 МІНІСТЕРСТВО ОСВІТИ I НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ   
ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

ФАКУЛЬТЕТ БІОМЕДИЧНОЇ ІНЖЕНЕРІЇ

КАФЕДРА БІОМЕДИЧНОЇ КІБЕРНЕТИКИ

**Комп’ютерний практикум №3**

з дисципліни «Методи та технології обчислювального інтелекту»

на тему: «Основи роботи з Pandas»

**Виконала:**

студентка гр. БС-01мп

Харченко Наталя

**Перевірив:**

доц. Федорін І.В.

Зараховано від \_\_\_.\_\_\_.\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(підпис викладача)

Київ-2020

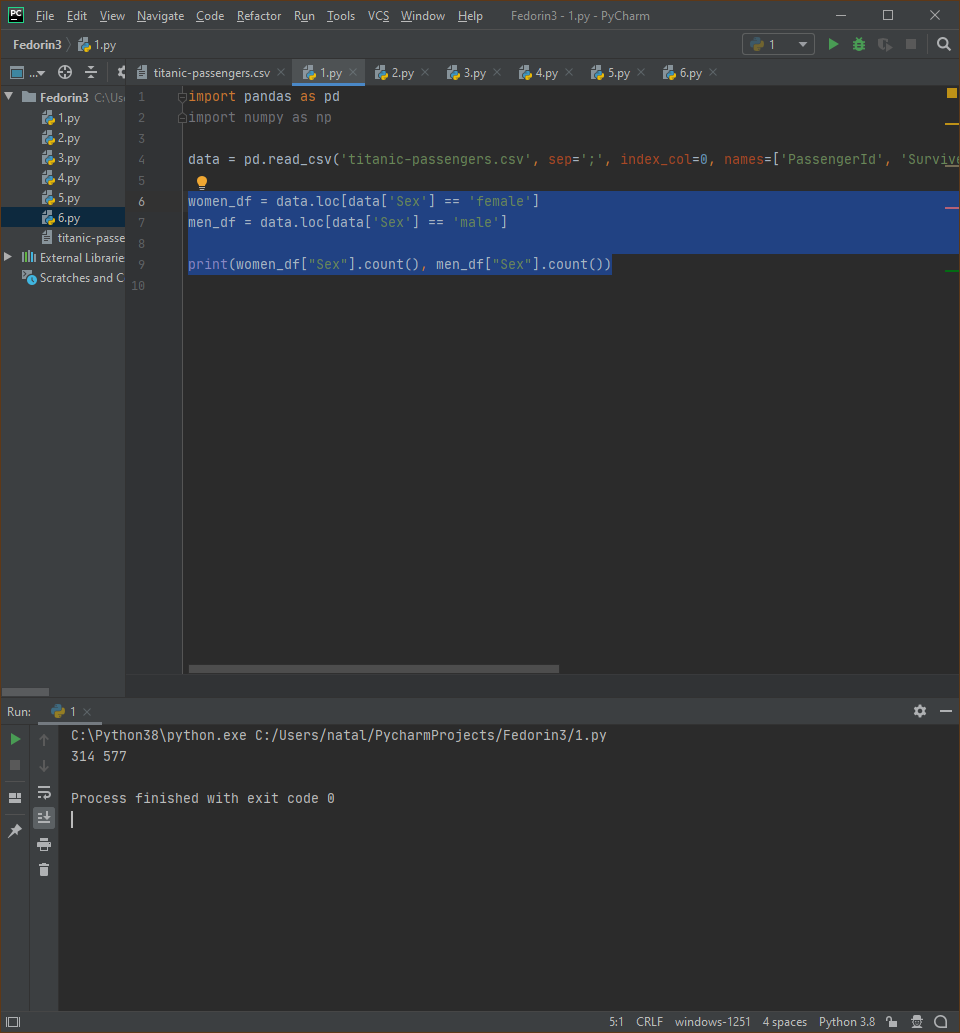
**Практичне завдання 1**

1. Яка кількість чоловіків і жінок їхала на кораблі? Приведіть два числа через пробіл.

**Програмна реалізація:**

import pandas as pd   
  
data = pd.read\_csv('titanic-passengers.csv', sep=';', index\_col=0, names=['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'])  
  
women\_df = data.loc[data['Sex'] == 'female']  
men\_df = data.loc[data['Sex'] == 'male']  
  
print(women\_df["Sex"].count(), men\_df["Sex"].count())

**Результат:**



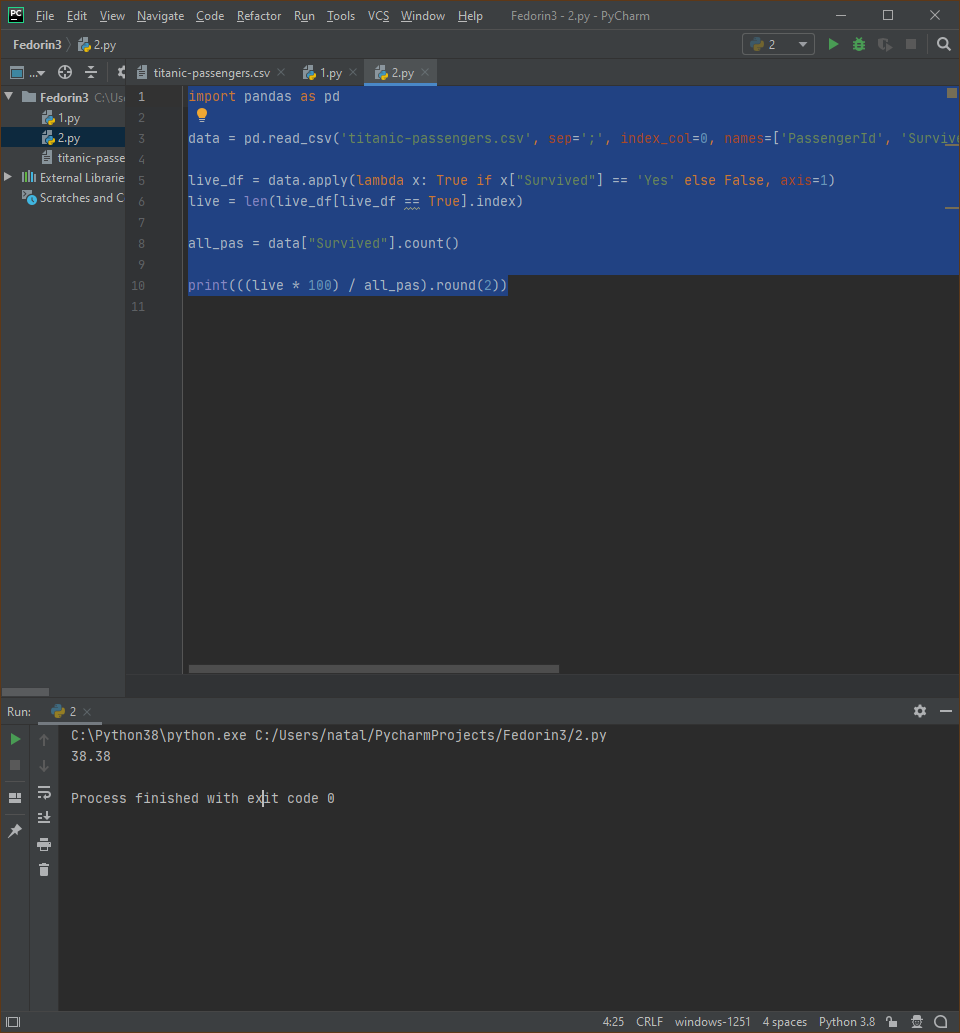
1. Якій частині пасажирів вдалося вижити? Порахуйте частку пасажирів, що вижили. Відповідь приведіть у відсотках (число в інтервалі від 0 до 100, знак відсотка не потрібний), округливши до двох знаків.

