**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-41*

*спеціальності 122 «Комп’ютерні науки»*

*ОП «Прикладне програмування»*

Терновцева В.О.

*Викладач:*

*Білий Р.О.*

**Київ – 2023**

**Лабораторна робота №1**

Агрегація, обробка пропусків та візуалізація даних пакетами Python

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

**Завдання**

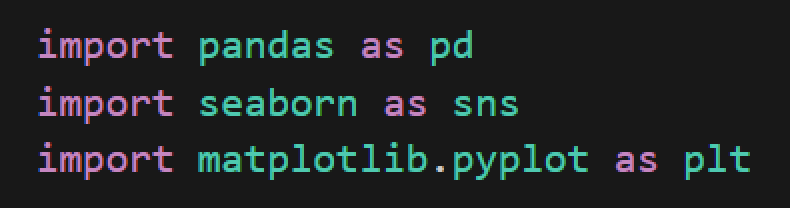
Завантажити файли з даними у папку проекту з посилання.

* 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.

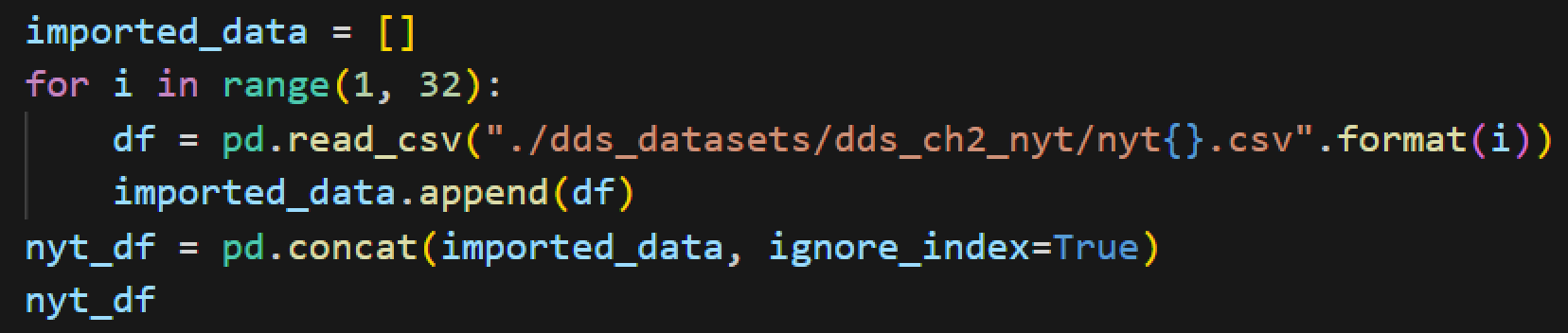
Завантажити файл .ipynb з виконаними завданнями на git в окрему папку з відповідною назвою лабораторної роботи

**Хід виконання**

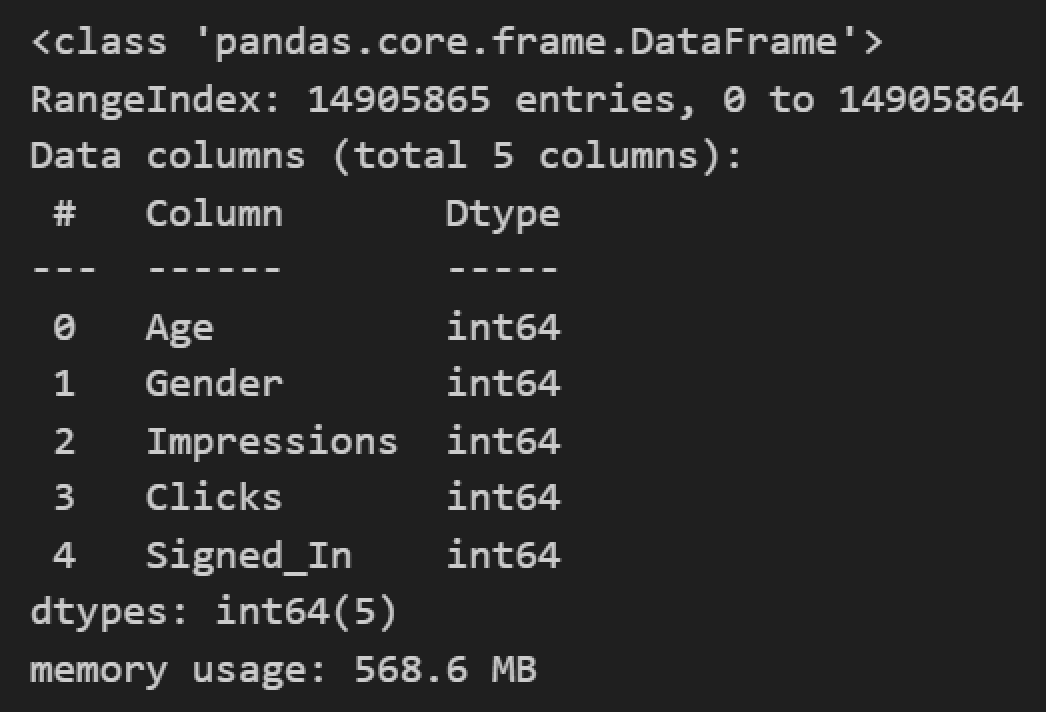
Імпортуємо потрібні бібліотеки.



