**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»

*Студентки 4 курсу*
*групи ПП-41*
*спеціальності 122 «Комп'ютерні науки»*
*ОП «Прикладне програмування»*
*Штонь Софії Максимівни*

*Викладач:*
Білий Р. О.

**Київ – 2023**

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

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

### *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 для виконання*

- 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.

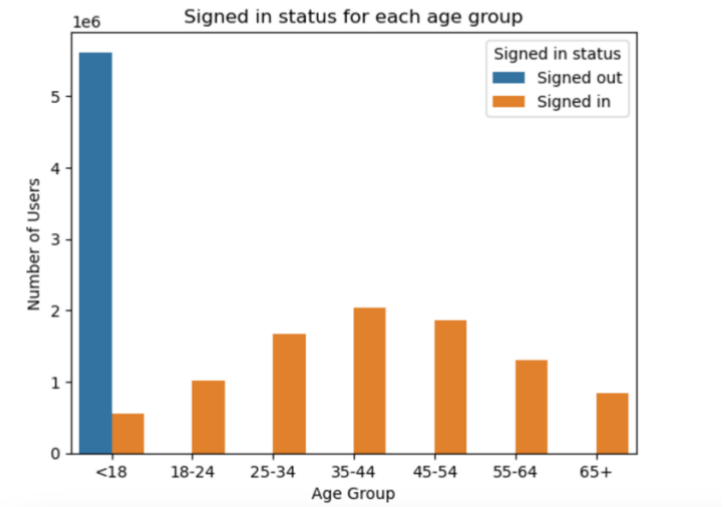
## *Хід роботи*

По-перше, імпортуємо необхідні пакети та створимо DataFrame з наявних датасетів:

```
In [1]: %matplotlib inline

In [2]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns

In [3]: #df = pd.read_csv("dataset/dds_datasets/dds_ch2_nyt/nyt1.csv")

        df = pd.concat(
            map(pd.read_csv, ['dataset/dds_datasets/dds_ch2_nyt/nyt1.csv', 'data
             , 'dataset/dds_datasets/dds_ch2_nyt/nyt6.csv', 'dataset/dds_dataset
              , 'dataset/dds_datasets/dds_ch2_nyt/nyt12.csv', 'dataset/dds_datas
              , 'dataset/dds_datasets/dds_ch2_nyt/nyt18.csv'   , 'dataset/dds_da
              , 'dataset/dds_datasets/dds_ch2_nyt/nyt24.csv'   , 'dataset/dds_da
              , 'dataset/dds_datasets/dds_ch2_nyt/nyt30.csv'   , 'dataset/dds_da
```

Створимо агрегацію користувачів за віком (<18, 18–24, 25–34, 35–44, 45–54, 55–64 та 65+), та відобразимо статус signed in/ signed out за віком користувачів на графіку:

```
In [4]: bins = [0, 18, 24, 34, 44, 54, 64, float('inf')]
        labels = ['<18', '18-24', '25-34', '35-44', '45-54', '55-64', '65+']

        df['age_group'] = pd.cut(df['Age'], bins=bins, labels=labels, include_lo

In [5]: sns.countplot(x='age_group', hue='Signed_In', data=df)
        plt.title('Signed in status for each age group')
        plt.xlabel('Age Group')
        plt.ylabel('Number of Users')
        plt.legend(title='Signed in status', labels=['Signed out', 'Signed in'])
        plt.show()
```



Бачимо, що щось не так – серед всі signed out користувачі знаходяться у віковій категорії <18. З датасету видно, що вік signed out користувачів автоматично 0, адже ця інформація невідома для цих користувачів.

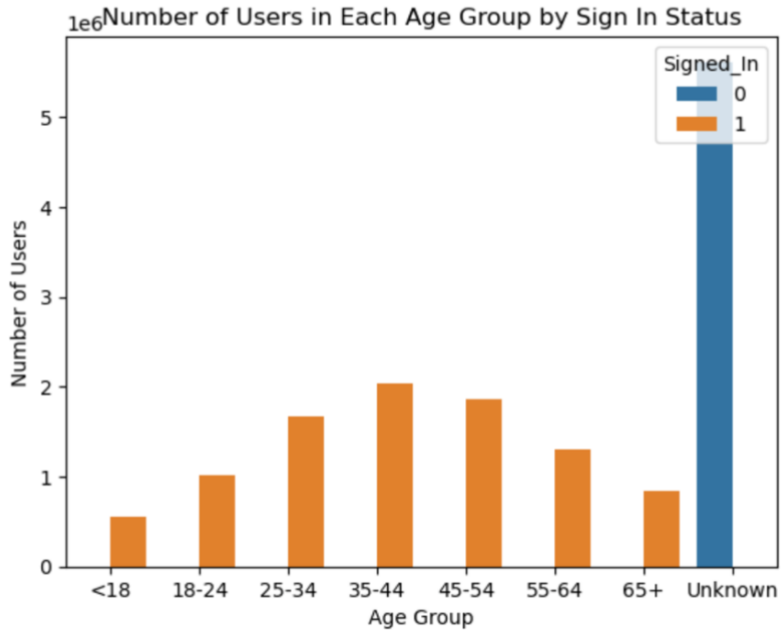
```
In [6]: signed_out_counts = df[df['Signed_In'] == False].groupby('age_group').si

        print("Number of signed out users for each age group:")
        print(signed_out_counts)
```

```
Number of signed out users for each age group:
age_group
<18      5613610
18-24          0
25-34          0
35-44          0
45-54          0
55-64          0
65+            0
dtype: int64
```

Проблема: Користувачі, що не увійшли в акаунт, мають вік 0 та додаються в age_group <18. Оскільки їх вік невідомий, вони повинні розглядатися як окрема категорія користувачів.

