

▼ Пошуковий аналіз даних

Мета

Ознайомитись з методами перевірки статистичних гіпотез. Після завершення цієї лабораторної роботи ви зможете:

- Досліджувати дані за допомогою візуалізацій
- Робити описовий аналіз
- Групувати дані для аналізу
- Знаходити зв'язок між ознаками
- Перевіряти гіпотези про значущість коефіцієнта кореляції та про вигляд закону розподілу
- Робити дисперсійний аналіз



Завдання, що оцінюються

1. Скачати дані із файлу, який зберегли наприкінці попередньої роботи (з виправленими помилками та заповненими пропусками). Записати дані у dataframe. Дослідити ознаки з метою виявлення зв'язку між ними, побудувавши їх візуалізації. Візуально оцініть наявність та силу зв'язку між ознаками.
2. Порахувати кореляцію між всіма кількісними ознаками
3. Побудувати діаграми розсіювання для кількісних ознак та 'CO2 emission'. Які кількісні ознаки можуть бути предикторами кількості викидів CO2?
4. Побудувати діаграму розмаху для 'CO2 emission' по регіонам.
5. Виконати дисперсійний аналіз для кількості викидів CO2, згрупувати дані по регіонам. Чи може регіон бути предиктором для кількості викидів CO2?




Завдання #1:

Дослідити ознаки з метою виявлення зв'язку між ними, побудувавши їх візуалізації

Зчитую дані з файлу у датафрейм

```
# Напишіть ваш код нижче та натисніть Shift+Enter для виконання
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

path = 'clean_data2.csv'
df = pd.read_csv(path)
df.head()
```

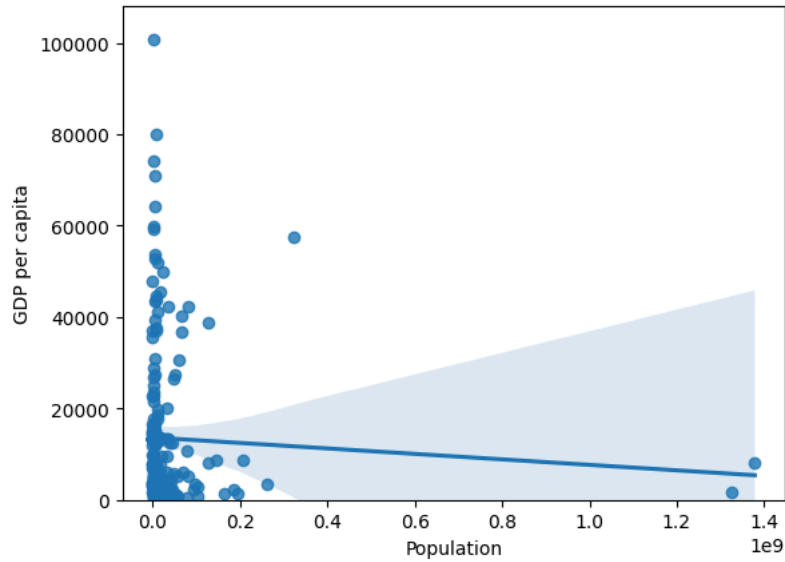


	Country	Name	Region	GDP per capita	Population	CO2 emission	Area	Population Density
0		Afghanistan	South Asia	561.778746	34656032.0	9809.225000	652860.0	53.083405
1		Albania	Europe & Central Asia	4124.982390	2876101.0	5716.853000	28750.0	100.038296
2		Algeria	Middle East & North Africa	3916.881571	40606052.0	145400.217000	2381740.0	17.048902
3		American Samoa	East Asia & Pacific	11834.745230	55599.0	165114.116337	200.0	277.995000
4		Andorra	Europe & Central Asia	36988.622030	77281.0	462.042000	470.0	164.427660

