Project 2 Approval & Proposal

To get a dataset approved ...

1. Send your instructor a description of the data, including a link to the exact dataset or a place from where it can be downloaded with one click without registering (or anything like that).
· The data needs to be in a raw text file format like json or csv (or at least very easy to convert to that format).
· Report the size of the dataset or the subset you plan on looking at. This is **not** a big data project.

Link to Dataset:

https://github.com/UC-Berkeley-I-School/mids-w200-project2-Aruna-GregREPO-

Sourced from: https://www.kaggle.com/rounakbanik/the-movies-dataset

Description:

IMDB movie dataset up until July 2017

The main Movies file has data on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.

2. Show that you've done some preliminary research and that there will be enough interesting questions to ask of the data. This should include column names and information about any missing data.

Columns:

['adult',
 'belongs_to_collection',
 'budget',
 'genres',
 'homepage',
 'id',
 'imdb_id',
 'original_language',
 'original_title',
 'overview',
 'popularity',

'poster_path',
'production_companies',
'production_countries',
'release_date',
'revenue',
'runtime',
'spoken_languages',
'status',
'tagline',
'title',
'video',
'vote_average',
'vote_count']

Fill Rates (percentage of filled values)

```
print("Fill Rates:")
movies.notna().sum() / len(movies) * 100
```

```
Fill Rates:

adult                    100.000000
belongs_to_collection      9.878363
budget                   100.000000
genres                   100.000000
homepage                  17.110617
id                       100.000000
imdb_id                   99.962607
original_language         99.975805
original_title           100.000000
overview                  97.901590
popularity                99.993401
poster_path               99.150958
production_companies      99.993401
production_countries      99.993401
release_date              99.808636
revenue                   99.993401
runtime                   99.428106
spoken_languages          99.993401
status                    99.815234
tagline                   44.898049
title                     99.993401
video                     99.993401
vote_average              99.993401
vote_count                99.993401
dtype: float64
```

Project Proposal

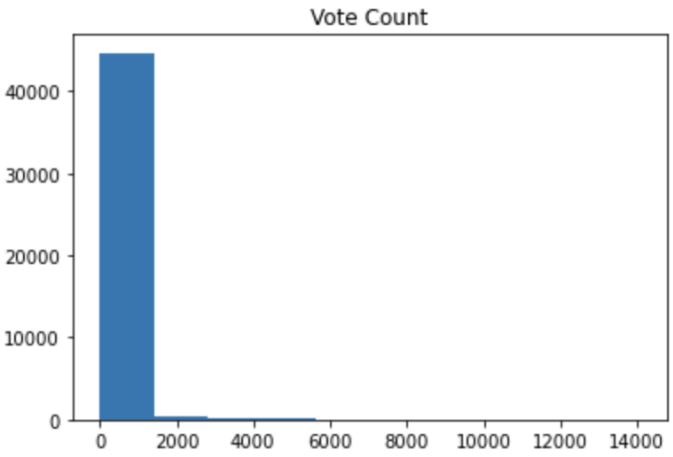
Team members: Greg Rosen, Aruna Bisht

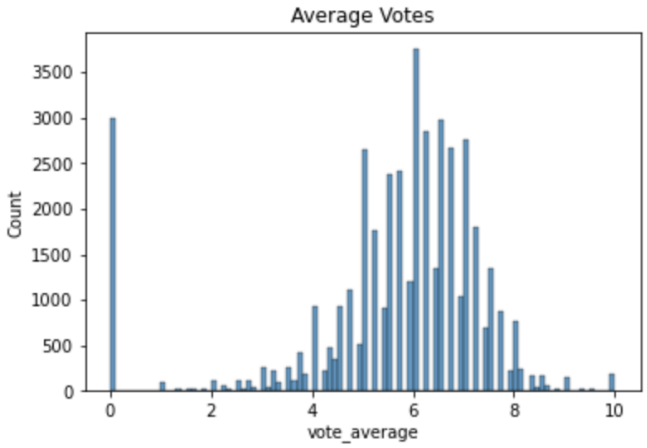
Name of GitHub repository: mids-w200-project2-Aruna-GregREPO-

