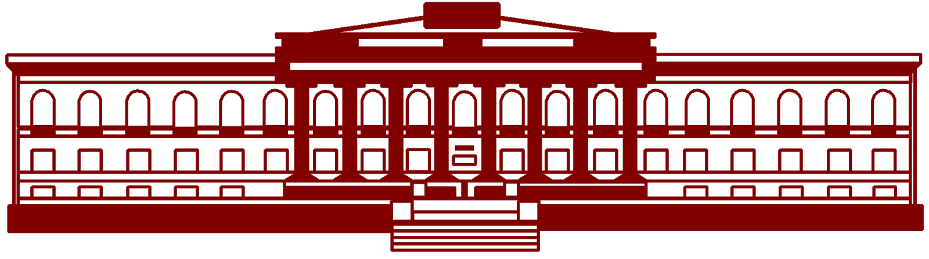
**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**імені ТАРАСА ШЕВЧЕНКА**

****

**ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**

**Кафедра прикладних інформаційних систем**

**Звіт до лабораторної роботи №1**

**з курсу**

**«Data Science та Big Data»**

*студента 4 курсу*

*групи ПП-43*

*спеціальності 122 «Комп'ютерні науки»*

*ОП «Прикладне програмування»*

Пилиповича Д.О.

*Викладач*

Білий Р.О.

**Київ – 2023**

**Тема:** Агрегація, обробка пропусків та візуалізація даних пакетами Python

**Мета:** Отримання практичних навичок у роботі з raw data, використовуючи пакети jupyter, pandas, seaborn.

**Завдання:**

*Контекст:*

У дата сеті знаходяться 31 набір даних з іменами nyt1.csv, nyt2.csv, ..., nyt31.csv.

Кожен із них демонструє один (симульований) день показів оголошень та переходів по них, записаних на головній сторінці газети The New York Times у травні 2012 року. Кожен рядок представляє одного користувача. Існує п'ять стовпців: вік, стать (0 = жінка, 1 = чоловік), кількість показів, кількість переходів та статус авторизації.

*Завдання до виконання*

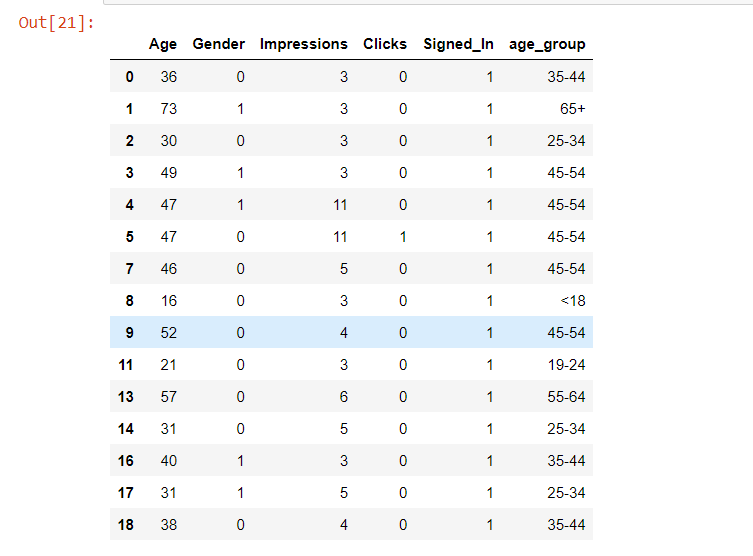
* Завантажити файли з даними у папку проекту з посилання: [**https://github.com/oreillymedia/doing\_data\_science**](https://github.com/oreillymedia/doing_data_science)
* Створіть нову змінну age\_group, яка агрегує користувачів як <18, 18–24, 25–34, 35–44, 45–54, 55–64 та 65+.
* Зафіксуйте на діаграмі кількість показів та показник переходів (CTR = #clicks/#inmpressions) для цих шести вікових категорій.
* Вивчіть дані та проведіть візуальні та кількісні порівняння між сегментами користувачів/демографічними групами (наприклад, чоловіки старше 18 років у порівнянні з жінками старше 18 років або авторизовані та неавторизовані користувачі).
* Створіть метрики/вимірювання/статистику, які підсумовують дані. Приклади можливих метрик включають CTR, квантил, середнє значення, медіану, дисперсію та максимальне значення. Ці показники потрібно розрахувати за різними сегментами користувачів. Подумайте про елементи, які важливо відстежувати з часом - що стискає дані, але, як і раніше, захоплює поведінку користувача.
* Результати статистичного дослідження подати у вигляді результуючого ДатаФрейма (одного), дивлячись на який можна зрозуміти і порівнювати дані за віковими підкатегоріями.
* Опишіть та інтерпретуйте будь-які закономірності, які знайдете.
* Завантажити файл ipynb з виконаними завданнями на git в окрему папку з відповідною назвою лабораторної роботи

