Загрузка данных

Для обработки данных был использован файл, содержащий информацию о фильмах. Таблица содержит следующие столбы:

1. Director (Режиссер);
2. Genre (Жанр);
3. Movie Title (Наименоввание фильма);
4. Release Date (Дата выхода);
5. Studio (Студия);
6. Adjusted Gross\_USDmill (Скорректированный общий бюджет);
7. Budget\_USDmill (Бюджет);
8. Gross\_USDmill (Общий бюджет);
9. IMDb Rating (Индекс в прокате. IMDb это американский сервис онлайн видео проката. Просмотрев фильм или сериал, зритель оценивает контент по шкале от одного до десяти.);
10. MovieLens Rating (Индекс в Интернете. MovieLens — это веб-система рекомендаций и виртуальноесообщество, которое рекомендует фильмы пользователям на основе ихпредпочтений.);
11. Overseas\_USDmill (Продажи билетов вне США);
12. Profit\_USDmill (Доход);
13. Profit\_proc (Доход в процентах);
14. US\_USDmill (Продажи билетов в США);

Данные были представлены в формате MS Excel в файле ‘data.xlsx’.

Для выполнения задания были установлены библиотеки pandas и openpyxl с помощью стандартного инструмента управления пакетами в Python – pip.

Для загрузки файлов была использована библиотека pandas. После импорта библиотеки pandas была использована функция read\_excel для чтения данных, содержащихся в файле.

Фрагмент программного кода выглядит следующим образом:

import pandas as pd

# Загрузка данных из файла

Excel df = pd.read\_excel('data.xlsx')

print(type(df))

Результат:

<class 'pandas.core.frame.DataFrame'>

Загруженные данные сохраняются в объект типа DataFrame, который представляет собой двумерную структуру данных, аналогичную таблице. Основными компонентами DataFrame являются строки, столбцы и значения. Каждый столбец в DataFrame в свою очередь представляет собой объект Series, который является одномерным массивом данных определенного типа. Для обращения к столбцу по его названию можно воспользоваться квадратными скобками и указать название столбца в виде строки внутри скобок.

Функция df.head() была использована для вывода нескольких первых строк данных из файла, по умолчанию выводятся 5 строк. Также была использована функция pd.set\_option('display.max\_columns', 14), позволяющая настроить максимальное количество выводимых столбцов, и функция pd.set\_option('display.width', 250), позволяющая отобразить все столбцы без переноса на следующие строки.

Фрагмент программного кода выглядит следующим образом:

import pandas as pd

# Загрузка данных из файла Excel

pd.set\_option('display.max\_columns', 14) # Настройка максимального количества выводимых столбцов

pd.set\_option('display.width', 250) # Настройка максимальной ширины экрана для вывода

df = pd.read\_excel('data.xlsx')

print(type(df))

print(df.head())

