Lab: Joining and Analyzing Movie and Review Data

Objective:

This lab exercise is designed to help students not only perform data joining operations but also to conduct deeper analysis using the pandas library in Python. By working with the movie and review datasets, students will learn how to derive insights and perform complex data manipulations.

Datasets Description:

1. **Movies Dataset** (`movies.csv`):
   - **Columns**:
     - `movie_id`: Unique identifier for each movie
     - `title`: Title of the movie
     - `genre`: Genre of the movie
     - `release_year`: Year the movie was released
2. **Reviews Dataset** (`reviews.csv`):
   - **Columns**:
     - `review_id`: Unique identifier for each review
     - `movie_id`: Identifier linking the review to a movie
     - `reviewer`: Name of the reviewer
     - `rating`: Rating given by the reviewer (out of 10)
     - `review_date`: Date when the review was posted

Tasks:

1. **Data Preparation and Exploration:**
   - Load the `movies.csv` and `reviews.csv` datasets into pandas dataframes.

- Display summary statistics and the first few rows of each dataframe to understand their structure.

2. **Highest Rated Movies:**
   - Perform an inner join on the `movies` and `reviews` dataframes based on the `movie_id` column.
   - Calculate the average rating for each movie.
   - Identify the top 10 highest-rated movies.
   - Display the titles and average ratings of these top 10 movies.
3. **Genre Analysis:**
   - Group the joined dataframe by `genre` and calculate the average rating for each genre.
   - Determine which genre has the highest average rating.
   - Display the genres along with their average ratings.
4. **Reviewer Analysis:**
   - Identify the top 5 reviewers who have given the most reviews.
   - Calculate the average rating given by each of these top 5 reviewers.
   - Display the reviewer names and their average ratings.
5. **Time-Based Analysis:**
   - Analyze how movie ratings have changed over time.
   - Group the joined dataframe by the month extracted from the `review_date` column and calculate the average rating for each month.
   - Create a line plot showing the trend of average movie ratings over time.
   - Discuss any observable trends.
6. **Movies with Most Reviews:**
   - Identify the top 10 movies that have received the most reviews.
   - Display the titles of these movies along with the number of reviews they have received.

7. **Distribution of Ratings:**
    - Create a histogram to visualize the distribution of ratings.
    - Analyze the distribution and discuss any skewness or patterns observed in the ratings.
8. **Impact of Release Year:**
    - Analyze if there is any correlation between the release year of a movie and its average rating.
    - Create a scatter plot showing the relationship between the release year and the average rating.
    - Discuss any observable patterns or correlations.
9. **Reviewer Consistency:**
    - For each reviewer, calculate the standard deviation of their ratings.
    - Identify the most and least consistent reviewers based on the standard deviation of their ratings.
    - Display the names of these reviewers along with their standard deviations.
10. **Challenge Task - Genre Popularity Over Time:**
    - Analyze how the popularity of different genres has changed over time.
    - Group the joined dataframe by `genre` and the year extracted from the `review_date` column.
    - Calculate the number of reviews for each genre per month.
    - Create a line plot for each genre showing the number of reviews over time.
    - Discuss any trends or shifts in genre popularity over time.

Submission:

- Submit a Jupyter notebook file (`.ipynb`) with the solutions to all tasks.
- Ensure that the notebook contains explanations and comments for each step.
- Include the `movies.csv` and `reviews.csv` files used in the lab.

Notes:

- Handle any edge cases, such as missing or duplicated data, appropriately.
- Use appropriate pandas functions and methods to accomplish each task.
- Write clean, readable code and include comments to explain your logic.
- Visualize data effectively using appropriate plots and charts.

Good luck, and enjoy the process of mastering data joining and analysis with pandas!