The goal in this project is to explore the connection between movie ratings and movie attributes such as genre, release year, and runtime by combining a ratings dataset with a metadata dataset. By consolidating user ratings with descriptive movie information, we can analyze trends in viewer preferences and construct a simple recommendation system.

Research Questions:

Do certain genres consistently receive higher average ratings?

How does the release year of a movie influence its ratings?

Teams:

Sheryl John

- Responsible for data cleaning and quality assessment
- Will handle tasks such as removing missing values, standardizing data types
- Leads documentation of cleaning steps, ethical considerations, metadata, and reproducibility notes.
- Also responsible for preparing visualizations that illustrate patterns in genres, release years, and rating distributions.

Xing

- Responsible for data integration
- Will implement the dataset joining process, specifically:
- Builds the reproducible pipeline (Jupyter Notebook) that automates loading, cleaning, joining, and analyzing the data.

Both of us will work together on exploratory analysis, deciding research questions, and preparing the final presentation.

Data Sets

This project will use two publicly available, complementary, reliable, and popular datasets used in research and instruction.

The first dataset is the MovieLens Ratings Dataset (100k) from GroupLens Research at the University of Minnesota. The dataset has been widely accepted as one of the most consistent benchmark datasets for recommender system projects and has been used in a huge number of academic studies. The one we are using has approximately 100,000 ratings from roughly 600 users on roughly 9,000 distinct movies. Each rating is recorded as a 0.5 to 5.0. There are four columns in the data: userId, an anonymized user id; movieId, the id of the movie being rated; rating, the rating provided; and timestamp, which records when it was performed. Because this data set is anonymized, it contains no personally identifiable information, so it is ethically acceptable to use. Its small size and neatly organized structure also render it ideal for in-class analysis, but still rich enough to facilitate meaningful observations.

The second data collection is the Movies Metadata Dataset, located on Kaggle and originally constructed from The Movie Database (TMDB), a highly dependable and regularly updated database of movie information. The dataset contains information regarding around 45,000 films and provides copious descriptive features that may be utilized to augment the MovieLens ratings. The dataset contains columns such as id (a unique identifier for every film), title, genres, release_date, runtime, budget, and revenue. Not every value is comprehensive, but the dataset's scope makes it an excellent one for enrichment. For example, release year and genre columns allow us to break ratings into categories by film type, and runtime and budget can be utilized to provide further context to deciphering viewer preferences.

The two datasets will be combined by merging the movieId column in MovieLens with the id column in the Kaggle dataset. Since the Kaggle id column will not always be an integer stored neatly, it will need to be standardized and cleaned before the join. When integrated, the two dataframes will allow us to link behavioral data (user ratings) to descriptive attributes (genres, years, runtimes), providing a more complete picture of film trends and enabling us to answer the research questions effectively.

Timeline:

In week one of the project, we will download both datasets from their original locations and carefully review their terms of use and ethical guidelines to ensure that they can be used within an academic environment. During week two, we will set up the GitHub repository directory structure and move the datasets into an isolated data/ directory, and with this explicit and well-structured directory, we are prepared to begin our project. Sheryl will pre-clean datasets in week three by handling missing values, normalizing identifiers, and standardizing date formats. Xing will merge the two datasets in week four by joining MovieLens movieId column with Kaggle id column, resolving mismatches and documenting the integration process.

After consolidating the data, the fifth week will be exploratory data analysis when both of us will examine trends in ratings across genre, year of release, and runtime. In the sixth week, Xing will

develop a basic recommender system that displays the top-rated movies by genre and year and how consolidated data can be utilized to form insights. The seventh week will be spent on workflow automation, where Xing will develop an automatable Jupyter Notebook pipeline that executes the whole process from loading the data to visualization. Lastly, during the eighth week, Sheryl will finalize the project documentation, such as ethical implications, metadata descriptions, and a refined final report. Collectively, these steps offer an orderly, week-by-week guide to finishing the project.

Constraints

This project only uses publicly available datasets, so our scope of analysis is solely based on what has been made public by MovieLens and Kaggle. Though this makes the data easy to use and accessible, it also limits us from using proprietary or more comprehensive datasets available elsewhere. Additionally, the MovieLens dataset is fully anonymized, which means we cannot examine any demographic information about users, such as gender, age, or location. Such a limitation restricts the types of inferences we can draw as we can only study rating activity without correlating it with user history. Finally, the movies metadata file may include missing or inconsistent values for features like release dates, runtime, budgets, or revenues that could hamper the completeness and credibility of parts of our analysis.

Gaps

There also exist gaps in the project scope that are missing due to the nature of the data. For example, there exists no external data in the form of advertising budgets, ad campaigns, or critic's opinions that are all typically key drivers of the success of a movie and the ratings provided by its audience. Without these kinds of factors, our analysis takes a more limited focus on ratings and overall movie attributes. Another constraint is with merging the datasets themselves: there will be some movie IDs that won't match precisely between MovieLens and Kaggle, even using the linking file. This excludes some subset of movies from the final merged dataset, potentially making our findings less inclusive.