**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 №4 з дисципліни

«Аналіз даних в інформаційних системах»

„**Вивідна 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 4](#_Toc131690473)

[2 Завдання 5](#_Toc131690474)

[2.1 Основне завдання 5](#_Toc131690475)

[2.2 Додаткове завдання 5](#_Toc131690476)

[3 Виконання основного завдання 6](#_Toc131690477)

[3.1 Подивитись, проаналізувати структуру 6](#_Toc131690478)

[3.2 Вказати, чи є параметри, що розподілені за нормальним законом 7](#_Toc131690479)

[3.3 Перевірити гіпотезу про рівність середнього і медіани для одного з параметрів 8](#_Toc131690480)

[3.4 Вказати, в якому регіоні розподіл викидів СО2 найбільш близький до нормального 8](#_Toc131690481)

[3.5 Побудувати кругову діаграму населення по регіонам 9](#_Toc131690482)

[4 Виконання Першого додаткового завдання 10](#_Toc131690483)

[4.1 Завантажити карту України Ukraine.jpg 10](#_Toc131690484)

[4.2 Розмістити бульбашки, що відповідають їх населенню, на довільних 5 містах (статистику взяти в інтернеті) 10](#_Toc131690485)

[4.3 Знайти найбільшу відстань між містами в пікселях та кілометрах 11](#_Toc131690486)

[5 Виконання другого додаткового завдання 14](#_Toc131690487)

[5.1 Завантажити shape-файл с областями України. 14](#_Toc131690488)

[5.2 Побудувати картограми для прибутку населення на 1 особу і ВВП по регіонам за 2016 рік. 14](#_Toc131690489)

[5.3 По даним за 2006-2015 роки для кожного регіону розрахувати коефіцієнт кореляції між прибутком населення на 1 особу та ВВП. Відобразити на картограмі. 17](#_Toc131690490)

[6 Висновок 19](#_Toc131690491)

# Мета лабораторної роботи

Мета роботи – ознайомитись з методами визначення точкових оцінок параметрів розподілу; дослідити, що впливає на якість точкових оцінок; методикою визначення інтервальних оцінок параметрів розподілу; дослідити, що впливає на якість інтервальних оцінок; методами перевірки статистичних гіпотез про вигляд закону розподілу; дослідити, що впливає на ширину критичної області.

# Завдання

## Основне завдання

Скачати дані із файлу Data2.csv

1. Подивитись, проаналізувати структуру
2. Вказати, чи є параметри, що розподілені за нормальним законом
3. Перевірити гіпотезу про рівність середнього і медіани для одного з параметрів
4. Вказати, в якому регіоні розподіл викидів СО2 найбільш близький до нормального
5. Побудувати кругову діаграму населення по регіонам

## Додаткове завдання

Відповісти на питання (файл Data2.csv):

Завдання 1

1. Завантажити карту України Ukraine.jpg
2. Розмістити бульбашки, що відповідають їх населенню, на довільних 5 містах (статистику взяти в інтернеті)
3. Знайти найбільшу відстань між містами в пікселях та кілометрах

Завдання 2

1. Завантажити shape-файл с областями України.
2. Побудувати картограми для прибутку населення на 1 особу і ВВП по регіонам за 2016 рік.
3. По даним за 2006-2015 роки для кожного регіону розрахувати коефіцієнт кореляції між прибутком населення на 1 особу та ВВП. Відобразити на картограмі.

# Виконання основного завдання

## Подивитись, проаналізувати структуру

Для початку треба завантажити дані у Python за допомогою бібліотеки pandas.

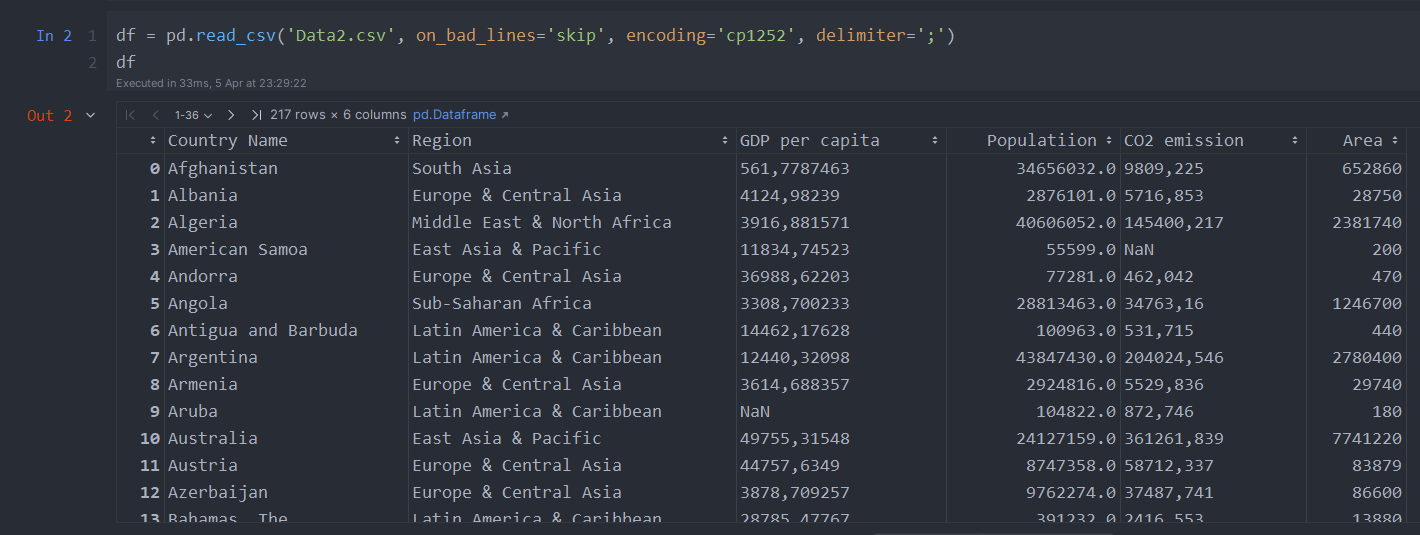


Рис. 3.1 – Завантаження даних

Бачимо, що в наборі даних є наступні стовпці: назва країни, регіон, ВВП на душу населення, кількість населення, викиди СО2 та площа. Як було описано в попередній лабораторній роботі, набір даних містить наступні помилки: дійсні числа записані через кому, є пропущенні значення ВВП на душу населення та викидів СО2 та від’ємні значення площі: виправимо ці помилки.

