



Unit 13

csv module

Spread sheet and corresponding CSV file

Name	Exam1	Exam2	Final Exam	Overall Grade
Bill	75.00	100.00	50.00	75.00
Fred	50.00	50.00	50.00	50.00
Irving	0.00	0.00	0.00	0.00
Monty	100.00	100.00	100.00	100.00
Average				56.25

FIGURE 14.2 A simple spreadsheet from Microsoft Excel 2008.

```
Name,Exam1,Exam2,Final Exam,Overall Grade
Bill,75.00,100.00,50.00,75.00
Fred,50.00,50.00,50.00,50.00
Irving,0.00,0.00,0.00,0.00
Monty,100.00,100.00,100.00,100.00

Average,,,,56.25
```

csv file

- <https://docs.python.org/3/library/csv.html>
- Là file text mà mỗi dòng các phần tử phân tách nhau bởi dấu phẩy (,), chấm phẩy (;), hay tab (\t).
- Python hỗ trợ module csv, gồm:
 - csv.field_size_limit – return maximum field size
 - csv.get_dialect – get the dialect which is associated with the name
 - csv.list_dialects – show all registered dialects
 - csv.register_dialect - associate dialect with name
 - csv.unregister_dialect - delete the dialect associated with the name the dialect registry

csv module

- `csv.reader` – read data from a csv file
- `csv.writer` – write data to a csv file
- **`csv.QUOTE_ALL`** - Quote everything, regardless of type.
- **`csv.QUOTE_MINIMAL`** - Quote fields with special characters
- **`csv.QUOTE_NONNUMERIC`** - Quote all fields that aren't numbers value
- **`csv.QUOTE_NONE`** – Don't quote anything in output

Read csv

```
import csv
workbook_file = open('Workbook1.csv', 'r')
workbook_reader = csv.reader(workbook_file)

for row in workbook_reader:
    print(row)

workbook_file.close()
```

```
>>>
['Name', 'Exam1', 'Exam2', 'Final Exam', 'Overall Grade']
['Bill', '75.00', '100.00', '50.00', '75.00']
['Fred', '50.00', '50.00', '50.00', '50.00']
['Irving', '0.00', '0.00', '0.00', '0.00']
['Monty', '100.00', '100.00', '100.00', '100.00']
[]
['Average', '', '', '', '56.25']
```

Read file csv into Dictionary

```
import csv
with ('employee_birthday.txt', mode='r') as csv_file:
    csv_reader = csv.DictReader(csv_file)
    for row in csv_reader:
        print(f'\t{row["name"]} - {row["age"]}.')
```

Write csv

```
import csv
with ('employee_file.csv', mode='w') as employee_file:
    employee_writer = csv.writer(employee_file, delimiter=',',
                                  quotechar='"', quoting=csv.QUOTE_MINIMAL)

    employee_writer.writerow(['John Smith', 'Accounting', 'November'])
    employee_writer.writerow(['Erica Meyers', 'IT', 'March'])
```

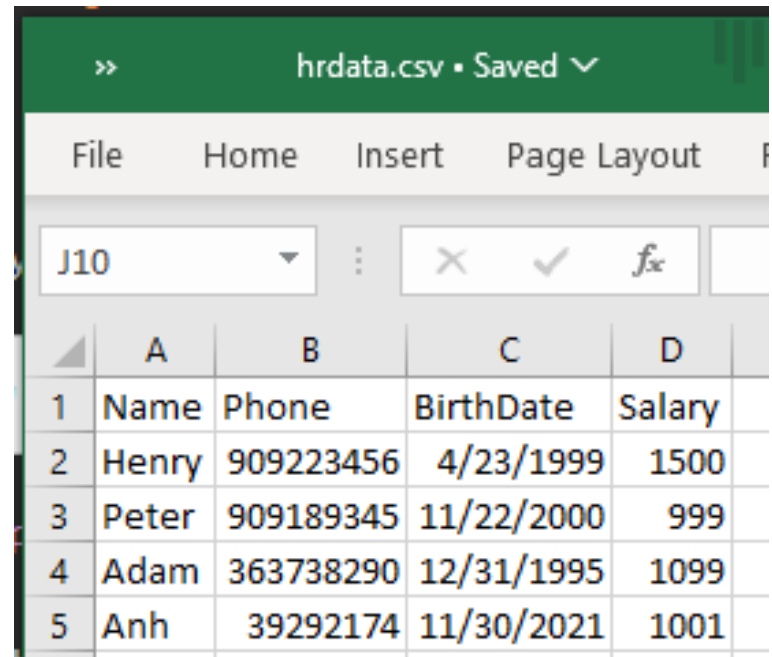
Write csv from Dict

```
import csv
with ('employee_file2.csv', mode='w') as csv_file:
    fieldnames = ['emp_name', 'dept']
    writer = csv.DictWriter(csv_file, fieldnames=fieldnames)
    writer.writeheader()
    writer.writerow({'emp_name': 'John Smith', 'dept': 'Accounting'})
    writer.writerow({'emp_name': 'Erica Meyers', 'dept': 'IT'})
```


