

# Data and Artificial Intelligence

## Cyber Shujaa Program

### Week 1 Assignment

### Web Scraping and Data Handling in Python

**Student Name:** Jeff Walter Maende

**Student ID:** CS-DA02-25001

#### Introduction

This week's assignment was to start extracting data using web scraping. I was new to the tools we were introduced to. I used Python's powerful libraries to extract data from the website using Jupyter Notebooks hosted on Google Colab. It was a very effective way of learning by seeing the 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

I have outlined the step-by-step sequence of tasks completed with explanations of code and code snippets backing my steps with screenshots as attached below:

Inbox - jeffmaende@gmail.com | DA2-2025: Week 1: Introduction | Web\_Scraping.ipynb - Colab

**Web\_Scraping.ipynb**

# Title: Web Scraping Project  
# Name: Jeff Walter Maende  
# Date: 21 September 2025  
# Description: Web Scraping using BeautifulSoup, requests and pandas python libraries

#Importing libraries needed for webscraping  
from bs4 import BeautifulSoup  
import requests  
import pandas as pd

from google.colab import drive  
drive.mount('/content/drive')

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force\_remount=True).

#set URL of website in a variable  
url = 'https://www.scrapethissite.com/pages/forms/'  
page = requests.get(url)

#Use BeautifulSoup to extract the HTML content  
soup = BeautifulSoup(page.text,'html')  
print(soup)

</tr>  
<tr class="team">  
<td class="name">

Boston Bruins  
1990  
44  
24  
0.55  
299  
264  
35

Buffalo Sabres

Variables Terminal Python 3

22°C Mostly cloudy Search ENG US 13:38 21/09/2025

Inbox - jeffmaende@gmail.com | DA2-2025: Week 1: Introduction | Web\_Scraping.ipynb - Colab

**Web\_Scraping.ipynb**

</tr>  
<tr class="team">  
<td class="name">

Boston Bruins  
1990  
44  
24  
0.55  
299  
264  
35

Buffalo Sabres

Variables Terminal Python 3

22°C Mostly cloudy Search ENG US 13:38 21/09/2025

File Edit View Insert Runtime Tools Help

Commands + Code + Text | Run all

```

<td class="losses"> 30
</td>
<td class="ot-losses">
</td>
<td class="pct text-danger"> 0.388
</td>
<td class="gf"> 292
</td>
<td class="ga"> 278
</td>
<td class="diff text-success"> 14
</td>
</tr>
<tr class="team">
<td class="name"> Calgary Flames
</td>
<td class="year"> 1990
</td>
<td class="wins"> 46
</td>
<td class="losses"> 26
</td>
<td class="ot-losses">
</td>

```

Variables Terminal Python 3

22°C Mostly cloudy Search ENG US 13:38 21/09/2025

File Edit View Insert Runtime Tools Help

Commands + Code + Text | Run all

```

<td class="gf"> 257
</td>
<td class="ga"> 236
</td>
<td class="diff text-success"> 21
</td>
</tr>
</table>
<div class="row pagination-area">
<div class="col-md-10 text-center">
<ul class="pagination">
<li>
<a href="/pages/forms/?page_num=1">
1
</a>
</li>
<li>
<a href="/pages/forms/?page_num=2">
2
</a>
</li>
<li>
<a href="/pages/forms/?page_num=3">
3
</a>

```

Variables Terminal Python 3

22°C Mostly cloudy Search ENG US 13:38 21/09/2025

Inbox - jeffmaende@gmail.com | DA2-2025: Week 1: Introduction | Web\_Scraping.ipynb - Colab

**Web\_Scraping.ipynb**  

File Edit View Insert Runtime Tools Help

Commands + Code + Text ▶ Run all ▶

```
[59] #Extract the table with the Hockey Scores
hockey_table = soup.find('table', class_='table')
#print(hockey_table)

[60] #Extract the column headings
table_titles = hockey_table.find_all('th')
hockey_table_title = [title.text.strip() for title in table_titles]
print(hockey_table_title)

[61] ['team Name', 'Year', 'Wins', 'Losses', 'OT Losses', 'Win %', 'Goals For (GF)', 'Goals Against (GA)', '+ / -']

[62] #Save the column headings onto a Pandas DataFrame
df = pd.DataFrame(columns=hockey_table_title)
df

[63] Team Name Year Wins Losses OT Losses Win % Goals For (GF) Goals Against (GA) + / -
```

