**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**імені ТАРАСА ШЕВЧЕНКА**

Факультет інформаційних технологій

**Кафедра прикладних інформаційних систем**

****

**Звіт**

до виконання лабораторної роботи №1

з дисципліни «Data Science та Big Data»

на тему:

## «Агрегація, обробка пропусків та візуалізація даних пакетами Python»

(*тема роботи*)

ВАРІАНТ №5

**Виконано:**

студ. групи ПП- 41, підгрупа 1

Горошко Андрія

**Перевірено:**

Білий Р.О.

**Київ – 2023**

**Мета роботи:**

Метою лабораторної роботи є отримання практичних навичок у роботі з raw data, використовуючи пакети jupyter, pandas, seaborn.

**Завдання:**

### Завантажити файли з даними у папку проекту 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 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

Зведемо і покращимо дані:

import numpy as np

import pandas as pd

import seaborn as sb

import matplotlib.pyplot as plt

arr = []

for i in range (1, 32):

    day = pd.read\_csv('/dataset/dds\_datasets/data\_nyt/nyt' + str(i) + '.csv')

    day['Day'] = i

    arr.append(day)

data = pd.concat(arr)

dataf = data.copy()

bounds = ['-1', '0', '18'  , '25', '35', '45', '55', '65', float('inf')]

labels = ['NoN', '<18', '18-24', '25-34', '35-44', '45-54', '55-64', '65+']

dataf['age\_group'] = pd.cut(dataf['Age'], bins=bounds, labels=labels)

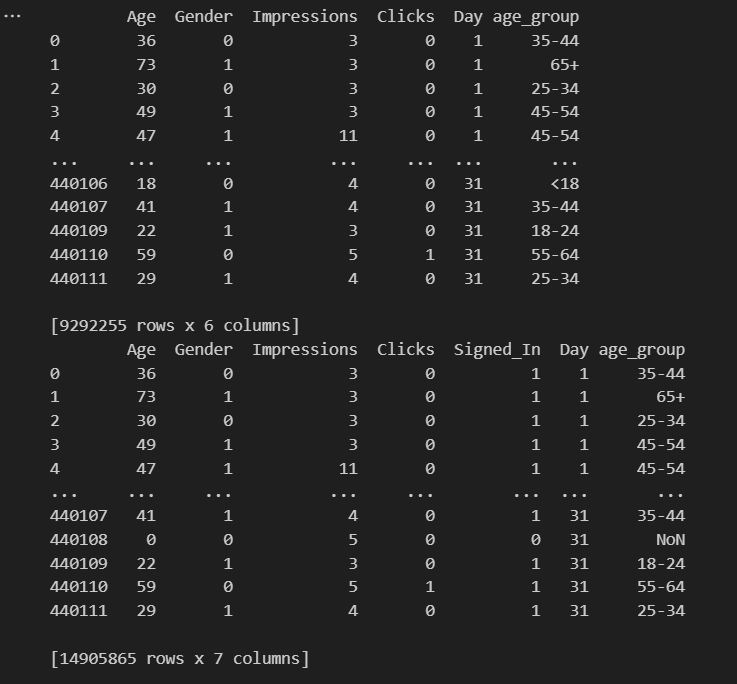
dataf\_signed = dataf[dataf['Signed\_In'] == 1]

dataf\_unsigned = dataf[dataf['Signed\_In'] == 0]

