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 інформаційнних системах»

на тему: «Описова статистика»

Викладач:
Олійник Ю.О.

Виконав:
студент 2 курсу
групи ІП-15 ФІОТ
Мєшков Андрій
Ігорович

Київ-2023

Практикум №3

Описова статистика

Мета роботи: ознайомитись з методикою первинної обробки статистичних даних; проаналізувати вплив способу представлення даних на їх інформативність.

Завдання:

Скачати потрібні дані.

Основне завдання

Скачати дані із файлу Data2.csv

1. Записати дані у data frame
2. Дослідити структуру даних
3. Виправити помилки в даних
4. Побудувати діаграми розмаху та гістограми
5. 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 CO2 на душу населення.

Хід роботи:

Основне завдання:

Інсталюємо pandas:

pip3 install pandas

Імпортуємо бібліотеку

import pandas as pd

1. Записати дані у data frame

```
data_path = 'Data2.csv'
df = read_dataset(data_path)
def read_dataset(path):
    df = pd.read_csv(path, sep=';', encoding='cp1252')
    return df
```

2. Дослідити структуру даних

```
def print_exploring(df):
    print('Data frame info:')
    df.info()
    print('\nFirst 5 rows:')
    print(df.head())
```

```
Data frame info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 217 entries, 0 to 216
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Country Name    217 non-null    object
 1   Region          217 non-null    object
 2   GDP per capita  190 non-null    object
 3   Populatiion     216 non-null    float64
 4   CO2 emission    205 non-null    object
 5   Area            217 non-null    object
dtypes: float64(1), object(5)
memory usage: 10.3+ KB

First 5 rows:
     Country Name                     Region GDP per capita  Populatiion CO2 emission     Area
0     Afghanistan                 South Asia    561,7787463   34656032.0     9809,225   652860
1         Albania       Europe & Central Asia     4124,98239    2876101.0     5716,853    28750
2         Algeria  Middle East & North Africa    3916,881571   40606052.0   145400,217  2381740
3  American Samoa         East Asia & Pacific    11834,74523      55599.0          NaN      200
4         Andorra       Europe & Central Asia    36988,62203      77281.0      462,042      470
```

Рисунок 1. Дослідження структури

3. Виправити помилки в даних

Виправляємо помилки

```
def remove_typo(df):
    df.rename(columns={"Populatiion": "Population"}, inplace=True)
    return df
```

Форматування типу даних

```
def clean_up(df):
    df['Area'] = df['Area'].str.replace(',', '.').astype(float)
    df["GDP per capita"] = df["GDP per capita"].str.replace(',', '.').astype(float)
    df["CO2 emission"] = df["CO2 emission"].str.replace(',', '.').astype(float)
    return df
```

Виправити негативні дані

```
def fix_negative(df):
    fix_gdp = df[df['GDP per capita'] < 0]
    area_gdp = df[df['Area'] < 0]
    fix_gdp['GDP per capita'] *= -1
    area_gdp['Area'] *= -1
    df[df['GDP per capita'] < 0] = fix_gdp
    df[df['Area'] < 0] = area_gdp
    return df
```

Виправити пропущені дані середніми

```
def fix_NaN(df):
    df = df.fillna(df.mean())
    return df
```

4. Побудувати діаграми розмаху та гістограми

Інсталюємо бібліотеку Matplotlib

```
pip3 install matplotlib
```

Імпортуємо бібліотеку

```
import matplotlib.pyplot as plt
```

Створюємо діаграми розмаху

```
def create_boxplot(df):
    fig, axs = plt.subplots(1, 4, figsize=(16, 4))
    fig.suptitle('Діаграми розмаху', fontsize=16)
    axs[0].set_title('GDP per capita')
    axs[0].boxplot(df['GDP per capita'])
    axs[1].set_title('Population')
    axs[1].boxplot(df['Population'])
    axs[2].set_title('CO2 emission')
    axs[2].boxplot(df['CO2 emission'])
    axs[3].set_title('Area')
    axs[3].boxplot(df['Area'])
```

Створюємо гістограми

```
def create_hist(df):
    fig, axs = plt.subplots(1, 4, figsize=(16, 4))

    fig.suptitle('Гістограми', fontsize=16)
    axs[0].set_title('GDP per capita')
    axs[0].hist(df['GDP per capita'])
    axs[1].set_title('Population')
    axs[1].hist(df['Population'])
    axs[2].set_title('CO2 emission')
    axs[2].hist(df['CO2 emission'])
    axs[3].set_title('Area')
    axs[3].hist(df['Area'])
```

