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

Практикум № 3

з курсу «Аналіз даних в інформаційних системах»

на тему: «Описова статистика»

|  |  |
| --- | --- |
| Викладач:  Ліхоузова Т. А. | Виконав:  студент 2 курсу  групи ІП-11 ФІОТ  Головня О. Р. |

Київ-2023

**Основне завдання**

1. Скачати дані із файлу Data2.csv
2. Записати дані у data frame
3. Дослідити структуру даних
4. Виправити помилки в даних
5. Побудувати діаграми розмаху та гістограми
6. Додати стовпчик із щільністю населення

**Додаткове завдання**

Відповісти на питання (файл Data2.csv):

1. Чи є пропущені значення? Якщо є, замінити середніми
2. Яка країна має найбільший ВВП на людину (GDP per capita)? Яка має найменшу площу?
3. В якому регіоні середня площа країни найбільша?
4. Знайдіть країну з найбільшою щільністю населення у світі? У Європі та центральній Азії?
5. Чи співпадає в якомусь регіоні середнє та медіана ВВП?
6. Вивести топ 5 країн та 5 останніх країн по ВВП та кількості СО2 на душу населення.

Для виконання лабораторної роботи було використано Python.

Код програми:

1. **Основне завдання:**

Записав дані у data frame, здійснив видалення мінусів, заміну ком на крапки, встановлення типів даних у стовпцях. Також здійснив заміну пропущених значень на середні (див. «Код програми»).

import pandas as pd

import matplotlib.pyplot as plt

# зчитування файлу та запис даних у DataFrame

df = pd.read\_csv("Data2.csv", delimiter=";")

# Зміна типів даних у стовпцях

df['Country Name'] = df['Country Name'].astype('string')

df['Region'] = df['Region'].astype('string')

df['GDP per capita'] = df['GDP per capita'].str.replace(',', '.').astype(float)

df['GDP per capita'] = df['GDP per capita'].astype('float')

mean\_populatiion = df['Populatiion'].mean()

df['Populatiion'] = df['Populatiion'].fillna(mean\_populatiion).astype(int)

df['Populatiion'] = df['Populatiion'].astype('int')

df['CO2 emission'] = df['CO2 emission'].str.replace(',', '.').astype(float)

df['CO2 emission'] = df['CO2 emission'].astype('float')

df['Area'] = df['Area'].str.replace(',', '').astype(int)

df['Area'] = df['Area'].astype('int')

# замінюємо від'ємні значення в стовпці Populatiion на середнє значення

mean\_populatiion = df['Populatiion'].mean()

df.loc[df['Populatiion'] < 0, 'Populatiion'] = mean\_populatiion

# замінюємо від'ємні значення в стовпці Сфзшеф на середнє значення

mean\_GDP = df['GDP per capita'].mean()

df.loc[df['GDP per capita'] < 0, 'GDP per capita'] = mean\_GDP

# замінюємо від'ємні значення в стовпці CO2 emission на середнє значення

mean\_CO2 = df['CO2 emission'].mean()

df.loc[df['CO2 emission'] < 0, 'CO2 emission'] = mean\_CO2

# замінюємо від'ємні значення в стовпці Area на середнє значення

mean\_area = df['Area'].mean()

df.loc[df['Area'] < 0, 'Area'] = mean\_area

# замінюємо NaN значення середнім значенням

mean\_GDP = df['GDP per capita'].mean()

df['GDP per capita'] = df['GDP per capita'].fillna(mean\_GDP)

# замінюємо NaN значення середнім значенням

mean\_CO2 = df['CO2 emission'].mean()

df['CO2 emission'] = df['CO2 emission'].fillna(mean\_CO2)

# заміна NaN значень у стовпці "Area" на середнє значення

mean\_area = df['Area'].mean()

df['Area'] = df['Area'].fillna(mean\_area).astype(int)

