Data and Artificial Intelligence

Cyber Shujaa Program

Week 1 Assignment
Web Scraping and Data Handling in Python

NAME: PAULINE KUNGU

PROJECT: WEB SCEAPING PROJECT

DATE: May 15, 2025

INTRODUCTION

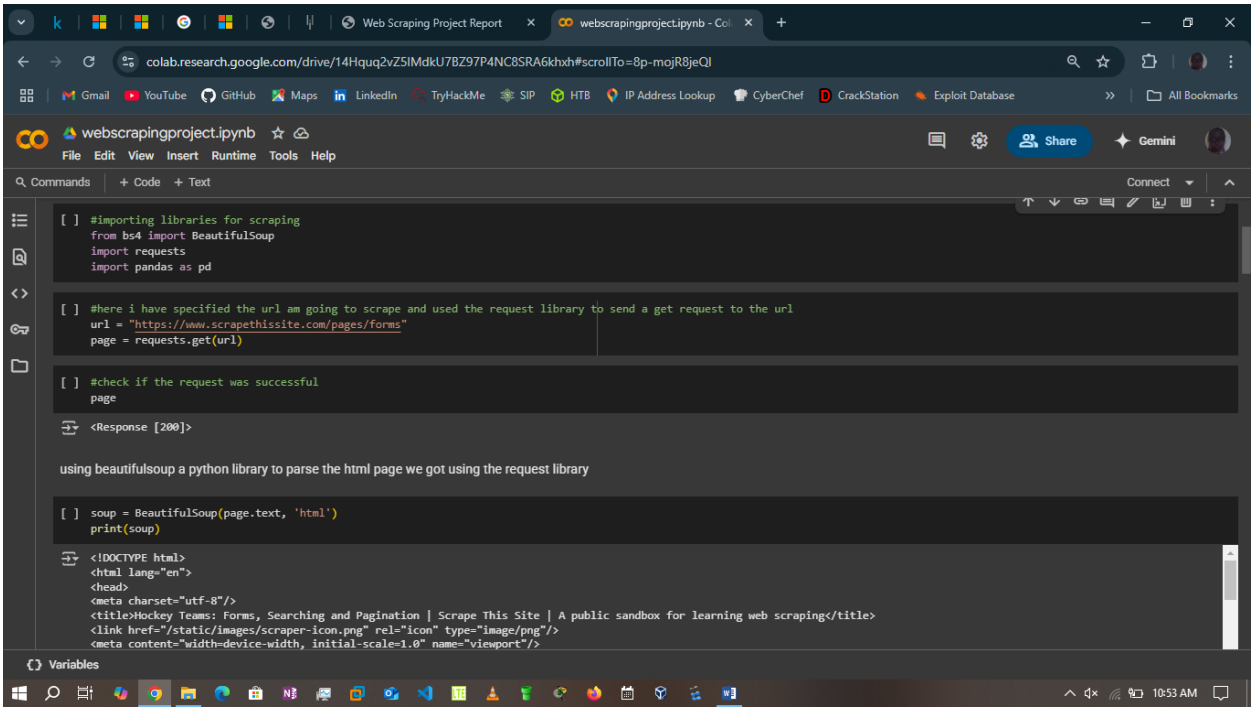
Data is important for machine learning algorithms to learn patterns, make predictions and improve their performance. Sufficient and high quality data is needed. Web scraping is the automated process of extracting data from websites.

What am going to cover:

