Міністерство освіти і науки України

Львівський національний університет імені Івана Франка

факультет електроніки та комп’ютерних технологій

кафедра системного проектування

**ЗВІТ**

про виконання лабораторної роботи №4

“Аналіз даних з використанням бібліотеки Pandas ”

виконала студентка групи ФеС-32

Богоніс Єлизавета

Перевірив старший викладач

Рибак А.В.

Львів-2019

**ТЕОРЕТИЧНІ ВІДОМОСТІ**

**Pandas** – програмна бібліотека, написана для мови програмування [Python](https://uk.wikipedia.org/wiki/Python" \o "Python) для маніпулювання даними та їхнього аналізу. Вона, зокрема, пропонує структури даних та операції для маніпулювання чисельними таблицями та [часовими рядами](https://uk.wikipedia.org/wiki/%D0%A7%D0%B0%D1%81%D0%BE%D0%B2%D0%B8%D0%B9_%D1%80%D1%8F%D0%B4). pandas є [вільним програмним забезпеченням](https://uk.wikipedia.org/wiki/%D0%92%D1%96%D0%BB%D1%8C%D0%BD%D0%B5_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BD%D0%B5_%D0%B7%D0%B0%D0%B1%D0%B5%D0%B7%D0%BF%D0%B5%D1%87%D0%B5%D0%BD%D0%BD%D1%8F), що випускається за трипунктовою [ліцензією BSD](https://uk.wikipedia.org/wiki/%D0%9B%D1%96%D1%86%D0%B5%D0%BD%D0%B7%D1%96%D1%8F_BSD). Ця назва походить від терміну [«панельні дані](https://uk.wikipedia.org/w/index.php?title=%D0%9F%D0%B0%D0%BD%D0%B5%D0%BB%D1%8C%D0%BD%D1%96_%D0%B4%D0%B0%D0%BD%D1%96&action=edit&redlink=1)» ([англ.](https://uk.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D1%96%D0%B9%D1%81%D1%8C%D0%BA%D0%B0_%D0%BC%D0%BE%D0%B2%D0%B0" \o "Англійська мова) *panel data*), який в [економетрії](https://uk.wikipedia.org/wiki/%D0%95%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D1%96%D1%8F) позначає багатовимірні структуровані набори даних.

**Завдання:**

1. Оглянути особливості структурних типів даних бібліотеки Pandas: Series, DataFrame і Index.

2. Імпорт даних з файлу \*.csv в структурний об’єкт DataFrame, валідація даних (відкидання або доповнення пропущених даних) і експорт валідованих даних у файл (\*.csv, \*.xlsx).

3. Виконати об’єднання даних (merge/join). Підготувати набір даних для двох таблиць і виконати такі типи об’єднань: left join, inner join, right join, full outer join.

4. Виконати наступні методи аналізу даних: агрегація (sum(), mean(), median(), min(), max()), групування, фільтрація тощо.

5. Висновок

**Хід роботи**

1. Оглянути особливості структурних типів даних бібліотеки Pandas: Series, DataFrame і Index.

**Series** - це одновимірний індексований масив, здатний зберігати дані будь-якого типу (цілі, рядкові, поплавкові, пітонні об'єкти тощо).

Якщо Series являє собою одновимірну структуру, яку для себе можна уявити як таблицю з одним рядком, то **DataFrame** - це вже двовимірна структура - повноцінна таблиця з безліччю рядків і стовпців.

2&4. Імпорт даних з файлу \*.csv в структурний об’єкт DataFrame, валідація даних (відкидання або доповнення пропущених даних) і експорт валідованих даних у файл (\*.csv, \*.xlsx). Виконати наступні методи аналізу даних: агрегація (sum(), mean(), median(), min(), max()), групування, фільтрація тощо.

Другий та четвертий пункти були зроблені разом, у одній програмі.

Лістинг програми:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from openpyxl.workbook import Workbook

df = pd.DataFrame(

{"Age": (5, 6, np.nan , 8, 9, 5),

"Income": (9, 8, 5, 2, np.nan , 9)} )

print(df)

# std - стандартне відхилення , count - к-сть чисел

print(df.describe().loc[["mean", "std", "min", "max", "count"]].T)

#зчитуємо дані з файлу

file = pd.read\_csv("test1.csv") #файл теж є у моєму репозиторії

kf = pd.DataFrame(file)

print(file)

print(kf.describe())

#ПУСТІ ЗНАЧЕННЯ, ЇХ ЗАМІНА/ВИДАЛЕННЯ

d = df.dropna() *#видаляє пусті значення*

print(d)

print([df.notnull()]) *#перевіряємо чи є пусті значення*

k = kf.fillna(1) *#заміняємо пусті значення на 1*

print(k)

print([kf.notnull()]) #*перевіряємо чи є пусті значення*

#СУМА ТА АГРЕГАЦІЯ

print(kf.sum(axis = 1, skipna = True)) *#сума*

print(kf.describe().loc[["mean", "std", "min", "max", "count"]].T)