#print(f"{'Country Name':<40}{'Region':<40}{'GDP per capita':<20}{'Populatiion':<20}{'CO2 emission':<20}{'Area':<20}")

# виведення даних у форматованому вигляді

#for index, row in df.iterrows():

# print(f"{row['Country Name']:<40}{row['Region']:<40}{row['GDP per capita']:<20}{row['Populatiion']:<20}{row['CO2 emission']:<20}{row['CO2 emission']:<20}")

# побудова діаграм розмаху та гістограм для кожної змінної

for column in df.columns:

if df[column].dtype in ['float64', 'int64']:

fig, ax = plt.subplots(1, 2, figsize=(12,6))

ax[0].boxplot(df[column])

ax[0].set\_title(f"{column} Boxplot")

ax[0].set\_ylabel(column)

ax[1].hist(df[column], bins=30)

ax[1].set\_title(f"{column} Histogram")

ax[1].set\_xlabel(column)

plt.show()

# Діаграма розмаху та гістограма для Area

fig, ax = plt.subplots(1, 2, figsize=(12,4))

ax[0].boxplot(df['Area'])

ax[0].set\_title("Area Boxplot")

ax[0].set\_ylabel("Area")

ax[1].hist(df['Area'], bins=30)

ax[1].set\_title("Area Histogram")

ax[1].set\_xlabel("Area")

plt.show()

df['Population Density'] = df['Populatiion'] / df['Area']

#print(f"{'Country Name':<40}{'Region':<40}{'GDP per capita':<20}{'Populatiion':<20}{'CO2 emission':<20}{'Area':<20}{'Population Density':<20}")

# виведення даних у форматованому вигляді

#for index, row in df.iterrows():

# print(f"{row['Country Name']:<40}{row['Region']:<40}{row['GDP per capita']:<20}{row['Populatiion']:<20}{row['CO2 emission']:<20}{row['CO2 emission']:<20}{row['Population Density']:<20} ")

max\_gdp\_row = df.loc[df['GDP per capita'].idxmax()]

print('\nmax gdp row:')

print(max\_gdp\_row)

print('\nmin area row:')

min\_row = df.loc[df['Area'].idxmin()]

print(min\_row)

mean\_area\_by\_region = df.groupby('Region')['Area'].mean().sort\_values(ascending=False)

print(mean\_area\_by\_region)

largest\_region = mean\_area\_by\_region.index[0]

print(f"The region with the largest average country area is {largest\_region}\n")

max\_pd\_row = df.loc[df['Population Density'].idxmax()]

print(max\_pd\_row)

# Групування даних за регіоном

region\_groups = df.groupby("Region")

# Створення списку регіонів, для яких середнє та медіанне значення співпадають

matching\_regions = []

for region, data in region\_groups:

mean\_gdp = data["GDP per capita"].mean()

median\_gdp = data["GDP per capita"].median()

if mean\_gdp == median\_gdp:

matching\_regions.append(region)

print('\nСпівпадіння середнього і медіани:')

print(matching\_regions)

# Вибір топ 5 країн та останніх 5 країн по GDP per capita

top\_gdp = df.sort\_values(by='GDP per capita', ascending=False).head(5)

last\_gdp = df.sort\_values(by='GDP per capita').head(5)

# Вибір топ 5 країн та останніх 5 країн по кількості CO2 викидів

top\_co2 = df.sort\_values(by='CO2 emission', ascending=False).head(5)

last\_co2 = df.sort\_values(by='CO2 emission').head(5)

# Виведення результатів

print('Топ 5 країн по GDP per capita:')

print(top\_gdp[['Country Name', 'GDP per capita']])

print('\nОстанні 5 країн по GDP per capita:')

print(last\_gdp[['Country Name', 'GDP per capita']])

print('\nТоп 5 країн по кількості CO2 викидів:')

print(top\_co2[['Country Name', 'CO2 emission']])

print('\nОстанні 5 країн по кількості CO2 викидів:')

print(last\_co2[['Country Name', 'CO2 emission']])