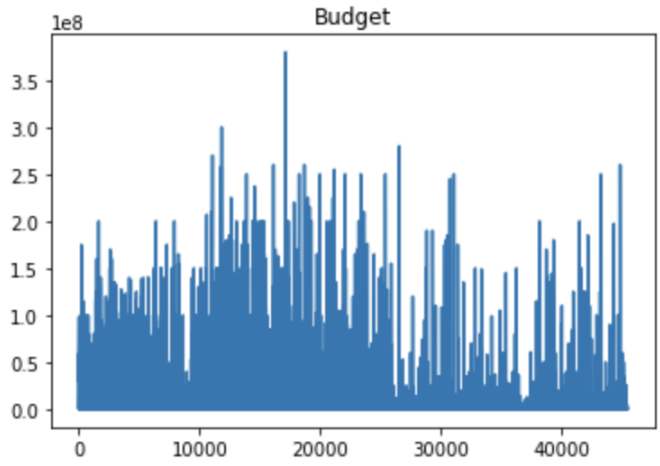
Primary dataset to analyze: IMDB movie dataset up until July 2017

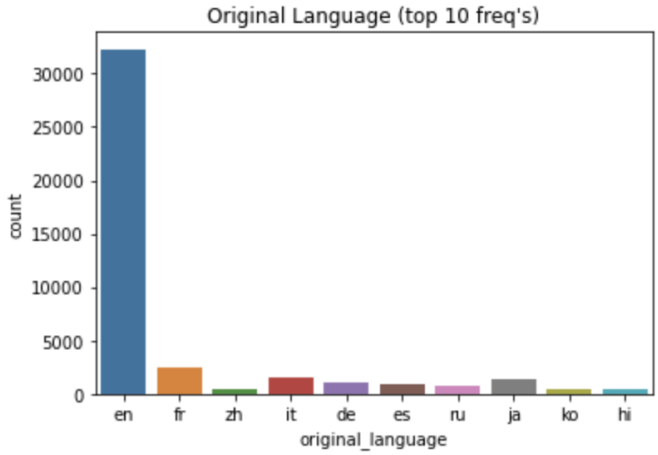
The main Movies file has data on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.

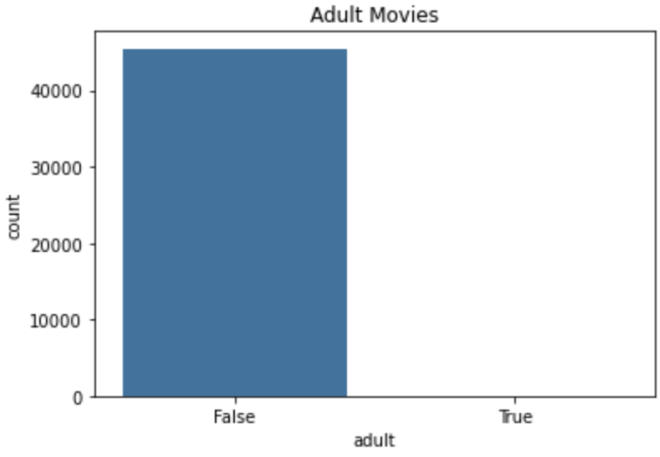
Initial plots, figures, tables:











Some variables to explore and insights expected to glean:

Some brainstormed questions, which include the attributes of interest (e.g. genre, budget):

- Which genres have the highest budget?
- What movie titles have the highest budgets per year?
- What are the differences in movie trends in the 1990s vs. 2000s?
- What production companies have the highest budgets?
- Which actors have the highest budget?
- What genres of movies get the highest ratings?
- Do ratings correlate with revenue generated?
- Which actors get the highest ratings?
- What are genre trends in different countries?

Supplemental datasets:

credits.csv - cast and crew information on each movie

What to cover in final report and how to organize it:

Theme: Entertainment is an integral part of our life -- almost everyone watches movies. But sometimes we make wrong selections based on popularity or marketing budget versus the actual ratings of the movie.

Also add in stats on how much society likes watching movies ($ spent/year). Can show trend of movie budgets by year.

- Do the highest grossing movies generate higher ratings?
- Do the most popular movies generate higher ratings?
- Do the most popular actors (top 30) generate higher ratings?
    - What is a “popular actor”?
        - Frequency of movies
    - Who are the most popular actors?
- Do the most popular actors (top 30) generate higher popularity?
- Compare results of popular actors’ relation to both ratings and popularity.
- Do the most popular production companies generate higher ratings?
- Do the most popular production companies generate higher ratings?
- Do the movies with the highest budgets generate higher ratings?