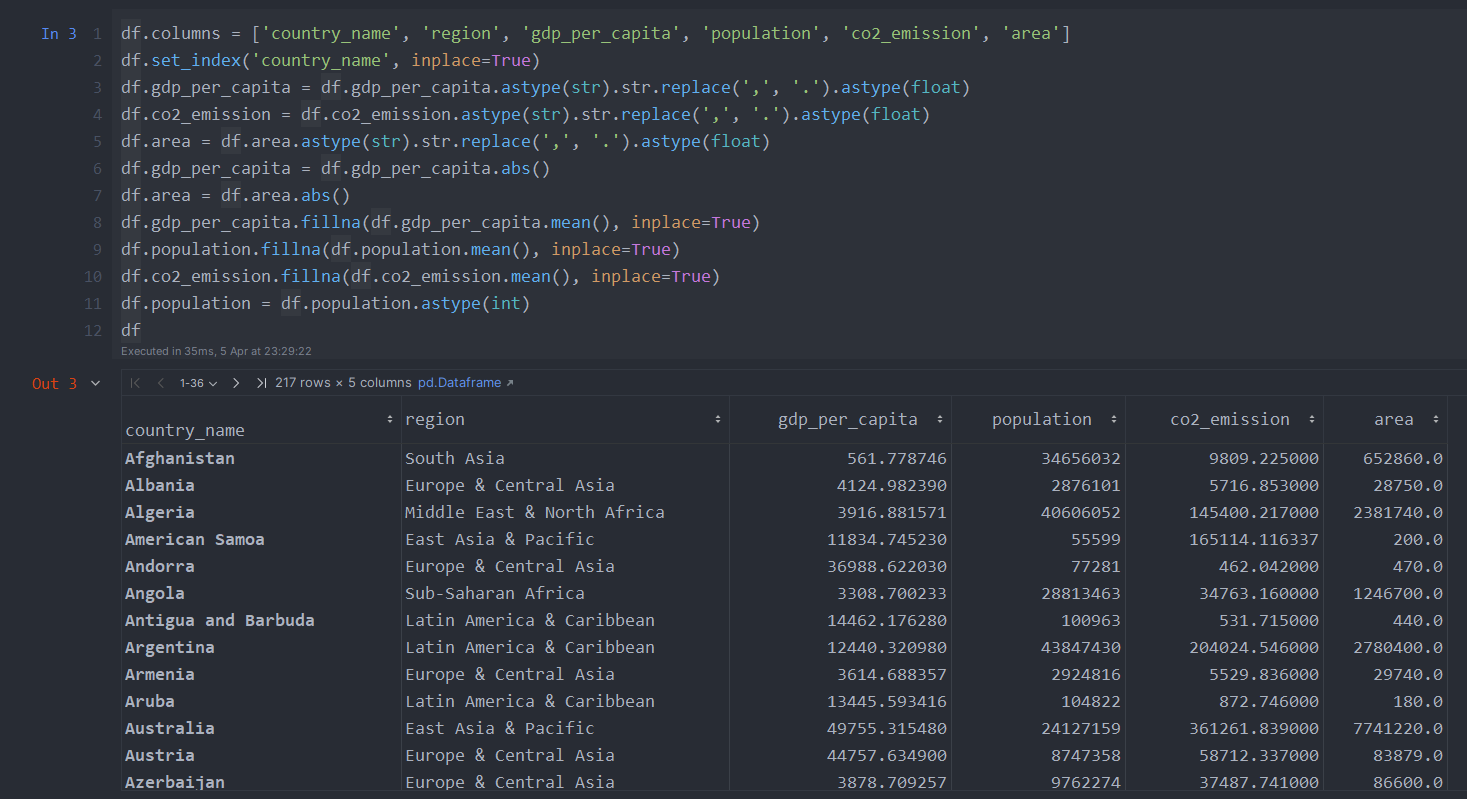


Рис. 3.2 – Виправлення помилок

## Вказати, чи є параметри, що розподілені за нормальним законом

Для перевірки такої гіпотези використаємо функцію бібліотеки scipy normaltest. Вона заснована на тесті Д’Агостіно і Пірсона. Рівень значущості оберемо стандартний – 0.05.

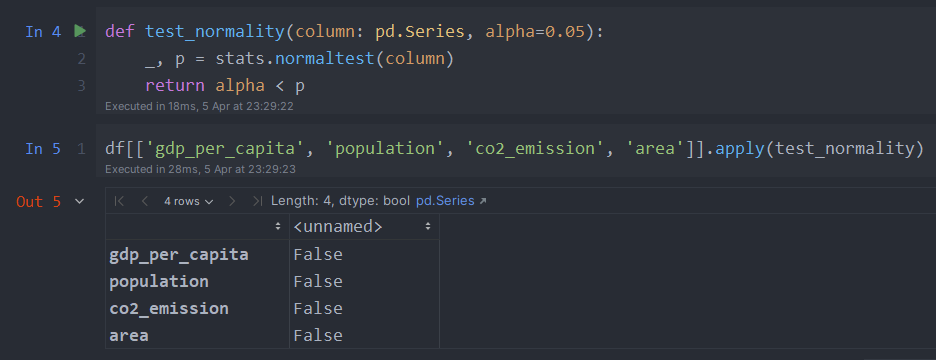


Рис. 3.3 – Перевірка гіпотези про нормальність розподілу стовпців набору даних

Як бачимо, жоден із стовпців не розподілений за нормальним законом. Відобразимо гістограми й перевіримо це.