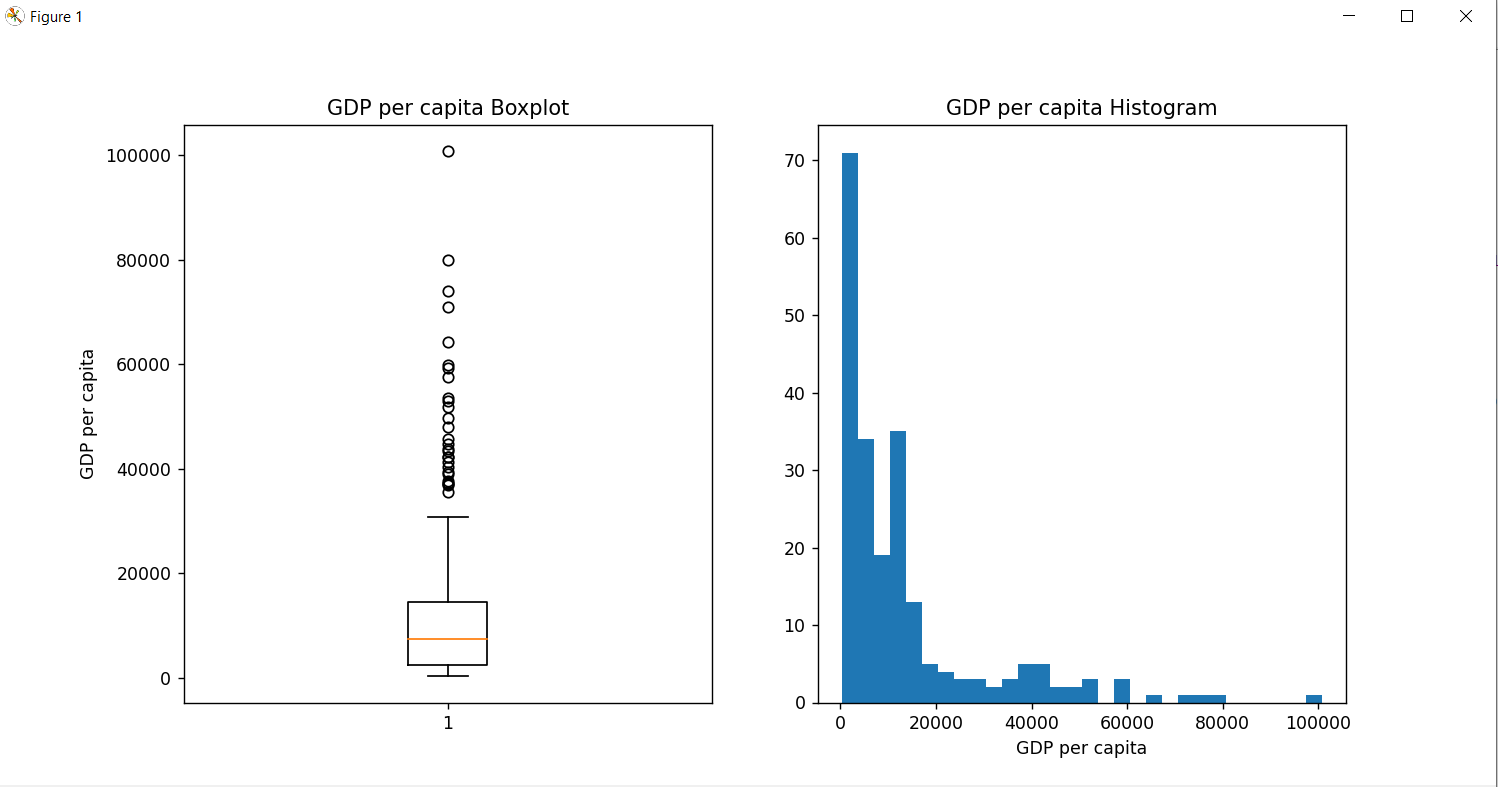


Рисунок 1.1. Діаграма розмаху та гістограма для значень ВВП на душу населення.

