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

Звіт

з комп'ютерного практикуму № 3 з дисципліни

«Аналіз даних в інформаційних системах»

на тему: «Описова статистика»

Виконав студент ІП-13, Дойчев Костянтин Миколайович
(шифр, прізвище, ім'я, по батькові)

Перевірив Олійник Юрій Олександрович
(прізвище, ім'я, по батькові)

Київ 2023

Комп’ютерний практикум 3

Тема – Описова статистика.

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

Завдання

Основне:

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

Основне завдання
DataFrame та його структура

За допомогою Python бібліотеки Pandas прочитаємо та завантажимо дані з даного Data2.csv файлу в dataframe. Проінспектуємо структуру наших даних.

```
In 33
import pandas as pd
import matplotlib.pyplot as plt
Executed at 2023.06.25 19:26:42 in 24ms

In 34
# read data
df = pd.read_csv('data/Data2.csv', sep=';', decimal=',', encoding='cp1252')
Executed at 2023.06.25 19:26:42 in 20ms

In 35
# Print rows
head_rows = 5
tail_rows = 6

h = df.head(5)
t = df.tail(6)

print('Head')
print(h)
print('Tail')
print(t)
Executed at 2023.06.25 19:26:42 in 40ms
```

```
Head
     Country Name                    Region  GDP per capita  Populatiion  \
0     Afghanistan                South Asia      561.778746   34656032.0
1         Albania       Europe & Central Asia     4124.982390    2876101.0
2         Algeria  Middle East & North Africa     3916.881571   40606052.0
3  American Samoa         East Asia & Pacific    11834.745230      55599.0
4         Andorra       Europe & Central Asia    36988.622030      77281.0

   CO2 emission       Area
0      9809.225   652860.0
1      5716.853    28750.0
2    145400.217  2381740.0
3           NaN      200.0
4       462.042      470.0
Tail
              Country Name                    Region  GDP per capita  \
211               Vietnam        East Asia & Pacific     2170.648054
212  Virgin Islands (U.S.)   Latin America & Caribbean             NaN
213     West Bank and Gaza  Middle East & North Africa     2943.404534
214           Yemen, Rep.  Middle East & North Africa      990.334774
215                Zambia         Sub-Saharan Africa     1269.573537
216              Zimbabwe         Sub-Saharan Africa     1029.076649

     Populatiion  CO2 emission      Area
211   92701100.0    166910.839  330967.0
212     102951.0           NaN     350.0
213    4551566.0           NaN    6020.0
214   27584213.0     22698.730  527970.0
215   16591390.0      4503.076  752610.0
```

На даному рисунку можна помітити загальну інформацію про датафрейм:

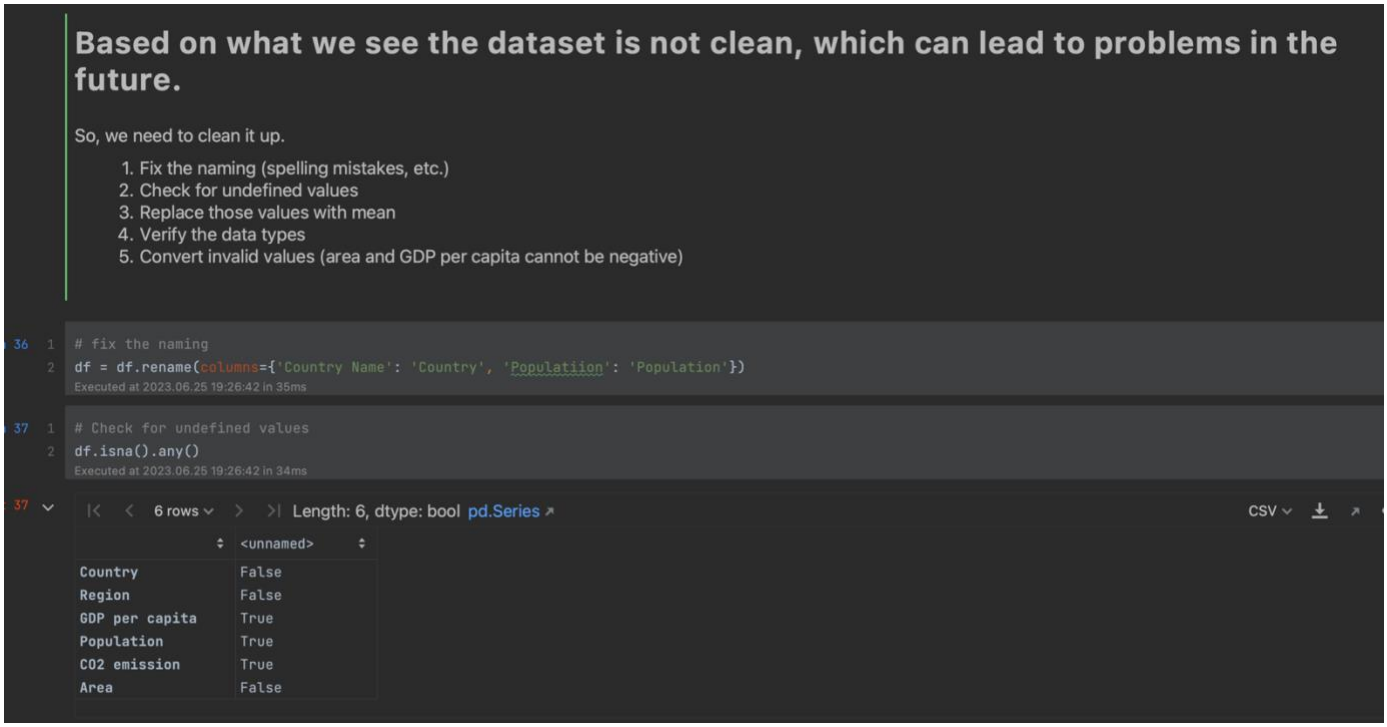
кількість рядків та колонок, назви всіх колонок, кількість записів в кожній з них, тип даних колонки та використання пам’яті.

Висновки:

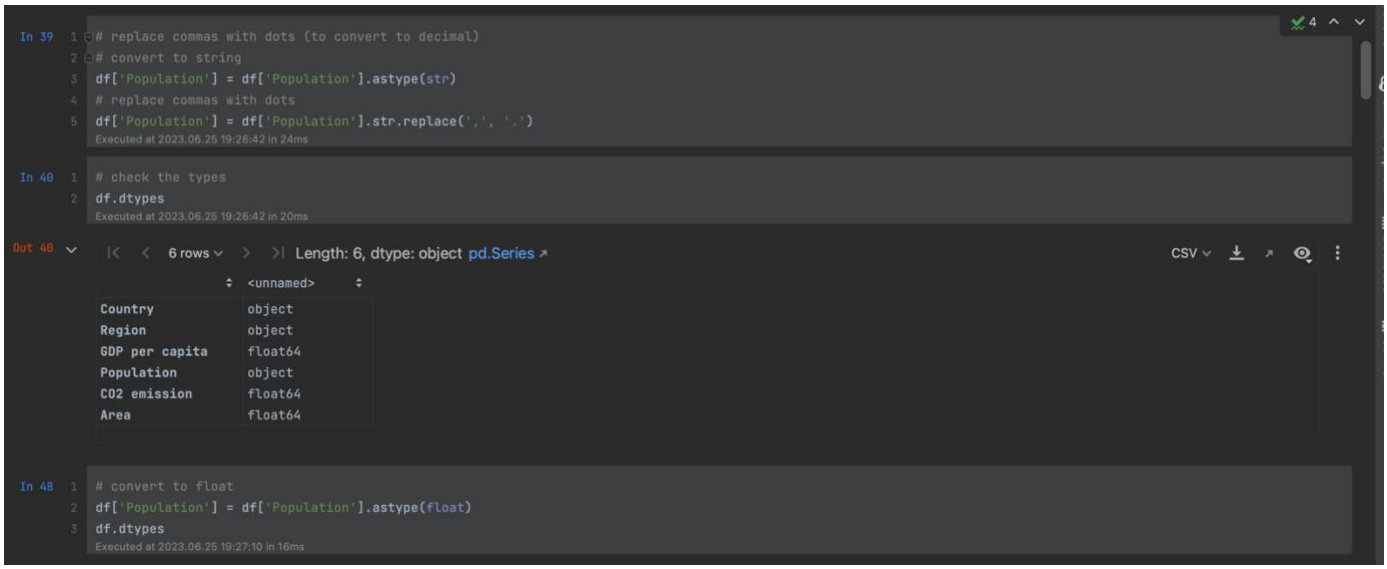
1. Помилки в правописі (spelling) колонки Population. Populatiion -> Population
2. Пусті дані (NaN)
3. Від’ємні значення де вони не мають бути. Площа не може бути від’ємною

