Міністерство освіти і науки України

Національний технічний університет України «Київський політехнічний

інститут імені Ігоря Сікорського"

Факультет інформатики та обчислювальної техніки

Катедра інформатики та програмної інженерії

Звіт

з комп’ютерного практикуму № 3 з дисципліни

«Аналіз даних в інформаційних системах»

«Описова статистика»

Виконав студент ІП-11 Лесів Владислав Ігорович

Перевірила Ліхоузова Тетяна Анатоліївна

Київ 2023

**Комп’ютерний практикум 3**

**Описова статистика**

**Мета роботи:** ознайомитись з методикою первинної обробки статистичних даних; проаналізувати вплив способу представлення даних на їх інформативність.

**Основне завдання**.

Скачати дані із файлу Data2.csv.

1. Записати дані у data frame
2. Дослідити структуру даних
3. Виправити помилки в даних
4. Побудувати діаграми розмаху та гістограми
5. Додати стовпчик із щільністю населення

**Додаткове завдання.**

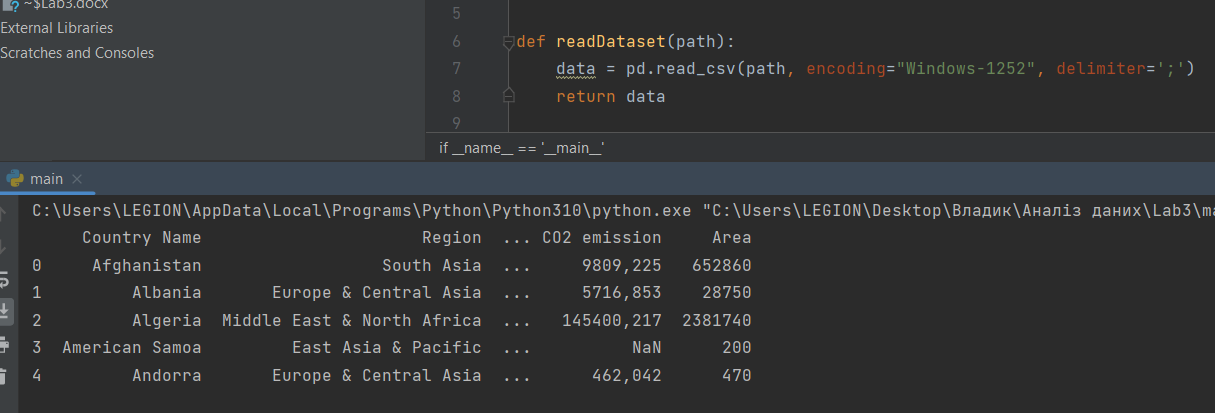
Відповісти на питання (файл Data2.csv):

1. Чи є пропущені значення? Якщо є, замінити середніми
2. Яка країна має найбільший ВВП на людину (GDP per capita)? Яка має найменшу площу?
3. В якому регіоні середня площа країни найбільша?
4. Знайдіть країну з найбільшою щільністю населення у світі? У Європі та центральній Азії?
5. Чи співпадає в якомусь регіоні середнє та медіана ВВП?
6. Вивести топ 5 країн та 5 останніх країн по ВВП та кількості СО2 на душу населення.

**Хід роботи.**

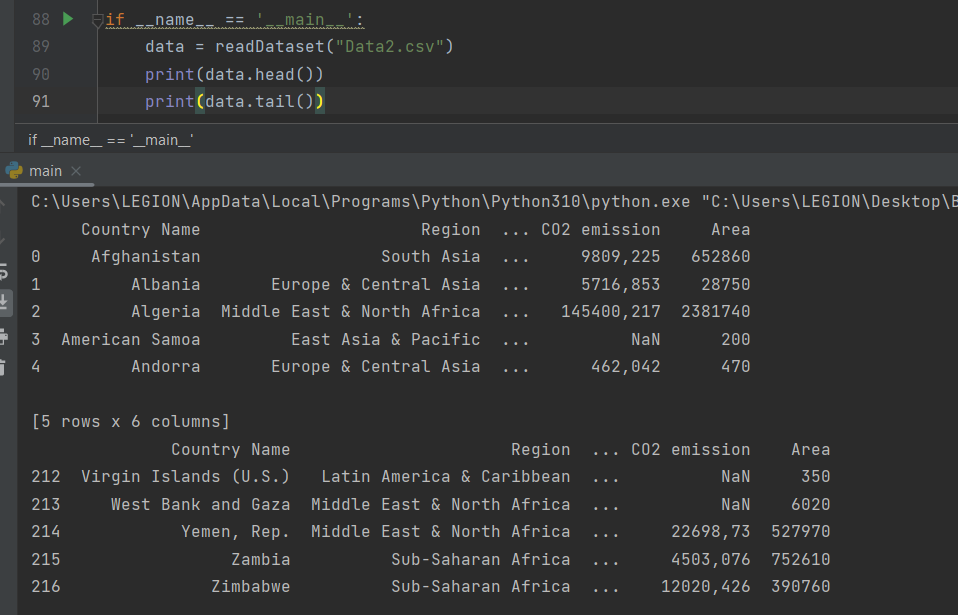
**Основне завдання.**

1. Записую дані у DataFrame за допомогою методу read\_csv.



