**Assignment #2: Data Scrubbing using Python**

The goals of this assignment are to

1. Select a data source that is relevant or meaningful for you to work with.
2. “scrub” or “clean” a datafile using Python
   * *NOTE -You must use Python 3.x (not Python 2.x) in order to get credit for this assignment.*
3. parse the text data file into an MS-Excel, Open Office Calc or Google *Sheet* for testing purposes

**Part 1: Data selection and retrieval**

First, you will need to select a **datafile** to work from. For this assignment, please select any reputable data source that is of interest to you. Sample data sets of interest you might consider (but are NOT required to use) can be found on our Resources Pages under "[Data in the Wild](https://cs.nyu.edu/courses/fall19/CSCI-UA.0060-001/notes/dbw_resources_python_dataInTheWild_fa19.html)".

*NOTES:*

* Do not use data that we used in class or datasets that are from the same source or very similar.
* You may not use Python's *pandas* module for this assignment.
* You will not get credit for a project that is a copy of the programs that we wrote in class.
* If you have any questions on your data source or code, you are welcome to check with me.

**Part 2: Data retrieval and manipulation**

In addition, take care not to simply copy and re-use the code that we have written in class but rather to work with data of interest to you and resolve any programming challenges that might arise!

NOTE: *The example we did in class used very clean data. I selected that sample so that you could see the conceptual framework and the steps. I imagine and expect that you will have to write significantly more code for this assignment in order to correct, verify or otherwise manipulate individual data fields; in order to create calculated fields; in order to select specific records; and/or in order to ensure that the data fields in the output reflect consistent data types and units of measurement, etc.*

In almost all cases, you will need to “clean up” or “scrub” the data before you can use it in a spreadsheet (MS-Excel, Google Sheets, etc). If you pick data that are already in a usable format, use Python to modify the data such as removing unnecessary columns; adding computed columns; selecting only records that meet specific criteria and other modifications. Either way, this assignment asks that you modify your dataset in at least five major and discernable, different ways. (In other words, removing five columns from an otherwise suitable data file "counts" as one change rather than five changes and does not meet the full requirements for the assignment.)

Use Python to write a program to read in the datafile which you have downloaded or copied from the Internet; or your program should obtain the data directly using python's urllib module. Be sure to include appropriate exception handling so that your program does not "crash" if the data file or web page is not found or cannot be opened. Feel free to use the Beautiful Soup library in python to assist in removing markup (e.g. HTML) from the data (See <http://www.crummy.com/software/BeautifulSoup/>).

Your python program should write out a new file which is ready for data analysis. Your program should be written to work consistently with the dataset of your choice regardless of the file size (in other words, the program should be scalable and not based on strategies e.g. “for the third row … for the fourth row … ” etc.)

We have explored some of these programs in class; see our [class notes for examples](https://cs.nyu.edu/courses/fall19/CSCI-UA.0060-001/notes/dbw_readings_pythonDataScrubbing_fa19.html).

Do not select data in a JSON format for this assignment; we will look at JSON data later in the semester.

If your dataset contains less than 3,000 records/rows, please see your instructor for permission.

If your dataset is very large and therefore difficult to work with in a spreadsheet environment, you may use Python to select only specific records for the spreadsheet portion of this project; and then use the full dataset in the next or a later assignment. Be sure to indicate this in your comments in the code and in your essay.

Extra Credit is available for examples which tackle especially large and complex data tables.

**Part 3: Verify your data in a spreadsheet environment**

Import your newly scrubbed datafile into a worksheet in Microsoft *Excel*, Open Office *Calc* or Google's *Sheets*.

Double-check for "outliers" or anomolies by creating two or more graphs and running some relevant statistics (e.g. to compare the mean and median or other calculations). Include these results along with the report described in Part 4.

**Part 4: Write a 1-2 page double-spaced summary of your work**

Write about your data source; why you selected this particular dataset; why you believe this is or is not a credible data source; and what this dataset could be used for. Be sure to include the URL in this report.

Include at least one additional paragraph to describe the test(s) that you did in your spreadsheet (your graphs amd statistics) to verify that you believe these are data that can be readily used for study. If you needed to make specific changes to the data (e.g. removing temperatures of 999.9 or values of '\*\*\*\*' or other custom changes), please be sure to note these changes and your reasoning. If you believe these data are not credible or usable even after running your corrections, feel free to say so and why.

In class we discussed the goals for a data set to be COMPLETE - COHERENT - CORRECT - ACCOUNTABLE. (See the final slide in [of the class notes.)](https://cs.nyu.edu/courses/fall19/CSCI-UA.0060-001/notes/dbw_Unit3_DataScrubbing_BadData.pdf) Be sure to account for these goals in your discussion.

**What to turn in:**

Zip together and post the following files to NYU/Classes:

1. The original textfile which you downloaded from the Internet BEFORE it was modified by your python program. If you are using python's urllib module, please provide an excerpt or sample data set here.
2. Your Python program which you wrote to “scrub” the data.
   * NOTE: Be sure to include comments so that the grader and I know what your goals are in working with these data.
3. Your spreadsheet results from Part 3.
4. Your narrative of 1-2 pages from Part 4 plus additional page(s) as needed for the images of the graphs and the statistics that you considered together in one PDF file.

Include the following information in the “comments” section on NYU/Classes with your submission:

* Provide the URL where you obtained your data set.