4 - Pandas - Lab Exercises

Create a new cell for each question. You will need the file `movie_dataset.csv` from Moodle. Whilst each question has a single correct solution/answer, there may be multiple ways to arrive to it.

Section 1

Exercise 1.1

Use Pandas to load the file `movie_dataset.csv`. Assign it to a variable called `movies`.

Exercise 1.2

Find out some basic information about `movies`. How many rows of data are there, how many columns? Are there columns that have missing data for some rows?

Exercise 1.3

What are the names of the columns between `vote_average` and `crew`?

Exercise 1.4

What are the first and last movies listed in the dataset?

Exercise 1.5

Sort the dataset by `budget`. Which is the film with the largest budget? By default Pandas will sort lowest to highest, but can you find a way to reverse this behaviour? Try Googling or looking at the Pandas documentation for the solution.

Exercise 1.6

What is the tagline for the film at index `3002`?

Section 2

Note: Make sure you still have your `movies` variable loaded. If not, repeat Exercise 1.1 before continuing.

Exercise 2.1

How many documentaries are in our `movies` dataset? Start by looking at the `genres` column.

Exercise 2.2

How many documentaries were originally in French? You'll need both the `genres` and `original_language` columns.

Exercise 2.3

How many films in our dataset had above average `budget` but below average `revenue`?

Exercise 2.4

Using the filtered data from Exercise 2.3, can you identify the movie with the highest `budget`, and the movie with the lowest `revenue`?