**Програмна реалізація:**

import pandas as pd  
  
data = pd.read\_csv('titanic-passengers.csv', sep=';', index\_col=0, names=['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'])  
  
live\_df = data.loc[data["Survived"] == "Yes"]  
live = live\_df["Survived"].count()

all\_pas = data["Survived"].count()  
  
print(((live \* 100) / all\_pas).round(2))

**Результат:**

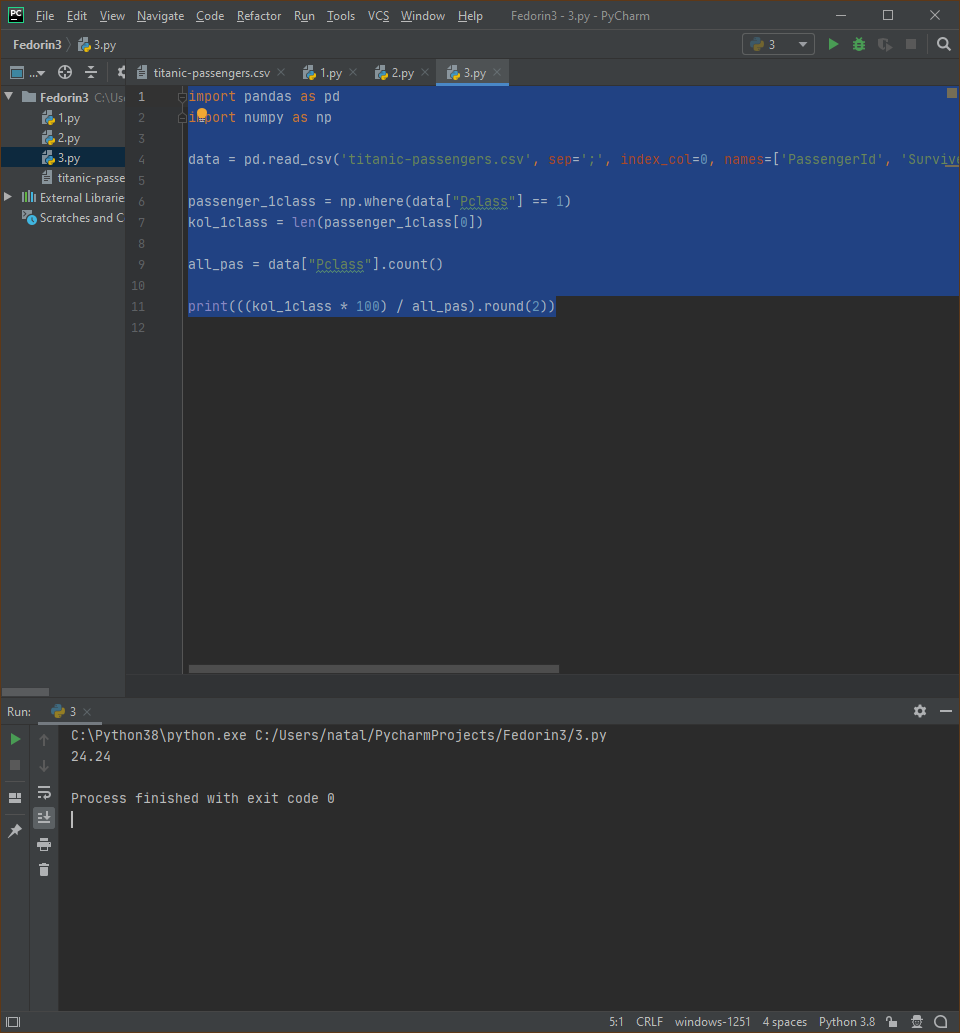


1. Яку частку пасажири першого класу складали серед всіх пасажирів? Відповідь приведіть у відсотках (число в інтервалі від 0 до 100, знак відсотка не потрібний), округливши до двох знаків.

**Програмна реалізація:**

import pandas as pd   
  
data = pd.read\_csv('titanic-passengers.csv', sep=';', index\_col=0, names=['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'])  
passenger\_1class = data.loc[data["Pclass"] == 1]  
kol\_1class = passenger\_1class["Pclass"].count()  
all\_pas = data["Pclass"].count()  
  
print(((kol\_1class \* 100) / all\_pas).round(2))

**Результат:**

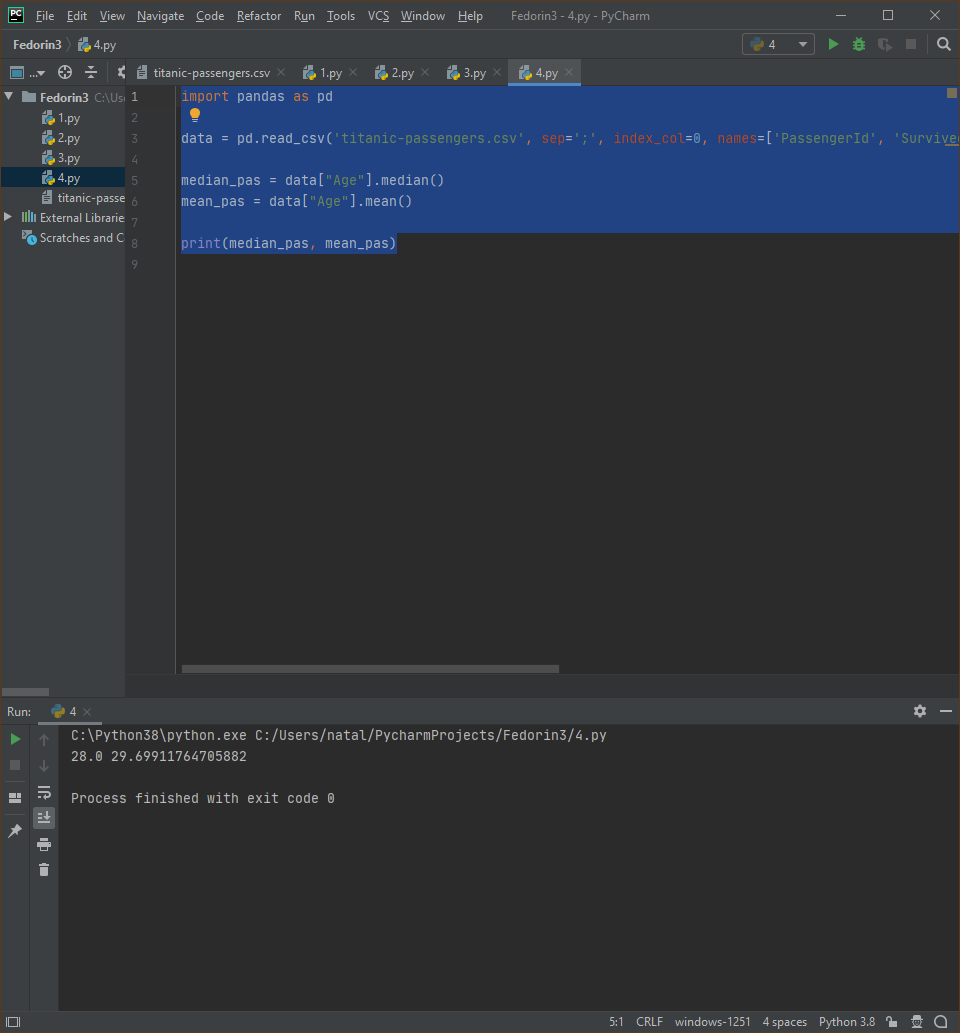


1. Якого віку були пасажири? Порахуйте середнє і медіану віку пасажирів. Як відповідь приведіть два числа через пробіл.

**Програмна реалізація:**

import pandas as pd  
  
data = pd.read\_csv('titanic-passengers.csv', sep=';', index\_col=0, names=['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'])  
  
median\_pas = data["Age"].median()  
mean\_pas = data["Age"].mean()  
  
print(median\_pas, mean\_pas)