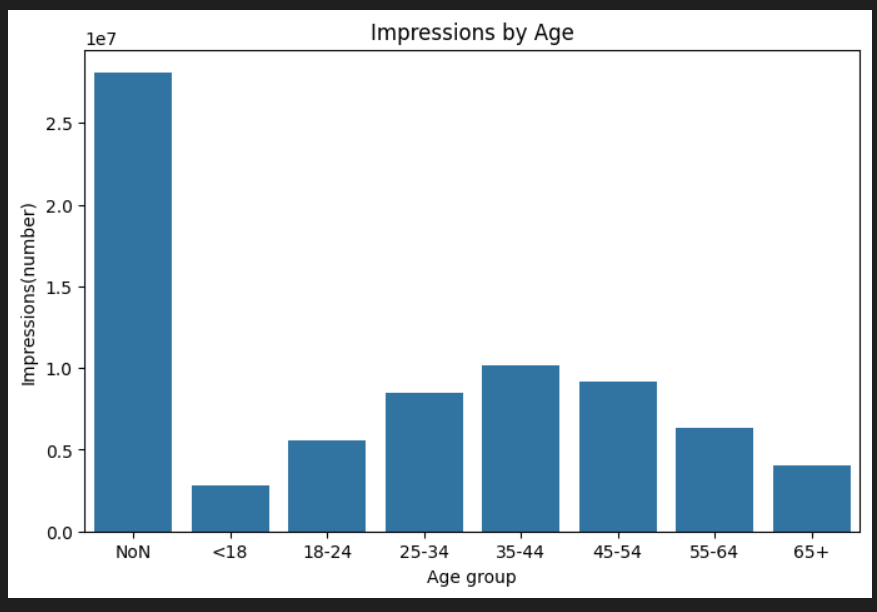
dataf\_signed = dataf\_signed.drop('Signed\_In', axis=1)

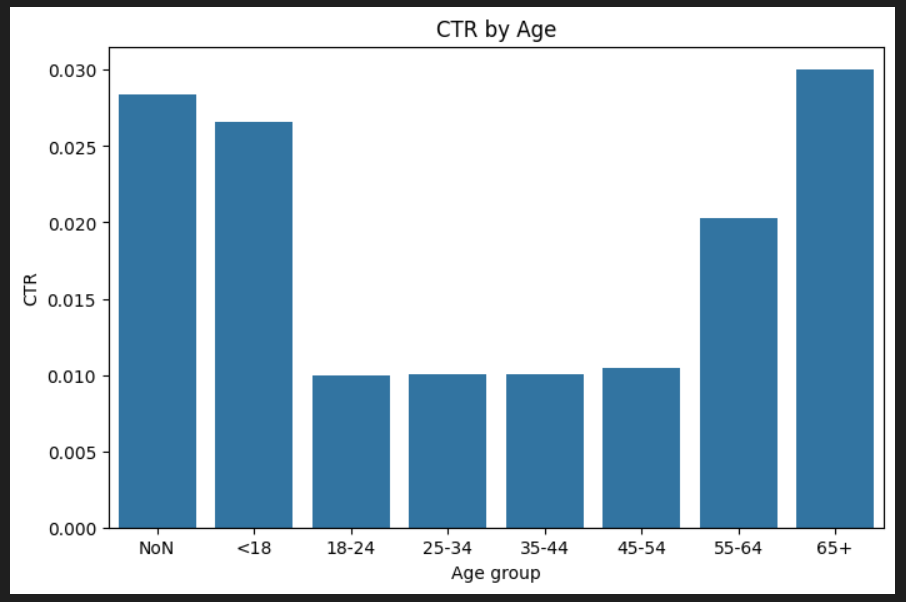
print(dataf\_signed)

print(dataf)