Показуємо діаграми

```
plt.show()
```

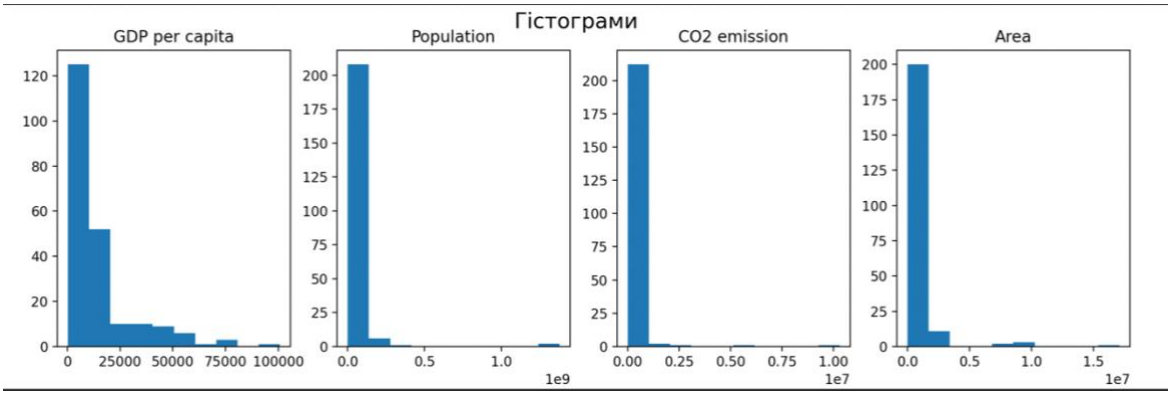


Рисунок 2. Гістограми

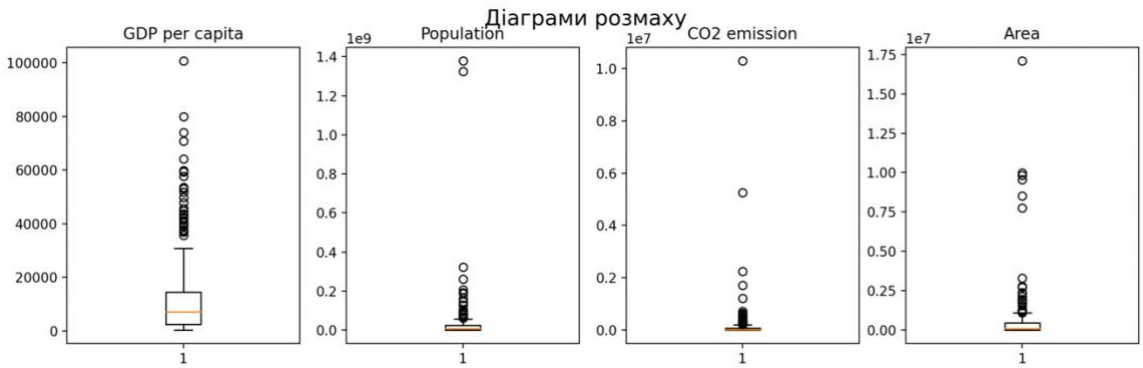
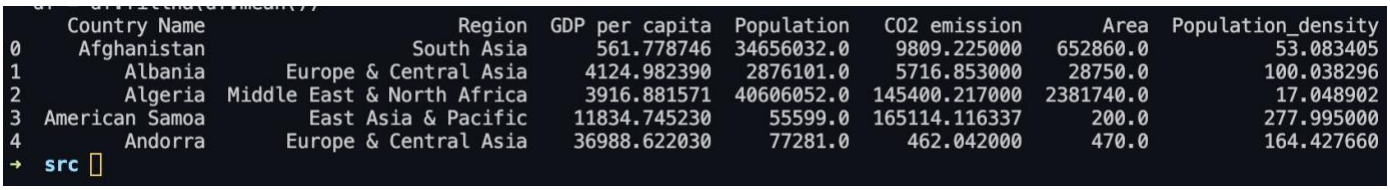


Рисунок 3. Діаграми розмаху

5. Додати стовпчик із щільністю населення

```
def add_population_density(df):
    df["Population_density"] = df["Population"] / df["Area"]
    print(df.head())
    return df
```



	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

→ src

Рисунок 4. Щільність населення

Додаткове завдання:

1. Чи є пропущені значення? Якщо є, замінити середніми

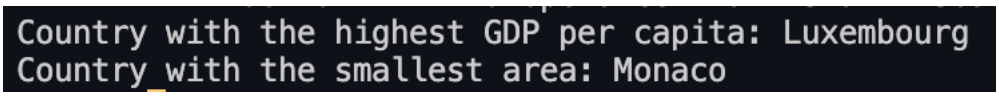
   Пропущені значення є, вони були замінені раніше за допомогою

   функції fix_NaN(df).

