Міністерство освіти і науки України

Національний технічний університет України «Київський політехнічний

інститут імені Ігоря Сікорського"

Факультет інформатики та обчислювальної техніки

Катедра інформатики та програмної інженерії

Звіт

з лабораторної роботи № 1 з дисципліни

«Прикладні задачі машинного навчання»

«Введення в Data Science»

Виконав студент ІП-11 Лесів Владислав Ігорович

Перевірив Нестерук Андрій Олександрович

Київ 2023

**Лабораторна робота 1**

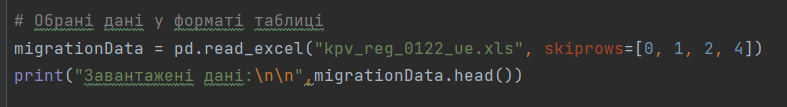
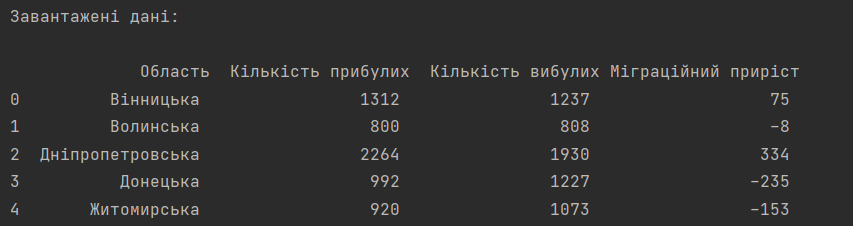
**Введення в Data Science**

**Постановка завдання.**

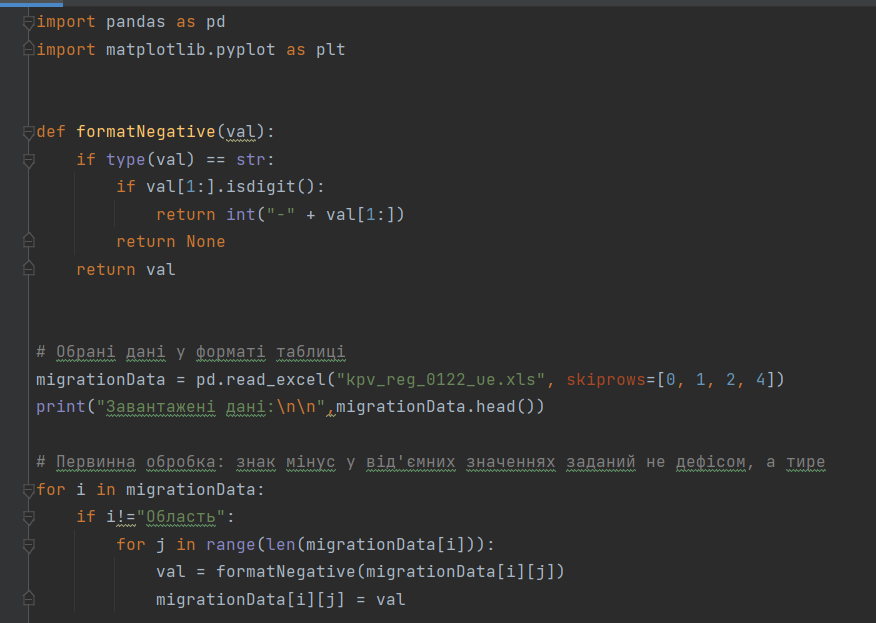
1. На сайті <http://www.ukrstat.gov.ua/> обрати дані, які є цікавими (можна використати будь-який ресурс з відкритими даними), та завантажити дані;
2. Знайти математичне сподівання, медіану, моду, дисперсію, середньоквадратичне відхилення (пояснити їх зміст);
3. Візуалізувати завантажені дані за допомогою гістограм;
4. Для даних проробити всі дії з пункту «Колекції Series і DataFrame бібліотеки pandas»;
5. Виконати первинну обробку даних;
6. Прочитати набір даних катастрофи «Титаніка»;
7. Завантажити набір даних катастрофи «Титаніка» за URL-адресою;
8. Переглянути рядки набору даних катастрофи «Титаніка»;
9. Налаштувати назви стовпців: замінити 'Unnamed: 0' на 'name' і скоротити 'passengerClass' до 'class';
10. Провести простий аналіз даних: визначити наймолодшого пасажира, найстаршого, який був середній вік пасажирів та статистику за пасажирами, які вижили. Відсортувати всіх жінок з кают 1-го класу, знайти наймолодшу та найстаршу серед них, кількість уцілілих;
11. Побудувати гістограму віку пасажирів.