Дані успішно записані.

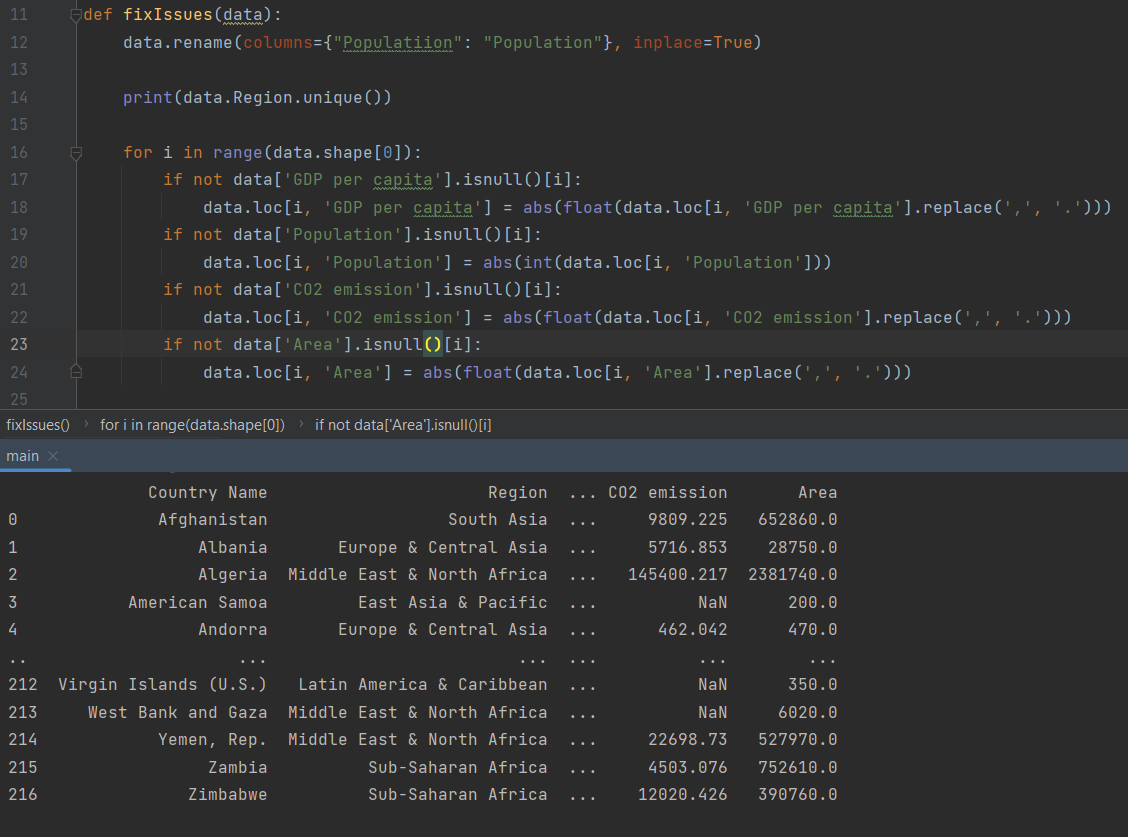
1. Досліджую структуру даних.



Бачимо, що отримали DataFrame, назви стовпців якого задані з csv-файлу, а назви рядків – пронумеровані рядки по порядку. Маємо стовпці Country Name, Region, GDP per capita, Populatiion, CO2 emission, Area.

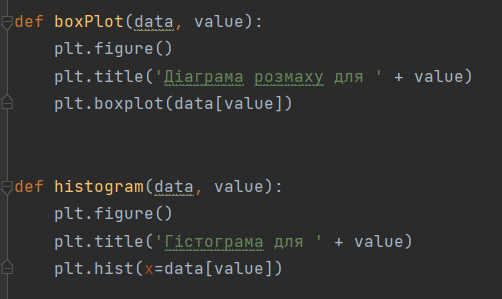
1. Маємо такі проблеми даних. Назва стовпця Populatiion написана з помилкою, необхідно виправити на Population. Значення GDP per capita, Population, CO2 emission, Area мають на увазі числові дані типу float, але написані у неправильному форматі через кому, а не через крапку. Необхідно замінити ці коми на крапки і перетворити значення у числовий формат. Також деякі значення є від’ємними, що неможливо для досліджуваних даних. Необхідно взяти модуль від значень.