**Результат:**

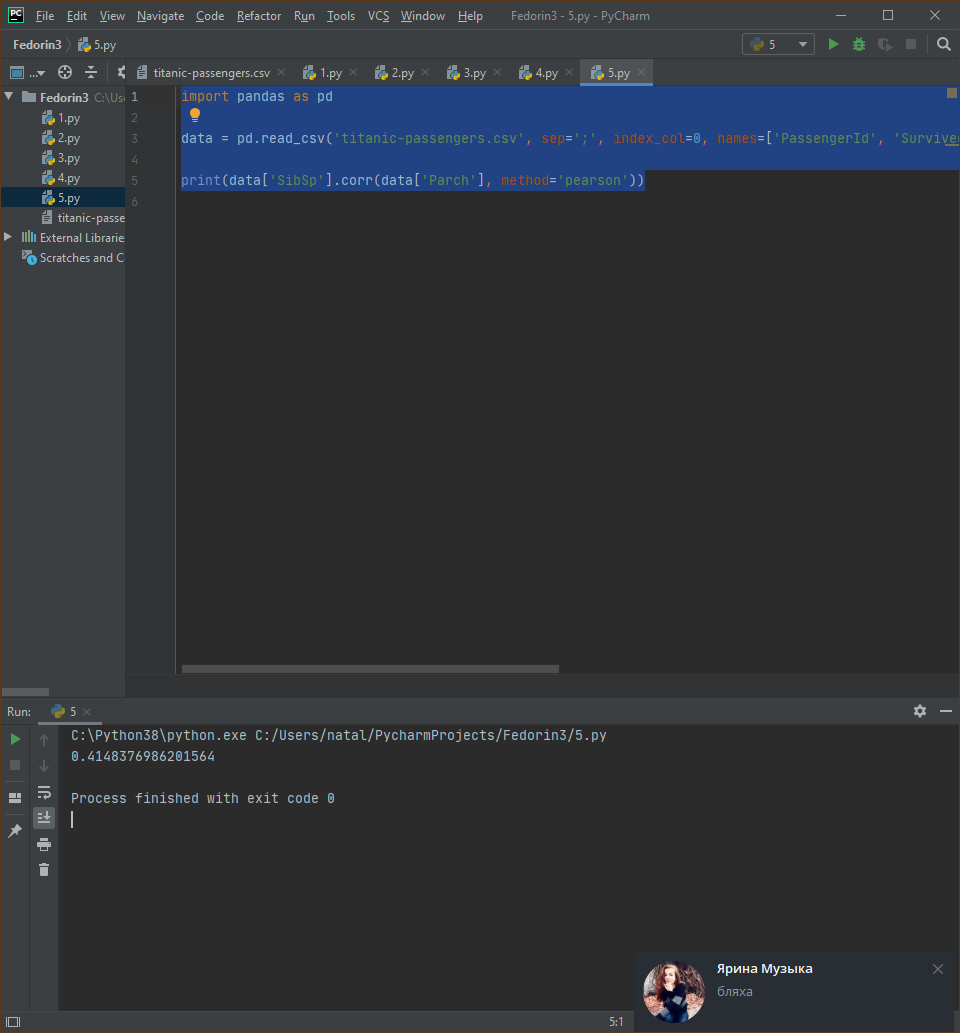


1. Чи корелює число братів / сестер / подружжя з числом батьків / дітей? Порахуйте кореляцію Пірсона між ознаками SibSp і Parch.

**Програмна реалізація:**

import pandas as pd  
  
data = pd.read\_csv('titanic-passengers.csv', sep=';', index\_col=0, names=['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'])  
  
print(data['SibSp'].corr(data['Parch'], method='pearson'))

**Результат:**

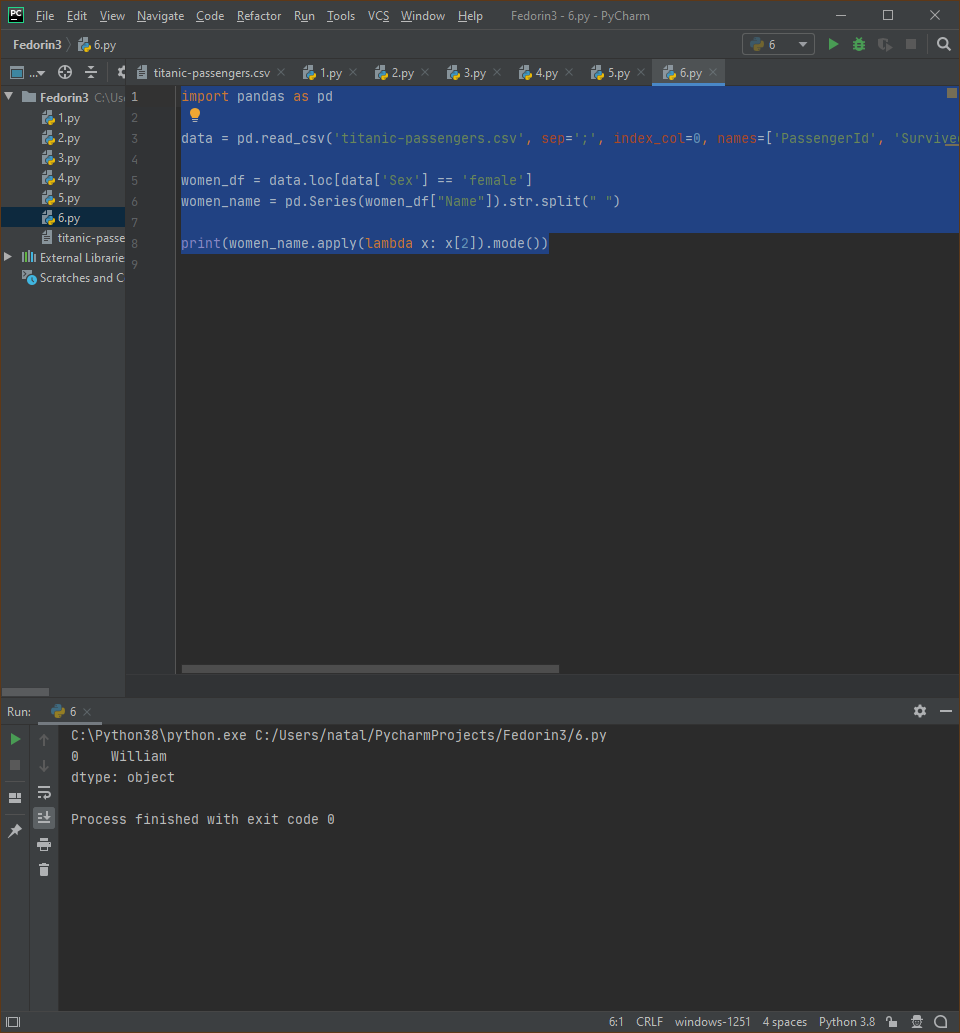


1. Яке найпопулярніше жіноче ім'я на кораблі? Вийміть з повного імені пасажира (колонка Name) його особисте ім'я (First Name). Спробуйте вручну розібрати кілька значень стовпця Name і виробити правило для вилучення імен, а також поділу їх на жіночі і чоловічі.

**Програмна реалізація:**

import pandas as pd  
  
data = pd.read\_csv('titanic-passengers.csv', sep=';', index\_col=0, names=['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'])  
women\_df = data.loc[data['Sex'] == 'female']  
women\_name = pd.Series(women\_df["Name"]).str.split(" ")  
  
print(women\_name.apply(lambda x: x[2]).mode())

**Результат:**



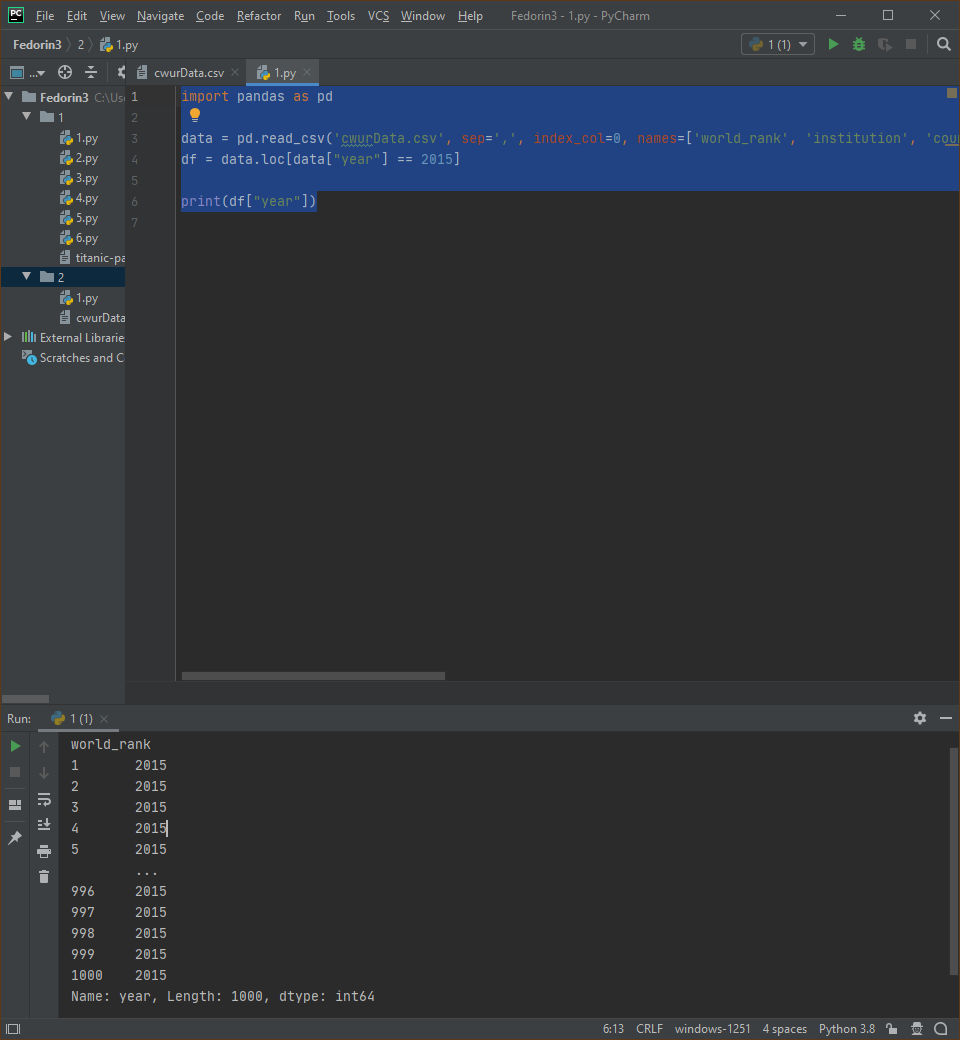
**Практичне завдання 2**

1. У рейтингу залиште тільки дані за 2015 рік.