Як вирішення проблеми, створюємо нову вікову категорію “Unknown”, щоб відобразити користувачів, вік яких невідомий:

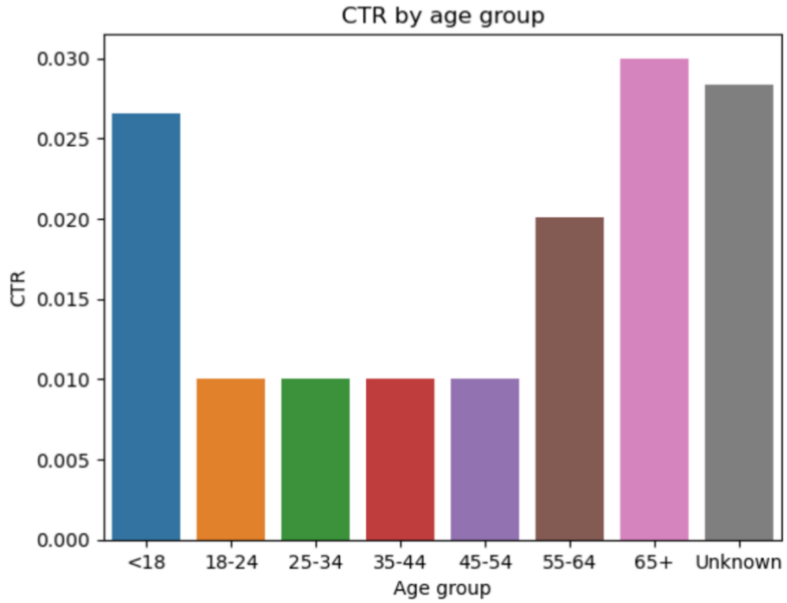
```
In [7]: # adding a new 'Unknown' category for signed out users with age=0
        df['age_group'] = pd.cut(df['Age'], bins=bins, labels=labels, include_lo
        df['age_group'] = df['age_group'].cat.add_categories('Unknown')
        df.loc[(df['Age'] == 0) & (df['Signed_In'] == 0), 'age_group'] = 'Unknow

        sns.countplot(x='age_group', hue='Signed_In', data=df)
        plt.title('Number of Users in Each Age Group by Sign In Status')
        plt.xlabel('Age Group')
        plt.ylabel('Number of Users')
        plt.show()
```




Тепер інформація відображається коректно.

Відобразимо clickthrough rate (CTR) по віковій категорії:

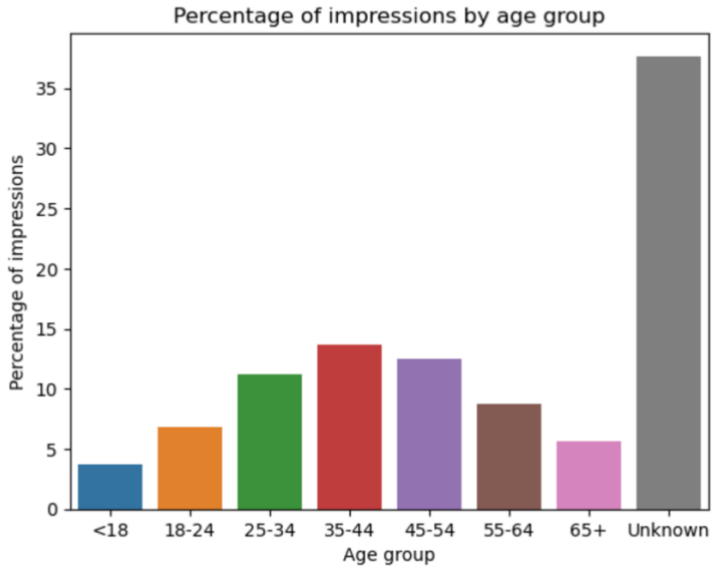
```
In [8]: # adding a CTR column
        df['CTR'] = df['Clicks'] / df['Impressions']

        sns.barplot(x='age_group', y='CTR', data=df, label='CTR', ci=None)
        plt.xlabel('Age group')
        plt.title('CTR by age group')
        plt.show()
```




Бачимо, що найвищі показники у категорії <18, 65+, а також у користувачів, вік яких невідомий.

