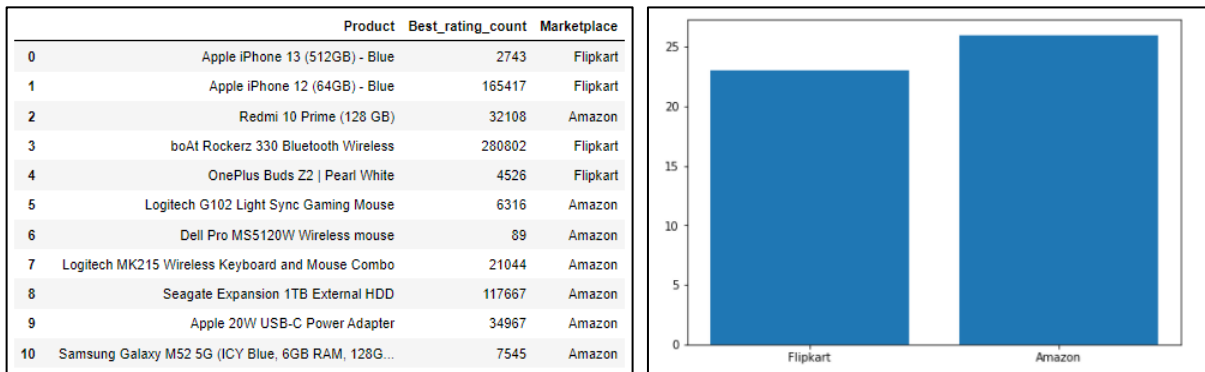


Figure 5.5: Best Rating Count Analysis

5.1.6 Analysis of data collected over a period of time

Here, we have collected the data of over one week and used this data for our analysis.

- **Amazon Rating count Analysis** – With reference to the figure 5.6 below, Rating count tells us the average number of sales of a product that could have taken place by a particular marketplace at a particular given time.

| Product | 11_04_2022_Amazon_rating_count | 12_04_2022_Amazon_rating_count | 13_04_2022_Amazon_rating_count | 14_04_2022_Amazon_rating_count | 15_04_2022_Amazon_rating_count | 16_04_2022_Amazon_rating_count | 17_04_2022_Amazon_rating_count |
|--|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|
| Apple iPhone 11 (512GB) - Blue | 3067 | 3070 | 3070 | 3070 | 3070 | 3071 | 3071 |
| Apple iPhone 12 (64GB) - Blue | 11138 | 11138 | 11138 | 11144 | 11144 | 11143 | 11155 |
| Redmi 10 Prime (128 GB) | 32098 | 32098 | 32101 | 32101 | 32101 | 32103 | 32108 |
| boAt Rockerz 330 Bluetooth Wireless | 63767 | 63769 | 63769 | 63771 | 63771 | 63775 | 63781 |
| OnePlus Buds Z2 Pearl White | 2308 | 2308 | 2308 | 2308 | 2308 | 2308 | 2308 |
| Logitech G102 Light Sync Gaming Mouse | 6315 | 6315 | 6315 | 6315 | 6316 | 6316 | 6316 |
| Dell Pro MS5120W Wireless mouse | 89 | 89 | 89 | 89 | 89 | 89 | 89 |
| Logitech MK275 Wireless Keyboard and Mouse Combo | 21042 | 21042 | 21042 | 21042 | 21042 | 21044 | 21044 |
| Seagate Expansion 1TB External HDD | 107650 | 107652 | 107652 | 107653 | 107653 | 107658 | 107667 |

Figure 5.6: Amazon Rating Count Analysis

- **Amazon Price Analysis** - With reference to the figure 5.7 below, Price Analysis tells us how the price changes over time. We can also conclude at what particular time the product was available at the lowest price.