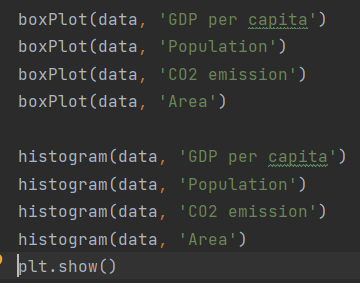
Ще одна проблема – відсутні деякі значення. Їх після всіх перетворень з існуючими значеннями ми заміняємо середніми від відповідного стовпця.



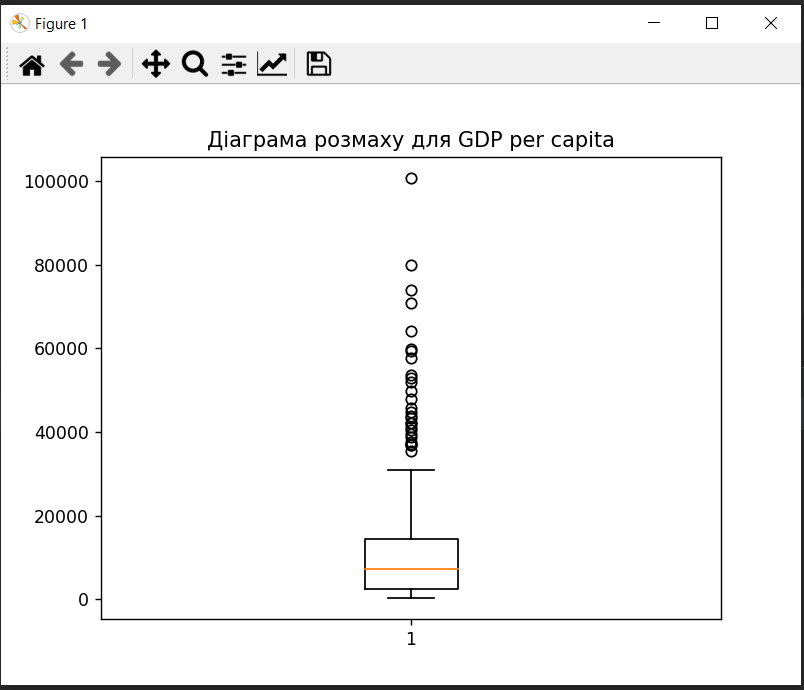
1. Будую діаграми розмаху та гістограми.

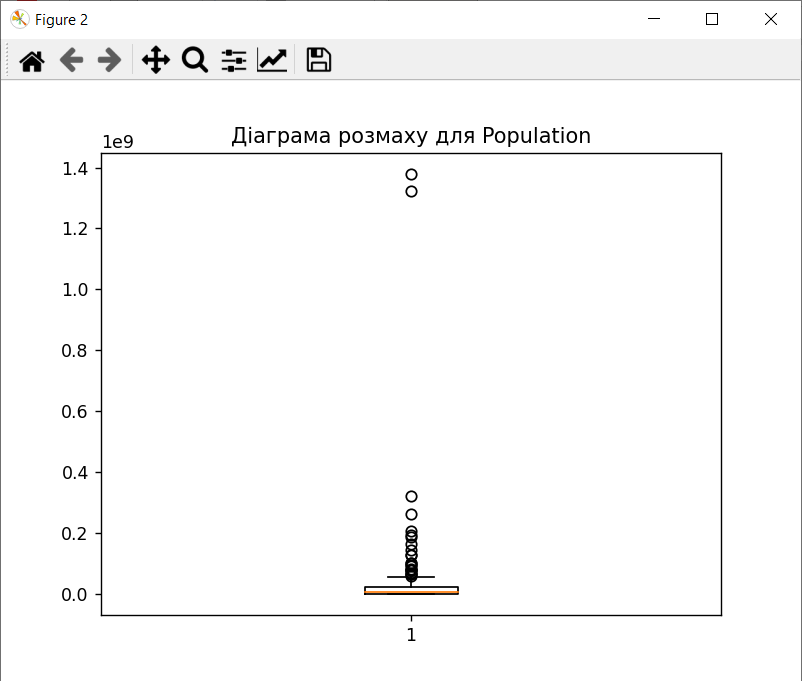
Для побудови використовую matplotlib.pyplot.boxplot() та matplotlib.pyplot.hist().