Xử lý csv trong thư viện Pandas

```
import pandas
df = pandas.read_csv('hrdata.csv',
index_col='Name')
print(df)
```

```
import pandas
df = pandas.read_csv('hrdata.csv',
index_col='Name',
parse_dates=['Birth Date'])
print(df)
```



The screenshot shows the Microsoft Excel interface with a file named 'hrdata.csv' open. The ribbon at the top includes 'File', 'Home', 'Insert', and 'Page Layout'. The formula bar shows 'J10'. The data is organized into columns labeled A, B, C, and D. The first row contains headers: Name, Phone, BirthDate, and Salary. The subsequent rows contain data for five individuals: Henry, Peter, Adam, and Anh.

	A	B	C	D
1	Name	Phone	BirthDate	Salary
2	Henry	909223456	4/23/1999	1500
3	Peter	909189345	11/22/2000	999
4	Adam	363738290	12/31/1995	1099
5	Anh	39292174	11/30/2021	1001

Xử lý csv trong thư viện Pandas

```
df = pandas.read_csv('hrdata.csv',  
    index_col='Employee',  
    parse_dates=['Hired'],  
    header=0,  
    names=['Employee', 'Hired', 'Salary', 'Sick Days'])
```

```
# Process data
```

```
# Write new file
```

```
df.to_csv('hrdata_modified.csv')
```



Xử lý file excel

Xử lý excel với openpyxl

- Có rất nhiều thư viện Đọc Ghi và đọc file Excel
- Thư viện **openpyxl**:
 - `pip install openpyxl`
 - <https://openpyxl.readthedocs.io/en/stable>

Ghi file excel với openpyxl

```
# File excel (Workbook), trong file sẽ có nhiều Worksheet,  
# trong worksheet có nhiều cell
```

```
from openpyxl import Workbook
```

```
wb = Workbook()
```

```
# Tạo worksheet có tên NhatNghe
```

```
ws = wb.create_sheet("NhatNghe", 1)
```

```
ws["A1"] = "Trung Tâm Nhất Nghệ"    # Ghi ô "A1"
```

```
ws.append([2, 3, 4])                # Thêm dòng mới
```

