**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 №6 з дисципліни

«Аналіз даних в інформаційних системах»

„**Класифікація та кластеризація**”

**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](#_Toc131726251)

[2 Завдання 4](#_Toc131726252)

[2.1 Основне завдання 4](#_Toc131726253)

[2.2 Додаткове завдання 4](#_Toc131726254)

[3 Виконання основного завдання 5](#_Toc131726255)

[4 Виконання додаткового завдання 15](#_Toc131726256)

[4.1 Визначити, який регіон домінує в кластерах по ВВП на душу населення та щільності населення 15](#_Toc131726257)

[4.2 Вивести частотні гістограми всіх показників файла Data2.csv, використовуючи цикл 28](#_Toc131726258)

[4.3 Створити функцію, яка на вхід отримує два набори даних, перевіряє чи є лінійна залежність та виводить True чи False (будемо розуміти під «є лінійна залежність», якщо коефіцієнт кореляції по модулю більше 0,8) 29](#_Toc131726259)

[5 Висновок 31](#_Toc131726260)

# Мета лабораторної роботи

Мета роботи – ознайомитись з методами класифікації та кластеризації; моделями, що використовують дерева прийняття рішень; інструментами факторного аналізу методом головних компонент та методом найбільшої подібності.

# Завдання

## Основне завдання

Для даних по титаніку titanic.csv побудувати модель, в якій можна визначити, чи виживе пасажир, заповнивши решту параметрів.

Використати декілька методів. Порівняти результати.

## Додаткове завдання

Використовуючи файл Data2.csv

1. визначити, який регіон домінує в кластерах по ВВП на душу населення та щільності населення
2. вивести частотні гістограми всіх показників файла Data2.csv, використовуючи цикл
3. створити функцію, яка на вхід отримує два набори даних, перевіряє чи є лінійна залежність та виводить True чи False (будемо розуміти під «є лінійна залежність», якщо коефіцієнт кореляції по модулю більше 0,8)

# Виконання основного завдання

Для початку завантажимо дані, перевіримо, чи є серед них пропущені.

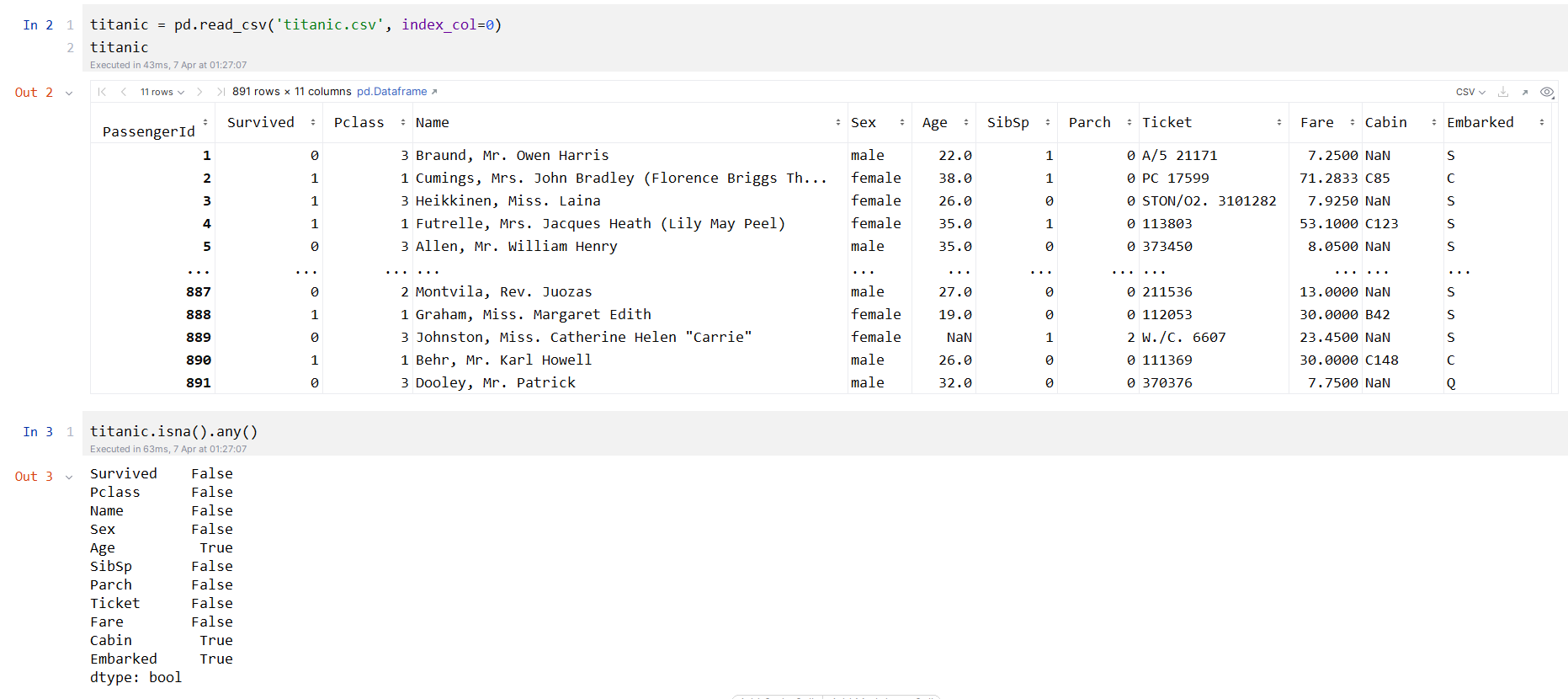


Рис. 3.1 – Завантаження даних, перевірка на пропущені значення

Бачимо, що є пропущені значення в стовпцях «Cabin», «Embarked» та «Age». Відкинемо рядкові стовпці, які ні на що не впливають.

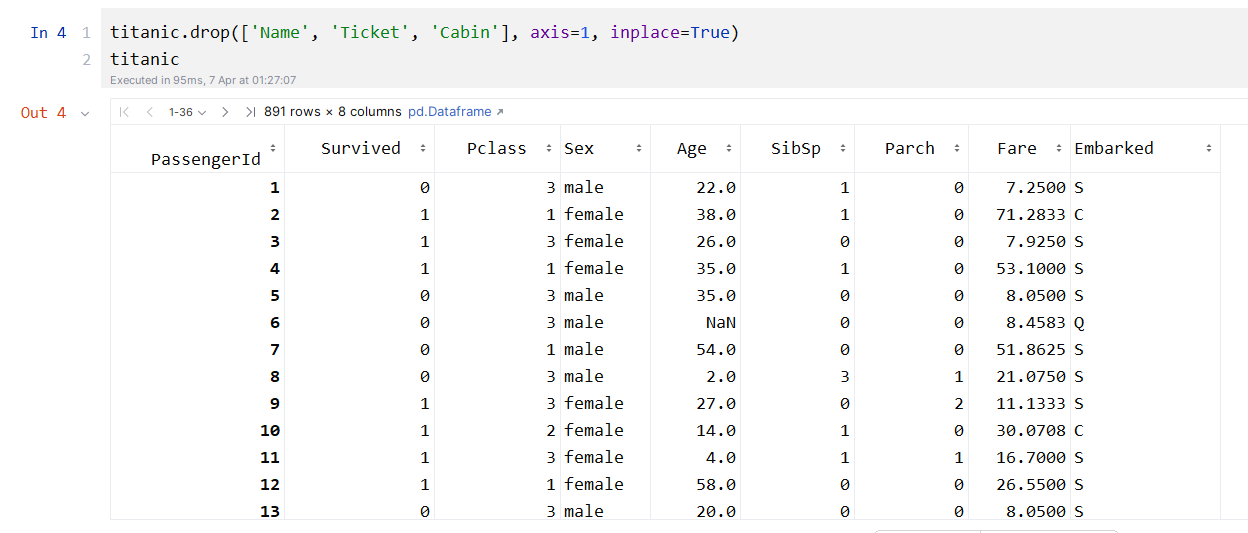


Рис. 3.2 – Відкидання зайвих стовпців

Тепер перевіримо, скільки саме значень пропущено в стовпці «Embarked».

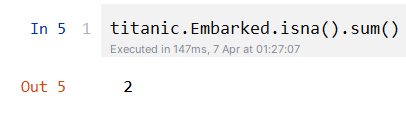


Рис. 3.3 – Кількість пропущених значень у стовпці «Embarked»

Бачимо, що таких усього 2, тому пропонується заповнити їх тим, що найчастіше повторюється. Виведемо кількість повторів кожного зі значень.

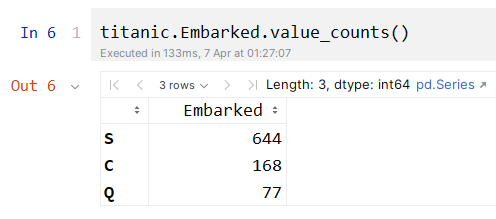


Рис. 3.4 – Кількість кожного значення в стовпці «Embarked»

Найчастіше повторюється значення «S», а отже заповнимо пропуски таким значенням.

