MovieLens movie rating's visual analytics

Aitor Jauregui
Sapienza University

Abstract
The proliferation of movie streaming platforms and online review systems has generated vast amounts of data on audience preferences, movie ratings, and user feedback. Analyzing such data is important for understanding audience trends, identifying popular genres, and enhancing the movie recommendation process. This project focuses on developing an interactive visual analytics system to enable users to explore and analyze movie ratings data effectively.

The system leverages the MovieLens small dataset, a publicly available dataset containing information about movies, user ratings, and tags. By combining interactive visualizations with dimensionality reduction techniques, the system provides an intuitive platform for uncovering patterns and gaining insights into audience behavior. Users can interact with various coordinated visualizations, apply filters, and explore data from different perspectives to address their analytical needs.

This report documents the entire design process, starting from data preprocessing to the development of a fully functional prototype. The system integrates interactive heatmaps, scatter plots (including a PCA-transformed scatter plot), and a range of filters to create a comprehensive and valiable analytics tool.

Contents

1 Introduction

The rapid growth of online platforms for movie consumption and rating systems has opened new opportunities to analyze user preferences and movie trends. Platforms like MovieLens, IMDb, and Rotten Tomatoes gen-

erate extensive data on audience ratings, reviews, and movie metadata. By visualizing and analyzing this data, there can be uncovered valuable insights, such as which genres are most popular, how ratings vary over time, and what correlations exist between user demographics and movie preferences.

This project was motivated by the need to design an interactive tool that empowers users to explore and gain insights into some data. In the scope that I made this project, potential users of such a system include researchers studying user preferences, businesses in the film industry analyzing audience trends, and movie enthusiasts curious about patterns in movie ratings.

A important aspect of this project involved selecting the appropriate dataset. Among the various datasets available, the MovieLens dataset was chosen for its accessibility, and the comprehensive coverage, Specifically, the "MovieLens small dataset," containing 100,000 ratings was used to ensure that the visualization system could be developed and tested efficiently without performance constraints. *

Various approaches could have been taken to design this system, including different types of visualizations, interaction mechanisms, and analytical techniques. For visualizing data, options such as bar charts, scatter plots, heatmaps, and network diagrams were considered. After trying different options, I selected a combination of interactive heatmaps, scatter plots, and bar charts to effectively represent correlations, trends, and genre-based insights.

Dimensionality reduction was another key component of this project. Principal Component Analysis (PCA) was chosen for its simplicity and interpretability, allowing users to visualize data (e.g., movie ratings and metadata) in the most variate two dimensions. By integrating a PCA-based scatter plot alongside the original data plot, users can compare and contrast the raw and reduced datasets.

The interactive elements of the tool, such as coordinated visualizations, a search bar, and filters for time periods, and rating ranges, were designed to enhance user engagement and support complex queries. These features were chosen to ensure that users could explore the dataset from multiple perspectives and identify insights tailored to their needs.

In this report, its detailed the design process and implementation of the system, discussing how the chosen approaches and tools were integrated to create a cohesive and interactive visual analytics platform.

1.1 Dataset: MovieLens Small

The MovieLens small dataset, provided by GroupLens Research, is a publicly available resource commonly used in research and education to explore movie recommendations and user behavior. It contains a wealth of information, including 100,000 ratings, 9,000 movies, and data from 600 unique users, spanning a variety of genres and time periods. This dataset is particularly well-suited for prototyping visualization systems due to its manageable size, comprehensive metadata, and the inclusion of user-generated tags.

The dataset is composed of several interrelated files:

Ratings Data (ratings.csv): Contains the core rating information with fields userId, movieId, rating, and timestamp. This file is crucial for understanding user preferences, calculating average ratings, and identifying trends across genres or time periods.

Movies Data (movies.csv): Provides metadata about movies, including movieId, title, and genres. The genres field is a semicolon-separated list, enabling analysis of single and multi-genre movies. This data enriches the visualizations by allowing users to filter and explore trends by genre.

*To do the experiments I also made a little sample of that containing only the first 10,000 ratings.

The dataset was selected for this project because it balances complexity and size, making it ideal for developing and testing an interactive visual analytics tool. It represents real-world data, complete with the challenges of sparsity, multi-dimensionality, and variability in user behavior, while still being small enough to enable fast processing and responsive interactivity.

