**Міністерство освіти і науки України**

**Львівський національний університет імені Івана Франка**

Факультет електроніки та комп’ютерних технологій

**Звіт**

Про виконання лабораторної роботи №3

# З курсу «Системи опрацювання даних»

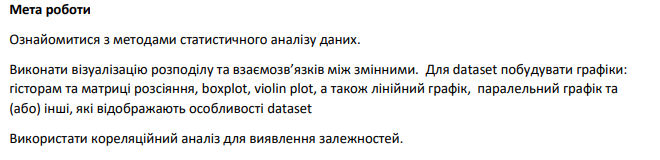
«Статистичний, візуальний та кореляційний аналіз даних»

Виконав:

Студент групи ФеС-21

Шавало Андрій

Львів-2025



**Завдання**

1. Завантажити набір даних dataset у форматі CSV (наприклад, data.csv).

2. Завантажити дані у pandas DataFrame та вивести перші 5 рядків. Зрозуміти і пояснити про щo дані.



 **Overall rank** – загальний рейтинг країни за рівнем щастя.

 **Country or region** – назва країни або регіону.

 **Score** – індекс щастя.

 **GDP per capita** – рівень ВВП на душу населення.

 **Social support** – рівень соціальної підтримки.

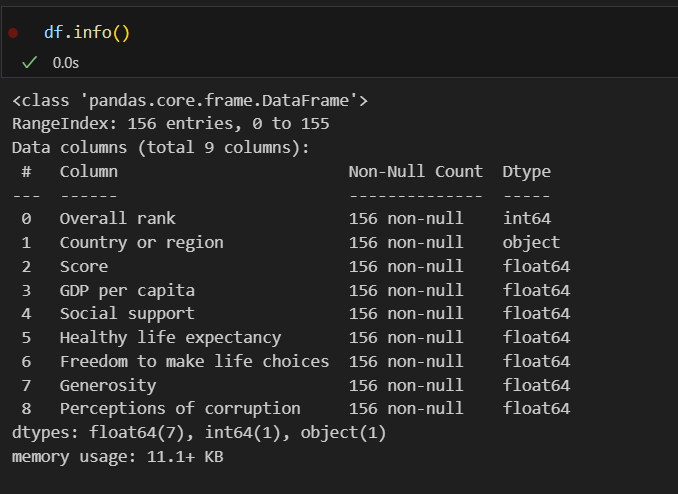
 **Healthy life expectancy** – очікувана тривалість здорового життя.

 **Freedom to make life choices** – рівень свободи у виборі життєвого шляху.

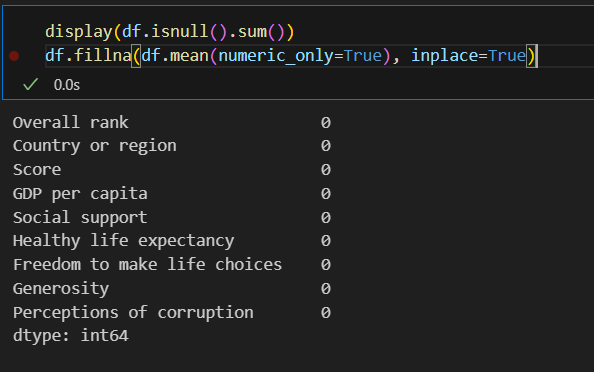
 **Generosity** – рівень благодійності.

 **Perceptions of corruption** – рівень сприйняття корупції.

3. Відобразити загальну інформацію про датасет (.info() …)

