**Лабораторна робота № 3.**

**Кореляційний аналіз даних**

**1. Мета роботи**

Метою лабораторної роботи є отримання практичних навичок з виконання кореляційного аналізу даних.

**2. Контекст**

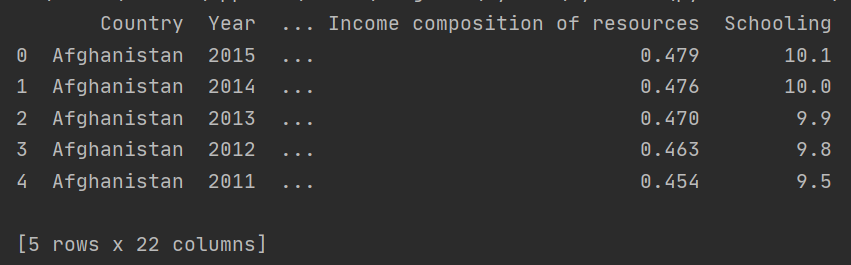
Хоча в минулому було проведено багато досліджень факторів, що впливають на очікувану тривалість життя, враховуючи демографічні змінні, структуру доходу та рівень смертності. Встановлено, що вплив імунізації та індекс людського розвитку раніше не враховувався. Крім того, деякі з минулих досліджень проводилися з урахуванням множинної лінійної регресії на основі набору даних за один рік для всіх країн. Отже, це дає мотивацію для вирішення обох факторів, зазначених раніше, шляхом формулювання регресійної моделі на основі моделі змішаних ефектів і множинної лінійної регресії, враховуючи дані за період з 2000 по 2015 рік для всіх країн. Також буде розглянуто важливу імунізацію, наприклад проти гепатиту В, поліомієліту та дифтерії. У двох словах, це дослідження буде зосереджено на факторах імунізації, факторах смертності, економічних факторах, соціальних факторах та інших факторах, пов’язаних зі здоров’ям. Оскільки спостереження в цьому наборі даних базуються на різних країнах, країні буде легше визначити фактор прогнозування, який сприяє нижчому значенню очікуваної тривалості життя. Це допоможе підказати країні, якій території слід надати значення, щоб ефективно покращити очікувану тривалість життя її населення.

Проект покладається на точність даних. Сховище даних Глобальної обсерваторії охорони здоров’я (GHO) під керівництвом Всесвітньої організації охорони здоров’я (ВООЗ) відстежує стан здоров’я, а також багато інших пов’язаних факторів для всіх країн. Набори даних стають доступними для громадськості з метою аналізу даних про здоров’я. Набір даних, пов’язаний із очікуваною тривалістю життя та факторами здоров’я для 193 країн, було зібрано з того самого веб-сайту сховища даних ВООЗ, а відповідні економічні дані – з веб-сайту Організації Об’єднаних Націй. Серед усіх категорій факторів, пов'язаних зі здоров'ям, були обрані лише ті критичні фактори, які є більш репрезентативними. Було помічено, що за останні 15 років у секторі охорони здоров’я відбувся величезний розвиток, що призвело до покращення рівня смертності людей, особливо в країнах, що розвиваються, порівняно з останніми 30 роками. Тому в цьому проекті ми розглянули дані за 2000-2015 роки для 193 країн для подальшого аналізу. Окремі файли даних були об’єднані в єдиний набір даних. При первинному візуальному огляді даних було виявлено деякі відсутні значення. Оскільки набори даних надходили від ВООЗ, ми не виявили очевидних помилок. Результат показав, що більшість відсутніх даних стосуються населення, гепатиту В і ВВП. Відсутні дані були з менш відомих країн, таких як Вануату, Тонга, Того, Кабо-Верде тощо. Знайти всі дані для цих країн було важко, тому було вирішено виключити ці країни з остаточного набору даних моделі. Остаточний об’єднаний файл (остаточний набір даних) складається з 22 стовпців і 2938 рядків, що означало 20 прогнозних змінних. Потім усі прогнозні змінні були розділені на кілька широких категорій: фактори, пов’язані з імунізацією, фактори смертності, економічні фактори та соціальні фактори.

