**Лабораторна робота №5**

**Тема «Аналіз даних даних і візуалізація Pyspark»**

Почтовий Богдан

**Виконання завдань:**

1.Прочитали CSV файл за допомогою spark.read.csv().

2.Вивели схему даних за допомогою printSchema.

3.Вивели назви стовпців і типи даних за допомогою dtypes.

4.Вивели перші п'ять рядків даних за допомогою методу head.

5.Вивели перший рядок даних за допомогою методу first().

6.Зробили опис датасету за допомогою describe().

7.Вивели список стовпців за допомогою методу columns.

8.Вивели загальну кількість рядків даних за допомогою методу count.

9.Перевірили, чи у всіх рядків вистачає даних. Замінили пропущені значення на NaN, а потім тип даних на числовий.

10.Замінили комірки з пропущеними значеннями на середні за стовпцем і тип даних на float.

11.Визначили, що датасет містить інформацію про 550 книг.

12.Змінили назви всіх колонок на більш зрозумілі та зручні для роботи.

13.Знайшли автора з найвищим та найнижчим рейтингом, а також автора з найбільшою кількістю рецензій.

14.Побудували гістограму рейтингу 10 найкращих книг.

15.Побудували матрицю кореляції для стовпців 'user\_rating', 'reviews', 'price'.

**Висновки:** У процесі виконання лабораторної роботи ми отримали практичні навички зчитування та аналізу даних у середовищі Python за допомогою бібліотеки Pandas. Ми ознайомилися з основними методами та функціями для обробки та аналізу даних, які допомагають отримати корисну інформацію з набору даних.

**КОД**

import pandas as pd

# Зчитування даних з CSV файлу

df = pd.read\_csv('bestsellers\_with\_categories.csv')

# Виведення схеми даних

print("Схема даних:")

print(df.dtypes)

print()

# Виведення назв стовпців і типів даних

print("Назви стовпців і типи даних:")

print(df.dtypes)

print()

# Виведення перших п'яти рядків

print("Перші п'ять рядків:")

print(df.head())

print()

# Виведення першого рядка

print("Перший рядок:")

print(df.iloc[0])

print()

# Опис датасету

print("Опис датасету:")

print(df.describe())

print()

# Список стовпців

print("Список стовпців:")

print(df.columns.tolist())

print()

# Загальна кількість рядків даних

print("Загальна кількість рядків даних:", len(df))

print()

# Заміна непридатних значень на NaN та типів даних на числовий

df = df.replace('NaN', pd.NA)

df[['User Rating', 'Reviews', 'Price']] = df[['User Rating', 'Reviews', 'Price']].apply(pd.to\_numeric, errors='coerce')

# Заміна пропущених значень на середні за стовпцем

df[['User Rating', 'Reviews', 'Price']] = df[['User Rating', 'Reviews', 'Price']].fillna(df[['User Rating', 'Reviews', 'Price']].mean())

# Зміна назв всіх колонок

df.columns = ['name', 'author', 'user\_rating', 'reviews', 'price', 'year', 'genre']

# Знайдення автора з найвищим рейтингом

print("Автор з найвищим рейтингом:")

print(df[df['user\_rating'] == df['user\_rating'].max()]['author'])

print()

# Знайдення автора з найнижчим рейтингом

print("Автор з найнижчим рейтингом:")

print(df[df['user\_rating'] == df['user\_rating'].min()]['author'])

print()

# Знайдення автора з найбільшою кількістю рецензій

print("Автор з найбільшою кількістю рецензій:")

print(df[df['reviews'] == df['reviews'].max()]['author'])

print()

# Побудова гістограми рейтингу 10 найкращих книг

print("Гістограма рейтингу 10 найкращих книг:")

top\_10\_books = df.nlargest(10, 'user\_rating')

top\_10\_books['user\_rating'].hist()

plt.show()

# Побудова матриці кореляції

print("Матриця кореляції:")

print(df[['user\_rating', 'reviews', 'price']].corr())

**Скріни**