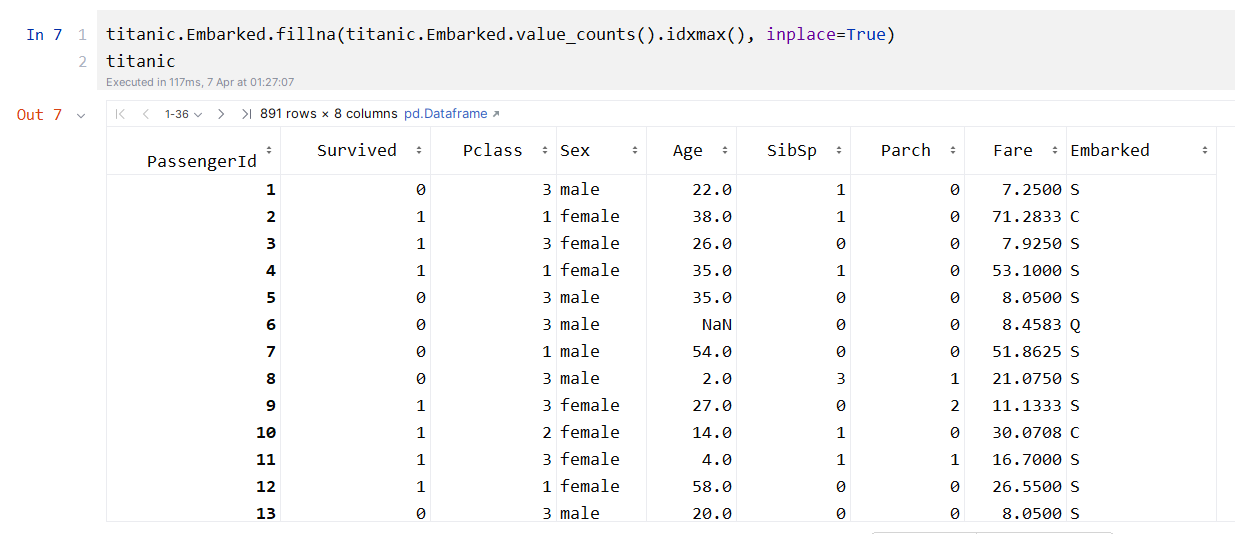


Рис. 3.5 – Заповнення пропущених значень стовпця «Embarked»

Далі займемося заповненням пропусків віку. Можемо припустити, що для чоловіків та жінок середній вік є різним, а отже обчислимо його.

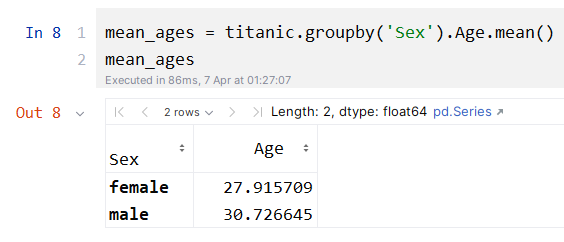


Рис. 3.6 – Розрахунок середнього віку в залежності від статі

Припущення справдилося, а тому заповнюватимемо пропуски в залежності від віку. Наостанок перевіримо, чи залишилися десь пропущені значення.

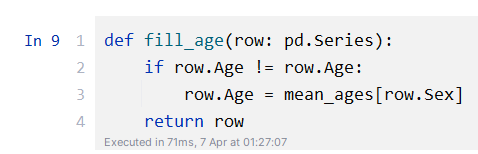


Рис. 3.7 – Функція заповнення пропусків у стовпці віку

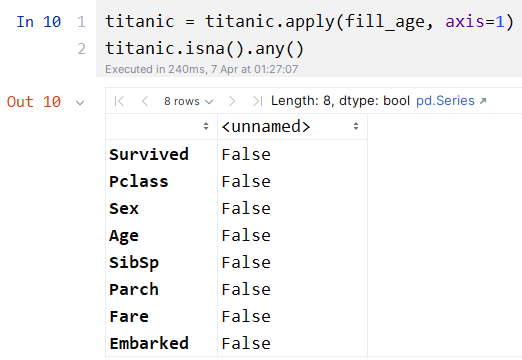


Рис. 3.8 – Заповнення пропущених значень у стовпці віку, перевірка на наявність пропущених значень у наборі даних

Далі, оскільки в нас є категоріальні значення («Sex», «Embarked»), використаємо метод pandas.get\_dummies(), який перетворить ці значення в матрицю з бінарними змінними.

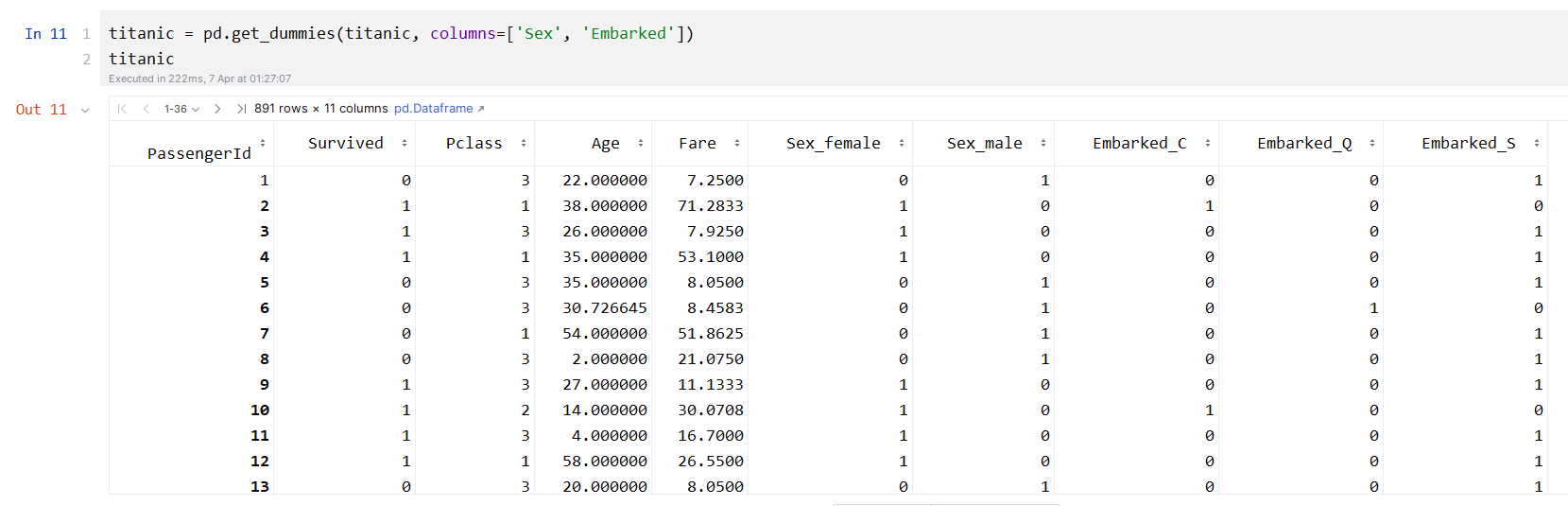


Рис. 3.9 – Перетворення категоріальних значень на матрицю з бінарними змінними

Після цього відобразимо теплову карту кореляції.

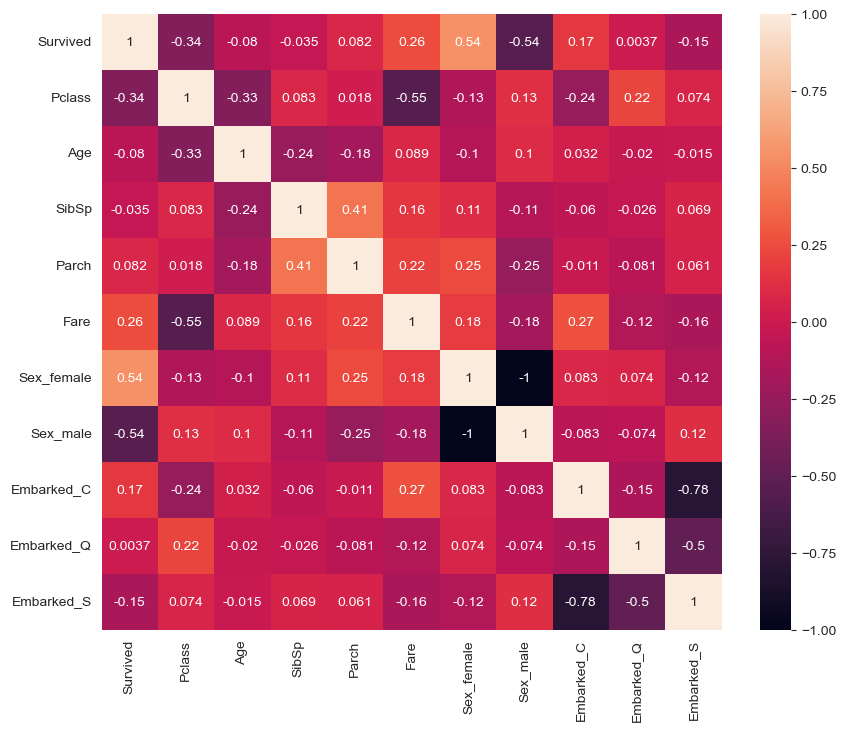


Рис. 3.10 – Теплова карта кореляції

Бачимо, що стовпці «SibSp», «Parch» мало впливають на те, чи виживе пасажир (коефіцієнти кореляції -0.035 та 0.082 відповідно), а отже їх можна відкинути.

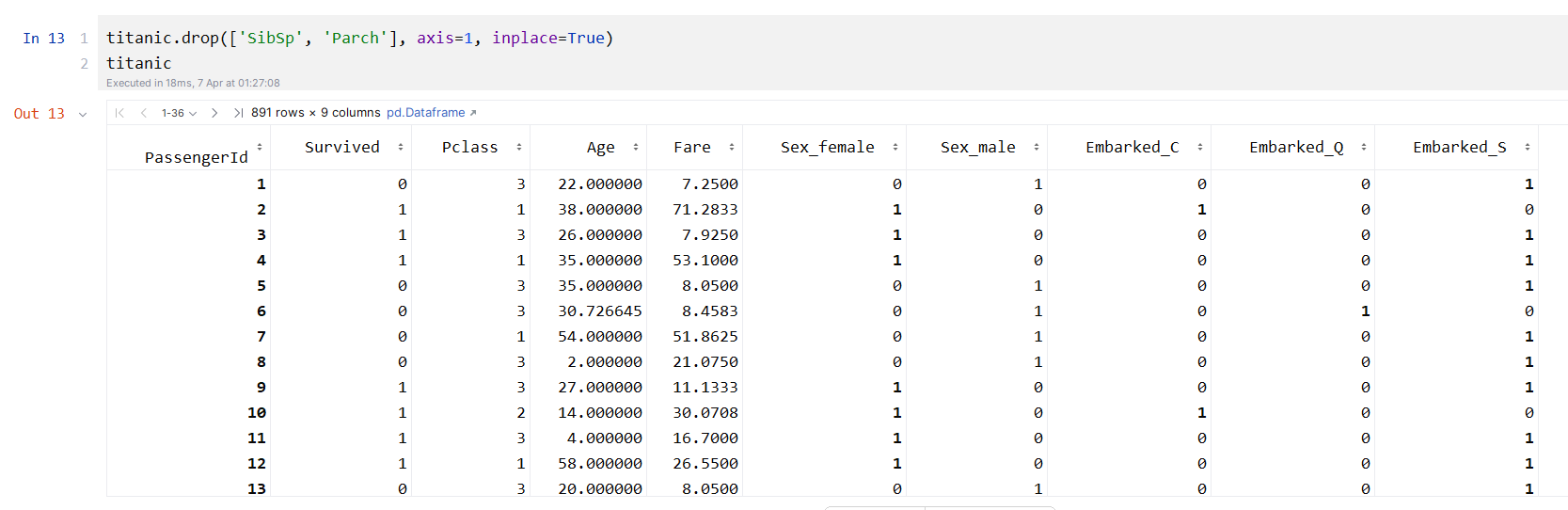


Рис. 3.11 – Видалення незначущих стовпців

Далі розділимо дані на навчальну та тестову вибірки в пропорції 75% на 25%.