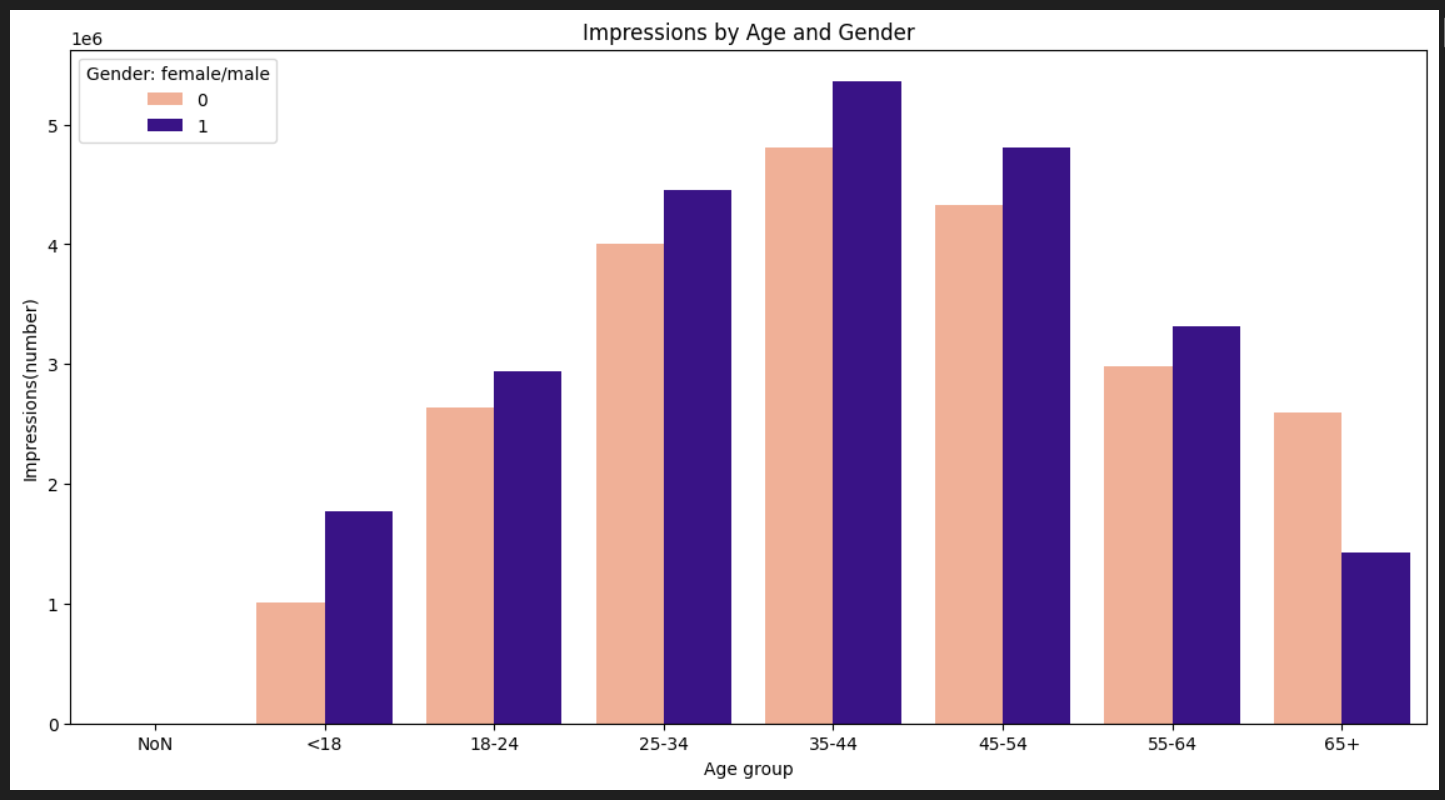
Проілюструємо перегляди по віковим групам