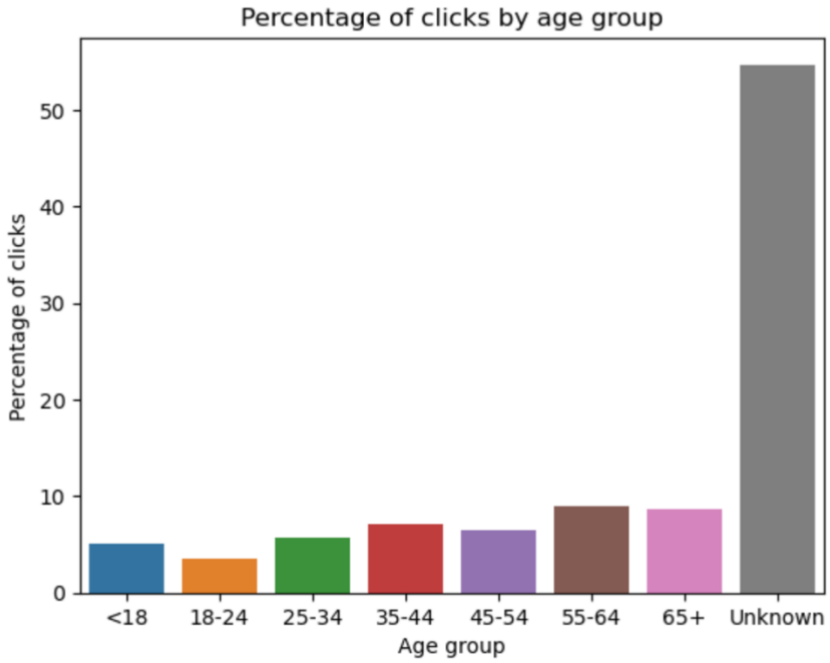
Далі, відобразимо відсоток показів за віковою категорією:

```
In [9]: impressions_by_age = (df.groupby('age_group')['Impressions']
                              .sum() / df['Impressions'].sum() * 100).reset

        sns.barplot(x='age_group', y='Impressions', data=impressions_by_age, ci=
        plt.ylabel('Percentage of impressions')
        plt.xlabel('Age group')
        plt.title('Percentage of impressions by age group')
        plt.show()
```

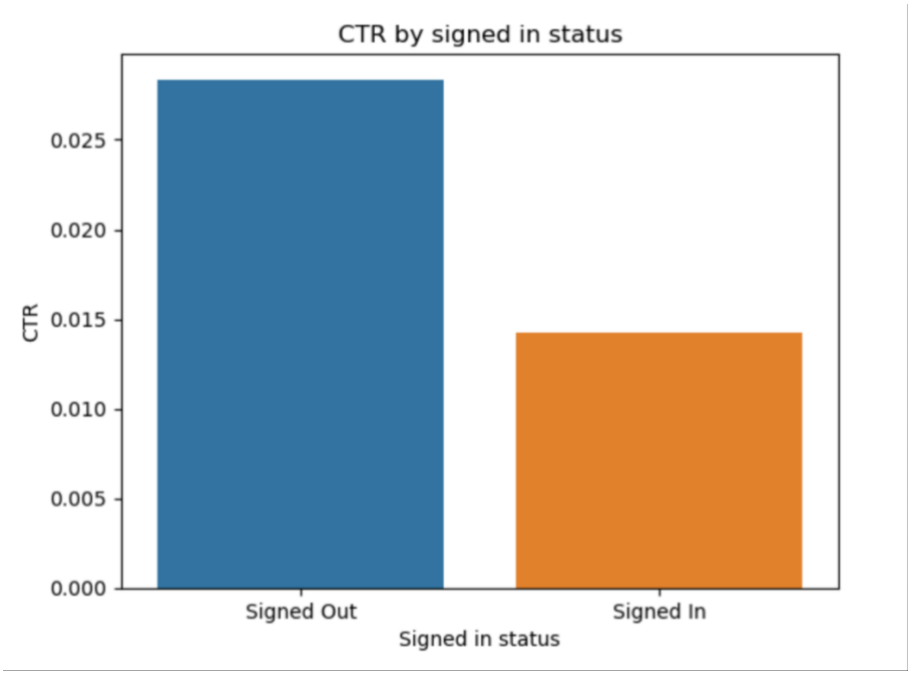



Бачимо, що відсоток показів набагато вищий у незареєстрованих користувачів. Відобразимо також відсоток кліків:

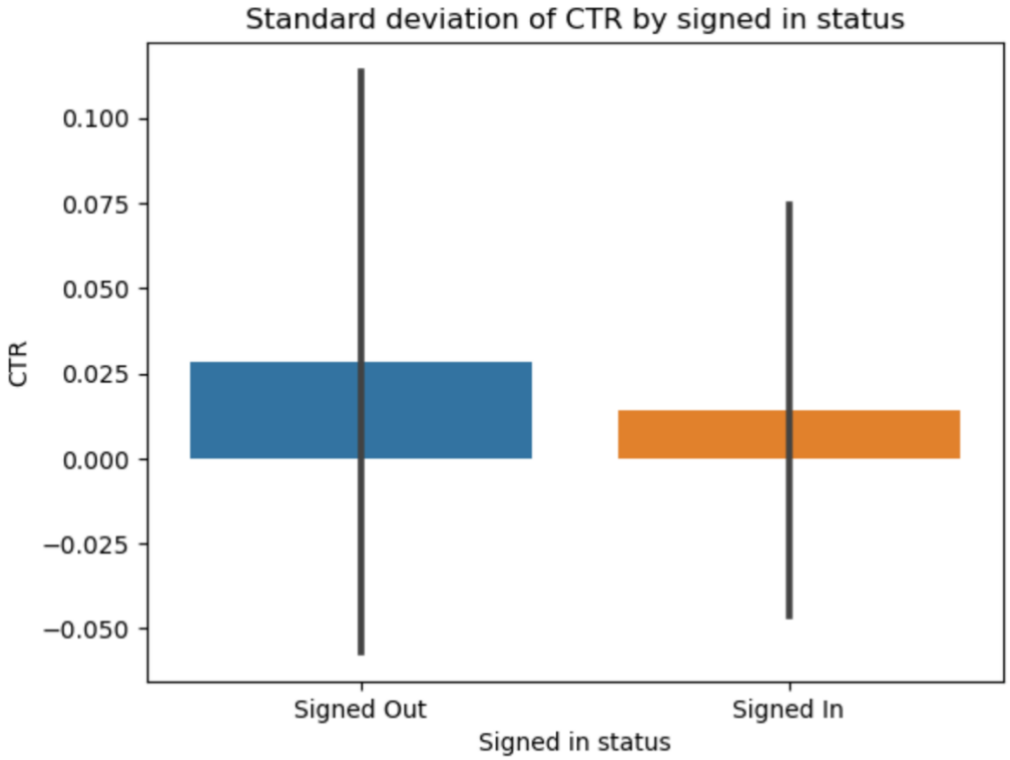
```
In [10]: clicks_by_age = (df.groupby('age_group')['Clicks']
                            .sum() / df['Clicks'].sum() * 100).reset_inde

sns.barplot(x='age_group', y='Clicks', data=clicks_by_age, ci=None)
plt.ylabel('Percentage of clicks')
plt.xlabel('Age group')
plt.title('Percentage of clicks by age group')
plt.show()
```