**Хід виконання:**

Завантажуємо файли, які знаходяться за посиланням, яке наведено у завданні. Потім щоб приступити до виконання інших завдань, треба імпортувати бібліотеки та створити з інформації у файлах датафрейм зчитуючи інформацію у об’єднаних файлів один з одним. Далі було поділено на вікові групи, які вказані у завданні. Для виведення результату було ще видалені дані у яких не було вказано вік, бо вони не поділяються на групи і виводиться NaN. На рис 1 продемонстрована реалізація виконання.



*рис 1 (створення датафрейму, поділ на вікові категорії та видалення непотрібних даних)*

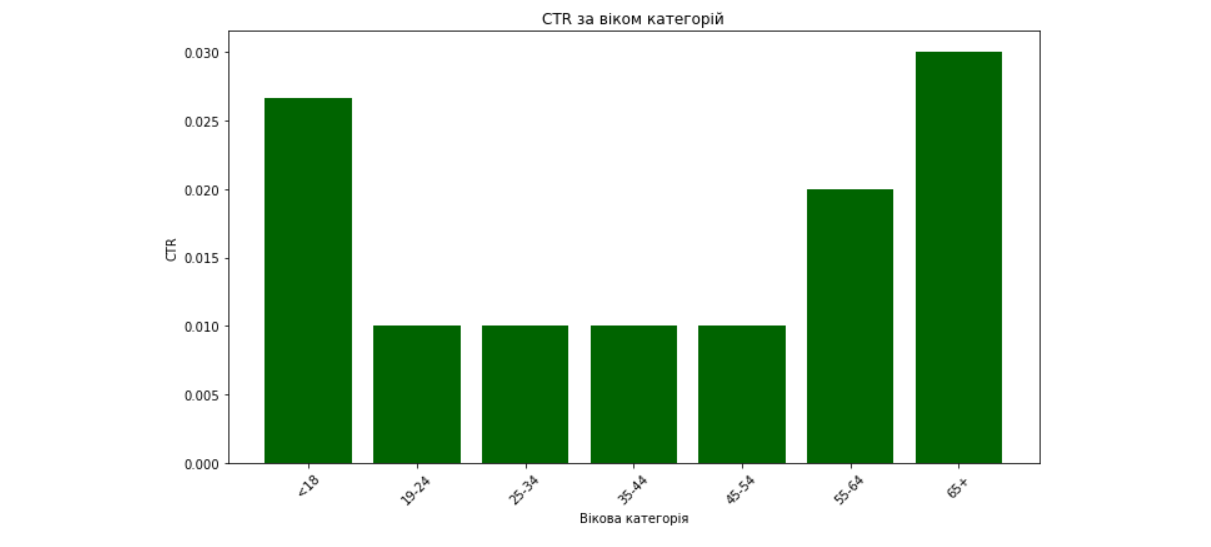


