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 лабораторної роботи № 3 з дисципліни

«Основи науки про дані»

«Робота з пропущеними даними»

Виконав студент \_\_\_\_ІП-11 Трикош І. В\_\_\_\_

Перевірив \_\_\_Новотарський М. А.\_\_\_

Київ 2024

**Завдання**

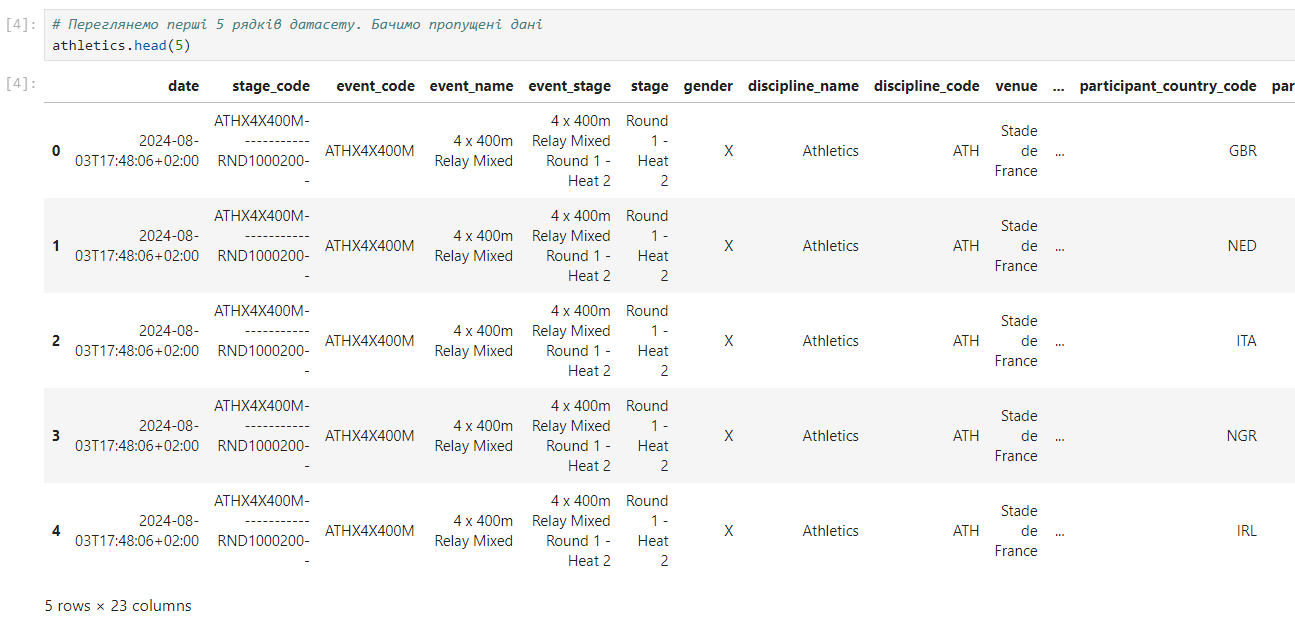
1. Детально розглянути наведені теоретичні відомості до даної лабораторної роботи.
2. Ознайомитися з кодом, що наведений у Jupyter notebook DD\_lab3.ipynb.
3. Виберіть набір даних відповідно до номеру у списку з таблиці 3.
4. Виконати етапи аналізу та попередньої обробки даних, наведені в даній лабораторній роботі.