 Проілюструємо кліки до показів по віковим групам

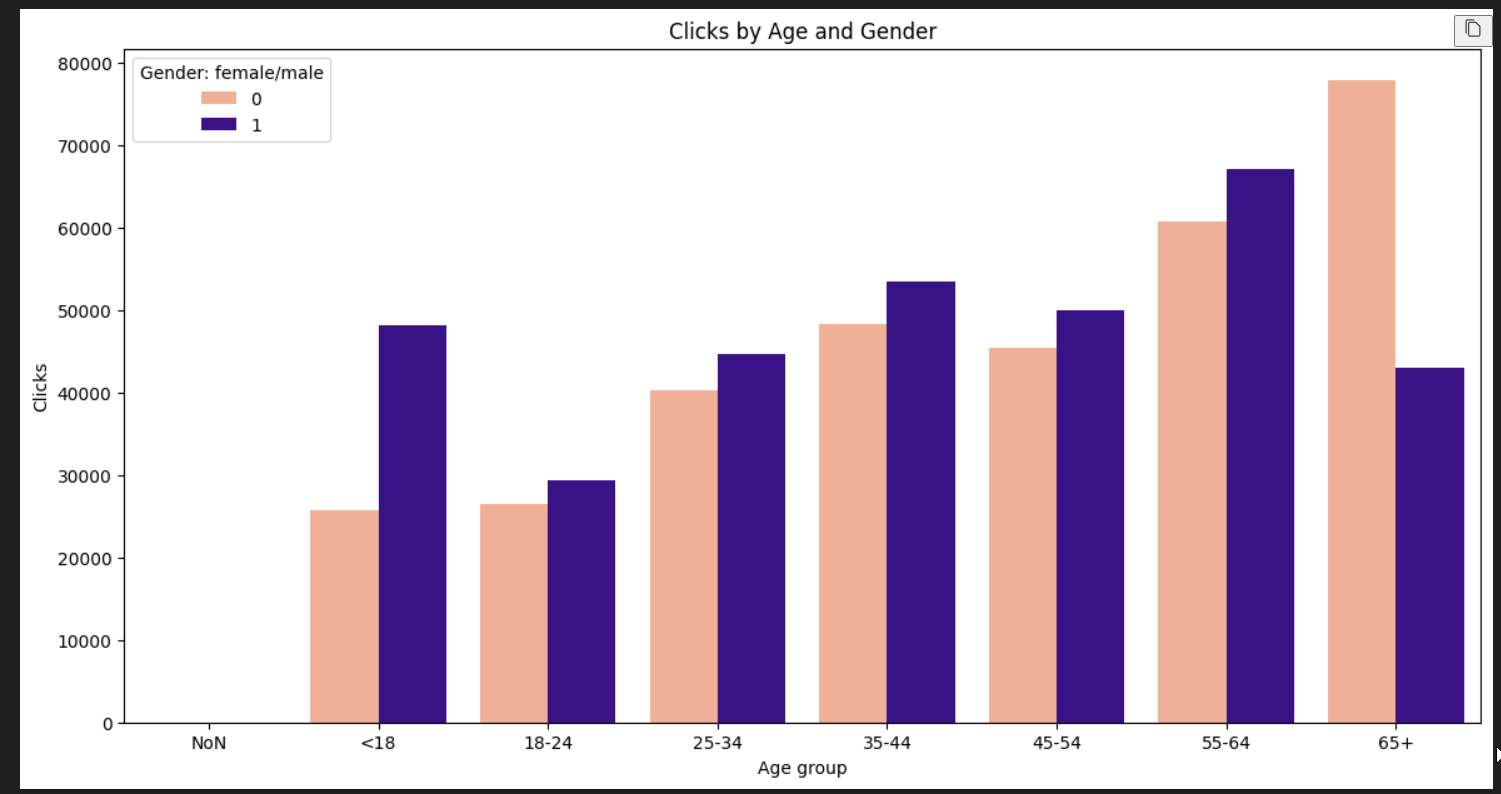


При аналізі отриманих рельтатів бачимо, що кількість переглядів досягає свого піку для середньої вікової групи (35-44), проте якщо звернути увагу на CTR, то стає очевидно, що для більшості вікових категорій не вистачає часу на споживання інформації, і на пік виходять групи до 18 та після 55, тобто вікові групи у яких кількість вільного часу значно більша через відсутність повноцінних робіт.

