IMDb Top-Rated Movies Web Scraping Project

Objective

This project is designed to scrape and analyze IMDb's top-rated movies from the IMDb website. Using Python, the script fetches movie data, parses HTML and JSON content, cleans and transforms the data, and provides insightful visualizations on movie durations and ratings.

Steps

1. **Importing Dependencies**:
    - **Libraries**: requests for handling HTTP requests, BeautifulSoup from bs4 for HTML parsing, pandas for data manipulation, json for parsing JSON data, and matplotlib.pyplot for visualizations.
2. **Fetching Movie Data from IMDb**:
    - The script sends an HTTP GET request to IMDb's top-rated movies chart URL, using a user-agent header to mimic a browser request.
3. **Parsing HTML Content**:
    - BeautifulSoup is used to parse the HTML content from IMDb.
    - The <script> tag containing structured JSON data (type="application/ld+json") is extracted for further data parsing.
4. **Extracting and Storing Movie Details**:
    - **Data Fields**: The JSON data provides details such as the movie title, duration, URL, description, and IMDb rating. These details are stored in individual lists.
    - **Data Storage**: A pandas DataFrame is created with columns for each data field and is saved as scrapped_IMDBmovies.csv.
5. **Data Cleaning and Transformation**:
    - **Handling Missing Values**: Checks for missing values (NaN and null) are performed and displayed.
    - **Duration Conversion**: ISO 8601 format (PTxHyM) is converted to minutes for consistency, allowing for numeric analysis and plotting.
6. **Data Visualization**:
    - **Box Plot**: Two box plots display the distribution of movie durations and IMDb ratings, offering insights into data spread and potential outliers.
    - **Scatter Plot**: A scatter plot shows the relationship between movie duration and IMDb rating, allowing for a visual correlation between these variables.
7. **Final Data Export**:

- The cleaned and transformed data is saved as final_IMDBmovies.csv.

Project Outcome

By the end of this project, the extracted IMDb data allows for an understanding of movie duration and rating trends. The project also serves as a foundation for further exploration, such as analyzing genre-specific durations or year-wise rating trends. The data transformation and visualization steps make it easier to derive insights and identify patterns in IMDb's top-rated movies.