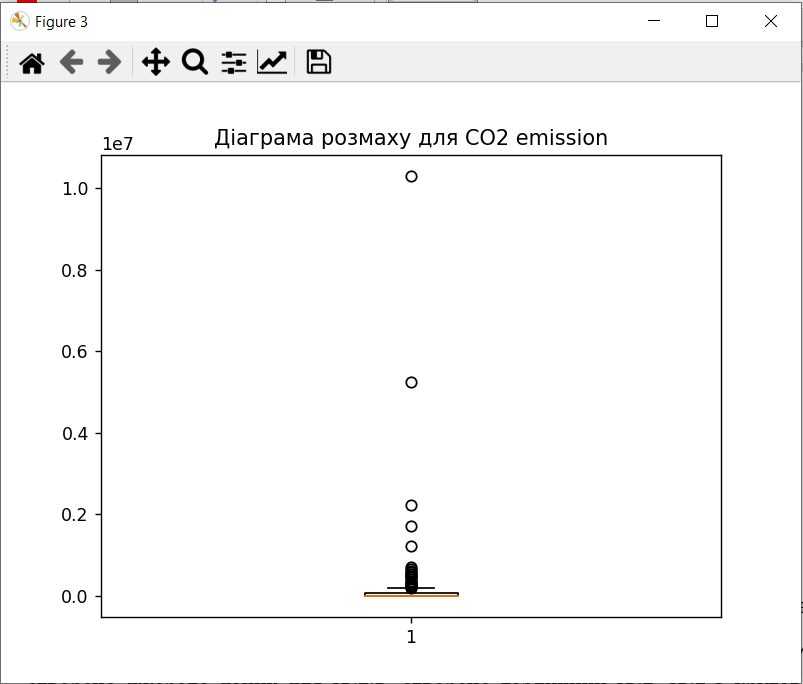


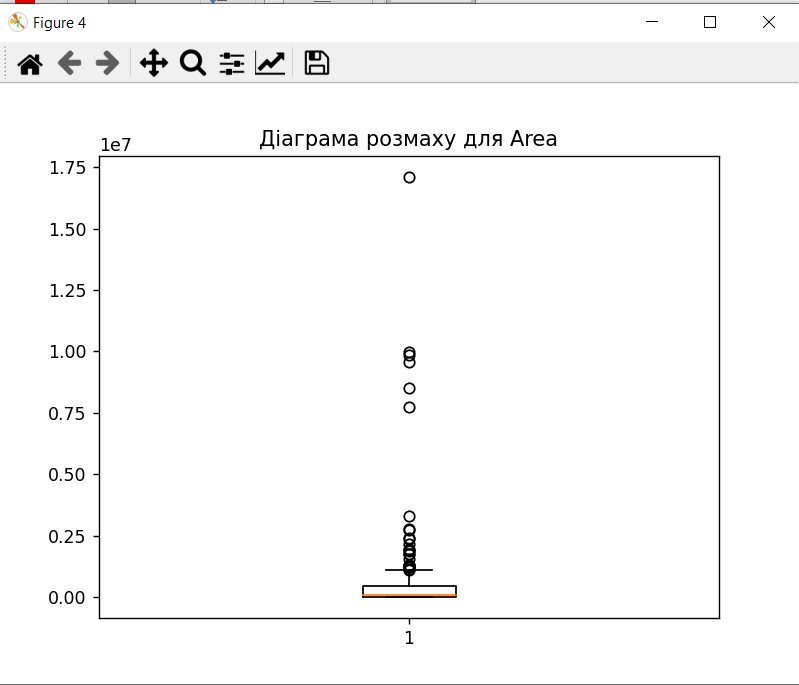


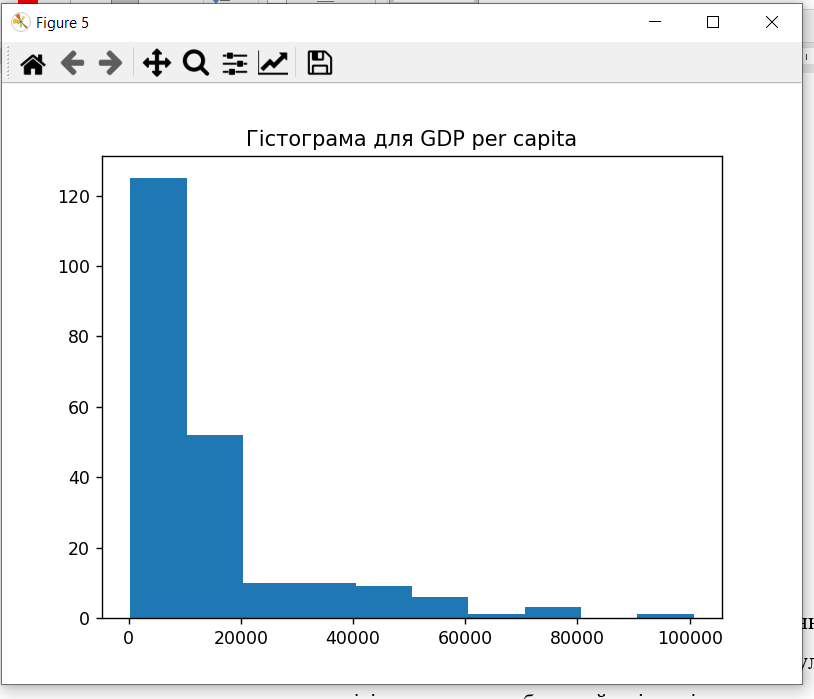
Отримуємо такі діаграми.

