**Problem Statement/Motivation:**

The goal of this project is to mine data from a dataset with information on over 950,000 films. Runtime has been identified as the primary feature we wish to explore. Some interesting questions we hope to answer from examination of the data in relation to this feature include such things as predicting the runtime of a film based on its genre, country of origin, or date of release. We also hope to identify a “sweet spot” for runtime of a film such that the film’s rating is maximized, and examine how this may differ from country to country or genre to genre. This knowledge could be applied within the film industry to produce movies with a greater likelihood of positive public reception, and therefore, likely an increased box office take. In a similar way, we hope to identify the runtime ranges which are too short or too long to result in good ratings for a film. Again, this knowledge could be used by film industry professionals to produce films more likely to become successful.

**Literature Survey:**

A survey of the existing literature suggests that some work has been done to examine the impacts of runtime on other facets of films, but that it’s been fairly minimal. A piece in Cognitive Research Journal speaks to runtime tangentially as it relates to trends in the pacing of films over film history, but the author examines only 210 movies and makes an arbitrary cutoff at the two and half hour mark for which movies are to be included in his survey. Another piece from Film and Digital Media does focus on runtime, but the piece lacks a lot of depth, and it is our belief that we could go further with our analysis. A last piece from IMDB does look to determine how the success of films is measured, but the piece does not speak to runtime at all, so again, we might be able to introduce a new angle and further the research in this area.

The evolution of pace in popular movies. <https://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-016-0029-0>

Time Isn’t Money: Does Runtime Affect a Film’s Box Office. <https://filmanddigitalmedia.wordpress.com/2017/10/19/time-isnt-money-does-runtime-affect-a-films-box-office/>

How is the success of films and TV shows measured? <https://pro.imdb.com/content/article/entertainment-industry-resources/featured-articles/how-is-the-success-of-films-and-tv-shows-measured/GLFTC8ZLBBUSNTM3>

MovieLens. <https://movielens.org/>

**Proposed Work:**

Our proposed work begins with EDA, including data cleaning and preprocessing concurrently. In regard to data cleaning, we’ve noticed that certain features, such as release date, need some work (for instance, there are a number of dates between 2025-2099 which will need to be corrected or removed). In our earlier topic presentation we mentioned bringing in external datasets for ratings and box office take from Metacritic and IMDB. We have subsequently decided to abstain from these additions - the primary dataset from Letterboxd does include ratings; and we determined that most analyses into features that determine success of a film already analyze budget vs. box office (as seen in the article from IMDB in our literature survey), and so we would be furthering the research less by focusing on this aspect, as it is already well-covered. The result of these decisions is that our proposed work includes much less data preprocessing and integration. That said, some integration is still required, as the Letterboxd dataset is comprised of 10 separate CSVs that share a primary key and will need to be merged together on that key as we build our dataframes. Once the EDA, cleaning, and preprocessing are completed, we will move to the evaluation stage, wherein we will be plotting our primary features to discover correlations. Some of this data (for instance, genre) will need to be one-hot encoded prior to plotting. We can then perform such evaluation tasks as regression analysis of MSE to predict runtimes and so forth.

**Data Set:**

Our primary dataset provides data from the website Letterboxd. The dataset was found on Kaggle at this link: <https://www.kaggle.com/datasets/gsimonx37/letterboxd/data?select=releases.csv>.

Graydon has the dataset downloaded on his machine. The dataset, as previously mentioned, contains data on over 950,000 films. It is 24.71 GB, and due to this size, we have decided to rule out the use of the data mining tool Orange - despite its inbuilt visualization tools and friendly interface, one of its known drawbacks is that performance degrades on larger datasets such as this one. The dataset includes 10 csv files which share “id” as the primary key, but which focus on different facets of the movie data - one focuses on genre, one on country of origin, one on actors, one on studios, etc. The dataset provides current data up through the present date, meaning we can provide the most up-to-date evaluation of the runtime feature compared to the current literature.

**Evaluation Methods:**

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 runtime), precision/recall (for evaluating success of model predictions), and F1 score for tasks like genre classification. Some existing tools we are considering using for evaluation of our results include visualization tools such as Matplotlib and Seaborn to produce visual plots to aid in evaluation. We might graph variance vs. covariance to help us visualize spread of ratings vs. runtime or other features.

**Tools:**

We intend primarily to make use of Python and associated data libraries for our mining - Pandas for producing dataframes for ease of manipulation of the data, NumPy to aid in calculations, and Seaborn or Matplotlib to produce our plots and other visualizations for our evaluation needs. We have decided not to use Orange as previously stated.

**Milestones:**

Identification: primary variables include runtime (prime), genre, ratings, country of origin (and possibly language in association with this), studio. Completed by 10/1.

EDA (October):

Cleaning (release dates, etc.) by 10/15.

Preprocessing - dataset comprised of 10 csvs that need to be merged (all share primary key of “id”, should be easy. One-hot encoding (many movies have multiple genres) by 10/31.

Analysis (November):

Apply evaluation methods and create visualizations.

Finalize Results (Late Nov, early Dec):

Wrap up evaluation, synthesize results, produce final presentation.