2. Яка країна має найбільший ВВП на людину (GDP per capita)? Яка

   має найменшу площу?

   highest_gdp_per_capita = df.loc[df['GDP per capita'].idxmax()]
   print("Country with the highest GDP per capita:", highest_gdp_per_capita['Country Name'])
   smallest_area = df.loc[df['Area'].idxmin()]
   print("Country with the smallest area:", smallest_area['Country Name'])

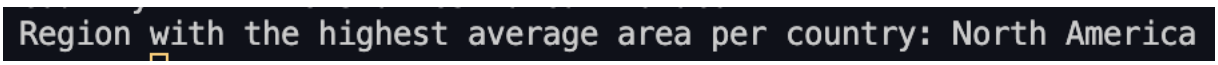


```
Country with the highest GDP per capita: Luxembourg
Country with the smallest area: Monaco
```

Рисунок 5. Відповідь на питання№2

3. В якому регіоні середня площа країни найбільша?

   mean_area_by_region = df.groupby('Region')['Area'].mean()
   region_with_highest_mean_area = mean_area_by_region.idxmax()
   print("Region with the highest average area per country:", region_with_highest_mean_area)



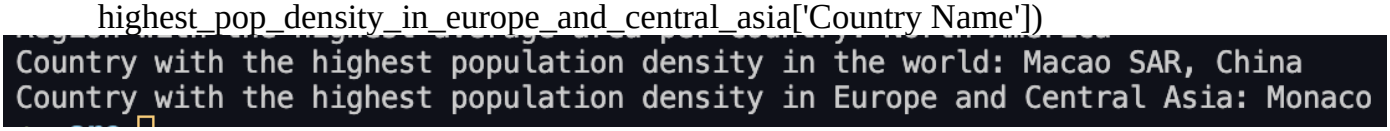
```
Region with the highest average area per country: North America
```

Рисунок 6. Відповідь на питання№3

4. Знайдіть країну з найбільшою щільністю населення у світі? У

   Європі та центральній Азії?

   highest_pop_density = df.loc[df['Population_density'].idxmax()]
   print("Country with the highest population density in the world:",
   highest_pop_density['Country Name'])
   europe_and_central_asia = df[df['Region'].isin(['Europe & Central Asia'])]
   highest_pop_density_in_europe_and_central_asia =
   europe_and_central_asia.loc[europe_and_central_asia['Population_density'].idxmax()]
   print("Country with the highest population density in Europe and Central Asia:",
   highest_pop_density_in_europe_and_central_asia['Country Name'])



```
Country with the highest population density in the world: Macao SAR, China
Country with the highest population density in Europe and Central Asia: Monaco
```

Рисунок 7. Відповідь на питання№4

5. Чи співпадає в якомусь регіоні середнє та медіана ВВП?

```
for region in df['Region'].unique():
    region_data = df[df['Region'] == region]
    mean_gdp = region_data['GDP per capita'].mean()
    median_gdp = region_data['GDP per capita'].median()
    print('\nMean GDP per capita in', region, ' - ' ,mean_gdp)
    print('Median GDP per capita in', region, ' - ' , median_gdp)
    if mean_gdp == median_gdp:
        print(f"In the {region} region, the mean and median GDP per capita are the same: {mean_gdp}")
```

```
Mean GDP per capita in South Asia  -  2795.2139349749996
Median GDP per capita in South Asia  -  1576.608412

Mean GDP per capita in Europe & Central Asia  -  22742.13551799658
Median GDP per capita in Europe & Central Asia  -  13445.593416057367

Mean GDP per capita in Middle East & North Africa  -  15459.162532674858
Median GDP per capita in Middle East & North Africa  -  13445.593416057367

Mean GDP per capita in East Asia & Pacific  -  15130.226548166813
Median GDP per capita in East Asia & Pacific  -  5910.620932

Mean GDP per capita in Sub-Saharan Africa  -  2878.6655206160854
Median GDP per capita in Sub-Saharan Africa  -  1034.3903605

Mean GDP per capita in Latin America & Caribbean  -  10485.343135639849
Median GDP per capita in Latin America & Caribbean  -  10833.201075

Mean GDP per capita in North America  -  37755.682535352455
Median GDP per capita in North America  -  42183.2951
```

Рисунок 8. Відповідь на питання№5. Середнє і медіана ВВП ніде не співпадає.