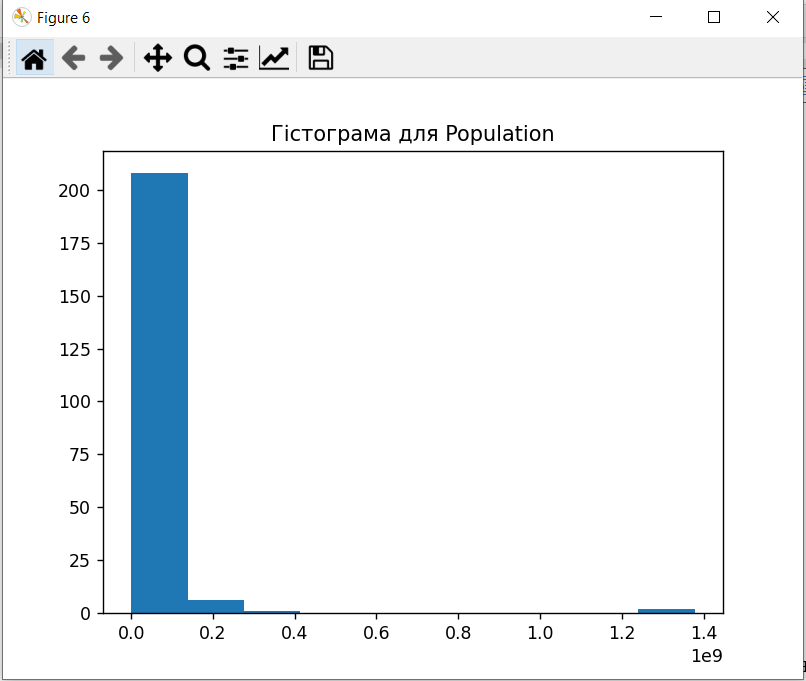


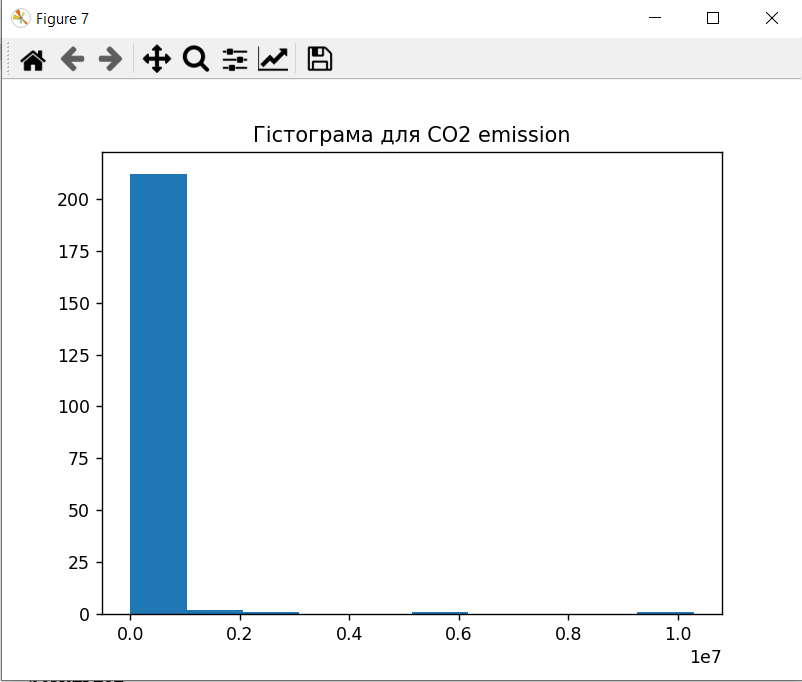


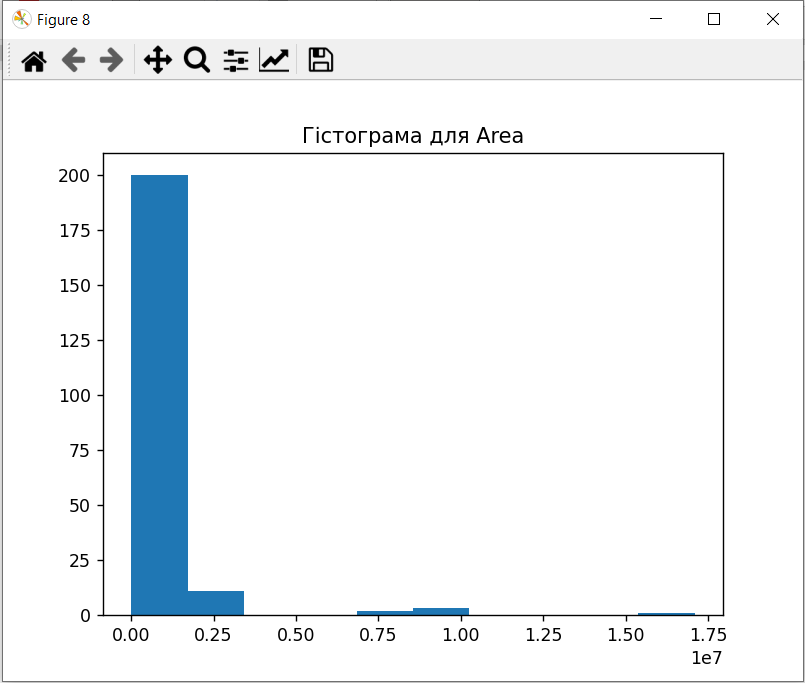




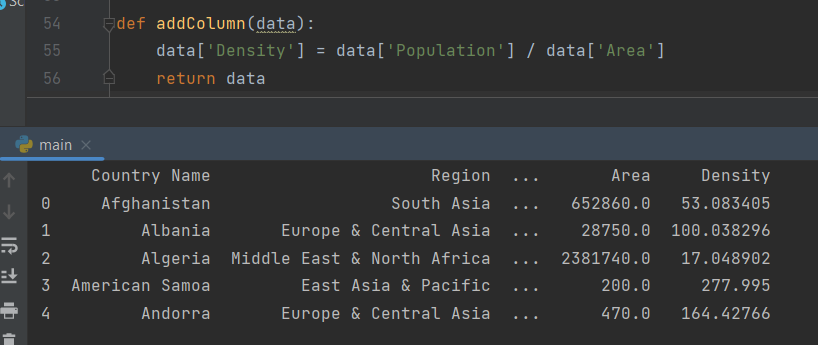






