Homework 1: Structured Data

Due: 24 hours before the live session in Week 5

You can choose to complete this assignment by yourself or with a group of at most two total participants. Each person must turn the assignment in for grading, and each person must contribute to the development of the program. Use the file Donors_Data.csv.

Structured Data Processing

For purposes of this writeup, we will use examples from the Donors data file.
The main outline of your assignment is to write a program that will read in the data from a file, such as the .csv file saved from Excel. This will be in a format that is structured with lines of data representing one type of unit (i.e., one donor in the donors file). Your program will represent the data as Python data structures. You may choose for the overall structure to be one or both of the following:

- A list of dictionaries, or some combination of lists, dictionaries, and NumPy arrays

- A pandas dataframe

You will do data exploration and cleaning on this data.

The program will do some processing to convert the data to a form that will answer at least two questions as described below, and write files with the data suitable for answering each question. Graphing is optional.

Data:

You may choose a data set to work with. As a guideline, data sets should have somewhere between 500 and 4,000 lines of data with some number of columns between 4 and 50. These guidelines are not exact limits, just guidance for selecting data.
If the data comes in an Excel spreadsheet with a lot of columns, it is okay to first edit the file to remove columns that you do not need for your processing. For example, in the Donors data, you might wish to create a separate spreadsheet with only a few columns of data.

Questions:

For this assignment, at least one question that you choose to answer should look at the data in a different unit of analysis than is present in the data file. For example, instead of looking at individual donors, you could look at the donors of each of the nine income or wealth types.

Simplest example question (you should do one more complex than this):

For each wealth type, what is the average home value of all the donors of that type?

- Unit of analysis: wealth types

- Comparison: for each wealth type, compute the average home values of the neighborhoods of all the donors of that type

- Output: should be in a file with nine rows of data (you may also produce header and label rows), where each row has an income type (1–9) and the average home values

One way to increase the complexity of this particular question would be to add more items to be compared to the income types, e.g., add columns to the output with average total gifts or values of the last gifts. Another way is to introduce a more detailed unit of analysis; for example, suppose that for each income level you reported by gender, giving the average home value for both men and women in each category.

Other ideas:

Compare donors in the various zip codes with various types or amounts of giving.

Compare donors by the number of promotions with the total amount of donations and the frequency of donations.

Compare the number of months since the last donation to the donation amounts.

What to Submit:

In addition to the program that you write, you should write a small report. In it you should provide:

- Data and its source

- Description of your data exploration and data cleaning steps

- At least two clearly stated comparison questions with the unit of analysis, the comparison values, and how they are computed

- Brief description of the program

- Description of the output files

For your program, you may use any of the code developed in class as a template, but it is absolutely essential that you use appropriate variable names and that you write original comments for what your program does. Recall that good comments demonstrate your understanding of the code that you write and the problem that you are trying to solve.

Group Work:

If you choose to work in a group (of two), you may write and submit a single program, but you must process the data for two additional comparison questions. Each member of your group should write some part of the program, even if edited together later. Your report should describe the roles of the group members and who did what parts of the assignment, possibly including the data exploration, data cleaning, formulating questions to answer, and debugging.

Submit the following in a single submission (for each person):

- Report (A Word document or PDF file)
  - I must be able to find the data that you utilized (link to the website, etc.)

- Program(s)
  - Must be submitted as stand-alone Python or Jupyter Notebook files that I can execute on my own machine

- Output files (may be included in the Report)