**Assignment-Data Scientist**

**Assignment: Web Scraping**

**Objective:**

Create a Python script to automate the process of scraping **all available transaction documents** for **Wakad** for the year **2023** from the given site. The process should include handling form submissions, scraping transaction data, handling pagination, and downloading HTML files of transaction documents.

**Steps to Complete the Assignment:**

1. **Understanding the Website Structure:**

Navigate to <https://freesearchigrservice.maharashtra.gov.in/>.

Select “Rest of Maharashtra” and analyze the form inputs (Year, Property No., Tehsil, Village etc.).

1. **Automate Form Filling:**

Use Selenium to fill in the form:

Year: 2023

District : Pune

Tahsil : Haveli

Village : Wakad

Iterate over Property Numbers.

Handle the “No Data Found” scenario gracefully.

Automate Filling the captcha

1. **Scrape Transaction Document:**

For each transaction:

Get Doc Num, SRO Code, and Year.

Identify the link for the index document in the last column.

1. **Download Index Files:**

Simulate clicking the last column’s link for each transaction to access the document page.

Save the HTML content of the document page locally using the format:

{DocNum}\_{SROCode}\_{Year}.html

Example: 12345\_567\_2023.html.

1. **Navigate Multiple Pages:**

Handle pagination for transactions (10 transactions per page).

Extract and process all transactions for a given property number.

1. **Organize and Store Files:**

Save all downloaded HTML files in a designated folder.

**Deliverables:**

1. Python script with comments explaining each step.
2. Folder containing the downloaded HTML files named as {DocNum}\_{SROCode}\_{Year}.html.

**Assessment Criteria:**

1. **Script Functionality:** The script should correctly automate form submissions and file downloads.
2. **File Organization:** Files should be named and saved in the required format.
3. **Error Handling:** Proper handling of “No Data Found” and other edge cases.
4. **Code Quality:** Code should be well-structured, commented, and easy to follow.