## Exercise 4: Building a Companies House Web Scraper using Python + Google Colab + Gemini AI

Objective

* In this exercise, you will create a Google Colab Notebook that:
* Accepts an uploaded Excel file containing a list of company names
* Automatically searches for each company on the Companies House Website
* Scrapes key information for each search result:
* Company Registration Number
* Registered Office Address
* Company Status
* Outputs the results into a structured table
* Saves the final dataset into a downloadable Excel file
* Uses Gemini AI (Code Assistant) within Google Colab to help generate working Python code
* Uses Chrome Developer Tools to locate the HTML element selectors needed for scraping

This exercise demonstrates how AI can support real-world data automation tasks such as compliance checks, due diligence, onboarding, or data enrichment.

Setup

1. Create/open a Google Colab notebook

* Go to https://colab.research.google.com/
* Log in with a Google account
* Start a New Notebook

2. Prepare an Excel spreadsheet

Create an Excel file with a single column (Column A):

|  |
| --- |
| St Albans Digital Media |
| John Smith Consulting |
| Acme Holdings Ltd |

Save as companies.xlsx

(You will upload this inside the Notebook.)

3. Use Gemini Code Assistance in Google Colab

Throughout this exercise, you will use Gemini with prompt-guided tasks (provided below).

# Prompt:

# Create Python code that lets me upload an Excel spreadsheet of company names.

# The spreadsheet will have company names in Column A.

# Read the file into a pandas DataFrame and display the list of names.

Before Prompt 2: Using Chrome Developer Tools to Inspect Companies House HTML

Before scraping, you must discover HTML selectors manually.

Instructions:

1. Go to a sample Companies House search URL such as:

https://find-and-update.company-information.service.gov.uk/search?q=St+Albans+digital+media

1. Open Chrome Developer Tools:
   * Right click → Inspect
   * Use the element selector tool (top-left of DevTools)
2. Identify important elements:

**Company Number:**

A black background with text

AI-generated content may be incorrect.

Selector examples:

#company-number strong

or p#company-number strong

**Registered Office Address**

A black background with white text

AI-generated content may be incorrect.

**Company Status**

A black background with green and yellow text

AI-generated content may be incorrect.

Prompt 2: Search Companies House & Scrape Data

Prompt:

# Write Python code that searches the Companies House website for each company name

# by constructing a search URL like:

# https://find-and-update.company-information.service.gov.uk/search?q=St+Albans+digital+media

#

# Parse the HTML using BeautifulSoup and extract:

# - Company Registration Number from: <p id="company-number"><strong>...</strong></p>

# - Registered Office Address from: <dt>Registered office address</dt> + <dd>

# - Company Status from the element: <dd id="company-status">...</dd>

#

# Store results in a pandas DataFrame.

Prompt 3: Display Results as DataFrame Grid

Prompt:

# Take the scraped company data and display it as a table/grid using pandas.

# The table should include:

# - Company Name

# - Registration Number

# - Registered Office Address

# - Company Status

Prompt 4: Save to Downloadable Excel File

Prompt:

# Convert the company data DataFrame into an Excel spreadsheet.

# Save it as companies\_report.xlsx and create a download link in the notebook.

Learning Objectives

By completing this exercise, you have learnt:

**Technical Skills**

1. How to upload files into Google Colab
2. Using AI (Gemini) to help break down coding tasks
3. How to inspect HTML structure using Chrome Developer Tools
4. How to scrape a website using requests and BeautifulSoup
5. How to identify CSS selectors from live HTML
6. Building structured DataFrames with pandas
7. Cleaning, formatting, and processing text-based data
8. Exporting Python data into Excel format
9. Understanding rate-limited scraping and using respectful request patterns
10. Sharing and publishing a completed notebook