IMDB5000 pandas exercise

# Preliminaries

* Import the file as a dataframe. The index is the names of the movies.
  + note: when using read\_csv() use the option encoding=’utf-8-sig’.
* Remove duplicates (identical rows)
* Remove directors with less than two movies
* Remove movies which do not have at least three actors.

# Warm up

* How many languages are represented?
* What is the oldest movie?
* Who is the director with the highest average imdb score?
* Add a column called ‘profits’ with gross-budget and answer:
  + What is the most profitable movie and what is the biggest failure?
  + Who is the most profitable director?

# Simple questions

* How much money was spent on all the films each year (budget)?
* Which director makes the longest movies (on average)? Create a dataframe with this information, and another column with the number of movies.
* How many unique actors are represented in the file? Consider all three columns.
* How many movies were released every year? Show it on a graph. Now separate the answer to color and B/W films.
* Which is the most prolific director, who’s average number of movies per year is the highest? Now consider only directors with more than 5 movies.

# Difficult questions

* Which actor participated in the highest number of movies? Consider all three columns.
* How many movies of each genre were produced? Visualize it.
  + Challenge: How many movies of each genre were produced every year?
* Which pair of director-actor is the most frequent? Consider all three columns.