**Хід роботи.**

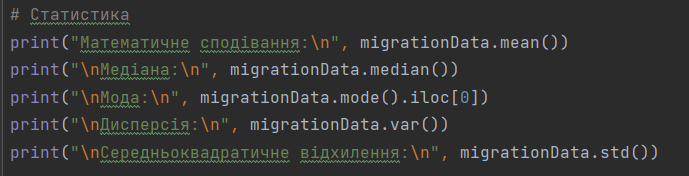
1. Для дослідження я обрав дані «Кількість прибулих, вибулих по регіонах у січні 2022 року». Оскільки дані в таблиці Excel, використовую метод read\_excel для прочитання.

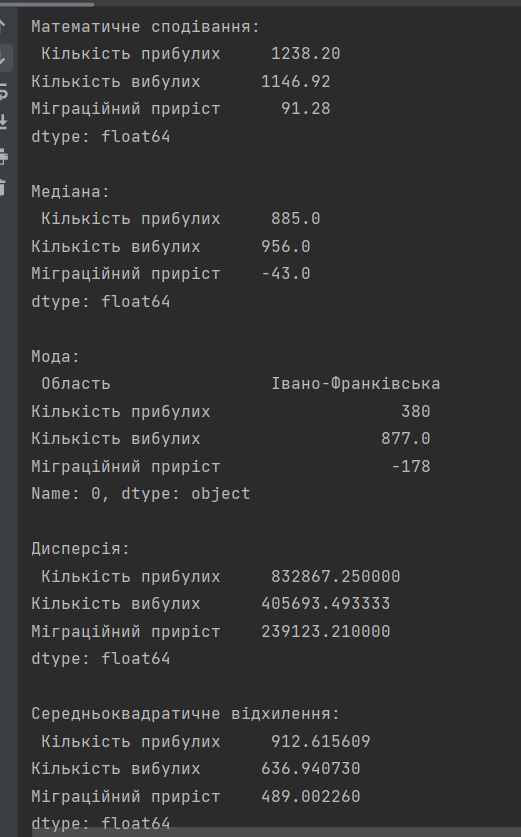
 

1. Спочатку виконаю первинну обробку даних для коректного прорахунку майбутніх досліджень. Під первинну обробку попадає той факт, що від’ємні значення міграційного приросту записані з довгою рискою як з мінусом, а не з короткою, тож такі значення інтерпретатор не зчитує, як числа. Отож заміняю усі такі значення відповідно на правильний мінус, після цього маємо від’ємні числа.

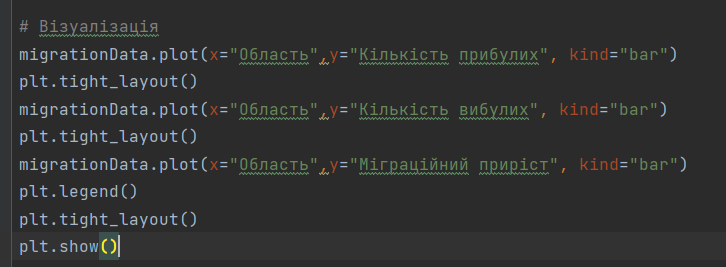
****

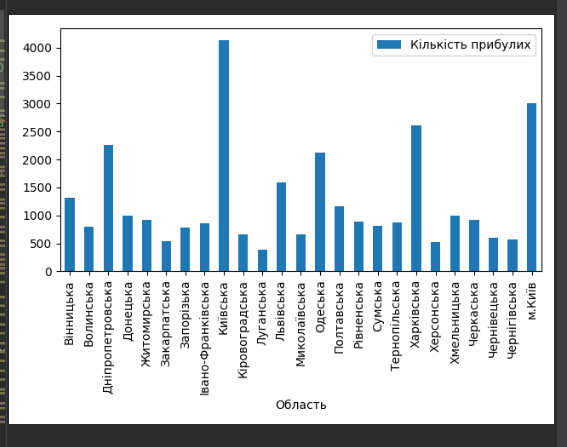
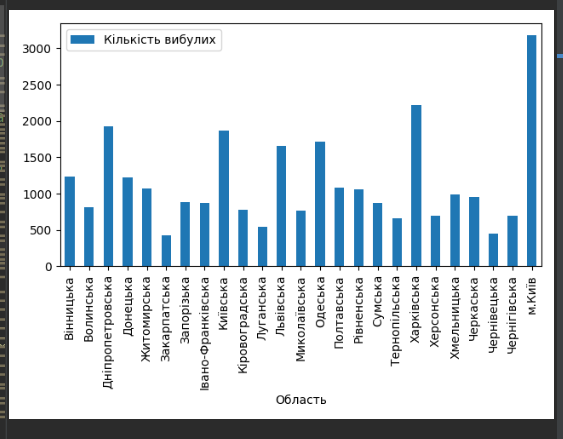
1. Знаходжу математичне сподівання (середнє значення), медіану (величина посередині відсортованого ряду), моду (найчастіше значення), дисперсію (розкид випадкової величини навколо її математичного очікування), середньоквадратичне відхилення (корінь з дисперсії) за допомогою відповідних методів. У моді обираємо показ першого рядка, адже значення моди виводяться у першому рядку, решта значень рядків залишаються NaN;