| Product | 11_04_2022_Amazon_price | 12_04_2022_Amazon_price | 13_04_2022_Amazon_price | 14_04_2022_Amazon_price | 15_04_2022_Amazon_price | 16_04_2022_Amazon_price | 17_04_2022_Amazon_price |
|--|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| Apple iPhone 13 (512GB) - Blue | 105900 | 105900 | 105900 | 105900 | 105900 | 105900 | 105900 |
| Apple iPhone 12 (64GB) - Blue | 53999 | 53999 | 53999 | 53999 | 53999 | 53999 | 53999 |
| Redmi 10 Prime (128 GB) | 14999 | 14999 | 14999 | 14999 | 14999 | 14999 | 14999 |
| boAt Rockerz 330 Bluetooth Wireless | 1499 | 1499 | 1499 | 1499 | 1499 | 1499 | 1499 |
| OnePlus Buds Z2 Pearl White | 4999 | 4999 | 4999 | 4999 | 4999 | 4999 | 4999 |
| Logitech G102 Light Sync Gaming Mouse | 1495 | 1495 | 1495 | 1495 | 1495 | 1495 | 1495 |
| Dell Pro MS5120W Wireless mouse | 2057 | 2057 | 2057 | 2057 | 2057 | 2057 | 2057 |
| Logitech MK275 Wireless Keyboard and Mouse Combo | 1195 | 1195 | 1195 | 1195 | 1195 | 1195 | 1195 |
| Seagate Expansion 1TB External HDD | 3899 | 3899 | 3899 | 3899 | 3899 | 3899 | 3899 |

Figure 5.6: Amazon Price Analysis

5.2 In-Depth Analysis of Data

- **Example 1 –**

We have chosen a product at random and try to compare the change in rating count available at different marketplace.

- ✓ Comparison of Rating Count Data available on different marketplace in the figure 5.8 and 5.9:-

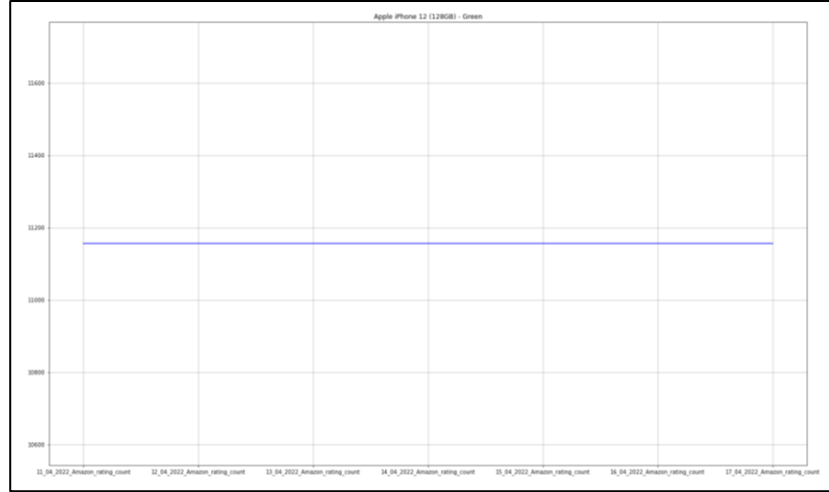


Figure 5.8: Apple iPhone 12 (128GB) – Green (AMAZON RATING COUNT)

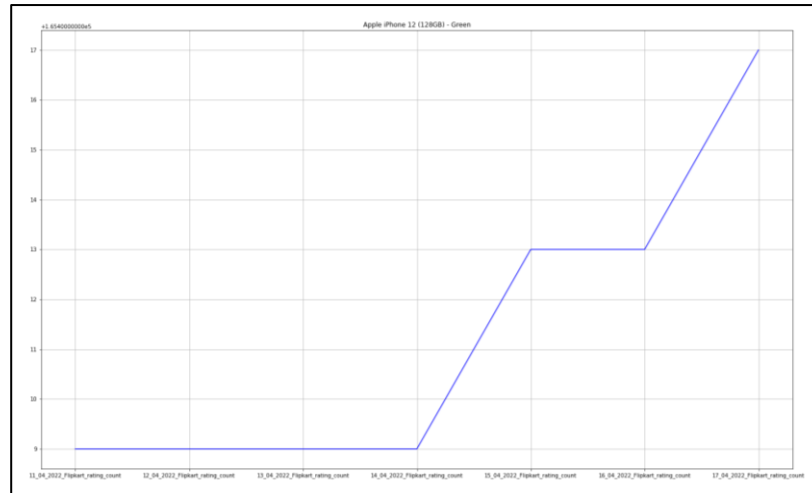


Figure 5.9: Apple iPhone 12 (128GB) – Green (FLIPKART RATING COUNT)

Here, we have selected Apple Iphone 12, 128GB of ROM & Green Color for our analysis. We can clearly see how rating count increases in Flipkart marketplace and a constant rating count in amazon marketplace over a period of one week. This shows consumers were interested in buying this product from Flipkart marketplace.