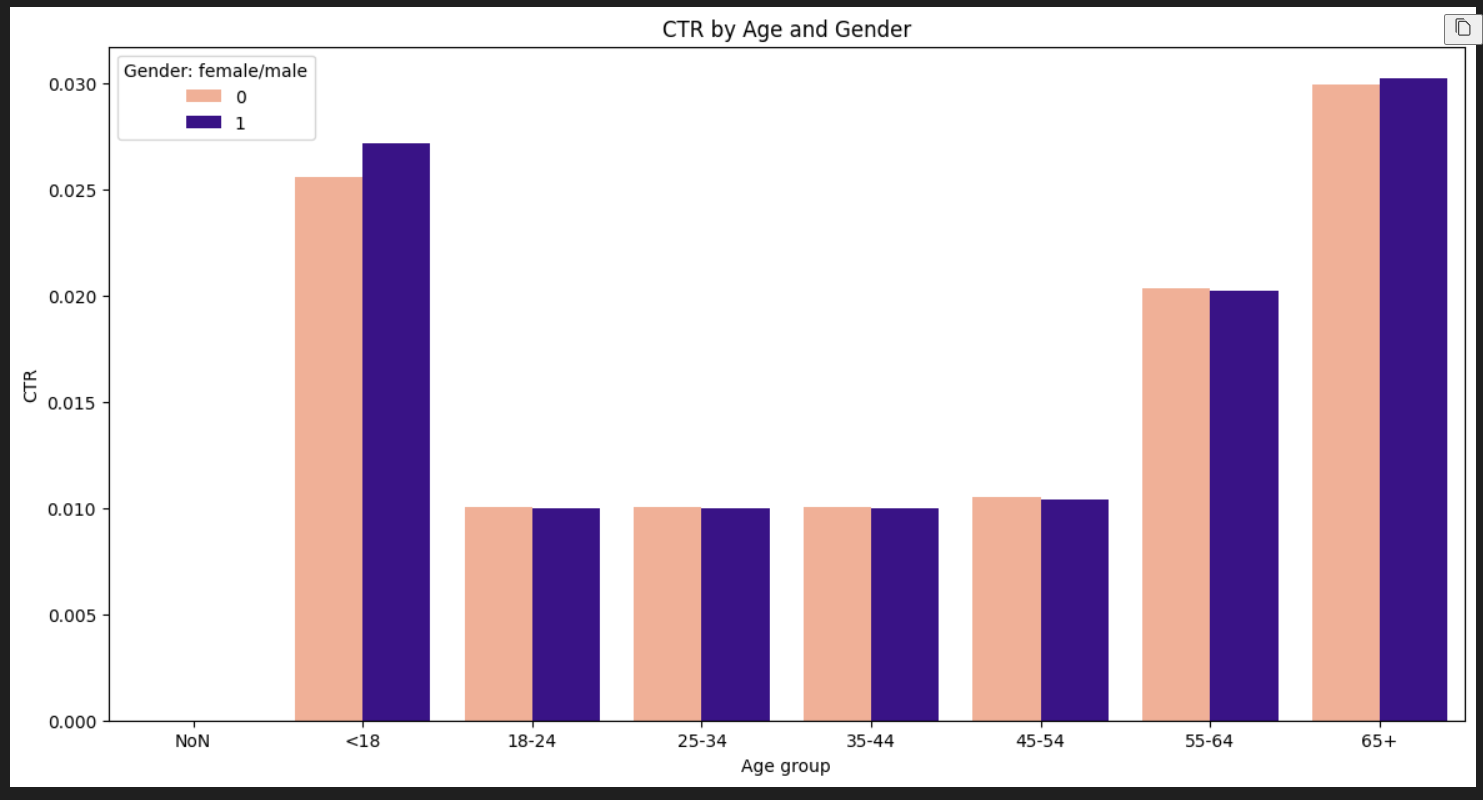
Проілюструємо перегляди по віковим групам поділенні на стать