Будую діаграми

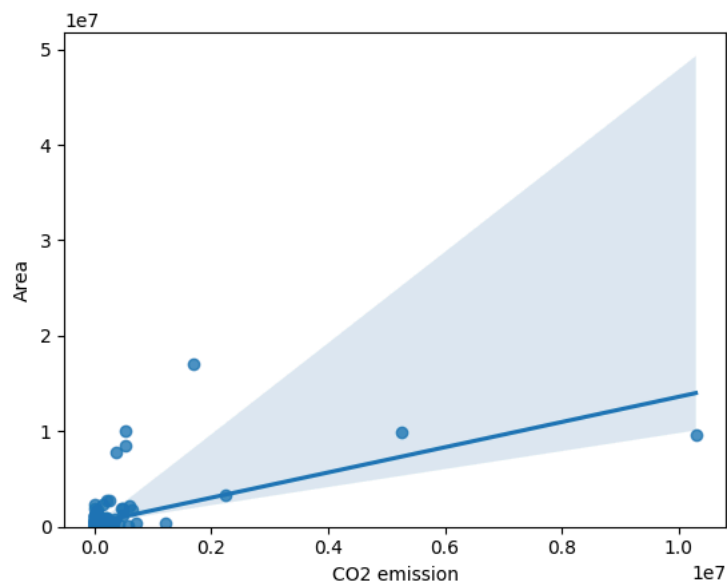
```
# Напишіть ваш код нижче та натисніть Shift+Enter для виконання
sns.regplot(x='Population', y='GDP per capita', data=df)
plt.ylim(0,)
```

↗ (0.0, 108027.32707098429)



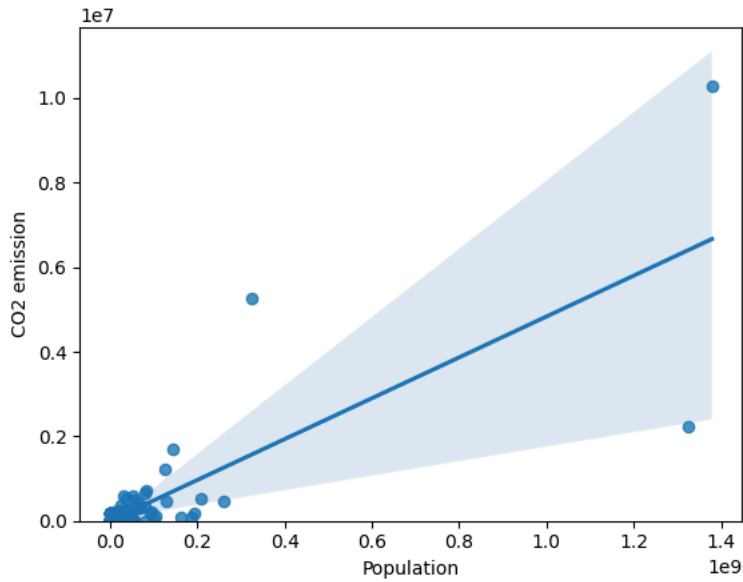
```
sns.regplot(x='CO2 emission', y='Area', data=df)
plt.ylim(0,)
```

↗ (0.0, 51736177.25081582)



```
sns.regplot(x='Population', y='CO2 emission', data=df)
plt.ylim(0,)
```

(0.0, 11662716.657024726)



Візуально оцініть наявність та силу зв'язку між ознаками.

▼

## Завдання #2:

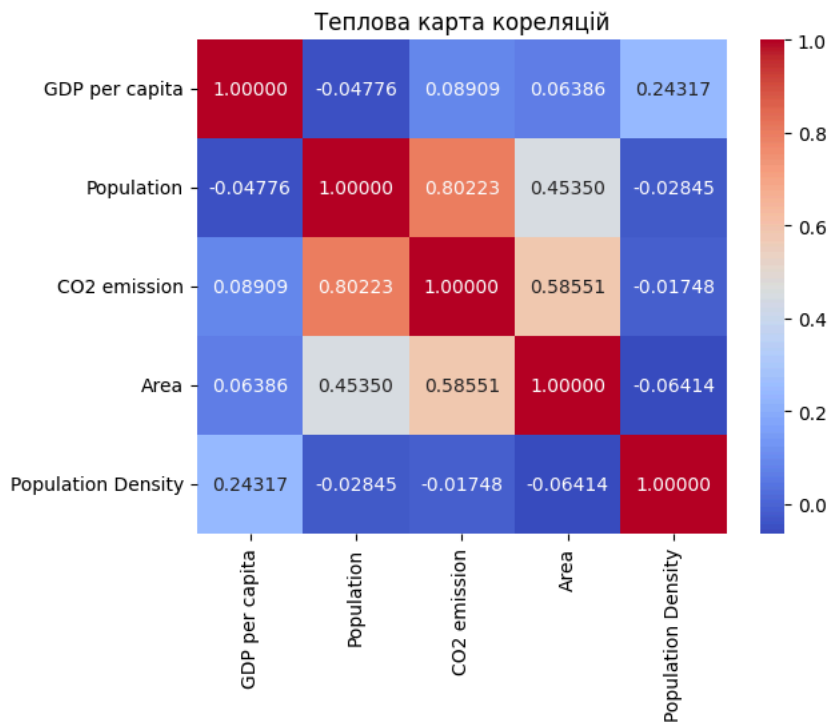
Порахувати кореляцію між всіма кількісними ознаками

