# Data and Artificial Intelligence

# Cyber Shujaa Program

# Week 1 Assignment Web Scraping and Data Handling in Python

**Student Name:DOROTHY AWINO**

**Student ID:** **CS-DA01-25075**

**NOTEBOOK LINK:** [**https://colab.research.google.com/drive/1GHt7ZYpL1ZqVXf2VUJYusKq3kGjfr0aH?usp=sharing**](https://colab.research.google.com/drive/1GHt7ZYpL1ZqVXf2VUJYusKq3kGjfr0aH?usp=sharing)

## 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

I Outlined a step-by-step sequence of tasks as they are completed with explanations of code and code snippets backed up by screenshots showing my work.

I Ensured to follow good coding practices by using appropriate names for variables, using comments and white space for code readability.

I finally Shared my Notebook by clicking the Share button on the top-right side of the page and Ensured I allow ‘Anyone with the link’ to be a Viewer

**Link to Code:**

## 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.

### Overview

This week’s assignment will develop hands-on experience automating web data gathering using Python and on Google Colab.

The task involves scraping structured data from a live website.

The link to the website: <https://www.scrapethissite.com/pages/forms/>

Three libraries will be used: **Requests**for handling HTTP requests, **BeautifulSoup**for parsing HTML, and **pandas**for storing and manipulating data.

Data will be collected and organized into a DataFrame, and export results to a .csv file.

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.

### Sample Code

'''

  Title: Web Scraping Project

  Name: DOROTHY AWINO ONGONGA

  Date: 8 JUNE 2025

  few comments about the project

'''

#Importing libraries that I will need for webscraping

from bs4 import BeautifulSoup

import requests

import pandas as pd

#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)

#Extract the table with the Hockey Scores

hockey\_table = soup.find('table', class\_='table')

#print(hockey\_table)

#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)

#Save the column headings onto a Pandas DataFrame

df = pd.DataFrame(columns=hockey\_table\_title)

df

#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)

  #saving each row data as it is generated into the pandas data frame

  length = len(df)

  df.loc[length] = each\_raw\_data

#Inspect the resulting DataFrame

df

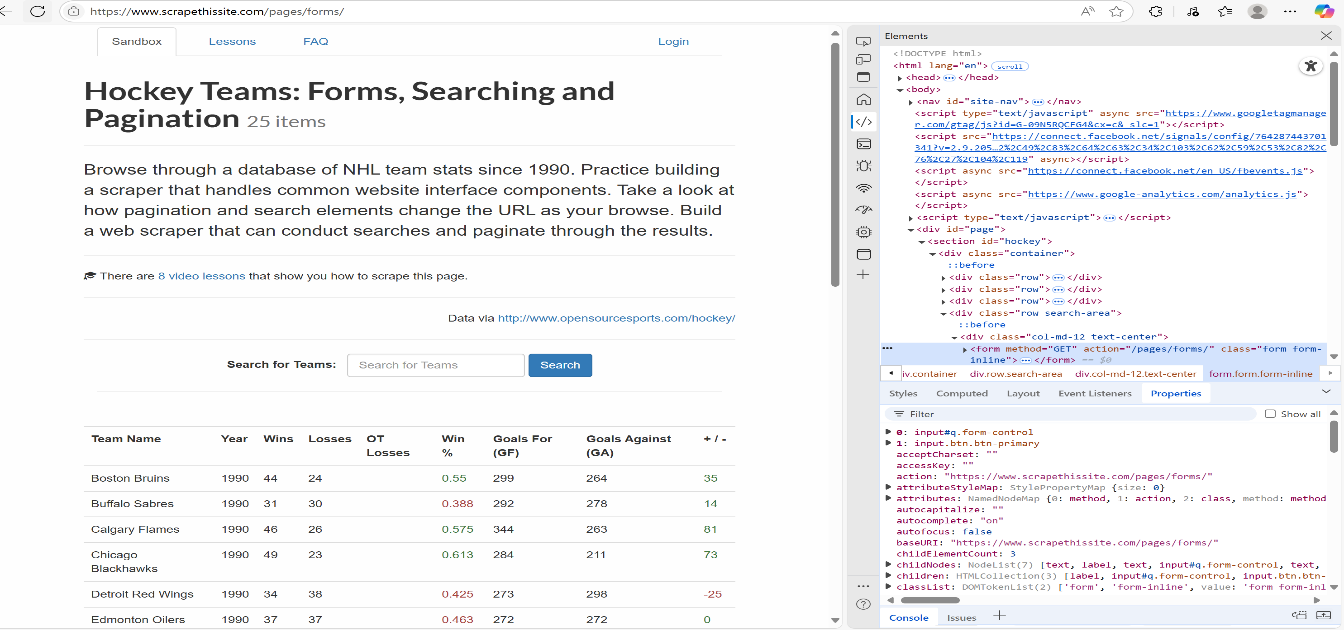
#Save to a .csv file in the current folder

