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 № 5 з дисципліни

«Основи науки про дані»

«Розв’язування задачі кластеризації»

Виконав студент \_\_\_\_ІП-11 Трикош І. В\_\_\_\_

Перевірив \_\_\_Новотарський М. А.\_\_\_

Київ 2024

**Завдання**

1. Детально розглянути наведені теоретичні відомості до даної лабораторної роботи.

2. Ознайомитися з кодом, що наведений у Jupyter notebook DD\_lab5.ipynb.

3. Завантажити з мережі Інтернет набір даних з розширенням \*.csv, \*.txt, \*.pdf або \*.json який орієнтований на вирішення задачі кластеризації методом k-середніх.

4. Вирішити задачу кластеризації, спираючись на приклад, який описаний у теоретичних відомостях та наведений у ноутбуці DD\_lab5.ipynb.

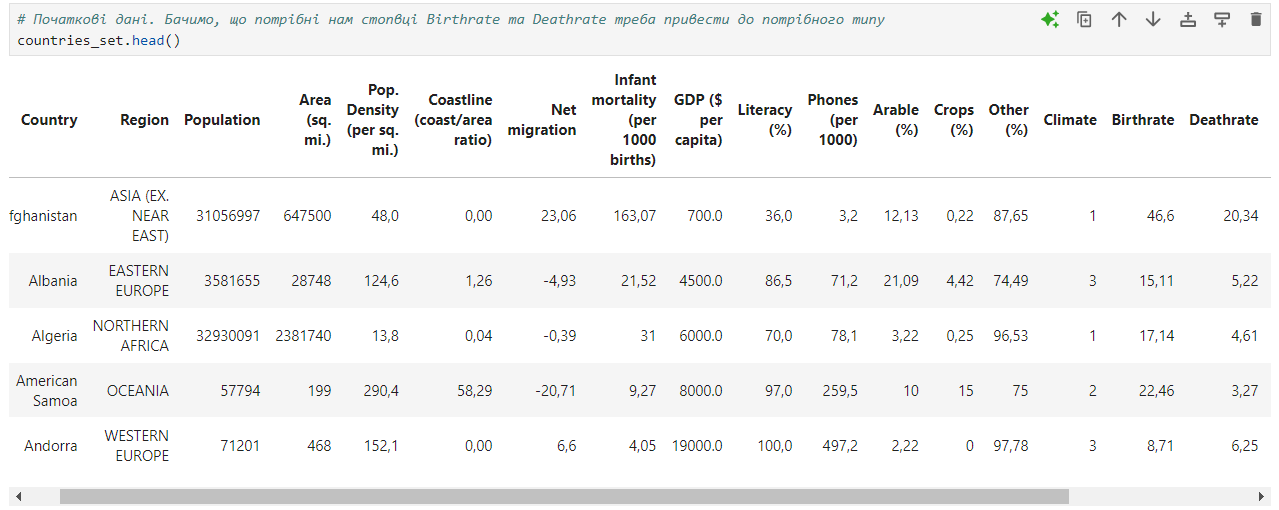
**Виконання**

Для виконання лабораторної роботи візьмемо датасет про країни світу, що містить такі дані, як народжуваність, смертність, населення, площа, тощо: <https://www.kaggle.com/datasets/fernandol/countries-of-the-world>

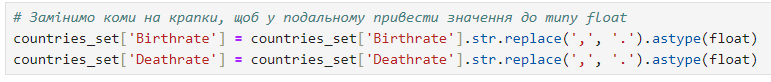
Зчитуємо дані з csv файлу:



