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 № 4 з дисципліни

«Аналіз даних в інформаційних системах»

«Вивідна статистика»

Виконав студент ІП-11 Лесів Владислав Ігорович

Перевірила Ліхоузова Тетяна Анатоліївна

Київ 2023

**Комп’ютерний практикум 4**

**Вивідна статистика**

**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.

**Основне завдання**.

Скачати дані 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.**

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 3

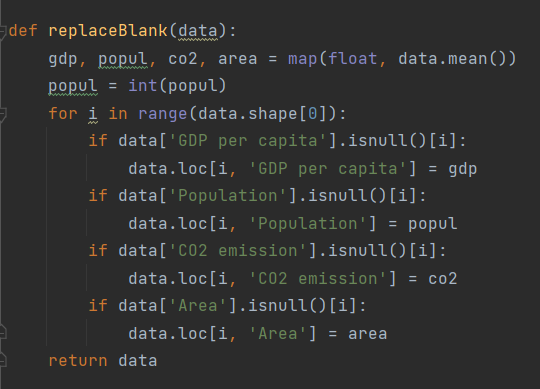
1. Завантажити shape-файл с областями України.
2. Побудувати картограми для прибутку населення на 1 особу і ВВП по регіонам за 2016 рік.
3. По даним за 2006-2015 роки для кожного регіону розрахувати коефіцієнт кореляції між прибутком населення на 1 особу та ВВП. Відобразити на картограмі.

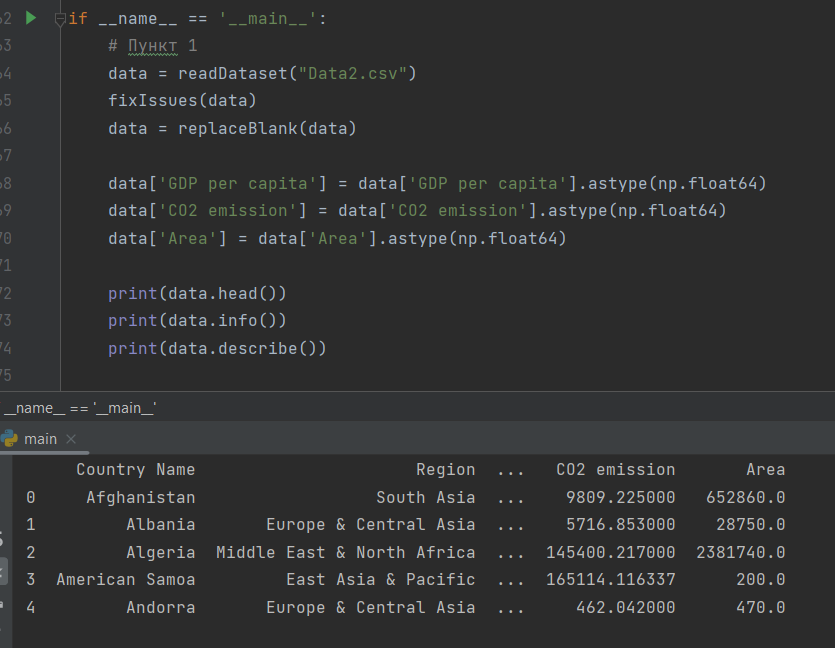
**Хід роботи.**

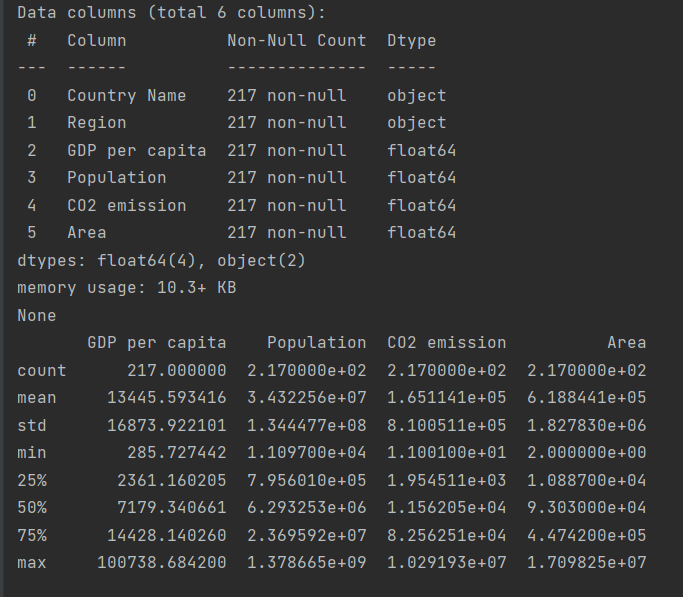
**Основне завдання.**

1. Як було зроблено в ході КП №3, записую дані у DataFrame за допомогою методу read\_csv, досліджую структуру даних (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). Виправляю проблеми даних. Назву стовпця Populatiion виправляю на Population. У числових значеннях замінюю коми на крапки і перетворюю у float. Виправляю від’ємні значення, а відсутні врешті заміняємо середніми від відповідного стовпця.