**Програмна реалізація:**

import pandas as pd  
  
data = pd.read\_csv('cwurData.csv', sep=',', index\_col=0, names=['world\_rank', 'institution', 'country', 'national\_rank', 'quality\_of\_education', 'alumni\_employment', 'quality\_of\_faculty', 'publications', 'influence', 'citations', 'broad\_impact', 'patents', 'score', 'year'])  
df = data.loc[data["year"] == 2015]  
  
print(df["year"])

**Результат:**

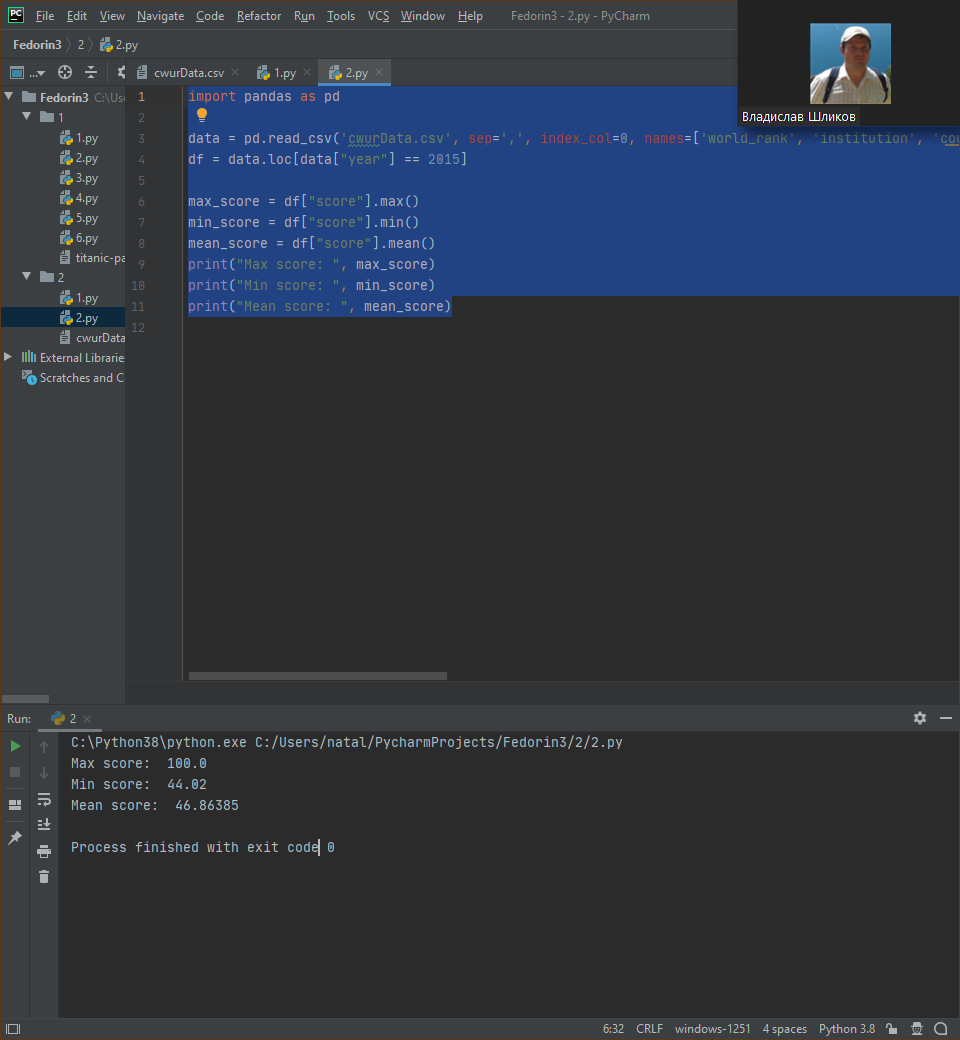


1. Обчисліть максимальний, мінімальний і середній загальний бал світового рейтингу.

**Програмна реалізація:**

import pandas as pd  
  
data = pd.read\_csv('cwurData.csv', sep=',', index\_col=0, names=['world\_rank', 'institution', 'country', 'national\_rank', 'quality\_of\_education', 'alumni\_employment', 'quality\_of\_faculty', 'publications', 'influence', 'citations', 'broad\_impact', 'patents', 'score', 'year'])  
df = data.loc[data["year"] == 2015]  
  
max\_score = df["score"].max()  
min\_score = df["score"].min()  
mean\_score = df["score"].mean()  
print("Max score: ", max\_score)  
print("Min score: ", min\_score)  
print("Mean score: ", mean\_score)

**Результат:**

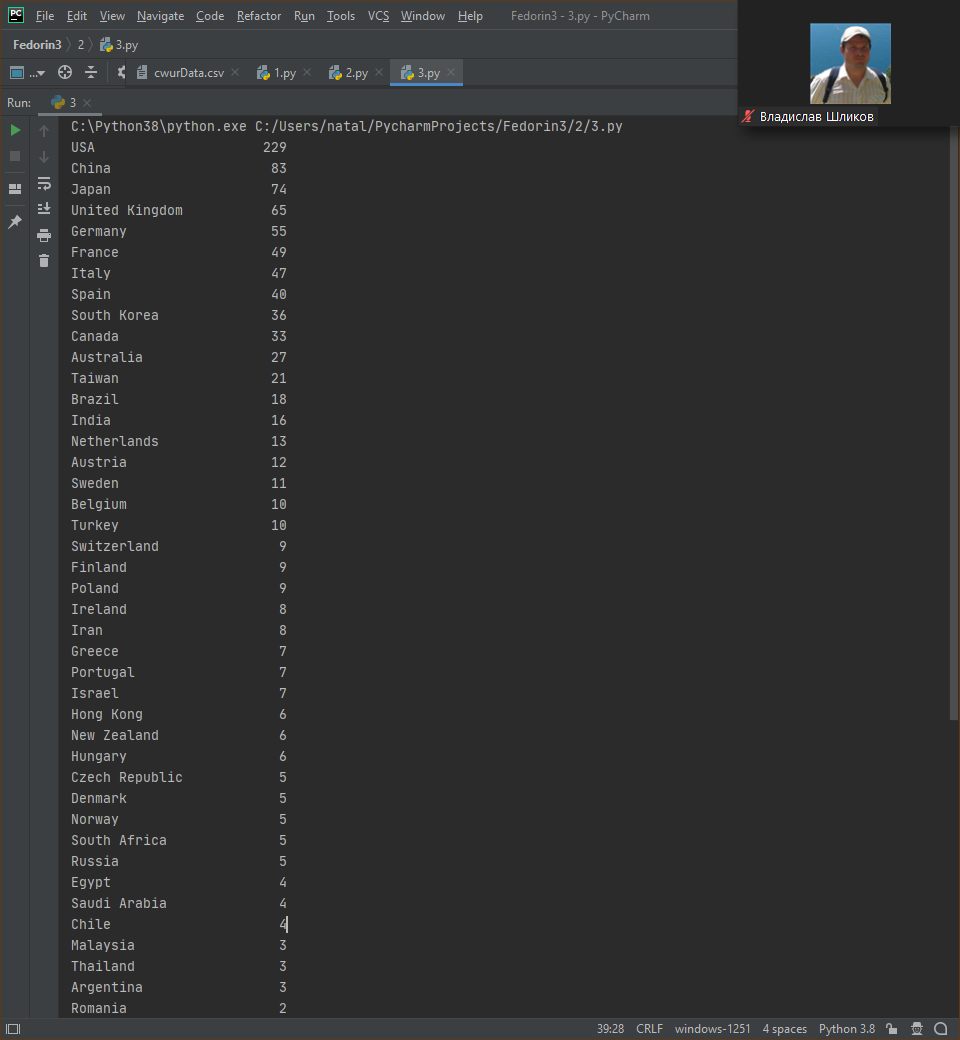
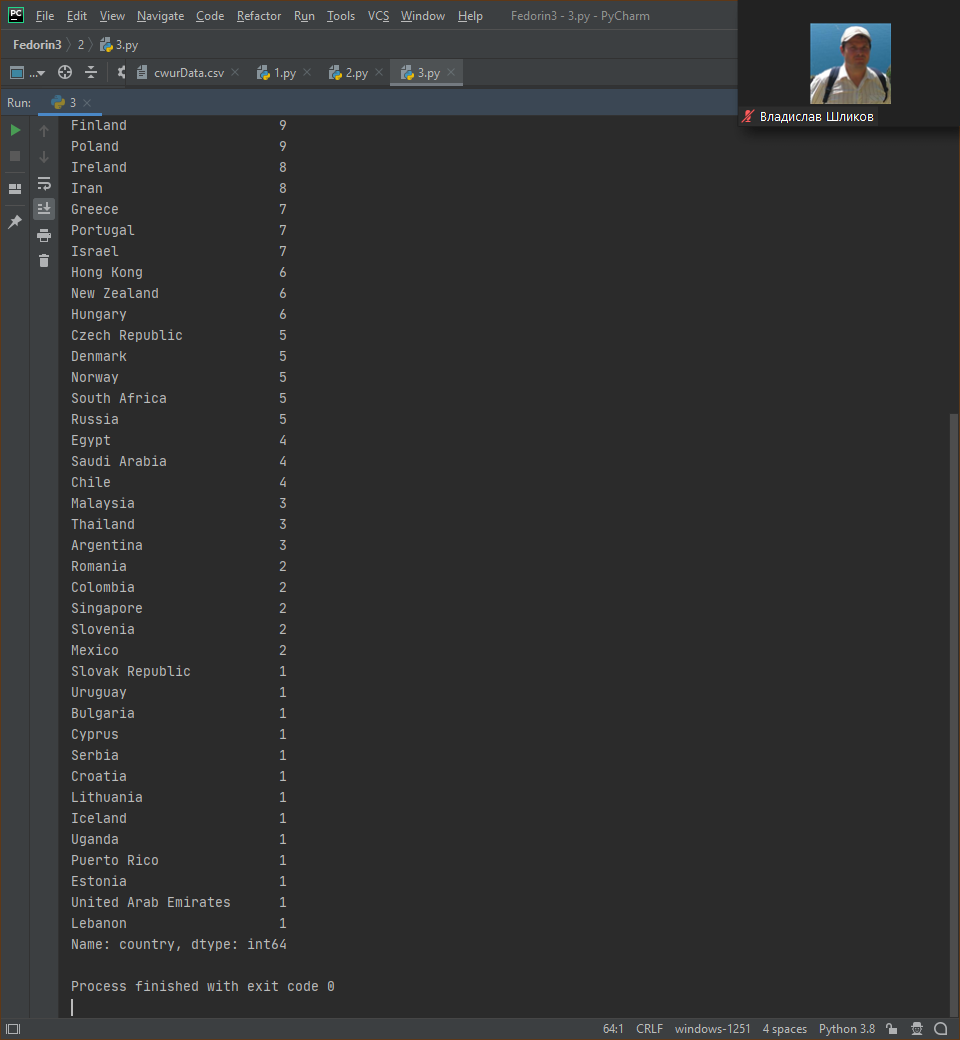