Рис. 3.12 – Розділення набору даних на навчальну та тестову вибірки

Після того, як усі описані дії проведено, можемо приступити до процесу класифікації. Використовуватимуться наступні методи: метод K-найближчих сусідів, дерево рішень та випадкові ліси. Для кожного з цих методів за допомогою крос-валідації підберемо найкращі гіперпараметри. Почнемо з методу K-найближчих сусідів.

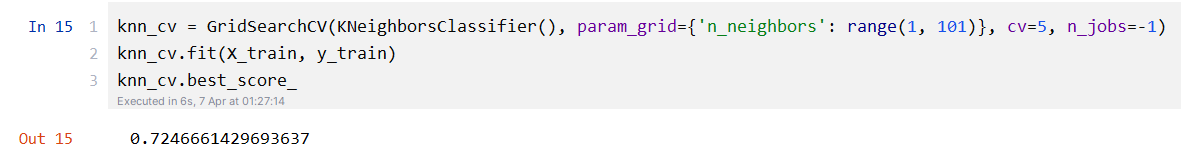


Рис. 3.13 – Крос-валідація для знаходження гіперпараметру К методу К-найближчих сусідів

Бачимо, що оцінка на навчальній вибірці склала 0.724, накреслимо ці оцінки на графіку залежності оцінки від кількості сусідів.

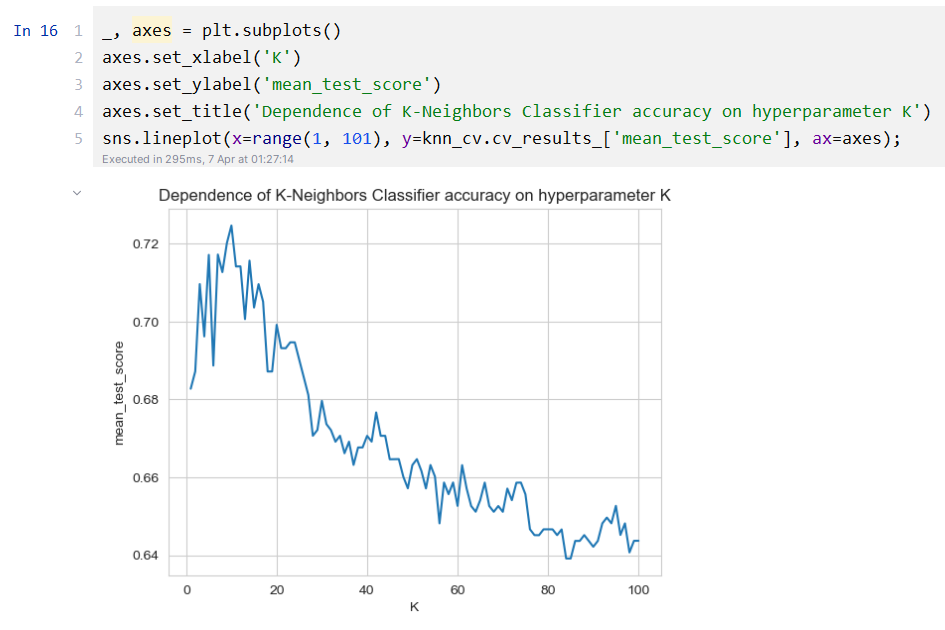


Рис. 3.14 – Графік залежності точності методу від гіперпараметру К

Далі визначимо середньоквадратичну похибку на тестовій вибірці.

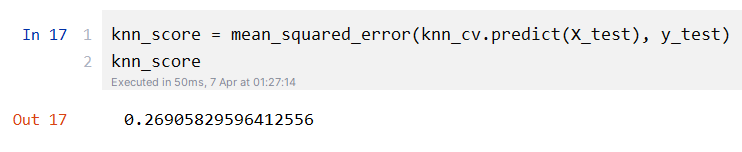


Рис. 3.15 – Середньоквадратична похибка на тестовій вибірці для методу К-найближчих сусідів

Вона склала 0.269, переходимо до методу дерев рішень.

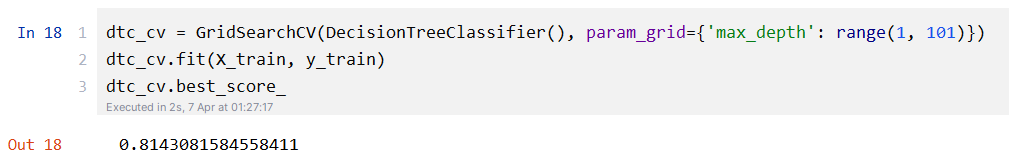


Рис. 3.16 – Крос-валідація для підбору значення максимальної глибини дерева рішень

Тут оцінка склала вже 0.814, накреслимо її на графіку залежності оцінки від максимальної глибини.

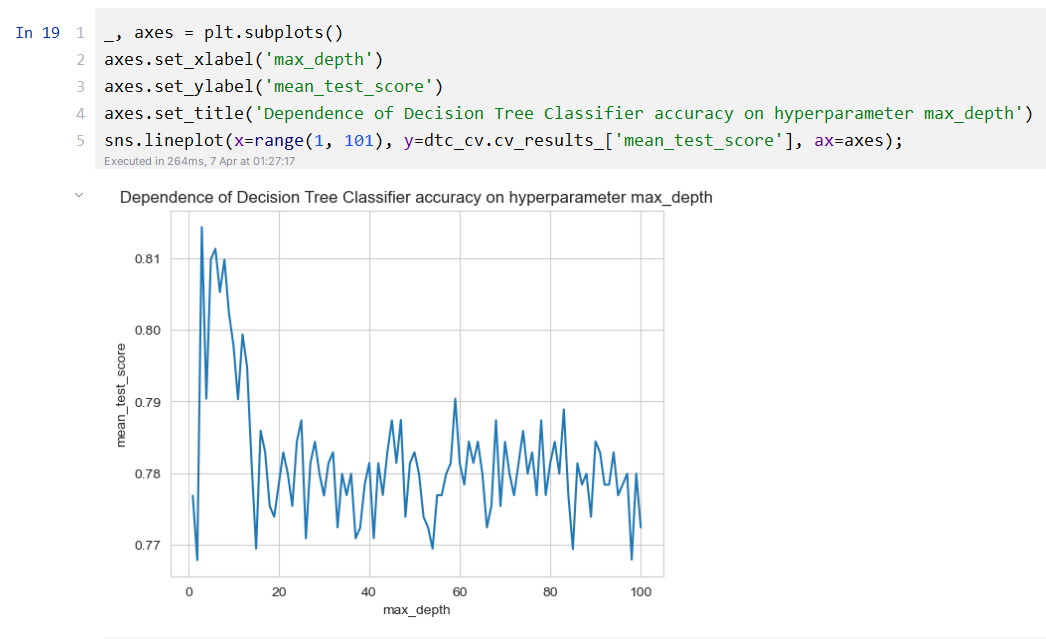


Рис. 3.17 – Графік залежності точності дерева прийняття рішень від максимальної глибини

Тепер обрахуємо середньоквадратичну похибку для цього методу.

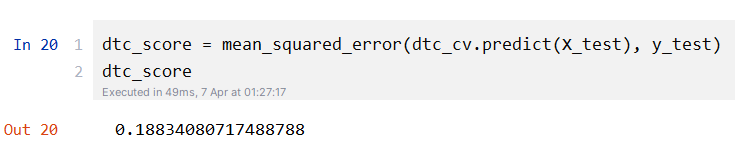


Рис. 3.18 – Середньоквадратична похибка на тестовій вибірці для дерева рішень

Вона склала 0.188, що є доволі хорошим результатом. Перейдемо до випадкових лісів.

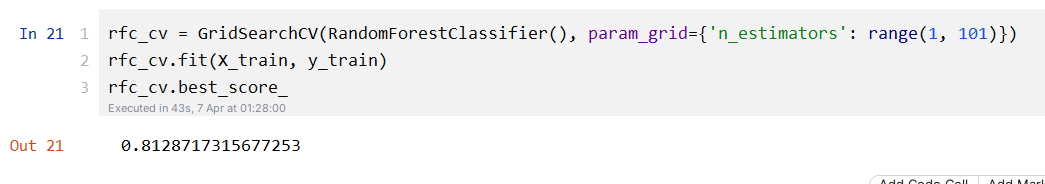


Рис. 3.19 – Крос-валідація для підбору гіперпараметру кількості дерев у методі випадкових лісів

Тут отримали оцінку на навчальній вибірці 0.812. Відобразимо залежність оцінки від кількості дерев на графіку.

