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 № 6 з дисципліни

«Аналіз даних в інформаційних системах»

«Класифікація та кластеризація»

Виконав студент ІП-11 Лесів Владислав Ігорович

Перевірила Ліхоузова Тетяна Анатоліївна

Київ 2023

**Комп’ютерний практикум 6**

**Класифікація та кластеризація**

**Мета роботи:** ознайомитись з

* методами класифікації та кластеризації;
* моделями, що використовують дерева прийняття рішень;
* інструментами факторного аналізу методом головних компонент та методом найбільшої подібності.

**Основне завдання**.

Скачати потрібні дані.

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;

1. Вивести частотні гістограми всіх показників файлу Data2.csv, використовуючи цикл;
2. 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).

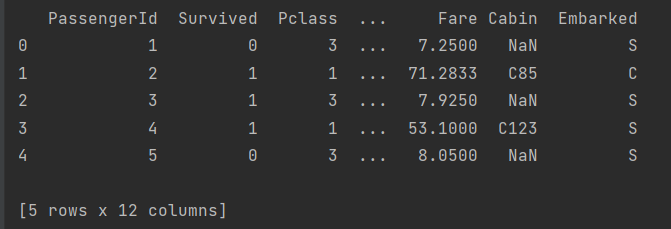
**Хід роботи.**

**Основне завдання.**

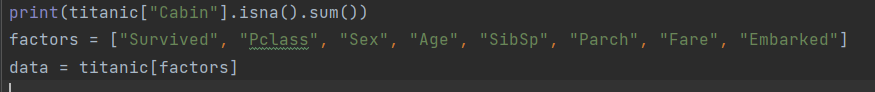
Для даних по титаніку titanic.csv побудувати модель, в якій можна визначити, чи виживе пасажир, заповнивши решту параметрів. Використати декілька методів. Порівняти результати.

Для початку завантажую датасет.





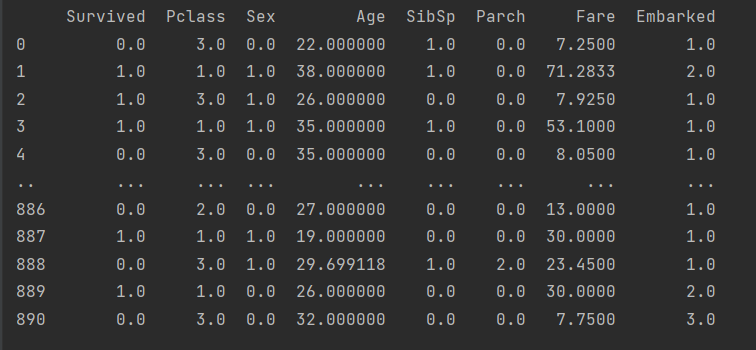
Далі опрацьовую його. Вводжу фактори, адже ID, ім’я, номер квитка та номер кімнати не впливає на те, виживе пасажир чи ні. Крім того, у колонці Cabin пропущена більша частина значень, тож її варто видалити.



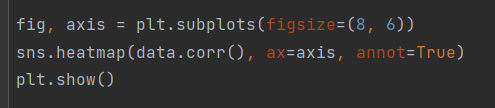


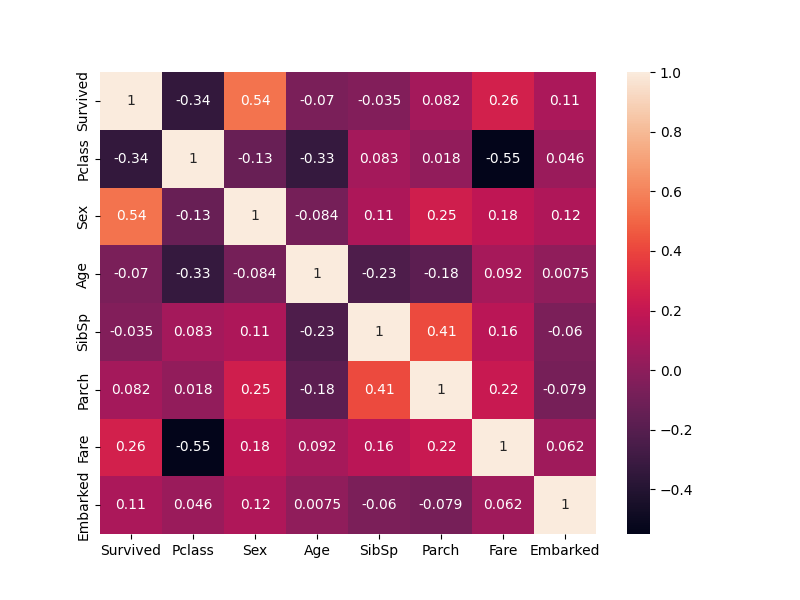
Заміняю пропущені значення віку середніми, а пропущене місце посадки – найчастішим значенням. Після підрахунку отримали найчастіше місце посадки – S. Також заміняю текстові дані числовими, тобто стать і місце посадки заміняю цифрами.



