

Experiment 1

Name: Asad Shaikh

Branch: MTech CE

Registration Id: 242050023

Aim: To write a program to read the data from any online website, excel file and CSV file.

Theory:

In Big Data Analytics, the first step of any data analysis process is data collection or ingestion. Data can come from various sources such as:

- Online websites (like Wikipedia, financial data sites, etc.)
- Excel files used in business environments
- CSV (Comma Separated Values) files, a standard format for tabular data

Python, with libraries like pandas, provides powerful tools to read and manipulate these data formats efficiently.

1. Reading Data from an Online Website (HTML Table):

- Websites often contain tables of structured data (e.g., stock prices, country statistics).
- **pandas.read_html(url)** reads all tables from a webpage and returns a list of DataFrames.
- You need an internet connection and a well-structured HTML table for this to work.

2. Reading Data from an Excel File:

- Excel is commonly used in data reporting and storage.
- Python uses **pandas.read_excel()** to load .xlsx files.
- Requires the openpyxl library (for .xlsx format).

3. Reading Data from a CSV File:

- CSV is one of the most popular text-based formats for storing tabular data.
- Each row in the file is a data record, and each field is separated by a comma.
- Python uses **pandas.read_csv()** to read CSV files easily and efficiently.

Code / Output:

Installing openpyxl, importing pandas and drive

```
!pip install pandas requests openpyxl

import pandas as pd
from google.colab import drive
drive.mount('/content/drive')
```

Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (2.2.2)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (2.32.3)
Requirement already satisfied: openpyxl in /usr/local/lib/python3.11/dist-packages (3.1.5)
Requirement already satisfied: numpy>=1.23.2 in /usr/local/lib/python3.11/dist-packages (from pandas) (2.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas) (2025.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests) (2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests) (2025.4.26)
Requirement already satisfied: et-xmlfile in /usr/local/lib/python3.11/dist-packages (from openpyxl) (2.0.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
Mounted at /content/drive

Reading data from Website:

```
[14] # 1. Read data from an online website (HTML table)
url = "https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)"
html_tables = pd.read_html(url)
gdp_data = html_tables[2] # Table with GDP data
print("=== Online Website Data ===")
gdp_data.head()
```

=== Online Website Data ===

	Country/Territory	IMF[1][12]		World Bank[13]		United Nations[14]	
	Country/Territory	Forecast	Year	Estimate	Year	Estimate	Year
0	World	113795678	2025	105435540	2023	100834796	2022
1	United States	30507217	2025	27360935	2023	25744100	2022
2	China	19231705	[n 1]2025	17794782	[n 3]2023	17963170	[n 1]2022
3	Germany	4744804	2025	4456081	2023	4076923	2022
4	India	4187017	2025	3549919	2023	3465541	2022

Reading data from Excel file (.xlsx):

```
# 2. Read data from an Excel file
# Download the Excel file first or use an existing local file
excel_file = "/content/drive/My Drive/sem 2/BDA/datasets/MultiplicationTable.xlsx" # Replace with your file path
excel_data = pd.read_excel(excel_file)
print("\n=== Excel File Data ===")
excel_data.head()
```

=== Excel File Data ===

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10
0	NaN	1	2	3	4	5	6	7	8	9	10
1	NaN	2	4	6	8	10	12	14	16	18	20
2	NaN	3	6	9	12	15	18	21	24	27	30
3	NaN	4	8	12	16	20	24	28	32	36	40
4	NaN	5	10	15	20	25	30	35	40	45	50

Reading data from .csv file:

```
✓ 0s # 3. Read data from a CSV file
csv_file = "/content/drive/My Drive/sem 2/BDA/datasets/iris.csv"
csv_data = pd.read_csv(csv_file)
print("\n=== CSV File Data ===")
csv_data.head()
```

↔

=== CSV File Data ===

	sepal.length	sepal.width	petal.length	petal.width	variety
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa

Conclusion:

In this experiment, we learnt about python and we also learnt about how to read data from websites, excel files and csv files and display them in tabular format.