Similarly, Iphone 20W charger is having different rating count on different

marketplace. We can clearly see in figure 5.10 how rating count changes over a period of time on Amazon marketplace on the other hand in figure 5.11 we can clearly see a constant graph of rating count from flipkart marketplace. This show people were more inclined towards amazon when it comes to buy iphone charger.

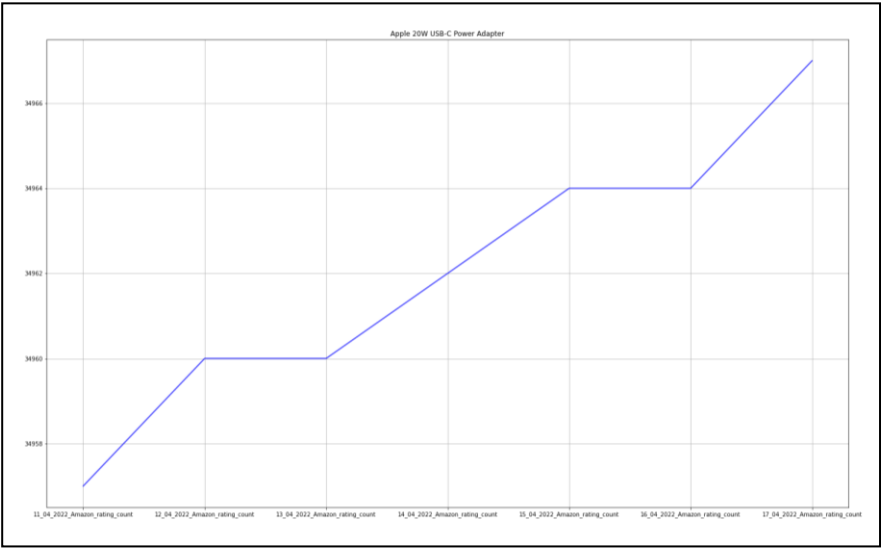


Figure 5.10: iphone 20W charger (AMAZON RATING COUNT)

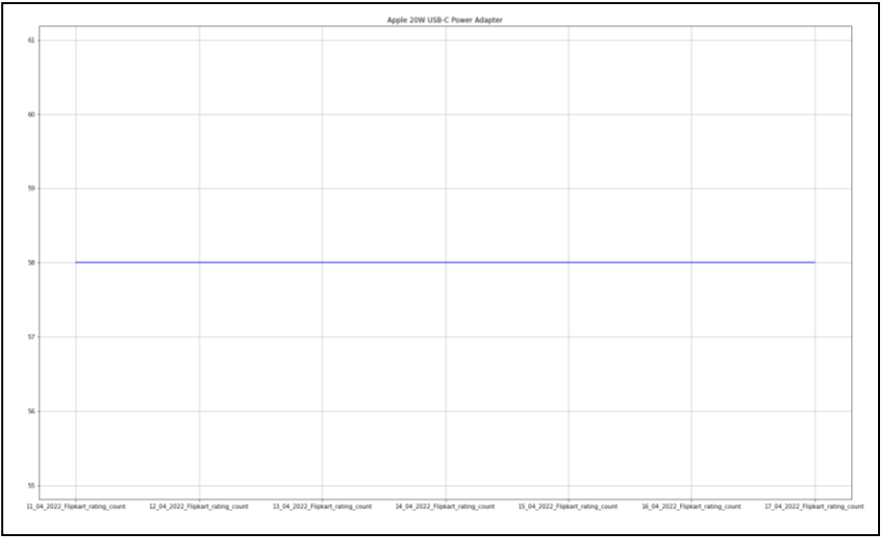


Figure 5.11: iphone 20W charger (FLIPKART RATING COUNT)

- **Example 2 –**

- ✓ Relation between Rating Count & Product Sales:-

- Product - boAt Rockerz 330 Bluetooth Wireless
- Factor – Rating Count
- MarketPlace – Amazon

Below graph 5.10 shows how rating count of the product increases over time, this also Shows how the demand of the product increases over time.