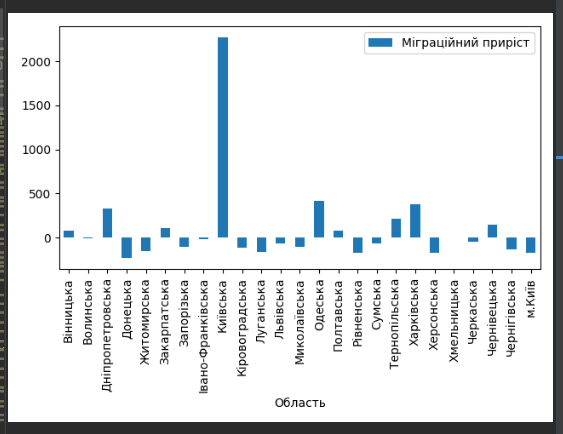




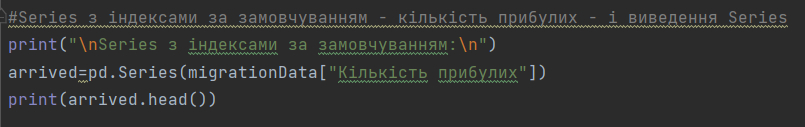
1. Візуалізую дані за допомогою гістограм. У гістограмах показані відповідні стовпці за областями.

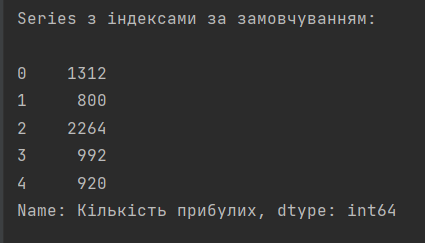
****

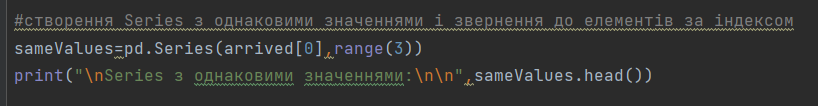
 

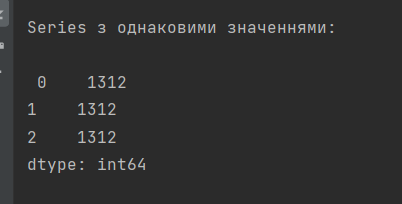
****

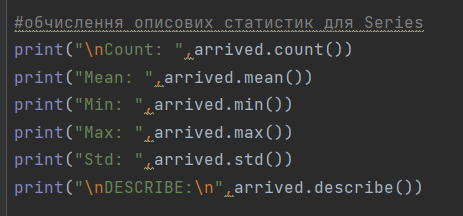
1. Для даних пророблюю всі дії з пункту «Колекції Series і DataFrame бібліотеки pandas». Для Series:

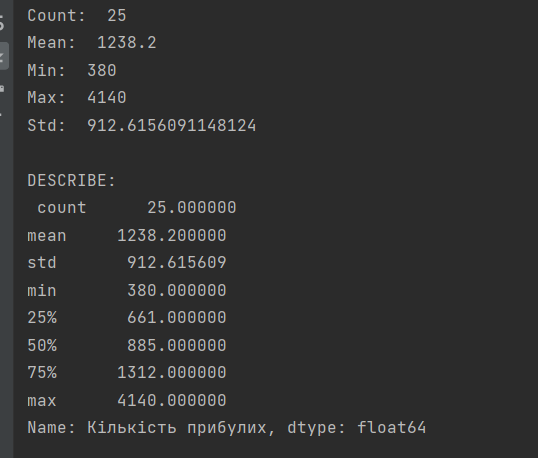


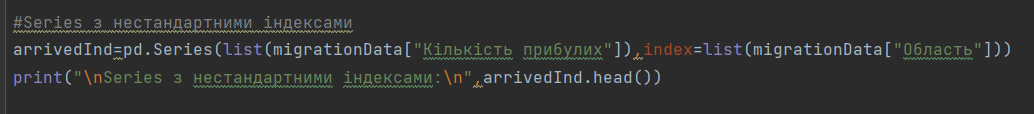


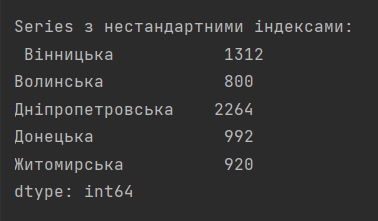


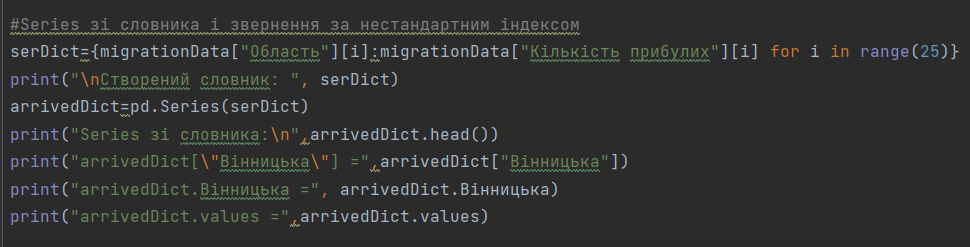


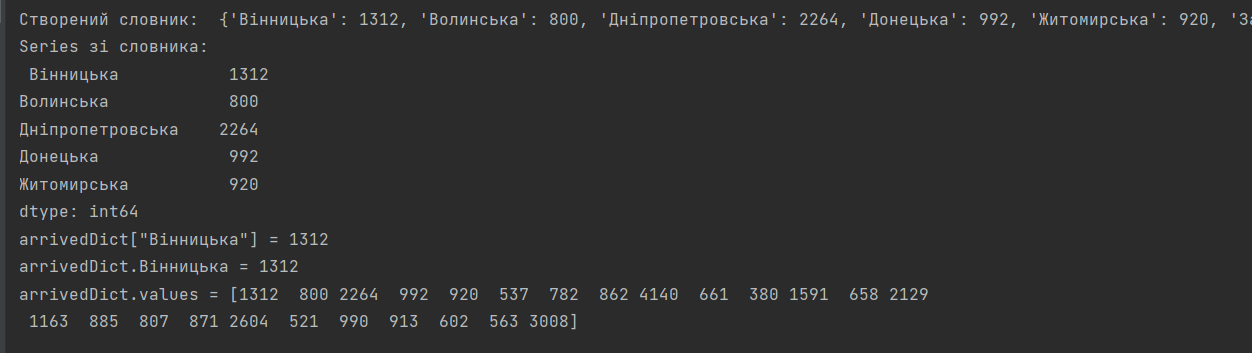


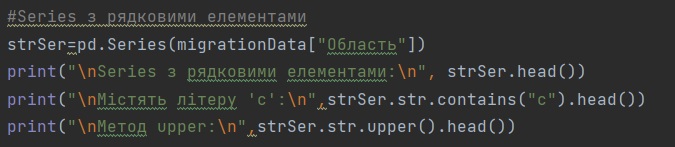


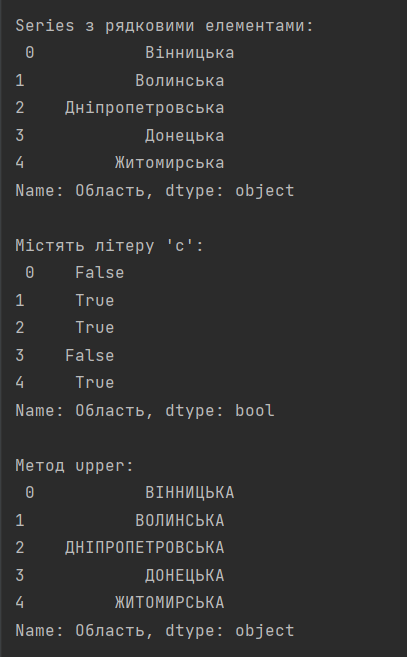




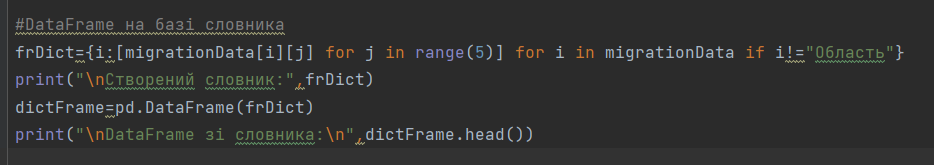


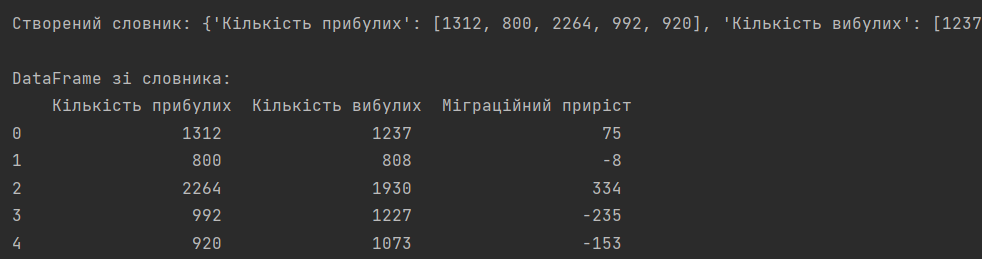


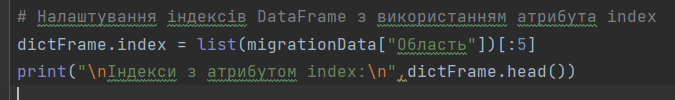


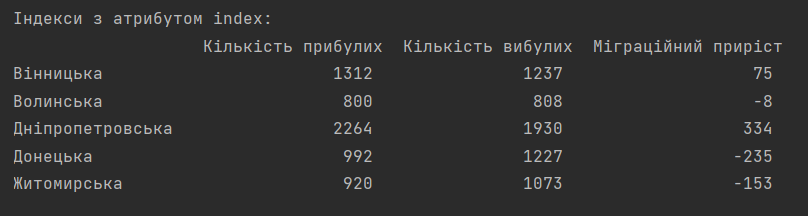


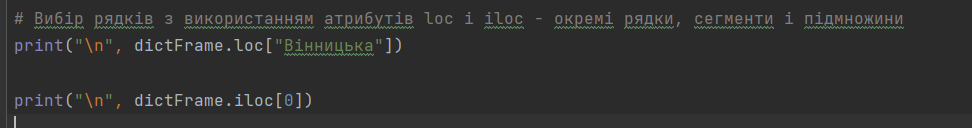
Для DataFrame:

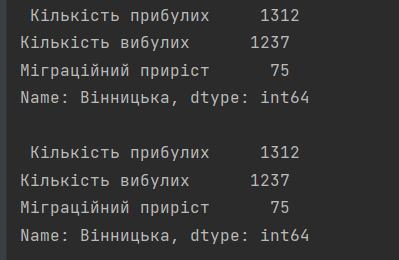


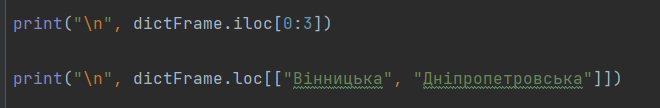


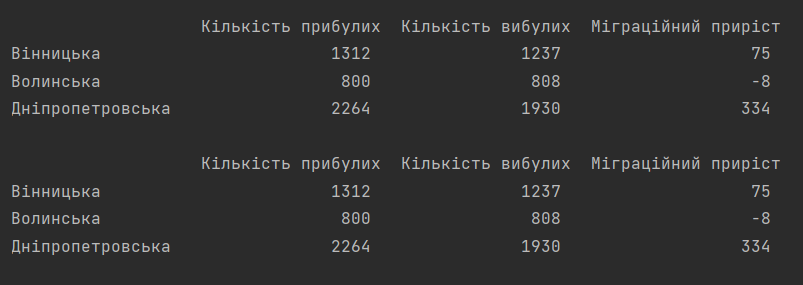


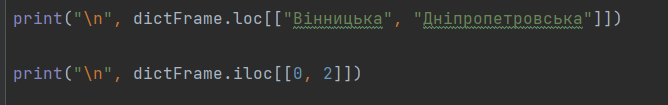


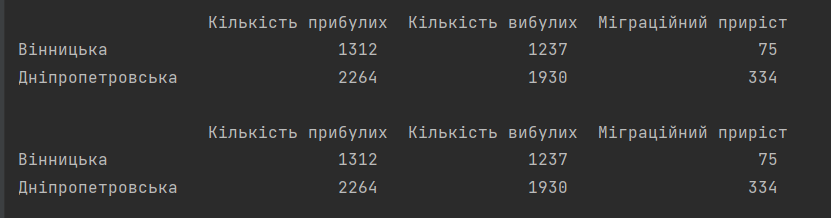


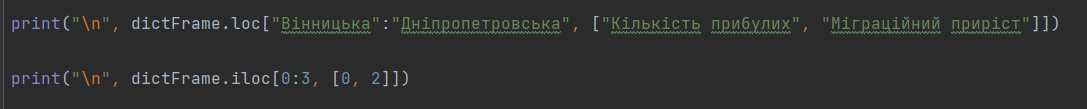


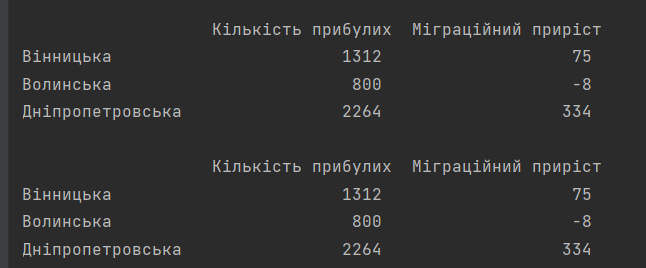


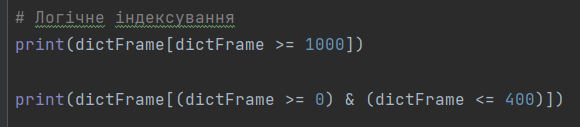