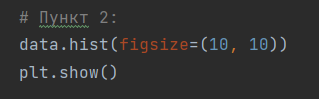


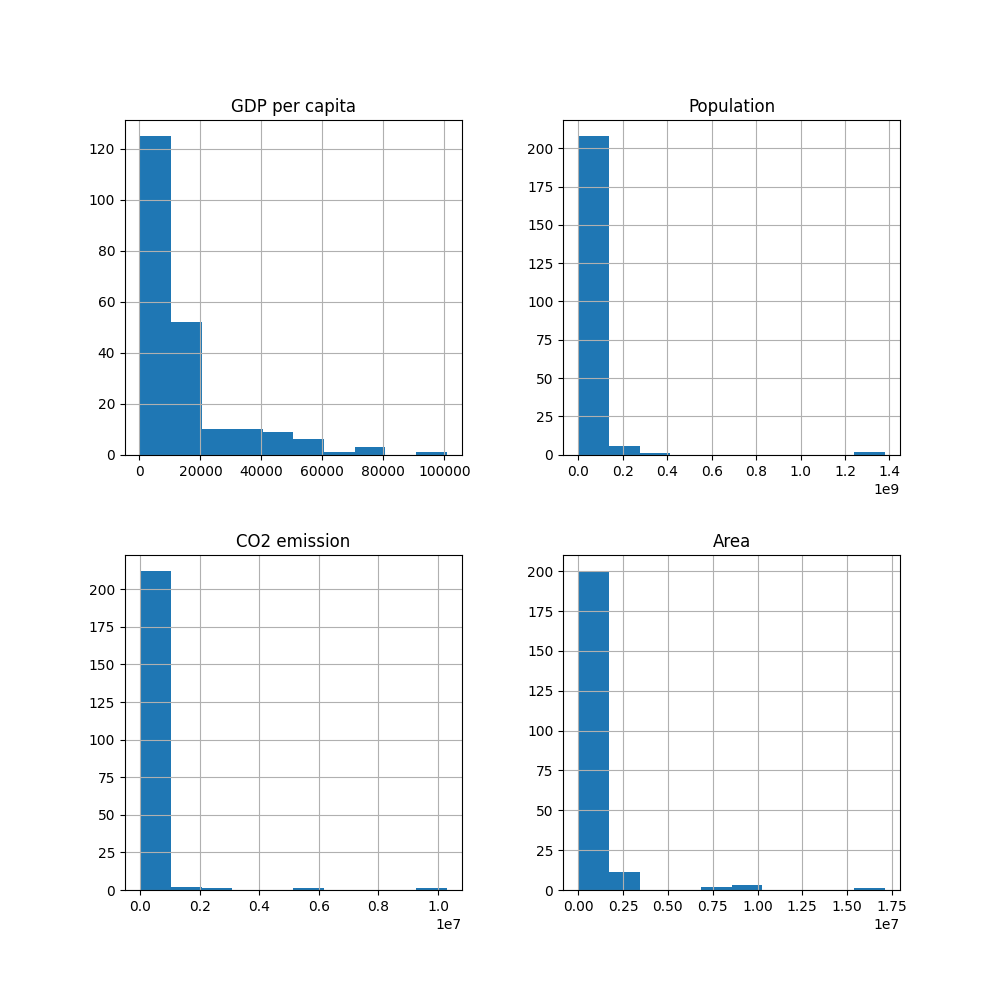




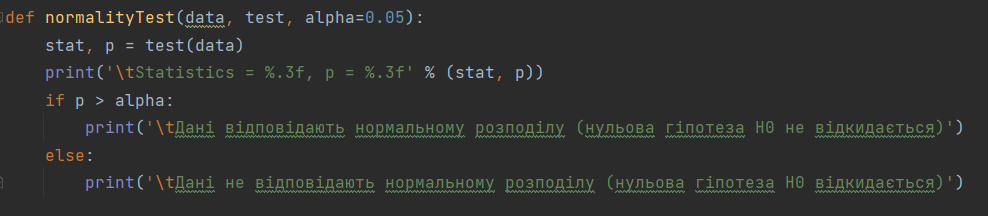
1. Вказати, чи є параметри, що розподілені за нормальним законом;

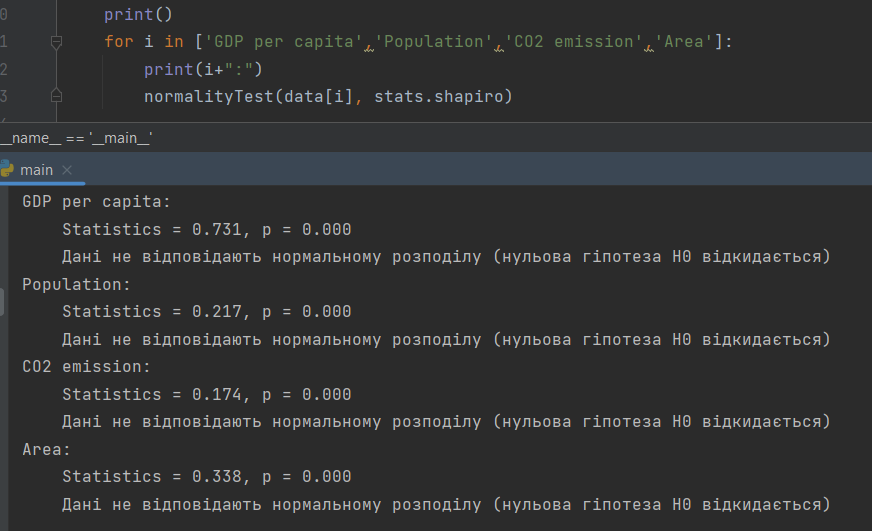
Для цього будую гістограми для кількісних параметрів датафрейму.





