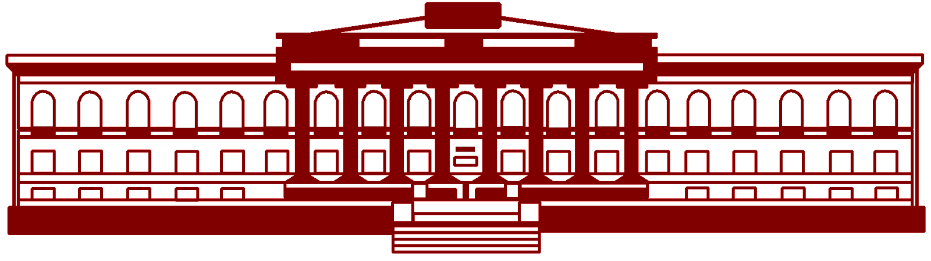
**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**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»

Цимбала Дмитра

Викладач:

Білий Р.О.

**Київ – 2022**

## Агрегація, обробка пропусків та візуалізація даних пакетами Python

**1. Мета роботи**

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

**2. 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.

### 3. Завдання до 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

**Виконання:**

1. **Кількість показів та показних переходів для шести вікових категорій**

age\_group\_data = df.groupby('age\_group').agg({'Impressions': 'sum', 'Clicks': 'sum'}).reset\_index()

age\_group\_data['CTR'] = age\_group\_data['Clicks'] / age\_group\_data['Impressions']

plt.figure(figsize=(10, 6))

sns.barplot(x='age\_group', y='Impressions', data=age\_group\_data, palette='plasma')

plt.title('Impressions by Age Group')

plt.xlabel('Age Group')

plt.ylabel('Number of Impressions')

plt.show()

plt.figure(figsize=(10, 6))

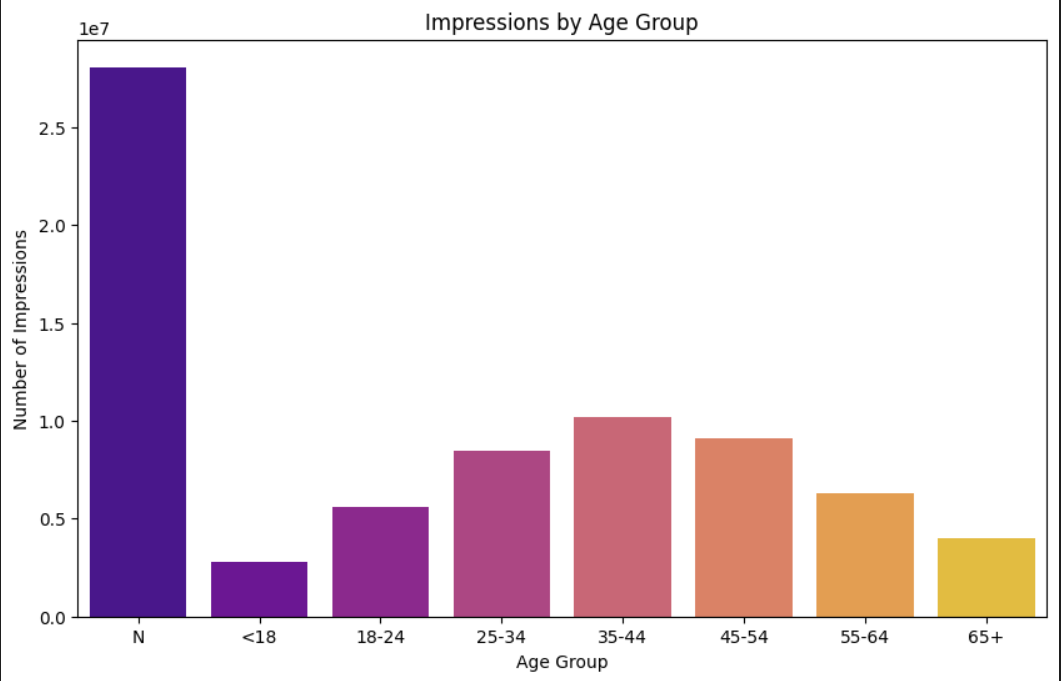
sns.barplot(x='age\_group', y='CTR', data=age\_group\_data, palette='plasma')