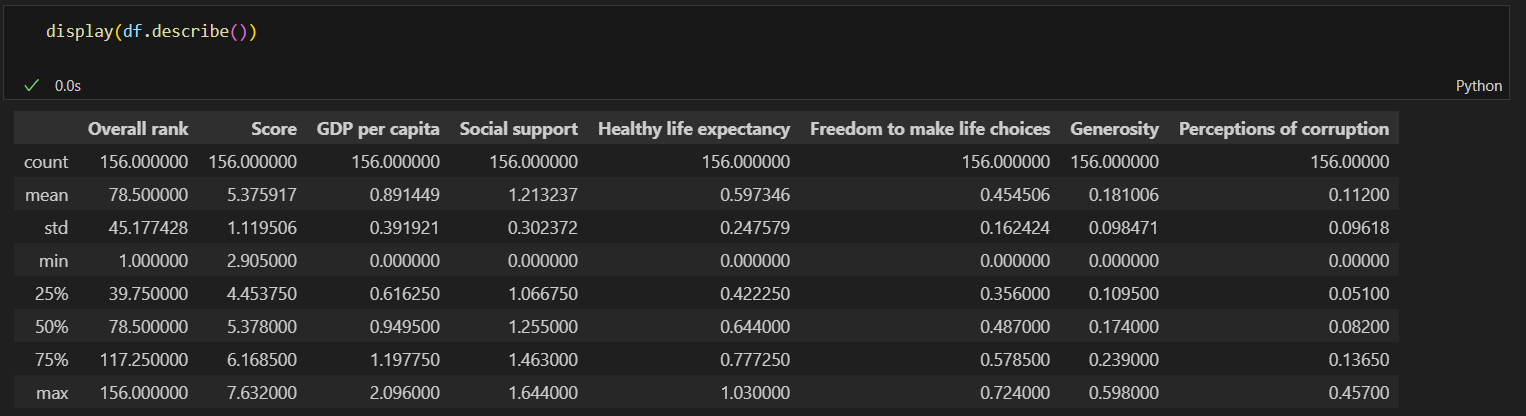


4. Перевірити наявність пропущених значень та обробити їх.

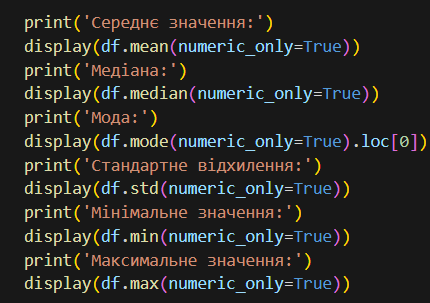


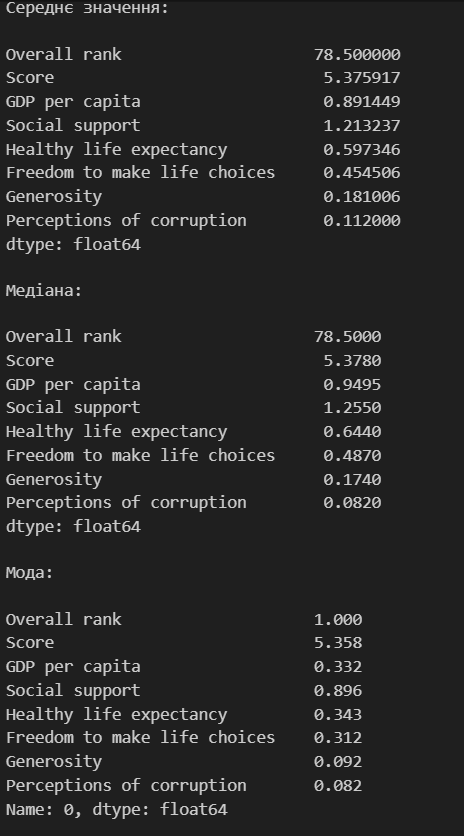
**Статистичний аналіз даних**

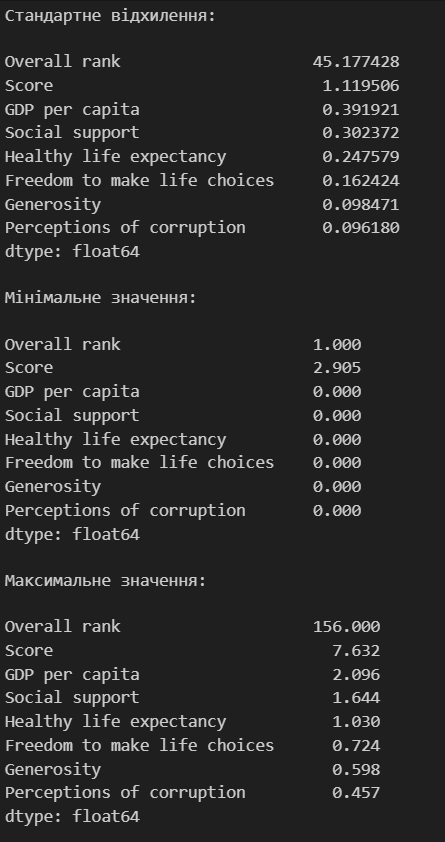
* Використати метод describe() для визначення статистичних характеристик.