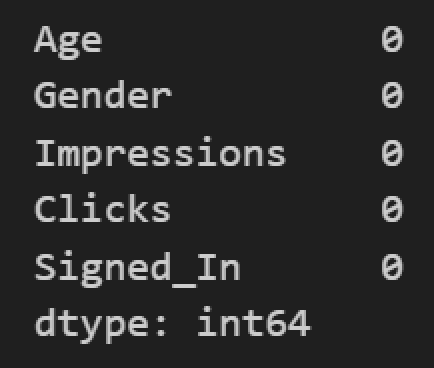
Завантажуємо усі набори даних з файлів та об’єднуємо у спільних DataFrame з назвою nyt\_df.



Переглядаємо типи даних у DataFrame за допомогою методу info(). Як можна побачити, усі дані описані за допомогою цілих чисел.



Перевіряємо DataFrame на наявність порожніх значень. Для кожного стовпчика ця кількість дорівнює нулю.



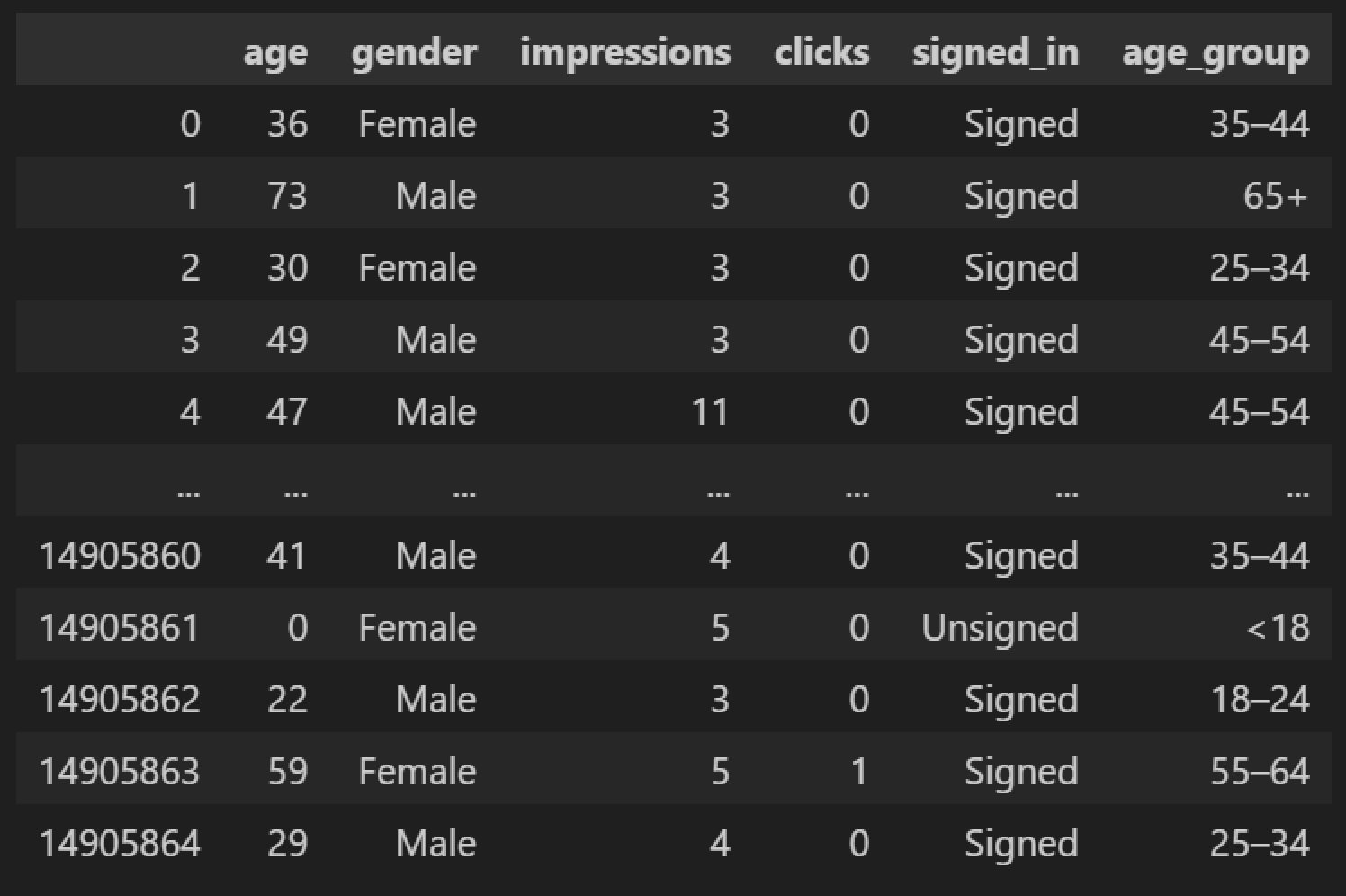
Для зручності звертання до фрейму перейменуємо назви стовпчиків.