df.to\_csv(r'./Hockey.csv')

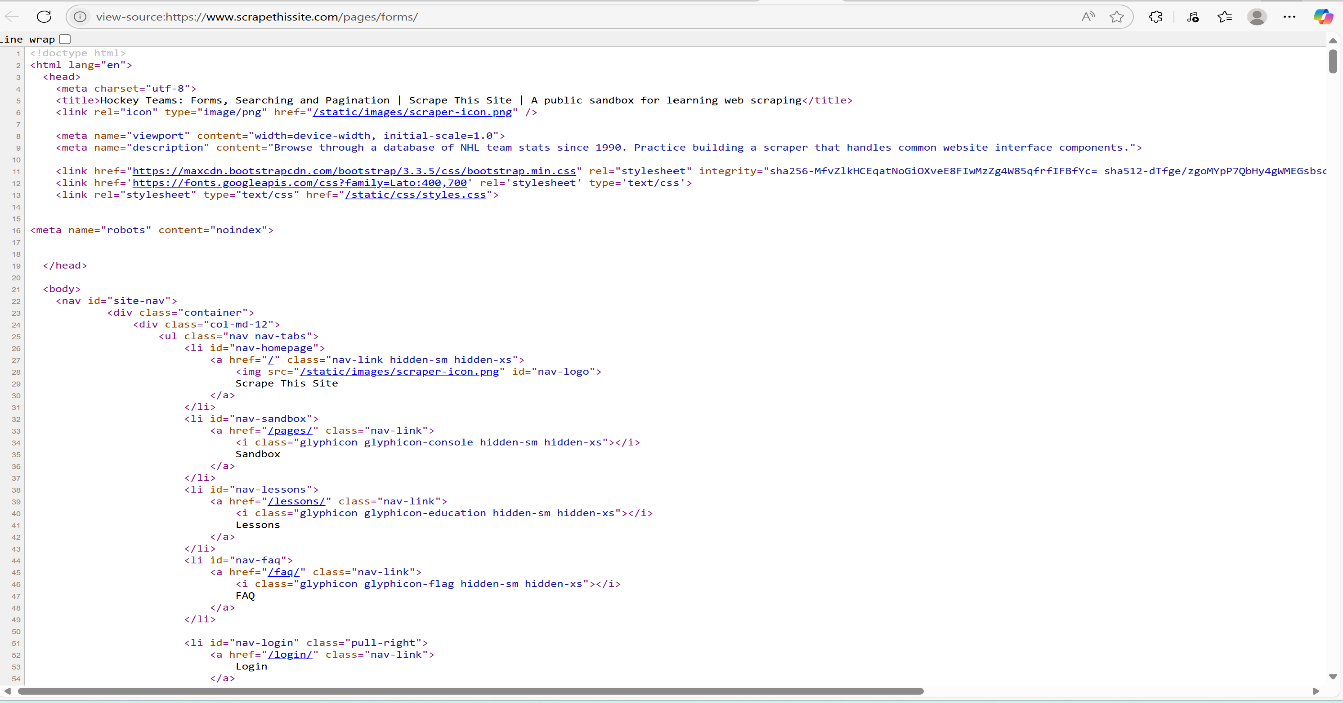
WEB SCRAPING

The Web Scraping Project URL is <https://www.scrapethissite.com/pages/forms/>, it is an open source site