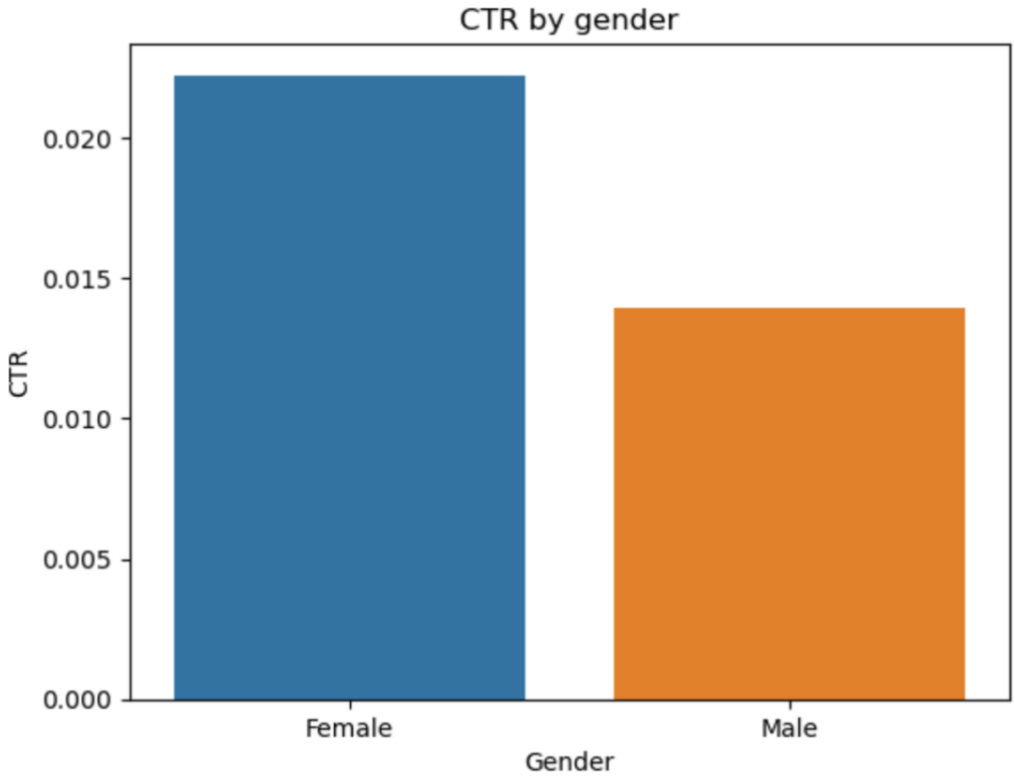

Бачимо, що відсоток кліків також набагато вищий у незареєстрованих користувачів. Далі, проілюструємо CTR за статусом signed out / signed in:



Видно, що CTR незареєстрованих користувачів набагато вищий, ніж у зареєстрованих. Також, відобразимо стандартне відхилення для цих показників (чорна лінія):




Тепер відобразимо показних CTR за статтю:



Той факт, що CTR значно вищий у жінок, викликає сумніви у тому, чи коректно відображені дані. Перевіримо кількість незареєстрованих жінок та чоловіків, та бачимо, що в датасеті є помилка:

```
In [14]: signed_out_users = df[df['Signed_In'] == 0]

         female_signed_out_count = signed_out_users[signed_out_users['Gender'] ==
         male_signed_out_count = signed_out_users[signed_out_users['Gender'] == 1

         print("Number of signed out female users:", female_signed_out_count)
         print("Number of signed out male users:", male_signed_out_count)

         Number of signed out female users: 5613610
         Number of signed out male users: 0
```