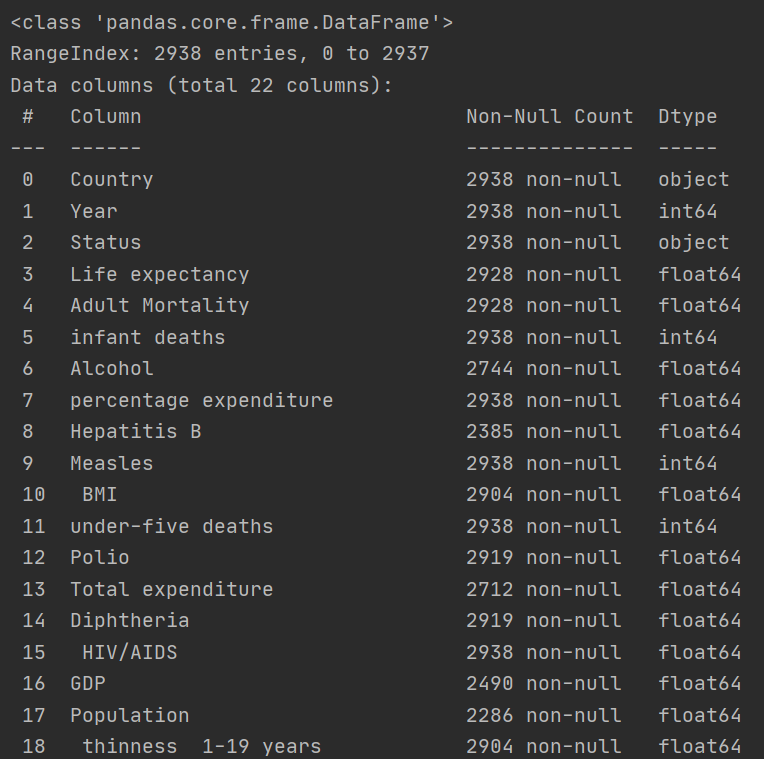
**2. Завдання до виконання**

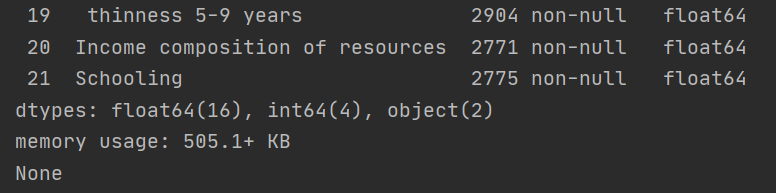
* Завантажити файл з даними у папку проекту

<https://drive.google.com/file/d/1jCvIDccQFednMlgA84ErFgS9EzDotHEQ/view?usp=sharing>

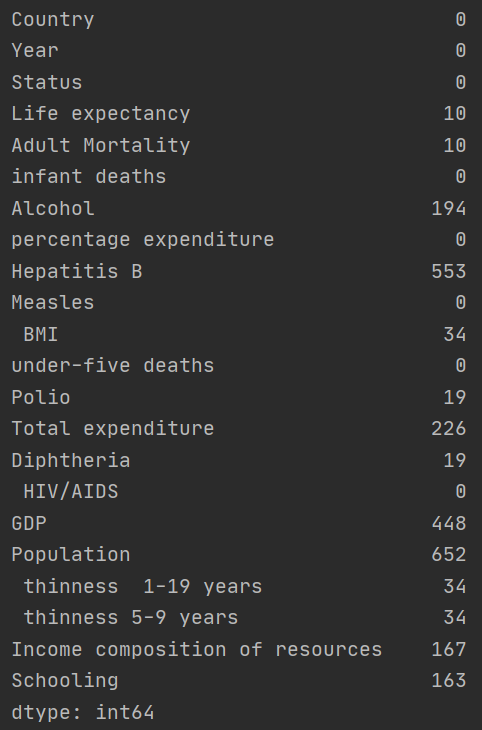


* Зробити розвідувальний аналіз даних (EDA). Побудувати графіки розподілів факторів – histplot (з доданням rugplot), boxplot з пакети seaborn.

