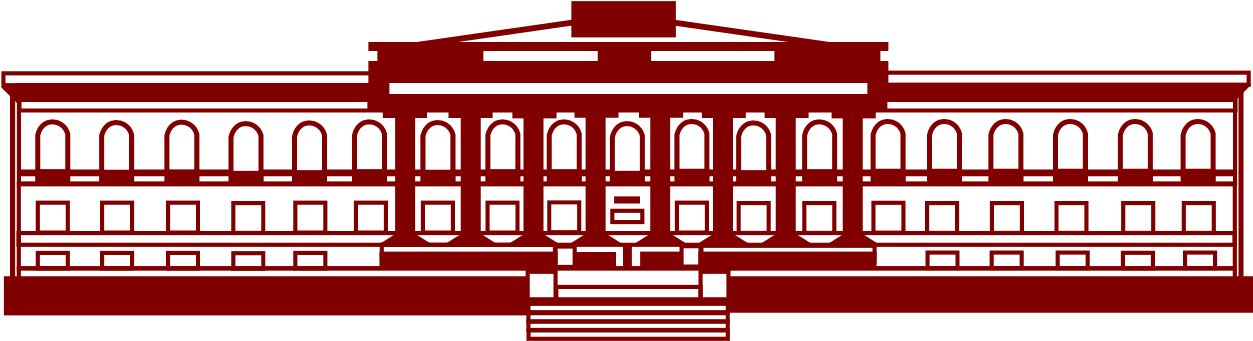
**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені ТАРАСА ШЕВЧЕНКА**



**ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**

**Кафедра прикладних інформаційних систем**

# Звіт до лабораторної роботи №1

# з курсу

**«Інтелектуальний аналіз даних»**

*Студентки 3 курсу групи ПП-22*

*спеціальності 122 «Комп'ютерні науки»*

*ОП «Прикладне програмування»*

Самчук Анастасії Олександрівни

*Викладач:*

Білий Р.О.

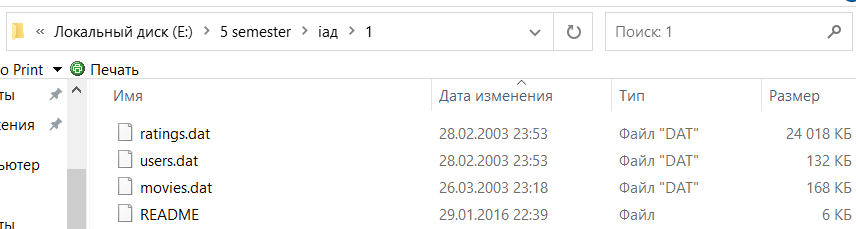
**Київ – 2023**

**Тема:** Розгортання середовища. Агрегація та дослідницький аналіз даних (EDA) пакетами Python.

**Мета:** Метою лабораторної роботи є отримання практичних навичок з розгортання робочого середовища, виконання агрегації та дослідницького аналізу даних, використовуючи пакети jupyter, pandas, seaborn, notebook.

**Хід виконання:**

1. *Встановлюю у віртуальне середовище пакети jupyter / pandas / seaborn, notebook, використовуючи conda.*
2. *Завантажую файли з даними MovieLens 1M Dataset з наданого лінку:* [*https://grouplens.org/datasets/movielens/1m/*](https://grouplens.org/datasets/movielens/1m/) *:*



1. *Завантажую файли movies.dat, ratings.dat та users.dat у змінні, використовуючи методи pandas:*

import pandas as pd

movies\_df = pd.read\_csv('movies.dat', sep='::', encoding='latin-1', names=['MovieID', 'Title', 'Genres'], engine='python')

ratings\_df = pd.read\_csv('ratings.dat', sep='::', encoding='latin-1', names=['UserID', 'MovieID', 'Rating', 'Timestamp'], engine='python')

users\_df = pd.read\_csv('users.dat', sep='::', encoding='latin-1', names=['UserID', 'Gender', 'Age', 'Occupation', 'Zip-code'], engine='python')

