

**Title:** Movie Mining

**Team Members:**

Caleb Stillman  
Graydon Sinclair  
Stephen Tynan?

**Description:**

Group 8's Movie Mining project intends to sift through a dataset of over 950,000 films. *What interesting questions do we intend to answer?*

**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.

**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 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:**

**Evaluation:**