A

PROJECT REPORT

ON

**AMAZON WEB SCRAPING AND VISUALIZATION**

A red and white logo

Description automatically generated

**SUBMITTED BY**

Jay Shukla 002619454

**SUBMITTED TO**

Northeastern University

April 17, 2024

**ACKNOWLEDGEMENT**

I would like to express our humble and sincere gratitude to my mentor, Huang Junwei sir for providing me the opportunity to indulge myself in this project which is enlisted in the syllabus of the course INFO7390 Advances Data Science.

**TABLE OF CONTENTS**

[**INTRODUCTION**](#_5g9l6b83lt4i)…………………………………………………………………………………….3

[**TOOLS AND TECHNOLOGIES USED**](#_y5z5p8wflqnk)…………………………..………………………………4

[**METHODOLOGY**](#_68tnu76ze9gb)……………………………………………………………………………………5

[**CONCLUSION**](#_k74zlti8ya6a)……………………………………………………………………………………...13

[**REFERENCES**](#_hg08rz5zv31y)……………………………………………………………………………………..13

# INTRODUCTION

This is the extension of my first project which was amazon web scrapping. Here in this project, I have used various tools and technologies to complete the given project tasks as mentioned in the objectives section. I have used the core concepts of data engineering like extract, transform and load in this project. To complete the project objectives, I have scraped my data from Amazon.com[[5]](https://www.amazon.com/s?k=laptop). I have scraped data related to Laptops in this project as it aligned with the project objectives. Then I processed the collected data using Python and Pandas and visualised using Python packages like Matplotlib and Seaborn. Finally, I prepared a findings report describing all the findings and insights we found from the data.

**PROJECT OBJECTIVES**

* To use Scrapy to collect data from Amazon
* To visualise the plot of price vs products (group by)
* To prepare the report of the findings (patterns)

**SCOPE AND APPLICATIONS**

The project basically contains visualization of laptop products and its relevant components with prices. So the visualization shows how the price differs with different brands of laptops and other components of the laptops. This visualization may have the following scope and applications:

* Help laymen to get laptop products according to their requirements and budget.
* Analyse pricing strategies of competitors and make informed decisions on pricing.
* Gather data on features, specifications and customer preferences to inform product development decisions.
* Gain insights into preferences through customer reviews.
* Track the performance, reputation and other products’ information of own brand and those of the competitors.

# TOOLS AND TECHNOLOGIES USED

1. **Python**

Python is a high-level, interpreted, and general-purpose programming language. Python offers various key features like simplicity, readability, versatility, etc. making it suitable for a wide range of applications like web development, scientific computing, data engineering, data analysis, data processing, machine learning and more. Also, Python is also an object-oriented programming language, meaning that it is built around the concept of objects and classes. This makes it easy to create complex and modular programs, as well as to reuse code and make modifications to existing code.

1. **Scrapy**

Web scraping is a technique used to extract data from websites. It involves making HTTP requests to a website's server, downloading the HTML content of the web page, and then parsing that data to extract the desired information. We have used Scrapy for web scraping in this project to collect data about Laptop from Amazon.com. Scrapy is a free and open-source web crawling framework for Python. It provides a way to extract the data from websites and save it in the desired format such as CSV, JSON, or XML. Scrapy is also useful for performing tasks such as crawling multiple pages, following links, and storing the data in a structured format[[1]](https://scrapeops.io/python-scrapy-playbook/scrapy-beginners-guide/).

1. **Pandas**

Pandas is a software library written for the Python programming language for data manipulation and analysis. It provides data structures for efficiently storing large datasets and tools for working with them. Pandas was used for analysing the collected data and for cleaning the data.

1. **Matplotlib**

Matplotlib is a data visualisation library in Python. It provides a high-level interface for creating static, animated, and interactive visualisations in Python. Matplotlib allows users to create a wide range of visualisations, including line plots, bar plots, scatter plots, histograms, pie charts, box plots, error bars, etc. It is a powerful tool for data exploration and analysis, and is widely used in data science and scientific computing. Matplotlib can be easily integrated with other libraries in the PyData ecosystem, such as Pandas, to enable the creation of complex visualisations with just a few lines of code. So, we have used Matplotlib to visualise the various relationships in the collected dataset.

1. **Seaborn**

Seaborn is a data visualisation library built on top of Matplotlib in Python. It provides a high-level interface for creating attractive and informative statistical graphics. Seaborn is designed to make it easier to create visualisations that show the relationships between variables in a dataset. It has several built-in functions for creating commonly used visualisations, such as scatter plots, line plots, bar plots, histograms, and density plots. Seaborn also includes advanced features for visualising the distribution of data and for comparing multiple variables in a single plot. In addition, Seaborn has built-in themes for styling and customising visualisations, making it easier to create aesthetically pleasing and informative graphics.