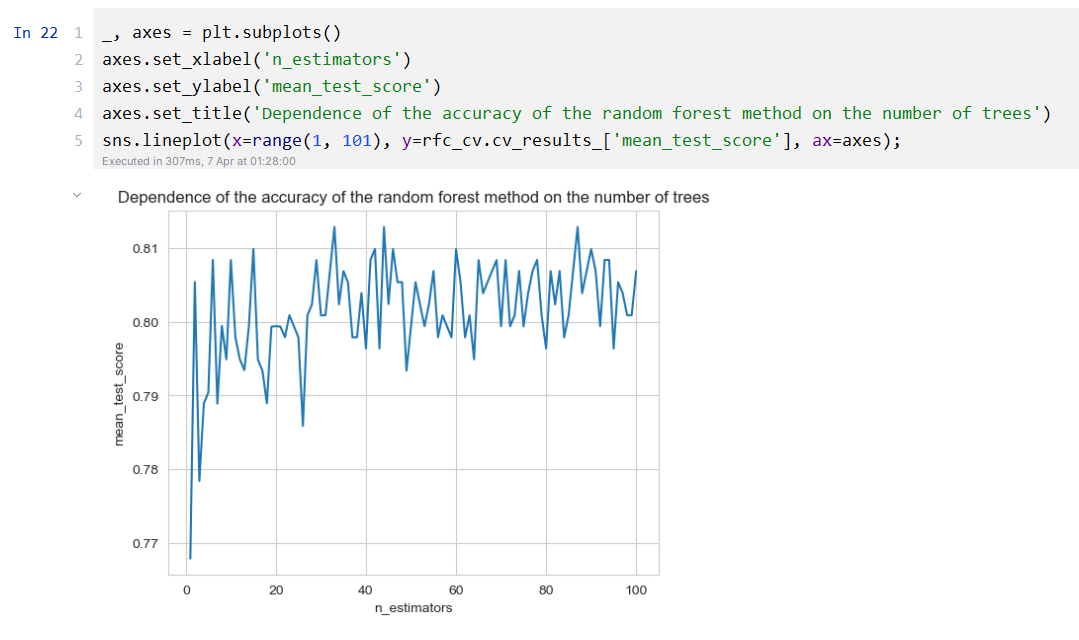


Рис. 3.20 – Залежність точності методу випадкових лісів від кількості дерев

Тепер обрахуємо середньоквадратичну похибку для випадкових лісів.

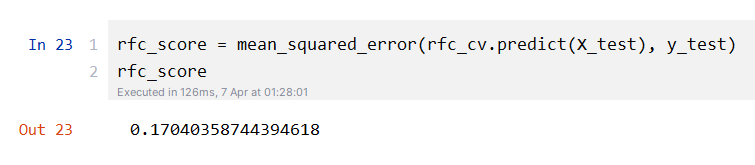


Рис. 3.21 – Середньоквадратична похибка на тестовій вибірці для методу випадкових лісів

Вона склала 0.170. Відобразимо стовпчасту діаграму середньоквадратичної похибки для кожного методу.



Рис. 3.22 – Стовпчаста діаграма середньоквадратичної помилки кожного з методів

Перевіривши різні методи, я дійшов до висновку, що найкращим на заданих даних є метод випадкових лісів із середньоквадратичною похибкою 0.179, що є доволі хорошим результатом.

# Виконання додаткового завдання

## Визначити, який регіон домінує в кластерах по ВВП на душу населення та щільності населення

Для початку проведемо все ті ж дії з цим набором, що й у попередніх лабораторних роботах: завантажимо дані, виправимо помилки, а також додамо стовпець щільності населення.

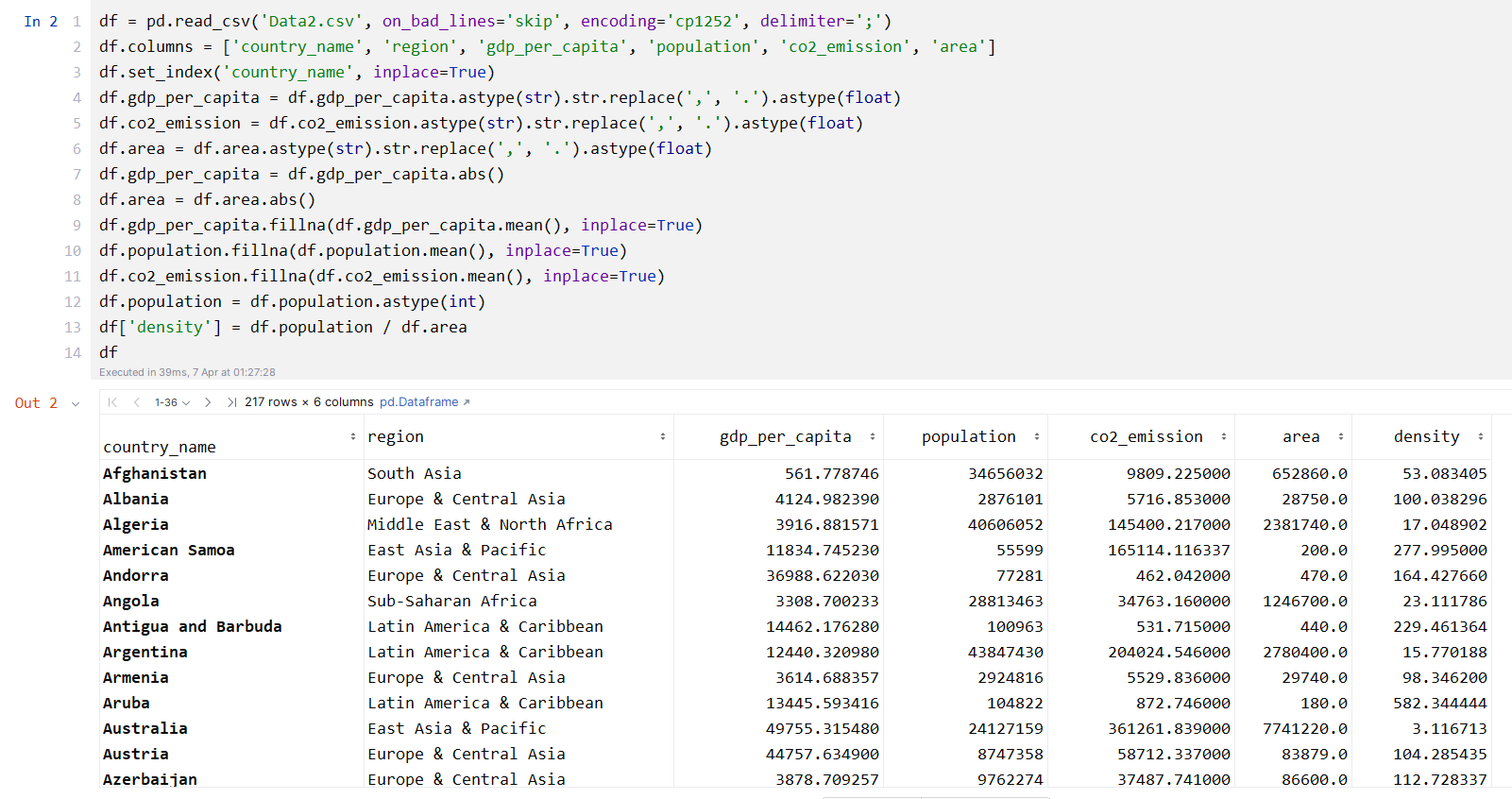


Рис. 4.1 – Завантаження даних та виправлення помилок

Далі виокремимо ознаки, за якими й будемо проводити кластеризацію: ВВП на душу населення.

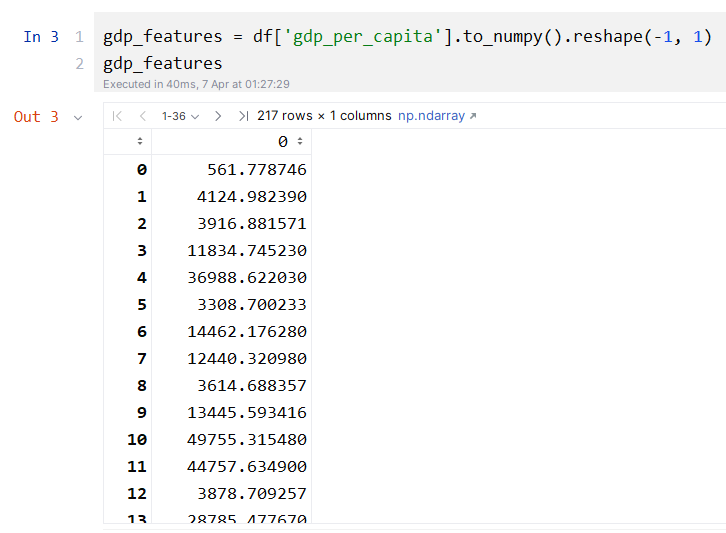


Рис. 4.2 – Виділення ознак для кластеризації за ВВП на душу населення

Після цього необхідно знайти точку локтя для моделі кластеризації методу K-середніх. Визначимо максимальну кількість кластерів як 13 та обрахуємо для кожної кількості кластерів інерцію.

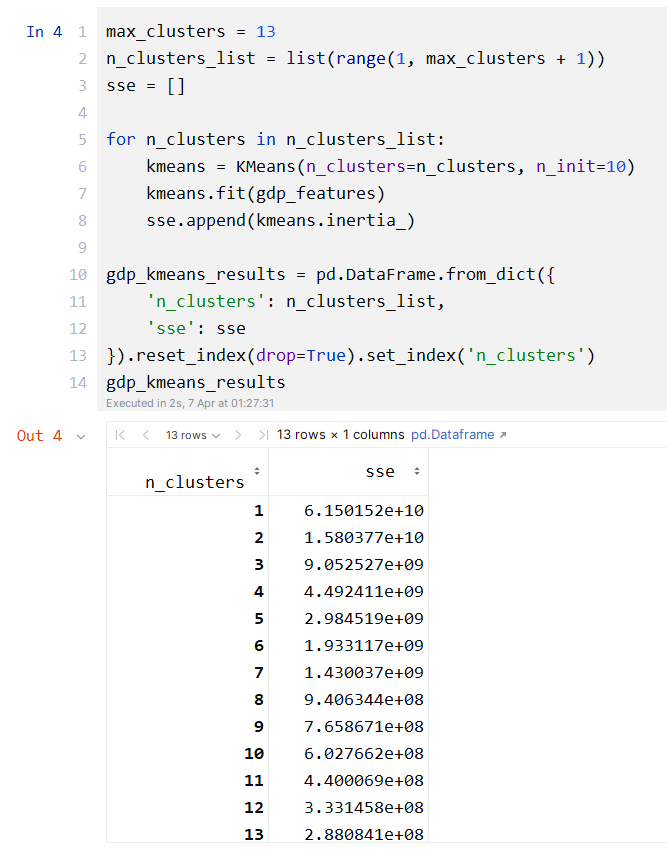


Рис. 4.3 – Обрахунок інерції для різної кількості кластерів

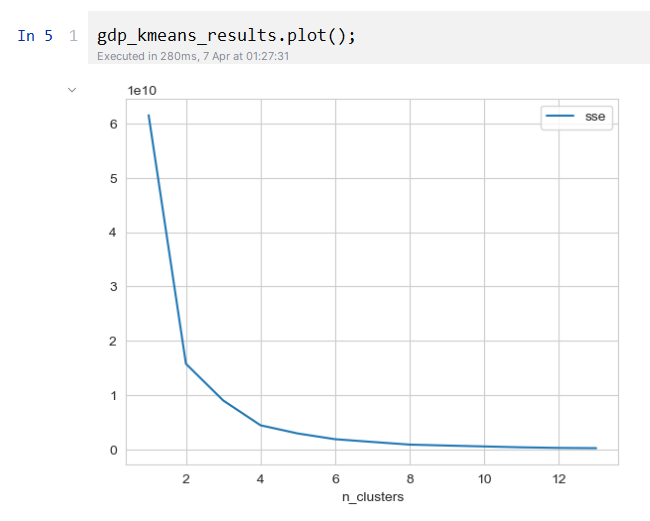


Рис. 4.4 – Графік залежності інерції від кількості кластерів

Обрахувавши інерцію, можемо знайти точку локтя.