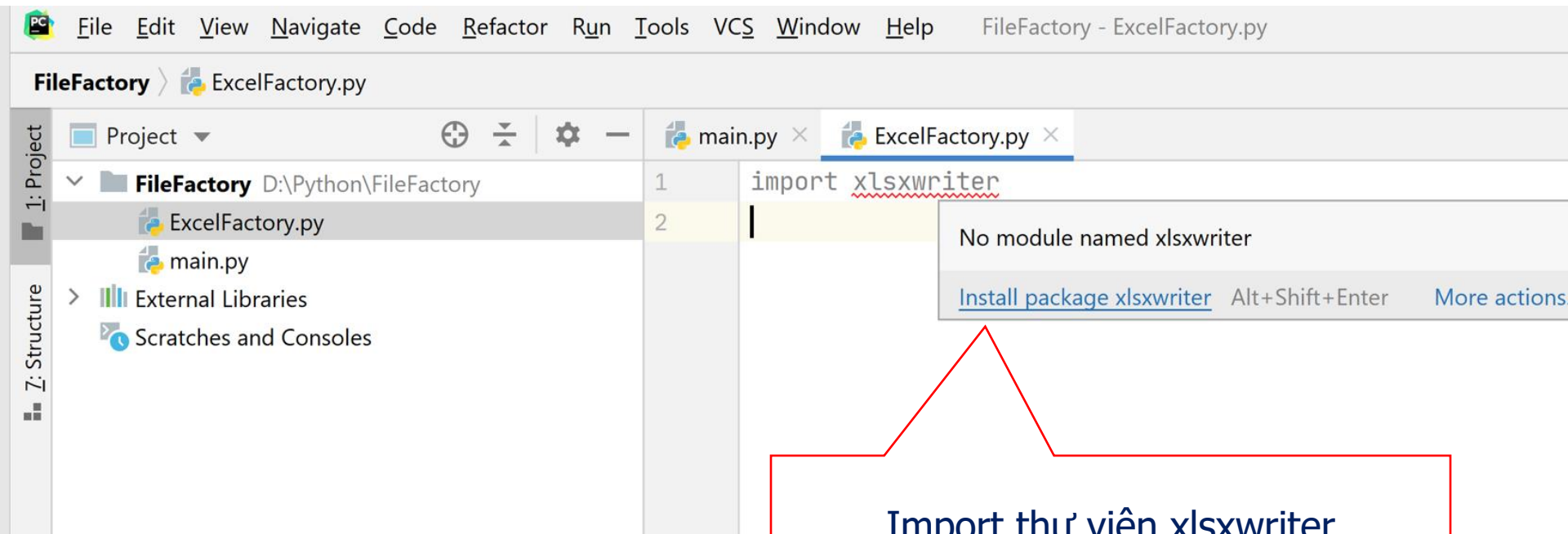
```
wb.save("DemoOpenpyxl.xlsx")        # Lưu file
```

Coding đọc dữ liệu trong file Excel

```
from openpyxl import load_workbook
wb = load_workbook('demo.xlsx')
print(wb.sheetnames)
ws = wb[wb.sheetnames[0]]
for row in ws.values:
    for value in row:
        print(value, "\t", end="")
    print("")
```

Xử lý file excel

- <https://xlsxwriter.readthedocs.io/> dùng thư viện này để tạo file Excel
`import xlsxwriter`



Import thư viện xlsxwriter
Di chuyển chuột tới dòng lệnh import
→ chọn Install package xlsxwriter

Xử lý file excel (tt)

```
import xlswriter
```

```
# Tạo một file excel cùng 1 sheet
```

```
workbook = xlswriter.Workbook('demo.xlsx')  
worksheet = workbook.add_worksheet()
```

```
# thiết lập các cột cho file
```

```
worksheet.set_column('A:A', 5)  
worksheet.set_column('B:B', 15)  
worksheet.set_column('C:C', 20)  
worksheet.set_column('D:D', 15)  
worksheet.set_column('E:E', 15)
```

```
# định dạng tiêu đề cột in đậm
```

```
bold = workbook.add_format({'bold': True})
```

```
# thêm dòng tiêu đề và định dạng in đậm
```

```
worksheet.write('A1', 'STT', bold)  
worksheet.write('B1', 'MÃ SẢN PHẨM', bold)  
worksheet.write('C1', 'TÊN SẢN PHẨM', bold)  
worksheet.write('D1', 'SỐ LƯỢNG', bold)  
worksheet.write('E1', 'ĐƠN GIÁ', bold)
```

```
#thêm một dòng dữ liệu
```

```
worksheet.write('A2', 1)  
worksheet.write('B2', 'SP1')  
worksheet.write('C2', 'Coca')  
worksheet.write('D2', '15')  
worksheet.write('E2', '15000')
```

```
#thêm một dòng dữ liệu
```

```
worksheet.write('A3', 2)  
worksheet.write('B3', 'SP2')  
worksheet.write('C3', 'Pepsi')  
worksheet.write('D3', '20')  
worksheet.write('E3', '18000')
```

```
#Chèn Logo vào
```

```
worksheet.insert_image('B5', 'HIENLTH.png')
```

```
workbook.close()
```