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

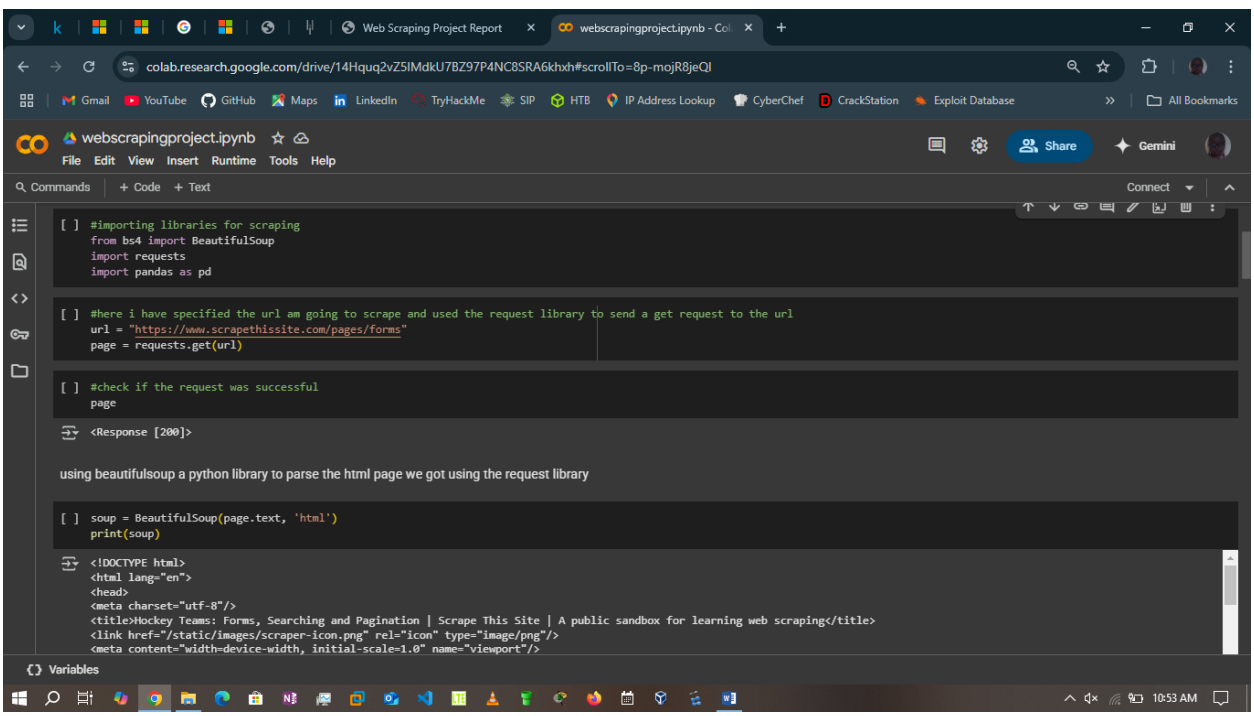
Step 1: Importing the libraries to use for web scraping

The first step is to import the necessary libraries am going to use. BeautifulSoup provides HTML parsing capabilities, the request library handles HTTP request and pandas is used for data manipulation.