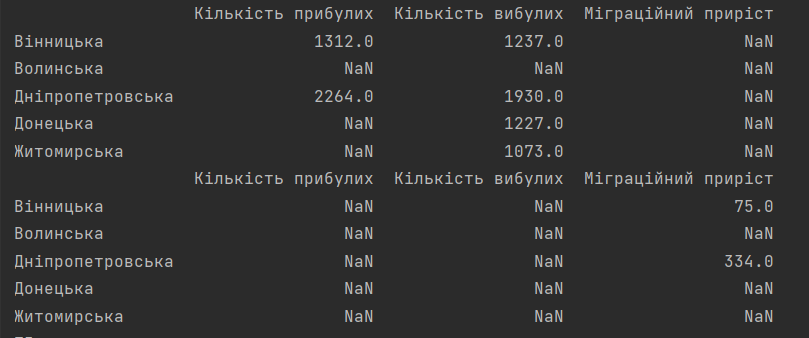


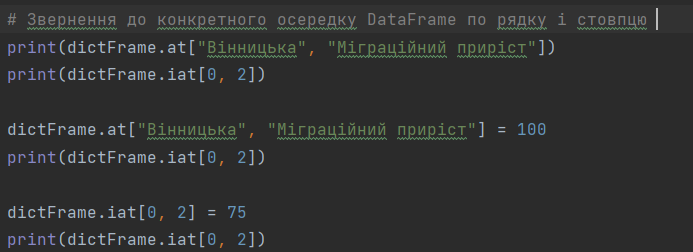


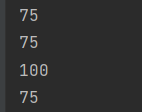


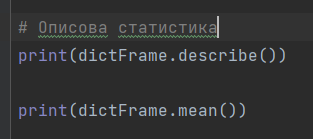


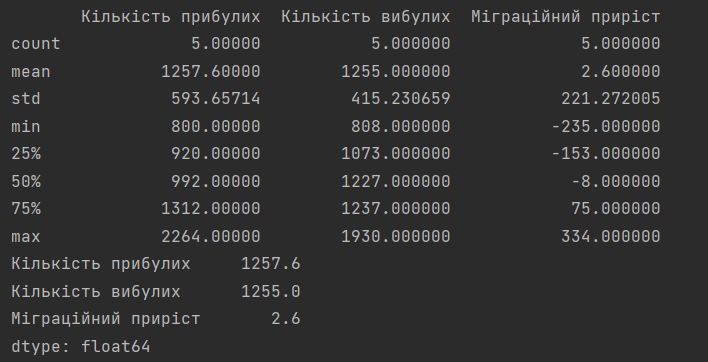


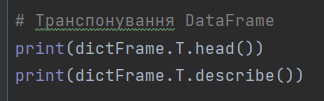


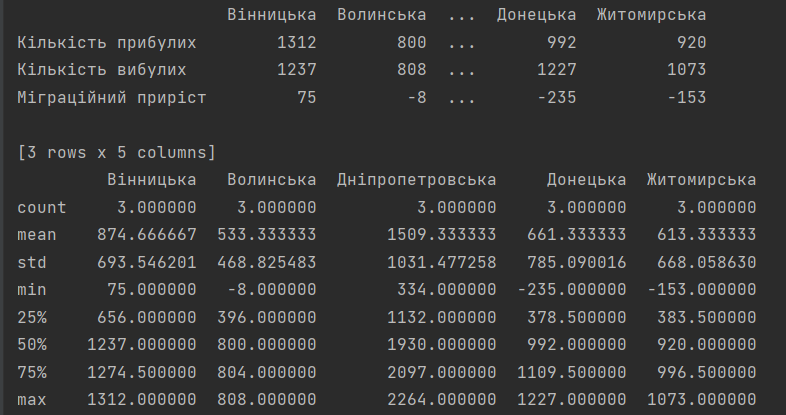