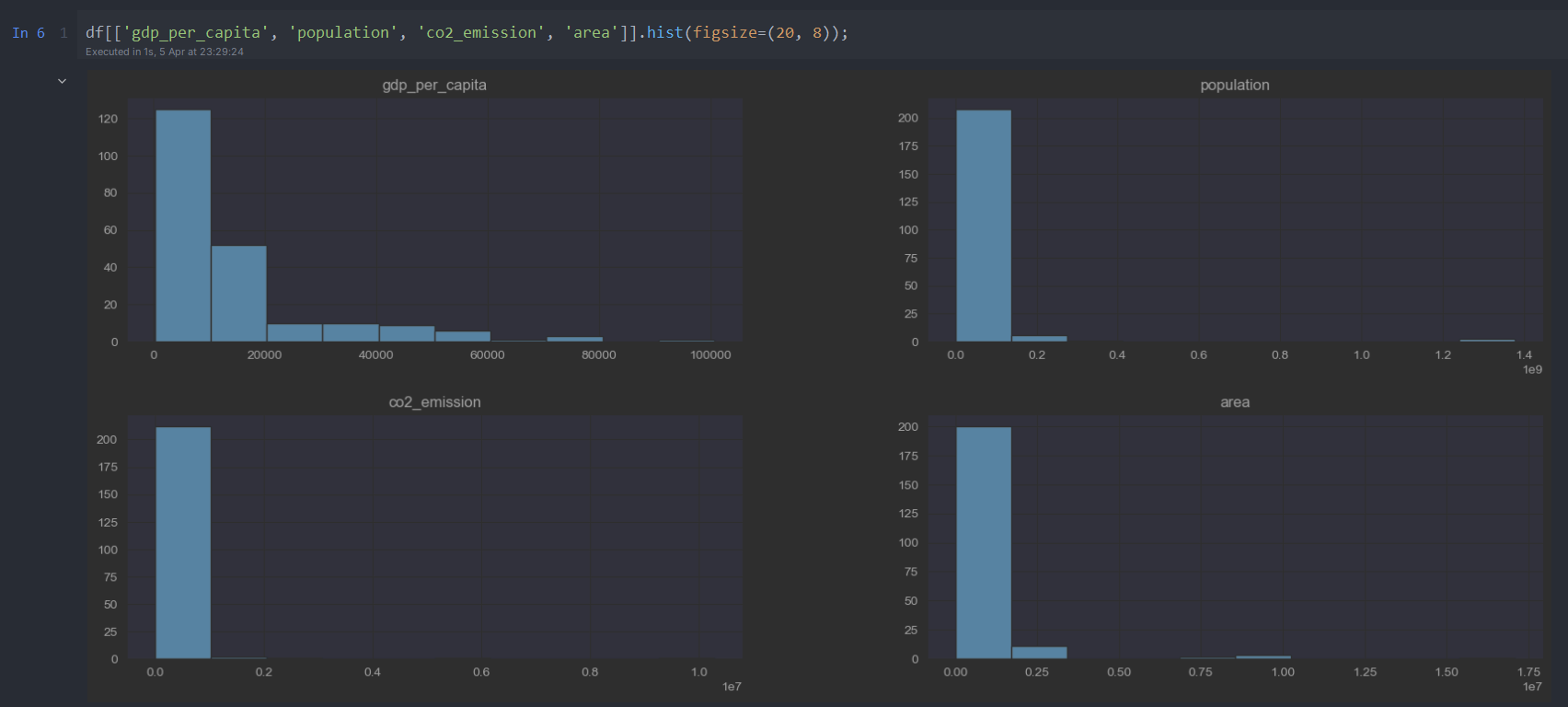


Рис. 3.4 – Гістограми для кожного стовпця

## Перевірити гіпотезу про рівність середнього і медіани для одного з параметрів

Для перевірки цієї гіпотези використаємо функцію бібліотеки scipy ttest\_1samp: вона використовує Т-критерій Стьюдента. Рівень значущості оберемо такий самий – 0.05.

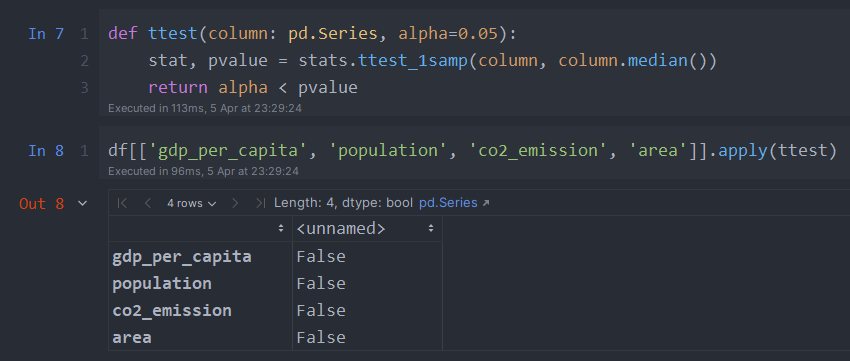


Рис. 3.5 – Перевірка гіпотези про рівність середнього і медіани

Як бачимо, гіпотеза не справджується для жодного зі стовпців.

## Вказати, в якому регіоні розподіл викидів СО2 найбільш близький до нормального

Для такої перевірки використаємо функцію бібліотеки scipy anderson, що задіює тест Андерсона-Дарлінга. Функція повертає статистику та критичні значення для наступних рівней значущості: 15%, 10%, 5%, 2.5%, 1%. Оберемо 5%, а потім віднімемо отриману статистику від цього критичного значення. Чим менше це значення буде, тим розподіл є ближчим до нормального

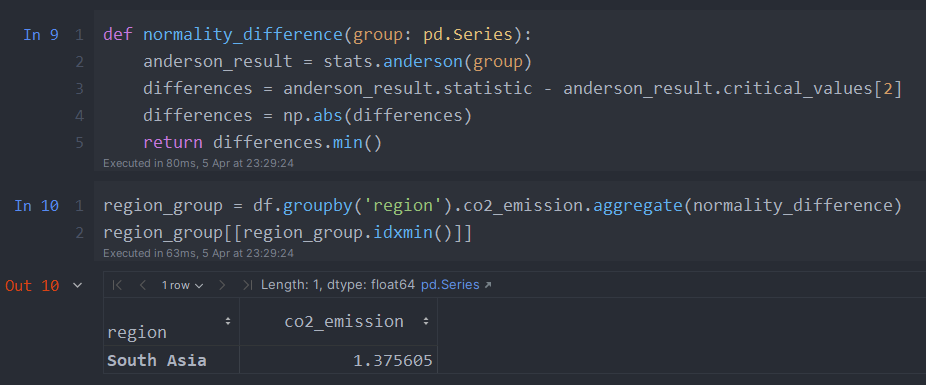


Рис. 3.6 – Пошук регіону, розподіл викидів СО2 якого є найбільш близьким до нормального

Бачимо, що для регіону «Південна Азія» така різниця дорівнює 1.3756, що все одно доволі багато.

## Побудувати кругову діаграму населення по регіонам

Для побудови кругової діаграми використаємо простий метод .pie об’єкту класу DataFrame, встановимо виведення значень у форматі відсотків.

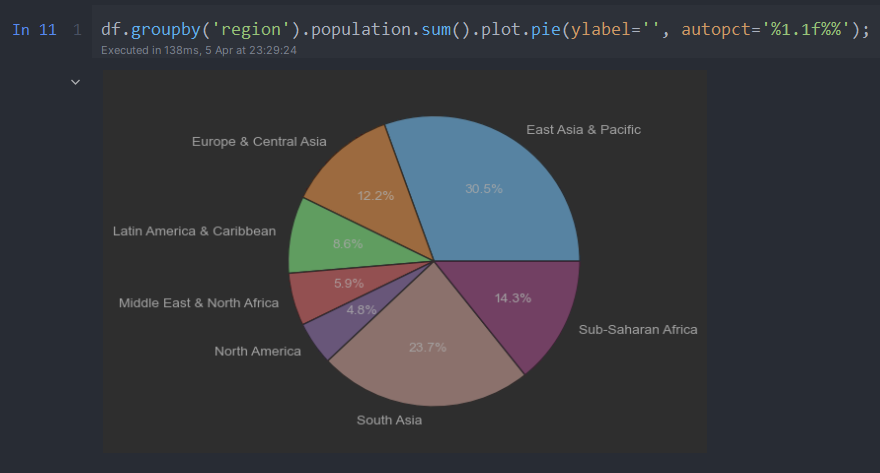


Рис. 3.7 – Побудова кругової діаграми населення за регіонами

# Виконання Першого додаткового завдання

## Завантажити карту України Ukraine.jpg

## Розмістити бульбашки, що відповідають їх населенню, на довільних 5 містах (статистику взяти в інтернеті)

Для початку знайдемо в інтернеті інформацію по населенню. Дані було взято з офіційного сайту Держстату. Також було знайдено в інтернеті дані про координати міст для подальшого розрахунку відстаней, а також прораховано координати точок міст на завантаженій карті: про це мова піде пізніше. Розмір бульбашок розрахуємо, поділивши кількість населення на 5000.