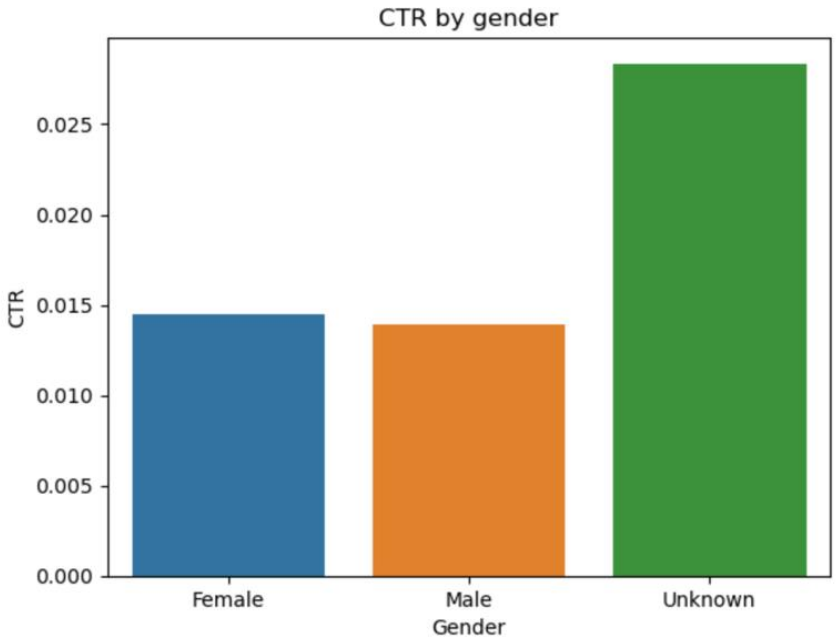
Проблема в тому, що користувачі, що не увійшли в акаунт, автоматично мають Gender = 0 та помилково розглядаються як жінки. Оскільки їх стать невідома, вони повинні розглядатися як окрема категорія користувачів.

Додаємо нову категорію для Gender – Unknown, та відображаємо дані про CTR знову:

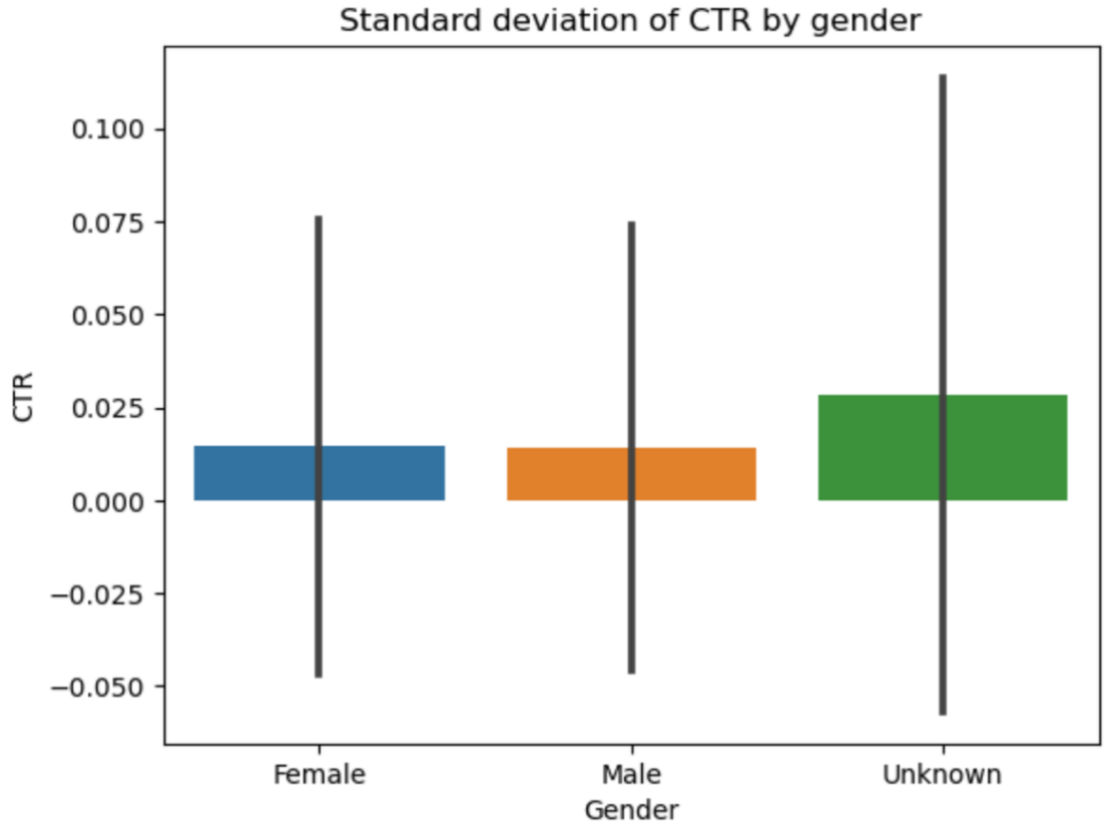
```
In [15]: # updating 'Gender' for signed out users with unknown gender to 'Unknown
         df.loc[(df['Signed_In'] == 0) & (df['Gender'] == 0), 'Gender'] = 'Unknow
```

Тепер маємо діаграму, яка коректно відображає CTR по статі:

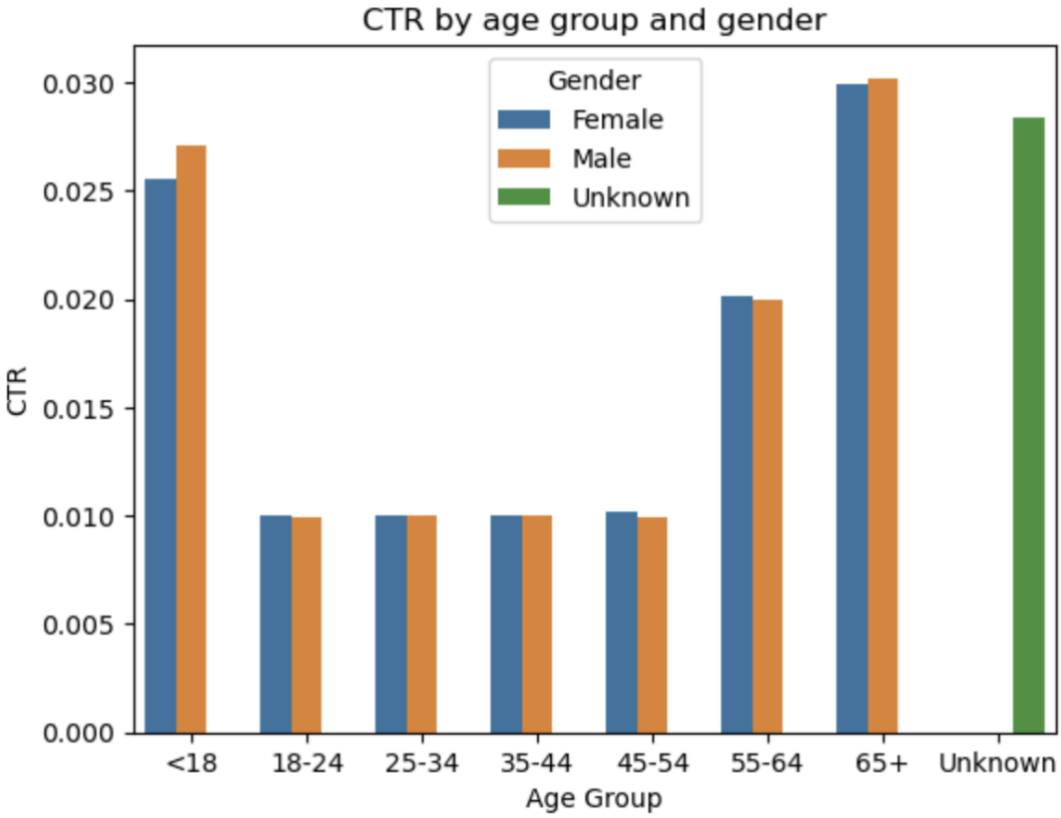
```
In [16]: sns.barplot(x='Gender', y='CTR', data=df, ci=None)
         plt.xlabel('Gender')
         plt.title('CTR by gender')
         plt.xticks(ticks=[0, 1, 2], labels=['Female', 'Male', 'Unknown'])
         plt.show()
```



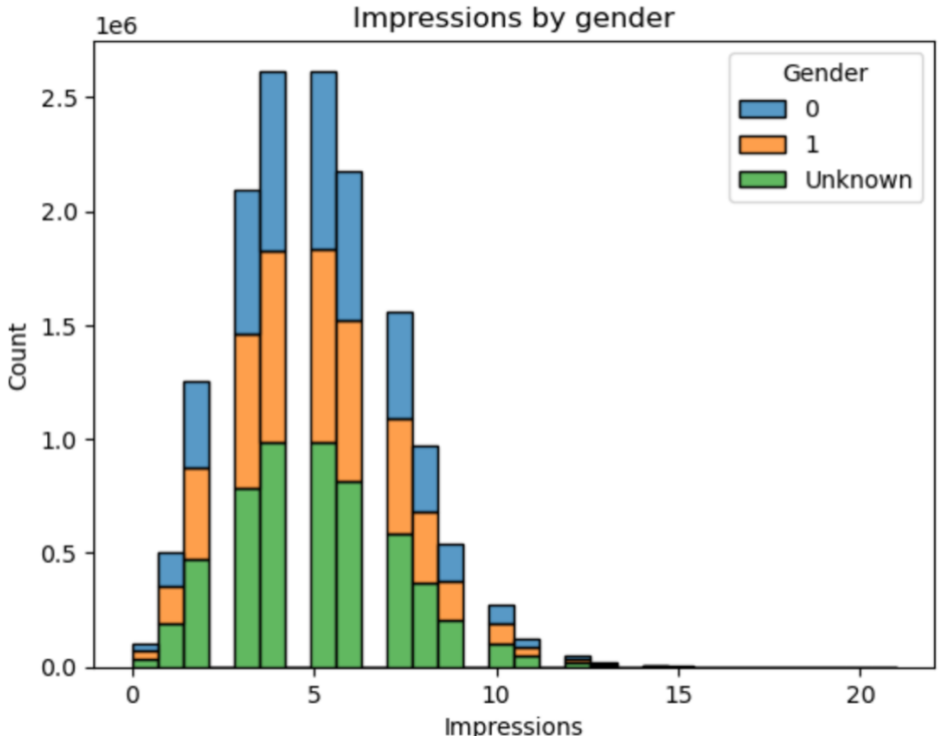
Також, відобразимо стандартне відхилення для цих показників (чорна лінія):



