# КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені ТАРАСА ШЕВЧЕНКА



## ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

## Кафедра прикладних інформаційних систем

**Звіт до лабораторної роботи №1**

# з курсу «Big Data»

*Студента 4 курсу*

*Групи ПП-41*

*спеціальності 122*

*«Комп'ютерні науки»*

*ОП«Прикладне програмування»*

Глинки Олега

## Київ – 2023

# Тема

## Агрегація, обробка пропусків та візуалізація даних пакетами Python

# Мета

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

# Виконання завдання

# Статистика CTR

# Picture 15

# З графіка можемо зробити висновок, що найбільший CTR у користувачів, яким більше 65 років, та у користувачів, яким менше 18 років. Найменший у користувачів, яким від 18 до 64 років.

# Статистика зареєстрованих користувачів

# Picture 5

# З графіка можемо зробити висновок, щонайбільше зареєстрованих користувачів у віковій категорії 35-44(близько 24%), а найменше у віковій категорії <18(близько 4%).

# Picture 6

# З графіка можемо зробити висновок, чоловіків серед зареєстрованих користувачів близько на 3% більше, ніж жінок.

# Підсумок статистики реакцій

# Picture 16

# Найбільша кількість реакцій у користувачів, яким менше 18(близько 30 мільйонів), а найменша у користувачів, яким більше 65(близько 5 мільйонів).

# Picture 17

# Середнє значення у всіх вікових груп однакове і дорівнює 5.

# Picture 1

# Бачимо, що найбільшу кількість реакцій робит вікова група 18-, а найбільша кількість реакцій це 4 та 6. Найменша кількість реакцій - 0, 1 та усі, що більше 12.

# Picture 8

# На графіку бачимо, що медіана реакцій для всіх вікових груп - 5.

# Picture 9

# З графіку робимо висновок, що максимальна кількість реакцій(21) серед вікової групи 0-18.

# Picture 10

# Найбільша дисперсія реакцій у вікової групи 0 - 18 = 2.240, а найменша у вікової групи 65+ = 2.233.

# Підсумок статистики кліків

# Picture 18

# Найбільша кількість кліків у користувачів, яким менше 18 років(близько 850000), а найменша у користувачів, яким від 18 до 24 років(близько 60000).

# Picture 19

# Найбільші середні значення кліків у вікових груп 18- та 65+(0.14 та 0.15 відповідно), а найменше у вікових груп від 25 до 54 років(0.05).Picture 2

# Бачимо, що найбільша кількість кліків - 0, найбільше робить їх вікова група 18-. Найменша кількість кліків - 3, вікова група 18- робить їх найбільше.

# Збільшені зображення на кліки 2 та 3 відповідно:

# Picture 4Picture 3

# Продовжуємо аналіз кліків.

# Picture 11

# Медіана кліків однакова у всіх вікових груп - 0.

# Picture 12

# Найбільша кількість кліків у вікової групи 0-18 = 6, а найменша у вікових груп 25-34 та 45-54 = 3.

# Picture 14

# Найбільша дисперсія кліків у вікових груп 0-18 та 65+ - близько 0.39 , а найменша у вікових груп 25-34, 35-44 та 45-54 - близько -0.23.

**Висновок**

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