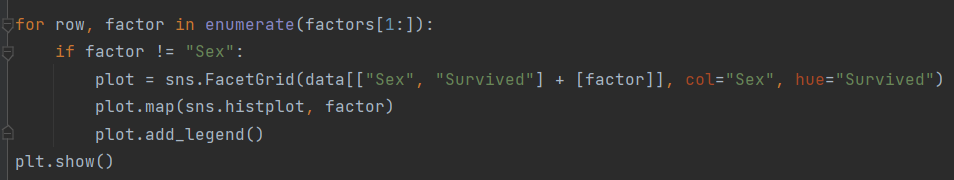


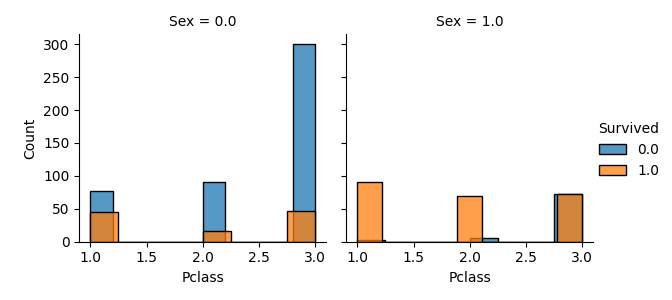
Далі будую графік матриці кореляції. Бачимо, що коефіцієнти кореляції невеликі між усіма значеннями.

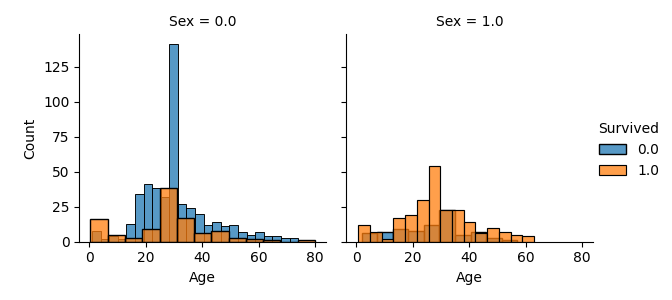


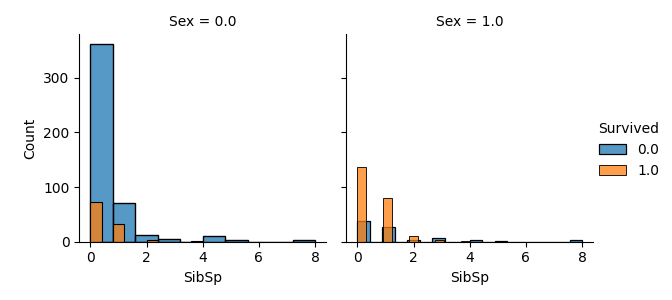


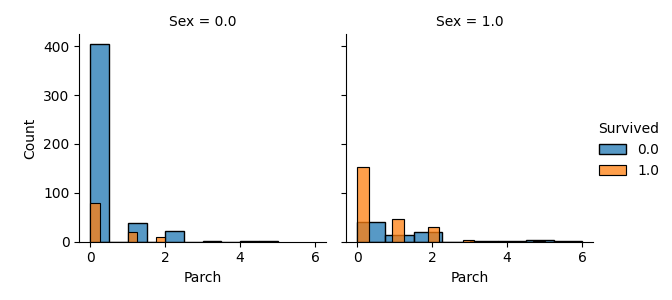
Також для візуалізації я обрав те, як виживали люди за статтю і відповідно до ще одного параметру. Маємо такі результати.