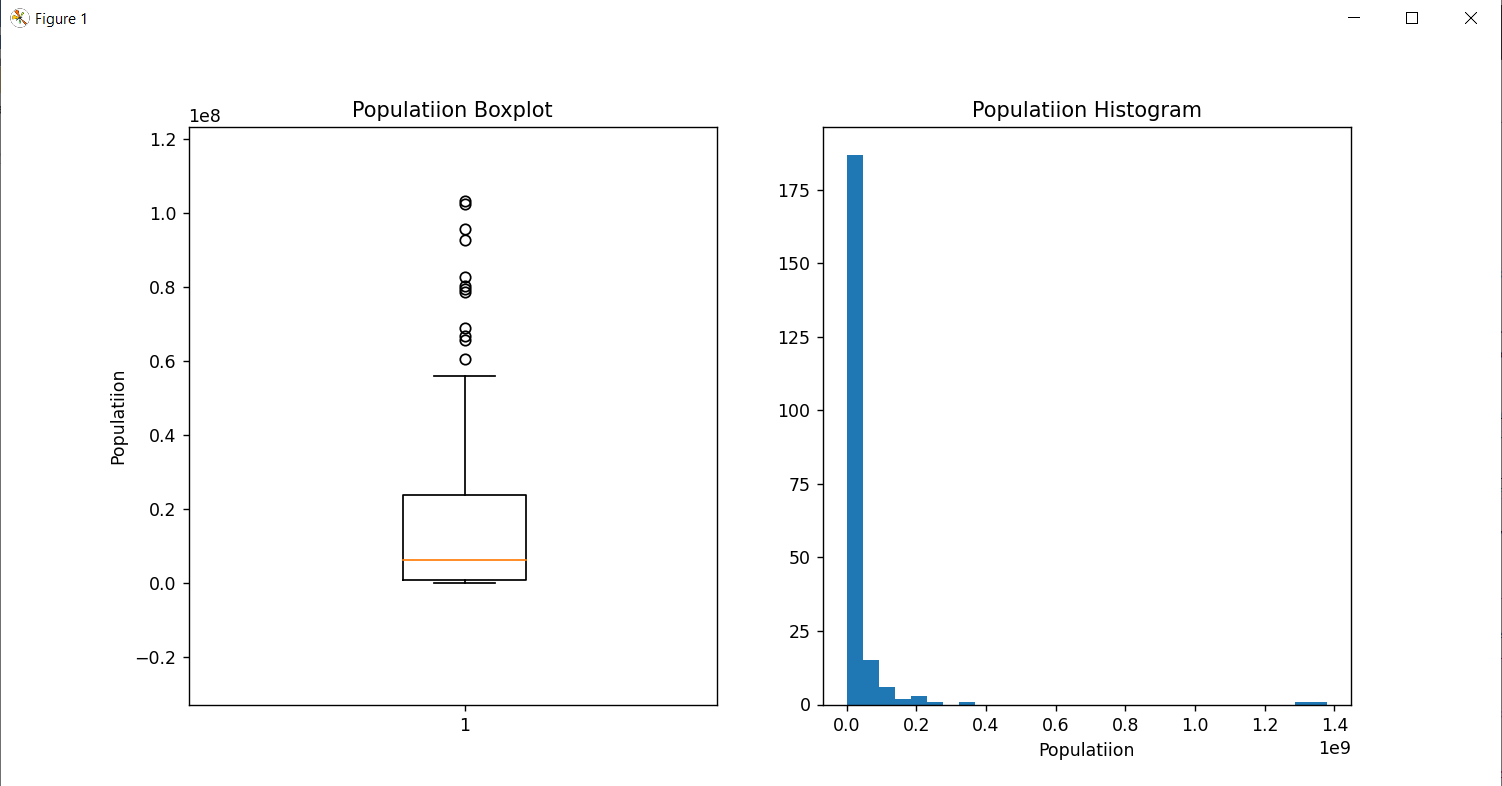


Рисунок 1.2. Діаграма розмаху та гістограма для значень кількості населення.