#ФІЛЬТРУВАННЯ

def filter\_func(x):

return x['Income'].sum() > 4 *#якщо більше 4 , то видаляється*

print(df)

print(df.groupby('Age').sum()) *# сума*

print(df.groupby('Age').filter(filter\_func)) *#фільтрується , видаляється те, що менше 4 і виводиться*

#ПЕРЕТВОРЕННЯ

k1 = kf.groupby('MSSubClass').transform(lambda x: x - x.mean()) *#цей стовпець видаляє, а від всіх інших віднімає своє середнє значення*

print(k1)

#ГРУПУВАННЯ

print('\t\tGROUPBY\n', kf.groupby('MoSold').mean()) #групує за "MoSold" та визначає середнє значення

#ЕКСПОРТ

d.to\_excel('df\_valid\_1.xlsx')

#створює таблицю в excel з валідованих даних

k.to\_excel('kf\_valid\_2.xlsx')

Результат програми:

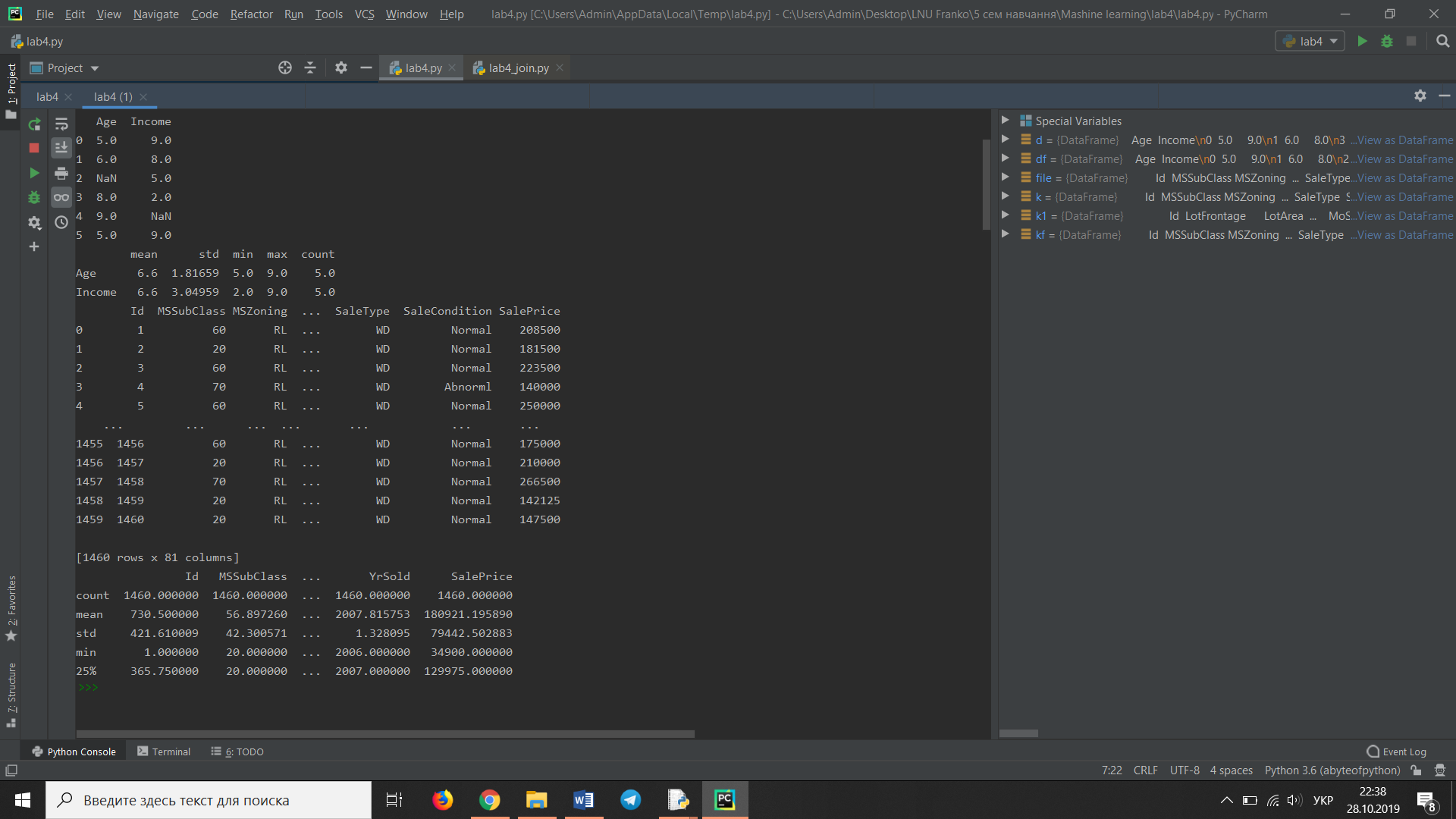


Рис.1 Імпорт, експорт даних та їх аналіз

3. Виконати об’єднання даних (merge/join). Підготувати набір даних для двох таблиць і виконати такі типи об’єднань: left join, inner join, right join, full outer join.