* Обчислити середнє значення, медіану, моду, стандартне відхилення для числових змінних. Визначити мінімальні та максимальні значення.







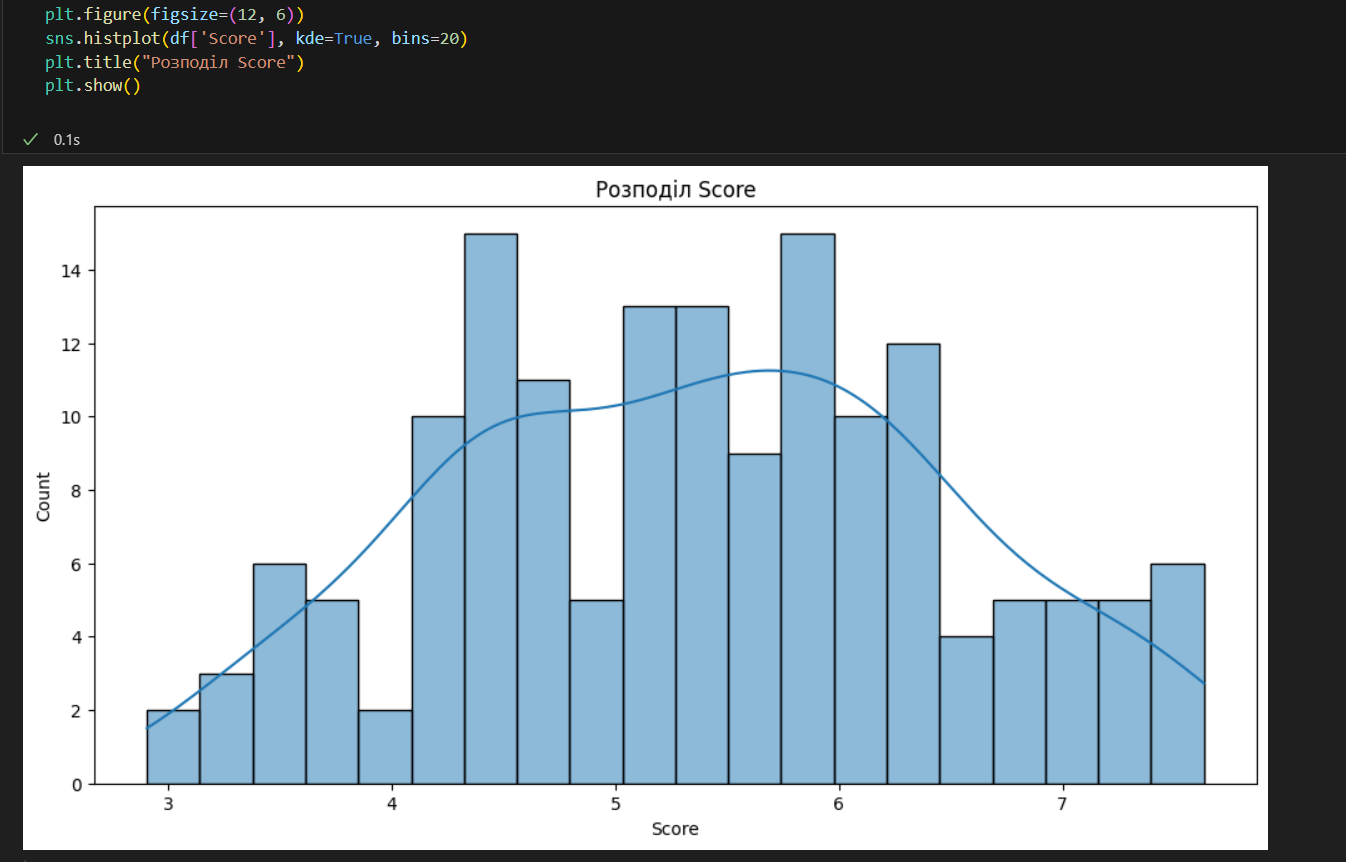
* Виявити аномальні значення (використати Z-score або IQR).