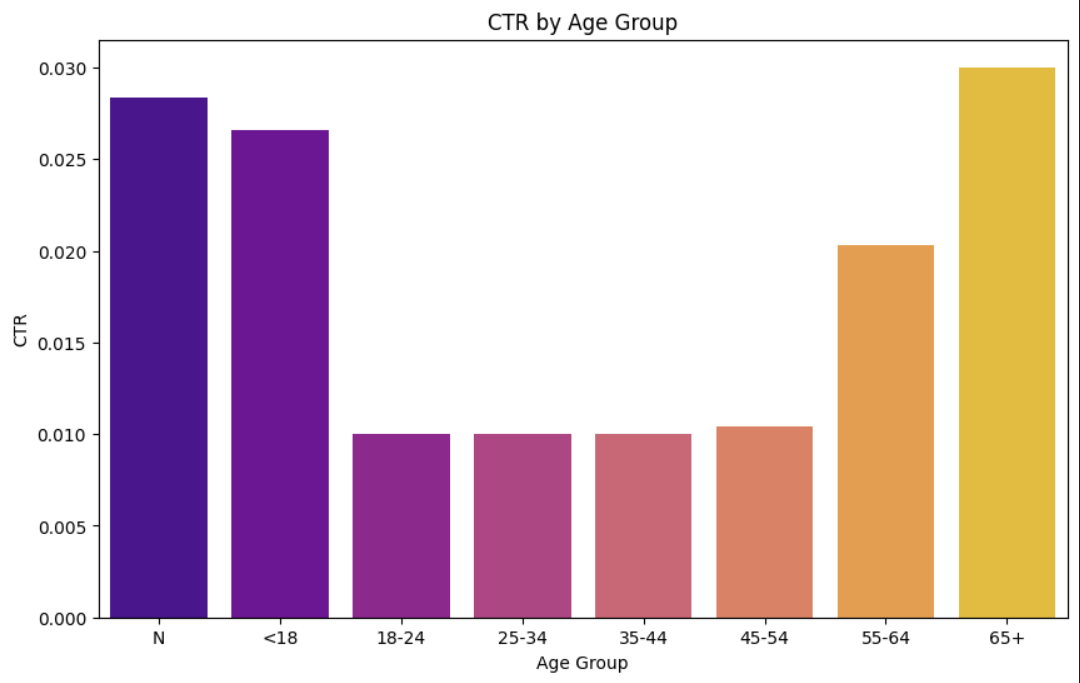
plt.title('CTR by Age Group')

plt.xlabel('Age Group')

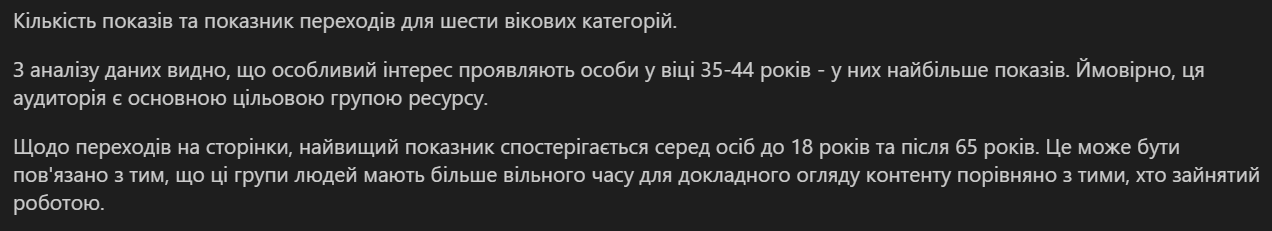
plt.ylabel('CTR')

plt.show()



****

**Висновок:**

****

1. **Кількість показів та показних переходів людей за віком і гендером**

#Кількість показів та показник переходів людей за віком і гендером.