1. Для всіх згаданих країн виведіть кількість університетів, які увійшли до світового рейтингу. Висновок організуйте спаданням кількості університетів.

**Програмна реалізація:**

import pandas as pd  
  
data = pd.read\_csv('cwurData.csv', sep=',', index\_col=0, names=['world\_rank', 'institution', 'country', 'national\_rank', 'quality\_of\_education', 'alumni\_employment', 'quality\_of\_faculty', 'publications', 'influence', 'citations', 'broad\_impact', 'patents', 'score', 'year'])  
df = data.loc[data["year"] == 2015]  
  
print(df["country"].value\_counts())

**Результат:**

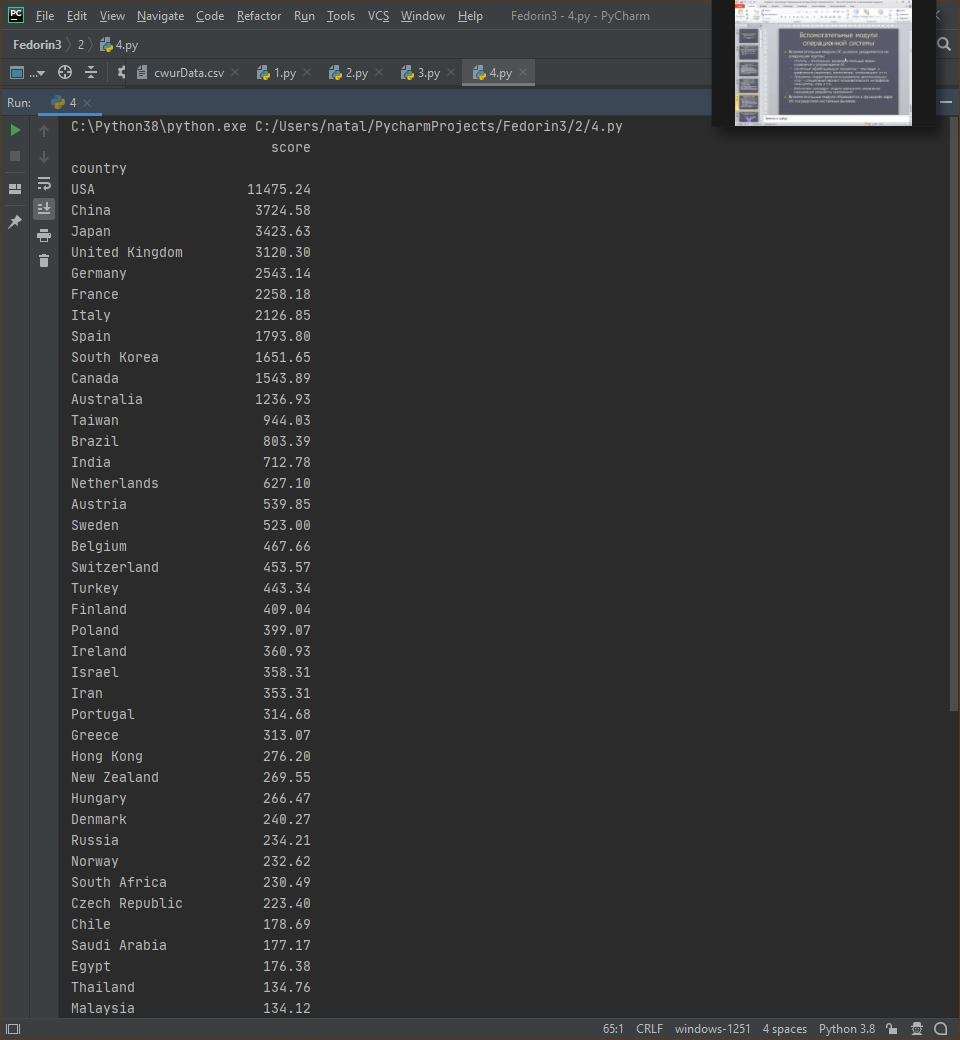
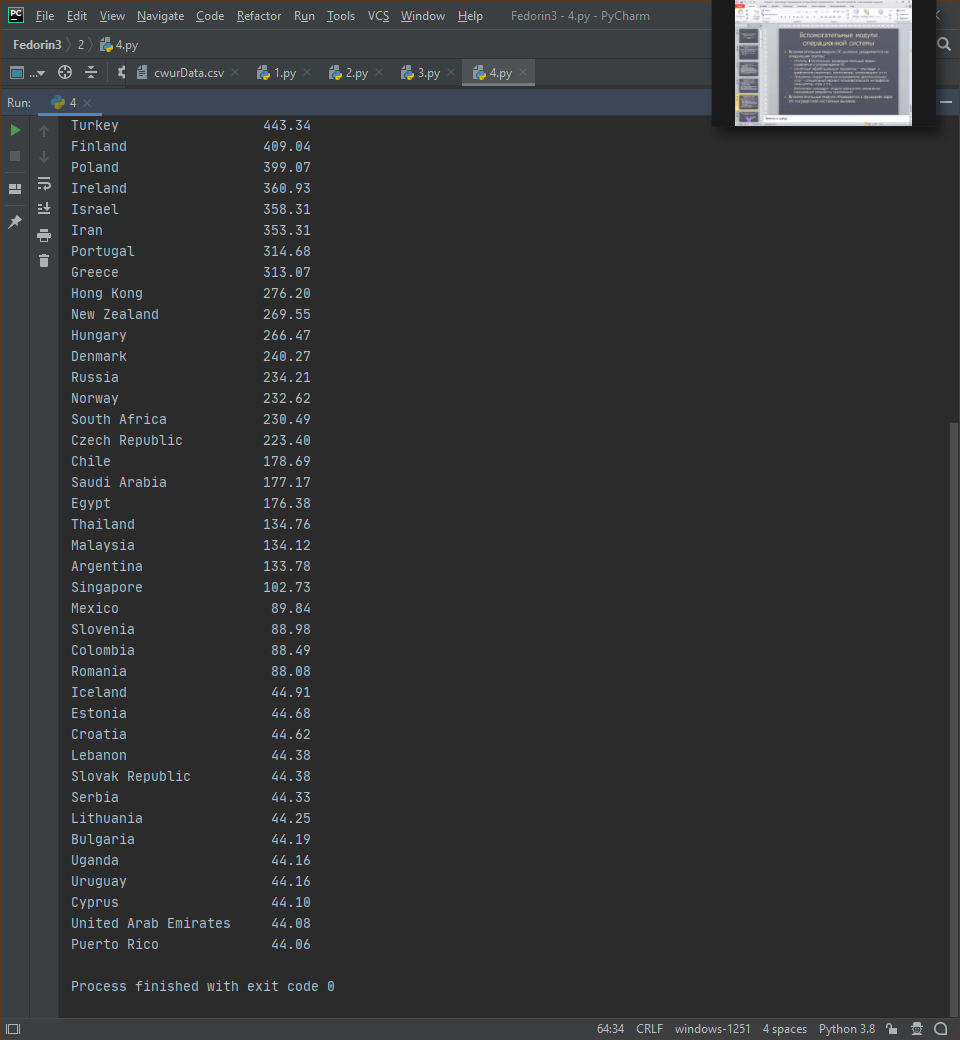
 

1. Обчисліть для кожної країни величину що є сумою загальних балів його університетів. Висновок організуйте по спадаючій цієї величини.