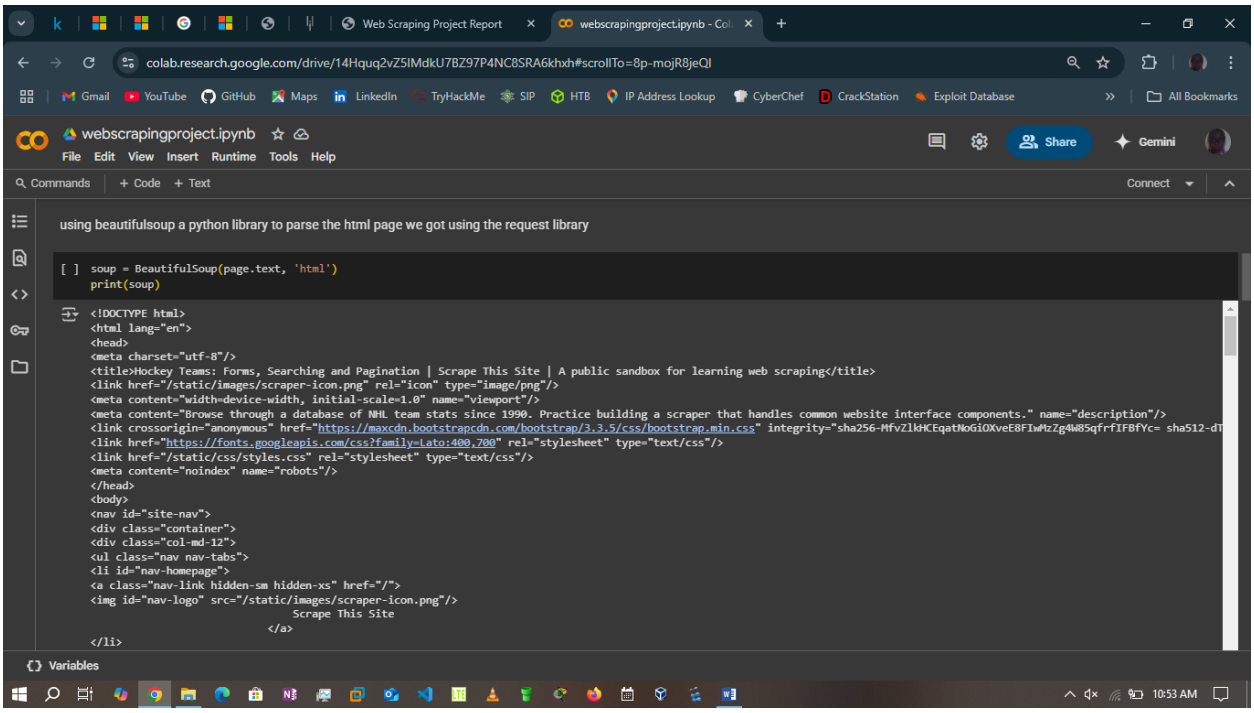
Step 2: Fetching the page

I specified the target URL and sent a GET request to retrieve the HTML content of the web page. The 200 is a response status code that confirms that the request was successful.



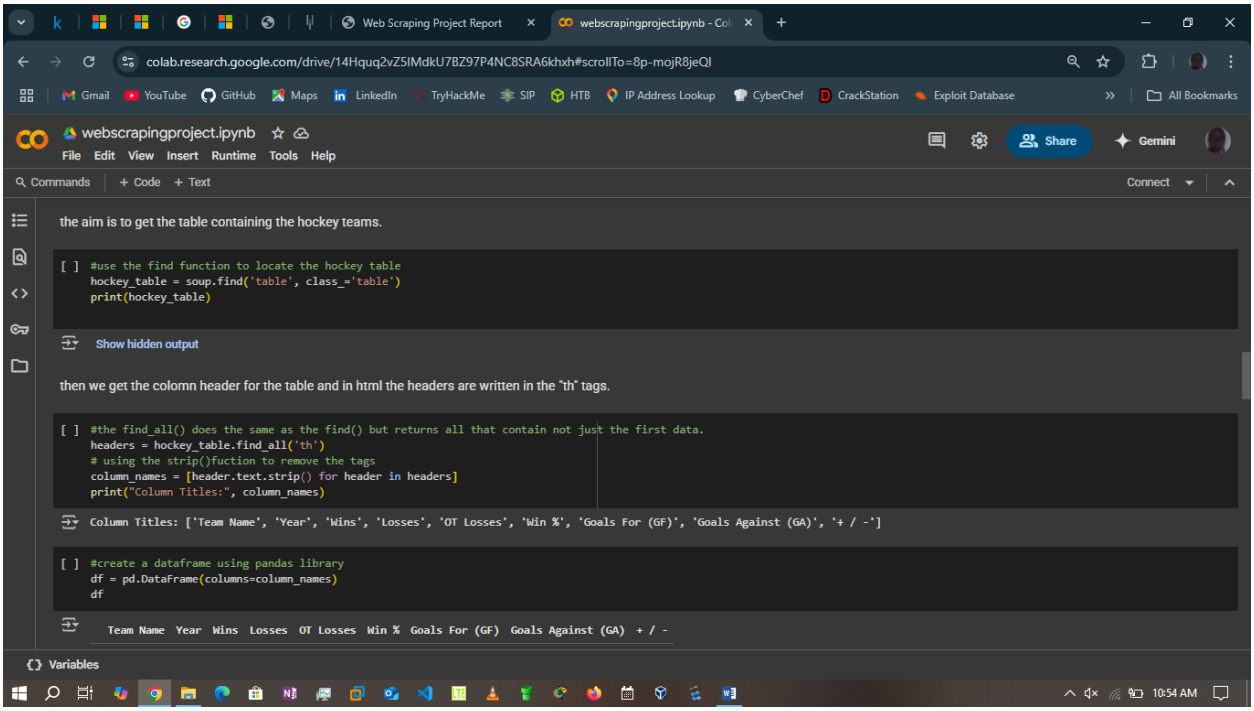
Step 3: parsing the HTML content

After retrieving the web page, I used beautifulsoup to parse the HTML content. Then using the find() function to locate the hockey table.



