Міністерство освіти і науки України

Національний технічний університет України

“Київський політехнічний інститут”

Кафедра АСОІУ

ЗВІТ

про виконання лабораторної роботи №2

з дисципліни

“ Аналіз даних в інформаційно-управляючих системах”

# СТВОРЕННЯ ВІ РІШЕННЯ

Виконав Студент

2 курсу групи ІП-11

Панченко Сергій

Київ 2023  
ОПИСОВА СТАТИСТИКА

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

**Перелік корисних джерел**

RStudio

* [Довідкові матеріали для початку роботи з R](https://drive.google.com/open?id=1_Dli7wxMFOuFVn7auz-P_p2fN2W0wib0)
* [Data visualization with ggplot2: cheatsheet](https://drive.google.com/open?id=1qF16Bmc5to7iq2IKu6lxUNwr7CZ3zZMD)
* [Гнатюк В. Вступ до R на прикладах](https://drive.google.com/open?id=1R1dWeO7i1VpiJZTkt2_orgyFQqe-LESY)
* [Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R](https://drive.google.com/open?id=104-kdkyB-g3s41L-AQsdZpbLJaSaBu_J)
* [​Data Mining for Business Analytics](https://drive.google.com/open?id=1d26sNsT5811wVPeBb6NzuUyt65i4accC)

Python

* [Довідкові матеріали для початку роботи з Python](https://drive.google.com/file/d/1XYTJEJYXkJUGf7NjGe-6FNrRXnntbhYj/view?usp=sharing)
* [Учебник по Python](https://github.com/rossant/ipython-minibook)
* [Подборка блокнотов по командам IPython](https://github.com/rossant/ipython-minibook)
* [Інтерактивний міні-гайд по візуалізації даних на Python](https://dataviz.dylancastillo.co/)
* [Элбон Машинное обучение с использованием Python 2019](https://drive.google.com/file/d/1bs0sZ97O7qawT7kXv5BouxZzQOT3QRb6/view?usp=sharing)
* [Python NumPy beginners](https://drive.google.com/file/d/1RPZ5NarLVsaZuwO0hdjQw5d4gpBX3w_r/view?usp=share_link)
* [Campesato Pandas Basics](https://drive.google.com/file/d/103WhlGlZ-pO6noYM63QRbx7KbKSTuXc4/view?usp=share_link)
* [Пособие по Matplotlib](http://nbviewer.jupyter.org/github/jrjohansson/scientific-python-lectures/blob/master/Lecture-4-Matplotlib.ipynb)
* [Первичный анализ данных с Pandas](https://habrahabr.ru/company/ods/blog/322626/)
* [Python Data Science Handbook](https://drive.google.com/open?id=1NzdTHgU_6HNHhDfx8yUs0xDlmbZVFNp1)

**Запитання для самоперевірки**

1. Чим відрізняються генеральна та вибіркова сукупності?
2. Які бувають способи відбору даних?
3. Як побудувати статистичний розподіл вибірки?
4. Що відносять до числових характеристик вибірки?
5. Як визначається мода та медіана?
6. Які бувають способи графічного зображення статистичних розподілів?
7. Що таке емпірична функція розподілу та які її властивості?

**ЗАВДАННЯ**

[Скачати потрібні дані](https://drive.google.com/drive/folders/1kYY4pxkeLXoszCX8icpFUrhFCnwgAyhz?usp=sharing).

**Завдання для самоперевірки**

Ознайомитися з:

Підключення бібліотек

Можливі формати вхідних даних

Перетворення форматів

Отримання інформації про структуру даних

Перетворення датафреймів

Додавання, видалення ознак (стовпчиків)

Перевірка та перетворення типу даних

Групування даних

Сортування

Фільтрація

Виділення підмножини

Об'єднання кількох датафреймів в один

Обчислення числових характеристик

Застосування функцій до елемента, стовпчика, рядка

[Приклад R](https://drive.google.com/file/d/19jYkMLyZiJl0NQGFk6WFgawgc3nguxEC/view?usp=sharing)

Поглиблено про графічне представлення інформації

кругові діаграми

діаграми розсіювання

діаграми розмаху

[Приклад R](https://drive.google.com/file/d/1MbD5jM2vX1ze5e4GkKCMeNlk7h8whETo/view?usp=sharing)

