**Міністерство освіти і науки України**

**Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського"**

**Факультет інформатики та обчислювальної техніки**

**Кафедра інформатики та програмної інженерії**

**Звіт**

з лабораторної роботи №3 з дисципліни

«Аналіз даних в інформаційних системах»

„**Описова статистика**”

**Виконав(ла)**

(шифр, прізвище, ім'я, по батькові)

*ІП-11 Тарасьонок Дмитро Євгенович*

**Перевірила**

(прізвище, ім'я, по батькові)

*Ліхоузова Т. А.*

Київ 2023

ЗМІСТ

[1 Мета лабораторної роботи 3](#_Toc130513001)

[2 Завдання 4](#_Toc130513002)

[2.1 Основне завдання 4](#_Toc130513003)

[2.2 Додаткове завдання 4](#_Toc130513004)

[3 Виконання основного завдання 5](#_Toc130513005)

[3.1 Записати дані у data frame 5](#_Toc130513006)

[3.2 Дослідити структуру даних 5](#_Toc130513007)

[3.3 Виправити помилки в даних 5](#_Toc130513008)

[3.4 Побудувати діаграми розмаху та гістограми 7](#_Toc130513009)

[3.5 Додати стовпчик із щільністю населення 9](#_Toc130513010)

[4 Виконання додаткового завдання 10](#_Toc130513011)

[4.1 Чи є пропущені значення? Якщо є, замінити середніми 10](#_Toc130513012)

[4.2 Яка країна має найбільший ВВП на людину (GDP per capita)? Яка має найменшу площу? 12](#_Toc130513013)

[4.3 В якому регіоні середня площа країни найбільша? 13](#_Toc130513014)

[4.4 Знайдіть країну з найбільшою щільністю населення у світі? У Європі та центральній Азії? 13](#_Toc130513015)

[4.5 Чи співпадає в якомусь регіоні середнє та медіана ВВП? 14](#_Toc130513016)

[4.6 Вивести топ 5 країн та 5 останніх країн по ВВП та кількості СО2 на душу населення. 16](#_Toc130513017)

[5 Висновок 19](#_Toc130513018)

# Мета лабораторної роботи

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

# Завдання

## 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.

# Виконання основного завдання

## Записати дані у data frame

Зробити дану операцію можна за допомогою функції read\_csv пакету pandas. На рисунку 3.1 наведено завантажений data frame.

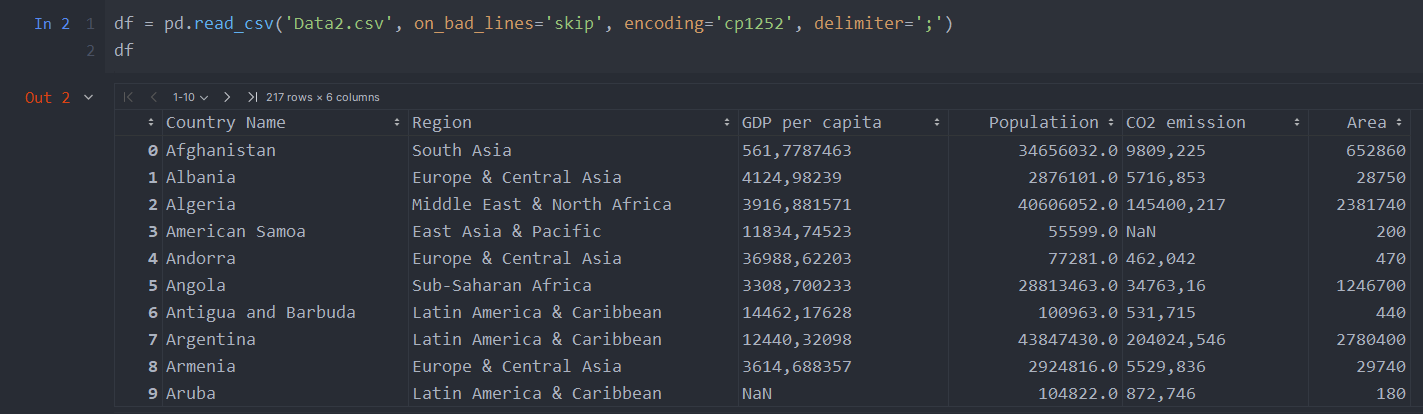


Рис. 3.1 – Data Frame

## Дослідити структуру даних

Було детально досліджено структуру даних: data frame містить 6 стовпців: назву країни, регіон, ВВП на душу населення, кількість населення, кількість викидів CO2, а також площу країни. Було помічено, що дійсні числа записані через роздільник «кома», що сприймається пакетом як рядок, було виявлено некоректно названий стовпець населення: «Populatiion» містить дві літери «i». Також помічено, що стовпці ВВП на душу населення й площі мають від’ємні значення, які, очевидно, бути такими не можуть.

## Виправити помилки в даних

Спочатку було перейменовано стовпці: записано їх у форматі snake\_case для зручності доступу до них як до атрибутів об’єкту, а також виправлено помилку в написанні назви стовпця кількості населення. На рисунку 3.2 наведено код та результат.

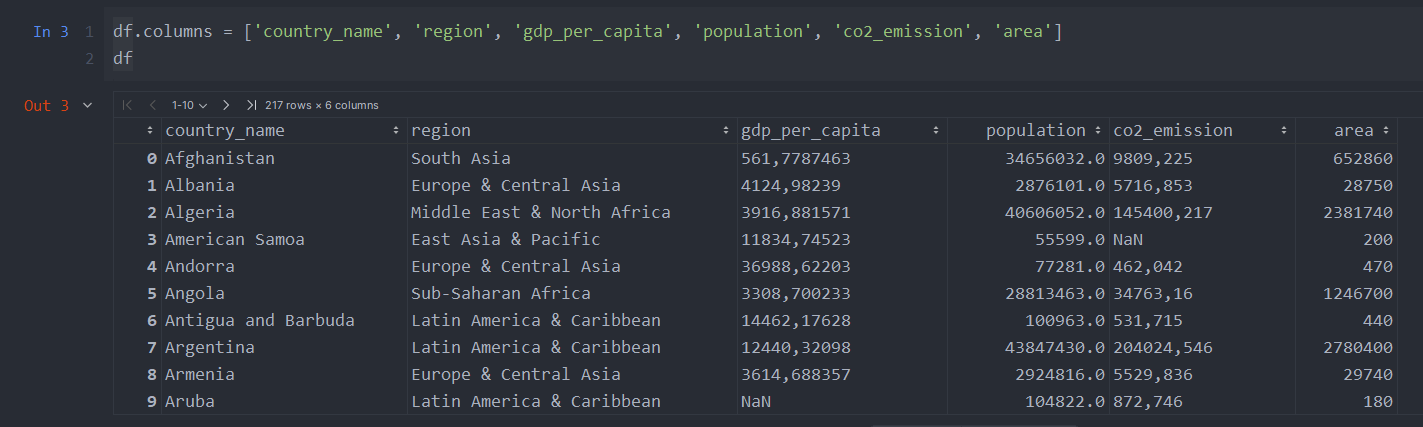


Рис. 3.2 – Перейменування стовпців

Далі було встановлено назву країни в якості індексу, оскільки пакет pandas за замовчуванням в якості індексу встановлює номер рядку.