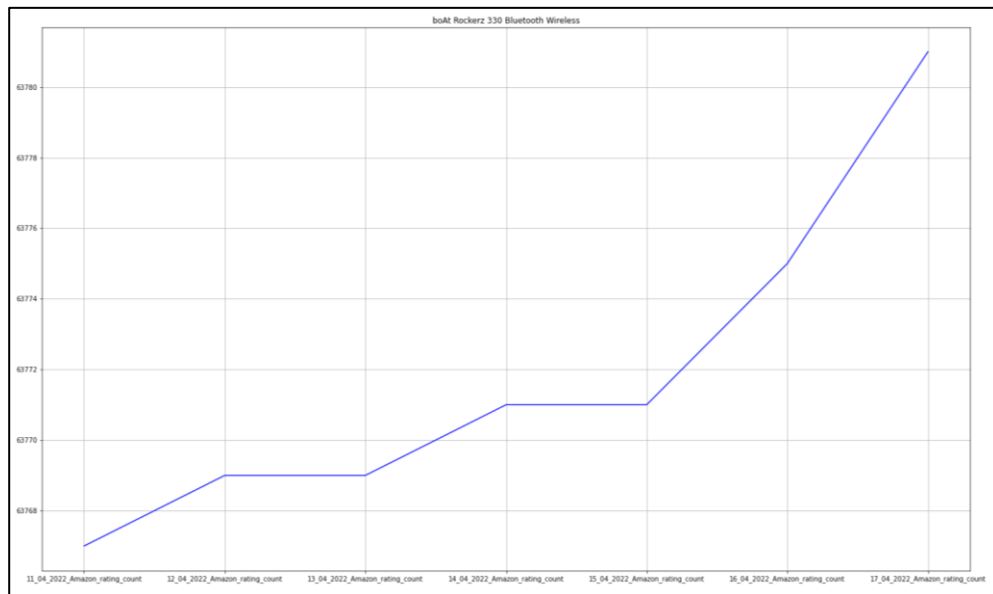


Figure 5.12: Apple iPhone 12 (128GB) – Green (FLIPKART RATING COUNT)

It is considered that out of 100 people buying a product approximately 10 to 15 people give their valuable reviews/ ratings. So, this means 10 to 15 reviews/ ratings symbolize the sales of approximately 100 products. Therefore, the change in rating count over a period of time, multiply by 10, gives the estimated

sales of a product over a period of time.

5.3 Conclusion

This project report provides a revolutionary collaborative Price filtering method for shopping sites based on Web scraping techniques. A relevant case study of analysis of price & comparison is done from marketplace or e-commerce websites, is given to show how the system works in practise. Compare prices on many web sources, if possible. Data can be scraped from a variety of sources, including social media, employment boards, and travel websites.

Prices may vary from time to time, so to get exact track of when the lowest price was available, we need to collect data for over a period of one year. Product price analysing system also helps in providing the life cycle of different type of products.

It also helps in comparison of sales of different types of products from different marketplace. Data extraction of product can help a business or startup to find a trending product, so, they can also boost their sales. The amount of data available on the internet is steadily expanding. Now this data can be stored in local data servers or computer for ease of retrieval. Information harvesting can be utilised in practically any case whereby knowledge needs to be transmitted from one area to another. Learning the foundations of data extraction is simple. Let's explore at how to make a simple data removal process in Excel.

The information taken here is not suitable for immediate use. Before we can use it, it need to go through some sort of purification procedure. Word transformation and query language can be used to accomplish this. It's worth

mentioning that extraction and alteration can be accomplished simultaneously. We may also leverage cloud computing to add a scheduler that will fetch the goods URL stored database as well as extract data such as sales price, rating, and name.

Web extraction and analysis of data will play a critical part in data extraction in the future due to the vast volumes of data available on the internet. Analysis of data also helps businesses to have a close look at their competitors' performance, they can also extract & analyse the available marketplace data & can conclude which marketplace is suitable for their product can compare the performance of the products with respect to other products.

5.4 Further enhancements that can be made:-

We can use multiple proxies we can distribute the requests made from user-end. Collecting of data & which type of data depends upon the user analysis. We can collect more data points such as meta data, tags, categories, breadcrumbs & other minute data which can help in better understanding and can analyse various KPIs (Key Performing Indicators). By using a scheduler we can schedule our requests at particular time & interval, this also prevents our requests from getting blocked.

Cloud servers are the remote storage or processing devices, which are used to run, automate tasks easily & help in data storage. By using cloud servers we can automate our bot to work at any particular time & interval.