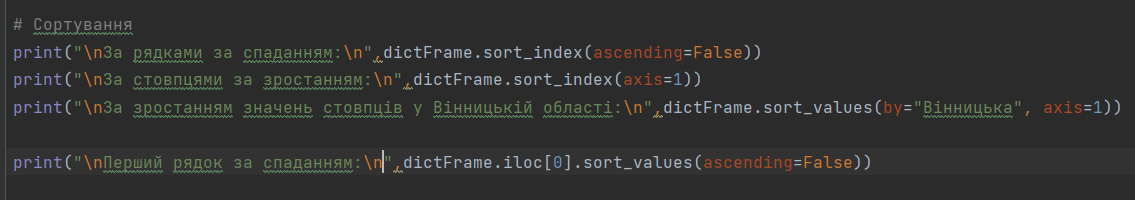


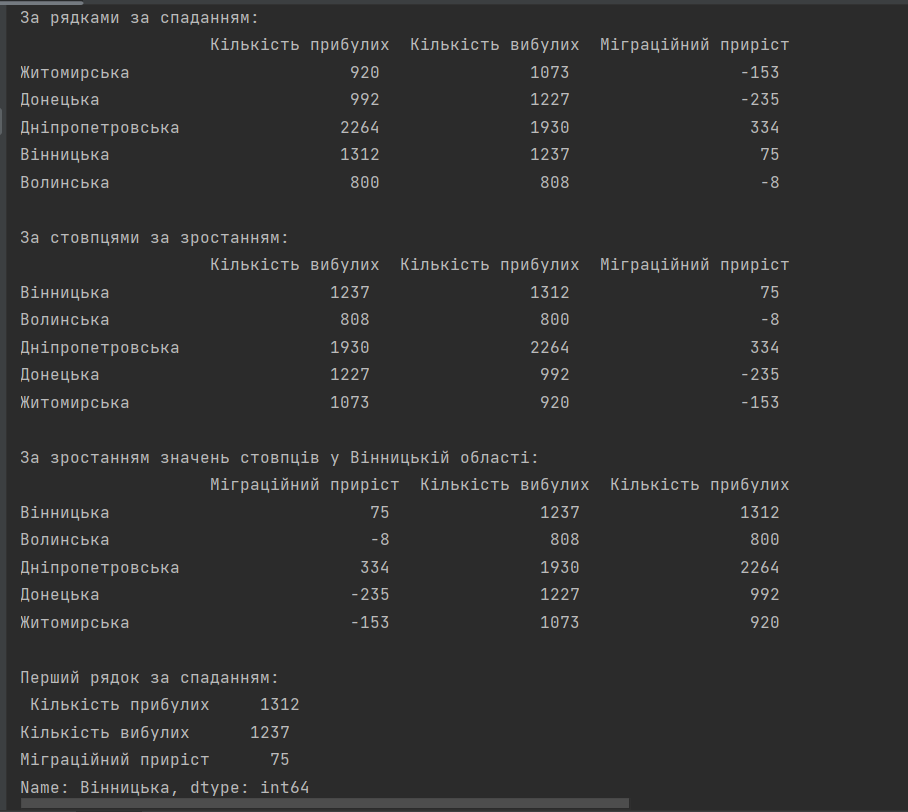




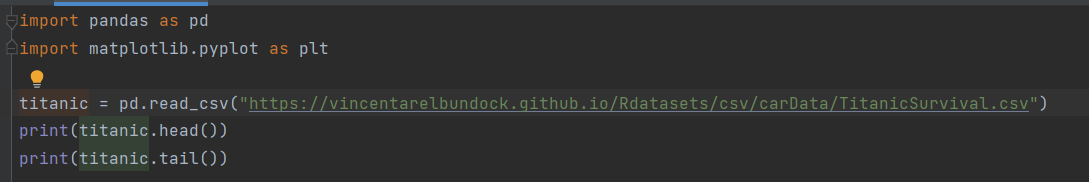


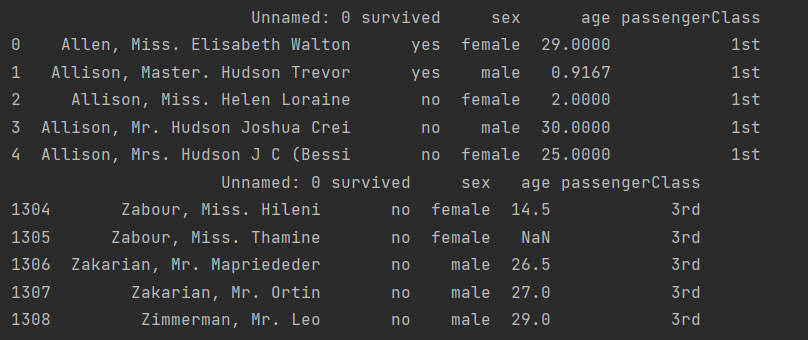




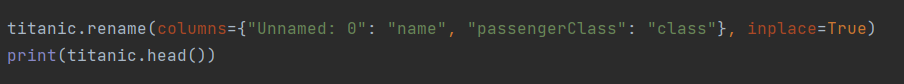


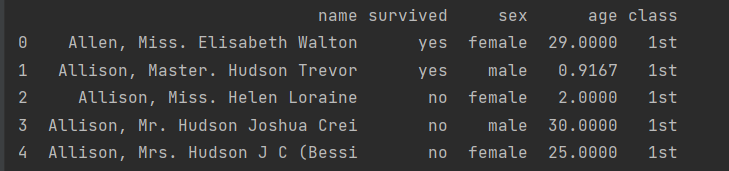
1. Завантажую набір даних катастрофи «Титаніка» за URL-адресою і переглядаю дані.



