CP32 – TEAM C – CARS24 WEB SCRAPING MINI PROJECT

Web Scraping and Analysis of Used Hyundai Car Data (Mumbai)

Team Members – Vijayabhaskar V (Team Lead), Adit Jain (Co-Lead), Harshit Kumar (Co-Lead), Deep Dhar, Punit Ayare, Ravi kant kumar, Arepalli Chandra Sekhar, Abhishri Pathak, Arun Singh, Deepak Melkani. Suraj Vishwakarma

CONTENTS

INTRODUCTION

- This **Web Scraping project** focuses on extracting structured data from the **Cars24** website using python.
- The main goal is to collect the details such as **kilometers driven, year of manufacture, fuel type, transmission, and price** for cars listed in the **Mumbai** location.
- This project contains Data Cleaning, Analysis, Visualization, export etc.., The Code identifies HTML elements, handles pagination to scrap multiple pages. As an output, Extracted the CSV file containing the data scraped.

TECHNOLOGICAL STACK & ROBUSTNESS

- Python – Primary Language for Web Scraping ; Beautiful Soup, Selenium & WebDriver for Scraping multi-page content.
- Requests handles HTTP requests to fetches webpages efficiently, Pandas for data manipulation, cleaning etc.., Jupyter notebook for effective team collaboration.
- Implemented error handling to manage missing data and inconsistent HTML tags. Incorporated try-except blocks to prevent runtime interruptions during scraping.
- Ensured smooth pagination and data continuity for complete dataset extraction.

WEBSITE STRUCTURE & WEB SCRAPING PROCESS

- Overview of **Cars24** website layout, How to locate brand filters, car details, and pagination, HTML tags identified (** , ** etc..)
- WEB Scraping Process : Imported necessary libraries, used Robots.txt for compliance check, Basic HTTP connectivity check for the website and checking the status, creating a directory for project structure setup, Advanced HTTP handling with retries, Logging, robust Error Handling.
- Testing & Advanced HTTP function (Unit Test, Mocked Response), Data Extraction, Converting the list to Pandas data frame, Data Cleaning and Preprocessing, Data Analysis & Export, Data Visualization & summary.

DATA EXTRACTION

- Website Loading using Selenium : We used Selenium WebDriver to open the cars24 webpage and scroll slowly, allowing all dynamically loaded car listings.
- HTML Parsing with Beautiful Soup : After the page is fully loaded, Beautiful Soup parses the page's HTML source to locate the required elements(car, prices, fuel)
- Identifying HTML tags for the particular data to extract details such as car name, kilometers driven, year, fuel etc..
- Data Extraction & Variable Storage : Each car's details are fetched from the HTML tags and stored in python variables using .find(), .findall() methods.
- All Extracted values are grouped into a dictionary with keys then appended to a list, converting the data to Data Frame to create structured tabular data.

DATA EXPORT, CLEANING, ANALYSIS & VISUALIZATION

- The list of Dictionaries is converted into a Pandas Data Frame. Each Key in the Dictionary becomes a column in the data frame, the raw, cleaned & final dataset is exported to a directory created initially.
- Removed missing or duplicate records, cleaned textual data by removing extra symbols, converted datatypes, standardized column names for better readability and uniformity.
- Performed basic Statistics and numerical analysis, for the numerical columns, analyzed distribution of some columns.
- Used Matplotlib & Seaborn to create charts from the cleaned dataset. Bar Charts, scatter plots, Histogram, & Pie chats for performing small Visualization.

CONCLUSION

- This project successfully scraped detailed car information from Cars24.com, using selenium and Beautiful soup.
- All required attributes – kilometers driven, year, fuel type, transmission, and price were efficiently extracted and organized.
- The collected data was cleaned, structured and stored in a CSV file for further analysis.
- The project improved understandings of dynamic website scraping
- Data Analysis, Visualization provided meaningful insights and overall this project enhanced practical skills in web scraping, data management etc.