**Програмна реалізація:**

import pandas as pd  
  
data = pd.read\_csv('cwurData.csv', sep=',', index\_col=0, names=['world\_rank', 'institution', 'country', 'national\_rank', 'quality\_of\_education', 'alumni\_employment', 'quality\_of\_faculty', 'publications', 'influence', 'citations', 'broad\_impact', 'patents', 'score', 'year'])  
df = data.loc[data["year"] == 2015]  
  
sum\_score\_country = df.groupby("country").sum()  
sum\_score\_country = sum\_score\_country.sort\_values(by="score", ascending=False)  
print(sum\_score\_country[['score']])

**Результат:**

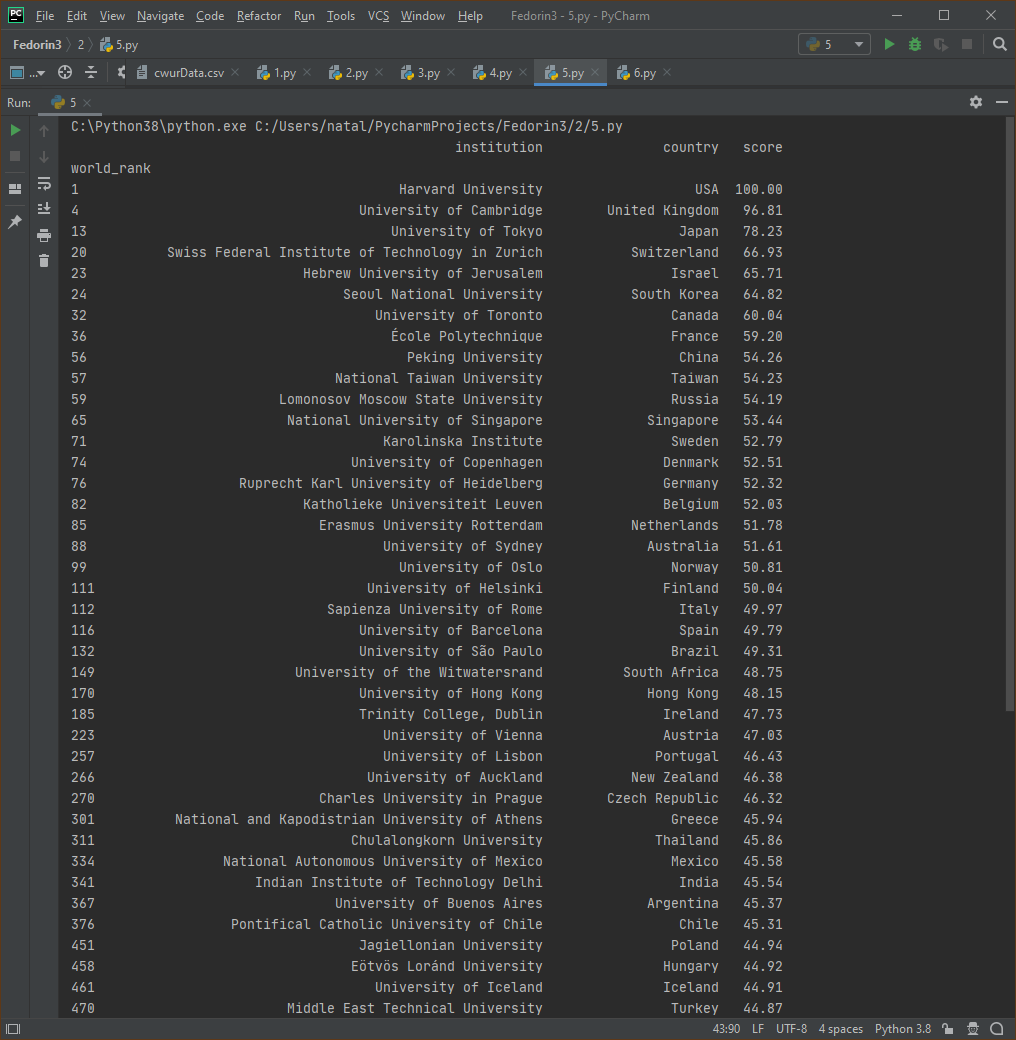
 

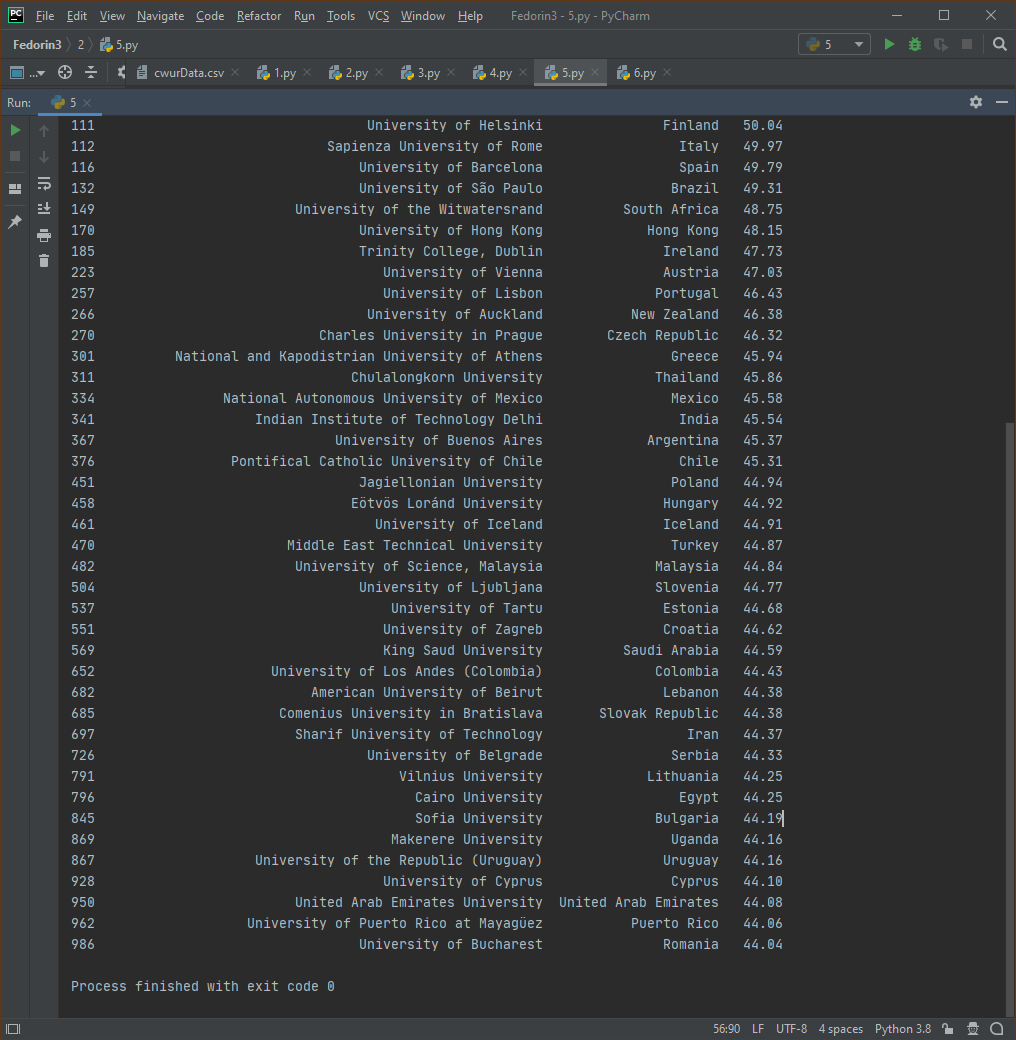
1. Обчисліть кращий університет для кожної країни, висновок організуйте по цій величині.