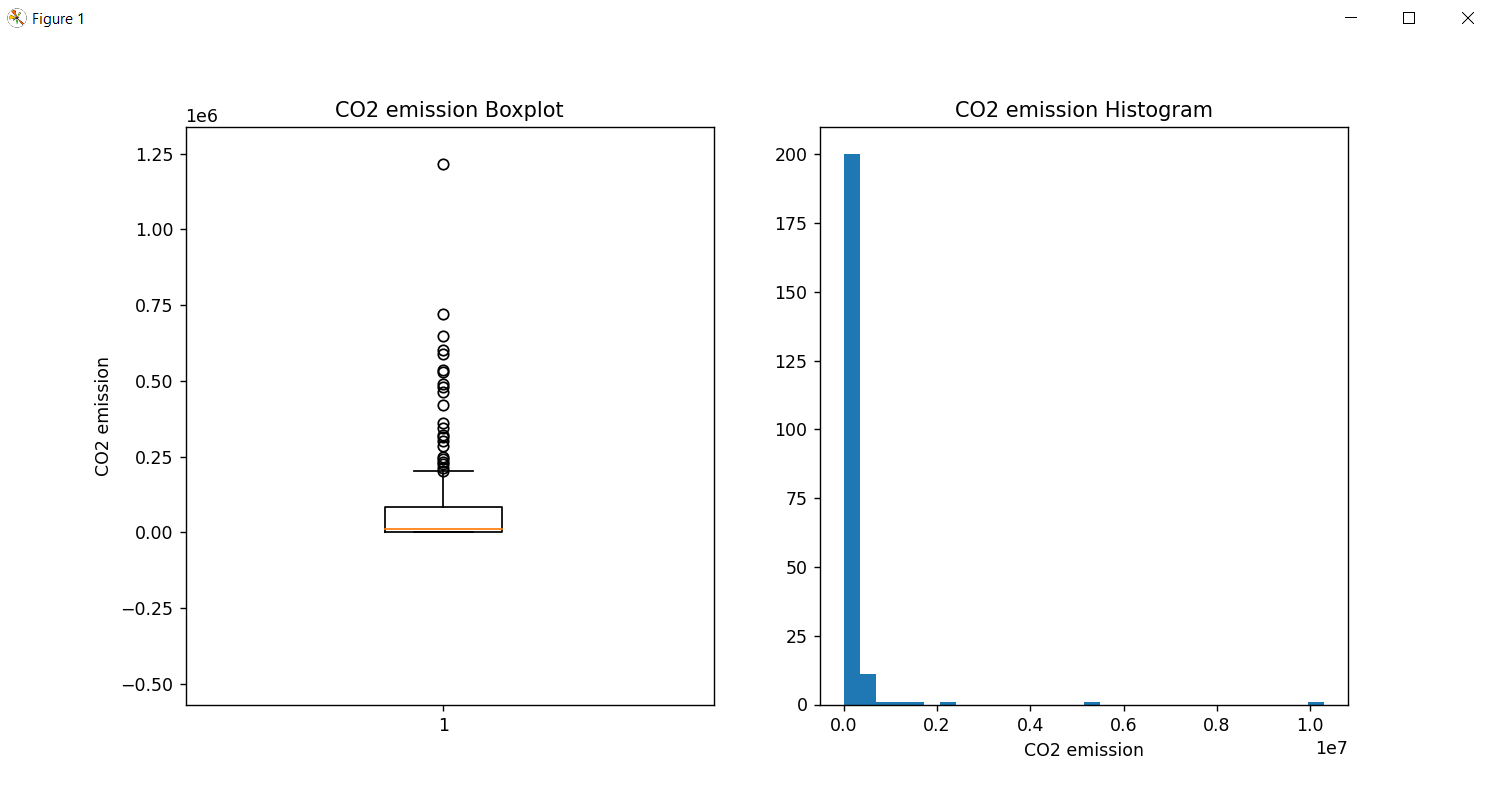


Рисунок 1.3. Діаграма розмаху та гістограма для значень викидів CO2.