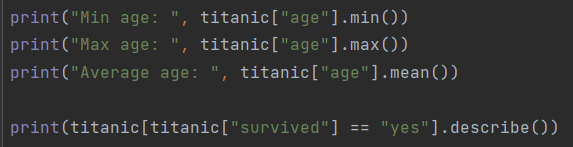
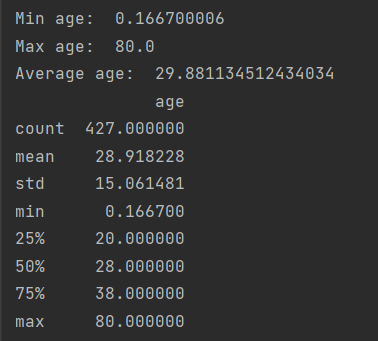


1. Налаштовую назви стовпців: замінюю 'Unnamed: 0' на 'name' і скорочую 'passengerClass' до 'class'; для цього я використав метод rename.

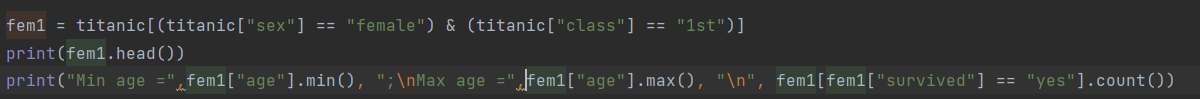


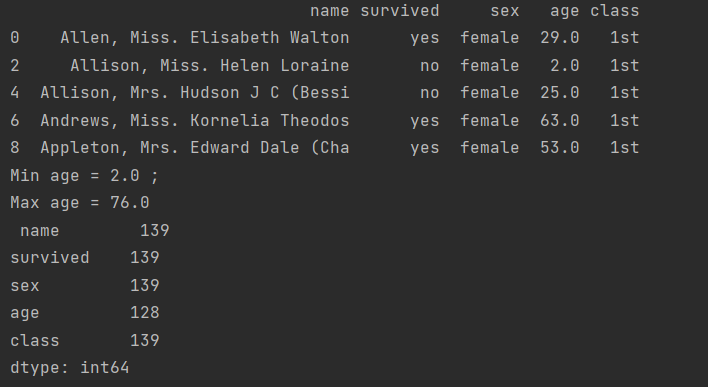


1. Проводжу простий аналіз даних: визначаю наймолодшого пасажира, найстаршого, який був середній вік пасажирів та статистику за пасажирами, які вижили.

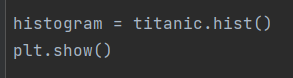
 

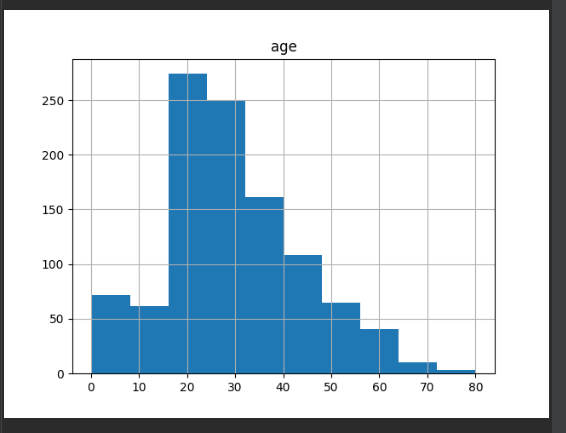
Сортую всіх жінок з кают 1-го класу, знаходжу наймолодшу та найстаршу серед них, кількість уцілілих;





1. Будую гістограму віку пасажирів.





**Висновок.**

Отже, у цій роботі я отримав навички роботи зі статистичними даними, їх обробки, виведення статистики, візуалізації та вчинення відповідних дій. Data Science може значно збільшити ефективність і продуктивність бізнесу та принести багато корисних результатів. У результаті лабораторної роботи було вивчено роботу з такими структурами даних модулю Pandas, як Series та DataFrame, створено різні Series I Dataframe з різних початкових структур даних, проведено дослідження різних статистик, створено гістограми для візуалізації віку пасажирів Титаніку й кількості міграцій людей за областями в січні 2022 року. Використовуючи програмний засіб PyCharm, мову програмування Python з модулями Pandas і matplotlib, отримуємо коректний результат.