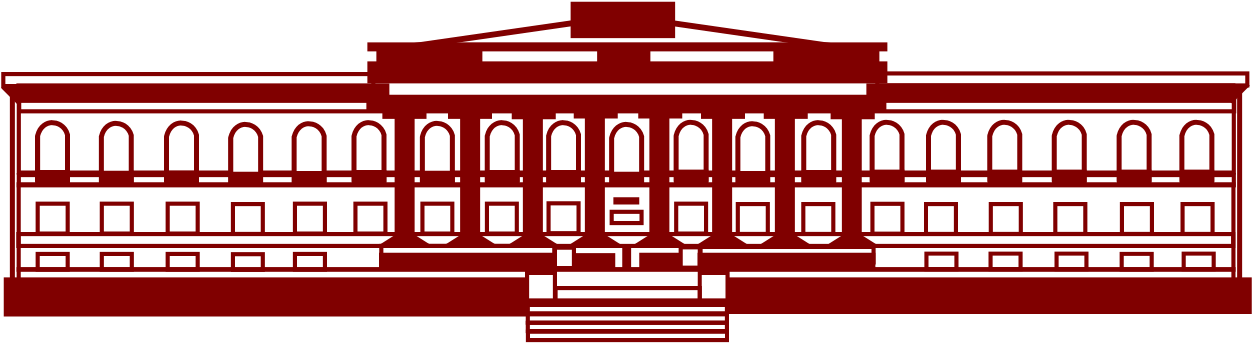
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені ТАРАСА ШЕВЧЕНКА



**ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**

**Кафедра прикладних інформаційних систем**

Звіт до практичної роботи №1

**з курсу**

**«Інтелектуальний аналіз даних»**

*студента 3 курсу групи ПП-32/3*

*спеціальності 122 «Комп'ютерні науки»*

*ОП «Прикладне програмування»*

Карпенка Глєба Олеговича

*Викладач:*

ас. Білий Р.О.

**Київ – 2023**

**Лабораторна робота №1**

**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 роботи.**

1. Перевіримо версії встановлених VS code. miniconda, Python 3.10+







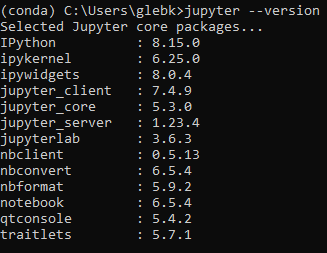
1. Створимо віртуальне середовище conda з python 3.12.



Версія python



1. Встановимо туди пакети jupyter / pandas / seaborn notebook



1. Завантажимо у папку проекту файли з даними MovieLens.
2. Прочитаємо README, завантажимо файли .dat у змінні.
3. import pandas as pd
4. import seaborn as sns
5. import matplotlib.pyplot as plt
6. import os
7. DIR\_NAME = 'ml-1m'
8. DAT\_ENCODING = 'ISO-8859-1'
9. data\_files = {
10. 'movies': {
11. 'filename': 'movies.dat',
12. 'column\_names': ['MovieID', 'Title', 'Genres']
13. },
14. 'ratings': {
15. 'filename': 'ratings.dat',
16. 'column\_names': ['UserID', 'MovieID', 'Rating', 'Timestamp']
17. },
18. 'users': {
19. 'filename': 'users.dat',
20. 'column\_names': ['UserID', 'Gender', 'Age', 'Occupation', 'Zip-code']
21. }
22. }
23. def read\_data\_from(filename, column\_names):
24. filepath = os.path.join(DIR\_NAME, filename)
25. df = pd.read\_csv(filepath, sep='::', header=None, names=column\_names, encoding=DAT\_ENCODING, engine='python')
26. return df
27. movies = read\_data\_from(data\_files['movies']['filename'], data\_files['movies']['column\_names'])
28. ratings = read\_data\_from(data\_files['ratings']['filename'], data\_files['ratings']['column\_names'])
29. users = read\_data\_from(data\_files['users']['filename'], data\_files['users']['column\_names'])

6. Зробимо злиття даних у DataFrame use\_ratings, який поєднує в собі рейтинги, які залишили користувачі, назви фільмів можна отримувати з movies DataFrame

user\_ratings = pd.merge(users, ratings, on='UserID', how='inner')

7. Отримаємо аналіз топ десяти фільмів для кожної демографічної групи(за віком і статтю), всього 14.