*рис 2 (результат виконання)*

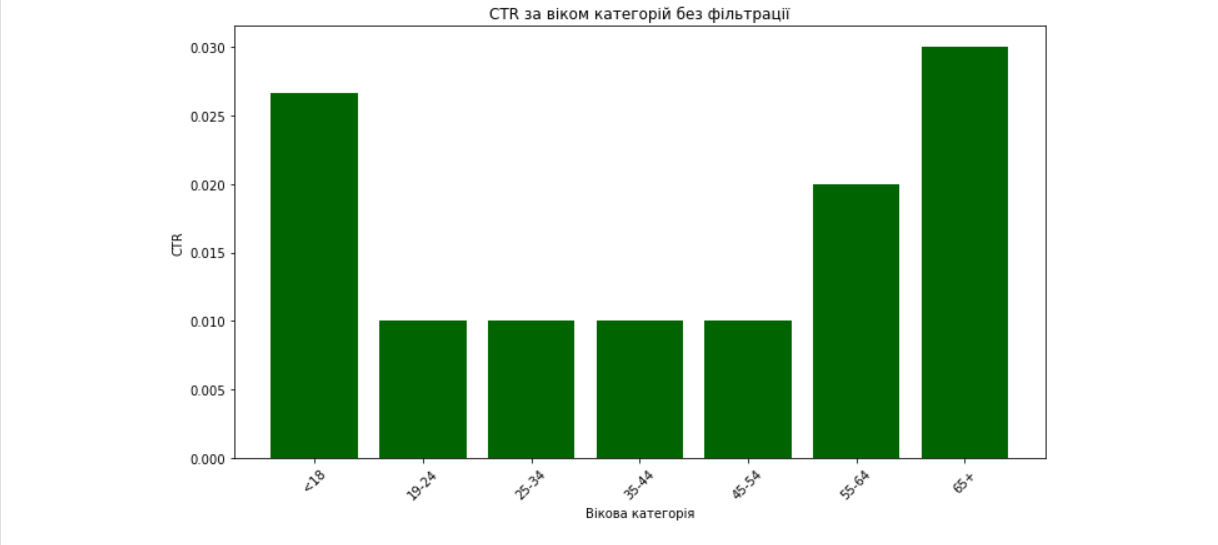
Далі було обрахованно CTR. CTR – це велечина, яка використовується наприклад, у сфері інтернет-маркетингу. Вона показує відношення кількості людей, які бачили рекламу і зробили те що вимагала вона. Щоб її обрахувати треба було кількість кліків поділити на кількість показів. Було обраховано для порівняння CTR для даних без NaN і з ним. Далі побудована була стовпчикова діаграма для виведення результату. На рис 3 продемонстрован виконання цього.



*рис 3 (виконання обрахунку CTR та побудова діаграми)*



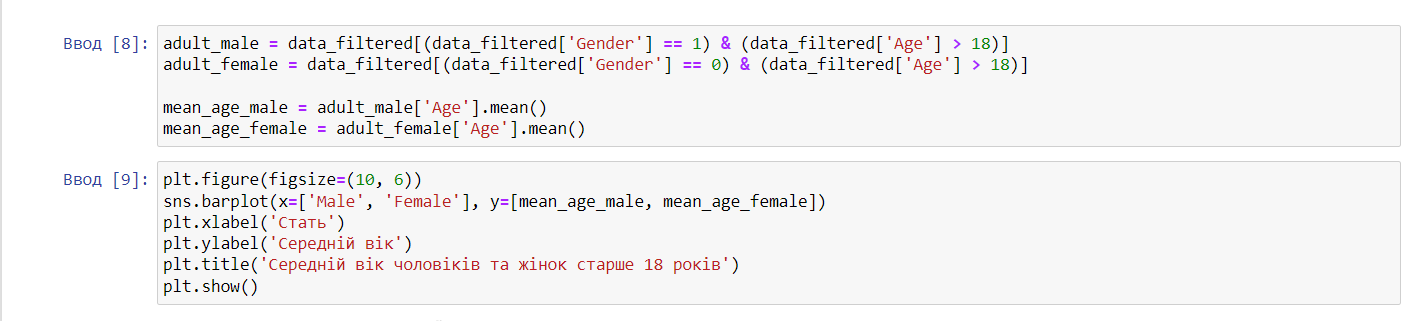
*рис 4 (діаграма відфільтрованих даних)*