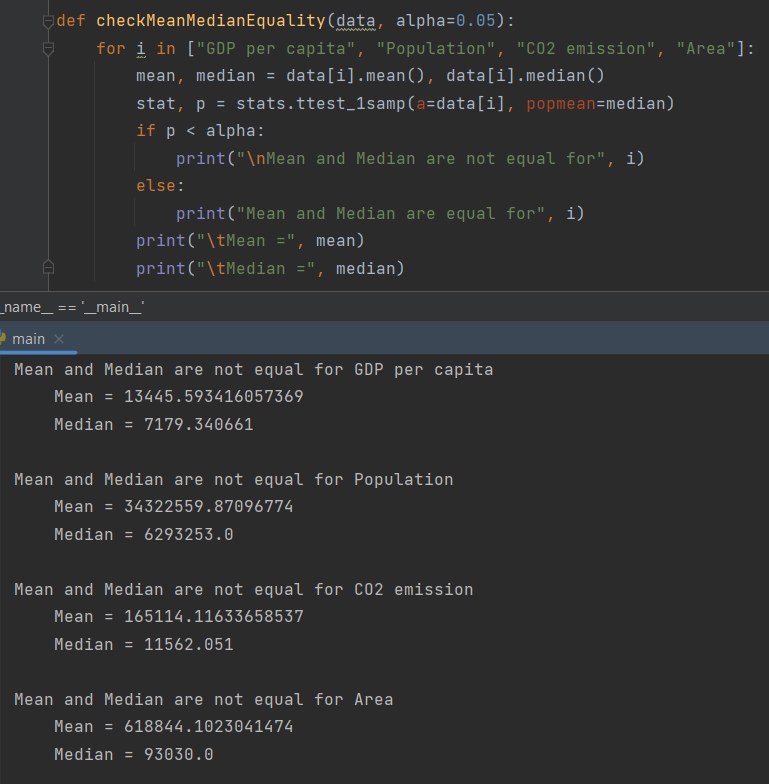
З гістограм видно, що жодна з величин не розподілена за нормальним законом. Також застосую критерії нормальности Shapiro-Wilk для перевірки.





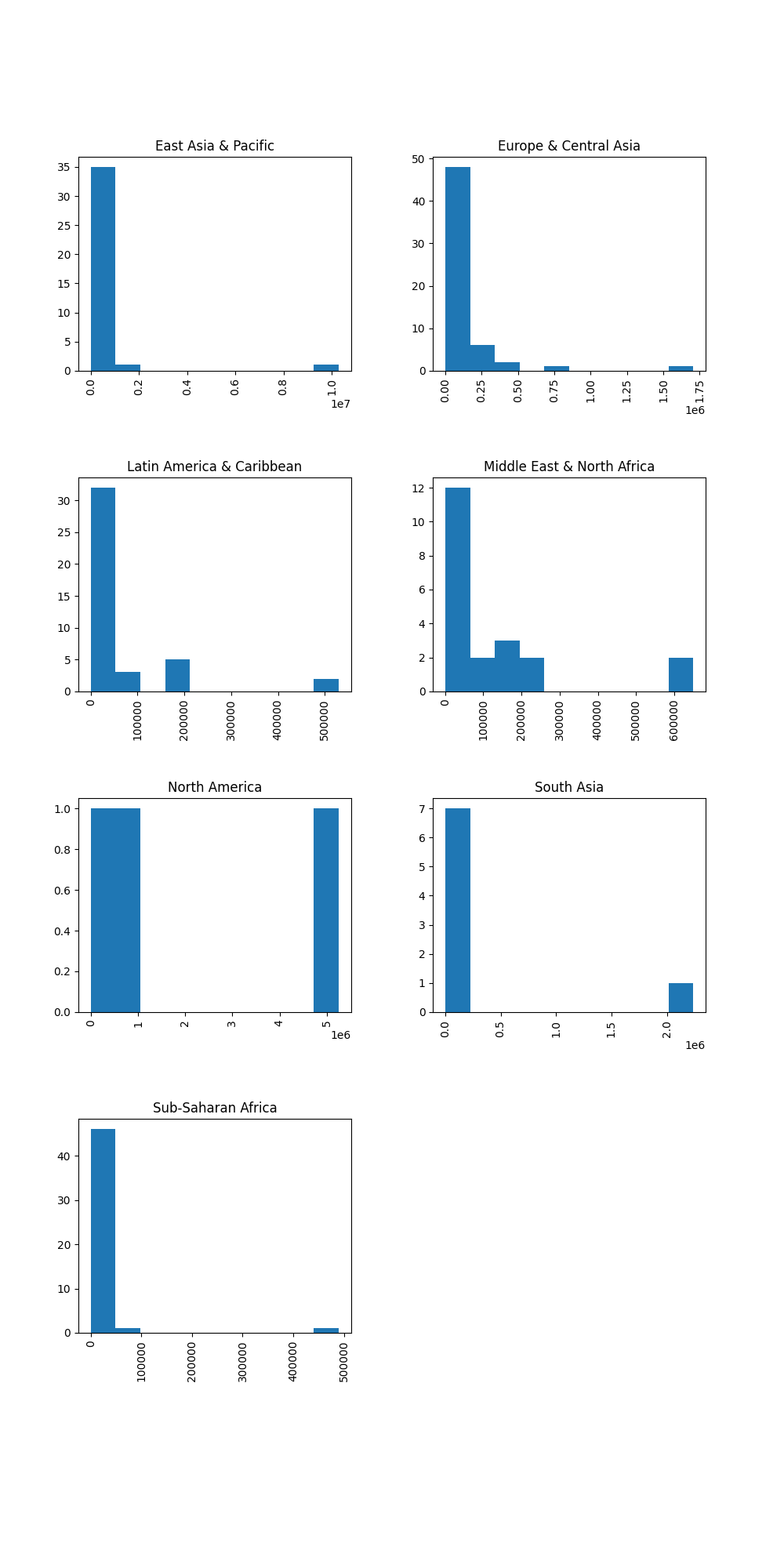
Переконуємося, що розподіл величин не відповідає нормальному.