Виправлення помилок

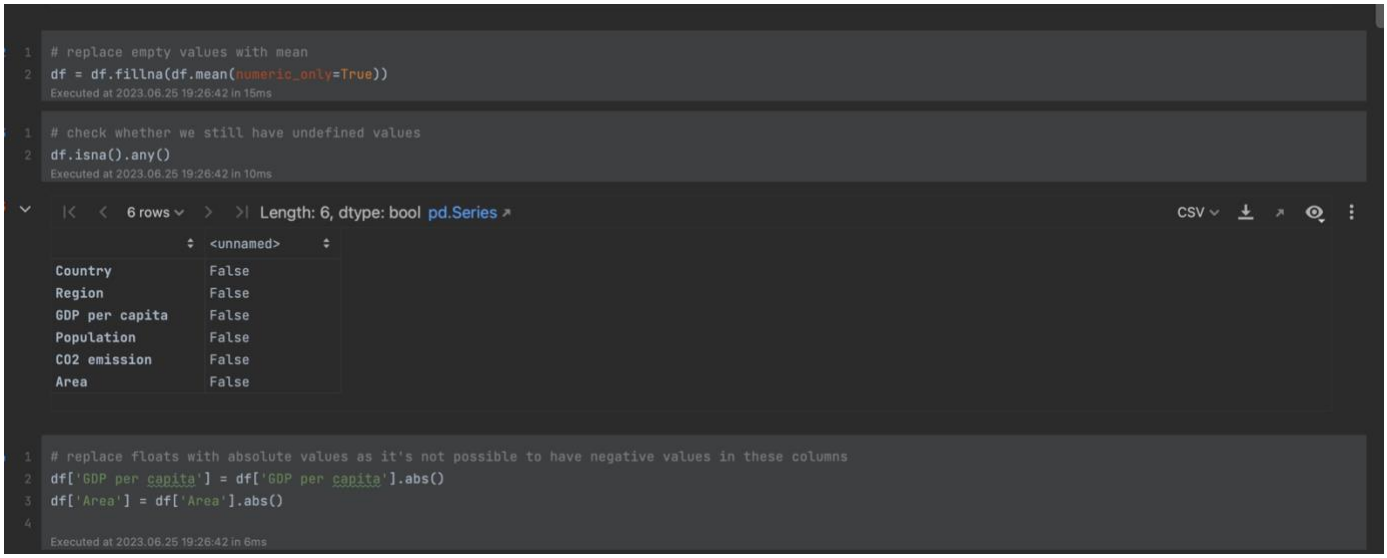
Змінимо назву колонки та перевіримо наявність пустих значень



Замінимо європейський стиль написання десяткових чисел (10,5) на американський (10.5)