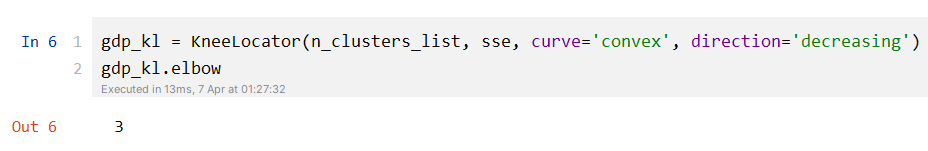


Рис. 4.5 – Знаходження точки локтя

Точка локтя дорівнює трьом, а отже використовуватимемо для цієї моделі 3 кластери. Навчимо цю модель та виведемо номери кластерів для кожного значення.



Рис. 4.6 – Навчання методу K-середніх та виведення міток кластеру для кожного значення

Підрахуємо кількість ознак у кожному з кластерів.

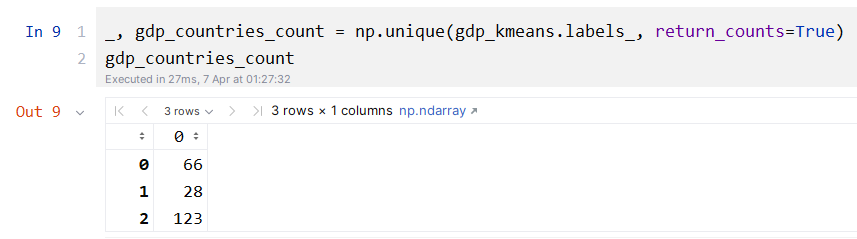


Рис. 4.7 – Кількість ознак у кожному кластері

Можемо вивести перелік країн із кожного кластеру.

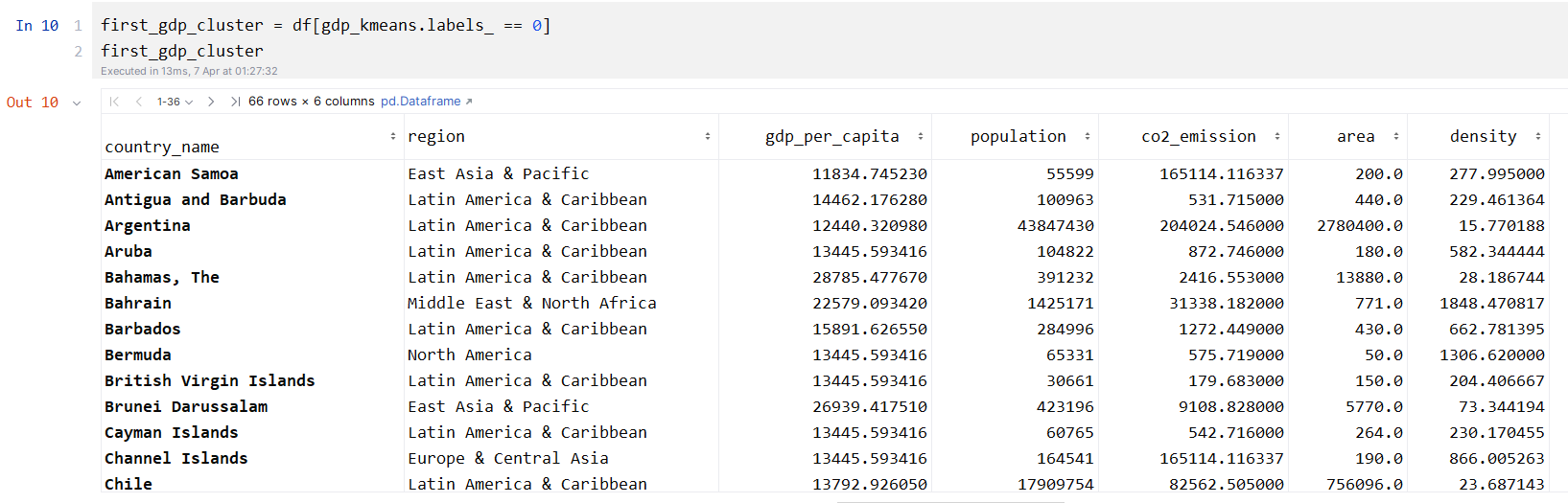


Рис. 4.8 – Перший кластер

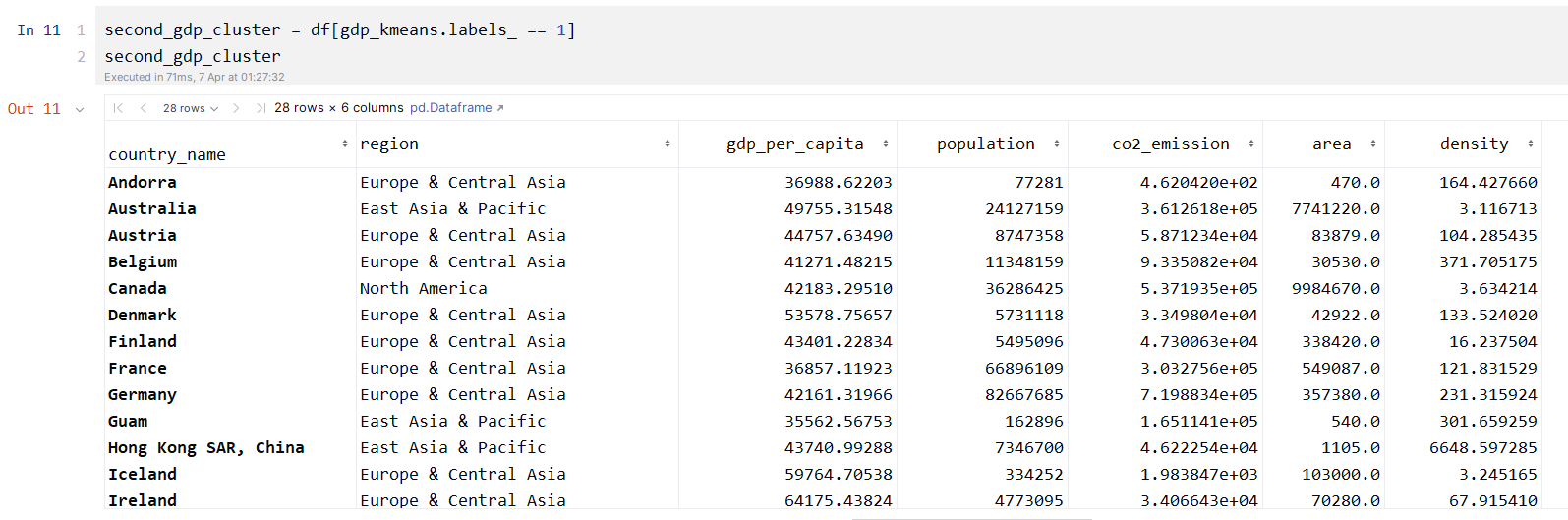


Рис. 4.9 – Другий кластер

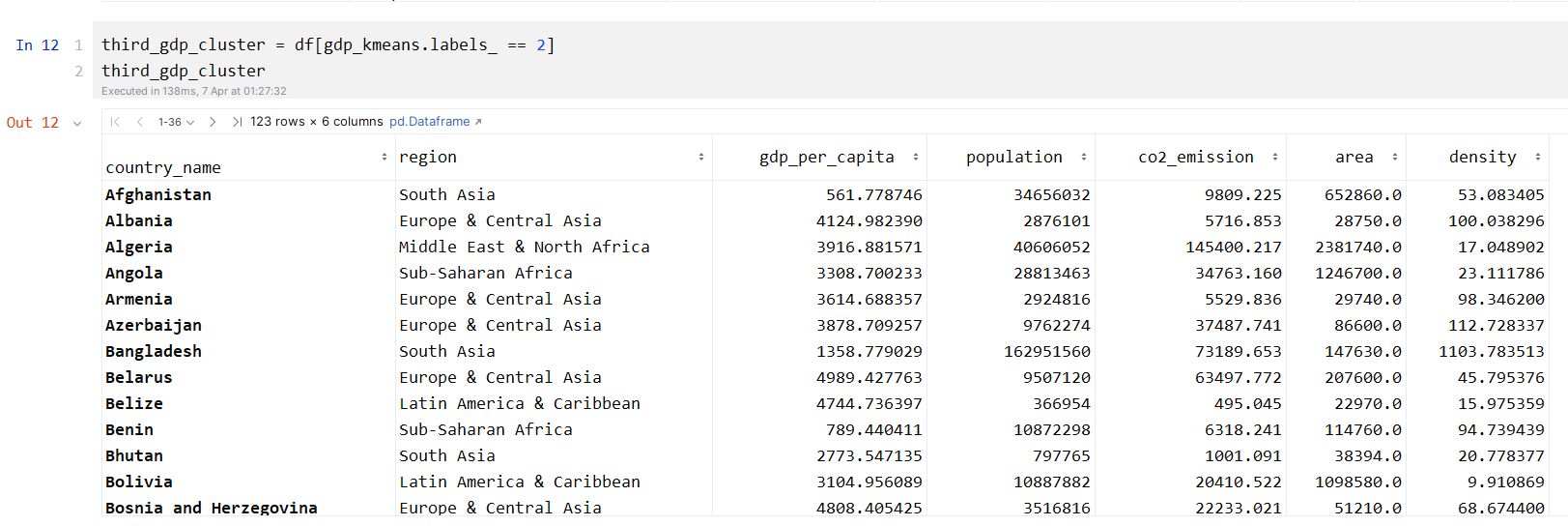


Рис. 4.10 – Третій кластер

Після цього можемо вивести максимальний ВВП на душу населення в кожному регіоні в кожному кластері.