1. Перевірити гіпотезу про рівність середнього і медіани для одного з параметрів; Для цього я застосую одновибірковий T-test (stats.ttest\_1samp())

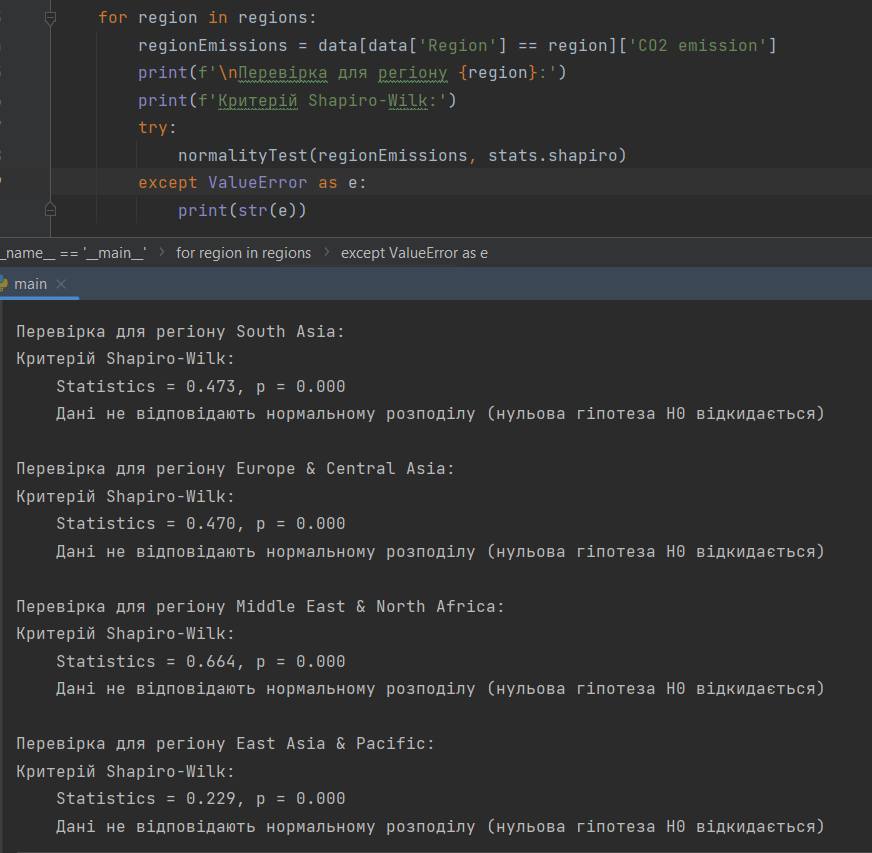


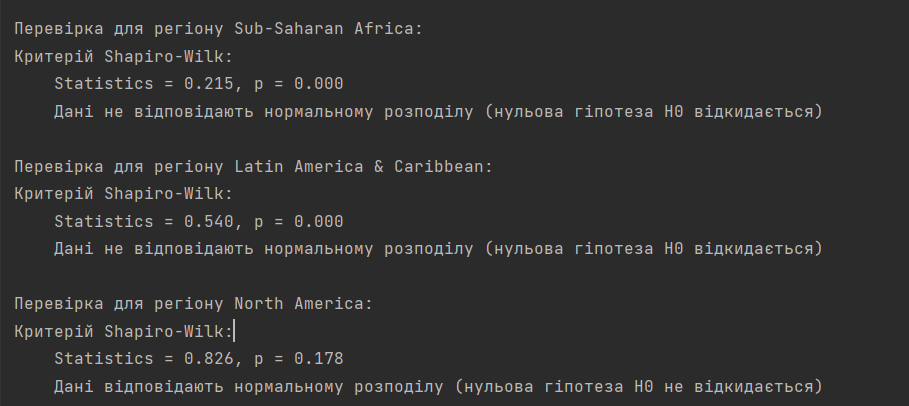
Бачимо, що в жодному стовпці гіпотеза не підтверджується.

1. Вказати, в якому регіоні розподіл викидів СО2 найбільш близький до нормального; Для цього так само побудуємо гістограми, а також переконаємося за допомогою критеріїв Shapiro-Wilk.



Візуально бачимо, що розподіли не є нормальними. Перевіримо.