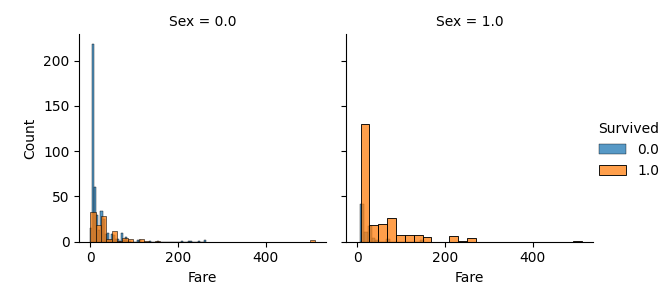






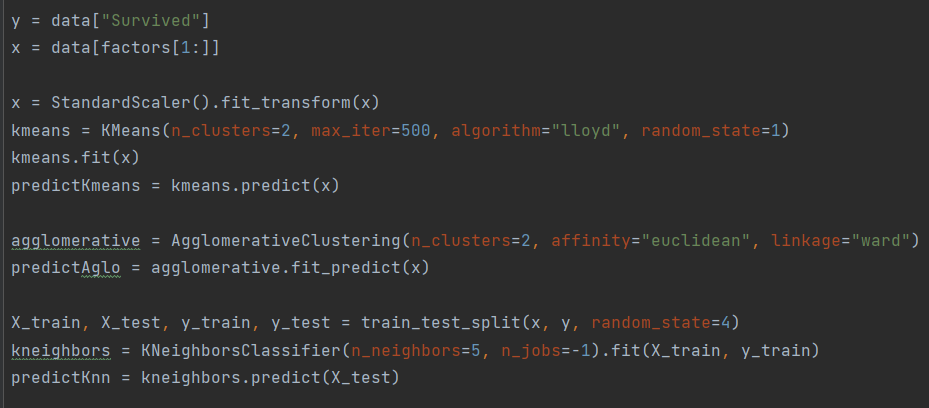




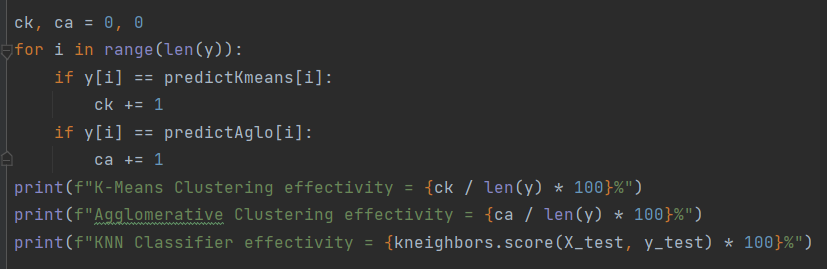


Бачимо, що жінок більше виживало, ніж помирало, а чоловіків навпаки. Причому незалежно від статі більше помирало людей з 3 класу, а найменше – з 1 класу. За віком найбільше померло людей молодих, між 20 і 40 років. Людей без родичів загинуло більше, ніж людей з 2+ родичами. І також тих, хто платив менше, загинуло більше.

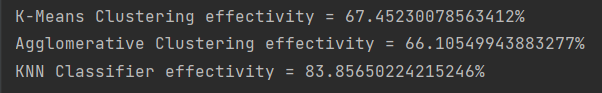
Для дослідження я обрав методи кластеризації KMeans і Agglomerative Clustering, а також класифікацію KNN.



Відповідно як правильний результат для кластеризації я відділяю колонку Survived, а решта даних кластеризую. Я обрав 2 кластери, адже ми будуємо модель для визначення, вижив пасажир чи ні, і тут можливо лише 2 варіанти. Тож кластеризую за KMeans, AgglomerativeClustering. Також розділяю для класифікації вибірку на навчальну й тестову, треную модель і передбачаю на тестовій вибірці.



Далі визначаю ефективність методів.

