Box Office Analysis
Brian Mann

Subject Area

This project will be taking a look at movies - how much they gross, how audiences and critics rate them, and what factors combine to create a blockbuster.

Data Sources

- **Flat file -** https://www.kaggle.com/datasets/andrezaza/clapper-massive-rotten-tomatoes-movies-and-reviews?select=rotten_tomatoes_movies.csv
    - A CSV file containing over 140,000 films from the Rotten Tomatoes movie database containing information including title, director, audience score, critics score, rating, etc.
- **Website -** https://www.boxofficemojo.com/chart/top_lifetime_gross/
    - A website containing the top 1,000 films ranked by total lifetime box office earnings - both worldwide and domestic.
- **API -** https://github.com/cinemagoer/cinemagoer
    - A python module that functions essentially as a free version of the IMDB Pro API, with options to get, set and search for data on the IMDB database, albeit at a slower speed.

Relationships

At a minimum, each dataset can be connected by movie title. However, due to some movies containing variations on how they are presented (hyphens, colons, roman numerals, sequels, remakes, etc.), there might not always be an exact 1-to-1 link between movie titles. In that case, further efforts will have to be made to use other information such as box office earnings, year, director, etc. to merge the datasets together. Additionally, each of these films should contain a unique numeric identifier on IMDB's database that could also serve as a key, although these are not shared by Rotten Tomatoes.

Project Description

The overall approach for this project will be to start with the top 1,000 highest grossing films dataset derived from the BoxOfficeMojo website, then expand upon it until all of the relevant information from Rotten Tomatoes and IMDB are added as additional columns. The website data merely contains the film titles, year made, and

information on domestic and international earnings. In contrast, the Rotten Tomatoes and IMDB data sets contain all sorts of information on ratings, cast, directors, reviews, runtime and much more. While little work will need to be done to clean up the website data, the bulk of the work will be in cleaning and integrating the data from the other two sources to flesh it out.

The biggest challenge will be in trying to join the datasets using movie titles. As discussed in the 'Relationships' section, the three datasets don't seem to be using the same naming conventions when it comes to special characters or subtitles. Due to this, more clever approaches will have to be made to merge the datasets using additional indicators such as partial titles, year, and earnings. Another challenge will be pulling the correct data from the IMDB API. In order to get the correct film, a search will need to be made based on the title and year. This could become difficult if there was also a TV show or other adaptation made at the same time that would throw off the results. This will also take a large amount of time to query all 1000 movies, so it will be imperative that the code works smoothly.

There are a couple of key ethical implications that might need to be considered when conducting this project. One is the intellectual property of IMDB and Rotten Tomatoes - are we compliant with the rights of these companies? The answer is yes - all of the data collected is free and publicly accessible on each of their respective websites. As far as can be assessed from the github, the Cinemagoer module does not dodge any of the licenses necessary for using the IMDB Pro API. Another topic to consider is that of the film ratings - are there biases in these assessments? Unfortunately, there is no further information on the demographics of reviewers to adequately come to any conclusions on how this could lead to biased reviews. Some audience or critic scores may be lower due to perceived cultural biases as opposed to the quality of the film. This will not be factored into our analysis.