

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
|  | | WEB SCRAPPING PROJECT REPORT | | | | |  | |
|  |  | | | | | | |  |
|  | | | |  |  | | | |
|  | | | | Hussain Raza RizviSp23-MSCS-0012Usman Bin HameedSP23-MSCS-0014 |  | | | |
|  | | | | 26-May-2023— — —APPLIED PROGRAMMING— — —Dr. Abdul Qadar Kara |  | | | |
|  | | |  | | |  | | |
|  | | | WEB PAGE: [www.priceoye.pk](http://www.priceoye.pk) | | |  | | |

1. Introduction To the Problem.

The Problem we fixed was to Get data from a webpage and use it for visualizations,  
This use case can be easily present in some rival companies so that they easily keep track of rate lists and helps customer in a better way compared to other competitors.

1. Explain the Tasks You Preformed.

* Web scrapping:

Web scraping is a technique used to extract data from websites, we used BeautifulSoup, which provides an easy-to-use interface for parsing HTML documents.

Initially we send a hit on "priceoye.pk", We Beautified it response by Html.parser. Due to this we were able to read Html code and figure out where we have our important Data, we used soup.find\_all("a", class\_="CLASS NAME") , we find divs having our exact data, and with the help of For Loop we were able to fetch that data in a list.

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated with medium confidence

* Data Cleaning:

We have performed Data Cleaning which is an essential step in the data analysis process. It focuses on identifying and rectifying errors, inconsistencies, and inaccuracies within a dataset. It encompasses a range of activities, including handling missing values, removing duplicates, addressing outliers, standardizing formats, and resolving inconsistencies

But according to our data we didn’t needed that, We added Column Names and removed inconsistency in Data as extra spaces, converted some of the columns into float64 and removed PKR from Amount so that it doesn’t create any problem in Visualization.

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

* Data visualization:

Data visualization is an important step in the data analysis process that aims to represent data in a visual format, such as charts, graphs, or plots. Visualization enables effective exploration, analysis, and communication of insights derived from the data, by Plots and Graphs understanding data is easy.

We made many plots so that we understand details about data easily.We are able to predict which companies have phones, laptops and smart watches in different price Ranges. We Also found about discounts they offered and Amongst all brands, Apple have most expensive products.

(Screenshots are attached in next section)

1. Mention Assumptions.
2. Report any Bug if Exists but you could not fix.

Initially While Scrapping Discounts and Retails Rates we faced issues, some products do have any retail rates and discounts. We were not able to fetch that data, so we search over the internet, so we explicitly initialized them as null, if value exists it updates it else its already NULL, that’s how we handled this error.

1. Explanation of Dataset.

The Data we Scrapped is Brand Name, Specific Phone Model Name, Price, Market Price, Discount (in percentage), Link of Ad and Link of Image.

Brand Name Helped us to Combine all data and find mean and visualize according to brands.

Price, Market Price and Discount helped us to Visualize data easily as how much discount is offered, and how much cheaper we are getting a specific product from market.

A picture containing text, screenshot, font, number

Description automatically generated

Furthermore, we have gathered Data about watches and laptops as well, we used these same Columns for them as well. A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, font, number

Description automatically generated

1. Summary of data you gathered using Visualization.
   1. Average Price of Each Range of Each company.

A screen shot of a graph

Description automatically generated with medium confidence

* 1. Box Plot of Prices of Each Brand

A screen shot of a graph

Description automatically generated with low confidence

* 1. Average Price Comparison of each Brands with Errors.

A screen shot of a graph

Description automatically generated with medium confidence

* 1. Brand Comparison: Average Discount

A screenshot of a graph

Description automatically generated with low confidence

* 1. Brand Comparison: Discount Distribution

A screenshot of a graph

Description automatically generated with medium confidence

* 1. Brand Comparison: Average Discount

A point plot can be used to show the average discount for each brand, along with confidence intervals to represent the variability.

A screen shot of a graph

Description automatically generated with medium confidence

1. Comments on your Learning Experience during the project.

* We learned how to use web scraping techniques to extract data from a specific website and how to understand it.
* Worked with BeautifulSoup.
* Data cleaning and preprocessing as it is most essential thing in Data Science.
* Handling missing data.
* Visualization Techniques.
* Problem-solving and resourcefulness, we have to do research on each and every error we encountered, and we Learned something new.