EX NO :

DATE :

Implementation of reading Datasets in different formats

AIM

To implement and read datasets in different formats (such as CSV, Excel, JSON etc) using jupyter notebook

QUESTIONS

1.Write a Python program to read a CSV file into a DataFrame.

```
import pandas as pd
q1=pd.read_csv('C:\\Users\\dines\\OneDrive\\Documents\\dep\\q1lab2.csv')
print(q1)
```

```
   Name        RegNo    Gender     Marks
0    Dinesh     2212046  Male         95
1    Asir       2212054  Male         96
2    Karan      2212047  Male         93
3    Shan       2212049  Male         94
4    Petchi     2212056  Male         95
```

2. Write a Pandas script to read multiple sheets from an Excel file into a dictionary of DataFrames.

```
import pandas as pd
q2=pd.read_excel('C:\\Users\\dines\\OneDrive\\Documents\\dep\\q2lab2.xlsx')
print(q2)
```

```
      Name   Dept     Email        City        State
0                                                NaN
1  Ravi   CSE      ravi@123    Sivakasi     Tamil...
2  Hari   ECE      hari@234    Sattur       Tamil...
3  Mark   IT       mark@456    Tuticorin    Tamil...
```

3. Write a Pandas script to read a CSV file from the URL https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv, and display the first 10 rows.

```
import pandas as pd
url='https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv'
df=pd.read_csv(url)
print(df.head(n=10))
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3
5            6         0       3
6            7         0       1
7            8         0       3
8            9         1       3
9           10         1       2

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0
5                                   Moran, Mr. James    male   NaN      0
6                            McCarthy, Mr. Timothy J    male  54.0      0
7                     Palsson, Master. Gosta Leonard    male   2.0      3
8  Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)  female  27.0      0
9                Nasser, Mrs. Nicholas (Adele Achem)  female  14.0      1

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

4. Write a Python program to read a JSON file data.json from C:\data\ and display the data types of each column in the DataFrame.

```
import pandas as pd
df_json=pd.read_json('C:\\Users\\dines\\OneDrive\\Documents\\dep\\p4.json')
print(df_json)
print(df_json.dtypes)
```

```
     name  age     car
0    John   30  Toyota
1  Dinesh   20    Benz
2     Ram   19    Audi
3   Karan   20    Tata
name    object
age      int64
car     object
dtype: object
```

5. Write a Python program to read all HTML tables from the URL https://www.w3schools.com/html/html_tables.asp and print the number of tables found

```
import pandas as pd
url = 'https://www.w3schools.com/html/html_tables.asp'
dfs = pd.read_html(url)
print("Number of tables found:", len(dfs))
for i, df in enumerate(dfs):
    print(f"\nTable {i + 1}")
    print(df.head())
```

```
Number of tables found: 2

Table 1
                          Company           Contact  Country
0            Alfreds Futterkiste     Maria Anders  Germany
1    Centro comercial Moctezuma   Francisco Chang   Mexico
2                  Ernst Handel     Roland Mendel  Austria
3                Island Trading     Helen Bennett       UK
4  Laughing Bacchus Winecellars   Yoshi Tannamuri   Canada

Table 2
         Tag                         Description
0    <table>                      Defines a table
1       <th>  Defines a header cell in a table
2       <tr>           Defines a row in a table
3       <td>          Defines a cell in a table
4  <caption>            Defines a table caption
```

6. Write a Pandas script to read a Zip Archive containing CSV files into Pandas DataFrames.

```
import pandas as pd
import os
import zipfile

zip_file_path = 'C:\\Users\\dines\\OneDrive\\Documents\\dep\\archive (2).zip'
extracted_dir = 'extracted_files'
os.makedirs(extracted_dir, exist_ok=True)
with zipfile.ZipFile(zip_file_path, 'r') as zip_ref:
    zip_ref.extractall(extracted_dir)
extracted_files = os.listdir(extracted_dir)
print('Extracted Files:', extracted_files)
```