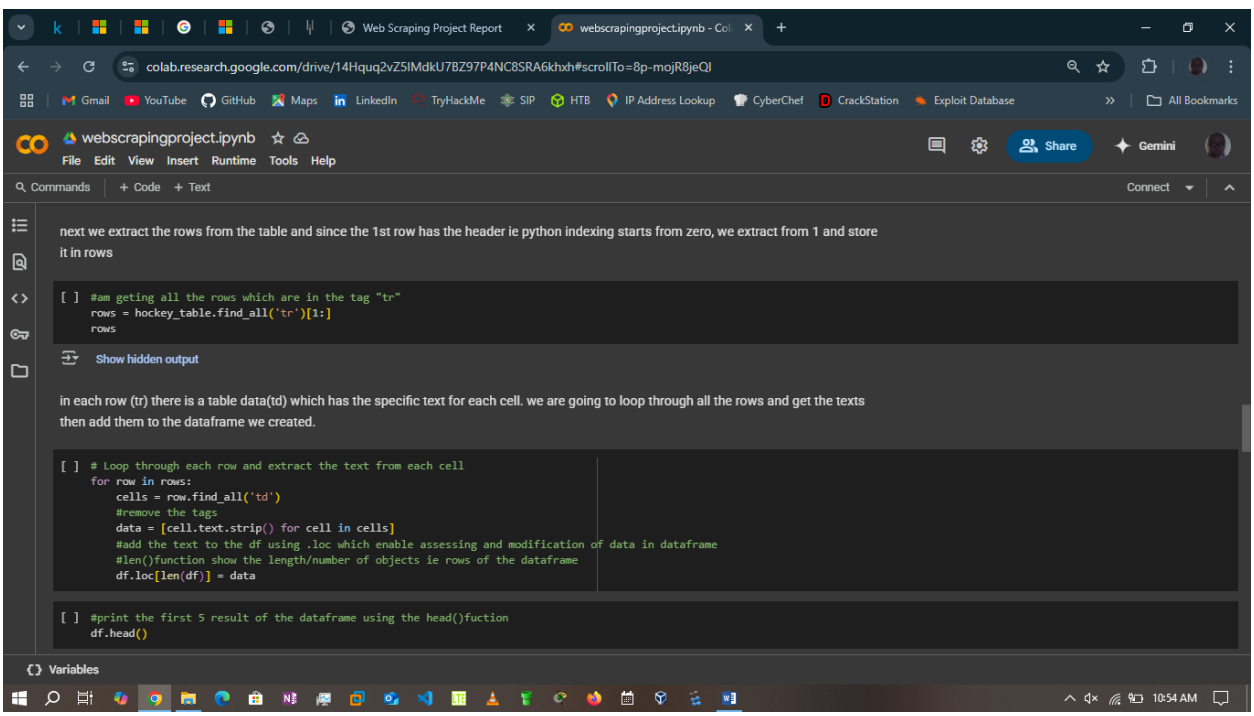
Step 4: Extracting the header column

I extracted the column header from the table to create the structure for the pandas dataframe.