Проілюструємо кліки по віковим групам поділенні на стать

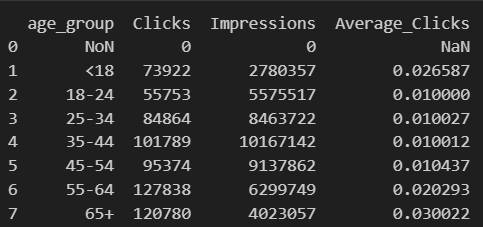


Проілюструємо кліки до показів по віковим групам поділенні на стать

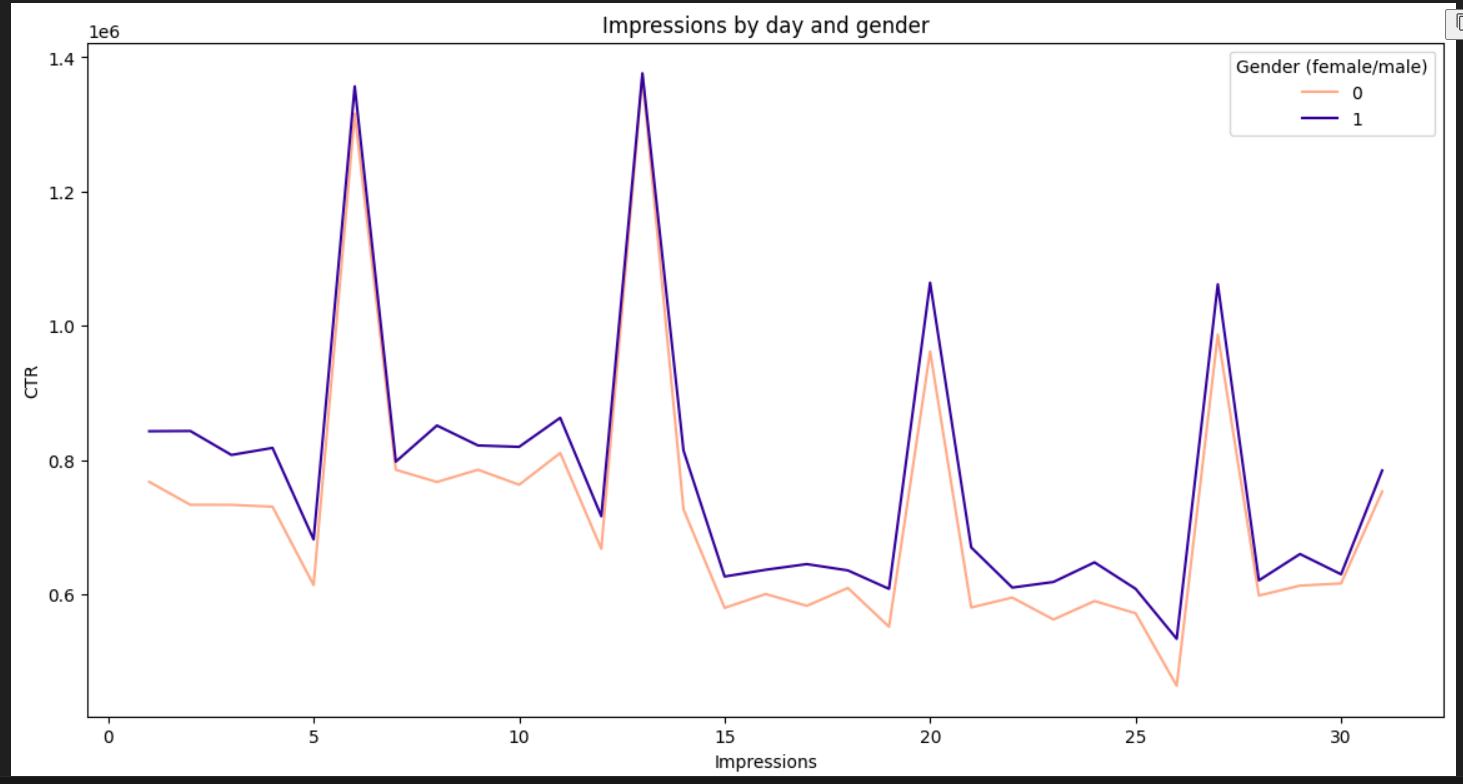


Після поділу попередніх даних на дві статі, бачимо, що суттьєвої інформації отримати нам не вдалося. Лише варто підмітити, що чоловіки до 18 років значно активніші в переглядах, після чого існує паритет аж до віку 65+, де жінки значно переважають чоловіків, що можна пов'язати з середньою тривалістю життя жінок та чоловіків