# METHODOLOGY

**Basic Workflow of Project**



Figure 1: Basic workflow of the project

As shown in the figure above, the project was completed in the following steps:

1. Web Scraping using Scrapy
2. Data Cleaning and Processing using Pandas
3. Data Visualization using Matplotlib and Seaborn
4. Finding Patterns and Insights from data
5. Preparing Project Documentation

**1. Web Scraping using Scrapy**

To collect data necessary for my project, I performed web scraping from Amazon.com[[5]](https://www.amazon.com/s?k=laptop) using a Python package named Scrapy. The aim of web scraping in this project is to collect data suitable for visualisation of products vs price. There are many products suitable for this task, but I collected data specifically related to laptops, as it was very relatable and useful to us. To perform web scraping using Scrapy, we performed the following steps:

1. Firstly, Scrapy was installed in a virtual environment.
2. A new Scrapy project was created using Scrapy.
3. A spider was created to scrape the necessary data from Amazon website.

The code for this Scrapy project.

After scraping about 3000 webpages, I collected 2843 data which was stored in a csv file.

The csv file contained following columns (laptop features):

Table 1: Information about originally scraped Laptop data

| SN | Column Names | Number of data |
| --- | --- | --- |
| 1 | Title | 2843 |
| 2 | Brand | 2810 |
| 3 | Screen\_Size | 2756 |
| 4 | CPU\_Model | 2631 |
| 5 | Ram | 2711 |
| 6 | Operating\_System | 2631 |
| 7 | Rating | 2631 |
| 8 | Series | 2501 |
| 9 | Colour | 2501 |
| 10 | Resolution | 182 |
| 11 | Weight | 66 |
| 12 | USB | 35 |
| 13 | Battery | 7 |
| 14 | Graphics | 2063 |
| 15 | Disk\_size | 2248 |
| 16 | Card\_desc | 2248 |
| 17 | Price | 1863 |

**2. Data Cleaning using Pandas**

The raw scraped data was noisy and contained many dirty and irrelevant values, so cleaning the data was necessary before performing visualisation. A subset of collected data can be seen in figure 2.

.

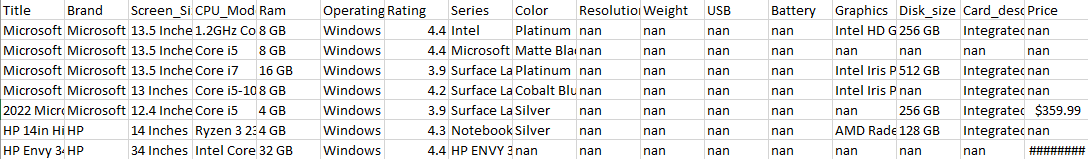


Figure 2: Original raw data collected from web scraping.

My data consists of many nan values in different columns, this was because most of the fields were not mentioned in the webpage on Amazon. After some analysis, I found that all the columns were not relevant for my visualisation purpose, so I only selected a subset of columns from the original dataset. These columns were as follows:

1. Title,
2. Price,
3. Rating,
4. Brand,
5. Screen\_Size,
6. Ram,
7. Disk\_size.

The subset can be seen in figure below:

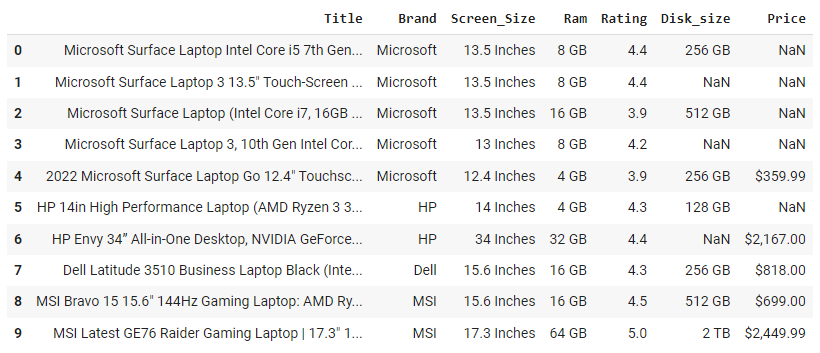


Figure 3: Subset of original data considering only relevant columns

The first step performed in data cleaning was removing all the nan values, either by dropping the rows or by filling the nan values. Except for the Rating column, we dropped the rows with nan values for all other columns. For the Rating column, we filled the nan values with a rating value of 0 which indicates that the product was not rated. The code snippet for cleaning nan values is shown below:

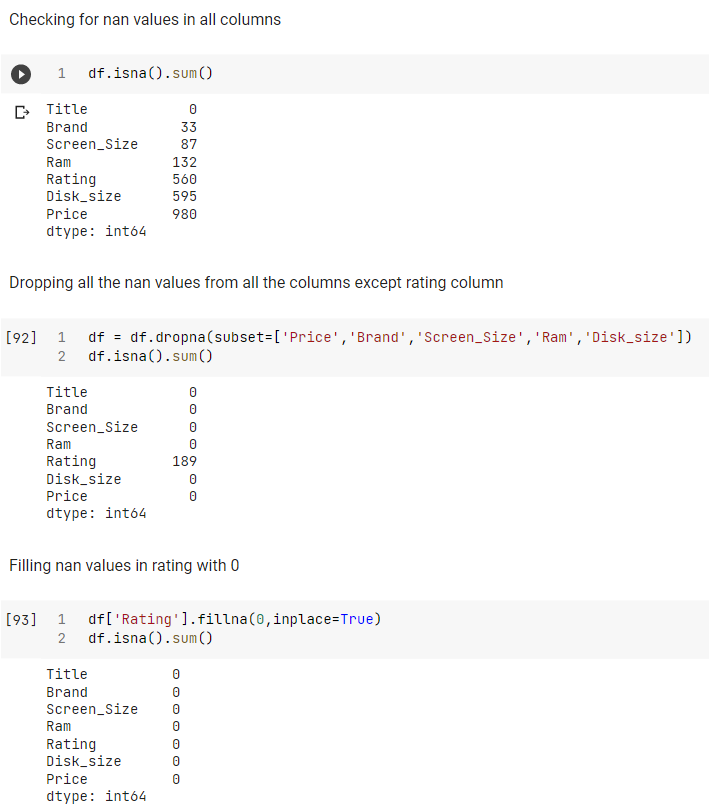
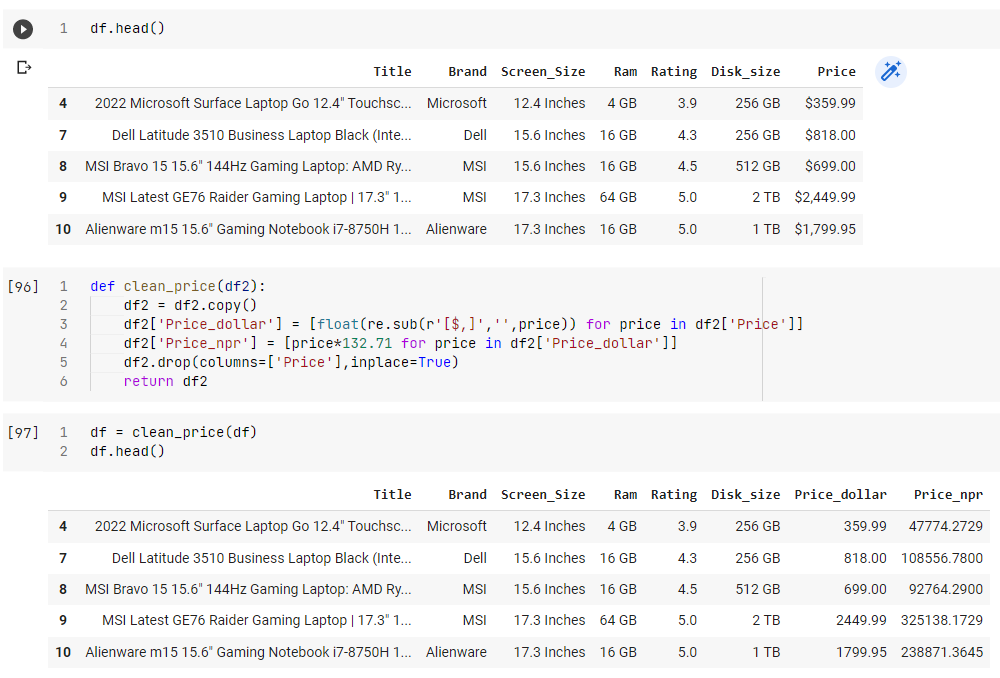


Figure 4: Cleaning nan values from the dataset

After removing all the rows with nan values, we obtained a dataset containing a total of 1448 rows and 7 columns.