Лістинг програми :

import numpy as np

import pandas as pd

df1 = pd.DataFrame({'name': ['Olya', 'Artur', 'Maria'],

'study': ['english', 'chemistry', 'mathematics']},columns=['name', 'study'])

df2 = pd.DataFrame({'name': ['Maria', 'Artur', 'Lilia'],

'drink': ['coffee', 'green tea', 'lemonade']}, columns=['name', 'drink'])

print("Tab1\n", df1)

print("Tab2\n", df2)

#INNER JOIN

k = pd.merge(df1, df2, how='inner') #співпадає Mary і Artur

print("Inner join:\n", k)

#LEFT JOIN

k1 = pd.merge(df1, df2, how='left') #з'єднання з ліва на право

print("Left join:\n", k1)

#RIGHT JOIN

k2 = pd.merge(df1, df2, how='right') #з'єднання з право на ліво

print("Right join:\n", k2)

#OUTER JOIN

k3 = pd.merge(df1, df2, how='outer') #з'єднання з право на ліво

print("Outer join:\n", k3)

Результат програми:

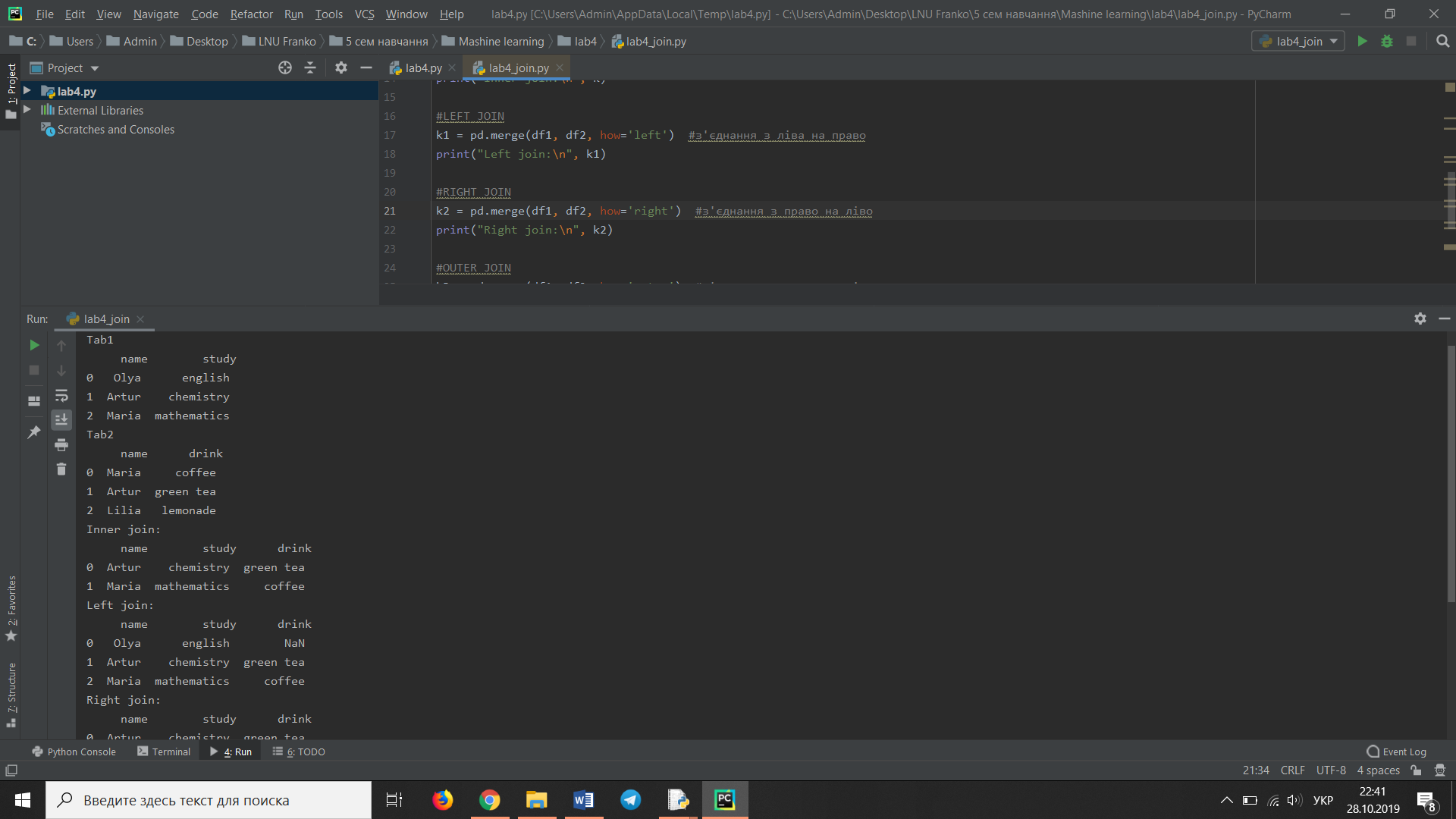


Рис.2 Об’єднання даних

**Висновок:** виконавши лабораторну роботу, ми оглянули особливості структурних типів даних бібліотеки Pandas: Series, DataFrame і Index.