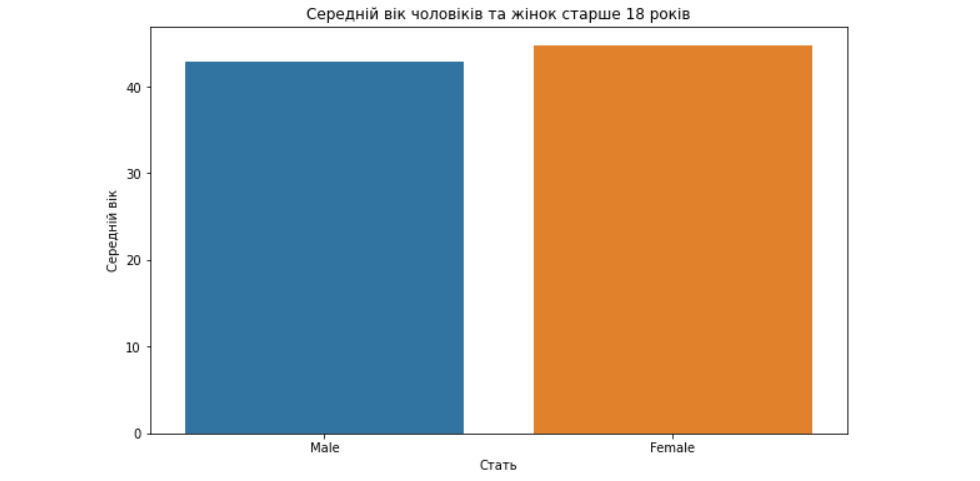
*рис 5 (діаграма невідфільтрованих даних)*

Як можна побачити ситуація не змінилась. Це сталось, бо користувачі без віку не стосуються жодної з категорій віку і тому змін не відбулось.

Далі була реалізація порівняння середнього віку чоловіків та жінок датафрейму. Для цього було взято окремо чоловіки і жінки та обраховано середній вік цих людей за допомогою функції mean(). Далі було побудована діаграма стовпчикова, яка відображає цю інформацію.

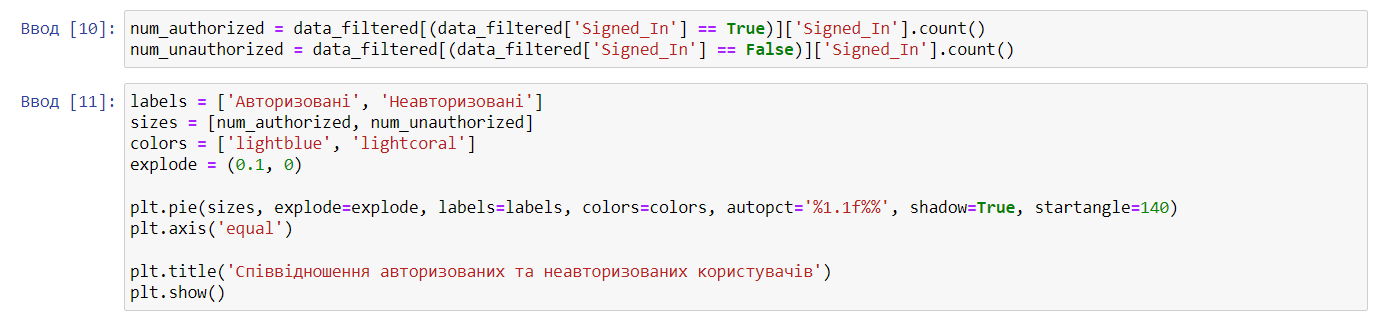


*рис 6 (виконання обрахунку середнього віку чоловіків і жінок та побудування діаграми)*



*рис 7 (діаграма середнього віку чоловіків та жінок)*

Потім було виконано обрахунки зареєстровані та незареєстровані користувачі. Для цього обраховуємо кількість таких користувачів та будуємо діаграму у вигляді кола, яке показує співвідношення зареєстрованих користувачів та незареєстрованих.