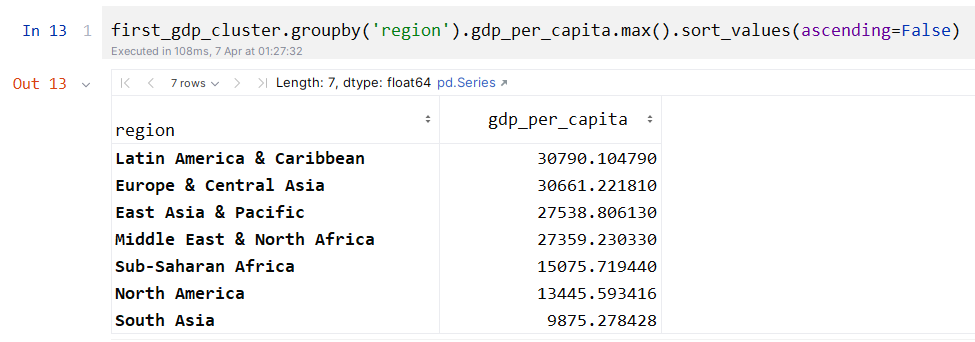


Рис. 4.11 – Максимальний ВВП на душу населення за регіонами в першому кластері

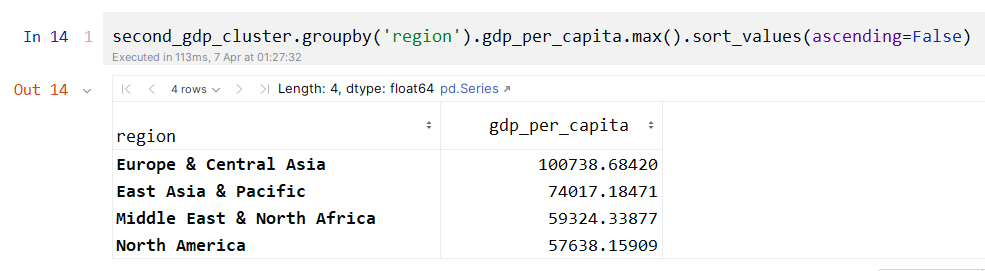


Рис. 4.12 – Максимальний ВВП на душу населення за регіонами в другому кластері

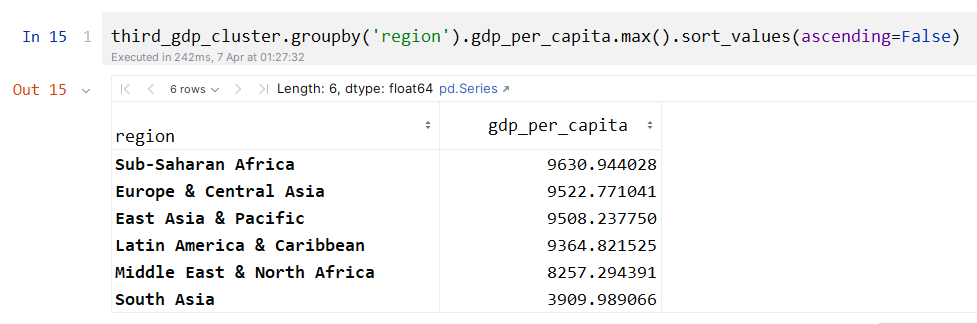


Рис. 4.13 – Максимальний ВВП на душу населення за регіонами в третьому кластері

Бачимо, що в першому кластері домінує Латинська Америка та Кариби, у другому – Південна Азія, а в третьому – Європа та Центральна Азія. Далі проведемо ці ж дії для щільності населення.

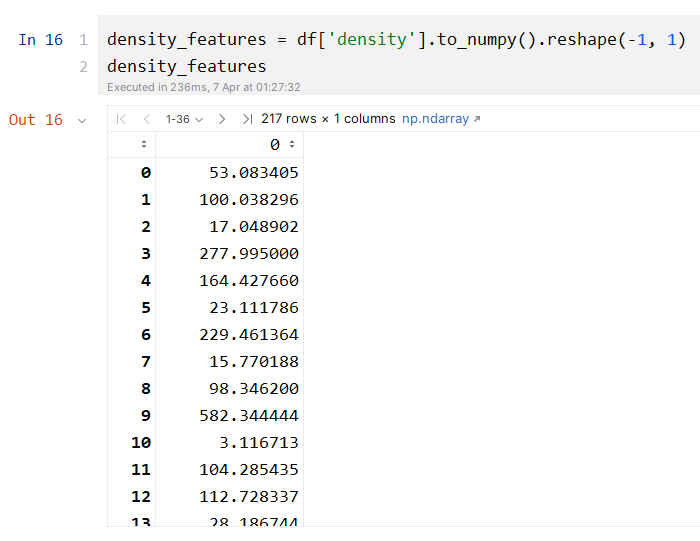


Рис. 4.14 – Виділення ознак для кластеризації за щільністю населення

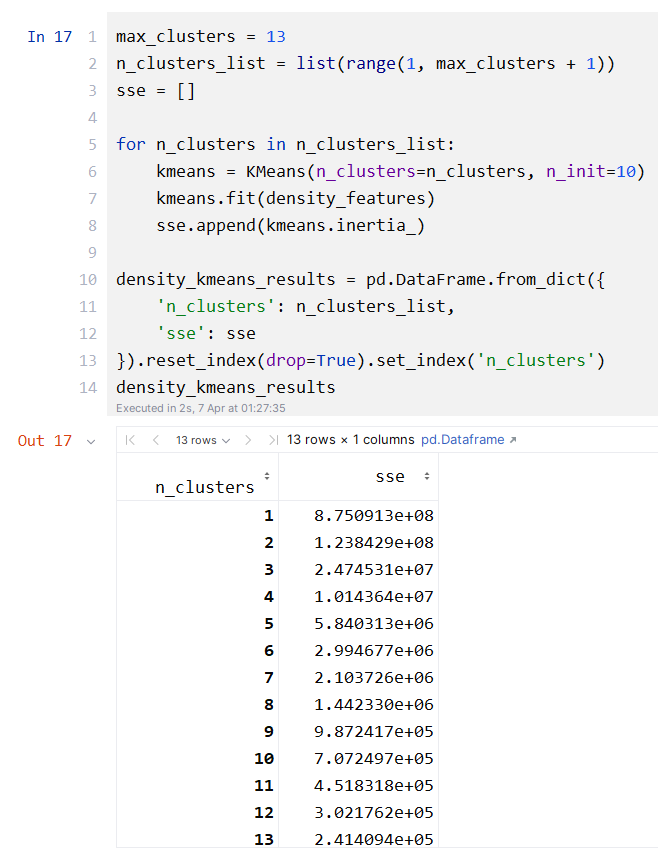


Рис. 4.15 – Обрахунок інерції для різної кількості кластерів

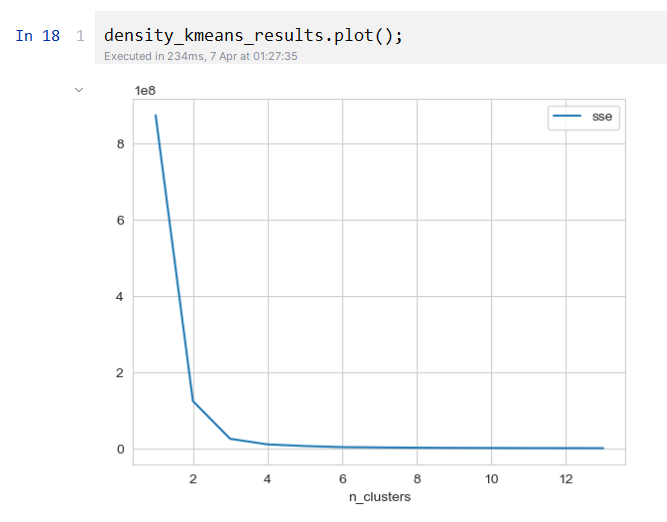


Рис. 4.16 – Графік залежності інерції від кількості кластерів

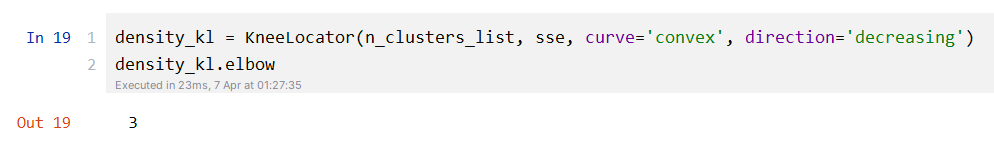


Рис. 4.17 – Точка локтя

Бачимо, що точка локтя знову дорівнює трьом, а отже знову використовуватимемо 3 кластери.