days\_signed = df\_s.groupby(['age\_group', 'Gender']).agg({'Impressions': 'sum', 'Clicks': 'sum'}).reset\_index()

days\_signed['CTR'] = days\_signed['Clicks'] / days\_signed['Impressions']

plt.figure(figsize=(12, 6))

sns.barplot(x='age\_group', y='Impressions', hue='Gender', data=days\_signed, palette=['red', 'blue'])

plt.title('Impressions by Age Group and Gender')

plt.xlabel('Age Group')

plt.ylabel('Number of Impressions')

plt.legend(title='Gender (female/male)', loc='upper right')

plt.show()

clicks\_amount = df\_s.groupby(['age\_group', 'Gender']).agg({'Impressions': 'sum', 'Clicks': 'sum'}).reset\_index()

plt.figure(figsize=(12, 6))

sns.barplot(x='age\_group', y='Clicks', hue='Gender', data=clicks\_amount, palette=['red', 'blue'])

plt.title('Clicks by age group (Signed)')

plt.xlabel('Age Group')

plt.ylabel('Clicks')

plt.legend(title='Gender (female\male)', loc='upper right')

plt.show()

plt.figure(figsize=(12, 6))

sns.barplot(x='age\_group', y='CTR', hue='Gender', data=days\_signed, palette=['red', 'blue'])

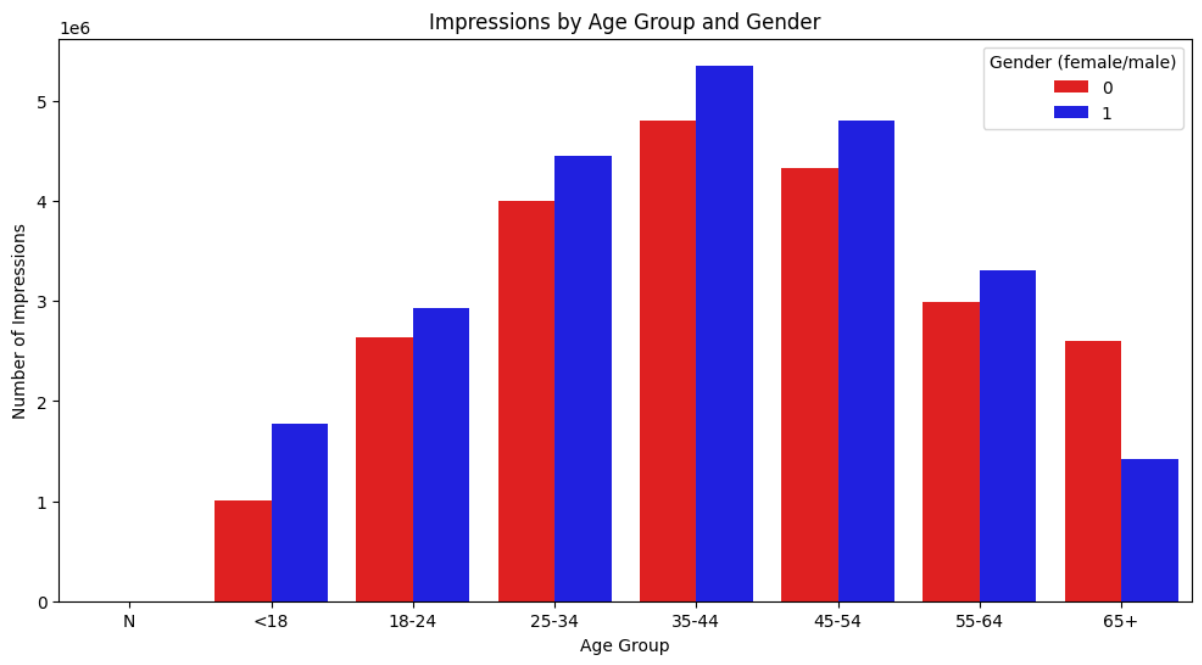
plt.title('CTR by Age Group and Gender')