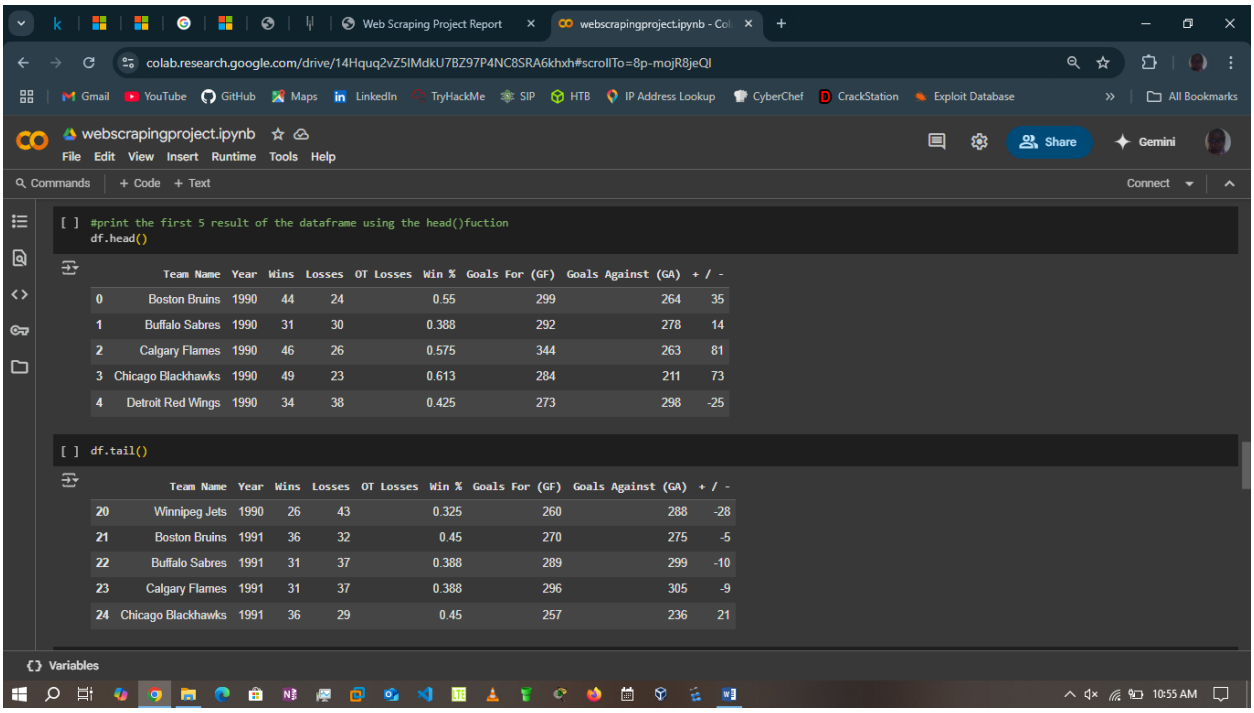
Step 5: Creating a dataframe and populating it

I then created the dataframe and extracted the column name in it. Then I also populated it data from the hockey table rows.



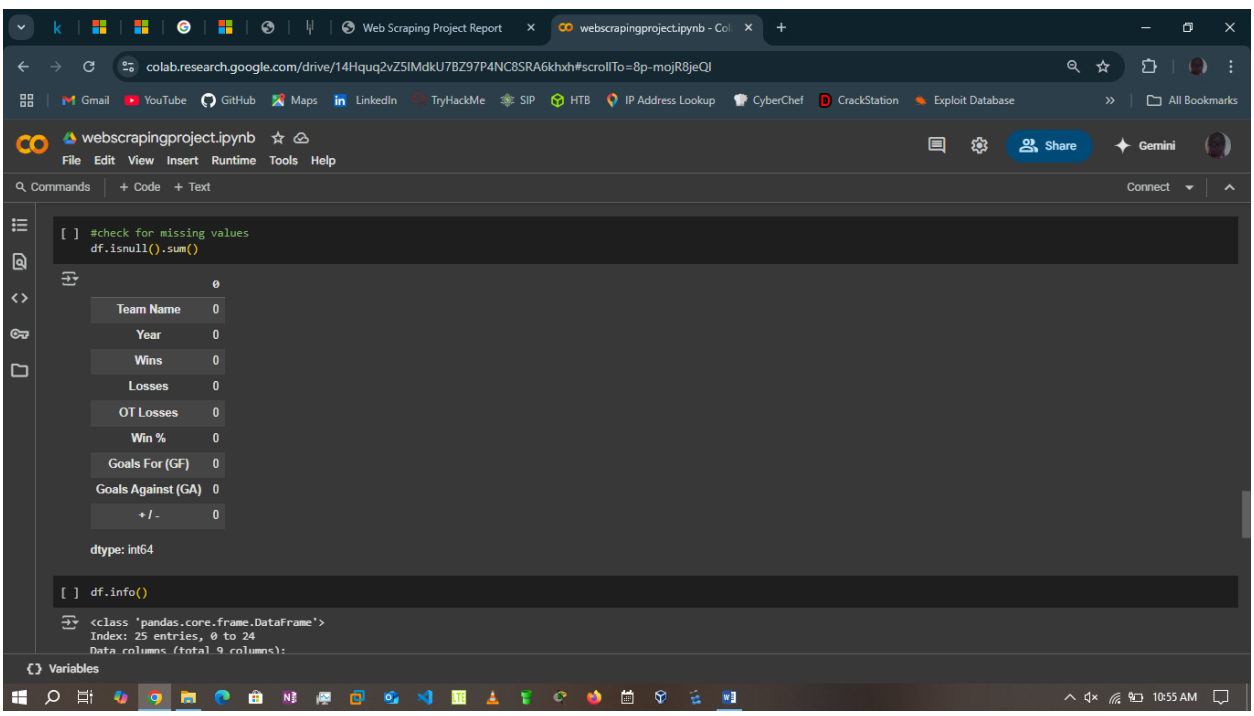
Step 6: Exploring the data

Here I was doing basic exploration of the data in the dataframe to understand the data in it. I used some common functions like head(), tail(), info(), unique() and checked for missing values.