```
Extracted Files: ['Athens 2004 Olympics Nations Medals.csv', 'Atlanta 1996 Olympics Nations Medals.csv', 'beijing_2022_Olympics_Nations_Medals.csv', 'heart_attack_dataset.csv', 'Lillehammer 1994 Olympics Nations Medals.csv', 'London 2012 Olympics Nations Medals.csv', 'Nagano 1998 Olympics Nations Medals.csv', 'Olympic_Games_(1994-2024).db', 'Paris 2024 Olympics Nations Medals.csv', 'PyeongChang 2018 Olympics Nations Medals.csv', 'Rio 2016 Olympics Nations Medals.csv', 'SaltLakeCity 2002 Olympics Nations Medals.csv', 'Sochi 2014 Olympics Nations Medals.csv', 'Sydney 2000 Olympics Nations Medals.csv', 'Tokyo 2020 Olympics Nations Medals.csv', 'Torino 2006 Olympics Nations Medals.csv', 'Vancouver 2010 Olympics Nations Medals.csv']
```

7. Write a Pandas script to read Feather files and output the column names and metadata of the resulting DataFrame.

```
import pandas as pd
import pyarrow.feather as feather
data = {
'mpg': [21.0, 21.0, 22.8, 21.4, 18.7],
'cyl': [6, 6, 4, 6, 8],
'disp': [160.0, 160.0, 108.0, 258.0, 360.0],
'hp': [110, 120, 93, 110, 175],
'drat': [3.90, 3.90, 3.85, 3.08, 3.15],
'wt': [2.620, 2.875, 2.320, 3.220, 3.440],
'qsec': [16.46, 17.02, 18.61, 19.44, 17.02],
'vs': [1, 0, 0, 0, 0],
'am': [1, 0, 1, 0, 0],
'gear': [3, 3, 4, 3, 4],
'carb': [1, 4, 3, 1, 2]  }
df = pd.DataFrame(data)
df.to_feather('mtcars.feather')
df_read = pd.read_feather('mtcars.feather')
colnames = df_read.columns.tolist()
import pyarrow as pa
table=pa.feather.read_table('mtcars.feather')
metadata = {
'schema': table.schema,
'num rows': table.num_rows,
'num columns': table.num_columns   }
print("Column names:", colnames)
print("Metadata", metadata)
```

```
Column names: ['mpg', 'cyl', 'disp', 'hp', 'drat', 'wt', 'qsec', 'vs', 'am', 'gear', 'carb']
Metadata {'schema': mpg: double
cyl: int64
disp: double
hp: int64
drat: double
wt: double
qsec: double
vs: int64
am: int64
gear: int64
carb: int64
```

8. Write a Python program to read an XML file named data.xml from C:\data\ and display the content in a DataFrame.

```
import pandas as pd
df=pd.read_xml('C:\\Users\\dines\\OneDrive\\Documents\\dep\\xmlfile.xml')
print(df)
```

```
   id       name  age gender admission_date  department
0   1  John Mark   21      M     2024-01-15  Cardiology
1   2      Smith   34      F     2024-02-20   Neurology
2   3    Johnson   29      F     2024-03-10    Oncology
```

9. How can you convert a YAML file into a panda DataFrame?

```
import yaml
with open('C:\\Users\\dines\\OneDrive\\Documents\\dep\\yamlfile.yml', 'r') as f:
    data=yaml.load(f,Loader=yaml.SafeLoader)
print(data)
```

```
{'company': 'spacelift', 'domain': ['devops', 'devsecops'], 'tutorial': [{'yaml': {'name': "YAML Ain't Markup Language", 'type': 'awesome', 'born': 200
1}}, {'json': {'name': 'JavaScript Object Notation', 'type': 'great', 'born': 2001}}, {'xml': {'name': 'Extensible Markup Language', 'type': 'good', 'bo
rn': 1996}}], 'author': 'omkarbirade', 'published': True}
```

10. Write a Pandas script to read fixed width formatted file into a panda DataFrame.

```
import pandas as pd
colspecs = [(0, 4), (4, 14), (14, 22), (22, 28)]
df=pd.read_fwf('C:\\Users\\dines\\OneDrive\\Documents\\dep\\fwffile.fwf',colspecs=colspecs, header=0)
print(df)
```

```
   101  Dinesh Man  ager 700   00 30
0  102  Ram Develo  per 6000    0 25
1  103  Asir Analy  st 55000     29
2  104  Karan Deve  loper 45  000 27
3  105  Petchi Ana  lyst 500  000 28
```

RUBRICS

Problem Understanding (10)	Implementation (20)	Viva (10)	Time Management (10)	Total (50)

RESULT

Thus the implementation to read datasets in different formats (such as CSV, Excel, JSON etc) using jupyter notebook was successfully executed and the output was verified