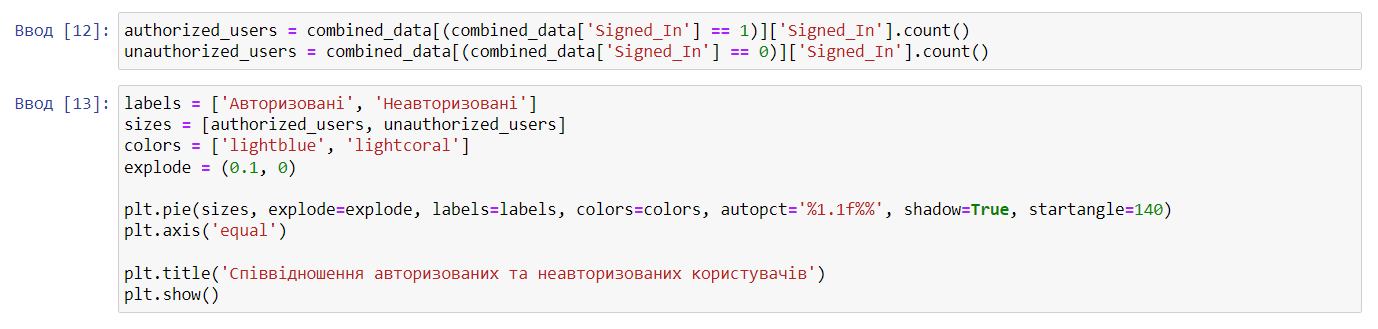


*рис 8 (виконання обрахунку співвідношення кількості авторизованих та неавторизованих користувачів)*



*рис 9 (співідношення авторизованих та неавторизованих користувачів)*

Як можна побачити, що всі користувачі авторизовані і тому спробував зробити це з користувачами, які не відносяться до жодної з вікових категорій.

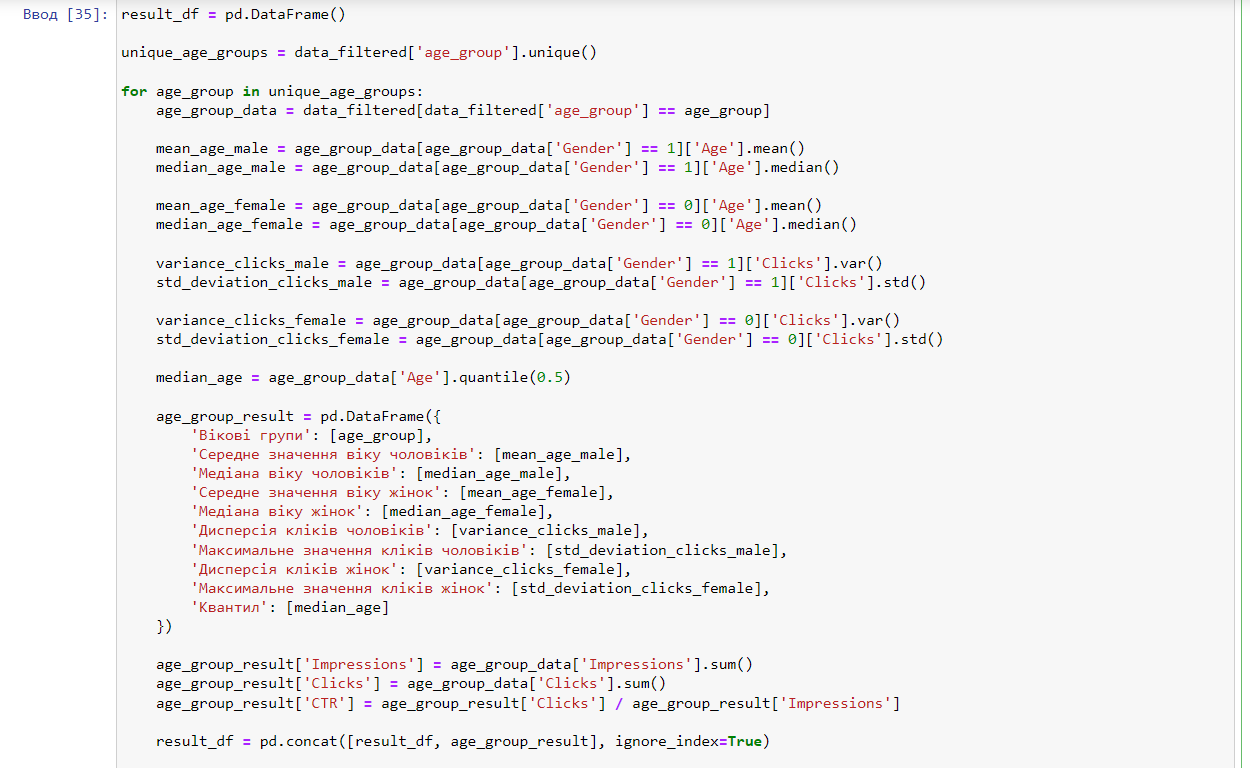


*рис 10 (виконання з урахуванням користувачів, які не відносяться до жодної з вікових категорій)*

