Міністерство освіти і науки України

Київський національний університет імені Тараса Шевченка

Факультет інформаційних технологій

Кафедра прикладних інформаційних систем

Звіт

з лабораторної роботи №1

з дисципліни “Data Science та Big Data”

Виконала:

Костюк Д.О.

група ПП-42

Перевірив:

Київ 2023

**Лабораторна робота №1**

## Тема роботи: Агрегація, обробка пропусків та візуалізація даних пакетами Python.

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

**Завдання**

У датасеті знаходиться 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.

Завантажити файли з даними у папку проєкту з посилання:

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.

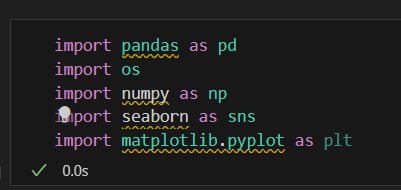
Результати статистичного дослідження подати у вигляді результуючого датафрейму (одного), дивлячись на який можна зрозуміти і порівнювати дані за віковими підкатегоріями.

Опишіть та інтерпретуйте будь-які закономірності, які знайдете.

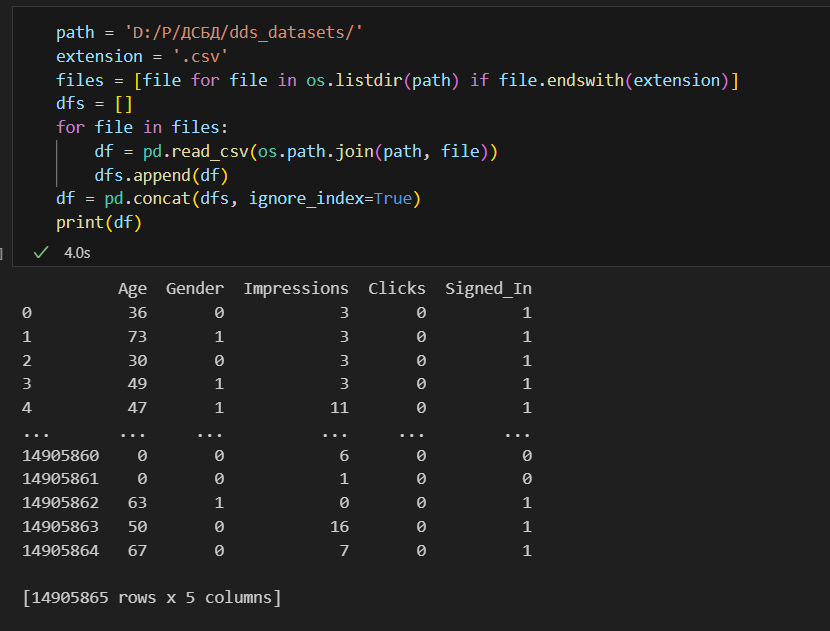
Завантажити файл ipynb з виконаними завданнями на git в окрему папку з відповідною назвою лабораторної роботи.

**Хід виконання**

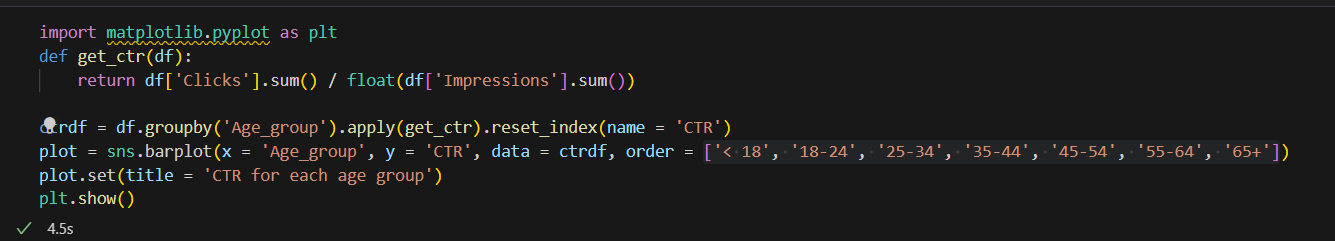
Імпортуємо усі використовувані пакети:



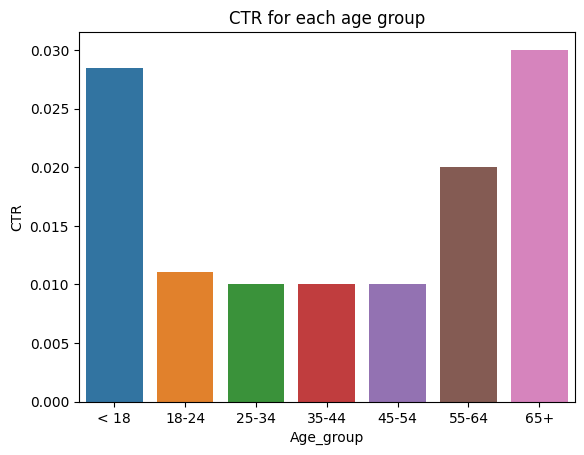
Створюємо датасет із усіх файлів:



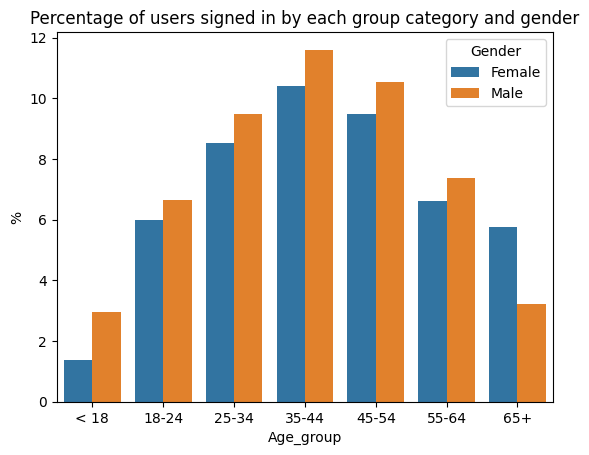
Обчислити CTR для кожної категорії:



Результат:



Обчислимо кількість зареєстрованих користувачів у кожній віковій групі по статі:

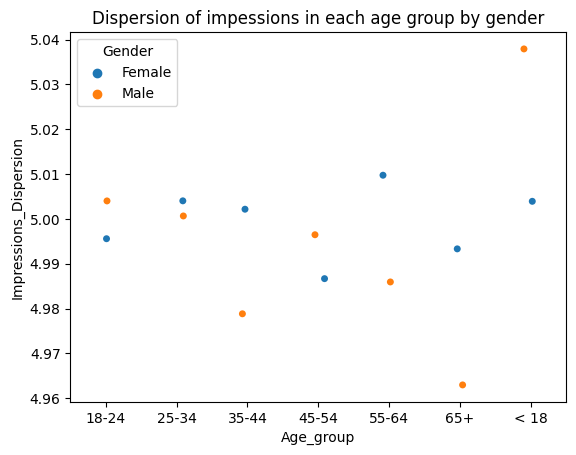


Отже, маємо, що найбільше зареєстрованих користувачів належать до категорії чоловіки 35-44.

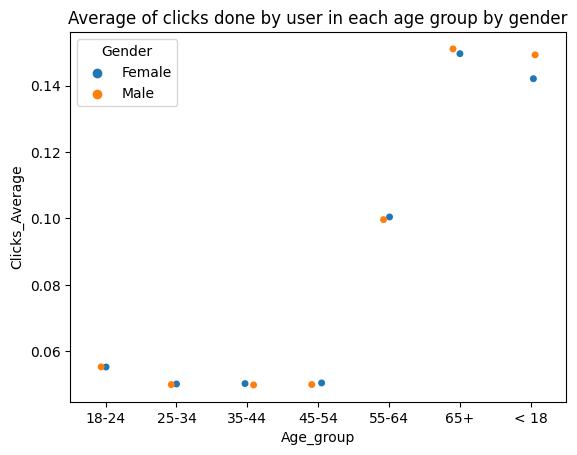
Найменше користувачів із категорії жінки < 18.