Then I cleaned the Price column, Price column originally has data in the form of $X as I only needed the numerical value X, so cleaning was needed. Also, I added a new column to check the price in INR (Indian Rupees). The code snippet for cleaning the price column is shown in Figure 5.

Figure 5: Cleaning Price column

Similarly, I cleaned the RAM and Disk size columns, which had the values in form X GB/TB to

numeric value X (in GB). Also, the Screen size originally had values in the form of X inches which was cleaned to numerical value X. The code snippet for cleaning RAM, Disk size and Screen size are shown in Figure 6 and 7 respectively.

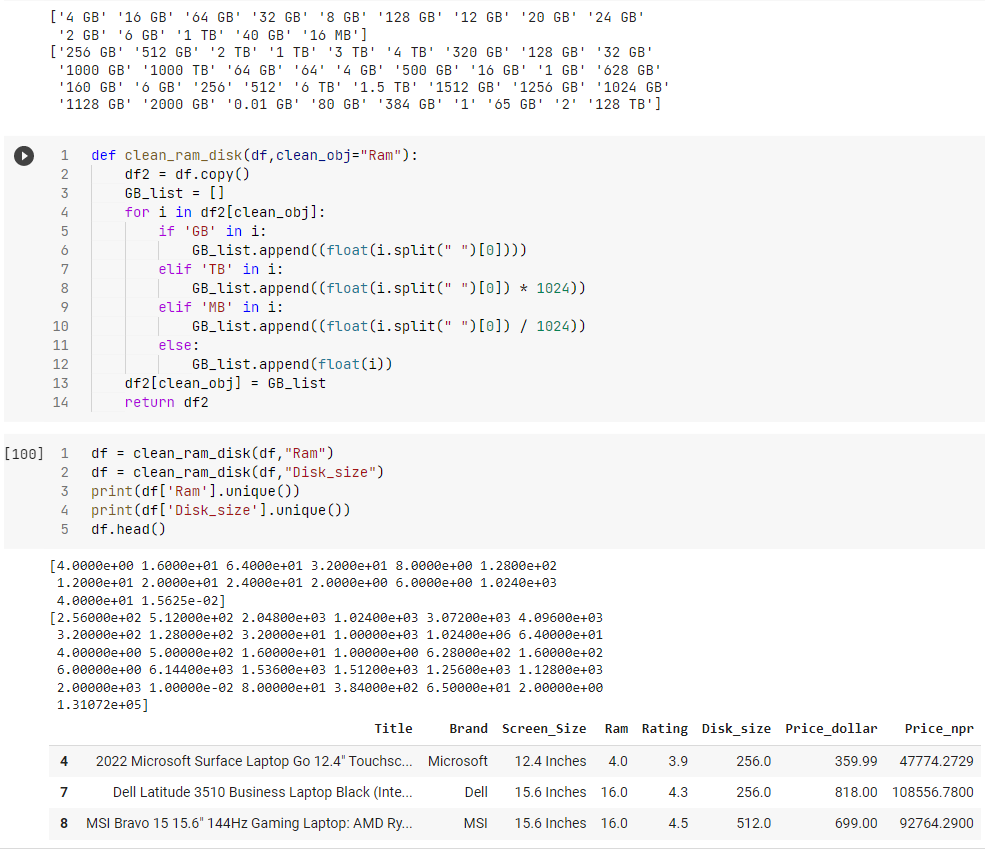


Figure 6: Cleaning RAM and Disk Size column

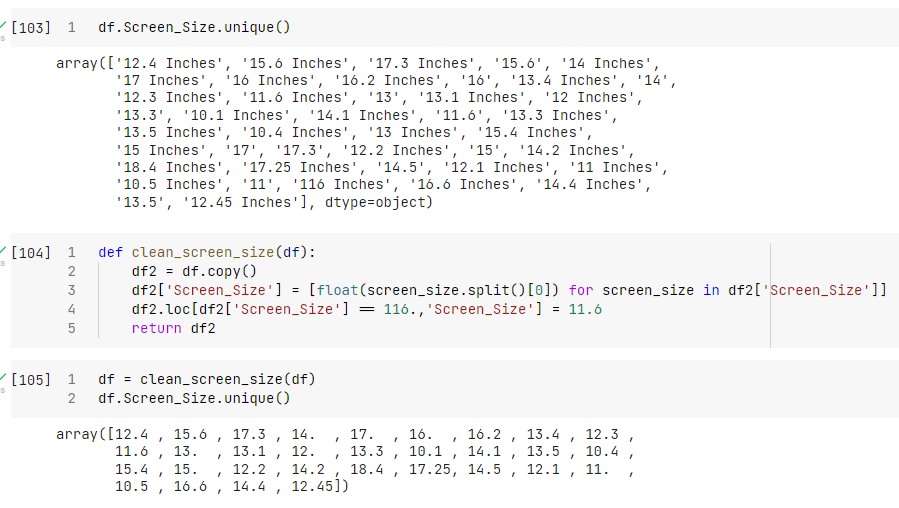


Figure 7: Cleaning Screen Size column

Finally, I checked for the duplicate rows in my dataset. I have considered rows with the same Title as duplicates. About 266 duplicate values were found, and all these duplicates were removed from the dataset. The code snippet for removing duplicates is shown in Figure 8. Finally, the cleaned data with 1182 rows and 7 columns was exported and used as a dataset for visualisation. The entire code for cleaning data can be found in the Jupyter notebook visualization.ipynb .

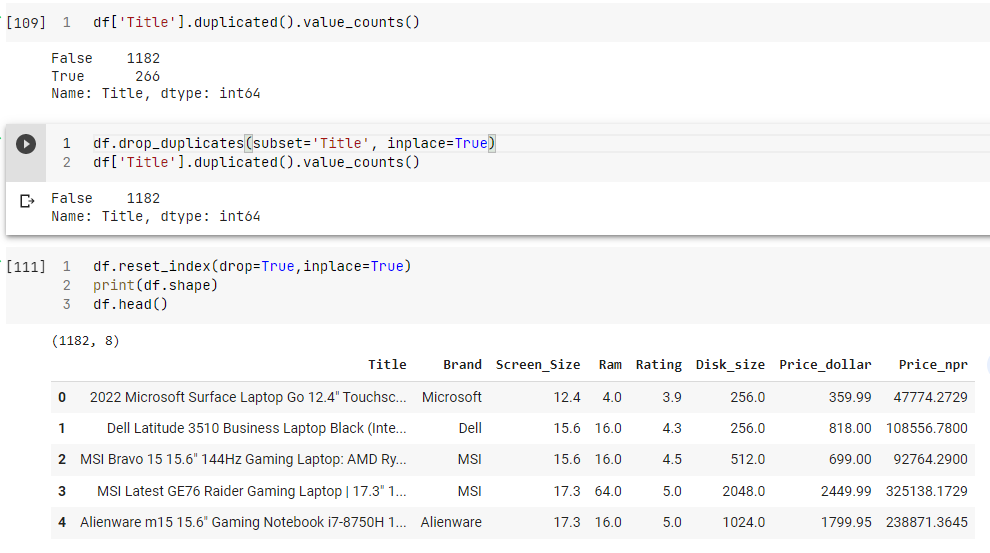


Figure 8: Removing duplicates from data

**3. Data Visualisation using Matplotlib and Seaborn**

After cleaning the data using pandas, the cleaned data was used for completing my second project objective, i-e to visualize the plot of price vs products. For visualisation, I have used Python’s Matplotlib and Seaborn library. I have performed various visualisations, which are shown in the Jupyter notebook visualization.ipynb .