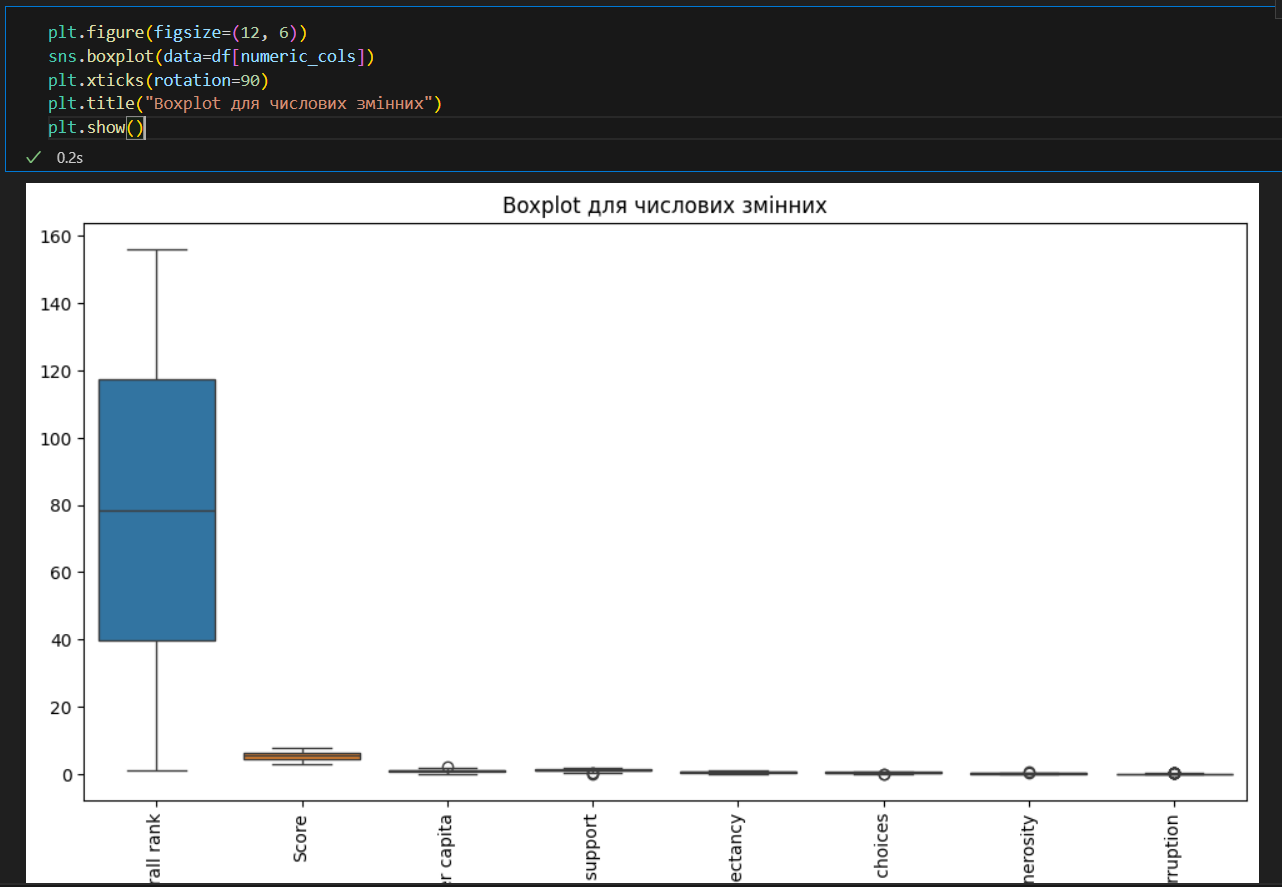


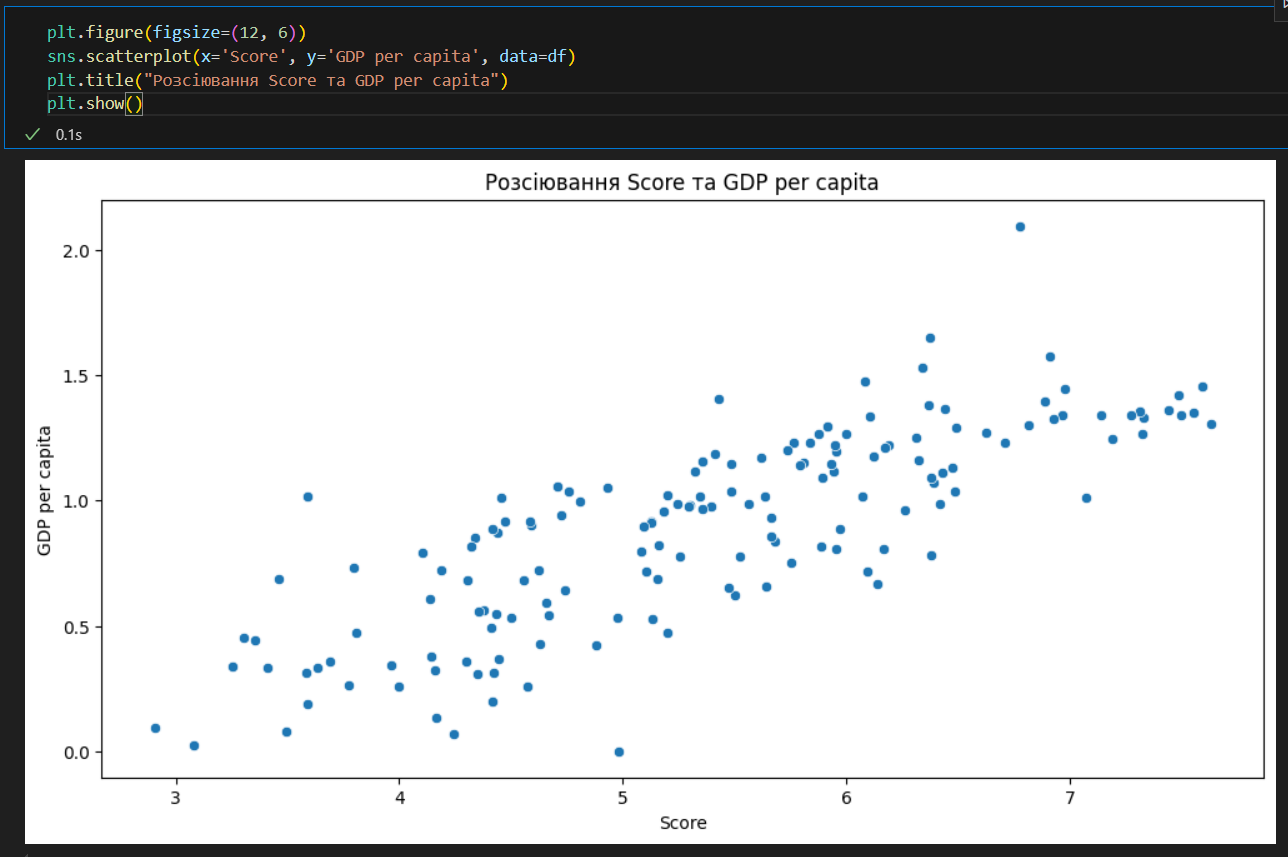
**Візуальний аналіз даних**

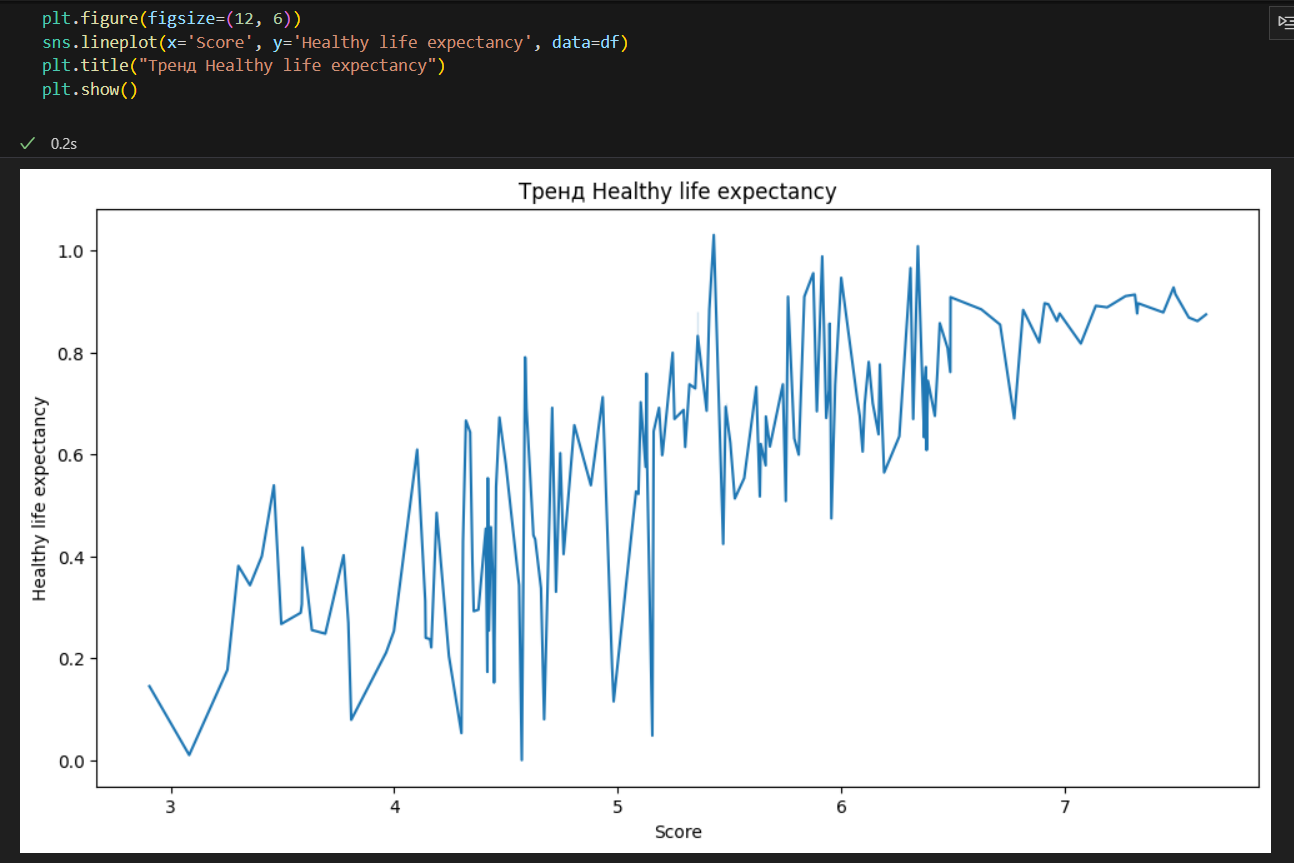