CHAPTER 6

REFERENCES

6.1 References

- [1]Andreas Mehlführer. Web scraping: A tooevaluation. Technische Universität Wien, 2009.
- [2]Ananda V.Saurkar, G.Kedar&ShewetaGowde, “Detailed Study on Web filtration & Storage Access,” Open Access National Index ofScience & Information Engineering, pp. 336-341, Vol. 5, 2018
- [3]Andrew S TanenbaumandAlbert S Woodhull. Operating Systems Design andImplementation (3rd Edition). Prentice-Hall, Inc., UpperSaddleRiver, NJ, USA, 2005.
- [4]Binay Rahsd. “Web Extraction & Valuation using Python: Evaluating data available from the marketplace websites. " Inc., 2015.
- [5]Chaulagain, RamSharan, SubrnaShakrya. "Cloud Computing based online scrapingfor big-dataapplications." In 2017 IEEE InternationalConferenceonSmartCloud (SmartCloud), pp. 139-144. IEEE, (2017).
- [6]Eric J.Farley and Lissa Pierote, “An evolving information analyzing Method for Criminal & Law Justice Researchers,” Justice Analytics Research & statistical association, pp. 2-10, 2020.
- [7]F.Zhao, J.Zhou, C.Nie, H.Huang, andA.Jin, (2016) IndexCrawler: A Three-stage spider & crawler for valuable HarvestingDeep online Interfaces,

IEEE transectionsoncomputingservices.

[8]Fedireco Polodiro, Ricardo Gianini, Rosana Lonte, StafenoMonsca& Francessa Rosetti, "Online Extraction ways to fetch info. on Commercial hardware and airlines for American HCIP," Statisstical Journal of the IOAS, pp. 156-167, 2015.

[9]Joacim Olofsson. Evaluation of web scraping tools for creating an embedded web wrapper. page 44. KTH, School of Computer Science and Communication (CSC), 2016.

[10]Jane Kine and Janna Axanbeck, "Online Minning of Corporate Web Pages", International Innovative Research Journal in Com. and Comm. Engineering, 2019.

[11]Kaushal Parikh, Dilip Singh, Dinesh Kumar and Mansingh Rathode, "Machine Intelligence in Online Extraction," KGMP internatinal journal of Engineering & commucation, pp.141-150, Vol. 3, 2018.

[12]M.Assuno, R.Calheiros, S.Bianchi, M. Netto, and R.Buyya, "Big data computing and clouds: Trends and future directions," Journal of Parallel and Distributed Computing, 59 (2015), pp.4 -16.

[13]M. K. Khakhani, S. Khakhani, and S.R.Biradar, Research challenges in big data analytics, International Journal of Application or Innovation in Engineering and Management, 2(8) (2015), pp.228-232.

[14]Mitchal, Rayn. "Online Extraction with .py: Collecting data available from the modern web pages. "O'Reilaly Plat., Inc., 2018.

[15]Madhusudan, Lumbhate Ponam, D. (2018) Deep-Web CrawlingEfficiently using static based python bot Crawlers, International Journal of Computer Science &Engineering (IRJET), 04(07), pp. 3304.

[16]Mitchal, Rayn. "Web Extraction & Valuation using Python: Evaluating data available from the marketplace websites. "O'Reilaly Plat., Inc., 2019.

- [17]S.Sirisuriya. Practical Working on online scrapping. International Research Conference, KDU, 8:135–139, 11 2015.
- [18]S. Del Rio, V. Lopez, J. M. Bentez, and F. Herrera, Information Sciences, 285 (2014), pp.112-137, on the application of mapreduce for imbalanced big data utilising random forest.
- [19]Sameer Padghan, Satish Chigle and Rahul Handoo, “Online Extraction of Media Websites,” International Innovative Research Journal in Com. and Comm. Engineering, pp. 692-696, Vol.16, 2018.
- [20]T. J. McCabe. A complexitymeasure. IEEE Transactionson Software Engineering, SE-2(4):308–320, Dec 1976.
- [21]X.Y.Chen and Z. G. Jin, Research on key technology and applications for the internet of things, Physics Procedia, 33, 561-566 (2012).
- [22]Yolande Neil. Web scraping the easyway. GeorgiaSouthern University, 2016.
- [23]Yen A. Gandomi and M. Haider, Beyond the Hype: Big Data Concepts, Methods, and Analytics, International Journal, 2012.