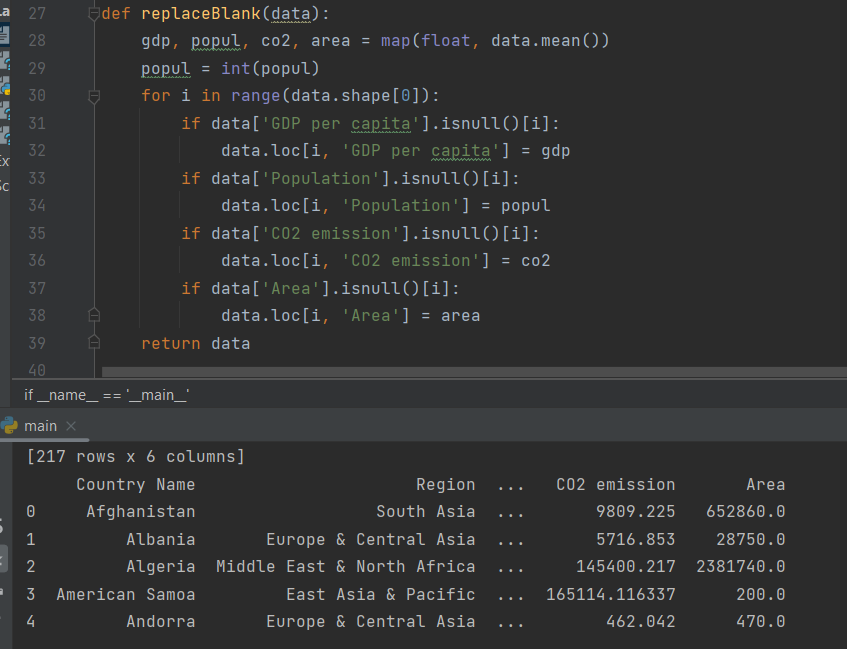


1. Додаю стовпчик із щільністю населення. Отримуємо відповідний останній стовпчик з розрахованими за формулою значеннями.