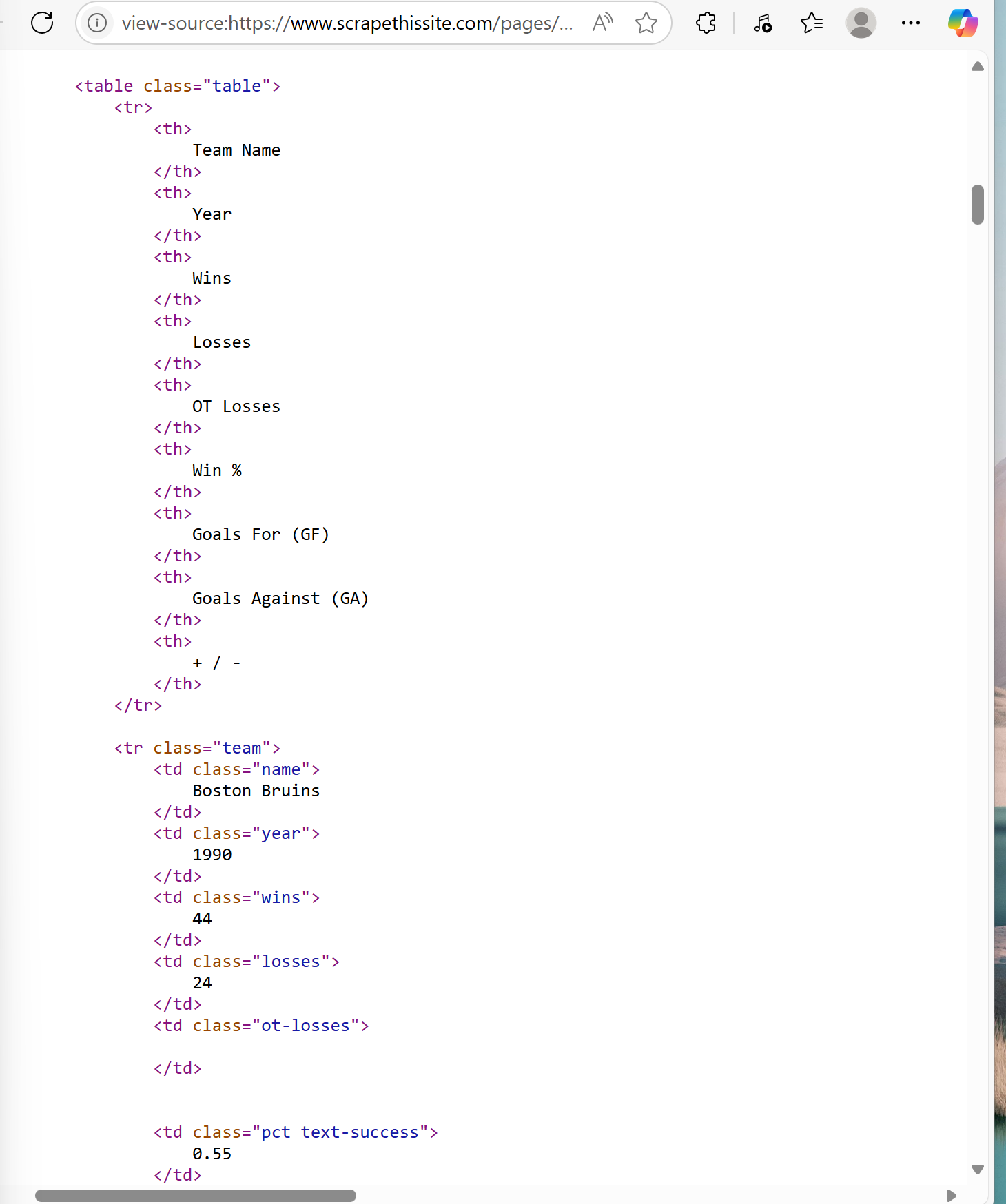
I have done a page inspection



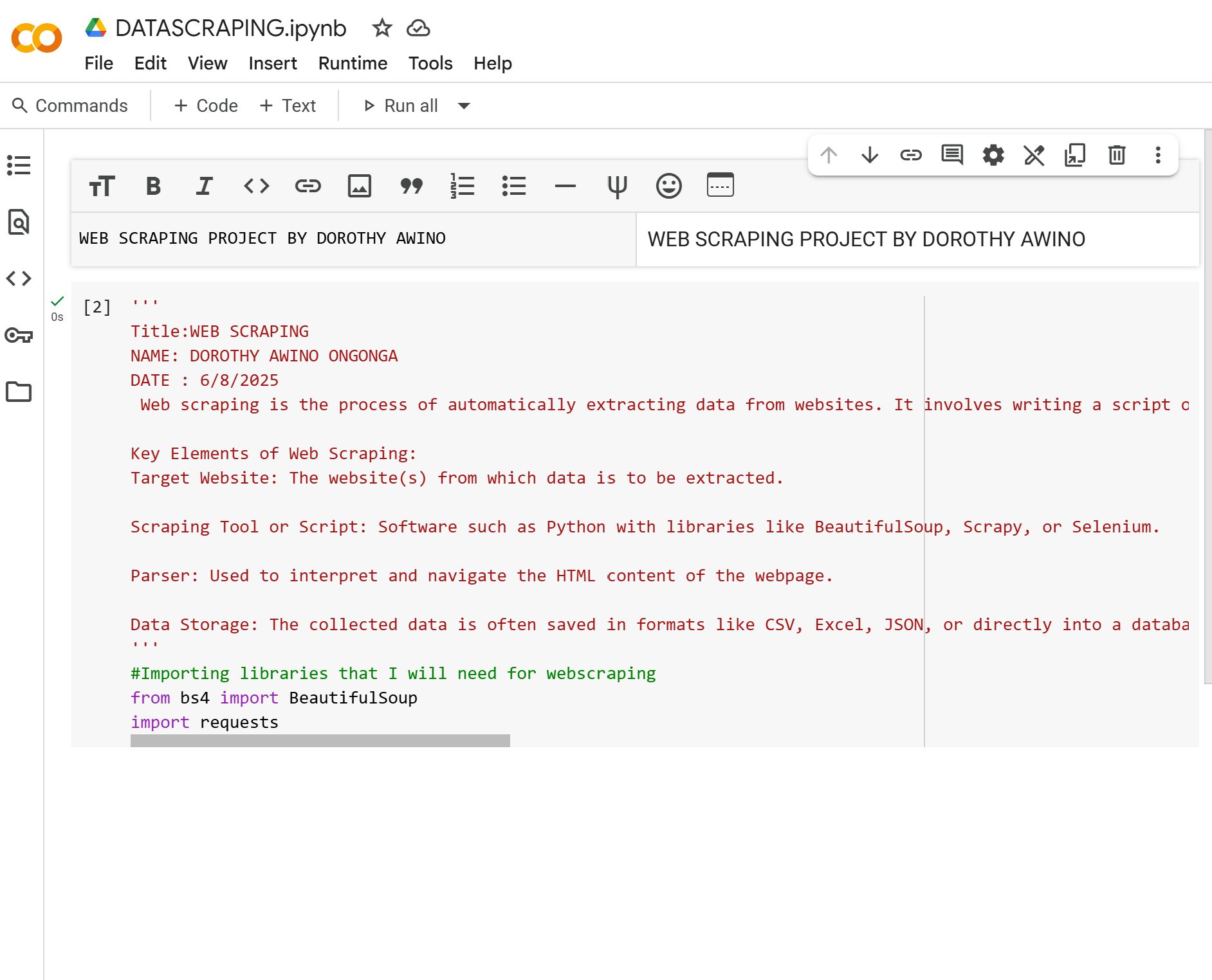
Below is the Page source with the Structure containing The Head,Body and other contents in HTML Language