**4. Finding patterns and insights**

Then I prepared the report on my findings from visualization of data. After visualizing the cleaned data, we found various important insights and findings from the data, which could be useful in the decision-making process. All these insights and findings were recorded and described in the findings report. The report was prepared using Google Docs.

**5. Preparing Project Documentation**

Finally, after completing all the project objectives, I prepared this project documentation using Google Docs. Here in this documentation, I have mentioned an introduction to my project, its main objectives, tools and technologies used and our basic project workflow.

# CONCLUSION

In conclusion, the project related to web scraping of Amazon's Product (laptops in my case) and their prices was a success. I was able to gather and organise a large amount of data on laptops and their prices using Scrapy and pandas libraries in Python. This data was then analysed through different plots plotted with the help of matplotlib to gain insights into the laptop market on Amazon. Insights on my analysis can be found on the finding report. This information can be useful for customers looking to purchase laptops, as well as for businesses looking to monitor the competition and adjust their own prices accordingly. The results of this project highlight the importance and usefulness of web scraping in today's data-driven world.

# REFERENCES

1. *Scrapy Beginners Series Part 1 - First Scrapy Spider | ScrapeOps*. (n.d.). https://scrapeops.io/python-scrapy-playbook/scrapy-beginners-guide/
2. *Amazon.com: Laptop*. (n.d.). https://www.amazon.com/s?k=laptop