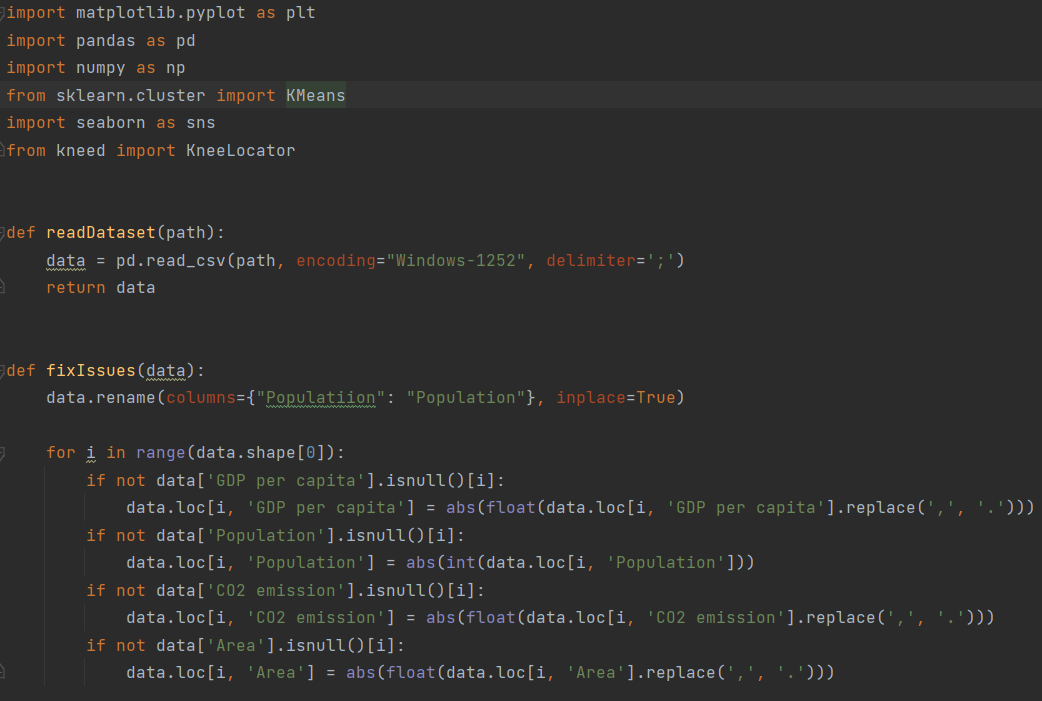


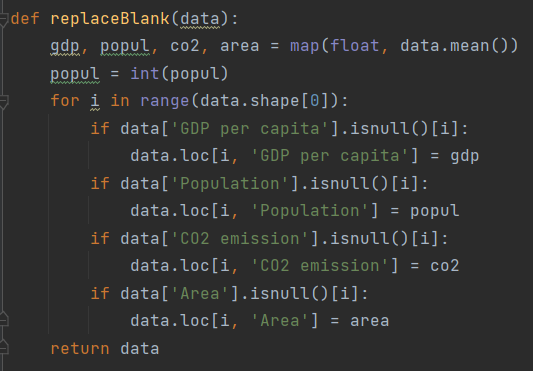
Бачимо, що для цих даних ефективність обох методів кластеризації приблизно однакова, хіба KMeans трошки ефективніше. Але у порівнянні з класифікацією, яку ми навчали на навчальних даних, кластеризація тут програє.

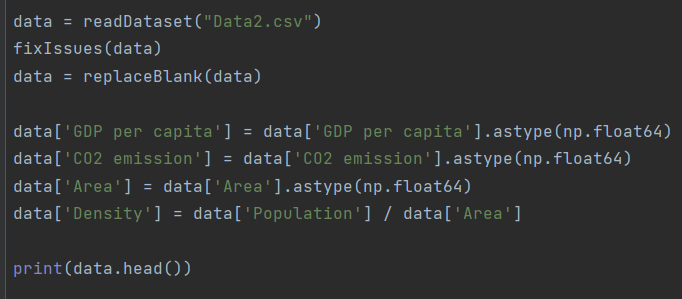
**Додаткове завдання.**

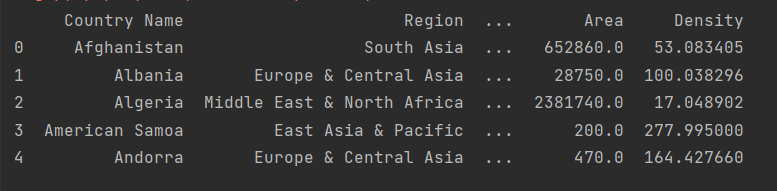
1. Визначити, який регіон домінує в кластерах по ВВП на душу населення та щільності населення;

Для початку завантажую дані і опрацьовую їх так, як я вже це робив у попередніх роботах: перейменовую колонку, у числах міняю кому на крапку, а пропущені значення заміняю середніми.