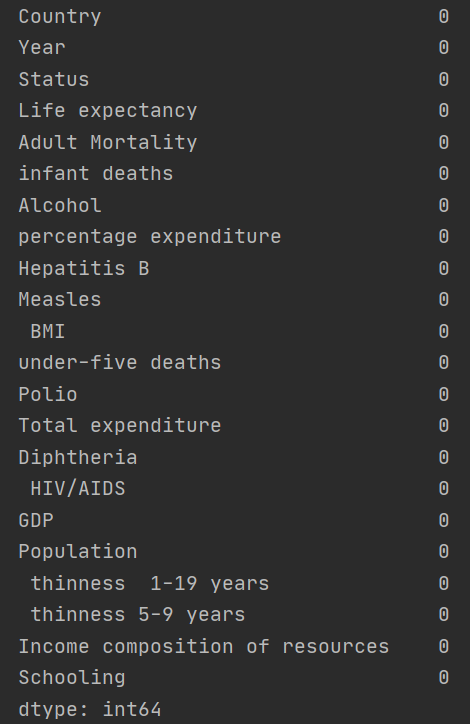


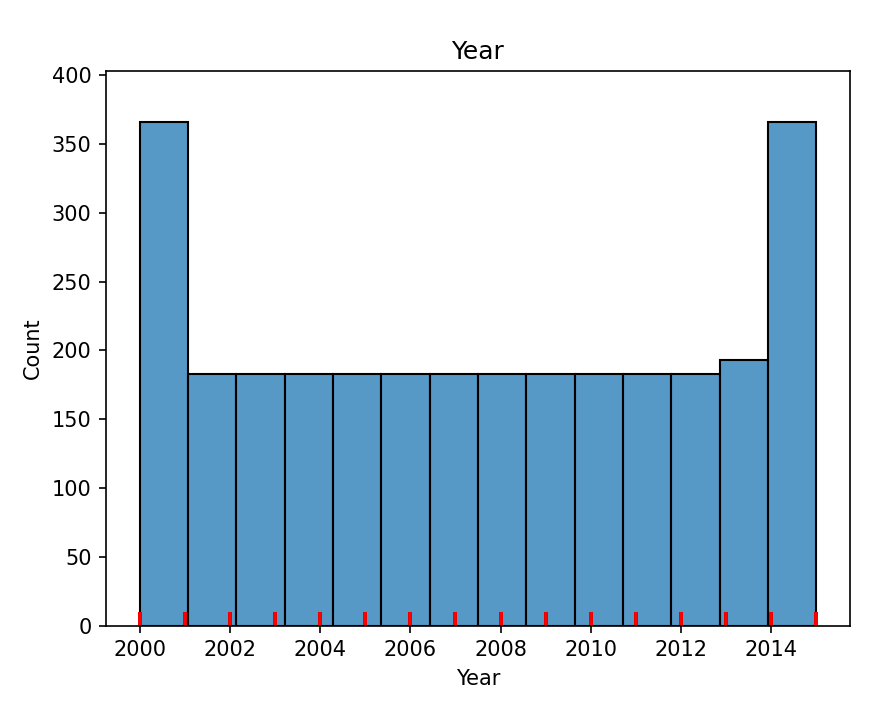


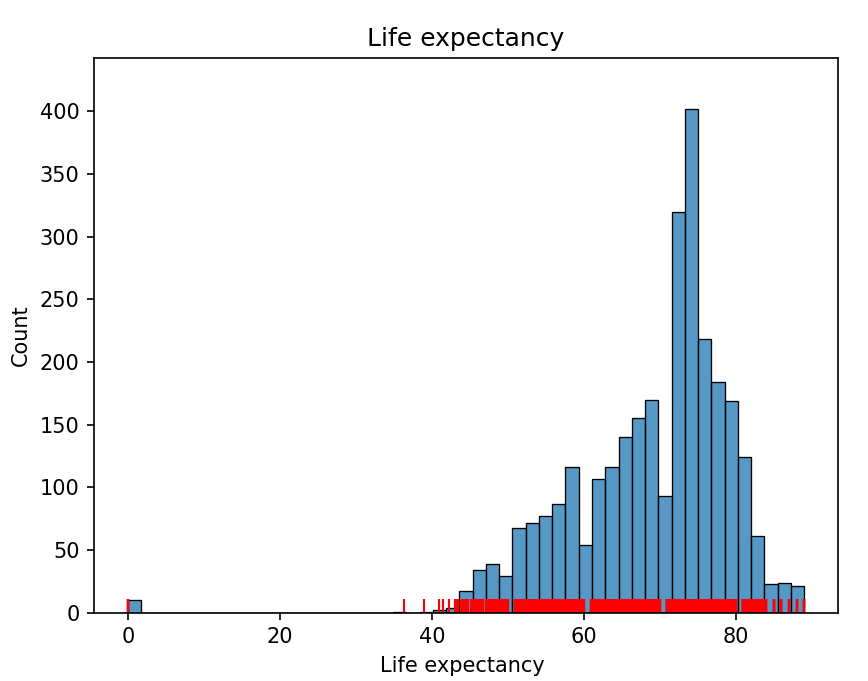
Кількість значень Nan

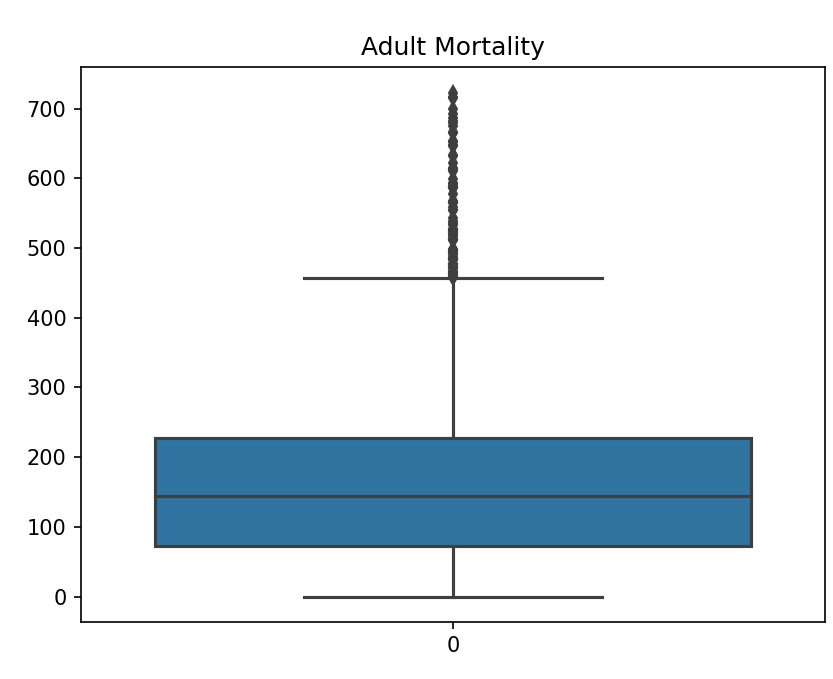


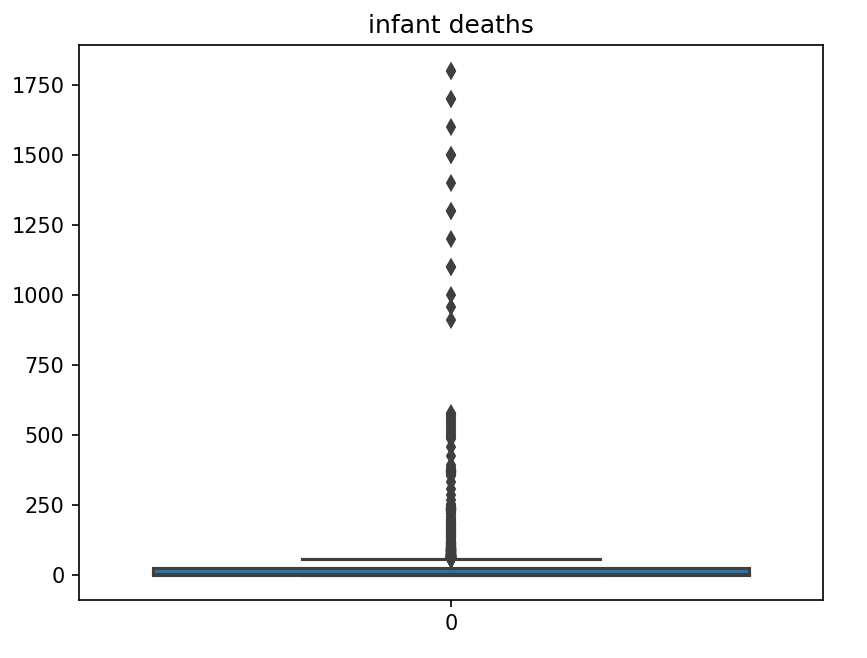
Заповнимо 0ми та перевіремо ще раз

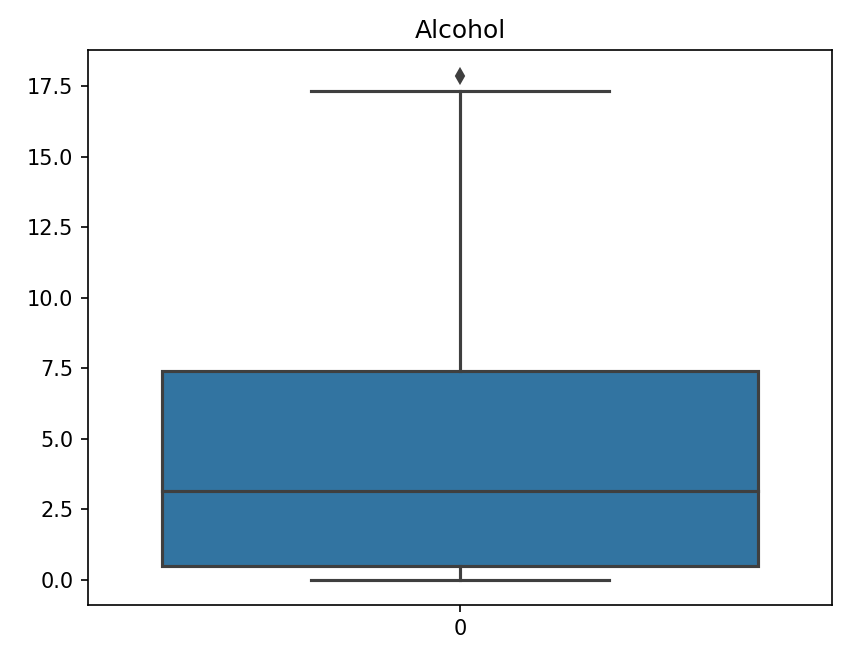