Рахую кореляцію між всіма кількісними ознаками

```
# Напишіть ваш код нижче та натисніть Shift+Enter для виконання
corr_matrix = df.corr(numeric_only=True)

sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.5f')
plt.title('Теплова карта кореляцій')
plt.show()

df.corr(numeric_only=True)
```



	GDP per capita	Population	CO2 emission	Area	Population Density
GDP per capita	1.000000	-0.047759	0.089094	0.063861	0.243174
Population	-0.047759	1.000000	0.802232	0.453500	-0.028449
CO2 emission	0.089094	0.802232	1.000000	0.585512	-0.017476
Area	0.063861	0.453500	0.585512	1.000000	-0.064138
Population Density	0.243174	-0.028449	-0.017476	-0.064138	1.000000



### Завдання #3:

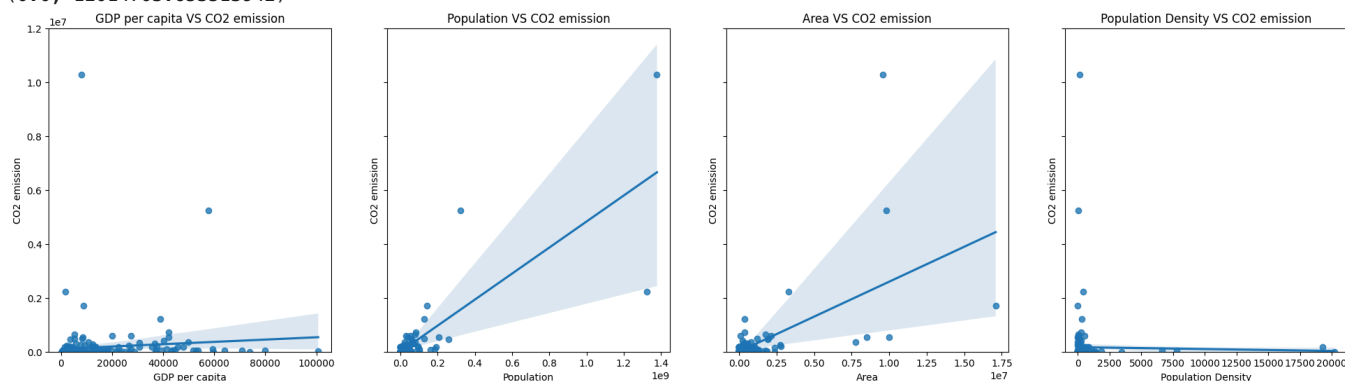
Побудувати діаграми розсіювання для кількісних ознак та 'CO2 emission'. Візуально оцінити наявність та силу зв'язку між цими ознаками.

Будую діаграму розсіювання для кількісних ознак та 'CO2 emission'

```
# Напишіть ваш код нижче та натисніть Shift+Enter для виконання
fig, axes = plt.subplots(1, 4, figsize=(24,6), sharey=True)
x_vars = ['GDP per capita', 'Population', 'Area', 'Population Density']
for ax, x_vars in zip(axes, x_vars):
    sns.regplot(x=x_vars, y='CO2 emission', data=df, ax=ax)
    ax.set_title(f'{x_vars} VS CO2 emission')
plt.ylim(0,)
```



(0.0, 12014703.653513942)



Які кількісні ознаки можуть бути предикторами кількості викидів CO2?

