LetterBoxd Film Analysis

Team: Clayton Brittan

Description

This project is to find interesting and potentially useful patterns and make predictions.

- Do movies with certain themes outperform others?
- What Genre of films gets the best ratings?
- What actors/directors are the best in each genre?

This project also aims to see if you can predict a films LetterBoxd rating based on it's features. My hope with this project is to find if there are any trends with how movies are rated. This project also aims to demonstrate trends in films over time and across genres and countries.

Prior Work

Another user on Kagle called Surume used this dataset to plot the distribution of movies by country, and the overall distribution of movie genres. Their work can be found at this link

https://www.kaggle.com/code/suntzunami/what-the-world-be-watchin

However, Surume's analysis only shows the popularity and distribution of producers and genres and does not address underlying trends in the data.

Dataset

The data set is hosted on kaggle: https://www.kaggle.com/datasets/gsimonx37/letterboxd/data

The dataset includes lots of information including movie posters for all 900k+ films.

The total file size is 24Gb, however since most of that is from movie posters we will be not using those. The dataset is further divided into CSV files and for our purposes we will be mainly using the movies, themes and genres csv files.

The dataset contains entries for every film posted on LetterBoxd and their accumulated score out of 5, as well as all relevant actor, director, producer, genre, release, language, country of origin, etc information.

The Dataset has been downloaded onto my computer and will be read using python and sql.

Proposed Work

Data Cleaning:

Since the dataset is split into multiple csv files I will need to combine them all into a merged dataset along their movie id key.

Data entries with missing values across multiple columns may need to be dropped for cohesion and to build predictive models

Data Preprocessing:

Create proper data types for movie entries, by combining similar rows

Data Reduction:

Determine relevant columns and remove irrelevant ones

Tools

-Python:

This will be the language used for the project and will include many internal libraries including plotting functionality (pyploy, matplotlib, seaborn), as well as math functionality and data wrangling tools (numpy, pandas)

-GitHub/Git:

Github will host the project and automatically keep track of any changes and updates made

-SQL:

Used to store and manage the dataset

-Textbook and class lectures:

Evaluation

- To evaluate the results we will be using the ‘rating’ column in the movies csv file as our target variable.
- We will use visualization methods to see how genres, themes and duration affect a films rating.
- We will also use regression analysis to determine how much these factors affect the score.
- We will utilize machine learning models to try to predict film scores and use CV-validation and precision and accuracy measures to evaluate our models efficacy.
- Use visualization techniques to see what genres actors and producers are most common in and how it affects their rating