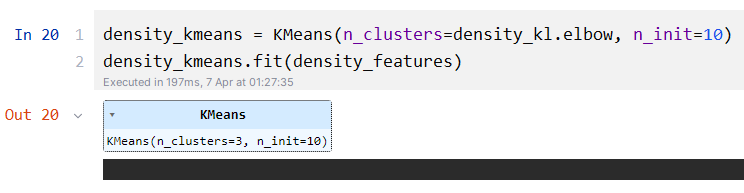


Рис. 4.18 – Навчання методу K-середніх

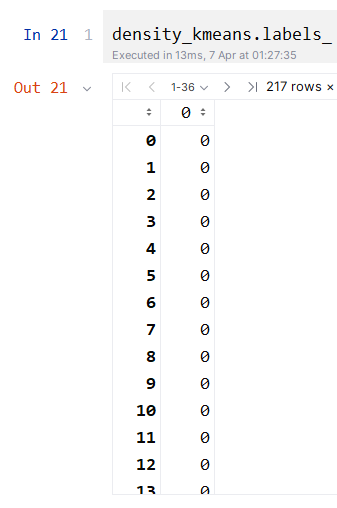


Рис. 4.19 – Мітки кластерів для кожної ознаки

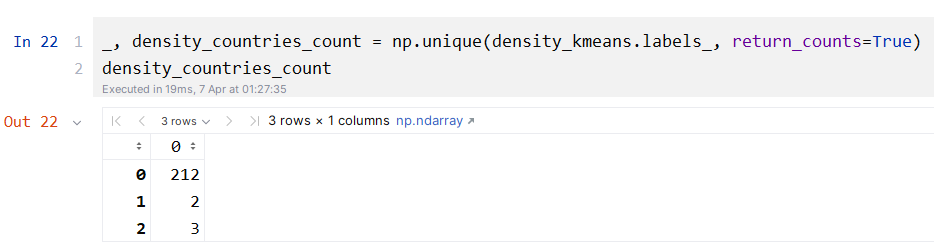


Рис. 4.20 – Кількість ознак у кожному кластері

Бачимо, що в першому кластері аж 212 країн, тоді як у другому – дві, а в третьому – 3. Оскільки кластери малі, можемо вивести перелік країн кожного з кластерів.

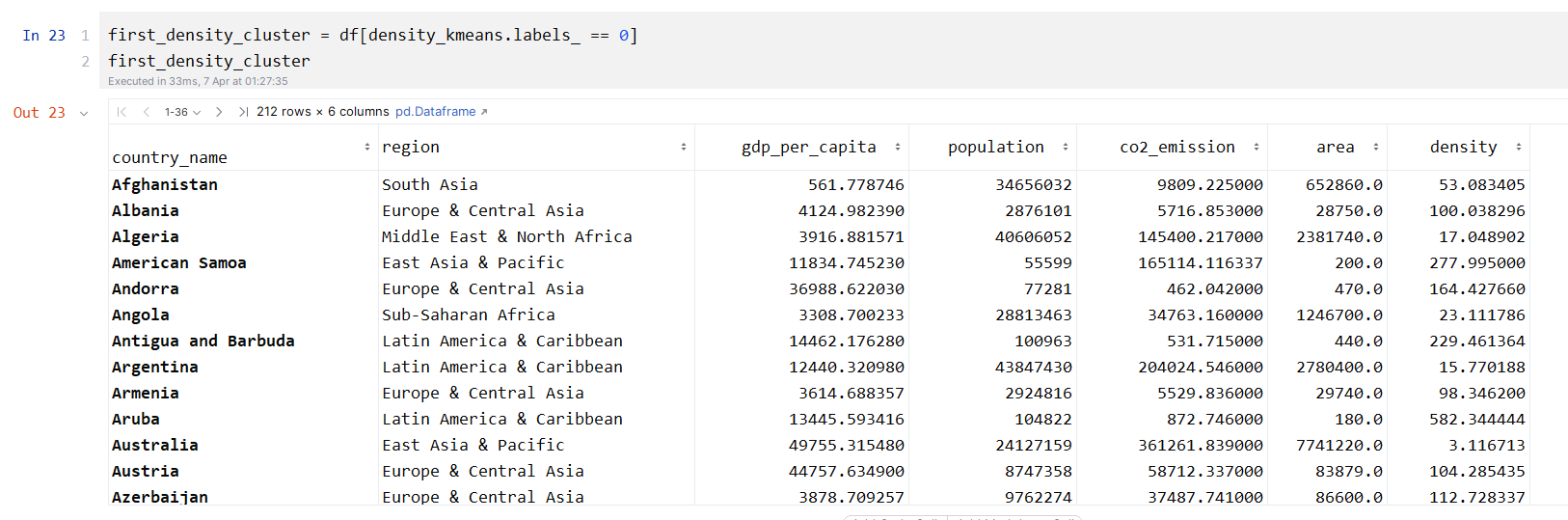


Рис. 4.21 – Перший кластер

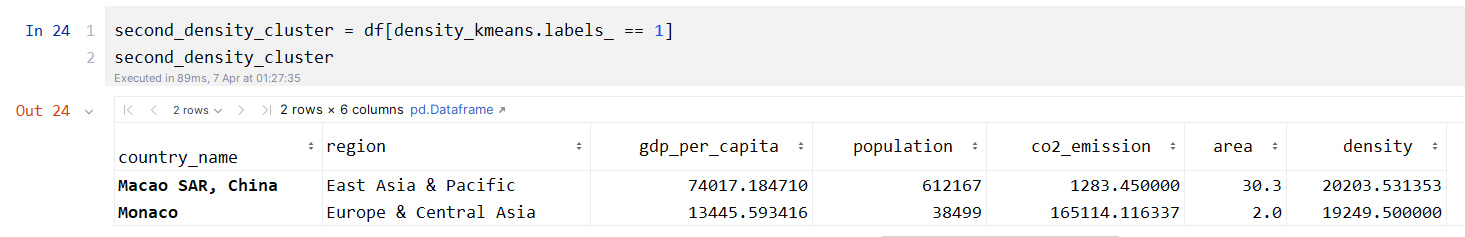


Рис. 4.22 – Другий кластер

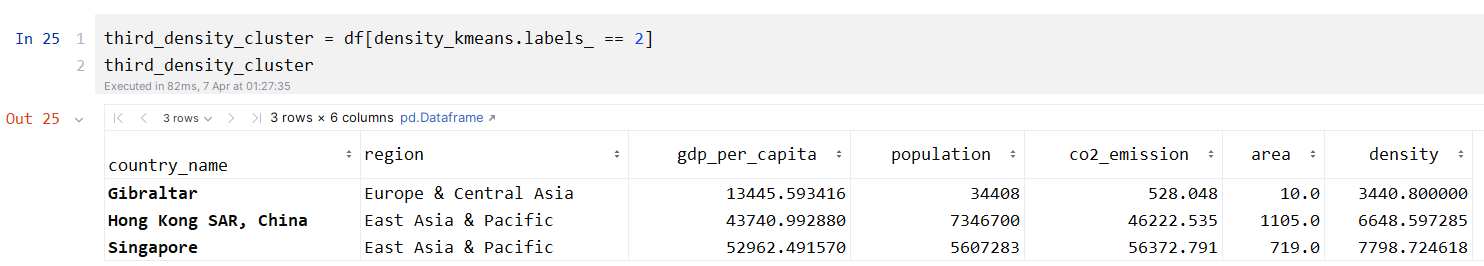


Рис. 4.23 – Третій кластер

Можемо зробити висновок, що, скоріше за все, кластеризація відпрацювала коректно, оскільки в другий кластер потрапили країни з дуже великою щільністю населення, а в третій – із більшою, ніж середня. Виведемо максимальну щільність населення в кожному регіоні в кожному кластері.

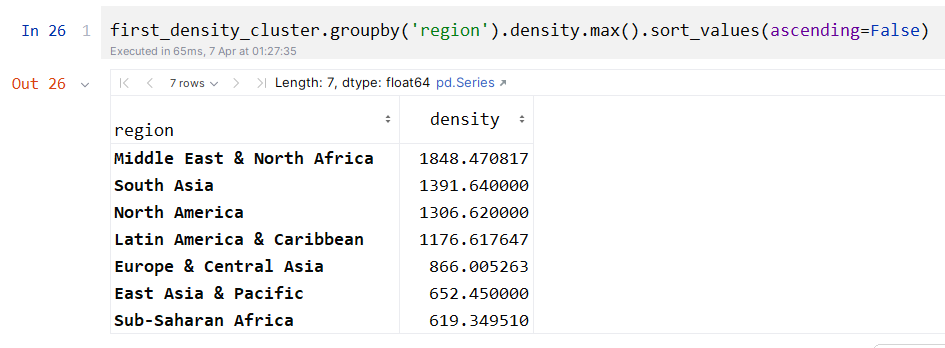


Рис. 4.24 – Максимальна щільність населення за регіонами в першому кластері

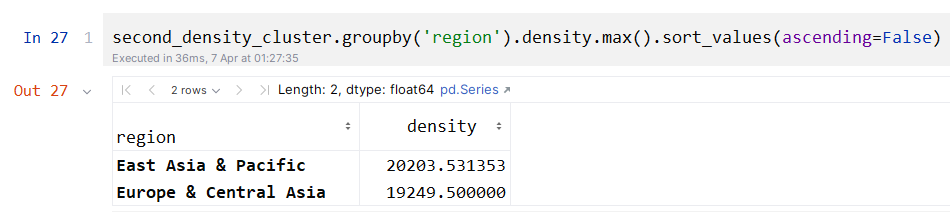


Рис. 4.25 – Максимальна щільність населення за регіонами в другому кластері

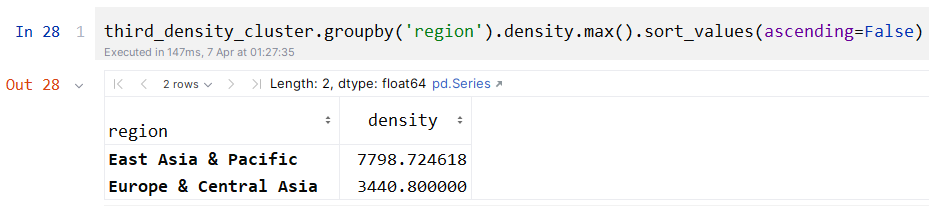


Рис. 4.26 – Максимальна щільність населення за регіонами в третьому кластері

Як бачимо, в першому кластері домінує регіон Близький Схід та Північна Африка, у другому та третьому – Східна Азія та Тихий океан.

Отримані кластери, особливо при кластеризації по щільності населення, було б добре перевірити на точність. Для цього можемо використати оцінку силуету.