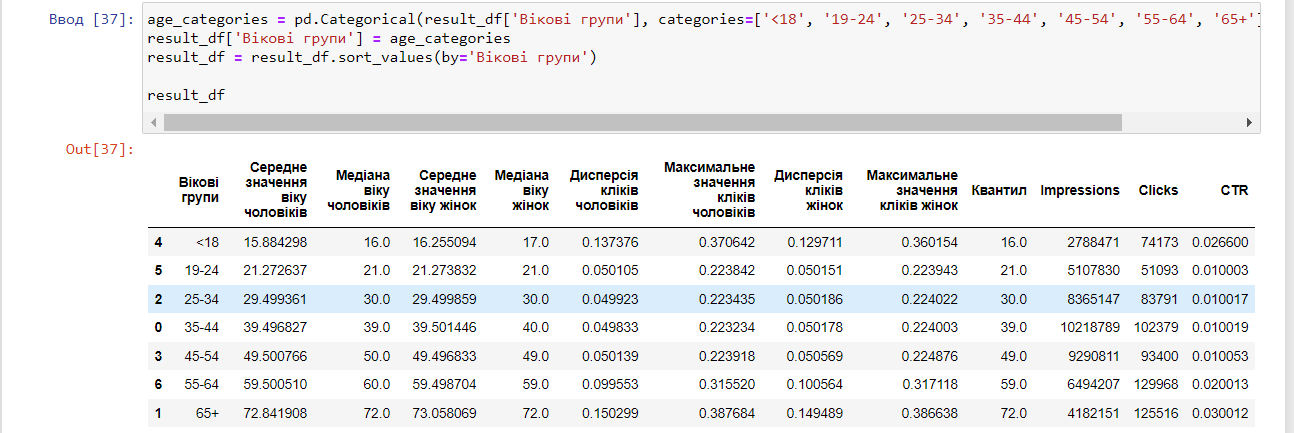


*рис 11 (діаграма з користувачами без вікової категорії)*

Далі обраховував середнє значення, медіану, дисперсію та максимальне значення, CTR та квантил. Середнє значення або середнє арифметичне - це сума всіх значень в масиві (або наборі даних), поділена на кількість цих значень. Це вимірюється у тих самих одиницях, що і вихідні дані. Наприклад, середній вік людей у групі. Медіана - це центральний елемент впорядкованого набору даних (або масиву). Щоб знайти медіану, спершу впорядкуйте дані за зростанням або спаданням, а потім знайдіть значення, яке розділяє набір даних на дві рівні частини. Якщо кількість значень парна, то медіана - середнє значення двох центральних значень. Дисперсія - це міра розсіювання значень у наборі даних відносно їх середнього значення. Вона обчислюється як середнє значення квадратів відхилень кожного значення від середнього значення. Максимальне значення - це найбільше число в наборі даних. Квантиль - це величина, яка розділяє набір даних на піднабори так, що певний відсоток даних знаходиться в кожному піднаборі. Наприклад, медіана - це 50-й квантиль, а перший квантиль (25-й перцентиль) розділяє набір даних так, що 25% значень менше або рівні йому. Виконання цих обрахунків та побудови нового датафрему продемонстровано на рис 12.



*рис 12 (обрахунки CTR, квантилу, середнього значення, медіану, дисперсію та максимальне значення за віковими категоріями)*



*рис 13 (вивід результату)*

З цих даних можна сказати що у жінок більше середні вік ніж у чоловіків, які є у цьому датафреймі, також більш сильне відхилення у середньому значені кліків там де більше було CTR.

**Висновок:** Отже, в цій лабораторній роботі було отриманно практичні навички у роботі з raw data, використовуючи пакети jupyter, pandas, seaborn.