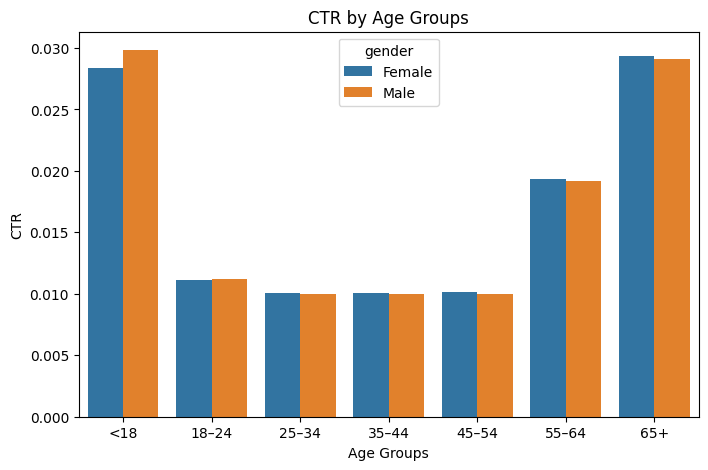
У стовпчиках gender та signed\_in замінимо нулі та одиниці на «Female», «Male» та «Unsigned», «Signed» відповідно. Таким чином буде легше сприймати ці значення.



Створюємо стовпчик age\_group для агрегації користувачів за наступними віковими категоріями: <18, 18-24, 25-34, 35-44, 45-54, 55-64 та 65+.

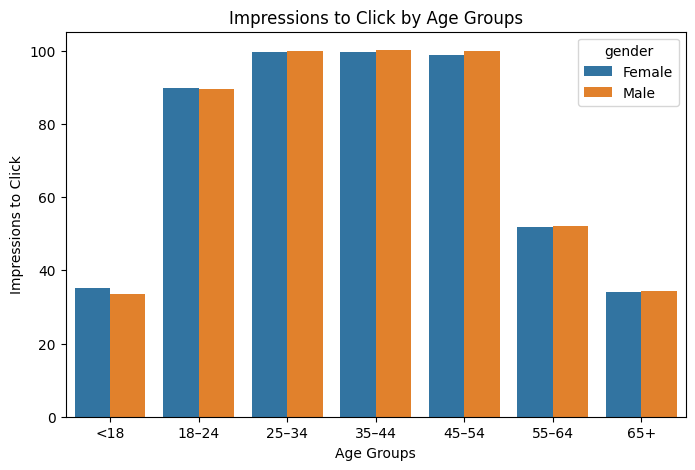


Для кожної вікової категорії (з розподілом на гендери) підраховуємо суму показів та суму переходів. Маючи ці дані, можна підрахувати для кожної категорії користувачів її CTR (Click-Through Rate), відобразивши відношення числа кліків до числа показів. Візуалізуємо отриманий результат.



