Data Scraping

Part 1: Selenium [15 Marks]

You are provided with the URL of a YouTube channel. Your task is to write Python code using Selenium to scrape to extract relevant information and gain insights from the collected data.

URL: https://www.youtube.com/@UnfoldDataScience

- Use Selenium to access the provided YouTube channel URL. Scrape the videos uploaded between Sep 10, 2019 and Sep 10, 2023.
- Extract the following information for each video on the channel's page:
  - Video Title
  - Views Count
  - Likes Count
  - Upload Date
  - Number of Comments
- Store the extracted data in a structured format.
- Create functions for the following tasks on the scraped data:
  - Calculate the average views count per video for videos uploaded in the last 30 days.
  - Identify the video with the highest likes-to-views ratio.
  - Find the correlation between the number of likes and the number of dislikes for the videos.
  - Determine the most common day of the week for video uploads.
  - Detect any outliers in the views count.

Part 2: Beautiful Soup [15 Marks]

You are provided with the URL of a website that lists the top-rated movies. Your task is to write Python code using Beautiful Soup to scrape to perform following tasks.

URL: https://www.imdb.com/

- Scrape the movies released between 2013 and 2023.
- Write a Python script using BeautifulSoup to scrape the following information for each movie:
  - Movie Title
  - Release Year
  - IMDb Rating
  - Director
  - Genre

- Store the scraped data in a structured format, such as a CSV file.
- Create functions for each of the tasks listed below:
  - Average IMDb rating for the top-rated movies.
  - The most common genre among the top-rated movies.
  - Identify the director with the highest average IMDb rating.
  - Determine the year with the highest number of top-rated movies.