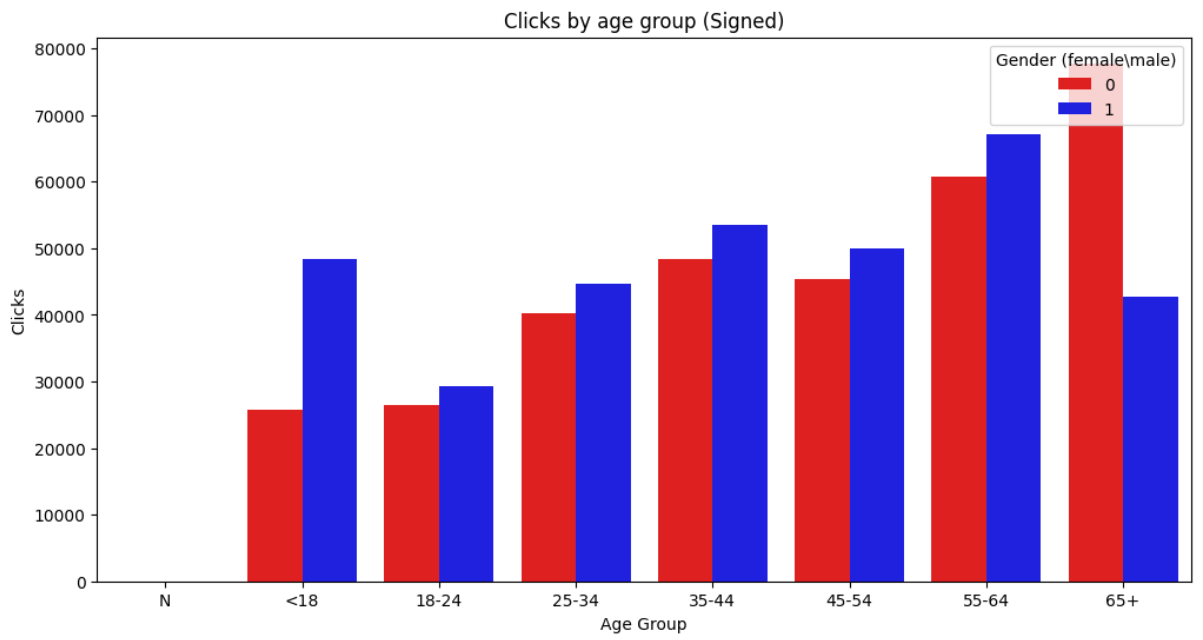
plt.xlabel('Age Group')

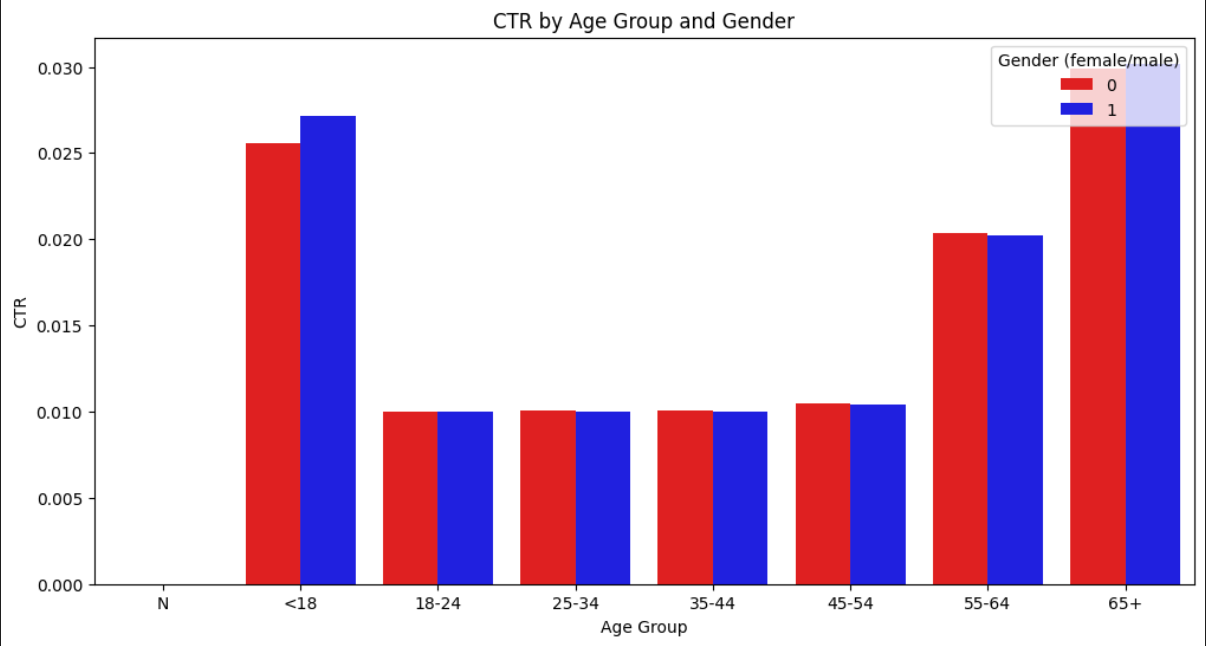
plt.ylabel('CTR')