Побудувати графіки за допомогою бібліотек Matplotlib, Pandas Visualization, Seaborn. Пояснити, що зображено на графіках, що відкладено по осях, як називається графік. Див. код вище. Теоретичні відомості.

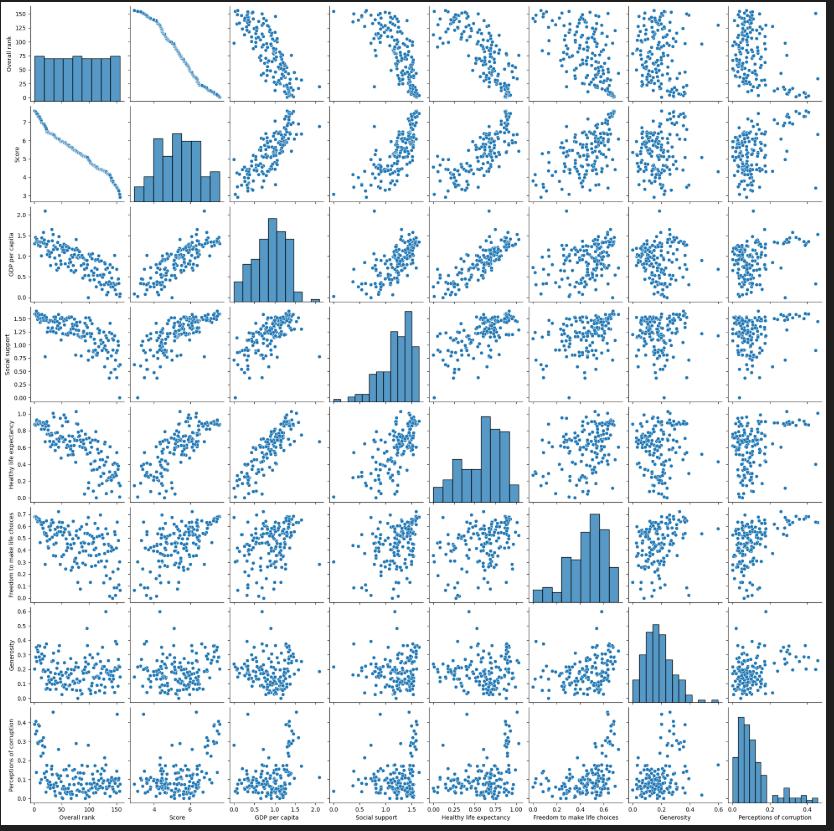
Для dataset побудувати щонайменше 8 видів графіків:





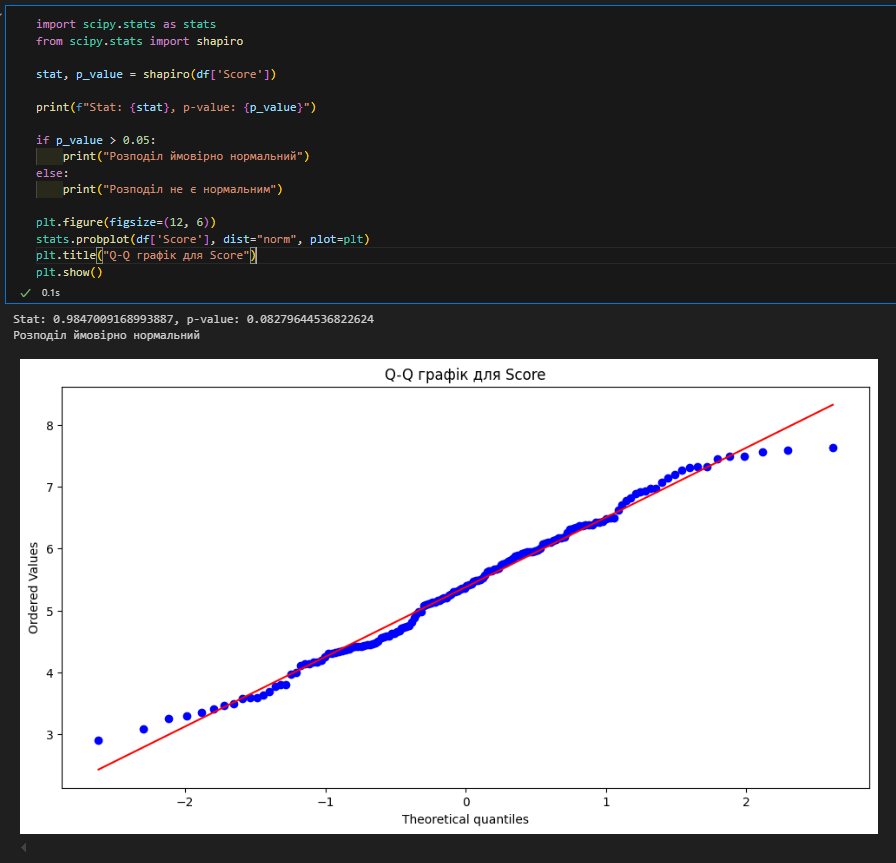






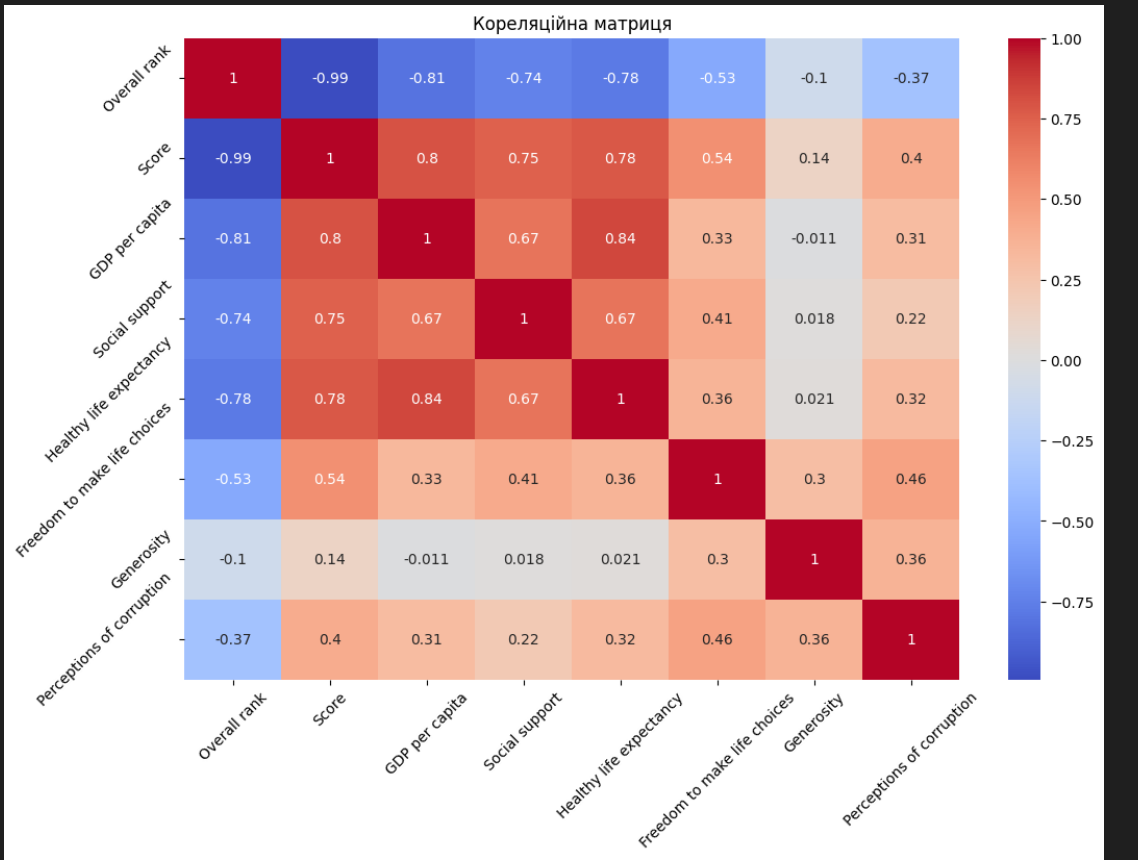
Перевірка нормальності розподілу даних у вибраних колонках Вибираємо колонку для аналізу Виконуємо Тест Шапіро-Уїлка або (Андерсона-Дарлінга чи Колмогорова-Смірнова). Через порівняння статистичних параметрів, візуалізації та результатів тестів робимо висновок про нормальний розподіл даних в колонці.