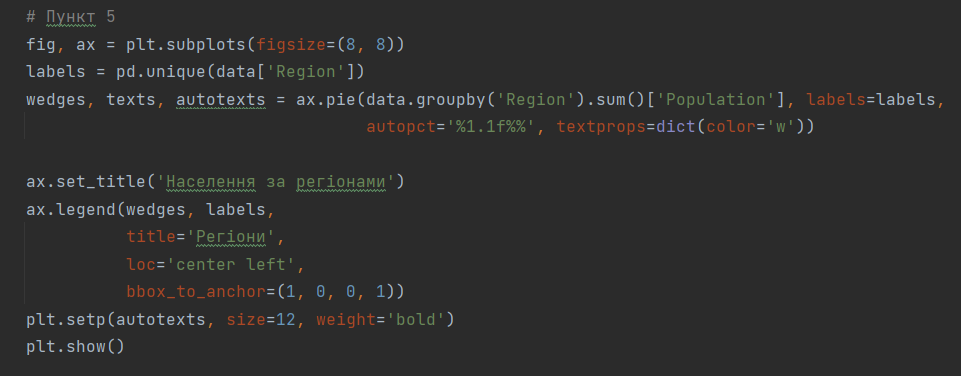


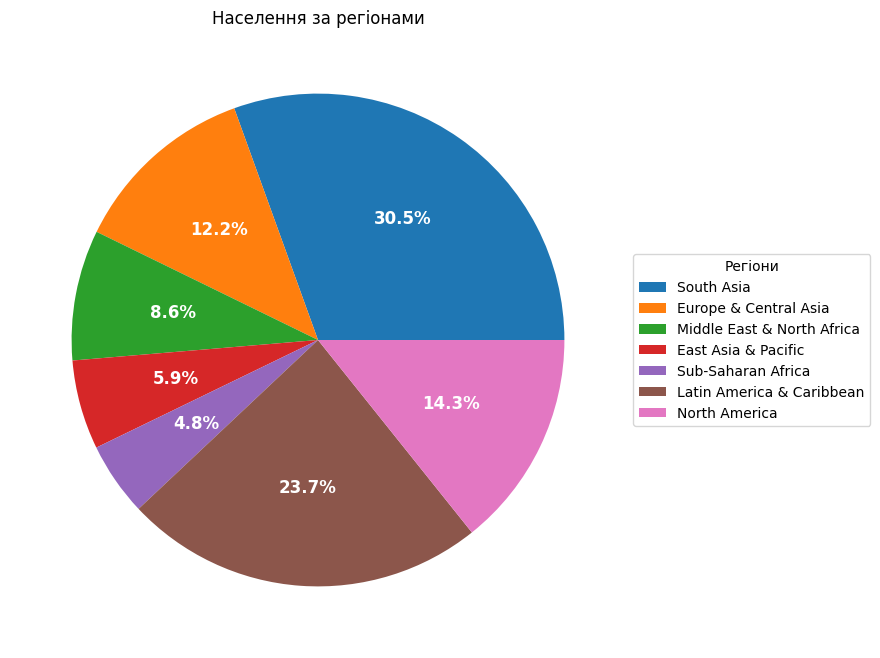


Бачимо, що за критерієм Shapiro-Wilk у Північній Америці розподіл найближчий до нормального, але оскільки у цьому регіоні всього 3 країни, то вибірка надто мала і можлива помилка у перевірці гіпотези. Тож розподіли не є нормальними.

1. Побудувати кругову діаграму населення по регіонам.

Для цього скористаюся функцією pie().





Отримали відповідну діаграму.

**Додаткове завдання.**

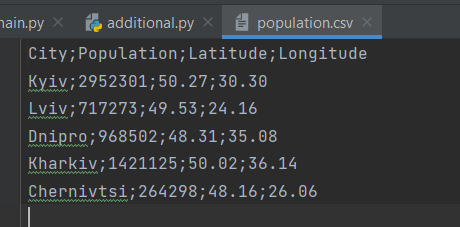
Завдання 1

1. Завантажити карту України Ukraine.jpg

Для даного завдання я використовуватиму модуль geopandas, який не підтримує розширення .jpg, тож замість такої карти завантажую карту UKR\_ADM1.shp.

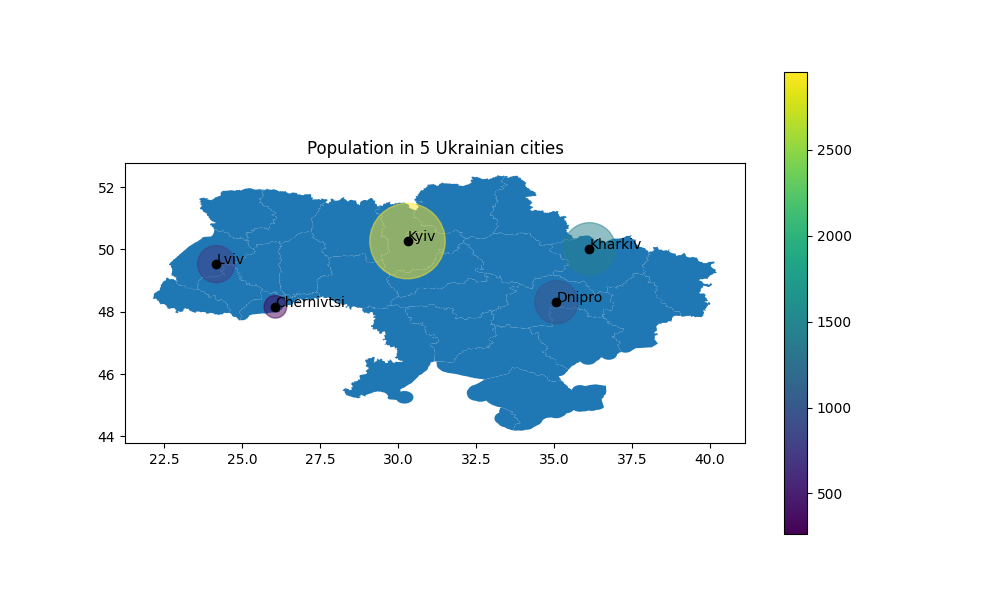


1. Розмістити бульбашки, що відповідають їх населенню, на довільних 5 містах (статистику взяти в інтернеті). Для цього завдання я знайшов статистику по 5 містах і створив власний невеликий csv-файл.