Замінимо пусті значення середнім арифметичним та приведемо числа до додатніх



Візуалізація

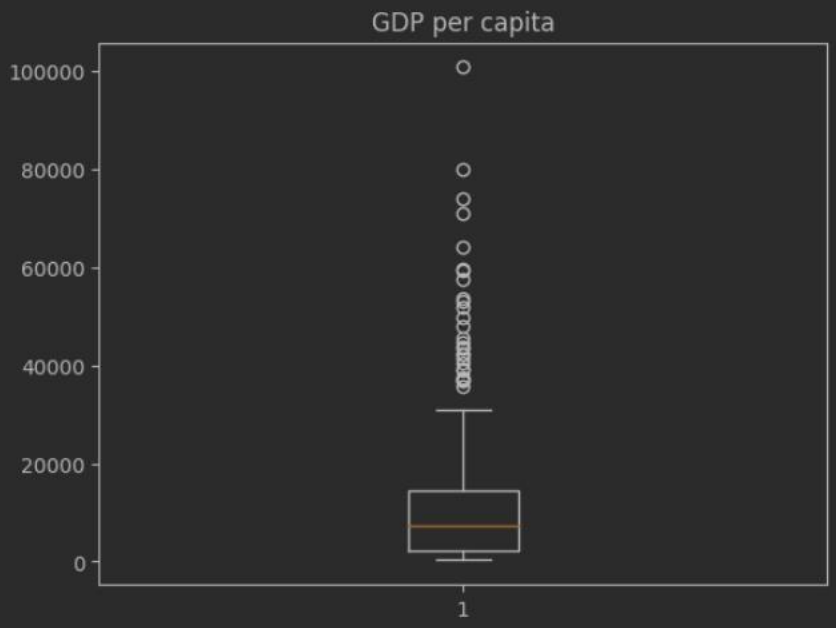
Виведемо діаграми розмаху та гістограми для кожного стовпця з чисельними даними.