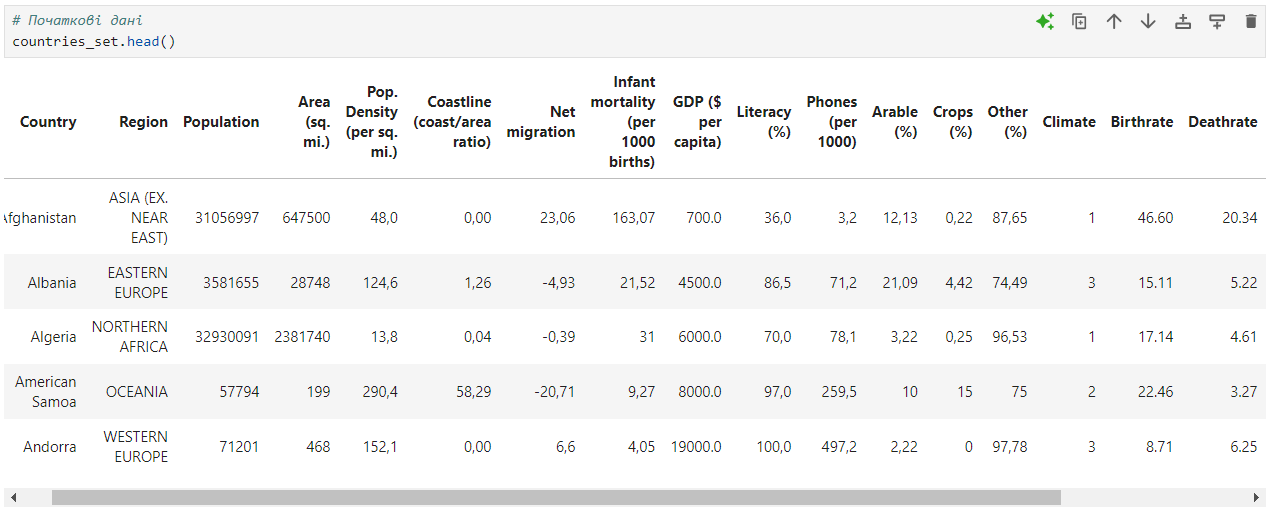
Виведемо перші 5 рядків даних з допомогою методу head. Бачимо, що потрібні нам стовпці Birthrate та Deathrate треба привести до типу float:



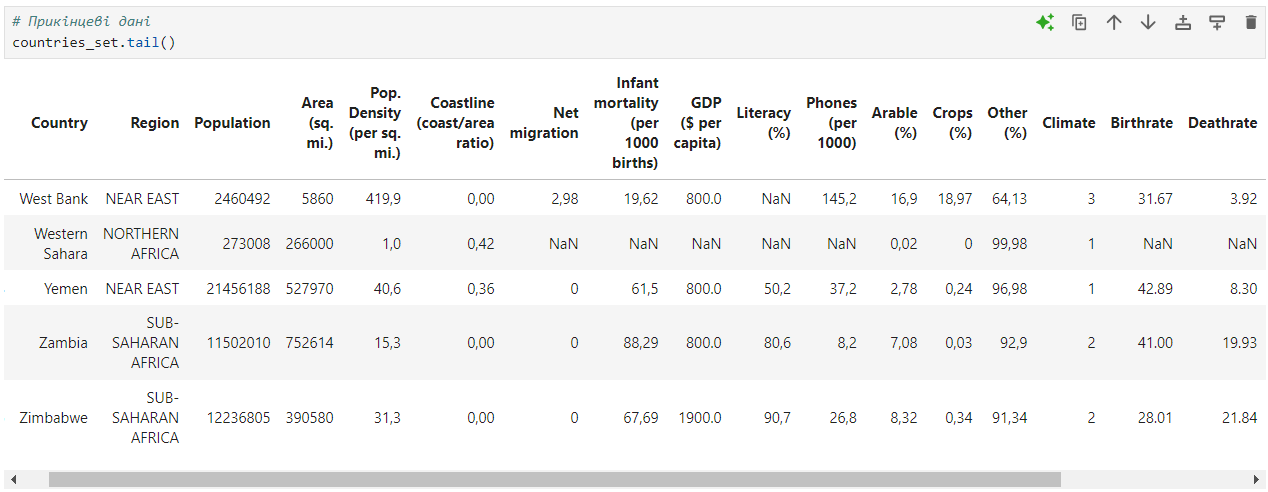
Замінимо у цих стовпцях коми на крапки, що перетворити дані до типу float, використовуючи метод replace (перший параметр методу – рядок, який треба замінити, другий – рядок, на який треба замінити) для заміни символів і метод astype, який приведе значення стовпця до вказаного типу:



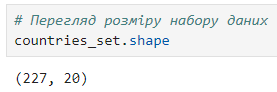
Виведемо перші 5 рядків даних з допомогою методу head. Бачимо, що тепер усе добре, дані потрібного типу:



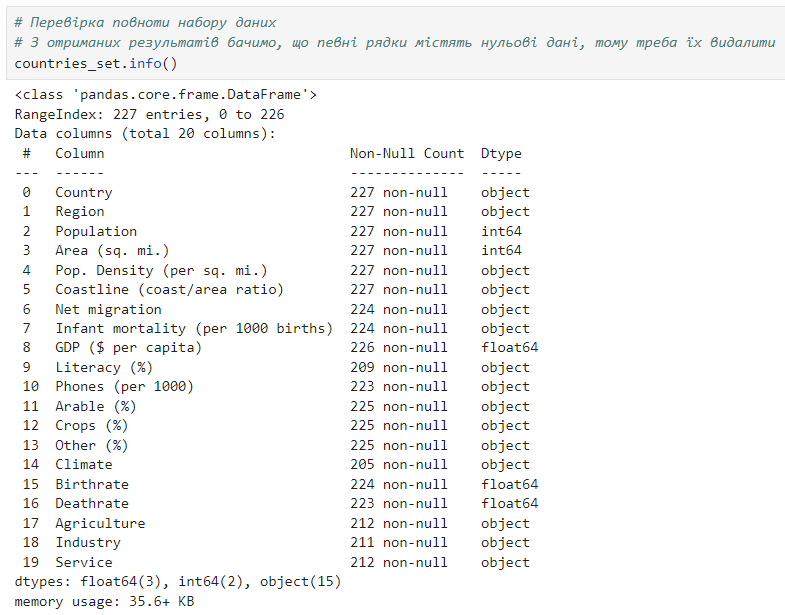
Виведемо останні 5 рядків даних з допомогою методу tail:



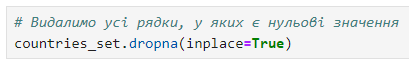
Переглянемо розмір набору даних, застосувуючи атрибут датафрейму shape:



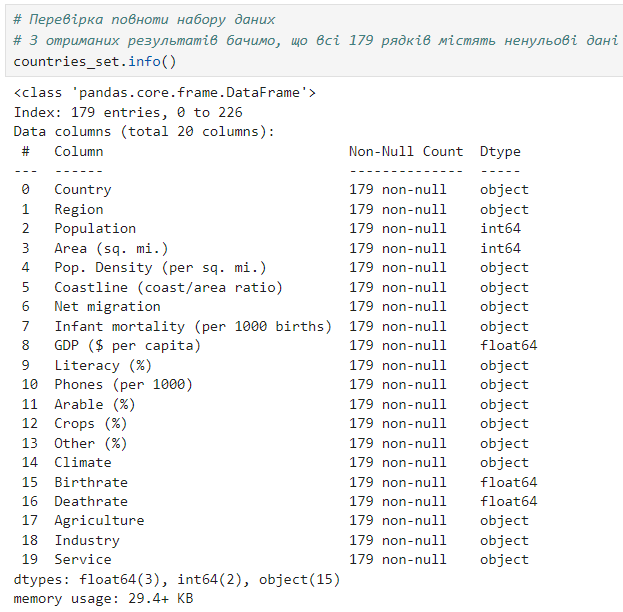
Переглянемо інформацію про датасет з допомогою методу info. Бачимо, що датасет містить пропущені значення:



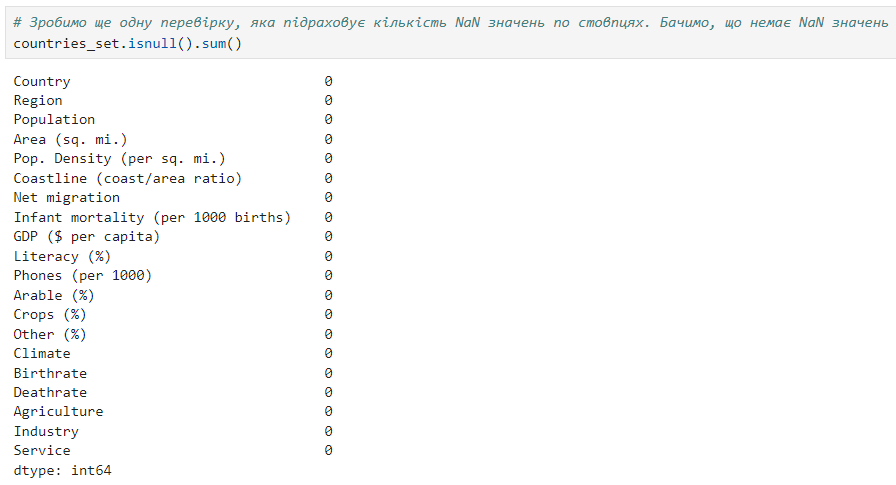
Видалимо усі рядки, які містять NaN значення з допомогою методу dropna:



Знову переглянемо інформацію про датасет з допомогою методу info. Бачимо, що тепер датасет не містить пропущений значень:



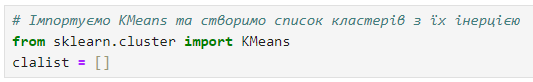
Виконаємо остаточну перевірку на NaN значення, підрахувавши їх кількість у кожному стовпці, використовуючи методи isnull (для перевірки на NaN) та sum (для сумування по стовпцях). Бачимо, що NaN значень немає, тому можна продовжити роботу:



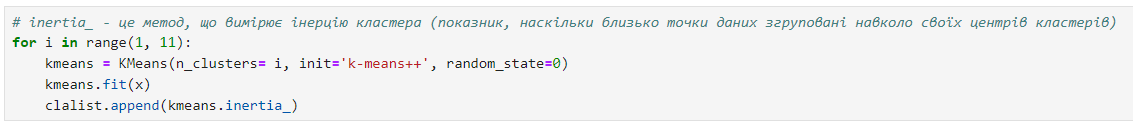
Тепер треба обрати ознаки, за якими буде відбуватися кластеризація. Було обрано стовпці Birthrate та Deathrate, використовуючи iloc та параметр values для отримання numpy масиву значень:



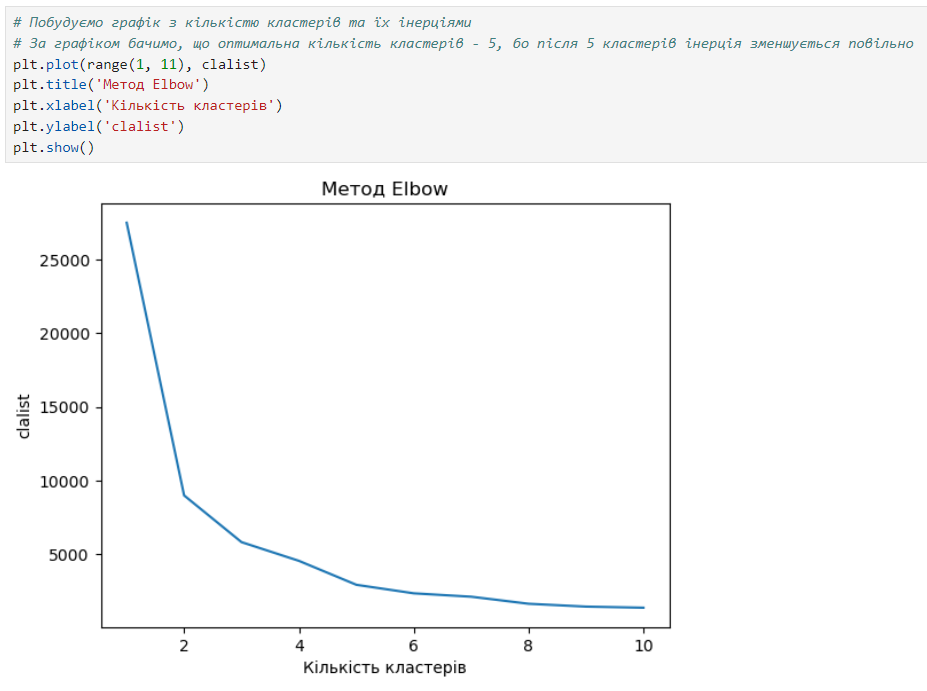
Імпортуємо модель KMeans з бібліотеки sklearn та створимо список кластерів з інерціями:



Визначимо оптимальну кількість кластерів. Для цього спочатку визначимо значення інерції (це показник того, наскільки близько точки даних згруповані навколо своїх центрів кластерів) для кількості кластерів від 1 до 10 включно. Кластери визначатимемо з допомогою класу KMeans, який приймає параметри кількості кластерів, стратегію ініціалізації центрів кластерів та значення генератора випадкових чисел. Тренуватимемо модель з допомогою методу fit, який приймає тренувальні дані як параметр. Інерцію кластера визначимо з допомогою методу inertia\_:



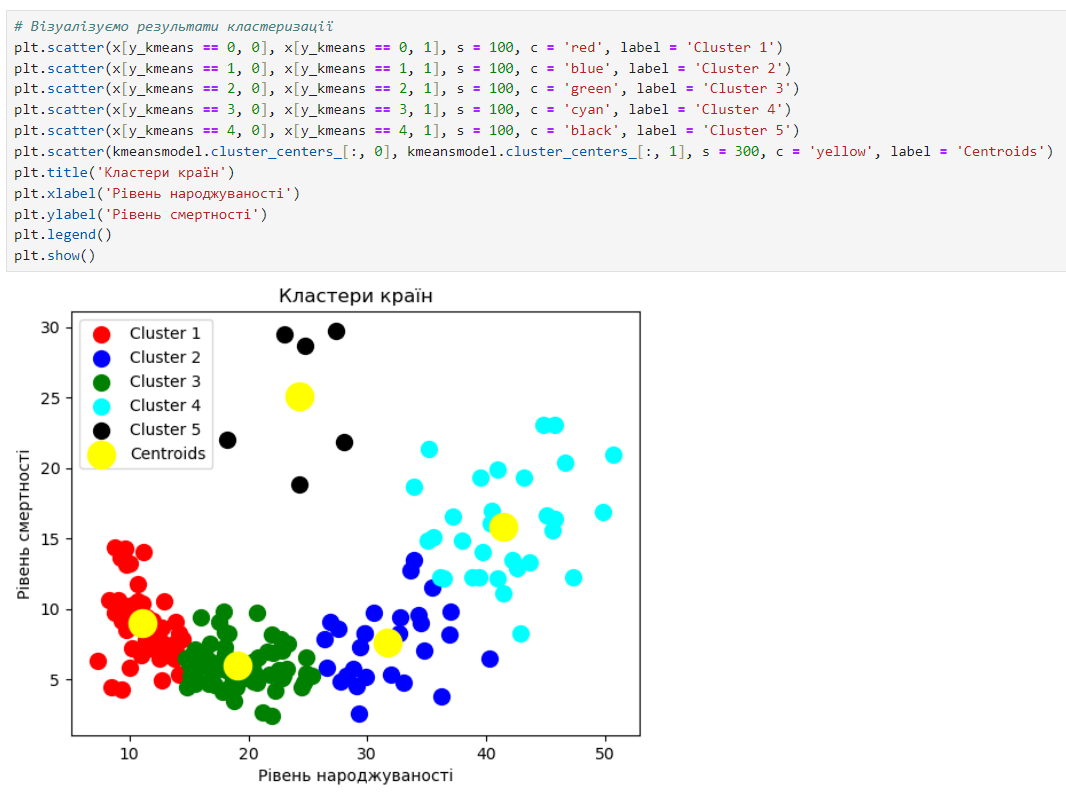
Тепер визначимо оптимальну кількість кластерів. Для цього побудуємо графік значень інерції для різної кількості кластерів. По графіку бачимо, що після 5 кластерів значення інерції повільно зменшується, а отже це оптимальне значення:



Створимо і натренуємо модель з оптимальною кількістю кластерів, використовуючи метод fit\_predict, адже в нас навчання без учителя:



Візуалізуємо результати кластеризації:



Пояснення такого розбиття:

Кластер 1 – сюди потрапили країни приблизно однаковим рівнем народжуваності та смертності.

Кластер 2 – сюди потрапили країни з високим рівнем народжуваності та порівняно нижчим рівнем смертності.

Кластер 3 – сюди потрапили країни з середнім рівнем народжуваності та низьким рівнем смертності.

Кластер 4 – сюди потрапили країни з високим рівнем народжуваності та високим рівнем смертності.

Кластер 5 – сюди потрапило лише декілька країн, що мають середній рівень народжуваності та високий рівень смертності.

**Висновки**

Під час виконання лабораторної роботи я навчився працювати з простим алгоритмом кластеризації k-means, виконавши навчання без учителя на даних про народжуваність та смертність у країнах світу.