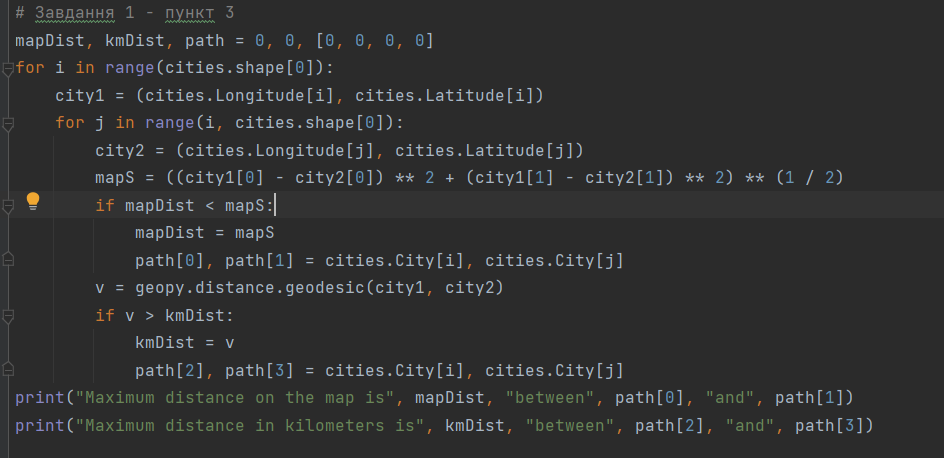
Завантажую файл, за допомогою широти і довготи створюю GeoDataFrame з колонкою розташування. Далі на графіку розміщую завантажену карту України, ділю населення кожного міста на тисячу, щоб бульбашки не були надто великі і розташовую бульбашки на графіку з розміром відповідно до населення. Потім підписую відповідні міста. Отримуємо такий результат.

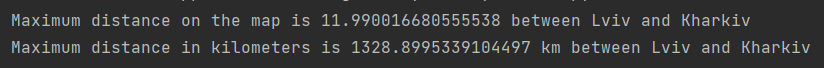


Бачимо, що результат коректний як графічно за розташуванням, так і за кількістю населення. З вибірки у Києві населення найбільше, у Чернівцях – найменше.

1. Знайти найбільшу відстань між містами в пікселях та кілометрах.

Оскільки відстань у пікселях за допомогою використовуваних програмних засобів знайти неможливо, то я буду знаходити відстань ніби на карті, використовуючи широту і довготу як координати точок, і в кілометрах, використовуючи geopy.distance.geodesic, який враховує форму Землі і подає максимально точні дані.