APPENDIX - I

SOURCE CODE

Extraction

#Extract product data from Flipkart

```
baseDataUrl = input_df['Flipkart_link'][i]
```

```
try:
```

```
    request.get(baseDataUrl)
```

```
except request.ConnectionError:
```

```
    #print("Link Not Working")
```

```
    F_L = "UPDATE"
```

```
    F_price_X = "LINK NOT WORKING"
```

```
    F_price = "LINK NOT WORKING"
```

```
    F_rating_X = "LINK NOT WORKING"
```

```
    F_rating = "LINK NOT WORKING"
```

```
    F_rating_count_X = "LINK NOT WORKING"
```

```
    F_rating_count = "LINK NOT WORKING"
```

```
else:
```

```
    print(str(input_df['Product'][i]) + " - FLIPKART Link Working")
```

```
    F_L = "OKAY"
```

```
    driver.get(baseDataUrl)
```

```
    xpath = input_df['Flipkart_price_xpath'][i]
```

```
    #print(xpath)
```

```

try:
    driver.find_element_by_xpath(xpath)
except NoSuchElementException:
    print("No such xpath")
    F_price_X = "UPDATE"
    F_price = "XPATH NOT WORKING"
else:
    print("xpath present")
    F_price_X = "OKAY"
    data = driver.find_element_by_xpath(xpath)
    print(data.text)
    F_price = data.text

```

```

xpath = input_df['Flipkart_rating_xpath'][i]

```

```

#print(xpath)

```

```

try:
    driver.find_element_by_xpath(xpath)
except NoSuchElementException:
    print("No such xpath")
    F_rating_X = "UPDATE"
    F_rating = "XPATH NOT WORKING"
else:
    print("Rating xpath present")
    F_rating_X = "OKAY"
    data = driver.find_element_by_xpath(xpath)
    print(data.text)
    F_rating = data.text

```

```

xpath = input_df['Flipkart_rating_count_xpath'][i]
#print(xpath)

try:
    driver.find_element_by_xpath(xpath)
except NoSuchElementException:
    print("No such xpath")
    F_rating_count_X = "UPDATE"
    F_rating_count = "XPATH NOT WORKING"
else:
    print("Rating count xpath present")
    F_rating_count_X = "OKAY"
    data = driver.find_element_by_xpath(xpath)
    print(data.text)
    F_rating_count = data.text

log_dic = {'Product':input_df['Product'][i],
           'Amazon_link':[A_L],
           'Amazon_price_xpath':[A_price_X],
           'Amazon_rating_xpath':[A_rating_X],
           'Amazon_rating_count_xpath':[A_rating_count_X],
           'Flipkart_link':[F_L],
           'Flipkart_price_xpath':[F_price_X],
           'Flipkart_rating_xpath':[F_rating_X],
           'Flipkart_rating_count_xpath':[F_rating_count_X]}

log_dic_df = pd.DataFrame(log_dic)

```



```
merge = LOG_df.append(log_dic_df)
```

Analysis

```
filter_df.rename(columns=dict, inplace=True)
master_df = pd.concat([master_df, filter_df], axis=1)
master_df.to_excel('Analyse/master.xlsx', index = False)
dt_list = ['11_04_2022', '12_04_2022', '13_04_2022', '14_04_2022',
'15_04_2022', '16_04_2022', '17_04_2022']
txt = '_Amazon_rating_count'
col = ['Product']
for i in range(len(dt_list)):
    col.append(dt_list[i] + txt)

amazon_df = pd.DataFrame(columns = col)
amazon_df['Product'] = master_df['Product'].copy()
amazon_df['11_04_2022_Amazon_rating_count'] =
master_df['11_04_2022_Amazon_rating_count'].copy()
amazon_df['12_04_2022_Amazon_rating_count'] =
master_df['12_04_2022_Amazon_rating_count'].copy()
amazon_df['13_04_2022_Amazon_rating_count'] =
master_df['13_04_2022_Amazon_rating_count'].copy()
amazon_df['14_04_2022_Amazon_rating_count'] =
master_df['14_04_2022_Amazon_rating_count'].copy()
amazon_df['15_04_2022_Amazon_rating_count'] =
master_df['15_04_2022_Amazon_rating_count'].copy()
```

```


amazon_df['16_04_2022_Amazon_rating_count'] =
master_df['16_04_2022_Amazon_rating_count'].copy()
amazon_df['17_04_2022_Amazon_rating_count'] =
master_df['17_04_2022_Amazon_rating_count'].copy()
amazon_df.to_excel('Analyse/Amazon/Amazon_rating_count.xlsx', index =
False)
analyse = pd.read_excel('Analyse/Amazon/Amazon_rating_count.xlsx')
#analyse

for i in range(0,len(analyse['Product'])):
    x = list(analyse.columns)
    x = x[1:]
    y = list(analyse.loc[i])
    y = y[1:]
    plt.figure(figsize=(25, 15))
    plt.grid()
    plt.plot(x, y, color = "Blue")
    title = str(analyse['Product'][i])
    plt.title(title)
    file_name = 'Analyse/Amazon/' + title + '.png'
    file_name = file_name.replace('"', "")
    plt.savefig(file_name)

```

APPENDIX - II

PUBLICATION PROOF

Akshat Goel <goelakshat.contact@gmail.com>


Thanks for your submission
1 message

info@coneco2009.com <info@coneco2009.com> 3 May 2022 at 09:32
To: goelakshat.contact@gmail.com

Dear Author,

We received your paper. We will send notification after review. Thanks for your submission.