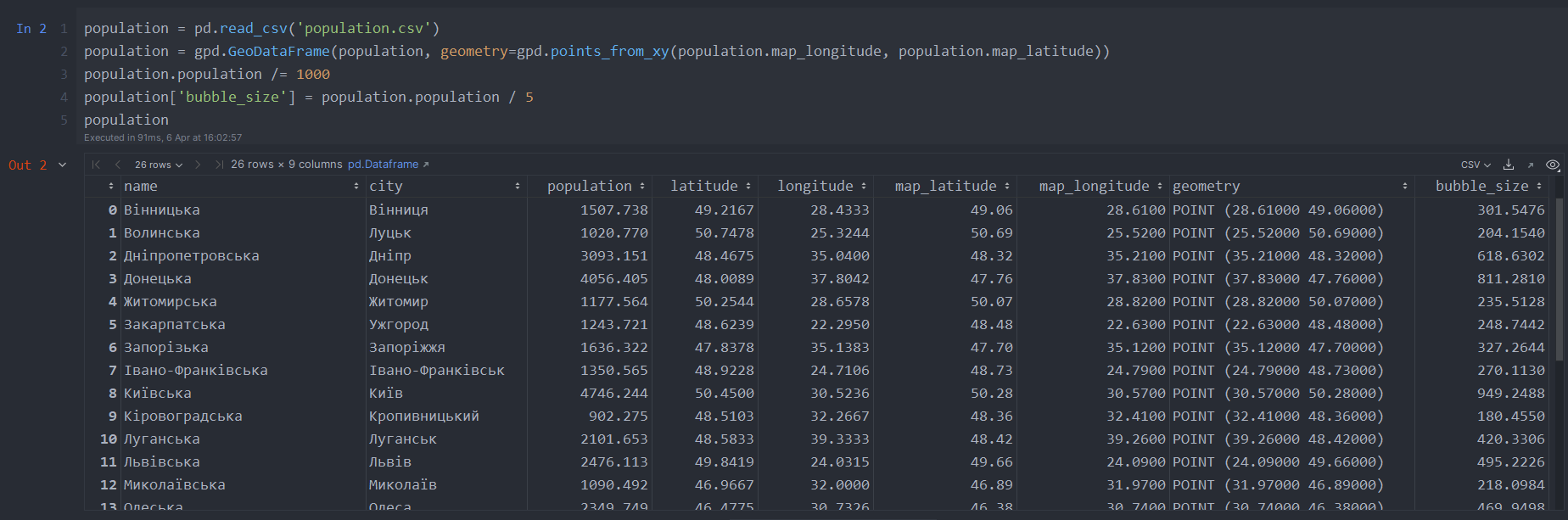


Рис. 4.1 – Завантаження даних про кількість населення в регіонах України

Дані ми завантажили за допомогою пакету geopandas, що є обгорткою над pandas для роботи з географічними даними. Для коректного відображення координат за допомогою сервісу Google Earth було знайдено точні координати крайніх точок України. Вони необхідні для коректного відображення координат на нашій карті. При першій спробі відобразити карту я стикнувся з проблемою, що географічні координати не відповідають точкам на завантаженій карті. Я вручну визначив координати цих точок та додав їх у набір даних і відображатиму бульбашки саме за цими координатами. Для відображення карти використаємо функцію imshow() пакету matplotlib, а для відображення бульбашок - .plot(). Також задамо обмеження координат для карти.

