Self-Practice Exercise – Regression Models on MovieLens

In this exercise, you will practice building and evaluating regression models using the MovieLens dataset and the workflow demonstrated in the regression lecture notebook.

Dataset

We will use the MovieLens dataset (public ZIP archive).

Download: http://files.grouplens.org/datasets/movielens/ml-latest-small.zip

Contained files:
- ratings.csv → columns: userId, movieId, rating, timestamp
- movies.csv → columns: movieId, title, genres

Step 1 – Download & Load the Data

Use Pandas to read the ratings and movies data directly from the ZIP file.

```python
import pandas as pd
import requests
from zipfile import ZipFile
from io import BytesIO

url = 'http://files.grouplens.org/datasets/movielens/ml-latest-small.zip'
r = requests.get(url)
z = ZipFile(BytesIO(r.content))

ratings = pd.read_csv(z.open('ml-latest-small/ratings.csv'))
movies  = pd.read_csv(z.open('ml-latest-small/movies.csv'))

ratings.head()
movies.head()
```

Step 2 – Data Preparation

1. Merge ratings and movies on movieId.
2. Convert timestamp into datetime and extract year and month.
3. Aggregate at the movie level:
   - Average rating per movie
   - Number of ratings per movie
4. One-hot encode genres into dummy variables.

Step 3 – Regression Models

Build models to predict the average rating of a movie from its features.

Models to implement:
1. Linear Regression
2. Ridge Regression (regularization)
3. Lasso Regression (feature selection)

Step 4 – Model Evaluation

Calculate the model:
- R^2 score
- RMSE (Root Mean Squared Error)

Summarize results in a comparison table.

Step 5 – Interpret Results

Identify which features (e.g., genres, number of ratings) are most predictive.
Compare coefficients (linear models) vs feature importances (tree models).

Step 6 – Extensions (Optional, Advanced 🌟)

1. Add user-level features (average ratings given by each user).
2. Predict individual ratings instead of movie-level average ratings.
3. Try time-based splits: train on earlier ratings, test on later ones.

Tasks

1. Download and preview ratings and movies.
2. Merge, preprocess, and create features (average rating, number of ratings, genre dummies).
4. Build and evaluate regression models (Linear, Ridge, Lasso).
5. Compare models in a summary table (R^2, RMSE).
6. Interpret key features driving the predictions.