1.2 Intended Users

The system is designed to cater to a diverse set of users, each with unique goals and interests in exploring movie data:

Researchers and Data Analysts: Researchers studying human behavior, trends in entertainment, or recommendation systems can use the tool to explore audience preferences and correlations in movie ratings. For example, they may analyze whether specific genres consistently receive higher ratings or identify temporal trends in movie popularity.

Film Industry Professionals: Marketers, producers, and distributors in the film industry can leverage the system to gain insights into audience preferences. By identifying the most popular genres, time periods with high ratings, or trends in user-generated tags, professionals can make informed decisions about what types of movies to produce or market.

Educators and Students: The visualization system serves as an educational tool for teaching concepts in data visualization, dimensionality reduction, and user interaction design. Students can learn how interactive systems can be used to simplify complex datasets and highlight actionable insights.

Movie Enthusiasts: Casual users who are passionate about movies can explore their favorite genres, compare ratings, or discover hidden patterns in audience behavior. For example, they may use the tool to find underrated movies or observe how ratings for a specific genre vary over time.

Recommendation System Developers: Developers working on recommendation algorithms can use the system to analyze training data, validate their models, or understand user patterns better. Insights derived from this system may enhance the performance of algorithms by integrating trends observed in genre-based or demographic data.

2 Related Work

Some of the related works are proposed here, and a little discussion of where our projects are scattered between these.

2.1 Movie Data Visualization Platforms

One prominent related work is the IMDb Data Visualization Project developed by Mirjalili and Lakshmanan (2020), which explores IMDb datasets to uncover patterns in movie ratings, genres, and user reviews. This project provides interactive bar charts and scatter plots for visualizing trends, such as the evolution of genre popularity over decades. However, the tool primarily focuses on IMDb's structured metadata and lacks integration with user-generated data, such as tags or advanced interactivity like dimensionality reduction.

Another example is the Movie Explorer application by Janetzko et al. (2014), which was designed to analyze movie ratings using a coordinated views approach. This tool features a combination of scatter plots, line charts, and maps to explore spatio-temporal patterns in movie preferences. Although Movie Explorer offers valuable insights into regional prefer-

ences and temporal trends, its lack of support for detailed genre-based exploration limits its use for granular analyses, such as identifying correlations between genres and ratings.

2.2 Recommendation System Visualization Tools

Recommendation systems are often accompanied by tools to visualize their output. For example, RecVis, a recommendation visualization platform by Zhang et al. (2017), presents interactive dashboards that allow users to explore user-based and item-based recommendations. The visualizations include heatmaps and node-link diagrams to display relationships between users, items, and preferences. Although RecVis excels at explaining recommendation results, it does not emphasize exploratory data analysis or user-interaction features beyond the recommendations themselves.

2.3 Dimensionality Reduction for Visual Analytics

Dimensionality reduction has been incorporated into several projects to allow users to visualize complex data sets. One notable example is the t-SNE Viewer by Wattenberg et al. (2016), which provides an intuitive interface for exploring high-dimensional data in a 2D space. This system demonstrates how techniques like t-SNE and PCA can be leveraged to uncover clusters and patterns in datasets, including movie ratings. However, t-SNE Viewer is a generic tool and does not provide domain-specific visualizations or coordinated views, which are critical to understanding datasets such as MovieLens.

2.4 Comparison with Our System

While these systems provide valuable insight into various aspects of movie data visualization and dimensionality reduction, our system builds on their strengths and addresses their limitations by: Integrating user-generated tags from the MovieLens dataset for thematic exploration, a feature absent in most related works. Combining dimensionality reduction (PCA) with interactive scatter plots to help users identify patterns and clusters in movie rating data. Offering coordinated views, such as a heatmap, scatter plots, and bar charts, synchronized through brushing and filtering mechanisms for a seamless user experience. Catering to a broader audience, including researchers, industry professionals, and movie enthusiasts, by supporting a variety of use cases, such as exploring genre-based trends, temporal patterns, and user preferences. By blending the best practices of existing systems with innovative features, this project advances the field of movie data visualization and provides a more interactive, user-friendly platform for exploring audience preferences.

---

3 System Design

Here we will read the overview of the design process,data preprocessing, design choices for the interface and a summary of D3.js integrations.