- Оскільки зв'язок між CO2 emission та GDP per capita не дуже виражений, доволі слабкий, що дає підставу вважати, що GDP per capita не є сильним та хорошим предиктором

- Зв'язки між Population та CO2 emission, Area та CO2 emission як ми бачимо на графіку є доволі сильним, хоч і у другому випадку зв'язок дещо слабший ніж у першому, проте цілком можна вважати, що Population та Area є хорошими предикторами CO2 emission

- Зв'язок між Population Density та CO2 emission є дуже слабкий, майже повністю відсутній, тому ознака Population Density не може слугувати предиктором CO2 emission

Обчислюю коефіцієнт кореляції Пірсона та P-value для всіх кількісних змінних та 'CO2 emission'

```
# Напишіть ваш код нижче та натисніть Shift+Enter для виконання
from scipy import stats
```

```
def pirson_corr(column):
    pearson_coef, p_value = stats.pearsonr(df[column], df['CO2 emission'])
    print(f'{column} VS CO2 emission')
    print('Pearson: {:.5f}'.format(pearson_coef))
    print('P-value: {:.5g}\n'.format(p_value))
```

```
numeric_df = df.select_dtypes(include='float64')
for column in numeric_df.columns:
    pirson_corr(column)
```

```
➡ GDP per capita VS CO2 emission
Pearson: 0.08909
P-value: 0.19106
```

```
Population VS CO2 emission
Pearson: 0.80223
P-value: 4.6379e-50
```

```
CO2 emission VS CO2 emission
Pearson: 1.00000
P-value: 0
```

```
Area VS CO2 emission
Pearson: 0.58551
P-value: 2.3157e-21
```

```
Population Density VS CO2 emission
Pearson: -0.01748
P-value: 0.79797
```

Кількісні ознаки, які можуть бути предикторами кількості викидів CO2: Population та Area