Thanks
Secretary

Akshat Goel <goelakshat.contact@gmail.com>

[EASR] Submission Acknowledgement
1 message

Editor of Engineering and Applied Science Research via Thai Journals Online (ThaiJO) 10 May 2022 at 20:36
<admin@tcj-thaijo.org>
Reply-To: Editor of Engineering and Applied Science Research <kku.enjournal@gmail.com>
To: Akshat Goel <goelakshat.contact@gmail.com>

Akshat Goel:

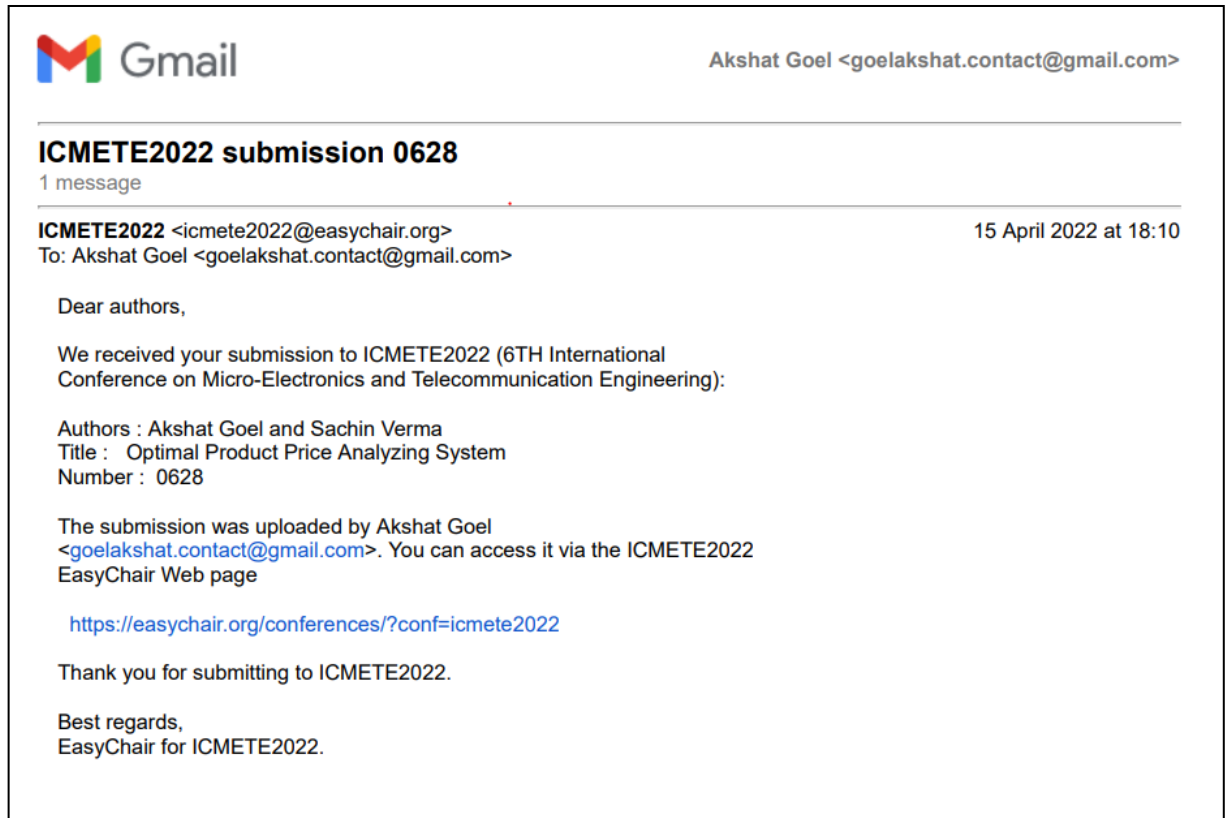
Thank you for submitting the manuscript, "Optimal Product Price Analyzing System" to Engineering and Applied Science Research. With the online journal management system that we are using, you will be able to track its progress through the editorial process by logging in to the journal web site:

Manuscript URL: <https://ph01.tci-thaijo.org/index.php/easr/authorDashboard/submission/248412>
Username: goelakshat.contact@gmail.com

If you have any questions, please contact me. Thank you for considering this journal as a venue for your work.