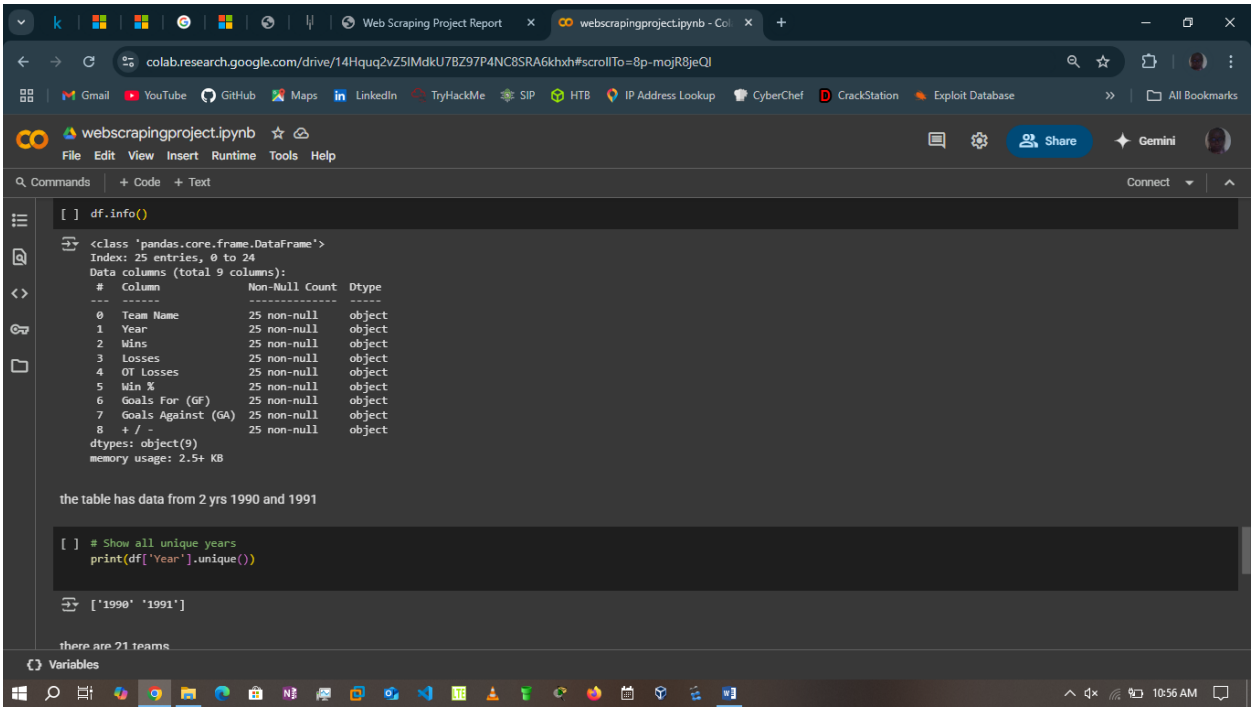
	Team Name	Year	Wins	Losses	OT Losses	Win %	Goals For (GF)	Goals Against (GA)	+ / -
0	Boston Bruins	1990	44	24		0.55	299	264	35
1	Buffalo Sabres	1990	31	30		0.388	292	278	14
2	Calgary Flames	1990	46	26		0.575	344	263	81
3	Chicago Blackhawks	1990	49	23		0.613	284	211	73
4	Detroit Red Wings	1990	34	38	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