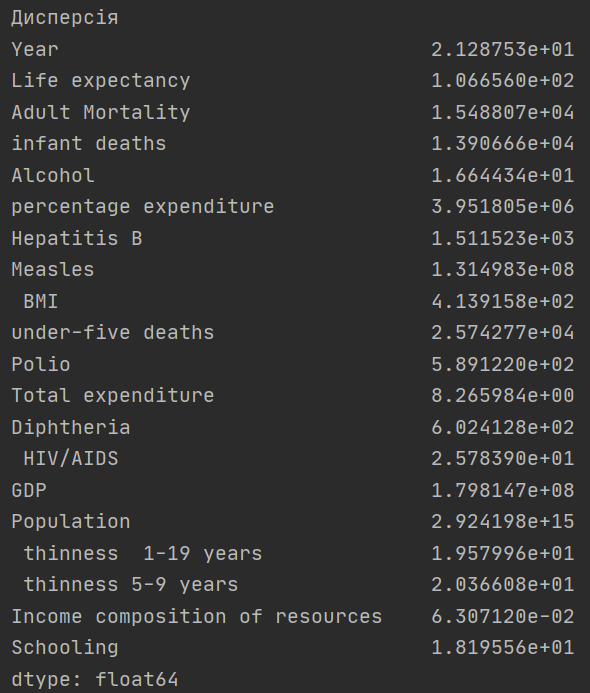


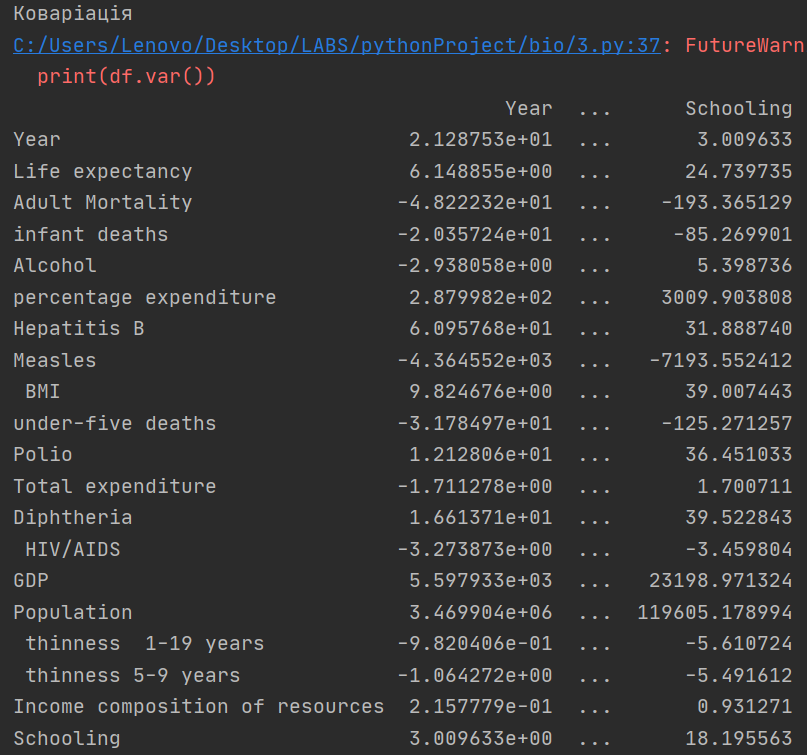


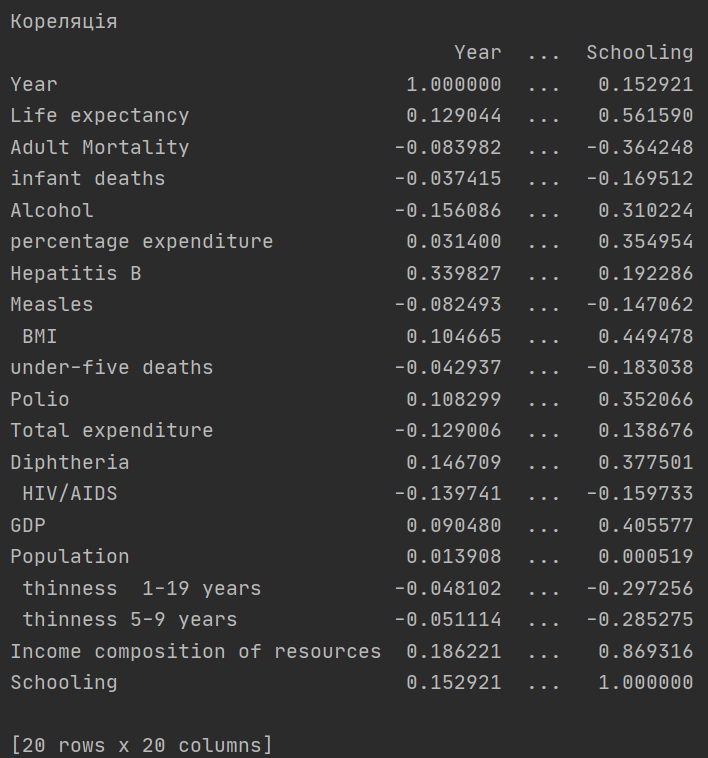


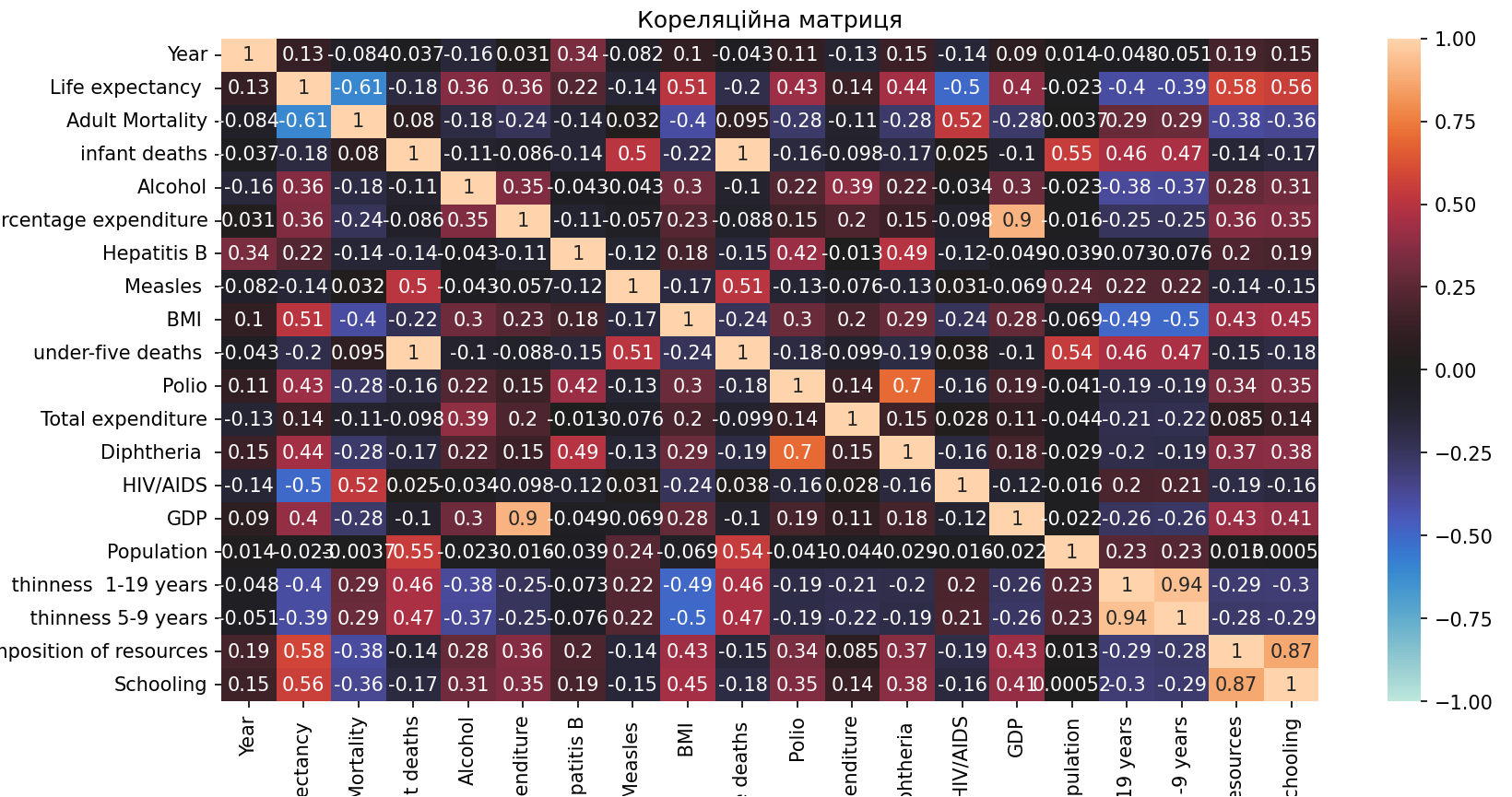


* Обрахувати – дисперсію, коваріацію та кореляцію по факторах. Зробити порівняльний аналіз між факторами.







* Провести кореляційний аналіз даних, шукаючи відповіді на наступні питання:  
    
  1. Чи справді різні фактори прогнозування, які були обрані спочатку, впливають на очікувану тривалість життя? Які передбачувані змінні насправді впливають на тривалість життя?



Найбільше впливає на тривалість життя infant deaths, BMI, HIV/AIDS, Income composition of resources,Schooling.

* 1. Чи повинна країна з нижчим значенням очікуваної тривалості життя (<65) збільшити витрати на охорону здоров’я, щоб покращити середню тривалість життя?

 Отже, так, чим більше витрат на охорону здоров'я, тим довша середня тривалість життя.

* 1. Як рівень дитячої та дорослої смертності впливає на тривалість життя?

Доросла смертність має найбільший вплив, , а дитяча – несуттєвий .

* 1. Чи має очікувана тривалість життя позитивну чи негативну кореляцію зі звичками в їжі, способом життя, фізичними вправами, курінням, вживанням алкоголю тощо.

Алкоголь має достатній вплив .Інших даних в таблиці немає.

* 1. Який вплив навчання в школі на тривалість життя людини?

Має великий вплив .