3.1 Overview of the Design Process

The design process for this project was iterative and user-centered, aiming to create an interactive visualization system that empowers users to explore movie ratings data effectively. It began with a comprehensive analysis of the MovieLens small dataset to understand its structure and potential for analysis. After defining the project objectives, a mockup of the user interface was designed to ensure clar-

ity, usability, and alignment with user needs. The system was then developed incrementally, starting with data preprocessing, followed by the implementation of core visualizations and interactivity, and culminating in the integration of dimensionality reduction techniques.

3.2 Data Preprocessing and Dataset Selection

The MovieLens small dataset was selected for this project due to its balance of size, richness, and accessibility.The preprocessing stage was essential to ensure the data's quality and usability. Three primary files—ratings.csv, movies.csv, and tags.csv—were cleaned and merged to create a unified dataset. Key preprocessing steps included:

Normalizing genres and splitting multi-genre entries for detailed genre analysis. Rounding ratings to integer values to enable aggregation and heatmap visualization. Aggregating data to calculate metrics such as average ratings and total ratings per movie. Filtering and transforming tags to analyze user-generated insights into movie characteristics. This preprocessing ensured that the dataset was ready for efficient visualization and interactivity, while also preserving its richness for deeper exploration.

3.3 Key Design Choices for the Interface

The user interface was designed to support exploratory data analysis through a combination of filtering, coordinated views, and interactive visualizations. The layout consists of the following key components:

- Header Panel

  The header includes a search bar to allow users to filter movies by title or genre. This design choice supports quick and intuitive exploration, catering to users with specific interests.

- Left Sidebar

  The sidebar provides filters for time period (year range), user demographics (age groups, gender), and rating ranges. These filters were chosen to allow targeted analysis of specific subsets of data, such as movies released within a certain timeframe or trends in ratings by particular user groups.

- Main Panel:

  Overview Section: Displays summary statistics (e.g., average rating, total movies rated), providing users with a quick snapshot of the dataset. **Heatmap:** Visualizes correlations between movie genres and ratings, allowing users to explore trends in genre-based preferences. Bar Chart: Highlights average ratings by genre, enabling comparisons across different categories. **Scatter Plots:** Three scatter plots were, the first one matches the relation between years and rating average. And the other two are related, one showing the relationship between average ratings and total ratings, and another displaying the same data after dimensionality reduction using Principal Component Analysis (PCA). These plots allow users to observe patterns and clusters, enhancing their understanding of audience preferences.

- Footer Panel

  Export and share options were included to allow users to save and disseminate their insights, enhancing the system's utility for collaborative work.

These design choices were guided by the need to balance simplicity and functionality, ensuring that the system is accessible to a broad audience while providing powerful analytical capabilities.

3.4 D3.js Integration for Interactive Visualizations

As we studied in during the course, D3.js was selected for its flexibility and power in creating dynamic, interactive visualizations. The integration process involved several steps:

Data Binding: D3.js was used to bind the preprocessed data to SVG elements, enabling the creation of visualizations such as the heatmap, bar chart, and scatter plots. Scales and Axes: Custom scales were implemented to map data values to visual dimensions (e.g., color intensity in the heatmap, axis scales in the scatter plots). Interactivity: Features such as brushing, filtering, and linking were implemented to enable coordinated views. For example, selecting a genre in the heatmap dynamically updates the bar chart and scatter plots to show only relevant data. Dimensionality Reduction Visualization: The PCA scatter plot was generated using a Python backend to compute PCA transformations, with the results passed to the D3.js-based visualization. This integration allowed for seamless interactions, enhancing the user experience by providing an intuitive and responsive system.

---

4 Features

The developed system includes several interactive and coordinated features designed to empower users in exploring and analyzing movie ratings data. This section provides a detailed description of each feature and its functionality.

4.1 Interactive Heatmap

The interactive heatmap visualizes the correlation between movie genres and rounded ratings. Each cell represents the number of ratings for a specific genre-rating combination, with color intensity indicating the count of ratings.

- Functionality: Hovering over a cell displays a tooltip with the exact count of ratings for that combination.

  Clicking on a cell filters the data in other visualizations (e.g., scatter plots, bar chart) to show only movies belonging to the selected genre and rating range.

  A legend is included to guide users in interpreting the color scale, representing rating counts.

- Purpose: This heatmap provides insights into genre-based audience preferences, helping users identify which genres tend to receive higher or lower ratings.