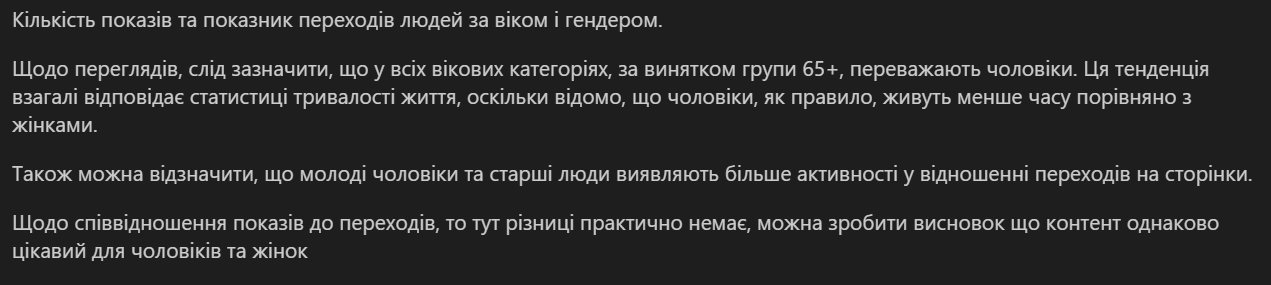
plt.legend(title='Gender (female/male)', loc='upper right')

plt.show()

****

****

****

**Висновок:  
**

1. **Кількість показів та показних переходів впродовж дня за гендером**

#Кількість показів та показник переходів впродовж дня за гендером.

days\_gender = df\_s.groupby(['Day', 'Gender']).agg({'Impressions': 'sum', 'Clicks': 'sum'}).reset\_index()

days\_gender['CTR'] = days\_gender['Clicks'] / days\_gender['Impressions']

plt.figure(figsize=(12, 6))

sns.lineplot(x='Day', y='Impressions', hue='Gender', data=days\_gender,palette=['red', 'blue'])

plt.title('Impressions per day (Gender)')

plt.xlabel('Day of month')

plt.ylabel('Number of Impressions')

plt.legend(title='Gender (female/male)', loc='upper right')

plt.show()

plt.figure(figsize=(12, 6))

sns.lineplot(x='Day', y='CTR', hue='Gender', data=days\_gender, palette=['red', 'blue'])

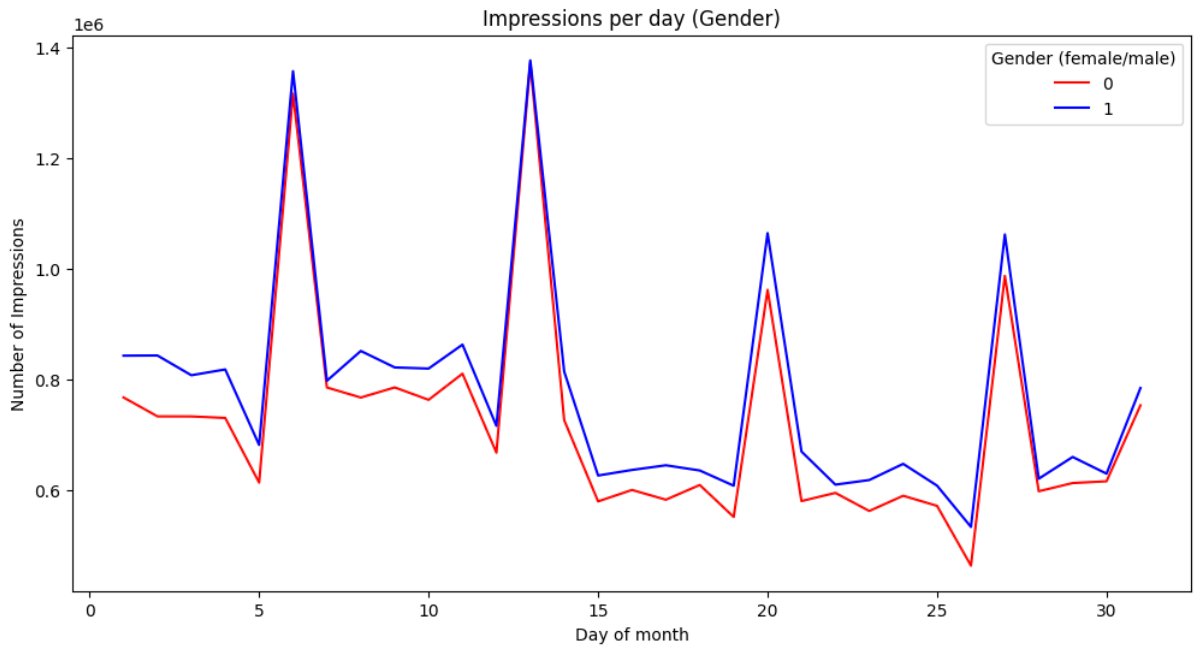
plt.title('CTR per day (Gender)')