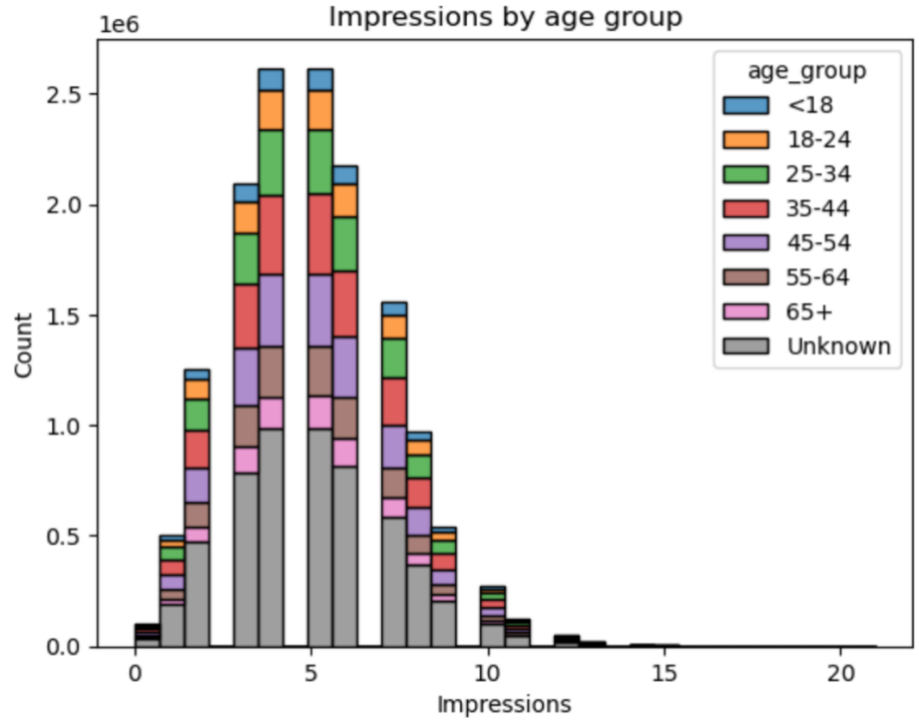

Зберемо цю інформацію в один графік:



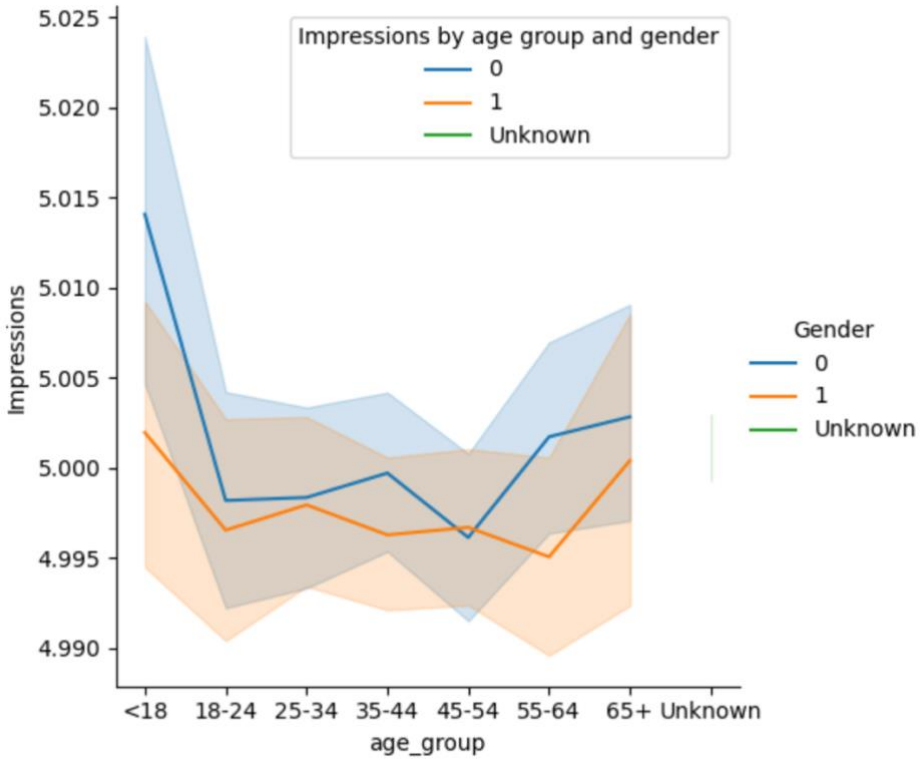
Далі, створимо гістограму, що відображає частоту показів за статтю:



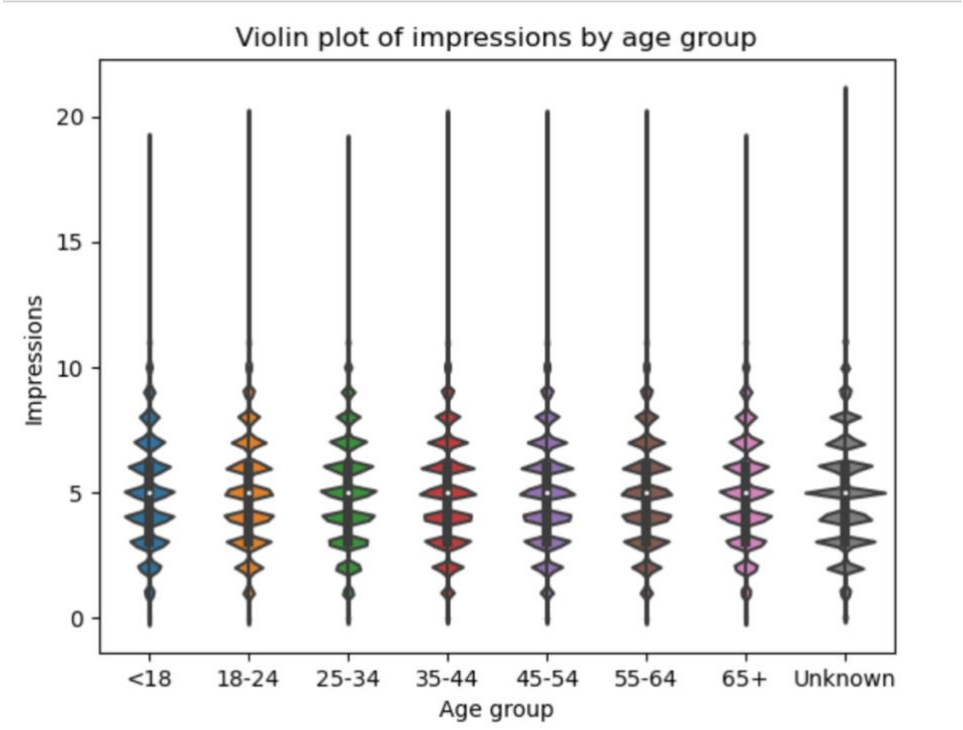

Схоже, що гістограма має нормальний розподіл. Відобразимо також покази за віковими категоріями:



Створимо графік, який відображає mean для кожної групи користувачів та статі:

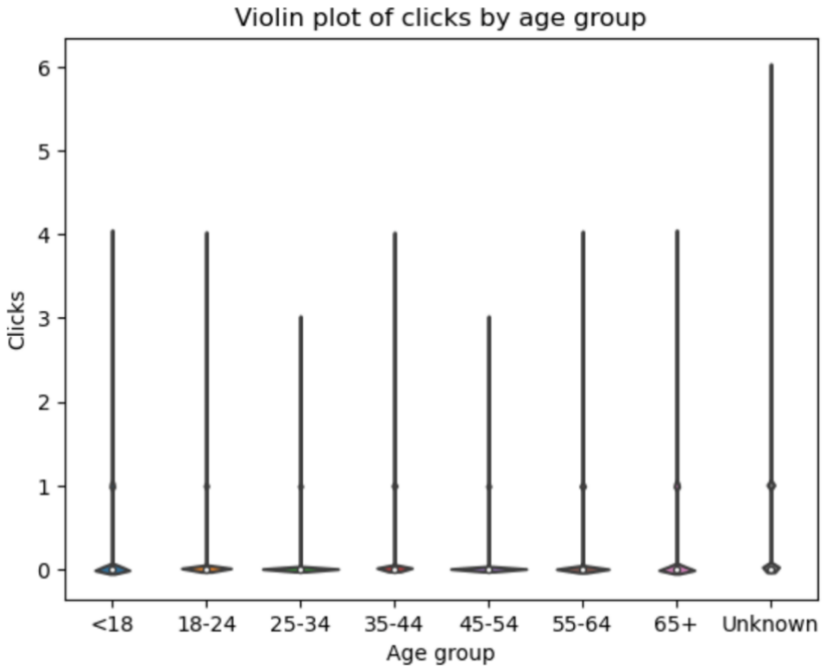



Також створимо violin plot для показів по віковим групам, що показує медіану, частотність, максимальне та мінімальне значення, а також квантили:

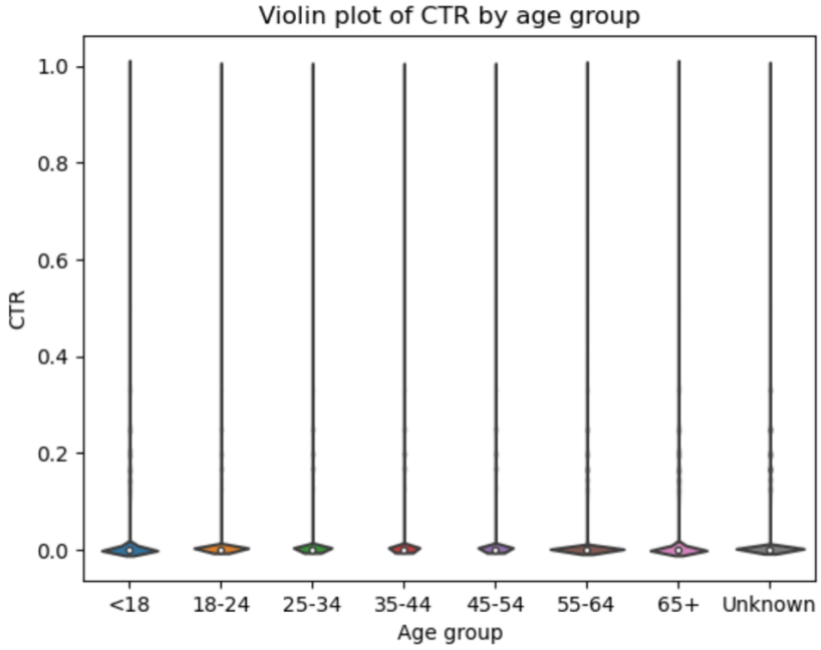


Бачимо, що медіана у всіх вікових груп дорівнює п’яти, та 0.25 та 0.75 квантили теж приблизно співпадають.

Зобразимо такі ж дані для кліків:




Та для CTR:




***Висновок:*** Отже, у ході цієї лабораторної роботи було отримано практичні навички у роботі з raw data, використовуючи пакети jupyter, pandas, seaborn.