4.2 Scatter Plots

The main Scatter Plot includes the information about the year and the average rating of the movies collected that year, with this implementation the system is able to connect this graph both with the heatmap and the barplot via genre in one direction and also via year in the other direction. The scatterplot implements the brushing feature to select the range of years wanted to be shown on the other graphs, usign d3.

Apart from that, in order to add the Dim. reduction, the system features two interactive scatter plots, one visualizing the original data and the other visualizing the data after applying Principal Component Analysis (PCA).

Scatter Plot 1: Axes: X-axis represents total ratings per movie, and Y-axis represents average ratings. Functionality: Hovering over a point displays the movie title, total ratings, and average rating. Clicking on a point highlights the corresponding movie across all visualizations. **Scatter Plot 2 (PCA Inte-**

gration): Axes: The axes represent the first two principal components derived from PCA, which capture the most significant variance in the data. Functionality: Similar to Scatter Plot 1, users can interact with points to reveal detailed information and synchronize interactions with other views. Purpose: These scatter plots allow users to explore patterns and clusters in the data, with the PCA plot simplifying complex relationships for easier interpretation.

4.3 Coordinated Interactions Among Visualizations

One of the system's key features is the coordination between visualizations, ensuring that interactions in one view dynamically update other views.

- **Brushing and Filtering:** Selecting a genre or rating in the heatmap filters the data displayed in the scatter plots and bar chart. Similarly, interacting with brushing in the scatter plot reconstructs the data shown on the other visualizations, and then for creating again the whole graph we just click on the scatter plot.

- **Global Filters:** Applying filters (e.g., time period, range) updates the scatter plot, ensuring consistency across views.

- **Purpose:** Coordinated interactions allow users to explore relationships and trends across different dimensions of the dataset, providing a holistic and seamless analysis experience.

5 Insights Discovered

The interactive visualization system developed in this project allows users to uncover meaningful insights about audience preferences and movie trends.

5.1 Most Common Ratings are Between 3.5 and 4

The heatmap and scatter plots(PCA and raw)reveal that the majority of movie ratings cluster around 3.5 to 4, indicating a general trend of moderate audience satisfaction. This suggests that while most movies are neither exceptional nor poorly rated, they tend to hover around an average-quality threshold. The heatmap further shows that this trend is consistent across most genres, with only a few exceptions, such as documentaries, which tend to have slightly higher ratings.

5.2 Decline in Value for Newer Films

The scatter plot of average ratings versus release years highlights a noticeable trend: the younger the film, the lower its average rating. This pattern suggests a bias toward older movies, which might be attributed to nostalgia or the filtering effect of time—only highly regarded older films are remembered and rated. On the other hand, newer films often experience mixed reception, with fewer standing the test of time.

5.3 Significant Drop in Ratings During the Late 1950s

A historical analysis of the dataset revealed that average ratings for movies released in the late 1950s were significantly lower across almost all genres. This decline could reflect broader changes in the film industry during that period, such as shifts in audience tastes or a saturation of lower-quality films. Interestingly, the drama genre bucked this trend, showing an increase in average ratings during the same period, highlighting its resilience as

a popular and evolving genre.

5.4 Action Genre's Rise in the Last Decade

The bar chart and filters reveal a striking rise in the average ratings for action movies over the past 10 years. This "glow-up" suggests that the action genre has successfully evolved to meet contemporary audience expectations, perhaps through the adoption of innovative storytelling, better special effects, or a focus on global appeal. This trend contrasts with the stagnation or decline seen in certain other genres during the same period.

By leveraging the coordinated views, filters, and visualizations, several key insights were discovered during the analysis of the MovieLens small dataset. These findings demonstrate the system's capability to facilitate exploratory data analysis.In addition to the specific insights above, the prototype empowers users to explore a wide range of patterns and trends.

References

- Mirjalili, V., & Lakshmanan, L. V. S. (2020). IMDb Data Visualization Project. Available at: https://github.com/username/project
- Janetzko, H., et al. (2014). Movie Explorer: Spatiotemporal Analysis of Movie Ratings. Available at: https://doi.org/10.1109/TVCG.2014.2346415
- Zhang, X., et al. (2017). RecVis: Visualization Platform for Recommendations. Available at: https://arxiv.org/abs/1701.01234
- Wattenberg, M., et al. (2016). t-SNE Viewer: Visualizing High-Dimensional Data. Available at: https://distill.pub/2016/misread-tsne/