Title: Movie Mining

Team Members:
Caleb Stillman
Graydon Sinclair
Stephen Tynan?

Description:
Group 8's Movie Mining project intends to sift through a dataset comprised of data on over 950,000 films. Some interesting questions we hope to answer include things such as predicting the runtime of a film based on its genre; identifying patterns regarding the most popular film genres in a particular country, which in turn could be used to predict what genre a country's next film is most likely to be; predicting the duration of a movie based on country of origin, genre, or date of release; or predicting the popularity of a movie within the Letterboxd community based on the actors, genre, or crew involved in a movie. A stretch goal that would require bringing in additional datasets might be finding whether there are stronger correlations between crew size and film budget or between actors and budget.

Prior Work:
I recall from the Information Visualization class a database that made use of similar information to the dataset we're using. However, instead of performing mining to discover interesting information from the dataset, the tool created an interactive visualization for discovering interesting correlations. A great many sites and projects perform EDA on datasets similar to ours and have uncovered interesting results about existing correlations or for use in predictions around films. Sentiment analysis is another way in which similar film datasets have been used, specifically around the ratings data.

Datasets:
We are using the Letterboxd dataset found here on Kaggle:
https://www.kaggle.com/datasets/gsimonx37/letterboxd/data?select=releases.csv.
Graydon has the dataset downloaded on his machine.

Proposed Work:
We don't expect to have to do much, if any, data cleaning on the Letterboxd dataset from Kaggle - fields appear complete and in a usable format. We do expect that we'll need to do some data preprocessing and integration - among the fields that this dataset is missing that we might find useful are each movie's budget and box office take, as well as critical reception/ratings. To address the former, we've talked about finding a dataset from IMDB, and for the latter, we've discussed finding a supplemental dataset with Metacritic ratings which would need to be integrated.

List of Tools:

- Python and associated data libraries (Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, etc.)
- Orange (open source data mining tool: https://orangedatamining.com/)

Evaluation:

Some means of evaluation that we might make use of include accuracy (for classification tasks such as predicting film genres), MSE (for regression tasks such as predicting box office revenue), precision/recall/F1 score (for evaluating success of model predictions). Some existing tools we are considering using for evaluation of our results include visualization tools such as Matplotlib and Seaborn, as well as Orange's integrated visualization tools.