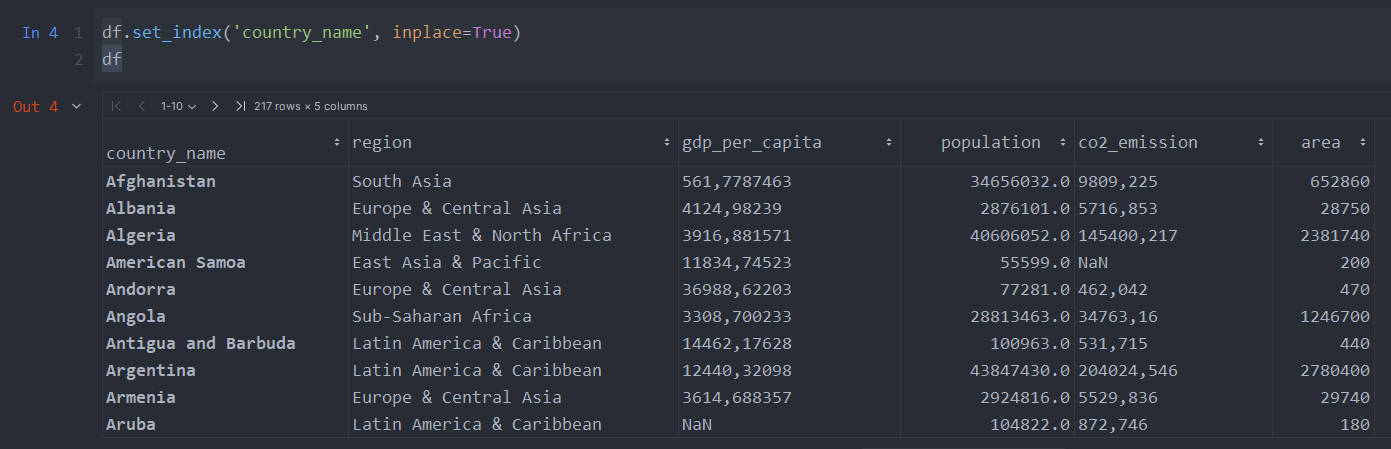


Рис. 3.3 – Встановлення індексу

Після цього було замінено коми в значеннях на крапки й приведено до дійсночисельного типу.

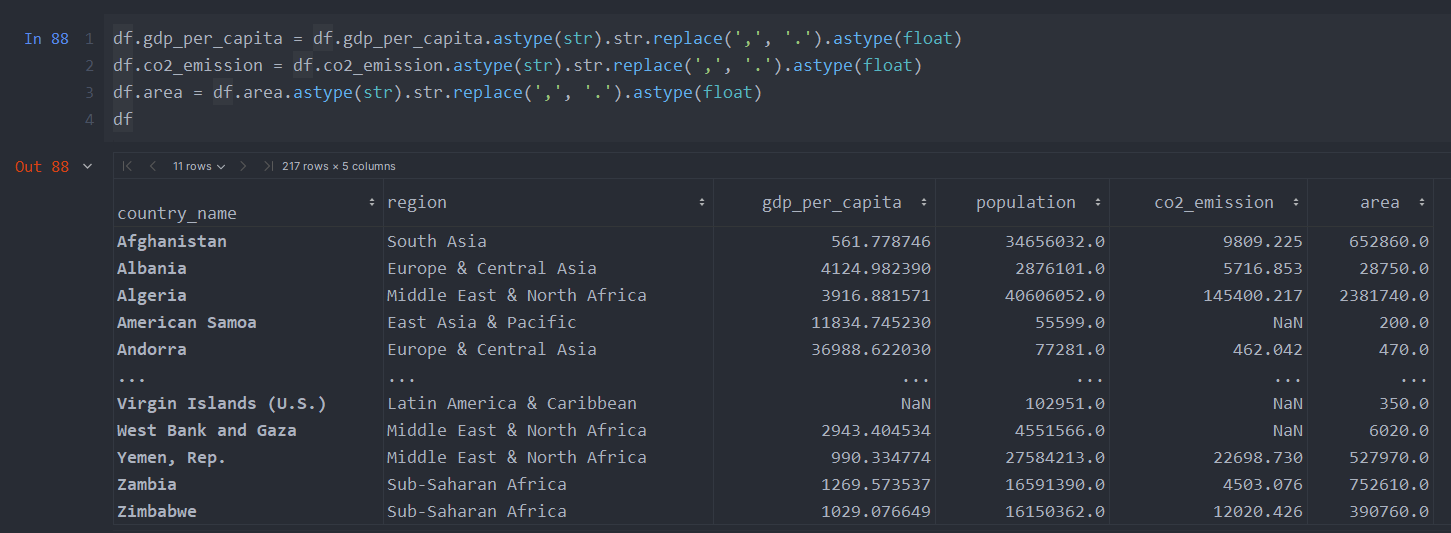


Рис. 3.4 – Виправлення дійсних чисел

Можемо перевірити типи стовпців за допомогою виклику методу «dtypes».

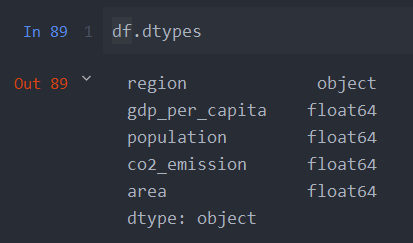


Рис. 3.5 – Виведення типів даних кожного стовпця

Після цього від’ємні значення було перетворено на додатні шляхом виклику методу модуля

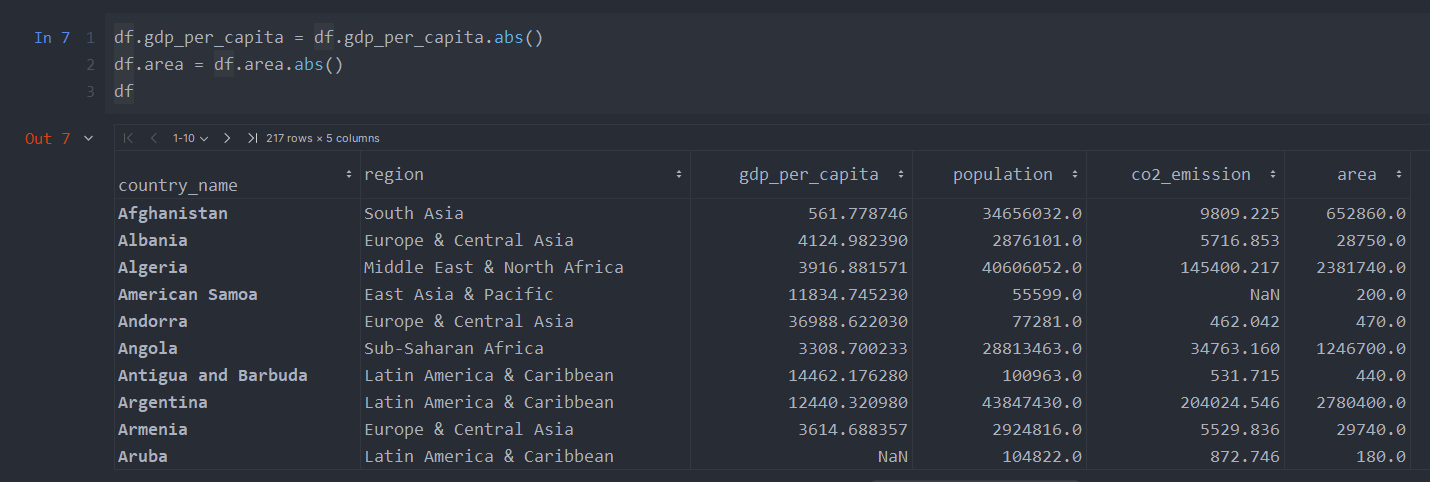


Рис. 3.6 – Заміна від’ємних значень ВВП на душу населення та площі на значення по модулю

