Homework 2: Load, Inspect, & Query Movie & Ratings Data

This assignment is designed to give you practice writing code and applying lessons and topics for the current module.

This homework deals with the following topics:

- The *pandas* module
- Loading data
- Inspecting data
- Querying data

The Assignment

In this assignment, you will analyze data about movies and ratings from the IMDB website and get to use functions from the *pandas* module for loading, inspecting and querying the data. For each question, there are clear instructions in each cell of the provided Jupyter Notebook file. Follow those instructions and write the code after each "# your code here".

We'll use nbgrader, a Jupyter Notebooks testing platform, to test whether each function implementation is correct. You can see the exact test we are running in the cell right below your solution.

About the Data

All of the data is contained within the "imdb.xlsx" file which contains 3 sheets:

- "imdb": contains records of movies and ratings scraped from IMDB website
    - There are 8 columns: movie_title, director_id, country_id, content_rating, title_year, imdb_score, gross, duration

1	movie_title	director_id	country_id	content_rating	title_year	imdb_score	gross	duration
2	The Shawshank Redemption	34	1	R	1994	9.3	28341469	142
3	The Godfather	33	1	R	1972	9.2	134821952	175
4	The Dark Knight	16	1	PG-13	2008	9	533316061	152
5	The Godfather: Part II	33	1	R	1974	9	57300000	220
6	The Lord of the Rings: The Return of the King	83	1	PG-13	2003	8.9	377019252	192
7	Pulp Fiction	85	1	R	1994	8.9	107930000	178
8	The Good, the Bad and the Ugly	98	2	Approved	1966	8.9	6100000	142
9	Schindler's List	103	1	R	1993	8.9	96067179	185

- "countries": contains the country (of origin) names
    - There are 2 columns: id, country

	id	country
1	1	USA
2	2	Italy
3	3	New Zealand
4	4	Japan
5	5	Brazil
6	6	Germany
7	7	France
8	8	UK
9	9	South Korea

- "directors": contains the director names
    - There are 2 columns: id, director_name

	id	director_name
1	1	Akira Kurosawa
2	2	Alejandro Amenabar
3	3	Alejandro G. Inarritu
4	4	Alfred Hitchcock
5	5	Andrew Stanton
6	6	Anthony Russo
7	7	Asghar Farhadi
8	8	Billy Bob Thornton
9	9	Billy Wilder
10	10	Brad Bird

Eventually, in a future assignment, you'll merge the "imdb" data with the "countries" data and the "directors" data. This will give you the country of origin and director for each movie.

Submission

Open the Jupyter Notebook directly in Coursera (which you will find in the item soon after this reading). The Coursera lab includes the imdb.xlsx file. To complete the assignment, complete the provided Jupyter Notebook file, following the detailed instructions in each cell. Test your submission before submitting by following the instructions on the assignment page in Coursera. When you're happy with your solutions, click the 'Submit Assignment' button in the top right.

Evaluation

Each question is worth 1 point except for Q7 - 2 points:

- 1 pt - counting the records
- 1 pt - removing the duplicates