#Extract the data row by row. First get all rows, then loop through each while stripping and saving data into the DataFrame

```
table_data = hockey_table.find_all('tr')
for row in table_data[1:]:
    raw_data = row.find_all('td')
    each_raw_data = [data.text.strip() for data in raw_data]
    print(each_raw_data)
```

```
['Boston Bruins', '1990', '44', '24', '', '0.55', '299', '264', '35']
['Buffalo Sabres', '1990', '31', '30', '', '0.388', '292', '278', '14']
['Calgary Flames', '1990', '46', '26', '', '0.575', '344', '263', '81']
['Chicago Blackhawks', '1990', '49', '23', '', '0.613', '284', '211', '73']
['Detroit Red Wings', '1990', '34', '38', '', '0.425', '273', '298', '-25']
['Edmonton Oilers', '1990', '37', '37', '', '0.463', '272', '272', '0']
[' Hartford Whalers', '1990', '31', '38', '', '0.388', '238', '276', '38']
['Los Angeles Kings', '1990', '46', '24', '', '0.575', '340', '254', '86']
['Minnesota North Stars', '1990', '27', '39', '', '0.338', '256', '266', '-10']
['Montreal Canadiens', '1990', '39', '30', '', '0.409', '273', '249', '24']
['New Jersey Devils', '1990', '32', '33', '', '0.4', '272', '264', '8']
['New York Islanders', '1990', '25', '45', '', '0.312', '223', '290', '67']
['New York Rangers', '1990', '36', '31', '', '0.45', '297', '265', '32']
['Philadelphia Flyers', '1990', '33', '37', '', '0.412', '252', '267', '-15']
['Pittsburgh Penguins', '1990', '41', '33', '', '0.512', '342', '305', '37]
['Quebec Nordiques', '1990', '16', '50', '', '0.2', '236', '354', '-118']
['St. Louis Blues', '1990', '47', '22', '', '0.588', '310', '250', '60']
['Toronto Maple Leafs', '1990', '23', '46', '', '0.287', '241', '318', '-72']
['Vancouver Canucks', '1990', '28', '43', '', '0.35', '243', '315', '-72']
['Washington Capitals', '1990', '37', '36', '', '0.463', '258', '258', '9']
['Winnipeg Jets', '1990', '26', '43', '', '0.325', '260', '288', '-28']
['Boston Bruins', '1991', '36', '32', '', '0.45', '270', '275', '-5']
['Buffalo Sabres', '1991', '31', '37', '', '0.388', '289', '299', '-10']
['Calgary Flames', '1991', '31', '37', '', '0.388', '296', '305', '-9']
['Chicago Blackhawks', '1991', '36', '29', '', '0.45', '257', '236', '21']
```

Variables Terminal Python 3

22°C Mostly cloudy

ENG US 13:38 21/09/2025

Inbox - jeffmaende@gmail.com | DA2-2025: Week 1: Introduction | Web\_Scraping.ipynb - Colab