For this Project,I specifically focused on the TABLE contents

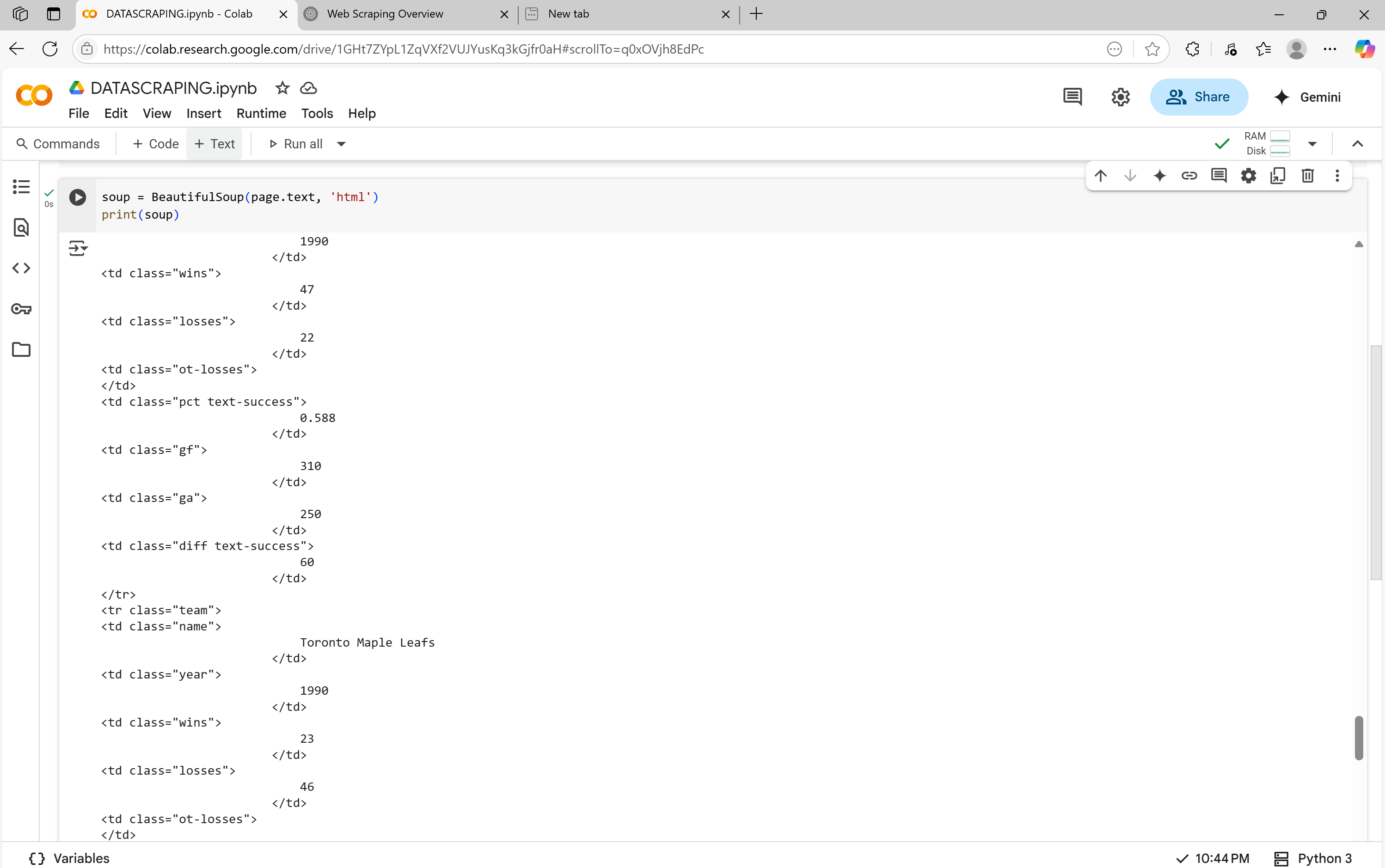


Using Google Colab,I created a Notebook which is made up of two types of cells: The Code cell and Text Cell and Named it DATASCRAPING.ipynb.Additionally,I imported Python LIBRARIES i.e Pandas as PD,BeautifulSoup from bs4 ,NUMPY and requests. I gave a short decription about webscraping as a pivot to the project using a DOC STRING and added a Text independent of the code.