def get\_top\_10\_movies():

    grouped = user\_ratings.groupby(['Gender', 'Age', 'MovieID'])

    average\_ratings = grouped['Rating'].mean().reset\_index()

    top\_10\_movies = average\_ratings.groupby(['Gender', 'Age']).apply(lambda x: x.nlargest(10, 'Rating')).reset\_index(drop=True)

    return top\_10\_movies

top\_10\_movies = pd.merge(get\_top\_10\_movies(), movies, on='MovieID', how='left')

8. Отримаємо “цільову групу”, візуалізуємо її методами seaborn.

sns.set(style="whitegrid")

plt.figure(figsize=(8, 24))

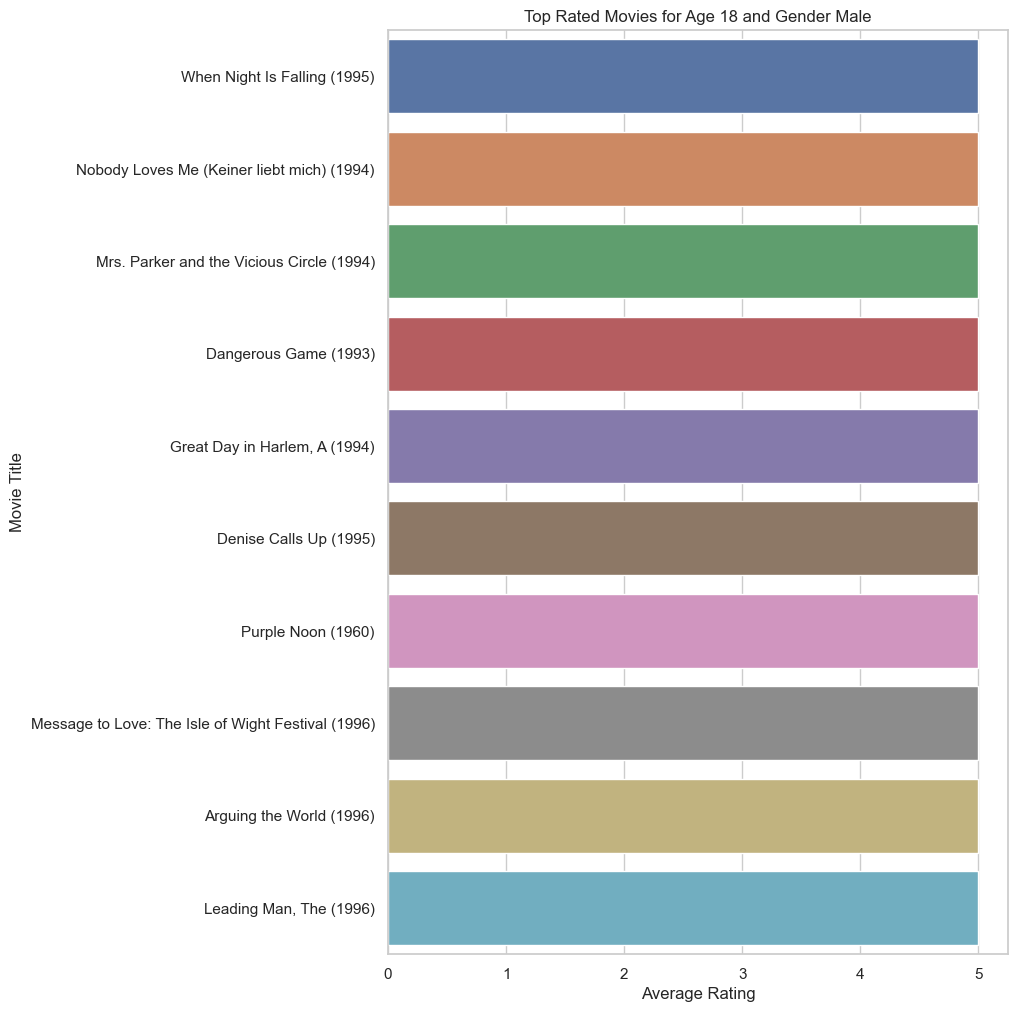
sns.barplot(data=target\_group, x='Rating', y='Title', ci=None)

plt.title('Top Rated Movies for Age 18 and Gender Male')

plt.xlabel('Average Rating')

plt.ylabel('Movie Title')

plt.show()



**Висновок**

На лабораторній роботі я розгорнув робоче середовище для аналізу даних використовуючи Python 3.10, miniconda3, jupyter, pandas, seaborn, notebook. Для практичного аналізу даних я завантажив дані понад 1млн. анонімних рейтингів приблизно 3,900 фільмів зроблених 6тис. користувачів. Використовуючі вказані пакети я зробив аналіз топ 10 стрічок для кожної демографічної групи і обрав одну цільову, котру я візуалізував методами pyplot і seaborn.