DA2402 Project : Movie Ratings Analysis

Data Curation and Visualization Course

1 Project Objective

Analyze movie ratings, user behavior patterns, and build recommendation insights using the MovieLens dataset.

2 Data Source and Collection

- **Download Page:** `https://grouplens.org/datasets/movielens/latest/`

- **Datasets:** Download "ml-latest-small.zip" containing:

  1. `movies.csv` - Movie metadata (movieId, title, genres)
  2. `ratings.csv` - User ratings (userId, movieId, rating, timestamp)
  3. `tags.csv` - User-generated tags (userId, movieId, tag, timestamp)
  4. `links.csv` - External database links (movieId, imdbId, tmdbId)

3 Data Storage

- Store each dataset in separate pandas DataFrames

- Create relationships between movies, ratings, and tags and merge the files into a single analysis DataFrame

4 Data Cleaning & Wrangling

- Handle duplicate ratings and missing values

- Extract release year from movie titles using regex

- Split pipe-separated genres into individual categories

- Convert UNIX timestamps to readable date-time format

- Calculate average ratings and rating counts per movie

- Analyze tag frequency and relevance

- And any other data handling required for visualizing the plots

5   Visualization & Dashboard

Build an interactive movie exploration dashboard:

- Rating distribution by genre and release year

- Tag cloud visualization of most common user tags

- User rating behavior analysis (most active users, rating patterns)

- Other interesting patterns you observed.