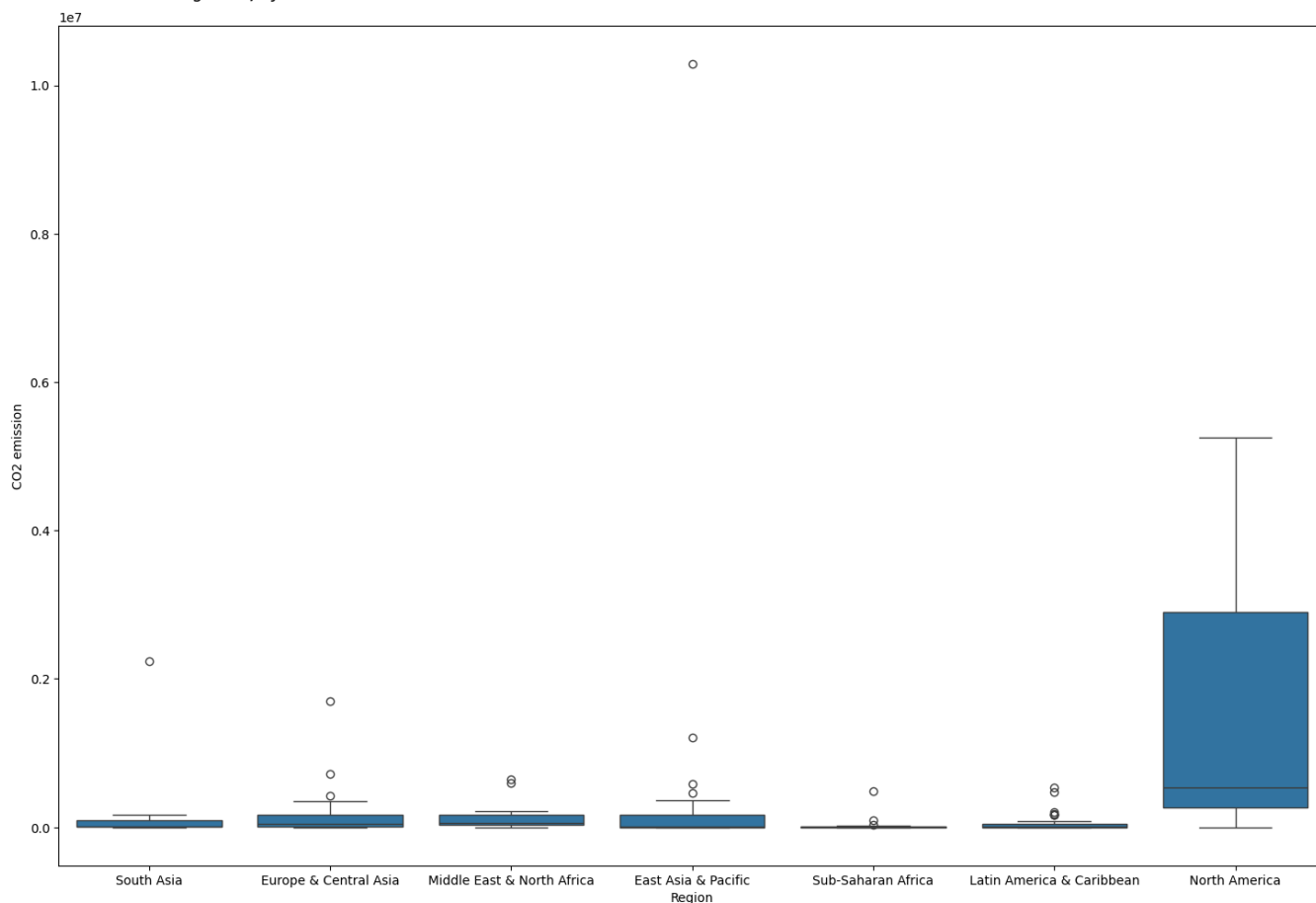
✓

## Завдання #4:

Побудувати діаграму розмаху для 'CO2 emission' по регіонам.

```
# Напишіть ваш код нижче та натисніть Shift+Enter для виконання
plt.figure(figsize=(18,12))
sns.boxplot(x='Region', y='CO2 emission', data=df)
```

<Axes: xlabel='Region', ylabel='CO2 emission'>



▼

## Завдання #5:

Виконати дисперсійний аналіз для кількості викидів CO2, згрупувати дані по регіонам

Групувати дані, щоб побачити чи впливає 'Region' на 'CO2 emission'.