Скачати дані із файлу Data1.csv

1. дослідити їх структуру
2. вивести перші 5 рядків
3. вивести останні 6 рядків
4. видалити стовпчик з абревіатурами
5. додати стовпчик з повним GDP, пропуски замінити нулями
6. вивести все summary
7. побудувати діаграму розмаху для GDP per capita
8. побудувати графік залежності High-technology exports від GDP

Приклад виконання [RStudio](https://drive.google.com/file/d/1uW7GNU7YgpQOCXsjqUYI0s__sZoVDTkD/view?usp=sharing), [Python](https://drive.google.com/file/d/19yd3Cv0UeDzkbDXDfFyc8cD4--r82qGd/view?usp=sharing)

**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** Приклад виконання [RStudio](https://drive.google.com/file/d/1CfQgC1aUJiBhX1T0mFyx3nje4l5yi5JO/view?usp=sharing), [Python](https://drive.google.com/file/d/1Zel3jIGZ6TM4xM-7vwMf_dzeC7cxZmbV/view?usp=sharing)

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. З яким населенням найчастіше зустрічаються країни у світі? У Європі?
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.

Завантажити вхідні дані в середовище, що буде використовуватись для їх обробки (розрахунки можна проводити з використанням R-Studio, Python, в пакетах STATISTICA, MathCad, MathLab, Excel тощо). Виконати необхідні розрахунки та оформити звіт.

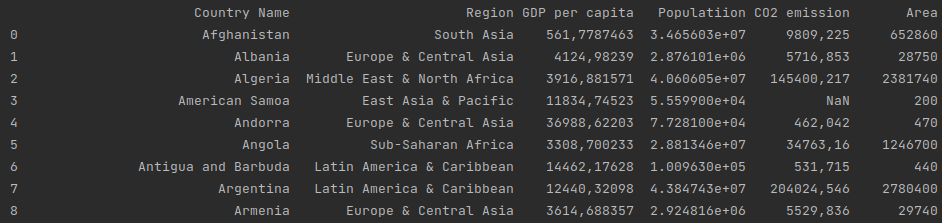
В звіт включити основне та додаткове завдання (самі завдання, числові та графічні відповіді на них, а також висновки по виконаному дослідженню; файл з кодом окремо).

Відповіді на теоретичні питання включати в звіт не потрібно - це просто підказка для вас, з чим треба розібратися до того, як виконувати роботу.

Загальний висновок по роботі у вигляді "розібрався, навчився" не потрібен, залишайте лише ваші висновки по проведеному дослідженню (наприклад, висновок щодо прийняття або відхилення гіпотези, яку ви перевіряли).

**Код:**

**1) Записати дані у дата фрейм:**



**2) Дослідити структуру даних:**

Бачимо, що датафрейм має NAN-значення, пусті клітинки. Також варто помітити, що «Populatiion» написана з помилкою, треба виправити на «Population». Також серед значень є відʼємні.

**3) Виправити помилки в даних:**

**Додаткове 1:**

Замінимо пропущені значення середніми

for column\_name in dataset.columns[2:]:  
 replace\_comma\_with\_dots(dataset, column\_name)  
 convert\_column\_to\_float(dataset, column\_name)  
 replace\_nan\_with\_mean(dataset, column\_name)  
 convert\_float\_with\_positive(dataset, column\_name)

def replace\_comma\_with\_dots(dataset: pd.DataFrame, column\_name: str) -> None:  
 dataset[column\_name] = dataset[column\_name].astype(str)  
 dataset[column\_name] = dataset[column\_name].str.replace(',', '.')  
  
  
def convert\_column\_to\_float(dataset: pd.DataFrame, column\_name: str) -> None:  
 dataset[column\_name] = dataset[column\_name].astype(float)  
  
  
def replace\_nan\_with\_mean(dataset: pd.DataFrame, column\_name: str):  
 mean\_value = dataset[column\_name].mean()  
 dataset[column\_name].fillna(value=mean\_value, inplace=True)  
  
def convert\_float\_with\_positive(dataset: pd.DataFrame, column\_name: str):  
 dataset[column\_name] = dataset[column\_name].abs()

