Exploratory Data Analysis

The exploratory data analysis section studies the dataset and makes some conclusions regarding the dataset, visualises the users and movies, rating distribution, sparsity, etc. The code analyses that the rating distribution is focused around 4 when the ratings are plotted against its count. A plot of userId vs no of ratings given by each user shows that each user has rated at least 20 movies. The plot of movie IDs vs the number of ratings follows the Long Tail phenomenon, where a small proportion of movies make up the popular part of the curve and a large number of movies belong to the tail, making them niche and obscure.

The majority of the code builds on the project built by Rohith Kumar Poshala (https://github.com/rposhala/Recommender-System-on-MovieLens-dataset). This corresponds to the initial part, loading the data and the estimation of sparsity using log count. The seaborn plots depicting rating distribution are visualised with the help of code from Jill Cates (https://github.com/topspinj/recommender-tutorial/blob/master/part-1-item-item-recommender.ipynb). The scatter plots are based on the implementation by Yash Patel. (https://www.kaggle.com/code/ysthehurricane/movie-recommendation-engine)

Dependencies and how to install them:

Numpy - pip install numpy
Pandas - pip install pandas
Matplotlib - pip install matplotlib
Seaborn - pip install seaborn