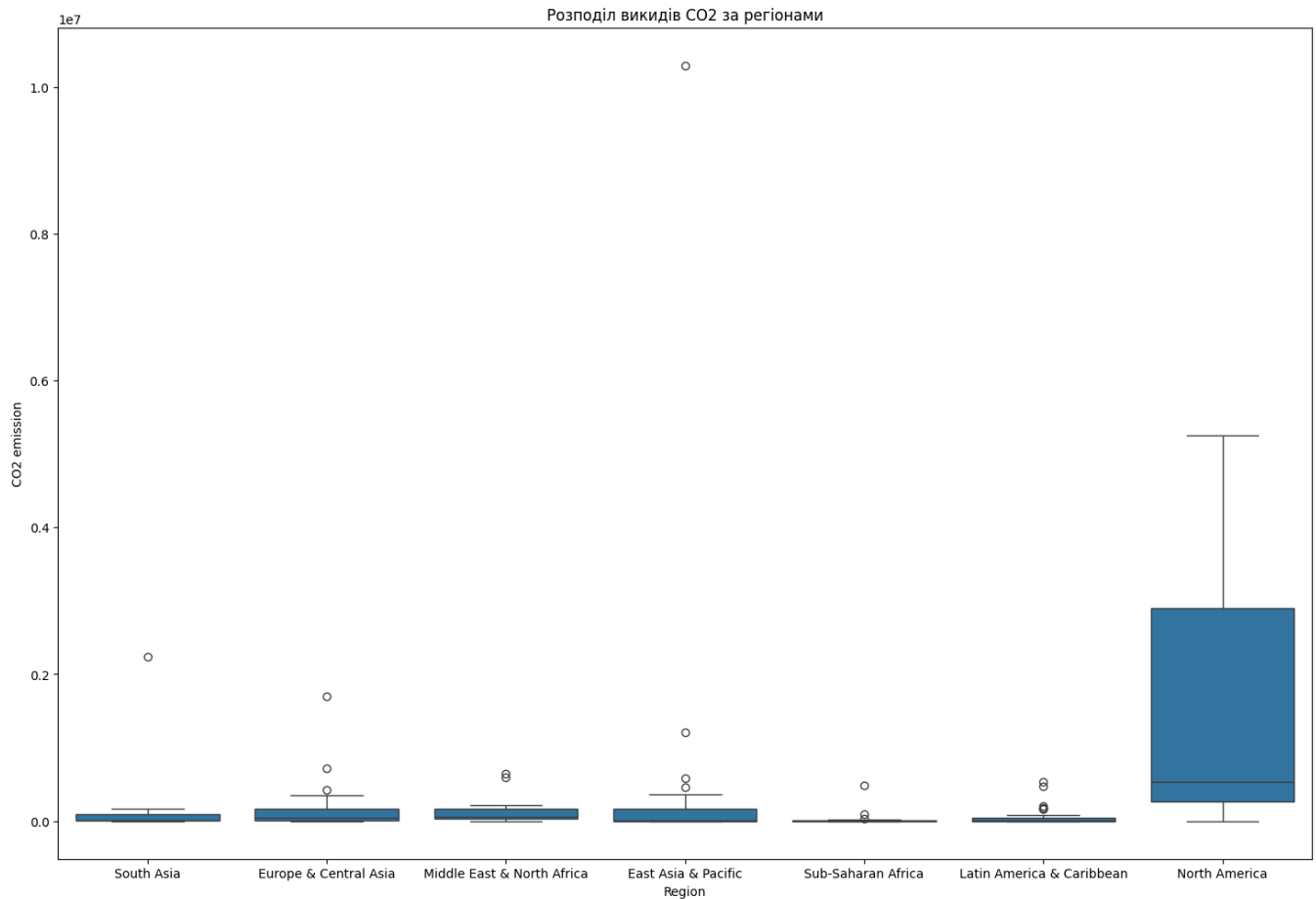
```
# Напишіть ваш код нижче та натисніть Shift+Enter для виконання
grouped_data = df.groupby('Region')['CO2 emission']
grouped_data.head(2)
```

```
0      9809.225000
1      5716.853000
2     145400.217000
3     165114.116337
4       462.042000
5      34763.160000
6       531.715000
7     204024.546000
10     361261.839000
14      31338.182000
15      73189.653000
20      6318.241000
21       575.719000
35     537193.498000
Name: CO2 emission, dtype: float64
```

Перевіряю розподіл даних в групах, щоб обрати вид дисперсійного аналізу.

```
# Напишіть ваш код нижче та натисніть Shift+Enter для виконання
plt.figure(figsize=(18,12))
sns.boxplot(x='Region', y='CO2 emission', data=df)
plt.title('Розподіл викидів CO2 за регіонами')
```

↗ Text(0.5, 1.0, 'Розподіл викидів CO2 за регіонами')



Для отримання F-test score та P-value скористаюсь функцією `f_oneway` з модуля "stats", якщо розподіл даних в групах дозволяє застосувати класичний дисперсійний аналіз, або `kruskal` з модуля "stats" для непараметричного дисперсійного аналізу Красскела-Уоліса.

```
from scipy import stats
```

```
# Напишіть ваш код нижче та натисніть Shift+Enter для виконання
f_test, p_value = stats.f_oneway(*grouped_data.apply(list))
print('F-statistic:', f_test)
print('P-value:', p_value)
```

↗ F-statistic: 3.567709637673427  
P-value: 0.0021855506878927533

Результат із `F_test` показником тесту, який показує сильну кореляцію, і `P-value` 0.0021 показує, що є статистична значущість. Але чи означає це, що досліджувані групи значуще відрізняються між собою?

