*Додаток 1*

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, Недельчев Євген Олександрович

(шифр, прізвище, ім'я, по батькові)

Перевірив Олійник Юрій Олександрович

( прізвище, ім'я, по батькові)

Київ 202 3

# **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

Додаткове:

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.

## **Основне завдання**

### **DataFrame та його структура**

За допомогою Python бібліотеки Pandas завантажимо дані з даного csv файлу в dataframe та досліджуємо структуру наших даних, використовуючи скрипти нижче. Підчас завантаження одразу змінимо назву колонки ‘Populatiion’ на ‘Population’.

Изображение выглядит как стол

Автоматически созданное описание

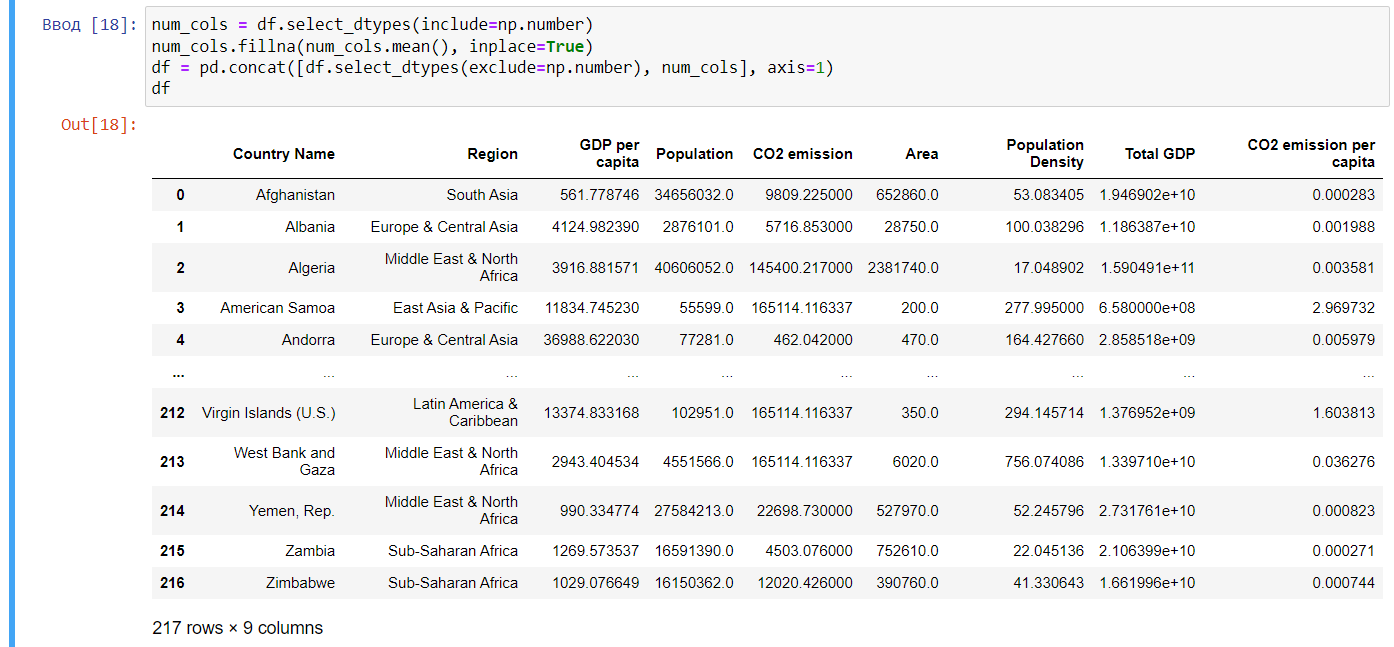
На даному рисунку можна помітити загальну інформацію про датафрейм: кількість рядків та колонок, назви всіх колонок, кількість записів в кожній з них, тип даних колонки та використання пам’яті.

### **Виправлення помилок**

Замінимо рядки, поля яких містять від’ємні елементи.



Також в даних наявні пусті елементи, які потрібно замінити середніми по стовпчику.