6. Вивести топ 5 країн та 5 останніх країн по ВВП та кількості CO2 на душу населення.

```
df_sorted_gdp = df.sort_values(by='GDP per capita',  ascending=False)
top5_gdp = df_sorted_gdp.head(5)
bottom5_gdp = df_sorted_gdp.tail(5)
print("\nTop 5 countries by GDP per capita:")
print(top5_gdp)
print("\nBottom 5 countries by GDP per capita:")
print(bottom5_gdp)

pd.set_option("display.max_columns", None)
df['CO2 emission per citizen'] = df['CO2 emission'] / df['Population']
df_sorted_co2 = df.sort_values(by='CO2 emission per citizen',  ascending=False)
top5_co2 = df_sorted_co2.head(5)
bottom5_co2 = df_sorted_co2.tail(5)
print("\n\nTop 5 countries by CO2 emissions per citizen:")
```

```
print(top5_co2)
print("\nBottom 5 countries by CO2 emissions per citizen:")
print(bottom5_co2)
```

```
Top 5 countries by GDP per capita:
        Country Name               Region  GDP per capita  Population  CO2 emission      Area  Population_density
115       Luxembourg  Europe & Central Asia    100738.68420    582972.0      9658.878    2590.0          225.085714
188      Switzerland  Europe & Central Asia     79887.51824   8372098.0     35305.876   41290.0          202.763333
116  Macao SAR, China    East Asia & Pacific     74017.18471    612167.0      1283.450      30.3        20203.531353
146           Norway  Europe & Central Asia     70868.12250   5232929.0     47626.996  385178.0           13.585742
92           Ireland  Europe & Central Asia     64175.43824   4773095.0     34066.430   70280.0           67.915410

Bottom 5 countries by GDP per capita:
               Country Name             Region  GDP per capita  Population  CO2 emission      Area  Population_density
118              Madagascar  Sub-Saharan Africa      401.742270  24894551.0      3076.613  587295.0           42.388495
37   Central African Republic  Sub-Saharan Africa      382.213174   4594621.0       300.694  622980.0            7.375230
134               Mozambique  Sub-Saharan Africa      382.069330  28829476.0      8426.766  799380.0           36.064795
119                   Malawi  Sub-Saharan Africa      300.307665  18091575.0      1276.116  118480.0          152.697291
31                   Burundi  Sub-Saharan Africa      285.727442  10524117.0       440.040   27830.0          378.157276


Top 5 countries by CO2 emissions per citizen:
               Country Name                     Region  GDP per capita  \
182  St. Martin (French part)  Latin America & Caribbean    13445.593416
163               San Marino      Europe & Central Asia    47908.561410
130                   Monaco      Europe & Central Asia    13445.593416
145  Northern Mariana Islands        East Asia & Pacific    22572.378820
3              American Samoa        East Asia & Pacific    11834.745230

     Population   CO2 emission   Area  Population_density  \
182     31949.0  165114.116337   54.4          587.297794
163     33203.0  165114.116337   60.0          553.383333
130     38499.0  165114.116337    2.0        19249.500000
145     55023.0  165114.116337  460.0          119.615217
3       55599.0  165114.116337  200.0          277.995000

     CO2 emission per citizen
182                  5.168053
163                  4.972867
130                  4.288790
145                  3.000820
3                    2.969732

Bottom 5 countries by CO2 emissions per citizen:
        Country Name             Region  GDP per capita    Population  \
44   Congo, Dem. Rep.  Sub-Saharan Africa      405.542501  7.873615e+07
38               Chad  Sub-Saharan Africa      664.295652  1.445254e+07
175           Somalia  Sub-Saharan Africa      434.208810  1.431800e+07
31            Burundi  Sub-Saharan Africa      285.727442  1.052412e+07
61            Eritrea  Sub-Saharan Africa    13445.593416  3.432256e+07

     CO2 emission       Area  Population_density  CO2 emission per citizen
44       4671.758  2344860.0           33.578189                  0.000059
38        729.733  1284000.0           11.255875                  0.000050
175       608.722   637660.0           22.453966                  0.000043
31        440.040    27830.0          378.157276                  0.000042
61        696.730   117600.0          291.858502                  0.000020
```

Рисунок 9. Відповідь на питання№6.

Висновок

За отриманими даними можна зробити висновок, що країни західної Європи мають високий ВВП, тоді як країни Африки мають найнижчі показники. Значення викидів вуглекислого газу на одну особу в країнах Африки є дуже малим. Проте, щодо країн з найбільшим значенням цього показника, не можна бути впевненим, оскільки серед топ-5 країн, які були отримані, всі мали пропущене значення в початковому датасеті. Це призвело до того, що їм було присвоєно середнє значення даного показника, яке може бути далеким від істинного. Крім того, всі ці топ-5 країн мають порівняно невелику кількість населення, що спричиняє великий показник викидів на одну особу. Тому необхідно здійснити додатковий збір даних для отримання достовірних результатів.