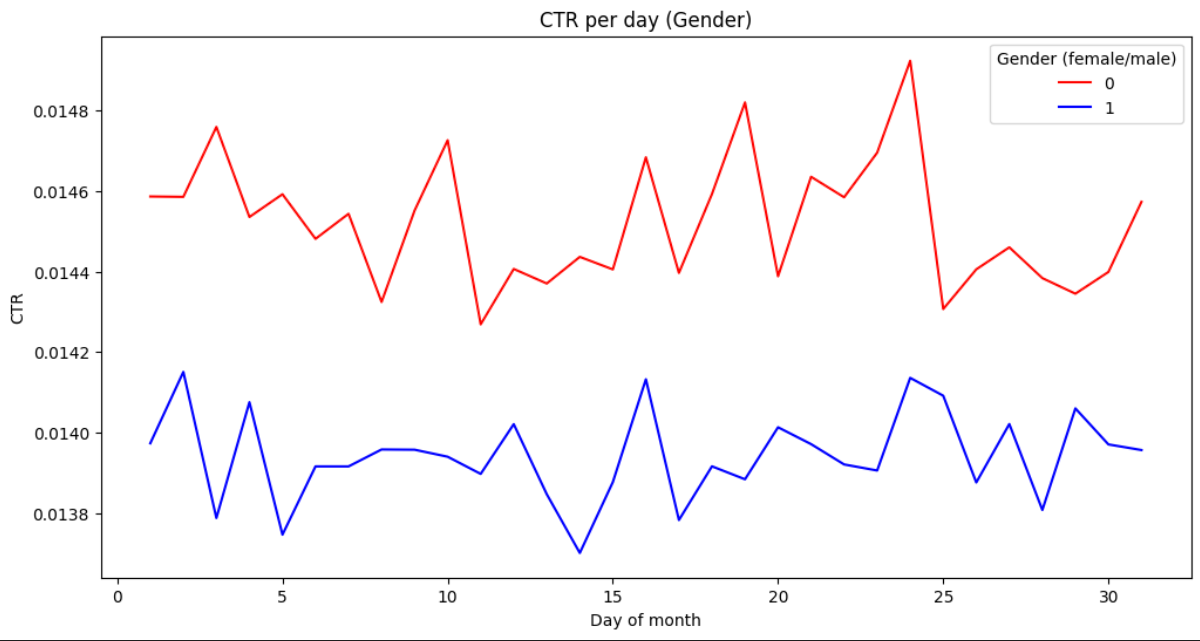
plt.xlabel('Day of month')

plt.ylabel('CTR')

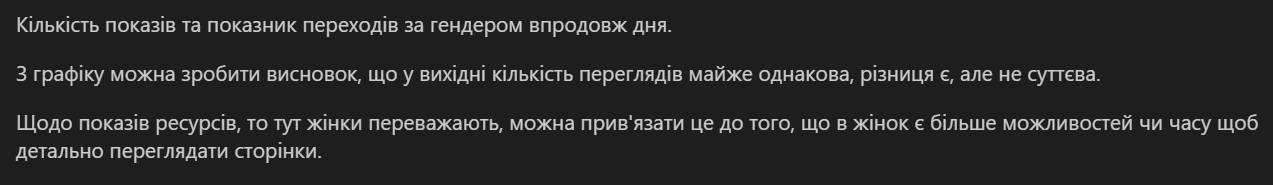
plt.legend(title='Gender (female/male)', loc='upper right')

plt.show()