## Побудувати діаграми розмаху та гістограми

Побудувати діаграми об’єктів класу DataFrame пакету pandas можна дуже легко: треба просто викликати метод об’єкту. Для побудови гістограми таким методом є .hist(), а для діаграми розмаху - .boxplot(). Для обох діаграм задається розмір: 20 по висоті й 12 по ширині для кращого відображення діаграм.

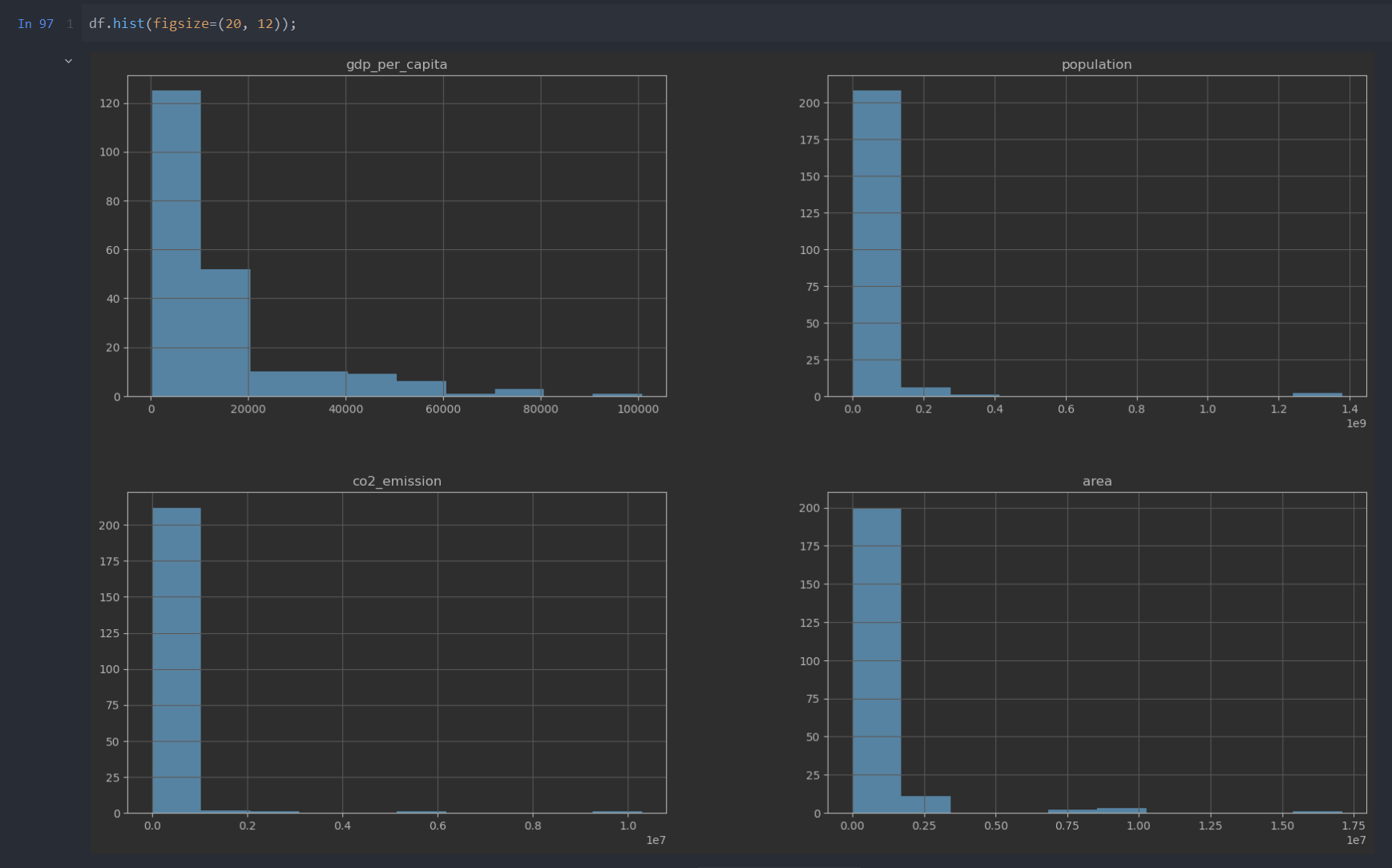


Рис. 3.7 – Побудова гістограм

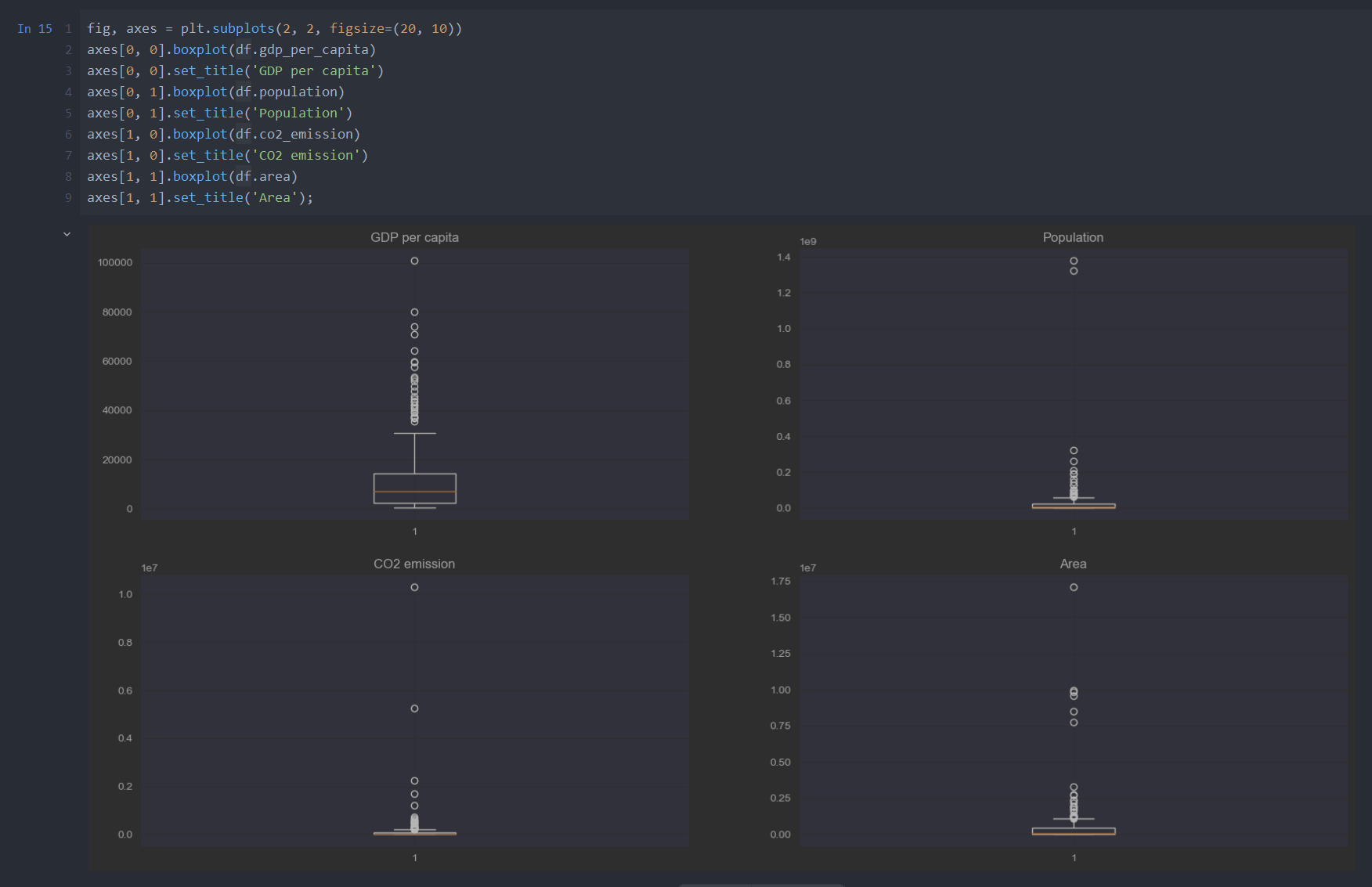


Рис. 3.8 – Побудова діаграм розмаху

## Додати стовпчик із щільністю населення

Зробити це можна дуже легко: треба просто провести операцію над стовпцями населення та площі країн і вказати назву нового стовпця так, ніби ми просто до нього звертаємося.

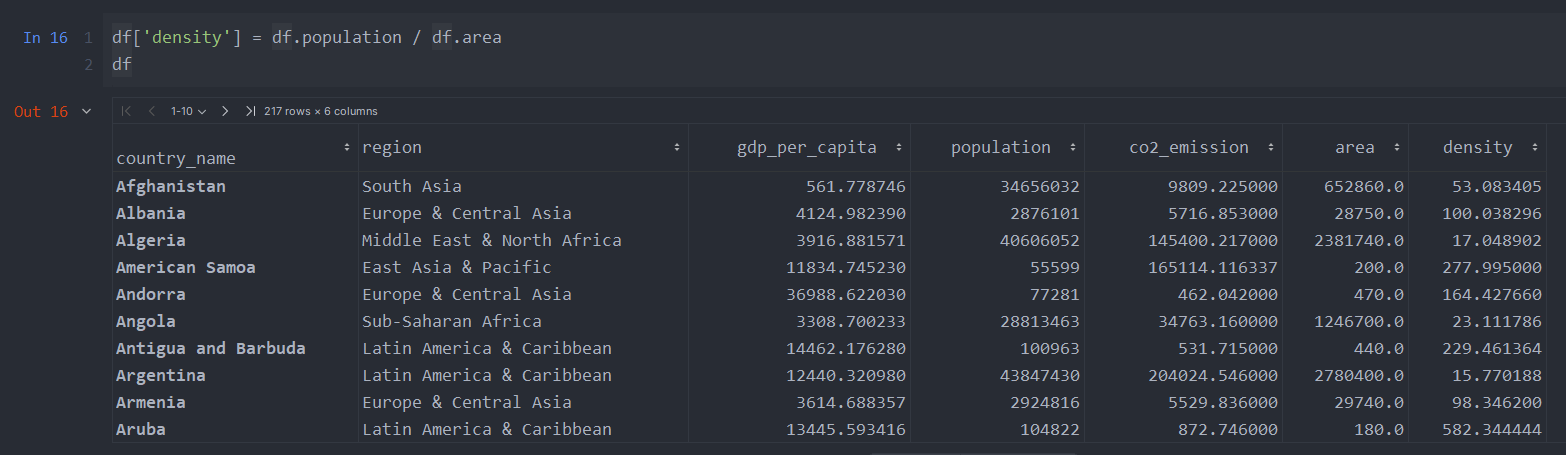


Рис. 3.9 – Додавання стовпця із щільністю населення

# Виконання додаткового завдання

## Чи є пропущені значення? Якщо є, замінити середніми

Варто зазначити, що виконання таких операцій, як побудова діаграм розмаху та створення