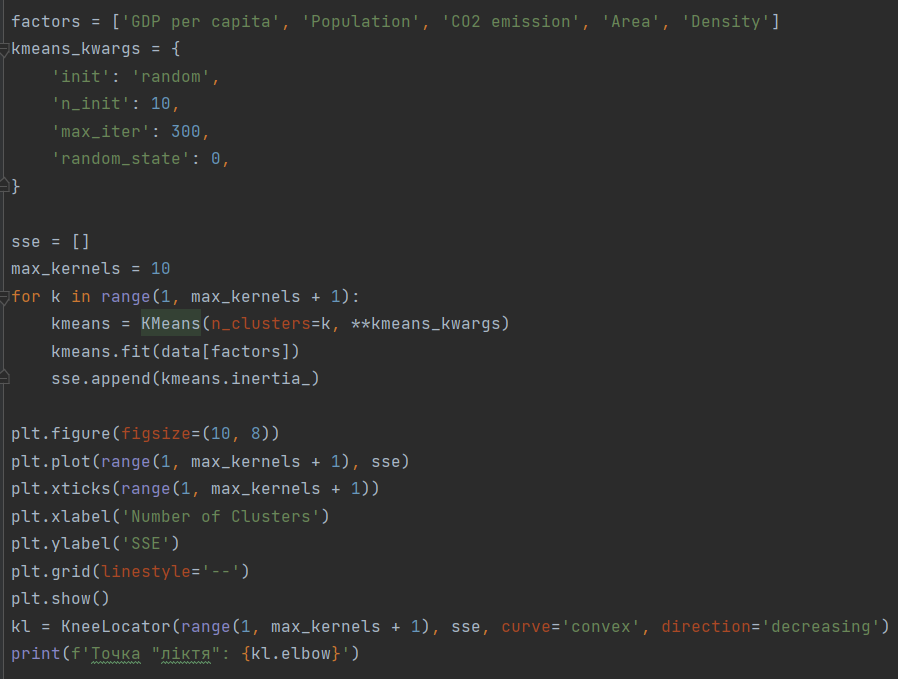




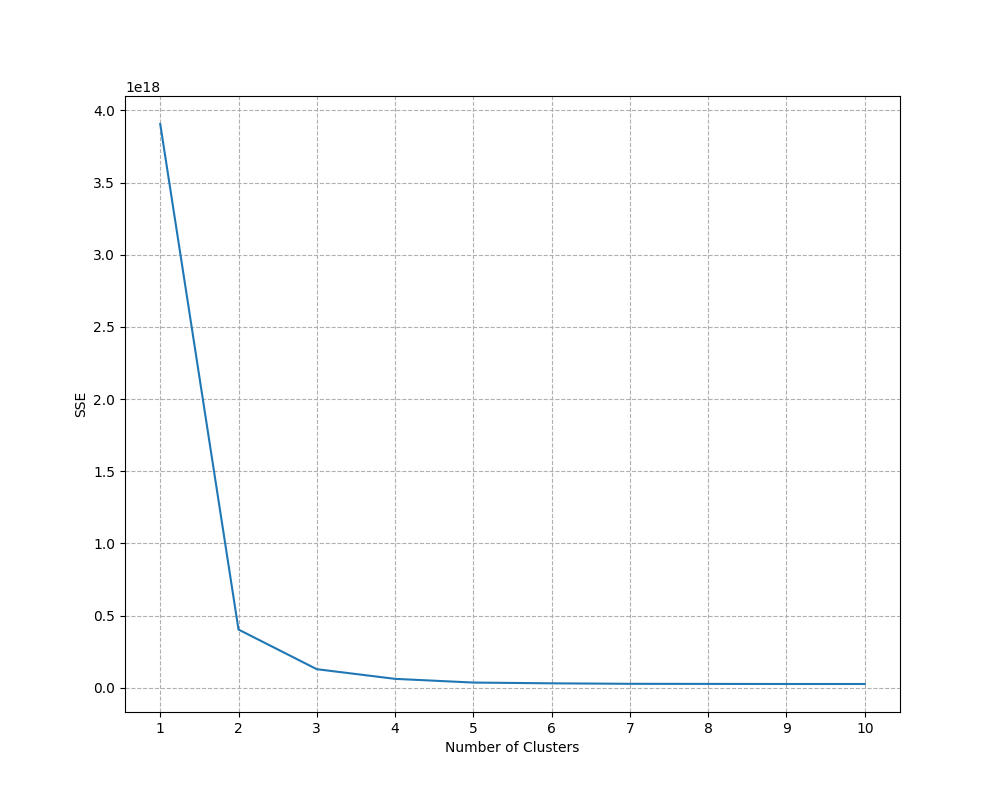




Виконаю кластеризацію даних. Вводжу фактори – це усі колонки, крім назв країн і регіонів. Спочатку визначаю задовільну кількість кластерів методом ліктя.