	Team Name	Year	Wins	Losses	OT Losses	Win %	Goals For (GF)	Goals Against (GA)	+ / -
20	Winnipeg Jets	1990	26	43		0.325	260	288	-28
21	Boston Bruins	1991	36	32		0.45	270	275	-5
22	Buffalo Sabres	1991	31	37		0.388	289	299	-10
23	Calgary Flames	1991	31	37		0.388	296	305	-9
24	Chicago Blackhawks	1991	36	29		0.45	257	236	21



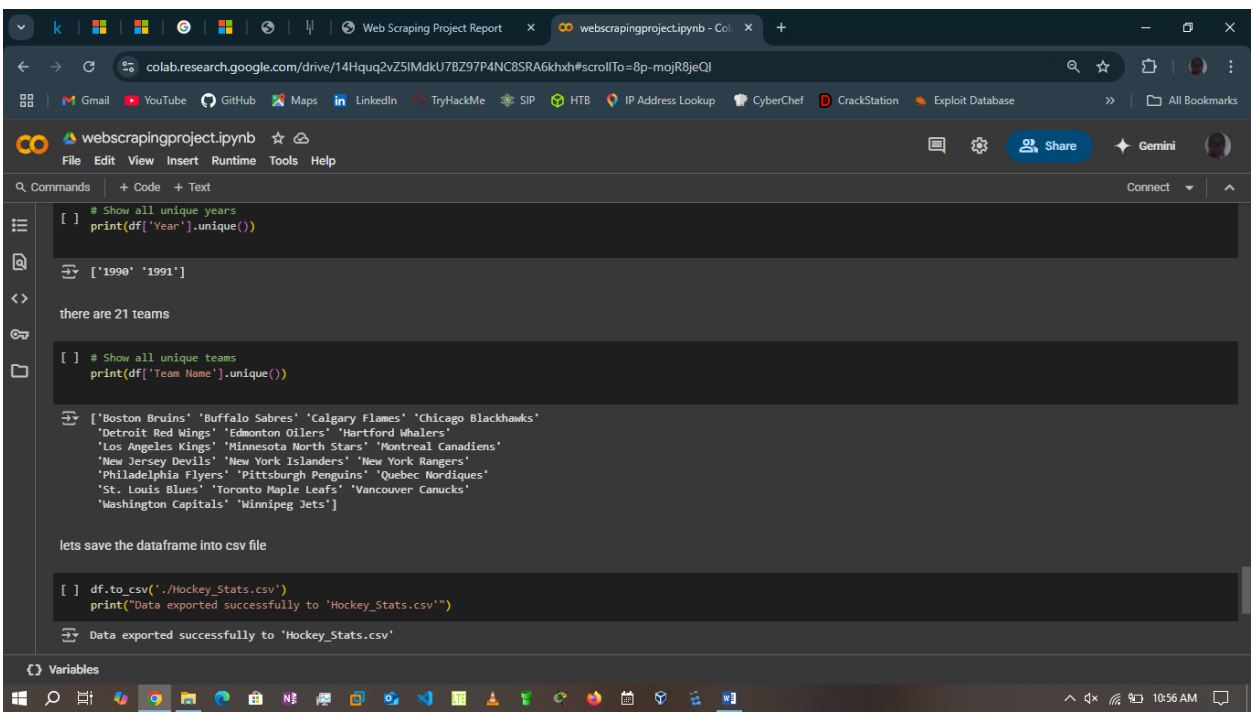
	0
Team Name	0
Year	0
Wins	0
Losses	0
OT Losses	0
Win %	0
Goals For (GF)	0
Goals Against (GA)	0
+ / -	0