стовпця щільності населення було б неможливим без заміни пропущених значень. Перевірити, чи є пропущені значення в кожному зі стовпців, можна просто викликавши метод .isna().any() об’єкту типу DataFrame.

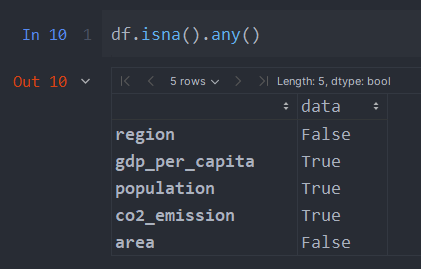


Рис. 3.10 – Перевірка, які стовпці містять пропущені значення

Як бачимо, пропущенні значення є в стовпцях ВВП на душу населення, кількості населення та кількості викидів CO2. Замінити пропущенні значення можна шляхом виклику методу .fillna() із вказанням значення, на яке треба замінити пропущене значення: за постановкою задачі, це середнє, а отже обрахуємо його шляхом виклику методу .mean(). У методі .fillna() встановимо аргумент inplace=True, щоб ці значення замінилися безпосередньо в DataFrame без зайвих операцій присвоєння.

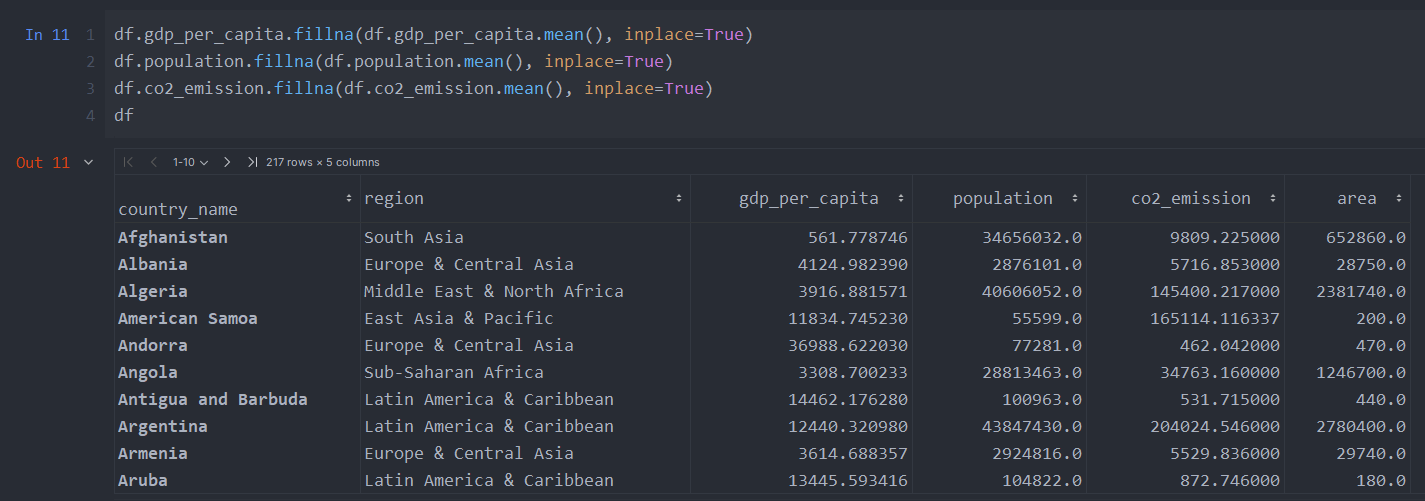


Рис. 3.11 – Заміна пропущених значень середніми значеннями

Можна після цього перевірити, що пропущенні значення відсутні викликавши вже згадані методи