У всіх вікових категоріях домінують чоловіки, окрім категорії 65+.

Обчислимо дисперсію impressions у кожній віковій категорії:



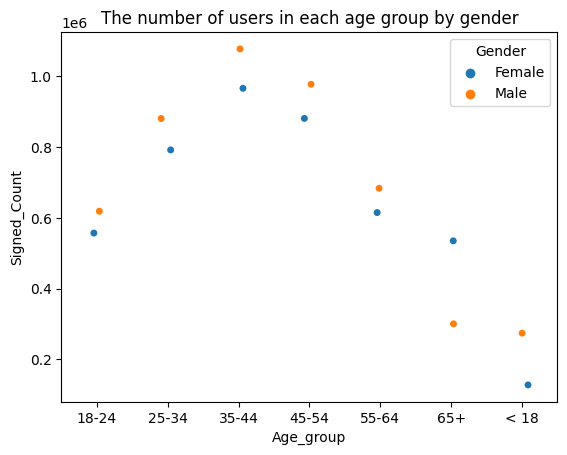
Обчислимо середню кількість натискань користувачів у кожній віковій групі:



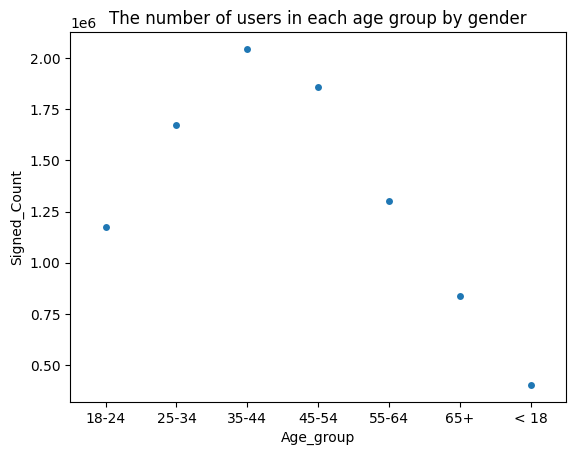
Найбільш активними користувачами (за кліками) є користувачі із категорій < 18 та 65+.

Найменш активними є категорії 25-54.

Обчислимо число користувачів всього:

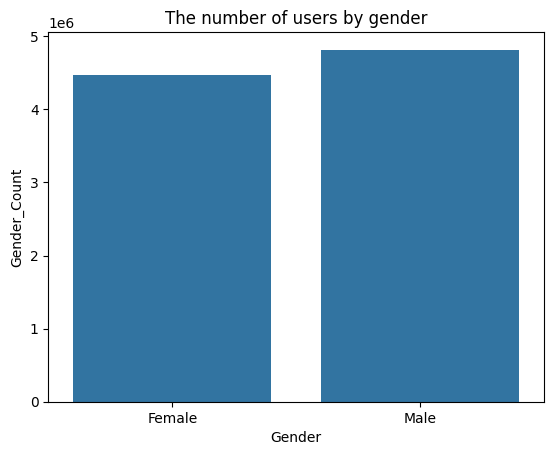


Якщо враховувати не лише зареєстрованих користувачів, а всіх, то залишаються ті самі тенденції:



Найбільше користувачів вікової категорії 35-44, найменше < 18.

Окрім цього сайт відвідує більша частка чоловіків, ніж жінок:



**Висновок:** у ході виконання даної лабораторної роботи я навчилась працювати із мовою програмування Python, пакетами jupyter, pandas, seaborn. За допомогою цих інструментів я навчилась працювати із даними, імпортувати їх, приводити до єдиного формату, виконувати запити до них, агрегувати та візуалізувати їх.