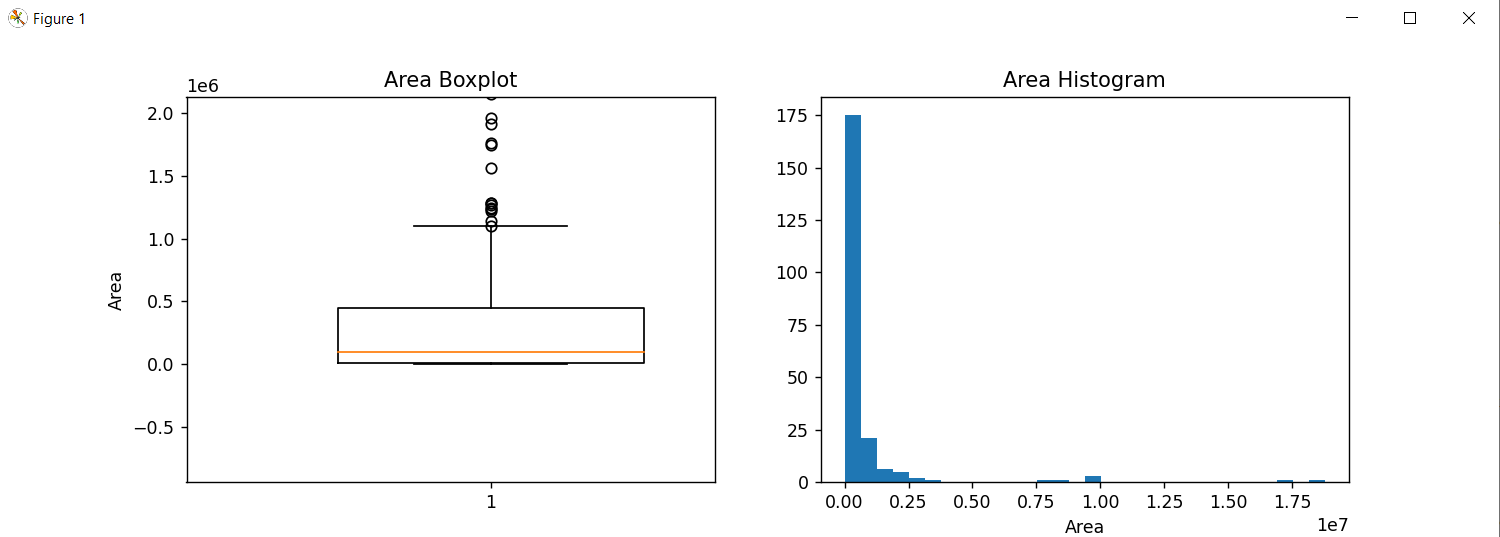


Рисунок 1.4. Діаграма розмаху та гістограма для значень площ країн.