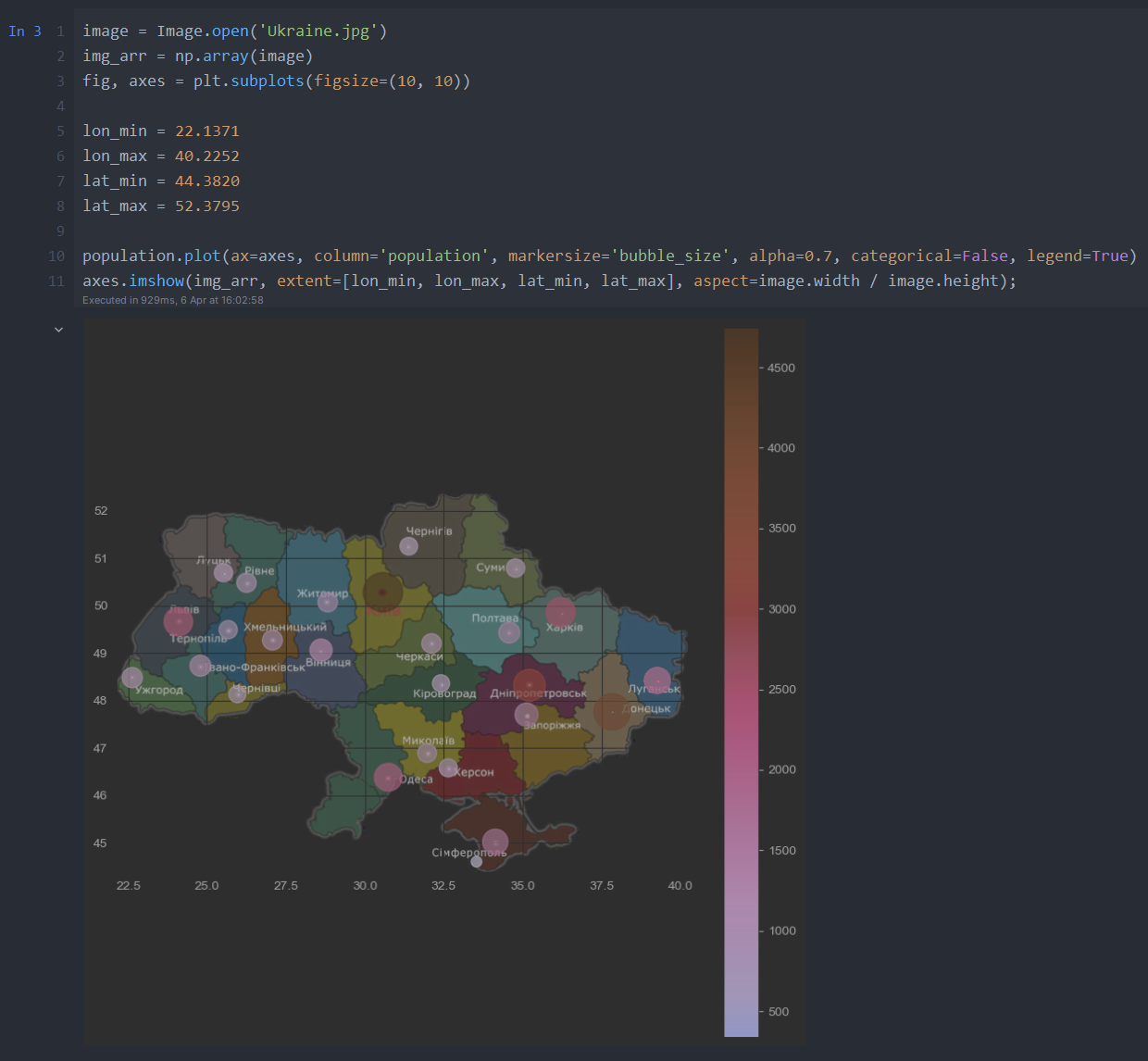


Рис. 4.2 – Зображення даних про кількість населення на мапі

## Знайти найбільшу відстань між містами в пікселях та кілометрах

Для початку було побудовано новий DataFrame, у якому є наступні стовпці: перше та друге міста, відстань у кілометрах та відстань у пікселях. Для розрахунку відстаней у кілометрах можна використати функцію distance пакету geopy, оскільки ця дія є досить складною для обчислення. Розрахунок відстані в пікселях було проведено наступним чином: було обраховано кількість пікселів на 1 градус широти та довготи й ці значення було поділено на різницю відповідних координат кожної з точок.

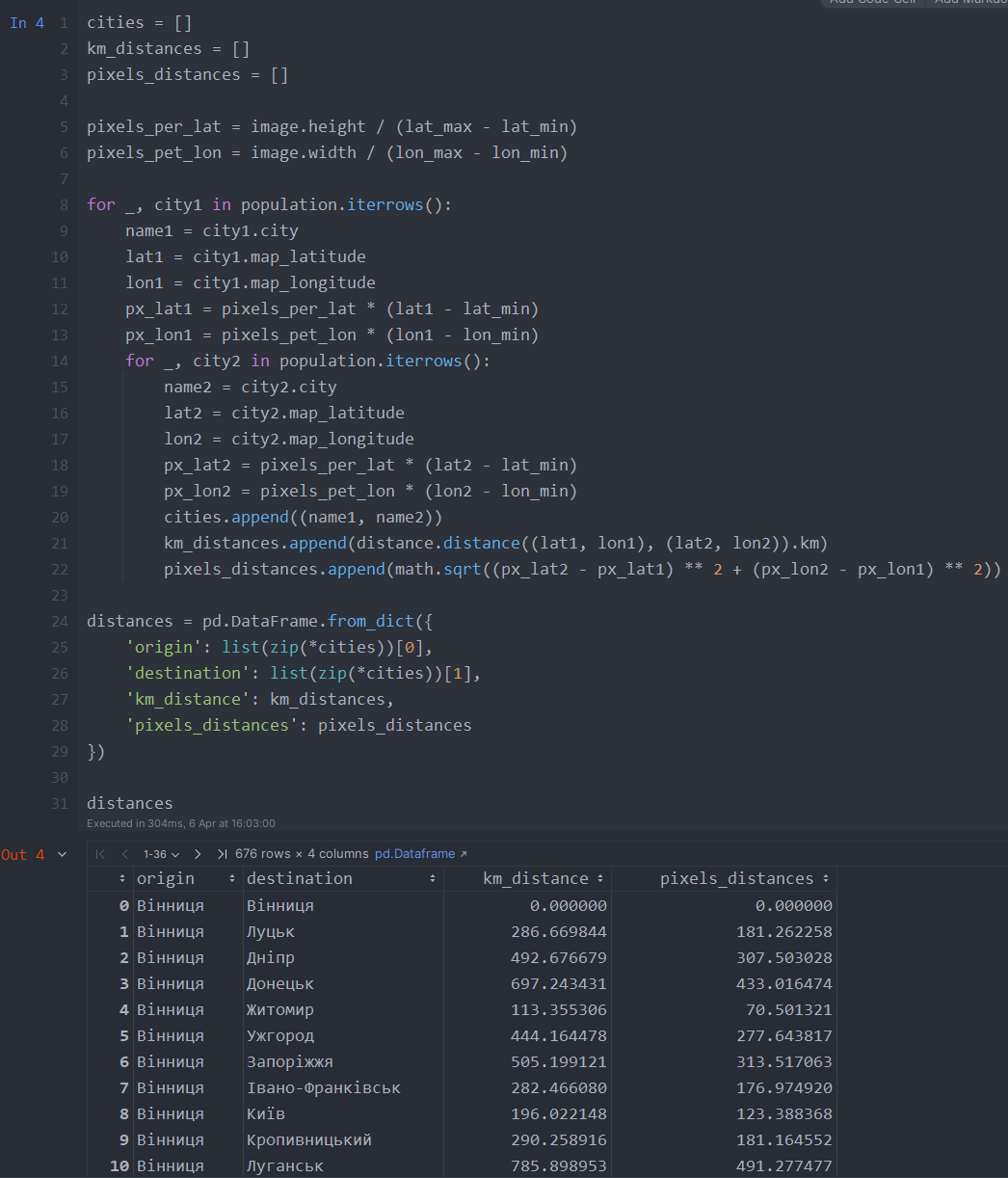


Рис. 4.3 – Обрахунок відстаней між містами

Для отримання найбільших відстаней між містами було використано методи .loc та idxmax()

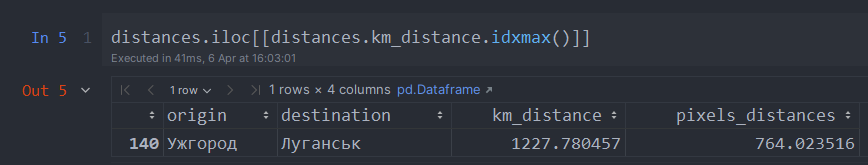


Рис. 4.4 – Пара міст із найбільшою відстанню в кілометрах

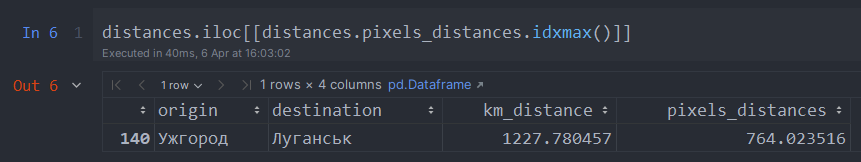


Рис. 4.5 – Пара міст із найбільшою відстанню в пікселях

# Виконання другого додаткового завдання

## Завантажити shape-файл с областями України.

Для роботи з shape-файлом було використано пакет geopandas.