Editor of Engineering and Applied Science Research

CONFERENCE PROOF



PLAGIARISM REPORT

| | | | |
|--------------------|--|--------------|----------------|
| sachin_akshatv3 | | | |
| ORIGINALITY REPORT | | | |
| 4% | 3% | 0% | 3% |
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |
| PRIMARY SOURCES | | | |
| 1 | Submitted to Victoria University Student Paper | 1 % | |
| 2 | www.smart-bear.eu Internet Source | <1 % | |
| 3 | origin.geeksforgeeks.org Internet Source | <1 % | |
| 4 | Submitted to Colorado State University, Global Campus Student Paper | <1 % | |
| 5 | www.edwardhk.com Internet Source | <1 % | |
| 6 | Submitted to Istanbul Aydin University Student Paper | <1 % | |
| 7 | Submitted to University of Greenwich Student Paper | <1 % | |
| 8 | Submitted to Institute of Technology, Nirma University Student Paper | <1 % | |
| 9 | files.webis.de | | |

| | | |
|---|---|------|
| | Internet Source | <1 % |
| 10 | www.slideshare.net Internet Source | <1 % |
| 11 | expskill.com Internet Source | <1 % |
| 12 | knowledgecommons.lakeheadu.ca Internet Source | <1 % |
| 13 | su-plus.strathmore.edu Internet Source | <1 % |
| <div> <div>Exclude quotes</div> <div>On</div> <div>Exclude matches</div> <div>< 10 words</div> </div> <div> <div>Exclude bibliography</div> <div>On</div> </div> | | |

| SRM INSTITUTE OF SCIENCE AND TECHNOLOGY (Deemed to be University u/s 3 of UGC Act, 1956) | | |
|--|--|--|
| Office of Controller of Examinations | | |
| REPORT FOR PLAGIARISM CHECK ON THE SYNOPSIS/THESIS/DISSERTATION/PROJECT REPORTS | | |
| 1 | Name of the Candidate (IN BLOCK LETTERS) | AKSHAT GOEL, SACHIN VERMA |
| 2 | Address of the Candidate | SRMIST, KATTANKULATHUR Mobile Number : 9896297313, 9205144564 |
| 3 | Registration Number | RA1811030010069, RA1811030010085 |
| 4 | Date of Birth | 06/09/2000, 24/02/2000 |
| 5 | Department | DEPARTMENT OF NETWORK AND COMMUNICATION |
| 6 | Faculty | DR. S. PRABAKERAN |
| 7 | Title of the Synopsis/ Thesis/ Dissertation/Project | OPTIMAL PRODUCT PRICE ANALYSING SYSTEM |
| 8 | Name and address of the Supervisor / Guide | Assistant Professor School of Computing - Department of Networking and Communications SRM Institute of Science & Technology (SRMIST) Kattankulathur, Chengalpattu District - 603 203 Tamil Nadu, India Mail ID : prabakes@srmist.edu.in Mobile Number : 9042394880 |
| 9 | Name and address of the Co-Supervisor / Co- Guide (if any) | Mail ID : Mobile Number : |
| 10 | Software Used | TURNITIN |
| 11 | Date of Verification | 12/05/2022 |

| 12 | Plagiarism Details: (to attach the final report) | | | |
|--|--|--|--|---|
| Chapter | Title of the Chapter | Percentage of similarity index (including self citation) | Percentage of similarity index (Excluding self citation) | % of plagiarism after excluding Quotes, Bibliography, etc., |
| 1 | Introduction | 2% | <1% | <1% |
| 2 | Literature Survey | 1.3% | <1% | <1% |
| 3 | System Analysis | 1% | <1% | <1% |
| 4 | System Design | <1% | <1% | <1% |
| 5 | Results and discussion | <1% | <1% | <1% |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| 9 | | | | |
| 10 | | | | |
| Appendices | | 4% | 3% | 1% |
| I / We declare that the above information have been verified and found true to the best of my / our knowledge. | | | | |
| Signature of the Candidate | | Signature of the Supervisor / Guide | | |
| Signature of the Co-Supervisor/Co-Guide | | Signature of the HOD | | |

Date : 12/05/2022