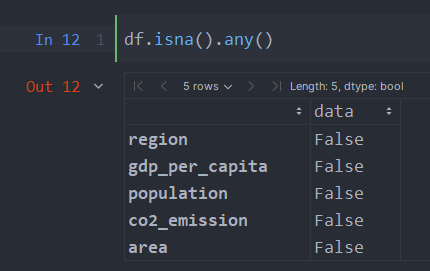


Рис. 3.12 – Перевірка, які стовпці містять пропущені значення

Також оскільки кількість населення – значення завжди ціле, а пропущенні значення відсутні (вони повинні обов’язково мати тип дійсного числа), можемо встановити тип стовпця населення як ціле число.

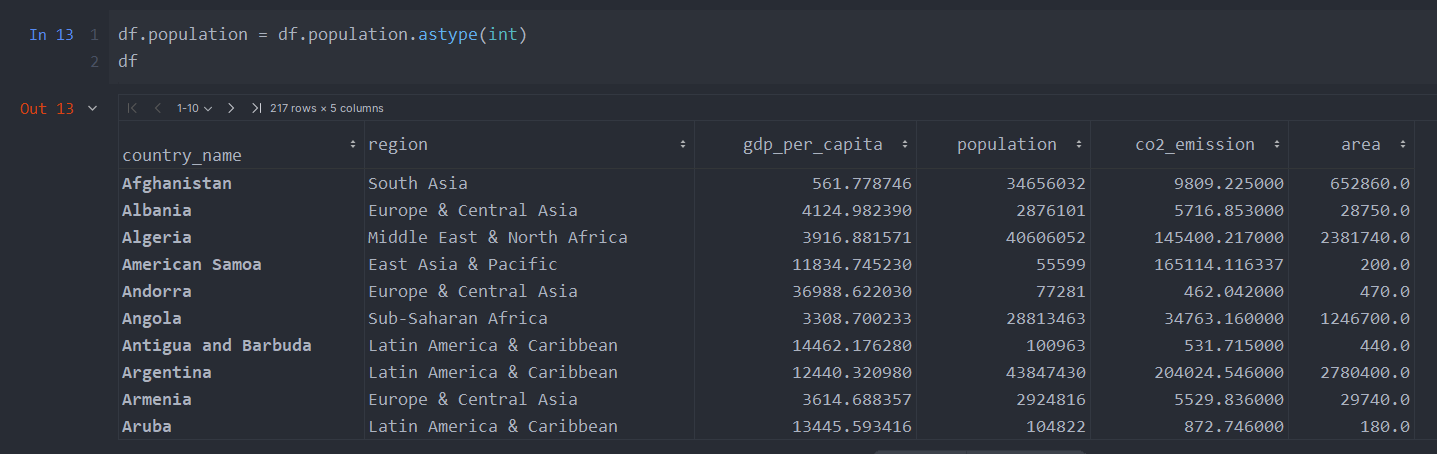


Рис. 3.13 – Перетворення типу кількості населення

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

Взнати, яка країна має найбільший ВВП на душу населення, можна викликавши спочатку метод .idxmax() для стовпця, за яким шукаємо максимальне значення, а потім викликати метод .loc[] нашого data frame, щоб відобразити повну інформацію щодо цієї країни.

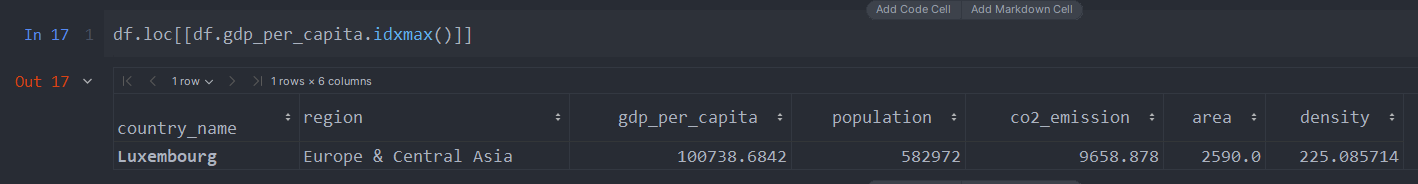


Рис. 3.14 – Виведення країни з найвищим ВВП на душу населення

Як бачимо, найвищий показник ВВП на душу населення в Люксембурзі. Проведемо такі ж операції для знаходження країни з найменшою площею, зміниться тільки метод: .idxmax() на .idxmin().

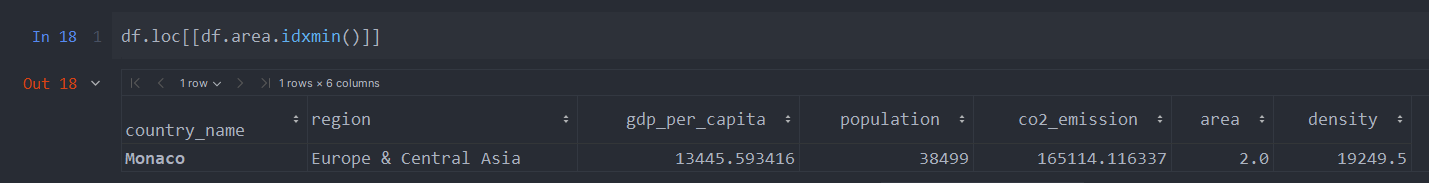


Рис. 3.15 – Виведення країни з найменшою площею

Як бачимо, Монако має найменшу площу серед нашого набору даних. Одразу ж можна зробити висновок про неповноту нашого набору даних, оскільки насправді країною з найменшою площею є Ватикан, а не Монако, але ця країна пропущена в нашому наборі даних.

## В якому регіоні середня площа країни найбільша?

Ця операція вже є складнішою. Треба спочатку згрупувати країни за регіоном (метод .groupby()), знайти середнє значення площі (.area.mean()), а потім вивести інформацію про цей континент, використавши вже згадані методи .loc[] та .idxmax().

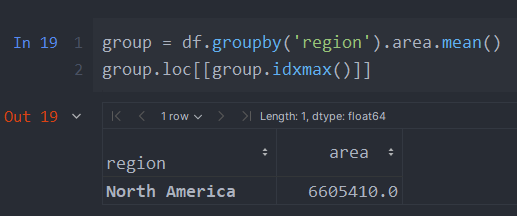


Рис. 3.16 – Виведення регіону із найбільшою середньою площею країн

Як бачимо, середня площа країн є найбільшою в Північній Америці.