### **Діаграми розмаху та гістограми**

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

Для формування графіку розмаху даних використаємо наступний скрипт:

Изображение выглядит как текст

Автоматически созданное описание

Отримали наступні графіки:

Изображение выглядит как диаграмма

Автоматически созданное описание

Для формування гістограм використаємо наступний скрипт:

Изображение выглядит как текст, письмо

Автоматически созданное описание

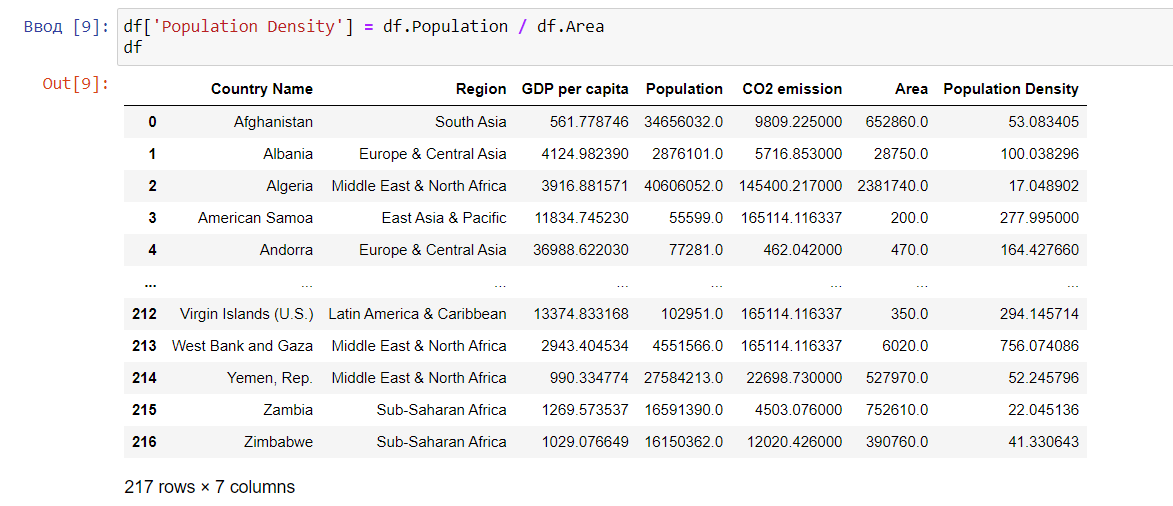
Отримали такі гістограми:

Изображение выглядит как стол

Автоматически созданное описание

### **Додавання стовпчику із щільністю населення**

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



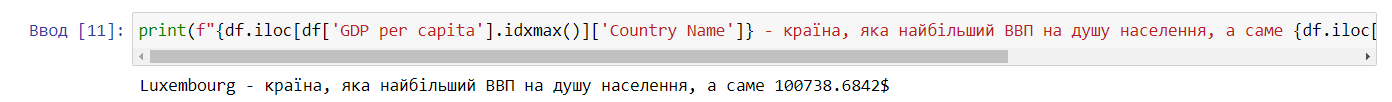
## **Додаткове завдання**

### **Заміна пропущених значень**

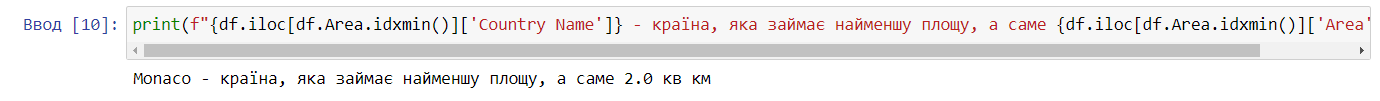
Демонстрація заміни пропущених значень описана в розділі «Виправлення помилок».