**Програмна реалізація:**

import pandas as pd  
  
data = pd.read\_csv('cwurData.csv', encoding='utf-8', sep=',', index\_col=0, names=['world\_rank', 'institution', 'country', 'national\_rank', 'quality\_of\_education', 'alumni\_employment', 'quality\_of\_faculty', 'publications', 'influence', 'citations', 'broad\_impact', 'patents', 'score', 'year'])  
df = data.loc[data["year"] == 2015]  
  
max\_score\_country = df.loc[df.groupby(["country"])['score'].idxmax()]  
max\_score\_country = max\_score\_country.sort\_values(by=["score"], ascending=False)  
pd.set\_option('display.width', None)  
print(max\_score\_country[['institution', 'country', 'score']])

**Результат:**



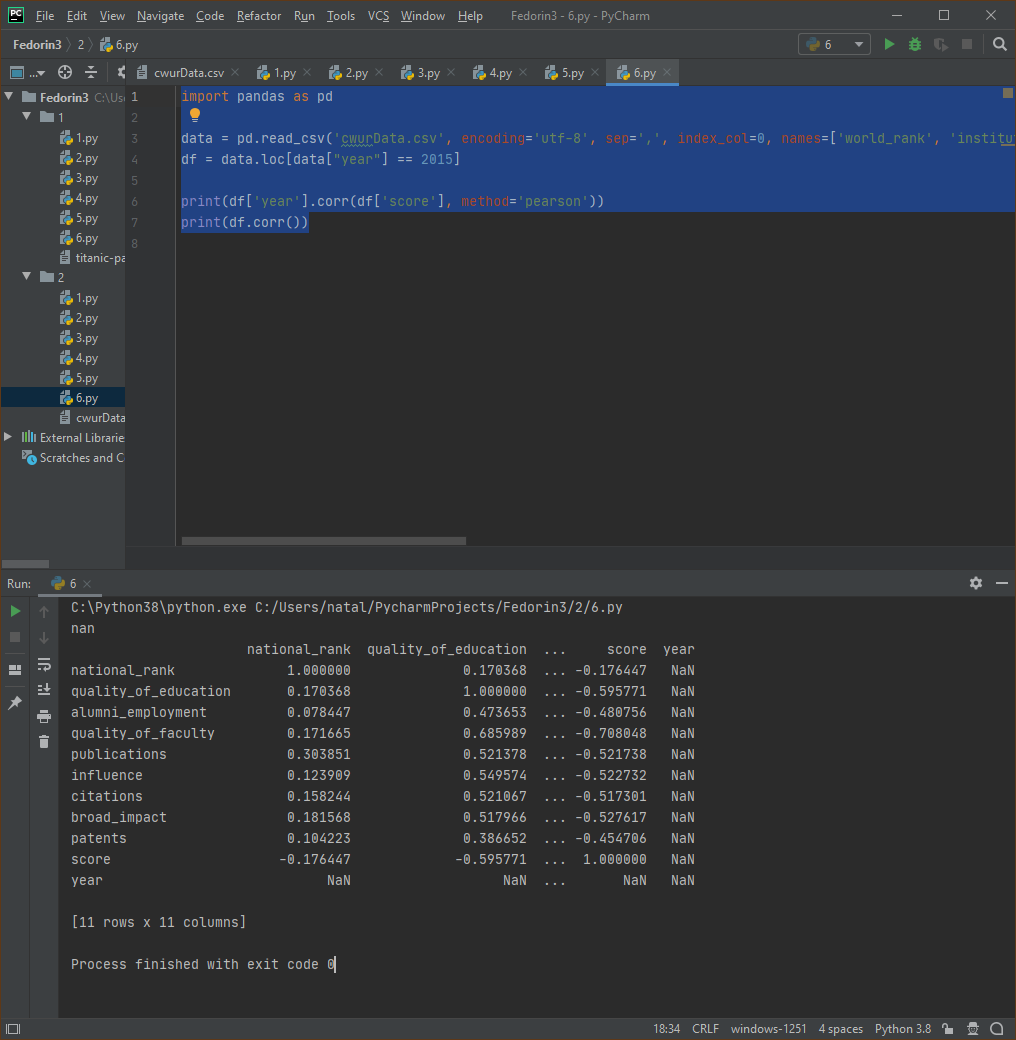


1. Зробіть висновок про те, чи є кореляція у двох останніх величин.

**Програмна реалізація:**

import pandas as pd  
  
data = pd.read\_csv('cwurData.csv', encoding='utf-8', sep=',', index\_col=0, names=['world\_rank', 'institution', 'country', 'national\_rank', 'quality\_of\_education', 'alumni\_employment', 'quality\_of\_faculty', 'publications', 'influence', 'citations', 'broad\_impact', 'patents', 'score', 'year'])  
df = data.loc[data["year"] == 2015]  
  
print(df['year'].corr(df['score'], method='pearson'))  
print(df.corr())

**Результат:**



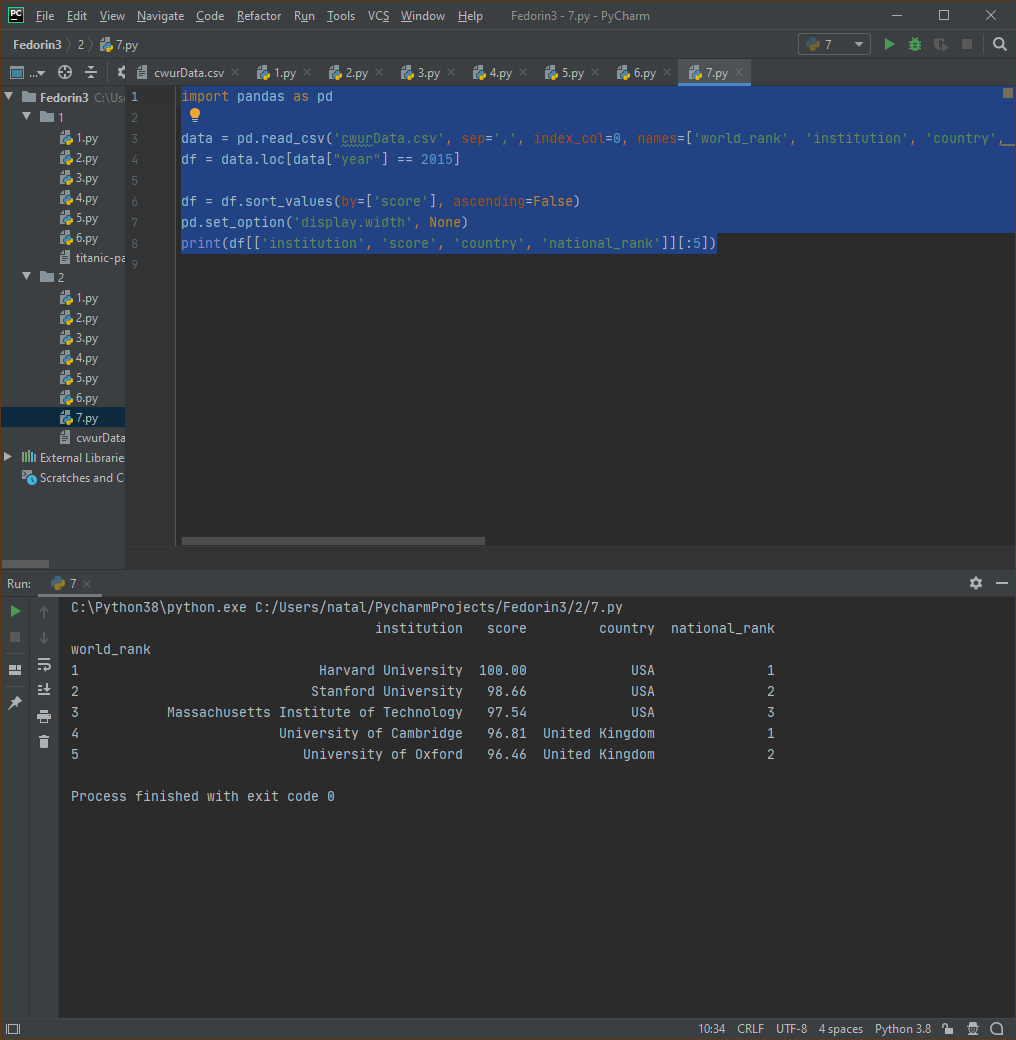
**Кореляція існує між усіма величинами окрім year, тому що це поле у нас незмінне.**

1. Обчисліть п'ять університетів з найбільшим рейтингом, виведіть їх назви, загальний бал, країни, рейтинги в своїй країні.