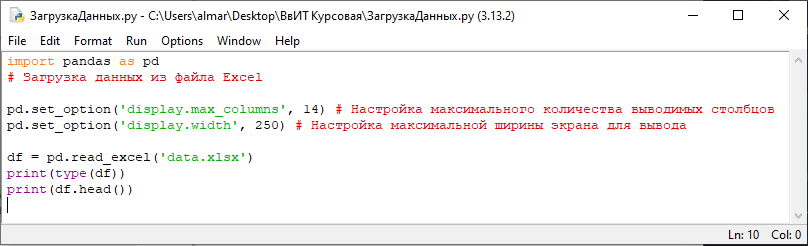


Рисунок 1.1 – Код программы в среде Python IDLE

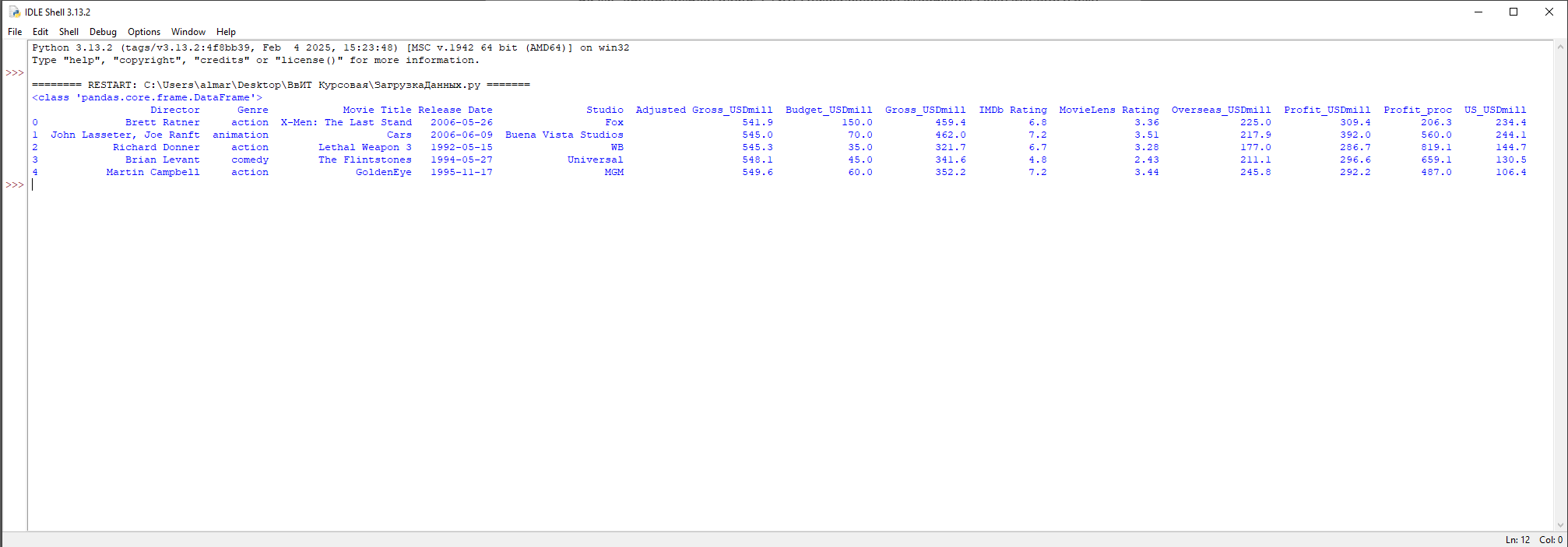


Рисунок 1.2 – Результат выполнения программы в среде Python IDLE

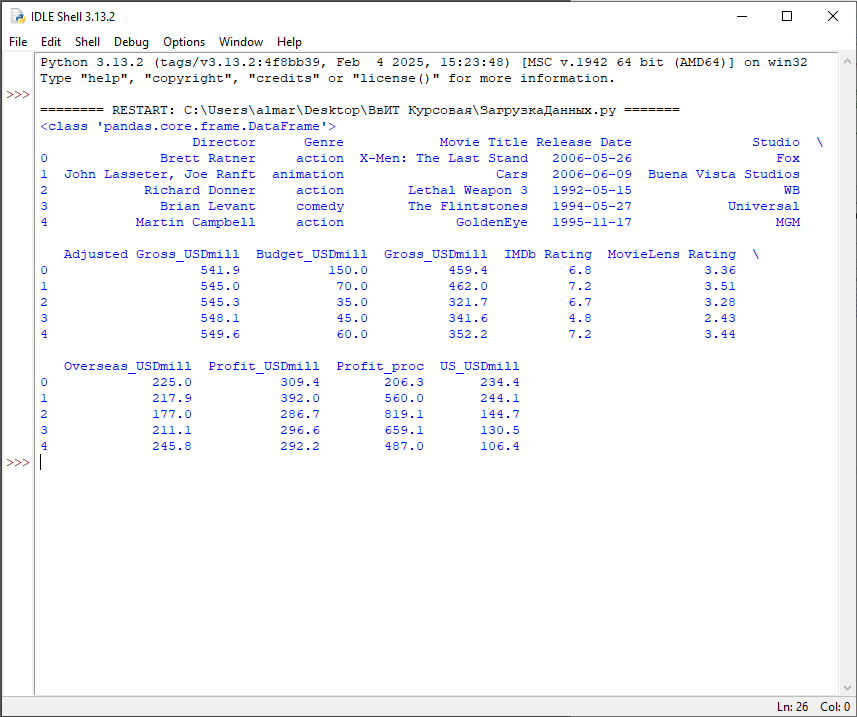


Рисунок 1.3 – Результат выполнения программы в среде Python IDLE со стандартной настройкой ширины экрана для вывода

Все данные в файле представлены в понятной и привычной форме, перевод не требуется.