Data and Artificial Intelligence
Cyber Shujaa Program

Week 1 Assignment
Web Scraping and Data Handling in Python

Student Name: Rodney Roy Gitonga

Student ID: CS-DA03-26025

Introduction

This week's assignment focused on extracting structured data from a live website using web scraping techniques. The specific objective was to automate the collection of hockey team performance data from scrapethissite.com.

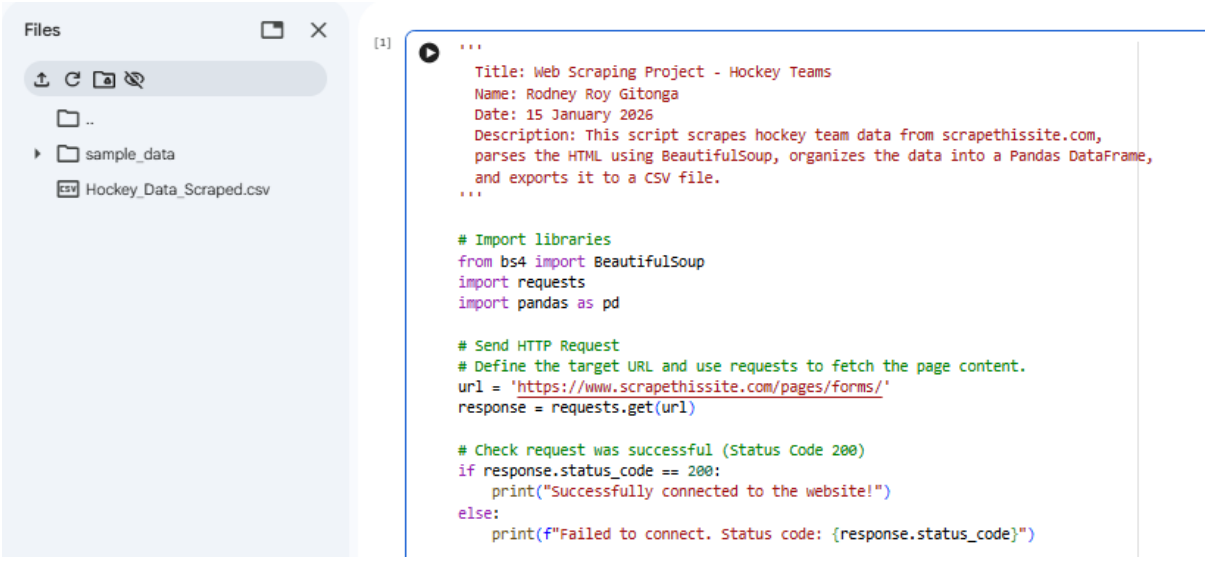
The objectives of the assignment were:

1. Perform practical Python coding on Jupyter Notebooks hosted on Google Colab.
2. Use the requests and beautifulSoup libraries to extract data from the web page
3. Parse and clean the extracted data for analysis
4. Store the structured data into a Pandas DataFrame.
5. Export the final dataset to a .csv file.

Tasks Completed

Step 1: Environment Setup and HTTP Request

Began by importing the libraries, requests for handling the connection, beautifulsoup for parsing HTML and pandas for data manipulation. I then defined the target URL and sent a GET request to the server.



Step 2: Parsing HTML Content

Once connection was established, I used beautifulsoup to parse the raw HTML text. This allowed me to navigate the structure of the page programmatically.

```
# Parse HTML Content
# Initialize BeautifulSoup to parse the text of the response
soup = BeautifulSoup(response.text, 'html.parser')
```

Step 3: Data Extraction and Cleaning

I identified the table headers to create column names then looped through the table rows(tr) extracting the data cells(td) stripping away whitespace and appending the clean data to a list ensuring operation efficiency.

```
# Extract Column Headers
# Find all table header cells ('th') and strip whitespace
header_tags = hockey_table.find_all('th')
columns = [header.text.strip() for header in header_tags]
print(f"Columns found: {columns}")

# Extract Row Data
# Find all table rows ('tr'). Skip the first row [1:] because it contains headers.
rows = hockey_table.find_all('tr')
extracted_data = []

for row in rows[1:]:
    # Find all data cells ('td') in the current row
    cells = row.find_all('td')
    # Clean the text for each cell
    row_data = [cell.text.strip() for cell in cells]
    # Append the clean row to our list
    extracted_data.append(row_data)
```

Step 4: Storing Data in a DataFrame

I converted the list of extracted data into Pandas DataFrame organising the data into a clear row and column format.

```
# Create DataFrame
# Convert the list of lists into a Pandas DataFrame
df = pd.DataFrame(extracted_data, columns=columns)
```

Step 5: Exporting to CSV

I finally exported the DataFrame to a CSV file named Hockey_Data_Scraped.csv to complete data handling cycle.

```
# Export to CSV
# Save the file without the pandas index column
csv_filename = 'Hockey_Data_Scraped.csv'
df.to_csv(csv_filename, index=False)
print(f"\nData successfully saved to {csv_filename}")
```

Link to Code:

https://colab.research.google.com/drive/1L6UXoipO2Ac0vGjnx_r9ylQip6E54qf5?usp=sharing

Conclusion

This week I gained a good grounding on the introductory concepts relating to data science and automated data gathering. By successfully scraping the hockey data, I demonstrated the ability to turn unstructured web content into a usable dataset. I am getting a better understanding of Python tools that I can build on as we work on more advanced concepts in later weeks. I look forward to adding this project to my portfolio as I continue to develop my skills in Data and AI.