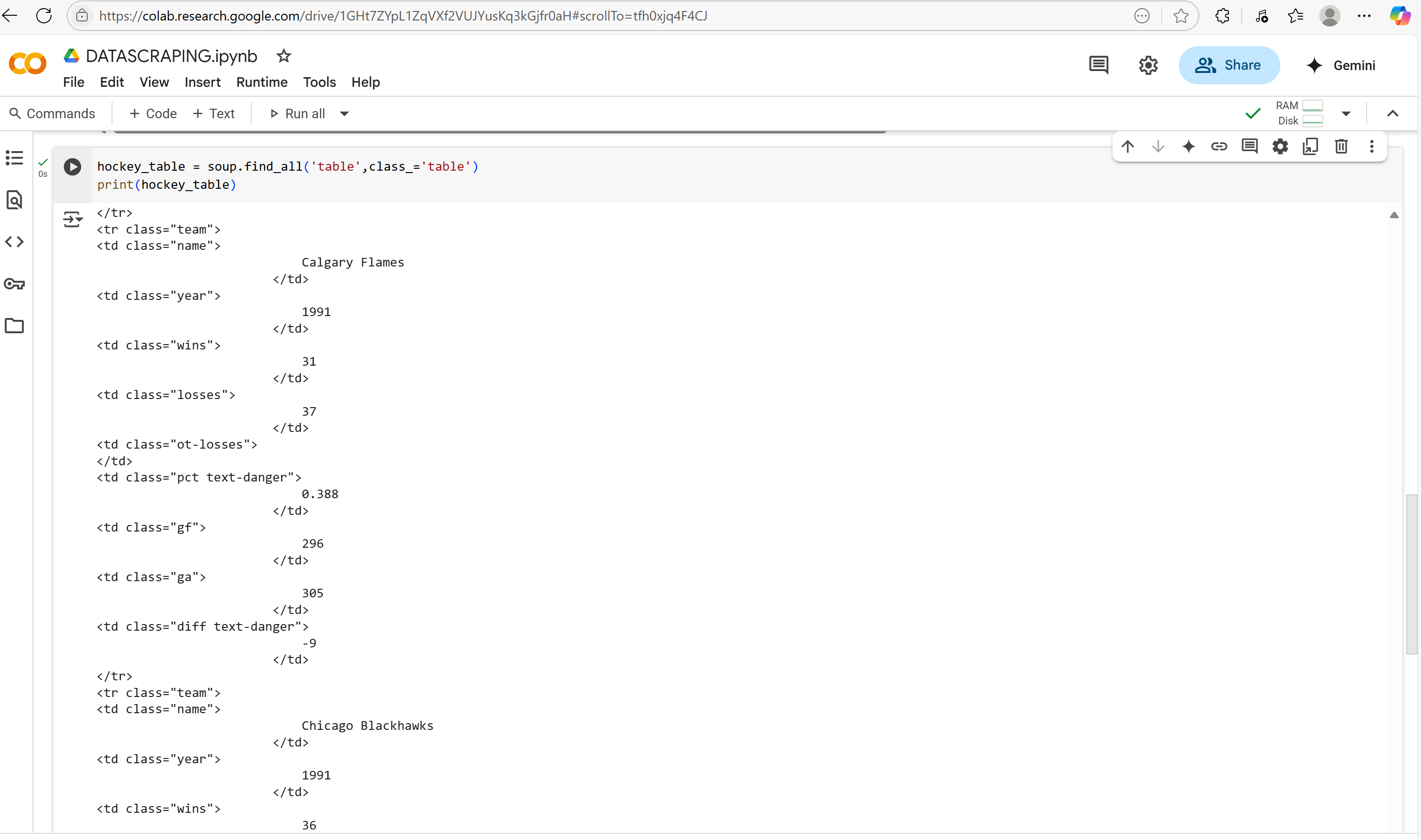
I Imported the website’s url and it was automatically saved as a string.I set is a variable and defined it as a code. 

The Library from Python called BeautifulSoup is now able to access all contents in html language from the URL I defined and retrieve specific tags in html.



BeautifulSoup is a Library that is able to grant me the ability to extract the page source that I wanted to scrape.

I Extracted Specifically The Table from the ‘SOUP’



In this table that has been extracted,it is made up of rows which are the tr and columns which are the td.