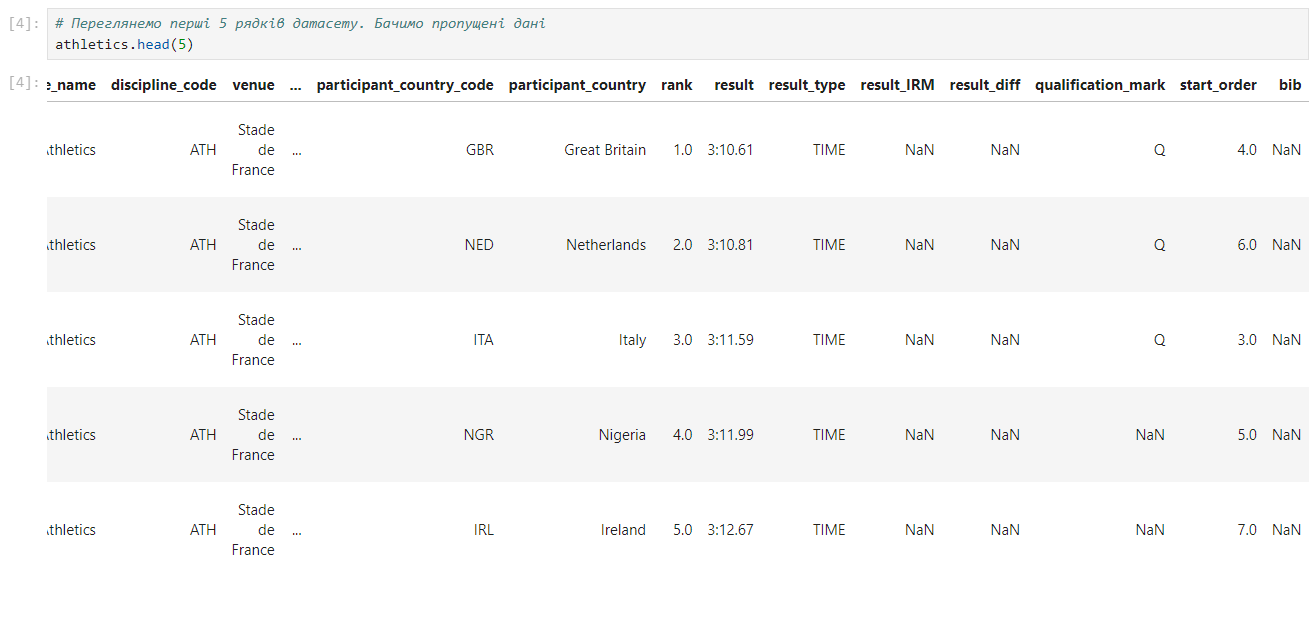
**Виконання**

Для виконання лабораторної роботи візьмемо датасет про Олімпійські ігри 2024: <https://www.kaggle.com/datasets/muhammadehsan02/olympic-summer-games-paris-2024/data> (<https://drive.google.com/file/d/1rtZkHbKiEhNY6wtvAbPYGKzfu28n0EVm/view?usp=sharing>)

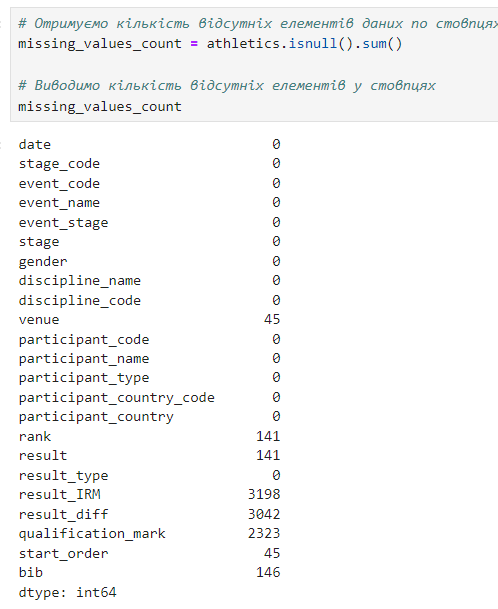
Працюватимемо з файлом Athletics.csv, що містить дані про змагання з атлетики.

Перше ознайомлення з даними. Переглянемо перші 5 рядків датасету з допомогою методу head, який бере перші N рядків з датасету:

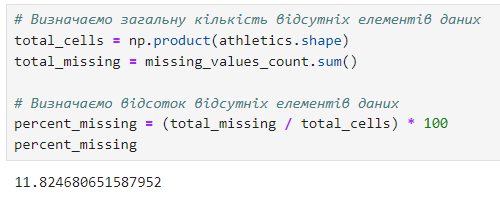




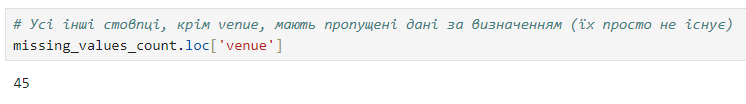
Визначимо кількість відсутніх даних по стовпцях. Для цього використаємо методи isnull та sum. isnull поверне датасет, у якому булеві значенння, що відповідають тому, чи пропущене значення. sum підсумує усі true значення в отриманому датасеті. Бачимо, що відсутніх даних доволі багато, а особливо у стовпцях result\_IRM, result\_diff і qualification\_mark:



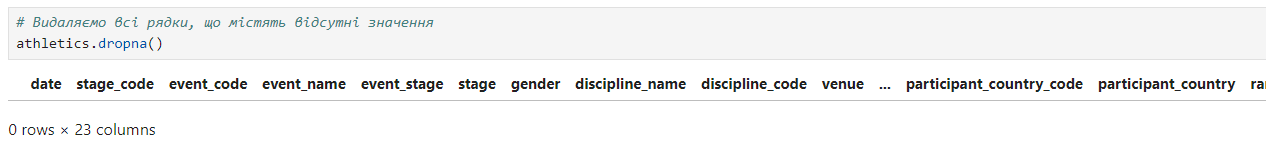
Тепер визначимо відсоток відсутніх елементів даних. Для цього поділимо кількість пропущених та кількість усіх елементів. Відсоток доволі великий:



Визначимо причину відсутності даних. Майже всі відсутні дані відсутні за визначенням. Наприклад, у стовпці qualification\_mark присутні значення тільки у подіях кваліфікацій. Але у стовпці venue є трохи пропущених значень, які варто заповнити:

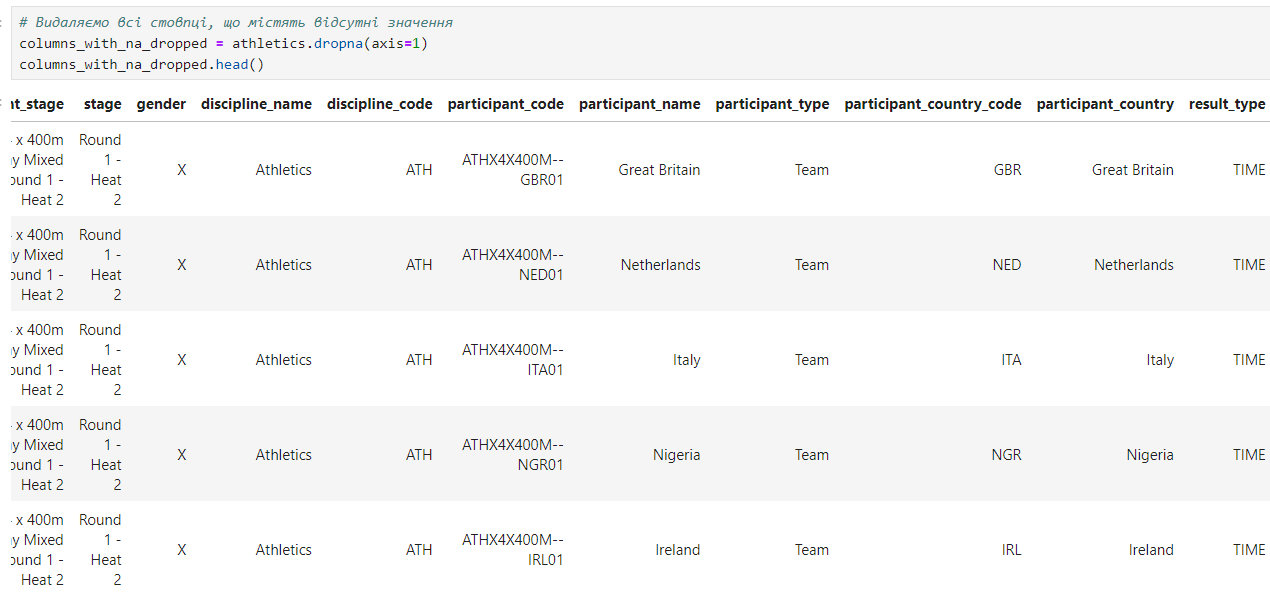


Видалимо усі рядки, де є пропущені значення. Для цього застосуємо метод dropna. Бачимо, що у всіх рядках є пропущені значення:

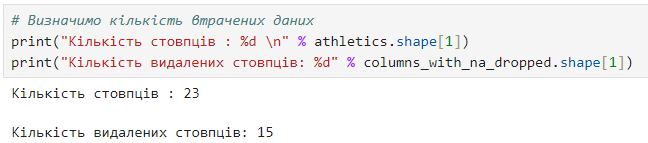


Видалимо усі стовпці, у яких є пропущені значення. Для цього застосуємо метод dropna з параметром axis=1, який і видалить стовпці:

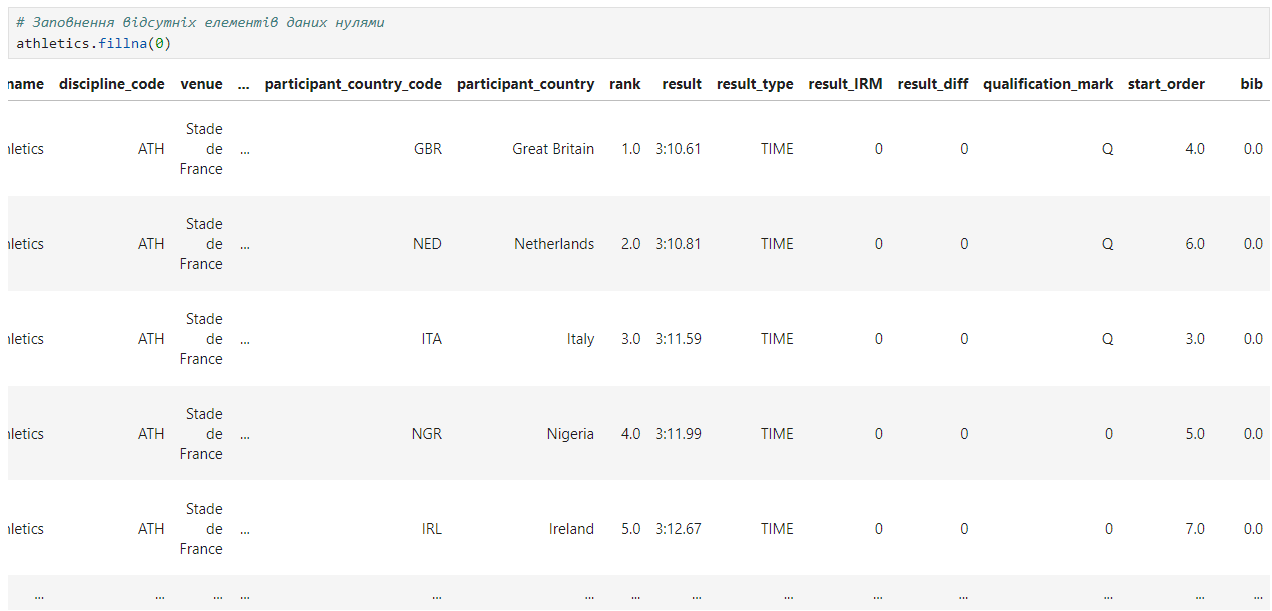




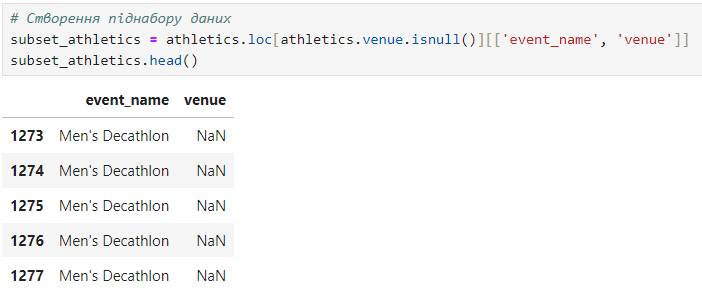
Визначимо кількість втрачених даних. Метод shape поверне кортеж, що репрезентує розмірність датафрейму. Бачимо, що видалилась значна кількість стовпців:



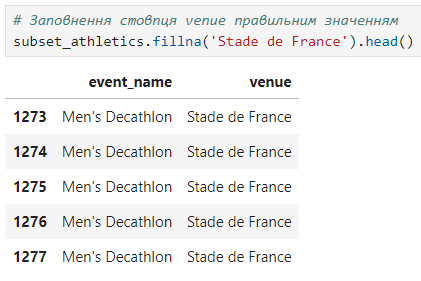
Заповнимо відсутні елементи даних нулями, скориставшись методом fillna. Заповнено зокрема стовпці result\_IRM, result\_diff, qualification\_mark, bib:



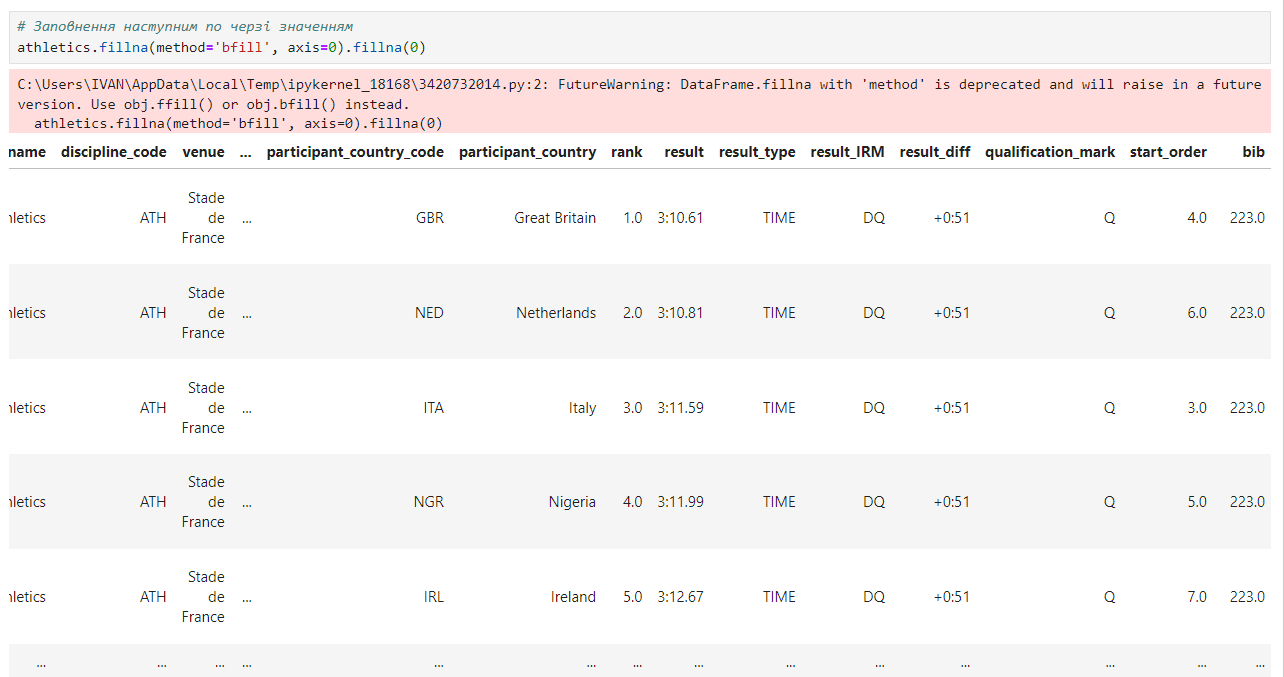
Створимо піднабір даних для заповнення лише потрібного нам стовпця. Для цього оберемо лише рядки, що не містять venue, та їх стовпці event\_name (для кращого сприйняття) та venue:



Заповнимо пропущені значення потрібним значенням (тим місцем, де відбувалися змагання). Для цього скористаємося методом fillna, куди передамо місце змагань. Бачимо, що значення успішно заповнені:



Заповнимо відсутні значення тими значеннями, які йдуть відразу після відсутніх, скориставшись методом fillna і передавши параметри method=’bfill’ (цей метод заповнює відсутні значеннями наступними невідсутніми значеннями) та axis=0 (заповнюємо по рядках, тобто братимемо наступне значення по стовпцях), і наостанок заповнимо нулями значення, що залишились. Бачимо, що тепер немає пропущених значень:



**Висновки**

Під час виконання лабораторної роботи я навчитися працювати з наборами даних, які містять відсутні або помилкові дані, пройшовши усі 4 етапи аналізу таких даних.