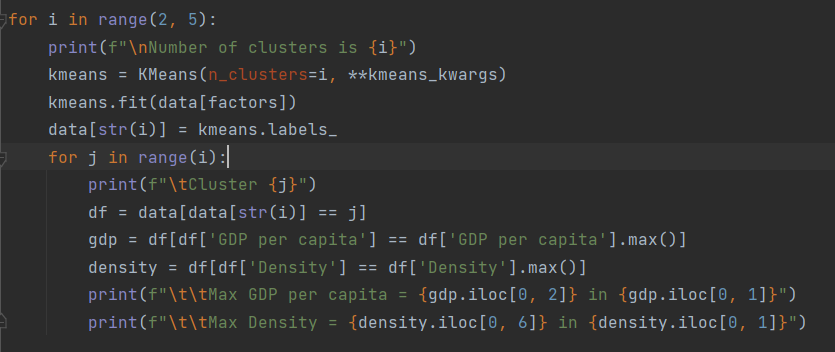


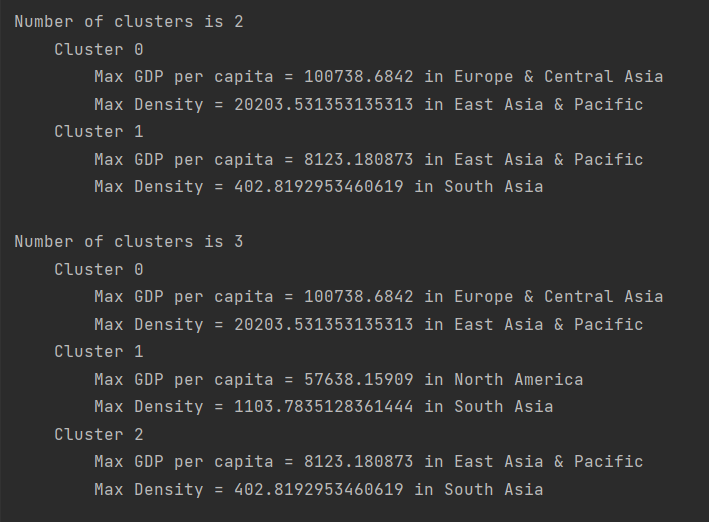


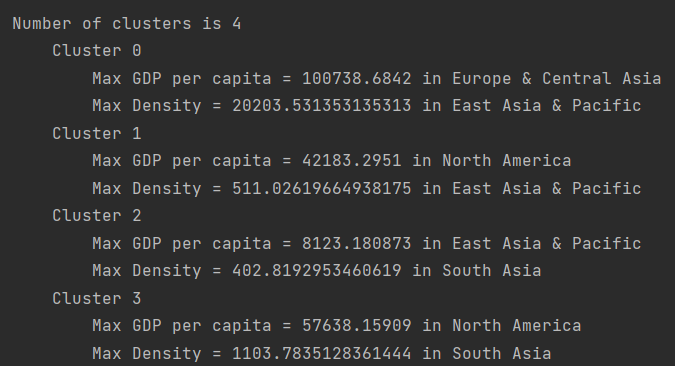


Переконуємося, що точка ліктя = 2.

Обираю кількість кластерів від 2 до 4, кластеризую дані і в кожному кластері шукаю регіон, у якому є найбільше значення GDP per capita i Density.