**Додаткове 2:**

# 2 Яка країна має найбільший ВВП на людину (GDP per capita)? Яка має найменшу площу?  
df\_max\_gdp = dataset.nlargest(1, ['GDP per capita'])  
max\_gdp = df\_max\_gdp['GDP per capita'].values[0]  
country\_max\_gdp = df\_max\_gdp['Country Name'].values[0]  
print(f'{bcolors.HEADER}Max GDP:\n{bcolors.ENDC}'  
 f'\t{bcolors.OKBLUE}Country: '  
 f'{bcolors.OKGREEN}{country\_max\_gdp}\n'  
 f'\t{bcolors.OKBLUE}GDP: {bcolors.OKGREEN}{max\_gdp}')

Max GDP:

Country: Luxembourg

GDP: 100738.6842

**Додаткове 3:**

# 3 В якому регіоні середня площа країни найбільша?df\_group\_by\_region = dataset.groupby('Region')df\_region\_area\_sum = df\_group\_by\_region.sum(numeric\_only=True)[['Area']]df\_region\_count = df\_group\_by\_region.count()[['Area']]df\_average\_area = df\_region\_area\_sum / df\_region\_countdf\_max\_average\_area = df\_average\_area.nlargest(1, ['Area'])max\_average\_area\_region = df\_max\_average\_area.axes[0][0]max\_average\_area = df\_max\_average\_area['Area'].values[0]print(f'{bcolors.HEADER}Max average area:{bcolors.ENDC}\n' f'\t{bcolors.OKBLUE}Region: ' f'{bcolors.OKGREEN}{max\_average\_area\_region}\n' f'\t{bcolors.OKBLUE}Value: ' f'{bcolors.OKGREEN}{max\_average\_area}')

Min Area:

Country: Monaco

Area: 2.0

Max average area:

Region: North America

Value: 6605410.0

**5. Додати стовпчик із щільністю населення та додаткове 4:**

# 4 знайдіть країну з найбільшою щільністю населення, у світі та в Європі  
dataset['Population'] = dataset['Populatiion']  
dataset.drop(['Populatiion'], axis=1)  
dataset['Density'] = dataset['Population'] / dataset['Area']  
df\_max\_density\_world = dataset.nlargest(1, ['Density'])  
max\_density\_world = df\_max\_density\_world['Density'].values[0]  
max\_density\_world\_country = df\_max\_density\_world['Country Name'].values[0]  
df\_europe = dataset[dataset['Region'] == 'Europe & Central Asia']  
df\_europe\_max\_density = df\_europe.nlargest(1, ['Density'])  
max\_density\_europe\_country = df\_europe\_max\_density['Country Name'].values[0]  
max\_density\_europe = df\_europe\_max\_density['Density'].values[0]  
print(f'{bcolors.HEADER}Max density:{bcolors.ENDC}\n'  
 f'\t{bcolors.OKBLUE}World: '  
 f'{bcolors.OKGREEN}{max\_density\_world\_country}\n'  
 f'\t{bcolors.OKBLUE}World Value: '  
 f'{bcolors.OKGREEN}{max\_density\_world}\n'  
 f'\t{bcolors.OKBLUE}Europe Country: '  
 f'{bcolors.OKGREEN}{max\_density\_europe\_country}\n'  
 f'\t{bcolors.OKBLUE}Europe Value: '  
 f'{bcolors.OKGREEN}{max\_density\_europe}')

Max density:

World: Macao SAR, China

World Value: 20203.531353135313

Europe Country: Monaco

Europe Value: 19249.5

**Додаткове 5:**

# 5 Чи співпадає в якомусь регіоні середнє та медіана ВВП?df\_region\_gdp\_mean = pd.DataFrame()df\_region\_gdp\_average = pd.DataFrame()df\_region\_gdp\_mean['Mean'] = df\_group\_by\_region.mean(numeric\_only=True)[['GDP per capita']]df\_region\_gdp\_average['Average'] = df\_group\_by\_region.median(numeric\_only=True)[['GDP per capita']]df\_region\_mean\_average = pd.concat([df\_region\_gdp\_mean, df\_region\_gdp\_average], axis=1)df\_region\_mean\_average['Difference'] = df\_region\_mean\_average['Mean'] - df\_region\_mean\_average['Average']df\_region\_mean\_average['Difference'] = df\_region\_mean\_average['Difference'].abs()df\_smallest\_mean\_average\_difference = df\_region\_mean\_average.nsmallest(1, ['Difference'])mean\_average\_info = bcolors.HEADER + 'Mean-Median Equality:\n' + bcolors.ENDCmean\_average\_info += f'\t{bcolors.OKBLUE}Region: {bcolors.OKGREEN}{df\_smallest\_mean\_average\_difference.axes[0][0]}'for column\_name in ['Mean', 'Average', 'Difference']: mean\_average\_info += f'\n\t{bcolors.OKBLUE}{column\_name}: ' \ f'{bcolors.OKGREEN}' \ f'{df\_smallest\_mean\_average\_difference[column\_name].values[0]}' \ f'{bcolors.ENDC}'print(mean\_average\_info)

Mean-Median Equality:

Region: Latin America & Caribbean

Mean: 10468.495457604762

Average: 10833.201075

Difference: 364.70561739523873

**Додаткове 6:**

# 6 Вивести топ 5 країн та 5 останніх країн по ВВП та кількості СО2 на душу населення.dataset\_gdp\_desc = dataset.sort\_values('GDP per capita', ascending=False)dataset\_gdp\_asc = dataset.sort\_values('GDP per capita', ascending=True)dataset\_co2\_desc = dataset.sort\_values('CO2 emission', ascending=False)dataset\_co2\_asc = dataset.sort\_values('CO2 emission', ascending=True)print(f'{bcolors.HEADER}GDP Top 5:{bcolors.ENDC}\n', dataset\_gdp\_desc.head(5).to\_string())print(f'{bcolors.HEADER}GDP Bottom 5:{bcolors.ENDC}\n', dataset\_gdp\_asc.head(5).to\_string())print(f'{bcolors.HEADER}CO2 Top 5:{bcolors.ENDC}\n', dataset\_co2\_desc.head(5).to\_string())print(f'{bcolors.HEADER}CO2 Bottom 5:{bcolors.ENDC}\n', dataset\_co2\_asc.head(5).to\_string())

GDP Top 5:

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

GDP Bottom 5:

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

CO2 Top 5:

Country Name CO2 emission

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

CO2 Bottom 5:

Country Name CO2 emission

201 Tuvalu 11.001

113 Liechtenstein 44.004

137 Nauru 47.671

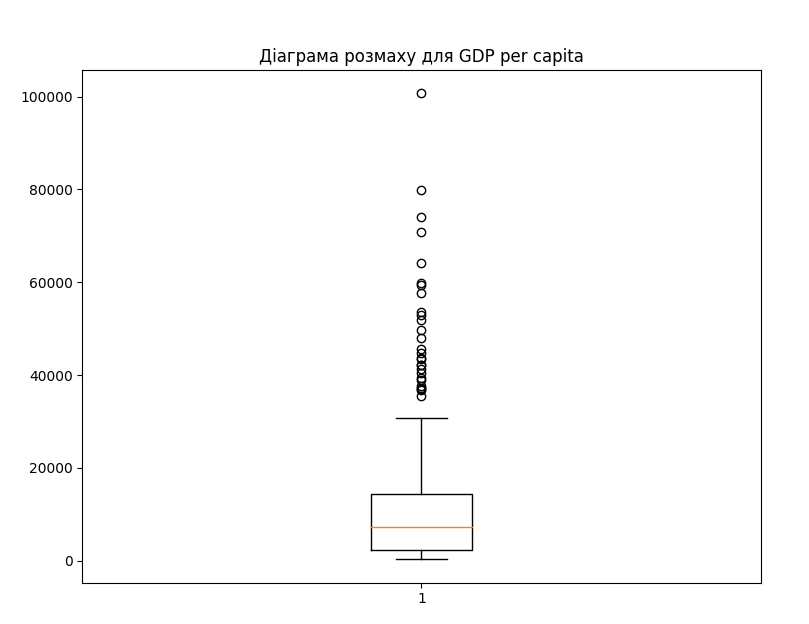
101 Kiribati 62.339

124 Marshall Islands 102.676

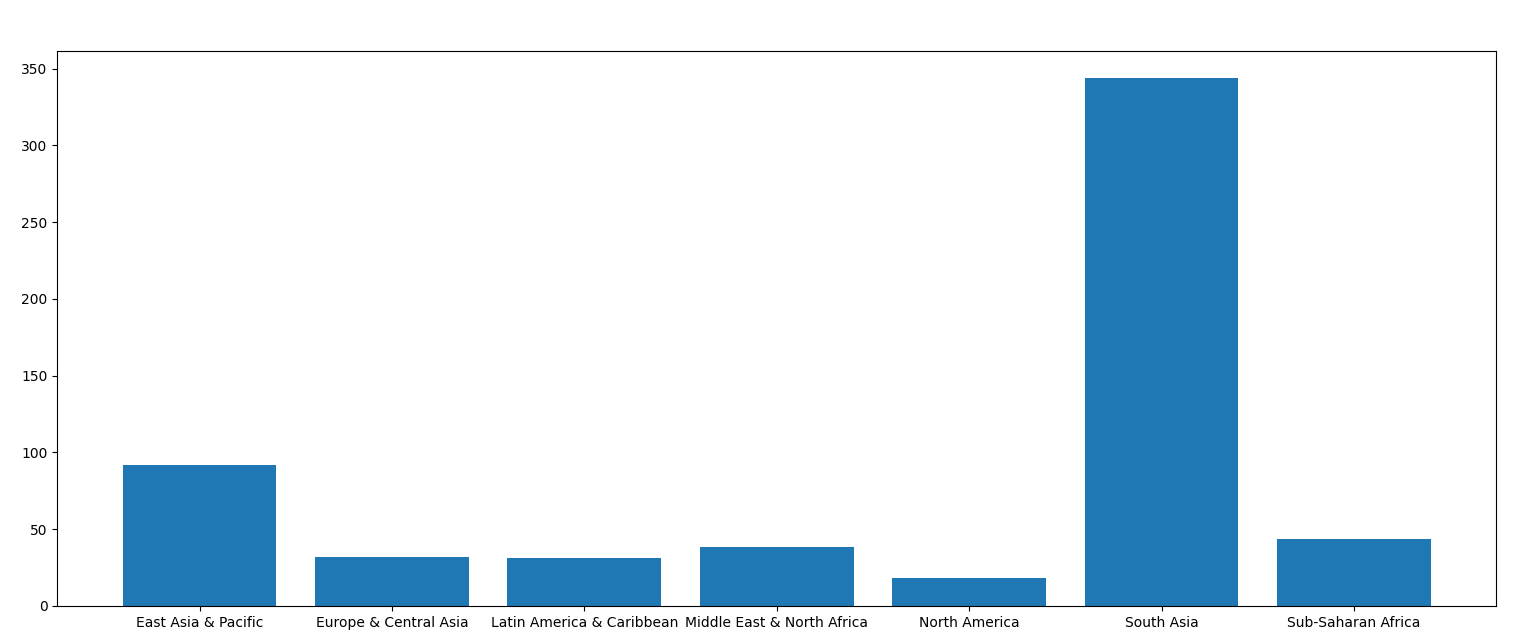
**6. Побудувати діаграми розмаху та гістограми**

#Plot average density by regions  
df\_group\_by\_region = dataset.groupby('Region')  
df\_region\_area\_sum = df\_group\_by\_region.sum(numeric\_only=True)[['Area']]  
df\_region\_population\_sum = df\_group\_by\_region.sum(numeric\_only=True)[['Population']]  
df\_region\_density = df\_region\_population\_sum['Population'] / df\_region\_area\_sum['Area']  
df\_region\_density: pd.DataFrame = df\_region\_density.to\_frame()  
df\_region\_density.columns = ['Density']  
print(df\_region\_density, df\_region\_density.axes)  
  
  
  
plt.bar(df\_region\_density.index, df\_region\_density['Density'])  
boxplot(dataset, 'GDP per capita')  
plt.show()

**Побудувати діаграми розмаху** GDP per capita:



**Гістограма щільності населення по регіонах**

****

**Висновок:**

**Відредагував дані, знайшов середні значення, медіани стопчиків датафрейму. Код написаний на пайтоні, графіки наведені.**