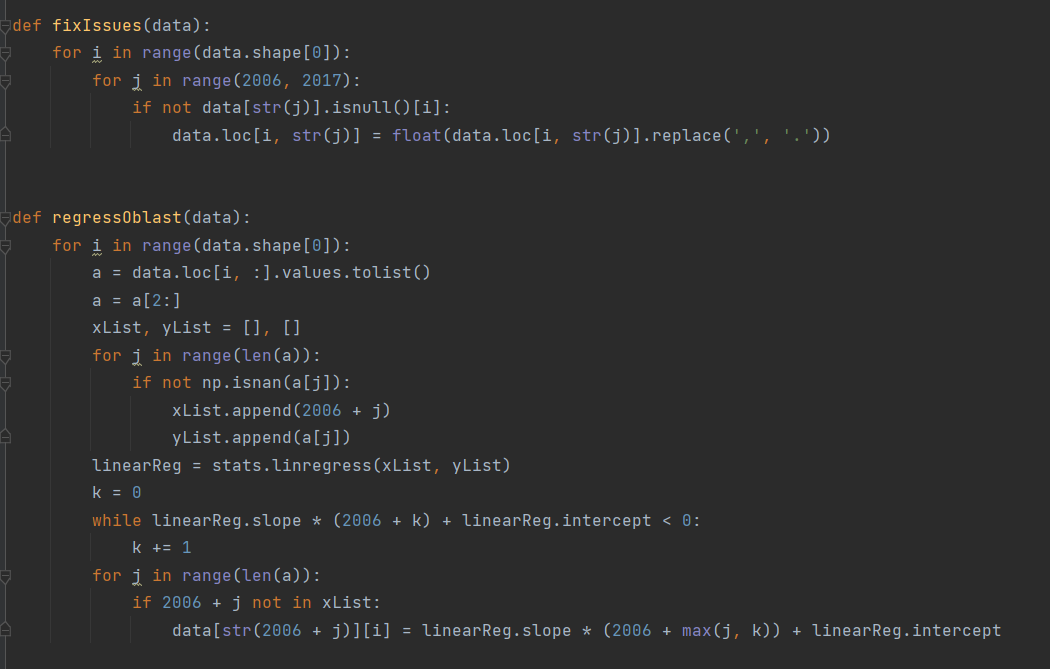


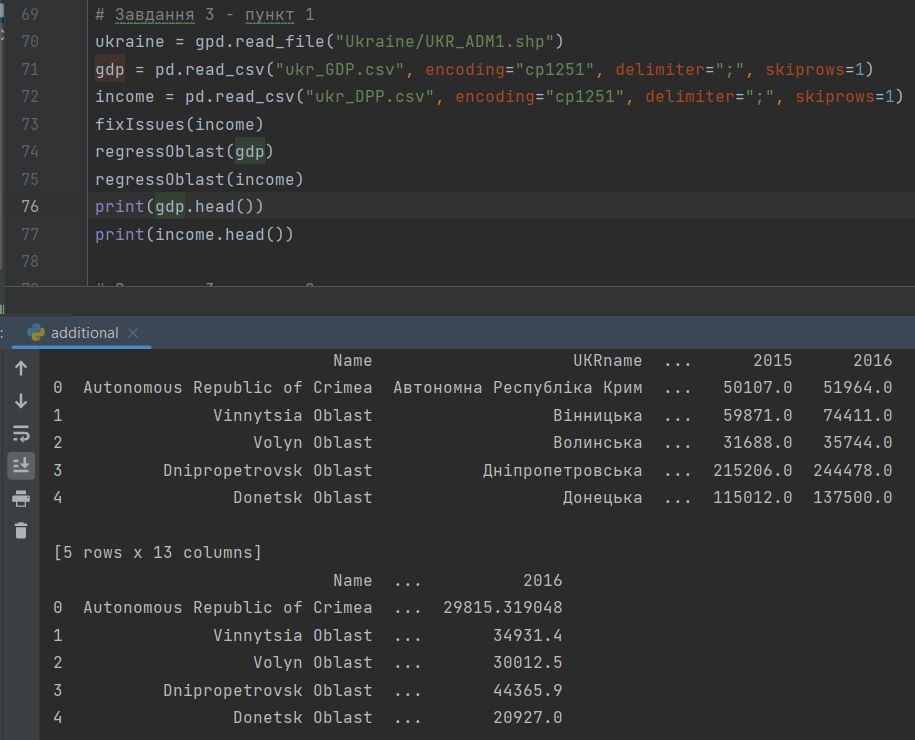


Завдання 3

1. Завантажити shape-файл с областями України.

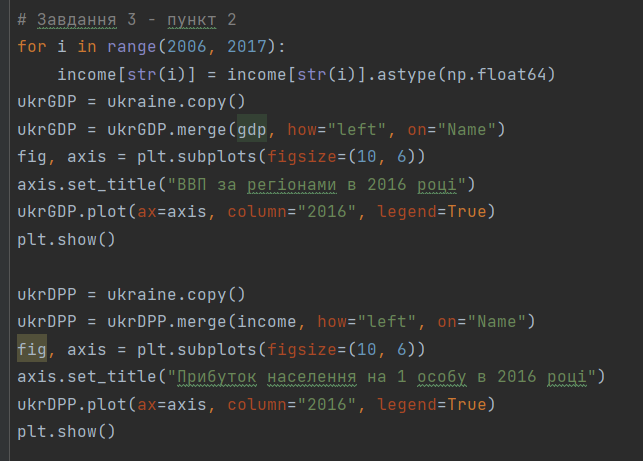
Для наступних завдань потрібно виправити помилки і заповнити пропущені значення. Заміняю коми в числових даних на крапки, змінюю тип даних на float. Для відсутніх даних я скористаюся лінійною регресією. Для кожного рядка я створю список тих значень, які існують, і на їх основі отримаю параметри регресійної прямої. Після цього усі відсутні значення відповідно заміню на прогнозовані. Якщо таке значення виходить менше 0, то я заміняю на перше додатне значення. У такому випадку вважаю, що дані змінюються настільки несуттєво щорічно, що це особливо не впливає на коректність даних.