### **Країна з найбільшим ВВП на людину, з найменшою площею**

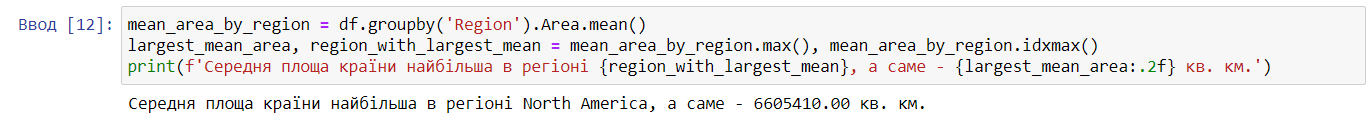
Виведемо країну з найбільшим ВВП на душу населення:



Виведемо країну з найменшою площею:



### **Регіон з найбільшою середньою площею країн**



### **Країна з найбільшою щільністю населення у світі; у Європі та центральній Азії**

Изображение выглядит как календарь

Автоматически созданное описание

### **Співпадіння середнього та медіани ВВП по регіонам**

Алгоритм дій тут доволі простий. Спочатку додамо стовпчик, який представлятиме загальний ВВП країни. Далі згрупуємо країни по регіонам та проітеруємося по цій групі, виводячи значення медіани та середнього. Якщо вони співпадають, то додаємо їх в список. І в кінці, якщо список не пустий, виводимо такі регіони.

Изображение выглядит как текст

Автоматически созданное описание

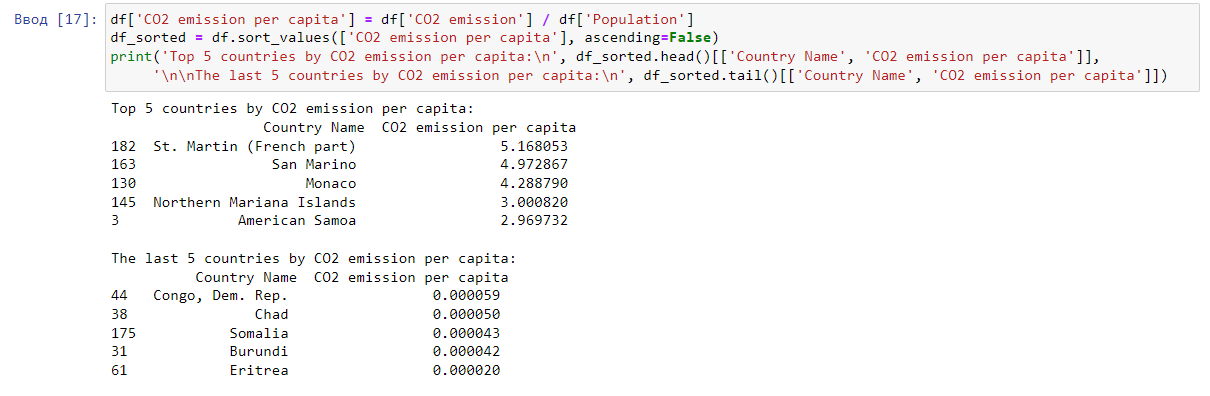
### **Топ 5 країн та 5 останніх країн по ВВП та кількості СО2 на душу населення**

Для виводу найкращих та найгірших країн за показником ВВП на душу населення відсортуємо наші дані по цьому стовпцю, та виведемо їх за допомогою методів head() і tail().



Для виводу найкращих та найгірших країн за показником кількості СО2 на душу населення додамо до наших даних стовпець «CO2 emission per capita», який по суті є результутатом ділення кількості населення країни на концентрацію СО2. Далі

відсортуємо наші дані по цьому стовпцю, та виведемо їх за допомогою методів head() і tail().



Висновок

У цьому комп’ютерному практикуму було вивчено можливості Python, а саме Pandas у роботі з даними. Вхідні дані було записано в DataFrame, структуру якого було вивчено та помічено нецілісність даних, тому я почистив дані від від’ємних значень, нульові замінив середніми для більш об’єктивної побудови гістограм та діаграм розмаху. 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 СО2, дані з ВВП на душу населення є найбільш кучними. Було визначено країну з найбільшим ВВП на душу населення у світі, з найменшою площею території, регіон з найбільшою середньою площею країн, країни з найбільшою густиною населення у світі та окремо в регіоні «Європа та центральна Азія». Регіонів з однаковими середньою та медіаною ВВП країн не виявилось, усі мають різні. Також було виведено 5 країн з найбільшим та найменшим ВВП на душу населення та 5 з найбільшою та найменшою кількістю викидів СО2.