Рис. 5.1 – Завантаження мапи

## Побудувати картограми для прибутку населення на 1 особу і ВВП по регіонам за 2016 рік.

Для початку завантажимо дані про ВВП за регіонами.

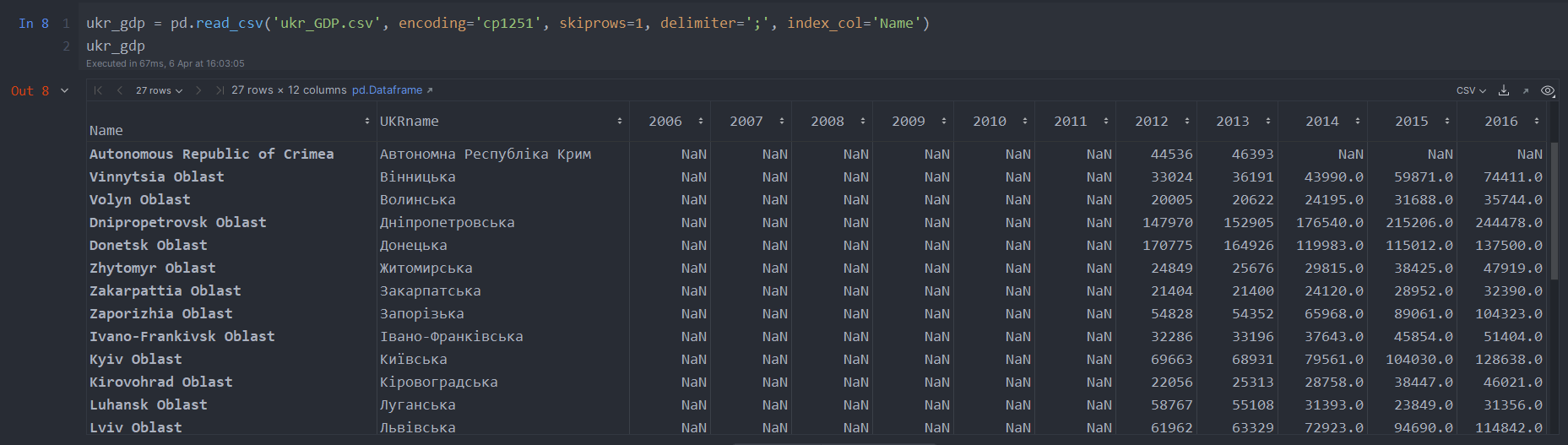


Рис. 5.2 – Завантаження даних про ВВП регіонів України

Бачимо, що є пропущені значення. Для їх заповнення використаємо поліноміальну регресію.

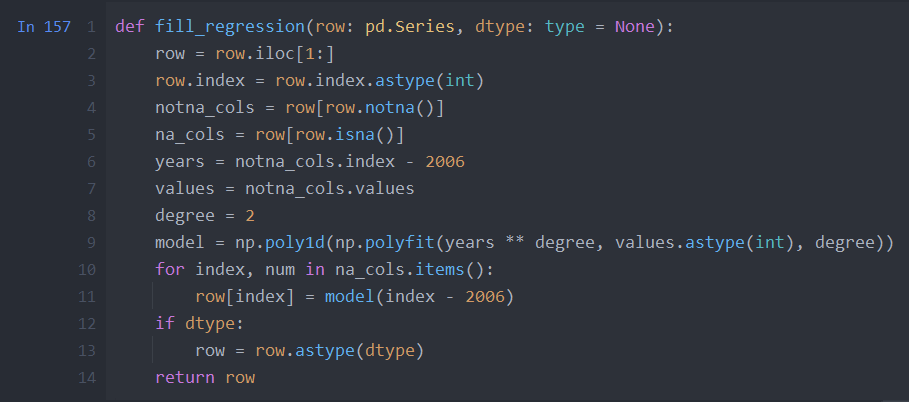


Рис. 5.3 – Побудова регресійної моделі

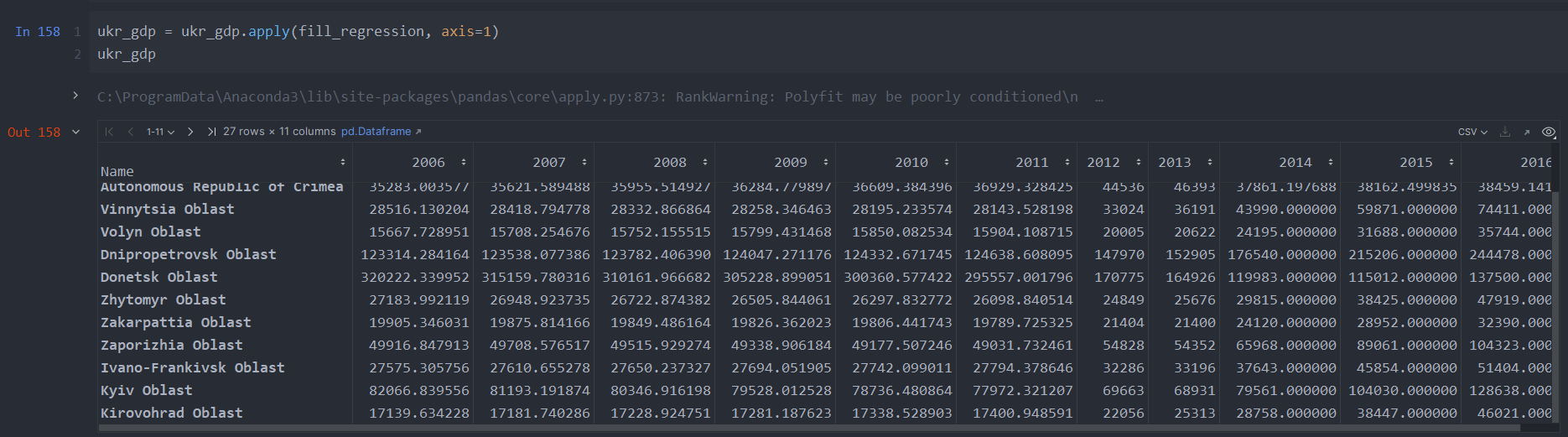


Рис. 5.4 – Заповнення пропущених значень

Далі об’єднаємо набір даних, створений на базі shape-файлу із набором даних про ВВП.