****

****

**Висновок:**



1. **Показних переходів на ресурси впродовж тижня в заложності від віку і статі**

#Показник переходів на ресурси впродовж тижня в незалежності від віку і статі.

df['CTR'] = df['Clicks'] / df['Impressions']

grouped\_data = df.groupby(['Day'])

ctr\_over\_time = grouped\_data['CTR'].mean().reset\_index()

plt.figure(figsize=(12, 6))

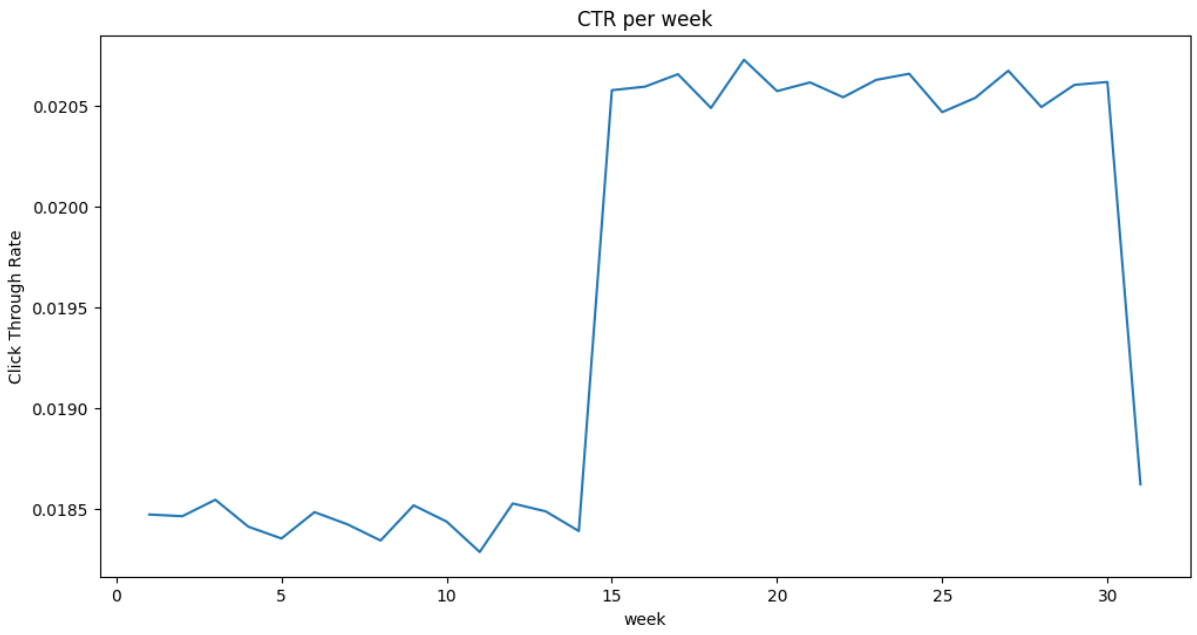
sns.lineplot(x='Day', y='CTR', data=ctr\_over\_time)

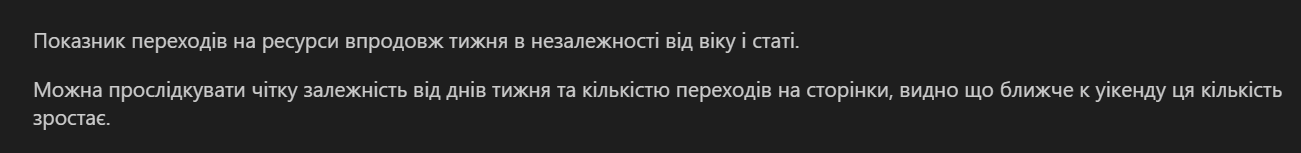
plt.title('CTR per week')

plt.xlabel('week')

plt.ylabel('Click Through Rate')

plt.show()

****

**Висновок:  
**

**Висновок:** Отримав практичні навички у роботі з raw data, використовуючи пакети jupyter, pandas, seaborn.