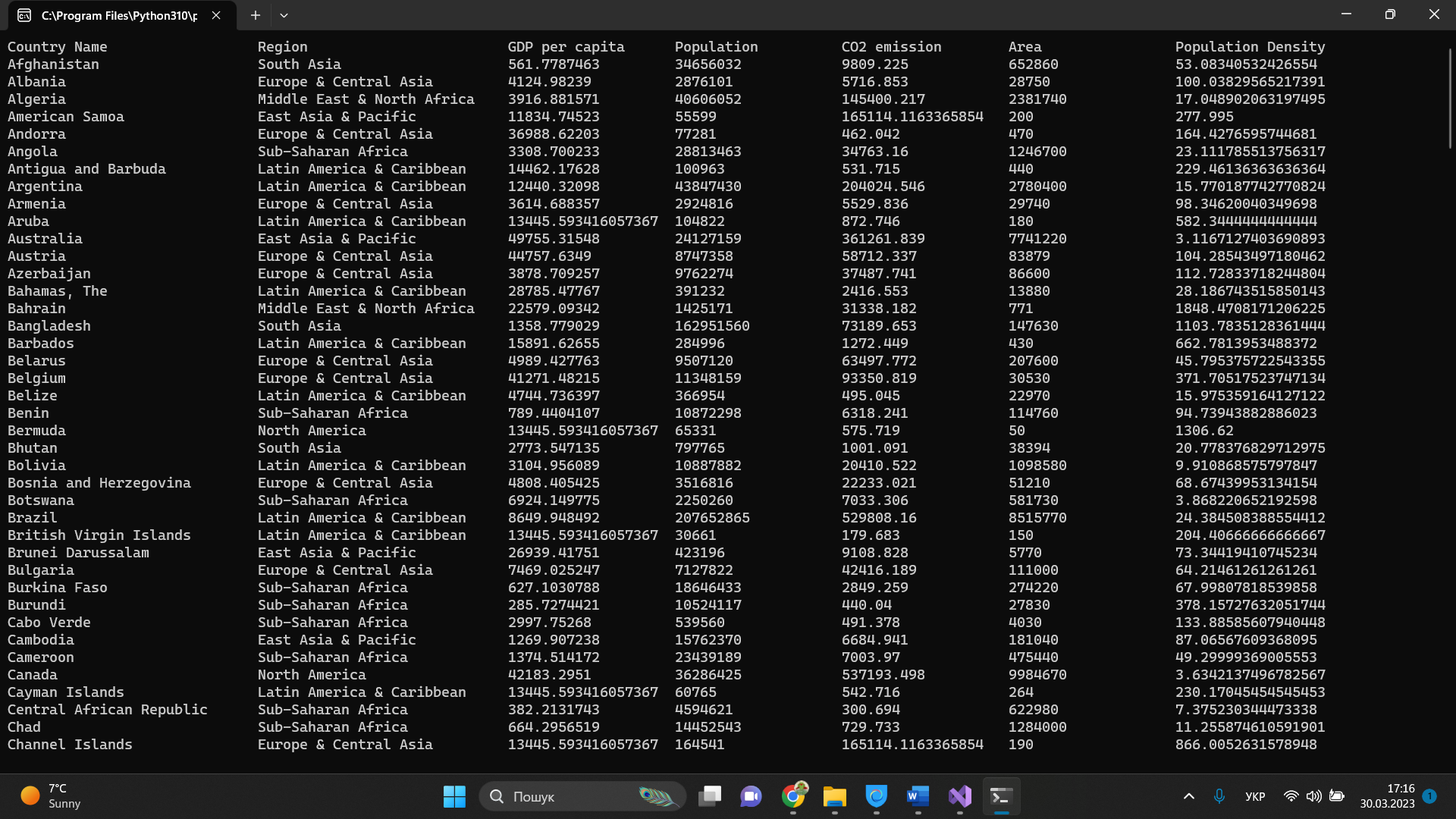
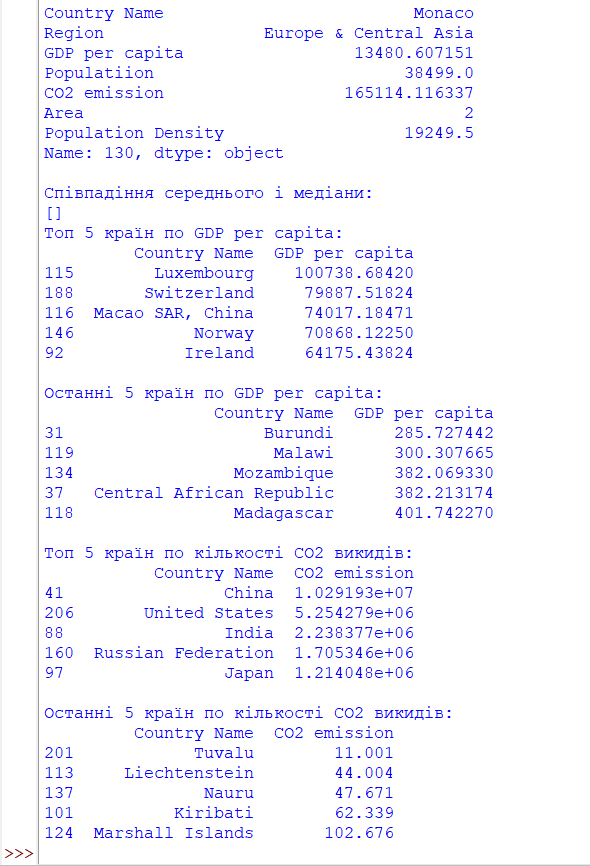