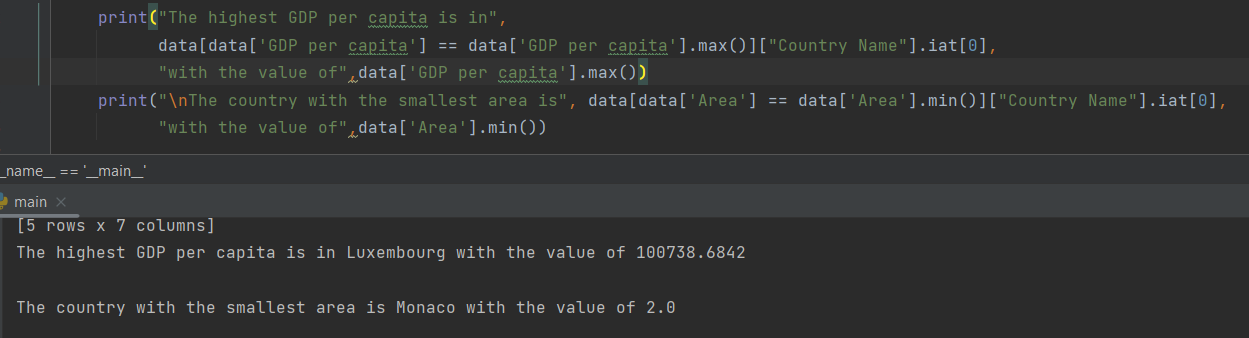


**Додаткове завдання.**

1. Як я написав вище, пропущені значення є, їх ми заміняємо середніми значеннями відповідних стовпців.



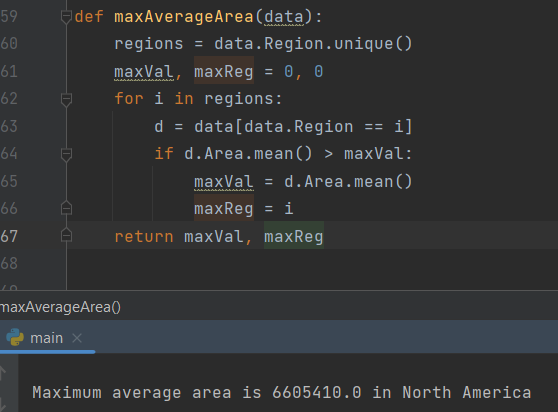
1. Яка країна має найбільший ВВП на людину (GDP per capita)? Яка має найменшу площу?



The highest GDP per capita is in Luxembourg with the value of 100738.6842

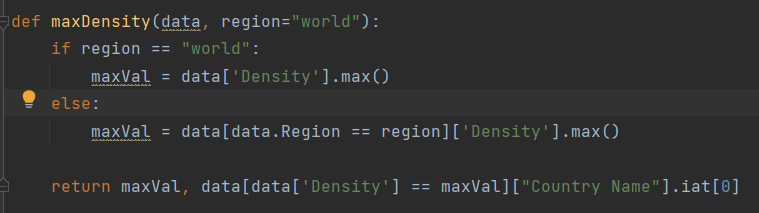
The country with the smallest area is Monaco with the value of 2.0

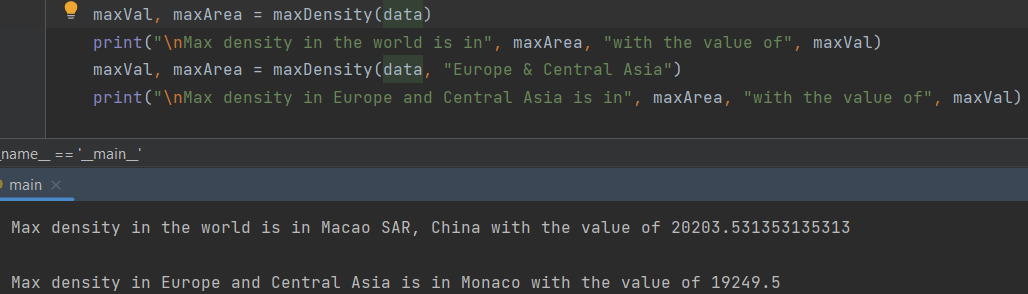
1. В якому регіоні середня площа країни найбільша?



Maximum average area is 6605410.0 in North America.

1. Знайдіть країну з найбільшою щільністю населення у світі? У Європі та центральній Азії?

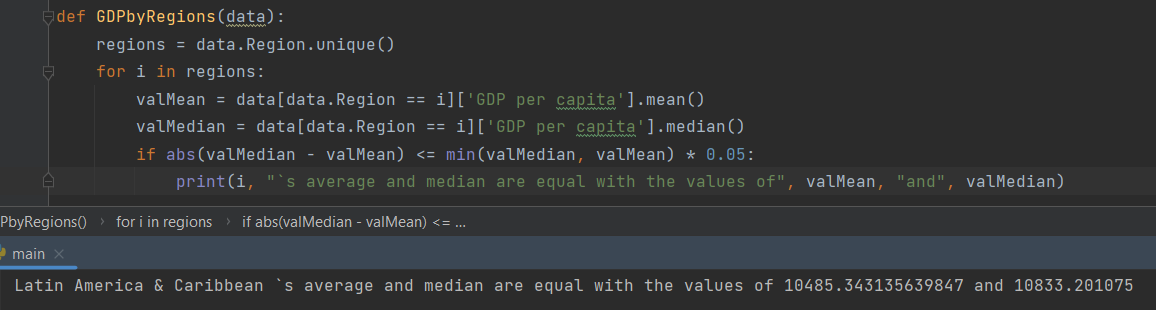




Max density in the world is in Macao SAR, China with the value of 20203.531353135313.