Додатково можна використати Q-Q графік змінної з вибраної колонки.

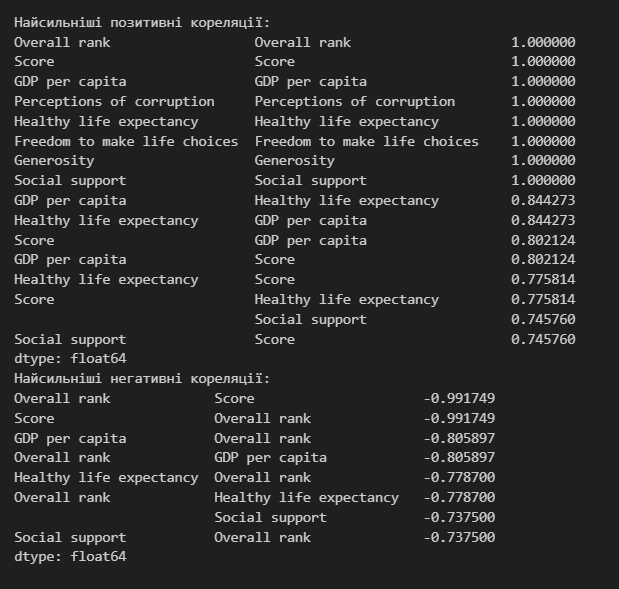


**Кореляційний аналіз**

1. Побудувати кореляційну матрицю (df.corr()).

2. Візуалізувати її за допомогою теплової карти (heatmap). 

3. Визначити найсильніші позитивні та негативні кореляції.



Кореляція та коваріація

df.corr(method='pearson') – кореляція Пірсона

df.corr(method='spearman') – кореляція Спірмена

df.corr(method='kendall') – кореляція Кендалла

df.cov() – коваріація

**Висновок**

Згідно з результатами проведеного аналізу даних, було виконано кілька важливих кроків для оцінки їх якості та нормальності розподілу. Спершу були перевірені наявні пропущені значення, які були заповнені середнім значенням для числових змінних. Далі було проведено статистичний аналіз, включаючи обчислення основних статистичних параметрів, таких як середнє, медіана, мінімум та максимум, а також виявлення аномальних значень через Z-score. Візуалізація даних включала кілька типів графіків, таких як гістограма, boxplot, теплова карта та графік парних порівнянь. Перевірка нормальності розподілу даних, проведена за допомогою тесту Шапіро-Уїлка, показала, чи можна припустити, що розподіл є нормальним, що має значення для подальшого статистичного аналізу. У результаті можна зробити висновок про нормальність або ненормальність розподілу вибраних змінних, що допомагає при виборі відповідних методів аналізу для подальшої роботи з даними.