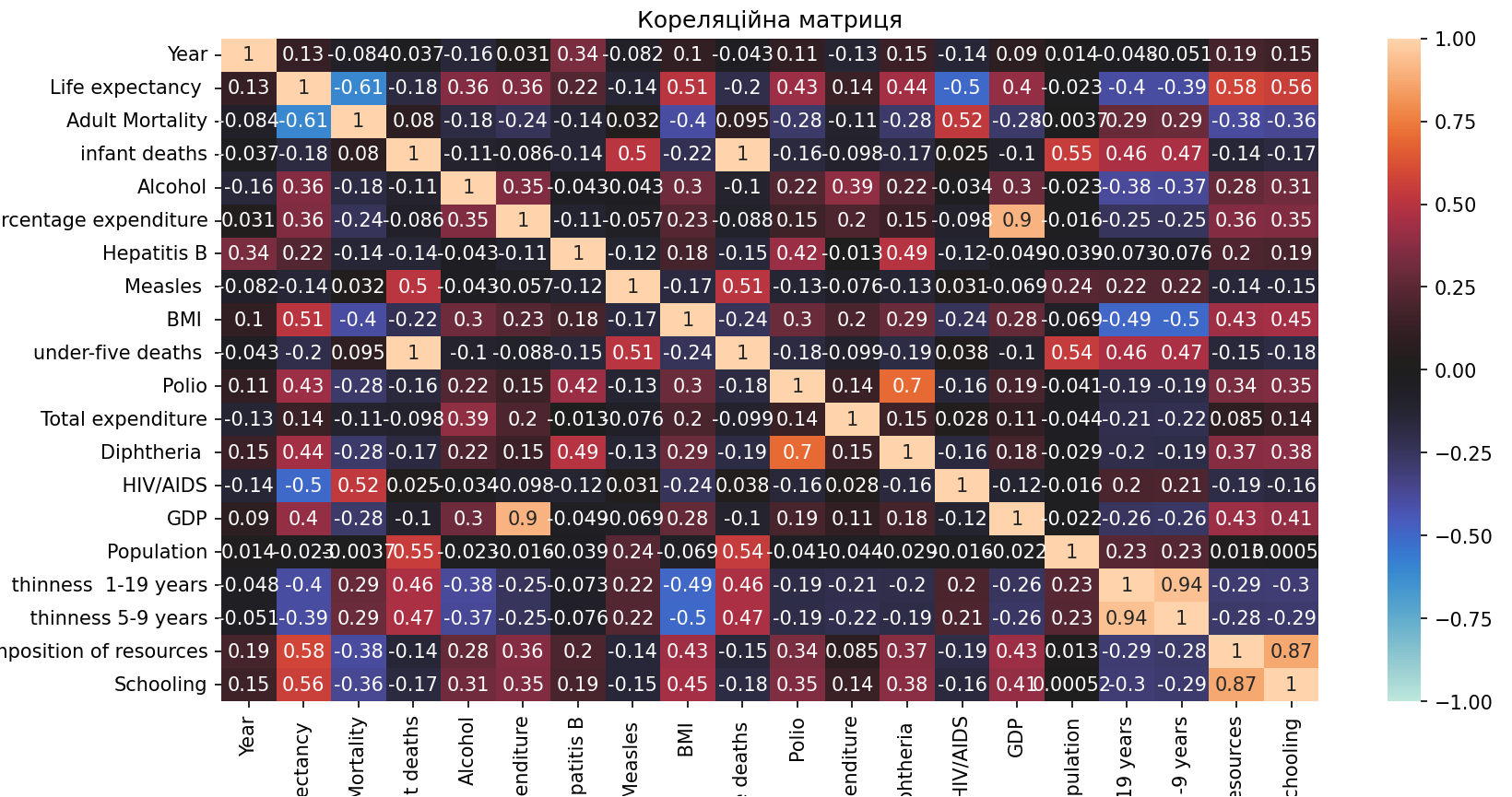
* 1. Позитивний чи негативний зв’язок очікуваної тривалості життя з вживанням алкоголю?

На диво, кореляція позитивна .

* 1. Чи зазвичай у густонаселених країнах тривалість життя нижча?

Кількість населення не має ніякого впливу на тривалість життя 

* 1. Який вплив охоплення імунізацією на очікувану тривалість життя?  
     Вплив гепатиту та кіру невеликий .Поліоеміліту достатній  Дифтерії та ВІЧ/СПІД також . Отже імунізація має великий вплив та покращить середню тривалість життя.
* Агрегувати розрахунки впливу факторів на тривалість життя у матрицю та зробити heatmap засобами seaborn для її візуалізації.



Лістинг програми:

import matplotlib.pyplot as plt  
import pandas as pd  
import seaborn as sns  
  
  
df = pd.read\_csv('Life Expectancy Data.csv')  
print(df.head())  
  
print(df.info())  
print(df.isnull().sum())  
df = df.fillna(0)  
print(df.isnull().sum())  
  
sns.histplot(df['Year'])  
sns.rugplot(df['Year'], color='red')  
plt.title('Year')  
plt.show()  
  
sns.histplot(df['Life expectancy '])  
sns.rugplot(df['Life expectancy '], color='red')  
plt.title('Life expectancy')  
plt.show()  
  
sns.boxplot(df['Adult Mortality'])  
plt.title('Adult Mortality')  
plt.show()  
  
sns.boxplot(df['infant deaths'])  
plt.title('infant deaths')  
plt.show()  
  
sns.boxplot(df['Alcohol'])  
plt.title('Alcohol')  
plt.show()  
  
print('Дисперсія')  
print(df.var())  
print('Коваріація')  
print(df.cov())  
print('Кореляція')  
corr = df.corr()  
print(corr)  
  
sns.heatmap(corr, annot = True, vmin=-1, vmax=1, center= 0)  
plt.title('Кореляційна матриця')  
plt.show()

**3. Необхідні пакети**

pandas, numpy, seaborn, jupyter

**4. Супровідні матеріали**<https://numpy.org/doc/stable/reference/generated/numpy.cov.html>

<https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html>

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html>

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html>

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kendalltau.html>