Рисунок 1.5. DataFrame з доданим стовпчиком із щільністю населення.

1. **Додаткове завдання:**

****

Рисунок 1.6. Відповіді на запитання до додаткового завдання.

Країна з найбільшим ВВП на душу населення - Люксембург з ВВП на душу населення 100738,6842.

Країна з найменшою площею - Монако з площею 2.

Регіон з найбільшою середньою площею країни - Північна Америка з середньою площею країни 6605410,0.

Країна з найбільшою щільністю населення - Монако з щільністю населення 19249,5.

Країна з найбільшою щільністю населення в Європі та Центральній Азії - Монако з щільністю населення 19249,5.

Немає жодного регіону, де середній ВВП на душу населення дорівнював би медіанному ВВП на душу населення.

Топ-5 країн з найбільшим ВВП на душу населення:

Country Name GDP per capita

115 Luxembourg 100738.68420

188 Switzerland 79887.51824

116 Macao SAR, China 74017.18471

146 Norway 70868.12250

92 Ireland 64175.43824

Топ-5 країн з найменшим ВВП на душу населення:

Country Name GDP per capita

31 Burundi 285.727442

119 Malawi 300.307665

134 Mozambique 382.069330

37 Central African Republic 382.213174

118 Madagascar 401.742270

Топ-5 країн з найбільшими викидами CO2 на людину:

41 China 1.029193e+07

206 United States 5.254279e+06

88 India 2.238377e+06

160 Russian Federation 1.705346e+06

97 Japan 1.214048e+06

Топ-5 країн з найменшим рівнем викидів CO2 на людину:

Country Name CO2 emission

201 Tuvalu 11.001

113 Liechtenstein 44.004

137 Nauru 47.671

101 Kiribati 62.339

124 Marshall Islands 102.676

**Висновок:**

Під час виконання лабораторної роботи я записав дані у data frame, здійснив видалення мінусів, заміну ком на крапки, встановлення типів даних у стовпцях. Також здійснив заміну пропущених значень на середні. Побудував діаграми розмаху та гістограми для ВВП на душу населення, кількості населення, викидів CO2 та площ країн.

Визначив країну з найбільшим ВВП на душу населення, країну з найменшою площею, регіон з найбільшою середньою площею країн, країну з найбільшою щільністю населення в світі, країну з найбільшою щільністю населення в Європі та Центральній Азії, вивів топ-5 країн з найбільшим ВВП на душу населення та 5 країн з найменшим ВВП на душу населення, топ-5 країн з найбільшими викидами CO2 на людину та 5 країн з найменшим рівнем викидів CO2 на людину.