Max density in Europe and Central Asia is in Monaco with the value of 19249.5.

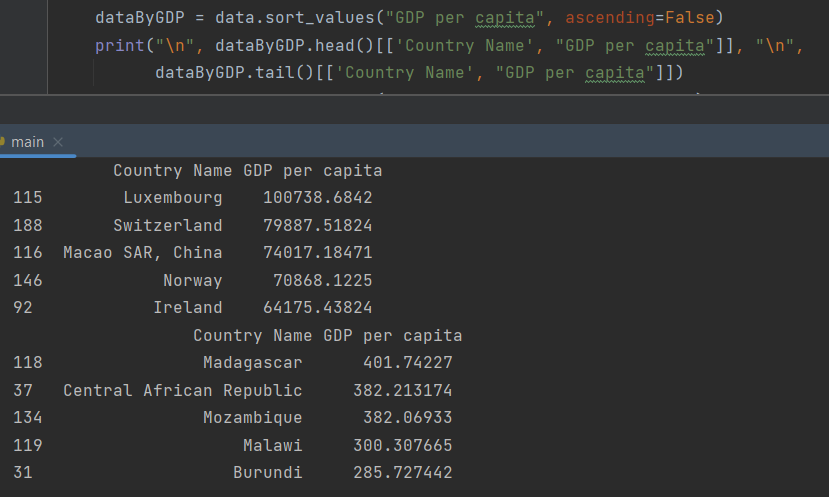
1. Чи співпадає в якомусь регіоні середнє та медіана ВВП? Як «співпадає» я позначив те, що різниця між значеннями становить не більше 5% від меншого значення.



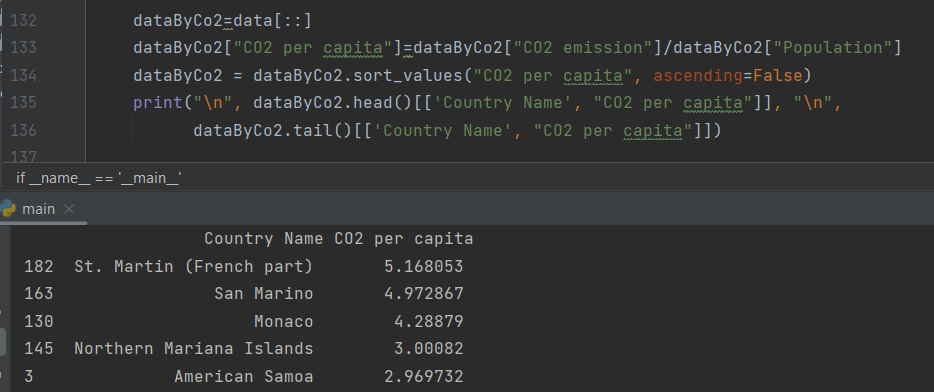
Latin America & Caribbean `s average and median are equal with the values of 10485.343135639847 and 10833.201075.

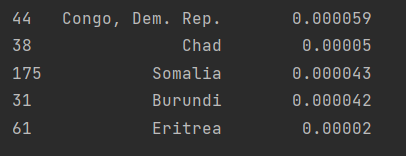
1. Вивести топ 5 країн та 5 останніх країн по ВВП та кількості СО2 на душу населення.

Топ країн за ВВП:



Топ країн за кількістю CO2 per capita:





**Висновок.**

Отже, у цій роботі я ознайомився з методикою первинної обробки статистичних даних; проаналізував вплив способу представлення даних на їх інформативність. У результаті лабораторної роботи було створено DataFrame для поданих даних, за допомогою якого я зміг зручно провести дослідження даних, представити візуалізацію, порахувати статистики, необхідні для аналізу. Я переконався, що використання DataFrame модулю pandas є доволі зручним і ефективним для дослідження поданих даних. Після досідження можемо побачити, що найбільше ВВП на душу населення в країнах переважно Європи (за винятком Макао), а найменше – у бідних країнах Африки. Щодо викидів СО2 на душу населення – найбільші викиди на островах та маленьких розвинених країнах, як Монако та Сан-Марино, а найменші – знову ж у бідних країнах Африки. Використовуючи PyCharm, отримуємо коректний результат.