Розглянемо їх окремо.

```
# Напишіть ваш код нижче та натисніть Shift+Enter для виконання
grouped_data_list = grouped_data.apply(list)
for i, region_item in enumerate(grouped_data_list.items()):
    region_name, region_data = region_item
```

```

for region_name2, region_data2 in grouped_data_list[i:].items():
    if region_name != region_name2:
        f_statistic, p_value = stats.f_oneway(region_data, region_data2)
        print(f"\n{region_name} - {region_name2}:")
        print("F-statistic:", f_statistic)
        print("P-value:", p_value)

```



F-statistic: 2.456104756327821  
P-value: 0.12029220254983197

Europe & Central Asia – Middle East & North Africa:  
F-statistic: 0.025434686706224247  
P-value: 0.8737062957355305

Europe & Central Asia – North America:  
F-statistic: 27.185776372971752  
P-value: 2.4871297063504576e-06

Europe & Central Asia – South Asia:  
F-statistic: 2.1449889489536647  
P-value: 0.14793124275570288

Europe & Central Asia – Sub-Saharan Africa:  
F-statistic: 7.930055837622809  
P-value: 0.0058179788422628425

Latin America & Caribbean – Middle East & North Africa:  
F-statistic: 4.00028376576551  
P-value: 0.04995035790485582

Latin America & Caribbean – North America:  
F-statistic: 24.47756779640105  
P-value: 1.2020263394688284e-05

Latin America & Caribbean – South Asia:  
F-statistic: 4.454047216844204  
P-value: 0.04005521026835715

Latin America & Caribbean – Sub-Saharan Africa:  
F-statistic: 3.9243756470814195  
P-value: 0.05071332517066309

Middle East & North Africa – North America:  
F-statistic: 10.785908142809197  
P-value: 0.003386665615272284

Middle East & North Africa – South Asia:  
F-statistic: 1.0781001825971532  
P-value: 0.30833268792682356

Middle East & North Africa – Sub-Saharan Africa:  
F-statistic: 14.816051348200512  
P-value: 0.00026718681820385405

North America – South Asia:  
F-statistic: 2.4462636855311657  
P-value: 0.1522433374176963

North America – Sub-Saharan Africa:  
F-statistic: 29.889184991922967  
P-value: 1.5319132629213808e-06

South Asia – Sub-Saharan Africa:  
F-statistic: 7.300524087267293  
P-value: 0.009192711202178152



## Додаткове завдання:

Дайте відповіді на питання

1. По результатам дисперсійного аналізу для кількості викидів CO2 по регіонам, вкажіть пару регіонів, що відрізняються найсильніше.
2. Створіть якісну ознаку 'Rich country', згрупувавши дані 'GDP per capita' в кілька категорій (багаті-бідні країни, 3-5 категорій). Побудуйте діаграму розмаху для 'CO2 emission' по категоріям 'Rich country'. Візуально оцініть наявність зв'язку між цими ознаками.
3. Виконайте дисперсійний аналіз для 'CO2 emission', згрупувавши дані по категоріям 'Rich country'.

► Натисніть тут, щоб побачити підказку

# Напишіть ваш код нижче та натисніть Shift+Enter для виконання



```

max_f_stat = float('-inf')
pair = None
max_p_value = None

group_names = list(grouped_data.groups.keys())

for i, region1 in enumerate(group_names):
    group1 = grouped_data.get_group(region1)
    for j, region2 in enumerate(group_names[i+1:], start=i+1):
        group2 = grouped_data.get_group(region2)
        f_stat, p_value = stats.f_oneway(group1, group2)

        if f_stat > max_f_stat and p_value < 0.05:
            max_f_stat = f_stat
            pair = (region1, region2)
            max_p_value = p_value

print(pair)

↗ ('North America', 'Sub-Saharan Africa')

plt.hist(df['GDP per capita'], bins='auto')
plt.xlabel('GDP per capita')
plt.ylabel('Count')
plt.grid()

df['Rich country'] = pd.cut(df['GDP per capita'], bins=[0, 20000, 50000, 100000], labels=['Poor', 'Middle', 'Rich'])
plt.figure(figsize=(10,10))
sns.boxplot(x='Rich country', y='CO2 emission', data=df)
plt.xlabel('Категорія країни за рівнем доходу')
plt.ylabel('CO2 emission')
plt.title('Діаграма розмаху CO2 emission для різних категорій країн')
plt.grid()

```