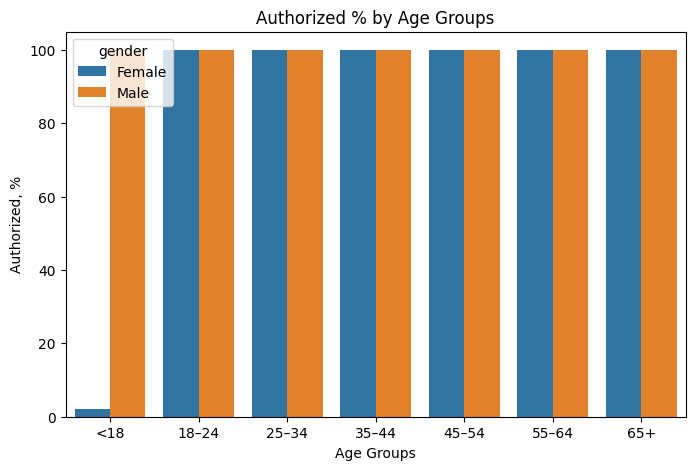
З графіку видно, що найчастіше клікають на оголошення люди до 18 та від 55 років. Користувачі в усіх інших вікових категоріях мають приблизно однаковий, найнижчий показник.

Далі переглянемо, скільки показів припадало на кожний перехід.



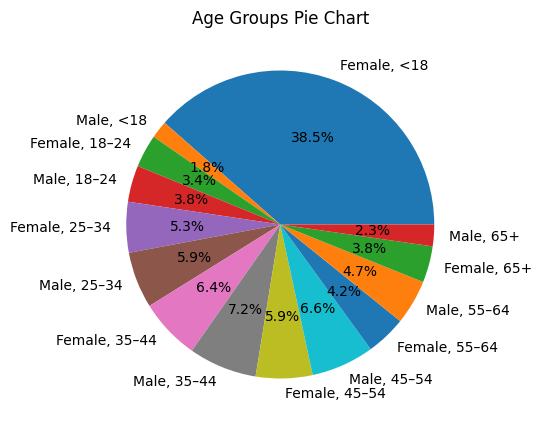
Бачимо, що для людей від 18 до 54 років потрібно, грубо кажучи, близько 100 показів оголошення для одного кліку.

Розглянемо відсоток авторизованих користувачів для кожної вікової категорії.



З отриманого результату можна зробити висновок, що, скоріше за все, набори даних є неповними, оскільки серед неавторизованих на сайті користувачів є лише неповнолітні жінки. Така статистика навряд чи відображає дійсність.

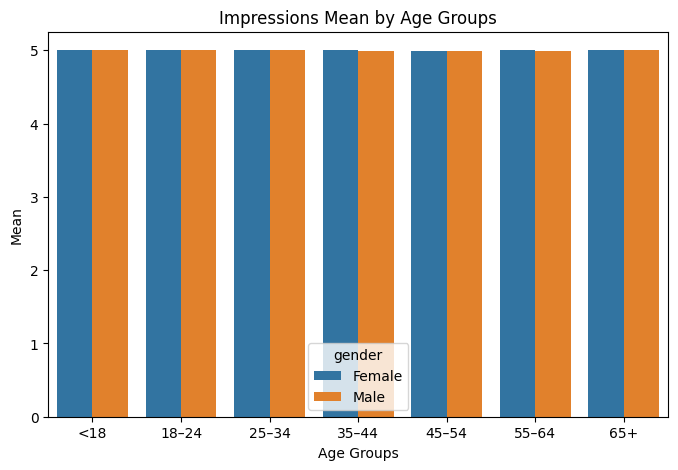
Тепер варто відмітити, який відсоток від усіх користувачів становить кожна вікова категорія.



Бачимо, що жінки віком до 18 років становлять майже 40% від усіх користувачів сайту. Такий великий відсоток може випливати з попереднього пункту – з відсутності у наборах даних інформації про неавторизованих повнолітніх користувачів.

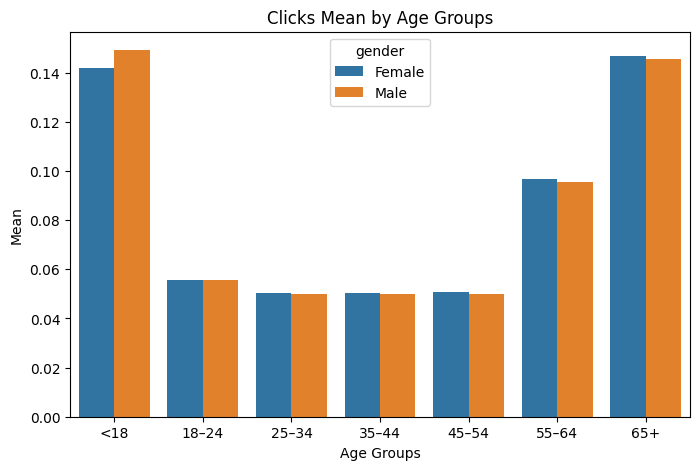
Якщо ж оцінювати інші вікові категорії, то можна сказати, що авдиторією сайту є переважно люди від 25 до 55 років.

