Assignment #1: Data table operations

By the end of this homework assignment, you should be able to examine raw CSV files as data tables and perform common operations on them in your language of choice.

See Assignment 0 for general instructions on what files to submit.

As an experiment for faster grading, submit the raw answers under Brightspace *Quizzes*. Codes and other attachments will still be submitted under Brightspace *Assignments*. **As a backup, upload a PDF file of your answers to Brightspace *Assignments* as well!**

For this assignment, we will be using data on restaurant inspection records in San Francisco. The dataset is composed of 4 files: `businesses.csv` contains information about each restaurant, `inspections.csv` contains inspection data, `violations.csv` contains violation data, and `legend.csv` contains the rating categories.

Because the focus of this assignment is on exploring data table operations through code, answer all questions assuming that the data has been cleaned (*spoiler alert: it has not!*). We will look at the data in more detail in the next assignment.

The points for each question are given in parentheses and the point breakdowns per question (if applicable) are given in brackets. The code for each question will also be given 1pt.

Read all CSV files as data tables.

1. (1pt) How many records (rows, *excluding column headers*) [0.5pt] and columns (*excluding unnamed index columns*) [0.5pt] are there for:
    - `businesses.csv`
    - `inspections.csv`
    - `violations.csv`
2. (1pt) How many duplicated records are there in each file? Count *all* duplicates – i.e., if rows A, B, and C are identical, and so are rows D and E, then write `5`.
    - `businesses.csv`
    - `inspections.csv`
    - `violations.csv`
3. (1pt) In `businesses.csv`, why is `business_id` a reasonable identifier for each restaurant while `name` is not?
4. (1pt) Each `business_id` in this dataset corresponds to a “restaurant” for all intents and purposes. How many restaurants are in:
    - `businesses.csv`

   - `inspections.csv`
   - `violations.csv`

5. (1pt) What are the top 5 restaurant names (values under the column `name`) with the most records in the file `businesses.csv`, and how many records does each one have? Present your result as a table.

6. (1pt) What are the top 5 addresses (values in the `address` column) that host the largest number of restaurants, and how many restaurants does each one host? Present your result as a table.

7. (1pt) How many cities are included in the dataset?

8. (1pt) How many restaurants are in postal code `94115`?

9. (1pt) How many restaurants lack latitude *and* longitude information?

10. (1pt) How many restaurants lack latitude *or* longitude information?

11. (1pt) Do all of the records in the violations data have a description of the violation?

12. (1pt) How many restaurants are in the business information file and have inspection data?

13. (1pt) How many restaurants have reported violations but whose names are unknown?

14. (1pt) How many restaurants have reported violations but have no inspection data?

15. (2pt) For every restaurant with inspection data, generate a *single* table where each restaurant is represented by a row, and where (a) the number of times it had an inspection [1pt] and (b) its highest inspection score [1pt] are included as columns. What values does it give for restaurants with IDs: `5634` and `2420`?

16. (2pts) Create a copy of the dataframe from `inspections.csv` but add a column `rating` that maps the inspection score to the rating given in `legend.csv`.

    - How many *unique* inspection records (rows) resulted in a rating of “Adequate”? [1pt]
    - How many *unique* inspection records resulted in a rating of “Good” or “Poor”? [1pt]

17. (2pt) Generate a table where each row corresponds to a restaurant ID, the columns are the ratings “Poor”, “Needs Improvement”, “Adequate” and “Good”, and the values are the number of inspection records corresponding to each restaurant-rating combination.

    - Output this data as a CSV file and **upload it**. There should be a header row but no line numbers. [1pt]

- How many restaurants *never* received a rating *below* “Adequate”? [1pt]