Chạy phần mềm và vào thư mục
phần mềm xem file Excel sẽ có kết
quả như mong muốn

FileFactory > ExcelOpenpyxl.py

Project
+
-
⚙️

FileFactory D:\Python\FileFacto

- demo.xlsx
- ExcelFactory.py
- ExcelOpenpyxl.py
- ExcelPandas.py
- logo_UEL.png
- main.py
- ReadExcelFile.py

> External Libraries
Scratches and Consoles

main.py x
ExcelFactory.py x
ReadExcelFile.py x
ExcelPandas.py x
ExcelOpenpyxl.py x

```

1 from openpyxl import load_workbook
2 wb = load_workbook('demo.xlsx')
3 print(wb.sheetnames)
4 ws = wb[wb.sheetnames[0]]
5 for row in ws.values:
6     for value in row:
7         print(value, "\t", end='')
8     print("")

```

for row in ws.values

Run: ExcelOpenpyxl x

C:\Python39\python.exe D:/Python/FileFactory/ExcelOpenpyxl.py

```

['Sheet1']

```

STT	MÃ SẢN PHẨM	TÊN SẢN PHẨM	SỐ LƯỢNG	ĐƠN GIÁ
1	SP1	Coca	15	15000
2	SP2	Pepsi	20	18000



Pandas

<https://pandas.pydata.org/index.html>

What is Pandas?

- Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures.
- The name Pandas is derived from the word Panel Data – an Econometrics from Multidimensional data.
- Pandas is used for:
 - Finance
 - Economics
 - Statistics
 - Analytics

Các thành phần Pandas

- Hai thành phần chính của Pandas là **Series** và **DataFrame**.
- Series về cơ bản là một cột.
- DataFrame là một bảng đa chiều được tạo thành từ một tập hợp các Chuỗi (Series).

Series

	apples
0	3
1	2
2	0
3	1

+

Series

	oranges
0	0
1	3
2	7
3	2

=

DataFrame

	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2

Cách tạo dataframe

- Có rất nhiều cách để tạo mới một DataFrame, một trong những lựa chọn tuyệt vời là sử dụng dict.
- Ví dụ: Chúng ta có một quầy bán táo và cam, giờ ta cần có một cột cho mỗi loại trái cây và một hàng cho mỗi lần mua hàng của khách hàng. Theo mục đích đó, ta sẽ có lệnh như sau:

```
data = {  
    'apples': [3, 2, 0, 1],  
    'oranges': [0, 3, 7, 2]  
}  
purchases = pd.DataFrame(data)
```

	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2

Cách tạo dataframe (tt)

- Tạo chỉ số cho dataframe

```
purchases = pd.DataFrame(data, index=['June', 'Robert', 'Lily', 'David'])
```


	Apple	Oranges
June	3	0
Robert	2	3
Lily	0	7
David	1	2

Cách tạo dataframe (tt)

- Đọc từ file csv:
 - `df = pd.read_csv('purchases.csv')`
 - `df = pd.read_csv('purchase.csv ', index_col = 0)`
- Đọc từ file json : `df = pd.read_json('purchases.json')`
- Đọc từ database

Các lệnh xử lý trên dataframe

- **df** = pandas.read_excel("data/FinancialSample.xlsx", header=None, skiprows=3)
- print(**df**) # Chỉ in tương trưng 10 dòng (5 đầu, 5 cuối)
- print("Số lượng dòng cột:", df.**shape**)
- print(**df.head**(10)) # In 10 dòng đầu
- print(**df.tail**(10)) # In 10 dòng cuối
- print(**df.describe**()) # Thống kê
- row_header = **df.head**() # Lấy dòng header
- print(row_header)

 **df.to_excel**("FinancialResult.xlsx") # Save dataframe
python

HIEUNLTH

Các lệnh xử lý trên dataframe

- `print(df.info())` # Lấy thông tin (entries?, column name?,...)
- `print(df.drop_duplicates())` # Loại bỏ dòng trùng
- `print(df.isnull())`
- `print(df.dropna())` # Drop null value
- `print(df.sum())`
- `print(df.max())`
- `print(df.mean())`
- ...

Tham khảo

- https://phamdinhhkhanh.github.io/deepai-book/ch_appendix/appendix_pandas.html