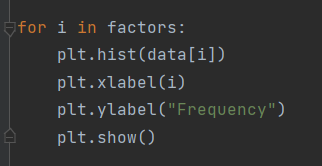


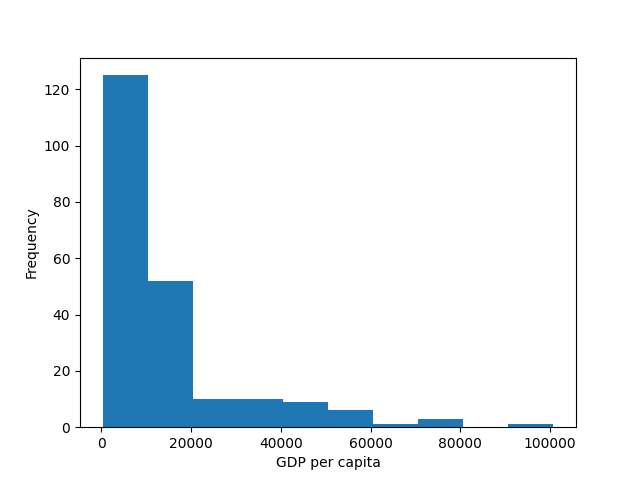


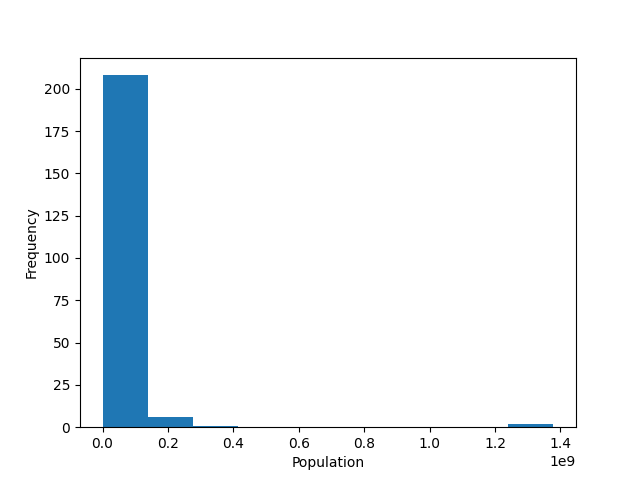


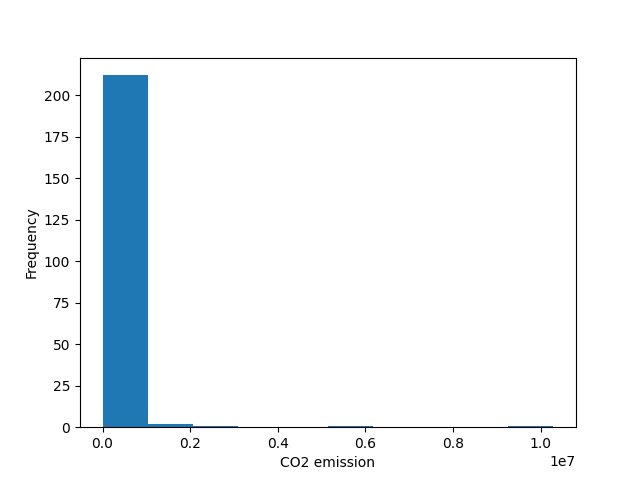
1. Вивести частотні гістограми всіх показників файлу Data2.csv, використовуючи цикл;

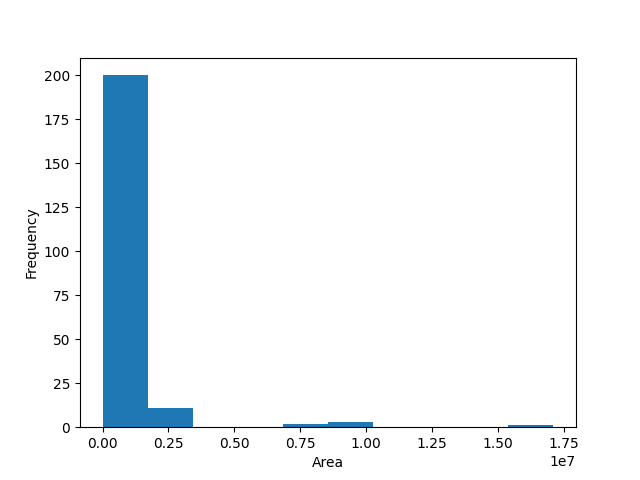
Для частотних гістограм використовую функцію plt.hist().

