SI 618 HW1

This homework is due September 8 right before class (3:59pm). Please turn in your Jupyter notebook (<uniqname>_si618_h1.ipynb and <uniqname>_si618_h1.html files) through Canvas.

You will use the book_data.csv obtained from https://www.kaggle.com/meetnaren/goodreads-best-books-of-2018. This dataset includes description of the best books in 2018 from Goodreads (book_data.csv) and the images of the covers *(you will not use this in this homework).*

The csv file includes the following columns:

- book_authors: The author(s) of the book, separated by '|'
- book_desc: A description of the book, as found on the Goodreads web page of the book
- book_edition: Edition of the book
- book_format: Format of the book, i.e., hardcover, paperback, etc.
- book_isbn: ISBN of the book, if found on the Goodreads page
- book_pages: No. of pages
- book_rating: Average rating given by users
- book_rating_count: No. of ratings given by users
- book_review_count: No. of reviews given by users
- book_title: Name of the book
- genres: Genres that the book belongs to; This is user-provided information
- image_url: URL of the book cover image

Please perform the following operations and turn in your Jupyter notebook titled uniqname_si618_hw1.ipynb and the corresponding html page (uniqname_si618_hw1.html) through Canvas.

- **Introduction: (10 points)**
  - Q1: Load the dataset. (1 points)
  - Q2: Find if there are any duplicates book titles. Remove duplicated from the data keeping only the first occurrence of a title. How many unique books are there? (7 points)
  - Q3: How many books have 2 authors? (2 points)

- **Length: (25 points)**
  - Q4: You will want to create a new column with the integer value of the number of pages. (If you remove rows in this process, please state why.) (10 points)
  - Q5: What is the median number of pages? (5 points)
  - Q6: What are the minimum and maximum numbers of pages? (5 points)
  - Q7: Does having more than 1 author result in a longer book on average? What is the average number of pages for books written by a single author? What is it

when there two authors? How about three authors? (We will do more careful analysis for these types of questions later. For now, we just want you to practice using some DataFrame functionalities). (5 points)

- **Ratings: (30 points)**
    - Q8: How many books have at least a rating of 4.5? (2.5 points)
    - Q9: How about at most a rating of 4? (2.5 points)
    - Q10: Discretize (i.e. round down) the ratings. The resulting ratings should have one of the following values: 1,2,3,4 or 5. (15 points)
    - Q11: For each of the discretized ratings (1,2,3,4,5), what is the average number of authors? (5 points)
    - Q12: How about the average number of reviews? (5 points)

- **Genres: (35 points)**
    - Q13: Create a new DataFrame, exploding the rows with multiple genres such that it is one row per genre/book. (15 points)
    - Q14: What is the average rating of books from different genres? What is the median? (10 points)
    - Q15: What is the average number of authors of books from different genres? What is the median? (10 points)