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

Для таких операцій треба знову ж таки використати метод .idxmax() для стовпця щільності населення, а потім метод .loc[] для виведення повної інформації про цю країну

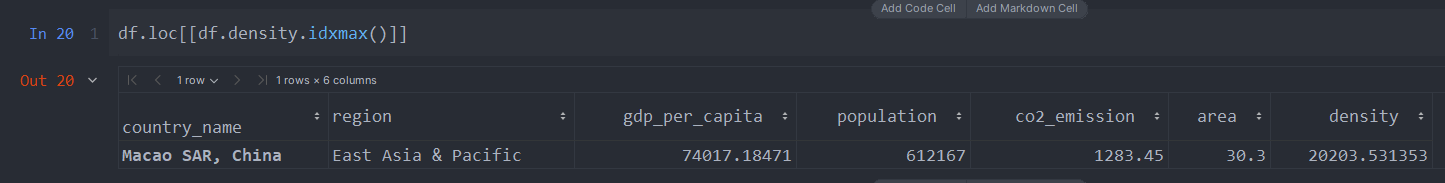


Рис. 3.17 – Виведення країни з найвищою щільністю населення

Як бачимо, Макао має найбільшу щільність населення у світі. Відфільтруємо країни за регіоном «Європа та Центральна Азія» й знайдемо країну з найбільшою щільністю населення, використавши ті ж самі вже згадані методи.

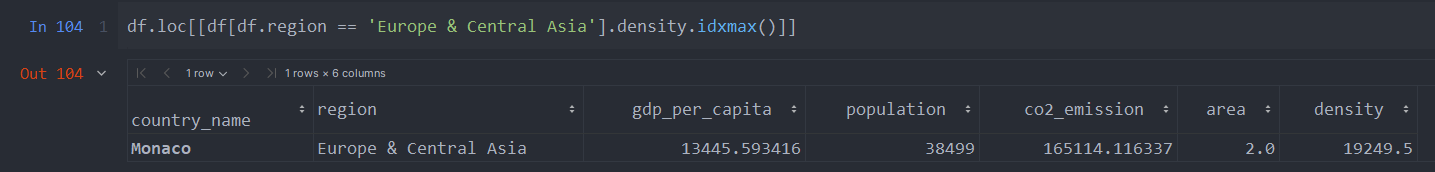


Рис. 3.18 – Виведення країни з регіону «Європа та Центральна Азія» із найбільшою щільністю населення

Робимо висновок, що Монако має найбільшу щільність населення в даному регіоні.

## Чи співпадає в якомусь регіоні середнє та медіана ВВП?

Зазначимо, що наш набір даних не містить стовпця із значенням ВВП, а отже його треба згенерувати за відомою формулою: кількість населення, помножена на ВВП на душу населення.

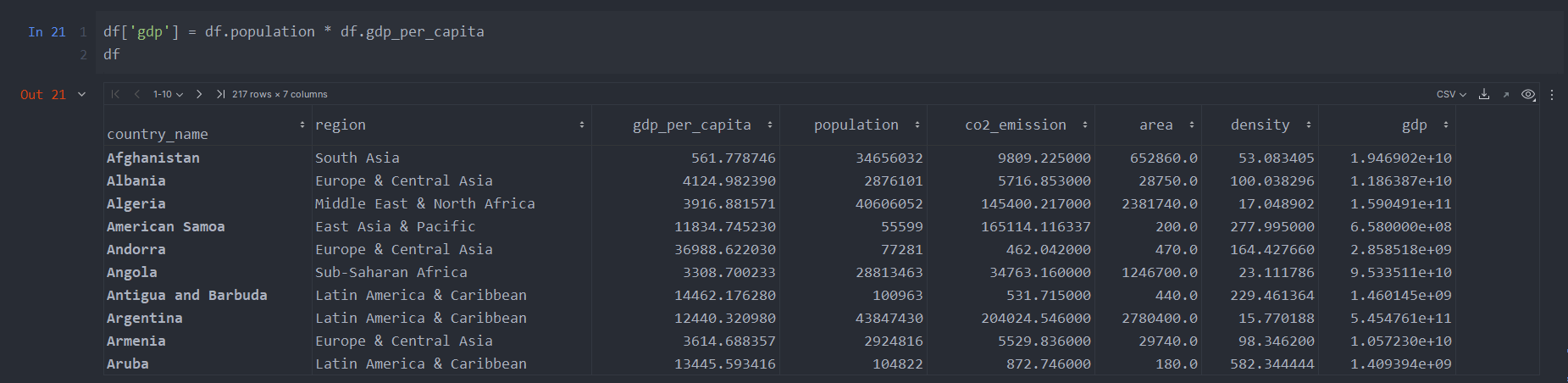


Рис. 3.19 – Обрахунок ВВП країн

Далі згрупуємо країни за регіоном, викликавши вже згаданий метод .groupby() із аргументом region (стовпець групування), потім оберемо стовпець, для якого обраховуємо середнє та медіану – у нашому випадку це ВВП й обрахуємо дані метрики.

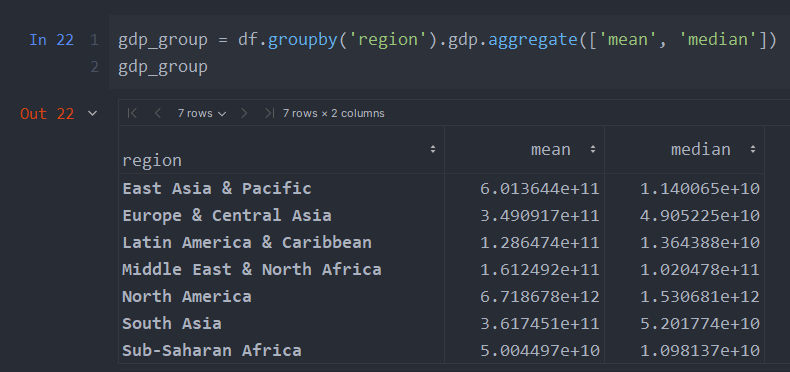


Рис. 3.20 – Обрахунок середніх та медіанних значень ВВП для кожного з регіонів

Далі для зручності додамо стовпець різниці між середнім та медіаною шляхом обчислення модуля різниці середнього та медіани для кожного зі стовпців

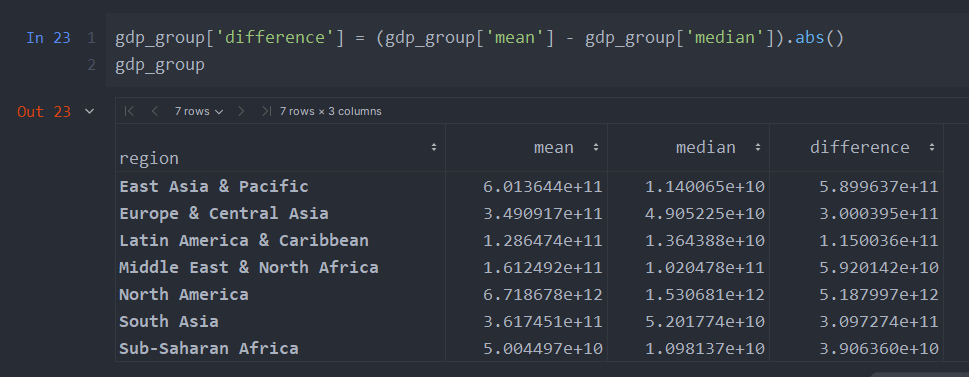


Рис 3.21 – Обчислення різниці між середнім та медіаною

Після цього необхідно прирівняти стовпці середнього та медіани й вивести ті, у яких вони рівні.