Візуалізуємо середнє значення показів для кожної з категорій. Це дасть можливість відслідкувати, скільки в середньому показів бачить користувач.



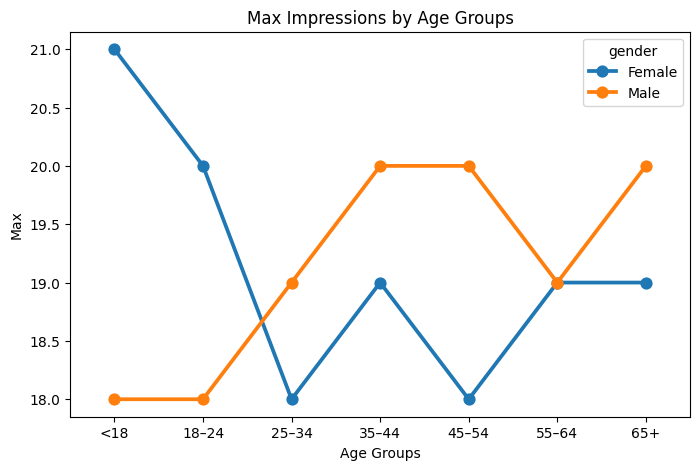
З графіку можна зробити висновок, що для абсолютно усіх користувачів, незалежно від віку та статі, середня кількість показів дорівнює 5.

Аналогічно проаналізуємо середню кількість переходів серед усіх вікових категорій.



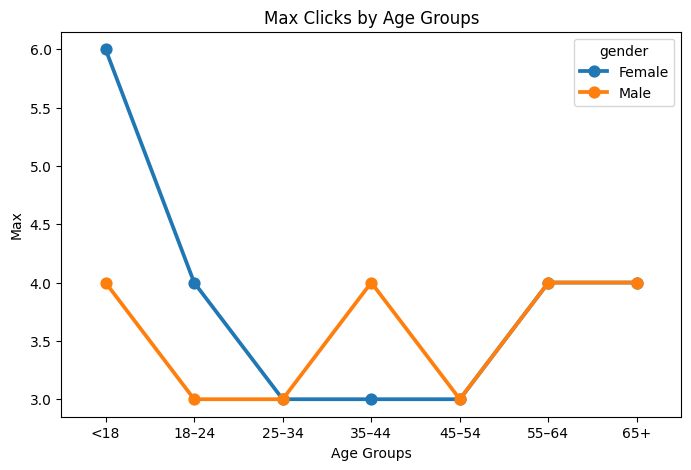
Можна зробити висновок, що користувачі в основному не переходять по оголошеннях, оскільки середнє значення близьке до нуля. Це також може бути зумовлене тим, що кожний користувач у середньому бачить близько 5 оголошень, чого, як було зазначено у графіках вище, не вистачає для зацікавлення людини.

Наступним кроком розглянемо, наскільки високих значень досягає кількість показів на людину.



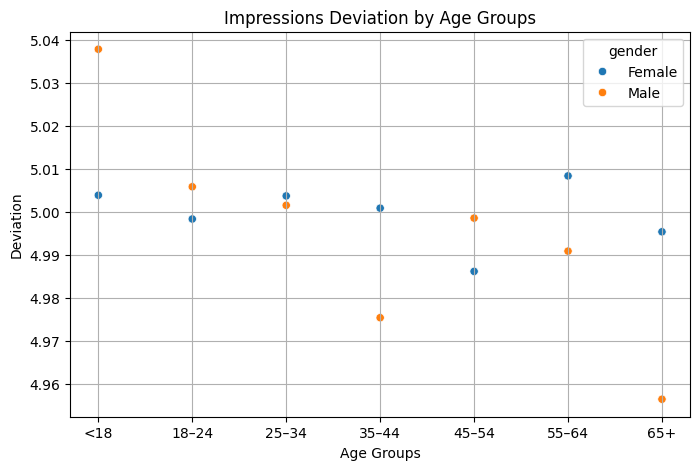
Таким чином, для чоловічої авдиторії максимальна кількість показів збільшується з віком людини, у той час як для жіночої навпаки – зменшується.