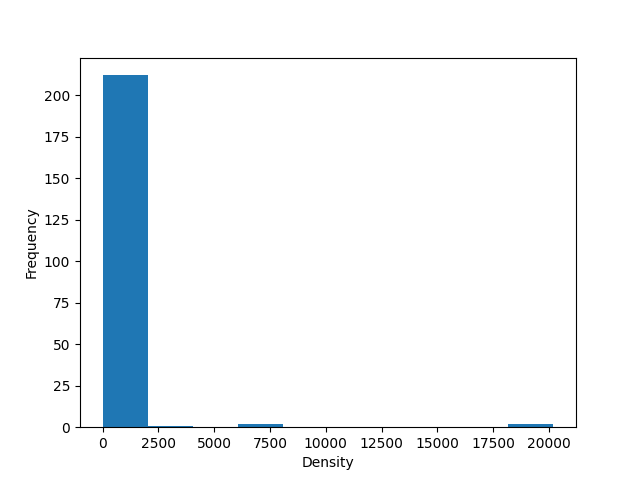








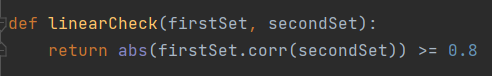




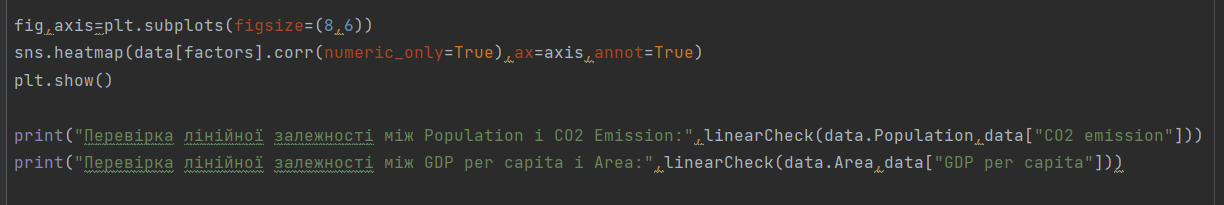
Бачимо, що найчастіше значення ВВП на душу населення становить від 0 – 20000 і далі спадає, майже в усіх країнах населення становить 0 – 0.2e9, і лише кілька країн-гігантів, як Китай чи Індія, мають 1.2-1.4e9. Схожа ситуація з викидами CO2, площею та густотою населення відповідно у своїх межах.

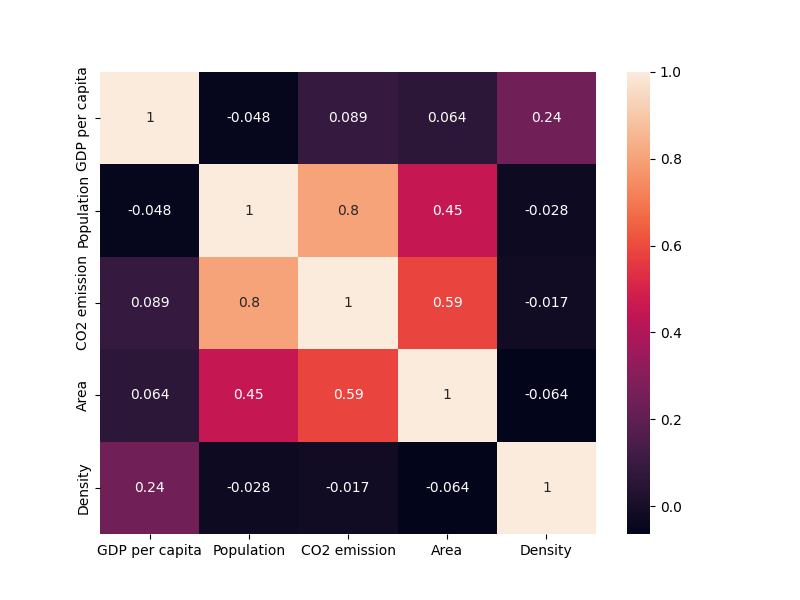
1. Створити функцію, яка на вхід отримує два набори даних, перевіряє чи є лінійна залежність та виводить True чи False (будемо розуміти під «є лінійна залежність», якщо коефіцієнт кореляції по модулю більше 0.8).

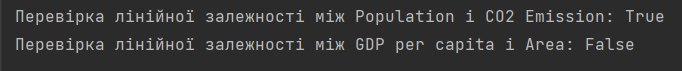
Для знаходження коефіцієнта кореляції використовую метод corr.



Виводжу матрицю кореляції і за її допомогою можемо перевірити правильність виконання функції.







Бачимо, що функція видає коректний результат.

**Висновок.**

Отже, у цій роботі я ознайомився з методами класифікації та кластеризації, моделями, що використовують дерева прийняття рішень, інструментами факторного аналізу методом головних компонент та методом найбільшої подібності.

У результаті лабораторної роботи було використано різні DataFrames, проведено їх дослідження, створено моделі класифікації KNN й кластеризацій KMeans й AgglomerativeClustering, які були навчені та протестовані за допомогою даних DataFrame. Було передбачено виживання пасажирів Титаніку відповідно до інших факторів. Отримали, що найкраще порається класифікація, адже у процесі навчання ми їй показуємо, який результат повинен вийти. А кластеризації для даного набору виявилися приблизно однаково ефективні.

У додатковому завданні зробили кластеризацію даних, побудували гістограми і створили функцію, яка визначає лінійну залежність між наборами. Використовуючи PyCharm, отримуємо коректний результат.