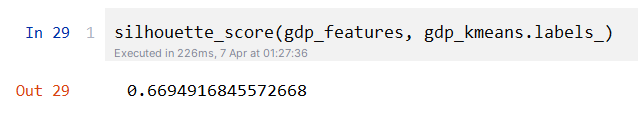


Рис. 4.27 – Оцінка силуету моделі кластеризації за ВВП на душу населення

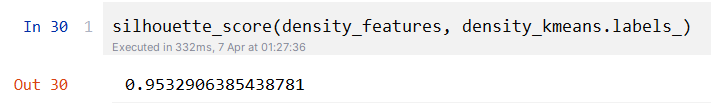


Рис. 4.28 – Оцінка силуету моделі кластеризації за щільністю

У випадку кластеризації за ВВП на душу населення вона дорівнює 0.669, що можна вважати непоганим результатом, а у випадку кластеризації за щільністю населення – аж 0.953, що є дуже високим результатом.

## Вивести частотні гістограми всіх показників файла Data2.csv, використовуючи цикл

Для початку відкинемо всі нечислові дані (регіон).

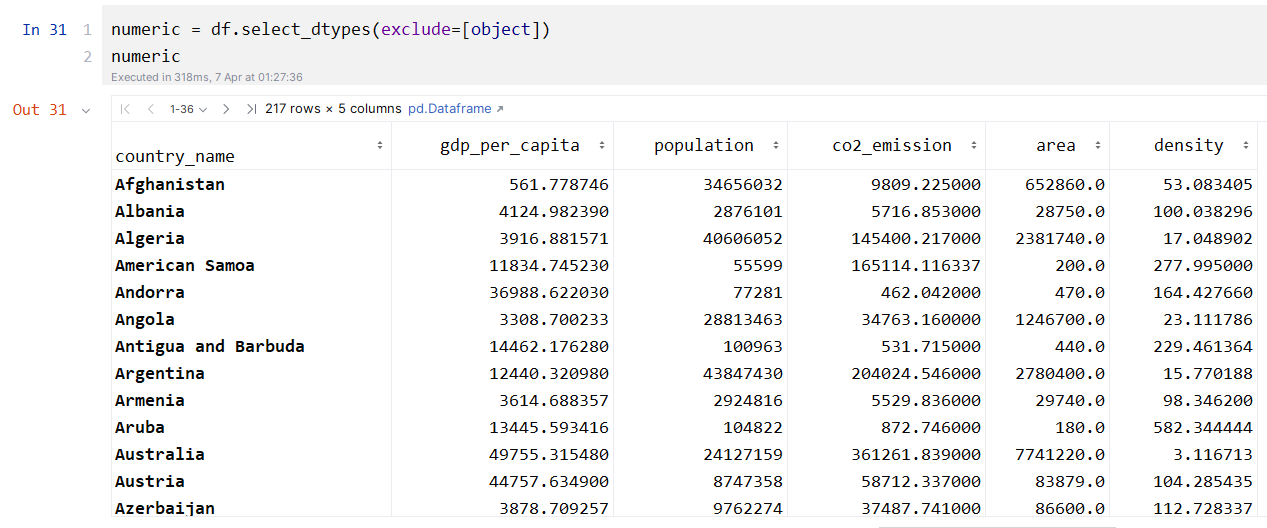


Рис. 4.29 – Цифрові дані набору даних

Потім просто пройдемося циклом по кожному стовпцю й відобразимо для них частотні гістограми за допомогою звичайного методу .hist() класу DataFrame.

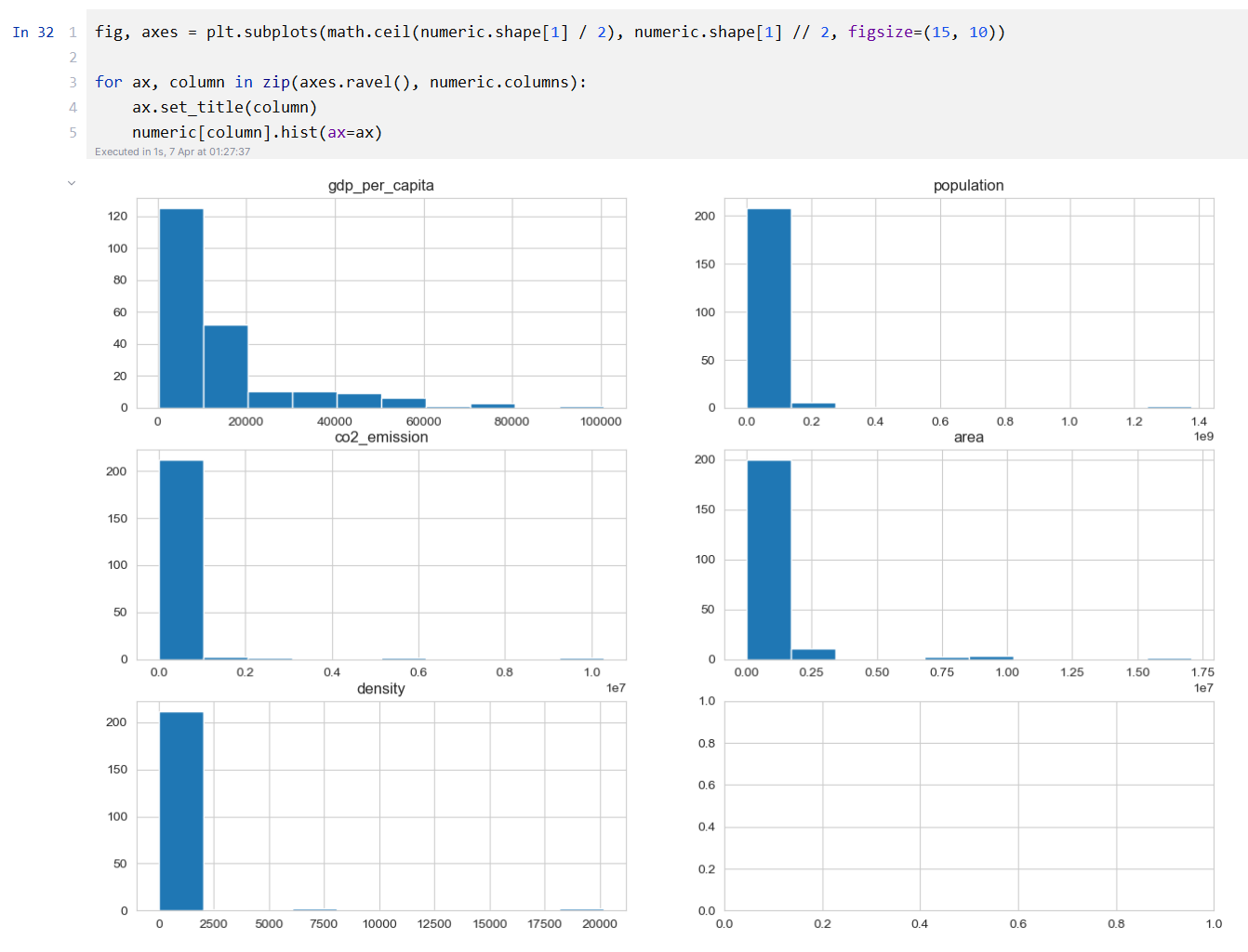


Рис. 4.30 – Частотні гістограми

## Створити функцію, яка на вхід отримує два набори даних, перевіряє чи є лінійна залежність та виводить True чи False (будемо розуміти під «є лінійна залежність», якщо коефіцієнт кореляції по модулю більше 0,8)

Таку перевірку зробити дуже просто: треба всього викликати метод .corrwith() класу DataFrame, задіяти для отриманих значень модуль та перевірити, чи є такі, що перевищують 0.8.

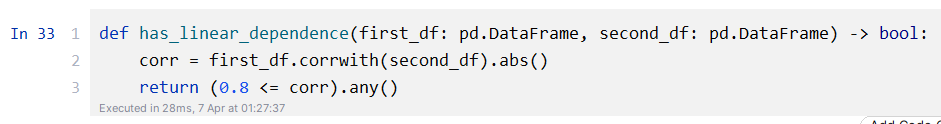


Рис. 4.31 – Функція перевірки на наявність лінійної залежності між двома наборами даних

Створимо тестові набори даних.

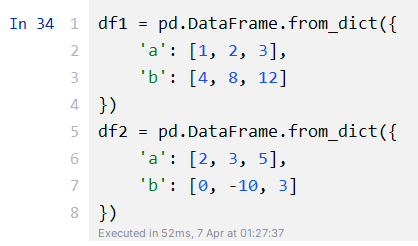


Рис. 4.32 – Створення тестових наборів даних

Бачимо, що в них явно є залежність між стовпцями «a» та вона явно відсутня між стовпцями «b». Перевіримо, що поверне створена функція.

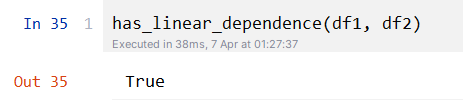


Рис. 4.33 – Тестування функції

Вона повернула True, як і очікувалося, а значить функція працює коректно.

# Висновок

У ході даної лабораторної роботи було досліджено набір даних пасажирів Титаніку, на основі якого побудовано декілька класифікаційних моделей, що були засновані на наступних методах: К-найближчих сусідів, дерева рішень, випадкових лісів. Для кожного методу було обраховано середньоквадратичну похибку, за значеннями якої було встановлено, що найкращим методом для такого набору даних є метод випадкових лісів, який показав точність у 81.28% на навчальній вибірці та середньоквадратичну помилку 0.17. Було помічено, що в методі дерева прийняття рішень дуже велику роль грає максимальна глибина.

Після цього було побудовано дві моделі кластеризації набору даних країн: за ВВП на душу населення та щільністю населення. Для кожної моделі було обраховано точку локтя: в обох випадках вона дорівнювала трьом. Потім було виведено, які країни входять до кожного з кластерів, а також які регіони домінують за ознакою кластеризації в кожному з кластерів. Також для обох моделей було обраховано оцінку силуету: для першої моделі вона склала 0.669, а для другої – 0.953.

Потім було зображено гістограми розподілу для кожної з ознак, а в кінці створено функцію, яка визначає, чи є лінійна залежність між двома наборами даних: її було перевірено на власноруч створеному тестовому наборі даних.