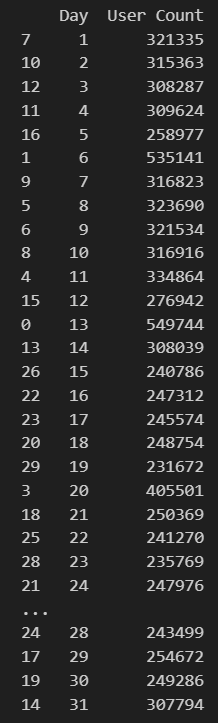
Загальна кількість кліків

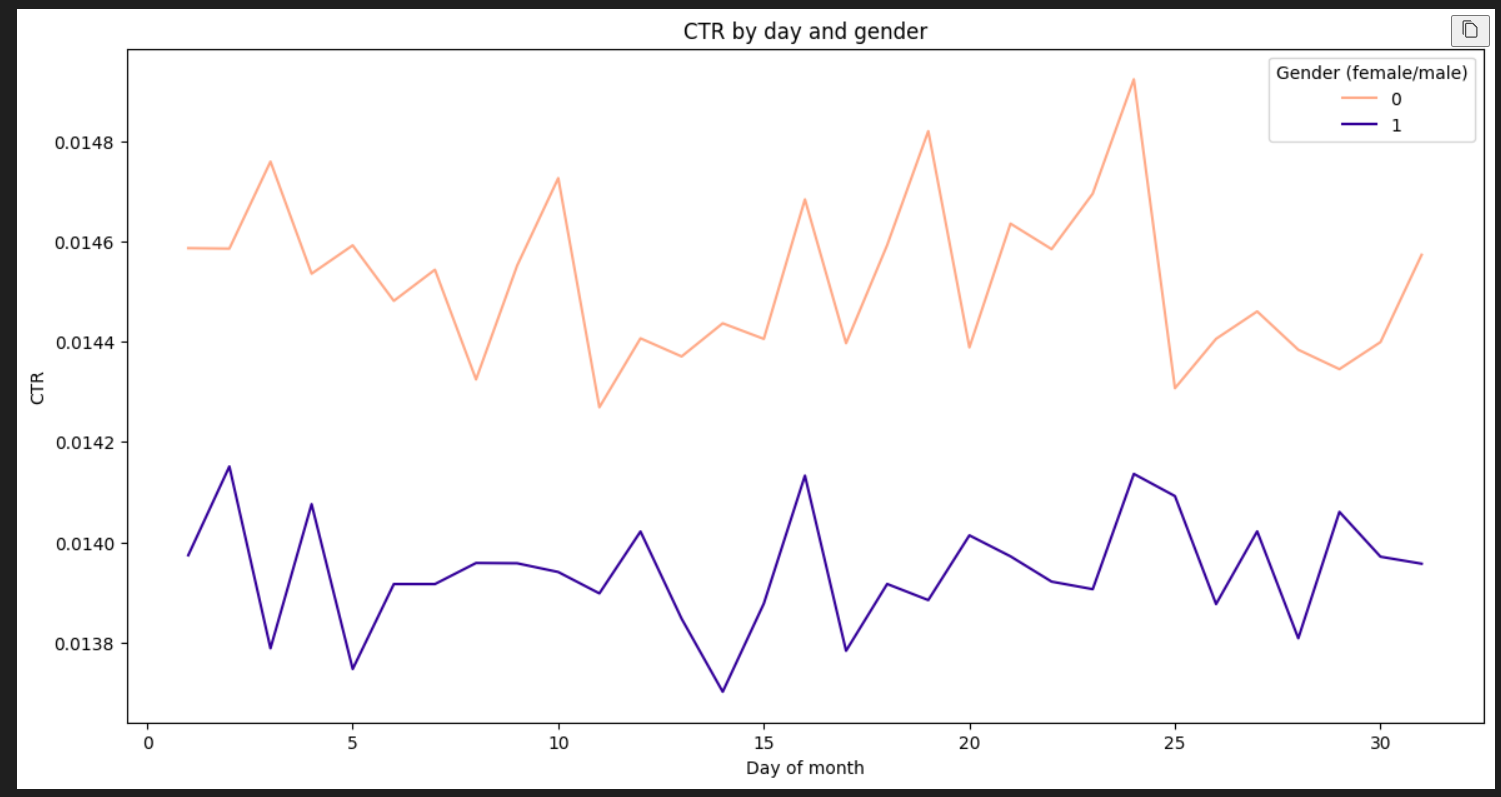


Далі я виконав сортування переглядів для чоловіків та жінок у загальному по дням, з чого видно, що у нашому тижні піки активності припадають на вихідні дні, адже періодичність у саме 7 днів

