Data and Artificial Intelligence
Cyber Shujaa Program

Week 1 Assignment
Web Scraping and Data Handling in Python

Student Name: Faith Jeptoo

Student ID: CS-DA02-25005

Table of Content

Introduction

This week's assignment was to start extracting data using web scraping. I was totally new to the tools we were introduced to. I had never written Python code before and had not created a Colab account. However, it was a very effective way of learning by seeing actual results.

The objectives of the assignment were:

1. Practical Python coding on Jupiter Notebooks hosted on Google Colab
2. Use requests and BeautifulSoup to extract data from a web page.
3. Parse and clean the extracted data.
4. Store structured data into a Pandas DataFrame.
5. Export the final dataset to a .csv file.

Tasks Completed

Step 1: Import Libraries

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
from google.colab import files
```

Step 2: Load the webpage

```
url = "https://www.scrapethissite.com/pages/forms/"
r = requests.get(url)
soup = BeautifulSoup(r.content, 'html.parser')
```

Step 3: Find the hockey score table

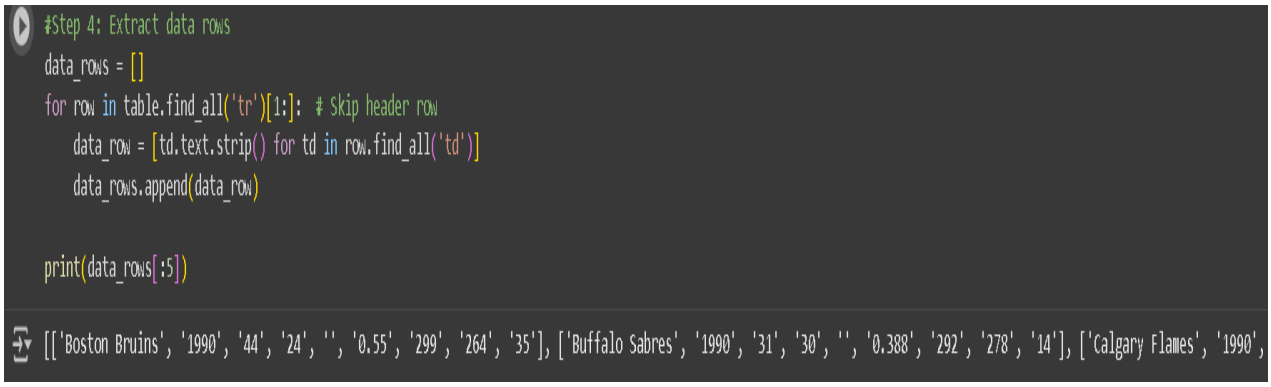
```
table = soup.find('table', class_='table')
```

Step 4: Get Column names

```
column_names = [th.text.strip() for th in table.find_all('th')]
```

Step 5: Extract data rows

```
data_rows = []
for row in table.find_all('tr')[1:]:    # Skip header row
   data_row = [td.text.strip() for td in row.find_all('td')]
   data_rows.append(data_row)
print(data_rows[:5]) # Preview first 5 rows
```



Step 6: Create an empty DataFrame

```
df = pd.DataFrame(columns=column_names)
print(df.head())
```

```
print(df.head())
```

```
            Team Name  Year Wins Losses OT Losses  Win % Goals For (GF)  \
0       Boston Bruins  1990   44     24           0.55            299
1      Buffalo Sabres  1990   31     30          0.388            292
2      Calgary Flames  1990   46     26          0.575            344
3  Chicago Blackhawks  1990   49     23          0.613            284
4   Detroit Red Wings  1990   34     38          0.425            273

  Goals Against (GA) + / -
0                 264    35
1                 278    14
2                 263    81
3                 211    73
4                 298   -25
```

Step 7: Fill the DataFrame with data

```
df = pd.DataFrame(data_rows, columns=column_names)
display(df.head())
```

```
#Step 6: Fill the DataFrame with data
df = pd.DataFrame(data_rows, columns=column_names)
display(df.head())
```

	Team Name	Year	Wins	Losses	OT Losses	Win %	Goals For (GF)	Goals Against (GA)	+ / -
0	Boston Bruins	1990	44	24		0.55	299	264	35
1	Buffalo Sabres	1990	31	30		0.388	292	278	14
2	Calgary Flames	1990	46	26		0.575	344	263	81
3	Chicago Blackhawks	1990	49	23		0.613	284	211	73
4	Detroit Red Wings	1990	34	38		0.425	273	298	-25

Step 8: Export to CSV

```
df.to_csv("Hockey.csv", index=False)
files.download("Hockey.csv") # Download the CSV file
```

Link to Code:

https://colab.research.google.com/drive/1LlNVwxj0FfRVTQG7wyKn1aQsd8ENoVuv?usp=sharing

Conclusion

This week I gained a good grounding on the introductory concepts relating to data science and artificial intelligence. I am getting a better understanding that I can build on as we work on more advanced concepts in later weeks. I have posted my writeup on my blog and I look forward to building a portfolio that I can showcase on my CV as I look for jobs in Data and AI.