1. *Роблю merge даних в єдиний DataFrame:*

merged\_df = pd.merge(pd.merge(ratings\_df, users\_df), movies\_df)

1. *Роблю аналіз топ-10 стрічок за рейтингом по кожному полу та кожній віковій підгрупі з розбивкою по статі (14 рейтингів):*

age = [1, 18, 25, 35, 45, 50, 56]

def top\_10\_movies\_by\_rating(gender):

    results = {}

    for i, ageGroup in enumerate(age):

        filtered\_data = merged\_df[(merged\_df['Gender'] == gender) & (merged\_df['Age'] == ageGroup)]

        sum\_ratings = filtered\_data.groupby('MovieID')['Rating'].count()

        movies = sum\_ratings[sum\_ratings >= 20].index.tolist()

        filtered\_movies = filtered\_data[filtered\_data['MovieID'].isin(movies)]

        rating = filtered\_movies.groupby(['Title'])['Rating'].mean().nlargest(10).reset\_index()

        results[f'{age[i]}'] = rating

    return results

female\_1 = top\_10\_movies\_by\_rating('F')['1']

top\_10\_movies\_by\_rating('F')['1']

1. *Візуалізую один із отриманих рейтингів засобами пакету seaborn:*

Код:

import seaborn as sns

import matplotlib.pyplot as plt

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

sns.barplot(x=female\_1.values, y=female\_1.index, palette='viridis')

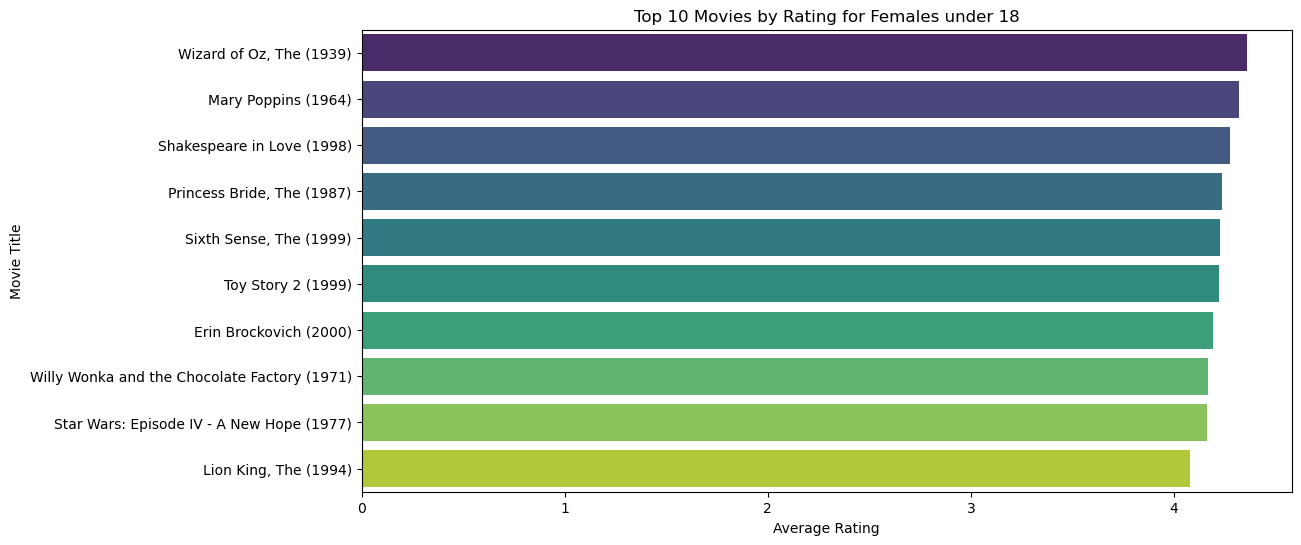
plt.xlabel('Average Rating')

plt.ylabel('Movie Title')

plt.title('Top 10 Movies by Rating for Females under 18')

plt.show()

Результат:



**Висновки**

Під час виконання лабораторної роботи я навчилася розгортати робоче середовище, виконувати агрегації та дослідницький аналіз даних, використовуючи пакети jupyter, pandas, seaborn, notebook.