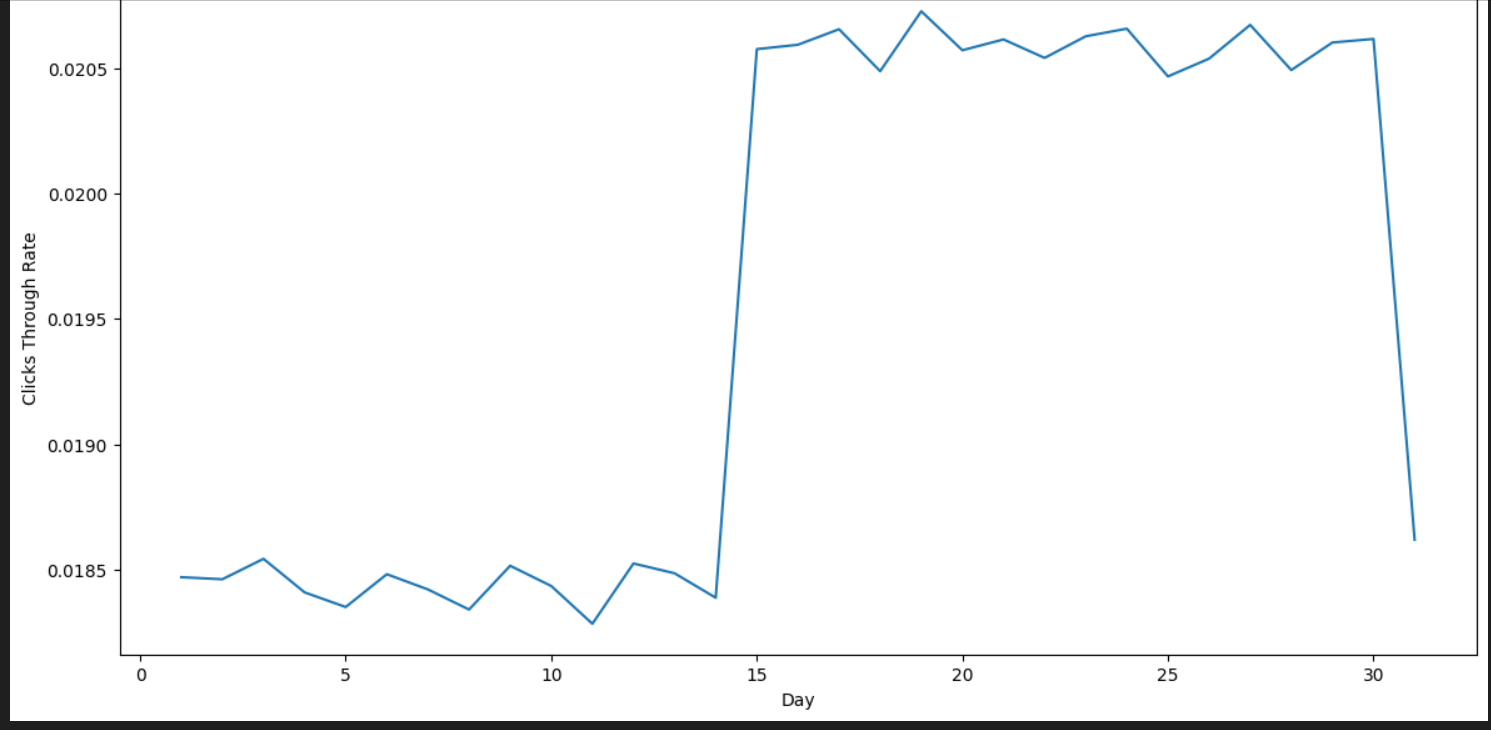


Алгебраїчне зображення графіку вище





З аналізу цього графіку видно, що 14 числа сталася якась значна подія, через яку і відбувся значний приріст аудиторії до сайту, який поступово до кінця місяця почав згасати



Висновок: Під час виконання роботи був отриманий досвід у роботі з raw data, використовуючи пакети jupyter, pandas, seaborn. А також покращено навички роботи з даними.