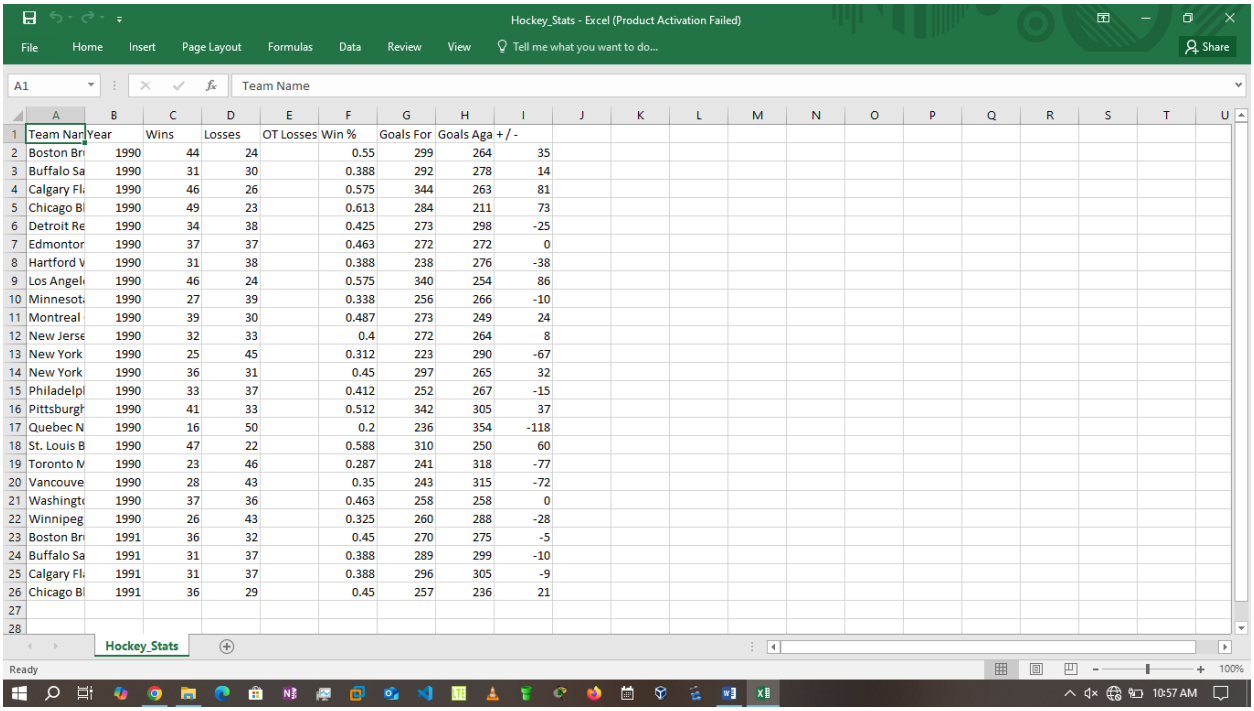
dtype: int64



Step 7: Exporting the csv

Finally I exported the dataframe to the CSV file then downloaded it from google colab. It opened as excel file as shown below:





Team Nan	Year	Wins	Losses	OT Losses	Win %	Goals For	Goals Aga	+ / -
Boston Br	1990	44	24		0.55	299	264	35
Buffalo Sa	1990	31	30		0.388	292	278	14
Calgary Fla	1990	46	26		0.575	344	263	81
Chicago Bl	1990	49	23		0.613	284	211	73
Detroit Re	1990	34	38		0.425	273	298	-25
Edmontor	1990	37	37		0.463	272	272	0
Hartford V	1990	31	38		0.388	238	276	-38
Los Angel	1990	46	24		0.575	340	254	86
Minnesota	1990	27	39		0.338	256	266	-10
Montreal	1990	39	30		0.487	273	249	24
New Jerse	1990	32	33		0.4	272	264	8
New York	1990	25	45		0.312	223	290	-67
New York	1990	36	31		0.45	297	265	32
Philadelpl	1990	33	37		0.412	252	267	-15
Pittsburgh	1990	41	33		0.512	342	305	37
Quebec N	1990	16	50		0.2	236	354	-118
St. Louis B	1990	47	22		0.588	310	250	60
Toronto M	1990	23	46		0.287	241	318	-77
Vancouve	1990	28	43		0.35	243	315	-72
Washingt	1990	37	36		0.463	258	258	0
Winnipeg	1990	26	43		0.325	260	288	-28
Boston Br	1991	36	32		0.45	270	275	-5
Buffalo Sa	1991	31	37		0.388	289	299	-10
Calgary Fla	1991	31	37		0.388	296	305	-9
Chicago Bl	1991	36	29		0.45	257	236	21

CONCLUSION

This web scraping project successfully demonstrated the process of extracting tabular data from a website and transforming it into a structured format for analysis. I was able to use python libraries like beautifulsoup, request and pandas to explore and extract data from the website HTML document. The data gathered can be used to train ML model that can be deployed for use.

NOTEBOOK

Link:

https://colab.research.google.com/drive/14Hquq2vZ5lMdkU7BZ97P4NC8SRA6khxh?usp=sharing