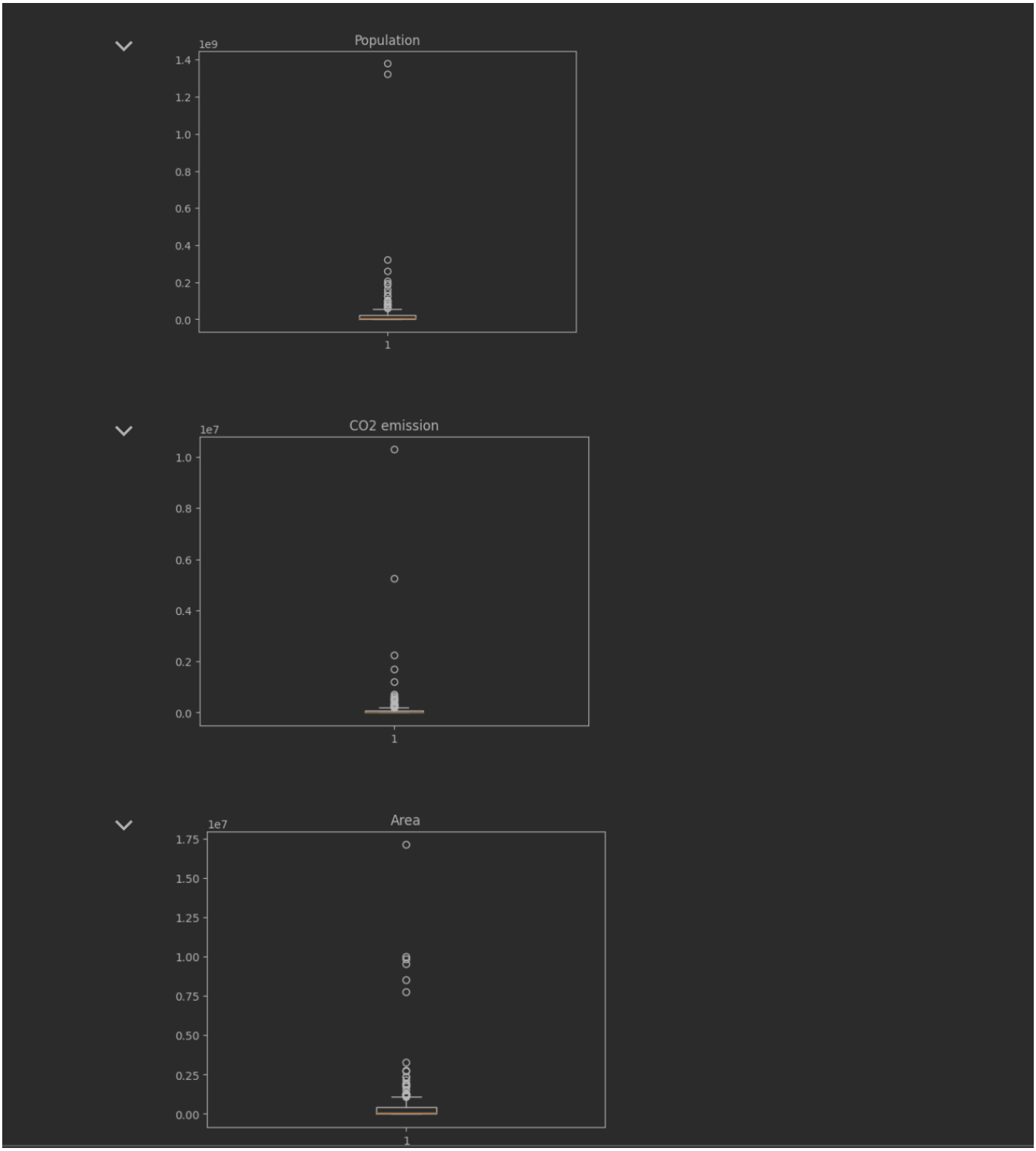
Data visualization

In 52

```
# boxplot
for column in df.columns:
    if df[column].dtype == float:
        plt.figure()
        plt.title(column)
        plt.boxplot(df[column])
```

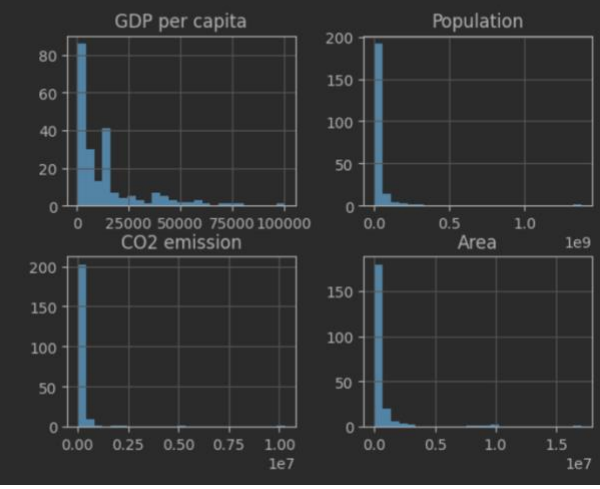
Executed at 2023.06.25 19:38:01 in 317ms




Population
1e9
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1
CO2 emission
1e7
1.0
0.8
0.6
0.4
0.2
0.0
1
Area
1e7
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1

```
In 46
# histogram

df.hist(bins=25)
plt.show()
Executed at 2023.06.25 19:26:43 in 209ms
```



Додавання стовпчику із щільністю населення

Додаємо стовпчик із щільністю населення кожної країни, який є просто представленням кількості населення поділеного на площу країни.

```
In 50
# Create a population density column

df['Population Density'] = df['Population'] / df['Area']
Executed at 2023.06.25 19:27:46 in 3ms

In 51
df
Executed at 2023.06.25 19:27:50 in 18ms

Out 51
1-10 >| 217 rows × 7 columns pd.Dataframe ↗ CSV
```

	Country	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
5	Angola	Sub-Saharan Africa	3308.700233	28813463.0	34763.160000	1246700.0	23.111786
6	Antigua and Barbuda	Latin America & Caribbean	14462.176280	100963.0	531.715000	440.0	229.461364
7	Argentina	Latin America & Caribbean	12440.320980	43847430.0	204024.546000	2780400.0	15.770188
8	Armenia	Europe & Central Asia	3614.688357	2924816.0	5529.836000	29740.0	98.346200
9	Aruba	Latin America & Caribbean	13374.833168	104822.0	872.746000	180.0	582.344444

Висновок

У цьому комп’ютерному практикуму було вивчено можливості Python, а саме Pandas у роботі з даними. Вхідні дані було записано в data frame, дані були не зовсім «чистими», тому було необхідно змінити назви, заповнити пусті значення, виправити помилки. На діаграмах розмаху було помічено великий розмах між даними. Наприклад, на діаграмі населення є дві країни з кількістю населення значно більшою за всі інші, так само і з викидами CO2, дані з ВВП на душу населення є найбільш кучними.