5 - 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

Drop all rows with any missing data from the `movies` dataframe, and assign it to a new variable called `movies_complete`. How many rows are there in `movies_complete`?

Exercise 1.3

Drop all rows where there is data missing in the `tagline` column only. Assign the result to a new variable called `taglines_df`. Bonus points if you can make it so that `taglines_df` only contains the columns `original_title`, `release_date` and `tagline`. You may need to refer to last week's material.

Exercise 1.4

Currently the `type` of data in the `release_date` column is `object` which implies that the dates are stored as strings. Can you transform the data in the `release_date` column so that you can easily find the earliest and latest release dates in the dataset?

Exercise 1.5

According to "Conspiracy Weekly", all movies that were supposedly directed by Ridley Scott were **actually** directed by a time travelling Justin Bieber. In the `director` column, replace all instances of `Ridley Scott` with `Justin Bieber` to correct this mistake in the data.

Section 2

Exercise 2.1

Which `director`s have the largest count of films in our dataset? Who are the top 5?

Exercise 2.2

Using `movies` can you group the data by director and then determine who has the highest **total** `revenue` across all their films?

Exercise 2.3

Using `movies` can you group the data by director and then determine who has the highest **average** `revenue` across all their films? How does this list compare with the list based on total revenue?

Exercise 2.4

Using `movies` can you group the data by director, and then create a dataframe that tells us each directors' `average_revenue`, `total_revenue`, `average_budget`, `total_budget` and `n_films`. Assign the result to a new variable called `director_performance` Afterwards experiment with sorting `director_performance` by different columns to see how it might change the ranking of the directors.

Exercise 2.5

Let's save our `director_performance` data so we can use it later, and perhaps share it widely. Let's also save our cleaned up and date transformed `movies` dataframe whilst we're at it in case we want to come back and do further analysis. Save each dataframe using the most appropriate file format for the job. Make sure you don't overwrite an existing file and ensure the filename ends with the correct file extension.