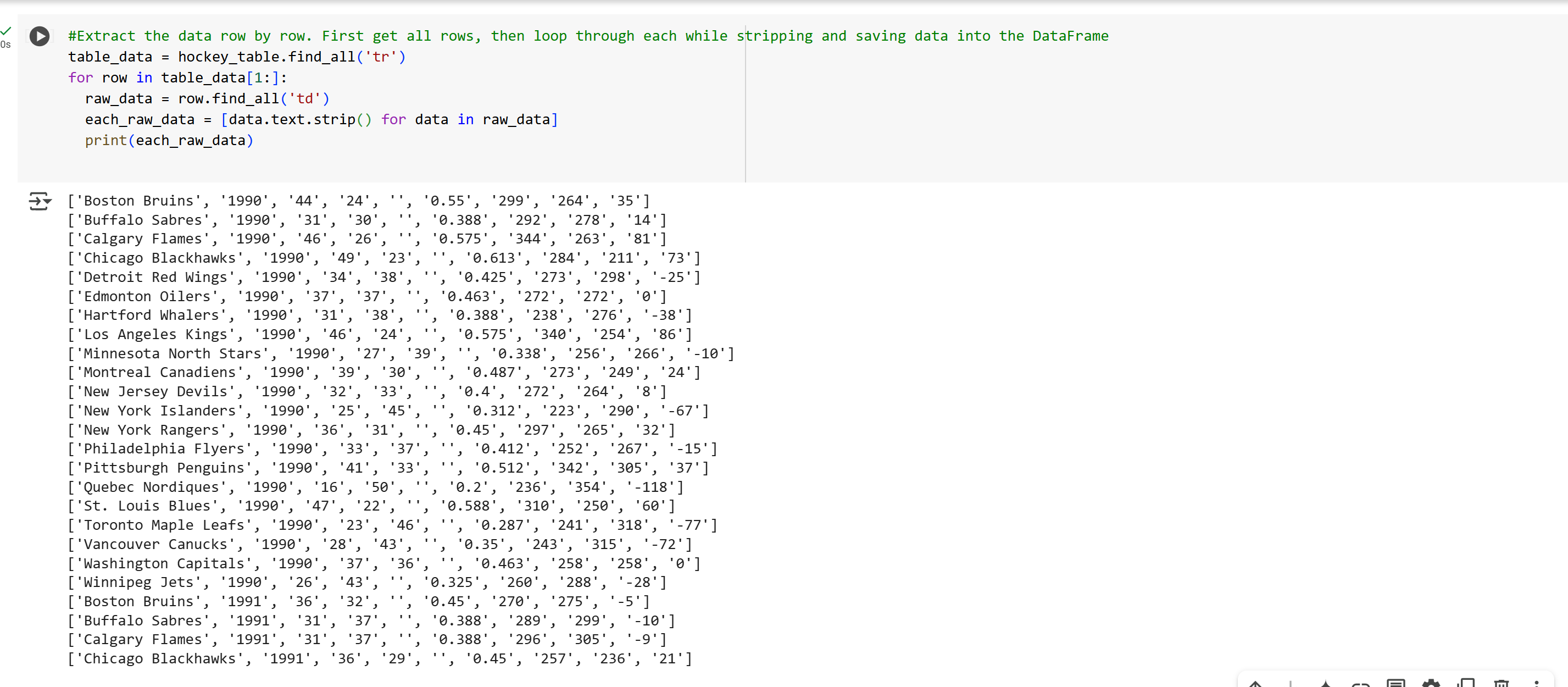
Now I want to work on One Row at a time i.e Extract the Column Headings



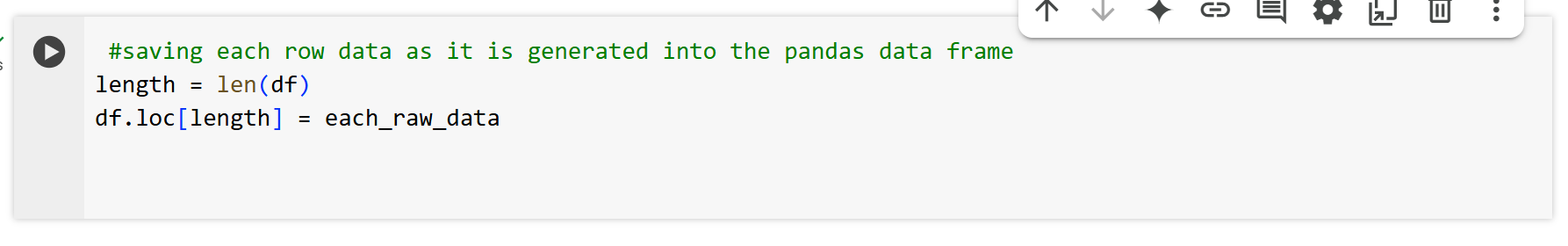
Save the column headings onto a Pandas DataFrame



I went on to Extract the Raw data Row by Row using Code.



Saved Each Row Data as it is generated into the Pandas Data Frame



I finally inspected the resulting DataFrame and saved it to a CSV file

