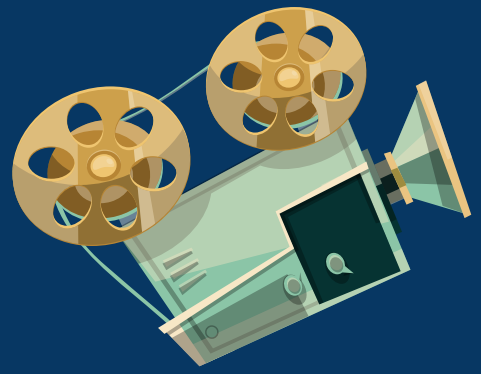
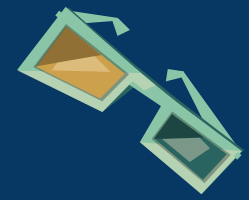
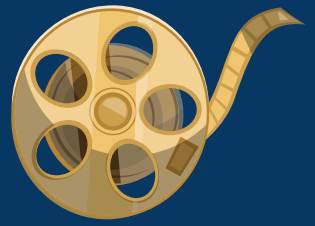


Non-English Film Engagement on IMDb



CSPB 4502



Emily Carpenter,
Janet Matthews-Derrico,
Marcus Almanza,
Yoshie Bell-Souder

Non-English Film Engagement on IMDb

Questions:

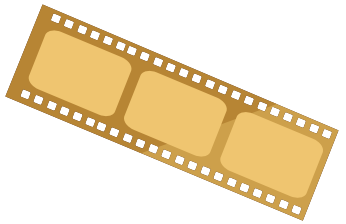
- Do non-English films have fewer reviews and ratings, regardless of their average rating?
- Do certain genres of movies do better (better could be ratings or gross/net profit) when released at certain times of the year (by month)?
- Do certain languages do better when it comes to ratings/reviews overall?
- Does the number of language speakers impact the average number of ratings per film for that language?
- Do international films have higher ratings due to fewer view counts?
- Which films are “underrated gems,” that have high ratings, but relatively low view counts?

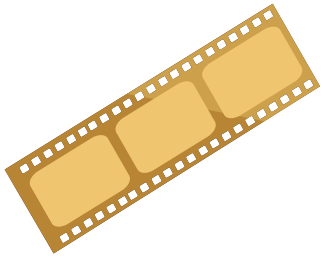
Prior Work: What prior work has been done on your idea

- Exploratory data analysis on IMDb movie titles searching for overall trends in popularity: https://www.kaggle.com/code/slayomer/eda-on-imbb-film-dataset
- Exploratory data analysis specifically on Netflix titles using IMDb dataset: https://www.kaggle.com/code/keswanirohit/netflix-visualization-and-eda
- Movie recommendation system created using the IMDb dataset: https://www.kaggle.com/code/jasonlei0420/ds5230-movie-recommendation-system

Dataset

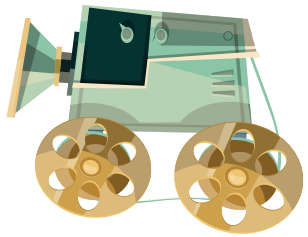
- **IMDb datasets:** title.basics.tsv.gz, title.ratings.tsv.gz, title.episode.tsv.gz
- **Dataset source:** data is provided by IMDb, the Internet Movie Database, and can be accessed here: https://www.imdb.com/interfaces/
- **Group access:** downloaded by Emily and Marcus





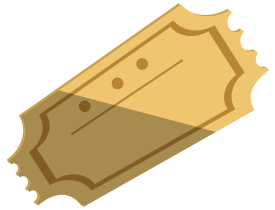
Proposed work: what do you need to do?

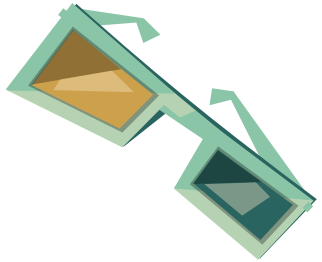
- **Data Cleaning:** Remove unnecessary rows, n/a data, entries without enough relevant information (e.g. “lost films” that have entries but cannot be viewed)
- **Data Integration:** Combine multiple IMDb datasets into one that can be used for data mining
- **Data Reduction:** Remove unnecessary attributes such as television shows, cast and crew information, etc
- **Data Processing:** Search for interesting patterns, determine whether correlations meet minimum support and confidence
- **Data Visualization:** Accurately display results in a way that is easy to understand

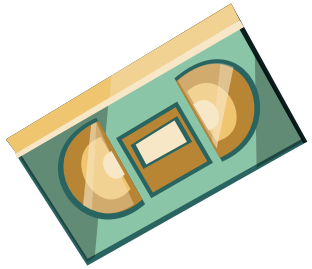


List of tool(s) you intend to use

- Github
- Discord
- Google Slides
- Tableau
- Python
  - Pandas
  - Numpy
- Matplotlib







Evaluation: How you can evaluate your results

- **Statistical analysis:** Utilize clustering, regression, pattern mining, and more as we learn new techniques and become more familiar with the dataset
- **Accuracy checks:** Develop test cases and comparison sets to ensure accurate data processing
- **Data visualization:** Utilize graphs, plots, and other visual tools to evaluate patterns
- **Critical Evaluation:** Scrutinize results for misleading strong association rules