# 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
+ Data visualization with ggplot2: cheatsheet
+ Гнатюк В. Вступ до R на прикладах
+ Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R
+ Data Mining for Business Analytics

Python

+ Довідкові матеріали для початку роботи з Python
+ Учебник по Python
+ Подборка блокнотов по командам IPython
+ Інтерактивний міні-гайд по візуалізації даних на Python
+ Элбон  Машинное обучение с использованием Python 2019
+ Python NumPy beginners
+ Campesato Pandas Basics
+ Пособие по Matplotlib
+ Первичный анализ данных с Pandas
+ Python Data Science Handbook

**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.

**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
Поглиблено про графічне представлення інформації
кругові діаграми
діаграми розсіювання
діаграми розмаху
Приклад R

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, Python

**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. 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 CO2 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).