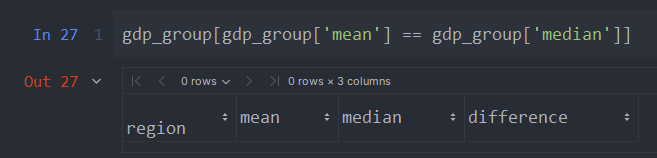


Рис. 3.22 – Виведення регіонів, середнє яких дорівнює медіані

Очікувано, такі регіони відсутні, але можемо подивитися, де різниця найменша.

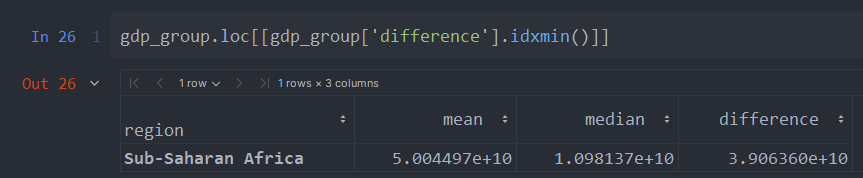


Рис. 3.23 – Виведення регіону із найменшою різницею між середнім та медіаною

Як бачимо, найменша різниця між середнім та медіаною є в Субсахарській Африці.

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

Знову ж зазначимо, що в наборі даних відсутній стовпець кількості CO2 на душу населення, натомість є стовпці кількості CO2 на країну та кількість населення. Проведемо такі ж дії, як і з обрахунком щільності населення.

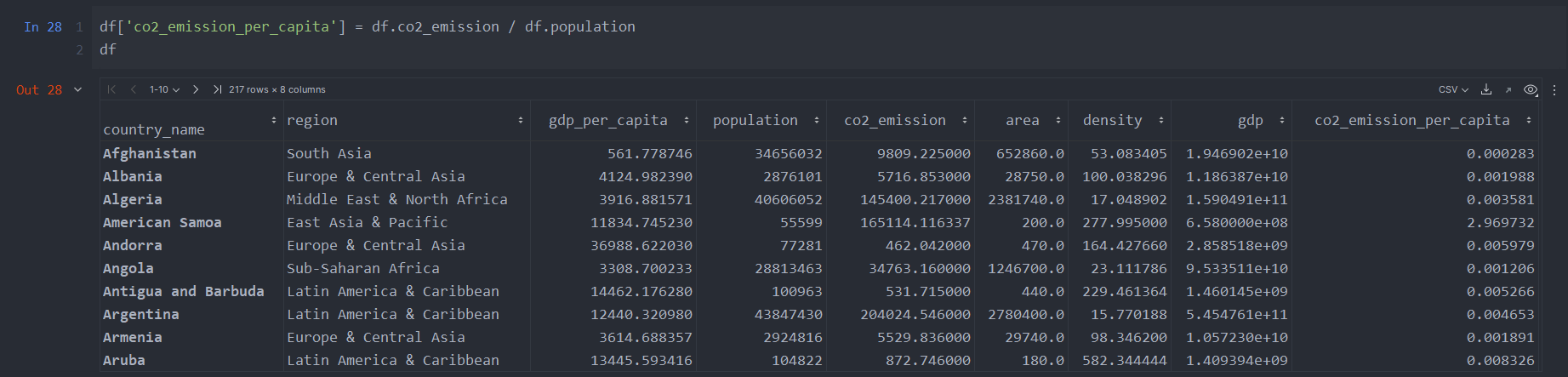


Рис. 3.24 – Обрахунок кількості викидів CO2 на душу населення

Далі відсортуємо набір даних за ВВП, викликавши метод .sort\_values() та вказавши стовпець, за яким іде сортування в якості аргументу by

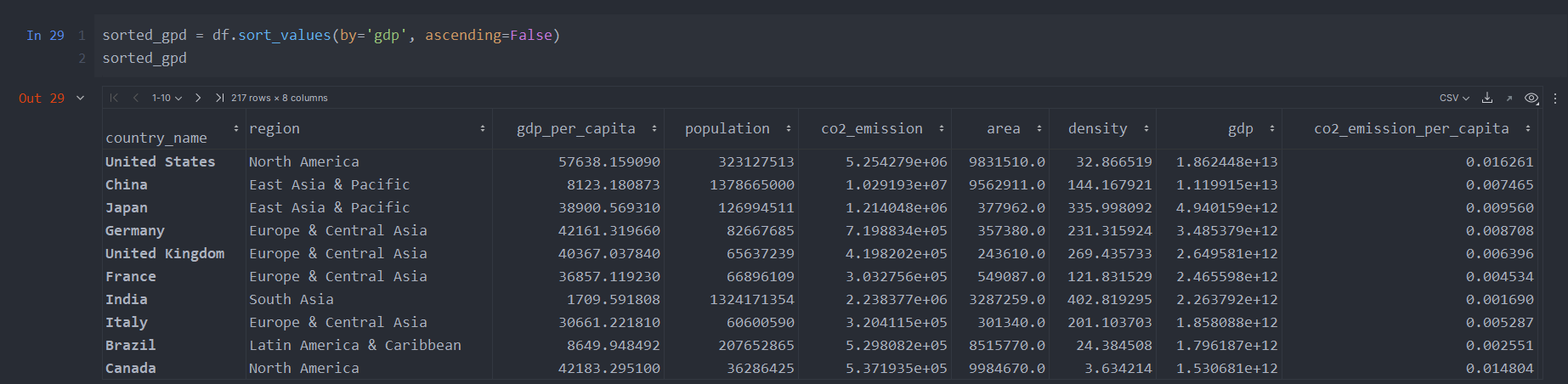


Рис. 3.25 – Сортування набору даних за ВВП

Після цього треба вивести 5 перших рядків, викликавши метод .head() із зазначенням кількості рядків (у нашому випадку – 5).

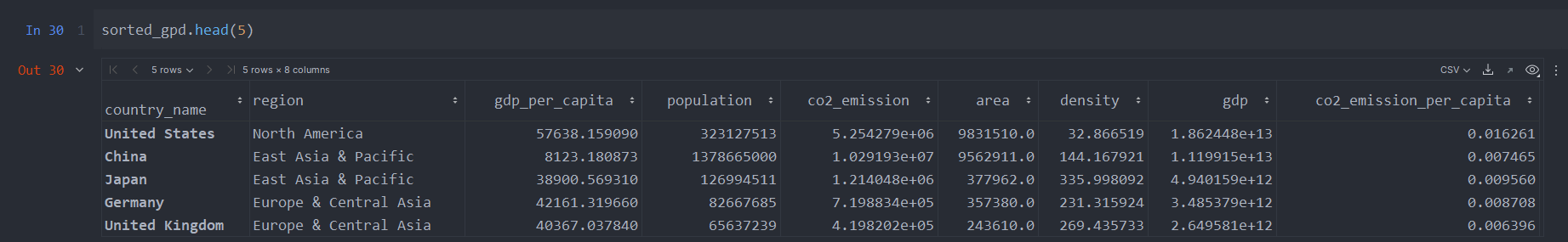


Рис. 3.26 – Виведення ТОП-5 країн за ВВП

Далі виведемо 5 останніх рядків, викликавши метод .tail() із зазначенням кількості рядків (у нашому випадку – 5) із поєднанням із методом .sort\_values() для того, щоб найвище стояв рядок із найменшим значенням ВВП.

