

Title: The Movies Dataset Visualization (Group H)

Names of all project participants:

Eunji Kim ek3223@columbia.edu

Dustin Nguyen dn2519@columbia.edu

Xintong Tang xt2249@columbia.edu

Frances Yang zy2479@columbia.edu

Abstract: This project seeks to understand if there is any relationship between a movie's genre, budget, revenue, and ratings. The dataset consists of movies released on or before July 2017, and data points include movie name, cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. The visualizations that are planned to be produced are ratings by budget and genre, allocated budget per genre, and a map of production country with amount of revenue generated. We would also like to explore whether certain directors generate more revenue or call for a higher budget, and we would also like to see if there is a minimum or maximum rating that occurs from a certain budget.

Techniques: ggplot2, NLP text analysis, interaction, geospatial map

Data Description:

Dataset: <https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset>

These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include genres, cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

The dataset contains the following files:

- **movies_metadata.csv**: The main Movies Metadata file. Contains information on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.
- **keywords.csv**: Contains the movie plot keywords for our MovieLens movies. Available in the form of a stringified JSON Object.
- **credits.csv**: Consists of Cast and Crew Information for all our movies. Available in the form of a stringified JSON Object.

- **links.csv**: The file that contains the TMDB and IMDB IDs of all the movies featured in the Full MovieLens dataset.
- **ratings.csv**: 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

Visualizations:

Word cloud: 1. what kinds of words are used to describe different types of movies from genres column (topic analysis on genres column)

Line chart: 1. Time series of revenue generated by types (genres) of movies, 2. Sheer number of types of movies are made over time

Bar chart: 1. Comparing movies by genres to see which genres bring in the most revenue and/or have the highest ratings, 2. Runtime of movies / genre/ budget/ revenue

Scatter plot: Comparing revenues by budget, different colors for each genre

Map: Choropleth for production country and movie revenue

Pie: Out of all the money spent on movies in the year - % budget per genre

Plotly: 1. Use ratings to compare movies - click rating we want, 2. Bubbles for directors for most revenue