Так само проаналізуємо і кількість кліків.



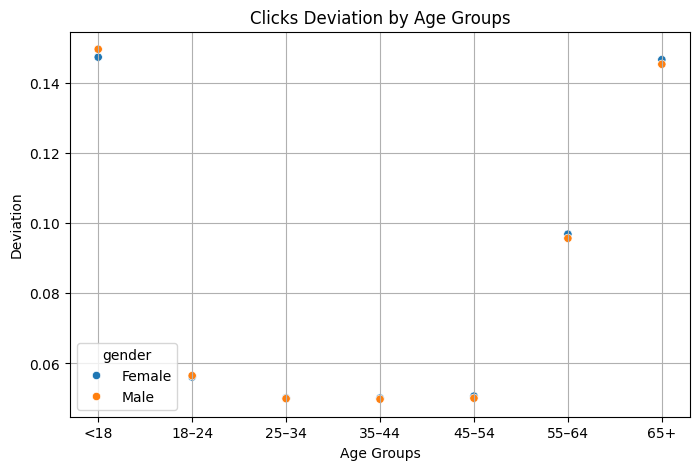
Максимальне значення переходів по оголошенням для жіночої авдиторії спочатку зменшується з віком, а потім збільшується на 1. Для чоловіків це значення тримається на позначках у 3-4 кліки.

За допомогою дисперсії кількості показів перевіримо, наскільки однорідною є авдиторія щодо цього показника.



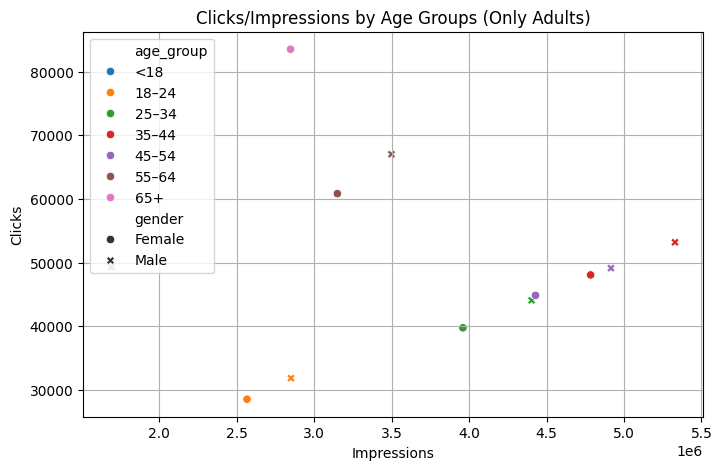
Оскільки значення для усіх вікових категорій є наближеними, робимо висновок, що дисперсія є низькою. Малий розкид значень свідчить про те, що люди приблизно однаково реагують на оголошення, незалежно від віку та гендеру.

Аналогічно перевіримо дисперсію для переходів по оголошенням.



Тут вже присутній певний розкид даних за віковою категорією. Це означає, що реакція людей віком від 18 до 54 років може відрізнятись від інших категорій. Дисперсія за гендером є незначною, тобто за цим критерієм реакція та кількість переходів є однорідними.

Заключним етапом буде побудова графіку залежності показів та переходів з розподілом на вікові категорії та гендери.



Як не дивно, між кількістю показів та кількістю переходів є лінійна залежність. Варто відмітити, що для людей віком від 65 років кількість кліків є найвищою при нижчій кількості показів. Гендер також має певний вплив, для чоловіків віком до 64 років кількість показів є більшою, ніж у жінок. Кластер справа на графіку може свідчити про те, що для користувачів у віці від 25 до 54 років покази спостерігаються найчастіше.

**Висновок**: під час виконання цієї лабораторної роботи я отримав практичні навички у роботі з raw data, використовуючи пакети jupyter, pandas, seaborn. Провів агрегацію, обробку пропусків, аналіз та візуалізацію даних. В процесі виконання лабораторної роботи було досліджено, як різні вікові та гендерні категорії людей взаємодіють з показами та переходами. Також отримані графіки вказують на існування загальної тенденції, що більша кількість показів часто супроводжується більшою кількістю переходів. Усі ці дослідження можуть бути важливою інформацією для визначення ефективності рекламних кампаній.