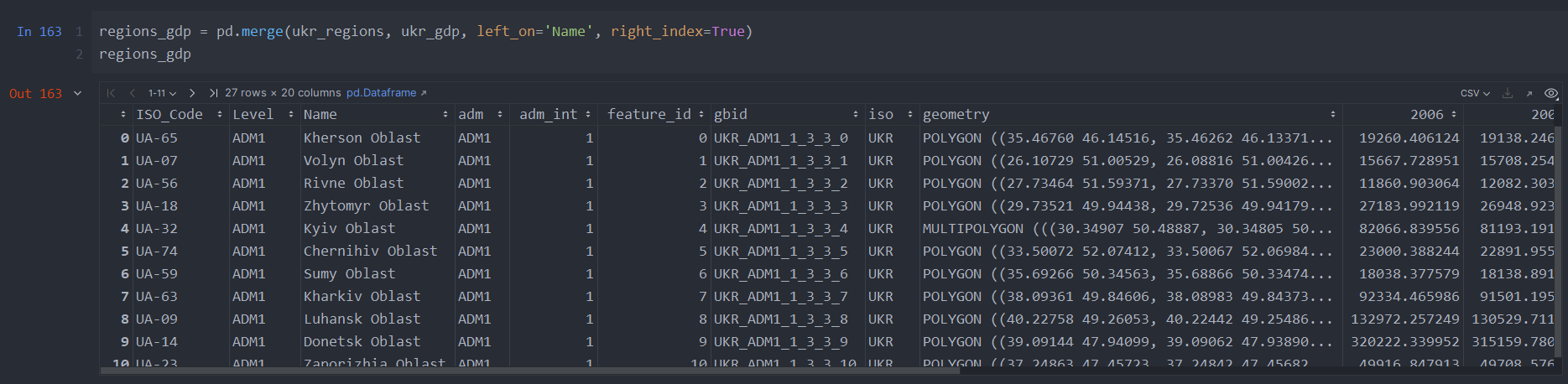


Рис. 5.5 – Об’єднаний DataFrame

Побудуємо картограму за допомогою методу .plot() нашого GeoDataFrame.

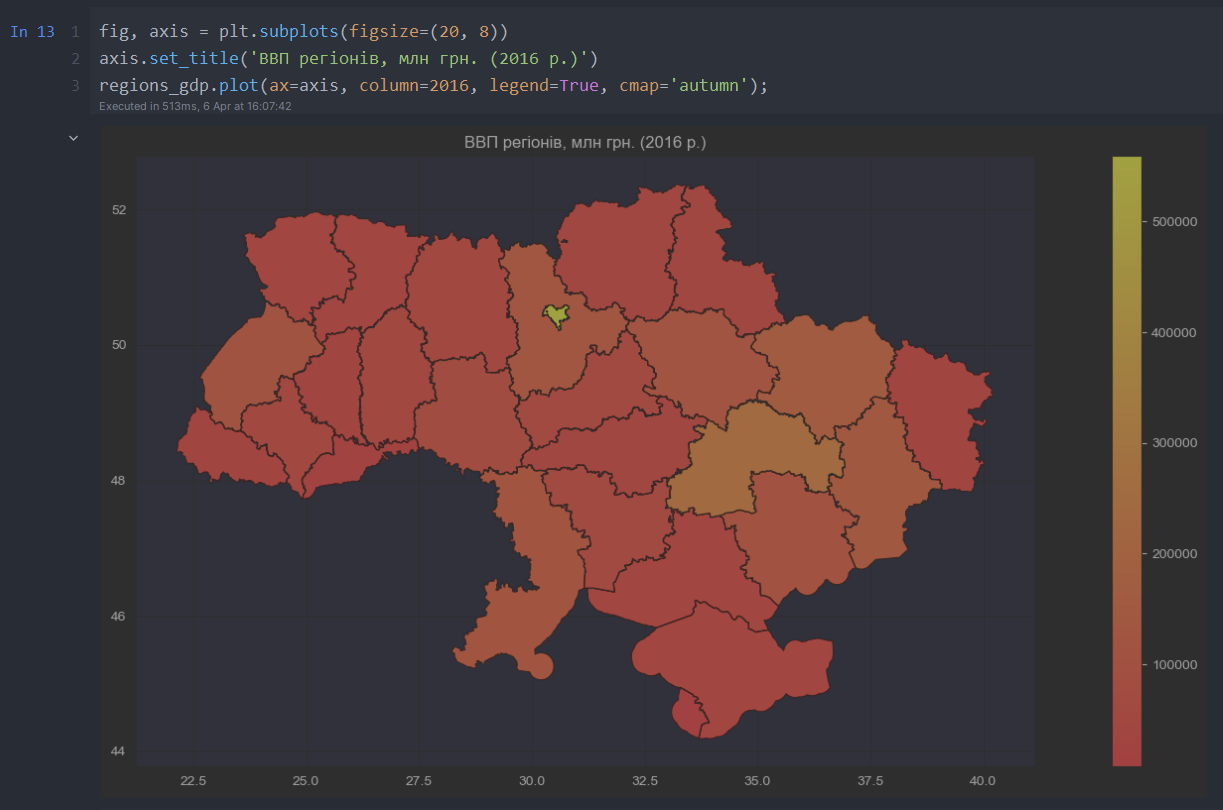


Рис. 5.6 – Картограма ВВП регіонів України в 2016 році

Як бачимо, найвищий ВВП має Київ.

Проведемо всі вищеописані дії для набору даних зарплат.