**Програмна реалізація:**

import pandas as pd  
  
data = pd.read\_csv('cwurData.csv', sep=',', index\_col=0, names=['world\_rank', 'institution', 'country', 'national\_rank', 'quality\_of\_education', 'alumni\_employment', 'quality\_of\_faculty', 'publications', 'influence', 'citations', 'broad\_impact', 'patents', 'score', 'year'])  
df = data.loc[data["year"] == 2015]  
  
df = df.sort\_values(by=['score'], ascending=False)  
pd.set\_option('display.width', None)  
print(df[['institution', 'score', 'country', 'national\_rank']][:5])

**Результат:**

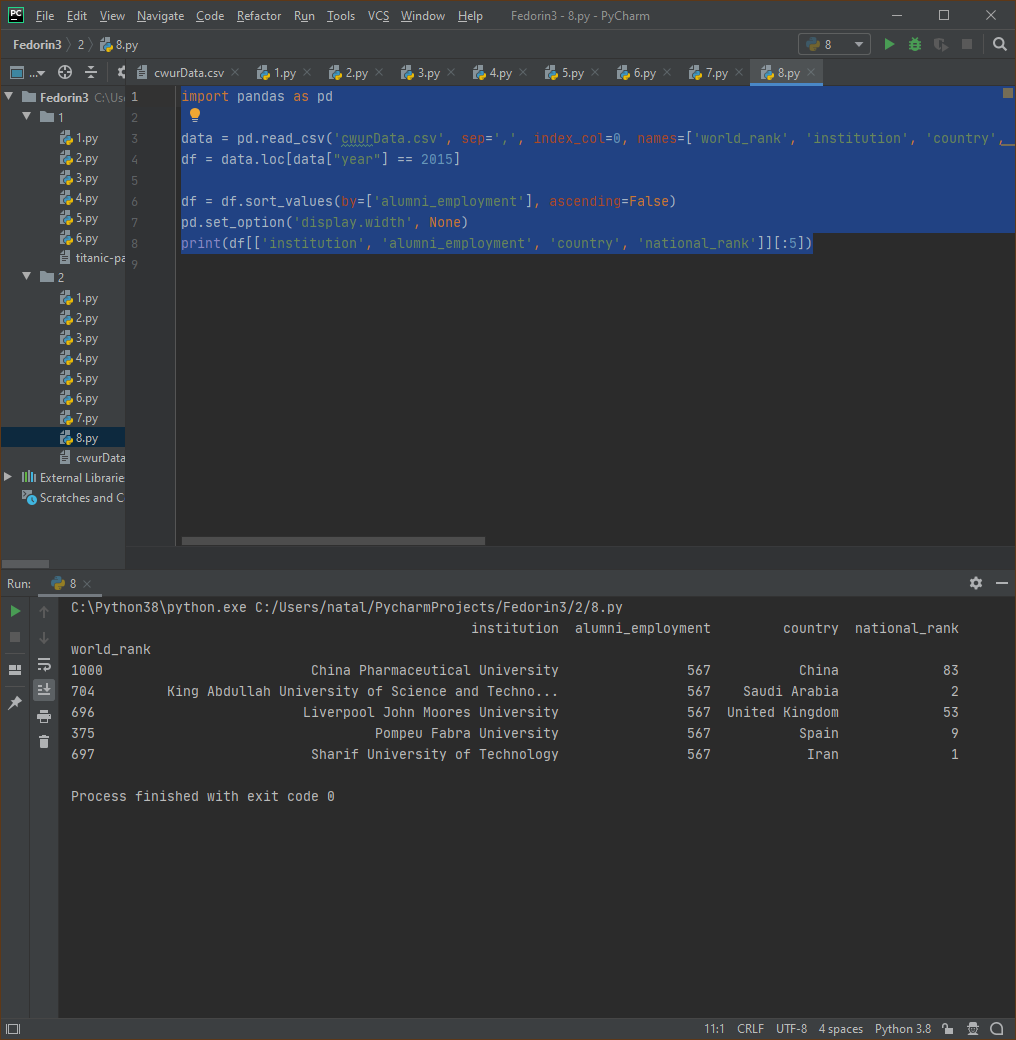


1. Випускники яких 5 університетів володіють найбільшим робочим рейтингом?

**Програмна реалізація:**

import pandas as pd  
  
data = pd.read\_csv('cwurData.csv', sep=',', index\_col=0, names=['world\_rank', 'institution', 'country', 'national\_rank', 'quality\_of\_education', 'alumni\_employment', 'quality\_of\_faculty', 'publications', 'influence', 'citations', 'broad\_impact', 'patents', 'score', 'year'])  
df = data.loc[data["year"] == 2015]  
  
df = df.sort\_values(by=['alumni\_employment'], ascending=False)  
pd.set\_option('display.width', None)  
print(df[['institution', 'alumni\_employment', 'country', 'national\_rank']].head(5))

**Результат:**

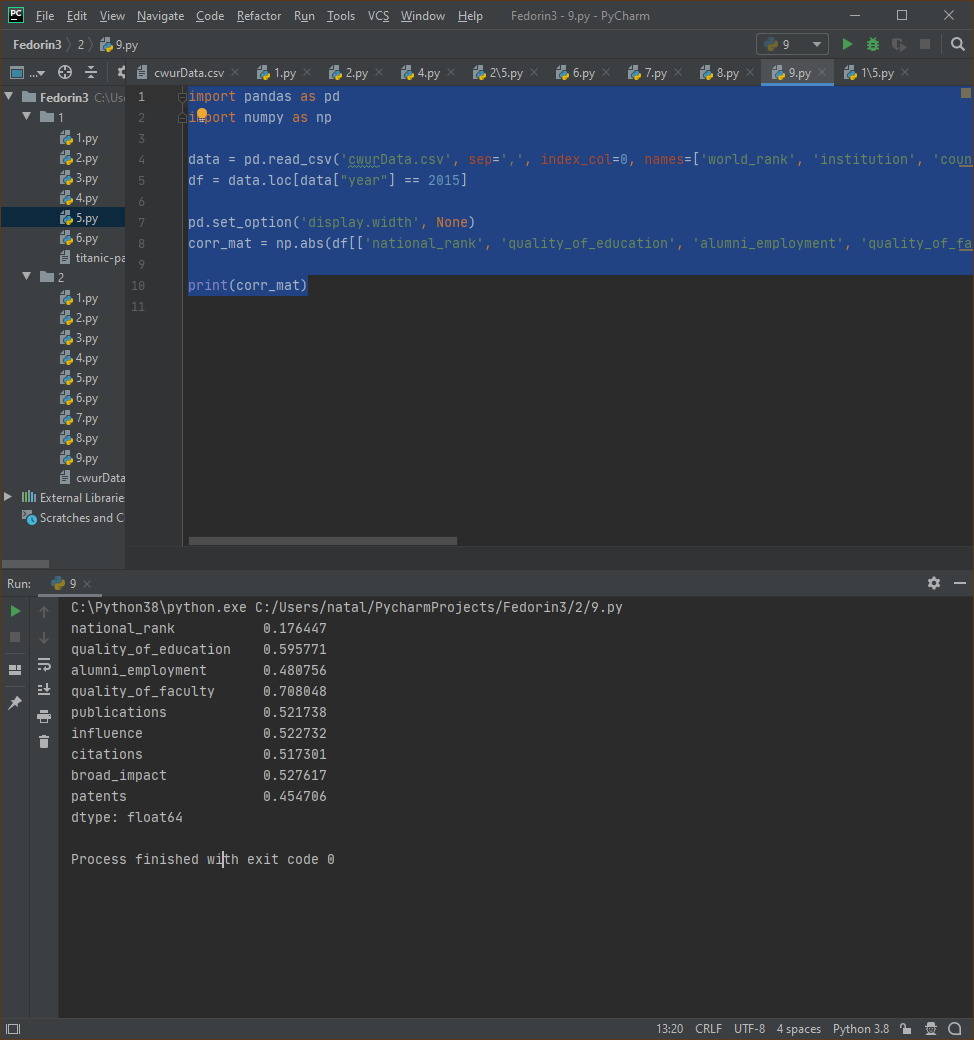


1. Як оцінити вплив окремих рейтингів на загальний. Який із зазначених рейтингів дає найбільший внесок у загальний рейтинг?

**Програмна реалізація:**

import pandas as pd  
import numpy as np  
  
data = pd.read\_csv('cwurData.csv', sep=',', index\_col=0, names=['world\_rank', 'institution', 'country', 'national\_rank', 'quality\_of\_education', 'alumni\_employment', 'quality\_of\_faculty', 'publications', 'influence', 'citations', 'broad\_impact', 'patents', 'score', 'year'])  
df = data.loc[data["year"] == 2015]  
  
pd.set\_option('display.width', None)  
corr\_mat = np.abs(df[['national\_rank', 'quality\_of\_education', 'alumni\_employment', 'quality\_of\_faculty', 'publications', 'influence', 'citations', 'broad\_impact', 'patents']].corrwith(df['score']))  
  
print(corr\_mat)

**Результат:**



**Як ми можемо побачити найбільший вплив на загальній рейтинг має Оцінка якості факультету, тому що це поле має найбільшу кореляцію.**