**Web\_Scraping.ipynb**  

File Edit View Insert Runtime Tools Help

Commands + Code + Text ▶ Run all ▶

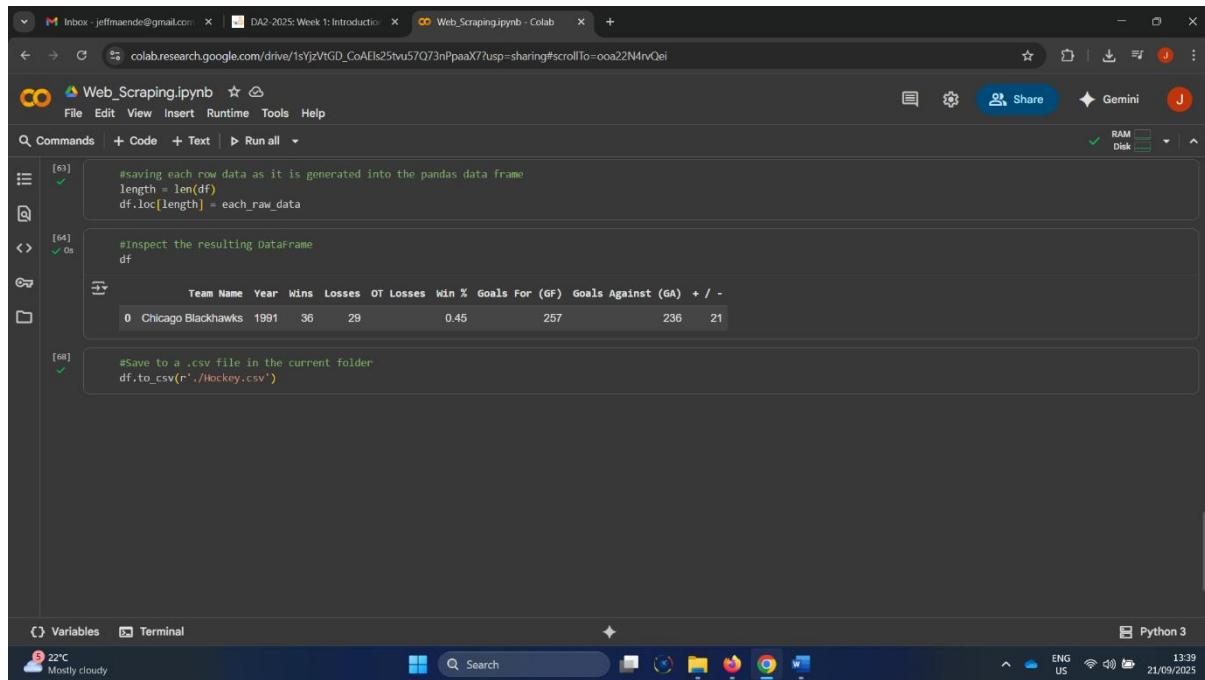
```
[62] #Extract the data row by row. First get all rows, then loop through each while stripping and saving data into the DataFrame
table_data = hockey_table.find_all('tr')
for row in table_data[1:]:
    raw_data = row.find_all('td')
    each_raw_data = [data.text.strip() for data in raw_data]
    print(each_raw_data)
```

```
['Boston Bruins', '1990', '44', '24', '', '0.55', '299', '264', '35']
['Buffalo Sabres', '1990', '31', '30', '', '0.388', '292', '278', '14']
['Calgary Flames', '1990', '46', '26', '', '0.575', '344', '263', '81']
['Chicago Blackhawks', '1990', '49', '23', '', '0.613', '284', '211', '73']
['Detroit Red Wings', '1990', '34', '38', '', '0.425', '273', '298', '-25']
['Edmonton Oilers', '1990', '37', '37', '', '0.463', '272', '272', '0']
[' Hartford Whalers', '1990', '31', '38', '', '0.388', '238', '276', '38']
['Los Angeles Kings', '1990', '46', '24', '', '0.575', '340', '254', '86']
['Minnesota North Stars', '1990', '27', '39', '', '0.338', '256', '266', '-10']
['Montreal Canadiens', '1990', '39', '30', '', '0.409', '273', '249', '24']
['New Jersey Devils', '1990', '32', '33', '', '0.4', '272', '264', '8']
['New York Islanders', '1990', '25', '45', '', '0.312', '223', '290', '67']
['New York Rangers', '1990', '36', '31', '', '0.45', '297', '265', '32']
['Philadelphia Flyers', '1990', '33', '37', '', '0.412', '252', '267', '-15']
['Pittsburgh Penguins', '1990', '41', '33', '', '0.512', '342', '305', '37]
['Quebec Nordiques', '1990', '16', '50', '', '0.2', '236', '354', '-118']
['St. Louis Blues', '1990', '47', '22', '', '0.588', '310', '250', '60']
['Toronto Maple Leafs', '1990', '23', '46', '', '0.287', '241', '318', '-72']
['Vancouver Canucks', '1990', '28', '43', '', '0.35', '243', '315', '-72']
['Washington Capitals', '1990', '37', '36', '', '0.463', '258', '258', '9']
['Winnipeg Jets', '1990', '26', '43', '', '0.325', '260', '288', '-28']
['Boston Bruins', '1991', '36', '32', '', '0.45', '270', '275', '-5']
['Buffalo Sabres', '1991', '31', '37', '', '0.388', '289', '299', '-10']
['Calgary Flames', '1991', '31', '37', '', '0.388', '296', '305', '-9']
['Chicago Blackhawks', '1991', '36', '29', '', '0.45', '257', '236', '21']
```

Variables Terminal Python 3

22°C Mostly cloudy

ENG US 13:38 21/09/2025



```
[63]: #saving each row data as it is generated into the pandas data frame
length = len(df)
df.loc[length] = each_raw_data

[64]: #Inspect the resulting DataFrame
df

[65]: Team Name Year Wins Losses OT Losses Win % Goals For (GF) Goals Against (GA) + / -
0 Chicago Blackhawks 1991 36 29 0.46 257 236 21

[66]: #Save to a .csv file in the current folder
df.to_csv(r'./Hockey.csv')
```

The screenshot shows a Google Colab notebook titled "Web\_Scraping.ipynb". The code cell [63] contains a snippet for saving raw data into a pandas DataFrame. Cell [64] displays the resulting DataFrame, which includes columns for Team Name, Year, Wins, Losses, OT Losses, Win %, Goals For (GF), Goals Against (GA), and a +/- column. A single row for the Chicago Blackhawks from 1991 is shown. Cell [66] demonstrates how to save the DataFrame to a CSV file named "Hockey.csv". The bottom status bar indicates the environment is Python 3.

I have shared my Notebook and can be accessed using the below link:

#### **Link to Code:**

[https://colab.research.google.com/drive/1sYjzVtGD\\_CoAEls25tvu57Q73nPpaaX7?usp=sharing](https://colab.research.google.com/drive/1sYjzVtGD_CoAEls25tvu57Q73nPpaaX7?usp=sharing)

#### **Conclusion**

This week I gained a solid grounding in the introductory concepts of data science and artificial intelligence, with a focus on their application in data analytics. I am developing a clearer understanding of how these tools can enhance data-driven audit procedures and risk assessment, which I will continue to build on as we explore more advanced concepts in the coming weeks.