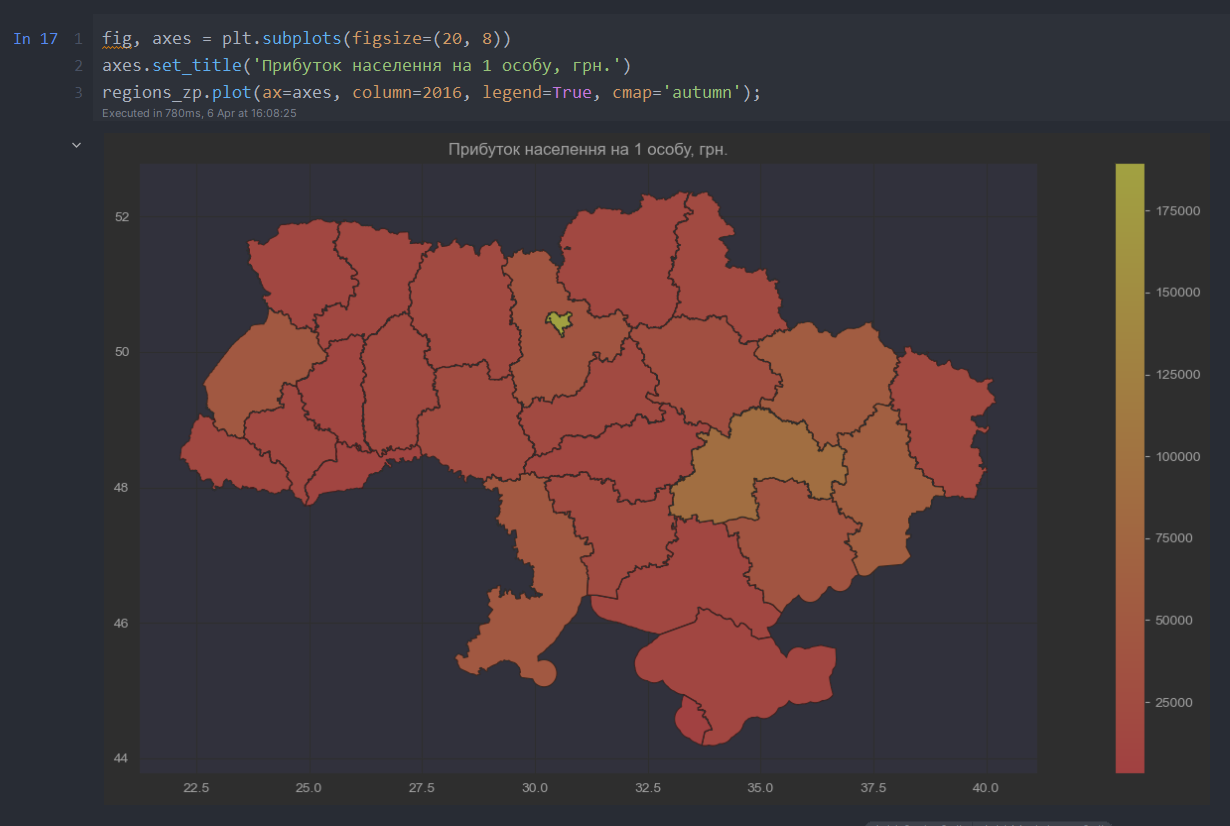


Рис. 5.7 – Картограма прибутку населення України в 2016 році

Як бачимо, найвищі зарплати знову в Києві.

## По даним за 2006-2015 роки для кожного регіону розрахувати коефіцієнт кореляції між прибутком населення на 1 особу та ВВП. Відобразити на картограмі.

Для обрахунку кореляції використаємо метод DataFrame .corrwith(), указавши необхідні роки. Виведемо дані про кореляцію.



Рис. 5.8 – Обрахунок коефіцієнтів кореляції

Далі побудуємо картограму за допомогою вже згаданого методу .plot().

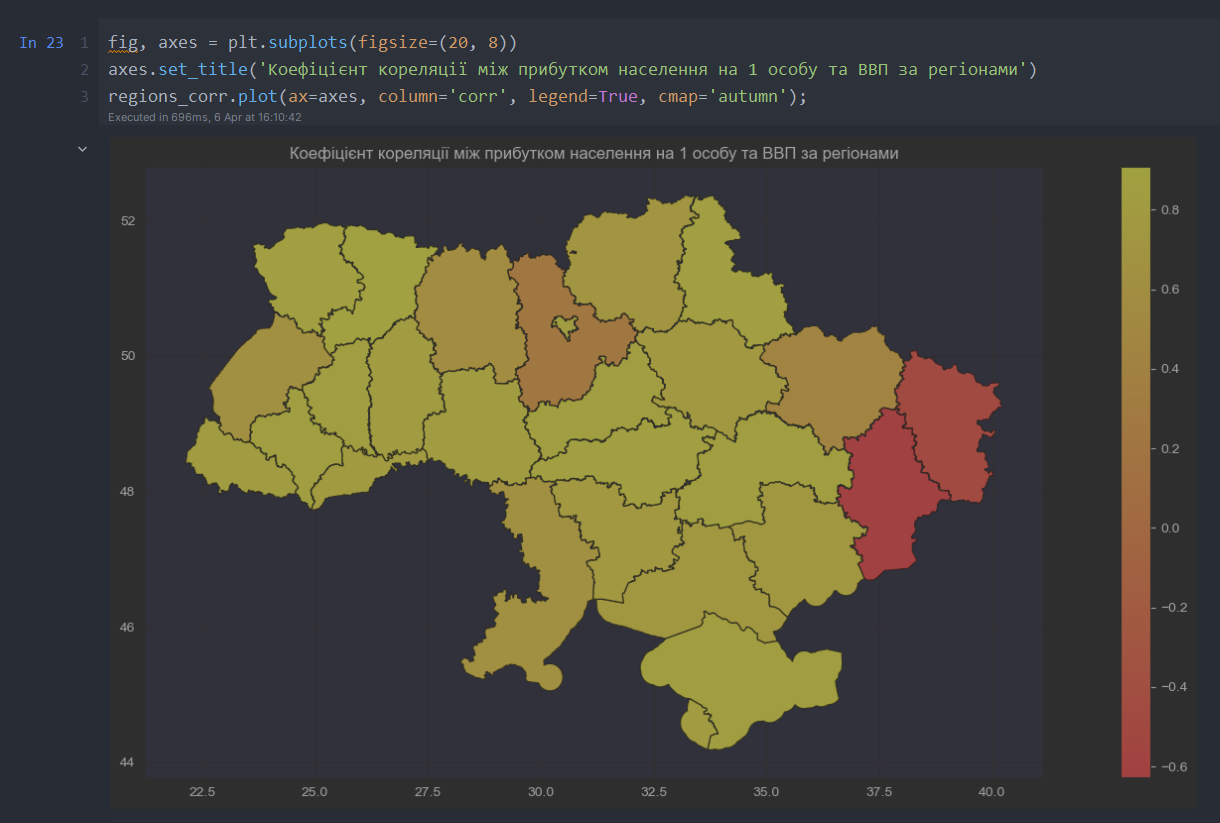


Рис. 5.9 – Відображення даних про кореляцію між прибутком населення та ВВП в регіонах України.

Можемо побачити, що всюди ВВП корелює з зарплатами, окрім Донецької та Луганської області. Спричинено це тим, що по цих регіонах відсутня достатня кількість інформації.

# Висновок

У ході даної лабораторної роботи було досліджено набір даних, що містить інформацію про країни світу. Усі стовпці було перевірено на нормальність розподілу, встановлено, що жоден зі стовпців не розподілений нормально, після цього було перевірено гіпотезу про рівність середнього та медіану, яку знову ж таки було відкинуто для всіх стовпців. Також за допомогою теста Андерсона-Дарлінга було встановлено, що найближчий до нормального розподіл викидів CO2 має регіон Південна Азія. Після проведення цих перевірок було побудовано кругову діаграму, що відображає кількість населення за регіонами.

У наступній частині лабораторної роботи було досліджено кількість населення в регіонах України та відображено JPG картограму з такими даними. Побачили, що найбільше населення в Києві. Також обрахували відстані між містами в кілометрах та пікселях і отримали, що найбільша відстань між Ужгородом та Луганськом – 1227 кілометрів або 764 пікселя.

Після цього було завантажено дані про ВВП та прибуток населення в регіонах України, заповнено пропущені значення за допомогою поліноміальної регресії другого порядку, після чого відображено картограми цих показників за 2016 рік (найвищі вони в Києві), а потім обраховано кореляцію між ВВП та прибутком населення й отримано, що всюди, окрім двох регіонів ці дані корелюють.