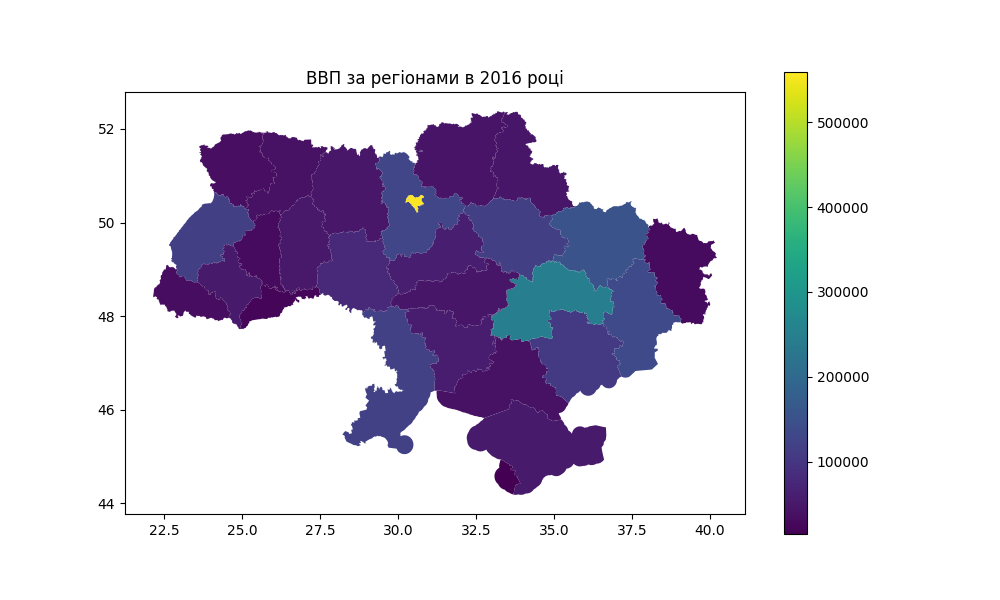


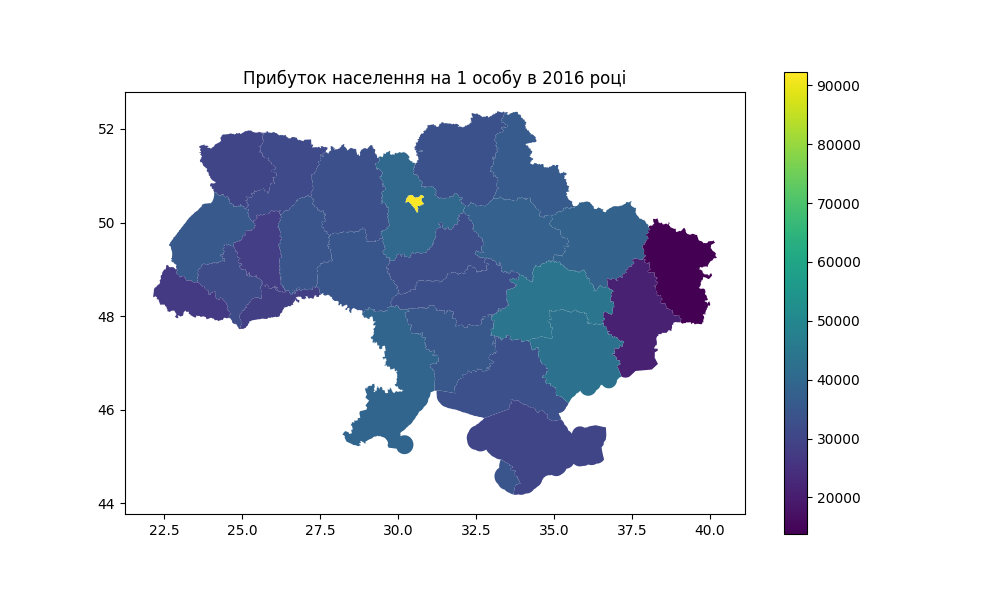
1. Побудувати картограми для прибутку населення на 1 особу і ВВП по регіонам за 2016 рік.

Для цього я роблю ліве з’єднання карти України й даних за назвою области і відповідно будую картограму за колонкою 2016 року.



Отримуємо такі картограми.

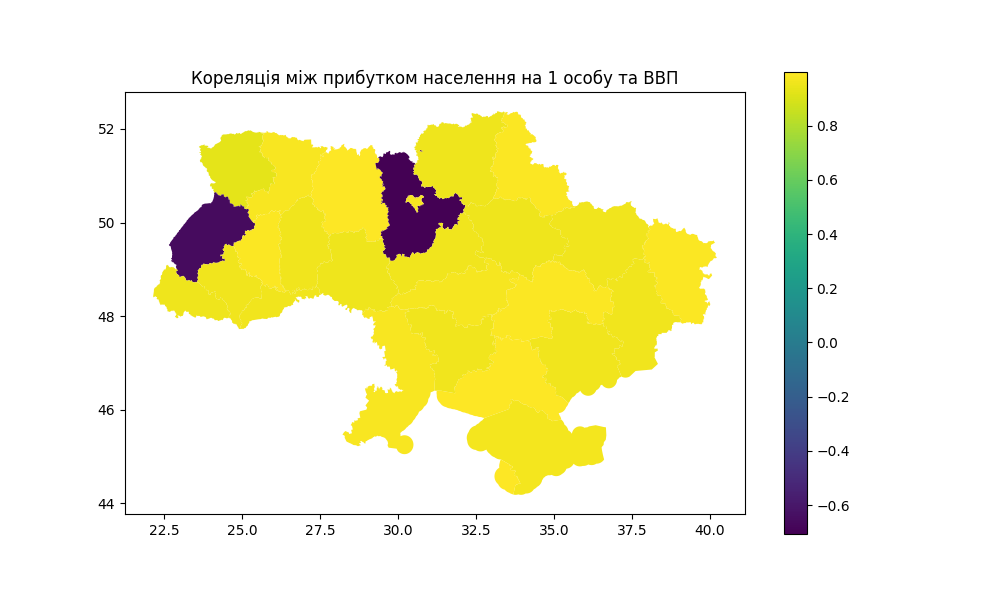




Бачимо, що найбільше ВВП було в Києві та Дніпропетровській області, у решті областей воно нижче. Прибуток так само найбільший в Києві, Дніпропетровській та Запорізькій областях. У решті областей прибуток менший, але особливо виділяється Донецька та Луганська області, частина яких була окупована ще на той момент і постраждала від бойових дій. Дані по Криму й Севастополю спрогнозовані, тому не виділяються серед інших регіонів.

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

Для цього я використовую функцію corrwith(), відповідно отримую кореляцію між прибутком 2006-2015 та ВВП за регіонами 2006-2015. Додаю стовпець кореляції до мапи і на її основі будую картограму.



Отримуємо такі результати. Бачимо, що коефіцієнт кореляції високий, тобто є залежність прибутків від ВВП. Є відхилення у Київській та Львівській областях. У Київській області це можна пояснити високим рівнем розвитку через близькість до столиці, але все ж невисокими звичайними регіональними доходами. Або ж це просто статистична аномалія.

**Висновок.**

Отже, у цій роботі я ознайомився з методами визначення точкових оцінок параметрів розподілу; дослідив, що впливає на якість точкових оцінок; методикою визначення інтервальних оцінок параметрів розподілу; дослідив, що впливає на якість інтервальних оцінок; методами перевірки статистичних гіпотез про вигляд закону розподілу; дослідив, що впливає на ширину критичної області.

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. Було проаналізовано дані країн на нормальний розподіл, виявилося, що жоден стовпець не розподілений нормально. За допомогою кругової діаграми наочно переконалися, що найбільша кількість населення – в Азії, оскільки маємо такі країни, як Китай, Індія та інші. У додаткових завданнях було проаналізовано ВВП і доходи населення за регіонами України, побачили, що між ними є кореляція і те, що деякі регіони відстають в економічному розвитку (війна, ближче\далі до економічних центрів тощо). Використовуючи PyCharm, отримуємо коректний результат.