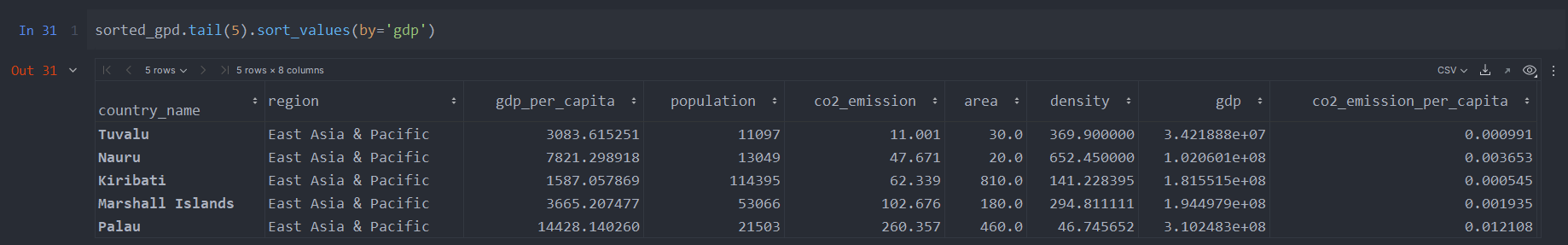


Рис. 3.27 – Виведення 5 країн із найменшим ВВП

Такі ж операції проведемо й для стовпця викидів CO2 на душу населення.

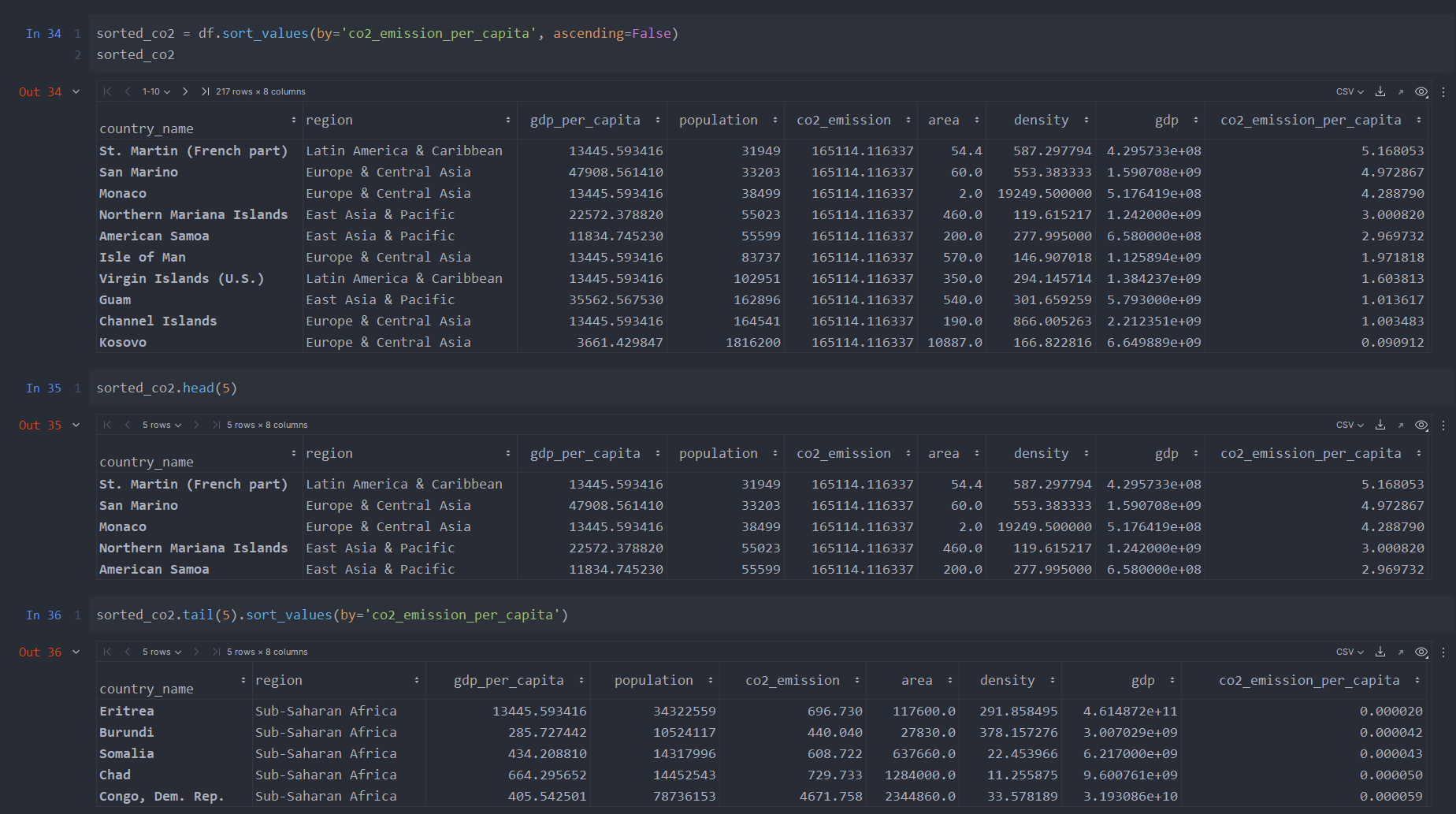


Рис. 3.28 – Виведення перших та останніх п’яти країн за викидами CO2 на душу населення

# Висновок

У ході даної лабораторної роботи я набув навичок в первинній обробці статистичних даних. Було проведено велику кількість операцій над набором даних, що містить інформацію про ВВП на душу населення, кількість населення, кількість викидів CO2 та площу: виправлено помилки, побудовано діаграми, додано стовпчики щільності населення, ВВП, викидів CO2 на душу населення, знайдено пропущені значення, замінено їх на середні, знайдено країни з найбільшою ВВП на душу населення (Люксембург), країну з найменшою площею (Монако), регіон із найбільшою середньою площею країн (Північна Америка), країну з найбільшою щільністю населення у світі (Макао), країну з найбільшою щільністю населення у Європі та Центральній Азії (Монако), перевірено, чи співпадає в якомусь із регіонів середнє та медіана ВВП (ні), виведено регіон із найменшою різницею між цими показниками (Субсахарська Африка), а також виведено ТОП-5 та 5 останніх країн за ВВП та кількістю CO2 на душу населення. Загалом отримані навички значно допоможуть у проведенні складнішого аналізу даних.